Iniciação à Estatística

roteiro de aulas

Adair José Regazzi[1]
Carlos Henrique Osório Silva[2]
Gerson Rodrigues dos Santos[3]
Paulo César Emiliano[4]
Eduardo Campana Barbosa[5]

2020 – Viçosa, MG

[1]Prof. Titular (aposentado), DET-UFV, adairreg@ufv.br
[2]Prof. Titular, DET-UFV, chos@ufv.br
[3]Prof. Associado I, DET-UFV, gerson.santos@ufv.br
[4]Prof. Adjunto III, DET-UFV, paulo.emiliano@ufv.br
[5]Prof. Adjunto I, DET-UFV, eduardo.barbosa@ufv.br

UNIVERSIDADE FEDERAL DE VIÇOSA
CCE – DEPARTAMENTO DE ESTATÍSTICA (DET)

EST 105 – Iniciação à Estatística – 1º Semestre de 2021

APRESENTAÇÃO

Estatística é uma mistura de ciência, tecnologia e arte. É uma ciência porque possui princípios e conceitos próprios, é uma tecnologia porque gera processos produtivos e é uma arte porque depende da razão indutiva e não está livre de controvérsias.

No curso de Iniciação à Estatística apresentaremos conceitos teóricos e também aplicações em tópicos básicos da ciência Estatística. As aulas são expositivas no quadro com resolução de exercícios após a apresentação da teoria. O curso é essencialmente preparatório para outras disciplinas mais avançadas e/ou aplicadas, que visam preparar os alunos para atuarem em atividades que requerem conhecimentos de Estatística.

A EST 105 é uma disciplina de massa, no atual período são **dez turmas regulares** com um total de 472 alunos matriculados. A página do Registro Escolar informa uma demanda de 128 vagas disponíveis. Portanto, os alunos devem estar atentos às normas do curso descritas a seguir, pois tratamentos diferenciados não serão praticados.

Sumário

1 Somatório e produtório **7**

1.1 Somatório 7

1.1.1 Número de termos do somatório 8

1.1.2 Propriedades de somatório 9

1.1.3 Exemplos resolvidos 10

1.1.4 Outras aplicações da notação de somatório 12

1.1.5 Exercícios propostos com respostas 13

1.2 Produtório 14

1.2.1 Número de termos do produtório 14

1.2.2 Alguns exemplos 14

1.3 Exercícios propostos com respostas 15

2 Estatística descritiva **28**

2.1 Introdução 28

2.2 Medidas de posição 29

2.2.1 Média aritmética 29

2.2.1.1 Propriedades da média aritmética 30

2.2.1.1.1 Exemplos de aplicação das propriedades da média . 32

2.2.2 Média geométrica 33

2.2.2.1 Exemplos de aplicação da média geométrica 34

2.2.3 Média harmônica 36

2.2.4 Mediana 37

2.2.4.1 Mediana para o caso em que n é ímpar 38

2.2.4.2 Mediana para o caso em que n é par 39

2.2.5 Moda 39

2.3 Medidas de dispersão 40

2.3.1 Variância amostral 41

2.3.1.1 Propriedades da variância 41

2.3.2 Desvio padrão amostral 42

2.3.3 Coeficiente de variação 42

2.3.4 Erro-padrão da média 43

2.3.5 Amplitude total 44

2.4 Coeficiente de correlação amostral 44

2.5 Exercícios propostos com respostas 46

3 Introdução à teoria da probabilidade **57**
3.1 Introdução . . . 57
3.2 Conceitos fundamentais . . . 58
3.2.1 Modelo determinístico . . . 58
3.2.2 Modelo probabilístico . . . 59
3.2.3 Experimentos probabilísticos ou aleatórios . . . 59
3.2.4 Espaço amostral . . . 59
3.2.5 Eventos . . . 60
3.2.5.1 Eventos mutuamente exclusivos . . . 61
3.2.5.2 Operações básicas entre subconjuntos ou eventos . . . 61
3.2.5.3 Propriedades das operações básicas . . . 62
3.3 Conceitos de probabilidade . . . 63
3.3.1 Conceito clássico ou probabilidade *a priori* . . . 63
3.3.1.1 Espaço amostral finito . . . 64
3.3.1.2 Espaço amostral finito e equiprovável . . . 64
3.3.1.3 Revisão: contagem por combinação . . . 64
3.3.2 Frequência relativa ou probabilidade *a posteriori* . . . 65
3.3.3 Conceito intuitivo . . . 66
3.3.4 Conceito moderno ou axiomático . . . 66
3.3.5 Probabilidade geométrica . . . 68
3.3.5.1 Exercícios propostos com respostas . . . 70
3.4 Teoremas do cálculo de probabilidades . . . 70
3.5 Exemplos resolvidos . . . 73
3.6 Probabilidade condicional e independência estocástica . . . 77
3.6.1 Probabilidade condicional . . . 77
3.6.2 Teorema do produto das probabilidades . . . 79
3.6.3 Independência estocástica (ou probabilística) . . . 79
3.6.3.1 Eventos independentes . . . 80
3.6.3.2 Eventos mutuamente independentes . . . 80
3.6.4 Teorema da probabilidade total . . . 81
3.6.5 Teorema (ou regra) de Bayes . . . 81
3.6.5.1 Exercícios propostos com respostas . . . 82
3.7 Exemplos resolvidos . . . 83
3.8 Exercícios propostos com respostas . . . 88

4 Variáveis aleatórias **103**
4.1 Conceito . . . 103
4.2 Variável aleatória discreta (v.a.d.) . . . 104
4.2.1 Função de probabilidade . . . 104
4.2.2 Variável aleatória discreta uniformemente distribuída . . . 104
4.2.3 Exemplos resolvidos . . . 105
4.3 Variável aleatória contínua (v.a.c.) . . . 106
4.3.1 Função densidade de probabilidade . . . 106
4.3.2 Variável aleatória contínua uniformemente distribuída . . . 107
4.3.3 Exercícios propostos com respostas . . . 107

4.4 Função de distribuição acumulada $[F(x)]$ . . . 108
4.4.1 $F(x)$ para X uma v.a.d. . . . 109
4.4.2 $F(x)$ para X uma v.a.c. . . . 110
4.5 Exercícios propostos com respostas . . . 110
4.6 Variáveis aleatórias bidimensionais . . . 111
4.6.1 Introdução . . . 111
4.6.2 Definição . . . 111
4.6.3 Distribuição conjunta, distribuições marginais e condicionais . . . 112
4.6.3.1 (X, Y) é v.a.d. bidimensional . . . 112
4.6.3.1.1 Função de probabilidade conjunta de X e Y . . . 112
4.6.3.1.2 Distribuição de probabilidade conjunta . . . 112
4.6.3.1.3 Distribuições marginais . . . 112
4.6.3.1.4 Distribuições condicionais . . . 113
4.6.3.2 (X, Y) é v.a.c. bidimensional . . . 113
4.6.3.2.1 Função densidade de probabilidade conjunta . . . 114
4.6.3.2.2 Distribuições marginais . . . 114
4.6.3.2.3 Distribuições condicionais . . . 114
4.6.4 Variáveis aleatórias independentes . . . 115
4.6.5 Exemplo resolvido . . . 115
4.6.6 Exercícios propostos com respostas . . . 117
4.7 Medidas de posição de uma variável aleatória . . . 118
4.7.1 Esperança matemática . . . 118
4.7.1.1 Caso em que X é uma v.a.d. . . . 118
4.7.1.2 Caso em que X é uma v.a.c. . . . 119
4.7.1.3 Propriedades da esperança matemática . . . 119
4.7.2 Mediana . . . 120
4.7.3 Moda . . . 121
4.8 Medidas de dispersão de uma variável aleatória . . . 121
4.8.1 Variância . . . 121
4.8.1.1 Propriedades da variância: . . . 122
4.8.1.2 Desvio padrão . . . 123
4.8.2 Covariância . . . 123
4.8.2.1 Propriedades da covariância . . . 123
4.9 Coeficiente de correlação . . . 124
4.10 Exercícios propostos com respostas . . . 124

5 Distribuições de variáveis aleatórias 142
5.1 Introdução . . . 142
5.2 Distribuições para variáveis aleatórias discretas . . . 142
5.2.1 Distribuição Bernoulli . . . 142
5.2.2 Distribuição binomial . . . 143
5.2.2.1 Exercícios propostos com respostas . . . 144
5.2.3 Distribuição de Poisson . . . 144
5.2.3.1 Exercícios propostos com respostas . . . 146
5.3 A distribuição normal para variáveis aleatórias contínuas . . . 146

5.3.1 Variável normal padronizada (Z) 148
5.3.1.1 Média e variância da variável normal padronizada 148
5.3.2 Tabela da distribuição normal padrão 149
5.3.3 Exercícios propostos com respostas 149
5.3.4 Teorema da combinação linear 150
5.3.5 Exercícios propostos com respostas 150
5.4 Exercícios propostos com respostas 151

6 Regressão linear simples **156**
6.1 Introdução 156
6.2 O modelo estatístico 156
6.3 Estimadores dos parâmetros 157
6.3.1 O método dos mínimos quadrados (MMQ) 158
6.4 Análise de variância 160
6.5 O coeficiente de determinação simples (r^2) 161
6.6 Exercícios propostos com respostas 163

7 Testes de hipóteses **170**
7.1 Introdução 170
7.2 Alguns conceitos 170
7.2.1 Parâmetro, estimador e estimativa 170
7.2.2 Hipóteses estatísticas 171
7.2.2.1 Hipótese de nulidade (H_0) 171
7.2.2.2 Hipótese alternativa (H_a ou H_1) 172
7.2.3 Nível de significância α 172
7.2.4 Região crítica 172
7.2.5 Erros de decisão 173
7.2.5.1 Erro tipo I 173
7.2.5.2 Erro tipo II 173
7.2.6 Valor$-p$ ou nível crítico ou probabilidade de significância 173
7.2.7 Poder de um teste 173
7.3 Etapas para a realização de um teste de hipóteses 174
7.4 Teste Z ("grandes amostras") 174
7.4.1 Teste Z para uma média 174
7.4.2 Outra aplicação do teste Z 175
7.4.3 Teste Z para duas médias 176
7.5 O teste de qui-quadrado (χ^2) 177
7.5.1 Teste de aderência 178
7.5.2 Teste de independência 179
7.5.3 Teste de homogeneidade 180
7.6 Teste F 181
7.7 Teste t de Student ("pequenas amostras") 182
7.7.1 Teste t para uma média 182
7.7.2 Teste t para duas médias 183
7.7.3 Teste t para dados pareados 186

7.8 Exercícios propostos com respostas-Lista 1 187
7.9 Exercícios propostos com respostas-Lista 2 193

8 Noções de amostragem **199**
8.1 Introdução 199
8.2 Amostra simples ao acaso 199
8.2.1 Variâncias 200
8.2.1.1 Variância da média 201
8.2.1.2 Intervalos de confiança 201
8.3 Dimensionamento da amostra 202
8.3.1 Amostra simples ao acaso 202
8.4 Amostra simples ao acaso para proporções ou porcentagens 206
8.5 Dimensionamento da amostra 208
8.5.1 Amostra simples ao acaso 208
8.6 Exercício resolvido 210
8.7 Exercício proposto 211

Referências **211**

A Tabelas estatísticas **214**

Capítulo 1

Somatório e produtório

Em tudo o que será exposto neste material, muitas somas e produtos estarão envolvidos. Neste capítulo apresentamos duas notações simplificadoras: o somatório na seção 1.1 e o produtório na seção 1.2. Estes temas serão utilizados nos capítulos posteriores e, sua compreensão é de suma importância.

1.1 Somatório

Muitos dos processos estatísticos exigem o cálculo da soma. Para simplificar a representação da operação de adição nas expressões algébricas, utiliza-se a notação Σ, letra grega sigma maiúsculo.

As principais representações utilizadas na Estatística são:

i) $\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$, soma simples;

ii) $\sum_{i=1}^{n} X_i^2 = X_1^2 + X_2^2 + \cdots + X_n^2$, soma dos quadrados (SQ);

iii) $\sum_{i=1}^{n} X_i Y_i = X_1 Y_1 + X_2 Y_2 + \cdots + X_n Y_n$, soma de produtos (SP);

iv) $\left(\sum_{i=1}^{n} X_i\right)^2 = (X_1 + X_2 + \cdots + X_n)^2$, quadrado da soma;

v) $\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{m} Y_j\right) = (X_1 + X_2 + \cdots + X_n) \cdot (Y_1 + Y_2 + \cdots + Y_m)$, produto das somas.

Lê-se $\sum_{i=1}^{n} X_i$ como: somatório de X índice i, com i variando de 1 até n, em que:

- n, é a ordem da última parcela ou limite superior (LS) do somatório;
- $i = 1$, é a ordem da primeira parcela da soma ou limite inferior do somatório (LI);
- i, é o índice que está indexando os valores da variável X. Outras letras como j, l, k podem ser utilizadas;

- X, é a variável cujos valores são somados. Outras letras comumente utilizadas são Y, W, Z, U e V.

Exemplo

Considere os valores X_i e Y_i informados na tabela a seguir.

i	1	2	3	4	5	6
X_i	90	95	97	98	100	60
Y_i	60	70	80	60	90	75

Com o auxílio de uma calculadora obtém-se:

a) $\sum_{i=1}^{6} X_i = 540$;

b) $\sum_{i=1}^{6} X_i^2 = 49738$;

c) $\left(\sum_{i=1}^{6} X_i\right)^2 = 291600$;

d) $\sum_{i=1}^{6} X_i Y_i = 39190$;

e) $\left(\sum_{i=1}^{6} X_i\right)\left(\sum_{i=1}^{6} Y_i\right) = 234900$.

1.1.1 Número de termos do somatório

O número de termos ou parcelas indexadas por um somatório (NT) pode ser obtido por:

$$NT = (LS - LI) + 1 - r.$$

em que r é o número de restrições ou de termos eliminados.

Exemplos

Obter o número de termos para os seguintes somatórios:

a) $\sum_{i=3}^{8} X_i$, $\qquad R : NT = (8-3)+1 = 6$

b) $\sum_{\substack{k=1 \\ k\neq 9,11}}^{15} Y_k$, $\qquad R : NT = (15-1)+1-2 = 13$

c) $\sum_{k=1}^{9}\sum_{l=5}^{8} Z_{kl}$, $\qquad R : NT = [(9-1)+1]\cdot[(8-5)+1] = 36$

Observação: Verifique $\sum_{k=1}^{9}\sum_{l=5}^{8} Z_{kl} = \sum_{l=5}^{8}\sum_{k=1}^{9} Z_{kl}$, ou seja, a ordem dos somatórios não altera o resultado.

$$\begin{aligned}\sum_{k=1}^{9}\sum_{l=5}^{8} Z_{kl} &= \sum_{k=1}^{9}(Z_{k5}+Z_{k6}+Z_{k7}+Z_{k8}) \\ &= (Z_{15}+Z_{16}+Z_{17}+Z_{18})+(Z_{25}+Z_{26}+Z_{27}+Z_{28})+\ldots \\ &+ (Z_{95}+Z_{96}+Z_{97}+Z_{98})\end{aligned}$$

1.1.2 Propriedades de somatório

As propriedades são simplesmente as regras da operação de adição. Elas facilitam o desenvolvimento das expressões algébricas com a notação do somatório. O objetivo é desenvolver as expressões até chegar às somas que não podem mais ser simplificadas ($\sum X, \sum X^2, \sum XY$), etc.

P.1) Somatório de uma constante k.

É igual ao produto do número de termos indexados pelo somatório pela constante. Uma constante é um termo não indexado ou cujo índice de indexação seja diferente do somatório que o precede.

$$\sum_{i=1}^{n} k = \underbrace{k + k + \ldots + k}_{n \text{ termos}} = nk.$$

Exemplos

a) $\sum\limits_{i=1}^{10} 5 = [(10-1)+1](5) = 10(5) = 50,$ c) $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} X_i = \sum\limits_{i=1}^{n} mX_i = m\sum\limits_{i=1}^{n} X_i.$

b) $\sum\limits_{i=3}^{12} Y_j = [(12-3)+1]Y_j = 10Y_j,$

P.2) Somatório do produto de uma constante por uma variável.

Basta retirar a constante do somatório (colocar em evidência).

$$\sum_{i=1}^{n} kX_i = kX_1 + kX_2 + \ldots + kX_n = k\left(X_1 + X_2 + \ldots + X_n\right) = k\sum_{i=1}^{n} X_i.$$

Exemplos

a) $\sum\limits_{i=1}^{n} \dfrac{X_i}{2} = \dfrac{1}{2}\sum\limits_{i=1}^{n} X_i, k = \dfrac{1}{2}$

b) $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} X_iY_j = \sum\limits_{i=1}^{n} X_i \sum\limits_{j=1}^{m} Y_j$, pois, X_i é uma constante para o somatório indexado com j.

c) Verifique para $X = \{2, 4, 6\}$ e $Y = \{3, 5\}$, que $\sum\limits_{i=1}^{3}\sum\limits_{j=1}^{2} X_iY_j = \sum\limits_{i=1}^{3} X_i \sum\limits_{j=1}^{2} Y_j = 96.$

P.3) Somatório de uma soma (ou subtração) de variáveis.

É igual à soma ou subtração dos somatórios dessas variáveis.

Exemplos

a) $\sum\limits_{i=1}^{n} (X_i + Y_i - W_i) = \sum\limits_{i=1}^{n} X_i + \sum\limits_{i=1}^{n} Y_i - \sum\limits_{i=1}^{n} W_i.$

b) $\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\sum\limits_{k=1}^{r} (X_i + Y_j - W_k) = mr\sum\limits_{i=1}^{n} X_i + nr\sum\limits_{j=1}^{m} Y_j - nm\sum\limits_{k=1}^{r} W_k.$

1.1.3 Exemplos resolvidos

1) Mostre que $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$, partindo-se de $\sum_{i=1}^{n} \left[(i+1)^2 - i^2\right]$.

Desenvolvendo-se este somatório temos:

$$\begin{aligned}
\sum_{i=1}^{n} \left[(i+1)^2 - i^2\right] &= \left(2^2 - 1^2\right) + \left(3^2 - 2^2\right) + \left(4^2 - 3^2\right) + \cdots + \left[(n-2)^2 - (n-1)^2\right] + \\
&+ \left[n^2 - (n-1)^2\right] + \left[(n+1)^2 - n^2\right] \\
&= \left(-1^2 + 2^2\right) + \left(-2^2 + 3^2\right) + \left(-3^2 + 4^2\right) + \cdots + \left[-(n-1)^2 + n^2\right] + \left[-n^2 + (n+1)^2\right] \\
&= -1^2 + \left(2^2 - 2^2\right) + \left(3^2 - 3^2\right) + \left(4^2 - 4^2\right) + \cdots + \left(n^2 - n^2\right) + (n+1)^2 \\
&= -1^2 + 0 + 0 + 0 + \cdots + 0 + (n+1)^2 = (n+1)^2 - 1 = n^2 + 2n + 1 - 1 \\
&= n^2 + 2n \qquad (1.1)
\end{aligned}$$

Podemos também resolver primeiro o quadrado e depois aplicar o somatório, assim:

$$\begin{aligned}
\sum_{i=1}^{n} \left[(i+1)^2 - i^2\right] &= \sum_{i=1}^{n} \left(i^2 + 2i + 1 - i^2\right) = \sum_{i=1}^{n} (2i+1) = \sum_{i=1}^{n} 2i + \sum_{i=1}^{n} 1 \\
&= 2\sum_{i=1}^{n} i + n \qquad (1.2)
\end{aligned}$$

Igualando-se (1.1) e (1.2) temos $2\sum_{i=1}^{n} i + n = n^2 + 2n$, logo $2\sum_{i=1}^{n} i = n^2 + n$ e assim

$$\sum_{i=1}^{n} i = \frac{n^2 + n}{2} = \frac{n(n+1)}{2},$$

como queríamos demonstrar.

2) Mostre que $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$, partindo-se de $\sum_{i=1}^{n} \left[(i+1)^3 - i^3\right]$.

A solução deste exemplo é similar àquela desenvolvida no exemplo **1)**. De fato, desenvolvendo-se este somatório temos:

$$\begin{aligned}
\sum_{i=1}^{n} \left[(i+1)^3 - i^3\right] &= \left(2^3 - 1^3\right) + \left(3^3 - 2^3\right) + \left(4^3 - 3^3\right) + \cdots + \left[(n-2)^3 - (n-1)^3\right] + \\
&+ \left[n^3 - (n-1)^3\right] + \left[(n+1)^3 - n^3\right] \\
&= \left(-1^3 + 2^3\right) + \left(-2^3 + 3^3\right) + \left(-3^3 + 4^3\right) + \cdots + \left[-(n-1)^3 + n^3\right] + \left[-n^3 + (n+1)^3\right] \\
&= -1^3 + \left(2^3 - 2^3\right) + \left(3^3 - 3^3\right) + \left(4^3 - 4^3\right) + \cdots + \left(n^3 - n^3\right) + (n+1)^3 \\
&= -1^3 + 0 + 0 + 0 + \cdots + 0 + (n+1)^3 = (n+1)^3 - 1 = n^3 + 3n^2 + 3n + 1 - 1 \\
&= n^3 + 3n^2 + 3n \qquad (1.3)
\end{aligned}$$

Podemos também resolver primeiro o termo ao cubo e depois aplicar o somatório, assim:

$$\begin{aligned}\sum_{i=1}^{n}\left[(i+1)^3-i^3\right] &= \sum_{i=1}^{n}\left(i^3+3i^2+3i+1-i^3\right)=\sum_{i=1}^{n}\left(3i^2+3i+1\right)\\ &= \sum_{i=1}^{n}3i^2+\sum_{i=1}^{n}3i+\sum_{i=1}^{n}1\\ &= 3\sum_{i=1}^{n}i^2+3\sum_{i=1}^{n}i+n \qquad (1.4)\end{aligned}$$

Igualando-se (1.3) e (1.4) temos

$$3\sum_{i=1}^{n}i^2+3\sum_{i=1}^{n}i+n=n^3+3n^2+3n,$$

logo

$$3\sum_{i=1}^{n}i^2=n^3+3n^2+2n-3\sum_{i=1}^{n}i,$$

e utilizando-se o resultado do exemplo **1)** temos:

$$\begin{aligned}3\sum_{i=1}^{n}i^2 &= n^3+3n^2+2n-3\frac{n(n+1)}{2}\\ &= \frac{2n^3+6n^2+4n-3n^2-3n}{2}\\ &= \frac{2n^3+3n^2+n}{2}\\ &= \frac{n\left(2n^2+3n+1\right)}{2}\\ &= \frac{n(n+1)(2n+1)}{2}\end{aligned}$$

logo

$$\sum_{i=1}^{n}i^2=\frac{n(n+1)(2n+1)}{6},$$

como queríamos demonstrar.

3) Dado que,

$$\sum_{k=1}^{n}k=\frac{n(n+1)}{2},\quad \sum_{k=1}^{n}k^2=\frac{n(n+1)(2n+1)}{6}\quad \text{e}\quad \sum_{k=1}^{n}k^3=\left[\frac{n(n+1)}{2}\right]^2,$$

utilize as propriedades de somatório para calcular:

$$\sum_{x=1}^{12}\sum_{y=1}^{10}(x-y)^3.$$

Temos que:

$$\begin{aligned}\sum_{x=1}^{12}\sum_{y=1}^{10}(x-y)^3 &= \sum_{x=1}^{12}\sum_{y=1}^{10}\left(x^3-3x^2y+3xy^2+y^3\right)\\ &= 10\sum x^3-3\sum x^2\sum y+3\sum x\sum y^2+12\sum y^3\\ &= 10\left(\frac{12\times13}{2}\right)^2-3\left(\frac{12\times13\times25}{6}\right)\left(\frac{10\times11}{2}\right)\\ &+ 3\left(\frac{12\times13}{2}\right)\left(\frac{10\times11\times21}{6}\right)+12\left(\frac{10\times11}{2}\right)^2\\ &= 60840-107250+90090-36300=7380.\end{aligned}$$

1.1.4 Outras aplicações da notação de somatório

É comum que seja necessário indexar uma variável com mais do que um único índice. Por exemplo, em tabelas de dupla entrada e em modelos Estatísticos, conforme os seguintes exemplos.

- Variável X avaliada num esquema fatorial $i \times j$ de modo que X_{ij} é o valor de X na combinação (i, j) em que $i = 1, 2, \ldots, r$ é o índice da linha e $j = 1, 2, \ldots, s$ é o índice da coluna, conforme apresentado na tabela a seguir:

linha (i)	coluna (j)						Total
	1	2	...	j	...	s	
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1s}	$X_{1.}$
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2s}	$X_{2.}$
...	...	...	...	...	...	...	...
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{is}	$X_{i.}$
...	...	...	...	...	...	...	...
r	X_{r1}	X_{r2}	...	X_{rj}	...	X_{rs}	$X_{r.}$
Total	$X_{.1}$	$X_{.2}$	...	$X_{.j}$	...	$X_{.s}$	G

G = total geral$=\sum_{i=1}^{r}\sum_{j=1}^{s}X_{ij}=\sum_{j=1}^{s}\sum_{i=1}^{r}X_{ij}=X_{..}$

Total da i-ésima linha: $\sum_{j=1}^{s}X_{ij}=X_{i.}$

Total da j-ésima coluna: $\sum_{i=1}^{r}X_{ij}=X_{.j}$

- Exemplo de um modelo estatístico com alguns somatórios simples, duplos e triplos que podem ser de interesse,

$$Y_{ijk}=\mu+\alpha_i+\beta_j+\varepsilon_{ijk},$$

$$\sum_{i=1}^{a}Y_{ijk},\qquad\sum_{i=1}^{a}\sum_{j=1}^{b}Y_{ijk},\qquad\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}Y_{ijk}.$$

1.1.5 Exercícios propostos com respostas

1) Considerando os seguintes valores:

$$X_1 = 2 \quad X_2 = 6 \quad X_3 = 7 \quad X_4 = 9$$

$$Y_1 = 1 \quad Y_2 = 4 \quad Y_3 = 5 \quad Y_4 = 11$$

Calcular:

a) $\sum_{i=1}^{3}\sum_{j=2}^{4}(X_i + 2)$

b) $\sum_{i=2}^{4}\sum_{j=2}^{3} 3\,(X_i - Y_j)$

c) $\sum_{i=1}^{3}(Y_i - 2)^2$

d) $\sum_{i=1}^{4}(X_i - 4Y_i)$

Respostas: a) 63 b) 51 c) 14 d) -60

2) Calcular:

a) $\sum_{i=-1}^{3}\left(i^2 + \frac{1}{j}\right)$

b) $\sum_{i=3}^{6}\sum_{j=0}^{2}(i+j)\,.\left(\frac{i-3}{i}\right)$

Respostas: a) $5(3 + \frac{1}{j})$ b)$\frac{429}{20}$

3) Calcule X_1 e X_3, dado que:

$$\sum_{i=1}^{6} X_i = 42 \qquad \sum_{i=1}^{6} X_i^2 = 364 \qquad \sum_{\substack{i=1 \\ i\neq 1,3}}^{6} X_i = 34 \qquad \sum_{\substack{i=1 \\ i\neq 1,3}}^{6} X_i^2 = 324$$

Resposta: $X_1 = 2$ e $X_3 = 6$ ou $X_1 = 6$ e $X_3 = 2$

4) Calcular:

a) $\sum_{i=1}^{5}\sum_{j=2}^{4}(i+j)$

b) $\sum_{j=5}^{9}\sum_{i=1}^{6} i \cdot j$

Respostas: a) 90 b)735

5) Sabendo-se que a soma dos n termos de uma progressão aritmética é dada por:

$$S_n = \frac{n(a_1 + a_n)}{2},$$

mostre que $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ partindo-se de $\sum_{i=1}^{n}(i+1)^2$.

1.2 Produtório

O símbolo $\prod$, letra grega pi maiúsculo, é utilizado para facilitar a representação dos produtos ou da operação de multiplicação. A representação básica de um produtório, lida como o produtório dos valores X índice i, com i variando de 1 a n, é a seguinte:

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

Alguns exemplos:

1) $b_1 \cdot b_2 \cdot \ldots \cdot b_n = \prod_{i=1}^{n} b_i$;

2) $\underbrace{b \cdot b \cdot b \cdot \ldots \cdot b}_{n \text{ termos}} = \prod_{i=1}^{n} b = b^n$

3) $\prod_{i=1}^{n} cX_i = cX_1 \cdot cX_2 \cdot \ldots \cdot cX_n = c^n \cdot X_1 \cdot X_2 \cdot \ldots \cdot X_n = c^n \prod_{i=1}^{n} X_i$

4) $\prod_{i=1}^{n} X_i Y_i = X_1 Y_1 \cdot X_2 Y_2 \cdot \ldots \cdot X_n Y_n = (X_1 \cdot X_2 \cdot \ldots \cdot X_n)(Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n)$
$= \left(\prod_{i=1}^{n} X_i\right)\left(\prod_{i=1}^{n} Y_i\right)$

5) $\prod_{i=1}^{n} i = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n = n!$

6) $\log\left(\prod_{i=1}^{n} X_i\right) = \log(X_1 \cdot X_2 \cdot \ldots \cdot X_n) = \log X_1 + \log X_2 + \ldots + \log X_n = \sum_{i=1}^{n} \log X_i$, sendo $X_i > 0,\ \forall\ i, 1 \leq i \leq n$.

1.2.1 Número de termos do produtório

O número de termos de um produtório (NT) é análogo ao de um somatório.

Exemplo

Determine o número de termos do produtório

$$\prod_{\substack{i=2 \\ i \neq 5,20,27}}^{30} k = k^{NT} = k^{26} \quad \text{pois, } NT = (30-2) + 1 - 3 = 26.$$

1.2.2 Alguns exemplos

1) Sabendo-se que:

$$X_1 = 2 \quad X_2 = 3 \quad X_3 = 5$$
$$Y_1 = 3 \quad Y_2 = 5 \quad Y_3 = 7$$

Calcule:

a) $\prod_{i=1}^{3} X_i = X_1 \cdot X_2 \cdot X_3 = 2 \cdot 3 \cdot 5 = 30$;

b) $\prod_{i=1}^{3} Y_i = Y_1 \cdot Y_2 \cdot Y_3 = 3 \cdot 5 \cdot 7 = 105$;

c) $\prod_{i=1}^{3} 3X_i = 3^3 \prod_{i=1}^{3} X_i = 27(30) = 810$;

d) $\prod_{i=1}^{3} X_i Y_i = \left(\prod_{i=1}^{3} X_i\right) . \left(\prod_{i=1}^{3} Y_i\right) = (30)(105) = 3150.$

A ordem de execução da expressão altera o resultado quando se tem somatório e produtório juntos. Por exemplo,

a) $\sum_{i=1}^{2} \prod_{k=1}^{3} [(k+1)\, i] = \sum_{i=1}^{2} (2i \cdot 3i \cdot 4i) = \sum_{i=1}^{2} (24i^3) = 24 \sum_{i=1}^{2} i^3 = 24\,(1^3 + 2^3) = 24 \cdot 9 = 216$

b)

$$\begin{aligned} \prod_{k=1}^{3} \sum_{i=1}^{2} [(k+1)\, i] &= \prod_{k=1}^{3} [(k+1) \cdot 1 + (k+1) \cdot 2] = \prod_{k=1}^{3} [3\,(k+1)] 3^3 \prod_{k=1}^{3} (k+1) \\ &= 27\,(2 \cdot 3 \cdot 4) = 27 \cdot 24 = 648 \end{aligned}$$

1.3 Exercícios propostos com respostas

1) Seja uma variável X, assumindo os seguintes valores:

$$X = \{5, 2, 3, 0, 1, 2, 6, 9, 4, 8\}, n = 10$$

Calcule:

a) $\sum_{i=1}^{10} X_i$

b) $\sum_{i=1}^{10} X_i^2$

c) $\left(\sum_{i=1}^{10} X_i\right)^2$

d) $\dfrac{\sum_{i=1}^{10} X_i^2 - \dfrac{\left(\sum_{i=1}^{10} X_i\right)^2}{10}}{10-1}$

e) $\sum_{i=1}^{10} (X_i - 4)$

f) $\sum_{i=1}^{10} (X_i - 4)^2$

g) $\dfrac{\sum_{i=1}^{10} (X_i - 4)^2}{10-1}$

h) $\dfrac{\sum_{i=1}^{10} X_i}{10}$

2) Sabendo-se que $\sum_{i=1}^{5} X_i = -6$ e $\sum_{i=1}^{5} X_i^2 = 12$,calcule:

a) $\sum_{i=1}^{5} (4X_i + 5)$ b) $\sum_{i=1}^{5} X_i\,(X_i - 2)$ c) $\sum_{i=1}^{5} (X_i - 3)^2$

3) Desenvolver e calcular:

a) $\sum_{i=1}^{3}\sum_{j=2}^{6}(i+bj)$

b) $\sum_{j=1}^{2}\sum_{i=1}^{5}(i-j)$

c) $\sum_{i=1}^{2}\sum_{j=0}^{2}(i+3j)^2$

d) $\sum_{i=1}^{7}\sum_{j=0}^{8}cb$

e) $\sum_{i=1}^{4}\sum_{j=1}^{5}i^2$

4) Utilizando os dados da Tabela abaixo, calcule:

$i\backslash j$	1	2	3	4
1	8	7	5	9
2	4	0	10	2

a) $\sum_{i=1}^{2}X_{i1}$

b) $\sum_{j=1}^{4}X_{1j}$

c) $\sum_{i=1}^{2}\sum_{j=1}^{4}X_{ij}$

d) $\sum_{\substack{j=1\\ j\neq 3}}^{4}X_{2j}$

e) $\sum_{j=2}^{3}X_{2j}$

f) $\sum_{\substack{j=1\\ j\neq 2}}^{4}\frac{1}{X_{2j}}$

g) $\prod_{\substack{j=1\\ j\neq 3}}^{4}6X_{1j}$

h) $\prod_{\substack{j=1\\ j\neq 2}}^{4}X_{2j}$

5) Escrever usando a notação de somatório ou produtório, conforme o caso:

a) $\left(\frac{X_1-Y_1}{2}+\frac{X_2-Y_2}{2}+\frac{X_4-Y_4}{2}\right)^2$

b) $a!$

c) $(X_1+Y_1)(X_1+Y_2)(X_1+Y_3)$

d) $(X_1Y_1)+(X_1Y_2)+(X_1Y_3)+(X_2Y_1)+(X_2Y_2)+(X_2Y_3)$

e) $(X_1Y_1)(X_2Y_2)\cdot\cdots\cdot(X_nY_n)$

6) Considere os seguintes valores:

$$\begin{array}{cccccccc} X_1=2 & X_2=4 & X_3=6 & X_4=8 & X_5=10 & X_6=12 & X_7=14 & X_8=16 \\ Y_1=1 & Y_2=3 & Y_3=5 & Y_4=7 & Y_5=9 & Y_6=11 & Y_7=13 & Y_8=15 \end{array}$$

Calcule os seguintes somatórios e produtórios:

a) $\sum_{i=1}^{8}\sum_{j=2}^{5}(X_i-3)$

b) $\sum_{i=1}^{8}\left(\frac{X_i}{2}-Y_i\right)^2$

c) $\sqrt{\prod_{i=1}^{4}X_i}$

d) $\prod_{i=2}^{4}\frac{X_iY_i}{3}$

7) Desenvolver:

a) $\sum_{i=-1}^{3}\left(i^2+\frac{1}{j}\right)$

b) $\sum_{\substack{j=1\\ j\neq 4}}^{5}\sum_{i=2}^{4}\frac{(i^2-j^2)\,j}{i+j}$

c) $\left[\sum_{i=1}^{5}(i+8)\right]^2$

d) $\prod_{i=1}^{5}(i+8)$

8) Se $\sum_{i=1}^{3}X_i=12$; $\sum_{i=1}^{3}X_i^2=56$; e $Y_1=3$; $Y_2=5$; $Y_3=6$, calcule:

a) $\sum_{i=1}^{3}9$

b) $\sum_{i=1}^{3}12X_i$

c) $\sum_{i=1}^{3}(X_i^2-2)$

d) $\sum_{i=1}^{3}X_iY_i$

9) Se $X_1=2$; $X_2=4$; $X_3=6$; e $Y_1=3$; $Y_2=5$; $Y_3=6$, calcule:

a) $\sum_{i=1}^{3}(X_iY_i)$

b) $\sum_{i=1}^{3}(X_i-2)(Y_i-5)$

10) Calcule X_9 e X_{21}, sabendo-se que:

$\sum_{i=1}^{50}X_i=200$, $\sum_{i=1}^{50}X_i^2=1206$, $\sum_{\substack{i=1\\ i\neq 9,21}}^{50}X_i=190$, $\sum_{\substack{i=1\\ i\neq 9,21}}^{50}X_i^2=1154$.

11) Dados:

i	f_i	X_i
1	3	10
2	5	11
3	9	15
4	10	19
5	2	21
6	1	26

Calcule as seguintes quantidades:

a) $\sum_{i=1}^{6}X_i$

b) $\sum_{i=1}^{6}f_i$

c) $\sum_{i=1}^{6}f_iX_i^2$

d) $\frac{\sum_{i=1}^{6}f_iX_i}{\sum_{i=1}^{6}f_i}$

12) Sabendo-se que:

$$X_1 = 3; \quad X_2 = 4; \quad X_3 = 8; \quad X_4 = 7; \quad X_5 = 6$$
$$Y_1 = 3; \quad Y_2 = 8; \quad Y_3 = 2; \quad Y_4 = 5; \quad Y_5 = 6$$

Calcule:

a) $\sum_{\substack{i=1 \\ i\neq 2}}^{5} X_i$

b) $\sum_{i=1}^{5} 4X_i$

c) $\sum_{i=3}^{5} (X_i + 6)$

d) $\sum_{i=2}^{4} (2X_i - 3)$

e) $\sum_{i=1}^{5} X_i Y_i$

f) $\sum_{i=1}^{5} (X_i + Y_i)$

13) Sabendo-se que $\sum_{x=1}^{n} x = \frac{n(1+n)}{2}$, calcule $\sum_{x=1}^{200} \frac{(x-100)}{2}$.

14) A variância (S^2) de uma amostra com n observações de uma variável aleatória X pode ser definida por,

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2. \tag{1.5}$$

Sendo $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, pede-se:

a) Utilize propriedades de somatório na equação (1.5) para obter a fórmula dada por,

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

b) Seja $Y_i = kX_i$, em que k é uma constante qualquer. Utilize propriedades de somatório na equação (1.5) para mostrar que $S_Y^2 = k^2 S_X^2$.

15) Considere os seguintes valores X_i e Y_i

i	1	2	3	4	5	6	7	8	9	10
X_i	2	5	7	9	8	6	4	5	2	10
Y_i	1	5	7	2	4	4	6	6	8	8

Calcule:

a) $\sum_{i=1}^{10} \dfrac{\left(X_i - \bar{X}\right)^2}{9}$

b) $\sum_{i=1}^{10} \left[\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)\right]$

c) $\prod_{\substack{i=1 \\ i\neq 2,3}}^{6} \left(\dfrac{X_i - Y_i}{2}\right)$

16) Verifique por indução matemática que $\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$.

17) Dados

$$\sum_{i=1}^{5} X_i = 2,6 \quad \sum_{i=1}^{5} X_i^2 = 1,84 \quad \sum_{j=3}^{8} Y_j = 11 \quad \sum_{j=3}^{8} Y_j^2 = 31.$$

Calcule $\sum_{i=1}^{5}\sum_{j=3}^{8} (2X_i - Y_j)^2$.

18) Utilize as **propriedades** para calcular os somatórios e produtórios a seguir:

a) Dado que $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ e $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$ calcule, $\sum_{\substack{x=1 \\ x\neq 2,4}}^{20} x(x+1)$.

b) $\sum_{i=1}^{3}\prod_{j=1}^{2} i^{2j-1}$;

c) Se $\sum_{i=1}^{3} X_i = 6$, $\sum_{i=1}^{3} X_i^2 = 20$, $\sum_{j=1}^{2} Y_j = 4$ e $\sum_{j=1}^{2} Y_j^2 = 10$ calcule,

$$\sum_{i=1}^{3}\sum_{j=1}^{2} \left[(X_i - 2)(Y_j - 1)^2\right];$$

d) $\prod_{k=1}^{5} \frac{(k+1)}{2}$.

19) Considere os elementos a_{ij} da matriz A, com $i = 1,2,3,4$ e $j = 1,2,3,4,5$ para indicar o elemento da i-ésima linha e j-ésima coluna,

$$A = \begin{bmatrix} -1 & 17 & 9 & -2 & 3 \\ 3 & 13 & 10 & 2 & 6 \\ 11 & -9 & 0 & -3 & 2 \\ -6 & -8 & 1 & 4 & 5 \end{bmatrix}$$

calcule:

a) $\sum_{i=1}^{3} a_{i2}$; b) $\sum_{\substack{i=1 \\ i\neq 2}}^{4}\sum_{\substack{j=2 \\ j\neq 4}}^{5} a_{ij}$; c) $\prod_{i=1}^{4} 2^{a_{i4}}$.

20) Utilize propriedades de somatório e produtório.

a) Calcule $\sum_{i=1}^{20}\sum_{j=1}^{10} 3\,(X_i - Y_j)$ dados $\sum_{i=1}^{20} X_i = 20$ e $\sum_{j=1}^{10} Y_j = 5$.

b) Calcule $\sum_{i=1}^{3}\sum_{j=1}^{4}\left(Y_{ij}-8\right)^2$, considerando-se a seguinte notação:

$$Y_{i\cdot}=\sum_{j=1}^{4}Y_{ij} \quad \text{e} \quad Y_{i\cdot}^2=\sum_{j=1}^{4}Y_{ij}^2 \quad \text{com,}$$

$$Y_{1\cdot}=30, \quad Y_{2\cdot}=32, \quad Y_{3\cdot}=38, \text{ e } Y_{1\cdot}^2=225, \quad Y_{2\cdot}^2=256, \quad Y_{3\cdot}^2=360$$

c) Calcule $\prod_{\substack{x=1\\x\neq 3}}^{5}\frac{\left(x-3\right)^2}{2}$.

21) Considere os seguintes valores,

$$m=50 \quad n=30 \quad k=3 \quad \sum_{j=1}^{m}Y_j=80 \quad \sum_{i=1}^{n}X_i=100 \quad \sum_{i=1}^{n}X_i^2=600.$$

Aplique propriedades de somatório e utilize os valores informados para calcular:

$$\sum_{j=1}^{m}\sum_{i=1}^{n}Y_j\left(X_i-k\right)^2.$$

22) Utilize as propriedades de somatório e produtório e os valores a seguir.

$$\sum_{i=1}^{3}X_{1i}=6 \qquad \sum_{i=1}^{3}X_{2i}=8 \qquad \sum_{j=1}^{5}Y_{1j}=10 \qquad \sum_{j=1}^{5}Y_{2j}=12.$$

Calcule,

a) $\sum_{k=1}^{2}\left[\sum_{i=1}^{3}\sum_{j=1}^{5}\left(X_{ki}-3\right)\left(Y_{kj}-2\right)\right]$;

b) $\sum_{i=1}^{2}\prod_{k=1}^{3}\left(2^k-1\right)i$;

c) $\sum_{i=1}^{5}\sum_{j=4}^{6}\left(2i-3j\right)^2$.

23) Considere as seguintes somas:

$$\sum_{j=1}^{10}Y_j=8 \text{ e } \sum_{\substack{i=3\\i\neq 5,9,11}}^{20}X_i=20.$$

Calcule:

$$\sum_{\substack{i=3\\i\neq 5,9,11}}^{20}\sum_{j=1}^{10}\left(X_i+Y_j-2\right).$$

24) Dados os seguintes valores e as respectivas somas,

$$X_1 = 2 \quad X_2 = 4 \quad X_3 = 6 \quad X_4 = 8 \quad X_5 = 10 \longrightarrow \sum X = 30 \quad \sum X^2 = 220$$

$$\begin{array}{lllll} Y_1 = 1 & Y_2 = 3 & Y_3 = 5 & Y_4 = 7 & Y_5 = 9 \\ Y_6 = 11 & Y_7 = 13 & Y_8 = 15 & Y_9 = 17 & Y_{10} = 19 \end{array} \longrightarrow \sum Y = 100 \quad \sum Y^2 = 1330$$

$$Z_1 = 12 \quad Z_2 = 20 \quad Z_3 = 30 \quad Z_4 = 40 \longrightarrow \sum Z = 102 \quad \sum Z^2 = 3044$$

Calcule: $\sum_{i=1}^{5}\sum_{j=1}^{10}\sum_{k=1}^{4}\left[(X_i - Y_j)^2 - Z_k\right]$.

25) Calcule: $\prod_{k=1}^{3}(3k-1)\,k^3$.

26) Utilize as propriedades de somatório e produtório .

a) $\sum_{i=1}^{20}\sum_{j=1}^{50}\left[(X_i - 2)(Y_j - 3) + Z_{ij}\right]$; $\sum_{i=1}^{20} X_i = 80$, $\sum_{j=1}^{50} Y_j = 30$, $\sum_{i=1}^{20}\sum_{j=1}^{50} Z_{ij} = 5520$.

b) $\sum_{i=1}^{60}\sum_{k=5}^{12}(Z_k - 5)^2$; $\sum_{k=5}^{12} Z_k^2 = 412$ e $\sum_{k=5}^{12} Z_k = 60$.

c) $\prod_{k=1}^{5}\left(\frac{2k+2}{2}\right)$.

27) Utilize as propriedades de somatório e produtório, dado:

$n = 50$, $\sum_{i=1}^{n} X_i = 20$, $\sum_{i=1}^{n} X_i^2 = 285$ e $e \approx 2,7183$ (base do logaritmo neperiano)

a) $\prod_{i=1}^{n}\left[e^{\left(2X_i + 5 - X_i^2\right)}\right]$;

b) $\sum_{k=1}^{3}\sum_{i=1}^{n}\left[(X_i - 2)^2\right]$.

28) Utilize as propriedades de somatório e produtório.

a) $\prod_{k=1}^{5}\left(\frac{2^{k-1}}{2}\right)$.

b) $\sum_{j=1}^{5}\sum_{\substack{k=2 \\ k\neq 3}}^{8}[(k-3)(j+1)]$.

c) $\sum_{k=1}^{3}\sum_{j=1}^{2}\sum_{i=1}^{10}\left[(2X_i - 1)^2 - 15\right]$, dado $\sum_{i=1}^{10} X_i = 15$ e $\sum_{i=1}^{10} X_i^2 = 50$.

29) Seja $SQD(a) = \sum_{i=1}^{n}(X_i - a)^2$, $0 < a < +\infty$. Mostre por propriedades de somatório que,

$$\sum_{i=1}^{n}(X_i - a)^2 = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - a\right)^2$$

e conclua que,

$$min_a\left(SQD(a)\right) = SQD(\bar{X}), \quad \text{em que,} \quad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

30) Calcule:

a) $\prod_{x=2}^{6}\left(\frac{x^2-2x+1}{x^2-1}\right)$; b) $\sum_{k=9}^{11}\sum_{\substack{x=1\\x\neq 4,5}}^{6}\left[(x-1)(x+1)-k\right]$.

31) Dados os seguintes somatórios,

$$\sum_{i=1}^{50} X_i = 100 \quad \sum_{i=1}^{50} X_i^2 = 125 \quad \sum_{j=5}^{12} Y_j = 18 \quad \sum_{\substack{k=3\\k\neq 6,10,12}}^{25} Z_k = 22$$

calcule:

$$\sum_{i=1}^{50}\sum_{j=5}^{12}\sum_{\substack{k=3\\k\neq 6,10,12}}^{25}\left[(X_i-2)^2 - Y_j\ Z_k\right].$$

32) Calcule os itens abaixo sabendo-se que,

$$\sum_{i=1}^{20} X_i = 11, \qquad \sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad \text{e} \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}.$$

a) $\sum_{k=1}^{4}\prod_{i=1}^{20}\left(k^{2X_i-1}\right)$. b) $\prod_{k=1}^{2}\sum_{x=1}^{4}\left(k^{2x-2}\right)$. c) $\sum_{x=1}^{60}\left[(x-1)^2-1161\right]$.

33) Nos itens a seguir assinale (V) se estiver inteiramente correto ou (F) caso contrário.

a)( ) $\prod_{i=1}^{n} X_iY_i = \left(\prod_{i=1}^{n} X_i\right)\left(\prod_{i=1}^{n} Y_i\right)$

b)( ) $\sum_{i=1}^{n}\frac{X_i}{Y_i} = \frac{\left(\sum_{i=1}^{n} X_i\right)}{\left(\sum_{i=1}^{n} Y_i\right)}$

c)( ) $\sum_{j=3}^{5}\sum_{\substack{k=1\\k\neq 6,7}}^{10}\sum_{i=1}^{n} X_i^2 = 30\sum_{i=1}^{n} X_i^2$

d)( ) $\prod_{i=1}^{n}\frac{X_i}{Y_i} \neq \frac{\left(\prod_{i=1}^{n} X_i\right)}{\left(\prod_{i=1}^{n} Y_i\right)}$

e)( ) $\sum_{i=1}^{n}\sum_{j=1}^{m} X_iY_j = \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{m} Y_j\right)$

f)( ) $\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{s}\left(2X_i + 3X_iY_j - Z_k\right) = 2ms\sum_{i=1}^{n} X_i + 3\sum_{i=1}^{n} X_i\sum_{j=1}^{m} Y_j - nm\sum_{k=1}^{s} Z_k$

34) Dado que $\sum_{i=1}^{10} X_i = 20$ e $\sum_{i=1}^{10} X_i^2 = 50$, calcule: $\sum_{i=1}^{10}\sum_{j=3}^{8}\left(X_i+2\right)^2$.

35) Dado que $\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$, calcule: $\sum_{n=3}^{5} \prod_{i=1}^{n} \left(2^{i-1}\right)$.

36) Dado que,

$$\sum_{i=1}^{20} X_i = 100, \qquad \sum_{i=1}^{20} X_i^2 = 250, \qquad \sum_{j=1}^{10} Y_j = 40,$$
$$\sum_{j=1}^{10} Y_j^2 = 65, \qquad \sum_{\substack{k=1 \\ k\neq 4,7}}^{15} Z_k = 12 \qquad \sum_{\substack{k=1 \\ k\neq 4,7}}^{15} Z_k^2 = 32.$$

Utilize as propriedades de somatório e calcule:

$$\sum_{i=1}^{20} \sum_{j=1}^{10} \sum_{\substack{k=1 \\ k\neq 4,7}}^{15} \left(X_i - Y_j + Z_k\right)^2$$

37) Dado que,

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \quad \text{e} \quad \sum_{k=1}^{n} k^3 = \left[\frac{n(n+1)}{2}\right]^2,$$

Utilize as propriedades de somatório e calcule:

$$\sum_{k=1}^{30} \frac{(2k-1)^3}{5}$$

38) Considere os seguintes pares de valores (X_i, Y_i),

i	1	2	3	4	5	6	7	8	9	10
X_i	2	4	6	8	3	5	8	2	9	3
Y_i	1	3	7	5	9	2	3	4	6	4

Utilize sua calculadora e as propriedades de somatório para calcular:

a) $\sum_{i=1}^{10} \left[\left(X_i - \bar{X}\right)\left(Y_i - \overline{Y}\right)\right]$. b) $\sum_{i=1}^{10} \left(X_i - \bar{X}\right)^2$.

39) Dado que,

$$\sum_{i=5}^{30} X_i = 200, \quad \sum_{j=8}^{15} Y_j = 150 \quad \text{e} \quad \sum_{\substack{k=1 \\ k\neq 3,5,8}}^{10} Z_k = 30,$$

Utilize as propriedades de somatório e calcule:

$$\sum_{i=5}^{30} \sum_{j=8}^{15} \sum_{\substack{k=1 \\ k\neq 3,5,8}}^{10} \left(2X_i + Y_j - 5Z_k\right).$$

40) Dado que,

$$\sum_{i=5}^{23} Y_i = 40, \quad \sum_{i=5}^{23} Y_i^2 = 65, \quad \sum_{k=1}^{10} Z_k = 30 \quad \text{e} \quad \sum_{k=1}^{10} Z_k^2 = 60,$$

utilize as propriedades de somatório e calcule,

$$\sum_{i=5}^{23} \sum_{k=1}^{10} \left[(Y_i - 3) \left(Z_k + 1 \right)^2 \right].$$

41) Dado que,

$$\sum_{x=1}^{n} x = \frac{n(n+1)}{2}, \qquad \sum_{x=1}^{n} x^2 = \frac{n(n+1)(2n+1)}{6} \qquad \text{e} \qquad \sum_{x=1}^{n} x^3 = \left[\frac{n(n+1)}{2} \right]^2,$$

utilize as propriedades de somatório e calcule,

$$\sum_{k=1}^{20} (k-2)^3$$

42) Dado que,

$$\sum_{i=1}^{10} X_{1i} = 40, \quad \sum_{i=1}^{10} X_{2i} = 40, \quad \sum_{i=1}^{10} X_{3i} = 50,$$

e também que,

$$\sum_{i=1}^{10} X_{1i}^2 = 200, \quad \sum_{i=1}^{10} X_{2i}^2 = 300, \quad \sum_{i=1}^{10} X_{3i}^2 = 500,$$

utilize as propriedades de somatório e calcule,

$$\sum_{k=1}^{3} \sum_{i=1}^{10} \sum_{j=1}^{5} \left(X_{ki} - 3 \right)^2.$$

43) Dado que,

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}, \qquad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \qquad \text{e} \qquad \sum_{k=1}^{n} k^3 = \left[\frac{n(n+1)}{2} \right]^2,$$

utilize as propriedades de somatório e calcule,

$$\sum_{\substack{k=1 \\ k \neq 5}}^{10} \left(k^3 - k^2 + 2k \right).$$

44) Desenvolva e calcule o somatório a seguir,

$$\sum_{k=25}^{150} \left(\frac{1}{k+4} - \frac{1}{k+5} \right)$$

45) Dado que,

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \qquad \text{e} \qquad \sum_{k=1}^{n} k^3 = \left[\frac{n(n+1)}{2}\right]^2,$$

utilize estas fórmulas e as propriedades de somatório para calcular,

$$\sum_{k=10}^{80} \left(k^3 + k^2\right).$$

46) Dado que,

$$\sum_{\substack{i=1 \\ i\neq 5,12,20}}^{30} X_i = 100, \qquad \sum_{j=5}^{50} Y_j = 200 \qquad \text{e} \qquad \sum_{k=1}^{10} Z_k = 80,$$

utilize as propriedades de somatório e calcule,

$$\sum_{\substack{i=1 \\ i\neq 5,12,20}}^{30} \sum_{j=5}^{50} \sum_{k=1}^{10} \left(2X_i - X_i Y_j + Z_k\right).$$

47) Desenvolva e calcule os seguintes somatórios:

a)

$$\sum_{k=0}^{3} \left(5 + \sqrt{4^k}\right)$$

b)

$$\sum_{k=1}^{50} \left[\ln\left(k+3\right) - \ln\left(k+2\right)\right]$$

Respostas dos exercícios propostos

1)a) 40 1)b) 240 1)c) 1600 1)d) 8,89 1)e) 0 1)f) 80 1)g) 8,89 1)h) 4

2)a) 1 2)b) 24 2)c) 93

3)a) $30\,(1+2b)$ 3)b) 15 3)c) 159 3)d) 63cb 3)e) 150

4)a) 12 4)b) 29 4)c) 45 4)d) 6 4)e) 10 4)f) $\frac{17}{20}$ 4)g) 108864 4)h) 80

5)a) $\left(\sum\limits_{\substack{i=1 \\ i\neq 3}}^{4} \frac{X_i - Y_i}{2}\right)^2$ 5)b) $\prod\limits_{i=1}^{a} i$ 5)c) $\prod\limits_{i=1}^{3} \left(X_1 + Y_i\right)$ 5)d) $\left(\sum\limits_{i=1}^{2} X_i\right)\left(\sum\limits_{j=1}^{3} Y_j\right)$ 5)e) $\prod\limits_{i=1}^{n} \left(X_i Y_i\right)$

6)a) 192 6)b) 140 6)c) 19, 59 6)d) 746, 66

7)a) $5\left(3+\frac{1}{j}\right)$ 7)b) -18 7)c) 3025 7)d) 154440

8)a) 27 8)b) 144 8)c) 50 8)d) Não é possível

9)a) 62 9)b) 4

10) 4 e 6

11)a) 102 11)b) 30 11)c)8098 11)d) $15,93$

12)a) 24 12)b) 112 12)c) 39 12)d) 29 12)e) 128 12)f) 52

13) 50

14)a) $\sum\left(X-\bar{X}\right)^2=\sum X^2-2\bar{X}\sum X+n\bar{X}^2=\sum X^2-2n\bar{X}^2+n\bar{X}^2=\ldots.$

14)b) $S_Y^2=\frac{1}{n-1}\sum\left(Y-\bar{Y}\right)^2=\frac{1}{n-1}\sum\left(kX-k\bar{X}\right)^2=\ldots.$

15)a) $\approx 7,51$ 15)b) $4,2$ 15)c) $\approx 3,5$

16) Verifique que para $n=2$ é verdadeiro e assuma que para n é verdadeiro e então prove que para $n+1$ também é!

17) $24\sum X^2-4\sum X\sum Y+5\sum Y^2=84,76.$

18)a) 3054 18)b) 98

18)c) $\sum X\sum Y^2-2\sum X\sum Y+2\sum X-6\sum Y^2+12\sum Y-12=0$ 18)d) $6!/32=22,5$

19)a) 21 19)b) 20 19)c) 2

20)a) 300 20)b) 9 20)c) 1

21) 21600

22)a) -2 22)b) 189 22)c) 1425

23) $10\sum X+15\sum Y-10.15.2=20.$

24) $10.4\sum X^2-2.4\sum X\sum Y+5.4\sum Y^2-5.10\sum Z=6300.$

25) $2.40.216=17280.$

26)a) $\sum X\sum Y-2.20\sum Y-3.50\sum X+6.20.50+\sum\sum Z=720$

26)b) $60\{\sum Z^2-10\sum Z+8.25\}=720$ 26)c) $6!=720.$

27)a) $e^5\approx 148,41$ 27)b) 1215

28)a) 32 28)b) 280 28)c) 0

29) basta mostrar que $\sum(X-a)^2=\sum(X-\bar{X})^2+n(\bar{X}-a)^2.$

30)a) 1/21 30)b) 18.

31) $= 20000 - 64000 + 32000 - 19800 = -31800$

32)a) 30 32)b) $4 \times 85 = 340$ 32)c) 550

33)a)( ) V 33)b)( ) F 33)c)( ) F 33)d)( ) F 33)e)( ) V 33)f)( ) F

34) 1020

35) 1096

36) -43400

37) 323820

38)a) 11 38)b) 62

39) 18500

40) -2210

41) 29240

42) 2450

43) 2640

44) $\frac{126}{4495}$

45) 10669170

46) -8640

47)a) 35 47)b) $\ln\left(\frac{53}{3}\right)$

Capítulo 2

Estatística descritiva

2.1 Introdução

A parte da estatística que trabalha com a organização, resumo e apresentação dos dados é chamada de estatística descritiva. Nesta fase inicial do estudo, o pesquisador busca detectar padrões, sumarizar os dados e, apresentá-los por meio de tabelas, gráficos e medidas descritivas.

A Estatística é uma ciência que trata de métodos científicos para coleta, organização, resumo, apresentação e análise de dados, bem como na obtenção de conclusões válidas e na tomada de decisões razoáveis baseadas em tais análises. Podemos dividir a Estatística em duas áreas: estatística indutiva (inferência estatística) e estatística descritiva.

- **Estatística indutiva (Inferência Estatística)**: Se uma amostra é representativa de uma população, conclusões importantes acerca da população podem ser inferidas de sua análise. A parte da estatística que desenvolve metodologias inferenciais e trata das condições sob as quais essas inferências são válidas chama-se estatística indutiva ou inferência estatística. No capítulo 6 abordaremos a Regressão Linear Simples (RLS) e no capítulo 7 alguns Testes de Hipóteses, que são procedimentos inferenciais clássicos desta área da Estatística. Neste capítulo, estudaremos procedimentos e medidas da Estatística Descritiva.

- **Estatística descritiva**: É a área da Estatística que procura somente descrever e resumir amostras, sem tirar quaisquer conclusões ou inferências acerca da população. A Estatística Descritiva pode ser resumida no fluxograma apresentado na Figura 2.1.

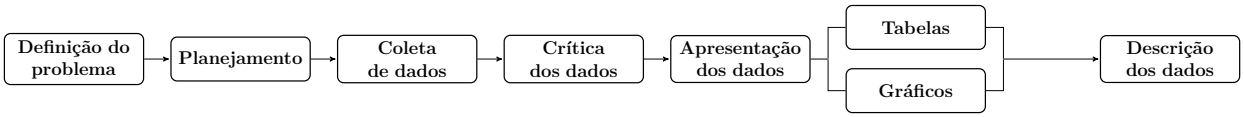


Figura 2.1: Fluxograma da Estatística Descritiva.

Após a definição do problema a ser estudado e o estabelecimento do planejamento da pesquisa, o passo seguinte é a coleta de dados.

- Coleta de dados: Consiste na busca ou compilação dos dados de variáveis. Podem ser experimentos conduzidos no campo ou em laboratórios; aplicação de questionários, etc.

- Crítica dos dados: Objetivando a eliminação de erros capazes de provocar futuros enganos de apresentação e análise, procede-se a uma revisão crítica dos dados, suprimindo os valores estranhos ao levantamento.

- Apresentação dos dados: Consiste em organizá-los de maneira prática e racional, para melhor entendimento do fenômeno que se está estudando. A apresentação dos dados pode ser feita por meio de tabelas e/ou gráficos, ambos podendo, ou não, estar associados a medidas descritivas.

- Descrição dos dados: A descrição dos dados é realizada com medidas que os representem de forma sumária, escolhidas de acordo com os objetivos do estudo.

