UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO
DEPARTAMENTO DE MATEMÁTICA APLICADA E ESTATÍSTICA

Noções de Estatística e Probabilidade

Prof: Vicente Garibay Cancho

-São Carlos, 29 de Março de 2020-

Conteúdo

Capítulo 1

Introdução

1.1 Introdução e Definição de Estatística

O termo estatística é derivado da palavra "estado", em virtude de ser função tradicional dos governos centrais levantar registros da população, tais como nascimentos, mortes, profissões e entre outras atividades. Contar e medir esses fatos gera muitas classes de dados numéricos.

A estatística é concebida popularmente como colunas de cifras ou gráficos, associadas geralmente com médias. Esse conceito se aproxima muito da definição tradicional de estatística: coleção, organização, resumo e apresentação de dados numéricos. Atualmente a estatística é uma ciência (ou método) baseada na teoria de probabilidades, cujo objetivo principal é auxiliar-nos a tomar decisões ou tirar conclusões em situações de incerteza, a partir de informações numéricas.

Como um procedimento de tomada de decisões, a estatística tem uma importância crescente em vários campos, por exemplo, na produção industrial, na medicina, na nutrição e biologia, na economia, na política, na psicologia, na análise de opinião pública e outras ciências sociais, na agricultura, na física, na química e na engenharia.

1.2 Populações e Amostras

Uma **população** é o conjunto maior de indivíduos ou objetos cujo estudo nos interessa ou acerca dos quais deseja ter informações. Os elementos desse conjunto se denominam dados ou observações. As observações mensuráveis denominam-se *dados quantitativos.* Por exemplo, altura de estudantes, idade de pessoas, a duração de uma lâmpada de luz (vida útil das lâmpadas) etc. Porém, o sexo, o estado civil das pessoas, a marca de cigarros são não mensuráveis e denominam-se *dados qualitativos.* Assim, uma *população estatística* é o conjunto de observações quantitativas ou qualitativas. A população sendo infinita, portanto, é impossível ter uma informação completa sobre ela, a população sendo numerosa talvez não seja possível estudar cada um dos seus elementos. Nesses casos, recorre-se à informação proporcionada por uma parte finita da população chamada **amostra.** Em estatística é freqüente trabalhar com as chamadas **amostras aleatórias**, nas quais todos os elementos da população têm a mesma chance de serem escolhidos para compor a amostra. Uma amostra aleatória tem a propriedade de refletir as características da população da qual foi sorteada. Alguns exemplos de população

- **população:** todos os eleitores do Brasil

  amostra: 2000 eleitores entrevistados em uma pesquisa pelo IBOPE.

- **população:** todas peças produzidas por uma maquina em um dia.

  amostra: 30 peças sorteadas ao acaso da produção de um dia maquina.

- **população:** um lote de artigos recebidas por uma empresa.

amostra: 20 artigos sorteados ao acaso para inspeção.

1.3 Parâmetro e Estatística

Um **parâmetro** é uma medida que descreve alguma característica de toda a população. Para determinar seu valor, é necessário utilizar a informação da população(censo). Com isso, as decisões são tomadas com certeza absoluta.

Uma **estatística** é uma medida que é obtida a partir dos dados amostrais e descreve alguma característica de uma amostra. As decisões nesse caso, tomadas com um grau de incerteza.

1.4 Etapas do Método de Análise Estatística

A estatística, como ciência, tem como objetivo desenvolver procedimentos que permitam obter conclusões acerca dos parâmetros de uma população a partir das informações contida na amostra. Para a aplicação objetiva e pragmática dos procedimentos e técnicas estatísticas é recomendável seguir as seguintes etapas:

i) Formulação do problema e definição de um objetivo

ii) Planejamento do experimento.

iii) Recolha de dados.

iv) Análise de dados.

v) Estabelecimento de inferência estatística acerca da população (com base na informação amostral).

1.4.1 Formulação do problema

É evidente a necessidade de encarar essa etapa com máximo rigor pois dela dependerá a forma como se desenvolverão todos os passos seguintes. Nesse sentido, deve-se determinar, nessa etapa, de forma clara, quais são os problemas apresentados e quais são os objetivos da investigação.

1.4.2 Planejamento do experimento

Nessa etapa deve-se definir que informações devem ser e como são recolhidos(amostra ou censo ?). O objetivo é obter um conjunto adequado de dados que permita alcançar os objetivos da pesquisa.

1.4.3 Recolha dos dados.

Nessa etapa se recolhem-se os dados de acordo com os planos estabelecidos na etapa anterior, tendo o cuidado de controlar a qualidade da informação que se recolhe. O sucesso de uma pesquisa depende muito da qualidade dos dados recolhidos.

1.4.4 Análise de dados

Nessa etapa classifica-se a informação segundo suas características e se resume mediante a aplicação de estimadores, para a análise posterior e interpretação e interpretação dos resultados.

1.4.5 Estabelecimento de inferência estatística acerca da população

Mediante a aplicação dos métodos de inferência estatística, as conclusões da pesquisa são generalizadas à população de onde se obteve a informação

Talvez, a contribuição mais importante, dada pela estatística para a realização de inferências seja justamente, a de permitir medir a confiança nas conclusões relativas às populações, obtidas a partir da informação contida na amostra. A figura 1.1, apresenta o esquema que sintetiza o método de análise estatística.

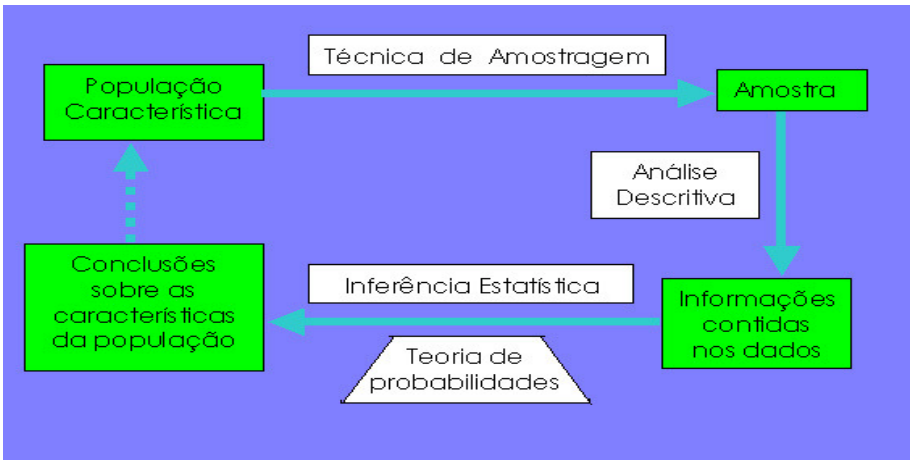


Figura 1.1: Etapas do Método de Análise Estatística.

1.5 Somatório

Dado um conjunto de observações de alguma característica ou variável X, representada por $X_1, X_2, \ldots, X_n$, a soma, $X_1 + X_2 + \cdots + X_n$, é expressado, em forma abreviada como:

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

Lê-se somatório de X_i, de $i = 1$ a $i = n$. O i denomina-se índice de adição da somatório.

1.5.1 Propriedades das somatórios

1. O número de termos da somatório, $\sum_{i=a}^{b} X_i$ é igual $b - a + 1$
2. Se c é uma constante qualquer, então $\sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i$
3. $$\sum_{i=1}^{n} (X_i + Y_i - Z_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Z_i$$
4. $\sum_{i=1}^{n} X_i = \sum_{j=1}^{n} X_j$

1.6 Somatório double

Freqüentemente em estatística deseja-se conhecer a interação entre duas variáveis, assim por exemplo, considere as 20 determinações de pressão sangüínea sistólica tomadas a um indivíduo que participa de um programa idealizado para estudar fontes e intensidade de variação de leituras da pressão sangüínea. A pressão do sangue foi medida por 4 médicos em cada uma das 5 visitas. Os dados são apresentados na seguinte tabela 1.1 Com a finalidade de

Tabela 1.1: Leituras da pressão sanguínea sistólica de um individuo tomadas em 5 visitas por 4 observadores

Número de visitas	número de médicos			
	1	2	3	4
1	118	112	116	118
2	120	116	112	112
3	114	120	112	117
4	118	116	118	116
5	118	108	122	116

ordenar linearmente essas duas classificações, utiliza-se um sistema de dois subíndices, isto é, usam-se um subíndice para o número de visitas e outro para o número de médicos. Em tais situações é freqüente utilizar as letras i e j para indicar o número da linha e o número da coluna, respectivamente. A cada observação denota-se por X_{ij} que indica o dado da i-ésima linha e j-ésima coluna. No conjunto de dados da tabela 1.1, $X_{34} = 117$, $X_{32} = 120$, por exemplo.

Considere agora, os diversos tipos de soma, por exemplo, a soma dos elementos da terceira linha é $\sum_{j=1}^{4} X_{3j}$. (na linha 3, o primeiro subíndice é fixo, o que muda é o segundo subíndice).

Para somar todos elementos da tabela 1.1, pode-se proceder de duas maneiras, primeiro somar os elementos correspondentes a cada linha e logo determinar a soma dessas somas ou somar cada coluna e logo somar essas somas.

por linhas temos:

$$\sum_{j=1}^{4} X_{1j} + \sum_{j=1}^{4} X_{2j} + \sum_{j=1}^{4} X_{3j} + \sum_{j=1}^{4} X_{4j} + \sum_{j=1}^{4} X_{5j} = \sum_{i=1}^{5}\sum_{j=1}^{4} X_{ij}$$

por colunas temos:

$$\sum_{1=1}^{5} X_{i1} + \sum_{i=1}^{5} X_{i2} + \sum_{i=1}^{5} X_{i3} + \sum_{i=1}^{5} X_{i4} = \sum_{j=1}^{4}\sum_{i=1}^{5} X_{ij}$$

No exemplo:

$$\sum_{i=1}^{5}\sum_{j=1}^{4} X_{ij} = 464 + 460 + 463 + 468 + 464 = 2319.$$

$$\sum_{j=1}^{4}\sum_{i=1}^{5} X_{ij} = 588 + 572 + 580 + 579 = 2319.$$

Em geral, suponha que a tabela 1.1, tenha n linhas e m colunas, então, soma de todos elementos da tabela é:

$$\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij}.$$

1.7 Exercícios

1. Verificar as seguintes expressões:

(a) $\sum_{i=1}^{n}[X_i(X_i+\bar{X})+(X_i-\bar{X})^2] = 2\sum_{j=1}^{n}X_j^2$, se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$.

(b) $\sum_{i=1}^{n}(X_i-\bar{X}) = 0$, se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$.

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

(d) $\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i-\bar{X})(Y_j-\bar{Y})^2 = 0$, se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ e $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n}Y_i$

(e) $\sum_{i=1}^{n}[X_i(X_i+\bar{X})-\bar{X}^2] = \sum_{i=1}^{n}X_i^2$, se $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$.

2. Na seguinte tabela tem-se a quantidade em toneladas de açúcar transportada desde os depósitos de uma distribuidora aos supermercados de Belo Horizonte.

Depósito	Supermercados		
	1	2	3
1	5	6	8
2	4	4	2
3	6	4	9
4	5	7	8
5	4	3	2

Se X_{ij} : é quantidade em toneladas de açúcar transportada desde o depósito i aos supermercados j. $i = 1,2,3,4.$ e $j = 1,2,3$ Representar em termos de somatório simplificada e determine o valor:

(a) Da quantidade total de açúcar transportada aos supermercados.

(b) Da quantidade total de açúcar transportada desde os depósitos 2 e 4 aos supermercados 1 e 3.

(c) Se os preços (em reais) por tonelada de açúcar nos supermercados 1, 2 e 3 são respectivamente: $P_1 = 450,0$, $P_2 = 500,0$ e $P_3 = 510,0$. Determine o ingresso da distribuidora para transportar aos supermercados 2 e 3.

(a) Suponha além da informação dada em (c) que os custos de transporte por tonelada desde os depósitos 1,2, 3, 4 e 5 são respectivamente: $C_1 = 1,5$, $C_2 = 0,90$, $C_3 = 1,2$, $C_4 = 1,5$ e $C_5 = 0,95$. Determine o lucro nos supermercados 1 e 3.

Capítulo 2

Análise Descritiva

2.1 Introdução

O objetivo da estatística descritiva, já identificado anteriormente, é o de representar de uma forma compreensível a informação contida nos dados. A necessidade de um esforço de classificação desses dados e de síntese da informação neles contida resulta da incapacidade que, normalmente, a mente humana tem de assimilar e interpretar conjuntos significativos de dados que sejam apresentados de uma forma desorganizada.

A forma de representar a informação contida numa amostra ou numa população depende antes de tudo, da escala na qual são expressos os dados que a integram. Por essa razão, antes de analisar as técnicas de estatística descritiva mais freqüentemente utilizadas, é apresentado uma classificação dos dados (ou variáveis).

2.2 Classificação dos Dados

Os dados podem ser classificados em **qualitativos** e **quantitativos**

2.2.1 Dados qualitativos

São aqueles dados cujos resultados não podem ser expressos em forma numérica. Esses tipos de dados classificam-se em:

Qualitativo ordinal

Para esses tipos de dados é possível estabelecer uma relação de ordem entre as possíveis categorias, por exemplo, grau de instrução de funcionários de uma empresa ($1^{\ 0}$ grau, 2^0 grau, superior), opinião de um grupo de pessoas sobre um programa de TV(ruim, regular, bom, muito bom).

Qualitativo nominal

Nesses tipos de dados não há uma relação de ordem entre as possíveis categorias. Por exemplo: cor de preferência, lugar de procedência dos estudantes de uma universidade.

2.2.2 Dados quantitativos

São aqueles cujos resultados são expressos em forma numérica e são de dois tipos:

Quantitativos discretos

São dados que tem um número finito ou infinito enumerável de possíveis valores. Usualmente são associados a processos de contagem, onde o resultado é representado mediante um número inteiro. Por exemplo; número de alunos por sala de aula, número de filhos por família na cidade de Ouro Preto, etc.

Quantitativos contínuos

São dados que têm um número infinito não enumerável de possíveis valores e são representados por números de um intervalo real. Por exemplo: Altura do aluno da turma 21, peso de crianças recém nascidas num hospital universitário etc.

2.3 Organização e Representação de Dados

2.3.1 Organização de dados qualitativos

Se os dados são qualitativos são simplesmente, agrupados segundo a freqüência e a proporção ou porcentagem de cada categoria e representados graficamente mediante barras horizontais ou verticais ou diagramas circulares (ou gráfico de pizza) .

Exemplo 2.3.1 *A 40 alunos que foram reprovados em alguma disciplina do semestre anterior. perguntado em quais disciplinas tinham sido reprovados e as respostas foram as seguintes:*

Cálculo II	*Cálculo II*	*Cálculo I*	*Álgebra*	*Estatística*	*Estatística*	*Cálculo II*
Biologia	*Química*	*Cálculo II*	*Estatística*	*Cálculo I*	*Estatística*	*Álgebra*
Álgebra	*Estatística*	*Cálculo II*	*Álgebra*	*Álgebra*	*Cálculo I*	*Cálculo I*
Estatística	*Cálculo II*	*Cálculo II*	*Cálculo II*	*Estatística*	*Cálculo I*	*Estatística*
Genética	*Mecânica*	*Economia*	*Estatística*	*Cálculo I*	*Bioquimica*	*Cálculo II*
Cálculo I	*Fisica*	*Cálculo II*	*Quimica*	*Física*		

A **freqüência absoluta** são o resultado de um processo de contagem das respostas obtidas entre os 40 alunos consultados. Assim, por exemplo, 10 alunos desaprovaram na disciplina de Cálculo II, 7 desaprovaram em cálculo I, etc. Observa-se que a soma das freqüências absolutas é igual ao número total de alunos consultados ou também chamada de tamanho da amostra a qual será denotado por n.

Suponha que um conjunto de dados qualitativos tenha k categorias (no exemplo $k = 5$) então $\sum_{i=1}^{k} f_i = n$

Considerando o número total de alunos consultados ($n = 40$ alunos), as freqüências relativas são obtidos dividindo cada freqüência absoluta por n, isto é, $f_{r_i} = \frac{f_i}{n}$. Por exemplo, para o caso cálculo II, sua freqüência relativa são obtidas da seguinte forma: $f_{r_1} = f_1/40 = 10/40 = 0,25$. Para cálculo I, $f_{r_2} = f_2/40 = 0,175$ e assim por diante.

Similarmente, as freqüências percentuais são obtidas dividindo cada freqüência absoluta por 40 e multiplicando por 100. Também é possível obter multiplicando cada freqüência relativa por 100, isto é, $p_i = \frac{f_i}{n} \times 100 = f_{r_i} \times 100$. Por exemplo, para cálculo II, $p_1 = \frac{10}{40} \times 100 = 25\%$ ou $p_1 = 100 \times f_{r_1} = 100 \times 0,25 = 25\%$ a freqüência percentual para cálculo I será: $p_2 = \frac{f_2}{40} \times 100 = \frac{7}{40} \times 100 = 17,5\%$, etc.

As freqüências relativas e percentuais têm uma interpretação similar e podem ser usadas indistintamente, por exemplo, para o caso de cálculo II, a freqüência relativa ou percentual indica que 25% dos alunos consultados desaprovaram em cálculo II. De maneira similar, são interpretados as outras freqüências relativas (ou percentuais). A vantagem do uso desse tipo de freqüências é que seu valor da informação sobre a incidência de uma resposta, sem requer do total de alunos consultados. A distribuição de freqüências do exemplo 2.3.1, é apresentado na tabela 2.1.

Para uma análise mais simples da informação é conveniente a representação dos dados mediante gráficos. Como foi dito anteriormente, existe uma grande diversidade de representações gráficas, sendo as mais simples e freqüentes

Tabela 2.1: Distribuição de alunos desaprovados numa disciplina no semestre 2003/1

Disciplina	Freqüência Absoluta f_i	Freqüência Relativa f_{r_i}	Freqüência Porcentual p_i
Cálculo II	10	0,250	25,0
Cálculo I	7	0,175	17,5
Álgebra	5	0,125	12,5
Estatística	9	0,225	22,5
Outras	9	0,225	22,5
Total	40	1,000	100

os gráficos de barras (horizontais e verticais) e os gráficos circulares (ou "pizza"). Para a elaboração do gráfico de barras é construído um sistema de eixos cartesianos XY. No eixo vertical se forma uma escala para representar a magnitude de algum tipo de freqüência; em geral, utilizam-se as freqüências percentuais. No eixo horizontal, uma escala para representar as respostas obtidas mediante barras verticais. A amplitude de cada barra é a mesma e é deixando um espaço entre cada barra. A altura de cada barra deve ser igual à magnitude da freqüência correspondente a cada um dos dados e que é representada no eixo vertical. É conveniente colocar rótulos aos eixos que permitam entender melhor a informação. O gráfico de barras verticais para os dados do exemplo 2.3.1, é apresentado na figura 2.1.

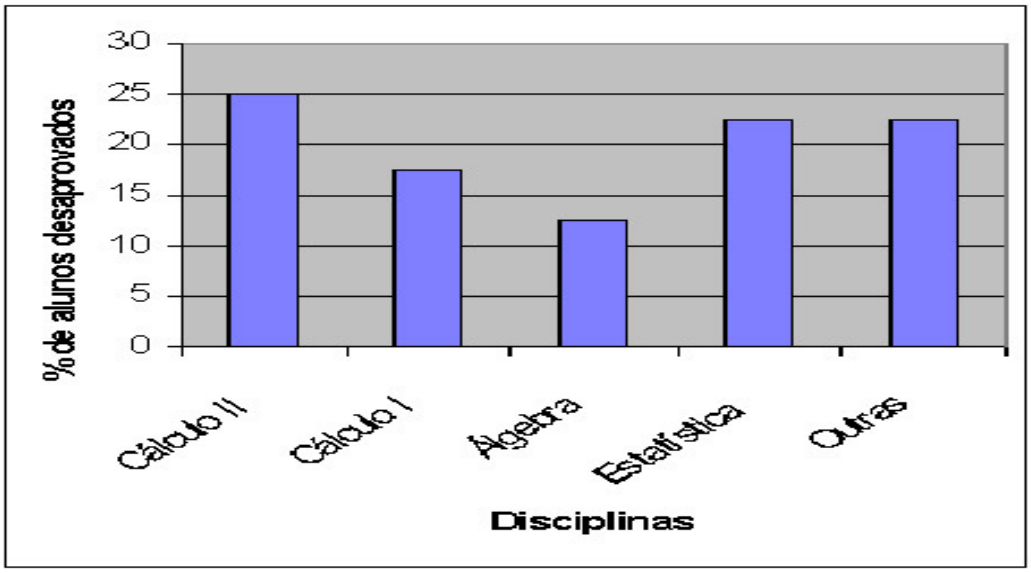


Figura 2.1: Distribuição de alunos desaprovados no semestre 2003/1.

Para a elaboração de gráficos circulares devem ser calculados os ângulos de cada região circular que são associados a cada resposta. Para isto, multiplica-se cada freqüência relativa por 360. Por exemplo, para o caso de cálculo II, o ângulo da região circular utilizada para representar essa resposta é $\alpha_1 = 360 f_{r_1} = 360 \times 0,25 = 90^o$. Uma vez determinados os ângulos das regiões o gráfico é construído partindo do eixo de referência, usualmente o eixo associado 0^o ou 90^o e representando as regiões circulares uma a uma. Para uma adequada identificação é conveniente colocar um rótulo de identificação ao lado de cada região e a freqüência que correspondente a cada resposta. O gráfico circular para os dados do exemplo 2.3.1 é apresentado na figura 2.2.

Podem ser utilizados, também, efeitos tridimensionais para obter uma melhor apresentação. Por exemplo, o gráfico anterior pode ser mostrado como:

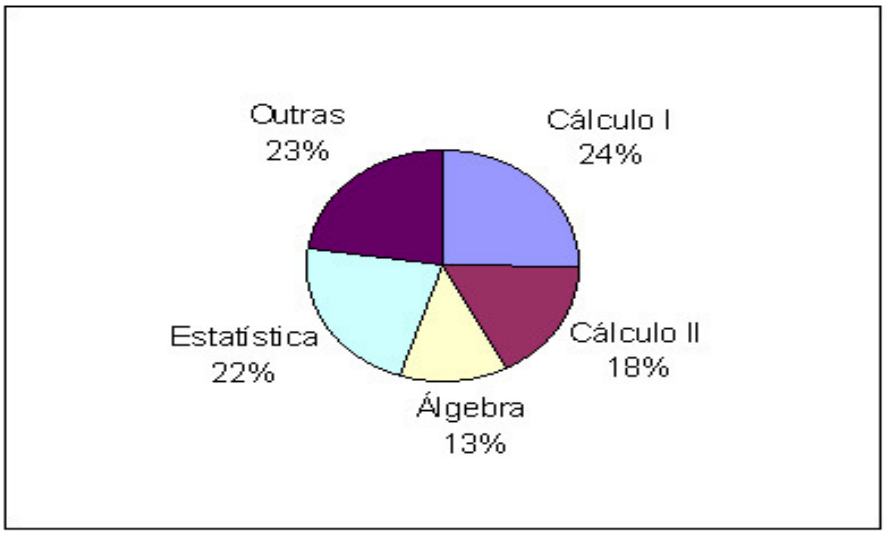


Figura 2.2: Distribuição de alunos desaprovados no semestre 2003/1

Para organizar e representar dados qualitativos **ordinais**, geralmente, ordena-se as categorias dos dados em ordem de maior a menor hierarquia.

2.3.2 Organização de dados quantitativos

Quantitativos discretos

Para dados quantitativos discretos cujo número de resultados possíveis não é grande (não é maior que 12 ou 15), a informação pode ser classificada e representada diretamente sem perda de informação da mesma.

Nesses casos, primeiro ordena-se a informação segundo sua magnitude e, em seguida obtém-se as *freqüências absolutas* associadas a cada valor observado. As freqüências relativas e percentuais são obtidas de forma similar à descrita na seção anterior.

Para representar, graficamente um conjunto de dados quantitativos discretos é construído um sistema de eixos cartesianos XY. No eixo vertical, utiliza-se uma escala para representar a magnitude de algum tipo de freqüência; em geral consideram-se as freqüências percentuais. No eixo horizontal, utiliza-se uma escala para representar os valores observados. Logo, para cada um dos dados na escala horizontal levanta-se um segmento de reta vertical cuja magnitude é determinada pela freqüência correspondente.

Exemplo 2.3.2 *Com a finalidade de estudar o número de emergências que chegam a um hospital por dia, o administrador de um hospital selecionou uma amostra 50 dias, ao acaso, dos arquivos de um hospital. Os dados são os seguintes:*

2	*2*	*1*	*1*	*3*	*4*	*6*	*7*	*0*	*0*	*0*	*1*	*1*	*1*	*2*	*2*	*1*	*0*
0	*0*	*0*	*5*	*5*	*1*	*2*	*2*	*1*	*1*	*1*	*2*	*1*	*3*	*4*	*4*	*4*	*1*
2	*1*	*1*	*1*	*2*	*2*	*2*	*4*	*5*	*0*	*0*	*0*	*2*	*1*				

Ao ordenar os dados em ordem crescente tem-se:

0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
2	2	3	3	4	4	4	4	4	5	5	5	6	7				

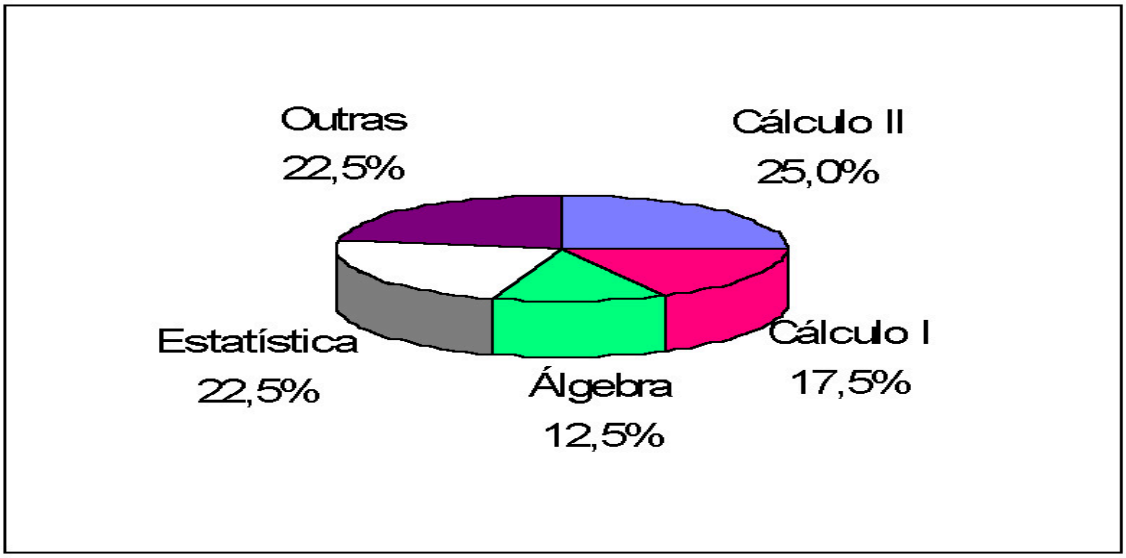


Figura 2.3: Distribuição de alunos desaprovados no semestre 2003/1

Tabela 2.2: Distribuição de freqüências do número de emergências atendidas pelo hospital

Número de emergências X_i	Frequência Absoluta f_i	Freqüência Relativa f_{r_i}	Freqüência Percentual p_i
0	10	0,20	20
1	16	0,32	32
2	12	0,24	24
3	2	0,04	4
4	5	0,10	10
5	3	0,06	6
6	1	0,02	2
7	1	0,02	2
Total	50	1,00	100

De maneira similar ao exemplo 2.3.1, as freqüências absolutas são o resultado de um processo de contagem das respostas obtidas nos 50 dias observados. Assim, por exemplo, em 12 dias (em cada um dos 12 dias) observou-se que o número de emergências atendidas pelo hospital foi igual 2, que em *dois* dias observou-se que o número de emergências foi igual a 3, etc. Na tabela 2.2, tem-se a correspondente distribuição de freqüências. E, na figura 2.4, é mostrada a representação gráfica dos dados do exemplo 2.3.2.

Quantitativos contínuos

Quando os dados em estudo são do tipo quantitativo contínuo, que assume muitos valores distintos, é conveniente agrupá-los em intervalos de classe. Mesmo correndo o risco de perder algum detalhe manifestado na ordenação de valores individuais, há vantagem em resumir os dados originais em uma distribuição de freqüência, onde os valores observados não mais aparecerão individualmente, mas agrupados em classe.

Quando se considera intervalos de classe de igual amplitude, o procedimento é o seguinte:

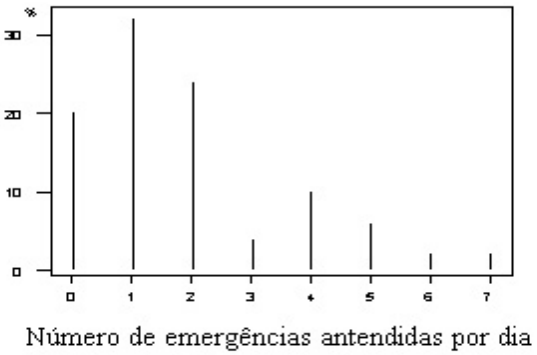


Figura 2.4: Distribuição do número de emergências atendidas pelo hospital

1. Deve-se estabelecer o número de intervalos de classe (k) que se vai utilizar. Tal número é recomendado que esteja entre 5 e 15. Não existe uma regra fixa para determinar o número ótimo de intervalos. O critério do pesquisador tem um papel importante na determinação do mesmo.

   Como referência, pode-se utilizar a regra de **Surges**, que indica que o número de intervalos de classe é dado por:

   $$k = 1 + 3,3 \log_{10}(n),$$

   onde n é o número de observações (ou tamanho da amostra). [O valor k deve ser arredondado ao número inteiro mais próximo].

2. Determinar o comprimento ou amplitude (A) dos dados, isto é,

   $$A = X_{max} - X_{min},$$

   onde X_{max} é o valor da observação de maior magnitude e X_{min} a observação de menor magnitude.

3. Determinar a *amplitude de cada intervalo de classe* (h):

   $$h = \frac{A}{k}$$

   quando o quociente A/k não é exato o valor de h deve ser arredondado ao valor superior mais próximo, segundo o número de cifras decimais dos dados.

4. Gerar os limites dos intervalos. Para o primeiro intervalo considere como limite inferior o valor da observação de menor magnitude, isto é, $LI_1 = X_{min}$.

   Os limites inferiores dos outros intervalos são obtidos da seguinte forma: $LI_i = LI_{i-1} + h$, para $i = 2, 3, \ldots, k$. Os limites superiores dos intervalos são obtidos: $LS_i = LI_{i+1}$, para $i = 1, 2, \ldots, k-1$ ou $LS_i = LS_{i-1} + h$, para $i = 2, 3, \ldots, k$.

5. Cada um dos intervalos é da forma $[LI_i; LS_i)$, isto é, fechado na esquerda e aberto na direita.

6. Obter as **marcas de classe** ou ponto médio (X_i') que são valores representativos da informação contida num intervalo. Numericamente são obtido como a média dos limites inferior e superior do intervalo. Isto é,

   $$X_i' = \frac{LI_i + LS_i}{2} = LI_i + \frac{h}{2}, \quad i = 1, \ldots, k$$

7. Uma vez definidos os intervalos de classe, o passo seguinte consiste em classificar cada observação em um dos ditos intervalos e determinar as **freqüências absolutas**, isto é, o número de observações que estão dentro de cada intervalo de classe. A partir dessas freqüências, as freqüências relativas e percentuais correspondentes a cada intervalo de classe são obtidos. Além disso, para o caso de dados quantitativos contínuos pode-se determinar a *densidade de freqüências* ou simplesmente *densidade* (d_i) definido pelo quociente das freqüências relativas (ou freqüências percentual) e amplitude de intervalo de classe, isto faz com que a área total do histograma seja igual a *um* (ou 100%).

8. Adicionalmente, quando se dispõe de dados quantitativos contínuos é conveniente obter as freqüências acumuladas procedendo da seguinte forma:

   (a) Freqüência acumulada absoluta (F_i):

   $$F_i = \sum_{j=1}^{i} f_j = f_1 + f_2 + \cdots + f_i = F_{i-1} + f_i;$$

   (b) Freqüência acumulada relativa (F_{r_i}):

   $$F_{r_i} = \sum_{j=1}^{i} f_{r_j} = f_{r_1} + f_{r_2} + \cdots + f_{r_i} = F_{r_{i-1}} + f_{r_i};$$

   (c) Freqüência acumulada percentual (P_i):

   $$P_i = \sum_{j=1}^{i} f_{r_j} = p_1 + p_2 + \cdots + p_i = P_{i-1} + p_i;$$

   (d) Densidade acumulada (D_i):

   $$D_i = \sum_{j=1}^{i} d_j = d_1 + d_2 + \cdots + d_i = D_{i-1} + d_i;$$

É necessário levar em conta que as freqüências estão associadas aos intervalos e não às observações, como foi considerado anteriormente para dados qualitativos e quantitativos discretos.

Para representar graficamente, a informação pode ser usada qualquer tipo de freqüência. Em especial, recomenda-se utilizar a freqüência relativa ou percentual que permite analisar a informação independente do número de observações. Além disso, é possível comparar os resultados com os obtidos em estudos similares sempre que os intervalos de classe forem iguais, ou, ao menos, similares.

O procedimento descrito anteriormente pode ser aplicado também quando se tem dados quantitativos discretos cujo número de resultados possíveis é grande (maior que 20) e sua representação gráfica, através dos procedimentos descritos na seção anterior não é apropriada.

Exemplo 2.3.3 *Os seguintes dados representam a quantidade de hemoglobina (Hb) em g/dl encontrados em 40 animais expostos a um produto tóxico.*

5,2	*10,2*	*7,0*	*7,1*	*10,2*	*8,3*	*9,4*	*9,2*	*5,4*	*8,1*
6,5	*7,1*	*6,6*	*7,8*	*6,8*	*7,2*	*8,4*	*9,6*	*8,7*	*7,3*
8,5	*5,7*	*6,4*	*10,1*	*8,2*	*9,0*	*7,8*	*8,2*	*7,8*	*6,6*
5,3	*6,2*	*9,1*	*8,6*	*7,0*	*7,7*	*8,3*	*7,5*	*9,8*	*7,5*

Para obter a tabela de distribuição de freqüências, procede-se da seguinte maneira:

$n = 40$, $k = 1 + 3,3 \log_{10}(40) = 6,2868 \approx 6$

$A = X_{\max} - X_{\min} = 10,2 - 5,2 = 5,0$,

$h = \frac{A}{k} = \frac{5}{6} = 0,8333 \approx 0,9$ (arredondamento por excesso a uma decimal, ou seja, à mesma precisão dos dados),

$$LI_1 = X_{min} = 5,2$$
$$LI_2 = LI_1 + h = 5,2 + 0,9 = 6,1 \quad LS_1 = LI_2 = 6,1 \quad X'_1 = \frac{LI_1+LS_1}{2} = 5,65$$
$$LI_3 = LI_2 + h = 6,1 + 0,9 = 7,0 \quad LS_2 = LI_3 = 7,0 \quad X'_2 = \frac{LI_1+LS_1}{2} = 6,55$$

De maneira similar obtém-se os outros limites de classe e suas marcas de classe.

Construídos os intervalos de classe, classificam-se as observações para serem obtidas as freqüências absolutas, relativas e densidades de forma similar ao indicado acima.

Para obter as freqüências acumuladas procede-se da seguinte forma:

$$F_1 = f_1 = 4 \qquad F_{r_1} = F_1/40 = 0,10 \quad P_1 = 100F_{r_1} = 10$$
$$F_2 = f_1 + f_2 = 4 + 6 = 10 \quad F_{r_2} = F_2/40 = 0,25 \quad P_2 = 100F_{r_2} = 25$$

De forma similar procede-se com os outros intervalos. Com os resultados anteriores é obtida a tabela 2.3, que contem a distribuição de freqüências para esse exemplo.

Tabela 2.3: Distribuição da quantidade de hemoglobina de 40 animais

Quantidade de Hb	X'_i	f_i	f_{r_i}	p_i	$d_i = \frac{p_i}{h}$	F_i	F_{r_i}	P_i
$5,2 \vdash 6,1$	5,65	4	0,100	10,0	11,11	4	0,100	10,0
$6,1 \vdash 7,0$	6,55	6	0,150	15,0	16,67	10	0,25	25,0
$7,0 \vdash 7,9$	7,45	12	0,300	30,0	33,33	22	0,550	55,0
$7,9 \vdash 8,8$	8,35	9	0,225	22,5	25,00	31	0,775	77,5
$8,8 \vdash 9,7$	9,25	5	0,125	12,5	13,89	36	0,900	90,0
$9,7 \vdash 10,6$	10,15	4	0,100	10,0	11,11	40	1,000	100,0
Total		40	1,00	100,0				

Histograma de freqüência

Primeiramente é construído um sistema de eixos cartesianos XY. No eixo vertical, é usada uma escala para representar a magnitude do tipo freqüência. Em geral, utilizam-se as freqüências relativas ou percentuais ou densidades. No eixo horizontal é usada uma escala para representar os intervalo de classe. Logo, para cada intervalo de classe na escala horizontal é construído um retângulo cuja altura é determinada pela freqüência usando. Por exemplo, com as freqüências percentuais da tabela 2.3, é obtida a seguinte representação gráfica:

Polígono de freqüências

No sistema de eixos cartesianos XY, no eixo vertical é usada uma escala para representar a magnitude de algum tipo de freqüência. Em geral, consideram-se as freqüências relativas ou percentuais. No eixo horizontal é usada uma escala para os valores da variável em estudo. Logo, plotam-se os pontos (X'_i, f_{r_i}), $i = 1, \ldots, k$. É considerando, também, dois intervalos adicionais: um anterior ao primeiro e outro posterior ao último intervalo de classe, cada um deles com uma freqüência zero. Por último, os pontos plotados são unidos por uma linha reta obtendo, assim, um polígono de freqüências. Por exemplo, com as freqüências percentuais da tabela 2.3 tem-se:

Polígono de freqüências acumuladas (ogiva)

No sistema de eixos cartesianos XY, no eixo vertical é usada uma escala para representar a freqüência acumulada. Em geral, consideram-se as freqüências relativas ou percentuais. No eixo horizontal é usada uma escala para os valores da variável em estudo. Logo, plotam-se os pontos (LS_i, F_{r_i}), $i = 1, \ldots, k$. É considerando, adicionalmente, o ponto (LI_1, F_{r_0}), com $F_{r_0} = 0$. Por último, unem-se os pontos plotados obtendo um polígono de freqüências

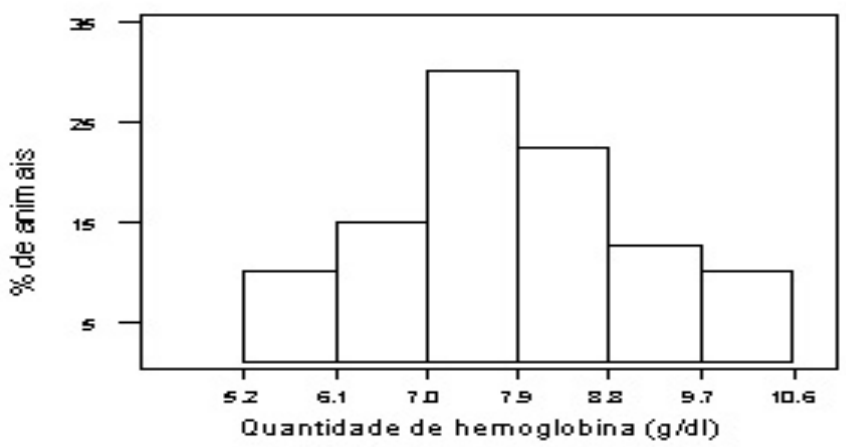


Figura 2.5: Distribuição da quantidade de hemoglobina de animais expostos a um produto tóxico

acumuladas (ogiva). Por exemplo, com as freqüências percentuais da tabela 2.3, tem-se a ogiva mostrada na figura 2.7

2.4 Medidas de Posição

Na seção anterior, foi apresentada a forma de representar a informação contida em conjunto de dados populacionais ou amostrais mediante tabelas de freqüências e gráficos. Essa informação constitui a informação básica do problema em estudo. Mas, é conveniente apresentar, além dos dados, medidas que mostrem a informação de maneira resumida. As medidas de posição ou tendência central, definidas nesta seção, são usadas para indicar um valor que tende a resumir ou representar melhor um conjunto de dados. As três medidas mais usadas são a média, a mediana e a moda.

2.4.1 Média

A média de um conjunto de observações é definida como a soma de todas as observações dividida pelo número de observações. Isto é,

$$\text{Média populacional} \quad : \quad \mu = \frac{1}{N}\sum_{i=1}^{N} X_i \tag{2.1}$$

$$\text{Média Amostral} \quad : \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.2}$$

onde

X_i: Valor da i-ésima observação da variável em estudo.

N: Tamanho da população.

n: Tamanho da amostra.

Essa medida de posição apresenta a desvantagem de ser fortemente influenciada por valores discrepantes, isto é, valores muito pequenos ou muito elevados. Portanto, nesse caso essa medida já não será um valor representativo do conjunto de dados.

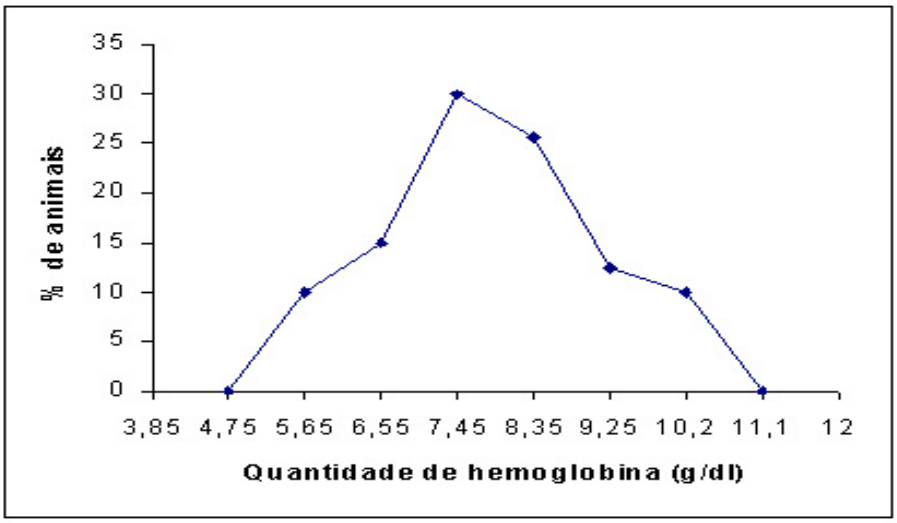


Figura 2.6: Polígono de freqüências para a quantidade de hemoglobina de animais expostos a um produto tóxico.

Exemplo 2.4.1 *Sejam as notas de quatro provas de um estudante:* $X_1 = 8.3$, $X_2 = 9.4$, $X_3 = 9.5$, $X_4 = 8,6$. *Determinar a nota média.*

$$\bar{X} = \frac{1}{4}\sum_{i=1}^{n} X_i = \frac{8,3 + 9,4 + 9,5 + 8.6}{4} = 8,95$$

Propriedades

1. A soma dos desvios das observações em relação à média é igual a zero. Isto é,

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

2. A soma de quadrados dos desvios das observações em relação à média é mínima, ou seja,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2, \text{ é um valor minímo.}$$

   Isto é,

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \leq \sum_{i=1}^{n}(X_i - k)^2,\ k \in R.$$

3. Para $k \neq 0 \in R$.
   - Se $Y_i = X_i \pm k$, então $\bar{Y} = \bar{X} \pm k$,
   - Se $Y_i = kX_i$, então $\bar{Y} = k\bar{X}$,
   - Se $Y_i = \frac{X_i}{k}$, então $\bar{Y} = \frac{\bar{X}}{k}$,

A demonstração dessas propriedades fica com exercício paro o leitor.

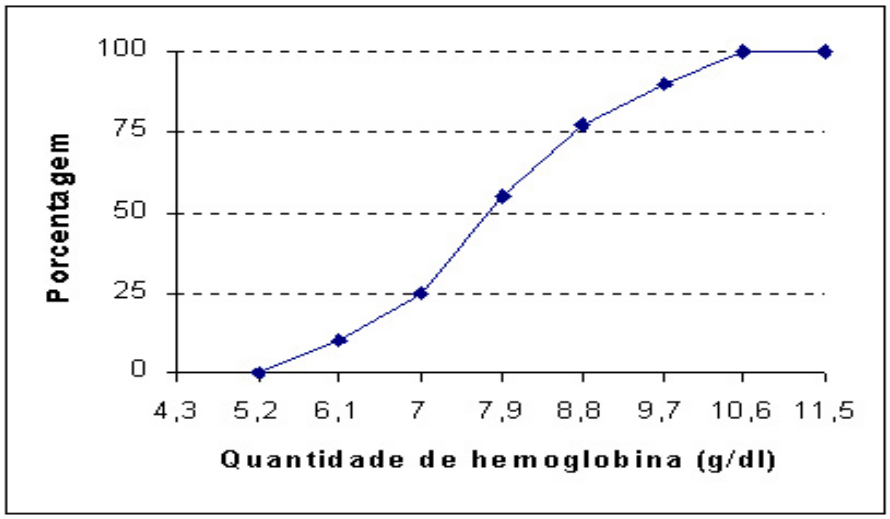


Figura 2.7: Polígono de freqüências acumuladas (ogiva) para a quantidade de hemoglobina de animais expostos a um produto tóxico

Quando tem-se dados **quantitativos contínuos** agrupados em uma tabela de distribuição de freqüências (TDF), a média pode ser calculada da seguinte forma:

$$\text{Média populacional} \quad : \quad \mu = \frac{1}{N}\sum_{i=1}^{k} f_i X_i' \tag{2.3}$$

$$\text{Média Amostral} \quad : \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{k} f_i X_i' = \sum_{i=1}^{k} f_{r_i} X_i' \tag{2.4}$$

onde

X_i': O i-ésima marca de classe da variável em estudo.

f_i: Freqüência absoluta do intervalo i.

f_{r_i}:Freqüência relativa do intervalo i.

k: Número de intervalos de classe.

N: Tamanho da população.

n: Tamanho da amostra.

Para dados **quantitativos discretos** em uma TDF a média é:

$$\text{Média populacional} \quad : \quad \mu = \frac{1}{N}\sum_{i=1}^{k} f_i X_i = \sum_{i=1}^{k} f_{r_i} X_i \tag{2.5}$$

$$\text{Média Amostral} \quad : \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{k} f_i X_i = \sum_{i=1}^{k} f_{r_i} X_i \tag{2.6}$$

onde

X_i: Valor observado i da variável em estudo.

f_i: Freqüência absoluta do valor observado i

f_{r_i}:Freqüência relativa do valor observado i.

k: Número de valores da variável em estudo.

Exemplo 2.4.2 *Considere os dados do exemplo 2.3.3, que representam a quantidade de hemoglobina (Hb) em g/dl encontrados em 40 animais expostos a um produto tóxico.*

5,2	*10,2*	*7,0*	*7,1*	*10,2*	*8,3*	*9,4*	*9,2*	*5,4*	*8,1*
6,5	*7,1*	*6,6*	*7,8*	*6,8*	*7,2*	*8,4*	*9,6*	*8,7*	*7,3*
8,5	*5,7*	*6,4*	*10,1*	*8,2*	*9,0*	*7,8*	*8,2*	*7,8*	*6,6*
5,3	*6,2*	*9,1*	*8,6*	*7,0*	*7,7*	*8,3*	*7,5*	*9,8*	*7,5*

(a) Achar a quantidade média de hemoglobina.

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} X_i}{n} \\ &= \frac{5,2 + 10,2 + \cdots + 7,5}{40} = \frac{311.4}{40} = 7,785\ g/dl.\end{aligned}$$

Logo, a quantidade média de hemoglobina em animais expostos a um produto tóxico é 7,785 g/dl

(b) Obtenha a tabela de distribuição de freqüências, e, em seguida, obtenha a quantidade média de hemoglobina dos 40 animais.

No exemplo 2.3.3 da seção anterior gerou-se a seguinte tabela de distribuição de freqüências da quantidade de hemoglobina em animais expostos a certo tóxico.

Quantidade de Hb	X_i'	f_i	f_{r_i}	F_i	F_{r_i}
$5,2 \vdash 6,1$	5,65	4	0,100	4	0,100
$6,1 \vdash 7,0$	6,55	6	0,150	10	0,25
$7,0 \vdash 7,9$	7,45	12	0,300	22	0,550
$7,9 \vdash 8,8$	8,35	9	0,225	31	0,775
$8,8 \vdash 9,7$	9,25	5	0,125	36	0,900
$9,7 \vdash 10,6$	10,15	4	0,100	40	1,000
Total		40	1,00		

Aqui, $k = 6$ e $n = 40$. Dessa forma,

$$\begin{aligned}\bar{X} &= \frac{\sum_{i=1}^{n} X_i' f_i}{n} \\ &= \frac{(5,65)(4) + (6,55)(6) + \cdots + (10,15)(4)}{40} = \frac{313.3}{40} = 7,8325\ g/dl.\end{aligned}$$

Os resultados anteriores (obtidos em (a) e (b)) não são iguais. Isto porque em (b) foram usadas as marcas de classe como valores representativos das observações. Quando tem-se dados agrupados em TDF, a média é obtida assumindo que a marca de classe é igual à média das observações classificadas em cada intervalo. Obviamente, na prática, isto ocorre raras vezes e, portanto, o valor obtido é uma aproximação do valor da média obtida como a soma de cada uma das observações.

Média ponderada

A média ponderada de um conjunto de observações $X_1, \ldots, X_n$, com pesos ou ponderações $W_1, \ldots, W_n$, é definida como:

$$\bar{X}_p = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i} = \frac{W_1 X_1 + \cdots + W_n X_n}{W_1 + \cdots + W_n}$$

Exemplo 2.4.3 *Suponha que os custos de produção e as quantidades produzidas por três filiais A, B e C de uma empresa são:*

Filial	*Custo de produção (X_i) unidades monetárias (u.m)*	*Quantidade produzida (W_i) (número de unidades)*
A	*1,20*	*500*
B	*1,60*	*200*
C	*1,05*	*900*

O custo médio de produção por unidade produzida para a empresa em seu conjunto é:

$$\bar{X}_p = \frac{(500)(1,20) + (200)(1,60) + (900)(1,05)}{500 + 200 + 900} = \frac{1865}{1600} = 1,1656\ (u.m)$$

Esse valor indica que o custo médio de produção por artigo para a empresa é de 1,1656 unidades monetárias por cada unidade produzida. Se, ao invés dessa média, fosse calculada a média aritmética,

$$\bar{X} = \frac{1,20 + 1,60 + 1,05}{3} = \frac{3,85}{3} = 1,2833\ (u.m)$$

Esse valor indicaria que o custo de produção por artigo das filiais é de 1,2833 unidades monetárias, supondo de que as três filiais produzem a mesma quantidade de artigos. Para nosso exemplo essa suposição não é verdadeira.

2.4.2 Média geométrica

A média geométrica de um conjunto de n observações positivas $X_1, \ldots, X_n$ define-se como:

$$\bar{X}_G = (X_1 \times X_2 \times \cdots \times X_n)^{1/n}$$

Essa média é usada na elaboração de números índices e para o cálculo de taxa média de variação.

Exemplo 2.4.4 *Suponha que uma fábrica teve um incremento em sua produção de: 15% no ano 1998, 10% em 1999 e 16% em 2001. Achar o crescimento médio anual.*

$$\bar{X}_G = ((1,15)(1,10)(1,16))^{1/3} = 1,136361.$$

Esse resultado indica que a produção é incrementada anualmente a um ritmo médio de 13,6461%.

2.4.3 Média harmônica

A média harmônica de n observações $X_1, \ldots, X_n$ é definida como:

$$\bar{X}_H = \frac{n}{\frac{1}{X_1} + \cdots + \frac{1}{X_n}}.$$

Essa média tem a particularidade de que os valores discrepantes a afetam em menor intensidade as outras médias.

Exemplo 2.4.5 *Suponha que um automóvel percorre os primeiros 10 quilômetros a 30 km/h e os outros 10 km a 60 km/h, a primeira vista pareceria que a velocidade média de 30 e 60 km/h é de 45 km/h. Mas esse tipo de medida é definido na Física como a distância total percorrida dividida pelo tempo total empregado para percorre-la. Como a distância total é 20 quilômetros e tempo total é $\frac{10}{30} + \frac{10}{60}$ hora. Daí tem-se que a velocidade média é:*

$$\bar{V} = \frac{20}{\frac{10}{30} + \frac{10}{60}} = \frac{120}{3} = 40\ km/h$$

É interessante observar que essa média pode ser calculada como uma média harmônica de 30 e 60 , isto é:

$$\bar{X}_H = \frac{2}{\frac{1}{30} + \frac{1}{60}} = 40\ km/h.$$

2.4.4 Mediana (Md)

É uma medida de posição que divide o conjunto de observações, previamente ordenadas de acordo a sua magnitude (crescente ou decrescente), em dois grupos de tal modo que 50% das observações são menores que a mediana e os outros 50% são maiores.

Suponha que $Y_1, Y_2, \ldots, Y_n$ seja um conjunto de n observações ordenadas em forma crescente, isto é, $Y_1 \leq Y_2 \leq \cdots \leq Y_n$. A mediana definida como

$$Md = \begin{cases} Y_{\frac{n+1}{2}}, & \text{se } n \text{ impar} \\ \frac{Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}}{2}, & \text{se } n \text{ par} \end{cases}$$

Exemplo 2.4.6 *Consideram-se duas amostras constituídas pelos dados apresentados a seguir e já ordenadas:*

a)$Y_1 = 2,0$,$Y_2 = 3,2$, $Y_3 = 4,5$, $Y_4 = 4,6$ $n = 4$; *é par então* $Md = \frac{Y_{\frac{4}{2}} + Y_{\frac{4}{2}+1}}{2} = \frac{Y_2 + Y_3}{2} = 4,5$

b)$Y_1 = 2$,$Y_2 = 3$, $Y_3 = 5$, $Y_4 = 6$, $Y_5 = 10$; $n = 5$ *é ímpar então* $Md = Y_{\frac{5+1}{2}} = Y_3 = 5$.

Propriedades

1. A soma dos desvios das observações em relação à mediana é mínima, ou seja,

$$\sum_{i=1}^{n} |X_i - Md|, \text{ é mínima}$$

Isto é,

$$\sum_{i=1}^{n} |X_i - Md| \leq \sum_{i=1}^{n} |X_i - h|, \ h \in R.$$

2. Para $k \neq 0 \in R$.
   - Se $Y_i = X_i \pm k$, então $Md_Y = Md_X \pm k$,
   - Se $Y_i = kX_i$, então $Md_Y = kMd_X$,
   - Se $Y_i = \frac{X_i}{k}$, então $Md_Y = \frac{Md_X}{k}$,

A mediana para dados **quantitativos contínuos** agrupados em TDF é obtida da seguinte forma:

$$Md = LI_i + \left[\frac{n/2 - F_{i-1}}{f_i}\right] h$$

onde

i :	é classe mediana, posição $(n+1)/2$. a classe mediana é o intervalo de classe onde na coluna das F_i superou o 50% dos dados.
LI_i :	limite inferior da classe mediana.
F_{i-1} :	freqüência acumulada absoluta da classe anterior à classe mediana.
f_i	freqüência absoluta da classe mediana

Exemplo 2.4.7 *Considerando os dados da TDF do exemplo 2.4.2, o intervalo que contém a classe mediana é* $i = 3$ *uma vez que a freqüência absoluta dessa classe é maior que 50% dos dados (maior a 20). Portanto*

$$m_e = LI_3 + \left(\frac{n/2 - F_2}{f_2}\right) h = 7,0 + \left(\frac{20 - 10}{12}\right)(0,9) = 7,75 \ g/l$$

Esse resultado indica que 50% dos animais expostos a um certo tóxico têm quantidades de hemoglobina menor que 7,75 g/dl e os outros 50% dos animais observados têm quantidades de hemoglobina superior a 7,75 g/dl.

2.4.5 Moda

A moda de um conjunto de observações é definida como o valor, classe ou categoria que ocorre com maior freqüência. A moda populacional é denotada por Mo e a moda amostral denotada por mo.

Exemplo 2.4.8 *Têm-se as seguintes observações amostrais:*

a)5, 8, 7, 9, 5, 4, 6.

b) 5, 8, 5, 9, 6, 5, 4, 9.

para (a) 4,5,5,6,7,8,9, então $mo = 5$

para (b) 4, 5, 5, 5, 6, 8, 9, 9, então $mo_1 = 5$ *e* $mo_2 = 9$

Propriedades

1. A moda pode não existir, ou pode existir mais de uma moda.
2. Aplica-se tanto para dados do tipo qualitativo quanto para do tipo quantitativo.
3. A moda é uma medida de tendência central instável e é difícil de estimar.

A moda para dados **quantitativos contínuos** agrupados em TDF é obtida da seguinte forma:

$$mo = LI_i + \left[\frac{d_1}{d_1 + d_2}\right] h$$

onde

i :	classe modal. A classe modal é idenditificada pela freqüência absoluta (f_i) com maior valor.
LI_i :	limite inferior da classe modal.
d_1 :	é a diferença entre a frequência absoluta da classe modal e freqüência absoluta anterior, ou seja, $d_1 = (f_i - f_{i-1})$.
d_2	é a diferença a freqüência absoluta da classe modal e freqüência absoluta posterior à classe modal, ou seja, $d_2 = (f_i - f_{i+1})$.

Exemplo 2.4.9 *Considerando os dados da TDF do exemplo 2.4.2, o intervalo que contém a classe modal é* $i = 3$ *uma vez que é o intervalo de classe de maior freqüência absoluta* ($f_3 = 12$). *Portanto,* $i = 3$, $d_1 = f_3 - f_2 = 12 - 6 = 6$ *e* $d_2 = f_3 - f_4 = 12 - 9 = 3$

$$mo = LI_i + \left[\frac{d_1}{d_1 + d_2}\right] h = 7,0 + \left[\frac{6}{6+3}\right] (0,9) = 7,6\ g/dl.$$

Esse valor indica que a quantidade de hemoglobina mais freqüente entre os animais observados estão ao redor de 7,6 g/dl.

2.4.6 Percentil e quartil

A mediana seja de uma população ou de uma amostra divide o conjunto de dados em duas partes iguais. Também é possível dividi-lo em mais de 2 partes.

Quando se divide um conjunto ordenado de dados em quatro partes iguais, os pontos da divisão são conhecidos como **quartil;** o primeiro quartil, Q_1, é o valor que divide aproximadamente, a quarta parte (25%) das observações abaixo dele, e os 75% restantes, acima dele. O segundo quartil é exatamente a mediana (Md). O terceiro quartil ou quartil inferior, Q_3, tem aproximadamente os três quartos (75%) das observações debaixo dele.

Exemplo 2.4.10 *A seguir são apresentada* 20 *observações do tempo de falha, em horas de um material, 204 228 252 300 324 444 624 720 816 912 1176 1296 1392 1488 1512 2520 2856 3192 3528 3710*

A mediana, já que $n = 20$ é pa é:

$$Md = Q_2 = \frac{912 + 1176}{2}$$

O primeiro quartil deve ter 25% dos dados abaixo dele ou, nesse exemplo, pelo menos 5 observações abaixo dele, e 75% dos dados acima dele ou menos de 15 de observações de seu valor acima dele. A quinta e sexta observação satisfazem essa definição de modo que Q_1 é definido como a média dessas observações

$$Q_1 = \frac{324 + 444}{2} = 384$$

Similarmente, o terceiro quartil deve ter s 75% dos dados abaixo dele ou pelo menos 15 observações abaixo de seu valor, e 25% dos dados acima ou pelo menos 5 observações acima dele. As observações 15 e 16 satisfazem essa definição. Portanto,

$$Q_1 = \frac{1512 + 2520}{2} = 2016$$

Definição 2.4.1 (Percentil) *O percentil P_p, é um valor que divide um conjunto de observações ordenados de forma crescente (ou decrescente) em duas partes, o* $100p\%$ *dessas observações com valores inferiores (superiores) a P_p, e o* $100(1-p)\%$ *com valores superiores (inferiores) a P_p. Sendo* $0 < p < 1$.

Observe que:

$Q_1 = P_{0,25}$

$Q_3 = P_{0,75}$

O percentil P_p para dados **quantitativos contínuos** agrupados em TDF é obtido da seguinte forma:

$$P_p = LI_i + \left[\frac{np - F_{i-1}}{f_i}\right] h,\ 0 < p < 1$$

onde

i :	classe percentil,
	a classe percentil é o intervalo de classe onde se supera
	por primeira vez o (np) dos dados, isto é, $F_i > np$ ou $F_{r_i} > p$
LI_i :	limite inferior da classe percentil.
F_{i-1} :	freqüência acumulada absoluta da classe
	anterior à classe percentil.
f_i	freqüência absoluta da classe percentil

Exemplo 2.4.11 *Considerando os dados da TDF do exemplo 2.4.2, o valor do percentil $P_{0,8}$ encontra-se na classe $i = 5$ pois sua freqüência acumulada é maior de $nk = 40 \times 0,8 = 32$. Isto é, $F_5 = 35 > nk = 32$. Portanto,*

$$P_{0,8} = LI_5 + \left[\frac{32 - F_4}{f_5}\right] h = 8,8 + \left[\frac{32 - 31}{5}\right] (0,9) = 8,98\ g/dl$$

Esse valor indica que em 80% dos animais observou-se uma quantidade menor que 8,89 g/dl e no 20% restante dos animais observou-se uma quantidade superior a 8,89 g/dl.

2.5 Medidas de Dispersão

As medidas de posição ou de tendência central não necessariamente proporcionam informação suficiente para descrever dados de maneira adequada. Por exemplo, considere os dados de resistência à tensão (em psi) de três amostras de alheação de alumínio-lítio.

Amostra 1:	130	150	145	158	165	140
Amostra 2:	148	148	148	148	148	
Amostra 3:	90	120	205	140	165	160

Vemos que a média das 3 amostras é, $\bar{X}_1 = \bar{X}_2 = \bar{X}_2 = 148$ psi. Porém, em relação ao diagrama de pontos da figura 2.8, observa-se que a dispersão ou variabilidade da amostra 3 é muito maior que da amostra 1 e que os dados da amostra 2 apresentam variabilidade nula. Nesta seção, são definidos e ilustrados várias medidas úteis de variabilidade:

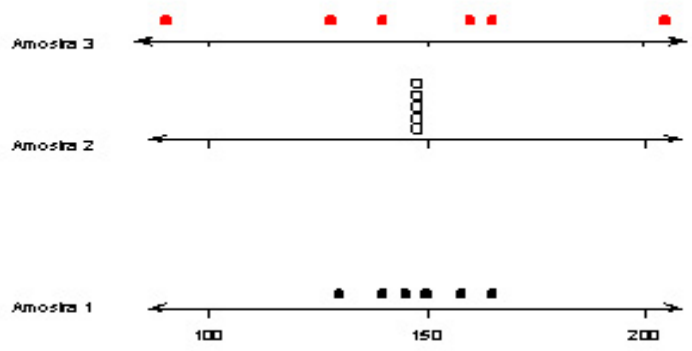


Figura 2.8: Diagrama de pontos dos dados da resistência à tensão

As medidas de dispersão ou variabilidade são medidas estatísticas que permitem conhecer o grau de homogeneidade ou heterogeneidade de um conjunto de dados. As medidas mais utilizadas são: amplitude, intervalo interquartil, variância, desvio padrão, e coeficiente de variabilidade. As três primeiras medidas são chamadas de medidas de variabilidade absoluta e a ultima é chamada de medida de variabilidade relativa.

2.5.1 Amplitude (A)

É a diferença entre a observação de maior e menor valor,

$$A = X_{\max} - X_{\min}.$$

Para as três amostras de resistência à tensão dadas anteriormente, a amplitude da primeira amostra é $A_1 = 165 - 130 = 35$, para a segunda amostra é $A_2 = 0$, enquanto para a terceira amostra é $A_3 = 205 - 90 = 115$. Desses resultados é claro que, quanto maior for a amplitude, maior será a variabilidade nos dados.

2.5.2 Intervalo interquartil (d)

É a diferença entre o terceiro quartil e o primeiro quartil,

$$d = Q_3 - Q_1$$

Considere os dados do exemplo 2.4.10, o intervalo interquartil é :

$$d = Q_3 - Q_1 = 2016 - 384 = 1632 \text{ horas}$$

O intervalo interquartil é menos sensível aos valores discrepantes ou extremos dos dados, que a amplitude.

2.5.3 Variância

É uma medida de dispersão absoluta das observações. É dada pela soma das diferenças quadráticas das observações em relação a sua média dividida pelo número total de observações. A variância populacional é denotada pela letra grega σ^2 e variância amostral por S^2

Populacional:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} = \frac{\sum_{i=1}^{N} X_i^2 - N\mu^2}{N} = \frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2.$$

Amostral:

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

onde

X_i: Valor da i-ésima observação da variável em estudo.

$\bar{X}$: Média amostral.

μ: Média populacional.

N: Tamanho da população.

n: Tamanho da amostra.

2.5.4 Desvio padrão

É a raíz quadrada positiva da variância. O desvio padrão populacional e amostral são denotados por σ e S respectivamente.

Populacional:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2 - N\mu^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N} X_i^2}{N} - \mu^2}.$$

Amostral:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} X_i^2 - \frac{\left[\sum_{i=1}^{n} X_i\right]^2}{n}}{n-1}}.$$

As unidades de medida da variância são iguais ao quadrado das unidades de medida da variável. Assim, se X é medido em libras por polegada quadrada (psi), a unidade da variância amostral são $(\text{psi})^2$. O desvio padrão tem a propriedade de medir a variabilidade nas mesmas unidades que a variável de interesse X.

Exemplo 2.5.1 *Na tabela 2.4, são apresentados as quantidades necessárias para cálculo da variância e do desvio padrão amostral, para os dados da amostra 1.*

Tabela 2.4: Cálculo da variância e o desvio padrão amostral

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	130	-18	324
2	150	2	4
3	145	-3	9
4	158	10	100
5	165	17	289
6	140	-8	64
	$\sum_{i=1}^{6} x_i = 888$	$\sum_{i=1}^{6}(x_i - \bar{x}) = 0$ $\bar{x} = \frac{888}{6} = 148$	$\sum_{i=1}^{6}(x_i - \bar{x})^2 = 790$

A variância amostral é:

$$S^2 = \frac{790}{6-1} = \frac{790}{5} = 158\ (psi)^2.$$

Enquanto que, o desvio padrão é:

$$S = \sqrt{158} = 12,57\ psi.$$

Alternativamente pode ser calculado a variância amostral utilizando a fórmula alternativa dada na definição de S^2:

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}. \tag{2.7}$$

Exemplo 2.5.2 *Na tabela 2.5, são apresentadas as quantidades necessárias para cálculo da variância usando a fórmula (2.7).*

Tabela 2.5: Cálculo da variância e o desvio padrão amostral

i	x_i	x_i^2
1	130	16900
2	150	22500
3	145	21025
4	158	24964
5	165	27225
6	140	19600
	$\sum_{i=1}^{6} x_i = 888$	$\sum_{i=1}^{6} x_i^2 = 132214$

Essa formula proporciona o seguinte:

$$S^2 = \frac{132214 - \frac{(888)^2}{6}}{6-1} = \frac{790}{5} = 158\ (psi)^2.$$

Essa quantidade é exatamente igual ao valor obtido anteriormente.

Observação 2.5.1 *A variância e o desvio padrão são utilizados para comparar a variabilidade de conjuntos de dados expressados nas mesmas unidades, com médias que sejam aproximadamente similares.*

Exemplo 2.5.3 *Deseja-se comparar a renda mensal do ano 2000 de duas empresas.*

Empresa A: $\mu_A = 450.000 \quad \sigma_A^2 = 2.500$
Empresa B: $\mu_B = 400.000 \quad \sigma_B^2 = 5.000$

Então pode-se afirmar que a renda mensal em 2000 da empresa B apresenta maior variabilidade que da empresa A ($\sigma_A^2 < \sigma_B^2$)

Exemplo 2.5.4 *A variância e o desvio padrão amostral para os dados das três amostras de alheação de alumínio-lítio do exemplo desta são apresentados abaixo:*

Amostra	Média	Variância	Desvio padrão
1	148	158	12,57
2	148	0	0
3	148	1502	38,8

Essas medidas confirmam a afirmação inicial de que a resistência à tensão da alheação de alumínio-lítio na amostra 3 apresenta uma maior dispersão que da amostra 1 e, que a resistência à tensão da alheação na amostra 2 não apresenta variabilidade. Esse último fato significa que as observações da resistência à tensão nessa amostra são todas iguais a sua média (148 psi).

2.5.5 Coeficiente de variabilidade

É uma medida de variabilidade adimensional expressa o número de vezes que o desvio padrão contém a média. Essa medida estatística é utilizada para comparar conjuntos de dados que têm diferentes unidades ou quando as médias são muito diferentes. Denota-se o coeficiente de variabilidade populacional e amostral por CV e cv, respectivamente.

