Universidade Federal de Pelotas

ESTATÍSTICA BÁSICA

Versão Preliminar

João Gilberto Corrêa da Silva

Universidade Federal de Pelotas

ESTATÍSTICA BÁSICA

Versão Preliminar

João Gilberto Corrêa da Silva

Pelotas, 2004

Conteúdo

Página

1.1. O Âmbito da Estatística..... 1
1.2. A Estatística e o Cotidiano..... 2
1.3. A Estatística na Pesquisa Científica..... 3
1.3.1. Especificação do objetivo..... 3
1.3.2. Coleta de informações..... 4
1.3.3. Análise de dados..... 4
1.3.4. Estabelecimento de descobertas..... 5
1.4. Situações Ilustrativas de Coleta e Análise de Dados..... 5
1.5. População e Amostra..... 6
1.6. Conceito da Estatística..... 9
1.7. A Interação da Estatística com Outros Campos..... 9
1.8. Mensuração e Escalas de Medida..... 10
1.9. Exercícios..... 13

2.1. Fontes de Dados Estatísticos..... 17
2.2. Descrição Sumária de Dados - Introdução..... 18
2.3. Descrição Sumária de Dados por Tabelas e Gráficos..... 19
2.3.1. Pequenos conjuntos de dados..... 19
2.3.2. Grandes conjuntos de dados..... 20
2.3.2.1. Distribuição de freqüências..... 20
2.3.2.2. Representação gráfica de uma distribuição de freqüências..... 22
2.4. Representação Simbólica de Conjuntos de Dados e da Operação de Adição..... 24
2.5. Medidas de Posição e de Dispersão..... 25
2.5.1. Medidas de centro..... 25
Propriedades da média aritmética..... 25
Média de uma distribuição de freqüências..... 26
2.5.2. Medidas de dispersão..... 26
Variância de uma distribuição de freqüências..... 28
2.6. Exercícios..... 29

3.1. Experimento Aleatório e Eventos..... 35
3.2. Conceito Clássico de Probabilidade..... 38
3.3. Contagem de Pontos do Espaço Básico..... 41
Regra da multiplicação..... 41
Regra das permutações..... 42
Regra das permutações com repetições..... 43
Regra dos arranjos..... 44

Regra das combinações..... 45
3.4. Conceito Empírico de Probabilidade..... 47
3.5. Conceito de Probabilidade em Espaço Básico Discreto..... 49
3.6. Operações de Eventos..... 50
3.7. Probabilidade Condicional..... 53
3.8. Independência Estatística..... 55
3.9. Conceito Geral de Probabilidade..... 57
3.10. Modelos de Probabilidade..... 58
3.11. Exercícios..... 58

4.1. Introdução..... 66
4.2. Variável Aleatória Discreta..... 69
4.2.1. Distribuição de probabilidade..... 69
4.2.2. Representação gráfica de uma distribuição de probabilidade..... 71
4.2.3. Média de uma distribuição de probabilidade..... 73
Propriedades do valor esperado..... 75
4.2.4. Variância de uma distribuição de probabilidade..... 75
Propriedades da variância..... 76
4.3. Distribuições Discretas Importantes..... 77
4.3*.1. Distribuição uniforme discreta..... 77
4.3.2. Distribuição de Bernoulli..... 78
- Média e variância..... 79
4.3.3. Distribuição binomial..... 79
- Média e variância..... 81
4.3.4. Amostragem de uma população dicotômica..... 82
Amostragem com reposição..... 82
Amostragem sem reposição..... 83
4.3.5. Distribuição hipergeométrica..... 84
- Média e variância..... 84
4.3.6. Distribuição geométrica..... 85
- Média e variância..... 86
4.3.7. Distribuição binomial negativa..... 86
- Média e variância..... 87
4.3.8. Distribuição de Poisson..... 87
4.4. Distribuição Conjunta de Duas Variáveis Aleatórias..... 88
Propriedades da função distribuição de probabilidade conjunta..... 92
Representação geométrica..... 92
4.5. Distribuição de uma Função de Duas Variáveis Aleatórias..... 93
4.6. Distribuição Marginal..... 93
4.7. Valor Esperado de uma Função de Duas Variáveis Aleatórias..... 94
4.8. Covariância e Correlação de Duas Variáveis Aleatórias..... 95

Propriedades do coeficiente de correlação..... 97
4.9. Distribuição Condicional e Independência Estatística..... 97
4.10. Distribuição Conjunta de n Variáveis Aleatórias..... 99
Propriedades do valor esperado..... 101
4.11. Distribuição Multinomial..... 102
4.12. Exercícios..... 103

5.1. Introdução..... 111
5.2. Densidade de Probabilidade..... 112
5.3. Distribuição Normal..... 116
5.3.1. Introdução..... 116
5.3.2. Cálculo de probabilidades para variável aleatória normal..... 118
5.4. Aproximação da Distribuição Binomial pela Distribuição Normal..... 122
5.5. Distribuição Normal Bivariada..... 124
5.6. Combinação Linear de Variáveis Aleatórias Normais..... 125
Casos particulares..... 126
5.7. Distribuição da Média de Variáveis Aleatórias..... 127
5.8. Exercícios..... 128

6.1. Introdução..... 133
6.2. Inferência Indutiva..... 133
6.3. População e Amostra..... 134
6.4. Distribuição Amostral..... 136
6.5. Média e Variância Amostrais..... 138
6.6. Teorema Central do Limite..... 139
6.7. Amostragem de Distribuições Normais..... 142
6.7.1. Introdução..... 142
6.7.2. Distribuição da média da amostra..... 143
6.7.3. Distribuição da variância da amostra..... 144
6.7.4. Distribuição t de Student..... 149
6.7.5. Distribuição F..... 153
6.8. Exercícios..... 157

7.1. Introdução..... 163
7.2. Estimação por Ponto..... 164
7.2.1. Conceitos..... 164
7.2.2. Propriedades de um estimador..... 165
7.2.3. Distribuição amostral da média de uma população normal..... 165
7.2.4. Distribuição amostral da variância de uma população normal..... 166
7.3. Estimação por Intervalo - Intervalo de Confiança..... 166
7.4. Teste de Hipótese..... 167
7.4.1. Conceitos..... 167

7.4.2. Testes de hipóteses referentes à media de uma população normal..... 170
7.4.3. Teste da hipótese de igualdade das médias de duas populações normais..... 172
Populações com mesma variância..... 172
7.4.4. Teste unilateral e teste bilateral..... 174
7.5. Exercícios..... 174

8.1. Introdução..... 186
8.1.1. Origens e importância da análise de regressão linear..... 186
8.1.2. Relações entre variáveis..... 187
- Relações deterministas..... 187
- Relações semideterministas..... 187
- Relações empíricas..... 188
8.1.3. Relações lineares..... 189
8.2. Relações de Duas Variáveis..... 190
8.3. Gráfico dos Dados..... 191
8.4. Análise de Regressão Linear Simples..... 193
8.4.1. Introdução..... 193
8.4.2. Modelo estatístico..... 194
8.4.3. Inferência estatística..... 196
8.4.4. Estimação (por ponto) dos parâmetros..... 196
Equação da linha reta ajustada..... 197
Resíduos e estimativa da variância do erro..... 198
Propriedades dos estimadores de quadrados mínimos..... 198
8.4.5. Teste de hipótese..... 201
8.4.5.1. Hipótese de relação linear entre Y e X..... 201
8.4.5.2. Hipótese referente à declividade da linha de regressão..... 202
8.4.5.3. Análise da variação..... 202
8.4.6. Coeficiente de determinação..... 205
8.4.7. Intervalo de confiança..... 206
8.4.7.1. Intervalo de confiança para o coeficiente de regressão b..... 205
8.4.7.2. Intervalo de confiança para $E(Y|X)$ 208
8.5. Correlação Linear Simples..... 209
Relação entre coeficiente de regressão e coeficiente de correlação..... 210
8.6. Exercícios..... 211

Apêndice..... 213
Tabela A-I Distribuição cumulativa binomial..... 215
Tabela A-II Pontos α-percentuais superiores da distribuição normal padrão..... 222
Tabela A-III Pontos α-percentuais superiores da distribuição qui-quadrado..... 223
Tabela A-IV Pontos percentuais da distribuição t (de Student)..... 224
Tabela A-V Pontos percentuais superiores da distribuição F..... 225

1 INTRODUÇÃO

Conteúdo

1.1 O Âmbito da Estatística....1
1.2 A Estatística e o Cotidiano....2
1.3 A Estatística na Pesquisa Científica....3
1.3.1 Especificação do objetivo....3
1.3.2 Coleta de informações....4
1.3.3 Análise de dados....4
1.3.4 Estabelecimento de descobertas....5
1.4 Situações Ilustrativas de Coleta e Análise de Dados....5
1.5 População e Amostra....6
1.6 Conceito da Estatística....9
1.7 A Interação da Estatística com Outros Campos....9
1.8 Mensuração e Escalas de Medida....10
1.9 Exercícios....13

1.1 O Âmbito da Estatística

Para a maioria das pessoas, o uso do termo "estatística" é familiar para designar e reportar fatos numéricos e figuras; por exemplo, as taxas de inflação mensais no Brasil no último quinquênio, os preços diários de mercadorias em uma região, quantidades de produtos exportados por um País em um ano, os valores das ações em uma bolsa de valores em um determinado dia, ou mesmo o número de pontos marcados pelas equipes em uma competição. Entretanto, este uso do termo não caracteriza o âmbito principal da estatística. A estatística trata, principalmente, de situações em que a ocorrência de algum evento não pode ser predita com certeza.

Freqüentemente, conclusões são incertas porque são baseadas em dados incompletos. A avaliação da presente taxa de desemprego em um estado, baseada em uma amostra de poucos milhares de pessoas, é um exemplo. A incerteza também decorre do fato de que observações repetidas de um fenômeno produzem resultados variáveis, embora sejam feitos esforços para controlar os fatores que governam o evento em observação. Por exemplo, pinheiros de um ano de idade não são todos da mesma altura, mesmo que eles se tenham originado de um mesmo lote de sementes e tenham sido cultivados sob semelhantes condições de solo e de clima. O tempo que é tomado para efetuar o corte de uma pastagem, os pesos de frangos de seis semanas criados em um

aviário, e o período de alívio de um sintoma de febre após a ingestão de um certo medicamento são outros exemplos de situações em que ocorre variabilidade em repetidas observações.

A estatística é um corpo de conceitos e métodos úteis para coletar e interpretar dados relativos a uma área particular de pesquisa, e extrair conclusões em situações em que incerteza e variação estão presentes.

Historicamente, a palavra "estatística" é derivada da palavra latina "status", que significa "estado". Por várias décadas, a estatística foi associada apenas com a apresentação de fatos e números referentes a situações econômicas, demográficas e políticas prevalecentes em um país. Mesmo hoje, a grande quantidade de relatórios de governo que contém documentação numérica abundante, com títulos tais como "Estatísticas da Produção Agrícola" e "Estatísticas do Trabalho", lembra a origem da palavra "estatística".

Um grande segmento do público, em geral, ainda tem a concepção errada de que a estatística associa-se exclusivamente com grandes conjuntos de números e algumas vezes complicadas séries de gráficos. Compilações de dados e apresentações numéricas e gráficas tornaram-se um aspecto secundário da estatística. O progresso científico e tecnológico tem demandado um desenvolvimento extraordinário da teoria e da metodologia estatística. Como um corpo de conhecimentos, a estatística, hoje, inclui conceitos e métodos de grande importância, alcance e aplicações em todas as áreas do conhecimento. Particularmente relevante é o papel da estatística no método científico cujo processo de inferência para a derivação de conhecimento é eminentemente baseado em fundamentos estatísticos.

1.2 A Estatística e o Cotidiano

A descoberta de fatos através de coleta e interpretação de dados não é confinada a pesquisadores profissionais. Ela faz parte da vida diária de todas as pessoas que se esforçam, consciente ou inconscientemente, para compreender assuntos de interesse referentes à sociedade, condições de vida, ambiente e ao mundo de modo geral. Ao tentar adquirir conhecimentos a respeito do estado do desemprego, poluição de resíduos industriais, eficácia de analgésicos, comportamento de equipes de futebol em competição, e outras questões de interesse da vida contemporânea, juntam-se fatos e números que devem ser interpretados, ou tenta-se compreender as interpretações que outros fazem. Assim, a aprendizagem ocorre diariamente através de análise de informação fatual freqüentemente implícita.

As fontes de informação fatual abrangem da experiência individual às notícias dos meios de comunicação, relatórios do governo e artigos em revistas profissionais. Previsões de tempo, relatórios do mercado, índices de custo de vida e resultados de pesquisas de opiniões são alguns exemplos específicos. Os métodos estatísticos são empregados extensivamente na preparação de tais relatórios. Relatórios baseados em argumento estatístico apropriado e cuidadosa interpretação das conclusões são verdadeiramente informativos.

Freqüentemente, entretanto, o mau uso, deliberado ou inadvertido, da estatística conduz a conclusões errôneas e distorções da verdade. Para o público geral, o consumidor básico desses

relatórios, alguma idéia do argumento estatístico é essencial para a interpretação apropriada de dados e a avaliação das conclusões que são extraídas.

O argumento estatístico fornece critérios para determinar as conclusões que são realmente suportadas pelos dados e aquelas que não o são. Em todos os campos de estudo onde inferências são extraídas de análises de dados, a credibilidade das conclusões também depende grandemente do uso de métodos estatísticos adequados, especialmente no estágio da coleta de dados.

1.3 A Estatística na Pesquisa Científica

A fundamental importância da estatística no método científico é aparente pela observação de que a inferência em ciências fatuais compreende generalizações baseadas em informações incompletas ou imperfeitas em face de incerteza. Nestas condições, a estatística permeia extensivamente os domínios de toda a pesquisa científica. Uma melhor apreciação é feita com a revisão do método científico, a seguir.

O método científico é um esforço para apreender acerca de regularidades escondidas de fenômenos do que parece ser às vezes um mundo caótico. Embora o método científico não seja rigidamente estruturado, ele compreende, essencialmente, em cada um de seus ciclos, o que segue: Hipóteses ou modelos são tentativamente postulados para explicar um fenômeno; deduções lógicas são derivadas do modelo postulado e então contrastadas com descobertas fatuais; o modelo é modificado, e a pesquisa continua para melhores explanações.

As especificidades do processo de aprendizagem são tão diversas quanto as disciplinas de estudo, mas alguns passos básicos que formam o â mago da maioria das pesquisas científicas são listados a seguir.

1.3.1 Especificação do objetivo

O método científico pode ser considerado para aumentar a compreensão referente a algum fenômeno de interesse cujo presente conhecimento é julgado inadequado. O emprego do método científico, entretanto, requer objetivos mais específicos, tal como provar uma hipótese ou investigar uma teoria existente com respeito à extensão da validade de deduções lógicas dela derivadas.

Em algumas situações, o objetivo pode ser simplesmente a criação de uma base de dados e a derivação de informações que reflitam a presente situação. Por exemplo, as quantidades médias de tempo que os estudantes despendem semanalmente em atividades de lazer podem ser coletadas para estudar aquele componente da dedicação dos estudantes. Outras vezes o objetivo pode ser muito mais extensivo; não apenas ganhar uma compreensão dos fatores que operam em um ambiente, mas, também, determinar as possibilidades de seu uso no controle ou modificação de algum aspecto de um fenômeno. Compreender a química de resíduos sólidos de uma fábrica e seu conseqüente uso para purificação de água de rio da vizinhança é um objetivo dessa forma.

1.3.2 Coleta de informações

A procura de informação objetiva referente ao propósito do estudo é crucial em toda a pesquisa. Esse processo pode envolver uma ampla variedade de atividades, abrangendo desde sofisticados experimentos em ambientes controlados a ensaios de campo, levantamentos sócio-econômicos, ou mesmo registros históricos. Na presente era de instrumentação e mecanização progressiva da manutenção de registros, a crescente quantificação de observações é um fato da vida. Informações são coligidas na forma de dados que exprimem algumas características de indivíduos ou elementos sob estudo.

Especificamente, no estágio de coleta de informação, a estatística guia o pesquisador para os meios apropriados para coletar dados informativos, incluindo a determinação do tipo e extensão dos dados, de modo que conclusões obtidas da análise possam ser formuladas com um grau de precisão desejado. Em áreas de estudo em que a experimentação é de alto custo, o tipo e a quantidade de dados requeridos para fornecer o desejado nível de credibilidade nas conclusões devem ser cuidadosamente determinados a priori. Em outras áreas, tais decisões também são cruciais para a validade e efetividade das conclusões obtidas de uma análise de dados. Os ramos da estatística que tratam do planejamento de experimentos e de coletas de dados informativos são designados, respectivamente, delineamento de experimento e delineamento de amostra.

1.3.3 Análise de dados

Dados coletados por um processo apropriado de experimentação ou observação servem como a fonte básica para a aquisição de novo conhecimento acerca da área sob estudo. É necessário, então, examinar o conjunto de dados e extrair informação relativa à questão levantada na especificação dos objetivos. Uma cuidadosa análise dos dados é crucial para verificar o novo conhecimento obtido e avaliar sua validade.

Neste estágio, há uma necessidade ainda maior dos métodos estatísticos. Alguns desses métodos destinam-se a resumir a informação contida nos dados para focar atenção nas características salientes e descartar os detalhes não essenciais. Um grupo mais importante de métodos para análise de dados destina-se a derivar generalizações ou inferências sobre o fenômeno em estudo.

> A **Estatística Descritiva** compreende o conjunto dos métodos estatísticos para resumir e descrever as características proeminentes de conjuntos de dados referentes a observações de fenômenos particulares de uma classe de fenômenos. O conjunto de métodos estatísticos para avaliação da informação contida em conjuntos de dados e derivação de inferência, ou seja, generalização dessa informação para a classe completa de fenômenos constitui a **Inferência Estatística**.

Embora historicamente a atividade primária, hoje a Estatística Descritiva é apenas uma pequena parte do amplo substrato da Estatística. Uma função importante da Estatística é,

correntemente, a Inferência Estatística, ou seja, a avaliação da informação presente nos dados e a determinação do novo conhecimento ganho a partir dessa informação. O uso dos métodos de inferência estatística fornece uma base de raciocínio para a interpretação lógica de fatos observados, para julgar a extensão em que esses fatos suportam ou contradizem um modelo postulado e para sugerir revisões particulares da teoria existente ou, talvez, planejar ulteriores pesquisas.

1.3.4 Estabelecimento de descobertas

A significação da informação fornecida pelos dados deve, então, ser verificada no contexto do que era conhecido no estágio inicial da pesquisa, quando os objetivos foram especificados. A análise de dados destina-se a responder questões como: "Que generalidades podem ser derivadas a respeito do fenômeno em estudo a partir da evidência fornecida pelos dados?" "É uma conjetura formulada contraditada pelos dados?" "Os dados sugerem uma nova teoria para explicar o fenômeno?" Os resultados da análise são, então, empregados para responder essas questões e, também, para pesar as incertezas envolvidas nas respostas obtidas. Freqüentemente, os resultados sugerem a revisão de uma teoria existente, que pode requerer ulterior pesquisa através de coleta e análise de fatos.

A natureza básica da aquisição de conhecimento é tipicamente uma repetição desse ciclo em uma forma ou outra. Raramente, se alguma vez, uma verdade é desvendada em uma ou mesmo em poucas operações do ciclo e, em muitos campos, a alteração de condições demanda uma continuação indefinida do processo de repetição.

As diferentes áreas da estatística não são entidades disjuntas destinadas para uso separado em estágios individuais de uma pesquisa. Pelo contrário, elas são integradas em um sistema entrelaçado de atividades onde os métodos usados em uma área podem ter forte conexão com aqueles usados em outras áreas. Para decidir sobre o processo de coleta de dados e sua extensão deve-se ter uma percepção dos processos de inferência considerados para uso e da solidez das inferências desejadas. Por outro lado, os métodos de análise de dados e derivação de conclusões são grandemente dependentes do processo de geração dos dados.

1.4 Situações Ilustrativas de Coleta e Análise de Dados

Para esclarecer as generalidades precedentes, alguns exemplos são apresentados a seguir. Eles ilustram algumas situações típicas em que o processo cognitivo da pesquisa de um fenômeno envolve a coleta e análise de dados em que métodos estatísticos são, Conseqüentemente, auxílios indispensáveis na aprendizagem.

- **Melhoramento de plantas**. Experimentos de fertilização cruzada de diferentes tipos genéticos de espécies de plantas para produzir híbridos de elevado rendimento são de considerável interesse para os pesquisadores agrícolas. Como um exemplo simples, suponha que os rendimentos de duas variedades híbridas devem ser comparados sob condições específicas de clima. O único

meio de obter informação sobre o comportamento relativo das duas variedades é cultivá-las em um número de locais, coletar dados de seus rendimentos e analisá-los.

- **Diagnose química**. A detecção precoce é de importância fundamental para o tratamento cirúrgico de muitos tipos de câncer. Tendo em conta que exames periódicos em hospitais são caros e inconvenientes, os médicos buscam processos de diagnóstico efetivos que os próprios pacientes possam administrar. Para determinar os méritos de um novo processo em termos de sua taxa de sucesso na detecção de casos verdadeiros e evitar detecções falsas, o processo deve ser testado em um grande número de pessoas, as quais devem, então, ser submetidas a exames hospitalares para comparação.

- **Programas de treinamento**. Programas de treinamento e ensino, em muitos campos, destinados a um tipo específico de clientela (estudantes universitários, trabalhadores na indústria, grupos minoritários, deficientes físicos, crianças excepcionais, etc.) são continuamente monitorizados, avaliados e modificados para o aumento de sua utilidade para a sociedade. A coleta de dados referentes à aquisição ou desenvolvimento de habilidade de indivíduos ao final de cada programa é essencial para obter informação sobre a eficácia comparativa de diferentes programas.

- **Migração animal**. Biologistas estudam os hábitos migratórios de pássaros e animais identificando-os individualmente em locais geográficos relevantes e, subseqüentemente, seguindo-os em outros locais. Os dados obtidos por tais métodos não apenas ajudam à compreensão do mundo animal, mas, também, alertam os conservacionistas para situações que requerem ação para a proteção de espécies em perigo.

- **Levantamentos socio-econômicos**. Estudos referentes ao bem-estar econômico de diferentes grupos étnicos, padrões de gastos de consumidores de diferentes níveis de rendimento e atitudes em relação à legislação, são conduzidos nas áreas interdisciplinares de sociologia, economia e ciência política. Tais estudos são, muitas vezes, baseados em dados obtidos através de entrevistas ou contato com amostras representativas de pessoas selecionadas, por um processo estatístico, de uma população enorme que forma o domínio de estudo. Os dados são então analisados para a derivação de interpretações do assunto em questão.

1.5 População e Amostra

Os exemplos anteriores são extraídos de campos amplamente diferentes. Embora apenas descrições sumárias da abrangência e dos objetivos dos estudos sejam fornecidas, algumas características comuns são imediatamente aparentes.

Primeiramente, a característica mais saliente em todas essas áreas de estudo é a essencialidade da coleta de dados por processos apropriados de experimentação ou observação. Em segundo lugar, alguma quantidade de variabilidade nos resultados é inevitável, apesar do esforço para que condições iguais ou similares prevaleçam durante repetições de cada experimento ou observação. Por exemplo, no experimento de melhoramento de plantas, é irrealista esperar que cada planta de uma variedade particular tenha o mesmo rendimento, porque a natureza não segue tal lei rígida. Semelhantemente, um programa de treinamento para indivíduos com experiências

similares produz variabilidade em medidas de desempenho. A presença de alguma variação nos resultados sob condições experimentais constantes tende a obscurecer o efeito de uma alteração nessas condições. Um importante ingrediente para a análise estatística de dados é a formulação de modelos apropriados que representem a variabilidade inerente encontrada na natureza.

Uma terceira característica notável dos exemplos anteriores é a impossibilidade ou impraticabilidade física de coletar e estudar um conjunto exaustivo de dados referentes a uma área específica de pesquisa. Não há limite para a quantidade de experimentos de laboratório ou ensaios de campo que possam ser conduzidos, não importa a quantidade de experimentação já efetuada. Em estudos de opinião pública e de gastos com consumo, um corpo completo de informação emergiria apenas se os dados fossem coletados de cada indivíduo da nação. Por exemplo, para coletar um conjunto exaustivo de dados relacionados a danos sofridos por todos os carros de um modelo e ano particular em colisões a uma velocidade específica, cada um dos carros daquele modelo que sai das linhas de produção teria de ser sujeito a uma colisão! O conjunto completo de observações que poderia ser coletado através de repetições ilimitadas de um experimento ou os registros exaustivos de todos os elementos dentro do estudo seria tão vasto que se pode, no máximo, visualizar em imaginação. Tal vasto conjunto de dados pode ser considerado como uma fonte de informação completa, mas as limitações de tempo, recursos e facilidades, e, algumas vezes, a natureza destrutiva do teste, significa que se deve trabalhar com informação incompleta - os dados que são realmente coletados no curso de um estudo experimental.

As idéias fundamentais emanadas dessa discussão salientam uma distinção entre o conjunto de dados que é realmente adquirido através do processo de observação e a vasta coleção de todas as observações potenciais que podem ser concebidas em um dado contexto. Ou seja, uma pesquisa trata com um conjunto de dados coletados em um subconjunto ou parte da coleção de todos os indivíduos para os quais é desejado derivar inferências através da pesquisa.

A coleção de todas as unidades (elementos, indivíduos) de interesse em uma pesquisa é denominada **universo**, **população**, ou, mais especificamente, **população objetivo**. Um subconjunto de unidades da população através da qual se coleta informações na pesquisa é denominada **amostra**. O processo de escolha ou seleção de uma amostra da população é denominado **amostragem**.

Assim, a unidade da amostra é a fonte elementar de informação referente à população objetivo. A unidade da amostra pode ser uma planta, um animal, um conjunto de plantas, um conjunto de animais, uma área (um potreiro, por exemplo), uma fazenda, ou outros elementos, dependendo da pesquisa em particular e do objetivo desta.

Uma população pode ser real (ou concreta) ou hipotética (ou conceitual). Uma população é real se suas unidades podem ser identificadas. A população dos animais de um rebanho é um exemplo de população concreta. Entretanto, a população das plantas de uma cultivar de soja é hipotética, pois não se pode identificar todas as plantas de uma cultivar. Um conjunto é finito ou infinito, se o respectivo número de elementos é finito ou infinito, respectivamente. O conjunto dos números inteiros, por exemplo, é infinito. Todas as populações na natureza são finitas. Entretanto,

populações de número muito elevado de indivíduos são, por conveniência teórica, consideradas infinitas.

Em uma pesquisa, geralmente, tem-se interesse específico em algumas propriedades ou atributos particulares que distinguem as unidades da população sob consideração. Por exemplo, raça, sexo, idade, peso, no caso de animais; cultivar, altura, peso da produção de grãos, no caso de uma planta. Cada unidade manifesta uma alternativa particular de uma característica que o distingue das demais unidades. Por exemplo, sexo é uma característica que pode manifestar-se em uma unidade como macho ou fêmea.

Uma propriedade ou atributo das unidades de uma população objetivo é uma **característica** dessas unidades ou dessa população. Cada uma das formas alternativas de manifestação de uma característica é um **nível** dessa característica.

Uma característica é avaliada através de um processo de mensuração. A mensuração de uma característica demanda o estabelecimento de uma correspondência entre os níveis da característica e os valores de um conjunto de números que leve em conta as relações entre os níveis da característica e as operações que podem ser efetuadas sobre eles. A regra de correspondência estabelecida para representação de uma característica determina a representação dessa característica através de uma função numérica (ou seja, uma função de valores numéricos) definida no conjunto dos níveis da característica.

Uma função numérica que estabelece uma correspondência entre os níveis de uma característica e os valores de um conjunto de números é denominada **variável**. O valor da variável que corresponde a um nível particular da característica é um **nível** da variável.

Freqüentemente, os termos característica e variável são empregados indistintamente.

É comum a referência às características ou às variáveis que as expressam, em vez das unidades. Neste contexto, os dados da amostra consistem de medidas correspondentes à coleção de unidades que são incluídas na pesquisa. Essa coleção forma uma parte de uma coleção de unidades muito maior sobre a qual se deseja fazer inferências. O conjunto de medidas que resultaria se todas as unidades da coleção maior pudessem ser observadas é definida como a população.

Nesse contexto, uma **população**, também denominada **população estatística**, é o conjunto completo das possíveis medidas de características correspondentes à coleção inteira de unidades para as quais inferências devem ser feitas. A população representa a meta de uma pesquisa, e o objetivo do processo de coleta de dados é extrair conclusões sobre a população. Uma **amostra** de uma população estatística é o conjunto de medidas que são realmente coletadas no curso de uma pesquisa.

A estatística provê a metodologia para fazer inferências indutivas sobre a população a partir da coleta e análise de dados da amostra. Esses métodos permitem derivar generalizações plausíveis e, então, avaliar a extensão da incerteza referente a essas generalizações. Conceitos

estatísticos também são essenciais durante o estágio de planejamento de uma pesquisa, quando decisões sobre a forma e a extensão do processo de amostragem devem ser feitas de modo que dados informativos adequados possam ser gerados dentro das limitações dos recursos disponíveis.

Os objetivos da estatística são, essencialmente: a) fazer inferências sobre uma população a partir de uma análise da informação contida em dados da amostra, e b) fazer avaliações da extensão da incerteza envolvida nessas inferências. Um terceiro objetivo, não menos importante, é delinear o processo e a extensão da amostragem de modo que as observações formem uma base para a extração de inferências válidas e exatas.

O planejamento do processo de amostragem é freqüentemente o passo mais importante, especialmente em experimentos controlados em que vários fatores que influenciam medidas podem ser planejados. Um bom plano para o processo de coleta de dados permite uma análise imediata e inferências eficientes, enquanto nem métodos sofisticados de análise de dados salvam muita informação de dados produzidos por um experimento pobremente planejado.

1.6 Conceito da Estatística

A estatística compreende, essencialmente, a arte e a ciência da coleta, análise e interpretação de dados, e a habilidade de derivar generalizações lógicas relativas a fenômenos. A estatística tem sido conceituada de muitas maneiras. Um conceito apropriado para os presentes propósitos é o seguinte:

> A **Estatística** é a ciência e a arte que trata do desenvolvimento e aplicações de métodos de:
>
> - planejamento da coleta de dados,
> - resumo de dados,
> - análise de dados e derivação de generalizações lógicas (inferências) a partir de dados.

Estas quatro funções da estatística correspondem, respectivamente, a:

- Delineamento de experimentos, de levantamentos por amostragem e de estudos observacionais,
- Estatística descritiva,
- Análise estatística e inferência estatística.

1.7 A Interação da Estatística com Outros Campos

O uso primitivo da estatística na compilação estereotípica e apresentação passiva de dados foi sucedido pelo moderno papel de prover ferramentas analíticas para a eficiente coleta, compreensão e interpretação de dados. Os conceitos e métodos estatísticos permitem a extração de

conclusões válidas sobre populações, a partir de amostras. Em decorrência de seu extenso objetivo, a estatística tem penetrado todos os ramos da investigação humana em que a substantivação de asserções e a ramificação de informações devem ser fundadas em evidência baseada em dados.

Os poucos exemplos anteriores não têm a intenção de demarcar o âmbito da aplicação da estatística. Eles são apresentados para ilustrar a diversidade de aplicações da estatística. O uso dos métodos estatísticos em várias áreas da ciência e da tecnologia tem produzido muitas áreas interativas, tais como bioestatística, biometria, psicometria, estatística de negócios, econometria e demografia. Em muitas outras áreas em que nomes compostos ainda não emergiram, tais como ciência política, meteorologia, florestas e ecologia, a estatística já desempenha um papel proeminente.

Os conceitos básicos e o âmbito da metodologia estatística são quase idênticos em todas as diversas áreas de suas aplicações. Diferenças em ênfase ocorrem porque certas técnicas são mais úteis em uma área do que em outra. Entretanto, em decorrência de fortes similaridades metodológicas, exemplos extraídos de uma ampla gama de aplicações da estatística são úteis para o desenvolvimento de uma compreensão básica dos vários métodos estatísticos, de seus usos potenciais e de suas vulnerabilidades para mau uso.

1.8 Mensuração e Escalas de Medida

Uma variável pode compreender poucos, muitos ou infinitos valores. Em particular, uma variável com um único valor, comum para todas as unidades de uma população, é denominada **constante**. Naturalmente, uma constante é uma variável que não distingue as unidades da população. Variáveis de poucos valores distinguem as unidades qualitativamente, segundo classes ou categorias; por exemplo, sexo (macho e fêmea), raça de um ovino (Corriedale, Romney Marsh e Ideal) e cor dos olhos (claros e escuros). Por outro lado, variáveis de muitos ou infinitos níveis distinguem os indivíduos quantitativamente; peso e altura, por exemplo.

> Variáveis cujos níveis distinguem ou classificam as unidades em classes ou categorias são denominadas **variáveis qualitativas** ou **variáveis categóricas**. Variáveis que distinguem as unidades quantitativamente são denominadas **variáveis quantitativas**.

A representação de uma variável é feita através de símbolos específicos, um para cada um de seus níveis. Esses símbolos podem ser um conjunto de letras ou numerais, isto é, símbolos que representam números. A representação numérica é, em geral, a mais conveniente, mesmo para a representação de variáveis categóricas; neste caso, entretanto, operações aritméticas não têm sentido.

O processo de atribuição de símbolos específicos ou numerais aos indivíduos de uma população segundo regras específicas denomina-se **mensuração**. A informação registrada para cada variável em cada indivíduo é um **dado**. O conjunto de dados registrados para cada indivíduo para as variáveis consideradas constitui uma **observação**.

Como em uma pesquisa pode-se ter interesse em uma ou mais características, correspondentemente, devem ser consideradas uma ou mais variáveis. Logo, a observação referente a cada indivíduo pode ser univariada ou multivariada, respectivamente.

Os níveis de mensuração que podem ser adotados são variados. Por exemplo, para qualquer ovelha de um rebanho de ovinos, pode-se medir as seguintes características: a) sexo, b) prolificidade, c) temperatura, e d) peso. É aparente que as mensurações dessas características envolvem quatro diferentes níveis de mensuração: a) pode-se apenas distinguir os animais quanto ao sexo; b) pode-se distinguir e ordenar os animais quanto à prolificidade e dizer que um animal é mais prolífero do que o outro; c) pode-se distinguir e ordenar os animais quanto à temperatura e dizer que um animal tem uma temperatura que supera a de outro em tantos graus; d) pode-se distinguir e ordenar os animais quanto ao peso, dizer que o peso de um animal supera em tantas unidades o peso de outro animal e, ademais, que o peso de um animal é tantas vezes o peso de outro.

Esses quatro exemplos sugerem as quatro **escalas de mensuração**: escala nominal, escala ordinal, escala de intervalo e escala de razão. Essas quatro escalas são definidas e discutidas a seguir.

Escala nominal. Esta escala é utilizada para a mensuração de uma característica que apenas categoriza os indivíduos na população. Para cada categoria é atribuído um símbolo específico de modo que duas diferentes categorias sejam identificadas por distintos símbolos. Por exemplo, o sexo classifica os indivíduos em duas categorias - machos e fêmeas, identificando-os por símbolos, como as próprias palavras "macho" e "fêmea", as iniciais M e F, ou dois números 1 e 2.

A estrutura da escala nominal não é distorcida por uma substituição um a um dos símbolos. Assim, por exemplo, pode-se substituir M pelo número 1 e F pelo número 2, ou M por 10 e F por 100. Entretanto, operações aritméticas não fazem sentido para a escala nominal. Dessa forma, a média aritmética é uma medida sem sentido nesta escala.

Escala ordinal. Esta escala é adotada para a mensuração de uma característica que classifica e ordena os indivíduos na população. Para cada categoria é atribuído um símbolo distinto de tal modo que a ordem dos numerais corresponde à ordem das categorias. Assim, se são atribuídas letras às categorias, elas estão em ordem alfabética; se são atribuídas palavras às categorias, a ordem é especificada pelo significado das palavras. Por exemplo, as cultivares de uma espécie vegetal cultivada podem ser classificadas quanto à resistência à uma infecção como de resistência baixa, média e alta. Se se decide ordenar estas categorias de baixa a alta, então pode-se assinalar os números 1 = baixa, 2 = média e 3 = alta, ou as letras X = baixa, Y = média e Z = alta,

ou chamar as categorias de "baixa", "média" e "alta". Alternativamente, pode-se ordenar as categorias de alta a baixa e atribuir os números: 1 = alta, 2 = média e 3 = baixa. A representação das categorias pode ser feita por qualquer conjunto de símbolos que exprima a idéia de ordem.

A estrutura da escala ordinal não é distorcida por uma substituição de símbolos um a um que preserve a ordem. Assim, a transformação $1 \rightarrow 2$, $2 \rightarrow 3$ e $3 \rightarrow$ qualquer número maior que 3 é permitida, enquanto a substituição $1 \rightarrow 2$, $2 \rightarrow 3$ e $3 \rightarrow 1$ não é permitida. Operações aritméticas também não fazem sentido para a escala ordinal.

Escala de intervalo, ou **escala intervalar**. É utilizada para a mensuração de uma característica que classifica e ordena os indivíduos e, também, quantifica a comparação entre as categorias. A escala de intervalo estabelece uma unidade de mensuração e um ponto zero arbitrário, o que permite determinar quanto uma categoria é mais do outra. Por exemplo, a temperatura de um indivíduo em graus Celsius (C) é mensurada em uma escala de intervalo, em que 0 graus C é a origem e 1 grau C é a unidade de mensuração. Assim, um indivíduo com uma temperatura de 39,5 graus C tem uma temperatura de 2,4 graus C acima do normal (37,1 graus C). A estrutura da escala de intervalo não é distorcida por transformações lineares da forma $x' = ax+b$, $a>0$. O efeito desta transformação é o deslocamento da origem b unidades e a multiplicação da unidade de medida por a. Por exemplo, a transformação $x' = 1,8x + 32$ transforma a escala Celsius na escala Farenheit.

Para esta escala operações aritméticas têm sentido. Assim, por exemplo, tem significado dizer que a temperatura média anual em uma localidade é 19,5 graus C.

Escala de razão ou **escala racional**. Esta escala tem as mesmas características da escala de razão e, adicionalmente, estabelece um ponto de zero absoluto. Portanto, ela permite determinar quantas vezes um indivíduo supera o outro quanto à característica em consideração. Por exemplo, o peso de um indivíduo em quilogramas é uma escala de razão em que 0 kg é um ponto zero fixo e 1 kg é a unidade de medida. Assim, um indivíduo de 110 kg tem o dobro do peso de um indivíduo de 55 kg.

A estrutura da escala de razão não é distorcida por uma transformação da forma $x'=cx$, $c>0$. Um exemplo dessa transformação é $x'=1.000x$ que transforma quilogramas em gramas.

Observe-se que o número de alternativas de uma mesma característica pode variar segundo o nível de distinção dos indivíduos que se deseja ou que é viável considerar. Conseqüentemente, muito freqüentemente, distintas variáveis alternativas podem representar uma mesma característica. A cada uma dessas variáveis corresponde uma distinta escala de mensuração. Assim, por exemplo, a incidência de uma doença em uma espécie cultivada pode ser mensurada em diversas escalas alternativas, dependendo da precisão com que o pesquisador deseja perceber a infecção e dos instrumentos de mensuração disponíveis. Ele pode registrar, simplesmente, a presença ou ausência da infecção - uma categorização rude, envolvendo um baixo nível de percepção. Ele pode escolher um nível mais elevado de percepção, pela utilização de um número mais elevado de categorias e ordenação das incidências de infecção segundo o grau de intensidade, dentro dos extremos correspondentes a ausência de infecção e o grau mais elevado de infecção. Se a doença apresenta um sintoma de infecção aparente nas folhas, ele pode expressar a infecção pelo

número de folhas infectadas - um nível de percepção mais elevado que a ordem. Um nível de percepção ainda mais elevado é obtido exprimindo a intensidade de infecção através de um número de um intervalo de números reais; neste caso, a variável pode assumir qualquer valor real do intervalo. Um diferente nível de percepção é utilizado em cada uma destas quatro situações, que diferem quanto ao grau em que a variável é quantificada em relação às escalas de medida.

As variáveis também podem ser classificadas em discretas e contínuas.

> Uma **variável discreta** tem um número finito ou infinito contável de níveis, enquanto uma **variável contínua** compreende um conjunto de níveis correspondente aos números reais de um intervalo.

As variáveis de escalas nominal e ordinal são necessariamente discretas, enquanto variáveis de escalas de intervalo e de razão podem ser discretas ou contínuas. Por exemplo, a temperatura (em graus Celsius ou em graus Farenheit) é uma variável contínua em uma escala de intervalo, enquanto que o número de frutos em uma planta é uma variável discreta em uma escala de razão.

Uma variável de intervalo ou de razão é intrinsecamente discreta quando seus valores constituem um conjunto numérico finito ou infinito contável; número de leitões em uma ninhada, número de frutos em uma árvore, por exemplo. O nível de percepção da característica (precisão da medida) é, então, implicitamente determinado. Por outro lado, é conveniente observar que, embora, conceitualmente, variáveis contínuas (peso, altura e comprimento, por exemplo) possam assumir qualquer valor de um determinado intervalo de números reais, na prática seus valores restringem-se a números com poucas casas decimais (números inteiros, em alguns casos), limitando-se a um subconjunto discreto de tal intervalo. Esse subconjunto depende da precisão desejada e da precisão do processo e dos instrumentos de mensuração utilizados.

A escolha da escala de medida implica no nível de percepção utilizado na mensuração da característica e, portanto, na quantidade da informação fornecida. Conseqüências não menos importantes decorrem para a escolha das técnicas de inferência estatística, já que as propriedades das distribuições de probabilidade são distintas. Assim, para variáveis em escala nominal, são aplicáveis apenas métodos estatísticos para categorias não ordenadas, como os testes qui-quadrado para distribuição multinomial e para associação, e inferências referentes à distribuição binomial. Para variáveis em escala ordinal são apropriados métodos baseados em ordem. Estes métodos compreendem a área geral conhecida como estatística não paramétrica. As técnicas estatísticas apropriadas para as escalas de intervalo e de razão são comuns e compreendem a grande maioria da metodologia estatística disponível.

1.9 Exercícios

1. Dê exemplos de cinco populações em sua área.

2. Dê um exemplo de população real e um exemplo de população hipotética, cada uma com uma variável discreta e uma variável contínua de interesse.
3. Para uma população de sua escolha descreva um parâmetro e indique como ele pode ser estimado.
4. Exemplifique uma situação em que um censo é apropriado.
5. Os seguintes dados provém de seis animais de um rebanho de ovinos de uma fazenda:

Animal	Data de nascimento	Sexo	Idade de desmame (dias)	Peso ao desmame (kg)	PVS	FINC	OPP	Cor	Caráter
1	28/08/82	M	128	18,5	2,8	5	60	5	4
2	12/09/82	M	143	17,9	2,5	4	50	4	5
3	25/09/82	F	156	12,1	3,6	3	40	3	5
4	05/09/82	M	136	15,8	2,4	5	50	3	4
5	02/10/82	F	163	12,1	1,8	6	80	2	3
6	22/08/82	F	122	23,3	3,5	3	70	4	4

PVS: Peso de velo sujo.

FINC: Finura comercial da lã: 1-Merina; 2-Amerinada; 3-Prima A; 4-Prima B; 5-Cruza 1; 6-Cruza 2; 7-Cruza 3; 8-Cruza 4.

OPP: Número de ondulações por polegada da fibra da lã.

Cor: Cor da lã: 1-Amarela;...; 5-Branca uniforme.

Caráter: Pontuação de um conjunto de características do velo: 1-Pior caráter;...; 5-Melhor caráter.

A que tipos de variáveis e escalas de medidas correspondem estes dados?

6. Que escala de medida é comumente utilizada na mensuração de cada uma das seguintes características:
   a) Longevidade de um indivíduo.
   b) Temperatura ambiente.
   c) Pressão barométrica.
   d) Presença do caráter mocho em um animal de um rebanho.
   e) Volume preenchido de uma lata de conserva.
   f) Ano da data de nascimento de um indivíduo.

g) Número de sementes que germinam em um lote de 100 sementes postas a germinar.

h) Colocação de um animal em um concurso de produção de leite.

7. Apresente três exemplos de cada uma das seguintes variáveis: qualitativa, quantitativa, discreta, contínua, nominal e ordinal.

8. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando com as letras V ou F entre parênteses, respectivamente. Se a sentença for falsa, explique porque.

( ) No processo de pesquisa de um problema, a primeira cousa que um cientista faz é estabelecer o problema.

( ) Um cientista completa uma pesquisa e, então, formula uma hipótese com base nos resultados da pesquisa.

( ) Decisões científicas não podem ser feitas na base de uma única observação.

( ) Em uma pesquisa, o cientista deve, sempre, coletar uma quantidade de dados tão grande quanto humanamente possível.

( ) Mesmo um especialista em uma área de pesquisa pode não ser capaz de obter uma amostra que seja verdadeiramente representativa; então, é melhor escolher uma amostra aleatória.

( ) Em geral, um levantamento é tão bom quanto um experimento.

( ) Uma fraqueza de muitos levantamentos é que há pouco controle de variáveis explanatórias.

( ) O objetivo da estatística é fazer inferências sobre uma população baseadas na informação de uma amostra da população.

( ) A utilização de uma amostra não é necessária quando se usam técnicas estatísticas.

( ) Uma amostra é uma coleção de indivíduos de uma população selecionados por um processo específico.

( ) Toda população na natureza é finita.

( ) Censo é uma enumeração parcial de uma população.

( ) Parâmetros são quantidades determinadas a partir de uma amostra.

( ) A estatística provê estimativas de parâmetros.

( ) Parâmetros são conhecidos quando se usa uma amostra em lugar da enumeração completa de uma população.

( ) Uma técnica de amostragem é um método de extrair uma amostra de uma população.

( ) Em uma população pode haver várias variáveis de interesse para o pesquisador.

( ) Uma população pode ter várias características de interesse para o pesquisador.

( ) Se o pesquisador esforça-se suficientemente, há sempre um meio de encontrar valores da variável de interesse para cada membro da população.

( ) Quando se escolhem as categorias para uma escala nominal, a única condição é que exista uma categoria para cada observação

( ) Um número expresso em uma escala ordinal indica a posição de uma quantidade em um ordenamento.

( ) Tem sentido a aplicação de operações aritméticas com números de uma escala nominal.

( ) Dados de uma escala numérica podem ser facilmente transformados para uma escala nominal.

( ) A escala ordinal é algumas vezes usada, mesmo que informação numérica mais precisa seja disponível.

( ) Dados de uma escala ordinal podem ser facilmente transformados para uma escala numérica.

( ) Pressão barométrica é usualmente registrada em escala ordinal.

( ) A escala de razão diferencia e ordena objetos de acordo com uma unidade padrão e com referência a um ponto zero absoluto.

( ) Operações aritméticas podem ser efetuadas sobre números de escalas nominal e de intervalo, mas não podem ser aplicadas a números de escalas de razão e ordinal.

( ) Medidas de uma variável contínua não podem ser representadas por frações decimais.

( ) Os valores de variáveis contínuas são sempre números reais.

( ) Números inteiros podem ser algumas vezes usados para representar variáveis discretas.

( ) Renda anual arredondada para cruzados é registrada em uma escala numérica discreta.

( ) Idade é realmente uma variável contínua, mas é registrada usando uma escala numérica discreta.

( ) Se uma população é de tamanho infinito, a variável de interesse é contínua.

( ) Variáveis contínuas podem, teoricamente, ser medidas tão precisamente quanto desejado, mas elas são usualmente arredondadas para uma unidade conveniente.

2 ESTATÍSTICA DESCRITIVA

Conteúdo

2.1 Fontes de Dados Estatísticos 17
2.2 Descrição Sumária de Dados - Introdução 18
2.3 Descrição Sumária de Dados por Tabelas e Gráficos 19
2.3.1 Pequenos conjuntos de dados 19
2.3.2 Grandes conjuntos de dados 20
2.3.2.1 Distribuição de freqüências 20
2.3.2.2 Representação gráfica de uma distribuição de freqüências 22
2.4 Representação Simbólica de Conjuntos de Dados e da Operação de Adição 24
2.5 Medidas de Posição e de Dispersão 25
2.5.1 Medidas de centro 25
Propriedades da média aritmética: 25
Média de uma distribuição de freqüências 26
2.5.2 Medidas de dispersão 26
Variância de uma distribuição de freqüências 28
2.6 Exercícios 29

2.1 Fontes de Dados Estatísticos

Os dados provenientes de qualquer fonte são usualmente designados **dados estatísticos**. Eles são números que correspondem a valores de variáveis resultantes de um processo de mensuração de características de indivíduos de uma população ou amostra. Um conjunto de dados estatísticos pode compreender poucos ou centenas ou mesmo milhares de dados, cada um com um ou muitos algarismos.

De modo geral, os dados estatísticos têm duas origens: pesquisa e registros históricos. Dados de pesquisa são gerados para algum propósito específico, através de algum plano preestabelecido. As diversas fontes de dados estatísticos são caracterizadas a seguir:

Censo. É a enumeração de todos os indivíduos de uma população com referência a uma ou mais características.

Levantamento por amostragem. É a enumeração parcial de uma população, isto é, a enumeração de uma parte (amostra) da população com referência a uma ou mais características. Os indivíduos a incluir na amostra são escolhidos por um procedimento específico que, em geral, envolve casualização, e a pesquisa é conduzida segundo um plano previamente estabelecido.

Experimento. Pesquisa em que condições alternativas são impostas com o propósito de avaliar, comparativamente, seus efeitos. A pesquisa é conduzida segundo um plano previamente estabelecido.

Estudo observacional. Estudo em que a escolha dos indivíduos da população para consideração é limitada. Os indivíduos são incluídos no estudo segundo as circunstâncias. Este método de pesquisa é usual em algumas áreas, tal como em medicina, onde o pesquisador utiliza os pacientes que lhe recorrem.

Dados históricos. Dados disponíveis, coletados e mantidos sem propósito específico de uso. São disponíveis em agências coletoras de dados (IBGE, por exemplo), em anuários, etc.

Os dados também são usualmente classificados, quanto à sua origem, em primários e secundários.

Dados primários são aqueles cujo presente uso corresponde a uma etapa da pesquisa que lhes gerou, ou seja, o presente uso foi especificado por um plano de pesquisa; em particular, os procedimentos de análise estatística para as inferências a serem derivadas dos dados foram previstos no plano original da pesquisa. **Dados secundários** são dados disponíveis de pesquisas anteriores ou em registros históricos, cujo presente uso não foi previsto no plano da pesquisa que os originou.

2.2 Descrição Sumária de Dados - Introdução

Um conjunto de dados e seu significado são as peças básicas de informação que a natureza fornece ao pesquisador. Entretanto, a mente não pode alcançar, prontamente, o conteúdo global da informação registrada em dados de elevada ordem de complexidade. Diante de uma lista de 500 medidas, por exemplo, o pesquisador é incapaz de formar uma imagem mental das informações mais relevantes fornecidas pelos dados, isto é, o valor em torno do qual os dados tendem a se aglomerar, a forma de aglomeração e a extensão da variabilidade. Um resumo da informação relevante fornecida pelos dados é de grande utilidade para a compreensão de um conjunto de dados e evitar o efeito caótico da apresentação de uma massa grande de dados.

A **Estatística Descritiva** trata do resumo e da apresentação dos aspectos importantes de um conjunto de dados. Ela inclui a condensação de dados em forma de tabelas, sua representação gráfica e a determinação de indicadores numéricos de centro e de variabilidade. Esses métodos podem ser aplicados em situações em que o conjunto de dados enumera completamente a população (caso de um censo) ou em que o conjunto de dados é obtido a partir de uma amostra, isto é, uma fração da população. Nesta última situação, uma descrição sumaria é seguida de um exame mais profundo e ulterior análise dos dados, de modo que inferências sobre a população possam ser feitas.

O tipo de representação sumária para um conjunto particular de dados depende do volume e complexidade dos dados e do propósito a que ela se destina. Em linhas gerais, ela pode ser feita em dois níveis: Um resumo parcial que permite o exame de características importantes da forma global da distribuição dos dados, incluindo simetria e desvios da simetria e identificação de observações não usuais, que se afastam da massa maior dos dados; e um resumo mais drástico através de poucas medidas, designadas estatísticas para indicação do centro do conjunto de dados e da quantidade de variação que eles apresentam.

Os procedimentos para descrição sumária de dados servem aos seguintes propósitos básicos, à luz dos quais cada conjunto de dados deve ser examinado:

a) Resumo e descrição da configuração global dos dados através de:
   - Apresentação de tabelas e gráficos;
   - Exame da forma global dos dados representados graficamente com referência a características importantes, incluindo simetria ou desvios da simetria;
   - Inspeção dos dados representados graficamente com relação a observações não usuais que possam se situar afastadas da massa maior dos dados.

b) Determinação de medidas numéricas para:
   - Um valor típico ou representativo que indica o centro do conjunto de dados;
   - A quantidade de dispersão ou variação presente nos dados.

Tratar-se-á, aqui, apenas da descrição de dados univariados.

2.3 Descrição Sumária de Dados por Tabelas e Gráficos

2.3.1 Pequenos conjuntos de dados

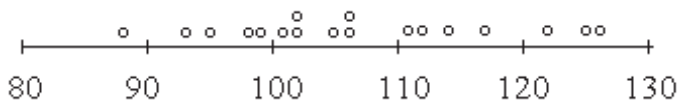
Para conjuntos de dados de relativamente poucas observações (menos que 20 ou 25), uma descrição sumária parcial não tem sentido. Uma representação gráfica do conjunto completo dos dados pode ser feita através de um diagrama de pontos.

> O **diagrama de pontos** é a representação do conjunto de dados sobre um eixo, com uma escala apropriada, de modo que o intervalo das observações seja abrangido pela figura.

Exemplo 2.1. Número de frutos em cada um de 18 pessegueiros:

117	95	122	88	126	112
111	105	98	106	114	93
99	102	106	101	102	125

A menor e a maior observação desse conjunto de dados são, respectivamente, 88 e 126. Logo, o intervalo dos dados é [88,126]. Desenha-se, então, o segmento de um eixo que abranja o intervalo [88,126]; por exemplo, (80,130). Sobre esse segmento, marca-se o ponto correspondente a cada uma das 18 observações:



Esse diagrama de pontos revela uma distribuição de pontos razoavelmente uniforme, com centro entre 100 e 110.

2.3.2 Grandes conjuntos de dados

Para um conjunto de dados numeroso é conveniente um resumo parcial que mantenha a informação relevante referente à distribuição do conjunto completo dos dados. Um tal resumo parcial é provido por uma distribuição de freqüências.

> Uma **distribuição de freqüências** é um número convenientemente pequeno de valores correspondentes às freqüências relativas das observações referentes a subintervalos adequadamente escolhidos do intervalo que compreende o conjunto dos dados, ou aos próprios valores distintos dos dados.

2.3.2.1 Distribuição de freqüências

No caso de variável numérica contínua ou variável numérica discreta com número elevado de níveis, uma distribuição de freqüências pode ser construída como segue:

a) Identifica-se o menor (m) e o maior (M) valor no conjunto de dados. Determina-se a **amplitude total** dos dados: A = M - m.

b) Escolhe-se um número de subintervalos de igual amplitude que cubram o intervalo dos dados (m, M) sem superposição. Esses subintervalos são chamados **intervalos de classe** e seus extremos são as **fronteiras de classe**.

O número de intervalos de classe deve ser convenientemente escolhido para a obtenção de um resumo adequado do conjunto de dados através da distribuição de freqüências. Usualmente, é um número entre 5 e 15, dependendo do número de observações. Uma regra prática é determinar o número de intervalos como:

$$n_c = \sqrt{n},$$

onde n é o número de dados.

c) Determina-se a amplitude de cada classe, dada por:

$$a = \sqrt{\frac{A}{n_c}},$$

e se constrói os intervalos de classe, como segue:

$[m, m+a), [m+a, m+2a),...,[m+(n_c -1)a, M].$

d) Conta-se o número de observações no conjunto de dados pertencentes a cada intervalo de classe, designado n_i para a i-ésima classe. Tal número é a **freqüência absoluta** da classe.

e) Determina-se a **freqüência relativa** de cada classe, dividindo a freqüência absoluta da classe pelo número total de observações no conjunto de dados; para a classe i:

$$f_i = \frac{n_i}{n}.$$

Exemplo 2.2. Dados de medidas de concentração de espermatozóides em sêmen de uma coleta de cada um de 45 touros, em unidades de 10 na potência 8 espermatozóides por ml:

13,8	12,3	9,7	8,6	10,2	12,8	14,1	12,6	12,7
9,0	14,2	13,3	14,8	14,2	13,9	11,0	9,7	14,3
9,5	12,3	11,3	13,9	15,1	11,0	9,6	13,2	12,0
11,8	7,5	12,5	12,1	11,7	10,5	12,2	11,0	12,2
14,2	10,4	15,9	12,0	13,8	9,3	12,9	12,4	10,4

Para construir uma distribuição de freqüências para esses dados, procede-se como segue:

a) $m = 7,5$

$M = 15,9$

$A = M-m$

$= 15,9-7,5 = 8,4;$

b) $n_c = \sqrt{n}$

$= \sqrt{45} = 6,7 \approx 7;$

c) $a = \sqrt{\frac{A}{n_c}}$

$= \frac{8,4}{7} = 1,2.$

É conveniente determinar as freqüências e apresentá-las através de uma **tabela de freqüências** (**Tabela 2.1**).

Tabela 2.1. Distribuição de freqüências relativas para os dados do **Exemplo 2.2**

N° da classe (i)	Intervalo de classe			Freq. absoluta n_i	Freq. relativa f_i
1	7,5	⊢—	8,7	2	1/45 = 0,044
2	8,7	⊢—	9,9	6	6/45 = 0,133
3	9,9	⊢—	11,1	7	7/45 = 0,155
4	11,1	⊢—	12,3	8	8/45 = 0,179
5	12,3	⊢—	13,5	10	10/45 = 0,222
6	13,5	⊢—	14,7	9	9/45 = 0,200
7	14,7	⊢—⊣	15,9	3	3/45 = 0,067
Soma				45	1

Se os dados correspondem à variável numérica discreta com pequeno número de níveis, o que é comum com variáveis nominais e ordinais e pode ocorrer mesmo com variáveis de intervalo e de razão discretas (como, por exemplo, número de leitões em uma ninhada), as classes são os próprios valores distintos dos dados. Então, uma distribuição de freqüências é construída, mais simplesmente, pela determinação das freqüências para esses distintos valores dos dados.

Exemplo 2.3. Dados de números de ovos postos por 60 galinhas em um período de 2 semanas:

3	7	8	9	7	9	4	8	10	8	10	10
10	7	11	7	9	5	11	9	9	9	8	8
5	12	10	8	6	8	11	9	9	7	8	4
9	10	9	6	6	9	8	11	8	9	9	9
9	8	10	10	7	9	8	12	7	9	9	9

Tabela 2.2. Distribuição de freqüências relativas para os dados do **Exemplo 2.3**.

Número da classe (i)	Classe	Freqüência absoluta (n_i)	Freqüência relativa (f_i)
1	3	1	1/60=0,0167
2	4	2	2/60=0,0333
3	5	2	2/60=0,0333
4	6	3	3/60=0,0500
5	7	7	7/60=0,1167
6	8	12	12/60=0,2000
7	9	19	19/60=0,3167
8	10	8	8/60=0,1333
9	11	4	4/60=0,0667
10	12	2	2/60=0,0333
Soma		60	1

2.3.2.2 Representação gráfica de uma distribuição de freqüências

Após o sumário de um conjunto de dados na forma de uma distribuição de freqüências, esta pode ser representada graficamente. No caso de variável numérica contínua ou variável numérica discreta com número elevado de níveis, a distribuição de freqüências é representada, graficamente, através de um histograma, que dá uma idéia da forma da distribuição.

> Um **histograma de freqüências** é a representação de uma distribuição de freqüências através de um gráfico de barras com os intervalos de classe no eixo horizontal e as freqüências (relativas ou absolutas) no eixo vertical.

Para a construção de um histograma de freqüências relativas marca-se os intervalos de classe no eixo horizontal de um gráfico. Em cada intervalo de classe, desenha-se um retângulo com base no intervalo e altura igual à freqüência relativa no intervalo. Para o conjunto de dados do **Exemplo 2.2**, o histograma de freqüências relativas é apresentado na **Figura 2.1**.

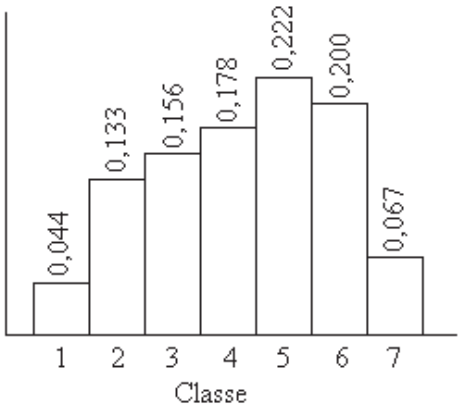


Figura 2.1. Histograma das freqüências relativas dos dados do **Exemplo 2.2**.

É mais conveniente construir os retângulos com alturas iguais às freqüências relativas divididas pela amplitude do intervalo de classe, em vez das freqüências relativas. Com essa convenção, a altura do retângulo sobre o i-ésimo intervalo de classe é f_i /a em vez de f_i . A área de cada retângulo representa, então, a proporção das observações que ocorrem no intervalo em que o retângulo tem a sua base, o que torna comparáveis histogramas construídos para diversos conjuntos de dados, com intervalos de classe de diferentes amplitudes. A única alteração no gráfico é a divisão da escala do eixo das ordenadas pela amplitude a.

Se os dados correspondem à uma variável numérica discreta com pequeno número de níveis, uma distribuição de freqüências é apresentada, graficamente, por um **diagrama de linhas**, construído como segue.

Um **diagrama de linhas de freqüências** é a representação de uma distribuição de freqüências através de um gráfico de linhas com os valores dos dados no eixo horizontal e as freqüências (relativas ou absolutas) no eixo vertical.

Os valores distintos dos dados são dispostos em um eixo horizontal. Desenham-se linhas verticais a partir desses pontos, de alturas iguais às correspondentes freqüências relativas. Nesse caso, linhas substituem retângulos para enfatizar que as freqüências não se espalham sobre intervalos.

O diagrama de linhas para os dados do **Exemplo 2.3** é apresentado na **Figura 2.2**.

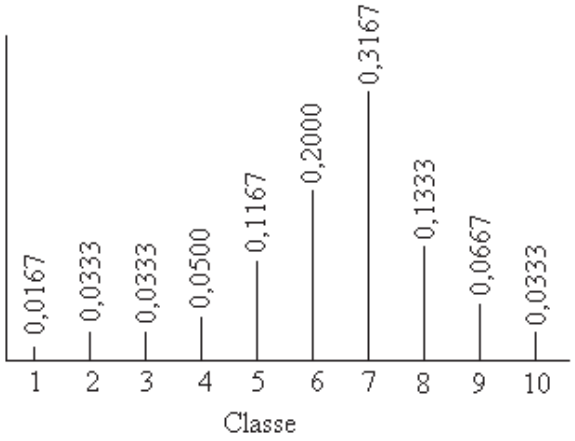


Figura 2.2. Diagrama de linhas das freqüências relativas dos dados do **Exemplo 2.3**.

2.4 Representação Simbólica de Conjuntos de Dados e da Operação de Adição

Para apresentar as idéias e as fórmulas correspondentes às medidas de posição e de dispersão é conveniente representar um conjunto de dados simbolicamente, para que a discussão não fique restrita à um conjunto específico de dados.

É usual representar uma variável por uma das últimas letras do alfabeto, em maiúsculo; seja X. Um valor particular da variável, ou seja, um dado, é representado pela correspondente letra minúscula com um índice para distinguí-lo de outros dados de um mesmo conjunto de dados. Então, um conjunto de n dados é representado, simbolicamente, por x_1, x_2,...,x_n. Por exemplo, um conjunto de dados constituído pelas seguintes medidas referentes a pesos de 5 frangos ao abate: 1,51; 1,82; 1,44; 1,65 e 1,78 kg, é representado, simbolicamente, por x_1, x_2, x_3, x_4 e x_5, onde x_1 =1,51; x_2 =1,82; x_3=1,44; x_4=1,65 e x_5 =1,78.

Operações de adição de dados e de expressões que envolvem dados são efetuadas com freqüência. Para evitar a apresentação detalhada e repetida desta operação, é conveniente representá-la, simbolicamente, através de uma notação apropriada. Assim, a soma dos dados x_1, x_2,...,x_n é representada pela notação $\sum_{i=1}^{n} x_i$, que é lida como: "a soma de todos os x_i com i de 1 a n", ou seja,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n.$$

O símbolo Σ denota o operador da soma, denominado **somatório**. O símbolo que o segue x_i denota os dados (valores da variável X) que são somados. As notações abaixo e acima de Σ especificam a amplitude dos termos que são adicionados. Por exemplo,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + x_4;$$

$$\sum_{i=1}^{n} (x_i - 2) = (x_1-2) + (x_2-2) + (x_3-2).$$

Se o conjunto de dados em consideração é claro no contexto, a soma de todos os n dados do conjunto pode ser indicada, mais simplesmente, pela notação $\sum_{i=1}^{n} x_i$, com a omissão dos extremos dos índices de x abaixo e acima do símbolo Σ do somatório, por ficarem subentendidos como 1 e n.

As seguintes três propriedades da soma são particularmente importantes para as aplicações a serem consideradas adiante. Sejam a e b dois números reais. Então,

a) $\sum_{i=1}^{n} a\,x_i = a\sum_{i=1}^{n} x_i$;

b) $\sum_{i=1}^{n} (a\,x_i + b) = a\sum_{i=1}^{n} x_i + nb$;

c) $\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} x_i^2 - 2a\sum_{i=1}^{n} x_i + na^2$.

2.5 Medidas de Posição e de Dispersão

Freqüentemente, necessita-se de um sumário mais drástico de um conjunto de dados que o propiciado por uma distribuição de freqüências. Tal sumário é obtido através da determinação de medidas que indicam o centro e a dispersão da distribuição dos dados.

Existem muitas medidas de centro e de dispersão. Tratar-se-á, aqui, apenas das medidas mais utilizadas - a média aritmética e a variância, e de medidas derivadas da variância.

2.5.1 Medidas de centro

Uma medida utilizada para representar o centro de um conjunto de dados é denominada **medida de posição** ou **medida de tendência central**. A medida de posição mais utilizada é a média aritmética ou, simplesmente, média:

A **média** de um conjunto de n dados de uma amostra $x_1, x_2, ..., x_n$, designada por $\bar{x}$, é a soma desses dados dividida por n:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i .$$

A média dos dados da amostra do **Exemplo 2.2** é determinada como segue:

$$\bar{x} = \frac{1}{45}(13{,}8 + 12{,}3 + ... + 10{,}4) = 12{,}042 .$$

Propriedades da média aritmética:

A média tem as seguintes propriedades importantes, algumas das quais a distingue de outras medidas de posição:

1) A média tem a mesma unidade de medida dos dados originais.

2) Um conjunto de dados tem apenas uma média.
3) O valor da média é influenciado por valores extremos do conjunto de dados.
4) A média não é aplicável para dados correspondentes a valores de uma variável qualitativa (nominal ou ordinal).
5) A soma dos desvios das observações em relação à média é nula:

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

6) A soma dos quadrados dos desvios das observações em relação a um valor de referência é mínima quando esse valor corresponde à média:

   $\sum_{i=1}^{n}(x_i - c)$ é mínima para $c = \bar{x}$.

7) A média é o centro de gravidade da figura correspondente aos dados dispostos sobre um eixo com um peso distribuído uniformemente sobre os correspondentes pontos sobre o eixo.

Média de uma distribuição de freqüências

Seja uma distribuição de freqüências com N intervalos de classe, e com freqüências absoluta e relativa na i-ésima classe respectivamente n_i e f_i. A média da distribuição de freqüências é dada por:

$$\bar{x}_f = \frac{1}{n}\sum_{i=1}^{N} n_i c_i$$

$$= \sum_{i=1}^{N} f_i c_i ,$$

onde c_i é o centro (média dos extremos) do i-ésimo intervalo de classe ou a i-ésima classe, respectivamente nos casos de distribuição de freqüências de variável contínua ou discreta com número elevado de níveis e de distribuição de freqüências de variável discreta com pequeno número de níveis. No primeiro caso, essa média não coincide necessariamente com a média da amostra, ou seja, com a média dos dados originais, pois esses dados são substituídos pelos pontos médios dos correspondentes intervalos de classe; no segundo caso, essa média coincide exatamente com a média da amostra.

Para o **Exemplo 2.2**, a média da distribuição de freqüências (**Tabela 2.1**) é obtida como segue:

$$\bar{x}_f = 0{,}044\times 8{,}1 + 0{,}133\times 9{,}3 + ... + 0{,}067\times 15{,}3 = 12{,}02,$$

valor diferente, mas bastante próximo, da média do conjunto de dados, $\bar{x} = 12{,}04$. A média da distribuição de freqüências da **Tabela 2.2**, para o **Exemplo 2.3**, é:

$$\bar{x}_f = 0{,}0167\times 3 + 0{,}0333\times 4 + ... + 0{,}0333\times 12 = 8{,}38,$$

que, se pode verificar, é a média do conjunto de dados.

2.5.2 Medidas de dispersão

A medida de dispersão mais utilizada é a variância da amostra:

A **variância da amostra**, também denominada **quadrado médio**, é a média dos quadrados dos desvios das observações em relação à média dos dados, designada por s^2:

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

Simbolicamente, a variância da amostra pode ser expressa por:

$$s^2 = \frac{SQX}{GL},$$

onde GL = n-1 (número de observações subtraído de uma unidade) é o número de **graus de liberdade** da amostra, e

$$SQX = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

é a soma de quadrados de X corrigida (para a média).

Uma expressão mais conveniente para o cálculo desta soma de quadrados é:

$$SQX = \sum_{i=1}^{n} x_i^2 - C$$

onde $\sum_{i=1}^{n} x_i^2$ é a **soma de quadrados não corrigida**, e

$$C = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = n\bar{x}^2$$

é designado **termo de correção** (para a média).

Para o conjunto de dados do **Exemplo 2.2**:

$$s^2 = \left[(13,8\text{-}12,042)^2 + (12,3\text{-}12,042)^2 + ... + (10,4\text{-}12,042)^2\right]$$

$$= \frac{1}{44}161,5298 = 3,6711 .$$

Ou, utilizando a expressão de cálculo:

$$\sum_{i=1}^{n} x_i = 541,9 ;$$

$$\sum_{i=1}^{n} x_i^2 = 13,8^2 + 12,3^2 + ... + 10,4^2 = 6.687,21$$

donde:

$$C = \frac{541,9^2}{45} = 6.525,68 ;$$

logo,

$$s^2 = \frac{1}{44}(6.687,21\text{-}6.525,68) = 3,6711 .$$

Note-se que, se todas as observações são iguais, $s^2=0$; e, quando as observações não são todas iguais, $s^2>0$. Por outro lado, o valor da variância da amostra cresce quando aumenta a variabilidade das observações.