No presente capítulo, apresentaremos as principais medidas descritivas de posição e dispersão, utilizadas em uma análise descritiva.

2.2 Medidas de posição

As medidas de posição ou de tendência central são medidas ou estatísticas empregadas para resumir um conjunto de dados (amostra) pela informação do posicionamento dos valores da amostra. Pode ser apresentado apenas um valor, ou preferivelmente, alguns valores que sejam "representativos" do conjunto de dados.

Dentre as principais medidas de posição utilizadas na Estatística, podem ser citadas, dentre outras:

- Médias (aritmética, geométrica e harmônica);
- Mediana;
- Moda;
- Quartis (25%, 50%, 75%);
- Decis (10%, 20%, ..., 90%);
- Percentis (1%, 2%, ..., 99%).

No presente texto abordaremos as três primeiras.

2.2.1 Média aritmética

A média aritmética é uma medida de tendência central dos dados, sendo portanto um valor em torno do qual os demais valores tendem a se concentrar.

Definição 1. Se $X_1, X_2, \ldots, X_n$ são n valores quaisquer da variável X, então a **média aritmética** de X, que denotaremos por $\bar{X}$, é dada por

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}, \tag{2.1}$$

ou seja, a média é simplesmente o total geral da amostral dividido igualmente entre todas as unidades.

Exemplo 1

Seja X_i a nota de um aluno em cada uma das $n = 4$ provas. As notas obtidas foram: Portanto

$$\bar{X} = \frac{\sum_{i=1}^{4} X_i}{4} = \frac{50 + 60 + 30 + 90}{4} = \frac{230}{4} = 57,5 \text{ pontos.}$$

Agora, se temos n observações da variável X, das quais f_1 são iguais a X_1, f_2 são iguais a $X_2, \ldots, f_k$ são iguais a X_k, então a média aritmética de X será dada por

$$\bar{X} = \frac{f_1X_1 + f_2X_2 + \cdots + f_kX_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum_{i=1}^{k} f_iX_i}{\sum_{i=1}^{k} f_i}. \tag{2.2}$$

Observação: Denota-se como **média aritmética ponderada** quando se considera f_i como sendo o peso ou ponderador de cada valor X_i.

Exemplo 2

Se um estudante obteve as seguintes notas nas quatro provas, qual foi a nota média?

Prova	Pesos (f_i)	Notas (X_i)
1ª	1	50
2ª	1	70
3ª	2	50
4ª	4	70

Neste exemplo é como se o peso fosse uma frequência, isto é, apesar de ter realizado apenas 4 provas, com pesos distintos, é como se ele tivesse realizado $n = \sum_{i=1}^{k} f_i = 8$ provas. Pela aplicação de (2.2) obtém-se

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{k} f_iX_i}{\sum_{i=1}^{k} f_i} = \frac{500}{8} = 62,5.$$

2.2.1.1 Propriedades da média aritmética

1) Se $Y_i = X_i + k$, então $\bar{Y} = \bar{X} + k$;

2) Se $W_i = kX_i$, então $\bar{W} = k\bar{X}$;

3) $SD_X\left(\bar{X}\right) = \sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$. A soma dos desvios (SD) com relação à média é igual a zero.

4) $\min_a (SQD_X(a)) = \min_a \left(\sum_{i=1}^{n} (X_i - a)^2 \right) = \sum_{i=1}^{n} (X_i - \bar{X})^2$. A soma dos quadrados dos desvios (SQD_X) com relação ao valor a é mínima, se $a = \bar{X}$.

As propriedades 1) e 2) podem ser facilmente demonstradas por propriedades de somatório.

1) $\bar{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n} = \dfrac{\sum_{i=1}^{n} (X_i + k)}{n} = \dfrac{\sum_{i=1}^{n} X_i}{n} + \dfrac{nk}{n} = \bar{X} + k;$

2) $\bar{W} = \dfrac{\sum_{i=1}^{n} kX_i}{n} = \dfrac{k \sum_{i=1}^{n} X_i}{n} = k\bar{X}.$

A demonstração da propriedade 3) por meio de propriedades de somatório é trivial, haja vista que, $\bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$, então $n\bar{X} = \sum_{i=1}^{n} X_i$, portanto,

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

Note que $\sum_{i=1}^{n} \left(X_i - \bar{X} \right) = \sum_{i=1}^{k} f_i \left(X_i - \bar{X} \right) = 0$. Sendo n e k definidos conforme a fórmula (2.2).

A demonstração da propriedade 4) é um pouco mais elaborada e se segue.

Seja $f(a) = SQD(a)$ a soma dos quadrados dos desvios em relação ao valor a. De um modo geral, para determinar os pontos críticos da função $f(\cdot)$ e determinar o(s) ponto(s) de máximo ou mínimo, o que se faz é derivá-la, isto é, obtém-se $f'(a) = \dfrac{df(a)}{da}$ e encontra-se o(s) valor(es) que a tornam igual a zero (ponto(s) crítico(s)). Depois, deve-se obter a segunda derivada $f''(a) = \dfrac{d^2 f(a)}{da^2}$ e verificar se $(a_0, f(a_0))$ é ponto de máximo, de mínimo, ou inflexão, pelo teste da segunda derivada.

Seja $f(a) = \sum_{i=1}^{n} (X_i - a)^2$. Sabemos que a derivada da soma é igual à soma das derivadas, então derivando, temos:

$$f'(a) = 2 \sum_{i=1}^{n} (X_i - a)(-1)$$

e,

$$f'(a) = -2 \sum_{i=1}^{n} (X_i - a)$$

Igualando $f'(a)$ a zero, para encontrarmos o valore crítico a_0, temos:

$$-2 \sum_{i=1}^{n} (X_i - a_0) = 0$$

$$\sum_{i=1}^{n} (X_i - a_0) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i = na_0 \quad \Rightarrow \quad a_0 = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}.$$

Podemos observar que o valor de a que anula a 1ª derivada é igual a $\bar{X}$. Apenas como um exercício ilustrativo, iremos verificar se $\bar{X}$ é a abscissa do ponto de máximo ou de mínimo. Isto será feito apenas por ilustração, pois uma soma dos quadrados não possui ponto de máximo. Neste caso, o valor de a que anula a 1ª derivada é obrigatoriamente um ponto de mínimo.

Obtendo-se a 2ª derivada, temos:

$$f''(a) = \frac{d}{da}\left(-2\sum_{i=1}^{n} X_i + 2na\right) = 2n.$$

Como $f''(a) > 0$, então, para $a_0 = \bar{X}$ temos a abscissa de um ponto de mínimo, ou seja, $f(a) = \sum_{i=1}^{n} (X_i - a)^2$ é um valor mínimo para $a = \bar{X}$.

Conforme já abordado no exercício 29) do Capítulo 1 de Somatório e Produtório, uma forma mais direta de se verificar esta propriedade da média aritmética é pela demonstração da seguinte igualdade (mostre e argumente!),

$$SQD(a) = \sum_{i=1}^{n} (X_i - a)^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - a\right)^2, 0 < a < +\infty.$$

Em outras palavras, isto significa que uma soma dos quadrados dos desvios tomados em relação à média aritmética é mínima. Esta propriedade, desejável em muitas situações, fornece um critério para obtenção de medidas mais representativas de um conjunto de dados chamado critério de mínimos quadrados.

No cálculo da média aritmética participam todos os valores observados, o que não acontece com todas as medidas de tendência central. Valores muito discrepantes dentro de um conjunto de dados tendem a exercer influência desproporcional sobre a média aritmética.

Para qualquer conjunto de dados, sempre será possível calcular a sua média aritmética e existirá somente uma média para cada conjunto de observações. É uma medida de fácil interpretação e presta-se muito bem a tratamentos adicionais. É o ponto de equilíbrio de uma distribuição de um conjunto de observações. Isto faz com que a média seja uma medida descritiva tão mais eficiente quanto mais simétrica for a distribuição das observações ao seu redor.

2.2.1.1.1 Exemplos de aplicação das propriedades da média

Exemplo 3

Um professor irá divulgar as notas (X_i) dos $n = 500$ alunos matriculados na disciplina. Porém, ele constatou que a média dos alunos foi $\bar{X} = 56,4$ e decidiu corrigir todas as notas antes de divulgá-las, de modo que a nota média fosse 60. Pede-se: o que pode se realizado?

Opção 1: ele pode acrescentar $k = 60 - 56,4 = 3,6$ em todas as notas, isto é, $Y_i = X_i + 3,6$, já que

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\left(\sum_{i=1}^{n} X_i + 3,6\right)}{n} = \frac{\sum_{i=1}^{n} X_i}{n} + \frac{\sum_{i=1}^{n} n \cdot 3,6}{n} = \bar{X} + 3,6 = 56,4 + 3,6 = 60.$$

Opção 2: ele pode multiplicar todas as notas por $k = \frac{60}{56,4} \cong 1,064$, isto é,$W_i = 1,064X_i$, já que

$$\bar{W} = \frac{\sum_{i=1}^{n} W_i}{n} = \frac{\sum_{i=1}^{n} 1,064X_i}{n} = \frac{1,064\sum_{i=1}^{n} X_i}{n} = 1,064 \cdot 56,4 \cong 60.$$

Observação: Veremos posteriormente, quando estudarmos as propriedades da variância amostral (S^2), que a opção 1 não altera a variância das notas, porém na opção 2 a variância será alterada, neste caso ela ficará multiplicada por $(1,064)^2$.

Exemplo 4

Numa empresa com $n = 1500$ funcionários, os salários (X_i) possuem uma média mensal de $\bar{X} = 5200$ reais, ou seja, este é o valor médio dos salários mensais. Se a empresa está em um momento de crise e decide reduzir todos os salários em $k = 150$ reais, para evitar demissões, qual será o novo salário médio?

Solução: Como $Y_i = X_i - k = X_i - 150$ serão os novos salários, então o novo salário médio será

$$\begin{aligned} \bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i=1}^{n} (X_i - 150)}{n} = \frac{\sum_{i=1}^{n} X_i}{n} - \frac{n \cdot 150}{n} \\ &= \bar{X} - 150 = 5200 - 150 = 5050. \end{aligned}$$

Exemplo 5

Considere o exemplo 4, mas ao invés de reduzir todos os salários em 150 reais, fosse decidido uma redução de 15% em todos os salários, qual seria o novo salário médio?

Solução: reduzir em 15% significa multiplicar por $k = 0,85$ (verifique isto utilizando regra de três), portanto

$$\begin{aligned} \bar{W} &= \frac{\sum_{i=1}^{n} W_i}{n} = \frac{\sum_{i=1}^{n} 0,85X_i}{n} = 0,85\frac{\sum_{i=1}^{n} X_i}{n} \\ &= 0,85\bar{X} = 0,85 \cdot 5200 = 4420. \end{aligned}$$

2.2.2 Média geométrica

Sejam $X_1, X_2, \ldots, X_k$, valores da variável X, associados as frequências $f_1, f_2, \ldots, f_k$, respectivamente, com $X_i > 0$, para todo i. A média geométrica de X, que denotaremos por $\bar{X}_G$ é dada por:

$$\bar{X}_G = \sqrt[\sum_{i=1}^{k} f_i]{X_1^{f_1} \cdot X_2^{f_2} \cdot \ldots \cdot X_k^{f_k}} = \sqrt[\sum_{i=1}^{k} f_i]{\prod_{i=1}^{k} X_i^{f_i}} \tag{2.3}$$

Em particular, se $f_1 = f_2 = \ldots = f_k = 1$, $(n = k)$ temos:

$$\bar{X}_G = \sqrt[n]{X_1 \cdot X_2 \cdot \ldots \cdot X_n}$$

sendo n o tamanho da amostra, ou o número total de observações e k é o número de distintos valores, em que f_i é a frequência do valor X_i.

Os resultados i) e ii) a seguir, são dois fatos verdadeiros cuja demonstração é deixada a cargo do leitor.

i) $\log \bar{X}_G = \dfrac{1}{\sum\limits_{i=1}^{k} f_i} \cdot \sum\limits_{i=1}^{k} f_i \cdot \log X_i$ e, caso $f_i = 1, \forall i, 1 \leq i \leq k$ então $\log \bar{X}_G = \dfrac{1}{n} \sum\limits_{i=1}^{n} \log X_i$.

ii) Se $Y_i = \dfrac{X_i}{\bar{X}_G}$ então $\prod\limits_{i=1}^{n} Y_i = 1$.

iii) A média geométrica deve ser utilizada para o cálculo da média de valores que formam uma série, isto é, cujos valores, se apresentam segundo uma progressão geométrica ou, revelam elementos "muito grandes", comparativamente com os demais, como por exemplo, a série $18, 20, 22, 24$ e 850 (em que a média geométrica é aproximadamente igual a 43,8, resultado que não foi tão influenciado assim pelo elemento 850), sendo também utilizada para o cálculo de índices do custo de vida.

A principal inconveniência da média geométrica, consiste no fato de ela ser grandemente influenciada, pelos elementos "pequenos" da série, se for o caso.

2.2.2.1 Exemplos de aplicação da média geométrica

Exemplificaremos, sua utilização na Geometria e na Economia.

Aplicação na Geometria

Considere um retângulo de lados medindo, p e q, a média geométrica destes dois números é o tamanho do lado de um quadrado cuja área é igual à área do retângulo com lados de tamanho p e q.

De forma similar, podemos considerar um paralelepípedo de lados medindo a, b, e c, a média geométrica destes três números nos dará o tamanho dos lados de um cubo (aresta), cujo volume, é o mesmo que o de um paralelepípedo com lados de tamanho igual aos três números dados.

Exemplo 6

Suponha que um paralelepípedo tenha dimensões 3 m, 8 m e 9 m. Deseja-se construir um cubo com o mesmo volume deste paralelepípedo e, assim sendo, qual deve ser a medida da aresta do cubo?

Neste caso devemos ter que a aresta d, será dada por:

$$d = \sqrt[3]{3 \cdot 8 \cdot 9} = 6.$$

Assim o paralelepípedo tem volume $V_p = 3 \cdot 8 \cdot 9 = 216$ m^3 e o cubo de aresta $d = 6$ m tem volume $V_c = 6^3 = 216$ m^3.

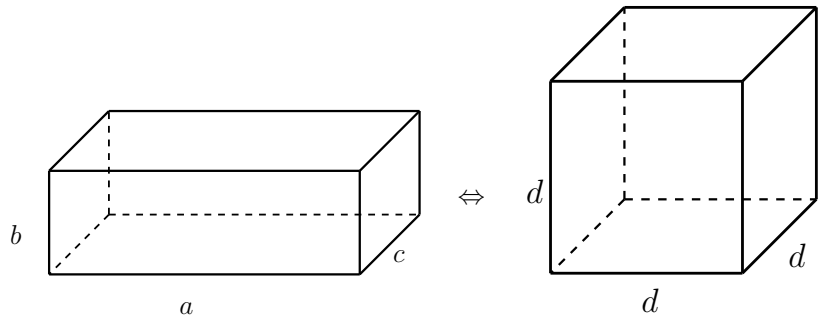


Figura 2.2: Paralelepípedo de dimensões a, b e c e cubo de aresta d, ambos com mesmo volume.

Aplicação na Economia: Índice médio

Considere um valor inicial VI e uma *taxa constante de juros* igual a $r \times 100\%$, com $0 < r < 1$. O valor final VF, após n juros é dado por,

$$VF = VI\,(1+r)^n, \quad \text{em que} \quad 1+r = \text{índice de juros.} \tag{2.4}$$

Por exemplo, um valor inicial $VI = 200$ com juros iguais a $1,5\%$ ao mês, após $n = 12$ meses resultará no seguinte valor final,

$$VF = 200\,(1+0,015)^{12} \approx 239,124$$

Considere agora que sobre um valor inicial $VI = 4000$ incidem *juros variáveis* por quatro mêses, conforme a tabela a seguir

	Valores		$(r \times 100\%)$	$(1+r)$
Mês	Inicial	Final	Juros%	Índice
1	4000,00	4060,00	1,5	1,015
2	4060,00	4222,40	4,0	1,040
3	4222,40	4631,97	9,7	1,097
4	4631,97	5336,03	15,2	1,152

O índice médio é o valor $1+r$ que aplicado na equação (2.4) com VI=4000, fornece o valor final correto VF=5336,03. Ou seja, seria como se os juros mensais fossem constantes (juros sobre juros).

O índice médio **aritmético** fornece o valor final incorreto,

$$\begin{aligned} \bar{I} &= \frac{1,015+1,040+1,097+1,152}{4} = 1,076\ (7,6\%) \\ VF &= 4000\,(1,076)^4 \approx 5361,78 \end{aligned}$$

O índice médio **geométrico** fornece o valor final correto,

$$\begin{aligned} \bar{I}_G &= \sqrt[4]{1,015.1,040.1,097.1,152} \approx 1,074706\ (7,47\%) \\ VF &= 4000\,(1,074706)^4 \approx 5336,03 \end{aligned}$$

2.2.3 Média harmônica

Sejam $X_1, X_2, \ldots, X_k$, valores distintos de X, associados às frequências absolutas f_1, f_2, $\ldots$, f_k, respectivamente. A média harmônica de X é dada por:

$$\bar{X}_H = \frac{f_1 + f_2 + \ldots + f_k}{\frac{f_1}{X_1} + \frac{f_2}{X_2} + \cdots + \frac{f_n}{X_k}} = \frac{\sum_{i=1}^{k} f_i}{\sum_{i=1}^{k} \frac{f_i}{X_i}}$$

Em particular, se $f_1 = f_2 = \ldots = f_n = 1$, então:

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

A média harmônica é particularmente recomendada para série de valores que são inversamente proporcionais, como por exemplo, para o cálculo da velocidade média e custo médio de bens comprados com uma quantia fixa.

Exemplo 7

Sejam os valores 2, 4 e 8. Calcular a média harmônica.

$$\bar{X}_H = \frac{3}{\frac{1}{2} + \frac{1}{4} + \frac{1}{8}} = \frac{3}{\frac{7}{8}} \cong 3,43$$

Exemplo 8

Considere que uma viagem da localidade A para a localidade B é realizada a uma velocidade média igual a V_1 e que o retorno de B para A, a uma velocidade média V_2. Considere que a distância entre A e B seja igual a d. Pede-se: qual é a velocidade média do trajeto completo $A \longrightarrow B \longrightarrow A$? Como velocidade é a razão distância pelo tempo, sejam t_1 e t_2 os tempos decorridos de tal forma que,

$$V_i = \frac{d}{t_i} \quad \text{para} \quad i = 1, 2$$

portanto,

$$\text{velocidade média} = \frac{2d}{t_1 + t_2} = \frac{2d}{\frac{d}{V_1} + \frac{d}{V_2}} = \frac{2}{\frac{1}{V_1} + \frac{1}{V_2}}$$

é a média harmônica das velocidades V_1 e V_2. De uma maneira geral, se forem considerados trajetos com distâncias d_i, cujas velocidades médias foram V_i, com respectivos tempos t_i para $i = 1, 2, \ldots, n$,

$$\text{velocidade média} = \bar{V}_H = \frac{d}{\sum_{i=1}^{n} t_i} = \frac{d}{\sum_{i=1}^{n} \frac{d_i}{V_i}} \quad \text{em que} \quad d = \sum_{i=1}^{n} d_i.$$

Este cálculo será equivalente à média harmônica das velocidades somente se as distâncias forem todas iguais ($d_1 = d_2 = \ldots = d_n$).

Para o exemplo anterior, se tomarmos $V_1 = 60$ km/h e $V_2 = 90$ km/h, a velocidade média correta é a média harmônica $\bar{V}_H$ e não a média aritmética $\bar{V}$.

$$\bar{V}_H = \frac{2}{\frac{1}{60} + \frac{1}{90}} = 72 \text{ km/h} \qquad \bar{V} = \frac{60 + 90}{2} = 75 \text{ km/h}.$$

Exemplo 9

Considere que se dispõe de um valor fixo (orçamento) para se adquirir um determinada item num período, por exemplo, de uma semana, ou um mês ou um ano. Os dados hipotéticos da tabela a seguir exemplificam esta situação:

Período	Orçamento	Valor do item
1°	500	20
2°	500	25
3°	500	50

Observe então que foram adquiridos $\frac{500}{20} = 25$ itens no 1° período, $\frac{500}{25} = 20$ itens no 2° período e, $\frac{500}{50} = 10$ itens no 3° período. Qual foi o custo médio do item considerando-se os três períodos?

Solução: o custo médio é, obviamente o total gasto dividido pelo número total de itens adquiridos, assim

$$\text{Custo médio} = \frac{1500}{25 + 20 + 10} \cong 27,273 \text{por item.}$$

Neste caso, o cálculo equivale ao de uma média aritmética dado pela fórmula (2.2), com $X_i =$ "valor do item" e $f_i =$ "número de itens adquiridos". Mas observe que,

$$\text{Custo médio} = \frac{1500}{25 + 20 + 10} = \frac{3 \cdot 500}{\frac{500}{20} + \frac{500}{25} + \frac{500}{50}} = \frac{3}{\frac{1}{20} + \frac{1}{25} + \frac{1}{50}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} = \bar{X}_H,$$

ou seja, é a média harmônica dos valores dos itens nos três períodos.

2.2.4 Mediana

A palavra mediana é sinônimo de "metade" e, para um conjunto de valores colocados em ordem crescente ou decrescente de grandeza, a mediana (Md) ou o valor mediano, é o elemento que ocupa a posição central. Numa amostra de n observações, a interpretação da mediana é a seguinte: pode-se afirmar que pelo menos 50% das observações da amostra são valores iguais ou superiores à mediana e, pelo menos 50% das observações da amostra são valores iguais ou inferiores ao valor mediano.

Considere a notação $X_{(i)}$ para indicar estatísticas de ordem da amostra, isto é: $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)}$, esta notação será utilizada nesta seção. Note que

- $X_{(1)}$ indica o menor valor do conjunto, isto é, $X_{(1)} =$ mínimo $\{X_1, X_2, X_3, \ldots, X_n\}$;
- $X_{(2)} =$ segundo menor valor de $\{X_1, X_2, X_3, \ldots, X_n\}$;

- $\vdots$

- $X_{(n)}$ indica o maior valor do conjunto, isto é, $X_{(n)} = \text{máximo}\{X_1, X_2, X_3, \ldots, X_n\}$.

2.2.4.1 Mediana para o caso em que n é ímpar

Neste caso, o valor mediano, Md é o elemento que ocupa a posição $\dfrac{n+1}{2}$ quando os dados estão ordenados, ou seja, $\text{Md}(X) = X_{\left(\frac{n+1}{2}\right)}$.

Exemplo 10

Considere uma variável assumindo os seguintes valores:

$$X = \{14, 0, 5, 10, 7, -8, 9\}.$$

Determine a mediana do conjunto X.

Primeiramente devemos colocar os dados em rol, o qual é

$$-8, 0, 5, 7, 9, 10, 14.$$

Como $n = 7$ é ímpar, a mediana é o elemento que ocupa a posição $\frac{n+1}{2} = \frac{7+1}{2} = 4$, quando os dados estão organizados em rol, assim:

$$\text{Md}(X) = X_{\left(\frac{7+1}{2}\right)} = X_{(4)} = 7.$$

Exemplo 11

Suponha que temos os dados abaixo e desejamos encontrar a mediana deste conjunto de dados.

Tabela 2.1: Conjunto de dados agrupados do exemplo 11

X_i	f_i
45	10
970	22
99	33
10	40

Note que nosso conjunto de dados é da seguinte maneira:

$$\underbrace{45, \cdots, 45}_{\text{10 vezes}}, \underbrace{970, \cdots, 970}_{\text{22 vezes}}, \cdots, \underbrace{10, \cdots, 10}_{\text{40 vezes}}$$

e, para determinarmos a mediana, a primeira coisa a se fazer é colocar os dados em rol, conforme na tabela a seguir,

Tabela 2.2: Rol do conjunto de dados agrupados do exemplo 11

$X_{(i)}$	f_i	$\sum f_i$
10	40	40
45	10	50
99	33	83
970	22	105

Como $n = 105$ é ímpar, a mediana é o elemento que ocupa a posição $\frac{n+1}{2} = \frac{105+1}{2} = 53$, quando os dados estão organizados em rol, assim:

$$\text{Md}(X) = X_{\left(\frac{105+1}{2}\right)} = X_{(53)} = 99.$$

A terceira coluna da tabela 2.2 é apenas um artifício utilizado para facilitar a identificação do valor X_i que ocupa a posição $\frac{n+1}{2} = 53$. $\sum f_i = 40$ indica que do 1° ao 40° valor, todos são iguais a 10; $\sum f_i = 50$ indica que do 1° ao 40° valor, todos são iguais a 50, e assim por diante.

2.2.4.2 Mediana para o caso em que n é par

Neste caso, a mediana Md é a média aritmética dos valores centrais de ordem $\frac{n}{2}$ e $\frac{n}{2} + 1$, isto é,

$$Md(X) = \frac{X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}}{2}.$$

Exemplo 12

Seja uma variável X assumindo os seguintes valores $X = \{8, 10, 5, 7, 15, 14\}$. Obter a mediana. (R: Md(X)=9)

Exemplo 13

Considerando a tabela a seguir, obter a mediana: (R: Md(X)=87)

X_i	87	90	82	89	85
f_i	15	4	5	8	10

2.2.5 Moda

O termo moda foi utilizado primeiramente por Karl Pearson, influenciado pelo uso popular do termo, com o significado de objeto que mais se está usando no momento.

A moda ou valor modal (Mo) é o valor mais frequente do conjunto de valores observados. Em algumas situações, a distribuição das observações é tal que as frequências são maiores nos extremos. Nesses casos, a utilização apenas da média e da mediana é contra-indicada, pois são valores pouco representativos do conjunto e o uso da moda poderá, ser considerado.

Com relação à moda, uma série de dados pode ser classificada em:

- Amodal: não possui moda;

- Unimodal: possui apenas uma moda;

- Bimodal: possui duas modas;

- Multimodal: possui mais de duas modas.

Com softwares que realizam análises estatísticas à disposição, numa análise descritiva obtém-se não somente o valor modal, mas sim os $\alpha\%$ valores mais frequentes e também os $\alpha\%$ menos frequentes. Por exemplo, os 5% mais e menos frequentes de cada amostra. Para dados climáticos como temperaturas máximas, precipitações máximas e etc., o valor modal é uma medida descritiva importante.

Exemplo 14

Identificar a moda para os dados apresentados a seguir:

X_i	0	1	2	3	4	5
f_i	4	5	7	3	2	1

(Resposta: $\text{Mo}(X) = 2$)

Exemplo 15

Considerando a Tabela abaixo, obter a moda:

X_i	3	5	9	10	15
f_i	1	4	2	4	3

(R: Série bimodal $\Rightarrow \text{Mo}_1(X) = 5$ e $\text{Mo}_2(X) = 10$)

Relação entre média, mediana e moda

1. Distribuição simétrica: $\bar{X} = Mo = Md$

2. Distribuição assimétrica: a média e a mediana se deslocam.

   (a) Assimetria positiva: $\bar{X} > Md > Mo$;

   (b) Assimetria negativa: $\bar{X} < Md < Mo$.

2.3 Medidas de dispersão

Introdução

As medidas de dispersão, ou de variabilidade, são estatísticas descritivas que quantificam de algum modo a variabilidade dos dados, geralmente utilizando como referência uma medida de posição.

Caracterizar um conjunto de dados apenas por medidas de posição é inadequado e perigoso, pois, conjuntos com medidas de posição semelhantes podem apresentar características muito diferentes.

2.3.1 Variância amostral

A variância mede a dispersão dos valores em torno da média. Ela é dada pela soma dos quadrados dos desvios em relação à média aritmética, dividida pelo número de graus de liberdade. É a medida de dispersão mais utilizada, fácil de calcular e compreender, além de ser bastante empregada na inferência estatística.

Para uma amostra de n valores, $X_1, X_2, \cdots, X_n$, a variância amostral é dada por:

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

Se aos valores $X_1, X_2, \cdots, X_k$ estiverem associados às frequências f_1, f_2, $\cdots$, f_k, a variância amostral será dada por:

$$S_X^2 = \frac{SQD_X}{n-1} = \frac{\sum_{i=1}^{k} f_i X_i^2 - \frac{\left(\sum_{i=1}^{k} f_i X_i\right)^2}{\sum_{i=1}^{k} f_i}}{\sum_{i=1}^{k} f_i - 1}.$$

Para denotar a variância da amostra de valores X_i, utiliza-se a notação S_X^2, porém denotar como $S^2(X)$, $\widehat{V}(X)$ ou mesmo $\widehat{\sigma}_X^2$, são outras alternativas comumente utilizadas em textos.

Fatos:

i) Muitos autores definem a variância da amostra usando n como denominador no cálculo de variância. Esta opção é igualmente válida. Neste caso, é comum se escrever S_n^2 e S_{n-1}^2, para distingui-las. Foge ao nível pretendido para este texto explicar a diferença entre as duas opções, entretanto, podemos citar que é possível demonstrar que, utilizando-se o denominador $n-1$, obtém-se um estimador não tendencioso da variância populacional, isto é, $E\left(S_{n-1}^2\right) = \sigma^2$ e $E\left(S_n^2\right) \neq \sigma^2$.

ii) De uma maneira geral, o número de graus de liberdade associados a uma estatística é o número de elementos da amostra, n, menos o número de parâmetros (medidas da população) já estimados. Existem $n-1$ desvios independentes no cálculo da variância amostral. Por exemplo, se $n = 3$ e $\bar{X}$ for utilizado no cálculo de uma estatística, então X_1 e X_2 podem ser quaisquer observações, porém X_3 está restrito, pois $\frac{X_1+X_2+X_3}{3} = \bar{X}$. Portanto, perde-se um grau de liberdade.

2.3.1.1 Propriedades da variância

Algumas propriedades úteis da variância são:

i) A variância é sempre maior ou igual a zero, isto é, $S_X^2 \geq 0$;

ii) Para $Y = X + k$, sendo k uma constante, $S_Y^2 = S_X^2$;

iii) Para $Y = kX$, sendo k uma constante, $S_Y^2 = k^2 S_X^2$.

Exemplo 16

Considere as amostras A e B a seguir e determine S_A^2 e S_B^2.

$$\begin{aligned} A &= \{4, 8, 3, 9, 7, 5\} \\ B &= \{1, 5, 2, 14, 3, 11\} \end{aligned}$$

Temos que

$$\begin{aligned} S_A^2 &= \frac{244 - \frac{(36)^2}{6}}{6-1} = 5,6 \\ S_B^2 &= \frac{356 - \frac{(36)^2}{6}}{6-1} = 28 \end{aligned}$$

Exemplo 17

Para os dados da Tabela a seguir, calcule a variância.

X_i	2	4	5	6	7	8
f_i	1	3	3	1	1	1

(R.: $S_X^2 \cong 2,89$)

2.3.2 Desvio padrão amostral

Como medida de dispersão, a variância tem a desvantagem de apresentar unidade de medida igual ao quadrado da unidade de medida dos dados. Assim, por exemplo, se os dados são medidos em metros, a variância é dada em metros ao quadrado. Para voltarmos à unidade de medida original, precisamos de uma outra medida de dispersão. Então, se define desvio padrão como a raiz quadrada positiva da variância.

$$S_X = \sqrt{S_X^2} > 0.$$

O desvio padrão é muito utilizado em procedimentos inferenciais, tais como na obtenção de intervalos de confiança (por exemplo: margens de erros em pesquisas eleitorais).

2.3.3 Coeficiente de variação

Frequentemente, se tem o interesse em comparar variabilidades de diferentes conjuntos de valores. A comparação se torna difícil em situações onde as médias são muito desiguais ou as unidades de medida são diferentes. Nesses casos, o coeficiente de variação, denotado por CV, é indicado por ser uma medida de dispersão relativa.

O CV expresso em percentagem é dado por:

$$CV_X(\%) = \frac{S_X}{\bar{X}} \cdot 100\%$$

Note que o CV é o desvio padrão expresso em percentagem do valor da média. É uma medida adimensional.

Aplicação:

- Utilizado para avaliação da precisão de experimentos;

- Utilizado para analisar qual amostra é mais homogênea (menor variabilidade relativa). Na situação em que as amostras possuem a mesma média, a conclusão pode ser feita a partir da comparação de suas variâncias. Para amostras com médias diferentes, aquela que apresentar menor CV, é a mais homogênea.

Exemplo 18

Duas turmas A e B da disciplina EST 105 apresentaram as seguintes estatísticas na primeira prova:

	Turmas	
Estatísticas	A	B
n	50	60
$\bar{X}$	65	70
S_X^2	225	235

Qual é a turma com uma distribuição mais homogênea das notas?

Solução:

$$CV_A = \frac{\sqrt{225}}{65} \cdot 100\% = 23,08\%$$

$$CV_B = \frac{\sqrt{235}}{70} \cdot 100\% = 21,90\%$$

Assim, a turma mais homogênea é a B, pois é a que possui menor coeficiente de variação.

2.3.4 Erro-padrão da média

É uma medida utilizada para avaliar a precisão da média. Tecnicamente, $S\left(\bar{X}\right)$ é a variância da distribuição amostral de $\bar{X}$. Ou seja, considere infinitas amostras de tamanho n obtidas de uma população. $S\left(\bar{X}\right)$ é a variância da distribuição dos valores $\bar{X}_1, \bar{X}_2, \cdots$. É dada por:

$$S\left(\bar{X}\right) = \sqrt{\frac{S_X^2}{n}} = \frac{S_X}{\sqrt{n}}.$$

X_i	X_1	X_2	$\cdots$	X_n
Y_i	Y_1	Y_2	$\cdots$	Y_n

Exemplo 19

Considerando $S_A^2 = 5,6$ e $S_B^2 = 28$, temos que:

$$\begin{aligned} S\left(\bar{X_A}\right) &= \frac{2,3664}{\sqrt{6}} = 0,966 \\ S\left(\bar{X_B}\right) &= \frac{5,2915}{\sqrt{6}} = 2,1602 \end{aligned}$$

Note que o erro-padrão da média é:

- Inversamente proporcional ao tamanho da amostra;
- Diretamente proporcional à variância da amostra.

2.3.5 Amplitude total

A amplitude total (AT) é dada pela diferença entre o maior e o menor valor de uma amostra ou de um conjunto de dados. Se X_1, X_2, X_3, ..., X_n é uma amostra de valores da variável X, então:

$$AT_X = X_{(n)} - X_{(1)}$$

Recorde que a notação $X_{(i)}$ indica estatísticas de ordem da amostra, isto é: $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)}$. Portanto, a amplitude total indica que o desvio entre duas observações quaisquer é no máximo igual a AT.

Exemplo 20

Considere as amostras A e B dadas por $A = \{4, 8, 3, 9, 7, 5\}$ e $B = \{1, 5, 2, 14, 3, 11\}$. Determine a amplitude total de A e B.

$$\begin{aligned} AT_A &= 9 - 3 = 6 \\ AT_B &= 14 - 1 = 13. \end{aligned}$$

2.4 Coeficiente de correlação amostral (Pearson)

O coeficiente de correlação (r ou $\widehat{\rho}$) mede o grau de associação linear entre duas variáveis aleatórias X e Y. O conceito de variável aleatória será apresentado adiante no texto (ver capítulo 4).

Sejam duas amostras relativas sá variáveis X e Y, dadas a seguir:

O coeficiente de correlação entre os valores de X e Y é dado por:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2 \cdot S_Y^2}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1}.\frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X \cdot SQD_Y}} \qquad -1 \leq r_{XY} \leq 1$$

em que:

$S_{XY} = \widehat{COV}(X,Y)$ é a covariância amostral, o estimador da covariância populacional, $COV(X,Y)$, conceito abordado na Seção 4.8.2 do Capítulo 4 deste texto.

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} \quad \text{e} \quad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

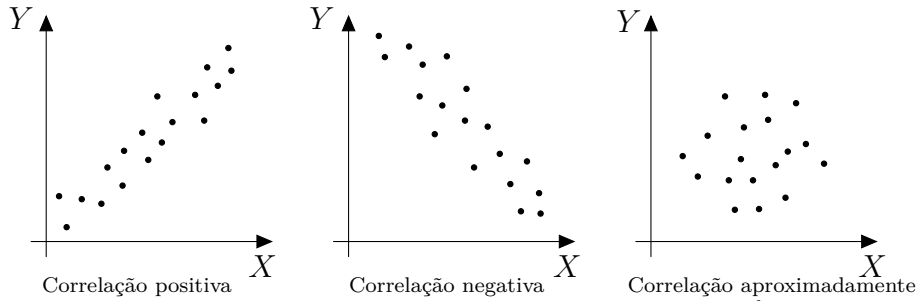


Figura 2.3: Representação gráfica de diversos coeficientes de correlação.

Exemplo 21

Considere a amostra a seguir e determine o coeficiente de correlação entre X e Y

Amostra X	4	8	3	9	7	5
Amostra Y	1	5	2	14	3	11

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 252 - \frac{(36)\,(36)}{6} = 36$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = 244 - \frac{(36)^2}{6} = 28$$

$$SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 356 - \frac{(36)^2}{6} = 140$$

$$r_{XY} = \frac{SP_{XY}}{\sqrt{SQD_X \cdot SQD_Y}} = \frac{36}{\sqrt{28 \cdot 140}} = 0,5750.$$

A correlação entre as variáveis X e Y foi $0,575$. Logo, a associação é positiva. A tendência é de que aumentando-se X, Y aumenta. Obviamente que, para verificar se esta correlação é estatisticamente significativa, torna-se necessário aplicar um teste de significância.

2.5 Exercícios propostos com respostas

1) A tabela a seguir apresenta os tempos de duração de chamadas telefônicas (em minutos), obtidos com uma amostra de oito telefonemas.

Telefonema	Tempo (min.)	Telefonema	Tempo (min.)
1	1	5	8
2	3	6	1
3	6	7	4
4	15	8	2

Calcule e interprete:

a) O tempo médio (aritmético).

b) O tempo mediano.

c) O tempo modal.

d) O erro-padrão da média.

e) O coeficiente de variação da amostra.

2) Assinale (V) se a afirmativa for totalmente verdadeira ou (F) caso contrário e indique onde deve ser corrigido.

a) ( ) Para valores $x_1, x_2, \ldots, x_n$ tais que $x_i > 0 \; \forall \; i$, tem-se que $\bar{X}_H \leq \bar{X}_G \leq \bar{X}$.

b) ( ) A variância amostral mede a dispersão em torno da média aritmética e resulta sempre em um valor não negativo.

c) ( ) Quanto ao valor mediano (Md) para uma amostra com n observações, pode-se afirmar que há $n/2$ observações maiores e também $n/2$ observações menores que Md.

d) ( ) O coeficiente de correlação linear é adimensional e o desvio padrão é expresso na mesma unidade de medida dos dados.

e) ( ) O erro-padrão da média é uma medida de dispersão que informa a precisão com que a média é estimada, pois representa o desvio padrão da distribuição amostral da média.

f) ( ) As amostras $A : \{15, 13, 10, 7, 4\}$ e $B : \{105, 103, 100, 97, 94\}$ possuem variâncias $S_A^2 = S_B^2 = 19,7$ e portanto são duas amostras com igual homogeneidade ou dispersão relativa.

3) Calcule as médias harmônica, geométrica e aritmética da seguinte amostra,

frequência	3	2	1	4
valor	2	3	5	1

4) Em um *Painel Sensorial* indivíduos treinados avaliam (degustam) determinado produto e atribuem uma nota de acordo com a percepção do sabor: 0=muito ruim, 1=ruim, 2=regular,

3=bom, 4=muito bom e 5=excelente. Na tabela a seguir são informadas as notas obtidas com um determinado azeite de oliva,

0	0	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	4	4	4	4	5	5	5	5	5	5	5	5

Sumarize as notas com duas medidas de posição e duas de dispersão e interprete os valores calculados.

5) Em 1930 foi disputada a primeira copa do mundo de futebol no Uruguai. Foram disputadas, até a copa de 1998 na França, um total de 16 copas, sendo que no período entre 1938-1950 a competição não foi realizada devido à segunda guerra mundial. Na tabela a seguir é informado o número de vezes que cada país terminou a competição entre os cinco primeiros colocados; observe que somente 24 países obtiveram tal desempenho. Os dados são reais e são consideradas todas as 16 copas disputadas no período 1930-1998. (FONTE: http://www.gazetaesportiva.net/copa2002/historia/indice.htm-acesso em maio de 2002)

PAÍS	N^o de vezes 1^o ao 5^o colocado
Alemanha	10
Brasil	11
Itália	8
Grupo A	5
Grupo B	4
Grupo C	3
Grupo D	2
Grupo E	1

Grupo A (2 países)-Argentina e Suécia; Grupo B (3 países)-França, Iugoslávia e Uruguai; Grupo C (4 países)- Tchecoslováquia, Holanda, Polônia e URSS; Grupo D (5 países)-Áustria, Chile, Espanha, Hungria e Inglaterra; Grupo E (7 países)-Bélgica, Bulgária, Croácia, EUA, País de Gales, Portugal e Suíça.

a) Calcule o número médio de participações terminando entre os cinco primeiros colocados, isto é, a média do N^o de vezes 1^o ao 5^o colocado dos países.

b) Calcule o erro-padrão da média.

c) Calcule o número mediano e o número modal.

d) A média aritmética é uma boa medida representativa (de posição) dos números da tabela? **SIM ou NÃO**? Justifique sua resposta.

6) Qual das duas amostras é a mais homogênea, isto é, a de menor dispersão relativa? Justifique sua resposta.

	valores x_i								$\sum x_i$	$\sum x_i^2$
Amostra 1:	95	90	84	82	79	73	71	60	634	51116
Amostra 2:	95	66	66	65	65	64	62	60	543	37727

7) Na tabela a seguir são informadas as notas de uma amostra de 18 alunos. Calcule as medidas de posição e dispersão abordadas e interprete o significado do valor encontrado, e/ou explique qual é a informação dada pela medida.

Nota	Nº de Alunos	Nota	Nº de Alunos
59	1	68	2
60	1	72	1
61	1	73	2
64	1	91	3
65	3	99	1
67	1	100	1

8) As estatísticas descritivas apresentadas na tabela a seguir são referentes à duas variáveis, X e Y, avaliadas em n unidades experimentais.

	Variáveis	
Estatísticas	X	Y
média aritmética	12	14
mediana	10	15
erro-padrão da média	0,6	1,12
coeficiente de variação	50%	80%

Assinale com V se a afirmativa estiver totalmente correta ou assinale F caso contrário e indique o(s) erro(s).

a) ( ) A amostra de valores X apresenta uma menor dispersão relativa ou maior homogeneidade.

b) ( ) $n = 150$ unidades experimentais foram avaliadas.

c) ( ) $S_X^2 = 36$ e $S_Y^2 = 11,2$.

d) ( ) Se for informado o valor de $\sum_{i=1}^{n} X_i Y_i$ pode-se calcular o coeficiente de correlação linear entre os valores das amostras X e Y.

e) ( ) A amplitude total da amostra X é maior porque a variância é maior.

f) ( ) O número de observações ≤ 10 na amostra de valores X é igual ao número de observações ≤ 15 na amostra de valores Y.

9) Uma reportagem intitulada: NEPOTISMO, DEPUTADOS CONTRATAM 151 PARENTES foi publicada no jornal O Estado de Minas no dia 07/09/2003. A reportagem informava que deputados federais contrataram 151 parentes como funcionários de seus gabinetes ou para ocupação de cargos da mesa diretora da casa e das lideranças dos partidos. Estes empregos consomem R$ 7,8 milhões por ano em salários. Na tabela a seguir são informados os totais de parentes com respectivos valores médios dos salários por categoria de parentesco,

PARENTESCO	MÉDIA SALARIAL (em R$ × 1000)
32 esposas	3,8
47 filhos	3,2
20 irmãos	2,6
18 cunhados	2,8
12 primos	2,4
11 sobrinhos	2,2
6 noras	2,7
2 netos	3,9
2 tios	3,3
1 mãe	3,8

Nos itens a seguir considere os valores de média salarial como sendo o valor do salário para cada integrante da categoria de parentesco. Calcule e interprete o valor calculado:

a) O salário médio dos parentes.

b) O salário mediano dos parentes.

c) O desvio padrão dos salários.

10) A revista VEJA do dia 05 de fevereiro de 2003 publicou uma reportagem intitulada Globalização Fase 2-como o Brasil vai enfrentar os outros países emergentes na corrida global. Nesta reportagem estão resultados de uma pesquisa do Monitor Group, empresa de consultoria estratégica especializada em competitividade, fundada em 1983 por professores da universidade americana Harvard. Duas das variáveis pesquisadas foram o cumprimento da lei e o controle da corrupção, as quais designaremos por X e Y, respectivamente. Numa escala de notas de 0 a 100 avaliou-se o índice de confiança da sociedade na qualidade e no cumprimento das leis e no controle da corrupção. Os resultados obtidos para sete países (Brasil-BRA, Coreia do Sul-COR, México-MEX, Chile-CHI, Índia-IND, China-CHN e Rússia-RUS) estão na tabela a seguir,

	Índices dos países						
Variáveis	BRA	COR	MEX	CHI	IND	CHN	RUS
Cumpr. da lei (X)	50	81	37	85	60	58	30
Contr. da corrupção (Y)	65	68	48	82	46	47	25

a) Calcule a nota média para o cumprimento da lei.

b) Calcule a nota mediana para o controle da corrupção.

c) Qual das duas amostras é a mais homogênea? justifique.

11) A Tabela a seguir mostra o resultado de um levantamento do IBGE a respeito do tamanho das famílias em certa região do Brasil. Para famílias de tamanho 7 ou mais utilize tamanho igual a 7 nos cálculos. Famílias de tamanho igual a 2 significa somente marido e mulher.

Tamanho	N^o de famílias
2	20300
3	12000
4	11000
5	6300
6	3000
7 ou +	2400

a) Calcule o tamanho médio das famílias.

b) Calcule o tamanho mediano das famílias.

c) Calcule o desvio padrão do tamanho das famílias.

12) A tabela a seguir apresenta parte do quadro final de medalhas dos jogos olímpicos de Atenas 2004. É apresentada a colocação final (posição) do país na competição, o número de medalhas de ouro, prata e bronze e o total de medalhas, para uma amostra dos países participantes.

Posição	País	Ouro	Prata	Bronze	Total
1	Estados Unidos	35	39	29	103
2	China	32	17	14	63
3	Federação Russa	27	27	38	92
4	Austrália	17	16	16	49
5	Japão	16	9	12	37
15	Grécia	6	6	4	16
18	Brasil	4	3	3	10
20	Espanha	3	11	5	19
28	Etiópia	2	3	2	7
38	Argentina	2	0	4	6
39	Chile	2	0	1	3
60	México	0	3	1	4
61	Portugal	0	2	1	3
66	Paraguai	0	1	0	1
69	Venezuela	0	0	2	2
71	Colômbia	0	0	1	1

a) Calcule o número mediano e o número modal de medalhas de ouro. Explique ou interprete os valores calculados.

b) Calcule o número médio de medalhas de ouro. A média é uma boa medida de posição para resumir os dados apresentados, SIM ou NÃO? justifique.

c) Calcule o erro-padrão da média do item b.

d) Qual das duas amostras é a mais homogênea, a do número total de medalhas ou a do número de medalhas de ouro? Justifique sua resposta.

13) O Brasil possui a sexta maior reserva geológica de urânio do mundo. O processo de coletar o urânio natural, contendo 0,7% de urânio-235, 99,3% de urânio-238 e traços de urânio-235, e retirar uma quantidade de 238 para aumentar a concentração de 235, é conhecido como enriquecimento. O enriquecimento do urânio brasileiro é feito no exterior. A Tabela a seguir informa os custos de geração por usina (US$ por megawatt) de algumas fontes de energia.

Fonte de energia	US$ por megawatt	Fonte de energia	US$ por megawatt
Hidrelétrica	30	Petróleo	57,4
A gás	39,7	Eólica em terra	66,2
Nuclear	40,4	Eólica em alto mar	99,1
A carvão	49	Tecn. de ondas e marés	119,1
Biomassa (bagaço de cana)	49	Solar	140

a) Calcule e interprete: A amplitude total dos custos e o custo mediano.

b) Calcule o desvio padrão dos custos.

c) Explique o que é uma **análise estatística descritiva** ou um **estudo descritivo** de um conjunto de dados?

14) Uma empresa avaliou 30 lotes de peças da indústria A e também 30 lotes da indústria B. O número de peças defeituosas por lote é apresentado na tabela a seguir.

Número de	Indústria A						Indústria B					
lotes	9	9	5	4	2	1	18	6	3	3	0	0
peças defeituosas	0	1	2	3	4	5	0	1	2	3	4	5

Calcule para as duas amostras (indústrias A e B):

a) O número médio de peças defeituosas por lote.

b) O desvio padrão do número de peças defeituosas por lote.

c) O número modal de peças defeituosas por lote.

d) Qual das duas amostras é a mais homogênea? Justifique sua resposta.

15) A copa do mundo de 2006 na Alemanha foi a 18ª edição da competição, sendo o Brasil o único país que participou de todas as edições. No quadro abaixo estão os nomes (conforme são popularmente conhecidos) dos 10 maiores artilheiros da nossa seleção com o respectivo número de gols marcados em copas do mundo, incluindo-se a de 2006.
(Fonte: http://200.159.15.35/brasilnacopa/index.aspx).

Nome do artilheiro	Gols
Ronaldo	15
Pelé	12
Ademir Menezes, Jairzinho, Rivaldo e Vavá	9
Leônidas da Silva	8
Bebeto e Careca	7
Rivelino	6

Pede-se, calcule e interprete o valor calculado:

a) O número médio de gols.

b) O número mediano de gols.

c) A amplitude total.

16) Faça as devidas associações. Alguns conceitos são do capítulo de Regressão Linear Simples.

A	Coeficiente de correlação	**G**	Média geográfica
B	Coeficiente de determinação	**H**	Extrapolação
C	Valor mediano	**I**	Estatística
D	Média harmônica	**J**	Estatística Descritiva
E	Desvio da regressão	**K**	Variância amostral
F	Regressão linear simples	**L**	Estatística Inferencial

a) ( ) Uma medida do grau de associação linear entre duas variáveis aleatórias.

b) ( ) Somente resumir, descrever e apresentar, sem inferir.

c) ( ) Percentual ou proporção da variabilidade observada sendo explicada pelo modelo ajustado.

d) ( ) Métodos científicos para planejar coleta, coletar, organizar, resumir, apresentar e analisar dados. Também inclui princípios e definições para validar resultados das análises e permitir conclusões válidas. É uma mistura de ciência, tecnologia e arte.

e) ( ) Utilizar a equação ou modelo ajustado para prever valores fora do intervalo investigado ou amostrado.

f) ( ) Mede a dispersão dos valores em torno da média aritmética.

g) ( ) É a diferença entre o valor observado e o estimado.

h) ( ) Estimar valores de uma variável dependente com base nos valores de uma variável independente.

i) ( ) Uma medida de posição adequada para valores tais como velocidades e custos.

j) ( ) O ponto de equilíbrio de uma amostra de valores que se apresentam em uma progressão geométrica.

k) ( ) Valor para o qual, pelo menos, metade dos valores são maiores ou iguais e, também, pelo menos metade dos valores são menores ou iguais.

l) ( ) Medir a dispersão dos pontos ajustados.

m) ( ) É a diferença entre o maior e o menor valor da regressão.

17) Quando duas variáveis linearmente relacionadas, X e Y, são avaliadas em n unidades experimentais, obtém-se os pares de valores (x_i, y_i) para $i = 1, 2, \ldots, n$ que possibilitam o cálculo do coeficiente de correlação linear, designado por r_{XY}. Se os valores da variável Y são multiplicados por uma constante k, positiva e finita, obtendo-se $Z = kY$, o coeficiente de correlação é designado r_{XZ}. Pede-se: Utilize as propriedades de somatório para verificar como o valor de r_{XZ} se compara ao valor r_{XY}.

18) Os registros médios mensais das temperaturas ($^\circ C$) mínimas (X_1) e máximas (X_2) de 7 meses, janeiro a julho, são apresentados na tabela a seguir.

Mês (i)	Jan($_1$)	Fev($_2$)	Mar($_3$)	Abr($_4$)	Mai($_5$)	Jun($_6$)	Jul($_7$)
T. Min. (X_{1i})	21	19	17	15	10	8	6
T. Max. (X_{2i})	36	37	35	29	26	25	24

Denomina-se amplitude térmica a diferença entre as temperaturas máximas e mínimas ($X_2 - X_1$). Pede-se: apresente os cálculos que justifiquem suas respostas,

a) Amplitude térmica mediana.

b) Amplitude total das amplitudes térmicas.

c) Amplitude térmica modal.

d) Se $\sum X_{1i}^2 = 1516$ e $\sum X_{2i}^2 = 6608$, qual é a amostra mais homogênea?

19) Em uma reportagem intitulada *stand-by eleva silenciosamente a conta de luz*, publicada no jornal O Globo em 26/08/2007, era informado que aparelhos em modo de espera podem representar 20% do consumo em uma residência. Na tabela a seguir são apresentados os consumos máximos dos equipamentos (X, potência em watts) em modo stand-by por 24 horas/dia, durante 30 dias e o respectivo gasto (Y, em reais com o valor do imposto incluído). **Atenção:** se utilizar os resultados diretamente da calculadora, indique a fórmula de cálculo.

Equipamento	$X(W)$	Y (R$)	Equipamento	$X(W)$	Y (R$)
TV normal	13	4,68	Videocassete	8	2,88
Som 3 em 1 completo	18	6,48	Recarreg. de bateria	4	1,44
Computador	4	1,44	Aparelho de fax	30	10,80
CD player	6	2,16	Home theater	12	4,32
Máquina de lavar	5	1,80	Decodif. TV a cabo	14	5,04
Decodif. parabólica	20	7,20	Modem de internet	20	7,15

a) Calcule a amplitude total dos gastos (em reais).

b) Calcule o consumo mediano (em watts).

c) Calcule o consumo médio (em watts).

d) Calcule o desvio padrão dos consumos (em watts).

e) Considere que HS=(W+2) / 5. Por exemplo, um equipamento com 5W equivale a 1,4HS de potência, e suponha que a potência de cada aparelho fosse expressa em HS. Pede-se: qual amostra seria a mais homogênea, a de valores em W ou em HS? justifique. Mostre como $\bar{HS}$ e S^2_{HS} se relacionam com $\bar{W}$ e S^2_W.

20) Considere que: consumo médio=distância total percorrida / total de combustível gasto. Kelly Quina vai e volta de carro de Santos a Bertioga em busca de seu cachorrinho. Seu carro faz 16 quilômetros por litro de gasolina na viagem de ida e 12 quilômetros por litro na viagem de volta. Se a distância de Santos a Bertioga é de 60 km, pede-se: calcule o consumo médio do trajeto total (ida e volta) e mostre que a média harmônica é a média correta a ser calculada. Calcule também a média aritmética para comparar.

21) A tabela a seguir apresenta o número de registros de pessoas doentes (n_i) nos meses de janeiro a maio, com os respectivos índices de aumento ($I_i = n_i \ / \ n_{i-1}$). Suponha que seja um surto epidêmico e que todas as condições permaneçam inalteradas. Pede-se: utilize o índice médio geométrico para prever o número de doentes no mês de junho.

Mês(i)	Jan(1)	Fev(2)	Mar(3)	Abr(4)	Mai(5)
número de doentes (n_i)	12	16	26	46	90
índice de aumento (I_i)	-	1,33	1,63	1,77	1,96

Respostas dos exercícios propostos

1)a) $\bar{X} = 5$ minutos, o tempo total dividido igualmente entre os 8 telefonemas 1)b) $Md = 3,5$ minutos, sendo 4 com duração acima e também 4 abaixo 1)c) $Mo = 1$ minuto, o valor mais frequente 1)d) S($\bar{X}$)$\approx 1,67$ minutos é uma estimativa do desvio padrão da distribuição amostral da média, uma medida de precisão da estimativa 1)e) CV(%)$\approx 94,4\%$ minutos é o valor do desvio padrão expresso em termos percentuais do valor da média.

2)a) V-pode ser demostrado pela desigualdade de Jensen 2)b) V 2)c) F, pelo menos $n/2 \geq$ e também $n/2 \leq$ ao valor mediano 2)d) V 2)e) V 2)f) F, $CV_A \approx 45,3\%$ e $CV_B \approx 4,4\%$, portanto a amostra B é mais homogênea.

3) $\bar{X}_H \approx 1,57$ $\bar{X}_G \approx 1,80$ $\bar{X} = 2,1$

4) Mo=2 Md=2,5 $\bar{X} \approx 2,82$ $S_X^2 \approx 2,513$ $S_X \approx 1,59$ $S\left(\bar{X}\right) \approx 0,27$ CV(%)=56,15 AT=5; com as devidas interpretações a cargo do leitor !!

5)a) $\bar{X} \approx 3,33$ 5)b) $S\left(\bar{X}\right) \approx 0,566$ 5)c) $Md_X = 2,5$ e $Mo_X = 1$ 5)d) Não. Note que 12 países (50%) aparecem no máximo 2 vezes entre os 5 primeiros (pouca representatividade da média); e também que 3 países (Ale, Bra e Ita) contribuem 29 vezes (36%) para o total 80. A média sem estes 3 é igual a $\approx 2,43$

6) $CV_1 \approx 14,08\%$ e $CV_2 \approx 16,43\%$, portanto a amostra 1 é a mais homogênea. O coeficiente de variação (CV) é o valor do desvio padrão expresso em percentual do valor da média.

7) **Medidas de posição:** $Mo = 65$ e 91 são as notas modais-as notas mais frequentes. $Md = 68$ é a nota mediana-antes de examinar a amostra pode-se afirmar que há pelo menos 9 alunos com nota maior ou igual a 68 e também pelo menos 9 alunos com nota menor ou igual a 68. No exemplo, após examinar a amostra verifica-se que há exatamente 10 alunos. $\bar{X} = 74$ é a nota média-o valor do total de pontos distribuídos igualmente entre os 18 alunos.
Medidas de dispersão: $AT = 41$ pontos é amplitude total das notas-diferença entre a maior e a menor nota. $S^2 \approx 189,88$ pontos ao quadrado é a variância das notas-soma dos quadrados dos desvios em relação à média dividida por 17, que é o número de graus de liberdade. $S \approx 13,78$ pontos é o desvio padrão, a raiz quadrada positiva da variância, é um desvio $(x_i - \bar{X})$ representativo da amostra. $CV \approx 18,62\%$ é o coeficiente de variação-veja resposta 6) $S\left(\bar{X}\right) \approx 3,25$ é o erro-padrão da média- estimativa do desvio padrão da distribuição amostral da média, *uma medida da precisão da estimativa da média.*

8)a) V 8)b) F, $n = 100$ 8)c) F, $S_Y^2 = 125,44$ 8)d) V 8)e) 8)f) F, pelo menos $n/2 = 50$ em cada amostra, valores não necessariamente iguais.

9)a) $\bar{X} = \frac{461,8}{151} \approx 3,06$ ou R$ 3060,00 é o total dos gastos com salários dividido igualmente entre os parentes 9)b) $Md_X = X_{(76)} = 3,2$ ou R$ 3200. Pelo menos 50% do total de parentes recebem este valor ou mais (84 parentes) e pelo menos 50% recebem este valor

ou menos (114 parentes). 9)c) $S_X^2 \approx 0,517$ ou R$ 517 é um desvio representativo da dispersão dos dados em torno do valor do salário médio.

10)a) $\bar{X} \approx 57,29$ 10)b) $Md_Y = Y_{(4)} = 48$ 10)c) $CV_X \approx 35,97\%$ e $CV_Y \approx 34,29\%$ portanto a amostra de valores Y é mais homogênea.

11)a) $\bar{X} = \frac{186.900}{55000} \approx 3,4$ 11)b) $Md_X = 3$, note que do total de 55 mil famílias avaliadas, há 32300 com $x \leq 3$ e 34700 com $x \geq 3$ 11)c) $S_X \approx 1,435$.

12)a) $Mo = 0$ medalhas, valor mais frequente, para 5 países, $Md =$ 2,5. Verifica-se que há 8 países com $\geq$ 2,5 e também 8 com $\leq$ 2,5 medalhas. 12)b) $\bar{X} = 9,125$ medalhas por país. Não, pois apenas 3 países (EUA, China e Rússia) são responsáveis por 64% (94 medalhas) do total, portanto é influenciada por valores altos da amostra. 12)c) $S\left(\bar{X}\right) = 3,07$ medalhas. 12)d) a do número total por apresentar menor CV. $CV_{\text{ouro}} \approx 134,6\%$ e $CV_{\text{total}} \approx 128,8.\%$.

13)a) $AT = 110$ US$ por megawatt, $Md = 53,2$ US$ por megawatt 13)b) $S \approx 37,41$ US$ por megawatt 13)c) é um estudo no qual se procura apenas resumir os dados por meio de tabelas e/ou gráficos e/ou medidas descritivas de posição e dispersão. Isto é, nenhum método inferencial é aplicado.

14)a) $\bar{X}_A \approx 1,467$ e $\bar{X}_B = 0,70$ 14)b) $S_A \approx 1,408$ e $S_B \approx 1,022$ 14)c) $Mo_A = 0$ e 1 (bimodal) e $Mo_B = 0$ 14)d) indústria A, menor CV, $CV_A \approx 95,98\%$ e $CV_B = 146\%$

15)a) $\bar{X} = 9,1$ gols por artilheiro é o total de gols dividido igualmente entre eles 15)b) $Md = 9$ gols, com 6 artilheiros $\geq$ e 8 $\leq$ deste valor 15)c) $AT = 15 - 6 = 9$ gols é a diferença entre o maior e o menor número de gols por artilheiro.

16) (A),(J),(B),(I),(H),(K),(E),(F),(D),( ),(C),( ),( ), é a sequência de cima pra baixo.

17) $SPD_{XZ} = kSPD_{XY}$ e $SQD_Z = K^2 SQD_Y$, portanto $r_{XZ} = r_{XY}$.