Populacional:

$$CV = \frac{\sigma}{\mu}$$

onde

μ: Média populacional.

σ: Desvio padrão populacional.

Amostral:

$$cv = \frac{S}{\bar{X}}$$

onde

$\bar{X}$: Média amostral.

S: Desvio padrão amostral

Observação 2.5.2 *O coeficiente de variabilidade geralmente é expressado em percentuais, isto é multiplica-se por 100 as expressões anteriores.*

Exemplo 2.5.5 *Considere a altura (em metros) e peso (em kg) de uma amostra de alunos.*

	Média	Desvio Padrão
Altura	1,70 m	0,085m
Peso	70 kg	7kg

Pode-se observar que as características (altura e peso)tem diferentes unidades e nada pode ser dito a respeito de sua variabilidade, mas,

$$cv_{Altura} = \frac{0,085}{1,70} \times 100\% = 5\%$$
$$cv_{Peso} = \frac{7}{70} \times 100\% = 10\%$$

Os alunos são duas vezes mais dispersos quanto ao peso do que à altura.

Exemplo 2.5.6 *Considere os pesos (em kg) de uma amostra de meninos de 11 anos de idade e de uma amostra de homens de 25 anos de idade.*

	Média	Desvio Padrão
Homens	66 ,0	4,5
Meninos	36,0	4,5

Aparentemente as duas amostras tem a mesma variabilidade, porem,

$$cv_H = \frac{4,5}{66,0} \times 100\% = 6,8\%$$
$$cv_M = \frac{4,5}{36,0} \times 100\% = 12,5\%$$

Os pesos dos meninos apresentam uma dispersão maior que dos adultos.

2.5.6 Medidas de variabilidade para dados agrupados

Suponha um conjunto de dados **quantitativos contínuos** agrupados em uma tabela de distribuição de freqüência com k intervalos de classes.

Amplitude

$$A = LS_k - LI_1$$

onde LS_k é o limite superior da k-ésima classe e LI_1 é o limite inferior da primeira classe.

Variância

Populacional:

$$\sigma^2 = \frac{\sum_{i=1}^{k}(X_i' - \mu)^2 f_i}{N} = \frac{\sum_{i=1}^{k} X_i'^2 f_i - N\mu^2}{N} = \frac{\sum_{i=1}^{k} X_i'^2 f_i}{N} - \mu^2.$$

Amostral:

$$S^2 = \frac{\sum_{i=1}^{k}(X_i' - \bar{X})^2 f_i}{n-1} = \frac{\sum_{i=1}^{k} X_i'^2 f_i - n\bar{X}^2}{n-1} = \frac{\sum_{i=1}^{k} X_i'^2 f_i - \frac{\left[\sum_{i=1}^{k} X_i' f_i\right]^2}{n}}{n-1}.$$

onde X_i' é a i-ésima a marca de classe (ou ponto médio do intervalo de classe), f_i é a i-ésima freqüência absoluta, n é o tamanho da amostra e N é o tamanho da população. Para dados **quantitativos discretos** organizados em TDF as expressões para a variância são similares mas considerando $X_i' = X_i$.

Desvio padrão

Populacional: $\sigma = \sqrt{\sigma^2}$ Amostral: $S = \sqrt{S^2}$

Exemplo 2.5.7 *Considere a TDF do exemplo 2.4.2, referente a quantidade de hemoglobina (g/dl) de animais expostos a certo tóxico:*

Quantidade de Hb	X_i'	f_i	$X_i' f_i$	$(X_i')^2 f_i$
$5,2 \vdash 6,1$	5,65	4	22,6	127,69
$6,1 \vdash 7,0$	6,55	6	39,3	257,415
$7,0 \vdash 7,9$	7,45	12	89,4	666,03
$7,9 \vdash 8,8$	8,35	9	75,15	627,5025
$8,8 \vdash 9,7$	9,25	5	46,25	427,8125
$9,7 \vdash 10,6$	10,15	4	40,60	412,09
Total		40	$\sum_{i=1}^{6} X_i' f_i = 313,3$	$\sum_{i=1}^{6} X_i'^2 f_i = 2518,54$

Amplitude

$$A = 10,6 - 5,2 = 5,4$$

Variância:

$$S^2 = \frac{2518,54 - (313,3)^2/40}{39} = 1,6569 \ (g/dl)^2$$

Desvio padrão:

$$S = 1,2872 \ g/dl$$

Esse resultado indica que a quantidade de hemoglobina em animais expostos a certo tóxico apresenta uma dispersão em relação a sua média (7,8325) de $1,2872$ g/dl.

Coeficiente de variabilidade:

$$cv = \frac{S}{\bar{X}} = \frac{1,2872}{7,8325} = 0,1643$$

Esse valor indica que a quantidade de hemoglobina em animais expostos a um certo tóxico, apresenta uma variabilidade relativa de $16,43\%$.

2.6 Boxplot

O boxplot é um gráfico que fornece uma visualização da distribuição dos dados, além de permitir detectar rapidamente uma possível assimetria dessa distribuição. Sua construção é baseada nas seguintes medidas: na mediana, no primeiro e terceiro quartis, e nos valores extremos. A forma desse gráfico tem as seguintes características (veja a figura 2.9):

a) A caixa ("box") é delimitada pelo primeiro (Q_1) e terceiro (Q_3) quartis. A linha interior da caixa corresponde a mediana ($me = Q_2$).

b) A partir dos limites da caixa, considera-se duas linhas auxiliares que distam 1,5 o intervalo interquartil $d = Q_3 - Q_1$. Essas linhas não aparecerão no gráfico final. Elas servem para caracterizar os valores discrepantes que são os valores menores que $Q_1 - 1,5d$ ou valores maiores que $Q_3 + 1,5d$. Os valores discrepantes serão representados no gráfico com asteriscos ($*$).

c) Os limites do gráfico, representados por uma linha à direita e à esquerda ("bigodes") da caixa, correspondem ao maior e ao menor valores não discrepantes do conjunto de dados.

Observação 2.6.1 *A caixa contém 50% dos valores (25% de cada lado da mediana). Outros 50% dos valores estão praticamente divididos entre o "bigode"direito e "bigode"esquerdo.*

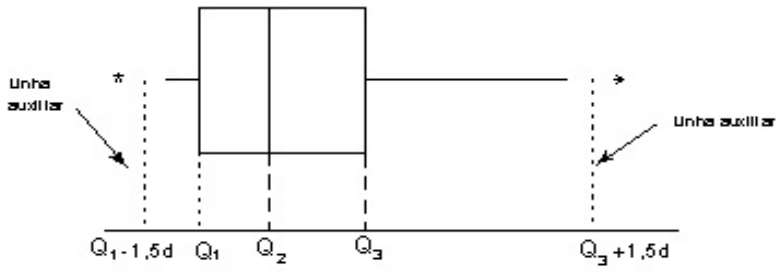


Figura 2.9: Desenho esquemático do Boxplot

Exemplo 2.6.1 (Exemplo de construção de um Boxplot) *Com a finalidade de aumentar o peso (em kg) um regime alimentar foi aplicado em 12 pessoas. Os resultados (ordenados) foram: -0,5 2,5 3,0 3,6 4,7 5,3 5,9 6,0 6,2 6,3 7,9 11,2*

Calculando as medidas temos:

mediana $(me$ ou $Q_2) = 5{,}6$ kg
1o.quartil $(Q_1) = 3{,}3$ kg
3o.quartil $(Q_3) = 6,25\, kg$

d =intervalo interquartil $= Q_3 - Q_1 = 2,95$ kg

Logo as linhas auxiliares correspondem aos pontos:

$Q_1 - 1,5d = -1,125$ kg
$Q_3 + 1,5d = 10,675$ kg

O gráfico de boxplot para o exemplo é mostrada na figura 2.10.

Da figura 2.10, pode-se observar que há uma observação discrepante no conjunto de dados, o que significa que há uma pessoa que teve um incremento de peso muito acima do resto das pessoas. Além disso, há uma maior concentração dos dados acima do peso mediano.

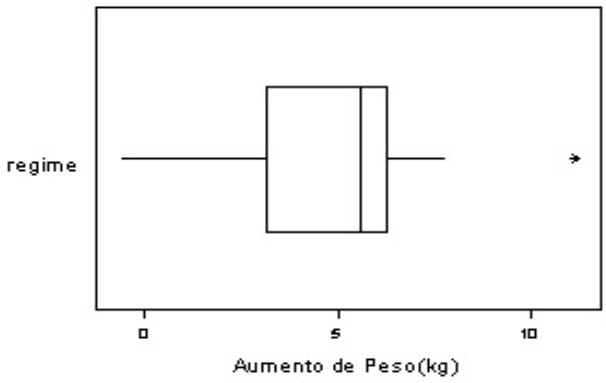


Figura 2.10: Gráfico de Boxplot para o regime alimentar

Observação 2.6.2 *O boxplot também pode-se representar em forma vertical, como mostra a figura 2.11.*

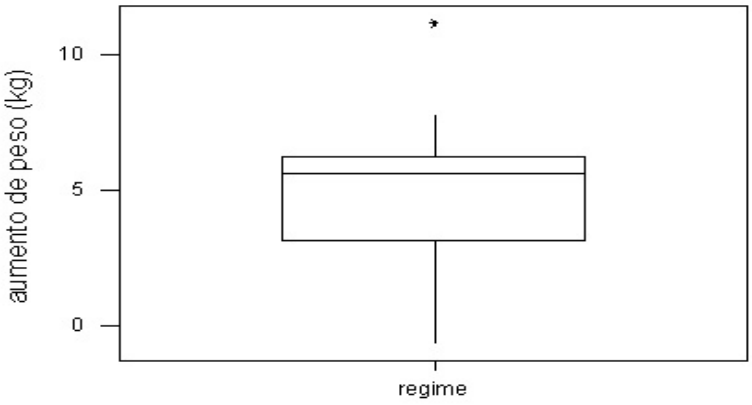


Figura 2.11: Gráfico de Boxplot para o regime alimentar

2.7 Exercícios Resolvidos

1. Uma pesquisa foi realizada numa cidade do interior de Minas Gerais, com o objetivo de determinar o número de horas por dia que as donas de casa se dedicam a assistir televisão. Obtendo-se os seguintes resultados:

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

(a) Construa a tabela de distribuição de freqüências com intervalos de classe do mesmo comprimento e usando a regra de Sturges.

(b) Interpretar:

(i) A marca de classe do segundo intervalo.

(ii) A freqüência absoluta de segundo intervalo de classe.

(iv) A freqüência relativa percentual do terceiro intervalo de classe.

(v) A freqüência acumulada relativa do quarto intervalo de class.

(c) Desenhe o histograma e polígono de freqüências relativas.

(d) Que porcentagem de donas de casa assistem televisão mais de 4,8 horas diárias?(considere a TDF)

(e) Qual é a quantidade mínima de horas que uma dona de casa deve assistir televisão para pertencer aos 14,2% das donas de casa que menos assistem televisão?

Solução

a) Construção da tabela de distribuição de freqüências absolutas e relativas:

(1) Cálculo do número de classe (k)

$$n = 49 \qquad k = 1 + 3,3\log(49) = 6.57765$$
$$\Rightarrow \quad k = 7 \ \text{(arredondamento simples)}$$

(2) Cálculo do comprimento ou amplitude (A)
$A = X_{\max} - X_{\min} = 6,3 - 2,5 = 3,8$

(3) Cálculo da amplitude (ou comprimento) de intervalo de classe (h)

$$h = \frac{A}{k} = \frac{3,8}{7} = 0,542857 \approx 0,6$$

(arredondamento por excesso ao um número igual ao de cifras decimais dos dados)

(3) Cálculos dos limites dos intervalos de classe

$LI_1 = X_{\min} = 2,5, \quad LS_1 = LI_1 + h = 2,5 + 0,6 = 3,1$
$LI_2 = LS_1, \quad LS_2 = LI_2 + h = 3,1 + 0,6 = 3,7$
$LI_3 = LS_2, \quad LS_3 = LI_3 + h = 3,7 + 0,6 = 4,3$
$LI_4 = LS_3, \quad LS_4 = LI_4 + h = 4,3 + 0,6 = 4,9$
$LI_5 = LS_4, \quad LS_5 = LI_5 + h = 4,9 + 0,6 = 5,4$
$LI_6 = LS_5, \quad LS_6 = LI_6 + h = 5,4 + 0,6 = 6,0$
$LI_7 = LS_6, \quad LS_7 = LI_7 + h = 6,0 + 0,6 = 6,7$

(4) Obtenção das marcas de classe (X_i'). É possível mostrar que a marca de classe satisfaz as seguintes relações que são de muita utilidade.

$$X_i' = \frac{LI_i + LS_i}{2};\ X_{i+1}' = X_i' + h;\ \ LS_i = X_i' + \frac{h}{2};\ \ LI_i = X_i' - \frac{h}{2}$$

Por exemplo:

$$X_i' = \frac{LI_i + LS_i}{2} = \frac{3,1 + 3,7}{2} = 3,4.$$

Desse modo calcula-se as marcas de classe restantes.

(5) Efetua-se a contagem para alocar cada observação (dado) ao intervalo que lhe corresponde. Determina-se as freqüências absolutas (f_i). Dos dados obtemos: $f_1 = 1$, $f_2 = 6$, $f_3 = 11$, $f_4 = 19$, $f_5 = 9$, $f_6 = 2$, $f_7 = 1$.

(6) Determinação das freqüência relativas(f_{r_i}) para cada intervalo "i"
$f_{r_i} = \frac{f_i}{n}$, Além disso, $\sum_{i=1}^{k} f_{r_i}$.
$f_{r_1} = 1/49 = 0,020, f_{r_2} = 0,122, \ldots, f_{r_7} = 0,020$

(7) Determinação das freqüências acumuladas absolutas(F_i)
$F_i = F_{i-1} + f_i$, $i = 1, 2, \ldots, k$, com $F_k = n$.
$F_1 = 1$, $F_2 = 1 + 6 = 7$, $F_3 = 7 + 11 = 18, \ldots, F_7 = 49 = n$

(8) Determinação das freqüências acumuladas relativas (F_{r_i})
Tem-se as seguintes relações para F_{r_i}:
$F_{r_i} = \sum_{j=1}^{i} f_{r_j}$; $F_{r_1} = f_{r_1}$, $F_{r_1} = \frac{F_1}{n}$, $F_{r_i} = F_{r_{i-1}} + f_{r_i}$, $i = 1, \ldots, k$
$F_{r_1} = 1/49 = 0,020$, $F_{r_2} = 0,020 + 0,122 = 0,142, \ldots, F_{r_7} = 1$

Na tabela 2.6, são apresentados a distribuição de freqüências do número de horas por dia que as 49 donas de casa entrevistadas assistem televisão:

Tabela 2.6: Distribuição do número de horas diárias que as 49 donas de casa entrevistadas assistem televisão

Número de horas	X_i'	f_i	f_{r_i}	p_i	F_i	F_{r_i}	P_i
$2,5 \vdash 3,1$	2,8	1	0,020	2,00	1	0,020	2,00
$3,1 \vdash 3,7$	3,4	6	0,122	12,20	7	0,142	14,20
$3,7 \vdash 4,3$	4,0	11	0,224	22,40	18	0,367	36,70
$4,3 \vdash 4,9$	4,6	19	0,388	38,80	37	0,755	75,50
$4,9 \vdash 5,5$	5,2	9	0,184	18,40	46	0,939	93,90
$5,5 \vdash 6,1$	5,8	2	0,041	4,10	48	0,979	97,9
$6,1 \vdash 6,7$	6,4	1	0,020	2,00	49	1,00	100
Total		49	1,00	100,0			

(b)Da tabela 2.6 tem-se:

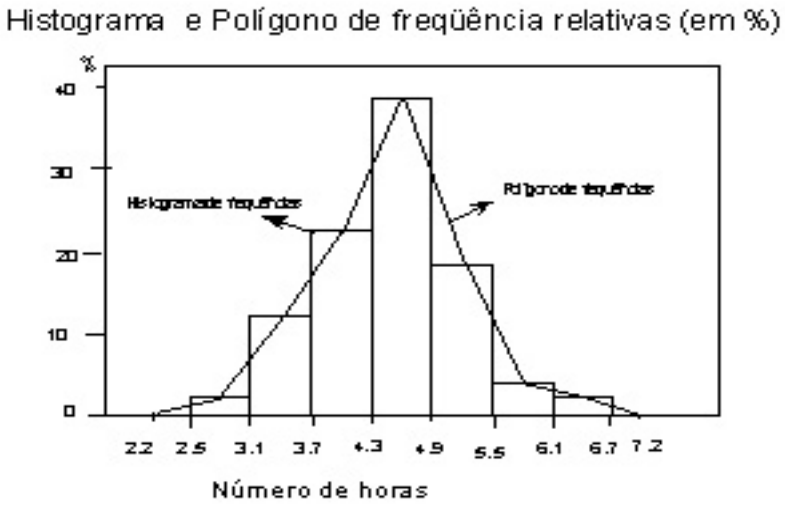
(i) $X_2' = 3,4$; há 6 donas de casa que em média assistem televisão 3,4 horas por dia.

(ii) $f_4 = 19$; há 19 donas de casa assistem televisão entre 4,3 e 4,8 horas por dia.

(iii) $p_3 = 22,4\%$; 22,4% das donas de casa assistem TV entre 3,8 e 4,3 horas por dia.

(iv) $P_4 = 75,5\%$; 75,5 % das donas de casa entrevistadas assistem TV menos de 4,8 horas ao dia.

(c) A partir da tabela 2.6, são construídos o histograma e o polígono de freqüências relativas em porcentagens.



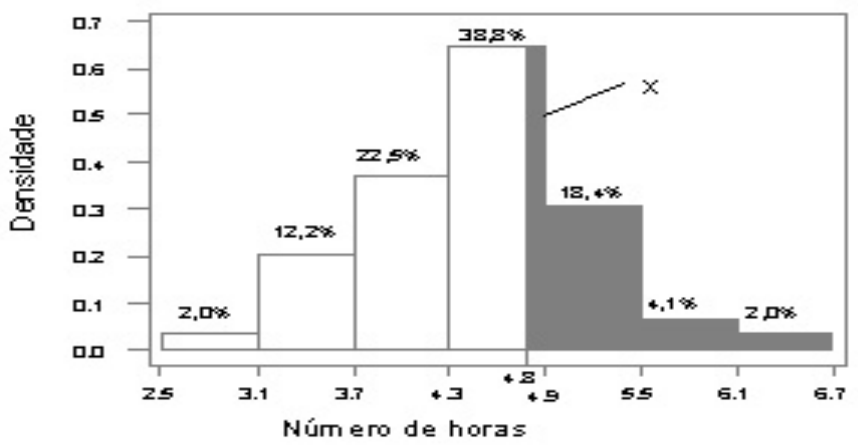


Figura 2.12: Gráfico de distribuição de densidade do números de horas que as donas de casa assistem TV.

(d) Para determinar a percentagem de donas de casa que assistem TV mais de 4,8 horas considere o gráfico do histograma de freqüência de densidade para esses dados. Essa freqüência é área hachurada no gráfico de densidade da figura 2.12, o qual é completamente determinada se obtemos o valor de x.

Da figura 2.12, tem-se

$$\frac{4,9-4,8}{x} = \frac{4,9-4,3}{38,8} \implies x = 6,5$$

Portanto, a porcentagens de donas de casa que assistem mais de 4,8 horas é aproximadamente 6,5+18,4+4,1+2=31

(e) Do gráfico de densidade na figura 2.12, observa-se que tempo máximo é 3,7 horas para ser incluído no grupo 14,2% das amas de casa que menos assistem televisão.

2. Um Biólogo estuda o comprimento em centímetros de peixes de uma espécie conhecida como carpa de Singapur (*cyprinus Cardio*). Para uma amostra aleatória, de tamanho 7, de peixes machos e 8, de peixes fêmeas, ele obteve os seguintes resultados:

Macho:	46	42	55	49	40	44	39	
Fêmea:	44	41	42	40	48	47	46	45

Faça uma análise descritiva dos dados e comente as principais diferenças.

Inicialmente na figura 2.13, é representado o boxplot para os comprimentos de peixes machos e fêmeas. Dessa figura, pode-se observar que há diferenças nos comprimentos de peixes machos e fêmeas. O valor mediano dos comprimentos dos peixes estão próximos, mas as medidas dos comprimentos dos peixes machos apresentam maior variabilidade que as dos peixes fêmeas.

Na tabela 2.7, são apresentados algumas medidas descritivas, para os dados do exemplo. A tabela confirma as afirmações feitas inicialmente.

Tabela 2.7: Medidas descritivas para o comprimento de peixes machos e fêmeas

Peixe	Média	Mediana	Desvio padrão
Macho	45,00	44,00	5,60
Fêmea	44,13	44,50	2,90

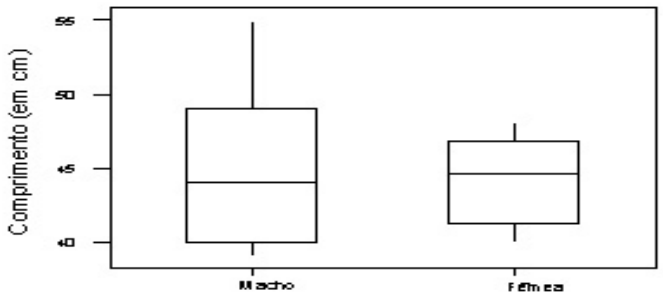


Figura 2.13: Boxplot dos comprimentos de peixes machos e fêmeas

2.8 Exercícios

1. Os seguintes dados são resultados de uma amostra aleatória de quantidade de hemoglobina (Hb) no sangue, em g/dl (gramas por decilitro),encontrados em 30 pacientes entre 15 - 20 anos, que foram ao laboratório central de um Hospital:

20.8	27.8	26.2	21.6	23.3	23.5	26.1	26.5	20.0	24.7	21.7	28.2
25.0	23.4	24.5	27.9	25.7	24.8	26.8	25.5	25.3	22.3	21.2	26.0
23.8	22.5	23.7	24.9	25.2	24.4						

   (a) Qual é a variável de estudo? Classifique-a.

   (b) Construa uma tabela de freqüências usando a regra de Sturges para determinar o número de intervalos de classe.

   (c) Faça o Histograma e o polígono de freqüências relativas.

   (d) Qual é o significado da freqüências acumulada percentual do quarto intervalo?

   (e) Faça o polígono de freqüências acumulada relativa (ogiva).

   (f) Qual é a porcentagem de pacientes com mais de 25.6 g/dl de hemoglobina no sangue?

   (g) Qual é a quantidade máxima de hemoglobina deve ter um paciente para pertencer aos 40% dos pacientes de menor quantidade de Hb no sangue.

2. Na elaboração de microcomprimidos de liberação gradual para um medicamento, coloca-se um cor que identifica o número de capas de recobrimento. O responsável da produção deseja ter uma representação gráfica da proporção em que se encontre cada cor. Para isto escolhe ao acaso uma amostra obtendo os seguintes resultados:

azul	verde	verde	verde	vermelho
azul	verde	azul	azul	verde
verde	azul	vermelho	verde	azul
vermelho	vermelho	vermelho	azul	verde
vermelho	vermelho	azul	azul	azul
verde	azul	verde	vermelho	verde

   (a) Classifique os dados obtidos.

   (b) Que tipo de gráfico você faria para estes dados. Faça-o.

3. Uma Empresa Farmacêutica classifica os seus empregados de acordo com o grau de instrução, assim foi obtido dos seguintes resultados:

Grau de instrução	N ^{o} de empregados	Gastos total mensal com remunerações
Primeiro grau	15	1950
Segundo grau	35	6650
Nível Superior	50	14000

(a) Que medida de posição recomendamos para a variável grau de instrução dos empregados?

(b) Achar a remuneração mensal média dos empregados.

(c) Se a empresa decidir dar um aumento mensal aos empregados de acordo com os seguinte critérios e apartir de 01/05/2000

(i) Cada empregado terá um aumento de 40 u.m. mensais

(ii) Adicionalmente ao aumento descrito em 1 os empregados teriam uma remuneração complementar ao total mensal, sendo 5% para os empregados com primeiro grau, 8% para empregados com segundo grau e 15% para empregados com instrução superior. Achar a remuneração média mensal prometido aos empregado apartir de 01/05/2000.

4. A continuação apresenta-se o rendimento (%) de uma reação para a fabricação de uma substância química, em 80 bateladas consecutivas produzidas por uma industria:

81,8	87,1	82,7	79,8	81,3	79,5	88,5	75,9	81,6	73,9
85,5	87,1	82,0	79,3	82,5	87,1	83,0	87,3	79,7	82,0
83,6	84,5	80,4	78,1	86,4	76,7	83,7	78,4	76,0	80,9
80,2	78,9	77,4	78,5	82,9	81,9	80,7	78,4	78,0	81,4
84,6	79,5	83,2	80,5	80,7	79,0	90,9	79,9	86,8	80,1
83,2	78,2	80,4	85,5	85,5	79,3	83,0	78,1	83,4	83,6
85,7	86,8	86,5	83,8	86,8	83,5	79,9	76,6	84,3	78,5
74,4	71,8	79,1	82,1	84,5	78,4	80,7	70,7	78,5	85,2

(a) Construa uma tabela de freqüências com intervalos de classe do mesmo comprimento considerando que k=7.

(b) Obtenha e interprete:

(i) A marca de classe do quarto intervalo de classe.

(ii) A freqüência absoluta do segundo intervalo de classe.

(iii) A freqüência acumulada percentual do segundo intervalo de classe.

(c) Desenhe o histograma de freqüências percentuais e descreve as principais características dos dados.

(d) Obtenha e interprete a média, mediana, moda e desvio padrão.

5. Para cinco volumes de uma solução foram medidos os tempos de aquecimento em um mesmo bico de gás e as respectivas temperaturas. O resultado foi a seguinte:

Tempo (min.):	22	20	19	23	17
Temperatura (0C):	75	80	78	84	78

Qual das duas variáveis apresenta uma maior variabilidade? Justifique.

6. Um artigo publicado na **Food Tecnology Journal** (1956), descreve um estudo sobre o conteúdo de protopectina em tomates durante o armazenamento. Para o qual considerou-se dois períodos de armazenamento e analisou-se as amostras de nove lotes de tomates em cada período, obtendo-se os dados abaixo:

Tempo de armazenamento.	lotes								
	1	2	3	4	5	6	7	8	9
7 Dias	1802.0	107.4	278.8	1275	544.0	672.2	818.0	406.8	461.6
21 dias	415.5	485.4	377.6	270.4	467.8	272.1	394.1	336.4	371.2

(a) Qual é a variável e de que tipo é ?

(b) Determine a média e mediana. Qual destas duas medidas é melhor para os dois grupos acima ?.

(c) Dos tempos de armazenamento, qual apresenta maior variabilidade? Justifique.

(d) Desenhe o Boxplot para cada um dos tempos de armazenamento. Quais são as principais diferenças?

(e) Considerando os ítens (b) a (d) , descreva as principais diferenças nos tempos de armazenamento.

7. Um hospital maternidade está planejando a ampliação dos leitos para recém nascidos. Para tal, fez um levantamento dos últimos 50 nascimentos obtendo a informação sobre o número de dias que os bebes permanecem no hospital, antes de terem alta. Os dados, já ordenados, são apresentados a seguir:

1	1	1	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4	4	5
5	5	5	5	5	6	7	7	8	15										

(a) Calcule média, moda e mediana.

(b) Determine o desvio padrão.

(c) Você identifica algum valor excepcional dentre os que foram observados? Se sim, remova-o e recalcule os ítens a) e b). Comente as diferenças encontradas.

(d) Dentre as medidas de posição calculadas em a), discuta quais delas seriam mais adequadas para resumir esse conjunto de dados.

8. O índice de germinação é um dos principais fatores para definir a qualidade de sementes. Ele é determinado em um experimento científico conduzido pelo fabricante e regulamentado pelos órgãos fiscalizadores. Um fabricante afirma que o índice de germinação de suas sementes de milho é mais de 85%. Para verificar tal afirmação uma cooperativa de agricultura sorteou 100 amostras com 100 sementes em cada uma e anotou a porcentagem de germinação em cada amostra. Os resultados estão na tabela de abaixo.

% de germinação	Freqüência
$70 \vdash 75$	5
$75 \vdash 80$	20
$80 \vdash 85$	40
$85 \vdash 90$	18
$90 \vdash 95$	12
$95 \vdash 100$	5

(a) Calcule e interprete a mediana, 1^0 quartil e 3^0 quartil . Comente a afirmação do fabricante.

(b) Desenhe o Boxplot

(c) Determine a proporção de sementes com índice de germinação menor de 82

(d) Suponha que outro fabricante produz sementes com índice de germinação média igual a 89% e desvio padrão igual a 5%, qual dos produtores apresentam maior variabilidade ?. Justifique

9. Uma maquina foi regulada para fabricar placas de 5 mm de espessura, em média, com uma variabilidade relativa de, no máximo, 3%. Iniciada a produção, foi colhida aleatoriamente uma amostra de tamanho 50, que forneceu a seguinte tabela de distribuição de freqüência com intervalos do mesmo comprimento.

Espessura (em mm)	N^0 de placas
$4,6 \vdash$	3
$\vdash$	18
$\vdash 4,8$	10
$\vdash$	18
$\vdash$	
$\vdash$	2

(a) Esboce o histograma de freqüências percentuais e descreva as principais características das placas amostradas.

(b) Que você pode afirmar a respeito da regulagem da maquina?

(c) Determinar e interpretar: a moda e a mediana.

(d) Qual deve ser a espessura das placas para ser considerado entre os 10% com maior espessura?

(e) Placas com espessuras menores ou iguais a 4,95 mm são vendidos a R$ 1,5 e placas com espessuras entre 4,95 mm e 5,15 mm são vendidos a R$ 2,0 e placas com espessuras maiores ou iguais a 5,15 mm são vendidos a R$ 1,0. Determinar o preço médio de venda de cada placa.

10. Um biólogo esta investigando qual o acasalamento de um determinado tipo de caramujo que produz o maior número mediano de ovos eclodidos. Nesse sentido desenvolve um experimento em que três grupos são investigados: Grupo1 (1 macho e 1 fêmea), Grupo 2 (2 machos e 1 fêmea) e Grupo 3 (1 macho e 2 fêmeas). Para cada grupo, 20 acasalamento são feitos e observados o número de ovos postos eclodidos após 14 dias de permanência. Os Boxplots correspondentes são apresentados na figura 2.14.

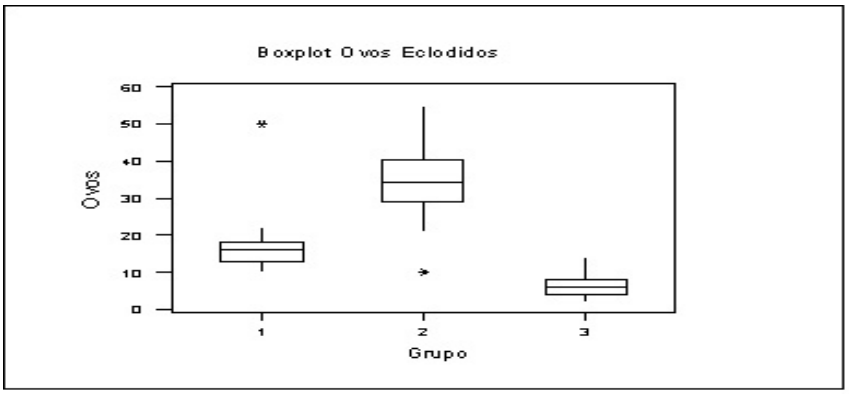


Figura 2.14: Boxplot do número de ovos eclodidos em três grupos.

(a) Qual grupo produz o maior número mediano de ovos eclodidos? Forneça uma estimativa desse número mediano de ovos eclodidos

(b) Qual são as principais diferenças entre os 3 grupos ?. Justifique.

11. Uma empresa química afirma que nenhum de seus funcionários estão contaminado por chumbo, para verificar isto a empresa faz um exame de rotina em 36 funcionários escolhido ao acaso, constatado as seguintes concentrações no sangue. Sabendo que o limite máximo de contaminação por chumbo é de 4,80 μmol /L(Concentrações em μmol de chumbo por litro de sangue).

3.35	3.67	4.27	5.11	5.55	2.83	3.29	3.63	4.15	4.96	5.50	2.81
3.26	3.58	3.94	4.58	5.42	2.52	3.15	3.55	3.90	4.49	5.28	2.32
3.09	3.49	3.82	4.43	5.25	1.53	3.03	3.45	3.76	4.36	5.20	1.28

(a) Construa uma tabela de distribuição de freqüências de classe usando a regra de Sturges ($k = 1 + 3,3 \log_{10}(n)$) para determinar o número de intervalos de classe .

(b) Calcule as medidas de posição e diga se o nível deste metal entre os funcionários é preocupante. Justifique.

(c) Determine a porcentagem de funcionários que se encontra no intervalo $(\bar{X} - S; \bar{X} + S)$.

12. O teste do pezinho é feito para se constatar em recém nascidos uma doença genética chamada de fenicetonúria. Este teste consiste em dosar a quantidade de um aminoácido, a fenilalanina, que em quantidades altas no organismo pode causar dano às células, principalmente as cerebrais. Numa maternidade, em um mesmo dia, o teste foi feito em 30 recém nascidos obtendo as seguintes concentrações de fenilalanina em μ mol / L

133,92	174,12	170,88	244,81	142,26	206,73	156,25
224,29	145,59	214,26	175,06	205,72	144,94	171,73
147,69	168,12	182,64	186,24	206,96	143,82	173,31
116,44	208,01	110,29	197,26	212,34	180,76	189,12
167,96	144,07					

(a) Construa uma tabela de distribuição de freqüências e a representação gráfica dos dados acima considerando a freqüência relativa em porcentagens. Comente as principais características destes dados.

(b) a concentração de fenilalanina permitida é de 70 a 210 μmol / L para um recém nascido sadio. Determine a porcentagens crianças que se encontra nessa faixa.

(c) Calcule e interprete as medidas de posição para esses dados.

(d) Numa outra maternidade a concentração de fenilalanina média foi de 2,99mg/dl e variância de $S^2 = 0,084mg^2/dl^2$. Qual das maternidade obteve maior variabilidade dos dados ?. Justifique.

13. O número de pessoas praticam a auto-medicação no Brasil são alarmantes. Para se constatar que essa atitude é praticada por pessoas de todos os níveis sócio-econômicos e graus de instrução, foi feito entrevistas com 20 pessoas de uma cidade do interior de Minas. Os dados obtidos foram organizados na tabela abaixo:

No	Automedicam	Grau de instrução	Nível sócio-econ.
1	Sim	1^o grau	Baixa
2	Sim	1^o grau	Baixa
3	Sim	2^o grau	Média
4	Sim	Superior	Média
5	Não	Superior	Alta
6	Não	Superior	Média
7	Sim	1^o grau	Baixa
8	Não	2^o grau	Baixa
9	Sim	2^o grau	Média
10	Não	Superior	Média
11	Sim	Superior	Alta
13	Sim	1^o grau	Baixa
14	Não	Superior	Média
15	Sim	1^o grau	Baixa
16	Sim	1^o grau	Baixa
17	Sim	2^o grau	Baixa
18	Sim	2^o grau	Média
19	Sim	1^o grau	Baixa
20	Não	Superior	Média

(a) classifique as variáveis em qualitativa nominal ou ordinal.

(b) Calcule a porcentagem dos entrevistados que praticam a auto-medicação, levando em conta a escolaridade e o nível social.

(c) De acordo com os dados, você acha que a auto-medicação não depende do nível sócio-econômico ou grau de instrução. Justifique.

14. O cloranfenicol é um antibiótico bacteriostático, pois inibe a síntese protéica. Apesar de agir somente em ribosomas bacterianos, este antibiótico produz efeitos colaterais e até a morte de pessoas com sensibilidade a esta família de antibiótico. Um grupo de 1400 pacientes com infecção por estreptococos tratado com o cloranfenicol e 800 foram retratadas com um novo antibiótico obtendo-se os seguintes dados relacionados com o aparecimento de efeitos colaterais e óbitos.

	Cloranfinicol	Novo Antibiótico
Não apresentaram	1279	613
Apresentaram	116	184
Óbito	5	3

(a) Classifique a variável em estudo. Qual dos dois antibióticos oferece menores riscos para a saúde dos pacientes?Justifique.

15. Em um laboratório de análises clínicas revelou os dados sobre o nível de glicose no soro de 50 pessoas que solicitaram esse exame. Os dados obtidos apresentados abaixo são em mg de glicose por decilitros de soro:

181,93	145,09	132,92	124,88	118,96	110,48	100,04	89,65
181,17	143,78	130,83	124,83	118,39	108,02	95,33	88,51
167,83	141,89	129,83	122,01	116,00	105,87	95,07	85,10
152,06	137,96	129,53	121,57	115,13	103,62	93,66	83,12
149,56	136,37	128,84	121,26	114,55	102,16	92,94	80,98
145,62	134,48	124,96	119,65	111,90	100,99	92,72	78,49
62,32	76,73						

(a) Construa uma tabela de freqüências usando a regra de Sturges para determinar o número de intervalo de classe.

(b) Faça o histograma de freqüências relativas e comente as principais características dos dados.

(c) Uma pessoa é considerado saudável, se o nível de glicose é maior o igual a 30mg/dl mais menor a 110 mg/dl. Qual é a porcentagens de pessoas saudáveis? (considere a TDF).

(d) Calcule e interprete média, mediana, 1^o quartil e 3^o quartil.

(e) Determina a porcentagens de pessoas que se encontram no intervalo $[Q_1 - 1,5d; Q_3 + 1,5d)$, onde $d = Q_3 - Q_1$.

16. Uma farmácia de manipulação encomendou lotes de ácido acetisalicílico(AAS) de duas empresas (A e B). Na análise da pureza da matéria prima constatou-se que havia ácido salicílico misturado ao AAS. Amostras dos lotes foram retiradas (100 mg), analisadas e organizadas na tabela abaixo:

Empresa A		Empresa B	
Lotes	% de Pureza	Lote	% de Pureza
1	96.793	1	93.808
2	98.381	2	94.651
3	96.590	3	93.073
4	96.458	4	95.169
5	97.335	5	95.376
6	95.778	6	94.606
7	94.941	7	94.410
8	97.578	8	93.691
9	94.764	9	95.614
10	96.197	10	94.194

(a) Determine a média e o desvio padrão e comente as principais diferenças.

(b) Construa o Boxplot e considerando o item (a) descreva as principais diferenças.

(c) Em uma das empresas estava especificado na embalagem do material que em média havia 1,175% de impurezas. Qual das duas empresas poderia, corretamente, informar isto?

17. Num hospital foi realizado exames para se determinar o nível de colesterol em pacientes com peso acima do normal. Os dados obtidos dos 36 pacientes examinados estão relacionados a seguir, em mg/dl.

180,31	213,99	227,53	246,87	264,67	275,18	182,41	214,41	235,22
254,43	266,19	288,08	188,43	218,06	235,40	257,57	266,52	290,89
191,71	219,67	237,98	260,42	269,72	292,66	204,24	220,42	241,23
262,83	271,95	327,64	212,81	225,22	246,38	264,42	274,00	336,47

(a) Construa uma tabela de freqüências usando a regra de Sturges.

(b) Faça o polígono de freqüências relativas.

(c) Calcule e interpreta a média e o desvio padrão.

(d) Qual o percentual de pacientes que pertencem ao intervalo de 150 a 240 mg/dl considerado para uma pessoa normal.

18. Na análise de vacinas contra a febre amarela, constatado uma possível fraude no volume especificado no rótulo dessas vacinas. Foram analisadas 30 ampolas de 0.50 ml, dando os seguintes resultados:

0,591	0,521	0,495	0,546	0,503	0,456	0,592	0,511	0,491
0,543	0,503	0,448	0,573	0,508	0,482	0,540	0,502	0,435
0,563	0,505	0,481	0,531	0,500	0,424	0,549	0,505	0,476
0,529	0,497	0,400						

(a) Calcule as medidas de posição e interprete-as.

(b) Faça uma representação gráfica.

(c) No rótulo dos lotes estava mencionado o volume de 0.5 ml e a variabilidade de 1% em volume nas ampolas. Diga se isto está correto, de acordo com os dados obtidos.

19. Em um laboratório de pesquisa genéticas foi feito cruzamentos entre camundongos pretos e albinos, o objetivo da pesquisa era se saber quais as cores dos filhotes e suas proporção; os dados obtidos foram organizados abaixo:

Preto	Marrom	albino	marrom	preto	marrom	albino	preto
Albino	Preto	preto	preto	preto	preto	preto	marrom
Preto	Albino	preto	albino	marrom	preto	albino	preto
Preto	Preto	marrom	preto	albino	preto	preto	albino

(a) Qual é a variável em estudo? Classifique-a.

(b) Calcule a medida de tendência central mais conveniente para os dados acima.

(c) Faça um gráfico adequado para os dados obtidos.

20. Hidatidose é uma doença causada por helmintos do gênero Echinococcus. O quadro abaixo mostra pacientes com cisto ciático operados em Azul (Província de Bueno Aires, Argentina)segundo grupos etários.

Grupos etários	No de pacientes operados
$0 \vdash 10$	29
$10 \vdash 20$	76
$20 \vdash 30$	88
$30 \vdash 40$	52
$40 \vdash 50$	42
$50 \vdash 60$	23
$60 \vdash 70$	12

Fonte: Adaptado do livro "Patologia"de Luís Rey

(a) Faça a representação gráfica dos dados considerando a freqüência relativas em percentuais e descreva as principais características.

(b) Calcule e interprete 1^o quartil, mediana e 3^o quartil.

(c) Qual é a idade média dos pacientes com cisto ciático operados em Azul.

(d) De acordo com os dados, qual o percentual que pacientes operados com menos de 18 anos.

21. Para cada uma das doses 0,20 0,32 0,50 e 0,80 (mg/cm^2) de um determinado inseticida foram submetidos seis grupos, cada um com dez besouros, e observado o número de sobreviventes. Os dados são resumidos na tabela abaixo.

0,20			0,32			0,50			0,80		
7	9	10	6	7	9	6	4	8	1	3	2
8	9	9	7	8	4	5	6	3	2	6	5

Para cada dose calcule a proporção de sobreviventes e calcule a média, mediana, desvio padrão e quartis para o número de sobreviventes. Compare o número médio com o número mediano de sobreviventes segundo as doses. Comente.

22. Um experimento é conduzido para comparar dois regimes alimentares no que diz respeito ao aumento de peso. Vinte indivíduos são distribuídos ao acaso entre dois grupos em que ao primeiro deles foi dado a dieta A e ao segundo a dieta B. Decorrido certo intervalo de tempo verifica-se que os aumentos de peso correspondentes foram as seguintes:

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

Análise os dados descritivamente e comente as principais diferenças.

23. Uma empresa construtora de equipamentos para indústria alimentar pretende adquirir termostatos para comandar a abertura de um certo tipo de fornos, contemplando a possibilidade de os adquirir a um dos fornecedores A ou B. O fornecedor B vende os termostatos mais caros, invocado que são mais fiáveis do mercado. Num ensaio de 9 termostatos de fornecedor A e 11 do fornecedor B, todos regulados à mesma temperatura, as temperaturas observadas de abertura dos fornos foram as seguintes.

Fornecedor A	423	425	401	430	417	425	416	421	419		
Fornecedor B	419	414	422	435	418	421	429	410	406	418	421

Você acha que o termostato do fornecedor B é mais confiável que do fornecedor A?. Justifique porque?

24. A qualidade de rebites é tanto melhor quanto maiores sua resistência média e sua homogeneidade. Com a finalidade de verificar qual das marcas A e B são melhores, 8 rebites da marca A foram ensaiados ao cisalhamento que forneceu uma média de 37,09 e desvio padrão de 4,05, ao passo que rebites da marca B forneceu, nas mesmas unidades, os seguintes valores: 38,5 39,0 40,7 37,8 41,4. A figura 2.15 mostra o gráfico do boxplot das cargas de ruptura das marcas A e B. Com a informação acima qual das marcas de rebites é

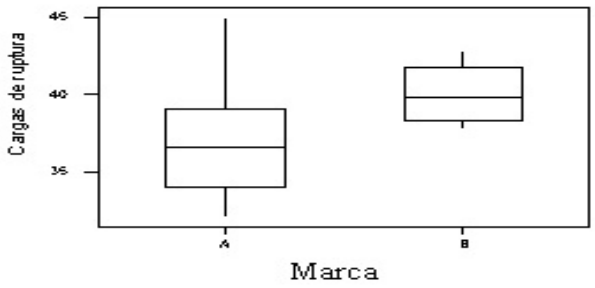


Figura 2.15: Boxplot das resistência dos rebites das marcas A e B.

melhor em pelo menos um aspecto? Justifique.

Capítulo 3

Introdução à Probabilidade

3.1 Introdução

A representação dos dados em forma sintética e compreensível, que foi o tema central do capítulo anterior, é um passo necessário, mas limitado, para viabilizar a utilização dos mesmos na análise e interpretação de processos ou na tomada de decisões.

Nesse capítulo é apresentado um conjunto de conceitos básicos da teoria de probabilidade, que constitui a parte fundamental sobre a qual se assenta a inferência estatística. Essa seria uma justificativa atribuída à teoria de probabilidade, mas, seu objetivo principal é modelar fenômenos ou processos nos quais interfere o acaso, pois faz dela um instrumento imprescindível para uma compressão dos fenômenos da natureza.

3.2 Conceitos Básicos

3.2.1 Experimentos aleatórios

Os fenômenos que ocorrem na natureza podem ser classificados em dois grupos: de um lado estão aqueles fenômenos que ocorrem naturalmente, sem a intervenção do homem. Enquanto de outro lado, estão aqueles fenômenos que ocorrem como conseqüência de experimentos realizados com a intervenção do homem. Nessas notas, a palavra experimento é usada para designar qualquer um dos dois tipos mencionados anteriormente. Pode-se dizer, portanto, que um experimento é qualquer procedimento que envolva observação. Assim, quando se efetuam medidas da massa de um elétron ou quando se observam as sucessivas posições da lua no espaço estão sendo realizados experimentos.

Um outro critério de classificação diz respeito à possibilidade de se prever ou não resultados particulares de um experimento que será realizado. Para certos experimentos, realizados sob determinadas condições, é possível prever um resultado particular. Quando a água é aquecida a 100^0 C, sob pressão normal, ela entra em ebulição. Um corpo colocado a 20m de altura e solto, cai por ação da gravidade. Esses experimentos são chamados experimentos determinísticos.

Para outros experimentos, realizados sob idênticas condições, não é possível prever um resultado particular. Se um dado é lançado sobre a superfície plana, não é possível afirmar que ocorra a face 6. Se esse experimento é realizado várias vezes, em condições idênticas, observaremos, em geral, resultados distintos. O número de pacientes que chegam a um hospital, num intervalo de tempo de uma hora, num dia varia de dia para dia. O número de lâmpadas que queimarão, 50 horas depois de 200 delas serem instaladas, não pode ser previsto com certeza. A estes experimentos denominamos de **experimentos aleatórios**(ε).

Exemplo 3.2.1 *Considere os seguintes experimentos:*

ε_1 : *Um dado é lançado sobre uma superfície plana e observamos a cara superior*

ε_2 : *Um moeda é lançada e observamos o resultado que aparece (cara ou coroa)*

Pode-se observar que um experimento aleatório tem as seguintes propriedades:

i. O experimento pode repetir-se, indefinidamente sem mudar as condições. .

ii. Cada experimento é não determinístico.

iii. Cada experimento tem vários resultados possíveis que são descritas com antecedência e com precisão. Por exemplo em ε_1 tal conjunto é $\{1, 2, 3, 4, 5, 6\}$ e, em ε_2 , é $\{cara, coroa\}$.

Exemplo 3.2.2 *Os seguintes experimentos são experimentos aleatórios:*

ε_3: *Escolher um representante ao acaso num grupo de 30 alunos.*
ε_4: *Examinar o sexo (feminino = M ou masculino = F) dos filhos em famílias com 3 filhos.*
ε_5: *Uma moeda é lançada três vezes sobre uma mesa e observado o número de caras.*
ε_6: *Observar o tempo de vida de uma lâmpada num período de um ano.*
ε_7: *Escolher ao acaso 2 vacinas de um lote que tem 2 tipos vacinas (A , B).*

3.2.2 Espaço amostral

O objetivo é construir um modelo matemático que descreva os experimentos aleatórios. Esse modelo deve ser genérico para englobar os exemplos mencionados e outros que, facilmente, possamos imaginar. Para este fim, introduzimos o conceito de espaço amostral.

Definição 3.2.1 *Denomina-se* **espaço amostral** *associado a um experimento aleatório, ao conjunto de resultados possíveis de dito experimento aleatório.*

O espaço amostral é denotado por Ω. Assim, por exemplo, os espaços amostrais associados aos respectivos experimentos dos exemplos 3.2.1-3.2.2, são:

ε_1 : $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$
ε_2 : $\Omega_2 = \{C, K\}$, C=cara e K = corõa
ε_3 : $\Omega_3 = \{R_1, \ldots, R_{30}\}$, R_i representa cada aluno: Pedro, João, Maria, etc.
ε_4 : $\Omega_4 = \{HHH, HHF, HFH, FHH, HMM, MHM, MMF, FFF\}$
ε_5 : $\Omega_5 = \{CCC, CCK, CKC, KCC, CKK, KCK, KKC, KKK\}$
ε_6 : $\Omega_5 = \{t \in R; t \geq 0\}$
ε_7 : $\Omega_6 = \{AA, AB, BA, BB\}$

3.2.3 Eventos aleatórios e operações

Muitas vezes, tem-se interesse na ocorrência de alguns resultados do experimento aleatório. Por exemplo, ao lançar um dado tem-se interesse em saber se o resultado é um número maior do que 4 ou, ao medir o tempo de vida de uma lâmpada, tem-se interesse em saber se ela durou mais de 100 horas.

Os pontos amostrais de Ω são chamados *eventos simples* e são denotados por w. Um *evento aleatório* será representado por um conjunto de eventos simples. Ou seja, um evento aleatório (ou simplesmente evento) será representado por um subconjunto de Ω e Denotado pelas letras A, B, C, etc .

Exemplo 3.2.3 *Considerando os experimentos aleatórios do exemplo 3.2.2 e os espaços amostrais respectivos, são apresentados exemplos de eventos aleatórios associados a seus respectivos* Ω.

Assim, A_i será o evento relacionado com o experimento cujo espaço amostral é Ω_i, $i = 1, \ldots, 7$.

A_1: o número observado é par;
A_2: resulte cara;
A_3: o representante escolhido seja João; $= \{\text{João}\}$
A_4: os filhos são do mesmo sexo; $=\{MMM, FFF\}$
A_5: o número de caras seja 3; $= \{3\}$
A_6: a lâmpada dure menos de 200 horas;
A_7: as 2 vacinas selecionadas sejam do tipo B; $= \{BB\}$.

Como o espaço amostral Ω é representado por um conjunto e os eventos são definidos como subconjuntos de Ω, são definidos operações entre eventos que correspondem às operações entre conjuntos. Ao se falar em eventos sempre se referira a eventos em relação a dado espaço amostral.

Um evento A ocorre quando observamos um evento simples, $w \in A$.

Sejam A e B dois eventos associados a um experimento aleatório cujo espaço amostral é Ω.

Definição 3.2.2 *A união dos eventos A e B é o evento que ocorre se pelo menos um dos eventos A ou B ocorre.*

A notação $A \cup B$ é usada para representar a união de A e B. Em notação matemática é representado por : $A \cup B = \{w \in \Omega; w \in A \text{ ou } w \in B\}$.

Definição 3.2.3 *A intersecção dos dois eventos A e B é o evento que ocorre se e somente se ambos ocorrem.*

É Denotado por AB ou $A \cap B$ o evento intersecção. Matematicamente, esse evento é representado por: $A \cap B = \{w \in \Omega; w \in A \text{ e } w \in B\}$

Exemplo 3.2.4 *Considere uma urna que contem bolas numeradas de 1 a 15. Uma bola é extraída da urna, sejam os eventos:*

A : *o número observado é múltiplo de 5 ;*
B : *o número observado é ímpar.*

Então, $\Omega = \{1, 2 \ldots, 15\}$, $A = \{5, 10, 15\}$ e $B = \{1, 3, 5, 7, 9, 11, 13, 15\}$. Assim,

$$A \cup B = \{1, 3, 5, 7, 9, 10, 11, 13, 15\},$$

ou seja, um ponto amostral pertence a $A \cup B$ se ele é ímpar ou se é múltiplo de 5. Para que um ponto amostral pertença a $A \cap B$ é necessário que ele seja ímpar e múltiplo de 5, logo, $A \cap B = \{5, 15\}$.

Definição 3.2.4 *O complementar de um evento A é o evento em que A não ocorre.*

A notação A^c ou $\bar{A}$ para designar o complementar de A e matematicamente é representada por : $A^c = \{w \in \Omega; w \notin A\}$.

No exemplo 3.2.4; $A^c = \{1, 3, 4, 6, 7, 8, 9, 11, 12, 13, 14\}$, $B^c = \{2, 4, 6, 8, 10, 12, 14\}$.

Definição 3.2.5 *Dois eventos A e B definidos no mesmo espaço amostral, são mutuamente exclusivos se não podem ocorrer juntos. Ou seja, a ocorrência de um exclui a ocorrência do outro. Em símbolos, $A \cap B = \emptyset$.*

O evento que contém todos os elementos de um espaço amostral e que, portanto, coincide com o espaço amostral é chamado *evento seguro*. Essa designação reflete o fato de que, na realização de um experimento aleatório correspondente, um dos resultados nele contido ocorre com certeza. O *evento impossível* representa-se através de um conjunto que não contém nenhum elemento do espaço amostral. Tal conjunto é representado por um *conjunto vazio*, ou seja, $\emptyset$.

3.3 Probabilidade

O conceito de probabilidade pode ser definido de diferentes maneiras. Apresenta-se seguidamente três definições distintas: a clássica, a frequentista e a axiomática.

3.3.1 Definição clássica ou a priori

Na origem, a teoria de probabilidade esteve associada aos jogos de azar (por exemplo, de dados ou de cartas). Dessa associação nasceu a definição clássica de probabilidade: se um experimento aleatório tiver $n(\Omega)$ resultados exclusivos e igualmente prováveis e se um acontecimento A tiver $n(A)$ desses resultados, então a probabilidade de ocorrer o evento A é dada por:

$$P(A) = \frac{n(A)}{n(\Omega)} \tag{3.1}$$

ou seja, a probabilidade de acorrer o evento A é a razão entre o número de resultados favoráveis à ocorrência de A e o número resultados possíveis do experimento aleatório.

Como resultado da definição acima, as probabilidades satisfazem algumas propriedades:

1. A probabilidade de ocorrência do evento A está compreendida entre 0 e 1.
2. $P(A) = 0$ se A é o evento impossível.
3. $P(A) = 1$ se A é o evento seguro.
4. Se todos os pontos amostrais de $\Omega = \{w_1, w_2, \ldots, w_n\}$ são igualmente prováveis tem-se: $P(\{w_i\}) = \frac{1}{n}$, $i = 1, \ldots, n$ e $P(\Omega) = 1$. Se A é um evento em Ω, então

$$P(A) = \sum_{w_i \in A} P(\{w_i\})$$

.

5. Se A e B são dois eventos em Ω e são mutuamente exclusivos, então

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

Exemplo 3.3.1 *Considere o lançamento de 2 dados balanceados. Calcular a probabilidade de*

(a) obter soma 7;

(b) obter soma 6;

(c) obter soma maior que 5;

(d) que o resultado do primeiro dado seja superior ao resultado do segundo.

Solução O experimento aleatório é "lançar dois dados". O espaço amostral associado a esse experimento aleatório é

$$\Omega = \left\{ \begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array} \right\}$$

onde cada ponto amostral é da forma (w_1, w_2), sendo w_1 o ponto amostral correspondente ao resultado do primeiro dado w_2, ao do segundo dado.

Sejam os seguintes eventos:

$A = \{(w_1, w_2) \in \Omega; w_1 + w_2 = 7\}$ = obter soma 7
$B = \{(w_1, w_2) \in \Omega; w_1 + w_2 = 6\}$ = obter soma 6
$C = \{(w_1, w_2) \in \Omega; w_1 + w_2 > 5\}$ = obter soma maior que 5
$D = \{(w_1, w_2) \in \Omega; w_1 > w_2\}$ = o resultado do primeiro dado ser maior que do segundo.

Uma simples contagem permite determinar $n_A = 6$, $n_B = 5$ $n_C = 26$ e $n_D = 15$. Então,

(a) $P(A) = \frac{6}{36}$ (b) $P(B) = \frac{5}{36}$
(c) $P(C) = \frac{26}{36}$ (d) $P(D) = \frac{15}{36}$

3.3.2 Definição frequentista ou a posteriori

A definição clássica não pode ser utilizada no cálculo da probabilidade de acontecimentos associados à realização da maioria dos experimentos com interesse prático, aos quais a equiprobabilidade dos resultados não se aplica. Por exemplo, se perguntamos qual é a probabilidade de que um paciente seja curado após o tratamento médico, ou qual é probabilidade de que uma determinada máquina produza artigos defeituosos. Uma forma de responder essas questões é obter alguns dados empíricos com a intenção de estimar as probabilidades.

Suponha que seja realizado um experimento n vezes (n grande) e o evento A ocorra exatamente $r \leq n$ vezes. Então, a freqüência relativa de vezes que ocorreu o evento A, "$f_{r_A} = \frac{r}{n}$", é a estimação da probabilidade que ocorra o evento A, ou seja,

$$P(A) = \frac{r}{n}.$$

Essa estimação da probabilidade por freqüência relativa de um evento A, $\frac{r}{n}$, é próxima da verdadeira probabilidade de ocorrência do evento A quando n tende ao infinito, isto é,

$$P(A) = \lim_{n\to\infty} f_{r_A} = \lim_{n\to\infty} \frac{r}{n}.$$

É imediato verificar, de acordo com a definição frequentista apresentada, que as probabilidades ainda satisfazem as propriedades apresentadas anteriormente.

Exemplo 3.3.2 *Suponha que uma moeda balanceada é lançado 1000 vezes. Os resultados desse experimento são apresentados na tabela 3.1*

Tabela 3.1: Lançamento de um moeda 1000 vezes.

Número de lançamento	Número de caras	Frequência relativa	Frequência acumulada	Freqüência ac. relativa
1 - 100	52	0,52	52	0,520
101-200	53	0,53	105	0,525
201-300	52	0,52	157	0,523
301-400	47	0,47	204	0,510
401-500	51	0,51	255	0,510
501-600	53	0,53	308	0,513
601-700	48	0,48	356	0,509
701-800	46	0,46	402	0,503
801-900	52	0,52	454	0,504
901-1000	54	0,54	508	0,508

Em um total de 1000 lançamentos ocorreram 508 caras, isto é, a freqüência relativa é aproximadamente 0,5. Portanto, baseada na definição frequentista, a probabilidade de cara em um lançamento de uma moeda balanceada é 0,5.

3.3.3 Definição axiomática

As definições anteriores são puramente empíricas ou experimentais, no entanto, após estabelecer uma forma de se determinar a probabilidade experimentalmente, pode-se deduzir leis ou propriedades da probabilidade em forma lógica ou computacional sob certas suposições chamadas de axiomas da probabilidade.

A probabilidade de um evento A é definida como o número $P(A)$, que satisfaz os seguintes axiomas:

Axioma 1 *A probabilidade $P(A)$ de qualquer evento satisfaz a relação*

$$0 \leq P(A) \leq 1$$

Axioma 2 *A probabilidade do evento certo (Ω) é*

$$P(\Omega) = 1$$

.

Axioma 3 *Se $A_1, A_2, \ldots, A_k$ são eventos mutuamente exclusivos, então*

$$P(A_1 \cup A_2 \cup, \ldots, \cup A_k) = P(A_1) + P(A_2) + \cdots + P(A_k)$$

Toda a teoria elementar da probabilidade está construída sob a base destes três simples axiomas.

A seguir, são apresentados propriedades que são conseqüência imediata dos axiomas acima.

Teorema 3.3.1 *1. Se $\emptyset$ é um evento impossível, então $P(\emptyset) = 0$*

2. Para um evento A, tem-se:

$$P(A^c) = 1 - P(A) \text{ ou } P(A) = 1 - P(A^c)$$

3. Se A e B são eventos tais que $A \subset B$, então

$$P(A) \leq P(B)$$

4. Se A e B são eventos em Ω, então

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

5. Se A, B e C são três eventos em Ω, então

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

Exemplo 3.3.3 *Na tabela 3.2 mostrada a seguir, são apresentados a composição por raça e sexo de uma população de certo país*

Tabela 3.2: Distribuição da população por raça e sexo de um país.

	Sexo		
Raça	Masculino	Feminino	Total
Branca	1726384	2110253	3836637
Outra	628 309	753125	1381434
Total	2354693	2863378	5218071

Suponha que seja selecionado um habitante desse país e considere os eventos:

H:	"o habitante selecionado é do sexo masculino"
H^c:	"o habitante selecionado é do sexo feminino"
B:	"o habitante selecionado é da raça branca"
B^c:	"o habitante selecionado é de outra raça"
$H \cap B$:	"o habitante selecionado é do sexo masculino e da raça branca"
$H \cup B$:	"o habitante selecionado é do sexo masculino ou da raça branca"
$H^c \cap B$:	"o habitante selecionado é do sexo feminino e da raça branca"
$H^c \cup B$:	"o habitante selecionado é do sexo feminino ou da raça branca"
$H^c \cap B^c$:	"o habitante selecionado é do sexo feminino e de outra raça "
$H^c \cup B^c$:	"o habitante selecionado é do sexo feminino ou de outra raça"

As probabilidades de ocorrência de cada um desses eventos são, respectivamente:

$$
\begin{aligned}
P(H) &= \tfrac{2354693}{5218071} = 0,451;\\
P(H^c) &= 1 - P(H) = 1 - 451 = 0,549;\\
P(B) &= \tfrac{3836637}{5218071} = 0,735;\\
P(B^c) &= 1 - P(B) = 1 - 0,735 = 0,265;\\
P(H \cap B) &= \tfrac{1726384}{5218071} = 0{,}331;\\
P(H \cup B) &= P(H) + P(B) - P(H \cap B)\\
&= 0,451 + 0,735 - 0,331 = 0,855;\\
P(H^c \cap B) &= \tfrac{2110253}{5218071} = 0{,}404;\\
P(H^c \cup B) &= P(H^c) + P(B) - P(H^c \cap B)\\
&= 0,549 + 0,735 - 0,404 = 0,880;\\
P(H^c \cap B^c) &= \tfrac{753125}{5218071} = 0,144\\
P(H^c \cup B) &= P(H^c) + P(B^c) - P(H^c \cap B^c)\\
&= 0,549 + 0,265 - 0,144 = 0,660.
\end{aligned}
$$

3.4 Probabilidade Condicional e Independência

Considere o exemplo 3.3.3, onde um indivíduo é selecionado, ao acaso, dentre os habitantes desse país. Caso se tenha a informação de que o indivíduo selecionado é do sexo masculino, a probabilidade de que seja da raça branca é $\frac{1726384}{2354693} = 0,73$. Esse porque do total de 2354693 de habitantes do sexo masculino, 1726384 são de raça branca. Este tipo de probabilidade chama-se probabilidade condicional e denota-se por $P(B|H)$. Lê-se a probabilidade de ocorrência do evento B dado que ocorreu o evento H.

Observe que, para o caso de experimentos aleatórios com resultados equiprováveis tem-se:

$$P(B|H) = \frac{n_{B\cap H}}{n_H} = \frac{n_{B\cap H}/n}{n_H/n} = \frac{P(B \cap H)}{P(H)} = \frac{0,331}{0,451} = 0,73.$$

Definição 3.4.1 (Probabilidade condicional) *Sejam A e B dois eventos em um mesmo espaço amostral Ω. A probabilidade condicional de A dado que ocorreu o evento B, é denotado por $P(A|B)$, é definido como:*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0. \tag{3.2}$$

Caso $P(B) = 0$, $P(A|B)$ pode ser definido arbitrariamente. Nesse texto será usado $P(A|B) = P(A)$

Exemplo 3.4.1 *Selecionamos uma semente, ao acaso, uma a uma e sem reposição, de uma sacola que contém 10 sementes de flores vermelhas e 5 de flores brancas. Qual é a probabilidade de que:*

(a) a primeira semente seja vermelha?

(b) a segunda seja branca se a primeira foi vermelha?

(c) a segunda seja vermelha se a primeira foi vermelha?

Sejam os eventos:

V_1: "a primeira semente selecionada é vermelha"
V_1^c: "a primeira semente selecionada é branca"
V_2: "a segunda semente selecionada é vermelha"
V_2^c: "a segunda semente selecionada é branca"

(a) A probabilidade de que a primeira semente seja vermelha é $\frac{10}{15} = \frac{2}{3}$. Pois há 10 sementes de flores vermelhas em um total de 15; isto é, $P(V_1) = \frac{2}{3}$.

(b) A probabilidade de que a segunda semente seja branca se a primeira foi vermelha é $\frac{5}{14}$, já que ainda existem 5 sementes brancas em um total de 14; isto é, $P(V_2^c|V_1) = \frac{5}{14}$.

(c) A probabilidade de que a segunda seja vermelha se a primeira foi vermelha é $\frac{9}{14}$, já que ainda existem 9 sementes vermelhas em um total de 14, isto é,$P(V_2|V_1) = \frac{9}{14}$.

Essas probabilidades podem ser representadas em um diagrama da árvore de probabilidades, que é mostrado na figura 3.1,

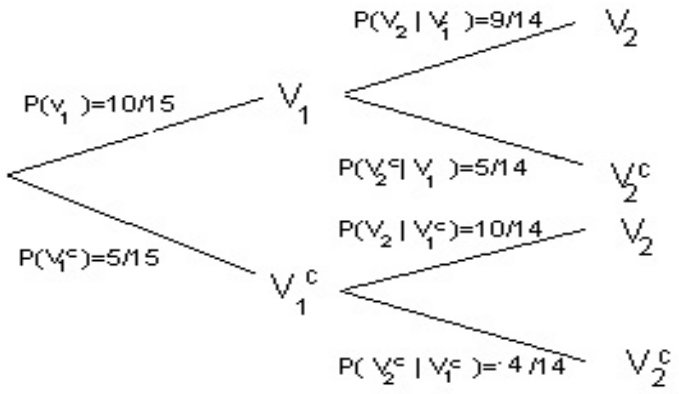


Figura 3.1: Diagrama da arvore de probabilidade

Da definição de probabilidade condicional e do teorema 3.3.1 podem ser mostrados o seguintes resultados:

Teorema 3.4.1 *Se B é um evento em Ω, tal que, $P(B) > 0$ então*

1. $P(\emptyset|B) = 0$
2. *o $A \subset \Omega$ então*
$$P(A^c|B) = 1 - P(A|B) \text{ ou } P(A|B) = 1 - P(A^c|B)$$
3. *Se A e C são eventos em Ω tal que, $A \subset C$, então*
$$P(A|B) \leq P(C|B)$$
4. *Se A e C são eventos em Ω, então*
$$P(A \cup C|B) = P(A|B) + P(C|B) - P(A \cap C|B)$$

Exemplo 3.4.2 *Em uma cidade, a probabilidade de chuva no primeiro dia de setembro é 0,50 e a probabilidade de chuva nos dois primeiros dias de setembro é 0,40. Se no primeiro dia de setembro choveu, qual é a probabilidade que no dia seguinte não chova ?*

Solução: definem-se os eventos: A : Chove no primeiro dia setembro. B: Chove no segundo dia de setembro. Do enunciado do problema tem-se : $P(A) = 0,50$ e $P(A \cap B) = 0,40$. A probabilidade pedida é $P(B^c|A)$. Pelo teorema 3.4.1, tem-se:

$$P(B^c|A) = 1 - P(B|A) = 1 - \frac{P(A \cap B)}{P(A)} = 1 - \frac{0,40}{0,50} = 0,20.$$

Exemplo 3.4.3 *Uma faculdade, em seu primeiro ano de funcionamento tem três cursos: Ciências, Administração e Engenharia. A classificação dos alunos por sexo, é apresentada na tabela a seguir.*

Tabela 3.3: Distribuição de alunos por curso e por sexo.

Sexo	Ciência	Administração	Engenharia	Total
Masculino	250	350	200	800
Feminino	100	50	50	200
Total	350	400	250	1000

Um estudante é selecionado ao acaso.

(a) Sabe-se que o estudante escolhido é do sexo masculino, qual é a probabilidade de que ele curse Ciências?

(b) Sabe-se que o estudante curse Engenharia, qual é a probabilidade de que seja do sexo feminino?

(c) Sabe-se que o estudante é do sexo feminino, qual é a probabilidade de que curse Ciências ou Administração?

Solução: Sejam os eventos:

B_1: O estudante selecionado do sexo masculino.
B_2: O estudante selecionado do sexo feminino.
A_1: O estudante é do curso de Ciências.
A_2: O estudante é do curso de Administração.
A_3: O estudante é do curso de Engenharia.

As probabilidade de ocorrência dos eventos são:

$$P(B_1) = \frac{800}{1000} = 0,80; \quad P(B_2) = \frac{200}{1000} = 0,20,$$

Essas probabilidade algumas vezes são chamadas de *probabilidades marginais*. Similarmente, $P(A_1) = \frac{250}{1000} = 0,35;$, $P(A_2) = \frac{400}{1000}$, e $P(A_3) = \frac{250}{1000} = 0,25$, são probabilidades marginais.