Uma medida de dispersão aparentemente natural é fornecida pela soma dos desvios das observações em relação à média. Entretanto, essa soma é sempre nula, o que foi salientado como uma das propriedades da média.

A variância da amostra é expressa em unidade de medida correspondente ao quadrado da unidade de medida dos dados originais. Esse fato dificulta a interpretação da variância como medida de dispersão. Uma medida mais interessante, expressa na mesma unidade de medida dos dados originais, é o desvio padrão da amostra:

O **desvio padrão da amostra** é a raiz quadrada da variância da amostra:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\,.$$

Para o conjunto de dados do **Exemplo 2.2**:

$$s = \sqrt{3,6711} = 1,9160\,.$$

O desvio padrão depende da unidade de medida em que os dados são registrados. Esse fato pode ser inconveniente quando desvios padrões de diversos conjuntos de dados, com diferentes unidades de medida, são comparados. Uma medida de dispersão mais conveniente, livre de unidade de medida, é o **coeficiente de variação** (CV), que é o desvio padrão expresso como percentagem da média:

$$CV = 100\times\frac{s}{\bar{x}}\,.$$

Para o **Exemplo 2.2**:

$$CV = 100\times\frac{1,9160}{12,042} = 15,91\%\,.$$

Variância de uma distribuição de freqüências

Seja uma distribuição de freqüências com N intervalos de classe e com freqüências absoluta e relativa, na i-ésima classe, respectivamente n_i e f_i. A variância da distribuição de freqüências é dada por:

$$s_f^2 = \frac{1}{n-1}\sum_{i=1}^{N} n_i\,(c_i - \bar{x}_f)^2\,,$$

onde $\bar{x}_f$ é a média, também determinada a partir da distribuição de freqüências, n é o número de dados originais e c_i é o centro do i-ésimo intervalo de classe ou a i-ésima classe, respectivamente para os casos de distribuição de freqüências de dados de variável contínua ou discreta de número elevado de níveis e de distribuição de freqüências de dados de variável discreta de pequeno número de níveis. No primeiro caso, o valor assim obtido não coincide, necessariamente, com a variância da amostra correspondente ao conjunto de dados original.

Para conjuntos de elevado número de dados, a variância pode ser determinada pela expressão:

$$s_f^2 = \frac{1}{n}\sum_{i=1}^{N} n\left(c_i - \overline{x}_f\right)^2$$

$$= \sum_{i=1}^{N} f_i \left(c_i - \overline{x}_f\right)^2 .$$

2.6 Exercícios

1. A seguir, é dada a tabela de freqüências das produções, em decagramas, de 229 plantas da variedade Richland:

Intervalo de classe	c_i	n_i	f_i
0,5 ⊢ 10,5		12	
10,5 ⊢ 20,5		25	
20,5 ⊢ 30,5		73	
30,5 ⊢ 40,5		62	
40.5 ⊢ 50,5		41	
50,5 ⊢ 60,5		12	
60,5 ⊢ 70,5		4	
Soma		229	

c_i: centro de classe

n_i: freqüência absoluta

f_i: freqüência relativa

a) Complete a tabela, preenchendo as colunas em branco.

b) Construa uma figura com o histograma e o polígono de freqüências, ambos para as freqüências relativas.

2. Construa a tabela de freqüências, o histograma e o polígono de freqüências para cada um dos conjuntos de observações dados a seguir:

a) Ganhos de peso, em kg, de 40 bovinos de corte, no período de engorde:

49,5	58,7	46,6	62,5	49,5	59,1	56,1	53,4	49,8
46,4	49,7	50,4	55,9	55,5	56,0	47,1	50,6	47,3
57,0	53,2	58,5	54,0	48,9	48,7	53,4	53,7	54,5
55,1	53,3	53,5	50,3	49,7	47,2	49,7	50,9	45,8
64,6	52,7	50,3	57,8					

b) Números de laranjas danificadas em 65 caixas recebidas em um supermercado, de uma mesma procedência:

5	4	8	8	5	3	3	5	11	5	7	7
9	4	6	13	8	4	14	8	2	5	16	8
14	7	8	3	8	9	10	6	12	5	5	5
6	5	16	2	8	6	10	9	7	6	6	2
11	11	8	9	9	8	11	5	7	15	5	2
5	11	4	8	3							

c) Período de engorda (em dias) de 25 leitões, até atingirem o peso de abate:

105	114	121	117	115
147	119	106	111	142
109	113	163	151	121
114	123	116	109	118
126	115	111	137	106

d) Pesos (em gramas) de 35 ratos ao início de um experimento:

403	387	416	429	406	421	412
369	394	428	406	389	400	468
424	414	418	399	411	407	391
437	402	426	411	392	407	416
398	416	410	409	419	396	413

3. As cinco observações em um conjunto de dados são $x_1=4$, $x_2=3$, $x_3=5$, $x_4=2$ e $x_5=3$. Determine os valores numéricos de:

   a) $\sum_{i=1}^{5} x_i$; b) $\sum_{i=1}^{5} 4$; c) $\sum_{i=1}^{5} 4x$; d) $\sum_{i=1}^{3} (x-3)$;

   e) $\sum_{i=1}^{5} x_i^2$; f) $\sum_{i=2}^{4} x_i$; g) $\sum_{i=1}^{5} (x_i-3)^2$;

   h) $\frac{1}{5}\sum_{i=1}^{5} x_i$; i) $\frac{1}{4}\left[\sum_{i=1}^{5} x_i^2 - \frac{1}{5}\left(\sum_{i=1}^{5} x_i\right)^2\right]$.

4. Seja $\{x_1, x_2,...,x_n\}$ um conjunto de dados. Exprima, simbolicamente, usando o símbolo de soma Σ, o seguinte:
   a) Soma de todos os dados do conjunto.
   b) Soma dos k primeiros dados (k<n).
   c) Soma dos quadrados dos dados.
   d) Soma dos quadrados dos desvios dos dados em relação a 8.

5. Os dados que seguem são as medidas de 5 pinheiros de 6 anos expressas em metros: 1,54; 1,87; 1,44; 1,62 e 1,77. Determine os valores das seguintes estatísticas para aquele conjunto de dados: média, variância, desvio padrão e coeficiente de variação, indicando as respectivas unidades de medida.

6. Os dados que seguem são as alturas, em cm, de cordeiros de sete meses que receberam suplemento de sal mineral:

   57; 59; 49; 62; 51; 50; 55; 48; 52; 42; 61; 57.

   Para esse conjunto de dados, determine as seguintes estatísticas:

   a) média aritmética;

   b) variância (quadrado médio);

   c) desvio padrão;

   d) coeficiente de variação.

7. Para cada um dos conjuntos de dados da questão 2, determine as seguintes estatísticas:

   a) média;

   b) variância (quadrado médio);

   c) desvio padrão;

   d) coeficiente de variação.

8. Escreva uma breve sentença (não a definição) referente a cada um dos seguintes conceitos:

   a) medida de posição;

   b) variância;

   c) soma dos desvios da média;

   d) amplitude dos dados.

9. Escreva em forma expandida:

   a) $\sum_{i=1}^{6}(x_i-a)^2$; b) $\sum_{i=1}^{n}(x_i-y_i)$; c) $\sum_{i=1}^{p}(x_i-b)^3$;

   d) $\sum_{u=1}^{k} x_u^u$; e) $\sum_{i=1}^{5}(-1)^i x_i y_i$.

10. Demonstre que para qualquer conjunto de dados $\{x_1,x_2,..., x_n\}$:

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

    b) $\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - C$, $\quad C = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = n\,\bar{x}^2$.

11. Suponha que, em vez do metro, as medidas das alturas dos pinheiros do exercício 5 houvessem sido expressas em centímetro. Quais seriam as alturas dos 5 pinheiros nessa nova escala.

    a) Determine as mesmas estatísticas indicadas naquele exercício para os dados expressos nessa nova escala.

    b) Para cada uma das estatísticas, verifique a relação entre seus valores para os dados nas duas escalas.

12. Para os dados do exercício 5, se a=10, mostre que:

    a) A média de aX é igual a a vezes a média de X.

    b) A média de a+X é igual a a mais a média de X.

    c) A variância de aX é igual a a vezes a variância de X.

    d) A variância de a+X é igual a variância de X.

e) O desvio padrão de aX é igual a a vezes o desvio padrão de X.

13. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando com as letras V ou F entre parênteses, respectivamente. Se a sentença for falsa, explique porque.

( ) De uma tabela de freqüências pode-se, sempre, derivar o número de dados menor que o limite superior de uma classe.

( ) Em uma tabela de freqüências, a soma das freqüências absolutas é sempre igual a um.

( ) Para qualquer conjunto de dados, a freqüência relativa em uma classe não pode ser zero.

( ) Distribuições de freqüência são sempre simétricas.

( ) A soma das áreas dos retângulos em um histograma de freqüências é sempre igual a um.

( ) É apropriado calcular a média de um conjunto de dados registrados em escala nominal.

( ) .A média da amostra é sempre um valor da amostra.

( ) Para qualquer amostra, Σ(x - média de x) = 0.

( ) Se Σ(x-c)=0 para uma amostra, então c é a média da amostra.

( ) Se a variável X é expressa em metros, então a média de X também é expressa em metros.

( ) Se uma variável X é expressa em metros, a unidade de medida do desvio padrão é metros quadrados.

( ) Se para cada valor de X em uma amostra Y = X+10, então: média de Y+10 = média de X.

( ) Se para cada valor de X em uma amostra Y = aX, então a média de Y é igual a a vezes a média de X e a variância de Y é a elevado ao quadrado vezes a variância de X.

( ) A média dos dados de um conjunto multiplicados por uma constante é igual à média dos dados originais.

( ) Se todas as observações de um conjunto de dados são divididas por uma constante, então sua média é dividida pela mesma constante.

( ) Se uma constante é subtraída de cada uma das observações, a média dos dados resultante é igual à média dos dados originais subtraída da constante.

( ) Se todas as observações de um conjunto de dados são multiplicadas por uma constante, então a correspondente variância é multiplicada pela mesma constante.

( ) O desvio padrão de um conjunto de dados divididos por uma constante é igual ao desvio padrão dos dados originais dividido pela mesma constante.

( ) Se uma constante é subtraída de cada observação, o desvio padrão das observações transformadas é subtraído da mesma constante relativamente ao desvio padrão das observações originais.

( ) A soma de quadrados corrigida é zero quando todas as observações são iguais.

( ) Se dois conjuntos de dados A e B têm a mesma média, o desvio padrão de A é 20 e o de B é 8, pode-se concluir que as observações no conjunto de dados A estão aglomeradas mais próximas da média do que as observações de B.

3 PROBABILIDADE

Conteúdo

3.1 Experimento Aleatório e Eventos 35
3.2 Conceito Clássico de Probabilidade 38
3.3 Contagem de Pontos do Espaço Básico 41
Regra da multiplicação 41
Regra das permutações 42
Regra das permutações com repetições 43
Regra dos arranjos 44
Regra das combinações 45
3.4 Conceito Empírico de Probabilidade 47
3.5 Conceito de Probabilidade em Espaço Básico Discreto 49
3.6 Operações de Eventos 50
3.7 Probabilidade Condicional 53
3.8 Independência Estatística 55
3.9 Conceito Geral de Probabilidade 57
3.10 Modelos de Probabilidade 58
3.11 Exercícios 58

3.1 Experimento Aleatório e Eventos

O conceito de probabilidade é relevante para o estudo de fenômenos cujas realizações não podem ser previstas exatamente. Ele se aplica a "experimentos" cujos resultados em diversas repetições são variáveis e, portanto, incertos, apesar dos esforços em manter as condições fixas. O termo "experimento" aqui não se refere estritamente a "experimento controlado", como é o caso de experimento de laboratório, mas a qualquer atividade que resulte na coleta de dados referentes a fenômenos aleatórios. Em tais experimentos, a variação dos resultados em realizações sucessivas é inevitável.

Um **experimento aleatório** $\mathcal{E}$ é o processo de obter dados relevantes a um fenômeno que satisfaz às seguintes condições:

i) o resultado de uma realização de $\mathcal{E}$ não pode ser predito com certeza;

ii) a coleção dos possíveis resultados de $\mathcal{E}$ é conhecida antes de sua realização;

iii) o experimento $\mathcal{E}$ pode ser repetido sob, essencialmente, as mesmas condições.

Assim, esse conceito de experimento aleatório não se restringe a experimentos de laboratório ou de campo, mas inclui qualquer atividade que implique na coleta de dados referentes a fenômenos aleatórios, isto é, fenômenos que exibem variação.

Para facilitar a compreensão dos conceitos importantes que seguem, considerar-se-á os seguintes exemplos de experimentos aleatórios:

a) pôr uma vaca em cria;

b) efetuar o plantio de cinco sementes de trigo;

c) pôr 100 sementes em uma câmara de germinação;

d) administrar um antibiótico a suínos que sofrem de uma mesma infecção viral até que um tenha uma reação adversa;

e) preparar vinho de uva de um vinhedo;

f) administrar um hormônio de crescimento a um bovino de corte para acelerar sua preparação para o abate.

Cada realização de um desses experimentos terá um resultado particular que se diferenciará dos demais resultados possíveis por um número de diferentes características. Em geral, entretanto, tem-se interesse em apenas uma ou poucas características. Nos experimentos utilizados para ilustração, pode-se ter o interesse restrito às seguintes características:

a) sexo do terneiro;

b) grau de infecção das folhas pela ferrugem;

c) número de sementes que germinam;

d) número de animais a que se administra o antibiótico;

e) concentração de açúcar no mosto;

f) peso e ganho de peso doze semanas após a administração do hormônio.

A especificação da característica ou das características de interesse deve fazer parte da descrição do experimento. Assim, as descrições completas dos experimentos que estão sendo utilizados para ilustração correspondem ao que segue:

a) pôr uma vaca em cria e observar o sexo do terneiro;

b) efetuar o plantio de cinco sementes de trigo e observar o grau de infecção de ferrugem nas folhas;

c) pôr 100 sementes em uma câmara de germinação e observar o número de sementes que germinam;

d) administrar um antibiótico a suínos que sofrem de uma mesma infecção viral até que um animal tenha uma reação adversa e registrar o número de animais a que se administra o antibiótico;

e) preparar vinho de uva de um vinhedo e determinar a concentração de açúcar no mosto;

f) administrar um hormônio de crescimento a um bovino de corte para acelerar sua preparação para o abate e registrar o peso e o ganho de peso doze semanas após a administração do hormônio.

> **Espaço básico** é a coleção de todos os resultados elementares possíveis do experimento. Cada um dos resultados elementares possíveis é um **evento simples**, ou **evento elementar**. O espaço básico é denotado por S. Um evento simples é um elemento de S, designado por s.

Assim, o espaço básico é um conjunto cujos elementos são os eventos simples do correspondente experimento. Dessa forma, toda a conceituação e propriedades derivadas são completamente análogas à conceituação e propriedades referentes a conjuntos.

Um evento simples do espaço básico S, correspondente a um experimento aleatório, é indicado pela notação $s \in S$, que se lê "s pertence a S". O espaço básico de um experimento é especificado pela listagem de todos os resultados possíveis do experimento, com o uso de símbolos convenientes para identificar os resultados, ou por uma sentença que caracteriza a coleção de todos os resultados possíveis. Assim, por exemplo, os espaços básicos para os experimentos que estão sendo utilizados para ilustração podem ser especificados como segue:

a) S = {M, F}, onde M e F representam sexo masculino e sexo feminino, respectivamente;

b) S = {SI, IFr, IM, IFo, IG}, onde SI, IFr, IM, IFo e IG representam ausência de infecção e os graus de infecção fraca, média, forte e grave, respectivamente.

c) S = {0, 1, 2,..., 100};

d) S = {C, SC, SSC, SSSC,...}, onde S designa sem reação adversa e C, com reação adversa e a seqüência SSC, por exemplo, indica que o terceiro animal que recebe o antibiótico é o primeiro com reação adversa;

e) Se a concentração de açúcar no mosto é medida como a proporção de açúcar em uma unidade de volume do mosto, a medida de concentração é um número real do intervalo [0; 1]. Então, o espaço básico para o experimento é $S = \{t: 0 \leq t \leq 1\}$, isto é, "o conjunto dos números reais t tal que t está compreendido entre 0 e 1, incluídos 0 e 1".

f) Nesse experimento, são consideradas duas características: peso e ganho de peso, ambas expressas em uma escala numérica contínua. Designando peso e ganho de peso por x e y, respectivamente, o espaço básico pode ser especificado por: S = {(x, y): x é um número real não negativo e y é um número real}.

Esses exemplos salientam que a complexidade do espaço básico depende da natureza do experimento. Os espaços básicos dos experimentos a), b) e c) são conjuntos finitos; são exemplos de **espaço básico finito**. O espaço básico do experimento d) não é finito, mas seus elementos podem ser enumerados, isto é, contados; é um exemplo de **espaço básico infinito contável**. A variável que exprime a característica considerada no experimento e) é numérica contínua; o

correspondente espaço básico, que consiste dos números reais do intervalo [0; 1], é um exemplo de **espaço básico contínuo**. O experimento f) ilustra **espaço básico contínuo bidimensional**, porque considera duas características, ambas expressas por variáveis de escalas numéricas contínuas.

> Um espaço básico cujos elementos constituem um conjunto finito ou infinito contável é um **espaço básico discreto**. Se o espaço básico inclui todos os números de um intervalo da linha reta, ele é chamado um espaço **básico contínuo**.

Cada vez que um experimento aleatório é realizado, pode ocorrer um e apenas um evento simples ao qual corresponde um e apenas um elemento do espaço básico. Em geral, entretanto, há interesse na ocorrência de algum aspecto descritivo comum a vários eventos simples. A correspondente coleção de eventos simples constitui um **evento composto**. Assim, no exemplo b), pode-se ter interesse nos eventos compostos correspondentes à ocorrência de plantas com infecção, ou seja, no evento {IFr, IM, IFo, IG}, e à ocorrência de plantas com infecção com grau abaixo do médio, ou seja, no evento {SI, IFr}. Por estes exemplos, fica claro que um evento composto ocorre quando qualquer um dos eventos elementares que o constitui ocorre.

Eventos simples e eventos compostos, designados, genericamente, eventos, são ambos subconjuntos do espaço básico. As primeiras letras maiúsculas do alfabeto, A, B, C,..., são comumente usadas para simbolizar eventos.

Para exemplificar mais amplamente a representação de eventos, recorre-se, a seguir, aos experimentos que estão sendo utilizados como ilustração:

a) evento "terneiro macho": {M};

b) evento "infecção acima da média": {IFo, IG};

c) evento "pelo menos noventa sementes germinam": {90,91,..., 100};

d) evento "no máximo três animais recebem o antibiótico": {C, SC,SSC};

e) evento "concentração de açúcar entre 0,4 e 0,6": {t: $0,4 \leq t \leq 0,6$};

f) evento "animal com peso acima de 400 k e com ganho de peso positivo": {(x, y): $x>400$; $y>0$}.

3.2 Conceito Clássico de Probabilidade

Intuitivamente, a probabilidade de um evento é uma medida da chance de sua ocorrência quando o correspondente experimento é realizado. É natural pensar que uma medida da chance de ocorrência de um evento seja provida pela proporção de vezes em que o evento é esperado ocorrer quando o experimento é repetido sob, essencialmente, as mesmas condições. Assim, a probabilidade depende da natureza do experimento e do espaço básico associado. Em algumas situações, a proporção de vezes que se espera cada um dos eventos simples ocorrer é determinada por derivação lógica, sem efetivamente realizar o experimento. Em outras situações, é necessário repetir o experimento um número elevado de vezes para obter informação sobre a proporção esperada de ocorrência dos eventos no espaço básico, através das correspondentes freqüências de ocorrências observadas. Nesse caso, as verdadeiras proporções (probabilidades) são desconhecidas e apenas aproximações podem ser obtidas.

Esse fato derivou muita discussão do conceito de probabilidade no passado, que ainda hoje se mantém polêmico entre diversas escolas. O conceito mais primitivo de probabilidade é o

chamado conceito clássico. Quando uma simetria da estrutura do experimento assegura que todos os eventos simples têm a mesma chance de ocorrer em uma realização do experimento, o espaço básico é dito equiprovável. Seja, por exemplo, o experimento "escolha aleatória de uma de 6 garrafas de uma caixa, identificadas pelos seis números inteiros 1, 2, 3, 4, 5 e 6". "Escolha aleatória" de um indivíduo de um conjunto é um processo de seleção que assegura igual chance de extração para todos os indivíduos. Nessas circunstâncias, sem realizar o experimento, pode-se, logicamente, concluir que é esperado que cada garrafa ocorra em 1/6 das vezes. Logo, pode-se expressar a probabilidade de cada um dos eventos simples desse experimento por: P(s) = 1\6, s∈ S, que se lê: "a probabilidade de s é igual a 1\6 para todo elemento s de S".

Se A é o evento "garrafa identificada por um número par", isto é, A = {2,4,6}, como a proporção de ocorrências do evento A é a soma das proporções das ocorrências de cada uma das garrafas identificadas pelos números 2, 4 e 6, deve-se ter:

$$P(A) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}.$$

Estrutura completamente análoga à desse experimento é a do experimento: "lançamento não tendencioso de um dado não viciado e observação do número da face que resulta virada para cima".

Essa derivação lógica de probabilidades para eventos de um espaço equiprovável correspondente a um experimento particular é generalizada como segue:

Conceito clássico de probabilidade: Se um espaço básico consiste de k eventos simples, $S = \{s_1, s_2, ..., s_k\}$, equiprováveis, a **probabilidade** de um evento A que consiste de m desses k eventos simples é $P(A) = \frac{m}{k}$.

Observe-se que k é o número de elementos do espaço básico, isto é, o número de distintos resultados possíveis do experimento, e m é o número de eventos simples que implicam na ocorrência de A, isto é, o número de resultados favoráveis a A. Assim, pode-se dizer que:

$$P(A) = \frac{\text{Número de resultados favoráveis a A}}{\text{Número de resultados possíveis}}.$$

Diz-se que um espaço básico cujos eventos simples são equiprováveis tem um **modelo de probabilidade uniforme**. Uma função P( ) que atribui igual probabilidade para todos os eventos simples do espaço básico S é designada **função de probabilidade uniforme**.

Apesar de aparentemente bastante restritiva, a condição de equiprobabilidade do espaço básico é de grande importância prática, pois corresponde a experimentos em que é feita a escolha de membros de um conjunto de objetos ou indivíduos por processo que atribui a cada um deles igual chance de seleção e, então, é procedida à mensuração de alguma característica específica. O processo de seleção de indivíduos de uma população que atribui a cada indivíduo a mesma probabilidade de constituir a amostra é denominado **amostragem aleatória simples**.

Conforme o próprio conceito indica, a probabilidade de um evento em um espaço básico equiprovável é determinada pelas contagens do número de resultados possíveis do experimento e do número de resultados favoráveis ao evento. A aplicação do conceito clássico de probabilidade será ilustrada, a seguir, através de exemplos simples. Regras gerais para a contagem de eventos ("análise combinatória") e outros exemplos do cálculo de probabilidades em espaço básico equiprovável serão dados adiante.

Exemplo 3.1. Experimento: Extração aleatória de um saco de sementes de uma pilha com 2 sacos de sementes da variedade V_1 e 3 da variedade V_2. Seja o seguinte evento: A: um saco de sementes da variedade V_1.

Há 5 resultados possíveis na seleção de um saco de uma pilha de 5 sacos. Dois desses 5 resultados são favoráveis ao evento A - qualquer dos dois sacos da variedade V_1 que seja extraído implica na ocorrência de A. Logo, P(A) = 2/5.

Exemplo 3.2. Experimento: Extração aleatória sucessiva de dois sacos de sementes da pilha de 5 sacos considerada no **Exemplo 3.1**, **com reposição** do primeiro saco extraído antes da extração do segundo saco. Sejam os seguintes eventos: A: dois sacos da variedade V_1; B: dois sacos da variedade V_2; C: um saco de cada variedade.

Na extração do primeiro saco, qualquer um dos 5 sacos de sementes pode resultar; há, portanto, 5 resultados possíveis na primeira extração. Como o segundo saco extraído é reposto antes da segunda extração, há, novamente, 5 resultados possíveis na segunda extração. O número de resultados possíveis do experimento é, portanto, 5x5=25.

Os resultados favoráveis ao evento A são aqueles que correspondem a dois sacos da variedade V_1. Na primeira extração, qualquer dos dois sacos da variedade V_1 é favorável ao evento A se, novamente, um dos dois sacos da variedade V_1 for obtido na segunda extração. Logo, o número de resultados favoráveis ao evento A é 2×2=4. Portanto,

$$P(A) = \frac{2\times 2}{5\times 5} = \frac{4}{25}.$$

Pelo mesmo raciocínio, pode-se determinar a probabilidade do evento B:

$$P(B) = \frac{3\times 3}{5\times 5} = \frac{9}{25}.$$

O evento C ocorre quando resulta um saco da variedade V_1 na primeira extração e um saco da variedade V_2 na segunda extração, ou um saco V_2 na primeira extração e um saco V_1 na segunda. Logo,

$$P(C) = \frac{2\times 3+3\times 2}{5\times 5} = \frac{12}{25}.$$

Observe-se que $P(A) + P(B) + P(C) = \frac{4}{25} + \frac{9}{25} + \frac{12}{25} = 1.$

Exemplo 3.3. Mesma situação do experimento do **Exemplo 3.2**, mas sem reposição do saco que resulta da primeira extração.

Como na situação do **Exemplo 3.2**, há 5 resultados possíveis na extração do primeiro saco. Entretanto, como o primeiro saco extraído não é reposto, há 4 resultados possíveis na segunda extração. Portanto, o número de resultados possíveis do experimento é, agora, $5\times4 = 20$.

A extração de qualquer um dos dois sacos V_1, seguida da extração do restante saco da variedade V_1, é favorável ao evento A. Portanto, o número de resultados favoráveis ao evento A é 2×1. Logo,

$$P(A) = \frac{2\times1}{5\times4} = \frac{2}{20} = \frac{1}{10}.$$

Por raciocínio análogo, obtém-se:

$$P(B) = \frac{3\times2}{5\times4} = \frac{6}{10} = \frac{3}{10}.$$

Um saco de cada uma das variedades ocorre quando se sucedem as extrações de 1 saco V_1 e 1 saco V_2 , ou 1 saco V_2 e 1 V_1 . Logo,

$$P(C) = \frac{2\times3+3\times2}{5\times4} = \frac{12}{20} = \frac{3}{5}.$$

Observe-se que $P(A)+P(B)+P(C) = \frac{1}{10}+\frac{3}{10}+\frac{3}{5} = 1$.

3.3 Contagem de Pontos do Espaço Básico

Modelos de probabilidade uniformes em espaço básico finito são úteis quando o esquema de amostragem é governado por algum mecanismo que torna todos os eventos elementares igualmente prováveis. Nesse espaço básico, a probabilidade de um evento A é dada pela razão:

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

onde n(A) é o número de eventos simples no evento A. Assim, nessa situação, a probabilidade de um evento reduz-se, essencialmente, às contagens do número de eventos simples que compõe o evento em consideração e do número de eventos simples no espaço básico. O conhecimento dos eventos e a sua listagem não são, em geral, necessários.

Nos exemplos simples anteriores, a contagem de eventos simples foi feita sem dificuldade. Em situações complicadas, entretanto, o processo de contagem pode se tornar difícil e trabalhoso. Métodos sistemáticos convenientes de contagem ou enumeração são, portanto, importantes para o cálculo direto de probabilidades em espaço básico finito equiprovável. As regras básicas de contagem são revisadas a seguir.

Regra da multiplicação

> Se uma operação consiste de duas etapas a primeira das quais pode ser procedida de n_1 modos distintos e para cada um desses modos a segunda etapa pode ser procedida de n_2 modos distintos, então a operação total pode ser efetuada em $n_1\times n_2$ diferentes modos.

Exemplo 3.4. Um agricultor tem disponíveis duas cultivares de arroz (C_1 e C_2) que podem ser semeadas em três épocas (E_1, E_2 e E_3). Quantas decisões alternativas o pesquisador pode tomar para o cultivo do arroz no que se refere à cultivar e época de semeadura?

Há $n_1 = 2$ cultivares cada uma das quais pode ser semeada em $n_2 = 3$ épocas. O número de diferentes decisões é, portanto, $n_1 \times n_2 = 2 \times 3 = 6$.

Uma lista completa das decisões (ou operações) possíveis pode ser obtida, de modo sistemático, por um **diagrama de árvore**, **Figura 3.1**. Esse diagrama mostra que há $n_1 = 3$ ramos (possibilidades) referentes à escolha de época de semeadura e para cada um desses ramos há $n_2 = 2$ ramos (possibilidades) referentes à escolha de cultivar. As 6 possíveis decisões são representadas pelos 6 distintos caminhos ao longo dos ramos da árvore.

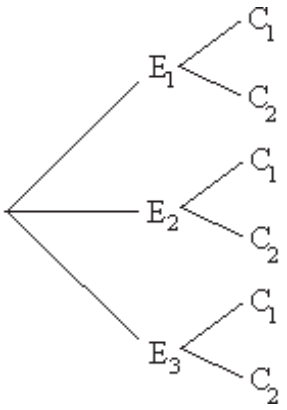


Figura 3.1. Diagrama de árvore que representa o processo de decisão para o cultivo de arroz quando são disponíveis duas cultivares que podem ser semeadas em três épocas.

A regra da multiplicação pode ser imediatamente estendida à situação de operação com mais de duas partes: Se uma operação consiste de k etapas, a primeira das quais pode ser procedida de n_1 modos distintos, para cada um dos quais há n_2 modos distintos da segunda etapa, para cada modo das duas primeiras etapas há n_3 modos da terceira etapa, e assim por diante, então, a operação total pode ser conduzida em $n_1 \times n_2 \times ... \times n_k$ modos.

Uma derivação da regra da multiplicação é a regra das permutações, que segue.

Regra das permutações

O número de diferentes **permutações** que podem ser formadas com n elementos ou objetos distintos, denotado por P_n e designado "permutações de n elementos", é o número de diferentes ordenações desses elementos, que é determinado pela expressão:

$$P = n(n-1)(n-2) \times ... \times 2 \times 1,$$

ou, equivalentemente,

$$P_n = 1 \times 2 \times ... \times n.$$

Esta expressão também é denotada por n! e designada **fatorial de n**.

Por exemplo, as permutações de três objetos identificados pelas letras a, b e c são: abc, acb, bac, bca, cab e cba. De modo geral, permutar n objetos equivale a colocá-los dentro de uma caixa com n compartimentos, em alguma ordenação. O número de permutações dos n objetos é o número de diferentes ordens de disposição dos n objetos.

1	2	...	n

Exemplo 3.5. Um experimento é conduzido para a comparação de cinco cultivares de feijão A, B, C, D e E quanto à produção de grãos. Ao fim do experimento, as médias observadas das cultivares são ordenadas. Quantas diferentes ordenações das cultivares podem resultar?

Qualquer uma das cinco cultivares pode resultar como mais produtiva. Para cada uma delas que resulte mais produtiva, há quatro cultivares restantes que podem resultar com a segunda maior produção. Portanto, há 5x4=20 diferentes arranjos das 5 cultivares para os dois primeiros lugares em produção. Para cada um desses 20 distintos resultados, há 3 cultivares restantes qualquer uma das quais pode ter a terceira produção mais elevada; portanto, há 5x4x3 = 60 distintos arranjos das 5 cultivares para as 3 primeiras colocações. Para cada um desses 60 arranjos, há duas cultivares restantes qualquer uma das quais pode ter a penúltima produção; logo, há 5x4x3x2 = 120 distintos arranjos para as 4 cultivares de maiores produções entre as 5. Como há apenas uma cultivar restante, o número total de ordenações ou permutações das cultivares quanto às grandezas de suas médias é 5! = 5x4x3x2x1 = 120.

Regra das permutações com repetições

O número de diferentes ordenações ou permutações de n elementos classificados em k categorias cada uma delas de elementos idênticos, com n_i elementos na i-ésima categoria ($n_1 + n_2 + ... + n_k = n$) é:

$$P = \frac{n!}{n_1! n_2! ... n_k!}.$$

Para verificar essa expressão, observe-se que das n! permutações dos n elementos $n_1!$ serão idênticas em decorrência da identidade dos n_1 elementos da categoria 1. Assim, apenas $n!/n_1!$ ordenações poderão diferir quanto aos elementos das demais categorias. Entretanto, $n_2!$ dessas ordenações serão idênticas, dada a identidade dos n_2 elementos da categoria 2, de modo que apenas $n!/n_1!n_2!$ ordenações poderão diferir quanto aos elementos das categorias 3,...,k. Estendendo-se esse raciocínio, conclui-se que apenas $n!/n_1!n_2!...n_k!$ ordenações dos n elementos serão distintas.

Uma generalização da regra das permutações é a regra dos arranjos.

Regra dos arranjos

O número de diferentes **arranjos** que podem ser formados com r elementos ou objetos selecionados de um grupo de n elementos distintos ($r \leq n$), denotado por A_n^r e denominado **número de arranjos de r elementos tomados de n**, é o número de ordenações que podem ser formadas com r desses n elementos, que é determinado pela expressão:

$$A_n^r = \underbrace{n(n-1)(n-2)\ldots(n-r+1)}_{\text{r fatores}}$$

Por exemplo, os arranjos de dois objetos tomados de três objetos identificados pelas letras a, b e c são: ab, ac, ba, bc, ca e cb. De modo geral, arranjar r de n objetos eqüivale a colocar um subconjunto de r dos n objetos dentro de uma caixa com r compartimentos, em alguma ordenação. O número de arranjos de r dos n objetos é o número de diferentes ordens de disposição de r dos n objetos.

Exemplo 3.6. Suponha-se que três antibióticos X, Y e Z devem ser testados através da aplicação de cada um deles a um animal diferente e que os três animais devem ser selecionados de um conjunto de doze animais. Quantas diferentes assinalações dos 3 antibióticos são possíveis.

Considerem-se as assinalações dos antibióticos X, Y e Z como três partes de um experimento. Qualquer um dos 12 animais disponíveis pode ser escolhido para a aplicação do antibiótico X. Para qualquer escolha de um animal para X, há 11 animais restantes, qualquer um dos quais pode ser escolhido para a assinalação do antibiótico Y. Assim, quanto à assinalação dos antibióticos X e Y, o número de escolhas possíveis, de acordo com a regra do produto, é $12 \times 11 = 132$. Por sua vez, após qualquer uma destas 132 diferentes assinalações dos antibióticos X e Y, há 10 animais restantes a um dos quais pode ser assinalado o antibiótico Z. Logo, o número total das possíveis assinalações dos três antibióticos é $12 \times 11 \times 10 = 1.320$.

Da observação das expressões das regras das permutações e dos arranjos, pode-se derivar a seguinte relação:

$$A_n^r = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!} \quad \text{ou} \quad \frac{P_n}{P_{n-r}}.$$

A regra dos arranjos trata da enumeração de todas as possíveis ordenações de r objetos tomados de uma coleção de n objetos distintos. Em algumas situações, há interesse apenas no número de escolhas possíveis de grupos de r objetos de um conjunto de n objetos, não importando as ordenações dos r objetos escolhidos. O número de "combinações" de r de n objetos, sem consideração para sua ordem de arranjamento, é determinado pela regra que segue.

Regra das combinações

O número das coleções possíveis de r objetos escolhidos de um grupo de n objetos distintos, denotado por C_n^r e designado **número de combinações de r elementos tomados de n**, é:

$$C_n^r = \frac{A_n^r}{r!} = \frac{n(n-1)(n-2)...(n-r+1)}{r(r-1)(r-2)...2.1}.$$

Essa expressão da regra das combinações pode ser derivada pela aplicação da regra da multiplicação a um processo de duas etapas, da seguinte forma: Os arranjos de r objetos tomados de n podem ser obtidos pela seleção de r objetos de n, seguida das permutações dos r objetos selecionados. Segue-se, portanto, que:

$$A_n^r = C_n^r\, r!; \text{ donde: } C_n^r = \frac{A_n^r}{r!}.$$

Propriedades das combinações:

1) $C_n^r = C_n^r = \frac{n!}{r!(n-r)!}$;

2) $C_n^1 = n$;

3) $C_n^n = 1$;

4) $C_n^0 = 1$.

Observe-se que:

$$C_n^0 \text{ e } C_n^n = \frac{n!}{n!0!} = \frac{1!}{0!}.$$

Assim, para que a expressão das combinações permaneça válida para r=0 e r=n, define-se 0! = 1.

Exemplo 3.7. Doze provadores são disponíveis dos quais 8 são mais experientes (designados M_1, M_2,..., M_8) e 4 são menos experientes (designados m_1, m_2, m_3, m_4). Três desses 12 provadores serão selecionados para uso em um experimento de degustação. a) Considerando a ordem de seleção como importante, quantos grupos distintos de 3 provadores poderão ser selecionados? b) Quantos grupos são possíveis com os dois primeiros provadores mais experientes e o último menos experiente? c) Se os 3 provadores são selecionados aleatoriamente (por um processo que atribui a todos os provadores restantes após cada seleção a mesma probabilidade de serem selecionados), qual é a probabilidade de que os dois primeiros provadores sejam mais experientes e o último menos experiente?

a) Há 12 provadores distintos dos quais 3 devem ser selecionados, um após o outro (sem repetição). Assim, o número de grupos que se distinguem pela ordenação dos provadores selecionados é: $A_{12}^3 = 12 \times 11 \times 10 = 1.320$.

b) O número de modos em que os dois primeiros provadores mais experientes podem ser selecionados é o número de arranjos de dois provadores do grupo de 8 provadores mais

experientes: M_1, M_2, ...,M_8, ou seja: $A_8^2 = 8\times7=56$. O número de modos em que o terceiro provador selecionado pode resultar menos experiente quando os dois primeiros são mais experientes é $A_4^1 = 4$. Logo, pela regra do produto, o número de modos em que podem resultar os dois primeiros provadores mais experientes e o terceiro menos experiente é: $A_8^2\times A_4^1 = 56\times4 =224$.

c) A escolha aleatória de 3 dos 12 provadores assegura que todos os A_{12}^3 distintos grupos ordenados são igualmente prováveis. Assim, o evento A: os dois primeiros provadores mais experientes e o terceiro menos experiente consiste de $A_8^2\times A_4^1$ grupos ordenados. Portanto, a probabilidade deste evento pode ser obtida pelo modelo de probabilidade uniforme:

$$P(A) = \frac{n(A)}{n(S)} = \frac{A_8^2\times A_4^1}{A_{12}^3} = \frac{224}{1.320}.$$

Exemplo 3.8. Considere-se a situação do exemplo anterior, mas suponha-se que, agora, a ordem de seleção dos três provadores é irrelevante (o que corresponde à seleção simultânea dos três provadores). a) Quantos distintos grupos são possíveis? b) Em quantos desses grupos dois provadores são mais experientes e um menos experiente? c) Se os 3 provadores são selecionados aleatoriamente (de modo que cada grupo de 3 provadores tenha igual chance de ser selecionado), qual é a probabilidade de que resultem dois provadores mais experientes e um menos experiente?

a) O número de grupos não ordenados de 3 dos 12 provadores é:

$$C_{12}^3 = \frac{12\times11\times10}{1\times2\times3} = 220.$$

b) O número de combinações de 2 provadores tomados dos 8 provadores mais experientes é: $C_8^2 = \frac{8\times7}{1\times2} = 28$, e o número de 8 combinações de 1 provador tomado dos 4 provadores menos experientes é: $C_4^1 = 4$. Pela regra do produto, o número de resultados possíveis que satisfazem às condições especificadas é: $C_8^2\times C_4^1 = 28\times4 = 112$.

c) Segundo o processo de seleção, os 220 grupos distintos de 3 provadores dos 12 provadores disponíveis são todos igualmente prováveis. O número de grupos com exatamente dois provadores mais experientes e 1 menos experiente é 112. Logo, pelo conceito de probabilidade em espaço básico finito equiprovável, a probabilidade do evento A: dois provadores mais experientes e um provador menos experiente é:

$$P(A) = \frac{C_8^2\times C_4^1}{C_{12}^3} = \frac{112}{220}.$$

Exemplo 3.9. Em uma região, há 15 agricultores associados a uma cooperativa dos quais 9 são favoráveis, 4 são desfavoráveis e 2 são indiferentes à adoção de uma nova tecnologia. Três desses agricultores devem ser selecionados, aleatoriamente, para entrevista em uma pesquisa de opinião sobre a adoção da referida tecnologia. a) Qual é a probabilidade de que pelo menos dois dos agricultores selecionados sejam favoráveis à adoção da nova tecnologia? Qual é a probabilidade de que os dois primeiros selecionados sejam favoráveis e o terceiro desfavorável à adoção da tecnologia?

a) O evento de interesse não envolve a ordem das pessoas no processo de amostragem. O número de diferentes grupos de 3 pessoas que podem ser selecionados de um conjunto de 15 pessoas é $C_{15}^3 = 455$. Em um processo de seleção aleatória esses 455 grupos podem ocorrer com uma mesma probabilidade. O evento A: pelo menos dois agricultores favoráveis, eqüivale ao evento A_1 ou A_2 , onde A_1 : dois agricultores favoráveis e A_2 : três agricultores favoráveis. Logo,

$$P(A) = P(A_1 \text{ ou } A_2) = P(A_1) + P(A_2).$$

Para calcular $P(A_1)$, observe-se que o evento A_1 significa, de fato, dois agricultores favoráveis e um não favorável (ou seja desfavorável ou indiferente). Como dos 15 agricultores 9 são favoráveis, o número de grupos de 2 agricultores favoráveis que pode resultar é $C_9^2 = \frac{9\times8}{2\times1} = 36$. Da mesma forma, 1 agricultor não favorável ocorre se qualquer um dos 6 agricultores desfavoráveis ou indiferentes é selecionado. Logo,

$$n(A_1) = C_9^2 \times C_6^1 = 36\times6 = 216.$$

Por semelhante raciocínio, obtém-se:

$$n(A_2) = C_9^3 \times C_6^0 = \frac{9\times8\times7}{3\times2\times1}\times1 = 84.$$

Empregando o conceito de probabilidade uniforme, obtém-se:

$$P(A_1) = \frac{n(A_1)}{n(S)} = \frac{C_9^2 \times C_6^1}{C_{15}^3} = \frac{216}{455},$$

$$P(A_2) = \frac{n(A_2)}{n(S)} = \frac{C_9^3 \times C_6^0}{C_{15}^3} = \frac{84}{455}.$$

Portanto,

$$P(A) = \frac{216}{455} + \frac{84}{455} = \frac{300}{455}.$$

b) Esta questão envolve a consideração da ordem das pessoas no processo de seleção. Portanto, a regra das permutações deve ser empregada. O número de grupos ordenados de 3 pessoas de um conjunto de 15 é $P_{15}^3 = 15\times14\times13 = 2.730$. Por outro lado, as duas primeiras pessoas serão favoráveis se provirem das 9 favoráveis e a terceira será desfavorável se provir das 4 desfavoráveis. Logo, o número de grupos de 3 agricultores os dois primeiros dos quais são favoráveis e o terceiro é desfavorável é: $P_9^2 \times P_4^1 = (9\times8)\times4 = 288$. Portanto, P(dois primeiros favoráveis e terceiro desfavorável) = $\frac{288}{2.370}$

3.4 Conceito Empírico de Probabilidade

Na maioria dos experimentos, o espaço básico não é equiprovável. É o caso por exemplo, do experimento de lançamento de um dado viciado. Nessas circunstâncias, a probabilidade de um evento não pode ser obtida por derivação lógica, usando a propriedade de simetria do experimento. Para a determinação da probabilidade de um evento, deve-se considerar com que freqüência um

evento ocorre quando o experimento é repetido sob, essencialmente, as mesmas condições. Quando um experimento é repetido n vezes, define-se a **freqüência relativa** de um evento A nas n realizações do experimento pela razão:

$$f_n(A) = \frac{\text{Numero de vezes em que A ocorre}}{n},$$

isto é, a proporção de vezes em que o evento A efetivamente ocorreu nas n repetições do experimento. Esse valor varia com diferentes conjuntos de n repetições do evento A. Para uma mesma seqüência de realizações do experimento, ele também flutua quando o número n de repetições varia. Entretanto, a experiência comum com um grande número de experimentos aleatórios, em muitos campos, indica que, se as condições experimentais permanecem razoavelmente constantes, a freqüência relativa f(A) tende a estabilizar-se em torno de um número único, quando o número de repetições do experimento cresce indefinidamente. Esse valor em torno do qual tende a estabilizar-se a freqüência relativa $f_n(A)$ é tomado como a **probabilidade** do evento A, designada por P(A).

Essa propriedade da freqüência relativa pode ser efetivamente verificada para um experimento particular. Seja, por exemplo, o experimento "lançamento de uma moeda e observação da face que resulta voltada para cima", com dois resultados possíveis: C = {cara} e c = {coroa}. Suponha-se que se realiza esse experimento, sucessivamente, n vezes e que se registra o resultado de cada uma de suas realizações e a correspondente freqüência relativa do evento {cara}. Pode-se dispor essas informações na forma da **Tabela 3.1**.

Tabela 3.1. Registro dos resultados de dezesseis repetições do lançamento de uma moeda e correspondentes freqüências relativas.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
C		x	x	x				x	x	x				x		
c	x				x	x	x				x	x	x		x	x
f_C	0	1/2	2/3	3/4	3/5	3/6	3/7	4/8	5/9	6/1	6/1	6/1	6/1	7/1	7/1	7/1

Para melhor apreciação, as freqüências relativas do evento C, até cada lançamento, são representadas no gráfico da **Figura 3.2**.

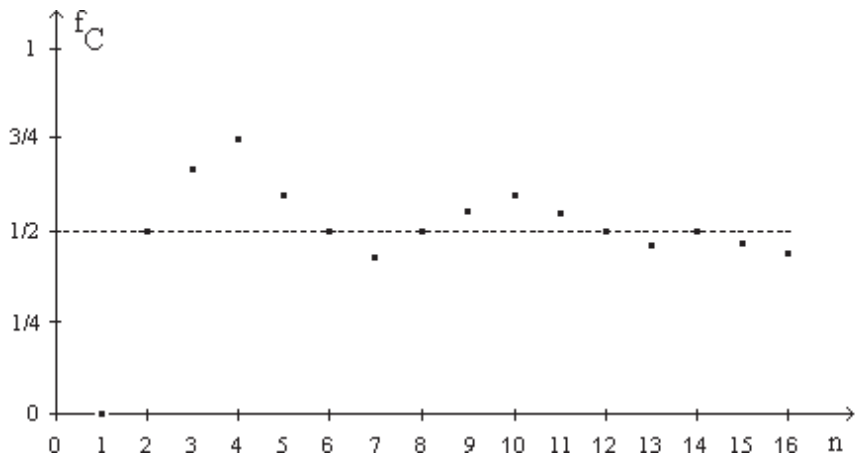


Figura 3.2. Representação gráfica das freqüências relativas do evento "cara" em uma sucessão de 16 lançamentos de uma moeda.

A observação da variação da freqüência relativa do evento {cara} indica que ela tende a aproximar-se de um número desconhecido, próximo de 1/2. Se a moeda é ideal (perfeitamente homogênea e simétrica) e o lançamento é não tendencioso, tal número é 1/2. Essa propriedade, que se observa para todos os experimentos aleatórios, é o fundamento para o **conceito empírico de probabilidade**:

A **probabilidade** de um evento C é o número real P(C) em torno do qual tende a estabilizar-se a freqüência relativa $f_n(C)$ quando o número n de realizações do experimento cresce indefinidamente:

$$f_n(C) \xrightarrow{n\uparrow} P(C).$$

3.5 Conceito de Probabilidade em Espaço Básico Discreto

A freqüência relativa satisfaz às seguintes propriedades importantes, que constituem a base para o estabelecimento do conceito de probabilidade para espaço básico discreto:

1) A freqüência relativa é o quociente de dois números inteiros não negativos em que o numerador não pode ser maior que o denominador; portanto, ela é um número racional entre 0 e 1; logo $0<f(A)<1$ para qualquer evento A.
2) Como algum resultado sempre deve ocorrer em qualquer realização do experimento, a freqüência relativa do espaço básico é 1: $f(S)=1$.
3) A freqüência relativa de um evento A é igual à soma das freqüências relativas de todos os eventos simples incluídos em A.

Dessas propriedades da freqüência relativa deriva-se o seguinte conceito de **probabilidade em espaço básico discreto**:

A **probabilidade** é uma função P( ) definida para eventos do espaço básico S que satisfaz às seguintes condições:

1) Para qualquer evento A, P(A) é um número real não negativo: P(A)>0.

2) A probabilidade do espaço básico é igual a 1: P(S)=1.

3) A probabilidade de qualquer evento A é a soma das probabilidades de todos os eventos simples incluídos em A: $P(A) = \sum_i p(s_i)$, $s_i \in A$.

Qualquer função que satisfaça a essas três condições é uma **função de distribuição de probabilidade**, também designada, mais simplesmente, **função de probabilidade** ou **distribuição de probabilidade**. Sob o ponto de vista prático, entretanto, há interesse apenas em funções cujos valores numéricos obtidos para probabilidades de eventos concordam com as freqüências relativas dos correspondentes eventos em uma longa série de realizações de um experimento particular.

Exemplo 3.10. Experimento: Observação dos sexos de gêmeos em partos duplos de ovelhas de um rebanho, cujo espaço básico é: S={MM, MF,FM,FF}. Considere-se as duas seguintes funções $P_1( )$ e $P_2( )$ definidas para os eventos de S:

a) $P_1(MM) = P_1(MF) = P_1(FM) = P_1(FF) = 1/4$; e

b) $P_2(MM) = P_2(FF) = 3/8$; $P_2(MF) = P_2(FM) = 1/8$.

Essas duas funções satisfazem às 3 condições de uma função de probabilidade; portanto, são funções de probabilidade. Entretanto, para um experimento particular, a atribuição de probabilidades deve corresponder com bastante aproximação às freqüências relativas quando um grande número de partos duplos de ovelhas do rebanho é observado. Em muitas aplicações práticas, a função de probabilidade $P_1( )$ é apropriada.

3.6 Operações de Eventos

A terceira condição de uma função de probabilidade em espaço básico discreto caracteriza que a probabilidade de qualquer evento em tal espaço básico é a soma das probabilidades dos eventos simples que implicam na ocorrência do evento. Quando o evento A é de natureza complexa, a determinação de P(A) através da adição das probabilidades dos eventos simples em A pode ser trabalhosa. Nesse caso, pode ser mais conveniente expressar o evento A em termos de outros eventos mais simples cujas probabilidades possam ser calculadas com menos trabalho. Assim, é importante saber as formas como um evento pode ser expresso em termos de outros eventos, através de operações de eventos. Para a melhor compreensão das operações de eventos, é interessante recorrer à representação gráfica do espaço básico como um conjunto de pontos em um diagrama onde cada ponto corresponde a um evento simples. Nesse diagrama, cada evento é representado como um conjunto de pontos que satisfazem à descrição do evento. Essa representação gráfica é designada **diagrama de Venn**.

Exemplo 3.11. Experimento: Administração de um antibiótico a três animais e observação do resultado para cada animal: reage (R) ou não reage (N). Considere-se os seguintes eventos:

A: Apenas o primeiro animal responde;

B: O primeiro animal responde;

C: Os dois primeiros animais não respondem.

Há dois resultados possíveis para o primeiro animal: R e N, e cada um desses dois resultados pode ser seguido de um de dois resultados possíveis do segundo animal: R e N. Cada um desses 2×2=4 resultados possíveis para os dois primeiros animais também pode ser seguido dos dois resultados possíveis para o terceiro animal. Portanto, para os 3 animais, há 2×2×2=8 resultados possíveis, isto é, 8 eventos simples para o experimento. Esses eventos simples podem ser identificados como segue: s_1 =RRR, s_2 =RRN, s_3 =RNR, s_4 =NRR, s_5 =RNN, s_6 =NRN, s_7 =NNR, s_8 =NNN. Com essa notação, os eventos A, B e C são:

$A = \{s_5\}$;

$B = \{s_1, s_2, s_3, s_5\}$;

$C = \{s_7, s_8\}$.

O diagrama de Venn é apresentado na **Figura 3.3**, onde o retângulo maior representa o espaço básico.

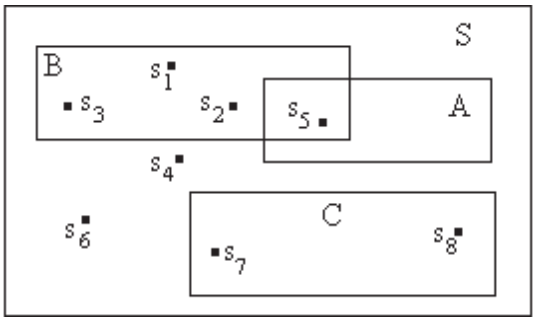


Figura 3.3. Diagrama de Venn para o espaço básico e os três eventos considerados no **Exemplo 3.11**.

Define-se, a seguir, três eventos resultantes de três operações básicas de eventos:

O evento **união** ou **reunião** de A e B, denotado por $A \cup B$, é o conjunto de todos os eventos simples que estão em A ou em B, ou em ambos A e B. O evento $A \cup B$ ocorre se pelo menos um dos dois eventos A e B ocorre.

O evento **interseção** de A e B, denotado por $A \cap B$, é o conjunto de todos os eventos simples que pertencem simultaneamente a ambos eventos A e B. O evento $A \cap B$ ocorre se ambos os eventos A e B ocorrem.

O evento **complemento** de A, denotado por A^c, é o conjunto de todos os eventos elementares que não estão contidos em A. O evento A ocorre se o evento A não ocorre.

Essas três operações básicas de eventos são ilustradas na **Figura 3.4**. Observe-se que o evento A∪B é o conjunto maior, contendo os eventos A e B, e o evento A∩B é o conjunto menor contido em ambos A e B.

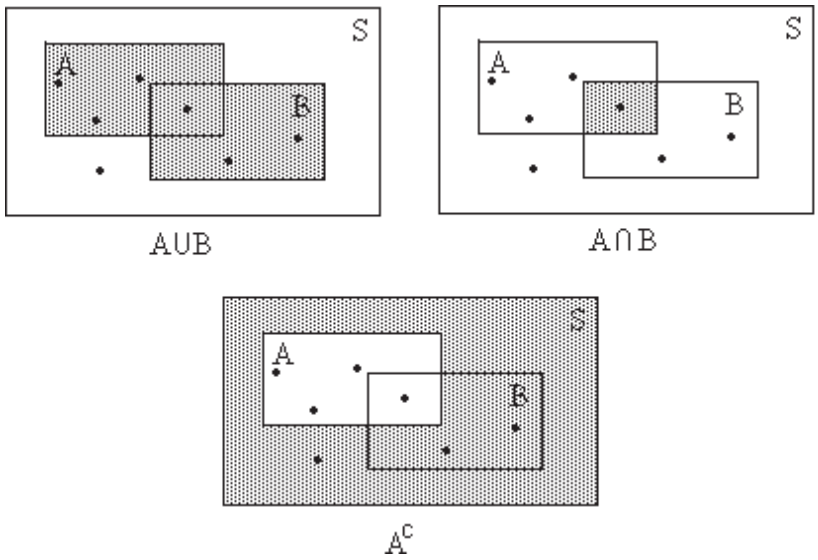


Figura 3.4. Ilustração das três operações básicas de eventos - reunião, interseção e complementação, através de Diagramas de Venn.

Os seguintes conceitos são importantes:

O espaço básico é designado **evento certo**, visto que sua probabilidade é igual a 1.

Os eventos A e B são **disjuntos** ou **mutuamente exclusivos** se não têm eventos simples em comum. Nesse caso, diz-se que o evento A∩B é um **evento impossível**, denotado por ∅.

Exemplo 3.12. Para ilustração, considere-se os eventos A, B e C do experimento do **Exemplo 3.11**, representados na **Figura 3.3**. De acordo com as definições das 3 operações básicas, obtém-se:

- $B \cup C = \{s_1, s_2, s_3, s_5, s_7, s_8\}$;
- $A \cap B = \{s_5\}$;
- $B^c = \{s_4, s_6, s_7, s_8\}$;
- $(B \cup C)^c = \{s_4, s_6\}$;
- $B^c \cap C^c = \{s_4, s_6\}$.

As operações de reunião e de interseção se estendem à qualquer número de eventos. Por exemplo, a reunião dos eventos A, B e C é o evento constituído por todos os eventos simples contidos em pelo menos um dos eventos A, B e C, designado por A∪B∪C; a **interseção dos**

eventos A, B e C é o evento correspondente à ocorrência simultânea dos três eventos A, B e C, designado por $A \cap B \cap C$.

3.7 Probabilidade Condicional

A probabilidade de um evento A pode alterar-se quando alguma informação é provida referente à ocorrência de um outro evento B.

> A probabilidade de um evento A quando é sabido que outro evento B ocorreu é denominada **probabilidade condicional** de A dado Be é denotada P(A|B).

Exemplo 3.13. Em um levantamento referente à adoção de uma nova tecnologia no cultivo do morangueiro, foram obtidos os seguintes dados referentes à 50 produtores de morango de um município, classificados em duas categorias quanto ao número de anos de cultivo do morangueiro:

Adoção da tecnologia	Anos de cultivo do morangueiro	
	<10 anos	>10 anos
Adota	16	4
Não adota	10	20

Considere-se o evento A: seleção de um produtor que adota a nova tecnologia. Se um indivíduo é selecionado aleatoriamente dessa população de 50 produtores, a probabilidade de que ele adote a nova tecnologia é:

$$P(A) = \frac{n(A)}{n(S)} = \frac{20}{50} = \frac{2}{5}.$$

Entretanto, a probabilidade de que um produtor de morango adote a nova tecnologia sob a condição de que ele cultive o morangueiro há mais de 10 anos é:

$$P(A|B) = \frac{4}{24} = \frac{1}{16},$$

onde B denota a seleção de um produtor de morangueiro há mais de 10 anos. Observe-se que, agora, o espaço básico é restrito à S' = B: produtores que cultivam o morangueiro há mais de 10 anos, de modo que n(S') = 24, enquanto que o número destes produtores que adotam a nova tecnologia é 4.

Observe-se que o numerador de P(A|B) é $n(A \cap B) = 4$, ou seja, o número de produtores de morangueiro por mais de 10 anos e que adotam a nova tecnologia, e o denominador é n(B), ou seja, o número de produtores de morangueiro há mais de 10 anos. Assim, a probabilidade condicional pode ser expressa por:

$$P(A|B) = \frac{n(A \cap B)}{n(B)},$$

ou, dividindo numerador e denominador por n(S),

$$P(A|B) = \frac{n(A \cap B) / n(S)}{N(B) / n(S)} = \frac{P(A \cap B)}{P(B)}.$$

Dessa forma, a probabilidade condicional P(A|B) pode ser expressa em termos de duas probabilidades definidas para o espaço básico completo S. Naturalmente, a consideração da probabilidade condicional faz sentido apenas se o evento condicionante não é nulo, ou seja, se $P(B) \neq 0$.

Essa expressão é válida para qualquer situação. Assim, se A e B são dois eventos quaisquer em um espaço básico S e $P(B) \neq 0$, a probabilidade condicional de A dado B é dada por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Semelhantemente, a probabilidade condicional de B dado A é expressa por:

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

Dessas duas expressões obtém-se, respectivamente:

$P(A \cap B) = P(B) \times P(A|B)$, $P(B) \neq 0$, e

$P(A \cap B) = P(A) \times P(B|A)$, $P(A) \neq 0$,

que exprimem a denominada "lei da multiplicação" da probabilidade. Assim, a **lei da multiplicação** estabelece que a probabilidade de que dois eventos A e B ocorram simultaneamente é o produto da probabilidade de A pela probabilidade condicional de B dado A. Alternativamente, é o produto da probabilidade de B pela probabilidade condicional de A dado B.

A lei da multiplicação pode ser imediatamente estendida à mais de dois eventos. Para três eventos, por exemplo, ela estabelece: se A, B e C são quaisquer três eventos em um espaço básico S, tais que $P(A) \neq 0$ e $P(A \cap B) \neq 0$, então: $P(A \cap B \cap C) = P(A) \times P(B|A) \times P(C|A \cap B)$.

Exemplo 3.14. Um fazendeiro tem uma caixa com 30 ovos dos quais 5 têm manchas de sangue. Ele retira 3 ovos da caixa aleatoriamente, um após o outro. Qual é a probabilidade de que os dois primeiros ovos tenham manchas de sangue e o terceiro seja limpo?

Denote-se por C_1, C_2 e S_3, respectivamente, os eventos de escolher ovos manchados na primeira e segunda extração e um ovo sem mancha na terceira. Deseja-se determinar a probabilidade do evento $C_1 \cap C_2 \cap S_3$. Segundo a lei da multiplicação,

$$P(C_1 \cap C_2 \cap S_3) = P(C_1) \times P(C_2|C_1) \times P(S_3|C_1 \cap C_2).$$

Como as extrações são procedidas aleatoriamente, tem-se $P(C_1)= 5/30$. Para determinar $P(C_2|C_1)$ note-se que quando C_1 ocorre há 29 ovos restantes na caixa, 4 dos quais são manchados. Assim, a probabilidade condicional de retirar um ovo manchado na segunda extração é dada por $P(C_1|C_2)=4/29$. Pelo mesmo raciocínio, encontra-se $P(S_3|C_1 \cap C_2)=25/28$. Logo, pela fórmula da multiplicação:

$$P(S_3|C_1 \cap C_2) = \frac{5}{30} \times \frac{4}{29} \times \frac{25}{28} = \frac{25}{1218}.$$

3.8 Independência Estatística

Uma situação particularmente importante é aquela em que a probabilidade condicional do evento A dado o evento B, P(A|B) é igual à probabilidade não condicional de A, ou seja, P(A|B) = P(A). Nessa situação, a ocorrência ou não ocorrência do evento B não afeta a probabilidade de A. Diz-se, então, que os eventos A e B são **independentes no sentido estatístico**:

> Dois eventos A e B são **independentes** (no sentido estatístico) se:
>
> P(A|B) = P(A).

Condições de independência equivalentes são: P(B|A) = P(B) e P(A∩B) = P(A).P(B), onde a última igualdade é a **lei da multiplicação** sob a condição de independência estatística. Na derivação desta expressão, pressupõe-se que P(A|B) ou P(B|A) existe e que, portanto, P(B) ≠ 0 ou P(A) ≠ 0. Entretanto, a lei da multiplicação também é válida quando P(A)=0 e P(B)= 0. A última condição mostra que a independência estatística é simétrica em A e B: Se A é independente de B, então, B é independente de A.

Se dois eventos não são independentes, eles são ditos **dependentes**.

Exemplo 3.15. Considere-se a situação de amostragem com reposição, ilustrada no **Exemplo 3.2**: Extração aleatória de dois sacos de sementes de uma pilha de 5 sacos, sendo 2 da variedade V_1 e 3 da variedade V_2 . Considerem-se os dois eventos A: um saco V_1 na primeira extração e B: um saco V_2 na segunda extração. Segundo aquele exemplo, tem-se:

$$P(A\cap B) = \frac{2\times 3}{5\times 5} = \frac{6}{25}.$$

Por outro lado, $P(A) = \frac{2}{5}$ e $P(B) = \frac{3}{5}$, donde resulta:

$$P(A).P(B) = \frac{2}{5}\times\frac{3}{5} = \frac{6}{25}.$$

Logo,

$$P(A\cap B) = P(A)\times P(B),$$

donde se conclui que os eventos A e B são estatisticamente independentes.

Exemplo 3.16. Considere-se, agora, os dois eventos A e B definidos no exemplo anterior para a situação de amostragem se reposição do **Exemplo 3.3**. Nesse caso, tem-se:

$$P(A\cap B) = \frac{2\times 3}{5\times 4} = \frac{6}{20},$$

enquanto que:

$$P(A) = \frac{2}{5} \text{ e } P(B) = \frac{3}{4}, \text{ donde: } P(A).P(B) = \frac{6}{25},$$

de modo que: P(A∩B) ≠ P(A).P(B). Logo, os dois eventos são, agora, dependentes.

O conceito de independência estatística é estendido para mais de dois eventos, como segue: Os eventos A_1, A_2,..., A_k são (**mutuamente**) **independentes** se e apenas se a probabilidade da interseção de qualquer 2, 3,...,k destes eventos é igual ao produto de suas respectivas probabilidades.

Para três eventos A, B e C, por exemplo, a independência implica que $P(A\cap B) = P(A)\times P(B)$, $P(A\cap C) = P(A)\times P(C)$, $P(B\cap C) = P(B)\times P(C)$ e $P(A\cap B\cap C) = P(A)\times P(B)\times P(C)$. Deve ser observado que três ou mais eventos podem ser independentes par a par sem serem mutuamente independentes.

Exemplo 3.17. Suponha que a iluminação de uma sala é controlada por três interruptores separados, de modo que as luzes estarão acesas quando os três interruptores estiverem ligados ou quando um dos interruptores estiver ligado e os outros dois desligados, com as mesmas probabilidades de iluminação da sala para essas quatro condições dos três interruptores. Considerem-se os eventos A: o primeiro interruptor está ligado, B: o segundo interruptor está ligado e C: o terceiro interruptor está ligado. Estes eventos e as correspondentes probabilidades associadas com as diversas situações dos interruptores estarem ligados ou desligados quando as luzes da sala estão acesas são representados no diagrama de Venn da **Figura 3.5**.

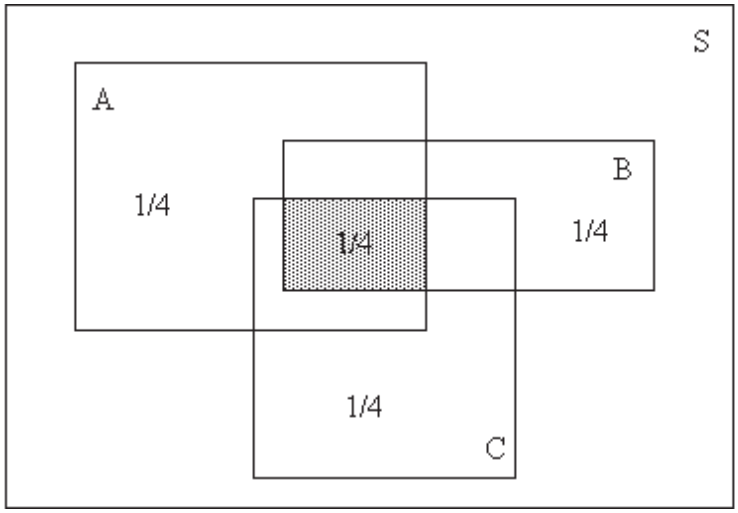


Figura 3.5. Diagrama de Venn para o **Exemplo 3.17**.

Pode-se verificar que cada um dos pares de eventos A, B e C são independentes, mas os três eventos A, B e C não são independentes. De fato, segundo o diagrama, $P(A)=P(B)=P(C) = 1/2$, $P(A\cap B) = P(A\cap C) = P(B\cap C) = 1/4$ e $P(A\cap B\cap C) = 1/4$. Logo, $P(A\cap B)= 1/4 =P(A)\times P(B)$, $P(A\cap C) = 1/4= P(A)\times P(C)$, $P(B\cap C) = 1/4 = P(B)\times P(C)$. Entretanto, $P(A\cap B\cap C) = 1/4$, enquanto que $P(A)\times P(B)\times P(C) = 1/8$.

Por outro lado, pode ocorrer que $P(A\cap B\cap C) = P(A)\times P(B)\times P(C)$ sem que A, B e C sejam independentes par a par.

Observação. Freqüentemente, os termos "mutua exclusividade" e "independência estatística" de eventos são confundidos. Estas duas propriedades de eventos são muito diferentes.

De fato, uma implica que a outra não pode ser verdadeira. Considere-se, por exemplo, dois eventos A e B com probabilidades positivas. Se eles são mutuamente exclusivos, a interseção $A \cap B$ é vazia e, portanto, $P(A \cap B)=0$. Se esses eventos fossem independentes, eles teriam que satisfazer à condição $P(A).P(B) = P(A \cap B)$, que não pode ser verdadeira porque o produto de dois números não nulos não pode ser zero. Como um exemplo extremo, os eventos A e A^c são mutuamente exclusivos, mas, intuitivamente, eles são muito dependentes, no sentido de que tão logo é dito que A ocorreu se está certo de que A^c não ocorreu. Pode-se verificar isto calculando: $P(A^c|A)=P(A^c \cap A)/P(A)=0/P(A)=0$.

3.9 Conceito Geral de Probabilidade

A discussão anterior foi confinada à situação de espaço básico discreto, em que o número de eventos simples é finito ou infinito contável. Em muitos experimentos, entretanto, a mensuração é efetuada em uma escala contínua, como é o caso, por exemplo, com características como altura, peso e temperatura. Nessa circunstância, o espaço básico é, usualmente, o conjunto de todos os números reais de um intervalo. Tal espaço básico é designado espaço básico contínuo. A interpretação da probabilidade de um evento como sua freqüência relativa em uma longa série de experimentos e a maior parte das propriedades de probabilidade permanecem válidas para esses espaços básicos. Uma notável exceção é o fato de que a probabilidade de um evento em um espaço básico contínuo não é obtida como a soma das probabilidades dos eventos simples que o compõe, já que a soma de um número infinito não contável de termos não tem sentido, ou seja, a operação de soma não é definida se os termos a serem somados não podem ser arranjados em uma seqüência. Este é o caso, por exemplo, quando o evento é um intervalo e os eventos elementares são os pontos do intervalo. Por essa razão, não se pode determinar a probabilidade de um evento pela adição das probabilidades de seus eventos simples. A estrutura de probabilidade de um espaço básico contínuo deve alterar-se para a atribuição de probabilidades a eventos (tipicamente subintervalos de um intervalo de números reais) de um outro modo.

O conceito de probabilidade para esse espaço básico mais geral é estabelecido através das condições que a atribuição de probabilidades deve satisfazer, em coerência com a experiência empírica. Essas condições, coerentes com as propriedades da freqüência relativa e coincidentes com as propriedades da probabilidade em espaço básico discreto, sugerem o seguinte **conceito axiomático de probabilidade**:

> **Probabilidade** é uma função real P( ) definida no espaço básico S que satisfaz às seguintes condições (postulados):
>
> 1) $P(A) > 0$ para qualquer evento A.
>
> 2) $P(S) = 1$.
>
> 3) Se $A_1, A_2,...$ é uma seqüência de eventos mutuamente exclusivos, então:
>
> $P(A_1 \cup A_2 \cup ...) = P(A_1)+P(A_2)+...$

As propriedades da probabilidade estabelecidas para espaço básico discreto podem, agora, ser derivadas como casos particulares destas três condições (axiomas). Assim, por exemplo, para

obter a **regra da complementação**, note-se que um evento A e seu complemento A são disjuntos e que $A \cup A^c = S$. Logo, pelos postulados 3) e 2), obtém-se:

$$P(A \cup A^c) = P(A)+P(A^c),$$

donde:

$$P(S) = P(A)+P(A^c);$$

logo,

$$P(A^c) = 1-P(A).$$

De modo semelhante, pode-se derivar as demais regras importantes da probabilidade para um espaço básico geral. As definições de probabilidade condicional e de independência estatística permanecem as mesmas de espaço básico discreto.

