

CE001 - BIOESTATÍSTICA - Turma N

Silvia Shimakura
LEG-UFPR
silvia.shimakura@ufpr.br

Este curso tem como objetivo apresentar os conceitos fundamentais e princípios básicos da Estatística através de exemplos.

Horários: Qua e Sex, 18:30-20:10 (Turma N)

Local: Setor de Ciências Biológicas - Anfiteatro 4

Monitoria no LABEST-Centro Politécnico: Horários a serem definidos

Ementa da disciplina: Para ver a ementa da disciplina clique aqui.

Avaliação: Quatro avaliações de igual peso e o exame final.

- Avaliação 1: 27/03
- Avaliação 2: 08/05
- Avaliação 3: 20/05, 22/05, 27/05, 29/05, 03/06: Seminários em grupo
- Avaliação 4: 12/06
- Exame Final: 08/07

Exemplos de provas: • Prova 1

- Prova 2
- Prova 3

Notas e avisos: Estarão disponíveis no aqui.

Programa computacional utilizado: Durante o semestre será utilizado, para fins didáticos, o programa estatístico R que é gratuito e de código aberto.

- Bibliografia:**
1. Pagano, M.; Gauvreau, K. (2004) *Princípios de Bioestatística*. Editora Pioneira Thomson Learning. (Livro texto do curso).
 2. Bussab, W.O.; Morettin, P.A. (1981) *Estatística Básica*. Editora Atual.
 3. Robert R. Sokal; F. James Rohlf (2012) *Biometry: the principles and practice of statistics in biological research*. Editora W. H. Freeman and Company.
 4. Nogueira, M.L.G.; Nunes, L.L.C.; Pinto, D.; Ribeiro, A.J.F.; Silva, C.Q.; Siqueira, A.L. Introdução à Bioestatística. Apostila da UFMG.

5. Reis, E.A.; Reis, I.A. (2001). Análise Descritiva de Dados - Tabelas e Gráficos. Relatório Técnico RTE-04/2001, Depto Estatística-UFMG.
6. Reis, E.A.; Reis, I.A. (2000). Exercícios resolvidos em Introdução à Bioestatística. Relatório Técnico RTE-03/2000, Depto Estatística-UFMG.

Página wiki do curso: O planejamento e cronograma das aulas estão disponíveis aqui.

Material do curso: As aulas serão dadas no estilo tutorial e estão disponíveis para download e/ou para acesso aqui.

Material extra:

1. Artigo, revista Superinteressante, Março 2009
2. Uso da distribuição binomial para quantificar evidência.
3. Exemplo de motivação para a distribuição Normal.
4. Tabela da Distribuição Normal Padrão
5. Tabelas de distribuições estatísticas.

1 Conteúdo

1. **Introdução:** O que é Estatística? Qual é o papel da Estatística na Ciência?
2. **Análise exploratória de dados:** sumário de dados, gráfico de barras, gráfico de setores, histograma, ramo-e-folhas, mediana, moda, desvio padrão, amplitude inter-quartis,...
3. **Noções de probabilidade:** definições, avaliação da qualidade de testes diagnósticos
4. **Distribuições de probabilidade:** distribuição binomial, distribuição normal
5. **Populações e amostras:** usando amostras para aprender sobre a população
6. **Intervalos de confiança:** estimando a média populacional a partir de uma amostra
7. **Testes de hipóteses:** idéia básica e testes para uma amostra
8. **Comparação de dois grupos:** As mensurações num grupo tendem a ser maiores em média do que em outro?
9. **Correlação:** verificando se os valores de duas quantidades tendem a ser relacionadas
10. **Regressão:** descrevendo como o comportamento de uma quantidade muda com o valor da outra

2 Introdução

2.1 O que é Estatística?

Estatística é um conjunto de métodos usados para se analisar dados. A Estatística pode ser aplicada em praticamente todas as áreas do conhecimento humano e em algumas áreas recebe um nome especial. Este é o caso da Bioestatística, que trata de aplicações da Estatística em Ciências Biológicas e da Saúde.

A palavra "Estatística" tem *peelo menos* três significados:

1. coleção de informações numéricas ou *dados*,
2. medidas resultantes de um conjunto de dados, como por exemplo médias,
3. métodos usados na coleta e interpretação de dados.

Razões para se estudar Estatística?

- A disponibilidade de aparelhos modernos, muitos dos quais acoplados a computadores, permitem a quantificação de muitos fenômenos. A massa de *dados* gerada precisa ser analisada adequadamente.
- Na Ciência, são realizados estudos *experimentais* ou *observacionais*, em que o interesse é comparar grupos/tratamentos ou ainda determinar fatores prognósticos/risco importantes.
- O material biológico estudado é sempre uma *amostra* e o objetivo final é tirar conclusões sobre toda a *população* de interesse com base na amostra.

Em geral, a disciplina de estatística refere-se a métodos para coleta e descrição dos dados, e então a verificação da força da evidência nos dados pró ou contra certas idéias científicas. A presença de uma *variação* não previsível nos dados faz disso uma tarefa pouco trivial.

2.2 Variação Amostral

Alguns exemplos em que a variação está presente nos dados.

1. Função pulmonar em pacientes com fibrose cística

A pressão inspiratória estática máxima (PI_{max}) é um índice de vigor respiratório muscular. Os seguintes dados mostram a idade (anos) e uma medida de PI_{max} (cm H₂O) de 25 pacientes com fibrose cística.

| Sujeito | Idade | PI _{max} |
|---------|-------|-------------------|
| 1 | 7 | 80 |
| 2 | 7 | 85 |
| 3 | 8 | 110 |
| 4 | 8 | 95 |
| 5 | 8 | 95 |
| 6 | 9 | 100 |
| 7 | 11 | 45 |
| 8 | 12 | 95 |
| 9 | 12 | 130 |
| 10 | 13 | 75 |
| 11 | 13 | 80 |
| 12 | 14 | 70 |
| 13 | 14 | 80 |
| 14 | 15 | 100 |
| 15 | 16 | 120 |
| 16 | 17 | 110 |
| 17 | 17 | 125 |
| 18 | 17 | 75 |
| 19 | 17 | 100 |
| 20 | 19 | 40 |
| 21 | 19 | 75 |
| 22 | 20 | 110 |
| 23 | 23 | 150 |
| 24 | 23 | 75 |
| 25 | 23 | 95 |

- (a) Todos os pacientes com fibrose cística têm o mesmo valor de PI_{max}?
- (b) Assumindo que a idade não afeta PI_{max}, qual é um valor de PI_{max} típico para pacientes com fibrose cística?
- (c) Quão grande é a variabilidade em torno deste valor típico?
- (d) Será que a suposição de que idade não afeta PI_{max} consistente com os dados?
- (e) Se idade na verdade afeta PI_{max}, como você descreveria o valor típico de PI_{max} e variabilidade?
- (f) Que tipo de representação gráfica poderia ser utilizada para visualizar adequadamente estes dados?

2. Conteúdo de gordura e proteína no leite

Cientistas mediram o conteúdo de gordura e proteína em amostras de leite de 10 focas cinza.

| Foca | Gordura % | Proteína % |
|------|-----------|------------|
| 1 | 57.2 | 10.4 |
| 2 | 58.3 | 9.4 |
| 3 | 53.9 | 11.9 |
| 4 | 48.0 | 12.4 |
| 5 | 57.8 | 12.1 |
| 6 | 54.1 | 8.5 |
| 7 | 55.6 | 10.4 |
| 8 | 49.3 | 11.6 |
| 9 | 48.8 | 11.4 |
| 10 | 53.8 | 10.8 |

- (a) Os percentuais são exatamente os mesmos de um animal para outro?
- (b) Baseado nesta amostra de 10 focas, os cientistas estimaram o conteúdo de gordura no leite de focas cinza com sendo 53.7%. Se eles agora coletarem mais amostras de leite de outras 10 focas, você esperaria que o novo valor estimado fosse 53.7%?
- (c) Como o tamanho de amostra influencia sua resposta?
- (d) O que aconteceria se eles tomassem um outro conjunto de amostras das mesmas 10 focas? Você esperaria obter a mesma estimativa neste caso?
- (e) O que aconteceria se uma fração do material coletado inicialmente das 10 focas fosse re-analisado? Você esperaria obter a mesma estimativa neste caso?

Pode-se dizer que cada medida pode ser constituída de três fontes de variação: Variação biológica, variação temporal e variação devido à erros de medida.

3 Estatística Descritiva - Tabelas e Gráficos

Edna A. Reis e Ilka A. Reis
Relatório Técnico RTE-04/2001
Departamento de Estatística-UFMG

A coleta de dados estatísticos tem crescido muito nos últimos anos em todas as áreas de pesquisa, especialmente com o advento dos computadores e surgimento de softwares cada vez mais sofisticados. Ao mesmo tempo, olhar uma extensa listagem de dados coletados não permite obter praticamente nenhuma conclusão, especialmente para grandes conjuntos de dados, com muitas características sendo investigadas.

Utilizamos métodos de Estatística Descritiva para **organizar**, **resumir** e **descrever** os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos.

As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

Ao se condensar os dados, perde-se informação, pois não se têm as observações originais. Entretanto, esta perda de informação é pequena se comparada ao ganho que se tem com a clareza da interpretação proporcionada.

A descrição dos dados também tem como objetivo **identificar anomalias**, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto. Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais freqüente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

Ao mesmo tempo que o uso das ferramentas estatísticas vem crescendo, aumenta também o **abuso** de tais ferramentas. É muito comum vermos em jornais e revistas, até mesmo em periódicos científicos, gráficos voluntariamente ou intencionalmente enganosos e estatísticas obscuras para justificar argumentos polêmicos.

3.1 Coleta e Armazenamento de Dados

Exemplo Inicial: Ursos Marrons

Pesquisadores do Instituto Amigos do Urso têm estudado o desenvolvimento dos ursos marrons selvagens que vivem em uma certa floresta do Canadá. O objetivo do projeto é estudar algumas características dos ursos, tais como seu peso e altura, ao longo da vida desses animais.

A ficha de coleta de dados, representada na Figura 1, mostra as características que serão estudadas na primeira fase do projeto. Na primeira parte do estudo, 97 ursos foram identificados (por nome), pesados e medidos. Os dados foram coletados através do preenchimento da ficha de coleta.

Para que os ursos possam ser identificados, medidos e avaliados, os pesquisadores precisam

INSTITUTO AMIGOS DO URSO

- Nome do animal: *Allen*
- Sexo: *macho*
- Idade: *19* (meses)
- Cabeça: - comprimento: *25,4* cm
- largura: *17,7* cm
- Pescoço: - perímetro: *38,1* cm
- Tórax: - perímetro: *58,4* cm
- Altura: *114,3* cm
- Peso: *29,51* kg

Data da coleta : *02 / 07 / 98*

Nome do funcionário responsável pela coleta:
Pedro Luis Rocha

Figura 1: Ficha de coleta de dados dos ursos marrons.

anestesiá-los. Mesmo assim, medidas como a do peso são difíceis de serem feitas (qual será o tamanho de uma balança para pesar ursos?). Desse modo, os pesquisadores gostariam também de encontrar uma maneira de estimar o peso do urso através de uma outra medida mais fácil de se obter, como uma medida de comprimento, por exemplo (altura, circunferência do tórax, etc.). Nesse caso, só seria necessária uma grande fita métrica, o que facilitaria muito a coleta de dados das próximas fases do projeto.

Geralmente, as coletas de dados são feitas através do preenchimento de fichas pelo pesquisador e/ou através de resposta a questionários (o que não foi o caso dos ursos é claro!). Alguns dados são coletados através de medições (altura, peso, pressão sanguínea, etc.), enquanto outros são coletados através de avaliações (sexo, cor, raça, espécie, etc.).

Depois de coletados, os dados devem ser armazenados e sistematizados numa planilha de dados, como mostra a Figura 2. Hoje em dia, essas planilhas são digitais e essa é a maneira de realizar a entrada dos dados num programa de computador.

VARIÁVEIS →

| | Nome | Mês Obs. | Idade | Sexo | Cabeça Comp. | Cabeça Larg. | Pescoço Perí. | Altura | Tórax Perí. | Peso | |
|----------------|---------|----------|-------|-------|--------------|--------------|---------------|--------|-------------|-------|-------|
| ELEMENTOS ↓ | 1 | Allen | jul | 19 | macho | 25,4 | 12,7 | 38,1 | 114,3 | 58,4 | 29,5 |
| | 2 | Berta | jul | 19 | fêmea | 27,9 | 16,5 | 50,8 | 120,7 | 61,0 | 31,8 |
| | 3 | Clyde | jul | 19 | macho | 27,9 | 14,0 | 40,6 | 134,6 | 66,0 | 36,3 |
| | 4 | Doc | jul | 55 | macho | 41,9 | 22,9 | 71,1 | 171,5 | 114,3 | 156,2 |
| | 5 | Quincy | set | 81 | macho | 39,4 | 20,3 | 78,7 | 182,9 | 137,2 | 188,9 |
| | 6 | Kooch | out | * | macho | 40,6 | 20,3 | 81,3 | 195,6 | 132,1 | 196,1 |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | 93 | Sara | ago | * | fêmea | 30,5 | 12,7 | 45,7 | 142,2 | 82,6 | 51,8 |
| | 94 | Lou | ago | * | macho | 30,5 | 14,0 | 38,1 | 129,5 | 61,0 | 37,2 |
| | 95 | Molly | ago | * | fêmea | 33,0 | 15,2 | 55,9 | 154,9 | 101,6 | 104,4 |
| 96 | Graham | jul | * | macho | 30,5 | 10,2 | 44,5 | 149,9 | 72,4 | 58,1 | |
| 97 | Jeffrey | jul | * | macho | 34,3 | 15,2 | 50,8 | 157,5 | 82,6 | 70,8 | |

Figura 2: Representação parcial da planilha de dados do exemplo dos ursos.

A planilha de dados é composta por linhas e colunas. Cada linha contém os dados de uma unidade experimental (urso), ou seja de uma ficha de coleta. As características (variáveis) são dispostas em colunas. Assim, a planilha de dados contém um número de linhas igual a número de participantes do estudo e um número de colunas igual ao número de variáveis sendo estudadas.

A planilha de dados dos ursos tem 97 linhas e 10 colunas. Alguns ursos não tiveram sua idade determinada. Esses dados são chamados dados faltantes e é comum representá-los por asteriscos (na verdade, cada software tem sua convenção para representar missing data).

3.2 Tipos de variáveis

Variável é a característica de interesse que é medida em cada elemento da amostra ou população. Como o nome diz, seus valores variam de elemento para elemento. As variáveis podem ter valores numéricos ou não numéricos.

Variáveis podem ser classificadas da seguinte forma:

1. **Variáveis Quantitativas:** são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser contínuas ou discretas.
 - (a) **Variáveis discretas:** características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens. Exemplos: número

de filhos, número de bactérias por litro de leite, número de cigarros fumados por dia.

- (b) **Variáveis contínuas**, características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido. Usualmente devem ser medidas através de algum instrumento. Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade.

2. **Variáveis Qualitativas (ou categóricas)**: são as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

- (a) **Variáveis nominais**: não existe ordenação dentre as categorias. Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
- (b) **Variáveis ordinais**: existe uma ordenação entre as categorias. Exemplos: escolaridade (1o, 2o, 3o graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).

As distinções são menos rígidas do que a descrição acima insinua.

Uma variável originalmente quantitativa pode ser coletada de forma qualitativa.

Por exemplo, a variável idade, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos, etc...), é qualitativa (ordinal). Outro exemplo é o peso dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado, etc.).

Outro ponto importante é que *nem sempre uma variável representada por números é quantitativa.*

O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável sexo passou a ser quantitativa!

Exemplo do ursos marrons (continuação):

No conjunto de dados ursos marrons, são qualitativas as variáveis sexo (nominal) e mês da observação (ordinal); são quantitativas contínuas as demais: idade, comprimento da cabeça, largura da cabeça, perímetro do pescoço, perímetro do tórax, altura e peso.

3.3 Estudando a Distribuição de Frequências de uma Variável

Como já sabemos, as variáveis de um estudo dividem-se em quatro tipos: qualitativas (nominais e ordinais) e quantitativas (discretas e contínuas). Os dados gerados por esses tipos de variáveis são de naturezas diferentes e devem receber tratamentos diferentes. Portanto, vamos estudar as ferramentas - **tabelas e gráficos** - mais adequados para cada tipo de dados, separadamente.

3.3.1 Variáveis Qualitativas - Nominais e Ordinais

Iniciaremos essa apresentação com os dados de natureza qualitativa, que são os mais fáceis de tratar do ponto de vista da análise descritiva.

No exemplo dos ursos, uma das duas variáveis qualitativas presentes é o **sexo** dos animais.

Para organizar os dados provenientes de uma variável qualitativa, é usual fazer uma tabela de frequências, como a Tabela 1, onde estão apresentadas as frequências com que ocorrem cada um dos sexos no total dos 97 ursos observados.

Cada categoria da variável sexo (feminino, masculino) é representada numa linha da tabela. Há uma coluna com as contagens de ursos em cada categoria (frequência absoluta) e outra com os percentuais que essas contagens representam no total de ursos (frequência relativa). Esse tipo de tabela representa a distribuição de frequências dos ursos segundo a variável sexo.

Como a variável sexo é **qualitativa nominal**, isto é, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma.

Tabela 1: Distribuição de frequências dos ursos segundo sexo.

| Sexo | Frequência Absoluta | Frequência Relativa (%) |
|-----------|---------------------|-------------------------|
| Feminino | 35 | 36,1 |
| Masculino | 62 | 63,9 |
| Total | 97 | 100,0 |

Quando a variável tabelada for do tipo **qualitativa ordinal**, as linhas da tabela de frequências devem ser dispostas na ordem existente para as categorias.

A Tabela 2 mostra a distribuição de frequências dos ursos segundo o mês de observação, que é uma variável qualitativa ordinal. Nesse caso, podemos acrescentar mais duas colunas com as frequências acumuladas (absoluta e relativa), que mostram, para cada mês, a frequência de ursos observados até aquele mês. Por exemplo, até o mês de julho, foram observados 31 ursos, o que representa 32,0% do total de ursos estudados.

A visualização da distribuição de frequências de uma variável fica mais fácil se fizermos um gráfico a partir da tabela de frequências. Existem vários tipos de gráficos, dependendo do tipo de variável a ser representada. Para as variáveis do tipo qualitativas, abordaremos dois tipos de gráficos: **os de setores e os de barras**.

Os gráficos de setores, mais conhecidos como gráficos de pizza ou torta, são construídos dividindo-se um círculo (pizza) em setores (fatias), um para cada categoria, que serão proporcionais à frequência daquela categoria.

Tabela 2: Distribuição de freqüências dos ursos segundo mês de observação.

| Mês de Observação | Freqüências Simples | | Freqüências Acumuladas | |
|-------------------|---------------------|-------------------------|-------------------------------|-------------------------------|
| | Freqüência Absoluta | Freqüência Relativa (%) | Freqüência Absoluta Acumulada | Freqüência Relativa Acumulada |
| Abril | 8 | 8,3 | 8 | 8,3 |
| Mai | 6 | 6,2 | 14 | 14,5 |
| Junho | 6 | 6,2 | 20 | 20,7 |
| Julho | 11 | 11,3 | 31 | 32,0 |
| Agosto | 23 | 23,7 | 54 | 55,7 |
| Setembro | 20 | 20,6 | 74 | 76,3 |
| Outubro | 14 | 14,4 | 88 | 90,7 |
| Novembro | 9 | 9,3 | 97 | 100,0 |
| Total | 97 | 100,0 | — | — |

A Figura 3 mostra um gráfico de setores para a variável sexo, construído a partir da Tabela 1. Através desse gráfico, fica mais fácil perceber que os ursos machos são a grande maioria dos ursos estudados. Como esse gráfico contém todas as informações da Tabela 1, pode substituí-la com a vantagem de tornar análise dessa variável mais agradável.