As probabilidade: $P(A_i \cap B_j)$, $i = 1, 2$ e $j = 1, 2, 3$ são chamados de probabilidades conjuntas. Essas probabilidades são mostradas na tabela 3.4.

Tabela 3.4: Distribuição de probabilidade conjunta e marginal do exemplo 3.4.3.

	A_1	A_2	A_3	$P(B_i)$
B_1	0,25	0,35	0,20	0,80
B_2	0,10	0,05	0,05	0,20
$P(A_j)$	0,35	0,40	0,25	1

(a) $P(A_1|B_1) = \frac{P(A_1 \cap B_1)}{P(B_1)} = \frac{0,25}{0,80} = 0,3125$

(b) $P(B_2|A_3) = \frac{P(A_3 \cap B_2)}{P(A_3)} = \frac{0,05}{0,025} = 0,20$

(c)

$$\begin{aligned} P(A_1 \cup A_2|B_2) &= P(A_1|B_2) + P(A_2|B_2) - P(A_1 \cap A_2|B_2) \\ &= \frac{P(A_1 \cap B_2)}{P(B_2)} + \frac{P(A_2 \cap B_2)}{P(B_2)} - \frac{P(A_1 \cap A_2 \cap B_2)}{P(B_2)} \\ &= \frac{0,10}{0,20} + \frac{0,05}{0,20} + 0 = 0,75. \end{aligned}$$

Da expressão (3.2), pode-se deduzir uma relação bastante útil,

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

Essa expressão é conhecida com a **regra do produto de probabilidade** ou probabilidade da intersecção que indica que a probabilidade de que ocorram os eventos A e B é igual à probabilidade de ocorrência do evento A vezes a probabilidade de que ocorrência do evento B , dado que o evento A ocorreu.

Exemplo 3.4.4 *No exemplo 3.4.1, suponha que se tenha interesse em determinar a probabilidade de que as duas sementes selecionadas sejam brancas*

Solução: O evento é $V_1^c \cap V_2^c$: "a primeira e a segunda sementes de flores são brancas"

$$P(V_1^c \cap V_2^c) = P(V_1^c)P(V_2^c|V_1^c) = \frac{5}{15} \times \frac{4}{14} = \frac{2}{21}$$

Teorema 3.4.2 *Se A, B e C são eventos de Ω, tais que $P(A) \neq 0$ e $P(A \cap B) \neq 0$, então*

$$P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$$

Exemplo 3.4.5 *Dois currais A e B têm 1000 cabeças de gado cada um. Existe uma epidemia que afeta os cascos e a boca do gado. 20% dos animais do curral A têm doença e 75% dos animais do curral B estão sadios. Escolhe-se um gado ao acaso .*

(a) Qual é a probabilidade de que o gado escolhido venha do curral A e tenha afecção aos cascos e a boca?

(b) Dos animais do curral B, afetados pela doença o 70% são menores de um ano. Qual é a probabilidade que o gado escolhido venha do curral B, tenha a doença e seja maior de um ano?

Solução: Sejam os eventos:

A: O gado escolhido é do curral A
B: O gado escolhido é do curral B
E: O gado escolhido estão afetados ao casco e boca
F: O gado escolhido tem idade acima de ano.

(a) Deve-se calcular

$$P(A \cap E) = P(A)P(E|A) = \frac{1000}{2000} \times 0,20 = 0,10.$$

(b) A probabilidade pedida é:

$$P(B \cap E \cap F) = P(B)P(E|B)P(F|B \cap E) = \frac{1000}{2000} \times (0,25) \times (0,30) = \frac{3}{80}.$$

Definição 3.4.2 (Independência de eventos) *Dois eventos A e B são independentes se a informação da ocorrência ou não de B não altera a probabilidade da ocorrência de A. Isto é,*

$$P(A|B) = P(A), \qquad P(B) > 0.$$

Conseqüentemente, dois eventos A e B são independentes se e somente se,

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

Exemplo 3.4.6 *Em uma escola 20% dos alunos tem problemas visuais, 8% problemas auditivos e 4% tem problemas visuais e auditivos. Seleciona-se um aluno dessa escola ao acaso:*

(a) os eventos de ter problemas visuais e auditivos são eventos independentes ?

(b) se o aluno selecionado tem problemas visuais, qual é a probabilidade de que tenha problemas auditivos?

(c) qual é a probabilidade de não ter problemas visuais ou o ter problemas auditivos ?

Solução: Sejam os eventos:

V: "o aluno tem problemas visuais"
A: "o aluno tem problemas auditivos"

Do enunciado do problema temos: $P(V) = 0,20$, $P(A) = 0,08$ e $P(A \cap V) = 0,04$. A partir desta informação, é possível construir a seguinte tabela:

	V	V^c	total
A	0,04	0,04	0,08
A^c	0,16	0,76	0,92
total	0,20	0,80	1,00

(a) $P(V)P(A) = 0,2 \times 0,08 = 0,16$
$P(V \cap A) = 0,04$.

Como $P(V \cap A) \neq P(V)P(A)$, A e V não são independentes.

(b) $P(A|V) = \frac{P(A \cap V)}{P(V)} = \frac{0,04}{0,20} = 0,20$

(c) $P(V^c \cup A) = P(V^c) + P(A) - P(V^c \cap A) = 0,8 + 0,08 - 0,04 = 0,84$

Uma conseqüência imediata da definição 3.4.2 é o teorema seguinte:

Teorema 3.4.3 *Se A e B, eventos em Ω, são eventos independentes, então*

(i) A e B^c são independentes;

(ii) A^c e B são independentes;

(iii) A^c e B^c são independentes.

O teorema mostra que se os eventos A e B são independentes então os complementares também são independentes. (A demonstração é deixada para o leitor)

Exemplo 3.4.7 *Sejam A e B dois eventos independentes, tais que a probabilidade de que ocorram simultaneamente os dois eventos é 1/6 e a probabilidade de que nenhum dos eventos ocorra é 1/3. Determine $P(A)$ e $P(B)$.*

Solução: Do enunciado tem-se: $P(A \cap B) = \frac{1}{6}$ e $P(A^c \cap B^c) = \frac{1}{3}$

Se A e B são independentes, então

$$P(A \cap B) = P(A)P(B) = \frac{1}{6} \tag{3.3}$$

Assim sendo A^c e B^c são também independentes (pelo teorema 3.4.3.iii). Isto é,

$$\begin{aligned}\frac{1}{3} = P(A^c \cap B^c) &= P(A^c)P(B^c) = [1-P(A)][1-P(B)] \\ &= 1 - P(A) - P(B) + P(A)P(B) = 1 - P(A) - P(B) + \frac{1}{6}. \text{ O qual implica} \\ P(B) &= \frac{5}{6} - P(A). \end{aligned} \tag{3.4}$$

Substituindo (3.4) em (3.3) vem:

$$\begin{aligned} P(A)\left[\frac{5}{6} - P(A)\right] &= \frac{1}{6} \\ P(A)^2 - \frac{5}{6}P(A) + \frac{1}{6} &= 0. \end{aligned}$$

Resolvendo a equação do segundo grau encontra-se $P(A) = 1/3$ ou $P(A) = 1/2$. Logo, o conjunto de soluções é: $\{P(A) = 1/3,\ P(B) = 1/2\}$ ou $\{P(A) = 1/2,\ P(B) = 1/3\}$.

Exemplo 3.4.8 *Um atirador acerta 80% de seus disparos e outro (na mesmas condições de tiro), 70%. Qual é a probabilidade de acertar se ambos atiradores disparam simultaneamente o alvo? Considere que o alvo foi acertado quando pelo menos uma das duas balas tenha feito impacto no alvo.*

Solução: sejam os eventos: B_i : "o atirador i acerta o alvo, $i = 1, 2$". $P(B_1) = 0,80$ e $P(B_2) = 0,70$. Logo,

$$\begin{aligned} P(B_1 \cup B_2) &= P(B_1) + P(B_2) - P(B_1 \cap B_2) \\ &= P(B_1) + P(B_2) - P(B_1)P(B_2) \\ &= 0,80 + 0,7 - (0,8)(0,7) = 0,94. \end{aligned}$$

Alternativamente, esse exemplo pode ser resolvido de uma segunda forma,

$$\begin{aligned} P(B_1 \cup B_2) &= 1 - P(B_1^c \cap B_2^c) \\ &= 1 - [1 - P(B_1)][1 - P(B_2)] \\ &= 1 - [1 - 0,80][1 - 0,70] = 0,94. \end{aligned}$$

Teorema 3.4.4 *Se $A_1, A_2, \ldots, A_n$ são n eventos em Ω independentes, então*

$$P(\bigcup_{i=1}^{n} A_i) = 1 - [1 - P(A_1)]\,[1 - P(A_2)] \ldots [1 - P(A_n)]$$

A demonstração se deixa para o leitor.

Exemplo 3.4.9 *A probabilidade de que falhe um motor em um avião é 0,10. Com quantos motores deve estar equipado um avião par ter uma seguridade de 0,999 de que o avião voe? (Suponha que é suficiente que um motor funcione para que o avião se mantenha em vôo)*

Solução: Sejam os seguintes eventos:

M_i: O motor i funciona perfeitamente, $i = 1, \ldots, n$,
A : O avião se mantém em vôo.

Os eventos M_i são independentes, e $P(M_i) = 0,9$ para $i = 1, \ldots, n$

O evento A é equivalente a: $A = \bigcup_{i=1}^{n} M_i$. Usando o teorema 3.4.4,

$$0,999 = P(A) = P(\bigcup_{i=1}^{n} M_i) = 1 - [1 - P(M_1)]\,[1 - P(M_2)] \ldots [1 - P(M_n)] = 1 - [0,1]^n.$$

Logo, $(0,1)^n = 0,001$. Daí, tem-se $n = 3$. Portanto, o avião deve ser equipado com três motores.

3.5 Teorema de Bayes

Definição 3.5.1 (Partição de um espaço amostral) *Uma coleção de eventos $B_1, B_2, \ldots, B_k$ formam uma partição do espaço amostral, se eles não tem intersecção entre si e sua união é igual ao espaço amostral completo*

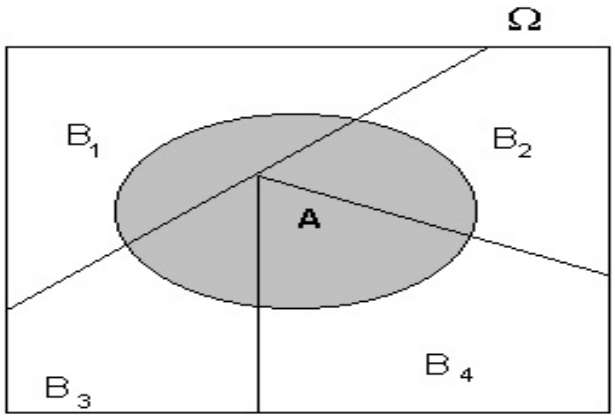


Figura 3.2: Condições de definição do teorema de Bayes para o caso de $k = 4$

Teorema 3.5.1 (Teorema da probabilidade total) *Se $B_1, B_2, \ldots, B_k$ formam uma partição do espaço amostral Ω, qualquer evento A, em Ω, satisfaz :*

$$P(A) = \sum_{i=1}^{k} P(B_i)P(A|B_i) = P(B_1)P(A|B_1) + \cdots + P(B_k)P(A|B_k)$$

Demonstração: Das condições do teorema temos que

1. $\Omega = B_1 \cup B_2 \cup, \cdots \cup B_k$, (hipóteses)
2. Para qualquer evento A em Ω tem-se
   $A = A \cap \Omega$
   $= A \cap (B_1 \cup B_2 \cup, \cdots \cup B_k)$
   $= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap A_k)$
3. Os eventos $(A \cap B_1), (A \cap B_2), \ldots, (A \cap A_k)$ são mutuamente exclusivos
4. Tomando probabilidades em ambos membros da igualdade da equação (2) vem

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap A_k) \\ &= P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \cdots + P(B_k)P(A|B_k) \\ P(A) &= \sum_{i=1}^{k} P(B_i)P(A|B_i). \end{aligned}$$

Teorema 3.5.2 (Teorema de Bayes) *Se $B_1, B_2, \ldots, B_k$ formam uma partição do espaço amostral, Ω e A é qualquer evento em Ω então*

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{k} P(B_i)P(A|B_i)}$$

Este teorema resulta de uma conseqüência imediata do teorema da probabilidade total

Exemplo 3.5.1 *Das pacientes de uma clínica de Ginecologia com idade acima de 40 anos, 70% são ou foram casadas e 30% são solteiras. E sendo solteira, a probabilidade de ter um distúrbio hormonal no último ano é 20% enquanto para as demais a probabilidade aumenta para 40%. Se um paciente é escolhido ao acaso de todas as pacientes da clínica,*

(a) qual é a probabilidade dela ter distúrbio hormonal?

(b) se a paciente escolhida resultou ter distúrbio hormonal qual é probabilidade dela ser solteira?

Solução: Sejam os eventos:

S : "a paciente sorteada seja solteira"
C : "a paciente sorteada seja casada"
D : "paciente sorteada com distúrbio hormonal"
D^C "paciente sorteada sem distúrbio hormonal."

Do enunciado tem-se: $P(S) = 0,30$, $P(C) = 0,70$, $P(D|S) = 0,20$ e $P(D|C) = 0,40$. Pelo teorema da probabilidade total dada em (3.5.1) vem:

(a) $P(D) = P(S)P(D|S) + P(C)P(D|C) = 0,30 \times 0,20 + 0,70 \times 0,40 = 0,34$ (ou 34%)

(b) Pelo teorema de Bayes tem-se :

$$P(S|D) = \frac{P(S)P(D|S)}{P(D)} = \frac{0,30 \times 0,20}{0,34} = \frac{3}{17}$$

3.6 Exercícios Resolvidos

1. Uma pesquisa de opinião determinou que a probabilidade de que uma pessoa consuma o produto A é 0,50, que consuma o produto B é 0,37 que consuma o produto C é 0,30, que consuma A e B é 0,12, que consuma somente o produto A e C é 0,08, que consuma somente B e C é 0,5 e que consuma somente C é 0,15. Obtenha a probabilidade de que uma pessoa consuma:

   (a) A ou B mas não C.

   (b) Somente A.

   Solução: Sejam os seguintes eventos:

   A: A pessoa consuma o produto A
   A^c: A pessoa não consume o produto A
   B: A pessoa consume o produto B
   B^c: A pessoa não consume o produto B
   C: A pessoa consuma o produto C
   C^c: A pessoa não consume o produto C

Do enunciado do problema tem-se:

$$P(A) = 0,50; \;\; P(B) = 0,37; \;\; P(C) = 0,30; \;\; P(A \cap B) = 0,12.$$

O evento somente A e C, escreve-se: $A \cap B^c \cap C$; logo, $P(A \cap B^c \cap C) = 0,08$.

Similarmente o evento somente B e C escreve-se: $A^c \cap B \cap C$; portanto, $P(A^c \cap B \cap C) = 0,05$.

E o evento somente C, escreve-se: $A^c \cap B^c \cap C$. Logo, $P(A^c \cap B^c \cap C) = 0,15$.

(a) Pede-se calcular a probabilidade do evento $(A \cup B) \cap C^c$.

Observe que

$$P((A \cup B) \cap C^c) = 1 - P((A^c \cap B^c) \cup C) \tag{3.5}$$

Pela propriedade de probabilidade tem-se:

$$P((A^c \cap B^c) \cup C) = P(A^c \cap B^c) + P(C) - P(A^c \cap B^c \cap C). \tag{3.6}$$

Mas,

$$\begin{aligned} P(A^c \cap B^c) &= 1 - P(A \cup B) \\ &= 1 - [P(A) + P(B) - P(A \cap B)] \\ &= 1 - [0,5 + 0,37 - 0,12] = 0,25 \end{aligned} \tag{3.7}$$

Substituindo esse valor em (3.6), vem:

$$P((A^c \cap B^c) \cup C) = 0,25 + 0,30 - 0,15 = 0,40$$

Finalmente, substituindo em (3.5) obtém-se a probabilidade pedida, ou seja,

$$P((A \cup B) \cap C^c) = 1 - 0,40 = 0,60.$$

(b) O evento somente A, escreve-se: $A \cap B^c \cap C^c$. Mas o evento A pode ser escrito como a união de eventos mutuamente exclusivos (disjuntos), isto é:

$$A = (A \cap B) \cup (A \cap B^c \cap C) \cup (A \cap B^c \cap C^c).$$

Portanto,

$$P(A) = P(A \cap B) + P(A \cap B^c \cap C) + P(A \cap B^c \cap C^c),$$

sendo

$$\begin{aligned} P(A \cap B^c \cap C^c) &= P(A) - P(A \cap B) - P(A \cap B^c \cap C) \\ &= 0,50 - 0,12 - 0,08 = 0,30. \end{aligned}$$

Uma forma prática de resolver esse exercício é levando os dados do problema para um diagrama de Venn, como se observa na figura 3.3. Além disso, observe que as probabilidades indicadas no diagrama correspondem a eventos mutuamente exclusivos. Logo,

(a) $P((A \cup B) \cap C^c) = 0,30 + 0,10 + 0,20 = 0,60$

(b) $P(A \cap B^c \cap C^c) = 0,3$

2. A probabilidade de que a construção de um prédio termine a tempo é 17/20, a probabilidade de que não haja greve é 3/4, a probabilidade de que a construção termine a tempo dado que não houve greve é 14/15 e a probabilidade de que haja greve e a construção não termine a tempo é 1/10. Qual é a probabilidade de que:

   (a) A construção termine a tempo e não haja greve?

   (b) Não haja greve dado que a construção terminou a tempo?

   (c) A construção não termine a tempo se houve greve?

   (d) A construção não termine a tempo se não houve greve?

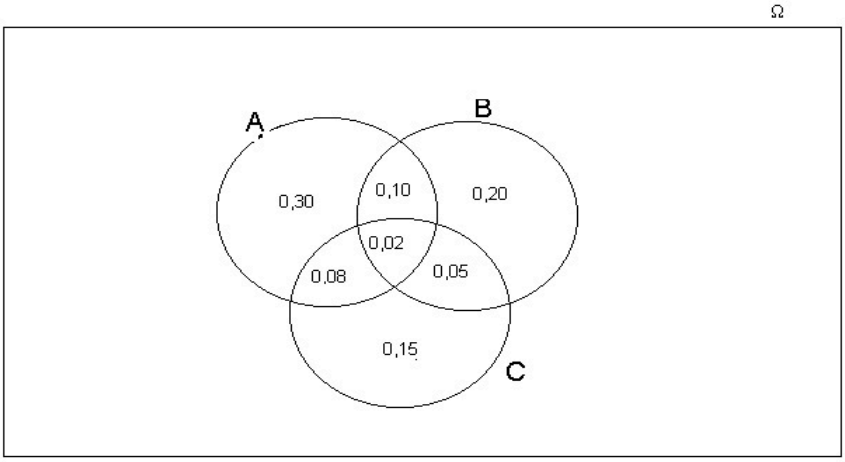


Figura 3.3: Digrama de Venn do exercício 1

Solução: Sejam os eventos

A: A construção termine a tempo,
B: Não haja greve.

Do enunciado do problema tem-se:

$$P(A) = \frac{17}{20};\ \ P(B) = \frac{3}{4};\ \ P(A|B) = \frac{14}{15}, P(A^c \cap B^c) = \frac{1}{10}$$

$$\begin{aligned}
\text{(a)} \quad P(A \cap B) &= P(B)P(A|B) = \frac{3}{4}\frac{14}{15} = \frac{7}{10} = 0,7 \text{ (pela regra do produto).} \\
\text{(b)} \quad P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{7/10}{17/20} = \frac{14}{17} \text{ (da definição de probabilidade condicional)} \\
\text{(c)} \quad P(A^c|B^c) &= \frac{P(A^c \cap B^c)}{P(B^c)} = \frac{P(A^c \cap B^c)}{1 - P(B)} = \frac{1/10}{1 - \frac{3}{4}} = \frac{2}{5}. \\
\text{(d)} \quad P(A^c|B) &= 1 - P(A|B) = 1 - \frac{14}{15} = \frac{1}{15} \text{ (pelo teorema 3.4.1.3)}
\end{aligned}$$

3. Os membros de um clube são médicos ou são advogados , 40% dos membros são médicos enquanto que 30% das mulheres, são médicas. 50% dos médicos e 30% dos advogados ganham mais de R$ 100.000 por ano. Porem, somente 20% das mulheres médicos e 10% das mulheres advogadas ganham mais de R$ 100.000, por ano. Se um membro do clube é sorteado ao acaso,

   (a) Qual é a probabilidade de que ganhe mais R$ 100.000 por ano?

   (b) Se a pessoa escolhida foi mulher, qual é a probabilidade de que ela ganhe mais de R$ 100.000 por ano.?

   Solução: Sejam os seguintes eventos:

   M: O membro do clube é médico.
   A: O membro do clube é advogado.
   F: O membro do clube é do sexo feminino.
   G: O membro do clube ganhe mais de R$ 100.000 por ano

   (a) Deve-se calcular $P(G)$.

$\Omega = A \cup M$ e $A \cap M =$. Assim, os eventos A e M formam uma partição do espaço amostral Ω (0 clube). Além disso, $G \subset \Omega$ e $G = (A \cap G) \cup (M \cap G)$. Aplicando o teorema de probabilidade total 3.5.1 temos,

$$\begin{aligned} P(G) &= P(A)P(G|A) + P(M)P(G|M) \\ &= (0,6)(0,3) + (0,4)(0,5) = 0,38. \end{aligned}$$

(b) Deve-se calcular $P(G|F)$. De (a) tem-se $G = (A \cap G) \cup (M \cap G)$. Logo,

$$\begin{aligned} P(G|F) &= P\left((A \cap G) \cup (M \cap G)|F\right) = P(A \cap G|F) + P(M \cap G|F) \\ &= P(A|F)P(G|A \cap F) + P(M|F)P(G|M \cap F) \\ &= (0,7)(0,1) + (0,30)(0,2) = 0,13. \end{aligned}$$

4. Uma empresa de desenvolvimento urbano está considerando a possibilidade de construir um centro comercial na região de Belo Horizonte. Uma condição para que essa obra seja realizada é a construção de uma estrada que une a região ao centro da cidade. Se a prefeitura aprova a construção da estrada, há uma probabilidade de 0,90 de que a empresa construa o centro comercial, no entanto se a estrada não é aprovada a probabilidade é de 0,20. Baseado na informação disponível, o presidente da empresa estima que há uma probabilidade de 0,60 de que a construção da estrada seja aprovada pela prefeitura.

   (a) Qual é a probabilidade de que a empresa construa o centro comercial ?

   (b) Se o centro comercial foi construído, qual é a probabilidade de que a estrada tenha sido aprovada pela prefeitura.?

   (c) Se o centro comercial foi construído, qual é a probabilidade de que a estrada não tenha sido aprovada pela prefeitura?

   Solução: define-se os eventos:

   A: A estrada é aprovada.
   B: O centro comercial é construído.

   (a) Deve-se calcular $P(B)$, aplicando o teorema de probabilidade total 3.5.1. O evento B é equivalente a: $B = (A \cap B) \cup (A^c \cap B)$. Logo,

$$\begin{aligned} P(B) &= P(A \cap B) + P(A^c \cap B) = P(A)P(B|A) + P(A^c)P(B|A^c) \\ &= (0,6)(0,90) + (0,4)(0,20) = 0,54 + 0,08 = 0,62. \end{aligned}$$

   (b) A probabilidade pedida é $P(A|B)$. do teorema de Bayes tem-se:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \frac{0,6 \times 0,9}{0,62} = \frac{54}{62} = 0,87$$

   (c) Deve-se calcular $P(A^C|B)$. Do teorema 3.4.1, tem-se:

$$P(A^c|B) = 1 - P(A|B) = 1 - 0,87 = 0,13.$$

5. O gerente da empresa EX viaja em um avião de 6 motores para assistir a uma reunião importante em EEUU. A probabilidade de que motor falhe é de 0,10 e cada um funciona independentemente dos outros. Precisa-se de que pelo menos um motor de cada lado do avião funcione. Qual é a probabilidade que o gerente esteja ausente na reunião por causa de um acidente com seu avião?

   Solução: Sejam os eventos:

   M_i: O i-ésimo motor funciona perfeitamente $i = 1, \ldots, 6$.
   A: O gerente esteja ausente na reunião por causa do acidente.
   A^c: O gerente não esteja ausente na reunião por causa do acidente.

   Deve-se determinar a probabilidade do evento A,isto é,

$$P(A) = 1 - P(A^c) \tag{3.8}$$

Do enunciado do problema tem-se: $P(M_i) = 0,90$, $i = 1, \ldots, 6$. Suponhamos que os motores M_1, M_2 e M_3 estejam de um lado e os motores M_4, M_5 e M_6 do outro lado. Além disso, os M_i são independentes $i = 1, \ldots, 6$.

O evento A^c é equivalente à ocorrência conjunta dos eventos,

E: Ao menos um dos motores M_i funcionam perfeitamente $i = 1, 2, 3$.
F: Ao menos um dos motores M_i funcionam perfeitamente $i = 4, 5, 6$

Ou seja $E = \bigcup_{i=1}^{3} M_i$ e $F = \bigcup_{i=4}^{6} M_i$, Portanto, $A^c = E \cap F$. Já que os eventos E e F são independentes, implica

$$\begin{aligned} P(A^c) &= P(E)P(F) = P\left(\bigcup_{i=1}^{3} M_i\right) P\left(\bigcup_{i=4}^{6} M_i\right) \\ &= (1 - [1 - P(M_1)][1 - P(M_2)][1 - P(M_3)])\,(1 - [1 - P(M_4)][1 - P(M_5)][1 - P(M_0)]) \\ &= (1 - (0,1)^3)(1 - (0.1)^3) = (0,999)^2 = 0,998001. \end{aligned}$$

A segunda igualdade da equação acima deve-se ao teorema 3.4.4. Substituindo este resultado em (3.8) temos que:

$$P(A) = 1 - 0,998001 = 0,001999.$$

6. A probabilidade de fechamento de cada relê do circuito apresentado na figura 3.4 é dado por p. Se todos os relê funcionarem independentemente, Qual é a probabilidade de que haja corrente entre os terminais L e R.?

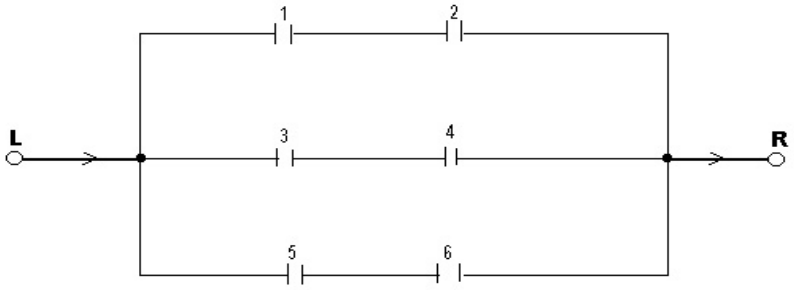


Figura 3.4: Diagrama de um circuito.

Solução: Sejam os eventos:

R_i: O relê i está fechado, $i = 1, \ldots, 6$.
A : A corrente passa por L e R.

Do enunciado do problema tem-se: $P(R_i) = p$ e $A = (R_1 \cap R_2) \cup (R_3 \cap R_4) \cap (R_5 \cap R_6)$ (observe que $(R_1 \cap R_2)$, $(R_3 \cap R_4)$ e $(R_5 \cap R_6)$ não são mutuamente exclusivos (disjuntos)). Se $B_1 = R_1 \cap R_2$, $B_2 = R_3 \cap R_4$ e $B_3 = R_5 \cap R_6$. Portanto,

$$P(A) = P(B_1 \cup B_2 \cup B_3) = P(B_1) + P(B_2) + P(B_3) - P(B_1 \cap B_2) - P(B_1 \cap B_3) - P(B_2 \cap B_3) + P(B_1 \cap B_2 \cap B_3)$$

Mas, $P(B_i) = p^2$, $i = 1, 2, 3$; $P(B_i \cap B_j) = p^4$, $i \neq j = 1, 2, 3$ e $P(\bigcap_{i=1}^{3} B_i) = p^6$. Daí tem-se:

$$P(A) = 3p^2 - 3p^4 + p^6$$

3.7 Exercícios

1. Determine um possível espaço amostral para experimentos descritos abaixo:
   (a) Um posto tem dois tipos de vacina (A e B). Três vacinas são selecionadas , uma de cada vez, ao acaso e com reposição , observando-se (i) o número de vacinas do tipo A; (ii) o número de vacinas do tipo B.
   (b) Lança-se duas moedas e anota-se a configuração
   (c) Conta-se o número de peças produzidas em um dia numa industria
   (d) Observa-se uma lâmpada até que se queime
   (e) Inspeciona-se três peças para verificar se são defeituosas ou não
2. Sejam A, B e C três eventos quaisquer no espaço amostral Ω. Expresse cada um dos eventos em termos de operações entre A, B e C.
   (a) Ocorre exatamente dois dos eventos.
   (b) Ocorre pelo menos um dos eventos.
   (c) Ocorre todos os eventos.
   (d) Não ocorre nenhum dos eventos.
   (e) Não ocorre A, ou não ocorre B ou não ocorre C.
   (f) Ocorre exatamente um dos eventos.
   (g) Ocorre pelo menos um dos eventos.
3. Um número é escolhido ao acaso, dentre os números 1,2,...,50. Qual é a probabilidade de que o número escolhido seja divisível por 6 ou por 8?
4. Sejam A e B eventos de Ω, tais que $P(A) = 0,5$, $P(B) = 0,25$ e $P(A \cap B) = 0,2$. Calcular $P(A \cup B)$, $P(A \cap B^c)$, $P(A^c \cap B^c)$, $P(A^c|B^c)$ e $P(B^c|A^c)$
5. Uma urna contém 30 bolas numeradas de 1 a 30. Três bolas são sorteadas ao acaso da urna. Qual é a probabilidade de que a soma dos números sorteados seja par?
6. Lança-se um dado 12 vezes. Determinar a probabilidade de obter:
   (a) dois "seis".
   (b) no máximo dois "seis".
7. Em um determinado exame de seleção foram propostos dois problemas. Sabendo-se que 132 indivíduos acertaram o primeiro, 86 erraram o segundo, 120 acertaram os dois e 54 acertaram apenas um problema, qual a probabilidade de que um indivíduo escolhido ao acaso dentre os que fizeram o exame:
   (a) Não tenha acertado nenhum problema.
   (b) Tenha acertado apenas o primeiro problema.
   (c) Tenha acertado apenas o segundo problema.
   (d) Tenha acertado pelo menos um problema.
8. Um número é escolhido ao acaso entre os inteiros de 1 a 20 (isto é, todos tem a mesma probabilidade). Considere os eventos: A : o número é múltiplo de 3 ; B : o número é ímpar.
   (a) Descreva os eventos: $A \cap B$, $A \cup B$ e $A \cup B^C$
   (b) calcule as probabilidades dos eventos em (a).
9. Um restaurante popular oferece dois tipos de refeições: salada completa ou um prato a base de carne. 20% dos fregueses do sexo masculino preferem salada e 30% das mulheres preferem carne . 75% dos fregueses são homens. Um freguês é escolhido ao acaso. Considere os seguintes eventos: H: freguês é homem; M : freguês é mulher; A: freguês prefere salada ; B: freguês prefere carne. Calcule as probabilidades: $P(H \cap A)$, $P(A|H)$, $P(H \cup B)$ e $P(A)$.

10. Duas ambulâncias são mantidas em um posto para atender emergência. Devido a vários problemas, como manutenção pôr exemplo, a probabilidade que cada ambulância esteja disponível é 0,9. A disponibilidade de uma ambulância é independente da outra.

    (a) Em um acidente qual é a probabilidade de que as duas ambulâncias estejam disponíveis?

    (b) Qual a probabilidade de que nenhuma esteja disponível ?

    (c) Se uma ambulância é chamada em um acidente, qual a probabilidade de que o chamado seja atendido?

11. Dois tipos de vacina foram aplicados em uma população de tal forma que 60% das pessoas receberam vacina do tipo A e as 40% restante receberam vacina do tipo B. Sabendo que a vacina do tipo A fornece 70% de imunização e a B fornece 80%, determine a probabilidade de que uma pessoa escolhida ao acaso, (i) esteja imunizado dado que foi vacinada por A; (ii) esteja imunizado; (iii) tenha sido vacinada pôr A dado que não esteja imunizado.

12. Um pedagogo deseja investigar se a "aversão"pela estatística está relacionada com o sexo. Um teste investigando atitude é administrado a 2000 estudantes para determinar seus níveis de ansiedade em relação à resolução de problemas de estatística . Cada estudante é classificado quanto a nível (alto ou baixo) de ansiedade e quanto ao sexo. Os resultados são apresentados na tabela abaixo.

Sexo/ Nível de Ansiedade	Alto	Baixo	Total
Feminino	270	630	900
Masculino	330	770	1100
Total	600	1400	2000

    (a) Se um aluno é selecionado qual é a probabilidade de que seja homem e tenha nível de ansiedade baixo?

    (b) Se o aluno selecionado é do sexo feminino, qual é a probabilidade de que tenha nível de ansiedade baixo?

    (c) Com base nesses dados verifique se o sexo e o nível de ansiedade são independentes.

13. O senhor X pode ir para sua casa usando a estrada A e a estrada B. Na estrada A ele tem probabilidade 0,25 de se atrasar devido a engarrafamento, enquanto que na estrada B essa probabilidade vale 0,35. Se ele escolhe o caminho A com probabilidade 0,7 e o caminho B com probabilidade 0,3: (i) Qual é a probabilidade de que ele se atrase devido a engarrafamento ?, (ii) se ele se atrasou qual é a probabilidade de que o senhor X tenha escolhido a estrada A.?

14. A probabilidade de uma pessoa contrair meningite durante certo ano é 0,001 se ela for vacinada 0,005 se ela não for vacinada. Se 95% da população for vacinada , (i) qual é a probabilidade de uma pessoa contrair meningite? (ii) se uma pessoa contrair meningite, qual a probabilidade dela ter sido vacinada?

15. Numa sorveteria 25% dos clientes são mulheres e o restante são homens. Dentre os homens 30% gostam de um novo sabor (jiló caramelizado) e, dentre as mulheres, apenas 20%. Escolhendo-se um cliente ao acaso

    (a) qual é a probabilidade dele ser homem e gostar desse novo sabor?

    (b) qual é a probabilidade de ser mulher ou não gostar desse novo sabor ?

    (c) qual é a proporção de clientes que gostam do novo sabor ?

    (d) se o cliente escolhido resultou mulher, qual é a probabilidade de que goste do novo sabor ?

16. Em uma universidade o 70% dos estudantes são de ciências e o 30% são de letras. Dos estudantes de ciências, 60% são homens e os de letras, 40% são homens. Escolhe-se ao acaso um estudante. Calcular a probabilidade que:

    (a) seja um estudante homem,

    (b) seja um estudante homem se é de ciências,

    (c) seja uma estudante de ciências, se é homem,

(d) seja um estudante de ciências e mulher.

17. Em uma linha de produção há dois processos A e B. No processo A há 20% de defeituosos e em B há 25%. Em um lote de 300 produtos há 200 do processo A e 100 do processo B.

    (a) Se um produto é sorteado ao acaso, qual é a probabilidade de que seja defeituoso.

    (b) Se o produto sorteado resultou ser defeituoso, qual é a probabilidade de que seja do processo B.

18. Um pesquisador desenvolveu um teste para detectar um certo tipo de doença. Ele usa o teste em pacientes com ou sem a doença. Suponha que ele aplica o teste em uma população onde a taxa de incidência da doença é igual a 2%. Sabe-se que em indivíduos sem a doença, a probabilidade do resultado do teste ser positivo é de 5% (taxa de falso positivo), enquanto que em indivíduos com a doença, a probabilidade do resultado do teste ser negativo é 20% (taxa de falso negativo). Selecionando-se um indivíduo, ao acaso, dessa população,

    (a) qual é a probabilidade de que o resultado do teste seja positivo?

    (b) qual é a probabilidade dele ser portador da doença se o resultado de seu teste foi positivo?

19. Num laboratório há três gaiolas. Na gaiola I há 2 coelhos pardos e 3 brancos, a gaiola II tem 4 coelhos pardos e 3 brancos e a gaiola III contem 5 coelhos pardos e 5 brancos . Seleciona-se, ao acaso, uma gaiola e tira-se um coelho ao acaso desta gaiola.

    (a) Qual é a probabilidade que o coelho escolhido seja branco ?

    (b) Se o coelho sorteado foi um coelho pardo, qual é probabilidade de que seja da gaiola III ?

20. No circuito elétrico dado na figura 3.5, em que consiste tensão entre os pontos A e B, determine a probabilidade de passar corrente entre A e B, sabendo-se que a probabilidade de cada chave estar fechada é 0,5 e que cada chave está aberta ou fechada independente de qualquer outra.

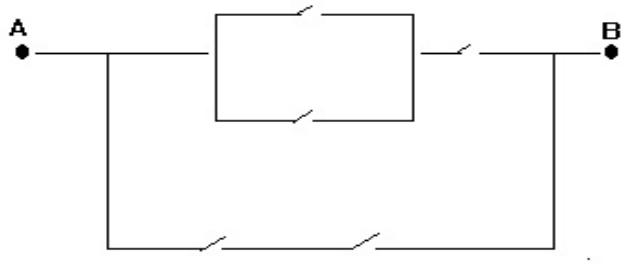


Figura 3.5: Diagrama de um circuito.

21. Em uma fábrica, a máquina 1 produz por dia o dobro de peças que máquina 2. Sabe-se que 4% das peças fabricadas pela maquina 1 tendem a ser defeituosas, enquanto 7% de defeituosas são produzidas pela máquina 2. A produção diária das maquinas é misturada.

    (a) Selecionando-se ao acaso uma peça da produção das maquinas, qual é a probabilidade que a peça seja defeituosa ?

    (b) Se a peça sorteada resultou (em (a)) ser não defeituosa, qual á probabilidade de que ela seja da maquina 1?

    (c) Se selecionamos uma amostra de 3 peças, qual é a probabilidade de que as 2 sejam defeituosas ? (considere que amostra é com reposição)

22. Uma cidade tem 30.000 habitantes e três jornais: A, B, e C. Uma pesquisa de opinião revela que 12.000 lêem A, 8.000 lêem B, 7.000 lêem A e B, 6.000 lêem C, 4.500 lêem A e C, 1.000 lêem B e C e 500 lêem A, B e C. Seleciona-se, ao acaso, um habitante dessa cidade. Qual a probabilidade de que ele leia: (a) pelo menos um jornal. (b) somente um jornal.

23. Os problemas de assédio sexual têm recebido muita atenção nos últimos anos. Em uma pesquisa, 420 trabalhadores (240 dos quais homens) consideram que uma simples batida no ombro como uma forma de assedio sexual, enquanto 580 trabalhadores (380 dos quais homens) não consideram isso como assédio sexual. Escolhido aleatoriamente um dos trabalhadores pesquisados, determine:

    (a) a probabilidade de obter alguém que não considere um simples tapa no ombro uma forma de assedio sexual.

    (b) De escolher um homem ou alguém que não considere uma simples batida no ombro como uma forma de assédio sexual.

24. Dois processadores, um do tipo A e outro do tipo B são colocados em teste por 50 mil horas. A probabilidade que um erro de cálculo aconteça em um processador do tipo A é de $2/60$, no tipo B, $1/80$ e em ambas, $1/1000$. Qual é a probabilidade de que somente o processador A ou apenas o processador B tenha apresentado erro.?

25. Uma montadora trabalha com 2 fornecedores (A e B) de uma determinada peça. As chances de que uma peça proveniente dos fornecedores A e B esteja fora das especificações são 10% e 5% respectivamente. A montadora recebe 30% das peças do fornecedor A e 70% de B.

    (a) Se uma peça do estoque inteiro é escolhida ao acaso, calcule a probabilidade de que ela esteja fora das especificações.

    (b) Se uma peça do estoque inteiro é escolhida ao acaso e verifica-se que ela está fora das especificações, de qual fornecedor ela é mais provável de ter vindo ?

26. Suponha que A e B são eventos independentes associados a um mesmo experimento aleatório, a $P(A \cup B)$ é de 0,6 enquanto que a probabilidade de que somente A ocorra é de $0,2$. Qual é probabilidade de que somente ocorra o evento B.?

27. Três maquinas A B e C apresentam, respectivamente, 10%, 20% e 30% de defeituosos na sua produção. Se as três maquinas produzem igual quantidade de peças e retiramos duas peças ao acaso da produção global qual é a probabilidade que as duas sejam perfeitas.?

28. Um dado é viciado de tal forma que a probabilidade de dar "seis"é $1/5$, sendo os demais resultados equiprováveis. Jogando-se esse dado juntamente com o dado normal, calcule a probabilidade de que

    (a) a soma dos pontos seja igual a 10.

    (b) tenha dado ponto 6 no dado viciado, sabendo que a soma dos pontos seja superior a 9.

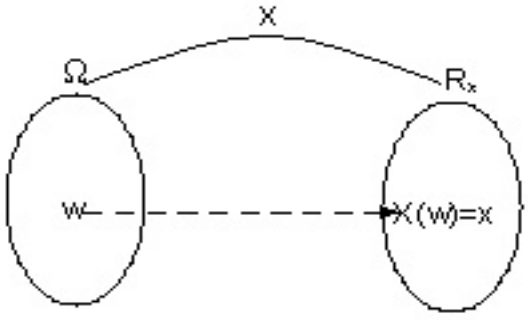
Capítulo 4

Variáveis Aleatórias

4.1 Introdução e Definição de Variável Aleatória

Na análise estatística de alguma característica (variável) de interesse da população é freqüente que seu valor numa futura observação não se pode predizer com certeza; assim por exemplo, quando se estuda o consumo dos clientes de uma loja, é difícil saber com precisão quanto gastará o seguinte cliente que ingresse na loja. Nesses casos, a análise será mas simples se for estabelecido o comportamento probabilístico da variável para assim, poder estabelecer uma metodologia para estimar seu comportamento futuro. Nesse capítulo são apresentados os procedimentos clássicos para avaliar e analisar o comportamento aleatório das variáveis.

Definição 4.1.1 (Variável aleatória) *Seja Ω o espaço amostral associado a um experimento aleatório. Uma variável aleatória, X, é uma função que tem como domínio em Ω e como contradomínio um subconjunto dos números reais,$R_X \subset R$.*



Por exemplo, retira-se, ao acaso, um artigo de um grande lote e definem-se as variáveis:

X : Número de falhas do artigo..

Y : Tempo de vida do artigo. .

O espaço amostral associado a esse experimento aleatório é:

$$\Omega = \{a_1, a_2, a_3, \ldots\}$$

Para o exemplo, os valores possíveis da variável X são $0, 1, 2, ...$, e os valores possíveis da variável Y serão números reais não negativos. Ou seja, o contradomínio das variáveis X, Y são:

$$\begin{aligned} R_X &= \{x;\ x = 0, 1, 2, 3, ...\} \\ R_Y &= \{y;\ y \geq 0, y\epsilon\mathcal{R}\} \end{aligned}$$

As variáveis aleatórias podem ser classificados, segundo o tipo de contradomínio em 2 tipos:

- *Variáveis aleatórias discretas.* Aquelas variáveis cujo contradomínio é um conjunto finito ou infinito enumerável de valores. No exemplo anterior, X é uma variável aleatória discreta pois seu contradomínio R_X é um conjunto infinito enumerável.
- *Variáveis aleatórias continuas.* Aquelas variáveis cujo contradomínio é um conjunto infinito não enumerável. No exemplo anterior, Y é uma variável aleatória continua pois seu contradomínio R_Y é o conjunto infinito não enumerável com infinitos de elementos.

4.2 Variáveis Aleatórias Discretas

4.2.1 Função de probabilidade

Se X é uma variável aleatória discreta que tem como contradomínio R_X, uma função $f(x)$ é chamada função de probabilidade da variável aleatória X se tem como domínio R_X, e como contradomínio um conjunto de número reais $P(X = x_i] = f(x_i)$ que satisfaz as seguintes condições:

1. $P[X = x_i] = f(x_i) \geq 0$, se $\quad x_i \in R_x$;
2. $0 \leq f(x_i) \leq 1, \qquad\qquad se \quad x_i \in R_x$;
3. $\sum\limits_{x_i \in R_X} f(x_i) = 1.$

Exemplo 4.2.1 *Suponha que* 3 *artigos são retirados ao acaso um a um e sem reposição de uma caixa que contém* 10 *unidades das quais* 2 *são defeituosos. Seja a variável aleatória, X : Número de artigos não defeituosos na amostra. Determinar a função de probabilidade de X.*

O espaço amostral,Ω, associado ao experimento aleatório é dado por:

$$\Omega = \{D_1D_2D_3^c, D_1D_2^cD_3, D_1^cD_2D_3 D_1D_2^cD_3^c, D_1^cD_2D_3^c, D_1^cD_2^cD_3, D_1^cD_2^cD_3^c\},$$

onde D_i e D_i^c representam respectivamente, o i-ésimo artigo defeituoso e não defeituoso, $i = 1, 2, 3$.

Como X conta o número de artigos não defeituosos, segue imediatamente que X pode assumir os valores 1, 2 e 3. Para deduzir a função de probabilidade de X, observe que o valor 1 ocorre nos eventos $\{D_1D_2D_3^c\}$, $\{D_1D_2^cD_3\}$ e $\{D_1^cD_2D_3\}$, enquanto que o valor 2, tem os eventos $\{D_1D_2^cD_3^c\}$,$\{D_1^cD_2D_3^c\}$ e $\{D_1^cD_2^cD_3\}$, e valor 3, tem apenas um evento a ele associado, ou seja, $\{D_1^cD_2^cD_3^c\}$. Segue, então, as probabilidades associadas aos valores X

$$\begin{aligned} f(1) = P[X = 1] &= P[(D_1, D_2, D_3^c) \cup (D_1, D_2^c, D_3) \cup (D_1^c, D_2, D_3)] \\ &= P[(D_1, D_2, D_3^c) + P[(D_1, D_2^c, D_3)] + P[(D_1^c, D_2, D_3)] \\ &= (2/10)(1/9)(8/8) + (2/10)(8/9)(1/8) + (8/10)(2/9)(1/8) = 1/15 \end{aligned}$$

$$\begin{aligned} f(2) = P[X = 2] &= P[(D_1, D_2^c, D_3^c) \cup (D_1^c, D_2, D_3^c) \cup (D_1^c, D_2^c, D_3)] \\ &= P[(D_1, D_2^c, D_3^c) + P[(D_1^c, D_2, D_3^c)] + P[(D_1^c, D_2^c, D_3)] \\ &= (2/10)(8/9)(7/8) + (8/10)(2/9)(7/8) + (8/10)(7/9)(2/8) = 7/15 \\ f(3) = P[X = 3] &= P[(D_1^c, D_2^c, D_3^c)] = (8/10)(7/9)(6/8) = 7/15. \end{aligned}$$

Conseqüentemente a função de probabilidade da variável aleatória X é dada por:

$$f(x) = P(X = x) \begin{cases} 1/15, & \text{se } x = 1 \\ 7/15, & \text{se } x = 2, 3 \\ 0, & \text{caso contrário} \end{cases} \tag{4.1}$$

O gráfico dessa distribuição de probabilidade é:

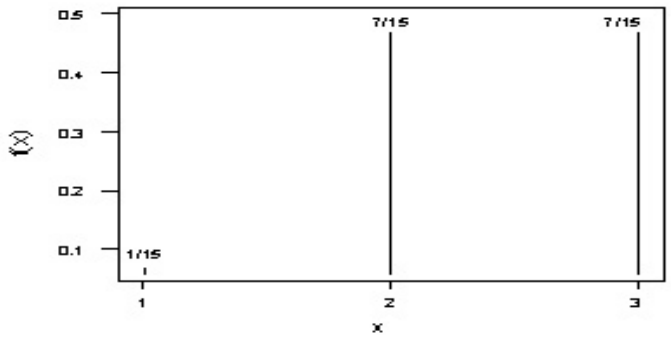


Figura 4.1: Gráfico da função de probabilidade da variável aleatória X.

4.2.2 Função de distribuição acumulada de uma variável aleatória discreta

Outro conceito importante no desenvolvimento dos seguintes capítulos é a função de distribuição acumulada ou simplesmente função de distribuição (FDA) de uma variável aleatória.

Definição 4.2.1 *Seja X uma variável aleatória discreta com contradomínio $R_X = \{x_1, x_2, \dots\}$ e função de probabilidade $f(x_i) = P(X = x_i)$. Seja $x \in R$, a função de distribuição acumulada de X denotado por $F(x)$, é definida como:*

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i) = \sum_{x_i \leq x} P(X = x_i) \ \ \textit{onde} \ \ x_i \in R_X$$

Exemplo 4.2.2 *Considere o exemplo 4.2.1. Determine a função de distribuição da variável aleatória X : número de artigos não defeituosos. Ou seja, $F(x)$.*

Neste caso $R_X = \{1, 2, 3\}$ portanto,

$$
\begin{aligned}
\text{Se } x<1 \qquad & F(x) = \quad P(X\leq x)=0\\
\text{Se } x=1 \qquad & F(1) = \quad P(X\leq 1)=\sum_{x_i\leq 1}P(X=x_i)=P(X=1)=f(1)=\frac{1}{15}\\
\text{Se } 1\leq x<2 \qquad & F(x) = \quad P(X\leq x)=\sum_{x_i\leq x}P(X=x_i)=P(X=1)=\frac{1}{15}=\frac{1}{15}\\
\text{Se } x=2 \qquad & F(2) = \quad P(X\leq 2)=\sum_{x_i\leq 2}P(X=x_i)=P(X=1)+P(X=2)=\frac{1}{15}+\frac{7}{15}=\frac{8}{15}\\
\text{Se } 2\leq x<3 \qquad & F(x) = \quad P(X\leq x)=\sum_{x_i\leq x}P(X=x_i)=P(X=1)+P(X=2)=\frac{1}{15}+\frac{7}{15}=\frac{8}{15}\\
\text{Se } x=3 \qquad & F(3) = \quad P(X\leq 3)=\sum_{x_i\leq 3}P(X=x_i)=P(X=1)+P(X=2)+P(X=3)\\
& \qquad = \qquad \frac{1}{15}+\frac{7}{15}+\frac{7}{15}=1\\
\text{Se } x\geq 3 \qquad & F(3) = \quad P(X\leq x)=\sum_{x_i\leq x}P(X=x_i)=P(X=1)+P(X=2)+P(X=3)=1
\end{aligned}
$$

Observação 4.2.1 *Pode-se observar, que se* $x\in[1;\ 2)$*, então* $F(x)=F(1)$*, se* $x\in[2;\ 3)$*,* $F(x)=F(2)$*. Em geral, se* $x\in[x_l;\ x_{l+1})$*, então* $F(x)=F(x_l)$*, onde* x_l *e* x_{l+1} *são elementos de* R_x*.*

Logo, a função de distribuição pode ser escrito como:

$$
F(x)=\begin{cases}0, & \text{se } x<1\\ \frac{1}{15}, & \text{se } 1\leq x<2\\ \frac{8}{15}, & \text{se } 2\leq x<3\\ 1, & \text{se } x\geq 3\end{cases} \tag{4.2}
$$

Na figura 4.2, é apresentado o gráfico da FDA da variável aleatória X.

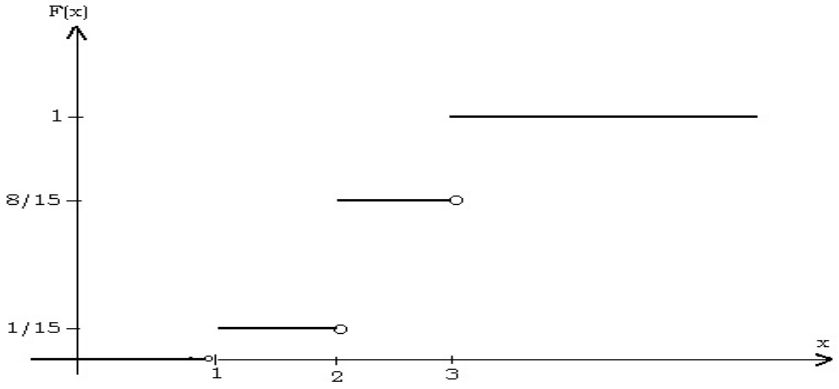


Figura 4.2: Gráfico da função de distribuição acumulada

Propriedades da função de distribuição

Sendo $F(x)$ a FDA da variável aleatória discreta X com contradomínio R_X, deve satisfazer as seguintes propriedades:

1. Para todo $x \in R$, $0 \leq F(x) \leq 1$.
2. $F(x)$ é uma função monótona não decrescente.
3. $$\lim_{x \to -\infty} F(x) = 0 \quad \text{e} \quad \lim_{x \to +\infty} F(x) = 1.$$
4. Se $R_x = \{x_1, x_2, \ldots, \}$ tal que, $x_1 < x_2 < \ldots$, então $f(x_i) = P(X = x_i) = F(x_i) - F(X_{i-1})$
5. Se $a, b \in R$ tal que $a < b$, então
   (i) $P(X \leq a) = F(a)$.
   (ii) $P(X \geq a) = 1 - P(X < a)$
   (iii) $P(a < X \leq b) = F(b) - F(a)$
   (iv) $P(a \leq X \leq b) = F(b) - F(a) + P(X = a)$
   (v) $P(a < X < b) = F(b) - F(a) - P(X = b)$

Exemplo 4.2.3 *A variável aleatória X tem a seguinte função de distribuição:*

$$F(x) = \begin{cases} 0, & se\ x < 0 \\ 1/8, & se\ 0 \leq x < 1 \\ 1/2, & se\ 1 \leq x < 2 \\ 5/8, & se\ 2 \leq x < 3 \\ 1, & se\ x \geq 3 \end{cases}$$

Calcular: (a) $P(1 < X \leq 3)$; (b) $P(X \geq 2)$; (c) A função de probabilidade da variável aleatória X.

Da propriedade 5.iii da FDA temos que

(a) $P(1 < X \leq 3) = F(3) - F(1) = 1 - 1/2 = 1/2$

(b) Da propriedade 5.i da FDA: $P(X \geq 2) = 1 - P(X < 2) = 1 - F(1) = 1 - 1/8 = 7/8$

(c) Da função da distribuição acumulada, tem-se $R_X = \{0, 1, 2, 3\}$. Considerando, a propriedade 4 da FDA, pode-se mostrar que a função de probabilidade da variável aleatória X é:

$$f(x) = P(X = x) = \begin{cases} 1/8, & \text{se } x = 0, 2 \\ 3/8, & \text{se } x = 1, 3 \\ 0, & \text{caso contrário} \end{cases}$$

4.3 Variáveis Aleatórias Contínuas

4.3.1 Função de probabilidade

Uma função $f(x)$ é chamada função de probabilidade ou função densidade de probabilidade da variável aleatória contínua X se satisfaz as seguintes condições.

1. $f(x) \geq 0$, se $x \epsilon \mathcal{R}$

2. $\int_{-\infty}^{\infty} f(x)\,dx = 1$

3. Seja o evento $A = \{x/\ \ a \leq x \leq b\}$. Assim, $P[A] = P[x\epsilon A] = P[a \leq x \leq b] = \int_a^b f(x)\,dx$

Exemplo 4.3.1 *Suponha que o tempo de produção de um artigo (em minutos) é uma variável aleatória (v.a.) X que tem como função densidade de probabilidade:*

$$f(x) = \begin{cases} \frac{(5-x)}{4}, & se\ \ 2 \leq x \leq 4 \\ 0 & caso\ contrário \end{cases} \tag{4.3}$$

Verificar se $f(x)$, é uma função de densidade de probabilidade e calcular a probabilidade do tempo de produção de um artigo, escolhido ao acaso ser menor que 3 minutos.

A figura 4.3.1, mostra o gráfico da função de probabilidade de X.

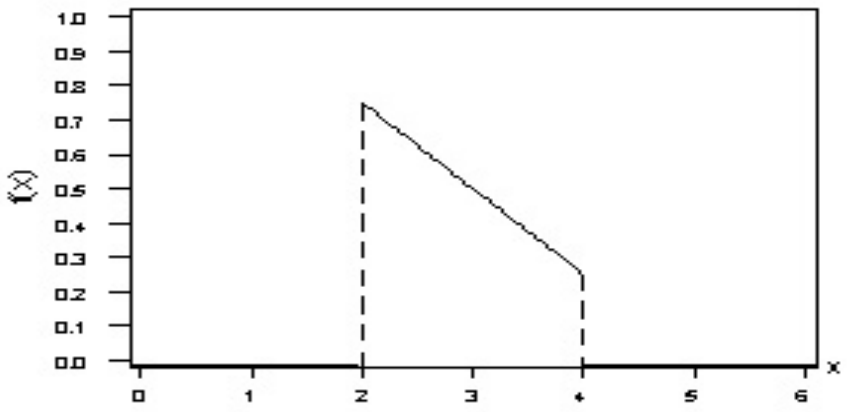


Figura 4.3: Função de densidade da va X do exemplo 4.3.1.

Da figura pode-se observar que a função, $f(x) \geq 0$ (é não negativa) para $x \in R$. Para que seja uma função de densidade é preciso verificar se a área sob eixo x e a função $f(x)$ é igual a 1. Isto é, a integral de $-\infty$ a $+\infty$ deve ser igual a *um.*

$$\begin{aligned} \int_{-\infty}^{\infty} f(x)\,dx &= \int_{-\infty}^{2} f(x)\,dx + \int_{2}^{4} f(x)\,dx + \int_{4}^{\infty} f(x)\,dx = \int_{2}^{4} f(x)\,dx \\ &= \int_{2}^{4} \frac{5-x}{4}\,dx = \frac{1}{4}(5x - \frac{x^2}{2})\ |_2^4 = 1 \end{aligned}$$

Logo, a probabilidade do tempo de produção de um artigo escolhido ao acaso ser menor que 3 minutos é a probabilidade do evento: $A = \{x \in R_X; x < 3\}$, ou seja,

$$\begin{aligned} P(A) = P(X < 3) &= \int_{-\infty}^{3} f(x)\,dx = \int_{-\infty}^{2} f(x)\,dx + \int_{2}^{3} f(x)\,dx = \int_{2}^{3} f(x)\,dx \\ &= \int_{2}^{3} \frac{5-x}{4}\,dx = \frac{1}{4}(5x - \frac{x^2}{2})\ |_2^3 = \frac{5}{8}. \end{aligned}$$

Observação 4.3.1 *Se X é uma variável aleatória contínua, então*

$P(X = x) = 0$, *para todo* $x \in R_X$

$P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b)$, *para todo* $a, b \in R_X$

$P(X \leq a) = P(X < a)$, *para todo* $a \in R$.

4.3.2 Função de distribuição acumulada de uma variável aleatória contínua

Definição 4.3.1 *Seja X uma variável aleatória contínua (VAC) com função densidade de probabilidade $f(x)$. A função de distribuição acumulada (FDA) da VAC X, é definida como*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt \quad \textit{para todo} \quad x \in R.$$

Exemplo 4.3.2 *Considere a variável aleatória X do exemplo 4.3.1. Determine a FDA de X.*

Dos intervalos da definição de $f(x)$ apresentados em (4.3), tem-se:

Se $x < 2$, tem-se $f(x) = 0$. Logo, $F(x) = 0$.

Se $2 \leq x \leq 4$ tem-se

$$F(x) = \int_{-\infty}^{x} f(t)dt = F(x) = \int_{-\infty}^{2} f(t)dt + \int_{3}^{x} f(t)dt = 0 + \int_{2}^{x} \frac{5-t}{4}dt = -\frac{(5-t)^2}{8}|_2^x = \frac{9-(5-x)^2}{8}.$$

Se $x > 4$ tem-se:

$$F(x) = \int_{-\infty}^{x} f(t)dt = \underbrace{\int_{-\infty}^{2} f(t)dt}_{0} + \int_{2}^{4} f(t)dt + \underbrace{\int_{4}^{x} f(t)dt}_{0} = \int_{2}^{4} f(t)dt = 1$$

Logo, a FDA da variável X é:

$$F(x) = \begin{cases} 0, & \text{se } x < 2 \\ \frac{9-(5-x)^2}{8}, & \text{se } 2 \leq x \leq 4 \\ 1, & \text{se } x \geq 4 \end{cases} \tag{4.4}$$

O gráfico da FDA da variável aleatória X:

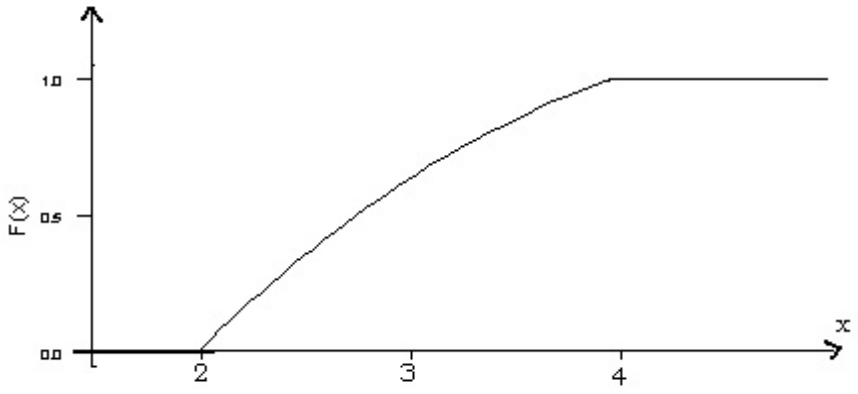


Figura 4.4: Função de distribuição acumulada da variável aleatória X, do exemplo 4.3.1.

Observação 4.3.2 *A FDA, além de caracterizar uma variável aleatória contínua X, permite o cálculo de probabilidades de eventos da forma $(a \leq X \leq b)$, onde $a < b \in R$. Isto é*

$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a)$$

Exemplo 4.3.3 *Considere A FDA, exemplo 4.3.2, obtenha: $P(X < 3)$ e $P(2,5 \leq X < 3,5)$*

Considerando a FDA apresentada em (4.4), tem-se:

$$\begin{aligned} P(X<3) &= F(3)=\frac{9-(5-3)^2}{9}=\frac{5}{8}. \\ P(2,5\leq X<3,5) &= F(3,5)-F(2,5)=\frac{9-(5-3,5)^2}{9}-\frac{9-(5-2,5)^2}{9}=0,5. \end{aligned}$$

Propriedades da função de distribuição

1. $0\leq F(x)\leq 1$, para todo $x\in R$.
2. $F(x)$ é uma função monótona não decrescente.
3. $$\lim_{x\to-\infty}F(x)=\lim_{x\to-\infty}\int_{-\infty}^{x}f(t)dt=0 \quad \text{e} \quad \lim_{x\to+\infty}F(x)=\lim_{x\to+\infty}\int_{-\infty}^{x}f(t)dt=1$$
4. F(x) é função contínua para todo $x\in R$
5. Do segundo teorema fundamental do cálculo tem-se:
$$f(x)=\frac{d}{dx}F(x)=\frac{d}{dx}\int_{-\infty}^{x}f(t)dt$$

.

Exemplo 4.3.4 *Suponha que o tempo de vida de um microorganismo seja uma variável aleatória X com a seguinte FDA:*

$$F(x)=\begin{cases} 1-ke^{-\frac{x}{2}}, & x\geq 0 \\ 0, & x<0 \end{cases}$$

(a) Para que valor de k, $F(x)$ é uma FDA da variável X.

(b) Determinar: $P(X\geq 2)$, $P(2<X\leq 4)$ e$P(X\geq -1)$.

(c) Determinar a função de densidade de X.

(c) Determinar a função de densidade da variável aleatória $Y=2X+1$.

(a) Uma vez que $F(x)$ é uma função contínua, para todo $x\in R$, tem-se que: $F(0)=0$, ou seja, $1-ke^{-0}=0$, o qual resulta em $k=1$. Logo,

$$F(x)=\begin{cases} 1-e^{-\frac{x}{2}}, & x\geq 0 \\ 0, & x<0 \end{cases}$$

é a FDA de X

(b_1) $P(X\geq 2)=1-P(X<2)=1-F(2)=1-[1-e^{-1}]=e^{-1}$.

(b_2) $P(2<X\leq 4)=F(4)-F(2)=1-e^{-2}-(1-e^{-1})=e^{-1}-e^{-2}$.

(b_3) $P(X>-1)=1-P(X\leq -1)=1-0$.

(c) Da propriedade 5, da FDA contínua, tem-se:

$$f(x)=\frac{d}{dx}F(x)=\begin{cases} \frac{1}{2}e^{-\frac{x}{2}}, & x\geq 0 \\ 0, & x<0 \end{cases}$$

(c) Seja $F_Y(y)$ a FDA da variável aleatória $Y = 2X + 1$, então,

$$F_Y(y) = P(Y \leq y) = P(2X + 1 \leq y) = P\left(X \leq \frac{y-1}{2}\right) = F(\frac{y-1}{2}) = \begin{cases} 1 - e^{-\frac{\frac{y-1}{2}}{2}}, & \frac{y-1}{2} \geq 0 \\ 0, & \frac{y-1}{2} < 0. \end{cases}$$

Logo,

$$f(y) = \frac{d}{dy}F_Y(y) = \begin{cases} \frac{1}{4}e^{-\frac{y-1}{4}}, & y \geq 1 \\ 0, & y < 1 \end{cases}$$

4.4 Valor Esperado e Variância

Definição 4.4.1 (Valor esperado de uma variável aleatória) *Seja X uma variável aleatória com função de probabilidade ou função densidade de probabilidade,$f(x)$. O valor esperado, ou esperança matemática ou média da variável aleatória, denotado por $E(X) = \mu_X$, é definida como:*

1. *Se X é uma variável aleatória discreta,*

$$E(X) = \sum_{x \in R_X} xf(x).$$

2. *Se X é uma variável aleatória contínua,*

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Nessa definição supõe-se que somatório e a integral convergem. Em caso contrário dizemos que o valor esperado da variável aleatória X não existe.

Definição 4.4.2 (Valor esperado de uma função de variável aleatória) *Seja $Y = g(X)$, sendo $g(.)$ uma função real e contínua na variável aleatória X. O valor esperado de $g(X)$, é definida como:*

1. *Se X é uma variável aleatória discreta,*

$$E(g(X)) = \sum_{x \in R_X} g(x)f(x),$$

2. *Se X é uma variável aleatória contínua,*

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx,$$

Como anteriormente, supõe-se que tanto a somatório quanto a integral convergem.

Definição 4.4.3 (Variância de uma variável aleatória) *Seja X uma variável aleatória com função de probabilidade $f(x)$, com média $E(X) = \mu_X$, a variância da variável aleatória, X, denotado por $Var(X) = \sigma^2$ é definida como o valor esperado da variável aleatória $(X - \mu_X)^2$.*

1. *Se X é uma variável aleatória discreta,*

$$Var(X) = E[(X - \mu_X)^2] = \sum_{x \in R_X} (x - \mu_X)^2 f(x).$$

2. *Se X é uma variável aleatória contínua,*

$$Var(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx.$$

4.4.1 Propriedades do valor esperado e variância de uma variável aleatória

Sejam X e Y duas variáveis aleatórias definidas no mesmo espaço amostral Ω e a e b duas constantes reais. É possível mostrar as seguintes propriedades:

1. $E(a) = a$.
2. $E(aX) = aE(X)$
3. $E(aX \pm b) = aE(X) \pm b$
4. $E(aX \pm bY) = aE(X) \pm bE(Y)$
5. $Var(a) = 0$
6. $Var(aX) = a^2 Var(X)$
7. Se X e Y são variáveis aleatórias independentes[1], $V(aX \pm bY) = a^2 Var(X) + b^2 Var(Y)$.

Teorema 4.4.1 *Se X é uma variável aleatória com média, μ_X, então*

$$Var(X) = E(X^2) - (\mu_X)^2$$

A demonstração é deixada por conta do leitor.

Exemplo 4.4.1 *Suponha que tem-se 3 caixas (C_1, C_2 e C_3) com dois tipos de ampolas (A e B). A caixa C_1 contem 40 ampolas dos quais 10 são do tipo A e 30 de B, a caixa C_2 tem 20 ampolas do tipo A e 20 do tipo B e a caixa C_3, somente tem ampolas do tipo B. Sorteia-se ao acaso, uma ampola de cada caixa e define-se a variável aleatória Y como número de ampolas escolhidos do tipo B.*

(a) Determine o espaço amostral e a função de probabilidade de Y.

(b) Calcule a média e variância do número de ampolas do tipo B.