3.10 Modelos de Probabilidade

Para analisar os resultados de um experimento aleatório, é conveniente construir um modelo, ou seja, uma representação abstrata que descreva a estrutura do experimento e possa ser usada na análise. Modelos para experimentos aleatórios são denominados **modelos de probabilidade**.

> Um **modelo de probabilidade** é estabelecido pela especificação de três componentes: o espaço básico S, a classe de todos os eventos em S e uma função que permita a determinação de probabilidades de eventos de S.

O segundo componente é facilmente caracterizado quando o espaço básico é finito. Nesse caso, a classe de eventos de S é a coleção de todos os eventos do espaço básico, incluídos o evento certo S e o evento impossível $\varnothing$. Assim, no exemplo a) utilizado para ilustrar experimento aleatório, a classe dos eventos de S é a coleção de eventos $\{\{M\},\{F\},S,\varnothing\}$. Essa definição é generalizada para espaço básico infinito contável e, portanto, abrange, em geral, espaços básicos discretos. Entretanto, a definição da classe de eventos de S não é tão simples para espaço básico infinito não contável, quando os pontos de S constituem um "continuum", como nos exemplos e) e f). Essa situação somente pode ser tratada com desenvolvimentos teóricos mais avançados, fora dos propósitos deste texto. Entretanto, ela não é um problema sério para os métodos estatísticos. Nas aplicações, é suficiente reconhecer que os eventos são subconjuntos do espaço básico.

3.11 Exercícios

1. Especifique o espaço básico e o espaço de eventos para os seguintes experimentos aleatórios:
   a) Plantio de uma semente e observação do resultado da germinação.
   b) Plantio de duas sementes e observação do resultado da germinação.
2. Especifique o espaço básico para cada um dos seguintes experimentos aleatórios:
   a) Número de animais que revelam resposta positiva à um hormônio de crescimento do total de 100 animais de um rebanho aos quais se administra o hormônio.
   b) A temperatura no próximo dia 29 de março às 9 horas.

c) Número de animais de um rebanho de 50 bovinos da raça Hereford com peso superior a 300 kg.

d) Em um levantamento referente ao consumo de vinho no Rio Grande do Sul, 1.000 pessoas serão solicitadas a responder "sim" ou "não" à questão: "Você bebeu vinho na semana passada?" Apenas o número de pessoas que responderem "sim" será registrado.

e) Escolha aleatória de uma entre 6 garrafas de vinho tinto das cultivares C_1, C_2, C_3, C_4, C_5 e C_6.

f) Escolha aleatória de uma entre 6 garrafas de vinho tinto das cultivares C_1, C_2, C_3, C_4, C_5 e C_6 de cada uma de 2 caixas, uma com rótulos brancos e a outra com rótulos verdes.

g) Classificação de 4 latas de compota de pêssego das cultivares A, B, C e D por um provador quanto à preferência com respeito a uma propriedade organolética, com omissão da identificação das cultivares para o provador.

h) Escolha aleatória de 2 sementes de um pacote contendo 5 sementes de idêntica aparência, sendo que 2 são de plantas de flores brancas e 3 de plantas de flores azuis, e observação das cores das flores das plantas resultantes.

i) Extração aleatória de 2 maçãs de um saco contendo 2 maçãs da cultivar A e 3 maçãs da cultivar B, com reposição da primeira maçã extraída, e identificação das maçãs resultantes.

j) Um provador afirma ser capaz de distinguir entre os vinhos Cabernet Franc de 3 diferentes cantinas. Um copo de vinho de cada uma das 3 cantinas é dado ao provador para identificação, em uma ordem desconhecida. O número de identificações corretas será registrado.

k) O número de horas que um aparelho novo de TV trabalhará sem necessidade de reparo.

l) O número de helmintos encontrados no corpo de um bovino da raça Ibagé na necropsia, após o período de 2 meses em uma pastagem infectada.

m) O número de mortes em acidentes de trânsito em um estado, no próximo ano.

3. Especifique o seguinte evento para o experimento aleatório do exercício 1.b): A={Duas sementes germinam}; B={Pelo menos uma semente germina}; e C={Nenhuma semente germina}.

a) {mais de 30 animais reagem à aplicação do hormônio}.

b) {Temperatura superior a 21 graus C}.

c) {Pelo menos 45 animais com peso superior a 300 kg}.

d) {Pelo menos 850 pessoas respondem "sim"}.

e) A={Garrafa de vinho de cultivar identificada com número par}; B={Garrafa de vinho de cultivar identificada com número menor que 4}; C = Complemento de B.

f) A={Ambas garrafas com vinho da cultivar C }; B={Garrafas cuja soma das identificações das cultivares é 5}; C={Garrafa cuja soma das identificações é no mínimo 10}.

4. Em um experimento aleatório de plantio de quatro sementes, especifique os seguintes eventos:
   a) A={Duas sementes germinam}.
   b) B={Pelo menos uma semente germina}.
   c) C={Nenhuma semente germina}.
5. Determine as probabilidades dos eventos especificados na questão 4.e).
6. Determine as probabilidades dos eventos especificados na questão 4.f).
7. Determine as probabilidades dos eventos especificados n a questão 5, sabendo que a probabilidade de uma semente germinar é 0,9.
8. Um mecânico sabe que um carro com seis velas tem duas velas com mau funcionamento. Se ele retira aleatoriamente duas delas para substituição, qual é a probabilidade de que ele efetue a substituição correta das duas velas defeituosas?
9. A probabilidade de um casal ter um filho homem é 0,5. Determine a probabilidade de que:
   a) Tendo dois filhos, ambos sejam mulheres.
   b) Em uma prole de 5 filhos, dois sejam homens.
   c) Em uma prole de 5 filhos, pelo menos dois sejam homens.
10. Em herança mendeliana simples, um caráter físico de uma planta ou animal é determinado por um par simples de genes. O caráter "cor" da ervilha é um exemplo. Ele é determinado por um par de genes A e V que representam, respectivamente, as cores amarela e verde. O gene A é dominante, de modo que plantas com o genótipo (par de genes) VV terão ervilhas verdes e aquelas com os genótipos AA e AV terão ervilhas amarelas. Os descendentes (progênie) recebem um gene de cada pai com igual probabilidade para cada gene de cada pai.
    a) Efetuado o cruzamento de ervilhas de coloração amarela, determine a probabilidade de que resulte uma planta com ervilhas: a) amarelas; b) verdes.
    b) Em um cruzamento entre ervilhas AV e AV, qual é a proporção da progênie que terá grãos amarelos? Qual é a proporção de ervilhas amarelas que terão o genótipo AA?
11. Um outro caráter mendeliano simples é a rugosidade. As ervilhas podem ser lisas ou rugosas. Esse caráter é determinado por um par de genes L e R, respectivamente para as alternativas lisa e rugosa. O caráter liso é dominante, de modo que ervilhas LL e LR são lisas, enquanto que ervilhas RR são rugosas.
    a) Se ervilhas AV-LR são cruzadas com ervilhas VV-RR, quais são os possíveis resultados? Quais são suas correspondentes probabilidades?
    b) Responda semelhantes questões para cruzamentos entre ervilhas AV-LR e VV-RR.
    c) Responda semelhantes questões para cruzamentos entre ervilhas AV-LR e AV-LR.
12. Na questão 2.h), determine as probabilidades dos seguintes eventos: a) A: duas plantas com flores da mesma cor; b) B: duas plantas com flores de cores diferentes.
13. Na questão 2.i), determine as probabilidades dos seguintes eventos: a) A: duas maçãs da cultivar A; b) B: duas maçãs da mesma cultivar; c) C: uma maçã de cada cultivar.

14. Em uma operação em uma linha de montagem de uma indústria, 1/5 dos itens produzidos são defeituosos. Se forem retirados aleatoriamente três itens para teste, qual é a probabilidade de que:
    a) exatamente um item seja defeituoso?
    b) pelo menos um item seja defeituoso?
15. Um saco contém 1.500 batatas das quais 1.100 são classificadas como de primeira e 400 como de segunda. Extraindo-se uma amostra de três batatas, determine a probabilidade de que: a) as três batatas sejam de primeira; b) pelo menos uma batata seja de segunda.
16. Um pesquisador em tecnologia de alimentos tem disponíveis 20 voluntários para avaliação organolética em um experimento e deseja extrair uma amostra aleatória de 10 desses voluntários. Para tal, ele decide escrever todas as combinações de 10 nomes em cartões a, então, extrair um dos cartões ao acaso. Quantas combinações de nomes ele deve anotar em cartões?
17. Um geneticista que pesquisa gado leiteiro tem quatro touros e oito vacas que podem ser usados em um experimento. Quantos diferentes pares constituídos de um macho e uma fêmea ele pode constituir?
18. Há evidência de que entre as formas inferiores de vida animal podem ser transmitidas características de um animal para outro junto com a transferência da substância química conhecida como RNA. Em um experimento referente a esse comportamento de transferência, oito salamandras são divididas em dois grupos de quatro animais. Um grupo constituirá o grupo experimental e ou outro o grupo controle.
    a) Mostre que os dois grupos podem se formados com 70 diferentes constituições.
    b) Qual é a probabilidade de que as quatro salamandras mais ágeis resultem em um mesmo grupo?
    c) Qual é a probabilidade de que três das quatro salamandras mais ágeis situem-se no mesmo grupo?
    d) Todas as salamandras em um grupo (o grupo experimental) recebem RNA de uma salamandra que foi treinada para maior agilidade. O outro grupo (o grupo controle) recebe RNA de uma salamandra não treinada. Para que se possa acreditar que o comportamento é transferido com o RNA, qual deve ser o número de animais mais rápidos no grupo experimental? Justifique a resposta.
19. Uma pessoa afirma que tem habilidade para distinguir entre café percolado e café instantâneo. Em um experimento para testar esta afirmada habilidade, o indivíduo deve provar 10 xícaras de café, cinco com cada um dos dois tipos de café. Para melhoria do experimento, são usados xícaras iguais e o café das duas origens é preparado simultaneamente. Então, as dez xícaras de café são oferecidas ao indivíduo, sem identificação de seu conteúdo, para prova e indicação das cinco que contém café percolado.
    a) De quantas formas ele pode selecionar 5 entre as dez xícaras de café?
    b) De quantas formas ele pode selecionar as 5 xícaras de café percolado?
    c) Qual é a probabilidade de que selecione corretamente as cinco xícaras com café percolado por palpite em vez de por habilidade sensorial?

20. Um lote é formado de 10 artigos bons, 4 com defeitos menores e 2 com defeitos graves. Um artigo é escolhido ao acaso. Determine a probabilidade de que: a) o artigo não tenha defeitos; b) o artigo não tenha defeitos graves; c) o artigo seja perfeito ou tenha defeitos graves.
21. Considere o mesmo lote de artigos do exercício anterior. Dois artigos são escolhidos ao acaso, repondo-se o primeiro artigo extraído antes da extração do segundo. Determine a probabilidade de que: a) ambos os artigos sejam perfeitos; b) ambos tenham defeitos graves; c) pelo menos um seja perfeito; d) no máximo um seja perfeito; e) exatamente um seja perfeito; f) nenhum deles tenha defeitos graves; g) nenhum deles seja perfeito.
22. Responda às mesmas perguntas da questão 16 no caso de extração sem reposição.
23. Em um exame para admissão de laboratoristas, os candidatos são solicitados a efetuar cinco análises químicas e redigir o relatório de seus resultados. Para diminuir o trabalho de avaliação, os examinadores decidem avaliar apenas uma amostra aleatória de dois dos cinco relatórios redigidos por cada candidato. Com base nos resultados da avaliação, os candidatos são classificados em três categorias: A, se seus dois relatórios estão corretos; B, se um relatório está correto e o outro incorreto; e C, se os dois relatórios contém incorreções.
    a) Qual é a probabilidade de um candidato receber o grau A quando ele submeteu cinco relatórios corretos?
    b) Qual é a probabilidade de um candidato receber o grau C quando ele submeteu: i) cinco relatórios corretos? ii) quatro relatórios corretos? iii) dois relatórios corretos? iv) um relatório correto? v) nenhum relatório correto?
24. Quatro provadores (sejam 1, 2, 3 e 4) deverão ordenar três diferentes marcas de vinho (sejam A, B e C) quanto à preferência, sem conhecimento das marcas que lhes são dadas para degustar. Cada provador classifica como 1 a marca que mais lhe agrada, como 2 a segunda e como 3 a que menos lhe agrada, e então são somados os pontos atribuídos para cada marca de vinho. Suponha que os provadores realmente não podem discriminar entre as marcas, de modo que cada um está atribuindo sua ordenação ao acaso. Determine a probabilidade de que: a) a marca A receba o total de pontos igual a 4; b) alguma marca receba o total de pontos igual a 4; c) alguma marca receba uma soma de pontos de 5 ou menos.
25. Em um experimento, foram semeados com trigo quatro conjuntos (blocos) de cinco talhões (parcelas) - um talhão com cada uma das cinco cultivares A, B, C, D e E. Em todos os blocos a cultivar C foi a mais produtiva. Qual é a probabilidade de que esse resultado tenha decorrido exclusivamente por acaso?
26. Em um experimento semelhante ao do exercício anterior, a cultivar C teve a produção mais elevada em três dos quatro blocos e a segunda produção mais elevada no bloco restante. Qual é a probabilidade de ela se ter destacado tanto ou mais do que neste experimento exclusivamente por acaso?
27. Em um experimento de vinte parcelas, instalam-se quatro repetições de cinco cultivares (A, B, C, D e E) atribuídas às parcelas completamente ao acaso. Na ordenação dos resultados nas vinte parcelas, as parcelas com o tratamento C ficaram

nos quatro primeiros lugares. Qual é a probabilidade de que isto tenha decorrido por mero acaso?

28. Um distribuidor de semente de melão determinou através de pesquisa que 4 porcento de seu estoque de sementes não germinará. Ele vende as sementes em saquinhos de 50 unidades e garante que ao menos 90 porcento de germinação. Qual é a probabilidade de que um saquinho de sementes particular viole a garantia?

29. Em uma pesquisa de laboratório em um labirinto em formato de T, um animal tem a escolha de entre ir para a esquerda e obter alimento ou ir para a direita e receber um choque elétrico leve. Suponha que na primeira tentativa (antes de qualquer condicionamento) os animais têm igual probabilidade de dirigirem-se para a direita ou para a esquerda. Após receber alimento em uma tentativa particular, as probabilidades de escolha dos caminhos para a direita e para a esquerda são, respectivamente, 0,4 e 0,6. Entretanto, após o animal receber um choque em uma tentativa particular, as probabilidades de ir para a direita e para a esquerda tornam-se 0,2 e 0,8, respectivamente.

    a) Qual é a probabilidade de que o animal dirija-se para esquerda na segunda tentativa?

    b) Qual é a probabilidade de que o animal dirija-se para a esquerda na terceira tentativa?

30. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando com as letras V ou F entre parênteses, respectivamente. Se a sentença for falsa, explique porque.

    ( ) O conceito de probabilidade é útil para o estudo de fenômenos que podem ser previstos exatamente.

    ( ) Um experimento aleatório é um processo de observação e coleta de dados relevantes referentes a um fenômeno aleatório.

    ( ) Um processo de observação e coleta de dados relevantes referentes a um fenômeno aleatório é um experimento aleatório.

    ( ) Um experimento aleatório pode ser repetido sob condições essencialmente semelhantes.

    ( ) A observação e o registro da hora do nascer do sol em um próximo determinado dia é um exemplo de experimento aleatório.

    ( ) A observação e o registro do número de frutos de um determinado pessegueiro na próxima safra é um exemplo de experimento aleatório.

    ( ) O espaço básico de um experimento aleatório é o conjunto dos resultados possíveis desse experimento.

    ( ) O espaço básico de um experimento aleatório é único.

    ( ) Um evento unitário de um experimento aleatório é um resultado possível desse experimento.

    ( ) O espaço básico de um experimento aleatório é discreto se e somente se o conjunto dos resultados elementares desse experimento é finito.

    ( ) Um espaço básico contínuo não é enumerável.

    ( ) Um evento de um experimento aleatório é um subconjunto de seu espaço básico.

( ) O conceito clássico de probabilidade é aplicável em amostragem aleatória.

( ) Um espaço básico equiprovável é finito.

( ) Como um dado tem 6 faces, então a probabilidade de qualquer de duas faces é 1/6.

( ) O conceito clássico de probabilidade é aplicável a experimentos cujas probabilidades dos resultados elementares são desconhecidas.

( ) A regra da multiplicação é útil para a contagem do número de composições de resultados parciais de experimentos aleatórios que compreendem duas ou mais etapas.

( ) A regra das permutações é uma aplicação da regra da multiplicação para a contagem do número de diferentes modos de extração de n elementos de um conjunto com reposição de cada elemento antes da extração do seguinte.

( ) A regra dos arranjos permite o cálculo das distintas ordenações de a elementos de um conjuntos de n elementos.

( ) O número de permutações de n elementos dispostos em linha é a n-ésima parte do número de permutações de n elementos dispostos em círculo.

( ) O número de combinações de r elementos tomados de um conjunto de n elementos é menor do que o número de correspondentes arranjos, porque combinações de mesmos elementos não se distinguem.

( ) abc e cba são duas combinações distintas de 3 das letras a, b, c, d, e, mas constituem um mesmo arranjo.

( ) A freqüência relativa de um evento A que ocorre em n_A das n realizações de um experimento aleatório é a razão n_A/n.

( ) O conceito empírico de probabilidade estabelece que a probabilidade de um evento A é o limite da freqüência relativa f_A desse evento quando o número de realizações do experimento tende a infinito.

( ) A probabilidade de um evento A de um espaço básico equiprovável é a soma das probabilidades dos resultados elementares que implicam na ocorrência de A.

( ) Se dois eventos A e B não têm resultado elementar em comum, então o evento "A e B" é um evento impossível.

( ) A probabilidade da ocorrência do evento "A ou B" é a soma das probabilidades desses eventos.

( ) A probabilidade de um evento elementar de um espaço básico finito pode ser nula.

( ) Se A e B são dois eventos tais que todo resultado possível de A é, também, um resultado possível de B, então $P(A) \leq P(B)$.

( ) A soma das probabilidades de dois eventos que não podem ocorrer simultaneamente pode ser maior do que 1.

( ) O número de eventos de um espaço básico de n resultados elementares é 2^n.

4 VARIÁVEL ALEATÓRIA E DISTRIBUIÇÃO DE PROBABILIDADE DISCRETA

Conteúdo

4.1 Introdução....66
4.2 Variável Aleatória Discreta....69
4.2.1 Distribuição de probabilidade....69
4.2.2 Representação gráfica de uma distribuição de probabilidade....71
4.2.3 Média de uma distribuição de probabilidade....73
Propriedades do valor esperado....75
4.2.4 Variância de uma distribuição de probabilidade....75
Propriedades da variância....76
4.3 Distribuições Discretas Importantes....77
4.3.1 Distribuição uniforme discreta....77
4.3.2 Distribuição de Bernoulli....78
- Média e variância:....79
4.3.3 Distribuição binomial....79
- Média e variância:....81
4.3.4 Amostragem de uma população dicotômica....82
Amostragem com reposição....82
Amostragem sem reposição....83
4.3.5 Distribuição hipergeométrica....84
- Média e variância....84
4.3.6 Distribuição geométrica....85
- Média e variância....86
4.3.7 Distribuição binomial negativa....86
- Média e variância....87
4.3.8 Distribuição de Poisson....87
4.4 Distribuição Conjunta de Duas Variáveis Aleatórias....88
Propriedades da função distribuição de probabilidade conjunta....92
Representação geométrica....92
4.5 Distribuição de uma Função de Duas Variáveis Aleatórias....93
4.6 Distribuição Marginal....93

4.7 Valor Esperado de uma Função de Duas Variáveis Aleatórias....94
4.8 Covariância e Correlação de Duas Variáveis Aleatórias....95
Propriedades do coeficiente de correlação....97
4.9 Distribuição Condicional e Independência Estatística....97
4.10 Distribuição Conjunta de n Variáveis Aleatórias....99
Propriedades do valor esperado....101
4.11 Distribuição Multinomial....102
4.12 Exercícios....103

4.1 Introdução

Os resultados de um experimento aleatório diferem quanto a um grande número de características das unidades da população em consideração. Usualmente, entretanto, há interesse em apenas uma ou poucas dessas características. Considerar-se-á, inicialmente, a situação de apenas uma característica.

O modelo de probabilidade para uma característica é especificado pela coleção de todos os resultados elementares registrados para a característica, ou seja, por seu espaço básico, e pelas probabilidades associadas com esses resultados, isto é, os eventos simples. Esses eventos simples não são necessariamente numéricos. Por exemplo, o espaço básico associado com o experimento de germinação de uma semente é S={G, g}, onde G e g descrevem os dois eventos simples que correspondem às duas alternativas da característica considerada, respectivamente, germina e não germina. Embora eventos simples possam descrever características qualitativas, geralmente se tem interesse em expressá-los através de uma variável numérica, já que o tratamento analítico de um modelo de probabilidade somente pode ser estabelecido através de sua expressão algébrica. No exemplo, é usual assinalar o número 1 ao evento simples G e o número 0 ao evento simples g, já que o evento "germina" é o de maior interesse. Por outro lado, muitas vezes não há interesse em detalhes associados com os eventos simples, mas apenas em alguma descrição numérica de algum aspecto quantitativo referente aos resultados do experimento. Por exemplo, o espaço básico para o experimento de germinação de três sementes é S=(GGG, GGg, GgG, gGG, Ggg, gGg, ggG, ggg}. Entretanto, pode-se ter interesse apenas no número de sementes que germinam. Neste caso, os valores numéricos 0, 1, 2 e 3 são assinalados aos pontos do espaço básico.

Dessa forma, é usualmente importante a representação de eventos no espaço básico S através de conjuntos de números reais. Essa representação é obtida através de um "mapeamento" dos pontos do espaço básico S sobre o conjunto R dos números reais, isto é, sobre os pontos de uma reta.

Os números assinalados aos eventos simples do espaço básico são valores particulares de uma variável numérica, denominada **variável aleatória**. A designação "**aleatória**" provém do fato de que as realizações ou valores particulares da variável estão associados com probabilidades que correspondem às chances de suas ocorrências. Um conceito formal de variável aleatória é dado a seguir.

Variável aleatória é uma função X( ) definida no espaço básico S que faz corresponder a cada ponto $s \in S$ um e um só número real x (isto é, um e um só ponto sobre a linha reta **R**), de modo que a imagem inversa sobre S de cada evento A de R_x (conjunto de pontos sobre **R** gerados pelo mapeamento) seja um evento de S.

O conceito de variável aleatória é ilustrado na **Figura 4.1**.

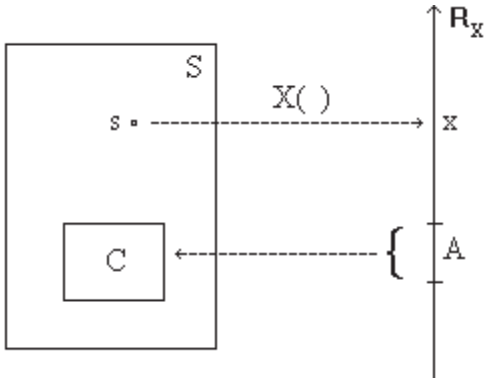


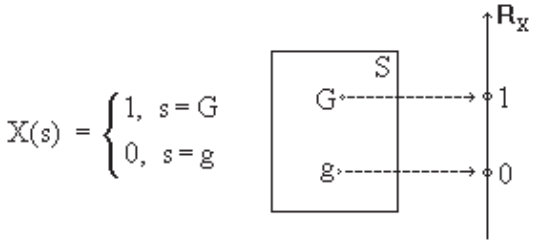
Figura 4.1. Ilustração do mapeamento efetuado por uma variável aleatória X, de pontos no espaço básico S em pontos sobre a linha reta **R**.

Dessa forma, probabilidades de eventos no espaço básico podem ser determinadas a partir de probabilidades de eventos sobre a linha reta, e vice-versa:

$$P_x[A] = P_x[C] .$$

Uma variável aleatória é usualmente denotada por X ou uma outra das últimas letras do alfabeto, em maiúsculo (Y, Z, T, U e V). Embora a notação X(.) seja mais completa, por enfatizar que a variável aleatória é uma função, a notação mais abreviada X é adotada mais freqüentemente. Um valor específico correspondente a uma realização de uma variável aleatória é denotado, genericamente, pela correspondente letra em minúsculo. Essa notação é acrescida de um índice quando diversos valores distintos são considerados. Assim, x é um valor genérico da variável aleatória X; x_1, x_2 e x_3 são três valores particulares de X.

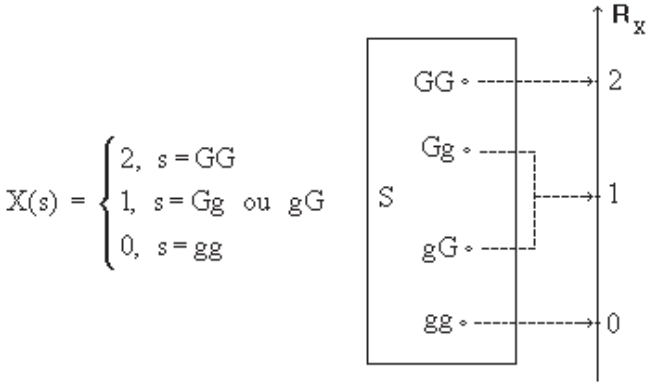
Exemplo 4.1. Seja o experimento "colocação de uma semente a germinar e observação do resultado: G (germina) e g (não germina)". Como há interesse na germinação, é comum definir a variável aleatória X(.) que atribui o número 1 ao evento {G} e o número 0 ao evento {g}, ou seja:



Exemplo 4.2. Considere-se o experimento constituído pela seqüência de dois experimentos especificados no Exemplo 4.1, ou seja, "colocação de duas sementes a germinar e observação do resultado para cada semente: G ou g". O espaço básico para este experimento é:

S = {GG, Gg, gG, gg}.

Uma variável aleatória de interesse é X: número de sementes que germinam. Esta variável aleatória atribui o número 0 ao evento simples {gg}, o número 1 a cada um dos eventos simples {Gg} e {gG}, e o número 2 ao evento simples {GG}:



O estabelecimento de um **modelo de probabilidade** para um experimento aleatório consiste em determinar o espaço $\mathbf{R}_x$ sobre a linha reta, resultante do mapeamento dos eventos simples do espaço básico através de uma variável aleatória X de interesse, e uma função f(x) que atribua probabilidades a eventos sobre a linha reta que representam os eventos sobre o espaço básico. Dessa forma, todas as considerações de interesse referentes a eventos no espaço básico podem ser feitas através de eventos sobre a linha reta. A ulterior interpretação dos resultados é reportada, novamente, a eventos do espaço básico.

O tratamento analítico que se inicia com o estabelecimento do modelo de probabilidade é caracteristicamente distinto para as situações em que o mapeamento define sobre a linha reta um número contável ou um número não contável de pontos. Por essa razão, essas duas situações devem ser consideradas separadamente.

Neste Capítulo, tratar-se-á de modelos de probabilidade discretos. A situação de variável aleatória contínua será abordada no **Capítulo 5**.

4.2 Variável Aleatória Discreta

4.2.1 Distribuição de probabilidade

> Uma variável aleatória X é uma **variável aleatória discreta** se ela assume um número finito ou infinito contável de valores distintos com probabilidades diferentes de zero cuja soma das probabilidades é igual a 1.

Os dois últimos exemplos são ilustrações de variável aleatória discreta.

Neste caso, um modelo de probabilidade para um experimento é uma lista dos valores da variável aleatória com probabilidades positivas, sejam x_1, x_2 ,..., e das probabilidades que lhes são associadas, ou seja:

$$P[X=x_i] = P_S[\{s \mid X(s)=x_i\}], \; i=1,2,...$$

> Uma função f(x) ou f(x) que atribui probabilidades aos pontos do espaço R_x de uma variável aleatória discreta X, tal que $f_x(x) = P[X=x]$ para $x \in \mathbf{R}_x$, é denominada **função distribuição de probabilidade**, ou **função de probabilidade**, ou **distribuição de probabilidade**.

Os valores da variável aleatória X são denominados **pontos de massa**, e a função f(x) denota a massa associada com o ponto de massa x. Por essa razão, a função f(x) também é denominada **função massa de probabilidade**. A notação p(x) é algumas vezes usada em lugar de f(x) para denotar função de distribuição de probabilidade discreta.

Uma função de distribuição de probabilidade discreta f(x) efetua um mapeamento dos pontos do espaço $\mathbf{R}_x$ da variável aleatória X sobre a reta **R** em pontos do intervalo [0,1]. Assim, seu domínio situa-se sobre a linha reta **R** e seu contradomínio, no intervalo [0,1].

Quando a variável aleatória assume um pequeno número de valores distintos com probabilidades positivas, a distribuição de probabilidade pode ser apresentada na forma da **Tabela 4.1**.

Tabela 4.1. Tabela de uma distribuição de probabilidade.

Valor de X	x_1	x_2	...	x_k
Probabilidade	$f(x_1)$	$f(x_2)$	...	$f(x_k)$

Freqüentemente, entretanto, é mais conveniente representar uma distribuição de probabilidade por uma fórmula para a função de probabilidade f(x) e a especificação dos valores possíveis de X. Essa representação é empregada mais extensivamente, como se verá adiante.

A definição da função distribuição de probabilidade f(x) pode ser estendida para todo o conjunto dos números reais, com a condição de que $f(x) = 0$ para todo $x \in \mathbf{R}_x$.

Uma função f(x) é uma **função distribuição de probabilidade** de uma variável aleatória discreta X se e somente se ela satisfaz às seguintes propriedades, derivadas dos axiomas do conceito de probabilidade (**Seção 3.9**):

1) $f(x) > 0$, para qualquer $x \in \mathbf{R}$;

2) $\sum_x f(x) = 1$, ou seja, a soma das probabilidades para todos os valores da variável aleatória é igual a 1.

Qualquer função que satisfaça a essas duas condições é uma função distribuição de probabilidade discreta. Entretanto, uma tal função somente terá sentido prático se for associada à um experimento aleatório de interesse.

Exemplo 4.3. Suponha-se que no experimento do **Exemplo 4.1** a probabilidade de uma semente germinar é 0,9. Então, as probabilidades da variável aleatória X lá considerada são:

$P[X=1] = P_s[s=G] = 0,9;$

$P[X=0] = P_s[s=g] = 0,1.$

Logo, a distribuição de probabilidade de X é:

$$f(x) = \begin{cases} 0,9, & x = 1 \\ 0,1, & x = 0 \end{cases}$$

Exemplo 4.4. Considere-se o experimento do **Exemplo 4.2**. Se a probabilidade de uma semente germinar é 0,9, como a germinação de uma semente é independente da germinação da outra semente, as probabilidades associadas com os valores da variável aleatória X: número de sementes que germinam, são determinadas como segue, onde os índices são utilizados para a identificação das sementes:

$P[X=2] = P_S[\{G_1\ G_2\}] = P_S[G_1] \times P_S[G_2] = 0,9 \times 0,9 = 0,81;$

$P[X=1] = P_s[\{G_1\ g_2\} \cup \{g_1\ G_2\}] = P_s[G_1\ g_2] + P_s[g_1\ G_2]$

$= P_s[G_1] \times P_s[g_2] + P_s[g_1] \times P_s[G_2]$

$= 0,9 \times 0,1 + 0,1 \times 0,9$

$= 2 \times 0,9 \times 0,1 = 0,18;$

$P[X=0] = P_s[\{g_1\ g_2\}] = P_s[g_1] \times P_s[g_2] = 0,1 \times 0,1 = 0,01.$

Logo, a função de probabilidade da variável aleatória X é:

$$f(x) = \begin{cases} 0,81, & x = 2 \\ 0,18, & x = 1 \\ 0,01, & x = 0 \end{cases}$$

A distribuição de probabilidade é a contraparte teórica da distribuição empírica de freqüências de que se tratou no **Capítulo 2**. A distribuição de probabilidade é um modelo de probabilidade teórico que assinala probabilidades a valores da variável aleatória. O modelo é derivado de alguma hipótese plausível referente ao mecanismo de chance concernente ao

fenômeno em estudo. Por outro lado, uma distribuição de freqüências é construída após a obtenção de um conjunto de dados através de um certo número de repetições de um experimento. Essas repetições constituem uma amostra aleatória de todos os possíveis conjuntos de repetições do experimento. Diferentes conjuntos proverão diferentes distribuições de freqüências, enquanto que uma distribuição de probabilidade é determinada pelas hipóteses postuladas. Suponha-se, por exemplo, que se postula a hipótese de que uma moeda é perfeita e lançada sem tendenciosidade. A distribuição de probabilidade da variável aleatória X: número de caras, é apresentada na **Tabela 4.2**. Suponha-se que, após o lançamento da moeda 100 vezes, foi obtida a distribuição de freqüências apresentada na **Tabela 4.3**. Para um número grande de lançamentos, as freqüências relativas aproximam-se das probabilidades teóricas (conceito empírico de probabilidade). As freqüências relativas obtidas podem constituir uma indicação para suspeita da adequabilidade do modelo que gerou a distribuição de probabilidade teórica.

Tabela 4.2. Distribuição de probabilidade do evento "cara" no experimento de lançamento de uma moeda perfeita sem tendenciosidade.

x_i	0	1
$f(x_i)$	0,5	0,5

Tabela 4.3. Distribuição de freqüências relativas do evento "cara" em 100 lançamentos de uma moeda.

x_i	0	1
n_i	41	59
f_i	0,41	0,59

4.2.2 Representação gráfica de uma distribuição de probabilidade

A distribuição de probabilidade pode ser representada, graficamente, por um **diagrama de linhas**. Os valores distintos da variável aleatória X com probabilidades positivas são marcados sobre um eixo horizontal. De cada um desses pontos x, traça-se uma linha vertical com altura igual à correspondente probabilidade f(x). A representação da distribuição de probabilidade do experimento de germinação de duas sementes, considerada no **Exemplo 4.4**, é apresentada na **Figura 4.2**.

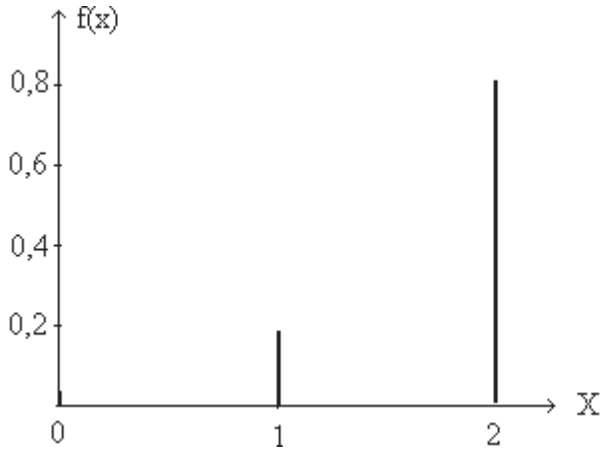


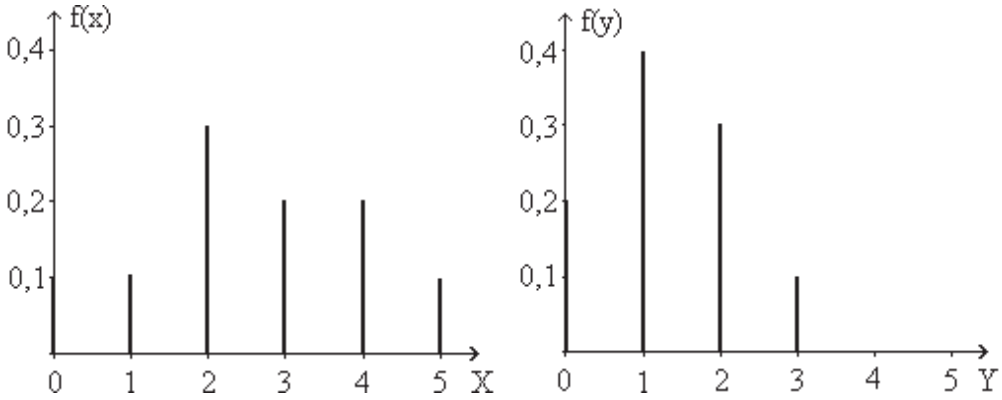
Figura 4.2. Representação gráfica da distribuição de probabilidades do número de sementes que germinam em um experimento de germinação de duas sementes em que a probabilidade de uma semente germinar é 0,9.

A representação gráfica de uma distribuição de probabilidade revela como a probabilidade total 1 é distribuída sobre os distintos valores da variável aleatória. Essa informação é muito importante, particularmente para propósito de comparação de duas ou mais distribuições de probabilidades quanto à extensão de suas similaridades e diferenças.

Exemplo 4.5. Em uma pesquisa de mercado, os registros da venda diária de compotas de dois fabricantes por um longo período são utilizados para avaliar as distribuições de probabilidades do número de latas de compotas vendidas das duas procedências. As distribuições de probabilidades das variáveis aleatórias X: número de latas de compotas do fabricante A, e Y: número de latas de compotas do fabricante B, são apresentadas na **Tabela 4.4**. Suas representações geométricas estão na **Figura 4.3**.

Tabela 4.4. Distribuições de probabilidades das quantidades de latas de compotas de dois fabricantes, vendidas diariamente nos supermercados de uma localidade.

Fabricante	Distribuição de probabilidades						
A	x	0	1	2	3	4	5
	f(x)	0,1	0,1	0,3	0,2	0,2	0,1
B	y	0	1	2	3		
	f(y)	0,2	0,4	0,3	0,1		



Fabricante A Fabricante B

Figura 4.3. Distribuições de probabilidades das quantidades de latas de compota de dois fabricantes vendidas em uma localidade.

Uma comparação de duas distribuições de probabilidades ou de suas representações gráficas é elucidativa, mas subjetiva, pois diferentes pessoas podem extrair diferentes conclusões sobre a extensão das similaridades. Assim, é desejável atribuir medidas numéricas concretas a aspectos específicos das distribuições, de modo que comparações possam ser efetuadas em avaliações objetivas. Medidas numéricas que caracterizam o centro e a dispersão de uma distribuição de probabilidade são importantes para tais propósitos. A medida de centro mais comumente usada é a **média**, ou **valor esperado**, e a de dispersão é a **variância**.

Em Estatística Descritiva, foi visto como determinar a média e a variância de uma distribuição de freqüências de um conjunto de dados. As distribuições de probabilidades são modelos teóricos em que as probabilidades são as freqüências relativas para a população. Assim, semelhantes medidas de centro e de dispersão podem ser definidas para elas como extensões das correspondentes medidas para amostras.

4.2.3 Média de uma distribuição de probabilidade

Recorde-se que a média de uma distribuição de freqüências é dada por:

$$\overline{x}_f = \sum_i c_i f_i ,$$

onde c_i é o centro da i-ésima classe, ou o i-ésimo nível da variável (para variável com um número pequeno de distintos níveis), e f_i é a correspondente freqüência relativa. A média da variável aleatória X ou de sua distribuição de probabilidade, ou média populacional, é definida como uma extensão desse conceito:

A **média** da variável aleatória X, ou da distribuição de probabilidade de X, também denominada **valor esperado** de X, denotada por $E(X)$ ou μ, é definida como:

$$E(X) = \sum_i x_i f(x_i),$$

ou seja, a soma dos produtos dos valores da variável aleatória X com probabilidades positivas pelas correspondentes probabilidades.

Observe-se que o somando $x_i f(x_i)$ é o i-ésimo valor possível da variável aleatória X multiplicado pela sua correspondente probabilidade, e $E(X)$ é a soma desses produtos. Assim, $E(X)$ é uma "média" dos valores que a variável aleatória X assume com cada um desses valores ponderado pela sua correspondente probabilidade. Pesos maiores são atribuídos a valores mais prováveis de X.

Exemplo 4.6. Considere-se a distribuição de probabilidade para o experimento germinação de duas sementes do **Exemplo 4.4**. A determinação do valor esperado da variável aleatória X: número de sementes que germinam, pode ser feita com o auxílio da **Tabela 4.5**.

Tabela 4.5. Determinação do $E(X)$ para a variável aleatória X do **Exemplo 4.4**.

x	0	1	2	Soma	
f(x)	0,01	0,18	0,81	1,00	
x f(x)	0	0,18	0,81	1,80	$= E(X)$

Ou seja,

$$E(X) = 0\times0{,}01+1\times0{,}18+2\times0{,}81 = 1{,}80.$$

O valor esperado não é, necessariamente, um valor possível da variável aleatória, como ilustra o **Exemplo 4.6**. Nesse caso, não se pode "esperar" obter o valor esperado. Isso é uma contradição à designação de "valor esperado". Uma designação mais apropriada seria "valor médio". Entretanto, neste texto adota-se a designação usual de valor esperado.

A média $\mu=E(X)$ é o centro de gravidade (centróide) da distribuição da massa unitária que é determinada por f(x). De fato, se a massa total 1 é distribuída sobre os segmentos do diagrama de linhas correspondente a uma distribuição de probabilidade, com a massa proporcional aos segmentos, então $E(X)$ situa-se sobre o ponto de equilíbrio da figura. Assim, $E(X)$ indica o "centro" dos pontos em que se situam os valores possíveis da variável aleatória X. Essa propriedade do valor esperado é uma justificativa para seu uso como medida de centro de uma variável aleatória, ou de sua distribuição de probabilidade. Outras propriedades do valor esperado como medida de posição de uma distribuição de probabilidade são análogas às propriedades da média aritmética como medida de posição de um conjunto de dados de uma amostra, ou de sua distribuição de freqüências.

O conceito de valor esperado pode ser estendido para uma função de uma variável aleatória: Se g(x) é uma função de uma variável aleatória X, então:

$$E[g(X)] = \sum_x g(x) f(x),$$

onde a soma é feita sobre todos os valores de X com probabilidades positivas, com g(x) e f(x) avaliadas nesses valores.

Propriedades do valor esperado

O símbolo *E*, que significa a operação de determinação do valor esperado, é um operador com as seguintes propriedades, onde a e b são duas constantes:

1) $E(a) = a$;

2) $E(bX) = b\,E(X)$;

3) $E(a+bX) = a+bE(X)$.

Essas propriedades são imediatas da definição de valor esperado de uma função g(X) de X. Para derivar a propriedade 3, por exemplo, toma-se g(X)=a+bX na expressão de *E*[g(X)]; obtém-se:

$$\begin{aligned} E(a+bX) &= \sum_x (a + bx) f(x) \\ &= a \sum_x f(x) + b \sum_x x f(x) \\ &= a + bE(X), \end{aligned}$$

visto que $\sum_x f(x) = 1$.

4.2.4 Variância de uma distribuição de probabilidade

A variância de uma distribuição de freqüências é:

$$s^2 = \sum_i (c_i - \overline{x}_f)^2 f_i .$$

A variância da variável aleatória X ou de sua distribuição de probabilidade, ou variância populacional, é definida como uma extensão desse conceito:

> A **variância** da variável aleatória X ou da distribuição de probabilidade de X, designada por Var(X) ou σ_x^2 ou σ^2, é definida como:
>
> $$\text{Var}(X) = \sum_i \left[x_i - E(X)\right]^2 f(x_i),$$
>
> com a soma sobre todos os valores de X com probabilidades positivas.

Logo, a expressão da variância de X é:

$$\begin{aligned} \text{Var}(X) &= E[X-E(X)]^2 = \\ &= E(X^2)-[E(X)]^2. \end{aligned}$$

Observe-se que o somando na expressão da Var(X) é o quadrado da diferença entre o i-ésimo valor possível da variável aleatória X e a média de X, multiplicado pela probabilidade de que X assuma seu i-ésimo valor. $E(X)$ é a soma desses produtos. Logo, Var(X) é a média ponderada desses quadrados de diferenças que atribui pesos mais elevados às diferenças elevadas ao quadrado mais prováveis. Assim, a variância de uma variável aleatória cujos valores tendem a ser afastados de sua média é maior do que a de uma variável aleatória com os mesmos valores mas que tendem a ser mais próximos de sua média. Essa propriedade justifica a variância como uma medida de variabilidade ou dispersão de uma variável aleatória, ou de sua distribuição.

Foi observado anteriormente que a média $E(X)$ é o centro de gravidade da massa unitária determinada pela função distribuição de probabilidade f(x). A variância também tem um significado físico. Ela representa o momento de inércia da mesma distribuição com respeito ao eixo perpendicular ao eixo x pelo centro de gravidade $E(X)$.

A unidade de medida da variância é o quadrado da unidade de medida da variável aleatória a que corresponde, o que pode ser inconveniente para a interpretação da variância como medida de dispersão em aplicações. Uma medida de dispersão alternativa, mais conveniente sob esse aspecto, é o **desvio padrão**, definido como a raiz quadrada da variância, ou seja:

$$\sigma_x = \sqrt{\text{Var}(X)}\,.$$

Exemplo 4.7. A variância da variável aleatória considerada no **Exemplo 4.4** pode ser determinada com o auxílio da **Tabela 4.6**.

Tabela 4.6. Determinação da Var(X) para a variável aleatória X do **Exemplo 4.4**.

x	0	1	2	soma	
f(x)	0,01	0,18	0,81	1,00	
xf(x)	0	0,18	1,62	1,80	= E(X)
$x^2f(x)$	0	0,18	3,24	3,42	= $E(X^2)$

Logo,

$$\text{Var}(X) = 3{,}42 - 1{,}80^2$$
$$= 0{,}18.$$

Donde:

$$\sigma_x = \sqrt{0{,}18} = 0{,}4243.$$

Propriedades da variância

As seguintes propriedades são derivadas da definição de variância: Se X é uma variável aleatória e a e b são duas constantes, então:

1) Var(X) > 0, ou seja, a variância é não negativa.

2) Var(a) = 0, ou seja, a variância de uma constante é nula.

3) $Var(bX) = b^2Var(X)$.

4) $Var(a+bX) = b^2Var(X)$.

4.3 Distribuições Discretas Importantes

Alguns modelos de probabilidade são comuns a experimentos de diversas áreas. Assim, por exemplo, o mesmo modelo estabelecido para o experimento "colocação de uma semente a germinar e observação do resultado: G ou g" aplica-se a qualquer outro experimento de duas alternativas (isto é, experimento com a mesma estrutura), tais como "observação do sexo de um animal", "observação da cor dos olhos de um indivíduo (claros e escuros)" e "observação do caráter aspas em um bovino da raça Angus". Dessa forma, é importante o estabelecimento dos modelos de probabilidade mais usuais para famílias de experimentos de estrutura comum.

Nos exemplos recém formulados, a probabilidade do evento "sucesso" é, em geral, desconhecida. Assim, o modelo de probabilidade postulado para um problema particular deve incluir a probabilidade desconhecida para a população como um valor desconhecido, ou seja, um parâmetro, e a probabilidade para qualquer valor individual da variável aleatória é especificada em termos deste parâmetro. De fato, nas aplicações, um modelo de probabilidade especifica uma **família paramétrica de distribuições de probabilidade**, indexada por um ou mais parâmetros.

Os modelos de probabilidade de variável aleatória discreta mais importantes são apresentados e estudados a seguir.

4.3.1 Distribuição uniforme discreta

Uma variável aleatória discreta X tem uma **distribuição uniforme** se ela atribui a mesma probabilidade para todos os valores $x \in \mathbf{R}$. Seja X uma variável aleatória que atribui probabilidades positivas iguais aos N primeiros números inteiros positivos. Então, X tem **distribuição uniforme discreta**, expressa pela equação:

$$f(x) = \frac{1}{N}, \quad x = 1,2,...,N,$$

onde o parâmetro N é definido para os inteiros positivos, ou seja: N=1,2,...

Essa função é representada, geometricamente, na **Figura 4.4**.

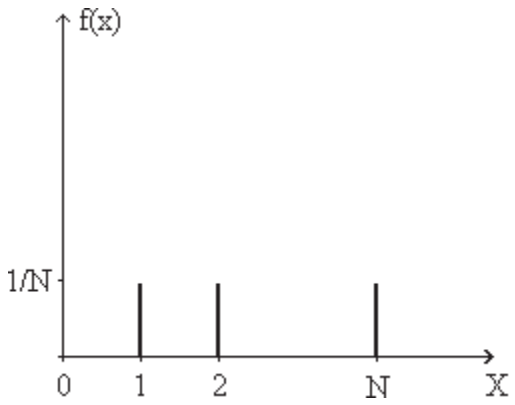


Figura 4.4. Distribuição de probabilidade de uma variável aleatória uniforme discreta.

Esse é o modelo de probabilidade para a extração aleatória de um elemento de uma coleção. - Média e variância:

$$E(X) = 1.\frac{1}{N} + 2.\frac{1}{N} + \ldots + N\frac{1}{N} = \frac{1}{N}(1 + 2 + \ldots + N) = \frac{1}{N}.\frac{N(N+1)}{2} = \frac{N+1}{2}.$$

$$\text{Var}(X) = \frac{N^2 - 1}{12}.$$

4.3.2 Distribuição de Bernoulli

Diz-se que um experimento é um **experimento de Bernoulli** se seu espaço básico compreende dois eventos simples. Como usualmente há interesse no registro da ocorrência de um dos dois eventos simples, o evento de interesse é designado "sucesso" e o outro evento, "falha" ou "insucesso". A variável aleatória de Bernoulli assinala o número 1 para o evento simples sucesso e o número zero para o evento simples falha:

$$X(s)\begin{cases} 1, & s = s_1 \quad \text{(sucesso)} \\ 0, & s = s_2 \quad \text{(fracasso)} \end{cases}$$

O modelo de probabilidade de Bernoulli é definido pela atribuição da probabilidade p ao sucesso e de seu complemento, 1-p, à falha. A função de probabilidade de Bernoulli tem a seguinte expressão:

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

$$= p^x (1\text{-}p)^x, \quad x=0,1,2, \ldots ,$$

onde o parâmetro p é um número real no intervalo [0,1].

A representação geométrica da distribuição de Bernoulli é dada na **Figura 4.5**.

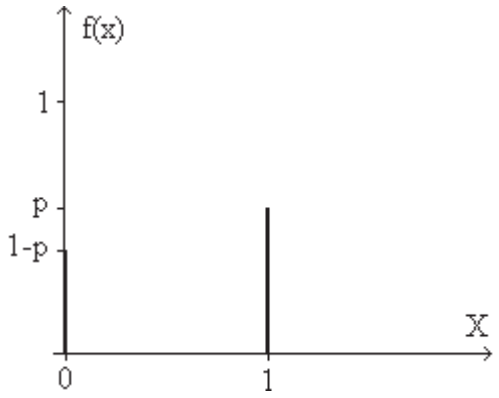


Figura 4.5. Distribuição de probabilidade de Bernoulli.

- Média e variância:

$$E(X) = 1\times p + 0\times(1-p) = p;$$
$$E(X^2) = 1^2\times p + 0^2 \times(1-p) = p.$$

Logo,

$$Var(X) = p-p^2 = p(1-p).$$

Exemplo 4.8. A distribuição do **Exemplo 4.3** ilustra a distribuição de Bernoulli:

$$f(x) = \begin{cases} 0,9, & x = 1 \\ 0,1, & x = 0 \end{cases}$$
$$= 0,9^x\times 0,1^{1-x}, \quad x= 0,1$$

Logo,

$$E(X) = 0,9.$$
$$Var(X) = 0,9(1-0,9) = 0,09.$$

4.3.3 Distribuição binomial

A realização sucessiva de um número fixo n de experimentos de Bernoulli independentes e com mesma probabilidade p de sucesso constitui um **experimento binomial**. A variável aleatória binomial X é a soma dos sucessos em n ensaios de Bernoulli com essas propriedades. Ou seja, uma **variável binomial** X é definida como: $X = X_1+X_2 +...+X_n$, onde X_i (i=1,2,...,n) são variáveis aleatórias com distribuição de Bernoulli, independentes e com mesma distribuição:

$$f(x_i) = \begin{cases} p, & x_i = 1 \\ 1-p, & x_i = 0 \end{cases}$$

Essa variável aleatória X: número de sucessos, tem **distribuição binomial**.

Para derivar a expressão da distribuição de probabilidade da variável aleatória binomial X, observe-se que a probabilidade de x sucessos e n-x falhas, em uma ordem específica dos n

ensaios de Bernoulli, é $p^x (1-p)^{n-x}$. Nesta expressão, há um fator p para cada um dos x sucessos e um fator 1-p para cada uma das n-x falhas; os x fatores p e os n-x fatores 1-p são multiplicados, pela pressuposição de independência. Como a probabilidade $p^x (1-p)^{n-x}$ aplica-se a qualquer seqüência dos n ensaios de Bernoulli com x sucessos e n-x falhas e essas seqüências são mutuamente exclusivas, essa probabilidade deve ser multiplicada pelo número dessas seqüências. O número de modos em que se pode selecionar os x ensaios com sucesso do total de n ensaios é C_n^x. Portanto, a probabilidade de x sucessos em n ensaios é $C_n^x\ p^x (1-p)^{n-x}$.

A variável aleatória X tem **distribuição binomial** se e apenas se sua distribuição de probabilidade é dada por:

$$f(x) = C_n^x\ p^x (1-p)^{n-x},\ \ x=0,1,2,...,n,$$

onde p e n são parâmetros, $0<p<1$ e n inteiro positivo, e C_n^x é o número de combinações de x elementos tomados de n:

$$C_n^x = \frac{n(n-1)(n-2)\ldots(n-x+1)}{x(x-1)(x-2)\ldots 2.1}.$$

Indica-se que uma variável aleatória X tem distribuição binomial correspondente à soma dos sucessos de n experimentos de Bernoulli independentes, com probabilidade de sucesso p, pela notação $X \sim b(n;p)$.

A designação "binomial" provém do fato de que a probabilidade para cada valor de x é igual ao correspondente termo da expansão binomial de $[p+(1-p)]^n = p^0 (1-p)^1 + p (1-p)^{n-1} + ... + p^i (1-p)^{n-i} + ... + p^{n-i} (1-p)^1 + p^n (1-p)^0$.

A distribuição binomial depende de dois parâmetros, p e n. Quando n=1, ela reduz-se à distribuição de Bernoulli. Para qualquer valor de n, ela é simétrica se e somente se p=0,5. A representação geométrica da distribuição binomial é ilustrada na **Figura 4.6**, para as variáveis aleatórias $X \sim b(4;0,5)$ e $Y \sim b(4;0,3)$, cujas distribuições de probabilidade são dadas na **Tabela 4.7**.

Tabela 4.7. Distribuições de probabilidade das variáveis aleatórias $X \sim b(4;0,5)$ e $Y \sim b(4;0,3)$.

Var. aleatória	0	1	2	3	4
$X \sim b(4;0,5)$	0,0625	0,25	0,375	0,25	0,0625
$Y \sim b(4;0,3)$	0,2401	0,4116	0,2646	0,0756	0,0081

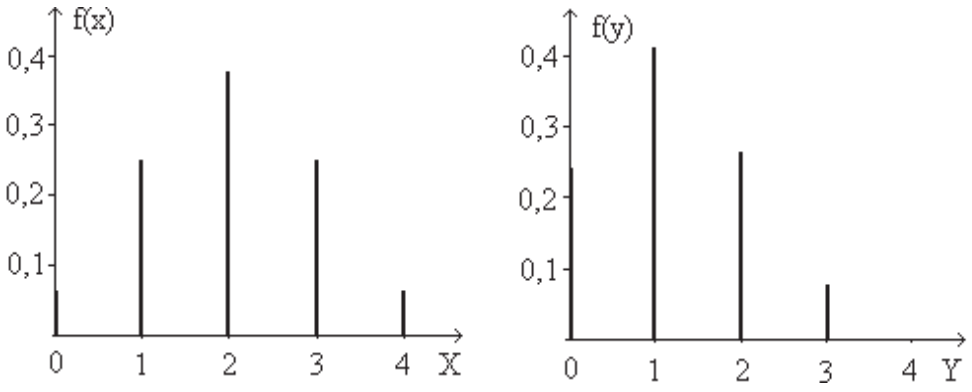


Figura 4.6. Distribuições de probabilidades das variáveis aleatórias X ~ b(4;0,5) e Y ~ b(4;0,3).

- Média e variância:

$E(X) = np.$

$Var(X) = np(1-p).$

Exemplo 4.9. Suponha-se que cinco sementes são selecionadas, aleatoriamente, de um saco de sementes cuja probabilidade de germinar é 0,9. Qual é a probabilidade de que 3 das 5 sementes germinem?

Esse experimento é binomial, já que para cada semente há duas alternativas: germina e não germina; p=0,9 é a probabilidade de uma semente germinar e n=5 é o número de sementes que são colocadas a germinar. Logo, a função de probabilidade de X: número de sementes que germinam, é:

$$f(x) = C_5^x .0,9^{\mathbf{x}} .0,1^{5-\mathbf{x}};$$

portanto, a probabilidade de que 3 das 5 sementes germinem é:

$$P[x = 3] = C_5^3 .0,9^3 .0,1^{3-2} =$$

$$= \frac{5.4.3}{3.2.1} .0,9^3 .0,1^2 = 0,0729.$$

A média e a variância do número de sementes que germinam são:

$E(X) = 5 \times 0,9 = 4,5;$

$Var(X) = 5 \times 0,9 \times 0,1 = 0,45.$

Probabilidades da distribuição binomial envolvem potências de p e 1-p cujo cálculo torna-se trabalhoso quando o parâmetro n é moderadamente elevado. Por essa razão, tabelas dessas probabilidades são disponíveis em vários textos. A **Tabela I** do **Apêndice** apresenta os valores das probabilidades acumuladas $P[X \leq c] = \sum_{x=0}^{c} C_c^n p^x (1-p)^{n-x}$ para alguns valores de n, p e c.

Probabilidades para outros eventos podem ser obtidas pela expressão desses eventos em termos de eventos da forma [X≤c] e aplicação de propriedades da probabilidade, como ilustram os seguintes exemplos:

$P[X>x] = 1-P[X\leq x];$

$P[X=x] = P[X\leq x] - P[X\leq x-1];$

$P[a\leq X\leq b] = P[X\leq b] - P[X\leq a-1].$

Exemplo 4.10. Suponha-se que a percentagem de germinação de sementes de uma cultivar de feijão disponíveis em um depósito é 80%. Se quinze sementes desse lote são plantadas, qual é a probabilidade de que dez ou mais sementes germinem?

Pressupondo que o resultado seja independente para cada semente, a variável aleatória X: número de sementes que germinam tem distribuição binomial com n=15 e p=0,8. Então a probabilidade de que dez ou mais sementes germinem é:

$$P[X\geq 10] = \sum_{x=10}^{15} C_{15}^{x} 0,8^{x} 0,2^{15-x}.$$

Entretanto, esta probabilidade pode ser calculada, facilmente, com o auxílio da **Tabela I**, como segue:

$P[X\geq 10] = 1-P[X\leq 9] = 1-0,061 = 0,939.$

4.3.4 Amostragem de uma população dicotômica

Uma população dicotômica é aquela em que a característica em consideração tem duas alternativas. Na amostragem de indivíduos de uma tal população, é freqüentemente de interesse a distribuição do número de indivíduos com uma das alternativas, denominada "sucesso".

Em um processo de amostragem aleatória, cada unidade da população tem a mesma probabilidade de ser selecionada. Na extração de uma amostra de duas ou mais unidades, um de dois procedimentos alternativos pode ser adotado: as unidades são selecionadas sucessivamente de modo que cada unidade extraída é recolocada na população antes da extração da unidade seguinte, ou as unidades extraídas não são recolocadas na população. Esses dois procedimentos de amostragem serão tratados a seguir através de um exemplo.

Exemplo 4.11. Suponha-se que um lote de um certo equipamento contém 12 unidades das quais 9 são perfeitas e 3 defeituosas. Qual é a probabilidade de que, selecionando-se, aleatoriamente, três unidades, duas sejam perfeitas?

Amostragem com reposição

A primeira unidade é selecionada, a qualidade do equipamento (perfeito ou defeituoso) é registrada e a unidade é recolocada no lote antes da próxima seleção. A reposição da unidade assegura a manutenção da probabilidade de seleção de um equipamento perfeito na segunda extração e a independência dos dois eventos: equipamento perfeito na primeira seleção e equipamento perfeito na segunda seleção. Trata-se, portanto, de uma sucessão de dois experimentos de Bernoulli independentes e com igual probabilidade de sucesso. Nestas condições,

a probabilidade do número de equipamentos perfeitos na seleção aleatória de três equipamentos tem distribuição binomial com probabilidade de sucesso p=9/12=0,75:

$$P[X=x] = C_3^x (0,75)^x (0,25)^{3-x}, \quad x = 0,1,2,3.$$

Logo,

$$P[X=2] = C_3^2 (0,75)^2 (0,25)^{3-2} = \frac{3\times2}{2\times1}\times 0,75^2 \times 0,25 = 0,4219$$

Amostragem sem reposição

A primeira unidade é selecionada e não é reposta no lote antes da seleção da segunda unidade. Semelhantemente, uma segunda unidade é selecionada e também não é recolocada no lote antes da seleção da terceira unidade, e assim sucessivamente. Com esse procedimento a condição de independência dos experimentos de Bernoulli é violada. De fato, a probabilidade de sucesso para a primeira unidade selecionada é: P[equipamento perfeito] = 9/12. Se a primeira unidade selecionada é perfeita, restam no lote para a segunda seleção 11 unidades das quais 8 são perfeitas. Então, a probabilidade de que a segunda unidade seja perfeita, dado que a primeira unidade selecionada é perfeita, é $\frac{8}{11} \neq \frac{9}{12}$, o que estabelece a falta de independência.

Desta forma, as condições do experimento binomial não se verificam na amostragem sem reposição. Portanto, a distribuição do número de sucessos para esse procedimento de amostragem não é binomial.

Observe-se, entretanto, que a probabilidade de sucesso na segunda seleção permanece inalterada, ou seja, a probabilidade (não condicionada) de sucesso é 9/12 em cada um dos experimentos de Bernoulli (isto é, em cada uma das seleções). Considere-se a probabilidade de selecionar um equipamento perfeito na segunda seleção. Denote-se por S e F os eventos sucesso e falha em cada uma das seleções e acrescente-se um índice para indicar a ordem da seleção. Observe-se que um sucesso na segunda seleção pode resultar quando ocorre sucesso ou insucesso na primeira seleção. Então, a probabilidade de sucesso na segunda seleção é:

$$P[S_2] = P[(S_1 \cap S_2) \cup (F_1 \cap S_1)] = P[S_1 \cap S_2] + P[F_1 \cap S_1],$$

dada a mútua exclusividade dos dois eventos. Mas

$$P[S_1 \cap S_2] = P[S_1] \times P[S_2|S_1] = \frac{9}{12} \times \frac{8}{11}$$

e, semelhantemente,

$$P[F_1 \cap S_1] = P[F_1] \times P[S_2|F_1] = \frac{3}{12} \times \frac{9}{11}.$$

Logo,

$$P[S_2] = \frac{9}{12} \times \frac{8}{11} + \frac{3}{12} \times \frac{9}{11} = 0,75.$$

4.3.5 Distribuição hipergeométrica

O processo de amostragem aleatória de indivíduos de uma população finita, sem reposição, o que corresponde a uma sucessão de n ensaios de Bernoulli em que não é satisfeita a condição de independência, é um **experimento hipergeométrico**. A variável aleatória X: número de sucessos em um total de n ensaios (ou número de indivíduos em uma amostra de tamanho n com uma específica das duas alternativas da característica em consideração), tem a **distribuição hipergeométrica**.

Seja uma população de N elementos dos quais M (M<N) manifestam uma alternativa A de uma certa característica, e, portanto, N-M não a manifestam. Suponha-se que se seleciona desta população uma amostra de n elementos, sem reposição. Seja X o número de elementos da amostra que manifestam a alternativa A. Para determinar a distribuição de X, observe-se que o número x de elementos da amostra que manifestam a alternativa A deve proceder do grupo dos M elementos da população com esta alternativa, e os n-x elementos que não manifestam a alternativa A devem vir do outro grupo dos N-M elementos. Então, a distribuição de X pode ser derivada pela regra das combinações, como segue.

O número total de amostras de n elementos que podem ser extraídos da população é C_N^x . O número de amostras de n elementos com x elementos com a alternativa A e n-x elementos com a outra alternativa da característica é $C_M^x\ C_{N-M}^{n-x}$. Logo, a distribuição do número de sucessos X é:

$$f(x)=\frac{C_M^x \times C_{N-M}^{n-x}}{C_N^n},\ x=0,\ 1,\ 2,...,\ n,$$

onde os parâmetros n, N e M satisfazem às seguintes condições: N é um inteiro positivo e M e n são inteiros não negativos tais que n≤M e n≤N-M.

- Média e variância

A média e a variância da distribuição hipergeométrica são, respectivamente:

$E(X) = np,\quad p = \frac{M}{N}$ (proporção de elementos na população com a alternativa A da característica)

e

$$Var(X) = np(1-p)\frac{N-n}{N-1}.$$

Observe-se que a média da distribuição hipergeométrica é a mesma da distribuição binomial, que corresponde à amostragem efetuada indivíduo por indivíduo, com reposição. Entretanto, a variância é a variância da distribuição binomial multiplicada pelo fator $\frac{N-n}{N-1}$, que é denominado "fator de correção para população finita". Quando n/N (proporção do tamanho da amostra em relação ao tamanho da população) é muito pequeno, o fator de correção é próximo de 1, de modo que a distribuição binomial aproxima a distribuição hipergeométrica.

Exemplo 4.12. Considere-se o experimento do **Exemplo 4.11** para a situação de amostragem sem reposição. Nestas circunstâncias, a distribuição de probabilidade da variável aleatória X: número de unidades perfeitas em 3 unidades selecionadas de um lote de 12 equipamentos dos quais 9 são perfeitos, é hipergeométrica:

$$f(x) = \frac{C_9^x \times C_3^{3-x}}{C_{12}^3}$$

$$= \frac{1}{220} C_9^x \times C_3^{3-x}, \quad x = 0,1,2,...$$

Então, a probabilidade de 2 unidades perfeitas entre as 3 selecionadas é:

$$P[X=2] = \frac{1}{220} C_9^2 \times C_3^{3-2} = \frac{1}{220} \times 36 \times 3 = 0,49.$$

4.3.6 Distribuição geométrica

O experimento geométrico é outro experimento derivado de uma seqüência de ensaios de Bernoulli, independentes e com a mesma probabilidade p de sucesso. Se o número n desses ensaios é fixo, o experimento é um experimento binomial e a variável aleatória X: número de sucessos, é uma variável aleatória com distribuição binomial b~(n,p). Se, em vez de fixar o número de ensaios antecipadamente, repete-se os ensaios de Bernoulli até a ocorrência do primeiro sucesso, o experimento é um **experimento geométrico**. Então, o número de sucessos é fixo em 1 e o número de ensaios é uma variável aleatória.

Seja X o número de ensaios efetuados até a ocorrência do primeiro sucesso. Os valores possíveis de X são 1, 2, ..., onde X=x se e somente se x-1 falhas ocorrem em seqüência, antes da ocorrência do primeiro sucesso. A seqüência de ensaios de Bernoulli encerra-se quando o primeiro sucesso ocorre. Assim, da condição de independência dos ensaios, segue-se:

$$f(x) = (1-p)^{x-1}p, \quad x=1,2,...,$$

onde o parâmetro p satisfaz $0<p\leq 1$.

Essa distribuição de probabilidade é denominada **distribuição geométrica**, também designada **distribuição discreta de espera de tempo**. Esta última denominação deriva-se do fato de que, se a realização de um ensaio de Bernoulli toma uma unidade de tempo, então a espera de tempo para a obtenção de um único sucesso é a variável aleatória com a distribuição geométrica.

A distribuição geométrica é útil no estudo de uma característica rara de uma população, como, por exemplo, a incidência de uma doença rara do sangue. Nesses casos, poder-se-ia examinar um número específico n de indivíduos e contar o número afetado, ou examinar indivíduos até que seja encontrada uma pessoa afetada. A primeira alternativa corresponde à distribuição binomial, mas a amostra de n indivíduos pode não conter uma única pessoa afetada, de modo que se obteria pouca informação sobre a incidência da doença. O segundo método de amostragem, correspondente à distribuição geométrica, garante a presença de um indivíduo com a incidência da doença na amostra.

- Média e variância

A média e a variância da distribuição geométrica são:

$E(X) = 1/p$,

$Var(X) = (1-p)/p^2$.

Exemplo 4.13. A incidência de uma doença que afeta 5 por cento dos animais de um rebanho de 50.000 bovinos deve ser avaliada pelo seguinte processo de amostragem dos animais: Animais do rebanho são selecionados aleatoriamente, um por um, e examinados até que o primeiro animal afetado pela doença seja encontrado. Determine a distribuição de probabilidade da variável aleatória X: número de ordem do primeiro animal que manifesta a doença. Quantos animais se espera sejam examinados utilizando este método de amostragem?

Se a população é consideravelmente grande, como no presente caso, pode-se supor que a percentagem de animais afetados permaneça praticamente constante nas seleções sucessivas de um animal. Logo, a distribuição da variável aleatória X é geométrica, ou seja:

$$P[X=x] = 0{,}95^{x-1} \times 0{,}05, \quad x=1,2,...$$

O número de animais que se espera sejam examinados até encontrar o primeiro que manifeste a doença é:

$$E(X) = 1/0{,}05 = 20.$$

4.3.7 Distribuição binomial negativa

Considere-se uma generalização do experimento geométrico: uma sucessão de ensaios de Bernoulli independentes e com mesma probabilidade de sucesso p, até a ocorrência do k-ésimo sucesso. Esse experimento é denominado um **experimento binomial negativo**, ou **experimento de Pascal**.