As vantagens da representação gráfica das distribuições de freqüências ficam ainda mais evidentes quando há a necessidade de comparar vários grupos com relação à variáveis que possuem muitas categorias, como veremos mais adiante.

Uma alternativa ao gráfico de setores é o gráfico de barras (colunas) como o da Figura 4. Ao invés de dividirmos um círculo, dividimos uma barra. Note que, em ambos os gráficos, as freqüências relativas das categorias devem somar 100%. Aliás, essa é a idéia dos gráficos: mostrar como se dá a divisão (distribuição) do total de elementos (100%) em partes (fatias).

Uma situação diferente ocorre quando desejamos comparar a distribuição de freqüências de uma mesma variável em vários grupos, como por exemplo, a freqüência de ursos marrons em quatro regiões de um país.

Se quisermos usar o gráfico de setores para fazer essa comparação, devemos fazer quatro gráficos, um para cada região, com duas fatias cada um (ursos marrons e ursos não marrons). Uma alternativa é a construção de um gráfico de barras (horizontal ou vertical) como na Figura 5, com uma barra para cada região representando a freqüência de ursos marrons naquela região. Além de economizar espaço na apresentação, permite que as comparações sejam feitas de maneira mais rápida (tente fazer essa comparação usando quatro pizzas e comprove!!)

A ordem dos grupos pode ser qualquer, ou aquela mais adequada para a presente análise. Frequentemente, encontramos as barras em ordem decrescente, já antecipando nossa intuição de ordenar os grupos de acordo com sua freqüência para facilitar as comparações. Caso a variável fosse do tipo ordinal, a ordem das barras seria a ordem natural das categorias, como na tabela de freqüências.

A Figura 6 mostra um gráfico de barras que pode ser usado da comparação da distribuição

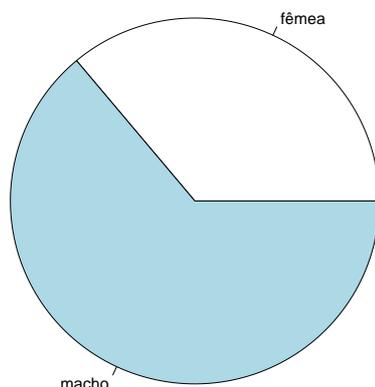


Figura 3: Gráfico de setores para a variável sexo.

de freqüências de uma mesma variável em vários grupos. É também uma alternativa ao uso de vários gráficos de setores, sendo, na verdade, a junção de três gráficos com os da Figura 4 num só gráfico.

Observação: Este tipo de gráfico só deve ser usado quando não houver muitos grupos a serem comparados e a variável em estudo não tiver muitas categorias (de preferência, só duas). No exemplo da Figura 6, a variável raça tem três categorias, mas uma delas é muito menos freqüente do que as outras duas.

Através desse gráfico, podemos observar que a população brasileira total, em 1999, dividia-se quase que igualmente entre brancos e negros, com uma pequena predominância de brancos. Porém, quando nos restringimos às classes menos favorecidas economicamente, essa situação se inverte, com uma considerável predominância de negros, principalmente na classe da população considerada indigente, indicando que a classe sócio-econômica influencia a distribuição de negros e brancos na população brasileira de 1999.

Freqüentemente, é necessário fazer comparações da distribuição de freqüências de uma variável em vários grupos simultaneamente. Nesse caso, o uso de gráficos bem escolhidos e construídos torna a tarefa muito mais fácil. Na Figura 7, está representada a distribuição de freqüências da reprovação segundo as variáveis sexo do aluno, período e área de estudo.

Analisando os três gráficos da Figura 7, podemos notar que o percentual de reprovação entre os alunos do sexo masculino é sempre maior do que o percentual de reprovação entre os alunos do sexo feminino, em todas as áreas, durante todos os períodos.

A área de ciências exatas é a que possui os maiores percentuais de reprovação, em todos os períodos, nos dois sexos.

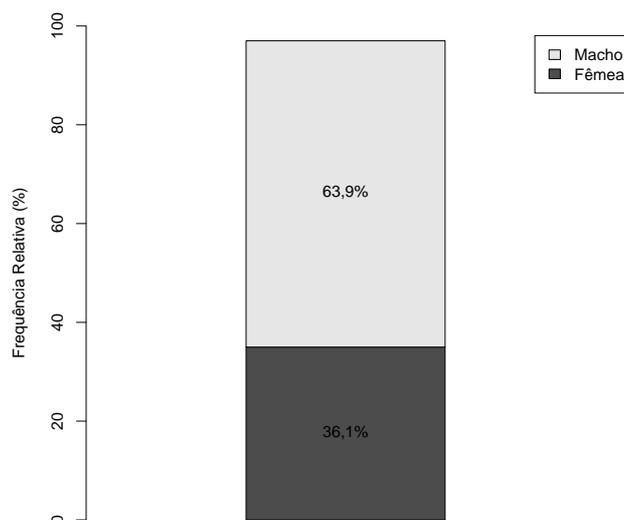


Figura 4: Gráfico de barras para a variável sexo.

Na área de ciências humanas, o percentual de reprovação entre os alunos do sexo masculino cresce com os períodos, enquanto esse percentual entre as alunas se mantém praticamente constante durante os períodos.

Na área de ciências biológicas, há uma diminuição do percentual de reprovação, a partir do segundo período, entre os alunos dos dois sexos, sendo mais acentuado entre os estudantes do sexo masculino.

Chegar às conclusões colocadas acima através de comparação numérica de tabelas de frequências seria muito mais árduo do que através da comparação visual possibilitada pelo uso dos gráficos. Os gráficos são ferramentas poderosas e devem ser usadas sempre que possível.

É importante observar que a comparação dos três gráficos da Figura 7 só foi possível porque eles usam a mesma escala, tanto no eixo dos períodos (mesma ordem) quanto no eixo dos percentuais de reprovação (mais importante). Essa observação é válida para toda comparação entre gráficos de quaisquer tipos.

3.3.2 Variáveis Quantitativas Discretas

Quando estamos trabalhando com uma variável discreta que assume poucos valores, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.

A Tabela 3 apresenta a distribuição de frequências do número de filhos por família em uma localidade, que, nesse caso, assumiu apenas seis valores distintos.

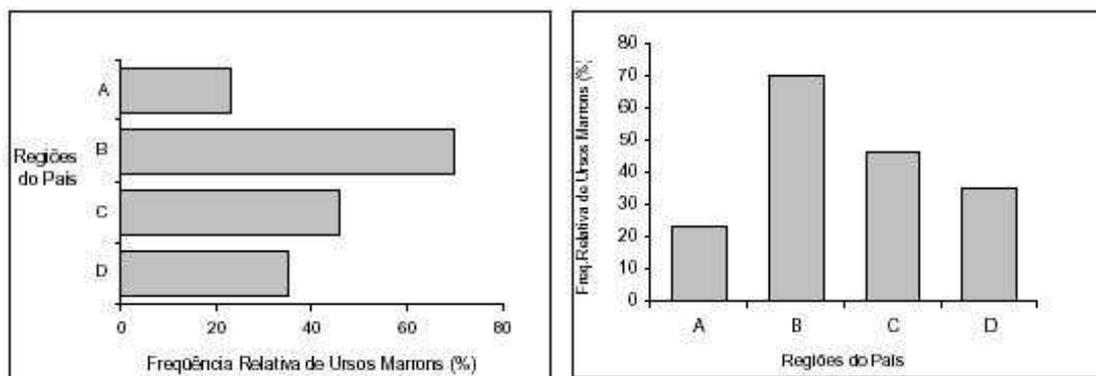


Figura 5: Gráfico de barras horizontais e verticais para a frequência de ursos marrons em quatro regiões.

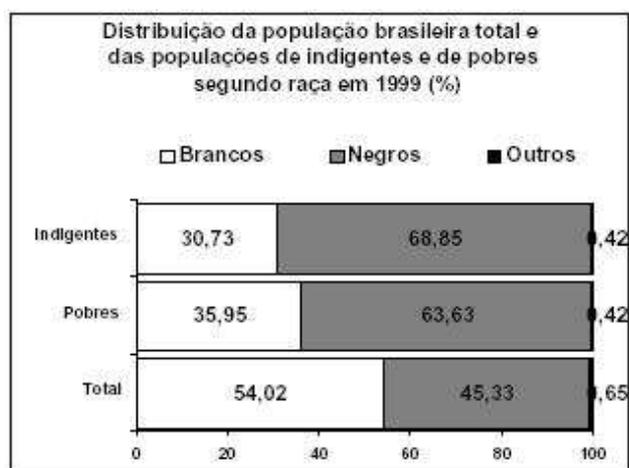


Figura 6: Gráfico de barras para comparação da distribuição de frequências de uma variável (raça) em vários grupos (indigentes, pobres e população total).

Analisando a Tabela 3, podemos perceber que as famílias mais frequentes são as de dois filhos (40%), seguida pelas famílias de três filhos. Apenas 16% das famílias têm mais de três filhos, mas são ainda mais comuns do que famílias sem filhos.

A Figura 8 mostra a representação gráfica da Tabela 3 no gráfico à esquerda e a distribuição de frequências do número de filhos por família na localidade B no gráfico à direita. Como o número de famílias estudadas em cada localidade é diferente, a frequência utilizada em ambos os gráficos foi a relativa (em porcentagem), tornando os dois gráficos comparáveis. Comparando os dois gráficos, notamos que a localidade B tende a ter famílias menos numerosas do que a localidade A. A maior parte das famílias da localidade B (cerca de 70%) têm um ou nenhum filho.

Importante: Na comparação da distribuição de frequências de uma variável entre dois ou

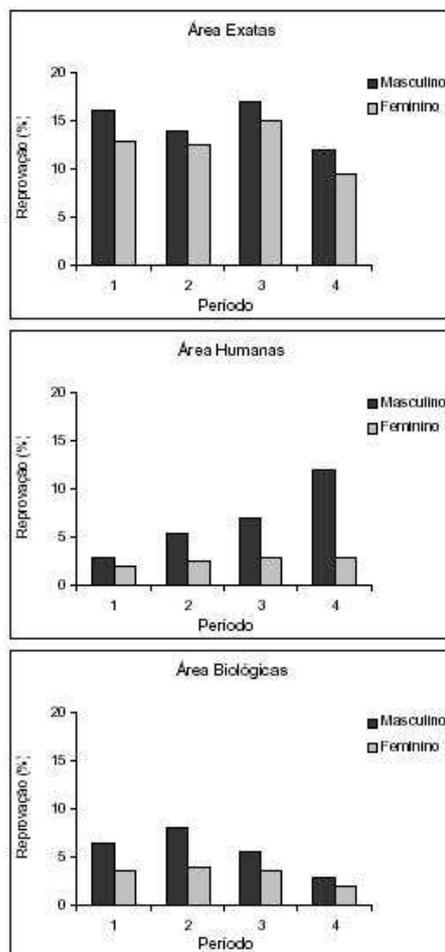


Figura 7: Distribuição de freqüências de reprovação segundo área, período e sexo do aluno. Fonte: A Evasão no Ciclo Básico da UFMG, em Cadernos de Avaliação 3, 2000.

mais grupos de tamanhos (número de observações) diferentes, devemos usar as freqüências relativas na construção do histograma. Deve-se, também usar a mesma escala em todos os histogramas, tanto na escala vertical quanto na horizontal.

Quando trabalhamos com uma variável discreta que pode assumir um grande número de valores distintos como, por exemplo, o número de ovos que um inseto põe durante sua vida, a construção da tabela de freqüências e de gráficos considerando cada valor como uma categoria fica inviável. A solução é agrupar os valores em classes ao montar a tabela, como mostra a Tabela 4.

A Figura 9 mostra o gráfico da distribuição de freqüências do número de ovos postos por 250 insetos ao longo de suas vidas. Podemos perceber que o número de ovos está concentrado em torno de 20 a 24 ovos com um ligeiro deslocamento para os valores maiores.

A escolha do número de classes e do tamanho das classes depende da amplitude dos valores

Tabela 3: Distribuição de freqüências do número de filhos por família em uma localidade (25 lares).

| Número de filhos | Freqüência Absoluta | Freqüência Relativa (%) | Freqüência Relativa Acumulada (%) |
|------------------|---------------------|-------------------------|-----------------------------------|
| 0 | 1 | 4,0 | 4,0 |
| 1 | 4 | 16,0 | 20,0 |
| 2 | 10 | 40,0 | 60,0 |
| 3 | 6 | 24,0 | 84,0 |
| 4 | 2 | 8,0 | 92,0 |
| 5 | 2 | 8,0 | 100,0 |
| Total | 25 | 100 | — |

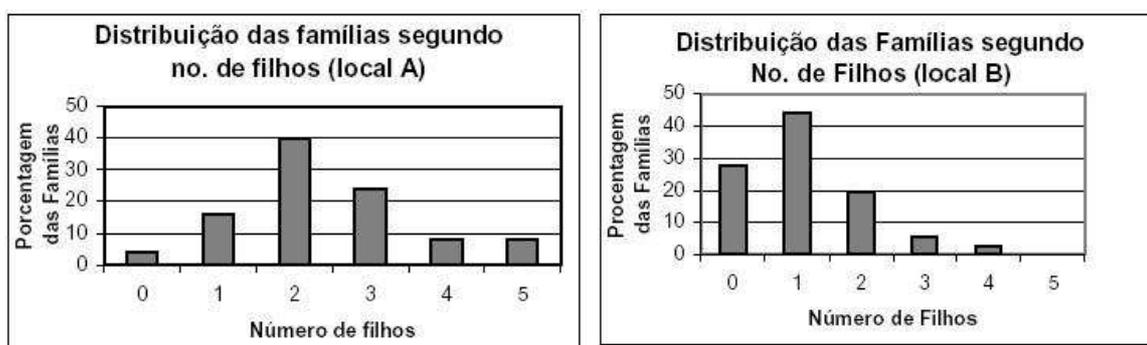


Figura 8: Distribuição de freqüências do número de filhos por família na localidade A (25 lares) e B (36 lares).

a serem representados (no exemplo, de 10 a 44) e da quantidade de observações no conjunto de dados.

Classes muito grandes resumem demais a informação contida nos dados, pois forçam a construção de poucas classes. No exemplo dos insetos, seria como, por exemplo, construir classes de tamanho 10, o que reduziria para quatro o número de classes (Figura 10).

Por outro lado, classes muito pequenas nos levaria a construir muitas classes, o que poderia não resumir a informação como gostaríamos. Além disso, para conjuntos de dados pequenos, pode ocorrer classes com muito poucas observações ou mesmo sem observações. Na Figura 11, há classes sem observações, mesmo o conjunto de dados sendo grande.

Alguns autores recomendam que tabelas de freqüências (e gráficos) possuam de 5 a 15 classes, dependendo do tamanho do conjunto de dados e levando-se em consideração o que foi exposto anteriormente.

Os limites inferiores e superiores de cada classe dependem do tamanho (amplitude) de classe escolhido, que deve ser, na medida do possível, igual para todas as classes. Isso facilita a interpretação da distribuição de freqüências da variável em estudo.

Com o uso do computador na análise estatística de dados, a tarefa de construção de tabelas

Tabela 4: Distribuição de freqüências do número de ovos postos por 250 insetos.

| Número de ovos | Freqüências Simples | | Freqüências Acumuladas | |
|----------------|---------------------|-------------------------|------------------------|------------------------|
| | Freqüência Absoluta | Freqüência Relativa (%) | Freq.Abs. Acumulada | Freq.Rel. Acumulada(%) |
| 10 a 14 | 4 | 1,6 | 4 | 1,6 |
| 15 a 19 | 30 | 12,0 | 34 | 13,6 |
| 20 a 24 | 97 | 38,8 | 131 | 52,4 |
| 25 a 29 | 77 | 30,8 | 208 | 83,2 |
| 30 a 34 | 33 | 13,2 | 241 | 96,4 |
| 35 a 39 | 7 | 2,8 | 248 | 99,2 |
| 40 a 44 | 2 | 0,8 | 250 | 100,0 |
| Total | 250 | 100 | — | — |

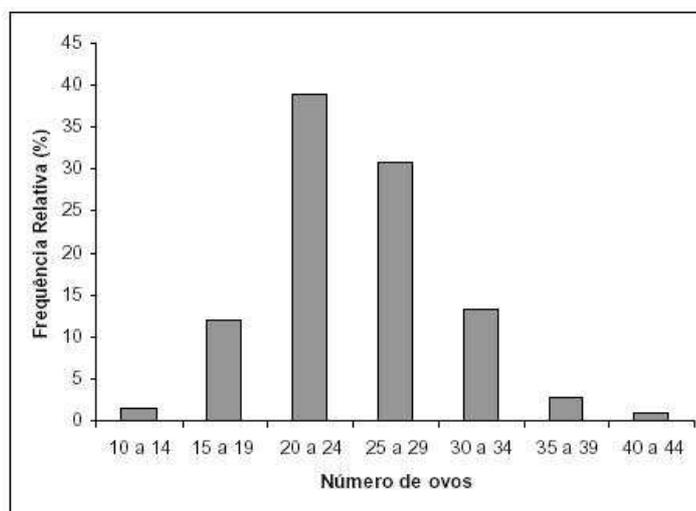


Figura 9: Distribuição de freqüências do número de ovos postos por 250 insetos.

e gráficos ficou menos trabalhosa e menos dependente de regras rígidas. Se determinado agrupamento de classes não nos pareceu muito bom, podemos construir vários outros quase que instantaneamente e a escolha da melhor representação para a distribuição de freqüências para aquela variável fica muito mais tranqüila.

3.3.3 Variáveis Quantitativas Contínuas

Quando a variável em estudo é do tipo contínua, que assume muitos valores distintos, o agrupamento dos dados em classes será sempre necessário na construção das tabelas de freqüências. A Tabela 5 apresenta a distribuição de freqüências para o peso dos ursos machos.

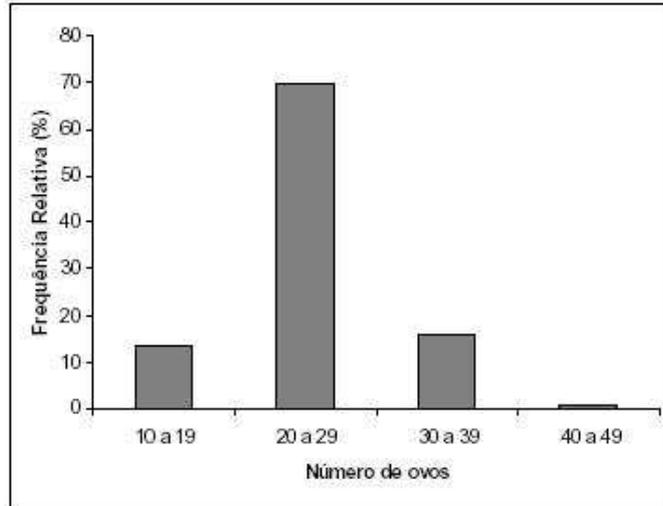


Figura 10: Distribuição de frequências do número de ovos postos por 250 insetos.(classes de tamanho 10)

Os limites das classes são representados de modo diferente daquele usado nas tabelas para variáveis discretas: o limite superior de uma classe é igual ao limite inferior da classe seguinte. Mas, afinal, onde ele está incluído?