18)a) 17 18)b) 4 18)c) 18 18)d) temp. máximas ($CV_1 \approx 42,04\%$ $CV_2 \approx 18,45\%$).

19)a) R$ 9,36 19)b) 12,5W 19)c) $\approx$ 12, 83W 19)d) $\approx$ 8,055 W 19)e) $\bar{HS} = (\bar{W} + 2)/5$ e $S_{HS}^2 = (1/25)S_W^2$, portanto $CV_W \approx 62,8\%$ e $CV_{HS} \approx 54,3\%$, valores HS mais homogênea.

20) distância total percorrida=$D = 2 \times 60$ e consumo total=$C = C_1 + C_2 = 60/16 + 60/12$, então verifica-se que $D/C \approx 13,7$ é a média harmônica dos consumos. A média aritmética é 14km/l.

21) $\bar{I}_G = \sqrt[4]{\prod_{i=1}^4 I_i} = \sqrt[4]{1,33 \times 1,63 \times 1,77 \times 1,96} \approx 1,656$, portanto $n_6 = \bar{I}_G \times n_5 \approx 149,04$.

Capítulo 3

Introdução à teoria da probabilidade

3.1 Introdução

Probabilidade pode ser conceituada como o ramo da Matemática que desenvolve modelos para se estudar fenômenos aleatórios, sendo que a complexidade matemática do modelo depende do fenômeno estudado. A teoria matemática da probabilidade dá-nos o instrumental para construção e análise de modelos matemáticos relativos a fenômenos aleatórios. Ao estudarmos um fenômeno aleatório, temos diante de nós um experimento cujo resultado não pode ser previsto. Ocorrem-nos então logo à mente, experimentos relacionados com jogos de azar. De fato, o estudo das probabilidades, teve suas origens no século XVII por meio dos estudos dos jogos de azar propostos pelo Cavalheiro de Mére aos matemáticos franceses Fermat e Pascal. No entanto, somente no século XX é que se desenvolveu uma teoria matemática rigorosa baseada em axiomas, definições e teoremas.

A história da teoria das probabilidades mostra uma interação estimulante entre a teoria e a prática; o progresso da teoria abre novos campos de aplicação e, por sua vez, as aplicações conduzem a novos problemas e a pesquisas produtivas. A teoria das probabilidades é, hoje em dia, aplicada num grande número de áreas distintas e a flexibilidade de uma teoria geral se faz necessária para atender a um tão grande número de solicitações diversas. Entretanto, a obtenção de valores numéricos de probabilidades não é o principal objetivo da teoria, e sim a descoberta de leis gerais e a construção de modelos teóricos satisfatórios. Como exemplos reais da importância das probabilidades no nosso cotidiano, podemos citar as previsões:

- climáticas quanto ao volume das precipitações pluviométricas e quanto à temperaturas máximas e mínimas;
- no esporte em geral, por exemplo, as *chances* de rebaixamento, do título, de vaga na taça libertadores, para os times de futebol do campeonato brasileiro;
- na economia, de inflação, da taxa de juros;
- financeiras, de preços de ações, da cotação de moeda estrangeira;
- eleitorais, pesquisas de intenção de voto;
- quanto à eficácia de vacinas e de medicamentos em geral.

Com o advento da teoria das probabilidades, foi possível estabelecer as distribuições de probabilidade, consideradas hoje a espinha dorsal da teoria estatística, pois todos os processos inferenciais são aplicações de distribuições de probabilidade. Para a ciência Estatística o significado prático de um valor de probabilidade é importante para um adequado entendimento de como as inferências são realizadas. Assim, o conhecimento dos conceitos advindos da teoria das probabilidades é de grande importância para uma correta utilização dos métodos estatísticos.

Como avaliar probabilidades

Para avaliar probabilidades há essencialmente duas alternativas,

1) métodos objetivos: utilizam conceitos, regras e modelos da teoria das probabilidades.

2) métodos subjetivos: utilizam a experiência, dados históricos, bom senso e etc.

Os métodos subjetivos são muitas das vezes a única alternativa para se avaliar probabilidades em situações reais. Eles obviamente fornecem valores (ou estimativas) válidos desde que a metodologia empregada seja coerente com os axiomas fundamentais (abordados na seção 3.3) para o cálculo das probabilidades.

Geralmente as intuições com relação à aleatoriedade de fenômenos são frequentemente errôneas ou mal interpretadas. Portanto, quando se avalia probabilidades deve-se utilizar teoremas e regras em contraposição aos cálculos baseados na intuição. Considere o seguinte exemplo para ilustrar a natureza intuitiva das interpretações: um paciente é informado pelo médico que possui uma doença grave e que a *chance* de vir a óbito nos próximos 6 meses é de 80%.

- O paciente interpreta como: *o médico me disse que vou morrer.*
- O médico interpreta como: segundo a literatura especializada, quando 100 pacientes como este são submetidos ao tratamento adequado, espera-se que apenas 20 sobrevivam.

Neste exemplo uma questão técnica é o conceito de *chance* empregado como uma probabilidade percentual, o que a rigor está incorreto. Se p é a probabilidade de um evento, então $\frac{p}{(1-p)}$ é definido como sendo a *chance* do evento. Por exemplo, se $p = 0,40$ ou 40% então $\frac{0,40}{1-0,40} = \frac{1}{1,5} \Rightarrow$ *a chance do evento é de* 1 *em* $1,5$ ou *a probabilidade do evento não ocorrer é* $1,5$ *vezes a probabilidade dele ocorrer.*

Pode-se afirmar então que as incertezas sobre qualquer quantidade ou estado desconhecidos da natureza, somente podem ser descritas probabilisticamente.

Para uma adequada compreensão dos princípios básicos da teoria das probabilidades, que é o nosso objetivo inicial, iniciaremos com alguns conceitos.

3.2 Conceitos fundamentais

3.2.1 Modelo determinístico

É aquele modelo em que, a partir das condições em que o experimento é realizado, pode-se determinar seu resultado. Sabe-se, por exemplo, que a expressão $e = -4,9t^2 + v_0t$ representa

a distância vertical percorrida por um objeto acima do solo, sendo v_0 a velocidade inicial e t o tempo gasto na queda. Portanto, conhecidos os valores de v_0 e t, o valor de e fica implicitamente determinado. É importante observar que existe uma relação definida entre t e e que determina unicamente a quantidade no primeiro membro da equação, se aquelas do segundo membro forem fornecidas.

3.2.2 Modelo probabilístico

É aquele modelo em que as condições de execução de um experimento não determinam o resultado final, mas sim o comportamento probabilístico do resultado observável.

Considere, por exemplo, a seguinte situação: deseja-se determinar qual a precipitação pluviométrica que ocorrerá numa determinada localidade como resultado de uma tempestade que se avizinha. Dispõe-se de informações sobre pressão barométrica em vários pontos, variação de pressão, velocidade do vento, etc. Embora sejam essas informações valiosas, não são capazes de responder a questão levantada, qual seja, a de qual será o volume da precipitação pluviométrica. Como se pode notar, este fenômeno não se coaduna com um tratamento determinístico; um modelo probabilístico se adapta à situação com mais propriedade.

3.2.3 Experimentos probabilísticos ou aleatórios

São aqueles experimentos cujos resultados podem não ser os mesmos, ainda que sejam repetidos sob condições essencialmente idênticas. São exemplos de experimentos probabilísticos:

i) E_1: Lançar uma moeda 10 vezes e observar o número de caras obtidas.

ii) E_2: Escolher, ao acaso, um ponto de um círculo de raio unitário, centrado em $(0, 0)$.

iii) E_3: Selecionar uma carta de um baralho com 52 cartas e observar seu "naipe".

iv) E_4: Lançar um dado e observar o número da sua face superior.

v) E_5: Amostrar n peças em um lote e verificar o número de defeituosas, designado como X.

vi) E_6: Registrar com o auxílio de um aparelho apropriado, o número de partículas radioativas emitidas por uma fonte em um período de 24 horas, designado como Y.

vii) E_7: Realizar um teste de vida útil com n componentes eletrônicos para registrar os tempos $t_i,\ i = 1, \ldots, n$ de operação contínua até a ocorrência da primeira falha.

3.2.4 Espaço amostral

Chama-se espaço amostral o conjunto de todos os possíveis resultados de um experimento aleatório ou, em outras palavras, é o conjunto universo relativo aos resultados de um experimento. Esse conjunto será representado pela letra S. Assim, pode-se dizer que, a cada experimento aleatório sempre estará associado um conjunto de resultados possíveis ou espaço amostral.

Aos experimentos aleatórios exemplificados anteriormente estão associados os seguintes espaços amostrais, respectivamente:

(*i*) $S_1 = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

(*ii*) $S_2 = \{(x, y) : x^2 + y^2 \leq 1\}$ é infinito e não enumerável.

(*iii*) $S_3 = \{ouro, paus, copas, espadas\} = \{\diamondsuit, \clubsuit, \heartsuit, \spadesuit\}$.

(*iv*) $S_4 = \{1, 2, 3, 4, 5, 6\}$.

(*v*) $S_5 = \{0, 1, 2, \ldots, n\}$ é finito e enumerável.

(*vi*) $S_6 = \{0, 1, 2, \ldots\}$ é infinito e enumerável.

(*vii*) $S_7 = \{t_i : 0 < t_i \leq t_{max}\}$ é finito porém não enumerável. O valor t_{max} poderá ser conhecido ou desconhecido, sendo que estimar t_{max} poderá ser um dos objetivos do experimento.

O espaço amostral não é único, já que em um mesmo experimento aleatório diversas características ou tipos de resultados podem ser registrados. No experimento aleatório *vii* o objetivo poderia ser verificar o número de componentes que não irão falhar antes de t_0 horas de funcionamento, portanto teríamos $S_{vii} = S_v$.

3.2.5 Eventos

Denomina-se evento a todo conjunto particular de resultados de S, ou ainda, a todo subconjunto de S. Será útil considerarmos o espaço amostral S e o conjunto vazio $\emptyset$ como eventos. O primeiro é denominado evento certo e o segundo, evento impossível.

Geralmente eventos são designados pelas primeiras letras do alfabeto: A, B, C, $D, \ldots, H$. Portanto, para os exemplos anteriores:

(*v*) seja $A =$ { nenhuma peça defeituosa foi amostrada } ou $A = \{X = 0\}$, sendo X o número de peças defeituosas amostradas;

(*vi*) seja $B =$ { no máximo 10 partículas } ou $B = \{Y \leq 10\}$, sendo Y o número de partículas emitidas;

(*vii*) seja $C = \{t_{(1)} > 300$ horas $\}$, em que $t_{(1)}$ é o mínimo valor t_i.

Observe que nos exemplos anteriores, em *v* e *vi*, procura-se motivar a ideia de uma variável (X e Y), um conceito importante em conexão com modelos para o cálculo das probabilidades, neste caso X e Y são variáveis aleatórias. Ou seja, avaliar a probabilidade do evento A em S, equivale a avaliar a probabilidade do evento $\{X = 0\}$ em S_X, o espaço amostral de X, isto é, $P_S(A) = P_{S_X}(X = 0)$. O conceito formal de P como uma função de probabilidade será apresentado a seguir. O estudo das variáveis aleatórias e dos respectivos modelos de distribuição de probabilidades, são apresentados em outros capítulos deste texto.

Em particular, se S é um espaço amostral discreto enumerável composto de n pontos amostrais, existem 2^n subconjuntos ou eventos que podem ser formados a partir de S (veja

adiante na subseção 3.3.1.3). O conjunto que reúne todos esses subconjuntos é chamado de espaço de eventos ou classe de eventos.

Exemplo: Seja $S = \{1, 2, 3\}$, temos então $n = 3 \Rightarrow 2^3 = 8$ eventos.

$$\begin{array}{cccc} A_1 = \emptyset, & A_2 = \{1\}, & A_3 = \{2\}, & A_4 = \{3\}, \\ A_5 = \{1, 2\}, & A_6 = \{1, 3\}, & A_7 = \{2, 3\}, & A_8 = \{1, 2, 3\}. \end{array}$$

3.2.5.1 Eventos mutuamente exclusivos

Dois eventos são mutuamente exclusivos (ou mutuamente excludentes) se, e somente se, a ocorrência de um impede a ocorrência do outro. Correspondentemente, caracterizam-se, na teoria dos conjuntos, por dois conjuntos disjuntos, isto é, que não possuem nenhum ponto em comum. Como exemplo, considere-se os seguintes casos:

(i) No lançamento de um dado, a ocorrência de uma face elimina a possibilidade de ocorrência das outras cinco.

(ii) Seja S o espaço amostral referente a retirada de uma carta de um baralho de 52 cartas. Seja A o evento "retirada de um ás" e B o evento "retirada de uma carta de ouro". Vê-se que a possibilidade de ocorrer A e B ao mesmo tempo não está descartada, ou seja, ocorrer ás de ouro. Logo, os eventos A e B não são mutuamente exclusivos.

Em outras palavras, dois eventos A e B são mutuamente exclusivos se o seu conjunto interseção for vazio, ou seja, $A \cap B = \emptyset \Leftrightarrow A$ e B são disjuntos.

3.2.5.2 Operações básicas entre subconjuntos ou eventos

Quando se avalia probabilidades deve-se utilizar conceitos, teoremas e regras em contraposição aos cálculos baseados na intuição, pois geralmente as intuições com relação à aleatoriedade de fenômenos são frequentemente errôneas ou mal interpretadas. Portanto, existem regras e relações básicas que são válidas para todas as situações. A seguir apresentamos as mais fundamentais. Sejam A, B, e C subconjuntos ou eventos de S. Inicialmente, duas relações básicas:

(i) Está contido ou contém,

$$A \subset B \text{ ou } B \supset A \iff w \in A \Rightarrow w \in B.$$

(ii) Igualdade,

$$A = B \iff A \subset B \text{ e } B \subset A.$$

As 3 operações básicas são:

(iii) União,

$$A \cup B = \{w : w \in A \text{ ou } w \in B \}.$$

Portanto é o evento somente A ou somente B ou ambos.

(iii*) União para uma sequência finita de eventos $\{A_i\}_{i=1}^n$ ou infinita,

$$A_1 \cup A_2 \cup \ldots = \bigcup_{i=1}^{+\infty} A_i \;=\; \{w:\; w \in A_i \;\text{ para pelo menos um valor } i\;\}.$$

(iv) Interseção,

$$A \cap B \;\;=\;\; \{w:\; w \in A \;\text{ e }\; w \in B\;\}.$$

Portanto são os eventos A e B simultaneamente.

(iv*) Interseção para uma sequência finita de eventos $\{A_i\}_{i=1}^n$ ou infinita,

$$A_1 \cap A_2 \cap \ldots = \bigcap_{i=1}^{+\infty} A_i \;=\; \{w:\; w \in A_i \;\text{ para todo } i\;\}.$$

(v) Complementação,

$$A^c = \bar{A} = \{w:\; w \notin A\}.$$

A^c é o evento complementar de A ou é o evento *não* A,

$$\text{não ocorre } A \iff \text{ ocorre } A^c.$$

Para finalizar uma definição:

(vi) Partição. Os eventos $\{A_i\}_{i=1}^n$ ou $\{A_i\}_{i=1}^\infty$ formam uma partição de S se:

$$A_i \cap A_j \;=\; \emptyset \;\;\forall\;\; i \neq j \;\text{ e também }\; \cup_i A_i = S$$

3.2.5.3 Propriedades das operações básicas

As propriedades a seguir são úteis tanto na demonstração de resultados (como por exemplo as desigualdades de Boole e Bonferoni, não abordadas neste texto) quanto na resolução de exercícios de probabilidades. Provar as propriedades significa demonstrar a igualdade das expressões por argumentos matemáticos lógicos. Entretanto, pode-se visualizar ou ilustrar as propriedades facilmente com o auxílio de um diagrama de Venn.

a) Comutativa: $A \cup B = B \cup A$ e $A \cap B = B \cap A$.

b) Associativa: $(A \cup B) \cup C = A \cup (B \cup C)$ e $(A \cap B) \cap C = A \cap (B \cap C)$.

c) Distributiva: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ e $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

d) Leis de DeMorgan:

d).i $(A \cup B)^c = A^c \cap B^c$ ou $A \cup B = (A^c \cap B^c)^c$ ou ainda $\bigcup_{i=1}^n A_i = \left(\bigcap_{i=1}^n A_i^c\right)^c$.

d).ii $(A \cap B)^c = A^c \cup B^c$ ou $A \cap B = (A^c \cup B^c)^c$ ou ainda $\bigcap_{i=1}^n A_i = \left(\bigcup_{i=1}^n A_i^c\right)^c$

e) Diferença entre A e B: $A - B = A \cap B^c = AB^c$

f) Diferença simétrica entre A e B: $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$

3.3 Conceitos de probabilidade

Como a teoria das probabilidades está, historicamente, ligada aos jogos de azar, esta associação gerou inicialmente um conceito chamado conceito clássico ou probabilidade "a priori", devido a Laplace. O conceito de frequência relativa como estimativa de probabilidade ou probabilidade "a posteriori" surgiu posteriormente através de Richard Von Mises. Já no século XX, como a conceituação até então existente não era apropriada a um tratamento matemático mais rigoroso, A. N. Kolmogorov conceituou probabilidade com axiomas[1] rigorosos.

3.3.1 Conceito clássico ou probabilidade *a priori*

Seja E um experimento aleatório e S um espaço amostral enumerável finito e equiprovável, a ele associado, composto de n pontos amostrais. Define-se a probabilidade de um evento $A \in S$, indicada por $P(A)$, como sendo a relação entre o número de pontos amostrais pertencentes ao evento A (f) e o número total de pontos (n):

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

Podemos complementar definindo a probabilidade do evento A não ocorrer, indicada por $P(\bar{A})$, como:

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

Obviamente, $n = f + c$, em que c é o número de pontos amostrais em S, não pertencentes ao evento A. Conclui-se também que $P(A) + P\left(\bar{A}\right) = 1 \;\Rightarrow\; P\left(\bar{A}\right) = 1 - P(A)$.

Considere os seguintes exemplos:

(i) Seja E o experimento relativo ao lançamento de um dado perfeitamente simétrico. Seja A o evento ocorrência de um número par. Considerando-se que os pontos de $S = \{1, 2, 3, 4, 5, 6\}$ são equiprováveis, isto é, cada ponto de S tem a mesma probabilidade de ocorrer, pois o dado não é *viciado*, tem-se que:

$$P(A) = \frac{3}{6} = 0,5 \;\text{ ou }\; 50\%$$

pois S possui $n = 6$ pontos amostrais dos quais $f = 3$ pertencem ao evento $A = \{2, 4, 6\}$.

(ii) Seja o espaço amostral referente ao número de caras obtidos em 3 lances de uma moeda e A o evento ocorrência de exatamente uma cara.

Neste caso: $S = \{0, 1, 2, 3\}$ e $A = \{1\}$. Aqui, o conceito clássico não pode ser imediatamente aplicado pois, os pontos de S não são equiprováveis, ou seja $P(A) \neq \frac{1}{4}$. Para aplicar o conceito clássico, deve-se considerar o seguinte espaço amostral:

$$S = \{ca\,ca\,ca, ca\,ca\,co, ca\,co\,ca, co\,ca\,ca, ca\,co\,co, co\,ca\,co, cococa, co\,co\,co\}$$

[1] Axioma é uma verdade absoluta, ou seja, um resultado que deve ser tomado como verdadeiro, não sendo demonstrável.

vê-se que $A = \{ca\,co\,co, co\,ca\,co, co\,co\,ca\}$, logo, $P(A) = \dfrac{3}{8}$ já que, nesse caso, S é equiprovável.

Fatos:

- O conceito clássico só pode ser utilizado em situações em que o espaço amostral é enumerável, finito e equiprovável.

- Sendo $P(A) = \dfrac{f}{n}$, no caso de S possuir infinitos pontos amostrais (enumeráveis ou não), pelo conceito clássico todos os eventos teriam probabilidade zero de ocorrer.

O espaço amostral de um experimento aleatório, conforme definido e exemplificado na Seção 3.2.4, pode ser finito ou infinito (enumerável ou não), entretanto os espaços amostrais finitos e equiprováveis facilitam o cálculo de probabilidades.

3.3.1.1 Espaço amostral finito

Um espaço amostral finito S pode ser indicado como: $S = \{a_1, a_2, \cdots, a_n\}$. Neste caso, a probabilidade de cada ponto $a_i \in S$ é um número real $p_i = P(a_i)$ que satisfaz às seguintes condições:

i) $p_i \geq 0$ para $i = 1, 2, 3, \cdots, n$;

ii) $p_1 + p_2 + \cdots + p_n = \sum_{i=1}^{n} p_i = 1$.

Portanto, a probabilidade $P(A)$, de qualquer evento $A \subset S$, é dada pela soma das probabilidades dos pontos $a_i \in A$.

3.3.1.2 Espaço amostral finito e equiprovável

Seja S um espaço amostral finito. Se cada ponto de S tem a mesma probabilidade de ocorrer, então o espaço amostral chama-se equiprovável ou uniforme. Em particular, se S contém n pontos, então a probabilidade de cada ponto será $\dfrac{1}{n}$. Portanto, se um evento A contém r pontos do espaço amostral ($r \leq n$), então $P(A) = \dfrac{r}{n}$, conforme o conceito clássico,

$$P(A) = \frac{\text{número de elementos de } A}{\text{número de elementos de } S}.$$

3.3.1.3 Revisão: contagem por combinação

No conceito a priori define-se a probabilidade de um evento A qualquer como,

$$P(A) = \frac{f}{n},$$

em que f é o número de resultados do espaço amostral S pertencentes ao evento A e n é o número total de resultados (finito) equiprováveis em S. Portanto, em muitas aplicações o

valor de f pode ser calculado com o auxílio de uma fórmula útil denotada $\binom{n}{r}$ que designa a combinação de n objetos distintos em grupos de tamanho r, na qual um grupo se distingue de outro pela natureza dos objetos e não pela ordem. Por exemplo, combinar os objetos A, B, 1, 2 em grupos de tamanho 3 resulta em 4 grupos dados por: $(A, B, 1)$, $(A, B, 2)$, $(A, 1, 2)$, $(B, 1, 2)$. A fórmula da combinação é a seguinte,

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}, \quad \text{com} \quad 0! = 1 \quad \text{e} \quad n! = n \cdot (n-1) \cdot (n-2) \cdots 1$$

Pode-se demonstrar por indução matemática que,

$$\sum_{k=0}^{n} \binom{n}{k} = \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{n} = 2^n.$$

Vale mencionar também que: $\binom{n}{r} = \binom{n}{n-r}$ e que $\binom{n}{2} = \frac{n(n-1)}{2}$.

3.3.2 Frequência relativa ou probabilidade *a posteriori*

Considere um experimento aleatório que possa ser repetido indefinidamente em idênticas condições e seja A um evento pertencente ao espaço amostral associado a este experimento aleatório. Se após n repetições do experimento, com n suficientemente grande, for observado $m \leq n$ vezes o evento A, então uma estimativa da probabilidade $P(A)$ é dada pela frequência relativa $\dfrac{m}{n}$. Esta definição é, às vezes, chamada de probabilidade empírica e tem por base o princípio estatístico da estabilidade, ou seja, à medida que o número de repetições do experimento cresce, a frequência relativa $\dfrac{m}{n}$ se aproxima da probabilidade $P(A)$.

A exigência n suficientemente grande é por demais vaga para que sirva como uma boa definição de probabilidade, além de impossibilitar, tal como o conceito clássico, o tratamento probabilístico de eventos associados a experimentos aleatórios com espaços amostrais infinitos.

Exemplo hipotético 1: Um dado perfeitamente simétrico foi lançado n vezes e registrou-se o número de vezes m que determinada face ocorreu (por exemplo a face com o número 6). Os resultados deste experimento aleatório são apresentados na Tabela 3.1. A frequência relativa observada é a probabilidade *a posteriori* e a frequência relativa esperada é a probabilidade *a priori* do evento, dada por $P(A) = \frac{1}{6} \approx 0,167$. Com o aumento do número de repetições n, as duas se aproximam.

Tabela 3.1: Resultados hipotéticos do lançamento de um dado n vezes, sendo m o número de vezes que determinada face ocorre

		Frequência relativa	
n	frequência (m)	observada	esperada
20	2	0,100	0,167
40	5	0,125	0,167
60	7	0,117	0,167
80	15	0,188	0,167
100	17	0,170	0,167
150	24	0,160	0,167
200	33	0,165	0,167

Exemplo hipotético 2: Em um ensaio clínico 2000 voluntários foram vacinados contra a gripe causada pelo vírus influenza C e observou-se que 1900 deles não contraíram a gripe naquela estação. Todos os voluntários eram homens entre 40 e 70 anos de idade, saudáveis e sem nenhum histórico de alergias à vacina ou quaisquer outras condições que os tornassem mais propensos a griparem. Se um homem com 50 anos de idade for vacinado, tendo ele o mesmo perfil de saúde dos voluntários do estudo, qual é a probabilidade dele vir a contrair a gripe causada pelo vírus influenza C? Neste caso o conceito de probabilidade como frequência relativa é a única saída e resulta em,

$$P(\text{gripar}) = \frac{100}{2000} = 0,05.$$

3.3.3 Conceito intuitivo

Seja $S = \{w_1, w_2, w_3 \ldots\}$ um espaço amostral enumerável e sejam valores $p_i > 0$ tais que $\sum_i p_i = 1$. Então, $\forall \ A \subset S$ tem-se que $P(A)$ é a probabilidade do evento ou subconjunto A que satisfaz a:

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

Em particular, o espaço amostral S e o conjunto vazio $\emptyset$ satisfazem a:

$$\begin{aligned} P(S) &= 1; \\ P(\emptyset) &= \sum_{w_i \in \emptyset} p_i = 0. \end{aligned}$$

O conceito intuitivo, apesar de ser mais abrangente que os dois conceitos anteriores, ainda é uma particularidade do conceito moderno descrito a seguir.

3.3.4 Conceito moderno ou axiomático

Seja E um experimento aleatório e S um espaço amostral associado a E. A cada evento A de S associaremos um número real $P(A)$, denominado probabilidade do evento A, se forem satisfeitas as seguintes condições ou axiomas:

(i) $P(A) \geq 0$, para qualquer evento A em S;

(ii) $P(S) = 1$;

(iii) Se A e B são dois eventos de S e são mutuamente exclusivos, então

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

O axioma *(iii)* pode ser generalizado para o caso de um número finito de eventos mutuamente exclusivos, ou seja,

(iii^*)

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n) = \sum_{i=1}^{n} P(A_i).$$

Para uniões infinitas o axioma *(iii)* também é verdadeiro,

(iii^{**}) Se $A_i \bigcap A_j = \emptyset \quad \forall\ i \neq j \quad$ e $\quad A_i \subset S \ \ \forall \ \ i$ então, $P(\cup_i A_i) = \sum_i P(A_i)$.

Os axiomas resultam em diversas regras ou teoremas (demonstrados adiante no texto na Seção 3.4) para o cálculo de probabilidades. Se A e B são dois eventos em S, segue que:

(1) $P(\emptyset) = 0$;

(2) $P(A^c) = 1 - P(A)$ e $0 \leq P(A) \leq 1$;

(3) $P(A \cap B^c) = P(A) - P(A \cap B)$;

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

(5) Se $A \subset B$ então $P(A) \leq P(B)$;

(6) $P(A) = \sum\limits_i P(A \cap B_i)$, se $\{B_i\}$ formam uma partição de S (finita ou infinita).

Pelo que se pode notar, o conceito axiomático não fornece formas e sim condições para o cálculo de probabilidade, ou seja, qualquer processo de cálculo da probabilidade é válido desde que satisfaça os axiomas. Facilmente se comprova que os conceitos *a priori*, *a posteriori* e intuitivo se enquadram dentro desse conceito. Estes axiomas fundamentais para o cálculo de probabilidades são conhecidos como os axiomas de Kolmogorov [Kolmogorov, A. N. (1933)] e foram publicados originalmente em russo, sendo traduzidos para o inglês como *foundations of the theory of probability* em 1950, ano em que passaram a ser conhecidos pela comunidade científica internacional.

Conceitos básicos complementares

A teoria geral da medida é fundamental para um perfeito entendimento da teoria de probabilidade, entretanto, um tratamento matemático mais formal da probabilidade, como uma medida em um espaço, não é o objetivo do presente capítulo. Os conceitos de álgebra e de sigma álgebra de eventos, são abordados resumidamente à seguir.

Definição: sigma álgebra

Uma σ-álgebra é uma classe de eventos ou de subconjuntos de Ω (ou de S, um espaço amostral associado a um experimento aleatório), a qual denotaremos por $\mathcal{A}$, que satisfaz a:

i) $\Omega \in \mathcal{A}$ (ou $\emptyset \in \mathcal{A}$).

ii) Se $A \in \mathcal{A}$ então $A^c \in \mathcal{A}$.

iii) Se $A_i \in \mathcal{A}$, então $\bigcup_{i=1}^{+\infty} A_i \in \mathcal{A}$.

Se somente uniões finitas satisfazem ao item iii), isto é, $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$, então $\mathcal{A}$ será uma **álgebra de eventos** e não uma sigma álgebra. Outros textos denotam uma sigma álgebra por $\mathcal{B}$ ou $\mathcal{F}$ devido ao termo campo de Borel ou *Borel field* em inglês, que se refere à menor sigma álgebra que contém todos os intervalos abertos em $\mathbb{R}$. Os elementos de $\mathcal{A}$ são denominados **conjuntos mensuráveis**.

Fatos:

a) Se cada $A_i \in \mathcal{A}$ então $\bigcap_{i=1}^{+\infty} A_i \in \mathcal{A}$.

b) $\mathcal{A} = \{\emptyset, \Omega\}$ é a sigma álgebra trivial.

c) Se Ω contém n elementos então $\mathcal{A}$ contém $\sum_{k=0}^{n} \binom{n}{k} = 2^n$ elementos.

d) Se $A \subset \Omega$ então a menor sigma álgebra gerada por A será $\mathcal{A} = \{A, A^c, \emptyset, \Omega\}$.

Finalmente, o conceito de probabilidade abordado a seguir, na forma de exemplos, se faz necessário para situações particulares no plano e no espaço, nas quais tanto os eventos quanto os espaços amostrais são formados por infinitos pontos amostrais.

3.3.5 Probabilidade geométrica

Definição básica

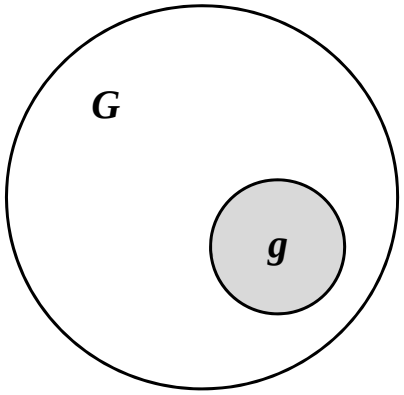
Suponhamos que um segmento l seja parte de um outro segmento L e que se tenha escolhido ao acaso um ponto de L. Se admitirmos que a probabilidade deste ponto pertencer a l é proporcional ao comprimento de l e não depende do lugar que l ocupa em L, então a probabilidade de que o ponto selecionado esteja em l será :

$$P = \frac{\text{comprimento de } l}{\text{comprimento de } L}.$$

Analogamente, suponhamos que uma figura plana g seja parte de uma outra figura plana G e que se tenha escolhido ao acaso um ponto de G. Se admitirmos que a probabilidade

deste ponto pertencer a g é proporcional à área de g e não depende do lugar que g ocupa em G, então a probabilidade de que o ponto selecionado esteja em g será:

$$P = \frac{\text{área de } g}{\text{área de } G}.$$



Exemplo 22

Sejam $\overline{AB}$ um segmento de reta e E um evento que se caracteriza pela escolha, ao acaso, de um ponto do segmento $\overline{AB}$, que pertença também ao segmento menor $\overline{ab}$, contido em $\overline{AB}$. Se os comprimentos de $\overline{AB}$ e $\overline{ab}$ são, respectivamente, 5 e 2 unidades, então a probabilidade da ocorrência de E é dada por:

$$P(E) = \frac{\text{comprimento de } \overline{ab}}{\text{comprimento de } \overline{AB}} = \frac{2}{5}.$$

Exemplo 23

Seja um experimento referente à escolha de um ponto ao acaso contido em um círculo de raio igual a r centrado na origem. Tem-se então:

$$S = \left\{(x, y) \in \mathbb{R}^2 \,:\, x^2 + y^2 \leq r^2\right\},$$

ou seja, os pares de valores (x, y) que satisfazem a condição $x^2 + y^2 \leq r^2$, são os pontos amostrais que compõe S.

Admita que o raio deste círculo seja igual a 2 e sejam os eventos,

$$A = \left\{(x, y) \in \mathbb{R}^2 \,:\, x^2 + y^2 \leq 2\right\} \quad \text{e} \quad B = \left\{(x, y) \in \mathbb{R}^2 \,:\, x^2 + y^2 \leq \frac{1}{2}\right\}.$$

Pelo conceito de probabilidade geométrica, segue que $P(A)$ e $P(B)$ são definidos como:

$$P(A) = \frac{\text{área do círculo de raio } \sqrt{2}}{\text{área do círculo de raio } 2} = \frac{\pi\left(\sqrt{2}\right)^2}{\pi(2)^2} = \frac{1}{2}$$

e,

$$P(B) = \frac{\text{área do círculo de raio } \sqrt{\frac{1}{2}}}{\text{área do círculo de raio } 2} = \frac{\pi\left(\sqrt{\frac{1}{2}}\right)^2}{\pi(2)^2} = \frac{1}{8}.$$

Como se pode observar, o conceito axiomático, por ser amplo, trouxe como principal vantagem a possibilidade da extensão do estudo aos eventos pertencentes aos espaços amostrais infinitos não enumeráveis. Intuitivamente, a probabilidade de um evento é a medida do conjunto que representa o evento e pode ser calculada por diversas formas.

3.3.5.1 Exercícios propostos com respostas

1) Tendo-se tomado, ao acaso, dois números positivos x e y, que não excedem a dois, determinar a probabilidade P de que o produto xy não exceda à unidade e o quociente $\frac{y}{x}$ não exceda a dois. (Resposta: $P \cong 0,385$).

2) Considere um círculo de raio igual a R e um quadrado de lado igual a L. Pede-se (utilize o conceito de probabilidade geométrica):

a) Se o círculo está inscrito ao quadrado, calcule a probabilidade de que um ponto escolhido ao acaso dentro do quadrado também esteja dentro do círculo. (Resposta: $\frac{\pi}{4} \approx 0,79$).

b) Se o círculo está circunscrito ao quadrado, calcule a probabilidade de que um ponto escolhido ao acaso dentro do círculo também esteja dentro do quadrado. (Resposta: $\frac{2}{\pi} \approx 0,64$).

3) Considere a escolha aleatória de dois números x e y reais e positivos tais que $0 \leq x, y \leq 2$. Qual é a probabilidade de que o produto seja menor do que 2 com x menor do que y? Isto é, $P(\{(x,y) : xy < 2,\ x < y\})$. (Resposta: $\frac{1+ln2}{4} \approx 0,423$).

4) Considere a escolha aleatória de dois números x e y reais e positivos tais que $0 \leq x, y \leq 4$. Qual é a probabilidade de que $|x-y| < 1$? (Resposta: $\frac{7}{16} = 0,4375$).

3.4 Teoremas do cálculo de probabilidades

Os teoremas enunciados a seguir são um poderoso instrumento de auxílio no cálculo de probabilidades. Os diagramas de Venn são úteis na compreensão tanto dos teoremas como dos processos de demonstração.

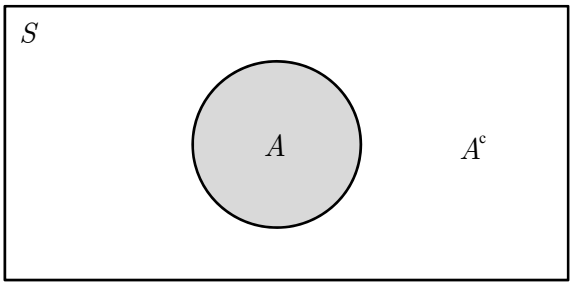
Teorema 1

Se $\emptyset$ é um conjunto vazio, então $P(\emptyset) = 0$.

Temos que $A = A \cup \emptyset$, então $P(A) = P(A \cup \emptyset) = P(A) + P(\emptyset)$ pelo axioma *(iii)* já que $A \cap \emptyset = \emptyset$, isto é, A e $\emptyset$ são mutuamente exclusivos. Portanto, $P(\emptyset) = P(A) - P(A) = 0$.
Fato: A recíproca deste teorema não é verdadeira, isto é, $P(A) = 0$ não implica que A seja um conjunto vazio.

Teorema 2

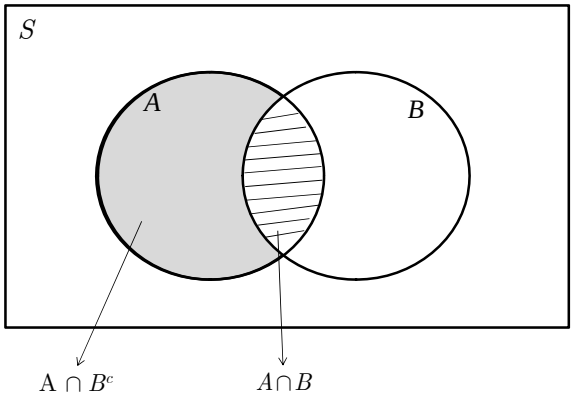
Se A^c é o complemento de A, então $P(A) = 1 - P(A^c)$.



Por definição, $A \cup A^c = S$, então $P(A \cup A^c) = P(S) = 1$, já que pelo axioma (ii) $P(S) = 1$. Mas A e A^c são mutuamente exclusivos, isto é, $A \cap A^c = \emptyset$, logo pelo axioma (iii) temos que $P(A \cup A^c) = P(A) + P(A^c) = 1 \therefore P(A) = 1 - P(A^c)$.

Teorema 3

Se A e B são dois eventos quaisquer e B^c é o complemento de B, então $P(A \cap B^c) = P(A) - P(A \cap B)$.



Pelo diagrama de Venn acima verifica-se que, $A = (A \cap B^c) \cup (A \cap B)$. Portanto,

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

pois $(A \cap B^c)$ e $(A \cap B)$ são mutuamente exclusivos, logo: $P(A \cap B^c) = P(A) - P(A \cap B)$.

Obviamente: $P(A^c \cap B) = P(B) - P(A \cap B)$.

Teorema 4 (Teorema da soma das probabilidades)

Se A e B são dois eventos quaisquer, então $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Pelo diagrama de Venn apresentado no **Teorema 3**, vê-se que: $A \cup B = B \cup (A \cap B^c)$, sendo B e $(A \cap B^c)$ mutuamente exclusivos. Assim: $P(A \cup B) = P(B) + P(A \cap B^c)$. Novamente pelo **Teorema 3**, $P(A \cap B^c) = P(A) - P(A \cap B)$, logo: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

FATO: Para três eventos quaisquer A, B e C, temos que:

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

Para verificar este resultado basta considerar que $A \cup B \cup C = (A \cup B) \cup C$, portanto,

$$P(A \cup B \cup C) = P[(A \cup B) \cup C].$$

Aplicando-se o **Teorema 4** vem:

$$\begin{aligned} P(A \cup B \cup C) &= P(A \cup B) + P(C) - P[(A \cup B) \cap C] \\ &= P(A) + P(B) - P(A \cap B) + P(C) - P[(A \cap C) \cup (B \cap C)] \\ &= P(A) + P(B) + P(C) - P(A \cap B) - \\ &\quad [P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)]. \end{aligned}$$

Assim,

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

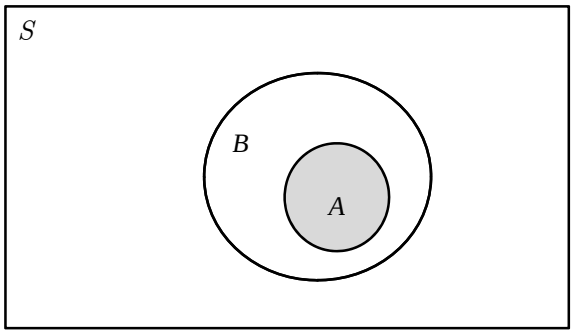
O **Teorema 4** pode ser generalizado para uma união finita ou infinita da seguinte forma,

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) \quad - \sum_{\substack{\forall\ m=2n \\ n=1,2,3,4,...}} P\left(\bigcap_{k=1}^{m} A_{i_k}\right) + \sum_{\substack{\forall\ m=2n+1 \\ n=1,2,3,4,...}} P\left(\bigcap_{k=1}^{m} A_{i_k}\right),$$

em que $\{A_{i_k}\}_{k=1}^{m}$ é qualquer subsequência de m dos eventos A_i da sequência finita de eventos $\{A_i\}_{i=1}^{n}$ ou infinita $\{A_i\}_{i=1}^{\infty}$. Observe que quando $m = 2n$ a interseção envolve um número par de eventos A_i e quando $m = 2n + 1$ um número ímpar.

Teorema 5

Se $A \subset B$, então $P(A) \leq P(B)$.



Pelo diagrama podemos escrever que: $B = A \cup (A^c \cap B)$. Segue que $P(B) = P(A) + P(A^c \cap B)$, pois A e $(A^c \cap B)$ são mutuamente exclusivos. Pelo axioma (i), $P(A^c \cap B) \geq 0 \quad \therefore P(A) \leq P(B)$.

FATO: Os axiomas (i) e (ii) juntamente com o **Teorema 5** resultam que para um evento A qualquer, $0 \leq P(A) \leq 1$.

3.5 Exemplos resolvidos

Exemplo 24

Um algarismo é escolhido dentre os algarismos $1, 2, 3, 4, 5$ e em seguida uma segunda seleção é feita entre os quatro algarismos restantes. Admita que os vinte resultados possíveis tenham a mesma probabilidade. Determine a probabilidade de que um algarismo ímpar seja escolhido:

a) Na primeira vez;

b) Na segunda vez;

c) Ambas as vezes;

d) Se X_i é o algarismo obtido na i-ésima seleção, calcular $P(2X_1 + X_2 \geq 8)$.

A solução deste exemplo será por meio do conceito clássico, já que até este ponto no texto o conceito de probabilidade condicional ainda não foi abordado. Inicialmente, vejamos o espaço amostral deste experimento aleatório (facilmente obtido por um diagrama em árvore):

$$S = \{(1,2),(1,3),(1,4),(1,5),(2,1),(2,3),(2,4),(2,5),(3,1),\ldots,(5,4)\}$$

Uma vez que os resultados são igualmente prováveis e S contém 20 pontos, cada ponto de S ocorre com probabilidade igual a $\frac{1}{20}$. Portanto, sejam os seguintes eventos:

$A =$\{ímpar na primeira vez\}, $B =$ \{ímpar na segunda vez\} e $A \cap B =$ \{ambas as vezes\}.

Portanto, $A = \{(1,2),(1,3),\cdots,(1,5),(3,1),(3,2),\cdots,(3,5),(5,1),(5,2),\cdots,(5,4)\}$ é formado por 12 pontos amostrais; $B = \{(1,2),(3,2),\cdots,(5,2),\cdots,(1,4),\cdots,(5,4),\}$ também é formado por 12 pontos amostrais; $A \cap B = \{(1,3),(3,1),(1,5),(5,1),(3,5),(5,3)\}$ por 6 pontos amostrais.

As probabilidades são:

$$P(A) = \frac{12}{20} = \frac{3}{5}, \quad P(B) = \frac{12}{20} = \frac{3}{5} \quad \text{e} \quad P(A \cap B) = \frac{6}{20} = \frac{3}{10}.$$

Para avaliar $P(2X_1 + X_2 \geq 8)$, basta considerar $Y = 2X_1 + X_2$ e transformar o espaço amostral S em S_Y, da seguinte forma: $(1,2) = 4, (1,3) = 5, (1,4) = 6$, obtendo-se assim o espaço amostral associado a Y,

$$S_Y = \{4,5,6,7,5,7,8,9,7,8,10,11,9,10,11,13,11,12,13,14\}.$$

Desta forma obtém-se que: $P(2X_1 + X_2 \geq 8) = \frac{13}{20}$.

Exemplo 25

Sejam A, B e C três eventos arbitrários. Escreva as expressões correspondentes aos eventos abaixo e as fórmulas para o cálculo das probabilidades em termos de

$P(A)$, $P(B)$, $P(C)$, $P(A \cap B)$, $P(A \cap C)$, $P(B \cap C)$ e $P(A \cap B \cap C)$.

a) Nenhum dos três eventos ocorre, ou seja, $A^c \cap B^c \cap C^c$.

Sabemos que $A^c \cap B^c \cap C^c = (A \cup B \cup C)^c$, assim

$$\begin{aligned} P(A^c \cap B^c \cap C^c) &= P[(A \cup B \cup C)^c] \\ &= 1 - P(A \cup B \cup C) \\ &= 1 - [P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C)]. \end{aligned}$$

b) Pelo menos um dos três eventos ocorre, isto é, $A \cup B \cup C$.

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &+ P(A \cap B \cap C). \end{aligned}$$

c) Somente A ocorre, isto é, $A \cap B^c \cap C^c$.

Sabemos que $A \cap B^c \cap C^c = A \cap (B \cup C)^c$, assim

$$\begin{aligned} P(A \cap B^c \cap C^c) &= P[A \cap (B \cup C)^c] \\ &= P(A) - P[A \cap (B \cup C)] \\ &= P(A) - P[(A \cap B) \cup (A \cap C)] \\ &= P(A) - [P(A \cap B) + P(A \cap C) - P(A \cap B \cap C)] \\ &= P(A) - P(A \cap B) - P(A \cap C) + P(A \cap B \cap C). \end{aligned}$$

d) A e B ocorrem mas, C não ocorre, isto é, $A \cap B \cap C^c = (A \cap B) \cap C^c$.

$$P[(A \cap B) \cap C^c] = P(A \cap B) - P(A \cap B \cap C).$$

e) Os três eventos ocorrem, isto é, $A \cap B \cap C$.

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

f) Exatamente um dos eventos ocorre (um e somente um dos eventos ocorre), isto é,

$$(A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C),$$

o qual denotaremos simplesmente por X.

$$\begin{aligned} X &= (A \cap B^c \cap C^c) \cup (A^c \cap B \cap C^c) \cup (A^c \cap B^c \cap C) \\ &= [A \cap (B \cup C)^c] \cup [B \cap (A \cup C)^c] \cup [C \cap (A \cup B)^c], \end{aligned}$$

assim

$$\begin{aligned}
P(X) &= P[A\cap(B\cup C)^c]+P[B\cap(A\cup C)^c]+P[C\cap(A\cup B)^c]\\
&= P(A)-P[A\cap(B\cup C)]+P(B)-P[B\cap(A\cup C)]+P(C)-P[C\cap(A\cup B)]\\
&= P(A)+P(B)+P(C)-P[(A\cap B)\cup(A\cap C)]-P[(B\cap A)\cup(B\cap C)]\\
&- P[(C\cap A)\cup(C\cap B)]\\
&= P(A)+P(B)+P(C)-[P(A\cap B)+P(A\cap C)-P(A\cap B\cap C)]\\
&- [P(A\cap B)+P(B\cap C)-P(A\cap B\cap C)]\\
&- [P(A\cap C)+P(B\cap C)-P(A\cap B\cap C)]\\
&= P(A)+P(B)+P(C)-2P(A\cap B)-2P(A\cap C)-2P(B\cap C)\\
&+ 3P(A\cap B\cap C).
\end{aligned}$$

g) Exatamente dois dos eventos ocorrem (dois e somente dois dos eventos ocorrem), ou seja,

$$(A\cap B\cap C^c)\cup(A\cap B^c\cap C)\cup(A^c\cap B\cap C),$$

o qual denotaremos simplesmente por Y.

$$\begin{aligned}
Y &= (A\cap B\cap C^c)\cup(A\cap B^c\cap C)\cup(A^c\cap B\cap C)\\
&= [(A\cap B)\cap C^c]\cup[(A\cap C)\cap B^c]\cup[(B\cap C)\cap A^c],
\end{aligned}$$

assim

$$\begin{aligned}
P(Y) &= P[(A\cap B)\cap C^c]+P[(A\cap C)\cap B^c]+P[(B\cap C)\cap A^c]\\
&= P(A\cap B)-P(A\cap B\cap C)+P(A\cap C)-P(A\cap B\cap C)+P(B\cap C)\\
&- P(A\cap B\cap C)\\
&= P(A\cap B)+P(A\cap C)+P(B\cap C)-3P(A\cap B\cap C).
\end{aligned}$$

h) Pelo menos dois dos eventos ocorrem, isto é, $(A\cap B)\cup(A\cap C)\cup(B\cap C)$, o qual denotaremos por W.

$$\begin{aligned}
P(W) &= P[(A\cap B)\cup(A\cap C)\cup(B\cap C)]\\
&= P(A\cap B)+P(A\cap C)+P(B\cap C)-P(A\cap B\cap C)\\
&\quad -P(A\cap B\cap C)-P(A\cap B\cap C)+P(A\cap B\cap C)\\
&= P(A\cap B)+P(A\cap C)+P(B\cap C)-2P(A\cap B\cap C).
\end{aligned}$$

i) No máximo dois eventos ocorrem, isto é, $(A\cap B\cap C)^c$.

$$P[(A\cap B\cap C)^c]=1-P(A\cap B\cap C).$$

Exemplo 26

De uma urna que contém quatro bolas brancas e três vermelhas, tiram-se três bolas de uma só vez. Pede-se:

a) Obter o espaço amostral referente ao número de bolas brancas retiradas.
Como são retiradas 3 bolas:

$$S = \{0, 1, 2, 3\}.$$

b) Obter a probabilidade de ocorrência de cada ponto do espaço amostral.
Como a retirada das três bolas é feita de uma só vez, (o que equivale a retirar uma após a outra sem reposição) o espaço amostral básico do experimento será composto por $C_7^3 = 35$ pontos amostrais, ou seja, grupos diferentes de 3 bolas que podem ser formados a partir do grupo de 7. Os resultados podem ser resumidos da seguinte forma:

Brancas	Vermelhas	Número	Probabilidade
0	3	$C_4^0 C_3^3 = 1$	1/35
1	2	$C_4^1 C_3^2 = 12$	12/35
2	1	$C_4^2 C_3^1 = 18$	18/35
3	0	$C_4^3 C_3^0 = 4$	4/35

c) Obter um processo geral para o cálculo das probabilidades.

Representando por X o número de bolas brancas e por x um seu valor qualquer, se terá:

$$P(X = x) = \frac{C_4^x \cdot C_3^{3-x}}{C_7^3}, \text{ com } x = 0, 1, 2, 3.$$

Exemplo 27

Uma caixa contém 20 peças, das quais 5 são defeituosas. Extraem-se duas ao acaso. Qual a probabilidade de:

a) Ambas serem perfeitas?

b) Ambas serem defeituosas?

c) Uma ser perfeita e outra defeituosa?

a) Se $A = \{$ambas serem perfeitas$\}$

$$P(A) = \frac{C_{15}^2 \cdot C_5^0}{C_{20}^2} = \frac{105}{190} = 0,5526.$$

b) Se $B = \{$ambas serem defeituosas$\} = \{$nenhuma perfeita$\}$

$$P(B) = \frac{C_{15}^0 \cdot C_5^2}{C_{20}^2} = \frac{10}{190} = 0,0526.$$

c) Se $C = \{$uma ser perfeita e outra defeituosa$\}$

$$P(C) = \frac{C_{15}^1 \cdot C_5^1}{C_{20}^2} = \frac{75}{190} = 0,3947.$$

FATO: Seja S o espaço amostral referente ao experimento aleatório descrito no **exemplo 27** e seja X o número de peças perfeitas, de modo que S_X é o espaço amostral associado a X. Então, $P_S(B) = P_{S_X}(X = 0)$, $P_S(C) = P_{S_X}(X = 1)$ e $P_S(A) = P_{S_X}(X = 2)$, ou seja:

$$S = \{B,\ C,\ A\} \quad \text{corresponde a} \quad S_X = \{0,\ 1,\ 2\}.$$

3.6 Probabilidade condicional e independência estocástica

3.6.1 Probabilidade condicional

Probabilidade condicional é um conceito importante da teoria das probabilidades. As considerações seguintes conduzem de modo natural à definição formal.

Exemplo preparatório: Suponha que a tabela a seguir represente uma possível divisão dos alunos matriculados em dado Instituto de Matemática, num dado ano (exemplo obtido em Bussab e Morettin (2011)).

Sexo/Curso	Homens(H)	Mulheres(F)	Totais
Matemática Pura (M)	70	40	110
Matemática Aplicada (A)	15	15	30
Estatística (E)	10	20	30
Computação (C)	20	10	30
Totais	115	85	200

Vamos indicar por A o evento que ocorre quando, escolhendo-se ao acaso um aluno do Instituto, ele for um estudante de Matemática Aplicada. Os eventos M, E, C, H e F têm significados análogos. Desta maneira, vemos que

$$P(A) = \frac{N_A}{N} = \frac{30}{200}, \quad P(H) = \frac{N_H}{N} = \frac{115}{200}, \quad P(A \cap H) = \frac{N_{AH}}{N} = \frac{15}{200},$$

$$P(A \cup H) = P(A) + P(H) - P(A \cap H) = \frac{30}{200} + \frac{115}{200} - \frac{15}{200} = \frac{130}{200}.$$

Podemos considerar agora a subpopulação formada pelos homens. A probabilidade de que um aluno do Instituto, escolhido ao acaso nessa subpopulação, seja um estudante de Matemática Aplicada é igual a $\frac{N_{AH}}{N_H}$, em que N_{AH} é o número de homens estudantes de Matemática Aplicada e N_H é o número total de homens. O resultado é então $\frac{15}{115}$. Em termos de eventos tem-se:

$$A = \{\text{estudante de Matemática Aplicada}\} \quad \text{e} \quad H = \{\text{estudante é Homem}\}.$$

A notação adotada é $P(A|H)$, que pode ser lido de diversas maneiras:

- probabilidade do evento A dado o evento H ou,
- probabilidade condicional do evento A dado H ou,
- probabilidade do evento A condicional ao evento H.

Definido como:

$$P(A|H) = \frac{N_{AH}}{N_H} = \frac{P(A \cap H)}{P(H)}.$$

É claro que cada subpopulação pode ser considerada como sendo uma nova população; o termo subpopulação é usado unicamente por conveniência de linguagem servindo para indicar que existe uma população maior sendo considerada.

Por analogia com a fórmula acima, introduziremos agora a seguinte definição formal.

Definição 2. Seja B um evento cuja probabilidade é positiva. Para um evento A, arbitrário, definimos.

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

A quantidade assim definida será chamada de probabilidade condicional de A na hipótese B. No caso de todos os pontos amostrais terem probabilidades iguais, a probabilidade condicional pode ser avaliada diretamente no espaço amostral reduzido pelo evento B como,

$$P(A|B) = \frac{N_{AB}}{N_B},$$

ou seja, é o quociente do número de pontos amostrais comuns a A e B, pelo número de pontos de B. Obviamente tem-se que:

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

Considerar probabilidades condicionais de vários eventos com relação a uma hipótese particular H é equivalente a escolhermos H como um novo espaço amostral, com probabilidades proporcionais às probabilidades originais; o fator de proporcionalidade $P(H)$ é necessário para que se tenha a probabilidade total do novo espaço igual a 1. Essa formulação mostra que todos os teoremas gerais sobre probabilidades são válidos também para probabilidades condicionais, com respeito a qualquer hipótese particular H. Por exemplo, a relação fundamental para a probabilidade da ocorrência de A ou B ou ambos, toma a forma,

$$P[(A \cup B)|H] = P(A|H) + P(B|H) - P[(A \cap B)|H].$$

Se A e B forem eventos mutuamente exclusivos, então,

$$P[(A \cup B)|H] = P(A|H) + P(B|H).$$

Ou ainda, o **Teorema 3** sob a hipótese H resulta em,

$$P[(A \cap B^c)|H] = P(A|H) - P[(A \cap B)|H].$$

Exemplo resolvido: Numa cidade 70% dos prédios foram bem planejados e 60% dos prédios foram bem planejados e bem construídos. Se um prédio for aleatoriamente selecionado, pede-se:

i) qual é a probabilidade dele ter sido bem planejado e não ter sido bem construído?

ii) se ele foi bem planejado, qual é a probabilidade condicional dele ter sido bem construído?

A solução é simples desde que se identifique corretamente os eventos e as respectivas probabilidades: $A =\{$ prédio bem planejado$\}$ e $B =\{$ prédio bem construído $\}$. Solução:

i) $P(A \cap B^c) = P(A) - P(A \cap B) = 0,70 - 0,60 = 0,10$;

ii) $P(B|A) = P(A \cap B)/P(A) = 0,60/0,70 \approx 0,86$.

3.6.2 Teorema do produto das probabilidades

Vimos que a probabilidade condicional de A na hipótese H (ou dado H) é,

$$P(A|H) = \frac{P(A \cap H)}{P(H)}. \tag{3.1}$$

A fórmula (3.1) é frequentemente usada na forma,

$$P(A \cap H) = P(A|H)\, P(H). \tag{3.2}$$

Esse resultado é conhecido pelo nome de teorema do produto das probabilidades (ou teorema das probabilidades compostas). Para A e B dois eventos quaisquer, pertencentes ao mesmo espaço amostral, tem-se então que:

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

Para generalizar esse resultado para três eventos A, B, C, tome em primeiro lugar $H = B \cap C$ como hipótese e, então, aplique (3.2) uma vez mais; segue-se então que,

$$P(A \cap B \cap C) = P[A\,|(B \cap C)] \cdot P(B\,|C) \cdot P(C).$$

A escolha da hipótese H pode ser considerada como irrelevante, portanto,

$$\begin{aligned} P(A \cap B \cap C) &= P(A)\, P(B|A) P[C|(A \cap B)], \text{ ou} \\ &= P(B)\, P(A|B) P[C|(A \cap B)], \text{ ou} \\ &= P(C)\, P(B|C) P[A|(B \cap C)]. \end{aligned}$$

As generalizações para quatro ou mais eventos saem diretamente pelo mesmo processo.

3.6.3 Independência estocástica (ou probabilística)

A probabilidade condicional $P(A\,|H)$ não é, em geral, igual à probabilidade $P(A)$. No caso em que $P(A\,|H) = P(A)$ diremos que A é estocasticamente (probabilisticamente) independente de H. A expressão $P(A \cap H) = P(A\,|H) \cdot P(H)$ mostra que a condição $P(A\,|H) = P(A)$ pode ser escrita na forma,

$$P(A \cap H) = P(A) \cdot P(H). \tag{3.3}$$

Essa equação é simétrica em A e H, e mostra que sempre que A for estocasticamente independente de H, o mesmo se pode dizer de H com relação a A. É preferível, portanto, começarmos com a seguinte definição simétrica.

3.6.3.1 Eventos independentes

Definição 3 (Eventos independentes)**.** Dois eventos A e B são ditos independentes (estocasticamente independentes) quando $P(A|B) = P(A)$ e também $P(B|A) = P(B)$, o que resulta em,

$$P(A \cap B) = P(A)\, P(B)\,. \tag{3.4}$$

Essa definição engloba a situação na qual $P(B) = 0$, caso em que $P(A|B)$ não é definida. O termo estatisticamente independente é sinônimo de independência estocástica.

Suponhamos agora três eventos A, B e C, tais que:

i) $\left.\begin{array}{l} P(A \cap B) = P(A) \cdot P(B) \\ P(A \cap C) = P(A) \cdot P(C) \\ P(B \cap C) = P(B) \cdot P(C) \end{array}\right\}$ são independentes dois a dois;

ii) $P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C)$

Se os eventos A, B e C satisfazem a $i)$ e $ii)$, eles são **mutuamente independente**.

3.6.3.2 Eventos mutuamente independentes

Definição 4 (Eventos mutuamente independentes)**.**

$$\{A_i\}_{i=1}^{n} \text{ são mutuamente independentes} \Leftrightarrow P\left(\bigcap_{k=1}^{m} A_{i_k}\right) = \prod_{k=1}^{m} P(A_{i_k}) \;\; \forall \;\; 2 \leq m \leq n.$$

Relembrando que $\{A_{i_k}\}_{k=1}^{m}$ é qualquer subsequência de m dos eventos A_i. Portanto, os n eventos $A_1, A_2, \cdots, A_n$ são mutuamente independentes, se para todas as combinações $1 \leq i < j < k < \cdots \leq n$, as regras de multiplicação,

$$\begin{aligned} P(A_i \cap A_j) &= P(A_i) \cdot P(A_j), \\ P(A_i \cap A_j \cap A_k) &= P(A_i) \cdot P(A_j) \cdot P(A_k), \\ &\cdots \\ P(A_1 \cap A_2 \cap \cdots \cap A_n) &= P(A_1) \cdot P(A_2) \cdot \ldots \cdot P(A_n), \end{aligned} \tag{3.5}$$

se aplicam. Vale mencionar que o número total de equações definidas por (3.5) é $2^n - n - 1$. A primeira linha corresponde a C_n^2 equações, a segunda C_n^3, etc. Temos, portanto $2^n - n - 1$ condições a serem satisfeitas para n eventos serem considerados mutuamente independentes. Por outro lado, as C_n^2 condições da primeira linha são suficientes para garantir *independência dois a dois*.

Fato: Se os eventos $\{A_i\}_{i=1}^{n}$ são mutuamente independentes, então $(\{A_i\}_{i=1}^{n}) \bigcup (\{A_i^c\}_{i=1}^{n})$ formam um conjunto (ou uma sequência) com $2n$ eventos mutuamente independentes.

3.6.4 Teorema da probabilidade total

Enunciaremos agora um resultado útil que relaciona a probabilidade de um evento com probabilidades condicionais. Este resultado é conhecido como o Teorema da Probabilidade Total.

