

Universidade Federal de Santa Catarina  
Centro Tecnológico  
Departamento de Informática e Estatística

**ESTATÍSTICA PARA AS CIÊNCIAS AGRÁRIAS E  
BIOLÓGICAS COM NOÇÕES DE EXPERIMENTAÇÃO**

Prof. Dalton Francisco de Andrade  
Prof. Paulo José Ogliari

FLORIANÓPOLIS  
Santa Catarina - Brasil  
Janeiro - 2006

# Sumário

<b>1</b>	<b>Introdução Geral</b>	<b>6</b>
1.1	A Estatística na Pesquisa Agrária e Biológica . . . . .	6
1.1.1	Forma de coleta dos dados . . . . .	6
1.1.2	Erro Experimental . . . . .	7
1.1.3	Análise Exploratória de Dados . . . . .	8
1.1.4	Inferência Estatística . . . . .	9
1.1.5	Teoria da Probabilidade e seus Modelos . . . . .	13
1.2	A Estatística e o Método Científico . . . . .	14
1.3	Aspectos do Planejamento Estatístico de um Experimento . . . . .	20
1.3.1	Experimentos com um fator e mais de um fator . . . . .	21
1.3.2	Tratamento Controle . . . . .	22
1.3.3	Variáveis e Covariáveis . . . . .	23
1.3.4	Repetição e Casualização . . . . .	23
1.4	Exercícios Propostos . . . . .	24
<b>2</b>	<b>Análise Exploratória de Dados</b>	<b>26</b>
2.1	Organização, Resumo e Representação de Dados . . . . .	26
2.1.1	Introdução . . . . .	26
2.1.2	Classificação das Variáveis Seleccionadas para o Estudo . . . . .	27
2.1.3	Distribuições de Frequências. Representação em Tabelas e Gráficos . . . . .	30
2.1.3.1	Distribuições de Frequências de Variáveis Qualitativas. Representação em Tabelas . . . . .	30
2.1.3.2	Representações Gráficas de Variáveis Qualitativas . . . . .	36
2.1.3.3	Exercícios Propostos . . . . .	39
2.1.3.4	Tabelas de Contingência . . . . .	42
2.1.3.5	Exercícios Propostos . . . . .	49
2.1.3.6	Coeficiente de Contingência de Pearson . . . . .	50
2.1.3.7	Exercícios Propostos . . . . .	55
2.1.3.8	Distribuições de Frequências de Variáveis Quantitativas: Diagrama de Pontos, Gráfico de Colunas e Histograma . . . . .	55
2.1.3.9	Exercícios Propostos . . . . .	67

2.1.3.10	Ramo-e-Folhas . . . . .	69
2.1.3.11	Exercícios Propostos . . . . .	73
2.2	Algumas Medidas e Modelo para Variáveis Quantitativas . . . . .	75
2.2.1	Medidas de Tendência Central . . . . .	75
2.2.1.1	Média, Mediana e Moda . . . . .	75
2.2.2	Medidas de Dispersão . . . . .	83
2.2.2.1	Variância, Desvio Padrão, Coeficiente de Variação e Desvio Interquartilico . . . . .	83
2.2.2.2	Exercícios Propostos . . . . .	91
2.2.3	O Uso da Mediana e dos Quartis na Interpretação de um Conjunto de Dados . . . . .	94
2.2.4	Desenho Esquemático (em inglês leia-se <i>Box Plot</i> ) . . . . .	97
2.2.4.1	Exercícios Propostos . . . . .	101
2.2.5	Coeficiente de Correlação . . . . .	102
2.2.5.1	Exercícios propostos . . . . .	109
2.2.6	Ajuste da Equação de uma Reta . . . . .	110
2.2.6.1	Exercícios propostos . . . . .	115
<b>3</b>	<b>Modelos de Probabilidades para Experimentos Simples</b>	<b>118</b>
3.1	Introdução . . . . .	118
3.2	Conceitos Fundamentais . . . . .	122
3.3	Operações com Eventos . . . . .	126
3.4	A Probabilidade de Um Evento . . . . .	129
3.5	Conceito de Probabilidade . . . . .	129
3.6	A Regra da Adição . . . . .	132
3.7	Probabilidade Condicional e Independência . . . . .	134
3.8	Teorema de Bayes . . . . .	139
3.9	Exercícios Propostos . . . . .	143
<b>4</b>	<b>Modelos de Probabilidades para Contagens</b>	<b>148</b>
4.1	Introdução . . . . .	148
4.2	O Conceito de Variável Aleatória Discreta e Função de Probabilidade . . . . .	149
4.3	A média e a Variância de Uma Variável Aleatória Discreta. Propriedades . . . . .	152
4.4	Alguns Modelos Probabilísticos para Contagens . . . . .	160
4.5	Distribuição de Bernoulli . . . . .	161
4.6	A Distribuição Binomial . . . . .	163

4.7	A Distribuição de Poisson . . . . .	172
4.8	Aproximação da Distribuição Binomial pela Distribuição de Poisson . . . . .	175
4.9	Exercícios Propostos . . . . .	182
<b>5</b>	<b>Modelos de Probabilidades para Medidas</b>	<b>191</b>
5.1	Introdução . . . . .	191
5.2	A Média de Uma Variável Aleatória Contínua . . . . .	196
5.3	A Distribuição Normal . . . . .	197
5.3.1	Introdução . . . . .	198
5.3.2	A Distribuição Normal Padrão . . . . .	202
5.3.3	O Uso da Tabela da Distribuição Normal Padrão . . . . .	204
5.3.4	A Distribuição Normal como Aproximação da Distribuição Binomial . . . . .	208
5.3.5	Exercícios propostos . . . . .	212
<b>6</b>	<b>Introdução à Inferência Estatística</b>	<b>219</b>
6.1	Introdução . . . . .	219
6.2	Como Selecionar Uma Amostra . . . . .	223
6.3	Amostragem Aleatória Simples . . . . .	225
6.3.1	Obtenção de Uma Amostra Aleatória Simples . . . . .	226
6.4	Distribuições Amostrais . . . . .	230
6.4.1	Distribuição Amostral da Proporção . . . . .	231
6.4.2	Distribuição Amostral da Média . . . . .	237
6.5	Exercícios Propostos . . . . .	244
<b>7</b>	<b>Estimação dos Parâmetros</b>	<b>247</b>
7.1	Introdução . . . . .	247
7.2	Propriedades dos Estimadores . . . . .	247
7.3	Métodos para Encontrar Estimadores . . . . .	250
7.3.1	Método da Máxima Verossimilhança . . . . .	251
7.4	Estimativas Pontuais e Intervalares . . . . .	252
7.4.1	Estimativas Pontuais . . . . .	252
7.4.2	Erro Amostral . . . . .	253
7.4.3	Estimativas por Intervalos de Confiança . . . . .	254
7.4.3.1	Intervalo de Confiança para a Média da População . . . . .	255
7.4.3.2	Intervalo de Confiança para Uma Proporção Populacional . . . . .	260
7.4.4	Determinação do Tamanho da Amostra . . . . .	262

7.4.4.1	Determinação do Tamanho da Amostra para Estimar uma Média Populacional . . . . .	262
7.4.4.2	Determinação do Tamanho da Amostra Para Estimar Uma Proporção Populacional . . . . .	264
7.4.4.3	Outro parâmetro: Determinação do Tamanho da Amostra para Estimar a Densidade de Primatas na Natureza . . . . .	267
7.5	Exercícios Propostos . . . . .	269
<b>8</b>	<b>Testes de Hipóteses Sobre os Parâmetros</b>	<b>274</b>
8.1	Introdução . . . . .	274
8.2	Construção de Testes de Hipóteses . . . . .	274
8.3	Teste Para Proporções Populacionais . . . . .	276
8.3.1	Teste Para Uma Proporção Populacional . . . . .	276
8.3.2	Poder do Teste e Tamanho da Amostra para Uma Proporção Populacional . . . . .	283
8.3.3	Teste de Duas ou Mais Proporções (Teste de Qui-Quadrado) . . . . .	288
8.3.4	Teste de Aderência . . . . .	293
8.3.5	Tamanho da Amostra para Duas Proporções Populacionais . . . . .	295
8.4	Teste para Médias Populacionais . . . . .	297
8.4.1	Teste para Uma Média Populacional . . . . .	298
8.4.1.1	Teste para Uma Média Populacional com Variância Conhecida	299
8.4.1.2	Teste para Uma Média Populacional com Variância Desconhecida . . . . .	300
8.4.1.3	Poder do Teste e Tamanho da Amostra para uma Média Populacional . . . . .	303
8.4.2	Teste de Duas Médias Populacionais com Variâncias Populacionais Desconhecidas . . . . .	306
8.4.2.1	Teste para Dados Pareados . . . . .	308
8.4.2.2	Teste para Comparação de Duas Médias, de Amostras Independentes e Variâncias Desconhecidas e Iguais . . . . .	312
8.4.2.3	Teste para Comparação de Duas Médias, de Amostras Independentes e Variâncias Desconhecidas e Desiguais . . . . .	316
8.4.2.4	Teste $F$ para Comparação de Variâncias Populacionais . . .	318
8.5	Exercícios Propostos . . . . .	320
	<b>Bibliografia</b>	<b>330</b>

Apêndice 1: Tabela da distribuição binomial	332
Apêndice 2: Tabela da distribuição de Poisson	342
Apêndice 3: Tabela da distribuição normal padrão	344
Apêndice 4: Tabela da distribuição $t$ de Student	346
Apêndice 5: Tabela da distribuição Qui-Quadrado	348
Apêndice 6: Tabela da distribuição $F$	350
Apêndice 7: Tabela de números aleatórios	353
Respostas da Maioria dos Exercícios	355

# 1 Introdução Geral

## 1.1 A Estatística na Pesquisa Agrária e Biológica

A estatística é uma ciência que se preocupa com o planejamento de uma pesquisa, envolvendo desde a forma de coleta das observações, obtidas em *experimentos* ou *levantamentos*, até a maneira como é feita a organização, a descrição e o resumo dos dados, assim como a avaliação e afirmação sobre características de interesse do pesquisador.

As análises estatísticas dependem da forma de como os dados são coletados, e o planejamento estatístico da pesquisa indica o esquema sob o qual os dados serão obtidos. Portanto, o planejamento da pesquisa e a análise estatística dos dados estão intimamente ligados.

Planejamento da pesquisa  $\Rightarrow$  Análise estatística

Dessa forma, o pesquisador deve possuir um razoável conhecimento de estatística para desenvolver suas pesquisas ou, então, consultar um estatístico para auxiliá-lo. É importante frisar que esta consulta deve ser feita antes do início da pesquisa, ainda durante a fase de elaboração do projeto.

### 1.1.1 Forma de coleta dos dados

Em alguns casos, como por exemplo, na descrição de novas espécies florestais, os dados são obtidos através da simples observação de como o fenômeno acontece na natureza. Neste caso, não são feitas alterações no fenômeno em estudo. Outro exemplo é o estudo sobre os efeitos da poluição ambiental sobre animais roedores de determinada espécie. Neste caso, um certo número de animais (amostra), capturados na região em estudo, será analisado, contando o número de micronúcleos existentes a cada cinco mil células (característica em estudo).

Levantamento: observa-se o fenômeno na natureza

Por outro lado, em muitas outras situações, as observações precisam ser geradas e são feitas comumente sob condições controladas pelo pesquisador, e os fatos, eventos ou fenômenos são forçados a sofrer variações sistemáticas, mediante a aplicação de *tratamentos*.

Tratamentos: as variações de um ou mais fatores de interesse em avaliar

Por exemplo, num estudo da produção de milho mediante a aplicação de diferentes doses de nitrogênio, as diferentes doses de nitrogênio constituiriam os tratamentos. Os outros fatores, como por exemplo, diferenças na fertilidade, umidade do solo e existência de pragas e ervas daninhas, que poderão influir nos resultados (produção final de milho), são minimizados, tanto quanto possível, do ponto de vista prático. Neste caso temos um *experimento*. A função do experimento é determinar as relações de causa e efeito, como por exemplo, verificar como as doses de nitrogênio (causa) influenciam na produção de milho (efeito).

Experimento: causa  $\Rightarrow$  efeito

### 1.1.2 Erro Experimental

Quando instalamos um experimento desejamos verificar o efeito de diferentes tratamentos. Os demais efeitos, que não os de tratamentos, devem ser controlados ao máximo do ponto de vista prático. Portanto, as pulverizações com produtos químicos, as capinas, as mensurações etc. devem ser feitas de modo o mais homogêneo possível em todo o experimento. Da mesma forma, a disponibilidade de água, a temperatura, a umidade e as sementes utilizadas devem ser as mais similares possíveis. Acontece que, na prática, por maiores que sejam os esforços dos pesquisadores para homogeneizar todos esses efeitos, isto normalmente não é possível, pois existem as variações casuais ou aleatórias. Isto pode ser verificado quando, ao repetir o experimento, sob condições similares, obtemos resultados diferentes. Essas variações são chamadas de *erro experimental* ou simplesmente *erro*. É bom chamar a atenção que apesar do termo *erro*, isto não significa que o experimento foi mal feito. Vale a pena ressaltar que essas variações também ocorrem para dados obtidos através de levantamentos.

Finney (1952) diz que o propósito da ciência estatística é fornecer uma base objetiva para a análise de problemas nos quais os dados estão sujeitos a variação do acaso. Por maiores que sejam os conhecimentos de um pesquisador sobre, por exemplo, nutrição e fisiologia animal, ele jamais será capaz de prever com exatidão qual vai ser o peso de um suíno criado sob determinadas condições. Existe um grande número de causas que fazem este resultado variar, como, por exemplo, variações genéticas, de temperatura ambiental, umidade, doenças



etc. Assim, quando o elemento *acaso* está presente em um problema, dificuldades reais são introduzidas. São os chamados fenômenos *aleatórios*.

É importante ressaltar que quase tudo que fazemos no nosso cotidiano são fenômenos aleatórios e, portanto, apresentam uma chance de ocorrência devido ao acaso. Assim sendo, é desejável determinar qual é a probabilidade de ocorrência dos eventos de interesse. Para tal finalidade, precisamos estabelecer o modelo probabilístico adequado.

Fenômeno aleatório  $\Rightarrow$  Modelo probabilístico

A seguir, vamos apresentar duas partes importantes da estatística, quais sejam: *análise exploratória de dados* e *inferência estatística*.

### 1.1.3 Análise Exploratória de Dados

Todo o trabalho de análise estatística inicia com a análise exploratória de dados, com o objetivo de obter conhecimento dos dados. Em muitos casos, com uma boa análise exploratória de dados, atende-se aos objetivos da pesquisa (principalmente em estudos realizados através de levantamentos), enquanto em outros casos, ela serve como uma primeira aproximação da análise final.

O conceito de estatística é bastante amplo e engloba a noção usual que as pessoas têm do que seja estatística. Esse conceito usual logo relaciona a estatística com tabelas, gráficos, taxas, índices etc. Assim, ouvimos falar da produtividade média do milho no Estado de Santa Catarina, índice pluviométrico mensal ou anual de uma determinada região, índice de inflação, taxa de desemprego, estatísticas de saúde pública, estatísticas da loteria, estatísticas do aumento da produção de maçã em Santa Catarina etc. Esse foco da estatística é apresentado nos livros como análise exploratória de dados. Ele se preocupa com a organização, apresentação, simplificação e descrição dos dados. Este enfoque de análise exploratória de dados foi introduzido por Tukey (1971), no seu livro: *Exploratory Data Analysis*. A análise exploratória de dados utiliza-se muito de técnicas visuais e procura vislumbrar alguma regularidade quase sempre presente num conjunto de dados, podendo sugerir modelos que possam ser utilizados na inferência estatística, discutida na seção 1.1.4.

**Exemplo 1.1** Foi feito um experimento para estudar a duração, em dias, do quinto estágio ninfal de *Triatoma klugi* com alimentação em galo. Os resultados são apresentados na Tabela 1.1. Esses resultados estão representados num gráfico de colunas dado na Figura 1.1. Duas

Tabela 1.1: Valores da duração em dias do quinto estágio ninfal de *Triatoma klugi* com alimentação em galo, Departamento de Microbiologia, UFSC, Florianópolis, SC, 2001.

26	37	39	40	41	42	42
43	44	45	45	45	45	45
47	48	48	48	48	48	49
49	49	51	51	51	52	53
53	53	56	57	60	62	62

interpretações podem ser levantadas desse gráfico: 1) um valor típico para representar os dados seria 48 dias; 2) a distribuição é levemente assimétrica à esquerda, ou seja, os valores acima do valor típico estão mais concentrados do que os valores abaixo do mesmo. A análise exploratória de dados será estudada no Capítulo 2.

#### 1.1.4 Inferência Estatística

Há, ainda, um campo bastante amplo da ciência estatística que se refere à análise e à interpretação do conjunto total de dados, considerando a observação de somente uma parte deles. Essa parte é conhecida como Estatística Indutiva ou Inferencial, e normalmente escapa a noção corrente da grande maioria das pessoas.

Para deixar mais clara a finalidade da inferência estatística é necessário fazer a apresentação de dois conceitos fundamentais, que são: *População* e *Amostra*.

Uma *população* consiste de todos os valores possíveis de uma característica desejada observados em unidades amostrais ou experimentais. Os valores que compõem uma população geralmente são diferentes entre si, e a mesma pode apresentar um tamanho finito, nem sempre conhecido, ou infinito. Na experimentação a definição de população é conceitual, isto é, não tem existência real. São exemplos de populações: todos os valores possíveis da produção de milho, em kg/ha, de uma cultivar; todos os pesos ao nascer de coelhos da raça gigante; todos os valores de diâmetro altura do peito (DAP) de uma espécie do manguezal do Itacorubi; todos os valores de micronúcleos por cinco mil células examinadas de roedores de uma determinada região. Nestes exemplos as unidades experimentais são, respectivamente: um canteiro de 5 m × 2 m, láparo (filhote de coelho) da raça gigante, uma árvore da espécie, cinco mil células do roedor.

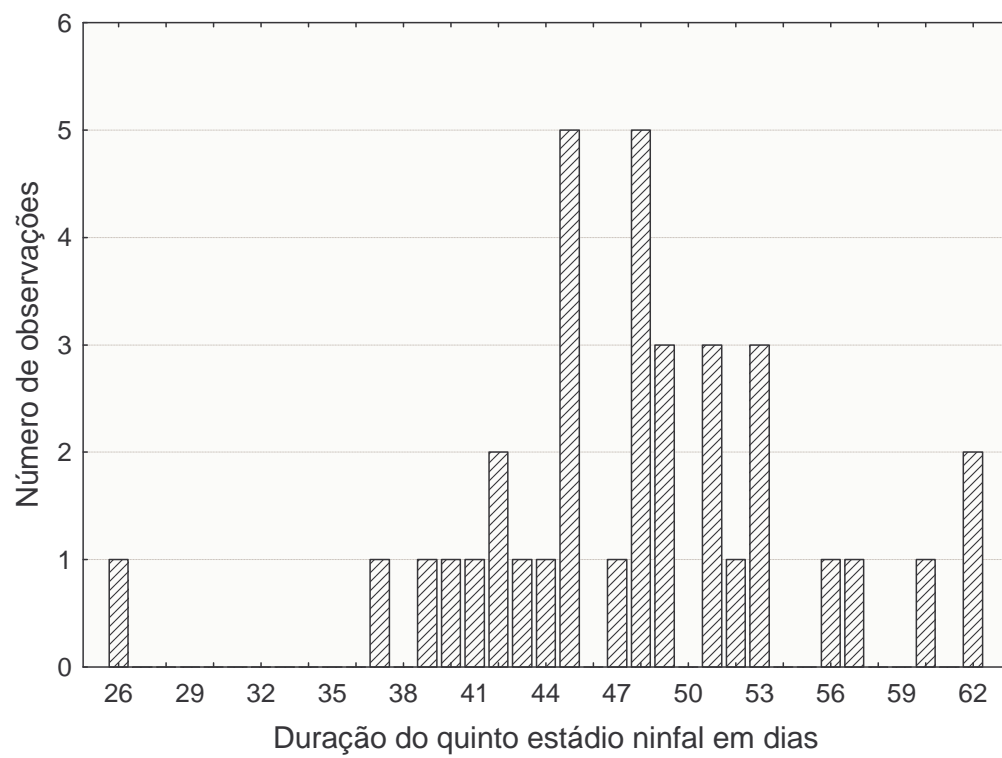


Figura 1.1: Gráfico de colunas da duração, em dias, do quinto estágio ninfal de *Triatoma klugi*

População: todos os valores possíveis de uma característica em estudo observados em unidades experimentais.

*Amostra* é uma parte (subconjunto) da população. Exemplos: os pesos ao nascer de coelhos de uma ninhada podem ser uma amostra de uma população; uma amostra formada por 100 pesos de pacotes de café selecionados; 50 valores de micronúcleos por cinco mil células examinadas de peixes do gênero bagre selecionados; cinco valores de produção de milho cultivado com a dose de 50 kg/ha de nitrogênio, cada um tomado num canteiro de 5 m  $\times$  2 m.

Amostra: um subconjunto da população

Portanto, com a inferência estatística, examinando apenas uma amostra tiramos conclusões sobre a população. É intuitivo que, quanto maior a amostra, mais precisas e confiáveis deverão ser as inferências realizadas sobre a população. Levando esse raciocínio ao extremo, concluímos que os resultados mais exatos seriam obtidos pelo exame completo de toda a população, procedimento denominado de *censo* ou *recenseamento*. Porém, a utilização de amostras pode ser feita de tal maneira que se obtenham resultados confiáveis, em termos práticos, de forma equivalente ou até mesmo superiores aos que seriam conseguidos através do censo. Na experimentação biológica e agrária, geralmente não temos acesso a toda a população, portanto, somos obrigados a trabalhar com amostras. Por exemplo, não podemos conhecer todos os valores possíveis de produção de milho em kg/ha de uma cultivar; não podemos determinar todos os diâmetros de caramujos de uma área. O fato é que não é necessário examinar toda a população para se chegar às conclusões desejadas.

Como exemplo, podemos citar o experimento que tem como objetivo verificar o comportamento da produção de milho sob o efeito de diferentes doses de nitrogênio: 0, 25, 50, 75 e 100 kg de nitrogênio por hectare. Essas doses foram definidas pelo pesquisador e correspondem aos cinco tratamentos envolvidos no experimento. A cultivar de milho escolhida para o experimento foi a cultivar mais plantada na região. Para cada dose de nitrogênio foram plantados cinco canteiros de terra de 10 m<sup>2</sup> cada, portanto, o tamanho da amostra para cada dose de nitrogênio é cinco (cinco repetições dos tratamentos). Veja os resultados dessas amostras na Tabela 1.2. Para cada amostra existe uma correspondente população e cada população é formada por todos os valores possíveis de produção, em kg/ha, para a dose

correspondente. Após realizada a análise estatística e a interpretação dos resultados, poderá ser indicada qual a melhor dose de nitrogênio para a cultivar de milho utilizada.

Tabela 1.2: Produção de milho, em kg/ha, submetidos à diferentes doses de nitrogênio (tratamentos), em kg/ha, no Oeste Catarinense, 1993.

Tratamentos	Repetições					Total	Média
	I	II	III	IV	V		
0	2850	1780	2100	2900	2010	11640	2328
25	3200	1980	2220	2850	2100	12350	2470
50	4150	2330	3700	4050	2500	16730	3346
75	4380	2830	3420	3900	3080	17610	3522
100	4000	2630	3150	3780	2670	16230	3246

A inferência estatística, dada a sua grande importância, apresenta um grande número de métodos de análise. Seriam necessários vários cursos de estatística para estudá-los. É bom deixar claro que para se realizar uma análise estatística, por mais sofisticado que seja o método utilizado, deve-se, primeiramente, proceder à análise exploratória dos dados.

Como já foi comentado anteriormente, uma característica dos experimentos biológicos e agrônômicos é que os seus resultados tendem a variar toda vez que se repetir o experimento, sob condições bastante semelhantes. Esta variabilidade nos resultados do experimento deixa o pesquisador com dúvidas quanto ao(s) melhor(es) tratamento(s). Neste caso, os métodos estatísticos irão auxiliá-lo.

Para ilustrar esta variabilidade vamos considerar os resultados da Tabela 1.2, relativos às produções de milho, em kg/ha. Note que, em todos os tratamentos, as produções de milho variaram. Por exemplo, tomando-se o tratamento 75, obtivemos cinco produções diferentes, variando de 2830 a 4380 kg/ha. Esta variação é decorrente de inúmeros fatores que interferem no resultado final, como já discutido na Seção 1.1.2. Por causa desta variabilidade, seria pouco confiável realizar apenas uma repetição. Intuitivamente, percebemos que as análises ficam mais consistentes se fizermos várias repetições e usarmos a média dos resultados.

No caso do experimento citado, desejamos comparar cinco tratamentos, incluindo o controle (dose 0), com relação à produção. Mais especificamente, podemos estabelecer dois objetivos para o experimento, quais sejam: 1<sup>o</sup>) testar a hipótese de que não existem diferenças entre os tratamentos; 2<sup>o</sup>) estimar a diferença de produção entre dois tratamentos.

Com respeito ao primeiro objetivo, podemos, por exemplo, comparar a média do tratamento 75 com o controle. Observamos uma diferença de 1194 (= 3522 - 2328) kg/ha entre

as duas médias obtidas com as amostras. Esta é uma diferença bem considerável, porém uma nova repetição do experimento poderia fornecer um outro resultado, para mais ou para menos, devido ao fato de estarmos trabalhando com fenômenos aleatórios. Em outras palavras, mesmo se o tratamento 75 não diferir do controle, em termos de produção média de milho, as produções médias obtidas a partir de amostras, provavelmente serão diferentes. Assim, ficamos na dúvida se os resultados obtidos decorreram de uma diferença real entre os tratamentos ou se foram originados pelo acaso. Este tipo de problema é solucionado através da aplicação dos *testes de hipóteses*, que serão estudados no capítulo 8. Basicamente, num teste de hipóteses, o pesquisador vai decidir se um resultado obtido em um experimento ocorreu devido ao acaso ou devido a ação do tratamento. Caso o pesquisador decida pela existência da ação do tratamento, as produções médias de milho do tratamento 75 e do controle são diferentes, seu interesse seria avaliar a magnitude desta diferença. Esse é o segundo objetivo referido acima. Descritivamente podemos dizer que para este experimento esta diferença foi de 1194 kg/ha em favor do tratamento 75. Mas esta é uma medida que por si só tem pouca importância, devido ao fato de ser o resultado de um experimento com cinco repetições. Métodos de estimação de uma média ou da diferença entre duas médias serão estudados nos capítulos 7 e 8.

### 1.1.5 Teoria da Probabilidade e seus Modelos

Vimos que o objetivo da inferência estatística é tirar conclusões sobre populações com base nos resultados de amostras extraídas dessas populações. Como vamos trabalhar com amostras, o processo não pode ser exato. Ao se fazer inferências sobre uma população, portanto, estamos sempre sujeitos a cometer erros, isto é, o pesquisador não pode fazer afirmativas com 100% de certeza. Isto, porém, não deve desesperançar o pesquisador, pois a inferência estatística permitirá dizer até que ponto se pode estar errando, em termos probabilísticos. Por exemplo, poderemos afirmar com 95% de confiança que a real diferença entre as médias de dois tratamentos é um valor no intervalo de 694 a 1589 kg/ha. Então é possível determinar limites dentro dos quais a verdadeira diferença deve encontrar-se, com um certo grau de confiança definido pelo pesquisador. Esses intervalos são conhecidos como intervalos de confiança, que serão estudados nos capítulos 7 e 8.

A espinha dorsal da inferência estatística é a teoria da probabilidade, com seus modelos probabilísticos.

O que é um modelo? Modelo é uma versão simplificada de algum evento, fenômeno ou acontecimento da vida real. Por exemplo, um globo terrestre é uma versão simplificada do

planeta terra. Uma maquete de um prédio e um layout (distribuição interna) também são exemplos de modelos.

Um dos propósitos da ciência é descrever e fazer previsões de eventos do mundo real, do mundo no qual nós vivemos. Uma maneira pela qual isto é feito, é construindo-se modelos matemáticos (que são expressões matemáticas) que possam descrever adequadamente os fenômenos do mundo real. Por exemplo, suponhamos que 10 vacas de mesma idade e raça são tratadas com uma determinada ração para aumentar a produção de leite (total da lactação). Admitimos que a probabilidade de aumento de produção na lactação de cada animal é de 0,65. Então podemos estar interessados em saber qual é a probabilidade de exatamente 8 vacas aumentarem a produção na lactação. O modelo que possibilita o cálculo desta probabilidade é denominado de *modelo binomial*. Esse e outros modelos serão estudados nos capítulos 4 e 5.

Um modelo deve simplificar as coisas, e certos pormenores devem ser desprezados. É claro que estes pormenores não devem ter importância para o entendimento do fenômeno em estudo. A resolução do problema matemático pode estar correta e, mesmo assim, estar em grande discordância com os dados observados, simplesmente porque as hipóteses básicas feitas não são confirmadas. Por isso, é muito importante deduzir certas consequências do modelo e, a seguir, comparar esses resultados previstos pelo modelo com dados reais (observados). É a validação do modelo.

## 1.2 A Estatística e o Método Científico

Nesta seção pretendemos mostrar como a estatística pode ajudar o pesquisador a resolver problemas.

Quando se faz uma pesquisa científica o procedimento geral é formular hipóteses e testá-las. Inicialmente essas hipóteses são formuladas em termos científicos, dentro da área de estudo (hipóteses científicas)<sup>1</sup>. Em seguida, as hipóteses devem ser expressas em termos estatísticos (hipóteses estatísticas). É claro que deve haver uma correspondência perfeita entre as hipóteses científica e estatística. Por exemplo, no caso do experimento de aplicação de diferentes doses de nitrogênio, cujos resultados estão apresentados na Tabela 1.2, podemos formular a seguinte hipótese científica: é possível aumentar a produção de milho (kg/ha)

---

<sup>1</sup>Hipótese científica: Dado um problema bem definido, identificado, vamos imaginar uma explicação para algum aspecto do problema que nos tenha despertado interesse. Essa é a hipótese, e deve ser coerente com as observações importantes já feitas, aliada aos conhecimentos teóricos que o pesquisador possui sobre o assunto.

através da aplicação de nitrogênio. E a seguinte hipótese estatística: existe diferenças entre as médias verdadeiras de produção de uma cultivar de milho quando a mesma é submetida a diferentes doses de nitrogênio. Para testar uma hipótese estatística, é preciso um conjunto de observações, isto é, é preciso coletar dados (valores a respeito da característica ou características que estamos estudando). Por exemplo, nós precisamos de dados sobre a produção de milho para as diferentes doses de nitrogênio, para que possamos testar a hipótese acima formulada. Como estamos tratando de experimentação, vamos obter nossas observações, nossos dados, através de um experimento, ou seja, as observações serão feitas sob condições controladas, os fatos ou fenômenos a serem estudados serão planejados a sofrer variações sistemáticas, mediante a aplicação de tratamentos. Por exemplo, a produção de milho sofre variações causadas pela aplicação de diferentes doses de nitrogênio. Os efeitos dos outros fatores que não serão analisados, serão minimizados tanto quanto possível (fertilidade do solo, ataque de pragas e doenças, plantas invasoras, sombreamento etc.). A hipótese acima formulada vai ser testada por meio de uma análise estatística. Esta, por sua vez, depende de como foi instalado o experimento, ou seja, de como as observações foram obtidas. Com isso podemos verificar a grande importância de um bom planejamento inicial do experimento, fase esta chamada de *planejamento estatístico do experimento*. Planejamento de experimento e análise estatística são feitos em seqüência e estão intimamente ligados. No planejamento do experimento, é importante a participação de especialistas de diversas áreas, pois quanto maior o conhecimento adquirido, melhor o entendimento sobre o fenômeno em estudo e, isto, facilitará a solução final do problema. "Por delineamento estatístico de experimento, entendemos o processo de planejamento do experimento, de tal forma que os dados obtidos possam ser analisados através de métodos estatísticos, resultando em conclusões válidas e objetivas" (Montgomery, 2005). Podemos resumir isto que foi dito, por meio da representação gráfica da circularidade do método científico (Peres e Saldiva, 1982), apresentada na Figura 1.2.

Para fixarmos melhor a idéia de pesquisa científica estatisticamente planejada, vamos considerar dois projetos de pesquisa.

### **Projeto 1: Recuperação de Ervais Nativos Através da Decepa<sup>2</sup>**

1) *Identificação do problema*. Inicia-se uma pesquisa científica com a definição do problema, juntamente com a formulação dos objetivos e hipóteses.

Os ervais explorados para a produção de erva-mate são na grande maioria plantas nativas. Dada a grande demanda por erva-mate, as ervaíras foram exploradas, isto é, extraída sua massa foliar, através de métodos inadequados; esta falta de manejo implicou na extinção de

---

<sup>2</sup>Projeto desenvolvido no Centro de Pesquisa para a Agricultura Familiar, EPAGRI S.A.



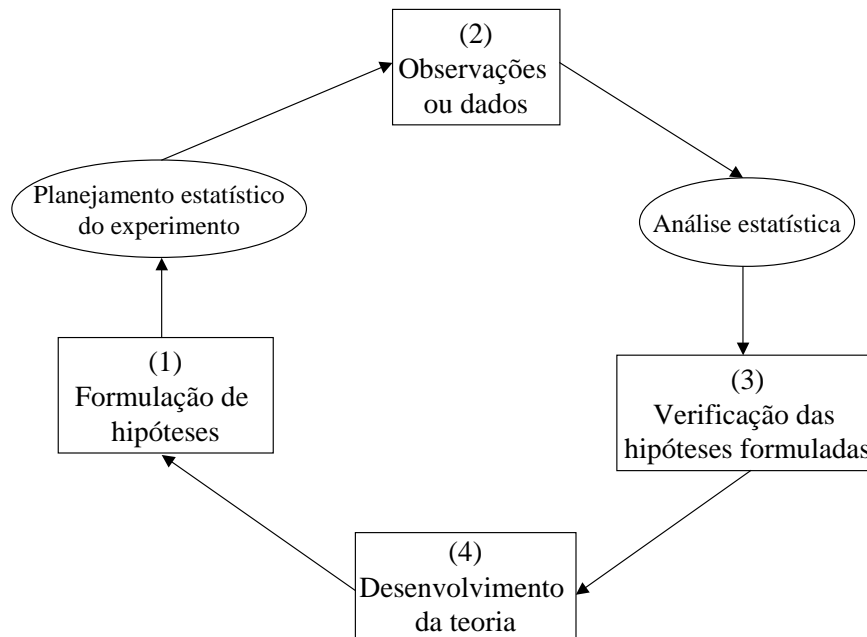


Figura 1.2: Circularidade do método científico

Fonte: modificado de Peres e Saldiva, 1982

espécies e outras estão em vias de extinção, então, os ervais nativos ficaram em sua quase totalidade comprometidos. Buscando recuperar os ervais nativos, irá aplicar-se a técnica da decepta total das erveiras em diferentes níveis de altura.

2) *Objetivo geral.* Determinar o efeito da decepta em plantas adultas e danificadas de erveiras.

3) *Objetivos específicos:*

- Verificar a capacidade e comprimento de brotação;
- estudar a produção de massa verde;
- verificar a sobrevivência das erveiras decepadas.

4) *Formulação da hipótese científica.* É viável a recuperação de ervais nativos e improdutivos através da prática da decepta.

5) *Escolha dos fatores que devem ser incluídos no estudo e seus correspondentes níveis (tratamentos).* O pesquisador deve escolher os fatores a serem estudados, a faixa na qual esses fatores serão variados, e os níveis específicos utilizados no experimento. Neste projeto o fator, também conhecido como variável independente, é a decepta, e as alturas em que será realizada a mesma são os tratamentos (níveis do fator). Foram utilizadas quatro alturas

de decepta (4 tratamentos), quais sejam: tratamento 1 - altura de decepta a 0,00 m do solo; tratamento 2 - altura de decepta a 0,30 m do solo; tratamento 3 - altura de decepta a 0,60 m do solo; tratamento 4 - altura de decepta a 0,90 m do solo. Este é um *experimento com um fator*.

6) *Escolha da unidade experimental*. As unidades experimentais são as que recebem os tratamentos e devem ser as mais homogêneas possíveis, para que quando submetidas a tratamentos diferentes, seus efeitos sejam facilmente detectados. Portanto, elas devem ser orientadas no sentido de minimizar o erro experimental. As unidades experimentais pode ser um animal, um conjunto de animais, uma pessoa, cinco mil células, uma planta, um conjunto de plantas, um vaso, um frango, cinco áreas de 5  $\text{cm}^2$  em um frango, um tubo de ensaio, etc. Nesse experimento, a unidade experimental é formada por 8 plantas de erva-mate de diâmetros bem próximos. Então, o tratamento 1 será aplicado a 8 plantas de erva-mate, o tratamento 2 a outras 8 plantas, e assim por diante, até o quarto tratamento. Embora tenhamos 8 medidas da variável resposta em cada unidade experimental, esses resultados não são repetições *independentes*. Na realidade, só temos um resultado *independente* para cada tratamento, a média das oito árvores para cada variável resposta em estudo.

A heterogeneidade das unidades experimentais é que determina os diferentes planos experimentais.

7) *Escolha das variáveis que serão medidas nas unidades experimentais*. Denominam-se de variáveis as características que serão mensuradas, avaliadas pelos pesquisadores nas unidades experimentais. As variáveis são pré-estabelecidas pelo pesquisador e devem medir diretamente os tratamentos de acordo com os objetivos do trabalho. Algumas variáveis medidas nesse experimento foram: percentagem de brotação, número de brotos, comprimento dos brotos, produção de massa verde para a indústria e sobrevivência das erva-mates deceptadas. O importante é que os dados devem ser objetivos, precisos e verdadeiros, isto é o mínimo que se espera de alguém que irá publicar um trabalho de pesquisa.

8) *Determinação das regras e procedimentos pelos quais os tratamentos são atribuídos às unidades experimentais: delineamentos experimentais*. Trata-se de normas de designar os tratamentos às unidades experimentais e que definem os delineamentos experimentais. A estatística é bastante rica em planos (delineamentos) experimentais. É nesse item que a estatística participa fortemente do planejamento da pesquisa, ou seja, a sua contribuição é bastante grande, de tal forma que pode-se chamar esta etapa de planejamento estatístico do experimento. Veja Figura 1.2. É importante na fase de planejamento da pesquisa, escolher adequadamente o delineamento, pois, caso contrário, pode-se ter muita dificuldade na análise estatística e, até mesmo, invalidar os resultados do experimento.

Neste experimento, os tratamentos foram atribuídos às unidades experimentais da seguinte forma. Primeiramente foram formados cinco blocos, onde cada bloco é constituído de 4 unidades experimentais (pois temos 4 tratamentos); como cada unidade experimental tem 8 plantas, então, um bloco tem 32 plantas. Os blocos estão controlando as diferenças de diâmetros entre as plantas, assim, o bloco I é formado por 32 plantas com diâmetros entre 10 e 13 cm (exclusive); o bloco II é formado por 32 plantas com diâmetros entre 13 e 16 cm (exclusive); o bloco III é formado por 32 plantas com diâmetros entre 16 e 19 cm (exclusive); o bloco IV é formado por 32 plantas com diâmetros entre 19 e 22 cm (exclusive) e o bloco V é formado por 32 plantas com diâmetros superiores a 22 cm. Portanto, cada tratamento será repetido 5 vezes, uma em cada bloco. Dentro de cada bloco houve o sorteio (aleatorização) de qual unidade receberá o tratamento 1, qual receberá o tratamento 2, e assim por diante. Este delineamento experimental é denominado de blocos completos ao acaso.

9) *Coleta dos dados.* Aqui são feitas as medidas das variáveis estabelecidas pelo pesquisador. As variáveis, logicamente, devem avaliar diretamente os efeitos dos tratamentos de acordo com objetivos do experimento. Também podem ser coletadas variáveis complementares, que serão úteis para explicar o comportamento dos tratamentos.

10) *Análise estatística dos resultados.* O objetivo da análise estatística é verificar as hipóteses formuladas no início da pesquisa científica. Por exemplo, comparar as médias de produção de massa foliar obtidas com cada uma das alturas de decepta, ou, estabelecer uma relação funcional entre a produção de massa foliar e a altura de decepta das plantas.

Existem excelentes softwares estatísticos para realizar as análises. A análise de resíduos é uma importante técnica para verificar, por exemplo, se o modelo é adequado.

11) *Relatório final (publicação).* Apresentar tabelas e gráficos de forma a mostrar os efeitos esperados, comparar os resultados obtidos com os objetivos do experimento para verificar se as questões propostas foram respondidas. Apresentar medidas de precisão das estimativas. Se possível, fazer referências a outras pesquisas similares e uma avaliação de todas as etapas com sugestões para possíveis alterações em pesquisas futuras.

A experimentação é uma importante fase do processo de aprendizagem, onde nós formulamos hipóteses, realizamos o experimento para pesquisar sobre essas hipóteses e, de acordo com os resultados, formulamos novas hipóteses, e assim sucessivamente. Isto sugere que a experimentação é iterativa.

## **Projeto 2 - Análise de Alguns Aspectos da Dinâmica de Populações de Duas Amostras de *Biomphalaria tenagophila* Submetidas a Diferentes Concentrações de Matéria Orgânica no Meio.**

1) *Identificação do problema.* A importância do gênero *Biomphalaria* no contexto da saúde pública brasileira deve-se ao fato de que dentre as 19 espécies que constituem este gênero, dez delas são encontradas no Brasil, sendo que três são hospedeiras intermediárias do *Schistosoma mansoni*, causador da esquistossomose mansoni, um dos mais importantes problemas de saúde pública em muitos países tropicais e subtropicais. Sua ocorrência é acentuada entre populações carentes de alguns países subdesenvolvidos. Segundo a Organização Mundial da Saúde, são estimados 200 milhões de pessoas como tendo sido contaminadas pelo *S. mansoni*, enquanto, outros 500 a 600 milhões correm o risco de contraí-la. As três espécies hospedeiras são: *B. straminea*; *B. glabrata* e *B. tenagophila*.

No Brasil, estima-se em 5,5 milhões de pessoas infectadas, isto parece ser uma subestimativa, pois o Instituto de Medicina Tropical (IMT), da Faculdade de Medicina da USP, estimou em 10 milhões o número de pessoas infectadas.

Em Santa Catarina, o primeiro foco de transmissão ocorreu em São Francisco do Sul, através da *B. tenagophila*. Hoje, tem-se registro de *B. tenagophila* em 26 municípios do nordeste do Estado.

Na Ilha de Santa Catarina, pesquisadores obtiveram registros de *B. tenagophila*, *B. oligoza* e *Drepanotrema sp.*, em 8 pontos estratégicos.

Em consequência da alta endemicidade da esquistossomose no país, a distribuição dos planorbídeos vem sofrendo constantes investigações, sendo que especial atenção tem sido dada, ainda, ao controle da expansão das espécies vetoras, apesar que a área ocupada por cada uma das espécies vetoras do *S. mansoni*, está aumentando.

O gênero *Biomphalaria* apresenta uma grande tolerância a diferentes condições ecológicas, isto permitiu sua ampla distribuição geográfica. Muitos autores analisaram o comportamento reprodutivo e crescimento em função de uma série de variáveis, sendo que os efeitos de temperatura, tipo ou ausência de alimentação, influência do fotoperíodismo e densidade populacional são alguns que receberam maior atenção.

As condições do meio exigidas pelas biomfalárias para colonizar um ambiente, são: riquezas de microflora e matéria orgânica, pouca turbidez, boa insolação, pH em torno de 6 a 8, teor de NaCl abaixo de 3 por 1000 e temperatura média entre 20°C e 25°C. Cabe aqui ressaltar, todavia, que as biomfalárias suportam modificações consideráveis nas características físicas, químicas e biológicas de seus ambientes, podendo, inclusive, utilizar-se da estivação como uma estratégia para suportar a adversidade do meio.

Sabendo-se que os representantes do gênero *Biomphalaria* são constantemente encontrados em grande variedades de coleções de água doce, paradas ou pouco correntes, natural ou artificialmente alagadas, que um único espécime é capaz de produzir, por autofecundação,

uma população de numerosos indivíduos e que *B. tenagophila* está em provável extensão em Santa Catarina, torna-se importante estudar os fatores bióticos e abióticos que interferem na biologia, distribuição e adaptação aos ambientes por eles explorados. Assim sendo, o objetivo do presente trabalho é avaliar crescimento, desempenho reprodutivo, sobrevivência e fecundidade de duas amostras de *B. tenagophila*, considerando-se diferentes condições do meio.

2) *Objetivo geral.* Determinar o efeito de diferentes condições do meio (poluição) sobre a biologia de *B. tenagophila*.

3) *Objetivos específicos:*

- Verificar (estudar) o crescimento de *B. tenagophila*.
- Estudar o desempenho reprodutivo da espécie.

4) *Hipótese científica.* Num meio com poluição, o desenvolvimento biológico da espécie é prejudicado.

5) *Escolha do fator que deve ser incluído no estudo e seus correspondentes níveis.* Nesse projeto, o fator em estudo é *poluição*, também chamado de *variável independente*, os diferentes níveis de poluição são os *tratamentos*. Nesse trabalho foram utilizados dois tratamentos, quais sejam: 1) Sem poluição (com troca de água) e 2) com poluição (completa a água).

6) *Escolha da unidade experimental.* Nesse experimento a unidade experimental é um caramujo. Cada unidade experimental é formada por um copo de vidro (mini-aquário), com 60 ml de água deionizada, contendo um caramujo.

7) *Escolha das variáveis que serão medidas nas unidades experimentais.* Algumas variáveis avaliadas neste experimento foram: diâmetro em três diferentes tempos (nascimento, na 1ª desova e no final do experimento); idade na 1ª desova; número de ovos desenvolvidos; número de ovos viáveis; número de ovos inviáveis; número total de ovos.

8) *Determinação das regras e procedimentos pelos quais os tratamentos são atribuídos às unidades experimentais.* Nesse experimento a amostra total foi composta por 51 caramujos (i.é. 51 unidades experimentais). Através de um processo aleatório foram escolhidos 23 caramujos para receberem o tratamento T e 28 caramujos para receberem o tratamento C. A regra de estabelecer os tratamentos às unidades experimentais foi completamente ao acaso.

9) *Coleta de dados; Análise estatística dos dados e Relatório final.* Idem ao que foi comentado para o projeto 1.

### 1.3 Aspectos do Planejamento Estatístico de um Experimento

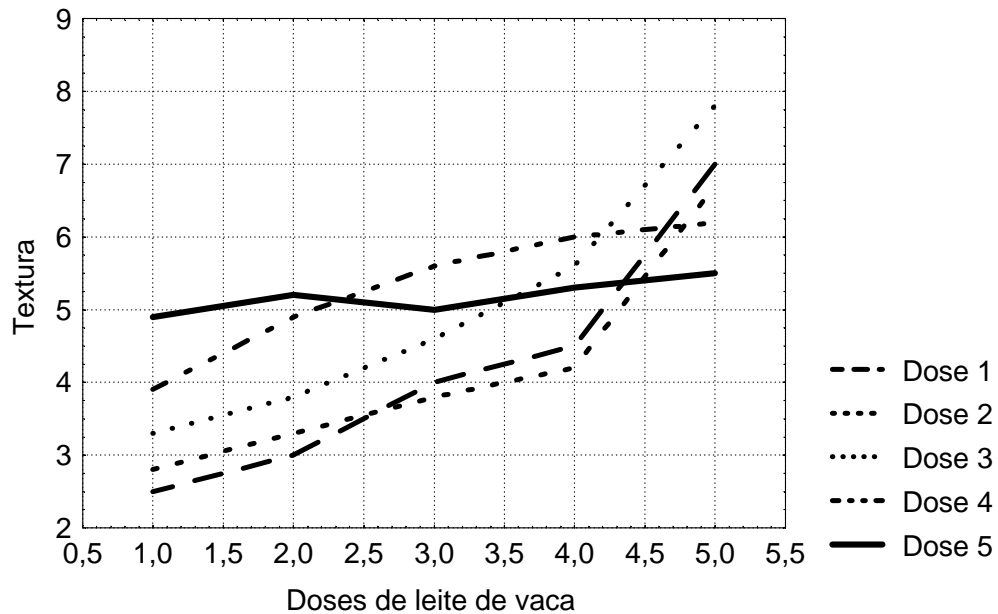


Figura 1.3: Efeito da interação entre doses de leite de vaca e doses de coagulante

### 1.3.1 Experimentos com um fator e mais de um fator

Nos projetos 1 e 2, vimos dois experimentos com apenas 1 fator. No projeto 1 o fator é a decape e, no projeto 2 é a poluição. Em muitos casos, temos experimentos com mais de um fator. Considere um experimento com cinco (5) doses de leite de vaca e cinco (5) doses de coagulante. Aqui temos 2 fatores em estudo, doses de leite e coagulante, com 5 níveis cada um, caracterizando um *experimento fatorial*. É interessante salientar que a escolha dos fatores e seus níveis é incumbência do pesquisador. Nos experimentos fatoriais é possível estudar-se a *interação* que existe entre os fatores, isto é, como é o comportamento dos níveis de um fator dentro dos níveis do outro fator. A Figura 1.3 ilustra o efeito da interação entre os fatores leite de vaca e coagulante.

Os fatores podem ser *quantitativos ou qualitativos*. A temperatura de um forno, os níveis de nitrogênio e de fósforo são exemplos de fatores quantitativos, enquanto que, fabricantes de drogas, diferentes locais e meios de cultura são fatores qualitativos.

É importante para o planejamento e a análise estatística distinguirmos as seguintes situações:

- a) um pesquisador deseja conduzir um experimento para verificar o efeito do starter "*Lactobacillus plantarum*" em salame tipo italiano sobre o tempo de maturação dos mesmos.

Para essa finalidade utilizou 3 concentrações do starter, quais sejam:  $2,5 \times 10^5$ ;  $5,0 \times 10^5$  e  $10,0 \times 10^5$  células viáveis/grama de massa. Ele deseja saber se há diferenças entre as 3 concentrações. Portanto, nesse experimento, temos um fator (starter) de *efeito fixo*, isto é, as concentrações foram definidas pelo pesquisador, ou seja, não foi feita uma escolha aleatória dos níveis, assim, as conclusões desse experimento se referem apenas as concentrações utilizadas no experimento.

b) um tecnologista quer comparar a qualidade de pão fabricado por diferentes padarias da cidade de Florianópolis. Neste caso, as padarias são os tratamentos. Ele deseja que seus resultados sejam válidos para todas as padarias de Florianópolis, então, em lugar de escolher intencionalmente algumas padarias que pretende compará-las, deverá sorteá-las a partir de algum procedimento que garanta a aleatoriedade, como por exemplo, a tabela de número aleatórios, dada no apêndice 7, assim, o pesquisador estará fazendo um experimento onde o fator é dito *aleatório*. Para saber se um efeito é aleatório, verifique se os tratamentos em comparação representam uma *amostra aleatória* de uma população. Se os tratamentos não são uma amostra aleatória, o efeito é fixo.

Sobre as conclusões pode-se dizer: 1) os fatores são fixos: neste caso, os resultados (conclusões) são válidos apenas para os níveis do fator que estão presentes no experimento e 2) os fatores são aleatórios: as conclusões são válidas para a população de níveis.

Os tratamentos são selecionados pelo pesquisador e deve ser feita de acordo com os objetivos do trabalho. Um bom conhecimento do material experimental e alguma idéia sobre os efeitos dos tratamentos são muito úteis para dar mais objetividade aos trabalhos.

As conclusões de um experimento dependem de como os dados foram coletados. No projeto 1 estudou-se 4 alturas de decepas e procurou-se controlar a idade das plantas por meio da blocagem, isto é, foram agrupadas as árvores com idades próximas). Caso não tivesse sido feito esse controle, o pesquisador não saberia dizer se as diferenças médias da variável resposta seriam devido às alturas de decepas ou da idade das árvores. Esse fato é conhecido como *confundimento* de fatores e a variável idade é conhecida como *variável de perturbação* (em inglês: *nuisance variable*).

### 1.3.2 Tratamento Controle

É necessário quando não se conhece a eficiência dos tratamentos em estudo, ou quando a eficiência dos tratamentos é conhecida mas não é consistente em todas as condições. Nem todos os experimentos necessitam do tratamento controle. Exemplo: alta pressão em tem-

peratura ambiente ( $25^{\circ}C$ ) e o experimento foi feito a  $2^{\circ}C$  (frango cru). O que é? O tratamento controle consiste em se realizar todos os procedimentos que são feitos nas unidades experimentais usadas para os outros tratamentos, exceto a aplicação do efeito em estudo. Exemplo: num estudo sobre aditivos em alimentos, um tratamento pode consistir de uma porção de um vegetal contendo um aditivo particular que é servido a um degustador. O tratamento controle consistiria de uma porção do mesmo vegetal servido ao degustador, na mesma situação experimental, exceto que não seria utilizado o aditivo no alimento. É fundamental que o tratamento controle seja conduzido nas mesmas condições experimentais dos outros tratamentos.

### 1.3.3 Variáveis e Covariáveis

O que pode constituir problema, as vezes, é a forma como a variável é medida, pois disso depende a precisão das observações e o tipo de análise a ser executada. Exemplo: se os valores da variável sabor de um alimento é dada numa escala de 1 a 10, pode-se aumentar a precisão e facilitar a análise, utilizando-se como observação, a média de 3 valores da mesma unidade experimental.

Quando temos uma variável que influencia as variáveis dependentes, chama-se a mesma de *covariável*. Exemplos: 1) se o tempo necessário para executar um experimento é 30 dias, e se a temperatura do ambiente tem influência na variável dependente (resposta), então, a temperatura deve ser mantida constante. Se isso não for possível, então, deve-se medir a temperatura para cada unidade experimental; 2) num experimento para comparar 4 meios de cultura em frangos congelados onde a variável dependente é a população de *Staphilococcus aureus*, cada frango apresenta uma população inicial de *Staphilococcus* diferente, neste caso, a população inicial de *Staphilococcus* é a covariável; 3) Num experimento para estudar a produção de 10 variedades de soja, o número de sementes que germinam nos canteiros é a covariável. Estas variáveis entram na análise como covariáveis e, observe que elas não podem ser controladas pelo pesquisador. Este fato é que as diferencia de uma variável de perturbação (*nuisance variable*).

### 1.3.4 Repetição e Casualização

Para que a metodologia estatística possa ser aplicada aos resultados de um experimento, é necessário obedecer a dois princípios básicos da experimentação, o da *repetição* e da *aleator-*



ização dos tratamentos. Um terceiro princípio, o *controle local*, pode ou não ocorrer num experimento.

A *Repetição* consiste, como o próprio nome indica, em repetir o mesmo tratamento várias vezes. O uso de repetições dos tratamentos é necessário para podermos calcular a variabilidade e, com isso, executar os testes estatísticos e, também, para fazer estimação intervalar dos efeitos dos tratamentos. De um modo geral, quanto maior o número de repetições, mais precisas vão ser as nossas estimativas. Na prática, o número de repetições vai depender muito dos recursos e material experimental disponível. O cálculo do tamanho da amostra é um dos principais itens do planejamento de um experimento e, a sua determinação, não é trivial, e exige que se tenha algum conhecimento sobre a variabilidade dos dados, a precisão e confiança desejadas nos resultados. Então, as principais finalidades do uso de repetições são: 1) dar uma estimativa do erro experimental; 2) aumentar a precisão de um experimento, reduzindo o desvio padrão das médias dos tratamentos e 3) estimação e testes de hipóteses. O que caracteriza uma repetição é que ela deve gerar um resultado independente.

A *aleatorização ou casualização* consiste no sorteio dos tratamentos às unidades experimentais por um processo bem definido, fixo; é necessária para termos certeza de que um tratamento não seja beneficiado ou prejudicado por alguma causa conhecida ou desconhecida, tais como: intensidade de luz, constituição genética, temperatura, umidade, ventilação, etc. É também fundamental para atender a suposição de que os dados são oriundos de uma amostra aleatória. Os métodos estatísticos requerem que as observações (ou os erros), sejam variáveis aleatórias independentemente distribuídas. A casualização faz com que esta suposição seja válida. O princípio da casualização é uma das principais contribuições dos estatísticos à ciência experimental, principalmente Ronald A. Fisher (1890 - 1962). Só a casualização garante que unidades com características diferentes tenham igual probabilidade de serem designadas para os diferentes tratamentos. Com a casualização, obtemos estimativas não tendenciosas das médias dos tratamentos e das diferenças entre as médias; obtemos uma estimativa não tendenciosa do erro experimental. Certas restrições podem ser incluídas na casualização (*controle local*), para levar em consideração alguma(s) fonte(s) de variação do material experimental. O delineamento em blocos completos ao acaso apresenta uma restrição.

## 1.4 Exercícios Propostos

- 1 - Planeje um experimento para comparar a produção de cinco variedades de milho.

2 - Planeje um experimento para testar o efeito da adubação nitrogenada (5 níveis), sobre a produção de milho.

3 - Planeje um experimento na sua área de pesquisa.

Para a resolução desses exercícios, entende-se que um experimento está planejado quando estão definidos:

1. enunciado do problema com formulação do objetivo geral, dos objetivos específicos e da(s) hipótese(s);
2. escolha do(s) fator(es) e de seus níveis que deve(m) ser incluído(s) no estudo; Faça uma descrição dos mesmos;
3. as variáveis respostas ou dependentes em análise e a forma como serão medidas;
4. a unidade experimental;
5. decidir sobre o número de unidades experimentais a serem associadas a cada tratamento;
6. a forma (maneira) como os tratamentos serão designados às unidades experimentais;
7. bibliografia.

## 2 Análise Exploratória de Dados

### 2.1 Organização, Resumo e Representação de Dados

#### 2.1.1 Introdução

Como já comentamos anteriormente, a parte da Estatística que trata da organização, apresentação, resumo e descrição dos dados é conhecida como Análise Exploratória de Dados. Esta parte, geralmente limitava-se a construção de alguns tipos de gráficos (linhas, colunas e setores) e ao cálculo de algumas medidas de tendência central e de variabilidade, como, por exemplo, a média e a variância. Atualmente, foram desenvolvidas muitas outras técnicas (Tukey, 1971), principalmente visuais, através das quais procura-se estudar a regularidade presente nos dados. Esta análise permite que o pesquisador adquira um bom conhecimento e senso crítico sobre os seus dados observados.

Neste capítulo, através da Análise Exploratória de Dados, procuraremos tirar o máximo de informações de um conjunto de dados, ou seja, fazer todas as interpretações necessárias para responder aos objetivos de uma pesquisa.

Quando estamos trabalhando com um conjunto de dados, é bastante provável que o mesmo apresente algum tipo de regularidade, ou seja, um padrão de variação. Devido a esta regularidade presente nos dados é possível ajustar-se um *modelo*. Este é um dos principais objetivos da análise exploratória de dados, isto é, procurar estabelecer um modelo para um conjunto de dados, o qual possa ser utilizado na análise estatística inferencial. Exemplo do que seja um modelo no aspecto geral: vamos verificar o tipo de relacionamento entre a taxa de crescimento de uma pastagem cultivada no Planalto Catarinense e a temperatura do solo a 10 cm de profundidade no período de junho a novembro. A Figura 2.1 mostra a distribuição dos pontos entre essas duas variáveis. De modo visual, podemos verificar que existe uma relação linear entre a temperatura do solo e a taxa de crescimento, desse modo, podemos traçar uma reta a "olhômetro", o mais próximo possível de todos os pontos, que será o nosso *modelo*. Evidentemente que os pontos não caem sobre a reta (pode ocorrer para alguns). A diferença entre os dados e o modelo é chamada de *erro*. Explicação sobre o termo do erro foi dado na Seção 1.1.2.

Chama-se de *modelo* a parte da variabilidade dos dados que é explicada pelo mesmo e *erro* a parte da variabilidade dos dados não explicada pelo modelo. A Figura 2.2 ilustra os Dados, a parte do modelo e a parte do erro para uma observação. Os dois componentes são igualmente importantes. O estudo denominado de Análise de Resíduos, nos fornece

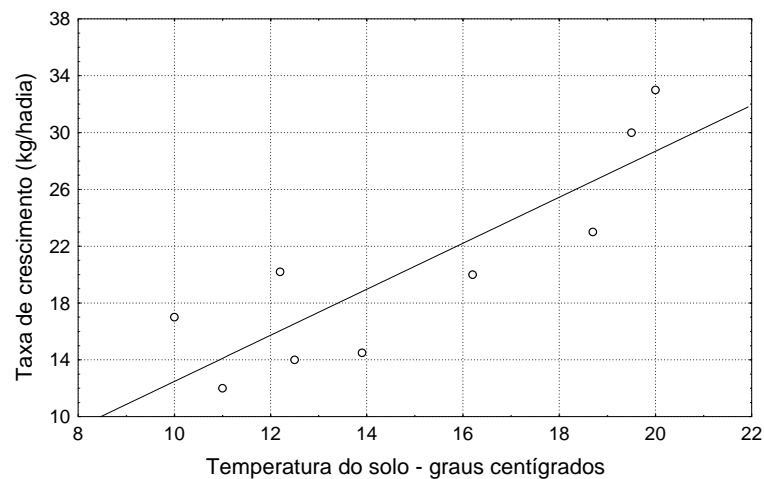


Figura 2.1: Relação entre temperatura do solo a 10 cm de profundidade e taxa de crescimento de uma pastagem de inverno no Planalto Catarinense

informação se a parte do modelo é adequada ou não para representar os dados, dentre outros aspectos importantes da análise de dados.

### 2.1.2 Classificação das Variáveis Seleccionadas para o Estudo

Um pesquisador quando está realizando um determinado experimento ou levantamento, necessita avaliar certas características nas plantas, nos animais, instrumentos, pessoas, etc. Por exemplo, registrar o diâmetro altura de peito (DAP) de Avicénias do manguezal do Itacorubi, a produção de milho por hectare, a resistência ao ataque de pragas do feijão, fazer a contagem do número de vagens por planta, número de grãos por vagem, etc. Estas características chamam-se variáveis, porque originam valores que tendem a variar quando se fazem medidas sucessivas, por exemplo, vamos supor que plantemos quatro áreas de  $30m^2$  de um híbrido de milho, com certeza vamos obter quatro valores diferentes de produção nessas áreas (unidades) experimentais.

Um pesquisador deve aprender a identificar quatro tipos de variáveis, que serão descritas através de um exemplo.

**Exemplo 2.1** Um pesquisador instalou um experimento para avaliação do comportamento de híbridos de milho, para isso, tomou algumas medidas agrônômicas que estão apresentadas na Tabela 2.1.

Muitas variáveis, tais como, tipo de grão e resistência à ferrugem, apresentam como

Tabela 2.1: Resultados de um experimento de competição de híbridos de milho para a região preferencial I com altitudes abaixo de 800m - safra:1987/1988

Híbridos	Rendimento médio (kg/ha)	Ciclo (dias)	Altura planta (cm)	Altura espiga (cm)	Tipo grão	Ferrugem <sup>1</sup> (escala)
1	6388	65	242	103	dentado	r
2	6166	65	258	134	semi-dentado	r
3	6047	65	240	104	semi-dentado	s
4	5889	66	243	108	semi-dentado	s
5	5823	69	257	128	dentado	ms
6	5513	68	241	108	semi-dentado	s
7	5202	64	235	108	dentado	r
8	5172	68	240	103	dentado	s
9	5166	69	253	123	dentado	ms
10	4975	70	250	117	semi-dentado	ms
11	4778	70	242	114	dentado	mr
12	4680	66	245	111	semi-duro	ms
13	4660	69	239	110	semi-duro	mr
14	5403	73	264	138	dentado	ms
15	5117	76	282	149	dentado	mr
16	5063	72	274	151	dentado	r
17	4993	71	279	134	semi-dentado	r
18	4980	72	274	140	dentado	ms
19	4770	73	244	140	dentado	r
20	4685	71	265	139	semi-duro	mr
21	4614	73	248	110	semi-dentado	r
22	4552	73	265	128	semi-dentado	r
23	3973	74	261	124	semi-dentado	mr
24	4550	71	259	129	semi-duro	s
25	5056	64	252	104	semi-duro	mr
26	4500	70	271	109	dentado	ms
27	4760	68	243	137	semi-duro	r
28	5110	66	252	141	semi-dentado	ms
29	4960	70	262	120	dentado	ms
30	4769	73	260	118	dentado	r
31	4849	74	250	119	semi-dentado	s
32	5230	71	255	138	semi-duro	s

<sup>1</sup>r=resistente; mr=moderadamente resistente; ms=m. susceptível; s=susceptível

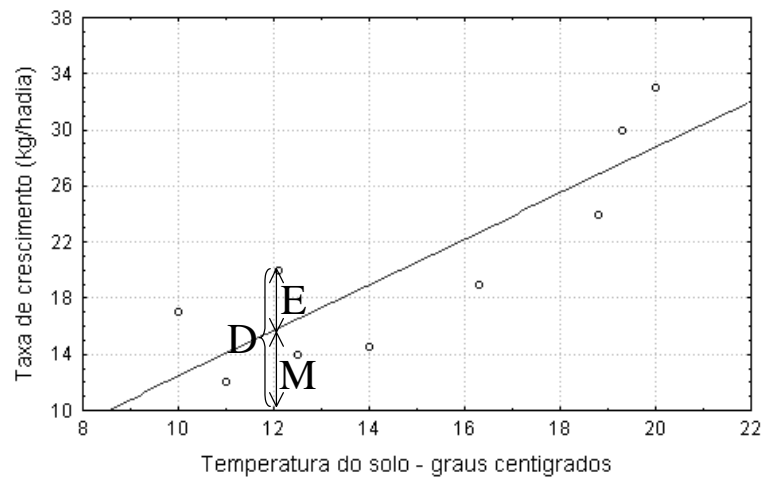


Figura 2.2: Os componentes de um modelo

resultado uma *qualidade* ou *atributo*, e outras variáveis, tais como, rendimento médio, ciclo da cultura, apresentam como resultado *medidas* ou *contagens*. As variáveis do primeiro tipo são chamadas de variáveis *qualitativas* e as do segundo tipo são chamadas de variáveis *quantitativas*.

Dentre as variáveis qualitativas podemos distinguir dois tipos:

- *variáveis qualitativas ordinais*: para estas variáveis existe uma ordem nos possíveis resultados da mesma. No exemplo, temos a resistência à ferrugem, dada numa escala. Outro exemplo podem ser, 1<sup>º</sup> grau, 2<sup>º</sup> grau, superior.
- *variáveis qualitativas nominais*: para estas variáveis não existe uma ordenação nos resultados. No exemplo, temos o tipo de grão como variável desse tipo. Outros exemplos, podem ser: germina/não germina, os cursos da UFSC (Agronomia, Ciências Biológicas, etc.), a cor da flor de soja.

Dentre as variáveis quantitativas também podemos distinguir dois tipos:

- *variáveis quantitativas discretas*: estas variáveis só podem assumir certos valores, em geral números inteiros e normalmente são resultantes de contagens. É possível formar uma lista (finita ou infinita) dos valores. No exemplo, a variável ciclo da cultura é discreta. Outros exemplos, podem ser: número de dias da emergência à floração, número de vagens por planta, número de grãos por vagem.
- *variáveis quantitativas contínuas*: estas variáveis assumem todos os valores possíveis dentro de um determinado intervalo. Esta variáveis, dependendo da precisão uti-

lizada na medição, são capazes de diferenciar animais, plantas para valores muito pequenos. São variáveis cujos resultados geralmente são fracionários. No exemplo, temos a variável rendimento médio de grãos de milho, altura da planta, altura de espiga.

A distinção entre variáveis contínuas e discretas é muitas vezes artificial, pois depende da aproximação (precisão) utilizada. Por exemplo, idade é uma variável de medida de tempo, portanto, por definição, é uma variável aleatória contínua, porém, em muitos casos ela é medida em anos completos (discretização da variável), o que a torna uma variável discreta. Serão tratadas como variáveis contínuas todas as que, pelo menos em teoria, possam assumir qualquer valor dentro de um intervalo.

Para cada tipo de variável existem técnicas mais apropriadas para resumir as informações; entretanto, vamos verificar que técnicas usadas num caso podem ser adaptadas para outros, pois é possível transformar variáveis quantitativas em qualitativas e vice-versa. Por exemplo, sexo:  $1 \rightarrow$  masculino e  $0 \rightarrow$  feminino; rendimento: colocar em categorias, por exemplo, baixa, média e alta produção. O estudo de probabilidades também apresenta os seus modelos de acordo com cada tipo de variável. Estes modelos serão estudados nas seções 3, 4 e 5.

### **2.1.3 Distribuições de Freqüências. Representação em Tabelas e Gráficos**

Feita a coleta dos dados, através de censos, de levantamentos por amostragem (*Survey*, em inglês), ou de experimentos, os mesmos apresentam-se, geralmente, de maneira desorganizada, ainda sem valor informativo sobre o fenômeno em estudo, portanto, os mesmos devem ser organizados e resumidos, para possibilitarem a obtenção de informações úteis para o trabalho de pesquisa.

O estudo das distribuições de freqüências nos permite conhecer a forma, a maneira como os valores de uma variável se comporta, isto é, é possível ter uma boa idéia global dos valores, ou seja, da distribuição. Uma distribuição de freqüências pode ser representada em forma de tabela ou gráfico.

#### **2.1.3.1 Distribuições de Freqüências de Variáveis Qualitativas. Representação em Tabelas**

Feita a coleta dos dados relativa às variáveis definidas no início da pesquisa, o investigador tem interesse, agora, em conhecer o comportamento dessas variáveis considerando os elementos pesquisados, por exemplo, conhecer a distribuição da variável resistência à ferrugem pesquisada em 32 híbridos de milho, na região de Chapecó, SC.

Este estudo pode ser feito através da construção de distribuições de frequências, chamando-se de frequência, por exemplo, o número de híbridos para a categoria S, MS, MR, e R, respectivamente. Chama-se, portanto, distribuição de frequências, a correspondência entre categorias ou valores possíveis de uma variável e as respectivas frequências. Inicialmente vamos representar as distribuições de frequências em tabelas.

Alguns aspectos importantes devem ser levados em consideração na construção de uma tabela, quais sejam:

1. Toda tabela deve conter um título completo; as três questões que devem ser respondidas num título são: o quê se está estudando? onde foi feito o estudo? e quando?. O título deve ser colocado na parte superior da tabela.
2. Se os dados não são próprios deve-se indicar a fonte dos mesmos. Ela vai na parte inferior da tabela.
3. As notas e chamadas são utilizadas para fazer esclarecimentos de ordem geral e específicas, respectivamente. Ambas são numeradas, geralmente em algarismos arábicos, pode-se, ainda, utilizar letras minúsculas ou símbolos, como por exemplo, asterisco. Também são colocadas na parte inferior da tabela.
4. Os totais e subtotais devem ser bem destacados.
5. De preferência usar o mesmo número de casas decimais para os algarismos.
6. Não devem ser fechadas lateralmente.
7. Quando algum valor da tabela é nulo pela própria natureza do fenômeno em estudo, deve-se utilizar o hífen (-) para substituí-lo. Quando não se tem informação sobre algum valor deve-se colocar três pontos (...). Se existe dúvida sobre a exatidão da informação deve-se usar ponto de interrogação (?). Se algum valor for omitido para evitar individualização, deve-se usar a letra  $x$ . O símbolo de parágrafo (§) é utilizado para retificar uma informação publicada anteriormente.

**Exemplo 2.2** A Tabela 2.1 apresenta os dados sobre resistência à ferrugem de 32 híbridos de milho (coluna 7). A distribuição de frequências para essa variável é apresentada na Tabela 2.2, cuja construção passamos a discutir.

Na Tabela 2.2, a primeira coluna mostra todas as categorias da variável resistência à ferrugem. Na segunda coluna temos as frequências absolutas, resultantes da contagem de



Tabela 2.2: Distribuição de frequências da resistência à ferrugem de 32 híbridos de milho recomendados para a região preferencial I com altitudes abaixo de 800m, 1987/88

Resistência à ferrugem	Frequência absoluta	Frequência relativa	Porcentagem	% acumulada de satisfeito
R	10	0,313	31,250	31,250
MR	6	0,188	18,750	50,000
MS	9	0,281	28,125	-
S	7	0,219	21,875	-
TOTAL	32	1,000	100,000	

quantas observações se identificam com cada categoria. A notação para as frequências absolutas é  $n_i$ , assim temos,  $n_1 = 10$ ,  $n_2 = 6$ ,  $n_3 = 9$ ,  $n_4 = 7$ . A terceira coluna apresenta uma medida relativa de cada frequência, obtida da divisão de cada frequência absoluta pelo número total de observações, resultando nas proporções ou frequências relativas ( $f_i = n_i/n$ ). Por exemplo, a frequência relativa da categoria  $R$  é calculada por:  $f_1 = 10 \div 32 = 0,3125$ . Multiplicando por 100 as frequências relativas, temos as porcentagens de ocorrências de cada resultado observado (coluna 4). A última coluna da tabela apresenta a porcentagem acumulada, dada pela porcentagem da respectiva coluna adicionada da(s) anterior(es) à ela. Por exemplo, os 50% da categoria MR é o resultado da adição de 18,75% e 31,25%. Observa-se na Tabela 2.2 que temos praticamente 50% de híbridos resistentes e 50% de susceptíveis. Deixamos as categorias MS e S com traços pois não tem sentido prático o cálculo das porcentagens acumuladas neste caso, ou seja, não tem sentido afirmarmos que 100% dos híbridos são susceptíveis.

Essas medidas relativas são particularmente importantes na comparação entre distribuições de frequências, como veremos a seguir em distribuições bidimensionais, isto é, quando temos duas variáveis em estudo. Quando temos mais do que duas variáveis o raciocínio é o mesmo.

### *Tabelas Bidimensionais*

A Tabela 2.3 mostra três distribuições de frequências, uma para cada região preferencial. Dizemos que esta tabela é bidimensional, pois apresenta a distribuição de duas variáveis, quais sejam: 1) resistência à ferrugem e 2) região.

Como os totais marginais da Tabela 2.3 são diferentes, e isto dificulta a interpretação, é interessante trabalhar com porcentagens, tornando, assim, os resultados comparáveis. As

Tabela 2.3: Distribuição de frequências da resistência à ferrugem de híbridos de milho, segundo as regiões preferenciais

Resistência à ferrugem	Regiões			TOTAL
	Chapecó	Campos Novos	Içara	
R	10	3	12	25
MR	6	12	2	20
MS	9	3	3	15
S	7	1	2	10
TOTAL	32	19	19	70

Tabela 2.4: Distribuição das porcentagens da resistência à ferrugem de híbridos de milho, para as regiões preferenciais (Perfis colunas)

Resistência à ferrugem	Regiões			TOTAL
	Chapecó	Campos Novos	Içara	
R	31,2	15,8	63,2	35,8
MR	18,8	63,2	10,5	28,6
MS	28,1	15,8	15,8	21,4
S	21,9	5,2	10,5	14,3
TOTAL	100,0	100,0	100,0	100,0

porcentagens podem ser calculadas de três formas: 1) fixando-se os totais de linhas em 100%; 2) fixando-se os totais de colunas em 100% e 3) fixando-se o total geral em 100%. Isto vai de acordo com o objetivo da pesquisa, uma delas será a mais adequada.

Na Tabela 2.4 fixamos os totais de colunas em 100%. Este tipo de distribuição serve para comparar a distribuição das regiões conforme a resistência à ferrugem, ou seja, para uma dada região podemos analisar a variável resistência à ferrugem. Com respeito a essa tabela, observamos que em Campos Novos a porcentagem de híbridos moderadamente resistentes é bem superior às outras duas regiões, 63,2% em Campos Novos contra 18,8% em Chapecó e 10,5% em Içara. Por outro lado, vemos que Içara apresenta a maior porcentagem de híbridos resistentes à ferrugem (63,2%), bem superior à Chapecó (31,2%) e Campos Novos (15,8%). Percebe-se que em Chapecó a distribuição é mais homogênea nas categorias da resistência.

Tabela 2.5: Distribuição das porcentagens da resistência à ferrugem de híbridos de milho (Perfis linhas)

Resistência à ferrugem	Regiões			TOTAL
	Chapecó	Campos Novos	Içara	
R	40,0	12,0	48,0	100,0
MR	30,0	60,0	10,0	100,0
MS	60,0	20,0	20,0	100,0
S	70,0	10,0	20,0	100,0
TOTAL	45,7	27,1	27,1	100,0

Tabela 2.6: Distribuição conjunta das porcentagens em relação ao total geral

Resistência à ferrugem	Regiões			TOTAL
	Chapecó	Campos Novos	Içara	
R	14,3	4,3	17,1	35,7
MR	8,6	17,1	2,9	28,6
MS	12,9	4,3	4,3	21,4
S	10,0	1,4	2,8	14,3
TOTAL	45,7	27,1	27,1	100,0

De outra forma, se desejamos saber, por exemplo, para os híbridos resistentes, qual a distribuição por região, devemos fixar os totais de linhas (Perfis linhas). Os resultados são dados na Tabela 2.5. Em Campos Novos observamos uma baixíssima porcentagem de híbridos resistentes à ferrugem (12%); Chapecó e Içara apresentam praticamente a mesma porcentagem, 40% e 48%, respectivamente. Para a categoria MR, Campos Novos se destaca com 60%. Dos MS e S, Chapecó se destaca com 60% e 70%, respectivamente.

Na Tabela 2.6, expressamos as porcentagens em relação ao total geral. Observamos que Campos Novos apresenta a maior porcentagem de híbridos moderadamente resistentes (17,1%). Içara apresenta a maior porcentagem de híbridos resistentes (17,1%), seguido de Chapecó com 14,3%.

### Exercício resolvido

1. Um estudo sobre o tempo de vida de duas amostras de *Biomphalaria straminea*

Tabela 2.7: Número de indivíduos classificados segundo o tempo de vida, em dias, e condição a que as populações foram submetidas

Tempo de vida em dias	Condição		Total
	Agrupadas	Isoladas	
58 a 179	3	6	9
180 a 300	6	19	25
301 a 422	36	24	60
Total	45	49	94

(Amostra A: indivíduos agrupados numa bacia e Amostra I: indivíduos isolados em copos de vidro), produziu os resultados da Tabela 2.7:

a) Dos indivíduos que tiveram tempo de vida entre 58 e 179 dias, qual a percentagem deles para a condição Agrupados? e Isolados?

b) Qual a percentagem de indivíduos para tempo de vida entre 180 e 300 dias e condição Isolados?

c) Dos indivíduos submetidos à condição de Agrupados, qual a percentagem de indivíduos teve tempo de vida entre 301 e 422?

Observou-se 9 indivíduos com tempo de vida entre 58 e 179 dias, portanto, as percentagens são dadas por:  $3/9=33,33\%$  e  $6/9=66,67\%$ , respectivamente. A resposta do item b) é:  $19/94=0,2021=20,21\%$ . Observou-se um total de 45 indivíduos submetidos a condição de Agrupados, portanto, a percentagem dos que tem tempo de vida entre 301 e 422 é:  $36/45=0,80=80\%$ .

### 2.1.3.2 Representações Gráficas de Variáveis Qualitativas

As representações gráficas de tabelas de distribuições de frequências permitem uma boa visualização da distribuição da variável. Podemos ter uma rápida e concisa visualização da variabilidade da variável. A utilização de gráficos para ilustrar os resultados de uma pesquisa sempre é recomendável. A construção de gráficos depende muito da habilidade artística de cada um. Não é objetivo deste livro entrar em detalhes sobre a construção de gráficos e, sim, somente comentar sobre alguns pontos que devem ser respeitados. Podemos citar:

1. Os gráficos devem ser claros, simples, atrair a atenção e inspirar confiança.
2. Servem para realçar certos aspectos importantes de uma pesquisa.
3. O tamanho deve ser adequado à sua publicação em revistas, periódicos, cartazes, livros, etc.
4. Sempre devem ter um título completo e deve ser colocado na parte inferior do gráfico.
5. Devem ser construídos numa escala que não desfigure os fatos ou as relações que se deseja destacar.
6. Devem ser mais largos do que altos.
7. Deve-se sempre especificar (dar nome) e graduar (criar escala) os eixos.
8. Quando os dados não são próprios deve-se citar a fonte. Esta deve ser colocada na parte inferior do gráfico.
9. Pode-se usar notas para esclarecimentos gerais.

Nesta seção veremos os *gráficos de barras*, *colunas*, *setores* e *linhas*, que são particularmente importantes na representação de dados categorizados. Mostraremos os tipos padrões de gráficos. Os programas estatísticos como o Statistica, Statgraphics, Minitab, SPSS, SAS e, também, planilhas eletrônicas como o Excel, fornecem uma gama enorme de gráficos.

Os *gráficos de barras* tem por finalidade comparar grandezas, por meio de retângulos de igual largura, dispostos horizontalmente, e com alturas proporcionais as grandezas. Deve-se deixar uma distância entre os retângulos. É usado quando as inscrições a serem colocadas sob os retângulos forem muito extensas. Sempre que possível ordenar as barras em ordem decrescente. A Figura 2.3 mostra o gráfico de barras da distribuição da Tabela 2.2.

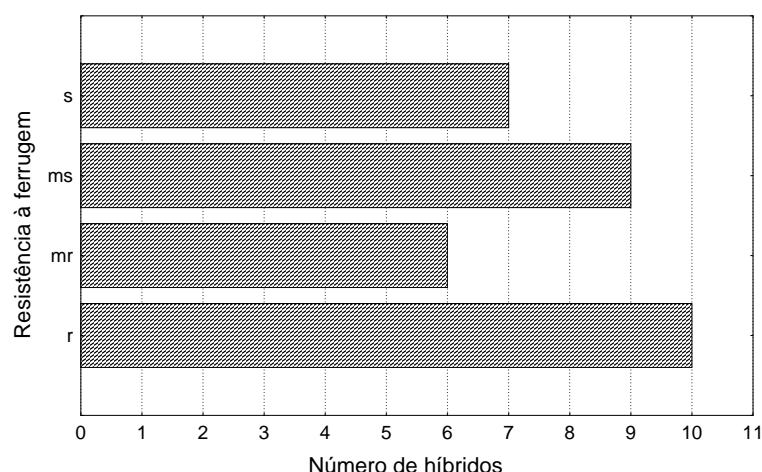


Figura 2.3: Distribuição de frequências da resistência à ferrugem de híbridos de milho, para a região preferencial I, 1987/88

Para efetuar uma análise comparativa de várias distribuições, podemos construir um *gráfico de barras múltiplo*. A Figura 2.4 mostra o gráfico de barras múltiplo da distribuição do tipo de grão e da resistência à ferrugem, cujos dados estão na Tabela 2.1. Observe a necessidade de construção de uma legenda.

Quando os retângulos são colocados na posição vertical, temos os *gráficos de colunas*. A finalidade desse tipo de gráfico é a mesma dos gráficos de barras, isto é, servem para comparar grandezas. Eles devem ser preferidos aos gráficos de barras quando as legendas a se inscreverem sob os retângulos forem pequenas. Na Figura 2.5 temos o gráfico de colunas para tipo de grão (os dados estão na Tabela 2.1, coluna 6).

O *gráfico de setores* consiste em dividir a área total de um círculo em subáreas (setores) proporcionais às frequências. É um gráfico ideal para representar dados de porcentagens. O número de setores deve ser adequado. Considerando o tipo de grão dentado da Tabela 2.1, temos as seguintes porcentagens, 42,9%, 35,7%, 14,3% e 7,1%, para as categorias, moderadamente susceptível, resistentes, moderadamente resistentes e susceptível, respectivamente, cujo gráfico é dado na Figura 2.6.

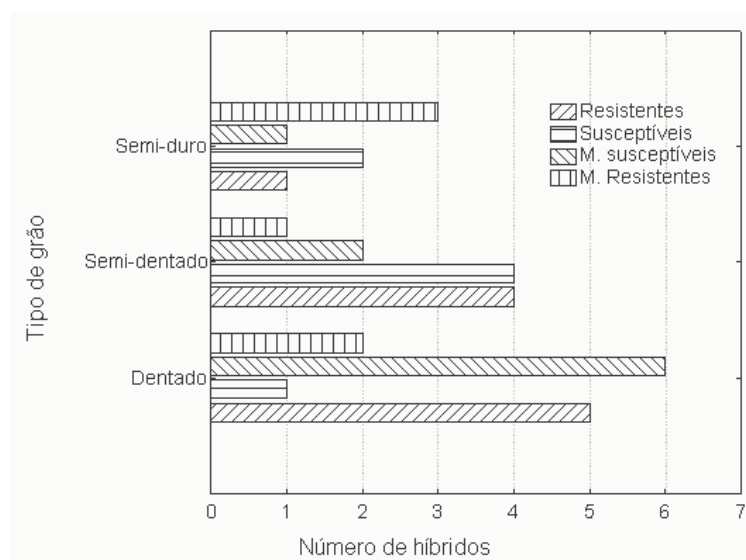


Figura 2.4: Distribuição das freqüências do tipo de grão e resistência à ferrugem de híbridos de milho, para a região I, 1987/88

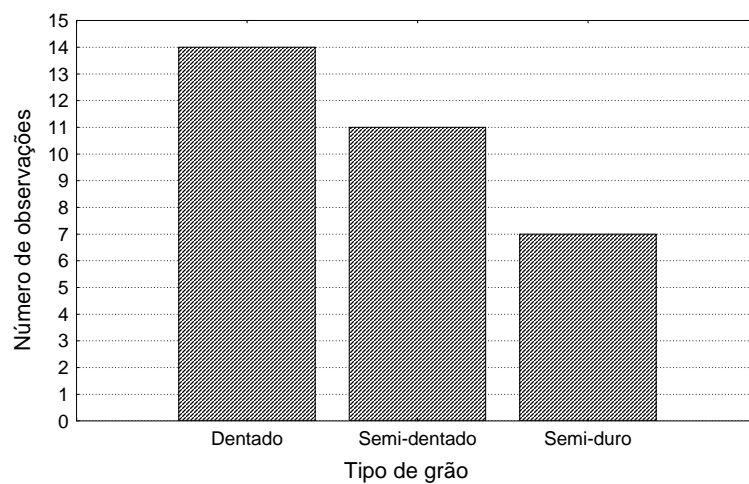


Figura 2.5: Distribuição das freqüências dos tipos de grão para 32 híbridos de milho, para a região I, 1987/88

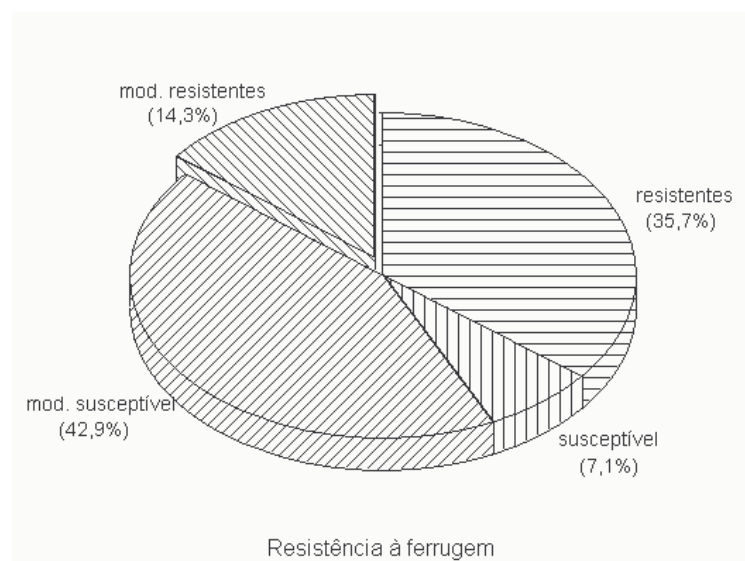


Figura 2.6: Distribuição das proporções (%) da resistência à ferrugem, para o tipo de grão dentado, para a região I, 1987/88

Os *gráficos de linhas* são adequados para dados ordenados ao longo do tempo (séries temporais) e servem para comparar distribuições. Na Figura 2.7, apresentamos o gráfico de linhas da variável acidez em ácido láctico (%) de 4 tratamentos, denominados aqui por PA, PB, PC e C.

### Exercícios resolvidos

1. Um pesquisador está procurando verificar se existe associação entre hábito de crescimento (3=indeterminado trepador e 4=indeterminado prostrado) e porte (Tr=trepador, EB=ereto na base e Pr=prostrado) na cultura de feijão de vagem. Para esse fim foi conduzido um experimento na UFSC, safra 1991/92, cujos resultados estão na Tabela 2.8.

a) Construa a distribuição de frequência conjunta para as variáveis hábito de crescimento e porte. Resposta Tabela 2.9.

b) Faça um gráfico para a distribuição de frequência conjunta do item a). Resposta Figura 2.8.

2. Faça um gráfico para a distribuição de frequências conjunta da tabela 2.7 usando no eixo das ordenadas as percentagens em relação ao total da tabela. Resultado: veja figura 2.9.

### 2.1.3.3 Exercícios Propostos

1. Um pesquisador está procurando verificar se existe associação entre local (1= Rio vermelho; 2= Costa da lagoa) e número de abortos para crianças do sexo feminino. Para



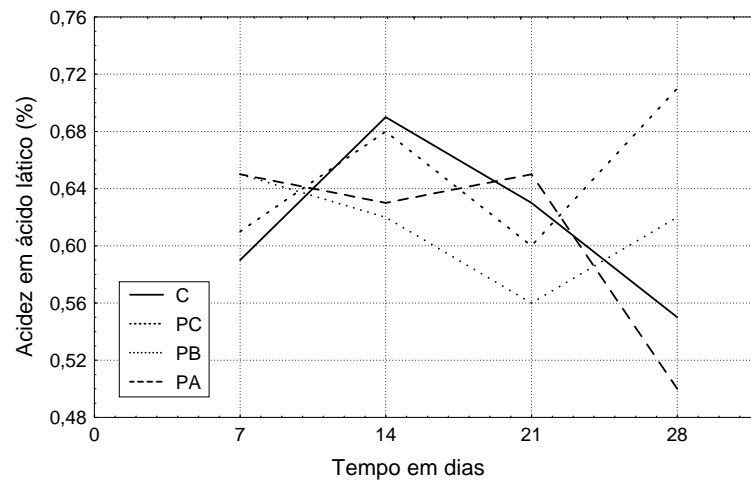


Figura 2.7: Comportamento da variável acidez em ácido láctico (%) nos diversos tratamentos durante a maturação de salame tipo italiano,UFSC, 1992

Tabela 2.8: Hábito e porte para 50 materiais de feijão de vagem.

H	P	H	P	H	P	H	P	H	P
4	Tr	4	Tr	4	Tr	4	Pr	4	Tr
4	EB	4	Tr	4	Tr	4	Tr	3	Pr
3	Pr	3	Pr	3	Tr	4	Pr	3	Pr
4	Tr	3	Pr	4	Tr	3	Pr	3	Pr
4	Tr	3	Pr	4	Tr	4	Tr	4	Tr
4	Tr	3	EB	4	Tr	3	Pr	4	Tr
3	Pr	4	EB	4	Tr	4	Pr	4	Tr
3	EB	4	EB	4	Tr	3	Pr	4	Tr
4	Tr	4	Tr	3	Pr	4	Tr	3	Pr
4	Tr	4	Tr	4	Tr	4	Tr	4	Tr

Tabela 2.9: Distribuição conjunta do hábito de crescimento e porte na cultura de feijão de vagem, UFSC, 1991/92

Porte	Hábito de crescimento		Total
	3	4	
Trepador	1 6,3%	28 82,4%	29 58,0%
Ereto na base	2 12,5%	3 8,8%	5 10,0%
Prostrado	13 81,3%	3 8,8%	16 32,0%
Total	16 100,0%	34 100,0%	50 100,0%

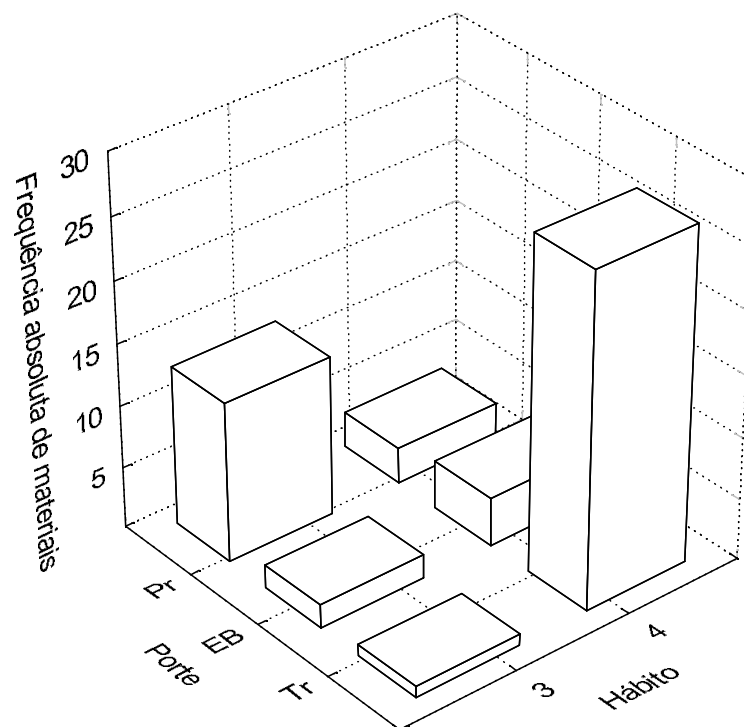


Figura 2.8: Gráfico da distribuição conjunta do item a).

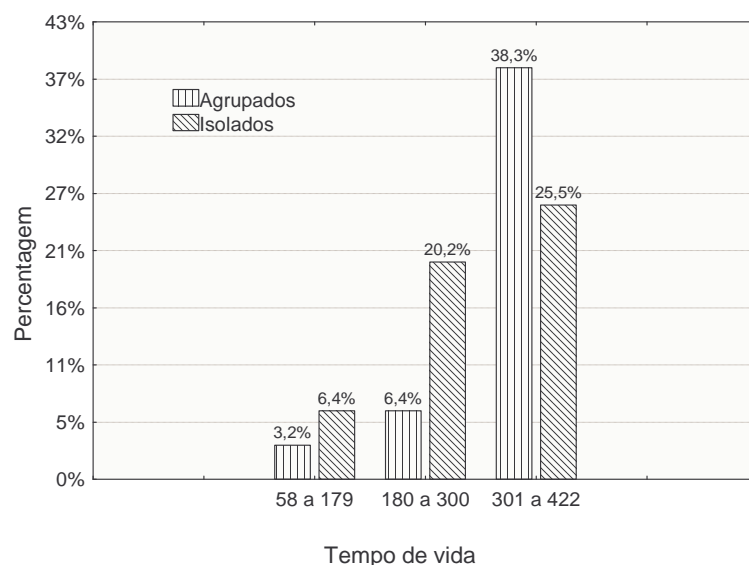


Figura 2.9: Gráfico da distribuição conjunta do item a).

esse fim foi feito um levantamento, cujos resultados estão apresentados na Tabela 2.10 (*Obs:* a unidade de amostragem é uma família):

a) Construa a distribuição de freqüências conjunta para as variáveis local e número de abortos.

b) Faça um gráfico para a distribuição de freqüências conjunta do item a).

2. Um economista agrícola está estudando os fatores que afetam a adoção de uma nova variedade de arroz altamente produtiva. Os resultados obtidos estão na Tabela 2.11.

a) Faça um gráfico para a Tabela 2.11, mostrando o comportamento da adoção segundo a situação de posse da terra (perfil linha). Faça a interpretação dos resultados.

3. Foi conduzido um experimento com o objetivo de avaliar o poder germinativo de duas cultivares de cebola: A) Bola Precoce-EMPASC 352 e B) Norte 14. Foram utilizadas para o teste de germinação 4 repetições de 100 sementes, totalizando 400 sementes para cada cultivar. A variável em estudo é o número de sementes que germinaram. Os resultados obtidos estão na Tabela 2.12. Faça um gráfico mostrando o comportamento das cultivares com relação à germinação das sementes. Faça a interpretação dos resultados.

4. A Tabela 2.13 representa a distribuição conjunta das variáveis local de coleta e peso de mexilhões. Faça um gráfico para representar a distribuição conjunta. Obtenha uma conclusão relevante.

#### 2.1.3.4 Tabelas de Contingência

É muito freqüente nas Ciências Biológicas o interesse em verificar se duas variáveis quali-

Tabela 2.10: Resultados de um estudo sobre locais e ocorrência de aborto

Local	Aborto	Local	Aborto	Local	Aborto	Local	Aborto	Local	Aborto
2	0	2	0	2	0	1	1	1	0
2	0	2	1	2	0	1	1	1	0
2	1	2	1	2	1	1	1	1	0
2	0	2	1	2	1	1	0	1	0
2	0	2	1	2	2	1	2	1	0
2	0	2	1	2	0	1	1	1	0
2	0	2	0	2	0	1	0	1	0
2	2	2	0	2	0	1	0	1	0
2	2	2	0	2	0	1	0	1	0
2	0	2	0	2	1	1	0		
2	1	2	0	1	0	1	0		
2	1	2	0	1	0	1	0		
2	0	2	0	1	0	1	0		
2	0	2	0	1	1	1	0		

Tabela 2.11: Distribuição conjunta de frequências

Posse	Adoção		Total
	Adota	Não adota	
Proprietário	102	26	128
Vários arrendatários	42	10	52
Único arrendatário	5	2	7
Total	149	38	187

Tabela 2.12: Germinação de sementes para duas cultivares de cebola.

Cultivares	Germinação		Total
	Germinaram	Não germinaram	
Bola Precoce	392	8	400
Norte 14	381	19	400
Total	773	27	800

Tabela 2.13: Distribuição conjunta das variáveis local e peso de mexilhões

Local	Peso			Total
	[7;20)	[20;33)	[33;46)	
Mangue	21	12	2	35
Sambaqui	1	14	19	34
Total	22	26	21	69

tativas apresentam-se *associadas*, isto é, se o conhecimento de uma variável ajuda a entender uma outra variável. Construindo uma distribuição de frequência conjunta das duas variáveis, ou seja, uma tabela de contingência, podemos satisfazer de forma exploratória esse objetivo. Vamos verificar que a simples construção da distribuição conjunta das frequências será um poderoso instrumento para ajudar na compreensão dos dados. Como veremos, as tabelas construídas na seção anterior, são chamadas de tabelas de contingência e serão agora introduzidas formalmente.

Para se construir uma tabela de contingência, deve-se observar conjuntamente as duas variáveis nos elementos em estudo. Por exemplo, vamos observar para cada aluno, o grau de satisfação com o curso e o preparo dos professores. Se desejamos descrever a variável grau de satisfação com o curso, sabendo-se que a mesma tem associação com o preparo dos professores, fica mais fácil compreender a primeira variável.

Vamos considerar que a variável  $A$  tenha  $s$  categorias,  $A_1, A_2, \dots, A_s$  e a variável  $B$  tenha  $r$  categorias,  $B_1, B_2, \dots, B_r$ . Por exemplo, a variável grau de satisfação com o curso tem 5 categorias, quais sejam: 1 - muito bom; 2 - bom; 3 - médio; 4 - baixo e 5 - muito baixo. A variável preparo dos professores também tem 5 categorias, no caso, idênticas à primeira variável.

Se observarmos as duas variáveis,  $A$  e  $B$ , em um grupo de  $n$  elementos, objetos ou pessoas, teremos uma classificação como a indicada na Tabela 2.14, a qual é chamada de *tabela de contingência*.

Na Tabela 2.14,  $n_{11}$  representa o número de elementos classificados na categoria  $A_1$  e  $B_1$  simultaneamente, em outras palavras, é a frequência observada de elementos que pertencem a categoria  $A_1$  e  $B_1$  simultaneamente. Os valores  $n_{i.} = \sum_{j=1}^r n_{ij}$ ,  $n_{.j} = \sum_{i=1}^s n_{ij}$  e  $n_{..} = \sum_{i=1}^s \sum_{j=1}^r n_{ij}$  representam os totais de linhas, colunas e o total geral, respectivamente.

**Exemplo.** Na Tabela 2.15 apresenta-se a distribuição conjunta da resistência à ferrugem e tipo de grão para os 32 híbridos de milho.

Cada casela dá a frequência observada de híbridos que pertencem a categoria  $A_i$  e  $B_j$ ,

Tabela 2.14: Distribuição conjunta de frequências das variáveis A e B, observados em  $n$  elementos

A	B				TOTAIS
	$B_1$	$B_2$	...	$B_r$	
$A_1$	$n_{11}$	$n_{12}$	...	$n_{1r}$	$n_{1.}$
$A_2$	$n_{21}$	$n_{22}$	...	$n_{2r}$	$n_{2.}$
.	.	.	...	.	.
$A_s$	$n_{s1}$	$n_{s2}$	...	$n_{sr}$	$n_{s.}$
TOTAIS	$n_{.1} = \sum_{i=1}^s n_{i1}$	$n_{.2} = \sum_{i=1}^s n_{i2}$	...	$n_{.r} = \sum_{i=1}^s n_{ir}$	$n_{..} = \sum_{i=1}^s \sum_{j=1}^r n_{ij}$

Tabela 2.15: Distribuição conjunta das frequências das variáveis resistência à ferrugem e tipo de grão para 32 híbridos de milho, 1987/88.

Tipo de grão	Resistência à ferrugem <sup>1</sup>				Total
	R	MR	MS	S	
Dentado	5	2	6	1	14
Semi-dentado	4	1	2	4	11
Semi-duro	1	3	1	2	7
Total	10	6	9	7	32

<sup>1</sup> R = resistente; MR = moderadamente resistente; MS = moderadamente susceptível; S = susceptível.

Tabela 2.16: Distribuição conjunta das percentagens das variáveis resistência à ferrugem e tipo de grão para 32 híbridos de milho, 1987/88.

Tipo de grão	Resistência à ferrugem <sup>1</sup>				Total
	R	MR	MS	S	
Dentado	35,7	14,3	42,9	7,1	100,0
Semi-dentado	36,4	9,1	18,2	36,4	100,0
Semi-duro	14,3	42,8	14,3	28,6	100,0
Total	31,2	18,8	28,1	21,9	100,0

<sup>1</sup>R = resistente; MR = moderadamente resistente; MS = moderadamente susceptível; S = susceptível.

$i = 1, 2, 3$  e  $j = 1, 2, 3, 4$ , simultaneamente. Assim, observamos na casela (1;1), 5 híbridos resistentes à ferrugem e com o tipo de grão dentado, e assim por diante.

Observem que os totais de colunas nos fornece a distribuição de frequências da variável resistência à ferrugem, enquanto os totais de linhas nos dão a distribuição de frequências da variável tipo de grão. Tecnicamente estas distribuições são chamadas de *distribuições marginais*, então, temos a distribuição marginal da variável resistência à ferrugem (totais de colunas) e a distribuição marginal do tipo de grão (totais de linhas), enquanto que a tabela toda representa a *distribuição conjunta* das duas variáveis.

Como os totais marginais da Tabela 2.15 são diferentes, torna-se difícil fazer alguma interpretação. Para facilitar, podemos incluir as frequências relativas, em porcentagem, que podem ser calculadas em relação aos totais de linhas, colunas ou em relação ao total geral. Os totais (linhas, colunas ou geral) em porcentagens são comparáveis, pois reduz-se as frequências a um mesmo total.

A Tabela 2.16 apresenta as porcentagens calculadas em relação aos totais de linhas. Sendo assim, vamos estudar a distribuição da resistência à ferrugem para cada um dos diferentes tipos de grãos (os perfis são as linhas).

Interpretação: observa-se, para os híbridos selecionados, um indicativo de *associação* entre a resistência à ferrugem e o tipo de grão, pois, enquanto o tipo de grão semi-duro apresenta a menor porcentagem de resistência à ferrugem (14,3%), o tipo de grão semi-dentado apresenta a menor porcentagem de moderadamente resistentes (9,1%) e o tipo de grão dentado apresenta a menor porcentagem de híbridos suscetíveis à ferrugem (7,1%). Podemos verificar esta *associação* de outra forma. Observamos na amostra selecionada que 31,2% dos híbridos são resistentes (R) à ferrugem. Ora, se não existe associação (dependência)

Tabela 2.17: Contagens de plantas segregando para dois caracteres numa progênie da espécie "X"

Ciclo	Virescência		Total
	Normal	Virescente	
Tardio	3470 77,11%	910 75,83%	4380 76,84%
Precoce	1030 22,89%	290 24,17%	1320 23,16%
Total	4500 100,00%	1200 100,00%	5700 100,00%

entre as variáveis, esperaríamos esta mesma porcentagem (31,2%) para todos os 3 tipos de grãos. Observamos, na amostra de híbridos, uma porcentagem de 35,7% para dentado, 36,4% para semi-dentado e 14,3% para semi-duro, este último resultado bastante abaixo dos 31,2% esperados. Isto nos leva a acreditar que realmente existe associação entre a resistência à ferrugem e o tipo de grão dos híbridos de milho amostrados.

### Exercícios resolvidos

1. Para os dados da Tabela 2.8 podemos considerar que o hábito está associado com o porte? justifique.

Sim, pois verificamos que para o hábito de crescimento 3, 81,3% dos materiais apresentam porte prostrado. Por outro lado, para hábito 4, 82,4% dos materiais apresentam porte trepador. Procure entender esta associação através do gráfico da Figura 2.8.

2. Os dados da tabela de contingência 2.17 tem por objetivo verificar se os caracteres ciclo e virescência, de uma progênie da espécie "X", segregam de forma independente. Sabe-se que a virescência é controlada por um único locus (locus V), cuja condição normal é determinada por um gene dominante (V). Similarmente, o ciclo é determinado por um outro locus (locus P), cuja condição tardia é determinada por um alelo dominante.

a) Fixe os totais de colunas em 100%. Resposta Tabela 2.17.

b) Faça um gráfico para a tabela, verificando a distribuição do ciclo segundo a virescência (perfil coluna). Resposta Figura 2.10.

c) Os dois pares de genes são herdados independentemente ou há evidência de associação? justifique com apresentação de valores. Respsota: existe uma associação fraca entre as variáveis, pois para virescência normal observa-se que 77,11% apresentam ciclo tardio. Observa-se quase a mesma porcentagem (75,83%) para virescente e tardio. Portanto, não houve uma mudança razoável quando mudamos de nível de virescência, indicando associação fraca.

3. A tabela de contingência 2.18 refere-se ao número de pássaros de uma particular



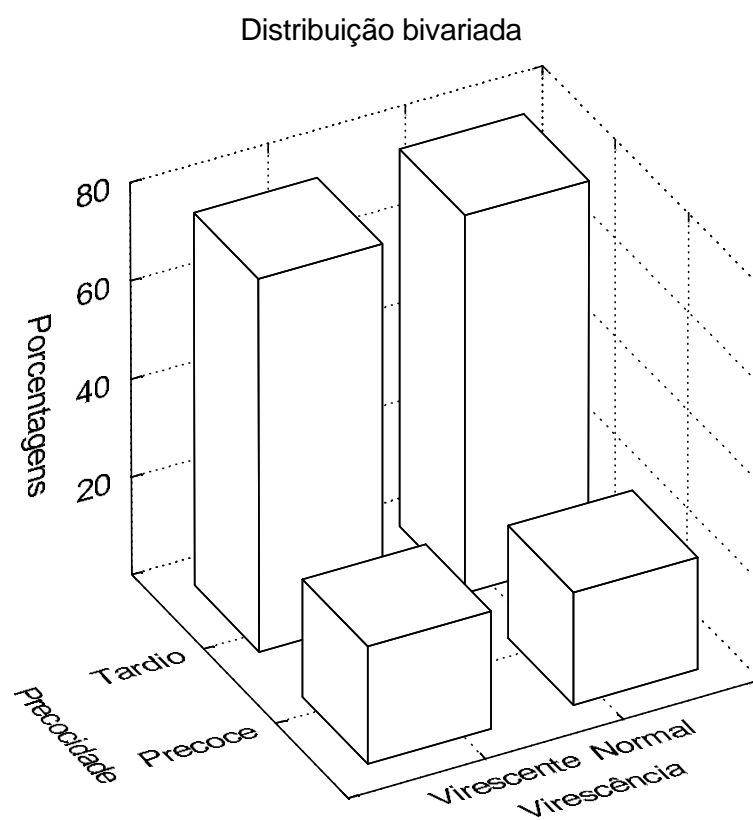


Figura 2.10: Distribuição da precocidade segundo a virescência

Tabela 2.18: Distribuição conjunta das variáveis local e estação

Estação do ano	Local da Floresta			Total
	Árvores	Arbusto	Chão	
Primavera	30 50,8%	20 33,9%	9 15,3%	59 100,0%
Outono	13 21,3%	22 36,1%	26 42,6%	61 100,0%
Total	43 35,8%	42 35,0%	35 29,2%	120 100,0%

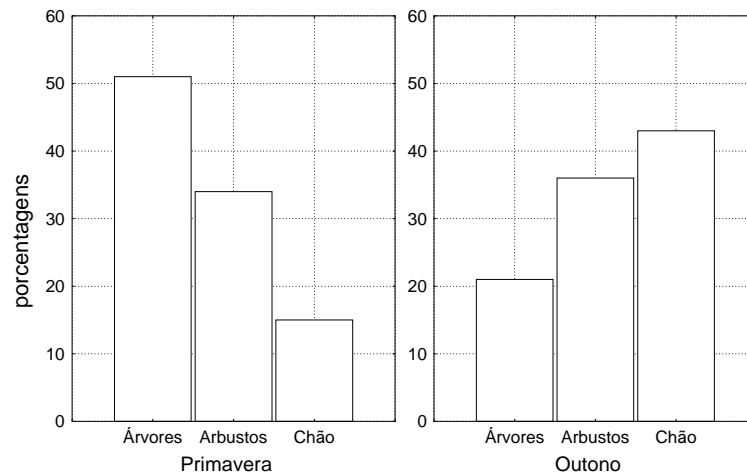


Figura 2.11: Associação entre local e estação

espécie, classificados de acordo com duas variáveis qualitativas, que são: 1º) local da floresta e 2º) estação do ano. Os pássaros foram observados alimentando-se de acordo com as duas variáveis. Um pesquisador levantou a hipótese (no início do trabalho), de que os pássaros alimentam-se nestes 3 locais da floresta nas mesmas proporções na primavera e no outono. Em outras palavras, o pesquisador formulou a hipótese de que não existe associação entre as duas variáveis. Você aceita ou rejeita a hipótese formulada pelo pesquisador? justifique. Resposta: rejeitamos a hipótese formulada. Na primavera 50,8% dos pássaros alimentam-se nas árvores, enquanto que no outono, o comportamento muda, isto é, a maior porcentagem (42,6%) alimentam-se no chão. Faça um gráfico de barras múltiplo para representar os dados da tabela acima e visualizar esta associação. Resposta na Figura 2.11.