O símbolo $|-$ resolve essa questão. Na segunda classe $(25| - 50)$, por exemplo, estão incluídos todos os ursos com peso de 25,0 a 49,9 kg. Os ursos que porventura pesarem exatos 50,0 kg serão incluídos na classe seguinte. Ou seja, ursos com pesos maiores ou iguais a 25 kg e menores do que 50 kg.

A construção das classes da tabela de frequências é feita de modo a facilitar a interpretação da distribuição de frequências, como discutido anteriormente. Geralmente, usamos tamanhos e limites de classe múltiplos de 5 ou 10. Isso ocorre porque estamos acostumados a pensar no nosso sistema numérico, que é o decimal. Porém, nada nos impede de construirmos classes de outros tamanhos (inteiros ou fracionários) desde que isso facilite nossa visualização e interpretação da distribuição de frequências da variável em estudo.

A representação gráfica da distribuição de frequências de uma variável contínua é feita através de um gráfico chamado histograma, mostrado na Figura 12. O histograma nada mais é do que o gráfico de barras verticais, porém construído com as barras unidas, devido ao caráter contínuo dos valores da variável.

Os histogramas da Figura 12 têm a mesma forma, apesar de serem construídos usando as frequências absolutas e relativas, respectivamente. O objetivo dessas figuras é mostrar que a escolha do tipo de frequência a ser usada não muda a forma da distribuição. Entretanto, o uso da frequência relativa torna o histograma comparável a outros histogramas, mesmo que os conjuntos de dados tenham tamanhos diferentes (desde a mesma escala seja usada!)

Analisando o histograma para o peso dos ursos machos, podemos perceber que há dois grupos de ursos: os mais leves, com pesos em torno de 50 a 75 Kg, e os mais pesados, com pesos em torno de 150 a 175 Kg. Essa divisão pode ser devida a uma outra característica dos ursos, como idades ou hábitos alimentares diferentes, por exemplo.

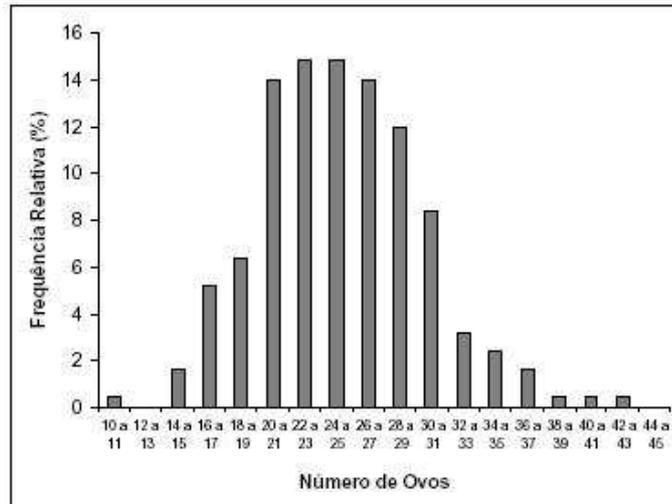


Figura 11: Distribuição de frequências do número de ovos postos por 250 insetos.(classes de tamanho 2)

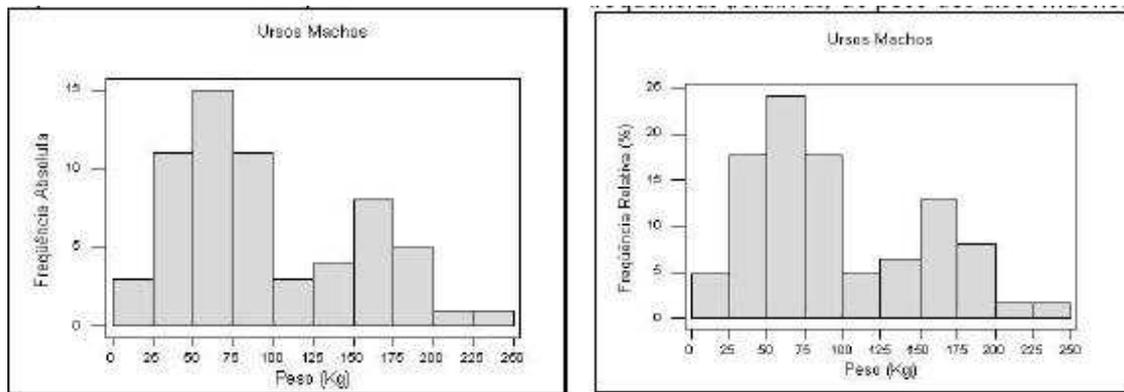


Figura 12: Histograma para a distribuição de frequências (absolutas e relativas) de pesos de ursos machos

A Tabela 6 apresenta a distribuição de frequências para o peso dos ursos fêmeas, representada graficamente pelo histograma à esquerda na Figura 13. Apesar de não haver, neste conjunto de dados, fêmeas com peso maior de que 175 Kg, as três últimas classes foram mantidas para que pudessemos comparar machos e fêmeas quanto ao peso.

A Figura 13 também mostra o histograma para o peso dos ursos machos (à direita). Note que ele tem a mesma forma dos histogramas da Figura 12, porém com as barras mais achatadas, devido à mudança de escala no eixo vertical para torná-lo comparável ao histograma das fêmeas.

Comparando as distribuições dos pesos dos ursos machos e fêmeas, podemos concluir que as fêmeas são, em geral, menos pesadas do que os machos, distribuindo-se quase

Tabela 5: Distribuição de frequências dos ursos machos segundo peso.

| Peso (kg) | Frequência | Frequência | Freq. Abs. | Freq. Rel. |
|------------|------------|--------------|------------|---------------|
| | Absoluta | Relativa (%) | Acumulada | Acumulada (%) |
| 0 – 25 | 3 | 4,8 | 3 | 4,8 |
| 25 – 50 | 11 | 17,7 | 14 | 22,6 |
| 50 – 75 | 15 | 24,2 | 29 | 46,8 |
| 75 – 100 | 11 | 17,7 | 40 | 64,5 |
| 100 – 125 | 3 | 4,8 | 43 | 69,4 |
| 125 – 150 | 4 | 6,5 | 47 | 75,8 |
| 150 – 175 | 8 | 12,9 | 55 | 88,7 |
| 175 – 200 | 5 | 8,1 | 60 | 96,8 |
| 200 – 225 | 1 | 1,6 | 61 | 98,4 |
| 225 – 250 | 1 | 1,6 | 62 | 100,0 |
| Total | 62 | 100,0 | - | - |

Tabela 6: Distribuição de frequências dos ursos fêmeas segundo peso.

| Peso (kg) | Frequência | Frequência | Freq. Abs. | Freq. Rel. |
|------------|------------|--------------|------------|---------------|
| | Absoluta | Relativa (%) | Acumulada | Acumulada (%) |
| 0 – 25 | 3 | 8,6 | 3 | 8,6 |
| 25 – 50 | 5 | 14,3 | 8 | 22,9 |
| 50 – 75 | 18 | 51,4 | 26 | 74,3 |
| 75 – 100 | 5 | 14,3 | 31 | 88,6 |
| 100 – 125 | 2 | 5,7 | 33 | 94,3 |
| 125 – 150 | 1 | 2,9 | 34 | 97,1 |
| 150 – 175 | 1 | 2,9 | 35 | 100,0 |
| 175 – 200 | 0 | 0 | 35 | 100,0 |
| 200 – 225 | 0 | 0 | 35 | 100,0 |
| 225 – 250 | 0 | 0 | 35 | 100,0 |
| Total | 35 | 100,0 | - | - |

simetricamente em torno da classe de 50 a 75 Kg . O peso das fêmeas é mais homogêneo (valores mais próximos entre si) do que o peso dos ursos machos.

Muitas vezes, a análise da distribuição de frequências acumuladas é mais interessante do que a de frequências simples, representada pelo histograma. O gráfico usado na representação gráfica da distribuição de frequências acumuladas de uma variável contínua é a ogiva, apresentada na Figura 14. Para a construção da ogiva, são usadas as frequências acumuladas (absolutas ou relativas) no eixo vertical e os limites superiores de classe no eixo horizontal.

O primeiro ponto da ogiva é formado pelo limite inferior da primeira classe e o valor zero, indicando que abaixo do limite inferior da primeira classe não existem observações. Daí por diante, são usados os limites superiores das classes e suas respectivas frequências acumuladas, até a última classe, que acumula todas as observações. Assim, uma ogiva deve começar no valor zero e, se for construída com as frequências relativas acumuladas,

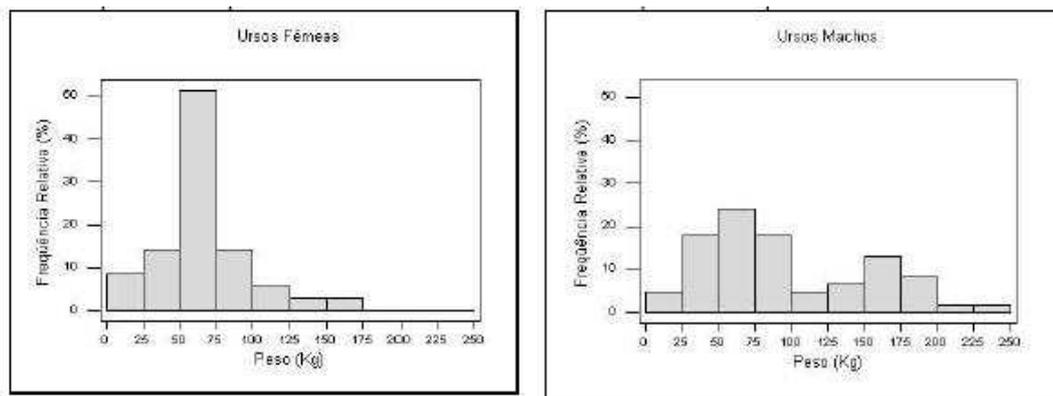


Figura 13: Histograma para a distribuição de frequências de pesos de ursos fêmeas (esquerda) e machos (direita)

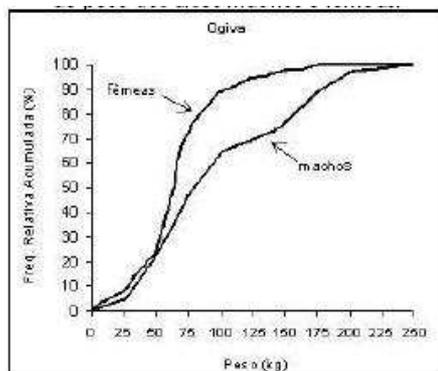


Figura 14: Ogivas para a distribuição de frequências de pesos de ursos machos e fêmeas

terminar com o valor 100

A ogiva permite que sejam respondidas perguntas do tipo:

a) Qual o percentual de ursos têm peso de até 125 Kg?

Na Figura 15(a), traçamos uma linha vertical partindo do ponto 120 kg até cruzar com cada ogiva (fêmeas e machos). A partir deste ponto de cruzamento, traçamos uma linha horizontal até o eixo das frequências acumuladas, encontrando o valor de 70% para os machos e 95% para as fêmeas.

Assim, 95% das fêmeas têm até 125 kg, enquanto 70% dos machos têm até 125 kg. É o mesmo que dizer que apenas 5% das fêmeas pesam mais que 125 kg, enquanto 30% dos machos pesam mais que 125 kg.

b) Qual o valor do peso que deixa abaixo (e acima) dele 50% dos ursos?

Na Figura 15(b), traçamos uma linha horizontal partindo da frequência acumulada de 50% até encontrar as duas ogivas. A partir destes pontos de encontro, traçamos uma linha vertical até o eixo do valores de peso, encontrando o valor de 80 kg para os machos

e 65 kg para as fêmeas.

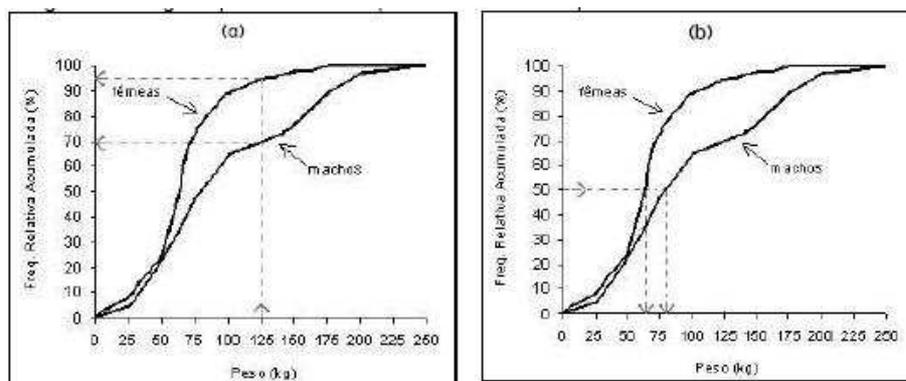


Figura 15: Ogivas para a distribuição de frequências de pesos de ursos machos e fêmeas

Assim, metade dos machos pesam até 80 kg (e metade pesam mais que 80 kg), enquanto metade das fêmeas pesam até 65 kg.

3.3.4 Outros Gráficos para Variáveis Quantitativas

Quando construímos uma tabela de frequências para uma variável quantitativa utilizando agrupamento de valores em classes, estamos resumindo a informação contida nos dados. Isto é desejável quando o número de dados é grande e sem um algum tipo de resumo ficaria difícil tirar conclusões sobre o comportamento da variável em estudo.

Porém, quando a quantidade de dados disponíveis não é tão grande, o resumo promovido pelo histograma não é aconselhável.

Para os casos em que o número de dados é pequeno, uma alternativa para a visualização da distribuição desses dados são os gráficos denominados diagrama de pontos e diagrama de ramo-e-folhas.

O Diagrama de Pontos

Uma representação alternativa ao histograma para a distribuição de frequências de uma variável quantitativa é o diagrama de pontos, como aqueles mostrados na Figura 16.

Neste gráfico, cada ponto representa uma observação com determinado valor da variável. Observações com mesmo valor são representadas com pontos empilhados neste valor.

Através da comparação dos diagramas de pontos da Figura 16, podemos ver que os ursos machos possuem pesos menos homogêneos (mais dispersos) do que as fêmeas, que estão concentradas na parte esquerda do eixo de valores de peso.

O Diagrama de Ramo-e-Folhas

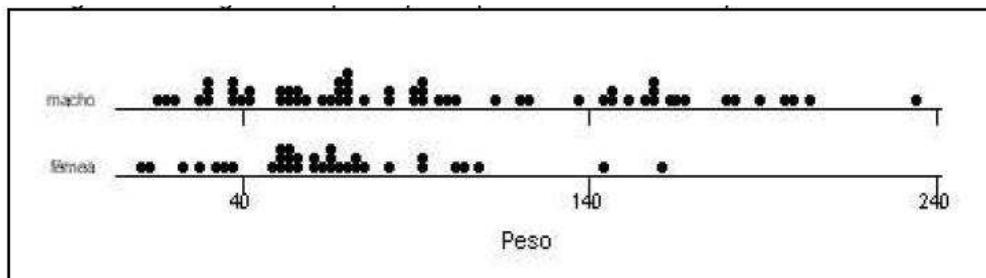


Figura 16: Diagrama de pontos para o peso de ursos machos e peso dos ursos fêmeas.

Outro gráfico útil e simples para representar a distribuição de freqüências de uma variável quantitativa com poucas observações é o diagrama de ramo-e-folhas. A sua sobre os demais é que ele explicita os valores dos dados, como veremos.

Exemplo dos ursos marrons (continuação):

Dos 35 ursos fêmeas observados, somente 20 puderam ter sua idade estimada. Para visualizar a distribuição dos valores de idade dessas fêmeas, usaremos um diagrama de ramo-efolhas, já que um histograma resumiria mais ainda algo que já está resumido.

Os 20 valores de idade (em meses) disponíveis, já ordenados são:

8 9 11 17 17 19 20 44 45 53 57 57 57 58 70 81 82 83 100 104

Podemos organizar os dados, separando-os pela dezenas, uma em cada linha:

```

8   9
11  17 17 19
20
44  45
53  57 57 57 58
70
81  82 83
100 104

```

Como muitos valores em cada linha tem as dezenas em comum, podemos colocar as dezenas em evidência, separando-as das unidades por um traço. Ao dispor os dados dessa maneira, estamos construindo um diagrama de ramo-e-folhas (Figura 17). O lado com as dezenas é chamado de ramo, no qual estão dependuradas as unidades, chamadas folhas.

Os ramos e as folhas podem representar quaisquer unidades de grandeza (dezenas e unidades, centenas e dezenas, milhares e centenas, etc). Para sabermos o que está sendo representado, um ramo-e-folhas deve ter sempre uma legenda, indicando o que significam os ramos e as folhas.

Se a idade estivesse medida em dias, por exemplo, usando esse mesmo ramo-efolhas, poderíamos estabelecer que o ramo representaria as centenas e as folhas, as dezenas. Assim, 0—8 seria igual a 80 dias e 10—4 seria igual a 1040 dias.

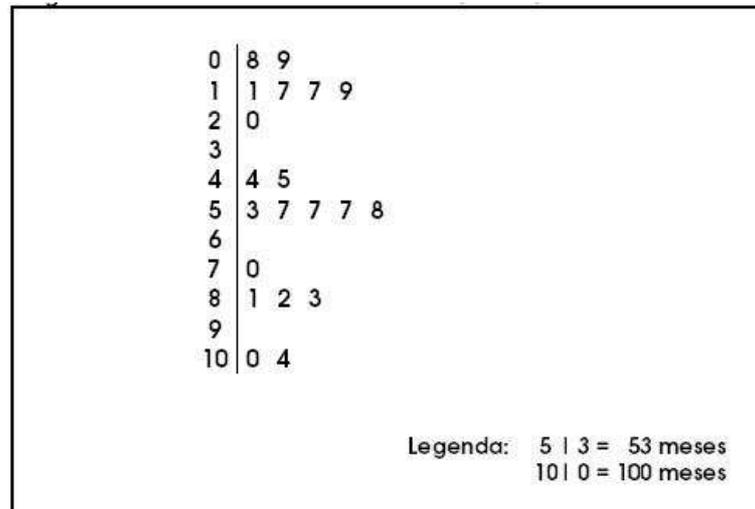


Figura 17: Ramo-e-folhas da idade (meses) dos ursos fêmeas.

Analisando o ramo-e-folhas para a idade dos ursos fêmeas, percebemos a existência de três grupos: fêmeas mais jovens (até 20 meses), fêmeas mais crescidas (de 44 a 58 meses) e um grupo mais velho (mais de 70 meses), com destaque para duas fêmeas bem mais velhas.

O ramo-e-folhas também pode ser usado para comparar duas distribuições de valores, como mostra a Figura 18. Aproveitando o mesmo ramo do diagrama das fêmeas, podemos fazer o diagrama dos machos, utilizando o lado esquerdo. Observe que as folhas dos ursos machos são dependuradas de modo espelhado, assim como explica a legenda, que agora deve ser dupla.

Observando a Figura 18, notamos que os ursos machos são, em geral, mais jovens do que os ursos fêmeas, embora possuam dois ursos bem idosos em comparação com os demais.

Importante: No ramo-e-folhas, estamos trabalhando, implicitamente, com freqüências absolutas. Assim, ao comparar dois grupos de tamanhos diferentes, devemos levar isso em conta. Caso os tamanhos dos grupos sejam muito diferentes, não se deve adotar o ramo-e-folhas como gráfico para comparação de distribuições.

3.3.5 Aspectos Gerais da Distribuição de Freqüências

Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

- Tendência Central;
- Variabilidade;
- Forma.