Solução:

(a) Seja B_i: a ampola do tipo B escolhida da caixa i e A_i : a ampola do tipo A escolhida da caixa i Logo, o espaço amostral é $\Omega = \{A_1A_2B_3, A_1B_2B_3, B_1A_2B_3, B_1B_2B_3\}$,

w_i	$A_1A_2B_3$	$A_1B_2B_3$	$B_1A_2B_3$	$B_1B_2B_3$
$P(\{w_i\})$	$\frac{10}{40} \times \frac{20}{40} \times 1$	$\frac{10}{40} \times \frac{20}{40} \times 1$	$\frac{30}{40} \times \frac{20}{40} \times 1$	$\frac{30}{40} \times \frac{20}{40} \times 1$
$Y(\{w_i\})$	1	2	2	3

Portanto, a variável aleatória Y, assume os valores 1, 2 e 3. Da tabela anterior as probabilidades associadas aos valores de Y são as seguintes:

$$\begin{aligned} f(1) = P[Y=1] &= P(A_1A_2B_3) = \frac{1}{8}. \\ f(2) = P[Y=2] &= P(\{A_1B_2B_3\} \cup \{B_1A_2B_3\}) = P(\{A_1B_2B_3\}) + P(\{B_1A_2B_3\}) = \frac{4}{8}. \\ f(3) = P[Y=3] &= P(\{B_1B_2B_3\}) = \frac{3}{8}. \end{aligned}$$

Logo, a função de probabilidade (f.p) da variável aleatória é dado por:

$$f(y) = P(Y=y) = \begin{cases} \frac{1}{8}, & \text{se } y = 1 \\ \frac{4}{8}, & \text{se } y = 2 \\ \frac{3}{8}, & \text{se } y = 3 \\ 0, & \text{caso contrário} \end{cases}$$

[1]Se as variáveis aleatórias X e Y são independentes a distribuição conjunta de probabilidades de X e Y ($f(x,y)$) é igual ao produto de cada uma das distribuições marginais ($f_X(x)$ e $f_Y(y)$.) Isto é, $f(x,y) = f_X(x)f_Y(x)$)

A f.p da variável aleatória Y, pode ser representada na tabela de distribuição de probabilidade:

y	1	2	3
$f(y) = P[Y = y]$	$\frac{1}{8}$	$\frac{4}{8}$	$\frac{3}{8}$

(b) A média e variância de Y.

$$\begin{aligned} E(X) = \sum_y yf(y) &= 1 \times \frac{1}{8} + 2 \times \frac{4}{8} + 3 \times \frac{3}{8} = 2,25 \\ E(X^2) = \sum_y y^2 f(y) &= 1^2 \times \frac{1}{8} + 2^2 \times \frac{4}{8} + 3^2 \times \frac{3}{8} = 5,5 \end{aligned}$$

Da definição da média e variância tem-se:

$$\begin{aligned} \mu_y = E(Y) &= 2,25 \\ \sigma_y^2 = Var(Y) &= E(Y^2) - \mu_y^2 = 5,5 - 2,25^2 = 0,4375 \end{aligned}$$

Exemplo 4.4.2 *Suponha que as vendas diárias de uma Drogaria (em dezenas de milhares de dólares) é uma variável aleatória com função de densidade;*

$$f(x) = \begin{cases} x, \ \textit{se}, & 0 \leq x < 1 \\ 2 - x, \ \textit{se}, & 1 \leq x < 2 \\ 0, \ \textit{caso contrário} \end{cases}$$

Escolhe-se ao acaso um dia de venda. Determine:

(a) A probabilidade de que as vendas da Drogaria seja maior de 5.000 dólares mais não superior a 1.5.000 dólares.

(b) A média e o desvio padrão das vendas diárias.

(c) Se o lucro diário é definido pela função $Y = 0,2X - 0,1$, calcule a média e variância do lucro diário.

Solução: Seja X : Vendas diárias de uma Drogaria (Dezenas de milhares de dólares)

(a) Seja o evento $A = \{x \in R_X; 0,5 < x \leq 1,5\}$, então se deseja determinar : $P(A) =?$

$$\begin{aligned} P(0,5 < X \leq 1,5) &= \int_{0,5}^{1,5} f(x)dx = \int_{0,5}^{1} xdx + \int_{1,0}^{1,5} (2-x)dx \\ &= \left(\frac{x^2}{2}\right) |_{0,5}^{1} + \left(2x - \frac{x^2}{2}\right) |_{1}^{1,5} = \frac{3}{4} \end{aligned}$$

(b) Da definição da esperança matemática temos

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 x^2 dx + \int_1^2 x\,(2-x)\,dx = 1,0 \\ E(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^1 x^3 dx + \int_1^2 x^2\,(2-x)\,dx = \frac{7}{6} \end{aligned}$$

Logo, a média e o desvio padrão de X são respectivamente:

$$\begin{aligned} \mu_X &= E(X) = 10.000,0 \text{ dólares.} \\ \sigma_X &= \sqrt{Var(X)} = \sqrt{E(X^2) - \mu_x^2} = \sqrt{\frac{7}{6} - 1^2} = \sqrt{\frac{1}{6}} = 4082,4829 \text{ dólares.} \end{aligned}$$

Esses valores indicam que, a longo prazo (um número elevado de dias), espera-se que as vendas diárias da drogaria mostrem um comportamento com uma média de 10.000 dólares e um desvio padrão de 4082,4829 dólares, mesmo que as vendas tenham flutuações aleatórias.

(c) Seja $Y = g(X) = 0,2X - 0,1$. Das propriedades da esperança matemática, vem

$$\begin{aligned} \mu_y &= E(g(X)) = E(0,2X - 0,1) = 0,2E(X) - 0,1 = 0,2(1) - 0,1 = 0,1. \\ \sigma_y^2 &= Var(0,2X - 0,1) = 0,2^2 Var(X) = 0,2^2(\frac{1}{6}) = 0,0067. \end{aligned}$$

Esses valores indicam que a longo prazo (um número elevado de dias), espera-se que os lucros diários da drogaria mostrem um comportamento com uma média de 1.000 dólares e uma variância 0,0067(dezenas de milhares de dólares)^2.

4.5 Principais Modelos Discretos

Algumas variáveis discretas geradas mediante processos de contagem podem ser associadas a funções de probabilidade que tenham um comportamento particular conhecido. Assim, por exemplo, quando se estuda o número de artigos defeituosos em um lote ou quando se estuda o número de pessoas que chegam a um estabelecimento comercial num certo período de tempo, entre outros. Nesses casos, é possível estudar o comportamento de tais variáveis através de funções de probabilidade particulares em cada caso. Nessa seção, são apresentadas algumas das principais funções de probabilidade ou distribuições de probabilidade, que podem ser utilizadas para analisar variáveis, tais como as descritas anteriormente.

4.5.1 Ensaio e distribuição de Bernoulli

Há muitos experimentos que tem somente dois resultados possíveis, chamado de *sucesso* (S) e *fracasso* (F). Logo, o espaço amostral para esse tipo de experimento é $\Omega = \{S, F\}$. Por exemplo, ao lançar uma moeda, obtém-se somente dois resultados possíveis, cara (C) ou coroa (K). Chama-se de sucesso ao evento de interesse. No exemplo, caso o interesse seja "cara", obtém-se um sucesso quando no ensaio ocorre cara. Caso contrário, obtém-se um fracasso.Um experimento com essa característica chama-se de experimento ou ensaio de **Bernoulli**.

Seja a variável aleatória X, definida como o número de sucessos num ensaio de Bernoulli. Então, o contradomínio de X é dado por $R_X = \{1, 0\}$. Isto é, $X(S) = 1$ se o resultado do ensaio é sucesso e $X(F) = 0$, se o resultado é fracasso. A variável aleatória assim definida chama-se *variável aleatória de Bernoulli.* Sejam $P(E) = p$ e $P(F) = q = 1 - p$ as probabilidade de sucesso e fracasso respectivamente. A distribuição de probabilidade da variável aleatória X de Bernoulli, é chamada de **distribuição de Bernoulli,** e é dada por

x	0	1
$f(x) = P[X = x]$	q	p

A distribuição de Bernoulli pode, também ser expressa como uma função $f(x)$, dada por

$$f(x) = P[X = x] = \begin{cases} p^x(1-p)^{1-x}, & x = 0, 1 \\ 0, & \text{caso contrário.} \end{cases}$$

A média e variância da variável aleatória X, são respectivamente

$$\begin{aligned} \mu_X &= E(X) = 0 \times q + 1 \times p = p. \\ \sigma_X^2 &= Var(X) = E(X^2) - \mu_x^2 = 0^2 \times q + 1^2 \times p - p^2 = p(1-p) \end{aligned}$$

Denota-se por $X \sim bernoulli(p)$ para indicar que a variável aleatória X tem distribuição Bernoulli com parâmetro p.

4.5.2 Distribuição Binomial

Existem muitos problemas, nos quais o experimento consiste em n ensaios (ou experimentos) de Bernoulli $\varepsilon_1, \ldots, \varepsilon_n$, uma seqüência de ensaios de Bernoulli forma um processo de Bernoulli ou experimento Binomial quando satisfazer as seguintes condições:

(i) Cada ensaio tem somente dois resultados possíveis S ou F.

(ii) Os ensaios são independentes. Isto é, o resultado (sucesso ou fracasso) de qualquer ensaio é independente do resultado de qualquer outro ensaio.

(iii) A probabilidade de sucesso, p, permanece constante de ensaio em ensaio. Logo, a probabilidade de fracasso $q = 1 - p$ também é constante.

Exemplo 4.5.1 *Suponha um experimento onde uma moenda é lançada três vezes e suponha que p seja a probabilidade de cara. Seja X a variável aleatória que representa o número de caras obtidas ao final dos três lançamentos. Achar a distribuição de probabilidade de X.*

Solução. O espaço amostral para experimento de lançar uma moeda três vezes é:

$$\Omega = \{KKK, KKC, KCK, CKK, KCC, CKC, CCK, CCC\}.$$

Seja X_i $(i = 1, 2, 3)$ a variável aleatória de Bernoulli que representa o número caras no lançamento i. Então a variável

$$X = X_1 + X_2 + X_3,$$

representa o número de caras nos 3 lançamento da moeda. Pode-se mostrar que $X_i \sim bernoulli(p)$.

w_i	$P(\{w_i\})$	$X_1(w_i)$	$X_2(w_i)$	$X_3(w_i)$	$X(w_i) = X_1(w_i) + X_2(w_i) + X_2(w_i)$
KKK	$(1-p)^3$	0	0	0	0
KKC	$(1-p)^2p$	0	0	1	1
KCK	$(1-p)^2p$	0	1	0	1
CKK	$(1-p)^2p$	1	0	0	1
KCC	$(1-p)p^2$	0	1	1	2
CKC	$(1-p)p^2$	1	0	1	2
CCK	$(1-p)p^2$	1	1	0	2
CCC	p^3	1	1	1	3

O contradomínio da variável X é: $R_X = \{0, 1, 2, 3\}$. Logo,

$$\begin{aligned} P[X=0] &= P(\{KKK\}] = (1-p)(1-p)(1-p) = (1-p)^3 \\ P[X=1] &= P(\{KKC\}) + P(\{KCK\}) + P(\{CKK\}) = 3p(1-p)^2 \\ P[X=2] &= P(\{KCC\}) + P(\{CKC\}) + P(\{CCK\}) = 3p^2(1-p) \\ P[X=3] &= P(\{CCC\}) = p^3 \end{aligned}$$

A distribuição de probabilidades da variável aleatória X é dada por

x	0	1	2	3
$f(x) = P[X = x]$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3

O comportamento de X fica completamente determinado pela função,

$$f(x) = \begin{cases} \binom{3}{x} p^x (1-p)^{3-x}, & x = 0, 1, 2, 3 \\ 0, & \text{caso contrário} \end{cases}$$

onde $\binom{3}{x} = \frac{3!}{x!(3-x)!}$. Observe que as probabilidades correspondem aos termos do desenvolvimento em binômio de Newton de $(p + (1-p))^2$, o que justiça o nome distribuição Binomial escolhido para esse modelo.

Definição 4.5.1 (Distribuição Binomial) *Considere a repetição de n ensaios de Bernoulli independentes todos com a mesma probabilidade de sucesso p. A variável aleatória que conta o número total de sucessos nos n ensaios de Bernoulli, é denominada de variável aleatória Binomial com parâmetros n e p e sua função de probabilidade é dado por*

$$f(x) = P[X = x] = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x = 0, 1, \ldots, n \\ 0, & \textit{caso contrário} \end{cases}$$

onde $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ *representa o coeficiente Binomial.*

A notação $X \sim B(n, p)$, é usado para indicar que a variável X tem distribuição Binomial com parâmetros n e p.

Propriedades da distribuição Binomial

Se $X \sim B(n,p)$ então:

(a) E(X)=np.

(b) Var(X)=np(1-p)

A demonstração dessas propriedades é deixada como exercício, para o leitor.

Exemplo 4.5.2 *Suponha que o nascimentos de menino e menina seja igualmente prováveis e que o nascimento de qualquer criança não afeta a probabilidade do sexo do próximo nascimento. Determine a probabilidade de:*

(a) Exatamente 4 meninos em 10 nascimentos.

(b) Ao menos 4 meninos em 10 nascimentos.

(c) No máximo um menino em 10 nascimentos.

Solução: Seja a variável aleatória X número de meninos em 10 nascimentos.

$$R_X = \{0, 1, \ldots, n\}$$

O evento de interesse é nascimento de menino. Então define-se

S : "nascimento de um menino."
F :"nascimento de uma menina."
$P(S) = P(F) = 1/2$

Do enunciado do problema a variável aleatória X tem distribuição Binomial (satisfaz as condições de um experimento Binomial) com parâmetros $n = 10$ e $p = 0,5$, com função de probabilidade é dada por:

$$P[X = x] = \begin{cases} \binom{10}{x}\left(\frac{1}{2}\right)^{10}, & x = 0, 1, \ldots, 10, \\ 0, & \text{caso contrário} \end{cases}$$

(a) $P(X = 4) = \binom{10}{4}\left(\frac{1}{2}\right)^{10} = \frac{210}{1024} = 0,205078$

(b) $P(X \geq 4) = 1 - p(X < 4) = 1 - (P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3]) = 1 - 0,05469 = 0,94531$

(c) $P(X \leq 1) = P[X = 0] + P[X = 1] = \frac{1}{1024} + \frac{10}{1024} = \frac{11}{1024} = 0.01074$

Exemplo 4.5.3 *O professor da disciplina de Estatística e probabilidade elaborou uma prova de múltipla escolha, constituída de 10 questões, cada uma com 4 alternativas. Suponha que todos estudantes que irão a fazer a prova não assistem as aulas e não estudaram para a mesma (o que é muito freqüente). O professor estabeleceu que para aprovar deve acertar ao menos 6 questões. Se 100 alunos se apresentaram, quantos alunos foram aprovados na disciplina?*

Solução. Uma vez que todos os estudantes, que farão a prova não assistem as aulas ou não estudaram, a escolha de cada resposta em cada uma das 10 questões será feita ao acaso. Portanto, a escolha da resposta de cada questão é considerada de um ensaio de Bernoulli, com

$$p = \text{Probabilidade de acertar a resposta correta} = \frac{1}{4}, \;\; q = 1 - p = \frac{3}{4}.$$

A variável aleatória definida, X : número de questões respondidas corretamente nas 10 questões com $R_X = \{0, 1, \ldots, n\}$, tem distribuição Binomial. Isto é, $X \sim B(10, 1/4)$.

$$P[X=x]=\begin{cases}\binom{10}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{10-x}, & x=0,1,\ldots,10,\\ 0, & \text{caso contrário}\end{cases}$$

Para ser aprovado o estudante deve responder ao menos 6 questões corretas. Isto é, a probabilidade de ser aprovado a prova é.

$$P(X\geq 6)=\sum_{x=6}^{10}\binom{10}{x}\left(\frac{1}{4}\right)^x\left(\frac{3}{4}\right)^{10-x}=0,0197.$$

Portanto, dos 100 alunos que se apresentaram para a prova, seriam aprovados $100(0,0197)\approx 2$ alunos.

Aplicações da Distribuição Binomial numa Amostra

O sorteio de uma amostra de n elementos de uma população pode ser considerada como um experimento que consiste de n ensaios (ou experimento) de Bernoulli. Os n ensaios serão independentes nos seguintes casos:

(a) Quando os elementos da amostra são sorteados com ou sem reposição de uma população infinita. Obviamente, o resultado de um sorteio qualquer é independente do outro sorteio e a proporção p de sucessos ($P(S)=p$) permanece constante em cada sorteio. Então, é aplicável a distribuição Binomial.

(b) Quando os elementos da amostra são sorteados com reposição de uma população finita. Suponha que a população tenha N elementos, dos quais k são de certa classe que temos interesse. Define-se, assim, a variável X : numero de elementos da classe de interesse na amostra de tamanho n.

Os sorteios individuais são ensaios de Bernoulli, onde elemento da classe de nosso interesse corresponde "sucesso"e o experimento de tomar uma amostra de tamanho n com reposição consiste nos n ensaios independentes de Bernoulli onde $p=P(sucesso)=\frac{k}{N}$; isto é, X tem distribuição binomial,

$$f(x)=\binom{n}{x}\left[\frac{k}{N}\right]^x\left[1-\frac{k}{N}\right]^{n-x}, \quad x=1,\ldots,n$$

Exemplo 4.5.4 *Numa população grande de Drosophila, o 25% das moscas tem mutação de asas. Seleciona-se, aleatoriamente 300 moscas da população para uma exame de mutação de asas. A variável aleatória X é definida como o número de moscas que têm mutação na amostra. Determinar o valor esperado e a variância de X*

Como a população é grande (infinita), não interessa se amostragem é com ou sem reposição, portanto, X tem distribuição Binomial com parâmetros $n=300$ e $p=0,25$, isto é $X\sim B(300,0,25)$

A função de probabilidade de X é

$$f(x)=\binom{300}{x}(0,25)^x(0,75)^{300-x}, \quad x=0,1,\ldots,n$$

A média

$$E(X)=np=300\times 0,25=75$$

Variância

$$Var(X)=np(1-p)=75\times\frac{3}{4}=\frac{225}{4}$$

4.5.3 Distribuição Hipergeométrica

Suponha uma população finita com N elementos , divididos em duas classes. Uma classe com M ($M<N$) elementos (sucesso) e a outra com $N-M$ elementos (fracasso). Por exemplo, no caso particular de N peças produzidas, podem ser consideradas as classe: M artigos defeituosos e (N-M) artigos não defeituosos.

Considere o seguinte experimento, uma amostra aleatória de tamanho n ($n < N$) sem reposição é sorteada da população finita de N elementos. A variável aleatória é definida da seguinte forma,

X : Número de elementos com a característica de interesse(sucessos) na amostra de tamanho n.

A variável aleatória assim definida chama-se variável aleatória Hipergeométrica e sua função de probabilidade é:

$$f(x) = P(X = x) = \begin{cases} \frac{\binom{M}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, & x = 0, 1, \ldots, \min\{n, M\} \\ 0, & \text{caso contrário} \end{cases}$$

A notação $X \sim H(N, M, n)$, indica que a variável aleatória X tem distribuição Hipergeométrica com parâmetros N, M e n.

Propriedades da distribuição Hipergeométrica

Se $X \sim H(N, M, n)$, então

(a) $E(X) = n\frac{M}{N}$

(b) $Var(X) = n\frac{M}{N}(1 - \frac{M}{N})(\frac{N-n}{N-1})$

Exemplo 4.5.5 *Suponha que o gerente de credito de um estabelecimento recebe 10 pedidos de credito, dos quais 4 têm documentação incompleta e devem ser devolvidas aos clientes. Escolhe-se, ao acaso 5 pedidos sem reposição obter:*

(a) a probabilidade de devolver mais de 3 pedidos de crédito.

(b) A média e o coeficiente de variabilidade de variável X.

Seja X : número de pedidos de crédito devolvidos numa amostra de 5 pedidos . Neste caso considera-se "sucesso", se o pedido de credito é devolvido . Portanto $X \sim H(10, 4, 5)$, ou seja,

$$f(x) = P(X = x) = \begin{cases} \frac{\binom{4}{x}\binom{6}{5-x}}{\binom{10}{5}}, & x = 0, 1, 2, 3, 4 \\ 0, & \text{caso contrário} \end{cases}$$

(a) A probabilidade pedida é:

$$P(X > 3) = P(X = 4) = f(4) = \frac{\binom{4}{4}\binom{6}{5-x}}{\binom{10}{5}} = \frac{1}{42} = 0,0238.$$

(b) $\mu_X = E(X) = 5 \times \frac{4}{10} = 2$ e $\sigma_X^2 = Var(X) = 5(\frac{4}{10})(1 - \frac{4}{10})(\frac{10-5}{10-1}) = \frac{2}{3} = 0,6667$ e $CV = \frac{\sigma_X}{\mu_X} \times 100\% = \frac{\sqrt{0,6667}}{2} \times 100\% = 40,28\%$

Distribuição binomial como aproximação da distribuição hipergeométrica

Nas distribuição binomial e hipergeométrica só há duas possibilidades mutuamente exclusivas, de ocorrência em cada prova; porem a primeira se refere à realização de n ensaios, em condições idênticas (extração com reposição) enquanto que na hipergeométrica a composição é alterada após a realização de cada prova. Verificamos, porém, que se N (tamanho da população) for muito grande em relação n ($f = n/N < 0,1$) praticamente não há variação nas condições dos ensaios, que podem então ser considerada como extração com reposição.

Assim, a distribuição binomial pode ser usada como limite da distribuição quando n for suficientemente pequeno em relação a N. Isto é, Se $X \sim H(N, M, n)$ e $f = \frac{n}{N} < 0,10$ então $X \sim B(n, \frac{M}{n})$.

Exemplo 4.5.6 *Foram colocados em uma caixa 100 peças, 40 dos quais foram fabricadas pela industria B e as outras pela indústria A. Retiradas, sem reposição, 8 peças, qual é a probabilidade de que sejam 4 da indústria A?*

Solução: Seja a variável aleatória X o número de peças da industria B. A distribuição exata de X é a hipergeométrica. Isto é, $X \sim H(100, 40, 8)$

A probabilidade pedida é :

$$P(X = 4) = \frac{\binom{40}{4}\binom{60}{4}}{\binom{100}{8}} = 0,2395$$

Já que $f = \frac{8}{100} = 0,08 < 0,10$ tem-se $X \sim B(8, \frac{40}{100})$. (aproximadamente) Logo,

$$P(X = 4) = \binom{8}{4} 0,4^4 0,6^4 = 0,2322.$$

4.5.4 Distribuição de Poisson

A distribuição de Poisson é uma das distribuições discretas mais importantes pois que se aplica a muitos problemas práticos. A distribuição de Poisson pode ser obtida de duas formas. A primeira se deduz a partir de um *processo de Poisson* e a segunda como limite da distribuição Binomial.

Inicialmente é apresentada a idéia intuitiva de um processo de Poisson. Muitos problemas consistem em observar a ocorrência de eventos discretos num intervalo contínuo (unidade de medida), por exemplo, o número de manchas (falhas) por unidade de medida (digamos $1m^2$) no esmaltado de uma geladeira. Pode-se encontrar 0 manchas, 1 mancha, 2 manchas, ou talvez mais, num metro quadrado. Isto é, podemos contar o número de falhas por unidade de medida. Sendo impossível contar o número de pontos sem manchas (é infinito não enumerável). Além disso, as falhas são eventos discretos, uma vez que ocorre em pontos isolados na área de 1 m^2. Ao se definir a variável aleatória X : número de manchas em um metro quadrado, o contradomínio é $R_X = \{0, 1, \ldots, \}$

Outro exemplo é contar o número de chamadas que chegam a uma central telefônica de uma empresa num intervalo de tempo (de 8,00 horas a 10,00 horas, por exemplo) num dia determinado. Podem chegar 0 chamadas, 1 chamada, 2 chamadas, etc. É um evento discreto, visto que o tempo de chegada de qualquer delas é um ponto isolado num período de 2 horas. Pode-se também contar, número de bactérias em um cm^3 de água. Nesse caso, o intervalo contínuo é o número de bactérias é um evento discreto supondo que se possa considerar cada bactéria como um ponto no espaço.

Os eventos discretos gerados num intervalo contínuo (unidade: comprimento, área, volume, tempo, etc.) formam um processo de Poisson com parâmetro λ se satisfazer as seguintes propriedades:

1. O número médio de ocorrência dos eventos numa unidade de medida (comprimento, área, volume, tempo, etc.) é conhecido e igual a λ.
2. A ocorrência de um evento numa unidade de medida h não afeta a ocorrência ou a não ocorrência em outra unidade de medida h contígua. Isto é, a ocorrência dos eventos em unidades de medida contíguas são independentes.
3. Seja uma unidade de medida suficientemente pequeno de comprimento h, logo:
   - a probabilidade de sucesso nessa unidade de medida é proporcional ao comprimento do intervalo , isto é, λh;
   - a probabilidade da ocorrência de 2 ou mais sucessos, nessa unidade de medida pequena é aproximadamente igual a zero.

Definição 4.5.2 *Uma variável discreta X tem distribuição de Poisson com parâmetro μ se sua função de probabilidade é dada por*

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}, \quad x = 0, 1, 2, \ldots, \tag{4.5}$$

onde

X	*numero de eventos discretos em t unidades de medida.*
λ	*é a média de eventos discretos em uma unidade de medida.*
t	*número de unidade de medida.*
$\mu = \lambda t$	*é a média de eventos discretos em t unidades de medidas.*

A notação $X \sim P_o(\mu)$ é para indicar que a variável aleatória X tem distribuição de Poisson com parâmetro μ. A média e a variância de variável aleatória com distribuição de Poisson com parâmetros μ são:

$$E(X) = \mu$$

$$Var(X) = \mu.$$

Exemplo 4.5.7 *Suponha que a central telefônica de empresa de grande porte recebe, em média, 3 chamadas cada 4 minutos. Qual é probabilidade que a central recepcione 2 ou menos chamadas em um intervalo de 2 minutos?*

Solução: Se, X: número de chamadas que recebe a central telefônica da empresa em intervalos de 2 minutos, então $X \sim P_o(\mu = \lambda t)$. Aqui, $\lambda = 3/4 = 0.75$, $t = 2$, então $\mu = \lambda t = 0,75 \times 2 = 1,5$. Daí, $X \sim P_o(1,5)$ ou seja, a variável aleatória X tem a seguinte função de probabilidade:

$$f(x) = P[X = x] = \frac{e^{-1,5}1,5^x}{x!}, \quad x = 0, 1, \ldots.$$

$P(X \leq 2) = P[X = 0] + P[X = 1] + P[X = 2] = e^{-1,5}[1 + 1,5 + \frac{1,5^2}{2}] = 0,808847.$

Exemplo 4.5.8 *Sabe-se que um líquido particular contem certas bactérias a razão de 4 bactérias por cm^3. Uma amostra de $1cm^3$ desse líquido é tomado. (a) Qual é a probabilidade que a amostra não contenha nenhuma bactéria.? (b) Qual é a probabilidade de que em $0,5cm^3$ do líquido haja pelo menos uma bactéria?*

Solução: (a) Seja a variável aleatória X : número de bactérias em $1cm^3$ do líquido. Aqui $\lambda = 4$, $t = 1$ e $\mu = \lambda t = (4)(1) = 1$. Então $X \sim P_o(4)$. A função de probabilidade da variável aleatória X é dada por:

$$f(x) = P(X = x) = \frac{4^x e^{-4}}{x!}, \quad x = 0, 1, \ldots.$$

$P(X = 0) = e^{-4} = 0,0183$

(b) X : O número de bactérias em $0,5cm^3$ do liquido. Aqui $\lambda = 4$, $t = 0,5$ e $\mu = \lambda t = (4)(0,5) = 2$. Então $X \sim P_o(2)$.

$$f(x) = P(X = x) = \frac{2^x e^{-2}}{x!}, \quad x = 0, 1, \ldots.$$

$P(X \geq 1) = 1 - P(X < 1) = 1 - P(X = 0) = 1 - e^{-2} = 0,864.$

Distribuição de Poisson com aproximação da distribuição Binomial

Será mostrado agora, a distribuição de Poisson como um limite da distribuição Binomial, com $\mu = np$ é considerado que $p = P(S)$ é suficientemente pequena ($p \longrightarrow 0$) e n é suficientemente grande ($n \longrightarrow \infty$), de tal forma que np permaneça constante. A distribuição binomial para x sucessos em n ensaio de Bernoulli é dada por:

$$P[X = x] = \binom{n}{x} p^x q^{n-x}, \quad x = 0, \ldots, n.$$

Considera-se $\mu = np$. Logo $p = \frac{\mu}{n}$ e $q = 1 - p = 1 - \frac{\mu}{n}$. Substituindo-se na função de probabilidade tem-se:

$$\begin{aligned} P[X = x] &= \frac{n!}{x!(n-x)!}(\frac{\mu}{n})^x(1-\frac{\mu}{n})^{n-x} \\ &= \frac{n!}{(n-x)!n^x} \times \frac{\mu^x}{x!} \times \frac{(1-\frac{\mu}{n})^n}{(1-\frac{\mu}{n})^x} \\ &= \frac{n(n-1)(n-2)\ldots(n-(x-1))(n-x)!}{n^x(n-x)!} \times \frac{\mu^x}{x!} \times \frac{(1-\frac{\mu}{n})^n}{(1-\frac{\mu}{n})^x} \\ &= (1-\frac{1}{n})(1-\frac{2}{n})\ldots(1+\frac{x+1}{n}) \times \frac{\mu^x}{x!} \times \frac{(1-\frac{\mu}{n})^n}{(1-\frac{\mu}{n})^x} \end{aligned}$$

(1) Se $n \longrightarrow \infty$, então $\frac{x}{n} \longrightarrow 0$ e,$\frac{x+1}{n} \longrightarrow 0$

(2) $p = \frac{\mu}{n}$ é pequeno ($p \longrightarrow 0$), então $(1-\frac{\mu}{n})^x \longrightarrow 1$

(3) sabe-se que $(1-\frac{\mu}{n})^n = e^{-\lambda}$ se $n \longrightarrow \infty$.

de (1), (2) e (3) temos que para $n \longrightarrow \infty$,

$$P[X = x] = \frac{\mu^x}{x!}e^{-\mu}$$

Observação 4.5.1 *Da forma como foi obtido essa aproximação, a distribuição de Poisson pode ser utilizado para aproximar probabilidades de uma distribuição Binomial quando n é suficientemente grande ($n \longrightarrow \infty$) e p é muito pequeno ($p \longrightarrow 0$). Na prática considera-se que a aproximação é aceitável se $np < 5$ ou $n(1-p) < 5$. Nesse caso, considera-se que $X \sim P_o(np)$.*

Exemplo 4.5.9 *Uma vacina imuniza contra polio num 99,99%. Supondo que a vacina foi administrada a 10.000 pessoas.*

(a) Qual é número esperado de pessoas não imunizados ?

(b) Qual é a probabilidade de se ter exatamente k pessoas não imunizadas?

(c) Qual é probabilidade de se ter menos de 2 pessoas não imunizadas?

Solução: X número de pessoas não imunizadas nas 10.000 vacinadas. $R_X = \{0, 1, \ldots, 10.000\}$. A probabilidade que uma pessoa não seja imunizado é 0,0001, ou seja $P(S) = p = 0,0001$ e $n = 10.000$, portanto $X \sim B(10.000, 0,0001)$

(a) $E(X) = np = (10.000)(0,0001) = 1$.

(b) $E(X) = 1 < 5$ então $X \sim P_o(1)$, portanto $P[X = k] = \frac{e^{-1}}{k!}$

(c) $P(X \leq 1) = P[X = 0] + P[X = 1] = 2e^{-1} = 0.7358$

Propriedade reprodutiva da distribuição de Poisson

A propriedade reprodutiva de algumas distribuições de probabilidades é a seguinte: em que, se duas ou mais variáveis aleatórias independentes, com a distribuições do mesmo tipo, se somam, a variável resultante tem uma distribuição do mesmo tipo da soma. Essa propriedade chama-se propriedade reprodutiva.

Teorema 4.5.1 *Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes , com distribuição de Poisson com parâmetros $\mu_1, \ldots, \mu_n$, respectivamente te, então a variável aleatória*

$$Y = X_1 + \cdots + X_n,$$

tem distribuição de Poisson com parâmetros $\mu = \mu_1 + \cdots + \mu_n$.

Exemplo 4.5.10 *Em uma fábrica foram registrados em três semanas a média de acidentes: 2,5 na primeira semana, 2 na segunda semana e 1,5 na terceira semana. Suponha que o número de acidentes por semana segue um processo de Poisson. Qual é a probabilidade de que haja 4 acidentes nas três semanas?*

Solução:

Definem-se as variáveis aleatórias com distribuição de Poisson com parâmetro μ_i, $(i = 1, 2, 3)$.

X_1: Número de acidentes na primeira semana.

X_2: Número de acidentes na segunda semana.

X_3: Número de acidentes na terceira semana.

As três variáveis aleatórias são independentes. A variável aleatória $X = X_1 + X_2 + X_3$ pelo teorema 4.5.1, tem distribuição de Poisson com parâmetro $\mu = 2,5 + 2 + 1,5 = 6$. Isto é, $X \sim P_o(6)$

$$P(X = 4) = \frac{6^4\ e^{-6}}{4!} = 0,1339.$$

4.6 Principais Modelos Contínuos

Nessa seção são apresentados algumas das principais distribuições contínuas.

4.6.1 Distribuição uniforme

Definição 4.6.1 *Uma variável aleatória contínua X tem distribuição uniforme com parâmetros α e β se sua função de densidade é dado por:*

$$f(x) = \begin{cases} \frac{1}{\beta-\alpha}, & \alpha \leq x \leq \beta \\ 0, & \textit{caso contrário.} \end{cases} \tag{4.6}$$

A função da distribuição acumulada de uma variável aleatória uniforme contínua é:

$$F(x) = \begin{cases} 0; & x < \alpha \\ \frac{x-\alpha}{\beta-\alpha}, & \alpha \leq x < \beta \\ 1, & x \geq \beta \end{cases} \tag{4.7}$$

Na figura 4.5, é mostrada a representação gráfica da função de densidade de probabilidade e da função de distribuição acumulada da variável aleatória uniforme contínua.

A média e variância de uma variável aleatória X, com distribuição uniforme no intervalo $\alpha \leq x \leq \beta$ são dadas por:

$$E(X) = \frac{\alpha + \beta}{2} \ \text{ e } \ Var(X) = \frac{(\alpha - \beta)^2}{12} \tag{4.8}$$

A notação $X \sim U(\alpha, \beta)$ é usada para indicar que X tem distribuição uniforme no intervalo (α, β).

4.6.2 Distribuição exponencial

Definição 4.6.2 *Uma variável aleatória contínua X tem distribuição exponencial com parâmetro λ, se sua função de densidade é dada por*

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{x}{\lambda}}, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{4.9}$$

A média e a variância de uma variável aleatória X, com distribuição exponencial são dadas por:

$$E(X) = \lambda \ \text{ e } \ Var(X) = \lambda^2. \tag{4.10}$$

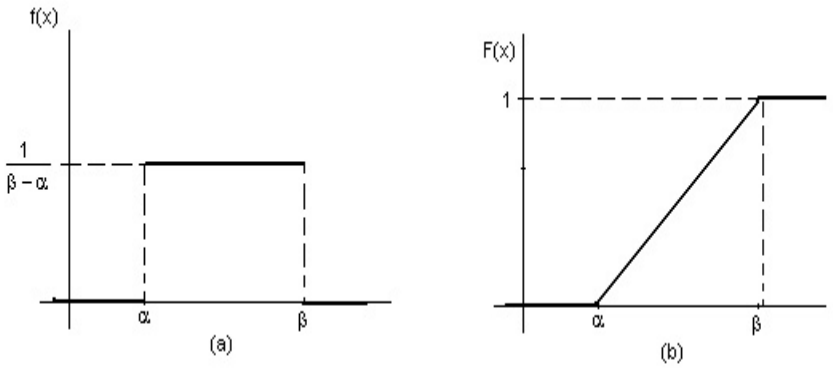


Figura 4.5: Função de: (a) densidade e (b) distribuição acumulada, da distribuição uniforme

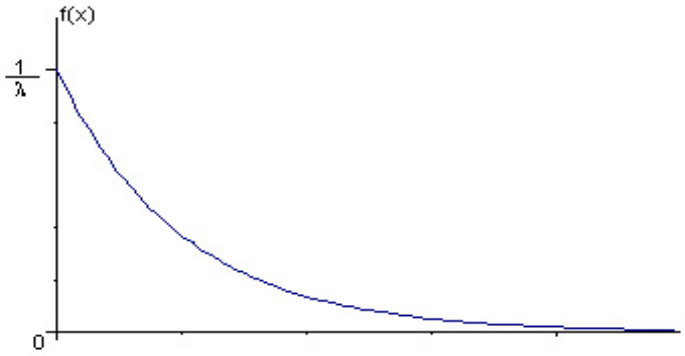


Figura 4.6: Função de densidade de probabilidade de $X \sim Ex(\lambda)$.

A notação $X \sim Ex(\lambda)$ indica que a variável aleatória X tem distribuição exponencial com parâmetro λ.

Na figura 4.6, é apresentado o gráfico da densidade

A função da distribuição acumulada de uma variável aleatória contínua com distribuição exponencial com parâmetro λ:

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\frac{x}{\lambda}}, & x > 0 \end{cases} \tag{4.11}$$

Exemplo 4.6.1 *O tempo de vida (em horas) de um transistor é uma variável aleatória X com f.d.p*

$$f(x) = \begin{cases} \frac{1}{500} e^{-\frac{x}{500}}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

(a) Qual é a média de vida do transistor ?

(b) Qual é a probabilidade de que o tempo de vida seja maior do que a média

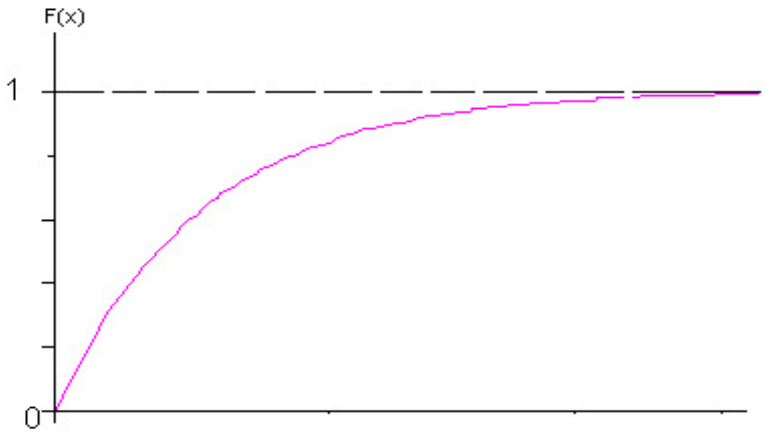


Figura 4.7: Função de distribuição acumulada, de $X \sim Ex(\lambda)$.

(c) Se um transistor em particular há durado mais 300 horas. Qual é a probabilidade de que dure outras 400 horas?

Solução(a)Já que $X \sim E(500)$, de (4.10) temos que: $E(X) = 500$ horas.

(b) Também temos que a função de distribuição acumulada de X é dado por:

$$F(x) = \begin{cases} 0, & x \leq 0 \\ 1 - e^{-\frac{x}{500}}, & x > 0 \end{cases}$$

Daí temos que: (b) $P(X > 500) = 1 - P(X \leq 500) = 1 - (1 - e^{-\frac{500}{500}}) = e^{-1}$.

(c)

$$\begin{aligned} P(X \geq 700 | X > 300) &= \frac{P(X \geq 700; X > 300)}{P(X > 300)} \\ &= \frac{P(X \geq 700)}{P(X > 300} = \frac{1 - [1 - e^{-7/5}]}{1 - [1 - e^{-3/3}]} \\ &= e^{-4/5}. \end{aligned}$$

4.6.3 Distribuição normal

A distribuição normal foi descoberta no século XVIII. Astrônomos e outros cientistas observaram, não sem certa surpresa, que mensurações repetidas de uma mesma quantidade (como distância entre a lua e terra ou a massa de um objeto) tendiam a variar e quando se coletava um grande número dessas mensurações, dispondo-as numa distribuição de freqüências, elas se apresentavam repetidamente com uma forma análoga da figura 4.8.

Definição 4.6.3 (Distribuição normal) *Uma variável aleatória contínua X tem distribuição normal com média μ e variância σ^2, se sua função de densidade é dado por:*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in R \tag{4.12}$$

Usaremos a notação $X \sim N(\mu, \sigma^2)$, para indicar que X tem distribuição normal com parâmetros μ e σ^2. A função de densidade da normal é representada na figura 4.8. Algumas propriedades da distribuição normal podem ser facilmente observadas de seu gráfico

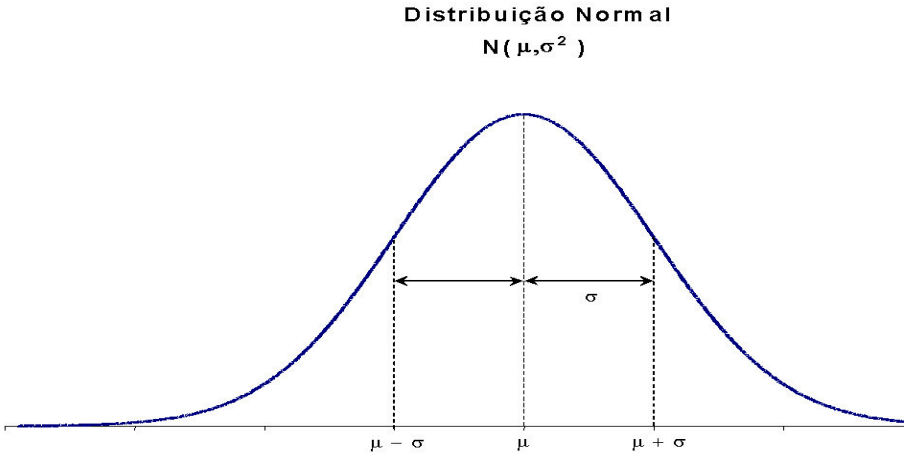


Figura 4.8: Distribuição normal com parâmetros μ e σ^2

1. $E(X) = \mu$ e $Var(X) = \sigma^2$.
2. A curva é simétrica em torno da média μ.
3. É assintótica em relação ao eixo horizontal.
4. A área total sob a curva é igual a *um* portanto, cada metade da curva tem $0,5$ da área total.

A figura 4.9 apresenta o comportamento da função de densidade para valores diferentes da média μ e variâncias iguais. A variância é uma medida de dispersão ou de variabilidade da variável aleatória. A maior variância, maior

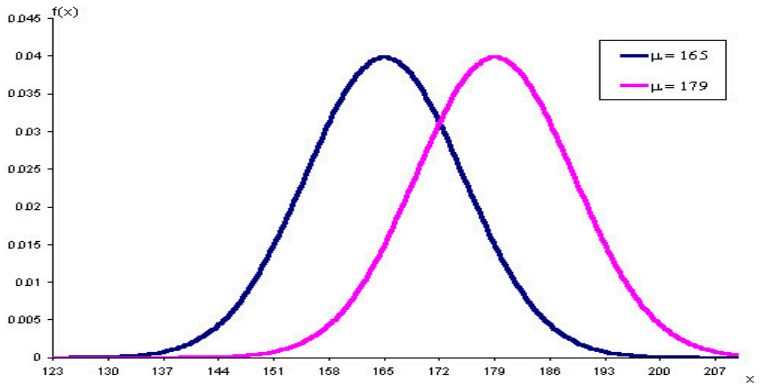


Figura 4.9: Distribuições normais com médias diferentes e variâcias iguais.

variabilidade. Isso pode ser observado graficamente na figura 4.10.

Definição 4.6.4 (Distribuição normal padrão ou reduzida) *Se Z é uma variável aleatória que tem distribuição normal com média $\mu = 0$ e variância $\sigma^2 = 1$, então Z é chamado de variável aleatória normal padrão, sua função de densidade dada por:*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}. \quad z \in R \tag{4.13}$$

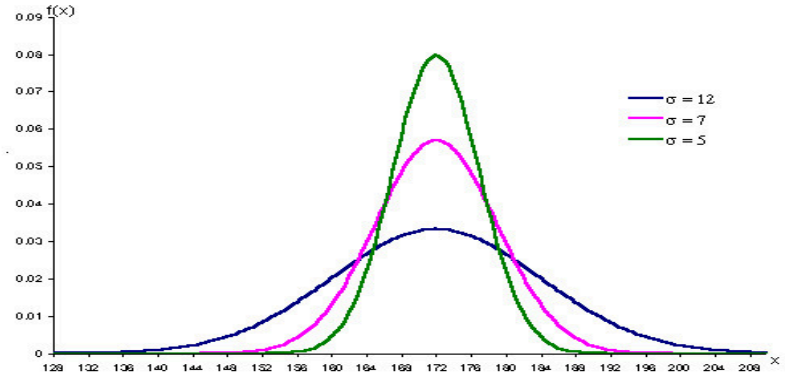


Figura 4.10: Distribuições normais com médias iguais e variâncias diferentes.

Teorema 4.6.1 (Transformação linear de uma variável normal) *Se X é uma variável aleatória normal com média μ e variância σ^2, então a variável, $Y = a+bX$ tem distribuição normal com média, $\mu_Y = a+b\mu$ e variância, $\sigma_Y^2 = b^2\sigma^2$.*

Uma conseqüência imediata do teorema 4.6.1 é a variável

$$Z = \frac{X - \mu}{\sigma} \tag{4.14}$$

que tem distribuição normal padrão, sendo $X \sim N(\mu, \sigma^2)$.

Uso da tabela normal padrão para o cálculo de probabilidade

A tabela de distribuição normal padrão (veja apêndice A) fornece a probabilidade da variável normal padrão Z assumir um valor menor ou igual a z. Isto é,

$$\Phi(z) = P(Z \leq z).$$

Essa probabilidade é representada pela área sombreada na figura 4.11. A função $\Phi(z)$ também recebe o nome de distribuição acumulada da distribuição normal padrão. A tabela A do apêndice A fornece os valores de $\Phi(z)$, para valores $0 \leq z < 3,99$ (os valores para $\Phi(z)$, para $-3,99 \leq z \leq 0$ são obtidos por simetria).

Exemplo 4.6.2 *Seja Z uma variável aleatória normal padrão. Determine:*

(a) $P(Z < 1,80)$;

(b) $P(0,80 \leq Z < 1,40)$;

(c) $P(Z \leq -0,58)$;

(d) $P(-0,58 \leq Z \leq 0,58)$;

(e) o valor de k tal $p(Z \leq k) = 0,95$.

Solução: Para o cálculo de probabilidades sob a distribuição de variáveis aleatórias contínuas (normal padrão) torna-se indiferente o uso de sinais $<$ ou $\leq$ bem como $>$ ou $\geq$, então temos:

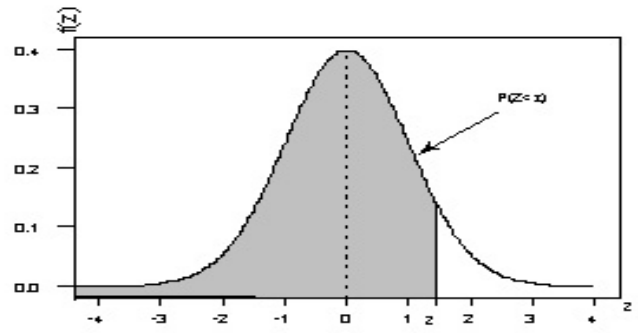


Figura 4.11: Probabilidade $\Phi(z) = P(Z \leq z)$

(a) $P(Z \leq 1,80) = 0,96784$

(b) $P(0,80 \leq Z < 1,40) = P(Z \leq 1,40) - P(Z \leq 0,80) = 0.91924 - 0.78814 = 0,1311$

(c) $P(Z \leq -0,58) = 1 - P(Z \leq 0,58) = 1 - 0.71904 = 0,28096$

(d) $P(-0,58 \leq Z \leq 0,58) = P(Z \leq 0,58) - P(Z \leq -0,58) = P(Z \leq 0,58) - [1 - P(Z \leq 0,58]$
$= 2P(Z \leq 0,58) - 1 = 2 \times 0,71904 - 1 = 0,43808$

(e) $p(Z \leq k) = 0,95$. da tabela normal padrão observa-se que $z = 1,64$

Observação 4.6.1 *Se* $Z \sim N(0,1)$ *então,*

- $P(Z \leq -z) = 1 - P(Z \leq z)$, *para todo* $z > 0$
- $P(-z < Z \leq < z) = 2P(Z \leq z) - 1$

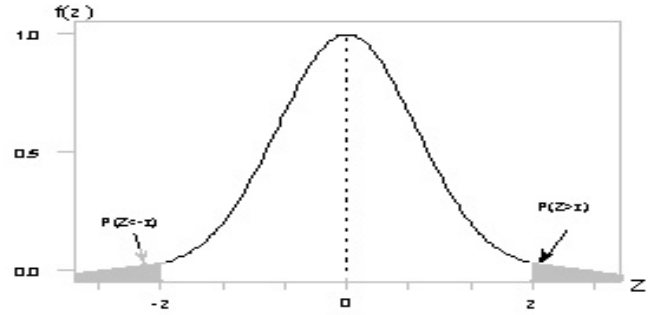


Figura 4.12: Probabilidade $P(Z \leq -z) = P(Z \geq z) = 1 - P(Z \leq z)$.

Exemplo 4.6.3 *Se* $X \sim N(90, 100)$ *determine*

(a) $P(70 \leq X < 90)$.

(b) $P(|X - 90| \leq 30)$.

(c) O valor de a tal que $P(90 - 2a \leq X \leq 90 + 2a) = 0,99$.

Solução: Utilizando a fórmula (4.14), tem-se

(a)

$$\begin{aligned} P(70 \leq X < 90) &= P\left(\frac{70-90}{10} \leq \frac{X-\mu}{\sigma} \leq \frac{90-90}{10}\right) = P(-2 \leq Z \leq 0) \\ &= P(Z \leq 0) - P(Z \leq -2) = P(Z \leq 0) - [1 - P(Z \leq 2)] \\ &= 0,5 - [1 - 0,97725] = 0,47725 \end{aligned}$$

(b)

$$\begin{aligned} P(|X - 90| \leq 30) &= P(-30 \leq X - 90 \leq 30) = P\left(\frac{-30}{10} \leq \frac{X-90}{10} \leq \frac{30}{10}\right) = P(-3 \leq Z \leq 3) \\ &= P(Z \leq 3) - P(Z \leq -3) = 2P(Z \leq 3) - 1 = 2 \times 0,99865 - 1 = 0,9973 \end{aligned}$$

(c)

$$\begin{aligned} P(90 - 2a \leq X \leq 90 + 2a) &= P(-2a \leq X - 90 \leq 2a) = P(\frac{-2a}{10} \leq Z \leq \frac{2a}{10}) \\ &= 2P(Z \leq \frac{a}{5}) - 1 = 0,99 \Rightarrow \quad P(Z \leq \frac{a}{5}) = 0,995 \end{aligned}$$

Portanto $\frac{a}{5} = 2,57 \rightarrow a = 12,85$.

Exemplo 4.6.4 *Os níveis de colesterol sérico em homens de 18 a 24 anos de idade tem distribuição normal com média de 178,1 mg/mL e desvio padrão de 40,7 mg/mL. Os dados se baseiam na "National Health Survey". Escolhido aleatoriamente um homem entre 18 e 24 anos, determine:*

(a) a probabilidade de que seu nível de colesterol esteja entre 200 mg/mL e 250 mg/mL.

(b) o nível de colesterol para ser incluído nos 10% dos homens com menor nível de colesterol.

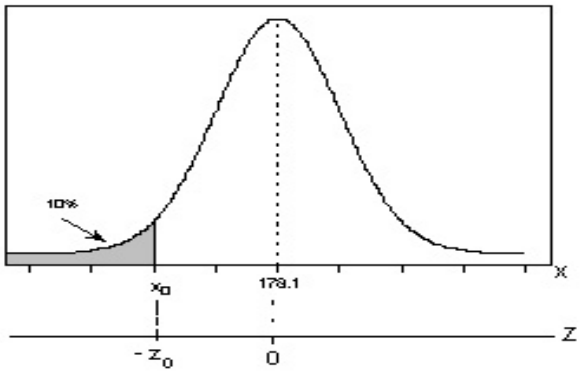
Solução: Seja a variável X : "nível de colesterol sérico em homens com idade entre 18 a 24 anos." $X \sim N(178,1; 40,7^2)$.

(a) $P(200 \leq X \leq 250) = P\left(\frac{200-178,1}{40,7} \leq \frac{X-\mu}{\sigma} \leq \frac{200-178,1}{40,7}\right) = P(0,54 \leq Z \leq 1,77) =$

$= P(Z \leq 1,77) - P(Z \leq 0,54) = 0,96164 - 0,70540 = 0,25624$

(b) Da figura, $P(X < x_0) = 0,10$

Portanto, $0,10 = P(X < x_0) = P\left(Z < \frac{x_0-178,1}{40,7}\right)$, $\Rightarrow P(Z < -z_0) = 0,10$, sendo $-z_0 = \frac{x_0-178,1}{40,7}$. Da observação 4.6.1, tem-se que $P(Z \leq z_0) = 0,90$. Isso implica em $z_0 = 1,28$. Daí $\frac{x_0-178,1}{40,7} = -1,28 \Longrightarrow x_0 = 126,004$



Teorema 4.6.2 (Combinação linear de variáveis aleatórias normais) *Sejam $X_1, \ldots, X_n$, n variáveis aleatórias independentes onde $X_i \sim (\mu_i; \sigma_i^2)$ para $i = 1, \ldots, n$ e sejam $a_1, \ldots, a_n$ constantes reias. Seja a variável aleatória Y uma combinação linear das variáveis aleatórias normais, $X_1, \ldots, X_n$. Isto é,*

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n.$$

Então a variável aleatória Y, tem distribuição normal com média

$$\mu_Y = a_1\mu_1 + a_2\mu_2 \cdots + a_n\mu_n = \sum_{i=1}^{n} a_i\mu_i$$

e variância

$$\sigma_Y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 \cdots + a_n^2\sigma_n^2 = \sum_{i=1}^{n} a_i^2\sigma_i^2.$$

Exemplo 4.6.5 *Uma empresa desenvolve um conjunto restrito de atividades,X_i ($i = 1, 2, 3$). Suponha que o lucro Y (em unidades monetárias) associado às diferentes atividades é dado pela seguinte equação: $Y = 2X_1 + 3X_2 + X_3$. Considerado que as diferentes atividades da empresa são variáveis aleatórias independentes com distribuição normal tais que: $X_1 \sim N(10, 5)$, $X_2 \sim N(15, 20)$ e $X_3 \sim N(12, 10)$, qual é a probabilidade de que empresa tenha um lucro de no máximo, 80 unidades monetárias.?*

Solução: Do teorema 4.6.2, tem-se $Y \sim N(\mu_Y, \sigma_Y^2)$ onde,

$$\mu_Y = 2E(X_1) + 3E(X_2) + E(X_3) = 2 \times 10 + 3 \times 15 + 12 = 77,$$

$$\sigma_Y^2 = 4Var(X_1) + 9Var(X_2) + Var(X_3) = 4 \times 5 + 9 \times 20 + 10 = 210.$$

Logo,

$$P(Y \leq 80) = P\left(Z \leq \frac{80 - 77}{\sqrt{210}}\right) = P(Z \leq 0,21) = 0,58317$$

Exemplo 4.6.6 *Suponha que a carga máxima suportada X_1 por um pilar de concreto armado durante sua vida é uma variável aleatória normal com média 110 kg e desvio padrão de 16 kg, além disso admite-se que sua resistência é outra variável aleatória X_2, com distribuição normal com média 215 kg e desvio padrão de 30 kg. Qual é a probabilidade de ruptura desse pilar?.*

Solução: Considere a variável Aleatória

$$Y = X_2 - X_1,$$

o pilar se romperá quando $X_1 > X_2$ o qual é equivalente a, $Y < 0$. Do teorema 4.6.2, $Y \sim N(\mu_Y, \sigma_Y^2)$, pois X_1 e X_2 são variáveis aleatórias normais independentes. Sendo

$$\mu_Y = E(X_2) - E(X_1) = 215 - 110 = 105$$

e

$$\sigma_Y^2 = Var(X_2) + (-1)^2 Var(X_1) = 30^2 + 16^2 = 1156$$

Daí tem-se que: $P(Y < 0) = P(\frac{Y-\mu_Y}{\sigma_Y^2} < \frac{0-105}{\sqrt{1156}}) = P(Z < -3,09) = 0,001$.

Um resultado imediato do teorema 4.6.2 está dado no seguinte corolário.

Corolário 4.6.1 (Propriedade reprodutiva da distribuição normal) *Se $X_1, \ldots, X_n$ são variáveis aleatórias independentes e identicamente distribuídos com distribuição normal com média μ e variância σ^2, isto é, $X_i \sim N(\mu, \sigma^2)$, então, a variável aleatória:*

$$Y = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i \tag{4.15}$$

tem distribuição normal com média $n\mu$ e variância $n\sigma^2$, ou seja, $Y \sim N(n\mu, n\sigma^2)$ ou

$$Z = \frac{\sum\limits_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

onde $\bar{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$.

Exemplo 4.6.7 *O peso de peixes pescados por uma embarcação tem distribuição normal com média de 4,5 kg e desvio padrão 0,5 kg. Se os peixes são embaladas em caixas que contem 20 peixes, qual é a probabilidade de que o peso total dos peixes contidos numa caixa seja maior de 92 kg?*

Solução: Seja a variável aleatória, X : peso de um peixe. Então $X \sim N(4,5, (0,5)^2)$, e seja Y : o peso total da caixa com 20 peixes, então $Y = X_1 + X_2 + \cdots + X_n$, onde X_i é o peso do i-ésimo peixe na caixa. Assim, $X_i \sim N(4,5, (0,5)^2)$, $i = 1, \ldots, 20$. Pelo corolário 4.6.1, $Y \sim (20 \times 4,5, 20 \times (0,5)^2) = N(90, 5)$

$$P(Y > 92) = P(\frac{Y - 90}{\sqrt{5}} > \frac{92 - 90}{\sqrt{5}}) = P(Z > 0,89) = 1 - p(Z \leq 0,89) = 1 - 0,81327 = 0,18673$$

4.7 Distribuições Amostrais

Definição 4.7.1 *As variáveis aleatórias $X_1, X_2, \ldots, X_n$ constituem uma amostra aleatória de tamanho n de uma população $X \sim f(x, \theta)$, se: (a) cada X_i é uma variável aleatória independente e (b) cada X_i, tem a mesma distribuição de probabilidade $f(x, \theta)$.*

A definição de amostra aleatória é satisfeita nos seguintes casos:

1. Quando a amostra provem de uma população infinita[2] e quando a amostra é sorteada ao acaso com reposição de uma população finita.
2. Quando as amostras se sorteia sem reposição de uma população finita, evidentemente não satisfaz a definição da amostra aleatória, pois as variáveis aleatórias $X_1, \ldots, X_n$ não são independentes. Porem, se o tamanho da amostra é muito pequena em comparação com o tamanho da população, a definição é satisfeita aproximadamente.

[2]Quando o tamanho da população não é mencionado neste texto será considerado como uma população infinita

Exemplo 4.7.1 *De uma população normal com média* 10 *e variância* 12 *selecionou-se uma amostra aleatória,* $X_1, X_2, \ldots, X_{10}$. *Calcular*

$$P(X_1 - X_5 + X_8 \geq 13).$$

Solução: Se X, é uma variável aleatória da população normal, $X \sim N(10, 12)$. Então, por ser $X_1, \ldots, X_{10}$ uma amostra aleatória, satisfaz: (a) X_i, $i = 1, \ldots, 10$ são variáveis aleatórias independentes e (b) $X_i \sim N(10, 12)$. Se, $Y = X_1 - X_5 + X_8$, então $Y \sim N(\mu_Y, \sigma_Y)$ por ser variáveis aleatórias normais independentes (pela teorema 4.6.2) onde

$$\mu_Y = E(X_1 - X_5 + X_8) = E(X_1) - E(X_5) + E(X_8) = 10 - 10 + 10 = 10$$
$$\sigma_Y^2 = Var(X_1 - X_5 + X_8) = Var(X_1) + Var(X_5) + Var(X_8) = 12 + 12 + 12 = 36.$$

Logo

$$\begin{aligned} P(X_1 - X_5 + X_8 \geq 13) &= P(Y \geq 13) = P(Z \geq \frac{13 - 10}{6}) \\ &= P(Z \geq 0,5) = 1 - P(Z \leq 0,5) = 1 - 0,69146 = 0,30854. \end{aligned}$$

Definição 4.7.2 (Estatística) *Um estatística é uma variável aleatória que depende somente da amostra observada*

Exemplo 4.7.2 *Sejam* $X_1, \ldots, X_n$ *uma amostra aleatória de uma população* X, *então* $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *e* $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ *são estatísticas.*

Definição 4.7.3 *A distribuição de probabilidade de uma estatística é chamada de* **distribuição amostral**

4.7.1 Distribuição da média amostral

Teorema 4.7.1 *Se de uma população com média* μ_X *e variância* σ_X^2 *se extraem amostras aleatórias de tamanho* n *e para cada amostra determinam-se a média*

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

então a média e variância da variável $\bar{X}$ *são dados por:*

a) Se a amostragem é com reposição de uma população finita (ou amostragem com ou sem reposição em uma população infinita).

$$\mu_{\bar{X}} = \mu_X \quad e \quad \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

b) Se a amostragem é sem reposição de uma população finita com N *elementos.*

$$\mu_{\bar{X}} = \mu_X \ e \ \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\left[\frac{N-n}{N-1}\right]$$

Observação 4.7.1 *Se a fração de amostragem* $f = \frac{n}{N}$ *é pequena* $(f < 0,1)$ *e o tamanho da população* (N) *é grande, a variância da média amostral em (b) é aproximado com a expressão do caso* (a), *isto é,*

$$\sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}$$

Exemplo 4.7.3 *Um auditor de uma empresa deseja determinar a quantidade de produtos existentes no estoque da empresa. Para isso determinou para cada produto do inventario, a diferença* (X) *entre o número artigos registrados e o número de artigos realmente existente. Se o inventario consta de* 5 *artigos e os valores de* X *em milhares de dólares são:*

Produto	A	B	C	D	E
X	0	-1	0	1	2

obter a distribuição amostral de $\bar{X}$ *para amostragem com ou sem reposição, quando* $n = 2$

Solução: A função de probabilidade de X é dado por:

$$f(x) = \begin{cases} 1/5, & \text{se } x = -1, 1, 2 \\ 2/5, & \text{se } x = 0 \\ 0, & \text{caso contrário} \end{cases}$$

Portanto,

$$E(X) = \frac{2}{5} = 0,4 \;\; e \;\; E(X^2) = \frac{6}{5} = 1,2$$

com o qual: $\mu_X = E(X) = 0,4$ e $\sigma_X^2 = E(X - \mu_X)^2 = E(X^2) - E(X)^2 = 1,2 - (0,4)^2 = \frac{26}{25} = 1,04$

Considerando o teorema 4.7.1 tem-se que a média e variância da distribuição da média amostral com $N = 5, n = 2$ é:

a) Para uma amostragem com reposição.

$$\mu_{\bar{X}} = \mu_X = 0,4 \text{ e } \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n} = \frac{1,04}{2} = 0,52$$

b) Para uma amostragem sem reposição.

$$\mu_{\bar{X}} = \mu_X = 0,4 \text{ e } \sigma_{\bar{X}}^2 = \frac{\sigma_X^2}{n}\left[\frac{N-n}{N-1}\right] = \frac{1,04}{2}\left[\frac{5-2}{5-1}\right] = 0,39$$

Para determinar a distribuição da média amostral deve-se determinar todas as amostras possíveis, suas respectivas médias e suas probabilidades de ocorrência considerando 2 casos:

(a) Para uma amostragem sem reposição

Quando a seleção dos elementos da amostra se efetua com probabilidades iguais, o número de amostras possíveis é igual á:

$$\text{Número de amostras possíveis} = \binom{N}{n} = \binom{5}{2} == 10$$

onde N é o tamanho da população e n é o tamanho da amostra. As amostras possíveis se apresentam na tabela seguinte:

Amostra possível	Valores observados de X	Média amostral $\bar{x}$	probabilidade
A, B	$(0; -1)$	$-0,5$	$0,1 = \frac{1}{10}$
A, C	(0;0)	0,0	0,1
A, D	(0;1)	0,5	0,1
A, E	(0;2)	1,0	0,1
B, C	(-1;0)	-0,5	0,1
B, D	(-1;1)	0,0	0,1
B, E	(-1;2)	0,5	0,1
C, D	(0;1)	0,5	0,1
C, E	(0;2)	1,0	0,1
D, E	(1;2)	1,5	0,1

Sendo a seleção com probabilidades iguais, todas as amostras possíveis tem a mesma probabilidade de ocorrência , e portanto a probabilidade de um valor da média amostral será igual a probabilidade de seleção de cada amostra ($\frac{1}{10}$) multiplicada por o número de amostras que geram dito valor.

Logo,a função de probabilidade da média amostrais $\bar{X}$, é:

$$f(\bar{x}) = \begin{cases} 0,1, & \text{se } \bar{x} = 1,5 \\ 0,2, & \text{se } \bar{x} = -0,5; 0,0; 1,0 \\ 0,3, & \text{se } \bar{x} = 0,5 \\ 0,0, & \text{caso contrário} \end{cases}$$

Pode-se mostrar que

$$\begin{aligned} E(\bar{X}) &= \sum \bar{x_i} f(\bar{x_i}) = (1,5)(0,1) + ... + (0,5)(0,3) = 0,4 \\ E(\bar{X}^2) &= \sum {\bar{X_i}}^2 f(\bar{x_i}) = (1,5)^2(0,1) + ... + (0,5)^2(0,3) = 0,55 \end{aligned}$$

$$\mu_{\bar{x}} = E[\bar{X}] = 0,4 \quad \text{e} \quad \sigma^2_{\bar{x}} = E[\bar{X}^2] - \mu^2_{\bar{x}} = 0,55 - (0,4)^2 = 0,39$$

(b) Para uma amostragem com reposição

Quando a seleção dos elementos da amostra se efetua com probabilidades iguais, o número de amostras possíveis é igual a $N^n = 5^2 = 25$, onde N é o tamanho da população e n é o tamanho da amostra. As amostras possíveis se apresentam na seguinte tabela:

Amostra possível	Valores observados de X	Média amostral	probabilidade
A, A	0; 0	0, 0	$0,04 = \frac{1}{25}$
A, B	0;-1	-0,5	0,04
A, C	0;0	0,0	0,04
A, D	0;1	0,5	0,04
A, E	0;2	1,0	0,04
B, A	-1;0	-0,5	0,04
B, B	-1;-1	-1,0	0,04
B, C	-1;0	-0,5	0,04
B, D	-1;1	0,0	0,04
B, E	-1;2	0,5	0,04
C, A	0;0	0,0	0,04
C, B	0;-1	-0,5	0,04
C, C	0;0	0,0	0,04
C, D	0;1	0,5	0,04
C, E	0;2	1,0	0,04
D, A	1;0	0,5	0,04
D, B	1;-1	0,0	0,04
D, C	1;0	0,5	0,04
D, D	1;1	1,0	0,04
D, E	1;2	1,5	0,04
E, A	2;0	1,0	0,04
E, B	2;-1	0,5	0,04
E, C	2;0	1,0	0,04
E, D	2;1	1,5	0,04
E, E	2;2	2,0	0,04

Como no caso anterior, a probabilidade de um valor de $\bar{X}$ é igual a probabilidade de seleção de cada amostra ($\frac{1}{25}$) multiplicada por o número de amostras que geram dito valor. Logo, a função de probabilidade das médias amostrais é:

$$f(\bar{X}) = \begin{cases} \frac{1}{25}, & \text{se } \bar{x} = -1,0; 2,0 \\ \frac{4}{25}, & \text{se } \bar{x} = -0,5 \\ \frac{6}{25}, & \text{se } \bar{X} = 0,0; 0,5 \\ \frac{5}{25}, & \text{se } \bar{X} = 1,0 \\ \frac{2}{25}, & \text{se } \bar{x} = 1,5 \\ 0, & \text{casso contrario} \end{cases}$$

Daí tem-se que:

$$\begin{aligned} E(\bar{X}) &= \sum \bar{x_i} f(\bar{x_i}) = (-1,0)(\frac{1}{25}) + ... + (1,5)(\frac{2}{25}) = 0,4 \\ E(\bar{X}^2) &= \sum \bar{x_i}^2 f(\bar{x_i}) = (-1,0)^2(\frac{1}{25}) + ... + (1,5)^2(\frac{2}{25}) = 0,68 \end{aligned}$$

$$\mu_{\bar{x}} = E[\bar{X}] = 0,4 \quad \text{e} \quad \sigma^2_{\bar{x}} = E[\bar{X}^2] - \mu^2_{\bar{x}} = 0,68 - (0,4)^2 = 0,52$$

No exemplo anterior, conseguimos enumerar as possíveis amostras e assim obter sua função de probabilidade da média amostral. Nem sempre isso será possível, por exemplo se X tem distribuição de Poisson com parâmetro $\mu = 5$, uma amostra aleatória de tamanho 2 desta população, X_1 e X_2 continuaram sendo independentes e identicamente distribuídos com função de probabilidade, $P_o(5)$. Mas, é complicado enumerar todas as possíveis amostras de tamanho 2, portanto é difícil de determinar a distribuição de probabilidade da média amostral.