### 2.1.3.5 Exercícios Propostos

1. Um economista está estudando os fatores que afetam a adoção de uma nova variedade de arroz altamente produtiva. Ele deseja saber se a adoção é afetada pela situação de posse da terra. Os resultados obtidos numa amostra de 187 agricultores estão organizados na

Tabela 2.19: Distribuição conjunta das variáveis

Estado de saúde	Vacinação			Total
	Não-vacinados	Uma dose	Duas doses	
Gripados	24	9	13	46
Não-gripados	289	100	565	954
Total	313	109	578	1000

tabela de contingência 2.11. O que o economista pode concluir?

2. Um estudo é realizado a fim de avaliar a eficiência de uma nova vacina antigripal, a qual foi administrada aos membros de uma pequena comunidade. A vacina foi administrada em duas doses, ao longo de duas semanas. Algumas pessoas tomaram as duas doses, outras tomaram apenas a 1ª dose e outras não tomaram qualquer dose. A Tabela 2.19 mostra os resultados obtidos para um total de 1000 habitantes dessa comunidade. Esses dados apresentam uma evidência suficiente para garantir que tal vacina foi bem sucedida, reduzindo o número de casos de gripe nessa comunidade? Em outras palavras, verifique se existe associação nesta tabela. Justifique com valores de proporções obtidas fixando-se os totais de linhas em 100%.

3. Utilizando os dados da Tabela 2.19, responda:

- faça um gráfico para a distribuição conjunta das porcentagens obtidas no exercício 2;
- calcule a proporção de não-vacinados entre os indivíduos não-gripados;
- calcule a proporção de não-gripados e que usaram duas doses de vacina.

4. De acordo com os dados da Tabela 2.10, podemos dizer que a ocorrência de aborto está associada com o local? Justifique.

5. Com os dados da Tabela 2.12, verifique se existe dependência (associação) entre as cultivares e a germinação de sementes. Justifique.

6. Para os resultados da Tabela 2.13 você concluiria que o peso está associado com o local? Justifique com os dados de percentagens.

7. Faça um gráfico para a Tabela 2.16 e observe a associação existente entre o tipo de grão e a resistência à ferrugem.

### 2.1.3.6 Coeficiente de Contingência de Pearson

Vamos agora pensar que estamos interessados em obter uma medida estatística que indique se existe ou não relação entre duas variáveis e qual a magnitude desta, isto é, a grandeza da associação. Aqui, desejamos fazer um estudo da associação ou dependência entre duas

variáveis categorizadas. No caso de estarmos trabalhando com variáveis quantitativas, estas podem ser transformadas em variáveis categorizadas. Exemplo de categorização de uma variável quantitativa: seja a variável rendimento de uma variedade transformada em três categorias:  $< 1500$  kg/ha (rendimento baixo),  $\geq 1500$  e  $\leq 3000$  (rendimento normal) e  $> 3000$  (rendimento alto).

Uma medida de associação que pode ser utilizada nesse caso é o *coeficiente de contingência de Pearson*, representado pela letra  $C^*$ . Podemos através deste coeficiente, por exemplo, verificar a grandeza da associação entre o local da floresta onde os pássaros se alimentam e a estação do ano, veja Tabela 2.18. Outro exemplo, verificar a magnitude da associação entre a adoção de tecnologia e a situação de posse da terra, veja a Tabela 2.11.

Antes de passarmos ao estudo detalhado do cálculo do coeficiente de contingência de Pearson, vamos, sem muitos detalhes, citar e indicar os cálculos de outras duas estatísticas para o estudo de associação.

A primeira delas é o coeficiente de associação de Yule, para tabelas  $2 \times 2$ , dado por:

$$Q = \frac{(f_{11}f_{22} - f_{12}f_{21})}{(f_{11}f_{22} + f_{12}f_{21})},$$

que varia de -1 a 1, onde,  $f_{11}$  representa a frequência de ocorrência da categoria 1 da variável  $A$  e a categoria 1 da variável  $B$ , e assim para os demais termos. Por exemplo, para a Tabela 2.17 temos:

$$Q = \frac{(3470.290 - 910.1030)}{(3470.290 + 910.1030)} = \frac{69000}{1943600} = 0,04.$$

Concluimos que existe uma associação muito fraca entre as duas variáveis, pois o valor de  $Q$  é próximo de zero.

Um outro coeficiente é o de Cramér, para tabelas  $l \times c$  ( $l$  = número de linhas e  $c$  = número de colunas da tabela), dado por:

$$V = \sqrt{\chi^2/n(\min(l-1, c-1))},$$

que varia de 0 (zero) a 1; onde  $\min(l-1, c-1)$  é o mínimo entre o número de linhas( $l$ ) menos 1 e o número de colunas( $c$ ) menos 1;  $n$  é o total geral de frequências da tabela; a estatística  $\chi^2$ , leia-se Qui-Quadrado, será mostrada em detalhes ainda nesta subseção. Para os dados da Tabela 2.18, temos  $\chi^2 = 15,05$ ,  $n = 120$ ,  $(\min(l-1, c-1) = \min(1; 2) = 1$ , portanto,  $V = 0,3541$ . De acordo com este coeficiente, a associação entre as duas variáveis é moderada para fraca.

Como foi dito inicialmente, a construção da distribuição conjunta das frequências, nos ajuda bastante na interpretação dos resultados. Na Tabela 2.18, temos a distribuição conjunta de frequências das variáveis local da floresta e estação do ano. Observe que para

facilitar a interpretação dos resultados, os totais de linhas foram fixadas em 100%. A interpretação já foi feita anteriormente e o resultado foi um indicativo de associação entre o local da floresta e a estação do ano. Vamos, agora, utilizar este exemplo para mostrar todos os passos no cálculo do coeficiente de contingência de Pearson.

### *Independência de Variáveis*

Quando se constroi uma distribuição conjunta de frequências, um dos principais objetivos, é procurar estabelecer a associação existente entre as variáveis, isto é, desejamos conhecer o grau de dependência entre as variáveis, pois conhecendo o grau de dependência entre elas, podemos prever o melhor resultado de uma variável sabendo-se o resultado da outra. Por exemplo, se desejamos saber o local da floresta onde os pássaros se alimentam, se nos tivermos informação sobre a estação do ano, vamos ter condições de estimar com maior precisão o local onde os pássaros se alimentam, pois existe uma dependência entre o local da floresta e a estação do ano.

A dependência não é no sentido de que uma *determina* a outra. Por exemplo, num estudo para verificar se existe associação entre a satisfação com o emprego e os salários percebidos, não significa que melhorando as condições de trabalho (satisfação), vai melhorar os salários.

Em primeiro lugar, observamos que, independentemente da estação do ano, 35,8% dos pássaros se alimentam nas árvores, 35% nos arbustos e 29,2% no chão.

Ora, se existe independência entre o local da floresta e a estação do ano, esperamos estas mesmas porcentagens para cada categoria da estação do ano. Comparando-se as frequências, podemos interpretar que existe dependência entre as variáveis.

Quando existe dependência entre as variáveis é interessante conhecer a magnitude dessa associação, ou seja, conhecer se a associação é fraca, moderada ou forte. Portanto, é importante termos uma medida de associação entre variáveis categorizadas.

### *Medida de Associação Entre Duas Variáveis Categorizadas*

Trataremos do coeficiente de contingência de Pearson, representado pela letra  $C^*$ , que descreve num único número a dependência entre duas variáveis.

Teoricamente este valor varia entre 0 (zero) e 1 (um),  $0 \leq C^* \leq 1$ , sendo nulo quando as variáveis são independentes. Quando existe uma associação perfeita entre as duas variáveis, o coeficiente de contingência de Pearson vale 1.

Para o cálculo do coeficiente é necessário, em primeiro lugar, calcularmos uma outra estatística, chamada de Qui-Quadrado, e representada pela letra grega  $\chi$  elevada a potência 2. Então, passamos de imediato ao cálculo do  $\chi^2$ , através do exemplo da Tabela 2.18.

Tabela 2.20: Frequências observadas e esperadas sob a hipótese de independência

Estação do ano	Local da Floresta			Total
	Árvores	Arbusto	Chão	
Primavera	30 21,14	20 20,65	9 17,21	59
Outono	13 21,86	22 21,35	26 17,79	61
Total	43	42	35	120

Tabela 2.21: Desvios entre as frequências observadas e esperadas, no caso de independência das variáveis

Estação do ano	Local da Floresta		
	Árvores	Arbusto	Chão
Primavera	8,86	-0,65	-8,21
Outono	-8,86	0,65	8,21

Na hipótese de independência, esperamos para local da floresta árvores e estação primavera,  $59 \times 0,358 = 21,122$  pássaros; para local da floresta árvores e estação outono, esperamos  $61 \times 0,358 = 21,838$  pássaros, e assim para todas as caselas restantes da tabela. Um modo prático de se encontrar as *frequências esperadas* sob a hipótese de independência, é dado por:

$$fe_{11} = \frac{n_{1.}n_{.1}}{n_{..}} = \frac{59 \times 43}{120} = 21,141.$$

A notação  $fe_{11}$  indica a frequência esperada para a estação primavera (linha 1) e local árvores (coluna 1). Se o leitor não está lembrado desses índices, veja novamente a Tabela 2.14.

Para a casela 21, isto é, estação outono (linha 2) e local árvores (coluna 1), temos:

$$fe_{21} = \frac{n_{2.}n_{.1}}{n_{..}} = \frac{61 \times 43}{120} = 21,858.$$

A diferença verificada na segunda casa decimal entre os dois procedimentos de cálculo é devido à aproximações. Fazemos a mesma operação para as demais caselas. Todas as frequências observadas e esperadas sob a hipótese de independência, estão demonstradas na Tabela 2.20. Encontre esses valores. Observando-se a Tabela 2.20 podemos verificar as discrepâncias existentes entre os valores observados e esperados caso as variáveis fossem independentes. Na Tabela 2.21, apresentamos os desvios entre os valores observados e os esperados. A estatística Qui-Quadrado ( $\chi^2$ ), que é uma medida de afastamento global da hipótese de independência, isto é, quanto maior o valor do  $\chi^2$ , maior será o grau de associação

entre as duas variáveis, é calculado através da seguinte expressão:

$$\chi^2 = \sum_{i=1}^{lc} \frac{(o_i - e_i)^2}{e_i} \quad (2.1)$$

onde,  $o_i$  é a frequência observada da  $i$ -ésima casela;  $e_i$  é a frequência esperada da  $i$ -ésima casela;  $l$  é o número de linhas e  $c$  é o número de colunas. Para o exemplo, temos:

$$\begin{aligned} \chi^2 &= \frac{8,86^2}{21,14} + \frac{-0,65^2}{20,65} + \frac{-8,21^2}{17,21} + \frac{-8,86^2}{21,86} + \frac{0,65^2}{21,35} + \frac{8,21^2}{17,79} \\ \chi^2 &= 3,7133 + 0,0205 + 3,9166 + 3,5910 + 0,0198 + 3,7889 \\ \chi^2 &= 15,0501. \end{aligned}$$

É fácil perceber que se  $\chi^2 = 0$ , as duas variáveis são independentes e que se  $\chi^2 > 0$  indica associação das variáveis. O valor de Qui-Quadrado não possui um limite superior, pois varia de 0 (zero) a  $+\infty$  (mais infinito). Na seção 8.3.3 iremos fazer o teste de qui-quadrado, que é um teste confirmatório, no sentido de poder afirmar se existe ou não associação significativa (estatisticamente comprovada) entre as variáveis.

Descritivamente, Karl Pearson propôs o chamado coeficiente de contingência, representado pela letra  $C$ , definido por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}, \quad (2.2)$$

onde  $n$  é o número total de observações. Para o exemplo em estudo temos:

$$C = \sqrt{\frac{15,0501}{15,0501 + 120}} = 0,3338,$$

indicando que existe associação, porém podemos dizer que a associação é fraca.

Quando existe uma associação perfeita, esse coeficiente não atinge o valor 1, por isso, foi sugerido uma correção, a qual consiste em calcular,

$$C^* = \frac{C}{\sqrt{(t-1)/t}}, \quad (2.3)$$

onde  $t$  é o mínimo entre o número de colunas e o número de linhas da tabela de contingência. Para o exemplo, onde  $t = 2$ , temos:

$$C^* = \frac{0,3338}{\sqrt{(2-1)/2}} = 0,4721.$$

Agora, temos uma associação moderada entre as variáveis. Devemos considerar as variáveis como moderadamente associadas.

### Exercício resolvido

1. Para os dados da Tabela 2.9, obter o coeficiente de contingência de Pearson e interpretar.

O valor de qui-quadrado é:

$$\chi^2 = 7,3877 + 3,4766 + 0,1000 + 0,0470 + 12,1278 + 5,7072 = 28,8464.$$

O valor do coeficiente de contingência é:

$$C = \sqrt{\frac{28,8464}{28,8464 + 50}} = 0,6049.$$

Portanto, com a correção temos:

$$C^* = \frac{0,6049}{\sqrt{1/2}} = 0,8554.$$

Conclusão: existe forte associação entre hábito e porte.

### 2.1.3.7 Exercícios Propostos

1. Com os dados da Tabela 2.10, calcule o coeficiente de contingência de Pearson e conclua.
2. Com os dados da Tabela 2.8, calcule o coeficiente de contingência de Pearson e conclua.
3. Com os dados da Tabela 2.12, calcule o coeficiente de contingência de Pearson e interprete.
4. Calcule o coeficiente de contingência de Pearson e o de Cramér para os dados da Tabela 2.13. Faça a interpretação.

### 2.1.3.8 Distribuições de Frequências de Variáveis Quantitativas: Diagrama de Pontos, Gráfico de Colunas e Histograma

Inicialmente faremos uma breve introdução sobre o estudo de distribuições de frequências para variáveis aleatórias quantitativas e, após, discutiremos as técnicas para este estudo.

Quando a variável em estudo é quantitativa, discreta ou contínua, as principais características a serem observadas numa distribuição de frequências são:

1. *valor típico ou representativo*, que como o próprio nome indica, corresponde a escolha de um único valor para representar todo o conjunto de valores;
2. *assimetria*, por exemplo, no estudo da distribuição da renda (em número de salários mínimos) das famílias brasileiras, a grande maioria das famílias apresentam baixo



rendimento familiar, enquanto que uma minoria apresenta altos rendimentos, isto provoca uma cauda longa à direita da distribuição, tornando-a assimétrica, veja Figura 2.38;

3. *dispersão*, é uma medida da concentração dos dados em torno do valor típico. É necessário ter um valor referência para poder compará-lo;
4. *valores discrepantes ou "outliers"*, são valores muito pouco prováveis de ocorrerem na distribuição, algumas vezes são valores que se distanciam demais dos outros. É importante realizar um estudo para saber a razão da ocorrência desses valores, pode-se citar 3 principais causas: 1) erro de transcrição de dados; 2) algum fato importante ocorreu durante o trabalho e 3) o valor é verdadeiro e deve ser considerado como tal;
5. *formação de subgrupos*, por exemplo, ao estudar-se a distribuição das alturas dos alunos, pode-se chegar a conclusão que existem dois grupos, formados de acordo com o sexo.

No caso do estudo de distribuições de freqüências de variáveis quantitativas podemos estabelecer duas situações, quais sejam:

- 1º) a variável é contínua, nesse caso, é necessário a criação de classes de ocorrências, pois não existem ou são poucos os valores que se repetem. Observação: quando temos poucos valores, isto é, o nosso conjunto de valores não é grande, em torno de 25, a distribuição pode ser representada por meio de um *diagrama de pontos*, ou seja, cada observação corresponde a um ponto na reta dos reais. Esta técnica será mostrada em seguida.
- 2º) a variável é discreta, nesse caso, temos duas situações: a) quando temos poucos resultados diferentes da variável, fazemos a contagem dos dados para cada valor da variável. Exemplos: 1) o número de plantas sadias de mandioca colhidas na área útil da parcela (área de 19,44 m<sup>2</sup>) varia no intervalo de 23 a 27 plantas (5 valores diferentes) e 2) número de grãos por vagem de soja. Nesse caso diz-se que não há perda de informação e as distribuições de freqüências são feitas de forma idêntica às variáveis qualitativas (categorizadas). Também pode-se fazer o diagrama de pontos; b) quando temos muitos valores diferentes da variável procedemos de forma idêntica ao de variável contínua, isto é, vamos criar faixas de ocorrência. Por exemplo, ciclo da cultura de 150 genótipos de feijão.

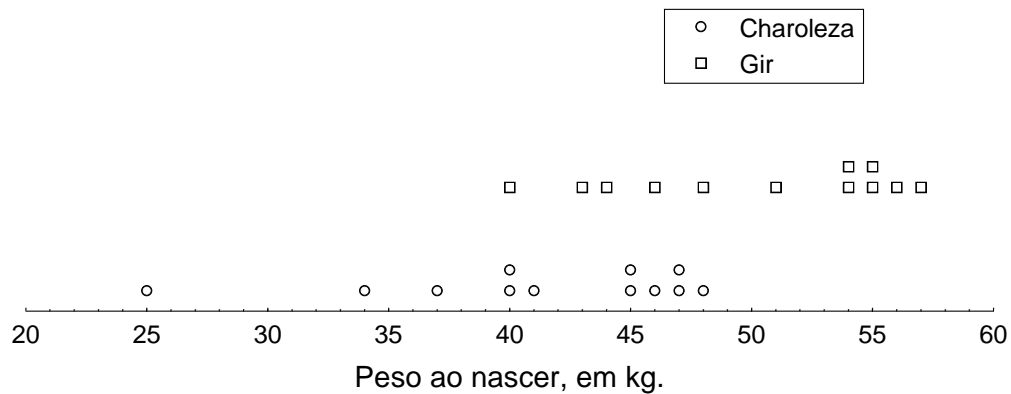


Figura 2.12: Distribuição de frequências do peso ao nascer de bezerros da raça Charoleza e Gir (kg)

Passamos, agora, ao estudo da construção das distribuições de frequências para variáveis aleatórias quantitativas. Os métodos abordados são: Diagrama de Pontos, Dados Agrupados em Classes e Ramo-e-Folhas.

### *Diagrama de Pontos*

Quando temos poucas observações de uma variável em estudo (até aproximadamente 25), a distribuição pode ser representada por um diagrama de pontos, isto é, cada observação corresponde a um ponto na reta dos números reais. É possível representar duas ou mais distribuições no mesmo diagrama, para isso basta identificar cada distribuição por um símbolo diferente (criar uma legenda). A Figura 2.12 ilustra esse diagrama com os pesos ao nascer de 24 bezerros machos das raças Charoleza e Gir, em kg. Os dados para as duas raças são:

Charoleza	47	45	37	41	46	47	34	25	40	45	48	40
Gir	40	43	44	46	48	51	54	55	56	57	55	54

Interpretação do diagrama de pontos:

- Observamos que os pesos ao nascer de bezerros da raça Charoleza são mais dispersos (estão mais espalhados) do que a raça Gir, principalmente devido ao valor 25. Portanto, os bezerros da raça Charoleza são mais heterogêneos quanto ao peso ao nascer.
- Ambas as distribuições são assimétricas à esquerda, pois temos poucos valores e mais espalhados no lado esquerdo das distribuições, isso gera uma cauda mais longa à esquerda.

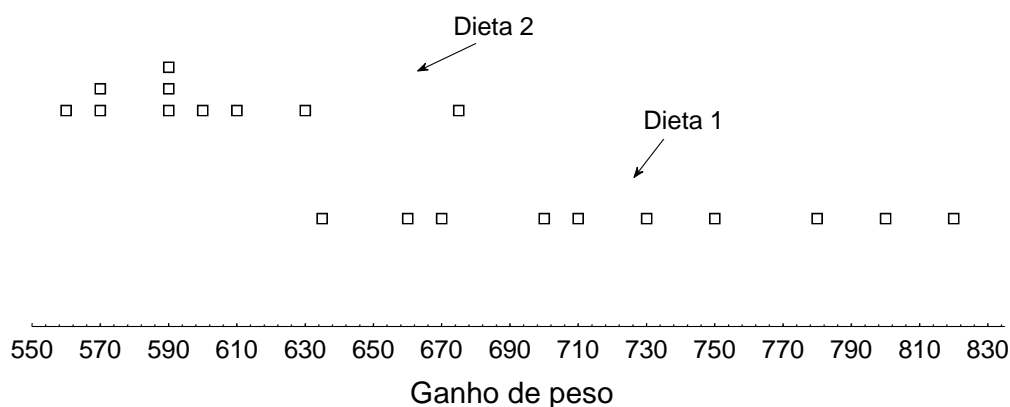


Figura 2.13: Diagrama de pontos para duas dietas de porcos

- Os valores representativos para as raças Charoleza e Gir são, 42 kg e 53 Kg, respectivamente. Estes valores dividem as distribuições aproximadamente ao meio, por exemplo, existem 6 pontos abaixo de 42 e 6 pontos acima de 42.
- O valor 25 kg, na raça Charoleza, está bem afastado do restante da distribuição, assim, podemos considerá-lo como um valor discrepante.
- Não observamos formação de subgrupos em nenhuma das distribuições (raças).

### Exercício resolvido

1. Para comparar o ganho médio de peso, GMP, em gramas por dia, de porcos submetidos a duas dietas diferentes,  $D_1$  e  $D_2$ , foi conduzido um experimento com 20 porcos. Dez porcos foram submetidos a cada uma das dietas. Os resultados para esse experimento foram:

Dieta 1		Dieta 2	
635	820	675	600
780	670	570	610
660	580	590	560
610	700	590	630
710	570	590	570

Construir um diagrama de pontos para verificar se existe evidência de diferenças entre os dois tratamentos (dietas). O diagrama de pontos é dado na Figura 2.13. Conclua para as duas dietas quanto à dispersão, assimetria, valores representativos e discrepantes.

A dieta 1 apresenta maior dispersão do que a dieta 2. A dieta 2 apresenta assimetria à direita (tem uma cauda mais longa para a direita), enquanto que a dieta 2 é aproximadamente simétrica. Os valores representativos são 715 e 590, para as dietas 1 e 2, respectivamente.

Tabela 2.22: Número de brotos por explante de abacaxi

Meio 1	47	35	23	21	23	26	18
	30	22	36	22	21	19	
Meio 2	13	11	15	24	20	20	19
	18	22	22	20	17	25	

O valor 675, na dieta 2, está mais afastado da distribuição dos demais valores. Os dados indicam que a dieta 1 é superior a dieta 2. Para fazermos uma afirmação sobre a diferença entre duas médias devemos realizar um teste de hipóteses, que será tratado no capítulo 8.

### Exercícios propostos

1. Os dados da Tabela 2.22 correspondem à variável número de brotos por explante de abacaxi avaliada em dois meios de cultura (dois tratamentos). Construir um diagrama de pontos e comparar os dois meios de cultura quanto à:

- a) dispersão;
- b) assimetria;
- c) valores representativos;
- d) valores discrepantes;
- e) formação de subgrupos.

## *Distribuição de Frequências de Variáveis Discretas Sem Perda de Informação*

A construção de distribuições de frequências de variáveis discretas, quando os diferentes valores observados da variável não são muitos, é feita de forma idêntica às variáveis qualitativas (categorizadas). Vamos fazer a contagem para cada valor observado da variável em estudo. Como exemplo, vamos estudar a distribuição do número de plantas sadias de mandioca colhidas na área útil da parcela. As frequências e porcentagens são dadas na Tabela 2.23. Na primeira coluna da tabela temos os valores observados da variável, quais sejam: 23, 24, 25, 26 e 27.

## *Representação Gráfica de Variáveis Discretas Sem Perda de Informação*

Nesse caso, podemos citar os gráficos de ordenadas e os gráficos de colunas. A seguir apresentamos os dois tipos de gráficos, Figuras 2.14 e 2.15, para o exemplo do número de plantas sadias de mandioca (Tabela 2.23)

Tabela 2.23: Distribuição de frequências e proporções (em %) do número de plantas sadias de mandioca, Chapecó, SC, 1984

Número de plantas	Frequência	Proporção	Porcentagem	Porcentagem acumulada
23	1	0,0333	3,33	3,33
24	3	0,1000	10,00	13,33
25	4	0,1333	13,33	26,67
26	8	0,2667	26,67	46,67
27	14	0,4667	46,67	100,00
TOTAL	30	1,0000	100,00	

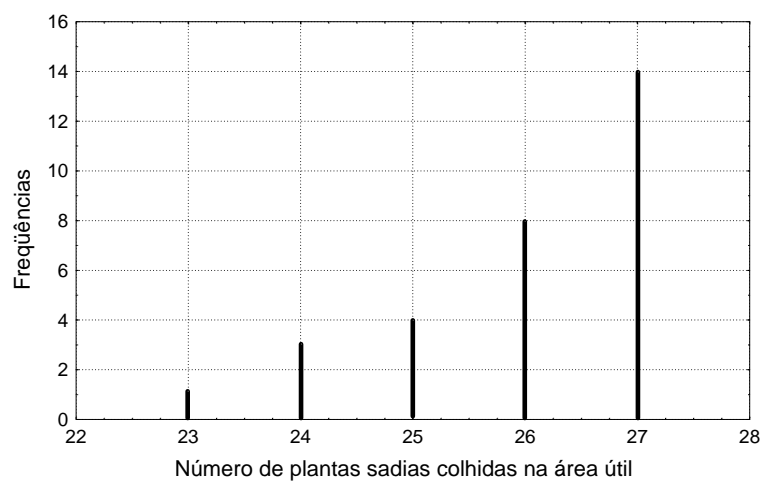


Figura 2.14: Diagrama de ordenadas do número de plantas sadias de mandioca

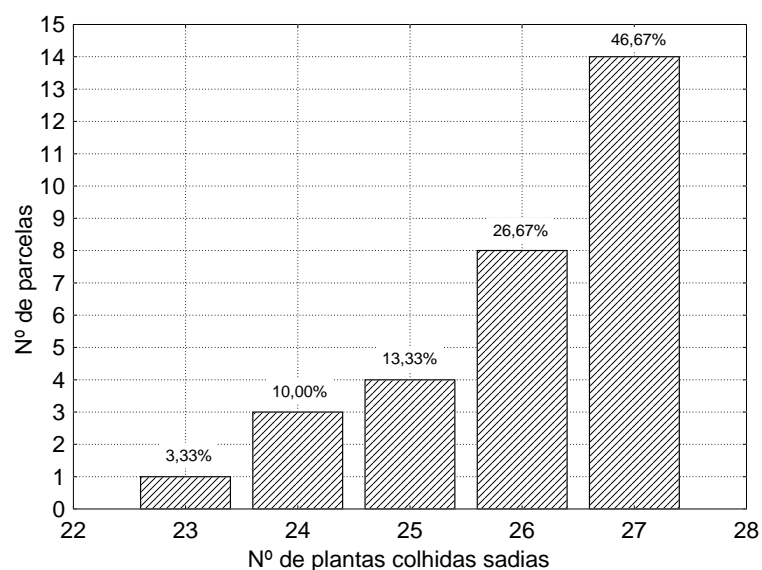


Figura 2.15: Gráfico de colunas do número de plantas sadias de mandioca, Chapecó, SC, 1984

Observamos que 47% das unidades experimentais (parcelas) deram como resultado 27 plantas sadias (total de plantas/unidade); somente 3% das unidades apresentaram como resultado 23 plantas sadias. A distribuição é assimétrica à esquerda (cauda longa para o lado esquerdo da distribuição). Um valor representativo seria o 27. Não observamos valores discrepantes. Não faremos uma conclusão para a dispersão pois não temos duas ou mais distribuições para comparação.

### Exercício proposto

1. Deseja-se estudar o comportamento da variável número de ovos inviáveis de *Biomphalaria taenagophila* (caramujo) em ambiente poluído. Para isso tomou-se uma amostra de 23 caramujos, obtendo-se os seguintes resultados:

9	11	10	0	4	4	5	12	4	1	2	8
4	7	1	11	10	3	14	3	2	28	4	

Faça uma representação gráfica para os dados (gráfico de ordenadas ou de colunas). Faça a interpretação do gráfico.

### *Dados Agrupados em Classes*

Quando temos muitas observações de uma variável quantitativa em estudo (acima de 25), é recomendável a formação de intervalos de valores, isto é, construir classes de ocorrências. Por exemplo, uma classe pode ser de 8 a 16 cm de comprimento de camarão. Quanto

Tabela 2.24: Rendimentos médios, em kg/ha, de 32 híbridos de milho, região Oeste, 1987/88

3973	4660	4770	4980	5117	5403	6166
4500	4680	4778	4993	5166	5513	6388
4550	4685	4849	5056	5172	5823	
4552	4760	4960	5063	5202	5889	
4614	4769	4975	5110	5230	6047	

ao número de classes que deve ser usado, existe alguma recomendação. Este número não poderá ser muito grande (maior que 15) e nem muito pequeno (menor que 5). Mas também não há um rigor muito grande quanto ao número de classes a ser usado. Pode-se deixar como compromisso do pesquisador decidir sobre o número de classes a ser usado. Ninguém melhor do que a pessoa que conhece o fenômeno em estudo para decidir sobre a melhor representação da distribuição. Sempre que for possível, recomenda-se utilizar classes com a mesma amplitude.

Os dados da Tabela 2.24 referem-se aos rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a região Oeste Catarinense. Vamos considerar estes dados para ilustrar a construção da distribuição de freqüências em classes. As classes podem ser definidas de acordo com resultados lógicos da variável, por exemplo, definir 5 classes com amplitudes de 500 kg. Outra forma é decidirmos pelo número de classes a ser utilizado e, a seguir, fazer a divisão da amplitude total dos dados pelo número de classes, sendo que, o resultado encontrado será a amplitude de cada classe. Por exemplo, amplitude total dos dados da Tabela 2.24 é  $\Delta = 6388 - 3973 = 2415$  kg/ha. Se desejamos ter 5 classes, isso implica que a amplitude de cada classe será igual a  $\Delta_i = 2415/5 = 483$  kg/ha. Veja a primeira coluna da Tabela 2.25. Nesta tabela temos a distribuição de freqüências completa da variável em estudo.

A freqüência absoluta ou simplesmente freqüência de classe (segunda coluna da tabela), representada por  $n_i$ , é o número indicativo da quantidade de valores, indivíduos, itens, elementos, etc. pertencentes a essa classe. A freqüência absoluta acumulada de uma classe é a soma da freqüência dessa classe com as freqüências das classes anteriores (coluna 3). A freqüência relativa ou proporção, representada por  $f_i$ , é definida pelo quociente da freqüência absoluta da classe e o número total de observações, ou seja,  $f_i = n_i/n$ , onde  $n = 32$  (coluna 4). Também temos a freqüência relativa acumulada, calculada da mesma forma que a freqüência absoluta acumulada (coluna 5). A porcentagem (coluna 6) é a forma que a maioria das pessoas entendem, é mais explícito, tem maior êxito, e é simplesmente a multiplicação das proporções por 100, isto é,  $100 \times f_i$ . Da mesma forma podemos ter as freqüências percentuais

Tabela 2.25: Distribuição de freqüências de 32 híbridos de milho recomendados para o Oeste Catarinense, 1987/88

Rendimento médio	Freqüência absoluta	Freqüência acumulada	Freqüência relativa	Freq. rel. acumulada	Porcentagem	Porcentagem acumulada
3973 ⊢ 4456	1	1	0,0313	0,0313	3,13	3,13
4456 ⊢ 4939	12	13	0,3750	0,4063	37,50	40,63
4939 ⊢ 5422	13	26	0,4063	0,8126	40,63	81,26
5422 ⊢ 5905	3	29	0,0937	0,9063	9,37	90,63
5905 ⊢ 6388	3	32	0,0937	1,0000	9,37	100,00
TOTAL	32		1,0000		100,00	

acumuladas (coluna 7).

A representação gráfica dessas distribuições recebe um nome especial, *histogramas*, e é mostrado na Figura 2.16. É um gráfico de colunas justapostas em que a altura de cada coluna é proporcional a freqüência absoluta ou porcentagem de ocorrências da classe, de modo que a soma das áreas dos retângulos seja igual a 1 ou 100%. A altura de cada retângulo é denominada de densidade de freqüência e é dada por:

$$d_i = \frac{n_i}{n \times \Delta_i}.$$

No exemplo da Figura 2.16 temos  $n \times \Delta_i = 32 \times 483 = 15.456$ , para todas as classes, pois estas apresentam a mesma amplitude. Para a classe 2,  $d_2 = 12/(15.456) = 0,000776$ .

As amplitudes de classes nem sempre são iguais para todas as classes da mesma distribuição de freqüência. As vezes isto ocorre pela própria natureza da pesquisa. Por exemplo, deseja-se discriminar melhor os baixos salários; ou as idades menores da primeira relação sexual de jovens. Neste caso, é necessário tomar alguns cuidados com a análise e construção do histograma, pois os  $\Delta_i$  assumem valores diferentes. Deve-se usar os valores da densidade de freqüência no eixo das ordenadas, também conhecido como eixo dos  $Y's$ . Não é possível fazer-se uma interpretação de um valor específico da densidade de freqüência, pois como o nome indica, só faz sentido verificar-se o comportamento dentro de um intervalo ou no geral.

Uma representação gráfica alternativa ao histograma é o *polígono de freqüências*, e o seu uso é particularmente importante quando deseja-se comparar duas ou mais distribuições, pois é possível representar-se diversas distribuições no mesmo gráfico, veja Figura 2.19. Para construir um polígono de freqüências basta colocar num gráfico os pares  $(n_i; s_i)$ , e uní-los com uma linha, onde  $s_i$  corresponde ao ponto médio da  $i$ -ésima classe. Por exemplo, o



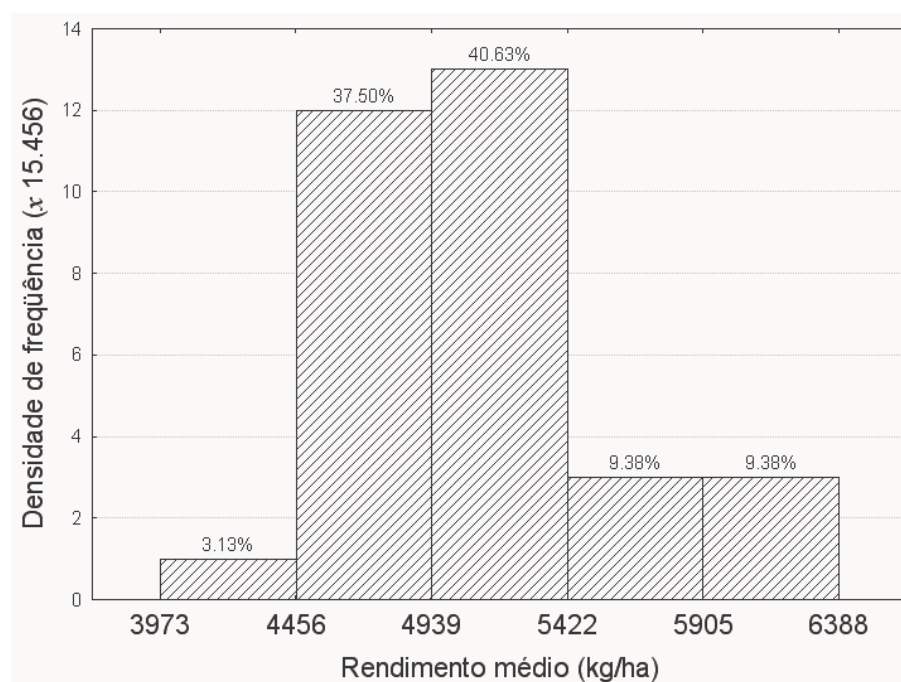


Figura 2.16: Histograma dos rendimentos médios, em kg/ha, de 32 híbridos recomendados para o Oeste Catarinense, 1987/88

ponto médio da primeira classe,  $s_1$ , é igual a  $(4456+3973)/2=4214,5$ , portanto, o par  $(n_1; s_1)$  é dado por  $(1; 4214, 5)$ . Usar o mesmo procedimento para os 4 pares restantes. Esse gráfico para a variável rendimento de grãos de híbridos de milho é mostrado na Figura 2.17.

**Interpretação.** A interpretação pode ser feita através da tabela de distribuição de freqüências, histograma ou do polígono de freqüências. A maioria dos híbridos (78%), apresentam uma produção entre 4456 e 5422 kg/ha. Um valor representativo para esses dados seria 5000 kg/ha. Temos aproximadamente 50% dos valores abaixo de 5000 kg/ha, conseqüentemente, 50% acima deste valor. Existe uma tendência dos dados se concentrarem próximo ao valor representativo e, a medida que se afastam do valor típico, a concentração diminui. A distribuição dos dados apresenta uma pequena assimetria à direita, pois a cauda à direita é um pouco mais longa. Desconsiderando o valor 3973 kg/ha, a amplitude de variação dos valores está dentro do esperado, vai de 4500 a 6388 kg/ha, isto é, a dispersão dos valores não é muito grande. Seria bom verificar por que um híbrido produziu razoavelmente menos que os demais. Não se observa formação de subgrupos.

#### Comentários:

- 1 Limites indeterminados ou classes abertas. É até comum aparecer tabelas de distribuições de freqüências em que a 1ª classe e/ou a última, apresentam o limite inferior e/ou

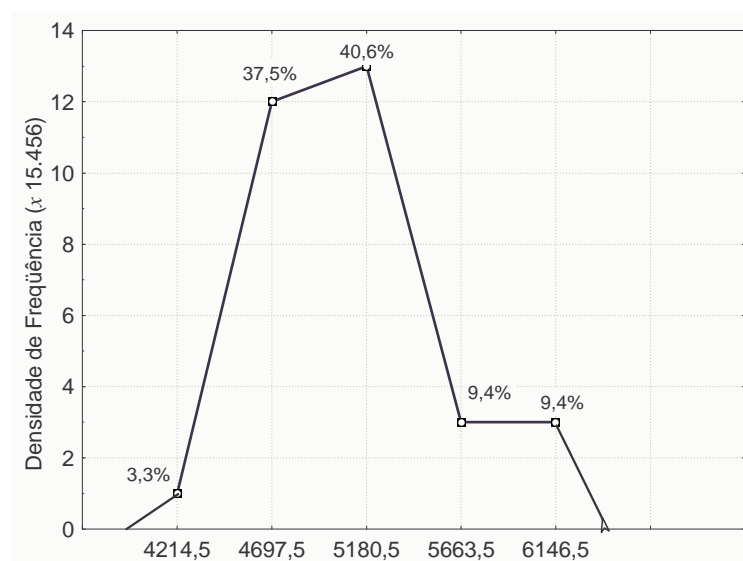


Figura 2.17: Polígono de frequências dos rendimentos médios, em kg/ha, de 32 híbridos recomendados para o Oeste Catarinense, 1987/88

Tabela 2.26: Altura de brotos de explantes de abacaxi

1,00	1,18	1,21	1,27	1,34	1,37	1,43	1,47	1,52	1,68
1,01	1,19	1,25	1,30	1,35	1,37	1,43	1,47	1,57	1,73
1,08	1,19	1,26	1,31	1,36	1,39	1,44	1,49	1,61	1,77
1,11	1,20	1,27	1,34	1,36	1,41	1,46	1,50	1,62	

superior indefinidos. Deve-se, sempre que possível, evitar esse tipo de limites, pois dificulta no trabalho descritivo dos dados, principalmente no que diz respeito ao cálculo de estatísticas.

- 2 O número de classes estabelecido para a distribuição de frequências é bastante subjetivo, pois trata-se de uma análise exploratória de dados. Então, o número de classes deve ser o suficiente para nos dar uma boa idéia do fenômeno, deve por em evidência a regularidade do fenômeno. Assim, vale a pena salientar, que a decisão sobre o número de classes cabe ao pesquisador.

### Exercícios resolvidos

1. Os dados da Tabela 2.26 foram obtidos de um experimento desenvolvido para avaliar o comportamento "In Vitro" de abacaxi (*Ananas comosus*) cv. Primavera e refere-se a variável altura dos brotos de explantes, em centímetros (dados ordenados crescentemente).
  - a) Preencha a tabela de distribuição de frequências 2.27.

Tabela 2.27: Distribuição de frequências da variável altura de brotos de explantes de abacaxi.

Altura	Frequência absoluta	Frequência relativa	Porcentagem
1,00 – 1,13			
1,13 – 1,26			
1,26 – 1,39			
1,39 – 1,52			
1,52 – 1,65			
1,65 – 1,78			
Total	39	1,0000	100,00

b) Construa um histograma.

c) Faça algumas interpretações relevantes.

d) Pode-se dizer que esta amostra é oriunda de uma população com distribuição normal? Justifique. Observação: leia mais adiante o tópico *O modelo normal*.

As frequências absolutas, relativas e porcentagens para as 6 classes são dadas por: 4, 6, 12, 10, 4, 3, 0,1026, 0,1538, 0,3077, 0,2564, 0,1026, 0,0769, 10,26, 15,38, 30,77, 25,64, 10,26 e 7,69, respectivamente. O histograma é dado na Figura 2.18. A distribuição é aproximadamente simétrica. Um valor representativo do conjunto de dados é 1,36. Não observa-se valor discrepante. Pode-se dizer que a variável altura de explantes de abacaxi segue aproximadamente uma distribuição normal.

2. Os dados da Tabela 2.28 são relativos a resposta à variação de luz no crescimento das plantas, medido através da área foliar específica após 60 dias, da espécie *Cecropia glazioui*, em amostras situadas em borda de mata e mata fechada, com intensidade de luz média de 10,6% e 0,87%, respectivamente. A espécie *C. glazioui* é uma planta pioneira, heliófita, perenifolia, seletiva higrófila, ocorrendo preferencialmente em capoeiras e capoeirões de derribadas recentes.

Compare as duas amostras através do polígono de frequências múltiplo. Por polígono de frequências múltiplo, entende-se que no mesmo gráfico (no mesmo plano cartesiano), vamos representar as duas amostras (borda da mata e mata fechada), por duas linhas poligonais fechadas. O polígono de frequências múltiplo é dado na Figura 2.19. O valor 1,47 é obtido através do cálculo:  $35 \times 0,04191$ , onde 0,04191 é a amplitude de classe.

Observa-se que o crescimento foi maior na mata fechada do que na borda da mata, com valores representativos de 0,70 e 0,40, respectivamente. A dispersão dos dados é um pouco maior na mata fechada. A distribuição é mais simétrica na amostra da borda da mata, com os dados concentrando-se mais na parte central.

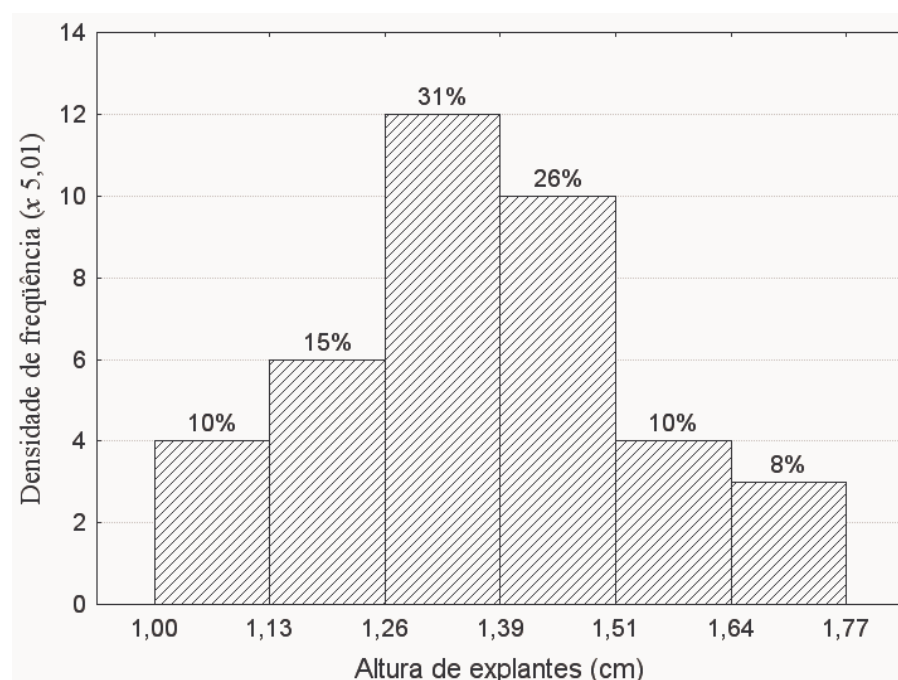


Figura 2.18: Histograma da altura de explantes de abacaxi

### 2.1.3.9 Exercícios Propostos

1. Os dados da Tabela 2.29 referem-se a biometria total, em mm, do *Macrobrachium potiuna* (Müller, 1880) da família Palaemonidae. Obs. os dados encontram-se ordenados.

a) Construa a tabela de distribuição de frequências com 6 classes para os dados.

b) Construa o histograma.

c) Indique um valor representativo para os dados; comente sobre a assimetria; esta amostra é oriunda de uma população com distribuição aproximadamente normal? justifique.

2. Dispõe-se de uma relação de 36 produções em kg/ha de milho do município de Chapecó e uma relação de 36 produções, também em kg/ha, do município de Campos Novos. A tabela de distribuição de frequências é dada em 2.30. Construa o polígono de frequências múltiplo. Com base no polígono de frequência múltiplo, discuta e compare as duas distribuições quanto as principais características estatísticas.

## *O Modelo Normal*

Uma distribuição de frequências muito importante em estatística, é aquela onde os dados se distribuem simetricamente em torno de um valor central, de tal forma que os dados estão em maior quantidade na região próxima desse valor e, a medida que se afastam desse valor, a concentração diminui. Um grande número de métodos de análise estatística se

Tabela 2.28: Área foliar específica de *Cecropia glazioui*

Borda da mata					Mata fechada				
0,2145	0,3458	0,3796	0,4125	0,4657	0,5521	0,6780	0,7126	0,7894	0,8564
0,2540	0,3482	0,3815	0,4142	0,4670	0,5841	0,6842	0,7154	0,7912	0,8654
0,2592	0,3487	0,3874	0,4182	0,4823	0,6284	0,6898	0,7179	0,8023	0,8665
0,2891	0,3490	0,3924	0,4326	0,4862	0,6357	0,6945	0,7256	0,8046	0,9214
0,2971	0,3547	0,3931	0,4358	0,4921	0,6489	0,6950	0,7321	0,8451	0,9421
0,3013	0,3574	0,3971	0,4573	0,5010	0,6570	0,6970	0,7783	0,8467	0,9573
0,3279	0,3648	0,4015	0,4582	0,5231	0,6704	0,7125	0,7884	0,8468	0,9689

Tabela 2.29: Biometria total de *Macrobrachium potiuna*

25,60	27,75	29,95	32,20	33,90	34,75	35,20	36,00	37,10	39,20	41,75	44,05
25,90	28,30	31,25	32,20	33,95	34,80	35,55	36,70	38,90	39,55	42,80	45,20
25,90	29,05	31,70	33,75	34,75	35,10	35,65	37,05	39,10	40,45	43,95	46,74

Tabela 2.30: Distribuição de frequências do rendimento de milho para Chapecó e Campos Novos

Rendimento Chapecó	Freq. absoluta	Rendimento Campos Novos	Freq. absoluta
4200 ┤ 4552	1	6613 ┤ 7095	1
4552 ┤ 4904	3	7095 ┤ 7577	2
4904 ┤ 5256	5	7577 ┤ 8059	4
5256 ┤ 5608	8	8059 ┤ 8541	13
5608 ┤ 5960	8	8541 ┤ 9023	7
5960 ┤ 6312	6	9023 ┤ 9505	7
6312 ┤ 6664	5	9505 ┤ 9987	2
Total	36		36

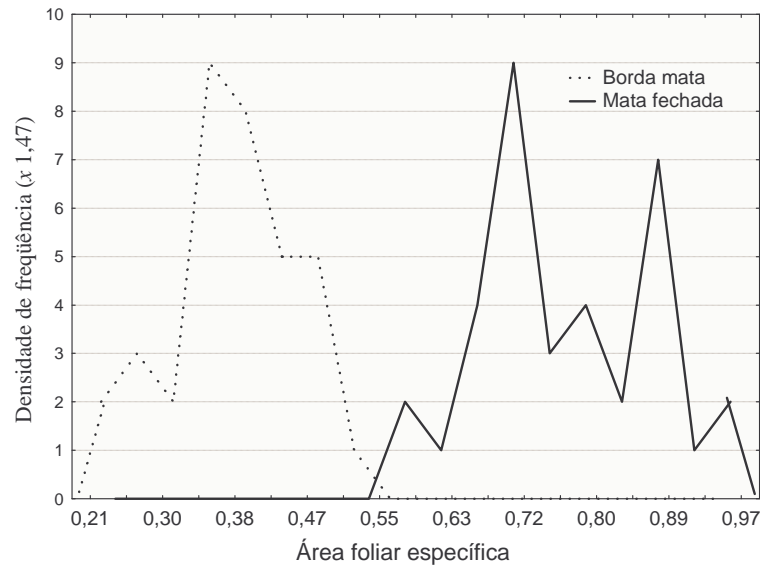


Figura 2.19: Polígono de frequências múltiplo da área foliar específica de *C. glazioui*

fundamentam nessa distribuição, denominada *Distribuição Normal ou Gaussiana*, devido a Karl Gauss (1777-1855). A representação gráfica deste modelo, que será estudado com detalhes no capítulo 5, é dada na Figura 2.20 .

A distribuição das alturas de explantes de abacaxi, em cm, está representada na Figura 2.21. A princípio, dá para dizer que os dados seguem um modelo normal, pois a distribuição é aproximadamente simétrica em torno do valor médio, com maior concentração em torno deste e pouca concentração nas extremidades.

### 2.1.3.10 Ramo-e-Folhas

Tanto os histogramas, como os demais tipos de gráficos vistos anteriormente, dão uma boa idéia da forma da distribuição da variável em estudo, isto é, do comportamento dos dados. Um outro modo de representação de um conjunto de valores, com o objetivo de se ter uma idéia geral dos dados, é o ramo-e-folhas. Uma vantagem do ramo-e-folhas sobre o histograma é que no primeiro não há perda de informação dos dados, pois trabalha-se com todos os dados originais. É possível, também, através do ramo-e-folhas obter-se algumas outras informações mais gerais sobre os dados, o que será mostrado através de um exemplo. O ramo-e-folhas é útil quando o conjunto de dados não é muito numeroso, pois, caso contrário, a interpretação torna-se difícil.

**Exemplo (com discussão sobre a construção do ramo-e-folhas).** Nas Figuras

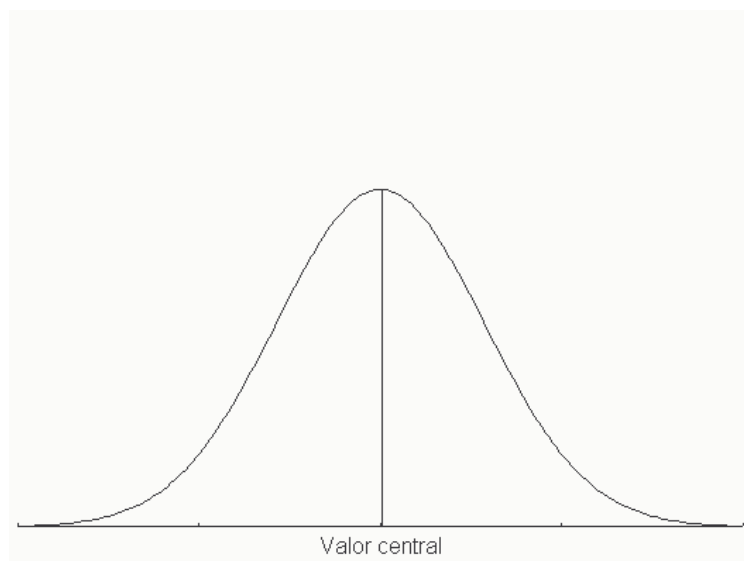


Figura 2.20: A representação gráfica do Modelo Normal ou Gaussiano

2.22, 2.23 e 2.24 temos os ramos-e-folhas da variável rendimento de grãos de híbridos de milho. A idéia básica para construir um ramo-e-folhas é dividir cada observação em duas partes; a primeira parte, que será denominada de ramo, é colocada à esquerda de uma linha vertical; a segunda parte, que serão denominadas de folhas, é colocada à direita dessa linha vertical. O critério de divisão da observação é decisão do pesquisador. Ele vai procurar um ramo-e-folhas que represente bem o fenômeno biológico/agronômico. Assim, para os rendimentos de híbridos de milho, considerando, por exemplo, o valor 3.973, o 3 (parte do milhar) é o ramo e 973 (centena) é a folha, fazemos esta mesma divisão para todos os demais valores.

Na Figura 2.22 temos apenas 4 ramos (fazendo analogia com o número de classes do histograma que deverá ser entre 5 e 15), ou seja, reduziu-se muito os dados. Para aumentar o número de ramos podemos subdividí-los em duas partes, garantindo que os ramos sejam equiprováveis, isto é, tenham a mesma chance de receber uma observação, assim, os ramos com o símbolo \* recebem valores de 000 a 499 e os ramos com o símbolo o recebem valores de 500 a 999. Esse ramo-e-folhas é mostrado na Figura 2.23.

Podemos ainda subdividir os ramos em 5 partes, da seguinte forma: os ramos com o símbolo ★ recebem valores na faixa 000 a 199; ramos com o símbolo ● recebem valores na faixa 200 a 399; ramos com o símbolo ▷ recebem valores na faixa 400 a 599; ramos com o símbolo o recebem valores na faixa 600 a 799 e ramos com o símbolo ▲ recebem valores na faixa 800 a 999. Este ramo-e-folhas é dado na Figura 2.24.

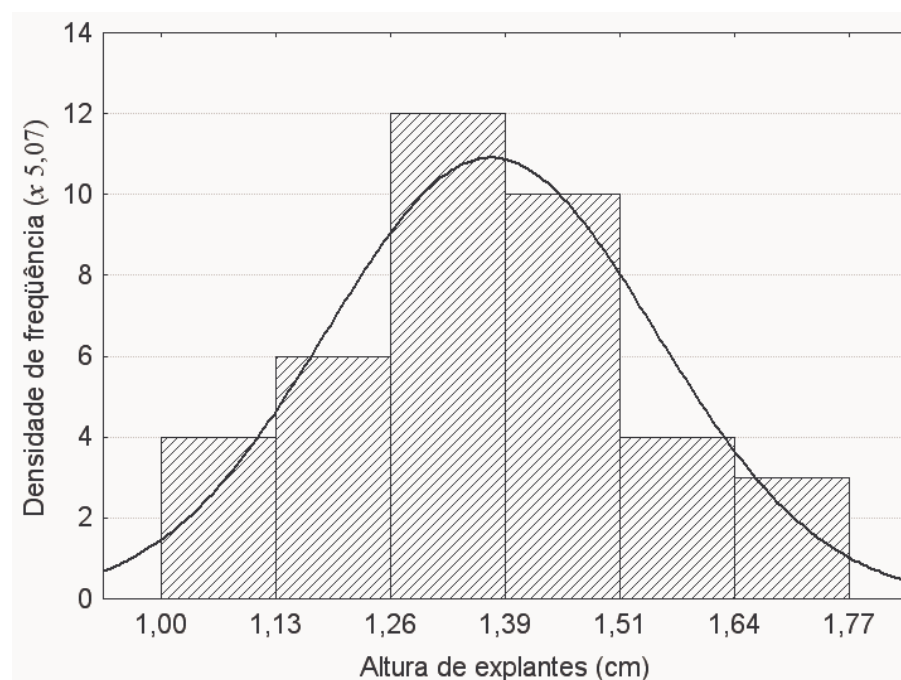


Figura 2.21: Comparação da distribuição das alturas de explantes de abacaxi e o modelo normal

**Interpretação:** No primeiro ramo-e-folhas não observa-se que o valor 3973 está distante da distribuição dos demais valores, isso indica que o ramo-e-folhas com apenas quatro ramos não é adequado. O segundo ramo-e-folhas nos indica claramente a existência de um híbrido (3.973) com produção abaixo dos demais. Observamos que existe uma queda brusca no lado esquerdo e uma queda mais suave no lado direito, portanto, a distribuição é assimétrica à direita. Esta distribuição parece não seguir o modelo normal. Os valores estão bastante concentrados na faixa entre 4500 e 5400 kg/ha. Um valor típico para representar este conjunto de dados é 5000 kg/ha. No ramo-e-folhas da Figura 2.24 observa-se 5 híbridos com altos rendimentos e cujos valores se distanciam dos demais. Se o interesse é encontrar híbridos com alta produtividade esses cinco são os de maior interesse no estudo. O híbrido com baixa produção tornou-se mais evidente, o que também é desejável. Observação: A escolha do número de ramos é equivalente a escolha do número de classes no histograma.

### Exercícios resolvidos

1. Os dados da Tabela 2.31 referem-se a uma avaliação da formação de calos, mais especificamente, da altura de calos em cm, após um período de 30 dias "in vitro" para os explantes de procedência do cerrado da espécie *Mandevilla velutina*. Obs. é importante que os calos não se desenvolvam, pois quanto menor eles forem, maior será o desenvolvimento





3 <sup>^</sup>	973
4 <sup>*</sup>	
4 <sup>•</sup>	
4 <sup>▷</sup>	500 550 552
4 <sup>°</sup>	614 660 680 685 760 769 770 778
4 <sup>^</sup>	849 960 975 980 993
5 <sup>*</sup>	056 063 110 117 166 172
5 <sup>•</sup>	202 230
5 <sup>▷</sup>	403 513
5 <sup>°</sup>	
5 <sup>^</sup>	823 889
6 <sup>*</sup>	047 166
6 <sup>•</sup>	388
6 <sup>▷</sup>	

Unidade 1,0

3<sup>^</sup> | 973 = 3973 kg/ha

Figura 2.24: Ramo-e-folhas do rendimento médio, em kg/ha, de 32 híbridos de milho, com 14 ramos

Tabela 2.32: Dados de peso da carne, em gramas, de mexilhões do manguezal

9,49	9,54	9,53	11,13	11,76	12,69	12,92
12,92	13,04	14,05	14,88	16,22	16,38	16,92
17,64	18,17	18,53	19,17	19,65	19,80	19,81
20,01	21,60	21,99	22,74	23,71	25,14	25,34
25,67	26,79	27,20	30,20	30,85	33,97	36,35

vão os valores na faixa de 15 até 19,99. E assim para os demais ramos. Esta construção do ramo-e-folhas foi necessária pois, se não tivéssemos feito esta quebra no ramo, teríamos muitos ramos, o que tornaria muito difícil a interpretação.

### 2.1.3.11 Exercícios Propostos

1. Os dados da Tabela 2.33 correspondem a variável altura de plantas, medida em cm, para dois tratamentos. O tratamento 1 é o controle (testemunha), isto é, não foi feita a inoculação do rizóbio nas plantas. O tratamento 2 corresponde à aplicação do rizóbio nas plantas. A fim de comparar os dois tratamentos faça um ramo-e-folhas para cada um dos tratamentos. Quais os valores para representar os dois tratamentos? Qual dos dois

0°	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	
0	2 3 3 3 3 3	
0▷	4 4 4 4 4 5 5 5 5	
0★	6 6 6 6 6	0° 0 = 0,0
0●	8 8	unidade: 0,10

Figura 2.25: Ramo-e-folhas da altura de calos

9°	49	
9	54 53	
1°	101 176 269 292 292 304 405 488	
1	622 638 692 764 814 853 914 965 980 981	
2°	001 160 199 274 371	
2	514 534 567 679 720	9° 49 = 9,49
3°	020 085 397	1° 101=11,01
3	635	unidade: 0,01

Figura 2.26: Ramo-e-folhas do peso da carne de mexilhões do manguezal

tratamentos apresenta maior dispersão? Comente sobre a assimetria das distribuições. Têm valores discrepantes? Pode-se dizer que estas duas amostras são oriundas de duas populações com distribuição pelo menos aproximadamente normal? justifique?

2. Faça um ramos-e-folhas da variável duração do primeiro estágio ninfal de *Triatoma klugi*, em dias, alimentadas em galo, cujos valores são:

21	21	21	22	22	22	22
23	23	23	23	23	25	26
28	28	28	30	30	33	35
35	36	39	39	40	40	42
42	45	46	46	48	50	59

- Qual o valor representativo do conjunto de valores?
- Comente sobre a assimetria da distribuição.

Tabela 2.33: Altura de plantas para dois tratamentos

Controle	25	29	29	30	31	31	32	32	33	33
	35	36	36	37	37	38	38	40	41	43
Com rizóbio	34	36	39	39	40	41	41	41	42	44
	44	45	45	46	46	47	47	49	49	51

c) Pode-se considerar algum(ns) valor(es) como sendo discrepante(s)?

## 2.2 Algumas Medidas e Modelo para Variáveis Quantitativas

### 2.2.1 Medidas de Tendência Central

#### 2.2.1.1 Média, Mediana e Moda

Procuramos até aqui interpretar um conjunto de dados através do estudo de distribuição de frequências, diagrama de pontos e do ramo-e-folhas, o que já foi um grande avanço no sentido de conhecer o comportamento da(s) variável(is) em estudo. Porém, um conjunto de dados pode reduzir-se a uma ou apenas algumas medidas numéricas que representam todo o conjunto original dos dados. Estas medidas são de muito mais fácil compreensão do que os dados originais, esta é uma grande virtude da estatística, isto é, reduzir um conjunto de dados em apenas algumas medidas facilmente compreensíveis. Por exemplo, para conhecer o rendimento típico de híbridos de milho, podemos calcular a média ou a mediana dos resultados da variável. É importante chamar a atenção desde já, que sempre que for apresentada uma medida de tendência central é necessário apresentar também uma medida de variabilidade, dispersão, que serão tratadas na próxima seção.

O objetivo dessa seção é apresentar as principais medidas de tendência central, as quais são assim chamadas devido ao fato dos dados naturalmente tenderem a se concentrar em torno desses valores centrais. As três medidas de tendência central mais utilizadas para resumir um conjunto de dados são:

- A média aritmética
- A mediana
- A moda

Estas medidas aplicam-se para dados isolados, como também para dados organizados numa distribuição de frequências em classes, sendo assim, será mostrado o procedimento de

cálculo dessas medidas para os dois casos, quais sejam: 1) dados isolados e 2) agrupados em classes.

A *média aritmética* é a idéia que ocorre a grande maioria das pessoas quando se fala em média; como ela possui certas propriedades importantes, ela é a medida de posição mais utilizada. Contudo, ela pode nos levar a erros de interpretação, assim, a *mediana* pode ser a medida mais recomendada em muitas situações. Ainda nesta seção indicaremos estas situações.

A média aritmética ( $Me$ ), é a soma dos valores numéricos de uma variável dividida pelo número deles. Por exemplo, considere os pesos ao nascer, em kg, de 10 bezerros da raça de gado Crioula:

47	51	45	50	50	52	46	49	53	51
----	----	----	----	----	----	----	----	----	----

Assim, a média, será:

$$Me = \frac{47 + 51 + 45 + 50 + 50 + 52 + 46 + 49 + 53 + 51}{10} = 49,4 \text{ kg.}$$

Genericamente, a  $Me$ , quando todos os valores são diferentes uns dos outros, é dada por:

$$Me(X) = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (2.4)$$

onde  $n$  é o número de dados da amostra e  $X$  é uma variável em estudo, por exemplo, peso ao nascer. A letra grega  $\sum$ , leia-se sigma (maiúscula), como pode ser visto na equação 2.4, representa um somatório, isto é, uma soma de valores.

Agora, quando temos um conjunto de  $n$  valores de uma variável  $X$ , dos quais  $n_1$  são iguais a  $x_1$ ,  $n_2$  são iguais a  $x_2$ , ...,  $n_k$  são iguais a  $x_k$ , então, a média aritmética de  $X$  é dada por:

$$Me(X) = \frac{n_1x_1 + n_2x_2 + \dots + n_kx_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i x_i}{\sum_{i=1}^k n_i} = \frac{\sum_{i=1}^k n_i x_i}{n} \quad (2.5)$$

onde  $k$  é o número de valores diferentes da variável em estudo. Se  $f_i = n_i/n$  representa a frequência relativa da observação  $x_i$ , então,  $Me$  pode ser escrita da seguinte maneira:

$$Me(X) = \sum_{i=1}^k f_i x_i.$$

**Exemplo.** Para os dados dos pesos ao nascer de bezerros da raça Crioula, com  $k = 8$  valores diferentes, temos:

$$Me(X) = \frac{(1 \times 45) + (1 \times 46) + (1 \times 47) + (1 \times 49) + (2 \times 50) + (2 \times 51) + (1 \times 52) + (1 \times 53)}{(1 + 1 + 1 + 1 + 2 + 2 + 1 + 1)}$$

$$\begin{aligned}
&= (0,10 \times 45) + (0,10 \times 46) + (0,10 \times 47) + (0,10 \times 49) + (0,20 \times 50) + \\
&+ (0,20 \times 51) + (0,10 \times 52) + (0,10 \times 53) \\
&= 49,4kg.
\end{aligned}$$

**Exercício.** Os valores da variável peso ao nascer de uma amostra de 10 bezerros da raça Nelore foram:

51	40	46	48	54	56	44	43	55	57
----	----	----	----	----	----	----	----	----	----

Encontre a média aritmética da amostra dos pesos ao nascer de bezerros da raça Gir.

O diagrama de pontos para as duas raças é dado na Figura 2.27. Percebe-se que os dois

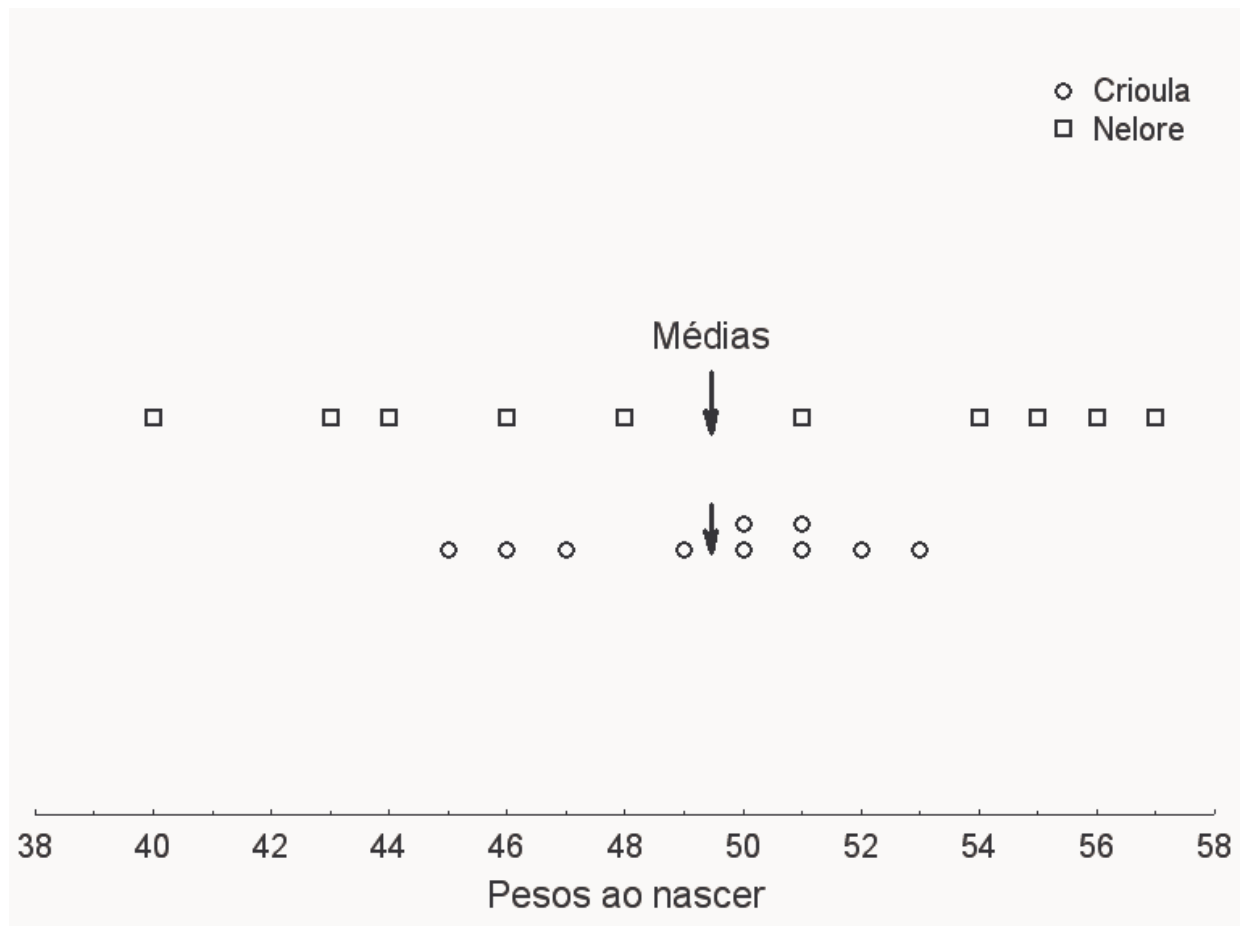


Figura 2.27: Diagrama de pontos para peso ao nascer das raças Crioula e Nelore

conjuntos de dados são bastante diferentes, entretanto, apresentam a mesma média, então, a média aritmética, por si só, tem muito pouco valor científico. É preciso alguma medida de variabilidade para acompanhar a média. Isto será visto na próxima seção.

A principal restrição ao uso da média aritmética é que a mesma é muito sensível a valores excessivamente altos ou baixos (valores discrepantes ou "outliers"). Ela é uma medida bastante adequada quando os dados apresentam pelo menos aproximadamente uma distribuição normal. Quando a distribuição é assimétrica deve-se utilizar preferencialmente a mediana.

No caso em que os dados estão agrupados em classes de ocorrências, a expressão da média aritmética é dada por:

$$Me = \frac{\sum_{i=1}^k n_i s_i}{n} = \sum_{i=1}^k f_i s_i$$

onde  $n_i$ ,  $f_i$  e  $s_i$  são a frequência absoluta, frequência relativa e o ponto médio da  $i$ -ésima classe, respectivamente, e  $k$  é o número de classes do histograma.

**Exemplo.** Para a distribuição de frequências da Tabela 2.25, que diz respeito aos rendimentos, em kg/ha, de híbridos de milho, a média aritmética tem como resultado,  $Me(X) = 5105,031 \text{ kg}$ . Vamos ao cálculo.

$$\begin{aligned} Me(X) &= 0,0313(4214,5) + 0,3750(4697,5) + 0,4063(5180,5) + 0,0937(5663,5) + 0,0937(6146,5) \\ &= 5105,031 \text{ kg} \end{aligned}$$

*Observação:* em termos computacionais, os programas estatísticos calculam a média facilmente, sem a necessidade de formar as classes.

Acontece, muitas vezes, que determinados valores de um conjunto de dados são mais importantes que os demais, ou seja, tem pesos diferentes, merecendo assim um tratamento especial.

**Exemplo.** Os tubérculos de batatas sementes são classificados, para efeito de comercialização, em quatro tipos de tamanhos (diâmetros), a saber:

- Tipo I - maior que 50 mm até 60 mm inclusive (50 a 60];
- Tipo II - maior que 40 mm até 50 mm inclusive (40 a 50];
- Tipo III - maior que 28 mm até 40 mm inclusive (28 a 40];
- Tipo IV - de 23 mm até 28 mm inclusive [23 a 28].

As batatas sementes são comercializadas em caixas de 30 kg. Um agricultor produziu 500 caixas em um hectare ( $10.000 \text{ m}^2$ ), assim distribuídas:

- 100 caixas do tipo I  $\rightarrow$  preço: 1500 u.m./cx;
- 180 caixas do tipo II  $\rightarrow$  preço: 3500 u.m./cx;

- 140 caixas do tipo III  $\rightarrow$  preço: 3000 u.m./cx;
- 80 caixas do tipo IV  $\rightarrow$  preço: 1600 u.m./cx.

Qual o preço médio, por caixa, obtido pelo agricultor? Podemos usar a expressão 2.5 para obter esse valor. Temos que a variável  $X$  é o preço da caixa de batata semente, portanto,  $x_1$  é o preço da caixa do tipo I, e assim por diante;  $n_i$  é o número de caixas produzidas por cada tipo,  $i = 1, 2, 3, 4$ . Vamos ao cálculo.

$$\begin{aligned} Me(X) &= \frac{\sum_{i=1}^4 n_i x_i}{\sum_{i=1}^4 n_i} = \frac{(100 \times 1500) + (180 \times 3500) + (140 \times 3000) + (80 \times 1600)}{100 + 180 + 140 + 80} \\ &= \frac{1328000}{500} = 2656,00 \text{ u.m.} \end{aligned}$$

### *Propriedades da Média Aritmética*

A média aritmética possui algumas propriedades importantes, dentre as quais vamos discutir duas. Antes de apresentá-las vamos ver o que significa um desvio ou resíduo de um dado em relação a sua média. Esse desvio é calculado como:

$$d_i = x_i - \bar{x}.$$

Assim, existem desvios positivos, negativos e nulos. Para os dados de pesos ao nascer de bezerros da raça Crioula, temos:

$d_1=45-49,4=-4,4$	$d_2=46-49,4=-3,4$	$d_3=47-49,4=-2,4$	$d_4=49-49,4=-0,4$	$d_5=50-49,4=0,6$
$d_6=50-49,4=0,6$	$d_7=51-49,4=1,6$	$d_8=51-49,4=1,6$	$d_9=52-49,4=2,6$	$d_{10}=53-49,4=3,6$

A primeira propriedade é que a soma dos desvios calculados em relação a média aritmética do conjunto de dados é nula:

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

No exemplo:  $-4,4-3,4-2,4-0,4+0,6+0,6+1,6+1,6+2,6+3,6=0$ .

A segunda propriedade é que a soma dos quadrados dos desvios em relação a média é um mínimo. Formalmente, temos:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n d_i^2 = \text{mínimo}.$$

Veremos a utilização dessas propriedades quando tratarmos do estudo de medidas de dispersão.



Uma segunda medida de tendência central é a *mediana*. A mediana divide um conjunto de dados ao meio, onde 50% dos valores se posicionam abaixo da mediana, e 50% dos valores se posicionam acima da mediana, portanto, a mediana, é uma medida baseada na ordenação dos dados (*rank*, em inglês).

**Definição:** a mediana de um conjunto de valores, é o valor  $Md$  que ocupa a posição  $(n + 1)/2$ , quando os dados estão ordenados crescentemente. Se  $(n + 1)/2$  for fracionário, toma-se o valor da posição imediatamente superior a  $(n + 1)/2$ .

**Exemplo:** Vamos calcular a mediana dos pesos ao nascer de 10 gatas da raça Crioula. A distribuição de frequências é dada na Tabela 2.28. O uso da fórmula para o cálculo da mediana é  $Md = (50 + 50)/2 = 50$ . Na Figura 2.28, a

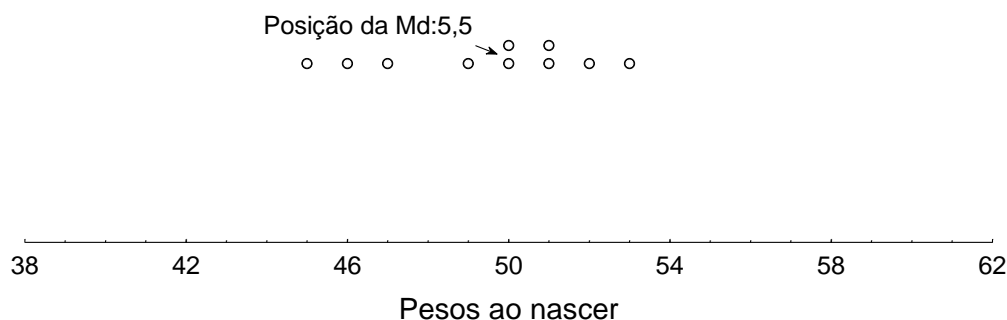


Figura 2.28: Diagrama de pontos para peso ao nascer da raça Crioula - cálculo da mediana

mediana será a média aritmética entre os valores que ocupam a 5ª e a 6ª posição, então,

$$Md = (50 + 50)/2 = 50 \text{ kg}.$$

Uma medida estreitamente relacionada com a mediana são os *quartis*. Embora não sejam medidas de tendência central, serão aqui tratadas devido a semelhança com o cálculo da mediana. Os quartis dividem um conjunto de dados em 4 partes iguais, do seguinte modo: aproximadamente 25% dos dados serão inferiores ao primeiro quartil ( $Q_1$ ), ou seja, 25% dos dados estão localizados em posição inferior ao primeiro quartil; 50% dos dados ocupam posição inferior ao segundo quartil ( $Md$ ), que é a mediana, e aproximadamente 75% dos dados ocupam posição inferior ao terceiro quartil ( $Q_3$ ), portanto, 25% dos valores estão localizados em posição superior ao terceiro quartil.

**Definição:** dado um conjunto de dados ordenados, podemos obter, de forma aproximada, o primeiro quartil, ( $Q_1$ ), como sendo a mediana dos valores de posições menores ou iguais

à posição da mediana. A mediana dos valores de posições maiores ou iguais à posição da mediana corresponde ao terceiro quartil, ( $Q_3$ ).

**Exemplo.** Vamos calcular o  $Q_1$  e o  $Q_3$  para os dados de pesos ao nascer, em kg, de bezerros da raça Crioula. Os dados são: 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100. A posição da mediana é  $i = (n + 1)/2 = (100 + 1)/2 = 50,5$ . A posição do primeiro quartil é  $j = (n + 1)/4 = (100 + 1)/4 = 25,25$ . A posição do terceiro quartil é  $k = 3(n + 1)/4 = 3(100 + 1)/4 = 75,75$ . O primeiro quartil é o valor na posição 25,25 e o terceiro quartil é o valor na posição 75,75.

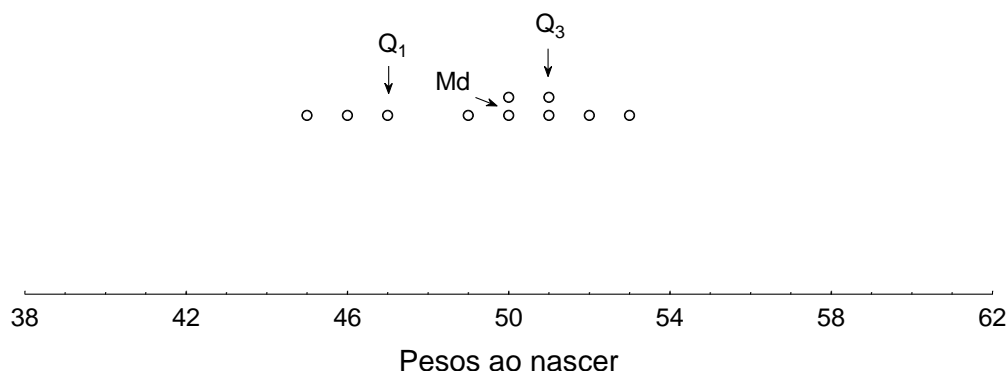


Figura 2.29: Diagrama de pontos para peso ao nascer da raça Crioula - 1º e 3º quartis

O cálculo da mediana e dos quartis para um histograma serão feitos por meio de argumentos geométricos, através da proporcionalidade existente entre área e base de retângulos. Geometricamente, a mediana é o valor da abcissa que determina uma linha vertical que divide o histograma em duas partes de áreas iguais, então, 50% da área do histograma está abaixo da mediana e 50% da área está acima da mediana. Da mesma forma, o 1º quartil é o valor da abcissa que determina uma linha vertical que divide o histograma em duas partes de áreas diferentes, ou seja, 25% da área está abaixo do 1º quartil e 75% da área está acima do 1º quartil. O terceiro quartil é o valor da abcissa que determina uma linha vertical que divide o histograma em duas partes de áreas diferentes, ou seja, 75% da área está abaixo do 3º quartil e 25% da área está acima do 3º quartil.

**Exemplo.** Vamos encontrar os valores da mediana e dos quartis para a variável rendimento de grãos, em kg/ha, de híbridos de milho. O histograma é dado na Figura 2.16. A classe mediana é aquela cuja porcentagem acumulada é, de pelo menos, 50%, então, no exemplo, a mediana é um valor que encontra-se na terceira classe, cujos limites inferior e superior são: 4939 e 5422, respectivamente. Agora, vamos usar a regra da proporcionalidade,

que é dada por:

$$\frac{5422 - 4939}{40,6} = \frac{Md - 4939}{9,4}.$$

No lado esquerdo do sinal de igualdade, temos o limite superior (5422) subtraído do limite inferior (4939) da classe mediana, dividido pela porcentagem de observações desta classe (40,6). No lado direito do sinal de igualdade, temos um outro retângulo cuja área vale 9,4%, obtida da seguinte forma: somando-se as áreas das classes anteriores à classe mediana, temos  $3,13 + 37,50 = 40,63\%$ , assim,  $50,0 - 40,6 = 9,4\%$ ; a base deste "novo" retângulo é  $Md - 4939$ , cujo valor  $Md$  estamos interessados em determinar.

Obtemos o valor da mediana fazendo-se:

$$Md = 11,8276(9,4) + 4939 = 5050,828 \text{ kg}.$$

O cálculo dos quartis é feito de maneira análoga ao cálculo da mediana. A classe que contém o 1º quartil é aquela cuja porcentagem acumulada é de pelo menos 25%, no exemplo, o 1º quartil encontra-se na segunda classe, cujos limites inferior e superior são: 4456 e 4939, respectivamente; a porcentagem de observações nesta classe é de 37,5%. Pela regra da proporcionalidade encontramos:

$$\frac{4939 - 4456}{37,5} = \frac{Q_1 - 4456}{21,9} \Rightarrow Q_1 = 4738,072 \text{ kg}.$$

O valor 21,9 é obtido fazendo-se  $25,0 - 3,1$ .

A classe que contém o 3º quartil é aquela cuja percentagem acumulada seja de pelo menos 75%, no exemplo, é a terceira classe, cujo limite inferior, superior e porcentagem de observações valem: 4939, 5422 e 40,6, respectivamente. Aplicando a regra obtemos:

$$\frac{5422 - 4939}{40,6} = \frac{Q_3 - 4939}{34,4} \Rightarrow Q_3 = 5348,241 \text{ kg}.$$

Na Figura 2.30 apresentamos o histograma da variável rendimento de grãos, em kg/ha, acompanhado dos quartis e mediana.

Finalmente, temos a *moda* (Mo), definida com o valor que ocorre com maior frequência num conjunto de dados. Pela própria definição, percebe-se que a moda pode não existir, pois pode não existir um valor mais freqüente que os demais, ou existindo, pode não ser a única, assim temos séries amodal, unimodal, bimodal, trimodal, etc.

**Exemplo.** Para os dados de pesos ao nascer de bezerros da raça Crioula temos duas modas:  $M_{o1} = 50$  e  $M_{o2} = 51$ , portanto, a série é bimodal. No caso do histograma, devemos encontrar a classe modal, isto é, a classe que apresenta a mais alta frequência. A moda é

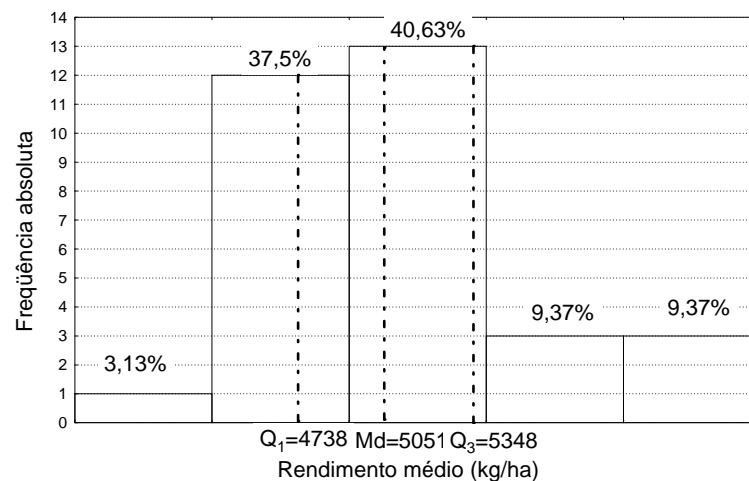


Figura 2.30: Representação geométrica da mediana, do primeiro quartil e do terceiro quartil

por definição o ponto médio dessa classe. A utilidade da moda ocorre quando num conjunto de dados, um, dois, ou um grupo de valores, ocorrem com muito maior frequência do que outros.

## 2.2.2 Medidas de Dispersão

### 2.2.2.1 Variância, Desvio Padrão, Coeficiente de Variação e Desvio Interquartilico

Quando apresentamos uma medida de tendência central para representar um conjunto de dados, é necessário que esta medida seja acompanhada de uma outra medida que resuma a variabilidade dos dados, ou seja, a dispersão dos dados. Na Figura 2.27 estão representados os pesos ao nascer de bezerros das raças Crioula e Nelore. Observa-se que as duas distribuições têm a mesma média amostral, porém, os valores para a raça Nelore estão bem mais esparramados (dispersos) do que os valores da raça Crioula. Dizemos que a variabilidade na raça Nelore é maior do que na raça Crioula. Enfim, os dois conjuntos de dados são bastante diferentes, apesar de apresentarem a mesma média amostral. Com isso, fica claro que são necessários, pelo menos dois tipos de medidas para descrever razoavelmente bem um conjunto de dados. Uma medida de dispersão quantifica a magnitude da variabilidade dos dados. Vamos apresentar as seguintes medidas de dispersão:

- Variância
- Desvio padrão
- Coeficiente de variação

$X_i$	5,4	5,4	5,8	6,4	6,4	6,6	6,6	6,8	6,8	7,0	7,3	7,3	7,5	8,2	8,8	8,8
$\bar{x}$	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94	6,94
$(x_i - \bar{x})$	-1,54	-1,54	-1,14	-0,54	-0,54	-0,34	-0,34	-0,14	-0,14	0,06	0,36	0,36	0,56	1,26	1,86	1,86
$(x_i - \bar{x})^2$	2,37	2,37	1,30	0,29	0,29	0,12	0,12	0,02	0,02	0,00	0,13	0,13	0,31	1,59	3,46	3,46

Figura 2.31: Cálculo da variância do diâmetro da roseta foliar de bromélias expostas ao sol

- Desvio interquartilico

Para os métodos estatísticos, a medida de dispersão é de fundamental importância, pois a necessidade do uso da estatística, é devida a existência de variabilidade nos dados observados e isto sempre ocorre para fenômenos aleatórios.

Para o cálculo da variância e do desvio padrão, o princípio básico é analisar os desvios das observações em relação à média aritmética. Em cada caso, o valor zero para a variância ou desvio padrão, indica ausência de variação; a variação vai aumentando à medida que aumenta o valor da medida de dispersão.

A variância é uma medida de dispersão que nos fornece uma idéia da variabilidade dos dados em torno da média. Ela é o quociente entre a soma dos quadrados dos desvios dos dados observados, tomados em relação a sua média aritmética, e o número de dados ( $n$ ) menos 1. É representada por  $s^2$  quando os dados são oriundos de uma amostra e por  $\sigma^2$ , leia-se sigma ao quadrado, quando os dados representam a população. Vamos ilustrar os passos para o cálculo da variância através de um exemplo.

**Exemplo.** Vamos calcular a variância para os dados de uma amostra de tamanho,  $n = 16$ , do diâmetro (em cm) da roseta foliar de bromélias expostas ao sol. Os dados amostrais obtidos foram:

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

Os passos para o cálculo da variância são dados na Figura 2.31. Precisamos do valor da média aritmética dos dados, no exemplo temos  $\bar{x} = 6,94$  cm. Após são calculados os desvios dos dados em relação à média,  $(x_i - \bar{x})$ , onde  $x_1 = 5,4$ ,  $x_2 = 5,4$ ,  $x_3 = 5,8$ , ...,  $x_{16} = 8,8$ ; em seguida estes desvios são elevados ao quadrado,  $(x_i - \bar{x})^2$ . Finalmente, aplicamos a expressão da variância amostral que é dada por:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}. \quad (2.6)$$

No exemplo, com base nos resultados de 2.31, temos:

$$s^2 = \frac{15,98}{15} = 1,065 \text{ cm}^2.$$

Foi também selecionada uma outra amostra de 16 valores de diâmetros da roseta foliar de bromélias em ambiente de sombra. Os resultados foram:

13,4	13,7	14,4	14,6	14,6	14,8	15,2	15,2
15,4	15,7	16,2	16,4	16,7	17,5	17,8	17,8

Para esses dados o valor da variância é  $s^2 = 1,893 \text{ cm}^2$ . Obtenha esse valor. A conclusão que podemos tirar é que, para a variável diâmetro, as bromélias em ambiente de sombra são mais heterogêneas (apresentam maior variabilidade) do que as expostas ao sol.

Para os dados de peso ao nascer de bezerros, representados na Figura 2.27, obtemos  $s_{CR}^2 = 6,92 \text{ kg}^2$  e  $s_N^2 = 36,48 \text{ kg}^2$ , para as raças Crioula e Nelore, respectivamente. Portanto, a raça Nelore é muito mais heterogênea do que a raça Crioula, para peso no nascimento.

A variância apresenta um inconveniente de ordem prática, pois como ela é expressa em unidades ao quadrado, isto causa problemas de interpretação. Uma outra medida de variabilidade, calculada através da variância, é o *desvio padrão* da amostra ( $s$ ). Na prática o desvio padrão é preferido em relação a variância, pois ele é expresso na mesma unidade dos dados originais. O desvio padrão nada mais é do que a raiz quadrada da variância, logo:

$$s = \sqrt{s^2}. \quad (2.7)$$

**Exemplo.** Para os dados amostrais do diâmetro da roseta foliar de bromélias, em cm, expostas ao sol e em ambiente de sombra, os valores do desvio padrão são,  $s_{Sol} = 1,032 \text{ cm}$  e  $s_{Sombra} = 1,376 \text{ cm}$ , respectivamente. O desvio padrão é uma medida relativa, assim, só faz sentido afirmar que um desvio é grande (ou pequeno) comparativamente à outro. Nesse exemplo, o desvio padrão para expostas ao sol é menor do que para ambiente de sombra. Podemos dizer que, para expostas ao sol, a dispersão dos valores em torno da média é, em média igual a 1,032 cm e, para ambiente de sombra, a dispersão dos valores em torno da média é, em média, igual a 1,376 cm.

Existe uma expressão mais geral para o cálculo da variância e desvio padrão dada por:

$$s^2 = \frac{\sum_{i=1}^k (x_i - \bar{x})^2 n_i}{n - 1}, \quad (2.8)$$

onde  $k$  é o número de valores diferentes de  $x_i$ ;  $n_i$  é a frequência de ocorrência do  $i$ -ésimo valor. Para os dados de diâmetro de roseta expostas ao sol, também podemos calcular a variância usando 2.8, da seguinte forma:

$$s^2 = \frac{(2 \times 2,37) + (1 \times 1,30) + (2 \times 0,29) + (2 \times 0,12) + \dots + (2 \times 3,46)}{16 - 1} = 1,065 \text{ cm}^2.$$

Em algumas situações, como por exemplo, quando a população não é muito grande, é preferível realizar o censo, isto é, obter as informações sobre todos os elementos, plantas, pessoas etc. que constituem esta população. Por exemplo, num estudo sobre a consangüinidade na comunidade da Costa da Lagoa da Conceição, Florianópolis, SC, foram levantados os dados de todos os moradores (população). Nesse caso temos a *variância populacional*, representada por  $\sigma^2$ , leia-se sigma ao quadrado, e é calculada através da expressão:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}, \quad (2.9)$$

onde  $\mu$  (leia-se "mi") é calculada por:  $\mu = \sum_{i=1}^N x_i / N$ , é a média obtida com todos os dados da população,  $N$  é o tamanho da população, isto é, o número total de dados. Da mesma forma, o desvio padrão populacional é obtido através da raiz quadrada da variância e é representado por  $\sigma$ .

O *coeficiente de variação* é utilizado quando temos interesse em comparar variabilidades em situações onde as médias são muito diferentes ou as unidades de medida são diferentes. Nesse caso, utilizamos o coeficiente de variação, pois é uma medida relativa percentual da variabilidade dos dados em torno da média, isto é,

$$CV(\%) = \frac{s}{\bar{x}} \times 100. \quad (2.10)$$

É uma medida de dispersão relativa porque estabelece uma relação entre o desvio padrão ( $s$ ), e a média ( $\bar{x}$ ). Sendo uma medida independente da unidade da variável é útil para se estudar comparativamente duas ou mais distribuições.

**Exemplo.** Os dois conjuntos de dados abaixo representados, referem-se ao comprimento do corpo, dado em  $mm$ , e peso de fêmeas, dado em  $g$ , de *Penaeus paulensis* (Crustacea, Decapoda, Penaidae), respectivamente, obtidos nas despescas dos viveiros do Centro de Ciências Agrárias (CCA) da Universidade Federal de Santa Catarina (UFSC).

27	26	26	25	25	25	25	23	23
30	30	33	33	33	35	35	35	36

0,14	0,16	0,14	0,12	0,12	0,12	0,11	0,09	0,07
0,18	0,23	0,28	0,28	0,32	0,31	0,33	0,36	0,33

A média e o desvio padrão para cada uma das amostras são:  $\bar{x}_C = 29,17 \text{ mm}$ ,  $\bar{y}_P = 0,21 \text{ g}$ ,  $s_C = 4,63 \text{ mm}$  e  $s_P = 0,10 \text{ g}$ . Assim, os coeficientes de variação valem:

$$cv_C = \frac{4,6305}{29,1667} \times 100 = 15,88\% \quad cv_P = \frac{0,0984}{0,2050} \times 100 = 48,00\%.$$

Portanto, a variabilidade na variável peso é muito maior do que na variável comprimento. A variabilidade é mais significativa para peso. Observe que pelos valores dos desvios padrões a conclusão seria diferente.

O coeficiente de variação é bastante utilizado em dinâmica de populações vegetais ou animais. Outra aplicação importante do coeficiente de variação é na estatística experimental, pois ele indica a precisão do experimento, ou seja, a capacidade de realizarmos novamente o experimento, sob as mesmas condições, e produzir resultados semelhantes. Quais são os valores de C.V. aceitáveis na experimentação? Os valores dos coeficientes de variação dependem do tipo de pesquisa e da variável em estudo, sendo assim, não existe uma orientação geral, deve-se fazer uma busca bibliográfica em pesquisas similares. Numa situação de ensaios agrícolas de campo, para culturas anuais como soja, milho e feijão e variável rendimento de grãos, temos a seguinte orientação:

$$\begin{aligned} C.V. \leq 10\% & \Rightarrow \text{baixo} \\ 10\% < C.V. \leq 20\% & \Rightarrow \text{médio} \\ 20\% < C.V. \leq 30\% & \Rightarrow \text{alto} \\ C.V. > 30\% & \Rightarrow \text{muito alto} \end{aligned}$$

Para dados agrupados em classes (histograma), podemos calcular a variância através das seguintes expressões:

$$s^2 = \frac{\sum_{i=1}^k (s_i - \bar{x})^2 n_i}{n - 1} \quad \text{para dados amostrais} \quad (2.11)$$

$$\sigma^2 = \frac{\sum_{i=1}^k (s_i - \mu)^2 n_i}{N} \quad \text{para dados populacionais} \quad (2.12)$$

onde:  $k$  é o número de classes;  $s_i$  é o ponto médio da  $i$ -ésima classe;  $n_i$  é o número de dados observados na  $i$ -ésima classe e  $N$  é o tamanho da população. A única alteração das fórmulas anteriores é a substituição dos valores originais,  $x_i$ , pelos pontos médios,  $s_i$ .

**Exemplo.** Para a distribuição de freqüências da Tabela 2.25, que diz respeito aos rendimentos, em kg/ha, de híbridos de milho, o valor da variância é dado por:

$$\begin{aligned} s^2 &= \frac{1}{32 - 1} \{ [(4214,5 - 5105,0)^2 \times 1] + [(4697,5 - 5105,0)^2 \times 12] + [(5180,5 - 5105,0)^2 \times 13] \\ &\quad + [(5663,5 - 5105,0)^2 \times 3] + [(6146,5 - 5105,0)^2 \times 3] \} \\ s^2 &= 227409,74 \text{ (kg/ha)}^2. \end{aligned}$$

O desvio padrão vale:

$$s = \sqrt{s^2} = 476,87 \text{ kg/ha}.$$



Da mesma forma que a média aritmética, a variância é uma medida de dispersão, que representa bem a realidade, quando os dados apresentam pelo menos aproximadamente uma distribuição normal. Para distribuições assimétricas, uma medida da variabilidade é dada pelo *desvio interquartilico*, calculada por:

$$Q_3 - Q_1.$$

**Exemplo.** Para os valores de rendimento de grãos, em kg/ha, de híbridos de milho, o valor do desvio interquartilico é 5348,2-4738,1=610,1. Temos que 50% dos valores encontram-se no intervalo de 4738,1 a 5348,2.

## *A Média e o Desvio Padrão Sob um Modelo Normal*

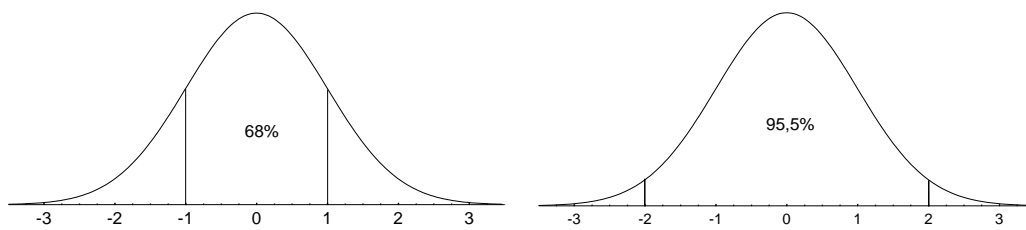
Um caso importante acontece quando os dados numa distribuição de frequências, apresentam o histograma com a forma de sino, esta distribuição é comumente chamada de distribuição normal ou Gaussiana (devido a Karl F. Gauss). Nessa distribuição sempre temos:

1. 68% dos dados estão compreendidos entre a média mais ou menos um desvio padrão  $(\mu - 1\sigma; \mu + 1\sigma)$ .
2. 95,4% dos dados estão compreendidos entre a média mais ou menos dois desvios padrões  $(\mu - 2\sigma; \mu + 2\sigma)$ .
3. 99,7% dos dados estão compreendidos entre a média mais ou menos três desvios padrões  $(\mu - 3\sigma; \mu + 3\sigma)$ .

Esses resultados podem ser visualizados na Figura 2.32.

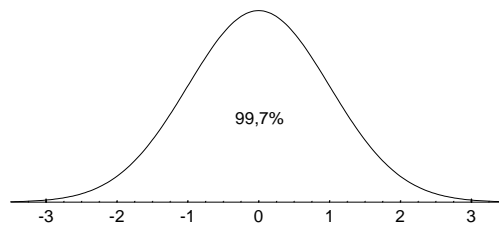
Nos trabalhos científicos, é bastante comum e aconselhável, a representação dos dados na forma  $\bar{x} \pm s$  ou  $\bar{x}(s)$ . Isso indica, que sob normalidade dos dados, 68% dos dados estão entre a média mais 1 desvio padrão e a média menos 1 desvio padrão. O responsável pelos dados (Agrônomo, Biólogo), deve avaliar se estes intervalos são amplos (pouco precisos), ou não (precisos), para o fenômeno real em estudo.

Vimos a importância de se identificar as observações discrepantes num conjunto de dados. Numa distribuição aproximadamente normal, algum(ns) valor(es) maior(es) que  $\bar{x} + 3s$ , ou, menor(es) que  $\bar{x} - 3s$ , são considerados valores discrepantes ou *outliers*. Para a amostra de peso de fêmeas de *Penaeus paulensis* encontramos  $\bar{x} = 0,2050$  e  $s = 0,0984$ . Temos que  $\bar{x} + 3s = 0,2050 + 3(0,0984) = 0,5000$  e  $\bar{x} - 3s = 0,2050 - 3(0,0984) = -0,0902$ . Como na amostra estudada não temos nem um valor acima de 0,50 e nem abaixo de 0 (zero) concluímos, por este critério, que não existe valor discrepante.



(a) Aproximadamente 68% dos dados estão entre  $\mu \pm \sigma$

(b) Aproximadamente 95,4% dos dados estão entre  $\mu \pm 2\sigma$



(c) Aproximadamente 99,7% dos dados estão entre  $\mu \pm 3\sigma$

Figura 2.32: Distribuição dos dados, sob o modelo normal, de acordo com  $\mu$  e  $\sigma$ .

## Uma Regra Empírica para a Dispersão

Existe uma regra empírica determinada por um matemático russo chamado Tchebysheff, que diz o seguinte: "dado um número  $k > 1$ , e uma amostra de  $n$  observações,  $x_1, x_2, \dots, x_n$  é certo que *pelo menos*  $(1 - \frac{1}{k^2})$  dessas observações pertencerão ao intervalo  $\bar{x} - ks$  e  $\bar{x} + ks$ . Esta regra é importante, principalmente quando não se conhece a distribuição dos dados.

**Exemplo.** O objetivo é estudar o número de insetos de determinada espécie por árvore de *Pinus elliotti*. Uma amostra de tamanho  $n = 16$  árvores foi coletada e os resultados foram:

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

A média e o desvio padrão valem  $\bar{x} = 3,6875$  insetos e  $s = 2,3866$  insetos, respectivamente. Para  $k = 2$  temos que pelo menos 75% das observações encontram-se no intervalo  $[0;8]$ . Verificando-se os dados da amostra encontramos 100% das observações nesse intervalo.

### Exercício resolvido

1. Utilizando os dados da Tabela 2.33, calcule a média, a variância, o desvio padrão e o coeficiente de variação para cada um dos tratamentos (Sem e Com rizóbio).

	Sem rizóbio	Com rizóbio
$\bar{x} = \sum_{i=1}^{20} x_i / 20$	34,35	43,30
$s^2 = \sum_{i=1}^{20} (x_i - \bar{x})^2 / (20 - 1)$	20,45	19,91
$s = \sqrt{s^2}$	4,58	4,46
$CV(\%) = \frac{s}{\bar{x}} \cdot 100$	13,16%	10,30%

Utilizando a média e o desvio padrão, como você descreveria os dados do tratamento 1 e do tratamento 2? ainda utilizando a média e o desvio padrão, verifique se existem dados suspeitos e discrepantes, para o tratamento 1 e 2. Indique os cálculos. Resultados:

	Descrição	Suspeitos	Discrepantes
	$\bar{x} \pm 1s$	$\bar{x} \pm 2s$	$\bar{x} \pm 3s$
Sem rizóbio	34, 35 $\pm$ 4, 52 [29,83;38,87]	34, 35 $\pm$ 2(4, 52) [25,31;43,39]	34, 35 $\pm$ 3(4, 52) [20,78;47,92]
Com rizóbio	43, 3 $\pm$ 4, 46 [38,84;47,76]	43, 3 $\pm$ 2(4, 46) [34,38;52,22]	43, 3 $\pm$ 3(4, 46) [29,92;56,68]

No tratamento 1 verificamos que o valor 25 é suspeito, porém, não é considerado discrepante; no tratamento 2 verificamos que o valor 34 é suspeito, porém, não é discrepante.

Calcule a mediana e os quartis para cada um dos tratamentos. Aplicando as definições encontramos:

	$Q_1$	Md	$Q_3$
Sem rizóbio	31	34	37,5
Com rizóbio	40,5	44	46,5

Fazer uma representação gráfica das médias e dos desvios padrões. A representação é dada na Figura 2.33. As alturas das colunas representam as médias dos tratamentos portanto, no tratamento com rizóbio a altura média é maior do que sem tratamento. Os desvios padrões dos tratamentos são praticamente iguais.

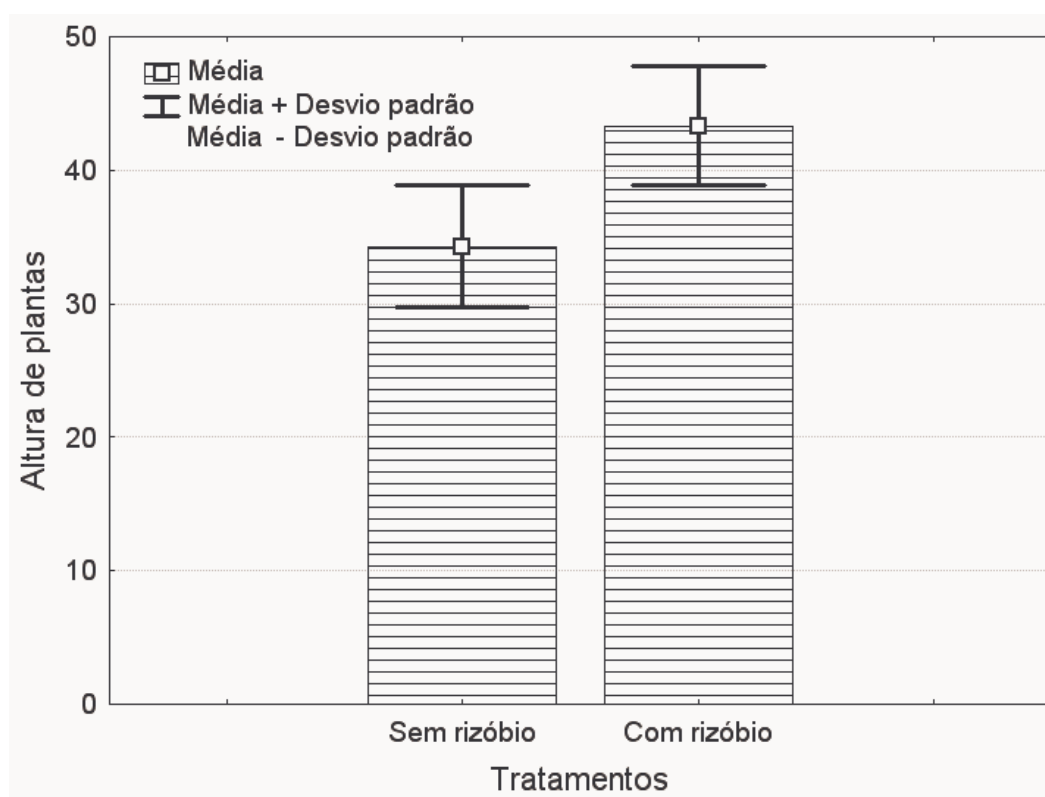


Figura 2.33: Representação gráfica das médias e dos desvios padrões

#### 2.2.2.2 Exercícios Propostos

1. A tabela de distribuição de frequências 2.34 foi construída a partir dos dados da Tabela 2.31. Observe que a distribuição é assimétrica à direita, pois possui uma cauda mais longa à direita. Calcular a média, a variância, o desvio padrão, a mediana e os quartis da distribuição de frequências. Considere o conjunto 1 formado pela média e variância, e

Tabela 2.34: Distribuição de frequência da variável altura de calos em cm, num período de 30 dias "in vitro" da espécie *Mandevilla velutina*

Altura de calos	Ponto médio ( $s_i$ )	Frequência absoluta ( $n_i$ )
0,00 $\vdash$ 0,14	0,07	15
0,14 $\vdash$ 0,28	0,21	1
0,28 $\vdash$ 0,42	0,35	5
0,42 $\vdash$ 0,56	0,49	8
0,56 $\vdash$ 0,70	0,63	7
0,70 $\vdash$ 0,84	0,77	2
Total		38

o conjunto 2 formado pela mediana e quartis. Qual dos dois conjuntos você recomendaria para essa distribuição? Justifique.