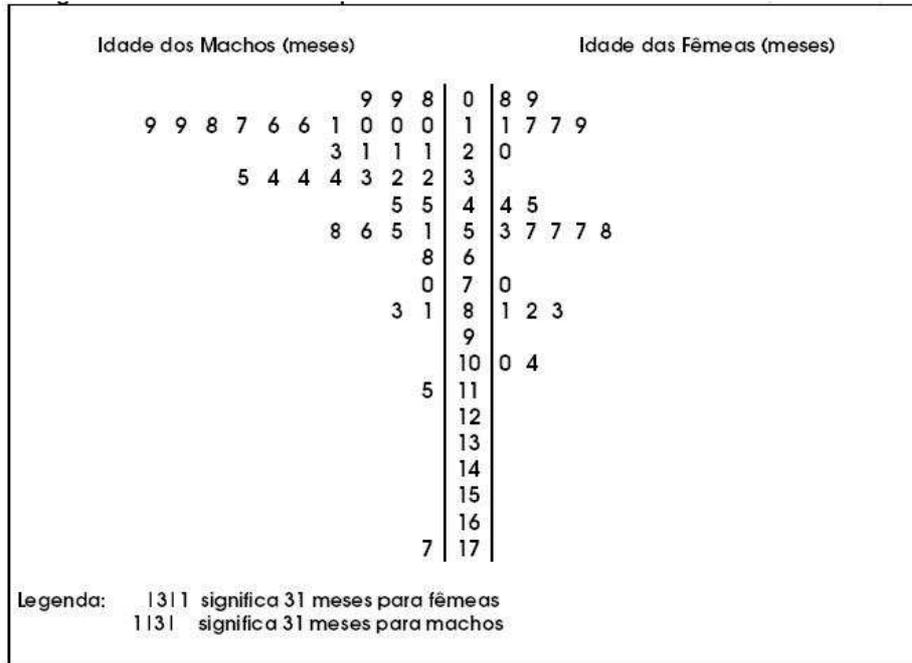


Figura 18: Ramo-e-folhas da idade (meses) dos ursos fêmeas.

O histograma (ou o diagrama de pontos, ou o ramo-e-folhas) permite a visualização destas características da distribuição de freqüências, como veremos a seguir. Além disso, elas podem ser quantificadas através das medidas de síntese numérica (não discutidas aqui).

Tendência Central

A tendência central da distribuição de freqüências de uma variável é caracterizada pelo valor (ou faixa de valores) típico da variável.

Uma das maneiras de representar o que é típico é através do valor mais freqüente da variável, chamado de **moda**. Ou, no caso da tabela de freqüências, a classe de maior freqüência, chamada de **classe modal**. No histograma, esta classe corresponde àquela com barra mais alta ("pico").

No exemplo dos ursos marrons (Figura 13), a classe modal do peso dos ursos fêmeas é claramente a terceira, de 50 a 75 kg. Assim, os ursos fêmeas pesam, tipicamente, de 50 a 75 kg. Entretanto, para os ursos machos, temos dois picos: de 50 a 75 kg e de 150 a 175 kg. Ou seja, temos um grupo de machos com peso típico como o das fêmeas e outro grupo, menor, formado por ursos tipicamente maiores.

Dizemos que a distribuição de freqüências do peso dos ursos fêmeas é **unimodal** (apenas uma moda) e dos ursos machos é **bimodal** (duas modas). Geralmente, um histograma bimodal indica a existência de dois grupos, com valores centrados em dois pontos diferentes do eixo de valores. Uma distribuição de freqüências pode também ser amodal, ou seja, todos os valores são igualmente freqüentes.

Variabilidade

Para descrever adequadamente a distribuição de freqüências de uma variável quantitativa, além da informação do valor representativo da variável (tendência central), é necessário dizer também o quanto estes valores variam, ou seja, o quão **dispersos** eles são.

De fato, somente a informação sobre a tendência central de um conjunto de dados não consegue representá-lo adequadamente.

A Figura 19 mostra um diagrama de pontos para os tempos de espera de 21 clientes de dois bancos, um com fila única e outro com fila múltipla, com o mesmo número de atendentes. Os tempos de espera nos dois bancos têm a mesma tendência central de 7 minutos. Entretanto, os dois conjuntos de dados são claramente diferentes, pois os valores são muito mais dispersos no banco com fila múltipla.

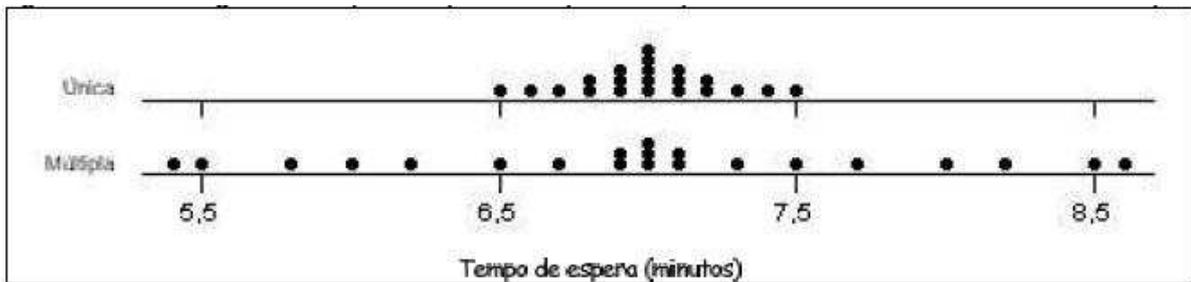


Figura 19: Ramo-e-folhas da idade (meses) dos ursos fêmeas.

Assim, quando entramos num fila única, esperamos ser atendidos em cerca de 7 minutos, com uma **variação** de, no máximo, meio minuto a mais ou a menos. Na fila múltipla, a variação é maior, indicando-se que tanto pode-se esperar muito mais ou muito menos que o valor típico de 7 minutos.

Forma

A distribuição de freqüências de uma variável pode ter várias formas, mas existem três formas básicas, apresentadas na Figura 20 através de histogramas e suas respectivas ogivas.

Quando uma distribuição é **simétrica** em torno de um valor (o mais freqüente), significa que as observações estão igualmente distribuídas em torno desse valor (metade acima e metade abaixo).

A **assimetria** de uma distribuição pode ocorrer de duas formas:

- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas preferem se referir ao lado onde está faltando dados (cauda). As duas denominações são alternativas.

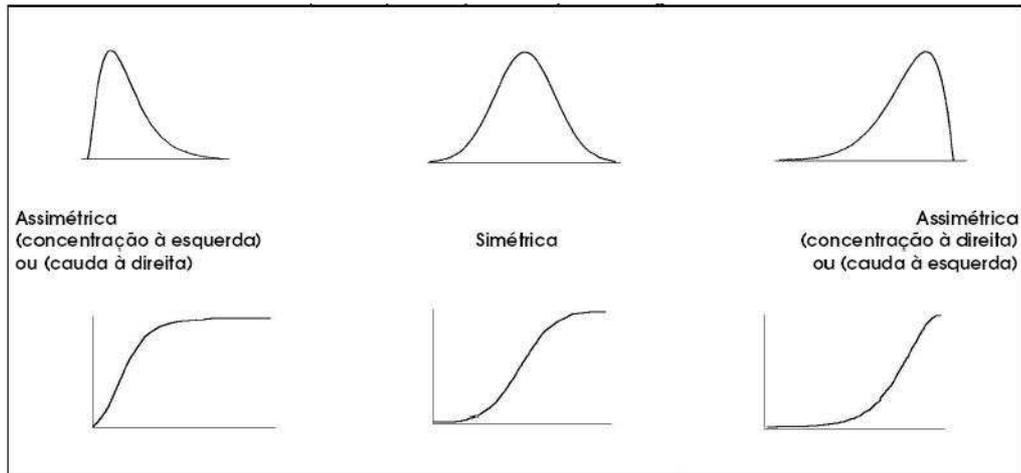


Figura 20: Ramo-e-folhas da idade (meses) dos ursos fêmeas.

Em alguns casos, apenas o conhecimento da forma da distribuição de frequências de uma variável já nos fornece uma boa informação sobre o comportamento dessa variável.

Por exemplo, o que você acharia se soubesse que a distribuição de frequências das notas da primeira prova da disciplina de Estatística que você está cursando é, geralmente, assimétrica com concentração à direita? Como você acha que é a forma da distribuição de frequências da renda no Brasil?

Note que, quando a distribuição é assimétrica com concentração à esquerda, a ogiva cresce bem rápido, por causa do acúmulo de valores do lado esquerdo do eixo. Por outro lado, quando a distribuição é assimétrica com concentração à direita, a ogiva cresce lentamente no começo e bem rápido na parte direita do eixo, por causa do acúmulo de valores desse lado. Quando a distribuição é simétrica, a ogiva tem a forma de um S suave e simétrico.

A ogiva para uma distribuição de frequências bimodal (Figura 21) mostra essa característica da distribuição através de um platô ("barriga") no meio da ogiva. A ogiva para o peso dos ursos machos (Figura 15) também mostra essa barriga .

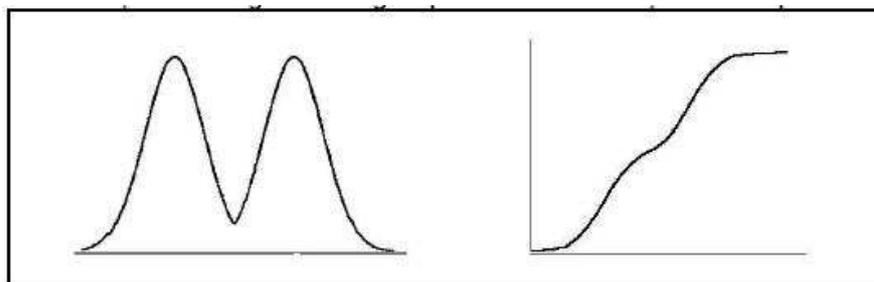


Figura 21: Ramo-e-folhas da idade (meses) dos ursos fêmeas.

Séries Temporais

Séries temporais (ou séries históricas) são um conjunto de observações de uma mesma variável quantitativa (discreta ou contínua) feitas ao longo do tempo.

O conjunto de todas as temperaturas medidas diariamente numa região é um exemplo de série temporal.

Um dos objetivos do estudo de séries temporais é conhecer o comportamento da série ao longo do tempo (aumento, estabilidade ou declínio dos valores). Em alguns estudos, esse conhecimento pode ser usado para se fazer previsões de valores futuros com base no comportamento dos valores passados.

A representação gráfica de uma série temporal é feita através do **gráfico de linha**, como exemplificado na Figura 22.

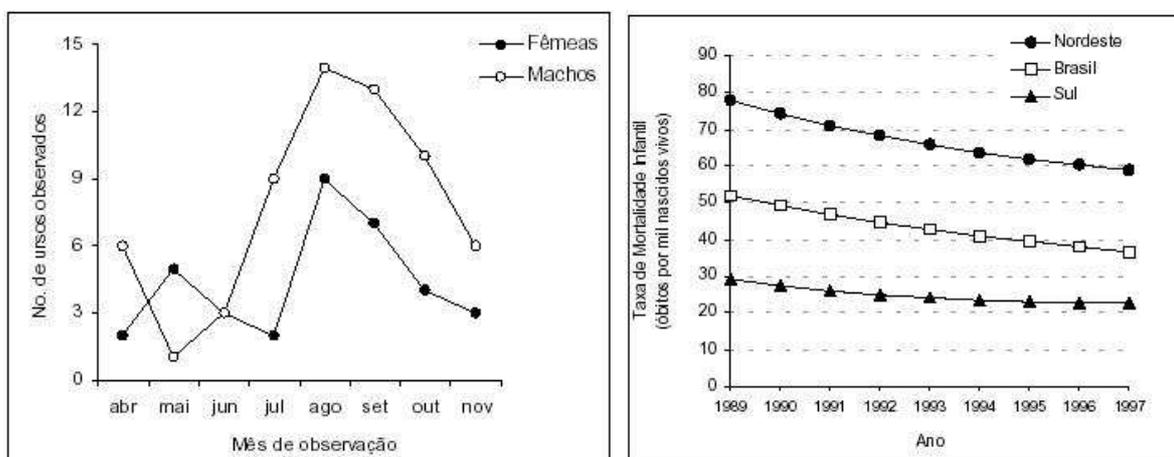


Figura 22: Gráfico de linha para o número de ursos machos e fêmeas observados ao longo dos meses de pesquisa (à esquerda) e taxa de mortalidade infantil de 1989 a 1997 nas Regiões Nordeste e Sul e no Brasil (à direita).

No eixo horizontal do gráfico de linha, está o indicador de tempo e, no eixo vertical, a variável a ser representada. As linhas horizontais pontilhadas são opcionais e só devem ser colocadas quando ajudarem na interpretação do gráfico. Caso contrário, devem ser descartadas, pois, como já enfatizamos antes, um gráfico deve ser o mais limpo possível.

No gráfico da Figura 22, podemos notar que a taxa de mortalidade infantil na região Nordeste esteve sempre acima da taxa da região Sudeste durante todo o período considerado, com um declínio das taxas nas duas regiões e também no Brasil como um todo ao longo do período.

Embora o declínio absoluto na taxa da região Nordeste tenha sido maior (aproximadamente 20 casos em mil nascidos vivos), a redução percentual na taxa da região Sudeste foi maior (cerca de 8 casos a menos nos 30 iniciais, ou seja, 27% a menos, enquanto 20 casos a menos nos 80 iniciais na região Nordeste representam uma redução de 25%).

Podemos observar ainda uma tendência à estabilização da taxa de mortalidade infantil da região Sudeste a partir do ano de 1994, enquanto a tendência de declínio permanece na região Nordeste e no Brasil.

Ao analisar e construir um gráfico de linhas, devemos estar atentos a certos detalhes que podem mascarar o verdadeiro comportamento dos dados.

A Figura 23(a) apresenta um gráfico de linhas para o preço médio do litro de leite entre os meses de maio e agosto de 2001. Apesar de colocar os valores para cada mês, o gráfico não mostra a escala de valores e não representa a série desde o começo da escala, o valor zero.

Essa concentração da visualização da linha somente na parte do gráfico onde os dados estão situados distorce a verdadeira de dimensão da queda do preço, acentuando-a. Ao compararmos com o gráfico da Figura 23(b), cujo escala vertical começa no zero, percebemos que houve mesmo uma queda, mas não tão acentuada quanto aquela mostrada no gráfico divulgado no jornal.

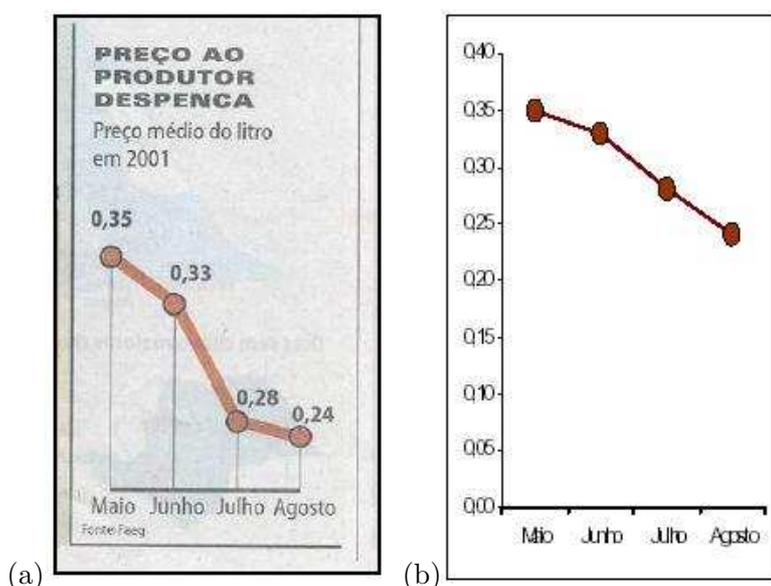


Figura 23: Gráfico de linhas para o preço médio do litro de leite: (a) original (jornal Folha de São Paulo, set/2001), (b) modificado, com a escala de valores mostrada e iniciando-se no zero.

Outro aspecto mascarado pela falta da escala é que as diferenças entre os valores numéricos não correspondem às distâncias representadas no gráfico.

Por exemplo, no gráfico de linha divulgado para a série do preço do leite, vemos que a queda no preço de maio para junho foi de R\$0,02 e, de julho para agosto, foi de R\$0,04, duas vezes maior. No entanto, a distância (vertical) entre os pontos de maio e julho é maior do que a distância (vertical) entre os pontos de julho e agosto!!

E mais, a queda de junho para junho foi de R\$0,05, pouco mais do que a queda de R\$0,04 de junho a agosto. Porém, a distância (vertical) no gráfico entre os pontos de junho e julho é cerca de quatro vezes maior do que a distância (vertical) dos pontos de julho e agosto!!

Examinando o gráfico apenas visualmente, sem nos atentar para os números, tenderemos a pensar que as grandes quedas no preço do leite ocorreram no começo do período de

observação (de maio a julho), enquanto, na verdade, as quedas se deram quase da mesma forma mês a mês, sendo um pouco maiores no final do período (de julho a agosto).

Além disso, a palavra despenca nos faz pensar numa queda abrupta, que é o que o gráfico divulgado parece querer mostrar. No entanto, analisando o gráfico da Figura 23(a), que corrige essas distorções, notamos que houve sim uma queda, mas não tão abrupta quanto colocada na Figura 23(b).

A Figura 23 mostra os efeitos na representação de uma série temporal quando mudamos o começo da escala de valores do eixo vertical. À medida que aproximamos o começo da escala do valor mínimo da série, a queda nos parece mais abrupta. A mesma observação vale para o caso em que o gráfico mostrar um aumento dos valores da série: quanto mais o início da escala se aproxima do valor mínimo da série, mais acentuado parecerá o aumento.

De maneira geral, um gráfico de linhas deve ser construído de modo que:

- O início do eixo vertical seja o valor mínimo possível para a variável que está sendo representada (para o caso do preço de leite, o valor zero, leite de graça), para evitar as distorções ilustradas na Figura 23;
- O final do eixo vertical seja tal que a série fica centrada em relação ao eixo vertical, como mostrado na Figura 24(a);
- Os tamanhos dos eixos sejam o mais parecidos possível, para que não ocorra a distorção mostrada nos gráficos (b) e (c) da Figura 24.

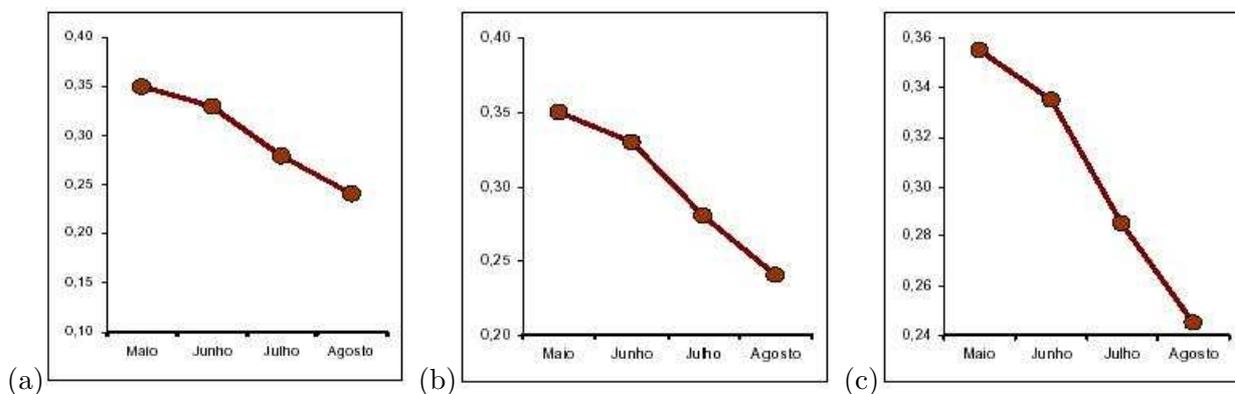


Figura 24: Efeitos da mudança no início e/ou final da escala do gráfico em linhas da série temporal do preço do leite.