4.7.2 Forma da distribuição da média amostral quando a população não é normal

Seja X uma variável aleatória que tem uma distribuição normal com média μ_X e variância σ^2_X. Se desta distribuição seleciona-se amostras aleatórias de tamanho n, a média amostral,

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

é uma combinação linear de variáveis X_i, todas elas com distribuição $N(\mu_X, \sigma^2_X)$ e independentes entre si (o fato da distribuição de X ser normal presume, em rigor que a população é infinita e que, portanto, não há diferença entre escolher uma amostra com e sem reposição). Foi visto na seção anterior, uma combinação linear de variáveis normais independentes é, também é normal, portanto, a média amostral segue uma distribuição normal com média μ_X e variância, σ^2_X. Isto é,

$$\bar{X} \sim N(\mu_X, \sigma^2_X/n).$$

Embora este resultado seja de extrema importância, eles são relativamente limitado, já que, somente permite especificar a distribuição da média amostral no caso de uma população normal. Na prática, muitas vezes não temos informação a respeito da distribuição das variáveis que constituem a amostra, o que nos impede utilizar o resultado apresentado. Felizmente, satisfeitas certas condições pode ser mostrado que para uma amostra suficientemente grande, a distribuição de probabilidade da média amostral pode ser aproximada por uma distribuição normal, com média e variância iguais àquelas calculadas anteriormente. Este fato é um dos teoremas mais importantes da estatística e probabilidade e é denominado o **teorema central do limite**.

A continuação enuncia-se o teorema central do limite considerando que a população é infinita.

Teorema 4.7.2 (Teorema Central do Limite) *Seja $X_1, \ldots, X_n$ uma amostra aleatória de tamanho n retirada de uma população com média μ_X é variância σ^2_X, finita. Então a média amostral, $\bar{X}$, tem distribuição aproximadamente normal com média μ_X e variância σ^2_X/n, para n suficientemente grande ($n \to \infty$). Isto é,*

$$Z = \frac{\bar{X} - \mu_X}{\sigma_X/n} \overset{n\to\infty}{\longrightarrow} N(0,1).$$

Neste texto consideraremos que o tamanho de amostra é suficientemente grande quando $n \geq 30$.

Exemplo 4.7.4 *Suponha que na produção em série de um artigo, o peso é uma variável aleatória com uma média de* 950 *g e uma variância de* 1600 g^2. *Seleciona-se aleatóriamente e com reposição* 36 *artigos, calcular a probabilidade que a média amostral seja maior de* 965 *g.*

Solução: Seja X o peso do artigo (em gramas), como, $\mu_X = 950$, $\sigma_X^2 = 1600$ e $n = 36$. Pelo teorema 4.7.2, tem-se que $\bar{X}$ aproximadamente normal com média, $\mu_{\bar{X}} = \mu_X = 950$ e variância $\sigma_{\bar{X}}^2 = 1600/36$. Portanto,

$$\begin{aligned} P(\bar{X} > 965) &= P\left(Z > \frac{965 - 950}{\frac{40}{\sqrt{36}}}\right) = P(Z > 2,25) \\ &= 1 - P(Z \leq 2,25) = 1 - 0,9878 = 0,0122 \end{aligned}$$

4.7.3 Distribuição da diferença de duas médias amostrais

Teorema 4.7.3 *$X_1, \ldots, X_n$ é uma amostra aleatória de tamanho n de uma população com característica X que tem distribuição normal com média μ_1 e variância σ_1^2 e que $Y_1, \ldots, Y_m$ é outra amostra aleatória de tamanho m, de uma população com a característica Y que tem distribuição normal com média μ_2 e variância σ_2^2. Se X e Y são independentes, então a diferença amostral $\bar{X} - \bar{Y}$ tem distribuição normal com média $\mu_1 - \mu_2$ e variância $\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$. Isto é,*

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \sim N(0,1), \tag{4.16}$$

Se as populações onde foram retiradas as amostras não tiveram distribuição normal, pelo teorema 4.7.2, segue válido o resultado se os tamanhos amostrais n e m são suficientemente grandes, isto é $n \geq 30$ e $m \geq 30$.

Exemplo 4.7.5 *Suponha que numa central de correios (A) o peso (em gramas) das cartas tem distribuição normal com média 350 g e desvio padrão de 56,27 g.*

(a) Qual deve ser o tamanho da amostra para que a probabilidade de que o peso médio das carta defira do peso médio verdadeiro em menos de 15 g, seja igual a 0,9426

(b) Em outra central de correio (B) encontrou-se que os peso (em gramas) das cartas tem distribuição normal com média de 320 g e desvio padrão de 50 g. Retiram-se ao acaso 20 cartas de cada central de correios, qual é probabilidade de que o peso médio das cartas retiradas do correio A seja maior ao peso médio das cartas do correio B em pelo menos 10 g?

Solução: Seja, X : peso das cartas do correio A, então $X \sim N(350, (56,27)^2)$

(a) $\bar{X} \sim N(350, (56,27)^2/n)$, do enunciado do problema temos que determinar $n =?$, tal que, $P(|\bar{X} - \mu| < 15) = 0,9426$

$$\begin{aligned} P(|\bar{X} - \mu| < 15) &= P(\frac{|\bar{X} - \mu|}{\sigma/\sqrt{n}} < 15/56,27/\sqrt{n}) \\ &= P(|Z| \leq 0,2666\sqrt{n}) = 0,9426, \end{aligned}$$

que é equivalente a:

$$\begin{aligned} P(|Z| < z_0) &= P(-z_0 \leq Z \leq z_0) = P(Z \leq z_0) - P(Z \leq -z_0) \\ &= 2P(Z \leq z_0) - 1 = 0,9426 \end{aligned}$$

portanto $P(Z \leq z_0) = 0,9713$, da tabela normal padrão, temos que, $z_0 = 1,90$. Portanto, $0,2666\sqrt{n} = 1,90$, $n = 51$.

(b) Y o peso de cartas do correio B, então, $Y \sim N(320, 50^2)$, que implica em

$$\begin{aligned} n = 20 \quad & \bar{X} \sim N(350, (56,27)^2/20) \\ m = 20 \quad & \bar{Y} \sim N(320, 50^2/20), \\ & \bar{X} - \bar{Y} \sim N(350 - 320, \frac{56,27^2}{20} + \frac{50^2}{20}) = N(30, 283, 31) \end{aligned}$$

$P(\bar{X} - \bar{Y} \geq 10) = P(Z \geq \frac{10-30}{\sqrt{283,31}}) = P(Z \geq -1,19) = P(Z \leq 1,19) = 0,88297$

4.7.4 Distribuição amostral de uma proporção amostral

Considere uma população dicotômica, constituída apenas por elementos de dois tipos, isto é, cada elemento pode ser classificado com *sucesso* ou *fracasso*. Suponha que a probabilidade de sucesso seja p e de fracasso seja $q = 1-p$. Se dessa população retira-se uma amostra aleatória de n observações $X_1, \ldots, X_n$. Seja a variável aleatória Y número de sucessos na amostra. Então,

1. $Y = \sum_{i=1}^{n} X_i$ tem distribuição Binomial com parâmetros n e p.

2. A proporção amostral de sucessos é: $\hat{p} = \frac{Y}{n} = \sum_{i=1}^{n} X_i/n = \bar{X}$. De (1) a distribuição de probabilidade de $\hat{p}$ é:

$$P(\hat{p} = \frac{y}{n}) = \binom{n}{y} p^y (1-p)^{n-y}.$$

   E para n suficientemente grande (teorema 4.7.2), tem distribuição aproximadamente normal com média p e variância $\frac{pq}{n}$. Isto é,

$$\hat{p} \sim N(p, \frac{pq}{n}).$$

Exemplo 4.7.6 *Uma empresa tem um número grande de funcionários. A probabilidade de que um empregado selecionado ao acaso, participe de um programa de treinamento é 0,40.*

(a) Se 10 funcionários são escolhidos ao acaso, qual é a probabilidade que proporção de participantes seja

(a1) exatamente 60%?

(a2) pelo menos 80%?

(b) suponha que 100 funcionários escolhidos ao acaso, participaram do treinamento qual é a probabilidade de que proporção de participantes do programa seja maior que 50%?

Solução: Seja Y : número de funcionários que participaram do programa de treinamento entre os 10 selecionados. Considere *sucesso:* "funcionário que participa do programa."Logo, $P(sucesso) = 0,40$. Portanto, $Y \sim B(10, 0,4)$.

(a1) $P(\hat{p} = 0,60) = P(\frac{Y}{10} = \frac{6}{10}) = P(Y = 6) = \binom{10}{6}(0,4)^6(0,6)^4 = 0,1115$

(a2) $P(\hat{p} \geq 0,8) = P(Y \geq 8) = 0,0123$.

(b) Y : número de funcionários que participaram do programa de treinamento entre os 100 selecionados. Então $Y \sim B(100, 0,4)$. Logo, $\hat{p} \sim N(0,4, 0,24/100)$

$$\begin{aligned} P(\hat{p} > 0,50) &= P\left(\frac{\hat{p} - p}{\sqrt{pq/100}} > \frac{0,5 - 0,40}{\sqrt{0,24}/10}\right) \\ &= P(Z > 2,04) = 1 - P(Z \leq 2,04) = 1 - 0,97932 = 0,02068. \end{aligned}$$

Observação 4.7.2 *Os resultados de acima são válidas também nos seguintes casos:*

1. *Para uma população infinita, qualquer que seja o tipo de amostragem.*

2. *Para população finita, com amostragem com reposição.*

Se a amostragem é sem reposição, em uma população finita de N elementos, a distribuição exata de probabilidade $\hat{p}$ é uma distribuição Hipergeométrica. Isto é,

$$P(\hat{p} = \frac{y}{n}) = \frac{\binom{N}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \tag{4.17}$$

A variância de $\hat{p}$ é ajustado através do fator de correção de população finita, isto é,

$$Var(\hat{p}) = \frac{pq}{n}\left(\frac{N-n}{N-1}\right).$$

Se, n é suficientemente grande, pelo teorema central do limite, a variável aleatória,

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

tem distribuição aproximadamente normal padrão.

Exemplo 4.7.7 *Informações anteriores mostram que 10% do lote de peças para uma máquina são defeituosos. Suponha que um lote de 5000 peças foi adquirido. Seleciona-se uma amostra de 400 peças, ao acaso e sem reposição. Que proporção da amostra terá*

(a) entre 9% e 10% de peças defeituosas ?

(b) menos de 8% de peças defeituosas

Solução: Seja a variável aleatória Y : número de peças defeituosas na amostra e $P(sucesso) = p = 0,10$. A população é finita pois $N = 5000$ e $\hat{p} = \frac{Y}{n}$ é a proporção de defeituosos na amostra. Já que, $n = 400$, grande, a variável aleatória, $\hat{p}$ tem distribuição aproximadamente normal com média $\mu_{\hat{p}} = 0,10$ e desvio padrão, $\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}\left(\frac{N-n}{N-1}\right)} = \sqrt{\frac{(0,10)(0,90)}{400}\left(\frac{5000-400}{5000-1}\right)} = 0,0144$.

(a)

$$\begin{aligned} P(0,09 < \hat{p} < 0,10) &= P\left(\frac{0,09-0,10}{0,0144} < \frac{\hat{p}-p}{\sqrt{\frac{pq}{n}(\frac{N-n}{N-1})}} < \frac{0,10-0,10}{0,0144}\right) \\ &= P(-0,69 < Z < 0) = P(Z \leq 0) - P(Z \leq -0,69) \\ &= 0,5 - 0,2451 = 0,2549. \end{aligned}$$

(b) $P(\hat{p} < 0,08) = P\left(\frac{\hat{p}-p}{\sqrt{\frac{pq}{n}(\frac{N-n}{N-1})}} < \frac{0,08-0,10}{0,0144}\right) = P(Z < -1,39) = 0,0823.$

4.8 Distribuições Utilizadas na Inferência Estatística

4.8.1 Distribuição Qui-quadrado

Definição 4.8.1 *Sejam $Z_1, \ldots, Z_k$ k variáveis aleatórias distribuídas normalmente e independentes com média $\mu = 0$ e variância $\sigma^2 = 1$. A variável aleatória,*

$$W = Z_1^2 + Z_2^2 + \cdots + Z_k^2 \tag{4.18}$$

tem distribuição Qui-quadrado com k graus de liberdade e sua função de densidade é dada por:

$$f(w) = \frac{1}{\Gamma(k/2)2^{k/2}} w^{\frac{k}{2}-1} e^{-\frac{w}{2}}, \quad w > 0 \tag{4.19}$$

onde $\Gamma(a)$ é uma função matemática definida

$$\Gamma(a) = \int_0^\infty x^{a-1}e^{-x}dx,$$

chamada de função gama essa função satisfaz as seguintes propriedades:

$$\begin{aligned} \Gamma(a) &= (a-1)\Gamma(a-1) \\ \Gamma(1/2) &= \sqrt{\pi} \\ \Gamma(a) &= (a-1)!, \text{ para } a \text{ enteiro} \end{aligned}$$

O gráfico da distribuição Qui-quadrado para $k = 2, 4, 6, 10$ graus de liberdade é mostrado na figura 4.13.

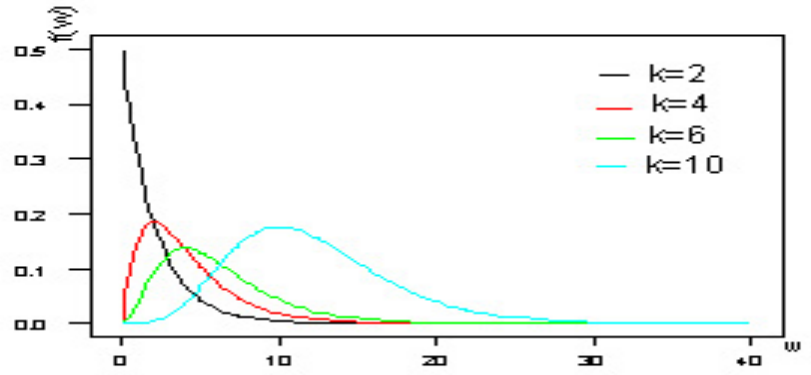


Figura 4.13: Funções de densidade de probabilidade de varias distribuições $\chi^2_{(k)}$

A notação $W \sim \chi^2_{(k)}$ é usada para indicar que a variável W tem distribuição Qui-quadrado com k graus de liberdade.

Propriedades

Se $W \sim \chi^2_{(k)}$

(a) $E(W) = k$ e $Var(W) = 2k$.

(b) A distribuição é assimétrica direita.

(c) A medida que aumentam-se os graus de liberdade, torna-se simétrica.

Uso da tabela Qui-quadrado

Na tabela B do apêndice A, tem-se os pontos críticos da distribuição $W \sim \chi^2_{(k)}$, denotado por $\chi^2_{\alpha,k}$ tal que a probabilidade

$$P(W > \chi^2_{\alpha,k}) = \int_{\chi^2_{\alpha,k}}^\infty f(w)dw$$

Essa probabilidade é representada pela área sombreada da figura 4.14. Para ilustrar o uso da tabela B, observe que as áreas α estão na primeira linha e na primeira coluna estão os graus de liberdade k. Portanto, o valor de χ^2 com 10 graus de liberdade e com área (probabilidade) 0,05 à direita é $\chi^2_{0,05,10} = 18,31$. Isto é,

$$P(W > \chi^2_{0,05,10}) = P(W > 18,31) = 0,05.$$

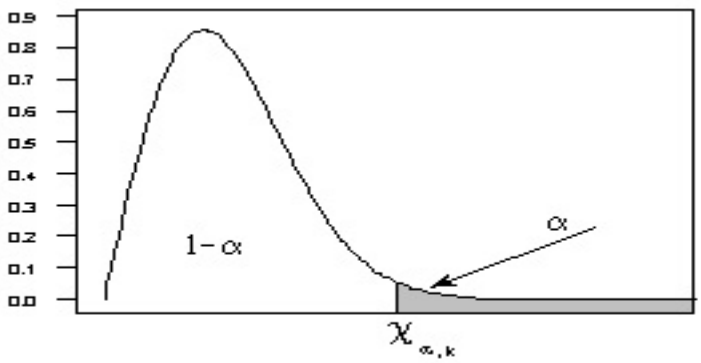


Figura 4.14: Pontos críticos $\chi^2_{\alpha,k}$ das distribuições $\chi^2_{(k)}$

Exemplo 4.8.1 *Se X é uma variável aleatória $\chi^2_{(17)}$, obtenha: (a) $P(X \geq 8,67)$; (b)$P(X \leq 867)$; (c) $P(6,41 < X < 27,59)$; (d) o valor de a tal que $P(X < a) = 0,025$.*

Solução

(a) $P(X \geq 8,67) = P(X \geq \chi^2_{0,95,17}) = 0,95$.

(b) $P(X \leq 867) = 1 - P(X \geq 8,67) = 1 - 0,95 = 0,05$.

(c) $P(6,14 < X < 27,59) = P(X \geq 6,41) - P(X \geq 27,59) = 0,99 - 0,05 = 0,94$

(d) $P(X < a) = 0,025$; implica que $P(X > a) = 0,975$. Logo, $a = \chi^2_{0,725,17} = 7,56$.

Teorema 4.8.1 (Propriedade reprodutiva) *Se $W_1, W_2, \ldots, W_n$ são variáveis aleatórias independentes distribuídas cada uma com distribuição Qui-quadrado com $k_1, k_2, \ldots, k_n$ graus de liberdade respectivamente, então, a variável*

$$W = W_1 + W_2 + \ldots, W_n$$

tem distribuição Qui-quadrado com $k = \sum_{i=1}^{n} k_i$ graus de liberdade

Exemplo 4.8.2 *Se W_1, W_2 e W_3 são variáveis aleatórias independentes com distribuição Qui-quadrado respectivamente com 2, 3 e 4 graus de liberdade respectivamente, então $W = W_1 + W_2 + W_3 \sim \chi^2_{(9)}$.*

Teorema 4.8.2 *Seja $X_1, \ldots, X_n$ uma amostra aleatória de uma população normal com média μ e variância, σ^2. Então a variável aleatória*

$$W = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \tag{4.20}$$

segue uma distribuição Qui-quadrado com $n-1$ graus de liberdade.

Prova: A variável $Z_i = \left(\frac{X_i - \mu}{\sigma}\right) \sim N(0,1)$, $i = 1, \ldots, n$ independentes entre si. Pela definição da distribuição Qui-quadrado, tem-se $\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_{(n)}$ e $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_{(1)}$, mas

$$\underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2}_{\chi^2_{(n)}} = \underbrace{\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2}_{\chi^2_{(n-1)}} + \underbrace{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)}_{\chi^2_{(1)}}$$

Pelo teorema 4.8.1, W tem distribuição Qui-quadrado com $n-1$ graus de liberdade. Uma forma equivalente da variável W, em (4.20), é:

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Exemplo 4.8.3 *Suponha que o tempo de atendimento por cliente em uma loja tem distribuição normal com variância de 0,81. Se uma amostra aleatória de 21 clientes foi retirada, obtenha: (a) $P(S^2 < 1,272)$; (b) $P(0,50625 < S^2 < 1,272)$;*

Solução: Seja X : o tempo de atendimento por cliente. Se $X \sim N(\mu, 0,81)$.

Então $W = \frac{(n-1)S^2}{\sigma^2} = \frac{(20-1)S^2}{0,81} \sim \chi^2_{(20)}$.

(a)

$$\begin{aligned} P(S^2 < 1,272) &= P\left(\frac{(n-1)S^2}{\sigma^2} < \frac{(21-1)(1,272)}{0,81}\right) \\ &= P(W < 31,41) = 1 - P(W \geq 31,41) \\ &= 1 - 0,05 = 0,95 \end{aligned}$$

(b)

$$\begin{aligned} P(0,50625 < S^2 < 1,272) &= P\left(\frac{(21-1)(0,50625)}{0,81} < \frac{(n-1)S^2}{\sigma^2} < \frac{(21-1)(1,272)}{0,81}\right) \\ &= P(12,5 < W < 31,41) = P(W > 12,5) - P(W > 31,41), \end{aligned}$$

Nesse caso, na tabela $\chi^2_{(20)}$, não há a probabilidade associada ao valor 12,5. Porém, essa probabilidade pode ser aproximada mediante um processo de interpolação linear da seguinte forma:

$P(W > \chi^2_{\alpha,20})$	$\rightarrow$	0,50	α	0,90	$(0,90-0,5)$	$\rightarrow$	$(12,44-19,34)$
$\chi^2_{\alpha,20}$	$\rightarrow$	19,34	12,5	12,44	$(\alpha-0,5)$	$\rightarrow$	$(12,5-19,34)$

onde

$$\alpha = 0,5 + \frac{(12,5-19,34)(0,90-0,5)}{12,44-19,34} = 0,896522.$$

Portanto, $P(0,50625 < S^2 < 1,272) = P(W > 12,5) - P(W > 31,41) = 0,896522 - 0,05 = 0,846522$

4.8.2 A distribuição t-Student

Definição 4.8.2 *Seja Z e W duas variáveis independentes com distribuição normal padrão e Qui-quadrado com k graus de liberdade, respectivamente. A variável aleatória,*

$$T = \frac{Z}{\sqrt{\frac{W}{k}}}$$

tem distribuição t-Student com k graus de liberdade. A função de densidade de probabilidade é dado por:

$$f(t) = \frac{\Gamma(\frac{k+1}{2})}{(k\pi)^{1/2}\Gamma(\frac{k}{2})}\left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}$$

A notação $T \sim t(k)$ é usada para indicar que a variável T tem distribuição t-Student com k graus de liberdade.

Na figura 4.15 é apresentado o gráfico da função de densidade de probabilidade, para $k = 5, 10, 20$ graus de liberdade.

Propriedades Se $T \sim t_{(k)}$.

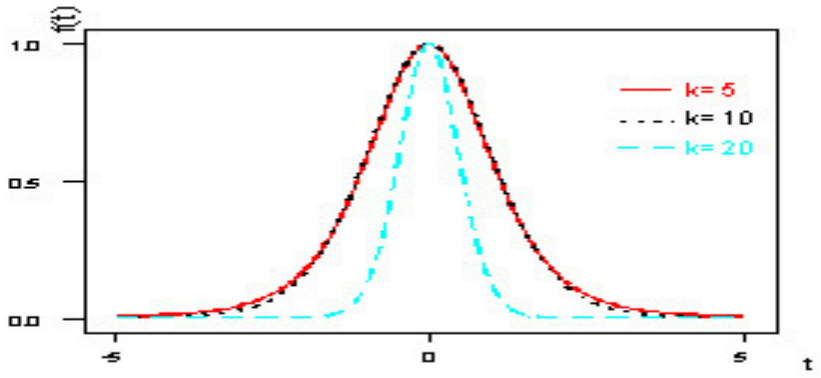


Figura 4.15: Função de densidade de probabilidade da distribuição t-Student.

(a)

$$\begin{aligned} E(T) &= 0 \\ Var(T) &= \frac{k}{k-2}, \quad k > 2 \end{aligned}$$

(b) A distribuição é simétrica em torno de sua média.

(c) Se $k \to \infty$, $T \sim N(0,1)$.

Uso da tabela t-Student

A tabela C, do apêndice A proporciona os pontos críticos da distribuição t-Student. Seja $t_{\alpha,k}$ o valor da variável aleatória T com k graus de liberdade para o qual tem-se uma área (probabilidade) α. Portanto, $t_{\alpha,k}$ é um ponto crítico na cauda superior da distribuição t-Student com k graus de liberdade. Este ponto crítico aparece na figura 4.16. Na tabela C do apêndice, os valores de α encontram-se na primeira linha da tabela, enquanto os graus de liberdade aparecem na primeira coluna da parte esquerda. Para ilustrar o uso da tabela, observe que o valor de t-Student com 10 graus de liberdade que tem área de 0,05 à direita é $t_{0,05,10}$. Isto é,

$$P(T > t_{0,05,10}) = P(T > 1,812) = 0,05$$

Como, a distribuição t-Student é simétrica com respeito a zero (média), tem-se que $t_{1-\alpha,k} = -t_{\alpha,k}$. Isto é, o valor da variável T que corresponde a uma área igual $(1-\alpha)$ à direita (e, portanto, uma área de α à esquerda) é igual ao negativo do valor de T, que tem área α na cauda direita da distribuição. Em conseqüência, $t_{0,95,10} = -t_{0,05,10} = -1,812$.

Exemplo 4.8.4 *Seja T uma variável aleatória com distribuição t-Student com 12 graus de liberdade (gl). Determine:*

(a) $P(T > -1,356)$

(b) $P(0,695 < T < 2,179)$

(c) $P(-2,179 < T < 2)$

(d) $P(-1,782 < T < 1,782)$

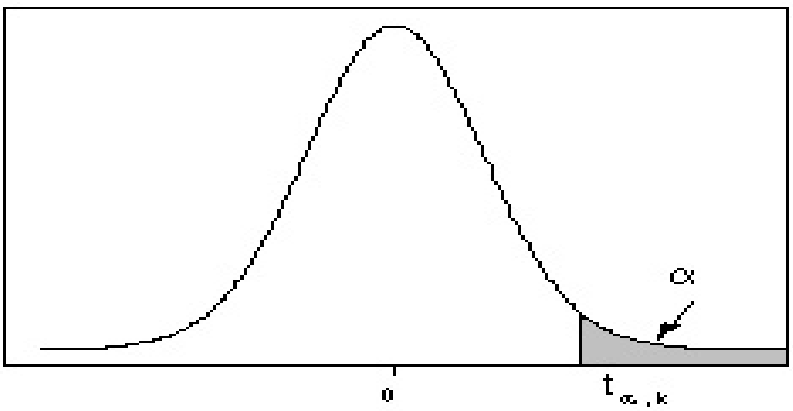


Figura 4.16: Pontos críticos, $t_{\alpha,k}$, da distribuição t-Student com k graus de liberdade

Solução: Se $T \sim t(12)$

(a) Da tabela t-Student tem-se: $P(T > 1,356) = 0,10$. Pela simetria da distribuição t-Student tem-se; $P(T > 1,356) = P(T < -1,356) = 0,10$. Portanto,

$$P(T > -1,356) = 1 - P(T < -1,356) = 1 - P(T > 1,356) = 1 - 0,10 = 0,90.$$

(b) $P(0,695 < T < 2,179) = P(T > 0,695) - P(T > 2,179) = 0,25 - 0,025 = 0,225$

(c) $P(-2,179 < T < 2) = P(T > -2,179) - P(T > 2)$. Mas na tabela t-Student não há o valor de 2 para 12 graus de liberdade (ou seja, não há $t_{\alpha,12}$). Porem, essa quantidade pode ser aproximado mediante uma interpolação linear.

$$\begin{array}{lcllllll} P(T > t_{\alpha,20}) & \rightarrow & 0,05 & \alpha & 0,025 & (0,05 - 0,025) & \rightarrow & (1,782 - 2,179) \\ t_{\alpha,20} & \rightarrow & 1,782 & 2 & 2,179 & (\alpha - 0,025) & \rightarrow & (2 - 2,179) \end{array}$$

daí tem-se:

$$\alpha = 0,025 + \frac{(0,05 - 0,025)(2 - 2,179)}{1,782 - 2,179} = 0,036272.$$

Logo,

$$\begin{aligned} P(-2,179 < T < 2) &= P(T > -2,179) - P(T > 2) = 1 - P(T > 2,179) - P(T > 2) \\ &= 1 - 0,025 - 0,036272 = 0,938728. \end{aligned}$$

(c)

$$\begin{aligned} P(-1,782 < T < 1,782) &= P(T > -1,782) - P(T > 1,782) = 1 - P(T < 1,782) - P(T > 1,782) \\ &= 1 - 2P(T > 1,782) = 1 - (2)(0,05) = 0,90. \end{aligned}$$

Observação 4.8.1 *Se* $T \sim t_{(k)}$ *e* $t_1 > 0 \in R$ *tem-se:*

1. $P(T > -t_1) = 1 - P(T > t_1)$
2. $P(-t_1 < T < t_1) = 1 - 2P(T > t_1)$

Teorema 4.8.3 *Seja $X_1 \ldots, X_n$ uma amostra aleatória de tamanho n de uma população normal com média μ e variância σ^2 (desconhecida). Assim, a variável aleatória*

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

tem distribuição t-Student com $n-1$ graus de liberdade

Exemplo 4.8.5 *De uma população normal com média μ, seleciona-se uma amostra aleatória de tamanho 16 sendo a variância amostral igual a 2,25. Qual é probabilidade de que média amostral difira da média real numa quantidade maior que 0,7543?*

Solução: $P(|\bar{X} - \mu| > 0,7543) =?$ Do teorema 4.8.3, tem-se

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\bar{X} - \mu}{2,25/\sqrt{16}} \sim t(15).$$

Logo,

$$P\left(\frac{|\bar{X} - \mu|}{S/\sqrt{n}} > \frac{0,7543}{2,25/\sqrt{16}}\right) = P(|T| > 1,341)$$

$$\begin{aligned} P(|T| > 1,341) &= 1 - P(-1,341 \leq T \leq 1,341) = 1 - [P(T > -1,341) - P(T > 1,341)] \\ &= 1 - [1 - P(T < 1,341) - P(T > 1,341) = 1 - [1 - 2P(T > 1,341)] \\ &= 2P(T > 1,341) = 2 \times 0,10 = 0,20 \end{aligned}$$

Teorema 4.8.4 *Seja $X_1, \ldots, X_n$ uma amostral aleatória de tamanho n de uma população com característica X, que tem distribuição normal com média μ_1 e variância σ^2 (desconhecida). Seja $Y_1, \ldots, Y_m$ outra amostra aleatória de tamanho m, de uma população com característica Y que tem distribuição normal com média μ_2 e variância σ^2 (desconhecida). Se X e Y são independentes, a variável aleatória:*

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}},$$

segue uma distribuição de t-student com $n + m - 2$ graus de liberdade, onde $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$, e é conhecida com a variância ponderada.

Prova: Se $X \sim N(\mu_1, \sigma^2)$ e $Y \sim N(\mu_1, \sigma^2)$ então $\bar{X} \sim N(\mu_1, \sigma^2/m)$ e $\bar{Y} \sim N(\mu_1, \sigma^2/m)$. Daí,

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}} \sim N(0,1) \tag{4.21}$$

Além disso,

$$W_1 = \frac{(n-1)S_1^2}{\sigma^2} \sim \chi^2_{(n-1)} \text{ e } W_2 = \frac{(m-1)S_2^2}{\sigma^2} \sim \chi^2_{(m-1)}$$

Pelo teorema 4.8.1, tem-se:

$$W = W_1 + W_2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2} \sim \chi^2_{(n+m-2)} \tag{4.22}$$

Além disso, as variáveis Z em (4.21) e W em (4.22) são independentes. Pela definição da distribuição t-Student tem-se:

$$T = \frac{Z}{\sqrt{\frac{W}{n+m-2}}} = \frac{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}}}{\sqrt{\frac{\frac{(n-1)S_1^2 + (m-1)S_2^2}{\sigma^2}}{n+m-2}}} = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}} \sim t_{(n+m-2)},$$

onde $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$.

4.8.3 Distribuição F-Snedecor

Definição 4.8.3 *Seja W_1 uma variável aleatória com distribuição Qui-quadrado com k_1 graus de liberdade e W_2 outra variável aleatória com distribuição Qui-quadrado com k_2 graus de liberdade. Se W_1 e W_2 são independentes, a variável aleatória,*

$$F = \frac{\frac{W_1}{k_1}}{\frac{W_2}{k_2}},$$

segue uma distribuição F-Snedecor com graus de liberdade, k_1 (numerador) e k_2 (denominador). A função de densidade de probabilidade é dada por:

$$h(f) = \frac{\Gamma(\frac{k_1+k_2}{2})}{\frac{\Gamma(k_1/2}{\Gamma(k_2/2)}} \frac{(\frac{k_1}{k_2})^{\frac{k_1}{2}} f^{\frac{k_1}{2}-1}}{(1+\frac{k_1}{k_2}f)^{\frac{k_1+k_2}{2}}}, \quad f > 0$$

A notação $F \sim F(k_1, k_2)$ indica que que a variável aleatória F tem distribuição F-Snedecor, com graus de liberdade k_1 e k_2.

Propriedades

Se $F \sim F(k_1, k_2)$ então

1. A distribuição é assimétrica direita.
2. A média e variância são respectivamente

$$\mu = \frac{k_2}{k_2-2}, \quad k_2 > 2 \ \ e \ \ \sigma^2 = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}, \quad k_2 > 4$$

Uso da tabela F-Snedecor

Os pontos críticos da distribuição F-Snedecor são apresentados na tabela D do apêndice. Seja $f_{\alpha,u,v}$ o ponto crítico da distribuição F com graus de liberdade numerador u e graus de liberdade denominador v, tal que a probabilidade de que variável aleatória F seja maior que este valor é

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} h(f)df = \alpha$$

Isto é ilustrado na figura 4.17. Por exemplo se $u = 5$ e $v = 10$, então da tabela C do apêndice, tem-se:

$$P(F > f_{0,05,5,10}) = P(F(5,10) > 3,33) = 0,05.$$

Isso é o ponto crítico do 5% superior de $F(3,5)$ é $f_{0,05,5,10} = 3,33$.

A tabela D contém, somente pontos críticos na cauda superior (valores de $f_{\alpha,u,v}$, para $\alpha \leq 0,25$) da distribuição F. Os pontos críticos na cauda inferior $f_{1-\alpha,u,v}$ podem ser obtidos da seguinte forma:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,v,u}}.$$

Por exemplo, para determinar o ponto crítico na cauda inferior $f_{0,95,5,10}$ observe que:

$$f_{0,95,5,10} = \frac{1}{f_{0,05,10,5}} = \frac{1}{4,74} = 0,211.$$

Exemplo 4.8.6 *Seja Y uma variável aleatória F-Snedecor.*

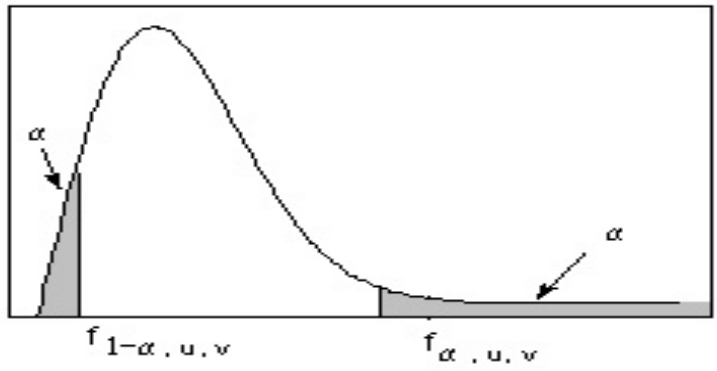


Figura 4.17: Pontos críticos, $f_{\alpha,u,v}$ e , $f_{1-\alpha,u,v}$ da distribuição F-Snedecor com u e v graus de liberdade

(a) Se $Y \sim F(8,12)$ *obtenha: (a1)* $P(Y > 2,85)$*; (a2)* $P(2,85 < Y < 4,50)$*; (a3)* y_1 *se* $P(y_1 < Y < 2,95) = 0,94$

(b) Se $Y \sim F(45,24)$*, achar* y_1 *tal que,* $P(Y \leq y_1) = 0,95$

Solução: Se $Y \sim F(8,12)$,

(a1) $P(Y > 2,85) = P(Y > f_{0,05,8,11}) = 0,05$.

(a2)$P(2,85 < Y < 4,50) = P(Y > 2,85) - P(Y < 4,50) = 0,05 - 0,01 = 0,04$

(a3)

$$\begin{aligned} P(y_1 < Y < 2,95) &= P(Y > y_1) - P(Y > 2,85) = 0,94 \\ &= P(Y > y_1) - 0,05 = 0,94. \end{aligned}$$

Dai tem-se: $P(Y > y_1) = 0,99$, $y_1 = f_{0,99,8,12}$. Logo, $y_1 = f_{0,99,8,12} = \frac{1}{f_{0,01,12,8}} = 1/5,67 = 0,176367$

(b) Se $Y \sim F(45,24)$, $P(Y \leq y_1) = 1 - P(Y > y_1) = 0,95$, daí tem-se: $P(Y > y_1) = 0,05$ e $y_1 = f_{0,05,45,24}$.

A tabela F-Snedecor não contem o valor crítico $f_{0,05,45,24}$. Esse valor pode ser aproximado mediante o processo de interpolação harmônica.

gl do numerador	40	45	60	$(1/45 - 1/60)$	$\rightarrow$	$(1/40 - 1/60)$
gl do denominador	24	24	24			
$f_{0,05,u,v}$	1,89	y_1	1,84	$(y_1 - 1,84)$	$\rightarrow$	$(1,89 - 1,84)$

Daí tem-se

$$y_1 = 1,84 + \frac{(1,89 - 1,84)(1/45 - 1/60)}{(1/40 - 1/60)} = 1,87333$$

Teorema 4.8.5 *Seja* $X_1, \ldots, X_n$ *uma amostra de tamanho n retirada de uma população, X que tem distribuição normal com média* μ_1 *(desconhecida) e variância,* σ_1^2*. Seja* $Y_1, \ldots, Y_m$ *uma amostra de tamanho m de uma população, Y, com distribuição normal com média* μ_2 *(desconhecida) e variância* σ_2^2 *e se X e Y são independentes, a variável aleatória,*

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

segue uma distribuição F-Snedecor com $n-1$ *e* $m-1$ *graus de liberdade.*

Exemplo 4.8.7 *Suponha que duas máquinas A e B produzem em forma independente um mesmo artigo. A maquina A é regulado produzir artigos com peso médio μ (desconhecido) e variância $\sigma_1^2 = 5$. Enquanto a maquina B foi regulada a produzir artigos com média μ a variância $\sigma_2^2 = 4$. Da produção da máquina A foi escolhidas ao acaso, uma amostra aleatória de $n = 11$ artigos e da máquina B uma amostra aleatória $m = 12$ artigos. Supondo que os pesos dos artigos produzidos pelas máquinas A e B seguem uma distribuição normal determine o valor de k tal que, $P(\frac{S_1^2}{S_2^2} > k) = 0,05$.*

Solução: Do teorema 4.8.5, tem-se que a variável

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{4}{5}\frac{S_1^2}{S_2^2},$$

segue uma distribuição F com 10 e 11 graus de liberdade, isto é, $F \sim F(10, 11)$. Portanto, $P(\frac{S_1^2}{S_2^2} > k) = P(\frac{4}{5}\frac{S_1^2}{S_2^2} > \frac{4}{5}k) = P(F(10, 11) > \frac{4}{5}k) = 0,05$, Da tabela C do apêndice temos que, $\frac{4}{5}k = 2,85 \longrightarrow k = 3,575$.

4.9 Exercícios

1. O Departamento de Matemática é formado por 35 professores, sendo 21 homens e 14 mulheres. Uma comissão de 3 professores será constituída, sorteando-se, ao acaso, três membros do departamento. Considere a variável aleatória X : número de mulheres na comissão. Determine:
   (a) A probabilidade a comissão ser formada por pelo menos duas mulheres.
   (b) O valor esperado e variância de X.
   (c) A função de distribuição acumulada de X.
   (d) A distribuição de probabilidades, valor esperado e a variância da variável $|1 - 2X|$

2. A produção diária de 850 peças contem 50 não satisfazem os requerimentos do cliente. Da produção de um dia escolhe-se ao acaso três peças uma a uma e sem reposição. Seja a variável aleatória X o número de peças da amostra que não cumpre com os requerimentos do cliente. Determine
   (a) a função de probabilidades de X e sua representação gráfica ,
   (b) a função de distribuição acumulada de X e sua representação gráfica,
   (c) Para variável $Y = 2X-1$, determine sua função de probabilidade e sua função de distribuição acumulada.
   (d) E(X) e Var(X).

3. Considere no exercício 2, que escolha foi com reposição. Determina a função de probabilidades da variável aleatória e a esperança de X e compare com os resultados do exercício anterior.

4. Num lote de 400 lâmpadas por experiências passadas se sabe que 20% são defeituosos. Do lote sortia-se uma amostra (sem reposição) de 3 lâmpadas. (i) Qual é probabilidade de obter no máximo 1 defeituosos na amostra. (ii) se cada lâmpada tem um custo de 1,5 unidades monetária (u.m) e vende-se a 3 u.m , qual é o lucro esperado na amostra.

5. Uma empresa Química paga a seus estagiários de acordo com o ano de curso do estudante. Para obter o salário mensal pago por 30 horas semanais, multiplica-se o salário mínimo pelo ano de curso do estagiário. Dessa forma, o estudante do primeiro ano ganha um salário mínimo, o do segundo recebe dois e assim por diante até o quinto ano. A empresa vai empregar 2 novos estagiários e admitimos que todos os anos têm igual número de estudantes interessados no estágio (considere a população de candidatos muito grande de modo a não haver diferença entre escolher com e sem reposição). Determinar:
   (a) O gasto médio da empresa nos estagiários ,
   (b) A probabilidade de que o empresário gaste mais de 5 salários mínimos nos estagiários?

(c) Determine a função de distribuição para variável aleatória gasto e faça sua representação gráfica.

6. Suponha que D, o número de medicamentos vendidos em uma semana, seja uma variável aleatória com a seguinte função de probabilidade:

$$f(d) = P(D = d) = \begin{cases} \frac{Cd^2}{d!}; & d = 1, 2, 3, 4 \\ 0; & \text{caso contrário} \end{cases}$$

(a) Determine: (i) A constante C para que f(d) seja a função de probabilidade de D. (ii) o número médio de medicamentos vendidos. (iii) a probabilidade do número de medicamentos vendidos em uma semana seja no máximo três peças.

(b) Se cada medicamento vendido o representante ganha uma comissão de R$ 12,00 e se o custo do medicamento é de R$ 3,00, qual é o lucro esperado em uma semana?

7. Uma variável aleatória X tem a seguinte função de probabilidade:

$$f(x) = \begin{cases} \frac{k}{2^x}; & x = 0, 1, \ldots \\ 0; & \text{caso contrário} \end{cases}$$

(a) Determine a constante k

(b) Determine a probabilidade que X assuma um valor par.

8. O tempo de duração(em anos) de certo microprocessador, é considerado uma variável aleatória contínua X, com a seguinte função de densidade de probabilidade

$$f(x) = \begin{cases} e^{-\frac{x-k}{10}}; & x \geq 2 \\ 0; & x < 2 \end{cases}$$

(a) Determine a constante k para que $f(x)$ seja uma função de densidade de probabilidade de X.

(b) Determine e interprete $E(X)$ e $Var(X)$,

(c) Qual é a probabilidade de um microprocessador dure mais de 5 anos em uma escolha aleatória?

(d) Determine a função de distribuição acumulada da variável tempo de vida,

(e) Se um microprocessador há durando mais de 7 anos, qual é a probabilidade que dure outros 2 anos?

9. Uma industria produz artigos cujos pesos (em kg) é uma variável aleatória contínua X, que tem a seguinte função de densidade de probabilidade

$$f(x) = \begin{cases} x - 8 & ; \quad 8 \leq x \leq 9 \\ 10 - x & ; \quad 9 < x < 10 \\ 0 & ; \quad \text{caso contrário} \end{cases}$$

(a) Determine a média e desvio padrão da variável aleatória X;

(b) O fabricante vende um artigo por um preço fixo de R$ 20,00 e garante o reembolso do preço de venda a qualquer cliente se o peso do artigo seja inferior a 8,25 kg. O custo de produção está relacionado ao peso do artigo de acorda com a expressão $0,05X + 0,50$. Expresse a variável lucro L, em termos da variável aleatória X.

(c) Determine o lucro esperado por artigo.

10. Sabe-se que com determinado tratamento alcança 60% de curas para certa doença quando o mesmo é administrado a pacientes em condições bem definidas. Se tratamento for aplicado a 20 pacientes nessas condições, qual é probabilidade de que:

(a) Ocorram no máximo 5 curas?

(b) Ocorram no mínimo 9 e no máximo 11 curas ?.

(c) Qual é o número esperado de curas? E qual a variância?.

11. O teste de DNA, feito numa clínica, tem 99.99% de confiabilidade nos resultados. Durante o último ano, num hospital, esse exame foi requisitado por 200 pessoas para a comprovação de paternidade. Com esses dados, calcule:

    (a) A probabilidade que 5 confirmações de paternidade estejam erradas.

    (b) A probabilidade que, ao menos, 2 confirmações estejam erradas.

12. Um fármaco usado para combater intoxicação causada pelo mercúrio, causa, em 45% dos pacientes, efeitos colaterais. Num teste feito em 10 pessoas contaminadas por mercúrio, obtenha:

    (a) A probabilidade de exatamente 5 pessoas apresentarem efeitos colaterais.

    (b) A probabilidade de menos de 2 pessoas apresentarem efeitos colaterais.

    (c) A probabilidade de ninguém apresentarem efeitos colaterais.

13. Num teste de laboratório para se medir a taxa de glicose no sangue, constatou-se que 25% das pessoas que fizeram o teste tinham glicose em torno de 100 mg/dl. Calcule a probabilidade de:

    (a) Em 10 pessoas que fizeram o teste, mais de 9 tenham glicose em torno de 100 mg/dl.

    (b) Em 50 pessoas que fizeram o teste, haja entre 5 e 10 pessoas com glicose em torno de 100 mg/dl.

14. Uma universidade processa 100.000 avaliações em determinado semestre, em ocasiões anteriores mostraram, que o 0,1% de todas avaliações estavam equivocadas. Suponha que uma pessoa faz cinco disciplinas nesta universidade em um semestre. Qual é a probabilidade que todas as avaliações estejam corretas?

15. Um exame de múltipla escolha consiste em 10 questões, cada uma com cinco possibilidades de escolha. A aprovação exige no mínimo 50%. Qual a chance de aprovação, se

    (a) O candidato comparece ao exame sem saber absolutamente nada, apelando apenas para o palpite.

    (b) O candidato estudou suficiente para poder eliminar três escolhas, devendo então apenas entre as duas escolhas restante.

16. Um time Mineiro de futebol tem probabilidade 0,70 de vitórias sempre que joga. Se o time atuar 4 vezes determine a probabilidade de que vença:

    (a) Todas as 4 partidas.

    (b) Exatamente 2 partidas.

    (c) Pelo menos uma partida.

    (d) No máximo 3 partidas.

    (e) Mais da metade das partidas.

17. Um corpo se encontra em repouso, no ponto (0,0). Lança-se um dado e por cada número primo que aparece o corpo se movimenta uma unidade de distância à direita, em caso contrário uma unidade à esquerda. Calcular a probabilidade que após 10 lançamentos o corpo se encontre:

    (a) a 8 unidades de distância à direita da origem;

    (b) a 3 unidades de distância à direita da origem;

    (c) a 2 unidades de distância à esquerda da origem;

    (d) a mais de uma unidade à direita da origem.

18. Um atirador faz três disparos a um alvo. Em cada um dos disparos a probabilidade de acertar é igual a 3/4. Acerta-se uma vez recebe $R\$50,0$, se acerta duas vezes recebe $R\$70,0$, se acerta três vezes recebe R$100,0 e nenhum dos disparos acertou o alvo, tem que pagar $R\$700$. Calcular o lucro esperado.

19. Uma mulher de 47 anos pretendia ter filhos através de inseminação artificial. Uma junta de técnicos da área fizeram testes para se saber qual o risco que ela poderia correr. Foi diagnosticado que, por ser uma mulher muito saudável, o único risco era de nascer uma criança com alguma doença genética. Assim, foi dado a probabilidade de 0,1 para ocorrer o nascimento de uma criança doente. Supondo que ela tenha 6 filhos, qual a probabilidade de 2 nascerem doentes.(Calcule usando a distribuição Poisson e a distribuição Binomial)

20. O número de partículas emitidas por uma fonte radiativa, durante o período especificado, é uma variável aleatória de Poisson. Se a probabilidade de não houver emissões for igual a 1/3, qual é a probabilidade de que 2 ou mais emissões ocorram?

21. Laminas de metal apresentam defeitos no cromado, segundo uma distribuição de Poisson, com uma média de um defeito por m^2. Essas laminas são usadas para construção de janelas para uma instalação industrial cuja dimensão são de, $150\,cm \times 200\,cm$.

    (a) Em um grupo 10 dessas janelas qual é a probabilidade de que no máximo 4 delas não tenha nenhum defeito?

    (b) Em um grupo de 3 dessas janelas, qual é a probabilidade de total de falhas nesse grupo seja no máximo três?

22. Em uma fabrica, a maquina 1 produz por dia o dobro de peças que a maquina 2 e, a maquina 3 o triplo da maquina 1. Sabe-se que 6% das peças fabricadas pela maquina 1 tendem a ser defeituosas, e o 4% das peças produzidas pela maquina 2 tendem a ser defeituosas, enquanto 8% de peças defeituosas da maquina 3. A produção diária é misturada. Extraída uma amostra aleatória (com reposição) de 20 peças, qual é a probabilidade de que essa amostra contenha:

    (a) No máximo duas peças defeituosas?

    (b) Entre três e cinco peças defeituosas?

    (c) Suponha que o total de peças produzidas por dia é de 1000 peças. Refaça o item (a) se amostragem foi sem reposição.

23. Foi analisada uma cultura de bactérias para se obter o número médio de bactérias por mm^2. Os dados obtidos, levaram a se prever a probabilidade de não se encontrar nenhuma bactéria escolhendo-se, aleatoriamente, um $1mm^2$ na placa de cultura que é igual a 0.006734. Calcule, assim, o valor médio de bactérias por mm^2, sabendo que a variável "no de bactérias $/mm^2$ da placa de cultura"constitui uma distribuição de Poisson.

24. Em uma comunidade isolada no himalaia, foram feitas medições de nível de colesterol no sangue nos moradores locais. O valor da média encontrado foi de 178 mg/dl e um desvio padrão igual a 10 mg/dl. Supondo que o nível de colesterol dessa população tem distribuição normal obtenha:

    (a) a probabilidade de um morador dessa comunidade apresentar taxa de colesterol igual a 180 mg/dl.

    (b) a probabilidade de um morador se encontrar entre 168 e 188 mg/dl.

25. Um vendedor de automóveis sabe que o número de carros vendidos por dia em sua loja comporta-se como uma variável de Poisson cuja média é 2 nos dias de bom tempo, e é 1 nos dias chuvosos. Se em 70% dos dias faz bom tempo, qual é a probabilidade de que em certo dia do ano sejam vendidos pelo menos três automóveis?

26. Considere um experimento que consiste em contar o número de partículas alfa emitidas, num intervalo de tempo de um segundo. Sabe-se por experiências passada que, em média, 3 de tais partículas são emitidas por segundo. Determinar a probabilidade de que não mais de duas partículas alfa sejam emitidas em um quarto de segundos.

27. Um determinado fármaco, usado para combater infecção, foi usado em cobaias para se verificar sua eficácia. Foi usado quantidades variáveis do fármaco que se assemelha de uma variável aleatória com distribuição normal. Assim, foi obtida a probabilidade de 99.9% de que os animais foram tratados com uma quantidade de fármaco igual ou menor a 171 mg. Calcule a média de fármaco utilizado nas cobaias, sabendo que por estudos similares $\sigma = 5$ mg.

28. A dureza H de uma peça de aço pode ser pensada como sendo uma variável aleatória com distribuição uniforme no intervalo (50,90) da escala de Rochkwel. Qual é a probabilidade que a peça tenha dureza entre 55 e 60.

29. O petróleo é separado por destilação nas frações, listados na tabela seguinte

Fração	Temperatura de destilação (^0C)	Preço de venda por galão (US \$)
Gás	Menos de 20	C_1
Petróleo éter	$20 \mid- 60$	C_2
Ligroin	$60 \mid- 100$	C_3

Suponha que C dólares é o custo de produzir um galão de petróleo e a temperatura de destilação T está distribuído uniformemente em $[0, 100]$. Achar o lucro esperado (por galão) pelas frações.

30. Suponha que um fabricante tenha que decidir entre dois processos de fabricação de certa componente eletrônica. O custo do processo A é de c dólares e do processo B é kc dólares por unidade de componente, onde $k > 1$. Os tempos de falhas das componentes eletrônicas pode ser consideradas como uma variável aleatória exponencial com média de falha de 200 horas para os fabricados pelo processo A e 300 horas para B. Admita-se, além disso, que se a componente dure menos de 400 horas, pagará uma multa de D dólares. Que processo deverá usar ?

31. O 5% das lâmpadas produzidas por certa maquina são defeituosos. O tempo de vida,T, de uma lâmpada defeituosa é uma variável exponencial com média 0,5 ano, enquanto que o tempo de vida T1 de uma lâmpada não defeituosa é uma variável aleatória exponencial com média 2 anos. Calcular a probabilidade de uma lâmpada:

    (a) Se queimar antes dos 2 anos.

    (b) Durar entre 2 e 4 anos.

32. Certo tipo de fusível tem duração de vida que segue uma distribuição exponencial com tempo médio de vida de 100 horas. Cada peça tem um custo de 10,0 unidades monetárias (u.m) e se durar menos de 20 horas, existe um custo adicional de 8.0 u.m.

    (a) Qual é a probabilidade de uma durar mais de 150 horas?

    (b) Determinar o custo esperado.

33. A fabrica de pneu "DURAMAS"produz um tipo de pneus que tem uma vida útil média de 80.000 km e um desvio padrão de 8.000 km. Supondo que essa vida útil tem distribuição normal :

    (a) qual é a probabilidade de que um pneu dure más de 96.000 km ?

    (b) O 50% dos pneus durem entre a e b quilômetros. Achar os valores a e b, sim eles são simétricos respeito à média.

34. Um combustível para foguetes vai a conter certo porcentagem (chamado de X) de um componente especial. As especificações exigem que X esteja compreendido entre 30 a 35 por cento. O fabricante terá um lucro liquido no combustível (por galão) que é a seguinte função de X:

$$T(X) = \begin{cases} -0,10 & \text{por galão} \quad \text{se} \quad 30 < x < 35 \\ 0,05 & \text{por galão} \quad \text{se} \quad 33 \leq x < 40 \ \text{ ou } \ 25 \leq x \leq 30 \\ 0,10 & \text{caso contrário} \end{cases}$$

    Se $X \sim N(33, 9)$. Calcular (a) a função de probabilidade de $T(X)$, (b) $E(T(X))$.

35. Um teste de aptidão feito por pilotos de aeronaves em treinamento inicial requer que uma série de operações seja realizada em uma rápida sucessão. Suponha que o tempo necessário para completar o teste seja distribuído de acordo com uma Normal de média 90 minutos e desvio padrão 20 minutos.

    (a) Para passar no teste, o candidato deve completá-lo em menos de 80 minutos. Se 65 candidatos tomam o teste, quantos são esperados passar no teste?

(b) Se os 5% melhores candidatos serão alocados para aeronaves maiores, quão rápido deve ser o candidato para que obtenha essa posição?

36. Estudos meteorológicos indicam que a precipitação pluviométrica mensal em períodos de seca numa certa região pode ser considerada como seguindo a distribuição Normal de média 30 mm e variância 16 mm^2.

    (a) Qual a probabilidade de que a precipitação pluviométrica mensal no período da seca esteja entre 24mm e 38mm?

    (b) Qual seria o valor da precipitação pluviométrica de modo que exista apenas 10% de chance de haver uma precipitação inferior a esse valor?

    (c) Construa um intervalo central em torno da média que contenha 80% dos possíveis valores de precipitação pluviométrica.

37. Numa certa população, o peso dos homens tem distribuição normal com média 75kg e desvio padrão 10kg, enquanto que o das mulheres é também normal com média 60kg e desvio padrão 4kg.

    (a) Sorteando-se um homem qualquer, qual é a probabilidade dele ter peso acima de 65kg?

    (b) Sorteando-se uma mulher qualquer, qual é a probabilidade dela ter peso acima de 65kg?

    (c) Qual é a probabilidade de uma pessoa ter peso acima de 65kg, sendo ela sorteada de um grupo em que o número de mulheres é o dobro do de homens?.

38. O diâmetro X de rolamentos de esfera fabricados por uma certa fábrica tem distribuição normal com média 0,614 cm e desvio padrão 0,0025. O lucro T de cada esfera depende de seu diâmetro, e $T = 0,10$ se a esfera é boa, isto é, se $(0,61 < X < 0,618)$; T=0,05 se a esfera é recuperável, isto é, se $(0,608 < X < 0,61)$ ou $(0,618 < X < 0,62)$; T=-0,10 se a esfera é defeituosa, isto é, $(X < 0,6080$ ou $X > 0,620)$. Calcular:

    (a) As probabilidades de as esferas serem boas, recuperáveis e defeituosas:

    (b) O valor médio do lucro T.

39. Supondo que numa população de pessoas normais a pressão de pulso é uma variável aleatória tem distribuição normal com média 40 mmHg e desvio padrão 16 mmHg. Se uma pessoa é selecionada dessa população obtenha:

    (a) a probabilidade da pessoa apresentar pressão de pulso menor a 45 mmHg a e maior 60 mmHg.

    (b) a probabilidade da pessoa sorteada apresentar pressão de pulso menor que 55 mmHg.

40. Em uma espécie animal, a taxa normal de hemoglobina é uma variável aleatória com distribuição normal com média $\mu = 150g/L$ de sangue e variância, $\sigma = 144g/L$ de sangue. Se uma animal dessa espécie é selecionada ao acaso, qual a probabilidade de que a taxa de hemoglobina normal, estar entre 146 e 153 g/L.?

41. Um estudo feito em duas cidades (A e B) de Minas obteve o valor médio e o desvio padrão da concentração de glicose no sangue de pessoas que não apresentavam distúrbios fisiológicos em relação a concentração de glicose no sangue.

    Cidade A $\mu_1 = 104.8mg/100mL$ de sangue $\sigma_1 = 6.4mg/100mL$ de sangue.

    Cidade B $\mu_2 = 102.3mg/100mL$ de sangue $\sigma_2 = 4.9mg/100mL$ de sangue.

    Admitindo que a concentração de glicose no sangue de pessoas das duas cidades tem distribuição normal,

    (a) calcule a probabilidade de uma pessoa da cidade A ter a concentração de glicose no sangue seja pelo menos 100 mg/100mL de sangue.

    (b) calcule a probabilidade de uma pessoa da cidade B ter a concentração de glicose no sangue pelo menos 100 mg/100mL de sangue.

    (c) Retirando-se uma pessoa de amostra contendo a proporção de 1:3 para moradores da cidade A e B, Qual a probabilidade dessa pessoa ter a concentração de glicose seja pelo menos 100 mg/100mL de sangue

    (a) se uma pessoa é sorteada ao acaso de cada uma das cidades, qual é probabilidade que a concentração de glicose da pessoa da cidade A seja maior ao da pessoa da cidade de B.?

42. A concentração de uma substância X no sangue tem distribuição normal com média 10 mg e desvio padrão 2 mg por unidade de volume. É considerado doente o indivíduo que tenha uma dosagem menor que 6,0 mg ou maior que 13,5 mg.

    (a) Se um indivíduo é escolhido ao acaso, qual é a probabilidade dele ser considerado doente ?

    (b) Em 100 pessoas escolhidas ao acaso, qual é a probabilidade de observamos no máximo 2 doentes?.

    (c) Se escolhemos ao acaso 30 pessoas, qual é a probabilidade de que a concentração média da substância das 30 pessoas ultrapasse 11 mg?

43. A capacidade máxima de um elevador é de 500 kg. Se a distribuição dos pesos dos usuários é suposta normal com média 70 kg e desvio padrão 10 kg. Qual é probabilidade de que 10 passageiros ultrapassem esse limite ?.

44. Um braço mecânico consta de três partes. Suponha que X, Y e Z são produzidos por diferentes fabricas e cuja longitude de cada um estão dado por : $X \sim N(12, 0, 02)$, $Y \sim N(24, 0, 03)$ e $Z \sim N(18, 0, 04)$, onde a média está dado em centímetros e variância em centímetros quadrados. Calcular a probabilidade do braço esteja compreendido entre 53.8 y 54.2.

45. Uma corretora de negocia título na Bolsa de Valores e utiliza um modelo probabilístico para avaliar o lucro seus lucros. Suas aplicações financeiras de compra e venda atingem três áreas: agricultura, industria e comércio. Admite que o seguinte modelo representa o comportamento do lucro diário da corretora (em milhares de dólares)$L = 3L_A + 5L_I + 4L_C$, com L_A, L_I e L_C representando respectivamente os lucros diários nos setores de agricultura, industria e comércio. As distribuições de probabilidade dessas variáveis aleatórias são $L_A \sim N(3, 5)$, $L_I \sim N(6, 9)$ e $L_C \sim N(4, 16)$. Supondo independência entre os três setores, qual será a probabilidade de um lucro diário acima de 50 mil ?.

46. O tempo gasto no exame de uma universidade tem distribuição normal com média 100 minutos e desvio padrão 10 minutos.

    (a) Qual é a porcentagem de vestibulandos que gastam no máximo 90 minutos no exame?

    (b) Qual é probabilidade de que um vestibulando gaste exatamente 160 minutos?

    (c) Qual deve ser o tempo da prova, de modo que 90% dos vestibulandos terminem no prazo estipulado?

    (d) Dez vestibulandos foram sorteados ao acaso, qual é a probabilidade que pelo menos dois alunos gastem no máximo 90 minutos?

    (e) Suponha que o total de vestibulandos foi 700. Refaça o item (d) se amostragem foi sem reposição.

47. A dimensão de hastes metálicas fabricadas em série é uma variável aleatória normalmente distribuída com média 60 cm e variância 4 cm. Ao se coletar uma amostra aleatória de 10 valores determine:

    (a) A probabilidade de que a média amostral esteja situada entre 59,5 a 60,5 cm.

    (b) A probabilidade de que variância amostra seja inferior a 3 cm.

    (c) Refaça os cálculos indicados nos ítens (a) e (b) supondo uma amostra com n=20.

48. Se tomarmos uma amostra de 20 elementos de uma variável aleatória X tal que $X \sim N(\mu, \sigma^2)$ e se nesta amostra obtivermos $S = 5$. Com que probabilidade podemos afirmar que a média da amostra não se afaste de em mais de uma unidade.

49. Suponhamos que uma central atacadista tenha como média para o montante de vendas o valor de 150 OTN's e como desvio padrão o valor 10 OTN's . Suponha ainda que 20% das vendas efetuadas tenha valor superior a 170 OTN's. Nestas condições ao se coletar uma amostra de 100 clientes calcular:

    (a) A probabilidade de que a média encontrada na amostra se distância da média real em mais de 2 unidades .

    (b) A probabilidade de que a amostra apresente mais de 26 clientes que efetuem compras com valor superior a 170 OTN's

50. Admitimos que em um lote de 800 motores apresente 200 com um determinado defeito. Ao coletarmos uma amostra de 50 motores sem reposição, qual é a probabilidade de que a mesma apresente menos de 10 motores com defeito.

51. Constatou-se que um lote de 20.000 faturas de uma grande cadeia de lojas apresenta média de 4,5 OTN's e como desvio padrão o valor 0,5 OTN's , sendo ainda que 30% das mesmas superior a 0,5 OTN's. Tomada uma amostra (sem reposição) de 225 faturas, calcular:

    (a) A probabilidade de que a média amostral se afastar em 0,01 OTN's da média real .

    (b) Qual a probabilidade de que dentre as 225 faturas observadas mais de 60 apresentem um valor superior a 5,0 OTN's

52. A maquina de empacotar um determinado produto o faz segundo uma distribuição normal, com média μ e desvio padrão 10 gr.

    (a) Em quanto deve ser regulado o peso médio para que apenas 10% dos pacotes tenham menos do que 500.

    (b) Com a maquina assim regulada qual é a probabilidade de que o peso total de 4 pacotes escolhidos ao acaso seja inferior a 2 kg ?

53. No exercício anterior, após a maquina estar regulada programou-se uma carta de controle de qualidade. De hora em hora, será retirada uma amostra de 4 pacotes, e estes serão pesados. Se a média da amostra foi inferior a 4095 gr ou superior a 520 gr, para-se a produção para reajustar a máquina isto é, reajustar o peso médio.

    (a) Qual a probabilidade de ser feita uma parada desnecessária ?

    (b) Se o peso médio da maquina desregulou-se para 500 gr, qual a probabilidade de continuar-se a produção fora dos padrões desejados. ?

54. Uma empresa recebe certo componente em grandes lotes. Sabendo-se que o fornecedor envia lotes com 10% de peças defeituosas, qual é a probabilidade de numa amostra com 100 ítens, a proporção defeituosa ser

    (a) 17% ou mais ?

    (b) entre 9,5% e 10% ?

    (c) menor que 8% ?

    (d) maior que 9 %?

55. Cerca de 15% dos bares em Ouro preto vendem fiado a seus clientes. Determine a probabilidade de, numa amostra aleatória de 64 bares:

    (a) 16% ou menos venderem fiado.

    (b) Entre 15% e 16% venderem fiando.

    (c) Mais de 15% e 17% venderem fiado.

56. Sabendo-se que 70% da população ativa do Brasil ganha menos de 3 salários mínimos, qual é a probabilidade de que uma amostra aleatória com 900 pessoas apresentar:

    (a) mais de 67% das pessoas da amostra recebendo menos de 3 salários mínimos ?

    (b) mais que 72% ou menos que 68% da amostra ganhando menos que 3 salários mínimos?

    (c) Entre 540 a 720 pessoas com renda menor que 3 salários mínimos ?

57. Suponha que tem-se 2 processos (A e B) para produzir um artigo, e que o tempo médio de produção para o processo A é 300 horas e desvio padrão 16 horas, enquanto que para o processo B tem o tempo médio de 306 horas e uma desvio padrão de 12 horas. Se sorteiam-se uma amostra aleatória de 64 artigos produzidos com processo A e 49 produzidos com o processo B, calcular a probabilidade que:

    (a) A diferença de médias amostrais seja superior a 2 horas.

(b) O tempo médio de produção da amostra do processo A seja menor ao correspondente processo B.

(c) Refaça os cálculos indicados nos ítens (a) e (b) supondo que as amostras foram selecionados sem reposição de um lote de 500 artigos produzidos pelo processo A, e de um lote de 480 artigos produzidos pelo processo B.

58. Suponha que uma empresa de comercialização tem 2 lojas A , B e que porcentagens de clientes que consideram que a atenção dada é boa na loja A de 70% entanto que na loja B é de 63%. Para avaliar a opinião dos clientes enquanto ao atendimento seleciona-se amostras aleatórias de tamanhos: 50 para a loja A e 60 para a loja B, calcular a probabilidade de que a proporção de clientes satisfeitos pela atenção recebida pela loja A na amostra supere aos dados pela loja B em menos de 0,05% se:

(a) A amostra é com reposição.

(b) A amostra sem reposição, tendo-se escolhida a amostra da loja A de uma total de 900 clientes e a de B de um total de 1400 clientes.

59. Suponha que os pesos de artigos produzidos por uma maquina tem distribuição normal com média μ e variância 25 gr. Se escolhe ao acaso 16 artigos, calcular:

(a) $P(S^2 > 32,128)$

(b) O valor de k tal que $P(S < k) = 0,6$

60. Suponha que 2 maquinas A e B produzem um mesmo artigo e que os pesos por artigo (em gramas) tem distribuição normais com médias: $\mu_1 = 550$ e $\mu_2 = 565$ e variâncias: $\sigma_1^2 = 144$ e $\sigma_2^2 = 256$ respectivamente. Escolhe-se ao acaso 21 artigos produzidos pela maquina A e 31 produzidos pela maquina B, calcular :

(a) a probabilidade de que o peso médio de produção da amostra da maquina A seja maior do peso médio dos produzidos pela maquina B em mais de 2 gr.

(b) $P(1,08563 \leq \frac{S_1^2}{S_2^2} \leq 1,4344)$

Capítulo 5

Inferência Estatística

5.1 Introdução

A inferência estatística é o processo que consiste em utilizar os resultados de uma amostra para tirar conclusões gerais de uma ou mais características de uma população. Ela compreende: estimação de parâmetros e teste de hipóteses estatística.

5.2 Estimação de Parâmetros

No capítulo anterior foram considerados diversas distribuições de probabilidade. Muitas vezes sabe-se ou admite-se que uma variável aleatória X (característica da população) segue uma certa distribuição de probabilidade, mas não são conhecidos os valores dos parâmetros da distribuição. Por exemplo, se X seguir a distribuição normal, pode-se querer saber o valor de seus parâmetros (a média e a variância). Para estimar os parâmetros, considera-se uma amostra aleatória de tamanho n e, utiliza-se os dados amostrais para estimar os parâmetros desconhecidos. Isso é conhecido como o problema de estimação. E esse problema pode ser dividido em duas categorias: estimação pontual e estimação por intervalos.

5.2.1 Estimação pontual

Para fixar os conceitos, seja X alguma característica da população com função de probabilidade ou função de densidade $f(x;\theta)$, onde θ é o parâmetro da distribuição. Suponha que conhecida a forma funcional de $f(x;\theta)$, como por exemplo, a distribuição normal, mas não se sabe o valor de θ. Portanto, sorteia-se uma amostra aleatória de tamanho n e desenvolve-se uma função dos valores amostrais

$$\hat{\theta} = h(X_1, \ldots, X_n)$$

que forneça uma estimativa de θ. $\hat{\theta}$ é conhecido como um *estimador*, e um valor numérico particular assumido pelo estimador é conhecido como uma *estimativa*. Note que $\hat{\theta}$ pode ser tratado como uma variável aleatória, pois é uma função dos dados amostrais. O estimador $\hat{\theta}$ fornece uma regra, ou fórmula, que diz como se pode estimar o θ verdadeiro. Assim, ao se admitir que

$$\hat{\theta} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \bar{X}$$

temos que $\bar{X}$, a média amostral, é um estimador do valor médio verdadeiro (ou populacional), μ. Se em um caso especifico, $\bar{X} = 50$, tem-se uma *estimativa* de μ. O estimador θ obtido anteriormente nos fornece uma única estimativa (pontual) de θ.