Para derivar a distribuição de probabilidade da variável aleatória X: número de ensaios para a produção de k sucessos, observe-se que a ocorrência do k-ésimo sucesso no x-ésimo ensaio significa que um sucesso no x-ésimo ensaio é precedido de k-1 sucessos e x-k falhas, em alguma ordem especificada. Como cada sucesso ocorre com probabilidade p e cada falha com probabilidade 1-p, a probabilidade do k-ésimo sucesso em uma ordem específica ocorrer no x-ésimo ensaio é:

$$p^{k-1}.(1-p)^{x-k} \times p = p^k \times (1-p)^{x-k}.$$

Mas o número de ordens em que k-1 sucessos ocorrem em x-1 ensaios é dado pela regra das combinações: C_{x-1}^{k-1}. Como esses eventos "o k-ésimo sucesso ocorre no x-ésimo ensaio" são mutuamente exclusivos para ordens de sucessos diferentes e têm a mesma probabilidade $p^k.(1-p)^{x-k}$, a probabilidade de ocorrer o k-ésimo sucesso no x-ésimo experimento é obtida pela multiplicação da probabilidade do k-ésimo sucesso em uma ordem específica ocorrer no x-ésimo ensaio pelo número de diferentes ordens dos k sucessos. Portanto, a distribuição de probabilidade da variável aleatória X: número do ensaio em que ocorre o k-ésimo sucesso, é:

$$f(x) = C_{x-1}^{k-1} p^k \times (1-p)^{x-k}, \quad x=k, k+1, k+2,...,$$

onde p e k são parâmetros tais que $0<p\leq 1$ e $k=0,1,2,...$

- Média e variância

A média e variância da distribuição de Pascal são:

$E(X) = k/p;$

$Var(X) = k(1-p)/p^2.$

Exemplo 4.14. Suponha que o experimento considerado no **Exemplo 4.13** seja modificado de modo que os animais são selecionados e examinados até a ocorrência de 10 animais afetados com a doença. Então, a distribuição de probabilidade da variável aleatória X: número de animais que devem ser examinados até encontrar 10 animais com o sintoma da doença, é:

$$P[X=x] = C^{9}_{x-1} \times 0{,}05^{10}. \times 0{,}95^{x-10}, \quad x=10,11,12,...$$

4.3.8 Distribuição de Poisson

Um **experimento de Poisson** é uma seqüência de ensaios de Bernoulli com as mesmas características de um experimento binomial, mas em que o interesse reside não no número de sucessos em n ensaios mas no número de sucessos em um dado intervalo de tempo ou de espaço. Por exemplo, número de chuvas de granizo que ocorrem em um ano; número de frutos de uma planta que caem em uma safra em decorrência de uma doença; número de partos gêmeos em um rebanho de ovinos em um período de parição; número de bactérias por milímetro cúbico em uma dada cultura; e número de lebres em uma área. Tais experimentos são caracterizados pelo número esperado de sucessos por unidade de tempo ou de espaço, enquanto o experimento binomial é caracterizado pelo número de sucessos na realização de n ensaios. A especificação do intervalo de tempo ou da região substitui a especificação do tamanho n de um experimento binomial.

Um experimento binomial é baseado em ensaios de Bernoulli, caracterizados pela probabilidade comum p de sucesso. Em um experimento de Poisson, o intervalo de tempo ou a região em consideração deve ser dividido em subintervalos de amplitudes suficientemente pequenas para garantir que em cada subintervalo de tempo ou sub-região seja efetuado apenas um ensaio de Bernoulli; por exemplo, ocorra apenas um granizo, ou caia apenas um fruto, ou apareça apenas uma bactéria. No exemplo do número de bactérias por milímetro cúbico, por menor que seja a unidade de área considerada, sempre há a possibilidade da ocorrência de mais de uma bactéria. Então, o espaço deve ser subdividido em unidades de áreas tão pequenas que seja garantido que a probabilidade da ocorrência de mais de uma bactéria em uma unidade de área seja insignificante. Quando um intervalo de tempo ou de espaço é dividido em um número n de subintervalos de modo que a ocorrência de mais de um sucesso seja insignificante, tem-se uma situação aproximada de um experimento de Poisson. Nessa situação, n é usualmente muito grande e a probabilidade p de sucesso muito pequena.

A variável aleatória X: número de sucessos no intervalo de tempo ou de espaço fixado, tem a **distribuição de Poisson**:

$$f(x) = \frac{e^{-k}\, k^{x}}{x!}, \quad x=0,1,2,...,$$

onde o parâmetro $k>0$ é o número médio de sucessos que ocorrem no dado intervalo de tempo ou de espaço, e $e=2{,}71828...$

Exemplo 4.15. O número médio de frutos que caem de um pessegueiro em uma safra com a duração de 20 dias é 2. Qual é a distribuição de probabilidade da variável aleatória X: número de frutos que caem de um pessegueiro durante uma safra? Determine a probabilidade de que em uma safra caia apenas um fruto de um pessegueiro?

A variável aleatória X tem distribuição de Poisson com k=2:

$$P[X=x] = \frac{e^{-2}2^x}{x!},\ x=0,1,2,...$$

Logo,

$$P[X=1] = \frac{e^{-2}2^1}{1!} = 2e^{-2} = \frac{2}{2,71828^2} = 0,7358.$$

A distribuição de Poisson também é útil como uma aproximação da distribuição binomial, quando o número n de ensaios de Bernoulli é muito grande e a probabilidade p de sucesso é muito pequena. Nessa situação, o cálculo direto de probabilidades binomiais pode envolver cálculos trabalhosos. De fato, demonstra-se que, se uma variável aleatória X tem distribuição binomial b~(n;p) com n muito grande e p muito pequeno de modo que np permanece constante, então a distribuição de X aproxima-se da distribuição de Poisson com k=np, quando n cresce indefinidamente.

Exemplo 4.16. Um medicamento para o tratamento de uma doença de bovinos manifesta um efeito colateral raro em 0,25% dos animais a que é administrado. Qual é a probabilidade de que 10 animais de um rebanho de 3.000 animais manifestem o efeito colateral?

Trata-se, claramente, de um experimento binomial com n=3.000 e p=0,0025. A probabilidade em questão é dada por:

$$P[X=10] = C_{3.000}^{10} \times 0,0025^{10} \times 0,9975^{2.990}\ !$$

Probabilidades para a variável aleatória X podem ser obtidas, com boa aproximação, pela distribuição de Poisson com k = np = 3.000×0,0025 = 7,5:

$$P[X=x] = \frac{e^{-7,5}7,5^x}{x!}.$$

Donde:

$$P[X=10] = \frac{e^{-7,5}7,5^{10}}{10!} = 0,0858.$$

4.4 Distribuição Conjunta de Duas Variáveis Aleatórias

Nas sessões anteriores, tratou-se da distribuição de uma variável aleatória. Muito freqüentemente, entretanto, o pesquisador tem interesse em duas ou mais características do resultado de um experimento aleatório e, portanto, deve considerar as distribuições de duas ou mais variáveis aleatórias. Assim, no experimento do **Exemplo 4.2**, pode-se ter interesse na altura da planta e no peso seco das raízes aos vinte dias, além da germinação da semente. Se as variáveis

observadas simultaneamente não são relacionadas, suas propriedades podem ser estabelecidas e estudadas através das respectivas distribuições univariadas. Muito comumente, entretanto, as variáveis são relacionadas e a determinação de seu inter-relacionamento constitui um dos propósitos da pesquisa. Neste caso, a distribuição conjunta das duas variáveis deve ser estabelecida e estudada.

O conceito de variável aleatória estende-se para essa situação como segue:

Sejam X e Y duas variáveis aleatórias que mapeam cada ponto do espaço básico S em um e somente um ponto sobre o plano $\mathbf{R}^2$, de modo que a imagem inversa sobre S de cada evento A do espaço R^2_{XY} (conjunto de pontos sobre $\mathbf{R}^2$ gerados pelo mapeamento efetuado por X e Y) seja um evento de S (**Figura 4.7**). Então, o par ordenado (X,Y) constitui uma **variável aleatória bidimensional**.

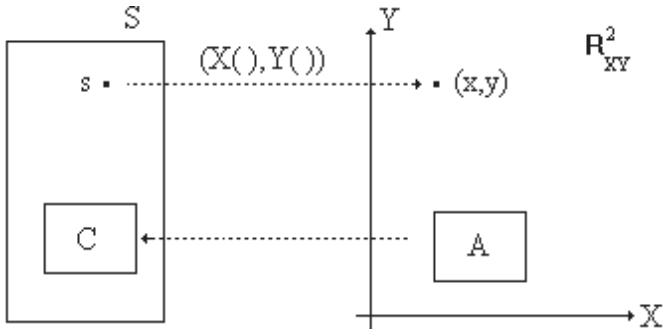


Figura 4.7. Ilustração do conceito de variável aleatória bidimensional

De modo geral, tem-se um espaço básico S e n variáveis aleatórias, definidas em S, que mapeam o espaço S em pontos do hiperplano $\mathbf{R}^n$, onde n é um número inteiro igual ou maior que 2. Nesta Seção, tratar-se-á da situação bivariada (n=2), concluindo-se com algumas observações referentes à situação multivariada mais geral (n>2).

Se X e Y são duas variáveis aleatórias discretas, a probabilidade de que X assuma o valor x e Y assuma o valor y é a probabilidade da interseção dos eventos $C_1 = \{s \in S \mid X(s)=x\}$ e $C_2 = \{s \in S \mid Y(s)= y\}$, denotada por P[X=x,Y=y]; ou seja:

$$P[X=x,Y=y] = P[\{s \in S \mid X(s)=x \text{ e } Y(s)=y].$$

Se X e Y são variáveis aleatórias discretas, a função f(x,y) = P[X=x,Y=y] definida para os pares de valores (x,y) no espaço de X e Y é denominada **função distribuição de probabilidade conjunta**, ou **função de probabilidade conjunta**, ou **distribuição de probabilidade conjunta** de X e Y.

Como na situação univariada, a função de probabilidade conjunta de X e Y pode ser especificada por uma lista dos pares de seus distintos valores possíveis (x,y) e suas respectivas probabilidades, dispostos em uma tabela. Mais especificamente, suponha-se que as variáveis aleatórias X e Y assumem, respectivamente, I e J valores distintos x_1 x_2...,x_I e y_1 y_2...,y_J com

probabilidades positivas. Então, a **variável aleatória bidimensional** (X,Y) poderá assumir probabilidades positivas em I×J pares de valores (x_i,y_i), i=1,2,...,I; j=1,2,...,J. A distribuição de probabilidade conjunta das duas variáveis aleatórias discretas X e Y pode ser especificada por uma tabela de dupla-entrada (**Tabela 4.8**), com os valores da variável aleatória X em uma das entradas e os valores da variável aleatória Y na outra entrada, e as probabilidades $f(x_i,y_j)$ correspondentes aos pares de valores (x_i,y_j) nas células.

Tabela 4.8. Distribuição de probabilidade conjunta de duas variáveis aleatórias X e Y

X	Y			
	y_1	y_2	...	y_J
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	...	$f(x_1,y_J)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	...	$f(x_2,y_J)$
...	...	...	...	...
x_I	$f(x_I,y_1)$	$f(x_I,y_2)$	...	$f(x_I,y_J)$

Exemplo 4.17. Considere-se o experimento de germinação de duas sementes, **Exemplo 4.2**, para o qual foi definida a variável aleatória X: número de sementes que germinam. Outras variáveis aleatórias podem ser consideradas para esse mesmo experimento, como, por exemplo, Y: germinação da primeira semente, e Z: germinação da segunda semente. As expressões dessas três variáveis aleatórias são:

$$X(s)=\begin{cases}0,\ s=gg\\1,\ s=Gg,gG\\2,\ s=GG\end{cases};\quad Y(s)=\begin{cases}0,\ s=gg,gG\\1,\ s=Gg,GG\end{cases}\quad\text{e}\quad Z(s)=\begin{cases}0,\ s=gg,Gg\\1,\ s=gG,GG\end{cases}$$

Os pares de valores das variáveis aleatórias X e Y, nesta ordem, são: (0,0), (0,1), (1,0), (1,1), (2,0), (2,1). Supondo que a probabilidade de uma semente germinar é 0,9 (**Exemplo 4.3**), as probabilidades para os dois primeiros pares de valores de X e Y são obtidas a seguir:

$$f(0,0) = P[X=0,Y=0] = P[\{gg\}\cap\{gg,gG\}] = P[\{gg\}] = 0{,}1\times 0{,}1 = 0{,}01;$$

$$f(0,1) = P[X=0,Y=1] = P[\{gg\}\cap\{GG,Gg\}] = P[\varnothing] = 0.$$

A determinação das probabilidades dos eventos definidos pelos demais pares de valores (x_i,y_j), de modo semelhante, produz a distribuição de probabilidade conjunta de X e Y da **Tabela 4.9**.

Tabela 4.9. Distribuição de probabilidade conjunta de X e Y, **Exemplo 4.17**.

X	Y		Soma
	0	1	
0	0,01	0,00	0,01
1	0,09	0,09	0,18
2	0,00	0,81	0,81
Soma	0,10	0,90	1,00

Semelhantemente, pode-se determinar a distribuição de probabilidade conjunta das variáveis aleatórias Y e Z, apresentada na **Tabela 4.10**.

Tabela 4.10. Distribuição de probabilidade conjunta de Y e Z, **Exemplo 4.17**.

Z	Y		Soma
	0	1	
0	0,01	0,09	0,10
1	0,09	0,81	0,90
Soma	0,10	0,90	1,00

Alternativamente, a distribuição conjunta f(x,y) pode ser especificada, mais convenientemente, por uma expressão analítica, ou seja, uma equação que exprima os valores f(x,y) = P[X=x,Y=y] para cada par de valores (x,y) do espaço da variável aleatória (X,Y). Por exemplo, a distribuição de probabilidade da **Tabela 4.10** pode ser especificada, alternativamente, pela equação:

$$f(y,z) = 0{,}9^{y+z}.0{,}1^{2-y-z},\ y, z = 0, 1; y+z=2.$$

A probabilidade de qualquer evento referente ao espaço de X e Y pode ser determinada a partir da função de probabilidade conjunta f(x,y). Assim, para **o Exemplo 4.17**:

P[X=1] = P[X=1,Y=0] + P[X=1,Y=1] = f(1,0) + f(1,1) = 0,09 + 0,09 = 0,18;

P[X+Y=2] = f(2,0) + f(1,1) = 0,0 + 0,09 = 0,09;

P[X>Y] = f(1,0) + f(2,0) + f(2,1) = 0,09 + 0,0 + 0,81= 0,90.

Propriedades da função distribuição de probabilidade conjunta

Uma função bivariada f(x,y) é uma **distribuição de probabilidade conjunta** de duas variáveis aleatórias X e Y se e somente se satisfaz às duas seguintes propriedades, análogas àquelas da função de probabilidade de uma variável aleatória discreta:

1) $f(x,y) > 0$ para qualquer par de valores de X e Y;

2) $\sum_x \sum_y f(x,y) = 1$, onde a soma dupla se estende sobre todos os pares de valores possíveis de X e Y do domínio de f(x,y).

Representação geométrica

A distribuição de probabilidade conjunta f(x,y) é representada, geometricamente, em um espaço de três dimensões. A probabilidade de cada ponto (x,y) do plano $\mathbf{R}^2$ é representada por um segmento de altura f(x,y). Para ilustração, considere-se o **Exemplo 4.18**.

Exemplo 4.18. A **Tabela 4.11** apresenta a distribuição de probabilidade conjunta, derivada de observações em um grande número de plantas de macieira, referente aos números de frutos em uma planta com dois tipos de defeitos raros D_1 e D_2 , denotados, respectivamente por X e Y.

Tabela 4.11. Distribuição conjunta das variáveis aleatórias X e Y do **Exemplo 4.18**.

X	Y			Soma
	0	1	2	
0	0,50	0,10	0,05	0,65
1	0,15	0,10	0,05	0,30
2	0,05	0,00	0,00	0,05
Soma	0,70	0,20	0,10	1,00

A representação geométrica da distribuição de probabilidade conjunta das variáveis aleatórias X e Y do **Exemplo 4.18** é apresentada na **Figura 4.8**.

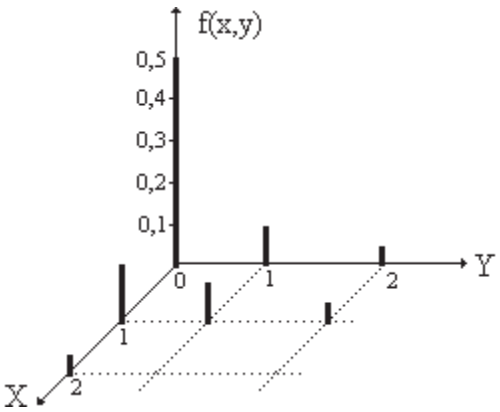


Figura 4.8. Representação geométrica da função de probabilidade conjunta do **Exemplo 4.18**.

4.5 Distribuição de uma Função de Duas Variáveis Aleatórias

A distribuição de probabilidade de qualquer função das variáveis aleatórias X e Y pode ser derivada da distribuição de probabilidade conjunta f(x,y). Para ilustração, considere-se, no **Exemplo 4.18**, o número de frutos de uma planta de macieira que apresentam defeitos de um dos tipos D_1 ou D_2 , ou seja, a variável aleatória Z = X+Y. Esta variável aleatória Z pode assumir os valores 0, 1, 2, 3 e 4. Probabilidades para Z podem ser determinadas como segue:

P[Z=0] = P[X+Y=0] = f(0,0) = 0,50;

P[Z=1) = P[X+Y=1] = f(1,0) + f(0,1) = 0,15+0,10 = 0,25;

P[Z=2] = P[X+Y=2] = f(2,0) + f(1,1) + f(0,2) = 0,05+0,10+0,05 = 0,20.

A distribuição de probabilidade de Z é apresentada na **Tabela 4.12**.

Tabela 4.12. Distribuição de probabilidade de Z=X+Y, **Exemplo 4.18**

Z:	0	1	2	3	4
P[Z]:	0,50	0,25	0,20	0,05	0,00

4.6 Distribuição Marginal

As distribuições de probabilidade das variáveis aleatórias X e Y podem ser obtidas da distribuição de probabilidade conjunta f(x,y). A probabilidade de cada valor possível x_i de X é a soma das probabilidades conjuntas $f(x_i, y_j)$ sobre os valores y_j de Y para $X=x_i$, ou seja: $f_X(x_i) = \sum_j f(x_i, y_j)$. Semelhantemente, a probabilidade de cada valor possível y_j de Y é a soma das probabilidades conjuntas sobre os valores x_i para $Y=y_j$: $f_Y(y_j) = \sum_i f(x_i, y_j)$. Para o **Exemplo 4.18**,

as probabilidades dos distintos valores de X estão na margem oposta à entrada para X na **Tabela 4.11**, ou seja, na última coluna, e as probabilidades para Y, na última linha. Como essas probabilidades são os totais marginais da tabela de dupla-entrada das probabilidades conjuntas, essas distribuições de X e Y são denominadas **distribuições marginais**.

Se X e Y são variáveis aleatórias discretas com distribuição de probabilidade conjunta f(x,y), as funções:

$$f_X(x_i) = \sum_j f(x_i, y_i) \quad \text{e}$$

$$f_Y(y_j) = \sum_i f(x_i, y_i)$$

são denominadas **distribuições marginais** de X e Y, respectivamente.

Essas distribuições são, simplesmente, as distribuições univariadas de X e Y, aqui denotadas por $f_X(x)$ e $f_Y(y)$ com os subscritos utilizados para distinguir os símbolos que denotam as duas funções de probabilidade. O qualificativo "marginal" na designação da distribuição de uma variável aleatória é, de fato, redundante. Quando aplicado à distribuição de X, por exemplo, ele apenas significa que outra variável aleatória, além de X, foi originalmente considerada no problema e que a distribuição de X foi obtida a partir da tabela de dupla-entrada da distribuição conjunta de X e Y.

4.7 Valor Esperado de uma Função de Duas Variáveis Aleatórias

O conceito de valor esperado pode ser estendido à situação de duas variáveis aleatórias discretas:

Se Z = g(X,Y) é uma função de duas variáveis aleatórias X e Y, com distribuição conjunta f(x,y), seu **valor esperado** é:

$$E[g(x,y)] = \sum_x \sum_y g(x,y) f(x,y).$$

São particularmente importantes as seguintes propriedades do valor esperado em distribuições de duas variáveis aleatórias.

a) O valor esperado (média) de cada uma das variáveis aleatórias X e Y, com distribuição conjunta f(x,y), pode ser determinado através das respectivas distribuições marginais. De fato, pela definição:

$$E(X) = \sum_x \sum_y x\, f(x,y) = \sum_x x \sum_y f(x,y) = \sum_x f_x(x);$$

$$E(Y) = \sum_x \sum_y y\, f(x,y) = \sum_x y \sum_y f(x,y) = \sum_y f_y(y).$$

Semelhante propriedade pode ser estabelecida para as variâncias e desvios padrões de X e Y.

b) O valor esperado da soma de duas variáveis aleatórias é a soma de seus valores esperados:

$$E(X+Y) = \sum_x \sum_y (x+y)\, f(x,y) = \sum_x \sum_y x\, f(x,y) + \sum_x \sum_y y\, f(x,y) =$$

$$= \sum_x x \sum_y f(x,y) + \sum_x y \sum_y f(x,y) = \sum_x x f_X(x) + \sum_y y f_Y(y) =$$

$$= E(X) + E(Y).$$

Entretanto, a variância da soma de duas variáveis aleatórias é igual à soma das respectivas variâncias apenas na situação particular de que se tratará adiante.

4.8 Covariância e Correlação de Duas Variáveis Aleatórias

A covariância das variáveis aleatórias X e Y é uma medida da variação conjunta de X e Y:

> A **covariância** de duas variáveis aleatórias X e Y, denotada por Cov(X,Y) ou σ_{XY}, é o valor esperado do produto dos desvios dessas variáveis aleatórias em relação aos seus respectivos valores esperados:
>
> $$\text{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\}.$$

Intuitivamente, pode-se dizer que X e Y variam na mesma direção se é alta a probabilidade de que valores elevados de X sejam associados com valores elevados de Y e valores baixos de X sejam associados com valores baixos de Y. Nessa situação, a probabilidade de que os desvios X-*E*(X) e Y-*E*(Y) sejam ambos positivos ou ambos negativos é elevada, de modo que o produto [X-*E*(X)][Y-*E*(Y)] é predominantemente positivo. conseqüentemente, o valor esperado desse produto é positivo e elevado. Por outro lado, se X e Y tendem a variar em sentidos opostos, valores positivos de X-*E*(X) associam-se mais freqüentemente com valores negativos de Y-*E*(Y), e vice-versa. Então, o produto [X-*E*(X)][Y-*E*(Y)] é predominantemente negativo e seu valor esperado é negativo. Dessa forma, o sinal e a grandeza de $E\{[X-E(X)][Y-E(Y)]\}$ refletem o sentido e o grau da **relação linear** entre as variáveis aleatórias X e Y. É nesse sentido que a covariância exprime a relação, ou associação, entre os valores de X e Y.

Pode-se verificar que:

$$\text{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\},$$

$$= E(XY) - E(X).E(Y),$$

onde a última expressão pode ser obtida pelo desenvolvimento da primeira. Assim, a covariância de X e Y pode ser determinada a partir dos valores esperados de X e Y, que podem ser obtidos através das respectivas distribuições (marginais), e do valor esperado do produto XY, que pode ser obtido a partir da distribuição conjunta de X e Y, como segue:

$$E(XY) = \sum_x \sum_y xy\, P[X=x, Y=y].$$

O valor da Cov(X,Y) depende das unidades de medida associadas com as variáveis aleatórias X e Y. Muito freqüentemente, é conveniente uma medida da relação de associação linear

entre as duas variáveis que não dependa de unidades de medida, como se logra com o uso do coeficiente de variação como medida de dispersão. Isso é obtido pela divisão da covariância pelos desvios padrões de X e Y, se estes desvios padrões são ambos positivos. A medida resultante é o coeficiente de correlação de X e Y:

Uma medida da associação linear de duas variáveis aleatórias X e Y é **coeficiente de correlação linear**, denotado por Corr(X,Y) ou ρ_{XY} e expresso por:

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y},$$

contanto que Cov(X,Y), σ_X e σ_Y existam, e $\sigma_X>0$ e $\sigma_Y>0$.

Exemplo 4.19. Para ilustração do cálculo da covariância e do coeficiente de correlação, considere-se a situação do **Exemplo 4.18**. As médias das variáveis aleatórias X e Y podem ser obtidas a partir das respectivas distribuições marginais da **Tabela 4.11**:

$$E(X) = 0\times0{,}65 + 1\times0{,}30 + 2\times0{,}05 = 0{,}4;$$

$$E(Y) = 0\times0{,}70 + 1\times0{,}20 + 2\times0{,}10 = 0{,}4.$$

O valor de $E(XY)$ é determinado como segue:

$$E(XY) = \sum\sum xy\, f(x,y)$$

$$= 0\times0\times0{,}50 + 0\times1\times0{,}10 + 0\times2\times0{,}05 + 1\times0\times0{,}15 + 1\times1\times0{,}10 +$$

$$+ 1\times2\times0{,}05 + 2\times0\times0{,}05 + 2\times1\times0 + 2\times2\times0 = 0{,}20.$$

Assim,

$$\text{Cov}(X,Y) = 0{,}20 - 0{,}4\times0{,}4 = 0{,}04.$$

Para determinar o coeficiente de correlação, necessita-se dos desvios padrões de X e Y, determinados como segue:

$$E(X^2) = 0^2\times0{,}65 + 1^2\times0{,}30 + 2^2\times0{,}05 = 0{,}5 \text{ e}$$

$$E(Y^2) = 0^2\times0{,}70 + 1^2\times0{,}20 + 2^2\times0{,}10 = 0{,}6;$$

logo:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = 0{,}5 - 0{,}4^2 = 0{,}34 \text{ e}$$

$$\text{Var}(Y) = E(Y^2) - [E(Y)]^2 = 0{,}6 - 0{,}4^2 = 0{,}44;$$

donde:

$$\sigma_X = \sqrt{0{,}34} = 0{,}583 \text{ e}$$

$$\sigma_Y = \sqrt{0{,}44} = 0{,}663;$$

portanto, o coeficiente de correlação é:

$$\rho_{XY} = \frac{0{,}04}{0{,}583\times0{,}663} = 0{,}103.$$

Propriedades do coeficiente de correlação

1) O coeficiente de correlação situa-se no intervalo entre -1 e 1: $-1 \leq \rho_{XY} \leq 1$.

2) O coeficiente de correlação não se altera se as variáveis aleatórias são adicionadas ou multiplicadas de constantes com o mesmo sinal.

4.9 Distribuição Condicional e Independência Estatística

No **Capítulo 3**, a probabilidade condicional de um evento A dado um evento B foi definida como:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \text{ contanto que } P(B) \neq 0.$$

Suponha-se que os pontos X=x e Y=y são os mapeamentos dos eventos A e B do espaço básico de um experimento, efetuados pelas variáveis aleatórias X e Y, respectivamente. Então, pode-se escrever:

$$P[X=x \mid Y=Y] = \frac{P\left[X = x, Y = y\right]}{P[Y = y]}$$

$$= \frac{f\left(x, y\right)}{f_Y(y)},$$

desde que $P[Y=y] = f_Y(y) \neq 0$, onde f(x,y) é o valor da função de distribuição conjunta de X e Y no ponto (x,y) e $f_Y(y)$ é o valor da distribuição marginal de Y em Y=y. Então, denotando por f(x | y) as probabilidades dos valores possíveis da variável aleatória X para um valor fixo qualquer y da variável aleatória Y, define-se a distribuição condicional de X dado um valor particular y de Y como segue:

A função:

$$f(x \mid y) = \frac{f\left(x, y\right)}{f_Y(y)}, \; f_Y(y) \neq 0,$$

definida para cada valor x no espaço de X, é denominada **distribuição condicional** de X dado Y=y.

Semelhantemente, a função:

$$f(y \mid x) = \frac{f\left(x, y\right)}{f_X(x)}, \; f_X(x) \neq 0,$$

para cada valor y no espaço de Y, é denominada **distribuição condicional** de Y dado X=x.

As variáveis aleatórias X e Y são **independentes** (no sentido estatístico) se os eventos $\{s\in S \mid X(s)=x_i\}$ e $\{s\in S \mid Y(s)=y_j\}$ são independentes para todos os pares de valores possíveis (x_i,y_j) de X e Y.

Sejam A e B os eventos no espaço básico S correspondentes aos pontos $X(s)=x_i$ e $Y(s)=y_j$. Então, como A e B são independentes, por definição: $P[A\cap B] = P[A].P[B]$, o que implica que:

$$P[X=x_i,Y=y_j] = P[X=x_i].P[Y=y_j].$$

Então, pode-se estabelecer a definição de independência estatística como segue:

As variáveis aleatórias X e Y são **independentes** (no sentido estatístico) se:

$$f(x_i,y_j) = f_X(x_i).f_Y(y_j)$$

para todos os pares de valores possíveis (x_i,y_j).

Isso significa que X e y são independentes se a probabilidade de cada ponto (x_i,y_j) no espaço $\mathbf{R}^2_{XY}$ é o produto das probabilidades marginais nos pontos $X=x_i$ e $Y=y_i$, nos correspondentes espaços $\mathbf{R}_X$ e $\mathbf{R}_Y$. Na representação tabular da distribuição de probabilidade conjunta, isto significa que as probabilidades nas células são os produtos das correspondentes probabilidades totais nas margens (probabilidades marginais).

Exemplo 4.20. Para ilustração, considere-se as distribuições de probabilidade especificadas na **Tabela 4.9** e na **Tabela 4.10**, do **Exemplo 4.17**. As variáveis aleatórias X e Y não são independentes, pois $P[X=0,Y=0] = 0{,}01$, enquanto que $P[X=0].P[Y=0] = 0\times 0 = 0$ (a falha da condição de independência para uma célula da tabela da distribuição conjunta é suficiente para implicar na dependência das variáveis aleatórias). Entretanto, é de esperar que as variáveis aleatórias Y e Z sejam independentes, já que elas se referem a duas partes fisicamente independentes de um experimento: o resultado da germinação da primeira semente não tem qualquer implicação sobre a germinação da segunda semente. De fato, pode-se verificar que os valores nas células da **Tabela 4.10** são os produtos dos valores nas correspondentes margens.

A independência de eventos implica nas seguintes propriedades importantes:

1) Se X e Y são variáveis aleatórias independentes, então:

$E(XY) = E(X)E(Y)$.

De fato, a independência de X e Y implica que $f(x,y) = f_X(x).f_Y(y)$; logo:

$$E(XY) = \sum_x \sum_y xy\, f(x,y) = \sum_x \sum_y xy\, f_X(x)\, f_Y(y) =$$

$$= \sum_x x\, f_X(x) \sum_y y\, f_Y(y) = E(X).E(Y).$$

2) Se X e Y são independentes,

$Cov(X,Y) = 0$ e

$Corr(X,Y) = 0$.

Estas propriedades são imediatas.

3) Se X e Y são independentes, então:

$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$.

Recorde-se que o valor esperado da soma de duas variáveis aleatórias é, sempre, igual à soma dos valores esperados das duas variáveis, sejam elas independentes ou não. Entretanto, propriedade semelhante não é sempre válida para a variância. De fato,

$$\begin{aligned}\text{Var}(X+Y) &= E[(X+Y) - E(X+Y)]^2 = E[\{X-E(X)\} + \{Y-E(Y)\}]^2 = \\ &= E[X-E(X)]^2 + E[Y-E(Y)]^2 + 2E[X-E(X)][Y-E(Y)] = \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y).\end{aligned}$$

Assim,

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

se e apenas se $\text{Cov}(X,Y)=0$, o que é implicado pela independência das variáveis aleatórias X e Y.

Semelhantemente, pode-se demonstrar que $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X,Y)$, de modo que $\text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)$, se e apenas se X e Y são independentes.

De acordo com a segunda propriedade, a independência de duas variáveis aleatórias implica na nulidade de sua covariância e, portanto, de sua correlação. Entretanto, a covariância e correlação nulas não implicam, necessariamente, na independência estatística.

4.10 Distribuição Conjunta de n Variáveis Aleatórias

Todos os conceitos referentes à distribuição de duas variáveis aleatórias discretas podem ser generalizados para a situação multivariada mais geral de n variáveis aleatórias. A dificuldade que surge é a impossibilidade de representação geométrica em espaços de mais de três dimensões. Tratar-se-á, aqui, apenas das extensões importantes para os propósitos deste texto.

Se $X_1,X_2,...,X_n$ são n variáveis aleatórias discretas definidas no mesmo espaço básico S, então, o vetor $(X_1,X_2,...,X_n)$ é uma **variável aleatória n-dimensional discreta**, que assume valores com probabilidades positivas em um número contável de pontos $(x_1,x_2,...,x_n)$ do espaço $\mathbf{R}^n$. Nesse caso, a **função distribuição de probabilidade conjunta** de $X_1,X_2,...,X_n$, denotada por $f_{X_1X_2\ldots X_n}(x_1,x_2,...,x_n)$ ou, mais simplesmente, $f(x_1,x_2,...,x_n)$, é definida como:

$$f_{X_1X_2\ldots X_n}(x_1,x_2,...,x_n) = P[X_1=x_1,X_2=x_2,...,X_n=x_n],$$

para todos os pontos do espaço $R^n_{X_1X_2\ldots X_n}$.

Uma função multivariada $f(x_1,x_2,...,x_n)$ é uma **função distribuição de probabilidade conjunta** de n variáveis aleatórias discretas $X_1,X_2,...,X_n$ se e apenas se satisfaz às seguintes duas propriedades, extensões do caso bivariado:

1) $f(x_1,x_2,...,x_n) \geq 0$ para todo ponto $(x_1,x_2,...,x_n)$ em R^n.

2) $\sum_{x_1}\sum_{x_2}...\sum_{x_n} f(x_1,x_2,...,x_n) = 1$, onde a soma estende-se sobre todos os valores possíveis de $X_1,X_2,...,X_n$.

Na situação de mais de duas variáveis aleatórias, pode-se considerar não apenas distribuições marginais de variáveis aleatórias individuais, mas, também, distribuições marginais conjuntas de várias variáveis aleatórias:

A **distribuição marginal** de X_1 apenas é dada por:

$$f_{X_1}(x_1) = \sum_{x_2}...\sum_{x_n} f_{X_1X_2...Xn}(x_1,x_2,...,x_n),$$

para todos os valores x_1 do espaço de X_1. A **distribuição marginal conjunta** de $X_1,X_2,...,X_p$, p<n, é dada por:

$$f_{X_1X2...X_p}(x_1,x_2,...,x_p) = \sum_{x_{p+1}}...\sum_{x_n} f_{X_1X_2...Xn}(x_1,x_2,...,x_n),$$

para todos os valores $x_1,x_2,...,x_p$ nos espaços de $X_1,X_2,...,X_p$, respectivamente.

Outras distribuições marginais podem ser definidas de modo semelhante.

Semelhantemente, pode-se considerar vários tipos de distribuições condicionais. Por exemplo, se $f(x_1,x_2,...,x_n)$ é o valor da distribuição conjunta das variáveis aleatórias $X_1,X_2,...,X_n$ em $(x_1,x_2,...,x_n)$, a **distribuição condicional** de X_1 dado $X_2,...,X_n$ é:

$$f_{X_1|X_2...X_n}(x_1|x_2,...,x_n) = \frac{f_{X_1X_2...X_n}(x_1,x_2,...,x_n)}{f_{X_2...X_n}(x_2,...,x_n)}.$$

A distribuição **condicional conjunta** de $X_1,X_2,...,X_p$ (p<n) dado $X_{p+1},...,X_n$ é:

$$f_{X_1...X_p|X_{p+1}...X_n}(x_1,...,x_p|x_{p+1},...,x_n) = \frac{f_{X_1X_2...X_n}(x_1,x_2,...,x_n)}{f_{X_{p+1}...X_n}(x_{p+1},...,x_n)}$$

Em situações de duas ou mais variáveis aleatórias, questões de independência são de grande importância. A definição de independência estatística é generalizada como segue:

Se $f(x_1,x_2,...,x_n)$ é a distribuição de probabilidade conjunta de n variáveis aleatórias discretas $X_1,X_2,...,X_n$ e $f_{X_i}(x_i)$, i=1,2,...,n, é a distribuição de probabilidade marginal de X_i , i=1,2,...,n, então as variáveis aleatórias $X_1,X_2,...,X_n$ são (estatisticamente) **independentes** se e somente se:

$$f_{X_1X_2\ldots X_n}(x_1,x_2,...,x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \ldots f_{X_n}(x_n)$$

para todos os pontos $(x_1,x_2,...,x_n)$ do espaço $R^{n}_{X_1X_2\ldots X_n}$.

Exemplo 4.21. Considere-se o experimento de germinação sucessiva de cinco sementes em que a probabilidade de uma semente germinar é p=0,9. Como a probabilidade de qualquer das cinco sementes germinar não é afetada pela germinação das demais sementes, a probabilidade de que as três primeiras sementes germinem e as duas últimas não germinem é: 0,9×0,9×0,9×0,1×0,1 = 0,0729.

Os conceitos de valor esperado e correspondentes propriedades podem ser estendidos para variáveis aleatórias n-dimensionais como segue:

O **valor esperado** de uma função $g(x_1,x_2,...,x_n)$ de uma variável aleatória n-dimensional discreta $(X_1,X_2,...,X_n)$ com distribuição de probabilidade conjunta $f(x_1,x_2,...,x_n)$ é:

$$E[g(X_1,X_2,...,X_n)] = \sum_{x_1}\sum_{x_2}\ldots\sum_{x_n} g(x_1,x_2,\ldots,x_n)\, f(x_1,x_2,\ldots,x_n) \ .$$

Propriedades do valor esperado

1) Se $X_1,X_2,...,X_n$ são quaisquer n variáveis aleatórias discretas:
$E(X_1+X_2+...+X_n) = E(X_1) + E(X_2) +...+ E(X_n)$.

2) Se $X_1,X_2,...,X_n$ são variáveis aleatórias independentes:
$E(X_1,X_2,...,X_n) = E(X_1)\, E(X_2)...E(X_n)$.

3) Se $c_1,c_2,...,c_n$ são constantes e $g_i(x_1,x_2,...,x_n)$, i=1,2, ...,n, são n funções das variáveis aleatórias $X_1,X_2,...,X_n$, então:

$$E[\sum_{i=1}^{n} c_i g_i(X_1,X_2,...,X_n)] = \sum_{i=1}^{n} c_i\, E[g_i(x_1,x_2,...,x_n)].$$

Valores esperados e variâncias de combinações lineares de variáveis aleatórias são freqüentemente úteis na teoria da amostragem e em problemas de inferência estatística. Sejam $X_1,X_2,...,X_n$ variáveis aleatórias e $Y = \sum_{i=1}^{n} a_i x_i$, onde $a_1,a_2,...,a_n$ são constantes. Então:

1) $E(Y) = \sum_{i=1}^{n} a_i\, E(X_i)$.

$$2)\ \mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2\,\mathrm{Var}(X_i) + \sum_{i<j}^{n}\sum^{n} a_i a_j\,\mathrm{Cov}(X_i,X_j)$$

onde a soma dupla estende-se para todos os valores de i e j de 1 a n, para i<j.

Se as variáveis aleatórias $X_1,X_2,...,X_n$ são independentes, então:

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} a_i^2\,\mathrm{Var}(X_i).$$

4.11 Distribuição Multinomial

Uma importante generalização do experimento binomial é uma seqüência de ensaios independentes cada um com mais de dois resultados alternativos possíveis, com a mesma probabilidade dos respectivos resultados para todos os ensaios. Uma seqüência de tais ensaios é denominada um **experimento multinomial**. Essa situação ocorre, por exemplo, quando plantas de um conjunto são classificadas em mais de duas categorias quanto ao nível de infecção de uma doença, como, por exemplo, sem infecção e com infecção fraca, regular e forte.

Considere-se uma seqüência de n ensaios independentes cada um com k resultados alternativos possíveis mutuamente exclusivos $a_1,a_2,...,a_k$, respectivamente com probabilidades $p_1,p_2,...,p_k$ ($p_1+p_2+...+p_k=1$). Sejam $X_1,X_2,...,X_n$ as variáveis aleatórias que exprimem os números dos resultados com as alternativas $a_1,a_2,...,a_k$, respectivamente, nos n ensaios, com $X_1+X_2+...+X_k=n$.

Procedendo como na derivação da expressão da distribuição binomial, observe-se, inicialmente, que a probabilidade de x_1 resultados com a alternativa a_1, x_2 resultados com a alternativa a_2, etc. e x_k resultados com a alternativa a_k, em uma ordem específica, é $p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$. O número de ordens em que podem ocorrer esses resultados alternativos é dado pela regra das permutações com repetições, ou seja, $P = \dfrac{n!}{x_1!x_2!\ldots x_k!}$. Assim, a probabilidade de $x_1,x_2,...,x_k$ resultados dos n ensaios com as alternativas $a_1,a_2,...,a_k$, respectivamente, é dada por: $\dfrac{n!}{x_1!x_2!\ldots x_k!}\, p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}$.

As variáveis aleatórias $X_1,X_2,...,X_n$ têm **distribuição multinomial** se e apenas se sua distribuição de probabilidade conjunta é:

$$f(x_1,x_2,...,x_k) = \frac{n!}{x_1!x_2!\ldots x_k!}\; p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k},$$

$x_i = 0,1,2,...,n$ (i=1,2,...k), $x_1+x_2+...+x_k=n$, onde n, $p_1,p_2,...,p_k$ são parâmetros satisfazendo às condições: n=1,2,..., $0 \le p \le 1$, i=1,2, ...,k, $p_1+p_2+...+p_k = 1$.

A designação "multinomial" provém do fato de que a probabilidade para cada conjunto de valores $x_1,x_2,...,x_k$ é igual ao termo correspondente da expansão multinomial de $(p_1+p_2+...+p_k)^n$.

Exemplo 4.22. De acordo com a teoria mendeliana da herança, o cruzamento de uma planta de ervilha de semente amarela redonda com uma planta de ervilha de semente verde angulosa produz uma planta que produz sementes amarela redonda, amarela angulosa, verde

redonda e verde angulosa com probabilidades 9/16, 3/16, 3/16 e 1/16, respectivamente. Então, a probabilidade de que de nove plantas assim obtidas quatro produzam semente amarela redonda, duas produzam semente amarela angulosa, três produzam semente verde redonda e nenhuma produza semente amarela angulosa é dada por:

$$P[X_1=4,X_2=2,X_3=3,X_4=0] = \frac{9!}{4!2!3!0!}\left(\frac{9}{16}\right)^4\left(\frac{3}{16}\right)^2\left(\frac{3}{16}\right)^3\left(\frac{1}{16}\right)^0$$
$$= 0{,}02923.$$

4.12 Exercícios

1. Supondo que a probabilidade de um casal ter um filho de olhos azuis é igual a 1/4, qual é a probabilidade de em uma prole de 6 filhos:
   a) Nenhum ter olhos azuis?
   b) Pelo menos um ter olhos azuis?
2. Uma caixa contém 4 pintos da raça A e 6 pintos da raça B. Suponha que se retira pintos da caixa através de um processo que atribui a cada pinto na caixa igual probabilidade de ser extraído:
   a) Extraindo-se um pinto, qual é a probabilidade de sair um pinto da raça A?
   b) Retirando-se dois pintos, simultaneamente (isto é, sem reposição), qual é a probabilidade de sair um pinto de cada raça?
   c) Extraindo-se dois pintos, com a reposição do primeiro antes da extração do segundo, qual é a probabilidade de sair um pinto de cada raça?
3. Um lote de sementes tem poder germinativo igual a 80%. Se 5 sementes são escolhidas ao acaso, quantas devem germinar para que se tenha no mínimo 60% de germinação?
4. Um produtor de semente de soja produz sementes com 95% de poder germinativo e as embala em sacos de 60 kg. Quando um lote é vendido, o comprador tem direito à indenização de um saco de semente para cada saco em que a germinação é menor que a indicada. Se esse produtor possui 500 sacos, determine:
   a) Quantos sacos deverá manter para indenização?
   b) Em um conjunto de 10 sacos que ele vende, qual é a probabilidade de que até 3 deles apresentem germinação inferior a 95%?
5. Em um pomar de macieira foram arrancadas 100 plantas para verificar o grau de incidência de Phytophtora nas raízes, obtendo-se os seguintes resultados: 60 plantas não atacadas, 30 plantas levemente atacadas e 10 com ataque severo. Suponha que desse mesmo pomar são arrancadas mais 10 plantas. Determine:
   a) A probabilidade de se obter pelo menos 2 plantas levemente atacadas e 1 severamente atacada.
   b) A média e a variância das variáveis aleatórias.

6. De acordo com a teoria mendeliana da herança de um caráter, a fertilização cruzada de duas espécies de plantas de flores vermelhas e brancas produz descendentes dos quais 25% são plantas de flor vermelha. Suponha que um horticultor cruza 5 pares de plantas vermelhas e brancas. Qual é a probabilidade de que dos resultantes 5 descendentes se obtenha:
   a) Nenhuma planta de flor vermelha.
   b) Quatro ou mais plantas de flor vermelha.
7. Sabe-se que do acasalamento de um touro com vacas de um rebanho podem originar-se descendentes aspados e mochos. Considere o experimento "acasalamento do touro com três vacas do rebanho e observação do caráter aspas dos descendentes".
   a) Especifique o espaço de amostra do experimento.
   b) Especifique o evento A: Pelo menos dois descendentes aspados.
   c) Determine a expressão da variável aleatória X: número de descendentes com aspas.
   d) Suponha que a probabilidade do caráter "aspas" em descendentes do acasalamento do touro com vacas do rebanho é 0,75. Determine a probabilidade do evento A especificado no item b.
8. A probabilidade de um leitão de uma leitegada de 10 leitões viver até a idade adulta é 0,3.
   a) Quantos leitões de uma leitegada de 10 leitões se pode esperar viverem até a idade adulta?
   b) Qual é a variância do número de leitões que chegam à idade adulta?
   c) Qual é a probabilidade de que seis leitões de uma leitegada de 10 leitões atinjam a idade adulta?
   d) Qual é a probabilidade de que 6 ou mais leitões cheguem à idade adulta?
9. Uma industria utiliza uma máquina para o enlatamento de compota de pêssego em latas de 500g que garante o peso da lata entre 485 e 515g. O controle de qualidade da indústria classifica as latas produzidas em três categorias: B-, A e B+, segundo o peso da lata se situe nos intervalos 485-495, 495-505 e 505-515, respectivamente.
   a) Considere o experimento aleatório: "retirada de duas latas de uma pilha e sua classificação". Especifique o espaço básico.
   b) Para o experimento "retirada de uma lata de uma pilha e classificação da lata", considere a seguinte variável aleatória:

$$X = \begin{cases} -1, & \text{se a classificação é B-} \\ 0, & \text{se a classificação é A} \\ 1, & \text{se a classificação é B+} \end{cases}$$

   cuja função de probabilidade é:

$$P[X=x] = \begin{cases} 3/4, \text{ se } x=0 \\ 1/8, \text{ se } x=-1,1 \end{cases}$$

   Determine: i) E(X); ii) Var(X); iii) P[X<1].

c) Para o experimento especificado no item a, determine a probabilidade de que ambas as latas sejam classificadas na categoria A.

10. A probabilidade de se obter, por autofecundações sucessivas, uma boa variedade de algodão é 1/20.000. Qual é a probabilidade de em 40.000 autofecundações se conseguir pelo menos uma boa variedade?

11. Três pintos são retirados da caixa com 10 pintos mencionada no exercício 2 (extração sem reposição). Suponha que X é a variável aleatória que denota o número de pintos da raça A extraídos da caixa.

a) Complete a seguinte tabela da distribuição de probabilidade:

x:	0	1	2	3
P[X=x]				

b) Determine: i) P[X=2]; ii) P[0,5<X<2,5]; iii) F(3)=P[X≤3]; iv) $E(X)$; v) Var(X).

12. A probabilidade de partos duplos em um rebanho de ovinos da raça Booroola é 0,1. Considere 5 partos de ovelhas desse rebanho. Determine:

a) $E(X)$.

b) Var(X).

c) Probabilidade de um parto duplo.

d) Probabilidade de pelo menos um parto duplo.

13. A função de probabilidade de uma variável aleatória é:

$$P[X = x] = \begin{cases} 1/3, \text{ se } x = 1,2,3 \\ 0, \text{ caso contrário} \end{cases}$$

Determine:

a) $E(X)$.

b) Var(X).

c) P[X≤2].

14. Supondo que uma planta de café tem a probabilidade de 0,30 de sobreviver a uma geada, qual é a probabilidade de que:

a) 9 em cada 10 plantas de um pomar resistam a uma geada;

b) pelo menos 7 plantas sobrevivam a uma geada.

15. Um engenheiro de segurança afirma que 1 em 10 acidentes de automóvel são devidos à fadiga do motorista. Determine as probabilidades de que entre 5 acidentes 0, 1, 2, 3, 4 e 5 sejam devidos à fadiga do motorista. Efetue a representação gráfica dessa função de probabilidade.

16. Uma pesquisa revelou que 15 porcento dos pacientes que usam um certo medicamento têm efeitos colaterais indesejáveis. Usando a tabela de probabilidades da distribuição binomial (Tabela I), determine a probabilidade de que entre 18 pacientes que recebem esse medicamento:

a) nenhum tenha qualquer efeito colateral indesejável;
b) exatamente 2 tenham efeitos colaterais indesejáveis;
c) pelo menos 5 tenham efeitos colaterais indesejáveis;
d) no máximo 8 tenham efeitos colaterais indesejáveis.

17. Um engenheiro de controle de qualidade inspeciona uma amostra aleatória de 2 liquidificadores de cada lote de 2 liquidificadores, e aceita o lote apenas se os dois aparelhos estão em boas condições de uso; caso contrário, todos os liquidificadores do lote são inspecionados com o custo imputado ao vendedor. Qual é a probabilidade de que um lote seja aceitado sem ulterior inspeção, se:
    a) o lote contém 5 liquidificadores que não estão em boas condições de uso;
    b) o lote contém 10 aparelhos que não estão em boas condições de uso;
    c) o lote contém 15 aparelhos que não estão em boas condições de uso.
18. Artigos de uma fábrica dos quais 0,5 por cento são defeituosos são embalados em caixas de papelão cada uma com 100 unidades.
    a) Que proporção de caixas está livre de artigos defeituosos?
    b) Qual é a proporção de caixas que contém dois ou mais artigos defeituosos?
19. Considere a distribuição de probabilidade de Poisson para cada um dos seguintes valores de λ: 0,25; 0,50 e 1,00.
    a) Determine a distribuição de probabilidade para esses quatro valores de λ. Arredonde as probabilidades para 4 algarismos decimais.
    b) Efetue a representação gráfica das 3 distribuições de probabilidade determinadas no item anterior.
    c) Determine o valor esperado e a variância da distribuição para λ=0,25. Porque esses valores são levemente diferentes de $E(X) = \lambda = 0,25$ e $Var(X) = \lambda = 0,25$?
20. Aditivos como antibióticos, vermífugos e inseticidas são incorporados em rações para animais em partes por milhão (ppm). Para mistura efetiva, o aditivo pode ser comprimido em bolinhas ("pellets") do tamanho de um grão na ração e então colorido com tintura vegetal para facilidade de identificação. O controle de qualidade da mistura é efetuado pela extração de uma amostra de volume fixo da ração com o aditivo e contagem do número de bolinhas coloridos do aditivo. Suponha conhecido que a contagem de bolinhas em um uma amostra de ração apropriadamente misturada segue uma distribuição de Poisson com λ=2,5.

    a) Determine a probabilidade de que: i) uma amostra não contenha nenhum bolinhas de aditivo; ii) uma amostra contenha exatamente uma bolinha de aditivo; iii) uma amostra contenha pelo menos uma bolinha de aditivo.
    b) Determine os resultados mais prováveis de ocorrem em aproximadamente 80% das amostras.
21. Suponha que um processo alternativo de controle de qualidade de misturas consideradas na questão anterior que requer uma amostra de dez volumes fixos

extraídos independentemente de cada partida de mistura. Em uma dessas amostras de ração apropriadamente misturada, determine:

a) o número total esperado de bolinhas coloridas;

b) a probabilidade de não aparecer qualquer dessas bolinhas.

22. Um centro telefônico recebe um número médio de 5 chamadas por hora. Admitindo que o número de chamadas recebidas nesse centro por hora tenha distribuição de Poisson, determine a probabilidade de que:

a) em uma hora particular sejam recebidas 5 chamadas;

b) decorra mais de meia hora entre duas chamadas sucessivas.

23. Uma loja vende, em média, 4 unidades de um certo por mês. A loja renova seu estoque desse artigo uma vez por mês. Supondo que as vendas ocorrem aleatória e independentemente, Até que número deve a loja completar seu estoque para reduzir a probabilidade de esgotamento do estoque para menos que 1/100?

24. Em um problema de hereditariedade, um geneticista não está seguro se as duas alternativas A e a de uma característica apresentam-se na razão 1:1 ou 3:1. Então, decide conduzir um experimento para basear sua conclusão. Se ele deseja que a diferença entre os valores esperados da alternativa A sob as duas hipóteses (1:1 e 3:1) seja mínima, qual é o tamanho mínimo da amostra que deve escolher para seu experimento?

25. Sejam X e Y variáveis aleatórias independentes. Demonstre que $E(XY) = E(X).E(Y)$.

26. Considere o experimento de plantio de três sementes de um lote de sementes com 80% de poder germinativo.

a) Especifique o espaço de amostra do experimento.

b) Especifique o evento "as duas primeiras sementes germinam" e determine a sua probabilidade.

c) Estabeleça as expressões das variáveis aleatórias:

X = número de sementes que germinam;

$$Y = \begin{cases} 1, & \text{se 2 ou mais sementes germinam} \\ 0, & \text{caso contrario} \end{cases}$$

d) Determine as funções de probabilidade de X e de Y.

e) Determine a função de probabilidade conjunta de X e Y.

f) Determine: i) $E(X)$, $E(Y)$ e $E(XY)$; ii) Cov(X,Y); iii) Corr(X,Y).

27. Sejam X e Y variáveis que exprimem as seguintes características referentes à produtividade de ovelhas de um rebanho da raça Corriedale:

X = cordeiros nascidos por ovelha parida;

Y = cordeiros desmamados por ovelha parida,

Suponha que as distribuições de probabilidade conjunta e marginais de X e Y sejam dadas na seguinte tabela:

Y	X		P[Y=y]
	1	2	
0	0,15	0,05	0,20
1	0,50	0,20	0,70
2	0,00	0,10	0,10
P[X=x]	0,65	0,35	1,00

a) Determine: i) $E(X)$; ii) $E(Y)$; iii) Var(X); iv) Var(Y); v) $E(X+Y)$; vi) $E(XY)$; vii) Cov(X,Y); viii) ρ_{XY}.

b) São as variáveis X e Y independentes? Justifique a resposta.

28. Considere o seguinte experimento aleatório: "Germinação de 3 sementes e observação do resultado do teste de germinação". (Represente a germinação de uma semente por G e a não germinação por g.)

a) Especifique o espaço básico S desse experimento.

b) Especifique as variáveis aleatórias X e Y que exprimem as duas seguintes características dos eventos elementares desse espaço básico S:

X: Ocorrência de germinação na primeira semente;

Y: Número de sementes que germinam.

c) Supondo que a probabilidade de uma semente germinar é 0,8, igual para as 3 sementes, determine as distribuições de probabilidade dessas 2 variáveis aleatórias.

d) Estabeleça a função de probabilidade conjunta e as funções de probabilidade marginais das variáveis aleatórias X e Y, em uma tabela de dupla entrada.

e) Determine: i) $E(X)$; ii) $E(Y)$; iii) Var(X); iv) Var(Y); v) $E(X+Y)$; vi) $E(XY)$; vii) Cov(X,Y); viii) ρ_{XY}.

f) Verifique se as variáveis aleatórias X e Y são estatisticamente independentes.

29. Se o resultado "cara" é um sucesso no lançamento de uma moeda, a obtenção da face 6 é um sucesso no lançamento de um dado e um ás é um sucesso na extração de uma carta de um baralho de 52 cartas, determine a média e o desvio padrão do número total de sucessos em um jogo com cada um dos seguintes esquemas alternativos, na suposição de que a moeda, o dado e o baralho não sejam viciados e os correspondentes processos de lançamento ou extração sejam não tendenciosos:

a) lançamento de uma moeda, rolamento de um dado e extração de uma carta de um baralho;

b) lançamento de uma moeda três vezes, rolamento de um dado duas vezes e extração de uma carta de um baralho.

30. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando as letras V ou F entre parênteses, respectivamente. Se a sentença for falsa, explique porque?

( ) A probabilidade de escolher uma amostra aleatória de três pessoas em que a primeira responde "sim" e a última pessoa responde "não" de uma população em que P["sim"]=0,7 é 0,7×0,7×0,3.

( ) Se um casal particular é normal, então a probabilidade de que seu primo filho seja uma menina é aproximadamente 0,5.

( ) A probabilidade de escolher uma amostra aleatória de três pessoas em que exatamente duas respondem "sim" de uma população em que P["sim"]=0,6 é 0,6×0,6×0,4.

( ) A probabilidade de obter duas faces "três" no lançamento de dois dados é 1/6+1/6 = 1/3.

( ) Em uma distribuição de probabilidade discreta, a área total compreendida entre a curva que representa a distribuição e o eixo horizontal é igual a um.

( ) Em uma distribuição de probabilidade discreta, o comprimento de uma linha vertical em um certo valor pode ser interpretado como a probabilidade de tal valor resultar em uma amostragem aleatória.

( ) Em uma distribuição de probabilidade, a probabilidade de uma variável aleatória assumir um valor particular é zero.

( ) Variáveis aleatórias têm sempre valores numéricos.

( ) Variáveis nominais não podem ser diretamente modeladas por uma distribuição de probabilidade.

( ) O valor esperado de uma distribuição de probabilidade pode ser interpretado como o centro de gravidade.

( ) A variância de uma distribuição de probabilidade pode ser definida simbolicamente como $E[X-E(X)]^2$ e também pode ser expressa na forma $E(X^2)-[E(X)]^2$.

( ) A variância de uma distribuição de probabilidade é uma medida de posição e o valor esperado indica a dispersão.

( ) Se duas distribuições de probabilidade têm a mesma variância, então seus valores esperados também são iguais.

( ) Em um experimento binomial, os resultados correspondem a duas classes mutuamente exclusivas.

( ) Em um experimento binomial com n ensaios de Bernoulli, a variável aleatória pode assumir qualquer dos n valores.

( ) Distribuições binomiais não são simétricas, exceto quando p=1-p.

( ) Sendo a distribuição binomial discreta, seu valor esperado é um valor inteiro.

( ) Se o parâmetro p da distribuição binomial é 0,6, a probabilidade de exatamente 60 sucessos em 120 ensaios é maior que a probabilidade de 72 sucessos em 120 ensaios.

( ) Se A e B são eventos mutuamente exclusivos, então $P[A \cup B] = P[A] \times P[B]$.

( ) A variância de distribuições discretas pode ser calculada pela fórmula V(X) = np(1-p).

( ) A regra da adição de probabilidades aplica-se apenas para eventos mutuamente exclusivos.

( ) A distribuição binomial é um exemplo de distribuição de probabilidade contínua.

( ) Para a determinação de probabilidades em uma distribuição binomial o número de ensaios n e o parâmetro p devem ser conhecidos.

( ) Em uma distribuição de Poisson, $E(X)=np$ e $Var(X)=np(1-p)$.

( ) Os valores de uma variável aleatória com distribuição de Poisson são discretos, contáveis.

( ) Dado que $E(X)$ é usualmente pequeno para uma distribuição de Poisson, um número relativamente grande de intervalos é necessário para estimar o parâmetro λ efetivamente.

( ) Uma característica típica da distribuição de Poisson é que o valor esperado é maior que a variância.

( ) A distribuição de Poisson é algumas vezes chamada "distribuição de eventos raros"; portanto, ela é raramente encontrada na pesquisa experimental.

( ) A forma da distribuição de probabilidade de Poisson é simétrica em relação a seu valor esperado.

( ) Há uma diferente distribuição de Poisson para cada par de valores de λ e n.

( ) A distribuição de Poisson pode ser sempre usada para aproximar probabilidades de uma distribuição binomial.

( ) Em amostragens de uma distribuição de Poisson, dado que λ é usualmente pequeno, valores pequenos de X são muito mais prováveis do que valores grandes.

( ) Se X_1 e X_2 são variáveis aleatórias com a mesma distribuição de probabilidade, então $E(X_1-X_2)=0$ e $Var(X_1-X_2)=0$.

( ) Se duas populações têm a mesma média, então elas também têm a mesma variância.

5 VARIÁVEL ALEATÓRIA CONTÍNUA E DENSIDADE DE PROBABILIDADE

Conteúdo

5.1 Introdução....111
5.2 Densidade de Probabilidade....112
5.3 Distribuição Normal....116
5.3.1 Introdução....116
5.3.2 Cálculo de probabilidades para variável aleatória normal....118
5.4 Aproximação da Distribuição Binomial pela Distribuição Normal....122
5.5 Distribuição Normal Bivariada....124
5.6 Combinação Linear de Variáveis Aleatórias Normais....126
Casos particulares....126
5.7 Distribuição da Média de Variáveis Aleatórias....127
5.8 Exercícios....128

5.1 Introdução

Recorde-se que uma variável aleatória é uma função real definida no espaço básico à qual é associada uma função distribuição de probabilidade. A variável aleatória é discreta se essa função distribui a probabilidade total 1 a um número finito ou infinito contável de seus valores.

Uma distribuição de probabilidade de variável aleatória discreta atribui probabilidades positivas a pontos isolados da reta. Dessa forma, probabilidades de intervalos podem ser determinadas pela soma das probabilidades dos pontos nele incluídos.

No caso de espaço básico não contável e variável aleatória não discreta, a definição de probabilidades envolve alguma complicação. Essa situação é ilustrada no exemplo que segue.

Exemplo 5.1. Considere-se a possibilidade de acidente com um caminhão que transporta uma carga de soja de uma propriedade rural até um armazém, distante 100 km, e suponha que há interesse na probabilidade de que ocorra um acidente em um certo trecho da estrada.

O espaço básico para este experimento consiste de um "continuum" de pontos no intervalo de 0 a 100. Suponha-se que, se um acidente ocorre, a probabilidade de que ele ocorra em um trecho de k quilômetros é k/100. Essa atribuição de probabilidades é consistente com os axiomas 1 e 2 do conceito de probabilidade (**Seção 3.9**), visto que as probabilidades k/100 são não negativas e P(S) = 100/100 = 1. A atribuição de probabilidades definidas aplica-se a intervalos do

segmento de reta compreendidos entre 0 e 100. Pelo axioma 3, pode-se, também, obter probabilidades para a reunião de qualquer seqüência finita ou infinita contável de intervalos mutuamente exclusivos. Por exemplo, a probabilidade de que um acidente ocorra em qualquer de dois intervalos que não se sobrepõem, com extensões de k_1 e k_2 km, é: $\frac{k_1 + k_2}{100}$ e, de modo mais geral, a probabilidade de que um acidente ocorra em algum intervalo de uma seqüência de intervalos que não se sobrepõe, de comprimentos k_1, k_2,..., é $\frac{k_1 + k_2 + \ldots}{100}$.

Probabilidades também podem ser calculadas para intervalos que se sobrepõe. Por outro lado, como a interseção de dois intervalos é um intervalo e o complemento de um intervalo é um intervalo ou a união de dois intervalos, pode-se calcular probabilidades para qualquer subconjunto do espaço básico que possa ser obtido por uniões ou interseções de um número finito ou infinito contável de intervalos, ou por complementos.

Dessa forma, nessa extensão do conceito de probabilidade ao caso contínuo, aplicam-se os três axiomas do conceito de probabilidade. Entretanto, em tais extensões, em geral, deve-se excluir da definição de "evento" todos os subconjuntos do espaço básico que não podem ser obtidos pela formação de reuniões ou interseções de números finitos ou infinitos contáveis de intervalos, ou por complementações. Demonstra-se que tais eventos existem sobre a linha reta. Entretanto, na prática, esse fato não tem conseqüências importantes, pois, simplesmente, não se atribui probabilidades a tais tipos de eventos sem interesse.

Por outro lado, considerando, ainda, o **Exemplo 5.1**, observe-se que a probabilidade de um acidente ocorrer em um intervalo de amplitude muito pequena, seja 1 cm, é apenas 0,000.000.1. Na medida que a amplitude do intervalo aproxima-se de zero, a probabilidade de um acidente no intervalo também se aproxima-se de zero. Por essa razão, atribui-se probabilidade zero a pontos individuais. Isto não significa que os eventos correspondentes não podem ocorrer, já que, quando um acidente ocorre em um trecho de 100 km de uma rodovia, ele tem de ocorrer em algum ponto, embora cada ponto tenha probabilidade zero.

5.2 Densidade de Probabilidade

O modo de assinalar probabilidades no **Exemplo 5.1** é muito particular, mas é de natureza semelhante ao modo em que se assinala probabilidades no caso de espaço básico discreto equiprovável, tal como seleção aleatória de uma garrafa de vinho de uma pilha. Para a melhor compreensão do processo geral de associação de probabilidades com variáveis aleatórias contínuas, suponha-se que uma indústria de vinho está interessada na quantidade real de vinho que é preenchida por uma máquina em garrafas de 630 ml. A quantidade de vinho preenchida nas garrafas será variável e poderá assumir qualquer valor de um certo intervalo em torno de 630 ml; portanto, é uma variável aleatória contínua. Entretanto, se o volume preenchido é registrado com precisão de mililitro, a correspondente variável aleatória será discreta, com uma distribuição de probabilidade que pode ser representada em um histograma, como o da **Figura 5.1.a**, em que as probabilidades correspondem às áreas dos retângulos. Se os volumes preenchidos são registrados com a precisão mais elevada de 1/5 de mililitro, a correspondente variável aleatória ainda será discreta, com uma distribuição de probabilidade como a representada no histograma da **Figura 5.1.b**.

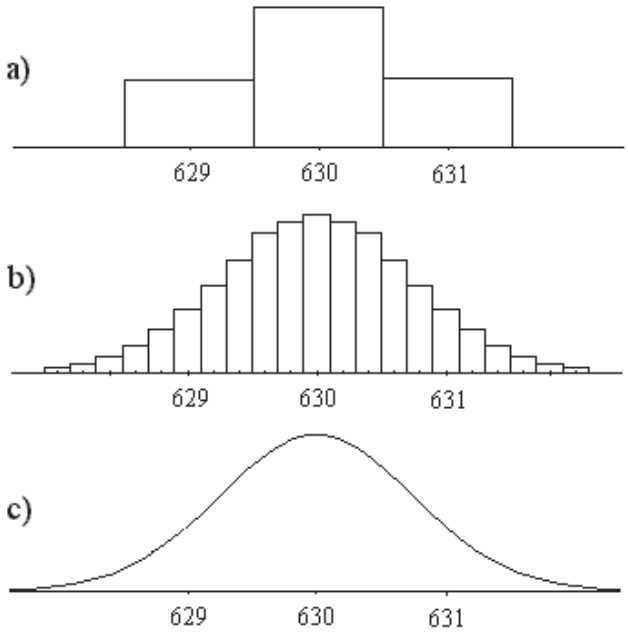


Figura 5.1. Funções de probabilidade das variáveis aleatórias do **Exemplo 5.1**, com precisão crescente da escala de mensuração.

É aparente que, se os volumes preenchidos são registrados com precisão crescente (de um milésimo de milímetro, de um milionésimo de milímetro, e assim por diante), os histogramas das distribuições de probabilidades das correspondentes variáveis aleatórias discretas se aproximarão de uma curva contínua, como a da **Figura 5.1.c**. Para cada um desses níveis de precisão, a probabilidade de que a quantidade de vinho preenchida situe-se em um intervalo específico é a soma das áreas dos retângulos assentados sobre o intervalo. Essa probabilidade aproxima-se da área sob a curva no correspondente intervalo quando o nível de precisão das medidas cresce. Dessa forma, probabilidades correspondentes à variável aleatória contínua que exprime o volume de vinho preenchido pela máquina em garrafas de 630 ml correspondem às áreas sob essa curva.

De modo geral, a definição de probabilidade para uma variável aleatória contínua X presume a existência de uma **função distribuição de probabilidade** f(x) que atribui como probabilidade associada com qualquer intervalo de valores do espaço de X a área sob a curva f(x), compreendida no correspondente intervalo. Em outras palavras, a probabilidade de que uma variável aleatória contínua X assuma um valor em um intervalo (a, b) é provida pela área sob a curva f(x), que é obtida pela **integração** da função f(x) no intervalo (a, b). Esse argumento é o fundamento para o conceito que segue.

Uma função f(x), definida no conjunto dos números reais, é uma **função distribuição de probabilidade** de uma variável aleatória contínua X se e somente se:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

para quaisquer números reais a e b (a≤b).

A função distribuição de probabilidade de uma variável aleatória contínua é, usualmente, designada **função densidade de probabilidade**, ou **densidade de probabilidade**, ou **função densidade**, ou, ainda mais simplesmente, **densidade**.

Observe-se que, agora, o valor da função densidade de probabilidade de X em um ponto c, ou seja, f(c), não corresponde à probabilidade P[X=c], como no caso de variável aleatória discreta. De fato, para qualquer variável aleatória contínua X, P[X=c] = 0 para qualquer valor c. Assim, a função de densidade f(x) associa probabilidades a intervalos; não a valores reais específicos.

Em decorrência dessa propriedade, as probabilidades associadas a uma variável aleatória contínua não se alteram pela alteração dos valores da correspondente função densidade de probabilidade para alguns valores particulares da variável aleatória. Dessa forma, a probabilidade de um intervalo é a mesma sejam os seus extremos incluídos ou excluídos, ou seja:

$P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = P(a < X < b)$.

Uma função densidade de probabilidade satisfaz às seguintes propriedades, análogas àquelas da função massa de probabilidade, que seguem dos axiomas da probabilidade:

1) $f(x) > 0$, para $-\infty < x < \infty$;

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

Reciprocamente, qualquer função real f(x) que satisfaça a essas duas propriedades é uma função densidade de probabilidade.

Exemplo 5.2. A função densidade para a variável aleatória contínua do **Exemplo 5.1**. é:

$$f(x) = \begin{cases} 1/100, & 0 \leq x \leq 100 \\ 0, & \text{caso contrario} \end{cases}$$

Essa função densidade é representada na **Figura 5.2**.

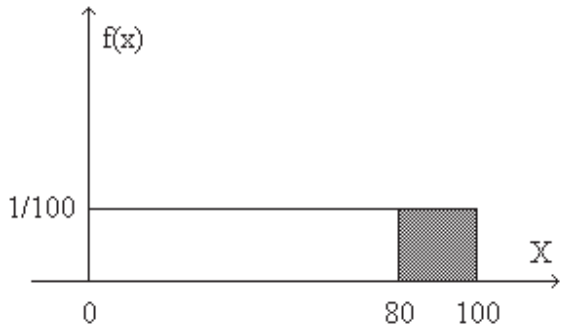


Figura 5.2. Representação geométrica da função densidade da variável aleatória do **Exemplo 5.1**.

A probabilidade de um acidente entre os quilômetros 80 e 100 é dada pela área do retângulo sobre o intervalo (80,100), ou seja: $(100\text{-}80).\frac{1}{100} = \frac{1}{5}$. Essa probabilidade (área) é a **integral definida** da função f(x) no intervalo (80,100), ou seja:

$$\int_{80}^{100} \frac{1}{100} dx = \frac{x}{100}\Big|_{80}^{100} = \frac{100-80}{100} = \frac{1}{5}$$

Mais comumente, o cálculo de probabilidades para variáveis aleatórias contínuas não pode ser efetuado, simplesmente, por processo geométrico, como o ilustrado para o exemplo. Em tais situações, o conhecimento de Cálculo Integral é imprescindível.