3.3.6 O Diagrama de Dispersão

O diagrama de dispersão é um gráfico onde pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada elemento do conjunto de dados.

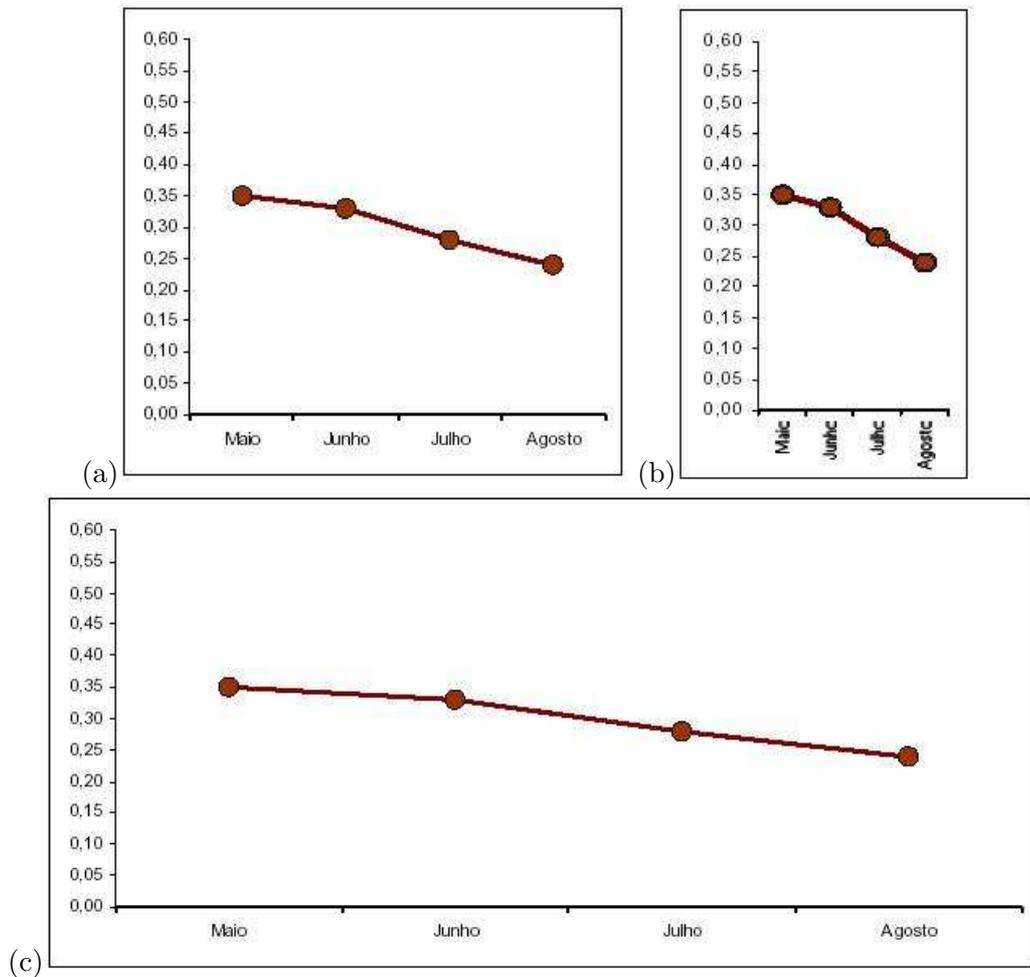


Figura 25: Efeitos de alterações na dimensão horizontal do gráfico de linhas da série do preço do leite.

A Tabela 7 e a Figura 26 mostram um esquema do desenho do diagrama de dispersão. Neste exemplo, foram medidos os valores de duas variáveis quantitativas, X e Y, em quatro indivíduos. O eixo horizontal do gráfico representa a variável X e o eixo vertical representa a variável Y.

O diagrama de dispersão é usado principalmente para visualizar a relação/associação entre duas variáveis, mas também para é muito útil para:

- Comparar o efeito de dois tratamentos no mesmo indivíduo.
- Verificar o efeito tipo antes/depois de um tratamento;

A seguir, veremos quatro exemplos da utilização do diagrama de dispersão. Os dois primeiros referem-se ao estudo da associação entre duas variáveis. O terceiro utiliza o

Tabela 7: Dados esquemáticos.

| Indivíduos | Variável X | Variável Y |
|------------|------------|------------|
| A | 2 | 3 |
| B | 4 | 3 |
| C | 4 | 5 |
| D | 8 | 7 |

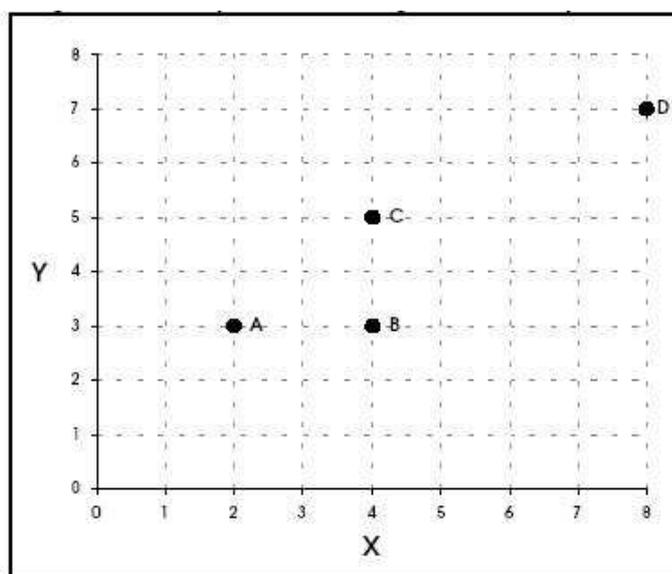


Figura 26: Esquema do diagrama de dispersão.

diagrama de dispersão para comparar o efeito de duas condições no mesmo indivíduo. O último exemplo, similar ao terceiro, verifica o efeito da aplicação de um tratamento, comparando as medidas antes e depois da medicação.

Exemplo dos ursos marrons (continuação):

Recorde que um dos objetivos dos pesquisadores neste estudo é encontrar uma maneira de conhecer o peso do urso através de uma medida mais fácil de se obter do que a direta (carregar uma balança para o meio da selva e colocar os ursos em cima dela) como, por exemplo, uma medida de comprimento (altura, perímetro do tórax, etc.).

O problema estatístico aqui é encontrar uma variável que tenha uma relação forte com o peso, de modo que, a partir de seu valor medido, possa ser calculado (estimado, na verdade) o valor peso indiretamente, através de uma equação matemática.

O primeiro passo para encontrar esta variável é fazer o diagrama de dispersão das variáveis candidatas (eixo horizontal) versus o peso (eixo vertical), usando os pares de informações de todos os ursos. Você pode tentar as variáveis: idade, altura, comprimento da cabeça, largura da cabeça, perímetro do pescoço e perímetro do tórax.

Na Figura 27, mostramos a relação entre peso e altura e entre peso e perímetro do tórax. Respectivamente.

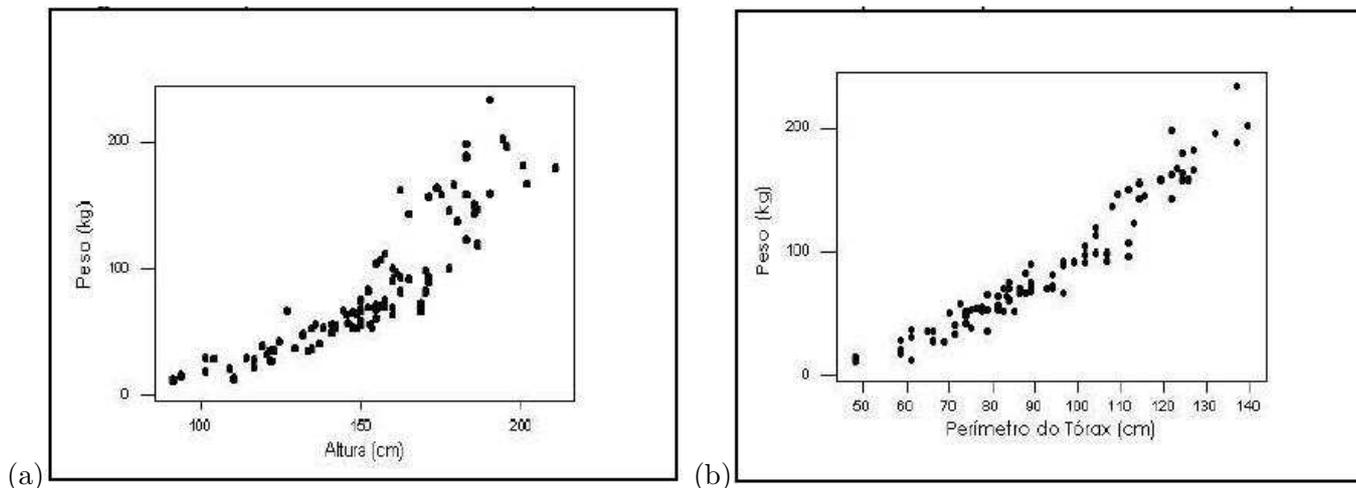


Figura 27: Diagrama de dispersão da altura versus o peso (a) e do perímetro do tórax versus o peso (b) dos ursos marrons.

Podemos ver que, tanto a altura quanto o perímetro do tórax são fortemente associados ao peso do urso, no sentido de que quanto mais alto o urso ou quanto maior a medida de seu tórax, mais pesado ele será.

Mas note que este crescimento é linear para o perímetro do tórax e não-linear para a altura.

Além disso, com os pontos estão mais dispersos no gráfico da altura, a variável mais adequada para estimar, sozinha, o peso é o perímetro do tórax (a técnica estatística adequada aqui chama-se Regressão Linear Simples).

Exemplo dos morangos:

Um produtor de morangos para exportação deseja produzir frutos grandes, pois frutos pequenos têm pouco valor mesmo no mercado interno. Além disso, os frutos, mesmo grandes, não devem ter tamanhos muito diferentes entre si. O produtor suspeita que uma dos fatores que altera o tamanho dos frutos é o número de frutos por muda.

Para investigar a relação entre o número de frutos que uma planta produz e o peso destes frutos, ele observou dados de 10 morangueiros na primeira safra (Tabela 8). O diagrama de dispersão é mostrado na Figura 28.

O diagrama de dispersão mostra-nos dois fatos. O primeiro, que há um decréscimo no valor médio do peso do fruto por árvore à medida que cresce o número de frutos na árvore. Ou seja, não é vantagem uma árvore produzir muitos frutos, pois ele tenderão a ser muito pequenos.

O segundo fato que percebemos é que, com o aumento no número de frutos na árvores, cresce também a variabilidade no peso, gerando tanto frutos muito grandes, como muito pequenos.

Tabela 8: Peso dos frutos e número de frutos por planta em 10 morangueiros na primeira safra.

| Muda | N | Peso dos Frutos (gramas) | | | | | | | | | | | | | |
|------|----|--------------------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1 | 5 | 15,2 | 15,5 | 15,6 | 15,7 | 16,4 | | | | | | | | | |
| 2 | 6 | 14,0 | 14,5 | 15,4 | 15,9 | 15,9 | 16,1 | | | | | | | | |
| 3 | 7 | 13,7 | 13,8 | 14,1 | 14,1 | 14,5 | 14,9 | 15,5 | | | | | | | |
| 4 | 8 | 11,0 | 11,5 | 12,4 | 12,4 | 12,9 | 14,5 | 15,5 | 16,6 | | | | | | |
| 5 | 9 | 10,2 | 11,1 | 12,1 | 12,4 | 13,5 | 13,8 | 14,0 | 15,4 | 16,0 | | | | | |
| 6 | 10 | 9,0 | 9,3 | 10,7 | 11,6 | 11,7 | 12,6 | 12,8 | 12,8 | 13,4 | 15,1 | | | | |
| 7 | 11 | 7,8 | 8,6 | 8,7 | 9,6 | 11,1 | 11,9 | 12,1 | 12,5 | 14,1 | 14,2 | 14,0 | | | |
| 8 | 12 | 7,3 | 9,4 | 10,2 | 10,3 | 10,8 | 10,6 | 11,1 | 11,5 | 11,5 | 12,9 | 13,4 | 15,0 | | |
| 9 | 13 | 6,9 | 7,6 | 8,5 | 10,0 | 10,9 | 11,0 | 11,4 | 11,6 | 12,0 | 12,0 | 12,7 | 13,5 | 14,0 | |
| 10 | 14 | 7,0 | 8,0 | 9,0 | 10,0 | 10,0 | 10,5 | 11,0 | 11,2 | 11,2 | 11,7 | 12,5 | 12,9 | 13,5 | 13,5 |

Assim, conclui-se que não é vantagem ter poucas plantas produzindo muito frutos, mas sim muitas plantas produzindo poucos frutos, mas grandes e uniformes. Uma análise mais detalhada poderá determinar o número ideal de frutos por árvore, aquele que maximiza o peso médio e, ao mesmo tempo, minimiza a variabilidade do peso.

Exemplo da Capacidade Pulmonar:

Captopril é um remédio destinado a baixar a pressão sistólica. Para testar seu efeito, ele foi ministrado a 12 pacientes, tendo sido medida a pressão sistólica antes e depois da medicação (Tabela 9).

Tabela 9: Pressão sistólica (mmHg) medida em 12 pacientes antes e depois do Captopril.

| Paciente | A | B | C | D | E | F | G | H | I | J | K | L |
|----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Antes | 200 | 174 | 198 | 170 | 179 | 182 | 193 | 209 | 185 | 155 | 169 | 210 |
| Depois | 191 | 170 | 177 | 167 | 159 | 151 | 176 | 183 | 159 | 145 | 146 | 177 |

Os mesmos indivíduos foram utilizados nas duas amostras (Antes/depois). Assim, é natural compararmos a pressão sistólica para cada indivíduo, comparando a pressão sistólica depois e antes. Para todos os pacientes, a pressão sistólica depois do Captopril é menor do que antes da medicação. Mas como podemos ver se estas diferenças são grandes? Através do diagrama de dispersão mostrado na Figura 29.

Cada ponto no diagrama de dispersão corresponde às medidas de pressão sistólica de um paciente, medida antes e depois da medicação.

A linha marcada no diagrama corresponde à situação onde a pressão sistólica não se alterou depois do paciente tomar o Captopril.

Veja que todos os pontos estão abaixo desta linha, ou seja para todos os pacientes o Captopril fez efeito. Grande parte destes pontos está bem distante da linha, mostrando que a redução na pressão sistólica depois do uso do medicamento não foi pequena.

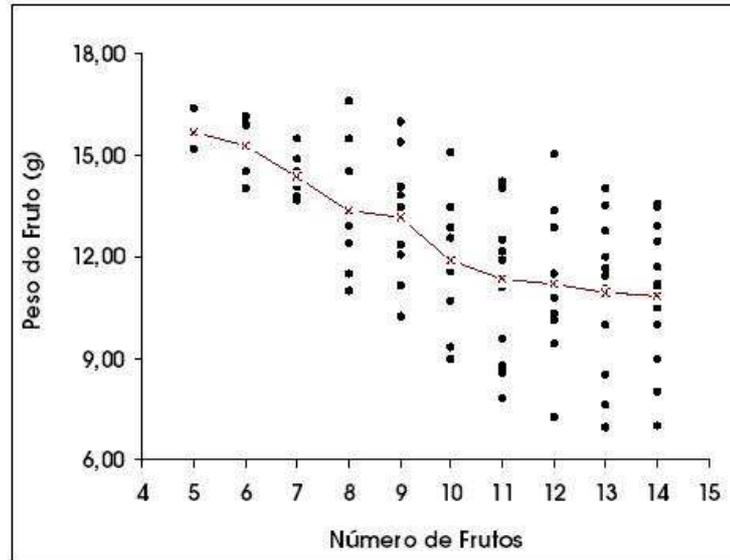


Figura 28: Diagrama de dispersão do número de frutos por árvore versus o peso do fruto e linha unindo os pesos médios dos frutos.

4 Estatística Descritiva - Medidas Resumo

4.1 Dados qualitativos

Para sumarizar dados qualitativos numericamente, utiliza-se **contagens, proporções, percentagens, taxas por 1000, taxas por 1.000.000**, etc, dependendo da escala apropriada.

Por exemplo, se encontrarmos que 70 de 140 estudantes de medicina são homens, poderíamos relatar a taxa como uma proporção (0.5) ou provavelmente ainda melhor como um percentual (50%).

Se encontrarmos que 7 de uma amostra de 5000 pessoas são portadores de uma doença rara poderíamos expressar isto como uma proporção observada (0.0014) ou percentual (0.14%), mas melhor seria 1.4 casos por mil.

4.1.1 Resumindo numericamente

Considere o seguinte conjunto de dados que mostra os escores de abundância médios *DAFOR* de ocorrência de *Nardus stricta* em 100 áreas investigadas em Exmoor, Inglaterra.

| | |
|-----------|----|
| Dominante | 8 |
| Abundante | 33 |
| Frequente | 32 |
| Ocasional | 17 |
| Raro | 10 |

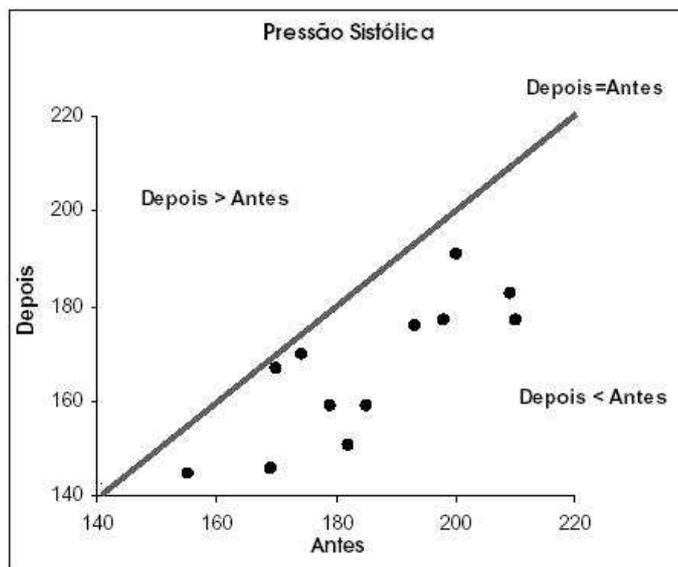


Figura 29: Diagrama de dispersão da pressão sistólica antes X depois da medicação e linha correspondendo ao não efeito individual da medicação.

A moda de um conjunto de dados categóricos é a categoria que tem o maior percentual de dados. Ela deve ser usada cuidadosamente como uma medida resumo global porque é muito dependente da forma como os dados são categorizados. Para os dados dos sexos dos ursos marrons a moda é *machos*. Para os dados acima, a categoria modal é “Abundante”, mas por muito pouco.

A mediana, bem como a moda, podem ser calculadas para **dados ordenados**. Este é valor do “meio”, mais comumente usado para dados quantitativos. A mediana não faz sentido para os dados dos sexos dos ursos.