Tabela 2.35: Produção de cana-de-açúcar em t/ha

Variedade 1		Variedade 2		
65	78	88	93	99
68	80	89	95	
75	80	90	96	
76	82	91	97	
77	86	92	97	

Tabela 2.36: Peso de carne de mexilhões, em gramas, em dois locais

Sambaqui			Manguezal		
30,61	42,88	27,94	25,34	9,49	19,17
28,89	36,22	41,45	25,67	16,92	21,60
32,21	28,86	42,59	17,64	12,91	20,01
24,25	22,56	15,25	33,97	14,05	19,81
25,63	22,92	33,29	11,13	14,88	16,22

2. Para se estudar o comportamento de duas variedades de cana-de-açúcar, realizou-se um experimento onde foram obtidos os resultados descritos na Tabela 2.35. Para decidir se a produção média das duas variedades de cana-de-açúcar são semelhantes ou não, adotou-se o seguinte teste:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{onde} \quad s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} \quad (2.13)$$

Caso  $|t| < 2$  as produções médias são semelhantes, caso contrário são diferentes. Qual é a sua conclusão?

3. Na Tabela 2.36 temos os resultados da variável peso de carne, em gramas, de mexilhões de dois locais: 1) Sambaqui e 2) Manguezal. a) calcule a média e a mediana para cada um dos locais. Onde houve maior crescimento?

b) Calcule o  $Q_1$  e o  $Q_3$  para cada um dos locais. Explique o significado destes números.

c) Compare os dois locais quanto a homogeneidade (calcule uma medida de dispersão e conclua).

d) Calcule o coeficiente de variação para cada local e interprete. A conclusão é a mesma do item c)? Qual das duas conclusões é a definitiva?.

4. A Tabela 2.37 apresenta uma amostra de valores de peso de carne de mexilhão do Sambaqui. a) Construa um histograma. A distribuição apresenta a forma aproximada do

Tabela 2.37: Distribuição de frequências para peso de mexilhões da localidade de Sambaqui

Peso	Frequências	Porcentagens
$8 < peso \leq 11$	3	8,57
$11 < peso \leq 14$	6	17,14
$14 < peso \leq 17$	5	14,29
$17 < peso \leq 20$	7	20,00
$20 < peso \leq 23$	4	11,43
$23 < peso \leq 26$	4	11,43
$26 < peso \leq 29$	2	5,71
$29 < peso \leq 32$	2	5,71
$32 < peso \leq 35$	1	2,86
$35 < peso \leq 38$	1	2,86

modelo normal? Justifique.

b) Localize no histograma a classe que contém o percentil de ordem 90 ( $P_{90}$ ). Interprete este valor.

c) Acima de que peso encontram-se 85% (Calcule o  $P_{15}$ ) dos mexilhões?

### 2.2.3 O Uso da Mediana e dos Quartis na Interpretação de um Conjunto de Dados

O objetivo do uso da mediana e dos quartis é obter informações sobre a forma, o valor representativo, a dispersão e os valores discrepantes da distribuição dos dados observados. Através destas estatísticas é possível obter-se todas as informações relevantes de uma distribuição, ou seja, podemos responder às principais questões da pesquisa.

Sabemos que a média e o desvio padrão são afetados, de forma exagerada, por valores extremos (valores altos ou baixos), portanto, não são medidas indicadas para distribuições assimétricas, pois não representam bem a realidade dos fatos. Além disso, somente com a média e o desvio padrão não temos idéia da forma como os dados se distribuem. A sugestão é fazer uso das seguintes medidas:

i) Mediana.

ii) Os valores extremos (o menor valor e o maior valor) do conjunto de dados.

iii) O 1º e 3º quartis.

Tabela 2.38: Dados de crescimento do pseudobulbo de *Laelia purpurata*, Florianópolis, SC.

Luz Direta	1,6	1,6	1,9	1,9	2,1	2,1	2,1	2,1	2,1	2,4	2,5	2,5	
	2,7	3,4	3,4	3,7	3,9	4,2	4,8	6,3	6,5	7,2	8,8	9,4	9,5
Luz Indireta	1,4	1,9	2,8	3,1	3,5	3,5	3,6	3,9	4,3	4,5	4,6	4,8	
	6,3	6,5	6,7	6,7	6,8	6,9	8,1	8,6	10,4	12,7	16,3	16,8	16,9

Tabela 2.39: Cálculo dos quartis e extremos para dados de crescimento do pseudobulbo de *Laelia purpurata*

	Md	$Q_1$	$Q_3$	$E_i$	$E_s$	$Q_1 - 1,5(Q_3 - Q_1)$	$Q_3 + 1,5(Q_3 - Q_1)$
Luz direta	2,7	2,1	4,8	1,6	9,5	-1,95	8,85
Luz indireta	6,3	3,6	8,1	1,4	16,9	-3,15	14,85

Obtemos, então, o que se denomina na literatura, por esquema dos cinco números ou esquema extremos-e-quartis.

**Exemplo:** Foram tomadas duas amostras de tamanhos igual a 25 observações, de crescimento do pseudobulbo de *Laelia purpurata*, sob duas condições de luminosidade (com luz direta e com luz indireta). Os dados estão apresentados na Tabela 2.38. Os resultados dos cálculos das medianas e dos quartis, juntamente com os extremos  $E_i$  e  $E_s$ , são apresentados na Tabela 2.39. Nesta tabela, as duas últimas colunas representam um critério para identificar a presença de valores discrepantes, o qual passamos a descrever.

### Valores Discrepantes (em inglês: *Outliers*)

Com o uso dos quartis também é possível verificar (detectar) se um ou mais valores da distribuição dos dados são considerados valores discrepantes. Se algum valor for menor do que  $Q_1 - 1,5(Q_3 - Q_1)$ , ou maior do que  $Q_3 + 1,5(Q_3 - Q_1)$ , então, esse valor é considerado *outlier*. Num conjunto de dados pode existir mais do que um valor discrepante. No exemplo, esses limites são dados por: -1,95 e 8,85, para luz direta e, -3,15 e 14,85, para luz indireta, respectivamente. Portanto, observa-se na Tabela 2.38, que os valores 9,4 e 9,5 são considerados *outliers* para luz direta, e que os valores 16,3, 16,8 e 16,9, são considerados *outliers* para luz indireta.

Uma justificativa para utilizarmos o valor 1,5 nas expressões do cálculo dos valores discrepantes (deixaremos a prova para a seção 5.3.2), é que a área entre a curva normal e os



pontos limites  $Q_1 - 1,5(Q_3 - Q_1)$  e  $Q_3 + 1,5(Q_3 - Q_1)$  é igual a 99,3%. Portanto, estamos considerando 0,7% dos valores da distribuição normal como sendo valores discrepantes ou outliers. A ilustração é dada na Figura 2.34.

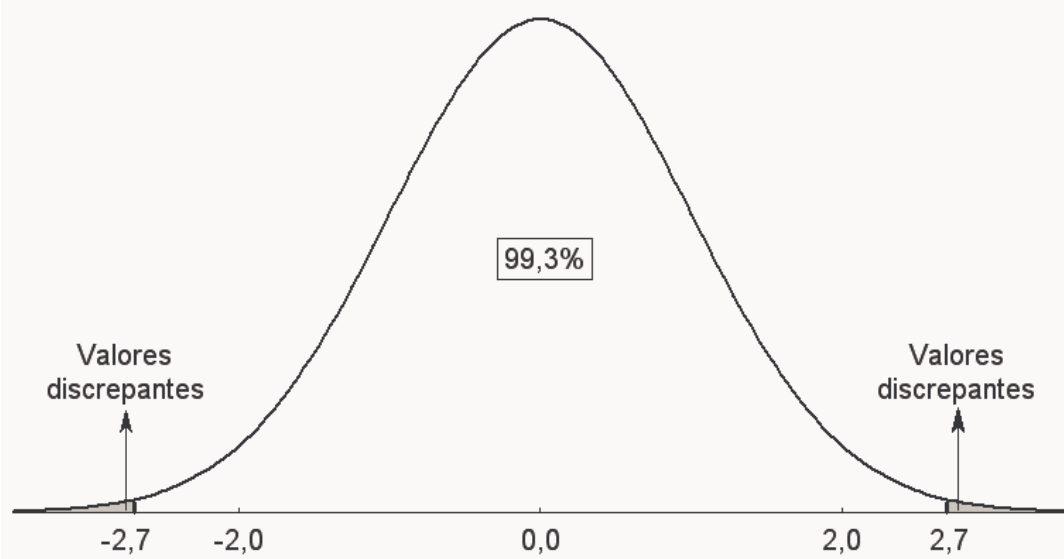


Figura 2.34: Área não sombreada (99,3%) entre os pontos limites na distribuição normal. A área sombreada corresponde aos valores discrepantes e é igual a 0,7%

Como vamos utilizar esses resultados para estudar a forma de uma distribuição de dados? Para uma distribuição simétrica, em forma de sino, a chamada distribuição normal, temos a Figura 2.35. Olhando-se para a Figura 2.35, esperamos intuitivamente que:

1.  $(Md - E_i) \cong (E_s - Md)$ , ou seja, a dispersão inferior é aproximadamente igual a dispersão superior;
2.  $(Md - Q_1) \cong (Q_3 - Md)$ ;
3.  $(Q_1 - E_i) \cong (E_s - Q_3)$ ;
4. As distâncias entre a mediana e os quartis sejam menores do que as distâncias entre os extremos e os quartis, ou seja,  $(Md - Q_1)$  e  $(Q_3 - Md)$  sejam menores do que  $(Q_1 - E_i)$  e  $(E_s - Q_3)$ .

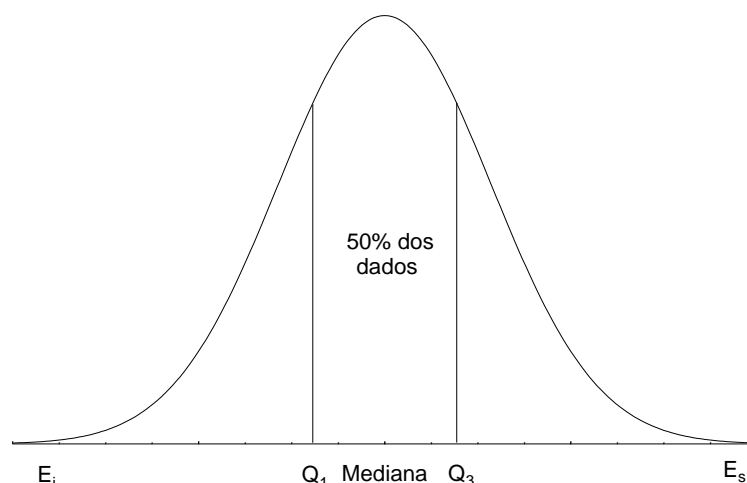


Figura 2.35: Forma da distribuição normal.

#### 2.2.4 Desenho Esquemático (em inglês leia-se *Box Plot*)

As informações obtidas na seção 2.2.3 podem ser representadas graficamente num desenho esquemático, como ilustrado na Figura 2.36. Os programas estatísticos de língua inglesa trazem esta figura com o nome de *box plot*. A seguir faremos comentários sobre a construção e interpretação do desenho esquemático da Figura 2.36.

O primeiro passo é construirmos o eixo e a escala para a variável resposta; no exemplo, para a variável comprimento do pseudobulbo, foi feita uma escala de 0 a 18 cm, com intervalos de 2 cm. O desenho esquemático fornece uma medida da posição central dos dados através da mediana. As medianas valem 2,7 cm e 6,3 cm, para luz direta e indireta, respectivamente, indicando que o crescimento de pseudobulbos é maior para luz indireta. Observe, na legenda da Figura 2.36, que as medianas estão representadas por quadradinhos vazios dentro das "caixas" (retângulos). O desenho esquemático, também, dá uma idéia da dispersão, ou contrariamente, da concentração dos valores, através do intervalo interquartílico ( $Q_3 - Q_1$ ). Os desvios interquartílicos valem 2,7 cm e 4,5 cm, para luz direta e indireta, respectivamente, indicando que o crescimento de pseudobulbos é mais disperso para luz indireta. Observe, na legenda da figura, que o desvio interquartílico contém 50% das observações, na faixa de 25% ( $Q_1$ ) a 75% ( $Q_3$ ) e é representado graficamente por uma "caixa". Os comprimentos das caudas são dados pelas linhas contínuas que vão da "caixa" (retângulo) aos valores mais afastados que não sejam *outliers*. Observe na Tabela 2.38 que os valores superiores mais afastados, que não são *outliers*, são iguais a 8,8 e 12,7, para luz direta e indireta,

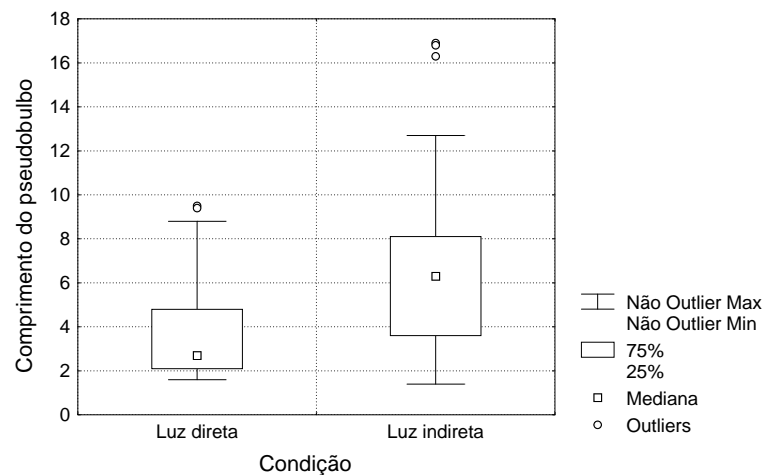


Figura 2.36: Desenho esquemático para comprimento de pseudobulbos de *Laelia purpurata* para luz direta e indireta.

respectivamente. Os valores inferiores mais afastados, que não são *outliers*, são iguais a 1,6 e 1,4, para luz direta e indireta, respectivamente. Os *outliers* estão representados por pequenos círculos vazios e observa-se na Figura 2.36 a presença dos mesmos em ambos os ambientes. Não se observa valores discrepantes para os valores inferiores das distribuições.

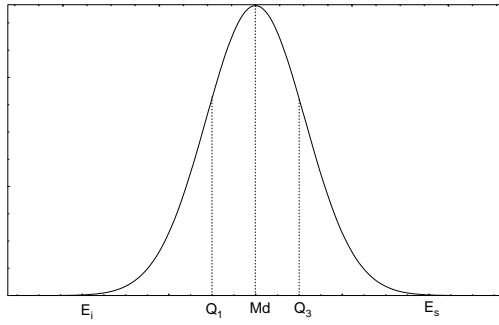
Nas Figuras 2.37 e 2.38 temos o comportamento dos quartis e mediana quanto à dispersão e à assimetria. As posições relativas dos valores  $Q_1$ ,  $Q_3$  e  $Med$  dão uma idéia da assimetria e dispersão da distribuição. No nosso exemplo, a distribuição com luz direta é mais assimétrica e menos dispersa do que com luz indireta.

Observe na Figura 2.39 como fica o desenho esquemático para uma distribuição normal. Podemos ver claramente que as amostras para luz direta e indireta não são originárias de populações com distribuição normal.

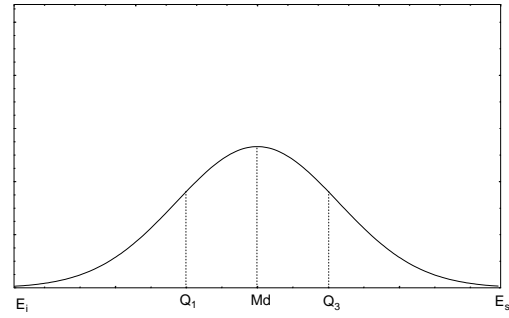
### Exercícios resolvidos

1. Construir o desenho esquemático para a variável rendimento médio de grãos, em kg/ha, da Tabela 2.1. Os valores da mediana e dos quartis foram calculados considerando os dados individualmente, isto é, não são agrupados em classes, e aparecem na legenda da Figura 2.40. Encontre estes valores. Este desenho esquemático foi feito no programa Statistica. Obtenha informações relevantes sobre a variável em estudo. Você saberia fazer esta figura na mão? O rendimento médio é 4986,5 kg/ha. O desvio interquartilístico vale 493,5 kg/ha. São observados 4 valores discrepantes, um híbrido com produção aquém dos demais e, 3 híbridos com produções acima dos demais. A distribuição é assimétrica à direita, portanto, não segue uma distribuição normal.

*Extremos-e-quartis e a dispersão de uma distribuição*



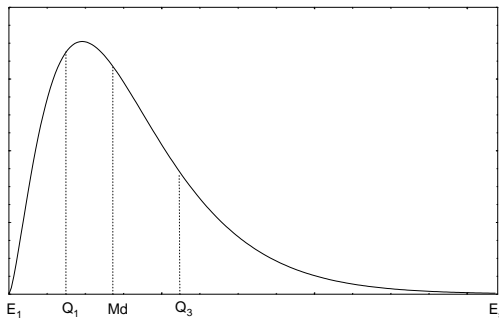
(a) Menor dispersão do que 2.37(b)



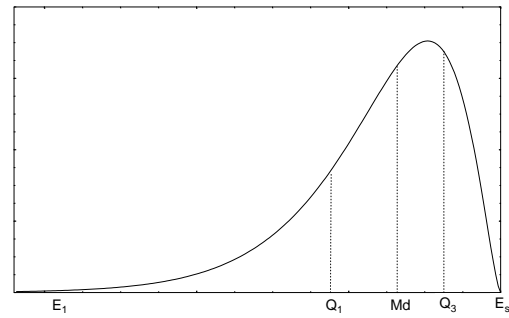
(b) Maior dispersão do que 2.37(a)

Figura 2.37: Em distribuições dispersas os valores dos quartis e extremos ficam mais afastados da mediana

*Extremos-e-quartis e a assimetria de uma distribuição*



(a) Assimetria à direita



(b) Assimetria à esquerda

Figura 2.38: Em distribuições assimétricas a distância entre a Md e  $Q_1$  ou  $E_i$  é diferente da distância entre Md e  $Q_3$  ou  $E_s$

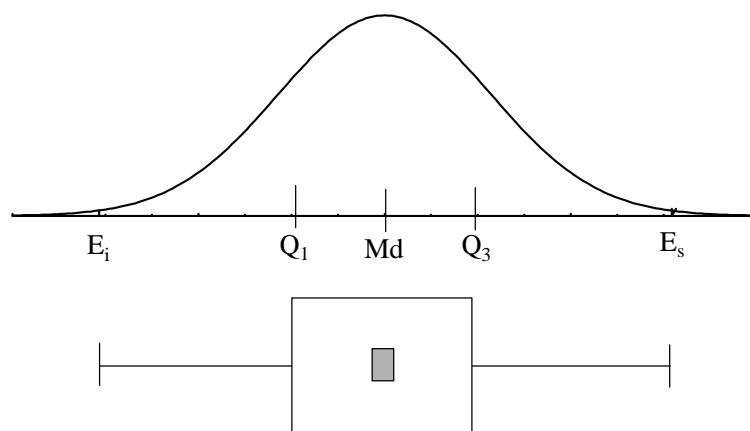


Figura 2.39: Desenho esquemático para uma distribuição normal.

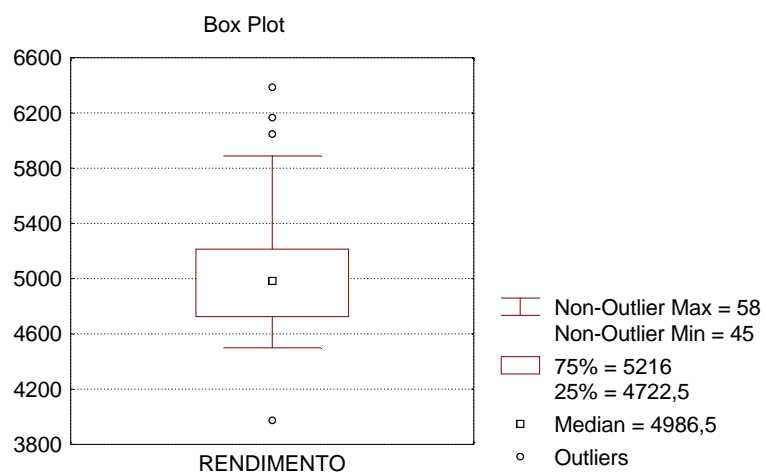


Figura 2.40: Desenho esquemático para rendimento de grãos.

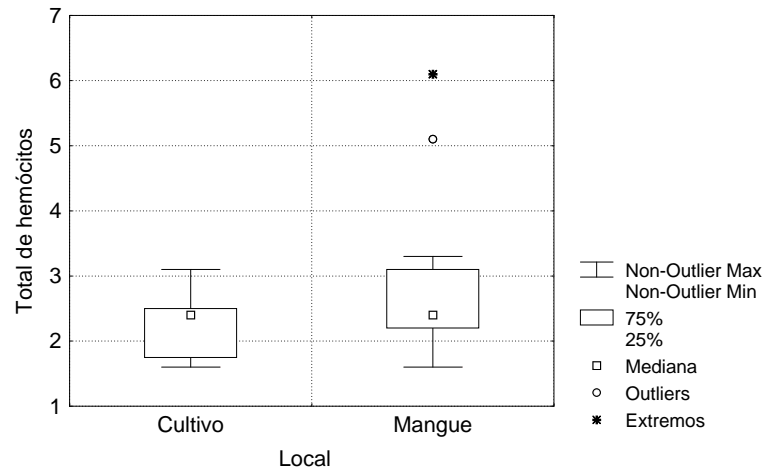


Figura 2.41: Desenho esquemático para número total de hemócitos em dois locais.

Tabela 2.40: Produção de cana-de-açúcar em toneladas por hectare

Variedade 1	65	68	75	76	77
	78	80	80	82	86
Variedade 2	88	89	90	91	92
	93	95	96	97	97
	99				

2. Na Figura 2.41 representa-se duas distribuições de freqüências, uma para local Cultivo e outra para Mangue. Os resultados referem-se a variável número total de hemócitos no bivalve *Crossostrea rhizophorae* coletados em ambiente de cultivo e mangue. Descreva e compare as principais informações (valor típico, simetria, dispersão, outliers) das duas distribuições.

O número total mediano de hemócitos nos dois locais é praticamente o mesmo e, em torno de  $2,5 \cdot 10^{-3}/mm^3$ . As duas distribuições são assimétricas. No mangue, a distribuição apresenta valores discrepantes.

#### 2.2.4.1 Exercícios Propostos

1. Para se estudar o comportamento de duas variedades de cana-de-açúcar, realizou-se um experimento onde foram obtidos os resultados indicados na Tabela 2.40.

- Calcular a  $Md$ ,  $Q_1$ ,  $Q_3$ ,  $Q_3 - Q_1$ ,  $Q_1 - 1,5(Q_3 - Q_1)$  e  $Q_3 + 1,5(Q_3 - Q_1)$  para cada uma das variedades.
- Faça o desenho esquemático múltiplo para os dados das variedades 1 e 2.
- Compare os dois conjuntos de dados através do desenho do item b).

Tabela 2.41: Tamanho de pecíolos de *Hydrocotille sp*

Área úmida				Área seca			
13,8	15,6	16,1	16,6	7,3	8,4	9,0	10,4
14,3	15,8	16,3	16,8	7,6	8,4	9,0	10,4
14,5	15,8	16,3	16,8	7,8	8,4	9,3	10,9
15,0	15,8	16,3	16,9	7,8	8,6	9,3	10,9
15,0	15,8	16,3	17,0	8,0	8,6	9,3	11,7
15,5	16,0	16,5	17,0	8,2	8,6	9,6	11,7
15,5	16,0	16,5	17,2	8,2	8,6	9,6	12,0
15,5	16,0	16,6	17,4	8,3	9,0	9,8	
15,6	16,1	16,6		8,3	9,0	9,8	

2. Para se estudar o comportamento de uma planta típica de dunas, a *Hydrocotille sp*, quanto ao seu desenvolvimento, mediu-se o tamanho do pecíolo (cm), em duas áreas: seca e úmida. Selecionou-se de cada uma dessas áreas, amostras aleatórias de plantas e mediu-se o tamanho dos pecíolos. Os dados são dados na Tabela 2.41.

a) Calcular a  $Md$ ,  $Q_1$ ,  $Q_3$ ,  $Q_3 - Q_1$ ,  $Q_1 - 1,5(Q_3 - Q_1)$  e  $Q_3 + 1,5(Q_3 - Q_1)$  para cada uma das variedades.

b) Faça o desenho esquemático múltiplo para os dados das áreas seca e úmida.

c) Compare os dois conjuntos de dados, quanto as principais características, através do desenho do item b).

3. Verifique, construindo os desenhos esquemáticos para os dados da Tabela 2.36, se existem valores discrepantes. Obs. utilize os cálculos já realizados neste exercício.

### 2.2.5 Coeficiente de Correlação

É comum na prática o interesse em se analisar o comportamento conjunto de duas ou mais variáveis quantitativas. Nessa seção trataremos do estudo de correlação entre duas variáveis quantitativas, porém, a generalização para mais de duas variáveis é possível. Vamos pensar que estamos interessados em obter uma medida estatística que indique se existe ou não relação entre duas variáveis, qual a magnitude e o sinal dessa relação. Muitas vezes estamos interessados em obter uma medida estatística que indique a força da *relação linear* entre duas variáveis quantitativas  $X$  e  $Y$ . Por exemplo, queremos verificar se existe correlação

Tabela 2.42: Dados de produção de matéria seca e radiação fotossintética ativa

Produção	10	60	110	160	220	280	340	400	460	520
Radiação	18	55	190	300	410	460	570	770	815	965

entre doses de nitrogênio e a produção de milho ou, então, estudar o relacionamento entre a porcentagem de nucleotídeos totais e a temperatura em graus centígrados. Uma medida de correlação linear muito utilizada na prática é o *coeficiente de correlação de Pearson*.

### *Diagrama de Dispersão*

Um primeiro passo para verificar se existe correlação entre duas variáveis quantitativas é construir um gráfico de dispersão, que nada mais é do que a representação gráfica dos pares de valores num sistema cartesiano. No eixo das abscissas colocamos a variável  $X$  e no eixo das ordenadas a variável  $Y$ .

**Exemplo.** Considere os dados referentes a produção de matéria seca de uma planta ( $Y$ ) e a quantidade de radiação fotossintética ativa ( $X$ ). Os dados obtidos experimentalmente são apresentados na Tabela 2.42.

O diagrama de dispersão é mostrado na Figura 2.42. Através da pura observação da distribuição dos pontos, concluímos que existe uma dependência entre as variáveis produção ( $Y$ ) e radiação ( $X$ ), porque no conjunto das observações, a medida que aumenta a radiação fotossintética, também aumenta a produção de matéria seca. Em termos práticos, isso significa que conhecendo-se a quantidade de radiação ajuda a prever a produção de matéria seca.

**Exemplo.** Deseja-se saber se existe correlação entre o espaçamento das linhas na cultura da soja ( $X$ ) e a fração da radiação solar extinta pela planta ( $Y$ ). Para atender a esse objetivo foram coletados pares de valores das duas variáveis. Os resultados obtidos foram:

Radiação	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0	1,1
Espaçamento	0,53	0,51	0,48	0,45	0,44	0,41	0,40	0,39	0,36	0,30

Observando-se o gráfico 2.43 verificamos uma dependência inversa entre espaçamento e radiação, ou seja, aumentando-se o espaçamento entre linhas ocorre uma diminuição na fração da radiação solar extinta pela planta.

**Exemplo.** Uma pesquisadora está interessada em verificar se existe dependência entre a salinidade (g/l) e a temperatura para a região III da Lagoa da Conceição. Os dados estão descritos na Tabela 2.43.



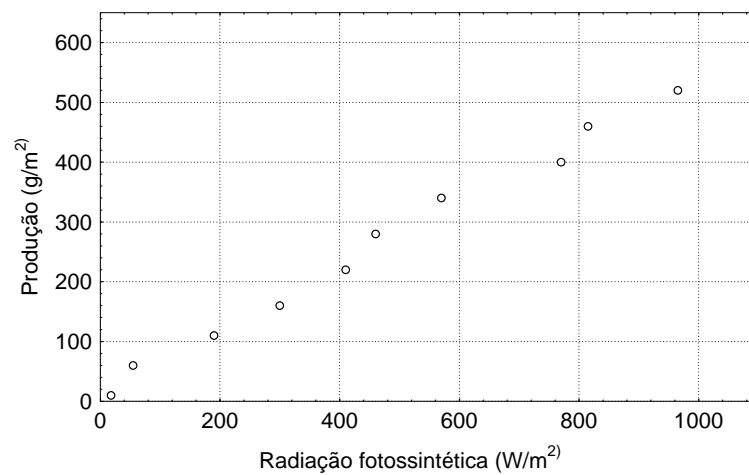


Figura 2.42: Diagrama de dispersão das variáveis produção e radiação fotossintética

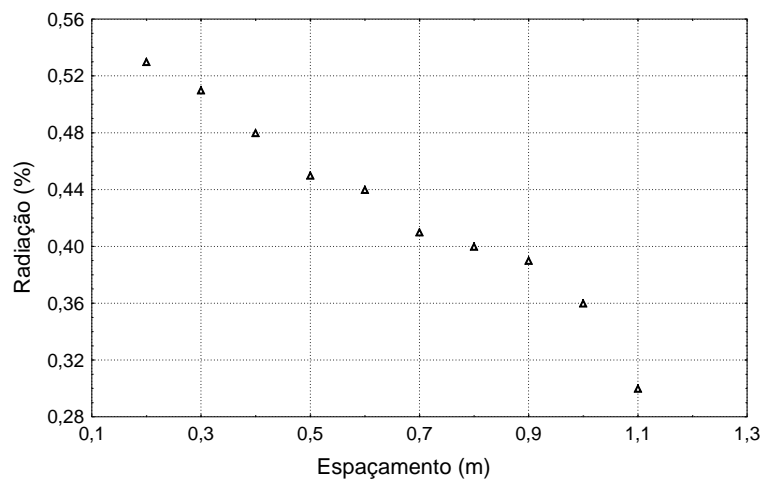


Figura 2.43: Diagrama de dispersão das variáveis radiação e espaçamento

Tabela 2.43: Valores de salinidade (g/l) e temperatura para a região III da Lagoa da Conceição

Estação	Temperatura ( $Y$ )	Salinidade ( $X$ )
23	24,0	3,85
23A	23,0	9,61
24	23,0	2,26
25	26,0	2,06
26	25,5	2,89
27	25,0	9,61
27A	24,3	10,58
28	23,0	11,40

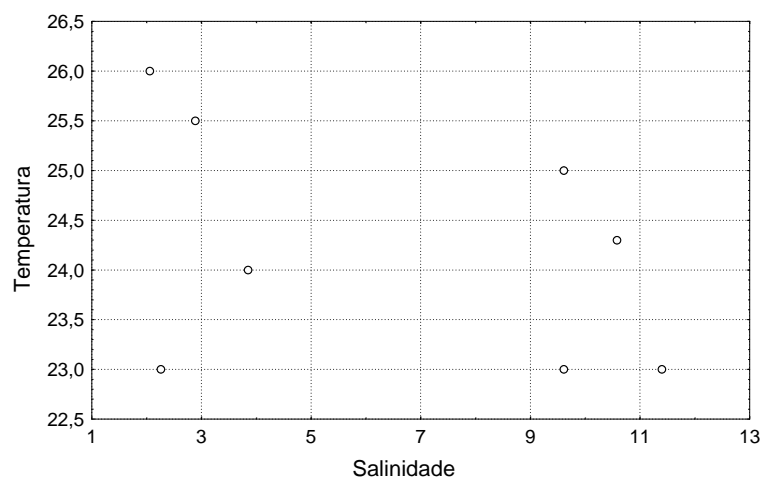


Figura 2.44: Diagrama de dispersão das variáveis salinidade e temperatura

O diagrama de dispersão é mostrado na Figura 2.44. Olhando-se para o gráfico 2.44, não observamos um relacionamento linear entre a salinidade e a temperatura, ou seja, os resultados não indicam que aumentando-se a salinidade existe uma diminuição ou um aumento da temperatura.

### *O Coeficiente de Correlação*

O coeficiente de correlação, representado pela letra  $r$ , é utilizado para quantificar a associação entre duas variáveis quantitativas. Indica o quanto a nuvem de pontos aproxima-se de uma reta. Salientamos que duas variáveis podem apresentar uma associação, por exemplo, curvilínea, ao invés de linear. Neste curso trataremos somente de associação linear

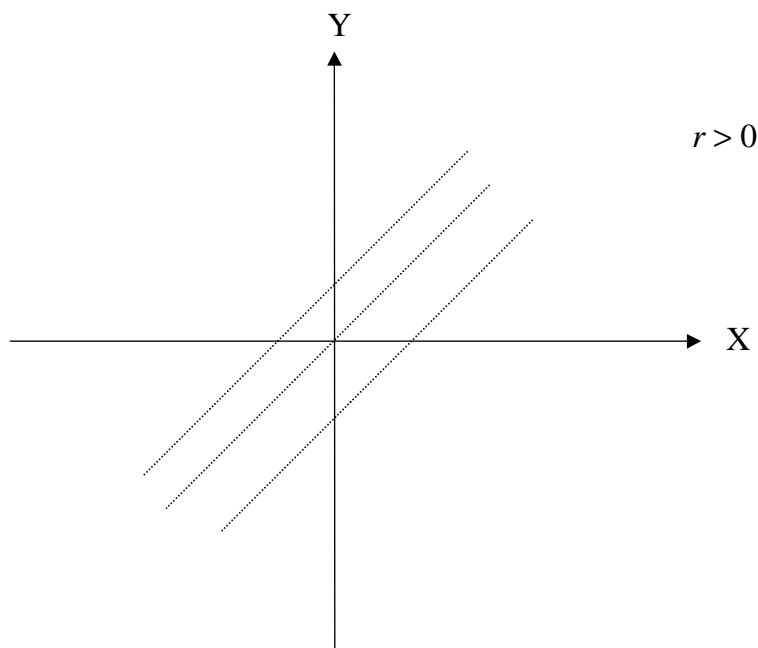


Figura 2.45: Correlação positiva

entre duas variáveis. O coeficiente de correlação,  $(r)$ , é uma medida estatística que assume valores entre

$$-1 \leq r \leq 1,$$

onde  $r = -1$  indica uma relação linear inversa perfeita e,  $r = 1$  indica uma relação positiva perfeita entre as variáveis. Quando o valor de  $r$  for igual a zero, significa que não existe associação entre as variáveis.

Considere a Figura 2.45. Aqui, através de uma transformação, a origem foi transportada para o centro da nuvem de pontos, então, a origem do sistema é dado pelo par  $(\bar{x}, \bar{y})$ . Tecnicamente foi feita uma translocação de eixos. Notamos que a grande maioria dos pontos estão situados no 1º e 3º quadrantes. Nesses quadrantes as coordenadas têm o mesmo sinal e, portanto, o produto será sempre positivo. Assim, se para cada ponto fizermos o produto de suas coordenadas e somarmos esses produtos, o resultado será um número positivo, pois existem mais produtos positivos do que negativos. Portanto,  $r > 0$ . Isto significa uma associação positiva entre as variáveis.

Para a dispersão da Figura 2.46 a soma dos produtos das coordenadas é na maioria negativa. Isto acarreta um  $r < 0$ , portanto, caracterizando uma relação negativa (uma cresce enquanto a outra decresce) entre as variáveis .

Para a dispersão da Figura 2.47 a soma dos produtos das coordenadas será próxima de zero. Portanto,  $r \cong 0$ , indicando que não existe associação entre as variáveis.

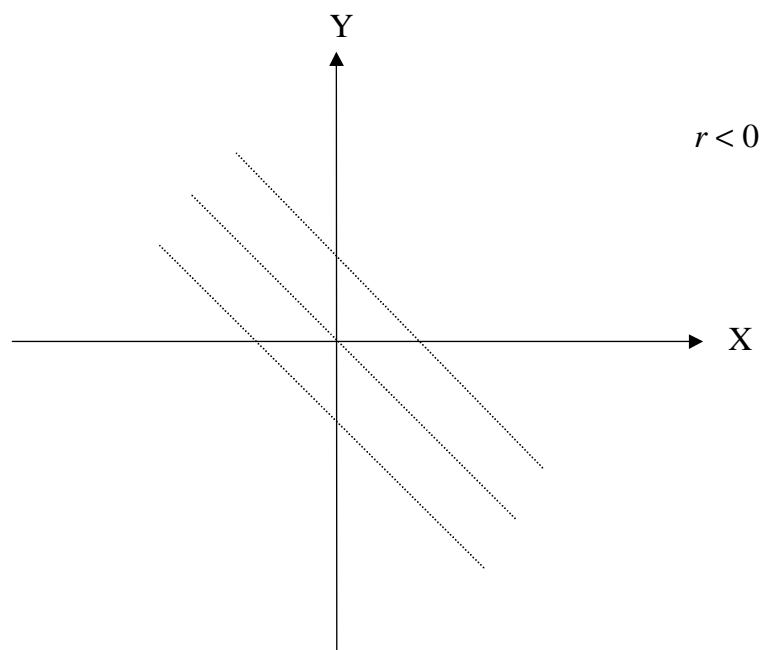


Figura 2.46: Correlação negativa

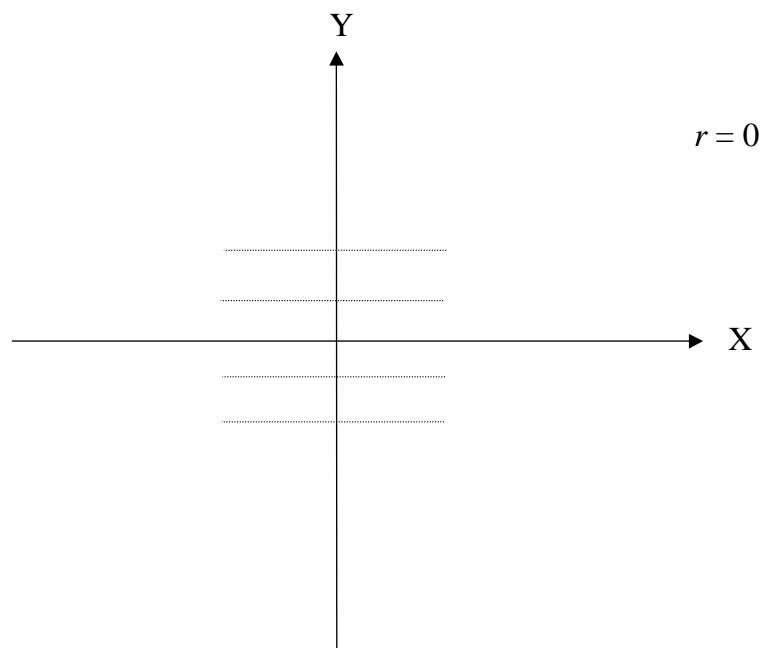


Figura 2.47: Correlação nula

Tabela 2.44: Etapas intermediárias para o cálculo do coeficiente de correlação

Observação	x	y	$(x - \bar{x})$	$(y - \bar{y})$	$Z_x = (x - \bar{x})/\sigma_x$	$Z_y = (y - \bar{y})/\sigma_y$	$Z_x Z_y$
1	18	10	-437,3	-246,0	-1,419	-1,496	2,123
2	55	60	-400,3	-196,0	-1,299	-1,192	1,548
3	190	110	-265,3	-146,0	-0,861	-0,888	0,764
4	300	160	-155,3	-96,0	-0,504	-0,584	0,294
5	410	220	-45,3	-36,0	-0,147	-0,219	0,032
6	460	280	4,7	24,0	0,015	0,146	0,002
7	570	340	114,7	84,0	0,372	0,511	0,190
8	770	400	314,7	144,0	1,021	0,876	0,894
9	815	460	359,7	204,0	1,167	1,240	1,448
10	965	520	509,7	264,0	1,654	1,605	2,656

**Exemplo.** Considere os dados referentes à produção de matéria seca de uma planta ( $Y$ ) e a quantidade de radiação fotossintética ativa ( $X$ ), mostrados na Tabela 2.42 e, que, desejamos saber se existe associação entre a produção e a radiação e qual a grandeza desta associação. Pela Figura 2.42 percebe-se um relacionamento linear entre elas, agora, desejamos quantificá-lo através do cálculo do coeficiente de correlação, o qual passamos a fazer passo a passo. Inicialmente, precisamos fazer a mudança de origem do sistema para o ponto  $(\bar{x}, \bar{y})$ , onde  $\bar{x} = 455,30$  e  $\bar{y} = 256,00$ . Os resultados encontram-se nas colunas 4 e 5 da Tabela 2.44. Existe, ainda, um problema quanto à escala usada. A variável  $X$ , é dada em  $g/m^2$ , e  $Y$ , é dada em  $W/m^2$ , como  $X$  tem maior dispersão, o produto ficará mais afetado pelos resultados de  $X$  do que de  $Y$ . Para corrigir isso, podemos reduzir as duas variáveis para uma mesma escala, isso é obtido dividindo-se os desvios pelos respectivos desvios padrões das variáveis, onde:  $\sigma_X = 308,1263$  e  $\sigma_Y = 164,4506$ . Observe que são os desvios padrões populacionais. Esses novos valores estão nas colunas 6 e 7 da Tabela 2.44. Na última coluna da tabela, indica-se os produtos das variáveis padronizadas; a soma dos mesmos, cujo valor é igual a 9,953 que, como esperávamos, é positiva. Para completar o cálculo do coeficiente de correlação basta dividir 9,953 pelo número de pares de valores (média), cujo valor é igual a:

$$\text{Correlação}(X,Y) = \frac{9,953}{10} = 0,9953.$$

Portanto, para esse exemplo, o grau de associação linear está quantificada em 0,9953. Indicando uma relação linear muito forte positiva entre as duas variáveis.

Tabela 2.45: Etapas intermediárias para o cálculo do coeficiente de correlação (fórmula operacional)

Observação	$x$	$y$	$x^2$	$y^2$	$xy$
1	18	10	324	100	180
2	55	60	3025	3600	3300
3	190	110	36100	12100	20900
4	300	160	90000	25600	48000
5	410	220	168100	48400	90200
6	460	280	211600	78400	128800
7	570	340	324900	115600	193800
8	770	400	592900	160000	308000
9	815	460	664225	211600	374900
10	965	520	931225	270400	501800
Totais	4553	2560	3022399	925800	1669880

Uma fórmula prática para o cálculo do coeficiente de correlação é dada por:

$$Corr(X, Y) = \frac{n(\sum xy) - (\sum x \sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}. \quad (2.14)$$

**Exemplo.** Vamos considerar os dados de produção de matéria seca e a quantidade de radiação fotossintética para mostrar os cálculos da estatística 2.14. Os cálculos intermediários estão desenvolvidos na Tabela 2.45. O valor do coeficiente de correlação por 2.14 é:

$$Corr(X, Y) = \frac{10(1669880) - (4553)(2560)}{\sqrt{10(3022399) - 4553^2} \sqrt{10(925800) - 2560^2}} = \frac{5043120}{5067155,33} = 0,9953.$$

### 2.2.5.1 Exercícios propostos

1. A Tabela 2.46 apresenta os valores de condutividade (mho) e salinidade (g/l) para a região III da Lagoa da Conceição.

- Construir o gráfico de dispersão. Conclua sobre a dependência entre as variáveis  $X$  e  $Y$ .
- Quantifique a dependência entre  $X$  e  $Y$  através do coeficiente de correlação. Existe uma dependência forte, média ou fraca?

2. Um estudo de vida de prateleira do café torrado e moído foi realizado. Os testes sensoriais foram iniciados a partir do 9º dia de estocagem e depois a intervalos de mais ou menos 7 dias. Em cada época de avaliação sensorial três amostras (pacotes) foram obtidas ao acaso. Seis provadores treinados avaliaram as três amostras simultaneamente, julgando

Tabela 2.46: Valores de condutividade (mho) e salinidade (g/l) para a região III da Lagoa da Conceição

Estação	Condutividade ( $Y$ )	Salinidade ( $X$ )
23	19,92	3,85
24	11,78	2,26
25	14,11	2,06
26	16,10	2,89
27	36,52	9,61
28	51,46	11,40

Tabela 2.47: Notas médias de aroma de café torrado e moído

Sessão de avaliação	Tempo de estocagem (dias)	Resultado médio da equipe de provadores para cada amostra		
	$x$	$y_1$	$y_2$	$y_3$
1	9	4,8	4,7	4,7
2	14	4,0	4,7	4,8
3	22	3,7	3,7	3,5
4	29	3,2	3,5	3,2
5	36	3,7	3,0	3,3
6	43	2,5	2,8	2,7

o produto quanto ao aroma em uma escala descritiva de 1 a 6 pontos: 6 = excelente; 5 = bom; 4 = aceitável; 3 = pouco aceitável; 2 = inaceitável e 1 = não bebível. Os resultados obtidos são dados na Tabela 2.47.

- faça o diagrama de dispersão entre as variáveis tempo e média de aroma;
- calcule o coeficiente de correlação entre as duas variáveis e faça uma conclusão.

### 2.2.6 Ajuste da Equação de uma Reta

O coeficiente de correlação dá um número que resume o grau de relacionamento linear entre duas variáveis, por exemplo,  $r = 0,90$ ; o ajuste de um modelo linear simples tem como resultado uma equação matemática que descreve esse relacionamento. A partir do momento que temos um modelo ajustado, o conjunto de dados passa a ser representado através deste. Vamos procurar estabelecer uma equação matemática linear, isto é, ajustar a equação de

uma reta para descrever o relacionamento entre duas variáveis. Olhando-se para a Figura 2.42, observa-se que existe uma relação linear positiva entre as variáveis produção ( $Y$ ) e radiação ( $X$ ), portanto, podemos ajustar a equação de uma reta aos dados. As variáveis  $Y$  e  $X$  são também denominadas de variáveis *dependentes* e *independentes*, respectivamente.

Utilização:

1. Temos duas variáveis que medem aproximadamente a mesma coisa, mas uma delas é relativamente dispendiosa, ou difícil de lidar, enquanto que a outra não. Por exemplo, a resistência e a dureza de um material podem estar relacionadas, de modo que conhecendo-se a dureza podemos estimar a resistência. Se o teste de resistência destroi o material, enquanto que o teste de dureza não o destroi, uma pessoa interessada em estimar a resistência, obviamente preferirá confiar nos resultados do teste de dureza para estimar a resistência. A finalidade de uma equação de regressão seria então *estimar* valores de uma variável, com base em valores conhecidos da outra. Esse estudo também é conhecido como calibração da equação de regressão.
2. Outra aplicação das equações de regressão é explicar valores de uma variável em termos da outra. Isto é, podemos suspeitar de uma relação de *causa e efeito* entre duas variáveis. Por exemplo, quantidade de fertilizante e produção de trigo.
3. Predizer valores *futuros* de uma variável. Por exemplo, predizer produções futuras.

## A Equação da Reta

Uma reta é dada pela equação matemática:

$$y_i = \alpha + \beta x_i$$

onde o parâmetro  $\alpha$  representa o ponto onde a reta corta o eixo dos  $y$  e o parâmetro  $\beta$  é o coeficiente angular, ou ainda, o quanto varia a média de  $y$  para o aumento de uma unidade da variável  $x$ . Esses parâmetros estão representados na Figura 2.48. Por exemplo, a reta estimada por,

$$\hat{y}_i = 5 + 3x_i$$

intercepta o eixo das ordenadas no ponto em que  $y = 5$ ; o coeficiente angular da reta vale 3, o que significa que a cada unidade de variação de  $x$ , correspondem 3 unidades de variação de  $y$ . Dando alguns valores para  $x$ , podemos resolver a equação para obter os valores de  $\hat{y}$



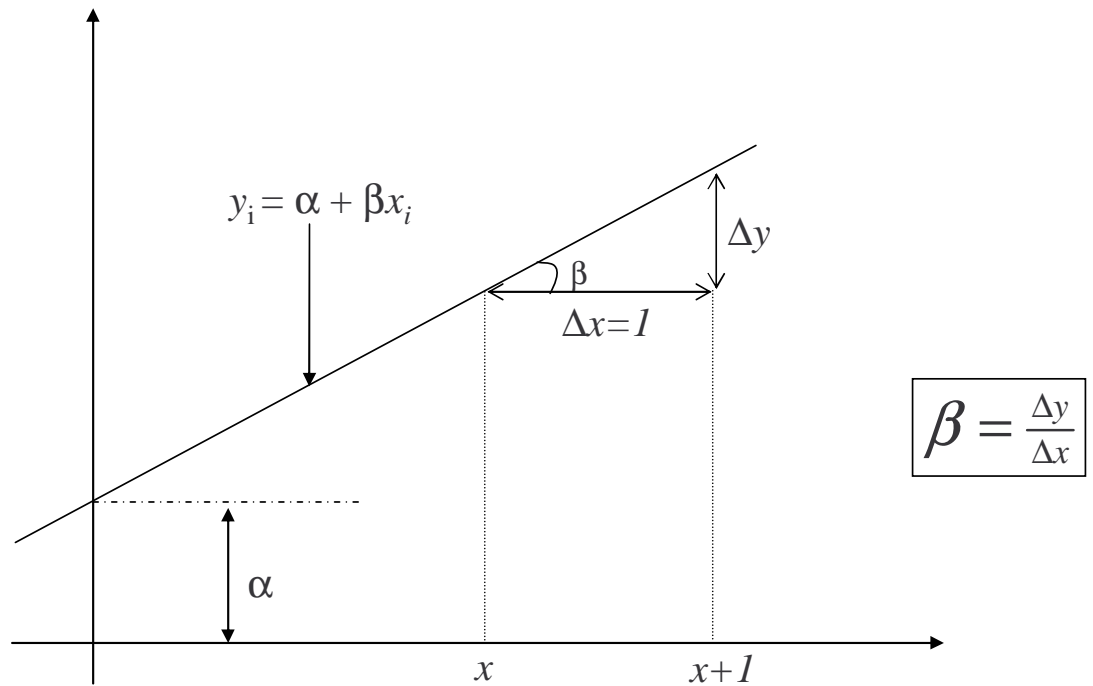


Figura 2.48: Interpretação dos parâmetros da equação de uma reta

(leia-se  $\hat{y}$  chapéu e representa os valores estimados da variável  $Y$ ).

Valores de $x_i$	$\hat{y} = 5 + 3x_i$
2,0	11
3,0	14
5,0	20

Devemos, através de um método adequado, estimar os parâmetros  $\alpha$  e  $\beta$ .

### *Estimativas dos Parâmetros $\alpha$ e $\beta$*

A idéia básica na estimativa da parte funcional do modelo,  $\alpha + \beta x_i$ , é encontrar a reta que passa o mais próximo possível de todos os pontos observados. Representaremos esta reta por:

$$\hat{y} = a + bx$$

e a denominaremos de reta estimada pelos dados. Um critério conhecido como método dos mínimos quadrados, fornece as seguintes expressões para as estimativas dos parâmetros da equação:

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Tabela 2.48: Cálculos intermediários para a estimação dos parâmetros

Dados		Cálculos intermediários	
$x$	$y$	$xy$	$x^2$
18	10	180	324
55	60	3300	3025
190	110	20900	36100
300	160	48000	90000
410	220	90200	168100
460	280	128800	211600
570	340	193800	324900
770	400	308000	592900
815	460	374900	664225
965	520	501800	931225
4553	2560	1669880	3022399

$$a = \frac{\sum y - b \sum x}{n}$$

onde  $n$  é o número de pares  $(x, y)$  observados (tamanho da amostra).

**Exemplo.** Ilustraremos a obtenção da equação da reta com as observações de produção ( $Y$ ) e radiação ( $X$ ) (descritos anteriormente). A Tabela 2.48 apresenta os dados originais e os cálculos intermediários para se encontrar as estimativas dos parâmetros. Encontramos:

$$b = \frac{10(1669880) - (4553)(2560)}{10(3022399) - (4553)^2} = 0,5312;$$

$$a = \frac{2560 - 0,5312(4553)}{10} = 14,1537.$$

A equação da reta estimada com base nos dados da amostra fica:

$$\hat{y}_i = 14,154 + 0,531x_i.$$

Os valores observados, juntamente com a equação estimada da reta estão na Figura 2.49. Como era esperado, observa-se um bom ajuste da equação da reta aos dados observados.

A distância de cada um dos valores observados até a equação da reta é chamada de *desvio* ou *resíduo* e são representados por  $e_i$ . O cálculo dos desvios é dado por:

$$e_i = y_i - \hat{y}_i.$$

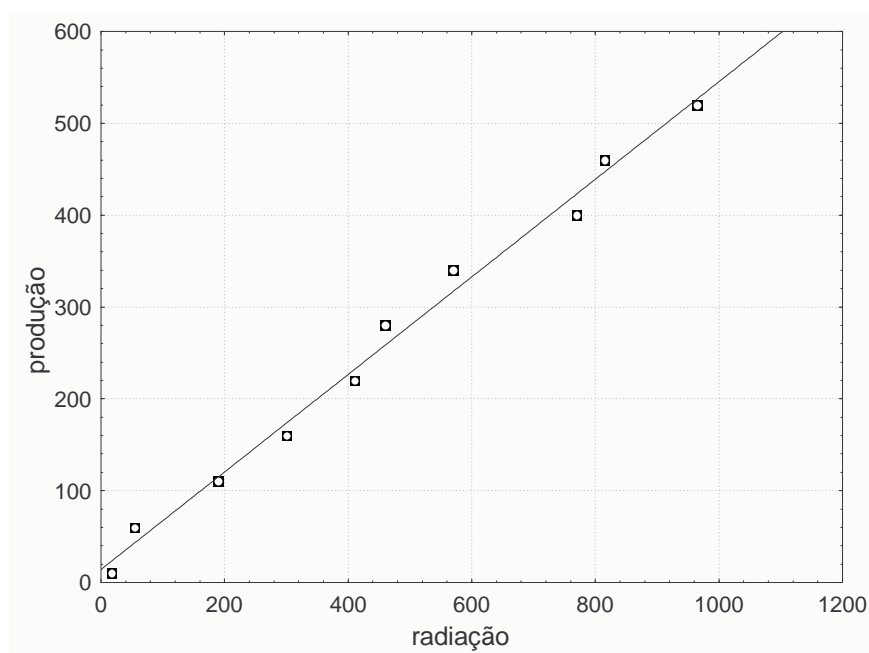


Figura 2.49: Reta ajustada aos dados observados

Por exemplo, o primeiro resíduo é dado por:

$$e_1 = 10 - (14,154 + 0,531 \times 18) = 10 - 23,71 = -13,71.$$

Para todos os dados de produção e radiação, dados na Tabela 2.48, os valores ajustados e os resíduos são dados por:

Valores estimados	Desvios ou resíduos
23,71	-13,71
43,37	16,63
115,08	-5,08
173,51	-13,51
231,94	-11,94
258,50	21,50
319,93	23,07
423,16	-23,16
447,07	12,93
526,74	-6,74

O gráfico dos valores ajustados, também denominados de valores estimados, versus os resíduos é dado na Figura 2.50. O padrão esperado é que os pontos estejam aleatoriamente

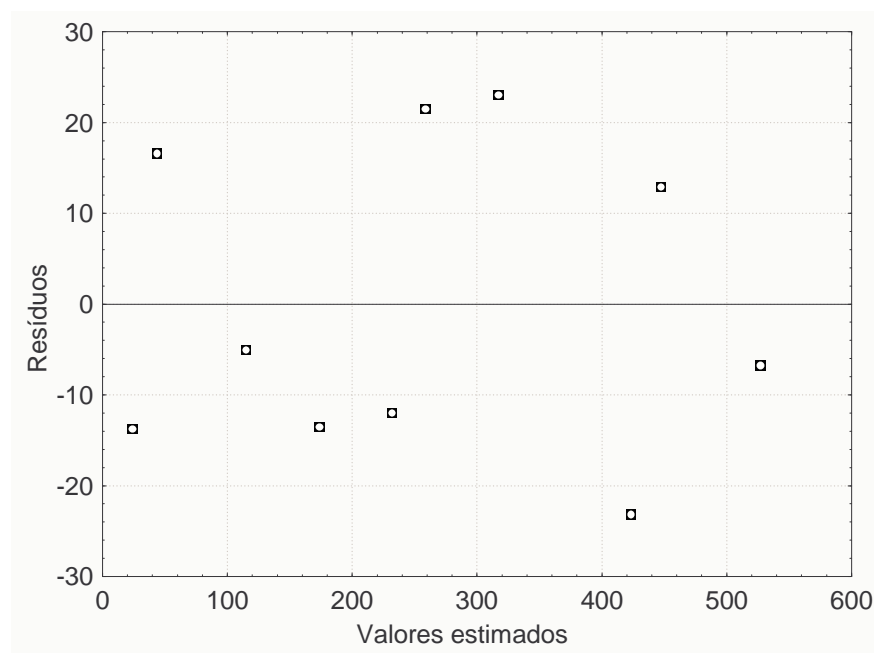


Figura 2.50: Gráfico dos valores ajustados versus resíduos

distribuídos em torno do valor zero, isto é, não existe nenhum padrão nos resultados. Se ocorrer este padrão, dizemos que o modelo é adequado para representar os nossos dados. Observamos na Figura 2.50 que os pontos encontram-se distribuídos ao acaso em torno do valor zero, portanto, o modelo linear simples (equação da reta), pode ser utilizado para os dados amostrados. Num outro caso, se o padrão de distribuição fosse curvilíneo, um polinômio de 2º grau seria o indicado. Assim, podemos querer estimar (avaliar) a produção para um valor de radiação igual a 100. Este valor é obtido fazendo-se:

$$\hat{y}_i = 14,154 + 0,531 \times 100 = 67,254.$$

### 2.2.6.1 Exercícios propostos

1. a) Obtenha a equação da reta para os dados da Tabela 2.46, isto é, para condutividade ( $Y$ ) e salinidade ( $X$ ).
- b) De acordo com esta função, quais seriam os valores estimados ( $\hat{y}$ ) para os valores de  $x$  da Tabela 2.46.
- c) O que você acha da discrepância entre os valores observados e estimados ( $y - \hat{y}$ )?
- d) Encontre o valor estimado ( $\hat{y}$ ), para salinidade igual a 5,0.
2. a) Encontre a equação da reta para os dados da Tabela 2.47, ou seja, a equação do tempo de estocagem ( $Y$ ) sobre o aroma.

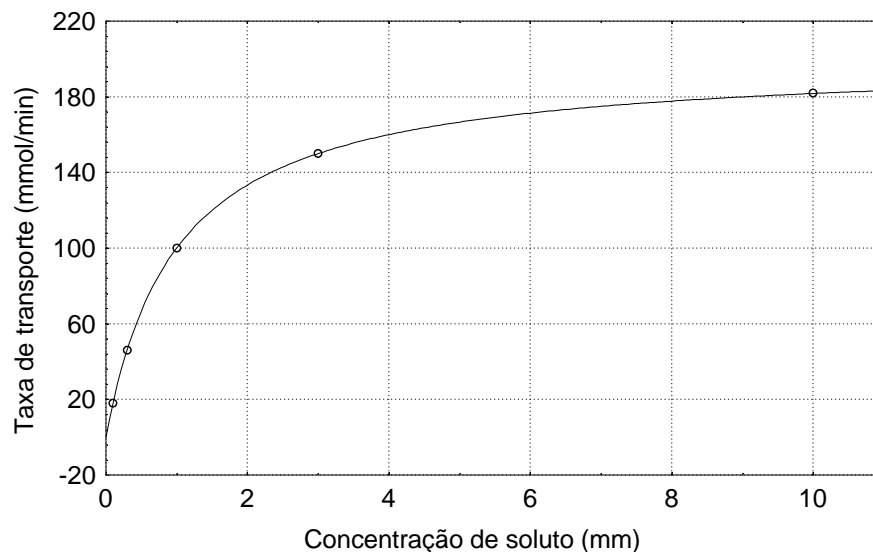


Figura 2.51: Cinética enzimática

b) Você considera que esta equação está explicando bastante da relação entre tempo de estocagem e aroma?

c) Faça o gráfico dos resíduos versus valores ajustados. Existe algum padrão na distribuição?

3. (Albert et.al. 1999, página 173 e 708). A curva mostrada na Figura 2.51 é descrita pela equação:

$$velocidade = V_{max}[S/(S + K_M)] \quad (2.15)$$

a qual é chamada de equação de Michaelis-Menten. A velocidade da reação enzimática ( $V$ ) aumenta à medida que a concentração do substrato ( $S$ ) é aumentada, até que um valor máximo ( $V_{max}$ ) é atingido. Para a maioria das enzimas, a concentração de substrato na qual a velocidade da reação é metade da máxima fornece uma medida direta da força de ligação do substrato à enzima ( $K_M$ ).

Para o transporte de acetado mediado por proteína carreadora, a relação entre concentração,  $S$ , e taxa de transporte,  $V$ , pode ser descrita pela equação 2.15, que descreve reações enzimáticas simples:

$$Taxa\ de\ transporte = V_{max}S/(S + K_M) \quad (2.16)$$

Desejamos, através do ajuste da equação de regressão, estimar os parâmetros  $V_{max}$  e  $K_M$ . Para facilitar a obtenção dos mesmos, fazemos uma transformação na equação 2.16, de tal forma que, os dados possam ser dispostos como uma linha reta. Fazendo-se a transformação inversa, temos:

$$1/taxa = (K_M/V_{max})(1/S) + 1/V_{max}. \quad (2.17)$$

Tabela 2.49: Valores de concentração (S) e taxa de transporte (V)

S	V	1/S	1/V
0,1	18	10,0	0,056
0,3	46	3,3	0,022
1,0	100	1,0	0,01
3,0	150	0,33	0,0067
10,0	182	0,1	0,0055

Portanto, temos a equação de uma reta:  $y = \alpha + \beta x$ , onde:  $y = 1/V$ ,  $x = 1/S$ ,  $\alpha = 1/V_{max}$  e  $\beta = K_M/V_{max}$ . Para os dados da Tabela 2.49, encontre as estimativas dos parâmetros  $V_{max}$  e  $K_M$ , ajustando a equação de uma reta. Conhecendo-se os valores de  $V_{max}$  e  $K_M$  você pode calcular as taxas de transporte para as concentrações de 0,5 mM e 100 mM de acetato usando a equação 2.16. Encontre estes valores.

## 3 Modelos de Probabilidades para Experimentos Simples

### 3.1 Introdução

O importante estudo de probabilidade é anterior aos anos de 1500, século XVI. Nessa época, as aplicações eram dirigidas aos jogos de azar. Pessoas se utilizavam do conhecimento da teoria das probabilidades para planejar estratégias de apostas. Conta-se que um senhor conhecido como "Cavaleiro de Meré", amigo do francês Pascal (1610), e grande apreciador de jogos, solicitava ao mesmo, que calculasse suas chances de ganho.

No entanto, somente no século XX, é que o cálculo de probabilidade teve um desenvolvimento bastante grande, e baseado numa teoria matemática através de axiomas (condições) rigorosos, definições e teoremas.

A probabilidade é um ramo da matemática cuja grande aplicação é na estatística, como veremos a seguir. Até o presente momento do nosso curso de estatística, estudamos, de forma empírica, isto é, sem uma justificativa científica (só descrevendo e não explicando), o comportamento dos fenômenos, eventos da natureza, através da construção das *distribuições de frequências*. Aqui, temos especial interesse em experiências aleatórias, casuais, ou seja, devemos realizar o experimento para saber qual será o resultado. Por exemplo, não é possível saber qual será a produção por hectare de uma linhagem "X" de feijão, se este não for plantado, colhido e mensurado. Outros exemplos de experimentos aleatórios: 1) germinação de sementes; 2) sobrevivência de enxertos; 3) número de micronúcleos/3000 células; 4) número de plantas numa determinada área.

Experimentos aleatórios: quando repete-se o experimento, em condições mais próximas possíveis, dão resultados geralmente diferentes.

Verificando-se os fenômenos aleatórios na natureza, construindo as suas distribuições de frequências (conhecimento empírico), e com o surgimento da teoria da probabilidade, foi possível matematizar a natureza, isto é, criar *modelos probabilísticos* (*distribuições de probabilidades*) que representam muito bem os fenômenos da natureza. Com isso, é possível, fazendo-se algumas suposições adequadas (hipóteses, conjecturas) e sem a necessidade de se observar diretamente o fenômeno, estabelecer distribuições de probabilidades que representam muito bem as distribuições de frequências, quando o fenômeno é observado diretamente.

Esses modelos probabilísticos são considerados hoje a espinha dorsal da estatística, pois, como veremos, todos os procedimentos inferenciais (amostra  $\rightarrow$  população), são aplicações de probabilidade. Vamos ilustrar a construção de modelos probabilísticos com dois exemplos.

**Exemplo 1:** de um grupo de duas mulheres (M) e três homens (H), uma pessoa será sorteada para presidir uma determinada reunião. Queremos estudar as probabilidades de que o presidente seja do sexo masculino ou feminino. Suposições:

- 1º) Só existem duas possibilidades.
- 2º) Cada pessoa tem a mesma chance de ser sorteada.

Tabela 3.1: Modelo de probabilidades para o experimento:

Sexo	M	H	Total
Frequência teórica	$\frac{2}{5}$	$\frac{3}{5}$	1

**Exemplo 2:** o objetivo é estudar a distribuição das proporções no lançamento de um dado. Vamos construir esta distribuição através de suposições teóricas;

- 1º) só podem ocorrer seis faces e,
- 2º) considerando o dado como sendo perfeito, cada face deve ocorrer o mesmo número de vezes e, portanto, a proporção de ocorrência de cada face é de  $\frac{1}{6}$ .

Assim, o modelo teórico de frequências para esse experimento é:

Faces do dado	1	2	3	4	5	6	Total
Frequências	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

A primeira grande aplicação de probabilidade nas biociências, foi com as leis da hereditariedade (Mendel, 1822-1884)<sup>3</sup>. Outras aplicações podem ser: chance de sobrevivência (germina ou não germina), distribuição e interação de espécies animais ou vegetais, risco de doenças, ocorrência de mutações, previsão de safras, na previsão de chuvas, geadas, granizos, etc.

Uma aplicação, particularmente importante para o nosso curso, é quando um pesquisador conduz um experimento. Quando um experimento é instalado, desejamos comparar os efeitos dos tratamentos (variações de um fator a ser estudado). Para se estimar os efeitos dos

---

<sup>3</sup>Gregor Johann Mendel, botânico da Morávia.



tratamentos e também para executar os testes estatísticos é necessário o uso de repetições (aplicação do mesmo tratamento em diversas unidades experimentais e que formará a amostra de estudo), através das quais vamos ter a possibilidade de calcular a variabilidade dos dados, ou seja, a variância. O pesquisador esforça-se ao máximo para manter todas as condições experimentais o mais constantes possíveis, só deixando variar os tratamentos. Por exemplo, num experimento agrícola de campo, vamos aplicar um inseticida, herbicida ou fungicida de forma homogênea em todo o experimento; a fertilidade do solo deve ser igual para todos os tratamentos. A temperatura, a umidade, a luminosidade, devem ser mantidas constante dentro de um laboratório. A temperatura deve ser constante em todos os pontos dentro de uma estufa. E assim por diante. Pois bem, por maior que seja esse esforço para manter todas as condições experimentais o mais constantes possíveis, os valores obtidos nas repetições dos tratamentos dificilmente resultam em valores iguais, é quase com certeza absoluta que todos os valores vão diferir uns dos outros. Pode-se concluir que sempre vão existir variações. Logo, um pesquisador não pode afirmar com 100% de certeza que o tratamento A é melhor que o B, e assim por diante. Portanto, todas as conclusões tomadas, ou as inferências realizadas, são feitas com certo grau de incerteza; essa incerteza nós expressamos em termos de *probabilidade*. Então, quando um pesquisador diz que o tratamento A é melhor que o B, ao nível de significância de 5%, ele admite que pode estar tomando uma decisão errônea e que a probabilidade de erro é de no máximo igual a 5%. Estes experimentos cujos resultados podem não ser os mesmos, de repetição para repetição, são chamados de *experimentos aleatórios, probabilísticos ou estocásticos*.

*"Todas as inferências realizadas tem certo grau de incerteza expressadas por probabilidades".*

A seguir apresentamos um exemplo de um experimento no delineamento inteiramente casualizado.

### **Avaliação do Comportamento "In Vitro" de abacaxi *Ananas comosus* cv. Primavera**

- Objetivo: determinação de um meio de cultura onde produza um maior número de mudas por explante (meristema) inoculado.

1 24 T3	7 20 T4	13 12 T3	19 26 T2	25 19 T5	31 18 T3
2 42 T1	8 22 T5	14 13 T6	20 11 T3	26 26 T1	32 24 T4
3 10 T5	9 20 T2	15 13 T4	21 9 T6	27 25 T5	33 14 T2
4 35 T1	10 14 T6	16 14 T2	22 21 T1	28 9 T5	34 23 T1
5 21 T2	11 18 T6	7 11 T4	23 15 T3	29 20 T6	35 15 T4
6 7 T6	12 14 T3	18 23 T1	24 19 T5	30 20 T4	36 21 T2

Figura 3.1: No canto superior esquerdo está indicado a unidade experimental; no centro à direita o tratamento e, no canto inferior esquerdo, está indicado o número de brotos por explante

Tratamentos:

---

Meio de cultura 1	(T1)
Meio de cultura 2	(T2)
Meio de cultura 3	(T3)
Meio de cultura 4	(T4)
Meio de cultura 5	(T5)
Meio de cultura 6	(T6)

---

- Unidade experimental: Um tubo de ensaio.
- Variáveis observadas: 1) número de brotos por explante inoculado; 2) tamanho dos brotos em cm; 3) número de folhas por broto, etc.
- Delineamento Experimental: delineamento inteiramente casualizado com 20 repetições. Obs: na Figura 3.1 apresentamos esse delineamento utilizando somente os resultados de 6 repetições. Condições experimentais controladas ao máximo: Temperatura de  $24^{\circ}C \pm 1^{\circ}C$  ; fotoperíodo de 16hs luz; intensidade luminosa de 4000 lux, meristemas uniformes (tamanho).

Independente de qual seja a aplicação, observa-se que quando utiliza-se do cálculo de probabilidade existe sempre um elemento do *acaso*, *aleatório* ou de *incerteza*, quanto a ocorrência ou não de um evento, fenômeno futuro. Assim sendo, parece improvável, em muitas situações , afirmar por antecipação, o que vai ocorrer, mas é perfeitamente possível quantificar quão provável é a ocorrência de determinado evento, fenômeno futuro.

*As probabilidades são utilizadas para exprimir a chance de ocorrência de determinado evento.*

### 3.2 Conceitos Fundamentais

Inicialmente, é interessante definirmos a notação a ser utilizada para representar algum fato, fenômeno. Existe uma relação entre a teoria dos conjuntos e a teoria das probabilidades, assim, será utilizada a notação e símbolos da teoria dos conjuntos. Os conjuntos são representados por letras maiúsculas e os seus elementos, são colocados entre chaves ( $A = \{a, e, i, o, u\}$ ).

Há duas maneiras pelas quais se pode descrever os elementos de um conjunto:

1<sup>o</sup>) Consiste em relacionar todos os elementos do conjunto, ou um número suficiente deles.

**Exemplo:** desejamos representar os híbridos de milho, de ciclo tardio, recomendados para a região de Chapecó, altitude menor que 800 m da safra 1988/89.

$A = \{C125, AG28, AG35, SAVE, 342-A, C408, AG401, C317\}$ .

Outro exemplo: principais espécies do gênero *Biomphalaria* em Santa Catarina.

$B = \{B.straminea; B.glabrata; B.tenagophila\}$ ;

2<sup>o</sup>) Consiste em formular uma regra que defina a(s) característica(s) comum(ns) aos membros do conjunto. **Exemplos:**

$A = \{\text{Todos os gêneros de formigas da família Formicidae}\}$ ;

$B = \{\text{Todas as espécies do gênero Biomphalaria}\}$ ;

$C = \{\text{Todas as plantas que produzem } O_2\}$ ;

$D = \{\text{Todas as espécies da família das Solanaceas}\}$ ;

$E = \{\text{Espécies do gênero Biomphalaria que são hospedeiras intermediárias do } Schistosoma mansoni\}$ ;

Essa notação será utilizada para representar *espaços amostrais* e *eventos*.

Inicialmente, dissemos que não é necessário verificar diretamente o fenômeno para entender o seu comportamento, a sua variabilidade, enfim, a sua distribuição de frequências e,

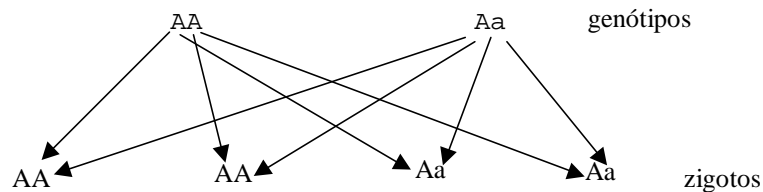


Figura 3.2: Resultados do cruzamento de genótipos

sim, que é possível, fazendo-se algumas suposições adequadas, *criar um modelo teórico* que represente muito bem essa distribuição, são os chamados *modelos probabilísticos*. Vamos ver a construção de mais um modelo de probabilidade.

**Exemplo:** Se cruzarmos indivíduos de genótipos  $AA$  e  $Aa$ . Queremos estudar as proporções dos resultados desse cruzamento. O gene  $A$  de um indivíduo  $AA$  encontra o gene  $A$  ou o gene  $a$  de um indivíduo  $Aa$ . As células fertilizadas tem genótipo  $AA$  e  $Aa$ , conforme Figura 3.2. Observamos:

1º) Que só existem essas duas possibilidades, pois  $Aa$  e  $aA$  não são ordenados.

2º) Não existe razão nenhuma para admitir que um dos dois resultados ocorra com maior frequência, sendo assim, teremos o seguinte modelo teórico de frequências para o experimento,

Cruzamentos	AA	Aa	Total
Frequência teórica	1/2	1/2	1

O espaço dos resultados desse experimento, ou simplesmente, *espaço amostral*, representado pela letra grega  $\Omega$  (ômega), fica:

$$\Omega = \{AA, Aa\}.$$

As vezes o espaço amostral é representado pela letra  $S$ . Portanto, espaço amostral, é o conjunto de todos os resultados possíveis do experimento. Cada um dos elementos, observações que compõem  $\Omega$  chama-se de *ponto amostral*.

Agora, suponha que para o espaço amostral,  $\Omega = \{AA, Aa\}$ , estamos interessados no evento homozigoto, então, temos:

$$A = \{AA\}.$$

Poderíamos também estar interessados no evento heterozigoto, então, temos:

$$B = \{Aa\}.$$

Portanto, *eventos* são subconjuntos do espaço amostral,  $\Omega$ , ou seja, é um conjunto de resultados de um experimento. Se um evento coincide com o espaço amostral,  $\Omega$ , ele se

chama *evento certo* e temos,

$$\begin{aligned} A &= \Omega \\ A &= \{AA, Aa\} \\ A &\rightarrow \text{evento dos fenótipos com dominância.} \end{aligned}$$

Observação: o alelo  $A$  é dominante sobre  $a$ , portanto,  $Aa$  tem o mesmo fenótipo que  $AA$ .

Se um evento não possui nenhum elemento do espaço amostral, temos o *evento impossível*,

$$A = \emptyset \text{ (phi).}$$

**Exemplo:**  $D = \{\text{homozigoto recessivo}\} = \{aa\}$ .

Quando o evento é constituído de apenas um elemento temos o *evento simples*. Como exemplos temos os eventos  $A$  e  $B$ .

Do exemplo acima, podemos fazer uma generalização. Todo o fenômeno ou experimento no qual está envolvido um elemento **casual, aleatório, ou de incerteza**, terá seu *modelo de probabilidades*.

Um modelo probabilístico fica definido, especificado, construído, no momento em que estabelecemos o espaço amostral ( $\Omega$ ) e as probabilidades dos pontos amostrais; para o exemplo, temos o seguinte modelo probabilístico:

Cruzamentos	$AA$	$Aa$	Total
Freqüências teóricas	1/2	1/2	1

Esse espaço amostral é *discreto*, pois podemos enumerar todos os resultados do experimento. Essa enumeração pode ser finita ou infinita.

**Exemplo de modelo.** Cruzamos o genótipo  $Aa$  (pai) e  $Aa$  (mãe). Os resultados dos cruzamentos estão indicados na Figura 3.3.

Observações:

1. As quatro recombinações  $AA$ ,  $Aa$ ,  $aA$  e  $aa$  são igualmente prováveis.
2. As duas recombinações  $Aa$  e  $aA$  não são ordenadas, isto é, não é possível distingui-las biologicamente.

Logo, o espaço de resultados é:

$$\Omega = \{AA, Aa, aa\}.$$

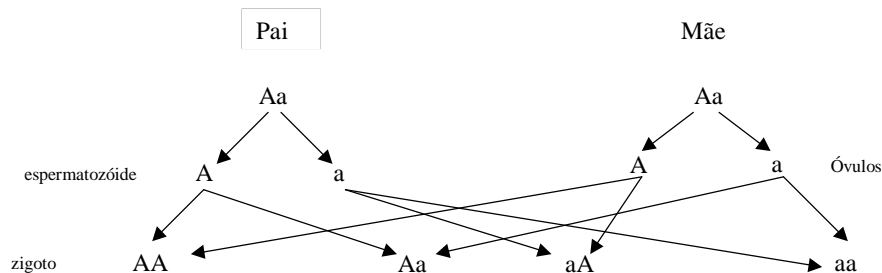


Figura 3.3: Cruzamento de genótipos

Como as recombinações são igualmente prováveis, associamos a cada uma delas a probabilidade  $1/4$ , logo, o modelo probabilístico para o experimento fica:

Cruzamentos	$AA$	$Aa$	$aa$	Total
Frequência teórica	$\frac{1}{4}$	$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$	$\frac{1}{4}$	1

Aqui, novamente, o espaço amostral é discreto.

**Exemplo de modelo:** Em tomateiros, vamos fazer o cruzamento entre dois indivíduos homozigóticos, um sendo recessivo,  $s_p s_p$ , e responsável pelo hábito de crescimento determinado e o outro dominante,  $S_p S_p$ , responsável pelo hábito de crescimento indeterminado. Em F1 obtêm-se:  $S_p s_p$ . Fazendo-se a autofecundação desses indivíduos F1 ( $F1 \times F1$ ), obtêm-se as seguintes recombinações:

$$S_p S_p; S_p s_p; s_p S_p \text{ e } s_p s_p$$

todas com a mesma frequência teórica e igual a  $1/4$ . Veja o quadro:

F1	F1	
	$S_p$	$s_p$
$S_p$	$S_p S_p$	$S_p s_p$
$s_p$	$s_p S_p$	$s_p s_p$

Observações:

1. As quatro recombinações são igualmente prováveis.
2. As duas recombinações,  $S_p s_p$  e  $s_p S_p$ , não são ordenadas, não sendo possível diferenciá-las biologicamente.

O modelo probabilístico fica:

Cruzamentos	$S_p S_p$	$s_p S_p$	$S_p s_p$	$s_p s_p$	Total
Frequência teórica	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$		1

**Observação:** Um espaço amostral pode ser definido de diferentes maneiras para um mesmo experimento, dependendo dos objetivos do problema a ser estudado. Por exemplo, suponha que lancemos uma moeda cinco vezes. Se estamos interessados apenas na sequência de caras e coroas obtida, um espaço amostral é:

$$\Omega_1 = \{ckkkk, kckkk, kkckk, kkkck, \dots\}.$$

Observação: são possíveis  $2^5=32$  pontos amostrais. Mas se estamos interessados no número de caras obtidas, então, um espaço amostral é:

$$\Omega_2 = \{0, 1, 2, 3, 4, 5\}.$$

Temos, aqui, outro exemplo de espaço amostral discreto.

**Exemplo:** Considere um experimento que consiste em medir as alturas  $H$  de homens adultos. Um espaço amostral conveniente é:

$$\Omega = \{H : H > 0\},$$

isto é, o conjunto de todos os números reais positivos. Se  $A$  indica o evento "a altura de homens adultos é superior a 150 cm e inferior a 200 cm", então  $A = \{H : 150 < H < 200\}$ . Esses são exemplos de espaços amostrais *contínuos*, pois o intervalo contém um número infinito de valores.

**Outro exemplo:** Considere um experimento que consiste em medir as alturas  $h$  de plantas de milho. Um espaço amostral conveniente é  $\Omega = \{h : h > 0\}$ , isto é, o conjunto de todos os números reais positivos.

### 3.3 Operações com Eventos

Como os eventos são subconjuntos do espaço amostral  $\Omega$ , são também conjuntos, logo, todas as operações realizadas com conjuntos são válidas para os eventos.

Se  $A$  e  $B$  são dois eventos de  $\Omega$ , então, a operação  $(A \cup B)$ , que lê-se: *A união com B*, significa "a ocorrência do evento  $A$  ou do evento  $B$  ou de ambos", ou seja, os pontos amostrais pertencem ao evento  $A$  ou ao evento  $B$  ou, ainda, a ambos. Essa operação é também chamada de *reunião dos eventos*. A representação gráfica é dada na Figura 3.4. Esta figura é chamada de Digrama de Venn. Em aritmética assemelha-se com  $(A+B)$ .

**Exemplo:** Se  $A = \{1, 2\}$  e  $B = \{1, 5, 6\}$ , então  $(A \cup B) = \{1, 2, 5, 6\}$ . Observe que o ponto 1 ocorre nos dois eventos mas está relacionado uma única vez. O ponto 2 só ocorre em  $A$  e os pontos 5 e 6 ocorrem em  $B$ .

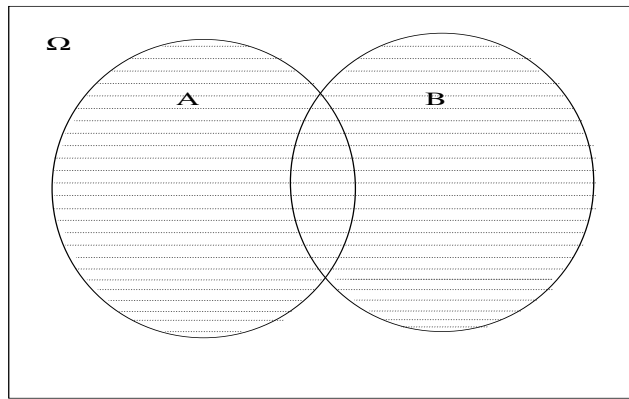


Figura 3.4: Diagrama de Venn para a reunião de eventos

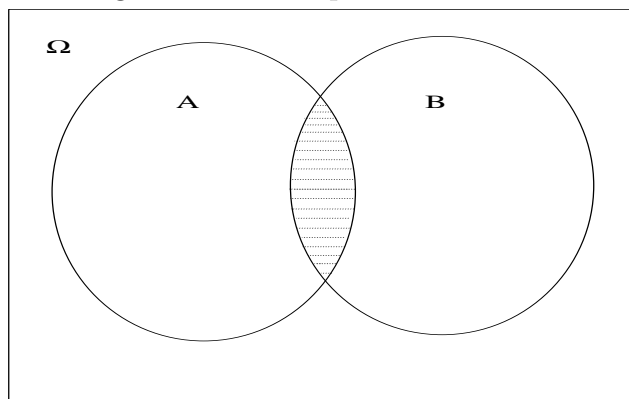


Figura 3.5: Diagrama de Venn para a intersecção de eventos

Uma outra operação entre eventos é a *intersecção*, representada pelo símbolo " $\cap$ ". Então,  $(A \cap B)$  é a ocorrência dos eventos  $A$  e  $B$ , simultaneamente, ou seja, é formado pelos pontos amostrais que pertencem ao eventos  $A$  e  $B$ . Em aritmética, assemelha-se com o produto  $(A \times B)$ . A representação gráfica da intersecção é dada na Figura 3.5.

**Exemplo:** Duas famílias, por exemplo, solanaceas (pimenta, pimentão, tomate) e mir-taceas (goiaba, pitanga, jaboticaba) de uma mesma ordem são por definição disjuntas, isto é, não possuem espécies em comum, logo sua intersecção é vazia, isto é,  $(A \cap B) = \Phi$ .

**Exemplo:** Para os eventos  $A = \{1, 2\}$  e  $B = \{1, 5, 6\}$ ,  $(A \cap B) = \{1\}$ .

Indica-se por  $A^c$  o *complementar* do evento  $A$ , e significa a não ocorrência de  $A$ . Outra forma de representação é  $\bar{A}$ . A representação gráfica do complemento de um evento é dada na Figura 3.6.

**Exemplo:** No lançamento de um dado podemos estar interessados no evento saiu face ímpar,  $A = \{1, 3, 5\}$ , qual é o evento que ocorre sempre que  $A$  não ocorre? Resposta:  $A^c = \{2, 4, 6\}$ .



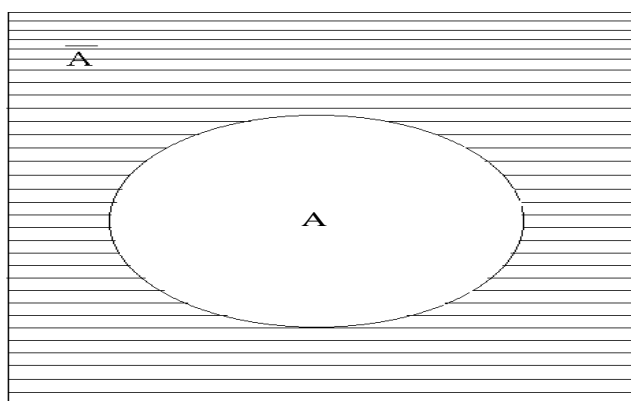


Figura 3.6: Complementar de um evento

**Exemplo:** Vamos considerar todas as espécies pertencentes a ordem Himenoptera como sendo o espaço amostral,  $\Omega$ . Estamos interessados na família Formicidae, este é o nosso evento. Qual será o complementar desse evento? Consistirá de todas as espécies da ordem Himenoptera, mas não pertencente a família Formicidae.

Algumas outras operações envolvendo intersecção, reunião e complemento são dadas por:

- a)  $(A \cap B)^c = A^c \cup B^c$  (Lei de Morgan);
- b)  $(A \cup B)^c = A^c \cap B^c$  (Lei de Morgan);
- c)  $A \cap A^c = \emptyset$ ;
- d)  $A \cup A^c = \Omega$ ;
- e)  $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ ;
- f)  $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$  (distributiva).

**Exercício.** Construa os diagramas de Venn para os itens a) e b).

### *Eventos Mutuamente Exclusivos e Independência*

Se dois eventos não podem ocorrer ao mesmo tempo, ou seja, se a ocorrência de um deles impede a possibilidade de ocorrência do outro, são chamados *eventos mutuamente exclusivos* ou *disjuntos*.

**Exemplo:** considere os resultados do lançamento de um dado,  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , e os eventos,  $A = \{1, 3, 5\}$  e  $B = \{2, 4, 6\}$ , então os eventos A e B são mutuamente exclusivos pois  $A \cap B = \emptyset$ .

**Exemplo:** Considere como sendo  $\Omega$  todas as espécies da ordem Himenoptera, e os eventos  $A = \{\text{espécies da família Formicidae}\}$  e  $B = \{\text{espécies da família Apidae}\}$ , então os eventos A e B são mutuamente exclusivos, pois  $A \cap B = \emptyset$ .

Dois eventos, associados a um experimento aleatório, são ditos *independentes* quando a ocorrência de um deles não interfere na ocorrência do outro.

**Exemplo:** vamos supor que dois alunos tentem resolver uma mesma questão em separado, ao mesmo tempo. Os eventos que consistem em que cada um dos alunos acerte a questão são independentes, pois o fato de um aluno acertar a questão não influencia no fato do outro também acertar.

**Outro exemplo:** Cegueira e surdez, espera-se que sejam independentes.

Em ensaios agrícolas de campo, as parcelas experimentais são independentes, pois o resultado de uma unidade experimental não interfere no resultado das demais.

A definição de independência é importantíssima em estatística. A maioria dos testes, pressupõem independência entre os eventos, como veremos na seção 8.

Eventos mutuamente exclusivos são independentes? Não, eventos mutuamente exclusivos são dependentes, pois a ocorrência de um deles impede a ocorrência do outro.

### 3.4 A Probabilidade de Um Evento

Seja qual for o evento, por exemplo, chuva, genótipos homozigóticos, produção de uma cultura, saiu face ímpar ,etc., a probabilidade de um evento  $A$ , denotada por  $P(A)$  é um número entre 0 e 1, que indica a chance de ocorrência de  $A$ . Quanto mais próximo de 1 é  $P(A) \implies$  maior é a chance de ocorrência de  $A$ , e quanto mais próxima de 0 é  $P(A) \implies$  menor é a chance de ocorrência do evento  $A$ . Definição:

*Seja  $\varepsilon$  um experimento aleatório e  $\Omega$  um espaço amostral associado a esse experimento. A cada evento  $A$  associamos um número real representado por  $P(A)$  e denominado probabilidade de  $A$ , que expressa a chance de ocorrência de  $A$  .*

Sempre temos as seguintes probabilidades: quando o evento é impossível,  $A=\phi \longrightarrow P(A)=0$ ; quando o evento é certo,  $A=\Omega \longrightarrow P(A)=1$ , portanto:

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

### 3.5 Conceito de Probabilidade

#### *Conceito de Freqüência Relativa*

Existem experiências que podem ser repetidas muitas vezes sob condições quase constantes. Sendo assim, observa-se que a frequência relativa estabiliza em certos valores e serve como estimativa da probabilidade. Nesse caso incluem-se as experiências genéticas, muito importante para a área de biociências.

Vamos utilizar um exemplo para ilustrar o conceito de probabilidade baseado na teoria freqüentista.

Vamos considerar o número de nascimentos de meninas e meninos,  $n$ . O espaço de resultados é  $\Omega = \{\text{masc.}, \text{fem.}\}$ . Neste experimento, vamos supor que estamos interessados no número de nascimentos de meninos,  $n_1$ , então,  $n_1$  é a frequência absoluta ou simplesmente a frequência de nascimentos de meninos. A frequência absoluta pode ser um número qualquer entre  $0 \leq n_1 \leq n$ . Mas, como vimos anteriormente, a probabilidade de um evento, é um número entre 0 e 1. Portanto, para nos aproximarmos do conceito de probabilidade, vamos considerar a frequência relativa,  $f_1$ , então,  $f_1 = \frac{n_1}{n}$ , isto é,

$$\text{Frequência relativa de um evento} = \frac{\text{frequência observada do evento}}{\text{número total de repetições do experimento}}.$$

Agora, pode-se ver facilmente que a frequência relativa, varia de 0 a 1 ( $0 \leq f_1 \leq 1$ ). A frequência relativa é freqüentemente expressada em percentagem, logo,  $0\% \leq f_1 \leq 100\%$ .

Se o número de repetições do experimento for muito grande, a frequência  $f_1$  estabiliza em certos valores, por exemplo, para  $n=100.000$  nascimentos, temos que  $f_1 = 53\%$ , este valor é usado como *estimativa da probabilidade*, a frequência relativa converge para a probabilidade. Existem regras que demonstram como tal estimativa é confiável.

**Definição:** Se após  $n$  repetições de um experimento, com  $n$  suficientemente grande, se verificar  $n_1$  ocorrências de um evento, então a probabilidade de ocorrência desse evento será a frequência relativa  $\frac{n_1}{n}$ .

**Exemplo 1:** Se cruzarmos dois indivíduos com genótipos AA e Aa, o gene A do indivíduo AA encontra o gene "A" ou o gene "a" do indivíduo Aa. O espaço dos resultados desse experimento é  $\Omega = \{AA, Aa\}$ . Fatores experimentais demonstram que os dois resultados ocorrem com a mesma probabilidade, isto é:

$$P(AA) = \frac{1}{2} = 50\% \quad \text{e} \quad P(Aa) = \frac{1}{2} = 50\%$$

Para completar nosso *modelo probabilístico*, duas condições sempre devem ser respeitadas:

1.  $0 \leq P(E_i) \leq 1$ ;
2.  $\sum P(E_i) = 1$ .

Onde,  $E_i$  representa um evento qualquer. Como vemos, a probabilidade total do experimento deve ser igual a 1. Estas duas condições são necessárias, a fim de que o nosso modelo seja consistente com o conceito de frequência relativa. No exemplo de cruzamentos de indivíduos, temos que  $0 \leq P(AA) \leq 1$  e  $0 \leq P(Aa) \leq 1$  e  $P(AA) + P(Aa) = 0.5 + 0.5 = 1$ .

**Exemplo 2:** foram colhidas amostras aleatórias de indivíduos de uma população e verificado os seus grupos sanguíneos: A, B, AB e O. Então o  $\Omega = \{A, B, AB, O\}$ . Foram encontradas as frequências 40%, 10%, 5%, 45% para os grupos A, B, AB e O, respectivamente, podemos dizer que a probabilidade de sortear um indivíduo da população com grupo sanguíneo A é de 40% e assim por diante.

**Exercício resolvido:** considere o experimento de cruzamentos de genótipos  $Aa \times Aa$ , e os eventos  $A = \{AA, aa\}$  e  $B = \{Aa\}$ , tais que  $P(A) = \frac{1}{2}$ ,  $P(B) = \frac{1}{2}$  e  $P(A \cap B) = 0$ . Calcular:

a)  $P(A^c)$ .

b)  $P(B^c)$ .

c)  $P(A^c \cap B^c)$ .

d)  $P(A^c \cup B^c)$ .

e)  $P(A^c \cap B)$

Solução:

a)  $1 - P(A) = 1 - \frac{1}{2} = \frac{1}{2}$ .

b)  $1 - P(B) = 1 - \frac{1}{2} = \frac{1}{2}$ .

c)  $P[(A \cup B)^c] = 1 - P(A \cup B) = 1 - 1 = 0$ .

d)  $P[(A \cap B)^c] = 1 - P(A \cap B) = 1 - 0 = 1$ .

e) Como não temos uma operação direta, descrevemos da seguinte forma:

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

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

logo,

$$P(A^c \cap B) = P(B) - P(A \cap B) = \frac{1}{2} - 0 = \frac{1}{2}.$$

### 3.6 A Regra da Adição

Introduziremos a noção de adição de eventos através de um exemplo. Considere um censo realizado nos domicílios da comunidade polonesa de Dom Pedro, próximo à Curitiba, para se estudar a mortalidade dos residentes na colônia. A mortalidade foi verificada sobre o marido, a mulher, marido e mulher e filhos. Em média o estudo foi realizado sobre três gerações. Os dados estão mostrados na Tabela 3.2.

Tabela 3.2: Mortalidade dos residentes na colônia polonesa de Dom Pedro, Curitiba, Paraná.

Idade	Sexo		Total
	Masculino(M)	Feminino(F)	
0 a 1(A)	33	28	61
2 a 4(B)	4	7	11
5 a 9(C)	2	2	4
10 a 14(D)	0	1	1
15 a 29(E)	1	6	7
> 29 (G)	7	8	15
Total	47	52	99

O evento A indica o evento que acontece quando, ocorrer uma morte na colônia, esta pessoa deve ter até um ano de idade. O evento M acontece quando, ocorrer uma morte na colônia, esta pessoa for do sexo masculino. Os demais eventos tem significados análogos. Sendo assim, a probabilidade de ocorrência do evento A, isto é, a pessoa falecida tiver idade entre 0 e 1 ano é dada por,

$$P(A) = \frac{61}{99} = 61,62\%.$$

A probabilidade de ocorrência do evento M, isto é, a pessoa falecida ser do sexo masculino vale,

$$P(M) = \frac{47}{99} = 47,47\%.$$

É fácil ver também que a probabilidade da ocorrência de A e M, simultaneamente vale,

$$P(A \cap M) = \frac{33}{99} = 33,33\%,$$

isto é, temos uma probabilidade de 33,33% da pessoa falecida ser do sexo masculino e ter até um ano de idade.

Agora, se desejarmos calcular a probabilidade de ocorrência de A e/ou de M, isto é,  $P(A \cup M)$ , como será o cálculo? O resultado será dado por:

$$\begin{aligned}P(A \cup M) &= P(A) + P(M) - P(A \cap M) \\P(A \cup M) &= 0,6162 + 0,4747 - 0,3333 \\P(A \cup M) &= 0,7576 \\P(A \cup M) &= 75,76\%.\end{aligned}$$

Devemos subtrair  $P(A \cap M)$ , pois, caso contrário, estaríamos contando duas vezes as pessoas falecidas do sexo masculino e com idade entre 0 e 1 ano.

De modo geral, se A e B são dois eventos quaisquer, a chamada **regra da adição** é dada por:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (3.1)$$

Esta regra é válida para mais do que dois eventos, ou seja, podemos generalizá-la. Por exemplo, para três eventos: A, B e C, a regra fica:

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

Se considerarmos os eventos A e C, a  $P(A \cup C)$  é calculada do seguinte modo:

$$\begin{aligned}P(A) &= \frac{61}{99} = 0,6162 = 61,62\%; \\P(C) &= \frac{4}{99} = 0,040 = 4,00\%; \\P(A \cup C) &= 0,6162 + 0,040 = 0,6562 = 65,62\%,\end{aligned}$$

pois, nesse caso, os eventos A e C são disjuntos ou mutuamente exclusivos, pois se A ocorre, impede a ocorrência de C, e vice-versa. Nesse caso, então,  $A \cap C = \emptyset$ , logo  $P(A \cap C) = 0$ .

Então, quando os dois eventos são **mutuamente exclusivos**, a expressão geral 3.1 se reduz para,

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

**Exemplo:** considere o lançamento de dois dados, sendo os eventos A={soma dos números obtidos igual a 9}, B={número do primeiro dado maior ou igual a 4} e C={soma dos números

menor ou igual a 4}. Enumere os elementos de A, B e C. Obtenha  $P(A \cup B)$  e  $P(A \cup C)$ .

	1	2	3	4	5	6
1	1 1	1 2	1 3	1 4	1 5	1 6
2	2 1	2 2	2 3	2 4	2 5	2 6
3	3 1	3 2	3 3	3 4	3 5	3 6
4	4 1	4 2	4 3	4 4	4 5	4 6
5	5 1	5 2	5 3	5 4	5 5	5 6
6	6 1	6 2	6 3	6 4	6 5	6 6

$$A = \{36 \ 45 \ 54 \ 63\}$$

$$B = \{41 \ 42 \ 43 \ 44 \ 45 \ 46 \ 51 \ 52 \ 53 \ 54 \ 55 \ 56 \ 61 \ 62 \ 63 \ 64 \ 65 \ 66\}$$

$$C = \{11 \ 21 \ 12 \ 31 \ 22 \ 13\}$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{36} + \frac{18}{36} - \frac{3}{36} = \frac{19}{36} = 0,5278.$$

$$P(A \cup C) = P(A) + P(C) - P(A \cap C) = \frac{4}{36} + \frac{6}{36} - \frac{0}{36} = \frac{10}{36} = 0,2778.$$

### 3.7 Probabilidade Condicional e Independência

Vamos voltar a Tabela 3.2 para explicar o que é uma probabilidade condicional. Dado que ocorra a morte de uma pessoa com idade entre 0 e 1 ano, a probabilidade dessa pessoa ser do sexo feminino é de  $\frac{28}{61} = 0,4590 = 45,9\%$ . Isto porque de um total de 61 falecimentos com idade entre 0 e 1 ano, 28 são do sexo feminino.

Representamos uma probabilidade condicional da seguinte forma:

$$P(\text{Feminino} | \text{idade entre 0 e 1 ano}) = \frac{28}{61} = 45,9\%.$$

Lê-se, assim: a probabilidade da pessoa falecida ser do sexo feminino, dado que (|) ela tem idade entre 0 e 1 ano é de 45,9%. A barra vertical indica o evento que ocorreu, o evento conhecido, isto é, sabe-se que a pessoa falecida tem entre 0 e 1 ano, dado isso, deseja-se saber qual é a probabilidade dela ser do sexo feminino. Podemos agora dar uma definição de *probabilidade condicional*.

**Definição:** Para dois eventos quaisquer A e B, associados a um experimento, sendo  $P(B) > 0$ , definimos a probabilidade condicional de A, quando B tiver ocorrido, como sendo:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \quad (3.2)$$

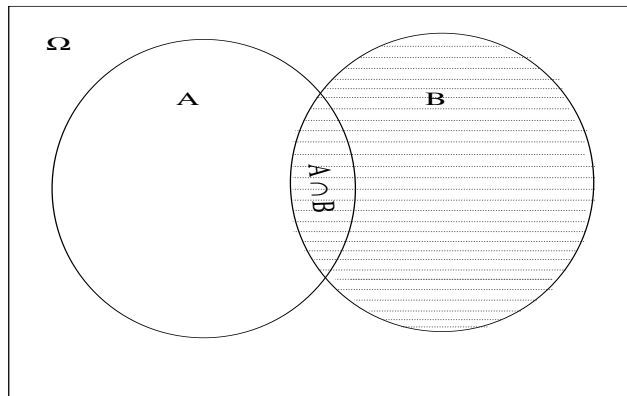


Figura 3.7: espaço reduzido da probabilidade condicional

Para se entender a expressão 3.2, basta pensar que o espaço amostral ficou reduzido, ou restrito ao evento B. Veja Figura 3.7. A presença da intersecção no numerador é justificada facilmente considerando que a ocorrência de A, quando se sabe que B ocorreu, corresponde à ocorrência de A e de B simultaneamente, isto é, de sua intersecção.

Para o exemplo mencionado, se A e F, indicam, respectivamente, os eventos "idade entre 0 e 1 ano e sexo feminino", temos:

$$P(F|A) = \frac{P(F \cap A)}{P(A)} = \frac{28 \div 99}{61 \div 99}.$$

Observe que  $\Omega=99$ , e  $A=61$ , sendo que  $\Omega$  desaparece (simplificação), portanto, o novo espaço de resultados é A. Assim,

$$P(F|A) = \frac{28}{61} = 45,9\%.$$

De modo geral, sempre que calculamos a  $P(A)$ , dado que B tenha ocorrido, estamos sempre calculando a  $P(A)$  em relação ao espaço amostral reduzido de B, isto é, o espaço amostral fica reduzido de  $\Omega$  para B.

**Exemplo:** Numa população os animais podem ser fecundos e não fecundos. Vinte por cento (20%) dos animais da raça X são fecundos; trinta por cento (30%) dos animais da raça Y são não fecundos e setenta e cinco por cento (75%) dos animais são da raça X. Considere os eventos:

$H=\{\text{o animal é da raça X}\}$

$M=\{\text{o animal é da raça Y}\}$



$A = \{\text{o animal é fecundo}\}$        $B = \{\text{o animal não é fecundo}\}$

Raça	Fecundidade		Total
	Não Fecundo(B)	Fecundo(A)	
X (H)	60	15	75
Y (M)	7,5	17,5	25
Total	67,5	32,5	100

$$\begin{aligned}
 P(M|A) &= P(M \cap A)/P(A); \\
 P(M|A) &= \frac{0,175}{0,325} = 0,5384. \\
 P(B|M) &= P(B \cap M)/P(M) = 0,075/0,25 = 0,30 = 30\%. \\
 P(A \cup H) &= P(A) + P(H) - P(A \cap H) = 0,325 + 0,75 - 0,15 = 0,925.
 \end{aligned}$$

### *Regra do Produto de Probabilidades*

Da relação 3.2 podemos obter a chamada Regra do Produto de Probabilidades, dada por:

$$P(A \cap B) = P(B) \times P(A|B). \quad (3.3)$$

**Exemplo:** para os dados do exemplo de população de animais, temos,

$$P(A \cap H) = P(A|H) \times P(H) = 0,20 \times 0,75 = 0,15 = 15\%.$$

### *Um Caso Importante: Independência dos Eventos*

Uma situação especial da fórmula 3.3 é muito importante. Essa situação ocorre quando os dois eventos A e B são independentes, isto é, quando a ocorrência do evento A não influencia na ocorrência do evento B, e vice-versa. Vamos através de um exemplo ilustrar a definição de independência entre eventos.

**Exemplo:** um cientista quer saber se existe dependência entre a cegueira para as cores e a surdez nos homens. Admite-se as seguintes probabilidades para os eventos :

Cegueira	Surdez		Frequência total
	Surdez(S)	Não Surdez( $\bar{S}$ )	
Cegueira para cores(C)	0,0004	0,0796	0,0800
Não Cegueira( $\bar{C}$ )	0,0046	0,9154	0,9200
Frequência total	0,0050	0,9950	1,0000

O evento  $C$  corresponde a homens com cegueira para cores, e o evento  $S$  corresponde a homens com surdez. Vamos calcular a probabilidade de surdez dado que o evento cegueira para cores ocorreu:

$$P(S|C) = \frac{P(S \cap C)}{P(C)} = \frac{0,0004}{0,0800} = 0,0050.$$

Por sua vez, a probabilidade de surdez é:

$$P(S) = 0,0050$$

então,

$$P(S|C) = P(S),$$

isto é, a surdez é *independente* da cegueira para cores, pois a ocorrência da cegueira para cores *não alterou* a probabilidade de ocorrência da surdez. Logo, a ocorrência de um evento não influenciou na ocorrência do outro.

Generalizando, "Se dois eventos são independentes, a  $P(A | B) = P(A)$ , então a regra do produto de probabilidades fica:

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

Essa fórmula pode ser tomada como definição de independência, ou seja,  $A$  e  $B$  são independentes se, e somente se,

$$P(A \cap B) = P(A)P(B). \quad (3.4)$$

**Exemplo:** Sabemos que cegueira para cores e surdez são dois eventos independentes. Qual é a probabilidade de um homem apresentar simultaneamente cegueira e surdez?

$$P(C \cap S) = P(C)P(S) = 0,08 \times 0,005 = 0,004.$$

**Exemplo com dependência:** Uma urna contém duas bolas brancas ( $B$ ) e três vermelhas ( $V$ ). Suponha que sorteemos duas bolas ao acaso, em sequência e *sem reposição*. Isto significa que escolhemos a primeira bola, verificamos a sua cor e não a devolvemos à urna, misturamos as bolas restantes e retiramos a segunda bola. O diagrama de árvore ilustra as possibilidades, veja Figura 3.8. Em cada "galho" da árvore estão indicadas as probabilidades de ocorrência, sendo que para segundas bolas temos probabilidades condicionais. A

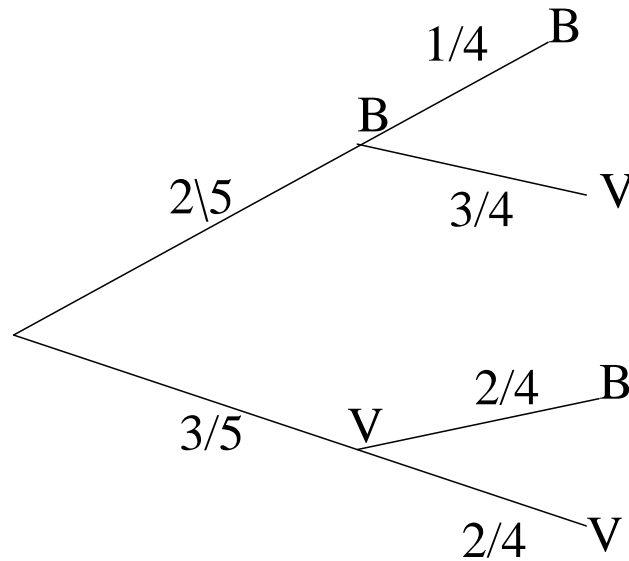


Figura 3.8: diagrama de árvore

distribuição de probabilidades do exemplo é dada por:

Resultados	Probabilidades
BB	$\frac{2}{5} \frac{1}{4} = \frac{2}{20}$
BV	$\frac{2}{5} \frac{3}{4} = \frac{6}{20}$
VB	$\frac{3}{5} \frac{2}{4} = \frac{6}{20}$
VV	$\frac{3}{5} \frac{2}{4} = \frac{6}{20}$
Total	1

As probabilidades calculadas no quadro foram obtidas através da probabilidade condicional. Por exemplo, a probabilidade de sair branca na primeira e branca na segunda é dada por:

$$P(B \cap B) = P(\text{Branca na primeira})P(\text{Branca na segunda} \mid \text{Branca na primeira}) = (2/5)(1/4).$$

Considere os eventos:  $B1 = \{\text{saiu bola branca na primeira retirada}\}$  e  $B2 = \{\text{saiu bola branca na segunda retirada}\}$ .

$$P(B2|B1) = \frac{2}{20} \div \frac{2}{5} = \frac{2}{20} \frac{5}{2} = \frac{1}{4}.$$

$$P(B2) = P(BB) + P(VB) = \frac{2}{20} + \frac{6}{20} = \frac{8}{20} = \frac{2}{5}.$$

Portanto, sendo que  $P(B2|B1) \neq P(B2)$ , implica que os eventos  $B1$  e  $B2$  são dependentes, pois a ocorrência de  $B1$  alterou a probabilidade de  $B2$ .