O conceito de valor esperado é estendido para variável aleatória contínua pela transformação de soma em integração:

A **média** (**valor esperado**) e a **variância** de uma variável aleatória contínua X são definidos como segue:

$$E(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx \text{ , e}$$

$$\text{Var}(X) = \int_{-\infty}^{+\infty} \left[x - E(x)\right]^2 f(x)\, dx .$$

As propriedades do valor esperado e da variância listada para variável aleatória discreta também se estendem para variável aleatória contínua:

1) Se X é uma variável aleatória contínua e *a* e *b* são constantes, então:

$E(aX+b) = a\,E(X) + b$.

$\text{Var}(aX+b) = a^2\,\text{Var}(X)$.

2) Se $g_i(X)$, i=1,2,...,n, são funções de uma variável aleatória contínua X e c_i, i=1,2,...,n, são constantes, então:

$$E[\sum_{i=1}^{n} c_i\, g_i(X)] = \sum_{i=1}^{n} c_i\, E[g_i(X)].$$

Os conceitos e propriedades referentes à distribuição de duas ou mais variáveis aleatórias também se estendem ao caso contínuo.

Uma função f(x,y) definida no plano R^2_{XY} é uma **função densidade de probabilidade conjunta** de duas variáveis aleatórias contínuas X e Y se e somente se:

$$P[(x,y) \in A] = \iint_A f(x,y)\, dx\, dy$$

para qualquer região (evento) A do plano R^2_{XY}.

Uma função densidade bivariada f(x,y) é representada, geometricamente, por uma superfície em um espaço de três dimensões. O volume sob essa curva acima do plano XY é igual a um. A probabilidade de uma região A do plano XY, ou seja, a probabilidade de que um valor da variável aleatória bidimensional (X,Y) situe-se na região A, é o volume do sólido reto sob a superfície f(x,y) com base em A. Essa probabilidade (volume) é a integral da função f(x,y) na área A.

Uma função bivariada f(x,y) é uma **função densidade de probabilidade conjunta** de duas variáveis aleatórias contínuas X e Y se seus valores satisfazem às duas propriedades que seguem, derivadas dos postulados do conceito de probabilidade:

1) $f(x,y) \geq 0$, $-\infty<x<+\infty$; $-\infty<y<+\infty$.

2) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\,dxdy = 1$.

Todos os conceitos e propriedades da distribuição de n variáveis aleatórias discretas estendem-se para a situação de n variáveis aleatórias contínuas.

Ao contrário de distribuições discretas, tais como binomial, binomial negativa e hipergeométrica, as distribuições de variáveis aleatórias contínuas, usualmente, não podem ser derivadas através de argumentos probabilistas simples. Em vez disso, para uma variável aleatória contínua, deve ser procedida uma escolha judiciosa da função densidade de probabilidade, com base em conhecimentos anteriores e em informações providas por dados disponíveis. Felizmente, são disponíveis algumas famílias paramétricas gerais de funções densidades de probabilidade que se têm revelado apropriadas ou boas aproximações para uma ampla gama de situações experimentais. Entre essas, salienta-se a função de densidade de probabilidade normal, da qual se tratará a seguir.

5.3 Distribuição Normal

5.3.1 Introdução

A distribuição normal é a mais importante e amplamente utilizada distribuição em estatística aplicada. As distribuições de freqüências de observações de muitas medidas físicas de fenômenos naturais têm a aparência muito próxima da distribuição normal. Por exemplo, as medidas de características biométricas de seres vivos, como peso e altura, e erros de mensuração em experimentos científicos. As representações gráficas de freqüências relativas dessas características, através de histogramas, aproximam-se da **curva normal** da **Figura 5.1.c**. Mas há outra razão mais fundamental para a importância da distribuição normal em estatística. Uma propriedade teórica da média da amostra, derivada do **teorema central do limite**, permite o uso da distribuição normal como uma aproximação dos modelos probabilistas para muitos experimentos cujas observações básicas constituem realizações de variáveis aleatórias discretas ou contínuas não normais.

A distribuição normal foi pela primeira vez pesquisada no século dezoito, quando cientistas observaram um extraordinário grau de regularidade em erros de mensuração. Eles descobriram que os padrões, ou distribuições, dos erros de medidas de fenômenos que observavam podiam ser muito bem aproximados por uma curva contínua que referiram como "curva normal dos erros". As propriedades matemáticas da **curva normal** foram estudadas por Abraham de Moivre (1667-1745), Pierre Laplace (1749-1827) e Karl Gauss (1777-1855).

Uma variável aleatória X tem **distribuição normal** e é referida como uma **variável aleatória normal**, se sua densidade de probabilidade é expressa por:

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

onde μ e σ^2 são parâmetros que satisfazem $-\infty<\mu<+\infty$ e $\sigma> 0$, *e* denota a base do sistema logarítmico neperiano, aproximadamente igual a 2,71828, e π é a conhecida constante matemática, aproximadamente igual a 3,14159.

O intervalo da variável aleatória normal é $-\infty<x<+\infty$. Isso pode parecer uma restrição à utilidade da distribuição normal, já que os erros de mensuração são, obviamente, demarcados em algum intervalo e a maioria das variáveis de interesse na natureza, em particular as variáveis biométricas, são não negativas. O argumento para essa aparente contradição é que, de fato, a distribuição normal, como qualquer modelo probabilista e, em geral, qualquer modelo matemático, é apenas uma aproximação às situações reais modeladas. No caso de erros de medida, por exemplo, o modelo probabilista normal pressupõe, tacitamente, que a probabilidade de obter erros de medida muito grandes é muito pequena e a probabilidade de obter erros de medida negativos é, também, muito pequena. O modelo probabilista normal aproxima a situação real do fenômeno físico que requer que essas probabilidades sejam nulas, assinalando probabilidades pequenas a esses eventos. Essa aproximação é semelhante à aproximação que se obtém quando se pressupõe que observações efetuadas através de instrumentos de medida discretos constituem realizações de variáveis aleatórias contínuas.

Dado que uma função densidade normal particular é especificada por um par de valores de μ e σ^2, é comum denotar que uma variável aleatória X tem distribuição normal com média μ e variância σ^2 por $X\sim N(\mu,\sigma^2)$.

A distribuição normal particular com média $\mu=0$ e variância $\sigma^2=1$ é denominada **distribuição normal padrão**. É comum designar a variável aleatória padrão por Z, de modo que: $Z\sim N(0,1)$.

Claramente, $f(x)>0$ para qualquer valor real x de X. Entretanto, conhecimentos de Cálculo Integral são necessários para demonstrar que $\int_{-\infty}^{+\infty} f(x,y)\,dx = 1$ para quaisquer valores de μ e σ^2, e que $E(X)=\mu$ e $Var(X)=\sigma^2$, o que justifica o uso dos símbolos μ e σ^2 para representarem os dois parâmetros da função densidade normal.

A **Figura 5.3** apresenta os gráficos da função densidade normal f(x) para: a) diferentes valores de μ e σ^2 fixo, e b) diferentes valores de σ^2 e μ fixo. A **Figura 5.3.a** mostra que uma alteração da média de μ_1 para μ_2 apenas desloca a curva μ_1-μ_2 unidades ao longo do eixo x. A forma da curva permanece inalterada. Por outro lado, um diferente valor do desvio padrão resulta em uma diferente altura máxima da curva e em uma diferente quantidade da área em qualquer intervalo fixo centrado em μ (**Figura 5.3.b**). Com a alteração de σ^2 apenas, a posição do centro da curva não muda.

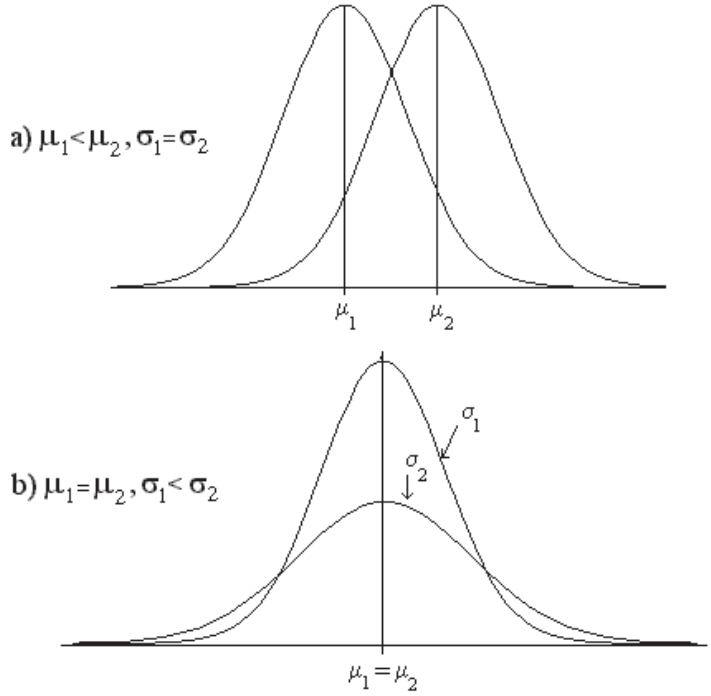


Figura 5.3. Gráficos de funções de densidades normais com: **a)** diferentes médias e iguais variâncias; **b)** iguais médias e diferentes variâncias.

O gráfico da função densidade de probabilidade normal é simétrico em relação à média μ e tem a forma de sino, com ponto de simetria em μ, que coincide com a moda e a mediana da distribuição. Gráficos para valores grandes de σ são espalhados em relação à μ, enquanto gráficos para valores pequenos de σ têm um pico elevado no ponto μ e a maioria da área sob o gráfico muito próxima de μ. Assim, valores grandes de σ implicam maior probabilidade para valores de X afastados de μ do que valores pequenos de σ. Intervalos de mesma amplitude têm probabilidade mais elevada quanto mais próximo da média μ se situam. Intervalos centrados em μ com extremos afastados 1, 2 e 3 desvios padrões de μ têm probabilidades, respectivamente, 0,683, 0,954 e 0,997, ou seja:

$P[\mu-\sigma<X<\mu+\sigma] = 0,683;$

$P[\mu-2\sigma<X<\mu+2\sigma] = 0,954;$

$P[\mu-3\sigma<X<\mu+3\sigma] = 0,997.$

Assim, áreas sob a curva externas a intervalos cujos extremos se afastam de μ mais de três desvios padrões são muito pequenas.

5.3.2 Cálculo de probabilidades para variável aleatória normal

A probabilidade de um evento quando X tem uma distribuição normal, por exemplo, $P[a<X<b]$, é provida pela integral definida da função de densidade no intervalo (a,b), ou seja:

$$P[a<X<b] = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx\,.$$

Essa integral não pode ser calculada pelas técnicas de integração usuais. Ademais, seu valor depende dos parâmetros μ e σ^2, o que demandaria a avaliação da integral para cada par de valores (μ,σ^2). Felizmente, entretanto, demonstra-se que a transformação de variável: $Z = \frac{x-\mu}{\sigma}$ conduz o cálculo de $P[a<X<b]$ a uma integração definida que independe de μ e σ^2. Essa transformação desloca a média de μ para o ponto zero e altera a escala pela divisão por σ, tornando o desvio padrão igual a um. De fato,

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

de modo que os seguintes eventos são equivalentes:

$$\{s \mid a<X(s)<b\} = \{s \mid a<Z(s)\sigma+\mu<b\} = \left\{s \mid \frac{a-\mu}{\sigma} < Z(s) < \frac{b-\mu}{\sigma}\right\}.$$

Então, substituindo $x = \sigma z+\mu$ na expressão que define $P[a<X<b]$ como a integral definida da função densidade da variável aleatória $X\sim N(\mu,\sigma^2)$, obtém-se:

$$P[a<X<b] = \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz\,.$$

Pode-se reconhecer que o integrando é, agora, a função de densidade da variável aleatória normal padrão Z, com média zero e variância 1, ou seja: $Z\sim N(0,1)$, e que:

$$P[a<X<b] = P\left[\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma}\right].$$

Desse modo, o cálculo de probabilidade para uma variável aleatória $X\sim N(\mu,\sigma^2)$, para um par de valores específicos de (μ,σ^2), reduz-se ao cálculo de probabilidade para a variável aleatória $Z\sim N(0,1)$, que independe de μ e σ^2.

Probabilidades para intervalos da variável aleatória Z são disponíveis em tabelas especiais. Uma dessas tabelas é a Tabela II, do Apêndice.

Em inferência estatística, freqüentemente, necessita-se de valores da estatística Z que demarcam pequenas áreas nas caudas da curva normal padrão. Denote-se por z_α o valor da estatística que limita à sua direita a área α sob a curva normal padrão (**Figura 5.4**).

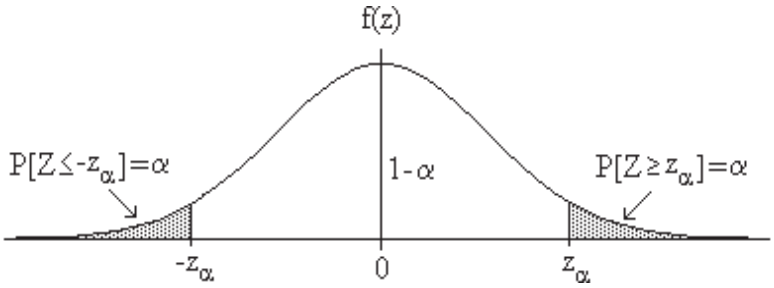


Figura 5.4. Gráfico da distribuição normal padrão, com a indicação do 100(1-α)-ésimo percentil.

Como α% da área sob a curva normal padrão situa-se à direita de z, (1-α)% da área fica à esquerda de z_α . Assim, z_α é o **100(1-α)-ésimo percentil** da distribuição normal padrão, também designado ponto **α-percentual superior** ou **desvio normal correspondente a** α. Por simetria, a área sob a curva normal padrão à esquerda de $-z_\alpha$ também é α. Os valores z_α são, usualmente, referidos como **valores críticos** da distribuição normal padrão.

A Tabela II contém os valores z_α para diversos valores de α. Através de consulta à Tabela II, pode-se obter o valor de $P[Z > z_\alpha]$ para um dado α, ou o valor de α para uma dada $P[Z > z_\alpha]$. Essa tabela permite a determinação de probabilidades para variáveis aleatórias normais mais genéricas $X \sim N(\mu, \sigma^2)$, para qualquer intervalo infinito da forma (b, ∞), já que:

$$P[X>b] = \int_b^\infty \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)}{2\sigma^2}}\, dx$$

$$= \int_{z_\alpha}^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\, dz .= P[X> z_\alpha],$$

onde $z_\alpha = \frac{b-\mu}{\sigma}$. Essa probabilidade corresponde à área da cauda à direita sob a curva f(x) da **Figura 5.4**.

Assim, para obter essa probabilidade, subtrai-se a média μ de *b* e divide-se o resultado pelo desvio padrão σ, ou seja, calcula-se $z = (b-\mu)/\sigma$. Entra-se com esse valor na Tabela II e lê-se a probabilidade na correspondente célula. Reciprocamente, dada a probabilidade de um intervalo da forma (b,∞) para uma variável aleatória $X \sim N(\mu,\sigma^2)$, pode-se obter o valor de *b*, também pela inspeção da Tabela II.

Exemplo 5.3. P[X>8], X~N(2,9).

Subtraindo a média μ=2 de ambos os membros da desigualdade X>8 e, logo após, dividindo ambos os membros por $\sigma=\sqrt{9}=3$, obtém-se, sucessivamente:

$$P[X>8] = P[X-2>8-2] = P\left[\frac{x-2}{3} > \frac{8-2}{3}\right]$$

$= P[Z \geq 2]$.

Entrando na Tabela II com z=2, obtém-se: $P[Z \geq 2] = 0,0228$ (valor de a correspondente a $z_\alpha = 2$). Portanto, $P[X>8] = 0,0228$.

A probabilidade de um intervalo infinito da forma $(-\infty, b)$ pode ser obtida pela relação:

$P[X<b] = 1-P[X \geq b]$.

Exemplo 5.4. $P[X<3]$, $X \sim N(2,9)$.

Pela relação anterior:

$P[X<3] = 1-P[X>3]$.

Mas

$$P[X>3] = P\left[\frac{x-2}{3} > \frac{3-2}{3}\right] = P[Z>0,33] = 0,3707,$$

onde 0,3707 é o valor da Tabela II para $z_\alpha = 0,33$. Logo,

$P[X<3] = 1-0,3707 = 0,6293$.

A Tabela II não inclui valores negativos de z_α, já que, pela simetria da distribuição normal padrão em relação à zero, probabilidades de intervalos infinitos com um extremo finito negativo podem ser obtidas de probabilidades de intervalos infinitos com extremo finito positivo, pelas relações:

$P[Z<-z_\alpha] = P[Z>z_\alpha]$ e

$P[Z>-z_\alpha] = 1-P[Z<-z_\alpha] = 1-P[Z>z_\alpha]$.

Essas relações podem ser verificadas com o auxílio da **Figura 5.4**.

Exemplo 5.4.

a) $P[Z<-1,5] = P[Z>1,5] = 0,0668$.

b) $P[Z>-1,5] = 1-P[Z \geq 1,5] = 1-0,0668 = 0,9332$.

A probabilidade de um intervalo finito, da forma $P[a<X<b]$, pode ser obtida por uma extensão simples dos argumentos anteriores. Esta probabilidade corresponde à área sombreada na **Figura 5.5**. Ela pode ser calculada através da relação:

$$P[a<X<b] = P[X>a] - P[X>b],$$

onde ambas as probabilidades do segundo membro podem ser calculadas com o auxílio da Tabela II, segundo procedimento indicado anteriormente.

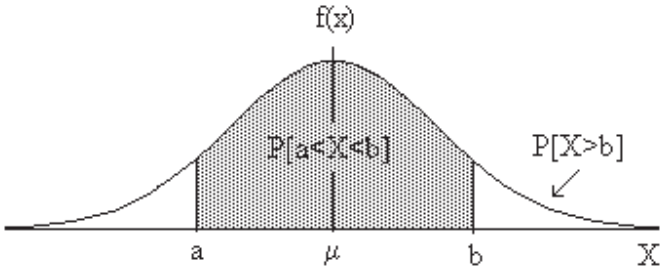


Figura 5.5. Gráfico da distribuição de uma variável aleatória $X \sim N(\mu,\sigma^2)$, com a indicação das áreas correspondentes a $P[a<X<b]$, $P[X>a]$.

Exemplo 5.5. $P[3<X<8]$, $X \sim N(2,9]$.

De acordo com a última relação e usando resultados de exemplos anteriores, obtém-se:

$$P[3<X<8] = P[X>3] - P[X>8] =$$
$$= 0,6293 - 0,0228 = 0,6065.$$

5.4 Aproximação da Distribuição Binomial pela Distribuição Normal

A distribuição binomial b(n,p) é a distribuição do número X de sucessos em n experimentos independentes de Bernoulli com probabilidade de sucesso comum p. Demonstra-se que a distribuição normal é uma boa aproximação para a distribuição binomial com parâmetros n elevado e p não muito próximo de 0 ou 1.

Recorde-se que a média e a variância da distribuição b(n,p) são: $\mu=np$ e $\sigma^2=np(1-p)$. Conforme foi visto na **Seção 4.3.8**, quando o parâmetro n é grande mas np é moderado, ou seja, quando $\mu=np$ ou $\sigma^2=np(1-p)$ é moderado, a distribuição de Poisson constitui uma boa aproximação da distribuição binomial. A aproximação normal é útil na situação mais típica em que n é elevado e p não é muito próximo de 0 ou 1. Essa propriedade permite tratar uma variável aleatória X com distribuição b(n,p) como se ela tivesse distribuição N(np, np(1-p)).

Nessas circunstâncias, a determinação da probabilidade de que $X \sim b(n,p)$ assuma valores em um intervalo $[a,b]$, ou seja:

$$P[a \le X \le b] = \sum_{x=a}^{b} C_n^x P^x (1-p)^{n-x}$$

pode ser procedida através do cálculo da probabilidade de que $X \sim N(np, np(1-p))$ esteja no intervalo $[a,b]$. Essa probabilidade pode ser determinada através da padronização de X, ou seja, através da variável aleatória:

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

que tem distribuição N(0,1), de modo que (aproximadamente):

$$P[a \le X \le b] \approx P\left[\frac{a - np}{\sqrt{np(1-p)}} \le Z \le \frac{b - np}{\sqrt{np(1 - p)}}\right]$$

Essa probabilidade pode ser calculada a partir da tabela de probabilidades da distribuição normal padrão (Tabela II), conforme visto na **Seção 5.3.2**. Como uma regra geral, esta aproximação é usualmente satisfatória quando np e n(1-p) são ambos maiores do que 15.

A aproximação lograda pela distribuição normal é ilustrada na **Figura 5.6** que mostra o histograma da distribuição b(15;0,4) e a curva da distribuição normal que a aproxima, com $\mu = 15\times0,4 = 6,0$ e $\sigma^2 = 15\times0,4\times0,6 = 3,6$. Essa ilustração mostra que a aproximação é razoável, mesmo para um valor pequeno de n, ou seja 5, que implica np=6, muito menor do que o menor valor de np indicado pela regra prática mencionada anteriormente.

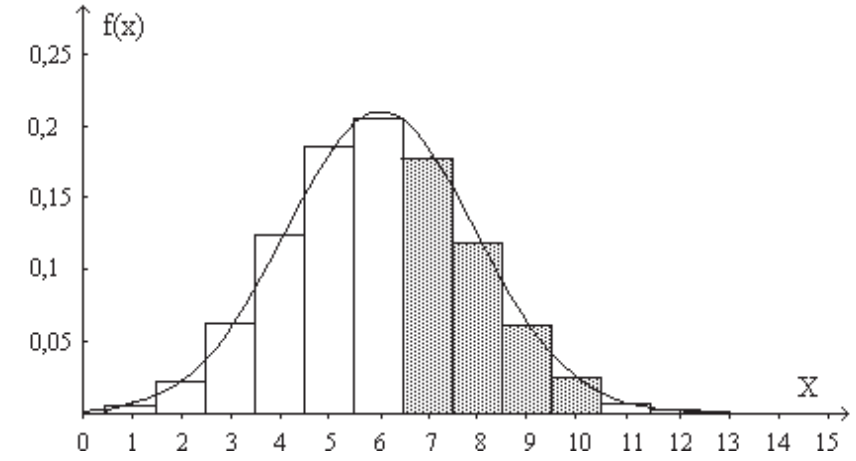


Figura 5.6. Histograma da distribuição b(15;0,4) e curva da distribuição normal que a aproxima.

Para uma ilustração mais concreta, considere-se a probabilidade de que X∼b(15;0,4) situe-se no intervalo [7,10]. O valor exato, obtido da tabela da distribuição binomial (Tabela I), é:

$$P[7\leq X\leq 10] = 0,991-0,610 = 0,381.$$

Por outro lado, a distribuição N(6; 3,6) que aproxima a distribuição b(15;0,4), dá:

$$P[7<X<10] \approx P\left[\frac{7-6}{1,9} \leq Z \leq \frac{10-6}{1,9}\right]$$

$$= P[0,526 \leq Z \leq 2,105]$$

$$= 0,982 - 0,700 = 0,282.$$

Essa aproximação pode ser melhorada através de um ajustamento no processo de cálculo. De fato, observe-se que a probabilidade $P[7\leq X\leq 10] = 0,381$ é a área das barras sombreadas na **Figura 5.6**. Como essas barras estendem-se de 6,5 a 10,5, uma melhor aproximação da área sob a curva é obtida se se considera o intervalo (6,5; 10,5). Com esse ajustamento, tem-se:

$$P[7<X<10] \approx P\left[\frac{6,5-6}{1,9} \leq Z \leq \frac{10,5-6}{1,9}\right] =$$

$$= P[0,263 \leq Z \leq 2,368] =$$

$$= 0,991 - 0,604 = 0,387.$$

Esse resultado é consideravelmente mais próximo do valor correto, 0,381, apesar de n ser pequeno.

Como através desse processo se está aproximando uma distribuição discreta através de uma distribuição contínua, esse ajustamento é denominado **correção de continuidade**.

Em resumo, se X~b(n,p) com os parâmetros n grande e p não muito próximo de 0 ou 1, a distribuição da variável aleatória padronizada $Z = \frac{X - np}{\sqrt{np(1-p)}}$ é aproximadamente N(0,1), de modo que:

$$P[a \leq X \leq b] \approx P\left[\frac{a - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - np}{\sqrt{np(1-p)}}\right] \text{ (sem correção de continuidade)}$$

$$\approx P\left[\frac{a - 0{,}5 - np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b - 0{,}5 - np}{\sqrt{np(1-p)}}\right] \text{ (com correção de continuidade)}$$

Exemplo 5.6. Um levantamento de grande escala, efetuado recentemente, revelou que 30% das árvores de uma floresta estavam infestadas com um certo parasita. Supondo que o nível de infestação da floresta se tenha mantido estável, qual é a probabilidade de que em uma amostra aleatória de 1.000 plantas o número de plantas infestadas seja menor do que 280?

Seja X o número de plantas infestadas na amostra de 1.000 plantas da floresta. Sob a pressuposição de que a proporção de infestação na população seja 0,3, X tem distribuição b(1.000; 0,3). Então, como np=300 e np(1-p)=210 são ambos valores elevados, uma boa aproximação da distribuição de X é provida pela distribuição N(300, 210). Assim, a probabilidade de que o número de plantas infestadas seja inferior a 280 pode ser obtida como segue:

a) Sem correção de continuidade:

$$P[X<279] \approx P\left[Z \leq \frac{279 - 300}{14{,}5}\right] = P[Z \leq -1{,}448] = 0{,}074.$$

b) Com correção de continuidade:

$$P[X<279] \approx P\left[Z \leq \frac{279{,}5 - 300}{14{,}5}\right] = P[Z \leq -1{,}414] = 0{,}079.$$

Observe-se que para a situação do **Exemplo 5.6** a correção de continuidade pouco alterou o resultado. Isso decorre do pouco efeito da subtração de 0,5 no numerador quando o denominador $\sqrt{np(1-p)}$ é grande. De modo geral, a correção de continuidade é dispensável para valores elevados de np que implicam em valores elevados de $\sqrt{np(1-p)}$.

5.5 Distribuição Normal Bivariada

A distribuição normal multivariada é de especial importância para os métodos estatísticos a serem tratados neste texto. A discussão que segue, entretanto, restringir-se-á à distribuição normal bivariada, já que a situação multivariada requer notação matricial.

Um par de variáveis aleatórias X e Y tem **distribuição normal bivariada**, se e apenas se sua função de densidade de probabilidade conjunta é dada por:

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]}$$

para $-\infty<x<\infty$ e $-\infty<y<\infty$, onde $\sigma_1^2>0$, $\sigma_2^2>0$, e $-1<\rho<1$.

Demonstra-se que os parâmetros μ_1, μ_2, σ_1^2 e σ_2^2 são, respectivamente, as médias e as variâncias das variáveis aleatórias X e Y, ou seja:

$E(X) = \mu_1$; $E(Y) = \mu_2$;

$\text{Var}(X) = \sigma_1^2$; $\text{Var}(Y) = \sigma_2^2$.

O parâmetro ρ, ou ρ_{XY}, é denominado **coeficiente de correlação** das variáveis aleatórias X e YY. Sua expressão é:

$$\rho = \frac{\text{Cov}(X,\ Y)}{\sigma_1\ \sigma_2}.$$

Dessa forma, o parâmetro ρ exprime a variação conjunta das variáveis aleatórias X e Y.

As funções densidades condicionais das variáveis de um par de variáveis aleatórias com distribuição normal bivariada também são importantes. Demonstra-se que, se X e Y têm distribuição normal bivariada, a função densidade condicional de Y dado X=x também é normal, com média:

$$E(Y\,|\,X=x) = \mu_{Y|x} = \mu_2 + \rho\,\frac{\sigma_2}{\sigma_1}(x-\mu_1)$$

e variância:

$$\text{Var}(Y\,|\,X=x) = \sigma_{Y|x} = \sigma_2^2(1-\rho^2);$$

e a densidade de probabilidade condicional de X dado Y=y é normal com média:

$$E(X|Y=y) = \mu_{X|y} = \mu_1 + \rho\,\frac{\sigma_1}{\sigma_2}(x-\mu_2)$$

e variância:

$$\text{Var}(X\,|\,Y=y) = \sigma_{X|y} = \sigma_1^2(1-\rho^2);$$

A distribuição normal bivariada tem algumas propriedades importantes. Entre elas a seguinte: Se duas variáveis aleatórias têm distribuição normal bivariada, então elas são independentes se e somente se $\rho = 0$. Se $\rho_{XY} = 0$, diz-se que as variáveis X e Y são **não correlacionadas**.

5.6 Combinação Linear de Variáveis Aleatórias Normais

Em seções anteriores, foram apresentados alguns conceitos referentes a funções de várias variáveis aleatórias. Em particular, o conceito de valor esperado de uma função $g(X_1,X_2,...,X_n)$ de n variáveis aleatórias. A distribuição de funções de variáveis aleatórias é essencial em inferência estatística. Por exemplo, a distribuição de combinações lineares de variáveis aleatórias, das quais a média da amostra é um caso especial, tem um importante papel. Nesta Seção, tratar-se-á da distribuição de combinações lineares de variáveis aleatórias normalmente distribuídas.

Preliminarmente, revisar-se-á as expressões da média e da variância de uma combinação linear de variáveis aleatórias, de que se tratou na **Seção 4.10**.

Sejam $X_1,X_2,...,X_n$ variáveis aleatórias independentes com médias μ_1, μ_2, ..., μ_n e variâncias σ_1^2, σ_2^2,..., σ_n^2, respectivamente. Seja Y uma combinação linear das vaiáveis aleatórias X_1, X_2, ...,X_n, ou seja, uma variável aleatória expressa por:

$$Y = a_1 X_1 + a_2 X_2 +...+ a_n X_n,$$

onde a_1, a_2, ...,a_n são constantes. Então, Y possui às seguintes propriedades:

1) $E(Y) = a_1 E(X_1) + a_2 E(X_2) +...+ a_n E(Xn)$,

isto é:

$$\mu_Y = a_1\mu_1 + a_2\mu_2 + ... + a_n\mu_n.$$

Esta propriedade significa que o valor esperado de uma combinação linear de variáveis aleatórias é igual à combinação linear dos valores esperados dessas variáveis aleatórias. Esta relação é válida mesmo que as variáveis aleatórias X_1, X_2, ...,X_n sejam dependentes.

2) $Var(X) = a_1 Var(X_1) + a_2 Var(X_2) +...+ a_n Var(X_n)$,

ou seja:

$$\sigma^2 = a_1^2 \sigma_1^2 + a_2^2 \sigma_{21}^2 +...+ a_n^2 \sigma_n^2.$$

Esta propriedade é válida apenas se as variáveis aleatórias X_1, X_2, ...,X_n são independentes.

A propriedade fundamental referente à combinação linear de variáveis aleatórias, na situação particular em que as variáveis têm distribuição normal, é a seguinte:

Se X_1, X_2, ...,X_n são variáveis aleatórias independentes normalmente distribuídas, então $Y = a_1 X_1 + a_2 X_2 +...+ a_n X_n$ também é uma variável aleatória normalmente distribuída, com média μ e variância σ^2, ou seja:

$$Y \sim N(a_1\mu_1+a_2\mu_2+...+a_n\mu_n,\ a_1^2\sigma_1^2+a_2^2\sigma_2^2+...+a_n^2\sigma_n^2).$$

Casos particulares

As duas propriedades da distribuição normal que seguem são importantes em muitas aplicações.

1) Se $X\sim N(\mu,\sigma^2)$ e *a* e *b* são duas constantes, então:

$$Y = a + bX \sim (a + b\mu,\ b^2\sigma^2).$$

Ou seja, a **transformação linear** de uma variável aleatória normal, gera uma variável aleatória também normal.

2) Se $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ e X e Y são estatisticamente independentes, então:

$X+Y \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$ e

$X-Y \sim N(\mu_1-\mu_2, \sigma_1^2+\sigma_2^2)$.

Isto é, a soma e a diferença de duas variáveis aleatórias normais também são variáveis aleatórias normais.

5.7 Distribuição da Média de Variáveis Aleatórias

O valor esperado e a variância da média de n variáveis aleatórias independentes podem ser derivados dos resultados da seção anterior. Sejam X_1, X_2,...,X_n n variáveis aleatórias independentes normalmente distribuídas. A média dessas n variáveis é:

$$\overline{X} = \frac{X_1+X_2+...+X_n}{n} = \frac{1}{n}X_1 + \frac{1}{n}X_2 + ... + \frac{1}{n}X_n .$$

O valor esperado e a variância da média $\overline{X}$ são imediatamente derivados das duas propriedades listadas na seção anterior:

$$\mu_{\overline{X}} = \frac{1}{n}\left(\mu_1+\mu_2+...+\mu_n\right) \text{ e}$$

$$\sigma^2_{\overline{X}} = \frac{1}{n}\left(\sigma_1^2+\sigma_2^2+...+\sigma_n^2\right).$$

Se as n variáveis aleatórias X_1, X_2, ...,X_n têm distribuição idêntica, com média e variância comuns μ e σ^2, respectivamente, as expressões do valor esperado e da variância de sua média simplificam-se:

$$\mu_{\overline{X}} = \frac{1}{n}\left(\mu+\mu+...+\mu\right) = \mu, \text{ e}$$

$$\sigma^2_{\overline{X}} = \frac{1}{n}\left(\sigma_1^2+\sigma_2^2+...+\sigma_n^2\right) = \frac{\sigma^2}{n}.$$

Dessa forma, o valor esperado da média $\overline{X}$ das variáveis aleatórias X_1, X_2,... ,X_n é o valor esperado comum das variáveis aleatórias individuais e a variância de $\overline{X}$ é a n-ésima parte da variância comum dessas variáveis aleatórias. Ademais, se as n variáveis aleatórias X_1, X_2,... ,X_n são normalmente distribuídas com média μ e variância σ^2, a média $\overline{X}$ é normalmente distribuída com média μ e variância $\frac{\sigma^2}{n}$, ou seja: $X \sim N(\mu, \frac{\sigma^2}{n})$.

Na **Seção 5.3.2**, foi visto que subtraindo de uma variável aleatória com distribuição normal sua média e, então, dividindo por seu desvio padrão, a variável aleatória resultante tem distribuição normal padrão, com média 0 e variância 1. Assim, a variável aleatória:

$$\frac{\overline{X}-\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}\left(\overline{X}-\mu\right)}{\sigma}$$

tem distribuição normal com média 0 e variância 1, contanto que as variáveis aleatórias X_1, X_2,..., X_n sejam todas normalmente distribuídas com média e variância comuns μ e σ^2, respectivamente.

5.8 Exercícios

1. Determine a média e a variância da variável aleatória X com a seguinte função de densidade de probabilidade:

$$f(x) = \begin{cases} \frac{1}{10}, & 0 \le x \le 10 \\ 0, & \text{caso contrario} \end{cases}$$

2. A quantidade real de café (em gramas) em uma jarra de 8 hectogramas enchida por uma máquina é uma variável aleatória com densidade de probabilidade:

$$f(x) = \begin{cases} 0, & x \le 7{,}9 \\ 5, & 7{,}9 < x < 8{,}1 \\ 0, & x \ge 8{,}1 \end{cases}$$

   Determine a probabilidade de que uma jarra de 8 hectogramas preenchida por essa máquina contenha:

   a) no máximo 7,93 hectogramas de café;

   b) entre 7,89 e 8,02 hectogramas de café;

   c) no mínimo 7,95 hectogramas de café.

3. A variável aleatória X que exprime o peso de indivíduos de uma população tem distribuição normal com média igual a 70 kg e desvio padrão igual a 10 kg. Determine a probabilidade de indivíduos dessa população com peso compreendido em cada um dos seguintes intervalos:

   a) (40 kg, 60 kg); b) $\{x \mid x<100\}$; c) $\{x \mid x> 25\}$; d) $\{x \mid x >100\}$.

4. O tempo de vida sem defeito de um aparelho para uma determinação de laboratório tem distribuição normal com média de 12 meses e desvio padrão de 2 meses. O fabricante garante a substituição das unidades que produz caso demonstrem defeitos no prazo de 15 meses. Se o fabricante produz 10.000 dessas unidades em um ano, quantas dessas unidades deve esperar substituir dessa produção após 15 meses ?

5. A temperatura média no dia 21 de março em um local tem distribuição normal com média de 22°C e desvio padrão de 5,5°C. Qual é a probabilidade da temperatura do próximo dia 21 de março situar-se: a) entre 26 e 32°C; b) abaixo de 17°C ?

6. Uma pesquisa mostrou que o tempo de vida de um certo tipo de bateria tem distribuição normal com média de 1248 dias e desvio padrão de 185 dias. Se o fabricante dessas baterias garante a bateria por 24 meses (ou seja, 730 dias), que percentagem das baterias terá que substituir para atender à garantia ?

7. Quanto maior o conteúdo de enxofre do carvão, ele é menos desejável como combustível para aquecimento. Sabendo que a variabilidade da determinação do enxofre de carvão produzido de uma certa mina é σ = 6 kg por tonelada e que o

conteúdo de enxofre desse carvão tem distribuição normal, responda às seguintes questões:

a) Se minas com 80 kg de enxofre por tonelada de carvão são inadequadas para produção de carvão combustível, qual é a probabilidade de que uma mina com conteúdo médio de enxofre de 62 kg por tonelada seja considerada na categoria de inadequada na base de uma amostra aleatória de uma tonelada de carvão ?

b) Algumas cidades não permitem a venda de carvão nos limites da cidade se o seu conteúdo de enxofre é superior a 34 kg por tonelada. Qual é a probabilidade de que carvão com conteúdo médio de enxofre de 40 kg por tonelada terá permissão de ser vendido na cidade na base de uma amostra aleatória de uma tonelada do carvão ?

8. São disponíveis dados de potência de tração de um trator, em kg. Suponha que os dados possam ser considerados provenientes de uma distribuição normal com média 30 e variância 25. Extraindo-se, aleatoriamente, uma observação simples, determine a probabilidade de que ela seja maior do que 45?

9. Uma indústria de pêssego em calda efetua o enlatamento de seu produto com o uso de equipamento que preenche latas com peso líquido nominal de 1 kg, mas que se distribui normalmente com média de 1 kg. O peso líquido de uma lata de compota é considerado apreciavelmente abaixo do peso nominal se for inferior a 985 gramas. Se o risco desse fato deve ser menor que 0,01, qual é o valor máximo permitido para o desvio padrão?

10. As especificações de uma lata cilíndrica para enlatamento de pêssego em calda estabelecem um diâmetro interno de 25cm ± 0,5cm. Se o diâmetro interno das latas produzidas por uma indústria dessas embalagens é distribuído normalmente com média μ=25,1cm e desvio padrão σ=0,3cm, qual é a proporção das latas de sua produção que satisfarão as especificações ?

11. Registros anteriores revelam que a produção anual de leite de vacas de um rebanho jérsei é aproximadamente normal com média μ=35 kg e desvio padrão σ=6 kg.

a) Qual é a probabilidade de que a produção de leite para uma vaca escolhida aleatoriamente desse rebanho situe-se no intervalo entre 30 e 45 kg?

b) Qual é a probabilidade de que a produção de leite de uma vaca selecionada aleatoriamente desse rebanho exceda 45 kg em um dado ano?

12. A quantidade de café instantâneo que uma máquina deposita em canecas com capacidade para 180 gramas é uma variável aleatória com distribuição normal com média igual a 180 gramas e desvio padrão de 2,4 gramas. Determine a probabilidade de que uma caneca preenchida por essa máquina contenha:

a) no mínimo 183 gramas;

b) menos do que 175 gramas;

c) entre 179 e 181 gramas.

13. O tempo de vida de lâmpadas produzidas por uma indústria é uma variável aleatória com distribuição normal com desvio padrão de 25 horas. Determine o tempo médio de vida dessas lâmpadas, se é sabido que a probabilidade de que uma lâmpada dessa procedência dure mais de 400 horas é 0,10.

14. Uma máquina produz resistores elétricos com resistência média de 50 ohms, com desvio padrão de 2 ohms. Admitindo-se que a distribuição da resistência é normal, quais são os limites de tolerância que devem ser impostos à resistência para assegurar que não mais do que 1/1000 dos resistores caiam fora da faixa de tolerância?
15. Outras resistências são produzidas a partir daquelas fabricadas pela máquina considerada na questão anterior, através da montagem em série de duas resistências simples produzidas pela máquina. Supondo que os pares para montagem dessas resistências são escolhidos aleatoriamente, quais são os limites de tolerância que devem ser impostos à resistência total para assegurar que não mais do que 1/1000 delas caiam fora desses limites? (A resistência total de dois resistores com resistências de R_1 e R_2 ohms montados em série é $R_1 + R_2$.)
16. Uma indústria produz tubos cilíndricos cujo diâmetro interno é uma variável aleatória com média de 3 cm e desvio padrão de 0,02 cm, e cuja espessura é uma variável aleatória com média de 0,3 cm e desvio padrão de 0,005 cm. Supondo que a espessura do tubo é independente do diâmetro, quais são a média e o desvio padrão da distribuição do diâmetro externo dos tubos produzidos por essa indústria?
17. Um fabricante sabe que, em média, dois porcento de seus produtos são defeituosos. Usando a aproximação provida pela distribuição normal, determine a probabilidade de que em um lote de 400 unidades ocorram:
    a) exatamente 10 unidades defeituosas;
    b) pelo menos 10 unidades defeituosas.
18. Se 60 por cento dos clientes de uma loja compram a prazo, qual é a probabilidade de que entre 200 clientes escolhidos aleatoriamente pelo menos 125 façam suas compras a prazo? (Use a aproximação provida pela distribuição normal.)
19. Uma estação de televisão afirma que seu horário de cinema de segunda-feira a noite tem, regularmente, 35 por cento da audiência total. Se essa afirmação é correta, qual é a probabilidade de que entre 100 telespectadores contatados por telefone em uma noite de segunda-feira menos do que 25 estejam assistindo o filme?
20. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando as letras V e F entre parêntesis, respectivamente. Se a sentença for falsa, explique porque.
    ( ) Teoricamente, variáveis contínuas podem ser medidas tão precisamente quanto desejado, mas usualmente elas são arredondadas para uma unidade conveniente.
    ( ) Em uma distribuição de probabilidade contínua, a área total entre a curva que representa a distribuição e o eixo horizontal é igual a 1.
    ( ) Em uma distribuição de probabilidade contínua, a probabilidade de qualquer valor particular da variável aleatória é a distância vertical, no ponto correspondente a este valor, entre a curva que representa a distribuição e o eixo horizontal.
    ( ) Nenhum dos dois parâmetros de uma distribuição normal pode ser negativo.
    ( ) $P(X>\mu) = 0,5$ apenas se X tem distribuição normal.
    ( ) Se as notas em uma disciplina de Estatística de uma classe seguem a distribuição normal N(7,0; 0,81), cerca de 95% da classe terá notas no intervalo entre 6,1 e 7,9.

( ) Pesquisa extensiva tem mostrado que a maioria dos podem ser considerados como realizações de variáveis aleatórias com distribuição normal, independentemente do campo de pesquisa.

( ) Todas funções de distribuições contínuas com representação geométrica em forma de sino são distribuições normais.

( ) Em uma distribuição normal, se a média tem valor numérico elevado, então a variância também tende a ser grande.

( ) Em uma distribuição normal, cerca de 95% dos valores ficam no intervalo de -2 a 2.

( ) Na distribuição normal padrão, a média e a variância são, respectivamente, np e np(1-p).

( ) Se X~N(4, 1), então $\sigma_{\overline{X}} = 1/\sqrt{n}$.

( ) $P(\overline{X} > a) \neq P[Z > (a - \mu_{\overline{X}})/\sigma_{\overline{X}}]$, se X tem distribuição normal.

6 AMOSTRAGEM ALEATÓRIA E DISTRIBUIÇÃO AMOSTRAL

Conteúdo

6.1 Introdução....133
6.2 Inferência Indutiva....133
6.3 População e Amostra....134
6.4 Distribuição Amostral....136
6.5 Média e Variância Amostrais....138
6.6 Teorema Central do Limite....139
6.7 Amostragem de Distribuições Normais....142
6.7.1 Introdução....142
6.7.2 Distribuição da média da amostra....143
6.7.3 Distribuição da variância da amostra....144
6.7.4 Distribuição t de Student....150
6.7.5 Distribuição F....154
6.8 Exercícios....158

6.1 Introdução

Nos capítulos anteriores, tratou-se dos conceitos e aplicações de probabilidade e de algumas distribuições de probabilidade importantes. O propósito deste Capítulo é introduzir o conceito de amostragem e as distribuições de variáveis aleatórias importantes derivadas pela amostragem. Ele conecta a matéria referente a distribuições de probabilidade com o corpo da estatística propriamente dita, ou seja, os métodos de inferência estatística, relacionados com estimação e teste de hipótese, a serem tratados a partir do próximo capítulo.

6.2 Inferência Indutiva

O progresso do conhecimento científico é freqüentemente atribuído à experimentação. A pesquisa experimental é conduzida para a busca de soluções para problemas referentes a sistemas de interesse. Com base nos dados gerados pela pesquisa, são extraídas conclusões que, usualmente, extrapolam as condições do experimento particular. Ou seja, o pesquisador deriva generalizações a partir do experimento particular para uma classe de experimentos similares. Essa extensão de uma situação particular para uma mais geral é denominada **inferência indutiva**.

A inferência indutiva é um processo de raciocínio fortuito, que gera conclusões incertas. Entretanto, se o experimento é conduzido de acordo com certos princípios, o grau de incerteza da inferência indutiva pode ser mensurado. De fato, uma das funções da estatística é a provisão de técnicas para a derivação de inferências indutivas que permitam a mensuração do grau de incerteza envolvido. A incerteza é expressa em termos de probabilidade. Por essa razão, os conceitos e as propriedades relacionados com probabilidade são fundamentais para a compreensão dos métodos de inferência estatística.

Enquanto que as conclusões derivadas da inferência indutiva são apenas prováveis, aquelas obtidas por **inferência dedutiva** são conclusivas. O processo de inferência dedutiva é mais comum nas ciências exatas. O método silogístico ilustra o processo de inferência dedutiva. A partir de duas proposições referentes a fatos aceitos, denominadas premissas principal e secundária, é derivada uma terceira proposição, a conclusão. Por exemplo:

Premissa principal: Os ângulos internos de um retângulo têm 90º.

Premissa secundária: O quadrado é um retângulo.

Conclusão: Os ângulos internos do quadrado têm 90º.

Embora a inferência dedutiva seja extremamente importante, o conhecimento do mundo real deriva-se, preponderantemente, pelo processo de inferência indutiva. O processo de inferência indutiva é ilustrado através do **Exemplo 6.1**.

Exemplo 6.1. Suponha-se que um produtor de sementes tem em um tonel cinco milhões de sementes de plantas de uma espécie que produz flores brancas e azuis e deseja saber a proporção dessas sementes que produzirão flores brancas. Esta questão pode ser respondida de modo correto apenas através do plantio de cada uma das cinco milhões de sementes e a observação da cor das flores de cada planta resultante. Entretanto, esse processo não é apropriado, já que o produtor deseja vender as sementes, ou quer obter a resposta sem utilizar todas as sementes de que dispõe ou sem despender tanto esforço. Um processo alternativo é plantar algumas poucas sementes e, com base nas cores das flores das plantas resultantes destas sementes, derivar uma conclusão referentes às cinco milhões de sementes. Naturalmente, esse processo não garantirá uma resposta certa, que poderia ser produzida apenas pela observação das cores das plantas produzidas por todas as cinco milhões de sementes. A observação das cores das flores de plantas resultantes de apenas algumas das sementes não permite uma predição exata. Entretanto, pode ser derivada uma conclusão probabilista satisfatória, se as sementes forem escolhidas através de processo apropriado. Esse é um processo de inferência indutiva. Ou seja, seleciona-se uma parte das cinco milhões de sementes, planta-se essas sementes, observa-se o número dessas sementes que produzem plantas de flores brancas, e, com base nessas poucas sementes, faz-se uma predição sobre a proporção das cinco milhões de sementes que correspondem a plantas de flores brancas. Dessa forma, a partir do conhecimento das cores de poucas sementes, generaliza-se para o conjunto total das cinco milhões de sementes. Esse processo não assegura a correção da resposta, mas provê um nível de confiança na conclusão em um sentido probabilista.

6.3 População e Amostra

Na seção anterior, viu-se que o processo geral de aquisição de conhecimento do mundo empírico é a observação de poucas unidades ou elementos do conjunto total de unidades de interesse e, na base dessas poucas unidades, derivações de conclusões para o conjunto maior. O

conjunto de todas as unidades de interesse em uma pesquisa é denominado **população objetivo**, ou, mais simplesmente, **população**. No **Exemplo 6.1**, a população objetivo é o conjunto das cinco milhões de sementes no tonel. Em outras situações, a população objetivo pode ser as lavouras plantadas com trigo em uma região, os animais adultos de uma raça de gado de corte em uma certa época, o conjunto hipotético das alturas de plantas de uma cultivar de soja, etc. A população objetivo pode ser real ou hipotética, mas deve ser bem definida, através da identificação de seus elementos ou da especificação das condições para que um elemento lhe pertença.

A abordagem da inferência indutiva, sob o ponto de vista da estatística, é a seguinte: O objetivo de uma pesquisa é desvendar algum aspecto da população objetivo. Geralmente, é impossível ou impraticável ou inconveniente, examinar a população inteira. Então, examina-se uma sua parte (ou seja, uma **amostra** da população) e, com base nesta pesquisa limitada, estabelece-se inferências com respeito à população objetivo.

O problema que se levanta imediatamente é o modo de seleção da amostra. A validade das conclusões probabilistas referentes à população objetivo depende do modo de seleção da amostra. Particularmente importante é o caso de amostra aleatória simples, que será definida adiante.

Como regra, o interesse de uma pesquisa não reside na população propriamente, mas em uma ou mais **características** de suas unidades. No **Exemplo 6.1**, a característica de interesse é a cor das flores das plantas produzidas pelas sementes; não há interesse genérico nas sementes ou nas plantas, ou em qualquer outra característica das unidades da população de sementes. Por essa razão, é usual restringir a designação de **população** à coleção das alternativas das características de interesse (no caso, cores das flores das plantas resultantes das sementes), expressas por uma variável numérica, no conjunto de unidades que a pesquisa visa abranger. Neste texto, tratar-se-á apenas da situação de uma característica.

Pressupõe-se que a característica de interesse na população é expressa por uma variável aleatória com uma distribuição de probabilidade conhecida, de modo que a cada elemento na população é associado um valor numérico. No **Exemplo 6.1**, as cinco milhões de sementes no tonel constituem a população da qual deve ser extraída uma amostra. Cada semente é uma unidade da população e produzirá uma planta de flor branca ou azul. Dessa forma, a característica de interesse é a cor da flor com duas alternativas: branca e azul.

Para a modelagem probabilista deve-se definir uma variável aleatória que associe um valor numérico com cada uma das alternativas da característica, para cada elemento da população. Seja a variável aleatória X_i que assume os valor 1 ou 0, dependendo se a i-ésima semente amostrada produz uma planta de cor branca ou azul, respectivamente, i=1,2,...,n (n = **tamanho da amostra**). Se o processo de amostragem das sementes é tal que as variáveis aleatórias $X_1, X_2,...,X_n$ são independentes e têm a mesma distribuição de probabilidade, ou seja, $f(x_1) = f(x_2) =... = f(x_n)$, então a amostra é denominada **aleatória simples**.

O significado das variáveis aleatórias $X_1, X_2,...,X_n$ é fundamental para a compreensão do conceito de amostra aleatória. A variável aleatória X_i é uma representação para o valor numérico que o i-ésimo elemento amostrado assumirá. Após a observação da amostra, os valores atuais específicos de $X_1, X_2,...,X_n$ são conhecidos. Esses valores observados, denotados por $x_1, x_2,...,x_n$, respectivamente, são considerados uma **realização** das variáveis aleatórias $X_1, X_2,...,X_n$.

Freqüentemente, não é possível selecionar uma amostra aleatória da população objetivo. Uma amostra aleatória é, então, selecionada de alguma população relacionada. Se $X_1, X_2,...,X_n$ é

uma amostra aleatória de uma população com distribuição de probabilidade f(x), então esta população é denominada **população amostrada**.

Através de amostras aleatórias, pode-se derivar sentenças probabilistas válidas referentes a populações amostradas. Entretanto, conclusões probabilistas sobre a população objetivo não são válidas no sentido probabilista de freqüência relativa, a menos que a população objetivo coincida com a população amostrada. Os seguintes exemplos ilustram a distinção entre população amostrada e população objetivo.

Exemplo 6.2. Suponha-se que um pesquisador conduz um levantamento para o estudo da adoção de tecnologias por agricultores, através de uma amostra dos agricultores associados às cooperativas de uma região. Neste caso, a população objetivo é o conjunto dos agricultores da região, enquanto a população amostrada compreende os agricultores cooperados da região. O pesquisador pode derivar conclusões probabilistas válidas sobre esta população amostrada. Entretanto, deverá usar seu julgamento pessoal para extrapolar para a população objetivo. A confiabilidade dessa extrapolação não pode ser avaliada em termos probabilistas de freqüência relativa.

Exemplo 6.3. Um pesquisador deve executar um experimento para a comparação de cultivares de arroz com o propósito de derivar recomendação de cultivares para os agricultores de uma região. O pesquisador tem a sua disposição oito locais da região onde pode conduzir o experimento. Nessa situação, a população amostrada compreende esses oito locais, enquanto a população objetivo é constituída por todos os locais de cultivo de arroz da região.

Daqui em diante, muito mais freqüentemente, se fará referência à população amostrada. Quando for usada a designação "população" sem o qualificativo "amostrada" ou "objetivo", se estará referindo à população amostrada.

A **inferência estatística** trata do problema de selecionar uma amostra de uma população com distribuição de probabilidade f(.) e, com base nas observações dessa amostra, derivar sentenças probabilistas referentes à f(.).

6.4 Distribuição Amostral

Seja X uma variável aleatória de uma população com distribuição de probabilidade f(x). Suponha-se que uma amostra de dois valores de X é extraída aleatoriamente. Os números resultantes são denotados por x_1 e x_2, na ordem de extração, e designados, respectivamente, **primeira** e **segunda observação**. A coleção de todos tais pares de números que poderiam ter resultado do processo de extração aleatória de uma amostra de tamanho dois da população constitui uma distribuição bivariada. Cada um desses pares de números (x_1,x_2) é um **valor** (ou uma **realização**) da variável aleatória bidimensional (X_1,X_2), e X_1, X_2 é uma **amostra aleatória de tamanho dois** da distribuição f(x). Pela definição de amostra aleatória, a distribuição conjunta de X_1 e X_2, que se denomina **distribuição da amostra de tamanho 2**, é dada por: $f_{X_1X_2}(x_1,x_2) = f_{X_1}(x_1)\,f_{X_2}(x_2)$.

Exemplo 6.4. Seja X uma variável aleatória com dois valores 0 e 1 que exprime uma característica com duas alternativas, com probabilidades p e 1-p, respectivamente. Então, X tem

distribuição discreta de Bernoulli e sua função de probabilidade é: $f(x) = p^x(1-p)^{1-x}$, x = 0, 1. A distribuição conjunta para uma amostra aleatória de dois valores de X, sejam x_1 e x_2, é:

$$f_{X_1X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) = p^{x_1+x_2}(1-p)^{2-x_1-x_2},\ x_1, x_2 = 0, 1.$$

Observe-se que a função $f_{X_1X_2}(x_1, x_2)$ é a distribuição da amostra na ordem de extração. A probabilidade de uma amostra em que é extraído primeiro 0 e após 1 é: $f_{X_1X_2}(0,1) = p(1-p)$. Assim, a função bivariada considerada neste exemplo não é a distribuição do número de sucessos, seja Y, em dois ensaios de Bernoulli, cuja função de densidade é: $f_Y(y) = C_y^2 p^y (1-p)^{2-y}$, y = 0,1,2.

O conceito de amostra aleatória generaliza-se para uma amostra de tamanho n:

> Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n de uma população com distribuição de probabilidade f(x). A **distribuição da amostra aleatória** $X_1, X_2, ..., X_n$ é a distribuição conjunta de $X_1, X_2, ..., X_n$, ou seja:
>
> $$f_{X_1X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1) f_{X_2}(x_2), ..., f_{X_n}(x_n)$$

Observe-se que a definição de variável aleatória exclui, automaticamente, amostragem de uma população finita sem reposição, já que neste caso os resultados das extrações não são independentes.

O problema básico em estatística é o seguinte: É desejado estudar uma população que tem uma distribuição de probabilidade cuja forma é conhecida mas dependente de uma ou mais características desconhecidas, denominadas **parâmetros**. Dessa forma, a função distribuição de probabilidade não é completamente conhecida; ela é uma função do parâmetro desconhecido. Ou seja, sua forma é $f(x;\theta)$, onde θ denota o parâmetro. O procedimento da estatística clássica é extrair uma amostra aleatória $X_1, X_2, ..., X_n$ de tamanho n desta distribuição e aproximar o valor desconhecido do parâmetro θ por um **estimador**, ou seja, uma função $t(X_1, X_2, ..., X_n)$ dos valores da amostra. Em geral, muitas funções podem cumprir esse propósito. O problema é determinar a melhor função para estimar θ. Este problema será formulado e tratado no próximo capítulo. Nesta seção, tratar-se-á de certas funções de uma amostra aleatória. Inicialmente, definir-se-á o conceito de estatística:

> Uma **estatística** é uma função de variáveis aleatórias observáveis que não depende de qualquer parâmetro desconhecido.

Assim, uma estatística é, também, uma variável aleatória. O qualificativo "observável" é requerido das variáveis aleatórias tendo em conta a intenção do uso da estatística na execução de inferências referentes à função de distribuição das variáveis aleatórias. Se as variáveis aleatórias não fossem observáveis, ou seja, se seus valores não pudessem ser observados, elas não seriam utilizáveis em inferências. Por exemplo, se $X \sim N(\mu,\sigma^2)$, onde μ e σ^2 são desconhecidos, $X-\mu$ e $(X-\mu)/\sigma^2$ não são estatísticas, visto que elas não são funções apenas da variável aleatória observável X, mas, também, de parâmetros desconhecidos. Entretanto, X, X-2 e X^2, por exemplo, são estatísticas.

6.5 Média e Variância Amostrais

Seja $X_1, X_2, ...,X_n$ uma amostra aleatória de uma distribuição de probabilidade f(x; θ). A **média amostral**, denotada por $\overline{X}_n$ ou $\overline{X}$, é definida como:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + ... + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

O valor da média amostral para uma amostra particular $(x_1, x_2,...,x_n)$, denotado por $\overline{x}$, é dado por:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

conforme definido na **Seção 2.5**.

A média amostral $\overline{X}$ é uma estatística, ou seja, uma função das variáveis aleatórias $X_1, X_2, ...,X_n$. Logo, ela também é uma variável aleatória com uma distribuição de probabilidade. Em geral, a distribuição de probabilidade de $\overline{X}$ depende da função distribuição de probabilidade f(.) da qual foi extraída a amostra aleatória.

Se $X_1, X_2,...,X_n$ é uma amostra aleatória da função de probabilidade f(.) com média μ e variância σ^2 e $\overline{X} = \frac{1}{n}(X_1+X_2+...+X_n)$, então (**Seção 5.7**):

$$E(\overline{X}) = \mu_{\overline{X}} = \mu \quad \text{e} \quad \text{Var}(\overline{X}) = \sigma^2_{\overline{X}} = \frac{1}{n}\sigma^2.$$

Estas propriedades da média amostral $\overline{X}$ sugerem a possibilidade de seu uso como uma estimativa da média μ da distribuição f(.). A propriedade $E(\overline{X}) = \mu$ significa que, na média, $\overline{X}$ é igual ao parâmetro μ que está sendo estimado, ou seja, que a distribuição de $\overline{X}$ está centrada em μ. A outra propriedade, $\text{Var}(X) = \frac{1}{n}\sigma^2$, significa que a dispersão dos valores de $\overline{X}$ em torno da média μ é a n-ésima parte da dispersão da distribuição populacional f(.), e é tão menor quanto maior o tamanho n da amostra. Por exemplo, a variância de $\overline{X}$ para uma amostra de tamanho 10 é a metade da variância para uma amostra de tamanho 5. Dessa forma, os valores de $\overline{X}$ tendem a ser mais concentrados em torno de μ para uma amostra de tamanho grande do que para uma amostra de tamanho pequeno.

Ademais, conforme visto na **Seção 5.7**, se $X \sim N(\mu,\sigma^2)$, i=1,2,...,n, então $\overline{X} \sim N(\mu,\sigma^2/n)$, e $\sqrt{n}(X-\mu)/\sigma \sim N(0,1)$. Dessa forma, se a distribuição populacional é normal, a média amostral também tem distribuição normal, com dispersão muito mais concentrada em torno da média do que a variável aleatória original. Essa propriedade é ilustrada na **Figura 6.1**.

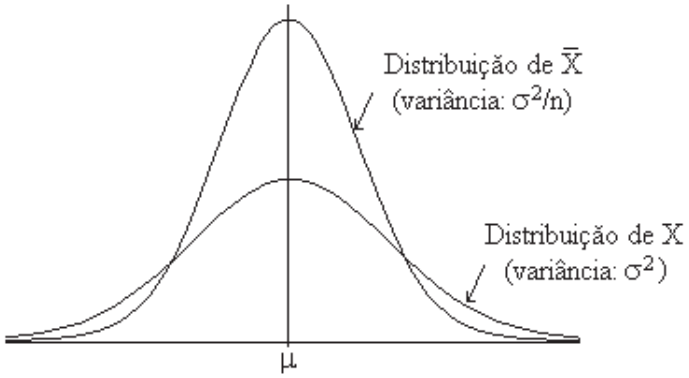


Figura 6.1. Distribuições de probabilidade de X e da correspondente média amostral $\overline{X}$.

Se X_1, X_2, ...,X_n é uma amostra aleatória de tamanho n de uma distribuição f(x), então, a **variância amostral**, denotada por S_n^2 ou S^2, é:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \quad n > 1$$

Conforme definido na **Seção 2.5**, a variância para uma amostra particular é denotada por s^2 e expressa por:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2 .$$

A variância amostral para uma amostra particular também é designada **quadrado médio** e denotada, simbolicamente, por:

$$s^2 = \frac{1}{GL}SQX,$$

onde $SQX = \sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ é a **soma dos quadrados dos desvios** dos valores amostrais x_i em relação à média amostral $\overline{x}$, ou **soma dos quadrados de X corrigida para a média**, e GL = n-1 denota os **graus de liberdade** associados com SQX.

A razão para o divisor n-1 em lugar de n, que poderia parecer mais natural na expressão da variância amostral S^2, é a propriedade conveniente que resulta para S^2, ou seja: $E(S^2) = \sigma^2$.

6.6 Teorema Central do Limite

Na seção anterior, foram referidas propriedades importantes da distribuição da média amostral $\overline{X}$, ou seja, que $E(\overline{X})=\mu$ e $Var(\overline{X})=\sigma^2/n$, qualquer que seja a distribuição populacional f(x), e que, se esta distribuição é normal, a distribuição de $\overline{X}$ também é normal. O **teorema central do limite**, enunciado a seguir, estabelece que, sob certas condições, a distribuição aproximada de $\overline{X}$ é normal, qualquer que seja a distribuição de probabilidade f(.).

Seja $\overline{X}_n = \frac{1}{n}(X_1+X_2+...+X_n)$ a média de uma amostra de tamanho n, X_1, X_2 ,...,X_n, de uma distribuição de probabilidade qualquer f(.) com média μ e variância σ^2 e defina-se:

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma}.$$

Então, a distribuição de Z_n aproxima-se da distribuição N(0,1) quando n tende para infinito.

Este teorema central do limite estabelece, de modo mais geral, que a média e a soma de n variáveis aleatórias **idêntica e independentemente distribuídas** é normal, qualquer que seja a distribuição comum das variáveis aleatórias.

Este é um dos mais importantes teoremas em probabilidade e estatística, por essa razão, denominado **teorema central do limite**. Ele estabelece que a distribuição limite de Z_n, ou seja da média $\overline{X}_n$ padronizada, é a distribuição normal padrão, e que a própria média $\overline{X}_n$ tem distribuição aproximada, ou **assintótica**, normal com média μ e variância σ^2/n, qualquer que seja a distribuição populacional comum (discreta ou contínua).

O extraordinário desse teorema é que ele não estabelece qualquer condição referente à forma da distribuição de probabilidade original. Sua importância, sob o ponto de vista prático, é que, para grandes amostras, a média amostral $\overline{X}_n$ de uma amostra de qualquer distribuição com média μ e variância σ^2 tem, aproximadamente, distribuição normal, com média μ e variância σ^2/n, desde que a média e a variância sejam finitas.

Sob certas condições, o teorema central do limite também é válido quando as variáveis aleatórias não são identicamente distribuídas. De modo geral, pode-se dizer que a generalização que ele estabelece é válida quando as contribuições das variáveis aleatórias individuais para a soma, ou média, são "pequenas" relativamente à soma total. Nessa situação, μ é substituído por $\frac{1}{n}(\mu_1+\mu_2+...+\mu_n)$ e σ^2 por $\frac{1}{n}(\sigma_1^2+\sigma_2^2+...+\sigma_n^2)$, onde μ_i e σ_i^2 e denotam a média e a variância da variável aleatória X_i.

A questão importante refere-se ao tamanho da amostra necessário para o uso da distribuição normal como uma aproximação para a distribuição de $\overline{X}$, ou ao grau de aproximação logrado com o uso de uma amostra de um dado tamanho n. O grau de aproximação depende do tamanho da amostra e da grandeza do desvio da distribuição populacional em relação à forma normal, ou seja, da forma da distribuição de probabilidade f(.) das variáveis aleatórias originais. Se a distribuição da população é normal, então a média $\overline{X}$ tem distribuição normal exata para qualquer valor de n. Na medida em que a distribuição da população afasta-se da normalidade, valores maiores de n são requeridos para uma boa aproximação. Como uma regra geral, pode ser esperado que para valores n>30 sejam logradas boas aproximações em situações em que a distribuição populacional tem uma única moda e não tem caudas demasiadamente longas.

A validade do teorema central do limite pode ser evidenciada empiricamente para qualquer distribuição particular, através de construção experimental como a descrita a seguir. Foram extraídas amostras sucessivas de tamanho quatro da **distribuição retangular** e determinadas as correspondentes médias amostrais. A distribuição de freqüências foi, então,

construída e representada geometricamente na **Figura 6.2**, onde os pontos substituem os retângulos, indicando as freqüências relativas nas classes em cujo centro se projetam. Através desses pontos, foi traçada uma curva suave, resultando um gráfico em forma aproximada de sino, característica da distribuição normal. A **Figura 6.3** apresenta o resultado de um experimento semelhante efetuado com a **distribuição triangular**.

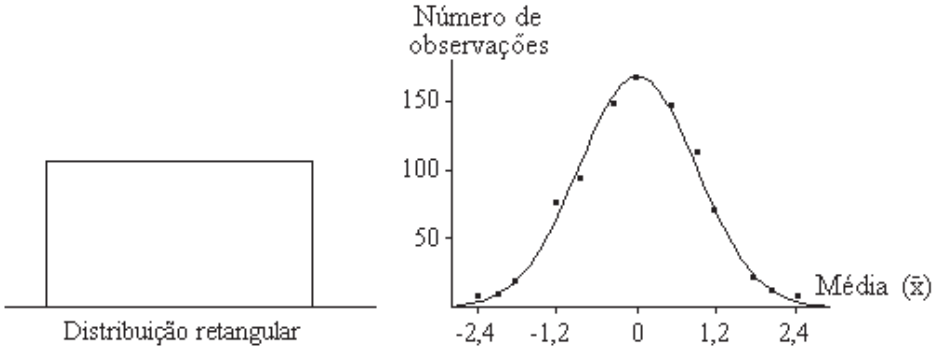


Figura 6.2. Distribuição empírica da média de amostras de tamanho quatro da distribuição retangular.

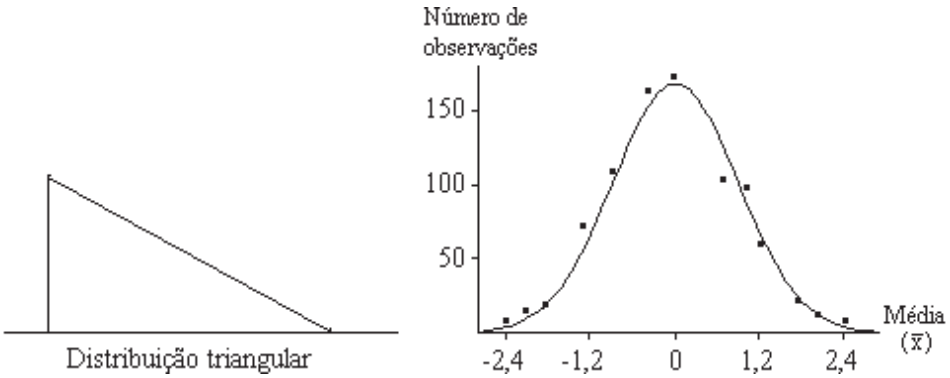


Figura 6.3. Distribuição empírica da média de amostras de tamanho quatro da distribuição triangular.

Os experimentos utilizados para ilustração demonstraram, empiricamente, que a distribuição amostral da média $\overline{X}$ é aproximadamente normal, mesmo para duas distribuições básicas drasticamente não normais, como o são as distribuições retangular e triangular, e amostras de tamanho bastante pequeno.

Assim, o teorema central do limite é fundamental para a importância da distribuição normal em estatística.

Na teoria dos erros, por exemplo, pode ser suposto, usualmente, que os erros têm distribuição normal, já que eles são comumente compostos da soma de um grande número de pequenos erros independentes.

A maioria dos procedimentos de inferência em estatística clássica, para estimação e teste de hipótese, são baseados na média amostral. Se as variáveis aleatória individuais que

compreendem a amostra são normalmente distribuídas, então a média amostral é certamente normal e a teoria para variáveis aleatórias normais é apropriada. Se as variáveis aleatórias individuais não são normalmente distribuídas, mas seu número é suficientemente grande, o teorema central do limite assegura que a média amostral tem distribuição aproximadamente normal, de modo que a teoria para variáveis aleatórias normais também é aplicável. Entretanto, se as variáveis aleatórias individuais que compreendem a amostra têm distribuição desconhecida e seu número não é suficientemente grande, a distribuição da média amostral não pode ser pressuposta normal. Nessa situação, a teoria baseada em distribuição normal não é adequada e deve-se recorrer a técnicas não paramétricas.