Já para os dados de abundância, a categoria mediana é “Frequente”, porque 50% dos dados estão em categorias superiores, e menos do que 50% estão em categorias inferiores. A mediana é mais **robusta** do que a moda pois é menos sensível à categorização adotada.

4.2 Dados quantitativos

4.2.1 Resumindo numericamente

Para resumir numericamente dados quantitativos o objetivo é escolher medidas apropriadas de **locação** (“qual o tamanho dos números envolvidos?”) e de **dispersão** (“quanta variação existe?”) para os tipos de dados.

Existem três escolhas principais para a medida de locação, a chamada “3 Ms”, as quais estão ligadas a certas medidas de dispersão como segue:

| M | 'Dispersão' |
|------------------------------------|---------------|
| média (o valor 'médio') | desvio padrão |
| mediana (o valor do 'meio') | IQR |
| moda (o valor 'mais comum') | proporção |

4.2.2 A moda

Nem todos os conjuntos de dados são suficientemente balanceados para o cálculo da média ou mediana. Algumas vezes, especialmente para dados de contagem, um único valor domina a amostra.

A medida de locação apropriada é então a **moda**, a qual é o valor que ocorre com maior frequência. A proporção da amostra a qual toma este valor modal deveria ser utilizada no lugar de uma medida formal de dispersão.

Algumas vezes, podemos distinguir claramente dois ou mais 'picos' na frequência dos valores registrados. Neste caso (chamado **bimodal/multimodal**) deveríamos apresentar ambas as localizações. Dados deste tipo são particularmente difíceis de resumir (e analisar).

Exemplo. Dez pessoas registraram o número de copos de cerveja que eles tomaram num determinado sábado:

$$0, 0, 0, 0, 0, 1, 2, 3, 3, 6$$

A moda é 0 copos de cerveja, a qual foi obtida pela metade da amostra. Poderíamos adicionar mais informação separando a amostra e dizendo que daqueles que tomaram cerveja a mediana foi de 3 copos.

4.2.3 A mediana e a amplitude inter-quartis

Uma outra forma de sumarizar dados é em termos dos **quantis** ou **percentis**. Essas medidas são particularmente úteis para dados não simétricos.

A **mediana** (ou percentil 50) é definida como o valor que divide os dados ordenados ao meio, i.e. metade dos dados têm valores maiores do que a mediana, a outra metade tem valores menores do que a mediana.

Adicionalmente, os quartis **inferior** e **superior**, Q1 e Q3, são definidos como os valores abaixo dos quais estão um quarto e três quartos, respectivamente, dos dados.

Estes três valores são frequentemente usados para resumir os dados juntamente com o mínimo e o máximo.

Eles são obtidos ordenando os dados do menor para o maior, e então conta-se o número apropriado de observações: ou seja é $\frac{n+1}{4}$, $\frac{n+1}{2}$ e $\frac{3(n+1)}{4}$ para o quartil inferior, mediana e quartil superior, respectivamente.

Para um número par de observações, a mediana é a média dos valores do meio (e analogamente para os quartis inferior e superior).

A medida de dispersão é a **amplitude inter-quartis**, $IQR = Q3 - Q1$, i.e. é a diferença entre o quartil superior e o inferior.

Exemplo. O número de crianças em 19 famílias foi

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

A **mediana** é o $(19+1) / 2 = 10^o$ valor, i.e. 3 crianças.

O quartil **inferior** e **superior** são os valores 5^o e 15^o , i.e. 2 e 6 crianças, portanto **amplitude inter-quartil** é de 4 crianças. Note que 50% dos dados estão entre os quartis inferior e superior.

Box-and-Whisker Plots

Box-and-Whisker plots ou simplesmente **box-plots** são simples representações diagramáticas dos cinco números sumários: (mínimo, quartil inferior, mediana, quartil superior, máximo).

Um box-plot para os dados geoquímicos fica como mostrado a seguir (Figura 30).

4.2.4 Média, variância e desvio padrão

Para resumir dados quantitativos aproximadamente **simétricos**, é usual calcular a **média** aritmética como uma medida de localização. Se x_1, x_2, \dots, x_n são os valores dos dados, então podemos escrever a média como

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n},$$

onde $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$ e frequentemente é simplificada para $\sum x_i$ ou até mesmo $\sum x$ que significa ‘adicione todos os valores de x ’.

A **variância** é definida como o ‘desvio quadrático médio da média’ e é calculada de uma amostra de dados como

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

A segunda versão é mais fácil de ser calculada, no entanto muitas calculadoras têm funções prontas para o cálculo de variâncias, e é raro ter que realizar todos os passos manualmente.

Comumente as calculadoras fornecerão a raiz quadrada da variância, o **desvio padrão**, i.e.

$$s = \sqrt{\text{variância}} = \sqrt{s^2}$$

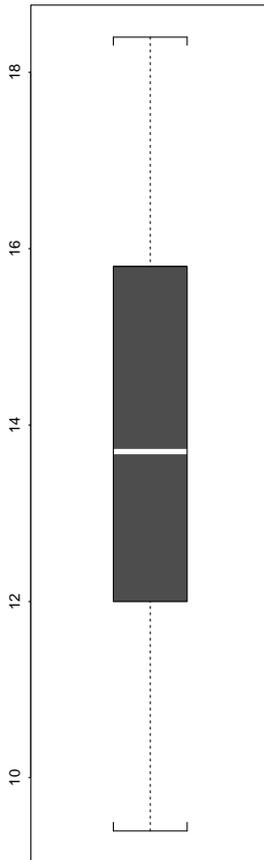


Figura 30: Representação dos 5 números sumários num box-plot

a qual é medida nas mesmas unidades dos dados originais.

Uma informação útil é que para qualquer conjunto de dados, pelo menos 75% deles fica dentro de uma distância de 2 desvios padrão da média, i.e. entre $\bar{x} - 2s$ e $\bar{x} + 2s$.

Exemplo. Sete homens foram pesados, e os resultados em kg foram:

57.0, 62.9, 63.5, 64.1, 66.1, 67.1, 73.6.

A **média** é $454.3/7 = 64.9$ kg,

a **variância** é $(29635.05 - 454.3^2/7)/6 = 25.16$ kg²

e o **desvio padrão** é $\sqrt{25.16} = 5.02$ kg.

4.2.5 Coeficiente de variação

Uma pergunta que pode surgir é: *O desvio padrão calculado é grande ou pequeno?*

Esta questão é relevante por exemplo, na avaliação da precisão de métodos.

Um desvio padrão pode ser considerado grande ou pequeno dependendo da ordem de grandeza da variável.

Uma maneira de se expressar a variabilidade dos dados tirando a influência da ordem de grandeza da variável é através do **coeficiente de variação**, definido por:

$$CV = \frac{s}{\bar{x}}$$

O CV é:

- interpretado como a **variabilidade dos dados em relação à média**. Quanto menor o CV mais homogêneo é o conjunto de dados.
- **adimensional**, isto é, um número puro, que será positivo se a média for positiva; será zero quando não houver variabilidade entre os dados, ou seja, $s = 0$.
- usualmente **expresso em porcentagem**, indicando o percentual que o desvio padrão é menor ($100\%CV < 100\%$) ou maior ($100\%CV > 100\%$) do que a média

Um CV é considerado baixo (indicando um conjunto de dados razoavelmente homogêneo) quando for **menor ou igual a 25%**. Entretanto, esse padrão varia de acordo com a aplicação.

Por exemplo, em medidas vitais (batimento cardíaco, temperatura corporal, etc) espera-se um CV muito menor do que 25% para que os dados sejam considerados homogêneos.

Pode ser difícil classificar um coeficiente de variação como baixo, médio, alto ou muito alto, mas este pode ser bastante útil na comparação de duas variáveis ou dois grupos que a princípio não são comparáveis.

Exemplos:

1. Em um grupo de pacientes foram tomadas as pulsações (batidas por minuto) e dosadas as taxas de ácido úrico (mg/100ml). As médias e os desvios padrão foram:

| Variável | \bar{x} | s |
|-------------|-----------|------|
| pulsação | 68,7 | 8,7 |
| ácido úrico | 5,46 | 1,03 |

Os coeficientes de variação são: $CV_p = 8,7/68,7 = 0,127$ e $CV_{a.u.} = 1,03/5,46 = 0,189$, o que evidencia que a pulsação é mais estável do que o ácido úrico.

2. Em experimentos para a determinação de clorofila em plantas, levantou-se a questão de que se o método utilizado poderia fornecer resultados mais consistentes. Três métodos foram colocados à prova e 12 folhas de abacaxi foram analisadas com cada um dos métodos. Os resultados foram os seguintes:

| Método (unidade) | \bar{x} | s | CV |
|------------------------|-----------|-------|-------|
| 1(100cm ³) | 13,71 | 1,20 | 0,088 |
| 2(100g) | 61,40 | 5,52 | 0,090 |
| 3(100g) | 337,00 | 31,20 | 0,093 |

Note que as médias são bastante diferentes devido às diferenças entre os métodos. Entretanto, os três CV são próximos, o que indica que a consistência dos métodos é praticamente equivalente, sendo que o método 3 mostrou-se um pouco menos consistente.

4.2.6 Escore padronizado

O escore padronizado, ao contrário do CV, é útil para comparação dos resultados individuais.

Por exemplo, um aluno que tenha obtido nota 7 numa prova cuja média da classe foi 5 foi melhor do que numa prova em que tirou 8 mas a média da classe foi 9.

Além da comparação da nota individual com a média da classe, também é importante avaliar em cada caso se a variabilidade das notas foi grande ou não.

Por exemplo, o desempenho deste aluno que obteve nota 7 seria bastante bom se o desvio padrão da classe fosse 2 e apenas razoável se o desvio padrão da classe fosse 4.

Sejam x_1, x_2, \dots, x_n os dados observados em uma amostra de tamanho n e \bar{x} e s a média e o desvio padrão, então

$$z_i = \frac{x_i - \bar{x}}{s}, i = 1, \dots, n$$

é denominado **escore padronizado**.

Os escores padronizados são muito úteis na comparação da posição relativa da medida de um indivíduo dentro do grupo ao qual pertence, o que justifica sua grande aplicação como medida de avaliação de desempenho.

Exemplo:

Os escores padronizados são amplamente utilizados em teste de aptidão física. Mathews (1980) compara testes de aptidão física e conhecimento desportivo.

Maria apresentou um desempenho muito acima da média em força abdominal (dois desvio padrão acima da média); sua capacidade aeróbica (corrida/caminhada) está acima da média mas não é notável e ela tem um conhecimento desportivo bastante bom comparado com o grupo.

No salto de extensão e na suspensão com flexão do braço sobre antebraço, Maria obteve escores abaixo das respectivas médias do grupo, sendo que o desempenho de Maria para salto em extensão é bastante ruim.

Tabela 10: Resultados obtidos por duas alunas do curso secundário, média e desvio padrão da turma em teste de aptidão física e conhecimento desportivo

| Teste | \bar{x} | s | x | | z | |
|------------------------------------|-----------|-----|-------|-------|-------|-------|
| | | | Maria | Joana | Maria | Joana |
| abdominais em 2 min | 30 | 6 | 42 | 38 | 2,00 | 1,33 |
| salto em extensão (cm) | 155 | 23 | 102 | 173 | -2.33 | 0,78 |
| suspensão braços flexionados (seg) | 50 | 8 | 38 | 71 | -1.50 | 2,63 |
| correr/andar em 12 min (m) | 1829 | 274 | 2149 | 1554 | 1,17 | -1,00 |
| conhecimento desportivo | 75 | 12 | 97 | 70 | 1,83 | -0,42 |

Descreva o desempenho de Joana.

5 Introdução à probabilidade e aplicação em testes diagnósticos

Nesta seção serão introduzidos conceitos probabilísticos aplicados a um problema de verificação da qualidade de um teste diagnóstico.

5.1 Probabilidade

De maneira informal, probabilidade é uma medida da certeza de ocorrência de um evento. Formalmente, existem duas definições de probabilidade: a definição clássica e a frequentista.

5.1.1 Definição clássica

Considere o seguinte experimento aleatório: lançar uma moeda e observar a face voltada para cima.

Este experimento possui dois resultados possíveis: cara e coroa. Ao conjunto dos resultados possíveis de um experimento chamamos de **espaço amostral** e será denotado pela letra E . O espaço amostral do experimento acima é $E = \{c, \bar{c}\}$, em que c denota cara e \bar{c} coroa.

Um subconjunto do espaço amostral é chamado de **evento** e é denotado por letras maiúsculas. Para o exemplo acima, podemos definir os eventos:

$$A = \{c\} = \{\text{ocorrer cara}\} \text{ e } B = \{\bar{c}\} = \{\text{ocorrer coroa}\}$$

O evento A acima é chamado de **evento simples** pois é constituído de apenas um elemento do espaço amostral. O mesmo se aplica para o evento B .

Seja A um evento qualquer do espaço amostral. Se os eventos simples são equiprováveis podemos calcular $P(A)$ como:

$$P(A) = \frac{\text{número de resultados favoráveis à ocorrência do evento } A}{\text{número de resultados possíveis}} \quad (1)$$

Para o experimento acima se a moeda é não viciada, os eventos A e B são equiprováveis e $P(A) = P(B) = 1/2$.

No lançamento de um dado não viciado, os eventos simples são equiprováveis com probabilidade $1/6$, $P(\text{sair um número par}) = 3/6 = 1/2$, $P(\text{sair número 1 ou 3}) = 2/6 = 1/3$ e $P(\text{sair número maior do que 2}) = 4/6 = 2/3$.

5.1.2 Definição frequentista

Na maioria das situações práticas, os eventos simples do espaço amostral não são equiprováveis e não podemos calcular probabilidades usando a definição clássica. Neste caso, vamos calcular probabilidades como a frequência relativa de um evento. Segue um exemplo que ilustra o método.

Tabela 11: Classificação de uma amostra de 6800 pessoas quanto à cor dos olhos e à cor dos cabelos

| Cor dos olhos | Cor dos cabelos | | | | Total |
|---------------|-----------------|----------|-------|-------|-------|
| | Loiro | Castanho | Preto | Ruivo | |
| Azul | 1768 | 807 | 189 | 47 | 2811 |
| Verde | 946 | 1387 | 746 | 53 | 3132 |
| Castanho | 115 | 438 | 288 | 16 | 857 |
| Total | 2829 | 2632 | 1223 | 116 | 6800 |

Exemplo 1: Uma amostra de 6800 pessoas de uma determinada população foi classificada quanto à cor dos olhos e à cor dos cabelos. Os resultados foram:

Considere o experimento aleatório que consiste em classificar um indivíduo quanto à cor dos olhos. O espaço amostral é $E = \{A, V, C\}$, em que:

$A = \{\text{a pessoa tem olhos azuis}\}$
 $V = \{\text{a pessoa tem olhos verdes}\}$
 $C = \{\text{a pessoa tem olhos castanhos}\}$

Os eventos acima são claramente equiprováveis. Então vamos calcular a probabilidade de ocorrer um evento como a frequência relativa deste evento:

$$P(A) = \frac{\text{número de pessoas de olhos azuis}}{\text{número de pessoas na amostra}} = \frac{2811}{6800} = 0,4134 \quad (2)$$

O valor obtido é na verdade uma **estimativa** da probabilidade. A qualidade desta estimativa depende do número de replicações do experimento, ou seja, do tamanho da amostra.

À medida que o tamanho da amostra cresce, a estimativa aproxima-se mais do valor verdadeiro da probabilidade. Vamos, no entanto, assumir que o número de replicações é suficientemente grande para que a diferença entre a estimativa e o valor verdadeiro da probabilidade seja desprezível.

As probabilidades dos eventos V e C são:

$$P(V) = \frac{3132}{6800} = 0,4606 \text{ e } P(C) = \frac{857}{6800} = 0,1260$$

Observe que $P(A) + P(V) + P(C) = 1$. Este resultado é geral, uma vez que a união destes eventos corresponde ao espaço amostral.

Seja \bar{A} o evento {a pessoa não tem olhos azuis}. O evento \bar{A} é chamado de evento complementar de A e $P(\bar{A}) = \frac{3132+857}{6800} = 0,5866 = 1 - P(A)$.

Estes resultados são propriedades de probabilidades. Seja A um evento qualquer no espaço amostral E. Então valem as propriedades:

1. $0 \leq P(A) \leq 1$
2. $P(E) = 1$

$$3. P(\bar{A}) = 1 - P(A)$$

Voltando ao exemplo, vamos calcular algumas probabilidades. Seja L o evento {a pessoa tem cabelos loiros}.

Qual a probabilidade de uma pessoa ter olhos azuis e cabelos loiros?

O evento {a pessoa tem olhos azuis e cabelos loiros} é chamado de **evento interseção**. Ele contém todos os elementos do espaço amostral pertencentes concomitantemente ao evento A e ao evento L e será denotado por $A \cap L$, e a probabilidade deste evento é:

$$P(A \cap L) = \frac{1768}{6800} = 0,26 \quad (3)$$

Qual a probabilidade de uma pessoa ter olhos azuis ou cabelos loiros?

O evento {a pessoa tem olhos azuis ou cabelos loiros} é chamado de evento união e será denotado por $A \cup L$. Ele contém todos os elementos do espaço amostral que estão em A , ou somente em L , ou em ambos, e a probabilidade deste evento é:

$$P(A \cup L) = P(A) + P(L) - P(A \cap L) = \frac{2811}{6800} + \frac{2829}{6800} - \frac{1768}{6800} = \frac{3872}{6800} = 0,5694 \quad (4)$$

Para quaisquer dois eventos A e B do espaço amostral, podemos calcular a probabilidade do evento união da seguinte forma: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Se os eventos são **mutuamente exclusivos**, isto é, eles não podem ocorrer simultaneamente, $P(A \cap B) = 0$ e conseqüentemente

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

Num exemplo de lançamento de um dado como os eventos $P = \{\text{sair número par}\}$ e $I = \{\text{sair número ímpar}\}$ são mutuamente exclusivos, $P(P \cup I) = P(P) + P(I) = 3/6 + 3/6 = 1$. Entretanto, os eventos $O = \{\text{sair número 1 ou 3}\}$ e $Q = \{\text{sair número maior que 2}\}$ não são mutuamente exclusivos, pois $O \cap Q = \{3\}$. Neste caso, $P(O \cup Q) = P(O) + P(Q) - P(O \cap Q) = 2/6 + 4/6 - 1/6 = 5/6$.

A propriedade acima pode ser estendida para mais de dois eventos. Para 3 eventos quaisquer (A, B e C) no espaço amostral, a probabilidade do evento união ($A \cup B \cup C$) é

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

Se os eventos A, B e C são mutuamente exclusivos

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

No exemplo da cor dos olhos, os eventos A, V e C são mutuamente exclusivos e $P(A) + P(V) + P(C) = 1$.

5.1.3 Probabilidade condicional

A probabilidade de um evento A ocorrer, dado que se sabe que um evento B ocorreu, é chamada probabilidade condicional do evento A dado B. Ela é denotada por $P(A|B)$ e calculada por:

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

Esta expressão pode ser reescrita como:

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

A probabilidade do evento \bar{A} (complementar de A) dado que o evento B ocorreu, isto é, $P(\bar{A}|B)$, é expressa por:

$$P(\bar{A}|B) = 1 - P(A|B)$$

Os eventos A e B são independentes se o fato de um deles ter ocorrido não altera a probabilidade da ocorrência do outro, isto é,

$$P(A|B) = P(A) \text{ ou } P(B|A) = P(B)$$

Da regra da multiplicação temos:

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

Exemplo 2: Considerando o Exemplo 1

- a. Qual a probabilidade de uma pessoa escolhida ao acaso da população ter olhos azuis dado que possui cabelos loiros?

$$P(A|L) = \frac{P(A \cap L)}{P(L)} = \frac{1768/6800}{2829/6800} = \frac{1768}{2829} = 0,6250$$

Observe que quando condicionamos em L, restringimos o espaço amostral ao conjunto das pessoas loiras. Note que $P(A) = 0,4134 < P(A|L) = 0,6250$ e que os eventos A e L não são independentes pois $P(A|L) \neq P(A)$.