Sejam $B_1, B_2, \cdots, B_n$ eventos mutuamente exclusivos e exaustivos. Isto é, os eventos $\{B_1, B_2, \cdots, B_n\}$ satisfazem às seguintes condições:

$$B_i \cap B_j = \emptyset, \quad \forall\, i \neq j \quad \text{e também} \quad B_1 \cup B_2 \cup \cdots \cup B_n = S.$$

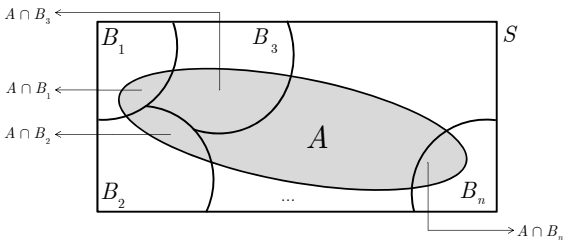
Então, um evento arbitrário A pode ser escrito como a união de eventos mutuamente exclusivos, isto é,

$$A = (A \cap B_1) \cup (A \cap B_2) \cup \ldots \cup (A \cap B_n),$$

logo,

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \ldots + P(A \cap B_n).$$

O diagrama de Venn a seguir ilustra os eventos no espaço amostral.



Mediante aplicação da definição de probabilidade condicional, temos

$$P(A|B_i) = \frac{P(A \cap B_i)}{P(B_i)} \quad \Rightarrow \quad P(A \cap B_i) = P(A|B_i)\, P(B_i).$$

assim

$$P(A) = P(A\,|B_1) \cdot P(B_1) + \cdots + P(A\,|B_n) \cdot P(B_n) = \sum_{i=1}^{n} P(A\,|B_i) \cdot P(B_i).$$

A utilidade deste resultado reside no fato de que as probabilidades que compõem o somatório acima são, em geral, conhecidas ou fáceis de serem calculadas.

3.6.5 Teorema (ou regra) de Bayes

Com base na definição de probabilidade condicional, pode-se estabelecer um resultado bastante útil conhecido como o Teorema de Bayes (ou Regra de Bayes ou ainda Fórmula de Bayes). Na formulação deste teorema os eventos B_i representam um conjunto de eventos mutuamente exclusivos e exaustivos, sendo geralmente considerados como possíveis causas

do evento A. Desta forma, para todo j pode-se avaliar a probabilidade condicional do evento B_j ter sido a causa da ocorrência do evento A como:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}. \tag{3.6}$$

O resultado dado pela Equação 3.6 é chamado Teorema de Bayes, em honra do filósofo inglês Thomas Bayes, e sua utilidade consiste em permitir-nos calcular a probabilidade a *posteriori* $P(B_j|A)$ em termos das probabilidades a *priori* $P(B_j)$ e $P(A|B_j)$, em geral conhecidas ou fáceis de serem avaliadas.

O resultado da seção 3.6.5 é a base teórica da Inferência Bayesiana, que se distingue da Inferência Clássica ou Fisheriana (abordada no Capítulo 7 acerca de Testes de Hipóteses), por admitir que a média (μ), a variância (σ^2) e quaisquer outros parâmetros desconhecidos da população, sejam variáveis aleatórias associadas a um modelo de distribuição de probabilidade. Este segmento ou alternativa para se inferir é computacionalmente intenso e tem se tornado bastante popular nos últimos trinta anos (desde 1980), devido ao avanço dos recursos computacionais.

3.6.5.1 Exercícios propostos com respostas

1) Quando mensagens codificadas são transmitidas por telégrafos podem ocorrer erros de transmissão. Por exemplo, o código Morse utiliza pontos e barras numa proporção de 3:4. Isto é, para qualquer símbolo enviado P(enviar ponto)=3/7 e P(enviar barra)=4/7 . Suponha que esteja havendo interferência na linha de transmissão de tal forma que 1/8 é a probabilidade de que um ponto recebido seja incorretamente interpretado como barra e vice-versa. Isto é, P(receber barra | enviado ponto)=P(receber ponto | enviado barra)=1/8. Qual é a probabilidade condicional de que um sinal tenha sido enviado como ponto, dado que ele foi recebido como ponto? (Resposta: $0,84$)

2) Um teste usado em amostras de sangue de indivíduos com suspeita de dengue é 95% acurado, isto é, indica o resultado correto em 95% dos casos (resultado positivo (+) indica com dengue e resultado negativo (−) caso contrário). Suponha que 10% da população esteja infectada com o vírus da dengue. Qual é a probabilidade condicional,

a) do teste resultar negativo (−) quando o indivíduo testado esta com dengue? (R:$0,05$)

b) do indivíduo testado estar com dengue quando o teste resultar negativo (−)?(R:$1/172 \approx 0,006$)

3) Numa usina nuclear somente 1 a cada 100 dias há algo errado com o reator. Em 99% dos dias em que há algo errado com o reator o alarme desta usina dispara, e, por outro lado, ele também dispara em 1% dos dias em que não há nada de errado com o reator. Qual é a probabilidade condicional de haver algo errado com o reator se o alarme disparou? (R: 0,50)

4) Suponha que um teste seja 99% acurado em indicar uma doença rara, que somente ocorre em 1 pessoa a cada 1 milhão. Avalie a probabilidade condicional de uma pessoa estar com esta doença rara se o teste indicar que sim (R: $9,9 \times 10^{-5} \approx 10^{-4}$ ou 1 chance em 10 mil).

3.7 Exemplos resolvidos

Exemplo 28

Extraem-se aleatoriamente duas cartas de um baralho comum de 52 cartas. Determine a probabilidade de serem ambas ases,

a) se a primeira carta é reposta (**com reposição**),

b) se a primeira carta não é reposta, isto é, **sem reposição**.

Sejam os eventos: $A_1 = \{$ ás na 1ª extração $\}$ e $A_2 = \{$ ás na 2ª extração$\}$. Sabemos que $P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1)$, assim

a) Na 1ª extração há 4 ases em 52 cartas, $P(A_1) = \frac{4}{52}$. Como a carta é resposta antes da 2ª extração, então $P(A_2 | A_1) = \frac{4}{52}$, pois há ainda 4 ases entre as 52 cartas na 2ª extração. Neste caso A_1 e A_2 são eventos independentes, então,

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2) = \left(\frac{4}{52}\right)\left(\frac{4}{52}\right) = \frac{1}{169}. \tag{3.7}$$

b) Como em a), $P(A_1) = \frac{4}{52}$. Mas, devido à ocorrência de um ás na 1ª extração, há agora apenas 3 ases entre as 51 cartas restantes, de modo que $P(A_2 | A_1) = \frac{3}{51}$. Neste caso A_1 e A_2 não são eventos independentes, então,

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 | A_1) = \left(\frac{4}{52}\right)\left(\frac{3}{51}\right) = \frac{1}{221}.$$

Exemplo 29

Sejam duas urnas I e II. A urna I contém três fichas vermelhas e duas fichas azuis, e a urna II contém duas fichas vermelhas e oito fichas azuis. Joga-se uma moeda "honesta". Se a moeda der "cara", extrai-se uma ficha da urna I; se der "coroa", extrai-se uma ficha da urna II. Pede-se:

a) Determine a probabilidade de escolha de uma ficha vermelha.

b) Dado que a ficha é vermelha, qual é a probabilidade condicional de ter vindo da urna I?

a) Seja o evento $V = \{$escolha de uma ficha vermelha$\}$, portanto,

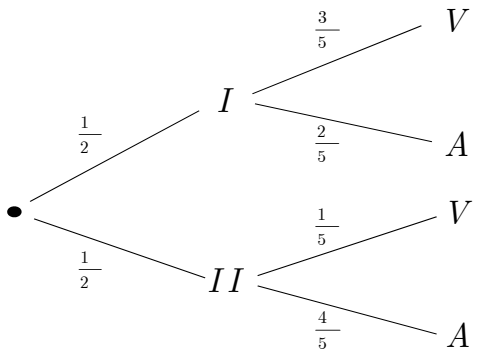
$$V = (V \cap I) \cup (V \cap II)$$

A escolha de uma das urnas são eventos equiprováveis, já que cara e coroa ocorrem com probabilidades iguais a, portanto, $P(I) = P(II) = \frac{1}{2}$. Basta fazer um diagrama em árvore para se determinar o espaço amostral deste experimento aleatório e as probabilidades associadas aos eventos de interesse. Portanto,

$$\begin{aligned} P(V) &= P(V \cap I) + P(V \cap II) \\ &= P(I)\ P(V|I) + P(II)\ P(V|II) \\ &= \frac{1}{2}\,\frac{3}{5} + \frac{1}{2}\,\frac{2}{10} \\ &= \frac{2}{5} = 0,4 \quad \text{ou} \quad 40\%. \end{aligned}$$

b) é uma aplicação direta do teorema de Bayes,

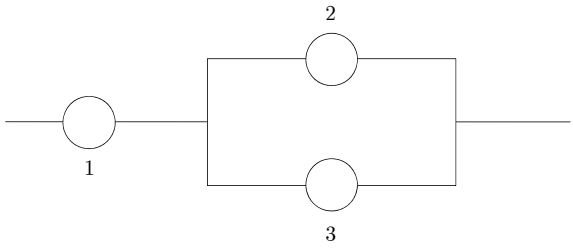
$$P(I|V) = \frac{P(V|I)P(I)}{P(V)} = \frac{\frac{3}{5}\frac{1}{2}}{\frac{2}{5}} = \frac{3}{4}.$$



Exemplo 30

Na Figura abaixo temos um sistema com três componentes funcionando independentemente, cada um com confiabilidade p_i. Considere as seguintes definições:

$$\begin{aligned} E = \{\text{o sistema funciona}\} &\rightarrow P(E) = \text{confiabilidade do sistema,} \\ R_i = \{\text{o } i\text{-ésimo componente funciona}\} &\rightarrow p_i = P(R_i) = \text{confiabilidade } i\text{-ésimo componente.} \end{aligned}$$



Pede-se: obtenha a confiabilidade do sistema.

Solução: considere os seguintes eventos: $A = R_1 \cap R_2$ e $B = R_1 \cap R_3$. Então o sistema funciona se pelo menos um dos dois eventos ocorrer. Ou seja, a confiabilidade do sistema é dada por,

$$\begin{aligned} P(E) &= P(A \cup B) \\ &= P(A) + P(B) - P(A \cap B) \\ &= P(R_1 \cap R_2) + P(R_1 \cap R_3) - P(R_1 \cap R_2 \cap R_3) \\ &= p_1p_2 + p_1p_3 - p_1p_2p_3. \end{aligned}$$

Exemplo 31

Uma água é contaminada se forem encontrados bacilos tipo A e/ou bacilos tipo B e C simultaneamente. As probabilidades de se encontrarem bacilos tipos A, B e C são respectivamente $0,30$, $0,20$ e $0,80$. Existindo bacilos tipo A, não existirão bacilos tipo B. Existindo bacilos tipo B, a probabilidade (condicional) de existirem bacilos tipo C é reduzida à metade. Pede-se:

a) Qual a probabilidade de aparecerem bacilos B ou C?

b) Qual a probabilidade da água estar contaminada?

c) Se a água está contaminada, qual é a probabilidade condicional de haverem bacilos do tipo B?

Considere o seguinte evento: $X = \{$ água contaminada $\} \Rightarrow X = A \cup (B \cap C)$. As probabilidades informadas são,

$$P(A) = 0,30, \quad P(B) = 0,20, \quad P(C) = 0,80 \quad \text{e} \quad P(C|B) = 0,40.$$

Sabemos também que $P(A \cap B) = 0$, pois foi informado que $A \cap B = \emptyset$. Portanto,

a) $P(B \cup C) = P(B) + P(C) - P(B \cap C)$, sendo que $P(B \cap C) = P(C|B)P(B) = 0,40 \times 0,20 = 0,08$. Logo, $P(B \cup C) = 0,20 + 0,80 - 0,08 = 0,92$.

b)

$$\begin{aligned} P(X) &= P[A \cup (B \cap C)] \\ &= P(A) + P(B \cap C) - P(A \cap B \cap C) \\ &= 0,30 + 0,08 - 0 = 0,38. \end{aligned}$$

c)

$$\begin{aligned} P(B|X) &= \frac{P(B \cap X)}{P(X)} \\ &= \frac{P\{B \cap [A \cup (B \cap C)]\}}{P(X)} \\ &= \frac{P[(B \cap A) \cup (B \cap C)]}{P(X)} \\ &= \frac{P(B \cap A) + P(B \cap C) - P(A \cap B \cap C)}{P(X)} \\ &= \frac{0 + 0,08 - 0}{0,38} = \frac{4}{19} \approx 0,21 \text{ ou } 21\%. \end{aligned}$$

Exemplo 32

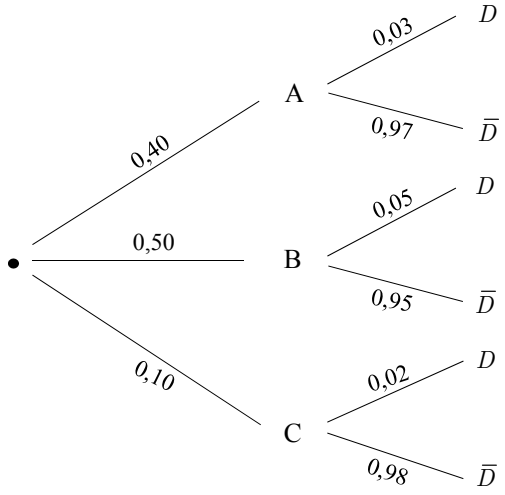
Em uma fábrica de peças, as máquinas A, B e C produzem 40, 50 e 10 por cento do total produzido, respectivamente. Da produção de cada máquina $3,5$ e 2 por cento, respectivamente, são peças defeituosas. Escolhida ao acaso uma peça da produção conjunta das três máquinas, pede-se:

a) Qual a probabilidade da peça escolhida ser defeituosa?

b) Sabendo-se que a peça escolhida é defeituosa, qual a probabilidade condicional de ter vindo da máquina B?

A solução deste exercício requer inicialmente que o espaço amostral seja convenientemente especificado. Isto pode ser realizado por meio de um diagrama em árvore com uma tabela auxiliar, conforme mostrados a seguir. $\overline{D} = D^c$ é o evento complementar de $D =$ { a peça escolhida é defeituosa },

	Classificação da peça		
Máquina	D	D^c	Total
A	0,012	0,388	0,40
B	0,025	0,475	0,50
C	0,002	0,098	0,10
Total	0,039	0,961	1,00



a) A probabilidade do evento D será obtida pelo teorema da probabilidade total com as seguintes probabilidades,

$$\begin{aligned} P(A) &= 0,40 \quad \text{e} \quad P(D|A) = 0,03, \\ P(B) &= 0,50 \quad \text{e} \quad P(D|B) = 0,05, \\ P(C) &= 0,10 \quad \text{e} \quad P(D|C) = 0,02. \end{aligned}$$

portanto,

$$\begin{aligned} P(D) &= P(D \cap A) + P(D \cap B) + P(D \cap C) \\ &= (0,03)\ (0,40) + (0,05)\ (0,50) + (0,02\ )(0,10) \\ &= 0,012 + 0,025 + 0,002 = 0,039. \end{aligned}$$

b) É uma aplicação do teorema de Bayes,

$$P(B|D) = \frac{P(B \cap D)}{P(D)} = \frac{(0,50)(0,05)}{0,039} = \frac{0,025}{0,039} = 0,641.$$

3.8 Exercícios propostos com respostas

1) Numa prova há 7 questões do tipo verdadeiro-falso (V ou F). Calcule a probabilidade de acertarmos todas as 7 questões se:

a) Escolhermos aleatoriamente as 7 respostas.

b) Escolhermos aleatoriamente as 7 respostas, mas, sabendo-se que há mais respostas V do que F (condicional).

2) Num exame de múltipla escolha há 3 alternativas para cada questão e apenas uma delas é correta. Portanto, para cada questão, um aluno tem probabilidade $\frac{1}{3}$ de escolher a resposta correta se ele está assinalando aleatoriamente e 1 se sabe a resposta. Um estudante sabe 30% das respostas do exame. Se ele assinalou corretamente uma das questões, qual é a probabilidade condicional de que ele tenha assinalado ao acaso?

3) Considere a escolha aleatória de um número entre os 10 primeiros números inteiros positivos (a partir de 1), e os eventos:

$$A = \{1, 2, 3, 4, 5\}, \quad B = \{4, 5, 6, 7\} \quad \text{e} \quad C = \{5, 9\}.$$

Pede-se: Os eventos são mutuamente independentes? Mostre porquê.

4) Uma urna contém 5 bolas vermelhas e 3 brancas. Uma bola é selecionada, aleatoriamente, dessa urna e não é reposta. Em seguida, duas bolas de cor diferente da bola extraída anteriormente (branca ou vermelha) são colocadas na urna. Se uma segunda bola é extraída aleatoriamente, qual é a probabilidade de:

a) A segunda bola ser vermelha?

b) A segunda bola ser da mesma cor da primeira?

5) Considerando o espaço amostral de um experimento constituído do lançamento de dois dados perfeitamente simétricos, pede-se:

a) Qual a probabilidade de que o primeiro dado mostre a face 2 e o segundo a face 3?

b) Qual a probabilidade de que ambos os dados mostrem a mesma face?

c) Qual a probabilidade de que o segundo dado mostre um número par?

6) Uma moeda perfeita é lançada 3 vezes e observado o número de caras. Qual é a probabilidade de ocorrer?

a) Pelo menos uma cara?

b) Somente cara ou somente coroa?

c) Exatamente uma cara?

7) Das 10 alunas de uma classe, 3 têm olhos azuis. Se duas delas são escolhidas aleatoriamente, qual é a probabilidade de:

a) Ambas terem olhos azuis?

b) Nenhuma ter olhos azuis?

c) Pelo menos uma ter olhos azuis?

8) Em um certo colégio, 25% dos estudantes foram reprovados em matemática, 15% em química e 10% em matemática e química ao mesmo tempo. Um estudante é selecionado aleatoriamente. Pede-se:

a) Qual é a probabilidade condicional de ter sido reprovado em matemática, se ele foi reprovado em química?

b) Se ele foi reprovado em matemática, qual é a probabilidade condicional de ter sido reprovado em química?

c) Qual é a probabilidade de ter sido reprovado em matemática ou química?

9) Um dado é viciado de tal forma que a probabilidade de ocorrer um certo número na face superior é proporcional ao seu valor. Pede-se:

a) Qual é a probabilidade condicional de ter ocorrido a face 3, sabendo-se que a face superior mostra um número ímpar?

b) Qual é a probabilidade condicional de ocorrer um número par, quando resultar na face superior um número maior que 3?

10) Sejam A, B e C três eventos de um mesmo espaço amostral S. Sabendo-se que:

$$P(A) = P(B) = \frac{1}{3}, \quad P(C) = \frac{1}{4}, \quad P(A \cap B) = \frac{1}{8},$$

$$P(A \cap C) = P(B \cap C) = \frac{1}{9}, \quad P(A \cap B \cap C) = \frac{1}{20}.$$

Calcular as probabilidades:

a) De ocorrer pelo menos um dos eventos A, B ou C;

b) De que não se realize nenhum dos eventos A, B ou C;

c) De que o evento A se realize, sabendo-se que já ocorreu B ou C (probabilidade condicional).

11) Seja $S = \{1, 2, 3, 4\}$ um espaço amostral equiprovável e os eventos $A = \{1, 2\}$, $B = \{1, 3\}$ e $C = \{1, 4\}$. Pede-se: verifique se os eventos A, B e C são mutuamente independentes.

12) Dois homens h_1 e h_2 e três mulheres m_1, m_2 e m_3 estão num torneio de xadrez. Os do mesmo sexo tem igual probabilidade de vencer, mas cada mulher tem duas vezes mais probabilidade de vencer o torneio do que qualquer um dos homens. Pede-se:

a) Qual é a probabilidade de que uma mulher vença o torneio?

b) Se h_1 e m_1 são casados, qual é a probabilidade de que um deles vença o torneio?

13) Um homem possui duas moedas, uma comum e a outra cunhada com duas caras. Ele apanhou uma moeda aleatoriamente e a lançou, se ocorreu a face cara, qual é a probabilidade condicional de que a moeda lançada tenha sido a de duas caras?

14) Jogam-se dois dados. Se as duas faces mostram números diferentes, qual é a probabilidade condicional de que uma das faces seja o 4?

15) Considere dois tipos de caixas de bombons, B e C. O tipo B contém 65% de bombons doces e 35% de bombons amargos, enquanto no tipo C essas percentagens de sabor são inversas. Além disso, 45% de todas as caixas de bombons são do tipo B e as restantes do tipo C. Escolhe-se, aleatoriamente, uma caixa e um bombom dessa caixa; se for constatado que ele é do tipo doce, qual é a probabilidade condicional de ter vindo de uma caixa do tipo C?

16) Definir e dar exemplos de: **a)** Eventos mutuamente exclusivos e, **b)** Eventos independentes.

17) Quatro urnas A, B, C e D contém bolas coloridas conforme especificado na tabela a seguir:

	COR DA BOLA		
URNA	VERMELHA	BRANCA	AZUL
A	1	6	3
B	6	2	2
C	8	1	1
D	0	6	4

Pede-se:

a) Se, aleatoriamente, extrai-se uma bola vermelha de uma das urnas, qual é a probabilidade condicional de ter sido da urna B?

b) Se forem extraídas duas bolas, sem reposição, da urna C, qual é a probabilidade de que ambas NÃO sejam vermelhas?

18) Numa placa de petri 20%, 40%, 25% e 15% do total das colônias bacterianas são dos tipos A, B, C e D, respectivamente. Sabe-se que 3%, 5%, 6% e 20% de cada colônia, respectivamente, são patogênicas.

a) Se for retirada uma amostra aleatória de uma única colônia bacteriana, qual é a probabilidade de que esta amostra contenha somente bactérias patogênicas?

b) Se for constatado que a amostra do item (a) possui somente bactérias patogênicas, qual é a probabilidade condicional de que as bactérias sejam do tipo D?

19) Quatro equipes A, B, C e D participam de um torneio que premiará uma única equipe campeã. Quanto às probabilidades de cada equipe vencer o torneio: as equipes C e D são equiprováveis, a equipe A é duas vezes mais provável do que B, a equipe B é duas vezes mais provável do que as equipes C e D. Pede-se: Qual é a probabilidade de que as equipes C ou D sejam campeãs?

20) Considere o seguinte experimento aleatório: Lançamento de um dado perfeitamente simétrico até que a face com o número 5 ocorra pela primeira vez. Pede-se:

a) O espaço amostral desse experimento.

b) Uma fórmula geral para o cálculo das probabilidades.

c) Mostre que a soma das probabilidades associadas aos pontos amostrais é um.

d) Qual é a probabilidade de ocorrer a face 5 no terceiro lançamento?

Dica: numa progressão geométrica $\{a_1, a_2, a_3, \ldots\}$ infinita ou ilimitada com $a_k = a_{k-1} \times q$, quando a razão $0 < q < 1$, a soma dos termos é dada por,

$$\sum_{i=1}^{+\infty} a_i = \frac{a_1}{1-q}.$$

21) Uma urna contém 5 bolas pretas, três vermelhas e duas brancas. Foram extraídas 3 bolas com reposição. Qual a probabilidade de terem sido duas bolas pretas e uma vermelha?

22) Uma caixa A contém 8 peças, das quais 3 são defeituosas e uma caixa B contém 5 peças, das quais 2 são defeituosas. Se uma peça é retirada aleatoriamente de cada caixa, pede-se:

a) Qual é a probabilidade de que ambas as peças não sejam defeituosas?

b) Qual é a probabilidade de que uma peça seja defeituosa e a outra não?

c) Se uma peça é defeituosa e a outra não, qual é a probabilidade condicional de que a peça defeituosa tenha vindo da caixa A?

23) Suponha que a probabilidade de que um vigia noturno num navio com luzes apagadas descubra um periscópio em certas condições de tempo seja igual a $0,7$. Pede-se: qual é a probabilidade de que uma combinação de dois vigias similares façam a descoberta? Qual é a pressuposição necessária para o cálculo desta probabilidade?

24) Suponha que A e B sejam eventos mutuamente exclusivos. Determine quais das relações abaixo são verdadeiras e quais são falsas. Justifique sua resposta.

a) $P(A|B) = P(A)$;

b) $P(A \cup B|C) = P(A|C) + P(B|C)$;

c) $P(A) = 0$, $P(B) = 0$, ou ambas;

d) $\dfrac{P(A|B)}{P(B)} = \dfrac{P(B|A)}{P(A)}$;

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

25) Repita o problema 24) supondo A e B eventos independentes.

26) Considere dois eventos, $A = \{$atirador A acerta o alvo$\}$ e $B = \{$ atirador B acerta o alvo $\}$, com probabilidades $P(A) = 0,51$ e $P(B) = 0,32$. Se os atiradores A e B atiram simultaneamente em um alvo, qual é a probabilidade do alvo ser atingido quando os eventos A e B:

a) são independentes?

b) são mutuamente exclusivos?

27) Considere uma moeda viciada de modo que 80% dos lançamentos mostram a face cara. Calcule as seguintes probabilidades:

a) Da face coroa ocorrer pelo menos uma vez em 5 lançamentos desta moeda.

b) De ocorrerem duas caras e uma coroa em 3 lançamentos desta moeda.

28) Os funcionários de uma empresa são classificados em 10% ótimos, 60% bons e 30% regulares. Um teste é proposto para classificar os funcionários em aprovado ou reprovado. Com base na classificação anterior, foram obtidas as seguintes probabilidades condicionais com o teste:

Classes	ótimos	bons	regulares
% de Aprovados	95	80	10

Pede-se: calcule a probabilidade condicional de um funcionário aprovado no teste pertencer a classe regulares.

29) Um teste usado em amostras de sangue de indivíduos com suspeita de dengue indica o resultado correto em 95% dos casos (resultado+indica com dengue e resultado − indica não doente). Suponha que em determinado bairro do RJ tenhamos 10% dos moradores infectados com o vírus da dengue. Calcule a probabilidade condicional,

a) do teste resultar − quando o morador do bairro está com dengue.

b) de um morador testado estar com dengue quando o teste resultar −.

30) Assinale **V** se a proposição for totalmente verdadeira e **F** caso contrário.

a) ( ) Se A e B são eventos independentes, então $P(A \cup B) = P(A) + P(B)$.

b) ( ) Para dois eventos quaisquer A e B, com $P(A) = 1/4$, $P(B) = 1/2$ e $P(A/B) = 1/3$, tem-se que $P(A \cap B) = 1/8$.

c) ( ) Para dois eventos quaisquer A e B, em que $\overline{A} = A^c$ e $\overline{B} = B^c$ são os respectivos complementos dos eventos, tem-se que $P(A^c \cap B^c) = 1 - P[(A \bigcup B)]$.

d) ( ) Para dois eventos quaisquer A e B, então $P(A \cap B^c) = P(B) - P(A \bigcap B)$.

e) ( ) Se A e B são dois eventos mutuamente exclusivos então $P(A \bigcap B) = P(A)\,P(B)$.

f) ( ) Se A, B e C são três eventos quaisquer com $P(C) \neq 0$, então $P[(A \bigcup B) \mid C] = P(A|C) + P(B|C)$.

g) ( ) Um espaço amostral finito consiste de três pontos amostrais s_1, s_2 e s_3, com probabilidades dadas respectivamente por $\frac{1}{2}p$, p^2 e p. Neste caso $p = 0,5$.

h) ( ) A probabilidade de que João resolva um problema é 1/3 e de que Pedro o resolva é 1/4. Se ambos tentam resolvê-lo independentemente, então a probabilidade de que o problema seja resolvido é 1/2.

31) Aproximadamente 1 em cada 90 nascimentos registrados são gêmeos, sendo 1/3 dos gêmeos idênticos (um óvulo) e 2/3 fraternos (dois óvulos). Gêmeos idênticos são necessariamente do mesmo sexo e com igual probabilidade para homens e mulheres. Gêmeos fraternos são 1/4 ambos mulheres, 1/4 ambos homens e 1/2 um de cada sexo. Considere os seguintes eventos: $A = \{$ nascimento de mulheres gêmeas $\}$, $B = \{$ nascimento de gêmeos idênticos $\}$ e $C = \{$ nascimento de gêmeos $\}$. Pede-se:

a) Defina o evento $A \cap B \cap C$ em palavras.

b) Calcule $P(A \cap B \cap C)$.

32) Sejam eventos A e B pertencentes a um mesmo espaço amostral com $P(A) = 1/3$ e $P(B^c) = 1/4$. Os eventos A e B podem ser mutuamente exclusivos? Explique com base nos axiomas de probabilidade.

33) Seja um evento A com $P(A) = p$. Define-se *chance* do evento A o quociente probabilidade de A ocorrer dividido pela probabilidade de A não ocorrer: $\frac{p}{1-p}$. Calcule a probabilidade do time A ser campeão este ano se os especialistas dizem que a chance de ser campeão é de:

a) "4 para 1";

b) "3 para 2".

34) Dois jogadores A e B disputam um jogo de azar no qual o vencedor recebe um prêmio de 40 moedas de ouro. O vencedor do jogo é aquele que primeiro vencer seis rodadas, sendo que em cada rodada do jogo a probabilidade de vitória é igual para ambos, portanto $P(A) = P(B) = 0,5$ por rodada. Pede-se: Se o jogo foi interrompido quando o jogador A havia vencido 5 rodadas e o B 3 rodadas, como deve ser dividido o prêmio?

35) Na última rodada da 1^a divisão do campeonato brasileiro de futebol do ano 2002, três times "grandes", A, B e C, corriam sério risco de serem rebaixados para a 2^a divisão. Um matemático calculou que $P(A) = P(B) = 0,4$ e $P(C) = 0,5$ eram as probabilidades de rebaixamento. Entretanto ele informava que time A ser rebaixado e time B ser rebaixado eram eventos mutuamente exclusivos. Ele também calculou que $P(A \bigcap C) = P(B \bigcap C) = 0,2$. Com base nestas informações calcule:

a) A probabilidade de rebaixamento de exatamente um dos três times (um e somente um).

b) A probabilidade de rebaixamento de pelo menos um destes três times.

36) Dois eventos A e B pertencentes a um mesmo espaço amostral possuem probabilidades $P(A) = 1/5$ e $P(B) = 1/6$. Se A e B são eventos independentes calcule a probabilidade de ocorrência de pelo menos um dos dois eventos.

37) Itens são inspecionados por uma firma antes de serem enviados aos compradores. Experiência demonstra que 8% dos itens inspecionados apresentam defeito do tipo A, 6% apresentam defeito do tipo B e 2% apresentam ambos os defeitos (A e B). Se um componente aleatoriamente selecionado é inspecionado, qual é a probabilidade dele apresentar **exatamente um** dos dois tipos de defeitos?

38) Considere que a probabilidade de um equipamento eletrônico falhar dependa da temperatura, de acordo com as seguintes **probabilidades condicionais**,

temperatura (t)	probabilidade de falhar
$t < 5^0C$	0,80
$5^0C \leq t \leq 15^0C$	0,40
$t > 15^0C$	0,10

Também considere que a temperatura no local de operação do equipamento se distribua de acordo com as seguintes probabilidades,

t	$t < 5^0C$	$5^0C \leq t \leq 15^0C$	$t > 15^0C$	Total
$P(t)$	0,10	0,50	0,40	1,00

Pede-se: Tendo-se observado uma falha no equipamento, qual é a probabilidade condicional de que ela tenha ocorrido com temperatura superior a 15^0C?

39) Nos itens a seguir assinale (V) se a afirmativa estiver totalmente correta ou (F) caso contrário. Se assinalar (F) indique aonde a afirmativa estiver incorreta.

a) ( ) Um espaço amostral de um experimento aleatório é um conjunto finito de elementos equiprováveis.

b) ( ) Uma moeda perfeitamente honesta foi lançada 10 vezes e observou-se 8 vezes a face cara. Conclui-se, pelo conceito de probabilidade *a posteriori*, que para esta moeda P(cara)=0,8 ou 80%.

c) ( ) Se dois eventos são mutuamente exclusivos, então a probabilidade da ocorrência simultânea destes dois eventos é igual a zero.

d) ( ) O conceito clássico ou *a priori* de probabilidade somente pode ser aplicado quando se considera um espaço amostral finito e equiprovável.

e) ( ) Um dado e uma moeda (não viciados) são lançados simultaneamente. A probabilidade de ocorrer a face cara na moeda e o número três no dado é inferior a 0,08 ou 8%.

40) Um réu foi a julgamento acusado de homicídio. Numa tentativa de inocentar seu cliente o advogado de defesa alega que ele é esquizofrênico (mentalmente doente) e portanto deve ser tratado e não preso. O advogado se baseia no resultado do exame de tomografia computadorizada (CAT) do réu que acusou atrofia cerebral. Um neurologista especialista em exames CAT informa que 30% dos esquizofrênicos são diagnosticados com atrofia cerebral enquanto que somente 2% dos indivíduos normais recebem o mesmo diagnóstico. Se 1,5% da população são esquizofrênicos, calcule a probabilidade condicional do réu ser um esquizofrênico, dado que seu exame CAT revelou atrofia cerebral?

41) Uma empresa irá pedir concordata se ocorrer o evento A ou se ocorrerem os eventos B e C simultaneamente. Os eventos A e B são mutuamente exclusivos e a probabilidade do evento B aumenta em 100% se o evento C ocorrer. As probabilidades são $P(A) = 0,05$, $P(B) = 0,10$ e $P(C) = 0,30$. Calcule a probabilidade desta empresa pedir concordata.

42) Assinale **V** se a afirmativa for inteiramente verdadeira ou **F** caso contrário. Justifique quando assinalar **F**.

a) ( ) Um matemático calculou que nas decisões de amanhã (18/04/2004) dos campeonatos carioca e mineiro respectivamente, poderão ser campeões Flamengo e Cruzeiro com probabilidade 0,43 ou Flamengo e Atlético com probabilidade 0,27 ou Vasco e Cruzeiro com probabilidade 0,21 ou Vasco e Atlético com probabilidade 0,20.

b) ( ) Em um jogo de azar sorteia-se dois números do conjunto $\{1, 2, 3, \ldots, 9, 10\}$ e portanto há 45 alternativas de sorteio. O apostador escolhe um número do mesmo conjunto e se ele acertar um dos dois sorteados ele ganha o prêmio. Então, pelo conceito clássico pode-se calcular que a probabilidade de ganhar é igual a $0,2$ ou 20%.

c) ( ) Pelo conceito axiomático de probabilidade tem-se que: (i) $P(A) \geq 0$ para qualquer evento A pertencente ao espaço amostral S; (ii) $P(S) = 1$ e (iii) Se A e B são dois eventos mutuamente exclusivos pertencentes a S, então $P(A \cup B) = P(A) + P(B)$.

d) ( ) Definido um experimento aleatório, tem-se que um espaço amostral associado a este experimento é um conjunto finito e equiprovável de possíveis eventos.

e) ( ) Considere a escolha aleatória de um número real x na reta $[0, 8]$. Ou seja $S = \{x : 0 \leq x \leq 8\}$. Seja o evento $A = \{x : 5 \leq x \leq 7\}$. Pelo conceito geométrico de probabilidade $P(A) = 0,25$.

43) Considere o lançamento de 3 dados perfeitamente simétricos e os seguintes eventos: $A =$ {os 3 dados mostram números pares na face superior}, $B =$ { os 3 dados mostram pelo menos um número diferente na face superior}. Calcule a probabilidade condicional do evento A dado o evento B. Isto é, calcule a probabilidade de que os três dados mostrem números pares, sabendo-se que eles mostram ou mostrarão pelo menos um número diferente.

44) Uma universidade possui um total de 50% de professores adjuntos, 30% de assistentes e 20% de professores auxiliares. Quanto ao tempo de serviço, 90% dos adjuntos e 50% dos assistentes possuem 10 ou mais anos, enquanto que 80% dos auxiliares têm **menos do que 10 anos** de serviço. Em uma pesquisa de opinião selecionou-se uma amostra aleatória de 50 professores desta universidade, todos eles com 10 ou mais anos de serviço. Pede-se: qual é o número mais provável de professores auxiliares desta amostra?

45) Considere um jogo de azar no qual dois números são simultaneamente sorteados do conjunto $\{1, 2, 3, \ldots, 10\}$. Calcule a probabilidade de um apostador acertar os dois números sorteados quando ele aposta em 3 números.

46) Nos itens a seguir assinale **V** somente se a afirmativa estiver totalmente correta ou assinale **F** e justifique ou explique ou indique o erro, caso contrário.

a) ( ) No lançamento simultâneo de um dado honesto e uma moeda honesta, a probabilidade de sair um número par e a face coroa é igual a 0,25.

b) ( ) Um grupo é formado por 5 homens casados, 3 mulheres casadas, 2 homens solteiros e 2 mulheres solteiras. Sorteia-se duas pessoas do grupo para formarem uma comissão, sem haver possibilidade de repetir a pessoas sorteada, de modo que o primeiro sorteado será

o presidente e o segundo o vice. Portanto, como podem ser formadas 132 comissões, a probabilidade de qualquer comissão (fulano como presidente e ciclano como vice) é igual a $1/132 \approx 0,008$.

c) ( ) Eventos mutuamente exclusivos são aqueles que ocorrem de modo independente um do outro.

d) ( ) Probabilidade *a posteriori* ou como frequência relativa é válida somente quando o espaço amostral do experimento é finito.

e) ( ) Considere a escolha aleatória de um ponto dentro de um círculo cujo raio é $r = 4$. Considere que dentro deste círculo há um círculo menor com $r = 2$. Então, pelo conceito de probabilidade geométrica, o ponto escolhido estará também dentro do círculo menor com probabilidade igual a 0,5. DICA: A área de um círculo é igual a πr^2.

47) Sejam X, Y e Z, 3 tipos de riscos associados a um investimento financeiro. Em uma análise de riscos, X, Y eZsão considerados eventos de um espaço amostral, cujas probabilidades são informadas na tabela a seguir.

Eventos	Probabilidades		Eventos	Probabilidades
X	0,35		X e Y	0,10
Y	0,20		X eZ	0,05
Z	0,15		Y eZ	0,06
			X e Y eZ	0,02

Pede-se: calcule as probabilidades dos eventos A, B e C, definidos a seguir:

$$\begin{aligned} \text{A} &= \{\textbf{Pelo menos um} \text{ dos eventos X,Y, Z}\}, \\ \text{B} &= \{\textbf{Exatamente dois} \text{ dos eventos X,Y,Z}\}, \\ \text{C} &= \{\textbf{Nenhum} \text{ dos eventos X,Y, Z}\}, \end{aligned}$$

48) Em uma grande empresa 60% do total de funcionários são do sexo masculino (homens). Sabe-se também que 10% dos homens e 25% dos funcionários do sexo feminino (mulheres), trabalham no setor de recursos humanos desta empresa. Pede-se: Se aleatoriamente for selecionado um funcionário do setor de recursos humanos, qual é a probabilidade condicional de que seja uma mulher?

49) Considere os eventos A e C definidos a seguir, com A^c e C^c os respectivos eventos complementares:

$$A = \{ \text{ o indivíduo é um fumante } \} \quad \text{e} \quad C = \{ \text{ o indivíduo desenvolve câncer } \}.$$

Assuma que as probabilidades associadas aos 4 eventos relacionados são conforme a seguir,

Evento	Probabilidade
$A \cap C$	0,15
$A \cap C^c$	0,25
$A^c \cap C$	0,10
$A^c \cap C^c$	0,50

Será que os dados anteriores sugerem uma relação entre o hábito de fumar e o desenvolvimento de câncer? Uma alternativa para se tentar responder esta questão é comparar duas probabilidades condicionais de interesse. Pede-se:

a) Calcule a probabilidade condicional de se desenvolver câncer dado que é fumante.

b) Calcule a probabilidade condicional de se desenvolver câncer dado que NÃO é fumante.

50) Em um procedimento de controle de qualidade amostra-se aleatoriamente 3 peças de caixas com 20 peças. Se a amostra revelar uma ou mais peças defeituosas então a caixa é separada e todas as 20 peças são rigorosamente inspecionadas, o que é denominado inspeção 100%. Calcule a probabilidade de haver inspeção 100% quando uma caixa contém 8 peças defeituosas.

51) Três eventos A, B e C são mutuamente independentes e ocorrem com probabilidades $1/8$, $1/4$ e $1/2$, respectivamente. Se um e somente um destes eventos ocorreu, calcule a probabilidade condicional de ter sido o evento A.

52) Numa espécie de inseto sabe-se que a população é formada por 70% de fêmeas e 30% de machos. Sabe-se também que 90% das fêmeas e 60% dos machos são estéreis. Calcule:

a) A probabilidade de se amostrar aleatoriamente um inseto não estéril desta espécie.

b) Quantos insetos devem ser amostrados para que se obtenha pelo menos 5 insetos não estéreis.

53) DEPUTADOS ACUSADOS DE CORRUPÇÃO. Na tabela abaixo estão indicados os nomes de 8 deputados que foram cassados, renunciaram ou que enfrenta processo, com respectivos partidos políticos. Outros 11 deputados acusados: 1 do PFL, 2 do PL, 2 do PP, 5 do PT e 1 do PTB, foram absolvidos das acusações (Fonte: Jornal Correio Braziliense, Brasília, 12 de julho de 2006).

	Partido				
Resultado	PL	PMDB	PP	PT	PTB
cassado	-	-	P.C.	J.D.	R.J.
renunciou	V.C.N. / B.R.	J.B.	-	P.R.	-
enfrenta processo	-	-	J.A.	-	-

Nomes: P.C.- Pedro Corrêa, J.D.- José Dirceu, R.J.- Roberto Jefferson, V.C.N.- Valdemar Costa Neto, B.R.- Bispo Rodrigues, J.B.- José Borba, P.R.- Paulo Rocha e J.A.- José Anene. Pede-se: Se aleatoriamente for sorteado um nome entre os 19 deputados acusados de corrupção,

a) Qual é a probabilidade de que ele seja filiado ao PT ou PL?

b) Qual é a probabilidade condicional de que ele seja do PL, dado que é um dos que renunciaram?

c) Qual é a probabilidade de que ele não tenha sido cassado e não enfrenta processo?

54) Um paciente está doente de uma entre 3 alternativas de doenças, A, B ou C, com probabilidades respectivamente iguais a 0,60; 0,30 e 0,10. Um exame laboratorial fornece resultado positivo $(+)$ para indicar paciente doente ou negativo $(-)$ para não doente de acordo com as seguintes probabilidades condicionais: Resultado positivo para 25% dos doentes com A, positivo para 70% dos doentes com B e positivo para 85% dos doentes com C.

a) Se o paciente realizar o exame laboratorial, qual é a probabilidade do teste resultar positivo $(+)$?

b) Qual é a probabilidade condicional do paciente estar com a doença C, dado que o resultado do exame laboratorial foi negativo $(-)$?

55) Tenta-se abrir uma porta escolhendo-se aleatoriamente uma chave de um chaveiro que contém n chaves e somente uma delas é a correta (consegue abrir). Qual é a probabilidade de se conseguir abrir a porta somente na $k-$ésima tentativa:

a) quando se descarta a chave usada após cada tentativa mal sucedida (amostragem sem reposição);

b) quando não se procede da maneira anterior, isto é, quando a mesma chave pode ser testada mais do que uma vez (amostragem com reposição).

c) Calcule as duas probabilidades anteriores para $n = 10$ e $k = 5$.

56) (clássico problema dos aniversários apresentado em diversos textos) Considere um grupo com n pessoas $(2 \leq n \leq 365)$ e calcule a probabilidade de pelo menos duas fazerem aniversário no mesmo dia (em anos iguais ou não). Assuma um ano com 365 dias de nascimentos equiprováveis. Verifique as probabilidades quando $n = 10, 20, 30$. DICA: calcule pelo complemento.

57) Considere que um veículo tenha se acidentado e sejam os seguintes eventos: $F = \{$ o freio foi a causa do acidente $\}$ e $C = \{$ a causa do acidente foi atribuída ao freio $\}$. Sabe-se que $0,04$ é a probabilidade do acidente ter sido causado pelo freio e também que a probabilidade condicional do acidente ser corretamente atribuído ao freio é igual a $0,82$ e a probabilidade condicional de ser incorretamente atribuída ao freio é $0,03$. Pede-se:

a) Calcule a probabilidade da causa do acidente ser atribuída ao freio.

b) Calcule a probabilidade condicional do acidente atribuído ao freio ser realmente devido ao freio.

58) Suponha que em uma partida de tênis entre os jogadores A e B, com rankings a e b respectivamente, que as probabilidades de vitórias sejam,

$$P(A) = \frac{b}{a+b} \qquad \text{e} \qquad P(B) = \frac{a}{a+b}$$

Por exemplo, se A é o 5^o do ranking e B é o 11^o então, $P(A) = 11/16 = 0,6875$ e $P(B) = 5/16 = 0,3125$. Em 2007 os dois jogos das semifinais do torneio Roland Garros estavam conforme a tabela a seguir,

Jogo	Jogador A (ranking) × Jogador B (ranking)
1	Federer (1) × Davydenko (4)
2	Nadal (2) × Djokovic (6)

Os vencedores dos jogos 1 e 2 se enfrentam no jogo final e o vencedor do jogo final é o campeão de 2007. Pede-se: calcule as probabilidades de cada semifinalista ser o campeão. DICA: faça um diagrama em árvore.

59) Uma pesquisa de mercado quanto à intenção de compra dos consumidores, mostrou que: 50% comprariam o produto do tipo A, 30% comprariam o produto do tipo B, 25% comprariam o produto do tipo C, 15% comprariam os produtos tipos A e B, $12,5\%$ comprariam os produtos tipos A e C, 10% comprariam os produtos tipos B e C, e, 2% comprariam os produtos tipos A, B e C. Seja X a variável aleatória discreta que represente o número de tipos de produtos que o consumidor estaria intencionado a comprar. Pede-se: calcule o percentual de consumidores intencionados a não comprar nenhum dos 3 tipos de produtos.

60) Do total de um tipo de medicamento encontrado no comércio, 80% são originais e legalizados, 15% são originais porém contrabandeados de outros países e 5% são falsificados no país. Se um doente realizar o tratamento com o medicamento original e legalizado considera-se probabilidade condicional igual a $0,95$ dele se curar, enquanto que se ele utilizar o medicamento original e contrabandeado a probabilidade é de apenas $0,50$, pois o medicamento é transportado em condições inadequadas. Se o doente utilizar o medicamento falsificado, além de nunca se curar ele ainda pode agravar a doença. Pede-se: Se um doente realizou o tratamento com o medicamento comprado todo de uma única vez, ou seja, de um mesmo tipo de procedência, e verificou-se que ele não se curou, calcule a probabilidade condicional de que tenha utilizado o medicamento falsificado.

Respostas dos exercícios propostos

1)a) $\frac{1}{128}$ 1)b)

2) $\frac{7}{16}$

3) Não, $P(B \cap C)$ e $P(A \cap B \cap C)$ não satisfazem a condição.

4)a) $\frac{41}{2}$ 4)b) $\frac{13}{36}$

5)a) 1/36 5)b) 1/6 5)c) 1/2

6)a) 7/8 6)b) 1/4 6)c) 3/8

7)a) 1/15 7)b) 7/15 7)c) 8/15

8)a) 2/3 8)b) 2/5 8)c) $0,3$

9)a) 1/3 9)b) 2/3

10)a) 223/360 10)b) 137/360 10)c) 67/170

11) Não são independentes porque a igualdade 3 a 3 não se verifica, isto é:
$P(A \cap B \cap C) \neq P(A) P(B) P(C)$

12)a) 3/4 12)b) 3/8

13) 2/3

14) 1/3

15) $0,3969$

17)a) 6/15 17)b) 1/45

18)a) $0,071$ 18)b) $0,4225$

19) $0,25$

20)a) $S = \{5, F5, FF5, \ldots\}$, sendo F qualquer face exceto a 5

20)b) A probabilidade de ocorrer a face 5 no n-ésimo lançamento do dado é: $P(n) = (\frac{5}{6})^{n-1}\frac{1}{6}$

20)c) $a_1 = \frac{1}{6}$, $q = \frac{5}{6}$ e $\sum_{i=1}^{+\infty} a_i = 1$

20)d) $\approx 0,116$

21) 9/40

22)a) 3/8 22)b) 19/40 22)c) 9/19

23) É necessário admitir independência dos eventos A_i, para $i = 1, 2$ sendo $A_i = \{$a descoberta do periscópio pelo i-ésimo vigia$\}$, neste caso o resultado é 0,91

24)a) F 24)b)V 24)c) F 24)d) V 24)e) F

25)a) V 25)b) F 25)c) F 25)d) F 25)e) V

26)a) $\approx 0,667$ 26)b) $0,83$

27)a) $\approx 0,672$ 27)b) $\approx 0,384$

28) $\frac{0,03}{0,605} \approx 0,0496$ ou $4,96\%$

29)a) $0,05$ 29)b) $1/172 \approx 0,006$

30)a)(F) 30)b)(F) 30)c)(V) 30)d)(F) 30)e)(F) 30)f)(F) 30)g)(V) 30)h)(V)

31)a) mulheres gêmeas idênticas 31)b) $1/540 \approx 0,19\%$

32) Não. $P(A \cup B) \leq 1 \Rightarrow P(A \cap B) \geq 1/12$

33)a) $0,80$ ou 80% 33)b) $0,60$ ou 60%

34) na proporção 7:1, 35 moedas para A e 5 moedas para B

35)a) $P(A \cap B^c \cap C^c) + P(A^c \cap B \cap C^c) + P(A^c \cap B^c \cap C) = 0,2 + 0,2 + 0,1 = 0,5$ 35)b) $P(A \cup B \cup C) = 0,9$

36) $P(A \cup B) = \frac{1}{3}$

37) $P(A \cap B^c) + P(A^c \cap B) = 0,06 + 0,04 = 0,1$

38) $P(t > 15^0 C/\mathrm{F}) = \frac{1}{8} = 0,125 \quad : \quad P(\mathrm{F}) = 0,32$

39)a) (F) conjunto finito e equiprováveis 39)b) (F) 10 vezes é pouco para conceito a posteriori e P(cara)=0,5 pois a moeda é honesta. 39)c) (V) 39)d)(V) 39)e)(F) $P(a \cap 3) = 1/12 = 0,083$ que é superior a $0,08$.

40) $\approx$ 0,186

41) $P\left[A \cup (B \cap C)\right] = 0,05 + 0,30.0,20 = 0,11$

42)a) pois $0,43 + 0,27 + 0,21 + 0,20 = 1,11 > 1$. 42)b) (V) 42)c) (V) 42)d) (F) S pode não ser finito e equiprovável 42)e) (V)

43) $P\left(A\right) = 27/216$, $P\left(B\right) = 1 - P(B^c) = 210/216$ e $P(A \cap B) = P\left(A\right) - P(A \cap B^c) = 27/216 - 3/216 = 24/216$ então $P(A/B) = 24/210 = 4/35 \approx 0,1143$

44) $P(\text{auxiliar}/ \geq 10) = 0,0625 \Longrightarrow$ número mais provável $= 0,0625 \times 50 = 3,125$ ou ≈ 3 a 4 professores.

45) Seja $P = P(\text{acertar os dois números sorteados})$, três soluções para o problema são: (1^a) seja A_i o evento acertar o i-ésimo número sorteado, então $P = P(A_1 \cap A_2) = P(A_1)P(A_2/A_1) = 3/10 \times 2/9 = 1/15$; (2^a) pelo conceito clássico e sob o ponto de vista do apostador $P = \frac{N_f}{N_t} = \frac{\binom{2}{2}}{\binom{8}{1}}\binom{10}{3} = 8/120 = 1/15$; (3^a) pelo conceito clássico e sob o ponto de vista do sorteador $P = \frac{N_f}{N_t} = \frac{\binom{3}{2}}{\binom{7}{0}}\binom{10}{2} = 3/45 = 1/15$. Portanto a probabilidade é igual a $1/15 \approx 0,067$.

46)a)(V) 46)b)(V) 46)c)(F) eventos mutuamente exclusivos são dependentes, pois se um ocorre o outro não ocorre 46)d)(F) qualquer S para *a posteriori*, requer somente n "grande" e prob. *a priori* requer S finito e equiprovável 46)e)(F) $\pi 2^2/\pi 4^2 = 0,25$

47) $P(A) = P(X \cup Y \cup Z) = 0,51$; $P(B) = P(X \cap Y \cap Z^c) + P(X \cap Y^c \cap Z) + P(X^c \cap Y \cap Z) = 0,15$; $P(C) = P(X^c \cap Y^c \cap Z^c) = 1 - P(A) = 0,49$. Um diagrama de Venn facilita a solução

48) Sejam os eventos: $R = \{\text{recursos humanos}\}$ e $M = \{\text{mulher}\}$. Pela fórmula de Bayes $P(M|R) = 0,10/0,16 = 0,625$

49)a) $P(C|A) = 0,375$ 49)b) $P(C/A^c) \approx 0,167$

50) $920/1140 \approx 0,81$ ou 81%

51) $P(A|E) = \frac{3}{31} \approx 0,097$ com $P(E) = P(A \cap B^c \cap C^c) + P(A^c \cap B \cap C^c) + P(A^c \cap B^c \cap C) = \frac{31}{64}$.

52)a) $0,19$ 52)b) $n \geq 5/0,19 \approx 26,32$

53)a) $11/19$ 53)b) $1/2$ 53)c) $15/19$

54)a) 0,445 54)b) $\approx$ 0,027

55)a) $1/n$ 55)b) $\left(\frac{(n-1)}{n}\right)^{k-1}\frac{1}{n}$ **c.** 0,1 e 0,06561

56) $p = 1 - \dfrac{\prod_{k=1}^{n-1}(365-k)}{365^{n-1}}$, $p \approx 0,1169$; $\approx 0,4114$; $\approx 0,7063$; para $n = 10, 20, 30$ respectivamente.

57)a) 0,0616 57)b) $\approx$ 0,5325

58) P(Federer)=4/7; P(Nadal)=3/10; P(Davydenko)=2/25 e P(Djokovic)=17/350

59) $30,5\%$

60) $0,303$

Capítulo 4

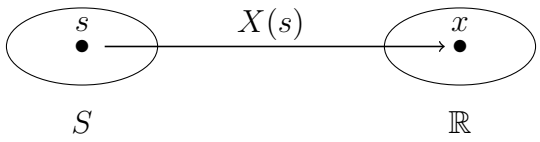
Variáveis aleatórias

4.1 Conceito

Uma variável aleatória é toda e qualquer variável cujos valores se distribuem segundo algum modelo de probabilidade, isto é, seus valores estão relacionados a um experimento aleatório. Uma variável aleatória é portanto uma função real definida em um espaço amostral, ou seja, se E é um experimento aleatório e S o espaço amostral associado a este experimento, uma função X que associe a cada elemento s pertencente a S um número real $X(s)$, é denominada variável aleatória:

$$\forall \ \ s \in S \quad \text{tem-se} \quad X(s) : S \mapsto S_x = \{x \in \mathbb{R} : X(s) = x\},$$

conforme a Figura a seguir, em que S_X é o espaço amostral associado a X, o conjunto de possíveis valores que X pode assumir.



Exemplo 33

Considere o lançamento de duas moedas e seja X o número de caras obtidas, denote $c =$ cara e $k =$ coroa. O espaço amostral deste experimento aleatório é dado por $S = \{kk, ck, kc, cc\}$ e o espaço amostral associado a X é dado por $S_X = \{0, 1, 2\}$. Portanto, $X(kk) = 0, X(ck) = X(kc) = 1$ e $X(cc) = 2$.

Observações:

a) Variável aleatória é uma função cujo domínio é S e contradomínio $\mathbb{R}$.

b) O uso de variáveis aleatórias permite descrever os resultados de um experimento aleatório por meio de números ao invés de palavras, o que apresenta a vantagem de possibilitar melhor tratamento matemático.

c) Nem toda função é uma variável aleatória.

As variáveis aleatórias são classificadas como **discretas** ou **contínuas**.

4.2 Variável aleatória discreta (v.a.d.)

Definição

Uma variável aleatória X é classificada como discreta (v.a.d.) se os valores x que X pode assumir formam um conjunto **enumerável**, finito ou infinito. Em geral o valor x é obtido mediante alguma forma de contagem. São exemplos de v.a.d.: **(i)** o número de acidentes ocorridos em uma semana; **(ii)** o número de peças defeituosas em uma amostra; **(iii)** o número de vitórias obtidas por uma equipe em um campeonato; **(iv)** o número de chamadas telefônicas recebidas por hora em uma central.

4.2.1 Função de probabilidade

Chama-se função de probabilidade (f.p.) da variável aleatória discreta X, a função $f(x)$ que fornece a probabilidade associada a cada valor x de X,

$$f(x) = P(X = x) = P(x) \quad \forall\ x.$$

Uma função de probabilidade deve satisfazer às seguintes condições:

i) $f(x) \geq 0$, para todo x;

ii) $\sum\limits_{x} f(x) = 1$.

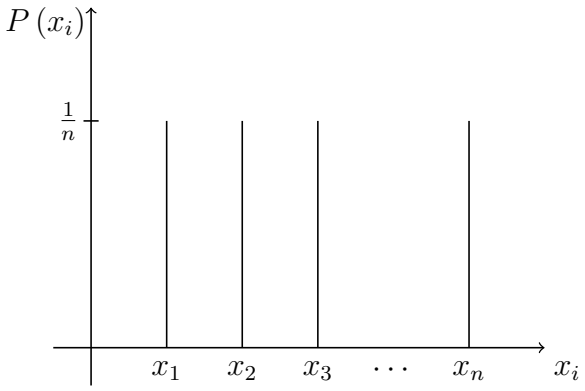
Distribuição de probabilidade: Os pares de valores $[x, P(x)]$ formam a distribuição de probabilidade da v.a.d. X, que pode ser representada por meio de tabelas e gráficos.

4.2.2 Variável aleatória discreta uniformemente distribuída

É a variável aleatória discreta X cujos valores $x_1, x_2, \ldots, x_n$, ocorrem com a mesma probabilidade. Portanto sua f.p. é dada por,

$$f(x) = \frac{1}{n}, \quad \text{para todo} \quad x.$$

i) Gráfico



ii) Tabela

x_i	x_1	x_2	x_3	x_4	...	x_n
$P(x_i)$	$1/n$	$1/n$	$1/n$	$1/n$	...	$1/n$

4.2.3 Exemplos resolvidos

1) Seja E: lançamento de um dado não-viciado e observar o número de pontos obtidos. O espaço amostral é: $S = \{1, 2, 3, 4, 5, 6\}$ e cada ponto de S tem probabilidade igual a $1/6$. Portanto, a variável aleatória X=\{número de pontos obtidos\}, tem distribuição uniforme discreta ou é uniformemente distribuída.
Tabela:

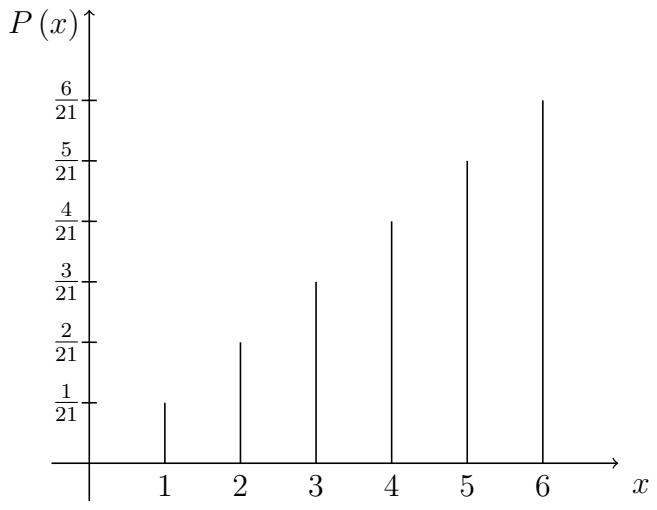
x_i	1	2	3	4	5	6
$P(x_i)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

2) Seja E: observar $X = \{$ número de pontos da face superior $\}$ no lançamento de um dado viciado, de tal forma que a probabilidade de cada face é proporcional ao número de pontos. Neste exemplo, o espaço amostral do experimento aleatório é o mesmo que aquele do exemplo anterior: $S = \{1, 2, 3, 4, 5, 6\}$, mas a probabilidade da face com um ponto é p, com dois pontos é $2p$ e assim por diante de modo que:

$$p + 2p + 3p + 4p + 5p + 6p = 1 \Rightarrow p = 1/21.$$

Para este exemplo a variável aleatória X não é uniformemente distribuída, tem-se:

(a) Gráfico:



(b) Tabela:

x_i	1	2	3	4	5	6
$P(x_i)$	$1/21$	$2/21$	$3/21$	$4/21$	$5/21$	$6/21$

A função de probabilidade é dada por,

$$f(x) = \begin{cases} \dfrac{x}{21}, & x = 1, 2, 3, 4, 5, 6, \\ \\ 0, & \text{para outros valores } x. \end{cases}$$

3) Uma urna contém 4 bolas azuis e 6 brancas. Duas bolas são retiradas sucessivamente: (I) com reposição e (II) sem reposição. Determinar, em cada caso, a distribuição de probabilidade e a função de probabilidade da v.a.d. X que representa o número de bolas brancas retiradas.

(I) Com reposição: $P(x) = \binom{2}{x}\left(\frac{6}{10}\right)^x\left(\frac{4}{10}\right)^{2-x}$, para $x = 0, 1, 2$.

x	0	1	2	total
$P(x)$	$\frac{4}{25}$	$\frac{12}{25}$	$\frac{9}{25}$	1,00

(II) Sem reposição: $P(x) = \dfrac{\binom{6}{x}\binom{4}{2-x}}{\binom{10}{2}}$, para $x = 0, 1, 2$.

x	0	1	2	total
$P(x)$	$\frac{6}{45}$	$\frac{24}{45}$	$\frac{15}{45}$	1,00

Convém salientar que o modelo (I) é o binomial e o (II) é o hipergeométrico, dois exemplos de modelos de distribuição para variáveis aleatórias discretas.