Os eventos excesso de peso ( $E$ ) e pressão elevada ( $A$ ), biologicamente são dependentes.

### 3.8 Teorema de Bayes

Uma das relações mais importantes envolvendo probabilidades condicionais é dada pelo teorema de Bayes, que expressa uma probabilidade condicional em termos de outras probabilidades condicionais e marginais. Vamos apresentar o Teorema de Bayes através de um exemplo.

**Exemplo.** Temos três profissionais: um Agrônomo, um Biólogo e um Engenheiro Civil. Cada um deles plantou 10 mudas de álamos. Das 10 plantadas pelo Agrônomo 9 sobreviveram; 5 do Biólogo e 2 do Engenheiro. Escolhe-se uma muda ao acaso, se a muda sobreviveu, qual a probabilidade de ela ter sido plantada pelo Engenheiro Civil? Veja a Figura 3.9. Queremos encontrar a  $P(E | S)$ . Sabemos que:

Marginais	Condicionais
$P(A) = \frac{10}{30} = \frac{1}{3}$	$P(S   A) = 90\%$
$P(B) = \frac{10}{30} = \frac{1}{3}$	$P(S   B) = 50\%$
$P(E) = \frac{10}{30} = \frac{1}{3}$	$P(S   E) = 20\%$

As probabilidades marginais são também chamadas de probabilidades *à priori*. Da definição de probabilidade condicional, temos:

$$P(E|S) = \frac{P(E \cap S)}{P(S)}.$$

Mas,

$$P(E \cap S) = P(E) \times P(S|E)$$

Assim:

$$P(E|S) = \frac{P(E)P(S|E)}{P(S)}. \quad (3.5)$$

As probabilidades no numerador sempre serão conhecidas. Precisamos encontrar o valor da probabilidade do denominador,  $P(S)$ . Como A, B e E são eventos mutuamente exclusivos,

	(A) Agrônomo	(B) Biólogo	(E) Engenheiro
S	* * *	* * *	* * ○
NS	* * *	* * ○	○ ○ ○
	○	○	○

Figura 3.9: Sobrevivência de mudas de álamos

e reunidos formam o espaço amostral completo, podemos decompor S, na reunião de três outros que também são mutuamente exclusivos, ou seja:

$$S = (A \cap S) \cup (B \cap S) \cup (E \cap S)$$

e então:

$$\begin{aligned} P(S) &= P(A \cap S) + P(B \cap S) + P(E \cap S) \\ P(S) &= P(A)P(S|A) + P(B)P(S|B) + P(E)P(S|E) \end{aligned}$$

Substituindo-se esse resultado em (3.5), obtem-se:

$$P(E|S) = \frac{P(E)P(S|E)}{P(A)P(S|A) + P(B)P(S|B) + P(E)P(S|E)}.$$

Esse valor é conhecido como probabilidade *à posteriori*. Assim;

$$\begin{aligned} P(E|S) &= \frac{(1 \div 3) \times 0,20}{((1 \div 3) \times 0,90) + ((1 \div 3) \times 0,50) + ((1 \div 3) \times 0,20)} \\ P(E|S) &= \frac{0,06667}{0,30 + 0,16667 + 0,06667} = \frac{0,06667}{0,5333} = 0,1250 \\ P(E|S) &= 12,50\%. \end{aligned}$$

Generalizando para  $n$  eventos.

Seja  $C_1, C_2, C_3, \dots, C_n$ , uma partição do espaço amostral  $\Omega$ , isto é,  $C_i \cap C_j = \emptyset$  para  $i \neq j$ , e  $C_1 \cup C_2 \cup C_3 \cup \dots \cup C_n = \Omega$ ; consideramos A um evento qualquer. Também são conhecidas  $P(C_i)$  e  $P(A|C_i)$  para  $i = 1, 2, 3, \dots, n$ . Então, temos:

$$P(C_i|A) = \frac{P(C_i)P(A|C_i)}{P(C_1)P(A|C_1) + P(C_2)P(A|C_2) + P(C_3)P(A|C_3) + \dots + P(C_n)P(A|C_n)} \quad (3.6)$$

para  $i = 1, 2, 3, \dots, n$ .

### Exercícios resolvidos

1. Num experimento com tomates em casa-de-vegetação, têm-se 26 vasos distribuídos segundo o seguinte delineamento.

Variedades	Adubos			Total
	1	2	3	
1	3	4	2	9
2	1	3	3	7
3	5	2	3	10
total	9	9	8	26

Sorteia-se um adubo (coluna) ao acaso, do qual sorteia-se uma variedade, verificando-se que ocorreu a variedade 2. Qual a probabilidade de que esta variedade esteja sendo tratada com o adubo 1?

Sejam os eventos:

$$\begin{aligned} A1 &= \text{ocorrer o adubo 1;} \\ A2 &= \text{ocorrer o adubo 2;} \\ A3 &= \text{ocorrer o adubo 3;} \\ B2 &= \text{ocorrer a variedade 2.} \end{aligned}$$

Verifica-se que:  $A1 \cap A2 \cap A3 = \emptyset$  e  $A1 \cup A2 \cup A3 = S$ .

Marginais	Condicionais
$P(A1) = \frac{9}{26}$	$P(B2   A1) = \frac{1}{9}$
$P(A2) = \frac{9}{26}$	$P(B2   A2) = \frac{3}{9}$
$P(A3) = \frac{8}{26}$	$P(B2   A3) = \frac{3}{8}$

$$P(A1 | B2) = \frac{P(A1)P(B2|A1)}{P(A1)P(B2|A1) + P(A2)P(B2|A2) + P(A3)P(B2|A3)}$$

$$P(A1 | B2) = \frac{1}{7}.$$

2. (Beiguelman, 2002) A frequência esperada de pessoas  $Rh_+$  em uma população é estimada em 90%. Qual a frequência esperada, nessa população de casais:

$$\begin{aligned} a) \quad Rh_+ &\times Rh_+ && (Rh_+ \text{ e } Rh_+)? \\ 0,90 &\times 0,90 &= & 0,81 &= & 81\%; \end{aligned}$$

$$\begin{aligned} b) \quad Rh_- &\times Rh_- && (Rh_- \text{ e } Rh_-)? \\ 0,10 &\times 0,10 &= & 0,01 &= & 1\%; \end{aligned}$$

$$\begin{aligned} c) \quad Rh_+ &\times Rh_- \\ H &\times M \\ 0,90 &\times 0,10 &= & 0,09 &= & 9\% \\ M &\times H \\ 0,90 &\times 0,10 &= & 0,09 &= & 9\% \\ &&& Resultado &= & 18\%; \end{aligned}$$

$$\begin{aligned}
d) \quad & \text{Marido } Rh_+ \times \text{Mulher } Rh_- \\
& 0,90 \times 0,10 = 0,09 = 9\%; \\
e) \quad & \text{Marido } Rh_- \times \text{Mulher } Rh_+ \\
& 0,10 \times 0,90 = 0,09 = 9\%.
\end{aligned}$$

Observação: os fatores são independentes do sexo.

3. (Beiguelman, 2002) Numa população a frequência de indivíduos  $Rh_-$  é estimada em 16%, a de  $Rh_+$  portadores do gene responsável pelo grupo  $Rh_-$  é estimada em 48%, e a de  $Rh_+$  não portadores desse gene em 36%. Qual a probabilidade de um indivíduo  $Rh_+$  dessa população ser portador do gene que determina o grupo  $Rh_-$ ?

$$A = \{\text{indivíduo ser } Rh_+\}$$

$$B = \{\text{indivíduo } Rh_+ \text{ e portador do gene que determina o grupo } Rh_-\}$$

$$\begin{aligned}
P(A) &= 0,48 + 0,36 = 0,84 \\
P(B|A) &= \frac{P(A \cap B)}{P(A)} = \frac{0,48}{0,84} = 0,5714
\end{aligned}$$

4. (Beiguelman, 2002) Entre os casais  $Rh_+$  e  $Rh_+$  da questão anterior, qual a proporção esperada daqueles capazes de gerar um(a) filho(a)  $Rh_-$ ?

Observação: O casal (homen e mulher) tem que ser portador do gene responsável pelo grupo  $Rh_-$ . Logo,  $0,5714 \times 0,5714 = 0,3265 = 32,65\%$ .

5. Defina um espaço amostral para cada um dos seguintes experimentos aleatórios:

a) Investigam-se famílias com quatro crianças, anotando-se a configuração segundo o sexo.

R: {MMMM, MMMF, MMFM, MFMM, ..., FFFF}. Temos um total de  $2^4 = 16$  eventos.

b) Três jogadores A, B e C disputam um torneio de tênis. Inicialmente, A joga com B e o vencedor joga com C, e assim por diante. O torneio termina quando um jogador ganha duas vezes em seguida ou quando são disputadas, ao todo, quatro partidas.

R: {AA, ACC, ACBB, BB, BCC, BCAA, ACBA, BCAB}.

c) Lance um dado até que a face 5 apareça pela primeira vez.

R: {5, F5, FF5, FFF5, FFFF5, .....}, onde F=face diferente de 5.

d) De todos os alunos do curso de estatística, escolhe-se um ao acaso e anota-se a sua altura.

R: {h |  $1,50 \leq h \leq 2,00$ } onde h é a altura.

6. Uma água é contaminada se forem encontrados bacilos tipo A e/ou bacilos tipo B e C simultaneamente. As probabilidades de se encontrarem bacilos tipo 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 de existirem bacilos tipo C é reduzida à metade.

Calcular:

a)  $P(B \cup C)$ ;

b)  $P(\text{água ser contaminada})$ ;

c)  $P(B | \text{água contaminada})$ .

R:

$$P(A) = 0,30$$

$$P(B) = 0,20$$

$$P(C) = 0,80$$

$$P(B | A) = \emptyset$$

$$P(C | B) = 0,40$$

$$\begin{aligned} a) \quad P(B \cup C) &= P(B) + P(C) - P(B \cap C) \\ &= 0,20 + 0,80 - (0,2 \times 0,40) \\ &= 1 - 0,08 \\ &= 0,92 \end{aligned}$$

$$\begin{aligned} b) \quad P(\text{contaminada}) &= P(A \cup (B \cap C)) \\ &= P(A) + P(B \cap C) - P[A \cap (B \cap C)] \\ &= 0,30 + 0,08 - 0,00 \\ &= 0,38 \end{aligned}$$

$$\begin{aligned} c) \quad P(B | \text{contaminada}) &= P(B \cap \text{contaminada}) \div P(\text{contaminada}) \\ &= P(B \cap C) \div 0,38 \\ &= 0,08 \div 0,38 \\ &= 0,21 \end{aligned}$$

### 3.9 Exercícios Propostos

1. Num levantamento em um município sobre a propriedade da terra e o tamanho do estabelecimento agrícola encontrou-se a seguinte situação:

- 45 agricultores proprietários com estabelecimentos menores que 50 hectares
- 10 agricultores arrendatários com estabelecimentos menores que 50 hectares
- 15 agricultores proprietários com estabelecimentos maiores que 50 hectares
- 2 agricultores arrendatários com estabelecimentos maiores que 50 hectares

Ao escolher ao acaso algum agricultor do município, qual é a probabilidade de que:

a) o estabelecimento agrícola tenha menos de 50 hectares?

b) o agricultor seja arrendatário e o estabelecimento agrícola menor de 50 hectares?



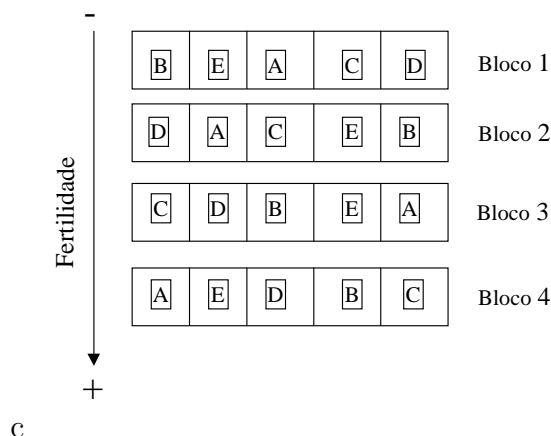


Figura 3.10: Delineamento blocos ao acaso

2. Em um experimento semeiam-se 4 blocos de 5 parcelas cada, com algodão. Em cada bloco, aparecem as variedades A, B, C, D e E, portanto, cada bloco contém 5 parcelas, uma para cada variedade, e estas serão designadas as parcelas através de um sorteio. Veja Figura 3.10. Observe que a construção dos blocos controla a diferença de fertilidade do solo. Assim, podemos estudar o efeito das variedades livre do efeito da fertilidade. Em todos os blocos a variedade C foi a mais produtiva. Qual a probabilidade de que isso tenha ocorrido por acaso, ou seja, as cinco variedades são igualmente produtivas e a variedade C foi a mais produtivas devido a fatores aleatórios no experimento? Qual a suposição feita? Observação: outra situação de uso do delineamento em blocos completos ao acaso, é com germinador de sementes. Para controlar diferenças de temperatura e luminosidade dentro do germinador, consideramos cada prateleira como um bloco.

3. No experimento anterior, suponha-se que a variedade C foi a mais produtiva em 3 blocos e teve o 2º lugar no outro bloco. Qual a probabilidade de ela se destacar igualmente ou mais do que nesse ensaio por simples acaso? Qual a suposição feita?

4. Em um certo locus podem ocorrer dois alelos  $C$  e  $D$ . Admitamos que os possíveis genótipos têm as seguintes probabilidades:

$$P(CC) = 0,46; \quad P(CD) = 0,31; \quad P(DD) = 0,23.$$

Qual é a probabilidade de que um genótipo contenha:

a) o alelo  $C$ ?

b) o alelo  $D$ ?

c) Qual a suposição feita em a) e b)?

5. Em um locus de um certo par de cromossomos, podem ocorrer alelos  $A$  e  $a$ . Os

genótipos  $AA$ ,  $Aa$ ,  $aa$  têm probabilidades:

$$P_{AA} = 0,11; \quad P_{Aa} = 0,37; \quad P_{aa} = 0,52.$$

Em um locus de outro par de cromossomos, podem ocorrer os alelos  $B$  e  $b$ . Os genótipos  $BB$ ,  $Bb$ ,  $bb$  têm as probabilidades:

$$P_{BB} = 0,35; \quad P_{Bb} = 0,25; \quad P_{bb} = 0,40.$$

Encontrar as probabilidades de combinações genéticas:

- a)  $AA$  junto com  $BB$ , isto é  $AA$  e  $BB$ .
- b)  $Aa$  junto com  $Bb$ , isto é  $Aa$  e  $Bb$ .
- c) Qual a suposição feita?

6. Um homen visita um casal que tem dois filhos. Uma das crianças, um menino, vem a sala. Encontre a probabilidade, de o outro também ser um menino, se:

- a) sabe-se que a outra criança é mais nova; observação: construa o espaço amostral de um casal que tem dois filhos de acordo com o sexo dos mesmos, e considere a ordem de nascimento.
- b) nada se sabe sobre a outra criança.

OBSERVAÇÃO: construa o espaço de resultados para um casal com dois filhos.

7. Numa certa cidade, 40% da população tem cabelos castanhos, 25% olhos castanhos e 15% tem cabelos e olhos castanhos. Uma pessoa da cidade é selecionada aleatoriamente.

- a) Se ela tem cabelos castanhos, qual é a probabilidade de ter também olhos castanhos?
- b) Se ela tem olhos castanhos, qual é a probabilidade de não ter cabelos castanhos?
- c) Qual é a probabilidade de não ter nem cabelos nem olhos castanhos?

8. Um lote é formado por 10 animais sadios, 4 com problemas menores e 2 com problemas graves. Todos os animais são numerados e é feita a escolha de um animal ao acaso. Ache a probabilidade de que:

- a) ele não tenha problemas;
- b) ele não tenha problemas graves;
- c) ele ou seja sadio ou tenha problemas graves.

9. Se do lote de animais descritos no problema 8, dois animais forem escolhidos (sem reposição), ache a probabilidade de que:

- a) ambos sejam sadios;
- b) ao menos um seja sadio;
- c) no máximo um seja sadio;
- d) exatamente um seja sadio;
- e) nenhum deles seja sadio.

10. Tem-se um pacote com 20 sementes, 8 das quais tem um alto poder germinativo (germinam), e 12 não germinam. As sementes são analisadas uma após a outra. Se essas sementes forem extraídas ao acaso, sem reposição, qual será a probabilidade de que:

- a) as duas primeiras sementes não germinam;
- b) as duas primeiras sementes germinam;
- c) das duas primeiras sementes analisadas, uma germina e a outra não germina;

OBS: *Estes eventos não são independentes.*

11. Um enxerto tem a probabilidade de sobreviver duas vezes maior do que não sobreviver. Plantados três enxertos, qual a probabilidade de exatamente dois sobreviverem?

12. De três eventos A, B e C, suponhamos A e B independentes, B e C mutuamente exclusivos. Suas probabilidades são:

$$P(A)=0,50, \quad P(B)=0,30 \text{ e } P(C)=0,10.$$

Calcular as probabilidades de:

- a) B e C ocorrerem (ambos);
- b) ocorrer ao menos um dentre A e B;
- c) B não ocorrer;
- d) ocorrerem os três.

13. Um produtor aceitará um lote com 100 sacos de sementes fiscalizadas, se uma amostra de 5 sacos escolhidos ao acaso do lote e inspecionada, não contenha nenhum com poder germinativo inferior ao especificado. Qual é a probabilidade que ele aceite o lote se este contém 10 sacos com poder germinativo abaixo do especificado?

14. Suponha que a probabilidade de se pegar gripe durante uma epidemia seja de 0,60. A experiência tem mostrado que uma vacina vem tendo sucesso de 80% na prevenção da gripe, quando aplicada em pessoas expostas a uma epidemia. Uma pessoa não vacinada tem probabilidade de 0,90 de pegar gripe, quando exposta a uma epidemia. Duas pessoas, uma vacinada e outra não, viajam a negócio. Suponha que elas não vão para o mesmo lugar, não vão estar em contato com as mesmas pessoas e não vão se encontrar (independentes). Qual é a probabilidade de que pelo menos uma delas fique gripada?

15. Num experimento genético é feito um cruzamento com *Drosophila*, no qual é esperado que  $\frac{1}{4}$  das progênes terão "olhos brancos" e  $\frac{1}{2}$  terão a característica chamada "olhos vermelhos". Assume que os dois locus gênicos segregam independentemente.

- a) Qual a proporção de progênes deveriam exibir ambas as características simultaneamente?
- b) Se 4 moscas são amostradas aleatoriamente, qual é a probabilidade de todas serem "olhos brancos"?

c) Qual é a probabilidade que nenhuma das 4 moscas tenham "olhos brancos" ou "olhos vermelhos"?

d) Se duas moscas são amostradas, qual é a probabilidade que pelo menos uma das moscas têm "olhos brancos" ou "vermelhos" ou ambas as características?

16. Uma empresa de sementes fiscalizadas, vende pacotes com 20 Kg cada. As máquinas A, B e C enchem 25, 35 e 40% do total produzido, respectivamente. Da produção de cada máquina 5, 4 e 2%, respectivamente, são pacotes fora do peso aceitável. Escolhe-se ao acaso um pacote e verifica-se que está fora do peso aceitável. Qual a probabilidade de que o pacote venha da máquina A?

17. Se ambos os pais tem genótipo  $Aa$  (heterozigotos), seus filhos tem genótipos  $AA$ ,  $Aa$  e  $aa$  com probabilidades

$$P(AA) = \frac{1}{4}, \quad P(Aa) = \frac{1}{2}, \quad P(aa) = \frac{1}{4}.$$

Qual a probabilidade de que dentre quatro crianças:

a) pelo menos uma tenha o genótipo  $aa$ .

b) Qual o número esperado de crianças com genótipo  $Aa$ ?

18. Em uma maternidade de Florianópolis nasceram em um determinado mês 220 crianças. No mesmo período nasceram em uma maternidade de Chapecó 197 crianças. Do total de 417 recém-nascidos nessas maternidades, 217 eram do sexo masculino e 200 eram do sexo feminino. Se a proporção de meninos e meninas entre os recém-nascidos for independente da sua procedência (as variáveis sexo e procedência são independentes), qual o número esperado de meninos e de meninas em cada um das maternidades durante o período em análise?

19. O estudo de uma tribo no Brasil revelou que 75% tinha sangue tipo A e o restante tinha sangue tipo O. Sessenta por cento (60%) de toda a população tinha fator  $Rh^-$ , enquanto 30% tinha  $Rh^+$  e sangue tipo A. Usando estas informações, encontre a probabilidade de que um membro da tribo tenha:

a) Sangue tipo A ou  $Rh^+$ .

b) Sangue tipo A e  $Rh^-$ .

c)  $Rh^+$  mas não sangue tipo A.

d) Sangue tipo O e  $Rh^-$ .

20. Num estudo sobre fecundidade de duas raças suínas, foram examinados 14 animais de cada raça, obtendo-se o resultado exposto na Tabela 3.3. a) A fecundidade é independente da raça? Justifique através da definição de independência de eventos.

b) Qual é a  $P(F | A)$ ? e c)  $P(F \cup A)$ ?

Tabela 3.3: Fecundidade de duas raças suínas

Raças	Fecundidade		Total
	Fecundas (F)	Não fecundas ( $\bar{F}$ )	
A (A)	12 0,42857	2 0,07143	14 0,5000
B (B)	8 0,28571	6 0,21428	14 0,5000
Total	20 0,71428	8 0,28572	28 1,0000

## 4 Modelos de Probabilidades para Contagens

### 4.1 Introdução

Até aqui introduzimos alguns modelos probabilísticos para experimentos simples, isto é, experimentos cujos os espaços de resultados ( $\Omega$ ) são simples. Esta teoria básica de probabilidade foi importante para uma boa compreensão do que é probabilidade e, também, para o conhecimento de algumas propriedades e regras as quais são úteis no estudo, por exemplo, de genética. Para variáveis qualitativas, os modelos vistos na seção anterior se adaptam muito bem.

No estudo de *Análise Exploratória de Dados*, tivemos contato com diversas variáveis obtidas de experimentos reais, como por exemplo, rendimento de grãos de híbridos de milho em kg/ha, número de dias da emergência à floração (ciclo da cultura), altura de plantas, número de plantas sadias colhidas na área útil de uma parcela, número de micronúcleos em 5000 células de peixes do tipo bagre, diâmetro de *Paepalanthus*. Para atender estas situações práticas mais gerais, precisamos ampliar os conceitos básicos vistos até o momento, para que tenhamos modelos probabilísticos que representem todos os tipos de variáveis vistas até aqui. Em outras palavras, precisamos sofisticar mais nossos modelos.

Inicialmente, dada a sua maior simplicidade estudaremos os modelos para variáveis aleatórias discretas. Alguns exemplos de variáveis aleatórias discretas são: número de plantas sadias (ou doentes) numa unidade experimental, número de insetos capturados numa armadilha, número de brotos por explante, número de sementes que germinam numa amostra de 400 sementes, proporção de enxertos sobreviventes. Já os modelos para variáveis contínuas necessitarão de um artifício matemático, que consiste na generalização do conceito de histograma, estudado na análise exploratória de dados. Esta generalização consiste em se fazer o número de classes tender para o infinito e será estudado no próximo capítulo.

Do ponto de vista prático, é desejável que se defina uma variável associada a um experimento, de tal modo que seus resultados sejam numéricos. Vejamos alguns exemplos.

**Exemplo 1.** No caso de um questionário, e as respostas possíveis são *sim* ou *não*, podemos associar uma variável que toma os valores 1 para sim e 0 para não.

**Exemplo 2.** Num estudo sobre sobrevivência de enxertos em ameixeiras, temos duas possibilidades, o exerto sobrevive ou morre, podemos atribuir o valor 1 para sobrevive e 0 para morte. Uma razão para isto, é que os recursos disponíveis para a análise estatística das variáveis quantitativas são maiores, mais amplas do que para variáveis qualitativas. Isto sugere o uso de artifícios para transformar as variáveis qualitativas em variáveis quantitativas.

Quando uma variável apresenta resultados que tendem a variar de uma observação para outra, em razão da variação do acaso, chama-se *variável aleatória*, às quais iremos associar modelos probabilísticos. Por exemplo, temos as seguintes variáveis aleatórias: número de indivíduos de determinada espécie vegetal encontrados em quadrados de  $10m \times 10m$  em um campo, a conversão alimentar de suínos em crescimento.

Vamos poder verificar até o final desse curso, a grande importância da construção de modelos probabilísticos para variáveis quantitativas.

## 4.2 O Conceito de Variável Aleatória Discreta e Função de Probabilidade

Vamos, agora, através de um exemplo, definir o que é uma variável aleatória discreta, função de probabilidade e distribuição de probabilidade, em termos matemáticos.

**Exemplo:** em um experimento genético com flores de ervilhas, os genótipos  $BB$  e  $VV$  produzem pétalas das flores de cor branca e vermelha, respectivamente. Em  $F_1$ , obtemos:

$$\boxed{BB \times VV = BV},$$

que são flores de cor rosa (genótipo que produz pétalas rosas). Fazendo-se a autofecundação de flores  $BV$ , obtemos os genótipos da Figura 4.1.

Vamos definir a variável  $X$  como sendo igual ao "número de alelos  $V$  em cada resultado da Figura 4.1 (genótipos)". Os resultados associados pela variável  $X$ , dada a suposição que os eventos são igualmente prováveis e  $BV$  e  $VB$  não são ordenados, são:

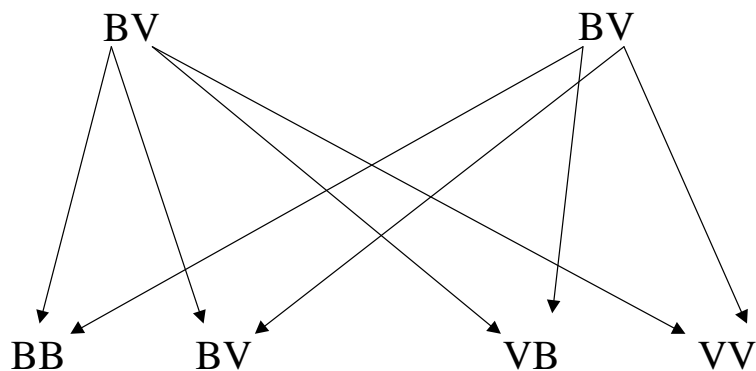


Figura 4.1: Experimento genético com flores de ervilhas

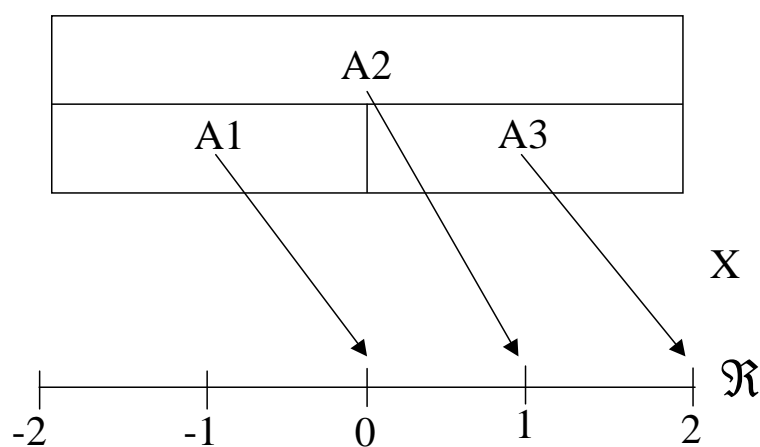


Figura 4.2: Esquema para o significado de variável aleatória

Resultados (eventos)	Probabilidades	$x_i$
$BB$	$\frac{1}{4}$	0
$BV$	$\frac{1}{4}$	1
$VB$	$\frac{1}{4}$	1
$VV$	$\frac{1}{4}$	2

Estes resultados são mostrados no esquema da Figura 4.2, considerando os eventos  $A1 = \{BB\}$ ;  $A2 = \{BV, VB\}$ ;  $A3 = \{VV\}$ . Vemos na Figura 4.2, pois, que a cada resultado do experimento (evento), a variável "X=número de alelos  $V$  em cada genótipo", *associa* um valor numérico. Essa associação em matemática é chamada de *função*.

**Definição:** uma variável aleatória é uma função que associa a todo evento pertencente a uma partição do espaço amostral ( $\Omega$ ), um único número real.

Notamos, que a variável aleatória para ser discreta deve assumir valores em um conjunto enumerável (finito ou infinito) de pontos do conjunto dos números reais ( $\mathbb{R}$ ).

Vimos como associar números aos eventos, agora, é muito importante que se associe probabilidades a estes valores da variável, isto é, como associar a cada valor  $x_i$  da variável aleatória  $X$  a sua probabilidade de ocorrência? Ela é dada pela probabilidade de ocorrência do evento correspondente (eventos equivalentes). Matematicamente, temos:

$$P(X = 0) = P(\{BB\}) = \frac{1}{4},$$

pois,  $X = 0$  se, e somente se, ocorre o evento  $BB$ ;

$$P(X = 1) = P(\{BV \text{ ou } VB\}) = \frac{1}{4} + \frac{1}{4} = \frac{2}{4},$$

pois  $X = 1$  se, e somente se, ocorrem eventos  $BV$  ou  $VB$ .

$$P(X = 2) = P(\{VV\}) = \frac{1}{4},$$

pois  $X = 2$  se, e somente se, ocorre o evento  $VV$ .

Agora vamos trabalhar com números, ao invés de conjuntos, isso implica que podemos usar funções, derivadas, integrais, que são muito utilizadas para representar os fenômenos da natureza.

Na tabela a seguir esquematizamos a *distribuição de probabilidades* da variável aleatória  $X$  (número de alelos  $V$  em cada genótipo)

$x_i$	0	1	2
$p(x_i)$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{4}$

Ao conjunto de pontos  $[x_i; p(x_i)]$  damos o nome de *Distribuição de Probabilidades* (Modelo probabilístico) da variável aleatória  $X$ .

É importante verificar que, para que haja uma distribuição de probabilidades de uma variável aleatória discreta  $X$ , é necessário que:

- a)  $p(x_i) \geq 0$  para todo  $i$ ;
- b)  $\sum_{i=1}^n p(x_i) = 1$ ;
- c)  $P(X = x_i) = p(x_i)$ .

**Definição:** chama-se *função de probabilidade* da variável aleatória discreta  $X$ , que assume os valores  $x_1, x_2, x_3, \dots, x_n$ , a função  $p(x_i)$  que a cada valor de  $x_i$  associa a sua probabilidade de ocorrência, isto é,

$$p(x_i) = P(X = x_i).$$



Nas próximas seções vamos mostrar algumas funções de probabilidades famosas.

### 4.3 A média e a Variância de Uma Variável Aleatória Discreta. Propriedades

Existem características numéricas que são muito importantes em uma distribuição de probabilidades de uma variável aleatória, discreta ou contínua. Essas características numéricas são os *parâmetros* das distribuições. Os dois parâmetros mais importantes são a *média* e a *variância*.

A média é também chamada de *valor esperado*, de *esperança matemática* ou de *média populacional* de uma variável aleatória, e a sua representação é  $E(X)$ , ou  $\mu_X$ , a qual lê-se: a esperança matemática da variável aleatória  $X$ .

Vamos, através de alguns exemplos, introduzir o conceito de esperança matemática.

**Exemplo 1.** O agricultor produtor de batatas-sementes (exemplo da seção 4.3 quer saber qual o lucro médio esperado por caixa da mesma. Vamos supor que:

Uma caixa do tipo I ( $50 < \text{diâmetro} \leq 60 \text{ mm}$ ) dá um lucro de -30,00 u.m.;

Uma caixa do tipo II ( $40 < \text{diâmetro} \leq 50 \text{ mm}$ ) dá um lucro de 100,00 u.m.;

Uma caixa do tipo III ( $28 < \text{diâmetro} \leq 40 \text{ mm}$ ) dá um lucro de 85,00 u.m.;

Uma caixa do tipo IV ( $23 < \text{diâmetro} \leq 28 \text{ mm}$ ) dá um lucro de -10,00 u.m.

As probabilidades (estimativas da frequência relativa) de um agricultor obter uma caixa do tipo:

$$\begin{aligned} I & \text{ é } 100/500 = 0,20; \\ II & \text{ é } 180/500 = 0,36; \\ III & \text{ é } 140/500 = 0,28; \\ IV & \text{ é } 80/500 = 0,16. \end{aligned}$$

Então, se chamarmos de  $X$ : o lucro por caixa de batata-semente e o lucro médio por caixa de  $E(X)$ , temos:

$$\begin{aligned} \mu_X &= E(X) = 0,20(-30,00) + 0,36(100,00) + 0,28(85,00) + 0,16(-10,00) \\ \mu_X &= E(X) = (-6,00) + 36,00 + 23,80 + (-1,6) \\ \mu_X &= E(X) = 52,20 \text{ u.m.} \end{aligned}$$

**Exemplo 2.** Considere a produção de pepinos para conserva, qual o lucro médio esperado pelo agricultor por caixa de pepino? Considere os seguintes valores de lucros para cada uma

das classificações:

1 <sup>a</sup> classe(pepinos de 6 à 9 cm)	lucro de	600	<i>u.m./cx.</i>
2 <sup>a</sup> classe(pepinos de 9 à 12 cm)	lucro de	500	<i>u.m./cx.</i>
3 <sup>a</sup> classe(pepinos maior que 12 cm)	lucro de	390	<i>u.m./cx.</i>
afilados	lucro de	-50	<i>u.m./cx</i>

As probabilidades (estimativas de frequências relativas) de um agricultor obter uma caixa do tipo:

$$\begin{aligned}
 1^{\text{a}} \text{ classe} & \quad \acute{e} \quad 300/500 = 0,60 \\
 2^{\text{a}} \text{ classe} & \quad \acute{e} \quad 100/500 = 0,20 \\
 3^{\text{a}} \text{ classe} & \quad \acute{e} \quad 50/500 = 0,10 \\
 \text{afilados} & \quad \acute{e} \quad 50/500 = 0,10 \\
 \sum p(x_i) & = 1,0
 \end{aligned}$$

Então, se  $X$  é o lucro por caixa de pepino para conserva e  $E(X)$  é o lucro médio esperado por cada caixa, temos que:

$$\begin{aligned}
 \mu_X &= E(X) = 0,60(600) + 0,20(500) + 0,10(390) + 0,10(-50) \\
 \mu_X &= E(X) = 360 + 100 + 39 - 5 \\
 \mu_X &= E(X) = 494 \text{ u.m./cx.}
 \end{aligned}$$

**Exemplo 3.** Em famílias com 4 crianças, vamos admitir, para simplificar, que a proporção de crianças que nascem do sexo masculino e feminino é 1:1. Seja a variável aleatória  $X$ ="número de meninos em famílias com quatro crianças." O espaço de resultados para famílias com 4 crianças é dado por ( $2^4$  pontos amostrais):

$$\Omega = \{FFFF, MFFF, FMFF, FFMF, FFFM, MMFF, FMFM, MFFM, FMMF, FMFM, FFMM, MMMF, MMFM, MFMM, FMMM, MMMM\}.$$

Os valores que  $X$  pode assumir são:

$$X = \{0, 1, 2, 3, 4\}.$$

As probabilidades são dadas por:

$$P(X = 0) = \frac{1}{16}; \quad P(X = 1) = \frac{4}{16}; \quad P(X = 2) = \frac{6}{16}; \quad P(X = 3) = \frac{4}{16}; \quad P(X = 4) = \frac{1}{16}.$$

As probabilidades correspondentes podem ser obtidas pela fórmula (função de probabilidade):

$$P(X = k) = \binom{n}{k} \frac{1}{2^n}.$$

onde:

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

onde  $n!$  significa "n fatorial". Por exemplo,  $3!=3.2.1=6$ .

Agora, representamos por  $E(X)$  ou  $\mu_X$ , o número médio de meninos em tais famílias. Qual é o valor de  $E(X)$ ?

$$E(X) = \mu_X = \left(\frac{1}{16} \times 0\right) + \left(\frac{4}{16} \times 1\right) + \left(\frac{6}{16} \times 2\right) + \left(\frac{4}{16} \times 3\right) + \left(\frac{1}{16} \times 4\right) = 2.$$

Portanto,

$$\begin{aligned}\mu_X &= E(X) = p_1x_1 + p_2x_2 + p_3x_3 + p_4x_4. \\ \mu_X &= E(X) = \sum_{i=1}^4 x_i p(x_i).\end{aligned}$$

**Definição:** se uma variável aleatória discreta  $X$ , toma os valores  $x_1, x_2, \dots, x_n$ , com as probabilidades correspondentes  $p(x_1), p(x_2), \dots, p(x_n)$ , então o seu valor esperado,  $E(X)$  ou  $\mu_X$ , é

$$\mu_X = E(X) = \sum x_i p(x_i), \text{ onde } i = 1, 2, 3, \dots, n.$$

$E(X)$  é uma "média" dos valores que a variável aleatória pode assumir, onde cada valor é ponderado pela probabilidade da variável aleatória ser igual a esse valor. No caso do agricultor, 494 *u.m.* é o valor esperado do lucro por caixa se ele executar o plantio do pepino um número grande de vezes, isto é, quando repetir a lavoura de pepino várias vezes.

A expressão do valor esperado é muito semelhante aquela da média aritmética apresentada na seção , onde:

$$Me(X) = \bar{X} = \sum f_i x_i.$$

A distinção entre  $p_i$ , a probabilidade da variável aleatória  $X$  assumir o valor  $x_i$ , e  $f_i$ , a frequência relativa do resultado  $x_i$ , é que a primeira corresponde a valores obtidos de um modelo teórico ajustado para os dados e, a segunda, corresponde as frequências observadas da variável. Como  $p_i$  e  $f_i$  têm a mesma interpretação, todas as medidas e gráficos discutidos no estudo de análise exploratória de dados para distribuições de frequências ( $f_i$ ), possuem um correspondente na distribuição de probabilidades ( $p_i$ ).

Um segundo parâmetro, é a *variância* de uma distribuição de probabilidade, que mede o grau de dispersão (ou de concentração) de probabilidades em torno da média verdadeira. Quanto menor a variância, maior o grau de concentração das probabilidades em torno da média e vice-versa, quanto maior a variância, maior o grau de dispersão das probabilidades em torno da média.

Chamamos de variância de  $X$  ao valor:

$$\sigma_X^2 = \sum [x_i - E(X)]^2 p(x_i) = \sum p(x_i)x_i^2 - \left[ \sum p(x_i)x_i \right]^2 = E(X^2) - [E(X)]^2.$$

Demonstração:

$$\begin{aligned} \sigma_X^2 &= \sum_{i=1}^n [x_i - E(X)]^2 p(x_i) \\ &= \sum_{i=1}^n \{x_i^2 - 2x_i E(X) + [E(X)]^2\} p(x_i) \\ &= \sum_{i=1}^n x_i^2 p(x_i) - 2E(X) \sum_{i=1}^n x_i p(x_i) + [E(X)]^2 \sum_{i=1}^n p(x_i) \\ &= \sum_{i=1}^n x_i^2 p(x_i) - 2 \left[ \sum_{i=1}^n x_i p(x_i) \right]^2 + \left[ \sum_{i=1}^n x_i p(x_i) \right]^2 \\ &= \sum_{i=1}^n x_i^2 p(x_i) - \left[ \sum_{i=1}^n x_i p(x_i) \right]^2 \\ &= E(X^2) - [E(X)]^2. \end{aligned}$$

O resultado acima foi obtido com as restrições:

$$\sum_{i=1}^n p(x_i) = 1 \quad e \quad E(X) = \sum_{i=1}^n x_i p(x_i).$$

Outra expressão para a variância de  $X$  é dada por:

$$\sigma_X^2 = E[X - E(X)]^2.$$

Para contornar o problema da unidade elevada ao quadrado, definimos o *desvio padrão*,

$$\sigma_X = \sqrt{\sigma_X^2}.$$

Para o exemplo do lucro por caixa de batata-semente, temos:

$x_i$ (lucro)	$p(x_i)$	$x_i^2$	$p(x_i)x_i^2$
-30,00	0,20	900,00	180
100,00	0,36	10.000,00	3.600
85,00	0,28	7.225,00	2.023
-10,00	0,16	100,00	16
Total	1,00		5819

Logo:

$$\begin{aligned}\sigma_X^2 &= 5819 - [52, 20]^2 \\ \sigma_X^2 &= 3.094 \text{ u.m.}^2\end{aligned}$$

e o valor do desvio padrão é:

$$\begin{aligned}\sigma_X &= D.P.(X) = \sqrt{3.094, 16} \\ \sigma_X &= D.P.(X) = 55,63 \text{ u.m.}.\end{aligned}$$

Para o exemplo da variável lucro esperado por caixa de pepino para conserva, temos:

$$\begin{aligned}\sigma_X^2 &= 281.400 - [494]^2 \\ \sigma_X^2 &= 37.364 \text{ u.m.}^2/cx.\end{aligned}$$

O desvio padrão fica:

$$\begin{aligned}\sigma &= \sqrt{37.364} \\ \sigma &= 193,29 \text{ u.m.}/cx.\end{aligned}$$

Para o exemplo da variável número de meninos em família com quatro crianças, temos:

$$\begin{aligned}\sigma_X^2 &= 5 - 2^2 = 1 \text{ menino}^2. \\ \sigma_X &= \sqrt{\sigma_X^2} = \sqrt{1} = 1 \text{ menino}.\end{aligned}$$

### *Algumas Propriedades da Média e da Variância*

1) Seja  $K$  uma constante, temos que  $E(K) = K$ .

Demonstração:

$$\begin{aligned}E(K) &= \sum_{i=1}^n Kp(x_i) \\ &= K \sum_{i=1}^n p(x_i) \\ &= K(1) = K.\end{aligned}$$

2) Sejam  $K$  uma constante e  $X$  uma variável aleatória, temos que:  $E(KX) = KE(X)$ .

Demonstração:

$$\begin{aligned}E(KX) &= \sum_{i=1}^n Kx_i p(x_i) \\ &= K \sum_{i=1}^n x_i p(x_i) \\ &= KE(X).\end{aligned}$$

3) Sejam  $X$  e  $Y$  duas variáveis aleatórias, temos que:  $E(X \pm Y) = E(X) \pm E(Y)$ .

Demonstração: para facilitar o entendimento da demonstração construímos a tabela a seguir.

Y	X			p(y)
	$x_1 = 0$	$x_2 = 1$	$x_3 = 2$	
$y_1=1$	3/20	3/20	2/20	8/20
$y_2=2$	1/20	1/20	2/20	4/20
$y_3=3$	4/20	1/20	3/20	8/20
p(x)	8/20	5/20	7/20	1,00

$$\begin{aligned}
 E(X \pm Y) &= \sum_{i=1}^m \sum_{j=1}^n (x_i \pm y_j) p(x_i, y_j) \\
 &= \sum_{i=1}^m \sum_{j=1}^n x_i p(x_i, y_j) \pm \sum_{i=1}^m \sum_{j=1}^n y_j p(x_i, y_j) \\
 &= \sum_{i=1}^m x_i \sum_{j=1}^n p(x_i, y_j) \pm \sum_{j=1}^n y_j \sum_{i=1}^m p(x_i, y_j).
 \end{aligned}$$

Agora, para um  $i$  fixo,

$$\sum_{j=1}^n p(x_i, y_j) = p(x_i)$$

e, para  $j$  fixo,

$$\sum_{i=1}^m p(x_i, y_j) = p(y_j),$$

logo, podemos escrever:

$$\begin{aligned}
 E(X \pm Y) &= \sum_{i=1}^m x_i p(x_i) \pm \sum_{j=1}^n y_j p(y_j) \\
 E(X \pm Y) &= E(X) \pm E(Y).
 \end{aligned}$$

**Exemplo.** Deve-se escolher um homen dentre um grupo de homens. O peso médio no grupo é de  $\mu_H = 79$  Kg, com desvio padrão de  $\sigma_H = 10$  Kg. Uma mulher deve ser escolhida de um grupo de mulheres com peso médio de  $\mu_M = 65$  Kg e desvio padrão de  $\sigma_M = 7$  Kg. Determine a  $E(X_H + X_M)$  e a variância  $Var(X_H + X_M)$  dos pesos combinados de um homen e de uma mulher.

$$E(X_H + X_M) = \mu_H + \mu_M$$

$$\begin{aligned}
&= 79 + 65 = 144 \text{ Kg.} \\
\sigma_{X_H+X_M}^2 &= \sigma_{X_H}^2 + \sigma_{X_M}^2 \text{ (} X_H \text{ independente de } X_M \text{)} \\
&= 100 + 49 = 149 \text{ Kg}^2. \\
\sigma_{X_H+X_M} &= \sqrt{149} = 74,5 \text{ kg.}
\end{aligned}$$

4) Se  $X$  e  $Y$  são variáveis aleatórias independentes, então:

$$E(XY) = E(X)E(Y).$$

Demonstração:

$$\begin{aligned}
E(XY) &= \sum_{i=1}^m \sum_{j=1}^n x_i y_j p(x_i, y_j) \\
&= \sum_{i=1}^m \sum_{j=1}^n x_i y_j p(x_i) p(y_j) \\
&= \sum_{i=1}^m x_i p(x_i) \sum_{j=1}^n y_j p(y_j) \\
&= E(X)E(Y).
\end{aligned}$$

5)  $Var(K) = 0$ .

Demonstração:

$$\begin{aligned}
Var(K) &= E(K^2) - [E(K)]^2 \\
&= K^2 - K^2 \\
Var(K) &= 0.
\end{aligned}$$

6)  $Var(KX) = K^2 Var(X)$ .

Demonstração:

$$\begin{aligned}
Var(KX) &= E(K^2 X^2) - [E(KX)]^2 \\
&= K^2 E(X^2) - [KE(X)]^2 \\
&= K^2 E(X^2) - K^2 [E(X)]^2 \\
&= K^2 \{E(X^2) - [E(X)]^2\} \\
&= K^2 Var(X).
\end{aligned}$$

**Definição:** Se  $X$  e  $Y$  são duas variáveis aleatórias, a covariância de  $X$  e  $Y$  é definida por:

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y),$$

ou seja, o valor médio do produto dos desvios de  $X$  e  $Y$  em relação aos seus valores médios. A covariância é uma medida de associação entre  $X$  e  $Y$ .

Demonstração:

$$\begin{aligned} Cov(X, Y) &= \sum_{i=1}^m \sum_{j=1}^n [x_i - E(X)][y_j - E(Y)]p(x_i, y_j) \\ Cov(X, Y) &= \sum_{i=1}^m \sum_{j=1}^n x_i y_j p(x_i, y_j) - \sum_{i=1}^m \sum_{j=1}^n x_i E(Y) p(x_i, y_j) \\ &\quad - \sum_{i=1}^m \sum_{j=1}^n E(X) y_j p(x_i, y_j) + \sum_{i=1}^m \sum_{j=1}^n E(X) E(Y) p(x_i, y_j). \end{aligned}$$

Por definição:

$$\sum_{i=1}^m \sum_{j=1}^n x_i y_j p(x_i, y_j) = E(XY).$$

Logo,

$$\begin{aligned} Cov(X, Y) &= E(XY) - E(Y) \sum_{i=1}^m x_i \sum_{j=1}^n p(x_i, y_j) \\ &\quad - E(X) \sum_{j=1}^n y_j \sum_{i=1}^m p(x_i, y_j) + E(X) E(Y) \sum_{i=1}^m \sum_{j=1}^n p(x_i, y_j) \\ Cov(X, Y) &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ Cov(X, Y) &= E(XY) - E(X)E(Y). \end{aligned}$$

**Exemplo:** vamos considerar a distribuição conjunta de probabilidades das variáveis  $X$ =”número de vagens por planta” e  $Y$ =”número de grãos por vagem”:

X	Y					$p(x)$
	0	1	2	3	4	
0	0/15	0/15	0/15	0/15	0/15	0/15
1	2/15	2/15	3/15	2/15	1/15	10/15
2	0/15	1/15	2/15	1/15	1/15	5/15
$p(y)$	2/15	3/15	5/15	3/15	2/15	1

Inicialmente, definiremos a variável  $W = XY$  e vamos obter a sua distribuição de probabilidade.



idades.

$W$	$P(W)$	$WP(W)$
0	2/15	0/15
1	2/15	2/15
2	4/15	8/15
3	2/15	6/15
4	3/15	12/15
6	1/15	6/15
8	1/15	8/15

Temos que:  $E(X) = 20/15$ ,  $E(Y) = 30/15$ , portanto,  $COV(X, Y) = (42/15) - (20/15)(30/15) = (42/15) - (588/225) = 2,8 - 2,6667 = -0,1333$ .

$$7) \text{ } Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X, Y)$$

onde  $Cov(X, Y) = E(XY) - E(X)E(Y)$ , como foi demonstrado.

Demonstração:

$$\begin{aligned}
 Var(X + Y) &= E[(X + Y)^2] - [E(X + Y)]^2 \\
 &= E(X^2) + 2E(XY) + E(Y^2) - [E(X) + E(Y)]^2 \\
 &= E(X^2) + 2E(XY) + E(Y^2) - [E(X)]^2 - 2E(X)E(Y) - [E(Y)]^2 \\
 &= E(X^2) - [E(X)]^2 - E(Y^2) - [E(Y)]^2 + 2E(XY) - 2E(X)E(Y) \\
 &= Var(X) + Var(Y) + 2Cov(X, Y).
 \end{aligned}$$

Se  $X$  e  $Y$  são duas variáveis aleatórias independentes, então  $Cov(X, Y) = 0$  e, portanto,

$$Var(X \pm Y) = Var(X) + Var(Y).$$

Observação: se  $Cov(X, Y) = 0$  não implica que  $X$  e  $Y$  são independentes. Pode ter uma outra dependência que não a linear.

Uma fórmula prática para o cálculo da covariância é dada por:

$$Cov(X, Y) = \sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}.$$

## 4.4 Alguns Modelos Probabilísticos para Contagens

Os diferentes tipos de distribuições de probabilidades podem ser considerados como *modelos para descrever situações reais* que envolvam resultados gerados pelo acaso. Uma consequência da simplificação, ou seja, da eliminação de detalhes de pouca importância dos

fenômenos reais, é que, em suas formas mais puras, elementares, poucos problemas são realmente únicos. Por isso, freqüentemente um pequeno número de modelos é suficiente para resolver muitos problemas que, a primeira vista, não parecem relacionados. Podemos, então, concluir que a maioria dos problemas podem ser resolvidos com o auxílio de poucos modelos básicos.

Esses modelos, cada qual apresentam uma série de hipóteses ou pressuposições, que devem ser atendidas para que o modelo possa ser utilizado validamente.

O ponto chave para a utilização de um modelo consiste em *confrontar* as hipóteses básicas do modelo e as condições da situação real. Se as hipóteses básicas são verificadas (atendidas), pode-se usar o modelo em questão.

Nesta seção, iremos estudar alguns desses modelos, procurando enfatizar como eles ocorrem na prática, sua função de probabilidade, parâmetros, e como calcular probabilidades.

## 4.5 Distribuição de Bernoulli

Seja um experimento onde desejamos verificar se uma semente de trigo, "germina" ou "não germina".

É claro que quando realizamos um experimento estamos interessados em verificar alguma(s) variável(is). Então, precisamos associar uma variável aleatória  $X$  que chamaremos de "germinação da semente de trigo" aos possíveis resultados do experimento; como temos apenas dois resultados possíveis, a variável aleatória  $X$  vale:

$$\begin{aligned}x &= 1 \text{ se a semente germinar;} \\x &= 0 \text{ se a semente não germinar.}\end{aligned}$$

Vamos estabelecer algumas condições (pressuposições):

1. Só é feita uma única tentativa (repetição) do experimento;
2. Só são possíveis dois resultados (germina ou não germina).

Vamos definir o evento  $G = \{\text{A semente germina}\}$ , sendo a sua probabilidade dada por  $P(G) = \pi$ , portanto, o valor do parâmetro  $\pi$  está entre 0 (zero) e 1 (um); a probabilidade da semente não germinar, é  $P(\bar{G}) = 1 - \pi = q$ .

A distribuição de probabilidades fica:

Resultados	$x_i$	$p(x_i)$
$\bar{G}$	0	$1 - \pi = q$
G	1	$\pi$
Total		$\pi + (1 - \pi) = 1$

Dizemos que a variável aleatória  $X$ ="germinação da semente", assim definida, tem distribuição de Bernoulli.

A função de probabilidade de uma variável Bernoulli é dada por:

$$P(X = x_i) = \pi^{x_i}(1 - \pi)^{1-x_i}. \quad (4.1)$$

**Exemplo:** Podemos obter as probabilidades da variável aleatória assumir os valores 0 e 1 através da aplicação da função de probabilidade:

$$\begin{aligned} P(X = 0) &= \pi^0(1 - \pi)^{1-0} = (1 - \pi); \\ P(X = 1) &= \pi^1(1 - \pi)^0 = \pi. \end{aligned}$$

### *Média e Variância de uma Variável Aleatória Bernoulli*

$$\begin{aligned} E(X) &= \sum x_i p(x_i) = (0 \times (1 - \pi)) + (1 \times \pi) = 0 + \pi = \pi. \\ VAR(X) &= E(X^2) - [E(X)]^2. \\ E(X^2) &= \sum x_i^2 p(x_i) = 0^2(1 - \pi) + 1^2\pi = \pi. \\ VAR(X) &= \pi - \pi^2 = \pi(1 - \pi). \end{aligned}$$

**Exemplo 1:** uma urna tem 30 bolas brancas e 20 verdes. Retira-se uma bola dessa urna. Seja  $X$ = saiu bola verde. Verifique se é um ensaio de Bernoulli (confronte as suposições do modelo com as condições do ensaio). Determinar  $P(X = x_i)$ , calcular  $E(X)$  e  $Var(X)$ .

Sim, pois  $\begin{cases} 1) \text{ uma única repetição} \\ 2) \text{ somente dois resultados (verde ou branco)} \end{cases}$

$$X = \begin{cases} 0 \longrightarrow (1 - \pi) = \frac{30}{50} = \frac{3}{5} \\ 1 \longrightarrow \pi = \frac{20}{50} = \frac{2}{5} \end{cases}$$

Podemos estruturar a função de probabilidade da seguinte forma:

$$P(X = x_i) = (0,40)^{x_i}(0,60)^{1-x_i}.$$

A média e a variância valem:

$$E(X) = 0,40 \text{ e } Var(X) = 0,24, \text{ respectivamente.}$$

**Exemplo 2:** em uma gaiola estão seis coelhos com uma mutação sanguínea letal e três outros com uma mutação óssea. Sorteia-se, aleatoriamente, um coelho dessa gaiola. Seja  $X$  = presença de mutação óssea. Verifique se é um ensaio de Bernoulli. Determinar a  $P(X = x_i)$ , calcular  $E(X)$ ,  $Var(X)$ .

Sim, pois:  $\left\{ \begin{array}{l} 1) \text{ uma única repetição;} \\ 2) \text{ somente dois resultados possíveis (mutação óssea ou sanguínea).} \end{array} \right.$

$$X = \left\{ \begin{array}{llll} 0 & \longrightarrow & (1 - \pi) & = \frac{6}{9} = 0,6667 \\ 1 & \longrightarrow & \pi & = \frac{3}{9} = 0,3333 \end{array} \right.$$

$$P(X = x_i) = (0,3333)^{x_i} \times (0,6667)^{1-x_i}.$$

$$E(X) = \sum x_i p(x_i) = 0 \times 0,6667 + 1 \times 0,3333 = 0,3333.$$

$$Var(X) = \pi(1 - \pi) = 0,3333 \times 0,6667 = \frac{6}{9} \times \frac{3}{9} = \frac{2}{9}.$$

## 4.6 A Distribuição Binomial

O objetivo inicial é encontrarmos a função de probabilidade (a expressão matemática) da distribuição binomial<sup>4</sup>. Para isso vamos ver como a distribuição binomial ocorre na prática.

Suponha, agora, que desejamos verificar a germinação de duas sementes de trigo. Vamos estabelecer quatro pressuposições nessa experiência: primeira pressuposição: o fato de uma semente germinar ou não, não tem influência no fato da outra semente germinar ou não, isto é, a germinação das sementes são *independentes*; segunda pressuposição: a probabilidade das sementes germinarem permanece constante (para isso, deve-se utilizar no experimento, sementes bastante homogêneas quanto às propriedades físicas, químicas e biológicas) e igual a  $\pi$  (identicamente distribuídas); terceira pressuposição: só há dois resultados possíveis, germina, ou não germina, e a quarta pressuposição: existe  $n = 2$  repetições. A variável ( $X$ ) pode ser definida como o "número de sementes que germinam" e, a probabilidade de uma semente germinar continua sendo igual a  $\pi$ .

---

<sup>4</sup>Esta distribuição foi estudada pelo matemático suíço Jacob Bernoulli (1664-1705)

Observe que vamos usar as definições de  $P(A \cup B)$  e  $P(A \cap B)$ , onde  $P(A \cup B) = P(A) + P(B)$  e  $P(A \cap B) = P(A)P(B)$ , dadas nas subseções 3.6 e 3.7.

A distribuição de probabilidade fica:

Resultados	$x_i$	$p(x_i)$
GG	2	$\pi\pi=\pi^2(1-\pi)^0 \longrightarrow p(2)=1\pi^2(1-\pi)^0 = \binom{2}{2} \pi^2(1-\pi)^{2-2}$
$G\bar{G}$	1	$\pi(1-\pi)=\pi^1(1-\pi)^1 \longrightarrow p(1)=2\pi^1(1-\pi)^1 = \binom{2}{1} \pi^1(1-\pi)^{2-1}$
$\bar{G}G$	1	$(1-\pi)\pi=\pi^1(1-\pi)^1$
$\bar{G}\bar{G}$	0	$(1-\pi)(1-\pi)=\pi^0(1-\pi)^2 \longrightarrow p(0)=1\pi^0(1-\pi)^2 = \binom{2}{0} \pi^0(1-\pi)^{2-0}$

Na tabela, o primeiro resultado mostrado é que as duas sementes germinam, isto é, temos o resultado  $GG$ . Neste caso, em que as duas sementes germinam, o valor assumido pela variável  $X$ , "número de sementes que germinam", é  $x = 2$ . Estamos considerando que a probabilidade de uma semente germinar é  $\pi$ , ou seja,  $P(G) = \pi$ . Assim, a probabilidade de duas sementes germinarem independentemente é dada por:  $P(G_1 \cap G_2) = P(G_1)P(G_2) = \pi\pi = \pi^2$ , onde  $G_1$  e  $G_2$  representam as sementes 1 e 2, respectivamente. Esta probabilidade pode ser reescrita como:

$$\pi.\pi = \pi^2(1-\pi)^0 =$$

onde os expoentes 2 e 0 indicam que duas sementes germinaram e nenhuma semente não germinou, respectivamente; Ainda podemos escrever:

$$= 1\pi^2(1-\pi)^{2-2} =$$

onde o valor 1 indica que existe somente uma sequência  $GG$ , e a diferença 2-2 indica que de duas sementes ensaiadas(o primeiro dois)as duas germinaram(o segundo dois), portanto, nenhuma não germinou. Finalmente podemos escrever:

$$= \binom{2}{2} \pi^2(1-\pi)^{2-2},$$

onde,  $\binom{2}{2}$  leia-se, combinação de 2(duas sementes ensaiadas), tomados 2 a 2(duas sementes germinaram), e é calculada por:

$$\binom{2}{2} = \frac{2!}{2!(2-2)!} = \frac{2!}{2!0!} = \frac{1 \times 2}{1 \times 2 \times (1)} = 1.$$

Para o segundo resultado, em que uma semente germina ( $G$ ) e a outra não germina ( $\bar{G}$ ), a probabilidade é dada por:

$$P(G \cap \bar{G}) = P(G)P(\bar{G}) = \pi(1-\pi) = \pi^1(1-\pi)^1.$$

Observe que, temos duas combinações possíveis em que uma semente germina e a outra não germina,  $G\bar{G}$  e  $\bar{G}G$ , portanto, a probabilidade vale:

$$P(G\bar{G} \cup \bar{G}G) = 2\pi^1(1-\pi)^1 = \binom{2}{1}\pi^1(1-\pi)^{2-1}.$$

Neste cálculo usamos combinações pois a ordem dos resultados não importa. Na verdade, em termos de resultados, corresponde a um único resultado, qual seja: uma semente germina e a outra não germina.

Usamos o mesmo procedimento para calcular a  $P(\bar{G} \cap \bar{G})$ .

Vamos, agora, verificar a germinação de três sementes de trigo (vamos repetir o experimento 3 vezes); considerando verdadeira a hipótese de independência e probabilidades constantes,  $\pi$ , a distribuição de probabilidade fica:

Resultados	$x_i$	$P(X = x_i) = p(x_i)$	
GGG	3	$\pi\pi\pi = \pi^3(1-\pi)^0$	$\longrightarrow 1\pi^3(1-\pi)^0 = \binom{3}{3}\pi^3(1-\pi)^{3-3}$
GG $\bar{G}$	2	$\pi\pi(1-\pi) = \pi^2(1-\pi)^1$	$\longrightarrow 3\pi^2(1-\pi)^1 = \binom{3}{2}\pi^2(1-\pi)^{3-2}$
G $\bar{G}$ G	2	$\pi(1-\pi)\pi = \pi^2(1-\pi)^1$	
$\bar{G}$ GG	2	$(1-\pi)\pi\pi = \pi^2(1-\pi)^1$	
G $\bar{G}$ $\bar{G}$	1	$\pi(1-\pi)(1-\pi) = \pi^1(1-\pi)^2$	$\longrightarrow 3\pi^1(1-\pi)^2 = \binom{3}{1}\pi^1(1-\pi)^{3-1}$
$\bar{G}$ GG $\bar{G}$	1	$(1-\pi)\pi(1-\pi) = \pi^1(1-\pi)^2$	
$\bar{G}$ $\bar{G}$ G	1	$(1-\pi)(1-\pi)\pi = \pi^1(1-\pi)^2$	
$\bar{G}$ $\bar{G}$ $\bar{G}$	0	$(1-\pi)(1-\pi)(1-\pi) = \pi^0(1-\pi)^3$	$\longrightarrow 1\pi^0(1-\pi)^3 = \binom{3}{0}\pi^0(1-\pi)^{3-0}$
Total		1	1

Generalizando para  $n$  ensaios. Vamos agora verificar a germinação de  $n$  sementes de trigo ( $n$  repetições do experimento, ou, também, podemos dizer,  $n$  ensaios independentes de Bernoulli), a probabilidade de  $k$  sementes de trigo germinar e, portanto,  $n - k$  sementes não germinar, nesta sequência:

$$\underbrace{G, G, \dots, G}_k, \underbrace{\bar{G}, \bar{G}, \dots, \bar{G}}_{n-k}$$

é dada por:

$$\pi^k(1-\pi)^{n-k}.$$

Mas, uma outra sequência de  $k$  sementes que germinam e  $n - k$  sementes que não germinam é:

$$G, G, G, \dots, \bar{G}, \bar{G}, G, \bar{G}, \dots, \bar{G}.$$

O valor da probabilidade continua sendo  $\pi^k(1 - \pi)^{n-k}$ .

Uma outra sequência poderia ser:

$$G, G, G, \dots, \bar{G}, G, \bar{G}, \bar{G}, \dots, \bar{G}.$$

Novamente, o valor da probabilidade nesta sequência é  $\pi^k(1 - \pi)^{n-k}$ .

Mas, existem:

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

de tais seqüências, onde  $k$  sementes germinam e  $n - k$  sementes não germinam, de modo que a probabilidade de  $k$  sementes germinarem é dada por:

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \quad (4.2)$$

para  $k = 0, 1, 2, 3, \dots, n$ .

*Observações:*

- 1) a denominação binomial decorre do fato de os coeficientes  $\binom{n}{k}$  serem exatamente os coeficientes do desenvolvimento binomial das  $n$  potências  $(a + b)$ ;
- 2) o cálculo dos coeficientes, para  $n$  e  $k$  grandes, são difíceis de serem realizados, por vezes necessita da ajuda de computadores, sendo assim, será estudado na Seção 4.8 o uso de uma aproximação para a distribuição binomial.

Estas probabilidades também podem ser indicadas por:

$$b(k : n; \pi).$$

Os possíveis valores de  $k = 0, 1, 2, 3, \dots, n$  e as probabilidades  $P(X = k)$ , dadas em 4.2 constituem a chamada *distribuição binomial*.

Quando uma variável aleatória  $X$  tem distribuição binomial com os parâmetros  $n$  e  $\pi$  escrevemos:

$$X \sim b(n; \pi).$$

Suposições do modelo binomial:

1. Existem  $n$  repetições ou provas idênticas do experimento. *Exemplo:* número de plantas sadias colhidas em parcelas de  $20m^2$  (foram plantadas 27 plantas em cada parcela),  $X : 0, 1, 2, \dots, 27$ , então,  $n$  é o número total de casos possíveis da variável que estamos estudando.
2. Só há dois tipos de resultados possíveis (plantas sadias ou doentes).

3. As probabilidades  $\pi$  de sucesso e  $1 - \pi$  de fracasso permanecem constantes em todas as repetições. Na prática não temos certeza absoluta disso, mas consideramos verdadeira esta suposição desde que as probabilidades sejam próximas.
4. Todos os resultados das repetições são independentes uns dos outros.

**Exemplo 1.** Num rebanho bovino 30% dos animais estão atacados de febre aftosa. Retira-se ao acaso, uma amostra de 10 animais.

- 1) Verifique se a variável "número de animais doentes" pode ser estudada pelo modelo binomial. Justifique.
- 2) Estruturar a função de probabilidade e representar a distribuição de probabilidade num gráfico.
- 3) Qual a probabilidade de se encontrar 6 animais doentes?

Primeiramente vamos verificar se a variável  $X$ : número de animais com febre aftosa, pode ser estudada pelo modelo binomial.

- 1) Temos  $n = 10$  animais, então  $X : 0, 1, 2, \dots, 10$ .
  - 2) Uma animal está ou não está com febre aftosa.
  - 3) A probabilidade para cada animal, de ter febre aftosa, é constante.
  - 4) Os 10 animais são selecionados aleatoriamente, ao acaso, isso garante a independência.
- Assim,

$$X : b(10; 0,30).$$

Temos:

$$\begin{aligned}\pi &= 0,30 \\ 1 - \pi = q &= 1 - 0,30 = 0,70 \\ n &= 10\end{aligned}$$

portanto, a estrutura da função de probabilidade é dada por:

$$P(X = k) = \binom{10}{k} 0,30^k 0,70^{10-k}.$$

No apêndice 1 temos a tabela das probabilidades binomiais. Como sabemos, estas probabilidades são dadas em função do número de repetições e da proporção de sucessos. Para uma distribuição binomial, temos que  $b(10;0,30)=0,037=3,7\%$ . Também, obtemos diretamente



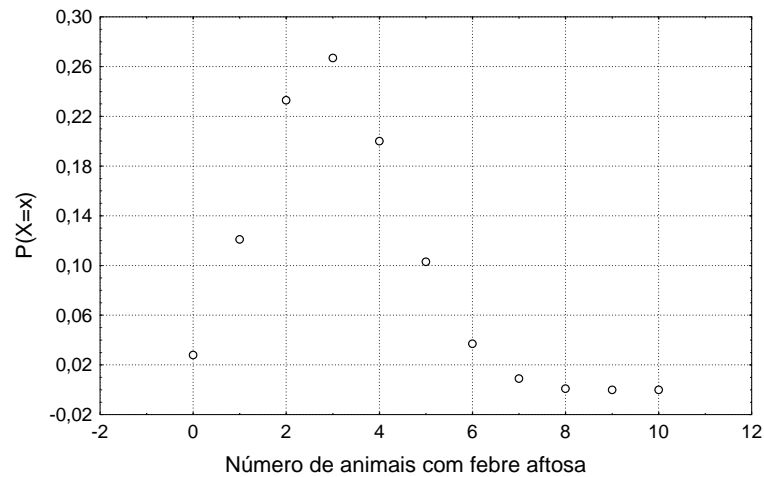


Figura 4.3: Distribuição de probabilidade para animais com febre aftosa

da tabela da distribuição binomial, que:

$$\begin{aligned}
 P(X = 0) &\longrightarrow b(0 : 10; 0,30) = 0,028 \\
 P(X = 1) &\longrightarrow b(1 : 10; 0,30) = 0,121 \\
 P(X = 2) &\longrightarrow b(2 : 10; 0,30) = 0,233 \\
 P(X = 3) &\longrightarrow b(3 : 10; 0,30) = 0,267 \\
 P(X = 4) &\longrightarrow b(4 : 10; 0,30) = 0,200 \\
 P(X = 5) &\longrightarrow b(5 : 10; 0,30) = 0,103 \\
 P(X = 6) &\longrightarrow b(6 : 10; 0,30) = 0,037 \\
 P(X = 7) &\longrightarrow b(7 : 10; 0,30) = 0,009 \\
 P(X = 8) &\longrightarrow b(8 : 10; 0,30) = 0,001 \\
 P(X = 9) &\longrightarrow b(9 : 10; 0,30) = 0,000 \\
 P(X = 10) &\longrightarrow b(10 : 10; 0,30) = 0,000
 \end{aligned}$$

O gráfico da distribuição de probabilidade é dado na Figura 4.3.

A probabilidade de encontrarmos 6 animais doentes, isto é, de  $k = 6$  vale:

$$\begin{aligned}
 P(X = 6) &= \binom{10}{6} (0,30)^6 (0,70)^4 = 210 \times 0,00073 \times 0,2401 \\
 P(X = 6) &= 0,037 \\
 P(X = 6) &= 3,7\%.
 \end{aligned}$$

**Exemplo 2.** Numa criação de coelhos, a taxa de nascimento de machos é de 40%. Qual a probabilidade de que nasçam pelo menos 2 coelhos machos, num dia em que nasceram 19 coelhos? Estruturar a função de probabilidade e representar a distribuição graficamente.

Primeiramente, vamos verificar se a variável  $X$ : número de coelhos machos, pode ser estudada pelo modelo binomial.

- 1) Temos  $n=19$  repetições do experimento.
- 2) Um coelho é macho ou é fêmea.
- 3) A probabilidade de ser macho é 0,40 e permanece constante.
- 4) Os resultados (macho e fêmea), em cada nascimento, são independentes.

Assim:

$$\begin{aligned} X &= 0, 1, 2, 3, \dots, 19 \\ \pi &= 0,40 \end{aligned}$$

logo:

$$X : b(19; 0,40).$$

Portanto, a estrutura da função de probabilidade é dada por:

$$P(X = k) = \binom{19}{k} 0,40^k 0,60^{19-k}.$$

A probabilidade desejada,  $P(X \geq 2)$ , é dada por:

$$\begin{aligned} P(X \geq 2) &= 1 - P(X < 2) = 1 - [P(X = 0) + P(X = 1)] \\ &= 1 - \left[ \binom{19}{0} (0,40)^0 (0,60)^{19} + \binom{19}{1} (0,40)^1 (0,60)^{18} \right] \\ &= 1 - (0,000 + 0,001) \\ &= 0,999. \end{aligned}$$

As probabilidades para  $X : b(19; 0,40)$ , valem:

$$\begin{aligned} P(X = 0) &= 0,000 & P(X = 10) &= 0,098 \\ P(X = 1) &= 0,001 & P(X = 11) &= 0,053 \\ P(X = 2) &= 0,005 & P(X = 12) &= 0,024 \\ P(X = 3) &= 0,017 & P(X = 13) &= 0,008 \\ P(X = 4) &= 0,047 & P(X = 14) &= 0,002 \\ P(X = 5) &= 0,093 & P(X = 15) &= 0,001 \\ P(X = 6) &= 0,145 & P(X = 16) &= 0,000 \\ P(X = 7) &= 0,180 & P(X = 17) &= 0,000 \\ P(X = 8) &= 0,180 & P(X = 18) &= 0,000 \\ P(X = 9) &= 0,146 & P(X = 19) &= 0,000. \end{aligned}$$

O gráfico da distribuição de probabilidade é dado na Figura 4.4.

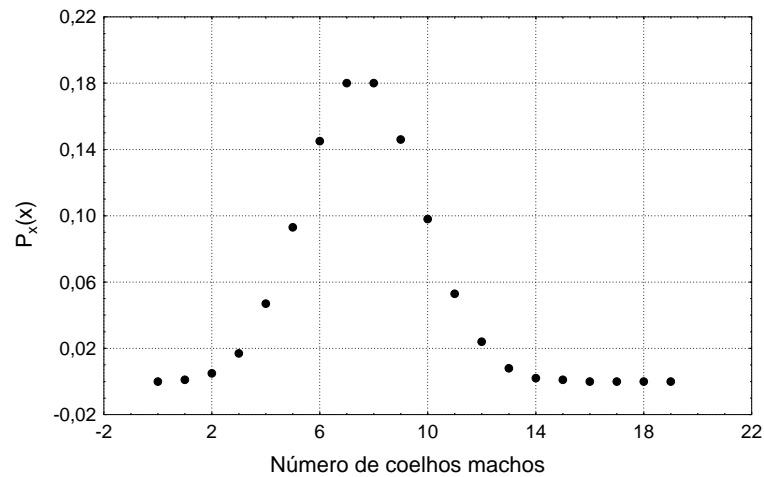


Figura 4.4: Distribuição de probabilidade para animais com febre aftosa

### *A Média e Variância de Uma Variável Binomial*

Uma variável aleatória binomial  $X$  é a soma de  $n$  variáveis *independentes* do tipo Bernoulli ( $Y$ ),

$$X = Y_1 + Y_2 + Y_3 + \dots + Y_n.$$

Aplicando as propriedades da esperança matemática e variância, e lembrando que  $E(Y_i) = \pi$  e  $Var(Y_i) = \pi(1 - \pi)$ , temos:

$$\begin{aligned} E(X) &= E(Y_1 + Y_2 + Y_3 + \dots + Y_n) \\ E(X) &= E(Y_1) + E(Y_2) + E(Y_3) + \dots + E(Y_n) \\ E(X) &= \pi + \pi + \pi + \dots + \pi \\ E(X) &= n\pi. \end{aligned}$$

No exemplo dos animais com febre aftosa, temos:

$$\begin{aligned} E(X) &= 10 \times 0,30 \\ E(X) &= 3. \end{aligned}$$

Portanto, em média, esperamos encontrar três animais com aftosa, dentre os dez selecionados.

A variância fica:

$$Var(X) = Var(Y_1 + Y_2 + Y_3 + \dots + Y_n),$$

como os ( $Y$ 's) são independentes, a variância de uma soma de variáveis aleatórias é a soma

das variâncias dessas variáveis, então:

$$\begin{aligned} Var(X) &= Var(Y_1) + Var(Y_2) + \dots + Var(Y_n) \\ Var(X) &= \pi(1 - \pi) + \pi(1 - \pi) + \dots + \pi(1 - \pi) \\ Var(X) &= n\pi(1 - \pi). \end{aligned}$$

No exemplo dos animais com febre aftosa, temos:

$$Var(X) = n\pi(1 - \pi) = 10 \times 0,30 \times 0,70 = 2,1 \text{ animais}^2.$$

O desvio padrão vale:

$$D.P.(X) = \sqrt{Var} = \sqrt{2,1} = 1,449 \text{ animais}.$$

Uma variável aleatória binomial, como vimos, é obtida de contagens, frequências de sucessos, todavia, em muitas situações, é importante expressar os dados na forma de *proporção*, por exemplo, representar que 18 sementes germinaram em 20 sementes testadas, corresponde a  $p = 18 \div 20 = 0,90$ . Neste caso (de proporções) a média e a variância são dadas por:

$$E(P) = \pi.$$

Demonstração:

$$\begin{aligned} E(P) &= E\left(\frac{X_1}{n}\right) = \frac{1}{n}E(X_1) = \frac{n\pi}{n} = \pi. \\ Var(P) &= \frac{\pi(1 - \pi)}{n}. \end{aligned}$$

Demonstração:

$$Var(P) = Var\left(\frac{X_1}{n}\right) = \frac{1}{n^2}Var(X_1) = \frac{1}{n^2}n\pi(1 - \pi) = \frac{\pi(1 - \pi)}{n}.$$

As probabilidades das proporções são exatamente iguais as probabilidades das contagens. Para o exemplo dos animais com febre aftosa, a probabilidade da proporção de animais com febre aftosa, numa amostra de 10 animais, ser igual a 0,30 é igual a 0,267, ou seja, exatamente igual a probabilidade de encontrarmos 3 animais com febre aftosa,  $P(X = 3) = 0,267$ . Assim, a distribuição de probabilidades para a proporção de animais com febre aftosa é dada por:

P(X=0)	P(X=1)	P(X=2)	P(X=3)	P(X=4)	P(X=5)	P(X=6)	P(X=7)	P(X=8)	P(X=9)	P(X=10)
0,028	0,121	0,233	0,267	0,200	0,103	0,037	0,009	0,001	0,000	0,000

## 4.7 A Distribuição de Poisson

Esta distribuição é largamente utilizada para *contagens* de indivíduos, plantas, colônias de bactérias, itens, objetos, dados num *intervalo de tempo*, numa *área*, num *volume*, num *comprimento*. A unidade de medida deve ser definida de tal modo que as contagens sejam baixas. Considera-se um número baixo como sendo menor que 10. Exemplos de variáveis Poisson: número de indivíduos por quadrante de  $1m^2$ ; número de colônias de bactérias de uma dada cultura por  $0,01\text{ mm}^2$  numa plaqueta de microscópio; número de defeitos por 100 *m* de tecido; número de acidentes numa esquina movimentada e bem sinalizada por dia; número de chamadas telefônicas numa central de PABX num intervalo de tempo de  $\frac{1}{2}$  minuto; número de partículas radioativas emitidas numa unidade de tempo; e número de micronúcleos/1000 células.

Uma aplicação importante da distribuição de Poisson<sup>5</sup> na área biológica, diz respeito ao estudo do padrão de dispersão de uma certa espécie vegetal ou animal num campo ou floresta, enfim, numa determinada área. Portanto, esta distribuição é muito utilizada nas disciplinas de dinâmica de populações e entomologia.

Inicialmente precisamos considerar que os indivíduos da espécie se distribuem *aleatoriamente* no campo. Isto significa que a posição de um indivíduo é completamente *independente* da posição de qualquer outro indivíduo da população, como mostra a Figura 4.5<sup>6</sup>.

Para o estudo da dispersão, supõe-se que uma grade é sobreposta à área de estudo, formando um grande número de quadrantes (quadrados ou retângulos). A ilustração é dada na Figura 4.7. Podemos contar o número de quadrantes com zero indivíduos, com um indivíduo, com dois indivíduos, e assim por diante. Observe que não há tendência de quadrantes com maior ou menor número de indivíduos e que o número de indivíduos num quadrante é *independente* do número de indivíduos em outro quadrante.

Pode este padrão de dispersão aleatória ser descrito matematicamente? A resposta é sim, e é feita através de um modelo, cuja função de probabilidade, ou seja, a probabilidade de encontrar  $X$  indivíduos por quadrante, é dada por:

$$P(X = k) = \frac{e^{-\lambda}(\lambda)^k}{k!} \quad k = 0, 1, 2, 3, \dots \quad (4.3)$$

Onde  $e$  é o número de Euler e vale 2,718282 e  $\lambda$  é o número médio de indivíduos por

---

<sup>5</sup>O nome desta distribuição está associado ao matemático francês S. Denis Poisson, (1781-1840).

<sup>6</sup>Outras duas formas de distribuições encontradas na natureza são a uniforme e por contágio, mostradas na figura 4.6. Esta última é mais comum de ocorrer, neste caso, a distribuição de probabilidade utilizada é a binomial negativa.

# Habitat

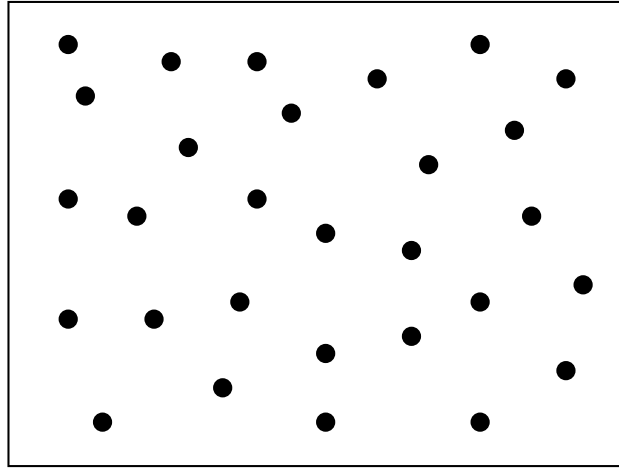
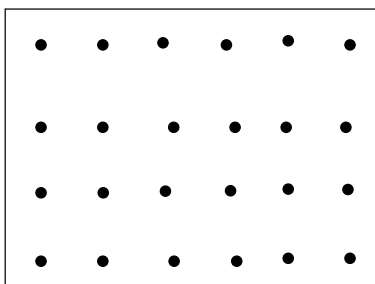
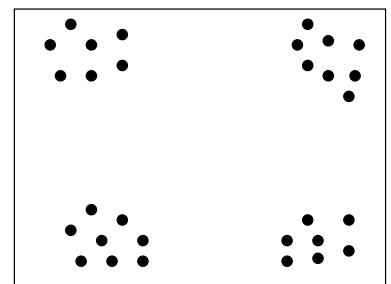


Figura 4.5: Distribuição aleatória dos indivíduos de uma determinada espécie numa área



(a) Distribuição uniforme



(b) Distribuição por contágio

Figura 4.6: Outras formas de distribuições

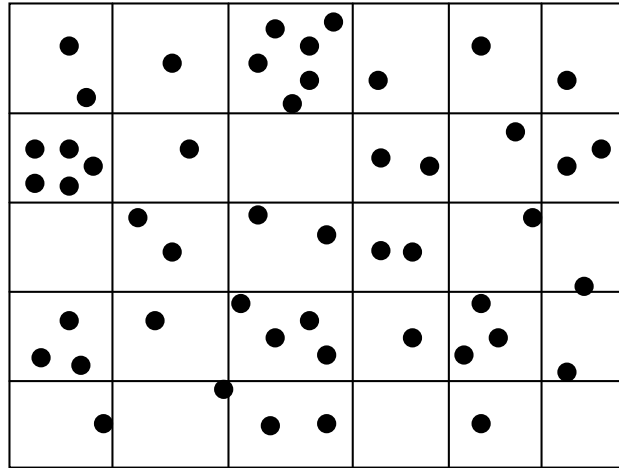


Figura 4.7: Distribuição em quadrantes

quadrante. O  $k!$  representa "k fatorial", por exemplo,  $3!=3.2.1=6$ . O  $0!$  é definido como sendo igual a 1. As probabilidades de encontrar  $X$  plantas por quadrante dadas pela função 4.3, juntamente com os valores de  $k = 0, 1, 2, 3, \dots$  constituem a distribuição de Poisson. Probabilidades, para alguns valores de  $\lambda$  são dadas no Apêndice 2.

Se o número de quadrantes com  $k$  indivíduos aproximam-se dos valores esperados de quadrantes com  $k$  indivíduos pelo modelo de Poisson, então a dispersão da espécie é *aleatória*. Este é um importante resultado para os Biólogos.

**Exemplo 1.** Numa área dividida em quadrantes de  $0,50m^2$ , foram encontradas em média 2,5 espécimes. Considerando que o modelo de Poisson é adequado, qual é a probabilidade de se encontrar num quadrante exatamente 4 espécimes? Seja  $X$  o número de espécimes por  $0,50m^2$ .

$$P(X = 4) = \frac{e^{-2,5}(2,5)^4}{4!} = 0,1336$$

Qual é a probabilidade de se encontrar no máximo 1 espécime por quadrante?

$$P(X = 0) = \frac{e^{-2,5}(2,5)^0}{0!} = 0,0821.$$

$$P(X = 1) = \frac{e^{-2,5}(2,5)^1}{1!} = 0,2052.$$

Portanto, a resposta é dada por:  $0,0821+0,2052=0,2873$ .

**Exemplo 2.** Numa placa de microscópio, dividida em quadrantes de  $1mm^2$ , encontra-se em média 5 colônias por  $mm^2$ . Considerando que a distribuição de Poisson é adequada, ou seja: 1) as colônias distribuem-se aleatoriamente na placa e, 2) o número médio de colônias

por  $mm^2$  permanece constante e é baixo, determine a probabilidade de um quadrante ter exatamente 1 colônia. Como  $\lambda = 5$ , temos:

$$P(X = 1) = \frac{e^{-5}(5)^1}{1!} = 0,0337.$$

Qual a probabilidade de se encontrar pelo menos 2 colônias por  $mm^2$ ? Pode-se calcular esta probabilidade como:

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)]$$

$$P(X = 0) = \frac{e^{-5}(5)^0}{0!} = 0,0067.$$

Portanto,

$$P(X \geq 2) = 1 - (0,0067 + 0,03369) = 0,9596.$$

Qual a probabilidade de se encontrar 8 colônias em  $2mm^2$ ? Neste caso,  $\lambda \times A = 5 \times 2 = 10$  colônias. Assim:

$$P(Y = k) = \frac{e^{-\lambda A}(\lambda A)^k}{k!}$$

onde  $\lambda$  é o número médio de colônias por  $1mm^2$  e  $A$  é a área de tamanho fixo. O resultado fica:

$$P(X = 8) = \frac{e^{-10}(10)^8}{8!} = 0,1126.$$

## 4.8 Aproximação da Distribuição Binomial pela Distribuição de Poisson

O modelo de Poisson pode ser considerado como limite da distribuição binomial, isto é, para determinados valores de  $n$  **grande** (fazendo-se  $n$  cada vez maior) e  $\pi$  **pequeno** (fazendo-se  $\pi$  cada vez menor), verifica-se a seguinte aproximação:

$$b(k : n, \pi) = \binom{n}{k} \pi^k (1 - \pi)^{n-k} \cong \frac{e^{-\lambda} (\lambda)^k}{k!}, \text{ com } k = 0, 1, 2, \dots \quad (4.4)$$

com um único parâmetro  $\lambda = n\pi$ , interpretado como a frequência média de ocorrência de um particular valor  $k$ . A distribuição limite é chamada de distribuição de Poisson. Para saber se a aproximação é boa, uma recomendação prática é verificar se a desigualdade  $n\pi \leq 7$  é válida.

Na Tabela 4.1 estão apresentadas algumas situações do cálculo das probabilidades usando os dois modelos a qual passamos a fazer algumas interpretações. O cálculo das probabilidades de Poisson foram calculadas para cinco valores de frequências médias, quais sejam:  $\lambda = 1$ ,



Tabela 4.1: Aproximação da distribuição binomial pela distribuição de Poisson, variando o tamanho da amostra  $n$  e as probabilidades  $\pi$

Distribuição	$n$	$\pi$	$k$								
			0	1	2	3	4	5	6	7	$\geq 8$
Binomial	500	0,0020	0,367511	0,368248	0,184124	0,061252	0,015252	0,003032	0,000501	0,000071	0,000010
	1000	0,0010	0,367695	0,368063	0,184032	0,061283	0,015290	0,003049	0,000506	0,000072	0,000010
	2000	0,0005	0,367787	0,367971	0,183986	0,061298	0,015309	0,003057	0,000509	0,000072	0,000010
	5000	0,0002	0,367843	0,367916	0,183958	0,061307	0,015321	0,003062	0,000510	0,000073	0,000010
	10000	0,0001	0,367861	0,367898	0,183949	0,061310	0,015324	0,003064	0,000510	0,000073	0,000010
Poisson	$\lambda = 1$		0,367879	0,367879	0,183940	0,061313	0,015328	0,003066	0,000511	0,000073	0,000010
Binomial	500	0,0040	0,134794	0,270670	0,271213	0,180809	0,090223	0,035944	0,011909	0,003375	0,001063
	1000	0,0020	0,135065	0,270670	0,270942	0,180628	0,090223	0,036017	0,011970	0,003406	0,001080
	2000	0,0010	0,135200	0,270671	0,270806	0,180537	0,090223	0,036053	0,012000	0,003422	0,001088
	5000	0,0004	0,135281	0,270671	0,270725	0,180483	0,090224	0,036075	0,012018	0,003431	0,001093
	10000	0,0002	0,135308	0,270671	0,270698	0,180465	0,090224	0,036082	0,012024	0,003434	0,001095
Poisson	$\lambda = 2$		0,135335	0,270671	0,270671	0,180447	0,090224	0,036089	0,01203	0,003437	0,001097
Binomial	500	0,0100	0,006570	0,033184	0,083631	0,140230	0,175995	0,176351	0,146959	0,104759	0,132320
	1000	0,0050	0,006654	0,033437	0,083929	0,140303	0,175731	0,175908	0,146590	0,104602	0,132848
	2000	0,0025	0,006696	0,033563	0,084077	0,140339	0,175599	0,175687	0,146406	0,104523	0,133110
	5000	0,0010	0,006721	0,033639	0,084165	0,140360	0,175520	0,175555	0,146296	0,104476	0,133267
	10000	0,0005	0,006730	0,033664	0,084195	0,140367	0,175494	0,175511	0,146259	0,104461	0,133319
Poisson	$\lambda = 5$		0,006738	0,033690	0,084224	0,140374	0,175467	0,175467	0,146223	0,104445	0,133372
Binomial	500	0,0140	0,000868	0,006161	0,021827	0,051447	0,090763	0,127841	0,149753	0,150057	0,401284
	1000	0,0070	0,000890	0,006272	0,022084	0,051789	0,090996	0,127780	0,149377	0,149527	0,401286
	2000	0,0035	0,000901	0,006327	0,022213	0,051960	0,091112	0,127748	0,149189	0,149264	0,401286
	5000	0,0014	0,000907	0,006361	0,022290	0,052061	0,091181	0,127729	0,149077	0,149107	0,401286
	10000	0,0007	0,000910	0,006372	0,022315	0,052095	0,091203	0,127723	0,149040	0,149055	0,401286
Poisson	$\lambda = 7$		0,000912	0,006383	0,022341	0,052129	0,091226	0,127717	0,149003	0,149003	0,401286
Binomial	500	0,0200	0,000041	0,000419	0,002132	0,007221	0,018310	0,037069	0,062413	0,089889	0,782507
	1000	0,0100	0,000043	0,000436	0,002200	0,007393	0,018614	0,037453	0,062737	0,089987	0,781137
	2000	0,0050	0,000044	0,000445	0,002235	0,007480	0,018765	0,037644	0,062897	0,090034	0,780457
	5000	0,0020	0,000045	0,000450	0,002256	0,007532	0,018856	0,037758	0,062992	0,090061	0,780050
	10000	0,0010	0,000045	0,000452	0,002263	0,007549	0,018886	0,037795	0,063024	0,090070	0,779915
Poisson	$\lambda = 10$		0,000045	0,000454	0,002270	0,007567	0,018917	0,037833	0,063055	0,090079	0,779779

$\lambda = 2$ ,  $\lambda = 5$ ,  $\lambda = 8$  e  $\lambda = 10$ . O cálculo das probabilidades binomiais foi realizado com combinações de  $n$  e  $\pi$ , cujos produtos fossem iguais as médias da Poisson. Para  $n = 500$ ,  $\pi = 0,0020$  ( $\lambda = 1$ ) e  $k = 4$ , observamos um erro relativo, ou seja, um desvio relativo da Poisson para a binomial de 0,50%, o que já é um valor considerável, indicando que a aproximação não é excelente para este tamanho de amostra. Para amostras de tamanhos iguais a 5.000 e 10.000, tal que  $\lambda = 1$ , os erros relativos para  $k = 0, 1, 2, 3, 4, 5, 6, 7$ , foram todos abaixo de 0,30%, valores baixos, indicando boa aproximação. Passando para uma média  $\lambda = 10$ , só encontramos erros relativos abaixo de 0,40% para  $n = 10.000$  e  $k = 1, 2, 3, 4, 5, 6, 7$ . Isto demonstra que a recomendação prática de se utilizar a aproximação para  $\lambda \leq 7$  é procedente. Observa-se, também, que os erros relativos para valores de  $k$  próximos de  $\lambda$  são baixos para amostras de tamanho mínimo igual a 1.000, para as médias tomadas como exemplos..

**Exemplo 1:** Mil pássaros tem a "chance", um de cada vez, de escolher entre duzentas gaiolas dispostas em um círculo. Admitimos que os pássaros não têm preferência direcional. Vamos definir a variável de interesse como sendo  $X$ ="número de pássaros por gaiola". Qual

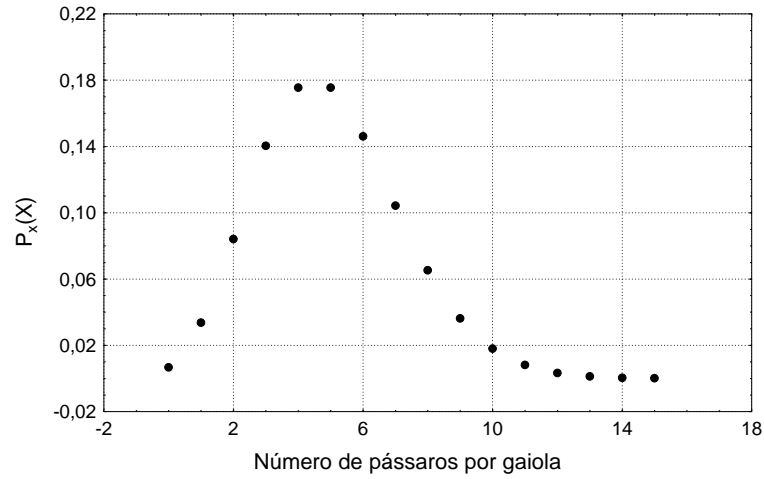


Figura 4.8: Gráfico da distribuição de probabilidades do número de pássaros

é a probabilidade de que uma gaiola específica seja escolhida 0, 1, 2, ..., vezes?

Como  $n = 1000$ ,  $\pi = 1/200 = 0,005$  e  $n\pi = 1000 \times 0,005 = 5 \leq 7$  podemos usar a distribuição de Poisson como aproximação. Temos que  $P(X = 0)$  é a probabilidade de uma gaiola não ser escolhida por nenhum pássaro, e assim por diante. Temos que:

$$\begin{aligned}
 P(X = 0) &= \frac{e^{-5}(5)^0}{0!} = \frac{0,00674 \times 1}{1} = 0,0067 \\
 P(X = 1) &= \frac{e^{-5}(5)^1}{1!} = \frac{0,00674 \times 5}{1} = 0,0337 \\
 P(X = 2) &= \frac{e^{-5}(5)^2}{2!} = \frac{0,00674 \times 25}{1 \times 2} = 0,0842 \\
 P(X = 3) &= \frac{e^{-5}(5)^3}{3!} = \frac{0,00674 \times 125}{1 \times 2 \times 3} = 0,1404 \\
 P(X = 4) &= \frac{e^{-5}(5)^4}{4!} = \frac{0,00674 \times 625}{1 \times 2 \times 3 \times 4} = 0,1755.
 \end{aligned}$$

Para os demais casos, temos:

$$\begin{aligned}
 P(X = 5) &= 0,1755 & P(X = 11) &= 0,0082 \\
 P(X = 6) &= 0,1462 & P(X = 12) &= 0,0034 \\
 P(X = 7) &= 0,1044 & P(X = 13) &= 0,0013 \\
 P(X = 8) &= 0,0653 & P(X = 14) &= 0,0005 \\
 P(X = 9) &= 0,0363 & P(X = 15) &= 0,0002 \\
 P(X = 10) &= 0,0181 & P(X \geq 16) &= 0.
 \end{aligned}$$

A distribuição de probabilidades, de acordo com a distribuição de Poisson, é dada na Figura 4.8.

**Exemplo 2:** a probabilidade de um indivíduo ter reação negativa a certa injeção é de 0,001. Suponha que 2.000 indivíduos receberam a injeção. Estruturar a função de probabilidade.

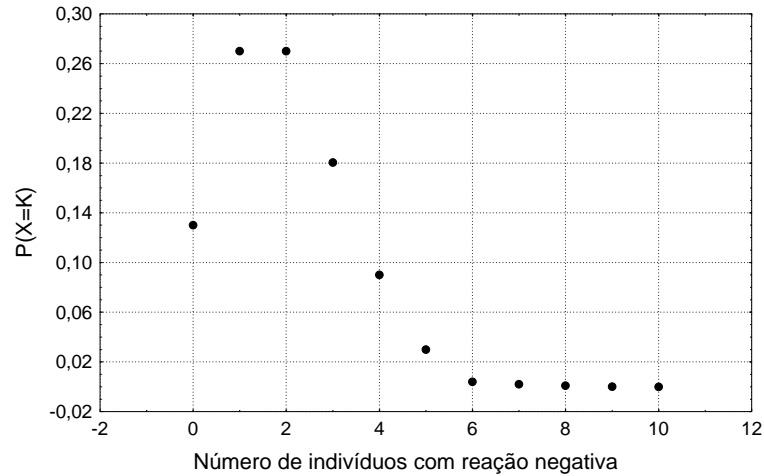


Figura 4.9: Distribuição de probabilidade para reação negativa

idade e fazer o gráfico da distribuição. Determinar a probabilidade de que, (a) exatamente três, (b) mais que quatro tenham reação negativa.

Com as informações do problema, temos:

$$\pi = 0,001 \quad n = 2000 \quad \lambda = 2000 \times 0,001 = 2$$

Assim:

$$P(X = k) = \frac{e^{-2}2^k}{k!} \quad \text{para } k = 0, 1, 2, 3, \dots$$

A distribuição de probabilidades de acordo com o modelo de Poisson é dada na Figura 4.9. A probabilidade de exatamente 3 apresentarem reação negativa é:

$$P(X = 3) = \frac{e^{-2}2^3}{3!} = 0,1804 = 18,4\%.$$

A probabilidade de mais de 4 apresentarem reação negativa é:

$$\begin{aligned} P(X > 4) &= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)] \\ &= 1 - [0,1353 + 0,2707 + 0,2707 + 0,1804 + 0,0902] = 1 - 94,73\% = 5,27\%. \end{aligned}$$

## *A Média e a Variância de Uma Variável Poisson*

Demonstra-se que:

$$E(X) = Var(X) = \lambda.$$

Portanto, na distribuição de Poisson, a média é igual a variância. Uma justificativa é que fazendo-se  $\pi$  tender a zero, implica que,  $q = 1 - \pi$ , tende para 1. Como a média,  $\lambda = n\pi$ , é um valor considerado fixo e finito, temos que a variância  $\sigma^2 = n\pi(1 - \pi)$ , tende para  $\sigma^2 = n\pi$ .

No exemplo dos pássaros:

$$E(X) = n\pi = \lambda \implies \text{média de 5 pássaros por gaiola.}$$

$$Var(X) = 5 \text{ pássaros}^2.$$

$$D.P.(X) = \sqrt{5} = 2,23 \text{ pássaros/gaiola.}$$

### Exercícios resolvidos

1) Num certo ano, o IBAMA registrou no litoral catarinense (área de reserva), 18 mortes de golfinhos.

a) Qual é a probabilidade de, num determinado mês do próximo ano, ocorrerem menos de 2 mortes?

b) Qual é a probabilidade de ocorrerem 2 mortes no próximo semestre?

a)

$$\lambda = 18/12 = 1,5 \text{ golfinhos/mês}$$

$$P(X = 0) = (e^{-1,5} 1,5^0)/0! = 0,223130 \times 1 = 0,22313$$

$$P(X = 1) = (e^{-1,5} 1,5^1)/1! = 0,334695$$

$$P(X = 0) + P(X = 1) = 0,22313 + 0,334695 = 0,557825.$$

b)

$$\lambda = 18/2 = 9 \text{ golfinhos}$$

$$P(X = 2) = (e^{-9} 9^2)/2! = 0,0050$$

2) Suponha que 80% de uma criação de suínos esteja atacada por leptospirose. Se uma amostra de 1000 suínos for examinada por um veterinário e seja  $Y$  o número de porcos doentes dessa amostra de 1000:

a) qual é o valor esperado de  $Y$ ?

b) qual é o desvio padrão de  $Y$ ?

Sendo:

$$n = 1000 \quad \pi = 0,80 \quad (1 - \pi) = 0,20$$

a)  $E(Y) = n\pi = 800$  suínos.

b)  $D.P.(Y) = \sqrt{n\pi(1 - \pi)} = \sqrt{160} = 12,649$  suínos.

$Var(Y) = 160$  suínos<sup>2</sup>.

3) Doze pares de animais experimentais são submetidos a duas dietas diferentes, A e B. A atribuição da dieta aos animais de cada par é feita ao acaso, aleatoriamente (princípio da

casualização). Após o experimento acha-se a diferença em ganho de peso entre os animais submetidos a dieta A e a dieta B. Se a diferença for positiva ( $A - B > 0$ ), o resultado será chamado de sucesso.

a) Verifique se é um experimento binomial (atender as 4 suposições).

b) Calcule  $E(X)$ ,  $Var(X)$  e  $D.P.(X)$ .

c) Qual a probabilidade de que pelo menos 9 sucessos ocorram se as duas dietas não possuem diferenças reais no que diz respeito a variável ganho de peso?

a)

- Têm-se  $n = 12$  repetições do experimento.
- $\pi$  constante, em cada repetição a probabilidade de sucesso é de 50%.
- Só há dois tipos de resultados possíveis.
- Os resultados das repetições são independentes.

b)

$$\begin{aligned} E(X) &= 12 \times \frac{1}{2} = 6 \\ Var(X) &= n\pi(1 - \pi) = 12 \times \frac{1}{2} \times \frac{1}{2} = 3 \\ D.P.(X) &= \sqrt{3} = 1,73. \end{aligned}$$

c)

$$\pi = 0,50 = 50\% \quad (1 - \pi) = 0,50 = 50\%.$$

$$\binom{12}{9} (0,50)^9 (0,50)^3 = 0,0540$$

$$\binom{12}{10} (0,50)^{10} (0,50)^2 = 0,0160$$

$$\binom{12}{11} (0,50)^{11} (0,50)^1 = 0,0030$$

$$\binom{12}{12} (0,50)^{12} (0,50)^0 = 0,000.$$

$$0,0540 + 0,0160 + 0,0030 + 0,000 = 0,073 = 73\%$$

4) Na Tabela 4.2,  $X$  significa número de filhos homens em famílias com 12 filhos. Calcule para cada valor da variável o número de famílias que você deveria esperar se  $X : b(12; 0,5)$  A probabilidade de encontrarmos 0 homens numa família é dada por:

$$P(X = 0) = \binom{12}{0} 0,50^0 0,50^{12-0} = 0,000.$$

Tabela 4.2: Número de famílias com  $X$  homens

X	Nº observado de famílias	Probabilidades	Nº esperado
0	5	0,000	0
1	35	0,003	32
2	180	0,016	171
3	599	0,054	577
4	1250	0,121	1294
5	1990	0,193	2063
6	2400	0,226	2416
7	2060	0,193	2063
8	1350	0,121	1294
9	600	0,054	577
10	179	0,016	171
11	35	0,003	32
12	7	0,000	0
Total	10.690	1,000	10690

A probabilidade de encontrarmos 1 homem numa família é dada por:

$$P(X = 1) = \binom{12}{1} 0,50^1 0,50^{12-1} = 0,003.$$

A probabilidade de encontrarmos 2 homem numa família é dada por:

$$P(X = 2) = \binom{12}{2} 0,50^2 0,50^{12-2} = 0,016.$$

Essas probabilidades e as demais são dadas na terceira coluna da Tabela 4.2. O número esperado de famílias com zero filhos homens, pela binomial, vale:  $10690 \times 0 = 0$ ; o número esperado de famílias com 1 filho homem, pela binomial, vale:  $10690 \times 0,003 = 32$ , e assim para os demais valores de  $X$ . Esses valores são encontrados na quarta coluna da Tabela 4.2.

Você acha que o modelo binomial é razoável para explicar o fenômeno? Justifique.  
Sim. Pois as frequências esperadas são próximas das frequências observadas. Apresentam o mesmo comportamento, regularidade.

## 4.9 Exercícios Propostos

1) Considerando que a probabilidade de nascer uma criança do sexo feminino é de 47%, qual a probabilidade de que uma família de 6 filhos seja constituída por seis crianças do sexo feminino?

2) Suponhamos que a porcentagem de germinação de sementes de feijão seja de 70%. Vão ser semeadas 4 sementes por cova, as quais serão espaçadas de 0,40m entre linhas e 0,20m entre covas. Supondo-se que cada canteiro a ser semeado conste de 6 linhas de 5m de comprimento, qual o número médio esperado de covas falhadas (nem uma semente germinou, das quatro semeadas) por canteiro?

3) Um contador eletrônico de bactérias registra, em média, 5 bactérias por  $\text{cm}^3$  de um líquido. Admitindo-se que esta variável tenha distribuição de Poisson;

a) qual é o desvio padrão do número de bactérias por  $\text{cm}^3$ ?

b) Encontre a probabilidade de que pelo menos duas bactérias ocorram num volume de líquido de  $1\text{cm}^3$ .

4) Em 30 bezerros nascidos, 60% foi do sexo feminino. Calcular o valor esperado (esperança) e o desvio padrão esperado.

5) Se a probabilidade de um indivíduo sofrer uma reação nociva, resultante da infecção de um determinado soro é 0,0001. Determinar a probabilidade de, entre 2.000 indivíduos:

a) exatamente três sofrerem a reação; b) mais de dois sofrerem a reação.

6) Um produtor de sementes vende pacotes com 20 sementes cada. Os pacotes que apresentarem mais de uma semente sem germinar serão indenizados. A probabilidade de uma semente germinar é 0,98. a) Qual é a probabilidade de um pacote não ser indenizado?

b) Se o produtor vende 1000 pacotes, qual é o número esperado de pacotes indenizados?

c) Quando o pacote é indenizado, o produtor tem um prejuízo de 1,20 u.m. (unidades monetárias) e, se o pacote não for indenizado, ele tem um lucro de 2,50 u.m. Qual o lucro líquido esperado por pacote? d) Calcule a média e a variância da variável "número de sementes que não germinam por pacote".

7) Sabe-se que 20% dos animais submetidos a um certo tratamento não sobrevivem. Se esse tratamento foi aplicado em 20 animais e se  $X$  é o número de não sobreviventes: a) qual é o número médio esperado de animais não sobreviventes,  $\text{VAR}(X)$  e  $\text{DP}(X)$ ? b) Calcular a  $P(2 < X \leq 4)$ ; c) Calcular a  $P(X \geq 2)$ .

8) Examinaram-se 2.000 ninhadas de 5 porcos cada uma, segundo o número de machos.

Os dados estão apresentados na tabela abaixo.

Número de machos (X)	Número de ninhadas
0	20
1	360
2	700
3	680
4	200
5	40
Total	2000

a) Calcule a proporção média de machos.

b) Calcule, para  $X=0$ ,  $X=1$  e  $X=2$ , o número de ninhadas que você deve esperar se  $X : b(5, \pi)$ , onde  $\pi$  é a proporção média de machos calculada em a).

c) Faça um gráfico, onde no eixo do Y vão as probabilidades ( probabilidade de encontrar 0 machos por ninhada, 1 macho por ninhada,.....,5 machos por ninhada) e no eixo da abcissas vai o número de machos.

9) (Bussab & Morettin, 1987). Uma certa região florestal foi dividida em 109 quadrados para estudar a distribuição de *Primula simenses* selvagem. A priori, supomos que este tipo distribui-se aleatoriamente na região. A tabela abaixo indica o número de quadrados com  $X$  *Primula Simenses*; o número médio de plantas por quadrado foi de 2,2 (encontre esse valor).