A aproximação normal da distribuição binomial é um exemplo particular da aplicação do teorema central do limite. Seja X_i uma amostra aleatória de uma distribuição de Bernoulli. Então, a distribuição de X_i é:

$$f(x_i) = \begin{cases} p, & x_i = 1 \\ 1-p, & x_i = 0 \end{cases}$$

com média p e variância p(1-p), conforme visto na **Seção 4.3.2**. Segundo o teorema central do limite, a distribuição da média amostral X é aproximadamente N(p, p(1-p)/n) quando n é grande. Agora, se $Y = X_1+X_2+ ...+X_n$ é o número de sucessos nos n ensaios de Bernoulli, então: $Y=n\overline{X}$. Logo, de acordo com propriedade da distribuição normal (**Seção 5.6**), Y∼N(np, np(1-p)), já que *E*(Y = np e Var(Y) = np(1-p). Portanto, $Z = (Y-np)/\sqrt{np(1-p)} \sim N(0,1)$, o mesmo resultado obtido na **Seção 5.4**.

Em problemas de inferência estatística, a proporção de sucessos na população p é desconhecida, constituindo, de fato, o propósito da pesquisa. Um procedimento de inferência estatística é, então, obter o número de sucessos Y, ou a proporção de sucessos $\hat{p}=nY$, a partir de uma amostra, e, usando este valor observado, determinar um intervalo de valores plausíveis para p.

6.7 Amostragem de Distribuições Normais

6.7.1 Introdução

A distribuição normal tem um papel predominante em estatística. O teorema central do limite é um argumento para a adequabilidade da distribuição normal em muitas situações. Entretanto, há outras razões importantes para a extensa utilização da distribuição normal. Em primeiro lugar, a experiência em pesquisas em muitos campos tem demonstrado que a distribuição normal é apropriada, pelo menos como uma boa aproximação à verdadeira distribuição, para muitos tipos de medidas físicas e de outras origens. Este fato tem sido considerado muito razoável com suporte no próprio teorema central do limite. A distribuição das distâncias dos pontos de impacto dos projéteis em relação a um alvo pode ser utilizada como ilustração. O percurso do projétil é afetado por inúmeros fatores, cada um de pequeno efeito. O desvio médio é o efeito líquido de todos esses fatores. Supondo que o efeito de cada um dos fatores é uma observação de alguma população, então o efeito total é, essencialmente, a média de um conjunto de observações de um conjunto de populações. Como os desvios observados têm a natureza de médias, pode-se esperar que eles tenham distribuição aproximadamente normal. Essa ilustração não tem a intenção

de implicar que a maioria das populações encontradas na prática são normais, o que não é o caso, mas que distribuições aproximadamente normais são muito freqüentes.

Um aspecto muito favorável à distribuição normal é que distribuições amostrais baseadas em uma distribuição normal são razoavelmente tratáveis analiticamente. A obtenção de distribuições de funções de observações amostrais é usualmente mais fácil para amostras de uma população normal do que para qualquer outra distribuição. Como conseqüência, distribuições amostrais de estatísticas derivadas da distribuição normal têm sido extensivamente estudadas e tabuladas, o que torna muito conveniente o uso da teoria e de tabelas disponíveis quando os dados podem ser pressupostos como razoavelmente normais.

Uma outra consideração favorável à distribuição normal é que muitas técnicas nela baseadas são robustas, ou seja, permanecem aproximadamente corretas para desvios razoáveis da normalidade. Deste modo, para muitos propósitos, a pressuposição de normalidade implica em pequeno erro, contanto que o desvio não seja demasiadamente elevado.

Como conseqüência dessas razões favoráveis à distribuição normal, ela tem sido freqüentemente utilizada de modo inadequado. Portanto, é conveniente alertar que a pressuposição de normalidade não deve ser admitida sem a necessária consideração e cuidado. Nesse sentido, testes apropriados são disponíveis que podem ser utilizados em situações em que a pressuposição de normalidade não seja assegurada pela experiência do pesquisador ou indicação da literatura.

Antes da aplicação de métodos estatísticos baseados na distribuição normal, o pesquisador deve conhecer, pelo menos aproximadamente, a forma geral da função de probabilidade que suas observações seguem. Se ela é normal, os métodos podem ser usados diretamente; caso contrário, algumas vezes, uma transformação de dados pode conduzir a observações transformadas que seguem uma distribuição normal. Se a normalidade não é lograda por uma transformação de dados ou se a forma da distribuição populacional não é conhecida, então deve-se recorrer a outros métodos de análise mais geral mas usualmente menos poderoso, denominados **métodos não paramétricos**.

As distribuições amostrais das estatísticas importantes utilizadas nos métodos de inferência a serem tratados neste texto são apresentadas a seguir.

6.7.2 Distribuição da média da amostra

A distribuição da média amostral de uma população normal foi considerada na **Seção 6.5**. Recorde-se que, se X_i , i=1,2,...,n, é uma amostra aleatória de uma distribuição $N(\mu,\sigma^2)$, a média $\overline{X}$ tem distribuição $N(\mu,\sigma^2/n)$. Em outras palavras, a distribuição exata da média amostral depende da distribuição populacional f(.) e tem a mesma forma desta, ou seja, é, também, normal com a mesma média e variância reduzida à n-ésima parte.

A utilidade da distribuição da média amostral em inferência estatística é evidenciada como segue. Como a média, μ, de uma distribuição normal é, em geral, desconhecida, é de interesse a obtenção de um seu **estimador**, que possa ser usada para a construção de uma aproximação da distribuição exata desconhecida. Esse estimador é construído através da informação provida por uma amostra aleatória da população. Intuitivamente, a média amostral $\overline{X}$ é um "bom" estimador da média populacional μ. Entretanto, a proximidade de $\overline{X}$ em relação a μ depende da distribuição de $\overline{X}$. Como essa distribuição é conhecida exatamente, pode-se determinar, por exemplo, a probabilidade (exata) de que o estimador $\overline{X}$ se situe em um intervalo

de uma certa amplitude em torno de μ, ou seja: $P[\overline{X}-a \leq \mu \leq \overline{X}+b]$, onde a e b são dois números reais, $a<0$ e $b>0$. Se a média μ e o desvio padrão σ^2 fossem conhecidos, esta probabilidade poderia ser determinada pela identidade:

$$P[\overline{X}-a \leq \mu \leq \overline{X}+b] = P[b \leq \overline{X}-\mu \leq a] =$$
$$= P\left[\frac{b\sqrt{n}}{\sigma} \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq \frac{a\sqrt{n}}{\sigma}\right] = P\left[\frac{b\sqrt{n}}{\sigma} \leq Z \leq \frac{a\sqrt{n}}{\sigma}\right],$$

onde $Z \sim N(0,1)$. Esta expressão pode ser avaliada para quaisquer valores de a, b, μ e σ^2, através de consulta à Tabela II.

Exemplo 6.5. Seja X_1, X_2, ...,X_n uma amostra aleatória de dez observações de uma distribuição normal com média μ e variância σ^2. O ponto superior da distribuição N(0,1) que limita uma área 0,025 na cauda direita desta distribuição é $z_{0,025}=1,96$ (Tabela I). Então, pode-se escrever a seguinte sentença probabilista:

$$P\left[-1,96 \leq \frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \leq 1,961\right] = 0,95,$$

que pode ser rescrita como:

$$P\left[\frac{-1,96\sigma}{\sqrt{n}} \leq \overline{X}-\mu \leq \frac{1,96\sigma}{\sqrt{n}}\right] = 0,95$$

Assim, a diferença entre o estimador $\overline{X}$ e o parâmetro estimado, μ, situa-se no intervalo $(-1,96\sigma/\sqrt{n};\ 1,96\sigma/\sqrt{n})$ com probabilidade 0,95; ou, em outras palavras, a probabilidade de que o estimador $\overline{X}$ situe-se a uma distância de μ inferior à $1,96\sigma/\sqrt{n}$ é 0,95.

6.7.3 Distribuição da variância da amostra

A distribuição normal tem dois parâmetros, μ e σ^2. Na **Seção 6.7.2**, foi feita referência à distribuição de $\overline{X}$, que "estima" a característica populacional desconhecida μ. Nesta Seção, procurar-se-á a distribuição da variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$, que estima a variância populacional σ^2.

Para maior simplicidade, suponha-se, inicialmente, que $X_i \sim N(0,1)$. Seja χ^2 ("qui-quadrado") a soma dos quadrados dos valores amostrais. Então, $\chi^2 = X_1 + X_2 + ... + X_n$ é, também, uma variável aleatória, já que é uma função de variáveis aleatórias. O contradomínio desta variável aleatória é o intervalo $[0, \infty)$, pois χ^2 é uma soma de quadrados de variáveis aleatórias com contradomínio $(-\infty, \infty)$. Demonstra-se que a distribuição desta nova variável aleatória é:

$$f_{\chi^2}\left(\chi^2\right) = \frac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)}\left(\chi^2\right)^{\frac{\nu-2}{2}} e^{-\frac{\chi^2}{2}}, \quad \chi^2 > 0,$$

onde o parâmetro ν ($\nu=n$ neste caso) é um número inteiro positivo e $\Gamma(.)$ denota a função gama, definida por:

$$\Gamma\left(\frac{\nu}{2}\right) = \left(\frac{\nu}{2}\right)! = \begin{cases} \left(\frac{\nu}{2}-1\right)\left(\frac{\nu}{2}-2\right)\times\ldots\times 3\times 2\times 1, & \text{para } \nu \text{ par e } \nu>2 \\ \left(\frac{\nu}{2}-1\right)\left(\frac{\nu}{2}-2\right)\times\ldots\times\frac{3}{2}\times\frac{1}{2}\times\sqrt{\pi}, & \text{para } \nu \text{ ímpar e } \nu>2 \end{cases}$$

e

$$\Gamma(1) = 1 \text{ e } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\,.$$

Uma variável aleatória contínua com esta função densidade é denominada uma **variável aleatória qui-quadrado com ν graus de liberdade**. Diz-se, então, que a variável aleatória χ^2 tem **distribuição qui-quadrado com parâmetro** ν, denominado **número de graus de liberdade**. O parâmetro ν é o número de variáveis aleatórias normais independentes cujos quadrados são somados para compor a variável aleatória χ^2. Portanto, ν é determinado pelo tamanho da amostra.

Denota-se que uma variável aleatória W tem distribuição qui-quadrado com ν graus de liberdade por $W \sim \chi^2_\nu$.

Essa distribuição é uma aproximação para as distribuições de variáveis aleatórias que exprimem medidas de características de muitos fenômenos físicos, não necessariamente justificada pela construção que motivou a definição da variável aleatória qui-quadrado como a soma dos quadrados de variáveis aleatórias independentes e com idêntica distribuição N(0,1). De fato, a distribuição qui-quadrado é um caso particular da **distribuição gama**, importante em muitas aplicações.

A representação gráfica da função densidade de χ^2 é apresentada na **Figura 6.4** para os casos específicos correspondentes a ν=1 e ν=3; a representação gráfica para ν=2 é semelhante à da **Figura 6.4.a**; a representação da **Figura 6.4.b** é a mais típica da distribuição χ^2 - as representações gráficas das densidades para ν>3 são semelhantes à esta forma. Observe-se que a forma típica da distribuição χ^2 é assimétrica, em vez de simétrica como a distribuição normal, e que a probabilidade total 1 distribui-se no intervalo (0, ∞). Demonstra-se que a média e a variância da variável aleatória qui-quadrado são dadas por: $E(\chi^2) = \nu$ e $\text{Var}(\chi^2) = 2\nu$. Dessa forma, ambas média e variância populacionais dependem dos graus de liberdade.

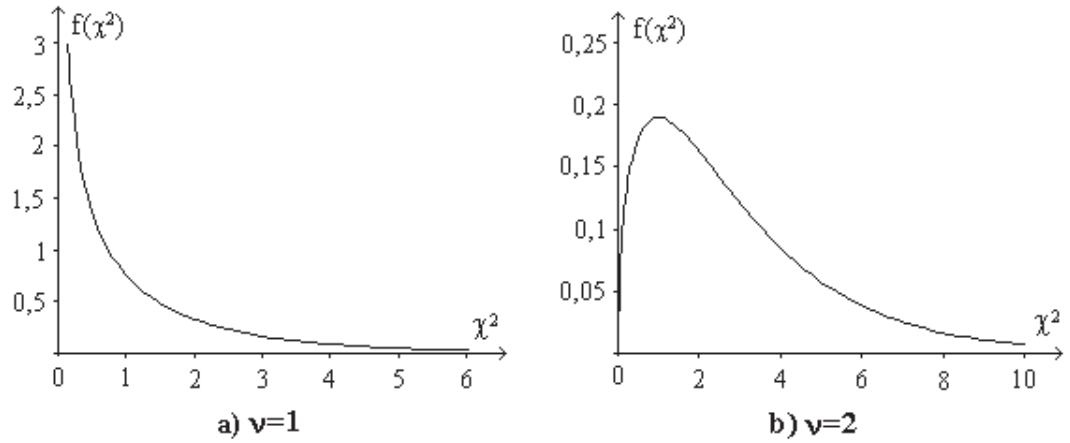


Figura 6.4. Função densidade da variável aleatória χ^2 para a) ν=1 e b) ν=2 graus de liberdade.

Em inferência estatística é freqüentemente de interesse a probabilidade de que χ^2 se situe à direita de um ponto $\chi^2_{\alpha,\nu}$ que limita uma área α na cauda superior da distribuição, expressa por:

$$P[\chi^2 \geq \chi^2_{\alpha,\nu}] = \int_{\chi^2_{\alpha,\nu}}^{\infty} f(\chi^2)\, d\chi^2 = \alpha$$

Essas probabilidades e os correspondentes valores $\chi^2_{\alpha,\nu}$, denominados pontos **100α percentuais superiores** da distribuição qui-quadrado, são disponíveis em tabelas, como a Tabela III do Apêndice. Por exemplo, para ν=5 e α=0,10, tem-se:

$$P[\chi^2 > \chi^2_{0,10;5}] = P[\chi^2 \geq 9{,}236] = 0{,}10.$$

Se X_1, X_2,...,X_n são variáveis aleatórias independentes e normalmente distribuídas com médias μ_1, μ_2, ..., μ_n e variâncias $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectivamente, a variável aleatória $X_1 + X_2 + \ldots + X_n$ não tem distribuição qui-quadrado. Isto porque as variáveis aleatórias X_i, i=1,2,...,n, não satisfazem à condição requerida de média zero e variância unitária. Entretanto, as variáveis aleatórias:

$$Z_i = \frac{X_i - \mu_i}{\sigma_i}, \quad i = 1, 2, \ldots, n,$$

têm distribuição normal padrão e são, também, estatisticamente independentes. Portanto,

$$\chi^2 = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

tem distribuição qui-quadrado com n graus de liberdade.

A distribuição qui-quadrado é uma das poucas distribuições que satisfazem à propriedade da aditividade, de grandes implicações para seu uso em inferência estatística, ou seja:

"Se duas variáveis aleatórias $\chi^2_{\nu_1}$ e $\chi^2_{\nu_2}$ têm distribuições qui-quadrado, com ν_1 e ν_2 graus de liberdade, a soma dessas variáveis aleatórias:

$$\chi^2 = \chi^2_{\nu_1} + \chi^2_{\nu_2},$$

também tem distribuição qui-quadrado, com $\nu = \nu_1+\nu_2$ graus de liberdade".

Este teorema é uma conseqüência imediata da definição da distribuição qui-quadrado. De fato, a distribuição $\chi^2_{\nu_1}$ é equivalente à distribuição da soma de quadrados de ν_1 variáveis independentes e normalmente distribuídas, cada uma com média 0 e variância unitária. Portanto, a variável aleatória $\chi^2_{\nu_1}$ pode ser expressa como:

$$\chi^2_{\nu_1} = X_1^2 + X_2^2 + \ldots + X^2_{\nu_1},$$

onde X_1, $X_2,\ldots,X_{\nu_1}$ são variáveis aleatórias independentes com a mesma distribuição N(0,1). Semelhantemente, $\chi^2_{\nu_2}$ pode ser expressa como:

$$\chi^2_{\nu_2} = Y_1^2 + Y_2^2 + \ldots + Y^2_{\nu_2},$$

onde Y_i, $i=1,2,\ldots,\nu_2$, são variáveis aleatórias independentes com distribuição N(0,1). Então, a variável aleatória χ^2 pode ser escrita como:

$$\chi^2 = X_1 + X_2 + \ldots + X_{\nu_1} + Y_1 + Y_2 + \ldots + Y_{\nu_2},$$

onde os dois conjuntos de variáveis aleatórias X_i, $i=1,2,\ldots, \nu_1$ e Y_i, $i=1,2,\ldots, \nu_2$, são independentes, em conseqüência da independência de $\chi^2_{\nu_1}$ e $\chi^2_{\nu_2}$. Assim, χ^2 é a soma de $\nu_1+\nu_2=\nu$ variáveis aleatórias normal e independentemente distribuídas, cada uma com média zero e variância unitária. Logo, por definição, a variável aleatória χ^2 tem distribuição qui-quadrado com ν graus de liberdade.

Por uma extensão do mesmo argumento, pode-se demonstrar que a propriedade da aditividade da distribuição qui-quadrado estende-se para qualquer número finito de variáveis aleatórias independentes com distribuição qui-quadrado.

Essa propriedade e a anterior definição da distribuição qui-quadrado são as bases para a derivação da distribuição da variância amostral S^2, que segue.

Como a média da distribuição da variância amostral S^2 é igual à variância populacional σ^2, é natural esperar que a variável aleatória:

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

seja um "bom" estimador da variância σ^2 de uma distribuição normal em que a variância σ^2 é desconhecida. Intuitivamente, o processo de estimação é o seguinte: Uma amostra aleatória de n observações é extraída da distribuição normal em que a variância σ^2 é desconhecida; uma realização σ^2 da variável aleatória S^2 é, então, determinada e usada como uma estimativa de σ^2. Essa estimativa será tão "melhor" quanto mais próxima do parâmetro σ^2 se situar. Como essa proximidade não pode ser assegurada para a estimativa obtida de uma amostra particular, busca-se um estimador cuja distribuição de probabilidade se concentre mais em torno de σ^2. Assim, uma medida da proximidade de S^2 em relação a σ^2 , que exprime a qualidade do estimador, é provida

pela probabilidade de que o estimador se situe em um intervalo de certa amplitude em torno de σ^2, ou seja, por:

$$P\left[a\,S^2 \le \sigma^2 \le b\,S^2\right],$$

onde *a* e *b* são dois números reais. Essa probabilidade depende da distribuição de S^2 , ou melhor, da distribuição de S^2/σ^2, que é relacionada à distribuição qui-quadrado.

Demonstra-se que, se $X_1, X_2,...,X_n$ são variáveis aleatórias independentes e com idêntica distribuição normal com média μ e variância σ^2, então, a variável aleatória $(n-1)S^2/\sigma^2$ tem distribuição qui-quadrado com n-1 graus de liberdade. Esse resultado é justificado a seguir.

A variável aleatória $(n-1)S^2/\sigma^2$ é expressa por:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

Observe-se que, se a média amostral $\overline{X}$ fosse substituída pela média populacional μ, a expressão resultante $\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$ teria distribuição qui-quadrado com n graus de liberdade, já que $\frac{X_i - \mu}{\sigma}$, i=1, 2,...n, são variáveis aleatórias independentes e identicamente distribuídas N(0,1). Agora, a variável aleatória $(n-1)S^2/\sigma^2$ pode ser escrita como segue:

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

$$= \frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{\sigma^2} - \frac{\left(\overline{X} - \mu\right)^2}{\sigma^2},$$

donde:

$$\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\sqrt{n}\left(\overline{X} - \mu\right)}{\sigma}\right)^2.$$

Mas

$$\frac{\sum_{i=1}^{n}\left(X_i - \mu\right)^2}{\sigma^2} \sim \chi^2_n \quad \text{e} \quad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1);$$

portanto, por definição, $\frac{\left(\overline{X} - \mu\right)^2}{\sigma^2/n} \sim \chi^2_1$.

Por outro lado, demonstra-se que S^2 e $\overline{X}$ são variáveis aleatórias independentes, o que implica que os dois termos do segundo membro da última igualdade são variáveis aleatórias

independentes. Assim, o primeiro membro dessa igualdade tem distribuição χ^2_n e o segundo membro é a soma de duas variáveis aleatórias independentes, cujo segundo termo $n(\overline{X}-\mu)^2/\sigma^2 \sim \chi^2_1$. Portanto, como conseqüência da aditividade da variável aleatória χ^2, $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$.

Uma ilustração do uso da distribuição dessa variável aleatória é provida **pelo Exemplo 6.6**.

Exemplo 6.6. Seja $X_1, X_2, ...,X_n$ uma amostra aleatória de dez observações de uma distribuição $N(\mu,\sigma^2)$. O ponto percentual superior da distribuição χ^2_9 que limita uma área na cauda direita de 0,05, provido pela Tabela III, é 16,919. Como $9S^2/\sigma^2$ tem distribuição qui-quadrado com 9 graus de liberdade, pode-se escrever:

$$P\left[\frac{9S^2}{\sigma^2} \leq 16{,}919\right] = 0{,}95,$$

que pode ser reescrita como:

$$P\left[\sigma^2 \geq \frac{9S^2}{16{,}919}\right] = 0{,}95.$$

Esta é, realmente, uma sentença probabilista referente à variância amostral S^2 e não à variância populacional σ^2, já que S^2 é uma variável aleatória e σ^2 é uma constante. O evento a que corresponde a probabilidade 0,95 é o intervalo de amplitude infinita com extremo inferior aleatório $9S^2/16{,}919$ (por depender da variável aleatória S^2). A sentença deve ser entendida como "a probabilidade de que o intervalo aleatório $(9S^2/16{,}919; \infty)$ inclua o parâmetro desconhecido σ^2 é 0,95".

Em capítulo ulterior, será visto que sentenças probabilistas desse tipo provém **estimadores de intervalo** (também denominados **intervalos de confiança**) para a variância populacional. Esta sentença probabilista também pode ser usada para testes de hipóteses referentes ao parâmetro σ^2. Por exemplo, para determinar se a variância é igual à um valor particular σ^2_0; ou seja, se a variância é igual a σ^2_0, quando $9S^2/\sigma^2$ deve ser usualmente menor do que 16,919.

A designação "graus de liberdade" pode ser interpretada como o número de termos independentes que são somados para formar a estatística qui-quadrado. A soma $\sum_{i=1}^{n}\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2$ compreende n termos independentes, mas a soma $\sum_{i=1}^{n}\left(\frac{X_i-\overline{X}_i}{\sigma_i}\right)^2$ tem apenas n-1 termos independentes, já que a relação $\sum_{i=1}^{n}(X_i-\overline{X})=0$ determina qualquer um dos desvios, dados os outros n-1 desvios.

6.7.4 Distribuição t de Student

Na **Seção 6.7.2**, foi mostrado que, se $X_1, X_2, ...,X_n$ é uma amostra aleatória da distribuição $N(\mu,\sigma^2)$, então a média amostral padronizada, ou seja, $\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}$ tem distribuição N(0,1). Esta distribuição é útil em inferências referentes à média μ quando a variância σ^2 é conhecida. Por exemplo, a sentença probabilista do **Exemplo 6.5**:

$$P\left[-\frac{1{,}96\sigma}{\sqrt{n}} \le \overline{X}-\mu \le \frac{1{,}96\sigma}{\sqrt{n}}\right] = 0{,}95,$$

é útil se o desvio padrão σ é conhecido, como pode ser o caso se informações anteriores são disponíveis. Entretanto, usualmente, σ é desconhecido. Nessa situação, parece natural utilizar uma estimativa S em lugar do parâmetro σ. Ocorre, todavia, que a variável aleatória resultante $\frac{\sqrt{n}(\overline{X}-\mu)}{S}$ não tem distribuição normal. Uma solução é provida pela estatística T, definida a seguir.

Se $X_1, X_2, ...,X_n$ é uma amostra aleatória de uma distribuição $N(\mu,\sigma^2)$, então a variável aleatória:

$$T = \frac{\sqrt{n}(\overline{X}-\mu)}{S}$$

tem **distribuição t** (de Student) com n-1 graus de liberdade, o que é denotado por: $T \sim t_{n-1}$

Observe-se que essa variável aleatória é, tradicionalmente, denotada pela letra minúscula t, contrariando à notação usual de letra maiúscula para variável aleatória.

A função densidade da variável aleatória T com ν graus de liberdade tem a seguinte expressão:

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad -\infty < t < \infty,$$

onde o parâmetro ν é um número inteiro positivo e $\Gamma(.)$ denota a função gama, definida na **Seção 6.7.3**. A representação geométrica da variável aleatória T é ilustrada na **Figura 6.5** para ν=1, ν=5 e ν=10.

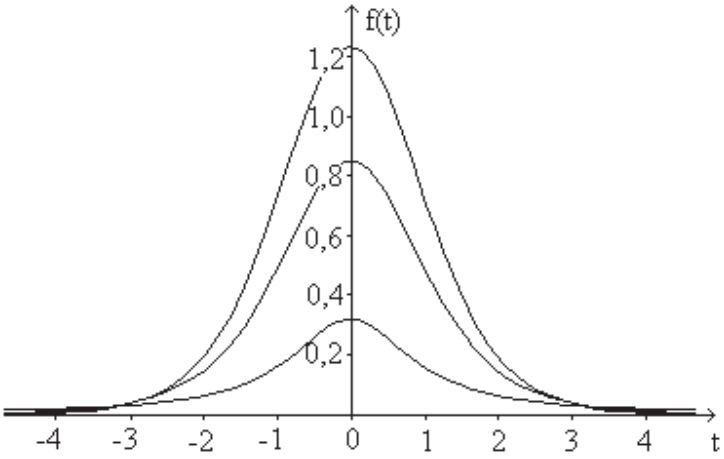


Figura 6.5. Distribuição de probabilidade da variável aleatória t para 3 valores particulares do parâmetro ν: 1, 5 e 10.

A curva da distribuição t tem a forma de sino e é simétrica em relação à origem 0, à semelhança da distribuição normal padrão. De fato, à medida em que o número de graus de liberdade ν tende à infinito, a distribuição t aproxima-se da distribuição normal padrão. A média da variável aleatória T é $E(T)=0$ para $\nu>1$, e a variância, $\text{Var}(T)= \frac{\nu}{\nu-2}$ para $\nu>2$. Probabilidades referentes à variável aleatória T de interesse em inferência estatística são disponíveis em tabelas, como a Tabela IV do Apêndice. A Tabela IV apresenta os pontos **100α-percentuais superiores $t_{\alpha;\nu}$**, que limitam uma área a na cauda direita, que satisfazem à condição:

$$P[T \geq t_{\alpha;\nu}] = \alpha.$$

Por exemplo, para $\nu=5$ graus de liberdade:

$$P[T \geq t_{0,05;5}] = P[T>2,015] = 0,05 .$$

Para $\nu>30$, $t_{\alpha;\nu}$ é próximo do ponto 100α-percentual da distribuição normal padrão, z_α, exceto para valores extremos de α, ou seja, valores de α muito próximos de 0 ou 1. Como a distribuição t é simétrica em relação à origem, valores $t_{\alpha:\nu}$ para $\alpha \geq 0,50$ podem ser obtidos pela relação $t_{\alpha:\nu} = - t_{1-\alpha;\nu}$.

A variável aleatória T pode ser utilizada em inferências estatísticas referentes à média populacional μ quando a variância populacional σ^2 é desconhecida.

Exemplo 6.7. Suponha-se, na situação do **Exemplo 6.5**, que a variância populacional σ^2 não é conhecida e que em seu lugar utiliza-se seu estimador S^2 correspondente a uma amostra aleatória de dez observações. Considerando o ponto da distribuição t que limita uma área de 0,025 na causa superior, pode-se escrever a seguinte sentença probabilista:

$$P\left[- t_{0,025;\,9} \leq \frac{\sqrt{10}\left(\overline{X} - \mu\right)}{S} \leq t_{0,025;\,9} \right] = 0,95,$$

ou seja:

$$P\left[-2{,}262 \le \frac{\sqrt{10}\left(\overline{X}-\mu\right)}{S} \le 2{,}262\right] = 0{,}95$$

que pode ser rescrita como:

$$P\left[\frac{-2{,}262\ S}{\sqrt{10}} \le \left(\overline{X}-\mu\right) \le \frac{2{,}262\ S}{\sqrt{10}}\right] = 0{,}95$$

Assim, a probabilidade de que a média amostral $\overline{X}$ se situe a uma distância $2{,}262\,S/\sqrt{10}$ da correspondente média populacional estimada μ é 0,95. A comparação desse valor com aquele obtido na situação em que a variância era suposta conhecida (**Exemplo 6.5**), ou seja, $1{,}960\sigma/\sqrt{10}$, mostra que a distribuição de $\overline{X}$ é mais concentrada em torno da média μ quando a variância σ^2 é conhecida do que quando ela é desconhecida e substituída por seu estimador S^2.

A distribuição de $\frac{\sqrt{n}(\overline{X}-\mu)}{S}$ também pode ser usada para teste da hipótese de que μ é igual a um valor específico μ_o quando a variância σ^2 é desconhecida. Se $\mu = \mu_0$, espera-se que a variável aleatória $\sqrt{n}(\overline{X}-\mu)/S$ se situe no intervalo $(-t_{\alpha/2;n-1};\ t_{\alpha/2;n-1})$ com probabilidade $1-\alpha$. Desse modo, para uma probabilidade $1-\alpha$ grande, um valor dessa variável aleatória fora desse intervalo é uma indicação de que μ não é igual a μ_0.

A distribuição t pode ser definida em um contexto mais geral do que o de amostragem aleatória de uma população normal, como segue:

Sejam X e χ^2 duas variáveis aleatórias independentes, definidas no espaço básico de um experimento aleatório, tais que X tem distribuição normal padrão e χ^2 tem distribuição qui-quadrado com ν graus de liberdade. Então, a variável aleatória:

$$T = \frac{X}{\sqrt{\chi^2_\nu/\nu}}$$

tem **distribuição t** (de Student) **com ν graus de liberdade**. A definição anterior é um caso particular dessa definição mais geral.

De fato, se $X_1, X_2, ..., X_n$ é uma amostra aleatória de uma distribuição $N(\mu,\sigma^2)$, então:

$$\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma} \sim N(0,1) \quad \text{e} \quad \frac{(n-1)\,S^2}{\sigma^2} \sim \chi^2_{n-1},$$

onde: $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$. Logo, pela definição mais geral,

$$T = \frac{\frac{\sqrt{n}(\overline{X}-\mu)}{\sigma}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}}}$$

$$= \frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim T_{n-1},$$

o que concorda com a primeira definição da variável aleatória T.

A definição geral da distribuição t é a base para a derivação de estatísticas úteis em inferências referentes às médias de duas ou mais distribuições normais. Para ilustração, considerar-se-á a estatística relacionada com a comparação de dois "tratamentos" em uma pesquisa experimental. Suponha-se que uma amostra aleatória de n_X unidades é extraída da população que usa um tratamento padrão e uma amostra aleatória de n_Y unidades é extraída da população que utiliza um novo tratamento. A comparação dos dois tratamentos é julgada com base na magnitude do valor observado da variável aleatória $\bar{X} - \bar{Y}$ nas duas amostras. Assim, para efetuar decisões a respeito da eficácia do novo tratamento comparativamente ao tratamento padrão, é necessário examinar a distribuição de $\bar{X} - \bar{Y}$.

Sejam $X_1, X_2, ..., X_{n_X}$ variáveis aleatórias independentes e com idêntica distribuição $N(\mu_X, \sigma^2)$ e $Y_1, Y_2, ..., Y_{n_Y}$ variáveis aleatórias independentes com idêntica distribuição $N(\mu_Y, \sigma^2)$ e suponha-se que os dois conjuntos de variáveis aleatórias são independentes. Então, conforme visto anteriormente,

$$\bar{X} = \frac{1}{n_X}\sum_{i=1}^{n_X} X_i \sim N(\mu_X, \sigma^2/n_X); \quad \bar{Y} = \frac{1}{n_Y}\sum_{i=1}^{n_X} Y_i \sim N(\mu_Y, \sigma^2/n_Y);$$

$$\frac{(n_X-1)S_X^2}{\sigma^2} = \frac{\sum_{i=1}^{n_X}(X_i-\bar{X})^2}{\sigma^2} \sim \chi^2_{n_X-1} \quad \text{e} \quad \frac{(n_Y-1)S_Y^2}{\sigma^2} = \frac{\sum_{i=1}^{n_Y}(Y_i-\bar{Y})^2}{\sigma^2} \sim \chi^2_{n_Y-1}.$$

Segundo propriedade de combinação linear de variáveis aleatórias com distribuição normal,

$$\bar{X} - \bar{Y} \sim N\left(\mu_X - \mu_Y, \frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}\right)$$

Então:

$$\frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}}$$

tem distribuição normal com média zero e variância um. Se a variância comum das duas populações σ^2 é conhecida, o problema está resolvido. Entretanto, para as situações mais usuais, em que σ^2 é desconhecida, a solução alternativa é a que segue.

De acordo com a propriedade da aditividade da distribuição qui-quadrado, a variável aleatória:

$$\frac{(n_X - 1)S_X^2}{\sigma^2} + \frac{(n_Y - 1)S_Y^2}{\sigma^2}$$

tem distribuição qui-quadrado com $n_X + n_Y - 2$ graus de liberdade e, ademais, é distribuída independentemente de $\overline{X} - \overline{Y}$. Então, segundo a definição da distribuição t, a variável aleatória:

$$\frac{\left[(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)\right] \Big/ \sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}\sigma}{\left(\frac{(n_X-1)S_X^2}{\sigma^2}+\frac{(n_Y-1)S_Y^2}{\sigma^2}\right) \Big/ \sqrt{n_X+n_Y-2}} = \frac{(\overline{X}-\overline{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{(n_X-1)S_X^2+(n_Y-1)S_Y^2}{n_X+n_Y-2}}\sqrt{\left(\frac{1}{n_X}+\frac{1}{n_Y}\right)}}$$

tem distribuição t com $n_X + n_Y - 2$ graus de liberdade.

6.7.5 Distribuição F

Na Seção anterior, foi apresentada a estatística T para a comparação das médias de duas populações normais com igual variância. Também é muito freqüente em inferência estatística a situação de comparação das variâncias de duas populações normais. Uma dessas situações é, naturalmente, a comparação de variâncias para verificação da condição de igualdade de variâncias necessária para a utilização da referida estatística T. O processo natural é a extração de uma amostra aleatória de cada uma das duas populações e, então, comparar as variâncias amostrais, o que pode ser procedido através do quociente S_X^2/S_Y^2. Se essa razão é próxima de um, as duas variâncias populacionais são julgadas iguais; caso contrário, não. Para efetuar decisões adequadas e quantificar a sentença "próxima de um", é necessário basear o processo de decisão na distribuição da estatística S_X^2/S_Y^2.

Sejam $X_1, X_2, ..., X_{n_X}$ uma amostra aleatória de uma população P_X de tamanho n_X, ou seja, n_X variáveis aleatórias independentes e normalmente distribuídas com média μ_X e variância σ_X^2, e $Y_1, Y_2, ..., Y_{n_Y}$ uma amostra aleatória de uma população P_Y de tamanho n_Y, ou seja, n_Y variáveis aleatórias independente e normalmente distribuídas, cada uma com média μ_Y e variância σ_Y^2, e suponha-se que essas $n_X + n_Y$ variáveis são independentes. Então, a distribuição de:

$$\frac{(n_X-1)S_X^2}{\sigma_X^2} = \frac{\sum_{i=1}^{n_X}(X_i - \overline{X})^2}{\sigma_X^2} \sim \chi^2_{n_X-1}$$

e, semelhantemente,

$$\frac{(n_Y-1)S_Y^2}{\sigma_Y^2} = \frac{\sum_{i=1}^{n_Y}(Y_i - \overline{Y})^2}{\sigma_Y^2} \sim \chi^2_{n_Y-1}$$

Além disso, estas duas variáveis qui-quadrado são independentes, já que ambos os conjuntos de variáveis aleatórias $X_1, X_2, ..., X_{n_X}$ e $Y_1, Y_2, ..., Y_{n_Y}$ são independentes. Então, a razão:

$$\frac{\dfrac{(n_X - 1)S_X^2}{\sigma_X^2} \Big/ (n_X - 1)}{\dfrac{(n_Y - 1)S_Y^2}{\sigma_Y^2} \Big/ (n_Y - 1)} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

é uma variável aleatória com **distribuição F** com $n_X - 1$ e $n_Y - 1$ graus de liberdade.

A distribuição F pode ser definida, de modo mais geral, como segue:

> Se χ_1^2 e χ_2^2 são duas variáveis aleatórias qui-quadrado independentes com ν_1 e ν_2 graus de liberdade, respectivamente, então, a variável aleatória $F = \dfrac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2}$ tem **distribuição F com ν_1 e ν_2 graus de liberdade**, nesta ordem.

O contradomínio dessa variável aleatória é o intervalo (0, ∞), já que as variáveis aleatórias χ_1^2 e χ_2^2 são ambas não negativas. A função densidade de probabilidade da variável aleatória F tem a seguinte expressão:

$$f_F(f) = \frac{\Gamma\left(\dfrac{\nu_1 + \nu_2}{2}\right) \nu_1^{\nu_1/2} \nu_2^{\nu_2/2}}{\Gamma\left(\dfrac{\nu_1}{2}\right)\Gamma\left(\dfrac{\nu_2}{2}\right)} \frac{f^{(\nu_1/2)-1}}{(\nu_2 + \nu_1 f)^{(\nu_2+\nu_2)/2}}, \qquad f > 0$$

onde os parâmetros ν_1 e ν_2 são números inteiros positivos e $\Gamma(.)$ denota a função gama, definida na **Seção 6.7.3**.

Uma variável aleatória contínua com essa função densidade de probabilidade é denominada uma **variável aleatória F com ν_1 e ν_2 graus de liberdade**; ou, alternativamente, é dita ter uma **distribuição F com ν_1 e ν_2 graus de liberdade**.

A representação geométrica da função de densidade da variável aleatória F é ilustrada na **Figura 6.6** para três combinações de graus de liberdade.

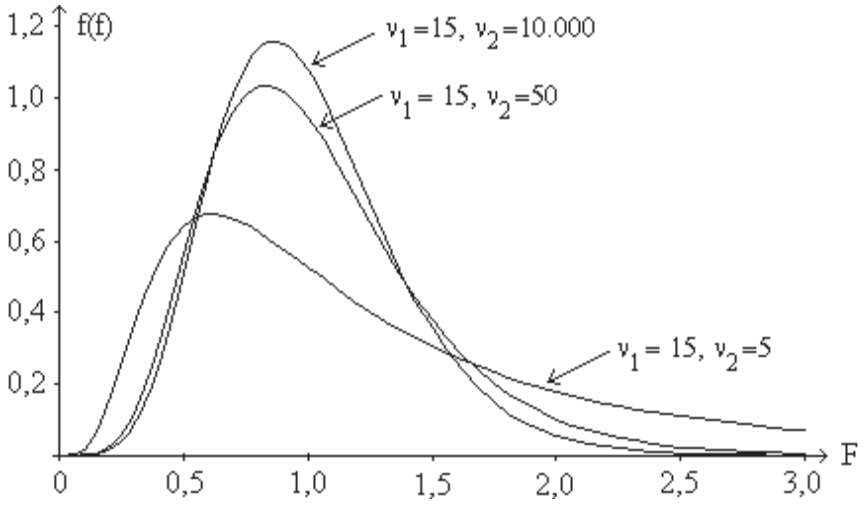


Figura 6.6. Distribuição de probabilidade da variável aleatória F para 3 diferentes combinações de graus de liberdade: ν_1=15, ν_2=5; ν_1=15, ν_2=50 e ν_1=15, ν_2=10.000.

Como a distribuição qui-quadrado, a distribuição F é assimétrica e distribui a probabilidade total 1 no intervalo infinito à direita da origem. A média e a variância da variável aleatória F são dadas por:

$$E(\mathrm{F}) = \frac{\nu_2}{\nu_2 - 2}, \quad \text{para } \nu_2 > 2 \text{ e}$$

$$\mathrm{Var}(\mathrm{F}) = \frac{2\nu_2^2(\nu_2 + \nu_1 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \quad \text{para } \nu_2 > 4$$

Em inferência estatística são usualmente de interesse probabilidades correspondentes à áreas sob a curva na cauda direita da distribuição F (**Figura 6.7**). Ou seja, probabilidades correspondentes a **pontos 100α-percentuais superiores da distribuição F**, que satisfazem à condição:

$$P[F > F_{\alpha;\nu_1;\nu_2}] = \alpha .$$

Como a distribuição F depende de dois parâmetros ν_1 e ν_2, uma tabela de três entradas é necessária para a tabulação dos valores de F correspondentes a diferentes probabilidades e valores de ν_1 e ν_2. A Tabela V do Apêndice é uma dessas tabelas.

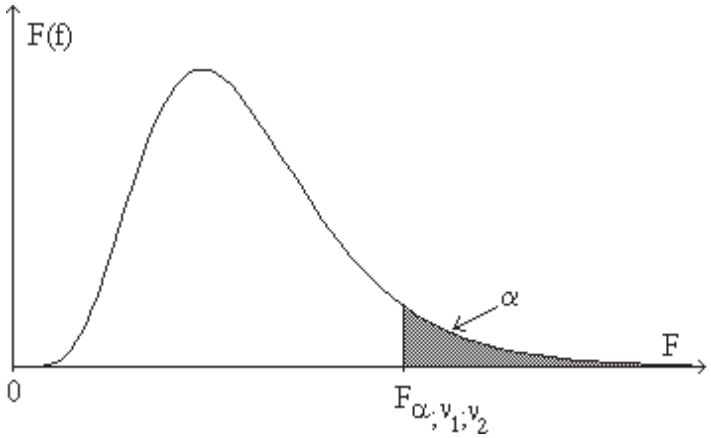


Figura 6.7. Pontos 100α–percentuais superiores da distribuição F.

Observe-se que, como a função densidade da variável aleatória F é assimétrica em ν_1 e ν_2, a ordem em que os graus de liberdade são dados é importante. Os graus de liberdade da variável χ^2 do numerador da estatística F são citados antes, seguidos dos graus de liberdade da variável χ^2 do denominador.

Pontos percentuais para valores de α>0,50 podem ser obtidos pela relação:

$$P\left[\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \geq F_{\alpha;\nu_1;\nu_2}\right] = P\left[\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} < \frac{1}{F_{\alpha;\nu_1;\nu_2}}\right] = \alpha \ .$$

Isso implica que:

$$P\left[\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \geq \frac{1}{F_{\alpha;\nu_1;\nu_2}}\right] = 1-\alpha \ .$$

Entretanto, $\left(\chi_2^2/\nu_2\right)/\left(\chi_1^2/\nu_1\right)$ também tem distribuição F mas com ν_2 e ν_1 graus de liberdade. Assim,

$$P\left[\frac{\chi_1^2/\nu_1}{\chi_2^2/\nu_2} \geq F_{1-\alpha;\nu_2;\nu_1}\right] = 1-\alpha$$

Portanto,

$$F_{1-\alpha;\nu_2;\nu_1} = \frac{1}{F_{\alpha;\nu_1;\nu_2}}$$

ou

$$F_{\alpha;\nu_1;\nu_2} = \frac{1}{F_{1-\alpha;\nu_2;\nu_1}}.$$

Por exemplo, o ponto percentual para α=0,95, ν_1=6 e ν_2=8, ou seja, $F_{0,95;6;8}$, pode ser obtido como segue:

$$F_{0,95;6;8} = \frac{1}{F_{0,05;8;6}} = \frac{1}{4,15} = 0,241.$$

Observe-se que, se ν_1=1, tem-se $F_{1;\nu_2} = T^2_{\nu_2}$, e, se ν_2=∞, $F_{\nu_1;\infty} = \chi^2_{\nu_1} / \nu_1$.

6.8 Exercícios

1. O supervisor de uma indústria de peças examina um décimo das unidades que saem da linha de produção. Liste algumas condições sob as quais este método pode não produzir uma amostra aleatória.
2. Uma amostra aleatória de tamanho 64 é tomada de uma população normal com média μ=51,4 e desvio padrão σ=6,8. Determine a probabilidade de que a média da amostra $\overline{X}$:
   a) exceda 53,1;
   b) situe-se no intervalo entre 50,4 e 52,4;
   c) seja menor que 50,8.
3. Suponha que o conteúdo de umidade por quilograma de um concentrado de proteína desidratada tem distribuição normal com média 3,5 e desvio padrão 0,5. Uma amostra aleatória de 16 porções de um quilograma é extraída dessa proteína desidratada. Seja $\overline{X}$ a média dessa amostra de medidas de conteúdo de umidade.
   a) Qual é a distribuição de $\overline{X}$? Qual é a média e a variância dessa distribuição?
   b) Determine a probabilidade de que: i) $\overline{X}$ exceda 3,7; $\overline{X}$ situe-se entre 3,34 e 3,66.
4. Uma amostra aleatória de tamanho 100 é extraída de uma população normal com variância 625. Qual é a probabilidade de que a média da amostra difira da média da população por 4 ou mais unidades?
5. São tomadas amostras independentes de tamanho 40 de duas populações normais com mesma média e variâncias σ_1^2=20 e σ_2^2=60. Qual é a probabilidade de que a diferença entre as médias das duas amostras seja menor do que 2?
6. São obtidas amostras aleatórias de tamanhos n_1 e n_2 de duas populações normais com médias μ_1 e μ_2 e variâncias As médias e variâncias σ_1^2=150 e σ_2^2=200. Qual é a probabilidade de que a média da primeira amostra exceda a média da segunda amostra em pelo menos 4,5 unidades?
7. A distribuição da percentagem de gordura do leite de vacas holandesas em uma região em um período de 5 anos foi aproximadamente normal com média 3,9 e desvio padrão 0,45.
   a) Qual é a proporção das vacas que tiveram percentagem de gordura no leite menor que 3?
   b) Qual é a proporção das vacas que tiveram percentagem de gordura no leite maior que 4,5?
   c) Determine o intervalo em que se situaram 90% das percentagens de gordura do leite.

8. Suponha que uma amostra aleatória de 25 vacas é selecionada da população de vacas holandesas da região referida na questão anterior.
   a) Descreva a distribuição da percentagem média de gordura do leite dessa amostra de 25 vacas.
   b) Compare a distribuição da média da amostra considerada no item anterior com a distribuição da média de uma amostra de 25 vacas do mesmo rebanho.
   c) Qual é a probabilidade de que a média da percentagem de gordura do leite da amostra considerada no item a exceda 4%.
9. Sabe-se que a aplicação da dose padrão de uma droga resulta na alteração média de 250 unidades no nível de ácido úrico de uma pessoa, com desvio padrão de 50 unidades. Uma nova dose da droga deve ser testada experimentalmente em 75 indivíduos, dos quais será medida a alteração de seus níveis de ácido úrico. A conjetura a ser testada é que a nova dose aumenta a alteração média do nível de ácido úrico da população em 20 unidades, mas sem alterar o desvio padrão.
   a) Supondo que a conjetura é verdadeira, qual é a probabilidade de que a média da amostra para 75 indivíduos seja menor que 260 ou maior que 280 unidades?
   b) Se além de aumentar a média em 20 unidades, a nova dose altera o desvio padrão de 50 para 64, qual é a probabilidade de que a média da amostra seja menor que 260 ou maior que 280?
10. Uma amostra aleatória de tamanho 64 é tomada de uma população infinita com média μ=112 e variância σ^2=144. Use o teorema central do limite para determinar a probabilidade de que a média da amostra resulte maior que 114,5.
11. Uma amostra aleatória de tamanho 100 é extraída de uma população infinita com média μ=53 e variância σ^2=444. Qual é a probabilidade de que a média da amostra situe-se entre 50 e 56?
12. Deverão ser efetuadas 80 medidas repetidas do ponto de fusão de uma nova liga metálica. Pelo conhecimento da precisão do aparelho de medida, sabe-se que o desvio padrão das medidas é 7 graus. Qual é a probabilidade de que a média da amostra de 80 medidas não se desvie do verdadeiro ponto de fusão por mais do que 1,54 graus?
13. A variável que exprime uma característica das unidades de uma população tem distribuição normal com variância σ^2=2.500. Determine a probabilidade de que a variância de uma amostra aleatória de tamanho 26 dessa variável aleatória situe-se entre 48 e 95.
14. A afirmativa de que a variância de uma população normal é σ^2=4 deve ser rejeitada se a variância de uma amostra aleatória de tamanho 9 excede 8,7675. Qual é a probabilidade de que essa afirmativa seja rejeitada mesmo que σ^2=4?
15. A afirmativa de que a variância de uma população normal é σ^2=25 deve ser rejeitada se a variância de uma amostra aleatória de tamanho 16 excede 41,66 ou é menor do que 7,67. Qual é a probabilidade de que a afirmativa referente à variância populacional seja correta?

16. Com base em amostras aleatórias independentes de tamanhos 10 de duas populações normais que provém as estimativas de variâncias s_1^2=12,8 e s_2^2=3,2, como se pode concluir sobre a afirmativa de que as duas populações têm a mesma variância?
17. Uma amostra aleatória de tamanho 25 de uma população normal tem média $\overline{x}$=47 e desvio padrão s=6. Com base na estatística T de Student, pode-se dizer que a informação provida pela amostra suporta a hipótese de que a média da população é μ=42?
18. A média e a variância de uma amostra aleatória de tamanho 9 de uma população normal são $\overline{x}$=27,7 e s^2 =3,24. Com base na estatística t, pode-se dizer essa informação suporta a afirmativa de que a média da população é μ=28,5?
19. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando as letras V e F entre parêntesis, respectivamente. Se a sentença for falsa, explique porque.
    ( ) Amostragem completamente aleatória e amostragem aleatória simples são dois diferentes métodos de amostragem.
    ( ) Um instrumento de loteria pode ser um modo aceitável de obter amostras completamente aleatórias.
    ( ) Quando se usa uma tabela de números aleatórios para selecionar uma amostra começa-se, sempre, pelo início da tabela.
    ( ) A escolha do delineamento de amostra não afeta a escolha do procedimento de análise estatística a usar.
    ( ) Se possível, amostras de tamanho maior do que um devem ser usadas para propósitos de inferência.
    ( ) Determinar um tamanho apropriado da amostra é nunca um problema quando se usam técnicas estatísticas.
    ( ) Para muitas amostras aleatórias a média amostral não é igual à média da população da qual a amostra foi extraída.
    ( ) Uma média amostral é calculada da mesma maneira que uma média de população.
    ( ) Uma variância amostral é calculada da mesma maneira que uma variância de população.
    ( ) Se uma população tem média 10 e desvio padrão 2, então a distribuição amostral das médias de amostras de tamanho n=2 tem média 10 e desvio padrão 1.
    ( ) A variância de uma distribuição amostral de médias é maior do que a variância da população.
    ( ) A distribuição da média $\overline{X}$ de uma amostra aleatória de X, quando X tem distribuição normal, tem desvio padrão menor que o de X, se n>1.
    ( ) Se amostras aleatórias de tamanho fixo n são extraídas de uma população normal, a distribuição das médias amostrais permanece normal quando n cresce.
    ( ) O pico da distribuição normal de $\overline{X}$ é cada vez menos pronunciado, na medida em que o tamanho da amostra n cresce.
    ( ) A média de qualquer amostra aleatória é mais provável situar-se mais próxima da média da população na medida em que o tamanho da amostra cresce.

( ) $\bar{X}$ é uma variável aleatória quando possui uma distribuição, e é um número quando toma um valor particular, denotado por $\bar{x}$.

( ) De acordo com o teorema central do limite, se n é grande, a distribuição amostral de médias é bem aproximada por uma distribuição normal.

( ) O teorema central do limite pode ser aplicado apenas a distribuições simétricas.

( ) A distribuição normal é uma ferramenta estatística valiosa em virtude do teorema central do limite.

( ) Se uma população tem variância igual a 12, então a variância das médias de todas as amostras de tamanho 3 extraídas aleatoriamente da população será igual a 4.

( ) A média de uma amostra aleatória de qualquer população tem distribuição normal com média m e variância σ^2.

( ) O "erro padrão da média" é o desvio padrão das médias de todas as amostras de um tamanho fixo de uma população.

( ) As outras condições permanecendo as mesmas, para reduzir o erro padrão da média pela metade é necessário dobrar o tamanho da amostra.

( ) Quando n é grande e p é próximo de 0,5, a distribuição binomial é aproximadamente uma distribuição normal.

( ) Há apenas uma distribuição χ^2.

( ) A estatística χ^2 não tem uma distribuição contínua, mas a distribuição contínua atribuída a Helmert fornece sentenças de probabilidade confiáveis.

( ) Muitas populações que ocorrem naturalmente podem ser modeladas pela distribuição χ^2.

( ) O valor esperado e a variância de uma distribuição χ^2 são iguais.

( ) Para cada número de graus de liberdade inteiro positivo, há uma diferente distribuição de t.

( ) O denominador de uma estatística T é algumas vezes chamado "erro padrão".

( ) A proporção da área da distribuição de t à direita de um valor de t especificado na cauda direita é maior do que a proporção além do mesmo valor na distribuição normal padrão.

( ) Há muitas distribuições de F diferentes, uma para cada par ordenado de números de graus de liberdade.

( ) As distribuições de F são simétricas.

( ) Todos os valores de F são não negativos.

7 INFERÊNCIA ESTATÍSTICA

Conteúdo

7.1 Introdução....163
7.2 Estimação por Ponto....164
7.2.1 Conceitos....164
7.2.2 Propriedades de um estimador....165
7.2.3 Distribuição amostral da média de uma população normal....165
7.2.4 Distribuição amostral da variância de uma população normal....166
7.3 Estimação por Intervalo - Intervalo de Confiança....166
7.4 Teste de Hipótese....167
7.4.1 Conceitos....167
7.4.2 Testes de hipóteses referentes à media de uma população normal....170
7.4.3 Teste da hipótese de igualdade das médias de duas populações normais....172
Populações com mesma variância....172
7.4.4 Teste unilateral e teste bilateral....174
7.5 Exercícios....174

7.1 Introdução

A **inferência estatística** ocupa-se das técnicas que permitem, a partir da observação de valores da variável aleatória em uma amostra aleatória de uma população, inferir sobre a natureza da distribuição que governa a população.

Uma **amostra aleatória** de uma população é uma parte da população gerada por processo que atribui a cada indivíduo da população a mesma probabilidade de integrar a amostra.

O procedimento de inferência estatística consiste no seguinte:

1) Postula-se que a distribuição que governa uma população é um membro de uma família de distribuições indexada por um ou mais parâmetros; isto é, que a distribuição da variável aleatória X é um membro de uma família de distribuições caracterizada por valores específicos desconhecidos de um ou mais parâmetros.
2) Com base em uma amostra aleatória de tamanho n da população, isto é, de uma realização particular $(x_1, x_2,...,x_n)$ de n variáveis aleatórias $X_1, X_2,...,X_n$, independentes e com mesma probabilidade, identifica-se um membro particular da família.

3) A partir daí, procede-se como se a distribuição identificada fosse a verdadeira distribuição que governa a população.

Conseqüências: O resultado da inferência dependerá da amostra e variará de amostra para amostra. Logo, a inferência está sujeita a erro.

A inferência estatística clássica se ocupa, basicamente, de três classes de problemas:

- Estimação por ponto,
- Estimação por intervalo, e
- Teste de hipótese

7.2 Estimação por Ponto

7.2.1 Conceitos

A estimação por ponto é o processo de inferência que determina um valor particular para uso em lugar do parâmetro (desconhecido), através de uma função dos elementos da amostra.

Um **estimador** de um parâmetro θ é uma função $t(X_1, X_2,...,X_n)$ dos elementos da amostra aleatória $X_1, X_2,...,X_n$ que atribui valores para "aproximação" de um parâmetro.

Uma **estimativa** de um parâmetro é o valor do estimador correspondente a uma realização particular $(x_1, x_2,...,x_n)$ das n variáveis aleatórias $X_1, X_2,...,X_n$ independente e identicamente distribuídas, gerada por uma amostra da população.

Exemplo 7.1. Estimadores e estimativas dos parâmetros da distribuição normal.

Um estimador da média μ de uma população com distribuição normal é:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n};$$

uma sua estimativa particular, obtida de uma amostra $(x_1, x_2,...,x_n)$ é:

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Um estimador da variância σ^2 é:

$$S^2 = \frac{1}{n-1}\left[(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2\right)\Big]$$

$$= \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

e uma estimativa particular:

$$s^2 = \sum_i \left(x_i - \overline{x}\right)^2$$

Assim, para uma população normal, tem-se:

Parâmetro	Nome	Estimador	Estimativa
μ	Média	$\overline{X}$	$\bar{x}$
σ^2	Variância	S^2	s^2

7.2.2 Propriedades de um estimador

Um estimador $t(X_1, X_2,...,X_n)$ de um parâmetro θ que indexa a família de distribuições postulada para uma população é **não tendencioso** ou **não viesado** se a média das correspondentes estimativas determinadas para todas as possíveis amostras aleatórias extraídas da população for igual ao (verdadeiro) valor do parâmetro, ou seja, se $E[t(X_1, X_2,...,X_n)] = \theta$.

Um estimador $t(X_1, X_2,...,X_n)$ é denominado de **variância mínima** do parâmetro θ se ele for o estimador de menor variância entre todos os estimadores desse parâmetro, ou seja, se: $Var[t(X_1, X_2,...,X_n)] \leq Var[t'(X_1, X_2,...,X_n)]$, para qualquer outro estimador $t'(X_1, X_2,...,X_n)$ de θ.

O **melhor estimador** ou **estimador ótimo** de um parâmetro θ é seu estimador não tendencioso e de variância mínima.

Exemplo 7.2. Os estimadores não tendenciosos e de variância mínima dos parâmetros μ e σ^2 da distribuição normal são a média e a variância amostrais, ou seja $\overline{X}$ e S^2, respectivamente.

De fato, pode-se demonstrar que:

$E(X) = \mu$ e $Var(X)\ \sigma^2 \leq$ Var(qualquer outro estimador de μ);

$E(S^2) = \sigma^2$ e $VAR(S^2) \leq$ Var(qualquer outro estimador de σ^2).

7.2.3 Distribuição amostral da média de uma população normal

Recorde-se que se X é uma variável aleatória com distribuição normal com média μ e variância σ^2 , o que se pode representar simbolicamente por $X \sim N(\mu, \sigma^2)$, então $Z = \dfrac{X - \mu}{\sigma}$ tem distribuição normal padrão, ou seja, distribuição normal com média zero e variância igual a 1, o que pode ser simbolicamente indicado por $Z \sim N(0, 1)$.

Considere-se uma amostra de tamanho n de uma população normal com média μ e variância σ^2. Tem-se, então, em consideração n variáveis aleatórias: $X_1, X_2,..., X_n$, independentes e com idêntica distribuição:

$$X_i \sim N(\mu, \sigma^2),\ i = 1, 2,..., n.$$

A média dessas n variáveis aleatórias é:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n};$$

Demonstra-se que a distribuição de X também é normal com a mesma média μ e variância σ^2/n, isto é, a variância da população reduzida à n-sima parte:

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n}).$$

Pela padronização da variável aleatória $\overline{X}$, obtém-se a variável aleatória normal padrão:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Distribuição t. A variância σ^2 é, em geral, desconhecida. Substituindo o desvio padrão da amostra s em lugar do desvio padrão da população σ na expressão de Z, obtém-se uma nova variável aleatória, designada por T que tem uma distribuição particular, denominada distribuição t com n-1 graus de liberdade (isto , com os graus de liberdade correspondentes aos g.l. da estimativa s^2 de σ^2):

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim T_{n-1}.$$

Probabilidades referentes a intervalos da variável aleatória T podem ser obtidos de tabelas especiais, como a Tabela de t em anexo.

7.2.4 Distribuição amostral da variância de uma população normal

O estimador S^2 de σ^2, determinado a partir de uma amostra de tamanho n, tem distribuição de χ^2 (qui-quadrado) com n-1 graus de liberdade. Melhor dito,

$$\frac{\nu S^2}{\sigma^2} \sim \chi^2_\nu,$$

onde ν=n-1 é o numero de graus de liberdade de S^2 . Probabilidades referentes à distribuição de S^2 podem ser obtidas de tabelas especiais, como a Tabela de χ^2 em anexo.

7.3 Estimação por Intervalo - Intervalo de Confiança

Um **intervalo de confiança** com **coeficiente de confiança** 1-α para o parâmetro θ de uma distribuição de probabilidade é um intervalo aleatório (a, b) tal que:

$$P(a < \theta < b) = 1-\alpha;$$

isto é, um intervalo aleatório cuja probabilidade de conter o parâmetro (desconhecido) θ é 1-α.

Um intervalo de confiança simétrico para a média de uma distribuição normal $N(\mu,\sigma^2)$ com coeficiente de confiança 1-α pode ser obtido como segue:

Os extremos do intervalo (-t', t') de valores da variável aleatória

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

correspondente à probabilidade 1-α podem ser obtidos da Tabela de t com a condição de que:

$$P(-t' < T < t') = 1-\alpha.$$

Substituindo $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ e transformando a dupla desigualdade -t' < T < t' de modo a obter μ em seu interior, obtém-se, sucessivamente:

$$-t' < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t';$$

donde:

$$-\frac{t'S}{\sqrt{n}} < \bar{X} - \mu < \frac{t'S}{\sqrt{n}}$$

e

$$-\bar{X} - \frac{t'S}{\sqrt{n}} < -\mu < \bar{X} - \frac{t'S}{\sqrt{n}}$$

multiplicando por -1 (do que resulta a inversão de sentido da dupla desigualdade), obtém-se:

$$\bar{X} - \frac{t'S}{\sqrt{n}} < \mu < \bar{X} + \frac{t'S}{\sqrt{n}}$$

Retornando à sentença de probabilidade inicial, resulta:

$$P\left(\bar{X} - \frac{t'S}{\sqrt{n}} < \mu < \bar{X} + \frac{t'S}{\sqrt{n}}\right) = 1-\alpha \ .$$

Logo, o intervalo

$$\left(\bar{X} - \frac{t'S}{\sqrt{n}},\ \bar{X} + \frac{t'S}{\sqrt{n}}\right)$$

é um intervalo de confiança para μ com coeficiente de confiança 1 - α.

7.4 Teste de Hipótese

7.4.1 Conceitos

Um **teste de hipótese** um processo de decisão estatística entre duas proposições alternativas referentes aos parâmetros de uma população. A primeira proposição, a hipótese básica, corresponde à hipótese científica formulada negativamente. Por essa razão ela é designada **hipótese de nulidade**. Uma hipótese estatística completamente formulada inclui a denominada **hipótese alternativa**, isto é, a decisão alternativa que deve ser adotada caso a hipótese de nulidade seja rejeitada no processo de decisão.

Essas duas hipóteses são, em geral, designadas por símbolos próprios: H_0 e H_A, respectivamente para representar as hipóteses de nulidade e alternativa.

A hipótese de nulidade propõe uma distribuição específica da família de distribuições postulada no processo de inferência como apropriada para a população em questão.

O teste de hipótese uma regra de decisão que, com base em dados da amostra, conduzirá à decisão em favor da hipótese H_0 ou da hipótese H_A, melhor dito, à aceitação ou rejeição de H_0 (e, portanto, à rejeição ou aceitação da H_A, respectivamente).

O processo consiste em escolher uma estatística apropriada com distribuição de probabilidade conhecida, denominada **critério de teste**, e decidir com base no valor da estatística determinado a partir de uma amostra. Se esse valor se situa em uma região de pequena área, isto é, de pequena probabilidade, arbitrariamente escolhida, na cauda da distribuição, a hipótese de nulidade é rejeitada, já que tal probabilidade pequena indica evidência de que a distribuição da estatística não pertence à distribuição especificada pela hipótese H_0. Essa região na cauda da distribuição do critério de teste é denominada **região de rejeição** (ou **região crítica**) e a região complementar, **região de aceitação**.

A regra de decisão é, portanto:

- Aceitar H_0 se o valor da estatística critério de teste estiver na região de aceitação;
- Rejeitar H_0 se o valor da estatística estiver na região de rejeição.

Exemplo 7.3. Suponha-se que resultados de pesquisas indiquem que a produção de grãos de uma cultivar de soja tem distribuição normal com média igual a 20 gramas e desvio padrão de 4 gramas. O melhoramento genético levado a efeito com a cultivar conduziu a uma nova cultivar cuja produção média o pesquisador espera seja superior (e base teórica assegura que não pode ser inferior) à da cultivar original, com o mesmo desvio padrão.

Com o intuito de verificar a validade dessa pressuposição, o pesquisador toma uma amostra de medidas da produção da cultivar melhorada para testar a seguinte hipótese:

$$\begin{cases} H_0: \mu = 20 \\ H_A: \mu > 20 \end{cases}$$

Com base no valor da média da amostra, o pesquisador deve decidir sobre H_0 ou H_A. Dois tipos de erro podem ser cometidos pelo pesquisador quando ele efetua um teste de hipótese:

Erro tipo I: Rejeitar a hipótese de nulidade H_0 quando ela é verdadeira;

Erro tipo II: Aceitar a hipótese de nulidade H_0 quando ela é falsa.

As probabilidades de cometer cada um desses erros são designadas simbolicamente por:

α = P(Erro tipo I),

β = P(Erro tipo II).

As quatro situações a que esse processo de decisão sob incerteza (dado o desconhecimento da verdadeira situação) pode conduzir o pesquisador são representadas na **Figura 7.1**:

Situação verdadeira	Decisão	
	H_0	H_A
H_o	Decisão correta	Erro tipo I
H_A	Erro tipo II	Decisão correta

Figura 7.1. Tabela de decisão em um processo de teste de hipótese estatística.

Em geral, o pesquisador tem alguma evidência empírica em favor da hipótese alternativa H_A, de modo que ele conduz o experimento com o propósito de comprovar tal evidência. Assim, é desejável um teste de hipótese que atribua alta probabilidade de rejeição da hipótese de nulidade H_0 se a hipótese alternativa for verdadeira, mantendo a probabilidade do erro tipo I adequadamente baixa. Tal probabilidade é denominada **potência do teste**. Logo,

Potência = P(Rejeitar H_0 quando H_0 é verdadeira)

= 1 - P(Aceitar H_0 quando H_A é verdadeira)

= 1 - β.

Assim, a potência de um teste exprime a habilidade do teste para rejeitar hipóteses de nulidade falsas. Em experimentos comparativos, isto significa sensibilidade para detectar diferenças reais.

Seria desejável um teste que assegurasse probabilidades convenientemente pequenas para ambos os erros tipo I e tipo II. Lamentavelmente, entretanto, essa condição não é viável, já que a diminuição de α implica em incremento de β, e vice-versa. O exemplo que segue ilustra a situação.

Exemplo 7.4. Considere-se a situação de melhoramento da cultura da soja, supondo, agora, que o pesquisador afirma que a pesquisa elevou a produção para 25 gramas. A hipótese a testar é, então:

$$\begin{cases} H_0: \mu = 20 \\ H_A: \mu = 25 \end{cases}$$

O teste dessa hipótese um processo de decisão entre duas alternativas simples: a média da população μ= 20 (H_0) ou μ=25 (H_A). As curvas das funções de densidade correspondentes a essas duas distribuições normais (ambas com mesmo desvio padrão de 4 gramas) são apresentadas na **Figura 7.2**. A região de rejeição correspondente a um erro tipo I com probabilidade α é a região indicada com sombreamento mais escuro na cauda direita da curva à esquerda; a região de aceitação é o complemento da área sob essa curva. Então, a probabilidade do erro tipo II (aceitar a hipótese H_0 quando H_A é a correta) é a região com sombreamento mais claro na cauda esquerda da curva da distribuição correspondente à hipótese H_A (curva à direita). Observe-se, então, que a tentativa de deslocar a fronteira entre as regiões de rejeição e de aceitação de H_0 de modo a diminuir a área α implica em incremento da área β, e vice-versa.

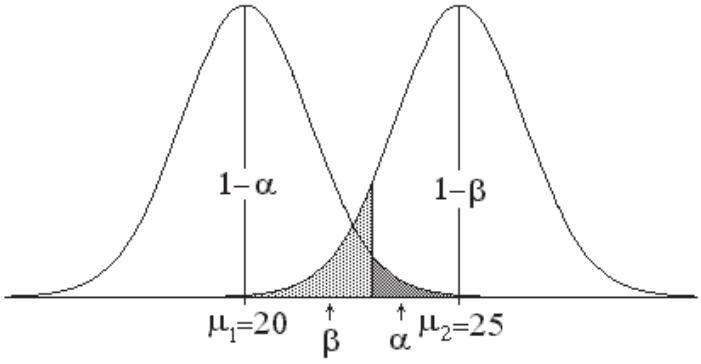


Figura 7.2. Representação gráfica das distribuições de probabilidades da produção de soja sob as hipóteses de nulidade e alternativa consideradas no **Exemplo 7.4**.

Como freqüentemente o pesquisador está interessado na rejeição da hipótese de nulidade H_0, dada a impossibilidade de fixar ambas as probabilidades α e β pequenas, é comum nos testes de hipótese fixar um valor convenientemente pequeno para α, sem consideração para o valor que resulta, correspondentemente, para β. Um teste de hipótese conduzido nessas condições é denominado **teste de significância**. A probabilidade α de erro tipo I é, então, designada **nível de significância** do teste.

A potência do teste a área não sombreada 1-β da curva à direita, isto é, o complemento da área achuriada (β) sob aquela curva.

7.4.2 Testes de hipóteses referentes à media de uma população normal

O exemplo anterior é uma ilustração de teste de hipótese referente à média de uma população com distribuição normal $X \sim N(\mu, \sigma^2)$ que, de modo geral, pode ser especificada por:

$$\begin{cases} H_0: \mu = c \ \ (\text{constante}) \\ H_A: \mu \neq c \end{cases}$$

Uma forma alternativa de especificar essa hipótese é:

$$\begin{cases} H_0: \mu - c = 0 \\ H_A: \mu - c \neq 0 \end{cases}$$

Nessa última forma, fica justificada a designação "hipótese de nulidade" para H_0.

Para o teste dessa hipótese, o pesquisador deve dispor de valores $\bar{x}$ e s^2 das estatísticas $\overline{X}$ e S^2, determinados a partir de uma amostra de tamanho n da população em referência. Então, um critério para teste dessa hipótese geral é a estatística:

$$T' = \frac{\overline{X} - c}{S_{\bar{x}}} \overset{H_0}{\sim} T_{n-1}$$

onde $S_{\bar{x}} = \sqrt{S^2/n}$ e n-1 é o número de graus de liberdade da estimativa da variância, s^2.

Exemplo 7.5. Suponha que seja conhecido que o coeficiente de digestibilidade de uma ração administrada a ovinos é 53,0 e que há indicações de que a suplementação de torta de soja pode alterar a digestibilidade da ração.