5.2.2 Estimação por intervalos

Ao invés de se obter uma única estimativa de θ, suponha que obtém-se duas estimativas de θ por meio da construção de dois estimadores, $\hat{\theta}_1$ e $\hat{\theta}_2$, e considera-se com alguma confiança (isto é, probabilidade) que o intervalo entre $\hat{\theta}_1$ e $\hat{\theta}_2$ inclui o verdadeiro θ. Assim, em um estimativa por intervalo, em contraste com a estimativa pontual, fornecemos uma classe de possíveis valores dentro do qual se pode encontrar o verdadeiro θ.

Definição 5.2.1 (Intervalos de confiança) *Seja $X_1, \ldots, X_n$ uma amostra aleatória de população com a característica X, cuja distribuição de probabilidade é $f(x;\theta)$. Seja $T_1 = G(X_1, \ldots, X_n)$ e $T_2 = H(X_1, \ldots, X_n)$ duas estatísticas tais que $T_1 < T_2$ e que*

$$P(T_1 < \theta < T_1) = \gamma = 1 - \alpha.$$

Então, o intervalo $(T_1; T_2)$ é chamado de intervalo de $100\gamma\%$ ou $(1-\alpha)100\%$ de confiança para θ.

Denota-se por $IC(\theta, 1-\alpha)$, o intervalo de $(1-\alpha)100\%$ de confiança para θ. Isto é,

$$IC(\theta; 1-\alpha) = (\ T_1\ ;\ T_2\ )$$

onde T_2 e T_1 são os limites superior e inferior de confiança respectivamente e $\gamma = 1 - \alpha$ é o coeficiente (ou nível) de confiança. A escolha do coeficiente de confiança depende do pesquisador e os valores mais utilizados são $\gamma = 1 - \alpha = 0,90;\ 0,95;\ 0,98;\ 0,99$.

Supondo que uma característica, X, da população tem distribuição normal ou qualquer outra distribuição e considerando as distribuições amostrais estudadas nos capítulos anteriores pode-se deduzir intervalos de confiança para: uma média populacional, uma proporção populacional, uma variância populacional, diferença de médias e razão de variâncias.

5.3 Intervalos de confiança para média de uma população (μ)

5.3.1 Quando variância σ^2 é conhecida

Suponha que $X_1, \ldots, X_n$, seja uma amostra aleatória de tamanho n extraída de uma população, com a característica X, que tem distribuição normal com média μ e variância σ^2. Foi visto que a média amostral $\bar{X}$ tem distribuição normal com média μ e variância σ^2/n. Assim

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

Logo, fixando um valor $(1-\alpha)$, encontrar-se $z_{\alpha/2}$ tal que:

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$$

ou, o que é equivalente,

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha. \tag{5.1}$$

Note que $z_{\alpha/2}$ pode ser obtida de tabela da distribuição normal padrão, utilizando-a de forma inversa aquela discutida no capítulo anterior e como mostra a figura 5.1, abaixo.

De (5.1)

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \Rightarrow \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Assim, o intervalo de confiança para μ, com *coeficiente de confiança* $(1-\alpha)$, é dado por

$$IC(\mu; 1-\alpha) = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right). \tag{5.2}$$

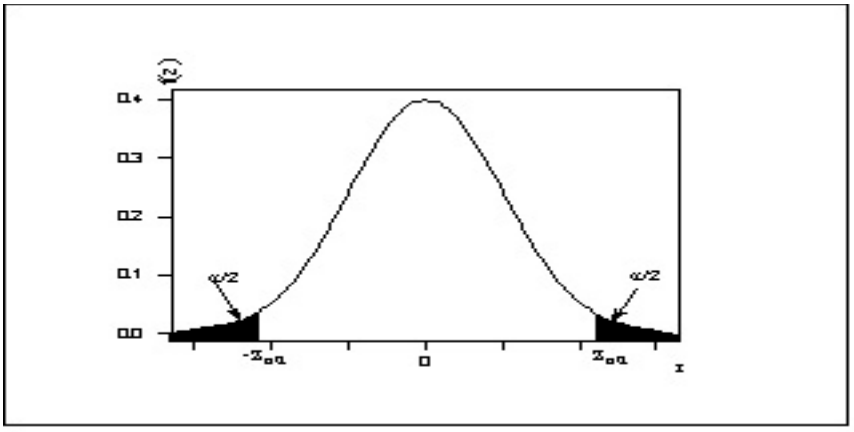


Figura 5.1: Distribuição normal padrão $P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha$.

Um erro muito comum que se comente ao interpretar o intervalo de confiança é dizer que a probabilidade de μ estar no intervalo é $1 - \alpha$. O erro resulta do fato de que μ não é uma variável aleatória e sim um parâmetro que caracteriza uma população. Ou seja, μ não varia e portanto, não tem uma distribuição de probabilidade. Deve ficar claro o que é aleatório (antes de que seja obtida a amostra e calculada os valores) é o intervalo. Portanto, o correto seria dizer que a probabilidade do intervalo a ser escolhido conter o verdadeiro valor da média é igual a $1 - \alpha$. Outra interpretação considerada é a seguinte: obtendo várias amostras e, para cada uma delas, calculando o correspondente intervalo de confiança para μ, teremos que um $100(1 - \alpha)\%$ das amostras conterão o valor de μ e $100\alpha\%$ das amostras não conterão a média populacional.

Exemplo 5.3.1 *Um pesquisador deseja estimar, com 99% de confiança a média da força máxima de um certo músculo de um grupo de indivíduos. Ele considera que os valores da força muscular estão distribuídos normalmente com variância de 144. Com esta finalidade selecionou-se uma amostra aleatória de 15 indivíduos da mesma faixa etária e do mesmo peso e obteve-se que* $\bar{X} = 84,3$. *Qual é o intervalo?*

Da tabela normal padrão temos que $z_{\alpha/2} = z_{0,005} = 2,57$. Substituindo em (5.2) temos que

$$\begin{aligned} IC(\mu; 0,99) &= \left(84,3 - 2,57\frac{12}{\sqrt{15}}; 84,3 + 2,57\frac{12}{\sqrt{15}}\right) \\ &= (84,3 - 7,9628; 84,3 + 7,9628) \\ &= (76,3372; 92,2672.) \end{aligned}$$

A interpretação deste intervalo de confiança é: dado o coeficiente de confiança de 99%, a longo prazo, em 99 de 100 casos, intervalos como $(76,3372; 92,2672)$ conterão a média verdadeira da força máxima de um certo músculo do grupo de indivíduos. Note, porém, que não se pode dizer que é 99% a probabilidade do intervalo especifico $(76,3372; 92,2672)$ conter a média verdadeira (μ) da força máxima de um certo músculo, pois, esse intervalo agora esta fixado, não é mais aleatório. Logo μ ou se encontra nele ou não se encontra: a probabilidade de o intervalo fixado especifico incluir o verdadeiro μ é portanto, de 1 ou 0.

Observação 5.3.1 *A continuação apresenta-se intervalos de confiança para o caso de populações finitas:*

(a) Se σ *é desconhecido e* $n \geq 30$, *pode-se utilizar o desvio padrão amostral* S *para aproximar* σ.

(b) No caso que a população é finita de N *elementos e* σ *é conhecido e amostragem é sem reposição, o intervalo de* $(1 - \alpha)100\%$ *de confiança para* μ *é:*

$$IC(\mu; 1 - \alpha) = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right). \tag{5.3}$$

Se σ é desconhecido e $n \geq 30$, por (a) o intervalo é

$$IC(\mu; 1-\alpha) = \left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}; \bar{X} + z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right). \tag{5.4}$$

(c) $z_{\frac{\alpha}{2}}$ é uma função crescente do coeficiente de confiança $\gamma = 1 - \alpha$. Portanto, se $\gamma \to 1$, o comprimento do intervalo de confiança é maior.

(d) O tamanho da amostra aparece no denominador de $z_{\frac{\alpha}{2}}\sigma$. Para amostras grandes os intervalos de confiança têm comprimentos mais curtos, portanto, informação mais precisa.

Exemplo 5.3.2 *De um lote de 2200 lâmpadas foram sorteadas 81 lâmpadas ao acaso, o tempo médio de duração das lâmpadas sorteadas foi 3200 horas com um desvio padrão de 900 horas. Construa um intervalo de 95% de confiança para o tempo médio das lâmpadas do lote.*

Solução: Já que $1 - \alpha = 0,95$, temos da tabela normal padrão $z_{\alpha/2} = z_{0,025} = 1,96$.

Como $\bar{X} = 3200$, $S = 900$ $n = 81$ e $N = 2200$ (tamanho da população finita), pela observação 5.3.1.b, tem-se:

$$\begin{aligned} IC(\mu; 1-\alpha) &= \left(\bar{X} - z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}; \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \right) \\ &= \left(3200 - \frac{(1,96)(900)}{\sqrt{81}} \sqrt{\frac{2200-81}{2200-1}}; 3200 + \frac{(1,96)(900)}{\sqrt{81}} \sqrt{\frac{2200-81}{2200-1}} \right) \\ &= (3008; 3396). \end{aligned}$$

Determinação do tamanho da amostra para estimar a média μ

A determinação do tamanho da amostra for muito importante, uma vez que, se a amostra for muito pequena não será significativa e, se a amostra for muito grande estão desperdiçando recursos. Utiliza-se o intervalo de confiança para calcular tamanho de uma amostra. Do intervalo de confiança para a média populacional

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

deseja-se que o comprimento do intervalo seja o mais curto possível, para isso tem-se duas opções:

(i) Diminuir o coeficiente de confiança: $1 - \alpha$

(ii) Aumentar o tamanho da amostra, o que diminui o erro padrão ($\frac{z_{\frac{\alpha}{2}}}{\sigma/\sqrt{n}}$), já que σ é fixo.

Dessas duas opções, a primeira não é recomendável porque aumenta-se α,que é o risco de que μ não esteja no intervalo.

Há uma conseqüência interessante que se desprende da relação entre o erro máximo de estimação (diferença entre o estimador e o parâmetro) e o risco (α definido anteriormente) que é a determinação do tamanho da amostra. O comprimento ou amplitude do intervalo é:

$$L = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Onde o erro máximo da estimação, denotado por E, é:

$$E = \frac{L}{2} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Dessa equação é possível obter n se o erro máximo de estimação E, o risco α e a variância populacional são conhecidos. Ou seja,

$$n = \frac{z^2_{\alpha/2}\sigma^2}{E^2}$$

Se a amostragem é sem reposição, é introduzido o fator de correção de população finita: $\sqrt{\frac{N-n}{N-1}}$, de onde:

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}},$$

que ao resolver em n, tem-se

$$n = \frac{Nz^2_{\alpha/2}\sigma^2}{E^2(N-1) + z^2_{\alpha/2}\sigma^2}.$$

Se o tamanho da população finita N é muito maior em comparação com n (isto é, $\frac{n}{N} < 0,10$) o fator de correção de população finita pode ser ignorado.

Exemplo 5.3.3 *Uma firma construtora deseja estimar a resistência média das barras de aço utilizadas na construção de casas. Qual o tamanho amostral necessário para garantir que haja um risco de 0,001 de ultrapassar um erro de 5 kg ou mais na estimação ? O desvio padrão da resistência para este tipo de barra é estimado em 25 kg.*

Solução: $E = 5$kg, $\sigma = 25$kg. Como o risco de ultrapassar esse erro é de 0,001, então, $\gamma = 1-\alpha = 1-0,001 = 0,999$. Logo, $z_{0,0005} = 3,29$. Daí, tem-se

$$n = \frac{z^2_{0,0005}\sigma^2}{E^2} = \frac{(3,29)^2(25^2)}{5^3} = 270,6025 \approx 271.$$

5.3.2 Quando a variância populacional σ^2 é desconhecida

Se $X_1, \ldots, X_n$ é uma amostra aleatória de tamanho n, de uma população normal com média μ e variância desconhecida σ^2 a variável aleatória,

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

tem distribuição t-Student com $n-1$ graus de liberdade. Seguindo o procedimento anterior, para o nível de confiança fixado,$100(1-\alpha)\%$ tal que $0 < \alpha < 1$, pode-se encontrar um valor de $t_{\frac{\alpha}{2},n-1}$, tal que

$$P\left(-t_{\frac{\alpha}{2},n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha, \tag{5.5}$$

onde $t_{\frac{\alpha}{2},n-1}$, é obtido da tabela de distribuição t-Student com $n-1$ graus de liberdade. Logo, o intervalo de confiança para μ, com coeficiente de confiança $100(1-\alpha)\%$ é dado por:

$$IC(\mu; 1-\alpha) = \left(\bar{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}; \bar{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}\right). \tag{5.6}$$

Exemplo 5.3.4 *Suponha que o gerente de produção de uma companhia que fornece petróleo para calefação de uso doméstico, deseja estimar o consumo médio anual(em galões) em casas onde moram somente uma família, numa área geográfica particular. Seleciona-se uma amostra de 36 casas em que moram somente uma família e o consumo médio para essa amostra resultou $\bar{x} = 1.122,7$ galões e um desvio padrão de $s = 295,72$ galões. Se o gerente de produção deseja ter 95% de confiança de que o intervalo obtido inclua o consumo médio anual de petróleo para calefação em casas de famílias que moram nessa área geográfica.*

Solução: Suponha que X : consumo de petróleo para calefação por família é tal que $X \sim N(\mu, \sigma^2)$. Para , $1-\alpha = 0,95$, $\alpha = 0,05$. Da tabela t-Student com 35 graus de liberdade tem-se que $t_{\frac{\alpha}{2},n-1} = t_{0,025,3-1} = 2,03$. Substituindo em (5.6)

$$\begin{aligned} IC(\mu; 0,99) &= \left(1.122,7 - 2,03\frac{295,7}{\sqrt{36}}; 1.122,7 + 2,03\frac{295,72}{\sqrt{36}}\right) \\ &= (1.122,7 - 100,5; 1.122,7 + 100,5) = (1,223,2; 1022,2.) \end{aligned}$$

O intervalo de confiança de 95% estabelece que existe uma seguridade de 95% de que a amostra selecionada é uma na qual a média populacional μ está localizada dentro do intervalo.

5.3.3 Para amostras grandes

A aplicação do Teorema Central do Limite permite a obtenção de intervalos de confiança para μ, quando a distribuição das variáveis aleatórias que constituem a amostra não é dada pelo modelo normal. Nesse caso, os intervalos terão coeficiente de confiança aproximadamente igual a $(1-\alpha) \times 100\%$, sendo que essa aproximação melhora à medida que aumenta o tamanho da amostra.

Exemplo 5.3.5 *Um provedor de acesso à internet está monitorando a duração do tempo das conexões de seus clientes com o objetivo de dimensionar seu equipamento. Suponha que são desconhecidos a média e a distribuição de probabilidade desse tempo, mas a variância, por analogia com outra serviços é considerada como sendo igual a 50 $(minutos)^2$. Uma amostra de 500 conexões resultou num valor observado médio de 25 minutos. O que dizer da verdadeira média com confiança de 95%.*

O Teorema Central do Limite garante que para amostras suficientemente grandes $\bar{X} \sim N(\mu; \sigma^2/n)$. Então o intervalo de confiança aproximado de 95% para o tempo médio de conexões, será dado por

$$\begin{aligned} IC(\mu; 0,95) &= \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \left(25 - 1,96\sqrt{\frac{50}{500}}; 25 + 1,96\sqrt{\frac{50}{500}}\right) \\ &= (25 - 0,62; 25 + 0,62) = (24,38; 25,62). \end{aligned} \tag{5.7}$$

5.4 Intervalo de Confiança para uma Proporção Populacional

Considere uma população dicotômica, constituída apenas por elementos de dois tipos (por exemplo, indivíduos doentes ou não doentes). O valor de p, que corresponde à proporção de elementos de um dos dois tipos na população (por exemplo, indivíduos doentes) é definido como **proporção populacional**. Se dessa população for retirada uma amostra aleatória de tamanho n, então $\hat{p} = Y/n$ será uma proporção amostral sendo Y o número de elementos de um tipo na amostra (por exemplo, número de indivíduos doentes), o que pode ser interpretado como número de sucesso em n ensaios de Bernoulli. Nessas condições a variável aleatória Y segue uma distribuição Binomial com parâmetros n e p.

De acordo com Teorema Central do Limite, para n suficientemente grande, a distribuição de Y (número de elementos de um tipo contidos na amostra) aproxima-se a uma distribuição normal com média np e variância $np(1-p)$. Daí é imediato verificar , que a proporção amostral $\hat{p}$ também aproxima-se da distribuição normal com média p e variância $p(1-p)/n$, ou seja,

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}), \tag{5.8}$$

Fixando o nível de confiança $(1-\alpha) \times 100\%$ tal que $0 < \alpha < 1$, o intervalo de confiança para p, para amostras suficientemente grandes, é dado por:

$$IC(p; 1-\alpha) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right). \tag{5.9}$$

Note que, nesse caso, os limites do intervalo dependem de p que é desconhecido. Assim sendo, o intervalo não pode ser calculado diretamente. Uma possível solução é substituir $p(1-p)$ por $\hat{p}(1-\hat{p})$ em (5.9). Assim o intervalo se reduz a

$$IC(p; 1-\alpha) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \tag{5.10}$$

Uma outra abordagem é baseada no fato de que a expressão $p(1-p)$ assume o valor máximo igual $1/4$ quando $0 \leq p \leq 1$. Como mostra a figura 5.3 abaixo.

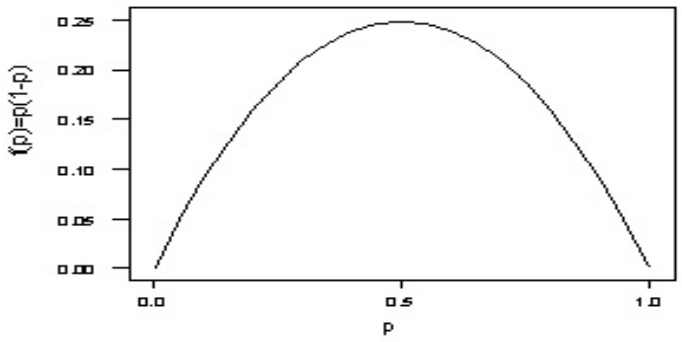


Figura 5.2: Gráfico da função $f(p) = p(1-p)$

Logo, o intervalo (5.9) se reduz a

$$IC(p; 1-\alpha) = \left(\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}} \right) \tag{5.11}$$

Foram apresentados acima duas alternativas para o cálculo do intervalo de confiança para p. A primeira dada em (5.10), é usualmente denominada de abordagem otimista, pois parte da crença que a estimativa obtida está suficientemente próxima de p, de tal forma que a variância $p(1-p)/n$ é bem aproximada por $\hat{p}(1-\hat{p})$. Já a abordagem, calculada em (5.11) é conhecida na literatura como abordagem conservativa, pois substitui-se a variância por um valor seguramente maior do que real. Assim assegura-se que o nível de confiança seja no mínimo $(1-\alpha) \times 100\%$.

Exemplo 5.4.1 *Um estudo foi feito para determinar a proporção de famílias em uma comunidade que tem telefone (p). Uma amostra de 200 famílias é selecionada, ao acaso, e 160 afirmam ter telefone. Que dizer de p com 95% de confiança?*

O Estimador (pontual) para p é dado por $\hat{p} = 160/200 = 0,8$.

Como $1-\alpha = 0,95$, $\alpha = 0,05$, portanto, $z_{0,025} = 1,96$. Logo, substituindo em (5.10), tem-se

$$\begin{aligned} IC_1(p; 0,95) &= \left(0,8 - 1,96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; 0,8 + 1,96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right) \\ &= (0,745; 0,855) \end{aligned}$$

E, em (5.11) tem-se

$$\begin{aligned} IC_2(p; 0,95) &= \left(0,8 - 1,96\sqrt{\frac{1}{4 \times 200}}; 0,8 + 1,96\sqrt{\frac{1}{4 \times 200}} \right) \\ &= (0,731; 0,869) \end{aligned}$$

Pode-se observar que o comprimento do intervalo de confiança otimista é menor que o comprimento do intervalo conservativo.

5.4.1 Determinação do tamanho da amostra para estimação de uma proporção populacional

A determinação do tamanho da amostra quando se quer estimar a proporção populacional é essencialmente a mesma descrita na seção 5.3 para a determinação do tamanho da amostra na estimação de uma média populacional. Para isto, considere o erro máximo de estimação

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Supondo que p e risco α são conhecidos, tem-se:

$$n = \frac{z^2_{\alpha/2}p(1-p)}{E^2}.$$

Se a população é finita e a amostragem é sem reposição

$$E = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}},$$

de onde tem-se:

$$n = \frac{Nz^2_{\alpha/2}p(1-p)}{E^2(N-1) + z^2_{\alpha/2}p(1-p)}$$

Quando não se tem nenhuma informação de p, considera-se $p = 0, .5$ Nesse caso o tamanho da amostra é:

$$n = \frac{0,25z^2_{\alpha/2}}{E^2}.$$

E

$$n = \frac{0,25Nz^2_{\alpha/2}}{E^2(N-1) + 0,25z^2_{\alpha/2}}$$

Se N é muito maior que n o fator de correção de população finita pode ser ignorado.

Exemplo 5.4.2 *O serviço social de um município deseja determinar a proporção de famílias com uma renda familiar inferior a R$ 200,00. Estudos anteriores indicam que esta proporção é de 20%.*

(a) Que tamanho de amostra se requer para assegurar uma confiança de 95% que o erro máximo de estimação desta proporção não ultrapasse o 0,05?

(b) Em quanto variara o tamanho da amostra se o erro máximo permissível é reduzido a 0,01.?

Solução: $p = 0,2$ e $1 - \alpha = 0,95$ da tabela normal padrão $z_{0,025} = 1,96$. Logo,

(a) O erro máximo é $E = 0,05$, então

$$n = \frac{(1,96)^2(0,2)(0,8)}{(0,05)^2} = 245,86 \approx 246.$$

(b) O erro máximo é $E = 0,01$, é então,

$$n = \frac{(1,96)^2(0,2)(0,8)}{(0,01)^2} = 6146,56 \approx 6147.$$

5.5 Intervalo de Confiança para a Variância (σ^2)

Se $X_1, \ldots, X_n$ é uma amostra aleatória de tamanho n, de uma população normal com média μ e variância σ^2, ambas desconhecidas, vimos que a variável aleatória

$$W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

Ou seja, a variável aleatória W tem distribuição Qui-quadrado com n-1 graus de liberdade.

Para um nível de confiança $(1-\alpha)\times 100\%$, é possível determinar $\chi^2_{1-\frac{\alpha}{2},n-1}$ e $\chi^2_{\frac{\alpha}{2},n-1}$, valores da distribuição Qui-quadrado com $n-1$ graus de liberdade, como é mostrado na figura.

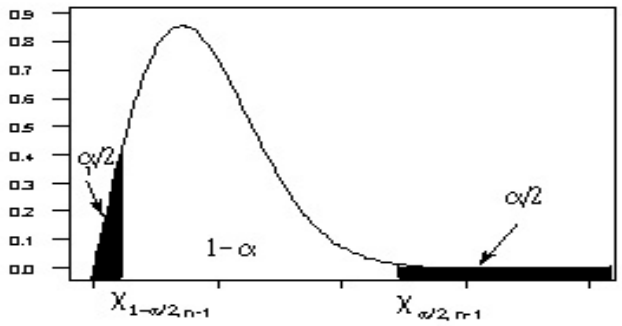


Figura 5.3: Distribuição Qui-quadrado com $n-1$ graus de liberdade

$$P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < W < \chi^2_{\frac{\alpha}{2},n-1}\right) = P\left(\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

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

$$IC(\sigma^2; 1-\alpha) = \left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}; \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right).$$

Exemplo 5.5.1 *Pretende-se avaliar a variabilidade associada ao resultado de um determinado método de análise química. Com esse objetivo, efetuaram-se 24 análises a uma determinada substância em que se segui o referido método, em condições perfeitamente estabilizadas. A variância amostral dos resultados (expressados numa determinada unidade) foi de 4,58. Admitindo que o resultado das análises segue uma distribuição normal. Um intervalo de confiança do* 90% *de confiança para variância, é dado por:*

$$IC(\sigma^2; 0.90) = \left(\frac{(24-1)4,58}{35,17}; \frac{(24-1)4,58}{13,09}\right) = (2,995; 8,047).$$

5.6 Intervalo de Confiança para a Diferença de Médias ($\mu_1 - \mu_2$)

Nesta seção considere que $X_1, \ldots, X_n$ é uma amostra aleatória de tamanho n de uma população com característica X que tem distribuição normal com média μ_1 e variância σ_1^2 e que $Y_1, \ldots, Y_m$ é outra amostra aleatória de tamanho m, de uma população com a característica Y que tem distribuição normal com média μ_2 e variância σ_2^2. Se X e Y são independentes foi apresentado distribuições amostrais para a diferença das médias amostrais, no caso quando as variâncias populacionais eram conhecidas e quando não são conhecidos mais iguais.

5.6.1 Quando as variâncias σ_1^2 e σ_2^2 são conhecidos

Foi visto que a variável

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

tem distribuição normal padrão. Considerando este resultado e seguindo o mesmo procedimento para o caso da média populacional, apresentada na seção 5.3, pode-se deduzir o intervalo de confiança para $\mu_1 - \mu_2$, para um nível de confiança $(1-\alpha)\times 100\%$ fixado. Ou seja,

$$P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right) = 1 - \alpha.$$

Logo, o intervalo de $(1-\alpha)\times 100\%$ de confiança para $\mu_1 - \mu_2$ é dado por:

$$IC(\mu_1 - \mu_2; 1-\alpha) = \left(\bar{X} - \bar{Y} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}; \bar{X} - \bar{Y} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right) \tag{5.12}$$

Exemplo 5.6.1 *Em um estudo em crianças com retardo mental, a 11 meninas e a 10 meninos, após um ano de educação especial acompanhado de terapia, foi aplicado um teste de conhecimentos. A média para meninas foi de 67,0 e para as meninos foi de 61,5 (em uma escala de 0 a 100). Supondo que as qualificações obtidas pelas meninas e meninos em estudo seguem uma distribuição normal com desvio padrão $\sigma_1 = 11$ e $\sigma_2 = 10$. Achar um intervalo de 90% de confiança para $\mu_1 - \mu_2$.*

Solução: Para o nível de confiança $1 - \alpha = 0,90$ temos que $\alpha = 0,10$. Obtemos da distribuição normal padrão o valor $z_{\alpha/2} = 1,64$, $\bar{X} = 67,0$, $n = 11$, $\bar{Y} = 61,5$ e $m = 10$. Substituindo em (5.12) o intervalo para $\mu_1 - \mu_2$ é dado por

$$\begin{aligned} IC(\mu_1 - \mu_2; 0,90)) &= \left(67,0 - 61,5 - 1,64\sqrt{\frac{121}{11} + \frac{100}{10}}; 67 - 61,5 + 1,64\sqrt{\frac{121}{11} + \frac{100}{10}}\right) \\ &= (-2,038; 13,038). \end{aligned}$$

5.6.2 Quando $\sigma_1^2 = \sigma_2^2 = \sigma^2$, mas desconhecidos

Mostrou-se que a variável T, definida por:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{n})}},$$

segue uma distribuição de t-student com $n + m - 2$ graus de liberdade, onde $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$ é conhecida com a variância ponderada. Neste caso o intervalo de confiança para $\mu_1 - \mu_2$, com0 um nível de confiança $(1-\alpha)$ é dado por:

$$IC(\mu_1 - \mu_2; 1-\alpha) = \left(\bar{X} - \bar{Y} - t_{\alpha/2,n+m-2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}; \bar{X} - \bar{Y} + t_{\alpha/2,n+m-2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}\right) \tag{5.13}$$

Exemplo 5.6.2 *O gerente de um banco está interessado em analisar a diferenças entre os saldos médios das contas à ordem de duas agências. De cada uma delas foi recolhida uma amostra aleatória de saldos (milhões de unidades monetárias), tendo-se registrado os seguintes resultados:*

Agência	*n*	*Média*	*Variância*
A	*10*	*17,8*	*30,3*
B	*13*	*14,2*	*28,7*

Supondo que saldos das agências tenha distribuição normal com variâncias iguais, mas desconhecidas. Determine um intervalo de 95% de confiança para $\mu_1 - \mu_2$.

Solução: Do enunciado do exemplo tem-se: $n = 10$, $\bar{X} = 17,8$, $S_1^2 = 30,7$, $m = 13$, $\bar{Y} = 14,2$, $S_2^2 = 28,7$ portanto a variância ponderada é, $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n+m-2} = \frac{(10-1)30,7+(13-1)28,7}{10+13-2} = 29,39$. Como $1-\alpha = 0,95$, $t_{0,025,21} = 2,08$. Logo, substituindo (5.13) temos um intervalo de 95% de confiança para $\mu_1 - \mu_2$ é dado por:

$$\begin{aligned} IC(\mu_1-\mu_2;0,95) &= \left(17,8-14,2-2,08\sqrt{29,39(\frac{1}{10}+\frac{1}{13})}\right. \\ &\quad ; \quad \left. 17,8-14,2+2,08\sqrt{29,39(\frac{1}{10}+\frac{1}{13})}\right) \\ &= (-1,14;8,34) \end{aligned}$$

5.6.3 Quando as variâncias são desconhecidas e diferentes

No caso em que as variâncias populacionais não são conhecidas e diferentes ($\sigma_1^2 \neq \sigma_2^2$) pode-se mostrar que a variável aleatória

$$T' = \frac{\bar{X}-\bar{Y}-(\mu_1-\mu_1)}{\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}} \sim t(\nu)$$

onde $\nu = \frac{\left(\frac{S_1^2}{n}+\frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n+1}+\frac{\left(\frac{S_2^2}{m}\right)^2}{m+1}} - 2$. Ou seja que T' tem distribuição t-Student com ν graus de liberdade.

Neste caso o intervalo de $(1-\alpha)\times 100\%$ de confiança para $\mu_1 - \mu_2$ é dado por:

$$IC(\mu_1-\mu_2;1-\alpha) = \left(\bar{X}-\bar{Y}-t'_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}};\bar{X}-\bar{Y}+t'_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n}+\frac{S_2^2}{m}}\right) \tag{5.14}$$

Exemplo 5.6.3 *Um artigo publicado no Food Technology Journal (1956) descreve um estudo sobre conteúdo de protopectina em tomates durante o armazenamento. Para o qual foram considerados dois períodos de armazenamento e analisou-se as amostras de nove lotes de tomates em cada período. Os dados sumariados apresentam-se a continuação:*

Tempo de armazenamento	*Média*	*Desvio Padrão*
7 Dias	*792*	*495,0*
21 Dias	*372,3*	*73,3*

Considerando que o conteúdo de propectina para os tempo de armazenamento tenha distribuição normal e que as variâncias verdadeiras são diferentes construa um intervalo de confiança do 95%, para diferença de médias entre o tempo de armazenamento de 7 dias e 21 dias.

Da tabela t-Student com $\nu = \frac{\left(\frac{S_1^2}{n}+\frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n+1}+\frac{\left(\frac{S_2^2}{m}\right)^2}{m+1}} - 2 = \frac{\left(495^2/9+73^2/9\right)^2}{\frac{(495^2/9)^2}{9+1}+\frac{(73^2/9)^2}{9+1}} - 2 \approx 8,0395 = 8$ graus de liberdade e nível de confiança $1-\alpha = 0,95$ obtém-se que $t'_{0,025,8} = 2,306$. Logo, substituindo em (5.14) o intervalo é calculado, ou seja:

$$\begin{aligned} IC(\mu_1-\mu_2,0,95) &= \left(729-3172-2,306\sqrt{\frac{495^2}{9}+\frac{73^3}{9}};729-3172+2,306\sqrt{\frac{495^2}{9}+\frac{73^3}{9}}\right) \\ &= (48,06;791,34). \end{aligned}$$

5.7 Intervalo de Confiança para Razão de Variâncias

Seja $X_1, \ldots, X_n$ uma amostra de tamanho n retirada de uma população com a característica X, que tem distribuição normal com μ_1 (desconhecida) e variância, σ_1^2. Considere $Y_1, \ldots, Y_m$ outra amostra de tamanho m de outra população com a característica Y, com distribuição normal μ_2 (desconhecida) e variância σ_2^2 e se X e Y são independentes, foi visto que a variável aleatória definida

$$F = \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \sim F_{(n-1;m-1)},$$

ou seja, que variável aleatória F tem distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdade, sendo S_1^2 e S_2^2 as variâncias amostrais calculadas com as n e as m amostras da população X e população Y, respectivamente.

Para um nível de confiança $(1-\alpha) \times 100$ fixado temos que

$$P\left(f_1 \leq F \leq f_2\right) = 1 - \alpha$$

ou seja,

$$P\left(f_1 \leq \frac{S_1^2}{S_2^2} \times \frac{\sigma_2^2}{\sigma_1^2} \leq f_2\right) = 1 - \alpha$$

Portanto, o intervalo de $(1-\alpha) \times 100\%$ de confiança para $\frac{\sigma_2^2}{\sigma_1^2}$ é dado por :

$$IC(\frac{\sigma_2^2}{\sigma_1^2}; 1-\alpha) = \left(f_1 \frac{S_2^2}{S_1^2} \; ; \; f_2 \frac{S_2^2}{S_1^2}\right). \tag{5.15}$$

onde f_1 e f_2 são valores da distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdade mostradas na figura 5.4, sendo $f_1 = \frac{1}{f_{\alpha/2,m-1,n-1}}$ e $f_2 = f_{\alpha/2,n-1,m-1}$.

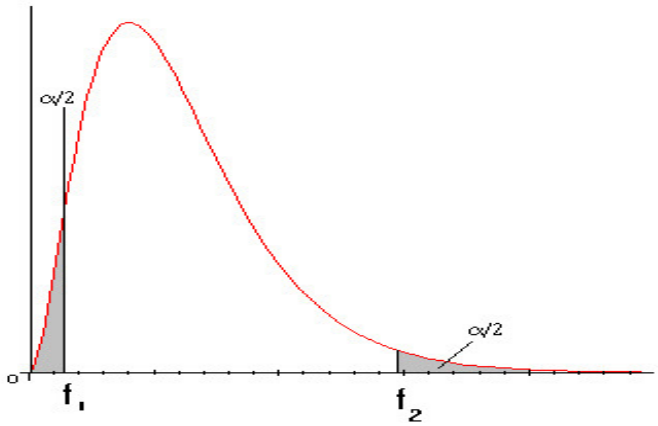


Figura 5.4: Distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdade

Exemplo 5.7.1 *Dois catalisadores podem ser usados em um processo químico em bateladas. Oito bateladas foram preparadas usando o catalaisador 1, resultado em rendimento médio de 86 e uma variância de 46,5. Dezessete batelados foram preparados com o catalisador 2, resultando um rendimento médio de 90 e uma variância de 23,4. Considerando que as medidas dos rendimentos sejam distribuidas aproximadamente normal. Determinar um intervalo do 90% de confiança para razão de variâncias dos rendimentos do catalisador 1 e o catalisador 2.*

Solução: Do enunciado temos que $n = 8$, $S_1^2 = 46,5$, $m = 17$, $S_2^2 = 23,4$ e da tabela F-Snedecor obtemos que $f_1 = \frac{1}{f_{0,05,7,16}} = 1/2,61 = 0,376$ e $f_2 = f_{0,05,16,7} = 3,49$. Substituindo essas quantidade em (5.15) temos que um

intervalo de 90% de confiança para a razão de variâncias, $\frac{\sigma_1^2}{\sigma_2^2}$:

$$IC(\frac{\sigma_1^2}{\sigma_2^2}; 0,90) = \left(0,376 \times \frac{46,5}{23,4} \; ; \; 3,49 \times \frac{46,5}{23,4}\right) = (0,7478; 6,935).$$

5.8 Teste de Hipóteses

5.8.1 Conceitos básicos

O teste de uma hipótese estatística é talvez a área mais importante da teória de decisão. Vamos introduzir os conceitos de teste de hipótese estatística através do exemplo seguinte.

Exemplo 5.8.1 *Considere que uma industria compra de um certo fabricante, pinos cuja resistência média à ruptura é especificada em 60 kgf (valor nominal da especificação). Em um determinado dia, a indústria recebeu um grande lote de pinos e a equipe técnica da industria deseja verificar se o lote atende as especifições.*

É claro que equipe técnica não espera que todos os pinos tenham examente uma resistência de 60 kgf. Alguma variabilidade em torno deste valor é esperada. A partir de experiência anterior a indústria sabe que a resistência à ruptura dos pinos desse fabricante segue uma distribuição normal com desvio padrão $\sigma = 5kgf$ e esta variabilidade é adequada para a industria. O interesse da industria consiste, então, em determinar se a resistência média dos pinos que constituem o lote entregue pelo fabricante pode ser ou não considerado igual a 60 kgf.

Do dito anteriormente considere que a resistência dos pinos do lote é uma variável aleatória X, tal que , $X \sim N(\mu, 25)$. Observe que equipe técnica da industria deseja testar:

$$H_0 : \mu = 60 \tag{5.16}$$

A seguir é apresentada a definição formal de hipótese estatística.

Definição 5.8.1 *Uma hipótese estatística é uma afirmação sobre os parametros de uma ou mais caracteristícas da população*

Em todo problema de teste de hipóteses, duas hipóteses complementares são consideradas. A hipótese que foi destacada na equação (5.16) denominada de **hipótese nula,** sendo representanda por H_0, (pois ela expressa que não há mudança). A outra hipótese, que será aceita caso H_0 seja rejeitada, é denominada **hipótese alternativa** e é denotada por H_1. Tem-se

$$\text{Rejeitar } H_0 \Rightarrow \text{Acietar } H_1$$

$$\text{Aceitar } H_0 \Rightarrow \text{Rejeitar } H_1$$

No exemplo, a hipótese alternativa H_1 é

$$H_1 : \mu \neq 60 \tag{5.17}$$

Essa hipótese é chamada de hipótese composta porque especifica mais de um valor para o parâmetro. No caso que especifique somente um único valor, a hipótese é chamada de hipótese simples, por exemplo a hipótese dada em (5.16).

Para realizar-se um **teste de uma hipótese estatística** retira-se uma amostra da população em estudo e com base na observação dos resultados dessa amostra toma-se a decisão de aceitar H_0 ou de rejeitar H_0.

Suponha que a equipe técnica da indústria tenha decidido retirar uma amostra aleatória de tamanho $n = 16$, do lote recebido, medir a resistência de cada pino e calcular a resistência média $\bar{X}$ (estimador de μ). Além disso, $\bar{X} \sim N(\mu, \frac{25}{16})$. Para quais valores de $\bar{X}$ a equipe técnica deve rejeitar H_0 e portanto não aceitar o lote?

Definição 5.8.2 *A variável aleatória cujo valor é utilizado para determinação da decisão a ser tomada em um teste de hipóteses é denominada estatística de teste*

Se o lote está fora de especificação , isto é , $H_1 : \mu \neq 60$, espera-se que $\bar{X}$ seja inferior ou superior a 60 kgf.

Suponha que equipe técnica tenha decidido adotar a seguinte regra: rejeitar H_0 se $\bar{X}$ for maior que 62.5 kgf e ou menor que 57.5 kgf. O conjunto $R_c = \{\bar{X} < 57,5 \,\text{ou}\, \bar{X} > 62.5\}$ é o conjunto de valores para os quais rejeita-se $H_0 : \mu = 0,5$, sendo denominado **região de rejeição ou região crítica** do teste. Os valores de $\bar{X}$ que não

pertencem ao intervalo [57,5 ; 62,5], constituem a **região de aceitação ($R_a = R_c^c$).** Os valores que estão na fronteiras entre a região crítica e a região de aceitação, são denominados **valores críticos.** Portanto, a regra consiste em, rejeitar H_0 a favor de H_1 se o valor assumido pela estatística de teste pertencer a região crítica. Isto é, se ocorrer o evento $(\bar{X} \in R_c)$, rejeita-se H_0. Caso contrário, se o valor assumido por $\bar{X}$ pertencer a região de aceitação R_c^c, isto é, se o evento $(\bar{X} \in R_c^c)$, ocorrer não rejeitar H_0.

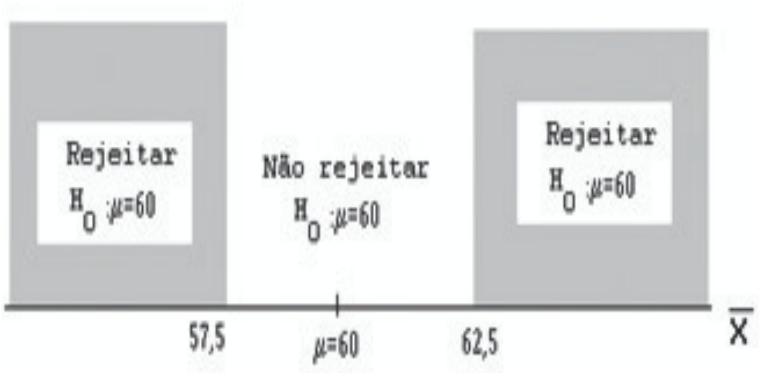


Figura 5.5: Regra de decisão para testar $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$

O procedimento de tomada de decisão em um teste de hipóteses pode resultar em dois tipos de conclusões incorretas. Por exemplo, é possível que a resistência média dos pinos que constituem o lote seja, de fato, igual a 60 kgf. Mas, pode acontecer que para os pinos selecionados para a composição da amostra aleatória, o valor observado para a estatística de $\bar{X}$ pertence a região crítica. Neste caso a hipótese nula H_0 seria rejeitada em favor da hipótese alternativa H_1, quando H_0 é de fato verdadeiro. Essa forma de conclusão incorreta é denominada de **erro tipo I.**

Por outro lado, poderia ocorrer situações na qual a hipótese H_0 é falsa, ou seja, na realidade a resistência média do lote de pinos é diferente de 60 kgf e a média amostral observada $\bar{x}$ pertença a região de aceitação, levando a aceitação de H_0 sendo ela falsa. Esta forma de conclusão incorreta é denominada de **erro tipo II.** Em resumo, em um teste de hipótese, podem ocorrer dois tipos de erros:

- Erro tipo I: rejeitar H_0 sendo H_0 verdadeira;
- Erro tipo II: Aceitar H_0 sendo H_0 falsa.

Portanto, ao testar qualquer hipótese estatística, existem quatro situações diferentes que determinam se a decisão final é correta ou incorreta. Essas situações aparecem na tabela 5.1.

Tabela 5.1: Decisões em um teste de hipóteses.

	Decisão real e desconhecida	
Decisão	H_0 verdadeira	H_0 falsa
Não rejeitar H_0	Decisão correta	Erro tipo II
Rejeita H_0	Erro tipo I	Decisão correta

Dado que a decisão tomada em um teste de hipóteses é baseada em variáveis aleatórias (estatística de teste), é possível calcular as probabilidades dos erros tipos I e II da tabela 5.1.

A probabilidade de erro tipo I é denominada de *nível de significância* do teste será denotada por α. Isto é,

$$\alpha = P(\text{Erro tipo I}) = P(\text{rejeitar } H_0 | \ H_0 \text{ é verdadeiro})$$

No exemplo 5.8.1, o erro tipo I irá ocorrer se $\bar{X} < 57,5$ ou $\bar{X} > 62,5$ quando a resistência média no lote de pinos for $\mu = 60$ kgf. Para este exemplo, observe que, se H_0 é verdadeira, isto é, $H_0 : \mu = 60$ então,$\bar{X}$ tem distribuição normal com média $\mu = 60$ e $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 1,25$. Portanto, a probabilidade do erro tipo I é calculada como:

$$\begin{aligned}\alpha &= P(\bar{X} < 57,5 \text{ ou } \bar{X} > 62,5|H_0 : \mu = 60) = P(\bar{X} < 57,5) + P(\bar{X} > 62,5|H_0 : \mu = 60)\\ &= P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{57,5-60}{1,25}) + P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{62,5-60}{1,25})\\ &= P(Z < -2) + P(Z > 2) = 0,02275 + 0,02275 = 0,0455.\end{aligned}$$

Este resultado, que está ilustrado na figura 5.6, significa que há 4,55% de chance que uma amostra aleatória extraida do lote de peças de pinos leve à rejeição da hipóteses nula $H_0 : \mu = 60$, quando a verdadeira resistência média dos pinos é, de fato, igual a 60 kgf.

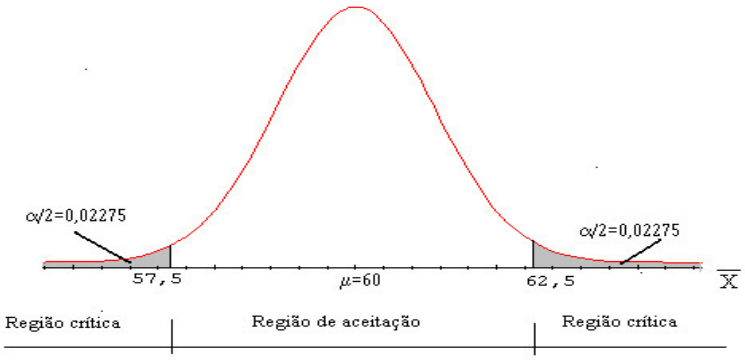


Figura 5.6: Região crítica e nível de significância para o teste de $H_0 : \mu = 60$ contra $H_1 : \alpha \neq 60$ com $n = 16$

Ao analisar a figura 5.6, pode-se observar que é possível diminuir α ao aumentar a amplitude da região de aceitação. Por exemplo, se no caso dos pinos, a região de aceitação fosse constituída pelo intervalo [56 ; 64], o valor de α será:

$$\begin{aligned}\alpha &= P(\bar{X} < 56) + P(\bar{X} > 64|H_0 : \mu = 60)\\ &= P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{56-60}{1,25}) + P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{64-60}{1,25})\\ &= P(Z < -3,2) + P(Z > 3,2) = 0,00069 + 0,00069 = 0,00138.\end{aligned}$$

Pode-se também diminuir o valor de α aumentando o tamanho da amostra. Se $n = 25$, a variância de $\bar{X}$ é $\sigma/\sqrt{n} = 5/\sqrt{25} = 1$. Ao utilizar a região crítica original da figura 5.6, tem-se:

$$\begin{aligned}\alpha &= P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{57,5-60}{1,0}) + P(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} < \frac{62,5-60}{1,0})\\ &= P(Z < -2,5) + P(Z > 2,5) = 0,00621 + 0,00621 = 0,01242.\end{aligned}$$

Ao avaliar um procedimento de teste de hipóteses é importante determinar a probabilidade de erro tipo II, o qual denota-se por β. Isto é,

$$\beta = P(\text{Erro tipo II}) = P(\text{aceitar } H_0|\ H_0 \text{ é falso})$$

Para o exemplo 5.8.1, o erro tipo II irá ocorrer se $57,5 \leq \bar{X} \leq 62,5$ quando a resistência média do lote é diferente de 60 kgf. Portanto, para que seja possível calcular o valor de β, deve-se considerar um valor particular para μ sob a hipótese alternativa. Como exemplo, suponha que é muito importante para a indústria rejeitar a hipótese nula $H_0 : \mu = 60$, quando a resistência dos pinos do lote μ for, igual a 56,5 kgf ou igual a 63,5 kgf. Nessa situação, para

verificar se o teste é de fato adequado, a industria poderia calcular o valor de β para $\mu = 56,5$ e $\mu = 63,5$ e então avaliar se esse valor é suficientemente baixo.

O cálculo de β para $\mu = 63,5$. Nesse caso, $\bar{X} \sim N(63,5, \frac{25}{16})$. Portanto, a probabilidade de erro tipo II é calculada como:

$$\beta = P(\text{Erro tipo II}) = P(57,5 \leq \bar{X} \leq 62,5 | H_1 : \mu = 63,5)$$

Os valores críticos 57, 5 e 62,5 padronizados com $\mu = 63,5$ são:

$$\begin{aligned} z_1 &= \frac{57,5 - 63,5}{1,25} = -4,80 \text{ e} \\ z_2 &= \frac{62,5 - 63,5}{1,25} = -0,80 \end{aligned}$$

Logo,

$$\begin{aligned} \beta &= P(57,5 \leq \bar{X} \leq 62,5 | H_1 : 63,5) = P(Z \leq -0,80) - P(Z \leq -4,80) \\ &= 0,21186 - 0,00 = 0,21186 \end{aligned}$$

Esse resultado, que está ilustrado na figura 5.7, significa que para o teste de $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$, com base na amostra de tamanho $n = 16$, quando o valor verdadeiro da resistência média dos pinos é $\mu = 63,5$, a probabilidade de que a hipóteses nula (que neste caso é falsa) não seja rejeitada é igual a 21,186%. Devido à simetria da distribuição normal, quando a verdadeira média é $\mu = 56,5$, a probabilidade de erro tipo II é igual 21,186%.

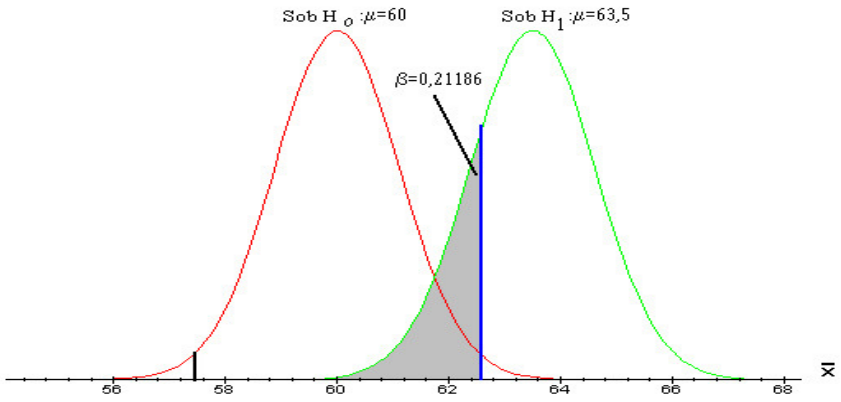


Figura 5.7: Probabilidade do erro tipo II (β) para o teste de $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$ com $n = 16$ e $\mu = 63,5$

A probabilidade de cometer erro tipo II aumenta rapidamente à medida que o valor verdadeiro de μ se aproxima do valor estabelecido sob a hipótese H_0. Para ilustrar essa afirmação, calcula-se o valor de β para o exemplo 5.8.1, no caso que o valor verdadeiro da resistência média dos pinos é $\mu = 61$ e que o teste de $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$ é conduzido baseando-se em uma amostra de tamanho $n = 16$, ou seja,

$$\begin{aligned} \beta &= P(57,5 \leq \bar{X} \leq 62,5 | H_1 : \mu = 61) \\ &= P(\bar{X} < 56) + P(\bar{X} > 64) = P(Z \leq 1,20) - P(Z \leq -2,80) \\ &= 0,88493 - 0,00256 = 0,88237. \end{aligned}$$

Esse resultado, que esta ilustrado na figura 5.8, significa que, para o teste de $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$, com base em amostras de tamanho $n = 16$, quando o valor verdadeiro da resistência média é igual a 61kgf, há 88,237% de chance que hipótese nula (que é falsa) não seja rejeitada.

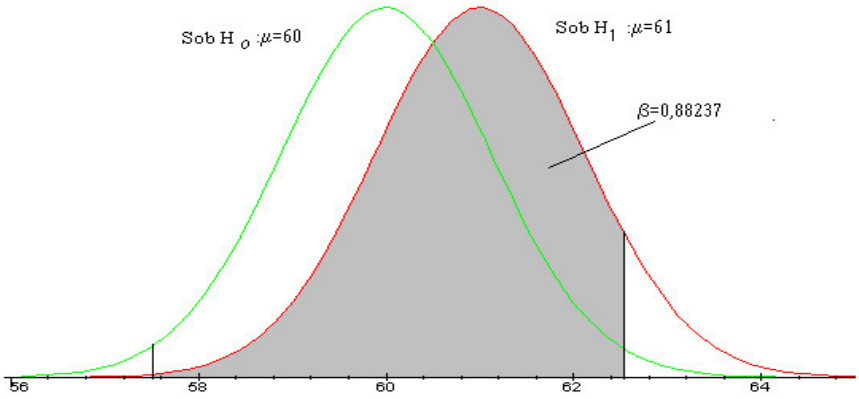


Figura 5.8: Probabilidade do erro tipo II(β) para o teste de: $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$ com $n = 16$ e $\mu = 61$

Portanto, a probabilidade de erro tipo II é muito maior para o caso em que a média verdadeira é $\mu = 61$ do que para a situação em que $\mu = 63,5$ kgf. No entanto, esse tipo de resultado não causa muita preocupação. Isso porque apenas diferenças de maior magnitude entre o valor verdadeiro de μ e o valor estabelecido sob H_0 são consideradas significativas sob o ponto de vista prático, devendo então ser detectadas com elevada probabilidade.

A probabilidade do erro tipo II também depende do tamanho da amostra (n). Para ilustrar este fato, refaz-se o cálculo de β, para exemplo 5.8.1, considerando que a hipótese nula é $H_0 : \mu = 60$ e a verdadeira média é $\mu = 63,5$ e que o tamanho da amostra aumenta de $n = 16$ para $n = 25$.

$$\beta = P(57,5 \leq \bar{X} \leq 62,5 | H_1 : \mu = 63,5)$$

Quando $n = 25$, $\bar{X} \sim N(63,5, \frac{25}{25})$ e os valores críticos de 57,5 e 62,5 padronizados são:

$$z_1 = \frac{57,5 - 63,5}{1} = -6 \text{ e}$$
$$z_2 = \frac{62,5 - 63,5}{1} = -1.$$

Logo,

$$\begin{aligned} \beta &= P(-6 \leq Z \leq -1) \\ &= P(Z \leq -1) - P(Z \leq -6) = 0,15866 - 0,0000 = 0,15866. \end{aligned}$$

Esse resultado é ilustrado na figura 5.9. Observa-se que o aumento do tamanho da amostra resulta em uma diminuição da probabilidade do erro tipo II.

A tabela 5.2 sumariza os resultados apresentados anteriormente conjuntamente com outros resultados obtidos de forma similar:

Tabela 5.2: Relacionamento entre n, α, β e região de aceitação para o exemplo 5.8.1.

Região de aceitação	Tamanho da amostra	α	β para $\mu = 61$	β para $\mu = 63,3$
$57,5 \leq \bar{X} \leq 62,5$	16	0,0455	0,88237	0,21186
$56,0 \leq \bar{X} \leq 64,0$	16	0,00138	0,99886	0,18843
$57,5 \leq \bar{X} \leq 62,5$	25	0,01242	0,93312	0,15866
$56,0 \leq \bar{X} \leq 64,0$	25	0,00003	0,99862	0,30209

A tabela 5.2 mostra as seguintes características dos testes de hipóteses:

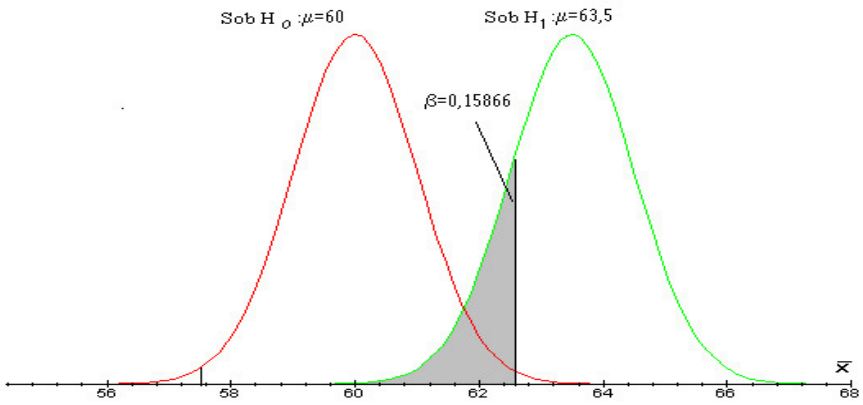


Figura 5.9: Probabilidade do erro tipo II(β) para o teste de: $H_0 : \mu = 60$ contra $H_1 : \mu \neq 60$ com $n = 25$ e $\mu = 63,5$

(i) Os erros tipo I e II estão relacionados. Se o tamanho de amostra permanece constante, uma diminuição da probabilidade de ocorrência de um dos erros implica em um aumento da probabilidade da ocorrência do outro erro.

(ii) A probabilidade de ocorrência do erro tipo I pode ser reduzida por meio de uma escolha apropriada da região crítica.

(iv) O valor de β aumenta à medida que valor verdadeiro de μ se aproxima do valor estabelecido sob a hipótese H_0.

(iv) Em geral, um aumento no tamanho da amostra reduz tanto α quanto β, desde que os valores críticos sejam mantidos constantes.

O ideal seria minimizar tantos o erros do tipo I quanto os do tipo II. Mas, infelizmente, para qualquer tamanho de amostra dado, não é possível minimizar ambos erros simultaneamente. A abordagem clássica deste problema considera que o erro tipo I é provavelmente ser o mais sério que o erro tipo II. Para tenta-se manter a probabilidade de cometer erro tipo I em um nível razoavelmente baixo, como 0,01, 0,05 ou 0,10 e então minimizar quanto possível a probabilidade do erro tipo II.

Definição 5.8.3 *O poder de um teste de hipóteses é a probabilidade de rejeitar H_0 quando H_0 é falsa.*

$$\begin{aligned} Poder &= P(\text{Rejeitar } H_0 | H_0 \text{ falsa}) \\ &= 1 - P(\text{Não rejeitar } H_0 | H_0 \text{ falsa}) = 1 - \beta \end{aligned}$$

O **poder de um teste de hipóteses** pode ser interpretado como a probabilidade de *rejeitar de maneira correta uma hipótese nula falsa,* o que representa a decisão correta. Em muitos casos, dois diferentes testes de hipóteses são comparados por meio de comparação do poder de cada um deles. Considere o exemplo 5.8.1, onde se testam as hipóteses

$$\begin{aligned} H_0 &: \mu = 60, \\ H_1 &: \mu \neq 60 \end{aligned}$$

onde μ é a resistência média dos pinos do lote. Suponha que o valor verdadeiro da média é $\mu = 63,5$. Para o tamanho da amostra $n = 16$, com região de aceitação $57,5 \leq \bar{X} \leq 62,5$ foi vista que $\beta = 0,21186$ (veja tabela 5.2). Logo, o poder do teste correspondente é:

$$Poder = 1 - \beta = 1 - 0,21186 = 0,78814$$

Já o poder do teste para $n = 25$, para a mesma região de aceitação é igual.

$$Poder = 1 - \beta = 1 - 0,15866 = 0,84135.$$

O poder do teste é uma medida capacidade do teste para detectar uma possível diferença existente entre o valor estabelecido para o parâmetro sob a hipótese H_0 e o valor assumido pelo parâmetro. Observe que o primeiro teste tem poder igual a 0,78814, para detectar a diferença entre resistência igual 60 kgf e a outra de 63,6 kgf estabelecida pela hipótese alternativa. Isso significa que, se a verdadeira resistência média dos pinos é 63,5 kgf, esse teste rejeitará de maneira correta $H_0 : \mu = 60$ e detectará essa diferença em 78,814% das vezes que for utilizado. O poder do segundo teste é um pouco maior (0,84135), como já era de se esperar, porque o tamanho da amostra é maior que aquele utilizado no primeiro. O poder de um teste pode ser aumentado por meio do aumento de n ou do aumento do nível de significância α.

5.8.2 Testes unilaterais e bilaterais

Se a hipótese nula e alternativa de um teste de hipóteses são:

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

onde μ_0 é uma constante conhecida, o teste é chamada de **teste bilateral**, pois é importante detectar diferenças a partir do valor hipotético da média μ_0 que se encontre em qualquer lado de μ_0 . Em um teste desse tipo a região crítica é dividida em duas partes, com a mesma probabilidade em cada cauda da distribuição da estatística de teste. O teste considerado no exemplo 5.8.1 é um teste bilateral . Em muitos problemas tem-se interesse em testar hipóteses do tipo:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0. \end{aligned}$$

Neste caso tem-se um teste **unilateral esquerdo**, porque a região de rejeição não é dividida em duas partes, ficando localizada apenas na cauda esquerda da distribuição da estatística de teste. Para exemplificar, considere o seguinte problema.

Exemplo 5.8.2 *Uma região do país é conhecida por ter uma população obesa. A distribuição de probabilidade do peso dos homens dessa região entre 20 e 30 anos é normal com média de 90 kg e desvio padrão de 10 kg. Um endocrinologista propõe um tratamento para combater a obesidade que consiste de exercícios físicos, dietas e ingestão de um medicamento. Ele afirma que com seu tratamento o peso médio da população da faixa em estudo diminuirá num período de três meses.*

Neste caso as hipóteses que deverão ser testados são:

$$\begin{aligned} H_0 &: \mu = 90\,kg \\ H_1 &: \mu < 90\,kg \end{aligned}$$

sendo μ a média dos pesos dos homens da faixa etária em estudo.

Em muitas situações, tem-se interesse em provar que a média populacional μ é maior do que valor de μ_0. Assim, tem-se um teste **unilateral direito**, para o qual as hipóteses assumem a forma:

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

No teste **unilateral direito** a região crítica fica localizada na cauda direita da estatística de teste. Para uma situação onde seria apropriado realizar um teste unilateral direito, considere o seguinte exemplo

Exemplo 5.8.3 *Um fabricante de uma certa peça afirma que o tempo médio de vida das peças produzidas é de 1000 horas. Suponha que os engenheiros de produção têm interesse em verificar se a modificação do processo de fabricação aumenta a duração das peças.*

Nesse caso as hipóteses que deverão ser testados são:

$$\begin{aligned} H_0 &: \mu = 1000\, horas \\ H_1 &: \mu > 1000\, horas \end{aligned}$$

sendo μ é o tempo médio de vida das peças produzidas pelo novo processo.

5.8.3 Procedimento básico de teste de hipóteses

O procedimento básico de teste de hipóteses relativo ao parâmetro θ de uma população, será decomposto em 4 passos:

(i) Definição das hipóteses:

$$\begin{aligned} H_0 &: \theta = \theta_0, \\ H_1 &: \theta < \theta_0 \;\; ou \;\; \theta > \theta_0 \;\; ou \;\; \theta \neq \theta_0 \text{ (qualquer alternativa)} \end{aligned}$$

(ii) Identificação da estatística do teste e caracterização da sua distribuição.

(iii) Definição da regra de decisão, com a especificação do nível de significância do teste.

(iv) Cálculo da estatística de teste e tomada de decisão.

Nas seguintes seções serão apresentados procedimentos básicos de teste de hipóteses para uma média populacional, diferenças de duas médias populacionais, variância populacional, igualdade de variâncias populacionaias, uma proporção populacional e a diferença de duas proporções populacionais.

5.9 Teste de Hipóteses para uma Média Populacional

Considere uma amostra aleatória de tamanho n de uma população normal com média μ (desconhecida) e variância σ^2. Inicialmente, considera-se o caso do teste unilateral esquerdo, para de imediato generalizar o procedimento. Suponha que tem-se interesse em verificar as seguintes hipóteses:

(i)

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

onde μ_0 é um valor numérico da média populacional.

(ii) A estatística do teste é a média amostral $\bar{X}$. Se população é normal (ou se amostra é grande $n \geq 30$, mesmo que a população não é normal) a distribuição de $\bar{X}$ é $N(\mu, \sigma^2/n)$ e a variável aleatória

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

(iii) É razoável, rejeitar H_0 em favor de H_1, se a média amostral $\bar{X}$ é demasiado pequena em relação μ_0. A região crítica, então poderia ser obtido, selecionando um k da média amostral, de maneira que $R_c = \{\bar{X} \leq k\}$ onde k é tal que $P(\bar{X} \leq k | H_0 : \mu = \mu_0) = \alpha$. Ou

$$P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) = P(Z \leq \frac{k - \mu_0}{\sigma/\sqrt{n}}) = \alpha$$

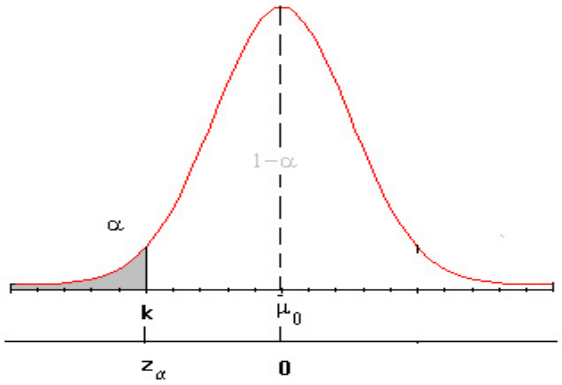


Figura 5.10: Região crítica para teste de hipóteses unilateral de uma média.

Da tabela normal padrão obtém-se z_α para um nível de significância α fixado (veja a figura 5.10)
Tem-se, $\frac{k-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$. Daí $k = \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}$. Logo, $R_c = \{\bar{X} \leq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}\}$.
(iv) Conclusão: se $\bar{x} \in R_c = \{\bar{X} \leq \mu_0 + \frac{z_\alpha \sigma}{\sqrt{n}}\}$, rejeita-se H_0, em caso contrário não se rejeita H_0.

Método alternativo

Um método alternativo prático é trabalhar diretamente na escala Z (veja a figura 5.10) de seguinte forma:
(i)

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

(ii) A estatística do teste é

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

se a hipótese nula é verdadeira $Z \sim N(0,1)$.
(iii) A região crítica, para um nível de significância α fixado é: $R_c = \{z \in Z \sim N(0,1); Z \leq z_\alpha\}$.
(iv) Calcula-se o valor da estatística do teste, Z_{obs} de acordo os dados amostrais e compara-se se Z_{obs} com z_α. Se $Z_{obs} \leq z_\alpha$ ($Z_{obs} \in R_c$) rejeita-se H_0 em caso contrário aceita-se H_0.

Exemplo 5.9.1 *Um comprador de tijolos acha que a qualidade dos tijolos está diminuindo. De experiências anteriores, considera-se a resistência média ao desmoronamento de tais tijolos é igual a 200 kg, com um desvio padrão de 10 kg. Uma amostra de 100 tijolos, escolhidos ao acaso, forneceu uma média de 195 kg. Ao nível de significância de 5%, pode-se afirmar que a resistência média ao desmoronamento diminuiu?*

Solução Seja μ é a resistência média ao desmoronamento dos tijolos. Nesse caso, tem-se interesse em testar as seguintes hipóteses:
(i)

$$\begin{aligned} H_0 &: \mu = 200 \text{ kg} \\ H_1 &: \mu < 200 \text{ kg}. \end{aligned}$$

(ii) A estatística do teste é $\bar{X}$. Sendo $n = 100$, sob H_0, $\bar{X}$ tem distribuição $N(200; \frac{100}{100}) = N(200; 1)$.

(iii) A região crítica: $R_c = \{\bar{X} \leq k\}$ onde k é tal que $P(\bar{X} \leq k|H_0) = \alpha$ ou seja,

$$P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{k - 200}{10/10}) = P(Z \leq k - 200) = 0,05.$$

Assim, $z_\alpha = k - 200 = -1,64$. Logo, $k = 198,36$. O que resulta então, a $R_c = \{\bar{X} \leq 198,36\}$.

(iv) Do enunciado do problema a média amostral é $\bar{x} = 195 \in R_c = \{\bar{X} \leq 198,36\}$. Nesse caso, rejeita-se H_0 ao nível de significância de 5%.

Método alternativo: uma solução alternativa ao problema obtém-se como segue: No passo (iii) a região crítica na escala Z é da forma $R_c = \{z \in Z \sim N(0,1); Z \leq z_\alpha\}$. Para $\alpha = 0,05$ tem-se $z_\alpha = -1,64$. Então, $R_c = \{z \in Z \sim N(0,1); Z \leq -1,64\}$.

No passo (iv) ao ínves de calcular $\bar{x}$, obtém-se o valor da estatística do teste com os dados,

$$Z_{obs} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{195 - 200}{1} = -5.$$

Como $Z_{obs} = -5 < z_\alpha = -1,64$, rejeita-se H_0 ao nível de significância de 5%.

Procedimento geral

A seguir é apresentado o procedimento geral de teste de hipóteses para uma média populacional considerando o **procedimento alternativo** descrito acima.

(i)

$H_0 : \mu = \mu_0 (ou\ \ \mu \geq \mu_0)$	$H_0 : \mu = \mu_0 (ou\ \ \mu \leq \mu_0)$	$H_0 : \mu = \mu_0$
$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$	$H_1 : \mu \neq \mu_0$

(ii) A estatística do teste é:

Quando a variância é conhecida

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}, \tag{5.18}$$

onde n representa o tamanho da amostra através da qual é calculada o valor da média amostral $\bar{X}$. Quando H_0 é verdadeira, a estatística de teste segue uma distribuição normal padrão ou reduzida. Esse resultado é válido também, quando o tamanho da amostra é suficientemente grande para qualquer população.

Quando a variância é desconhecida

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}, \tag{5.19}$$

sendo S o desvio padrão amostral calculado com as n observações da amostra aleatória.

Se H_0 é verdadeira, a estatística (5.19) segue uma distribuição t-Student com $n - 1$ graus de liberdade.