X plantas por quadrado	Número de quadrados com X plantas
0	26
1	21
2	23
3	14
4	11
5	4
6	5
7	4
8	1
> 8	0
Total	109

a) Se as plantas se distribuem aleatoriamente na região, qual a probabilidade de encontrarmos pelo menos 2 Primulas num quadrante?



- b) Dê as frequências esperadas de quadrados para os valores de  $X=0$ ,  $X=1$  e  $X=2$ .
- c) Apenas comparando os resultados de b) com as frequências observadas, podemos concluir que a distribuição da espécie é aleatória?
- d) Faça um gráfico, onde no eixo das ordenadas vão as probabilidades ( probabilidade de encontrar 0 plantas por quadrado, 1 planta por quadrado,.....,8 plantas por quadrado), e no eixo das abscissas vão o número de plantas/quadrado.

10) Quando contamos eritrócitos (células sanguíneas vermelhas) pode ser utilizado um reticulado, sobre o qual uma gota de sangue é igualmente distribuída. Sob microscópio é observada uma média de 8 eritrócitos por cada quadrado. É razoável que contagens por quadrado, frequentemente, se desviem de 1 ou 2 da média 8, e ocasionalmente de 3 ou mais?

11) Oito ratos experimentais movem-se "aleatoriamente" sobre uma superfície que é subdividida em vinte quadrantes de mesmo tamanho, numerados 1, 2, 3, ...,20. Uma fotografia é tirada. Admitindo-se a distribuição de Poisson,

- a) qual é a probabilidade de que exatamente um rato seja encontrado no quadrante  $n^{o}1$ ?
- b) quantos quadrantes devem conter 0, 1, 2,...ratos?

12) Uma armadilha para pegar vespa foi testada. Somente 128 das 720 vespas atraídas pela isca puderam ser apanhadas. Portanto, a probabilidade de pegarmos uma vespa é estimada como sendo  $\pi = 0,178$ . Qual é a probabilidade de que dentre três vespas escolhidas aleatoriamente:

- a) nenhuma seja capturada,
- b) pelo menos uma seja capturada?

13) Admitindo que cada criança tenha 51% de probabilidade de ser menino:  
Determine a probabilidade de uma família de 4 filhos ter:

- a) exatamente um menino;
- b) ao menos um menino.

14) Avaliaram-se 30 parcelas (área de terra de  $20m^2$ ) de 27 plantas de mandioca cada uma, de acordo com o número de plantas doentes colhidas. Os dados estão apresentados na tabela abaixo:

Número de plantas doentes colhidas (X)	Número de parcelas com X plantas doentes
0	14
1	8
2	4
3	3
4	1
Total	30

- a) Calcule a proporção média de plantas doentes.
- b) Calcule para cada valor de  $X$ , o número de parcelas que você deve esperar se  $X : b(27, \pi)$ , onde  $\pi$  é a proporção média calculada em a).
- c) Existe uma boa aproximação entre as proporções observadas e as estimadas pelo modelo binomial?
- d) Faça o gráfico, onde no eixo do  $Y$  têm-se as probabilidades (probabilidade de encontrar 0 plantas doentes por parcela, 1 planta doente por parcela,..., 4 plantas doentes por parcela) no eixo do  $X$  têm-se o número de plantas doentes colhidas (0,1,2,3,4).

15) Determine o número esperado de meninos em uma família com 8 crianças, supondo ser a distribuição do sexo igualmente provável. Qual é a probabilidade de ocorrer o número esperado de meninos?

16) Suponhamos que o processo de esterelização para um experimento biológico compreenda  $n$  estágios diferentes, cada um com probabilidade  $\pi$  de sucesso. Se uma falha em qualquer dos estágios ocasiona contaminação, qual a probabilidade de contaminação se  $n = 10$  e  $\pi = 0,99$ ?

17) Suponha que  $X_t$  seja o número de partículas emitidas em  $t$  horas por uma fonte radioativa, tenha uma distribuição de Poisson com parâmetro  $20t$ . Qual será a probabilidade de que exatamente 5 partículas sejam emitidas durante um período de 15 minutos?

18) Supondo-se que durante o abate de suínos, em um determinado frigorífico do Oeste Catarinense, o número de suínos descartados pelo Serviço de Inspeção Federal seja uma variável aleatória com média de 90 animais por mês (30 dias). Pergunta-se, qual a probabilidade de serem descartados:

- a) quatro suínos por dia?
- b) pelo menos 2 suínos por dia?

19) Numa família de 4 filhos, seja  $X$  = número de meninos e  $Y$  = número de variações na sequência de sexos. Relacionar o espaço amostral e, então:

- a) construir a distribuição conjunta de  $X$  e  $Y$ ;
- b)  $X$  e  $Y$  são independentes?
- c) quanto vale a covariância entre  $X$  e  $Y$ ?

20) Segundo a teoria de Mendel, no enxerto de duas espécies de plantas com flores amarelas e brancas, 30% das plantas resultantes têm flor amarela. Em sete pares de plantas enxertadas, qual a probabilidade (a) de não resultar flor amarela, (b) de haver 4 ou mais plantas com flor amarela?

21) Um laboratório é contratado para fornecer, a um distribuidor, lotes de vacinas para gado. Ocasionalmente, algumas vacinas se revelam ineficazes. O distribuidor quer proteger-

se contra o risco de receber um número indesejável de vacinas ineficazes. Como não é possível testar todas as vacinas (já que o teste inutiliza a vacina), o distribuidor adota o seguinte processo de seleção: extrai de cada lote uma amostra aleatória de 10 ampolas, contendo um número  $X$  de vacinas estéreis. Se  $X=0$ , o lote é aceito. Se  $X \geq 1$ , o lote é rejeitado. Este processo é designado plano de amostragem simples com amostra de tamanho  $n = 10$  e número de aceitação  $c = 0$ . Admitamos que o tamanho do lote seja suficientemente grande para que a distribuição de  $X$  seja (aproximadamente) binomial, com  $n = 10$  e  $\pi$ =fração desconhecida de vacinas estéreis em cada lote.

a) Se  $\pi = 0,20$ , qual a probabilidade de o lote ser aceito?

b) Calcule a probabilidade,  $P(A)$ , de aceitar um lote, para  $\pi = 0,05$ ,  $\pi = 0,10$ ,  $\pi = 0,20$  e  $\pi = 0,40$ . Grafe  $P(A)$  como função de  $\pi$ , unindo os pontos do gráfico por uma curva contínua (esta curva é chamada curva característica de operação (CCO) para o plano de amostragem).

22) Sementes certificadas de feijão são vendidas em um saco de 15 Kg ao preço de 20 u.m. cada. É característica de produção que 20% das sementes apresentem poder germinativo abaixo do especificado. Um comprador fez a seguinte proposta ao produtor de sementes: de cada saco escolhe 25 sementes, ao acaso e paga por saco:

- 25,00 u.m. se todas as sementes germinarem;
- 17,00 u.m. se uma ou duas sementes não germinarem;
- 10,00 u.m. se três ou mais sementes não germinarem.

O que é melhor para o produtor, manter o seu preço de 20,00 u.m. por saco ou aceitar a proposta do comprador?

*Sugestão:* encontrar o preço médio esperado pelo produtor.

23) O número de mortes por afogamento em fins de semana, numa cidade praiana é de 2 para cada 50.000 habitantes. Qual a probabilidade de que em 200.000 habitantes ocorram 5 afogamentos?

24) Num lote de sementes cujo percentual de germinação é 0,70, foi retirada uma amostra de 10 sementes e colocada num germinador. Se  $X$  é o número de sementes germinadas:

- a) estruturar a função de probabilidade de  $X$ , e fazer o gráfico da distribuição de probabilidade;
- b) determinar a probabilidade de germinarem duas ou mais sementes;
- c) determinar a média, a variância e o desvio padrão da distribuição de  $X$ .

25) Seja  $X$  o número de crianças não imunizadas numa campanha de vacinação contra uma determinada doença, onde a probabilidade de não-imunização é  $\pi = 10^{-3}$ . De

5000 crianças vacinadas, numa determinada localidade, qual a probabilidade de não ficarem imunes:

- a) uma criança;
- b) uma criança ou mais;
- c) estruture a função de probabilidade e faça um gráfico da distribuição de probabilidade.

26) Um grafólogo diz que pode determinar o sexo de uma pessoa com base em sua letra. Para testar tal afirmação, dão-se 15 pares de amostras de caligrafias a esse técnico, cada par sendo por um texto escrito por um homem e por uma mulher. Se esse técnico for, na verdade, incapaz de determinar seus julgamentos com base em palpites, qual é a probabilidade de ele identificar corretamente a caligrafia masculina em 10 ou mais desses 15 pares?

27) Um produtor de sementes afirma que apenas 5% das sementes que vende não germinam. Um comerciante compra regularmente um lote de sementes desse produtor, mas sob a seguinte condição: ele aceita o lote se, em 10 sementes escolhidas ao acaso, no máximo uma não germinar; caso contrário o lote todo é rejeitado.

- a) Se o produtor de fato tem razão, qual a probabilidade de um lote ser rejeitado?
- b) Suponha que o produtor esteja mentindo, isto é, na verdade a proporção de sementes que não germinam é de 10%. Qual a probabilidade de um lote ser aceito, segundo o critério acima?

28) Verificou-se que o número de quebras cromossômicas em um roedor, em qualquer período de um dia, num local poluído, pode ser considerado como uma variável aleatória que tenha distribuição de Poisson com parâmetro 0,1. (Isto é, em média haverá uma quebra cromossômica a cada 10 dias). Vamos supor que este roedor ficará 20 dias neste local poluído para experiência.

- a) Qual a probabilidade de se encontrar menos de três quebras cromossômicas?
- b) Qual a probabilidade de se encontrar mais de duas quebras cromossômicas?

29) Em uma determinada região ocorre em média 1 geada a cada 5 anos. Vamos definir  $X$  como sendo o número de geadas nos próximos 5 anos.

- a) Qual é a probabilidade de ocorrência de no máximo 1 geada nos próximos 5 anos?
- b) Qual é a probabilidade de ocorrência de pelo menos duas geadas nos próximos 5 anos?

30) Um agricultor tem duas alternativas para a venda de seu produto que é fornecido em lotes de 500 sacos de sementes certificadas cada.

- a) O comprador A, que paga 8,00 u.m. por saco e não exige nenhum teste.
- b) O comprador B, que, para cada lote recebido retira 10 sacos ao acaso e os examina: se todos os 10 estão dentro das especificações, quanto ao poder germinativo, paga 5.000,00 u.m. pelo lote; se entre os 10 sacos houver um fora da especificação, paga 4.000,00 u.m. pelo lote;

e se entre os 10 houver dois ou mais sacos fora das especificações paga apenas 2.500,00 u.m. pelo lote. Sabendo o agricultor ser de 10% a porcentagem real de sacos fora da especificação que produz, qual a melhor alternativa para a venda de seu produto?

*sugestão:* calcule o preço de venda esperado para cada alternativa.

31) Um Engenheiro Agrônomo pode atender, normalmente, 4 estabelecimentos agrícolas por dia. Em média ( $\lambda$ ), ele atende 3 estabelecimentos agrícolas por dia. Se o agrônomo tiver que atender mais de 4 estabelecimentos agrícolas em um dia ele deverá fazer horas-extras para atender a essas solicitações. Qual é a probabilidade do agrônomo ter que fazer horas-extras? Sendo  $\lambda = 3$ .

32) Um produtor de camarões de água-doce tratados com uma alimentação especial, deseja comparar o sabor proporcionado por esta nova alimentação com o sabor produzido pela alimentação tradicional. A cada um de 4 provadores são fornecidos 3 porções exatamente iguais, em ordem aleatória, duas das quais com a alimentação tradicional, e a outra com alimentação especial. Cada um desses provadores é inquirido sobre a porção que prefere. Suponha que essas duas fórmulas sejam igualmente saborosas. Seja  $Y$  o número de provadores que preferem camarões tratados com alimentação especial.

a) Qual é a probabilidade de pelo menos 3 dos 4 provadores preferirem a fórmula nova?

b) Calcule o valor esperado de  $Y$ .

33) Um casal deseja ter 8 crianças. Chamamos de  $X$  a variável número de filhos homens e que a probabilidade de nascer filho do sexo masculino ou feminino é de 50% para cada um, responda:

a) Verifique se esta situação pode ser estudada pela distribuição binomial (atender as quatro suposições do modelo). b) Calcule a  $E(X)$  e  $DP(X)$ .

c) Qual a probabilidade que nasçam no mínimo 4 filhos homens?

d) Faça um gráfico, onde no eixo  $Y$  são as probabilidades (probabilidade de zero sucesso, 1 sucesso,..., 8 sucessos), e no eixo das abcissas vai o número de filhos homens.

34) O número de partículas gama, emitidas por segundo, por certa substância radioativa é uma variável aleatória com distribuição de Poisson com média  $n\pi = 3$ . Se um instrumento registrador torna-se inoperante quando há mais de 4 partículas por segundo,

a) Qual a probabilidade de isto acontecer em um dado segundo?

b) Qual é o desvio padrão dessa variável aleatória?

35) Uma fonte radioativa é observada durante 7 intervalos de tempo, cada um de dez segundos de duração. O número de partículas emitidas  $X$  durante cada período observado, tem uma distribuição de Poisson com parâmetro 5.

a) Qual é a probabilidade de quem em cada um dos 7 intervalos de tempo, 4 ou mais

partículas sejam emitidas?

b) Qual é a probabilidade de que em cada um dos sete intervalos de tempo, três ou menos partículas sejam emitidas?

c) Qual é a probabilidade de que pelo menos um dos sete intervalos, 4 ou mais partículas sejam emitidas?

36) Suponha que o registro meteorológico mostra que em média 5 dos 30 dias do mês de novembro são chuvosos.

a) Supondo a distribuição binomial com cada dia visto como uma prova independente, encontre a probabilidade de que no próximo mês de novembro no máximo 3 dias sejam chuvosos.

b) Dê razões para justificar o uso da distribuição binomial na solução de a).

37) Um certo tratamento é capaz de imunizar 78% dos animais contra uma determinada doença. Uma nova amostra de 50 animais é testada. Seja  $X$  o número de animais que se tornarão imunes.

a) Verifique se é um experimento binomial (confronte as hipóteses do modelo com a situação real).

b) Qual é a  $E(X)$ ?

c) Qual é o  $DP(X)$ ?

d) qual é a  $P(X < 2)$ ?

38) Numa campanha de vacinação contra a aftosa, um animal de cada 1000 vacinados não fica imune à moléstia. Se forem vacinados 5000 animais, qual a probabilidade de não ficarem imunes:

a) cinco animais;

b) dois animais ou mais.

39) Uma fonte mineral contém um número médio de 4 bactérias por  $cm^3$ . Dez tubos de ensaio, de  $1cm^3$ , são encheidos com este líquido. Supondo que a distribuição de Poisson é aplicável, encontre a probabilidade:

a) de que todos os 10 tubos de ensaio apresentem bactérias, isto é, contenham ao menos uma bactéria cada;

b) que exatamente 8 tubos de ensaio apresentem bactérias.

40) A experiência mostra que somente  $\frac{1}{3}$  dos pacientes tendo certa doença recupera-se sob um tratamento padrão. Uma nova droga promissora será administrada a um grupo de 12 pacientes que tem a doença. Se a clínica requer que ao menos 7 dos pacientes se recuperem antes de aceitar a nova droga como o melhor tratamento, qual é a probabilidade de que a droga seja desacreditada, mesmo quando a taxa de recuperação aumenta para  $\frac{1}{2}$ ?

41) Considere uma anomalia metabólica que atinge aproximadamente 1 em cada 100

bebês. Se 4 nascem em um hospital específico, em certo dia, qual é a probabilidade de:

a) nenhum apresentar esse problema?

b) não mais de um apresentar esse problema?

Resolva o exercício pela distribuição binomial e Poisson. Compare os resultados.

## 5 Modelos de Probabilidades para Medidas

### 5.1 Introdução

Vamos ver agora como criar um modelo probabilístico, o que é uma função densidade de probabilidade (f.d.p.), e como calcular probabilidades no caso de variáveis aleatórias contínuas.

De um modo geral, podemos dizer que as variáveis aleatórias, cujos os valores resultam de mensurações ou medições (geralmente são valores fracionários), são variáveis aleatórias contínuas. Exemplos: rendimento de milho em Kg/ha, diâmetro de *Paepalanthus*, diâmetro de *Biomphalaria*, altura de plantas, etc.

Como foi tratado na Subseção 2.1.3.8, para representar a distribuição dos dados de variáveis aleatórias contínuas construímos um histograma. Um exemplo de um histograma para a variável aleatória contínua altura de plantas de eucalipto é dado na Figura 5.1(a). A construção de modelos probabilísticos para variáveis aleatórias contínuas envolve a idéia da generalização do histograma, isto é, fazer o número de classes tender para o infinito. Fazendo-se o número de classes tender para o infinito, implica que os intervalos de classe tornam-se extremamente pequenos, podemos dizer, ínfimos. A forma do histograma sofre uma modificação, passando a apresentar uma curva suave, ou seja, uma linha contínua, como o da Figura 5.1(b). Esta curva contínua é a representação gráfica de uma função, representada por  $f_X(x)$ , chamada de *função densidade de probabilidade* (f.d.p.), equivalente as funções de probabilidades para variáveis aleatórias discretas. Esta  $f_X(x)$  é a expressão matemática do modelo probabilístico para a variável aleatória quantitativa contínua  $X$ . Na verdade, a variável aleatória  $X$  define a curva, ou seja, o modelo.

Na Figura 5.1(a), a área de um determinado retângulo, indica a porcentagem de valores dentro daquela classe. Se somarmos as áreas de todos os retângulos chegaremos a 100%. Na Figura 5.1(b), a área total sob a curva equivale a 1 ou 100% e, portanto, a área entre dois pontos, digamos  $x_1$  e  $x_2$ , e a curva, corresponde a probabilidade de se encontrar um valor entre aqueles pontos ou a proporção de valores dentro daquele intervalo.

Já foram construídos muitos modelos teóricos para variáveis aleatórias contínuas; o que devemos fazer é escolher aquele adequado para a variável aleatória em estudo. A seguir apresentamos alguns exemplos de variáveis aleatórias contínuas com suas respectivas funções.

**Exemplo 1:** num estudo de radioatividade, as cintilações (que são contagens) seguem



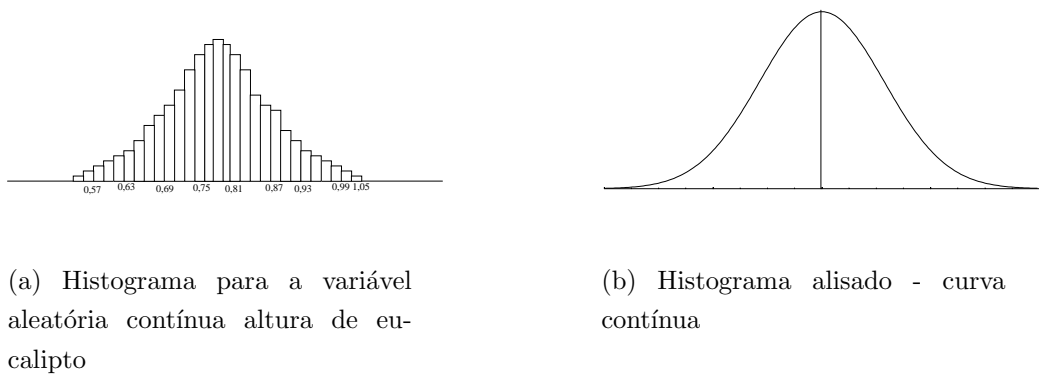


Figura 5.1: A variável aleatória contínua altura de plantas de eucalipto varia continuamente

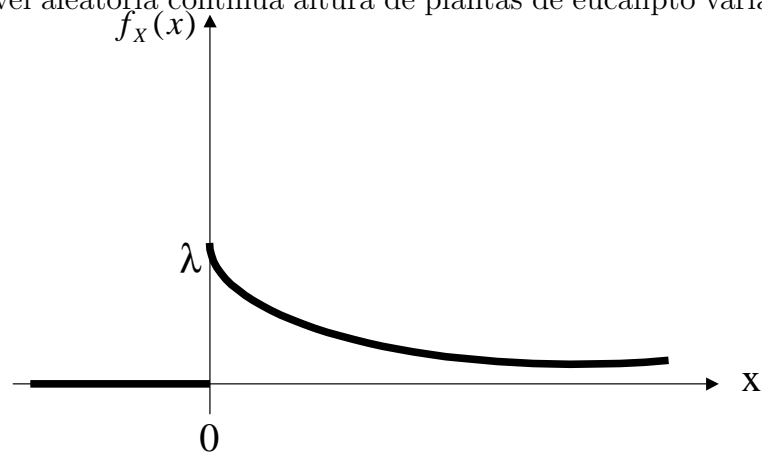


Figura 5.2: Representação gráfica do modelo exponencial

a distribuição de Poisson. Os instantes, quando os núcleos se desintegram, são representados por  $t_1, t_2, t_3, \dots, t_n$ . Em adição, os intervalos de tempo entre valores de  $t$  consecutivos,  $\Delta t_1, \Delta t_2, \dots$ , é uma variável aleatória contínua e segue o modelo exponencial.

Definição: se a função densidade de probabilidade de uma variável aleatória  $X$  for definida da seguinte maneira:

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{para } x \geq 0 \\ 0 & \text{para } x < 0 \\ \lambda > 0 \end{cases}$$

dizemos que  $X$  tem distribuição exponencial ou  $X$  segue o modelo exponencial com parâmetro  $\lambda$ , o número médio de ocorrências (igual a distribuição de Poisson). A representação gráfica desse modelo é dado na Figura 5.2.

**Exemplo 2:** num estudo de comportamento animal, pássaros são libertados um de cada vez, sob circunstâncias que tornam difícil a orientação. Espera-se que os pássaros escolham

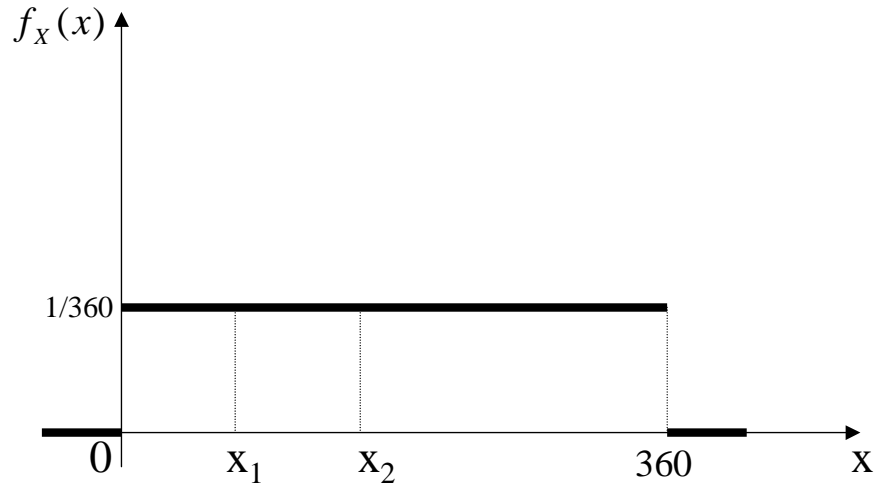


Figura 5.3: Modelo de distribuição da variável ângulo (graus)

direções aleatórias. Está-se medindo o ângulo entre o norte e a direção tomada pelo pássaro, no sentido horário (azimute). A direção é dita aleatória se cada azimute de  $0^0$  a  $360^0$  tem a mesma chance de ser escolhido.

Variável em estudo:  $X = \text{ângulo em graus}$ .

Espaço amostral:  $\Omega = \{x \in \mathbb{R} | 0 \leq x \leq 360\}$ .

Portanto, podemos estabelecer o seguinte modelo para a variável ângulo em graus  $X$ :

$$f_X(x) = \begin{cases} \frac{1}{360} & \text{para } 0 \leq x \leq 360 \\ 0 & \text{para outros casos.} \end{cases}$$

A representação gráfica desse modelo é dada na Figura 5.3.

Definição: se a função densidade de probabilidade de uma variável aleatória  $X$  é dada por:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{para } a \leq x \leq b \\ 0 & \text{para outros casos.} \end{cases}$$

$a < b$ , ambos reais.

dizemos que  $X$  tem distribuição uniforme, ou segue o modelo uniforme com parâmetros  $a$  e  $b$ . A representação gráfica desse modelo é dada na Figura 5.4.

Continuando com o exemplo 2, podemos verificar que:

$$\int_a^b f_X(x) dx = 1$$

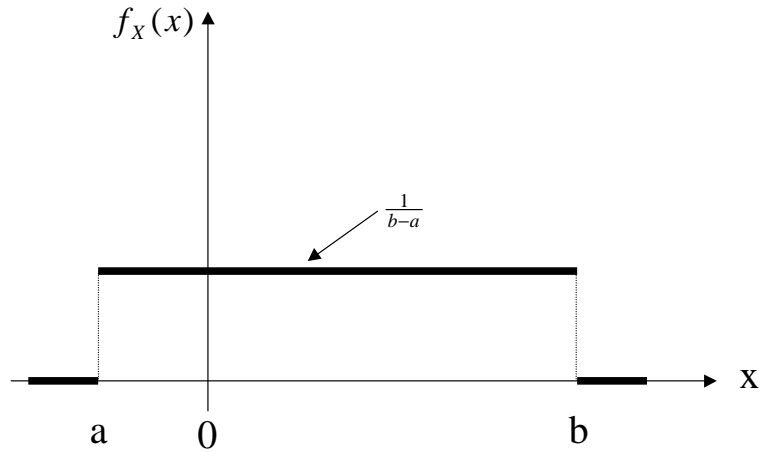


Figura 5.4: Representação gráfica do modelo uniforme

Resolvendo a integral da nossa função, para  $a = 0$  e  $b = 360$ , obtemos:

$$\int_0^{360} \frac{1}{360} dx = \frac{1}{360} x \Big|_0^{360} = \frac{1}{360} (360 - 0) = \frac{360}{360} = 1.$$

A probabilidade da variável aleatória estar dentro de um intervalo  $[x_1; x_2]$ , corresponde a área delimitada pelo intervalo e a curva da função. Para sabermos a área usamos o cálculo integral. Por exemplo, a probabilidade do ângulo estar entre  $0^0$  e  $90^0$  é dada por:

$$P(0^0 \leq X \leq 90^0) = \int_0^{90} \frac{1}{360} dx = \frac{1}{360} \int_0^{90} dx = \frac{1}{360} x \Big|_0^{90} = \frac{1}{360} (90 - 0) = \frac{90}{360} = \frac{1}{4}.$$

Observe que a probabilidade de obter exatamente um determinado valor é zero. Por exemplo:

$$P(X = 90^0) = \int_{90}^{90} \frac{1}{360} dx = \frac{1}{360} \int_{90}^{90} dx = \frac{1}{360} x \Big|_{90}^{90} = \frac{1}{360} (90 - 90) = \frac{0}{360} = 0.$$

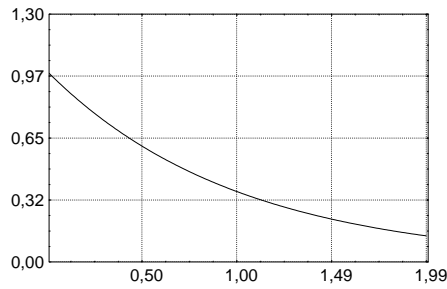
De modo geral, dados dois números  $x_1$  e  $x_2$  quaisquer a  $P(x_1 \leq X \leq x_2) = \frac{x_2 - x_1}{360^0}$ .

**Exemplo 3:** A distribuição de Weibull tem muitas aplicações em teoria da confiabilidade, onde estuda-se sistemas e seus componentes, por exemplo, sistemas biológicos, como o corpo humano.

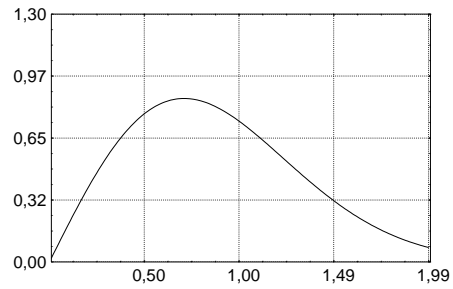
Definição: se a função densidade de probabilidade de uma variável aleatória for dada por:

$$f_X(x) = \begin{cases} \beta x^{\beta-1} e^{-\beta x} & \text{para } x \geq 0 \\ 0 & \text{para } x < 0 \end{cases}$$

onde  $\beta$  é uma constante positiva, dizemos que  $X$  tem distribuição de Weibull, ou segue o modelo de Weibull. Os gráficos para  $\beta = 1$  e  $\beta = 2$  são dados na Figura 5.5.



(a) Distribuição de Weibull para  $\beta = 1$



(b) Distribuição de Weibull para  $\beta = 2$

Figura 5.5: A distribuição de Weibull para  $\beta = 1$  e  $\beta = 2$

Como vimos, obtemos a *probabilidade* de que a variável aleatória  $X$  caia no intervalo  $[x_1, x_2]$ , calculando a área entre  $[x_1, x_2]$  e a curva, e isso é feito através do cálculo da integral da função  $f_X(x)$  entre os pontos  $x_1$  e  $x_2$ , então:

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x) dx.$$

Atualmente temos softwares que calculam estas probabilidades. Tabelas com as probabilidades são dadas nos apêndices deste livro.

Qual a probabilidade de uma medida de  $\alpha$ -globulina ser exatamente igual a 0,86666?

$$P(X = 0,86666) = \int_{0,86666}^{0,86666} f_X(x) dx = 0$$

pois a área sob um ponto do eixo  $x$  é zero. Não há contradição nesse resultado, pois é extremamente improvável que  $X$  assuma esse valor particular. Além do mais, existem infinitos outros números na vizinhança de 0,86666, de tal forma que a probabilidade é tão pequena que tende a desaparecer. Então, para variáveis aleatórias contínuas, tanto faz escrever:

$$P(x_1 \leq X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 < X < x_2).$$

A seguir apresentamos de forma formal (matemática) a definição de variável aleatória contínua.

**Definição:** diz-se que  $X$  é uma variável aleatória contínua, se existir uma função  $f(\cdot)$  denominada *função densidade de probabilidade (fdp)* de  $X$  que satisfaça as seguintes condições:

- a)  $f_X(x) \geq 0$  para todo  $x$  (não-negativa)

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

$$c) \text{ para quaisquer } x_1 \text{ e } x_2, \text{ com } -\infty < x_1 < x_2 \leq \infty, \text{ teremos: } P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x)dx.$$

Pode-se construir modelos teóricos probabilísticos para variáveis aleatórias, escolhendo-se adequadamente as funções densidades de probabilidades. Teoricamente, qualquer função  $f(\cdot)$ , que seja não-negativa e cuja área total sob a curva seja igual à unidade, caracterizará uma variável aleatória contínua.

## 5.2 A Média de Uma Variável Aleatória Contínua

Podemos estender todas as definições feitas para variável aleatória discreta, de modo equivalente, às variáveis aleatórias contínuas.

Se  $X$  é uma variável aleatória contínua, define-se a média de  $X$ , como sendo:

$$E(X) = \mu_X = \int_{-\infty}^{+\infty} x f_X(x)dx.$$

A média de  $X$  também é conhecida como esperança matemática e pode ser entendida como um "centro de distribuição de probabilidade".

A expressão para a *variância* de variáveis aleatórias contínuas é dada por:

$$Var(X) = \sigma_X^2 = E[(X - E(X))^2] = \int_{-\infty}^{+\infty} [X - E(X)]^2 f_X(x)dx$$

ou,

$$Var(X) = \sigma_X^2 = E(X^2) - [E(X)]^2$$

onde:

$$E(X^2) = \int_{-\infty}^{+\infty} x^2 f_X(x)dx.$$

O desvio padrão é dado por:

$$DP(X) = \sigma_X = \sqrt{Var(X)}.$$

**Exemplo.** Determine a esperança e a variância da variável aleatória  $X$ , ângulo entre o norte e a direção tomada pelos pássaros (azimute), em graus, cuja f.d.p. é dada por:

$$f_X(x) = \begin{cases} \frac{1}{360}, & 0 \leq x \leq 360 \\ 0, & \text{caso contrário.} \end{cases}$$

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_0^{360} x \frac{1}{360} dx.$$

$$E(X) = \frac{1}{360} \int_0^{360} x dx$$

$$E(X) = \frac{1}{360} \frac{x^2}{2} \Big|_0^{360}$$

$$E(X) = \frac{1}{360} \frac{(360)^2}{2}$$

$$E(X) = \frac{360}{2}$$

$$E(X) = 180^0$$

Passamos agora para o cálculo da variância.

$$E(X^2) = \int_0^{360} x^2 \frac{1}{360} dx.$$

$$E(X^2) = \frac{1}{360} \int_0^{360} x^2 dx$$

$$E(X^2) = \frac{1}{360} \frac{x^3}{3} \Big|_0^{360}$$

$$E(X^2) = \frac{1}{360} \frac{360^3}{3}$$

$$E(X^2) = \frac{360^2}{3}$$

$$E(X^2) = 43200^0 \tag{5.1}$$

Retornando, temos que:

$$Var(X) = 43200 - 180^2$$

$$Var(X) = 10800.$$

O desvio padrão vale  $103,92^0$ .

### 5.3 A Distribuição Normal

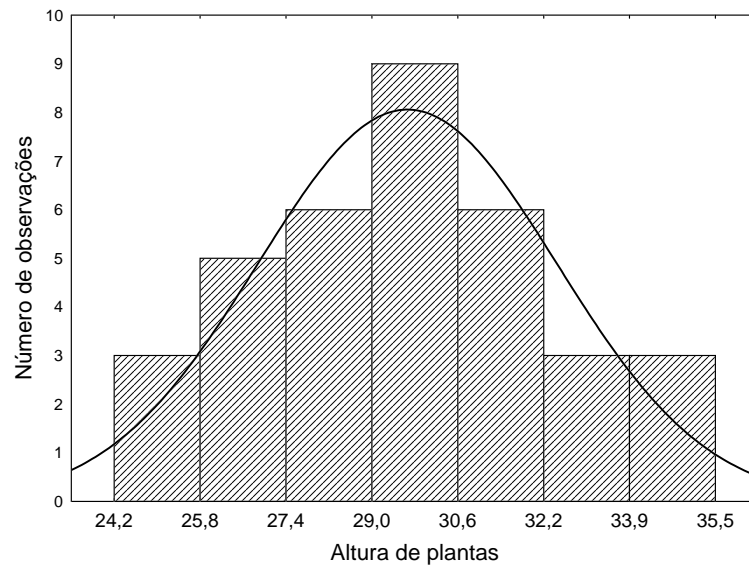


Figura 5.6: Distribuição de frequência da altura de plantas de *Amaranthus*, em cm

### 5.3.1 Introdução

O modelo normal ocupa uma posição de grande destaque tanto a nível teórico como prático, isso porque o modelo normal representa com boa aproximação muitos fenômenos da natureza como, por exemplo, a característica altura de plantas de *Amaranthus*, cuja distribuição de frequência é dada na Figura 5.6. Observe que existe uma tendência das observações se concentrarem próximo do valor central, ou seja, da média da distribuição, e esta concentração vai diminuindo a medida que os valores de altura vão aumentando e diminuindo, ou seja, existe baixa concentração de plantas baixas, assim como de plantas altas. A distribuição é aproximadamente simétrica, isto é, tomando a média como ponto central, a lado esquerdo é aproximadamente igual ao lado direito.

Outra razão da importância do modelo normal é que as distribuições amostrais de estatísticas como médias e proporções, podem ser aproximadas pela distribuição normal, isto é muito importante para o estudo de *inferência estatística*. Esses resultados serão vistos com mais detalhes no estudo de distribuições amostrais, no próximo capítulo.

O cálculo de probabilidades das distribuições binomial e Poisson, pode ser feito com boa aproximação, através da distribuição normal. A aproximação da distribuição binomial pela distribuição normal será vista na Seção 5.3.4.

A distribuição normal surgiu a aproximadamente duas centenas de anos passados, e de que forma? Cientistas coletando um grande número de observações de uma variável

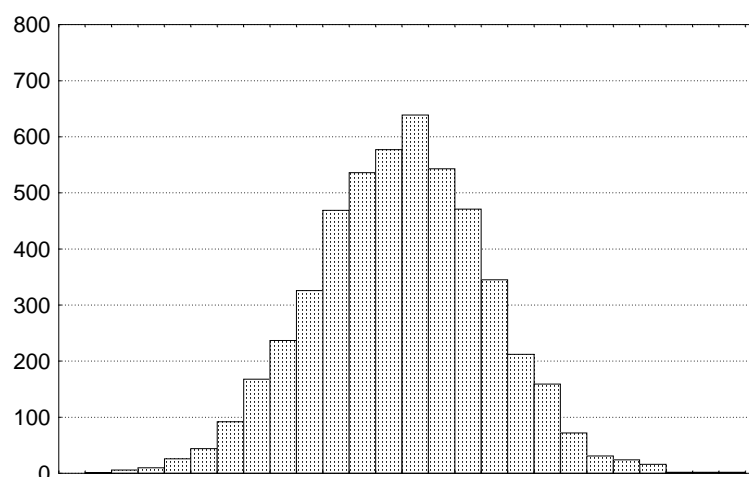


Figura 5.7: Distribuição de frequência em forma de sino

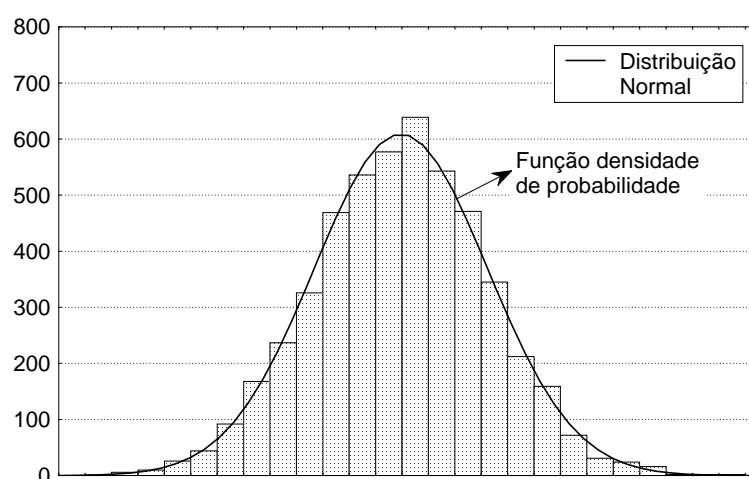


Figura 5.8: Uma curva contínua que aproxima a distribuição de frequências observadas

e construindo sua distribuição de frequência, verificaram que repetidamente o histograma apresentava a forma da Figura 5.7. A distribuição de frequências da Figura 5.7 é conhecida como distribuição em forma de sino.

Mais tarde, esse fato foi transformado em termos matemáticos, isto é, numa expressão matemática que representasse aqueles fenômenos de forma bem aproximada. Esse modelo matemático pode ser visto na Figura 5.8. Observa-se na Figura 5.8 uma curva contínua, simétrica em torno do seu ponto central, isto é, da sua média. Costuma-se designar essa distribuição por distribuição Gaussiana devido ao seu criador, Karl F. Gauss (1777-1855).

Em termos formais, matemáticos, temos a seguinte definição para variável aleatória contínua com distribuição normal.



**Definição:** dizemos que uma variável aleatória contínua  $X$  tem distribuição normal, com parâmetros  $\mu$  e  $\sigma^2$ , onde  $-\infty < \mu < +\infty$  e  $0 < \sigma^2 < +\infty$ , representam a média e a variância da população  $X$ , respectivamente, se a sua função densidade de probabilidade for dada por:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < +\infty$$

onde  $e$  representa a base dos logaritmos naturais e vale aproximadamente 2,7182,  $\pi = 3,1416$  e  $\sigma$  é o desvio padrão. Veja a  $f_X(x)$  na Figura 5.9. Costuma-se representar essa variável aleatória por  $X \sim N(\mu, \sigma^2)$ .

Pode-se demonstrar que:

- i)  $E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx = \mu.$
- ii)  $Var(X) = E[X - E(X)]^2 = \sigma^2.$
- iii)  $f_X(x)$  é simétrica ao redor de  $x = \mu$ , isto é,  $f(\mu + x) = f(\mu - x)$  para todo  $x$ .

Assim, observamos que os dois parâmetros  $\mu$  e  $\sigma^2$ , que caracterizam a distribuição normal, são a média e a variância de  $X$ . O desvio padrão é calculado a partir da variância de  $X$  e vale:

$$DP(X) = \sigma = \sqrt{\sigma^2}.$$

A distribuição normal fica completamente definida por dois parâmetros: *média* ( $\mu$ ) e *desvio padrão* ( $\sigma$ ), então, é fácil perceber que para cada combinação de uma média e um desvio padrão existe uma distribuição normal. Exemplos de possíveis variáveis com distribuição normal podem ser: pesos de bovinos, pesos de frangos, quantidade de chuva (mm), quantidades de  $\alpha$ -globulina no plasma sanguíneo de pessoas, comprimentos do ante-braço de morcegos, etc. Portanto, na prática, temos um número ilimitado de distribuições normais.

O importante é que sob qualquer curva normal a área total é igual a 1 ou 100%. Uma propriedade importante da distribuição normal é que a porcentagem de observações posicionados entre  $\mu - 1\sigma$ ,  $\mu + 1\sigma$  e a curva é igual a 68%, ou seja, a área é igual a 68%, independentemente dos valores de  $\mu$  e  $\sigma$ . A Figura 5.9 ilustra a situação. Também, a porcentagem de observações, ou a área entre os pontos  $\mu - 1,96\sigma$ ,  $\mu + 1,96\sigma$  e a curva é igual a 95%. Também, a porcentagem de observações ou a área entre os pontos  $\mu - 2,58\sigma$ ,  $\mu + 2,58\sigma$  e a curva é igual a 99%. Estas áreas ou porcentagens independem dos valores de  $\mu$  e  $\sigma$ . A ilustração é dada na Figura 5.10. Por exemplo, suponha que a espécie *Harengula clupeiola*, conhecida como sardinha cascuda, distribuída na Lagoa da Conceição, Florianópolis, SC,

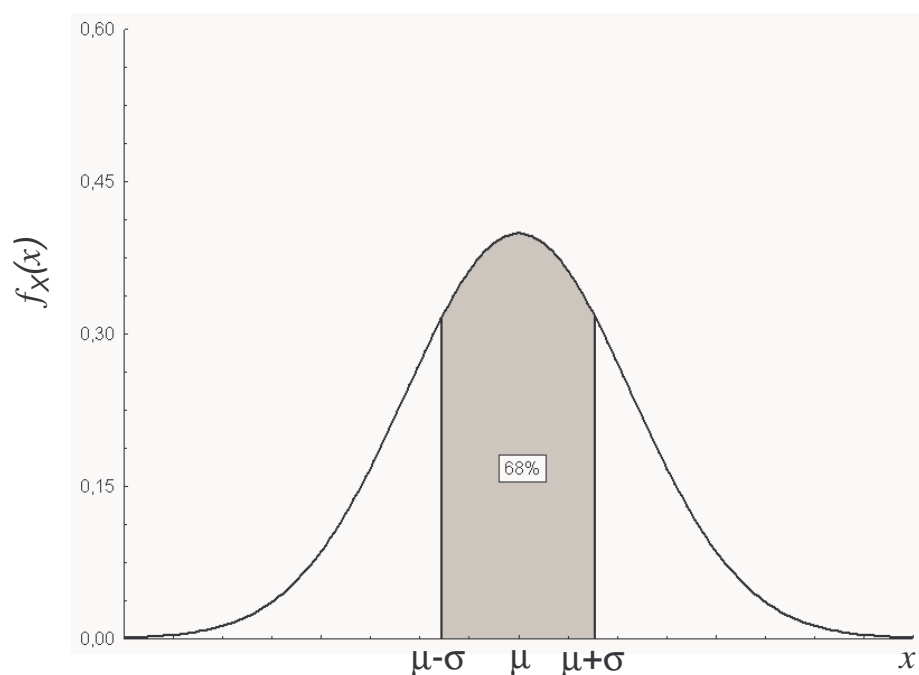


Figura 5.9: A distribuição normal com média  $\mu$  e pontos de inflexão  $\mu \pm \sigma$

apresente crescimento médio igual a 75 mm com desvio padrão igual a 5 mm. Assumindo que a variável em estudo apresente distribuição normal, podemos esperar que aproximadamente 95% dos valores, ou das espécimes, encontram-se no intervalo  $[65,2;84]$ . Como a curva normal é simétrica em relação a média, temos que a probabilidade de obtermos um valor acima da média é 50%, da mesma forma, a probabilidade de obter um valor abaixo da média é 50%.

A probabilidade de uma variável aleatória com distribuição normal tomar um valor entre dois pontos quaisquer, por exemplo, entre os pontos  $x_1$  e  $x_2$ , é igual à área sob a curva normal compreendida entre aqueles dois pontos. Veja a Figura 5.11. Suponha, então, que  $X \sim N(\mu, \sigma^2)$  e queiramos determinar a probabilidade de  $X$  estar entre  $x_1$  e  $x_2$ , portanto, como estamos interessados em obter uma área, devemos realizar o seguinte cálculo:

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

Acontece que essa integral não pode ser calculada exatamente, conseqüentemente, a probabilidade só pode ser obtida aproximadamente, e por métodos numéricos. Podemos obter estas probabilidades com o uso de programas computacionais estatísticos, entre os quais podemos citar o Statistica, Minitab e o Statgraphics.

Para padronizar todas as distribuições normais numa única distribuição e apresentar as

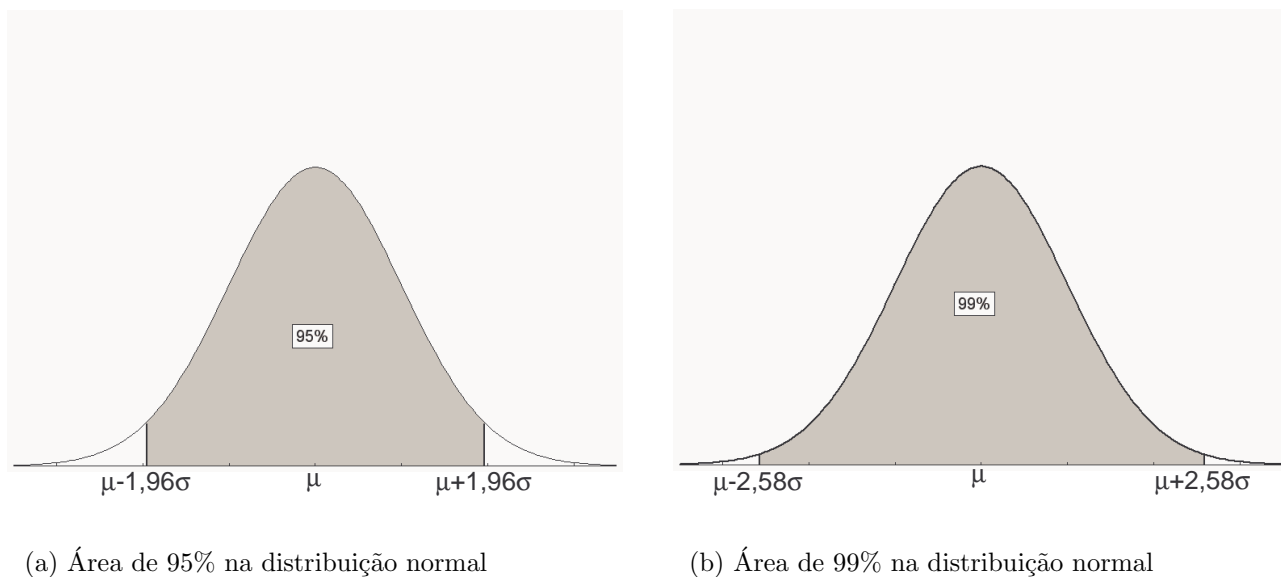


Figura 5.10: Áreas na distribuição normal

probabilidades numa única tabela, foi realizado uma transformação na variável aleatória  $X$ , originando uma nova variável aleatória, denominada de variável aleatória normal padronizada e representada pela letra  $Z$ , cuja distribuição denomina-se de *distribuição normal padrão*, a qual passaremos a estudar.

### 5.3.2 A Distribuição Normal Padrão

A transformação realizada para padronizar as distribuições normais é feita tendo a média como origem, referência do novo sistema, e o desvio padrão como medida de afastamento a contar da média, ou seja, a nova unidade de medida será o desvio padrão, isto é, o quanto em desvios padrões o valor de  $X$  se afasta da média  $\mu$ . Esta nova escala é chamada de escala  $Z$  e é definida como sendo:

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

onde:

- $Z$  = número de desvios padrões a contar da média;
- $X$  = é a variável na unidade original;
- $\mu$  = é a média da população;
- $\sigma$  = é o desvio padrão da população.

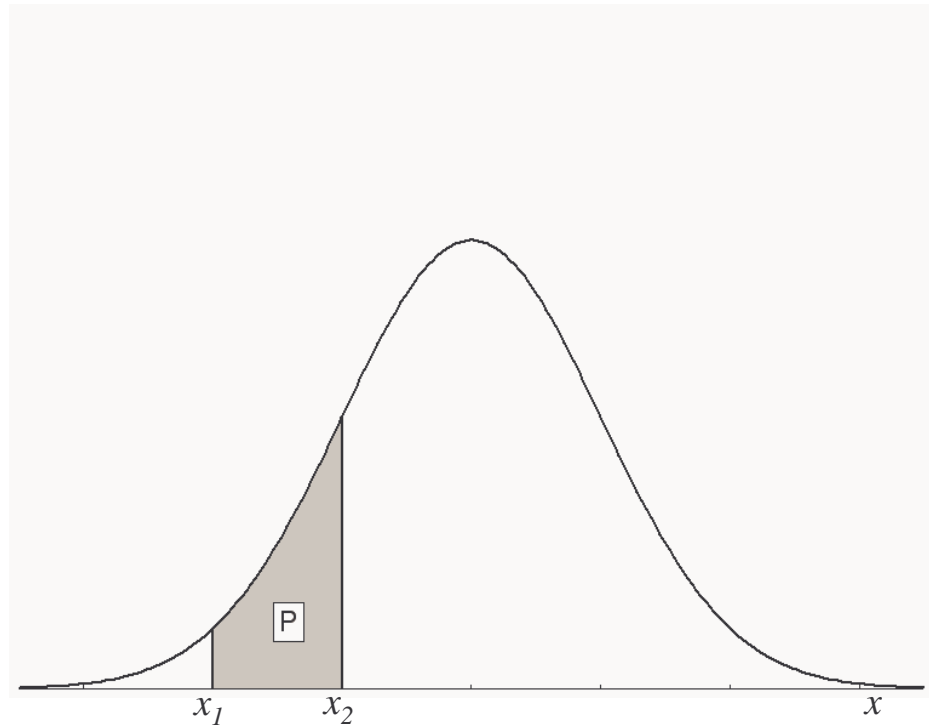


Figura 5.11: A probabilidade de um valor de  $x$  estar entre os pontos  $x_1$  e  $x_2$  e a curva, corresponde a área achurada da Figura.

**Exemplo:** consideremos uma distribuição normal com média  $\mu = 100$  e desvio padrão  $\sigma = 10$ . Para  $x = 100$ , temos:

$$z = \frac{100 - 100}{10} = 0.$$

Para  $x = 120$ , temos:

$$z = \frac{120 - 100}{10} = 2,$$

e assim por diante. A Figura 5.12 apresenta a distribuição de  $X$  e de  $Z$ .

**Definição:** se  $X : N(\mu, \sigma^2)$ , então a variável aleatória  $Z$  definida por:

$$Z = \frac{X - \mu}{\sigma}$$

tem uma distribuição  $N(0, 1)$ , isto é, tem distribuição normal com média  $\mu = 0$  e variância  $\sigma^2 = 1$ , cuja função densidade de probabilidade é dada por:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad -\infty \leq z \leq +\infty.$$

A seguir fazemos a demonstração do valor média e da variância na distribuição normal padrão.

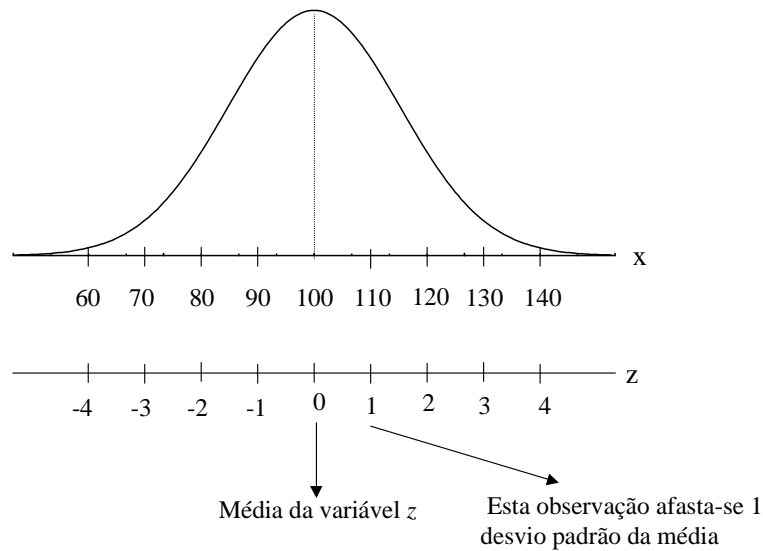


Figura 5.12: Distribuição das variáveis  $X$  e  $Z$

### *Média e Variância da Variável Aleatória $Z$*

Demonstração:

$$\begin{aligned}
 E(Z) &= E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}E(X - \mu) = \frac{1}{\sigma}[E(X) - E(\mu)] = \frac{1}{\sigma}(\mu - \mu) = 0. \\
 V(Z) &= E(Z^2) - [E(Z)]^2 \\
 &= E(Z^2) \\
 &= E\left(\frac{X - \mu}{\sigma}\right)^2 = \frac{1}{\sigma^2}E(X - \mu)^2 = \frac{1}{\sigma^2}E(X^2 - 2\mu X + \mu^2) \\
 &= \frac{1}{\sigma^2}[E(X^2) - 2\mu E(X) + \mu^2] \\
 &= \frac{1}{\sigma^2}[E(X^2) - \mu^2] = \frac{1}{\sigma^2}[E(X^2) - [E(X)]^2] \\
 &= \frac{1}{\sigma^2}\sigma^2 = 1.
 \end{aligned}$$

#### 5.3.3 O Uso da Tabela da Distribuição Normal Padrão

No apêndice 3, temos a tabela da distribuição normal padrão. Esta tabela dá a área entre a média (zero), o valor de  $z$  e a curva, isto é, a tabela dá a probabilidade de um valor cair no intervalo  $[0$  a  $z]$ . Veja a Figura 5.13, na qual a área achurada corresponde a probabilidade.

A distribuição normal padrão ou reduzida, é absolutamente simétrica em relação a média (origem), em função disso, a tabela só é dada para valores positivos de  $z$ . Se o valor calculado

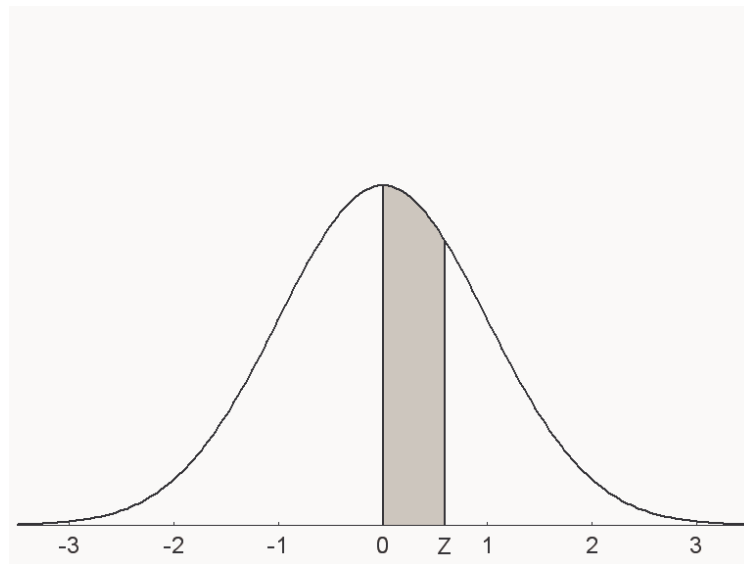


Figura 5.13: Distribuição normal padrão. A área achurada corresponde a probabilidade de uma observação estar no intervalo  $[0; z]$

de  $z$  for negativo, basta considerá-lo como positivo e entrar diretamente na tabela. A área total vale 100% ou 1, assim, acima de  $\mu$  temos 50% da área e abaixo de  $\mu$  temos os outros 50%. A seguir vamos calcular probabilidades usando a tabela da distribuição normal padrão.

**Exemplo 1.** Qual a probabilidade de uma observação estar entre 0 e 1,25? Isto é,  $P(0 \leq z \leq 1,25)$ . O que desejamos é a área entre 0 e 1,25 e a curva da Figura 5.14. Com o auxílio da tabela para  $z = 1,25$  temos  $P(0 \leq z \leq 1,25) = 0,3944$ .

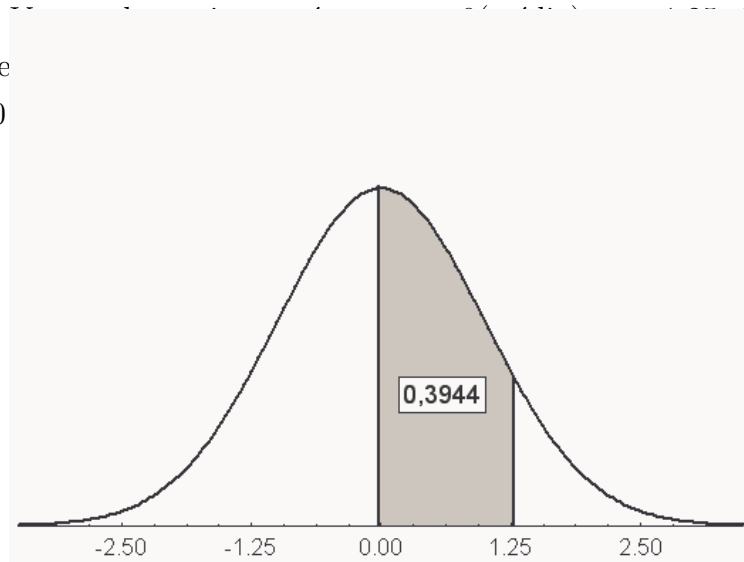


Figura 5.14:  $P(0 \leq z \leq 1,25) = 0,3944$

**Exemplo 2.** Qual a área entre 0 (média) e  $z = 1,73$ , isto é, a  $P(0 \leq z \leq 1,73)$ . O que desejamos é a área entre 0 e 1,73 e a curva da Figura 5.15. Com o auxílio da tabela para  $z = 1,73$  temos  $P(0 \leq z \leq 1,73) = 0,4582$ .

**Exemplo 3.** Qual é a probabilidade de  $z$  estar entre -1,73 e 0 ( $P(-1,73 \leq z \leq 0)$ )? Como a curva da distribuição normal padrão é absolutamente simétrica, essa probabilidade pode ser obtida calculando-se a  $P(0 \leq z \leq 1,73)$ . Como vimos no exemplo 2 essa probabilidade vale 0,4582.

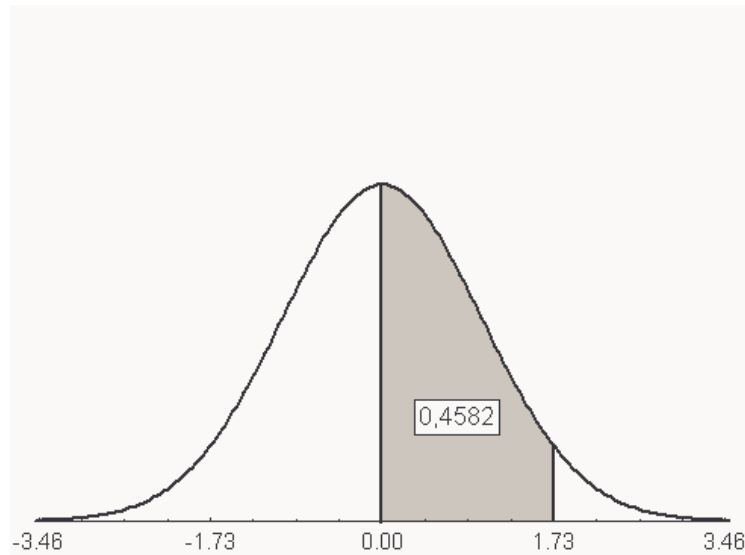


Figura 5.15:  $P(0 \leq z \leq 1,73) = 0,4582$

**Exemplo 4.** Qual é a probabilidade de  $z$  estar entre 0 e 1,73? A área achurada da Figura 5.15 indica essa probabilidade.

corresponde a área achurada da Figura 5.15, logo a

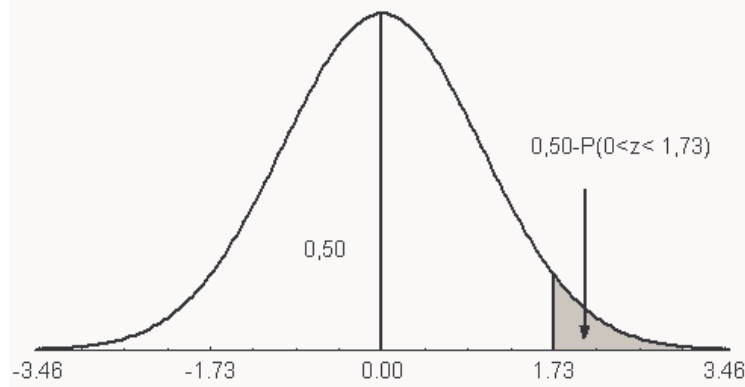


Figura 5.16:  $P(z \geq 1,73) = 0,0418$

área além de  $z$  é 50% menos o valor tabelado de  $z=1,73$ , então:

$$P(z \geq 1,73) = 0,50 - P(0 \leq z \leq 1,73) = 0,50 - 0,4582 = 0,0418.$$

**Exemplo 5.** Qual é a probabilidade de  $z$  estar entre -1 e 1 ( $P(-1 \leq z \leq 1)$ )? A área achurada da Figura 5.17 indica essa probabilidade.

$$\begin{aligned} P(-1 \leq z \leq 1) &= P(0 \leq z \leq 1) + P(0 \leq z \leq 1) \\ &= 2P(0 \leq z \leq 1) \\ &= 2(0,3413) \\ &= 0,6826. \end{aligned}$$

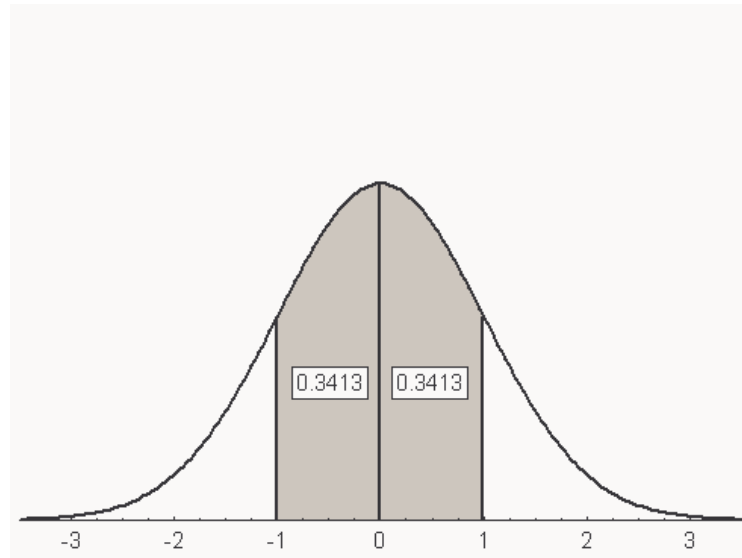


Figura 5.17:  $P(-1 \leq z \leq 1) = 0,6826$

**Exemplo 6.** Qual a probabilidade correspondente a  $P(1 \leq z \leq 3)$ ? Essa probabilidade corresponde à área sob a curva normal entre  $z = 1$  e  $z = 3$ . Essa

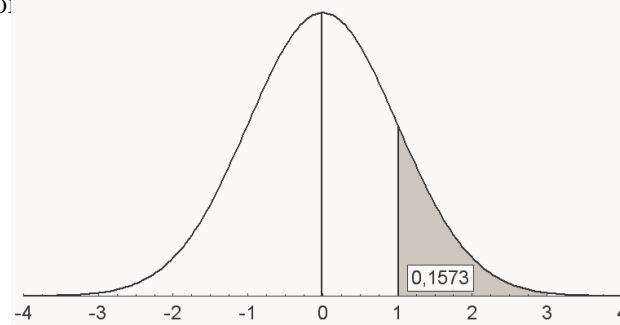


Figura 5.18:  $P(1 \leq z \leq 3) = 0,1573$

$$\begin{aligned} P(1 \leq z \leq 3) &= P(0 \leq z \leq 3) - P(0 \leq z \leq 1) \\ &= 0,4986 - 0,3413 \\ &= 0,1573. \end{aligned}$$

**Exemplo 7.** Suponha que  $X$  seja uma variável aleatória com distribuição normal de média  $\mu = 3$  e variância  $\sigma^2 = 16$ , e desejamos calcular a  $P(2 \leq X \leq 5)$ . O primeiro passo é calcularmos a variável  $z$ .

$$\begin{aligned} P(2 \leq X \leq 5) &= P\left(\frac{2 - \mu}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{5 - \mu}{\sigma}\right) \\ &= P\left(\frac{2 - 3}{4} \leq z \leq \frac{5 - 3}{4}\right) \end{aligned}$$



$$= P\left(\frac{-1}{4} \leq z \leq \frac{1}{2}\right).$$

Portanto,

$$\begin{aligned} P(-0,25 \leq z \leq 0,5) &= P(0 \leq z \leq 0,25) + P(0 \leq z \leq 0,50) \\ &= 0,0987 + 0,1915 \\ &= 0,2902. \end{aligned}$$

### 5.3.4 A Distribuição Normal como Aproximação da Distribuição Binomial

Muitas situações reais podem ser convenientemente descritas pelo modelo binomial. Acontece que quando o número de repetições do experimento ( $n$ ) for muito grande, para se calcular as probabilidades binomiais, são necessários cálculos extensos e, as vezes, somente é possível a resolução com o uso de computadores. Para  $n$  suficientemente grande e  $\pi$  não próximo de 0 (zero) e de 1 (um), obtém-se boas aproximações da normal à binomial. Para saber se a aproximação é boa, uma sugestão é verificar se  $n\pi \geq 5$  e  $n(1 - \pi) \geq 5$ .

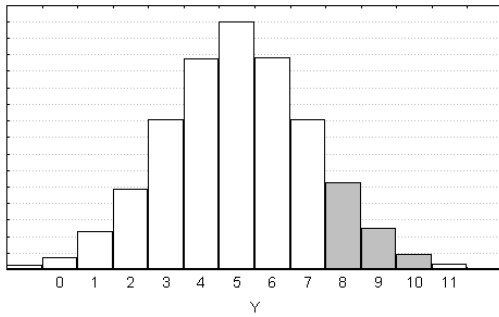
Como o modelo normal é para variável aleatória contínua (pensar em intervalo), enquanto o modelo binomial é para variável aleatória discreta, pode-se melhorar a aproximação fazendo-se um pequeno ajuste, chamado de correção de continuidade, que consiste em subtrair e/ou somar  $1/2$  aos valores da variável aleatória discreta. Mostraremos o procedimento através de um exemplo.

**Exemplo.** Consideremos a variável aleatória  $Y : b(10; 1/2)$ , isto é, temos uma variável aleatória binomial com parâmetros  $n = 10$  e  $\pi = 1/2$ , e desejamos calcular  $P(8 \leq Y \leq 10)$ , a Figura 5.19 mostra a distribuição de  $Y$  e a curva normal, respectivamente. Observe na Figura 5.19(b) a compensação de pequenas áreas que ocorre fazendo-se a correção de continuidade, ou seja, os retângulos possuem base unitária (um) com ponto médio, por exemplo, igual a 8, assim a base do retângulo vai de 7,5 a 8,5. Chamando  $X$  a variável com distribuição normal, temos:

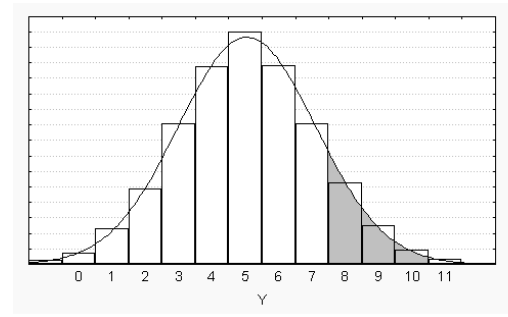
$$P(8 \leq Y \leq 10) = P(7,5 \leq X \leq 10,5) = P\left(\frac{7,5 - n\pi}{\sqrt{n\pi(1 - \pi)}} \leq z \leq \frac{10,5 - n\pi}{\sqrt{n\pi(1 - \pi)}}\right)$$

onde,  $n\pi$  e  $\sqrt{n\pi(1 - \pi)}$  são a média e o desvio padrão da distribuição binomial e valem  $n\pi = 10(0,50) = 5,0$  e  $\sqrt{10(0,50)(0,50)} = 1,58113$ , respectivamente. Logo,

$$P\left(\frac{7,5 - 5,0}{1,58113} \leq z \leq \frac{10,5 - 5,0}{1,58113}\right) = P(1,5811 \leq z \leq 3,4785) = 0,49975 - 0,44295 = 0,057.$$



(a) Distribuição de uma variável binomial, de média 5



(b) Distribuição normal como aproximação da binomial

Figura 5.19: Aproximação normal à distribuição binomial

Resolvendo o problema pela distribuição binomial, obtemos:

$$P(8 \leq Y \leq 10) = 0,044 + 0,010 + 0,001 = 0,055.$$

Observamos que os dois resultados são bem próximos, mesmo com  $n$  pequeno, porém,  $\pi = 0,50$ .

Para verificarmos em que condições a aproximação da binomial pela normal é bastante satisfatória, foram calculadas as probabilidades pelos dois modelos, combinando-se diversos valores de  $n$  e de  $\pi$ . Os resultados são mostrados na Tabela 5.1.

### Exercícios resolvidos

1) O comprimento, em cm, de *Pnnaeus schmitti* (camarão marinho), em condições normais no estuário (Lagoa do Ibiraquera), tem distribuição aproximadamente normal, com média de 6,0 cm e variância de 0,2  $cm^2$ .

a) Qual o intervalo simétrico em torno da média, que conterá 75% dos comprimentos dos camarões?

b) Qual o valor de comprimento,  $c$ , que é superado por 7% dos valores?

Do problema temos que:

$$\mu = 6,0 \quad \sigma^2 = 0,2 \quad \sigma = 0,4472.$$

Com o auxílio de um programa estatístico, ou com a tabela da distribuição normal padrão, para  $P(Z_1 < Z < Z_2) = 75\%$ , simétrico em torno de  $Z$ , obtemos  $Z_1 = -1,150349$  e  $Z_2 = 1,150349$ . Portanto,

$$-1,150349 = \frac{X_1 - 6}{0,4472} \implies X_1 = 5,4856.$$

Tabela 5.1: Aproximação da distribuição binomial pela distribuição normal, variando o tamanho da amostra  $n$  e as probabilidades  $\pi$

$n$	$\pi$	$n\pi$	$\sqrt{n\pi(1-\pi)}$	$P[0; 3]$		$P[(1/5)n-2; (1/5)n+2]$		$P[(1/2)n-2; (1/2)n+2]$		$P[(4/5)n-2; (4/5)n+2]$		$P[$
				Bin	Nor	Bin	Nor	Bin	Nor	Bin	Nor	
50	0.01	0.50	0.70	0.9984	0.9224	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
50	0.05	2.50	1.54	0.7604	0.7160	0.0032	0.0006	0.0000	0.0000	0.0000	0.0000	0.0000
50	0.10	5.00	2.12	0.2503	0.2350	0.1211	0.1191	0.0000	0.0000	0.0000	0.0000	0.0000
50	0.30	15.00	3.24	0.0000	0.0002	0.2156	0.2099	0.0122	0.0103	0.0000	0.0000	0.0000
50	0.50	25.00	3.54	0.0000	0.0000	0.0002	0.0002	0.5201	0.5205	0.0002	0.0002	0.0000
50	0.60	30.00	3.46	0.0000	0.0000	0.0000	0.0000	0.2179	0.2201	0.0132	0.0150	0.0000
50	0.75	37.50	3.06	0.0000	0.0000	0.0000	0.0000	0.0010	0.0005	0.4657	0.4488	0.0005
50	0.93	46.50	1.80	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0220	0.0133	0.5327
50	0.97	48.50	1.21	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.9372
100	0.01	1.00	0.99	0.9816	0.9282	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.05	5.00	2.18	0.2578	0.2398	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.10	10.00	3.00	0.0078	0.0149	0.0099	0.0062	0.0000	0.0000	0.0000	0.0000	0.0000
100	0.30	30.00	4.58	0.0000	0.0000	0.0457	0.0477	0.0001	0.0001	0.0000	0.0000	0.0000
100	0.50	50.00	5.00	0.0000	0.0000	0.0000	0.0000	0.3827	0.3829	0.0000	0.0000	0.0000
100	0.60	60.00	4.90	0.0000	0.0000	0.0000	0.0000	0.0580	0.0575	0.0001	0.0002	0.0000
100	0.75	75.00	4.33	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.2487	0.2402	0.0000
100	0.93	93.00	2.55	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0000	0.0744
100	0.97	97.00	1.71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6472
200	0.01	2.00	1.41	0.8580	0.8190	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
200	0.05	10.00	3.08	0.0090	0.0171	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
200	0.10	20.00	4.24	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
200	0.30	60.00	6.48	0.0000	0.0000	0.0026	0.0032	0.0000	0.0000	0.0000	0.0000	0.0000
200	0.50	100.00	7.07	0.0000	0.0000	0.0000	0.0000	0.2762	0.2763	0.0000	0.0000	0.0000
200	0.60	120.00	6.93	0.0000	0.0000	0.0000	0.0000	0.0055	0.0052	0.0000	0.0000	0.0000
200	0.75	150.00	6.12	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0906	0.0897	0.0000
200	0.93	186.00	3.61	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0003
200	0.97	194.00	2.41	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.1472
500	0.01	5.00	2.22	0.2636	0.2434	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.05	25.00	4.87	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.10	50.00	6.71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.30	150.00	10.25	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.50	250.00	11.18	0.0000	0.0000	0.0000	0.0000	0.1769	0.1769	0.0000	0.0000	0.0000
500	0.60	300.00	10.95	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.75	375.00	9.68	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0072	0.0078	0.0000
500	0.93	465.00	5.71	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
500	0.97	485.00	3.81	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002
1000	0.01	10.00	3.15	0.0101	0.0190	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.05	50.00	6.89	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.10	100.00	9.49	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.30	300.00	14.49	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.50	500.00	15.81	0.0000	0.0000	0.0000	0.0000	0.1256	0.1256	0.0000	0.0000	0.0000
1000	0.60	600.00	15.49	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.75	750.00	13.69	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0002	0.0000
1000	0.93	930.00	8.07	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1000	0.97	970.00	5.39	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

$$1,150349 = \frac{X_2 - 6}{0,4472} \implies X_2 = 6,5144.$$

b) Para uma probabilidade de 43% (50%-7%), obtemos na tabela da distribuição normal padrão aproximadamente  $z = +1,48$ . Portanto,

$$1,48 = \frac{c - 6}{\sqrt{0,2}} \Rightarrow c = 6,66.$$

Assim, 7% dos camarões maiores tem comprimento igual ou superior a 6,66 cm.

2) O diâmetro de certa espécie de árvore é uma variável aleatória com distribuição normal de média 50 cm e desvio padrão de 6 cm. Se o diâmetro de uma árvore diferir da média de mais de 10 cm, esta árvore é vendida por 10 u.m; caso contrário, é vendida por 20 u.m. Qual o preço médio de venda de cada árvore?

Do problema temos que:

$$\mu = 50 \text{ cm} \quad \sigma = 6 \text{ cm}.$$

A probabilidade de uma árvore diferir mais do que 10 cm da média populacional é:

$$\begin{aligned} P(X > 60 \text{ cm}) + P(X < 40 \text{ cm}) &= P(Z > \frac{60 - 50}{6}) + P(Z < \frac{40 - 50}{6}) \\ P(Z > 1,67) + P(Z < -1,67) &= 2P(Z > 1,67) \\ &= 2[0,5 - P(0 < Z < 1,67)] \\ &= 2[0,5 - 0,45254] \\ &= 0,09492. \end{aligned}$$

A probabilidade de uma árvore apresentar diâmetro entre 40 e 60 cm é  $1 - 0,09492 = 0,90508$ . Portanto, o preço médio de venda de cada árvore é:

$$E(X) = (10 \times 0,09492) + (20 \times 0,90508) = 0,9492 + 18,11 = 19,05.$$

3) Os ovos da produção de uma granja são classificados em grande ou pequenos, conforme seu diâmetro. Verificou-se que 45% dos ovos são considerados grandes. Supondo que os ovos são colocados em caixas de 60, aleatoriamente, pergunta-se:

a) Em que porcentagem de caixas teremos pelo menos 50% de ovos grandes? (50% é igual a 30 ovos).

b) Em que porcentagem de caixas teremos exatamente 50% de ovos grandes?

Sendo:

$$\pi = 0,45 \quad n = 60 \quad n\pi = 0,45 \times 60 = 27.$$

O desvio padrão do número de ovos por caixa, é dado por:

$$\sqrt{n\pi(1 - \pi)} = \sqrt{60 \times 0,45 \times 0,55} = 3,85356 \text{ ovos}.$$

a) Usando aproximação normal à binomial, a probabilidade de uma caixa conter 30 ou mais ovos, vale:

$$\begin{aligned} P(Y \geq 30) = P(X \geq 29,5) &= P\left(Z \geq \frac{29,5 - 27}{3,85356}\right) \\ P(Z \geq 0,65) &= 0,5 - P(0 \leq Z \leq 0,65) \\ &= 0,5 - 0,24215 \\ &= 0,25785. \end{aligned}$$

b) A probabilidade de encontrarmos exatamente 30 ovos, é:

$$\begin{aligned} P(Y = 30) = P(29,5 \leq X \leq 30,5) &= P\left(\frac{29,5 - 27}{3,85356} \leq Z \leq \frac{30,5 - 27}{3,85356}\right) \\ P(0,65 \leq Z \leq 0,91) &= 0,31859 - 0,24215 \\ &= 0,0764. \end{aligned}$$

4) Sabe-se que a probabilidade de um indivíduo inoculado contra um surto de gripe vir a ter uma reação séria indesejável é de 0,05. Usando a aproximação normal à distribuição binomial, calcule a probabilidade de que mais de 16 indivíduos dentre 200 indivíduos inoculados tenham tais reações.

Com as informações do problema temos que:

$$\begin{aligned} \pi &= 0,05 & n &= 200 & n\pi &= 200 \times 0,05 = 10 \\ n(1 - \pi) &= 200 \times 0,95 = 190 & \sqrt{n\pi(1 - \pi)} &= 3,0822. \end{aligned}$$

A probabilidade de que mais de 16 indivíduos tenham reações, é:

$$\begin{aligned} P(Y > 16) = P(Y \geq 17) \simeq P(X \geq 16,5) &= P\left(Z \geq \frac{16,5 - 10}{3,0822}\right) \\ &= P(Z \geq 2,1089) \\ &= 0,5 - 0,48257 \\ &= 0,0174 \\ &= 1,74\%. \end{aligned}$$

### 5.3.5 Exercícios propostos

1) Determinar a área limitada pela curva normal em cada um dos casos abaixo:

a) entre  $z=0$  e  $z=1,2$ ;

b) entre  $z=-0,68$  e  $z=0$ ;

- c) entre  $z=0,46$  e  $z=2,21$ ;
- d) entre  $z=-0,81$  e  $z=1,94$ ;
- e) à esquerda de  $z=-0,6$ ;
- f) à direita de  $z=-1,23$
- g) à direita de  $z=2,05$  e à esquerda de  $z=1,44$ .
- h) entre  $z=-1$  e  $z=+1$
- i) entre  $z=-1,96$  e  $z=+1,96$
- j) entre  $z=-2,56$  e  $z=+2,56$

2) A altura dos indivíduos de uma população distribui-se normalmente com média de 1,56 m e desvio padrão de 0,09 m. Qual a percentagem nesta população de indivíduos com altura de 1,80 m ou mais?

3) O peso médio das reses, que se encontram num curral de uma determinada fazenda, é de 200 kg, e o desvio padrão é de 10 kg. Em 120 animais retirados ao acaso do curral, quantos pesarão mais de 185 kg? Vamos considerar que o peso das reses tenha distribuição normal.

4) Foi feito um estudo sobre a altura de plantas de milho de certo híbrido, observando-se que ela se distribui normalmente com média 2,20 m e desvio padrão de 0,20 m. Qual a percentagem de plantas com altura: a)entre 2,15m e 2,25m; b)entre 2,00m e 2,40m; c)acima de 2,30m.

5) Em indivíduos sadios, o consumo renal de oxigênio tem distribuição normal de média  $12 \text{ cm}^3/\text{min}$  e desvio padrão  $1,5 \text{ cm}^3/\text{min}$ .

- a) Determinar a proporção de indivíduos sadios com consumo: inferior a  $10 \text{ cm}^3/\text{min}$ ; superior a  $8 \text{ cm}^3/\text{min}$ ; entre 9,4 e  $13,2 \text{ cm}^3/\text{min}$ ; igual a  $11,6 \text{ cm}^3/\text{min}$ .
- b) Determinar o valor do consumo renal que é superado por 98,5% dos indivíduos sadios.
- c) Determinar uma faixa simétrica em torno do valor médio que contenha 90% dos valores do consumo renal.

6) A observação dos pesos,  $X$ , de um grande número de espigas de milho mostrou que essa variável é normalmente distribuída com média  $\mu = 90g$  e desvio padrão  $\sigma = 7g$ . Num programa de melhoramento, entre outras características, uma cultivar deve satisfazer a condição  $78 < X < 104g$ , onde  $X$  é a variável aleatória peso da espiga. Nessas condições, tendo -se 100 linhagens, pergunta-se:

- a) Qual a proporção de linhagens que deverá ser aceita?
- b) Qual o número de linhagens que continuará participando do programa de melhoramento?

7) Suponha que o conteúdo de bactérias de um tipo particular, presentes em um recipiente de água de 1 mililitro, tenha distribuição aproximadamente normal, com média de 85

bactérias e desvio padrão de 9 bactérias. Qual a probabilidade de uma dada amostra de 1ml conter mais de 100 bactérias?

8) A distribuição dos pesos de coelhos criados numa granja pode muito bem ser representado por uma distribuição normal, com média de 5 kg e desvio padrão de 0,8 kg. Um abatedouro comprará 5.000 coelhos e pretende classificá-los de acordo com o peso, do seguinte modo: 20% dos leves como pequenos, os 55% seguintes como médios, os 15% seguintes como grandes e os 10% mais pesados como extras. Quais os limites de peso para cada classificação?

9) Num povoamento florestal de distribuição aproximadamente normal dos diâmetros à altura do peito (D.A.P.), a média é de 16,2 cm, e o desvio padrão é de 3,4 cm. Determinar o valor  $D$  tal que tenhamos  $P(D.A.P. > D) = 5\%$

10) Suponha que a variável diâmetro de *Paepalanthus* tenha distribuição normal com média 10mm e variância  $4mm^2$ .

a) Qual é a probabilidade de um *Paepalanthus* aleatoriamente retirado dessa população ter diâmetro maior que 14?

b) Se dois *Paepalanthus* forem selecionados aleatoriamente dessa população, qual é a probabilidade de ambos serem maiores que 14?

11) Num povoamento florestal temos uma distribuição aproximadamente normal dos diâmetros à altura do peito (D.A.P.) das árvores, com média de 12,6 cm e desvio padrão de 3,1 cm. Se cortarmos todas as árvores de menos de 15 cm de diâmetro, qual a percentagem de árvores que restarão de pé?

12) Uma máquina de empacotar determinado produto apresenta variações de peso com desvio padrão de 20 g. Em quanto deve ser regulado o peso médio do pacote, para que apenas 10% tenham menos de 400 g?

13) O peso bruto de latas de conserva é uma variável aleatória, com média 1000 gramas e desvio padrão 20 gramas. As latas tem peso médio de 100 gramas e desvio padrão de 10 gramas, também com distribuição normal.

a) Qual a probabilidade de uma lata conter menos de 850 gramas de peso líquido?

b) Qual a probabilidade de uma lata conter mais de 920 gramas de peso líquido? OBS. Considere  $X$  como sendo o peso bruto e  $Y$  como sendo o peso da lata. Como  $X$  e  $Y$  são independentes, a  $VAR(X - Y) = VAR(X) + VAR(Y)$  e o  $D.P.(X - Y) = \sqrt{VAR(X - Y)}$ , assim  $D.P.(X - Y) = 22,3606$ .

14) As vendas de sementes de milho tem distribuição aproximadamente normal com média 500 sacos e desvio padrão 50 sacos. Se a empresa decide produzir 600 sacos no mês em estudo, qual é a probabilidade de que não possa atender a todos os pedidos do mês, por estar com a produção esgotada?

15) Se  $X$  é uma variável aleatória  $N(95; 56, 25)$ , determinar  $b$  tal que

a)  $P(X > b) = 0,975$  e

b)  $P(X < b) = 0,3708$ .

16) Sabe-se que o comprimento de pétalas de uma população de plantas da espécie  $X$  é normalmente distribuída com média  $\mu = 3,2\text{cm}$  e  $\sigma = 1,8\text{cm}$ . Qual a proporção na população é esperada ter um comprimento de pétalas.

a) maior do que  $4,5\text{cm}$ ?

b) entre  $2,9$  e  $3,6\text{cm}$ ?

c) Determinar o valor do comprimento de pétalas que é superado por 65% das plantas.

17) Suponha que a altura em  $\text{cm}$  de uma cultivar de trigo é normalmente distribuída com média de  $100\text{cm}$ . Se um melhorista exige que ao menos 90% das plantas tenham altura superior a  $80\text{cm}$ , qual o maior valor que  $\sigma$  pode assumir, satisfazendo as exigências do melhorista?

18) Em um exame de estatística, a nota média foi 70, com  $\sigma = 4,5$ . Todos os alunos que obtiveram nota 75 a 89 receberam conceito  $B$ . Se as notas têm distribuição aproximadamente normal, e se 10 estudantes obtiveram conceito  $B$ , quantos se submeteram ao exame?

19) Uma enchedora automática de garrafas de vinho está regulada para que o volume médio de líquido em cada garrafa seja de  $\mu = 1000\text{cm}^3$  e o desvio padrão  $\sigma = 10\text{cm}^3$ . Pode-se admitir que a distribuição da variável seja normal.

a) Qual a porcentagem de garrafas em que o volume de líquido é menor que  $990\text{cm}^3$ ?

b) Qual a porcentagem de garrafas em que o volume de líquido não se desvia da média em mais que dois desvios padrões, em ambos os lados?

c) O que acontecerá com a porcentagem do item (b) se a máquina for regulada de forma que a média seja  $1.200\text{cm}^3$  e o desvio padrão  $20\text{cm}^3$ ? Explique.

20) A observação dos pesos,  $X$ , de um grande número de espigas de milho mostrou que essa variável é normalmente distribuída com média  $\mu = 90\text{g}$  e desvio padrão  $\sigma = 7\text{g}$ . Num programa de melhoramento, entre outras características, uma cultivar deve satisfazer a condição  $78 < X < 104\text{g}$ , onde  $X$  é a variável aleatória peso da espiga. Nessas condições, tendo-se 100 linhagens, pergunta-se:

a) qual a proporção de linhagens que deverá ser aceita?

b) qual a proporção de linhagens que deverá ser rejeitada?

c) qual o número de linhagens que continuará participando do programa de melhoramento?

21) 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  $w$ . Supondo que foram



abatidas apenas 4200 cabeças,

a) qual o número esperado de bovinos com peso entre 17 e 19?

b) qual o valor  $w$ ?

c) qual a  $P(w < X < \mu)$ ?

22) Suponha que os diâmetros de *Paepalanthus* sejam distribuídos com média  $12\text{cm}$  e desvio padrão  $5\text{cm}$ . Um *Paepalanthus* é considerado pequeno se seu diâmetro for menor que  $4\text{cm}$  ou grande se seu diâmetro for maior que  $19\text{cm}$ .

a) Encontre a porcentagem de *Paepalanthus* considerados pequenos e grandes.

b) Quinze por cento dos *Paepalanthus* maiores são classificados como grandes. Encontre o diâmetro mínimo para ser classificado como grande.

23) O tempo de validade de um princípio ativo de um determinado inseticida tem distribuição normal de média 803 dias e variância  $1581(\text{dias})^2$ . Determinar a probabilidade de validade deste inseticida:

a) antes de 750 dias;

b) entre 700 e 900 dias;

c) depois do tempo médio.

24) A distribuição das relações altura/comprimento de conchas de mexilhões *Perna perna*, num ambiente de costão batido, pode ser representada por uma distribuição aproximadamente normal, com média de 0,5 e desvio padrão de 0,02414. Um pesquisador pretende classificá-los de acordo com a relação acima, do seguinte modo: 25% dos mais leves como pequenos, os 50% seguintes como médios e os 25% restantes como grandes. Quais os valores de altura/comprimento que classificam os mexilhões como sendo pequenos, médios e grandes?

25) O comprimento do corpo de um animal da espécie X segue uma distribuição normal, com média  $\mu$  e desvio padrão  $\sigma = 10\text{mm}$ .

a) De quanto deve ser o comprimento médio do corpo,  $\mu$ , para que apenas 10% dos animais tenham menos do que 500 mm?

b) Com o comprimento médio  $\mu$  obtido no item a), qual a probabilidade de que o comprimento total de 4 animais escolhidos ao acaso seja inferior a 2000 mm?

26) Uma máquina automática para encher sacos de sementes de milho híbridos está regulada para que o peso médio de sementes em cada saco seja de 20 Kg e o desvio padrão 0,2 Kg. Pode-se admitir que a distribuição da variável seja normal.

a) Qual a porcentagem de sacos em que o peso de sementes não se desvia da média em mais que dois desvios padrões?

b) O que acontecerá com porcentagem do item a) se a máquina for regulada de forma que a média seja 30 Kg e o desvio padrão 0,30 Kg?

c) Qual a probabilidade de encontrar um saco com mais de 20,5  $Kg$ ?

27) O diâmetro de certa espécie de árvore é uma variável aleatória com distribuição normal de média 40 $cm$  e desvio padrão 5 $cm$ . Se o diâmetro de uma árvore diferir da média de mais de 8 $cm$ , esta árvore é vendida por 10 $u.m$ ; caso contrário, é vendida por 20 $u.m$ .

a) Qual o preço médio de venda de cada árvore?

b) Qual é probabilidade de encontrar uma árvore com diâmetro entre 35 $cm$  e 45 $cm$ ?

28) Uma máquina de encher sacos de sementes de milho híbridos pode ser regulada de modo a descarregar uma média de  $\mu$   $Kg$  de sementes por saco. Admitindo que a quantidade em  $Kg$  necessária para encher um saco de sementes tenha distribuição normal, com desvio padrão de 0,3  $Kg$  por saco, calcule o valor de  $\mu$  de modo que sacos com mais de 22  $Kg$  ocorram apenas 1% das vezes.

29) Numa espécie de planta, por exemplo, a *Grevilha robusta*, a altura das plantas tem distribuição aproximadamente normal com desvio padrão de 4 $m$  e tal que 20% das plantas possuem alturas inferiores a 10 $m$ .

a) Calcule a proporção de árvores com altura superior a 18 $m$ .

b) Nesta população, qual é um valor da altura,  $A$ , cuja probabilidade de se encontrar uma árvore com altura maior ou igual a  $A$  seja de 10%, isto é,  $P(\text{altura} > A) = 10\%$ ?

30) Num povoamento florestal, considere que  $X$ , a medida do diâmetro à altura do peito (DAP), tenha distribuição normal (aproximadamente), com média  $\mu = 16,2cm$  e variância  $\sigma^2 = 11,56cm^2$ . Tomando-se ao acaso uma árvore do povoamento, pergunta-se: qual a probabilidade do DAP da árvore selecionada casualmente ser:

a) menor que 19,6 $cm$ ?

b) entre 14,0 e 16,0 $cm$ ?

c) qual o valor de DAP, que é superado por 10% dos valores?

### Exercícios Sobre Aproximação Normal à Binomial

31) De um lote de sementes, extraímos uma amostra de 100 sementes ao acaso; se 10% das sementes do lote não germinam, calcular a probabilidade de 12 sementes da amostra não germinarem.

32) De um pomar de pessegueiros são colhidos 2000 frutos; se a probabilidade de um fruto colhido ser classificado como ideal para a indústria de conservas é de 0,45, qual a probabilidade de 950 ou mais pêssegos obterem a classificação ideal?

33) Dois pesquisadores desenvolveram a técnica do transplante nuclear, segundo a qual,

um núcleo de uma célula de um dos últimos estágios de um embrião é transplantado para um zigoto (célula única, o óvulo fertilizado), para se verificar se o núcleo pode desenvolver-se normalmente. Se a probabilidade de que um transplante singelo seja bem sucedido é de 0,65, qual é a probabilidade de mais de 70 transplantes, em 100, obterem sucesso?

34) Para comparar a produção de duas cultivares,  $A$  e  $B$ , de uma determinada espécie vegetal, foram plantadas 60 áreas de terras de  $5 \times 3 = 15m^2$ , separadas em pares (30 pares) segundo a fertilidade do solo (isto é, existe homogeneidade da fertilidade dentro do par; entre os pares há diferença de fertilidade). Uma área de cada par recebeu a cultivar  $A$  e a outra área do par recebeu a cultivar  $B$ . Dentro de cada par foi feito um sorteio para saber qual a área que receberia a cultivar  $A$  e a  $B$ , respectivamente. Veja Figura 5.20. As produções nas 60 áreas (30 pares) das cultivares foram anotadas, sendo que em 19 pares a cultivar  $A$  apresentou uma maior produção do que a cultivar  $B$ . Se não há qualquer diferença entre as duas cultivares quanto a produção, isto é, o fato da cultivar  $A$  produzir mais do que a  $B$ , ou a cultivar  $B$  produzir mais do que a  $A$ , é aleatória, casual. Nesse caso, a probabilidade  $p$  de que a produção de uma cultivar de um par seja maior que a produção de uma cultivar de um par seja maior que a produção da outra cultivar será de 0,50.

a) Calcular a probabilidade desse resultado (em 19 pares a cultivar  $A$  teve maior produção) ter ocorrido por acaso. Use a aproximação normal à binomial.

b) Caso essa probabilidade for menor ou igual a 5%, considera-se que esse resultado obtido no experimento indica que a cultivar  $A$  é realmente mais produtiva que a cultivar  $B$ . Qual a sua conclusão?

35) Dez por cento dos pés de milho que foram inoculados contra um certo vírus tiveram reações sérias indesejáveis devido a inoculação. Calcule a probabilidade de que mais de 20 pés de milho dos 150 pés inoculados tenham tais reações.

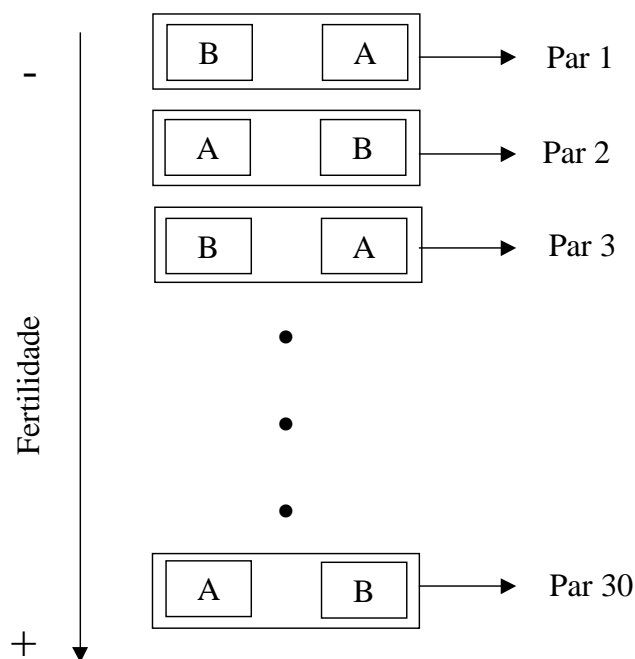


Figura 5.20: Projeto experimental para comparação de duas cultivares

## 6 Introdução à Inferência Estatística

### 6.1 Introdução

Na seção 1.1, foi dito que um curso básico de estatística deve contemplar, de forma esquemática, os seguintes assuntos: Análise Exploratória de Dados, Modelos Probabilísticos, Amostragem e Inferência Estatística. Esquemáticamente, isso foi ilustrado na Figura ??.

Agora, nós vamos ver como reunir a Análise Exploratória de Dados, Modelos Probabilísticos e Amostragem, para podermos desenvolver um estudo importantíssimo dentro da estatística, conhecido pelo nome de Inferência Estatística, isto é, como tirar conclusões sobre parâmetros da população (por exemplo, sobre médias ( $\mu$ ), proporções ( $\pi$ ), variâncias ( $\sigma^2$ )) com base no estudo de somente uma parte da *população*, ou seja, com base em uma *amostra*. Veja Figura 6.1. Uma população, em estatística, é formada por todos os valores possíveis de uma característica desejável. Esses valores não precisam ser todos diferentes, nem um número finito. Exemplos de populações: 1) todos os valores possíveis da produção de milho em quilogramas por hectare (kg/ha); 2) todos os pesos ao nascer de coelhos da raça gigante, em gramas; 3) todos os valores de diâmetros de Biomphalarias do Poção do Córrego Grande; 4) todos os valores de micronúcleos de roedores de uma região poluída. Por outro

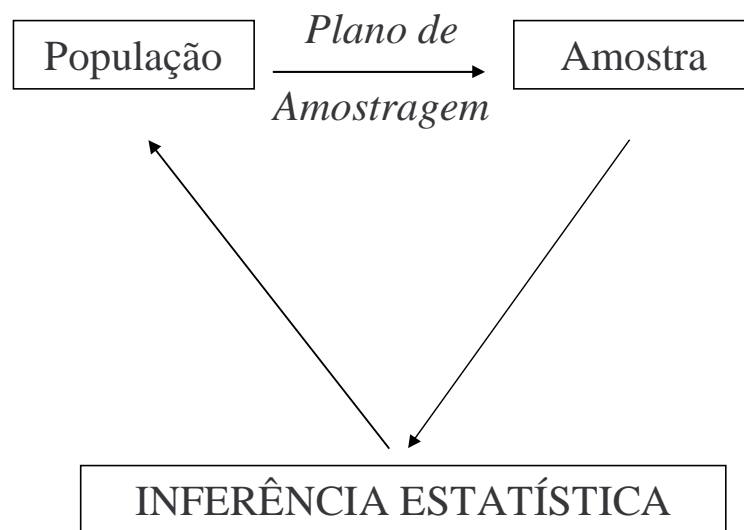


Figura 6.1: O raciocínio indutivo da inferência estatística

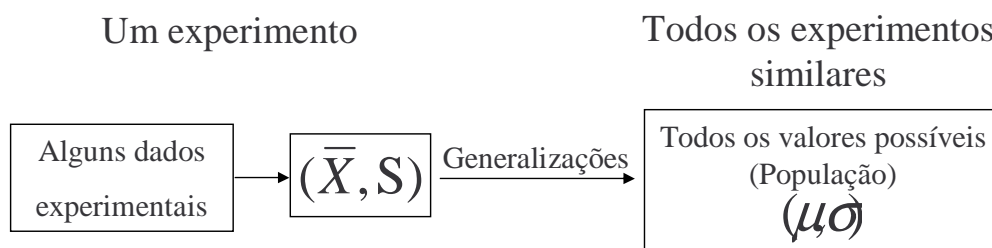


Figura 6.2: O processo de generalização com experimentos

lado, uma amostra, é uma parte (subconjunto) da população. Exemplos de amostras: 1) os rendimentos de milho, em kg/ha, de uma amostra de 5 unidades experimentais (canteiros); 2) os pesos ao nascer de uma ninhada de coelhos da raça gigante; 3) os diâmetros de uma amostra de 30 *Biomphalaria* do Poço do Córrego Grande; 4) os valores de micronúcleos de uma amostra de 25 roedores.

Num experimento, obtemos alguns dados experimentais (amostra) através do uso de repetições dos tratamentos, e generalizamos os resultados para todos os experimentos similares (população conceitual). A Figura 6.2 ilustra o processo de generalização numa situação experimental.

Em inferência estatística utilizamos uma notação própria para diferenciar medidas usadas para descrever características da amostra e da população.

Uma *estatística* é uma medida usada para descrever uma característica da amostra. Assim, por exemplo,  $\bar{X}$  a média da amostra;  $S$  o desvio padrão da amostra e  $P$  a proporção da amostra, são estatísticas.

Por outro lado, um *parâmetro* é uma medida usada para descrever uma característica da população. Geralmente são representados por letras gregas, assim, por exemplo,  $\mu$  representa a média populacional;  $\pi$  representa a proporção populacional e  $\sigma$  representa o desvio padrão populacional.

Quando uma estatística assume um determinado valor, temos o que denomina-se de *estimativa*. Temos os dados de uma particular amostra, calculamos o valor da estatística de interesse, este valor é a nossa estimativa. Por exemplo, a estimativa da produção média por planta da cultivar Gala é de  $\bar{x} = 84 \text{ kg/planta}$ . Outro exemplo, a estimativa da proporção de peixes com comprimento total menor do que 50 mm é  $p = 46\%$ .

Um resumo da notação utilizada é apresentada no quadro a seguir.

Notação utilizada para representar parâmetros, estatísticas (ou estimadores) e estimativas			
Característica	Parâmetros	Estatísticas ou Estimadores	Estimativas
Média	$\mu$	$\bar{X}$	$\bar{x}$
Proporção	$\pi$	$P$	$p$
Variância	$\sigma^2$	$S^2$	$s^2$
Desvio padrão	$\sigma$	$S$	$s$

Os dois problemas básicos da inferência estatística são: Estimação e Testes de Hipóteses. Vamos, através de um exemplo, ilustrar estas duas situações.

**Exemplo:** um pesquisador está interessado em avaliar a produção média por planta,  $\mu$ , da cultivar de maçã denominada Gala, para as seguintes condições: plantas com idade de aproximadamente 5 anos, em bom estado fitossanitário, cultivadas com alta tecnologia e para a região I do zoneamento agroclimático de Santa Catarina. A população é formada por todas as plantas da cultivar Gala nas condições citadas. Mais especificamente, a população é constituída por todos os valores de produção por planta. Para essa finalidade, o pesquisador vai coletar uma amostra aleatória de, por exemplo, 10 plantas<sup>7</sup>, da referida cultivar nas condições descritas. Uma amostra de valores de produções por planta, em kg, foi:

Plantas	1	2	3	4	5	6	7	8	9	10	Média ( $\bar{x}$ )	Desvio padrão ( $s$ )
Produção	84	82	90	86	80	91	85	79	81	82	84	4,0552

Com os 10 valores de produção/planta podemos calcular uma *estimativa* da produção média verdadeira por planta,  $\bar{x} = 84 \text{ kg}$ . Portanto, estamos usando a média da amostra,  $\bar{X}$ , como estimador da média verdadeira,  $\mu$ . Essa estimativa é chamada de estimativa *pontual*, pois origina um único valor. Esse é um raciocínio tipicamente indutivo, onde se parte do particular (amostra) para o geral (população). Esse é um exemplo de problema de estimação.

<sup>7</sup>O tamanho da amostra deve ser determinado com o auxílio da estatística.

Tabela 6.1: Produção por planta, em Kg, de maçãs das cultivares Gala e Golden

Variedades	Plantas										Média ( $\bar{x}$ )	Desvio padrão(s)
	1	2	3	4	5	6	7	8	9	10		
Gala	84	82	90	86	80	91	85	79	81	82	84,0	4,06
Golden	95	102	85	93	104	89	98	99	107	106	97,8	7,32

Um fato importante que se observa quando trabalhamos com amostras, é que sempre vamos ter que a média verdadeira,  $\mu$ , é igual a média na amostra,  $\bar{X}$ , mais um erro de amostragem. A representação disso é dada por:

$$\mu = \bar{X} + \text{erro amostral}$$

onde o termo *erro amostral* é a diferença entre a estatística ( $\bar{X}$ ) e o parâmetro ( $\mu$ ). Apesar do nome *erro*, isto não quer dizer que a amostragem foi feita de forma errada e, que, portanto, deve-se coletar uma nova amostra. Esse valor pode ser negativo ou positivo, pequeno, nulo ou grande. Em todas as pesquisas vamos estar envolvidos com o erro amostral. Dizemos que uma estimativa é *precisa*, se tivermos alto grau de *confiança* de que o erro amostral associado a estimativa em questão, é pequeno. A *precisão* e a *confiança* são dois conceitos chaves nesse estudo. A precisão pode ser entendida como a diferença máxima entre a estimativa e o parâmetro que o pesquisador deseja considerar no seu estudo. Voltaremos a tratar deste assunto posteriormente.

Uma outra forma de estimação é através da construção de intervalos de confiança. Nesse caso, temos uma estimativa intervalar, isto é, temos um intervalo, dentro do qual esperamos que o valor populacional se encontre. Por exemplo, para os dados de produção/planta da cultivar Gala ao invés de dizer que a estimativa é de 84 kg/planta, podemos dizer que a média está no intervalo de 81,10 a 86,90.

Essa forma de estimação é muito mais informativa que a estimativa pontual. O pesquisador pode verificar se esse intervalo é curto (preciso, informativo) ou se é muito amplo (pouco informativo). Isto será estudado no capítulo 7.

O segundo problema é o de teste de hipóteses sobre os parâmetros. Por exemplo, um pesquisador deseja saber se a produção média/planta da cultivar Gala é a mesma da produção média/planta da cultivar Golden. Para isso, foi obtida uma outra amostra aleatória de 10 plantas da cultivar Golden sob as mesmas condições descritas para a cultivar Gala. Os dados das duas amostras aleatórias estão apresentadas na Tabela 6.1.

As estimativas da produção média das duas cultivares, calculadas com os dados das duas amostras foram 84 Kg/planta e 97,8 kg/planta para as cultivares Gala e Golden, respectiva-

mente. Portanto, a diferença verificada entre as duas cultivares, com essas duas amostras, foi de 13,8 kg/planta a favor da cultivar Golden.

Observando-se os dados individualmente, verificamos que para as plantas 3 e 6, as produções na cultivar Gala foram superiores a da Golden. Portanto, podemos pensar que é perfeitamente possível obtermos um par de amostras, dentre todas as amostras possíveis de serem sorteadas, no qual a produção média da cultivar Gala é superior a da Golden. Isso devido simplesmente a amostragem, ou seja, variações devido a amostragem.

Assim, o problema que se apresenta, é o de decidir o que é uma diferença real, isto é, devido à cultivar, ou uma diferença casual, isto é, devido a variação casual na amostra. Este tópico será estudado no capítulo 8.

Logicamente, o pesquisador pretende generalizar os resultados obtidos na análise estatística, isto é, ele deseja saber se há diferença significativa entre as médias verdadeiras  $\mu_{Gala}$  e  $\mu_{Golden}$  (desconhecidas pelo pesquisador). Como ele está trabalhando com duas amostras aleatórias, dentre um grande número de possíveis amostras, ele não pode fazer afirmações com 100% de certeza, mas ele pode perfeitamente fazer uma afirmação probabilística, indicando a probabilidade de erro ao fazer uma afirmação sobre uma hipótese em teste. Para isso, utilizaremos as distribuições de probabilidades, como veremos nas próximas seções.

## 6.2 Como Selecionar Uma Amostra

Como vimos nos exemplos anteriores, um pesquisador trabalha com apenas uma parte da população, isto é, com uma amostra. A maneira como é selecionada uma amostra é de extrema importância, pois é através dos dados amostrais que serão calculadas as estimativas dos parâmetros desconhecidos da população e, também, serão feitas afirmativas sobre os mesmos.

É fácil de se imaginar que é fundamental o pesquisador ter um bom conhecimento da população, pois quanto mais ele conhecer a população, mais informativa será a amostra selecionada. Por exemplo, se um Biólogo deseja fazer um estudo sobre a poluição de uma baía em peixes do gênero Bagre, ele vai precisar coletar uma amostra de peixes desse gênero e avaliar o número de micronúcleos em 5000 células. Para se ter uma amostra que permita tirar conclusões válidas, é necessário ter um bom conhecimento sobre o comportamento desse tipo de peixe.