- b. Qual a probabilidade de uma pessoa escolhida ao acaso da população não ter cabelos loiros dado que tem olhos castanhos?

$$P(\bar{L}|C) = 1 - P(L|C) = 1 - \frac{115/6800}{857/6800} = 1 - 0,1342 = 0,8658$$

Exemplo 3: Um casal possui 2 filhos. Qual a probabilidade de ambos serem do sexo masculino?

Os eventos $M = \{\text{nascer uma criança do sexo masculino}\}$ e $F = \{\text{nascer uma criança do sexo feminino}\}$ são equiprováveis. Logo, a probabilidade de nascer um filho do sexo masculino é $1/2$. A ocorrência do evento $A = \{\text{o primeiro filho é do sexo masculino}\}$ não influencia a ocorrência do evento $B = \{\text{o segundo filho é do sexo masculino}\}$, e então:

$$P(A \cap B) = P(A)P(B) = 1/2 \times 1/2 = 1/4$$

5.2 Avaliação da qualidade de testes diagnósticos

Ao fazer um diagnóstico, um clínico estabelece um conjunto de diagnósticos alternativos com base nos sinais e sintomas do paciente. Progressivamente ele reduz suas alternativas até chegar à uma doença específica.

Alternativamente, ele pode ter fortes evidências de que o paciente tem uma determinada doença e deseja apenas sua confirmação. Para chegar à uma conclusão final o clínico utiliza-se de testes diagnósticos:

- exames de laboratório (ex. dosagem de glicose)
- exame clínico (ex. auscultação do pulmão)
- questionário (ex. CDI (Children's Depression Inventory))

Um teste diagnóstico é um instrumento capaz de diagnosticar a doença com determinada precisão. Para cada teste diagnóstico existe um **valor de referência** que determina a classificação do resultado do teste como **negativo** ou **positivo**.

Um teste diagnóstico é considerado útil quando ele identifica bem a presença da doença. Antes de ser adotado o teste deve ser avaliado para verificar sua capacidade de acerto. Esta avaliação é feita aplicando-se o teste a dois grupos de pessoas: um grupo doente e outro não doente. Nesta fase, o diagnóstico é feito por outro teste chamado padrão ouro.

Os resultados obtidos podem ser organizados de acordo com a tabela abaixo:

Tabela 12: Resultados de um teste para pacientes doentes e não doentes

| Doença | Teste | | Total |
|-----------------------|-------|-----|-------|
| | + | - | |
| Presente (D) | a | b | a+b |
| Ausente (\bar{D}) | c | d | c+d |
| Total | a+c | b+d | n |

O teste é aplicado a n indivíduos, dos quais sabidamente $(a+b)$ são doentes e $(c+d)$ são não doentes.

Exemplo 3: Em um estudo sobre o teste ergométrico, Wriner et al. (1979) compararam os resultados obtidos entre indivíduos com e sem doença coronariana. O teste foi definido como positivo se foi observado mais de 1mm de depressão ou elevação do segmento ST, por pelo menos 0,08s, em comparação com os resultados obtidos com o paciente em repouso. O diagnóstico definitivo (classificação como doente ou não-doente) foi feito através de angiografia (teste padrão ouro).

Tabela 13: Resultados do teste ergométrico aplicado a 1023 pacientes com doença coronariana e 442 pacientes sem a doença

| Doença | Teste Ergométrico | | | | Total | |
|-----------|-------------------|-------|-----|-------|-------|-------|
| | Coronariana | + | - | | | |
| D | 815 | (a) | 208 | (b) | 1023 | (a+b) |
| \bar{D} | 115 | (c) | 327 | (d) | 442 | (c+d) |
| Total | 930 | (a+c) | 535 | (b+d) | 1465 | (n) |

Sejam os eventos:

- $D = \{\text{a pessoa tem doença coronariana}\}$
- $\bar{D} = \{\text{a pessoa não tem doença coronariana}\}$
- $+ = \{\text{o resultado do teste ergométrico é positivo}\}$
- $- = \{\text{o resultado do teste ergométrico é negativo}\}$

Temos interesse em responder duas perguntas:

1. Qual a probabilidade do teste ser positivo dado que o paciente é doente?
2. Qual a probabilidade do teste ser negativo dado que o paciente não é doente?

Em outras palavras, interessa conhecer as probabilidades condicionais:

$$s = P(+|D) = \frac{P(+ \cap D)}{P(D)} = \frac{a}{a+b}$$

e

$$e = P(-|\bar{D}) = \frac{P(-|\bar{D})}{P(\bar{D})} = \frac{d}{c+d}$$

Estas probabilidades são chamadas **sensibilidade** e **especificidade**. Numa situação ideal a sensibilidade e a especificidade deveriam ser 1.

Alternativamente, duas outras medidas que são de mais fácil interpretação são definidas por:

$$PFP = P(+|\bar{D}) = 1 - e$$

e

$$PFN = P(-|D) = 1 - s$$

a **proporção de falsos positivos** e a **proporção de falsos negativos**.

Exercício: Calcule s , e , PFP e PFN para o exemplo do teste ergométrico.

5.3 Valor de predição de um teste

Além dos índices de sensibilidade e especificidade, o clínico precisa decidir se considera o paciente doente ou não uma vez tendo o resultado do teste daquele paciente. A ele interessa conhecer o valor de predição positiva e o valor de predição negativa de um teste:

$$VPP = P(D|+) = \frac{a}{a+c} \text{ e } VPN = P(\bar{D}|-) = \frac{d}{b+d}$$

Exercício: Calcule os valores de VPP e VPN para o teste ergométrico.

Note que os valores de predição são afetados pela **prevalência da doença**, a proporção de pessoas com a doença na população que é estimada por $(a + b)/n$. Já a sensibilidade e especificidade não são afetados pela prevalência da doença.

6 Distribuições teóricas de frequências

Como visto anteriormente, as distribuições dos dados (que são variáveis aleatórias) podem ter uma variedade de formas, incluindo formas simétricas e não simétricas. Introduziremos aqui alguns dos modelos probabilísticos mais comumente usados para tais dados.

6.1 A distribuição Binomial

Considere um experimento realizado n vezes, sob as mesmas condições, com as seguintes características:

1. cada repetição do experimento (ou ensaio) produz um de dois resultados possíveis, denominados tecnicamente por sucesso (S) ou fracasso (F), ie os resultados são dicotômicos.
2. a probabilidade de sucesso, $P(S) = p$, é a mesma em cada repetição do experimento. (Note que $P(F) = 1 - p$).
3. os ensaios são independentes, ie o resultado de um ensaio não interfere no resultado do outro.

As quantidades n e p são os parâmetros da distribuição binomial. O número total de sucessos X é uma variável aleatória com distribuição binomial com parâmetros n e p e é por denotada $X \sim B(n, p)$.

A probabilidade de $X = x$, pode ser encontrada como:

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, \quad x = \{0, 1, 2, \dots\} \quad (5)$$

A **média** de um variável aleatória binomial é np e a **variância** é $np(1-p)$.

Para melhor entendimento considere o seguinte exemplo:

Suponha que num pedigree humano envolvendo albinismo (o qual é recessivo), nós encontraremos um casamento no qual sabe-se que ambos os parceiros são heterozigotos para o gene albino. De acordo com a teoria Mendeliana, a probabilidade de que um filho desse casal seja albino é um quarto. (Então a probabilidade de não ser albino é $\frac{3}{4}$.)

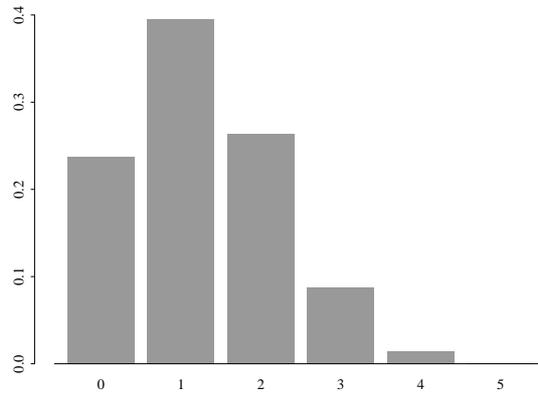
Agora considere o mesmo casal com 2 crianças. A chance de que ambas sejam albinas é $(\frac{1}{4})^2 = \frac{1}{16} = 0.0625$. Da mesma forma, a chance de ambas serem normais é $(\frac{3}{4})^2 = \frac{9}{16} = 0.5625$. Portanto, a probabilidade de que somente uma seja um albina deve ser $1 - \frac{1}{16} - \frac{9}{16} = \frac{6}{16} = \frac{3}{8} = 0.375$.

Alternativamente, poderíamos ter usado a formula acima definindo como variável aleatória X o número de crianças albinas, com $n = 2$, $p = \frac{1}{4}$, e estaríamos interessados em $P(X = 1)$.

Se agora considerarmos a família com $n = 5$ crianças, as probabilidades de existam $x = 0, 1, 2, \dots, 5$ crianças albinas, em que a probabilidade de albinismo é $p = \frac{1}{4}$, são dadas por

$$P(X = x) = \frac{5!}{x!(5-x)!} \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{5-x} \quad (6)$$

as quais ficam como segue.



O número esperado (ou média) de crianças albinas em famílias com 5 crianças para casais heterozigotos para o gene albino é $np = 5 \times \frac{1}{4} = 1,25$.

Exercício: Você leva sua cadela ao veterinário e descobre através de um exame de ultrasonografia que ela está grávida com uma ninhada de 8 filhotes.

- Qual é a probabilidade de que exatamente 3 dos filhotes sejam fêmeas?
- Qual é a probabilidade de que existam um número igual de machos e fêmeas?
- Qual é a probabilidade de que existam mais machos do fêmeas?

6.2 A distribuição Normal

A **distribuição Normal** é a mais familiar das distribuições de probabilidade e também uma das mais importantes em estatística.

Exemplo: O peso de recém-nascidos é uma variável aleatória contínua. A Figura 31 e Figura 32 abaixo mostram a distribuição de frequências relativas de 100 e 5000 pesos de recém-nascidos com intervalos de classe de 500g e 125g, respectivamente.

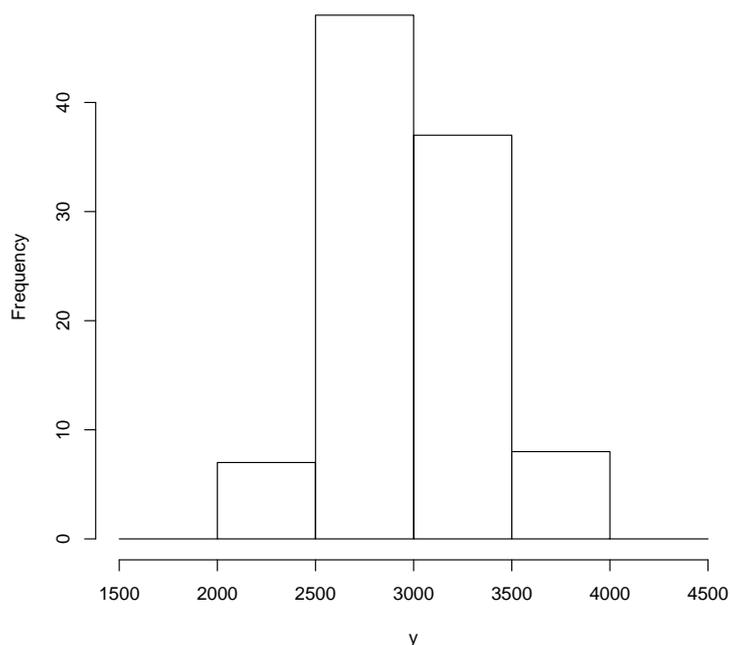


Figura 31: Histograma de frequências relativas a 100 pesos de recém-nascidos com intervalo de classe de 500g

O segundo histograma é um refinamento do primeiro, obtido aumentando-se o tamanho da amostra e reduzindo-se a amplitude dos intervalos de classe. Ele sugere a curva na Figura 33, que é conhecida como curva normal ou Gaussiana.

A variável aleatória considerada neste exemplo e muitas outras variáveis da área biológica podem ser descritas pelo modelo normal ou Gaussiano.

A equação da curva Normal é especificada usando 2 parâmetros: a **média** μ , e o **desvio padrão** σ .

Denotamos $N(\mu, \sigma)$ à curva Normal com média μ e desvio padrão σ .

A média refere-se ao centro da distribuição e o desvio padrão ao espalhamento (ou achatamento) da curva.

A distribuição normal é simétrica em torno da média o que implica que a média, a mediana

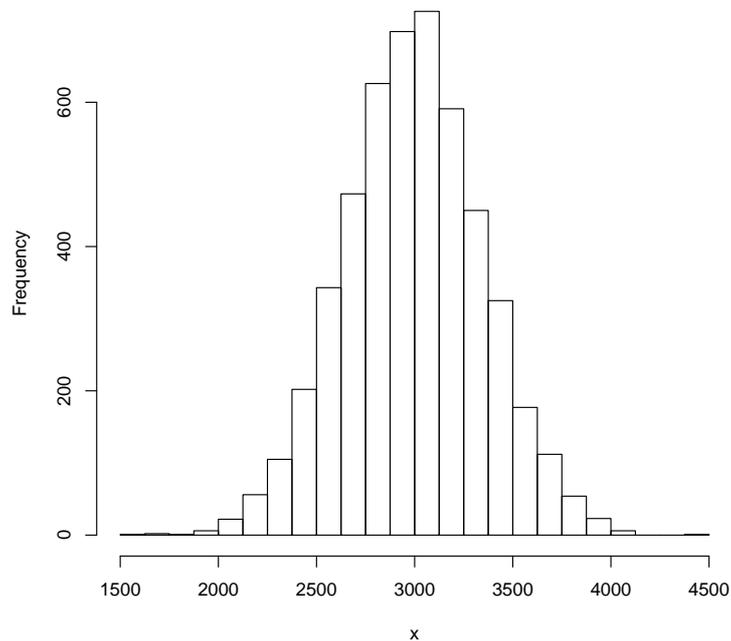


Figura 32: Histograma de frequências relativas a 5000 pesos de recém-nascidos com intervalo de classe de 125g

e a moda são todas coincidentes.

Para referência, a equação da curva é

$$f(x) = \frac{1}{\sqrt{(2\pi\sigma^2)}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}. \quad (7)$$

Felizmente, você não tem que memorizar esta equação. O importante é que você entenda como a curva é afetada pelos valores numéricos de μ e σ . Isto é mostrado no diagrama da Figura 34.

A área sob a curva normal (na verdade abaixo de qualquer função de densidade de probabilidade) é 1. Então, para quaisquer dois valores específicos podemos determinar a proporção de área sob a curva entre esses dois valores.

Para a distribuição Normal, a proporção de valores caindo dentro de um, dois, ou três desvios padrão da média são:

| Range | Proportion |
|-------------------|------------|
| $\mu \pm 1\sigma$ | 68.3% |
| $\mu \pm 2\sigma$ | 95.5% |
| $\mu \pm 3\sigma$ | 99.7% |

Exemplo: Suponhamos que no exemplo do peso do recém-nascidos $\mu = 2800g$ e $\sigma = 500g$.

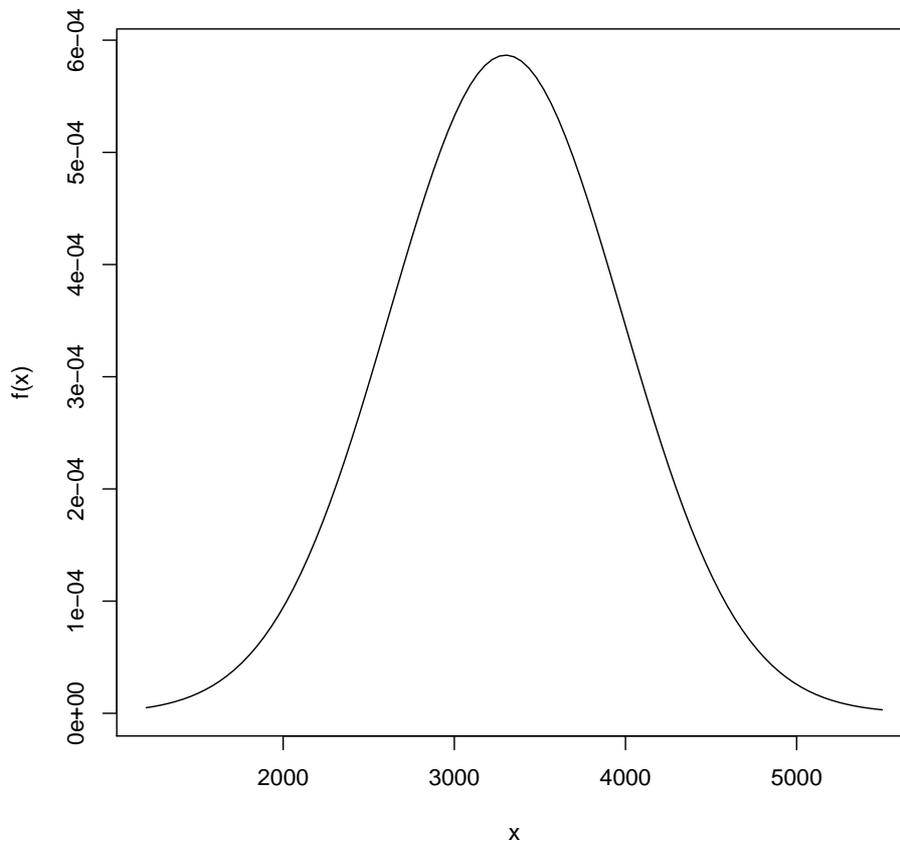


Figura 33: Função de densidade de probabilidade para a variável aleatória contínua X =peso do recém-nascido (g)

Então:

$$P(2300 \leq X \leq 3300) = 0,683$$

$$P(1800 \leq X \leq 3800) = 0,955$$

$$P(1300 \leq X \leq 4300) = 0,997$$

Usando este modelo podemos dizer que cerca de 68% dos recém-nascidos pesam entre 2300g e 3300g. O peso de aproximadamente 95% dos recém-nascidos está entre 1800g e 3800g. Praticamente todos os bebês desta população nascem com peso no intervalo (1300,4300).

Na prática desejamos calcular probabilidades para diferentes valores de μ e σ .

Para isso, a variável X cuja distribuição é $N(\mu, \sigma)$ é transformada numa forma padronizada Z com distribuição $N(0, 1)$ (**distribuição normal padrão**) pois tal distribuição é tabelada.