(iii) As regiões críticas, para um nível de significância α fixado, são os valores da distribuição da estatística do teste (Z ou $t(n-1)$) se a hipóteses nula é verdadeira.

Para o teste de hipóteses unilateral esquerdo, a região crítica ou região de rejeição é representada pela parte hachurada da figura 5.11.a. Ela concentra valores na cauda esquerda da distribuição da estatística do teste, isto é, o conjunto, tal que: $R_c^{(z)} = \{c \in Z \sim N(0,1); Z \leq -c\}$ ou $R_c^{(t)} : \{c \in T \sim t(n-1); T \leq -c\}$. Para o teste unilateral direita (ou de cauda direita), a região crítica é representada pela parte hachurada da figura 5.11.b, e representa o conjunto de valores tal que $R_c^{(z)} = \{c \in Z \sim N(0,1); Z \geq c\}$ ou $R_c^{(t)} : \{c \in T \sim t(n-1); T \geq c\}$. Para o teste

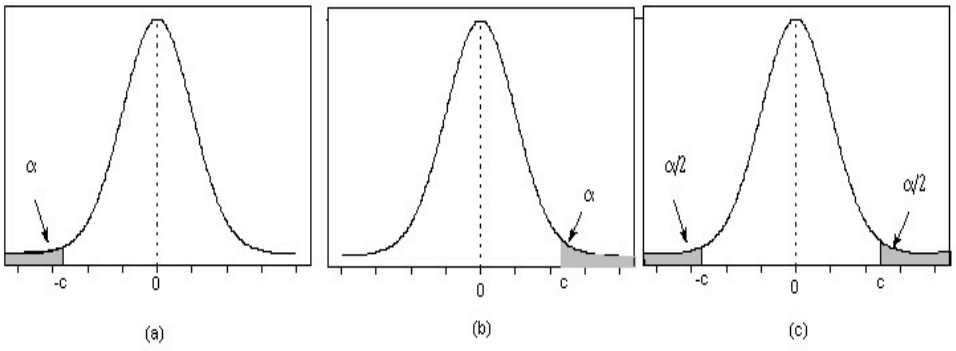


Figura 5.11: Regiões críticas para testes de hipóteses de uma média populacional :(a) unilateral esquerdo, (b) unilateral direito e (c) bilateral

bilateral, a região crítica é representada pela parte hachurada da figura 5.11.c, e representa o conjunto de valores tal que $R_c^{(z)} = \{c \in Z \sim N(0,1); |Z| \leq c\}$ ou $R_c^{(t)} : \{c \in T \sim t(n-1); |T| \leq c\}$.

(iv) Rejeita-se H_0, ao nível de significância, α se a estatística do teste observada (calculada com os dados da amostra) pertenece à região crítica, ou seja, se $Z_{obs} \in R_c^{(z)}$ ou $T_{obs} \in R_c^{(t)}$.

Exemplo 5.9.2 (Teste para um média populacional) *No exemplo 5.8.2, suponha que 25 homens na faxia etária entre 20 e 30 anos escolhidos ao acaso dessa população, foram tratados com o novo tratamento durante um período de três meses. Sendo o peso medio dos 25 homens igual a 84 kg, pode-se afirmar que o novo medicamento no combate da obesidade é eficaz. Use* $\alpha = 0,05$.

Solução: Seja X : Peso de homens da faixa etária entre 20 e 30 anos numa região do pais. Pelo enunciado tem-se, $X \sim N(90, 100)$. Deseja-se verificar as seguintes hipóteses:

$$\begin{aligned} H_0 &: \mu = 90 \quad \text{(o tratamento não é eficaz)} \\ H_1 &: \mu < 90 \quad \text{(o tratamento é eficaz).} \end{aligned}$$

onde μ é o peso médio de homens da faixa etária entre 20 e 30 anos tratados com o novo tratamento. Considerando que a variabilidade dos pesos dos homens tratados com a novo tratamento é a mesma da população a estatística de teste é (5.18), pois a população é normal, ou seja,

$$Z = \frac{\bar{X} - 90}{10/\sqrt{n}} \quad \text{sob} \overset{\sim}{} H_0 \quad N(0,1)$$

A região crítica é parte representada pela região hachurada da figura 5.12, para $\alpha = 0,05$:

Do enunciado tem-se: $\bar{X} = 84$ e $n = 25$. Logo a estatística de teste resulta

$$Z_{obs} = \frac{\bar{X} - 90}{10/\sqrt{n}} = \frac{84 - 90}{10/\sqrt{25}} = -2.$$

Como $Z_{obs} < -1,64$ rejeita-se H_0. Pode-se concluir para $\alpha = 0,05$ que o novo tratamento, proposto pelo endocrinologista, para perda de peso da população obesa dessa região, é eficaz.

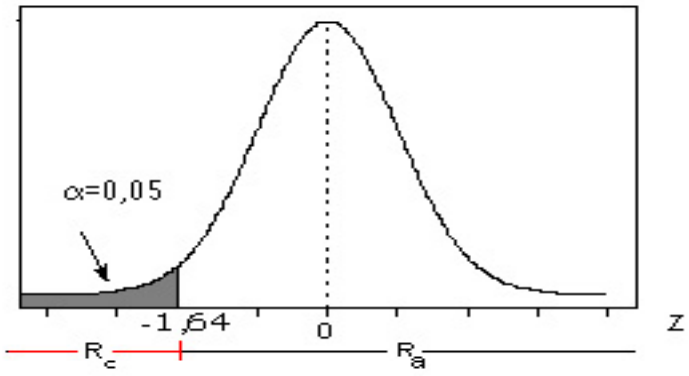


Figura 5.12: Região crítica para teste de hipóteses: $H_0 : \mu = 90$ contra $H_1 : \mu < 90$

5.10 Teste de Hipóteses para uma Variância Populacional

Suponha se tenha uma amostral aleatória de tamanho de uma população normal com média μ e variância σ^2 (ambas desconhecidas), e tem-se interesse em verificar as seguintes hipóteses estatísticas:

(i) $\begin{array}{lll} H_0 : \sigma^2 = \sigma_0^2 (ou\ \ \sigma^2 \geq \sigma_0^2) & H_0 : \sigma^2 = \sigma_0^2 (ou\ \ \sigma^2 \leq \sigma_0^2) & H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 & H_1 : \sigma^2 > \sigma_0^2 & H_1 : \sigma^2 \neq \sigma_0^2 \end{array}$

onde σ_0^2 é uma constante conhecida.

(ii) A estatística de teste é :

$$W = \frac{(n-1)S^2}{\sigma_0^2} \tag{5.20}$$

onde n é tamanho da amostra e S^2 é variância amostral calculada a partir das n observações amostrais. A estatística de teste, apresentada (5.20), tem distribuição qui-quadrado com $n-1$ graus de liberdade se a hipótese nula for verdadeira.

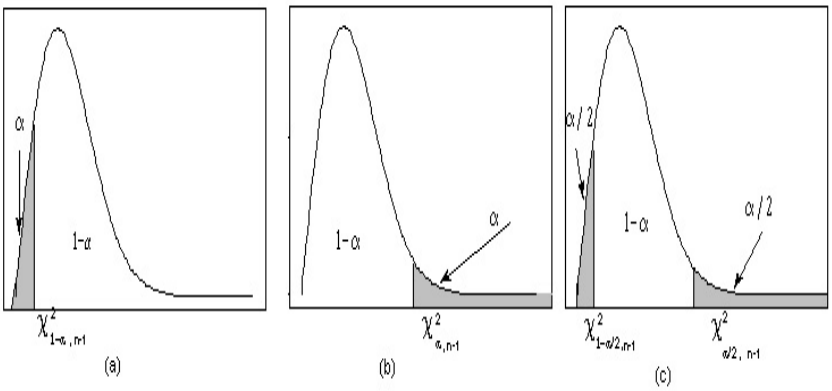


Figura 5.13: Região crítica para teste de hipóteses de uma variância populacional: (a) unilateral esquerdo, (b) unilateral direito e (c) bilateral

(iii) A região crítica para o teste de hipóteses unilateral é a parte hachurada da figura 5.13.a, que concentra valores na cauda esquerda da distribuição da estatística do teste, isto é, o conjunto tal que :$R_c = \{\chi^2_{n-1} \leq \chi^2_{1-\alpha,n-1}\}$. Para o teste unilateral de cauda direita , a região crítica é representada pela parte hachurada da figura 5.13.b, e representa o conjunto de valores da distribuição qui-quadrado com $n-1$ graus de liberdade, tal que $R_c = \{\chi^2_{n-1} \geq \chi^2_{\alpha,n-1}\}$. Para o teste bilateral a região crítica é representada pela parte hachurada da figura 5.13.c, e representa o conjunto de valores da distribuição qui-quadrado, estatística de teste, tal que $R_c = \{\chi^2_{n-1} \leq \chi^2_{1-\alpha/2,n-1}$ ou $\chi^2_{n-1} \geq \chi^2_{\alpha/2,n-1}\}$.

(iv) Rejeita-se H_0, ao nível de significância α, se a estatística de teste observada (calculada com os dados da amostra) pertence à região crítica, ou seja, se $W_{obs} \in R_c$.

Exemplo 5.10.1 (Teste hipóteses para uma variância populacional) *No exemplo 5.8.2, suponha que tem-se interesse em verificar se houve mudança no desvio padrão dos pesos na população. Com essa finalidade, 15 homens na faixa etária entre 20 e 30 anos foram escolhidos ao acaso dessa população. O desvio padrão dos 15 homens resultou em 8,5 kg. Use* $\alpha = 0,05$.

Solução: Como no exemplo 5.9.2, seja X : Peso de homens da faixa etária entre 30 e 30 anos numa região do pais. Portanto, $X \sim N(90, 100)$, deseja-se verificar as seguintes hipóteses:

$$\begin{aligned} H_0 &: \sigma = 10 \implies H_0 : \sigma^2 = 100, \\ H_1 &: \sigma \neq 10 \implies H_1 : \sigma^2 \neq 100 \end{aligned}$$

A estatística do teste é (5.20),

$$W = \frac{(n-1)S^2}{100} \overset{\sim}{\text{sob}} H_0 \quad \chi^2_{n-1}$$

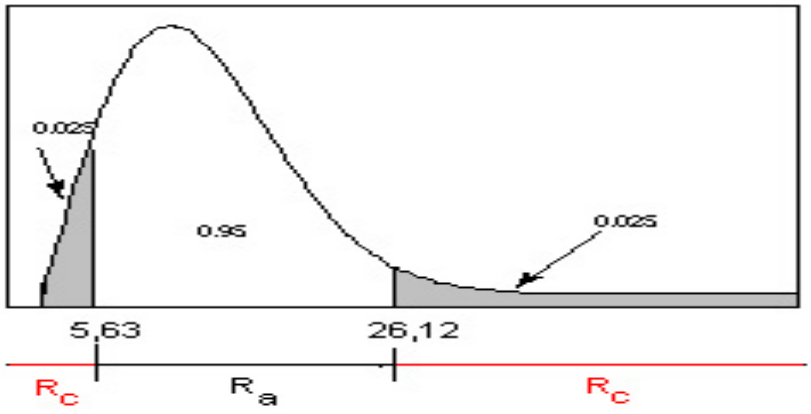


Figura 5.14: Região crítica para teste de hipóteses: $H_0 : \sigma^2 = 100$ contra $H_1 : \sigma^2 \neq 100$

A região crítica, para $\alpha = 0,05$, é o conjunto de valores da distribuição qui-quadrado com $n - 1 = 14$ graus de liberdade, tal que $R_c = \{\chi^2_{14} \leq 5,63$ ou $\chi^2_{14} \geq 26,12\}$, e é representada na figura 5.14, pela parte hachurada.

O valor da estatística calculada com os dados da amostra é :

$$W_{obs} = \frac{(15-1) \times 8,5^2}{100} = 10,115.$$

Como $W_{obs} \notin R_c$ aceita-se H_0 ao nível de significância de $\alpha = 0,05$

5.11 Teste de Hipótese para a Diferença de Médias Populacionais $(\mu_1 - \mu_2)$

Como no caso da construção de intervalos de confiança para a diferença de duas médias populacionais, considere que $X_1, \ldots, X_n$ é uma amostral aleatória de tamanho n de uma população com característica X, que tem distribuição normal com média μ_1 e variância σ_1^2. Considere que $Y_1, \ldots, Y_m$ é uma amostra aleatória de tamanho m, de uma população com característica Y que tem distribuição normal com média μ_2 e variância σ_2^2. Se X e Y são independentes foram apresentadas distribuições amostrais para a diferença das médias amostrais, quando as variâncias populacionais conhecidas e quando não são conhecidos mais iguais. Suponha que tem-se interesse em verificar se existe ou não uma diferença significativa entre as médias populacionais μ_1 e μ_2. O procedimento básico de teste, neste caso é a seguinte:

(i) As hipóteses estatística são:

$$\begin{array}{lll} H_0 : \mu_1 - \mu_2 = \Delta & H_0 : \mu_1 - \mu_1 = \Delta & H_0 : \mu_1 - \mu_2 = \Delta \\ H_1 : \mu_1 - \mu_2 < \Delta & H_1 : \mu_1 - \mu_2 > \Delta & H_1 : \mu_1 - \mu_2 \neq \Delta \end{array}$$

onde Δ é uma constante conhecida. Observa-se se $\Delta = 0$ tem-se o teste de hipóteses para a igualdade de duas médias populacionais.

(iii) A estatística do teste é:

Quando as variâncias σ_1^2 e σ_1^2 são conhecidas

$$Z = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \overset{\text{sob}}{\sim} H_0 \quad N(0,1) \tag{5.21}$$

Quando as variâncias $\sigma_1^2 = \sigma_2^2 = \sigma^2$ mas desconhecidas

$$T = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \overset{\text{sob}}{\sim} H_0 \quad t(n+m-2), \tag{5.22}$$

onde $S_p^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$, sendo S_1^2 e S_2^2 são as variâncias amostrais calculadas com as n e m das amostras da população X e população Y, respectivamente.

Quando as variâncias $\sigma_1^2 \neq \sigma_2^2$ e desconhecidas

$$T' = \frac{\bar{X} - \bar{Y} - \Delta}{\sqrt{(\frac{S_1^2}{n} + \frac{S_2^2}{m})}} \overset{\text{sob}}{\sim} H_0 \quad t(\nu), \tag{5.23}$$

onde $\nu = \dfrac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n+1} + \frac{\left(\frac{S_2^2}{m}\right)^2}{m+1}} - 2.$

Os passos (iii) e (iv) do procedimento de teste de hipóteses, são similares ao procedimento de teste de hipóteses para uma média populacional.

Exemplo 5.11.1 (Teste de hipóteses para diferença de duas médias populacionais) *Estuda-se o conteúdo de nicotina de duas marcas de cigarros (A e B), obtendo-se os seguintes resultados.*

A	*17;*	*20;*	*23;*	*20*	
B	*18;*	*20;*	*21;*	*22;*	*24*

Admitindo que o conteúdo de nicotinas das duas marcas tem distribuição normal e que as variâncias populacionais são iguais, com $\alpha = 0,05$, pode-se afirmar que existe alguma diferença significativa no conteúdo médio de nicotina nas duas marcas?

Solução: Sejam X : o conteúdo de nicotina da marca A. Y : o conteúdo de nicotina da marca B, tais que; $X \sim N(\mu_1, \sigma^2)$ e $Y \sim N(\mu_2, \sigma^2)$

As hipóteses estatística são:

$$\begin{aligned} H_0 : \mu_1 &= \mu_2 \iff H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 &\neq \mu_2 \iff H_1 : \mu_1 - \mu_2 \neq 0 \end{aligned}$$

(ii) A estatística do teste é dada em (5.22), pois as variâncias são iguais mais desconhecidas, ou seja

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \overset{\text{sob } H_0}{\sim} t(n+m-2),$$

onde é $n = 4$, $m = 5$ e S_p^2 é a variância ponderada.

(iii) A região crítica, para $\alpha = 0,05$, (parte achurada) representa os valores corresponte da distribuição t-Student com $n + m - 2 = 4 + 5 - 2 = 7$ graus de liberdade com mostra a figura 5.15:

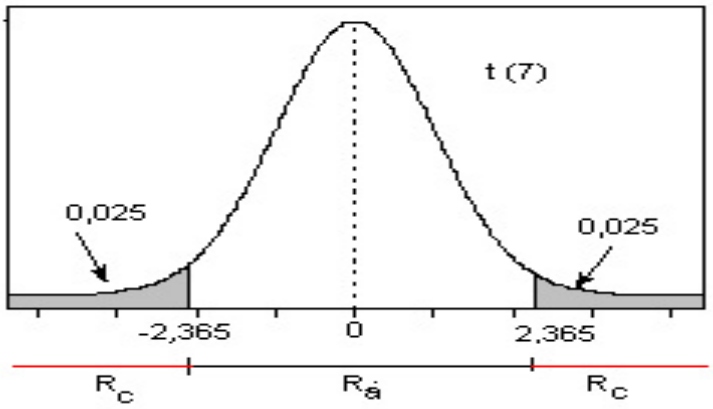


Figura 5.15: Região crítica para teste de hipóteses: $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \neq \mu_2$

Ou seja, é o conjunto: $R_c = \{t \in t(7); t \leq -2,365, \text{ ou } t \geq 2,365\}$

(iv) Dos dados do problema tem-se: $\bar{X} = 20$, $S_1^2 = 6$, $\bar{Y} = 21$, $S_2^2 = 5$ e $S_p^2 = \frac{(4-1)\times 6 + (5-1)\times 5}{4+5-2} = \frac{38}{7}$. Logo a estatística calculada ou observada é:

$$T_{obs} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} = \frac{20 - 21}{\sqrt{\frac{38}{7}\left(\frac{1}{4} + \frac{1}{5}\right)}} = -0,641$$

Como, $T_{obs} \notin R_c$, não se rejeita H_0. Portanto, não existe diferença significativa no conteúdo médio de nicotina nas duas marcas de cigarro ao nível de significância de $\alpha = 0,05$.

5.12 Teste de Hipóteses para a Igualdade de Duas Variâncis Populacionais

Na seção anterior foi apresentado o procedimento de teste de hipóteses para diferença de duas médias populacionais independentes. Em muitas outras situações, porém, pode-se estar interessado, também em verificar se as duas

populações independentes têm a mesma variância. Ou, pode-se estar interessado em estudar as variâncias de duas populações com a finalidade de verificar se a suposição de igualdade de variâncias. Para a escolha da estatística do teste no teste de hipóteses de diferença de duas médias, ou seja, para a escolha da estatística T dada em (5.22) ou T', dada em (5.23). Supõe-se que tem-se dois conjuntos de dados obtidos de duas populações independentes e distribuídos normalmente. Nesta seção apresenta-se o procedimento de teste de hipóteses estatístico para a igualdade de variâncias (homogeneidade).

(i)

$$H_0 : \sigma_1^2 = \sigma_2^2, \quad \text{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

(ii) A estatística do teste:

$$F = \frac{S_1^2}{S_2^2} \quad \text{sob}^{\sim} H_0 \quad F(n-1, m-1), \tag{5.24}$$

ou seja, F tem distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdade, sendo n o tamanho da amostra da população 1, S_1^2 a variância amostral da população 1, m o tamanho da amostra da população 2 e S_2^2 é variância amostral da população 2.

(iii) A região crítica para um nível significância α fixado é apresentada na figura 5.16. Ela representa o con-

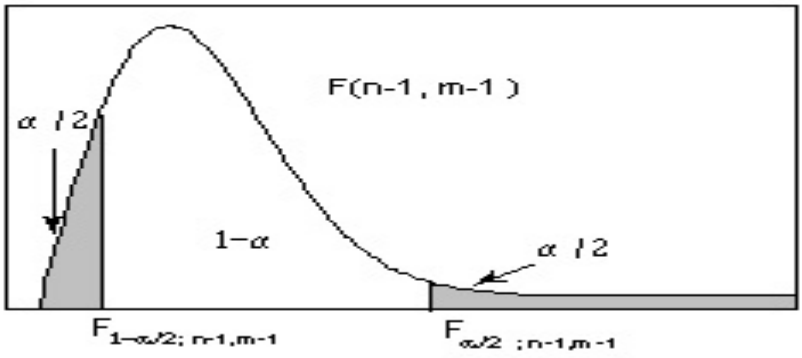


Figura 5.16: Região crítica para teste de hipóteses: $H_0 : \sigma_1^2 = \sigma_2^2$ contra $H_1 : \sigma_1^2 \neq \sigma_2^2$

junto de valores da distribuição F-Snedecor com $n-1$ e $m-1$ graus de liberdade, tal que: $R_c = \{F \leq F_{1-\alpha/2;n-1,m-1}$ ou $F \geq F_{\alpha/2;n-1,m-1}\}$.

(iv) Rejeita-se H_0 se a estatística calculada ou observada F_{obs} pertence à R_c.

Exemplo 5.12.1 *Um artigo publicado na* **Food Tecnology Journal** *(1956), descreve um estudo sobre o conteúdo de protopectina em tomates durante o armazenamento. Considerou-se dois períodos de armazenamento e analisou-se as amostras de nove lotes de tomates em cada período, obtendo-se os dados abaixo:*

Tempo de armazenamento	*Média*	*Desvio Padrão*
7 Dias	*792*	*495,0*
21 Dias	*372,3*	*73,3*

Admitindo que os conteúdos de protopectina para os 2 tempos de armazenamento tenha distribuição normal.

(a) Pode-se afirmar que as variâncias verdadeira de conteúdo de protopectina nos dois tempos de armazenamento são similares (ou homogêneas)? Use $\alpha = 0,10$

(b) Com probabilidade de cometer erro tipo I de 0,05, pode-se afirmar que o conteúdo médio de protopectina em tomates com tempo de 7 dias de armazenamento supera o conteúdo médio de protopectina em tomates armazenadas durante 21 dias em mais de 150 unidades ?

(c) Construa e interprete um intervalo de 90% de confiança para a razão de variâncias verdadeiras do conteúdo de protopectina armazenadas por um período de tempo de 7 dias e 21 dias.

Solução: (a) Sejam X : o conteúdo de protopectina em tomates armazenados em períodos de 7 dias e Y : o conteúdo de protopectina em tomates armazenados em períodos de 21 dias, tais que $X \sim N(\mu_1, \sigma_2^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$. Tem-se interesse em provar as seguintes hipóteses:

(i)
$$H_0 : \sigma_1^2 = \sigma_2^2, \quad \text{versus} \quad H_1 : \sigma_1^2 \neq \sigma_2^2$$

(ii) A estatística do teste é:
$$F = \frac{S_1^2}{S_2^2} \quad \text{sob}^{\sim} H_0 \quad F(n-1, m-1),$$
onde $n = m = 9$ e $F(8, 8)$ é a distribuição F-Snedecor com 8 e 8 graus de liberdade.

(ii) Para $\alpha = 0,10$ a região crítica R_c é tal que: $F_{0,05,8,8} = 3,44$ ou $F \leq F_{0,95,8,8} = \frac{1}{F_{0,05,8,8}} = \frac{1}{3,44} = 0,290698$. Ou seja, $R_c = \{F \leq 0,29069 \text{ ou } F \geq 3,44\}$.

(iv) Dos dados do problema tem-se $S_1 = 495$ e $S_2 = 73.3$. Com isso,
$$F_{obs} = \left(\frac{495}{73,3}\right)^2 = 45.6039.$$
Como $F_{obs} \in R_c$ rejeita-se H_0. Portanto, pode-se afirmar que as variâncias do conteúdo do protopectina em tomates armazenados em períodos de 7 dias e 21 dias não são similares.

(b) Nesse caso deseja-se verificar as seguintes hipóteses:
$$H_0 : \mu_1 - \mu_2 \leq 150 \quad \text{contra} \quad H_1 : \mu_1 - \mu_2 > 150$$
onde μ_1 é o conteúdo médio verdadeiro de protopectina em tomates armazenados durante 7 dias e μ_2 é o conteúdo médio verdadeiro de protopectina em tomates armazenados durante 21 dias.

A estatística de teste é (5.23) pois, no item (b), foi verificado que as variâncias são diferentes. Isto é,
$$T' = \frac{\bar{X} - \bar{Y} - 150}{\sqrt{(\frac{S_1^2}{n} + \frac{S_2^2}{m})}},$$
tem distribuição t-Student com $\nu = \dfrac{\left(\frac{S_1^2}{n} + \frac{S_2^2}{m}\right)^2}{\frac{\left(\frac{S_1^2}{n}\right)^2}{n+1} + \frac{\left(\frac{S_2^2}{m}\right)^2}{m+1}} - 2$, graus de liberdade.

Para $\alpha = 0,05$ e
$$\nu = \frac{\left(495^2/9 + (73^2/9)\right)^2}{\frac{(495^2/9)^2}{9+1} + \frac{(73^2/9)^2}{9+1}} - 2 \approx 8,0395 = 8,$$
a região crítica, é tal que: $R_c = \{T' \geq t_{0,025,8} = 1,860\}$.

Dos dados experimentais tem-se $\bar{X} = 792$ e $\bar{Y} = 372,3$; daí a estatística calculada ou observada é :
$$T'_{obs} = \frac{792 - 372,3 - 150}{\sqrt{(\frac{495^2}{9} + \frac{73,3^2}{9})}} = \frac{269,7}{166,779} = 1,61691$$

Como $T'_{obs} \notin R_c$, a hipóteses nula, H_0, não é rejeitado. Portanto, concluí-se que há evidência estatística suficiente para afirmar que conteúdo médio de protopectina de tomates armazenada em períodos de 7 dias não supera o conteúdo médio de protopectina de tomates armazenados em períodos de 21 dias, ao nível de significância de 5%.

(c) Como $1-\alpha = 0,90$ tem-se $\alpha = 0,10$ e como $n = m = 9$, da tabela da distribuição F-Snedecor com 8 e 8 graus de liberdade encontra-se os valores de $f_2 = 3,44$ e $f_1 = \frac{1}{3,44} = 0,290698$. Substituindo esss quantidades em (5.15) tem-se que um intervalo de 90% de confiança para a razão de variâncias, $\frac{\sigma_1^2}{\sigma_2^2}$, dado por:

$$IC(\frac{\sigma_1^2}{\sigma_2^2}; 0,90) = \left(0,29069 \times \left(\frac{495}{73,3}\right)^2 \; ; \; 3,44 \times \left(\frac{495}{73,3}\right)^2 \right) = (13,2570; 156.878).$$

Observa-se que esse intervalo de 90% de confiança não contém o valor de *um*, portanto pode-se afirmar com 90% de confiança que as variâncias do conteúdo de protopectina de tomates armazenadas em períodos de 7 dias e 21 dias não são homogêneas e essa mesma conclusão foi obtida através do procedimento de teste de hipóteses. Em geral pode-se utilizar intervalos de confiança para testar hipóteses bilaterais.

5.13 Teste Hipóteses para uma Proporção Populacional, para Amostras Grandes

O procedimento para os testes de hipóteses para proporção populacional é basicamente igual ao procedimento para o teste para uma média populacional. Considere o problema de testar a hipótese que a proporção de sucessos de um ensaio de Bernoulli é igual a valor especifico, p_0. Isto é, testar as seguintes hipóteses:

$$\begin{array}{lllll} H_0 & : & p = p_0 & H_0 : p \geq p_0 & H_0 : p \leq p_0 \\ H_1 & : & p \neq p_0 & H_1 : p < p_0 & H_1 : p > p_0, \end{array}$$

A estatística de teste é :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{\sim}{\text{sob}} H_0 \quad N(0,1), \tag{5.25}$$

ou seja, a estatística do teste (5.25) tem distribuição normal padrão. Na expressão acima, $\hat{p}$ a proporção amostral calculada com as n observações amostrais ($n \geq 30$).

Exemplo 5.13.1 *Um estudo é realizado para determinar a relação entre uma certa droga e certa anomalia em embriões de frango. Injetou-se 50 ovos fertilizados com a droga no 4^0 dia de incubação. No vigésimo dia de incubação, os embriões foram examinados e 7 apresentaram a anomalia. Suponha que deseja-se averiguar se a proporção verdadeira é inferior a 25% com um nível de significância de 0,05.*

Solução: Seja Y : número de embriões que apresentam anomalia nos 50 ovos fertilizados com a droga. Então, $Y \sim B(50, p)$, onde p é proporção populacional (ou verdadeira) de embriões que apresentam anomalia. Deseja-se verificar as seguintes hipóteses:

$$H_0 : p = 0,25 \quad \text{contra} \quad H_1 : p < 0,25.$$

A estatística de teste é apresentada em (5.25). Com $p_0 = 0,25$. Tem-se

$$Z = \frac{\hat{p} - 0,25}{\sqrt{\frac{0,25(1-0,25)}{n}}} \overset{\sim}{\text{sob}} H_0 \quad N(0,1).$$

A região crítica, para $\alpha = 0,05$ é o conjunto de valores da distribuição normal padrão menores ou iguais a $-1,64$ como mostra a figura 5.17. Isto é, $R_c = \{z \in Z; Z \leq -1,64\}$.

Temos que $n = 50$ e $Y = 8$. Portanto, $\hat{p} = \frac{Y}{n} = \frac{7}{50} = 0,14$ é a proporção estimada de embriões que apresentam a anomalia. A estatística calculada ou observada é:

$$Z_{obs} = \frac{\hat{p} - 0,25}{\sqrt{\frac{0,25(1-0,25)}{n}}} = \frac{0,14 - 0,25}{\sqrt{\frac{0,25 \times 0,75}{50}}} = -1,7963.$$

Como $Z_{obs} < -1,64$, rejeita-se H_0. Conclui-se ao nível de significância de 5% que a proporção de embriões que apresentam anomalia ao serem fertilizados com a droga é significativamente inferior a 25%.

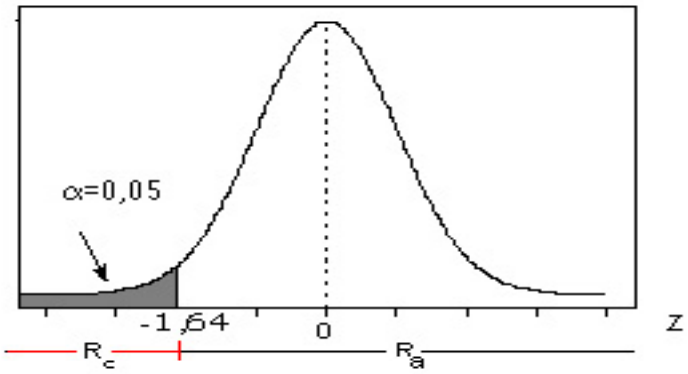


Figura 5.17: Regiões críticas para teste de hipóteses: $H_0 : p \geq 0,25$ contra $H_1 : p < 0,25$

5.14 Teste de Hipóteses de Igualdade de Duas Proporções Populacionais para Amostras Grandes

Suponha que tem-se duas amostras independentes de tamanhos n e m suficientemente grandes ($n > 30$ e $m > 30$), de duas populações Bernoulli, com probabilidades de sucessos p_1 e p_2 respectivamente. E sejam X : o número de sucessos na amostra de tamanho n e Y : o número de sucessos na amostra de tamanho m. Portanto, $X \sim B(n, p_1)$ e $Y \sim B(m, p_2)$. Há interesse em verificar as seguintes hipóteses estatística:

$$\begin{array}{lcccc} H_0 & : & p_1 = p_2; & H_0 : p_1 \geq p_2; & H_0 : p_1 \leq p_2; \\ H_1 & : & p_1 \neq p_2; & H_1 : p_1 < p_2; & H_1 : p_1 > p_2, \end{array}$$

A estatística do teste é, então

$$Z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n} + \frac{1}{m})}}, \tag{5.26}$$

que tem distribuição normal padrão se H_0 for verdadeira. Onde $\hat{p_1} = \frac{X}{n}$, $\hat{p_2} = \frac{Y}{n}$ e $\bar{p} = \frac{X+Y}{n+m}$.

Exemplo 5.14.1 *Um experimento foi conduzido com a finalidade de estudar a efetividade da vacina Salk contra a pólio. Para isso, considerou-se um grupo de 100 camundongos com as mesmas características (idade, peso, etc), os quais foram distribuídos ao acaso em dois grupos iguais. Ao primeiro grupo aplicou-se uma vacina similar sem o composto mais importante da vacina salk (placebo), e observou-se que 40 dos 50 camundongos foram imunizados. No outro grupo aplicou-se a vacina salk e observou-se que 45 dos 50 foram imunizados. Pode-se afirmar que a vacina Salk é efetiva contra a pólio. Use* $\alpha = 0,05$.

Solução: Sejam X : número de camundongos imunizados com a vacina Salk no grupo de 50 e Y : número de camundongos imunizados com a vacina placebo no grupo de 50. Então $X \sim B(50, p_1)$ e $Y \sim B(50, p_2)$. Tem-se interesse em verificar as seguintes hipóteses:

$$H_0 : p_1 \leq p_2 \quad \text{contra} \quad H_1 : p_1 > p_2$$

A estatística de teste é dada em (5.26), tem distribuição normal padrão. Para $\alpha = 0,05$, a região critica é a parte hachurada mostrada na figura 5.18. Ou seja, $R_c = \{z \in Z; Z \geq 1,64\}$.

Como $\hat{p_1} = \frac{45}{50} = 0,90$ e $\hat{p_2} = \frac{40}{50} = 0,80$ e $\bar{p} = \frac{45+40}{100} = 0,95$ a estatística apresentada em (5.26), avaliada com os dados amostrais é,

$$Z_{obs} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\bar{p}(1-\bar{p})(\frac{1}{n} + \frac{1}{m})}} = \frac{0,90 - 0,80}{\sqrt{0,95 \times 0,05(\frac{1}{50} + \frac{1}{50})}} = 2,294.$$

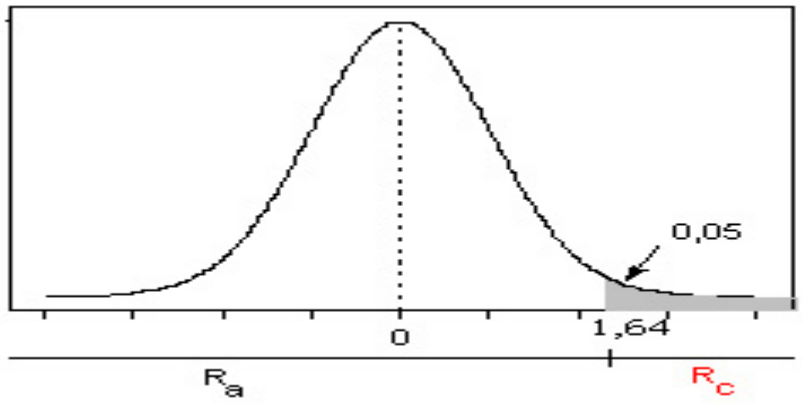


Figura 5.18: Regiões críticas para testar: $H_0 : p_1 \leq p_2$ contra $H_1 : p_1 > p_2$.

Como $Z_{obs} \in R_c$ rejeita-se H_0. Conclui-se, ao nível de significância de 5% que a vacina Salk é efetiva contra pólio.

5.15 Nível Descritivo

De acordo com o procedimento descrito anteriormente para o teste de hipóteses, no final toma-se uma decisão de *rejeição* ou de *não-rejeição* da hipótese nula. Esta dicotomia é, na realidade, artificial. De fato

(i) a fixação de um nível de significância é arbitrária e

(ii) os dados amostrais podem contradizer a hipótese nula em maior ou menor grau.

O **nível descritivo** denotado por α^* (ou **P-value**) constitui uma medida do grau com que os dados amostrais contradizem a hipótese nula. A sua definição é a seguinte: o **nível descritivo** corresponde à probabilidade da estatística de teste tomar um valor igual ou mais extremo do que aquela que, de fato, é observado. Alternativamente, pode-se definir o nível descritivo como o menor nível de significância para o qual a estatística de teste determina a rejeição da hipótese nula H_0. Note-se que, tal como a estatística de teste, o nível descritivo é calculado admitindo que H_0 seja verdadeira.

Exemplo 5.15.1 *No exemplo 5.13.1, a estatística de teste observada é,* $Z_{Obs} = -1,7963$ *(recorde-se que o nível de significância do teste era* $\alpha = 0,05$ *e o correspondente valor crítico* $z_{0,05} = -1,64$*).*

De acordo com a definição apresentada, o nível descritivo da prova, α^*, é:

$$\alpha^* = P(Z \leq -1,7963|H_0) = 0,0362 \text{ (veja a tabela normal padrão)}$$

Portanto, o nível descritivo é de 3,62% que indica a probabilidade de encontrarmos valores da estimativa mais desfavoráveis à hipótese nula. Note que o valor do nível descritivo se relaciona diretamente com o nível significância.

Nesse exemplo, se o nível de significância fosse fixado em qualquer valor igual ou superior a 3,62%, a conclusão seria pela rejeição de H_0 ao passo que valores inferiores a 3,62% conduziriam à aceitação da hipótese nula. O significado do nível descritivo é ilustrado na figura 5.19,

Como é evidente, quanto menor for o valor do nível descritivo maior será o grau com que a hipótese nula é contradita. Dada a relevância da informação contida no nível descritivo, é recomendável a sua inclusão explícita nos resultados de qualquer teste de hipóteses. Por exemplo, muito mais esclarecedor do que dizer que uma hipóteses nula foi rejeitada, ao nível de significância de 5%, é afirmar que isso sucedeu e que o nível descritivo foi de 0,3%.

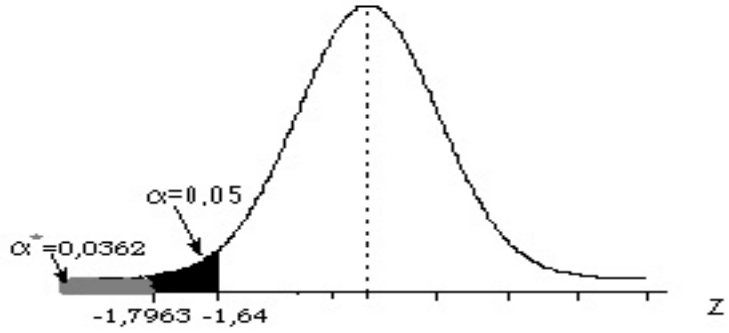


Figura 5.19: Valor do nível descritivo para testar: $H_0 : p = 0,25$ contra $H_1 : p < 0,25$.

Para os testes de hipóteses na qual a distribuição normal é a estatística do teste, o nível descritivo nesta caso é dado por:

$$\alpha^* = \begin{cases} 2(1 - \Phi(|z_{obs|})); & \text{para teste bilateral} \\ 1 - \Phi(z_{obs}); & \text{para teste unilateral de cauda superior} \\ \Phi(z_{obs}); & \text{para teste unilateral de cauda inferior} \end{cases}$$

onde z_{obs} é o valor da estatística do teste e $\Phi(\ )$ é a função da distribuição acumulada normal padrão definida no capítulo anterior.

5.16 Exercícios

1. Com a finalidade de estudar os efeitos do feijão no consumo humano examinou-se o incremento de peso de 20 indivíduos ao final de 3 dias. O pesquisador por experiências anteriores conhece que variância do incremento de peso de qualquer grupo de pessoas é 16 gramas. Os dados apresenta-se a continuação:

8,0	20,4	15,1	11,2	16,0	12,5	19,2	17,4	14,2	19,3
19,2	16,6	10,1	8,1	18,0	9,5	13,1	21,2	15,0	16,2

   (a) Construa um intervalo do 98% de confiança para o incremento de peso médio verdadeiro.

   (b) Suponha que deseja-se saber quantos indivíduos tem que ser examinados, para que o erro da média amostral não exceda 1,5 gr, com 99% de confiança.

2. Uma pesquisa é feita com a finalidade de estimar a proporção de estudantes da UFOP, usuários de algum tipo de droga (p) com um margem de erro de três pontos percentuais, a porcentagem de estudantes usuários de algum tipo de droga. Supondo que se pretende um nível de confiança de 99% nos resultados, quantos estudantes devem ser pesquisados ?.

   (a) Suponha que tenhamos uma estimativa com base em estudo anterior, que mostrou que 67% dos estudantes tinham consumido algum tipo de droga.

   (b) Suponha que não tenhamos qualquer informação que possa sugerir o valor de p.

   (c) Sabendo-se que a amostra obtida no item (a), forneceu uma estimativa de que 70% dos estudantes tinham consumido algum tipo de droga, obtenha e interprete um intervalo de 95% de confiança para a verdadeira proporção de estudantes que consumem algum tipo de droga.

3. Um artigo publicado no Journal Of Heat Transfer (1974) descreve um novo método para medir a condutividade térmica do ferro Armco. Ao utilizar uma temperatura de 100^oF e uma potência de entrada de 550 W,

resultaram as seguintes medições de condutividade (em Btu/hr-ft-^{o}F): 41,60; 41,48; 42,34; 41,95; 41,86; 42,18; 42,26; 41,48; 42,04; 41,72. Supondo que a condutividade térmica a 100^{o}F e 550 W se distribui normalmente com desvio padrão, $\sigma = 0,30 Btu/hr - ft - ^o F$. Obtenha um intervalo do 95% de confiança da condutividade média deste material.

4. De um lote de 2200 lâmpadas foram sorteadas 81 lâmpadas ao acaso, o tempo médio de duração das lâmpadas sorteadas foi de 3200 horas e um desvio padrão de 900 horas. Construa um intervalo de 95% de confiança para o tempo médio das lâmpadas do lote (suponha que tempo de duração das lâmpadas é normal).

5. A resistência média à tensão de uma fibra sintética é uma característica importante de qualidade de interesse do fabricante, o qual deseja encontrar um intervalo de 95% de confiança para estimar a média. O fabricante supõe, com base na resistência à tensão está distribuída aproximadamente normal. Embora, se desconheça a resistência média à tensão e seu desvio padrão. Selecionou-se uma amostra aleatória de 16 troços de fibra e determinou-se sua resistência (em psi, lb/plg2). A média e desvio padrão amostrais resultaram respectivamente; 49,86 psi e 1,66 psi. Que dizer ao respeito à resistência média da fibra sintética?.

6. Uma firma construtora deseja estimar a resistência média das barras de aço utilizadas na construção de casas. Qual tamanho amostral se requer para garantir que haja um risco de 0,001 de ultrapassar um erro de 5 kg ou mais na estimação? O desvio padrão da resistência para este tipo de barra é considerado 25 kg.

7. Uma psicóloga elabora um novo teste de percepção espacial e deseja estimar o escore médio alcançado por pilotos do sexo masculino. Quantas pessoas ela deve testar para o que o erro da média amostral não exceda 2,0 pontos, com 95% de confiança ?. Estudo anterior mostro sugere que $\sigma = 21,2$.

8. As alturas de estudantes mulheres do primeiro ano de uma universidade têm distribuição normal com média de 1,65 m, e desvio padrão de 0,5 m. Quantas estudantes devem ser pesquisadas, se queremos estimar a porcentagem das que têm mas 1,60 m. de altura ?. Admita um nível de confiança de 99% , em que o erro não supere 2,5 pontos percentuais.

9. Um fabricante da área farmacêutica produz frascos de certo produto. A quantidade de certo principio ativo em cada frasco é uma variável aleatória com média desconhecida e variância 30 mg. Um comprador tomou uma amostra de 11 elementos e mediu a quantidade dessa substância em cada frasco a média da amostral foi 263 mg. Supondo normalidade, determine um intervalo de 95% de confiança para a quantidade média da substância em cada frasco.

10. Uma agência governamental está encarregada de fiscalizar a contaminação de um certo produto alimentício, através da análise de uma amostra dos pacotes desse produto. Uma porcentagem de contaminação de 7% é considerado tolerável. Se a porcentagem de contaminação for maior que este valor o produtor deverá ser atuado. Uma norma da agência estabelece que, se no exame de 20 pacotes desse produto forem detectados pelo menos 4 pacotes contaminados, então a fabrica deve ser multado. Seja p a proporção de contaminação do produto.

    (a) Formule as hipóteses estatísticas especificando as hipóteses nula e alternativa .

    (b) Qual é o significado do erros tipo I e do tipo II neste problema.?

    (c) Qual é região crítica escolhida ?

    (d) Qual é nível de significância do teste ?

    (e) Qual é a probabilidade de se atuar o produto se a proporção de contaminação de seu produto for 15%.?

11. O encarregado do controle de trafego aéreo da Companhia de aviação ASA afirma que 95% dos vôos dessa Companhia chegam ao lugar de destino no máximo com 30 minutos de atraso. Uma instituição de defesa do consumidor recebeu queixas dos clientes da ASA que afirmam que a porcentagem de vôos que chegam no máximo com 30 minutos de atraso é muito menor. Eles examinam uma amostra selecionada ao acaso de 200 registros de vôos da ASA e verificaram que 182 vôos chegaram no máximo com 30 minutos de atraso.

    (a) Formule as hipóteses nula e alternativa . Faça o teste usando o nível descritivo (P-value)

    (b) Construa um intervalo do 98% de confiança para a verdadeira proporção .

12. As companhias de seguros estão ficando preocupados com o fato de que o número crescente de telefones celulares resulte em maior colisões de carros; estão por isso, pensando em cobrar prêmios mais elevados para os motorista que utilizam celulares. Desejamos estimar, com um margem de erro de três pontos percentuais, a porcentagem de motoristas que falam ao celular enquanto estão dirigindo. Supondo que se pretende um nível de confiança de 95% nos resultados, quantos motoristas devem ser pesquisados ?.

    (a) Suponha que tenhamos uma estimativa com base em estudo anterior, que mostrou que 18% dos motoristas falavam ao celular.

    (b) Suponha que não tenhamos qualquer informação que possa sugerir o valor de .

13. O rótulo de remédio contra resfriado Dozenol indica a presença de 600 mg de acetaminofem em cada onça fluida. A Food and Drug Administration (FDA) selecionou aleatoriamente 65 amostra de uma onça e constatou que o conteúdo médio de acetaminofen é de 585 mg, com um desvio padrão de 21 mg. Ao nível de significância de 1%, testa a afirmação da Medassist Pharmaceutical Company de que a média populacional é igual a 600 mg.

14. Determinou-se o custo de operação por cliente para cada uma de 12 organizações. Os 12 valores têm média de $2133 e desvio padrão de $345 .Ao nível de 0,01 de signifcância, teste a afirmação de uma acionista, que se queixa de que a média para todas as organizações desse tipo excede $ 1800 por cliente.

15. Em um estudo de 71 fumantes que estavam procurando deixar de fumar utilizando uma terapia especial, 32 não estavam fumando uma após o tratamento. Ao nível de 0,10 de siginificância, teste a afirmação de que, dos fumantes que procuram deixar de fumar com aquela terapia, a maioria está fumando um após o tratamento. Esses resultados sugerem que a terapia não é eficaz?

16. A Medassit Pharmaceutical Company utiliza uma maquina para encher frascos com um remédio, de tal maneira que o desvio padrão dos pesos é de 0,15 oz. Testou-se uma nova maquina em 71 frascos e, para essa amostra, o desvio padrão é 0,12 oz. A Dayton Machine Company, fabricante da nova maquina, afirma que ela enche os frascos com menor variação.

    (a) Teste a afirmação da Dayton Machine Company, ao nível de 0,05 de significância de. Se a máquina na Dayton está sendo usanda como experiência, deve-se cogitar de sua aquisição ?

    (b) Determine um intervalo de 95% de confiança para o desvio padrão dos pesos nos frascos.

17. Pesquisadores de Johns Hopkins fizeram um estudo de empregadas da IBM que estavam gravidas. De 30 empregadas que lidavam com éter-glicol, 10 tiveram aborto (espontâneo) mas, de 750 que não estavam expostas ao éter-glicol, apenas 120 abortaram.

    (a) No nível de 0,01 de significância, teste a afirmação de que as mulheres expostas ao éter-glicol apresentam maior taxa de aborto.

    (b) Qual é o nível descritivo para o teste de hipóteses em (a) ?.

18. A empresa "Duramas"garante que, se os pneus forem utilizados com condições normais, têm uma vida média superior a 40000 km. Uma amostra constituída por 30 pneus utilizados nas condições acima referidas proporcionou os seguintes resultados: $\bar{X} = 43200$ km e $S = 8000$ km. Teste, ao nível de significância de 5% se os pneus têm a vida média que o fabricante reivindica.

19. Um certo analgésico adotado em determinado hospital é eficaz em 70% dos casos. Um grupo de médicos chineses em vista a esse hospital afirma que a utilização da acupuntura produz melhores resultados. A direção do hospital resolve testar o método alternativo em 80 pacientes sorteados ao acaso, com a finalidade de adotá-lo em definitivo se ele apresentar eficiência satisfatória numa proporção de casos maior que do anestésico atual. Seja p a probabilidade de que a o método de acupuntura apresente a eficiência satisfatória quando aplicada a um paciente.

    (a) Formule este problema como um problema de testes de hipóteses especificando as hipóteses nula e alternativa.

    (b) Quais os erros de tipo I e II (em palavras) ?

(c) Supondo que o critério para rejeitar a hipóteses nula seja: número de pacientes, com resultado satisfatório, no mínimo 64, qual é a probabilidade do erro tipo I ? Interprete o resultado.

(d) Se dentre os 80 pacientes submetidos à nova técnica em 69 deles apresentaram eficiência satisfatória, qual é a decisão a ser tomada ?. (Use $\alpha = 0,01$)

20. Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de 23 mg por cigarro. Um laboratório realiza 6 análises desse índice, obtendo: 27; 24; 21; 25; 26; 22. Sabe-se que o índice de nicotina se distribui normalmente, com variância igual a 4,86 mg^2.

(a) Pode-se aceitar, no nível de 10%, a afirmação do fabricante?

(b) Determine o nível descritivo e qual é sua conclusão?

21. Um fabricante de um certo tipo de aço especial afirma que seu produto tem um severo serviço de controle de qualidade, traduzido pelo desvio padrão da resistência à tensão que não é maior do que 5 kg por cm^2. Um comprador, querendo verificar a veracidade da afirmação, tomou uma amostra de 11 cabos e submeteu-a a um teste de tensão. Os resultados foram as seguintes: $\bar{x} = 263$ e $S^2 = 48$. Esses resultados trazem alguma evidência contra a afirmação do fabricante? Use $\alpha = 0,05$.

22. Karl Pearson, que elaborou muitos conceitos importantes em estatística, coletou dados sobre crimes que 1909. Dos condenados por incêndio criminoso, 50 bebiam 43 eram abstêmios. Dos condenados por crime de fraude, 63 bebiam e 144 eram abstêmios. Com o nível de 0,01 de significância, teste a afirmação de que a proporção dos que bebem entre os incidiarios é maior do que proporção dos bebedores condenados por fraude. A bebida parece ter algum efeito sobre o tipo de crime?. Por que?

23. Realiza-se um experimento para comparar a média da absorção de medicamentos em espécimens de tecido muscular. Divide-se 72 espécimens em dois grupos iguais, seguindo um procedimento aleatório. Cada grupo foi ministrada uma de das 2 medicamentos (A e B), as médias amostrais foram respectivamente :$\bar{X_A} = 7,9$ e $\bar{X_B} = 8,5$. Admitindo que a absorção dos medicamentos tem distribuição normal e, que a variância de absorção para este tipo de medicamentos é 0,10.

(a) Construa um intervalo de 99% de confiança para diferença de médias de absorção do medicamento A e B.

(b) No nível de 1% de significância, pode-se afirmar que absorção dos medicamentos são os mesmos ?

(c) Teste o item (a) usando o nível descritivo ?

24. Dividem-se 50 pacientes de epilepsia em duas amostra aleatórias iguais, Ao grupo A se lhe deu tratamento que incluía doses diárias de vitamina D. Ao grupo B se lhes deu o mesmo tratamento com exceção que não recebeu vitamina D ao invés recebeu placebo em seu lugar. Os dados sumariados do número de ataques experimentados são apresentados na tabela embaixo:

Tratamento	Média	Variância
Vitamina D	15	8
Placebo	24	18

(a) Pode-se afirmar que as variâncias do número ataques dos 2 tratamentos são similares ou homogêneos. Use a=0,10.

(b) Há suficiente evidência que indique que a vitamina D reduz o número de ataques epilépticos ?. Use a=0,05.

(c) Construa um intervalo do 95% para diferença de médias de ataques do tratamento com vitamina D e com placebo.

25. Um artigo publicado no *Journal of Sport Science* (1987) apresenta os resultados de uma pesquisa sobre o nível de hemoglobina dos jogadores do jockey sobre gelo na olimpíada de Canada. Os resultados que aparecem no artigo são as seguintes (em g/dl):

15,3 16,0 14,5 16,2 14,9 15,7 15,3 14,6 14,5 16,2
15,7 16,0 15,0 15,7 16,2 14,7 14,8 14,6 15,6 15,2

Outro pesquisador mediu o nível de hemoglobina de 20 pessoas normais não esportistas escolhidos ao acaso. Os dados (em g/dl) são os seguintes:

12,5	13,0	10,3	11,6	10,6	11,2	13,4	10,2
11,8	14,0	11,2	11,9	12,2	10,9	11,1	9,8

Supondo que os dados têm distribuições normal.

(a) Pode-se afirmar que a variâncias do nível de hemoglobina em pessoas esportistas e não esportistas são as mesma. Use a=0,10.

(b) Determine um intervalo de 95% de confiança para a razão de variâncias do nível de hemoglobina entre os que são esportistas os que não são .

(c) Com probabilidade de cometer erro tipo I de 0,05, você poderia afirmar que existe alguma diferença no nível de hemoglobina entre pessoas esportistas e não esportistas. ?

(d) Considerando o item (a), construa um intervalo de 95% de confiança para diferença de médias do nível de hemoglobina entre pessoas esportistas e não esportistas.

26. Uma pesquisa é feita com a finalidade de verificar se filtros de cigarros realmente diferença, ou apenas são truques de venda sem qualquer efeito real. A continuação apresentam-se os dados sumariados dos conteúdos alcatrão e nicotina em uma amostra aleatória de cigarros tamanho padrão, com filtro e sem filtro. Todas as medidas em miligramas.

	Com filtro		Sem filtro	
	Alcatrão	Nicotina	Alcatrão	Nicotina
Tamanho da amostra	21	21	8	8
Média	13,3	0,94	24,0	1,65
Desvio padrão	3,7	0,31	1,7	0,16

Supondo que os dados tem distribuição normal.

(a) Construa e interprete um intervalo de 98% de confiança para desvio padrão da quantidade de nicotina em cigarros com filtro.

(b) Construa e interprete um intervalo do 95% de confiança , para a quantidade media de nicotina em cigarros sem filtro.

(c) Pode-se afirmar que a variâncias da quantidade de nicotina em cigarros com filtro e sem filtro são as mesma.? Use a=0,10.

(d) Com probabilidade de cometer erro tipo I de 0,05, você poderia afirmar que quantidade alcatrão em cigarros sem filtros é maior à quantidade alcatrão em cigarros com filtro ?

(e) Considerando o item (c), construa um intervalo de 95% de confiança para diferença de médias da quantidade média de alcatrão de cigarros com filtro e sem filtro.

27. Em estudo recente de 22.000 médicos, metade tomou doses regulares de aspirina, e à outra metade foi administrado um placebo. O estudo se estendeu por seis anos, a um custo total de $ 4,4 milhões. Entre os que tomaram aspirina, 104 tiveram ataque cardíacos, e dos que receberam um placebo 189 tiveram ataques.

(a) Esses resultados mostram uma redução estatisticamente significativa dos ataques cardíacos no grupo que tomaram aspirina ?. (Use o nível descritivo).

(b) Construa e interprete um intervalo do 98% de confiança para a proporção de médicos que tomaram aspirina e não tiveram ataques cardíacos.

28. Uma peça de um certo equipamento elétrico é fornecido, sob encomenda, por duas empresas externas (A e B). A dimensão desta peça é uma característica de qualidade importante no momento da montagem do produto. Para examinar se há diferença nas dimensões das peças da empresa A e empresa B, forem extraídas amostras aleatórias das respectivas fabricas, obtendo-se os dados abaixo (em mm):

Empresa A						Empresa B					
12,5	12,6	12,4	12,8	12,7	12,6	13, 0	13,1	13,0	13,2	13,1	12,7
12,6	12,5	12,6	12,4	12,3	12,7	13,0	12,1	12 ,9	12,9		

Supondo que os dados tem distribuição aproximada normal.

(a) Para um nível de significância de 5%, pode-se afirmar que variâncias são homogêneas ?

(b) Considerando o item (a), Existe diferencias significativas entre a média da dimensão fornecida pêlos dois fornecedores ?. Use $\alpha = 0.05$.

(c) Obtenha o nível descritivo do teste em (b).? Qual é sua conclusão ?

(d) Obtenha e interprete um intervalo de 95% de confiança para a diferença de media da dimensão do fornecedores A e B.

29. Numa determinada empresa industrial, uma peça é fabricada automaticamente, em grandes quantidades, por duas maquinas A e B, que se distinguem apenas pelo fato da maquina B ser mais velha (e mais usada) do que a maquina A. Com a finalidade de avaliar se as duas maquinas estão produzindo peças da mesma qualidade, avaliou-se o tempo (em segundos) de operação de cada maquina em produzir uma peça e, também foi verificado se peça satisfaz os requerimentos de Engenharia (se a peça é defeituosas ou não). Da linha de produção da maquina A obteve-se uma amostra aleatória de 31 peças e, da maquina B uma amostra aleatória de 41 peças obtendo-se os seguintes resultados .

Maquina	Tempo médio	Variância	N^o de peças defeituosos
A	45,020	31,393	6
B	48,041	6,758	6

(a) Pode-se dizer, ao nível de 5% de significância, que o tempo médio de operação da maquina B supera o tempo médio de operação da maquina A em mais de 2 segundos ?

(b) Quais são os pressupostos necessários para a resolução de item (a) ?

(c) Para um nível de 5% de significância, pode-se afirmar que maquina A produz a mesma proporção de peças defeituosas que a maquina B ?.

(d) Qual é o nível descritivo em (c), ? Qual é sua conclusão ?

30. Um experimento é conduzido para comparar dois regimes alimentares no que diz ao aumento de peso. Vinte indivíduos são distribuídos ao acaso entre dois grupos em que ao primeiro deles foi dada a deita A ao segundo a B. Decorrido certo intervalo de tempo verifica-se que os aumentos de peso correspondentes foram os seguintes:

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

Supondo que incrementos de peso tem distribuição normal.

(a) Construa e interprete um intervalo do 95% de confiança para o desvio padrão do incremento do peso de indivíduos alimentados com a dieta B.

(b) Ao um nível de 10% de significância pode-se afirmar que as variâncias verdadeiras dos incrementos de pesos de pessoas alimentadas com a dieta A e B são similares ?

(c) Com probabilidade de cometer erro tipo I de 0,05, você poderia afirmar que dieta B é melhor que a dieta A.

(d) Considerando o item (b), construa um intervalo de 95% de confiança para diferença de médias do incremento de peso de pessoas alimentadas com dieta B e A .

Capítulo 6

Análise de regressão e correlação

6.1 Introdução

Em diversas áreas de aplicação, freqüentemente há interesse em estudar a relação entre duas variáveis, como quantidade de fertilizante; e a produção com o uso do fertilizante, a concentração de uma droga injetada em um animal de laboratório e o batimento do coração após a injeção; a dureza de um plástico tratado com calor durante diferentes períodos de tempo , etc. A natureza e o grau de relação entre variáveis podem ser analisadas pelas técnicas de *Regressão* e *Correlação* respectivamente, mesmo que essas técnicas estão relacionadas têm propósitos e interpretações diferentes como será mostrado mais adiante.

O termo *regressão* foi introduzido pelo cientista inglês Francis Galton em 1880. Em um famoso ensaio, Galton verificou que embora houvesse uma tendência de pais altos terem filhos altos e pais baixos terem filhos baixos, a altura média dos filhos de uma dada altura tendia a se deslocar ou "regredir"até a altura média da população como um todo. Em outras palavras, a altura dos filhos de pais extraordinariamente altos ou baixos tende a se mover para a altura média da população. A *lei de regressão universal* de Galton foi confirmada por seu amigo Karl Pearson, que coletou mais de mil registros de alturas dos membros de uma família. Ele verificou que a altura média dos filhos de um grupo de pais altos era inferior à altura de seus pais e que a altura média dos filhos de um grupo de pais baixos era superior à altura de seus pais. Assim, tanto os filhos altos quanto os baixos "regrediram"em direção à altura média de todos os homens.

A moderna interpretação da regressão é, porém, bem diferente. Em linha gerais, podemos dizer: a *análise de regressão* ocupa-se do estudo da dependência de uma variável, a *variável dependente* (ou variável resposta), em relação a uma ou mais variáveis, as *variáveis explicativas* (ou variáveis independentes), com o objetivo de estimar e/ou prever a média (da população) ou valor médio da variável dependente em termo dos valores conhecidos ou fixos das variáveis explicativas.

A *análise de correlação*, por outro lado, consiste na medição do grau ou intensidade de associação entre duas variáveis. Quando se pode demonstrar que a variação de uma variável está de algum modo associada com a variação da outra, então podemos dizer que as duas variáveis estão *correlacionadas.* Uma correlação pode ser positiva (quando ao aumentar uma variável a outra também aumenta) ou negativa (quando ao aumentar uma variável a outra diminui). Por outro lado, se a variação de uma variável não corresponde em absoluto à variação da outra, então não existe nenhuma associação e portanto, nenhuma correlação entre as duas variáveis. Assim por exemplo, se um investigador deseja determinar o grau de associação que existe entre a biomassa do fitoplacton e a quantidade de clorofila "x", o investigador retira repetidas amostras de água do lugar amostrado na lagoa e mede a clorofila "x"e a biomassa em cada amostragem. Nessa situação, o investigador não tem controle sobre uma ou outra variável, já que os valores de clorofila "x"e da biomassa encontrados em cada amostra serão "os que natureza lhe provê". Portanto, deduz-se que as duas variáveis acima são aleatórias e a análise de correlação é o procedimento estatístico adequado.

6.2 Análise de Regressão

Conforme foi apresentado na seção anterior, a análise de regressão é uma das técnicas estatísticas mais utilizadas para pesquisar e modelar o relacionamento existente entre duas ou mais variáveis. O estudo da análise de regressão será iniciado considerando o exemplo abaixo.

Exemplo 6.2.1 *Um administrador de uma cadeia de supermercados deseja desenvolver um modelo com a finalidade de estimar as vendas médias semanais (em milhares de dólares) de cada supermercado, Para isto, selecionou-se uma amostra aleatória de 20 supermercados entre todos os que formam a cadeia. Ao desenvolver o modelo foi considerado entre outras variáveis explicativas (ou independentes) a variável "o número de clientes por semana."Os dados são apresentados na tabela 6.1:*

Tabela 6.1: Número de clientes e vendas semanais para uma amostra de 20 supermercados.

Supermercado	N^o de clientes (X)	Vendas semanais (Y)
1	907	11,20
2	926	11,05
3	506	6,84
4	741	9,21
5	789	9,42
6	889	10,08
7	874	9,45
8	510	6,73
9	529	7,24
10	420	6,12
11	679	7,63
12	872	9,43
13	924	9,46
14	607	7,64
15	452	6,92
16	729	8,95
17	794	9,33
18	844	10,23
19	1010	11,77
20	621	7,41

Na figura 6.1, é apresentado o **diagrama de dispersão** das vendas semanais e o número de clientes. O diagrama é somente um gráfico em que cada par (x_i, y_i) está representado como um ponto no sistema de coordenadas bidimensionais. A análise desse diagrama indica que uma curva não passa exatamente por todos os pontos, mas existe uma forte evidência que os pontos estão dispersos de maneira aleatória em torno de uma linha reta. Portanto, é razoável supor que a média da variável aleatória Y, está relacionada com X pela seguinte relação

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

onde β_o e β_1, são respectivamente, o intercepto e a inclinação da reta e recebem o nome de coeficientes de regressão. Mesmo que a média de Y seja uma função linear de X, o valor observado de y não cai de maneira exata sobre a reta. A maneira apropriada para generalizar este fato como um modelo probabilístico linear, é supor que o valor esperado de Y seja uma função linear, mas, para um valor fixo de X o valor real de Y será determinado pelo valor médio da função linear ($\mu_{Y|x}$) mais um termo que representa um erro aleatório, assim:

$$Y = \mu_{Y|x} + \varepsilon = \beta_0 + \beta_1 x + \varepsilon, \tag{6.1}$$

onde ε é o erro aleatório. É importante observar que ε leva em conta a falha desse modelo em se ajustar exatamente aos dados. Isso pode ser devido ao efeito de outras variáveis que afetam as vendas semanais. O modelo (6.1) recebe

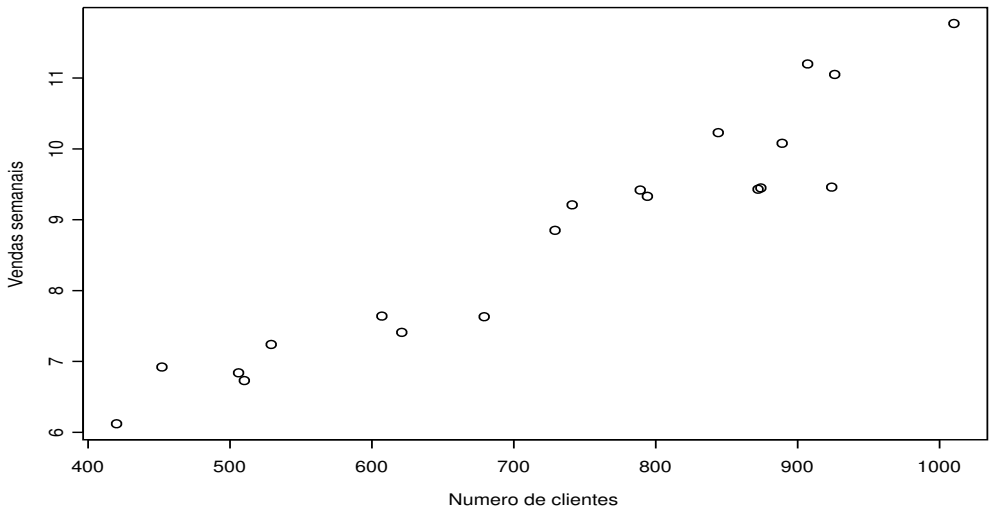


Figura 6.1: Diagrama de dispersão das vendas semanais e o número de clientes

o nome de **modelo de regressão linear simples**, pois tem somente uma variável explicativa ou variável regressora ou variável independente . Em muitas situações, os modelos desse tipo surgem de uma relação teórica. Em outras, não há nenhum conhecimento teórico da relação existente entre x e y. A seleção dos modelos se baseia na análise do diagrama de dispersão, tal como foi feito com os dados de vendas semanais. Nesses casos, o modelo de regressão se considera como um **modelo empírico**.

Em geral, a variável resposta pode estar relacionada com k variáveis explicativas $X_1, \ldots X_k$ obedecendo à equação :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon, \tag{6.2}$$

Em nosso exemplo, as variáveis $X_1, \ldots X_k$ poderia ser, por exemplo, número de promoções por semana, formas de pagamento e outras.

A equação (6.2) é denominada modelo de regressão linear múltipla, porque envolve mais uma variável explicativa. O adjetivo "linear"é usado para indicar que o modelo é linear nos parâmetros $\beta_1, \ldots, \beta_k$ e não porque Y é função linear dos X's. Por exemplo, uma expressão da forma $Y = \beta_o + \beta_1 \log X_1 + \beta_2 X_2^3 + \varepsilon$ é um modelo de regressão linear múltipla, mas o mesmo não acontece com a equação $Y = \beta_0 + \beta_1 X_1^{\beta_2} + \beta_3 X_2^2 + \varepsilon$.

Na seção seguinte é apresentado o caso mais simples em que apenas duas variáveis estarão envolvidas, o qual corresponde à regressão linear simples.

6.3 Modelo de Regressão Linear Simples

Conforme foi mencionado anteriormente, um modelo de regressão linear simples (MRLS) descreve uma relação entre uma variável independente (explicativa ou regressora) X e uma variável dependente (resposta) Y, nos termos seguintes:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{6.3}$$

onde β_0 e β_1 são constantes (parâmetros) desconhecidas e ε é o erro aleatório dado pela diferença entre o valor observado Y e a média de Y.

Como é mostrado na equação (6.3), os erros considerados no MRLS incidem diretamente sobre os valores observados de Y; a teoria da regressão assenta nas seguintes suposições:

1. Os erros têm média zero e a mesma variância desconhecida, σ^2.
2. Os erros são não correlacionados, ou seja, o valor de um erro não depende de qualquer outro erro.
3. A variável explicativa X é controlada pelo experimentador e é medida sem erro, ou seja, não é uma variável aleatória.
4. Os erros tem distribuição normal.

Se as suposições (1)-(4) se verificarem, atendendo à relação na equação (6.3), a variável dependente Y é uma variável aleatória com distribuição normal com variância σ^2 e média $\mu_{Y|x}$, sendo

$$E(Y|X = x) = \mu_{Y|x} = \beta_0 + \beta_1 x. \tag{6.4}$$

Observe em (6.4) que para um acréscimo de uma unidade em X há um acréscimo de β_1 unidades na média de Y. Se os valores de X incluem $X = 0$, então o intercepto β_0 é a média de Y quando $X = 0$. Em caso contrário, β_0 não tem interpretação prática.