As situações de amostragem, nas ciências biológicas e agrônômicas, na prática, são



inúmeras e podem ser mais ou menos complexas. Vejamos alguns exemplos de seleção de amostras.

**Exemplo 1.** Desejamos desenvolver uma pesquisa sócio-econômica, sobre os agricultores integrados à uma empresa, onde estamos interessados em diversos parâmetros relativos a atividade agropecuária. A população é constituída por todos os agricultores integrados à empresa. Podemos definir várias variáveis associadas a cada agricultor (propriedade agrícola). Essas variáveis serão observadas sobre uma amostra de 200 agricultores. Para selecionar os agricultores que participarão da amostra, sorteamos, com regras bem definidas, 200 fichas de um fichário, onde conste todos os agricultores integrados.

**Exemplo 2.** Estamos interessados numa pesquisa para tratar a glicemia<sup>8</sup> em peixes Mandi<sup>9</sup> através da injeção de glucagônio. Um parâmetro de interesse seria o teor de açúcar no sangue desses peixes. A população é constituída por todos os peixes Mandi do local de estudo. Nesse caso, o pesquisador não tem acesso a toda a população de peixes Mandi. Assim, ele vai selecionar (amostra) aqueles que estão acessíveis. Por exemplo, se desejamos uma amostra de 15 peixes, o pesquisador poderá utilizar os peixes capturados com uma tarrafa, ou selecionar alguns peixes em um laboratório. É importante que os peixes selecionados para participarem da amostra, tenham as mesmas características da população.

**Exemplo 3.** Numa pesquisa sobre propriedades químicas de uvas da cultivar Cabernet, com idade de  $\pm 6$  anos, nas seguintes condições (tratamentos):

- Tratamento 1: plantas sadias;
- Tratamento 2: plantas com infecção média da virose do enrolamento;
- Tratamento 3: plantas com infecção forte da virose do enrolamento,

podemos estar interessados no parâmetro teor médio de açúcar da cultivar, sob cada uma das três condições. As populações são formadas por todas as plantas adultas da cultivar cabernet na região considerada para cada um dos tratamentos. Nesse caso, o pesquisador seleciona intencionalmente 10 plantas, para cada tratamento, de um parreiral. Portanto, temos 3 amostras, cada uma de tamanho igual a 10.

**Exemplo 4.** Em estudos sobre a distribuição e abundância de primatas na natureza, deseja-se conhecer o número de animais em uma determinada área e a área de ocorrência atual desses animais.

---

<sup>8</sup>Presença de glicose no sangue em cota fisiológica.

<sup>9</sup>Designação comum a várias espécies de peixes siluriformes, especialmente da família dos pimelodídeos.

Necessita-se fazer a abertura de algumas trilhas (transectos) na mata, por onde serão feitas as observações e caminhadas. O tamanho da amostra corresponde ao número de vezes que cada trilha deve ser caminhada, por exemplo, 40 repetições. A contagem por transectos lineares é um dos métodos mais usados para a estimativa da densidade de várias espécies de animais.

Em resumo, a obtenção de soluções adequadas para problemas de amostragem exige, em geral, muito bom senso, conhecimento do fenômeno biológico e experiência. Como percebemos, existem muitas maneiras de se retirar uma amostra de uma população. Existe um campo dentro da estatística que trata especificamente dos projetos para retirar amostras. Entretanto, o plano mais simples e talvez o mais importante em estatística é o plano conhecido como Amostragem Casual Simples e, é desse plano, que trataremos agora.

### 6.3 Amostragem Aleatória Simples

O que caracteriza este tipo de amostragem, é que todos os elementos da população tem probabilidade igual de pertencer a amostra e, todas as possíveis amostras, tem igual probabilidade de serem sorteadas.

Vamos, a partir de um exemplo, apresentar com detalhes uma definição precisa, ou seja, matemática, de amostra aleatória simples. Essa é a definição mais adequada à experimentação.

**Exemplo:** considere a população das 17 alunas da turma de estatística básica do curso de Ciências Biológicas. Na Tabela 6.2 está representada a distribuição dos valores de altura ( $X$ ) das alunas da turma juntamente com a distribuição de probabilidades.

Obter uma amostra aleatória simples dessa população, equivale matematicamente a colocar numa caixa de papel, 17 pedacinhos de papéis numerados de 1 a 17 e, após, proceder a um sorteio. A altura  $x$  do primeiro elemento selecionado pode assumir qualquer valor da primeira coluna da Tabela 6.2, com probabilidades apresentadas na terceira coluna da mesma. Vamos representar essa primeira observação por  $X_1$ . Vamos repor esse papel na caixa. O segundo papel escolhido aleatoriamente, também pode assumir qualquer valor  $x$  da primeira coluna da Tabela 6.2, com as probabilidades dadas na terceira coluna. Vamos representar essa segunda observação por  $X_2$ . Repomos esse papel na caixa. Repetimos esse procedimento até selecionar todos os  $n$  elementos da amostra  $(X_1, X_2, \dots, X_n)$ .

Observe que a distribuição de probabilidades de  $X_1$ , é a mesma distribuição da população, isto é, de  $X$ ; a distribuição de probabilidades de  $X_2$ , é a mesma distribuição de probabilidades

Tabela 6.2: Distribuição de frequências das alturas das alunas do curso de estatística básica.

Altura	Frequências absolutas	Frequências relativas
1,56	1	0,059
1,60	4	0,235
1,61	2	0,117
1,62	1	0,059
1,63	1	0,059
1,65	1	0,059
1,67	2	0,117
1,68	1	0,059
1,70	1	0,059
1,72	2	0,117
1,73	1	0,059
Total	45	1,000

de  $X$ ; e, assim, para  $X_3, \dots, X_n$ .

Então, diz-se que  $X_1, X_2, \dots, X_n$  tem a mesma distribuição de probabilidades, e qual é esta distribuição? a distribuição da população, ou seja, de  $X$ . Dizemos que  $X_1, X_2, \dots, X_n$  são *identicamente distribuídas*.

De um modo geral, as observações  $X_1$  (altura da primeira aluna sorteada),  $X_2$  (altura da segunda aluna sorteada), ...,  $X_n$  (altura da  $n$ -ésima aluna sorteada) serão *independentes* caso a amostragem seja feita com reposição. Na prática, em particular na experimentação, a amostragem é feita sem reposição, o que acarreta perda da independência, porém para populações grandes os resultados (propriedades) importantes para a inferência podem ser considerados equivalentes, como será mostrado mais adiante. Agora, podemos definir amostra aleatória simples de forma precisa.

**Definição:** uma amostra aleatória simples de tamanho  $n$ , de uma variável aleatória  $X$ , é aquela cujas  $n$  observações  $X_1, X_2, \dots, X_n$  são independentes e identicamente distribuídas.

### 6.3.1 Obtenção de Uma Amostra Aleatória Simples

A obtenção de uma amostra verdadeiramente aleatória vai depender muito da situação da população de interesse. Frequentemente, não é possível obter-se uma amostra aleatória -

aquela em que cada membro da população pode participar da amostra - pois a mesma não se adapta aos dados da pesquisa. Por exemplo, na pesquisa para tratar a glicemia em peixes Mandi através da injeção de glucagônio ( $200 \mu g$ ), os peixes que formam a amostra são os acessíveis, que estão no laboratório ou que serão capturados com uma tarrafa.

Na pesquisa para estudar o teor médio de açúcar da cultivar cabernet, em três condições (tratamentos) diferentes, quais sejam: 1) plantas sadias; 2) plantas com infecção média da virose do enrolamento e 3) plantas com infecção forte da virose do enrolamento, o pesquisador selecionou intencionalmente 10 plantas de cada tratamento de um parreiral.

Uma outra situação, é o caso da pesquisa sócio-econômica dos agricultores de um determinado município, em que temos uma listagem de todos os agricultores desse município. Nesse caso, podemos numerar todos os agricultores de 1 até  $N$ , sorteando-se, a seguir, por meio de um processo bem definido, que nos garanta a aleatoriedade,  $n$  agricultores dessa seqüência, os quais formarão a amostra.

Um procedimento útil para realizar o sorteio é usar a tabela de números aleatórios. Esta tabela é dada no apêndice 7. As tabelas de números aleatórios contém os 10 algarismos (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) e são formadas por sucessivos sorteios do conjunto deles. Essas tabelas foram exaustivamente testadas e a sua aleatoriedade é garantida.

Na tabela de números aleatórios, esses números podem ser lidos isoladamente, ou em grupos de dois, três, quatro, ou mais algarismos; podem ser lidos em qualquer ordem: por colunas em qualquer sentido, por linhas em qualquer sentido, diagonalmente. Vamos, através de um exemplo, fazer uma ilustração do uso da tabela de números aleatórios.

**Exemplo.** A finalidade da pesquisa é determinar o peso médio, o peso total, em toneladas, e a porcentagem de fornecimentos abaixo de  $X$  toneladas, de cana-de-açúcar, entregue pelos fornecedores à indústria. A população é formada por 250 fornecedores de cana-de-açúcar. Deseja-se sortear uma amostra de 21 fornecedores.

Primeiramente, os 250 fornecedores são listados por ordem alfabética. Após isso, a população é numerada de 1 até 250. Como a identificação exigiu números com 3 algarismos, é necessário lermos números de 3 algarismos na tabela de números aleatórios. Utilizando a tabela (Apêndice 7), sorteia-se as colunas 8, 9 e 10, e, vamos, ler percorrendo as colunas de

cima para baixo, os números encontrados foram:

826	<i>97</i>	820	<i>176</i>	452	328	868	817	402	253
<i>232</i>	497	<i>174</i>	397	419	519	699	<i>13</i>	415	772
406	<i>207</i>	313	348	963	476	932	870	827	718
<i>140</i>	490	499	<i>134</i>	719	<i>209</i>	681	471	612	424
537	<i>56</i>	331	742	981	662	625	957	450	331
902	910	<i>7</i>	929	775	785	935	<i>30</i>	573	941
611	264	893	441	809	813	<i>113</i>	<i>79</i>	933	<i>136</i>
883	<i>233</i>	640	655	520	815	440	<i>69</i>	798	
928	741	<i>234</i>	746	312	616	<i>173</i>	995	496	
<i>138</i>	694	806	952	876	<i>3</i>	876	436	379	

A amostra selecionada é formada pelos números em *itálico* na matriz. Através desses números identifica-se os agricultores que vão fazer parte do estudo.

Em resumo, para selecionarmos uma amostra, com a utilização da tabela de números aleatórios, devemos seguir os seguintes passos:

- 1) faça uma lista dos itens da população;
- 2) numere os itens da lista, a começar de 1;
- 3) leia os números na tabela de números aleatórios de modo que o número de algarismos em cada um seja igual ao número de algarismos do último número da sua listagem. Exemplo, se o último número é 1370, devem ser lidos números com 4 algarismos. Sortear o início da leitura dos números, por exemplo, iniciar na quinta linha da esquerda para a direita;
- 4) despreze os números que não correspondem aos números da lista ou que sejam repetições de números lidos anteriormente (amostragem sem reposição). Continue o processo até ter o número desejado de observações;
- 5) use os números assim identificados para coletar os itens que vão fazer parte da amostra.

**Exemplo.** Um pesquisador deseja comparar os teores médios de proteína de três cultivares de cevada. Para executar o experimento ele dispõe de uma área de terra homogênea (mesma fertilidade, mesma umidade, etc.) de tamanho  $288\text{ m}^2$ . Portanto, as três cultivares vão ser comparadas em igualdade de condições. Um princípio básico da experimentação

<b>1</b> Cultivar 1	<b>2</b> Cultivar 2	<b>3</b> Cultivar 2
<b>4</b> Cultivar 3	<b>5</b> Cultivar 2	<b>6</b> Cultivar 3
<b>7</b> Cultivar 3	<b>8</b> Cultivar 2	<b>9</b> Cultivar 1
<b>10</b> Cultivar 3	<b>11</b> Cultivar 1	<b>12</b> Cultivar 1

Figura 6.3: Croqui de campo, indicando o processo de aleatorização do experimento

é o uso de repetições, ou seja, são necessários pelo menos dois valores para cada cultivar. Assim, a área total vai ser dividida em 12 canteiros de tamanhos  $6m \times 4m$ , totalizando  $24m^2$ /canteiro. O número de repetições (tamanho da amostra) por cultivar é 4. O croqui da área é dado na Figura 6.3.

A aleatorização das três cultivares foi feita do seguinte modo:

- enumerou-se os canteiros de 1 a 12, como indicado no canto superior direito da Figura 6.3;
- leu-se números na tabela de número aleatórios (Apêndice 7), com dois algarismos, selecionando-se apenas os valores que estão na faixa de 1 a 12 e desprezando-se os valores repetidos. Sorteou-se a sexta linha para iniciar a leitura dos números, da esquerda para a direita, produzindo a seguinte listagem: {12, 1, 11, 9, 5, 8, 3, 2, 7, 10, 6, 4};
- os canteiros 12, 1, 11 e 9 receberam a cultivar 1; os canteiros 5, 8, 3 e 2 receberam a cultivar 2 e os canteiros 7, 10, 6 e 4 receberam a cultivar 3.

**Exercício.** A finalidade da pesquisa é o de determinar o número de insetos/quadrante e o total de insetos na área. Sortear 15 quadrados de um total de 100, de  $10 m^2$  cada.

As técnicas estatísticas que veremos nas seções subseqüentes deste livro, pressupõem que a amostra foi selecionada aleatoriamente ou por algum processo equivalente. Embora, para os dados de uma pesquisa, nem sempre uma amostra aleatória é possível, tomando-se algumas precauções especiais no delineamento da pesquisa, esta suposição (amostra aleatória) pode ser pertinente, por exemplo, num experimento, fazendo-se a casualização (sorteio) dos tratamentos às unidades experimentais (Box *et al.*, 1978).

## 6.4 Distribuições Amostrais

O objetivo pelo qual selecionamos uma amostra, é para obter estimativas de parâmetros desconhecidos da população ( $\mu$ ,  $\sigma^2$ ,  $\pi$ ), fazer afirmações sobre os mesmos e, também, para auxiliar na verificação da forma (distribuição) como a estatística se distribui.

Vamos raciocinar em cima da seguinte pergunta, relativa ao exemplo em que o pesquisador está interessado em avaliar a produção média ( $\mu$ ), por planta, da cultivar de maçã Gala. O valor da produção média por planta ( $\bar{x}$ ), numa amostra de 10 plantas, vai ser um valor próximo da verdadeira produção média referente a todas as plantas da cultivar Gala?<sup>10</sup>

Como o valor da verdadeira produção média por planta é desconhecido, responderemos a pergunta através do estudo de como se distribuem os possíveis valores de  $\bar{x}$ .

Quando se extraem repetidas amostras de uma mesma população, provavelmente os valores das médias amostrais são todos diferentes e, também, diferentes da verdadeira média da população, simplesmente devido ao fato de serem considerados elementos diferentes da população. Como exemplo, veja a terceira coluna (amostras) e a sexta coluna (médias amostrais) da Tabela 6.4. Esta variabilidade é conhecida como *variabilidade amostral*. Portanto, torna-se imprescindível, quando vamos fazer inferências sobre parâmetros populacionais, levar em consideração a distribuição amostral. Estas são descritas pelas distribuições de probabilidades, como a distribuição normal e a binomial, por exemplo. Como exemplo de distribuição amostral temos os 16 valores das médias amostrais dadas na Tabela 6.4, obtidas das 16 amostras possíveis de tamanho  $n = 2$ .

Os valores da estatística, calculados nas amostras, vão formar uma nova população (de médias, de variâncias, de proporções), cuja distribuição recebe o nome de distribuição amostral (de médias, de variâncias, de proporções). Assim o conjunto:

- $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$  é denominado de distribuição amostral das médias ou da média; o conjunto:

---

<sup>10</sup>Depende da estatística, do tamanho da amostra e da variabilidade na população.

Tabela 6.3: Pesos de 4 suínos

Suínos	Peso (kg)
A	68
B	80
C	84
D	87

- $p_1, p_2, \dots, p_k$  é denominado de distribuição amostral das proporções ou da proporção; o conjunto:
- $s_1^2, s_2^2, \dots, s_k^2$  é denominado de distribuição amostral da variância.

onde  $k$  é o número de todas as amostras possíveis. No exemplo temos  $k = 16$ .

Para ilustrar o desenvolvimento do estudo de distribuição amostral da proporção, da média e da variância, vejamos um exemplo simples. Salientamos, que esse exemplo, no qual a população é bastante pequena ( $N = 4$ ), e a amostragem é feita com reposição, tem fim puramente didático. Para a nossa finalidade, suponhamos conhecidos os pesos de todos os suínos, embora o agricultor só irá conhecer os pesos dos suínos amostrados. Os pesos dos animais são dados na Tabela 6.3.

Os valores dos parâmetros proporção( $\pi$ ) de suínos com peso inferior a 75 kg, peso médio( $\mu$ ), variância( $\sigma^2$ ) e desvio padrão( $\sigma$ ) são:

$$\begin{aligned}
 \pi &= 1/4 = 0,25 \\
 \mu &= \sum x/N = 79,75 \text{ kg.} \\
 \sigma^2 &= \frac{\sum (x - \mu)^2}{N} = \frac{208,75}{4} = 52,18 \text{ kg}^2. \\
 \sigma &= \sqrt{\sigma^2} = \sqrt{52,18} = 7,22 \text{ kg.}
 \end{aligned} \tag{6.1}$$

#### 6.4.1 Distribuição Amostral da Proporção

Vamos considerar que um suíno com menos de 75 kg obtenha um preço bastante inferior na agroindústria. Sabemos que existe apenas 1 suíno nessa situação na população, portanto, a proporção, ( $\pi$ ), de suínos com peso inferior a 75 kg na população é de  $1/4=0,25$ . O agricultor deseja estimar esta proporção de suínos com peso inferior a 75 kg observando alguns suínos da população. Para isso, ele decide selecionar uma amostra aleatória de 2



Tabela 6.4: Todas as amostras possíveis de tamanho igual a dois, com reposição, da população dada na Tabela 6.3, com as respectivas proporções, médias e variâncias amostrais.

Amostras	Amostras	Pesos	Proporção < 75 kg ( $P$ )	Média peso ( $\bar{X}$ )	Variância ( $\div 1$ ) ( $S^2$ )	Variância ( $\div 2$ ) ( $S^2$ )
1	A,B	68;80	1/2=0,50	74,0	72,0	36,00
2	A,C	68;84	1/2=0,50	76,0	128,0	64,00
3	A,D	68;87	1/2=0,50	77,5	180,5	90,25
4	B,C	80;84	0/2=0	82,0	8,0	4,00
5	B,D	80;87	0/2=0	83,5	24,5	12,25
6	C,D	84;87	0/2=0	85,5	4,5	2,25
7	B,A	80;68	1/2=0,50	74,0	72,0	36,00
8	C,A	84;68	1/2=0,50	76,0	128,0	64,00
9	D,A	87;68	1/2=0,50	77,5	180,5	90,25
10	C,B	84;80	0/2=0	82,0	8,0	4,00
11	D,B	87;80	0/2=0	83,5	24,5	12,25
12	D,C	87;84	0/2=0	85,5	4,5	2,25
13	A,A	68;68	2/2=1	68,0	0,0	0,00
14	B,B	80;80	0/2=0	80,0	0,0	0,00
15	C,C	84;84	0/2=0	84,0	0,0	0,00
16	D,D	87;87	0/2=0	87,0	0,0	0,00

suínos, com reposição, e usar a proporção encontrada nesta amostra, como estimativa da proporção na população. Isto nos sugere algumas perguntas fundamentais:

- quais as proporções amostrais que o agricultor pode obter?
- qual a probabilidade de cada uma?
- qual a forma da distribuição das proporções amostrais?
- qual a média da distribuição amostral das proporções?
- qual a variância da distribuição amostral das proporções?

Para responder a essas perguntas, precisamos construir a distribuição amostral das proporções. Na Tabela 6.4 temos todas as 16 amostras possíveis de tamanho igual a dois suínos ( $N^n = 4^2$ ) e as correspondentes 16 proporções amostrais. Observe, por exemplo, para a amostra (A,A), que não temos uma ordenação nos resultados, pois só temos um suíno A.

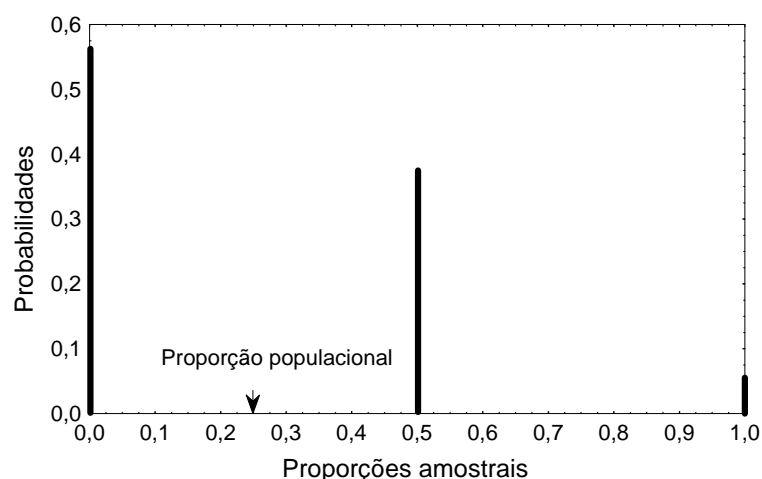


Figura 6.4: Distribuição amostral da proporção de suínos com peso abaixo de 75 kg, tamanho da amostra igual a 2.

A distribuição amostral da proporção de suínos com peso abaixo de 75 kg, para  $n = 2$ , fica:

$x_i$	0	1	2
$p = x_i/2$	0	0,50	1
$Prob(P = p)$	$9/16=0,5625$	$6/16=0,3750$	$1/16=0,0625$

onde  $x_i$  representa os possíveis valores da variável aleatória  $X$  "número de suínos com peso inferior a 75 kg na amostra".

A representação gráfica da distribuição amostral da proporção de suínos com peso inferior a 75 kg, para amostras de tamanho 2, extraídas de uma população com 4 suínos e, proporção populacional igual a 0,25, é dada na Figura 6.4.

Como a nossa amostra é de tamanho dois, é impossível obter uma proporção amostral com valor igual a 0,25. Porém, podemos observar que a verdadeira proporção (populacional) encontra-se entre 0 e 0,50, que são os valores que apresentam as maiores probabilidades de ocorrência na amostra. A forma da distribuição é assimétrica.

Como vimos no estudo de variáveis aleatórias discretas, contagens ou proporções, podem ser estudadas pela distribuição binomial. Vamos, para a situação em estudo, verificar se as pressuposições do modelo binomial são atendidas.

- 1) Temos duas repetições ( $n = 2$ ).
- 2) Em cada repetição, somente temos duas possibilidades: peso abaixo de 75 kg, ou, então, igual ou acima de 75 kg.

3) As probabilidades permanecem constantes em todas as repetições do experimento.

4) Os resultados das repetições são independentes.

Portanto, as pressuposições do modelo binomial são todas atendidas para a variável aleatória  $X$ . Da tabela da distribuição binomial (Apêndice 1), para  $n = 2$  e  $\pi = 0,25$ , obtemos:

$$Prob(P = 0) = P(X = 0) = 0,5625;$$

$$Prob(P = 0,50) = P(X = 1) = 0,3750;$$

$$Prob(P = 1) = P(X = 2) = 0,0620.$$

Como podemos ver, obtemos exatamente os mesmos valores da distribuição amostral das proporções. Assim, na prática, não é preciso construir empiricamente a distribuição amostral das proporções, pois, a mesma, pode ser obtida pela distribuição binomial.

Conforme visto na Seção 5.3.4 para determinadas condições de  $n$  e de  $\pi$ , a distribuição binomial pode ser bem aproximada pela distribuição normal com parâmetros

$$\mu_X = n\pi \text{ e } \sigma_X^2 = n\pi(1 - \pi)$$

conseqüentemente a distribuição das proporções também pode ser aproximada pela distribuição normal com parâmetros

$$\mu_P = \pi \text{ e } \sigma_P^2 = \frac{\pi(1 - \pi)}{n}$$

Observa-se que a média das proporções amostrais é sempre igual a proporção populacional ( $\mu_P = \pi$ ), e a medida que  $n$  aumenta, a variabilidade amostral diminui, isso implica que amostras grandes têm forte tendência a fornecer uma estatística amostral próxima do verdadeiro valor do parâmetro populacional.

O desvio padrão da distribuição amostral das proporções é dado por:

$$\sigma_P = \sqrt{\pi(1 - \pi)/n}.$$

Como pode ser verificado com os dados da Tabela 6.4:

$$\mu_P = (0,5 + 0,5 + 0,5 + 0 + 0 + 0 + 0,5 + 0,5 + 0,5 + 0 + 0 + 1 + 0 + 0 + 0)/16 = 4/16 = 0,25$$

e

$$\begin{aligned} \sigma_P^2 &= (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 \\ &+ (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 \\ &+ (1 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 / 16 = 1,5/16 = 0,09375. \end{aligned}$$

Note que  $\mu_P = 0,25 = \pi$  e  $\sigma_P^2 = 0,09375 = 0,18750/2 = \pi(1 - \pi)/n$ .

Usualmente os processos de seleção de amostras são feitos sem reposição e isto implica em uma correção na fórmula da variância e do desvio padrão da proporção, mas não na média. Dos dados da Tabela 6.4 vamos calcular a média, a variância e o desvio padrão das 12 primeiras amostras, que representam os possíveis valores se o processo fosse sem reposição. Temos:

$$\mu_P = (0,5 + 0,5 + 0,5 + 0 + 0 + 0 + 0,5 + 0,5 + 0,5 + 0 + 0 + 0)/12 = 3/12 = 0,25$$

e

$$\begin{aligned}\sigma_P^2 &= (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 \\ &+ (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0,5 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 + (0 - 0,25)^2 \\ &= 0,75/12 = 0,0625.\end{aligned}$$

A correção na variância é feita através do termo  $(N - n)/(N - 1)$ <sup>11</sup>, um valor menor ou igual a 1 (o valor 1 somente ocorre para  $n = 1$ ). Assim, a variância e o desvio padrão das proporções com a correção são dadas por:

$$\begin{aligned}\sigma_P^2 &= \frac{\pi(1 - \pi)}{n} \frac{N - n}{N - 1} \\ \sigma_P &= \sqrt{\frac{\pi(1 - \pi)}{n}} \sqrt{\frac{N - n}{N - 1}}.\end{aligned}$$

Podemos observar que a variância das proporções calculada com as 12 amostras (sem reposição),  $\sigma_P^2 = 0,0625$ , corresponde a  $(N - n)/(N - 1) = (4 - 2)/(4 - 1) = 2/3$ , da variância das proporções calculada com as 16 amostras (com reposição),  $\sigma_P^2 = 0,09375$ , ou seja,  $0,625 = (2/3) \times 0,09375$ .

No caso de populações infinitas, este termo pode ser desconsiderado, pois ele é aproximadamente igual a 1. Nos casos de populações finitas, este termo também pode ser ignorado quando o tamanho da amostra for menor do que 5% do tamanho da população ( $n < 0,05N$ ).

**Exemplo 1.** Um processo de encher garrafas de vinho dá em média 10% de mal enchidas. Extraída uma amostra de 125 garrafas, qual a probabilidade de que a proporção amostral de garrafas mal enchidas esteja entre 9% e 11%?

Como  $n$  é grande,  $\pi$  não é próximo de zero e  $n\pi = 125(0,10) = 12,5$ , podemos considerar a distribuição normal para aproximar probabilidades binomiais. Nesse exemplo, a população

---

<sup>11</sup>Este termo é proveniente da variância da distribuição hipergeométrica, distribuição exata dos dados quando a amostragem é feita sem reposição. Detalhes em Bussab & Morettin (1987)

é infinita, pois ela é formada por todas as garrafas enchidas e que virão a ser enchidas, portanto, não há necessidade de se fazer a correção para população finita no desvio padrão. No exemplo, a média e o desvio padrão valem:

$$\pi = 0,10 \quad (1 - \pi) = 1 - 0,10 = 0,90 \quad \sigma_P = \sqrt{\frac{\pi(1 - \pi)}{n}} = \sqrt{\frac{0,10(0,90)}{125}} = 0,0268.$$

Vamos calcular os valores de  $z_1$  e  $z_2$ :

$$z_1 = \frac{p - \pi}{\sigma_P} = \frac{0,09 - 0,10}{0,0268} = -0,37.$$

$$z_2 = \frac{p - \pi}{\sigma_P} = \frac{0,11 - 0,10}{0,0268} = 0,37.$$

Portanto, a probabilidade vale:

$$Prob(0,09 \leq P \leq 0,11) = Prob(-0,37 \leq z \leq 0,37) = 2Prob(0 \leq z \leq 0,37) = 2(0,1443) = 0,2886 = 28,86\%.$$

**Exemplo 2.** Sabe-se que 5% das latas de pêssego industrializados pela indústria A, em uma jornada de trabalho, apresentam peso drenado fora dos padrões. Qual a probabilidade de, em uma amostra de 200 latas, apresentarem-se fora dos padrões: a) 6% ou mais? e b) 4% ou menos?

Temos  $n$  suficientemente grande e  $n\pi = 200(0,05) = 10$ , portanto, em termos práticos, podemos usar a distribuição normal como aproximação da distribuição binomial. Vamos considerar a população como sendo infinita. A média e o desvio padrão são dados por:

$$\pi = 0,05 \quad (1 - \pi) = (1 - 0,05) = 0,95 \quad \sigma_P = \sqrt{\frac{0,05(0,95)}{200}} = 0,0154.$$

Para o item a) a probabilidade vale:

$$z = \frac{0,06 - 0,05}{0,0154} = 0,65.$$

$$Prob(P \geq 0,06) = Prob(z \geq 0,65) = 0,5 - Prob(0 \leq z \leq 0,65) = 0,5 - 0,2422 = 0,2578 = 25,78\%.$$

Para o item b) a probabilidade vale:

$$z = \frac{0,04 - 0,05}{0,0154} = -0,65.$$

$$Prob(P \leq 0,04) = Prob(z \leq -0,65) = 0,5 - 0,2422 = 0,2578 = 25,78\%.$$

**Exemplo 3. (Bussab & Morettin, 1987).** Um procedimento de controle de qualidade foi planejado para garantir um máximo de 10% de itens defeituosos na produção. A cada

15 minutos sorteia-se uma amostra de 50 peças e, havendo mais de 15% de defeituosas, pára-se a produção para verificações. Qual a probabilidade de uma parada desnecessária? Observação: os diferentes valores de  $p$  são obtidos devido à variação aleatória na amostra. Mesmo a população tendo uma proporção  $\pi = 10\%$ , podemos obter uma amostra com  $p$  acima de 15% devido puramente ao "azar" de selecionar a amostra. Por isso a pergunta acima: qual a probabilidade de uma parada desnecessária? ou seja, encontrar um  $p \geq 15\%$  quando, na verdade, temos  $\pi = 10\%$ .

Temos que  $n\pi = 50(0,10) = 5$ . Obtemos para a média e o desvio padrão os seguintes valores:

$$\pi = 0,10 \quad (1 - \pi) = (1 - 0,10) = 0,90 \quad \sigma_P = \sqrt{\frac{0,10(0,90)}{50}} = 0,0424.$$

$$z = \frac{0,15 - 0,10}{0,0424} = 1,18.$$

$$Prob(P \geq 0,15) = Prob(z \geq 1,18) = 0,1190 = 11,90\%.$$

#### 6.4.2 Distribuição Amostral da Média

Passamos agora a estudar a distribuição amostral da estatística  $\bar{X}$ , a média amostral.

Para mostrarmos os principais resultados (propriedades) de uma distribuição amostral de médias, retornemos ao exemplo da população de 4 suínos, dada na Tabela 6.3, e que, o agricultor, agora, deseja estimar o peso médio dos suínos ( $\mu$ ).

Com o objetivo de avaliar a média populacional ( $\mu = 79,75 \text{ kg}$ ), o agricultor decide selecionar uma amostra aleatória simples de tamanho 2, com reposição. De forma análoga ao que foi feito para a proporção amostral, devemos observar todas as possíveis amostras de tamanho  $n = 2$ , com reposição, e para cada uma das amostras calcular a média ( $\bar{x}$ ). Os resultados possíveis são apresentados na Tabela 6.4.

Os valores da média, variância e desvio padrão da distribuição amostral das médias são:

$$\begin{aligned} \mu_{\bar{X}} &= \frac{(74,0 + 76,0 + 77,5 + \dots + 80,0 + 84,0 + 87,0)}{16} \\ &= 79,75 \text{ kg}. \end{aligned}$$

$$\begin{aligned} \sigma_{\bar{X}}^2 &= \frac{(74,0 - 79,75)^2 + (76,0 - 79,75)^2 + \dots + (84,0 - 79,75)^2 + (87,0 - 79,75)^2}{16} \\ &= 26,09 \text{ kg}^2. \end{aligned}$$

$$\sigma_{\bar{X}} = \sqrt{26,09} = 5,11 \text{ kg}.$$

Observamos que o valor da média da distribuição amostral das médias é o mesmo da média da população ( $\mu_{\bar{X}} = \mu$ ). A variância da distribuição amostral das médias é dada por:

$$\sigma_{\bar{X}}^2 = 26,09 = \frac{52,18}{2} = \frac{\sigma^2}{n}.$$

O desvio padrão da distribuição da média é dado por:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}},$$

No exemplo  $\sigma_{\bar{X}} = 7,22/\sqrt{2} = 5,11$  kg.

No caso de população finita de tamanho  $N$  e amostragem aleatória simples sem reposição, a média da distribuição amostral da média é igual a média da população, porém a variância e o desvio padrão ficam dados por:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \frac{N-n}{N-1}.$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Na Tabela 6.4 vamos calcular a variância e o desvio padrão das 12 médias de peso de suínos obtidas com amostras de tamanho  $n = 2$ . Estas 12 médias são obtidas de amostras sem reposição de uma população finita.

$$\begin{aligned}\sigma_{\bar{X}}^2 &= \frac{(74,0 - 79,75)^2 + (76,0 - 79,75)^2 + \dots + (83,5 - 79,75)^2 + (85,5 - 79,75)^2}{12} \\ &= 17,40 \text{ kg}^2. \\ \sigma_{\bar{X}} &= \sqrt{17,40} = 4,17 \text{ kg}.\end{aligned}$$

O termo  $\sqrt{(N-n)/(N-1)}$  é denominado fator de correção para população finita. Quando ocorrer a desigualdade  $n < 0,05N$  este termo pode ser omitido da expressão da variância.

Mostramos como fica a média e a variância da distribuição amostral da média. O próximo passo é verificarmos a forma da distribuição amostral das médias, isto é, qual modelo probabilístico poderia ser utilizado para descrever a distribuição amostral das médias? Nas Figuras 6.5 e 6.6 temos os histogramas das médias amostrais possíveis para amostras de tamanhos 2 e 3, respectivamente, para os dados da Tabela 6.3. Observamos que a medida que o tamanho da amostra aumenta de 2 para 3, as médias amostrais ficam mais concentradas em torno da média verdadeira ( $\mu$ ), porque o desvio padrão diminui. Os casos extremos, de médias muito grandes ou pequenas, passam a ter uma menor frequência de ocorrência. É sabido que se a população sobre a qual desejamos selecionar uma amostra aleatória simples ( $X_1, \dots, X_n$ ), tem

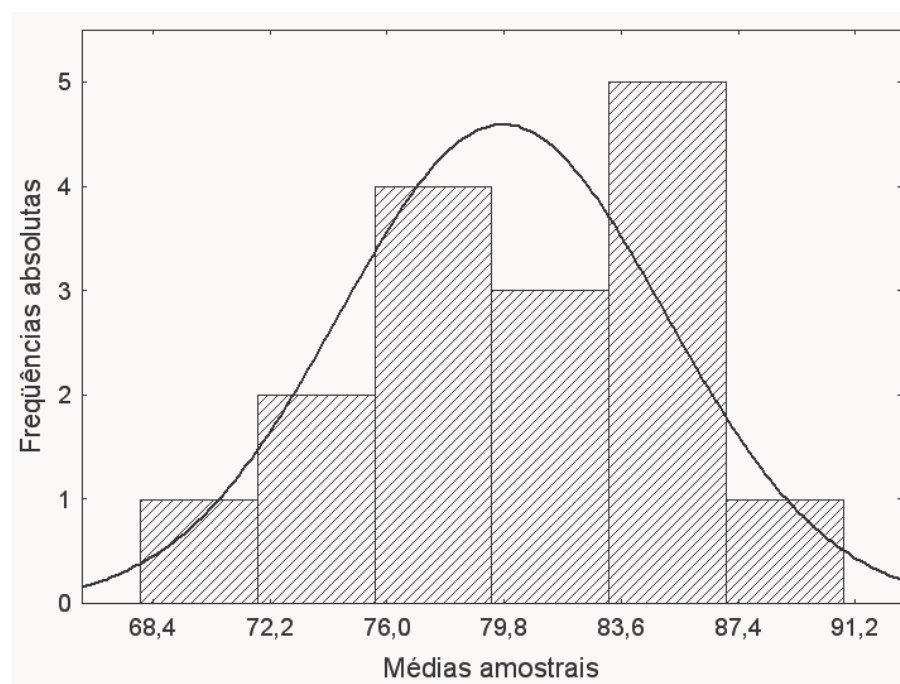


Figura 6.5: Histograma da distribuição amostral das médias para amostras de tamanho 2

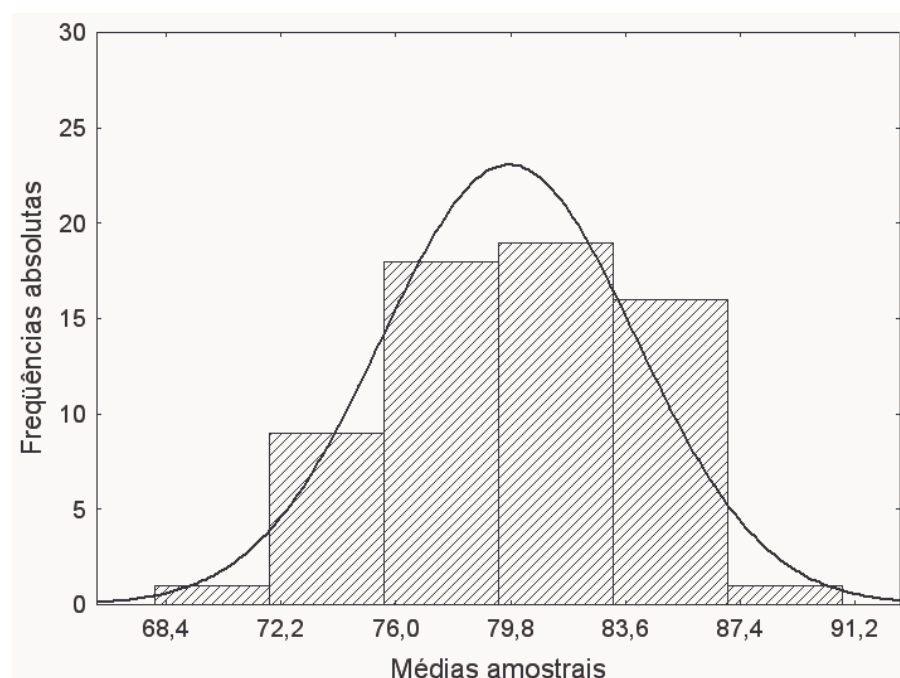


Figura 6.6: Histograma da distribuição amostral das médias para amostras de tamanho 3



distribuição normal com média  $\mu$  e variância  $\sigma^2$ , a distribuição amostral das médias também será normal, para qualquer tamanho de amostra, com média  $\mu$  e variância  $\sigma^2/n$ ;

Passamos agora a analisar o comportamento dos histogramas das médias para várias distribuições de  $X$ , quando vamos aumentando o tamanho da amostra. Na Figura 6.7 têm-se os histogramas correspondentes à distribuição amostral das médias relativas a 3 diferentes distribuições: Poisson, Uniforme e Bimodal.

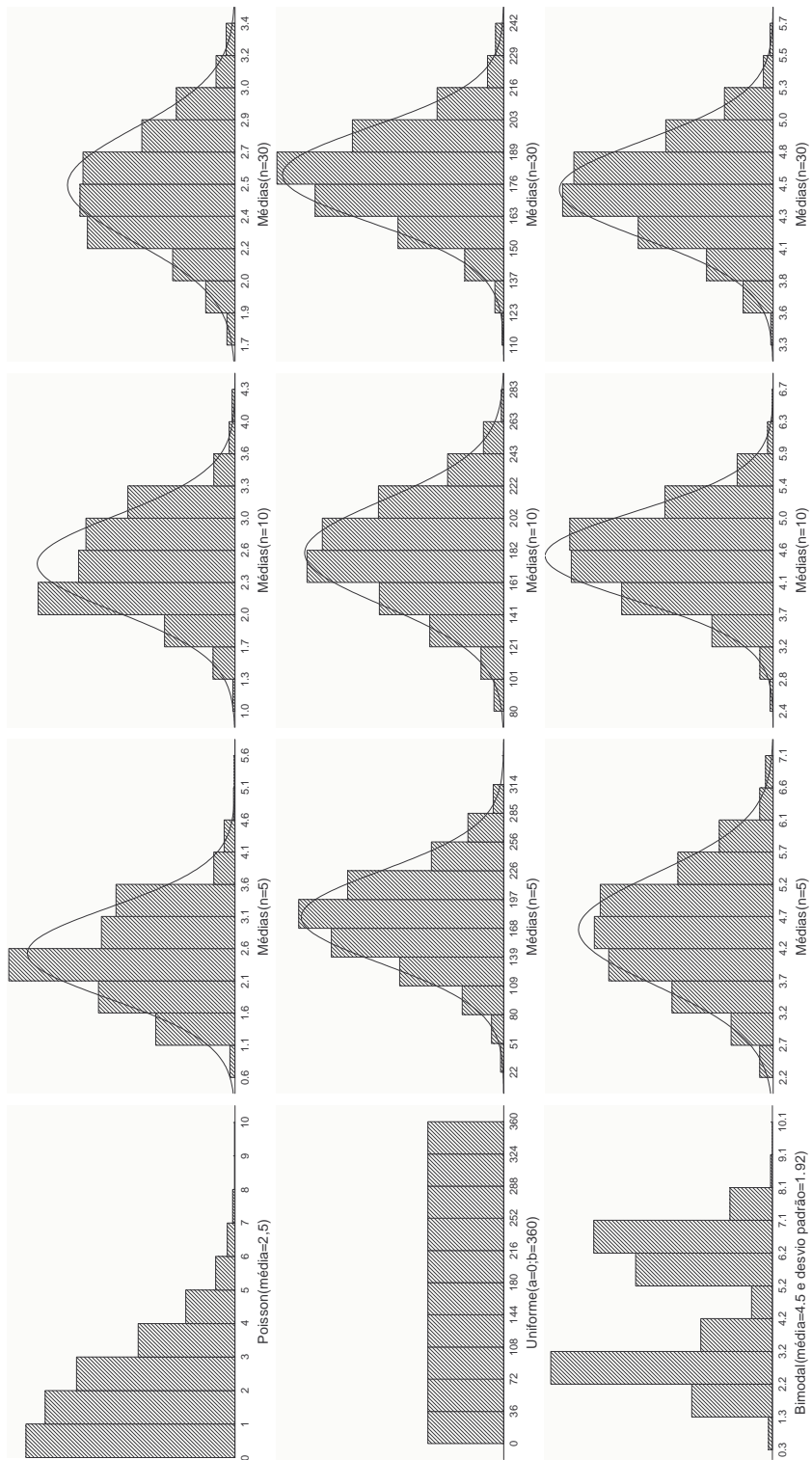


Figura 6.7: Histogramas da distribuição amostral das médias para três populações(Poisson, Uniforme e Bimodal) e para três tamanhos de amostras( $n=5$ ,  $n=10$  e  $n=30$ )

Os exemplos da Figura 6.7, nos indicam que quando o tamanho da amostra vai aumentando, independente da distribuição da variável aleatória  $X$ , a distribuição amostral das médias aproxima-se cada vez mais de uma distribuição normal, com média igual a média populacional (2,5 para a distribuição de Poisson, 180 para a Uniforme e 4,5 para a Bimodal) e variância decrescente. A rapidez dessa convergência depende da distribuição da população original da qual a amostra é retirada. Se a população original é uniforme, sua convergência é rápida. Já, se a distribuição da população tem a forma bimodal ou assimétrica, essa convergência é mais demorada. Este resultado, fundamental na teoria de inferência estatística, é conhecido como *teorema central do limite*. Esse teorema diz o seguinte:

- se a população sobre a qual desejamos selecionar uma amostra aleatória simples  $(X_1, \dots, X_n)$ , tem uma distribuição qualquer com média  $\mu$  e variância  $\sigma^2$ , a distribuição amostral das médias será aproximadamente normal com média  $\mu$  e variância  $\sigma^2/n$ , para amostras suficientemente grandes. Como sugestão, aceita-se que para amostras com mais de 30 observações a aproximação já pode ser considerada muito boa.

Em muitas das situações práticas com  $n$  grande e  $n/N$  pequeno, o teorema central do limite também se aplica quando a amostragem é sem reposição.

**Exemplo 1.** Suponha que o peso dos 35.000 alunos da UFSC, matriculados no primeiro semestre de 2005, é normalmente distribuído com média de 60,5 kg e desvio padrão de 12 kg.

- Qual será a média e o desvio padrão da distribuição amostral da média, na hipótese de utilizar amostras de tamanho  $n = 36$  sem reposição?
- Qual é a probabilidade da média de uma amostra de tamanho  $n = 36$  se encontrar entre 58 e 62 kg?
- e abaixo de 55 kg? e
- e acima de 65 kg?

Resolução:

- A média e o desvio padrão da distribuição amostral valem:

$$\mu_{\bar{X}} = 60,5 \text{ kg} \quad \sigma_{\bar{X}} = \frac{12}{\sqrt{36}} = 2 \text{ kg}.$$

pois  $(35.000 - 36)/(35.000 - 1) \cong 1$ , não necessitando da correção para população finita.

- A probabilidade vale:

$$z_1 = \frac{58 - 60,5}{2} = -1,25 \quad z_2 = \frac{62 - 60,5}{2} = 0,75.$$

$$\begin{aligned}
P(58 \leq \bar{X} \leq 62) &= P(-1,25 \leq z \leq 0,75) \\
&= 0,39435 + 0,27337 \\
&= 0,6677.
\end{aligned}$$

c) A probabilidade vale:

$$\begin{aligned}
z &= \frac{55 - 60,5}{2} = -2,75 \\
P(\bar{X} < 55) &= P(z < -2,75) \\
&= 0,0030.
\end{aligned}$$

d) A probabilidade vale:

$$\begin{aligned}
z &= \frac{65 - 60,5}{2} = 2,25 \\
P(\bar{X} > 65) &= P(z > 2,25) \\
&= 0,0122.
\end{aligned}$$

**Exemplo 2.** Uma nova ração para aumento de peso deve ser aplicada a uma amostra aleatória de 25 frangos ( $n = 25$ ), de um grande lote (o lote é a população). Supondo que o desvio padrão ( $\sigma$ ) do ganho de peso no período de um mês seja de 2 kg, qual a probabilidade de que a média dessa amostra ( $\bar{X}$ ) difira da média de todo o lote ( $\mu$ ) em 0,5 kg ou mais ( $|\bar{X} - \mu| \geq 0,5$ )? considerando que os frangos tratados com a nova ração não tem um ganho de peso maior do que os tratados com a ração padrão.

Inicialmente, passamos a calcular o valor da variável  $z$ .

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{X}}} = \frac{0,5}{2/\sqrt{25}} = \frac{0,5}{0,4} = 1,25.$$

Portanto, a probabilidade pedida vale (lembre do módulo):

$$\begin{aligned}
P(z \leq -1,25) + P(z \geq 1,25) &= 2P(z \geq 1,25) \\
&= 0,2112.
\end{aligned} \tag{6.2}$$

A probabilidade de  $\bar{X}$  diferir de  $\mu$  em mais do que 0,5 kg, devido ao acaso, é de 21,12%.

### Exercícios resolvidos

1. O governo coleta uma amostra aleatória de 400 estudantes da UFSC, para obter uma indicação dos que são a favor do ensino superior pago. Determine a probabilidade de

obter uma proporção amostral que difira por mais de 3 pontos percentuais da proporção populacional de estudantes que são a favor do ensino pago, se esta proporção é 10%.

$$\begin{aligned} Prob(0,07 \leq P \leq 0,13) &= Prob\left(\frac{0,07 - 0,10}{\sqrt{\frac{0,10(0,90)}{400}}} \leq z \leq \frac{0,13 - 0,10}{\sqrt{\frac{0,10(0,90)}{400}}}\right) \\ &= Prob(-2 \leq z \leq 2) = 0,9545. \end{aligned}$$

Portanto, a resposta é  $1 - 0,9545 = 0,0455$  ou 4,55%.

2. Os camarões machos da espécie *Penaeus paulensis* para serem considerados adultos devem apresentar um comprimento total maior ou igual a 22 mm. Se numa população de camarões machos, a média dos comprimentos foi igual a  $\mu = 27,3mm$  e desvio padrão  $\sigma = 7,8mm$ , portanto, considerada como sendo uma população adulta. Pergunta-se:

a) Qual a probabilidade de, para uma amostra de  $n = 35$  camarões, obtermos uma média  $\bar{X} < 22mm$ ?

b) Qual deve ser o valor para a média do comprimento total,  $\mu$ , a fim de que a  $P(\bar{X} \leq 22) = 0,05$  ou 5%?

a)

$$\begin{aligned} P(\bar{X} < 22) &= P\left(z < \frac{22 - 27,3}{7,8/\sqrt{35}}\right) = P(z < -4,02) \\ &= 0,5 - 0,49997 = 0,00003 \text{ ou } 0,003\%. \end{aligned}$$

b) Para uma área de 5%, no lado esquerdo da curva, o valor de  $z$  é -1,645.

$$-1,645 = \frac{22 - \mu}{1,3184} \Rightarrow \mu = 24,1688mm.$$

## 6.5 Exercícios Propostos

1. Se vamos extrair amostras de  $n=100$  observações de uma população muito grande, em que a proporção populacional é 20%, que percentagem de proporções amostrais poderemos esperar nos intervalos?

a) 16% a 24%

b) maior que 24%.

2. A proporção de estômatos da epiderme abaxial da folha de macieira da variedade M-9 com tamanho acima de um determinado valor é  $\pi = 0,12$  ou 12%. Extraída uma amostra de 35 folhas, qual a probabilidade de que a proporção esteja entre 8% e 13%?

3. Sabe-se que 46% de peixes *Xenomelaniris brasiliensis*, na localidade da praia da Barra da Lagoa, Florianópolis, apresentam comprimento total acima de 50mm. Qual é a probabilidade de uma amostra aleatória com 35 peixes apresentar mais que 53% ou menos que 40% de peixes com comprimento total acima de 50 mm?
4. Um distribuidor de sementes determina, através de testes, que 5% das sementes não germinam. Ele vende pacotes de 200 sementes com garantia de 90% de germinação. Qual a probabilidade de um pacote não satisfazer a garantia?
5. Os níveis de glutathione na região do cerebelo em ratos apresentam média de  $1,7\mu M$  e desvio padrão de  $0,4\mu M$ . Os níveis de glutathione de uma amostra de 8 ratos foram mensurados. A média da amostra é usada para decidir se os ratos estão estressados. Aplica-se a seguinte regra de decisão: se o nível de glutathione da amostra de 8 ratos é igual ou maior do que  $2,0\mu M$  os ratos são considerados estressados. Qual a probabilidade de se considerar os ratos como estressados?
6. Suponha que o rendimento de milho, em kg/ha, no Oeste Catarinense, é uma variável normalmente distribuída, com média de 7069 kg/ha e desvio padrão de 593,15 kg. Que valores espera-se encontrar para a média e o desvio padrão da distribuição amostral das médias, na hipótese de utilizar amostras de tamanho  $n=36$ ? Qual é a probabilidade da média de uma amostra se encontrar: a) entre 6809 e 7395 kg/ha; b) abaixo de 6816 kg/ha; c) acima de 7389 kg/ha.
7. O número de hemócitos/ $\mu l$  de hemolinfa (THC) em fêmeas de camarão *Macrobrachium roseberguii* é normalmente distribuída com média ( $\mu$ ) de 18.000 e desvio padrão ( $\sigma$ ) de 5125. Em uma amostra aleatória de  $n = 30$  fêmeas, qual a probabilidade de que o número médio de hemócitos/ $\mu l$  de hemolinfa (THC) seja maior do que 19.000?
8. O número de estômatos da epiderme abaxial em áreas aleatórias de  $0,24\text{ mm}^2$  de macieira da variedade M9 é normalmente distribuída com média  $\mu = 159,8$  e desvio padrão de  $\sigma = 34,5582$ . Em uma amostra aleatória de  $n = 35$  folhas, qual a probabilidade de que o número médio de estômatos seja menor do que 140?
9. Uma variável aleatória  $X$  tem distribuição normal, com média 100 e desvio padrão 10. a) Qual a  $P(90 < X < 110)$ ? b) Se  $\bar{X}$  é a média de uma amostra de 16 elementos retirados dessa população, calcule  $P(90 < \bar{X} < 110)$ .
10. A proporção esperada de filhotes machos de tartaruga *Caretta caretta* conhecida como cabeçuda é de 0,50. Uma fêmea desta espécie bota 200 ovos numa determinada praia. a) Qual a probabilidade da proporção amostral de machos, nesta ninhada, estar entre 0,48 e 0,53? b) Se a proporção de nascimentos de machos for maior do que 0,60 pode indicar algum

desequilíbrio ambiental, como, por exemplo, sombreamento da praia. Qual a probabilidade disso ocorrer?

## 7 Estimação dos Parâmetros

### 7.1 Introdução

Estimação é o nome técnico para o processo que consiste em utilizar os dados de uma amostra para avaliar parâmetros populacionais desconhecidos, ou, como o próprio nome indica, estimar os mesmos. Dentre as diversas características (parâmetros) de uma população que podem ser estimadas, nós vamos estudar as mais utilizadas, quais sejam: a *média*( $\mu$ ), a *proporção*( $\pi$ ) e a *variância*( $\sigma^2$ ).

Um pesquisador sempre está desenvolvendo um processo de estimação. Por exemplo, um Biólogo pode estar interessado na proporção de micronúcleos em 5000 células sanguíneas em peixes do gênero bagre; um Agrônomo pode estar interessado na produção média de uma cultura. Outros exemplos, os prejuízos causados pelo ataque de uma praga ou doença; o diâmetro de caramujos; o tamanho de Lulas encontradas no trato digestivo de Atuns; tamanho de crustáceos da classe Malacostraca e sub-classe Eumalacostraca, popularmente conhecida com o nome de Caprelas; parâmetros estatísticos genéticos (variância genética, ambiental, fenotípica).

Os resultados estatísticos apresentados neste capítulo têm um papel fundamental para os pesquisadores, no sentido de responder aos objetivos do trabalho.

### 7.2 Propriedades dos Estimadores

Um *estimador* é uma estatística que será usado para a estimação de um parâmetro populacional. Os estimadores mais freqüentes são a média, a proporção e a variância amostrais, representados por:  $\bar{X}$ ,  $P$  e  $S^2$ , respectivamente, utilizados para estimar os parâmetros  $\mu$ ,  $\pi$  e  $\sigma^2$ , respectivamente. É desejável que os estimadores apresentem certas propriedades, ou seja, certas qualidades. Dentre as propriedades desejáveis de um estimador podemos citar:

- *não-tendencioso ou não-viesado*, quer dizer, a média da sua distribuição amostral é igual ao parâmetro populacional que desejamos estimar. Em termos formais, temos que: para  $\theta$  e  $\hat{\theta}$ , o parâmetro e o estimador de uma característica qualquer, respectivamente, a não tendenciosidade implica que  $E(\hat{\theta}) = \theta$ , isto é, sua média deve ser igual ao parâmetro. Por exemplo,  $E(\bar{X}) = \mu$  e  $E(P) = \pi$ ;
- *precisão*, quer dizer, deseja-se que o estimador seja altamente concentrado, isto é, que



tenha pequena variância amostral (cada observação próxima da média de todas as observações).

- *acurácia*, quer dizer, que o estimador seja o mais próximo possível do parâmetro, isto é, que o erro amostral,  $e = \hat{\theta} - \theta$ , seja o menor possível. Cada observação o mais próximo do parâmetro.

O exemplo a seguir, tirado de Bussab & Morettin (1987), serve bem para ilustrar as propriedades de um estimador. No exemplo, supõe-se que desejamos comprar um rifle e, dentre muitos, foram selecionados 4 deles, denominados de A, B, C e D. Com o objetivo de testá-los, foram dados 15 tiros com cada um deles. A representação gráfica é dada na Figura 7.1. Com o objetivo de selecionar uma arma dentre as quatro, devemos adotar alguns critérios. Os rifles B e D são viesados, isto é, os tiros estão deslocados do alvo, além disso, o rifle B tem pouca precisão. O rifle A não é viesado, porém, apresenta baixa precisão, isto é, os tiros estão muito espalhados. Os rifles A, B e D apresentam baixa acurácia, isto é, os tiros não acertam o alvo. De acordo com esses critérios o rifle selecionado seria o C, pois ele é não-viesado, com boa precisão e acurácia.

Como já vimos, os estimadores  $\bar{X}$  e  $P$  são não-tendenciosos para os parâmetros  $\mu$  e  $\pi$ , respectivamente. Para o parâmetro  $\sigma^2$  um estimador não-tendencioso é a variância amostral dada por:

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

Podemos notar que a divisão é feita por  $n - 1$  e não por  $n$ , como seria esperado. A divisão por  $n$  forneceria um estimador tendencioso, como podemos verificar com os dados das duas últimas colunas da Tabela 6.4.

$$E(S^2) = \frac{72,0 + 128,0 + \dots + 0,0 + 0,0}{16} = 52,18 = \sigma^2$$

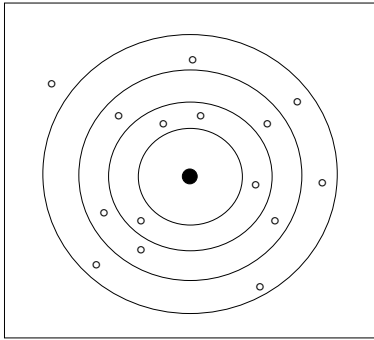
usando  $n - 1 = 2 - 1 = 1$  no denominador do cálculo das variâncias amostrais, e

$$E(S^2) = \frac{36,0 + 64,0 + \dots + 0,0 + 0,0}{16} = 26,09 \neq \sigma^2$$

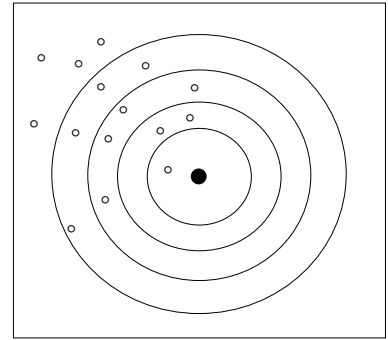
usando  $n = 2$  no denominador do cálculo das variâncias amostrais.

Nas situações onde temos dois estimadores  $T_1$  e  $T_2$  não-tendenciosos para o mesmo parâmetro  $\theta$ , e desejamos decidir por um deles, podemos utilizar o conceito de eficiência relativa definida por:

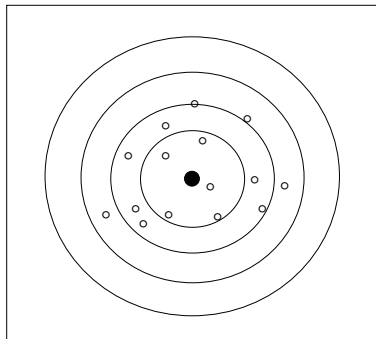
$$EF(T_1|T_2) = \frac{Var(T_1)}{Var(T_2)}.$$



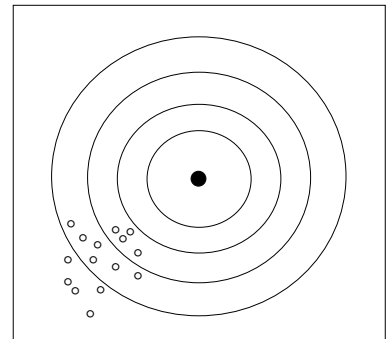
(a) Rifle A: não-viesado, pouca precisão, pouca acurácia



(b) Rifle B: viesado, pouca precisão e pouca acurácia



(c) Rifle C: não-viesado, boa precisão e acurácia



(d) Rifle D: viesado, boa precisão e baixa acurácia

Figura 7.1: Resultados de 15 tiros dos rifles A, B, C e D (ilustração das propriedades dos estimadores). Fonte: Bussab & Morettin (1987).

Dizemos que  $T_2$  é mais eficiente do que  $T_1$  se  $EF(T_1|T_2)$  for maior do que 1. Por exemplo, para estimar a média populacional, para uma variável aleatória com distribuição normal, a média  $\bar{X}$  e a mediana  $Md$  amostrais são dois estimadores não-tendenciosos e pode-se mostrar que

$$EF(Md|\bar{X}) = \frac{Var(Md)}{Var(\bar{X})} = \frac{(\pi/2)(\sigma^2/n)}{\sigma^2/n} = \frac{\pi}{2} = 1,57,$$

indicando que a média amostral é 57% mais eficiente do que a mediana amostral como estimador da média da população.

### 7.3 Métodos para Encontrar Estimadores

Se desejamos estimar a proporção populacional  $\pi$ , é intuitivo pensarmos em utilizar a proporção amostral como estimador. O mesmo valeria para a estimação da média populacional. Porém, existem situações que não se tem uma indicação tão clara de qual estimador é que deve ser usado. Por exemplo, um experimento foi instalado de acordo com o delineamento inteiramente casualizado. Neste delineamento as unidades experimentais são as mais homogêneas possíveis, portanto, espera-se que a variação seja devida ao fator em estudo. Como exemplos de fatores de interesse podemos citar: locais, variedades, espécies, dietas, etc. O modelo matemático desse delineamento é dado por:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

onde  $y$  é a característica que estamos interessados, por exemplo, comprimento do corpo de um animal;  $\mu$  é uma média geral;  $\tau_i$  é o efeito do  $i$ -ésimo nível do fator em estudo e  $\epsilon$  são variações aleatórias que ocorrem no experimento não devidas ao fator de estudo. O pesquisador tem interesse em estimar os parâmetros  $\mu$  e  $\tau$ . Para estimar  $\mu$ , uma sugestão é usar a média geral do experimento ( $\bar{Y}$ ). Para estimar  $\tau$ , o que devemos usar? Demonstra-se que um estimador do  $i$ -ésimo nível do fator em estudo é dado por:

$$\hat{\tau}_i = \bar{Y}_i - \bar{Y}.$$

Faz-se então necessário a existência de métodos apropriados para a obtenção de estimadores dos parâmetros que possam ser de interesse. Os três métodos mais utilizados são: 1) *método da máxima verossimilhança* e 2) *método dos momentos* e 3) *método dos mínimos quadrados*. Vamos apresentar o método da máxima verossimilhança.

### 7.3.1 Método da Máxima Verossimilhança

Suponha, por exemplo, que dentre  $n = 3$  sementes em teste, obtemos 2 sucessos (duas sementes germinaram) e 1 fracasso (uma semente não germinou). Vamos considerar que a probabilidade de uma semente germinar é  $\pi$ , onde  $0 \leq \pi \leq 1$ . Vamos definir a variável  $X$  como sendo o número de sementes que germinaram. O que devemos utilizar como estimador de  $\pi$ ? Este é um experimento binomial (verifique as pressuposições do modelo binomial), e a sua função de probabilidade é dada por:

$$P(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}.$$

Para  $n = 3$  e  $k = 2$ , temos:

$$P(X = 2) = \binom{3}{2} \pi^2 (1 - \pi)^1 = 3\pi^2(1 - \pi).$$

Vamos chamar esta expressão de *função de verossimilhança* representada por:

$$L(\pi) = 3\pi^2(1 - \pi). \quad (7.1)$$

A estimativa de máxima verossimilhança de  $\pi$  é o valor  $p$  que maximiza a função 7.1. Do Cálculo Diferencial sabemos que o valor desejado é aquele que anula a primeira derivada e faz com que a segunda derivada seja negativa. Apresentamos a seguir a primeira ( $L'(\pi)$ ) e a segunda derivada ( $L''(\pi)$ ) de  $L(\pi)$  em relação a  $\pi$ :

$$\begin{aligned} L'(\pi) &= 2\pi(1 - \pi) + \pi^2(-1) \\ &= 2\pi(1 - \pi) - \pi^2 \\ &= 2\pi - 2\pi^2 - \pi^2 \\ &= 2\pi - 3\pi^2 \\ &= \pi(2 - 3\pi) \end{aligned}$$

e

$$L''(\pi) = (2 - 3\pi) + \pi(-3) = 2 - 6\pi.$$

Os valores  $p = 0$  e  $p = 2/3$  são os dois únicos valores que anulam a primeira derivada. Como  $L''(0)=2$ , um valor maior que zero, e  $L''(2/3)=-2$ , um valor menor que zero, então a estimativa de máxima verossimilhança será  $p = 2/3$ .

De modo geral, o estimador de máxima verossimilhança é o valor do parâmetro de interesse que maximiza uma função, denominada função de verossimilhança, que é obtida a partir dos dados e da função de probabilidade associada aos mesmos.

## 7.4 Estimativas Pontuais e Intervalares

De modo geral, vamos supor que os valores da população se distribuem segundo um dado modelo probabilístico, cujos parâmetros são desconhecidos e, portanto, precisam ser estimados a partir dos dados de uma amostra aleatória simples extraída da população em estudo. Os estimadores são estatísticas e o conhecimento de suas correspondentes distribuições amostrais é fundamental para o desenvolvimento dos resultados que vamos apresentar nesta seção.

### 7.4.1 Estimativas Pontuais

Na estimação por ponto, utiliza-se de um único valor para a estimação do parâmetro. Este valor é determinado por um dos métodos citados na Seção 7.3. Como exemplos de estimadores pontuais temos a média aritmética da amostra  $\bar{X}$ , para a média populacional  $\mu$ , a proporção na amostra  $P$ , para a proporção populacional  $\pi$ , a variância da amostra  $S^2$ , para a variância populacional  $\sigma^2$ , e o desvio padrão da amostra  $S$ , para o desvio padrão populacional  $\sigma$ .

**Exemplo 1.** Uma amostra aleatória simples de 10 isolados de fungos ectomicorrízicos *P. tinctorius* cultivados *in vitro*, apresentaram os seguintes resultados de biomassa, em gramas.

0,034	0,033	0,029	0,034	0,027
0,034	0,029	0,035	0,032	0,028

A estimativa pontual da média populacional é dada por:

$$\bar{x} = \frac{0,034 + 0,033 + \dots + 0,032 + 0,028}{10} = \frac{0,315}{10} = 0,0315 \text{ gramas.}$$

A estimativa do desvio padrão da população, calculado na amostra, é dada por:

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}} = \sqrt{\frac{(0,034 - 0,0315)^2 + \dots + (0,028 - 0,0315)^2}{10 - 1}} = 0,00295 \text{ gramas.}$$

**Exemplo 2.** A espécie *Harengula clupeola* (sardinha cascuda) apresenta tamanho médio de maturação sexual com 150 milímetros de comprimento. Para se estimar a proporção de sardinhas que apresetam maturação sexual, foi selecionada uma amostra aleatória simples de 20 indivíduos e anotados os seus comprimentos, em *mm*, cujos resultados foram:

78	80	80	110	158	123	110	112	96	86
83	79	71	98	108	125	105	71	92	76

A estimativa pontual da proporção populacional é dada por:

$$p = \frac{2}{20} = 0,10.$$

Um problema de se utilizar estimadores pontuais é que eles não fornecem uma medida de precisão, isto é, o quanto poderíamos estar errando ao utilizarmos este único valor para estimar o parâmetro. Portanto o processo de estimação deve também levar em conta o possível erro amostral do estimador, que depende da sua distribuição amostral, como será visto a seguir.

#### 7.4.2 Erro Amostral

A média dos valores de uma amostra ( $\bar{X}$ ) dificilmente vai ser igual a média verdadeira ( $\mu$ ), apesar de esperarmos obter valores próximos. A diferença entre a estatística amostral e o parâmetro é denominada erro amostral da média, representado por,

$$e = \bar{X} - \mu.$$

Apesar do valor do erro amostral não poder ser calculado, pois o seu cálculo depende do valor do parâmetro que estamos estimando, sua magnitude pode ser quantificada a partir da distribuição amostral da média ( $\bar{X}$ ). Da Seção 6.4.2 sabemos que essa distribuição pode ser aproximada pela distribuição normal com média  $\mu$ , o parâmetro a ser estimado, e variância  $\sigma^2/n$ , conforme representada na Figura 7.2. Da Figura 7.2 podemos dizer que 68% dos possíveis valores da média de uma amostra aleatória simples de tamanho  $n$  não se afastam do verdadeiro valor da média em mais do que  $1\sigma/\sqrt{n}$ , e que 95% dos possíveis valores não se afastam em mais do que  $1,96\sigma/\sqrt{n}$ . Em outras palavras, podemos dizer que não esperamos um erro amostral maior do que  $1\sigma/\sqrt{n}$ , com 68% de confiança, e um erro amostral maior do que  $1,96\sigma/\sqrt{n}$ , com 95% de confiança. Estes resultados são apropriados para situações de populações infinitas.

**Exemplo.** Os pesos de pacotes de sementes de milho, enchidos automaticamente por uma máquina, apresentam desvio padrão ( $\sigma$ ) igual a 200 gramas. Se utilizarmos a média aritmética ( $\bar{X}$ ) de uma amostra aleatória simples de 25 pacotes para estimar o peso médio ( $\mu$ ) dos pacotes, esperamos cometer um erro de no máximo  $(1,96 \times 200)/\sqrt{25} = 78,4$  gramas, com 95% de confiança, para mais ou para menos.

O mesmo raciocínio pode ser aplicado no cálculo do erro amostral máximo esperado na estimação da proporção populacional ( $\pi$ ), pela proporção amostral ( $P$ ). Neste caso, tem-se  $\pi$  ao invés de  $\mu$  e  $\pi(1 - \pi)/n$  ao invés de  $\sigma^2/n$ , como parâmetros da distribuição amostral.

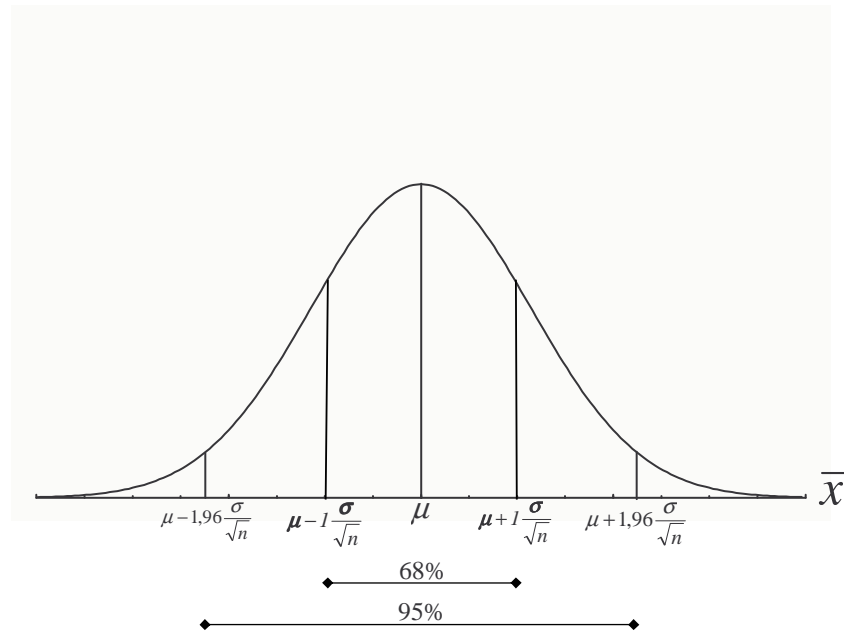


Figura 7.2: Intervalo para uma média com 68% e 95% de confiança

**Exemplo.** A proporção de germinação de determinada espécie é da ordem de 0,80. Se utilizarmos a proporção ( $P$ ) de uma amostra aleatória simples de 400 sementes para estimar a proporção ( $\pi$ ) de germinação, esperamos cometer um erro de no máximo  $(1,96 \times \sqrt{0,80(1 - 0,80)})/\sqrt{400} = 0,039$ , com 95% de confiança, para mais ou para menos. Neste caso, é mais comum dizer que o erro é de no máximo 3,9 pontos percentuais para mais ou para menos.

De uma forma geral pode-se escrever a expressão do valor do erro amostral máximo como:

$$e_{max} = Z_{\gamma/2} \frac{\sigma}{\sqrt{n}} \text{ na estimação de } \mu \quad (7.2)$$

e

$$e_{max} = Z_{\gamma/2} \frac{\sqrt{\pi(1 - \pi)}}{\sqrt{n}} \text{ na estimação de } \pi, \quad (7.3)$$

onde  $\gamma$  representa o grau de confiança associado ao valor do erro amostral. Na prática é comum estabelecermos  $\gamma$  igual a 0,95, ou seja 95% de confiança, fazendo  $Z_{\gamma/2} = Z_{0,475} = 1,96$ .

### 7.4.3 Estimativas por Intervalos de Confiança

Diferentes pesquisadores selecionando amostras de uma mesma população, deverão obter estimativas pontuais diferentes para o mesmo parâmetro populacional. Isto está relacionado

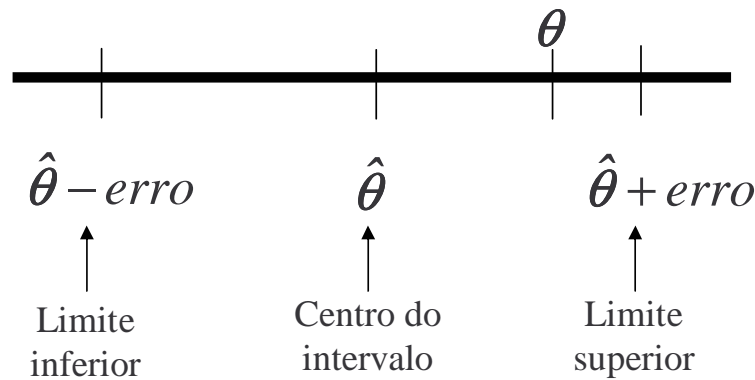


Figura 7.3: Erro de estimação associado a um intervalo de confiança

com o que denominamos de variabilidade amostral do estimador pontual. Uma forma mais apropriada seria construir um estimador que também levasse em conta essa variabilidade. Este estimador é o estimador por intervalo, que combina o estimador pontual com o erro amostral máximo esperado. Esse estimador por intervalo nos dá um intervalo dentro do qual esperamos, com um alto grau de confiança, que o parâmetro se encontre.

Os limites inferior (LI) e superior (LS) de um intervalo de confiança para um parâmetro  $\theta$ , cuja distribuição amostral do seu estimador pontual  $\hat{\theta}$  pode ser aproximada pela distribuição normal, são dados por:

$$LI = \hat{\theta} - \text{erro amostral máximo}$$

e

$$LS = \hat{\theta} + \text{erro amostral máximo},$$

onde o erro amostral máximo é calculado a partir da distribuição amostral do estimador pontual, conforme discutido na Seção 7.4.2. Na Figura 7.3 temos uma ilustração de um intervalo de confiança.

#### 7.4.3.1 Intervalo de Confiança para a Média da População

Vamos dividir a construção do intervalo de confiança para a média da população,  $(\mu)$ , usando o estimador pontual  $\bar{X}$ , em dois casos:

- quando o desvio padrão da população ( $\sigma$ ) é conhecido;
- quando o desvio padrão da população ( $\sigma$ ) não é conhecido

#### *Intervalo de Confiança para a Média Populacional Quando o Desvio Padrão da População é Conhecido*



Quando o desvio padrão populacional é conhecido, o erro amostral máximo é dado como em 7.2, e o intervalo de confiança para a média da população, com grau de confiança de  $100\gamma\%$ , é dado por:

$$\boxed{\text{I.C.}(\mu; \gamma): \bar{X} \pm z_{(\gamma/2)} \frac{\sigma}{\sqrt{n}},} \quad (7.4)$$

**Exemplo 1.** A distribuição dos pesos de pacotes de sementes de milho, enchidos automaticamente por uma certa máquina, é normal, com desvio padrão ( $\sigma$ ) conhecido e igual a 0,20 kg. Uma amostra de 15 pacotes retirada ao acaso apresentou os seguintes pesos, em kg,

20,05	20,10	20,25	19,78	19,69	19,90	20,20	19,89
19,70	20,30	19,93	20,25	20,18	20,01	20,09	

Construir os intervalos de confiança de 95% e 99% para o peso médio dos pacotes de sementes de milho.

Temos:

$$n = 15 \quad \bar{x} = 20,02 \text{ kg} \quad \frac{\sigma}{\sqrt{n}} = \frac{0,20}{\sqrt{15}} = 0,0516 \text{ kg}.$$

Assim, o intervalo de confiança de 95% ( $\gamma = 0,95$  e  $Z_{(\gamma/2)} = 1,96$ ), fica:

$$\begin{aligned} \text{I.C.}(\mu; 95\%) &: 20,02 \pm 1,96(0,0516) \\ &: 20,02 \pm 0,1012 \\ &: [19,92 \text{ kg}; 20,12 \text{ kg}]. \end{aligned}$$

Portanto, com 95% de confiança, podemos dizer que a média da população é um valor do intervalo acima.

Para o grau de confiança de 99% ( $\gamma = 0,99$  e  $Z_{(\gamma/2)} = 2,58$ ), o intervalo fica:

$$\begin{aligned} \text{I.C.}(\mu; 99\%) &: 20,02 \pm 2,58(0,0516) \\ &: 20,02 \pm 0,1332 \\ &: [19,89 \text{ kg}; 20,15 \text{ kg}]. \end{aligned}$$

Portanto, com 99% de confiança, podemos dizer que a média da população é um valor do intervalo acima. Notamos que para um mesmo tamanho de amostra, o intervalo de 99% tem uma amplitude maior do que o intervalo de 95%.

### *Intervalo de Confiança para a Média Populacional Quando o Desvio Padrão da População é Desconhecido*

Na maioria das situações práticas, o desvio padrão da população,  $\sigma$ , não é conhecido, usa-se, nesse caso, o desvio padrão da amostra,  $s$ , como estimativa do desvio padrão populacional, substituindo-se nas fórmulas,  $\sigma$  por  $s$ . Esta substituição pura e simples, no intervalo de confiança dado em 7.4, causa uma alteração na distribuição de probabilidade a ser considerada e como consequência no nível de confiança, isto é, o pesquisador pensa que está trabalhando com um nível de confiança, por exemplo, de 95%, quando, na realidade, o nível de confiança é menor, portanto, deve-se utilizar um valor maior do que o valor de  $z$ . Como é feito esse ajuste? Sabemos que a distribuição amostral da média pode ser aproximada pela distribuição normal com média  $\mu$  e desvio padrão  $\sigma/\sqrt{n}$ , logo, quando conhecemos  $\sigma$ , a distribuição da estatística

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

pode ser aproximada por terá distribuição normal padrão com média 0 (zero) e variância 1 (um) e representa-se por:  $Z \sim N(0, 1)$ , conforme discutido na Seção 6.4.2. E, assim, usamos esse modelo para resolvermos o nosso problema de estimação. Porém, quando usamos o desvio padrão da amostra,  $S$ , no lugar do desvio padrão da população,  $\sigma$ , obtemos uma nova estatística, denominada de estatística  $t$ , dada por:

$$t_{(n-1)} = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Essa estatística  $t$  tem distribuição  $t$  de Student<sup>12</sup> com parâmetro  $(n - 1)$ . Este parâmetro é conhecido como *graus de liberdade* ( $gl$ )<sup>13</sup> Assim como a distribuição normal, ela é simétrica, com média zero, porém, apresenta maior quantidade de dados nos extremos da distribuição. Na Figura 7.4 podemos verificar a forma da distribuição das estatísticas  $t$  e  $Z$ . A principal diferença entre as duas distribuições é que a distribuição  $t$  tem maior área nas extremidades, isto é, ela é mais densa. Isso significa que, para um dado nível de confiança, o valor de  $t$  será maior que o correspondente valor de  $Z$ . Pode-se perceber que a medida que  $n$  aumenta os valores de  $t$  se aproximam dos valores de  $z$ .

Assim, o intervalo de confiança para a média, quando usamos o desvio padrão da amostra

---

<sup>12</sup>O criador da distribuição  $t$  de Student foi W. S. Gossett, o qual adotou o pseudônimo de Student (século XX). Posteriormente essa distribuição foi estudada por Ronald A. Fisher.

<sup>13</sup>O parâmetro (graus de liberdade) da distribuição  $t$  depende diretamente do tamanho da amostra ( $n$ ). Como o cálculo do desvio padrão da amostra ( $S$ ), envolve o uso da média da amostra ( $\bar{X}$ ), que também é calculada a partir dos mesmos dados, o valor do parâmetro é  $(n - 1)$ .

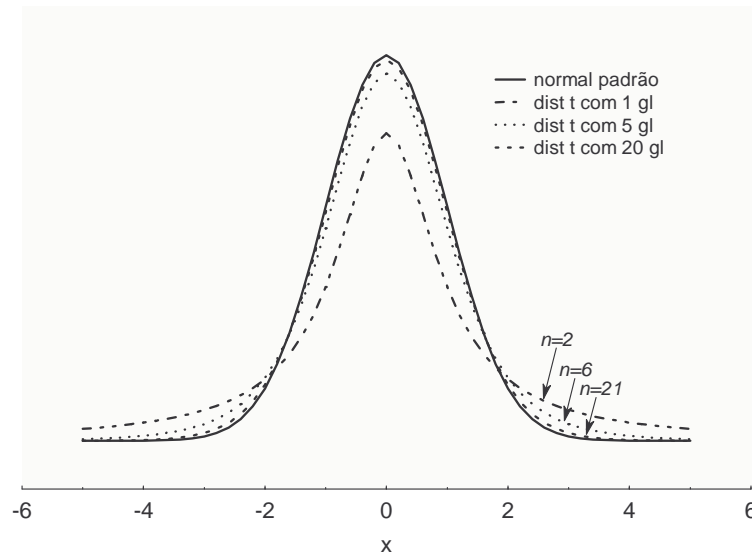


Figura 7.4: Forma das distribuições  $t$  de Student e normal padrão

é dado pela expressão:

$$\boxed{\text{I.C.}(\mu; \gamma): \bar{X} \pm t_{n-1; \gamma} \frac{S}{\sqrt{n}}} \quad (7.5)$$

onde  $\bar{X}$  e  $S$  são a média e o desvio padrão da amostra, respectivamente;  $n$  é o tamanho da amostra;  $\gamma$  é o nível de confiança e  $t$  é um valor de tabela da distribuição  $t$  de Student, obtido com  $n - 1$  graus de liberdade e nível de confiança  $\gamma$ .

A distribuição  $t$  só é teoricamente adequada quando a variável em estudo apresentar distribuição normal.

### *Uso da Tabela da Distribuição $t$*

No apêndice 4, temos a tabela da distribuição  $t$  de Student. A distribuição  $t$  de Student é ligeiramente diferente para cada amostra, ou seja, ela é dependente do tamanho da amostra. Para obter um valor de  $t$  precisamos de duas informações: 1) o valor de  $p$  e 2) o número de graus de liberdade, isto é, o tamanho da amostra menos um,  $(n - 1)$ . Por exemplo, para  $\gamma = 0,95$  e  $n = 28$ ,  $p = 1 - 0,95 = 0,05$ ,  $n - 1 = 27$  graus de liberdade e  $t = 2,052$ . Encontre este valor na tabela.

Nota-se que esta tabela está disposta de forma diferente da tabela da distribuição normal padrão. As áreas (probabilidades) estão na primeira linha da tabela e não no seu corpo, os valores de  $t$  estão no corpo da tabela e os graus de liberdade estão relacionados na primeira coluna (margem esquerda).

**Exemplo 1.** O peso médio, ao nascer, de bezerros da raça Ibagé, examinada uma amostra de 20 partos, foi de 26 kg com um desvio padrão de 2 kg. Dê a estimativa por intervalo do verdadeiro peso médio utilizando um nível de confiança de 95% ( $\gamma = 0,95$ ).

Da tabela da distribuição  $t$  obtemos  $t_{19;0,95} = 2,093$ . O intervalo de confiança fica:

$$\begin{aligned} I.C.(\mu; 95\%) &: 26 \pm 2,093 \frac{2}{\sqrt{20}} \\ &: 26 \pm 0,936 \\ &[25,064 \text{ kg}; 26,936 \text{ kg}]. \end{aligned}$$

Portanto, com 95% de confiança, podemos afirmar que a média da população é um valor do intervalo [25,064 kg; 26,936 kg].

**Exemplo 2.** Os resíduos industriais jogados nos rios, muitas vezes, absorvem o oxigênio necessário à respiração dos peixes e outras formas de vida aquática. Uma lei estadual exige um valor médio não inferior a 5 ppm de oxigênio dissolvido, a fim de que o conteúdo do mesmo seja suficiente para manter a vida aquática. Seis amostras de água retiradas de um rio revelaram os índices: 4,9, 5,1, 4,9, 5,0, 5,0 e 4,7 ppm de oxigênio dissolvido. Construir o intervalo com 95% de confiança para a verdadeira média de oxigênio, em ppm, e interpretar. A média e o desvio padrão da amostra valem:  $\bar{x} = 4,933$  e  $s = 0,137$ . O valor tabelado de  $t$ , com 5 graus de liberdade e  $\gamma = 0,95$  é 2,571. O intervalo de confiança fica:

$$\begin{aligned} IC(\mu; 95\%) &: 4,933 \pm 2,571 \frac{0,137}{\sqrt{6}} \\ &: 4,9333 \pm 0,144 \\ &[4,79 \text{ ppm}; 5,08 \text{ ppm}]. \end{aligned}$$

Podemos afirmar com 95% de confiança que a verdadeira média de oxigênio dissolvido é um valor entre 4,79 e 5,08 ppm. Portanto, não podemos afirmar que o conteúdo de oxigênio dissolvido é suficiente para manter a vida aquática.

### *Interpretação do Intervalo de Confiança para a Média*

De uma população com a característica  $X$  de interesse do estudo, se retirarmos todas as amostras possíveis de tamanho  $n$  e, para cada uma das amostras, calcularmos o intervalo de confiança da forma:  $\bar{X} \pm 1,96\sigma/\sqrt{n}$ , então, 95% deles deverão conter o parâmetro populacional,  $\mu$ . Na Tabela 7.1 estão representados todos os intervalos de confiança para uma amostra aleatória de tamanho  $n = 2$  da população dada na Tabela 6.3. Das colunas 6 e 7, podemos ver que 15 dos 16 possíveis intervalos contêm o verdadeiro valor da média 79,75 kg,

Tabela 7.1: Todas as amostras possíveis de tamanho igual a dois, com reposição, da população dada na Tabela 6.3, com os respectivos intervalos de confiança(IC).

Amostras	Amostras	Pesos	Média peso ( $\bar{X}$ )	Desvio padrão	IC $\sigma$ conhecido		IC $\sigma$ desconhecido	
					LI	LS	LI	LS
1	A,B	68;80	74,0	8,49	63,99	84,01	-2,24	150,24
2	A,C	68;84	76,0	11,31	65,99	86,01	-25,65	177,65
3	A,D	68;87	77,5	13,44	67,49	87,51	-43,21	198,21
4	B,C	80;84	82,0	2,83	71,99	92,01	56,59	107,41
5	B,D	80;87	83,5	4,95	73,49	93,51	39,03	127,97
6	C,D	84;87	85,5	2,12	75,49	95,51	66,44	104,56
7	B,A	80;68	74,0	8,49	63,99	84,01	-2,24	150,24
8	C,A	84;68	76,0	11,31	65,99	86,01	-25,65	177,65
9	D,A	87;68	77,5	13,44	67,49	87,51	-43,21	198,21
10	C,B	84;80	82,0	2,83	71,99	92,01	56,59	107,41
11	D,B	87;80	83,5	4,95	73,49	93,51	39,03	127,97
12	D,C	87;84	85,5	2,12	75,49	95,51	66,44	104,56
13	A,A	68;68	68,0	0,00	57,99	78,01	-	-
14	B,B	80;80	80,0	0,00	69,99	90,01	-	-
15	C,C	84;84	84,0	0,00	73,99	94,01	-	-
16	D,D	87;87	87,0	0,00	76,00	97,01	-	-

representando 93,75%, valor bem próximo do grau de confiança teórico estabelecido na construção dos intervalos para populações infinitas. Na prática ao selecionarmos uma amostra estamos automaticamente selecionando um dos possíveis intervalos de confiança.

As colunas 8 e 9 apresentam os limites dos intervalos de confiança no caso de  $\sigma$  desconhecido. Podemos ver que os intervalos não são informativos, apesar da grande maioria deles conter o verdadeiro valor da média. Isto ocorre porque o parâmetro  $\sigma$  é estimado a partir de uma amostra muito pequena. Nos casos em que  $\sigma$  é desconhecido amostras pequenas não fornecem intervalos convenientes.

#### 7.4.3.2 Intervalo de Confiança para Uma Proporção Populacional

O intervalo de confiança para uma proporção populacional ( $\pi$ ), é muito semelhante ao intervalo de confiança para uma média populacional com  $\sigma$  conhecido. A principal diferença está no desvio padrão da distribuição amostral das proporções. O erro amostral máximo é dado como em 7.3, assim o intervalo de confiança para a proporção da população, com grau

de confiança  $100\gamma\%$  é dado por:

$$\boxed{\text{I.C.}(\pi;\gamma):P \pm z_{\gamma/2}\sqrt{\frac{\pi(1-\pi)}{n}}.} \quad (7.6)$$

Uma dificuldade encontrada para o cálculo do desvio padrão, é que precisamos saber o valor de  $\pi$ , o qual é desconhecido e desejamos estimar. Na prática utilizamos duas possibilidades para superar esta dificuldade. Uma delas é a substituição de  $\pi(1 - \pi)$  pelo seu valor máximo que é 0,25, valor atingido quando  $\pi = 0,50$ . A outra é a substituição de  $\pi(1 - \pi)$  pelo seu estimador  $P(1 - P)$ . No primeiro caso podemos construir intervalos de confiança de amplitude maior do que o seu verdadeiro valor, isto ocorre quando  $\pi$  for muito diferente de 0,50. Na maioria dos problemas utilizamos a segunda alternativa que fornece o intervalo:

$$\boxed{\text{I.C.}(\pi;\gamma):P \pm z_{\gamma/2}\sqrt{\frac{P(1-P)}{n}}.} \quad (7.7)$$

**Exemplo 1.** Em certo lago, uma amostra aleatória de 1000 peixes acusou 290 tilápias. Construa um intervalo de 95% de confiança para a verdadeira proporção de tilápias na população piscosa do lago. Interpretar o intervalo.

A proporção verificada na amostra (estimativa pontual) vale  $p = 290/1000 = 0,29$ . Para  $\gamma = 0,95$ , o valor de  $z$  é igual a 1,96. O intervalo de confiança é:

$$\begin{aligned} IC(\pi; 95\%) &: 0,29 \pm 1,96\sqrt{\frac{0,29(0,71)}{1000}} \\ &: 0,29 \pm 0,03 \\ &[0,26\%; 0,32\%]. \end{aligned}$$

Podemos afirmar, com 95% de confiança, que a porcentagem de tilápias é um valor entre 26% e 32%.

**Exemplo 2.** Uma amostra de 35 peixes da espécie *Xenomelaniris brasiliensis* coletada na localidade Praia da Barra da Lagoa, Florianópolis, SC, apresentou 45,7% de peixes com comprimento total acima de 50 mm. Encontre um intervalo, com 99% de confiança, dentro do qual deve estar a verdadeira proporção de peixes com comprimento acima de 50 mm. O valor de  $z$  para  $\gamma = 0,99$  é 2,58. O intervalo fica:

$$IC(\pi; 99\%) : 0,457 \pm 2,58\sqrt{\frac{0,457(0,543)}{35}}$$

$$: 0,46 \pm 0,217$$

$$[0,243\%; 0,677\%].$$

Podemos afirmar, com 99% de confiança, que a porcentagem de peixes da espécie *Xenomelaniris brasiliensis* com comprimento total acima de 50 mm, é um valor entre 24,3% e 67,7%.

Observação: a amplitude do intervalo é muito grande. Para obter um intervalo com maior precisão é necessário aumentar o tamanho da amostra.

#### 7.4.4 Determinação do Tamanho da Amostra

Uma das perguntas mais freqüentes em estatística é: qual é o tamanho da amostra necessário para estimar um parâmetro populacional? A resposta a esta pergunta, só é possível de ser dada, após o pesquisador da área de interesse, fornecer algumas informações, como veremos a seguir para amostras aleatórias simples, no caso de estimação da média e da proporção. Resultados para outros métodos de amostragem podem ser encontrados em Bart et al.(1998).

Conforme discutimos na Seção 7.4.2 o processo de estimação envolve um erro que denominamos de erro amostral. As expressões 7.2 e 7.3 mostram como a magnitude desses erros está relacionada com o tamanho da amostra. Exemplos foram apresentados para o cálculo desses erros para determinados tamanhos de amostras. Agora, como o que desejamos é determinar o tamanho de amostra, o que precisamos estabelecer é a magnitude desejada do erro amostral.

##### 7.4.4.1 Determinação do Tamanho da Amostra para Estimar uma Média Populacional

A fórmula para a determinação do tamanho da amostra para a estimação da média da população é obtida a partir da expressão 7.2, e dada por:

$$n = \left( \frac{z_{\gamma/2}\sigma}{e_{max}} \right)^2. \quad (7.8)$$

Na expressão 7.8, podemos ver que o tamanho da amostra depende do grau de confiança desejado, através do valor  $z$ , do desvio padrão da população ( $\sigma$ ) e do erro amostral máximo desejado. O grau de confiança e o erro deverão ser estabelecidos pelo pesquisador, e o desvio padrão conhecido.

**Exemplo 1.** Continuação do exemplo dos pesos, em kg, de pacotes de sementes de milho. Que tamanho de amostra será necessário coletar para garantir um erro amostral de no máximo 0,05 kg, com 95% de confiança, na estimação da verdadeira média. Portanto, desejamos determinar um tamanho de amostra, de modo que tenhamos 95% de confiança, de que a média da amostra difira de no máximo 0,05 kg, para mais ou para menos, da média da população. Aplicando a fórmula do tamanho da amostra e sabendo que  $\sigma = 0,20$  kg, obtemos:

$$n = \left( \frac{1,96(0,20)}{0,05} \right)^2 = 61,46 \cong 62.$$

Portanto, vamos necessitar de 62 pacotes de milho para estimar a média populacional com a precisão e a confiança desejadas.

Na prática, geralmente o desvio padrão populacional ( $\sigma$ ) é desconhecido. Nesse caso, poderíamos usar o desvio padrão ( $S$ ) de uma amostra piloto de  $n'$  elementos e a distribuição  $t$  de Student, empregando-se a seguir a expressão:

$$n = \left( \frac{t_{n'-1;\gamma}^S}{e_{max}} \right)^2, \quad (7.9)$$

onde  $t_{n'-1;\gamma}$  é o valor de tabela, com  $n' - 1$  graus de liberdade (tamanho da amostra piloto menos um), e confiança  $\gamma$ . Se  $n \leq n'$  implica que a amostra piloto já é suficiente para a estimação da média, caso contrário, devemos retirar mais elementos da população para completar o tamanho mínimo da amostra.

A definição do erro amostral máximo pode ser feita de duas maneiras, através de um valor absoluto ou de um valor relativo. O valor absoluto independe do verdadeiro valor da média, enquanto que o valor relativo depende desse valor. Como ele é desconhecido, na prática é comum utilizar-se da média da amostra piloto para a sua definição, conforme apresentado no exemplo a seguir.

**Exemplo 2.** Continuação do exemplo dos pesos ao nascer de bezerros da raça Ibagé. Que tamanho de amostra será necessário para produzir um intervalo de confiança de 95% para a verdadeira média, com uma precisão de 5% da média da amostra piloto?

A amostra piloto de tamanho  $n' = 20$  nos forneceu  $\bar{x} = 26$  kg e  $s = 2$  kg. Temos ainda que a precisão desejada vale  $e_{max} = 0,05(26) = 1,3$  kg e  $t_{(19;0,95)} = 2,093$ . Portanto, o tamanho da amostra vale:

$$n = \left( \frac{2,093(2,0)}{1,3} \right)^2 = 10,37 \cong 11.$$



Necessitamos de uma amostra de 11 bezerros para a precisão e confiança estipuladas pelo pesquisador. Como a amostra piloto tem tamanho  $n' = 20$ , maior que o tamanho da amostra necessário  $n = 11$  bezerros, temos que a amostra piloto já é suficiente para o estudo. Na realidade, a utilização de um tamanho de amostra maior que o determinado, garante uma confiança maior do que 95%, para a precisão dada ou, então, uma precisão maior (erro amostral menor), para uma confiança de 95%.

Podemos ainda usar dois outros procedimentos para estimar o desvio padrão populacional,  $\sigma$ , quais sejam:

- sabemos que aproximadamente 96% dos valores de uma variável aleatória com distribuição normal, encontram-se no intervalo  $(\mu - 2\sigma \text{ e } \mu + 2\sigma)$ , assim, a amplitude deste intervalo vale  $\Delta = 4\sigma$ . Estimando a amplitude a partir dos valores máximo e mínimo esperados para a variável em estudo, obtemos para a estimativa de  $\sigma$  o valor  $\Delta/4$ . Exemplo: para a altura de uma população de pessoas adultas, podemos estimar a amplitude pelo valor  $0,40=1,90-1,50$ , fornecendo uma estimativa para  $\sigma$  de  $0,40/4=0,10\text{m}$ ;
- uma estimativa de  $\sigma$  também pode ser obtida em artigos, livros, dissertações, teses, ou seja, através de referências bibliográficas.

Os resultados apresentados até agora são válidos para populações grandes. Para situações de populações finitas de tamanho  $N$  e amostragem aleatória simples sem reposição, deve-se proceder a correção para população finita, utilizando a seguinte expressão:

$$n_0 = \frac{n}{1 + \frac{n}{N}}. \quad (7.10)$$

Quando ocorrer a desigualdade  $(n \leq 0,05N)$  a correção pode ser desconsiderada. Preliminarmente, a amostra é dimensionada para população infinita, obtendo-se o tamanho  $n$  e, numa segunda fase, corrigimos para população finita, obtendo-se o tamanho  $n_0$ , dado por 7.10.

#### 7.4.4.2 Determinação do Tamanho da Amostra Para Estimar Uma Proporção Populacional

Para encontrarmos o tamanho necessário de uma amostra para estimarmos uma proporção da população, procedemos de forma análoga ao que foi feito para o caso de estimação de uma média da população.

De 7.3 obtemos a expressão para o tamanho da amostra, dado por:

$$n = \left( \frac{z_{\gamma/2}}{e_{max}} \right)^2 \pi(1 - \pi). \quad (7.11)$$

Diferente da estimação da média, neste caso o valor de  $n$  depende do parâmetro  $\pi$  que desejamos estimar. Uma alternativa, quando não temos qualquer conhecimento do valor desse parâmetro, é usar  $\pi = 0,50$ , assim o produto  $\pi(1 - \pi)$  é máximo e igual a  $0,25=1/4$ . Se substituirmos esse valor na expressão 7.11, obtemos:

$$n = \frac{z_{\gamma/2}^2}{4e_{max}^2}. \quad (7.12)$$

Com este procedimento, estamos superestimando o tamanho da amostra, quando o verdadeiro valor de  $\pi$  for muito diferente de 0,50, ou seja, podemos estar trabalhando com uma amostra muito maior do que o necessário. Dependendo do custo de amostragem esta alternativa poderá encarecer muito a pesquisa.

Uma segunda alternativa seria obtermos uma estimativa de  $\pi$  através de uma amostra piloto de  $n'$  elementos. Com base nessa amostra, determinamos o valor de  $n$  através da expressão 7.13.

$$n = \left( \frac{z_{\gamma/2}}{e_{max}} \right)^2 P(1 - P), \quad (7.13)$$

obtida de 7.11 com  $\pi$  substituído pelo seu estimador  $P$ .

Se  $n \leq n'$ , a amostra piloto já é suficiente para a estimação. Caso contrário, deveremos retirar da população, os elementos necessários à complementação do tamanho desejado da amostra.

Uma terceira alternativa seria a utilização de limites para o valor de  $\pi$ . Por exemplo, se sabemos que a proporção de sementes que germinam, de uma determinada espécie, seguramente, é superior a 0,90, então, usamos esse valor para  $\pi$ , isto é, substituímos em 7.11  $\pi(1 - \pi)$  por  $0,90(1 - 0,90) = 0,09$ . Podemos notar que este mesmo valor seria utilizado em uma situação de determinada espécie cuja proporção de germinação é inferior a 0,10.

**Exemplo 1.** Continuação do exemplo de peixes da espécie *Xenomelaniris brasiliensis* da localidade da Praia da Barra da Lagoa, Florianópolis, SC. Se quisermos estimar a proporção

de peixes com comprimento acima de 50mm, qual o tamanho da amostra necessário para que tenhamos 99% de confiança de que o erro de nossa estimativa não seja superior a 5 pontos percentuais(0,05), sabendo que uma amostra piloto de 35 desses peixes apresentou 45,7% deles superiores a 50mm?

O valor de  $z$  para  $\gamma = 0,99$  é 2,58. O tamanho da amostra necessário é:

$$\begin{aligned}n &= \frac{2,58^2}{0,05^2} 0,457(1 - 0,457) \\n &= 2662,56(0,2482) \\n &= 660,71 \\n &\cong 661.\end{aligned}$$

Portanto, precisamos de uma amostra de 661 peixes para obtermos uma proporção amostral que não difira da proporção verdadeira em mais do que 5 pontos percentuais para mais ou para menos.

Neste mesmo exemplo, suponha que não saibamos nada sobre  $\pi$ , qual deve ser o tamanho da amostra pesquisada?

Da expressão 7.13, obtemos:

$$n = \frac{2,58^2}{4(0,05^2)} = 665,64 \cong 666.$$

**Exemplo 2.** É comum na estimação da proporção ( $\pi$ ) de quebra cromossômica em roedores de uma região poluída, cujo valor é da ordem de 0,0001, utilizar uma amostra de tamanho 100.000, valor este que teria um erro amostral máximo igual a:

$$e_{max} = 1,96 \frac{\sqrt{0,0001(1 - 0,0001)}}{\sqrt{100.000}} = 0,00006.$$

Este erro, em termos relativos, representa 62% do valor de  $\pi$ . Em geral, nas pesquisas, o erro amostral máximo não deve ser superior a 10% do verdadeiro valor do parâmetro. No caso de  $\pi = 0,0001$  o erro amostral máximo deveria ser igual a 0,00001, o que implicaria em um tamanho de amostra de:

$$n = \frac{1,96^2}{0,00001^2} 0,0001(1 - 0,0001) = 3.802.119,75 \cong 3.802.120.$$

As expressões apresentadas são válidas para populações grandes. Do mesmo modo que para a estimação de uma média, podemos aplicar a correção para populações finitas de tamanho  $N$  e amostragem aleatória simples sem reposição dada em 7.10.

#### 7.4.4.3 Outro parâmetro: Determinação do Tamanho da Amostra para Estimar a Densidade de Primatas na Natureza

##### *Uma Aplicação em Ecologia*

A utilização de trilhas (transectos, preferencialmente linhas retas) abertas no local é um dos métodos mais usados para a estimativa da densidade (contagem) de várias espécies de animais. É importante que as trilhas estejam bem localizadas, pois elas devem amostrar as diferentes vegetações do local em estudo.

Aqui, o tamanho da amostra corresponde ao número de vezes que devemos percorrer cada trilha selecionada, e é calculada através da expressão:

$$n = \frac{(2ts)^2}{e^2}$$

Os termos que aparecem nesta fórmula já foram discutidos. Ressaltamos que a estimativa do desvio padrão da amostra ( $s$ ), pode ser obtida através de uma amostra piloto (preliminar).

**Exemplo.** Desejamos determinar o tamanho da amostra (número de vezes que devemos percorrer as trilhas), para obter um intervalo de confiança com 90% e erro de 0,2 *animais/km<sup>2</sup>*. De uma amostra piloto com 5 repetições em todas as trilhas obteve-se uma estimativa do desvio padrão com sendo igual a 0,3 indivíduos/km<sup>2</sup>. Para o cálculo do tamanho da amostra temos as seguintes informações:

$t(\text{com } 5-1=4 \text{ graus de liberdade e } \gamma = 0,90)=2,132$

$s=0,3$

$e=0,2$

O tamanho da amostra é dado por:

$$n = \frac{(2.2,132.0,3)^2}{(0,2)^2} = 40,90 \cong 41$$

De acordo com as exigências do intervalo de confiança, precisamos de 41 repetições. Como já foram feitas 5 repetições, necessitamos de mais 36 repetições. Maiores detalhes sobre este procedimento pode ser encontrado em Cullen Jr. et al. (2003)

#### **Exercícios resolvidos**

1. De uma amostra de 100 peixes da espécie *Xenomelaniris brasiliensis*, coletada na Armação do Pântano do Sul, Florianópolis, SC, verificou-se que 57 deles apresentavam comprimento total maior que 50 mm. Com base nessa informação, determine o intervalo de confiança de 99% para a verdadeira proporção de peixes com comprimento total acima de 50 mm. Qual o tamanho de amostra necessário para estimar a verdadeira proporção com precisão de 5%, usando uma confiança de 95%?

Temos  $n = 100$ ,  $p = 57/100 = 0,57$  e  $q = 1 - p = 0,43$ .

$$\begin{aligned} IC(\pi; 99\%) &: 0,57 \pm 2,58 \sqrt{\frac{0,57(0,43)}{100}} \\ &: 0,57 \pm 0,12773 \\ &[0,44227; 0,6977]. \end{aligned}$$

O cálculo do tamanho da amostra fica:

$$\begin{aligned} n &= \left( \frac{1,96}{0,05} \right)^2 0,57(0,43) \\ &= 376,63 \cong 377. \end{aligned}$$

2. O diâmetro médio de *Biomphalaria tenagophila*, examinada uma amostra de 35 animais, foi de 0,871 mm com um desvio padrão de 0,057 mm.

a) Dê a estimativa por intervalo do verdadeiro diâmetro médio utilizando um nível de confiança de 95%.

b) Que tamanho de amostra será necessário para produzir um intervalo de confiança de 95% para a verdadeira média, com uma precisão de 2% da média da amostra preliminar?

Temos,  $n = 35$ ,  $\bar{x} = 0,871$  mm,  $s = 0,057$  mm e  $e = 0,02(0,871) = 0,0174$ .

$$\begin{aligned} IC(\mu; 95\%) &: 0,871 \pm 2,032 \frac{0,057}{\sqrt{35}} \\ &: 0,871 \pm 0,0196 \\ &0,8514 \leq \mu \leq 0,8906. \end{aligned}$$

$$\begin{aligned} n &= \left( \frac{2,032(0,057)}{0,0174} \right)^2 \\ &= 44,21 \cong 45. \end{aligned}$$

3. Em um experimento, 320 de 400 sementes germinaram. Determine o intervalo de confiança de 98% para a verdadeira proporção de sementes que germinam. Para realizar o teste de germinação, quantas sementes serão necessárias utilizar, se desejamos um intervalo de confiança de 95%, com precisão de 4%?

Temos  $p = 320/400 = 0,80$ .

$$\begin{aligned} IC(\pi; 98\%) &; 0,80 \pm 2,33 \sqrt{\frac{0,80(0,20)}{400}} \\ &: 0,80 \pm 0,0465 \\ &[0,7535; 0,8465]. \end{aligned}$$

$$n = \left( \frac{1,96}{0,04} \right)^2 0,80(0,20) = 384,16 \cong 385.$$

4. Sorteou-se uma amostra aleatória simples sem reposição de 100 estudantes do curso de Engenharia Agrônômica e, solicitou-se aos mesmos, que anotassem suas despesas com alimentação no período de um mês. Há 500 estudantes matriculados no curso. O resultado da amostra foi uma despesa média de 40 u.m. e um desvio padrão de 10 u.m. (u.m.= unidades monetárias).

a) Construa um intervalo de 95% de confiança para a verdadeira média.

b) Qual o tamanho da amostra necessário para que tenhamos 95% de confiança de que o erro na nossa estimativa não seja superior a 1,0 u.m.?