4.3 Variável aleatória contínua (v.a.c.)

Definição

Uma variável aleatória X é classificada como contínua (v.a.c.) se puder assumir todo e qualquer valor x em algum intervalo, digamos $a \leq x \leq b$, em que o comportamento probabilístico de X é determinado por uma função $f(x)$ denominada densidade de probabilidade. Portanto, uma v.a.c. está associada a um espaço amostral infinito não enumerável.

4.3.1 Função densidade de probabilidade

A função que denotaremos por $f(x)$, definida para $a \leq x \leq b$, será chamada função densidade de probabilidade (f.d.p.) se satisfizer às seguintes condições:

i) $f(x) \geq 0$, para todo $x \in [a, b]$;

ii) $\int_a^b f(x)\,dx = 1$.

Observações:

a) Para $c < d, P(c < X < d) = \int_c^d f(x)\,dx$;

b) Para um valor fixo de X, por exemplo, $X = x_0$, temos que $P(X = x_0) = \int_{x_0}^{x_0} f(x)\,dx = 0$; sendo assim, as probabilidades abaixo são todas iguais, se X for uma v.a.c.:

$$P(c \leq X \leq d) = P(c \leq X < d) = P(c < X \leq d) = P(c < X < d).$$

c) A função densidade de probabilidade $f(x)$, não representa probabilidade. Somente quando a função for integrada entre dois limites, ela produzirá uma probabilidade, que será a área sob a curva da função entre os valores considerados.

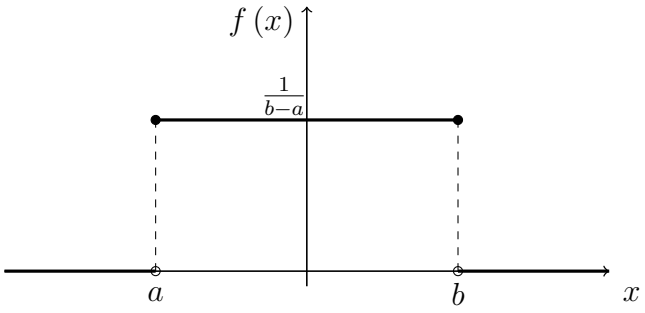
d) Se o conjunto de valores de X não estiver contido no intervalo $[a, b]$, isto é, $x \notin [a, b]$, então tem-se que $f(x) = 0$.

4.3.2 Variável aleatória contínua uniformemente distribuída

Definição 5. A variável aleatória contínua X tem distribuição uniforme no intervalo $[a, b]$, sendo a e b finitos, se a sua função densidade de probabilidade é dada por:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & \text{para } a \leq x \leq b; \\ 0, & \text{para outros valores de } x. \end{cases}$$

Gráfico:



4.3.3 Exercícios propostos com respostas

1) Seja uma v.a.c. X definida pela seguinte f.d.p.:

$$f(x) = \begin{cases} 0, & \text{para } x < 0, \\ kx, & \text{para } 0 \leq x \leq 2, \\ 0, & \text{para } x > 2. \end{cases}$$

a) Determinar o valor de k.

b) Traçar o gráfico da f.d.p.

c) Calcular $P(X \leq 1)$.

d) Calcular $P(X > 1)$.

Respostas: a) k=1/2, c) 1/4, d) 3/4.

2) Uma v.a.c. X tem a seguinte f.d.p.

$$f(x) = \begin{cases} 0, & \text{para } x < 0, \\ kx, & \text{para } 0 \leq x < 5, \\ k(10-x), & \text{para } 5 \leq x \leq 10, \\ 0, & \text{para } x > 10. \end{cases}$$

a) Determinar o valor de k.

b) Traçar o gráfico da f.d.p.

c) Calcular $P(X \geq 3)$.

Respostas: a) k=1/25, c) 41/50.

3) Uma v.a.c. X possui a seguinte função:

$$f(x) = \begin{cases} k, & \text{para } 0 \leq x < 1 \\ k(2-x), & \text{para } 1 \leq x < 2 \\ 0, & \text{para outros valores de } x \end{cases}$$

Pede-se:

a) A constante k para que $f(x)$ seja uma f.d.p.

b) $P(\frac{1}{2} \leq X < \frac{3}{2})$,

c) $P(X = 1)$,

d) $P(\frac{1}{2} \leq X \leq 1)$,

e) $P(X \geq 2)$,

f) Gráfico de $f(x)$.

Respostas: a) k=2/3, b) 7/12, c) 0, d) 1/3, e) 0.

4.4 Função de distribuição acumulada $[F(x)]$

Definição

Chamaremos de função de distribuição acumulada ou simplesmente função de distribuição, à função $F(x)$ associada à variável aleatória X, tal que: $F(x) = P(X \leq x)$. Observe que o domínio de F é todo o conjunto real.

Propriedades da $F(x)$

(i) $0 \leq F(x) \leq 1$ para todo x.

(ii) Se $x_1 \leq x_2$ então $F(x_1) \leq F(x_2)$, isto é, $F(x)$ é não decrescente.

(iii) $\lim_{x \to -\infty} F(x) = 0$ e $\lim_{x \to +\infty} F(x) = 1$

4.4.1 $F(x)$ para X uma v.a.d.

Para X uma variável aleatória discreta, temos que:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(x_i).$$

Fatos: Quando X é uma v.a.d. e para valores $x_1 < x_2$, tem-se:

$$\begin{aligned} P(x_1 < X \leq x_2) &= F(x_2) - F(x_1) \\ P(x_1 \leq X < x_2) &= F(x_2) - F(x_1) - P(x_2) + P(x_1) \\ P(x_1 \leq X \leq x_2) &= F(x_2) - F(x_1) + P(x_1) \\ P(x_1 < X < x_2) &= F(x_2) - F(x_1) - P(x_2) \end{aligned}$$

Exemplo 34

Seja X uma v.a.d. com a seguinte distribuição de probabilidade:

x_i	-2	-1	2	4	Total
$P(x_i)$	1/4	1/8	1/2	1/8	1,00

Pede-se:

a) Traçar o gráfico da distribuição de probabilidade de X.

b) Obter a função de distribuição acumulada e traçar seu gráfico.

Respostas:

a) Fica como exercício

b) Note que $P(x)$ é igual ao salto que $F(x)$ faz em x, daí a denominação de função escada, comumente utilizada em textos para $F(x)$ quado X é uma v.a.d.

$$F(x) = \begin{cases} 0 & , \text{ se } x < -2 \\ \frac{1}{4} & , \text{ se } -2 \leq x < -1 \\ \frac{3}{8} & , \text{ se } -1 \leq x < 2 \\ \frac{7}{8} & , \text{ se } 2 \leq x < 4 \\ 1 & , \text{ se } x \geq 4 \end{cases}$$

4.4.2 $F(x)$ para X uma v.a.c.

Para X uma v.a.c. temos que:

$$F(x) = P(X \leq x) = P(-\infty < X \leq x) = \int_{-\infty}^{x} f(u)du. \tag{4.1}$$

Fatos: Quando X é uma v.a.c. e para valores $x_1 < x_2$, tem-se:

$$\begin{aligned} P(x_1 < X \leq x_2) &= P(x_1 \leq X < x_2) = P(x_1 < X < x_2) = P(x_1 \leq X \leq x_2) \\ &= F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\, dx. \end{aligned}$$

Pode-se observar de (4.1) que a derivada da função de distribuição acumulada é a função densidade de probabilidade,

$$\frac{dF(x)}{dx} = f(x), \quad \text{em todos os pontos de continuidade de } f(x).$$

4.5 Exercícios propostos com respostas

1) Seja X uma v.a.c. com a seguinte f.d.p.:

$$f(x) = \begin{cases} \frac{1}{2}, & \text{para } 0 \leq x \leq 2 \\ 0, & \text{caso contrário} \end{cases}$$

Pede-se:

a) Traçar o gráfico da f.d.p.

b) Obter $F(x)$ e traçar seu gráfico.

c) Calcular $P(X \leq 1)$.

d) Calcular $P(1 \leq x \leq \frac{3}{2})$.

Respostas: b) $F(x) = 0$, se $x < 0$, $F(x) = \frac{x}{2}$, se $0 \leq x < 2$ e $F(x) = 1$, se $x \geq 2$ c) 1/2, d) 1/4.

2) Seja X uma v.a.c. e seja $f(x)$ uma função associada a X dada por:

$$f(x) = \begin{cases} ax & \text{se } 0 \leq x < 1 \\ a & \text{se } 1 \leq x < 2 \\ -ax + 3a & \text{se } 2 \leq x \leq 3 \\ 0 & \text{se } x < 0 \text{ ou } x > 3 \end{cases}$$

Pede-se:

a) Determinar a constante a para que $f(x)$ seja uma f.d.p.

b) Traçar o gráfico da f.d.p.

c) Obter $F(x)$ e traçar seu gráfico

d) $P\left(0 < X < \dfrac{3}{2}\right)$

e) Se X_1, X_2 e X_3 forem três observações independentes de X, qual será a probabilidade de exatamente um desses três números ser maior que $1,5$?

Respostas: a) $a = \frac{1}{2}$ c) $F(x) = 0$, se $x < 0$, $F(x) = \frac{x^2}{4}$, se $0 \leq x < 1$, $F(x) = \frac{(2x-1)}{4}$, se $1 \leq x < 2$ $F(x) = \frac{(-x^2+6x-5)}{4}$, se $2 \leq x < 3$ e $F(x) = 1$, se $x \geq 3$ d) 1/2, e) 3/8.

4.6 Variáveis aleatórias bidimensionais

4.6.1 Introdução

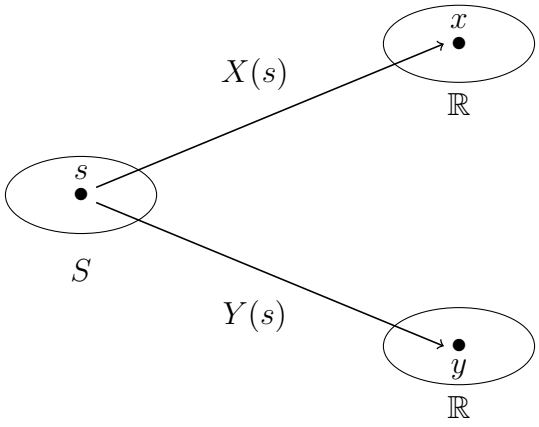
Variáveis aleatórias bidimensionais surgem quando para um determinado experimento aleatório, cada resultado s do espaço amostral S se associa a dois resultados $X(s) = x$ e $Y(s) = y$. Por exemplo, estudar a estatura X e o peso Y, de alguma pessoa escolhida ao acaso, o que forneceria o resultado (x, y). Portanto, cada resultado é identificado por cada um dos valores que as variáveis aleatórias unidimensionais assumem.

Em determinadas situações, X e Y não estão necessariamente ligadas a um mesmo experimento aleatório, mas existe uma razão bem definida para considerar X e Y conjuntamente.

Em situações reais de pesquisas, mais do que duas variáveis aleatórias $(X, Y, Z, W, \cdots)$ podem ser de interesse. Neste caso, tem-se um cenário multivariado ou multidimensional. As definições de distribuições que serão abordadas neste texto para o caso bidimensional (que é o multidimensional mais simples), são facilmente generalizadas para os casos multidimensionais.

4.6.2 Definição

Sejam E um experimento aleatório e S um espaço amostral associado a E. Sejam $X = X(s)$ e $Y = Y(s)$, duas funções, cada uma associando um número real a cada resultado $s \in S$. Denominamos (X, Y) uma variável aleatória bidimensional. Para o nosso estudo vamos considerar que X e Y sejam ambas discretas ou sejam ambas contínuas.



Do mesmo modo que no caso unidimensional, devemos associar à variável aleatória bidimensional (X, Y) um modelo de distribuição de probabilidade.

4.6.3 Distribuição conjunta, distribuições marginais e condicionais

4.6.3.1 (X, Y) é v.a.d. bidimensional

(X, Y) será uma v.a.d. bidimensional se os valores possíveis de X e Y forem finitos ou infinitos enumeráveis. Isto é, se os valores possíveis de (X, Y) podem ser representados por (x_i, y_j), $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, s$.

4.6.3.1.1 Função de probabilidade conjunta de X e Y

Chama-se de função de probabilidade conjunta da v.a.d. bidimensional (X, Y) a função:

$$P(X = x_i, Y = y_j) = P(x_i, y_j) = p_{ij}.$$

que a cada valor (x_i, y_j) associa sua probabilidade de ocorrência. Para que $P(x_i, y_j)$ seja uma função de probabilidade conjunta é necessário que satisfaça às seguintes condições:

i) $P(x_i, y_j) \geq 0$, para todo par (x_i, y_j),

ii) $\sum_i \sum_j P(x_i, y_j) = 1$.

4.6.3.1.2 Distribuição de probabilidade conjunta

É o conjunto de pares de valores $\{(x_i, y_j), P(x_i, y_j)\}$, $i = 1, 2, \ldots, r$ e $j = 1, 2, \ldots, s$.

	y				
x	y_1	y_2	...	y_s	$P(X = x_i)$
x_1	$P(x_1, y_1)$	$P(x_1, y_2)$	...	$P(x_1, y_s)$	$P(X = x_1)$
x_2	$P(x_2, y_1)$	$P(x_2, y_2)$	...	$P(x_2, y_s)$	$P(X = x_2)$
...	...	...	...	...	...
x_r	$P(x_r, y_1)$	$P(x_r, y_2)$	...	$P(x_r, y_s)$	$P(X = x_r)$
$P(Y = y_j)$	$P(Y = y_1)$	$P(Y = y_2)$	...	$P(Y = y_s)$	$1,00$

4.6.3.1.3 Distribuições marginais

Dada uma distribuição conjunta de duas variáveis aleatórias X e Y, podemos determinar a distribuição de X considerando-se todas as possíveis alternativas de valores para Y, ou alternativamente, a de Y considerando-se X. São as chamadas distribuições marginais. A distribuição marginal é constituída pelos valores da variável aleatória e as suas respectivas probabilidades marginais. A probabilidade marginal para cada valor é obtida da seguinte forma:

- para X: $P(X = x_i) = P(x_i) = \sum_{j=1}^{s} P(x_i, y_j)$;

- para Y: $P(Y = y_j) = P(y_j) = \sum_{i=1}^{r} P(x_i, y_j)$.

Com as probabilidades marginais para cada valor, podemos construir a distribuição marginal para cada uma das variáveis aleatórias.

- para X:

x_i	x_1	x_2	...	x_r	Total
$P(x_i)$	$P(x_1)$	$P(x_2)$	...	$P(x_r)$	$1,00$

- para Y:

y_j	y_1	y_2	...	y_s	Total
$P(y_j)$	$P(y_1)$	$P(y_2)$	...	$P(y_s)$	$1,00$

4.6.3.1.4 Distribuições condicionais

Seja x_i um valor de X, tal que $P(X = x_i) = P(x_i) > 0$. A probabilidade condicional de $Y = y_j$, dado que $X = x_i$ é dada por,

$$P(Y = y_j|X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}.$$

Assim, para x_i fixado, os pares $[y_j, P(Y = y_j|X = x_i)]$ definem a distribuição condicional de Y, dado que $X = x_i$.

y_j	y_1	$\cdots$	y_s	Total
$P(Y = y_j\|X = x_i)$	$P(Y = y_1\|X = x_i)$	$\cdots$	$P(Y = y_s\|X = x_i)$	$1,00$

Analogamente para a variável X:

$$P(X = x_i|Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}, \quad P(Y = y_j) > 0.$$

x_i	x_1	...	x_r	Total
$P(X = x_i\|Y = y_j)$	$P(X = x_1\|Y = y_j)$	...	$P(X = x_r\|Y = y_j)$	$1,00$

4.6.3.2 (X, Y) é v.a.c. bidimensional

A variável (X, Y) será uma v.a.c. bidimensional se (X, Y) puder assumir todos os valores em algum conjunto não enumerável.

4.6.3.2.1 Função densidade de probabilidade conjunta

Seja (X, Y) uma v.a.c. bidimensional. Dizemos que $f(x, y)$ é uma função densidade de probabilidade conjunta de X e Y, se satisfazer às seguintes condições:

i) $f(x, y) \geq 0$, para todo (x, y);

ii) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dxdy = 1$;

$f(x, y) = 0$ para (x, y) não pertencente aos intervalos de x e y.

Temos ainda que:

$P(a \leq X \leq b, c \leq Y \leq d) = \int_c^d \int_a^b f(x, y)\, dxdy = \int_a^b \int_c^d f(x, y)\, dydx.$

4.6.3.2.2 Distribuições marginais

As f.d.p.'s marginais de X e Y são dadas por:

$g(x) = \int_{-\infty}^{\infty} f(x, y)\, dy$ e $h(y) = \int_{-\infty}^{\infty} f(x, y)\, dx$, respectivamente.

Temos ainda que:

$P(a \leq X \leq b) = \int_a^b g(x)\, dx$ e $P(c \leq Y \leq d) = \int_c^d h(y)\, dy.$

4.6.3.2.3 Distribuições condicionais

Sejam X e Y v.a.c. com f.d.p. conjunta $f(x, y)$ e f.d.p. marginais dadas por $g(x)$ e $h(y)$.

A f.d.p. condicional de X, dado que $Y = y$ é definida por:

$$f(x|y) = \frac{f(x, y)}{h(y)}, h(y) > 0.$$

Analogamente, a f.d.p condicional de Y, dado $X = x$ é definida por:

$$f(y|x) = \frac{f(x, y)}{g(x)}, g(x) > 0.$$

As f.d.p.'s condicionais acima, satisfazem a todas condições impostas para uma f.d.p. unidimensional.

Deste modo, para y fixado, teremos:

i) $f(x|y) \geq 0$;

ii) $\int_{-\infty}^{\infty} f(x|y)\, dx = \int_{-\infty}^{\infty} \frac{f(x, y)}{h(y)} dx = \frac{1}{h(y)} \int_{-\infty}^{\infty} f(x, y)\, dx = \frac{h(y)}{h(y)} = 1.$

4.6.4 Variáveis aleatórias independentes

(X, Y) é uma v.a.d. bidimensional

Definição 6. Seja (X, Y) v.a.d. bidimensional. Dizemos que X e Y são independentes se, e somente se, para todo o par de valores(x_i, y_j) tem-se:

$$P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j).$$

Basta que esta condição não se verifique para um par(x_i, y_j) para que X e Y não sejam independentes. Neste caso diremos que X e Y são dependentes.

Definição 7. Seja (X, Y) uma v.a.d. bidimensional. Neste caso X e Y serão independentes se, e somente se:

$$P(X = x_i|Y = y_j) = P(X = x_i), \text{ para todo } i \text{ e } j.$$

ou equivalente se, e somente se:

$$P(Y = y_j|X = x_i) = P(Y = y_j), \text{ para todo } i \text{ e } j.$$

(X, Y) é uma v.a.c. bidimensional

Definição 8. Seja (X, Y) v.a.c. bidimensional. Diremos que X e Y são independentes, se e somente se a f.d.p. conjunta puder ser fatorada no produto das duas f.d.p.'s marginais:

$$f(x, y) = g(x)\, h(y).$$

Definição 9. Seja (X, Y) uma v.a.c. bidimensional. Neste caso X e Y serão independentes se, e somente se:

$$f(x|y) = g(x).$$

Neste caso é evidente que $f(y|x) = h(y)$.

4.6.5 Exemplo resolvido

Sejam X e Y variáveis aleatórias contínuas (v.a.c.) com função densidade de probabilidade (f.d.p.) conjunta dada por:

$$f(x, y) = \begin{cases} 3x, & \text{se } 0 \leq y \leq x \leq 1, \\ 0, & \text{para outros valores de } x \text{ e } y. \end{cases}$$

Pede-se:

a) $P(0 \leq X \leq 1/2, Y \geq 1/4)$;

b) $P(0 \leq X \leq 1/2)$;

c) $P(Y \geq 1/4)$;

d) X e Y são v.a. independentes? Justifique.

e) $P(0 \leq X \leq 1/2 \mid Y = 1/4)$;

f) $P(Y \geq 1/4 \mid X = 1/2)$;

g) $P(Y \leq \frac{X}{2})$.

Solução: Para se calcular as probabilidades deve-se inicialmente estabelecer os limites de integração corretamente. Para tanto, deve-se plotar o plano $X \times Y$ com a indicação do domínio $0 \leq y \leq x \leq 1$ da função $f(x, y)$.

a) $P(0 \leq X \leq 1/2, Y \geq 1/4)$. Esta probabilidade conjunta é obtida integrando-se a f.d.p. conjunta:

$$\int_{1/4}^{1/2} \int_{1/4}^{x} 3x \; dy \; dx = \frac{5}{128} \approx 0,04 \; (4\%) \quad \text{ou} \quad \int_{1/4}^{1/2} \int_{y}^{1/2} 3x \; dx \; dy = \frac{5}{128}.$$

b) $P(0 \leq X \leq 1/2)$. Esta probabilidade marginal para X é obtida integrando-se $g(x)$, a f.d.p. marginal de X:

$$\int_{0}^{1/2} g(x) \; dx = \frac{1}{8} = 0,125 \; (12,5\%) \quad \text{em que} \quad g(x) = \int f(x, y) \; dy = \int_{0}^{x} 3x \; dy = 3x^2.$$

c) $P(Y \geq 1/4)$. Esta probabilidade marginal para Y é obtida integrando-se $h(y)$, a f.d.p. marginal de Y:

$$\int_{1/4}^{1} h(y) \; dy = \frac{81}{128} \approx 0,63 \; (63\%)$$

em que $h(y) = \int f(x, y) \; dx = \int_{y}^{1} 3x \; dx = \frac{3}{2}(1 - y^2)$.

d) X e Y não são v.a. independentes, pois o produto das f.d.p.'s marginais não resulta na f.d.p. conjunta. Isto é:

$$g(x) \, h(y) = \frac{9}{2} x^2 (1 - y^2) \neq f(x, y) = 3x.$$

Portanto, note que o produto das probabilidades calculadas nos itens b) e c) não é igual à probabilidade calculada no item a):

$$P(0 \leq X \leq 1/2, Y \geq 1/4) \neq P(0 \leq X \leq 1/2) P(Y \geq 1/4).$$

e) $P(0 \leq X \leq 1/2 \mid Y = 1/4)$. Esta probabilidade condicional para X dado $Y = 1/4$ é obtida integrando-se $f(x|y)$, a f.d.p. condicional de X.

$$f(x|y = 1/4) = \left. \frac{f(x, y)}{h(y)} \right|_{y=1/4} = \frac{3x}{\frac{3}{2}(1 - \frac{1}{16})} = \frac{32}{15} x.$$

Note que a restrição $y \leq x$ deve ser atendida, portanto, $P(0 \leq X \leq 1/2 \mid Y = 1/4) = P(1/4 \leq X \leq 1/2 \mid Y = 1/4)$ é dada por:

$$\int_{1/4}^{1/2} \frac{32}{15} x \; dx = \frac{1}{5} = 0,2 \; (20\%).$$

f) $P(Y \geq 1/4 \mid X = 1/2)$. Esta probabilidade condicional para Y dado $X = 1/2$ é obtida integrando-se $f(y|x)$, a f.d.p. condicional de Y.

$$f(y|x = 1/2) = \left.\frac{f(x,y)}{g(x)}\right|_{x=1/2} = \left.\frac{3x}{3x^2}\right|_{x=1/2} = 2.$$

Novamente note que a restrição $y \leq x$ deve ser atendida, portanto, $P(Y \geq 1/4 \mid X = 1/2) = P(1/4 \leq Y \leq 1/2 \mid X = 1/2)$ é dada por:

$$\int_{1/4}^{1/2} 2\ dy = \frac{1}{2} = 0,5\ (50\%).$$

g) $P\left(Y \leq \frac{X}{2}\right)$. Esta probabilidade conjunta é obtida integrando-se $f(x,y)$, a f.d.p. conjunta, na região em que $y \leq \frac{x}{2}$. Os limites de integração podem ser estabelecidos plotando-se a reta $y = \frac{x}{2}$ no plano $X \times Y$, o que permite estabelecer as seguintes integrais:

$$P(Y \leq \frac{X}{2}) = \int_0^1 \int_0^{x/2} 3x\ dy\ dx = \int_0^{1/2} \int_{2y}^1 3x\ dx\ dy = \frac{1}{2} = 0,50\ (50\%).$$

4.6.6 Exercícios propostos com respostas

1) Dada a distribuição de probabilidade conjunta de (X, Y) na Tabela abaixo,

	Y		
X	0	1	2
0	0,10	0,20	0,20
1	0,04	0,08	0,08
2	0,06	0,12	0,12

Pede-se:

a) As distribuições de $X, Y, W = X + Y$ e $V = XY$.

b) Verifique que X e Y são independentes e obtenha a distribuição conjunta a partir das distribuições marginais de X e de Y.

c) As distribuições condicionais de X dado que $Y = 0$ e Y dado que $X = 1$.

d) $P(X \geq 1, Y \leq 1)$

e) $P(X \leq 1 \mid Y = 0)$

Respostas: a) $P(W = w) = \sum_{\substack{x,y \\ x+y=w}} P(x,y)$ d) $0,30$ e) $0,70$

2) Sejam X e Y v.a.c.'s com f.d.p. conjunta dada por:

$$f(x,y) = \begin{cases} k(2x+y), & \text{se } 2 \leq x \leq 6,\ \ 0 \leq y \leq 5, \\ 0, & \text{para outros valores de } x \text{ e } y. \end{cases}$$

Pede-se:

a) O valor de k.

b) $P(X \leq 3, 2 \leq Y \leq 4)$

c) $P(Y < 2)$

d) $P(X > 4)$

e) X e Y são v.a. independentes? Justifique.

f) $f\left(x|y\right)$

g) $f(x|Y = 1)$

h) $P\left(X > 3|0 \leq Y < 2\right)$.

Respostas: a) 1/210, b) 16/210, c) 72/210, d) 125/210, e) não, f) $f\left(x|y\right) = \frac{(2x+y)}{(32+4y)}$, g) $f(x|1) = \frac{2x+1}{36}$, h) 60/72.

4.7 Medidas de posição de uma variável aleatória

4.7.1 Esperança matemática

O valor esperado ou a esperança matemática de uma variável aleatória X é o valor médio calculado de acordo com o modelo de probabilidade associado a X. O valor esperado é um parâmetro denominado de tendência central. Parâmetro é uma característica de uma população e corresponde ao valor que se espera observar em média para uma variável aleatória X, cujos valores se distribuem segundo algum modelo de probabilidade.

4.7.1.1 Caso em que X é uma v.a.d.

Seja X uma v.a.d. com a seguinte distribuição de probabilidade:

x_i	x_1	x_2	$\cdots$	x_n	Total
$P\left(x_i\right)$	$P(x_1)$	$P(x_2)$	$\cdots$	$P(x_n)$	$1,00$

Define-se esperança matemática de X por:

$$E\left(X\right) = \mu_X = \mu = x_1P(x_1) + x_2P(x_2) + \cdots + x_nP(x_n),$$

$$E\left(X\right) = \sum_{i=1}^{n} x_iP(x_i).$$

Exemplo: Um fabricante produz peças tais que 10% delas são defeituosas e 90% não são defeituosas. Se uma peça defeituosa for produzida, o fabricante perde U$ $1,00$, enquanto uma peça não defeituosa lhe dá um lucro de U$ $5,00$. Seja a variável aleatória X que informa o lucro líquido por peça. Calcular a média do lucro líquido por peça. (R: $E\left(X\right) = 4,40$)

4.7.1.2 Caso em que X é uma v.a.c.

A esperança matemática de uma v.a.c. X é definida por:

$$E(X) = \int_{-\infty}^{+\infty} x\ f(x)\, dx.$$

Exemplo: Uma v.a.c. X apresenta a seguinte f.d.p.:

$$f(x) = \begin{cases} 0, & \text{para } x < 0 \\ \frac{x}{2}, & \text{para } 0 \leq x \leq 2 \\ 0, & \text{para } x > 2 \end{cases}$$

Calcular $E(X)$. (R: $E(X) = 4/3$)

4.7.1.3 Propriedades da esperança matemática

As propriedades a seguir apesar de estarem apenas demonstradas para quando X é uma v.a.c., valem também para quando X é uma v.a.d..

P1) Se X é uma v.a. com $P(X = k) = 1$, então $E(X) = k$, sendo k uma constante (ou a média de uma constante é a própria constante).

Prova: $E(X) = \int_{-\infty}^{+\infty} kf(x)\, dx = k \int_{-\infty}^{+\infty} f(x)\, dx = k.$

P2) A esperança matemática do produto de uma constante por uma variável é igual ao produto da constante pela esperança matemática da variável, ou seja, multiplicando-se uma variável aleatória por uma constante, sua média ficará multiplicada por essa constante.

Prova: $E(kX) = \int_{-\infty}^{\infty} kxf(x)\, dx = k \int_{-\infty}^{+\infty} xf(x)\, dx = kE(X).$

P3) A esperança matemática do produto de duas variáveis aleatórias independentes é igual ao produto das esperanças matemáticas das variáveis, ou seja, a média do produto de duas variáveis aleatórias independentes é o produto das médias.

Prova: $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\ y\ f(x,y)\ \ dx\ dy$

Se X e Y são v.a. independentes, a f.d.p. conjunta pode ser fatorada no produto das f.d.p. marginais de X e Y. Assim:

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x\ yg(x)\ \ h(y)\, dx\ dy = \int_{-\infty}^{\infty} xg(x)\, dx \int_{-\infty}^{\infty} yh(y)\, dy$$
$$E(XY) = E(X)\, E(Y).$$

Obs: $E(XY) = E(X)\, E(Y)$ não implica que X e Y sejam variáveis aleatórias independentes.

P4) A esperança matemática da soma ou da subtração de duas v.a. quaisquer é igual a soma ou a subtração das esperanças matemáticas das duas v.a., ou seja, a média da soma ou da subtração de duas v.a. é igual a soma ou subtração das médias.

Prova:

$$E(X \pm Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x \pm y) f(x,y)\, dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\, f(x,y)\, dxdy \pm \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y\, f(x,y)\, dxdy,$$

$$E(X \pm Y) = \int_{-\infty}^{\infty} x\, g(x)\, dx \pm \int_{-\infty}^{\infty} y\, h(y)\, dy = E(X) \pm E(Y).$$

P5) A esperança matemática da soma ou subtração de uma v.a. com uma constante é igual a soma ou subtração da esperança matemática com a constante, ou seja, somando-se ou subtraindo-se uma constante a uma v.a., a sua média ficará somada ou subtraída da mesma constante.

Prova: $E(X \pm k) = \int_{-\infty}^{\infty} (x \pm k) f(x)\, dx = \int_{-\infty}^{\infty} x f(x)\, dx + \int_{-\infty}^{\infty} k f(x)\, dx = E(X) \pm k.$

P6) A média de uma v.a. centrada é zero, ou seja, a média dos desvios dos valores da v.a. em relação a sua média é zero.

Obs: Dizemos que a v.a. está centrada quando todos os valores são expressos como desvios em relação à respectiva média, $(X - \mu_X)$.

Assim: $E[X - \mu_X] = E(X) - E[\mu_X] = \mu_X - \mu_X = 0$.

4.7.2 Mediana

A mediana de uma distribuição é o valor Md que satisfaz à seguinte condição:

$$P(X \leq Md) \geq \frac{1}{2} \quad \text{e} \quad P(X \geq Md) \geq \frac{1}{2}. \tag{4.2}$$

Quando X é uma v.a.d. a definição (4.2) pode resultar em mais de um valor Md, portanto no caso discreto pode existir mais do que um valor mediano na distribuição.
Exemplo: Obtenha o valor Md da seguinte distribuição,

x	0	2	4	6	11	20	total
$P(x)$	0,05	0,05	0,20	0,20	0,30	0,20	1,00

Solução: como $P(X \leq 11) = 0,80 > 1/2$ e também $P(X \geq 11) = 0,50 = 1/2$, então o valor 11 atende a condição. Mas observe que $P(X \leq 6) = 0,50 = 1/2$ e também $P(X \geq 6) = 0,70 > 1/2$, então o valor 6 também atende a condição. Portanto $Md_1 = 6$ e $Md_2 = 11$.

Para X uma v.a.c. o valor de Md satisfaz a:

$$\int_{-\infty}^{Md} f(x)\, dx = \frac{1}{2} \quad \text{e} \quad \int_{Md}^{+\infty} f(x)\ dx = \frac{1}{2}.$$

4.7.3 Moda

A moda ou valor modal de uma distribuição é o valor que possui maior probabilidade no caso discreto ou maior densidade de probabilidade no caso contínuo.

Exemplo: Seja X uma v.a.c., tal que:

$$f(x) = \begin{cases} 2x, & \text{para } 0 \leq x \leq 1 \\ 0, & \text{para outros valores} \end{cases}$$

Determinar:

a) $E(X)$

b) Moda

c) Mediana

d) Para $Y = 3X + 8$, calcule $E(Y)$

Respostas: a) 2/3, b) 1, c) $\sqrt{2}/2$, d) 10.

4.8 Medidas de dispersão de uma variável aleatória

4.8.1 Variância

A variância de uma distribuição é a medida que quantifica a dispersão dos valores em torno da média.

A variância de uma v.a. X é definida por:

$$V(X) = \sigma_X^2 = E[X - E(X)]^2 = E[X - \mu_X]^2.$$

- Para X v.a.d.: $V(X) = \sum_i (x_i - \mu_X)^2 P(x_i)$;

- Para X v.a.c.: $V(X) = \int_{-\infty}^{+\infty} (x - \mu_X)^2 f(x)\, dx$.

Uma fórmula mais prática para calcular a variância é:

$$V(X) = E\left(X^2\right) - [E(X)]^2,$$

pois,

$$\begin{aligned} V(X) &= E[X - E(X)]^2 \\ &= E\left\{X^2 - 2XE(X) + [E(X)]^2\right\} \\ &= E\left(X^2\right) - 2E(X)E(X) + [E(X)]^2 \\ &= E\left(X^2\right) - [E(X)]^2, \end{aligned}$$

em que:

- para X v.a.d.: $E(X^2) = \sum_i x_i^2 P(x_i)$,
- para X v.a.c.: $E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$.

4.8.1.1 Propriedades da variância:

As propriedades a seguir valem para X uma v.a.d. ou uma v.a.c.

P1) A variância de uma constante é igual a zero.

Prova: $V(k) = E[k - E(k)]^2 = E(k-k)^2 = 0$.

P2) Somando-se ou subtraindo-se uma constante à uma v.a., sua variância não se altera.

Prova:

$$\begin{aligned} V(X \pm k) &= E\left[(X \pm k) - E(X \pm k)\right]^2 \\ &= E[X - E(X) \pm (k-k)]^2 \\ &= E\left[X - E(X)\right]^2 = V(X) \\ V(X \pm k) &= V(X). \end{aligned}$$

P3) Multiplicando-se uma v.a. por uma constante, sua variância é multiplicada pelo quadrado da constante.

Prova:

$$\begin{aligned} V(kX) = E[kX - E(kX)]^2 & \\ &= E[kX - kE(X)]^2 \\ &= k^2 E\left[X - E(X)\right]^2 \\ V(kX) &= k^2 V(X). \end{aligned}$$

P4) A variância da soma de duas v.a. independentes é igual a soma das variâncias das duas variáveis.

Prova:

$$\begin{aligned} V(X+Y) &= E[X+Y]^2 - [E(X+Y)]^2 \\ &= E\left(X^2 + 2XY + Y^2\right) - [E(X) + E(Y)]^2 \\ &= E\left(X^2\right) + 2E(XY) + E\left(Y^2\right) - \{[E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2\} \\ &= E\left(X^2\right) + 2E(XY) + E(Y^2) - [E(X)]^2 - 2E(X)E(Y) - [E(Y)]^2. \end{aligned}$$

Se X e Y são independentes; $E(XY) = E(X)E(Y)$, assim:

$$\begin{aligned} V(X+Y) &= \{E\left(X^2\right) - [E(X)]^2\} + \{E(Y^2) - [E(Y)]^2\} \\ V(X+Y) &= V(X) + V(Y). \end{aligned}$$

Do mesmo modo:

$$V(X-Y) = V(X) + V(Y).$$

4.8.1.2 Desvio padrão

O desvio padrão de uma variável aleatória X é a raiz quadrada positiva da variância de X.

$$\sigma_X = \sqrt{V(X)} > 0.$$

4.8.2 Covariância

Sejam X e Y duas variáveis aleatórias. A covariância denotada por $Cov(X,Y)$ é definida por:

$$Cov(X,Y) = E\{[X - E(X)][Y - E(Y)]\}.$$

Desenvolvendo a expressão acima, temos:

$$\begin{aligned} Cov(X,Y) &= E\{XY - XE(Y) - YE(X) + E(X)E(Y)\} \\ &= E(XY) - E(X)E(Y) - E(Y)E(X) + E(Y)E(X) \\ &= E(XY) - E(X)E(Y), \end{aligned}$$

em que:

$$E(XY) = \sum_i \sum_j x_i y_j P(x_i, y_j), \text{ para } (X,Y) \text{ discreta e,}$$
$$E(XY) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xyf(x,y)\,dxdy, \text{ para } (X,Y) \text{ contínua.}$$

Fato: Para que haja covariância é necessário que existam pelo menos duas variáveis aleatórias. A covariância nos dá uma ideia da relação de dependência entre as variáveis. Observe que a variância é um caso particular da covariância, pois, $Cov(X,X) = V(X)$.

4.8.2.1 Propriedades da covariância

Proposições: Se X, Y, Z e W são variáveis aleatórias e a, b e k são constantes, então:

P1) $Cov(X,k) = 0$;

P2) $Cov(X,Y) = Cov(Y,X)$, isto é, a covariância é simétrica;

P3) Se $V(X) = 0$ ou $V(Y) = 0$, então $Cov(X,Y) = 0$;

P4) $Cov(aX,Y) = aCov(X,Y)$;

P5) $Cov(aX,bY) = abCov(X,Y)$;

P6) $Cov(X+Z, Y-W) = Cov(X,Y) - Cov(X,W) + Cov(Z,Y) - Cov(Z,W)$;

P7) $V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X,Y)$.

Exemplo ilustrativo: As propriedades da variância e da covariância resultam que,

$$V(3X + 2Y - W) = 9V(X) + 4V(Y) + V(W) + 12Cov(X,Y) - 6Cov(X,W) - 4Cov(Y,W).$$

4.9 Coeficiente de correlação

O coeficiente de correlação (de Karl Pearson, 1857-1936) é um parâmetro (ρ) que mede o grau de associação linear entre duas variáveis aleatórias. Para as variáveis aleatórias X e Y, o coeficiente de correlação é dado por:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)\,V(Y)}}, \quad \text{sendo que} \quad -1 \leq \rho_{XY} \leq 1.$$

Fatos:

- Se X e Y são variáveis aleatórias independentes então $Cov(X,Y) = 0$ e consequentemente $\rho_{XY} = 0$.
- Se $Cov(X,Y) = 0$ não implica que X e Y sejam variáveis aleatórias independentes, a não ser que X e Y tenham distribuição normal bivariada, ou seja, X e Y não correlacionados ($\rho_{XY} = 0$) não equivale, em geral, que X e Y sejam independentes.
- Mudanças de escala não alteram a correlação, por exemplo, se $Z = aX$ e $W = bY$, então $\rho_{ZW} = \rho_{XY}$.

4.10 Exercícios propostos com respostas

1) Sabendo-se que $Y = 3X - 5$ e que $E(X) = 2$ e $V(X) = 1$, calcule:

a) $E(Y)$
b) $V(Y)$
c) $E(X + 3Y)$
d) $E(X^2 + Y^2)$
e) $V(3X + 2Y)$
f) ρ_{XY}

2) Uma urna contém 5 bolas brancas e 7 bolas pretas. Três bolas são retiradas simultaneamente dessa urna. Se ganharmos R$ 200,00 por bola branca retirada e perdermos R$ 100,00 por bola preta retirada, qual é o lucro esperado?

3) Uma moeda honesta é lançada sucessivamente até sair a face cara ou até serem realizados 3 lançamentos. Obtenha a distribuição de X="número de lançamentos", calcule sua média, moda e variância.

4) Uma máquina de apostar tem 2 discos independentes. Cada disco tem 10 figuras: 4 maçãs, 3 bananas, 2 peras e 1 laranja. Paga-se 80 para acionar a máquina. Se aparecerem 2 maçãs ganha-se 40; 2 bananas 80; 2 peras 140 e 2 laranjas 180. Qual é o resultado esperado após inúmeras jogadas?

5) Um determinado artigo é vendido em caixas ao preço de 8 U.M. por caixa. Sabe-se que 20% dos artigos vendidos apresentam algum defeito de fabricação. Um comprador faz a seguinte proposta: Pede para amostrar, ao acaso, 10 artigos por caixa e pagará por caixa 10 U.M. se nenhum dos artigos amostrados for defeituoso; 5 U.M. se um ou dois artigos forem

defeituosos e 4 U.M. se três ou mais forem defeituosos. O que é melhor para o vendedor, manter o seu preço de 8 U.M. por caixa ou aceitar a proposta do comprador? Mostre porquê. Considere X="número de artigos defeituosos", com a seguinte distribuição de probabilidade: (**sugestão:** utilize a variável Y=valor pago por caixa)

x_i	0	1	2	≥ 3	Total
$P(x_i)$	0,1074	0,2684	0,3019	0,3223	1,00

6) (X, Y) é uma variável aleatória bidimensional discreta com a seguinte distribuição conjunta:

	Y		
X	-3	2	4
1	0,1	0,2	0,2
3	0,3	0,1	0,1

Pede-se calcular:

a) $E(X)$, $V(X)$ e σ_X

b) $E(Y)$, $V(Y)$ e σ_Y

c) $E(X+Y)$, $\text{Cov}(X,Y)$ e ρ_{XY}

d) X e Y são independentes?

7) Seja X uma v.a.c. com a seguinte f.d.p.:

$$f(x) = \begin{cases} \frac{2}{3}, & x \in [0,1] \\ \frac{2}{3}(2-x), & x \in (1,2] \\ 0, & \text{caso contrário} \end{cases}$$

Calcule:

a) A esperança matemática de $(X-1)^2$.

b) O desvio padrão de X.

c) A mediana.

8) Uma v.a.c. X possui f.d.p. dada por:

$$f(x) = \begin{cases} 6(x-x^2), & \text{se } 0 \leq x \leq 1, \\ 0, & \text{para outros valores de } x. \end{cases}$$

Calcular:

a) $P[\mu - 2\sigma \leq X \leq \mu + 2\sigma]$, $\mu = E(X)$, $\sigma = \sqrt{V(X)}$, dado $E(X^2) = 3/10$;

b) $F(x)$, a função de distribuição acumulada.

9) Suponha que X e Y tenham a seguinte distribuição conjunta:

	X		
Y	-1	1	2
-2	0,1	0,1	0,0
0	0,1	0,2	0,3
3	0,1	0,1	0,0

Sabendo-se que $V(X) = 1,41$ e $E(Y) = 0,2$; pede-se:

a) Moda da v.a. X. Justifique.

b) X e Y são v.a. independentes? Mostre.

c) $E\left(\frac{1}{2}X - 3Y^2 + 8\right)$

d) A correlação entre as v.a. X e Y (ρ_{XY}).

e) $V\left(\frac{1}{2}X - 3Y^2 + 8\right)$

10) Dada a função densidade de probabilidade abaixo:

$$f(x) = \begin{cases} \frac{1}{2}, & \text{se } 0 \leq x < 1 \\ -\frac{1}{4}(x-3), & \text{se } 1 \leq x \leq 3 \\ 0, & \text{para outros valores de } x \end{cases}$$

Calcule:

a) Mediana

b) $V(12X - 8)$, dado $E(X^2) = 5/3$

c) A função de distribuição acumulada.

d) $P(0,5 < X < 1,5)$

11) Uma v.a.c. X possui a seguinte f. d. p.:

$$f(x) = \begin{cases} 0, & \text{se } x < -1 \text{ ou } x \geq 4/3 \\ \frac{1}{2}\left(1 - x^2\right), & \text{se } -1 \leq x < 0 \\ \frac{1}{2}, & \text{se } 0 \leq x < 4/3 \end{cases}$$

Determinar:

a) $F(x)$, a função de distribuição acumulada

b) Mediana

c) $P(-0,5 \leq X \leq 0,5)$

12) Dada a distribuição conjunta abaixo, parcialmente indicada:

	X			
Y	-3	-2	-1	$P(y)$
-2	$1/15$	$1/15$		$7/30$
0	$8/30$		$2/15$	
1		$1/30$		$7/30$
$P(x)$	$6/15$	$7/30$		

Pede-se:

a) Verifique se X e Y são v.a. independentes.

b) $E\left(\frac{X^2}{3} - \frac{2Y}{5} - 10\right)$

c) $V(8 - 15X)$

13) Cite as propriedades da:

a) Esperança matemática;

b) Variância;

c) Covariância.

14) Conceitue:

a) Variável aleatória discreta;

b) Variável aleatória contínua.

15) Considere uma v.a.c. X associada à seguinte função:

$$f(x) = \begin{cases} kx^2, & \text{se } 2 \leq x < 5 \\ k(8 - x), & \text{se } 5 \leq x \leq 8 \\ 0, & \text{se } x \text{ assume outros valores} \end{cases}$$

Determinar:

a) O valor da constante k para que $f(x)$ seja uma f. d. p.

b) $P\left(\frac{8}{2} \leq X \leq \frac{27}{2}\right)$

16) Suponha que X e Y tenham a seguinte distribuição conjunta:

	X				
Y	1	2	4	5	$P(y)$
2	0,2	0,1	0,1	0,2	
3	0,1	0,1	0,1	0,1	
$P(x)$					

Pede-se:

a) $E\left(-\frac{1}{3}X\right)$ b) $V\left(5X-3Y\right)$

c) ρ_{XY}, o coeficiente de correlação linear.

d) X e Y são v.a. independentes? Justifique.

17) Sejam X e Y variáveis aleatórias independentes com,

$$E\left(X\right)=5,\quad V\left(X\right)=2,\quad E\left(Y\right)=8\quad \text{e}\quad V\left(Y\right)=3,$$

calcule:

a) $E\left(X-Y+3\right)$

b) $E\left[\left(X-Y\right)^2\right]$

c) $V\left(X-\frac{1}{3}Y\right)$

d) $V\left(3Y+2\right)$

e) Admita X e Y não independentes com $\rho_{XY}=0,7$ e calcule $V\left(2X+Y\right)$.

18) Seja (X,Y) uma variável aleatória bidimensional discreta, com a seguinte função de probabilidade:

$$P\left(x_i,y_j\right)=\begin{cases}\dfrac{2x_i+y_j}{42}, & \text{para } x_i=0,1,2 \text{ e } y_j=0,1,2,3,\\ 0, & \text{caso contrário.}\end{cases}$$

Pede-se:

a) A tabela com a distribuição de probabilidade conjunta;

b) A tabela com a distribuição marginal de X e também a de Y;

c) $E\left(X-2Y+4\right)$

d) Moda de X.

19) Dada a seguinte função:

$$f\left(x\right)=\begin{cases}k\left(2-x\right), & \text{se } 0\leq x<1,\\ k, & \text{se } 1\leq x\leq 2,\\ 0, & \text{caso contrário.}\end{cases}$$

Determinar

a) O valor de k para que $f\left(x\right)$ seja uma f.d.p.

b) $E\left(X^3\right)$

c) Mediana de X

d) Moda de X

e) $P\left(X\geq\frac{3}{2}\right)$

f) $P(X=1)$

20) Se X e Y são duas v.a. tais que, $E\left(X^2\right)=6$, $E\left(Y\right)=2$, $V\left(X\right)=5$, $V\left(Y\right)=7$, $\rho_{XY}=\frac{1}{2}$ e $Z=\frac{3}{2}Y-\frac{2}{5}X+\frac{3}{4}$.

a) $E\left(3Z-6\right)$ b) $V\left(9-2Z\right)$

c) X e Y são v.a. independentes? Justifique.

21) Seja X uma v.a.c. com f. d. p. dada por $f\left(x\right)=3x^2$, se $-1\leq x\leq 0$ e $f\left(x\right)=0$, caso contrário. Se $-2<a<0$ calcule:

$$P\left(X>a\left|X<\frac{a}{2}\right.\right).$$

22) Dado $f\left(x,y\right)=\begin{cases}\dfrac{3-y}{16}, & \text{se } 0\leq y\leq 2 \text{ e } 0\leq x\leq 4\\ 0, & \text{caso contrário}\end{cases}$. Determinar:

a) As funções marginais de X e Y.

b) Se X e Y são v.a. independentes.

23) Seja $f\left(x,y\right)=\begin{cases}(x+y-2xy), & \text{se } 0\leq x\leq 1 \text{ e } 0\leq y\leq 1,\\ 0, & \text{caso contrário.}\end{cases}$

Pede-se:

a) Mostre que $f(x,y)$ é uma f.d.p.

b) Obtenha as f.d.p.'s marginais de X e Y.

24) Suponha que as dimensões, X e Y, de uma chapa retangular de metal, possam ser consideradas variáveis aleatórias contínuas independentes, com as seguintes f.d.p.'s:

$$g\left(x\right)=\begin{cases}x-1, & \text{se } 1<x\leq 2\\ -x+3, & \text{se } 2<x<3\\ 0, & \text{para outros valores}\end{cases} \quad \text{e } h\left(y\right)=\begin{cases}\frac{1}{2}, & \text{se } 2<y<4\\ 0, & \text{para outros valores}\end{cases}$$

Pede-se: Ache a f.d.p. da área da chapa.

25) Seja (X,Y) uma variável aleatória discreta bidimensional com a seguinte função de probabilidade conjunta:

$$P(x,y)=\begin{cases}\binom{3}{x}\dfrac{1}{16}, & \text{para } \ x=0,1,2,3 \quad \text{e} \quad y=10,20,\\ \\ 0, & \text{para outros valores } (x,y).\end{cases}$$

Pede-se:

a) Calcule $P(Y=10)$.

b) X e Y são variáveis aleatórias independentes? Justifique sua resposta.

c) Apresente a tabela da distribuição conjunta das probabilidades.

26) Uma variável aleatória contínua X possui a seguinte função densidade de probabilidade:

$$f(x) = \begin{cases} k(1-x^2), & \text{se } -1 \leq x < 0, \\ k, & \text{se } 0 \leq x \leq \dfrac{4}{3}, \\ 0, & \text{para outros valores de } x. \end{cases}$$

Pede-se:

a) O valor de k.

b) A função de distribuição acumulada de X.

c) Calcule $P(X \leq 2/3)$.

27) Considere um jogo de azar no qual o jogador paga **determinado valor** para jogar e depois retira aleatoriamente **duas bolas** de uma urna que contém 10 bolas, sendo sete brancas, duas vermelhas e uma preta. O jogador recebe um prêmio para cada bola obtida, de acordo com a cor, conforme a tabela abaixo,

COR	branca	vermelha	preta
PRÊMIO	1	5	10

Pede-se: Qual deve ser o valor pago para jogar, de modo que o jogo seja justo? Isto é, de modo que a probabilidade do jogador perder ou ganhar algum valor sejam iguais? Explique seu raciocínio.

28) Considere a seguinte função densidade de probabilidade conjunta

$$f(x,y) = \begin{cases} 4xy, & \text{se } 0 \leq x \leq 1, \quad 0 \leq y \leq 1, \\ 0, & \text{outros valores.} \end{cases}$$

Pede-se:

a) Calcule $V(X-Y)$.

b) Justifique porquê X e Y são variáveis aleatórias independentes.

29) Considere a distribuição conjunta de probabilidades a seguir. Seja $W = X + Y$, calcule $V(W)$.

	Y	
X	2	4
0	0,3	0,1
1	0,5	0,1

30) Se um dado perfeitamente simétrico é lançado até sair a face com o número 6 ou até serem realizados no máximo 3 lançamentos, calcule o número médio de lançamentos.

31) O número de anos de serviço dos funcionários de uma grande empresa é uma variável aleatória discreta X, cuja função de probabilidade $f(x) = P(X = x)$ é dada na tabela a seguir,

$$f(x) = \begin{cases} 0,08, & x = 1, \ldots, 5 \\ 0,09, & x = 6, \ldots, 10 \\ 0,01, & x = 11, \ldots, 25 \\ 0, & \text{outros valores } x \end{cases}$$

a) Obtenha a função de distribuição acumulada $F(x) = P(X \leq x)$.

b) Qual é o percentual de funcionários com no máximo 10 anos de serviço.

c) Dentre os funcionários com no mínimo 10 anos de serviço, calcule o percentual com no mínimo 20 anos (probabilidade condicional).

32) Considere a variável aleatória discreta bidimensional, (X, Y), com a seguinte distribuição de probabilidades,

	y			
x	1	2	3	4
0	0,06	0,24	0,12	0,18
1	0,04	0,16	0,08	0,12

a) Calcule $P(1 \leq Y < 3)$.

b) Calcule $P(1 \leq Y < 3 | X = 1)$.

c) Explique os resultados encontrados nos itens a. e b.

33) A produção diária de uma peça resulta em Y itens defeituosos, cuja distribuição possui parâmetros média e variância, ambos iguais a 2. O lucro diário com a venda das peças é uma variável X dada por $X = 50 - 2Y - Y^2$. Calcule o valor esperado do lucro diário.

34) Sejam X e $\mathcal{E}$ variáveis aleatórias com

$$V(X) = 5, \quad V(\mathcal{E}) = 4 \text{ e } COV(X, \mathcal{E}) = -4,5.$$

Seja Y uma variável dada por $Y = b_0 + b_1 X + \mathcal{E}$. Para $b_0 = 20$ e $b_1 = 2$ calcule $V(Y)$, a variância de Y.

35) Considere a seguinte função densidade de probabilidade conjunta,

$$f(x, y) = \begin{cases} x + y, & 0 \leq x \leq 1 \quad \text{e} \quad 0 \leq y \leq 1, \\ 0, & \text{outros valores.} \end{cases}$$

a) Calcule a probabilidade conjunta: $P\left(X < 0,5\ ,\ Y > 0,25\right)$.

b) Calcule $E(2X - \frac{1}{6})$.

c) Calcule a probabilidade condicional: $P\left(X \geq 0,8 | Y = 0,5\right)$.

36) Considere a seguinte distribuição conjunta,

	X_2	
X_1	2	4
0	0,10	0,30
2	0,27	0,33

a) Calcule $P\left(\frac{X_1 + X_2}{2} \geq 2\right)$.

b) X_1 e X_2 são variáveis aleatórias independentes? Justifique sua resposta.

c) Calcule $V\left(2X_1 - X_2\right)$.

37) Seja $f\left(x, y\right)$ uma função densidade de probabilidade conjunta dada por,

$$f\left(x, y\right) = \begin{cases} \frac{6}{33}\left(x^2 + y^2 x\right), & 0 \leq x \leq 1 \quad \text{e} \quad 0 \leq y \leq 3, \\ \\ 0, & \text{outros valores.} \end{cases}$$

a) Calcule a probabilidade conjunta: $P\left(X < 1/2\ ,\ Y > 2\right)$.

b) Calcule a probabilidade marginal: $P\left(X > 3/4\right)$.

c) Calcule o valor médio de Y.

38) Painéis de madeira são oferecidos com duas opções de comprimento e três opções de largura, em metros, conforme a distribuição conjunta apresentada na tabela a seguir,

	Largura (Y)		
Comprimento (X)	1	2	3
2,5	0,05	0,05	0,10
5	0,10	0,50	0,20

Os painéis são comercializados com as bordas envolvidas em uma fita protetora de modo que $T = 2(X + Y)$ é a variável aleatória que representa o total de fita gasto para proteger um painel. Calcule a média e a variância de T por propriedades de $E(T)$ e $V(T)$ com base na distribuição de (X, Y) e também com base na distribuição de T.

39) Este é um problema com nomes e fatos reais. Vou a um churrasco e encontro o meu amigo Luiz Abrantes com as suas três filhas: Luiza, Paula e Bruna. Eu sei os nomes das filhas dele mas não tenho a menor idéia *de quem é quem* e portanto de forma completamente aleatória falarei os nomes. Considere que a variável aleatória X represente o número de nomes que eu acerto. Pede-se: Construa a tabela com a distribuição das probabilidades de X.

40) Considere a seguinte distribuição de probabilidades conjuntas:

$$P(x,y) = P(X = x, Y = y): \quad P(-2,2) = P(-1,1) = P(0,0) = P(1,1) = P(2,2) = 0,2.$$

a) Calcule a probabilidade condicional: $P\left(X = -2|Y = 2\right)$.

b) Calcule a média ou esperança matemática de W, sendo $W = X - 5Y + 6$.

41) Seja $f\left(x,y\right)$ uma função densidade de probabilidade conjunta dada por,

$$f\left(x,y\right) = \begin{cases} \dfrac{x^2}{14}\left(y+2\right), & \text{se } 1 \leq x \leq 2 \quad \text{e} \quad 0 \leq y \leq 2 \\ 0, & \text{para outros valores } (x,y). \end{cases}$$

Pede-se: X e Y são variáveis aleatórias independentes? Justifique sua resposta.

42) Uma máquina que produz componentes para discos rígidos de computadores pode operar a duas velocidades, lenta ou rápida. Na velocidade lenta o custo por peça é igual a 20,75 e na rápida é igual a 20,45. Na velocidade rápida mais peças são produzidas (menor custo), entretanto 5,48% das peças são defeituosas. Na velocidade lenta são produzidas menos peças porém somente 0,86% são defeituosas. Para cada peça defeituosa produzida na velocidade lenta ou na rápida, há um custo adicional igual a 10,40 para reparar a peça. Considere que as variáveis aleatórias X e Y representem respectivamente o custo de uma peça nas velocidades lenta e rápida. Calcule os custos esperados, ou seja, $E\left(X\right)$ e $E\left(Y\right)$.

43) Considere a função de distribuição acumulada da variável aleatória discreta X dada a seguir,

$$F\left(x\right) = \begin{cases} 0, & \text{se} \quad x < 0, \\ 2/6, & \text{se} \quad 0 \leq x < 1, \\ 5/6, & \text{se} \quad 1 \leq x < 3, \\ 1, & \text{se} \quad 3 \leq x. \end{cases}$$

Construa a tabela com a distribuição das probabilidades e calcule $E(10X - 5)$.

44) Sejam X e Y duas variáveis aleatórias tais que:

$$E\left(X\right) = 0, \qquad V\left(X\right) = 1 \quad \text{e} \quad Y = 5 - 2X.$$

Calcule:

a) $E\left(2X - 3Y - 4\right)$.

b) $V\left(3X - \frac{Y}{2} + 2\right)$.

c) ρ_{XY}, o coeficiente de correlação linear entre X e Y.

45) Considere a distribuição conjunta de probabilidades a seguir:

	Y		
X	1	2	3
0	0,03	0,05	0,02
2	0,27	0,45	0,18

a) X e Y são variáveis aleatórias independentes? Justifique sua resposta.

b) Se $W = XY$, calcule $E(W)$.

c) Se $W = XY$, calcule $V(W)$.

46) Seja (X, Y) uma variável aleatória contínua bidimensional tal que as duas f.d.p. marginais são dadas por,

$$g(x) = \frac{x}{2}, \quad 0 \leq x \leq 2 \quad \text{e} \quad h(y) = \frac{3y^2}{26}, \quad 1 \leq y \leq 3$$

Se possível, calcule $P(X \leq 1,\ Y \leq 2)$ e explique qual pressuposição é necessária para validar o cálculo.

47) O fabricante de um equipamento eletromecânico de cozinha conduziu um estudo com um grande número de consumidores, que utilizaram a assistência técnica autorizada, e verificou que todas as reclamações quanto ao produto podem ser classificadas em 6 categorias, conforme a distribuição das probabilidades apresentada na tabela a seguir.

	Natureza do defeito (Y)		
Prazo (X)	Elétrico	Mecânico	Estético
dentro da garantia	15%	13%	44%
fora da garantia	5%	6%	17%

a) A natureza do defeito e o prazo são variáveis aleatórias independentes? Justifique sua resposta.

b) Calcule a distribuição das probabilidades condicionais da natureza do defeito, quando o produto está dentro do prazo de garantia.

48) Um sistema eletrônico opera com dois componentes que funcionam simultaneamente. Sejam X e Y as duas variáveis aleatórias que denotam as vidas úteis destes componentes (em centenas de horas). Se $f(x, y)$ dada a seguir é a função densidade de probabilidade conjunta de (X, Y) calcule a seguinte probabilidade conjunta: $P(X > 1,\ Y > 1)$.

$$f(x,y) = \begin{cases} \dfrac{1}{8} x e^{-(x+y)/2}, & 0 < x < +\infty \quad \text{e} \quad 0 < y < +\infty \\ 0, & \text{para outros valores } x, y. \end{cases}$$

DICA: Os resultados a seguir podem ser úteis:

$$\lim_{x \to \infty} x e^{-x} = 0, \qquad \int x e^{Kx} dx = \frac{(Kx-1)}{K^2} e^{Kx} \qquad \text{e} \qquad \int e^{Kx} dx = \frac{1}{K} e^{Kx}.$$

49) Seja X a vida útil de um componente eletrônico, que representa o tempo de funcionamento em horas até ele apresentar a primeira falha. A função densidade de probabilidade de X é dada por,

$$f(x) = \begin{cases} K e^{-x/200}, & 0 \leq x < +\infty, \\ 0, & \text{para outros valores } x. \end{cases}$$

Pede-se:

a) O valor de K.

b) A probabilidade de um componente durar pelo menos 300 horas.

c) A probabilidade condicional de um componente durar pelo menos 700 horas sabendo-se que durar 300 horas é certo.

d) A função de distribuição acumulada de X.

e) Qual deve ser a garantia dada pelo fabricante de modo que no máximo 10% dos componentes tenham vida útil inferior à garantia?