6.3.1 Estimação dos parâmetros do MRLS através do método de mínimos quadrados

Suponha que tem-se n pares de observações $(x_1, y_1), \ldots, (x_n, y_n)$. A figura 6.2, mostra uma representação gráfica dos dados observados e um candidato para a linha de regressão. As estimações de β_0 e β_1 devem dar como resultado uma linha que (em algum sentido) se "ajuste melhor"aos dados. O cientista alemão Karl Gauss (1777-1855) propôs estimar os parâmetros de β_0 e β_1 de equação (6.3) de modo que se minimize a soma de quadrados dos desvios verticais da figura 6.2.

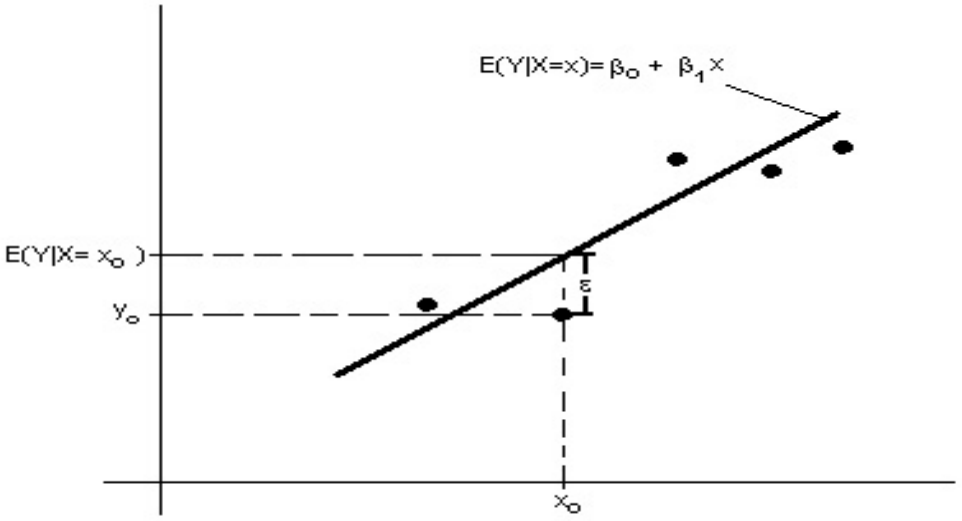


Figura 6.2: Desvio dos dados do modelo de regressão linear $n = 5$.

Este critério de estimação dos coeficientes de regressão é conhecido como **método de mínimos quadrados**. Ao utilizar o modelo (6.3), é possível expressar as n observações da amostra como:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n. \tag{6.5}$$

E a soma de quadrados dos desvios das observações em relação à linha de regressão é:

$$Q = \sum^{n} \varepsilon_i^2 = \sum^{n} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{6.6}$$

Os estimadores de mínimos quadrados (EMQ) de β_0 e β_1 denotados por $\hat{\beta}_0$ e $\hat{\beta}_1$ devem satisfazer as seguintes equações:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0}|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \\ \frac{\partial Q}{\partial \beta_1}|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0. \end{aligned} \tag{6.7}$$

Após simplificar as expressões anteriores, tem-se:

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 x_i^2 &= \sum_{i=1}^{n} x_i y_i. \end{aligned} \tag{6.8}$$

As equações (6.8) recebem o nome de equações normais de mínimos quadrados. A solução dessas equações fornece os EMQ, $\hat{\beta}_0$ e $\hat{\beta}_1$, dados por:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{6.9}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}. \tag{6.10}$$

onde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Portanto, a linha de regressão estimada ou ajustada é :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

e estima a média da variável dependente para um valor da variável explicativa $X = x$, $\mu_{Y|x}$. Note que cada par de observações satisfaz a relação:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n$$

onde $e_i = y_i - \hat{y}_i$ recebe o nome de **resíduo**. O resíduo descreve o erro no ajuste do modelo na i-ésima observação. Nesta seção, utilizamos os resíduos para o estudo da adequação do modelo ajustado.

Conforme o ponto de vista da notação, em certas situações é conveniente ter notações especiais no MRLS. Dados $(x_1, y_i), \ldots, (x_n, y_n)$ sejam:

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \\ S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}, \\ S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})y_i = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2. \end{aligned}$$

Os EMQ de β_0 e β_1 em termos da notação acima são:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}.$$

Exemplo 6.3.1 *Considere os dados do exemplo 6.2.1, apresentado ao inicio desta seção, no qual o gerente de supermercado estava interessado em estimar as vendas médias semanais de cada supermercado, dado o número de clientes por cada supermercado.*

Conforme já visto na figura 6.1, existe indicação da existência de um relacionamento linear entre as vendas semanais (Y) e o número de clientes (X) dos supermercados. Para determinar o modelo de regressão estimada foram calculados as seguintes quantidades:

$$\begin{aligned}
n &= 20 \\
\sum_{i=1}^{n} x_i &= 907 + 926 + \cdots + 621 = 14.623; \;\; \bar{x} = 731,15 \\
\sum_{i=1}^{n} y_i &= 11,20 + 11,05 + \cdots + 7,41 = 176,11; \;\; \bar{y} = 8,8055 \\
\sum_{i=1}^{n} x_i^2 &= (907)^2 + (926)^2 + \cdots + (621)^2 = 11.306.209 \\
\sum_{i=1}^{n} y_i^2 &= (11,20)^2 + (11,05)^2 + \cdots + (7,41)^2 = 1.602,0971 \\
\sum_{i=1}^{n} x_i y_i &= (907)(11,20) + (11,05)(926) \cdots + (7,41)(621) = 134.127,90 \\
S_{xx} &= \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2 = 11.306.209 - 20(731,15)^2 = 614.603 \\
S_{xy} &= \sum_{i=1}^{n} x_i y_i - n(\bar{x})(\bar{y}) = 134.127,90 - 20(8,8055)(731,15) = 5.365,08 \\
S_{yy} &= \sum_{i=1}^{n} y_i^2 - n(\bar{y})^2 = 1.609,0971 - 20(8,8055) = 51,3605.
\end{aligned}$$

Os EMQ dos parâmetros do MRLS são:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{5.365,08}{614.603} = 0,00873; \;\; \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 8,8055 - (0,00873)(731,15) = 2,423.$$

Portanto, a linha de regressão ajustada ou estimada para esses dados são:

$$\hat{y} = 2,423 + 0,00873x. \tag{6.11}$$

O gráfico desse modelo aparece na figura 6.3, junto com os dados da amostra.

A estimativa do coeficiente de regressão $\hat{\beta}_1$ foi 0,00873. Isto significa que, para cada incremento de uma unidade de X, estimamos que o valor da média de Y aumenta em 0,00873 unidades. Isto é, para cada incremento de um cliente, o modelo prevê uma estimação de um aumento nas vendas de 0,00873 mil dólares (ou 8,73 dólares). Portanto, para cada 100 clientes, esperamos que as vendas semanais aumentem, em média $ 873 dólares.

A estimativa do intercepto $\hat{\beta}_0$ foi de 2,423 mil dólares. Essa estimativa representa o valor médio Y, quando $X = 0$. Como é improvável que o número de clientes seja zero, esse valor pode ser visto como a proporção média das vendas semanais que variam em relação a fatores diferentes ao número de clientes.

Se o modelo de regressão ajustado aos dados (6.11) for aceitável, pode ser usado para prever os valores futuros da venda semanal.

Por exemplo, suponha que tem-se interesse em prever as vendas semanais para um supermercado com 600 clientes. No modelo de regressão ajustado em (6.11), é feito $X = 600$ e tem-se:

$$\hat{y} = 2,423 + (0,00873)(600) = 7,661.$$

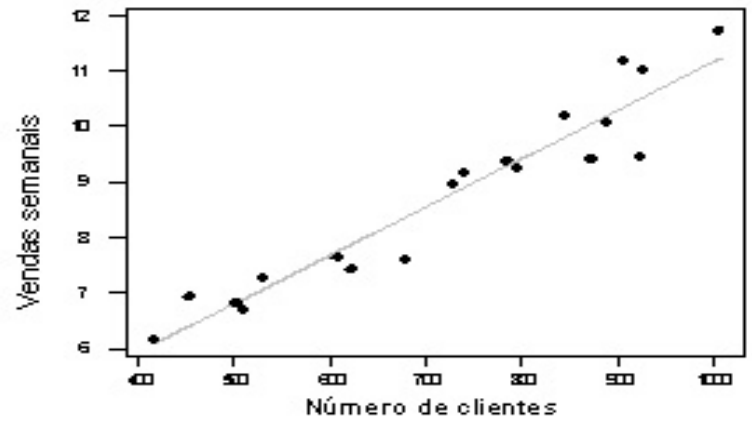


Figura 6.3: Gráfico de dispersão da venda semanal e o número de clientes, e o modelo de regressão ajustado: $\hat{y} = 2,423 + 0,00873x$

A venda semanal de 7,661 mil dólares pode ser interpretada com uma estimação da venda média semanal verdadeira dos supermercados com $X = 600$ clientes, ou como uma estimação de uma futura venda de um supermercado quando o número de clientes for $X = 600$. Claro que essas estimações estão sujeitas a um erro, isto é, é pouco provável que uma venda futura seja exatamente 7661 dólares quando o número de clientes do supermercado seja 600. Em seções subseqüentes, será visto como utilizar os intervalos de confiança e as previsões para descrever o erro ao fazer estimações a partir do modelo de regressão.

6.3.2 Propriedades dos estimadores de mínimos quadrados de β_0 e β_1 e a estimação de σ^2

Supondo que as suposições do modelo de regressão sejam válidas é possível demonstrar as seguintes propriedades:

$$E(\hat{\beta}_1) = \beta_1 \tag{6.12}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{6.13}$$

$$E(\hat{\beta}_0) = \beta_0 \tag{6.14}$$

$$Var(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]. \tag{6.15}$$

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \bar{x}}{S_{xx}} \tag{6.16}$$

Para realizarmos inferências com relação aos parâmetros do MRLS β_0 e β_1, é necessário estimar o parâmetro σ^2 que aparece nas expressões de $Var(\hat{\beta}_0)$ e $Var(\hat{\beta}_1)$. O parâmetro σ^2, que é a variância do termo aleatório ε no MRLS, reflete a variação aleatória ao redor da verdadeira linha de regressão.

Os resíduos, $e_i = y_i - \hat{y}_i$ são empregados na estimação de σ^2. A soma de quadrados residuais ou soma de quadrados dos erros, denotado por SQR é:

$$SQR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Pode-se demonstrar que o valor esperado da soma de quadrados dos residuais SQR, é dado por:

$$E(SQR) = (n-2)\sigma^2$$

Portanto,

$$\hat{\sigma^2} = \frac{SQR}{n-2}, \tag{6.17}$$

é um estimador não viciado de σ^2, isto é, $E(\hat{\sigma^2}) = \sigma^2$. A quantidade $\frac{SQR}{n-2}$ é denominado quadrado médio residual (QMR).

Uma fórmula mais conveniente para o cálculo da SQR é dada por:

$$SQR = S_{yy} - \hat{\beta}_1 S_{xy}. \tag{6.18}$$

Exemplo 6.3.2 *Com os dados do exemplo 6.3.1, é feita a estimação da variância σ^2. Nesse caso, $S_{yy} = 51,3605$, $S_{xy} = 5.365,08$ e $\hat{\beta}_1 = 0,00873$.*

Portanto, da equação (6.17),

$$\begin{aligned} \hat{\sigma^2} &= \frac{SQR}{n-2} \\ &= \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} \\ &= \frac{51,3605 - (0,00873)(5.365,08)}{20-2} = 0,2513. \end{aligned}$$

A estimativa de σ ($\hat{\sigma}^2 = 0,2513$) poderia ser utilizada na estimação da equação (6.13) para ter uma estimativa da variância do estimador do coeficiente de inclinação, e também na equação (6.15) para estimar a variância do intercepto. As raízes quadradas dos estimadores de variância resultantes se conhecem como **erros padrões estimados** da inclinação e do intercepto, respectivamente.

Definição 6.3.1 *No modelo de regressão linear simples, o* **erro padrão estimado** *da inclinação é dado por:*

$$EP(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

e o erro padrão do intercepto é dado por:

$$EP(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right]}$$

onde $\hat{\sigma}^2$ é calculada com a equação (6.17),

6.3.3 Teste de hipóteses em regressão linear simples

Um parte importante ao avaliar a adequação de um MRLS é o teste de hipóteses sobre os parâmetros do modelo e a construção de certos intervalos de confiança. Nessa seção são apresentados o procedimentos de teste de hipóteses e métodos para construir intervalos de confiança. Para realizar testes é necessário que a suposição dos erros serem independentes e identicamente distribuídos normalmente com média zero e variância σ^2 ($\varepsilon_i \sim NID(0, \sigma^2)$) seja válida. Na próxima subseção será discutido como a validade dessa suposição pode ser verificada através da análise de resíduos.

Teste de hipóteses sobre β_1 e β_0

Suponha que se deseje testar a hipótese de que a inclinação é igual a uma constante representada por $\beta_{1,0}$. As hipóteses apropriadas são:

$$\begin{aligned} H_0 : \beta_1 &= \beta_{1,0} \\ H_1 : \beta_1 &\neq \beta_{1,0} \end{aligned} \tag{6.19}$$

onde é considerada uma alternativa bilateral. Mas se os $\varepsilon \sim N(0, \sigma^2)$, de maneira imediata é possível demonstrar que a variável $Y_i \sim NID(\beta_0 + \beta_1 x_i,\ \sigma^2)$. Da equação (6.10) observa-se que $\hat{\beta}$ é uma combinação linear de variáveis aleatórias normais independentes e conseqüentemente, $\hat{\beta}_1 \sim N(\beta_1; \sigma^2/S_{xx})$. Além disso, $(n-2)\hat{\sigma}^2/\sigma^2$ tem distribuição qui-quadrado com $n-2$ graus de liberdade e $\hat{\beta}_1$ é independente de $\hat{\sigma}^2$. Como resultado destas propriedades, a estatística

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}}, \tag{6.20}$$

tem distribuição t-Student com $n-2$ graus de liberdade sob $H_0 : \beta_1 = \beta_{1,0}$. Rejeita-se H_0 se

$$|T_{obs}| > t_{\alpha/2,\ n-2}$$

onde T_{obs} é calculado a partir da equação (6.20).

Um procedimento similar pode ser utilizado para testar hipóteses sobre o intercepto. Para testar

$$\begin{aligned} H_0 : \beta_0 &= \beta_{0,0} \\ H_1 : \beta_0 &\neq \beta_{0,0} \end{aligned} \tag{6.21}$$

usamos a estatística

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}} \tag{6.22}$$

que tem distribuição t-Student com $n-2$ graus de liberdade. Rejeitamos a hipóteses nula se $|T_{obs}| > t_{\alpha/2,\ n-2}$.

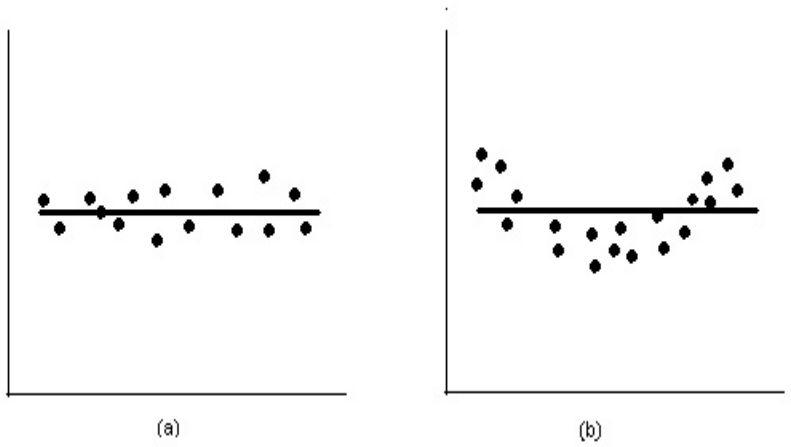


Figura 6.4: A hipótese $H_0 : \beta_1 = 0$ não é rejeitada.

Um caso particular muito importante das hipóteses dadas em (6.19) é:

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0 \end{aligned} \tag{6.23}$$

Esse teste está relacionado com a significância do modelo de regressão. Deixar de rejeitar $H_0 : \beta_1 = 0$ é equivalente a concluir que não há nenhuma relação linear entre X e Y. Na figura 6.4, é ilustrada essa situação. Note que esse resultado pode implicar que X é pouco importante para explicar a variação Y e o melhor estimador de Y para qualquer X é $\hat{Y} = \bar{Y}$ (figura 6.4a), ou que a verdadeira relação entre X e Y não é linear (figura 6.4b). Como alternativa, se $H_0 : \beta_1 = 0$ é rejeitado, implica que X tem importância ao explicar a variabilidade de Y (veja a figura 6.5). Contudo, a rejeição de $H_0 : \beta_1 = 0$ pode significar que o modelo linear é adequado (figura 6.5a), ou que, mesmo havendo um efeito linear de X, melhores resultados podem ser obtidos com a adição de termos polinomiais de ordem maior em X (figura 6.5b).

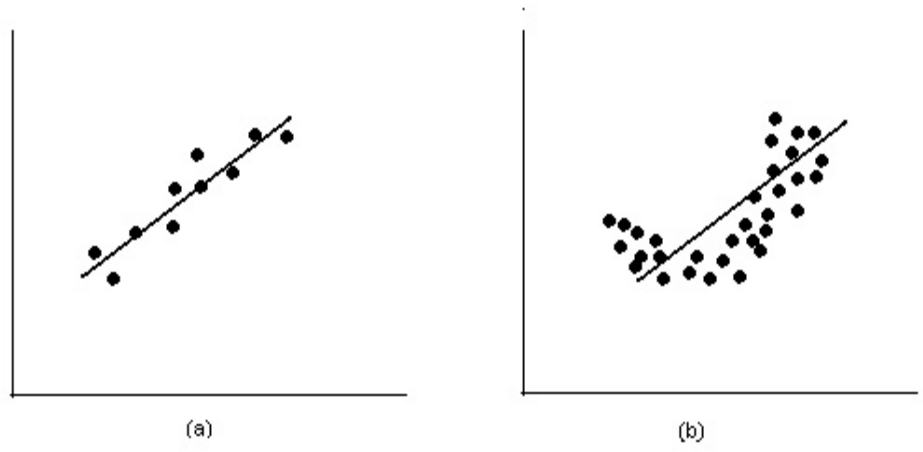


Figura 6.5: A hipótese $H_0 : \beta_1 = 0$ é rejeitada.

Exemplo 6.3.3 *Aqui é apresentado o teste de significância para o MRLS para os dados do exemplo 6.3.1.*

As hipóteses são

$$\begin{aligned} H_0 : \beta_1 &= 0 \\ H_1 : \beta_1 &\neq 0. \end{aligned}$$

Com $\alpha = 0,05$. Dos exemplos 6.3.1 e 6.3.2, tem-se:

$$\hat{\beta}_1 = 0,00873, \;\; n = 20 \;\; S_{xx} = 614,603, \;\; \hat{\sigma}^2 = 0,2512,$$

De modo que a estatística de teste, dada em (6.22), é:

$$T_{obs} = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{0,00873}{\sqrt{0,2513/614.603}} = 13,65.$$

Como $T_{obs} = 13,65 > t_{0,03,18} = 2,101$, rejeita-se a hipótese $H_0 : \beta_1 = 0$. Portanto, conclui-se ao nível de significância de 5%, que existe uma relação linear significativa entre o número de clientes e as vendas semanais.

Análise de variância para o teste de $H_0 : \beta_1 = 0$

Para testar a significância do modelo de regressão ($H_0 : \beta_1 = 0$,) pode-se utilizar o método conhecido como **análise de variância**. O método consiste em decompor a variabilidade da variável resposta em componentes mais manejáveis. Considere a seguinte *identidade*:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i - \bar{Y} + \hat{Y}_i) \tag{6.24}$$

Elevando ao quadrado a igualdade e somando as n observações em (6.24) vem:

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i-(\bar{Y}-\hat{Y}_i))^2 \\ &= \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2. \end{aligned} \tag{6.25}$$

Os dois componentes do membro direito da equação (6.25) medem, respectivamente, a quantidade de variabilidade em Y_i, explicada pela linha de regressão e variação residual que não é explicada pela linha de regressão. É usual chamar a $SQR=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ de *soma de quadrados dos residuais* e $SQreg=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$, *soma de quadrados da regressão*. Portanto, a equação (6.25) pode ser escrita como:

$$S_{yy}=SQreg+SQR \tag{6.26}$$

onde $S_{yy}=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$ é a soma de quadrados total de Y, representando por SQT. Comparando a equação (6.26) com a equação (6.18), observa-se que a soma de quadrados devido à regressão $SQreg$ é:

$$SQreg=\hat{\beta}_1 S_{xy}. \tag{6.27}$$

Pode-se mostrar que a soma de quadrado total, SQT, tem $n-1$ graus de liberdade e, SQR e $SQreg$ têm respectivamente 1 e $n-2$ graus de liberdade.

Também é possível demonstrar que:

$$\begin{aligned} E\left[\frac{SQreg}{1}\right] &= \sigma^2+\beta_1^2 S_{xx}, \\ E\left[\frac{SQR}{n-2}\right] &= \sigma^2 \end{aligned}$$

e que $SQreg/\sigma^2$ e SQR/σ^2 são variáveis aleatórias qui-quadrado independentes com 1 e $n-2$ graus de liberdade respectivamente. Portanto, se a hipótese nula $H_0:\beta_1=0$ é verdadeira, a estatística

$$F=\frac{SQreg/1}{SQR/(n-2)}=\frac{QMreg}{QMR}, \tag{6.28}$$

tem distribuição F com 1 e $(n-2)$ graus de liberdade. Portanto, rejeita-se H_0 se $F_{0bs}>F_{\alpha,\,1,\,n-2}$. As quantidades $QMreg=SQreg/1$ e $QMR=SQR/(n-2)$ são denominadas respectivamente **quadrado médio devido à regressão** e **quadrado médio devido aos residuais**. O procedimento do teste é usualmente representado em uma tabela de análise de variância, como mostrada na tabela 6.2 abaixo.

Tabela 6.2: Análise de variância para o teste de $H_0:\beta_1=0$

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	$SQreg=\hat{\beta}_1 S_{xy}$	1	$QMreg$	$QMreg/QMR$
Residual	$SQR=SQT-SQreg$	$n-2$	QMR	
Total	SQT	$n-1$		

Exemplo 6.3.4 *A seguir é apresentado o procedimento de análise de variância para testar se de fato existe relação linear entre o número de clientes (X) e as vendas semanais (Y), no modelo proposto para os dados do exemplo 6.3.1. (Use $\alpha=0,05$)*

Relembre que $S_{yy} = 51,3605$, $\hat{\beta}_1 = 0,00873$, $S_{xy} = 5.365,08$ e $n = 20$. A soma de quadrados da regressão é

$$SQreg = \hat{\beta}_1 S_{xy} = (0,00873)(5.365,08) = 46,8371$$

enquanto a soma de quadrados dos residuais é:

$$SQR = SQT - \hat{\beta}_1 S_{xy} = 51,3605 - 46,8371 = 4,5234$$

Na tabela 6.3, é apresentado um resumo da análise de variância para testar $H_0 : \beta_1 = 0$. Nesse caso, a estatística de teste é $F_{0bs} = QMreg/QMR = 46,837148/0,2512 = 186,4536$. Como $F_{obs} = 186,4536 > F_{0,05,1,18} = 4,41$ rejeita-se H_0, ao nível de significância de 5%.

Tabela 6.3: Análise de variância para o teste de $H_0 : \beta_1 = 0$ do exemplo 6.3.1

Fonte de variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	F
Regressão	46,8371	1	46,8371	186,4536
Residual	4,5234	18	0,2513	
Total	51,3605	19		

Note, que o procedimento de análise de variância para testar a significância da regressão é equivalente o teste t dada no início desta seção. Portanto, qualquer desses procedimentos conduz às mesmas conclusões. Não é difícil demonstrar que a estatística do teste T da equação (6.20), com $\beta_{1,0} = 0$,

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}}, \tag{6.29}$$

é equivalente ao teste F da equação (6.28). Elevando ao quadrado ambos membros da equação (6.29) e considerando que $\hat{\sigma}^2 = QMR$, tem-se que:

$$T^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/S_{xx}} = \frac{\hat{\beta}_1 S_{xy}}{QMR} = \frac{QMreg}{QMR}, \tag{6.30}$$

Observe que o termo T^2 da equação (6.30) é idêntico à F da equação 6.28. É verdade, em geral, que o quadrado de uma variável aleatória t-Student com ν graus de liberdade é uma variável aleatória F, com um e ν graus de liberdade no numerador e denominador , respectivamente. Portanto, o teste que utiliza T é equivalente ao teste baseado em F. Mas, o teste t é um pouco mais flexível , pois que permite testar hipóteses unilaterais, enquanto que o teste F é restrito ao teste bilateral.

6.3.4 Intervalos de confiança para β_1 e β_0

Além das estimativas pontuais para a inclinação e o intercepto da linha de regressão, é possível obter estimações por intervalos de confiança para esses parâmetros. O comprimento desses intervalos é uma medida da qualidade total da linha de regressão. Se para o MRLS é válida a suposição de que os $\varepsilon_i \sim NID(0, \sigma^2)$, então

$$(\hat{\beta}_1 - \beta_1)/\sqrt{QMR/S_{xx}} \text{ e } (\hat{\beta}_0 - \beta_0)/\sqrt{QMR[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}$$

são variáveis aleatórias com distribuição t-Student com $n-2$ graus de liberdade. Isso conduz à seguinte definição de intervalo de $100(1-\alpha)\%$ de confiança para a inclinação β_1 :

$$IC(\beta_1; 1-\alpha) = \left(\hat{\beta}_1 - t_{\frac{\alpha}{2},\, n-2}\sqrt{\frac{QMR}{S_{xx}}} \; ; \; \hat{\beta}_1 + t_{\frac{\alpha}{2},\, n-2}\sqrt{\frac{QMR}{S_{xx}}} \right) \tag{6.31}$$

De modo similar, um intervalo de $100(1-\alpha)\%$ de confiança para a inclinação β_0 é dado por:

$$IC(\beta_0; 1-\alpha) = \left(\hat{\beta}_0 - t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]}\ ;\ \hat{\beta}_0 + t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}]} \right) \tag{6.32}$$

Exemplo 6.3.5 *A seguir é obtido um intervalo de 95% de confiança para a inclinação do MRLS com os dados do exemplo 6.3.1,*

Relembre que $n = 20$, $\hat{\beta}_1 = 0,00873$, $S_{xx} = 614,603$ e $QMR = 0,2513$. Para $1-\alpha = 0,95$, tem-se $t_{0,025,\,18} = 2,101$. Então da equação (6.31), vem:

$$\begin{aligned} IC(\beta_1; 0,95) &= \left(\hat{\beta}_1 - t_{0,025,18}\sqrt{\frac{QMR}{S_{xx}}}\ ;\ \hat{\beta}_1 + t_{0,025,18}\sqrt{\frac{QMR}{S_{xx}}} \right) \\ &= \left(0,00873 - 2,101\sqrt{\frac{0,2513}{614.603}}\ ;\ 0,00873 + 2,101\sqrt{\frac{0,2513}{614.603}} \right) \\ &= (0,00873 - 0,00134; 0,00873 + 0,00134) \end{aligned}$$

Ou seja,

$$IC(\beta_1; 0,95) = (0,00739; 0,01007).$$

6.3.5 Intervalo de confiança para a resposta média

Também é possível construir intervalos de confiança para a resposta média correspondente a um valor especificado da variável explicativa, que representaremos por x_0. Ou seja, o interesse consiste em estimar um intervalo de confiança para $E(Y|X = x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$. Um estimador pontual de $\mu_{Y|x_0}$ pode ser obtido a partir do modelo de regressão ajustado

$$\hat{\mu}_{Y|x_o} = \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_0.$$

Considerando que a suposição de que os $\varepsilon_i \sim NID(0, \sigma^2)$ é válida, pode-se demonstrar que $E(\hat{\mu}_{Y|x_0}) = \hat{\mu}_{Y|x_0}$. A variância de $\hat{\mu}_{Y|x_o}$ é:

$$Var(\hat{\mu}_{Y|x_o}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Além disso, $\hat{\mu}_{Y|x_o}$ tem distribuição normal. Já que $\hat{\beta}_0$ e $\hat{\beta}_1$ são normalmente distribuídos. Também podemos demonstrar que a variável aleatória

$$T = \frac{\hat{\mu}_{Y|x_o} - \mu_{Y|x_o}}{\sqrt{QMR\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$$

tem distribuição t-Student com $n-2$ graus de liberdade. Portanto, um intervalo de $100(1-\alpha)\%$ de confiança para $\mu_{Y|x_0}$ é dado

$$IC(\hat{\mu}_{Y|x}; 1-\alpha) = \left(\hat{\mu}_{Y|x_o} - t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}; \hat{\mu}_{Y|x_o} + t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]} \right) \tag{6.33}$$

Observe que o comprimento de intervalo de confiança para $\hat{\mu}_{Y|x}$ é mínimo quando $x_0 = \bar{x}$ e aumenta. à medida que $|x_0 - \bar{x}|$ aumenta.

Exemplo 6.3.6 *Para o problema dos supermercados do exemplo 6.3.1, suponha que tem-se interesse em construir um intervalo de 95% de confiança da venda, média, semanal para todos supermercados com 600 clientes.*

No modelo ajustado $\hat{\mu}_{Y|x_0} = 2,423 + 0,00873x_0$. Para $x_0 = 600$, obtém-se $\hat{\mu}_{Y|x_0} = 7,661$. Também,

$$\bar{x} = 731,15, \quad QMR = 0,2513, \quad S_{xx} = 614.603, \quad n = 20 \text{ e } 1-\alpha = 0,95 \Rightarrow t_{0,05,18} = 2,101.$$

Substituindo esses valores na equação (6.33), obtém-se o seguinte intervalo de confiança:

$$\begin{aligned} & IC(\mu_{Y|x_0}; 0,95) = \\ = & \left(7,661 - 2,101\sqrt{0,2513[\frac{1}{20} + \frac{(600-731,15)^2}{614.603}]}; 7,661 + 2,101\sqrt{0,2513\left[\frac{1}{20} + \frac{(600-731,15)^2}{614.603}\right]}\right) \\ = & (7,661 - 0,292; 7,661 + 0,292) \\ = & (7,369; 7,935). \end{aligned}$$

Portanto, a partir do intervalo construído, conclui-se, com 95% de confiança, que as vendas médias semanais poderiam variar de 7.369 dólares a 7.953 dólares para supermercados com 600 clientes.

Ao repetir os cálculos anteriores para valores diferentes de x_0, obtém-se os limites de confiança para cada $\mu_{Y|x_0}$. Na figura 6.6, é mostrado o diagrama de dispersão com o modelo de regressão ajustado e os correspondentes limites de confiança de 95% (bandas de confiança). Observe que o comprimento do intervalo de confiança para $\mu_{Y|x_0}$ aumenta a medida que $|x_0 - \bar{x}|$ aumenta.

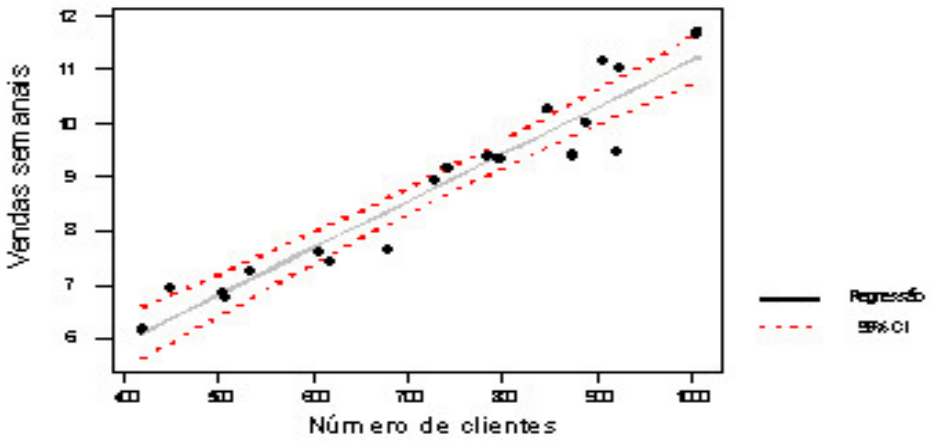


Figura 6.6: Diagrama de dispersão dos dados dos supermercados do exemplo 6.3.1, conjuntamente com a linha de regressão ajustada e as bandas de confiança do 95% para $\mu_{Y|x_0}$.

6.3.6 Previsão de novas observações

Uma aplicação muito importante de um modelo de regressão é a previsão de novas ou futuras observações de Y, (Y_0) correspondente a um dado valor da variável explicativa X, x_0, então

$$\hat{Y_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{6.34}$$

é o melhor estimador pontual de Y_0.

A nova observação Y_0 é independente das observações usadas para o desenvolvimento do modelo de regressão. Portanto, o intervalo de confiança para $\mu_{Y|x_0}$ da equação (6.33) é inadequado nesta situação, pois esse intervalo

se baseia somente nos dados utilizados para ajustar o modelo de regressão. O intervalo de confiança ao redor de $\mu_{Y|x_0}$ se refere à resposta média em x_0 (um parâmetro populacional), e não a observações futuras.

Seja Y_0 a observação futura quando $X = x_0$ e $\hat{Y_0}$, dado pela equação (6.34), o estimador pontual de Y_0. Note que o erro de previsão

$$\Psi = Y_0 - \hat{Y_0}$$

é uma variável aleatória com distribuição normal, com média zero e variância

$$\begin{aligned} Var(\Psi) &= Var(Y_0 - \hat{Y_0}) \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

dado que Y_0 é independente de $\hat{Y_0}$. Se é usado QMR como estimador de σ^2, pode-se demonstrar que

$$T = \frac{Y_0 - \hat{Y_0}}{\sqrt{QMR\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$$

tem distribuição t-Student com $n-2$ graus de liberdade. Portanto um intervalo de $100(1-\alpha)\%$ de confiança para uma futura observação é dado por:

$$IC(Y_0; 1-\alpha) = \left(\hat{Y} - t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}; \hat{Y} + t_{\frac{\alpha}{2},\, n-2}\sqrt{QMR[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}]}\right) \quad (6.35)$$

Observe que o comprimento do intervalo de confiança para a nova observação é mínimo quando $x_0 = \bar{x}$ e aumenta a medida que $|x_0 - \bar{x}|$ aumenta. Ao comparar as equações (6.35) e (6.33) observa-se que o comprimento do intervalo de predição em que $X = x_0$ é sempre maior que o comprimento do intervalo de confiança para a resposta média obtido quando $X = x_0$. Esse resultado é conseqüência do fato de que o intervalo de previsão depende tanto do erro associado ao ajuste do modelo quanto do erro associado à observação futura.

Exemplo 6.3.7 *Para ilustrar a construção de um intervalo de previsão, considere os dados do exemplo 6.3.1 e suponha agora, tem-se interesse em encontrar um intervalo de previsão de 95% das vendas semanais de um supermercado com 600 clientes.*

Considerando a equação (6.35) e os dados do exemplo 6.3.6, $\hat{Y} = 7,661$ e o intervalo de predição é:

$$\begin{aligned} IC(Y_0; 0,95) &= \left(7,661 - 2,101\sqrt{0,2513[1 + \frac{1}{20} + \frac{(600 - 731,15)^2}{614.603}]}; \right. \\ &\quad \left. 7,661 + 2,101\sqrt{0,2513\left[1 + \frac{1}{20} + \frac{(600 - 731,15)^2}{614.603}\right]}\right) \\ &= (7,661 - 1,084; 7,661 + 1,084) \\ &= (6,577; 8,745). \end{aligned}$$

Portanto, a partir do intervalo construído, conclui-se, com 95% de confiança, que as vendas médias semanais poderiam variar de 6.577 dólares a 8.745 dólares para um supermercado que tem 600 clientes.

Ao repetir os cálculos anteriores para diferentes valores de x_0, podemos obter os intervalos de previsão de 95%, que estão representados na figura 6.7. Observe que esse gráfico também apresenta os limites de confiança do 95% para $\mu_{Y|x_0}$, calculados com os dados do exemplo 6.3.1. Isto ilustra que os limites de previsão sempre são mais amplos que os limites de confiança da $\mu_{Y|x_0}$.

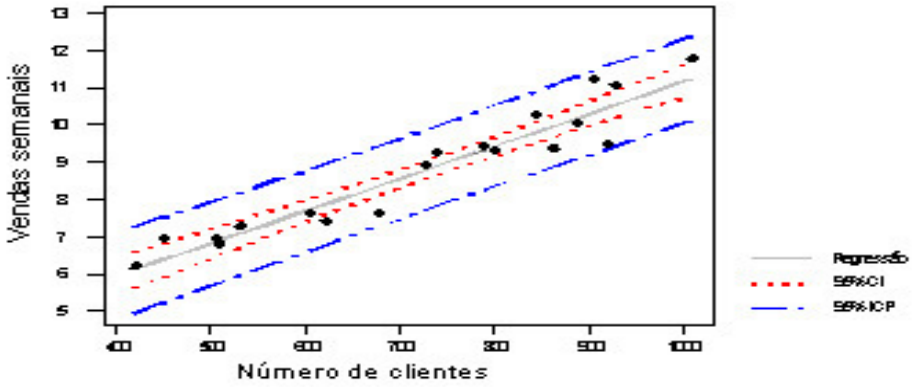


Figura 6.7: Digrama de dispersão dos dados dos supermercados do exemplo 6.3.1, conjuntamente com a linha de regressão ajustada e as bandas de confiança do 95% para $\mu_{Y|x_0}$ (CI) e Y_0 (ICP).

6.3.7 Estudo da adequação do modelo de regressão

O ajuste de um modelo de regressão requer várias suposições. A estimação dos parâmetros do modelo requer a suposição de que os erros são variáveis aleatórias não correlacionadas com média zero e variância constante. A construção de intervalos de confiança e testes de hipóteses requer que os erros sejam normalmente distribuídos. Além disso, é assumindo que a ordem do modelo é correta; isto é, se ajustamos um modelo de regressão linear simples, considera-se que o fenômeno realmente se comporta dessa forma.

O pesquisador deve sempre questionar a validade dessas suposições e realizar análises para verificar a adequação do modelo adotado. Nesta subseção serão discutidos métodos úteis para o estudo da adequação do modelo de regressão.

Análise residual

Os resíduos de um modelo de regressão são definidos como

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

onde y_i é uma observação real de Y e $\hat{y}_i$ é o valor correspondente estimado através do modelo de regressão. Freqüentemente a análise de resíduos é útil para verificar a suposição de que os erros são não correlacionados e têm uma distribuição que é aproximadamente normal com média zero e variância constante, assim como para determinar se é necessária a adição de termos adicionais ao modelo.

A análise da adequação do modelo será feita pelo gráfico de resíduos. Como uma verificação aproximada da normalidade, pode-se construir os histogramas de freqüências dos resíduos ou um gráfico de probabilidade normal dos resíduos. Muitos programas computacionais produzem gráficos de probabilidade normal dos resíduos (por exemplo, Minitab), já que, os tamanhos das amostra em um modelo de regressão geralmente são pequenos para que os histogramas sejam de utilidade por isso que o gráfico de probabilidade é o método preferido. Além desses métodos gráficos, existem procedimentos de testes para verificar a normalidade, como por exemplo o teste de aderência, teste de Shapiro-Wilk, teste de Kolgomorov, entre outras.

Também é possível padronizar os resíduos mediante o cálculo de:

$$d_i = \frac{e_i}{\sqrt{QMR}}, \quad i = 1, \ldots, n$$

Se os erros tem distribuição normal, então aproximadamente 95% dos resíduos padronizados devem pertencer ao intervalo $(-2, 2)$. Os resíduos fora desse intervalo podem indicar a presença de um valor atípico ("outlier"). Isto é, uma observação que não é comum do restante da massa de dados. Na literatura, foram propostas várias regras para descartar valores atípicos. Porém , muitas vezes, os "outliers"fornecem informações importantes sobre situações pouco usuais que são de interesse para o pesquisador e não devem ser descartadas. Para um estudo de valores atípicos, veja Montgomery e Peck, (1992).

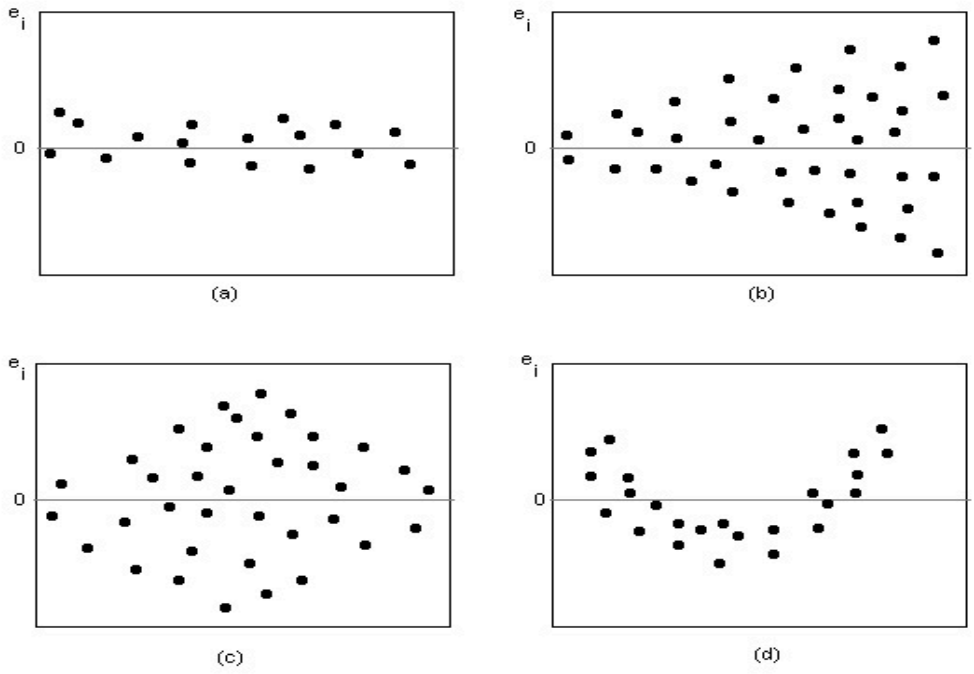


Figura 6.8: Padrões para gráficos de resíduos: (a) satisfatório, (b) funil, (c) laço duplo, (d) não linear.

Geralmente é útil fazer um gráfico dos resíduos (i) com uma seqüencia no tempo (se é conhecida); (ii) em relação aos $\hat{y}$ e (iii) em função da variável independente x. Usualmente, esses gráficos tem aspecto similar aos quatro padrões gerais que aparecem na figura 6.8. O padrão (a) dessa figura representa a situação ideal, enquanto que os padrões (b), (c) e (d) representam anomalias. Se os resíduos aparecem como em (b), a variância das observações pode aumentar com o tempo ou com a magnitude de Y ou X. Usualmente uma transformação nos dados sobre a resposta Y elimina este problema. Entre as transformações mais usadas para estabilizar a variância se inclui o emprego de: $\sqrt{y}$, $\ln y$ ou $1/y$. (veja Montgomery e Peck (1992) para mais detalhes). Se um gráfico dos resíduos com o tempo tem o aspecto da figura 6.8b, então a variância das observações aumenta com o tempo. Os gráficos dos resíduos com $\hat{y}$ ou com x, semelhantes (c) também indicam uma desigualdade da variância. Gráficos dos resíduos semelhantes ao de figura 6.8d, indicam que modelo é inadequado, isto é, que é necessário adicionar ao modelo termos de ordem superior, considerar uma transformação da variável x ou da variável y (ou ambas), ou considerar outras variáveis explicativas.

Exemplo 6.3.8 *A seguir é apresentado a análise residual para o modelo de regressão ajustado os dados de exemplo 6.3.1.*

Na tabela 6.4, são apresentados os valores observados e ajustados de Y para cada valor de x que aparece no conjunto

aos dados . Esses valores foram obtidos com o aplicativo MINITAB.

Tabela 6.4: Dados do exemplo 6.3.1, valores ajustados, resíduos e resíduos padronizados,

Supermercado	Número de clientes	Vendas Semanais	Valor Ajustado ($\hat{y}_i$)	Resíduo $e_i = y_i - \hat{y}_i$	Resíduo padronizado $d_i = e_i/\sqrt{QMR}$
1	907	11,20	10,3356	0,86438	1,72804
2	926	11,05	10,5015	0,54852	1,09658
3	506	6,84	6,8350	0,00499	0,00997
4	741	9,21	8,8865	0,32351	0,64675
5	789	9,42	9,3055	0,11449	0,22888
6	889	10,08	10,1785	-0,09848	-0,19688
7	874	9,45	10,0475	-0,59754	-1,19457
8	510	6,73	6,8699	-0,13993	-0,27974
9	529	7,24	7,0358	0,20421	0,40824
10	420	6,12	6,0843	0,03574	0,07145
11	679	7,63	8,3452	-0,71525	-1,42989
12	872	9,43	10,0301	-0,60008	-1,19965
13	924	9,46	10,4840	-1,02402	**-2,04718**
14	607	7,64	7,7167	-0,07671	-0,15335
15	452	6,92	6,3636	0,55639	1,11232
16	729	8,85	8,7817	0,06827	0,13648
17	794	9,33	9,3492	-0,01916	-0,03831
18	844	10,23	9,7856	0,44435	0,88833
19	1010	11,77	11,2348	0,53523	1,07000
20	621	7,41	7,8389	-0,42892	-0,85749

Na figura 6.9, são apresentado os gráficos da análise residual do exemplo 6.3.1. A figura 6.9a mostra um gráfico de probabilidade normal dos resíduos. Como esses resíduos estão localizados aproximadamente ao longo de uma linha reta, conclui-se que não há uma forte indicação de que a suposição de normalidade dos erros não seja adequada. Na figura 6.9b, mostra o gráfico de resíduos com os valores ajustados ($\hat{y}_i$) , enquanto na figura 6.9c, representa-se número de clientes (x_i). Nenhum desses gráficos fornecem indicação de algum problema sério quanto à adequação do modelo. Finalmente, na figura 6.9d é representado o gráfico de resíduos com os valores ajustados. O padrão do gráfico é semelhante ao da figura 6.9b. Mas, a figura 6.9d, mostra uma observação (o supermercado 13)os resíduos foram do intervalo $(-2, 2)$ o qual poderia ser considerado como um valor atípico.

Coeficiente de determinação (R^2)

A quantidade:

$$R^2 = \frac{SQreg}{SQT} = 1 - \frac{SQR}{SQT} \tag{6.36}$$

recebe o nome de **coeficiente de determinação** que é usado para julgar a adequação do modelo de regressão. Mas, no caso em que as variáveis X e Y sejam variáveis aleatórias distribuídas de maneira conjunta, R^2 é o quadrado do coeficiente de correlação entre X e Y. Da identidade da análise de variância, dadas em (6.25)-(6.26), temos que $0 \leq R^2 \leq 1$. Daí, o coeficiente de determinação pode ser interpretado como a proporção da variabilidade presente nas observações da variável resposta Y, que é explicada pela variável independente X no modelo de regressão. A proporção não explicada pela variável regressora X, recebe o nome de coeficiente de não determinação e é dada por $1 - R^2$.

Exemplo 6.3.9 *Para os dados dos supermercados do exemplo 6.3.1, determinar* R^2.

Da equação (6.36) tem-se:

$$R^2 = \frac{SQreg}{SQT} = \frac{46,8371}{51,3605} = 0,912$$

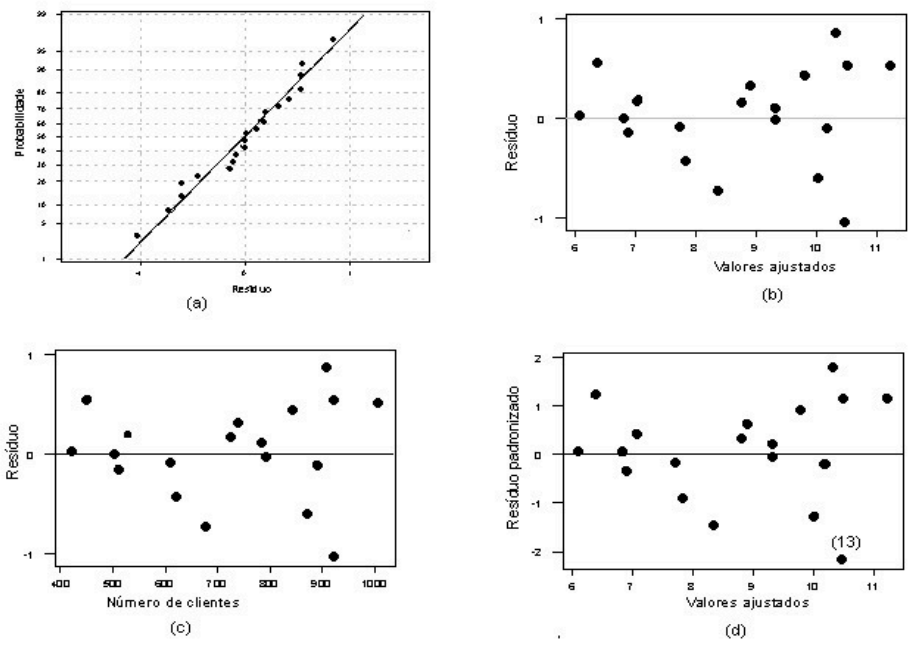


Figura 6.9: Gráfico de resíduos (e_i) para o exemplo 6.3.1 : (a) de probabilidade normal, (b) e_i contra $\hat{y}_i$ (c) e_i contra x_i e (d) resíduos padronizados (d_i) contra $\hat{y}_i$.

Esse resultado significa que o modelo ajustado explicou 91,2% da variação na variável resposta Y (vendas semanais). Isto é, 91,2% da variabilidade de Y é explicada pela variável regressora X (número de clientes).

A estatística R^2 deve ser empregada com cuidado, já que sempre é possível fazer R^2 igual a *um* mediante a adição ao modelo de um número suficiente de termos. Por exemplo, podemos obter um ajuste "perfeito"para os n pontos com um polinômio de grau $n-1$. Além disso, R^2 sempre aumenta por meio da adição de novas variáveis explicativas, o que não implica, necessariamente em que o novo modelo seja melhor que o anterior. Ao menos que a soma de quadrados dos residuais desse novo modelo esteja diminuído de uma quantidade igual ou menor que o quadrado médio residual do modelo original. Dessa forma, o novo modelo terá o quadrado médio do residual maior que o anterior, devido à perda de graus de liberdade no erro. Portanto, o novo modelo será pior que o anterior.

Existem várias idéias errôneas quanto a R^2. Em geral, R^2 não mede a magnitude da inclinação da reta de regressão. Um grande valor de R^2 não implica em um valor alto para inclinação da reta de regressão. Por outro lado, R^2 não mede a adequação do modelo, já que, isto pode ser inflacionado de maneira artificial com a adição ao modelo de termos polinomiais em X de maior ordem. A magnitude de R^2 pode ser grande mesmo que X e Y estejam relacionados de forma não linear. Por exemplo, o R^2 para a equação de regressão da figura 6.5b é relativamente grande, mesmo que a aproximação linear seja pobre. Finalmente, mesmo que R^2 seja grande, não, implica necessariamente, que o modelo de regressão proporcione previsões precisas de observações futuras.

6.4 Análise de correlação

Conforme foi mencionado no inicio deste capítulo a *análise de regressão* é usada quando tem-se interesse em estabelecer o tipo de relação que há entre uma variável dependente e uma ou mais variáveis independentes. Mas, quando tem-se interesse estabelecer o grau dessa relação é usada a *análise de correlação.*

No desenvolvimento da análise de regressão foi suposto que X seja uma variável controlada (ou fixa) e medida com erro desprezível, e que Y é uma variável aleatória. Muitas aplicações da análise de regressão envolvem situações em que tanto X quanto Y são variáveis aleatórias. Neste casos, a suposição usual é que as observações (X_i, Y_i), $i = 1, \ldots, n$ são variáveis aleatórias distribuídas de maneira conjunta obtidas da distribuição $f(x, y)$.

Por exemplo, suponha que se deseja desenvolver um modelo de regressão que relacione a resistência ao corte dos pontos de soldadura com o diâmetro dos mesmos. Nesse exemplo, não é possível controlar o diâmetro de soldadura. O que pode ser feito é selecionar ao acaso n pontos de soldadura e observar o diâmetro (X_i) e a resistência ao corte (Y_i) de cada um deles. Portanto, (X_i, Y_i) são variáveis aleatórias distribuídas de maneira conjunta.

Suponha que a distribuição conjunta de X_i e Y_i tenha uma distribuição normal bivariada cuja função de densidade é dada por

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ \frac{1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) \right] \right\} \tag{6.37}$$

onde μ_1 e σ_1^2 são a média e variância de X e μ_2 e σ_2^2 são a média e variância de Y e, ρ é *coeficiente de correlação* entre X e Y. O coeficiente de correlação é definido como:

$$\rho = \frac{E[(X-\mu_1)(Y-\mu_1)]}{\sigma_1\sigma_2} \tag{6.38}$$

O coeficiente de correlação é uma quantidade adimensional que mede a força da associação linear entre duas variáveis aleatórias.

De (6.37) é possível demonstrar que a função de densidade condicional de Y para um valor dado $X = x$ é dado por

$$f(y|x) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left\{ -\frac{1}{2}\left(\frac{y_i - \beta_0 - \beta_1 x}{\sigma^2_{Y|x}} \right)^2 \right\} \tag{6.39}$$

onde

$$\beta_0 = \mu_2 - \mu_1\rho\frac{\sigma_2}{\sigma_1}, \tag{6.40}$$

$$\beta_1 = \frac{\sigma_2}{\sigma_1}\rho \tag{6.41}$$

e a variância da distribuição condicional de Y para um $X = x$ é dado por:

$$\sigma_{Y|x} = \sigma_2^2(1-\rho^2). \tag{6.42}$$

Isto é, a distribuição condicional de Y dado $X = x$ é normal com média

$$E(Y|X = x) = \beta_0 + \beta_1 x \tag{6.43}$$

e variância $\sigma^2_{Y|x}$. Portanto, a média da distribuição condicional dado $X = x$ ($E(Y|X = x)$ é o modelo de regressão linear simples. Além disso, existe uma relação entre o coeficiente de correlação ρ e a inclinação β_1. Na equação (6.41), observe que se $\rho = 0$, existe $\beta_1 = 0$, que implica na não existência de regressão de Y sobre X. Isto é, o conhecimento de X não é suficiente para prever Y.

É possível demonstrar que os estimadores de máxima verossimilhança dos parâmetros β_0 e β_1 são:

$$\beta_0 = \bar{Y} - \hat{\beta}_1\bar{X} \tag{6.44}$$

e

$$\beta_1 = \frac{\sum_{i=1}^n Y_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{S_{XY}}{S_{XX}} \tag{6.45}$$

Note que os estimadores do intercepto e da inclinação dados acima são idênticos as equações (6,9) e (6.10) respectivamente, os quais foram obtidos pelo método de mínimos quadrados onde se supõe que a variável X é uma

variável controlável. Isto é, o modelo de regressão Y e X com distribuição conjunta normal bivariada, é equivalente ao modelo na qual X não é uma variável aleatória. Portanto, os métodos já apresentados na seção anterior podem ser empregados para análise de modelos onde X e Y são variáveis aleatórias com distribuição normal bivariada.

É possível realizar inferência sobre o coeficiente de correlação ρ desse modelo. Um estimador de ρ é o coeficiente de correlação amostral, representado por r e definido por

$$r = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \tag{6.46}$$

Das equações (6.45) e (6.46) é fácil demonstrar que:

$$\hat{\beta}_1 = \left(\frac{S_{YY}}{S_{XX}}\right)^{1/2} r. \tag{6.47}$$

Portanto, a inclinação $\hat{\beta}_1$ é igual ao coeficiente de correlação amostral r multiplicado por um fator de escala que é a raiz quadrada do quociente entre uma medida da dispersão dos valores de Y (S_{YY}) e a medida equivalente da dispersão dos valores de X (S_{XX}). No entanto, apesar de $\hat{\beta}_1$ e r estarem diretamente relacionados, eles fornecem diferentes tipos de informação. O coeficiente de correlação amostral r mede a força da associação linear entre X e Y, enquanto $\hat{\beta}_1$ mede a alteração esperada em Y quando X sofre uma variação unitária. No caso em que X não é uma variável aleatória, o coeficiente de correlação r deixa de ter sentido, uma vez que a magnitude de r depende da escolha feita para o espaçamento dos valores de X. Da equação (6.47), é possível demonstrar que:

$$r^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{S_{YY}} = \frac{SQreg}{SQT} = R^2.$$

onde R^2 é o coeficiente de determinação definido na equação (6.36). Isto é o coeficiente de determinação R^2 é igual ao quadrado do coeficiente de correlação amostral entre X e Y.

Em análise de correlação, freqüentemente, o interesse testar se o coeficiente de correlação é igual a zero, já que, $\rho = 0$ significa ausência de relacionamento linear entre Y e X. As hipóteses a serem testadas são:

$$\begin{aligned} H_0 &: \rho = 0 \\ H_1 &: \rho \neq 0. \end{aligned} \tag{6.48}$$

A estatística de teste apropriada é

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{6.49}$$

que tem distribuição t-Student com $n-2$ graus de liberdade se $H_0 : \rho = 0$ é verdadeira. Logo, a hipótese nula deverá ser rejeitada se $|T_{obs}| \leq t_{\alpha/2,\, n-2}$. Esse teste é equivalente ao teste de hipóteses $H_0 : \beta_1 = 0$, apresentado na seção anterior.

O procedimento para o teste das hipóteses

$$\begin{aligned} H_0 &: \rho = \rho_0 \\ H_1 &: \rho \neq \rho_0. \end{aligned} \tag{6.50}$$

onde $\rho_0 \neq 0$, é um pouco mais complicado. Para amostras de tamanho moderado grande ($n \geq 30$), a estatística

$$Z_r = \operatorname{arctanh} r = \frac{1}{2} \ln \frac{1+r}{1-r} \tag{6.51}$$

tem distribuição aproximadamente normal com média

$$\mu_{Z_r} = \operatorname{arctanh} \rho = \frac{1}{2} \ln \frac{1+\rho}{1-\rho}$$

e variância

$$\sigma^2_{Z_r} = (n-3)^{-1}.$$

Portanto, para testar a hipóteses $H_0 : \rho = 0$ a estatística de teste apropriada é:

$$Z = (\text{arctanh } r - \text{arctanh } \rho_0)\,(n-3)^{1/2}. \tag{6.52}$$

Se $H_0 : \rho = \rho_0$ é verdadeira, a estatística Z tem, aproximadamente, distribuição normal padrão. Portanto, H_0 deverá ser rejeitada se $|Z_{obs}| \geq z_{\alpha/2}$.

Além disso, é possível construir um intervalo aproximado de $100(1-\alpha)\%$ de confiança para o coeficiente de correlação ρ, que é dado por:

$$IC(\rho; 1-\alpha) = \left(\tanh\left[\text{arctanh } r - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]; \tanh\left[\text{arctanh } r + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right]\right), \tag{6.53}$$

onde

$$\tanh w = \frac{e^w - e^{-w}}{e^w + e^{-w}}.$$

Exemplo 6.4.1 *Suponha que se tenha interesse em medir a força da relação linear de dois produtos diferentes com relação ao preço em várias cidades do mundo. O preço de uma caixa de suco com seis latas de uma certa marca (X) e de uma libra de frango (Y) foram determinados em um supermercado localizado em uma amostra aleatória de nove cidades. Supondo que o preço da caixa de suco e de uma libra de frango são variáveis aleatórias com distribuição conjunta normal bivariada verifique se há relação linear entre X e Y. Os resultados são apresentados na tabela 6.5:*

Tabela 6.5: Preço (em dólares) de uma caixa de suco e de uma libra de frango em nove cidades.

Cidade	Caixa com seis sucos (X)	Uma libra de frango (Y)
Frankfurt	3,27	3,06
Hong Kong	2,22	2,34
Londres	2,28	2,27
Manila	3,04	1,51
México	2,33	1,87
Nova York	2,69	1,65
París	4,07	3,09
Sidney	2,78	2,36
Tokyo	5,97	4,85

Dos dados da tabela 6.5, são obtidos os valores seguintes:

$$n = 9;\ \sum_{i=1}^{n} X_i = 28,65;\ \ \bar{X} = 3,183;\ \ \sum_{i=1}^{n} X_i^2 = 28,65 = 102,66;\ \ S_{XX} = 11,4594;\ \ \sum_{i=1}^{n} Y_i = 23,00;$$

$$\bar{Y} = 2,5566;\ \ \sum_{i=1}^{n} Y_i^2 = 67,132;\ \ S_{YY} = 8,3522;\ \ \sum_{i=1}^{n} X_i Y_i = 81,854;\ \ S_{XY} = 8,6437$$

Com a equação (6.46)

$$r = \frac{8,6437}{\sqrt{(11,4594)(8,3522)}} = 0,883.$$

0 coeficiente de correlação $r = 0,883$, entre o preço de uma caixa de sucos e de uma libra de frango indica que há uma forte associação entre essas variáveis. Um maior preço da caixa de suco está associado fortemente com um

preço maior de uma libra de frango. Para verificar se essa associação é significativa, testa-se as hipóteses seguintes:

$$\begin{aligned} H_0 &: \rho = 0 \text{ (não relação linear entre } X \text{ e } Y) \\ H_1 &: \rho \neq 0 \text{ (há relação linear entre } X \text{ e } Y) \end{aligned}$$

O valor calculado para a estatística do teste foi

$$T_{obs} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0,883\sqrt{9-2}}{\sqrt{1-(0,883)^2}} = 4,98.$$

Para $\alpha = 0,05$, tem-se que $t_{0,025,7} = 2,365 < T_{obs} = 4,98$, logo, rejeita-se $H_0 : \rho = 0$ ao nível de significância de $\alpha = 5\%$. Isto é, há evidência estatística da existência de um relacionamento linear significativa entre o preços de suco e frango nas diferentes cidades.

6.5 Exercícios

1. Uma determinada peça que compõe aparelhos de ar condicionado tem sido produzida periodicamente em lotes de tamanhos variados. O fabricante deseja estudar a relação existente entre o tamanho do lote (X) e o número de horas de trabalho necessárias para a produção do lote (Y). Nos últimos 6 meses, 25 lotes foram produzidos observando-se os valores apresentados na tabela 6.6.

Tabela 6.6: Tamanho do lote e Número de horas de trabalho de 25 lotes.

Lote	Tamanho do Lote (X)	Número de horas de Trabalho (Y)	Lote	Tamanho do Lote (X)	Número de horas de Trabalho (Y)
1	80	399	14	20	113
2	30	121	15	110	435
3	50	221	16	100	420
4	90	376	17	30	212
5	70	361	18	50	268
6	60	224	19	90	377
7	120	224	20	110	421
8	80	352	21	30	273
9	100	353	22	90	468
10	50	157	23	40	244
11	40	160	24	80	342
12	70	252	25	70	323
13	90	389			

   (a) Construa o diagrama de dispersão e interprete-o.

   (b) Supondo que X e Y tenha distribuição conjunta normal bivariada, estime o coeficiente de correlação e verifique estatísticamente se existe relação linear entre as variáveis X e Y. (Use $\alpha = 0,05$)

   (c) Ajuste os dados a uma reta de regressão para a relação entre as variáveis X e Y.

   (d) Considerando a reta regressão ajustada dada no item (c). Estime o número médio de horas de trabalho para produzir um lote de 70 peças. Obtenha também uma estimativa por intervalo, de 98% de confiança.

2. É esperado que a massa muscular de uma pessoa diminua com a idade. Para estudar essa relação uma nutricionista selecionou 16 mulheres entre 40 e 79 anos, observou em cada uma delas a idade (X) e massa muscular (Y).

Tabela 6.7: Massa muscular e idade de 16 mulheres

Massa muscular	Idade	Massa muscular	Idade
82	71	65	76
91	91	84	65
100	43	116	45
68	67	76	58
87	56	97	45
73	73	100	53
78	68	105	49
80	56	77	78

   (a) Construa um diagrama de dispersão e interprete-o.

   (b) Ajuste uma linha de regressão para a relação entre as variáveis massa muscular (Y) e idade (X).

(c) Faça uma análise residual e verifique as suposições do modelo de regressão linear.

(d) Verifique estatísticamente se há relação entre X e Y. Use $\alpha = 0,05$

(e) Considerando a reta ajustada dada no item (b), estime a massa muscular média de mulheres com 50 anos idade.

(f) Estimar mediante um intervalo, um intervalo de 90% de confiança, a , massa muscular de uma mulher com 50 anos de idade.

(g) Supondo que X e Y tenha distribuição conjunta normal bivariada, estime o coeficiente de correlação.

(h) Obtenha um intervalo de 95% de confiança para o coeficiente de correlação de X e Y. O que você pode dizer ao respeito do item (d).

3. Um experimento foi feito com a finalidade de estudar a relação existente entre a densidade do óleo de milho (em gr/L) e temperatura de ebulição (em graus centígrados). Para uma amostra aleatória de 10 observações foram obtidos os seguintes resultados.

Densidade (Y)	910	915	867	908	902	875	889	899	878	869
Temperatura (X)	30	25	100	30	40	80	60	40	75	90

(a) Ajuste os dados a um modelo de regressão linear simples e interprete as estimativas dos parâmetros do modelo.

(b) Efetue a análise de variância e expresse suas conclusões com um nível de significância de 5%.

(c) Calcule e interprete o coeficiente de determinação e não determinação do modelo.

(d) Estimar, mediante um intervalo de 90% de confiança, a densidade média de óleo de milho, a uma temperatura de 60^o C. Interprete o resultado.

(e) Estimar, mediante um intervalo de 90% de confiança, a densidade de óleo de milho, a uma temperatura de 60^oC. Interprete seus resultados ? (Você poderia dizer porque o comprimento deste intervalo é maior que o item (d)).

(f) Com nível de significância de 5%, você pode afirmar, quando a temperatura é 0^o C, que a densidade média do óleo de milho é superior a 920 gr/L?

(g) Provar com $\alpha = 0,01$, se existe evidência estatística que permite afirmar que a cada incremento da temperatura em 1^o C, a densidade média de óleo de milho decresce em média mais de 0,6 gr/L.

(h) Estimar mediante um intervalo de 90% de confiança a variância da distribuição de densidades de óleo de milho, para uma temperatura de 45^oC.

(i) Para $\alpha = 0,05$, pode-se afirmar que a densidade média de óleo de milho é superior 900 gr/L, quando a temperatura é 60^oC?

(j) Supondo que Y e X tenha distribuição normal bivariada: (i) estime e interprete o coeficiente de correlação entre Y e X. (ii) Pode-se concluir para $\alpha = 0,05$, que a correlação existente entre a densidade do óleo de milho e a temperatura é diferente de -0,9 ?

4. O gerente de comercialização de uma cadeia de supermercados gostaria de determinar o efeito do espaço en estantes sobre as vendas de ração para animais de estimação. Selecionou-se uma amostra aleatória de 12 supermercados de igual tamanho e os resultados são apresentados a seguir:

(a) Construa o diagrama de dispersão e interprete-o.

(b) Supondo que existe uma relação linear entre X e Y, obtenha a linha de regressão ajustada. E interprete as estimativas do parâmetro.

(c) Faça um estudo da adequação do modelo ajustado.

(d) Ao nível de signicância de 5%, verifique se existe relação linear entre as variáveis X e Y.

(e) Considerando a reta ajustada dada no item (b), estime a venda média semanal em lojas com espaço em estantes de 8 pés .

(f) Estimar mediante um intervalo de 90%, a venda semanal de uma loja com espaço em estantes de 8 pés.

(g) Supondo que Y e X tem distribuição conjunta normal bivariada, estime e interprete o coeficiente de correlação entre Y e X.

Tabela 6.8: Espaço em estantes e vendas de ração para animais de estimação em 12 supermercados

Loja	Espaço em estantes (X) (pés)	Vendas semanais,(Y) (centos de dólares)
1	5	1,6
2	5	2,2
3	5	1,4
4	10	1,9
5	10	2,4
6	10	2,6
7	15	2,3
8	15	2,7
9	15	2,8
10	20	2,6
11	20	2,9
12	20	3,1

Bibliografia

[1] Bussab, W. O. e Morettin, P.A. (1987). *Estatística Básica*, 4ª Ed., São Paulo.

[2] Botter, D.A. , Paula, G.A., Liete, J.G. e Cordani, L.k. (1996). *Noções de Estatística*. São Paulo:IME/USP.

[3] Montgomery, D.C. e Runger, G.C. (1996) *Applied statistics and probability for engineers* John Wiley &Sons, Inc.

[4] Montgomery, D.C. e Peck, E.A. (1992) *Introduction to linear regression analysis* John Wiley &Sons, Inc.

[5] Montgomery, D.C. (1991) *Design and analysis of experiments* John Wiley &Sons, Inc.

[6] Moore David S. (1995). *The Basic Practice of Statístics*

[7] Fernadez, P.J. (1973). *Introbução à teória de probabilidades*. Rio de Janeiro: Livro Técnico.

[8] Meyer, Paul, L. (1977),*Probabilidade: aplicações à estatística* Livros técnicos e cientificos editora s.a.

[9] Peres , C.A., Saldiva, C.D. (1982). *Planejamento de Experimentos* 5^a Sinape, São Paulo,