Como a amostragem é sem reposição e a população é finita ( $N = 500$ ), com  $n = 100$  maior do que  $25 = 0,05N$  o erro padrão da média vale:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{10}{\sqrt{100}} \sqrt{\frac{500-100}{500-1}} = 0,8953.$$

Como usamos o desvio padrão da amostra, vamos usar a distribuição  $t$  de Student. O intervalo de confiança fica:

$$\begin{aligned} IC(\mu; 95\%) &: 40 \pm 1,99(0,8953) \\ &: 40 \pm 1,782 \\ &[38,218; 41,782]. \end{aligned}$$

Notamos que para a amostra de tamanho  $n = 100$  e grau de confiança de 95% o erro amostral máximo esperado é 1,782 u.um.

Para um erro amostral de 1 u.m., com o mesmo grau de confiança, o tamanho da amostra é dado por:

$$n = \left( \frac{1,99 \times 10}{1} \right)^2 = 396,01 \cong 397.$$

O tamanho da amostra, com a correção para população finita fica:

$$n_0 = \frac{397}{1 + \frac{397}{500}} = 221,29 \cong 222.$$

## 7.5 Exercícios Propostos

1) Um antropólogo mediu as alturas de uma amostra aleatória de 100 homens de determinada população, encontrando a média amostral de 173 cm. Se a variância da população

for de  $9 \text{ cm}^2$ :

- a) calcular um intervalo de 95% de confiança para a altura média de toda a população. Interpretar o I.C.
- b) determinar um intervalo de 99% de confiança para a altura média de toda a população. Interpretar o I.C.
- c) Qual é a probabilidade de encontrar um homem com altura superior a 177 cm.

2) Uma máquina enche pacotes de café com uma variância igual a  $100 \text{ g}^2$ . Ela estava regulada para enchê-los com 500 g, em média. Agora, ela se desregulou e queremos saber qual a nova média verdadeira (populacional). Uma amostra de 25 pacotes apresentou uma média igual a 485 g.

- a) Construir intervalos de confiança de 95% e 99% de confiança para a média verdadeira. Interpretar os intervalos de confiança
- b) Qual o erro máximo associado aos intervalos encontrados em a). Interpretar.
- c) Que tamanho de amostra será necessário para produzir um intervalo de confiança para a verdadeira média populacional, com uma precisão de 3,5 gramas de café em qualquer dos sentidos, dado que o desvio padrão da população é conhecido e igual a 10 gramas de café. Use uma confiança de 99%.

3) De um povoamento de eucaliptos, sorteou-se 30 árvores e determinou-se o diâmetro, em cm, com a finalidade de estimar o diâmetro médio do povoamento. Diâmetros de 30 eucaliptos em cm:

10,1	15,8	18,5	22,3	23,5	17,2	17,8	18,7
16,7	29,1	28,0	30,3	26,8	28,0	17,8	18,9
28,9	27,9	22,5	32,9	29,5	28,3	34,2	38,5
38,5	35,5	34,2	31,8	32,5	41,8		

Com base nessa amostra calcule:

- a) os intervalos de confiança aos níveis de 95% e 99%;
- b) qual o tamanho da amostra necessário para estimar o diâmetro médio de plantas de eucalipto se o erro máximo deve ser de 5% da média para um nível de confiança de 95%?

4) De 1000 lavouras de arroz, foi levantada uma amostra de 25 lavouras e a informação a respeito da produtividade permitiu o cálculo do rendimento médio, por hectare ( $1 \text{ ha} = 10.000 \text{ m}^2$ ), que foi de 3400 kg com um desvio padrão de 150 kg.

- a) Determine o intervalo de confiança a 95% e 99%, para o verdadeiro rendimento médio.
- b) Que tamanho deve ter a amostra para que seja de 95% a confiança na estimativa  $3400 \pm 100$ ?

5) Do rebanho bovino de determinado município, em maio, constatou-se que 30% de 20 animais examinados estavam com febre aftosa, e em junho, examinaram-se outros 20 animais, constatou-se que 18% estavam acometidos daquela moléstia.

Deseja-se saber entre que limites esteve a verdadeira proporção de animais com febre aftosa, tanto no mês de maio como no mês de junho. Utilize um nível de confiança de 95%.

Qual o erro máximo associado aos intervalos encontrados. Interpretar.

Num trabalho futuro, qual o tamanho de amostra necessário para obter um intervalo de 95% de confiança para a proporção, com um erro tolerável de 0,08, em cada um dos meses?.

6) Da produção de compotas de pêssego, da safra 79/80, de uma determinada indústria, foram examinadas 30 latas e 3 delas estavam fora dos padrões para a exportação. Determinar o intervalo de confiança, a 99%, para a verdadeira proporção de latas que estão fora dos padrões para exportação.

7) De uma partida de 1000 sacos de sementes de trigo tomou-se uma amostra de 50 sacos e verificou-se que 10 deles apresentavam teor de umidade acima do máximo permitido para o armazenamento. Com base nessa informação, determine o intervalo de confiança, de 99%, para a verdadeira proporção de sacos com teor de umidade acima do limite máximo permitido e o número máximo e mínimo de sacos de sementes com esse problema.

8) Num experimento de campo, o tratamento A, com nove repetições, tem média de 1500 kg/ha de feijão. Temos uma estimativa do desvio padrão  $s=180$  kg/ha, com 30 graus de liberdade. Obter o intervalo de confiança para a média, ao nível de 95% de confiança.

9) De uma amostra de 26 valores de número de colônias de *Salmonella typhimurium*, em placas que contém a infusão de *Bauhinia fortificata*, encontrou-se um desvio padrão igual a 47 colônias, e média igual a 76,15 colônias.

a) Qual o intervalo de confiança de 95% para a média populacional?

b) Que tamanho deve ter uma amostra para que o intervalo  $76,15 \pm 5$  tenha 95% de confiança?

10) Estime (Intervalo de Confiança), o comprimento médio ( $\mu$ ) de *Pnaeus schmitti* em condições normais no estuário (Lagoa de Ibiraquera) e num viveiro de cultivo (Fazenda de engorda de Laguna), sabendo-se que uma amostra de 30 indivíduos apresentou os resultados da Tabela 7.2. (Use  $\gamma = 95\%$ ).

11) Uma amostra aleatória de 60 progênies indica que 70% delas apresentam resistência à antracnose. Construir um intervalo de confiança para  $\pi$ , a proporção de progênies resistentes à antracnose, com  $\gamma = 95\%$ .

12) O diâmetro de altura do peito (DAP) de *Tabebuia ochracea*, o Ipê-Amarelo do Campo, família Bigoniáceas, é uma variável aleatória com distribuição aproximadamente normal de média 0,70 cm e desvio padrão de 0,40 cm. Encontre um intervalo de confiança em torno da



Tabela 7.2: Valores de comprimento

Lagoa de Ibiraquera		Fazenda de engorda de Laguna	
Classes	Frequências	Classes	Frequências
4,70 ┤ 5,07	2	7,6 ┤ 8,2	2
5,07 ┤ 5,44	2	8,2 ┤ 8,8	0
5,44 ┤ 5,81	11	8,8 ┤ 9,4	9
5,81 ┤ 6,18	11	9,4 ┤ 10,0	13
6,18 ┤ 6,55	3	10,0 ┤ 10,6	4
6,55 ┤ 6,92	1	10,6 ┤ 11,2	2

média que contenha 95% dos valores de DAP.

13) De um experimento anterior sabe-se que o desvio padrão da altura de plantas de um determinado híbrido de milho é  $\sigma = 0,80$  m.

a) Seleccionada uma amostra de 72 plantas deste híbrido, observou-se uma média de 2,10 m. Qual o intervalo de confiança de 95% para a média populacional?

b) Que tamanho deve ter uma amostra para que o intervalo  $2,10 \pm 0,20$  tenha 99% de confiança?

14) Um agrônomo realizou um levantamento para estudar o desenvolvimento de duas espécies de árvores, a Bracatinga e a Canafístula. Para esta finalidade foram coletadas duas amostras de tamanhos igual a 10 árvores. Os resultados para altura, em metros, estão descritos abaixo para as duas amostras:

Bracatinga					Canafístula				
6,5	6,9	6,9	8,6	8,7	8,2	10,0	10,3	13,4	14,4
9,3	10,1	11,4	15,2	17,2	14,8	15,9	20,6	21,9	23,8

Para verificar a hipótese de que as alturas das duas espécies são diferentes, o agrônomo adotou o seguinte critério. Construir os intervalos com 95% de confiança, para cada uma das espécies. Se os intervalos se sobrepõem (se interceptam) concluir que não há diferenças significativas entre as duas alturas médias, caso contrário, concluir que há diferenças entre as mesmas. Baseado neste critério qual a conclusão do agrônomo?

15) As bromélias são plantas epífitas que vivem sobre galhos e troncos das árvores, denominados forófitos. O estágio arbóreo pioneiro é formado pela vegetação que proporciona o sombreamento efetivo do solo, originando um microclima úmido e sombreado. Uma amostra de 35 forófitos no estágio arbóreo pioneiro da Floresta Ombrófila na Ilha de Santa Catarina, apresentou 40% sem bromélias.

1. Encontre um intervalo de confiança com 99% para a verdadeira proporção de forófitos sem bromélias. Faça a interpretação do intervalo.
2. Qual o erro máximo associado ao intervalo? Faça a interpretação.
3. Para estimar a proporção de forófitos sem bromélias ( $\pi$ ), qual o tamanho de amostra necessário para que tenhamos 99% de confiança de que o erro de nossa estimativa não seja maior do que 5%? Faça a interpretação do  $n$ . Considere a amostra com 35 forófitos como sendo uma amostra preliminar (piloto).
4. Uma outra amostra de 35 forófitos, no estádio de floresta secundária, apresentou 2% sem bromélias. Encontre o intervalo de confiança com 99% para a verdadeira proporção ( $\pi$ ), de forófitos sem bromélia.
5. Para concluir se existe diferença estatística entre as duas proporções, o(a) Biólogo(a), usou o seguinte critério: se os intervalos se sobrepõem (se cruzam, interceptam) conclui que não há diferença entre as duas proporções. Qual a conclusão do(a) Biólogo(a)? Faça a representação gráfica dos intervalos.

16) Os valores de DAP (Diâmetro à Altura do Peito), em cm, de forófitos possuidores de *Vriesea incurvata* (bromélia, conhecida como espada de Davi), em vegetação primária da Floresta Tropical Atlântica, em Santo Amâro da Imperatriz, SC, foram:

10,02	11,70	13,92	14,63	4,75	7,54	12,30
7,80	2,66	6,13	9,20	9,90	14,20	2,27
6,40	9,93	4,60	22,20	12,65	33,60	7,20
8,90	16,90	3,60	10,75	18,00	6,00	4,15
31,80	33,00					

1. Construa o intervalo de confiança em torno da média verdadeira que contenha 95% dos valores de DAP.
2. Que tamanho de amostra ( $n$ ) será necessário coletar para produzir um intervalo de confiança com 95% para a verdadeira média, com precisão de 5% da média da amostra?

## 8 Testes de Hipóteses Sobre os Parâmetros

### 8.1 Introdução

Em muitas situações práticas o interesse do pesquisador está relacionado com a verificação de uma afirmação sobre um ou mais parâmetros populacionais. Por exemplo:

1. a produtividade média de milho em Santa Catarina, é de 2300 kg/ha;
2. os comprimentos médios dos ante-braços de duas espécies de morcegos são iguais;
3. a proporção de fixação de fitoplâncton em dois tipos de solos é a mesma;
4. a produção média de duas cultivares de feijão é a mesma;
5. a sobrevivência de mudas depende da época de plantio.

Estas afirmações são consideradas hipóteses estatísticas, porque se referem a parâmetros populacionais, e comumente precisam ser verificadas a partir de amostras. O ramo da estatística que trata deste problema é conhecido como teste de hipóteses. Neste capítulo apresentamos alguns testes relacionados com médias e proporções.

Para exemplificar, vamos considerar a situação na qual um produtor precisa decidir pela compra ou não de sementes de milho fornecidas por um distribuidor, que afirma que a proporção de germinação das sementes é  $\pi = 0,94$ . Para tanto ele observou a proporção de germinação de uma amostra aleatória simples de 100 sementes e encontrou o valor de  $p = 0,93$ . A partir desse resultado o produtor deveria discordar do distribuidor? Em outras palavras, o resultado obtido pelo produtor na amostra selecionada seria improvável de ocorrer sendo a afirmação do distribuidor verdadeira? Aqui também devemos levar em consideração a variabilidade amostral, já vista no estudo de distribuições amostrais estudadas na Seção 6.4, e utilizadas no capítulo 7.

O objetivo de um teste estatístico de hipóteses é fornecer ferramentas que nos permitam aceitar ou rejeitar uma hipótese estatística através dos resultados de uma amostra.

### 8.2 Construção de Testes de Hipóteses

A construção de um teste de hipóteses requer a especificação de duas hipóteses, denominadas de hipótese nula ( $H_0$ ) e hipótese alternativa ( $H_1$ ), e de um critério para a rejeição da hipótese nula.

As hipóteses estatísticas são formuladas da seguinte maneira:

1. **hipótese nula( $H_0$ ):** é a hipótese que sugere um valor para o parâmetro populacional ou a igualdade dos parâmetros em teste. No nosso exemplo sobre germinação, a hipótese nula é que a verdadeira proporção de germinação de sementes é de  $\pi = 0,94$  e a representamos por:

$$H_0 : \pi = 0,94.$$

Um outro exemplo, seria a comparação da produção média de duas cultivares de feijão, representadas por  $\mu_1$  e  $\mu_2$ . Neste caso a hipótese nula seria representada por:

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

2. **Hipótese alternativa( $H_1$ ):** é a hipótese que sugere que a afirmação que estamos fazendo na hipótese nula é falsa. No nosso exemplo sobre germinação, a hipótese alternativa é que a proporção de germinação do lote é menor que 0,94, pois devemos nos precaver contra o lote ter menos do que 94% de germinação e, portanto, a distribuidora não está certa, e a representamos por:

$$H_1 : \pi < 0,94.$$

A não rejeição da hipótese nula  $\pi = 0,94$  automaticamente implicaria na não rejeição de um valor de  $\pi > 0,94$ .

Para o exemplo da produção das cultivares de feijão, poderíamos formular três diferentes hipóteses alternativas:

- (a)  $H_1 : \mu_1 \neq \mu_2$  (temos um teste bilateral);
- (b)  $H_1 : \mu_1 > \mu_2$  (temos um teste unilateral à direita);
- (c)  $H_1 : \mu_1 < \mu_2$  (temos um teste unilateral à esquerda).

dependendo do problema. Portanto, a construção da hipótese alternativa, depende do grau de conhecimento biológico ou agrônômico sobre o fenômeno, ou das informações que se têm do problema em estudo.

Qualquer que seja a decisão tomada, de rejeitar ou não rejeitar uma hipótese nula, estamos correndo o risco de uma decisão errônea, pois estamos nos baseando em resultados de uma amostra, os quais estão sujeitos as variações amostrais. Podemos rejeitar uma hipótese nula quando deveríamos aceitá-la, ou aceitar uma hipótese nula quando ela deveria

Tabela 8.1: Erros tipo I e II em testes de hipóteses

Ação	Se $H_0$ é:	
	verdadeira	falsa
não rejeitar $H_0$	decisão correta	<i>erro tipo II</i>
rejeitar $H_0$	<i>erro tipo I</i>	decisão correta

ser rejeitada. Por exemplo, suponha que a afirmação do distribuidor de sementes é verdadeira e o nosso critério indica que a hipótese nula deve ser rejeitada, isto nos levaria a cometer o que chamamos de *erro do tipo I*. Por outro lado, suponha que a afirmação do distribuidor é falsa e o nosso critério indica que a hipótese nula não deve ser rejeitada, isto nos levaria a cometer o que chamamos de *erro tipo II*. O critério para a rejeição da hipótese nula é construído de modo a garantir que as probabilidades de cometermos estes dois tipos de erros sejam pequenas. Esquemáticamente temos:

A probabilidade de *erro tipo I* é usualmente representada pela letra grega  $\alpha$  e denominada *nível de significância do teste*. Um valor aceitável para esta probabilidade é 0,05 ou 5%. Dependendo do problema este valor pode ser maior ou menor. A probabilidade de *erro tipo II* é usualmente representada pela letra grega  $\beta$ . Usualmente este valor não deve exceder a 0,20 ou 20%. No decorrer deste capítulo, apresentaremos uma discussão mais detalhada destas probabilidades.

### 8.3 Teste Para Proporções Populacionais

Aqui, desejamos avaliar certas afirmações feitas sobre proporções ou porcentagens populacionais. Vamos considerar o estudo de teste de hipóteses sobre proporções em duas situações, quais sejam: 1) teste para uma proporção populacional e 2) teste para duas ou mais proporções.

#### 8.3.1 Teste Para Uma Proporção Populacional

Feita uma afirmação sobre uma proporção populacional( $\pi$ ), desejamos estabelecer um critério para rejeitá-la ou não a partir dos dados de uma amostra aleatória simples. Este critério será baseado na proporção amostral, a qual sabemos ter uma distribuição normal,

quando o tamanho da amostra é suficientemente grande, com média  $\pi$  e desvio padrão,

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}},$$

onde  $n$  é o tamanho da amostra. Voltando ao exemplo do poder germinativo de sementes no qual a hipótese nula afirma que a proporção verdadeira é igual a 0,94, ou seja:

$$H_0 : \pi = 0,94.$$

A hipótese alternativa é dada por:

$$H_1 : \pi < 0,94.$$

Vamos considerar o critério de rejeição da hipótese nula quando a proporção amostral, baseada em uma amostra de tamanho 100, for menor do que 0,94. Podemos verificar que a probabilidade de rejeição indevida da hipótese nula é de 0,50 (50%), conforme segue:

$$Prob(P < 0,94 | \pi = 0,94) = P\left(Z < \frac{0,94 - 0,94}{\sqrt{\frac{0,94(1-0,94)}{100}}}\right) = P\left(Z < \frac{0,94 - 0,94}{0,02375}\right) = P(Z < 0) = 0,50.$$

Este valor, como vimos, é a probabilidade de cometermos o *erro tipo I*, caso esse fosse o critério utilizado para a rejeição da hipótese. Em outras palavras, existe uma probabilidade igual a 0,50 de obtermos uma proporção amostral menor do que 0,94 em um lote que efetivamente tem uma proporção de 0,94 de germinação. Precisamos então estabelecer um outro valor, menor do que 0,94, de modo a garantir uma probabilidade do erro tipo I pequena, ou seja, *nível de significância* do teste baixo.

Vamos considerar que o *nível de significância* fixado pelo pesquisador seja  $\alpha = 0,05$  ou 5%. Na distribuição normal padrão, o valor de  $z$ , abaixo do qual temos 5% das observações, vale -1,65. Observe que o nosso teste é feito no lado esquerdo da curva, isto é, ele é um teste unilateral à esquerda. Você pode obter este valor diretamente da tabela da distribuição normal padrão, basta encontrar o valor de  $z$  que corresponde a uma probabilidade de 0,45; ou, então, usar um programa (software) estatístico. Não esqueça que as probabilidades são dadas no corpo da tabela da distribuição normal padrão. Encontre este valor. Podemos encontrar o valor crítico do teste ( $p_c$ ), que corresponde ao valor de  $z = -1,65$ , da seguinte forma:

$$\begin{aligned} z &= \frac{p_c - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \\ -1,65 &= \frac{p_c - 0,94}{0,02375} \\ p_c &= 0,90. \end{aligned} \tag{8.1}$$

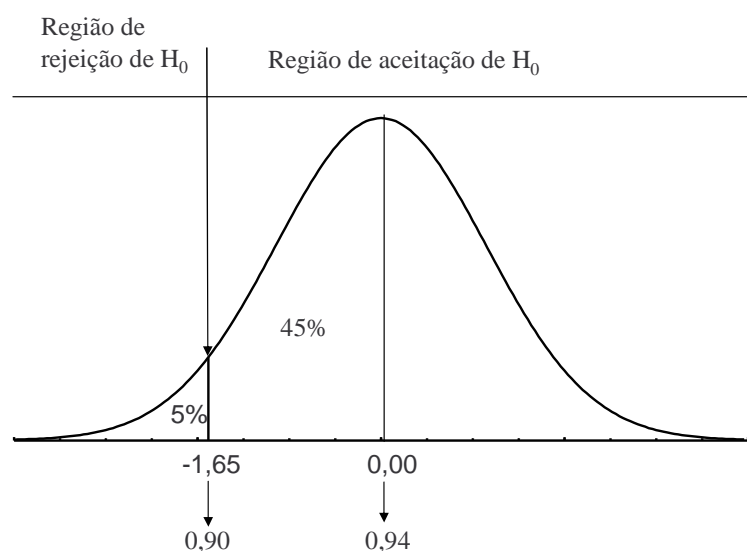


Figura 8.1: Regiões de aceitação e de rejeição da hipótese nula para uma probabilidade de erro de 5%

A *região de rejeição* do teste é composta por todos os valores menores a 0,90. Por outro lado, a *região de aceitação* é composta por todos os valores maiores ou igual a 0,90. Em outras palavras, a hipótese nula será rejeitada caso a proporção amostral de uma amostra de tamanho 100 seja menor do que 0,90. A Figura 8.1 ilustra as regiões de aceitação e de rejeição da hipótese nula para um nível de significância de 5%. Embora esperamos que o poder germinativo seja de 0,94, nós temos 5% das proporções amostrais com valores abaixo de 0,90. Portanto, se aceitarmos tomar  $z = -1,65$  ou  $p_c = 0,90$  como linha divisória entre o que é uma diferença casual e uma diferença real, há um risco de 5% de rejeitarmos a hipótese nula quando ela é de fato verdadeira.

A proporção verificada na amostra foi igual a  $p = 0,93$ , um valor pertencente a região de aceitação. Portanto, concluímos que não devemos rejeitar a hipótese nula de que  $\pi = 0,94$ , ao nível de significância de 5%.

Pelo que foi exposto, podemos estabelecer uma seqüência lógica para a realização de um teste de hipótese. Os itens dessa seqüência são:

1. todo o trabalho inicia com a formulação das hipóteses estatísticas. Como vimos, todo teste estatístico envolve duas hipóteses: hipótese nula ( $H_0$ ) e a hipótese alternativa ( $H_1$ );
2. o pesquisador precisa decidir qual a estatística amostral ou estimador que ele vai usar ( $\bar{X}, P, S^2, \bar{X}_1 - \bar{X}_2$ ) e, ter conhecimento dos resultados da distribuição amostral do

estimador, pois como vimos, ela descreve a variabilidade amostral;

3. o pesquisador precisa fixar um nível de significância ( $\alpha$ ) para o teste e, assim, delimitar as regiões de aceitação e de rejeição da hipótese nula;
4. calcular o valor do estimador com os dados da amostra;
5. consiste em verificar se o valor do estimador pertence ou não à região de rejeição de  $H_0$ . Se pertence à região de rejeição, rejeitamos  $H_0$ , caso contrário, não rejeitamos  $H_0$ .

Chamamos a atenção para o fato do valor crítico depender não só do nível de significância do teste, como também do tamanho da amostra utilizado. Na Seção 8.3.2 iremos discutir com detalhes a determinação do tamanho de amostra para a realização de testes de hipóteses, o qual envolverá também o controle do erro do tipo II.

No teste que realizamos, a hipótese alternativa é unilateral à esquerda ( $\pi < \pi_0$ ), onde  $\pi_0$  é um valor conhecido, no exemplo,  $\pi_0 = 0,94$ . Duas outras possibilidades para a hipótese alternativa seriam:

$$\begin{aligned} H_1 &: \pi > \pi_0 \text{ (teste unilateral à direita)} \\ H_1 &: \pi \neq \pi_0 \text{ (teste bilateral)} \end{aligned} \tag{8.2}$$

A Figura 8.2 ilustra essas três possibilidades para a hipótese alternativa. Observe as respectivas regiões de aceitação e de rejeição da hipótese nula em cada caso. No teste bilateral, as áreas de rejeição correspondem a  $\alpha/2$  cada uma, enquanto que nos testes unilaterais, a área de rejeição corresponde a  $\alpha$ . Observe que o sinal  $>$  ou  $<$  aponta para o lado da curva a ser utilizado. Quando definimos o nível de significância ( $\alpha$ ) do teste, conseqüentemente temos o ponto crítico, ou seja, o ponto limitrófico entre as duas regiões.

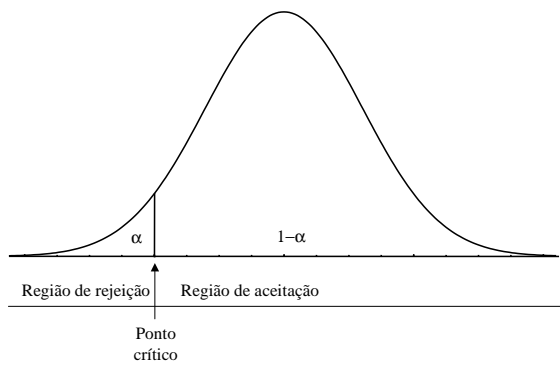
**Exemplo.** Um(a) Biólogo(a), com base em conhecimentos teóricos e práticos, afirma que a proporção ( $\pi$ ) de forófitos sem bromélias no estágio arbóreo pioneiro da Floresta Ombrófila na Ilha de Santa Catarina é igual a 0,47. Em uma amostra de 35 forófitos, 24 ( $p = 0,686$ ) não apresentaram bromélias. Teste a afirmativa do(a) Biólogo(a) ao nível de significância de 5%.

Hipóteses:

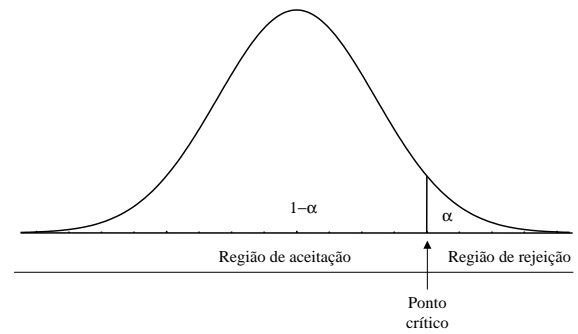
$$H_0 : \pi = 0,47 \quad H_1 : \pi \neq 0,47$$

Determinação da região de rejeição: como a hipótese alternativa é bilateral, precisamos determinar dois valores críticos, denominados de  $p_{c1}$  e  $p_{c2}$ , conforme a Figura 8.2(c). Os

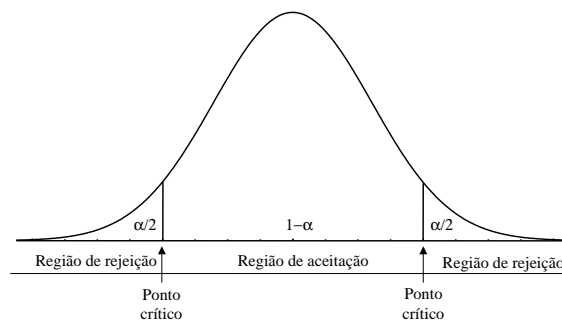




(a) Teste unilateral à esquerda



(b) Teste unilateral à direita



(c) Teste bilateral

Figura 8.2: Testes de hipóteses unilaterais e bilateral

valores dos pontos críticos são dados por:

$$-1,96 = \frac{p_{c1} - \pi}{\sqrt{\frac{\pi}{1-\pi}}} = \frac{p_{c1} - 0,47}{\sqrt{\frac{0,47(1-0,47)}{35}}} \Rightarrow p_{c1} = 0,305$$

$$+1,96 = \frac{p_{c2} - \pi}{\sqrt{\frac{\pi}{1-\pi}}} = \frac{p_{c2} - 0,47}{\sqrt{\frac{0,47(1-0,47)}{35}}} \Rightarrow p_{c2} = 0,635$$

O valor 1,96 é obtido diretamente da tabela da distribuição normal padrão e corresponde ao valor de Z acima do qual temos  $5\% \div 2$  das observações. O valor -1,96 corresponde ao valor de Z abaixo do qual temos  $5\% \div 2$  das observações. A hipótese nula será rejeitada se a proporção amostral for menor do que 0,305 ou maior do que 0,635. Como o Biólogo encontrou em sua amostra de 35 forófitos uma proporção igual a 0,686 sem bromélias, rejeitamos a afirmação do Biólogo sobre a verdadeira proporção de forófitos sem bromélias, ao nível de significância de 5%.

### *Teste de Hipóteses Através do valor p*

Uma forma mais direta para realizar um teste de hipóteses é através do cálculo da probabilidade de se obter um valor mais extremo do que o obtido pelo pesquisador dado que a hipótese nula é verdadeira. Esta probabilidade é denominada *valor p* e um *valor p* pequeno significa duas coisas: 1) ou o pesquisador observou um resultado pouco provável de ocorrer, supondo a hipótese nula verdadeira e 2) ou a hipótese nula é falsa. Na prática ficamos com a alternativa 2, isto é, rejeitamos a hipótese nula quando o *valor p* é menor do que o nível de significância( $\alpha$ ) fixado pelo pesquisador, nível este que representa a probabilidade de rejeição indevida da hipótese nula.

**Exemplo.** Para o caso da germinação de sementes, o *valor p* representa a probabilidade de obtermos um valor de proporção de germinação menor ou igual do que 0,93, de uma amostra de 100 sementes, extraída de um lote (população) com proporção de germinação de 0,94. Com o auxílio da tabela da distribuição normal padrão, dada no apêndice 3, obtemos:

$$\begin{aligned} \text{valor } p = Prob(P \leq 0,93) &= Prob\left(z \leq \frac{0,93 - 0,94}{0,02375}\right) \\ &= Prob(z \leq -0,42) \\ &= 0,5 - Prob(0 \leq z \leq 0,42) \\ &= 0,5 - 0,1628 \\ &= 0,3372. \end{aligned}$$

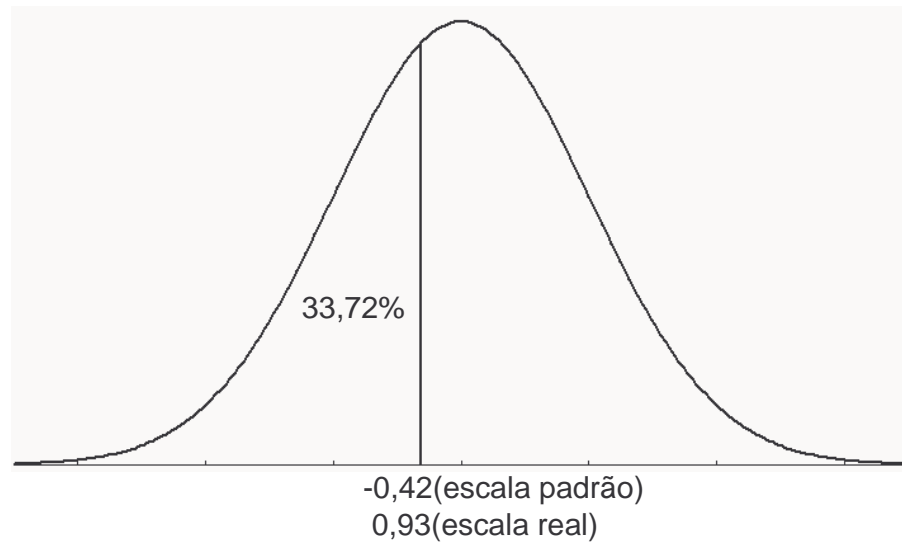


Figura 8.3: Probabilidade de encontrar um valor menor que 93% de um lote com 94% de poder germinativo

A Figura 8.3 mostra a área da distribuição normal padrão correspondente ao *valor p*. Portanto, nós temos uma probabilidade bastante alta, maior do que  $\alpha = 0,05$ , de obter um valor de proporção de germinação igual ou menor que 0,93 devido à variação amostral, ou seja, temos 33,72% de chance de obter uma amostra com uma proporção de germinação igual ou menor do que a encontrada, de um lote com 0,94 de proporção de germinação.

**Exemplo.** Para o caso dos forófitos sem bromélias, o *valor p* representa a probabilidade de obtermos um valor de proporção de forófitos em bromélias igual ou mais extremo do que o valor encontrado pelo pesquisador, que poderá ser maior ou menor do que 0,47, tendo em vista a hipótese alternativa bilateral formulada. No exemplo o valor encontrado pelo pesquisador na amostra foi 0,686, numa amostra de 35 forófitos, fornecendo um *valor p* igual a:

$$\begin{aligned}
 \text{valor } p &= 2 \times \text{Prob}(P \geq 0,686) \\
 &= 2 \times \text{Prob}\left(z \geq \frac{0,686 - 0,47}{0,0844}\right) \\
 &= 2 \times \text{Prob}(z \geq 2,56) \\
 &= 2 \times [0,5 - \text{Prob}(0 \leq z \leq 2,56)] \\
 &= 2 \times (0,5 - 0,4948) \\
 &= 2 \times 0,0052 = 0,0104.
 \end{aligned}$$

Como o *valor p* é menor do que o nível de significância fixado em 0,05, rejeitamos a hipótese

nula a este nível de significância. Notamos que a hipótese nula também seria rejeitada caso o nível de significância tivesse sido fixado em 0,02 ou 2%. Na prática, o valor do nível de significância a ser estabelecido pelo pesquisador, depende do problema em estudo.

Os testes para duas ou mais proporções serão estudados na Seção 8.3.3.

### 8.3.2 Poder do Teste e Tamanho da Amostra para Uma Proporção Populacional

Quando rejeitamos a hipótese nula, sabemos a probabilidade de podermos estar tomando uma decisão incorreta através da fixação do nível de significância do teste, ou seja, rejeitar uma hipótese nula quando ela é verdadeira. Porém não sabemos qual seria a probabilidade de rejeitar a hipótese nula sendo ela falsa, em outras palavras, qual seria o *poder* que o teste teria para detectar uma hipótese nula falsa. Veja estas possibilidades na Tabela 8.1. O poder do teste está relacionado com o erro do tipo II através da relação:  $\text{poder} = 1 - \beta$ , onde  $\beta$  representa a probabilidade de erro tipo II. Por exemplo, no problema de germinação de sementes, qual seria a probabilidade de rejeitarmos um lote com proporção de germinação de  $\pi = 0,93$  ao invés de  $\pi = 0,94$ , como afirmado pelo distribuidor, com base no critério estabelecido? Lembremos que, baseados em uma amostra de 100 sementes, estabelecemos que a hipótese nula deveria ser rejeitada caso a proporção encontrada na amostra fosse inferior a 0,90, ao nível de significância  $\alpha = 0,05$ . O cálculo desta probabilidade é dado por:

$$Prob(P < 0,90 | \pi = 0,93) = Prob\left(Z < \frac{0,90 - 0,93}{\sqrt{\frac{0,93(1-0,93)}{100}}}\right) = Prob(Z < -1,18) = 0,50 - 0,381 = 0,12 \quad (8.3)$$

e está representado na Figura 8.4.

Este valor de probabilidade significa que temos somente 12% de chance de rejeitar um lote com um ponto percentual menor do que a porcentagem de germinação indicada pelo distribuidor.

A Tabela 8.2 apresenta valores desta probabilidade (para  $n = 100$ ) para diferentes valores da verdadeira proporção( $\pi$ ), sabendo que a hipótese nula deveria ser rejeitada caso a proporção encontrada na amostra fosse inferior a 0,90. Da Tabela 8.2 podemos concluir que o poder do teste aumenta quanto maior for a diferença entre o valor que o distribuidor afirma ter( $\pi = 0,94$ ) e o verdadeiro valor da proporção de germinação do lote. Por exemplo, temos somente 12% de chance de detectar uma diferença de um ponto percentual,  $\delta = 0,010$ , conforme cálculo apresentado. Mesmo para detectar uma diferença de quatro pontos percentuais,  $\delta = 0,040$ , a chance seria só de 50%. Em outras palavras, este teste não fornece uma boa garantia para o produtor rejeitar um lote de sementes com proporção de germinação bem inferior a 0,94, com uma amostra de tamanho  $n = 100$  sementes. É necessário uma amostra de tamanho maior.

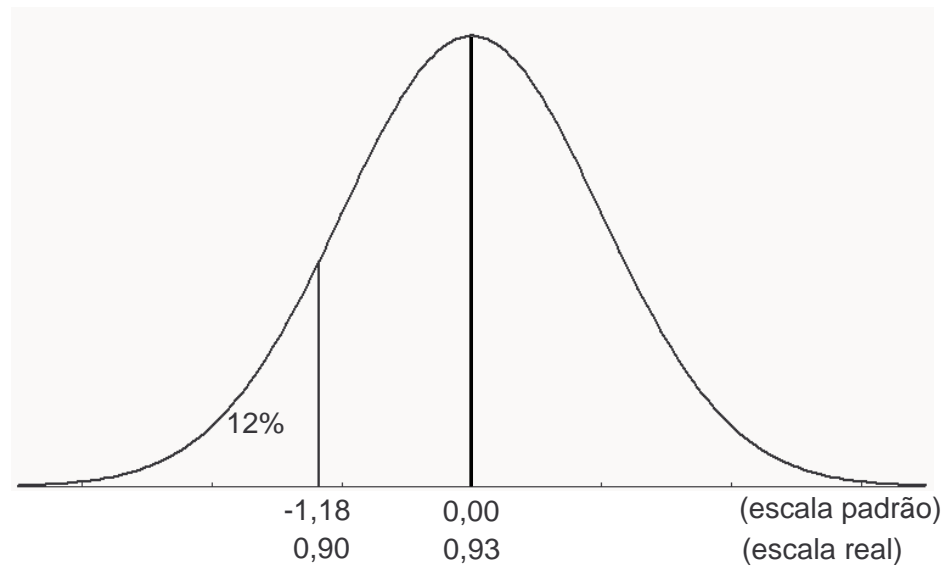


Figura 8.4: Probabilidade de rejeitar um lote com  $\pi = 0,93$  e  $p_c = 0,90$

Tabela 8.2: Poder do teste para diversos valores de proporções populacionais, para amostras de tamanhos  $n = 100$  e  $n = 400$

Proporção populacional ( $\pi$ )	Diferença a detectar( $\delta$ )	Poder( $n = 100$ )	Poder( $n = 400$ )
0,935	0,005	0,08	0,11
0,930	0,010	0,12	0,22
0,925	0,015	0,17	0,35
0,920	0,020	0,23	0,50
0,915	0,025	0,30	0,64
0,910	0,030	0,36	0,76
0,905	0,035	0,43	0,85
0,900	0,040	0,50	0,91

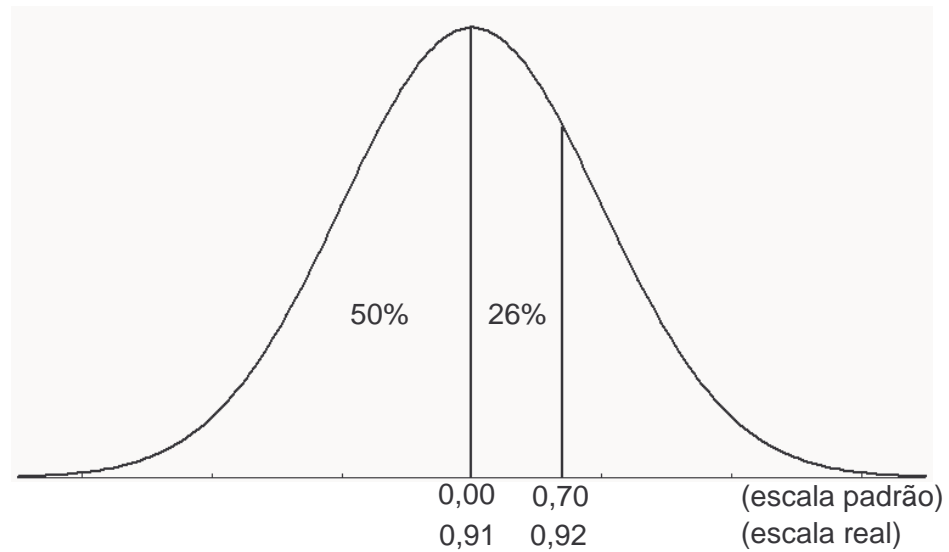


Figura 8.5: Poder do teste (50%+26%=76%) para  $\pi = 0,91$  e  $p_c = 0,92$  com  $n = 400$  sementes

Vamos trabalhar agora com uma amostra de tamanho 400 sementes para verificar como ficam os valores do poder do teste. Para um nível de significância de 5%, o valor crítico do teste( $p_c$ ) é obtido conforme 8.1 e seu valor é:

$$\begin{aligned}
 z &= \frac{p_c - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \\
 -1,65 &= \frac{p_c - 0,94}{0,0119} \\
 p_c &= 0,92.
 \end{aligned}$$

onde  $\sqrt{\pi(1-\pi)/n} = \sqrt{0,94(1-0,94)/400} = 0,0119$ . A região de rejeição do teste é formada por todos os valores menores a 0,92, ou seja, vamos rejeitar a hipótese nula ( $\pi = 0,94$ ), quando a proporção encontrada na amostra for menor do que 0,92, ao nível de significância de 5%. Este é o critério do nosso teste, baseado numa amostra de tamanho  $n = 400$  sementes. Os valores do poder do teste, isto é, as probabilidades de rejeitar a hipótese nula quando ela é falsa, estão dadas na Tabela 8.2. As probabilidades foram calculadas como em 8.3, usando 0,92 no lugar de 0,90 como valor crítico.

A Figura 8.5 representa o poder do teste para  $\pi = 0,91$ , para uma amostra de 400 sementes, o qual vale 76%.

Comparando os valores das duas últimas colunas da Tabela 8.2, verificamos que o poder do teste para detectar uma diferença de 4 pontos percentuais, aumenta de 0,50 para 0,91. Isto significa que uma diferença dessa magnitude tem uma grande chance de ser detectada, o que

não ocorreria com uma amostra de 100 sementes. Por outro lado, mesmo com uma amostra de 400 sementes (Tabela 8.2) a chance de detectar uma diferença de 2 pontos percentuais ( $\delta = 0,020$ ) não seria muito grande (0,50). Faz-se necessário a determinação do tamanho apropriado da amostra para detectar uma diferença  $\delta$ , de interesse do pesquisador, com alto poder do teste e nível de significância  $\alpha$ .

O cálculo do tamanho da amostra, para um *teste de hipótese unilateral à esquerda* ( $H_1 : \pi < \pi_0$ ), é dado pela expressão 8.4

$$n = \frac{Z_\alpha \sqrt{\pi_0(1-\pi_0)} + Z_\beta \sqrt{(\pi_0 - \delta)(1 - \pi_0 + \delta)}}{\delta^2}^2, \quad (8.4)$$

onde  $Z_\alpha$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $\alpha$ ,  $Z_\beta$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $1 - \text{poder} = \beta$ ,  $\pi_0$  é o valor da proporção populacional supondo-se que a hipótese nula é verdadeira e  $\delta$  é a diferença a ser detectada.

**Exemplo.** O pesquisador não deseja ficar com um lote que tenha uma proporção de germinação menor do que três pontos percentuais ( $\delta = 0,03$ ), daquela indicada pelo distribuidor ( $\pi_0 = 0,94$ ). Para a determinação do tamanho da amostra precisamos ainda especificar o nível de significância e o poder do teste. Os valores usualmente considerados na prática são 0,05 e 0,80, respectivamente.

Para  $\alpha = 0,05$ ,  $Z_{0,05} = 1,65$ , e para  $\text{poder} = 0,80$  temos  $\beta = 1 - 0,80 = 0,20$  e  $Z_{0,20} = 0,85$ . O valor do tamanho da amostra vale:

$$n = \frac{\left(1,65\sqrt{0,94(1-0,94)} + 0,85\sqrt{(0,94-0,03)(1-0,94+0,03)}\right)^2}{0,03^2} = 443,6.$$

Então, o tamanho da amostra necessário para detectar uma diferença de três pontos percentuais, caso exista, com poder de 80% é de 444 sementes, com nível de significância de 5%.

O cálculo do tamanho da amostra, para um *teste de hipótese unilateral à direita* ( $H_1 : \pi > \pi_0$ ), é dado pela expressão 8.5

$$n = \frac{Z_\alpha \sqrt{\pi_0(1-\pi_0)} + Z_\beta \sqrt{(\pi_0 + \delta)(1 - \pi_0 - \delta)}}{\delta^2}^2, \quad (8.5)$$

onde os elementos da fórmula são como em 8.4.

**Exemplo.** Um fabricante indica que os seus cigarros contém uma proporção de nicotina igual a 0,08. Um consumidor deseja verificar se tal afirmativa é verdadeira ou se a proporção de nicotina é maior do que o afirmado pelo fabricante. O tamanho da amostra necessário para detectar uma diferença de pelo menos dois pontos percentuais ( $\delta = 0,02$ ), caso exista, com nível de significância de 0,05 ( $Z_{0,05} = 1,65$ ) e poder do teste de 0,80 ( $Z_{0,20} = 0,85$ ) é dado por:

$$n = \frac{\left(1,65\sqrt{0,08(1-0,08)} + 0,85\sqrt{(0,08+0,02)(1-0,08-0,02)}\right)^2}{0,02^2} = 1.234,24.$$

Então, o tamanho da amostra necessário para detectar uma diferença de dois pontos percentuais, caso exista, com poder de 80% é de 1.235 cigarros, com nível de significância de 5%.

Para hipóteses bilaterais ( $H_1 : \pi \neq \pi_0$ ), a determinação do tamanho da amostra envolve os cálculos das expressões 8.4 e 8.5, com  $Z_\alpha$  substituído por  $Z_{\alpha/2}$ . O tamanho da amostra será o maior dos dois valores obtidos de  $n$ , dados por:

$$n_1 = \frac{Z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)} + Z_\beta\sqrt{(\pi_0-\delta)(1-\pi_0+\delta)}}{\delta^2}^2, \quad (8.6)$$

e

$$n_2 = \frac{Z_{\alpha/2}\sqrt{\pi_0(1-\pi_0)} + Z_\beta\sqrt{(\pi_0+\delta)(1-\pi_0-\delta)}}{\delta^2}^2. \quad (8.7)$$

**Exemplo.** Estudos têm indicado que a proporção de polinização das flores de maçã da variedade Fuji é igual a  $\pi_0 = 0,75$ . Um Agrônomo deseja verificar se esta proporção se mantém numa localidade diferente. Porém, ele não tem mais informações para afirmar que a proporção será maior ou menor, assim o teste de hipótese adequado é um teste bilateral. O tamanho da amostra necessário para detectar uma diferença de pelo menos três pontos percentuais ( $\delta = 0,03$ ), para mais ou para menos, caso exista, com nível de significância de 0,05 ( $Z_{0,05/2} = 1,96$ ) e poder do teste de 0,80 ( $Z_{0,20} = 0,85$ ) é dado por:

$$n_1 = \frac{\left(1,96\sqrt{0,75(1-0,75)} + 0,85\sqrt{(0,75-0,03)(1-0,75+0,03)}\right)^2}{0,03^2} = 1.682,08.$$

$$n_2 = \frac{\left(1,96\sqrt{0,75(1-0,75)} + 0,85\sqrt{(0,75+0,03)(1-0,75-0,03)}\right)^2}{0,03^2} = 1.602,17.$$



Tabela 8.3: Distribuição conjunta das frequências das variáveis época de plantio e sobrevivência de enxertos de ameixeiras

Época	Raízes		Total
	Sobreviventes	Mortas	
Fora Primavera	263	217	480
Na Primavera	115	365	480
Total	378	582	960

Tabela 8.4: Contagem de plantas segregando para dois caracteres numa progênie da espécie "X"

Ciclo	Virescência		Total
	Normal	Virescente	
Tardio	3470	910	4380
Precoce	1030	290	1320
Total	4500	1200	5700

Então, o tamanho da amostra necessário para detectar uma diferença de três pontos percentuais, caso exista, com poder de 80% é de 1.683 flôres, o maior valor entre 1.682,08 e 1.602,17, com nível de significância de 5%.

### 8.3.3 Teste de Duas ou Mais Proporções (Teste de Qui-Quadrado)

Nos trabalhos de pesquisa, é muito freqüente os dados representarem frequências (contagens) de ocorrências de determinada característica, que podem classificar-se em categorias de variáveis qualitativas. Os dados de frequências, classificados por categorias de variáveis qualitativas, podem ser representados por uma tabela de contingência, como as Tabelas 8.3 e 8.4. Os dados da Tabela 8.3 representam contagens de sobrevivência de enxertos de ameixeiras, provenientes de um experimento para comparar duas épocas de plantio: na primavera e fora da primavera. Para cada época de plantio foram transplantados 480 enxertos de ameixeiras e foi avaliado o número de enxertos sobreviventes após 30 dias. Os dados da Tabela 8.4 representam a contagem de plantas segregando para dois caracteres genéticos: ciclo e virescência, num total de 5700 plantas de uma progênie "X".

No primeiro experimento, os totais marginais das épocas foram pré-estabelecidos, e o

Tabela 8.5: Proporções populacionais numa tabela de contingência  $2 \times 2$

A	B		Total
	$B_1$	$B_2$	
$A_1$	$\pi_{11}$	$\pi_{12}$	$\pi_{1.}$
$A_2$	$\pi_{21}$	$\pi_{22}$	$\pi_{2.}$
Total	$\pi_{.1}$	$\pi_{.2}$	$\pi_{..}$

objetivo do pesquisador é verificar se as proporções de sobrevivência são as mesmas nas duas épocas de plantio. No segundo experimento, somente o total geral da tabela é que foi pré-estabelecido, e o objetivo do pesquisador é verificar se existe associação entre ciclo e virescência. Em ambos os casos, as hipóteses do pesquisador são respondidas através de testes de hipóteses envolvendo duas ou mais proporções. Na tabela de contingência 8.5 estão representadas as proporções populacionais correspondentes as quatro combinações das categorias das variáveis envolvidas nos estudos. onde  $\pi_{ij}$  é a proporção de elementos classificados na categoria  $i$  da variável  $A$  e na categoria  $j$  da variável  $B$ . Nas marginais estão representados totais dessas proporções. Por exemplo,  $\pi_{1.} = \pi_{11} + \pi_{12}$ .

No primeiro experimento, como os totais de linhas foram previamente fixados, temos duas distribuições binomiais, uma para cada época, fazendo com que  $\pi_{1.} = 1$  e  $\pi_{2.} = 1$ , e as hipóteses de interesse do pesquisador são dadas por:

$$H_0 : \pi_{1j} = \pi_{2j}$$

e

$$H_1 : \pi_{1j} \neq \pi_{2j}$$

com  $j = 1, 2$ . A hipótese nula ( $H_0$ ), indica que a proporção de sobreviventes na primavera é igual a proporção de sobreviventes fora da primavera. Este teste é conhecido na literatura como *teste de homogeneidade*.

No segundo experimento, como o total geral foi pré-fixado, temos uma única distribuição multinomial com quatro categorias fazendo com que  $\pi_{..} = 1$ , e as hipóteses de interesse do pesquisador são dadas por:

$$H_0 : \pi_{ij} = \pi_{i.}\pi_{.j}$$

e

$$H_1 : \pi_{ij} \neq \pi_{i.}\pi_{.j}$$

com  $i = 1, 2$  e  $j = 1, 2$ . A hipótese nula ( $H_0$ ), indica que a proporção de elementos classificados na categoria  $i$  da variável  $A$  e categoria  $j$  da variável  $B$  é igual ao produto das marginais dessas categorias. Este teste é conhecido na literatura como *teste de independência* entre as variáveis.

Na prática é comum encontrarmos situações que envolvem variáveis qualitativas com mais do que duas categorias. Por exemplo, poderíamos ter interesse em comparar as quatro estações do ano, ao invés de primavera e fora da primavera, e considerar mais de duas categorias de ciclo, tais como: tardio, normal e precoce. Assim, as hipóteses de homogeneidade e de independência podem ser generalizadas para  $s$  categorias da variável  $A$  e  $r$  categorias da variável  $B$ , ou seja,  $i = 1, 2, \dots, s$  e  $j = 1, 2, \dots, r$ .

Embora os experimentos tenham duas hipóteses diferentes, a estatística utilizada para testá-las é a mesma e é dada por:

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(n_{ij} - e_{ij})^2}{e_{ij}}, \quad (8.8)$$

onde  $n_{ij}$  é a frequência observada de elementos na categoria  $i$  da variável  $A$  e categoria  $j$  da variável  $B$  e  $e_{ij}$  é a frequência esperada de elementos nessas categorias e dada por:

$$e_{ij} = \frac{n_{i.} n_{.j}}{n_{..}}$$

com  $n_{i.}$ ,  $n_{.j}$  e  $n_{..}$  representando as frequências marginais e o total da tabela de contingência a ser analisada.

A estatística  $\chi^2$ , leia-se *Qui-Quadrado*, segue uma distribuição de qui-quadrado, sob a hipótese nula, com  $(s - 1) \times (r - 1)$  graus de liberdade, onde  $s$  é o número de linhas e  $r$  é o número de colunas da tabela de contingência. Esta distribuição teórica é dada no Apêndice 5, denominado Tabela da Distribuição Qui-Quadrado.

A hipótese nula é rejeitada se o valor da estatística  $\chi^2$ , calculada com os dados observados, for maior do que um valor crítico obtido da tabela da distribuição Qui-Quadrado, ao nível de significância  $\alpha$  e  $(s - 1) \times (r - 1)$  graus de liberdade, ou seja,  $\chi^2 > \chi_c^2$ . Os graus de liberdade estão representados na primeira coluna e os níveis de significância estão no cabeçalho da tabela. No exemplo da Tabela 8.3, o número de graus de liberdade é  $(2 - 1) \times (2 - 1) = 1$  e o valor crítico, para nível de significância  $\alpha = 0,05$ , é  $\chi_c^2 = 3,841$ . Este também é o valor crítico para o exemplo da Tabela 8.4.

Para o exemplo da Tabela 8.3, os valores observados no estudo foram:  $n_{11} = 263$ ,  $n_{12} = 217$ ,  $n_{21} = 115$  e  $n_{22} = 365$ ; os totais marginais são dados por:  $n_{1.} = 263 + 217 = 480$ ,

$n_{2.} = 115 + 365 = 480$ ,  $n_{.1} = 263 + 115 = 378$  e  $n_{.2} = 217 + 365 = 582$ ; o total geral é dado por:  $n_{..} = 263 + 217 + 115 + 365 = 960$ . Os valores esperados são dados por:

$$e_{11} = \frac{378 \times 480}{960} = 189, \quad e_{12} = \frac{582 \times 480}{960} = 291, \quad e_{21} = \frac{378 \times 480}{960} = 189 \text{ e } e_{22} = \frac{582 \times 480}{960} = 291.$$

O valor da estatística  $\chi^2$  é obtido de 8.8, com  $s = r = 2$  e é dado por:

$$\begin{aligned} \chi^2 &= \frac{(263 - 189)^2}{189} + \frac{(217 - 291)^2}{291} + \frac{(115 - 189)^2}{189} + \frac{(365 - 291)^2}{291} \\ &= \frac{(74)^2}{189} + \frac{(-74)^2}{291} + \frac{(-74)^2}{189} + \frac{(74)^2}{291} \\ &= 95,593. \end{aligned}$$

Portanto, como o valor calculado da estatística de  $\chi^2$  é maior do que o valor crítico do teste ( $95,593 > 3,841$ ), devemos rejeitar a hipótese nula e, assim, podemos afirmar que existe diferença significativa entre as duas épocas de plantio, ao nível de significância de 5%. Da Tabela 8.3 observamos que 54,8% dos enxertos plantados fora da primavera sobrevivem, enquanto que, 24,0% dos enxertos plantados na primavera sobrevivem, portanto, devemos preferir plantar fora da primavera.

Para o exemplo da Tabela 8.4, o valor da estatística  $\chi^2$  é obtido de 8.8, com  $s = r = 2$  e é dado por:

$$\begin{aligned} \chi^2 &= \frac{(3470 - 3457,90)^2}{3457,90} + \frac{(910 - 922,11)^2}{922,11} + \frac{(1030 - 1042,11)^2}{1042,11} + \frac{(290 - 277,90)^2}{277,90} \\ &= \frac{(12,1)^2}{3457,90} + \frac{(-12,1)^2}{922,11} + \frac{(-12,1)^2}{1042,11} + \frac{(12,1)^2}{277,90} \\ &= 0,0423 + 0,159 + 0,1405 + 0,5268 \\ &= 0,869. \end{aligned}$$

Portanto, como o valor calculado da estatística de  $\chi^2$  é menor do que o valor crítico do teste, não devemos rejeitar a hipótese nula e, assim, podemos afirmar que não existe associação entre ciclo e virescência, ao nível de significância de 5%. Da Tabela 8.4 observamos que dos materiais normais, 77,11% e 22,89% tardios e precoces, respectivamente. Para os materiais virescentes, observamos 75,83% e 24,17% para tardio e precoce, respectivamente. Então, observamos que as porcentagens nos dois ciclos são praticamente iguais para as categorias normal e virescente, portanto, as duas variáveis são independentes.

Como discutido na Seção 8.3.1, podemos utilizar o *valor p* para testar as hipóteses estatísticas de homogeneidade e independência. Nestes casos, o *valor p* é obtido a partir da

distribuição qui-quadrado, como sendo a probabilidade de se encontrar um valor maior ou igual ao valor de  $\chi^2$  calculado com os dados observados, representado por:

$$\text{valor } p = P(\chi_\nu^2 \geq \chi^2)$$

onde  $\nu$  representa os graus de liberdade. A hipótese nula é rejeitada quando o *valor p* for menor do que o nível de significância estabelecido pelo pesquisador. Os pacotes estatísticos fornecem diretamente o *valor p* dos testes.

Para o exemplo da Tabela 8.3, obtivemos um valor de  $\chi^2 = 95,593$ , com 1 grau de liberdade. Utilizando a tabela da distribuição qui-quadrado (Apêndice 5), podemos verificar que o *valor p* é menor do que 0,001. Para o exemplo da Tabela 8.4, obtivemos um valor de  $\chi^2 = 0,869$ , com 1 grau de liberdade, fornecendo um *valor p* maior do que 0,30. Para um nível de significância  $\alpha = 0,05$ , no primeiro exemplo a hipótese nula é rejeitada (*valor p* < 0,05), enquanto que no segundo exemplo a hipótese nula não é rejeitada (*valor p* > 0,05), conclusões idênticas aquelas obtidas utilizando-se o valor crítico.

### Restrições do Teste de Qui-Quadrado

Dado que a distribuição da estatística de qui-quadrado (8.8), neste caso, é uma distribuição aproximada de  $\chi^2$ , precisamos tomar certos cuidados na sua aplicação. Para tabelas  $2 \times 2$ , sugere-se aplicar uma correção, chamada *correção de continuidade ou correção de Yates*, que consiste em subtrair 0,5, das diferenças entre as freqüências observadas e esperadas, tomadas em módulo. Isto porque em tabelas  $2 \times 2$ , o teste de  $\chi^2$  tende a indicar mais significância do que deveria. Com esta correção, diminui-se o valor de  $\chi^2$  calculado, melhorando a aproximação com a distribuição teórica de  $\chi^2$ . A estatística de  $\chi^2$  com a correção de continuidade é dada por:

$$\chi^2 = \sum_{i=1}^s \sum_{j=1}^r \frac{(|n_{ij} - e_{ij}| - 0,5)^2}{e_{ij}}. \quad (8.9)$$

Esta correção não altera muito o valor da estatística  $\chi^2$  quando estamos trabalhando com amostras grandes, caso do exemplo 8.3, no qual temos:

$$\begin{aligned} \chi^2 &= \frac{(74 - 0,5)^2}{189} + \frac{(74 - 0,5)^2}{291} + \frac{(74 - 0,5)^2}{189} + \frac{(74 - 0,5)^2}{291} \\ &= 94,29. \end{aligned}$$

um valor bem próximo daquele obtido sem a correção.

Para situações onde o número de categorias das variáveis envolvidas é maior do que 2, é bastante comum encontrar-se caselas com valores de freqüências esperadas inferiores a

Tabela 8.6: Segregação mendeliana em ervilhas

Tipos de ervilhas	Frequências observadas	Frequências esperadas sob $H_0$
Amarelas lisas	315	$312,75 = 556 \times (9/16)$
Verdes lisas	108	$104,25 = 556 \times (3/16)$
Amarelas rugosas	101	$104,25 = 556 \times (3/16)$
Verdes rugosas	32	$34,75 = 556 \times (1/16)$
Total	556	556,00

5. Aumentar o tamanho da amostra é uma solução. A outra solução é agrupar uma ou mais linhas (ou colunas), a fim de obter frequências esperadas superiores a 5, tomando-se o cuidado de não prejudicar os objetivos do trabalho.

Uma alternativa para o uso da estatística  $\chi^2$ , quando temos amostras pequenas, seria usar o *Teste Exato de Fisher* (Siegel, 1975), o qual está implementado em programas estatísticos computacionais.

#### 8.3.4 Teste de Aderência

Uma aplicação importante do teste de  $\chi^2$  é verificar se frequências observadas seguem (estão de acordo) as proporções previstas por teorias ou hipóteses. Um exemplo clássico, é o caso da segregação mendeliana em ervilhas. Um outro exemplo seria verificar se os resultados observados seguem uma dada distribuição de probabilidade, como em estudos de dispersão de espécies em determinada área. Nestes estudos desejamos verificar se uma espécie se distribui aleatoriamente numa determinada área, ou seja, se os resultados seguem uma distribuição de Poisson.

### *Aplicação à Teoria Mendeliana*

Na descendência obtida de cruzamentos entre plantas com sementes amarelas lisas e outras de sementes verdes rugosas, obtemos na primeira geração ervilhas amarelas lisas. Estas, cruzadas entre si, produzem na segunda geração ervilhas de quatro tipos indicadas na Tabela 8.6.

Pela teoria mendeliana, as proporções esperadas para esses tipos de ervilhas são: 9/16, 3/16, 3/16 e 1/16, respectivamente. Com base nessa teoria, calculamos as frequências esperadas dadas na Tabela 8.6. Desejamos testar as seguintes hipóteses através do teste de qui-quadrado:

- $H_0$  : os dados *concordam* com a teoria mendeliana
- $H_1$  : os dados *não concordam* com a teoria mendeliana.

A estatística de  $\chi^2$  vale,

$$\chi^2 = \frac{(315 - 312,75)^2}{312,75} + \frac{(108 - 104,25)^2}{104,25} + \frac{(101 - 104,25)^2}{104,25} + \frac{(32 - 34,75)^2}{34,75} = 0,47.$$

O valor de  $\chi^2$  é bastante baixo, sugerindo que não devemos rejeitar a hipótese nula. Entretanto, devemos encontrar o *valor p* do teste, o qual vale,

$$P(\chi^2_3 > 0,47) = 0,925431 \text{ ou } 92,54\%.$$

Este valor é muito superior a 5%, portanto, podemos afirmar que os dados concordam com a teoria mendeliana, ao nível de significância de 5%.

Em termos do valor crítico, ao nível de significância de 5% e com 3 graus de liberdade, temos:

$$\chi^2_c = 7,815.$$

Como o valor calculado de  $\chi^2$  é menor do que o valor da distribuição de  $\chi^2_c$ , não rejeitamos  $H_0$ . Portanto, as frequências observadas seguem as proporções estabelecidas por Mendel.

### *Aplicação à Dispersão Espacial*

Os dados da Tabela 8.7 representam o número de árvores por quadrante da espécie *Guapira opposita*, obtidos de um estudo realizado com o objetivo de verificar a distribuição espacial desta espécie num local de restinga. Foram utilizados um total de 94 quadrantes e contou-se o número de quadrantes com zero árvores, uma árvore, duas árvores, e assim por diante. Na última categoria estão contabilizados todos os quadrantes que apresentaram 9 ou mais árvores. A hipótese nula ( $H_0$ ) é que a espécie se distribui aleatoriamente na região, ou seja, o número de quadrantes segue uma distribuição de Poisson com média 2,9 árvores/quadrante, e a hipótese alternativa ( $H_1$ ) é que a espécie não se distribui aleatoriamente na região. O valor da média corresponde a média do número de árvores ponderado pelo número de quadrantes. Utilizamos o valor 9 para a última categoria da tabela.

O valor da estatística para o teste é dado por:

$$\begin{aligned} \chi^2 &= \frac{(6 - 5,170)^2}{5,170} + \frac{(18 - 15,040)^2}{15,040} + \frac{(23 - 21,714)^2}{21,714} + \frac{(19 - 21,056)^2}{21,056} \\ &+ \frac{(11 - 15,228)^2}{15,228} + \frac{(6 - 8,836)^2}{8,836} + \frac{(5 - 4,230)^2}{4,230} + \frac{(4 - 1,786)^2}{1,786} \\ &+ \frac{(1 - 0,658)^2}{0,658} + \frac{(1 - 0,188)^2}{0,188} = 9,646 \end{aligned}$$

Tabela 8.7: Número de árvores por quadrante da espécie *Guapira opposita*

$X$ árvores por quadrante	Número de quadrantes com $X$ árvores	Probabilidades da Poisson	Número esperado de quadrantes com $X$ árvores
0	6	0.055	5.170
1	18	0.160	15.040
2	23	0.231	21.714
3	19	0.224	21.056
4	11	0.162	15.228
5	6	0.094	8.836
6	5	0.045	4.230
7	4	0.019	1.786
8	1	0.007	0.658
9 ou mais	1	0.002 <sup>1</sup>	0.188

<sup>1</sup>Calculada para  $k = 9$ 

Devemos encontrar o *valor p* do teste, o qual vale,

$$P(\chi_9^2 > 9,646) > 0,30 \text{ ou } > 30\%.$$

Este valor é muito superior a 5%, portanto, podemos afirmar que a distribuição da espécie é aleatória, ao nível de significância de 5%.

Em termos do valor crítico, ao nível de significância de 5% e com 9 graus de liberdade, temos:

$$\chi_c^2 = 16,919.$$

Como o valor calculado de  $\chi^2$  é menor do que o valor da distribuição de  $\chi_c^2$ , não rejeitamos  $H_0$ . Portanto, as frequências observadas são próximas das estabelecidas pelo modelo de Poisson.

### 8.3.5 Tamanho da Amostra para Duas Proporções Populacionais

Para situações nas quais desejamos comparar duas proporções populacionais a partir de amostras independentes, o que corresponderia a tabelas de contingência  $2 \times 2$  com os totais marginais fixados (Teste de homogeneidade), o tamanho da amostra, para cada uma das amostras independentes, pode ser determinado utilizando-se expressões semelhantes a (8.4) - (8.7).



Para testes de hipóteses unilaterais, o tamanho da amostra é dado por:

$$n = \frac{(Z_\alpha + Z_\beta)^2 [\pi_{11}(1 - \pi_{11}) + \pi_{21}(1 - \pi_{21})]}{\delta^2}, \quad (8.10)$$

onde  $Z_\alpha$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $\alpha$ ,  $Z_\beta$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $1 - \text{poder do teste} = \beta$ ,  $\pi_{11}$  e  $\pi_{21}$  são definidos como na Tabela 8.5, com os totais marginais da variável  $A$  fixados, e  $\delta$  é a diferença entre estas duas proporções a ser detectada.

A utilização da expressão 8.10 requer, por parte do pesquisador, a especificação dos valores para  $\alpha$  e  $\beta$ , discutidos na Seção 8.3.2, e para  $\pi_{11}$ ,  $\pi_{21}$  e  $\delta$ . Como  $\delta$  representa a diferença entre estas duas proporções, basta que o pesquisador especifique somente uma das duas proporções e  $\delta$ .

**Exemplo.** Um pesquisador deseja realizar um experimento com o objetivo de comparar duas formas de corte, longo (população 1) e curto (população 2), de enxertos de macieiras com relação à sobrevivência dos mesmos. Qual deveria ser o tamanho da amostra para detectar uma diferença, caso exista, de pelo menos 10 pontos percentuais, isto é,  $\delta = 0,10$ , ao nível de significância de  $\alpha = 0,05$  e poder de 0,80, isto é,  $\beta = 0,20$ ? De acordo com conhecimentos, ele sabe que a proporção de sobrevivência de enxertos com a forma de corte longo é da ordem de  $\pi_{11} = 0,60$  e que, caso haja diferença entre estas proporções, a proporção com a forma de corte longo é maior do que a proporção com a forma de corte curto. Assim,  $Z_\alpha = 1,65$ ,  $Z_\beta = 0,85$  e  $\pi_{21} = \pi_{11} - \delta = 0,60 - 0,10 = 0,50$ , fazendo com que:

$$n = \frac{(1,65 + 0,85)^2 [0,60(1 - 0,60) + 0,50(1 - 0,50)]}{0,10^2} = 306,25.$$

Portanto, o tamanho da amostra necessário para detectar uma diferença de 10 pontos percentuais, caso exista, com poder de 80%, é de 307 enxertos para cada tipo de corte, com nível de significância de 5%.

Nas situações em que o pesquisador não tem conhecimento a respeito do valor de nenhuma das proporções, ele deverá utilizar o valor 0,50 para uma das proporções e  $0,50 - \delta$  ou  $0,50 + \delta$  para a outra proporção, dependendo da hipótese alternativa unilateral.

Para testes de hipóteses bilaterais,  $H_1 : \pi_{11} \neq \pi_{21}$ , o tamanho da amostra será o maior

valor dos dois valores obtidos de  $n$ , dados por:

$$n_1 = \frac{(Z_{\alpha/2} + Z_{\beta})^2 [\pi_{11}(1 - \pi_{11}) + (\pi_{11} - \delta)(1 - \pi_{11} + \delta)]}{\delta^2}, \quad (8.11)$$

$$n_2 = \frac{(Z_{\alpha/2} + Z_{\beta})^2 [\pi_{11}(1 - \pi_{11}) + (\pi_{11} + \delta)(1 - \pi_{11} - \delta)]}{\delta^2}, \quad (8.12)$$

onde os termos destas expressões são como definidos em 8.10.

**Exemplo.** Um pesquisador deseja realizar um experimento com o objetivo de comparar duas cultivares de cebola, Bola Precoce (população 1) e Norte 14 (população 2), quanto a germinação de sementes. Qual deveria ser o tamanho da amostra para detectar uma diferença, caso exista, de pelo menos 5 pontos percentuais, isto é,  $\delta = 0,05$ , ao nível de significância de  $\alpha = 0,05$  e poder de 0,80, isto é,  $\beta = 0,20$ ? De acordo com conhecimentos, ele sabe que a proporção de germinação de sementes da cultivar Bola Precoce é da ordem de  $\pi_{11} = 0,90$  e que, caso haja diferença entre estas proporções, qualquer uma delas pode apresentar ser superior. Assim,  $Z_{\alpha/2} = 1,96$ ,  $Z_{\beta} = 0,85$  fazendo com que:

$$n_1 = \frac{(1,96 + 0,85)^2 [0,90(1 - 0,90) + (0,90 - 0,05)(1 - 0,90 + 0,05)]}{0,05^2} = 686,96$$

e

$$n_2 = \frac{(1,96 + 0,85)^2 [0,90(1 - 0,90) + (0,90 + 0,05)(1 - 0,90 - 0,05)]}{0,05^2} = 434,28.$$

Portanto, o tamanho da amostra necessário para detectar uma diferença de 5 pontos percentuais, para mais ou para menos, caso exista, com poder de 80%, é de 687 sementes para cada uma das cultivares, com nível de significância de 5%.

A determinação do tamanho da amostra para as outras situações apresentadas nesta seção, está fora da abrangência deste livro.

## 8.4 Teste para Médias Populacionais

O objetivo de testar-se hipóteses sobre médias verdadeiras é avaliar certas afirmações feitas sobre as mesmas. Por exemplo, podemos desejar verificar a afirmação de que as alturas médias de plantas de feijão, para sementes de alto e baixo vigor, são iguais.

Existem, basicamente, três tipos de afirmações que se podem fazer quando se estuda médias populacionais, quais sejam:

1. a afirmação diz respeito a uma média populacional, então, temos o teste de uma média populacional. Exemplo, o peso médio ao nascer de bezerros da raça Nelore, no planalto Catarinense, em agosto, é de 25,5 kg;
2. a afirmação diz que as médias de duas populações (dois tratamentos) são iguais, temos, então, um teste de comparação de duas médias. Exemplos: 1) as produções médias de batatinhas de duas variedades são iguais e 2) as áreas foliares específicas médias da espécie *Cecropia glaziovii*, cujo nome vulgar é embaúba ou embaúva, situadas na borda da mata e na mata fechada são iguais.
3. a afirmação diz que as médias de mais de duas populações (mais do que dois tratamentos) são todas iguais, temos, então, o teste de comparação de  $k$  médias, com  $k > 2$ . Por exemplo, desejamos saber se há diferenças entre três locais (Baía Norte, Baía Sul e Pântano do Sul), quanto ao número médio de micronúcleos por 5000 células sanguíneas de peixes do gênero bagre. Neste caso, devemos fazer uma Análise de Variância. Neste livro somente serão considerados testes para uma e duas médias. Existem diversos livros especializados em planejamento e análise de experimentos, que tratam desse tipo de análise, podemos citar, os livros de Vieira (1999) e Steel & Torrie (1960).

As técnicas que nós vamos estudar pressupõem uma distribuição normal da distribuição amostral da estatística ou estimador  $(\bar{X}, \bar{X}_1 - \bar{X}_2)$ . Como sabemos, essa suposição será válida se a distribuição da variável em estudo seguir uma distribuição normal e a amostragem for aleatória e, com boa aproximação, para outras distribuições se a amostra for suficientemente grande (sugestão:  $n \geq 30$ ).

#### 8.4.1 Teste para Uma Média Populacional

Para o teste de uma média, a hipótese nula é dada por:

$$H_0 : \mu = \mu_0.$$

Com conhecimento a respeito do assunto em estudo, o pesquisador irá selecionar uma das seguintes hipóteses alternativas:

$$H_1 : \mu \neq \mu_0 \text{ (teste bilateral)}$$

$$H_1 : \mu > \mu_0 \text{ (teste unilateral à direita)}$$

$$H_1 : \mu < \mu_0 \text{ (teste unilateral à esquerda)}$$

A estatística apropriada para o teste de hipóteses sobre uma média populacional é baseada na média amostral e sua distribuição depende da variância populacional ser conhecida ou não, conforme já discutido na Seção 7.4.3.1.

#### 8.4.1.1 Teste para Uma Média Populacional com Variância Conhecida

Sabemos que a distribuição amostral da média ( $\bar{X}$ ), de uma amostra aleatória simples de  $n$  elementos de uma população com média  $\mu$  e variância  $\sigma^2$ , pode ser, em geral, representada pela distribuição normal com média  $\mu$  e variância  $\sigma^2/n$ . Assim, quando a hipótese nula,  $\mu = \mu_0$ , for verdadeira, a estatística,

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}, \quad (8.13)$$

segue a distribuição normal padrão com média zero e variância um. Valores grandes positivos ou grandes negativos indicariam que a hipótese nula é falsa, pois o valor obtido para a média amostral foi muito diferente do valor  $\mu_0$  afirmado como sendo a média verdadeira. Num teste de hipóteses feito através do *valor p*, um valor grande positivo ou negativo da estatística 8.13, estará associado a um *valor p* pequeno. Na prática o *valor p* é calculado como segue:

- *valor p* =  $P(Z < Z_{obs})$  para  $H_1 : \mu < \mu_0$ ;
- *valor p* =  $P(Z > Z_{obs})$  para  $H_1 : \mu > \mu_0$ ;
- *valor p* =  $2 \times P(Z > |Z_{obs}|)$  para  $H_1 : \mu \neq \mu_0$ ,

onde  $Z_{obs}$  é o valor observado da estatística dada por (8.13) e a probabilidade pode ser obtida em tabelas da distribuição normal padrão (Apêndice 3), ou através de programas computacionais estatísticos. A hipótese nula será rejeitada quando o *valor p* for menor do que  $\alpha$ , o nível de significância do teste.

**Exemplo.** Uma balança para encher pacotes de sementes automaticamente, está programada para produzir pacotes com peso médio de 20 kg e desvio padrão de 0,20 kg. Periodicamente é feita uma inspeção para verificar se o peso médio está sob controle. Para este fim, foi selecionada uma amostra de 8 pacotes de sementes, cujos resultados foram:

20,3	19,8	20,3	19,7	19,8	19,7	19,8	19,8
------	------	------	------	------	------	------	------

Teste a hipótese de que a balança se desregulou e está produzindo um peso médio inferior a 20 kg. Use nível de significância de 5%.

As hipóteses são dadas por:

$$H_0 : \mu = 20 \quad H_1 : \mu < 20$$

Através dos dados da amostra obtemos:  $\bar{x} = 19,9 \text{ kg}$ . Estes resultados aplicados na expressão 8.13 fornece:

$$Z = \frac{19,9 - 20}{\frac{0,20}{\sqrt{8}}} = -1,41.$$

O *valor p* do teste, utilizando a tabela da distribuição normal padrão (Apêndice 3), vale:

$$\text{valor } p = P(Z < -0,01) = 0,50000 - 0,4207 = 0,0793.$$

Como o *valor p* é maior do que o nível de significância, não rejeitamos a hipótese nula.

Como vimos para o caso do teste de uma proporção populacional, aqui também podemos estabelecer um critério para rejeitar a hipótese nula. Este critério é estabelecido a partir da fixação do nível de significância ( $\alpha$ ) do teste e, como sabemos, para  $\alpha = 0,05$  e uma hipótese alternativa unilateral, obtemos da tabela da distribuição normal padrão,  $Z = -1,65$ . Assim, o valor crítico do teste é obtido a partir de 8.13, e dado por:

$$\begin{aligned} -1,65 &= \frac{\bar{X}_c - 20}{\frac{0,20}{\sqrt{8}}} \\ \bar{X}_c &= 19,88. \end{aligned} \tag{8.14}$$

Portanto, se a média encontrada a partir dos dados, for inferior a 19,88 kg, rejeitamos a hipótese nula ao nível de significância de 5%. No nosso exemplo, a média encontrada vale 19,90 kg, valor este superior ao valor crítico, portanto não rejeitamos a hipótese nula ao nível de significância de 5%.

#### 8.4.1.2 Teste para Uma Média Populacional com Variância Desconhecida

Como já foi dito no estudo dos intervalos de confiança, é comum na prática não conhecermos o valor da variância populacional ( $\sigma^2$ ). Então, devemos estimá-la a partir dos valores obtidos na amostra, ou seja, através da variância amostral ( $S^2$ ).

Neste caso, a estatística apropriada para o teste de hipóteses é obtida de (8.13) com  $\sigma$  substituído por  $S$ , o desvio padrão amostral, e dada por:

$$t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}, \tag{8.15}$$

a qual tem distribuição  $t$  de Student com  $n - 1$  graus de liberdade, quando a hipótese nula for verdadeira e a distribuição da variável em estudo, representada pela letra  $X$ , for normal com média  $\mu$  e variância  $\sigma^2$ .

Como no caso da estatística (8.13) valores grandes positivos ou grandes negativos de  $t$  indicariam que a hipótese nula é falsa, pois o valor obtido para a média amostral foi muito diferente do valor  $\mu_0$  afirmado como sendo média verdadeira. Num teste de hipóteses feito através do *valor p*, um valor grande positivo ou negativo da estatística (8.15), estará associado a um *valor p* pequeno. Na prática o *valor p* é calculado como segue:

- *valor p* =  $P(t_{n-1} < t_{obs})$  para  $H_1 : \mu < \mu_0$ ;
- *valor p* =  $P(t_{n-1} > t_{obs})$  para  $H_1 : \mu > \mu_0$ ;
- *valor p* =  $2 \times P(t_{n-1} > |t_{obs}|)$  para  $H_1 : \mu \neq \mu_0$ ,

onde  $t_{obs}$  é o valor observado da estatística dada por (8.15) e a probabilidade pode ser obtida em tabelas da distribuição  $t$  de Student (Apêndice 4), ou através de programas computacionais estatísticos. A hipótese nula será rejeitada quando o *valor p* for menor do que  $\alpha$ , o nível de significância do teste.

Vejamos agora a aplicação do teste de hipótese de que uma média populacional,  $\mu$ , é igual a um valor fixo,  $\mu_0$ , supondo-se que a variância populacional,  $\sigma^2$ , seja desconhecida.

**Exemplo 1.** Um agrônomo afirma que a produtividade média do feijão da safra das lavouras de agricultores familiares de um determinado ano é de 800 kg/ha. Para investigar a veracidade dessa afirmação, selecionou-se uma amostra de nove lavouras onde obteve-se os seguintes valores de produtividade de feijão:

Lavoura	1	2	3	4	5	6	7	8	9
Produtividade	767,8	764,1	716,8	750,2	756,0	692,5	736,1	746,1	731,4

a) Qual a conclusão ao nível de significância de 5%?

b) Dê a estimativa da média verdadeira, com confiança de 95%.

A nossa variável em estudo é a produtividade de feijão em kg/ha e é bem razoável assumir que a distribuição da variável é normal.

Como não temos informações adicionais se a produtividade média é maior ou menor do que 800 kg, vamos construir as hipóteses do seguinte modo:

$$H_0 : \mu = 800 \text{ kg/ha} \quad \text{versus} \quad H_1 : \mu \neq 800 \text{ kg/ha}.$$

A amostra nos fornece,  $\bar{x} = 740,11$  kg/ha e  $s = 24,07$  kg/ha. Como vamos fazer um teste sobre uma média populacional, com variância estimada com os dados de uma amostra, usamos a estatística 8.15, a qual segue uma distribuição  $t$  de Student com  $9 - 1 = 8$  graus de liberdade, sob a hipótese nula.

Com os valores da amostra vamos encontrar o valor da estatística teste,

$$t = \frac{740,11 - 800}{\frac{24,07}{\sqrt{9}}} = \frac{-59,89}{8,02} = -7,468.$$

O *valor p* do teste é dado por:

$$\text{valor } p = 2 \times P(t_8 > |-7,468|)$$

Se esse valor for igual ou maior do que 5% devemos aceitar a hipótese nula, caso contrário, rejeitamos.

Com o uso da tabela da distribuição *t* de Student (Apêndice 4), obtemos:

$$P(t_8 > |-7,468|) < 0,001,$$

indicando um *valor p* menor do que 0,002. Assim, rejeitamos a hipótese nula.

Podemos encontrar o exato *valor p* usando um programa computacional estatístico que disponibilize o cálculo da distribuição *t* de Student. No nosso exemplo, obtemos:

$$P(t_8 > |-7,468|) = 0,000036,$$

assim, o *valor p* do teste vale 0,000072 (= 2 × 0,000036). Portanto, rejeitamos a hipótese nula.

O intervalo de confiança, para a média verdadeira, fica:

$$\begin{aligned} IC(\mu; 95\%) &: 740,11 \pm 2,306 \frac{24,07}{\sqrt{9}} \\ &: 740,11 \pm 18,5 \\ &[721,61; 758,61]. \end{aligned}$$

**Exemplo 2.** Foi retirada uma amostra de 10 bezerros da raça Nelore, aos 210 dias de idade, com o objetivo de verificar se o peso médio dos animais atingiu ou não 186 kg. Os valores obtidos, em kg, foram os seguintes:

178	199	182	186	188	191	189	185	174	158
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Teste as hipóteses:

$$H_0 : \mu = 186 \text{ vs } H_1 : \mu < 186,$$

ao nível de significância de 5%.

A variável em estudo é do tipo contínua e vamos assumir que a mesma tenha, pelo menos aproximadamente, distribuição normal. Na amostra obtemos  $\bar{x} = 183 \text{ kg}$  e  $s = 11,18 \text{ kg}$ . Estes resultados aplicados em 8.15 fornecem:

$$t_9 = \frac{183 - 186}{\frac{11,18}{\sqrt{10}}} = -0,847$$

Com o auxílio da tabela da distribuição  $t$  de student, obtemos:

$$P(t_9 < -0,847) > 0,15.$$

Como 0,30 é maior do que os 5% definido pelo pesquisador, não devemos rejeitar a hipótese nula.

Podemos encontrar o exato *valor p* do teste usando um programa computacional estatístico que disponibilize o cálculo da distribuição  $t$  de Student. No nosso exemplo, obtemos:

$$P(t_9 < -0,847) = 0,2095.$$

Portanto, não rejeitamos a hipótese nula.

### 8.4.1.3 Poder do Teste e Tamanho da Amostra para uma Média Populacional

Quando não rejeitamos a hipótese nula, podemos estar incorrendo no erro tipo II, que corresponde a aceitação de uma hipótese nula falsa. No exemplo da balança para encher pacotes de sementes automaticamente, ela pode estar desregulada e produzindo pacotes de sementes com peso médio igual a, por exemplo, 19,96 kg, ao invés do peso médio de 20 kg, como programado. A probabilidade de não detectarmos isso é dada por:

$$\beta = P(\bar{X} \geq 19,88 | \mu = 19,96) = P\left(Z \geq \frac{19,88 - 19,96}{0,2/\sqrt{8}}\right) = P(Z \geq -1,13) = 0,8708,$$

onde 19,88 é o valor crítico do teste para um nível de significância de 5%, calculado em 8.14. O valor encontrado para a probabilidade do erro tipo II, indica que temos aproximadamente 87% de chance de não detectarmos uma redução de 40 gramas no peso médio dos pacotes de sementes. Se essa redução de 40 gramas for um valor dentro dos padrões de controle, esse valor alto para  $\beta$  não significa que o teste realizado com  $n = 8$  observações seja ruim.

Vamos supor agora uma redução de 100 gramas no peso médio de pacotes, ou seja, um peso médio verdadeiro de 19,90 kg, ao invés de 20,00 Kg. Nesta situação, a probabilidade de não detectarmos isso é dada por:

$$\beta = P(\bar{X} \geq 19,88 | \mu = 19,90) = P\left(Z \geq \frac{19,88 - 19,90}{0,2/\sqrt{8}}\right) = P(Z \geq -0,28) = 0,6103.$$

O valor encontrado para a probabilidade do erro tipo II, indica que temos aproximadamente 61% de chance de não detectarmos uma redução indesejada de 100 gramas no peso médio dos pacotes de sementes. Esse valor alto de  $\beta$  indica que o teste realizado com  $n = 8$  observações não é apropriado.



Tabela 8.8: Poder do teste para diversos valores de médias populacionais, para amostras de tamanhos  $n = 8$  e  $n = 25$

Média populacional ( $\mu$ ) (em kg)	Diferença a detectar( $\delta$ ) (em gramas)	Poder ( $n = 8$ )	Poder ( $n = 25$ )
19,98	20	0,08	0,11
19,96	40	0,13	0,23
19,94	60	0,20	0,40
19,92	80	0,29	0,60
19,90	100	0,39	0,77
19,88	120	0,50	0,89
19,86	140	0,61	0,96
19,84	160	0,71	0,99

Na prática nós gostaríamos de realizar um teste que nos garantisse uma probabilidade do erro do tipo II de no máximo 0,20, que corresponde a um poder de teste,  $1 - \beta$ , de no mínimo 0,80, para detectar uma diferença relevante para o problema em estudo, conforme referido na Seção 8.3.2.

A Tabela 8.8 apresenta valores do poder do teste (para  $n = 8$ ) para diferentes valores do verdadeiro peso médio( $\mu$ ), sabendo que a hipótese nula deveria ser rejeitada caso a média encontrada na amostra fosse inferior a 19,88 kg.

Da Tabela 8.8 podemos concluir que o poder do teste aumenta quanto maior for a diferença entre o peso médio programado ( $\mu = 20$ ) e o verdadeiro valor do peso médio dos pacotes de sementes. Por exemplo, temos somente 13% de chance de detectar uma diferença de 40 gramas,  $\delta = 40$ , conforme cálculo apresentado. Mesmo para detectar uma diferença de 100 gramas,  $\delta = 100$ , a chance seria só de 39%. Em outras palavras, este teste, baseado em uma amostra de 8 pacotes, não fornece uma boa garantia para detectar que a balança está desregulada. É necessário uma amostra de tamanho maior.

Vamos trabalhar agora com uma amostra de tamanho 25 pacotes de sementes para verificar como ficam os valores do poder do teste. Para um nível de significância fixado em 5%, o valor crítico do teste( $\bar{X}_c$ ) é obtido conforme 8.14 e seu valor é:

$$\begin{aligned} -1,65 &= \frac{\bar{X}_c - 20}{0,20/\sqrt{25}} \\ \bar{X}_c &= 19,93. \end{aligned}$$

A região de rejeição do teste é formada por todos os valores menores a 19,93, ou seja, vamos

rejeitar a hipótese nula ( $\mu = 20$ ), quando a média encontrada na amostra for menor do que 19,93, ao nível de significância de 5%. Este é o critério do nosso teste, baseado numa amostra de tamanho  $n = 25$  pacotes. Os valores do poder do teste, isto é, as probabilidades de rejeitar a hipótese nula quando ela é falsa, estão dadas na Tabela 8.8. Por exemplo, o valor do poder do teste, para o primeiro caso, é calculado como:

$$Poder = P(\bar{X} < 19,93 | \mu = 19,98) = P\left(Z < \frac{19,93 - 19,98}{0,2/\sqrt{25}}\right) = P(Z < -1,25) = 0,11.$$

Comparando os valores das duas últimas colunas da Tabela 8.8, verificamos que o poder do teste para detectar uma diferença de 120 gramas, aumenta de 0,50 para 0,89. Isto significa que uma diferença dessa magnitude tem uma grande chance de ser detectada, o que não ocorreria com uma amostra de 8 pacotes. Por outro lado, mesmo com uma amostra de 25 pacotes a chance de detectar uma diferença de 80 gramas não seria muito grande (0,60). Faz-se necessário a determinação do tamanho apropriado da amostra para detectar uma diferença  $\delta$ , de interesse do pesquisador, com alto poder do teste e nível de significância  $\alpha$ .

O cálculo do tamanho da amostra, para um *teste de hipótese unilateral* ( $H_1 : \pi < \pi_0$  ou  $H_1 : \pi > \pi_0$ ), é dado pela expressão 8.16

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\delta^2}, \quad (8.16)$$

onde  $Z_\alpha$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $\alpha$ ,  $Z_\beta$  é o valor da distribuição normal padrão que corresponde ao valor de  $Z$  acima do qual a área é igual a  $1 - poder = \beta$  e  $\delta$  é a diferença a ser detectada.

**Exemplo.** O tamanho da amostra, isto é, o número de pacotes que deveria ser selecionado para detectarmos uma redução de pelo menos 50 gramas no peso médio dos pacotes, caso tenha ocorrido, ao nível de significância de 5% ( $Z_\alpha = 1,65$ ) e poder do teste de 0,80 ( $Z_\beta = 0,85$ ), é dado por:

$$n = \frac{(1,65 + 0,85)^2 0,20^2}{0,05^2} = 100.$$

Para hipóteses bilaterais ( $H_1 : \mu \neq \mu_0$ ), a determinação do tamanho da amostra é obtida de 8.16 com  $Z_\alpha$  substituído por  $Z_{\alpha/2}$ . O tamanho da amostra será dado por:

$$n = \frac{(Z_{\alpha/2} + Z_\beta)^2 \sigma^2}{\delta^2}, \quad (8.17)$$

**Exemplo.** No exemplo anterior, considerando que pode ter ocorrido uma redução ou um aumento no peso médio dos pacotes, de pelo menos 50 gramas no peso médio dos pacotes, caso tenha ocorrido, o tamanho da amostra necessário para a detecção de tal desvio, ao nível de significância de 5% ( $Z_{\alpha/2} = 1,96$ ) e poder do teste de 0,80 ( $Z_{\beta} = 0,85$ ), é dado por:

$$n = \frac{(1,96 + 0,85)^2 0,20^2}{0,05^2} = 126,34.$$

Neste caso, devemos selecionar uma amostra de 127 pacotes.

As expressões 8.16 e 8.17 dependem do conhecimento do verdadeiro valor do desvio padrão ( $\sigma$ ). Em muitas situações práticas, esse valor é desconhecido e uma solução é a utilização do desvio padrão ( $S'$ ) de uma amostra piloto de  $n'$  elementos nestas expressões. Se, após a obtenção da amostra de tamanho  $n$ , for obtido um desvio padrão amostral muito superior ao desvio padrão da amostra piloto, novo tamanho da amostra deve ser calculado e a amostra deverá ser completada. Caso a amostra não seja completada, os resultados da pesquisa terão um poder do teste inferior ao estabelecido para o cálculo de  $n$ .

#### 8.4.2 Teste de Duas Médias Populacionais com Variâncias Populacionais Desconhecidas

Vamos, nesta seção, estender o procedimento anterior para o caso de comparação de duas médias populacionais, quando as variâncias populacionais são desconhecidas. A fundamentação básica continua sendo a mesma, só se farão algumas alterações quanto à estatística teste a ser utilizada.

Inicialmente, vamos supor que temos duas amostras selecionadas de populações normais (variável em estudo presente, pelo menos aproximadamente, uma distribuição normal). A questão da normalidade é especialmente importante quando o tamanho da amostra é pequeno. Para cada amostra, vamos calcular a média e a variância dos dados.

O objetivo do teste, neste caso, é decidir se as médias de duas populações são iguais, ou seja,

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

Como no caso do teste de uma média populacional, temos três possibilidades para a hipótese alternativa, quais sejam,

$$H_1 : \mu_1 \neq \mu_2 \quad (\text{teste bilateral})$$

$$H_1 : \mu_1 > \mu_2 \quad (\text{teste unilateral à direita})$$

$$H_1 : \mu_1 < \mu_2 \quad (\text{teste unilateral à esquerda}).$$

Num teste de comparação de duas médias, temos dois casos a considerar, quais sejam:

1. **dados pareados ou amostras dependentes.** Os dados de duas amostras constituem dados pareados quando estão relacionados dois a dois, segundo algum critério que introduz uma influência marcante entre os diversos pares de valores. Também é importante observar que deve haver independência entre os pares.

Exemplo: vamos submeter seis animais de determinada espécie a uma nova dieta. Afim de determinar o efeito da dieta foram tomados os pesos *antes* e *depois* da dieta. Nesse caso, as observações deverão ser feitas nos mesmos animais, medindo uma característica antes e depois deles serem submetidos a uma dieta. Outro exemplo: desejamos comparar a capacidade auditiva dos ouvidos direito e esquerdo de pessoas acima de 50 anos de idade. Neste caso as observações dos ouvidos direito e esquerdo deverão ser realizadas na mesma pessoa.

2. **dados não pareados ou amostras independentes.** Neste caso, os dados das duas amostras não estão relacionados por nenhum fator que possa confundir (misturar) o efeito do fator em estudo. Além disso deve haver independência entre observações dentro das amostras.

Exemplo: vamos fazer um estudo com o objetivo de verificar se o comprimento do antebraço de morcegos de duas espécies são iguais. Para isso, tomamos uma amostra de morcegos da espécie A e uma outra amostra de morcegos da espécie B. Podemos perceber que os dados das duas amostras não estão relacionados, ou seja, as duas amostras são independentes, pois os resultados de uma amostra não interfere nos resultados da outra amostra. Também verificamos que as observações dentro das amostras são independentes, pois são morcegos distintos.

Para o caso de *amostras independentes*, podemos, ainda, distinguir duas situações, quais sejam:

1. quando as variâncias são desconhecidas mas podem ser consideradas iguais, ou seja, a ordem de grandeza não difere muito;
2. quando as variâncias são desconhecidas e consideradas desiguais, ou seja, a ordem de grandeza difere bastante.

Vamos fazer dois testes de significância para comparação de duas médias de amostras independentes, um para variâncias homogêneas (consideradas iguais) e outro para variâncias heterogêneas (consideradas desiguais). Portanto, percebemos que também há a necessidade

de se fazer um teste de hipóteses para as variâncias populacionais, para podermos decidir se as mesmas podem ser consideradas iguais ou não.

#### 8.4.2.1 Teste para Dados Pareados

Inicialmente, vamos apresentar um exemplo que está muito relacionado com o planejamento de um experimento. Na verdade, vamos exemplificar a construção do delineamento, talvez o mais utilizado na experimentação, denominado de *blocos ao acaso*.

Vamos supor que desejamos colocar em teste duas cultivares de feijão (cultivar 1 e cultivar 2). Decidiu-se que serão feitas sete repetições de cada cultivar, portanto, os tamanhos das amostras vão ser iguais e igual a 7,  $n_1 = n_2 = 7$ . Também, decidiu-se que as unidades experimentais vão ser formadas por áreas de terra (canteiros) de  $20\text{ m}^2$  e serão utilizadas 250 gramas de sementes por unidade experimental. Como temos duas cultivares, cada uma com 7 repetições, vamos precisar de 14 unidades experimentais. Vamos imaginar, que a área que o pesquisador tem a disposição para implantar o experimento, apresenta uma considerável inclinação, fazendo com que exista heterogeneidade entre as unidades experimentais, quanto à fertilidade do solo. Na parte alta do terreno a fertilidade é menor do que na parte baixa do mesmo. É fundamental que o pesquisador encontre uma forma de dispor as unidades experimentais, de tal forma que a diferença de fertilidade seja eliminada da comparação entre as duas cultivares. Portanto, já de início, devemos descartar a possibilidade de se fazer um sorteio aleatório das duas cultivares nas 14 unidades experimentais. Imagine, que no sorteio, a maioria das unidades experimentais que estão localizadas na faixa mais fértil da área receba a cultivar 1. Dessa forma, não saberemos distinguir o que é variação devido a cultivar ou devido à fertilidade do solo (dizemos que os efeitos estão confundidos). A forma correta do desenho experimental é dado na Figura 8.6. Foram construídos sete blocos, dentro de cada bloco as unidades são homogêneas quanto à fertilidade, pois estão na mesma faixa do terreno. Entre os blocos pode e deve existir diferenças de fertilidade. O sorteio das cultivares às unidades experimentais é feito independentemente para cada bloco. Portanto, a formação de blocos constitui um pareamento.

Um outro critério para a formação dos pares pode ser, por exemplo, obtido de formulações de um comprimido. Deseja-se comparar dois métodos de determinação da quantidade de princípio ativo presente nos comprimidos, quais sejam: 1) HPLC e 2) Titulação. Para cada formulação aplica-se os dois métodos. Se temos, por exemplo, sete formulações, teremos sete pares.

Vamos apresentar o teste de duas médias para dados pareados através de um exemplo. Como fizemos para o caso do teste de hipótese de uma média, aqui, também vamos fazer

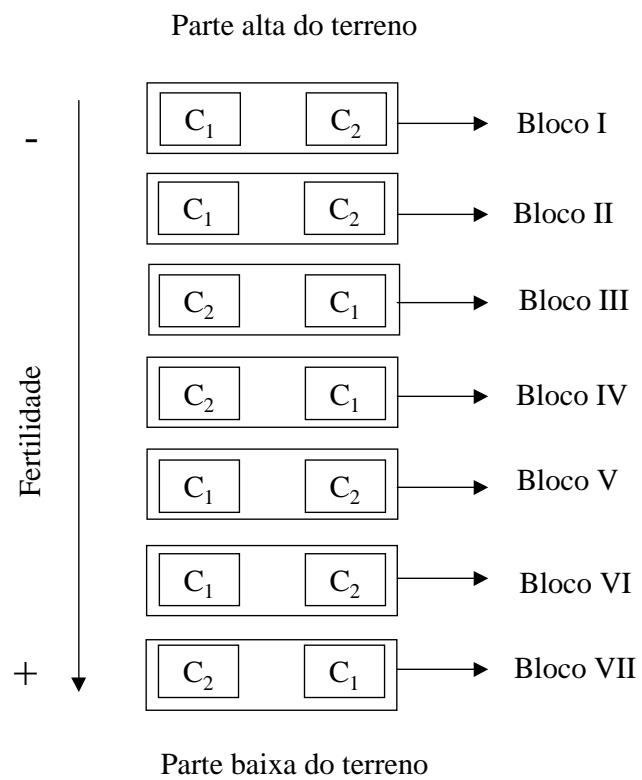


Figura 8.6: Um exemplo de Delineamento em Blocos ao Acaso

o teste através do valor crítico e do *valor p*. Como poderá ser verificado, o teste para comparação de duas médias de amostras pareadas, é um teste para comparação de uma média de valores de diferenças dos pares.

**Exemplo.** Foi conduzido um experimento para estudar o conteúdo de hemoglobina no sangue de suínos com deficiência de niacina. Aplicou-se 20 mg de niacina em 8 suínos. Podemos afirmar que o conteúdo de hemoglobina no sangue diminuiu, com a aplicação de niacina, ao nível de significância de 5%? Encontre o intervalo de confiança com 95% para a verdadeira redução no teor médio da hemoglobina. Foram mensurados os níveis de hemoglobina no sangue antes e depois da aplicação da niacina, os resultados obtidos no experimento foram:

Suínos	Antes (A)	Depois (B)	Diferenças (A-B)
1	13,6	11,4	2,2
2	13,6	12,5	1,1
3	14,7	14,6	0,1
4	12,1	13,0	-0,9
5	12,3	11,7	0,6
6	13,2	10,3	2,9
7	11,0	9,8	1,2
8	12,4	10,4	2,0

Vamos considerar que temos um suporte teórico para esperar que o conteúdo de hemoglobina no sangue diminua com a aplicação de niacina, portanto, as hipóteses ficam:

$$H_0 : \mu_A = \mu_B \quad \text{versus} \quad H_1 : \mu_A > \mu_B.$$

As mesmas hipóteses, dadas em forma de diferenças entre as duas médias,  $\mu_D = \mu_A - \mu_B$ , podem ser escritas como,

$$H_0 : \mu_D = 0 \quad \text{versus} \quad H_1 : \mu_D > 0.$$

Trabalhando-se com as diferenças, na verdade estamos eliminando as diferenças entre os indivíduos (no exemplo, suínos). Vamos admitir que, sob  $H_0$ , as diferenças de conteúdo de hemoglobina, seguem, pelo menos aproximadamente, uma distribuição normal com média 0 (zero) e variância  $\sigma_D^2$ , e a representamos por:  $D \sim N(0, \sigma_D^2)$ . Tomando-se os valores das diferenças entre  $A$  e  $B$ , caímos no caso do teste de uma média com variância desconhecida,

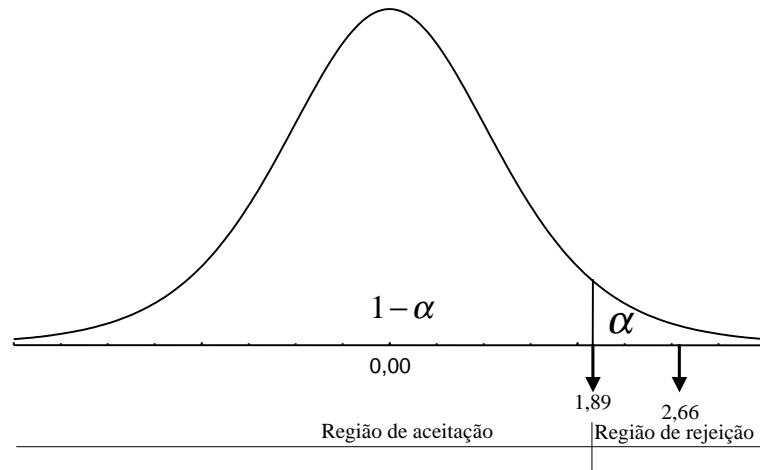


Figura 8.7: Teste de hipótese para comparar duas médias de dados pareados

cuja estatística teste é como em 8.15 e dada por:

$$t = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{n}}} = \frac{\bar{D}}{\frac{s_D}{\sqrt{n}}}, \quad (8.18)$$

a qual segue uma distribuição  $t$  de Student com  $n - 1$  graus de liberdade. Aqui,  $\bar{D}$  é a média das diferenças entre os valores de  $A$  e  $B$  e  $S_D$  é o desvio padrão das diferenças. Com os dados da amostra obtemos:  $\bar{d} = 1,15$  e  $s_{\bar{d}} = 1,225$ .

Para um nível de significância fixado em 5%, e teste unilateral, obtemos na tabela da distribuição  $t$  de Student o valor  $t_7 = 1,895$ . A estatística teste 8.18 calculada com os valores da amostra vale,

$$t = \frac{1,15}{1,225/\sqrt{8}} = 2,655.$$

As regiões do teste de hipótese e o valor da estatística teste estão indicadas na Figura 8.7, para o nível de significância desejado. Como o valor de  $t$  calculado na amostra cai na região de rejeição de  $H_0$ , concluímos, ao nível de significância de 5%, que o conteúdo de hemoglobina diminui com a aplicação de 20 mg de niacina.

Podemos, facilmente, obter o *valor p* do teste, determinando a probabilidade de se encontrar um valor igual ou mais extremo do que 2,655 devido ao acaso. Da tabela da distribuição  $t$  de Student (Apêndice 4), com 7 graus de liberdade, obtemos o *valor p* entre 0,01 e 0,02. Portanto, rejeitamos a hipótese nula ao nível de significância de 5%.

Com o auxílio de um programa computacional estatístico, podemos obter o exato *valor*



$p$  do teste, o qual vale:

$$P(t > 2,655) = 0,016352.$$

Importante: devemos tomar muito cuidado quando estamos usando um programa estatístico para fazer um teste de hipótese, no sentido de verificar se o programa está fazendo um teste bilateral ou unilateral.

O intervalo de confiança para  $\mu_D$  é dado por:

$$\begin{aligned} IC(\mu_D; 95\%) &: 1,15 \pm 2,365 \frac{1,225}{\sqrt{8}} \\ &: 1,15 \pm 1,0243 \\ &[0,1257; 2,1743]. \end{aligned} \quad (8.19)$$

Do intervalo construído podemos concluir que o uso de 20 gramas de niacina ocasiona uma redução na taxa média de hemoglobina do sangue de pelo menos 0,1257 e de no máximo 2,1743, com grau de confiança de 95%. Observe que o valor de  $t$  vale 2,365, obtido diretamente na tabela para  $\alpha = 5\%$ , pois o intervalo de confiança é bilateral, assim como a tabela por nós utilizada.

#### 8.4.2.2 Teste para Comparação de Duas Médias, de Amostras Independentes e Variâncias Desconhecidas e Iguais

Como vimos, o procedimento de teste de hipóteses sempre segue uma mesma seqüência, o que muda são as hipóteses de interesse do pesquisador e as estatísticas de teste a serem utilizadas.

Vamos trabalhar com amostras independentes, oriundas de duas populações com distribuição normal e com variâncias desconhecidas e iguais. Também deve existir independência entre as observações de cada amostra. A verificação da igualdade das duas variâncias pode ser feita através de um outro teste de hipóteses que está apresentado na Seção 8.4.2.4.

Para estudar se duas médias populacionais são iguais ou não, conforme as hipóteses descritas no início da Seção 8.4.2, a estatística teste a ser utilizada é dada por:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}} \quad (8.20)$$

a qual tem distribuição  $t$  de Student com  $n_1 + n_2 - 2$  graus de liberdade, onde  $n_1$  é o número de observações da amostra da população 1,  $n_2$  é o número de observações da amostra da população 2, e  $S_1^2$  e  $S_2^2$  são as respectivas variâncias amostrais.

A hipótese nula é rejeitada quando o valor da estatística  $t$ , em valor absoluto, é maior do que o valor crítico ( $t_{n_1+n_2-2}$ ), obtido da tabela da distribuição  $t$  de Student com nível de significância  $\alpha$ .

**Exemplo 1.** Com o objetivo de comparar as produções médias, em toneladas por hectare, de duas variedades de milho (Variedade A e Variedade B), foram observadas 5 unidades experimentais para cada variedade e os resultados obtidos foram os seguintes:

Variedade A	1,3	1,4	1,1	1,4	1,5
Variedade B	1,8	1,6	1,9	1,9	1,8

Dos dados das amostras obtemos,  $\bar{x}_A = 1,34$ ,  $\bar{x}_B = 1,80$ ,  $s_A^2 = 0,0231$  e  $s_B^2 = 0,0150$ . Assumindo que as variâncias populacionais sejam iguais, o que será verificado na Seção 8.4.2.4, a estatística 8.20 vale:

$$t = \frac{(1,34 - 1,80)}{\sqrt{\frac{(5-1)0,0231 + (5-1)0,0150}{5+5-2} \left[\frac{1}{5} + \frac{1}{5}\right]}} = \frac{-0,46}{\sqrt{0,0191(0,40)}} = -5,263.$$

Da tabela da distribuição  $t$  de Student, para uma hipótese alternativa bilateral, com 8 graus de liberdade, o valor crítico do teste correspondente a um nível de significância de 5% vale  $t_8 = 2,306$ . Como o valor absoluto de  $t$  é 5,263, e maior do que 2,306, um valor pertencente a região de rejeição, conforme a Figura 8.8, rejeitamos a hipótese de igualdade das produções médias das duas variedades de milho, ao nível de significância de 5%. Portanto, a cultivar B é mais produtiva do que a cultivar A.