50) Seja X uma variável aleatória contínua com a seguinte função densidade de probabilidade,

$$f(x) = \begin{cases} k, & -2 \leq x < 0, \\ k + \dfrac{3x}{125}, & 0 \leq x \leq 5, \\ 0, & \text{para outros valores } x. \end{cases}$$

a) Calcule o valor k e obtenha a $F(x)$.

b) Calcule $P(X \geq 0 | -1 < X < 3)$.

51) Seja Y uma variável aleatória discreta com função de probabilidade dada por,

$$P(Y = y_i) = \begin{cases} \dfrac{y_i}{N}, & \text{para } i = 1, 2, 3, 4 \\ \\ 0, & \text{para outros valores } i \end{cases}$$

em que,

$$N = \sum_{i=1}^{4} y_i \qquad \text{com} \qquad y_i = \sum_{k=i+1}^{i+2} k$$

Pede-se: Calcule $E(Y)$, o valor médio de Y.

52) Seja X a variável aleatória discreta que represente o número de artigos defeituosos por caixa, com função de distribuição acumulada dada por,

$$F(x) = \begin{cases} 0, & x < 0, \\ 0,68, & 0 \leq x < 1, \\ 0,95, & 1 \leq x < 2, \\ 0,98, & 2 \leq x < 3, \\ 1, & 3 \leq x. \end{cases}$$

Pede-se: Calcule o número médio de artigos defeituosos por caixa.

53) Calcule o valor de k na seguinte função densidade de probabilidade conjunta,

$$f(x,y) = \begin{cases} kx, & 0 \leq y \leq x \leq 2 \\ \\ 0, & \text{para outros valores } x \text{ e } y. \end{cases}$$

54) Considere a distribuição de probabilidades da v.a.d. tridimensional (X, Y, Z) dada na tabela a seguir,

	$X=1$		$X=4$	
Z	$Y=0$	$Y=1$	$Y=0$	$Y=1$
1	0,10	0,34	0,06	0,10
2	0,06	0,27	0,02	0,05

Pede-se:

a) Calcule a seguinte probabilidade condicional, $P(Y=0|X=4,\ Z=2)$.

b) Seja $W = \dfrac{X+Y}{2}$, calcule $E(W)$ e $V(W)$ diretamente pela distribuição de W (tente também pela distribuição conjunta de X e Y).

55) Seja X a variável aleatória contínua que represente o tempo (em segundos) que um rato de laboratório demora para executar uma tarefa e alcançar a comida, como recompensa pela tarefa. Quanto menor o tempo considera-se que maior é a inteligência do rato. Seja $f(x)$ uma função associada a X dada por,

$$f(x) = \begin{cases} \dfrac{t}{x^2}, & t \leq x < +\infty, \\ \\ 0, & \text{outros valores } x, \end{cases}$$

em que t é o menor valor possível do tempo para execução da tarefa. Pede-se:

a) Mostre que $f(x)$ é uma função densidade de probabilidade.

b) Calcule $P(X \geq t+h)$ para uma constante positiva h.

c) Para $t=5$, calcule $P(X \geq 7 \mid 5 < X < 10)$.

56) Seja (X, Y) uma variável aleatória contínua bidimensional com a seguinte função densidade de probabilidade conjunta,

$$f(x,y) = \begin{cases} 6(1-y), & 0 \leq x \leq y \leq 1, \\ \\ 0, & \text{para outros valores } x, y. \end{cases}$$

a) Obtenha as f.d.p.'s marginais de X e Y.

b) Calcule $P\left(Y \leq \frac{1}{2} | X \leq \frac{3}{4}\right)$.

c) Obtenha $f(x|y)$, a f.d.p. condicional de X dado $Y = y$.

d) Obtenha $f(y|x)$, a f.d.p. condicional de Y dado $X = x$.

e) Calcule $P\left(Y \geq \frac{3}{4} \mid X = \frac{1}{2}\right)$.

Respostas dos exercícios propostos

1)a) 1 1)b) 9 1)c) 5 1)d) 15 1)e) 81 1)f) 1.

2) R$75,00

3) $E(X) = 1,75$; Moda= 1; $V(X) = 11/16$

4) $E(X) = -59$

5) O vendedor deve manter o seu preço $[E(Y) \cong 5,21]$

6)a) 2; 1 e 1 6)b) 0,6; 9,24 e 3,04 6)c) 2,6; $-1,2$ e $-0,395$ 6)d) não são

7)a) 5/18 7)b) 0,478 7)c) 3/4

8)a) $\cong 0,9855$ 8)b) $F(x) = \begin{cases} 0, & \text{se } x < 0, \\ x^2(3-2x), & \text{se } 0 \leq x < 1 \\ 1, & \text{se } x \geq 1 \end{cases}$

9)a) 1 9)b) não 9)c) 0,55 9)d) $\cong -0,0737$ 9)e) $\cong 119,57$

10)a) 1 10)b) 71 10)c) $F(x) = \begin{cases} 0, & \text{se } x < 0, \\ \dfrac{x}{2}, & \text{se } 0 \leq x < 1 \\ -\dfrac{1}{8}(x^2 - 6x + 1), & \text{se } 1 \leq x < 3 \\ 1, & \text{se } x \geq 3 \end{cases}$ 10)d) 15/32

11)a) $F(x) = \begin{cases} 0, & \text{se } x < -1, \\ \dfrac{1}{6}(-x^3 + 3x + 2), & \text{se } -1 \leq x < 0 \\ \dfrac{1}{6}(2 + 3x), & \text{se } 0 \leq x < 4/3 \\ 1, & \text{se } x \geq 4/3 \end{cases}$ 11)b) 1/3 11)c) 23/48

12)a) não 12)b) $-3723/450$ 12)c) 689/4

15)a) 2/87 15)b) 149/261

16)a) -1 16)b) 72,16 16)c) 0 16)d) não

17)a) 0 17)b) 14 17)c) 7/3 17)d) 27 17)e) $\approx 17,86$

18)c) 70/42 18)d) 2

19)a) 2/5 19)b) 81/50 19)c) $(4 - \sqrt{6})/2$ 19)d) 0 19)e) 0,2 19)f) 0

20)a) 81/20 20)b) $\left(331 - 12\sqrt{35}\right)/5 \cong 52$ 20)c) não

21) $\dfrac{-7a^3}{(a^3 + 8)}$

22)a) $g(x) = \begin{cases} \frac{1}{4}, & 0 \leq x \leq 4 \\ 0, & \text{c.c.} \end{cases}$ $h(y) = \begin{cases} \frac{1}{4}(3-y), & 0 \leq y \leq 2 \\ 0, & \text{c.c} \end{cases}$ 22)b) Sim

23)a) $\int_0^1 \int_0^1 f(x,y)dxdy = 1$ 23)b) $g(x) = \begin{cases} 1, & \text{se } 0 \leq x \leq 1 \\ 0, & \text{c.c.} \end{cases}$ e $h(y) = \begin{cases} 1, & \text{se } 0 \leq y \leq 1 \\ 0, & \text{c.c.} \end{cases}$

24) $f(x,y) = \begin{cases} \frac{x-1}{2}, & \text{para } 1 < x < 2 \text{ e } 2 < y < 4 \\ \frac{-x+3}{2}, & \text{para } 2 < x < 3 \text{ e } 2 < y < 4 \\ 0, & \text{fora destes intervalos} \end{cases}$

25)a) 0,5 25)b) sim 25)c)

Y	X 0	1	2	3	$P(y)$
10	1/16	3/16	3/16	1/16	8/16
20	1/16	3/16	3/16	1/16	8/16
$P(x)$	2/16	6/16	6/16	2/16	1,00

26)a) 0,5 26)b) $F(x) = \begin{cases} 0, & x < -1 \\ \frac{1}{6}(-x^3+3x+2), & -1 \leq x < 0 \\ \frac{1}{6}(2+3x), & 0 \leq x < 4/3 \\ 1, & 4/3 \leq x \end{cases}$ 26)c) 2/3

27) 5,4

28)a) 1/9 28)b) porque $f(x,y) = g(x)\,h(y)$

29) 0,80

30) $546/216 \approx 2,53$

31)a) $F(x) = \begin{cases} 0, & x < 1 \\ 0,08x, & 1 \leq \lfloor x \rfloor < 6 \\ 0,40 + (x-5)0,09, & 6 \leq \lfloor x \rfloor < 11 \\ 0,85 + (x-10)0,01, & 11 \leq \lfloor x \rfloor \leq 25 \\ 1, & 25 < x \end{cases}$

em que $\lfloor x \rfloor = max\{m \in \mathbb{Z} | m \leq x\}$, isto é, o maior número inteiro que seja menor ou igual a x, 31)b) $F(10) = 0,85$ 31)c) $P(X \geq 20 | X \geq 10) =$ $[1 - F(20) + P(20)] / [1 - F(10) + P(10)] = 0,06/0,24 = 0,25$, portanto 25%.

32)a) 0.50 32)b) 0.50 32)c) são iguais porque as variáveis são independentes, isto é, $P(x,y) = P(x)\,P(y)$ ou $P(x|y) = P(x)$ e $P(y|x) = P(y)$.

33) $E(X) = 40$.

34) $V(Y) = 6$.

35)a) $\frac{21}{64} \approx 0,33$ 35)b) $2E(X) - \frac{1}{6} = 1$ 35)c) $\frac{7}{25} = 0,28$

36)a) $0,9$ 36)b) Não, $P(x_1, x_2) \neq P(x_1)P(x_2)$ 36)c) $4V(X_1)+V(X_2)-4COV(X_1, X_2) \approx 5,5$

37)a) $\frac{5}{33} \approx 0,152$ 37)b) $\frac{163}{352} \approx 0,463$ 37)c) $E(Y) = \frac{279}{132} \approx 2,11$

38)

t	7	9	11	12	14	16	total
$P(t)$	0,05	0,05	0,10	0,10	0,50	0,20	1,00

$E(X) = 4,5$, $V(X) = 1$, $E(Y) = 2,15$, $V(Y) = 0,4275$, $COV(X,Y) = -0,05$, portanto $E(T) = 13,3$ m e $V(T) = 5,31$ m^2.

39) Seja {LPB} a ordem correta dos nomes, então o espaço amostral S pode ser indicado da seguinte forma, $S = \{\text{LPB}, \text{LBP}, \text{PLB}, \text{PBL}, \text{BLP}, \text{BPL}\}$ o que resulta em S_X enumerável dado por, $S_X = \{3, 1, 1, 0, 0, 1\}$ e X uma v.a.d. com a seguinte distribuição:

x	0	1	2	3	total
$P(x)$	2/6	3/6	0	1/6	1,00

40)a) 0,5 40)b) 0.

41) São independentes pois $f(x,y) = g(x)\ h(y)$, em que $g(x) = \frac{3}{7}x^2$ e $h(y) = \frac{1}{6}(y+2)$.

42) $E(X) \approx 20,84$ e $E(Y) \approx 21,02$.

43)

x	0	1	2	3	total
$P(x)$	2/6	3/6	0	1/6	1,00

$E(X) = 1$ e $E(10X - 5) = 5$.

44)a) -19 44)b) 16 44)c) -1

45)a) sim, $P(x,y) = P(x)\,P(y)\ \forall$ par (x,y) 45)b) $= E(X)\,E(Y) = 3,42$ 45)c) 3,0636

46)

$$\begin{aligned} P(X \leq 1,\ Y \leq 2) &= \int_0^1 \int_1^2 f(x,y)\,dy\,dx = P(X \leq 1)\,P(Y \leq 2) \\ &= \int_0^1 g(x)\ dx \int_1^2 h(y)\ dy = 1/4 \times 7/26 \approx 0,07 \end{aligned}$$

se X e Y são v.a.c. independentes.

47)a) Não são v.a. independentes pois $P(x,y) \neq P(x)\,P(y)\ \forall$ par (x,y)

47)b)

defeito	elétrico	mecânico	estético	total
P(defeito|dentro)	0,2083	0,1806	0,6111	1,00

48) $P(X > 1, Y > 1) = \frac{1}{8}\int_1^{+\infty} e^{-y/2}\left(\int_1^{+\infty} xe^{-x/2}dx\right)\ dy = \frac{3}{2}e^{-1} \approx 0,552$

49)a) $1/200$ 49)b) $e^{-300/200} \approx 0,22$ ou 22% 49)c) $e^{-400/200} \approx 0,135$ ou $13,5\%$ 49)d) $F(x) = 1 - e^{-x/200}, 0 \leq x < \infty$ $F(x) = 0, x < 0$, $F(x) = 1, x \to \infty$ 49)e) $\leq -200ln0,9 \approx 21$ horas

DICA: Inicialmente obtenha a fórmula geral para $P(X \geq x)$.

50)a) $k = 1/10$, $F(x) = \begin{cases} 0 & , \quad x < -2 \\ \frac{1}{10}(x+2) & , \quad -2 \leq x < 0 \\ \frac{1}{10}\left(\frac{3}{25}x^2 + x + 2\right) & , \quad 0 \leq x < 5 \\ 1 & , \quad 5 \leq x \end{cases}$

50)b) $[F(3) - F(0)] / [F(3) - F(-1)] = 0,408/0,508 \approx 0,803$, ou pode ser calculado da forma usual, integrando a $f(x)$, $\left(\int_0^3 f(x)\,dx\right) / \left(\int_{-1}^3 f(x)\,dx\right)$.

51)

y	5	7	9	11	Total
$P(y)$	5/32	7/32	9/32	11/32	1,00

$E(Y) = 69/8 = 8,625$.

52) $E(X) = 0 + 0,27 + 0,06 + 0,06 = 0,30$.

53) $k = 3/8$ atende a $\int_0^2 \int_0^x f(x,y)\ dy\ dx = 1$.

54)a) $0,02/0,07 \approx 0,286$ 54)b) $E(W) = 1,225 \approx 1,23$ e $V(W) = 0,406875 \approx 0,41$.

55)a) $t > 0 \implies f(x) \geq 0\ \forall\ x$ e $\int_t^\infty f(x)\,dx = -t(\frac{1}{+\infty} - \frac{1}{t}) = 1$ 55)b) $\frac{t}{t+h}$ 55)c) $3/7$.

56)a) $g(x) = 3(1-x)^2$, se $0 \leq x \leq 1$ e $h(y) = 6y(1-y)$ se $0 \leq y \leq 1$ 56)b) $32/63$ 56)c) $f(x|y) = \frac{1}{y}$, para $0 \leq x \leq y$ 56)d) $f(y|x) = \frac{2(1-y)}{(1-x)^2}$, se $x \leq y \leq 1$ 56)e) $1/4$.

Capítulo 5

Distribuições de variáveis aleatórias

5.1 Introdução

Após uma introdução aos conceitos básicos da teoria das probabilidades e das variáveis aleatórias, o estudo dos principais modelos para o cálculo de probabilidades é uma extensão natural necessária para um melhor entendimento de como as inferências estatísticas são realizadas.

Algumas variáveis aleatórias adaptam-se muito bem a uma série de problemas práticos e aparecem com bastante frequência. Portanto, um estudo pormenorizado das mesmas permite a construção dos correspondentes modelos de probabilidade, bem como a determinação dos seus parâmetros, isto é, dos valores que especificam um modelo ou uma distribuição de probabilidades (normal, χ^2, t, F, Poisson, binomial, etc). Assim, para um dado problema, tenta-se verificar se ele satisfaz às condições de algum modelo de probabilidade conhecido, pois isso facilita muito o trabalho do pesquisador, por permitir caracterizar as variáveis aleatórias por meio do valor médio $E(X)$, da variância $V(X)$ e de probabilidades de interesse.

5.2 Distribuições para variáveis aleatórias discretas

5.2.1 Distribuição Bernoulli

A distribuição Bernoulli, denominação que homenageia Jacques Bernoulli, é utilizada para modelar experimentos dicotômicos, isto é, experimentos aleatórios cujos resultados são do tipo ocorre ou não ocorre ou, como é muitas vezes denominado, sucesso e falha. Nesse caso, a variável vale um se o evento ocorre e vale zero se não ocorre.

Definição 10. Dizemos que uma variável aleatória X tem distribuição de Bernoulli se sua função de probabilidade é dada por

$$f(x) = P(X = x) = p^x(1-p)^{1-x}, x = 0 \text{ ou } 1.$$

A tabela com a distribuição de probabilidades é trivial,

x	0	1	Total
$P(x)$	$1-p$	p	$1,00$

portanto

$$E(X) = \sum_x P(x) = p, \tag{5.1}$$

$$V(X) = E\left(X^2\right) - [E(X)]^2 = p(1-p) = pq, \tag{5.2}$$

sendo que p, a probabilidade de ocorrer um sucesso, é o parâmetro do modelo e, $q = 1 - p$ é a probabilidade de ocorrer um fracasso.

Se uma variável aleatória tem distribuição de Bernoulli, com probabilidade de sucesso p, denotaremos por $X \sim \text{Ber}(p)$. Parâmetro(s) de um modelo de probabilidade refere-se ao(s) valor(es) que precisa(m) ser conhecido(s), para que se tenha o modelo completamente especificado para o cálculo de probabilidades.

5.2.2 Distribuição binomial

Esta distribuição é também conhecida como sequência de Bernoulli, e é considerada a mais importante distribuição de variáveis aleatórias discretas. Vamos procurar caracterizar esta distribuição a partir da seguinte situação: Considere um experimento aleatório consistindo em N tentativas independentes de modo que a probabilidade de ocorrer sucesso em cada uma das tentativas é sempre igual a p. Ocorrer sucesso significa ocorrer algum evento de interesse. Seja X o número de sucessos nas N tentativas, então X pode assumir os valores $0, 1, 2, 3, \cdots, N$. Nestas condições a v.a.d. X tem distribuição Binomial com parâmetros N e p, denotado geralmente como,

$$X \sim \text{Binomial}(N, p).$$

Exemplos: Considere os seguintes experimentos aleatórios (E):

- E_1: N de lançamentos de uma moeda e X=nº de caras;
- E_2: N lançamentos de um dado e X=nº de vezes que ocorre a face 5;
- E_3: Amostragem de N peças de uma linha de produção que produz um percentual $p \times 100\%$ de peças defeituosas. X=nº de peças defeituosas.

O modelo binomial exige que as tentativas sejam independentes e p constante. Esta distribuição é caracterizada por dois parâmetros, p e N, em que $0 < p < 1$ é um parâmetro contínuo e $N > 0$ é discreto (número inteiro), ou seja, para cada combinação de valores para p e N vamos ter uma distribuição específica.

Se $X \sim \text{Binomial}(N, p)$, pode-se demonstrar que:

i) **Função de probabilidade:**

$$P(X = x) = C_N^x \, p^x \, q^{N-x} = \frac{N!}{x!\,(N-x)!} \, p^x \, q^{N-x},$$

para $x = 0, 1, 2, \cdots, N$; $C_N^x = \binom{N}{x} = \dfrac{N!}{x!(N-x)!}$ é a combinação de N elementos x a x; e $p + q = 1$. Um resultado conhecido como o teorema do Binômio de Newton e que pode ser demonstrado por indução matemática é o seguinte:

$$(p+q)^N = \sum_{x=0}^{N} C_N^x \, p^x \, q^{N-x} = 1$$

ii) **Distribuição de probabilidade:**

x_i	0	1	2	$\cdots$	N	Total
$P(X = x_i)$	q^N	$C_N^1 p^1 q^{N-1}$	$C_N^2 p^2 q^{N-2}$	$\cdots$	p^N	1

iii) **Média e variância de** X,

a) Média: $E(X) = \sum\limits_{x} xP(x) = Np$;

b) Variância: $V(X) = E(X^2) - [E(X)]^2 = Npq$.

Os resultados a) e b) podem ser obtidos facilmente por propriedades da variância $V(X)$, e do valor esperado $E(X)$, tomando-se $X = \sum\limits_{i=1}^{N} Y_i$, em que $Y_1, Y_2, \cdots, Y_N$ são variáveis aleatórias discretas independentes e identicamente distribuídas (amostra aleatória) do modelo Bernoulli(p), denotado como $Y_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, $i = 1, 2, \cdots, N$. Este procedimento descrito é exatamente o que será utilizado no exercício 9), conforme será percebido ao efetuar sua resolução.

5.2.2.1 Exercícios propostos com respostas

1) Se 20% dos parafusos produzidos por uma máquina são defeituosos, determinar a probabilidade de, entre 4 parafusos escolhidos ao acaso, no máximo 2 deles serem defeituosos.(R: $0,9728$)

2) Um fabricante garante que uma caixa de suas peças conterá no máximo 2 itens defeituosos. Se a caixa contém 20 peças e a experiência tem demonstrado que esse processo de fabricação produz 2 por cento de itens defeituosos, qual a probabilidade de que uma caixa de suas peças não vá satisfazer a garantia? (R: $0,0071$)

3) De 2.000 famílias com 4 crianças cada uma, quantas se esperaria que tivessem:

a) Pelo menos um menino? ($R: 1875$)

b) Exatamente dois meninos? ($R: 750$)

5.2.3 Distribuição de Poisson

A distribuição de Poisson é conhecida classicamente como a lei dos fenômenos raros. Esta distribuição é útil para descrever as probabilidades do número de ocorrências num campo ou intervalo contínuo (em geral tempo ou espaço). Alguns exemplos de variáveis que podem ter como modelo a distribuição de Poisson são:

- número de defeitos por cm^2;

- número de acidentes por dia;
- número anual de suicídios;
- número de chamadas erradas por hora, num circuito telefônico, etc.

Note-se que a unidade de observação (tempo, área, etc) é contínua mas a variável aleatória (número de ocorrências) é discreta. Pode-se verificar também que as falhas não são contáveis. Por ex.; não é possível contar o número de acidentes que não ocorreram em um dia, nem tão pouco o número de chamadas telefônicas que não foram realizadas. A distribuição de Poisson é uma forma limite da distribuição binomial, quando N tende a infinito e p tende a zero e o termo $Np = m$ permanece constante. Matematicamente, pode-se demonstrar da seguinte forma:

$$\lim_{N\to+\infty} \left(1 - \frac{m}{N}\right)^N = e^{-m} \tag{5.3}$$

e em seguida tome

$$p = \frac{m}{N}, \tag{5.4}$$

então,

$$\lim_{\substack{N\to+\infty \\ p=\frac{m}{N}}} \binom{N}{x} p^x (1-p)^{N-x} \overset{(5.4)}{=} \lim_{\substack{N\to+\infty \\ p=\frac{m}{N}}} \left[\frac{1}{x!} \underbrace{\frac{N}{N} \frac{(N-1)}{N} \frac{(N-2)}{N} \cdots \frac{[N-(x-1)]}{N}}_{x \text{ termos, cada um tendendo a um}} \frac{(N-x)!}{(N-x)!} \right.$$

$$\left. \times \quad m^x \times \underbrace{\left(1 - \frac{m}{N}\right)^N}_{\text{converge a } e^{-m} \text{ por (5.3)}} \times \underbrace{\left(1 - \frac{m}{N}\right)^{-x}}_{\text{converge para um}} \right] = \frac{e^{-m} m^x}{x!}$$

Na prática, a distribuição de Poisson é utilizada para aproximar a distribuição binomial quando o número de observações de um experimento aleatório é muito grande (ex: $N \geq 50$) e a probabilidade de sucesso é muito pequena (ex: $p \leq 0,1$). A distribuição de Poisson fica completamente caracterizada pelo parâmetro m, que representa o número esperado de ocorrências do evento por unidade de observação do processo. Se uma variável aleatória X tem os valores distribuídos segundo o modelo Poisson, utiliza-se a seguinte notação:

$$X \sim \text{Poisson}(m) \quad \text{em que} \quad m > 0.$$

Pode-se demonstrar que:

i) **Função de probabilidade:**

$P(X = x) = \dfrac{e^{-m} m^x}{x!}$, para $x = 0, 1, 2, \cdots$, em que: X é a variável aleatória discreta que representa o número de ocorrências por unidade contínua de observação, e=base do logaritmo neperiano ($e \approx 2,718$).

ii) **Média e variância:** $m = E(X) = V(X)$.

5.2.3.1 Exercícios propostos com respostas

1) Num livro de 800 páginas há 800 erros de impressão.

a) Qual a probabilidade de que uma página contenha pelo menos 3 erros de impressão? (R:$\approx 0,0803$)

b) Estime o número provável de páginas por livro que não contêm erros de impressão. (R:≈ 294)

2) Numa indústria há uma média de 3 acidentes por mês.

a) Qual a probabilidade de ocorrerem 2 acidentes no próximo mês? (R:$\approx 22,4\%$)

b) Qual a probabilidade de ocorrerem 10 acidentes nos próximos 6 meses? (R:$\approx 0,015$)

3) Demonstre que no modelo Poisson há a seguinte fórmula de recorrência:

$$P(x+1) = P(x) \cdot \frac{m}{x+1}.$$

5.3 A distribuição normal para variáveis aleatórias contínuas

A distribuição normal é uma das mais importantes distribuições de probabilidades, sendo aplicada em inúmeros fenômenos e constantemente utilizada para o desenvolvimento teórico da inferência estatística. É também conhecida como distribuição de Gauss, Laplace ou Laplace-Gauss.

Dizemos que uma v.a.c. X tem distribuição normal, ou é normalmente distribuída, se possuir a seguinte f.d.p.:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \text{ para } -\infty < \mu < +\infty, -\infty < x < +\infty \text{ e } \sigma > 0.$$

Notação: $X \sim N(\mu, \sigma^2)$.

Se X tem distribuição normal, então a média e a variância são os dois parâmetros que especificam a distribuição. Pode-se demonstrar que:

1) A média de X, é dada por

$$E(X) = \int_{-\infty}^{+\infty} x\ f(x)\ dx = \mu;$$

2) A variância de X é dada por

$$V(X) = E(X^2) - [E(X)]^2 = \sigma^2.$$

A importância do modelo normal para a Ciência Estatística, dentre outros, deve-se à importantes resultados que permitem a realização (é a base teórica) de diversos procedimentos inferenciais, tais como o teste Z para uma média (capítulo 7, seção 7.4.1); o conhecido Teorema Central do Limite (TCL), enunciado da seguinte forma simplificada a seguir.

Teorema 6 (TCL)

Médias amostrais para amostras de tamanho n, $(\bar{X}_n)$ são variáveis aleatórias normalmente distribuídas com média igual à média da população (μ) e variância igual à variância da população dividida por n, $\left(\frac{\sigma^2}{n}\right)$, $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

i) **Representação gráfica:**

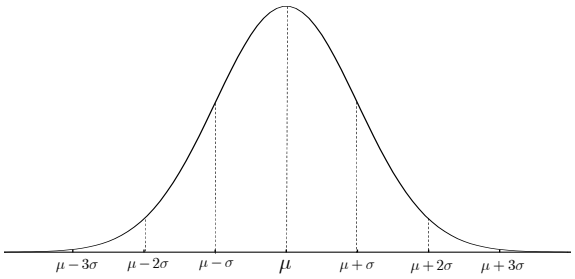


Figura 5.1: Representação gráfica da distribuição normal com média μ e variância σ^2.

É um gráfico em forma de sino (Figura 5.1). O seu posicionamento em relação ao eixo das ordenadas e seu achatamento serão determinados pelos parâmetros μ e σ, respectivamente.

A função de distribuição acumulada é dada por:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du.$$

ii) **Propriedades:**

1) $f(x)$ possui um ponto de máximo para $x = \mu$.

2) $f(x)$ tem dois pontos de inflexão cujas abscissas valem $\mu+\sigma$ e $\mu-\sigma$.

3) $f(x)$ é simétrica em relação a $x = \mu$. E ainda, μ=Mo=Md.

4) $f(x)$ tende a zero quando x tende para $\pm\infty$ (assintótica em relação ao eixo x).

5) Para quaisquer valores σ e μ finitos, tem-se que,

i) $P\left(\mu - \sigma \leq X \leq \mu + \sigma\right) \approx 0,6827$

ii) $P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) \approx 0,9545$

iii) $P\left(\mu - 3\sigma \leq X \leq \mu + 3\sigma\right) \approx 0,9973.$

Portanto, de 5)ii) verifica-se que em uma distribuição normal, aproximadamente 95% das observações estão distantes da média em no máximo dois desvios padrões.

iii) **Cálculo de probabilidades:**

Para o cálculo da probabilidade da v.a.c. assumir um valor em determinado intervalo, surgem dois problemas:

1º) Integração de $f(x)$, pois para o seu cálculo é necessário o desenvolvimento em séries.

2º) Elaboração de tabelas do tipo $P(X \leq x) = \int_{-\infty}^{x} f(u)du$ se torna inviável, pois a f.d.p. depende de dois parâmetros, μ e σ^2, o que acarreta a necessidade do estabelecimento de todas as possíveis combinações de valores desses parâmetros.

Estes problemas são resolvidos pela padronização dos valores, obtendo-se assim a distribuição normal padronizada ou reduzida.

5.3.1 Variável normal padronizada (Z)

É obtida por meio de uma transformação linear da variável normal X, obtendo-se assim uma escala relativa de valores na qual a média é tomada como ponto de referência e o desvio padrão como medida de afastamento da média:

$$Z = \frac{X - \mu}{\sigma} \text{ ou } z = \frac{x - \mu}{\sigma},$$

em que:
z=valor da variável normal padronizada Z,
x=valor de X,
μ=média de X,
σ=desvio padrão de X.

5.3.1.1 Média e variância da variável normal padronizada

i) Média:

$$E(Z) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}\left[E(X - \mu)\right] = \frac{1}{\sigma}\left[E\left(X\right) - E(\mu)\right],$$

$$E(Z) = \frac{1}{\sigma}(\mu - \mu) = 0.$$

ii) Variância:

$V(Z) = V\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}V\left(X\right)$, já que μ é uma constante. Vide seções 4.8.1.1 e 4.8.2.1.

$$V(Z) = \frac{1}{\sigma^2}\left(\sigma^2\right) = 1.$$

Conclusão: Se $X \sim N\left(\mu, \sigma^2\right)$, e $Z = \dfrac{X - \mu}{\sigma}$, então $Z \sim N\left(0, 1\right)$, para quaisquer valores de μ e σ^2. Portanto, será possível tabelar as probabilidades, $P(X \leq x) = P(Z \leq z)$, em função dos valores de Z.

A f.d.p. da variável Z é dada por:

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}, \text{ para } -\infty < z < +\infty.$$

5.3.2 Tabela da distribuição normal padrão

Há vários tipos de tabelas que nos fornecem as probabilidades sob a curva normal. A tabela que vamos utilizar é aquela que fornece a probabilidade da variável Z assumir um valor entre zero e um particular z_0, ou seja:

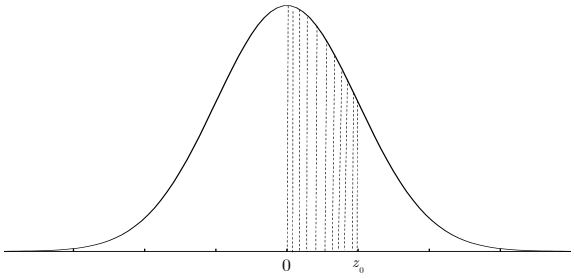


Figura 5.2: Tabela da Distribuição Normal Padrão.

$$P(0 \leq Z \leq z_0) = \int_0^{z_0} \phi(z)dz = \frac{1}{\sqrt{2\pi}} \int_0^{z_0} e^{-\frac{1}{2}z^2} dz,$$

é a área hachurada sob a curva normal $[\Phi(Z)]$.

5.3.3 Exercícios propostos com respostas

1) Calcule:

a) $P(Z \leq 1,82)$

b) $P(Z \leq -2,03)$

c) $P(-2,55 \leq Z \leq 1,20)$

d) $P(Z \geq 1,93)$

R: a) $0,9656$ b) $0,0212$ c) $0,8795$ d) $0,0268$

2) Se $X \sim \mathrm{N}(100, 25)$, calcule:

a) $P(X > 110)$

b) $P(95 \leq X \leq 105)$

c) Encontre x tal que $P(X \leq x) = 0,3446$

R: a) 0,0228 b) 0,6826 c) 98

5.3.4 Teorema da combinação linear

A combinação linear de variáveis normais independentes é também uma variável normal. Assim se X e Y são variáveis normais independentes, então $W = aX + bY + c$ é também uma variável normal com média $\mu_W = a\mu_X + b\mu_Y + c$ e variância $\sigma^2_W = a^2\sigma^2_X + b^2\sigma^2_Y$, em que a, b e c são constantes. Em particular, devemos notar que a soma ou subtração de duas ou mais variáveis aleatórias normais, também é uma variável normalmente distribuída.

Exemplo 35

Sejam X e Y variáveis aleatórias normais e independentes com $\mu_X = 12$, $\mu_Y = 6$, $\sigma^2_X = 4$ e $\sigma^2_Y = 25$. Para $W = 3X - 6Y + 10$ pede-se obter μ_W e σ_W. (R: $\mu_W = 10$ e $\sigma_W = \sqrt{936}$)

5.3.5 Exercícios propostos com respostas

1) Um pesquisador decidiu que, para facilitar a classificação das aves em experimentos de nutrição, deve-se dividir as poedeiras, no início da postura, em três grupos de peso equiprováveis, a saber: poedeiras pesadas, poedeiras médias e poedeiras leves. Encontre os pesos correspondentes à cada classe, sabendo-se que o peso médio das aves nessa idade é 1,5 kg, com desvio padrão de 0,170 kg (supor distribuição normal).
(R: Leves: peso $< 1,43$ kg; médias: $1,43 \leq$ peso $\leq 1,57$ kg; pesadas: peso $> 1,57$ kg)

2) O diâmetro de um cabo elétrico é normalmente distribuído com média 0,8 mm e variância $0,0004$. Dentre uma amostra de 1000 cabos, quantos esperamos que tenha diâmetro:

a) maior ou igual a $0,81$ mm.

b) entre $0,73$ e $0,86$ mm.

c) menor que $0,78$ mm.

(R: a) $308,5$ b) $998,7$ c) $158,7$)

3) Em uma distribuição normal 28% dos elementos são superiores a 34 e 12% dos elementos são inferiores a 19. Encontrar a média e a variância da distribuição. (R: $\mu \cong 29,03$ e $\sigma^2 \cong 73,4$)

4) X é uma v.a.c. tal que $X \sim$N$(12, 25)$. Qual a probabilidade de uma observação ao acaso ser menor do que $-2,5$? (R: $0,0019$).

5.4 Exercícios propostos com respostas

1) Determine a probabilidade de que, em 5 lançamentos de um dado, apareça a face 3:

a) Duas vezes;

b) No máximo uma vez;

c) Ao menos duas vezes.

2) Quantas vezes se deverá jogar um dado honesto para que se tenha a probabilidade igual a $0,5$, de ocorrer a face 3, pelo menos uma vez?

3) Considere a amostragem de 3 peças que saem de uma linha de produção. Sabe-se que são produzidas 20% de peças defeituosas. Calcule as seguintes probabilidades:

a) 2 peças defeituosas;

b) 2 peças não defeituosas;

c) Quantas peças defeituosas espera-se amostrar, considerando 500 peças?

4) Sabe-se que 24% dos indivíduos que recebem o medicamento X sofrem certos efeitos colaterais. Se o medicamento X for ministrado a quatro pacientes, qual a probabilidade de que:

a) Nenhum sofra efeitos colaterais;

b) Pelo menos um sofra efeitos colaterais;

c) Três não sofram efeitos colaterais.

5) Uma firma determina o sexo de pintos de um dia com 95%de probabilidade.

a) Se comprarmos 5 pintinhos tidos como do sexo feminino, qual é a probabilidade de que pelo menos um seja macho?

b) Quantos machos espera-se encontrar num lote de 500 pintinhos tidos como do sexo feminino?

6) Numa prova com 10 questões de múltipla escolha, cada uma com 5 alternativas e somente uma correta, pede-se:

a) Quantas questões acerta, em média, um aluno que assinala todas as questões inteiramente ao acaso?

b) Qual a probabilidade dele acertar 5 questões?

7) Num teste do tipo certo-errado, com 100 questões, qual a probabilidade de um aluno, respondendo as questões ao acaso, acertar 70% das questões?

8) Se $X \sim$ Binomial $(16; 0,75)$ determine:

a) A média de X;

b) A variância de X;

c) Se $Z = \dfrac{X - 12}{\sqrt{3}}$, calcule $E(Z)$ e $V(Z)$.

9) Seja $I(x_i) = 1$ se a i-ésima tentativa é um sucesso e $I(x_i) = 0$ caso contrário. Admita que $P\left[I(x_i) = 1\right] = p \ \ \forall \ \ i$ e que $Y = \sum_{i=1}^{N} I(x_i)$. Pede-se:

a) Obtenha $E(Y)$ e $V(Y)$;

b) Qual é a distribuição de Y?

10) Um processo de fabricação de fitas magnéticas produz, em média, fitas com um defeito a cada 200 m de rolo. Qual a probabilidade de que:

a) Em 500 m de fita não ocorra nenhum defeito?

b) Em 800 m de fita ocorram pelo menos 3 defeitos?

11) A experiência mostra que de cada 400 lâmpadas, 2 se queimam ao serem ligadas. Qual a probabilidade de que numa instalação de:

a) 600 lâmpadas, no mínimo 3 se queimem?

b) 900 lâmpadas, exatamente 8 se queimem?

12) Na pintura de paredes aparecem defeitos na proporção média de um defeito por metro quadrado. Qual a probabilidade de aparecerem 3 defeitos numa parede de 2 x 2 m?

13) Numa central telefônica são atendidas 300 chamadas por hora. Qual a probabilidade de:

a) Serem atendidas duas chamadas num período de 2 minutos?

b) Em T minutos, não ocorrerem chamadas telefônicas?

14) Estima-se em 1% a proporção de canhotos numa população. Qual a probabilidade de termos pelo menos um canhoto numa classe de 30 alunos?

15) Na revisão tipográfica de um livro acharam-se, em média, $1,5$ erros por página. Das 800 páginas do livro, estime quantas não apresentam erros?

16) O departamento de trânsito registrou num certo ano, numa determinada via pública, 30 acidentes fatais, com um movimento médio diário de 200 veículos. Qual é a probabilidade de que num determinado mês, do próximo ano, ocorram 3 acidentes fatais?

17) Seja X o número de crianças não imunizadas numa campanha de vacinação contra uma determinada doença, em que a probabilidade de não imunização é $0,001$. De 5000 crianças vacinadas, qual a probabilidade de não ficarem imunes:

a) Uma criança?

b) Pelo menos uma criança?

18) Na fabricação de peças de determinado tecido aparecem defeitos ao acaso, um a cada 250 m.

a) Qual a probabilidade de que não haja defeitos na produção de 1000 m de tecido?

b) Se a produção diária é de 625 m, num período de 80 dias de trabalho, em quantos desses dias poderemos esperar uma produção diária na qual não haja defeitos?

19) As notas de uma prova são normalmente distribuídas com média 73 e variância 225. Os 15% melhores alunos recebem o conceito A e os $11,9\%$ piores alunos recebem conceito R. Pede-se:

a) Nota mínima para receber A?

b) Nota mínima para ser aprovado?

c) $P\left(X \geq 55,3\right)$

20) Se $X \sim \mathrm{N}(3;4)$ encontre um valor x tal que:$P\left(X \geq x\right) = 2\,P\left(X \leq x\right)$.

21) A observação dos pesos X, de um grande número de espigas de milho, mostrou que essa variável é normalmente distribuída com média $\mu = 120$ g e desvio padrão $\sigma = 10$ g. Num programa de melhoramento genético da cultura do milho, entre outras características, uma linhagem deve satisfazer à condição $112 < X < 140$. Num programa envolvendo 450 linhagens, qual deve ser o número provável de linhagens que atende à essa condição?

22) Sabe-se que o peso médio, em arrobas, de abate de bovinos é normalmente distribuído com média 18 e variância $2,25$. Um lote de 5000 cabeças, com essa característica, foi destinado ao frigorífico que abate só a partir de um peso mínimo W. Sabendo-se que foram abatidas 4200 cabeças, pede-se:

a) O número esperado de bovinos com peso entre 17 e 19 arrobas?

b) Qual o valor de W?

23) O volume de correspondência recebido por uma firma quinzenalmente é normalmente distribuído com média de 4000 cartas e desvio padrão de 200 cartas. Qual a porcentagem de quinzenas em que a firma recebe menos de 3400 cartas?

24) O peso médio de um cigarro é a soma dos pesos do papel e do fumo, e vale em média $1,200$ g com desvio padrão $0,060$ g. O peso médio do papel é $0,040$ g com desvio padrão $0,020$ g. Os cigarros são feitos em uma máquina automática que pesa o fumo a ser colocado no cigarro, coloca-o no papel e enrola o cigarro, portanto admita que os pesos do papel e do fumo são independentes. Pede-se:

a) Determinar o peso médio do fumo em cada cigarro e o desvio padrão.

b) Qual a probabilidade de que um cigarro tenha menos de 1,130g de fumo?

25) Numa indústria a montagem de um certo item é feita em duas etapas. Os tempos necessários para cada etapa são independentes e têm as seguintes distribuições:

$$X_1 : N\left(75 \text{ seg}; 16 \text{ seg}^2\right), X_1 \text{ tempo da } 1^{\text{a}} \text{ etapa}$$

$$X_2 : N\left(125 \text{ seg}; 100 \text{ seg}^2\right), X_2 \text{ tempo da } 2^{\text{a}} \text{ etapa}$$

Qual a probabilidade de que sejam necessários para montar a peça:

a) mais de 210 seg?

b) menos de 180 seg?

26) Suponha que X, a carga de ruptura de um cabo (kg), tenha distribuição N$(100; 16)$. Cada rolo de 100 m de cabo dá um lucro de 25 u.m., desde que $X > 95$. Se $X \leq 95$, o cabo poderá ser utilizado para uma finalidade diferente, a um lucro de 10 u.m. por rolo. Determine o lucro esperado por rolo?

27) Um avião de turismo de 4 lugares pode levar uma carga útil de 350 kg. Supondo que os passageiros têm peso normalmente distribuído com média de 70 kg e desvio padrão de 20 kg e que a bagagem de cada passageiro também é normalmente distribuída com média 12 kg e desvio padrão de 5 kg. Calcule a probabilidade de:

a) Haver sobrecarga se o piloto não pesar os passageiros e respectivas bagagens?

b) Que o piloto tenha que retirar pelo menos 50 kg de gasolina para evitar sobrecarga?

Respostas dos exercícios propostos

1)a) 625/3888 1)b) 3125/3888 1)c) 763/3888

2) 4 vezes

3)a) 0,096 3)b) 0,384 3)c) 100

4)a) 0,3336 4)b) 0,6664 4)c) 0,4214

5)a) 0,2263 5)b) 25

6)a) 2 6)b) 0,0264

7) $\left(\begin{array}{c} 100 \\ 70 \end{array} \right) (0,5)^{100} \cong 2,32 \ . \ 10^{-5}$

8)a) 12 8)b) 3 8)c) 0 e 1

9)a) $E(Y) = Np$ e $V(Y) = Np(1-p)$, 9)b) Binomial

10)a) 0,0821 10)b) 0,7619

11)a) 0,5768 11)b) 0,0463

12) 0,1953

13)a) 0,00227 13)b) e^{-5T}

14) 0,2592

15) 179 páginas

16) 0,2138

17)a) 0,0337 17)b) 0,9933

18)a) 0,0183 18)b) 6,57 dias

19)a) 88,6 19)b) 55,3 19)c) 0,8810

20) 2,14

21) 345

22)a) 2486 22)b) 16,52

23) 0,13%

24)a) 1,160 g e 0,0566 g 24)b) 0,2981

25)a) 0,1762 25)b) 0,0314

26) 23,42 u.m.

27)a) 0,2981 27)b) 0,0401

Capítulo 6

Regressão linear simples

6.1 Introdução

Uma equação de regressão linear simples permite determinar, a partir das estimativas dos parâmetros, como uma variável independente (X) exerce, ou parece exercer, influência sobre outra variável (Y), chamada de variável dependente. Por exemplo, qual a influência do diâmetro à altura do peito (DAP) sobre o volume de árvores de Eucalipto? Esta pergunta poderia ser respondida a partir de uma regressão linear simples entre as variáveis Y (volume das árvores) e X (DAP das árvores). Logicamente, quanto maior o diâmetro, maior o volume, entretanto, é necessário determinar em que proporção isto ocorre e qual o modelo estatístico mais apropriado.

O problema básico da teoria da regressão consiste em: (a) estimar os parâmetros do modelo estatístico admitido; (b) deduzir testes de significância para esses parâmetros, e (c) calcular intervalos de confiança para esses parâmetros, com base na equação obtida. Neste texto será abordado apenas o primeiro item com algumas complementações.

6.2 O modelo estatístico

Dados n pares de valores de duas variáveis, X_i e Y_i, com $i = 1, 2, \ldots, n$, se admitirmos que Y é função linear de X, podemos estabelecer uma regressão linear simples, cujo modelo estatístico é $Y_i = \beta_0 + \beta_1 X_i + e_i$, em que β_0 e β_1 são os parâmetros do modelo, e e_i são os erros aleatórios.

O coeficiente angular da reta (β_1) é também chamado de coeficiente de regressão e o coeficiente linear da reta (β_0) é também conhecido como intercepto sendo o termo constante da equação de regressão.

Ao estabelecermos o modelo de regressão linear simples pressupomos que:

a) a relação entre X e Y é linear;

b) os valores de X são fixos, isto é, X não é uma variável aleatória;

c) a média dos erros é nula, isto é, $E(e_i) = 0$;

d) para um dado valor de X, a variância do erro é sempre σ^2, denominada de variância residual, isto é, $V(e_i) = \sigma^2$; dizemos que os erros são homocedásticos;

e) o erro de uma observação é independente do erro em outra observação, isto é, $E(e_i e_j) = 0$ para $i \neq j$;

f) os erros têm distribuição normal.

Combinado as pressuposições *c), d), e), f)* temos que $e_i \sim \text{NID}(0, \sigma^2)$, em que NID significa Normal e Independentemente Distribuído.

Devemos, ainda, verificar se o número de observações disponíveis é maior que o número de parâmetros do modelo de regressão. Para ajustarmos uma regressão linear simples, devemos ter, no mínimo, 3 observações. Se disporemos de apenas duas observações, teremos um problema de geometria analítica, não sendo possível nesse caso, fazer nenhuma análise estatística.

As pressuposições a, b e c, permitem escrever que $E(Y_i) = \beta_0 + \beta_1 X_i$, ou seja, as médias condicionais de Y dado X, isto é, $E(Y|X)$, estão sobre a reta $\beta_0 + \beta_1 X_i$.

A pressuposição d corresponde ao fato de que as distribuições de Y para diferentes valores de X apresentam todas a mesma dispersão. Graficamente temos:

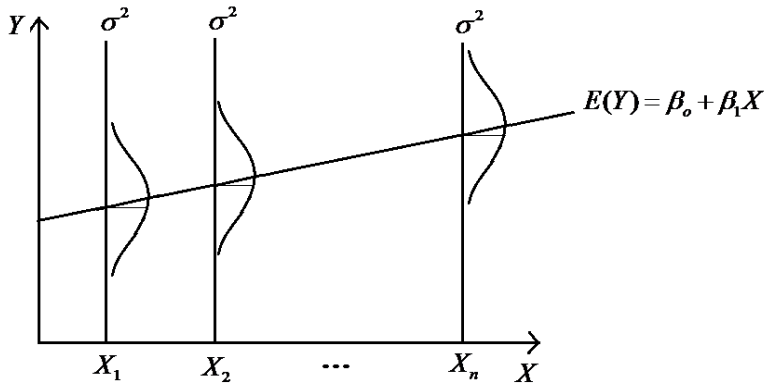


Figura 6.1: Representação gráfica da homogeneidade de variâncias.

A pressuposição f é necessária para que possamos utilizar as distribuições t e/ou F para testar hipóteses a respeito dos valores dos parâmetros, ou construir intervalos de confiança.

6.3 Estimadores dos parâmetros

O primeiro passo na análise de regressão linear simples (RLS) é obter as estimativas dos parâmetros β_0 e β_1. Essas estimativas são obtidas a partir de uma amostra de tamanho n, isto é, a partir de n pares X_i, Y_i. Graficamente temos:

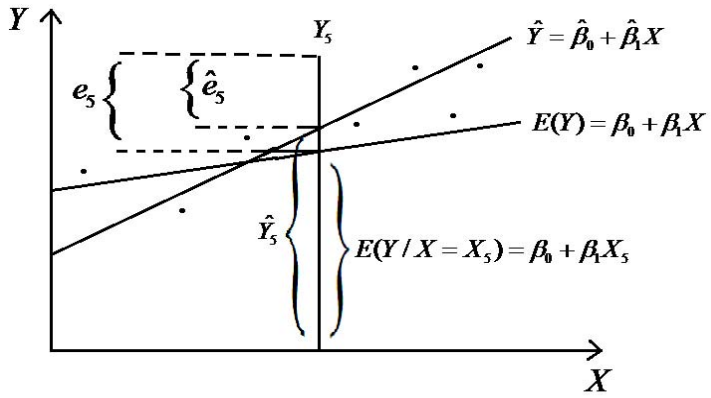


Figura 6.2: Interpretação geométrica de parâmetros e estimadores na RLS.

Y_i é o i-ésimo valor observado,
$\hat{Y}_i$ é o i-ésimo valor estimado (valor médio),
$Y_i = E[Y|X = X_i] + e_i \Rightarrow e_i = Y_i - \beta_0 - \beta_1 Xi$ é o i-ésimo erro aleatório,
$Y_i = \hat{Y}_i + \hat{e}_i \Rightarrow \hat{e}_i = Y_i - \hat{Y}_i$ é o i-ésimo desvio ou resíduo da regressão.

6.3.1 O método dos mínimos quadrados (MMQ)

O método usual para a obtenção das estimativas dos parâmetros de um modelo de regressão é o Método dos Mínimos Quadrados (MMQ). Este método consiste em adotar como estimativas dos parâmetros os valores que minimizam a soma dos quadrados dos erros aleatórios.

Sejam:

$$e_i = Y_i - \beta_0 - \beta_1 X_i$$

$$Z = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

A função Z terá mínimo quando suas derivadas parciais em relação a β_0 e a β_1 forem nulas (Observe que Z não tem máximo, por ser uma soma dos quadrados). Então,

$$\frac{\partial Z}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Z}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Assim, as estimativas de β_0 e β_1 são dadas por:

$$\begin{cases} \sum\limits_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \\ \sum\limits_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \end{cases} \tag{6.1}$$

A partir de sistema (6.1), podemos escrever o seguinte sistema de equações normais:

$$\begin{cases} \hat{\beta}_0 n + \hat{\beta}_1 \sum\limits_{i=1}^{n} X_i = \sum\limits_{i=1}^{n} Y_i \\ \hat{\beta}_0 \sum\limits_{i=1}^{n} X_i + \hat{\beta}_1 \sum\limits_{i=1}^{n} X_i^2 = \sum\limits_{i=1}^{n} X_i Y_i \end{cases} \tag{6.2}$$

Resolvendo este sistema de equações normais obtém-se:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} X_i Y_i - \frac{\left(\sum\limits_{i=1}^{n} X_i\right)\left(\sum\limits_{i=1}^{n} Y_i\right)}{n}}{\sum\limits_{i=1}^{n} X_i^2 - \frac{\left(\sum\limits_{i=1}^{n} X_i\right)^2}{n}} = \frac{SPD_{XY}}{SQD_X}$$

Na prática, determina-se $\hat{\beta}_1$ em primeiro lugar e depois $\hat{\beta}_0$.

A estimativa do coeficiente de regressão $\hat{\beta}_1$ mede o quanto muda na variável dependente $\hat{Y}$ por mudança unitária na variável independente X.

Existem duas relações bastante úteis que podem ser obtidas a partir do sistema (6.1):
Lembrando que $Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) = Y_i - \hat{Y}_i = \hat{e}_i$, então:

$$\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \;\Rightarrow\; \sum_{i=1}^{n} \hat{e}_i = 0 \tag{6.3}$$

$$\sum_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) = 0 \;\Rightarrow\; \sum_{i=1}^{n} X_i \hat{e}_i = 0 \tag{6.4}$$

Temos também que:

$$\sum_{i=1}^{n} \hat{Y}_i \hat{e}_i = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) \hat{e}_i = \hat{\beta}_0 \sum_{i}^{n} \hat{e}_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i \hat{e}_i$$

De acordo com (6.3) e (6.4), concluímos que:

$$\sum_{i=1}^{n} \hat{Y}_i \hat{e}_i = 0 \tag{6.5}$$

As relações (6.3), (6.4) e (6.5) mostram, respectivamente, que:

1. a soma dos desvios é nula;

2. a soma dos produtos dos desvios pelos correspondentes valores da variável independente é igual a zero; e

3. a soma dos produtos dos desvios pelos respectivos valores estimados da variável dependente é igual a zero.

Estas relações podem ser utilizadas para verificar se as estimativas dos parâmetros foram corretamente calculadas e verificar o efeito dos erros de arredondamento.

Como $Y_i = \hat{Y}_i + \hat{e}_i$ de (6.3), concluímos que:

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i=1}^{n} \hat{Y}_i}{n} = \bar{Y}$$

ou seja, a média dos valores observados de Y é igual a média dos valores estimados de Y. Isto ocorre em modelos de regressão linear que incluem a constante $\hat{\beta}_0$.

Para o modelo $Y_i = \beta_0 + \beta_1 X_i + \hat{e}_i$ com $i = 1, 2, \ldots n$, uma estimativa da variância residual é dada por:

$$\widehat{\sigma}^2 = S_e^2 = \frac{\text{SQTotal} - \text{SQRegressão}}{n-2} = \frac{\text{SQResíduo}}{n-2} = \text{QMResíduo}$$

em que,

$$SQTotal = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}$$

$$\text{SQRegressão} = \frac{\left[\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}\right]^2}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{(SPD_{XY})^2}{SQD_X} = \hat{\beta}_1 SPD_{XY}$$

6.4 Análise de variância

O esquema da análise de variância da regressão linear simples considerando-se o modelo $Y_i = \beta_0 + \beta_1 X_i + e_i$, é apresentado na Tabela a seguir:

Fontes de Variação	GL	SQ	QM	F
Regressão	1	SQRegressão	QMRegressão=V_1	V_1/V_2
Resíduo	$n-2$	SQResíduo	QMResíduo=V_2	
Total	$n-1$	SQTotal		

Em que, GL é o número de graus de liberdade e $\text{QMRegressão} = \frac{\text{SQRegressão}}{1} = \text{SQRegressão}$. Logo, no caso de uma RLS, QMRegressão=SQRegressão. E, ainda, $\text{QMResíduo} = \frac{\text{SQResíduo}}{n-2}$.

A estatística $F = \dfrac{V_1}{V_2}$, na Tabela anterior, testa a hipótese $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. Sob H_0 verdadeira, esta estatística F tem distribuição F central com 1 e $n-2$ graus de liberdade.

Regra de Decisão: Se F calculado for $\geq F$ tabelado, rejeita-se H_0 ao nível de significância α. Neste caso, diz-se que o resultado é significativo ($p < \alpha$), e a regressão linear existe. Caso contrário, não se rejeita H_0, e então, o resultado é não significativo ($p > \alpha$).

6.5 O coeficiente de determinação simples (r^2)

O coeficiente de determinação de uma regressão linear simples, denotado por r^2 e expresso em porcentagem, é dado por:

$$r^2 = \frac{SQ_{\text{Regressão}}}{SQTotal} 100, 0 \leq r^2 \leq 100\%$$

O r^2 indica a proporção da variação de Y que é "explicada" pela regressão, ou quanto da $SQTotal$ está sendo "explicada" pela regressão, ou quanto da variação na variável dependente Y está sendo "explicada" pela variável independente X. Quanto maior for o r^2, melhor. Além do coeficiente de determinação, outros critérios devem ser adotados na escolha de modelos.

Exemplo de Aplicação:

Sejam os dados a seguir:

X	4	7	10	12	17
Y	10	16	20	24	30

Admite-se que as variáveis X e Y estão relacionadas de acordo com o modelo $Y_i = \beta_0 + \beta_1 X_i + e_i$.

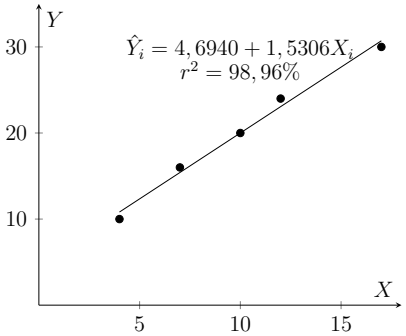
a) Determine as estimativas dos parâmetros da regressão linear e trace o gráfico.

$$n = 5, \sum_{i=1}^{n} X_i = 50, \sum_{i=1}^{n} X_i^2 = 598, \sum_{i=1}^{n} Y_i = 100, \sum_{i=1}^{n} Y_i^2 = 2232, \sum_{i=1}^{n} X_i Y_i = 1150$$

$$\hat{\beta}_1 = \frac{SPD_{XY}}{SQD_{XY}} = \frac{150}{98} = 1,5306$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{100}{5} - 1,5306\frac{50}{5} = 4,6940$$

A equação ajustada é: $\hat{Y}_i = 4,6940 + 1,5306 X_i$.



FATO: Fazer uma extrapolação significa utilizar a equação ajustada para prever valores fora do intervalo coberto pela amostra. Frequentemente a equação ajustada é razoável para o intervalo coberto pela amostra, mas é absolutamente inapropriada para fazer uma extrapolação.

b) Calcule o coeficiente de determinação e interprete.

$$SQTotal = 2232 - \frac{(100)^2}{5} = 232$$

$$SQ\text{Regressão} = \frac{(150)^2}{98} = 229,5918$$

$$r^2 = \frac{229,5918}{232}100 = 98,96\%$$

Interpretação: $98,96\%$ da variação observada em Y está sendo "explicada" pela regressão linear ajustada.

c) Interprete a estimativa obtida para $\hat{\beta}_1$.

$\hat{\beta}_1 = 1,5306$. Assim, para um aumento de uma unidade em X tem-se um acréscimo de $1,5306$ em Y, ou melhor, para um aumento de uma unidade em X, estima-se um aumento médio de $1,5306$ na variável dependente Y.

d) Determine a estimativa de Y para $X = 9$.

$\hat{Y} = 4,6940 + 1,5306(9) \cong 18,47$.

e) Faça a análise de variância da regressão, tomando $\alpha = 1\%$ no teste F.

O resultado da análise é dado a seguir:

FV	GL	SQ	QM	F
Regressão	1	229,5918	229,5918	286,02**
Resíduo	3	2,4082	0,8027	
Total	4	232,0000		

** Significativo ao nível de 1% de probabilidade $F_{1\%}(1;3) = 34,12$.

Com estes resultados pode-se concluir que a hipótese $H_0 : \beta_1 = 0$ foi rejeitada ($p < 0,01$). Logo, a regressão linear existe.

Exercício proposto

1) Os dados a seguir provêm de um experimento para testar o desempenho de uma máquina industrial. O experimento utilizou uma mistura de óleo diesel e gás, derivados de materiais destilados orgânicos. O valor da capacidade da máquina em cavalo vapor (HP) foi coletado a diversas velocidades medidas em rotações por minuto ($rpm \times 100$).

X	Y	X	Y	X	Y	X	Y
22,0	64,03	15,0	46,85	18,0	52,90	15,0	45,79
20,0	62,47	17,0	51,17	16,0	48,84	17,0	51,17
18,0	54,94	19,0	58,00	14,0	42,74	19,0	56,65
16,0	48,84	21,0	63,21	12,0	36,63	21,0	62,61
14,0	43,73	22,0	64,03	10,5	32,05	23,0	65,31
12,0	37,48	20,0	62,63	13,0	39,68	24,0	63,89

X=Velocidade Y=capacidade

Admitindo-se que as variáveis X e Y estão relacionadas de acordo com o modelo

$Y_i = \beta_0 + \beta_1 X_i + e_i$ pede-se:

a) Obter a equação ajustada e traçar seu gráfico. Mostre também o diagrama de dispersão;

b) Calcule o coeficiente de determinação e interprete;

c) Verifique que $\sum_{i=1}^{n} \hat{e}_i = 0$;

d) Verifique que $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$;

e) Interprete a estimativa obtida para β_1;

f) Determine a estimativa de Y para $X = 15,5$;

g) Faça a análise de variância da regressão e no teste F adotar $\alpha = 5\%$.

Nota: Os itens (c) e (d) são sempre verdadeiros para os modelos que incluem o termo constante β_0 (intercepto).

6.6 Exercícios propostos com respostas

1) Em regressão linear simples utiliza-se o modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ que após ajustado pelo método dos mínimos quadrados (MMQ) é representado por $\widehat{Y}_i = b_0 + b_1 X_i$ ou $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta_1} X_i$.

a) Explique as diferenças entre **erros aleatórios** e **desvios da regressão**.

b) Mostre que $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.

c) Para o modelo $Y_i = \beta_1 X_i + \varepsilon_i$, com $\beta_0 = 0$, obtenha o estimador $\widehat{\beta}_1$ pelo MMQ.

2) A tabela a seguir apresenta dados de uma amostra de 10 pacientes de um estudo médico conduzido para se pesquisar o relacionamento entre as variáveis idade (X) em anos e o número máximo de batimentos cardíacos por minuto (Y).

Idade	10	20	20	25	30	30	30	40	45	50
N° de batimentos	210	200	195	195	190	180	185	180	170	165

Dados:

$$SQD_X = 1350 \quad SQD_Y = 1710 \quad SPD_{XY} = -1475 \quad \bar{X} = 30 \quad \bar{Y} = 187$$

Assinale (V) se a afirmativa for totalmente verdadeira ou (F) caso contrário.

a) ( ) A equação de regressão linear simples ajustada é: $\widehat{Y}_i = 219,78 + 1,093 X_i$.

b) ( ) Aproximadamente 94,2% da variação observada nos valores do número máximo de batimentos cardíacos por minuto é explicada pela regressão linear nos valores da idade.

c) ( ) O coeficiente de correlação linear aproximadamente igual a 0.971 (correlação positiva) indica que com o aumento da idade espera-se uma diminuição do número máximo de batimentos cardíacos por minuto e vice-versa .

d) ( ) Uma estimativa do valor médio do número máximo de batimentos cardíacos por minuto para um indivíduo com idade igual a 50 anos é $\approx 165,14$.

e) ( ) A estimativa do correspondente (item d)) erro ou o desvio da regressão é igual a 0,14.