Para testar essa hipótese:

$$\begin{cases} H_0: \mu = 53,0 \\ H_A: \mu \neq 53,0 \end{cases}$$

o pesquisador administra a ração suplementada com torta de soja a 7 animais e determina os coeficientes de digestibilidade que seguem: 57,8; 56,2; 61,9; 54,4; 53,6; 56,4 e 53,2.

Os valores das estatísticas $\overline{X}$ e S^2 calculados para esta amostra são:

$\overline{x} = 56,21$ e $s^2 = 9,015$;

logo:

$$s_{\overline{x}} = \sqrt{s^2/n}$$
$$= \sqrt{9,015/7}$$

Supondo que a distribuição do coeficiente de digestibilidade é normal, o teste da hipótese especificada pode ser procedido através da estatística t:

$$T' = \frac{\overline{X} - 53,0}{S_{\overline{x}}} \overset{H_0}{\sim} T_6 .$$

Tem-se, no caso:

$$t' = \frac{56,21 - 53,0}{1,134} = \frac{3,21}{1,134} = 2,83 .$$

Consultando a Tabela de t, obtém-se, para 6 graus de liberdade e P = 0,025 (α=0,05, já que a área de cada cauda da curva da distribuição de t é 0,025), obtém-se:

$t_6 = 2,447$ ($\alpha = 0,05$, bilateral).

Como

t' (calculado) > t (tabelado),

rejeita-se a hipótese H_0. A **Figura 7.3**, mostra as regiões de rejeição e de aceitação e indica que o valor observado da estatística T situa-se na região de rejeição.

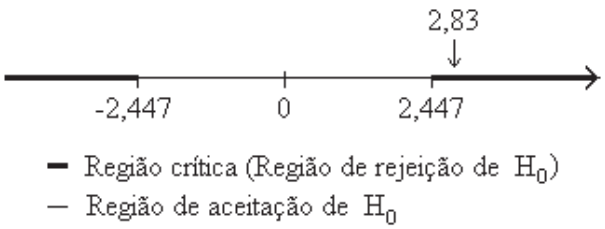


Figura 7.3. Regiões de rejeição e de aceitação do teste de hipótese ilustrado no Exemplo 7.4.

7.4.3 Teste da hipótese de igualdade das médias de duas populações normais

Populações com mesma variância

Em muitas situações, o pesquisador tem interesse em testar a hipótese de igualdade das medias μ_1 e μ_2 e de duas populações P_1: (μ_1, σ^2) e P_2: (μ_2, σ^2) com mesma variância σ^2, ou seja, a hipótese:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_2 \neq \mu_2 \end{cases}$$

ou, o que é o mesmo,

$$\begin{cases} H_0: \mu_1 - \mu_2 = 0 \\ H_A: \mu_2 - \mu_2 \neq 0 \end{cases}$$

Para o teste dessa hipótese, ele deve obter duas amostras aleatórias independentes, uma de cada uma das duas populações. Sejam (X_{11}, X_{12} ,...,$X_{1\,n_1}$) e (X_{21},X_{22},...$X_{2\,n_2}$) as duas amostras, de tamanhos n_1 e n_2, respectivamente, das populações P_1 e P_2. A partir dessas amostras, ele determina as estatísticas $\overline{X}_1$, $\overline{X}_2$, S_1^2 e S_2^2. Então, um critério para o teste da hipótese em referência é a estatística:

$$T' = \frac{\overline{X}_1 - X_2}{S_{\overline{x}_1 - \overline{x}_2}} \overset{H_0}{\sim} T_{n_1+n_2-2},$$

onde $S_{\overline{x}_1 - \overline{x}_2}$ é a estimativa do desvio padrão da estimativa da diferença entre as medias das duas populações, ou seja, a raiz quadrada de

$$S^2_{\overline{x}_1 - \overline{x}_2} = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) S^2,$$

onde S^2 é o estimador da variância comum das duas populações, determinado como a média ponderada dos estimadores das variâncias das duas amostras, tomando como pesos os respectivos graus de liberdade, ou seja:

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

$$= \frac{SQX_1 + SQX_2}{GLX_1 + GLX_2},$$

onde SQX_i é a soma de quadrados dos valores amostrais de X_i e GLX_i é o correspondente número de graus de liberdade (i=1,2).

Exemplo 7.6. Os coeficientes de digestibilidade determinados para duas rações A e B administradas a ovinos foram os seguintes:

X_1 (Ração A)	X_2 (Ração B)
57,8	64,2
56,2	58,7
61,9	63,2
54,4	62,5
53,6	59,8
56,4	59,2
53,2	
393,5	367,6
$\overline{x}_1 = 56,21$	$\overline{x}_2 = 61,26$
$SQX_1 = 54,09$	$SQX_2 = 26,52$
$s_1^2 = 9,015$	$s_2^2 = 5,304$

O teste a hipótese de que as médias populacionais dos coeficientes de digestibilidade das duas rações são iguais contra a alternativa de que elas diferem, supondo que as variâncias são iguais (mas desconhecidas), ou seja:

$$\begin{cases} H_0: \mu_1 = \mu_2 \\ H_A: \mu_2 \neq \mu_2 \end{cases}$$

é procedido como segue:

A estimativa da variância da estimativa da diferença entre as médias populacionais dos coeficientes de digestibilidade das duas rações é:

$$s^2 = \frac{54,09 + 26,52}{(7-1)(6-1)} = 7,328,$$

donde:

$$s^2_{\overline{x}_1 - \overline{x}_2} = \left(\frac{1}{7} + \frac{1}{6}\right) 7,3282 = 2,268;$$

logo,

$$s^2_{\overline{x}_1 - \overline{x}_2} = \sqrt{2,268} = 1,5061.$$

Portanto, o valor da estatística T para a amostra dos treze ovinos é:

$$t = \frac{56,21 - 61,26}{1,5061} = -3,353.$$

Por outro lado, o ponto superior que limita uma área (probabilidade) de 0,05/2 = 0,025 na cauda direita da curva da distribuição t para $n_1+n_2-2 = 11$ graus de liberdade, obtido da tabela de t de Student, é $t(0,05;11) = 2,201$.

Como o valor absoluto do valor observado de t, ou seja, 3,353, é superior a 2,201, o valor observado da estatística T situa-se na região de rejeição. Logo, a hipótese de nulidade é rejeitada. Conclui-se que as duas medias diferiram significativamente no nível de significância α=0,05.

7.4.4 Teste unilateral e teste bilateral

Há duas formas de teste de significância: a bilateral e a unilateral. O **teste bilateral**, o mais freqüentemente utilizado na prática, é apropriado quando o julgamento inicial, com base no raciocínio lógico referente às condições do problema, é de que o verdadeiro valor do parâmetro pode ser maior ou menor do que o valor sob a hipótese de nulidade. Nessa situação, um valor do parâmetro distante daquele correspondente à hipótese de nulidade pode revelar-se através de um valor t' grande e positivo ou pequeno e negativo. Assim, a região de rejeição para um teste com erro tipo I igual a α compreende as áreas simétricas nas duas caudas da distribuição t, cada uma correspondente à probabilidade $\alpha/2$. A hipótese de nulidade é rejeitada quando o valor observado t' da estatística t situa-se em qualquer uma das duas caudas. Nesse caso, o procedimento para o teste é calcular o valor absoluto de t', denotado $|t'|$, ignorando seu sinal. Procura-se na tabela de t a probabilidade de obter um valor de t maior do que o valor observado t' em qualquer das direções, isto é, a probabilidade de que $|t|$ exceda o valor observado $|t'|$ ($|t|$ significa "valor absoluto" de t).

O **teste unilateral** é apropriado apenas quando é sabido a priori o sinal da diferença entre o parâmetro e seu valor sob a hipótese de nulidade, se esta deve ser rejeitada. Assim, por exemplo, em um experimento para verificar a eficácia de um inseticida sobre o controle de uma praga, o pesquisador pode esperar, na pressuposição de que o inseticida não tem efeito tóxico prejudicial sobre a produção, que a aplicação do inseticida, se este tem algum efeito, somente pode resultar em aumento da produção. Nesse caso, a aplicação do teste unilateral é apropriada; ela implica em que o inseticida ou aumenta a produção ou não tem efeito sobre a produção; ele não pode prejudicar a produção. A região de rejeição para esse teste com erro tipo I igual a α é a cauda direita da distribuição t; logo, a hipótese de nulidade é rejeitada quando o valor observado t' da estatística t situa-se nessa cauda. Nesse caso, o procedimento para o teste é calcular o valor de t'. Se o sinal de t' for o esperado, procura-se na tabela de t a probabilidade de obter um valor de t maior ou menor do que o valor observado t', conforme o sinal de t'. Se o sinal de t' for o oposto do esperado, a hipótese de nulidade aceita, não havendo necessidade de recorrer à tabela de t.

O valor de α em um teste unilateral é exatamente a metade daquele de um teste bilateral com o mesmo t'. Assim, a hipótese de nulidade é rejeitada com maior freqüência para testes unilaterais do que para testes bilaterais.

7.5 Exercícios

1. Um pesquisador deseja conduzir um teste bilateral da hipótese de que a média de uma população igual a 12, sabendo que a variância da população 40. O pesquisador extrai uma amostra aleatória de 20 itens da população e determina a média da amostra.
   a) Especifique as hipóteses de nulidade e alternativa apropriadas.
   b) Suponha que a média da amostra 9,1. Determine a estatística para o teste da hipótese e calcule, aproximadamente, a probabilidade de um valor mais extremo, positivo ou negativo, do que o valor observado da estatística.
   c) Se o pesquisador escolheu o nível de significância α=0,05 para o teste, deve ele aceitar ou rejeitar a hipótese de nulidade? Escreva uma breve conclusão.

2. Considerando os mesmos valores para os parâmetros e estatísticas dados no problema 1, efetue o teste de hipótese com cada uma das seguintes alterações, comentando como o teste de hipótese alterado:
   a) α=0,01 em lugar de α=0,05.
   b) Teste unilateral em lugar de teste bilateral.
3. Suponha que conhecido que o conteúdo de proteína do pão de trigo tem distribuição normal com média igual a 12,0% e variância igual a 2,25%. Uma amostra de 100 pães com mistura de trigo e centeio analisada e revela o conteúdo médio de 11,8% de proteína. Supondo que a mistura de centeio não altera a variância do conteúdo de proteína do pão, verifique se a amostra de 100 pães revela evidência de que a mistura de centeio ao trigo não altera o conteúdo médio de proteína do pão. (Use o nível de significância α=0,05.)
4. O dano causado por uma praga à lavoura de feijão uma preocupação para os agricultores de uma região. A pesquisa recomenda a aplicação de um inseticida para a prevenção de perdas econômicas quando o número de ovos do inseto por planta excede 70 e afirma que o desvio padrão do número de ovos em uma planta igual a 40. Um agricultor colhe uma amostra aleatória de 80 plantas e obtém a média de 79 ovos por planta. Com base nessa informação, deve ele aplicar inseticida em sua lavoura?
5. O conteúdo médio de água considerado como padrão no controle de qualidade de salsichas enlatadas de uma indústria 30%. Partidas de salsichas com mais de 30% de água devem ser rejeitadas. A análise de uma amostra de 7 latas de salsichas de uma partida forneceu as seguintes percentagens de água: 32, 29, 33, 29, 33, 30 e 31. A partida deve ou não ser rejeitada?
6. Os dados seguintes são medidas de pressão do sangue de dois grupos de indivíduos: um grupo controle e um grupo tratado com sucrose (50 g/dia), após 13 semanas de tratamento:

	Tratados com sucrose	Controle
	110	106
	134	98
	122	108
	104	104
	118	120
	131	124
	114	108
		96
		100
$\sum X$	833	964
SQX	99.837	103.976

O pesquisador deseja saber se a administração de sucrose eleva a pressão do sangue. Efetue o teste de hipótese apropriado para responder à questão do pesquisador.

7. Para o estudo do efeito de um contraceptivo oral sobre a elevação de colesterol no sangue foi feita uma pesquisa utilizando dois grupos de dez mulheres, um dos quais recebeu o contraceptivo e o outro serviu como testemunha. Os resultados estão resumidos a seguir:

Contraceptivo	Observações						SX	SQX
Com:	163	157	245	169	93	110	937	160.553
Sem:	132	147	177	126	213	219	1.014	179.568

Efetue o teste de hipótese apropriado no nível de significância a=0,05.

8. Um pesquisador suspeita que o conteúdo de glicogênio no tecido do peito de frango é reduzido pelo cozimento em microondas. Para verificar sua suspeita, ele efetua um experimento com dois grupos de 5 peças de peito de frango. Os resultados foram os seguintes:

	Observações					ΣX	SQX
Não cozido:	28	17	36	23	27	131	3.627
Cozido em microondas:	12	7	11	10	11	51	535

a) Esses resultados parecem confirmar a natural expectativa de que a diminuição do conteúdo de glicogênio também resulta em uma menor variância. Efetue o teste de hipótese apropriado.

b) Tendo em conta o resultado obtido no item anterior, efetue o teste de hipótese apropriado para responder à questão de interesse do pesquisador.

9. Efetue o teste de homogeneidade de variância para os dados do exercício 6.
10. Foi efetuado um experimento para verificar se a infiltração de água de chuva difere entre solo que recebeu plantio continuado de soja e de milho. Os resultados obtidos em 7 talhões com cada um dos dois sistemas de cultivo foram os seguintes:

Soja contínua:	1,16	0,87	0,93	0,88	1,23	0,40	1,55
Milho contínuo:	0,84	1,96	0,73	1,42	1,41	0,89	0,73

Efetue o teste da hipótese formulada no plano do presente experimento.

11. Em propriedades de um município, observou-se o consumo de folhas de soja (em cm^2) por duas populações de anticárcia, obtendo-se os seguintes resultados:

População 1:	327	265	65	441	236	198	242	155
População 2:	40	125	233	148	103	79	188	97

Através de teste de hipótese apropriado, verifique se os dados são consistentes com a suspeita de que o consumo médio difere entre as duas populações.

12. Os seguintes dados foram obtidos num experimento executado para verificar se existe diferença sistemática na leitura da pressão sangüínea feita por dois aparelhos diferentes:

Aparelho A:	136	142	140	147	150	147				
Aparelho B:	141	167	141	145	127	146	135	152	135	152

Verifique se os dados confirmam a pressuposição.

13. Um experimento foi conduzido para determinar se o uso de um aditivo químico especial junto com um fertilizante acelera o crescimento das plantas. Ao final de quatro semanas, as plantas foram medidas e os resultados foram os seguintes:

Sem aditivo:	20	31	16	22	19	32	25	18	20	19
Com aditivo:	23	34	15	21	22	31	29	20	24	23

Verifique se esses dados são consistentes com a hipótese de que o uso do aditivo realmente acelera o crescimento das plantas.

14. Fenos de duas origens - feno de alfafa e feno de quicuio - foram testados em terneiros machos da raça Charolês de aproximadamente mesma idade, durante duas semanas. Ao final do período, foram registrados os seguintes dados de ganho de peso dos animais utilizados no experimento:

Feno de alfafa:	7,1	2,3	9,0	3,7	8,8		
Feno de quicuio:	3,6	4,5	2,3	1,2	4,9	3,2	5,7

Verifique se existe diferença significativa entre os dois tipos de feno quanto ao ganho de peso dos animais.

15. Uma amostra de 40 vasilhames de vinagre de uma partida de certa marca revelou o conteúdo médio líquido de 975 ml, com desvio padrão 10 ml. Teste a hipótese apropriada para verificar se o conteúdo médio das garrafas da referida partida deve ser considerado menor que 1.000 ml, conforme as exigências legais.
16. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando as letras V e F entre parêntesis, respectivamente. Se a sentença for falsa, explique porque.

( ) Os dois principais tipos de inferência são estimação por ponto e estimação por intervalo.

( ) Dois métodos de estimação são intervalo de confiança e teste de hipótese.

( ) Dado que a média amostral é um estimador não tendencioso da média da população, então essas duas médias são iguais.

( ) Uma das vantagens de usar uma média amostral em vez de uma única observação para estimar uma média populacional é que a variância amostral é mais provável estar próxima da média da população.

( ) Gosset (Student) descobriu que quando n é pequeno, a estimativa da variância tende a superestimar a variância.

( ) Um intervalo de confiança para a média com coeficiente de confiança de 0,95 contém 95% de todas as médias da população.

( ) Um intervalo de confiança baseado na distribuição de t é mais estreito do que o correspondente intervalo de confiança baseado na distribuição normal padrão.

( ) Se a variância de uma distribuição normal é desconhecida e sua estimativa é utilizada, então duas amostras aleatórias separadas de igual tamanho podem produzir dois intervalos de confiança de diferentes amplitudes.

( ) Em intervalo de confiança, as outras condições permanecendo iguais, quanto maior o tamanho da amostra, mais estreito é o intervalo.

( ) Se X tem distribuição binomial, o melhor estimador por ponto $\hat{p}=X/n$ do parâmetro p situa-se exatamente na metade do intervalo de confiança para p com coeficiente de confiança de 95%.

( ) Se o coeficiente de confiança é aumentado de 0,95 para 0,99, o intervalo de confiança torna-se mais estreito.

( ) Intervalos de confiança baseados em amostras grandes são mais prováveis de incluir o parâmetro populacional do que aqueles baseados em amostras pequenas.

( ) Para distribuições binomiais, as outras condições permanecendo constantes, quanto maior o valor de p maior será a amplitude do intervalo de confiança.

( ) As outras condições permanecendo iguais, quanto maior o coeficiente de confiança adotado maior será a amplitude do intervalo de confiança.

( ) No processo de estimação de um parâmetro populacional, um intervalo de confiança é mais provável estar correto do que uma estimativa por ponto.

( ) O parâmetro binomial p deve ser conhecido para se obter limites de confiança para p.

( ) Amostras repetidas do mesmo tamanho obtidas de uma mesma população binomial produzirão, sempre, intervalos de confiança com coeficiente de confiança de 99% de mesma amplitude para o parâmetro p.

( ) Intervalos de confiança para o parâmetro de Poisson são simétricos em relação à estimativa por ponto do parâmetro.

( ) Intervalos de confiança referem-se a parâmetros; intervalos de predição referem-se a observações.

( ) A hipótese alternativa é sempre a hipótese de pesquisa.

( ) ( ) A hipótese de nulidade pode ser a mesma hipótese científica.

( ) Em uma situação prática, a hipótese de nulidade, a hipótese alternativa e o nível de significância devem ser especificados antes da execução do experimento.

( ) Uma decisão feita na base de um procedimento estatístico será sempre correta.

( ) Se um procedimento estatístico correto é utilizado, é impossível rejeitar uma hipótese de nulidade verdadeira.

( ) Se a hipótese de uma pesquisa é verdadeira, então o acaso não interfere no resultado da pesquisa.

( ) Uma das principais razões para usar amostragem aleatória é determinar a probabilidade de que um experimento produza um resultado particular por acaso se a hipótese de nulidade é verdadeira.

( ) Dizer que a hipótese de nulidade é rejeitada não significa necessariamente que ela é falsa.

( ) Uma conclusão com base em um procedimento estatístico corretamente aplicado é baseada somente em probabilidade.

( ) O nível de significância em um procedimento estatístico depende do campo de pesquisa, do custo e da gravidade dos erros de decisão; entretanto, os níveis tradicionais são freqüentemente utilizados.

( ) Se X tem distribuição binomial e observa-se a freqüência $x/n = 0,5$, a hipótese de nulidade H_0: $p=0,05$ pode ser rejeitada.

( ) "Nível de significância" e "Região de rejeição" são duas expressões para a mesma cousa.

( ) O erro tipo I é definido como "a probabilidade de rejeitar a hipótese de nulidade quando ela é verdadeira".

( ) Quando a hipótese de nulidade é verdadeira, a probabilidade de cometer um erro tipo I é igual a α.

( ) Usando uma boa técnica estatística, hipóteses de nulidade verdadeiras são provavelmente rejeitadas tão freqüentemente quanto hipóteses de nulidade falsas.

( ) É impossível cometer um erro tipo I quando a hipótese de nulidade é falsa.

( ) A probabilidade de rejeitar a hipótese H_0 quando H_0 é falsa é designada pela letra grega β.

( ) A potência de um teste de hipótese é $1-\alpha$.

( ) É impossível cometer um erro tipo II quando a hipótese de nulidade é falsa.

( ) É impossível cometer um erro tipo II quando a hipótese de nulidade é rejeitada.

( ) Se se usa tamanho de amostra grande, há menor probabilidade de um erro tipo I e de um erro tipo II.

( ) Se um experimento é bem delineado e tanto α como β são pequenos, ele deve ser um bom experimento.

( ) Mesmo quando um procedimento estatístico correto é usado, é impossível aceitar a hipótese de nulidade quando ela é falsa.

( ) Quanto maior a região de rejeição, maior é a potência do experimento.

( ) P (X está na região de rejeição) = α se a hipótese de nulidade é verdadeira.

( ) Se a variância de uma população com distribuição normal é conhecida, então, se necessário, um teste de hipótese referente à média pode ser realizado com uma amostra de tamanho n=1.

( ) Um teste de hipótese envolvendo a estatística Z é freqüentemente usado porque a maioria das populações experimentais seguem distribuições normais com variâncias conhecidas.

( ) Para um teste de hipótese usando a estatística Z, a região de rejeição é unicamente determinada pela hipótese alternativa e o tamanho da amostra.

( ) A região de rejeição é baseada na hipótese de nulidade.

( ) O perigo do mau uso de um teste unilateral quando um teste bilateral deve ser usado é que ele faz o erro tipo I maior do que o para o teste apropriado.

( ) O perigo do mau uso de um teste bilateral quando um teste unilateral deve ser usado é que ele faz o erro tipo II maior do que o para o teste apropriado.

( ) A potência de um teste não pode ser menor que o erro tipo I.

( ) Quando se suspeita que a hipótese de nulidade é falsa, seria aconselhável diminuir a de 0,05 para 0,01 para ser menos provável rejeitá-la por acaso.

( ) Em um teste de hipótese, as outras condições permanecendo iguais, quanto maior o tamanho da amostra, menor o erro tipo I.

( ) Se o intervalo de confiança para a média com coeficiente de confiança 1-a não contém a média correspondente a hipótese de nulidade, então um teste bilateral conduziria rejeição da hipótese de nulidade no nível de significância a.

( ) Se um intervalo de confiança não contém o valor p_0 do parâmetro binomial p definido pela hipótese de nulidade H_0, essa hipótese pode ser rejeitada.

( ) Em testes de hipótese referentes ao parâmetro de Poisson λ, a hipótese alternativa pode ser unilateral ou bilateral.

( ) A potência de um teste de hipótese para o parâmetro l de uma distribuição de Poisson é aumentada quando se aumenta o número de intervalos amostrados.

( ) Uma hipótese referente o parâmetro binomial p testada pela distribuição binomial exata e pela aproximação normal dá exatamente as mesmas probabilidades.

( ) A distribuição de t é apropriada para amostras de pequeno tamanho, seja a variância conhecida ou não.

( ) Quando a variância de uma distribuição normal é conhecida, uma distribuição de t não deve ser usada para propósitos de inferência.

( ) Para um teste t de uma amostra, a região de rejeição é unicamente determinada pela hipótese alternativa e o tamanho da amostra.

( ) Para um teste de hipótese referente a média de uma distribuição normal, a distribuição de t é usada em lugar da distribuição normal padrão quando a variância é desconhecida e n é pequeno.

( ) O uso de um teste unilateral ou bilateral depende da questão perguntada pelo experimentador.

( ) Testes unilaterais são usualmente escolhidos para reduzir a probabilidade de um erro tipo I.

( ) Para um teste t em um nível a fixo, o valor absoluto do valor crítico cresce quando cresce o número de graus de liberdade.

( ) Se duas amostras consistem de pares de dados, o experimentador pode escolher o teste t para dados emparelhados ou o teste t para duas amostras independentes.

( ) Desde que testes t para dados emparelhados têm maior potência do que testes t para duas amostras independentes, o experimentador pode tentar planejar delineamento de pares combinados, se possível.

( ) Em um teste t para dados emparelhados, o parâmetro correspondente à hipótese de nulidade deve ser igual a zero.

( ) Em um teste t para comparação de médias com dados emparelhados envolvendo 20 pares de gêmeos, há 38 graus de liberdade.

( ) Mesmo quando as médias de duas populações normais são iguais, o valor calculado de t pode ser suficientemente grande devido a erro de amostragem para causar um erro tipo I.

( ) Se um teste t deixa de indicar significância para a diferença entre duas médias amostrais, resulta um erro tipo II.

( ) Se as outras condições permanecem as mesmas, é mais fácil rejeitar a hipótese de igualdade das médias de duas populações normais quando a diferença entre elas é 1,5 do que quando ela é 4,5.

( ) A variância ponderada de duas amostras é a média ponderada dos dois correspondentes desvios padrões.

( ) Um teste t para comparações com dados emparelhados deve ser sempre usado quando as variâncias das duas populações são iguais.

( ) Se um teste t determina que a diferença entre as médias de duas amostras é significativa, então o experimentador deve concluir que duas diferentes populações foram amostradas.

( ) Se as médias de duas populações normais são iguais, o valor calculado de t em um teste t para duas amostras será exatamente igual a zero.

( ) Se as variâncias de duas populações são iguais, a melhor estimativa da variância comum é a média das estimativas de variância das correspondentes amostras, independentemente de outras considerações.

( ) Se o experimentador está inseguro de que as médias de duas populações normais são iguais, esta hipótese pode ser testada antes da realização do teste t para duas amostras.

( ) Se a hipótese de igualdade das médias de duas populações normais é verdadeira para o teste t para a comparação de grupos, então a estatística t deve situar-se próxima de zero.

( ) Se o valor calculado de qui-quadrado é maior que o valor crítico, a hipótese de nulidade é rejeitada.

( ) A hipótese H_0: p=0,7 com Há: p≠0,7 pode ser testada tanto com a distribuição binomial como com a distribuição c^2. Se o tamanho da amostra é grande, a conclusão deve ser a mesma para os dois testes.

( ) Se há 75 observações de um experimento binomial e no teste qui-quadrado multinomial x_1-e_1 =25, então também x_2-e_2 =25.

( ) Quanto maior o número de graus de liberdade, menos provavelmente os valores amostrais de qui-quadrado estarão na região de rejeição.

( ) Dizer que um valor calculado de qui-quadrado é "significativo" indica que ele é numericamente menor do que o valor crítico com o qual é comparado.

( ) Em um experimento binomial, para testar H_0: p_1=0,25, p_2=0,50, p_3=0,25, devem ser usados 3 graus de liberdade.

( ) Se o tamanho da amostra é menor do que 25, uma correção para continuidade deve ser feita quando se efetua teste de uma razão 1:2:1.

( ) Com amostragem aleatória, valores significativos da estatística qui-quadrado serão obtidos apenas quando a hipótese de nulidade for falsa.

( ) A probabilidade de rejeitar a hipótese de nulidade aumenta na medida em que aumenta o número de graus de liberdade para a distribuição de qui-quadrado.

( ) Com amostragem aleatória, pode ser obtido um valor calculado de qui-quadrado maior do que o valor crítico, mesmo quando a hipótese de nulidade é falsa.

( ) O valor da estatística qui-quadrado deve ser relativamente grande se há estreita concordância entre as freqüências observadas e esperadas.

( ) O valor crítico para um teste multinomial de qui-quadrado, no nível de significância α=0,05, da hipótese de uma razão genética 27:9:9:3:3:3:1 é 14,067.

( ) O teste de qui-quadrado multinomial é um teste de qualidade de ajustamento.

( ) Para testar se um conjunto de amostras pode ser modelado por uma distribuição de Poisson, o experimentador deve especificar o parâmetro l antes da amostragem.

( ) Se a hipótese de nulidade para um teste de qualidade de ajustamento não é rejeitada, podemos concluir que os dados provém de uma população com a distribuição de probabilidade especificada.

( ) Um teste de qui-quadrado para uma tabela de contingência não é apropriado se se suspeita que as categorias nas filas e nas colunas não são independentes.

( ) Rejeitar a hipótese de nulidade em um teste de independência através da estatística χ^2 é decidir que as categorias nas filas são independentes das categorias nas colunas.

( ) O teste de χ^2 de homogeneidade pode ser usado se as razões sob a hipótese de nulidade são desconhecidas mas podem ser iguais para toda população amostrada.

( ) Um teste de χ^2 de independência para uma tabela $k \times 2$ tem k-1 graus de liberdade associado a ele.

( ) Um teste χ^2 de homogeneidade pode ser usado para testar a igualdade de parâmetros de duas distribuições binomiais.

( ) Um teste de χ^2 de independência testa a hipótese de nulidade de que há uma associação entre as categorias em filas e em colunas contra a alternativa de que elas não são relacionadas.

( ) Se a hipótese de igualdade das variâncias de duas populações é verdadeira, então a estatística F deve situar-se próxima de zero.

( ) O teste F é um procedimento para testar hipóteses referentes a variâncias apenas quando as médias são iguais.

( ) Quando as variâncias de duas populações são diferentes e desconhecidas e as amostras são pequenas, não há teste exato para testar uma hipótese de igualdade de médias das duas populações.

8 ANÁLISE DE REGRESSÃO E CORRELAÇÃO LINEAR SIMPLES

Conteúdo

8.1 Introdução....186
8.1.1 Origens e importância da análise de regressão linear....186
8.1.2 Relações entre variáveis....187
- Relações deterministas....187
- Relações semideterministas....187
- Relações empíricas....188
8.1.3Relações lineares....189
8.2 Relações de Duas Variáveis....190
8.3 Gráfico dos Dados....191
8.4 Análise de Regressão Linear Simples....193
8.4.1 Introdução....193
8.4.2 Modelo estatístico....194
8.4.3 Inferência estatística....196
8.4.4 Estimação (por ponto) dos parâmetros....196
Equação da linha reta ajustada....197
Resíduos e estimativa da variância do erro....198
Propriedades dos estimadores de quadrados mínimos....198
8.4.5 Teste de hipótese....201
8.4.5.1 Hipótese de relação linear entre Y e X....201
8.4.5.2 Hipótese referente à declividade da linha de regressão....202
8.4.5.3 Análise da variação....202
8.4.6 Coeficiente de determinação....205
8.4.7 Intervalo de confiança....206
8.4.7.1 Intervalo de confiança para o coeficiente de regressão b....206
8.4.7.2 Intervalo de confiança para $E(Y|X)$....208
8.5 Correlação Linear Simples....209
Relação entre coeficiente de regressão e coeficiente de correlação....210
8.6 Exercícios....211

8.1 Introdução

8.1.1 Origens e importância da análise de regressão linear

Em muitas áreas da pesquisa científica, a variação de características respostas de interesse é influenciada, em grande parte, por outras características cujas magnitudes variam no curso da pesquisa. A incorporação na análise estatística de informações referentes a estas características explanatórias é freqüentemente importante para a descrição e a derivação de inferências referentes às características respostas. O conhecimento de relações entre características também é útil para a predição de umas característica a partir de informações sobre as outras, e seu controle e otimização através da manipulação de fatores influentes.

A **análise de regressão** é um conjunto de técnicas estatísticas que tratam da formulação de modelos estatísticos que especificam relações entre variáveis, e do uso desses modelos para propósitos de inferências, particularmente predição. Os métodos mais usuais de análise de regressão, que serão abordados neste texto, tratam da situação de uma única variável resposta.

A palavra "regressão" foi empregada pela primeira vez no contexto aqui utilizado por Francis Galton, na análise da relação entre alturas de filhos e alturas médias de seus pais. De seu estudo, Galton concluiu que filhos de pais de estatura muito extrema (muito altos ou muito baixos) eram geralmente de estatura mais extrema (maior e menor, respectivamente) do que a média, mas não tão extrema como a de seus pais. Em sua publicação de 1885, "Regressão para a mediocridade em herança de estatura", Galton usou o termo "regressão" para significar que a altura do filho tende para a média em vez de para valores mais extremos.

Desta origem, o termo "análise de regressão" evoluiu para o contexto atual que diz respeito à análise de dados envolvendo duas ou mais variáveis com o objetivo de descobrir a natureza de sua relação e explorá-la para propósitos de predição.

O estudo de relações entre variáveis é importante em muitos campos da pesquisa científica, em particular da pesquisa experimental. Exemplos de relações de interesse são dados a seguir.

a) Produção de uma cultura e quantidades de fertilizantes aplicados ao solo, com o objetivo de estabelecer a forma da relação ou predizer a combinação ótima de fertilizantes.

b) Rendimento de uma cultura e várias características de clima, com o objetivo de obter a compreensão a cerca de possíveis mecanismos de influência de elementos do clima sobre o desenvolvimento da cultura.

c) Curva de lactação de vaca leiteira e níveis de componentes da dieta alimentar, para a compreensão da forma da tendência da produção de leite, ou estabelecimento da forma geral da curva de lactação com o propósito de subsequente exame dos efeitos de tratamentos sobre a curva de lactação.

d) Área foliar e peso da folha de uma cultivar em vários estádios de desenvolvimento da planta, para a predição da área foliar, uma característica de difícil mensuração, através do peso da folha, de fácil mensuração.

e) Área celular afetada e duração e intensidade de exposição a raios X, para monitorização de terapia de radiação.

f) Taxa de descarga de um rio medida em um local particular e quantidade de precipitação recente na correspondente bacia hidrográfica, para a predição de enchentes.

8.1.2 Relações entre variáveis

Podem ser caracterizados três tipos de relações entre variáveis: determinista, semideterminista e empírica.

- Relações deterministas

Certas relações correspondem a leis conhecidas, expressas por funções matemáticas exatas. Bases teóricas universalmente reconhecidas justificam a forma funcional. Desvios de observações que ocorrem para algumas dessas relações são considerados erros experimentais sem importância para a maioria dos propósitos. Assim, por exemplo, se X cruzeiros são depositados em uma conta de poupança ao juro anual de 100r%, a quantidade Y disponível nessa conta ao encerramento do n-ésimo ano é relacionada a X, r e n pela equação exata:

$$Y = X(1+r)^n.$$

Como um segundo exemplo, o tempo t para um objeto atingir a superfície da terra quando lançado de uma altura h é relacionado com h pela lei física da gravitação:

$$t = \sqrt{2h / g},$$

onde g é a constante gravitacional, que depende do local sobre a superfície. A origem dessa expressão matemática foi a postulação por Galileu de que a altura da queda de um objeto é proporcional ao quadrado do tempo para atingir o solo. Estimativas exatas da constante gravitacional foram ulteriormente obtidas por experimentação. Embora essa relação teórica seja aproximada, em decorrência da variação das condições ambientais, para a maioria dos propósitos tornam-se desnecessárias novas pesquisas experimentais. Dessa forma, esses casos são excluídos do domínio da análise da regressão.

- Relações semideterministas

Em algumas situações, a forma da lei que relaciona as variáveis é estabelecida por uma teoria conhecida, mas dependente de valores particulares de constantes desconhecidas (parâmetros) que aparecem em sua expressão matemática. Aproximações para a lei podem ser obtidas pela substituição dos parâmetros por estimativas, determinadas através de pesquisa experimental. O estabelecimento da relação exata é inviabilizado pela limitada precisão de instrumentos de medida, perturbações não controláveis das condições experimentais e outros fatores que introduzem erros experimentais.

Exemplo 8.1. A pressão (P) e o volume (V) de um gás, sob calor constante, são relacionados pela "lei dos gases ideais":

$$PV^g = \text{constante},$$

onde g é a taxa de calor específico do gás particular, que deve ser estimada através de pesquisa experimental.

Em alguns casos, há uma base teórica que sugere uma forma plausível para a relação, mas a base não é exata ou não é aceita universalmente. Ademais, freqüentemente, flutuações adicionais são produzidas por variáveis não controláveis, não incluídas na relação.

Exemplo 8.2. Suponha-se que uma fábrica de compota de pêssego produz latas de compota por lotes e que o produtor deseja relacionar o custo de produção de um lote (Y) com o tamanho do lote (X, número de latas no lote). Em um intervalo de variação realista de X, uma parcela do custo (custos de instalações e administração, por exemplo) é praticamente constante, independentemente do tamanho do lote. Um outro componente, que inclui matéria prima e o trabalho para produzir a compota, é diretamente proporcional ao número de latas produzidas. Denote-se o componente fixo do custo por F e o componente variável por c. Na ausência de qualquer outro fator, pode-se esperar uma relação custo - tamanho do lote determinista, expressa por:

$$Y = F+cX.$$

Entretanto, há um terceiro componente do custo a considerar cuja magnitude é de natureza não previsível. Por exemplo, o equipamento pode quebrar, ocasionalmente, do que pode resultar variação de custos de reparo e tempo despendido. A ocorrência de variação na qualidade da matéria prima também pode afetar o processo de produção. Dessa forma, uma relação determinista pode ser mascarada por componentes aleatórios. Conseqüentemente, uma relação adequada entre Y e X deve ser estabelecida através de pesquisa experimental.

- Relações empíricas

Em contraste com as situações anteriores, muito freqüentemente a relação não é governada por uma lei conhecida. Isso é o que ocorre com muitos fenômenos naturais, a que correspondem variáveis mutuamente relacionadas, ou uma variável que é dependente de um número de variáveis causais. Nessas situações, a forma da relação é comumente completamente desconhecida. Após a obtenção de suficiente conhecimento empírico sobre a relação, é muitas vezes possível a formulação de uma teoria que conduza a uma fórmula matemática correspondente a uma relação semideterminista.

Exemplo 8.3. Suponha que é desejado estudar a relação entre o rendimento de uma cultivar de tomate (Y) e a quantidade (X) de um certo fertilizante aplicado ao solo, com as demais condições tão constantes quanto possível. Para tal, pode ser conduzido um experimento, com a aplicação de diversas doses do fertilizante a várias parcelas em um intervalo de interesse. Diferentes doses do fertilizante produzirão diferentes rendimentos (o que ocorrerá, também, com diferentes parcelas com uma mesma dose do fertilizante), mas a relação entre rendimento e dose não segue uma fórmula matemática exata. Nesse caso, além das variações aleatórias não previsíveis, também não há qualquer base teórica conhecida para a determinação da forma da relação.

Exemplo 8.4. O desempenho de um novo provador de vinho depende da duração do período de treinamento e da natureza do programa de treinamento. A relação entre habilidade para degustação e duração do treinamento não é determinista, pelo simples fato de que dois seres humanos não são exatamente idênticos. Assim, para a avaliação da eficácia de um programa de treinamento, deve-se conduzir uma pesquisa experimental da relação entre habilidade de

degustação Y e duração do treinamento X. Ademais, a própria análise dos dados referentes a essas duas variáveis pode auxiliar na pesquisa da natureza da relação e em seu uso na avaliação e planejamento de um programa de treinamento.

Esses poucos exemplos ilustram o âmbito de aplicação da análise da regressão no contexto mais simples do estudo da relação de uma variável de interesse com uma única outra variável. Em situações mais complexas, uma variável de interesse pode depender de várias variáveis causais, ou várias variáveis podem inter-relacionar-se. Por exemplo, o rendimento de tomate pode ser estudado em relação à dosagem de um fertilizante e ao espaçamento entre plantas. A utilidade da análise de regressão estende-se a essas situações multivariadas. Ela fornece os métodos para a construção de modelos para as relações de interesse, estimação de parâmetros de interesse, determinação de variáveis de importância e de variáveis redundantes, e o emprego dos modelos para propósitos de predição e controle.

8.1.3 Relações lineares

Na análise de regressão, postula-se, para a população, que a relação entre uma variável de interesse Y e um conjunto de outras k variáveis $\{X_1,X_2,...,X_k\}$ é representada, , por um membro particular (desconhecido) de uma família de equações que exprimem Y como uma função de X_1, $X_2,...,X_k$ e um conjunto de constantes $\{a_1,a_2,...,a_p\}$:

$$Y = f(X_1,X_2,...,X_k; a_1,a_2,...,a_p),$$

onde Y é denominada **variável dependente**, ou **variável resposta**, $X_1,X_2,...,X_k$ são denominadas **variáveis independentes**, ou **variáveis explanatórias**, **ou variáveis preditoras**, e $a_1,a_2,...,a_p$ são constantes desconhecidas, denominadas **parâmetros.**

Cada membro dessa família de equações é especificado pela fixação de valores particulares para o conjunto de constantes $\{a_1,a_2,...,a_p\}$. Entretanto, a determinação do membro específico desta família de equações que corresponde a uma população de interesse é, em geral, inviabilizada pelo desconhecimento dos valores particulares dessas constantes apropriados para a população. O propósito da inferência estatística em análise de regressão é a obtenção de uma aproximação e de outras informações referentes aos parâmetros e, conseqüentemente, sobre a equação particular para a população. Essa equação particular (desconhecida) para a população pode ser adequadamente expressa por:

$$E(Y) = f(X_1,X_2,...,X_k; a_1,a_2,...,a_p),$$

onde $E(Y)$ indica **valor esperado** (ou valor populacional) da variável Y.

O processo de inferência estatística se baseia em um conjunto de n observações $\{x_{1j},x_{2j},...,x_{kj}, y_j : j=1,2,...,n\}$ das k+1 variáveis $X_1,X_2,...,X_k$ e Y, providas por uma amostra de n indivíduos ou unidades (amostra de tamanho n) da população de interesse. Um modelo de regressão é especificado com o estabelecimento da equação que exprime a relação de cada valor observado da variável dependente Y com os correspondentes valores das variáveis independentes, e das pressuposições referentes aos termos e símbolos da equação.

Um modelo de regressão é denominado um **modelo linear** se sua equação é um caso particular da seguinte forma geral:

$$Y = a_0 + a_1X_1 + a_2X_2 +...+ a_kX_k.$$

Um caso particular dessa forma geral é a equação do **modelo de regressão polinomial**:

$$Y = a_0 + a_1X + a_2X^2 + ... + a_kX^k,$$

em que: $X_1=X$, $X_2=X^2$,..., $X_k=X^k$. Esta equação exprime a relação de uma variável resposta Y com uma única variável explanatória X, representada por uma curva polinomial de grau k em um plano (espaço de duas dimensões).

Um caso ainda mais particular desse modelo linear é **o modelo de regressão linear simples**, que corresponde à relação de uma variável resposta Y com uma única variável explanatória X representada por uma linha reta, cuja equação é:

$$Y = a + bX.$$

8.2 Relações de Duas Variáveis

De modo geral, os modelos de relações de duas variáveis podem ser classificados em três categorias:

- **Modelo linear**, com equação da forma: $Y = aX + b$.

- **Modelo linear por anamorfose**: Modelo de equação não linear que pode tornar-se linear através de uma transformação de variáveis. Por exemplo, a equação $Y = ab^x$ pode ser transformada, através de uma transformação logarítmica, em

$$Z = A + BX,$$

onde: $Z = \log Y$, $A = \log a$ e $B = \log b$.

Outros exemplos de funções lineares por anamorfose são apresentados na **Figura 8.1**.

- **Modelo não linear**: Modelo não linear cuja equação pode ser tornada linear. Por exemplo, o modelo de Mitscherlich, com a seguinte equação exponencial:

$$Y = A[1-10^{-c(X+b)}].$$

Embora na maioria das situações de interesse prático a relação de duas variáveis seja não linear, a relação linear é importante por duas razões:

a) Em um modelo linear por anamorfose, uma transformação da equação não linear que relaciona as variáveis conduz a uma relação linear. Esse é o caso de algumas relações representadas por equações multiplicativas ou exponenciais que podem ser linearizadas por transformações logarítmicas, como exemplificado na **Figura 8.1**.

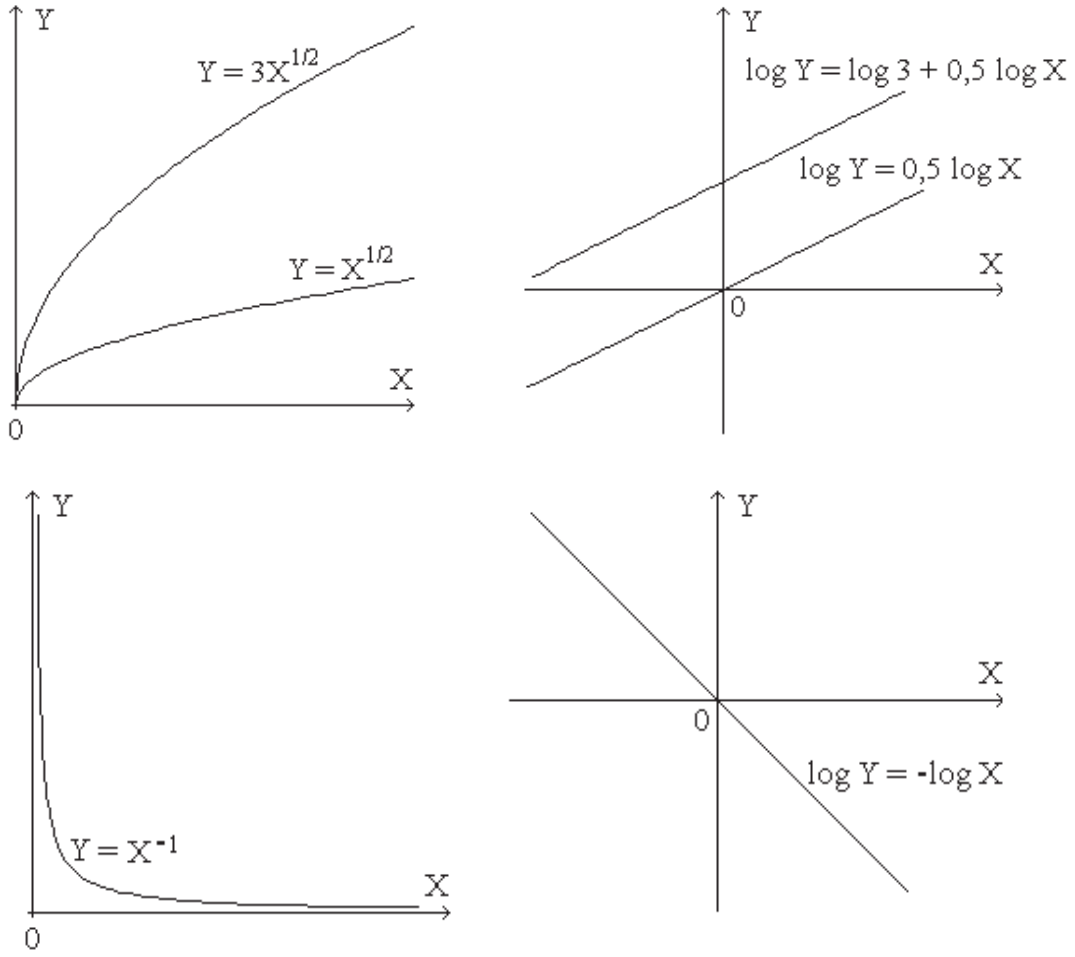


Figura 8.1. Várias formas de funções de equações não lineares e suas transformadas logarítmicas lineares, mostrando como logaritmos podem converter curvas não lineares em linhas retas.

b) A linha reta pode ser uma boa aproximação para um segmento de uma curva não linear em um intervalo limitado da variável independente.

8.3 Gráfico dos Dados

No estudo da relação de duas variáveis Y e X os dados são n pares de valores observados dessas duas variáveis: (x_1,y_1), (x_2,y_2), ...,(x_n,y_n), que podem ser arranjados na forma da **Tabela 8.1**.

Tabela 8.1. Forma dos dados em um estudo de relação de duas variáveis.

Observação:	1	2	...	n
X:	x_1	x_2	...	x_n
Y:	y_1	y_2	...	y_n

O primeiro passo no estudo da relação de duas variáveis é a representação gráfica dos dados através de um diagrama de dispersão de pontos. A inspeção desse diagrama pode permitir uma caracterização da relação entre as variáveis quanto à sua existência e forma; indica se os pontos se aglomeram em torno de alguma curva geométrica particular, como uma linha reta, por exemplo; e fornece uma apreciação visual da extensão da variação em torno de tal linha ou curva. O diagrama de dispersão é particularmente importante quando, como ocorre em muitas situações, não há uma relação teórica conhecida a priori. Em tais situações, ele é útil na pesquisa de um modelo apropriado para exprimir a relação.

Exemplo 8.5. Para uma ilustração concreta, considere-se que os dados da **Tabela 8.2** são as observações em 10 parcelas de um experimento de fertilização do solo com nitrogênio para a cultura do tomate, onde doses de nitrogênio e produção são expressos em unidades de medida convenientes.

Tabela 8.2. Dose de nitrogênio e produção de tomate em 10 parcelas de um experimento.

Dose (X):	1	1	2	3	4	4	5	6	6	7
Produção (Y):	2,1	2,5	3,1	3,0	3,8	3,2	4,3	3,9	4,4	4,8

O diagrama de dispersão para essas observações é apresentado na **Figura 8.2**. Este diagrama mostra que os pontos parecem se aglomerar em torno de uma linha reta, o que revela que a relação particular é de natureza aproximadamente linear.

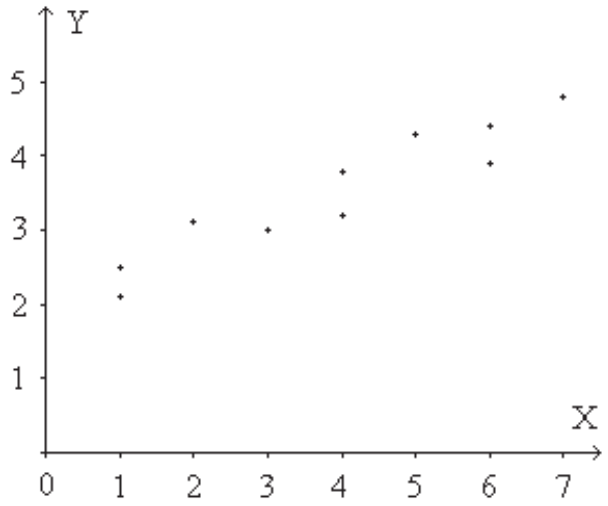


Figura 8.2. Diagrama de pontos para os dados da **Tabela 8.2**.

O **Exemplo 8.5** ilustra a relação mais simples entre duas variáveis, representada pela linha reta.

8.4 Análise de Regressão Linear Simples

8.4.1 Introdução

Recorde-se que, se a relação entre uma variável dependente Y e uma variável independente X corresponde exatamente a uma linha reta, então ela é expressa algebricamente pela equação:

$$Y = a + bX,$$

representada na **Figura 8.3**, onde a é a ordenada da interseção da reta com o eixo Y, denominada **ordenada na origem**, ou **interseção**, ou **coeficiente linear** da reta, e b é a tangente trigonométrica do ângulo que a reta forma com o eixo X, denominada **declividade**, ou **coeficiente angular** da reta. O coeficiente angular b corresponde à alteração na grandeza de Y por uma unidade de alteração no valor de X. Assim, a linha reta é a curva com a propriedade particular importante de taxa de alteração constante, isto é, acréscimos de Y para intervalos de X de mesma amplitude são iguais.

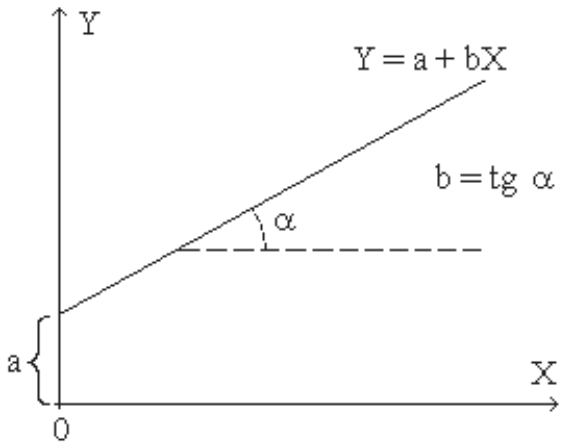


Figura 8.3. Representação gráfica da linha reta.

8.4.2 Modelo estatístico

Em situações não deterministas, relações lineares são mascaradas por variações aleatórias ou erros experimentais, de modo que, em um diagrama de dispersão, os pontos correspondentes a uma série de observações não se situam exatamente sobre uma reta.

A análise de regressão linear simples corresponde ao estudo da relação entre duas variáveis Y e X da forma $E(Y) = a+bX$, em que a variável explanatória X é controlada pelo experimentador, ou não sujeita a erro, a variável resposta Y é sujeita a fontes de erro não controláveis, e a e b são constantes desconhecidas (parâmetros).

A questão estatística fundamental em análise de regressão é a derivação de inferências referentes aos parâmetros a e b. Como qualquer processo de inferência estatística, inicia-se pela obtenção de uma amostra de observações - pares de valores de X e Y - da população de interesse.

Seja $\{(x_1,y_1), (x_2,y_2),..., (x_n,y_n)\}$ um conjunto de n observações referentes às variáveis X e Y, obtidas de uma amostra de tamanho n (ou seja, de n unidades) de uma população. A análise de regressão linear simples pressupõe que a relação entre a variável resposta Y e a variável explanatória X para a j-ésima observação é expressa pela seguinte equação:

$$y_j = a + bx_j + e_j, \quad j=1,2,...,n,$$

onde x_j e y_j são os valores das variáveis X e Y referentes à observação j, a e b são constantes específicas para a população em consideração, e e_j é o desvio entre o valor observado e o valor esperado da variável resposta (correspondente à verdadeira relação linear postulada para a população): $e_j = y_j - E(y_j)$.

O modelo estatístico para regressão linear simples completa-se com as seguintes pressuposições referentes aos termos e símbolos da equação:

a) A variável X é medida sem erro, ou seja, $\{x_1,x_2,...,x_n\}$ é um conjunto de valores selecionados pelo pesquisador para o estudo, não sujeitos a erro, logo, um conjunto de constantes conhecidas.

b) Os coeficientes a e b, que determinam a linha reta, são constantes desconhecidas, isto é, parâmetros.

c) e_1, e_2,...,e_n são realizações de n variáveis aleatórias com média zero: $E(e_j)=0$, que satisfazem às seguintes pressuposições:

c_1) Homogeneidade de variância: $Var(e_j) = \sigma^2_{Y:X}$ (constante, comum para todas as observações).

c_2) Distribuição normal.

c_3) Independência estatística.

De acordo com esse modelo, o valor y_j da variável resposta, observado para o nível x_j da variável explanatória, é uma realização de uma variável aleatória com distribuição normal de média $E(Y) = a+bX$ e variância $Var(Y) = \sigma^2_{Y:X}$. Isto significa que a observação do valor sobre a verdadeira linha de regressão é impedida pelo desvio (erro) aleatório. Essa estrutura de erro é ilustrada na **Figura 8.4**, que mostra que a distribuição de Y para cada nível de X tem média sobre a verdadeira linha reta desconhecida a+bX e variância comum $\sigma^2_{Y:X}$, também desconhecida. Um importante propósito da análise de regressão é estimar essa reta e essa variância.

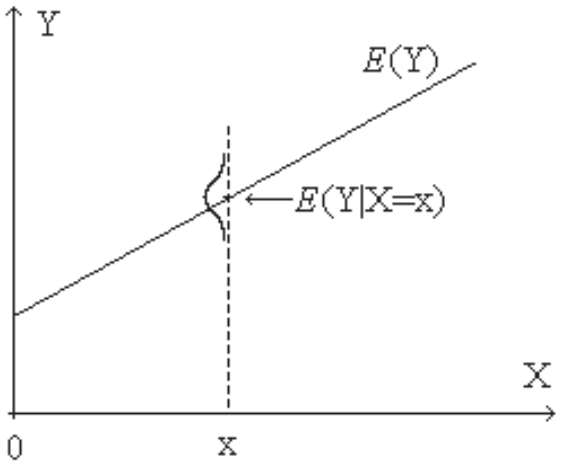


Figura 8.4. Distribuição normal de Y com médias sobre a linha reta $E(Y) = a+bX$.

Exemplo 8.6. Para a ilustração que segue utilizar-se-á os dados da **Tabela 8.3**, provenientes do registro de informações anuais de quantidade de frango comercializado e correspondente preço, nos Estados Unidos, no período de 1950 a 1959.

Tabela 8.3. Dados anuais de quantidade de frango comercializado (em milhões, X) e correspondente preço (em US$ por 50 kg, Y), nos Estados Unidos, no período de 1950 a 1959.

j	1	2	3	4	5	6	7	8	9	10
x_j	73	79	80	69	66	75	78	74	74	84
y_j	18,0	20,0	17,8	21,4	21,6	15,0	14,4	17,8	19,6	14,1

A "lei da oferta e da procura" em Economia estabelece que o valor da produção varia inversamente com a quantidade produzida, ou seja, o aumento da produção implica em diminuição do valor do produto. Dessa forma, segundo esta lei, há uma relação linear negativa entre valor da produção e quantidade do produto ofertada. Entretanto, essa relação é afetada por outras circunstâncias do mercado, tal como oferta de outros produtos substitutivos, de modo que pode ser interessante a verificação de sua ocorrência em uma situação prática.

A relação entre preço e quantidade de frango produzido revelada pelos dados do exemplo é mostrada no diagrama de pontos da **Figura 8.5**.

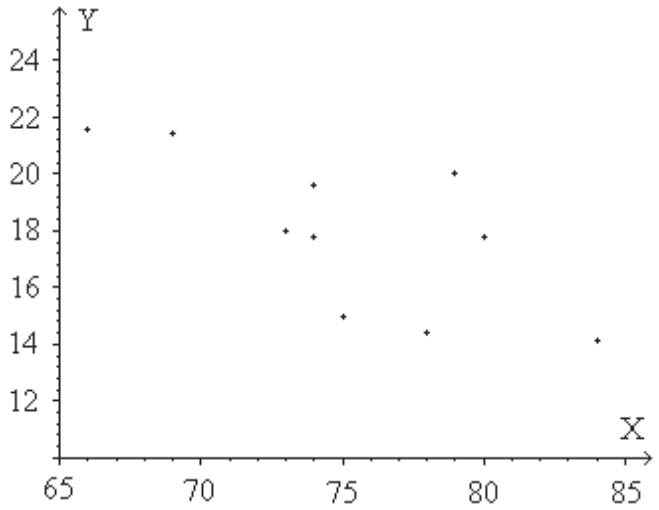


Figura 8.5. Diagrama de pontos que mostra a relação entre preço do frango (Y) e número de frangos comercializados anualmente (X) nos Estados Unidos (1950-1959).

8.4.3 Inferência estatística

Em regressão linear simples, como, em geral, em qualquer aplicação da estatística, interessam as seguintes inferências a respeito dos parâmetros do modelo estatístico:

- Estimação por ponto,
- Estimação por intervalo - Intervalo de confiança,
- Teste de hipótese.

8.4.4 Estimação (por ponto) dos parâmetros

Os parâmetros a e b da equação do modelo estatístico podem ser estimados pelo **método dos quadrados mínimos**, que determina para estimadores dos parâmetros os valores de a e b que minimizam a soma de quadrados dos erros (como função de a e b), isto é:

$$f(a,b) = \sum_{j=1}^{n} e_j^2$$

$$= \sum_{j=1}^{n} (y_{ij} - a - bx_j)^2 .$$

O processo de minimização dessa soma de quadrados dos erros conduz ao seguinte sistema de duas equações com as duas incógnitas a e b:

$$\begin{cases} n\hat{a} + (\sum_{j=1}^{n} x_j)\hat{b} = \sum_{j=1}^{n} y_j \\ (\sum_{j=1}^{n} x_j)\hat{a} + (\sum_{j=1}^{n} x_j^2)\hat{b} = \sum_{j=1}^{n} x_j y_j \end{cases}$$

cuja solução $(\hat{a}, \hat{b})$ corresponde aos estimadores dos parâmetros a e b:

$$\hat{b} = \frac{SPXY}{SQX}$$

e

$$\hat{a} = \bar{y} - \hat{b}\bar{x},$$

onde:

$$SPXY = \sum_{j=1}^{n} (x_j - \bar{x})(y_j - \bar{y}) = \sum_{j=1}^{n} x_j y_j - \frac{1}{n}(\sum_{j=1}^{n} x_j)(\sum_{j=1}^{n} y_j)$$

e

$$SQX = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}(\sum_{j=1}^{n} x_j)^2 .$$

Equação da linha reta ajustada

A substituição dos parâmetros na equação do modelo de regressão pelas correspondentes estimativas fornece a **equação da linha reta ajustada** (também denominada **equação ajustada**, **equação predita** ou **equação de quadrados mínimos**):

$$\hat{y} = \hat{a} + \hat{b}x .$$

Uma outra forma da equação de quadrados mínimos pode ser obtida pela substituição da expressão do estimador do parâmetro a nesta equação:

$$\hat{y} = \bar{y} + \hat{b}(x - \bar{x}) .$$

O valor $\hat{y}_0$ para um valor particular x_0 da variável explanatória X estima o correspondente ponto sobre a reta (desconhecida) para a população: $E(Y: X=x_0) = a+bx_0$, isto é, estima a média de Y para a população que corresponde a um dado valor x_0 de X.

Em particular, pode-se determinar os valores preditos (ou ajustados) da variável resposta Y correspondentes aos valores experimentais ou observados da variável explanatória: x_j, j=1,2,...,n. Esses valores preditos são as projeções sobre a reta de quadrados mínimos, ortogonais ao eixo X, dos pontos no plano que correspondem aos pares de valores observados (x_j, y_j), j=1,2,...,n.

Resíduos e estimativa da variância do erro

Os valores ajustados, em geral, diferem dos valores observados. A diferença entre um valor observado da variável resposta e seu correspondente valor ajustado é uma estimativa do desvio da regressão para a correspondente observação, denominada **resíduo** da observação:

$$\hat{e}_j = y_j - \hat{y}_j,\ j=1,2,...,n.$$

Uma estimativa da variância comum dos desvios da regressão é fornecida pela soma dos quadrados dos resíduos dividida pelo correspondente número de graus de liberdade, n-2 (onde 2 é o número de parâmetros estimados do modelo):

$$s^2_{Y:X} = \frac{1}{n-2}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2 .$$

Essa estimativa da variância dos desvios da regressão pode ser mais convenientemente obtida através do procedimento da análise da variação (que se verá adiante), usualmente utilizado para efetuar testes de hipóteses de interesse.

Propriedades dos estimadores de quadrados mínimos

Os estimadores de quadrados mínimos têm as seguintes propriedades:

1) A soma dos resíduos é nula:

$$\sum_{j=1}^{n}\hat{e}_j = \sum_{j=1}^{n}(y_j - \hat{y}_j) = 0.$$

2) A soma dos quadrados dos resíduos é mínima, ou seja, a soma dos quadrados das distâncias entre os valores observados (x_j,y_j), j=1,2,...n, e suas correspondentes projeções sobre um plano no espaço de três dimensões, ortogonais ao plano X_1X_2, é mínima quando esse plano é o plano de quadrados mínimos.

Exemplo 8.6 (continuação). Para os dados do **Exemplo 8.6**, as estimativas dos parâmetros a e b podem ser obtidas com o auxílio da **Tabela 8.4**.

Tabela 8.4

Tabela 8.4. Dados e cálculos para a determinação das somas de quadrados e somas de produtos.

j	x_j	y_j	$x_j - \overline{x}$	$y_j - \overline{y}$	$(x_j - \overline{x})^2$	$(y_j - \overline{y})^2$	$(x_j - \overline{x})(y_j - \overline{y})$
1	73	18,0	-2,2	0,03	4,84	0,0009	-0,066
2	79	20,0	3,8	2,03	14,44	4,1209	7,714
3	80	17,8	4,8	-0,17	23,04	0,0289	-0,816
4	69	21,4	-6,2	3,43	38,44	11,7649	-21,266
5	66	21,6	-9,2	3,63	84,64	13,1769	-33,396
6	75	15,0	-0,2	-2,97	0,04	8,8209	0,594
7	78	14,4	2,8	-3,57	7,84	12,7449	-9,996
8	74	17,8	-1,2	-0,17	1,44	0,0289	0,204
9	74	19,6	-1,2	1,63	1,44	2,6569	-1,956
10	84	14,1	8,8	-3,87	77,44	14,9769	-34,056
Soma	752	179,7	0	0	253,60	68,321	-93,040
Média	75,2	17,97					

Obtém-se:

$$\hat{b} = \frac{-93,04}{253,60} = -0,367$$

e

$$\hat{a} = 17,97-(-0,367)75,2 = 45,57.$$

A equação de regressão ajustada é, então:

$$\hat{y} = 45,57-0,367x.$$

Essa equação fornece o valor estimado (ou predito) do preço de frango para uma produção anual particular. Por exemplo, o preço de frango (em US$/50 kg) estimado para uma produção anual de 73 milhões de frangos é:

$$\hat{y}_{X=73} = 45,57-0,367(73) = 18,779.$$

Esse valor ajustado difere do correspondente valor observado (18,0). O resíduo para essa observação é:

$$\hat{e}_1 = 18,0-18,78 = -0,78.$$

A **Tabela 8.5** apresenta os resíduos para as 10 observações da amostra e correspondentes valores de outras estatísticas de que se tratará adiante.

Tabela 8.5. Produção anual de frangos (em milhões, X) e correspondente preço (em US$/50 kg, Y), nos Estados Unidos, no período 1950-59, correspondentes valores ajustados e resíduos da regressão de Y em relação a X, e desvios padrões e intervalos de confiança para os valores ajustados.

j	x_j	y_j	$\hat{y}_j$	$\hat{e}_j$	$s_{\hat{y}_j}$	EI_j	ES_j
1	73	18,0	18,78	-0,78	0,713	17,14	20,42
2	79	20,0	16,58	3,42	0,819	17,69	18,47
3	80	17,8	16,21	1,59	0,903	14,13	18,29
4	69	21,4	20,24	1,16	1,036	17,85	22,63
5	66	21,6	21,35	0,25	1,361	18,21	24,49
6	75	15,0	18,04	-3,04	0,654	16,53	19,55
7	78	14,4	16,94	-2,54	0,748	15,22	18,67
8	74	17,8	18,41	-0,61	0,672	16,86	19,96
9	74	19,6	18,41	1,19	0,672	16,86	19,96
10	84	14,1	14,74	-0,64	1,316	11,71	17,78
Soma	752	179,7	179,70	0			

A **Figura 8.6** mostra a representação gráfica da reta de quadrados mínimos e dos resíduos correspondentes aos valores observados.

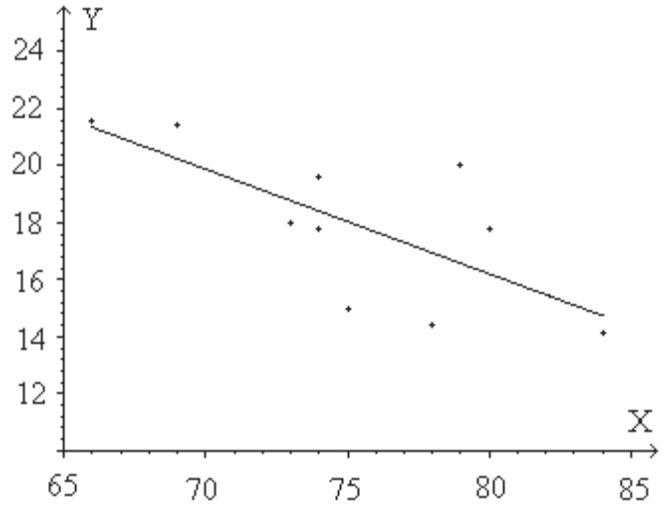


Figura 8.6. Representação gráfica da reta de quadrados mínimos ajustada para representar a relação entre preço de frango comercializado e quantidade de frango produzido, e dos resíduos da regressão.

A estimativa da variância do desvio da regressão é, então,

$$s^2_{Y:X} = \frac{1}{8}[(-0,78)2+3,422+...+(-0,64)2] = \frac{1}{8}34,132 = 4,2665\,.$$

8.4.5 Teste de hipótese

8.4.5.1 Hipótese de relação linear entre Y e X

Uma hipótese fundamental na análise de regressão linear simples diz respeito à linearidade da forma da relação entre a variável resposta e a variável explanatória. A hipótese estatística correspondente é especificada como segue:

$$\begin{cases} H_0: b = 0 \\ H_A: b \neq 0 \end{cases}$$

Um critério para o teste dessa hipótese é provido pela seguinte estatística:

$$T = \frac{\hat{b}}{s_{\hat{b}}},$$

onde $s_{\hat{b}}$ é a estimativa do desvio padrão da estimativa do coeficiente de regressão b, isto é, a raiz quadrada da estimativa da variância da estimativa de b:

$$s^2_{\hat{b}} = \frac{1}{SQX} s^2_{Y:X}\,.$$

Sob a hipótese de nulidade (H_0 :b=0), esta estatística tem distribuição t (de Student) com os graus de liberdade da estimativa da variância do erro, ou seja, n-2. Então, em um teste de hipótese com probabilidade de erro tipo I α=0,05, a hipótese de nulidade H_0: b=0 é rejeitada se o valor da estatística T observado para a amostra particular situar-se em uma das duas caudas da distribuição de t cada uma com área igual a $\alpha/2$=0,025: na cauda superior acima do valor t_α e na cauda inferior abaixo do valor $-t_\alpha$.

Exemplo 8.6 (continuação). Para o **Exemplo 8.6**, tem-se $\hat{b}$ = -0,367 e:

$$s^2_{\hat{b}} = \frac{4,267}{253,6} = 0,01684;$$

donde:

$$s_{\hat{b}} = \sqrt{0,01684} = 0,12976\,.$$

Então, a hipótese H_0: b=0 é rejeitada se o valor observado da estatística:

$$t = \frac{|-0,367|}{0,12976} = 2,8283,$$

superar o ponto percentual bilateral superior da distribuição t (Tabela IV) para 8 graus de liberdade da estimativa da variância do resíduo, ou seja:

$$t_{8;P} = \begin{cases} 2,306, & P = 0,05 \\ 3,355, & P = 0,01 \end{cases}$$

Como esse valor observado t=2,8283 está compreendido entre estes dois valores da tabela, rejeita-se a hipótese de nulidade, concluindo-se que há uma relação linear significativa (P<0,05) entre o preço do frango comercializado e a quantidade de frango produzido.

8.4.5.2 Hipótese referente à declividade da linha de regressão

Em algumas circunstâncias, há interesse em testar a hipótese de que a linha de regressão tem uma dada declividade, com base em alguma condição ou conjetura; ou seja, em testar a seguinte hipótese referente à declividade da linha reta:

$$\begin{cases} H_0: b=b_0 \\ H_A: b\neq b_0 \end{cases}$$

onde b_0 é a declividade conjeturada.

Um critério para o teste dessa hipótese é provido pela seguinte estatística T:

$$T = \frac{\hat{b}-b_0}{s_{\hat{b}}}.$$

Sob a hipótese de nulidade (H_0: $b=b_0$), esta estatística tem distribuição t (de Student) com n-2 graus de liberdade.

8.4.5.3 Análise da variação

O desvio total de uma observação particular em relação à média de todas as observações pode ser decomposto em dois desvios, conforme ilustrado na **Figura 8.7**:

$$y_j - \overline{y} = (\hat{y}_j - \overline{y}) + (y_j - \hat{y}_j),$$

onde:

$\hat{y}_j - \overline{y}$: desvio atribuível à regressão, isto é, à grandeza do coeficiente de regressão b;

$y_j - \hat{y}_j$: desvio devido ao afastamento do valor observado de Y em relação à linha de regressão ajustada, ou seja, o resíduo.