Para realizarmos o teste em termos do *valor p*, este é obtido entrando na tabela da distribuição  $t$  com 8 graus de liberdades (1ª coluna). Seguimos pela linha dos 8 graus de liberdade até encontrar um valor igual ou mais próximo de 5,263. O valor mais próximo encontrado é 5,041, cuja probabilidade vale 0,001. Como o valor absoluto encontrado da estatística teste (5,263) é maior do que 5,041, implica que o *valor p* é menor do que 0,001. Portanto, rejeitamos a hipótese nula.

Usando um software estatístico, para 8 graus de liberdade e teste bilateral, obtemos exatamente *valor p* igual a  $P(|t| > 5,263) = 0,000762$ .

Podemos calcular o intervalo de confiança para a diferença entre as médias,  $\mu_B - \mu_A$ . O intervalo de confiança é dado pela expressão,

$$IC(\mu_B - \mu_A; 95\%) : (\bar{x}_B - \bar{x}_A) \pm t_{n_A+n_B-2; 5\%} s_{\bar{x}_B - \bar{x}_A},$$

onde  $s_{\bar{x}_B - \bar{x}_A}$  é o desvio padrão da diferença entre as duas médias, e é dado por:

$$\sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2} \left[\frac{1}{n_A} + \frac{1}{n_B}\right]}.$$

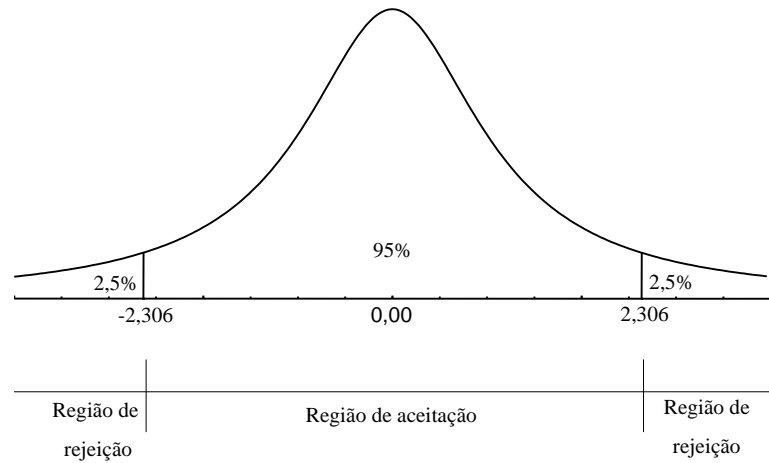


Figura 8.8: Teste de hipótese para comparar duas médias dos dados de produção de duas variedades de milho

Para o exemplo, o intervalo de confiança é dado por,

$$\begin{aligned} IC(\mu_B - \mu_A; 95\%) &: 0,46 \pm 2,306(0,0873) \\ &: 0,46 \pm 0,20131 \\ &[0,2587; 0,6613]. \end{aligned}$$

Portanto, podemos afirmar com 95% de confiança, que a produção média de milho da variedade B é superior a produção média da variedade A em no mínimo 0,2587 toneladas por hectare e no máximo 0,6613 toneladas por hectare.

Podemos também construir intervalos de confiança para a produção média de cada uma das variedades. Para a variedade A o intervalo de confiança é dado pela expressão:

$$IC(\mu_A; 95\%) : (\bar{x}_A) \pm t_{(n_A+n_B-2; 5\%)} \frac{s_{pool}}{\sqrt{n_A}},$$

onde:

$$s_{pool} = \sqrt{\frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}}.$$

A partir dos dados,  $s_{pool} = \sqrt{0,0191} = 0,1382$  e o intervalo é dado por:

$$\begin{aligned} IC(\mu_A; 95\%) &: 1,34 \pm 2,306 \frac{0,1382}{\sqrt{5}}, \\ &: 1,34 \pm 0,1425 \\ &: [1,1975; 1,4825] \end{aligned}$$

Com 95% de confiança podemos afirmar que a produção média de milho da variedade A é um valor entre 1,1975 e 1,4825 t/ha.

Para a variedade B o intervalo de confiança é dado pela expressão:

$$IC(\mu_B; 95\%) : (\bar{x}_B) \pm t_{(n_A+n_B-2; 5\%)} \frac{s_{pool}}{\sqrt{n_B}},$$

A partir dos dados o intervalo é dado por:

$$\begin{aligned} IC(\mu_B; 95\%) &: 1,80 \pm 2,306 \frac{0,1382}{\sqrt{5}}, \\ &: 1,80 \pm 0,1425 \\ &: [1,6575; 1,9425] \end{aligned}$$

Com 95% de confiança podemos afirmar que a produção média de milho da variedade B é um valor entre 1,6575 e 1,9425 t/ha.

**Exemplo 2.** Foi realizado um experimento com o objetivo de comparar os tempos gastos na manobra com os arados Fuçador e Erechim. Ambos os arados são de tração animal. Os dados obtidos com 11 repetições para cada arado, foram os seguintes:

Fuçador	0,20	0,22	0,18	0,23	0,12	0,20	0,13	0,12	0,13	0,22	0,17
Erechim	0,36	0,48	0,33	0,43	0,40	0,43	0,33	0,36	0,35	0,40	0,35

Espera-se que o arado Fuçador produza melhores resultados (gaste menos tempo médio na manobra). Testar a hipótese de que a média de tempo gasto com o arado Fuçador ( $\mu_F$ ) é menor do que a média do tempo gasto com o arado Erechim ( $\mu_E$ ), ao nível de 5%. Vamos assumir que as variâncias populacionais são iguais. Isto será verificado na Seção 8.4.2.4.

Com os dados amostrais obtemos:  $\bar{x}_F = 0,1745$ ,  $\bar{x}_E = 0,3836$ ,  $s_F^2 = 0,0018$  e  $s_E^2 = 0,0023$ .

As hipóteses sobre as médias verdadeiras, de interesse do pesquisador, são formuladas como:

$$H_0 : \mu_F = \mu_E \quad H_1 : \mu_F < \mu_E.$$

O valor da estatística 8.20, vale:

$$t = \frac{(0,1745 - 0,3836)}{\sqrt{\left[ \frac{(11-1)0,0018 + (11-1)0,0023}{11+11-2} \right] \left[ \frac{1}{11} + \frac{1}{11} \right]}} = -10,831.$$

Da tabela da distribuição  $t$  de Student temos  $t_{20;5\%} = 1,725$  como valor crítico do teste, ao nível de significância de 5%. Como o valor absoluto de  $t$  calculado é 10,831, e é maior do

que o valor crítico, rejeitamos a hipótese nula. Conclusão, o tempo médio gasto na manobra com o arado Fuçador é menor do que com o arado Erechim, ao nível de significância de 5%.

O intervalo de confiança de 95%, para a diferença entre as duas médias, é dado por:

$$IC(\mu_E - \mu_F; 95\%) : 0,2091 \pm 2,086(0,01931) \\ [0,1688; 0,2494].$$

Concluimos, com confiança de 95%, que a manobra realizada com o arado Erechim, leva em média de 0,17 a 0,25 minutos a mais do que com o arado Fuçador.

#### 8.4.2.3 Teste para Comparação de Duas Médias, de Amostras Independentes e Variâncias Desconhecidas e Desiguais

Suponhamos que as amostras sejam oriundas de duas populações com distribuição normal, independentes e variâncias desconhecidas e desiguais. Neste caso, a estatística de teste da igualdade entre duas médias, é dada por:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \quad (8.21)$$

Prova-se que esta variável tem uma distribuição *aproximada*  $t$  de Student, com graus de liberdade corrigido, dados por:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\frac{S_1^2}{n_1}}{\frac{n_1}{n_1-1}} + \frac{\frac{S_2^2}{n_2}}{\frac{n_2}{n_2-1}}}.$$

**Exemplo 1.** Em um centro agrícola, deseja-se testar o efeito de determinado fertilizante sobre a produção de trigo. Para isso, escolheram-se 24 áreas de terra, cada uma com  $5 \times 2 = 10m^2$ , de uma grande área homogênea. Metade dessas unidades são tratadas com um fertilizante, enquanto a outra metade não recebe o fertilizante, este é denominado de tratamento controle. A produção média de trigo sem fertilizante foi de 1260 kg/ha com desvio padrão de 730 kg/ha, enquanto que a produção média com fertilizante foi de 1710 kg/ha com desvio padrão de 280 kg/ha. Podemos afirmar que houve aumento significativo na produção de trigo devido a utilização do fertilizante?

As hipóteses, sobre a médias populacionais,  $\mu_F$  e  $\mu_C$ , relativas ao fertilizante e ao controle, respectivamente, são formuladas do seguinte modo:

$$H_0 : \mu_F = \mu_C \quad H_1 : \mu_F > \mu_C.$$

Supondo variâncias diferentes, o que será verificado na Seção 8.4.2.4, o valor da estatística  $t$  é:

$$t = \frac{(1710 - 1260)}{\sqrt{\frac{532900}{12} + \frac{78400}{12}}} = 1,99,$$

com graus de liberdade,

$$v = \frac{\left(\frac{532900}{12} + \frac{78400}{12}\right)^2}{\frac{\left(\frac{532900}{12}\right)^2}{11} + \frac{\left(\frac{78400}{12}\right)^2}{11}} = 14,17 \cong 15.$$

Da distribuição  $t$ , com 15 graus de liberdade, o valor crítico para um teste unilateral, com nível de significância de 5% vale 1,753. Como o valor absoluto de  $t$  calculado é 1,99, e é maior do que o valor crítico, concluímos que o uso de fertilizante provocou um aumento real na produção de trigo, ao nível de significância de 5%.

Como nos exemplos anteriores, intervalos de confiança poderiam ser construídos para diferenças das médias e também para cada uma delas.

**Exemplo 2.** As seguintes medidas de *Cytochrome oxidase* foram determinadas em machos de peixes *Periplaneta* em  $mm^3$  por 10 minutos por miligrama, em um estudo para comparar dois tratamentos, quais sejam: 1) 24 horas após injeção de methoxyclor e 2) controle, ou seja, sem injeção de methoxyclor.

Tratamentos	tamanho amostra	média	desvio padrão	variância
24 horas após injeção de methoxyclor	5	24,8	0,9	0,81
Controle	3	19,7	2,8	7,84

Verifique se existe efeito significativo da aplicação de methoxyclor quanto as médias de *Cytochrome oxidase*.

Para a comparação entre as duas médias,  $\mu_T$  e  $\mu_C$ , relativas ao tratamento e ao controle, respectivamente, vamos usar a estatística 8.21, supondo que as variâncias são desiguais.

As hipóteses, sobre as médias verdadeiras, são dadas por:

$$H_0 : \mu_T = \mu_C \quad H_1 : \mu_T \neq \mu_C.$$

O valor da estatística dada em 8.21 é:

$$t = \frac{(24,8 - 19,7)}{\sqrt{\frac{0,81}{5} + \frac{7,84}{3}}} = 3,06,$$

com, graus de liberdade,

$$v = \frac{\left(\frac{0,81}{5} + \frac{7,84}{3}\right)^2}{\frac{\left(\frac{0,81}{5}\right)^2}{4} + \frac{\left(\frac{7,84}{3}\right)^2}{2}} = 2,25 \cong 3.$$

Da distribuição  $t$ , com 3 graus de liberdade, o valor crítico para um teste bilateral, com nível de significância de 5% vale 3,182. Como o valor absoluto de  $t$  calculado é 3,06, e é menor do que o valor crítico, concluímos que o uso de aplicação de methoxyclor não forneceu um resultado diferente do controle, ao nível de significância de 5%.

#### 8.4.2.4 Teste $F$ para Comparação de Variâncias Populacionais

Como qualquer outro teste, devemos sempre iniciar pela formulação das hipóteses, que nesse caso são dadas por,

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_1 : \sigma_1^2 > \sigma_2^2.$$

Onde  $\sigma_1^2$  e  $\sigma_2^2$  são as variâncias da população 1 e população 2, respectivamente. A hipótese nula admite que as duas populações tem a mesma variância. A hipótese alternativa admite que a variância da população 1 é maior do que a variância da população 2, ou seja, as variâncias são heterogêneas. Portanto, vamos aqui, sempre fazer um teste unilateral.

Quando desejamos comparar variâncias, devemos utilizar a estatística  $F$ , dada pelo quociente entre as duas estimativas de variâncias,  $s_1^2$  e  $s_2^2$ , de  $\sigma_1^2$  e  $\sigma_2^2$ , respectivamente, distintas e supostas independentes. Temos,

$$F = \frac{s_1^2}{s_2^2}. \quad (8.22)$$

As variâncias  $s_1^2$  e  $s_2^2$  são calculadas com os  $n_1$  e  $n_2$  dados das amostras, respectivamente. Assim, dizemos que a  $s_1^2$  estão associados  $n_1 - 1$  graus de liberdade (numerador) e, da mesma forma, para  $s_2^2$  estão associados  $n_2 - 1$  graus de liberdade (denominador). Vamos admitir sempre que  $s_1^2$  é maior do que  $s_2^2$ , ou seja, no numerador vamos usar a variância maior, de tal modo que tenhamos  $F \geq 1$ .

Como nos outros testes de hipóteses já discutidos neste livro, a veracidade da hipótese nula pode ser verificada tanto através do *valor p* quanto do valor crítico da distribuição  $F$  apresentada no Apêndice 6.

Com os dados das amostras calculamos  $F$  da expressão 8.22, e comparamos o seu valor com o valor de  $F$  crítico, obtido na tabela da distribuição  $F$ , dada no apêndice 6, com  $n_1 - 1$  graus de liberdade no numerador e  $n_2 - 1$  graus de liberdade no denominador, e com um nível de significância  $\alpha$  fixado pelo pesquisador. Então, por este procedimento, rejeitamos  $H_0$  se  $F > F_{n_1-1, n_2-1, \alpha}$ . A tabela fornecida no apêndice 6 apresenta o nível de significância ( $\alpha$ )

fixado em 5%. Existem tabelas para outros valores de  $\alpha$ . Porém, os programas estatísticos já nos fornecem as probabilidades, através da distribuição de  $F$ , ou o *valor p* do teste.

**Exemplo 1.** Dos dados do exemplo para comparação de duas variedades de milho, A e B, apresentados na Seção 8.4.2.2, obtemos,  $s_A^2 = 0,0231$  com 4 graus de liberdade e  $s_B^2 = 0,0150$  também com 4 graus de liberdade. Assim, o valor da estatística  $F$  é,

$$F = \frac{0,0231}{0,0150} = 1,54,$$

a qual tem 4 graus de liberdade no numerador e 4 graus de liberdade no denominador. Da tabela da distribuição  $F$  (apêndice 6), com um nível de significância de 5%, com 4 e 4 graus de liberdade no numerador e denominador, respectivamente, obtemos o valor de  $F$  crítico como sendo igual a 6,39. Como o valor de  $F$  é 1,54, e é menor do que o  $F$  crítico, aceitamos a hipótese nula, portanto, as variâncias são homogêneas.

**Exemplo 2.** Dos dados do exemplo para comparação de dois arados, Fuçador e Erechim, apresentados na Seção 8.4.2.2, obtemos  $s_F^2 = 0,0023$  com 10 graus de liberdade e  $s_E^2 = 0,0018$  também com 10 graus de liberdade. A estatística  $F$  vale,  $F = s_E^2/s_F^2 = 0,0023/0,0018 = 1,2778$ , com 10 graus de liberdade no numerador e 10 graus de liberdade no denominador. Da tabela da distribuição  $F$  (apêndice 6), com um nível de significância de 5%, com 10 e 10 graus de liberdade no numerador e denominador, respectivamente, obtemos o valor de  $F$  crítico como sendo igual a 2,98. Como o valor de  $F$  é 1,28, e é menor do que o  $F$  crítico, aceitamos a hipótese nula, portanto, as variâncias são homogêneas.

**Exemplo 3.** Dos dados do exemplo para testar o efeito de determinado fertilizante sobre a produção de trigo, apresentados na Seção 8.4.2.3, obtemos  $s_C^2 = 532900$  com 11 graus de liberdade e  $s_F^2 = 78400$  também com 11 graus de liberdade. A estatística  $F$  vale,  $F = s_E^2/s_F^2 = 532900/78400 = 6,7972$ , com 11 graus de liberdade no numerador e 11 graus de liberdade no denominador. Da tabela da distribuição  $F$  (apêndice 6), com um nível de significância de 5%, com 11 e 11 graus de liberdade no numerador e denominador, respectivamente, obtemos o valor de  $F$  crítico como sendo igual a 2,82. Como o valor de  $F$  é 6,80, e é maior do que o  $F$  crítico, rejeitamos a hipótese nula, portanto, as variâncias são heterogêneas.

**Exemplo 4.** Dos dados do exemplo para testar o efeito de injeção de methoxycloer em peixes Periplaneta, apresentados na Seção 8.4.2.3, obtemos  $s_T^2 = 0,81$  com 4 graus de liberdade e  $s_C^2 = 7,84$  com 2 graus de liberdade. A estatística  $F$  vale,  $F = s_C^2/s_T^2 = 7,84/0,81 = 9,68$ , com 2 graus de liberdade no numerador e 4 graus de liberdade no denominador. Da tabela da distribuição  $F$  (apêndice 6), com um nível de significância de 5%, com 2 e 4 graus de liberdade no numerador e denominador, respectivamente, obtemos o valor de  $F$  crítico



como sendo igual a 6,94. Como o valor de  $F$  é 9,68, e é maior do que o  $F$  crítico, rejeitamos a hipótese nula, portanto, as variâncias são heterogêneas.

## 8.5 Exercícios Propostos

1. Sabe-se que a proporção de sementes de soja, com danos mecânicos, provocados pelo beneficiamento, é  $\pi = 0,18$ . Foram feitas 40 observações durante um dia de trabalho e se constatou que  $p=0,20$ . Verificar se a máquina precisa ser regulada, ao nível de significância de 5%.

2. Foram as seguintes as produções médias de duas variedades de soja, em t/ha: variedade A: média=3,8 t/ha com variância =  $0,36 (t/ha)^2$ ; variedade B: média=4,6 t/ha com variância =  $0,04 (t/ha)^2$ . As informações obtidas, para as variedades A e B, foram baseadas em amostras de tamanhos 30 e 35 respectivamente. Teste a hipótese de que não há diferença significativa entre as produções médias, ao nível de 1% de probabilidade.

3.(Lindstrom,1918; Snedecor & Cochran,1972). Em um experimento envolvendo o cruzamento de dois tipos de milho, Lindstrom encontrou quatro tipos distintos: verde(T1), dourado(T2), listras verdes (T3) e lisas verdes e douradas (T4). De acordo com a hereditariedade mendeliana, as probabilidades de obter esses quatro tipos são: 9/16, 3/16, 3/16 e 1/16, respectivamente. Em 1301 indivíduos da segunda geração, Lindstrom encontrou as seguintes frequências:

T1	773
T2	231
T3	238
T4	59
total	1301

Teste a hipótese de que a distribuição segue as leis da hereditariedade Mendeliana, contra a hipótese alternativa de que a distribuição não segue as leis da hereditariedade.

4. Que suposições são feitas acerca das populações de onde se extraem amostras aleatórias independentes, quando se utiliza a distribuição  $t$  de Student exata para realizar inferências a respeito da diferença entre médias populacionais?

5. A área foliar média da espécie *Laguncularia ranceirosa* do Manguezal do Bairro Santa Mônica (não poluído) é de  $50,76 \text{ cm}^2$ . Espera-se que a área foliar do Manguezal do Bairro Itacorubi seja maior devido a poluição do ambiente. Para verificar se essa afirmação é verdadeira coletou-se uma amostra de tamanho  $n = 20$ , do Manguezal do Itacorubi, cujos

resultados foram:

39,4	39,6	39,9	45,6	45,6	46,1	46,1	50,2	50,2	51,0
51,2	54,6	54,8	54,6	55,1	55,1	55,5	56,2	66,3	66,5

a) Esses resultados trazem evidências estatísticas de que houve aumento da área foliar? Use nível de significância de 5%.

b) Construir o intervalo de confiança com 90%, para a verdadeira média da área foliar do Manguezal do Itacorubi. Interpretar o resultado.

6. Faz-se um estudo com o objetivo de conhecer melhor a biologia e a distribuição de fêmeas de *Macrobrachium potiuna* na ilha de Santa Catarina, comparando-se dois locais em condições ambientais diferentes. Para isso coletou-se duas amostras, uma no Córrego do Chico em Rationes (A) e a outra no Poção do Córrego Grande (B). os resultados de comprimento total, em mm, de fêmeas ovígeras, foram:

Local	Amostra	Média	Desvio padrão
Córrego do Chico	30	32,25	3,26
Poção do Córrego	30	20,32	4,02

a) Determinar o intervalo de confiança de 99% para a verdadeira média de comprimento total para cada um dos locais. Interpretar.

b) Qual deve ser o tamanho da amostra para que a amplitude total de um intervalo de 99% de confiança para a média populacional do Poção do Córrego Grande seja no máximo igual a 4 mm?

c) Você diria que a diferença de comprimentos nos dois locais é estatisticamente significativa? use  $\alpha = 5\%$ .

7. Foi conduzido um experimento com o objetivo de avaliar o poder germinativo de duas cultivares de cebola: a) Bola Precoce-Empasc 352 e b) Norte 14. foram utilizadas para o teste de germinação, 4 repetições de 100 sementes, totalizando 400 sementes para cada cultivar. A variável de estudo é o número de sementes que germinaram. os resultados estão apresentados na Tabela 8.9. Teste a hipótese de que não há dependência entre cultivares e número de sementes que germinam, ao nível de significância de 5%. De outro modo, poderia perguntar: existe diferença entre as duas cultivares quanto à germinação?

8. Um agricultor que planta frutíferas deseja testar um novo tipo de inseticida, cujo fabricante garante reduzir os prejuízos causados por certo tipo de inseto. Para verificar essa afirmação do fabricante do inseticida, o agricultor pulveriza 200 árvores com o produto novo

Tabela 8.9: Poder germinativo de duas cultivares de cebola

Cultivares	Germinação		Total
	Germinaram	Não germinaram	
Bola precoce	392	8	400
Norte 14	381	19	400
Total	773	27	800

e 200 árvores com o produto que normalmente usa, obtendo os resultados:

	Inseticida novo	Inseticida padrão
Produção média (kg/planta)	240	227
Variância	980	820

Estes dados indicam evidência suficiente de que o inseticida novo é melhor do que o padrão (normalmente usado)?

9. Deseja-se testar a hipótese da possibilidade da quantidade de proteínas totais no plasma, depois de determinada operação em portadores de esquistossomose mansônica, ser diferente da quantidade antes da operação. Foi utilizada uma amostra de 17 pacientes, cujos

resultados foram:

Paciente	Antes	Depois
1	6,9	6,9
2	7,8	8,6
3	6,6	8,7
4	5,9	7,3
5	7,8	7,8
6	6,4	8,2
7	8,8	9,3
8	7,3	7,3
9	8,0	7,6
10	8,6	7,8
11	7,7	7,6
12	7,9	7,8
13	8,7	8,1
14	5,8	6,8
15	9,2	8,3
16	9,3	10,2
17	8,9	9,1

Faça o teste de hipótese e conclua.

10. Um biólogo está interessado em determinar o diâmetro médio da roseta foliar da bromélia de sol (ambiente de sol). o diâmetro foi medido numa amostra de tamanho  $n = 20$ , obtendo-se os seguintes resultados:

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

- Estime o diâmetro médio da roseta foliar utilizando um intervalo de confiança de 95%.
- Sabe-se que o diâmetro médio da roseta foliar das bromeliaceas em área sombreada é de 16 cm. Os dados amostrais obtidos evidenciam que as bromélias de sol desenvolvem-se menos do que as de sombra? Faça o teste de hipótese com nível de significância de 5%.
- Num trabalho futuro com bromélias de sol, determine o tamanho da amostra necessário para que o erro cometido na estimação seja de, no máximo, 0,70, com probabilidade de 95%.

11. Para comparar o peso vivo e peso jejum do gado Crioulo Lageano, um pesquisador selecionou aleatoriamente uma amostra de 15 animais e anotou o seus pesos. Os dados resultantes estão descritos abaixo. Esses dados têm evidência suficiente para garantir que

exite diferença entre peso vivo e peso jejum? Indique o nível mínimo significativo do teste ( $\alpha$ ).

Animal	Peso vivo	Peso jejum	Animal	Peso vivo	Peso jejum
1	498	453	9	398	353
2	510	466	10	410	380
3	540	491	11	450	412
4	580	480	12	495	453
5	440	405	13	508	462
6	350	315	14	515	477
7	595	557	15	560	505
8	513	477			

12. Realizou-se uma pesquisa para comparar dois locais, Itajaí e Araranguá, quanto à produção de arroz irrigado, em t/ha, na safra 88/89. Dez progênies foram utilizadas nos dois locais e os seus resultados anotados.

Itajaí		Araranguá	
5,3	9,3	6,0	8,4
5,7	9,2	6,3	8,3
7,0	6,9	7,2	7,2
6,2	7,0	6,8	7,3
8,0	6,5	7,8	7,0

Teste a hipótese de igualdade entre as produções médias.

13. Um agrônomo realizou um levantamento para estudar o desenvolvimento de duas espécies de árvores, a bracatinga (*Mimosa scabrella*) e Canafístula (*Peltophorum dubium*). Para esta finalidade foram coletadas duas amostras de tamanhos igual a 30 árvores. Os resultados para altura (dados fictícios), em metros, foram:

Bracatinga						Canafístula					
6,4	7,0	9,0	10,2	16,1	20,1	8,2	10,1	14,1	20,2	25,7	40,1
6,8	8,3	9,1	11,4	16,3	20,3	9,7	10,3	14,2	20,3	30,9	40,2
6,9	8,6	9,3	13,7	17,2	21,4	9,8	11,2	14,4	20,6	35,5	40,5
6,9	8,7	9,9	14,8	18,4	22,8	10,0	13,2	14,8	29,9	38,2	41,8
6,9	8,7	10,1	15,2	20,0	22,8	10,0	13,4	15,9	23,8	40,0	42,3

A hipótese levantada pelo pesquisador é que a Canafístula deve apresentar uma altura média maior do que a Bracatinga. Esses dados indicam evidência suficiente para suportar a hipótese do pesquisador?

14. Um biólogo deseja fazer um estudo para verificar a especificidade do fungo *Gigaspora gigantea* com as plantas *Spartina sp* e *Tibouchina sp*. Foram coletadas 20 amostras de solos em volta da *Spartina sp* e 20 amostras de *Tibouchina sp* e verificada a quantidade de fungo presente. os resultados foram:

<i>Spartina sp</i>				<i>Tibouchina sp</i>			
303	78	59	15	305	74	60	18
116	93	43	10	110	95	40	11
117	63	41	5	110	62	40	4
115	63	35	6	112	60	32	9
84	55	25	3	85	53	22	4

Esses dados tem evidência suficiente para indicar uma diferença entre as quantidades médias de *Gigaspora gigantea* encontrada em plantas de *Spartina sp* e *Tibouchina sp*?

15. Para determinar qual de duas cultivares de trigo é mais produtiva, um Centro de Pesquisas em Agricultura selecionou quatro áreas de terra, cada área em uma região recomendada para a produção de trigo no Estado de Santa Catarina. Cada área é totalmente homogênea e foi dividida em duas partes. Foram semeadas as duas cultivares, uma em cada parte da área, através de um sorteio. Entre as regiões deverá haver diferenças de produção. As produções, em kg/ha, foram:

Município	Cultivares	
	CEP 15-Batouí	CEP 18-Cavera
Abelardo Luz	2912	3854
Chapecó	2528	3295
Campos Novos	3309	3916
Mafra	2324	2548

Qual é a cultivar de trigo mais produtiva? Use nível de significância de 5%. Interpretar.

16. Foi conduzido um experimento, no qual avaliou-se a utilização de uma, duas e três armadilhas (raíz de tayuyá) por  $400 m^2$ , com e sem inseticida para capturar a vaquinha (*Diabrotica speciosa*) na cultura do feijão. Os dados obtidos, relativos ao número de vaquinhas

capturadas, foram:

Armadilhas	Inseticida		Total
	Com	Sem	
Uma armadilha	114	197	301
Duas armadilhas	1067	362	1429
Três armadilhas	715	273	988
Total	1896	822	2718

Podemos afirmar que existem diferenças significativas entre o número de armadilhas utilizadas com relação ao uso de inseticida, ao nível de significância de 5%?

17. Um agricultor que planta frutas deseja testar um novo tipo de inseticida, cujo fabricante garante reduzir o prejuízo por danos causados por certo tipo de inseto, consequentemente aumenta a produção. Para verificar essa afirmação do fabricante do inseticida, o agricultor pulveriza 200 árvores com o novo inseticida e 200 árvores com o inseticida normalmente utilizado. O agricultor anotou a produção média por árvore para os dois inseticidas.

a) Formule as hipóteses do teste ( $H_0$  e  $H_1$ ), em termos de médias populacionais.

b) Dê a definição de  $\alpha$  para esse experimento.

18. Foi realizado um experimento para avaliar o comportamento "in vitro" da espécie *Mandevilla velutina* (Apocinácea), provenientes de duas regiões diferentes: Cerrado e Restinga. Após isolar os explantes, com um nó com duas gemas axilares, obtidos das plantas matrizes, foi instalado o experimento com delineamento inteiramente casualizado com 20 repetições (20 explantes para o Cerrado e 20 para a Restinga), portanto, temos um total de 40 unidades experimentais. Esse experimento produziu um desvio padrão  $s = 1,5611$  com 38 graus de liberdade. A variável utilizada foi a altura em cm dos explantes de *Mandevilla* cultivadas "in vitro" durante 45 dias, cujos resultados foram:

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

a) Teste a hipótese de que não há diferença entre as duas regiões, ao nível de significância de 5%, para altura de explantes de *Mandevilla*.

b) Construa o intervalo de confiança de 95%, para a diferença das médias populacionais.

19. Foi desenvolvido um estudo para verificar a qualidade dos vinhos em um Estado. Uma propriedade física avaliada nesse estudo foi a estabilidade dos vinhos, dada em 4 categorias:

péssima, problemática, regular e aceitável. Numa amostra de 188 garrafas de vinho, foram encontrados os seguintes resultados:

Vinho	Estabilidade				Total
	Péssima	Problemática	Regular	Aceitável	
Branco	8	29	28	7	72
Rosado	10	22	10	3	45
Tinto	29	21	13	8	71
Total	47	72	51	18	188

- a) Teste a hipótese de que não há relação (dependência) entre tipos de vinhos e estabilidade.  
b) Teste a hipótese de que os vinhos tinto e branco são independentes da estabilidade. Qual dos dois vinhos é pior quanto à estabilidade?

20. Nitrogênio é o elemento mais comum aplicado no solo. Em regiões tropicais, apenas uma parte do nitrogênio aplicado é aproveitado pelas culturas. Informação sobre  $\mu$ , a porcentagem média de nitrogênio perdido, é importante para pesquisas sobre as condições ótimas de crescimento das plantas. Os dados a seguir descritos, representam a quantidade de nitrogênio perdido (dado em porcentagem do total de nitrogênio aplicado):

10,8	13,5	11,8	9,0	14,7
10,5	8,0	10,0	9,8	10,3
14,0	9,5	8,7	13,8	12,8

- a) Faça o teste de hipótese, ao nível de significância de 5%, para verificar se os dados da amostra suportam a hipótese de que a porcentagem média de nitrogênio perdido ( $\mu$ ) é menor do que 13%.  
b) Calcule o *valor p* do teste e interprete.

21. Para verificar a efetividade de um novo tratamento contra infestação de pulgões que atacam as folhas das plantas, em 100 plantas atacadas e tratadas com o novo inseticida, foram encontradas 9 plantas com pulgões depois de uma semana do tratamento. Desejamos saber se os resultados observados justificam a afirmação de que menos de 15% da população de plantas tratadas terão infestação de pulgões. Utilize um nível de significância de 5%. Calcule o *valor p* e interprete.

22. Sabe-se que o mar na Baía Norte, em Florianópolis, é poluído. Os mexilhões do gênero *Perna perna* são animais filtradores, ou seja, alimentam-se de matéria orgânica dissolvida na água. Para mexilhões cultivados em ambiente sem poluição, neste estudo é o local de Sambaqui, admite-se que o peso médio de mexilhões é igual a 30 g. Os valores medidos em



20 mexilhões na Baía Norte foram:

26	26,5	27	28	28,5	29	29,4	31	32	32,5
33,5	34	35	36,4	37,6	38,5	38,7	41	40,6	41,6

Verifique, através de um teste de hipótese, se realmente o peso dos mexilhões na Baía Norte é maior do que em Sambaqui. (Use  $\alpha = 0,05$ ).

Estime, através da construção do intervalo de confiança com 95%, o peso médio populacional de mexilhões na Baía Norte.

23. Queremos comparar dois métodos de poda em plantas de erva-mate, denominados de  $A$  e  $B$ . Vamos considerar que dispomos de 40 árvores de erveiras numa mata nativa. Considere duas maneiras no planejamento deste experimento:

1. sorteamos 20 árvores aleatoriamente para receber o método  $A$ , e as restantes receberam o método  $B$ . Após medimos a produção de massa verde de todas as árvores.
  2. medimos, inicialmente, os diâmetros dos caules das 40 árvores. Em seguida, numeramos as árvores, em ordem crescente de seus diâmetros, de 1 até 40. Consideremos os 20 pares formados: (1;2), (3;4), (5;6),..., (39;40), e de cada par sorteamos uma árvore para receber o método  $A$ , a outra árvore do par recebe o método  $B$ . Obtemos, assim, duas amostras de 20 árvores cada. Após medimos a produção de massa verde de todas as árvores.
- Qual o teste estatístico de hipóteses em cada caso?
  - Qual dos dois procedimentos você preferiria? Por quê?

24. Foi realizado um estudo para verificar a influência da incidência solar sobre a produção de espiguetas nas gramíneas da espécie *Paspalum notatum* Flüggé, conhecida como grama batatais. Levantou-se a hipótese de que a reprodução sexuada da espécie é prejudicada em áreas com menor luminosidade. Com esta finalidade efetuou-se a contagem das espiguetas produzidas pelas plantas em dois locais, quais sejam: adjacentes ao sol e a sombra leve. Os resultados de 20 amostras (plantas) por local foram:

Ambiente iluminado					Ambiente sombra				
44	54	60	60	70	44	47	52	55	57
75	79	80	81	83	59	60	61	62	63
86	88	88	89	90	64	66	67	67	68
91	95	101	107	108	68	69	71	73	76

Teste a hipótese levantada de que a média de espiguetas em local sombreado é menor do que a média em local adjacente ao sol. Use nível de significância de 5%.

25. Se a proporção de nascimentos de filhotes machos da tartaruga *Caretta caretta* for maior do que 0,50 pode indicar algum desequilíbrio ambiental, como, por exemplo, sombreamento da praia. Teste a hipótese de que a proporção de nascimentos de filhotes machos é maior do que 50%, ao nível de significância de 5%, dado que em 200 nascimentos observou-se 120 filhotes machos.

26. Uma certa região florestal foi dividida em 109 quadrados para estudar a distribuição de *Primula simenses* selvagem. A priori, supomos que este tipo distribui-se aleatoriamente na região. A tabela abaixo, extraída de Bussab & Morettin (1987), indica o número de quadrados com  $X$  *Primula Simenses*; o número médio de plantas por quadrado foi de 2,2.

X plantas por quadrado	Número de quadrados com X plantas
0	26
1	21
2	23
3	14
4	11
5	4
6	5
7	4
8	1
> 8	0
Total	109

Teste a hipótese de que as frequências observadas são próximas das esperadas utilizando um modelo de Poisson, ou seja, que a distribuição espacial de *Primula simmensis* é aleatória, ao nível de significância de 5%.

## Bibliografia

- ALBERTS,B., BRAY,D., JOHNSON,A., LEWIS,J., RAFF,M., ROBERTS,K., WALTER,P. *Fundamentos da biologia celular. Uma introdução à biologia molecular da célula*. Editora Artes Médicas Sul. Porto Alegre, 1999.
- BARBETTA,P.A. *Estatística aplicada às ciências sociais*. 4a. ed. Florianópolis: Editora da Universidade Federal de Santa Catarina, 2001.
- BART,J., NOTES,W.J., FLINGER,M.A. *Sampling and Statistical Models for Behavioral Ecologists*. United Kingdom:Cambridge University Press, 1998.
- BATSCHELET,E. *Introdução à matemática para biocientistas*. São Paulo: Editora da Universidade de São Paulo, 1978.
- BEIGUELMAN,B. *Curso Prático de Bioestatística*. 5a. ed. Ribeirão Preto - SP: FUNPEC-Editora , 2002.
- BOX,G.E.P., HUNTER,W.G., HUNTER,J.S. *Statistics for experimenters*. Canadá: John Wiley, 1978.
- BROWN,D., ROTHERY,P. *Models in Biology: mathematics, statistics and computing*. New York: John Wiley & Sons, 1994.
- BROWER,J.E., ZAR,J.H. *Fields & Laboratory Methods for General Ecology*. 2. ed. Dubuque, Iowa: WCB, 1984.
- BUSSAB,W.O., MORETTIN,P.A. *Estatística básica*. 4. ed. Coleção Métodos Quantitativos. São Paulo: Editora Atual, 1987.
- CENTENO,A.J. *Curso de estatística aplicada à biologia*. Goiânia: Editora da Universidade Federal de Goiás, 1982.
- FINNEY,D.J. *An introduction to statistical science in agriculture*. New York: John Wiley & Sons, 1953.
- HEATH,O.V.S. *A estatística na pesquisa científica*. v. 1. São Paulo: Editora da USP, 1981.
- JUNIOR,P.S., MACHADO,A.A., ZONTA,E.P. da SILVA,J.B. *Curso de estatística*. v. 1 e 2. Pelotas: Editora Universitária, 1992.

- CULLEN Jr, L., RUDRAN,R., VALLADARES-PADUA,C. *Métodos de estudos em biologia da conservação e manejo da vida silvestre*. Curitiba: Editora da UFPR e Fundação Boticário de Proteção à Natureza, 2003.
- MENDENHALL,N. *Probabilidade e estatística*. v. 1 e 2. Rio de Janeiro: Editora Campos, 1985.
- MONTGOMERY,D.D. *Design and analysis of experiments*. 6. ed. New York: John Wiley & Sons, Inc. 2005.
- MORETTIN,L.G. *Estatística básica: probabilidade*. 7. ed. São Paulo: Makron Books,1999.
- PERES,C.A., SALDIVA,C.D. *Planejamento de experimentos*. São Paulo - SP. 5º Simpósio Nacional de Probabilidade e Estatística.
- SIEGEL,S. *Estatística Não-Paramétrica(Para as Ciências do Comportamento)*. São Paulo.: Editora McGraw-Hill, 1975.
- SNEDECOR,G.W., COCHRAN,W.S. *Statistical methods*. 6. ed. Ames: Iowa State Univ., 1972.
- SOARES,J.F., FARIAS,A.A., CESAR,C.C. *Introdução à estatística*. Rio de Janeiro: Editora Guanabara Koogan S.A., 1991.
- SOARES,J.F., SIQUEIRA,A.L. *Introdução à estatística médica*. 1. ed. Belo Horizonte: Editora da Universidade Federal de Minas Gerais, 1999.
- STEEL,R.G.D., TORRIE,J.H. *Principles and procedures of statistics*. New York: Willey, 1960.
- TUKEY,J.W. *Exploratory data analysis*. Reading, Mass.: Addison-Wesley, 1971.
- VIEIRA,S. *Estatística experimental*. 2. ed. São Paulo: Editora Atlas S.A., 1999.
- WONNACOTT,T.H., WONNACOTT,R.J. *Estatística aplicada à economia e à administração*. 1. ed. Rio de Janeiro: Livros Técnicos e Científicos Editora, 1981.

## Apêndice 1

### Tabela da distribuição binomial

Tabela 1: Distribuição binomial: probabilidade de cada valor  $k$  em função de  $n$  e  $\pi$

$n$	$k$	$\pi$									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156

*continua na próxima página*

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	0	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
	1	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
	2	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
	3	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
	4	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	5	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
	6	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
	7	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
	8	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	0	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
	1	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
	2	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
	3	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
	4	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
	5	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
	6	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
	7	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
	8	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
	9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	0	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
	1	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
	2	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
	3	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
	4	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
	5	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
	6	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
	7	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
	8	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
12	0	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
	1	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
	2	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
	3	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
	4	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
	5	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
	6	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
	7	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
	8	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
	9	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
13	0	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
	1	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
	2	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
	3	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
	4	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
	5	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
	6	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
	7	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
	8	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
	9	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
	10	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
	1	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009

continua na próxima página



(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
	2	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
	3	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
	4	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
	5	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
	6	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
	7	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
	8	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
	9	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
	10	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
15	0	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
	1	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
	2	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
	3	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
	4	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
	5	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916
	6	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
	7	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
	8	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
	9	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
	10	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
	11	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
	15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,97
1	0	0,4500	0,4000	0,3500	0,3000	0,2500	0,2000	0,1500	0,1000	0,0500	0,0300
	1	0,5500	0,6000	0,6500	0,7000	0,7500	0,8000	0,8500	0,9000	0,9500	0,9700
2	0	0,2025	0,1600	0,1225	0,0900	0,0625	0,0400	0,0225	0,0100	0,0025	0,0009
	1	0,4950	0,4800	0,4550	0,4200	0,3750	0,3200	0,2550	0,1800	0,0950	0,0582
	2	0,3025	0,3600	0,4225	0,4900	0,5625	0,6400	0,7225	0,8100	0,9025	0,9409
3	0	0,0911	0,0640	0,0429	0,0270	0,0156	0,0080	0,0034	0,0010	0,0001	0,0000
	1	0,3341	0,2880	0,2389	0,1890	0,1406	0,0960	0,0574	0,0270	0,0071	0,0026
	2	0,4084	0,4320	0,4436	0,4410	0,4219	0,3840	0,3251	0,2430	0,1354	0,0847
	3	0,1664	0,2160	0,2746	0,3430	0,4219	0,5120	0,6141	0,7290	0,8574	0,9127
4	0	0,0410	0,0256	0,0150	0,0081	0,0039	0,0016	0,0005	0,0001	0,0000	0,0000
	1	0,2005	0,1536	0,1115	0,0756	0,0469	0,0256	0,0115	0,0036	0,0005	0,0001
	2	0,3675	0,3456	0,3105	0,2646	0,2109	0,1536	0,0975	0,0486	0,0135	0,0051
	3	0,2995	0,3456	0,3845	0,4116	0,4219	0,4096	0,3685	0,2916	0,1715	0,1095
	4	0,0915	0,1296	0,1785	0,2401	0,3164	0,4096	0,5220	0,6561	0,8145	0,8853
5	0	0,0185	0,0102	0,0053	0,0024	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000
	1	0,1128	0,0768	0,0488	0,0284	0,0146	0,0064	0,0022	0,0005	0,0000	0,0000
	2	0,2757	0,2304	0,1811	0,1323	0,0879	0,0512	0,0244	0,0081	0,0011	0,0003
	3	0,3369	0,3456	0,3364	0,3087	0,2637	0,2048	0,1382	0,0729	0,0214	0,0082
	4	0,2059	0,2592	0,3124	0,3602	0,3955	0,4096	0,3915	0,3281	0,2036	0,1328
	5	0,0503	0,0778	0,1160	0,1681	0,2373	0,3277	0,4437	0,5905	0,7738	0,8587
6	0	0,0083	0,0041	0,0018	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000
	1	0,0609	0,0369	0,0205	0,0102	0,0044	0,0015	0,0004	0,0001	0,0000	0,0000
	2	0,1861	0,1382	0,0951	0,0595	0,0330	0,0154	0,0055	0,0012	0,0001	0,0000

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,97
	3	0,3032	0,2765	0,2355	0,1852	0,1318	0,0819	0,0415	0,0146	0,0021	0,0005
	4	0,2780	0,3110	0,3280	0,3241	0,2966	0,2458	0,1762	0,0984	0,0305	0,0120
	5	0,1359	0,1866	0,2437	0,3025	0,3560	0,3932	0,3993	0,3543	0,2321	0,1546
	6	0,0277	0,0467	0,0754	0,1176	0,1780	0,2621	0,3771	0,5314	0,7351	0,8330
7	0	0,0037	0,0016	0,0006	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0320	0,0172	0,0084	0,0036	0,0013	0,0004	0,0001	0,0000	0,0000	0,0000
	2	0,1172	0,0774	0,0466	0,0250	0,0115	0,0043	0,0012	0,0002	0,0000	0,0000
	3	0,2388	0,1935	0,1442	0,0972	0,0577	0,0287	0,0109	0,0026	0,0002	0,0000
	4	0,2918	0,2903	0,2679	0,2269	0,1730	0,1147	0,0617	0,0230	0,0036	0,0008
	5	0,2140	0,2613	0,2985	0,3177	0,3115	0,2753	0,2097	0,1240	0,0406	0,0162
	6	0,0872	0,1306	0,1848	0,2471	0,3115	0,3670	0,3960	0,3720	0,2573	0,1749
	7	0,0152	0,0280	0,0490	0,0824	0,1335	0,2097	0,3206	0,4783	0,6983	0,8080
8	0	0,0017	0,0007	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0164	0,0079	0,0033	0,0012	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000
	2	0,0703	0,0413	0,0217	0,0100	0,0038	0,0011	0,0002	0,0000	0,0000	0,0000
	3	0,1719	0,1239	0,0808	0,0467	0,0231	0,0092	0,0026	0,0004	0,0000	0,0000
	4	0,2627	0,2322	0,1875	0,1361	0,0865	0,0459	0,0185	0,0046	0,0004	0,0001
	5	0,2568	0,2787	0,2786	0,2541	0,2076	0,1468	0,0839	0,0331	0,0054	0,0013
	6	0,1569	0,2090	0,2587	0,2965	0,3115	0,2936	0,2376	0,1488	0,0515	0,0210
	7	0,0548	0,0896	0,1373	0,1977	0,2670	0,3355	0,3847	0,3826	0,2793	0,1939
	8	0,0084	0,0168	0,0319	0,0576	0,1001	0,1678	0,2725	0,4305	0,6634	0,7837
9	0	0,0008	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0083	0,0035	0,0013	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0407	0,0212	0,0098	0,0039	0,0012	0,0003	0,0000	0,0000	0,0000	0,0000
	3	0,1160	0,0743	0,0424	0,0210	0,0087	0,0028	0,0006	0,0001	0,0000	0,0000
	4	0,2128	0,1672	0,1181	0,0735	0,0389	0,0165	0,0050	0,0008	0,0000	0,0000
	5	0,2600	0,2508	0,2194	0,1715	0,1168	0,0661	0,0283	0,0074	0,0006	0,0001
	6	0,2119	0,2508	0,2716	0,2668	0,2336	0,1762	0,1069	0,0446	0,0077	0,0019

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,97
	7	0,1110	0,1612	0,2162	0,2668	0,3003	0,3020	0,2597	0,1722	0,0629	0,0262
	8	0,0339	0,0605	0,1004	0,1556	0,2253	0,3020	0,3679	0,3874	0,2985	0,2116
	9	0,0046	0,0101	0,0207	0,0404	0,0751	0,1342	0,2316	0,3874	0,6302	0,7602
10	0	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0042	0,0016	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0229	0,0106	0,0043	0,0014	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000
	3	0,0746	0,0425	0,0212	0,0090	0,0031	0,0008	0,0001	0,0000	0,0000	0,0000
	4	0,1596	0,1115	0,0689	0,0368	0,0162	0,0055	0,0012	0,0001	0,0000	0,0000
	5	0,2340	0,2007	0,1536	0,1029	0,0584	0,0264	0,0085	0,0015	0,0001	0,0000
	6	0,2384	0,2508	0,2377	0,2001	0,1460	0,0881	0,0401	0,0112	0,0010	0,0001
	7	0,1665	0,2150	0,2522	0,2668	0,2503	0,2013	0,1298	0,0574	0,0105	0,0026
	8	0,0763	0,1209	0,1757	0,2335	0,2816	0,3020	0,2759	0,1937	0,0746	0,0317
	9	0,0207	0,0403	0,0725	0,1211	0,1877	0,2684	0,3474	0,3874	0,3151	0,2281
	10	0,0025	0,0060	0,0135	0,0282	0,0563	0,1074	0,1969	0,3487	0,5987	0,7374
11	0	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0021	0,0007	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0126	0,0052	0,0018	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0462	0,0234	0,0102	0,0037	0,0011	0,0002	0,0000	0,0000	0,0000	0,0000
	4	0,1128	0,0701	0,0379	0,0173	0,0064	0,0017	0,0003	0,0000	0,0000	0,0000
	5	0,1931	0,1471	0,0985	0,0566	0,0268	0,0097	0,0023	0,0003	0,0000	0,0000
	6	0,2360	0,2207	0,1830	0,1321	0,0803	0,0388	0,0132	0,0025	0,0001	0,0000
	7	0,2060	0,2365	0,2428	0,2201	0,1721	0,1107	0,0536	0,0158	0,0014	0,0002
	8	0,1259	0,1774	0,2254	0,2568	0,2581	0,2215	0,1517	0,0710	0,0137	0,0035
	9	0,0513	0,0887	0,1395	0,1998	0,2581	0,2953	0,2866	0,2131	0,0867	0,0376
	10	0,0125	0,0266	0,0518	0,0932	0,1549	0,2362	0,3248	0,3835	0,3293	0,2433
	11	0,0014	0,0036	0,0088	0,0198	0,0422	0,0859	0,1673	0,3138	0,5688	0,7153
12	0	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,97
	2	0,0068	0,0025	0,0008	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0277	0,0125	0,0048	0,0015	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000
	4	0,0762	0,0420	0,0199	0,0078	0,0024	0,0005	0,0001	0,0000	0,0000	0,0000
	5	0,1489	0,1009	0,0591	0,0291	0,0115	0,0033	0,0006	0,0000	0,0000	0,0000
	6	0,2124	0,1766	0,1281	0,0792	0,0401	0,0155	0,0040	0,0005	0,0000	0,0000
	7	0,2225	0,2270	0,2039	0,1585	0,1032	0,0532	0,0193	0,0038	0,0002	0,0000
	8	0,1700	0,2128	0,2367	0,2311	0,1936	0,1329	0,0683	0,0213	0,0021	0,0003
	9	0,0923	0,1419	0,1954	0,2397	0,2581	0,2362	0,1720	0,0852	0,0173	0,0045
	10	0,0339	0,0639	0,1088	0,1678	0,2323	0,2835	0,2924	0,2301	0,0988	0,0438
	11	0,0075	0,0174	0,0368	0,0712	0,1267	0,2062	0,3012	0,3766	0,3413	0,2575
	12	0,0008	0,0022	0,0057	0,0138	0,0317	0,0687	0,1422	0,2824	0,5404	0,6938
13	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0036	0,0012	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0162	0,0065	0,0022	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0495	0,0243	0,0101	0,0034	0,0009	0,0001	0,0000	0,0000	0,0000	0,0000
	5	0,1089	0,0656	0,0336	0,0142	0,0047	0,0011	0,0001	0,0000	0,0000	0,0000
	6	0,1775	0,1312	0,0833	0,0442	0,0186	0,0058	0,0011	0,0001	0,0000	0,0000
	7	0,2169	0,1968	0,1546	0,1030	0,0559	0,0230	0,0063	0,0008	0,0000	0,0000
	8	0,1989	0,2214	0,2154	0,1803	0,1258	0,0691	0,0266	0,0055	0,0003	0,0000
	9	0,1350	0,1845	0,2222	0,2337	0,2097	0,1535	0,0838	0,0277	0,0028	0,0004
	10	0,0660	0,1107	0,1651	0,2181	0,2517	0,2457	0,1900	0,0997	0,0214	0,0057
	11	0,0220	0,0453	0,0836	0,1388	0,2059	0,2680	0,2937	0,2448	0,1109	0,0502
	12	0,0045	0,0113	0,0259	0,0540	0,1029	0,1787	0,2774	0,3672	0,3512	0,2706
	13	0,0004	0,0013	0,0037	0,0097	0,0238	0,0550	0,1209	0,2542	0,5133	0,6730
14	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0002	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0019	0,0005	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0093	0,0033	0,0010	0,0002	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

continua na próxima página

(Probabilidades binomiais) continuação da página anterior

$n$	$k$	$\pi$									
		0,55	0,60	0,65	0,70	0,75	0,80	0,85	0,90	0,95	0,97
14	4	0,0312	0,0136	0,0049	0,0014	0,0003	0,0000	0,0000	0,0000	0,0000	0,0000
	5	0,0762	0,0408	0,0183	0,0066	0,0018	0,0003	0,0000	0,0000	0,0000	0,0000
	6	0,1398	0,0918	0,0510	0,0232	0,0082	0,0020	0,0003	0,0000	0,0000	0,0000
	7	0,1952	0,1574	0,1082	0,0618	0,0280	0,0092	0,0019	0,0002	0,0000	0,0000
	8	0,2088	0,2066	0,1759	0,1262	0,0734	0,0322	0,0093	0,0013	0,0000	0,0000
	9	0,1701	0,2066	0,2178	0,1963	0,1468	0,0860	0,0352	0,0078	0,0004	0,0000
	10	0,1040	0,1549	0,2022	0,2290	0,2202	0,1720	0,0998	0,0349	0,0037	0,0006
	11	0,0462	0,0845	0,1366	0,1943	0,2402	0,2501	0,2056	0,1142	0,0259	0,0070
	12	0,0141	0,0317	0,0634	0,1134	0,1802	0,2501	0,2912	0,2570	0,1229	0,0568
	13	0,0027	0,0073	0,0181	0,0407	0,0832	0,1539	0,2539	0,3559	0,3593	0,2827
	14	0,0002	0,0008	0,0024	0,0068	0,0178	0,0440	0,1028	0,2288	0,4877	0,6528
15	0	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	1	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	2	0,0010	0,0003	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	3	0,0052	0,0016	0,0004	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
	4	0,0191	0,0074	0,0024	0,0006	0,0001	0,0000	0,0000	0,0000	0,0000	0,0000
	5	0,0515	0,0245	0,0096	0,0030	0,0007	0,0001	0,0000	0,0000	0,0000	0,0000
	6	0,1048	0,0612	0,0298	0,0116	0,0034	0,0007	0,0001	0,0000	0,0000	0,0000
	7	0,1647	0,1181	0,0710	0,0348	0,0131	0,0035	0,0005	0,0000	0,0000	0,0000
	8	0,2013	0,1771	0,1319	0,0811	0,0393	0,0138	0,0030	0,0003	0,0000	0,0000
	9	0,1914	0,2066	0,1906	0,1472	0,0917	0,0430	0,0132	0,0019	0,0000	0,0000
	10	0,1404	0,1859	0,2123	0,2061	0,1651	0,1032	0,0449	0,0105	0,0006	0,0001
	11	0,0780	0,1268	0,1792	0,2186	0,2252	0,1876	0,1156	0,0428	0,0049	0,0008
	12	0,0318	0,0634	0,1110	0,1700	0,2252	0,2501	0,2184	0,1285	0,0307	0,0085
	13	0,0090	0,0219	0,0476	0,0916	0,1559	0,2309	0,2856	0,2669	0,1348	0,0636
	14	0,0016	0,0047	0,0126	0,0305	0,0668	0,1319	0,2312	0,3432	0,3658	0,2938
	15	0,0001	0,0005	0,0016	0,0047	0,0134	0,0352	0,0874	0,2059	0,4633	0,6333

## Apêndice 2

### Tabela da distribuição de Poisson

Tabela 3: Distribuição de Poisson: probabilidade de cada valor  $k$  em função de  $\lambda$

$k$	$\lambda$												
	0,005	0,01	0,015	0,02	0,025	0,03	0,035	0,04	0,045	0,05	0,055	0,06	0,07
0	0,9950	0,9900	0,9851	0,9802	0,9753	0,9704	0,9656	0,9608	0,9560	0,9512	0,9465	0,9418	0,9324
1	0,0050	0,0099	0,0148	0,0196	0,0244	0,0291	0,0338	0,0384	0,0430	0,0476	0,0521	0,0565	0,0653
2	0,0000	0,0000	0,0001	0,0002	0,0003	0,0004	0,0006	0,0008	0,0010	0,0012	0,0014	0,0017	0,0023
3	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

$k$	$\lambda$												
	0,09	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	1,2	1,4
0	0,9139	0,9048	0,8187	0,7408	0,6703	0,6065	0,5488	0,4966	0,4493	0,4066	0,3679	0,3012	0,2466
1	0,0823	0,0905	0,1637	0,2222	0,2681	0,3033	0,3293	0,3476	0,3595	0,3659	0,3679	0,3614	0,3452
2	0,0037	0,0045	0,0164	0,0333	0,0536	0,0758	0,0988	0,1217	0,1438	0,1647	0,1839	0,2169	0,2417
3	0,0001	0,0002	0,0011	0,0033	0,0072	0,0126	0,0198	0,0284	0,0383	0,0494	0,0613	0,0867	0,1128
4	0,0000	0,0000	0,0001	0,0003	0,0007	0,0016	0,0030	0,0050	0,0077	0,0111	0,0153	0,0260	0,0395
5	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0004	0,0007	0,0012	0,0020	0,0031	0,0062	0,0111
6	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0003	0,0005	0,0012	0,0026
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005
8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
9	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

$k$	$\lambda$												
	1,8	2	2,5	3	3,5	4	4,5	5	6	7	8	9	10
0	0,1653	0,1353	0,0821	0,0498	0,0302	0,0183	0,0111	0,0067	0,0025	0,0009	0,0003	0,0001	0,0000
1	0,2975	0,2707	0,2052	0,1494	0,1057	0,0733	0,0500	0,0337	0,0149	0,0064	0,0027	0,0011	0,0005
2	0,2678	0,2707	0,2565	0,2240	0,1850	0,1465	0,1125	0,0842	0,0446	0,0223	0,0107	0,0050	0,0023
3	0,1607	0,1804	0,2138	0,2240	0,2158	0,1954	0,1687	0,1404	0,0892	0,0521	0,0286	0,0150	0,0076
4	0,0723	0,0902	0,1336	0,1680	0,1888	0,1954	0,1898	0,1755	0,1339	0,0912	0,0573	0,0337	0,0189
5	0,0260	0,0361	0,0668	0,1008	0,1322	0,1563	0,1708	0,1755	0,1606	0,1277	0,0916	0,0607	0,0378
6	0,0078	0,0120	0,0278	0,0504	0,0771	0,1042	0,1281	0,1462	0,1606	0,1490	0,1221	0,0911	0,0631
7	0,0020	0,0034	0,0099	0,0216	0,0385	0,0595	0,0824	0,1044	0,1377	0,1490	0,1396	0,1171	0,0901
8	0,0005	0,0009	0,0031	0,0081	0,0169	0,0298	0,0463	0,0653	0,1033	0,1304	0,1396	0,1318	0,1126
9	0,0001	0,0002	0,0009	0,0027	0,0066	0,0132	0,0232	0,0363	0,0688	0,1014	0,1241	0,1318	0,1251
10	0,0000	0,0000	0,0002	0,0008	0,0023	0,0053	0,0104	0,0181	0,0413	0,0710	0,0993	0,1186	0,1251
11	0,0000	0,0000	0,0000	0,0002	0,0007	0,0019	0,0043	0,0082	0,0225	0,0452	0,0722	0,0970	0,1137
12	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0034	0,0113	0,0263	0,0481	0,0728	0,0948
13	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0013	0,0052	0,0142	0,0296	0,0504	0,0729
14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0005	0,0022	0,0071	0,0169	0,0324	0,0521
15	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009	0,0033	0,0090	0,0194	0,0347
16	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0045	0,0109	0,0217
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0021	0,0058	0,0128
18	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0029	0,0071
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0037
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0006	0,0019
21	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0009
22	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004
23	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002

*continua na próxima página*

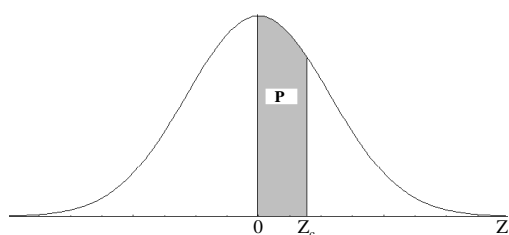
<i>(Probabilidades de Poisson) continuação da página anterior</i>													
$k$	$\lambda$												
	1,8	2	2,5	3	3,5	4	4,5	5	6	7	8	9	10
24	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000



### Apêndice 3

#### Tabela da distribuição normal padrão

Tabela 4: Distribuição normal padrão.  $P$  é a probabilidade da variável aleatória  $Z$  estar entre 0 e  $Z_c$ , ou seja,  $Prob(0 \leq Z \leq Z_c) = P$



$z$	Segunda decimal dos valores de $z$									
	0	1	2	3	4	5	6	7	8	9
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	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545

*continua na próxima página*

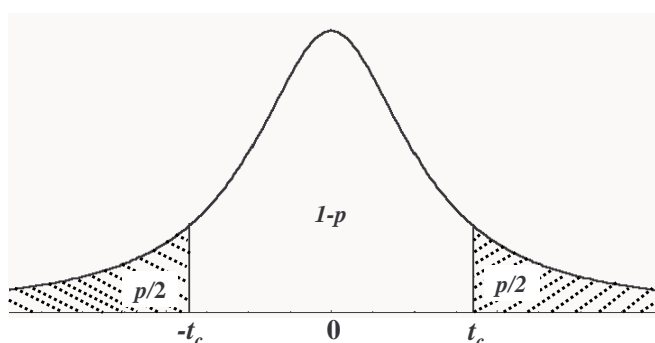
(Probabilidades da normal padrão) continuação da página anterior

$z$	Segunda decimal dos valores de $z$									
	0	1	2	3	4	5	6	7	8	9
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
3,1	0,4990	0,4991	0,4991	0,4991	0,4992	0,4992	0,4992	0,4992	0,4993	0,4993
3,2	0,4993	0,4993	0,4994	0,4994	0,4994	0,4994	0,4994	0,4995	0,4995	0,4995
3,3	0,4995	0,4995	0,4995	0,4996	0,4996	0,4996	0,4996	0,4996	0,4996	0,4997
3,4	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4997	0,4998
3,5	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998	0,4998
3,6	0,4998	0,4998	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,7	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,8	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999	0,4999
3,9	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000
4,0	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000	0,5000

## Apêndice 4

### Tabela da distribuição $t$ de Student

Tabela 5: Distribuição  $t$  de Student. Valores críticos de  $t$  tais que a probabilidade da variável aleatória  $t$  estar entre  $-t_c$  e  $t_c$  vale  $1-p$ , ou seja,  $P(-t_c \leq t \leq t_c) = 1 - p$



Graus de liberdade	Valores de $p$								
	0,30	0,20	0,10	0,05	0,04	0,02	0,01	0,002	0,001
1	1,963	3,078	6,314	12,706	15,894	31,821	63,656	318,289	636,578
2	1,386	1,886	2,920	4,303	4,849	6,965	9,925	22,328	31,600
3	1,250	1,638	2,353	3,182	3,482	4,541	5,841	10,214	12,924
4	1,190	1,533	2,132	2,776	2,999	3,747	4,604	7,173	8,610
5	1,156	1,476	2,015	2,571	2,757	3,365	4,032	5,894	6,869
6	1,134	1,440	1,943	2,447	2,612	3,143	3,707	5,208	5,959
7	1,119	1,415	1,895	2,365	2,517	2,998	3,499	4,785	5,408
8	1,108	1,397	1,860	2,306	2,449	2,896	3,355	4,501	5,041
9	1,100	1,383	1,833	2,262	2,398	2,821	3,250	4,297	4,781
10	1,093	1,372	1,812	2,228	2,359	2,764	3,169	4,144	4,587
11	1,088	1,363	1,796	2,201	2,328	2,718	3,106	4,025	4,437
12	1,083	1,356	1,782	2,179	2,303	2,681	3,055	3,930	4,318
13	1,079	1,350	1,771	2,160	2,282	2,650	3,012	3,852	4,221
14	1,076	1,345	1,761	2,145	2,264	2,624	2,977	3,787	4,140
15	1,074	1,341	1,753	2,131	2,249	2,602	2,947	3,733	4,073

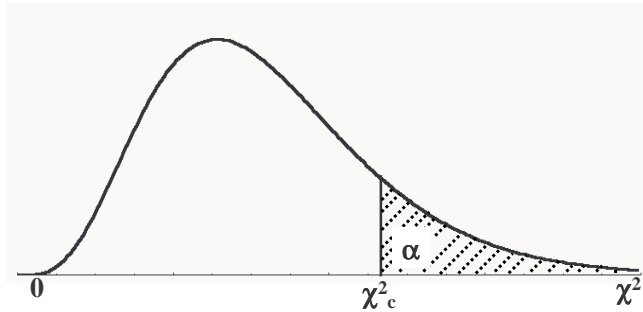
*continua na próxima página*

<i>(Probabilidades da distribuição t de Student) continuação da página anterior</i>									
<i>Graus de liberdade</i>	Valores de <i>p</i>								
	0,30	0,20	0,10	0,05	0,04	0,02	0,01	0,002	0,001
16	1,071	1,337	1,746	2,120	2,235	2,583	2,921	3,686	4,015
17	1,069	1,333	1,740	2,110	2,224	2,567	2,898	3,646	3,965
18	1,067	1,330	1,734	2,101	2,214	2,552	2,878	3,610	3,922
19	1,066	1,328	1,729	2,093	2,205	2,539	2,861	3,579	3,883
20	1,064	1,325	1,725	2,086	2,197	2,528	2,845	3,552	3,850
21	1,063	1,323	1,721	2,080	2,189	2,518	2,831	3,527	3,819
22	1,061	1,321	1,717	2,074	2,183	2,508	2,819	3,505	3,792
23	1,060	1,319	1,714	2,069	2,177	2,500	2,807	3,485	3,768
24	1,059	1,318	1,711	2,064	2,172	2,492	2,797	3,467	3,745
25	1,058	1,316	1,708	2,060	2,167	2,485	2,787	3,450	3,725
26	1,058	1,315	1,706	2,056	2,162	2,479	2,779	3,435	3,707
27	1,057	1,314	1,703	2,052	2,158	2,473	2,771	3,421	3,689
28	1,056	1,313	1,701	2,048	2,154	2,467	2,763	3,408	3,674
29	1,055	1,311	1,699	2,045	2,150	2,462	2,756	3,396	3,660
30	1,055	1,310	1,697	2,042	2,147	2,457	2,750	3,385	3,646
31	1,054	1,309	1,696	2,040	2,144	2,453	2,744	3,375	3,633
32	1,054	1,309	1,694	2,037	2,141	2,449	2,738	3,365	3,622
33	1,053	1,308	1,692	2,035	2,138	2,445	2,733	3,356	3,611
34	1,052	1,307	1,691	2,032	2,136	2,441	2,728	3,348	3,601
35	1,052	1,306	1,690	2,030	2,133	2,438	2,724	3,340	3,591
40	1,050	1,303	1,684	2,021	2,123	2,423	2,704	3,307	3,551
45	1,049	1,301	1,679	2,014	2,115	2,412	2,690	3,281	3,520
50	1,047	1,299	1,676	2,009	2,109	2,403	2,678	3,261	3,496
55	1,046	1,297	1,673	2,004	2,104	2,396	2,668	3,245	3,476
60	1,045	1,296	1,671	2,000	2,099	2,390	2,660	3,232	3,460
100	1,042	1,290	1,660	1,984	2,081	2,364	2,626	3,174	3,390

## Apêndice 5

### Tabela da distribuição Qui-Quadrado ( $\chi^2$ )

Tabela 6: Distribuição de Qui-Quadrado. Valores críticos de qui-quadrado tais que a probabilidade da variável aleatória  $\chi^2$  ser maior do que  $\chi_c^2$  vale  $\alpha$ , ou seja,  $P(\chi^2 \geq \chi_c^2) = \alpha$



Graus de liberdade	Valores de $\alpha$								
	0,30	0,20	0,10	0,05	0,04	0,02	0,01	0,002	0,001
1	1,074	1,642	2,706	3,841	4,218	5,412	6,635	9,549	10,827
2	2,408	3,219	4,605	5,991	6,438	7,824	9,210	12,429	13,815
3	3,665	4,642	6,251	7,815	8,311	9,837	11,345	14,796	16,266
4	4,878	5,989	7,779	9,488	10,026	11,668	13,277	16,923	18,466
5	6,064	7,289	9,236	11,070	11,644	13,388	15,086	18,908	20,515
6	7,231	8,558	10,645	12,592	13,198	15,033	16,812	20,791	22,457
7	8,383	9,803	12,017	14,067	14,703	16,622	18,475	22,601	24,321
8	9,524	11,030	13,362	15,507	16,171	18,168	20,090	24,352	26,124
9	10,656	12,242	14,684	16,919	17,608	19,679	21,666	26,056	27,877
10	11,781	13,442	15,987	18,307	19,021	21,161	23,209	27,721	29,588
11	12,899	14,631	17,275	19,675	20,412	22,618	24,725	29,354	31,264
12	14,011	15,812	18,549	21,026	21,785	24,054	26,217	30,957	32,909
13	15,119	16,985	19,812	22,362	23,142	25,471	27,688	32,536	34,527
14	16,222	18,151	21,064	23,685	24,485	26,873	29,141	34,091	36,124
15	17,322	19,311	22,307	24,996	25,816	28,259	30,578	35,627	37,698

*continua na próxima página*

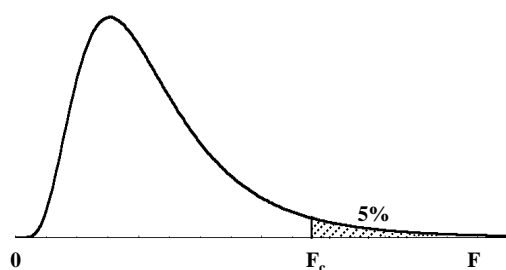
(Probabilidades da distribuição Qui-Quadrado) continuação da página anterior

Graus de liberdade	Valores de $\alpha$								
	0,30	0,20	0,10	0,05	0,04	0,02	0,01	0,002	0,001
16	18,418	20,465	23,542	26,296	27,136	29,633	32,000	37,146	39,252
17	19,511	21,615	24,769	27,587	28,445	30,995	33,409	38,648	40,791
18	20,601	22,760	25,989	28,869	29,745	32,346	34,805	40,136	42,312
19	21,689	23,900	27,204	30,144	31,037	33,687	36,191	41,610	43,819
20	22,775	25,038	28,412	31,410	32,321	35,020	37,566	43,072	45,314
21	23,858	26,171	29,615	32,671	33,597	36,343	38,932	44,522	46,796
22	24,939	27,301	30,813	33,924	34,867	37,659	40,289	45,961	48,268
23	26,018	28,429	32,007	35,172	36,131	38,968	41,638	47,392	49,728
24	27,096	29,553	33,196	36,415	37,389	40,270	42,980	48,811	51,179
25	28,172	30,675	34,382	37,652	38,642	41,566	44,314	50,223	52,619
26	29,246	31,795	35,563	38,885	39,889	42,856	45,642	51,627	54,051
27	30,319	32,912	36,741	40,113	41,132	44,140	46,963	53,022	55,475
28	31,391	34,027	37,916	41,337	42,370	45,419	48,278	54,411	56,892
29	32,461	35,139	39,087	42,557	43,604	46,693	49,588	55,792	58,301
30	33,530	36,250	40,256	43,773	44,834	47,962	50,892	57,167	59,702
31	34,598	37,359	41,422	44,985	46,059	49,226	52,191	58,536	61,098
32	35,665	38,466	42,585	46,194	47,282	50,487	53,486	59,899	62,487
33	36,731	39,572	43,745	47,400	48,501	51,743	54,775	61,256	63,869
34	37,795	40,676	44,903	48,602	49,716	52,995	56,061	62,607	65,247
35	38,859	41,778	46,059	49,802	50,928	54,244	57,342	63,955	66,619
40	44,165	47,269	51,805	55,758	56,946	60,436	63,691	70,617	73,403
45	49,452	52,729	57,505	61,656	62,901	66,555	69,957	77,179	80,078
50	54,723	58,164	63,167	67,505	68,804	72,613	76,154	83,656	86,660
55	59,980	63,577	68,796	73,311	74,662	78,619	82,292	90,061	93,167
60	65,226	68,972	74,397	79,082	80,482	84,580	88,379	96,403	99,608

## Apêndice 6

### Tabela da distribuição $F$

Tabela 7: Distribuição  $F$ : valores críticos de  $F$ . A área achurada corresponde a 5% da área total e, representa, a probabilidade de se obter um valor de  $F$  maior do que um valor  $F_{0,05;n_1-1;n_2-1}$ , onde  $n_1 - 1$  e  $n_2 - 1$  são os graus de liberdade do numerador e denominador, respectivamente.



<i>Graus de liberdade do denominador</i>	<i>Graus de liberdade do numerador</i>								
	1	2	3	4	5	6	7	8	9
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59

*continua na próxima página*

(Probabilidades da distribuicao  $F$ ) continuação da página anterior

Graus de liberdade do denominador	Graus de liberdade do numerador								
	1	2	3	4	5	6	7	8	9
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21

continua na próxima página



(Probabilidades da distribuicao *F*) continuação da página anterior

<i>Graus de liberdade do denominador</i>	<i>Graus de liberdade do numerador</i>								
	10	11	12	13	14	15	16	17	18
1	241,88	242,98	243,90	244,69	245,36	245,95	246,47	246,92	247,32
2	19,40	19,40	19,41	19,42	19,42	19,43	19,43	19,44	19,44
3	8,79	8,76	8,74	8,73	8,71	8,70	8,69	8,68	8,67
4	5,96	5,94	5,91	5,89	5,87	5,86	5,84	5,83	5,82
5	4,74	4,70	4,68	4,66	4,64	4,62	4,60	4,59	4,58
6	4,06	4,03	4,00	3,98	3,96	3,94	3,92	3,91	3,90
7	3,64	3,60	3,57	3,55	3,53	3,51	3,49	3,48	3,47
8	3,35	3,31	3,28	3,26	3,24	3,22	3,20	3,19	3,17
9	3,14	3,10	3,07	3,05	3,03	3,01	2,99	2,97	2,96
10	2,98	2,94	2,91	2,89	2,86	2,85	2,83	2,81	2,80
11	2,85	2,82	2,79	2,76	2,74	2,72	2,70	2,69	2,67
12	2,75	2,72	2,69	2,66	2,64	2,62	2,60	2,58	2,57
13	2,67	2,63	2,60	2,58	2,55	2,53	2,51	2,50	2,48
14	2,60	2,57	2,53	2,51	2,48	2,46	2,44	2,43	2,41
15	2,54	2,51	2,48	2,45	2,42	2,40	2,38	2,37	2,35
16	2,49	2,46	2,42	2,40	2,37	2,35	2,33	2,32	2,30
17	2,45	2,41	2,38	2,35	2,33	2,31	2,29	2,27	2,26
18	2,41	2,37	2,34	2,31	2,29	2,27	2,25	2,23	2,22
19	2,38	2,34	2,31	2,28	2,26	2,23	2,21	2,20	2,18
20	2,35	2,31	2,28	2,25	2,22	2,20	2,18	2,17	2,15
21	2,32	2,28	2,25	2,22	2,20	2,18	2,16	2,14	2,12
22	2,30	2,26	2,23	2,20	2,17	2,15	2,13	2,11	2,10
23	2,27	2,24	2,20	2,18	2,15	2,13	2,11	2,09	2,08
24	2,25	2,22	2,18	2,15	2,13	2,11	2,09	2,07	2,05
25	2,24	2,20	2,16	2,14	2,11	2,09	2,07	2,05	2,04
26	2,22	2,18	2,15	2,12	2,09	2,07	2,05	2,03	2,02
27	2,20	2,17	2,13	2,10	2,08	2,06	2,04	2,02	2,00
28	2,19	2,15	2,12	2,09	2,06	2,04	2,02	2,00	1,99
29	2,18	2,14	2,10	2,08	2,05	2,03	2,01	1,99	1,97
30	2,16	2,13	2,09	2,06	2,04	2,01	1,99	1,98	1,96

## Apêndice 7

### Tabela de números aleatórios

Tabela 8: Tabela de números aleatórios

---

27	07	94	02	42	62	53	83	62	79	15	75	04	91	21	38	34	35	59	13	70	84	98	56	25
42	96	93	24	13	96	82	62	25	57	42	29	23	92	63	53	27	27	10	84	00	34	36	09	04
70	95	03	73	48	56	76	22	91	51	46	53	44	31	76	50	65	69	80	34	30	94	67	20	24
27	81	35	60	34	19	50	21	75	27	98	05	90	10	96	97	12	42	06	62	80	29	07	39	73
09	32	50	76	89	91	84	59	49	09	00	65	86	01	09	12	48	02	87	11	04	06	51	75	77
70	12	86	80	61	01	11	82	78	81	56	16	65	62	70	99	72	39	70	14	09	53	19	86	68
15	64	65	71	71	76	24	05	67	27	54	41	17	35	31	05	91	43	96	57	73	54	44	58	35
81	30	32	73	24	83	96	99	50	62	14	08	47	13	27	15	29	46	97	56	11	01	00	66	80
78	75	03	19	52	60	47	72	34	71	36	56	71	37	91	98	32	02	14	07	43	20	97	09	82
60	33	68	60	92	90	66	36	47	78	80	25	95	29	39	30	26	27	97	68	83	19	99	07	10
63	99	46	22	81	49	07	57	58	29	76	61	33	19	60	42	83	53	25	21	75	80	90	35	80
15	02	13	93	86	97	49	62	84	83	08	30	28	23	25	81	71	21	17	83	02	14	47	24	24
51	35	40	08	38	95	61	33	12	21	38	97	55	44	39	73	08	29	24	13	06	45	47	87	03
25	39	96	35	84	10	74	08	78	15	60	12	46	44	34	94	41	67	29	02	45	80	50	66	78
29	84	49	10	84	68	45	53	43	43	85	11	06	45	99	45	93	60	63	83	65	01	85	73	79
82	26	39	87	76	89	15	92	85	99	74	52	78	68	91	48	91	45	91	85	56	66	97	92	92
90	59	29	64	18	21	85	92	60	82	70	29	85	29	15	86	74	63	68	28	23	96	44	78	00
25	59	50	94	53	24	35	86	25	16	19	85	33	55	39	88	18	79	97	41	77	36	96	50	06
92	53	95	44	88	58	29	83	26	53	32	09	97	77	25	22	07	78	72	09	86	87	82	84	37
98	35	54	71	83	74	60	46	51	22	38	12	26	83	34	82	05	51	59	76	10	48	75	74	10
91	99	67	88	34	54	59	76	44	97	29	73	26	72	67	12	29	17	96	73	17	09	66	75	80
62	09	09	30	45	21	10	55	19	83	74	16	08	08	72	85	80	49	82	65	49	89	49	52	60
94	18	43	36	42	66	04	24	03	92	65	21	05	84	23	14	87	38	17	92	73	49	71	35	83
43	75	50	76	77	21	43	07	94	54	84	08	01	83	00	11	53	52	23	16	85	26	42	81	82

---

*continua na próxima página*

---

*continuação da página anterior*

---

14 44 78 26 46 51 23 22 42 70 53 79 42 32 12 20 53 05 32 17 23 96 26 99 29

24 22 31 90 73 75 81 82 42 25 18 42 48 65 79 51 02 68 54 55 90 32 69 29 11  
97 34 87 79 10 61 34 44 80 36 37 53 30 67 26 40 88 80 00 00 67 24 09 72 14  
35 81 40 66 59 27 41 10 16 68 34 78 88 04 56 81 07 96 70 86 07 27 75 50 65  
42 04 38 95 55 86 32 49 52 44 95 77 87 00 45 99 26 07 14 72 89 05 82 05 09  
30 00 99 20 59 96 36 98 21 04 44 53 16 90 08 66 85 49 63 76 95 79 00 72 85

18 82 08 05 46 55 11 13 63 06 62 28 43 26 23 54 26 52 53 42 48 44 41 42 99  
48 57 16 92 57 82 91 96 04 63 43 40 76 23 83 58 65 23 77 76 24 57 88 00 29  
35 36 82 27 32 11 00 32 11 62 36 60 46 76 75 48 78 01 07 57 81 49 29 44 40  
37 22 32 84 16 83 22 82 28 52 97 62 55 02 54 83 53 61 83 04 10 91 04 68 36  
84 38 83 76 85 94 02 05 20 53 92 96 17 73 10 60 08 68 81 18 35 25 77 92 12

56 23 64 29 17 92 70 93 19 94 19 66 42 72 89 26 85 37 92 00 25 09 71 48 40  
09 33 99 94 34 11 34 17 90 53 43 14 10 96 82 49 40 65 45 58 02 36 21 47 42  
31 63 35 04 15 76 95 80 81 30 22 95 72 62 79 83 76 25 32 08 80 41 97 16 76  
40 08 27 39 70 80 04 10 58 96 70 51 29 36 51 89 33 30 07 60 90 74 51 96 92  
21 25 84 89 97 44 70 33 04 19 12 66 72 52 42 49 15 56 21 50 21 99 71 97 10

---

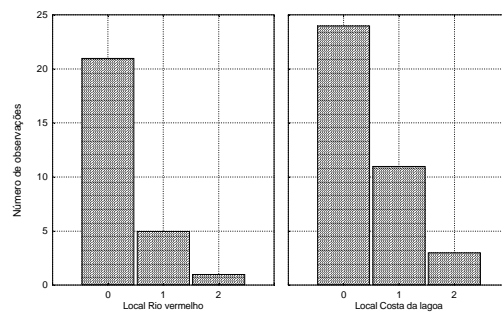


Figura 1: Gráfico da distribuição conjunta de local e aborto

## Respostas da Maioria dos Exercícios

Respostas dos exercícios sobre análise exploratória de dados

Respostas dos exercícios da seção 2.1.3.3

1. a)

Local	Número de abortos			Total
	0	1	2	
1	21	5	1	27
2	24	11	3	38
Total	45	16	4	65

b) Resposta Figura 1.

2. Veja Figura 2.

4. Veja Figura 3.

Respostas dos exercícios da seção 2.1.3.5

2. Sim. Para os gripados, 52,17% não foram vacinados, ou seja, mais da metade. Para os não gripados, 59,22% aplicaram duas doses da vacina, ou seja, a maioria dos não gripados foram vacinados.

3. b)30,29% c)56,50%.

4. Não.

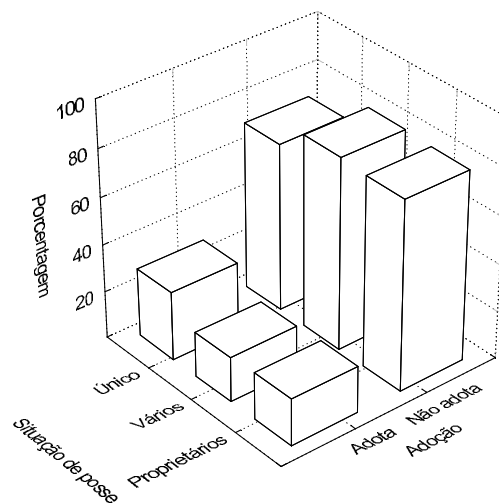


Figura 2: Gráfico da distribuição conjunta da situação de posse e adoção

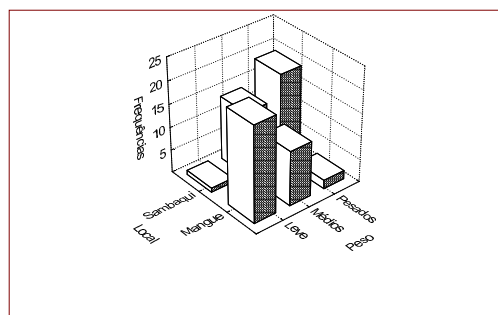


Figura 3: Gráfico da distribuição conjunta do local e peso de mexilhões

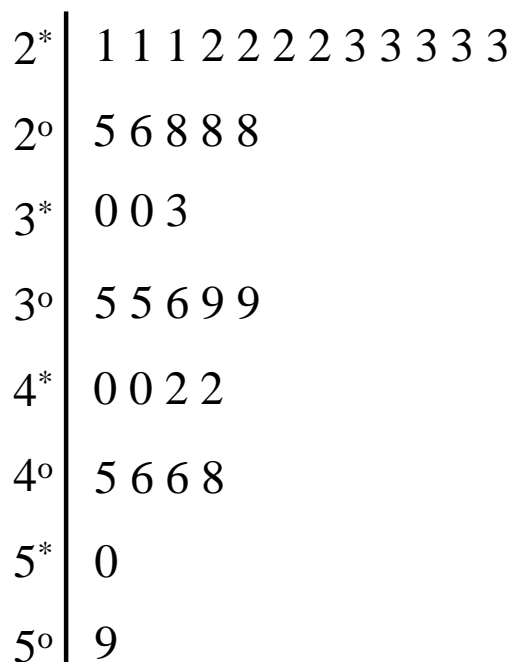


Figura 4: Ramo-e-folhas da duração do primeiro estágio ninfal

6. Sim, pois fixando-se os totais de locais em 100%, verificamos no manguezal que 60% dos mexilhões são leves e somente 5,71% são pesados. Por outro lado, no Sambaqui, verificamos somente 2,94% de leves e 55,88% de pesados. Portanto, no Sambaqui temos animais mais pesados.

#### Respostas dos exercícios da seção 2.1.3.7

1.  $C=0,1567$   $C^*=0,2216$ . O que demonstra que local e aborto estão fracamente associados.
2.  $C = 0,6008$ ;  $C^*=0,8496$ . O que demonstra que porte e hábito estão associados.
4.  $C = 0,5634$ ;  $C^* = 0,7967$ ;  $\chi^2 = 32,0898$ ;  $V = 0,6819$ .

#### Respostas dos exercícios da seção 2.1.3.11

2. Ramo-e-folhas na Figura 4. a) 30 dias; b) Assimétrica à direita; c) Visualmente não apresenta .

#### Respostas dos exercícios da seção 2.2.5.1

1. b)  $r=0,9840$ , portanto existe uma forte dependência entre as duas variáveis.

#### Respostas dos exercícios da seção 2.2.6.1

1. a)  $\hat{y} = 4,777 + 3,780x$ ; b)  $\hat{y} = 19,33 \ 13,32 \ 12,56 \ 15,70 \ 41,10 \ 47,87$ .
3.  $\hat{y} = 0,005 + 0,005x$ ;  $V_{max} = 200 \text{ mmol/min}$  e  $K_M = 1,0 \text{ mM}$ . Os resultados das taxas de transportes são:  $67 \text{ mmol/min}$  e  $198 \text{ mmol/min}$ .

### Respostas dos exercícios propostos sobre modelos de probabilidades para experimentos simples, Seção 3.9

- 1) a)  $\frac{55}{72}$  b)  $\frac{10}{72}$
- 2) 0,16%. Independência entre os resultados dos blocos.
- 3)  $\frac{2}{625} = 0,32\%$ . Idem ao exercício 2.
- 4) a) 0,77 b) 0,54 c) Os eventos são mutuamente exclusivos.
- 5) a) 0,0385 b) 0,0925 c) os pares de cromossomos são independentes.
- 6) i)  $P(A) = 1/2$  ii)  $P(A) = 1/3$ .
- 7) a) 37,5% b)  $\frac{0,10}{0,25} = 0,40 = 40\%$  c) 50%.
- 8) a)  $\frac{5}{8}$  b)  $\frac{7}{8}$  c)  $\frac{3}{4}$ .
- 9) a)  $\frac{3}{8}$  b)  $\frac{7}{8}$  c)  $\frac{5}{8}$  d)  $\frac{1}{2}$  e)  $\frac{1}{8}$
- 10) a)  $\frac{33}{95}$  b)  $\frac{14}{95}$  c)  $\frac{48}{95}$
- 11)  $\frac{4}{9}$
- 12) a) 0,00 b) 0,65 c) 0,70 d) 0,0
- 13) 0,59
- 14) 0,5952
- 15) a) 12,5% b) 0,391% c) 0,0198 d) 85,94%
- 16) 36,23%
- 17) a) 0,6836 b)  $E(X) = 2$
- 18) fazer a tabela
- 19) a) 0,85 b) 0,45 c) 0,10 d) 0,15
- 20) a) Não são independentes.  $P(F | A) \neq P(F)$ , altera as probabilidades. b) 85,71% e c) 78,57%.

### Respostas dos exercícios propostos sobre modelos de probabilidade para variáveis aleatórias discretas, Seção 4.9

- 1) 1,08%
- 2)  $E(X) = np = 156 \times 0,0081 = 1,26$  covas.
- 3) a) 2,236 bactérias/cm<sup>3</sup> b) 0,95957.
- 4)  $E(X)=18$   $DP(X)=2,7$ .
- 5) a) 0,11% b) 0,119%
- 6) a) 0,94 b)60 c) 2,28 u.m. d) $E(X)=0,4$   $VAR(X)=0,392$ .
- 7) a) 4 animais  $VAR(X)=3,2$  animais<sup>2</sup>  $DP(X)=1,8$  animais b) 0,4236 c) 0,931.
- 8) a)  $E(X)=np=2,4$  então, espera-se em média 2,4 machos por ninhada. A proporção média é:  $2,4/5=0,48$ . b) Para  $X=0$  76,  $X=1$  351 e  $X=2$  648,  $X=3$  598,  $X=4$  276,  $X=5$  51.
- 9) a) 0,6454 b) 12 - 27 - 29. c) Concluo que a D. de Poisson está razoavelmente bem adequada a esta situação.
- 10)  $DP(X) = \sqrt{8} = 2,828$ . Desvios de 1, 2 ou 3, ocorrem frequentemente. Desvios maiores que três são raros.
- 11) a)0,2681 b) 0 rato:13 quadrantes 1 rato:5,36 quad. 2 ratos:1,07 quad.
- 12) a)55,54% b)  $100-55,54\%=44,46\%$
- 13) a)=24% b)=94,24%.
- 14) a) Número médio de plantas doentes colhidas = 0,965, logo, proporção média =  $0,965/27 = 0,0357 = 3,57\%$  b)Para  $X=0$  11,24;  $X=1$  11,24;  $X=2$  5,41;  $X=3$  1,67 e  $X=4$  0,36 c)Não é muito boa a aproximação. Acreditamos modelo de Poisson seja melhor.
- 15)  $E(X) = n\pi = 8.(1/2) = 4$ .  $P(X=4)=0,27$ .
- 16)9,56%
- 17) Para 15 minutos o parâmetro (Valor médio) vale 5.  $P(X=5)=0,17546$
- 18) a) 0,1680 b) 0,800852



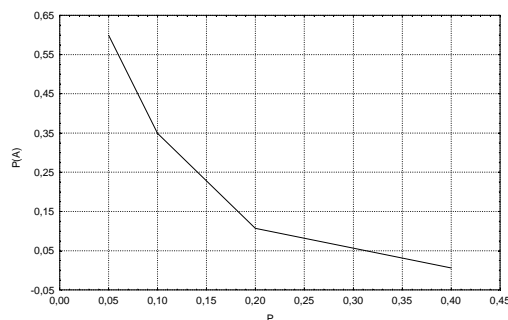


Figura 5: Curva característica de operação (CCO)

19) a)

X	Y				Total
	0	1	2	3	
0	1	0	0	0	1
1	0	2	2	0	4
2	0	2	2	2	6
3	0	2	2	0	4
4	1	0	0	0	1
Total	2	6	6	2	16

b) Não são independentes, pois  $P(X \text{ e } Y)$  não é igual a  $P(X) \cdot P(Y)$  c)  $E(X) = 32/16$ ;  $E(Y) = 24/16$ ; fazendo  $Z = X \cdot Y$  temos que  $E(Z) = 48/16$ , assim  $\text{COV}(X, Y) = 48/16 - (32/16 \cdot 24/16) = -45$ .

20)  $P(X=0) = 0,0824$   $P(X \geq 4) = 0,126$

21) a) 0,107 b) 0,599; 0,349; 0,107 e 0,006. O gráfico é dado na Figura 5.

22) 10,72 u.m., que o preço médio por saco da proposta, logo o produtor deve manter seu preço de 20,00 u.m. por saco.

23) 0,091603.

24) a)  $P(X=k) = \binom{10}{k} 0,70^k 0,30^{n-k}$ . Veja Figura 6. b)  $P(X \geq 2) = 1 - [P(X=0) + P(X=1)] = 1 - 0,00015 = 0,99985$  c)  $E(X) = 7$   $\text{VAR}(X) = 2,1$ .

25) a) 0,0337 b) 0,9933 c)  $P(X=k) = (e^{-5} \cdot 5^k) / k!$ . Veja Figura 7.

26) 15,1%.

27) a) 0,086138 b) 0,736098.

28) a) 0,676676 b) 0,323324.

29) a) 73,57% b) 26,42%.

30) a) 4.000,00 u.m. para o lote b) 3.952,434 u.m. para o lote. Assim, é melhor a opção A.

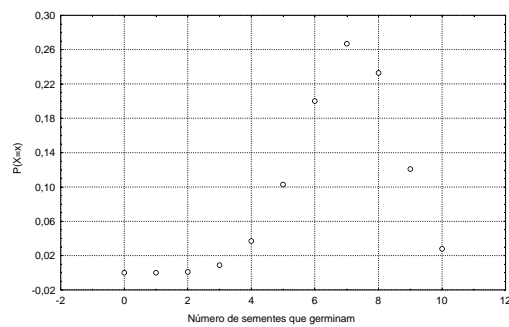


Figura 6: Gráfico da distribuição de probabilidades do número de sementes que germinam

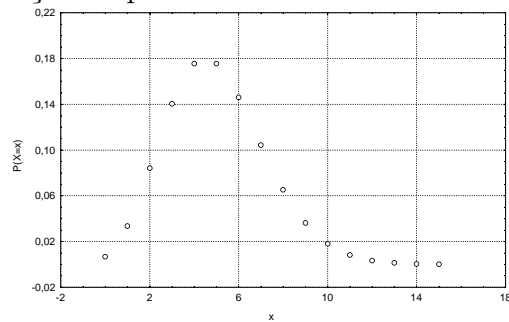


Figura 7: Gráfico da distribuição de probabilidades do número de crianças imunizadas

31) 18,48%

32) a) 11,11% b) 1,333

33) a)

- 1) Existe  $n=8$  repetições do experimento.
- 2) Há independência entre os nascimentos.
- 3) Só há dois tipos de resultados possíveis.
- 4)  $\pi$  constante, em cada nascimento a probabilidade de sucesso é de 50%.

b)  $E(X)=4$   $DP(X)=1,4142$  c) 63,6% d)

34) a) 18,47% b) 1,73.

35) a) 11,53% b) 0,00929% c) 73,45%.

36) a) 23,94%

b)

- 1) Cada dia é independente
- 2) Existe  $n=30$  repetições igualmente prováveis

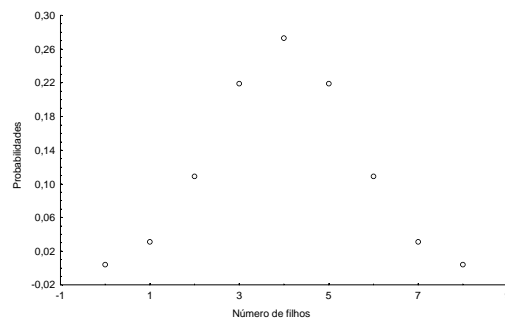


Figura 8: Distribuição de probabilidade para número de filhos homens

- 3)  $\pi$  constante
- 4) Só há 2 resultados possíveis: chove ou não chove.

37) a) Sim; b) 39 c) 2,929 d) 0%

38) a) 17,54% b) 95,95724%.

39) a) 0,5% b) 2,98%.

40) 0,61.

41) a) 96,059% (Binomial) e 0,96078% (Poisson) b) 99,94%.

### Respostas dos exercícios propostos para modelos de probabilidade para variáveis aleatórias contínuas, Seção 5.3.5

1) a) 0,3848 b) 0,2517 c) 0,3092 d) 0,7648 e) 0,2743 f) 0,8907 g) 0,9453 h) 0,68 i) 0,95 j) 0,99.

2) 0,0038 ou 0,38%

3)  $P(X > 185kg) = 0,9332$ . Logo, têm-se  $120 \times 0,9332 = 112$  animais.

4) a) 0,1974 b) 0,6826 c) 0,3085

5) a) 0,0918; 0,9962; 0,7463; 0; b) 8,745 cm<sup>3</sup>/min c) 9,5325 a 14,4675

6) a) 0,9336 b) 93,36 progênicos.

7) 0,047

8) 4,33; 5,54; 6,02

9)  $z = 1,645$  logo  $D = 21,79$  cm

10) a) 0,0228 b) 0,053%

11) 22,06%

12) média = 425,7 g.

- 13) a) 0,01222 b) 0,18673
- 14)  $0,02275 = 2,275\%$
- 15) a)  $b=80,3$  b)  $b=92,53$
- 16) a) 0,2358 b) 15,7% c) 2,507cm
- 17) 15,625
- 18) 75
- 19) a) 15,87% b) 95,45% c) A porcentagem permanece a mesma. Sempre existirá das observações entre  $-2\sigma$  e  $+2\sigma$  da média, se a distribuição é normal padrão.
- 20) a) 0,9336 b) 0,0664 c) 93
- 21) a) 2486 b)  $w=16,52$  c) 0,34
- 22) a) 13,56% b)  $D=17,20$
- 23) a) 0,0968 b) 0,9879 c) 0,5
- 24)  $x_1=0,484$   $x_2=0,516$
- 25) a) 512,85mm b) 0,51%.
- 26) a) 95,45% b) A porcentagem continua sendo a mesma, pois, se a distribuição é normal, então, entre a média  $\pm 2$  desvios padrões, tem-se 95,45% do dados. c) 0,621%
- 27) a) 18,904 b) 68,3%
- 28)  $\mu=21,303$
- 29) 12,30% b)  $A=18,52m$
- 30) a) 0,8413 b) 0,218230 c) 20,569
- 31) 0,10427 ou 10,43%
- 32)  $0,01321=1,32\%$ .
- 33)  $0,1736=17,36\%$ .
- 34) a) 0,04977 b) Conclui-se que a cultivar  $A$  é mais produtiva que a  $B$ .
- 35) 34,15%.

### Respostas dos exercícios propostos sobre distribuições amostrais, Seção 6.5

1. a) 68,3% b)  $0,15866 = 16\%$ .
2. 33,87%.
3.  $Prob(P \leq 0,40) + Prob(P \geq 0,53) = 0,2388 + 0,20327 = 0,4421 = 44,21\%$ .
4. 0,06% [prob.germinar=0,95; prob.de não-germinar=0,05;  $n=200$ ]. Utilizar distribuição normal padrão.

5.  $P(\bar{X} \geq 2,0) = 1,7\%$ .
6. A média da distribuição amostral é igual a 7069 kg/ha e o desvio padrão igual a 98,8583 kg/ha. a) 99,53% b) 0,52% c) 0,06%.
7. Média populacional=18000; desvio padrão populacional=5125; n=30. Resp.0,1423=14,23%.
8.  $P(\bar{X} \leq 140) = 0,035\%$ .
9. a)68% b)100%.
- 10) a)  $z_1=-0,57$   $z_2=0,85$ , assim,  $0,2157+0,3023=0,5180$  ou 51,8%. b) 0,0021 ou 0,21%

### Respostas dos exercícios propostos sobre avaliação de parâmetros (estimação), Seção 7.5

- 1) a)  $li=172,410$   $ls=173,589$  b)  $li=172,226$   $ls=173,774$  c) 9,18%.
- 2) a)  $li=481$   $ls=489$   $li=479,84$   $ls=490,16$  b) 54,34=55 pacotes.
- 3) a)  $li=23,5963$   $ls=29,5037$   $li=22,5694$   $ls=30,5306$  b)  $n=148,52=149$ .
- 4) a)  $li=3338,00$   $ls=3461,92$   $li=3316,09$   $ls=3483,91$  b)  $n=9,585=10$ .
- 5) Para o mês de maio:  $Li=0,0991$   $Ls=0,5009$  erro=0,20 e  $n=127$ . Para o mês de junho:  $Li=0,0116$   $Ls=0,3489$  erro=0,1683 e  $n=89$ .
- 6)  $li=zero$   $ls=0,2413$ .
- 7)  $li=0,054$   $ls=0,3459$  mínimo=54 máximo=346.
- 8)  $li=1377,6$   $ls=1622,4$  kg/ha.
- 9)  $57,16 \leq \mu \leq 95,14$  b)  $n=375$
- 10) Lagoa:  $5,73 \leq \mu \leq 6,51$  Fazenda:  $9,31 \leq \mu \leq 9,81$ .
- 11)  $0,584 \leq \pi \leq 0,816$ .
- 12)  $-0,084 \leq \mu \leq 1,484$ .
- 13) a)  $1,9152 \leq \mu \leq 2,2848$  b)  $n=106,5 \sim 107$ .
- 14)  $7,5076 \leq \mu_{Bra} \leq 12,6524$   $11,5418 \leq \mu_{Can} \leq 19,1182$ . Concluir que não há diferenças significativas entre as duas espécies.
- 15) a)  $0,1864 \leq \pi \leq 0,6136$ . Baixa precisão, aumentar o tamanho da amostra  $n$ . b)  $e = 0,2136$ , temos 99% de confiança que a proporção obtida na amostra não difere mais do que 0,2136 de  $\pi$ . c)  $n = 640$  d)  $0,041 \leq \pi \leq 0,081$ . e) veja Figura 9.
- 16) a)  $8,7168 \leq \mu \leq 15,0632$  b)  $n \cong 855$

### Respostas dos exercícios propostos sobre testes de hipóteses, Seção 8.5

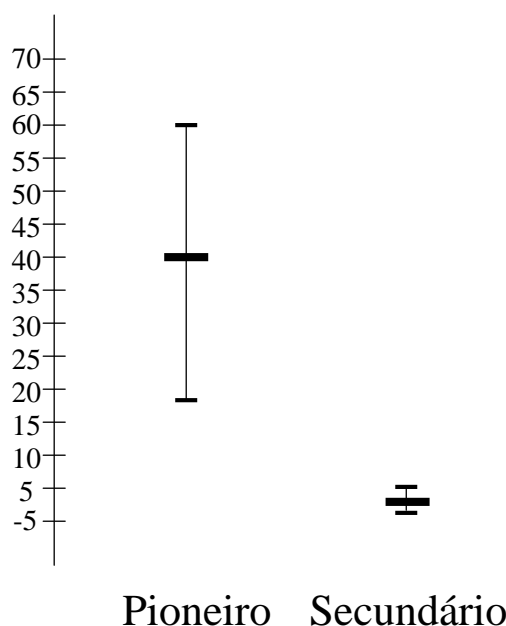


Figura 9: Representação gráfica dos intervalos de confiança

- 1)  $Z_c=0,329$   $Z_t=1,645$  - conclui-se que a máquina não precisa ser regulada
2.  $g.l. = 34,52 \approx 35$ ,  $T_t=2,724$  —  $T_c=6,98$  Rejeita-se  $H_0$ . A variedade B é superior.
3.  $\chi^2 = 9,27$ ,  $\chi_{0,05;3}^2 = 7,815$ , portanto, a distribuição não segue as leis da hereditariedade.
4. 1º),  $X_A \sim N(\mu_A, \sigma_A^2)$  e  $X_B \sim N(\mu_B, \sigma_B^2)$ ; 2º),  $\sigma_A^2 = \sigma_B^2$  e 3º) as amostras são independentes.
5.  $t = 0,2479$ . Aceita-se a hipótese nula.  $48,25 \leq \mu \leq 54,11$
6.  $30,6097 \leq \mu \leq 33,8904$ ;  $18,2972 \leq \mu \leq 22,3427$  b)  $n = 31$  c)  $t=12,623$ , rejeita-se a hipótese nula.
7.  $\chi^2 = 3,8347$ . Aceita-se a hipótese nula. As duas variedades são semelhantes quanto à germinação de sementes.
8.  $t = 4,333$ . Rejeita-se a hipótese nula, o inseticida novo é melhor do que o padrão.
9.  $t = -1,598$ , aceita-se a hipótese nula, a quantidade de proteínas no plasma é a mesma.
10. a)  $5,9373 \leq \mu \leq 7,8227$ . b)  $t = -20,2482$ , portanto, rejeitamos a hipótese nula. c)  $n = 36,27 \cong 37$ .
11.  $t = 10,588$ ,  $P(t > 10,588) < 0,1\%$ . Rejeita-se a hipótese nula.
12.  $P(F > 3,02) = 0,057602$ , portanto, aceita-se  $H_0$ .  $t = 0,24$ ,  $P(|t| > 0,24) = 0,811620$ , portanto, aceita-se a hipótese nula.
13.  $F=4,974$ ,  $P(F > 4,974) = 0,000022$ , assim, as variâncias são heterogêneas.  $t = 3,8977$ ,  $P(t > 3,8977) = 0,000181$ , existe diferença significativa entre as alturas das duas espécies.

14.  $F = 1,006$ ,  $P(F > 1,006) = 0,494867$ , variâncias homogêneas.  $t = 0,05417$ ,  $P(|t| > 0,05417) = 0,957084 = 95,71\%$ , aceita-se  $H_0$ .
15.  $t = -4,149$ ,  $P(|t| < -4,149) = 0,025441 = 2,54\%$ . A cultivar CEP 18-Cavera é mais produtiva.
16.  $\chi^2 = 164,62$ , com 2 graus de liberdade,  $P(\chi^2 > 164,62) = 0,000000$ . Existe diferença significativa entre o número de armadilhas utilizadas na captura da vaquinha com relação ao uso de inseticida.
17. a)  $H_0 : \mu_N = \mu_U$  e  $H_0 : \mu_N > \mu_U$ . b) É a probabilidade do agricultor aceitar como melhor o inseticida novo, sendo que, na verdade, não existe diferença entre os dois inseticidas, quanto à produção.
18. a)  $t = -0,688$ , com 38 graus de liberdade,  $P(|t| < -0,688) = 0,495631 = 49,56\%$ , aceita-se a hipótese nula. b)  $-0,34 - 2,021(0,49366) \leq \mu_C - \mu_R \leq -0,34 + 2,021(0,49366)$ .
19.  $\chi^2 = 22,352$ , com 6 graus de liberdade,  $P(\chi^2 > 22,352) = 0,001047 = 0,1047\%$ , existe diferenças entre os três tipos de vinhos quanto à estabilidade. b)  $\chi^2 = 18,755$ , com 3 graus de liberdade,  $P(\chi^2 > 18,755) = 0,000308 = 0,0308\%$ , existe diferenças entre os dois tipos de vinhos quanto à estabilidade. O vinho tinto é pior que o branco.
20. a)  $t(\text{calculado}) = -3,35$ ,  $t(\text{tabelado}) = -1,761$ , rejeita-se a hipótese de nulidade. b)  $\text{valor } p = 0,002382$ , tomando-se como referência o valor 0,05, rejeita-se a hipótese nula.
21. a)  $z = -1,68$ , Rejeita-se  $H_0$ . b)  $\text{valor } p = 0,046479$ , rejeita-se  $H_0$ .
22.  $t_{19} = 2,91$   $P(t \geq 2,91)$  está entre 0,1% e 0,5%, portanto, rejeita-se a hipótese de nulidade e conclui-se que a média na Baía Norte é realmente maior do que em Sambaqui. O intervalo fica:  $31,0 \leq \mu \leq 35,8$
23. a) no caso 1 devemos usar o teste  $t$  para comparação de duas médias independentes com variâncias desconhecidas e, no caso 2, devemos usar o teste  $t$  para comparação de duas médias para dados pareados. b) Prefereria o planejamento 2 pois devemos comparar as duas médias sob igualdade de condições.
24.  $\bar{x}_I = 81,45$  e  $\bar{x}_S = 62,45$ .  $H_0 : \mu_I = \mu_S$   $H_a : \mu_I > \mu_S$   $F = 4,16$  valor  $p = 0,003$  (variâncias heterogêneas)  $t = 4,48$   $gl = 27,63$  e valor  $p = 0,00006$ , portanto, rejeita-se a hipótese nula, ou seja, realmente a reprodução sexuada é prejudicada em área sombreada.
25.  $H_0 : \pi = 0,50$   $H_a : \pi > 0,50$   $z = 2,83$   $P(z \geq 2,83) = 0,50 - 0,4977 = 0,0023$  portanto, rejeita-se a hipótese nula.
26.  $\chi^2 = 50,50$  com 9 graus de liberdade;  $P(\chi^2_9 > 50,50) < 0,001$ , portanto, devemos rejeitar a hipótese nula, ou seja, a distribuição da espécie não é aleatória.

## Índice Remissivo

Experimento, 6, 7  
fatorial, 21

Repetição, 24

Tratamento  
controle, 23