3) Suponha que se estimou o coeficiente de correlação entre as notas da primeira prova (X) e as notas médias finais (Y) de um curso e obteve-se $r_{X,Y} = 0,73$. Os $n = 600$ pares de notas $(X_i, Y_i),\ i = 1, 2, \ldots, n$ apresentaram as seguintes estatísticas:

$$\begin{array}{rll} \text{primeira prova:} & \bar{X} = 72,8 & S_X = 8,1 \\ \text{médias finais:} & \bar{Y} = 76,4 & S_Y = 7,0 \end{array}$$

a) Verifique que a seguinte fórmula é uma alternativa para se estimar β_1:
$\widehat{\beta}_1 = r_{X,Y}\frac{S_Y}{S_X}$.

b) Ajuste uma equação de **regressão linear simples** e interprete a estimativa do coeficiente da regressão $b_1 = \widehat{\beta}_1$

c) Sabendo-se que a média final mínima para ser aprovado no curso é $Y = 60$, qual deve ser a decisão de um aluno que obteve $X = 55$ como nota da primeira avaliação? Continuar no curso ou desistir do curso? **justifique sua resposta com base na equação ajustada no item a)**.

d) Qual é a proporção da variabilidade nas notas médias finais explicada pela regressão nas notas da primeira prova?

4) Um economista interessado em estudar a relação entre o valor da renda familiar extra (X) ou disponível para gastos extras (chamada de *disposable income* na literatura em inglês) e o valor dos gastos com alimentação (Y) conduziu um estudo preliminar com 8 famílias, todas compostas por marido, esposa e dois filhos. Os resultados estão na tabela a seguir com valores X em milhares de dólares por ano e Y em centenas de dólares por ano.

X	30	36	27	20	16	24	19	25	SQD_X=291,88	$\bar{X} = 24,63$
Y	55	60	42	40	37	26	39	43	SQD_Y=783,50	$\bar{Y} = 42,75$

a) Ajuste a equação de regressão linear simples.

b) Interprete o valor do coeficiente da regressão (β_1) em termos do problema anunciado.

c) Calcule o coeficiente de determinação e interprete o valor calculado.

5) (proposto por E.B., monitor em 2001). Pode-se determinar o teor de proteínas (mg/ml) no leite de uma forma indireta analisando-se a absorvância de luz, medida em um aparelho denominado fotocolorímetro. A absorvância consiste na fração da luz incidente que a amostra

é capaz de absorver. Por exemplo, uma absorvância de 0,70 indica que a solução absorveu 70% da luz incidente. Por razões históricas, este método é denominado Método do Biureto. A tabela a seguir apresenta os resultados obtidos em um teste com cinco amostras padrão, de concentração previamente conhecida.

Conc. (mg/ml)	1,00	2,00	3,00	4,00	5,00
Absorvância	0,12	0,31	0,49	0,64	0,77

Pede-se:

a) Ajuste a reta de regressão linear simples para estimar a absorvância (Y) em função da concentração de proteínas (X).

b) Interprete o valor da estimativa do coeficiente de regressão (b_1).

c) Para uma absorvância igual a 0,58 estime a concentração média de proteínas, em mg/ml (regressão inversa).

d) Pode-se utilizar o modelo ajustado em a) para se estimar Y quando $X = 8,00$? explique.

e) Calcule o coeficiente de determinação e interprete.

6) Exemplo extraído de D.S. Falconer (1977), Introdução à genética quantitativa, 1ªedição. Os dados abaixo ilustram o efeito do gene anão em ratos com 6 semanas de idade, sendo X o número de genes e Y o peso médio dos ratos em gramas. O objetivo é relacionar as duas variáveis com um modelo de regressão linear simples (RLS): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

i	1	2	3
X_i	0	1	2
Y_i	15	12	6

a) Apresente a equação de RLS ajustada.

b) Interprete as estimativas dos parâmetros β_0 e β_1.

c) Apresente os desvios da regressão e mostre que a soma deles é igual a zero.

d) Calcule e interprete o coeficiente de determinação.

7) (I/2006). A eficiência de uma enzima utilizada no processo de fabricação de medicamentos é avaliada pela quantidade do princípio ativo do medicamento que é produzido na reação química catalizada pela enzima. Considere que a quantidade do princípio ativo (Y, em mg/kg do soluto) produzido em função da concentração do soluto (X, em g/kg do solvente) pode ser explicado por um modelo de regressão linear simples: $Y_i = \beta_1 X_i + \varepsilon_i$, ($\beta_0 = 0$). Os resultados obtidos por uma empresa que conduziu testes com duas enzimas, A e B, estão apresentados na tabela a seguir. Note que não ocorre reação química e portanto nenhum princípio ativo é produzido quando não há soluto.

ENZIMA	MODELO AJUSTADO	$r^2(\%)$
A	$\widehat{Y}_i = 6,5X_i$	91
B	$\widehat{Y}_i = 10,2X_i$	98

a) Qual das duas enzimas foi a mais eficiente nos testes? Justifique sua resposta com base nos modelos ajustados.

b) O modelo ajustado explicou melhor o fenômeno estudado para qual das duas enzimas? Justifique sua resposta.

c) Quando a concentração do soluto for igual a 20 g/kg do solvente e a reação for catalizada pela enzima A, qual é a estimativa da quantidade média do princípio ativo produzida (mg/kg do soluto)?

8) (II/2006). O preço de um modelo de motocicleta usada está linearmente relacionado ao ano de fabricação. A tabela a seguir apresenta os valores do preço (em milhares de reais, $R\$ \times 1000$) e o respectivo ano de fabricação (1993 a 1999, exceto 1996) de 6 motocicletas pesquisadas,

Motocicleta (i)	1	2	3	4	5	6
Ano (X_i)	93	94	95	97	98	99
Preço (Y_i)	6,3	7,0	8,2	9,0	10,5	12

Pede-se:

a) A estimativa do acréscimo médio no preço da motocicleta, para cada aumento de um ano (mais nova), é igual a $R\$$.....

b) O percentual do valor da variância observada nos preços, representado pelo valor da variância dos preços estimados, ou *explicado* pela regressão nos valores do ano de fabricação, é igual a .....%.

c) $R\$$.....é o preço mediano das motocicletas pesquisadas.

d) $R\$$..... é a amplitude total dos preços das motocicletas pesquisadas.

e) Estimar que o preço médio de uma motocicleta ano 1992 seja igual a $R\$$..... seria uma ..... com o modelo.

f) Os estimadores b_0 e b_1 foram obtidos pelo método.....
....

g) ..... é o desvio da regressão para o ano 1997.

h) O ..... do modelo, indicado por ε_i, é não observável e representa o efeito de todas as variáveis explicativas não incluídas no modelo, além das causas não controláveis de variação.

i) $R\$$..... é uma estimativa do preço médio de uma motocicleta 1996.

9) (II/2006). Estudou-se o relacionamento entre o tempo de uma reação química, expresso em minutos (Y) e o valor da concentração, expressa em %, de um composto ativador da reação (X). Os valores testados para X variaram de 0% a 51%, tendo este último valor causado reação *instantânea*. O estudo possibilitou o ajuste da seguinte equação de regressão linear simples,

$$\widehat{Y}_i = 10,2 - 0,20X \qquad r^2 = 0,9362$$

a) Interprete a estimativa do coeficiente da regressão.

b) Interprete a estimativa da constante da regressão.

c) Interprete a estimativa do coeficiente de determinação.

10) (I/2007). (Exemplo obtido de http://statmaster.edu.dk) Em um estudo sobre nutrição infantil em países em desenvolvimento, avaliou-se mensalmente as alturas (Y, em cm) de crianças com 18 a 30 meses de idade (X, em meses) da vila de Kalama no Egito. O objetivo do estudo era modelar por regressão linear simples (RLS), o relacionamento entre idade e altura com o propósito de compará-lo com outros países investigados no estudo.

X	18	19	20	21	22	23	24	25	26	27	28	29	30
Y	76,1	77,0	78,1	78,2	78,8	79,7	79,9	81,1	81,2	81,8	82,8	83,5	84,6

Pede-se:

a) Informe os valores das somas a seguir, $\sum X$ $\sum X^2$ $\sum Y$ $\sum Y^2$ e $\sum XY$.

b) Estime o acréscimo médio na altura, para cada aumento de um mês na idade.

c) Calcule o percentual do valor da variância observada nas alturas, representado pelo valor da variância das alturas estimadas pelo modelo de RLS, ou *explicado* pela regressão nos valores das idades.

d) Calcule os desvios da regressão para as idades 18 e 30 meses.

Respostas dos exercícios propostos

1)a) $\widehat{\varepsilon}_i$ são os desvios da regressão, valores estimados após o ajuste do modelo, ε_i são os erros aleatórios, não observáveis, do modelo e se referem aos efeitos de todas as fontes de variação não consideradas no modelo, essencialmente outras variáveis explicativas e causas aleatórios não controláveis. 1)b) trabalhe por propriedades de somatório até obter $\sum \widehat{\varepsilon} = \sum(Y - \bar{Y}) - b_1 \sum(X - \bar{X})$ 1)c) $SQE = f(\beta_1) = \sum_{i=1}^{n} (Y_i - \beta_1 X_i)^2 \Longrightarrow \frac{d\ f(\beta_1)}{d\ \beta_1}|_{\widehat{\beta}_1} \equiv 0$ resulta em $\widehat{\beta}_1 = \frac{\sum XY}{\sum X^2}$.

2)a) F 2)b) V 2)c) F 2)d) V 2)e) F

3)

a) $b_1 = \frac{SPD_{XY}}{SQD_X} = \frac{r_{XY}\sqrt{SQD_X\ SQD_Y}}{SQD_X} = \frac{r_{XY}\sqrt{S_X^2\ S_Y^2\ (n-1)^2}}{S_X^2\ (n-1)} = r_{XY}\frac{S_Y}{S_X}$

b) $\widehat{Y}_i = 30,47 + 0,63X_i$. A estimativa $b_1 = 0,63$ significa que para cada ponto obtido na primeira prova estima-se um aumento médio de 0,63 pontos na média final.

c) $\widehat{Y} = 30,47 + 0,63(55) \approx 65,1$. Deve continuar pois a média final estimada é superior a 60.

d) $r^2 = 0,73^2 = 0,5329$ ou $53,29\%$. r^2 é o coeficiente de determinação e r é o coeficiente de correlação.

4) a) $\widehat{Y}_i = 12,8 + 1,2X_i$ b) Estima-se aumento médio de 120 dólares nos gastos com alimentos para cada 1000 dólares de aumento na renda extra. c) $r^2 = 54,88\%$ é o percentual da variabilidade observada nos gastos sendo *explicada* pela RLS nos valores de renda extra.

5)a) $\widehat{Y}_i = -0,023 + 0,163X_i$ 5)b) Para cada aumento de 1 mg/ml na conc. de proteína estima-se aumento médio de 0,163 ou 16,3% na absorvância. 5)c) $\widehat{X}_i = \frac{0,023}{0,163} + \frac{1}{0,163}Y_i$ portanto $b_0^* = 0,141$ e $b_1^* = 6,135$ fornece $\widehat{X} = 3,699$ mg/ml 5)d) Sim, $\widehat{Y}_i = -0,023 + 0,163 \times 8 = 1,281$ ou $128,1\%$ mas além de ser uma extrapolação, o valor estimado supera 100% 5)e) $r^2 = 99,4\%$ é o percentual da variabilidade observada nos valores da absorvância *explicado* pela RLS nos valores da conc. de proteínas.

6)a) $\widehat{Y}_i = 15,5 - 4,5X_i$ 6)b) $b_0 = 15,5$ gramas é uma estimativa do peso médio dos ratos que não possuem o gene anão e $b_1 = -4,5$ é uma estimativa do decréscimo médio no peso para cada um gene anão de acréscimo. 6)c) $\widehat{\varepsilon}_1 = 15 - 15,5 = -0,5$, $\widehat{\varepsilon}_2 = 12 - 11 = 1$ e $\widehat{\varepsilon}_3 = 6 - 6,5 = -0,5$, portanto $\sum_{i=1}^{3} \widehat{\varepsilon}_i = 0$ 6)d) $r^2 \approx 96,4\%$ é o percentual da variabilidade nos valores de peso sendo *explicados* pela RLS nos valores do número de genes anão.

7)a) Enzima B, por apresentar maior valor b_1, o que significa maior aumento médio estimado do P.A. para cada aumento de uma unidade do soluto 7)b) Enzima B, maior r^2 7)c) $\widehat{Y} = 6,5 \times 20 = 130$ mg/kg do soluto.

8)a) $R\$890,00$ 8)b) $96,22\%$ 8)c) $R\$8600,00$ 8)d) $R\$5700,00$ 8)e) $R\$5276,00$ seria uma extrapolação 8)f) dos mínimos quadrados 8)g) -0,72 8)h) erro aleatório 8)i) $R\$8830,00$.

9)a) $b_1 = -0,20$, para cada acréscimo de 1% na conc. do composto, estima-se uma diminuição média de 0,20 minutos no tempo da reação (aumento de velocidade) 9)b) $b_0 = 10,2$, estima-se um tempo médio de 10,2 minutos quando nenhum composto (0%) é utilizado 9)c) $r^2 = 93,62\%$ é o percentual da variabilidade observada nos valores do tempo de reação que foram explicados pela RLS nos valores da concentração do composto.

10)a) $\sum X = 312$ $\sum X^2 = 7670$ $\sum Y = 1042,8$ $\sum Y^2 = 83727,74$ e $\sum XY = 25146,5$. 10)b) $b_1 = 0,6555$ 10)c) $r^2 = 98,8\%$ 10)d) $\widehat{\varepsilon}_{18} = -0,1824$ e $\widehat{\varepsilon}_{30} = 0,4516$

Capítulo 7

Testes de hipóteses

7.1 Introdução

Um dos principais objetivos da Estatística é inferir, ou concluir, para a população com base em dados amostrais. As inferências são realizadas basicamente de duas formas: pela estimação de valores dos parâmetros da população ou por testes de hipóteses à respeito de seus valores. No presente capítulo trataremos da segunda alternativa.

O teste estatístico de hipótese(s) é uma regra decisória que nos permite rejeitar, ou não rejeitar, uma hipótese estatística com base nos resultados de uma amostra. Estas hipóteses são, em geral, acerca dos parâmetros (valores da população) e a realização do teste se baseia na distribuição amostral dos respectivos estimadores, sob a pressuposição da hipótese sendo testada ser verdadeira. Alguns dos principais conceitos e a metodologia para condução de um teste de hipóteses serão abordados no presente capítulo. Trataremos apenas do enfoque clássico (ou frequentista) para os testes.

7.2 Alguns conceitos

7.2.1 Parâmetro, estimador e estimativa

Definição 11. Parâmetro é uma função de valores populacionais, sendo em geral, um valor desconhecido associado à população.

Exemplo: Na distribuição normal os parâmetros são a média $\mu = E(X)$ e a variância $\sigma^2 = V(X)$.

Definição 12. Um **estimador** de um parâmetro é qualquer função das observações da amostra aleatória $X_1, X_2, \cdots, X_n$. Ele representa uma dada fórmula de cálculo que fornecerá valores (ou estimativas) que serão diferentes, conforme a amostra selecionada.

Exemplos:

a) O estimador da média μ é $\hat{\mu} = \bar{X} = \dfrac{\sum_{i=1}^{n} X_i}{n}$;

b) O estimador da variância σ^2 é $\hat{\sigma}^2 = s^2 = \dfrac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1} = \dfrac{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$.

Definição 13. A **estimativa** é o valor numérico assumido pelo estimador, quando os valores $x_1, x_2, \ldots, x_n$ são considerados.

Exemplos: $\bar{X} = 10,42$ e $s^2 = 4,67$.

7.2.2 Hipóteses estatísticas

Hipótese estatística é uma suposição quanto ao valor de um parâmetro, ou uma afirmação quanto à natureza da população. As hipóteses estatísticas devem ser formuladas de modo a minimizar os erros de decisão, entretanto, estes aspectos da formulação das hipóteses não será abordado no presente texto.

Exemplos:

a) A média populacional da altura dos brasileiros adultos é 1,65 m, isto é, $\mu = 1,65$ m;

b) A distribuição dos pesos dos alunos da Universidade Federal de Viçosa é normal;

c) A proporção de indivíduos com alguma doença é 3%, $p = 0,03$;

d) A opinião acerca de uma lei e a renda mensal, dos moradores de uma cidade, são duas variáveis independentes.

Em um teste de hipótese, formulam-se duas hipóteses, a primeira chamada de hipótese de nulidade e a segunda denominada hipótese alternativa, estudadas nas seções 7.2.2.1 e 7.2.2.2, respectivamente.

7.2.2.1 Hipótese de nulidade (H_0)

É a hipótese estatística a ser testada. A hipótese H_0 é formulada com o "expresso propósito de ser rejeitada", e os testes são construídos sob a pressuposição de H_0 ser verdadeira. O teste de hipótese consiste em verificar se a amostra observada difere significativamente do resultado esperado sob H_0.

Exemplos:

a) Um fabricante informa que a tensão média de ruptura dos cabos é 50 kgf. $H_0 : \mu = 50$;

b) A informação de um fabricante, quanto à durabilidade média de suas lâmpadas, é de 6000 horas. $H_0 : \mu = 6000$;

c) Duas marcas de rações, X_1 e X_2, para leitões em fase de crescimento, propiciam em média o mesmo ganho de peso. $H_0 : \mu_1 = \mu_2$.

Para os três exemplos anteriores o raciocínio é que, enquanto não houver evidência amostral sugerindo que a informação não deve ser verdadeira, toma-se a informação como verdadeira.

7.2.2.2 Hipótese alternativa (H_a ou H_1)

É uma hipótese que contraria H_0, formulada com base no conhecimento prévio do problema, informações de pesquisas, etc. Para os três exemplos anteriormente citados, teríamos as seguintes possibilidades:

a) $H_{a1} : \mu \neq 50$ (teste bilateral), ou, $H_{a2} : \mu < 50$ (teste unilateral à esquerda), ou, $H_{a3} : \mu > 50$ (teste unilateral à direita).

b) $H_{a1} : \mu \neq 6000$ (teste bilateral), ou, $H_{a2} : \mu < 6000$ (teste unilateral à esquerda), ou, $H_{a3} : \mu > 6000$ (teste unilateral à direita).

c) $H_{a1} : \mu_1 \neq \mu_2$ (teste bilateral), ou, $H_{a2} : \mu_1 < \mu_2$ (teste unilateral à esquerda), ou, $H_{a3} : \mu_1 > \mu_2$ (teste unilateral à direita).

Conforme já explicado, o teste de hipóteses é um procedimento que, mediante informações obtidas de amostras, permite decidir rejeitar ou não rejeitar H_0. Portanto, o resultado final do teste de hipóteses (ou a decisão do teste) é enunciado em termos da hipótese de nulidade. Apesar de serem expressões sinônimas, geralmente os livros textos preferem que seja dito **não rejeitar** H_0 ao invés de **aceitar** H_0. Isto porque pode parecer que o termo aceitar signifique que a hipótese seja verdadeira, o que não é verdade, pois os erros de decisão são inerentes à metodologia do teste.

7.2.3 Nível de significância α

O nível de significância de um teste de hipóteses é a seguinte probabilidade (condicional), denotada como α,

$$\alpha = P\left(\text{rejeitar } H_0 \,|\, H_0 \text{ é verdadeira}\right).$$

Geralmente o valor de α é pequeno, $\alpha = 0,05$ ou $\alpha = 0,01$, de modo que uma decisão errônea do tipo rejeitar uma hipótese H_0 verdadeira seja minimizada. Entretanto, ao se minimizar este tipo de erro, automaticamente se aumenta a probabilidade da não rejeição de uma hipótese H_0 que seja falsa. Neste contexto, existem outras definições importantes relacionadas às propriedades desejáveis de um teste de hipóteses, tais como o poder de um teste e também o conceito de testes uniformemente mais poderosos (UMP). A teoria dos testes UMP não será tratada neste texto introdutório.

7.2.4 Região crítica

A região crítica em um teste de hipótese é a faixa de valores que nos levam à rejeição da hipótese H_0. Isto é, caso o valor observado da estatística do teste (Z, t, χ^2, F) pertença à região crítica, rejeita-se H_0, caso contrário não se rejeita H_0. Qualquer decisão tomada implica na possibilidade de cometer basicamente dois tipos de erros.

A estatística do teste fornece o valor calculado do teste, o valor de uma variável aleatória, obtido sob H_0, cujo modelo de probabilidade ou distribuição é conhecido. Portanto a estatística do teste é um valor calculado com os dados amostrais e que pode-se provar ser uma variável aleatória com distribuição conhecida (Z, t, χ^2, F, etc.).

7.2.5 Erros de decisão

Após a tomada de decisão no teste, (rejeitar ou não rejeitar a hipótese H_0), podem ocorre dois tipos de erros, dados nas seções 7.2.5.1 e 7.2.5.2.

7.2.5.1 Erro tipo I

O erro tipo I é caracterizado pelo fato de rejeitarmos H_0 quando esta é verdadeira. Portanto, α é a máxima probabilidade de se cometer o erro tipo I. Em geral, os valores mais utilizados de α são 1% e 5%. Entretanto, na prática, os softwares estatísticos são empregados na execução dos testes (R, SAS, SAEG, Minitab, SPlus, etc) e o valor$-p$ informado pelo software é utilizado nas inferências.

7.2.5.2 Erro tipo II

O erro tipo II é caracterizado pelo fato de não rejeitarmos (ou aceitarmos) H_0 quando esta é falsa. Designaremos por β a probabilidade de se cometer o erro tipo II.

A Tabela 7.1 sumariza as decisões tomadas em um teste de hipóteses, apresentado as decisões corretamente tomadas e indicando os erros cometidos.

Tabela 7.1: Erros cometidos em um teste de hipóteses

Decisão	Realidade	
	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	Erro tipo I	Não há erro
Não rejeitar H_0	Não há erro	Erro tipo II

7.2.6 Valor$-p$ ou nível crítico ou probabilidade de significância

O valor$-p$ em um teste de hipótese é o menor valor de α (nível de significância) para o qual se rejeita H_0. Uma definição mais geral: o valor$-p$ é a probabilidade de se obter um valor da estatística do teste tão, ou mais extremo, do que o valor observado, em favor da hipótese alternativa H_a ou H_1. Portanto, na prática, se valor$-p \leq \alpha$ a decisão é rejeitar H_0.

7.2.7 Poder de um teste

É a probabilidade de rejeitar H_0 quando esta é falsa.

$$\text{Poder} = 1 - \beta.$$

O poder de um teste frente à determinada hipótese é uma informação utilizada para o dimensionamento de tamanhos de amostras tendo-se em vista o controle dos dois tipos de erros.

A Tabela 7.2 a seguir sintetiza as probabilidades no cenário decisão *versus* realidade.

Tabela 7.2: Probabilidade dos erros cometidos em um teste de hipóteses

Decisão	Realidade	
	H_0 é verdadeira	H_0 é falsa
Rejeitar H_0	α	$1-\beta$
Não rejeitar H_0	$1-\alpha$	β

7.3 Etapas para a realização de um teste de hipóteses

Obviamente que na "vida real" o teste será realizado com o auxílio de algum software e, neste caso, não haverão valores tabelados, e sim o valor$-p$ será informado pelo software. Hipóteses e nível de confiança estarão implícitas para o executor do teste.

As etapas listadas a seguir resumem o procedimento a ser executado pelos alunos nas avaliações.

a) Enunciar as hipóteses H_0 e H_a;

b) Fixar o nível de significância α e identificar a estatística do teste;

c) Determinar a região crítica e a região de não rejeição de H_0 em função do nível α pelas tabelas estatísticas;

d) Por meio dos elementos amostrais, calcular o valor da estatística do teste;

e) Concluir pela rejeição ou não-rejeição de H_0, caso o valor da estatística obtido na 4ª etapa pertença ou não pertença, respectivamente, à região crítica determinada na 3ª etapa e descrever a conclusão prática do teste de hipóteses.

7.4 Teste Z ("grandes amostras")

7.4.1 Teste de hipótese de uma média populacional

Seja X uma variável aleatória com média $E(X)=\mu$ e variância $V(X)=\sigma^2$.

Em grandes amostras pode-se demonstrar que $\bar{X}$, a média amostral, é normalmente distribuída com média μ e variância $\dfrac{\sigma^2}{n}$. Isto é: $\bar{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$. Este resultado é conhecido como Teorema Central do Limite (TCL).

Usando-se a variável normal padronizada ou reduzida Z, temos:

$$Z = \frac{\bar{X}-\mu}{\sigma_{\bar{X}}}$$

em que,

$$\sigma_{\bar{X}} = \sqrt{V\left(\bar{X}\right)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

Logo,

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

portanto, para testar $H_0 : \mu = \mu_0$ calcula-se o valor de Z sob H_0, $Z_0 = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, em que $\bar{X}$ é a média da amostra $X_1, X_2, \cdots, X_n$ e σ é o desvio padrão da "população" que forneceu a amostra. Na prática, σ^2 também é desconhecido e utiliza-se a variância da amostra ($S^2 = \sigma^2$) no cálculo de Z_0. Este procedimento é válido em grandes amostras, por exemplo, $n > 30$.

Se o tamanho da amostra não é suficientemente grande (adotaremos $n < 30$), para supor que S^2 (variância da amostra) seja aproximadamente igual a σ^2, então a distribuição t de Student é utilizada. Isto será abordado adiante no texto no item 7.7.

Exercícios

1) A tensão de ruptura dos cabos produzidos por um fabricante apresenta a média de 1800 kg e o desvio padrão de 100 kg. Mediante nova técnica no processo de fabricação, proclamou-se que a tensão de ruptura pode ter aumentado. Para testar essa declaração, ensaiou-se uma amostra de 50 cabos, tendo-se determinado a tensão média de ruptura de 1850 kg. Pode-se confirmar a declaração ao nível de significância 0,05?

Resposta: $H_0 : \mu = 1800$ kg *versus* $H_a : \mu > 1800$ kg. $z_{\text{Calculado}} = 3,53, z_{\text{Tabelado}} \cong 1,64$. Rejeita-se H_0.

2) Considerando o mesmo problema anterior, testar:

i) $H_0 : \mu = 1800$ kg *versus* $H_a : \mu \neq 1800$ kg;

ii) Idem, utilizando $\alpha = 1\%$.

A hipótese alternativa $H_0 : \mu \neq 1800$ deve ser utilizada quando não houverem suspeitas de que a nova técnica seja melhor, isto é, pode ser pior.

Respostas:

i) $z_{\text{Calculado}} = 3,53, z_{\text{Tabelado}} = 1,96$. Rejeita-se H_0.

ii) $z_{\text{Calculado}} = 3,53, z_{\text{Tabelado}} \cong 2,57$. Rejeita-se H_0.

7.4.2 Outra aplicação do teste Z

O teste Z para uma média é um exemplo de uma aplicação do Teorema Central do Limite (TCL). Pelo TCL, as médias $\bar{X}$ de "grandes" (geralmente $n > 30$ atende, mas depende de outras características da população) amostras de tamanho n obtidas de uma população com média μ e variância σ^2 são normalmente distribuídas com média $E\left(\bar{X}\right) = \mu$ e variância $V\left(\bar{X}\right) = \sigma^2/n$. Então sabemos que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1).$$

Portanto, na tabela da distribuição normal padrão (Tabela A.1 do Apêndice), pode-se obter o valor positivo z tal que

$$P\left(-z < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z\right) = 1-\alpha. \tag{7.1}$$

A equação (7.1) pode ser reescrita como

$$P\left(\bar{X} - D < \mu < \bar{X} + D\right) = 1-\alpha, \tag{7.2}$$

em que $D = z\sqrt{\sigma^2/n}$ é geralmente designado como o erro de estimação e $1-\alpha$ é o nível de confiança, sendo α o nível de significância (em geral, 1%, 5% ou 10%). O resultado (7.2) é simplesmente uma afirmação probabilística futura: com probabilidade $1-\alpha$ o intervalo especificado irá conter o valor de μ. Após obtidos os dados, tem-se um intervalo com confiança $100\,(1-\alpha)\,\%$ para μ.

Exercício proposto

1) **(Freund e Simon (2000))** Os técnicos de uma grande indústria precisam determinar o tempo médio de montagem de uma peça. Eles pretendem utilizar a média de uma amostra aleatória de 150 operários para estimar este tempo médio. Com base em experiência, admite-se $\sigma = 6,2$. Pede-se:

a) Calcule o erro de estimação se for adotado 99% como nível de confiança;

b) Construa o respectivo intervalo de confiança para μ se a amostra fornece $\bar{X} = 69,5$.

7.4.3 Teste que envolve diferença de duas médias populacionais

Caso em que n_1 e n_2 são "grandes amostras".

Sejam $\bar{X}_1$ e $\bar{X}_2$ as médias obtidas em duas amostras de tamanhos n_1 e n_2, retiradas de duas populações X_1 e X_2, respectivamente, com variâncias σ_1^2 e σ_2^2 conhecidas e médias μ_1 e μ_2 desconhecidas.

Considerando-se as variáveis aleatórias $\bar{X}_1$ e $\bar{X}_2$ independentes, tem-se que:

$$\left(\bar{X}_1 - \bar{X}_2\right) \sim N\left(\mu_1 - \mu_2; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right), \tag{7.3}$$

em grandes amostras.

Portanto, ao nível de significância α pode-se, testar:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \text{ contra } H_{a_1} : \mu_1 < \mu_2, \text{ ou} \\ H_0 : \mu_1 = \mu_2 \text{ contra } H_{a_1} : \mu_1 > \mu_2, \text{ ou} \\ H_0 : \mu_1 = \mu_2 \text{ contra } H_{a_1} : \mu_1 \neq \mu_2. \end{cases}$$

O resultado (7.3) também é uma aplicação do TCL, mencionado anteriormente.

Utilizaremos então, a estatística:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - (\mu_1 - \mu_2)}{\sqrt{V\left(\bar{X}_1 - \bar{X}_2\right)}}.$$

Sob H_0 verdadeira, segue que $\left(\bar{X}_1 - \bar{X}_2\right) \sim N\left(0; \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$, e assim virá:

$$Z = \frac{\left(\bar{X}_1 - \bar{X}_2\right) - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \qquad \text{ou ainda,} \qquad Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}.$$

No teste Z para duas médias admite-se que as amostras n_1 e n_2 sejam "grandes amostras". Adotaremos na EST 105 que $n \geq 30$ atende essa condição. Portanto ao se mencionar σ_1^2 e σ_2^2 como "variâncias conhecidas", fica subentendido nos exercícios e no texto, que na verdade $S_1^2 \approx \sigma_1^2$ e $S_2^2 \approx \sigma_2^2$, ou seja, que as estimativas de variâncias obtidas das amostras são iguais aos valores dos parâmetros.

Exercício

1) Dois métodos X_1 e X_2 para execução de determinada tarefa são propostos. Deseja-se saber se são igualmente eficientes, no sentido do tempo exigido para a execução desta. Sabe-se que os tempos de execução em minutos através dos métodos X_1 e X_2, são normalmente distribuídos com variâncias $\sigma_1^2 = 8$ e $\sigma_2^2 = 10$. A fim de chegar a uma decisão, 48 operários, selecionados ao acaso, foram treinados para executar a tarefa através do método X_1 e 36 através do método X_2, durante certo tempo. A seguir os tempos de execução foram medidos obtendo-se as seguintes médias amostrais (em minutos): $\bar{X}_1 = 40$ e $\bar{X}_2 = 42$. A que conclusão chegar ao nível $\alpha = 5\%$?

Resposta: $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$. $z_{\text{Calculado}} = -3, z_{\text{Tabelado}} = 1,96$. Rejeita-se H_0.

7.5 O teste de qui-quadrado (χ^2)

Uma medida da discrepância existente entre as frequências observadas e esperadas é proporcionada pela estatística χ^2, expressa por:

$$\chi^2_{cal} = \frac{(F_{o_1} - F_{e_1})^2}{F_{e_1}} + \frac{(F_{o_2} - F_{e_2})^2}{F_{e_2}} + \cdots + \frac{(F_{o_k} - F_{e_k})^2}{F_{e_k}}$$

$$\chi^2_{cal} = \sum_{i=1}^{k} \frac{(F_{o_i} - F_{e_i})^2}{F_{e_i}}, \tag{7.4}$$

sendo F_o: frequência observada e F_e: frequência esperada sob H_0.

A expressão (7.4) nos dá um valor sempre positivo e tanto menor quanto maior for a concordância entre as frequências observadas e as frequências esperadas, calculadas com base

em H_0. Pode-se demonstrar que a estatística (7.4) tem uma distribuição "qui-quadrado", com v graus de liberdade, isto é

$$\sum_{i=1}^{k} \frac{\left(F_{o_i} - F_{e_i}\right)^2}{F_{e_i}} \sim \chi^2_v.$$

Regra decisória: Se $\chi^2_{cal} \geq \chi^2_{\text{Tabelado}}$, rejeita-se H_0.
O teste de qui-quadrado pode ser usado principalmente como:

a) Teste de aderência;

b) Teste de independência;

c) Teste de homogeneidade.

Exercício

1) Explique como um software calcula o valor$-p$ no teste de χ^2.

7.5.1 Teste de aderência

Pode ser utilizado para testar o ajustamento dos dados observados a um modelo ou a uma função (dividida em k categorias) de frequências.

Nota: Neste caso o número de graus de liberdade é dado por:

a) $v = k - 1$, quando as frequências esperadas puderem ser calculadas sem que se façam estimativas dos parâmetros populacionais a partir das distribuições amostrais. Tem-se que "k" é o número de categorias em que foi dividida a amostra.

b) $v = k - 1 - r$, quando para determinação das frequências esperadas, r parâmetros tiverem suas estimativas calculadas a partir das distribuições amostrais.

Exemplos

1) Uma pesquisa feita junto a 320 famílias, de 5 filhos cada, revelou a distribuição apresentada a seguir. Tais resultados são consistentes com a hipótese de igual probabilidade de nascimento para ambos os sexos? Usar $\alpha = 5\%$.

Número de meninos	Número de meninas	Número de famílias
5	0	18
4	1	56
3	2	110
2	3	88
1	4	40
0	5	8
		Total: 320

Resposta: $\chi^2_{\text{Calculado}} = 11,96; \chi^2_{\text{Tabelado}} = \chi^2_{5\%}(5) = 11,07$. Rejeita-se H_0.

2) Em seus experimentos com ervilhas, Mendel observou 315 lisas e amarelas, 108 lisas e verdes, 101 rugosas e amarelas, 32 rugosas e verdes. De acordo com sua teoria de hereditariedade, os números deveriam apresentar-se na proporção 9:3:3:1. As observações estão de acordo com esta teoria, ao nível de 1% de probabilidade?

Resposta: H_0 : proporção $= 9:3:3:1$ *versus* H_a : proporção $\neq 9:3:3:1$. $\chi^2_{\text{Calculado}} = 0,47; \chi^2_{\text{Tabelado}} = \chi^2_{1\%}(3) = 11,345$. Não se rejeita-se H_0.

7.5.2 Teste de independência

Este teste é usado em conexão com as tabelas de contingência. Essas tabelas são construídas com o propósito de se estudar a relação de dependência (associação) entre duas variáveis classificadas segundo critérios qualitativos.

Coloca-se à prova as hipóteses:

H_0 as variáveis são independentes, contra H_a as variáveis não são independentes, ou seja, elas apresentam algum grau de associação entre si.

Para investigar a concordância entre frequências observadas e frequências esperadas, utilizamos a estatística:

$$\chi^2_{cal} = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{\left(F_{o_{ij}} - F_{e_{ij}}\right)^2}{F_{e_{ij}}},$$

com os dados obtidos em tabelas com h linhas por k colunas (tabelas de contingência).

Pode-se demonstrar que χ^2_{cal} tem distribuição de qui-quadrado com v graus de liberdade, isto é: $\chi^2_{cal} \sim \chi^2_v$.

Quanto ao número de graus de liberdade v, devemos observar:

a) $v = (h-1)(k-1)$, se as frequências esperadas podem ser calculadas sem necessidade de estimação de parâmetros da população.

b) $v = (h-1)(k-1) - r$, se as frequências esperadas só podem ser avaliadas estimando-se r parâmetros populacionais.

Uma restrição ao uso do teste: $F_{e_{ij}} \geq 5$.

Exemplo proposto

1) A Tabela a seguir exibe as notas obtidas acerca de estudantes de Estatística (X) e Cálculo (Y). São apresentados o número de estudantes com notas (x e y) em três categorias, 0 a 5, 5 a 7 e, 7 a 10 pontos. Testar a hipótese de que os resultados em Estatística são independentes dos resultados em Cálculo, ao nível de significância de 2,5%.

Nota em Cálculo (Y)	Nota em Estatística (X)			Total
	$0 \leq x < 5$	$5 \leq x < 7$	$7 \leq x \leq 10$	
$0 \leq y < 5$	75	35	13	123
$5 \leq y < 7$	29	120	32	181
$7 \leq y \leq 10$	15	70	46	131
Total	119	225	91	435

Resposta: H_0 : As variáveis X e Y são independentes *versus* H_a : As variáveis X e Y não são independentes. $\chi^2_{\text{Calculado}} = 111,64; \chi^2_{\text{Tabelado}} = \chi^2_{2,5\%}(4) = 11,143$. Rejeita-se H_0.

7.5.3 Teste de homogeneidade

No teste de homogeneidade, uma das variáveis praticamente representa uma classificação dos elementos em populações distintas. Teremos então várias amostras, cada uma retirada de uma população diferente, e estaremos testando pelo χ^2 a hipótese de que a variável em estudo se distribui de forma homogênea nas várias populações. Embora o teste seja formalmente o mesmo, quando encarado dessa forma, iremos considera-lo como um teste de homogeneidade.

Para o caso de k amostras, podemos considerar:

$$\begin{cases} H_0: & \text{As populações são homogêneas;} \\ H_a: & \text{Pelo menos uma das populações não é homogênea com as demais.} \end{cases}$$

Exercício

1) Suponhamos que certo bairro possua 2 colégios A e B, igualmente procurados por crianças de todos os níveis econômicos, e que alguém afirme que a direção de um dos colégios faz certa discriminação quanto à aceitação de alunos, no sentido de que crianças de nível econômico mais elevado, tem mais chances de ser escolhidas. A fim de verificar este fato, selecionou-se uma amostra ao acaso de 100 crianças do colégio A e outra de 120 crianças do colégio B. Os resultados estão na Tabela a seguir (exemplo obtido em Gattás (1978)). Testar a hipótese de não haver discriminação por parte dos colégios. Usar $\alpha = 5\%$.

Colégio	Nível econômico			Total
	Inferior	Médio	Superior	
A	20	40	40	100
B	50	40	30	120
Total	70	80	70	220

Resposta: $\begin{cases} H_0: & \text{Não há discriminação por parte dos colégios;} \\ H_a: & \text{Há discriminação por parte dos colégios.} \end{cases}$

$\chi^2_{\text{Calculado}} = 12,57; \chi^2_{\text{Tabelado}} = \chi^2_{5\%}(2) = 5,991$. Rejeita-se H_0.

7.6 Teste de comparação de variâncias de duas populações

Sejam U e V variáveis aleatórias independentes, com distribuição de qui-quadrado com n_1 e n_2 graus de liberdade, respectivamente. Denomina-se F a variável aleatória definida pelo quociente:

$$F = \frac{U/n_1}{V/n_2}.$$

Considerando duas amostras de tamanhos n_1 e n_2 das variáveis aleatórias normais X_1 e X_2, respectivamente, pode-se demonstrar que :

$$U = \frac{(n_1-1)\,s_1^2}{\sigma_1^2} = \frac{\nu_1 s_1^2}{\sigma_1^2} \qquad \text{e} \qquad V = \frac{(n_2-1)\,s_2^2}{\sigma_2^2} = \frac{\nu_2 s_2^2}{\sigma_2^2}.$$

Então, sob $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$, a estatística

$$F = \frac{\dfrac{U}{\nu_1}}{\dfrac{V}{\nu_2}} = \frac{\left[\dfrac{(n_1-1)\,s_1^2}{(n_1-1)\,\sigma^2}\right]}{\left[\dfrac{(n_2-1)\,s_2^2}{(n_2-1)\,\sigma^2}\right]} = \frac{s_1^2}{s_2^2}$$

ou seja,

$$F = \frac{s_1^2}{s_2^2},$$

tem distribuição F, de Fisher-Snedecor, com $\nu_1 = (n_1 - 1)$ e $\nu_2 = (n_2 - 1)$ graus de liberdade.

Assim, para testar

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \text{ contra } H_{a_1} : \sigma_1^2 < \sigma_2^2, \text{ ou} \\ H_0 : \sigma_1^2 = \sigma_2^2 \text{ contra } H_{a_2} : \sigma_1^2 > \sigma_2^2, \text{ ou} \\ H_0 : \sigma_1^2 = \sigma_2^2 \text{ contra } H_{a_3} : \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

usamos a estatística F dada acima.

Observação: Nesse texto, vamos adotar sempre colocar a maior variância no numerador, de modo a obter um $F_{\text{Calculado}}$ maior que 1, e usaremos a Tabela unilateral para $F > 1$. Assim se:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_a : \sigma_1^2 > \sigma_2^2$$

com $F_{\text{Calculado}} = \dfrac{s_1^2}{s_2^2}$, e F_{Tabelado} dado por $F_\alpha(n_1 - 1; n_2 - 1)$

ou

$$H_0 : \sigma_2^2 = \sigma_1^2 \quad \text{versus} \quad H_a : \sigma_2^2 > \sigma_1^2$$

com $F_{\text{Calculado}} = \dfrac{s_2^2}{s_1^2}$, e F_{Tabelado} dado por $F_\alpha(n_2 - 1; n_1 - 1)$.

Decisão: Se $F_{\text{Calculado}} \geq F_{\text{Tabelado}}$ rejeita-se H_0.

Exercício

1) Na aplicação de dois métodos X_1 e X_2, obteve-se os resultados fornecidos abaixo. Testar a hipótese de igualdade das variâncias, ao nível de 5% de probabilidade.

Método	s^2	n
X_1	40	11
X_2	16	19

Resposta: $H_0 : \sigma_1^2 = \sigma_2^2$ *versus* $H_a : \sigma_1^2 > \sigma_2^2$. $F_{5\%}(10;18) = 2,41; F_{\text{Calculado}} = 2,50$. Rejeita-se H_0.

7.7 Teste t de Student ("pequenas amostras")

7.7.1 Teste de hipótese de uma média populacional

Caso em que X é normalmente distribuída com variância desconhecida.

Se selecionarmos uma amostra aleatória de tamanho n de determinada população, de sorte que $X_1, X_2, \cdots, X_n$ sejam independentes, então:

$$t = \frac{\bar{X} - \mu}{s\left(\bar{X}\right)}$$

tem distribuição t de Student com $n-1$ graus de liberdade.

Mas, $s\left(\bar{X}\right) = \sqrt{\hat{V}\left(\bar{X}\right)} = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$,

logo,

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}.$$

Desse modo, podemos testar:

$$\begin{cases} & H_{a_1} : \mu < \mu_0, \text{ ou} \\ H_0 : \mu = \mu_0 \text{ contra} & H_{a_2} : \mu > \mu_0, \text{ ou} \\ & H_{a_3} : \mu \neq \mu_0, \end{cases}$$

em que usaremos a estatística t, dada anteriormente.

Decisão:

a) Teste bilateral

Se $|t_{\text{Calculado}}| \geq t_{\text{Tabelado}}$ rejeita-se H_0;

b) Teste unilateral à direita

Se $t_{\text{Calculado}} \geq t_{\text{Tabelado}}$ rejeita-se H_0;

c) Teste unilateral à esquerda

Se $t_{\text{Calculado}} \leq -t_{\text{Tabelado}}$ rejeita-se H_0.

Observação: O valor de t_{Tabelado} é obtido em tabelas apropriadas

$$\text{Tabela bilateral} \begin{cases} \text{Teste bilateral: entrar com } \alpha \\ \text{Teste unilateral: entrar com } 2\alpha \end{cases}$$

$$\text{Tabela unilateral} \begin{cases} \text{Teste bilateral: entrar com } \dfrac{\alpha}{2} \\ \text{Teste unilateral: entrar com } \alpha \end{cases}$$

Observação: Consultando a tabela bilateral no teste t (Tabela A.3 do Apêndice), com nível de significância α e o número de graus de liberdade, vamos encontrar o valor de $t_{\frac{\alpha}{2}}$, que é o quantil de ordem $100\left(1-\frac{\alpha}{2}\right)\%$ da distribuição t de Student.

Exercício

1) Determinada firma desejava comprar cabos tendo recebido do fabricante a informação de que a tensão média é de 8000 kgf. Efetuou-se um ensaio em 6 cabos e obteve a tensão média de ruptura 7750 kgf, com um desvio padrão de 145 kgf. Efetuar um teste unilateral para analisar se a afirmação do fabricante é verdadeira ao nível de 5% de probabilidade.

Resposta: $H_0 : \mu = 8000$ kgf *versus* $H_a : \mu < 8000$ kgf. $t_{5\%}(5) = 2,015$, $t_{\text{Calculado}} = -4,22$. Rejeita-se H_0.

7.7.2 Teste de hipótese para o caso de duas amostras independentes

Muitos problemas aparecem quando se deseja testar hipóteses sobre médias de diferentes populações. Por exemplo, um experimentador pode estar investigando um novo tipo de adubo, comparando a produção média de períodos em que foi usado o adubo antigo e o novo.

Quando as variâncias das populações são substituídas pelas variâncias das amostras, isto é, s^2 em lugar de σ^2, o teste Z passa ao teste t. Uma questão crucial é a pressuposição das variâncias das populações serem iguais entre si. Este problema é conhecido na literatura como *Behrens-Fisher problem* e até a presente data não há uma solução definitiva e sim várias propostas. No presente texto consideraremos dois casos:

Caso A: O teste t (de student) na sua forma original. Válido para o caso em que as variâncias são iguais.

Caso B: A solução proposta por Welch para o caso em que as variâncias não são iguais. Neste caso, os graus de liberdade (n^*) são calculados pela aproximação de Welch-Satterthwaite (Satterthwaite (1946) e Welch (1947)).

Sejam X e Y normalmente distribuídas, sendo suas variâncias desconhecidas. Desejamos testar:

$$H_0: \quad \mu_1 = \mu_2 \quad \text{versus} \qquad H_a : \begin{cases} \mu_1 > \mu_2, & \text{ou} \\ \mu_1 < \mu_2, & \text{ou} \\ \mu_1 \neq \mu_2. \end{cases}$$

Antes devemos testar:

$$H_0': \quad \sigma_1^2 = \sigma_2^2 \quad \text{versus} \qquad H_a': \begin{cases} \sigma_1^2 > \sigma_2^2, & \text{ou} \\ \sigma_1^2 < \sigma_2^2, & \text{ou} \\ \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

Caso A: Se H_0' não for rejeitada, vamos admitir que as variâncias são iguais e que, consequentemente, os valores assumidos por s_1^2 e s_2^2 serão estimativas de um mesmo valor σ^2 que é a variância (comum) de ambas as populações. Sendo assim, vamos combinar s_1^2 e s_2^2 a fim de obter um melhor estimador para σ^2.

Temos que:

$$s_1^2 = \frac{SQD_{X_1}}{n_1 - 1} = \frac{\sum\limits_{i=1}^{n} X_{1i}^2 - \frac{\left(\sum\limits_{i=1}^{n} X_{1i}\right)^2}{n_1}}{n_1 - 1} \quad \text{e} \quad s_2^2 = \frac{SQD_{X_2}}{n_2 - 1} = \frac{\sum\limits_{i=1}^{n} X_{2i}^2 - \frac{\left(\sum\limits_{i=1}^{n} X_{2i}\right)^2}{n_2}}{n_2 - 1},$$

de modo que s^2 é o estimador de σ^2 obtido pela combinação de S_1^2 e S_2^2, sendo dado por:

$$s^2 = \frac{SQD_{X_1} + SQD_{X_2}}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}.$$

A seguir utilizaremos para o nosso teste, a variável aleatória

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

que tem distribuição t de student com $(n_1 + n_2 - 2)$ graus de liberdade.

Caso B: Se H_0' for rejeitada, vamos admitir que as variâncias não são iguais, portanto não tem sentido combinarmos s_1^2 e s_2^2.

Neste caso, utilizaremos para o nosso teste, a variável aleatória:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}},$$

que segue, aproximadamente, a distribuição t de student com n^* graus de liberdade, em que

$$n^* = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}.$$

Observação: Vamos adotar como graus de liberdade, o maior inteiro que não supere o valor calculado.

Decisão: conforme seção 7.7.1.

a) Teste bilateral

Se $|t_{\text{Calculado}}| \geq t_{\text{Tabelado}}$ rejeita-se H_0;

b) Teste unilateral à direita

Se $t_{\text{Calculado}} \geq t_{\text{Tabelado}}$ rejeita-se H_0;

c) Teste unilateral à esquerda

Se $t_{\text{Calculado}} \leq -t_{\text{Tabelado}}$ rejeita-se H_0.

Exercícios

1) Suponhamos que duas técnicas de memorização X_1 e X_2 deverão ser comparadas, medindo-se a eficiência pelo tempo exigido para decorar certo tipo de material. O mesmo material foi apresentado a $n_1 = 18$ e $n_2 = 13$ pessoas que o decoraram através das técnicas X_1 e X_2, respectivamente. Verificar se há diferença significativa entre as duas técnicas de memorização, adotando-se $\alpha = 5\%$. Os resultados foram:

$$\begin{array}{ll} \bar{X}_1 = 20 \text{ min} & \bar{X}_2 = 17 \text{ min} \\ s_1^2 = 12 \text{ min}^2 & s_2^2 = 15 \text{ min}^2 \\ n_1 = 18 & n_2 = 13 \end{array}$$

Resposta:
$H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$
$H_0' : \sigma_2^2 = \sigma_1^2$ *versus* $H_a' : \sigma_2^2 > \sigma_1^2$
$F_{\text{Calculado}} = 1,25; F_{5\%}(12;17) = 2,38$. Não se rejeita H_0'.
$|t_{\text{Calculado}}| = 2,27$, $t_{2,5\%}(29) = 2,045$. Rejeita-se H_0.
Fato: O valor $t_{\text{Tabelado}} = 2,045$ é o quantil de ordem $97,5\%$ da distribuição t de Student. Pode-se utilizar a notação $t_{2,5\%}(29) = 2,045$, ou mesmo $t_{5\%}(29) = 2,045$, o importante é consultar a Tabela corretamente. A notação t_α é bastante utilizada em textos, em qualquer situação.

2) Desejando-se saber se duas rações alimentares X_1 e X_2 para determinada raça de suínos são equivalentes, ou se a ração X_1 é superior à ração X_2 no sentido de causar um maior aumento de peso, a 11 animais sorteados ao acaso foi dada a ração X_1, e a outros 19 a ração X_2. Os resultados foram:

$$\begin{array}{ll} \bar{X}_1 = 66 \text{ kg} & \bar{X}_2 = 63 \text{ kg} \\ s_1^2 = 40 \text{ kg}^2 & s_2^2 = 16 \text{ kg}^2 \\ n_1 = 11 & n_2 = 19 \end{array}$$

A que conclusão chegar se adotarmos um nível de significância $\alpha = 5\%$.

Resposta:
$H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 > \mu_2$
$H_0' : \sigma_1^2 = \sigma_2^2$ *versus* $H_a' : \sigma_1^2 > \sigma_2^2$
$F_{\text{Calculado}} = 2,5; F_{5\%}(10;18) = 2,41$. Rejeita-se H_0'.
$|t_{\text{Calculado}}| = 1,418$, $t_{5\%}(14) = 1,761$. Não se rejeita-se H_0.

7.7.3 Teste de hipótese para o caso de dados emparelhados

Os resultados de duas amostras constituem dados emparelhados quando estão relacionados dois a dois segundo algum critério que introduz uma influência marcante entre os diversos pares, que supomos, influir igualmente sobre os valores de cada par. Por exemplo, medidas sobre o mesmo indivíduo, antes e depois da aplicação de algum medicamento ou uma ração, etc.

Sejam por exemplo:

X_{1i}: representa o peso de determinado animal i antes de receber uma ração,

X_{2i}: representa o peso do mesmo animal i depois que recebeu a ração.

$$d_i = X_{2i} - X_{1i}.$$

Tomando n animais nestas condições, podemos montar a seguinte Tabela:

Nº	X_{1i}	X_{2i}	$d_i = X_{2i} - X_{1i}$
1	X_{11}	X_{21}	d_1
2	X_{12}	X_{22}	d_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	X_{1n}	X_{2n}	d_n

Nesse teste estaremos testando a hipótese de que a diferença entre as médias das duas populações emparelhadas seja igual a um certo valor Δ, o que equivale a testar a hipótese de que a média de todas as diferenças, $\bar{D}$, seja igual a Δ.

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}, \quad \text{é um estimador de } \bar{D}.$$

Por exemplo, para $\bar{D} = 0$, as hipóteses seriam:

$$H_0 : \bar{D} = 0 \quad \text{versus} \quad H_a : \begin{cases} \bar{D} > 0, & \text{ou} \\ \bar{D} < 0, & \text{ou} \\ \bar{D} \neq 0 \end{cases}$$

A estatística do teste é: $t = \dfrac{\bar{d} - \bar{D}}{s\left(\bar{d}\right)}$.

Sob $H_0 : \bar{D} = 0$, teremos

$$t = \frac{\bar{d}}{s\left(\bar{d}\right)}$$

em que

$$s^2(d) = \hat{V}(d) = \frac{SQD(d)}{n-1} = \frac{\sum_{i=1}^{n} d_i^2 - \frac{\left(\sum_{i=1}^{n} d_i\right)^2}{n}}{n-1}$$

$$\hat{V}\left(\bar{d}\right) = \frac{\hat{V}(d)}{n} = \frac{s^2(d)}{n}$$

$$s\left(\bar{d}\right) = \sqrt{\hat{V}\left(\bar{d}\right)} = \frac{s(d)}{\sqrt{n}}.$$

A estatística do teste é:

$$t = \frac{\bar{d}}{s(d)/\sqrt{n}}.$$

Decisão: Note que ao trabalharmos com as n diferenças d_i, o problema será testar uma única média, conforme resolvido em 7.7.1, pela comparação do t de student calculado com o valor tabelado obtido em tabelas em função do α e $n-1$ graus de liberdade.

Exercício:

1) A Tabela a seguir mostra uma sequência de observações sobre os valores das pressões de sete indivíduos antes e depois da aplicação de um medicamento que tem por finalidade a diminuição da pressão arterial. Verificar se o medicamento teve efeito significativo ao nível de 1% de probabilidade.

Indivíduo	Pressão	
	Antes	Depois
1	1,1	0,0
2	3,9	1,2
3	3,1	2,1
4	5,3	2,1
5	5,3	3,4
6	3,4	2,2
7	5,0	3,2

Resposta: $d_i = X_{2i} - X_{1i}$, $H_0 : \bar{D} = 0$ *versus* $H_a : \bar{D} < 0$. $t_{\text{Calculado}} = -5,795$, $-t_{1\%}(6) = -3,143$. Rejeita-se H_0.

7.8 Exercícios propostos com respostas-Lista 1

1) Sabe-se que o consumo mensal per capita de um determinado produto tem distribuição normal, com desvio padrão 2 kg. A diretoria de uma firma que fabrica esse produto resolveu que retiraria o produto da linha de produção se a média de consumo per capita fosse menor que 8 kg. Caso contrário, continuaria a fabricá-lo. Foi realizada uma pesquisa de mercado, tomando-se uma amostra de 25 indivíduos, e verificou-se que a soma dos valores coletados foi de 180 kg. Pede-se:

a) Utilizando um nível de significância de 5%, e com base na amostra colhida determine a decisão a ser tomada pela diretoria.

b) Utilizando um nível de significância de 1%, a decisão seria a mesma? (Justifique a sua resposta.)

2) Estamos desconfiados de que a média das receitas municipais per capita das cidades pequenas (0-20000 habitantes) é maior do que a das receitas do estado, que é de 1229 unidades monetárias. Para comprovar, ou não, esta hipótese, sorteamos dez cidades pequenas, e obtivemos os seguintes resultados: 1230; 582; 576; 2093; 2621; 1045; 1439; 717; 1838; 1359. A que conclusão chegar a um nível de 5% de probabilidade?

3) Deseja-se comparar a qualidade de um produto produzido por duas fábricas. Esta qualidade será definida pela uniformidade com que é produzido o produto por cada fábrica. Tomaram-se duas amostras, uma de cada fábrica, medindo-se o comprimento dos produtos (o resumo dos resultados está na Tabela a seguir).

Estatísticas	Fábrica	
	X_1	X_2
Tamanho da amostra	21	17
Média	21,15	21,12
Variância	0,0412	0,1734

Pede-se:

a) A qualidade das duas fábricas é a mesma, ao nível de 5%? (em termos de variância)

b) Pode-se afirmar que o comprimento médio dos produtos produzidos pelas duas fábricas são iguais, ao nível de 5%?

4) Uma grande cadeia de magazines está interessada em saber se o valor médio das compras é maior em suas lojas do centro da cidade do que no "Shopping center" de certa localidade. O desvio padrão populacional para ambos os casos é de $ 10,00. Teste a afirmação de que ambas são iguais, contra a alternativa de que ambas não são iguais, ao nível de 0,01. Uma amostra aleatória das transações nos dois locais deu os seguintes resultados:

	Centro	"Shopping center"
Média	$ 45,00	$ 43,50
Tamanho da amostra	100	100

5) Uma fábrica de embalagens para produtos químicos está estudando dois processos para combater a corrosão de suas latas especiais. Para verificar o efeito dos tratamentos, foram usadas amostras cujos resultados estão na Tabela a seguir. Qual seria a conclusão sobre os dois tratamentos, ao nível de 5% de significância?

Método	Amostra	Média	Desvio padrão
X_1	15	48	10
X_2	12	52	15

6) Suponhamos que um pesquisador, desejando colocar à prova a hipótese de que a idade da mãe tem certa influência sobre o nascimento de criança prematura, verificou que, dentre 90 casos de prematuridade, 40 envolviam mães com idade inferior a 18 anos; 15 envolviam mães de 18 a 35 anos e 35 mães com idade acima de 35 anos. Isto leva o pesquisador a manter sua hipótese? Use nível de significância de 0,01.

7) No período de um ano, determinada firma teve 50 acidentes. Um dos aspectos de uma investigação levada a efeito pelo engenheiro de segurança diz respeito ao dia de ocorrência do acidente. Pelos dados da Tabela a seguir, pode-se dizer que o dia da semana tenha alguma influência? Teste a hipótese nula, de que os dias são igualmente prováveis, ao nível de 10% de probabilidade.

Dia	Segunda	Terça	Quarta	Quinta	Sexta
Nºde acidentes	15	6	4	9	16

8) A associação dos proprietários de indústrias metalúrgicas está preocupada com o tempo perdido com acidentes de trabalho, cuja média, nos últimos tempos, tem sido da ordem de 60 horas/homem por ano e desvio padrão de 20 horas/homem por ano. Tentou-se um programa de prevenção de acidentes e, após o mesmo, tomou-se uma amostra de 9 indústrias e mediu-se o número médio de horas/homem perdidas por acidente, que foi 50 horas por ano. Você diria, ao nível de 5%, que há evidência de melhoria?

9) Uma firma de produtos farmacêuticos afirma que o tempo médio para certo remédio fazer efeito é de 24 minutos. Numa amostra de 19 casos, o tempo médio foi de 25 minutos, com desvio padrão de 2 minutos. Teste a alegação, contra a alternativa de que o tempo médio é superior a 24 minutos, a um nível de significância de 1%.

10) Uma máquina automática enche latas com base no peso líquido, com variabilidade praticamente constante e independente dos ajustes, dada por um desvio padrão de 5 g. Duas amostras retiradas em dois períodos de trabalho consecutivos, de 10 e de 20 latas, forneceram pesos líquidos médios de, respectivamente, 184,6 e 188,9 gramas. Desconfia-se que a regulagem da máquina quanto ao peso médio fornecido possa ter sido modificada no período entre a coleta das duas amostras. Qual a conclusão?

a) Ao nível de 5% de significância?

b) Ao nível de 1% de significância?

11) Para investigar a influência da opção profissional sobre o salário inicial de recém formados, investigaram-se dois grupos de profissionais: um de liberais em geral e outro de formados em Administração de empresas. Com os resultados abaixo, expressos em salários mínimos, quais seriam suas conclusões ao nível de 5% de significância?

Liberais	6,6	10,3	10,8	12,9	9,2	12,3	7,0		
Administradores	8,1	9,8	8,7	10,0	10,2	10,8	8,2	8,7	10,1

12) Num estudo comparativo do tempo médio de adaptação, uma amostra aleatória, de 50 homens e 50 mulheres de um grande complexo industrial, produziu os seguintes resultados:

Estatísticas	Homens	Mulheres
Média	3,2 anos	3,7 anos
Desvios padrão	0,8 anos	0,9 anos

Que conclusões você poderia tirar para a população de homens e mulheres desta indústria, ao nível de 5% de significância?

13) Foram entrevistados 125 proprietários de certa marca de automóvel acerca do desempenho e do consumo de combustível de seus carros. O resultado da pesquisa de opiniões é resumido na Tabela a seguir:

Consumo	Desempenho		
	Péssimo	Regular	Bom
Alto	29	27	42
Baixo	4	6	17

Verificar, ao nível de 5% de significância, se devemos considerar que, no consenso geral, desempenho e consumo não guardam relação entre si.