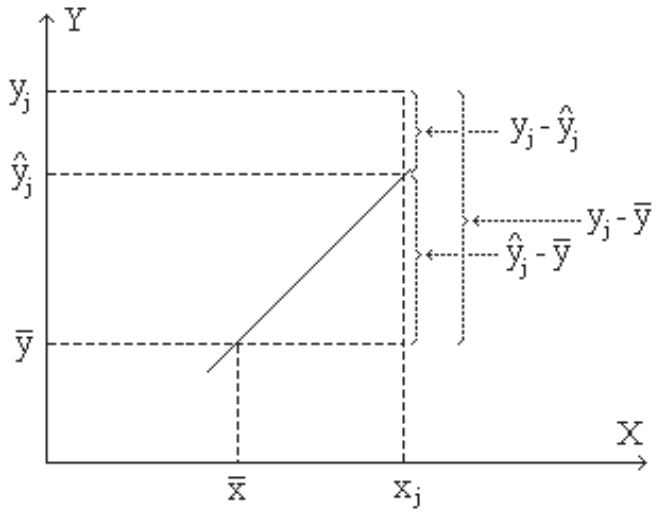


Figura 8.7. Decomposição do desvio total de uma observação no desvio atribuível à regressão e no resíduo.

Elevando ao quadrado ambos os membros da igualdade, obtém-se:

$$(y_j - \bar{y})^2 = (\hat{y}_j - \bar{y})^2 + (y_j - \hat{y}_j)^2 + 2(\hat{y}_j - \bar{y})(y_j - \hat{y}_j)$$

e, somando para todas as n observações:

$$\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 ,$$

dado que a soma dos produtos dos dois desvios é nula.

Assim, a variação total de Y, expressa pela soma dos quadrados (dos desvios) de Y, denotada por SQY, é decomposta em dois componentes:

$\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2$: variação atribuível à regressão - soma de quadrados da regressão: SQRegressão, e

$\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$: variação atribuível ao desvio da regressão - soma de quadrados do desvio ou resíduo: SQDesvio ou SQResíduo.

Demonstra-se que:

$$\text{SQRegressão} = \hat{b}^2 \text{SQX} =$$
$$= \hat{b}\,\text{SPXY},$$

onde as expressões de SQX e SPXY foram dadas anteriormente. De fato, da equação que exprime o valor ajustado para a j-ésima observação:

$$\hat{y}_j = \bar{y} + \hat{b}(x_j - \bar{x})$$

obtém-se a seguinte expressão:

$$\hat{y}_j - \bar{y} = \hat{b}(x_j - \bar{x})$$

que, elevada ao quadrado e somada para todas as n observações, produz:

$$\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 = \hat{b}^2 \sum_{j=1}^{n}(x_j - \bar{x})^2 =$$

$$= \hat{b}^2 \, SQX.$$

Correspondentemente à decomposição da variação total, os n-1 graus de liberdade para o total das n observações também são decompostos em 1 grau de liberdade correspondente à estimativa do coeficiente de regressão e n-2 graus de liberdade correspondentes ao resíduo.

Essa decomposição da variação da variável resposta Y é usualmente efetuada com o auxílio da seguinte "tabela da análise da variação":

Fonte de variação	GL	SQ	QM
Regressão	1	SQRegressão	QMRegressão
Desvio	n-2	SQDesvio	QMDesvio
Total	n-1	SQY	

onde:

SQDesvio = SQY-SQRegressão,

QMRegressão = SQRegressão,

QMDesvio = SQDesvio/(n-2).

Demonstra-se que o QMDesvio = $s^2_{Y:X}$ é um estimador não tendencioso da variância do desvio da regressão $\sigma^2_{Y:X}$, isto é, que:

E(QMDesvio) = $\sigma^2_{Y:X}$

e que:

$$E(\text{QMRegressão}) = \sigma^2_{Y:X} + b^2 \sum_{j=1}^{n}(x_j - \bar{x})^2 .$$

Por outro lado, ambos os quadrados médios QMDesvio e QMRegressão são estatisticamente independentes. Logo, a estatística F=QMRegressão/QMDesvio tem distribuição F com 1 e n-2 graus de liberdade correspondentes à QMR e QMD, respectivamente. Dessa forma, a estatística:

$$F = \frac{QM\,\text{Regressão}}{QM\,\text{Desvio}}$$

é um critério para o teste da hipótese de linearidade da relação entre Y e X:

$$\begin{cases} H_0 : b = 0 \\ H_A : b \neq 0 \end{cases}$$

Essa estatística F é um critério equivalente à estatística t para o teste desta mesma hipótese. De fato, demonstra-se que $F = T^2$.

Exemplo 8.6 (continuação). Para os dados do **Exemplo 8.6**, tem-se:

SQRegressão = (-0,367)(-93,04) = 34,16 ;

SQTotal = 68,32 ;

SQResíduo = SQTotal - SQRegressão

= 68,32 - 34,16 = 34,16

A análise da variação é completada na seguinte tabela:

Tabela da análise da variação:

Fonte de variação	GL	SQ	QM	F	F(0,05)	F(0,01)
Regressão	1	34,16	34,16	8,00	5,32	11,26
Resíduo	8	34,16	4,27			
Total	9	68,32				

Se o nível de significância (erro tipo I) escolhido para o teste foi α=0,05, dado que F>F(0,05), rejeita-se a hipótese H_0. Conclui-se que há uma relação linear significativa (P<0,05), no caso negativa, entre a variável Y (preço de frango) e a variável X (quantidade de frango comercializado), isto é, que o preço de frango decresce significativamente com o aumento da quantidade de frango produzido.

Essa conclusão é a mesma obtida anteriormente pelo teste t. De fato, pode-se verificar que: $T_8^2 = F_{1;8}$.

8.4.6 Coeficiente de determinação

O **coeficiente de determinação** é definido como:

$$r^2 = \frac{SQ\,Regressão}{SQTotal},$$

isto é, a proporção da variação total da variável resposta Y que é levada em conta pelo ajuste do modelo de regressão linear simples, isto é, que é explicada pela variação da variável independente X através da relação linear.

Segundo sua expressão de definição, o coeficiente de determinação é uma grandeza com o seguinte intervalo de variação:

$$0 \le r^2 \le 1,$$

já que a SQRegressão e a SQTotal são ambas, por definição, não negativas e, necessariamente, SQRegressão ≤ SQTotal.

O coeficiente de determinação é uma medida da qualidade do ajuste da equação de regressão linear simples. Nas situações extremas, tem-se:

- Ajuste perfeito → SQRegressão = SQTotal → r^2 =1,
- Ajuste extremamente mau → SQRegressão = 0 → r^2 =0.

Exemplo 8.6 (continuação). Tem-se, para o **Exemplo 8.6**, SQRegressão = 34,16 e SQTotal = 68,32. Logo,

$$r^2 = \frac{34,16}{68,32} = 0,50.$$

Assim, 50% da variação total do preço do frango é explicada pela variação da quantidade de frango produzido, através da relação linear.

8.4.7 Intervalo de confiança

8.4.7.1 Intervalo de confiança para o coeficiente de regressão b

Um **intervalo de confiança** para o coeficiente de regressão linear b **com coeficiente de confiança** $1-\alpha$ é o intervalo com extremos inferior (EI) e superior (IS) aleatórios que satisfaz a seguinte condição de probabilidade:

$$\text{Prob}[EI < b < ES] = 1-\alpha.$$

Uma expressão para esse intervalo de confiança, ou seja, para os extremos EI e ES desse intervalo, pode ser derivada do fato de que a estatística:

$$T = \frac{\hat{b}-b}{s_{\hat{b}}},$$

tem distribuição T (de Student) com ν graus de liberdade da estimativa $s_{\hat{b}}$ do desvio padrão de $\hat{b}$.

Então, a probabilidade de que a estatística T se situe no intervalo $(-t_{\nu;\alpha}, t_{\nu;\alpha})$ é igual a $1-\alpha$, onde $t_{\nu;\alpha}$ é o ponto superior da distribuição t que limita a cauda direita de área $\alpha/2$; ou seja:

$$\text{Prob}\left[-t_{\nu;\alpha} < \frac{\hat{b}-b}{s_{\hat{b}}} < t_{\nu;\alpha}\right] = 1-\alpha.$$

A relação expressa pela dupla desigualdade:

$$-t_{\nu;\alpha} < \frac{\hat{b}-b}{s_{\hat{b}}} < t_{\nu;\alpha.}$$

não se altera se seus três membros são multiplicados por $s_{\hat{b}}$, obtendo-se:

$$-t_{\nu,\alpha}s_{\hat{b}} < \hat{b}-b < t_{\nu,\alpha}s_{\hat{b}};$$

então, isolando o parâmetro b no membro interno da dupla desigualdade, tem-se:

$$-\hat{b}-t_{\nu,\alpha}s_{\hat{b}} < -b < -\hat{b}+t_{\nu,\alpha}s_{\hat{b}},$$

e, multiplicando esta dupla desigualdade por –1, vem:

$$\hat{b}-t_{\nu,\alpha}s_{\hat{b}} < b < \hat{b}+t_{\nu,\alpha}s_{\hat{b}}.$$

Portanto, a sentença de probabilidade anterior pode ser posta sob a forma:

$$\text{Prob}\left[\hat{b}-t_{\nu,\alpha}s_{\hat{b}} < b < \hat{b}+t_{\nu,\alpha}s_{\hat{b}}\right] = 1-\alpha.$$

Essa é uma sentença de probabilidade porque o intervalo $(\hat{b}-t_{\nu,\alpha}s_{\hat{b}},\ \hat{b}+t_{\nu,\alpha}s_{\hat{b}})$ é aleatório, já que seus extremos são funções de $\hat{b}$ e $s_{\hat{b}}$ que são estatísticas, ou seja, funções das observações de qualquer amostra aleatória que possa ser derivada da população sob consideração. Essa sentença de probabilidade é válida a priori, ou seja, antes da obtenção da amostra e correspondente determinação das estimativas $\hat{b}$ e $s_{\hat{b}}$ particulares para essa amostra. Após a obtenção de uma amostra particular, ela não tem mais sentido, pois, então, os extremos não são mais variáveis aleatórias, mas realizações particulares dessas variáveis aleatórias, ou seja, constantes conhecidas.

Nessas circunstâncias, diz-se que o intervalo (EI=$\hat{b}-t_{\nu,\alpha}s_{\hat{b}}$, ES-=$\hat{b}+t_{\nu,\alpha}s_{\hat{b}}$) determinado a partir de uma amostra particular é um **intervalo de confiança** para o coeficiente de regressão linear b com **coeficiente de confiança 1-α**, o que pode ser denotado por:

$$\text{IC b } (1\text{-}\alpha)\text{: } \begin{cases} EI = \hat{b} - t_{\nu;\alpha}\, s_{\hat{b}} \\ ES = \hat{b} + t_{\nu;\alpha}\, s_{\hat{b}} \end{cases}$$

onde $t_{\nu;\alpha}$ é o valor obtido da tabela de pontos percentuais superiores bilaterais da distribuição t (Tabela IV) para ν graus de liberdade da estimativa da variância do desvio da regressão $s^2_{Y:X}$ e probabilidade α.

De fato, a sentença de probabilidade Prob(EI<b<ES)=1-α significa o seguinte: Se fossem extraídas todas as amostras aleatórias possíveis da população e para cada uma dessas amostras se determinasse um intervalo de confiança para o parâmetro b com coeficiente de confiança 1-α, então a proporção 1-α do total dos intervalos de confiança determinados conteria o parâmetro desconhecido b. Posto de modo mais prático: se fossem determinados os intervalos de confiança para b com coeficientes de confiança de 95% para cada um conjunto particular de, por exemplo, 100 amostras aleatórias da população, então, o (valor real do) parâmetro estaria contido em aproximadamente 95% dos intervalos determinados.

Como usualmente é desejado coeficiente de confiança elevado, escolhe-se para a probabilidade α um valor convenientemente pequeno, muito freqüentemente:

$\alpha = 0{,}05 \rightarrow$ coeficiente de confiança = 0,95 (95%),

$\alpha = 0{,}01 \rightarrow$ coeficiente de confiança = 0,99 (99%).

Observe-se, entretanto, que a amplitude do intervalo de confiança, ou seja:

$$ES - EI = \hat{b}+t_{\nu,\alpha}s_{\hat{b}} - (\hat{b}-t_{\nu,\alpha}s_{\hat{b}}) = 2\,t_{\nu,\alpha}s_{\hat{b}},$$

será tão maior quanto menor for a probabilidade α e, portanto, quanto maior for o coeficiente de confiança 1-α. Dessa forma, a fixação de um coeficiente de confiança exageradamente grande implicaria em conseqüente intervalo de confiança de amplitude também exageradamente grande, que proverá informação demasiadamente vaga sobre a localização do parâmetro b.

Exemplo 8.6 (continuação). Intervalo de confiança para o parâmetro b com coeficiente de confiança de 95% para a situação do **Exemplo 8.6**.

Tem-se: $\hat{b}$=-0,367; $s_{\hat{b}}$=0,12976; $t_{8;0,05}$= 2,306. Logo:

EI = -0,367 - 2,306x0,12976 = -0,6662;

ES = -0,367 + 2,306x0,12976 = -0,0678.

Assim, um intervalo de confiança para o coeficiente de regressão linear b com coeficiente de confiança 0,95 é: (-0,6662; -0,0678).

8.4.7.2 Intervalo de confiança para *E*(Y|X)

Um **intervalo de confiança para a ordenada de um ponto sobre a reta postulada para a população**, para um valor particular x_0 de X, isto é, um intervalo de confiança para $E(Y|X=x_0)$, **com coeficiente de confiança** $1-\alpha$ é um intervalo com extremos inferior EI e superior IS aleatórios que satisfaz a condição de probabilidade:

$$\text{Prob}(EI < E(Y|X=x_0) < ES) = 1-\alpha,$$

onde os extremos têm as seguintes expressões: $EI = \hat{y} - t_{\nu;\alpha}\, s_{\hat{y}:x_0}$ e $ES = \hat{y} + t_{\nu;\alpha}\, s_{\hat{y}:x_0}$, e ν=n-2 e $s_{\hat{y}:x_0}$ é a estimativa do desvio padrão do valor ajustado $\hat{y}$ para um valor particular x_0, ou seja, a raiz quadrada da variância do valor ajustado $\hat{y}_{x_0}$:

$$s^2_{\hat{y}:x_0} = \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SQX}\right] s^2_{Y:X}.$$

Esse intervalo de confiança pode ser denotado por:

$$\text{IC } E(Y|X=x_0)\ (1-\alpha): \begin{cases} EI = \hat{y} - t_{\nu;\alpha}\, s_{\hat{y}:x_0} \\ ES = \hat{y} + t_{\nu;\alpha}\, s_{\hat{y}:x_0} \end{cases}$$

Exemplo 8.6 (continuação). Intervalo de confiança para o ponto sobre a reta de regressão $E(Y) = a + bX$ para o valor de X correspondente à primeira observação, ou seja, X=73.

Para X=73, obtém-se:

$$s^2_{\hat{y}:X=73} = 4,27\left[\frac{1}{10} + \frac{(73-75,2)^2}{253,60}\right] = 0,5085;$$

donde:

$$s_{\hat{y}:X=73} = \sqrt{0,5085} = 0,7131.$$

Então, u intervalo de confiança para $E(Y|X=73)$ com coeficiente de confiança de 95% é obtido como segue:

EI = 18,78-(2,306)(0,7131) = 17,14 ,

ES = 18,78+(2,306)(0,7131) = 20,42 .

Logo:

IC $E(Y|X=73)$ (95%): (17,14; 20,42).

8.5 Correlação Linear Simples

O coeficiente de correlação linear simples de duas variáveis aleatórias Y e X exprime o grau de associação linear entre essas variáveis. Ele é denotado pela letra grega ρ e expresso por:

$$\rho = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\ \text{Var}(Y)}}.$$

O coeficiente de correlação linear simples é estimado pela estatística:

$$r = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \sum_{j=1}^{n}(y_j - \bar{y})^2}} = \frac{SPXY}{\sqrt{SQX\ SQY}}.$$

Exemplo 8.6 (continuação). Retorne-se aos dados do **Exemplo 8.6**, referentes à quantidade anual de frangos comercializados e correspondentes preços no período de 1950 a 1959. A questão básica de interesse é a seguinte: os dados confirmam a associação linear negativa esperada entre preço e oferta nesse período (ou seja, maiores produções são acompanhadas de preços mais baixos)? Esse fato é aparentemente evidenciado pelo diagrama de pontos da **Figura 8.5** que revela a tendência geral dos pontos à esquerda situarem-se em posição mais elevada do que os pontos à direita. Entretanto, para quem esperasse uma associação bastante elevada entre preço e oferta, a dispersão dos pontos mostrada nessa figura pode ser desapontadora. Os pontos parecem situar-se dentro de uma elipse, o que é típico de diagramas que correspondem à associação linear de grau médio.

Utilizando resultados anteriores, obtém-se a estimativa do coeficiente de correlação linear:

$$r = \frac{-93,04}{\sqrt{(253,60)(68,32)}} = -0,7068.$$

A **Figura 8.8** ilustra diagramas de dispersão de pontos correspondentes a associações de vários graus. Em uma associação perfeita os pontos situam-se exatamente sobre uma reta não paralela ao eixo X. Nesse caso, a associação é positiva se a declividade da reta (em relação ao eixo horizontal X) é positiva, e negativa se a declividade é negativa. Na situação usual em que os pontos não se situam todos sobre uma reta, o sinal da associação corresponde ao sinal da declividade do eixo maior da elipse que contorna os pontos. A associação nula corresponde à situação em que a elipse reduz-se a um círculo.

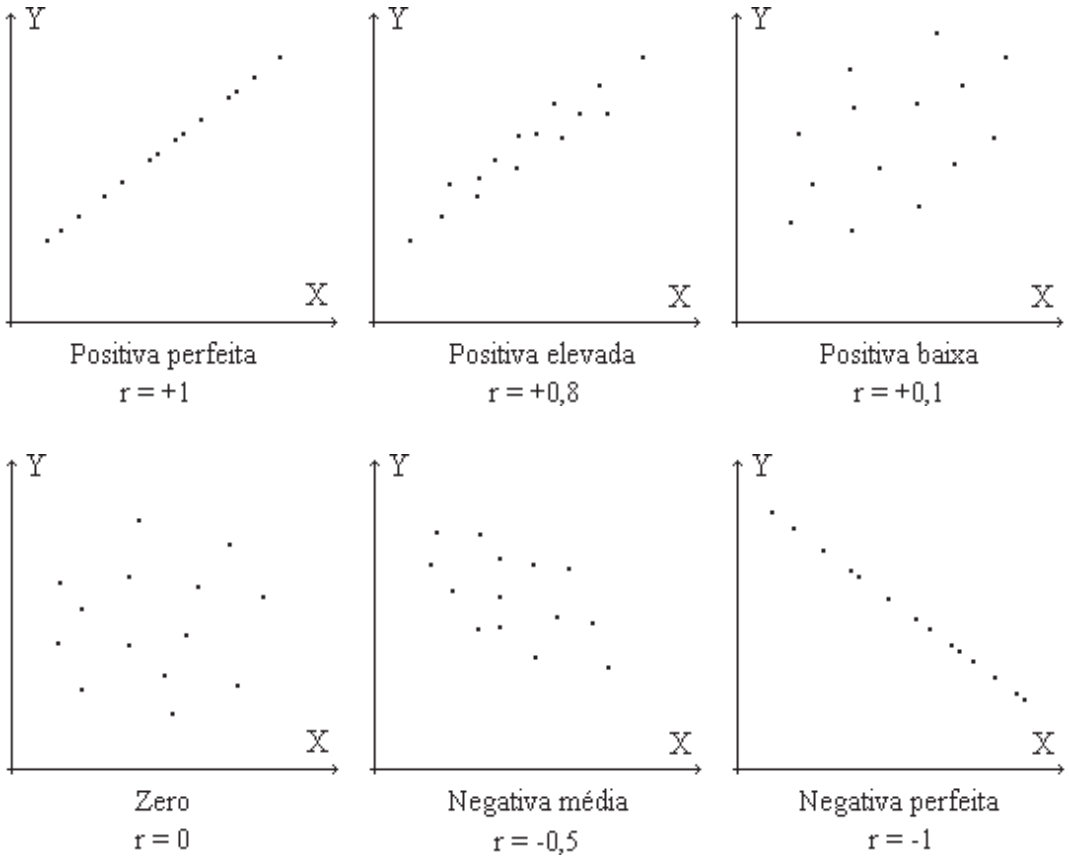


Figura 8.8. Diagramas de pontos correspondentes a diferentes graus de associação de duas variáveis e respectivos coeficientes de correlação.

Demonstra-se que o coeficiente de correlação linear simples entre duas variáveis aleatórias X e Y é a raiz quadrada do coeficiente de determinação correspondente ao ajuste da regressão linear simples de Y em relação a X. Essa relação entre coeficiente de correlação e coeficiente de determinação é a justificativa para o uso dos correspondentes símbolos: r e r^2. No exemplo, o coeficiente de determinação linear simples foi $r^2=0,50$. Assim, o coeficiente de correlação linear simples também pode ser obtido como segue:

$$r = \sqrt{0,50} = 0,707 .$$

Relação entre coeficiente de regressão e coeficiente de correlação

Demonstra-se que:

$$r = \hat{b}\sqrt{\frac{SQX}{SQY}},$$

de modo que o coeficiente de correlação linear tem o mesmo sinal do coeficiente de regressão linear.

Exemplo 8.6 (continuação).

$$r = -0{,}367\sqrt{\frac{253{,}60}{68{,}32}} = -0{,}7068.$$

8.6 Exercícios

1. Um experimento foi realizado com o objetivo de verificar se existe uma relação linear entre a quantidade de fósforo no tecido da planta de uma cultivar de feijão e a quantidade de fósforo adicionada ao solo como fertilizante. A quantidade média de fósforo no tecido das plantas colhidas das parcelas para análise (em ppm, Y) e as correspondentes doses de fósforo aplicadas ao solo (em kg/ha, X) são dadas a seguir:

X:	0	20	40	60	80
Y:	0,2	0,3	0,5	0,6	0,8

   a) Estime a equação de regressão: $E(Y) = a + bX$.

   b) Efetue a análise da variação e teste a hipótese H_0: b=O; H_A: b≠O, através da estatística F.

   c) Calcule o coeficiente de determinação r^2.

   d) Calcule os valores ajustados de Y para as doses de fósforo utilizadas no experimento. Determine o coeficiente de correlação entre os valores observados e os valores ajustados de Y. Verifique que o quadrado desse coeficiente de correlação é igual ao valor do coeficiente de determinação encontrado no item anterior.

   e) Redija a conclusão referente à adequabilidade da relação linear, com base nos resultados obtidos nos itens b) e c).

2. Em um experimento em que se estudou o efeito do nível de água aplicada através de irrigação sobre o rendimento de grãos de arroz irrigado foram obtidos os seguintes resultados:

								Soma
Nível de água (X):	12	18	24	30	36	42	48	210
Produção (Y):	5,27	5,68	6,25	7,21	8,02	8,71	8,42	49,56

   A soma dos produtos para as variáveis X e Y, corrigida para as correspondentes médias, é 103,68.

   a) Especifique as pressuposições para o ajustamento do modelo de regressão linear $E(Y) = a+bX$.

   b) Efetue o ajustamento do modelo de regressão linear especificado no item anterior.

   c) Teste a hipótese H_0: b=0 contra a alternativa H_A: b≠0.

   d) Interprete o significado geométrico dos coeficientes a e b do modelo de regressão linear simples.

   e) Calcule e interprete o coeficiente de determinação.

   f) Redija as conclusões.

3. No estudo da relação linear entre duas variáveis negativamente correlacionadas foram obtidos os seguintes dados:

   $\overline{x}=0;\ \overline{y}=12;\ s_x=8;\ s_y=10;\ \text{e}\ r^2=0{,}64.$

   Determine a estimativa da equação de regressão linear de Y em relação a X.
4. No estudo da relação linear entre duas variáveis negativamente correlacionadas foram obtidos os seguintes dados:
5. Decida se cada uma das seguintes sentenças é verdadeira ou falsa, indicando as letras V e F entre parênteses, respectivamente. Se a sentença for falsa, explique porque.

   ( ) A soma dos quadrados dos desvios da regressão não pode ser maior que a soma de quadrados total.

   ( ) Se a declividade da reta de regressão ajustada é um valor numérico grande, a relação entre Y e X é forte.

   ( ) Se todos os valores da amostra ficam sobre a reta de regressão ajustada, a soma de quadrados total é igual a zero.

   ( ) Uma das pressuposições da análise de regressão é que a variável dependente tem distribuição normal.

   ( ) Em testes de significância referentes ao coeficiente de regressão b pressupõe-se que Y tem a mesma variância para todos os valores fixos de X.

   ( ) Aceitar a hipótese de nulidade da declividade da reta de regressão é decidir que não há relação linear entre Y e X.

   ( ) Aceitar a hipótese de nulidade do coeficiente de correlação linear é decidir que não há relação entre Y e X.

   ( ) Quando cresce o grau de associação de duas variáveis, melhora o ajustamento da reta de regressão.

   ( ) Quanto maior é a magnitude de r, mais forte é a relação linear entre X e Y.

   ( ) Quando calcula um coeficiente de correlação, o experimentador pressupõe que há uma relação de causa-efeito entre X e Y.

   ( ) Em inferências referentes ao coeficiente de correlação, os valores de X podem ser fixos ou provenientes de amostragem aleatória, mas os valores de Y devem provir de uma amostra aleatória.

   ( ) Se os papéis de X e Y são trocados, a equação de regressão permanece a mesma.

   ( ) Se os papéis de X e Y são trocados, o coeficiente de correlação permanece o mesmo.

APÊNDICE

Tabela A-I. Distribuição cumulativa binomial..... 215
Tabela A-II. Pontos α-percentuais superiores da distribuição normal padrão..... 222
Tabela A-III. Pontos α-percentuais superiores da distribuição qui-quadrado..... 223
Tabela A-IV. Pontos percentuais da distribuição t (de Student)..... 224
Tabela A-V. Pontos percentuais superiores da distribuição F..... 225

Tabela A-I. Distribuição cumulativa binomial: $F(x; n, p) = P[X \leq x] = \sum_{t=0}^{x} C_t^n \, p^t (1-p)^{n-t}$.

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
1	0	0,990	0,950	0,900	0,800	0,750	0,700	0,600	0,500	0,400	0,300	0,200	0,100	0,050
	1	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
2	0	0,980	0,902	0.810	0,640	0,563	0,490	0,360	0,250	0,160	0,090	0,040	0,010	0,002
	1	1,000	0,997	0,990	0,960	0,938	0,910	0,840	0,750	0,640	0,510	0,360	0,190	0,097
	2		1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3	0	9,970	0,857	0,729	0,512	0,422	0,343	0,216	0,125	0,064	0,027	0,008	0,001	0,000
	1	1,000	0,993	0,972	0,896	0,844	0,784	0,648	0,500	0,352	0,216	0,104	0,028	0,007
	2		1,000	0,999	0,992	0,984	0,973	0,936	0,875	0,784	0,657	0,488	0,271	0,143
	3			1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
4	0	0,961	0,815	0,656	0,410	0,316	0,240	0,130	0,063	0,026	0,008	0,002	0,000	0,000
	1	1,000	0,986	0,948	0,819	0,738	0,652	0,475	0,313	0,179	0,084	0,027	0,004	0,000
	2		1,000	0,996	0,973	0,949	0,916	0,821	0,688	0,525	0,348	0,181	0,052	0,014
	3			1,000	0,998	0,996	0,992	0,974	0,938	0,870	0,760	0,590	0,344	0,185
	4				1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
5	0	0,951	0,774	0,590	0,328	0,237	0,168	0,078	0,031	0,010	0,002	0,000	0,000	0,000
	1	0,999	0,977	0,919	0,737	0,633	0,528	0,337	0,188	0,087	0,031	0,007	0,000	0,000
	2	1,000	0,999	0,991	0,942	0,896	0,837	0,683	0,500	0,317	0,163	0,058	0,009	0,001
	3		1,000	1,000	0,993	0,984	0,969	0,913	0,813	0,663	0,472	0,263	0,081	0,023
	4				1,000	0,999	0,998	0,990	0,969	0,922	0,832	0,672	0,410	0,226
	5					1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
6	0	0,941	0,735	0,531	0,262	0,178	0,118	0,047	0,016	0,004	0,001	0,000	0,000	0,000
	1	0,999	0,967	0,886	0,655	0,534	0,420	0,233	0,109	0,041	0,011	0,002	0,000	0,000
	2	1,000	0,998	0,984	0,901	0,831	0,744	0,544	0,344	0,179	0,070	0,017	0,001	0,000
	3		1,000	0,999	0,983	0,962	0,930	0,821	0,656	0,456	0,256	0,099	0,016	0,002
	4			1,000	0,998	0,995	0,989	0,959	0,891	0,767	0,580	0,345	0,114	0,033
	5				1,000	1,000	0,999	0,996	0,984	0,953	0,882	0,738	0,469	0,265
	6						1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
7	0	0,932	0,698	0,478	0,210	0,133	0,082	0,028	0,008	0,002	0,000	0,000	0,000	0,000
	1	0,998	0,956	0,850	0,577	0,445	0,329	0,159	0,063	0,019	0,004	0,000	0,000	0,000
	2	1,000	0,996	0,974	0,852	0,756	0,647	0,420	0,227	0,096	0,029	0,005	0,000	0,000
	3		1,000	0,997	0,967	0,929	0,874	0,710	0,500	0,290	0,126	0,033	0,003	0,000
	4			1,000	0,995	0,987	0,971	0,904	0,773	0,580	0,353	0,148	0,026	0,004
	5				1,000	0,999	0,996	0,981	0,938	0,841	0,671	0,423	0,150	0,044
	6					1,000	1,000	0,998	0,992	0,972	0,918	0,790	0,522	0,302
	7							1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
8	0	0,923	0,663	0,430	0,168	0,100	0,058	0,017	0,004	0,001	0,000	0,000	0,000	0,000
	1	0,997	0,943	0,813	0,503	0,367	0,255	0,106	0,035	0,009	0,001	0,000	0,000	0,000
	2	1,000	0,994	0,962	0,797	0,679	0,552	0,315	0,145	0,050	0,011	0,001	0,000	0,000
	3		1,000	0,995	0,944	0,886	0,806	0,594	0,363	0,174	0,058	0,010	0,000	0,000
	4			1,000	0,990	0,973	0,942	0,826	0,637	0,406	0,194	0,056	0,005	0,000
	5				0,999	0,996	0,989	0,950	0,855	0,685	0,448	0,203	0,038	0,006
	6				1,000	1,000	0,999	0,991	0,965	0,894	0,745	0,497	0,187	0,057
	7						1,000	0,999	0,996	0,983	0,942	0,832	0,570	0,337
	8							1,000	1,000	1,000	1,000	1,000	1,000	1,000
9	0	0,913	0,630	0,387	0,134	0,075	0,040	0,010	0,002	0,000	0,000	0,000	0,000	0,000
	1	0,996	0,929	0,775	0,436	0,300	0,196	0,071	0,020	0,004	0,000	0,000	0,000	0,000
	2	1,000	0,992	0,947	0,738	0,601	0,463	0,232	0,090	0,025	0,004	0,000	0,000	0,000
	3		0,999	0,992	0,914	0,834	0,730	0,483	0,254	0,099	0,025	0,003	0,000	0,000
	4		1,000	0,999	0,980	0,951	0,901	0,733	0,500	0,267	0,099	0,020	0,001	0,000
	5			1,000	0,997	0,990	0,975	0,901	0,746	0,517	0,270	0,086	0,008	0,001
	6				1,000	0,999	0,996	0,975	0,910	0,768	0,537	0,262	0,053	0,008
	7					1,000	1,000	0,996	0,980	0,929	0,804	0,564	0,225	0,071
	8							1,000	0,998	0,990	0,960	0,866	0,613	0,370
	9								1,000	1,000	1,000	1,000	1,000	1,000
10	0	0,904	0,599	0,349	0,107	0,056	0,028	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,996	0,914	0,736	0,376	0,224	0,149	0,046	0,011	0,002	0,000	0,000	0,000	0,000
	2	1,000	0,988	0,930	0,678	0,526	0,383	0,167	0,055	0,012	0,002	0,000	0,000	0,000
	3		0,999	0,987	0,879	0,776	0,650	0,382	0,172	0,055	0,011	0,001	0,000	0,000
	4		1,000	0,998	0,967	0,922	0,850	0,633	0,377	0,166	0,047	0,006	0,000	0,000
	5			1,000	0,994	0,980	0,953	0,834	0,623	0,367	0,150	0,033	0,002	0,000
	6				0,999	0,996	0,989	0,945	0,828	0,618	0,350	0,121	0,013	0,001
	7				1,000	1,000	0,998	0,988	0,945	0,833	0,617	0,322	0,070	0,012
	8						1,000	0,998	0,989	0,954	0,851	0,624	0,264	0,086
	9							1,000	0,999	0,994	0,972	0,893	0,651	0,401
	10								1,000	1,000	1,000	1,000	1,000	1,000
11	0	0,895	0,569	0,314	0,086	0,042	0,020	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,995	0,898	0,697	0,322	0,197	0,113	0,030	0,006	0,001	0,000	0,000	0,000	0,000
	2	1,000	0,985	0,910	0,617	0,455	0,313	0,119	0,033	0,006	0,001	0,000	0,000	0,000
	3		0,998	0,981	0,839	0,713	0,570	0,296	0,113	0,029	0,004	0,000	0,000	0,000
	4		1,000	0,997	0,950	0,885	0,790	0,533	0,274	0,099	0,022	0,002	0,000	0,000
	5			1,000	0,988	0,996	0,922	0.753	0,500	0,247	0,078	0,012	0,000	0,000
	6				0,998	0,992	0,978	0,901	0,726	0,467	0,210	0,050	0,003	0,000
	7				1,000	0,999	0,996	0,971	0,887	0,704	0,430	0,161	0,019	0,002
	8					1,000	0,999	0,994	0,967	0,881	0.687	0,383	0,090	0,015
	9						1,000	0,999	0,994	0,970	0,887	0,678	0,303	0,102
	10							1,000	1,000	0,996	0,980	0,914	0,686	0,431
	11									1,000	1,000	1,000	1,000	1,000

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
12	0	0,886	0,540	0,282	0,069	0,032	0,014	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,994	0,882	0,659	0,275	0,158	0,085	0,020	0,003	0,000	0,000	0,000	0,000	0,000
	2	1,000	0,980	0,889	0,558	0,391	0,253	0,083	0,019	0,003	0,000	0,000	0,000	0,000
	3		0,998	0,974	0,795	0,649	0,493	0,225	0,073	0,015	0,002	0,000	0,000	0,000
	4		1,000	0,996	0,927	0,842	0,724	0,438	0,194	0,057	0,009	0,001	0,000	0,000
	5			0,999	0,981	0,946	0,882	0,665	0,387	0,158	0,039	0,004	0,000	0,000
	6			1,000	0,996	0,986	0,961	0,842	0,613	0,335	0,118	0,019	0,001	0,000
	7				0,999	0,997	0,991	0,943	0,806	0,562	0,276	0,073	0,004	0,000
	8				1,000	1,000	0,998	0,985	0,927	0,775	0,507	0,205	0,026	0,002
	9						1,000	0,997	0,981	0,917	0,747	0,442	0,111	0,020
	10							1,000	0,997	0,980	0,915	0,725	0,341	0,118
	11								1,000	0,998	0,986	0,931	0,718	0,460
	12									1,000	1,000	1,000	1,000	1,000
13	0	0,878	0,513	0,254	0,055	0,024	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,993	0,865	0,621	0,234	0,127	0,064	0,013	0,002	0,000	0,000	0,000	0,000	0,000
	2	1,000	0,975	0,866	0,502	0,333	0,202	0,058	0,011	0,001	0,000	0,000	0,000	0,000
	3		0,997	0,966	0,747	0,584	0,421	0,169	0,046	0,008	0,001	0,000	0,000	0,000
	4		1,000	0,994	0,901	0,794	0,654	0,353	0,133	0,032	0,004	0,000	0,000	0,000
	5			0,999	0,970	0,920	0,835	0,574	0,291	0,098	0,018	0,001	0,000	0,000
	6			1,000	0,993	0,976	0,938	0,771	0,500	0,229	0,062	0,007	0,000	0,000
	7				0,999	0,994	0,982	0,902	0,709	0,426	0,165	0,030	0,001	0,000
	8				1,000	0,999	0,996	0,968	0,867	0,647	0,346	0,099	0,006	0,000
	9					1,000	0,999	0,992	0,954	0,831	0,579	0,253	0,034	0,003
	10						1,000	0,999	0,989	0,942	0,798	0,498	0,134	0,025
	11							1,000	0,998	0,987	0,936	0,766	0,379	0,135
	12								1,000	0,999	0,990	0,945	0,746	0,487
	13									1,000	1,000	1,000	1,000	1,000
14	0	0,869	0,488	0,229	0,044	0,018	0,007	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,992	0,847	0,585	0,198	0,101	0,047	0,008	0,001	0,000	0,000	0,000	0,000	0,000
	2	1,000	0,970	0,842	0,448	0,281	0,161	0,040	0,006	0,001	0,000	0,000	0,000	0,000
	3		0,996	0,956	0,698	0,521	0,355	0,124	0,029	0,004	0,000	0,000	0,000	0,000
	4		1,000	0,991	0,870	0,742	0,584	0,279	0,090	0,018	0,002	0,000	0,000	0,000
	5			0,999	0,956	0,888	0,781	0,486	0,212	0,058	0,008	0,000	0,000	0,000
	6			1,000	0,988	0,962	0,907	0,692	0,395	0,150	0,031	0,002	0,000	0,000
	7				0,998	0,990	0,969	0,850	0,605	0,308	0,093	0,012	0,000	0,000
	8				1,000	0,998	0,992	0,942	0,788	0,514	0,219	0,044	0,001	0,000
	9					1,000	0,998	0,982	0,910	0,721	0,416	0,130	0,009	0,000
	10						1,000	0,996	0,971	0,876	0,645	0,302	0,044	0,004
	11							0,999	0,994	0,960	0,839	0,552	0,158	0,030
	12							1,000	0,999	0,992	0,953	0,802	0,415	0,153
	13								1,000	0,999	0,993	0,956	0,771	0,512
	14									1,000	1,000	1,000	1,000	1,000

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
15	0	0,860	0,463	0,206	0,035	0,013	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,990	0,829	0,549	0,167	0,080	0,035	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	2	1,000	0,964	0,816	0,398	0,236	0,127	0,027	0,004	0,000	0,000	0,000	0,000	0,000
	3		0,995	0,944	0,648	0,461	0,297	0,091	0,018	0,002	0,000	0,000	0,000	0,000
	4		0,999	0,987	0,836	0,686	0,515	0,217	0,059	0,009	0,001	0,000	0,000	0,000
	5		1,000	0,998	0,939	0,852	0,722	0,403	0,151	0,034	0,004	0,000	0,000	0,000
	6			1,000	0,982	0,943	0,869	0,610	0,304	0,095	0,015	0,001	0,000	0,000
	7				0,996	0,983	0,950	0,787	0,500	0,213	0,050	0,004	0,000	0,000
	8				0,999	0,996	0,985	0,905	0,696	0,390	0,131	0,018	0,000	0,000
	9				1,000	0,999	0,996	0,996	0,849	0,597	0,278	0,061	0,002	0,000
	10					1,000	0,999	0,991	0,941	0,783	0,485	0,164	0,013	0,001
	11						1,000	0,998	0,982	0,909	0,703	0,352	0,056	0,005
	12							1,000	0,996	0,973	0,873	0,602	0,184	0,036
	13								1,000	0,995	0,965	0,833	0,451	0,171
	14									1,000	0,995	0,965	0,794	0,537
	15										1,000	1,000	1,000	1,000
16	0	0,851	0,440	0,185	0,028	0,010	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,989	0,811	0,515	0,141	0,063	0,026	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,999	0,957	0,789	0,352	0,197	0,099	0,018	0,002	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,993	0,932	0,598	0,405	0,246	0,065	0,011	0,001	0,000	0,000	0,000	0,000
	4		0,999	0,983	0,798	0,630	0,450	0,167	0,038	0,005	0,000	0,000	0,000	0,000
	5		1,000	0,997	0,918	0,810	0,660	0,329	0,105	0,019	0,002	0,000	0,000	0,000
	6			0,999	0,973	0,920	0,825	0,527	0,227	0,058	0,007	0,000	0,000	0,000
	7			1,000	0,993	0,973	0,926	0,716	0,402	0,142	0,026	0,001	0,000	0,000
	8				0,999	0,993	0,974	0,858	0,598	0,284	0,074	0,007	0,000	0,000
	9				1,000	0,998	0,993	0,942	0,773	0,473	0,175	0,027	0,001	0,000
	10					1,000	0,998	0,981	0,895	0,671	0,340	0,082	0,003	0,000
	11						1,000	0,995	0,962	0,833	0,550	0,202	0,017	0,001
	12							0,999	0,989	0,935	0,754	0,402	0,068	0,007
	13							1,000	0,998	0,982	0,901	0,648	0,211	0,043
	14								1,000	0,997	0,974	0,859	0,485	0,189
	15									1,000	0,997	0,972	0,815	0,560
	16										1,000	1,000	1,000	1,000
17	0	0,843	0,418	0,167	0,023	0,008	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,988	0,792	0,482	0,118	0,050	0,019	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,999	0,950	0,762	0,310	0,164	0,077	0,012	0,001	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,991	0,917	0,549	0,353	0,202	0,046	0,006	0,000	0,000	0,000	0,000	0,000
	4		0,999	0,978	0,758	0,574	0,389	0,126	0,025	0,003	0,000	0,000	0,000	0,000
	5		1,000	0,995	0,894	0,765	0,597	0,264	0,072	0,011	0,001	0,000	0,000	0,000
	6			0,999	0,962	0,893	0,775	0,448	0,166	0,035	0,003	0,000	0,000	0,000
	7			1,000	0,989	0,960	0,895	0,641	0,315	0,092	0,013	0,000	0,000	0,000
	8				0,997	0,988	0,960	0,801	0,500	0,199	0,040	0,003	0,000	0,000
	9				1,000	0,997	0,987	0,908	0,685	0,359	0,105	0,011	0,000	0,000

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
	10					0,999	0,997	0,965	0,834	0,552	0,225	0,038	0,001	0,000
	11					1,000	0,999	0,989	0,928	0,736	0,403	0,106	0,005	0,000
	12						1,000	0,997	0,975	0,874	0,611	0,242	0,022	0,001
	13							1,000	0,994	0,954	0,798	0,451	0,083	0,009
	14								0,999	0,988	0,923	0,690	0,238	0,050
	15								1,000	0,998	0,981	0,882	0,518	0,208
	16									1,000	0,998	0,977	0,833	0,582
	17										1,000	1,000	1,000	1,000
18	0	0,835	0,397	0,150	0,018	0,006	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,986	0,774	0,450	0,099	0,039	0,014	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,999	0,942	0,734	0,271	0,135	0,060	0,008	0,001	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,989	0,902	0,501	0,306	0,165	0,033	0,004	0,000	0,000	0,000	0,000	0,000
	4		0,998	0,972	0,716	0,519	0,333	0,094	0,015	0,001	0,000	0,000	0,000	0,000
	5		1,000	0,994	0,867	0,717	0,534	0,209	0,048	0,006	0,000	0,000	0,000	0,000
	6			0,999	0,949	0,861	0,722	0,374	0,119	0,020	0,001	0,000	0,000	0,000
	7			1,000	0,984	0,943	0,859	0,563	0,240	0,058	0,006	0,000	0,000	0,000
	8				0,996	0,981	0,940	0,737	0,407	0,135	0,021	0,001	0,000	0,000
	9				0,999	0,995	0,979	0,865	0,593	0,263	0,060	0,004	0,000	0,000
	10				1,000	0,999	0,994	0,942	0,760	0,437	0,141	0,016	0,000	0,000
	11					1,000	0,999	0,980	0,881	0,626	0,278	0,051	0,001	0,000
	12						1,000	0,994	0,952	0,791	0,466	0,133	0,006	0,000
	13							0,999	0,985	0,906	0,667	0,284	0,028	0,002
	14							1,000	0,996	0,967	0,835	0,499	0,098	0,011
	15								0,999	0,992	0,940	0,729	0,266	0,058
	16								1,000	0,999	0,986	0,901	0,550	0,226
	17									1,000	0,988	0,982	0,850	0,603
	18										1,000	1,000	1,000	1,000
19	0	0,826	0,377	0,135	0,014	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,985	0,755	0,420	0,083	0,031	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,999	0,933	0,705	0,237	0,111	0,046	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,987	0,885	0,455	0,263	0,133	0,023	0,002	0,000	0,000	0,000	0,000	0,000
	4		0,998	0,965	0,673	0,465	0,282	0,070	0,010	0,001	0,000	0,000	0,000	0,000
	5		1,000	0,991	0,837	0,668	0,474	0,163	0,032	0,003	0,000	0,000	0,000	0,000
	6			0,998	0,932	0,825	0,666	0,308	0,084	0,012	0,001	0,000	0,000	0,000
	7			1,000	0,977	0,923	0,818	0,488	0,180	0,035	0,003	0,000	0,000	0,000
	8				0,993	0,971	0,916	0,667	0,324	0,088	0,011	0,000	0,000	0,000
	9				0,998	0,991	0,967	0,814	0,500	0,186	0,033	0,002	0,000	0,000
	10				1,000	0,998	0,989	0,912	0,676	0,333	0,084	0,007	0,000	0,000
	11					1,000	0,997	0,965	0,820	0,512	0,182	0,023	0,000	0,000
	12						0,999	0,998	0,916	0,692	0,334	0,068	0,002	0,000
	13						1,000	0,997	0,968	0,837	0,526	0,163	0,009	0,000
	14							0,999	0,990	0,930	0,718	0,327	0,035	0,002
	15							1,000	0,998	0,977	0,867	0,545	0,115	0,013

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
	16								1,000	0,995	0,954	0,763	0,295	0,067
	17									0,999	0,990	0,917	0,580	0,245
	18									1,000	0,999	0,986	0,865	0,623
	19										1,000	1,000	1,000	1,000
20	0	0,818	0,358	0,122	0,012	0,003	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,983	0,736	0,392	0,069	0,024	0,008	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,999	0,925	0,677	0,206	0,091	0,035	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,984	0,867	0,411	0,225	0,107	0,016	0,001	0,000	0,000	0,000	0,000	0,000
	4		0,997	0,957	0,630	0,415	0,238	0,051	0,006	0,000	0,000	0,000	0,000	0,000
	5		1,000	0,989	0,804	0,617	0,416	0,126	0,021	0,002	0,000	0,000	0,000	0,000
	6			0,998	0,913	0,786	0,608	0,250	0,058	0,006	0,000	0,000	0,000	0,000
	7			1,000	0,968	0,898	0,772	0,416	0,132	0,021	0,001	0,000	0,000	0,000
	8				0,990	0,959	0,887	0,596	0,252	0,057	0,005	0,000	0,000	0,000
	9				0,997	0,986	0,952	0,755	0,412	0,128	0,017	0,001	0,000	0,000
	10				0,999	0,996	0,983	0,872	0,588	0,245	0,048	0,003	0,000	0,000
	11				1,000	0,999	0,995	0,943	0,748	0,404	0,113	0,010	0,000	0,000
	12					1,000	0,999	0,979	0,868	0,584	0,288	0,032	0,000	0000
	13						1,000	0,994	0,942	0,750	0,392	0,087	0,002	0,000
	14							0,998	0,979	0,874	0,584	0,196	0,011	0,000
	15							1,000	0,994	0,949	0,762	0,370	0,043	0,003
	16								0,999	0,984	0,893	0,589	0,133	0,016
	17								1,000	0,996	0,965	0,794	0,323	0,075
	18									0,999	0,992	0,931	0,608	0,264
	19									1,000	0,999	0,988	0,878	0,642
	20										1,000	1,000	1,000	1,000
25	0	0,778	0,277	0,072	0,004	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,974	0,642	0,271	0,027	0,007	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,998	0,873	0,537	0,098	0,032	0,009	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	3	1,000	0,966	0,764	0,234	0,096	0,033	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	4		0,993	0,902	0,421	0,214	0,090	0,009	0,000	0,000	0,000	0,000	0,000	0,000
	5		0,999	0,967	0,617	0,378	0,193	0,029	0,002	0,000	0,000	0,000	0,000	0,000
	6		1,000	0,991	0,780	0,561	0,341	0,074	0,007	0,000	0,000	0,000	0,000	0,000
	7			0,998	0,891	0,727	0,512	0,154	0,022	0,001	0,000	0,000	0,000	0,000
	8			1,000	0,953	0,851	0,667	0,274	0,054	0,004	0,000	0,000	0,000	0,000
	9				0,983	0,929	0,811	0,425	0,115	0,013	0,000	0,000	0,000	0,000
	10				0,994	0,970	0,902	0,586	0,212	0,034	0,002	0,000	0,000	0,000
	11				0,998	0,980	0,956	0,732	0,345	0,078	0,006	0,000	0,000	0,000
	12				1,000	0,997	0,983	0,846	0,500	0,154	0,017	0,000	0,000	0,000
	13					0,999	0,994	0,922	0,655	0,268	0,044	0,002	0,000	0,000
	14					1,000	0,998	0,966	0,788	0,414	0,098	0,006	0,000	0,000
	15						1,000	0,987	0,885	0,575	0,189	0,017	0,000	0,000
	16							0,996	0,946	0,726	0,323	0,047	0,000	0,000

Tabela A-I (continuação).

n	x	p												
		0,01	0,05	0,10	0,20	0,25	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
	17							0,999	0,978	0,846	0,488	0,109	0,002	0,000
	18							1,000	0,993	0,926	0,659	0,220	0,009	0,000
	19								0,998	0,971	0,807	0,383	0,033	0,001
	20								1,000	0,991	0,910	0,579	0,098	0,007
	21									0,998	0,967	0,766	0,236	0,034
	22									1,000	0,991	0,902	0,463	0,127
	23										0,998	0,973	0,729	0,358
	24										1,000	0,996	0,928	0,723
	25											1,000	1,000	1,000

Tabela A-II. Pontos α-percentuais superiores da distribuição normal padrão - Probabilidade da variável aleatória $Z = (X-\mu)/\sigma$ ser maior do que o valor nas margens (algarismo inteiro e primeira decimal na margem esquerda; segunda decimal na margem superior).

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0223
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0011
3,1	0,0010	0,0010	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
3,6	0,0002	0,0002	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
3,9	0,0000									

Tabela A-III. Pontos α-percentuais superiores da distribuição qui-quadrado - Probabilidade de um valor de χ^2 maior do que o valor na tabela.

n	Prob > χ^2												
	0,995	0,990	0,975	0,950	0,900	0,750	0,500	0,250	0,100	0,050	0,025	0,010	0,005
1	0,00003	0,00015	0,00098	0,00393	0,0158	0,102	0,455	1,32	2,71	3,84	5,02	6,63	7,88
2	0,01	0,0201	0,0506	0,103	0,211	0,575	1,39	2,77	4,61	5,99	7,38	9,21	10,6
3	0,0717	0,115	0,216	0,352	0,584	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,06	1,92	3,36	5,39	7,78	9,49	1,14	13,28	14,86
5	0,412	0,554	0,831	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,2	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,7	3,33	4,17	5,9	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,6	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,6	3,05	3,82	4,57	5,58	7,58	10,33	13,7	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,3	8,44	11,33	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,33	15,98	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	10,17	13,33	17,12	21,06	23,68	26,12	29,14	31,32
15	4,6	5,23	6,26	7,26	8,55	11,04	14,33	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,33	19,37	23,54	26,30	28,85	32,00	34,27
17	5,7	6,41	7,56	8,67	10,09	12,79	16,33	20,49	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	13,68	17,33	21,6	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,33	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,33	23,83	28,41	31,41	34,17	37,57	40,00
21	8,03	8,9	10,28	11,59	13,24	16,34	20,33	24,93	29,62	32,67	35,48	38,93	41,40
22	8,64	9,54	10,98	12,34	14,04	17,24	21,33	26,04	30,81	33,92	36,78	40,29	42,80
23	9,26	10,20	11,69	13,09	14,85	18,14	22,33	27,17	32,01	35,17	38,08	41,64	44,18
24	9,89	10,86	12,40	13,85	15,66	19,04	23,33	28,24	33,20	36,42	39,36	42,98	45,56
25	10,52	11,52	13,12	14,61	16,47	19,94	24,33	29,34	34,38	37,65	40,65	44,31	46,93
26	11,16	12,20	13,84	15,38	17,29	20,84	25,33	30,43	35,56	38,89	41,92	45,64	48,29
27	11,81	12,88	14,57	16,15	18,11	21,75	26,33	31,53	36,74	40,11	43,19	46,96	49,64
28	12,46	13,56	15,31	16,93	18,94	22,66	27,33	32,62	37,92	41,34	44,46	48,28	50,99
29	13,12	14,26	16,05	17,71	19,77	23,57	28,33	33,71	39,09	42,56	45,72	49,59	52,34
30	13,79	14,95	16,79	18,49	20,60	24,48	29,33	34,8	40,26	43,77	45,98	50,89	53,67
40	20,71	22,16	24,43	26,51	29,05	33,66	39,33	45,62	51,80	55,76	59,34	63,69	66,77
50	27,99	29,71	32,36	34,76	37,69	42,94	49,33	56,33	63,17	67,50	71,42	76,15	79,49
60	35,53	37,48	40,48	43,19	46,46	52,29	59,33	66,98	74,40	79,08	83,30	88,38	91,95
70	43,28	45,44	48,76	51,74	55,33	61,70	69,33	77,58	85,53	90,53	95,02	100,42	104,22
80	51,17	53,54	57,15	60,39	64,28	71,14	79,33	88,13	96,58	101,88	106,63	112,33	116,32
90	59,20	61,75	65,65	69,13	73,29	80,62	89,33	98,64	107,56	113,14	118,14	124,12	128,30
100	67,33	70,06	74,22	77,93	82,36	90,13	99,33	109,14	118,50	124,34	129,56	153,81	140,17

Tabela A-IV. Pontos percentuais da distribuição t (de Student).

ν	Pontos bilaterais superiores: Prob.($\|t\| > t_p$)								
	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
35	0,682	0,852	1,052	1,306	1,690	2,030	2,438	2,724	3,591
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
45	0,680	0,850	1,048	1,301	1,680	2,014	2,412	2,690	3,520
50	0,680	0,849	1,047	1,299	1,676	2,008	2,403	2,678	3,496
55	0,679	0,849	1,047	1,297	1,673	2,004	2,396	2,669	3,476
60	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
70	0,678	0,847	1,045	1,294	1,667	1,994	2,381	2,648	3,435
80	0,678	0,847	1,044	1,293	1,665	1,989	2,374	2,638	3,416
90	0,678	0,846	1,043	1,291	1,662	1,986	2,369	2,631	3,402
100	0,677	0,846	1,042	1,290	1,661	1,982	2,365	2,625	3,390
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
200	0,676	0,844	1,039	1,286	1,653	1,972	2,345	2,601	3,340
500	0,676	0,843	1,037	1,284	1,648	1,965	2,334	2,586	3,310
1000	0,675	0,842	1,037	1,283	1,647	1,962	2,330	2,581	3,301
Inf.	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291
ν	0,25	0,20	0,15	0,10	0,05	0,025	0,01	0,005	0,0005
	Pontos unilaterais superiores: Prob.($t > t_p$)								

Tabela A-V. Pontos percentuais superiores da distribuição F: Prob.[$F(\nu_1,\nu_2) > F_P$].

ν_2	P	ν_1																			
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
1	0,05	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,0	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
	0,025	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6	976,7	984,9	984,9	993,1	997,2	1001,	1006,	1010,	1014,	1018,
	0,01	4052,	5000,	5403,	5625,	5764,	5859,	5928,	5982,	6022,	6056,	6082,	6106,	6157,	6209,	6235,	6261,	6287,	6313,	6339,	6366,
	0,001	4053*	5000*	5404*	5625*	5764*	5859*	5929*	5981*	6023*	6056*	6084*	6107*	6158*	6209*	6235*	6261*	6287*	6313*	6340*	6366*
2	0,05	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
	0,025	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40	39,41	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
	0,01	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,41	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
	0,001	998,5	999,0	999,2	999,2	999,3	999,3	999,4	999,4	999,4	999,4	999,4	999,4	999,4	999,4	999,5	999,5	999,5	999,5	999,5	999,5
3	0,05	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
	0,025	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42	14,34	14,25	39,43	14,17	14,12	14,08	14,04	13,99	13,95	13,90
	0,01	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,13	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
	0,001	167,0	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2	128,8	128,3	127,4	126,4	125,9	125,4	125,0	124,5	124,0	123,5
4	0,05	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
	0,025	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84	8,75	8,66	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
	0,01	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,45	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
	0,001	74,14	61,25	56,18	53,44	51,71	50,53	49,66	49,00	48,47	48,05	47,70	47,41	46,76	46,10	45,77	45,43	45,09	44,75	44,40	44,05
5	0,05	6,61	5,79	5,41	5,19	5,05	4,95	4,88	8,82	4,77	4,74	4,70	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
	0,025	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62	6,52	6,43	6,46	6,33	6,28	6,23	6,18	6,12	6,07	6,02
	0,01	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,96	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
	0,001	47,18	37,12	33,20	31,09	29,75	28,84	28,16	27,64	27,24	26,92	26,64	26,42	25,91	25,39	25,14	24,87	24,60	24,33	24,06	23,79
6	0,05	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
	0,025	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46	5,37	5,27	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
	0,01	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
	0,001	35,51	27,00	23,70	21,92	20,81	20,03	19,46	19,03	18,69	18,41	18,18	17,99	17,56	17,12	16,89	16,67	16,44	16,21	15,99	15,75
7	0,05	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	6,60	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
	0,025	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76	4,67	4,57	3,51	4,47	4,42	4,36	4,31	4,25	4,20	4,14
	0,01	12,25	9,55	8,45	7,85	4,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
	0,001	29,25	21,69	18,77	17,19	16,21	15,52	15,02	14,63	14,33	14,08	13,88	13,71	13,32	12,93	12,73	12,53	12,33	12,12	11,91	11,70
8	0,05	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
	0,025	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30	4,20	4,10	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
	0,01	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,74	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
	0,001	25,42	18,49	15,83	14,39	13,49	12,86	12,40	12,04	11,77	11,54	11,35	11,19	10,84	10,48	10,30	10,11	9,92	9,73	9,53	9,33
9	0,05	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
	0,025	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96	3,87	3,77	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
	0,01	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
	0,001	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89	9,72	9,57	9,24	8,90	8,72	8,55	8,37	8,19	8,00	7,81
10	0,05	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
	0,025	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72	3,62	3,52	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
	0,01	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,78	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
	0,001	21,04	14,91	12,55	11,28	10,48	9,92	9,52	9,20	8,96	8,75	8,59	8,45	8,13	7,80	7,64	7,47	7,30	7,12	6,94	6,76
11	0,05	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
	0,025	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53	3,43	3,33	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
	0,01	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
	0,001	19,69	13,81	11,56	10,35	9,58	9,05	8,66	8,35	8,12	7,92	7,76	7,63	7,32	7,01	6,85	6,68	6,52	6,35	6,17	6,00
12	0,05	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
	0,025	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37	3,28	3,18	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
	0,01	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
	0,001	18,64	12,97	10,80	9,63	8,89	9,38	8,00	7,71	7,48	7,29	7,14	7,00	6,71	6,40	6,25	6,09	5,93	5,76	5,59	5,42

* Estes valores devem ser multiplicados por 100.

Tabela A-V (continuação)

ν_2	P	ν_1																			
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
13	0,05	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
	0,025	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25	3,15	3,05	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
	0,01	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
	0,001	17,81	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80	6,65	6,52	6,23	5,93	5,78	5,63	5,47	5,30	5,14	4,97
14	0,05	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,56	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
	0,025	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15	3,05	2,95	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
	0,01	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
	0,001	17,14	11,78	9,73	8,62	7,92	7,43	7,08	6,80	6,58	6,40	6,26	6,13	5,85	5,56	5,41	5,25	5,10	4,94	4,77	4,60
15	0,05	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
	0,025	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06	2,96	2,86	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
	0,01	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
	0,001	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08	5,94	5,81	5,54	5,25	5,10	4,95	4,80	4,64	4,47	4,31
16	0,05	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
	0,025	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99	2,89	2,79	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
	0,01	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
	0,001	16,12	10,97	9,00	7,94	7,27	6,81	6,46	6,19	5,98	5,81	5,67	5,55	5,27	4,99	4,85	4,70	4,54	4,39	4,23	4,06
17	0,05	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
	0,025	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92	2,82	2,72	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
	0,01	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
	0,001	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58	5,44	5,32	5,05	4,78	4,63	4,48	4,33	4,18	4,02	3,85
18	0,05	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
	0,025	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87	2,77	2,67	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
	0,01	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,44	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
	0,001	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39	5,25	5,13	4,87	4,59	4,45	4,30	4,15	4,00	3,84	3,67
19	0,05	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
	0,025	5,92	4,51	3,90	3,36	3,33	3,17	3,05	2,96	2,88	2,82	2,72	2,62	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
	0,01	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,36	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
	0,001	15,08	10,16	8,28	7,26	6,62	6,18	5,85	5,59	5,39	5,22	5,08	4,97	4,70	4,43	4,29	4,14	3,99	3,84	3,68	3,51
20	0,05	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
	0,025	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77	2,68	2,57	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
	0,01	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,30	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
	0,001	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08	4,94	4,82	4,56	4,29	4,15	4,00	3,86	3,70	3,54	3,38
21	0,05	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
	0,025	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73	2,64	2,53	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
	0,01	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,24	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
	0,001	14,59	9,77	7,94	6,95	6,32	5,88	5,56	5,31	5,11	4,95	4,81	4,70	4,44	4,17	4,03	3,88	3,74	3,58	3,42	3,26
22	0,05	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
	0,025	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70	2,60	2,50	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
	0,01	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,18	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
	0,001	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83	4,70	4,58	4,33	4,06	3,92	3,78	3,63	3,48	3,32	3,15
23	0,05	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
	0,025	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67	2,57	2,47	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
	0,01	7,88	5,66	4,76	4,26	3,94	9,71	3,54	3,41	3,30	3,21	3,14	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
	0,001	14,19	9,47	7,67	6,69	6,08	5,65	5,33	5,09	4,89	4,73	4,60	4,48	4,23	3,96	3,82	3,68	3,53	3,38	3,22	3,05
24	0,05	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
	0,025	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64	2,54	2,44	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
	0,01	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,09	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
	0,001	14,03	9,34	7,55	6,59	5,98	5,55	5,23	4,99	4,80	4,64	4,51	4,39	4,14	3,87	3,74	3,59	3,45	3,29	3,14	2,97

Tabela A-V (continuação)

ν_2	P	ν_1																			
		1	2	3	4	5	6	7	8	9	10	11	12	15	20	24	30	40	60	120	Inf.
	0,05	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
25	0,025	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61	2,51	2,41	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
	0,01	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	3,05	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
	0,001	13,88	9,22	7,45	6,49	5,88	5,46	5,15	4,91	4,71	4,56	4,42	4,31	4,06	3,79	3,66	3,52	3,37	3,22	3,06	2,89
	0,05	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
26	0,025	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59	2,49	2,39	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
	0,01	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	3,02	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
	0,001	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48	4,35	4,24	3,99	3,72	3,59	3,44	3,30	3,15	2,99	2,82
	0,05	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,16	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
27	0,025	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57	2,47	2,36	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
	0,01	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,98	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
	0,001	13,61	9,02	7,27	6,33	5,73	5,31	5,00	4,76	4,57	4,41	4,28	4,17	3,92	3,66	3,52	3,38	3,23	3,08	2,92	2,75
	0,05	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
28	0,025	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55	2,45	2,34	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
	0,01	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,95	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
	0,001	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35	4,22	4,11	3,86	3,60	3,46	3,32	3,18	3,02	2,86	2,69
	0,05	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
29	0,025	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53	2,43	2,32	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
	0,01	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,92	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
	0,001	13,39	8,85	7,12	6,19	5,59	5,18	4,87	4,64	4,45	4,29	4,16	4,05	3,80	3,54	3,41	3,27	3,12	2,97	2,81	2,64
	0,05	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,12	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
30	0,025	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51	2,41	2,31	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
	0,01	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,90	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
	0,001	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24	4,11	4,00	3,75	3,49	3,36	3,22	3,07	2,92	2,76	2,59
	0,05	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
40	0,025	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39	2,29	2,18	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
	0,01	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,73	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
	0,001	12,61	8,25	6,60	5,70	5,13	4,73	4,44	4,21	4,02	3,87	3,75	3,64	3,40	3,15	3,01	2,87	2,73	2,57	2,41	2,23
	0,05	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
60	0,025	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27	2,17	2,06	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
	0,01	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,56	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
	0,001	11,97	7,76	6,17	5,31	4,76	4,37	4,09	3,87	3,69	3,54	3,42	3,31	3,08	2,83	2,69	2,55	2,41	2,25	2,08	1,89
	0,05	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,86	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
120	0,025	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16	2,05	1,94	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
	0,01	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,40	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
	0,001	11,38	7,32	5,79	4,95	4,42	4,04	3,77	3,55	3,38	3,24	3,12	3,02	2,78	2,53	2,40	2,26	2,11	1,95	1,76	1,54
	0,05	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,79	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
Inf.	0,025	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05	1,94	1,83	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00
	0,01	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,24	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00
	0,001	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96	2,84	2,74	2,51	2,27	2,13	1,99	1,84	1,66	1,45	1,00

BIBLIOGRAFIA

BARBETTA, P.A. **Estatística Aplicada às Ciências Sociais**. Florianópolis: Editora da UFSC. 1998. 316p.

BARBETTA, P.A.; REIS, M.M.; BORNIA, A.C. **Estatística para Cursos de Engenharia e Informática**. São Paulo: Atlas. 2004. 416p.

BUSSAB, W. O. MORETTIN, P. A. **Estatística Básica.** 5. ed. São Paulo: Saraiva. 2002. 526 p.

BLACKWELL, D. **Estatística Básica**. São Paulo: McGraw-Hill do Brasil. 1974. 143p.

CALZADA B., J. Introducción a la estadística. Lima, Perú: El Estudiante. 1969. 244p.

CORRÊA DA SILVA, J.G. **Estatística Básica**. Versão preliminar. Instituto de Física e Matemática, Universidade Federal de Pelotas. Pelotas, 1992. 173p.

DEVORE, J. **Probability and Statistics for Engineering and the Sciences**. Monterrey, California: Brooks/Cole. 1982. 640p.

FONSECA, J.S.; MARTINS, G.A. **Curso de Estatística**. 6. ed. São Paulo: Atlas. 1996. 318p.

FREUND, J.E. Mathematical statistics. 2. ed. London: Prentice-Hall. 1972. 463p.

FREUND, J.E.; SIMON, G.A. **Estatística Aplicada - Economia, Administração e Contabilidade**. 9. ed. Porto Alegre: Bookman. 2000.404p.

FREUND, J.E.; WALPOLE, R.E. **Mathematical Statistics**, 3. ed. New Jersey: Englewood Cliffs. 1980. 548p.

HOEL, P.G. **Estatística Elementar**. São Paulo: Atlas. 1980. 430p.

LAPIN, L. **Statistics - Meaning and Method**. 2. ed. New York: Harcourt Brace Jovanovich. 1980. 543p.

LEVIN, J. **Estatística Aplicada a Ciências Humanas**. 2. ed. São Paulo: Harper & Row do Brasil. 1987. 392p.

LOPES, P.A. **Probabilidades & Estatística - Conceitos, Modelos, Aplicações em Excel**. Rio de Janeiro: Reichmann & Affonso. 1999. 174p.

MENDENHALL, W. **Introduction to Probability and Statistics**. 2. ed. Belmont, California: Wadsworth. 1969. 393p.

MEYER, P.L. **Probabilidade - Aplicações à Estatística**. 2. ed. Rio de Janeiro: Livros Técnicos e Científicos. 2000. 426p.

MILLER, I.; FREUND, J.E. **Probability and Statistics for Engineers**. New Jersey: Englewood Cliffs. 1965. 432p.

MONTGOMERY, D.C. e RUNGER, G.C. **Estatística Aplicada e Probabilidade para Engenheiros.** Rio de Janeiro: Livros Técnicos e Científicos. 2003. 720p.

MONTGOMERY, D.C.; RUNGER, G.C.; HUBELE. N.F. **Estatística Aplicada à Engenharia**. 2. ed. Rio de Janeiro: Livros Técnicos e Científicos. 2004. 335p.

MOORE, D.S. **A Estatística Básica e sua Prática**. Rio de Janeiro: Livros Técnicos e Científicos. 2000. 482p.

MOORE, D.S. **The Active Practice of Statistics - A Text for Multimedia Learning**. New York: W. H. Freeman. 1997. 432p.

MOORE, D.S. **Statistics: Concepts and Controversies**. San Francisco: W. H. Freeman. 1979. 313p.

MORETTIN, P.A. **Introdução à Estatística para Ciências Exatas**. São Paulo: Atual Editora Ltda. 1981. 211p.

NICK, E.; KELLNER, S.R.O. **Fundamentos de Estatística para as Ciências do Comportamento**. Rio de Janeiro: Renes. 1971. 316p.

PARADINE, C.G.; RIVETT, B.H.P. **Métodos Estatísticos para Tecnologistas**. São Paulo: Polígono/EDUSP. 1974. 350p.

PARADINE, C.G.; RIVETT, B.H.P. **Métodos Estatísticos para Tecnologistas**. São Paulo: Polígono/Universidade de São Paulo. 1974. 350p.

PIMENTEL GOMES, F. **Iniciação à Estatística**. 6. ed. São Paulo; Livraria Nobel S. A. 1978. 211p.

SILVEIRA JÚNIOR, P.S.; MACHADO, A.A.; ZONTA, E.P.; SILVA, J.B. **Curso de Estatística**, vol. 1. Pelotas: Editora Universitária, UFPEL. 1989. 135p.

SILVEIRA JÚNIOR, P.S.; MACHADO, A.A.; ZONTA, E.P.; SILVA, J.B. **Curso de Estatística**, vol. 2. Pelotas: Editora Universitária, UFPEL. 1992. 234p.

SPIEGEL, M.R. **Estatística**. São Paulo: McGraw-Hill do Brasil. 1975. 580p.

WALPOLE, R.E. **Introduction to Statistics**. New York: Macmillan. 1968. 365p.

WALPOLE, R.E.; MYERS, R.H. **Probability and Statistics for Engineers and Scientists**. 2. ed. New York: Macmillan. 1978. 580p.