14) Uma pesquisa sobre a qualidade de certo produto foi realizada enviando-se questionários a donas-de-casa através do correio. Suspeitando-se que os respondentes voluntários tenham um particular vício de respostas, fizeram-se mais duas tentativas com os não respondentes. Os resultados estão indicados a seguir. Você acha que existe relação entre a opinião e o número de tentativas? (Utilize o nível de significância de 5%)

Opinião	Número de respondentes (donas-de-casas)		
	Tentativas		
	1ª	2ª	3ª
Excelente	62	36	12
Satisfatório	84	42	14
Insatisfatório	24	22	24

15) Uma das maneiras de medir o grau de satisfação dos empregados de uma mesma categoria quanto à política salarial é através do desvio padrão de seus salários. A fábrica X_1 diz ser mais coerente na política salarial do que a fábrica X_2. Para verificar essa afirmação, sorteou-se uma amostra de 10 funcionários não especializados de X_1, e 15 de X_2, obtendo-se os desvios padrão $s_1 = 1,0$ SM (salário médio) e $s_2 = 1,6$ SM. Qual seria a sua conclusão, ao nível de 1%?

16) Dez cobaias adultas foram submetidas ao tratamento com certa ração durante uma semana. Os animais foram perfeitamente identificados e mantidos em gaiolas individuais. Os pesos, em gramas, antes e após uma semana, designados respectivamente, por x_{1i} e x_{2i} são dados a seguir. O uso da ração contribuiu para o aumento do peso médio dos animais? ($\alpha = 1\%$)

Cobaia	1	2	3	4	5	6	7	8	9	10
X_{1i}	635	704	662	560	603	745	698	575	633	669
X_{2i}	640	712	681	558	610	740	707	585	635	682
$d_i = X_{1i} - X_{2i}$										

Respostas dos exercícios propostos-Lista 1

1) $z_{\text{Calculado}} = -2,00$; $H_0 : \mu = 8$ *versus* $H_a : \mu < 8$

a) $z_{5\%} \cong -1,64$, Rejeita-se H_0

b) $z_{1\%} \cong -2,33$, Não se rejeita H_0

2) $H_0 : \mu = 1229$ *versus* $H_a : \mu > 1229$;
$t_{\text{Calculado}} = 0,566$; $t_{5\%}(9) = 1,833$; Não se rejeita H_0

3)a) $H_0 : \sigma_2^2 = \sigma_1^2$ *versus* $H_a : \sigma_2^2 > \sigma_1^2$
$F_{\text{Calculado}} = 4,21$; $F_{5\%}(16;20) = 2,18$; Rejeita-se H_0

3)b) $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$
$|t_{\text{Calculado}}| = 0,272$; $t_{2,5\%}(22) = 2,074$; Não se rejeita H_0

4)$H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 > \mu_2$
$z_{\text{Calculado}} = 1,06$; $z_{1\%} = 2,33$; Não se rejeita H_0

5)$H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$

$H_0' : \sigma_2^2 = \sigma_1^2$ *versus* $H_a' : \sigma_2^2 > \sigma_1^2$

$F_{\text{Calculado}} = 2,25$; $F_{5\%}(11;14) = 2,56$; Não se rejeita H_0'

$|t_{\text{Calculado}}| = 0,829$; $t_{2,5\%}(25) = 2,060$; Não se rejeita H_0

6) H_0 : Proporção=1:1:1 *versus* H_a : Proporção $\neq$ 1:1:1
$\chi^2_{\text{Calculado}} : 11,667$; $\chi^2_{1\%}(2) = 9,210$; Rejeita-se H_0

7) H_0 : Proporção=1:1:1:1:1 *versus* H_a : Proporção $\neq$ 1:1:1:1:1
$\chi^2_{\text{Calculado}} : 11,400$; $\chi^2_{10\%}(4) = 7,779$; Rejeita-se H_0

8) $H_0 : \mu_A = 60$ *versus* $H_a : \mu_A < 60$
$z_{\text{Calculado}} = -1,5$; $z_{5\%} = -1,64$ Não se rejeita H_0

9) $H_0 : \mu_A = 24$ *versus* $H_a : \mu_A > 24$
$t_{\text{Calculado}} = 2,179$; $t_{1\%}(18) = 2,552$; Não se rejeita H_0
$H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$
$z_{\text{Calculado}} = -2,22$

10)a) $z_{2,5\%} = 1,960$; Rejeita-se H_0

10)b) $z_{0,5\%} = 2,57$; Não se rejeita H_0

11) $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$

$H_0' : \sigma_1^2 = \sigma_2^2$ *versus* $H_a' : \sigma_1^2 > \sigma_2^2$

$F_{\text{Calculado}} = 6,13$; $F_{5\%}(6;8) = 3,58$; Rejeita-se H_0'

$|t_{\text{Calculado}}| = 0,481$; $t_{2,5\%}(7) = 2,365$; Não se rejeita H_0

12) $H_0 : \mu_1 = \mu_2$ *versus* $H_a : \mu_1 \neq \mu_2$

$H_0' : \sigma_1^2 = \sigma_2^2$ *versus* $H_a' : \sigma_1^2 > \sigma_2^2$

$F_{\text{Calculado}} = 1,26;\ 1,53 < F_{5\%}(49;49) < 1,69;$ Não se rejeita H_0'

$|t_{\text{Calculado}}| = 2,926;\ 1,98 < t_{2,5\%}(98) < 2,00;$ Rejeita-se H_0

13) H_0: Desempenho e consumo são independentes *versus* H_a: "não H_0"
$\chi^2_{\text{Calculado}} : 3,791;\ \chi^2_{5\%}(2) = 5,991;$ Não se rejeita H_0

14) H_0: Opinião e número de tentativas são independentes *versus* H_a: "não H_0"
$\chi^2_{\text{Calculado}} : 26,2;\ \chi^2_{5\%}(4) = 9,488;$ Rejeita-se H_0

15) $H_0 : \sigma_2^2 = \sigma_1^2$ *versus* $H_a : \sigma_2^2 > \sigma_1^2$
$F_{\text{Calculado}} = 2,56;\ F_{1\%}(14;9) = 5,00;$ Não se rejeita H_0

16) $H_0 : \mu_d = \bar{D} = 0$ *versus* $H_a : \mu_d = \bar{D} < 0$
$|t_{\text{Calculado}}| = 2,96;\ t_{1\%}(9) = 2,821;$ Rejeita-se H_0

7.9 Exercícios propostos com respostas-Lista 2

1) Considere o teste Z para uma média e seja Z_0 o valor da estatística do teste calculado sob a pressuposição de que a hipótese de nulidade é verdadeira,

$$H_0 : \mu = \mu_0 \quad \Rightarrow \quad Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

O valor$-p$ depende do valor z_0 e também se o teste é unilateral ($H_1 : \mu < \mu_0$ ou $H_1 : \mu > \mu_0$) ou bilateral ($H_1 : \mu \neq \mu_0$). Então, se $f(z)$ é a função densidade de probabilidade da variável aleatória Z, para $Z \sim N(0,1)$ e $z^* = |z_0|$ tem-se:

$$\begin{aligned} \text{teste unilateral : valor} - p &= P(Z \geq z^*) = \int_{z^*}^{+\infty} f(z)\, dz; \\ \text{teste bilateral : valor} - p &= 2 \times P(Z \geq z^*) = 2 \times \int_{z^*}^{+\infty} f(z)\, dz. \end{aligned}$$

Informe o valor$-p$ e a decisão do teste quando:

a) $z_0 = -1,99$ e $\alpha = 5\%$, teste bilateral.

b) $z_0 = 1,83$ e $\alpha = 5\%$, teste unilateral.

c) $z_0 = 2,20$ e $\alpha = 1\%$, teste unilateral.

d) $z_0 = -2,20$ e $\alpha = 1\%$, teste bilateral.

2) Considere o teste de qui-quadrado e seja χ_0^2 o valor da estatística do teste sob H_0. Para uma tabela $r \times c$, com r linhas e c colunas, tem-se que $\nu = (r-1)(c-1) - s$ ou $\nu = k - 1 - s$ (k colunas ou tabela $1 \times k$) é o número de graus de liberdade, em que k é o número de classes e s é o número de parâmetros estimados para se obter as frequências esperadas ($s = 0$ em geral). Portanto, valor$-p$=$P(\chi_\nu^2 \geq \chi_0^2)$. Pede-se: informe o valor$-p$ quando,

a) A tabela é 3×4 com $s = 0$ e $\chi_0^2 = 16,812$.

b) A tabela é 1×6 com $s = 2$ e $\chi_0^2 = 9,348$.

3) Com o objetivo de tornar seu produto mais competitivo, um fabricante de automóveis deseja extender a garantia oferecida em peças e serviços de 15000 km para 30000 km. Entretanto, esta mudança somente será viável se for comprovado que os custos não irão aumentar devido à esta extensão na garantia. A garantia atual é com base em estudos que indicam um custo médio de 100 u.m. (u.m.=unidades monetárias) por carro na garantia.

a) Suponha que um estudo preliminar trabalhou com uma amostra de 120 carros cujo custo total devido à serviços cobertos pela garantia tenha sido superior a 100 u.m. por carro. Esta amostra incluía carros na garantia e também fora da garantia e para cada um destes carros havia o registro da quilometragem, de modo que obteve-se quilometragem média igual a 30890 km. Pede-se:

i) A garantia deve ser estendida ou não? Utilize $\sigma = 5836$ km para realizar um teste de hipóteses e escreva as hipóteses estatísticas em termos do problema em questão e também em termos de um teste de hipóteses.

ii) Conclua para $\alpha = 1\%$ e também para $\alpha = 5\%$.

iii) Para averiguar se haverá um aumento no custo médio por carro, devido à extensão na garantia, quais mudanças devem ser realizadas no estudo? explique.

b) Se um novo estudo for realizado conforme descrito em a)iii) considerando-se os valores $n = 120$ e $\sigma = 25$, informe os valores de $\bar{x}$ para se rejeitar $H_0 : \mu = 100$ em favor de $H_1 : \mu > 100$ a 5% e 1%.

4) Suponha que a média da distribuição das estaturas (X) dos adultos seja $1,60$ m com desvio padrão igual a $0,18$ m. Se uma amostra aleatória de 200 adultos fornece média $\bar{X}$, qual deve ser o valor $\bar{x}$ para se declarar que:

a) a média aumentou a 1% de significância.

b) a média se alterou a 5% de significância.

5) Um fabricante informa que um medicamento é 90% eficiente para curar dores de cabeça (explique o significado desta informação). Um estudo com 200 pacientes resultou em 170 pacientes curados. Pede-se: Utilize $\sigma = 0,3$ para testar a informação do fabricante, (informe o valor$-p$) e conclua para $\alpha = 0,01$.

6) Uma máquina está regulada para fornecer $\mu = 500$ g por pacote e seja $\sigma^2 = 25$ g^2. Seja $d = |\bar{X} - \mu|$, calcule o tamanho da amostra para se concluir a 5% de probabilidade que a máquina não está bem regulada quando: $d = 1$; $d = 0,8$; $d = 0,6$; $d = 0,4$ e $d = 0,2$.

7) O número de defeitos nas placas de circuito impresso é suposto seguir a distribuição Poisson. Uma amostra aleatória de 80 circuitos forneceu os dados informados na tabela abaixo.

Número de defeitos	0	1	2	≥ 3
Número de placas	35	22	13	10

Realize um teste a 5% de probabilidade e conclua.

8) Uma amostra aleatória de 300 notas forneceu média $\bar{X} = 56$ e desvio padrão $S_X = 22$. Deseja-se testar a hipótese de que estas notas sejam de uma distribuição normal(μ, σ^2). A tabela a seguir mostra resultados parciais do teste de hipótese, utilizando-se 6 classes de notas equiprováveis.

Classe(i)	notas(X_i)	Frequências	
		observadas (O_i)	esperadas (E_i)
1	$0 \leq X_i < 34,88 (L_1)$	41	
2	$34,88 \leq X_i < 46,54 (L_2)$	55	
3	$46,54 \leq X_i < 56,00 (\hat{\mu})$	59	
4	$56,00 \leq X_i < 65,46 (L_3)$	48	
5	$65,46 \leq X_i < 77,12 (L_4)$	60	
6	$77,12 \leq X_i \leq 100$	37	

Pede-se: Calcule as frequências esperadas e explique como os valores das notas que dividem as classes (L_i) foram calculados, execute o teste e conclua para $\alpha = 0,025$ e também para $\alpha = 0,05$.

9) Utilize a notação da tabela a seguir para demonstrar como se obtém a fórmula,

$$E_{ij} = \hat{n}_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}, \text{ frequências esperadas } (E_{ij}) \text{ nos testes de } \chi^2 \text{ para,}$$

independência: $H_0 : X$ e Y são independentes, e
homogeneidade: H_0 : As populações são homogêneas para as categorias de Y.

População	ou	X	Y				Total
			y_1	y_2	$\cdots$	y_C	
1		x_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
2		x_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
$\vdots$		$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
L		x_L	n_{L1}	n_{L2}	$\cdots$	n_{LC}	$n_{L.}$
		Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	$n_{..}$

10) Para testar a hipótese de que um dispositivo adicionado ao cinto de segurança diminua a força média exercida no peito do passageiro durante o impacto, testou-se 12 carros com e sem o dispositivo. Nestes testes o carro colide com um muro de concreto a uma velocidade de 60 km/h e após a colisão avalia-se a força exercida no peito dos passageiros (bonecos *dummy*), registrada no aparelho apropriado.

a) Com base nos dados da tabela a seguir execute um teste unilateral a 1% de significância e conclua sobre a eficácia do dispositivo em reduzir a força média do impacto em 2 kg.

Dispositivo	Força (kg)					
Sem	5,0	8,0	9,3	10,2	4,8	6,5
Com	2,5	4,0	5,0	7,2	3,2	4,5

b) Se um novo estudo for conduzido com 14 carros ($n = 7$ para cada amostra) e forem obtidos os valores da tabela a seguir, conclua sobre a eficácia do dispositivo em reduzir a força média do impacto em 3 kg, em um teste bilateral a 5%.

Dispositivo	Média	Variância
Sem	8,8	3,25
Com	4,5	2,98

Respostas dos exercícios propostos-Lista 2

1) (Z uma média) A decisão do teste será rejeitar H_0 quando valor$-p \leq \alpha$. 1)a) $2 \times 0,0233 = 0,0466 = 4,67\%$, 1)b) $0,0336 = 3,36\%$, 1)c) $0,0139 = 1,39\%$, 1)d) $0,0278 = 2,78\%$.

2) (χ^2) 2)a) valor$-p = 0,01$, 2)b) valor$-p = 0,025$

3) (Z uma média) 3)a)i) $H_0 : \mu = 30000$, a amostra é de uma população com quilometragem média igual a 30 mil km e portanto os custos irão aumentar com a extensão da garantia, já que nesta amostra o custo foi superior a 100 u.m. por carro. Se $H_1 : \mu > 30000$ e a decisão for rejeitar H_0 (optar por H_1), não se pode concluir quanto ao aumento do custo médio por carro com a extensão da garantia, já que a amostra é de carros com custo superior a 100 u.m. por carro e veio de uma população com quilometragem média superior a 30 mil. Alternativamente, se $H_1 : \mu < 30000$ não há necessidade do teste já que $\bar{X} = 30890$. 3)a)ii) $Z_0 = 1,67$ e valor$-p = 0,0475(4,75\%)$, portanto rejeita-se H_0 a 5% ($Z = 1,64$ ou $1,65$ tabelado) e não se rejeita a 1% ($Z = 2,33$ tabelado) 3)a)iii) trabalharia com uma amostra aleatória de carros com quilometragem maior do que 15 mil e no máximo 30 mil km e com os registros dos custos por carro (dos itens na garantia), para testar $H_0 : \mu = 100$ *versus* $H_1 : \mu > 100$. Neste caso, não rejeitar H_0 significa mudar a garantia, ou extender para 30 mil km, já que os carros amostrados são uma amostra aleatória da nova população de carros que será acrescida à população da garantia atual. 3)b) $\bar{x} \approx 103,8(5\%)$ e $\bar{x} \approx 105,3(1\%)$.

4) (Z uma média)4)a) $\bar{x} \geq 1,629$ 4)b) $\bar{x} \leq 1,575$ ou $\bar{x} \geq 1,625$.

5) (Z uma média) Seja $X_i = 1$, se paciente curou e $X_i = 0$ se paciente não curou, após tomar o medicamento, para $i = 1, 2, \cdots, 200$. Então, $\bar{X} = \widehat{p} = 0,85$. $H_0 : \mu = 0,9$ e a informação do fabricante é verdadeira, $H_1 : \mu < 0.9$, não é verdadeira. $Z_0 \approx -2,36$, portanto valor$-p = 0,5 - 0,4909 = 0,0091(0,91\%)$ e rejeita-se H_0 a 1%.

6) (Z uma média) $n = 96$; $n = 151$; $n = 267$; $n = 600$ e $n = 2401$.

7) (χ^2 aderência) H_0 : O modelo Poisson(m) é adequado e H_1 : Poisson não é adequado. $\hat{m} = 78/80 = 0.975$ defeitos por placa, estimado com os dados ($s = 1$) e utilizando-se $x = 3$ na classe ≥ 3. $E_i = P(x_i) \times 80$, $P(x_i) = e^{-\hat{m}}\hat{m}^{x_i}/x_i!$, para os valores $x_i = 0, 1, 2$ e $P(\geq 3) = 1 - P(0) - P(1) - P(2)$. Obtém-se $\chi_0^2 \approx 5,3$ com $\nu = 4 - 1 - s = 2$ graus de liberdade. A hipótese H_0 não deve ser rejeitada a 5% de significância, $\chi^2(5\%, 2) = 5,991$. Abaixo a solução obtida com o sistema SAS (software estatístico)

x_i	0	1	2	$\geq$3
$P(x_i)$	0,37719	0,36776	0,17928	0,07576
O_i	35	22	13	10
E_i	30,1754	29,4210	14,3427	6,0609
χ_i^2	0,77139	1,87184	0,12570	2,56015

8)(χ^2 aderência) H_0 : o modelo normal(μ, σ^2) é adequado e H_1 : o modelo normal não é adequado. $E_1 = \cdots = E_6 = 300/6 = 50$, $\chi_0^2 = 9,20$ com $\nu = k - 1 - s = 6 - 1 - 2 = 3$

graus de liberdade, pois μ e σ foram estimados por $\bar{X} = 56$ e $S_X = 22$. Os valores L_i indicados na tabela são obtidos por $X = Z\sigma + \mu \implies L_i = Z_i S_X + \bar{X}$, em que os valores Z_i são os correspondentes às classes com probabilidade $1/6 \approx 0,166$, $z_1 = -0,96$; $z_2 = -0,43$; $z_3 = 0,43$; $z_4 = 0,96$, portanto $l_1 = 34,88$; $l_2 = 46,54$; $l_3 = 65,46$; $l_4 = 77,12$. Os valores tabelados são $\chi^2(5\%, 3) = 7,815$ e $\chi^2(2,5\%, 3) = 9,348$, portanto rejeita-se H_0 a 5% e não se rejeita a 2,5%.

9) No teste para independência, $H_0 : \ P(x_i, y_j) = P(x_i)\, P(y_j) \ \ \forall \ \ (x_i, y_j)$,

$$\begin{aligned} P(x_i) &= \tfrac{n_{i\cdot}}{n_{\cdot\cdot}}, \ \text{ probabilidade marginal para } X \\ P(y_j) &= \tfrac{n_{\cdot j}}{n_{\cdot\cdot}}, \ \text{ probabilidade marginal para } Y, \end{aligned}$$

portanto,

$$\frac{n_{ij}}{n_{\cdot\cdot}} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \frac{n_{\cdot j}}{n_{\cdot\cdot}} \Rightarrow \hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}.$$

No teste para homogeneidade $H_0 : P(i, y_j) = \frac{n_{\cdot j}}{n_{\cdot\cdot}} \ \ \forall \ \ (i, y_j)$,

$$P(i, y_j) = \frac{n_{ij}}{n_{i\cdot}}, \ \text{ proporção da categoria } Y_j \text{ na população } i$$

portanto,

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{n_{\cdot\cdot}} \Rightarrow \hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}}.$$

10) (t duas médias) 10)a) Tem-se $H_0 : \mu_{SEM} - \mu_{COM} = 2$ e $H_1 : \mu_{SEM} - \mu_{COM} < 2$. Observe que $\bar{X}_{SEM} - \bar{X}_{COM} = 2,9 \ > \ 2$ e portanto não há nenhuma evidência contrária a H_0 e o teste não é necessário. Observe também que o valor tabelado $t(1\%, 10) = 2,76$ se refere ao valor simétrico à direita da curva, isto é, $P(t_{10} \leq -2,76) = P(t_{10} \geq 2,76) = 0,01$. Apenas por curiosidade, o valor calculado seria,

$$t_0 = \frac{(\bar{X}_{SEM} - \bar{X}_{COM}) - 2}{\sqrt{\left(\frac{S^2_{SEM} + S^2_{COM}}{2}\right)\left(\frac{1}{6} + \frac{1}{6}\right)}} = \frac{(7,3 - 4,4) - 2}{\sqrt{(5,016 + 2,684)/6}} \approx 0,794$$

10)b) tem-se $H_0 : \mu_{SEM} - \mu_{COM} = 3$ e $H_1 : \mu_{SEM} - \mu_{COM} \neq 3$. t tabelado igual a $t_{12}(5\%) = 2,179 = 2,18$ e o calculado igual a,

$$t_0 = \frac{(8,8 - 4,5) - 3}{\sqrt{(3,25 + 2,98)/7}} \approx 1,38$$

A decisão será de não rejeitar H_0. Conclui-se que não há evidências para se suspeitar que o dispositivo seja ainda mais eficiente do que se suspeitava inicialmente, ou seja, ele não diminui a força média em mais do que 3 kg (já que $8,8 - 4,5 = 4,3$).

Capítulo 8

Noções de amostragem

8.1 Introdução

População e amostra: Do ponto de vista estatístico consideramos uma **população** o conjunto de todos os indivíduos (elementos) sobre os quais desejamos desenvolver certos estudos. Quando consideramos apenas uma parte deles, temos o que se denomina **amostra**.

Para fins de Amostragem, certas populações, embora finitas, são consideradas infinitas; isto ocorre sempre que a nossa amostra se constitui de no máximo 5% dos elementos da população.

Na população obtemos os parâmetros verdadeiros e na amostra, suas estimativas. Quando trabalhamos com todos os dados de uma população procedemos a um CENSO, e quando consideramos uma amostra, procedemos a uma AMOSTRAGEM.

Assim sendo, definimos a Amostragem como a parte da Estatística que estuda as populações através de amostras representativas. A dimensão de uma amostra é variável de conformidade com o grau de precisão que desejamos no estudo e, principalmente, com a homogeneidade dos elementos na população.

É de suma importância que definamos qual a nossa **unidade amostral**. Assim, por exemplo, para se estimar o volume de madeira de um povoamento de eucalipto, a unidade é uma árvore, uma parcela com 20 árvores, etc. Quando consideramos um levantamento por amostragem, normalmente levamos em conta que todas as amostras possíveis da população têm a mesma probabilidade de serem selecionadas. Na prática não organizamos todas as amostras possíveis para depois sortear uma delas; isto seria impraticável. O que usualmente fazemos é o sorteio das unidades amostrais, até constituir a amostra propriamente dita.

8.2 Amostra simples ao acaso

É uma amostragem probabilística, onde todos os elementos têm a mesma probabilidade de serem selecionados.

Quando fazemos uma amostragem, geralmente temos em mira:

a) Estimar o valor médio;

b) Estimar o valor total;

c) Estimar a porcentagem ou proporção de ocorrência ou incidência de um determinado fator;

d) Estimar a variância dos dados;

e) Determinar intervalos de confiança do valor médio e do valor total, assim como da porcentagem ou proporção.

Convencionaremos a seguinte notação:

	População	Amostra
Dados:	$X_1, X_2, \cdots, X_N$	$X_1, X_2, \cdots, X_n$
Média:	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
Total	$Y = N\mu = \sum_{i=1}^{N} X_i$	$\widehat{Y} = N\bar{X} = \frac{N}{n}\sum_{i=1}^{n} X_i$

O fator $\frac{N}{n}$ é o "fator de expansão" ou de "crescimento". Através dele é que expandimos resultados da amostra para a população. O seu inverso $\dfrac{n}{N}$ é o "fator de amostragem".

8.2.1 Variâncias

Normalmente a variância de X, em uma população finita, é definida por

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

Vamos apresentar um pouquinho diferente, usando o divisor $N-1$ em lugar de N, conforme Cochran (1977).

Na população temos:

$$S^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1} = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N-1}$$

e, na amostra, sua estimativa:

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

8.2.1.1 Variância da média

$$V\left(\bar{X}\right) = \frac{N-n}{N} \cdot \frac{S^2}{n}$$

Sua estimativa na amostra é dada por: $\hat{V}\left(\bar{X}\right) = \frac{N-n}{N} \cdot \frac{s^2}{n}$.

O termo $\frac{N-n}{N} = 1 - \frac{n}{N}$ é o "fator de correção para população finita". Se n é muito pequeno em relação a N (Quando $n < 0,05N$, não há necessidade de aplicar este fator), esta relação é praticamente 1 e neste caso teremos, como nas populações infinitas:

$$V\left(\bar{X}\right) = \frac{S^2}{n} \text{ e } \hat{V}\left(\bar{X}\right) = \frac{s^2}{n}.$$

Daí se obtém:

$$S\left(\bar{X}\right) = \sqrt{V\left(\bar{X}\right)} = \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}$$

e sua estimativa:

$$s\left(\bar{X}\right) = \sqrt{\hat{V}\left(\bar{X}\right)} = \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}}.$$

A partir da variância da estimativa da média, podemos determinar a variância da estimativa do total, ou seja:

$$V(\hat{Y}) = V(N\bar{X}) = N^2 V\left(\bar{X}\right) = N^2 \left(\frac{N-n}{N} \cdot \frac{S^2}{n}\right).$$

Sua estimativa é:

$$\hat{V}(\hat{Y}) = N^2 \left(\frac{N-n}{N} \cdot \frac{s^2}{n}\right),$$

consequentemente:

$$S(\hat{Y}) = N\, S\left(\bar{X}\right) = N \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{S}{\sqrt{n}}$$

e,

$$s(\hat{Y}) = N\, s\left(\bar{X}\right) = N \cdot \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}}.$$

8.2.1.2 Intervalos de confiança

Uma vez conhecidas as estimativas $s\left(\bar{X}\right)$e $s(\hat{Y})$, podemos determinar os extremos dos intervalos de confiança para a média e para o total a um coeficiente de confiança $1-\alpha$, ou seja:

Para a média:

$$IC(\mu)_{1-\alpha} : \bar{X} \pm t_{\frac{\alpha}{2}} s\left(\bar{X}\right) = \bar{X} \ \pm \ t_{\frac{\alpha}{2}} \sqrt{\frac{N-n}{N}} \cdot \frac{s}{\sqrt{n}}$$

ou

$$IC(\mu)_{1-\alpha} : \bar{X} \pm t_{\frac{\alpha}{2}} s\left(\bar{X}\right) = \bar{X} \ \pm \ t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \qquad \text{(população infinita)}.$$

Para o total:

$$\begin{aligned} IC(Y)_{1-\alpha} : \hat{Y} \pm t_{\frac{\alpha}{2}} s(\hat{Y}) &= N\ \bar{X} \ \pm \ t_{\frac{\alpha}{2}}\, N\, s\left(\bar{X}\right) \\ &= N\left[\bar{X} \pm t_{\frac{\alpha}{2}} s\left(\bar{X}\right).\right] \end{aligned}$$

Observação: $t_{\frac{\alpha}{2}}$ é o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição t de Student obtido em tabelas apropriadas com $n-1$ graus de liberdade. Numa tabela bilateral entra-se diretamente com α e o número de graus de liberdade.

8.3 Dimensionamento da amostra

8.3.1 Amostra simples ao acaso

Para dimensionarmos uma amostra necessitamos de um conhecimento prévio da variância da população ou de sua estimativa, e do grau de precisão desejado.

Sabemos que a variância amostral é dada por:

$$s^2 = \frac{\sum\limits_{i=1}^{n} (X_i - \bar{X})^2}{n-1} = \frac{\sum\limits_{i=1}^{n} X_i^2 - \frac{\left(\sum\limits_{i=1}^{n} X_i\right)^2}{n}}{n-1}.$$

Para população finita de tamanho N e amostras sem reposição, temos que

$$\hat{V}\left(\bar{X}\right) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right),$$

sendo $\left(1 - \frac{n}{N}\right)$ um fator de correção para população finita.

Logo, $s\left(\bar{X}\right) = \sqrt{\hat{V}\left(\bar{X}\right)}$.

Admitamos preliminarmente o caso de uma população infinita, onde:

$$s\left(\bar{X}\right) = \sqrt{\hat{V}\left(\bar{X}\right)} = \frac{s}{\sqrt{n_0}},$$

com n_0 o tamanho da amostra.

Assim, o intervalo de confiança para a média populacional μ fica:

$$\bar{X} \pm t_{\frac{\alpha}{2}} s\left(\bar{X}\right) = \bar{X} \pm t_{\frac{\alpha}{2}} \ \cdot \ \frac{s}{\sqrt{n_0}},$$

onde $t_{\frac{\alpha}{2}}$ é o quantil de ordem $1 - \frac{\alpha}{2}$ da distribuição t de Student obtido em tabelas apropriadas com $n_0 - 1$ graus de liberdade.

Tomemos $d = \dfrac{ts}{\sqrt{n_0}}$, e então:

$$IC(\mu)_{1-\alpha} : \bar{X} \pm d.$$

Desta forma, conforme o grau de precisão desejado teremos o valor de $\underline{d}$. Então:

$$d = \frac{t\,s}{\sqrt{n_0}} \;\Rightarrow\; d^2 = \frac{t^2\,s^2}{n_0} \quad \therefore \quad n_0 = \frac{t^2\,s^2}{d^2}. \tag{8.1}$$

Para população finita teremos:

$$d = \frac{ts}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \frac{ts}{\sqrt{n}} \sqrt{\frac{N-n}{N}}. \tag{8.2}$$

Confrontando (8.1) e (8.2) virá:

$$\frac{t\,s}{\sqrt{n_0}} = \frac{t\,s}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N}},$$

donde obtemos:

$$\frac{1}{n_0} = \frac{N-n}{n \cdot N} \quad \Rightarrow \quad n = n_0 - \frac{n\,n_0}{N}$$

$$n\left(1 + \frac{n_0}{N}\right) = n_0 \quad \therefore \quad n = \frac{n_0}{1 + \frac{n_0}{N}}.$$

Observações

a) Preliminarmente a amostra é dimensionada para a população infinita, obtendo-se o tamanho n_0 e, numa 2^{a} etapa, corrigimos para população finita, obtendo o tamanho $\underline{n}$.

b) Devemos salientar que se n_0 for inferior a 5% do tamanho da população, é desnecessário proceder à correção ou ajuste de n_0. Para fins de amostragem, a população é considerada infinita.

c) Quando não dispomos da variância da população (o que em geral acontece) utilizamos sua estimativa s^2, obtida de alguma situação anterior, geralmente numa amostra prévia.

d) O valor de $\underline{d}$ poderá ser fixado ou tomado como uma fração da estimativa da média, como por exemplo, $d = 0,10\bar{X}$.

Um exemplo ilustrativo

De uma amostra prévia de 10 elementos obtivemos: $\bar{X} = 6,38$ e $s\left(\bar{X}\right) = \dfrac{s}{\sqrt{10}} = 0,45$. Dimensione uma nova amostra nos seguintes casos (usar $\alpha = 5\%$):

i) A população é infinita e o grau de precisão é $d = 0,5$.

ii) A população é infinita e $d = 0,10\,\bar{X}$.

iii) A população é finita ($N = 150$) e $d = 0,075\,\bar{X}$.

Solução:

i) Se $n = 10$, $s\left(\bar{X}\right) = \dfrac{s}{\sqrt{10}} = 0,45 \;\Rightarrow\; s = 1,42$

Da Tabela t, para $\alpha = 5\%$ e 9 graus de liberdade: $t = 2,2622$. Assim, virá:

$$n = n_0 = \frac{t^2\, s^2}{d^2} \;=\; \frac{(2,2622)^2\,(1,42)^2}{(0,5)^2} \;=\; 41,3 \;\cong\; 42 \text{ elementos.}$$

ii) $d = 0,10\,\bar{X} = 0,10\,(6,38) = 0,638$

$$n = n_0 = \frac{(2,2622)^2\,(1,42)^2}{(0,638)^2} = 25,3 \;\cong 26 \text{ elementos.}$$

iii) $d = 0,075\,\bar{X} = 0,075(6,38) = 0,4785$

$$n_0 = \frac{(2,2622)^2\,(1,42)^2}{(0,4785)^2} = 45,07 \;\cong\; 46 \text{ elementos.}$$

Corrigindo para população finita:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} \;=\; \frac{46}{1 + \frac{46}{150}} = 35,2 \;\cong\; 36 \text{ elementos.}$$

Observação: Se o tamanho da amostra prévia (amostra piloto) for maior que o tamanho da nova amostra, ela é suficiente, caso contrário, devemos completar para o $\underline{n}$ dimensionado.

Outro exemplo

Um povoamento de eucalipto, de 10 ha de área, foi subdividido em parcelas de 400 m^2. Com o fim de se estimar o volume de madeira desse povoamento, foi colhida uma amostra ao acaso de 25 parcelas. Procedido ao corte das árvores, obteve-se o seguinte resultado em esteres/parcela:

11,2	8	8,5	10,5	11,3
11,5	7,5	10	14,5	12
10	12	11,2	12,6	10,5
11	11	11,3	11,3	9
9	10	11,5	11,5	10,5

Pede-se:

a) Calcular a média da amostra

$$\begin{aligned} \bar{X} &= \frac{\sum_{i=1}^{n} X_i}{n} = \frac{11,2 + 11,5 + \cdots, + 9,0 + 10,5}{25} = \frac{267,4}{25} \\ &= 10,7 \text{ esteres/parcela.} \end{aligned}$$

b) Calcular o erro padrão da média

Cálculo de N:

$$\begin{array}{rcl} 1 \text{ parcela} & \text{———} & 400 \text{ m}^2 \\ 25 \text{ parcelas} & \text{———} & x \end{array}$$

e, efetuando-se a regra de três chegamos a $x = 10000 \text{ m}^2 = 1$ ha.

$$\begin{array}{rcl} 25 \text{ parcelas} & \text{———} & 1 \text{ ha} \\ N & \text{———} & 10 \text{ ha} \end{array}$$

$N = 250$ (População Finita) e $n = 25 > 0,05N$.

$$s^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1} = \frac{2914,96 - \frac{(267,4)^2}{25}}{24} = 2,29$$

$$\hat{V}\left(\bar{X}\right) = \frac{s^2}{n} \cdot \frac{N-n}{N} = \frac{2,29}{25} \cdot \frac{250-25}{250} = 0,0824$$

$$s\left(\bar{X}\right) = \sqrt{\hat{V}\left(\bar{X}\right)} = \sqrt{0,0824} = 0,28 \text{ esteres/parcela.}$$

c) Estimar o volume do povoamento

O total da população é dado por $Y = N\mu$ e a estimativa é dada por $\hat{Y} = N\bar{X}$. Logo,

$$\hat{Y} = N\bar{X} = 250\,(10,7) = 2675 \text{ esteres.}$$

d) Obter o intervalo de confiança para a média da população, para $\alpha = 5\%$.

$$\text{I.C.}_{1-\alpha}\,(\mu) : \bar{X} \pm t_{\frac{\alpha}{2}}\, s\left(\bar{X}\right)$$

O valor de t correspondente a 24 graus de liberdade e $\alpha = 5\%$ é 2,0639. Assim, virá:

$$\begin{array}{rcl} \text{I.C.}(\mu)_{95\%} & : & 10,7 \pm 2,0639\,(0,28) \\ & : & 10,7 \pm 0,6 \\ & & 10,1 \leq \mu \leq 11,3. \end{array}$$

e) Obter o intervalo de confiança para o total da população adotando $\alpha = 5\%$.

Vimos que $\hat{Y} = N\,\bar{X}$. Logo, $\hat{V}(\hat{Y}) = N^2\hat{V}\left(\bar{X}\right)$ e $s(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}$

$$\text{I.C.}(Y)_{1-\alpha} : \hat{Y} \pm t_{\frac{\alpha}{2}} s(\hat{Y})$$

$$\hat{Y} = 2675$$

$$\alpha = 5\% \text{ e } 24 \text{ graus de liberdade } \Rightarrow t = 2,0639$$

$$\hat{V}(\hat{Y}) = (250)^2 (0,0824) = 5150$$

$$s(\hat{Y}) = \sqrt{5150} = 71,76$$

$$\begin{aligned} \text{I.C.}(Y)_{95\%} &: 2675 \pm 2,0639\,(71,76) \\ &: 2675 \pm 148,1 \\ &\quad 2526,9 \leq Y \leq 2823,1 \end{aligned}$$

f) Dimensione uma nova amostra supondo $d = 10\%$ da média e $\alpha = 5\%$.

$$d = 0,10\,\bar{X} = 0,10\,(10,7) = 1,07$$

$$\alpha = 5\% \text{ e } 24 \text{ graus de liberdade } \Rightarrow t = 2,0639$$

$$s^2 = 2,29$$

$$n_0 = \frac{t^2 \cdot s^2}{d^2} = \frac{(2,0639)^2\,(2,29)}{(1,07)^2} = 8,52 \cong 9 \text{ elementos.}$$

Como $n_0 < 0,05N$, a correção para população finita é desnecessária.

Assim, $n = 9$ elementos.

Com base neste dimensionamento ($n = 9$ elementos), podemos concluir que a amostra colhida ($n = 25$ elementos) é suficiente, logo, ela é representativa da população.

8.4 Amostra simples ao acaso para proporções ou porcentagens

Frequentemente num levantamento, desejamos determinar a proporção de unidades da população que possuem uma determinada característica.

Na maioria das vezes, este atributo pode ser levantado pela resposta "SIM" ou "NÃO" à determinada pergunta.

Os dados da população se enquadram em duas categorias C e C_1, a saber:

C: dados que possuem o atributo;

C_1: dados que não possuem o atributo.

A seguinte notação será empregada:

A=número de unidades de C na população;

a=número de unidades de C na amostra;

N=Tamanho da população;

n=tamanho da amostra;

$P = \frac{A}{N}$ = proporção de unidades de C na população;

$p = \frac{a}{n}$ = proporção de unidades de C na amostra.

Verificamos facilmente que p é uma estimativa de P e Np é uma estimativa de A. A fim de quantificar os resultados, cada unidade X_i da população ou da amostra valerá:

$$X_i = \begin{cases} 1, & \text{se } \in \text{ a classe } C \\ 0, & \text{se } \in \text{ a classe } C_1 \end{cases}$$

Assim teremos:

i) Na população

$$A = \sum_{i=1}^{N} X_i$$

$$\text{Média: } \mu = \frac{A}{N} = P$$

$$\text{Total: } A = NP$$

$$\text{Variância: } S^2 = \frac{\sum_{i=1}^{N} X_i^2 - \frac{\left(\sum_{i=1}^{N} X_i\right)^2}{N}}{N-1} = \frac{A - \frac{A^2}{N}}{N-1} = \frac{A\left(1 - \frac{A}{N}\right)}{N-1}$$

Tomando $A = NP$, vem:

$$S^2 = \frac{NP\,(1-P)}{N-1} = \frac{NPQ}{N-1}.$$

ii) Na amostra

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

$$\text{Média: } \bar{X} = \frac{a}{n} = p$$

Estimativa para o total na população: $\hat{A} = Np$.

$$s^2 = \frac{npq}{n-1}$$

$$V(p) = \frac{S^2}{n}\left(1 - \frac{n}{N}\right) = \frac{N-n}{N-1} \cdot \frac{PQ}{n},$$

onde $1 - \frac{n}{N}$ = fator de correção para população finita (Quando $n < 0,05N$ não há necessidade de aplicar este fator).

$$\hat{V}(p) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right) = \frac{N-n}{N(n-1)} \cdot pq$$

$$s(p) = \sqrt{\hat{V}(p)}$$

$$V\left(\hat{A}\right) = N^2 V(p), \qquad \hat{V}(\hat{A}) = N^2 \hat{V}(p), \qquad s(\hat{A}) = \sqrt{\hat{V}\left(\hat{A}\right)}.$$

O intervalo de confiança para P é:

$$I.C.(P)_{1-\alpha} : p \pm t_{\alpha/2} \cdot s(p).$$

O intervalo de confiança para A é:

$$I.C.(A)_{1-\alpha} : \hat{A} \pm t_{\alpha/2} \cdot s\left(\hat{A}\right).$$

8.5 Dimensionamento da amostra

8.5.1 Amostra simples ao acaso

Nossa pergunta seria: qual o tamanho da amostra para termos um determinado grau de precisão nas nossas estimativas?

Temos que:

$$I.C.(P)_{1-\alpha} : p \pm t_{\alpha/2} s(p) = p \pm d$$

e então:

$$d = t\sqrt{\frac{N-n}{N}} \cdot \sqrt{\frac{pq}{n-1}}.$$

Para **população infinita**, teremos:

$$d = t\sqrt{\frac{pq}{n'-1}}$$

e assim,

$$d^2 = \frac{t^2 pq}{n'-1}$$

e, finalmente:

$$n' = \frac{t^2 pq}{d^2} + 1$$

e, se tomarmos:

$$n_0 = \frac{t^2 pq}{d^2}, \text{ virá: } n' = n_0 + 1.$$

Para **população finita**, teremos:

$$d^2 = t^2 \frac{N-n}{N} \cdot \frac{pq}{n-1}$$

e daí:

$$\frac{d^2}{t^2 pq} = \left(1 - \frac{n}{N}\right)\frac{1}{n-1}$$

ou, desde que $n_0 = \dfrac{t^2pq}{d^2}$,

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

ou ainda:

$$\begin{aligned} \frac{n-1}{n_0} &= 1 - \frac{n}{N} \\ \frac{n}{n_0} + \frac{n}{N} &= 1 + \frac{1}{n_0} \\ n\left(\frac{1}{n_0} + \frac{1}{N}\right) &= 1 + \frac{1}{n_0} \end{aligned}$$

e, finalmente:

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

Sendo $n' = n_0 + 1$, virá:

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

ou, aproximadamente:

$$n = \frac{n'}{1 + \frac{n'}{N}}.$$

Exemplo ilustrativo

Numa Empresa de Celulose e Papel com 1600 funcionários foi tomada uma amostra simples ao acaso de 200 deles, verificando-se que apenas 32 eram fumantes. Pede-se:

a) Estimar a proporção e o número total de fumantes e seus respectivos intervalos de confiança ($\alpha = 0,05$).

Temos que $N = 1600$, $n = 200$, $a = 32$. Consequentemente:

$$p = \frac{a}{n} = \frac{32}{200} = 0,16$$

e,

$$q = 1 - p = 0,84$$

$$s^2 = \frac{npq}{n-1} = \frac{200\,(0,16)\,(0,84)}{199} = 0,1351$$

$$s = 0,37$$

$$s\,(p) = \sqrt{\frac{N-n}{N\,(n-1)}pq} = \sqrt{\frac{1600-200}{1600\,(199)}\,(0,16)\,(0,84)} = 0,0243$$

$$\hat{V}\,(p) = 0,000591.$$

Da Tabela t, para α=5% e 199 graus de liberdade: $t \cong 1,98$.

$$\begin{aligned} I.C.(P)_{95\%} &: 0,16 \pm 1,98\,(0,0243) \\ &: 0,16 \pm 0,05, \end{aligned}$$

ou seja:

$$0,11 \leq P \leq 0,21$$

$$\hat{A} = Np = 1600\,(0,16) = 256$$

$$\hat{V}\left(\hat{A}\right) = N^2\,\hat{V}\,(p) = (1600)^2.0,000591 = 1512,96$$

$$s\left(\hat{A}\right) = \sqrt{1512,96} = 38,89$$

$$\begin{aligned} I.C.(A)_{95\%} &: 256 \pm 1,98\,(38,89) \\ &: 256 \pm 77 \end{aligned}$$

ou seja:

$$179 \leq A \leq 333.$$

b) Dimensione uma nova amostra admitindo-se $d = 0,10p$ e $\alpha = 5\%$.

$$N = 1600, \quad p = 0,16, \quad q = 0,84, \quad t = 1,98, \quad d = 0,10(0,16) = 0,016.$$

Assim,

$$n' = \frac{t^2pq}{d^2} + 1 = \frac{(1,98)^2\,(0,16)\,(0,84)}{(0,016)^2} + 1 = 2059$$

e daí:

$$n = \frac{n'}{1 + \frac{n'}{N}} = \frac{2059}{1 + \frac{2059}{1600}} = 900.$$

Se tivéssemos fixado, por exemplo, $d = 0,05$ viria:

$$\begin{aligned} n' &= 212, \\ n &= 187. \end{aligned}$$

Existem outros tipos de amostragem, como por exemplo, Amostragem Estratificada, Amostragem Sistemática, Estimativas por Razão, Estimativas por Regressão, etc., que não serão abordadas aqui.

8.6 Exercício resolvido

Com o objetivo de dimensionar o tamanho da amostra (número de observações) para avaliar a viscosidade (cP), tomou-se uma amostra prévia de 10 madeiras de clones de ***Eucalyptus***, cujos resultados estão apresentados a seguir:

42,6 38,7 31,1 45,0 40,8 40,1 36,0 50,7 45,7 41,3

Considerando os dados acima, pede-se:

1) Estimativa da média;

2) Intervalo de confiança para a média ao nível de confiança de 95%;

3) Dimensionar uma nova amostra considerando o grau de precisão desejado igual a 5% da média e α=1%.

Respostas:

1) $\bar{X} = \frac{42,6+38,7+\cdots+41,3}{10} = \frac{412}{10} = 41,2.$

2)

$$\begin{aligned} IC(\mu)_{95\%} &: \quad 41,2 \pm 2,2622 \cdot \frac{5,41}{\sqrt{10}} \\ &: \quad 41,2 \pm 3,9 \\ &\quad\; 37,3 \leq \mu \leq 45,1. \end{aligned}$$

3) $n = \frac{(3,2498)^2(29,29)}{(2,06)^2} = 72,89 \cong 73$ elementos.

Neste caso, devemos completar a amostra prévia (amostra piloto) para o n dimensionado. Tomaríamos mais 63 unidades amostrais para termos uma amostra que, dentro dos critérios previamente estabelecidos, seja representativa da população.

8.7 Exercício proposto

Com o objetivo de dimensionar o tamanho da amostra para avaliar o teor de extrativos (%), tomou-se uma amostra prévia de 8 madeiras de *Eucalyptus*, cujos resultados são apresentados a seguir:

1,3 3,5 2,4 3,2 4,2 2,6 3,8 3,0.

Considerando os dados acima, pede-se:

1) Estimativa da média; (R: $\bar{X} = 3$)

2) Dimensionar uma nova amostra considerando o grau de precisão desejado igual a 5% da média e $\alpha = 1\%$. (R: $n \approx 450$)

Referências

BUSSAB, W. O.; MORETTIN, P. A. **Estatística básica.** 7ª ed. São Paulo, SP: Atual, 2011. 540 p.

COCHRAN, W. G. Sampling techniques. 3 ed. New York, NY: John Wiley and Sons, 1977. 428 p.

COSTA NETO, P. L. O. **Estatística.** São Paulo, SP: Edgard Blucher, 1977. 264 p.

FALCONER, D. S. **Introduction to quantitative genetics**. 3 ed. New York, NY: Longmans Green, 1989. 438 p.

FONSECA, J. S.; MARTINS, G. A. **Curso de estatística.** 6ª ed., São Paulo, SP: Atlas, 2008. 320 p.

FREUND, J. E.; SIMON, G. A. **Estatística aplicada.** 9ª ed. Porto Alegre, RS: Bookman, 2000. 404 p.

GATTÁS, R. R. **Elementos de probabilidade e inferência.** São Paulo, SP: Atlas, 1978. 267 p.

KOLMOGOROV, A. N. **Grundbegriffe der wahrscheinlichkeitsrechnung**. Springer, Berlin, 1933. 62 p.

MEYER, P. L. **Probabilidade:** aplicações à estatística. 2ª ed. Rio de Janeiro, RJ: Livros técnicos e científicos, 2006. 426 p.

MORETTIN, L. G. **Estatística básica:** probabilidade e inferência. São Paulo, SP: Makron Books, 2010. 390 p.

PIMENTEL-GOMES, F. **Curso de estatística experimental.** 15ª ed. Piracicaba, SP: ESALQ, 2009. 451 p.

PINHEIRO, J. I. D.; CARVAJAL, S. S. R.; CUNHA, S. B.; GOMES, G. C. **Probabilidade e estatística:** quantificando a incerteza. Rio de Janeiro, RJ: Elsevier, 2012. 568 p.

SATTERTHWAITE, F. E. An approximate distribution of estimates of variance components. **Biometrics Bulletin**, v. 2, n. 6, p. 110-114, 1946.

SPIEGEL, M. R. **Probabilidade e estatística.** Coleção Schaum. São Paulo, SP: McGraw-Hill do Brasil, Ltda, 1978. 528 p.

SPIEGEL, M. R. **Estatística.** 3ª ed. São Paulo: Makron books, 1994. 644 p.

TRIOLA, M. F. **Introdução à estatística.** 11ª ed. Rio de Janeiro: Livros técnicos e científicos, 2013. 740 p.

WELCH, B. L. The generalization of "student's" problem when several different population variances are involved. **Biometrika**, v. 34, n. 1, p. 28-35, 1947.

Apêndice A

Tabelas estatísticas

Tabela A.1: Áreas de uma distribuição normal padrão entre $z = 0$ e um valor positivo de z. As áreas para os valores de z negativos são obtidas por simetria

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2703	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4006	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Adaptada de Costa Neto, P. L. O. (1977) Estatística, Editora Edgard Blucher.

Tabela A.2: Valores χ^2 na distribuição de qui-quadrado com n graus de liberdade tais que $P\left(\chi_n^2 \geq \chi^2\right) = p \times 100\%$

n	p=99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	n
1	$0{,}0^3$ 16	$0{,}0^3$ 63	0,001	0,004	0,016	0,064	0,148	0,455	1,074	1,642	2,706	3,841	4,218	5,024	5,412	6,635	9,550	10,827	1
2	0,020	0,040	0,051	0,103	0,211	0,446	0,713	1,386	2,408	3,219	4,605	5,991	6,438	7,378	7,824	9,210	12,429	13,815	2
3	0,115	0,185	0,216	0,352	0,584	1,005	1,424	2,366	3,665	4,642	6,251	7,815	8,311	9,348	9,837	11,345	14,796	16,266	3
4	0,297	0,429	0,484	0,711	1,064	1,649	2,195	3,357	4,878	5,989	7,779	9,488	10,026	11,143	11,668	13,277	16,924	18,467	4
5	0,554	0,752	0,831	1,145	1,610	2,343	3,000	4,351	6,064	7,289	9,236	11,070	11,644	12,832	13,388	15,086	18,907	20,515	5
6	0,872	1,134	1,237	1,635	2,204	3,070	3,828	5,348	7,231	8,558	10,645	12,592	13,198	14,449	15,033	16,812	20,791	22,457	6
7	1,239	1,564	1,690	2,167	2,833	3,822	4,671	6,346	8,383	9,803	12,017	14,067	14,703	16,013	16,622	18,475	22,601	24,322	7
8	1,646	2,032	2,180	2,733	3,490	4,594	5,527	7,344	9,524	11,030	13,362	15,507	16,171	17,534	18,168	20,090	24,352	26,125	8
9	2,088	2,532	2,700	3,325	4,168	5,380	6,393	8,343	10,656	12,242	14,684	16,919	17,608	19,023	19,679	21,666	26,056	27,877	9
10	2,558	3,059	3,247	3,940	4,865	6,179	7,267	9,342	11,781	13,442	15,987	18,307	19,021	20,483	21,161	23,209	27,722	29,588	10
11	3,053	3,609	3,816	4,575	5,578	6,989	8,148	10,341	12,899	14,631	17,275	19,675	20,412	21,920	22,618	24,725	29,354	31,264	11
12	3,571	4,178	4,404	5,226	6,304	7,807	9,034	11,340	14,011	15,812	18,549	21,026	21,785	23,337	24,054	26,217	30,957	32,909	12
13	4,107	4,765	5,009	5,892	7,042	8,634	9,926	12,340	15,119	16,985	19,812	22,362	23,142	24,736	25,472	27,688	32,535	34,528	13
14	4,660	5,368	5,629	6,571	7,790	9,467	10,821	13,339	16,222	18,151	21,064	23,685	24,485	26,119	26,873	29,141	34,091	36,123	14
15	5,229	5,985	6,262	7,261	8,547	10,307	11,721	14,339	17,322	19,311	22,307	24,996	25,816	27,488	28,259	30,578	35,628	37,697	15
16	5,812	6,614	6,908	7,962	9,312	11,152	12,624	15,338	18,418	20,465	23,542	26,296	27,136	28,845	29,633	32,000	37,146	39,252	16
17	6,408	7,255	7,564	8,672	10,085	12,002	13,531	16,338	19,511	21,615	24,769	27,587	28,445	30,191	30,995	33,409	38,648	40,790	17
18	7,015	7,906	8,231	9,390	10,865	12,857	14,440	17,338	20,601	22,760	25,989	28,869	29,745	31,526	32,346	34,805	40,136	42,312	18
19	7,633	8,567	8,906	10,117	11,651	13,716	15,352	18,338	21,689	23,900	27,204	30,144	31,037	32,852	33,687	36,191	41,610	43,820	19
20	8,260	9,237	9,591	10,851	12,443	14,578	16,266	19,337	22,775	25,038	28,412	31,410	32,321	34,170	35,020	37,566	43,072	45,315	20
21	8,897	9,915	10,283	11,591	13,240	15,445	17,182	20,337	23,858	26,171	29,615	32,671	33,597	35,479	36,343	38,932	44,522	46,797	21
22	9,542	10,600	10,982	12,338	14,041	16,314	18,101	21,337	24,939	27,301	30,813	33,924	34,867	36,781	37,659	40,289	45,962	48,268	22
23	10,196	11,293	11,688	13,091	14,848	17,187	19,021	22,337	26,018	28,429	32,007	35,172	36,131	38,076	38,968	41,638	47,391	49,728	23
24	10,856	11,992	12,401	13,848	15,659	18,062	19,943	23,337	27,096	29,553	33,196	36,415	37,389	39,364	40,270	42,980	48,812	51,179	24
25	11,524	12,697	13,120	14,611	16,473	18,940	20,867	24,337	28,172	30,675	34,382	37,652	38,642	40,646	41,566	44,314	50,223	52,620	25
26	12,198	13,409	13,844	15,379	17,292	19,820	21,792	25,336	29,246	31,795	35,563	38,885	39,889	41,923	42,856	45,642	51,627	54,052	26
27	12,879	14,125	14,573	16,151	18,114	20,703	22,719	26,336	30,319	32,912	36,741	40,113	41,132	43,194	44,140	46,963	53,022	55,476	27
28	13,565	14,847	15,308	16,928	18,939	21,588	23,647	27,336	31,319	34,027	37,916	41,337	42,370	44,461	45,419	48,278	54,411	56,893	28
29	14,256	15,574	16,047	17,708	19,768	22,475	24,577	28,336	32,461	35,139	39,087	42,557	43,604	45,722	46,693	49,588	55,792	58,302	29
30	14,953	16,306	16,791	18,493	20,599	23,364	25,508	29,336	33,530	36,250	40,256	43,773	44,834	46,979	47,962	50,892	57,167	59,703	30
n	p=99%	98%	97,5%	95%	90%	80%	70%	50%	30%	20%	10%	5%	4%	2,5%	2%	1%	0,2%	0,1%	n

Adaptada de Bussab, W. O. e Morettin, P. A. (2011) Estatística Básica-Métodos Quantitativos, Editora Atual.

Tabela A.3: Valores positivos t na distribuição t_n de Student com n graus de liberdade em níveis de 10% a 0,1% de probabilidade $= 2 \times P\,(t_n \geq t)$, tabela bilateral

n	nível de probabilidade bilateral					
	10%	5%	2%	1%	0,5%	0,1%
1	6,31	12,71	31,82	63,66	127,32	636,62
2	2,92	4,30	6,97	9,92	14,09	31,60
3	2,35	3,18	4,54	5,84	7,45	12,94
4	2,13	2,78	3,75	4,60	5,60	8,61
5	2,02	2,57	3,37	4,03	4,77	6,86
6	1,94	2,45	3,14	3,71	4,32	5,96
7	1,90	2,36	3,10	3,50	4,03	5,41
8	1,86	2,31	2,90	3,36	3,83	5,04
9	1,83	2,26	2,82	3,25	3,69	4,78
10	1,81	2,23	2,76	3,17	3,58	4,59
11	1,80	2,20	2,72	3,11	3,50	4,44
12	1,78	2,18	2,68	3,06	3,43	4,32
13	1,77	2,16	2,65	3,01	3,37	4,22
14	1,76	2,14	2,62	2,98	3,33	4,14
15	1,75	2,13	2,60	2,95	3,29	4,07
16	1,75	2,12	2,58	2,92	3,25	4,02
17	1,74	2,11	2,57	2,90	3,22	3,97
18	1,73	2,10	2,55	2,88	3,20	3,92
19	1,73	2,09	2,54	2,86	3,17	3,88
20	1,73	2,09	2,53	2,84	3,15	3,85
21	1,72	2,08	2,52	2,83	3,14	3,82
22	1,72	2,07	2,51	2,82	3,12	3,79
23	1,71	2,07	2,50	2,81	3,10	3,77
24	1,71	2,06	2,49	2,80	3,09	3,75
25	1,71	2,06	2,49	2,79	3,08	3,73
26	1,71	2,06	2,48	2,78	3,07	3,71
27	1,70	2,05	2,47	2,77	3,06	3,69
28	1,70	2,05	2,47	2,76	3,05	3,67
29	1,70	2,04	2,46	2,76	3,04	3,66
30	1,70	2,04	2,46	2,75	3,03	3,65
40	1,68	2,02	2,42	2,70	2,97	3,55
60	1,67	2,00	2,39	2,66	2,92	3,46
120	1,65	1,98	2,36	2,62	2,86	3,37
$+\infty$	1,65	1,96	2,33	2,58	2,81	3,29

Adaptada de Frederico Pimentel Gomes (2009), Curso de Estatística Experimental, 12ª ed.

Tabela A.4: Valores de F ao nível de 1% de probabilidade. $P\left(F_{\nu_1,\nu_2} \geq F\right) = 0,01$. Número de graus de liberdade: ν_1 = numerador e ν_2 = denominador

	n_1																						
n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20	24	30	40	60	120	$+\infty$
1	4052	5000	5403	5625	5764	5859	5928	5982	6022	6056	6082	6106	6125	6142	6157	6169	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,41	99,42	99,42	99,43	99,43	99,44	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,13	27,05	26,98	26,92	26,87	26,83	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,30	14,24	14,20	14,15	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,83	9,77	9,72	9,68	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,66	7,60	7,56	7,52	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	8,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,41	6,35	6,31	6,27	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,74	5,67	5,61	5,56	5,52	5,48	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	5,05	5,00	4,96	4,92	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,78	4,71	4,65	4,60	4,56	4,52	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,34	4,29	4,25	4,21	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,10	4,05	4,01	3,98	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,90	3,85	3,82	3,78	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,75	3,70	3,66	3,62	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,61	3,56	3,52	3,48	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55	3,50	3,45	3,41	3,37	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,40	3,35	3,31	3,27	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,44	3,37	3,32	3,27	3,23	3,19	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,24	3,19	3,15	3,12	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,30	3,23	3,18	3,13	3,09	3,05	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,12	3,07	3,03	2,99	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	3,07	3,02	2,98	2,94	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,14	3,07	3,02	2,97	2,93	2,89	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,98	2,93	2,89	2,85	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,05	2,99	2,94	2,89	2,85	2,81	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,91	2,86	2,81	2,77	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,98	2,93	2,88	2,83	2,78	2,74	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,95	2,90	2,85	2,80	2,75	2,71	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,92	2,87	2,82	2,77	2,73	2,68	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,90	2,84	2,79	2,74	2,70	2,66	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,61	2,56	2,52	2,49	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,45	2,40	2,35	2,32	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40	2,34	2,29	2,24	2,19	2,16	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	2,18	2,12	2,07	2,04	1,99	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Adaptada de Bussab, W. O. e Morettin, P. A. (2011) Estatística Básica-Métodos Quantitativos, Editora Atual.

Tabela A.5: Valores de F ao nível de 5% de probabilidade. $P\left(F_{\nu_1,\nu_2} \geq F\right) = 0,05$. Número de graus de liberdade: ν_1 = numerador e ν_2 = denominador

	n_1																						
n_2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	244,4	245,0	245,9	246,0	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,72	8,71	8,70	8,69	8,66	8,64	8,62	9,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,89	5,87	5,86	5,84	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,52	3,51	3,49	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,25	3,23	3,22	3,20	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,04	3,02	3,01	2,98	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,88	2,86	2,85	2,82	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,57	2,55	2,53	2,51	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,56	2,53	2,50	2,48	2,46	2,44	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,43	2,40	2,39	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,39	2,37	2,35	2,33	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,33	2,31	2,29	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,29	2,27	2,25	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,23	2,20	2,18	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,20	2,18	2,15	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,18	2,15	2,13	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,17	2,14	2,13	2,10	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,13	2,11	2,09	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,13	2,11	2,09	2,06	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	2,10	2,07	2,05	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,16	2,13	2,10	2,08	2,06	2,03	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	2,06	2,04	2,02	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,07	2,05	2,03	2,00	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,12	2,09	2,06	2,04	2,01	1,99	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,95	1,92	1,90	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,89	1,86	1,84	1,81	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,86	1,83	1,80	1,77	1,75	1,73	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,72	1,69	1,67	1,64	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Adaptada de Bussab, W. O. e Morettin, P. A. (2011). Estatística Básica-Métodos Quantitativos, Editora Atual.