A quantidade Z é dada por

$$Z = \frac{X - \mu}{\sigma} \tag{8}$$

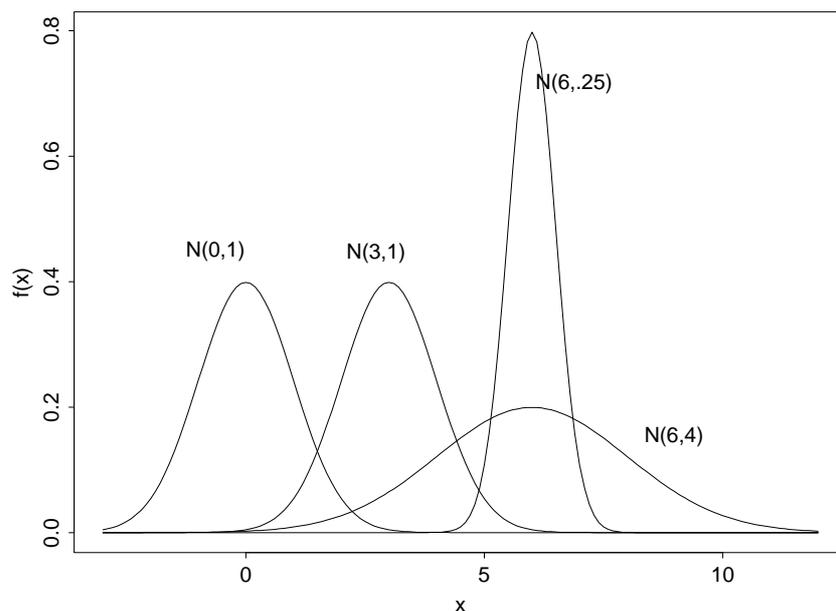


Figura 34: distribuições normais com mesma média μ e vários valores de σ

Exemplo: A concentração de um poluente em água liberada por uma fábrica tem distribuição $N(8,1.5)$. Qual a chance, de que num dado dia, a concentração do poluente exceda o limite regulatório de 10 ppm?

A solução do problema resume-se em determinar a proporção da distribuição que está acima de 10 ppm, ie $P(X > 10)$. Usando a estatística Z temos:

$$P(X > 10) = P\left(Z > \frac{10 - 8}{1.5}\right) = P(Z > 1.33) = 1 - P(Z \leq 1.33) = 0.09 \quad (9)$$

Portanto, espera-se que a água liberada pela fábrica exceda os limites regulatórios cerca de 9% do tempo.

Exercício: A concentração de cádmio em cinzas de um certo lixo radioativo tem distribuição $N(1,0.72)$. Quais são as chances de que uma amostra aleatória das cinzas tenha uma concentração de cádmio entre 0.5 e 1.75 ppm?

7 Inferência Estatística

A Estatística envolve métodos para o planejamento e condução de um estudo, descrição dos dados coletados e para tomada de decisões, predições ou inferências sobre os fenômenos representados pelos dados.

A qualidade dos resultados de um estudo depende basicamente do planejamento e condução do estudo e da análise dos dados. Os métodos estatísticos para análise de dados podem ser classificados como métodos descritivos - **Estatística Descritiva** - já vistos no início do curso e métodos inferenciais - **Inferência Estatística**.

A Inferência Estatística consiste de procedimentos para fazer generalizações sobre as características de uma população a partir da informação contida na amostra.

Exemplo:

Suponha que sementes geneticamente similares sejam selecionadas ao acaso e cultivadas em um ambiente enriquecido (tratamento) ou sob condições padrão (controle). Após determinado período de tempo, as plantas são cortadas, secas e pesadas.

Os resultados, expressos como o peso seco em gramas, para amostras de 10 plantas em cada ambiente são dadas abaixo:

| | | | | | | | | | |
|------------|------|------|------|------|------|------|------|------|------|
| Controle | 4,17 | 5,58 | 6,11 | 4,50 | 4,61 | 5,17 | 4,53 | 5,33 | 5,14 |
| Tratamento | 4,81 | 4,17 | 3,59 | 5,87 | 3,83 | 6,03 | 4,32 | 4,69 | 4,89 |

Neste exemplo podemos identificar duas populações e duas amostras:

População 1: Todas as possíveis plantas crescendo sob as mesmas condições do grupo tratamento

População 2: Todas as possíveis plantas crescendo sob as mesmas condições do grupo controle

Amostra 1: As 10 plantas cultivadas no ambiente enriquecido

amostra 2: As 10 plantas cultivadas no ambiente padrão

Interessa ao pesquisador verificar se existe efeito de tratamento e qual a magnitude deste efeito.

Esta pergunta será respondida com base na informação amostral.

O pesquisador deseja saber qual o melhor tratamento para a população, e não saber apenas o que aconteceu em suas amostras. Ele deseja generalizar, fazer inferências para a população.

Com este objetivo introduziremos dois procedimentos inferenciais a partir deste capítulo: **Estimação e Testes de hipóteses**.

7.1 Estimação

No exemplo acima interessa saber se existe efeito de fertilizante.

Mas o que é existir efeito de fertilizante?

Num mesmo tratamento, plantas diferentes respondem de formas diferentes (variabilidade). O peso seco das plantas é uma variável aleatória!

Vamos considerar que existe efeito de fertilizante quando o peso seco médio das plantas cultivadas em ambiente fertilizado diferir do peso seco médio das plantas cultivadas em ambiente padrão. Isto é, quando as distribuições do peso seco para o grupo controle e grupo tratamento apresentam médias, digamos μ_c e μ_t , diferentes.

As quantidades μ_c e μ_t são desconhecidas e chamadas **parâmetros**, e só podem ser conhecidas se observarmos toda a população, o que é quase sempre impossível.

O que fazemos é estimar os parâmetros a partir de uma amostra da população.

As médias μ_c e μ_t podem ser estimadas pelas médias amostrais \bar{X}_c e \bar{X}_t , que são funções dos valores da amostra e são chamadas de **estimadores** de μ_c e μ_t .

Os valores de \bar{X}_c e \bar{X}_t observados na amostra

$$\bar{x}_c = 5,03 \text{ g e } \bar{x}_t = 4,66 \text{ g}$$

são chamados de **estimativas** dos parâmetros. Observe que denotamos estimativas por letras minúsculas e estimadores por letras maiúsculas.

Exemplo: Exemplo 6.1.2 pág 122

Dois diferentes tipos de secagem foram usados na preparação de sementes. Duzentas sementes foram aleatoriamente selecionadas para serem submetidas a dois processos de secagem A e B. Após a secagem, as sementes foram observadas quanto à sua germinação. Os resultados foram:

| Processo de secagem | Germinação | | |
|---------------------|------------|-----|-------|
| | Sim | Não | Total |
| A | 70 | 30 | 100 |
| B | 62 | 38 | 100 |
| Total | 132 | 68 | 200 |

Neste caso interessa saber se existe diferença entre os métodos de secagem quanto à germinação de sementes. Vamos considerar que existe efeito de método de secagem quando as proporções populacionais de sementes germinadas pelos métodos A, p_A , e B, p_B , diferem.

Os parâmetros de interesse p_A e p_B são estimados pelas proporções amostrais

$$\hat{p}_A = \frac{x_A}{n_A} \text{ e } \hat{p}_B = \frac{x_B}{n_B}$$

em que

x_A é o número de sementes submetidas ao processo A que germinaram;

n_A é o número total de sementes submetidas ao processo A;

x_B é o número de sementes submetidas ao processo B que germinaram;

n_B é o número total de sementes submetidas ao processo B;

As estimativas dos parâmetros p_A e p_B são $\hat{p}_A = 0,70$ e $\hat{p}_B = 0,62$.

Nos exemplos acima, os parâmetros de interesse forma médias e proporções, mas poderíamos estar interessados em estimar medianas, desvios-padrão, etc.

Diferentes amostras podem ser retiradas de uma mesma população, e amostras diferentes podem resultar em estimativas diferentes. Isto é, um estimador é uma variável aleatória, podendo assumir valores diferentes para cada amostra.

Então, ao invés de estimar o parâmetro de interesse por um único valor, é muito mais informativo estimá-lo por um intervalo de valores que considere a variação presente na amostra e que contenha o seu verdadeiro valor com determinada confiança. Este intervalo é chamado de **intervalo de confiança**.

Para construir um intervalo de confiança precisamos conhecer a distribuição de probabilidade do estimador. Lembre que um estimador é uma variável aleatória e que uma variável aleatória é completamente caracterizada por sua distribuição de probabilidade.

Na próxima seção serão apresentados resultados sobre a distribuição de probabilidade da média amostral.

7.1.1 Teorema Central do Limite

Uma razão para a distribuição Normal ser considerada tão importante é porque *qualquer que seja* a distribuição da variável de interesse **para grande amostras, a distribuição das médias amostrais serão aproximadamente normalmente distribuídas**, e tenderão a uma distribuição normal à medida que o tamanho de amostra crescer. Então podemos ter uma variável original com uma distribuição muito diferente da Normal (pode até mesmo ser discreta), mas se tomarmos várias amostras grandes desta distribuição, e então fizermos um histograma das médias amostrais, a forma se parecerá como uma curva Normal.

A distribuição da média amostral \bar{X} é aproximadamente Normal com média μ e desvio padrão σ/\sqrt{n} .

Aqui μ e σ são a média e o desvio padrão populacionais das medidas individuais X , e n é o tamanho amostral. Denota-se

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n}).$$

A aproximação para a normal melhora à medida que o tamanho amostral cresce. Este resultado é conhecido como o **Teorema Central do Limite** e é notável porque permite-nos conduzir alguns procedimentos de inferência sem qualquer conhecimento da distribuição da população.

Exemplo simulado: Podemos ilustrar o Teorema Central do Limite por um exemplo simulado. O diagrama na Figura 35 sumariza os resultados de um experimento no qual foi

utilizado um computador para gerar 2000 observações de duas distribuições bem diferentes (linha superior). Nós então geramos uma amostra de tamanho 2 de cada distribuição e calculamos a média. Este procedimento foi repetido 1999 vezes e a segunda linha mostra os histogramas das médias resultantes das amostras de tamanho dois. Isto foi repetido com médias amostrais onde as amostras são de tamanhos 5 (terceira linha) e 10 (quarta linha).

Note como a forma da distribuição muda à medida que se muda de uma linha para a próxima, e como as duas distribuições em cada linha tornam-se mais similares nas suas formas à medida que o tamanho das amostras aumenta. Ainda mais, cada distribuição parece mais e mais com uma distribuição Normal. Não é necessário uma amostra de tamanho muito grande para ver uma forma Normal.

As médias populacionais para as duas distribuições são 5 e 3 respectivamente. Note como, quanto maior o tamanho de amostra mais perto as médias amostrais tendem a estar da média populacional.

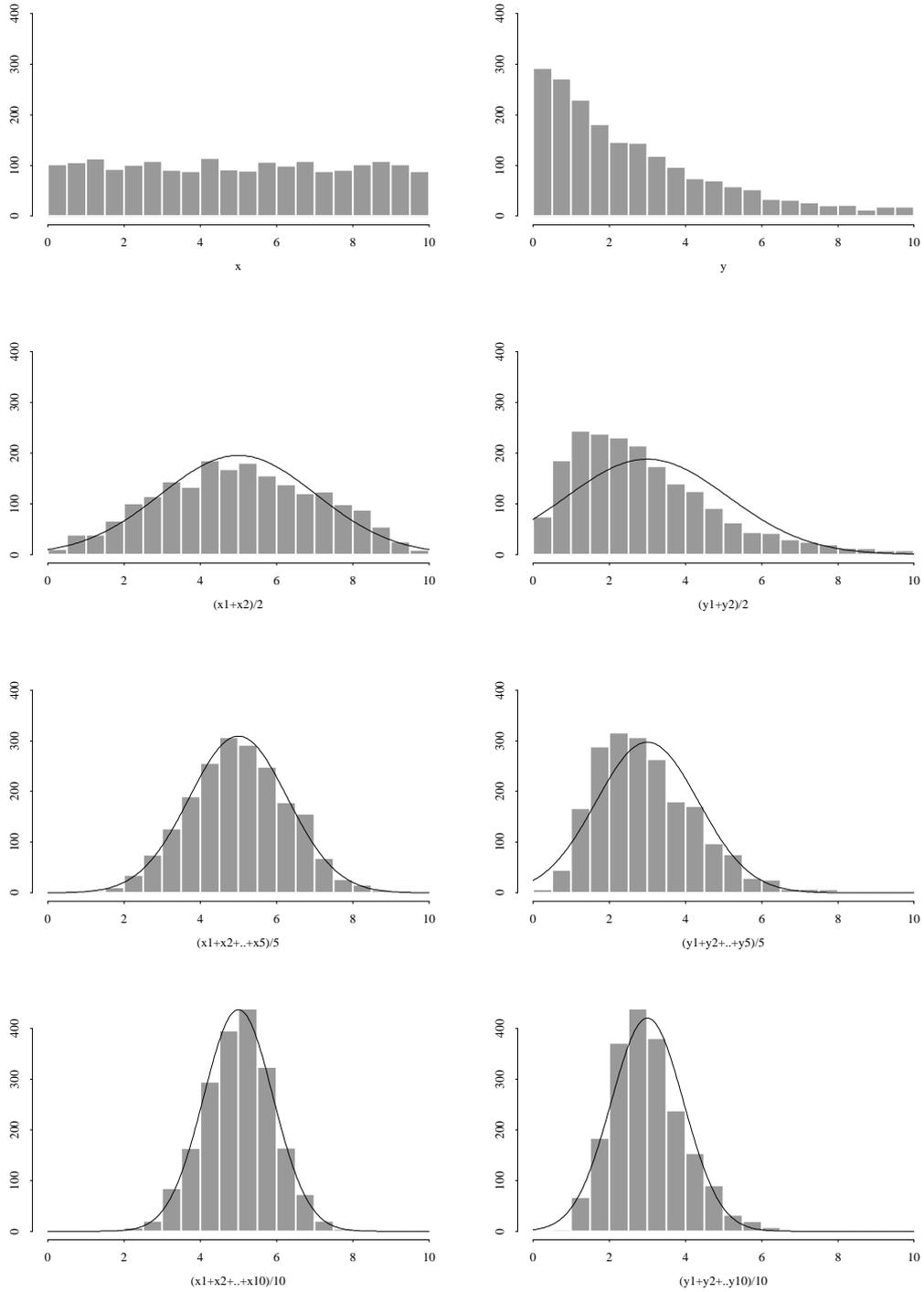


Figura 35: Teorema Central do Limite

Exemplo:

Suponha que para crianças nascidas com peso abaixo de 750g, o nível de bilirrubina sérico tem distribuição Normal com média 8,5mg/dl e desvio-padrão 3,5 mg/dl.

1. Calcule a probabilidade de que a média amostral \bar{X} , para uma amostra de 16 crianças:
 - (a) seja menor do que 8 mg/dl
 - (b) esteja entre 7,5 e 9,5 mg/dl
2. Encontre um intervalo simétrico em torno da média que contenha 95% dos valores de \bar{X} .

7.1.2 Intervalos de confiança de 95% para uma média

Na seção anterior vimos que para uma amostra suficientemente grande a distribuição das médias amostrais em torno da média populacional é Normal com desvio padrão σ/\sqrt{n} . Chamamos de σ/\sqrt{n} o **erro padrão** (SE) da média, uma vez que quanto menor seu valor tanto mais próximas estarão as médias amostrais da média populacional μ (i.e. tanto menor será o *erro*).

$$\begin{aligned}\text{média populacional} &= \mu \\ \text{desvio padrão populacional} &= \sigma \\ \text{SE da média} &= \sigma/\sqrt{n}\end{aligned}$$

Isto significa que 68.3% de todas as médias amostrais cairão dentro de ± 1 SE da média populacional μ . Similarmente 95% de todas as médias amostrais cairão dentro de $\pm 1.96 \times$ SE de μ .

Então intervalos da forma

$$\left(\bar{x} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right)$$

conterão a verdadeira média populacional μ 95% das vezes.

Um **problema** com a construção de tais intervalos é que não sabemos o verdadeiro desvio padrão populacional σ . Para grandes tamanhos amostrais, contudo, o desvio padrão amostral s será uma boa estimativa de σ . Portanto, podemos substituir σ por s de modo que podemos calcular o erro padrão como

$$\text{SE} = s/\sqrt{n},$$

e um intervalo de confiança de aproximadamente 95% para μ é:

$$\left(\bar{x} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{x} + 1.96 \times \frac{s}{\sqrt{n}}\right).$$

Este tipo de intervalo de confiança para a média pode ser usado para grandes amostras, independentemente da distribuição da variável original.

7.1.3 Intervalos de confiança mais exatos

Para amostras pequenas, onde s é uma estimativa menos confiável de σ , devemos construir nosso intervalo de confiança de uma forma ligeiramente diferente.

Ao invés de usar o valor 1.96, usamos um valor ligeiramente maior para refletir nossa redução na confiança. Obtemos o valor requerido da tabela de distribuição t . Tomamos o valor correspondente à linha $r = n - 1$ graus de liberdade. Note que quanto menor n , maiores os valores de t .

Então um intervalo de confiança exato é

$$\left(\bar{x} - t_{(n-1,0.05)} \times \frac{s}{\sqrt{n}} , \bar{x} + t_{(n-1,0.05)} \times \frac{s}{\sqrt{n}} \right).$$

Note ainda que à medida que n cresce, o valor de t torna-se próximo a 1.96.

Repare que se a distribuição da variável original é muito distante de uma normalmente distribuída, e o tamanho amostral é excessivamente pequeno, então as médias amostrais não terão uma distribuição aproximadamente normal e portanto este tipo de intervalo de confiança não deveria ser utilizado.

A distribuição t

Valores de t para que $P(|T| > t) = p$, onde T tem um distribuição T de Student com r graus de liberdade.

| | p | | | | |
|----------|-------|-------|--------|--------|---------|
| | 0.20 | 0.10 | 0.05 | 0.01 | 0.001 |
| 1 | 3.078 | 6.314 | 12.706 | 63.657 | 636.619 |
| 2 | 1.886 | 2.920 | 4.303 | 9.925 | 31.599 |
| 3 | 1.638 | 2.353 | 3.182 | 5.841 | 12.924 |
| 4 | 1.533 | 2.132 | 2.776 | 4.604 | 8.610 |
| 5 | 1.476 | 2.015 | 2.571 | 4.032 | 6.869 |
| 6 | 1.440 | 1.943 | 2.447 | 3.707 | 5.959 |
| 7 | 1.415 | 1.895 | 2.365 | 3.499 | 5.408 |
| 8 | 1.397 | 1.860 | 2.306 | 3.355 | 5.041 |
| 9 | 1.383 | 1.833 | 2.262 | 3.250 | 4.781 |
| 10 | 1.372 | 1.812 | 2.228 | 3.169 | 4.587 |
| 11 | 1.363 | 1.796 | 2.201 | 3.106 | 4.437 |
| 12 | 1.356 | 1.782 | 2.179 | 3.055 | 4.318 |
| 13 | 1.350 | 1.771 | 2.160 | 3.012 | 4.221 |
| 14 | 1.345 | 1.761 | 2.145 | 2.977 | 4.140 |
| 15 | 1.341 | 1.753 | 2.131 | 2.947 | 4.073 |
| 16 | 1.337 | 1.746 | 2.120 | 2.921 | 4.015 |
| r 17 | 1.333 | 1.740 | 2.110 | 2.898 | 3.965 |
| 18 | 1.330 | 1.734 | 2.101 | 2.878 | 3.922 |
| 19 | 1.328 | 1.729 | 2.093 | 2.861 | 3.883 |
| 20 | 1.325 | 1.725 | 2.086 | 2.845 | 3.850 |
| 21 | 1.323 | 1.721 | 2.080 | 2.831 | 3.819 |
| 22 | 1.321 | 1.717 | 2.074 | 2.819 | 3.792 |
| 23 | 1.319 | 1.714 | 2.069 | 2.807 | 3.768 |
| 24 | 1.318 | 1.711 | 2.064 | 2.797 | 3.745 |
| 25 | 1.316 | 1.708 | 2.060 | 2.787 | 3.725 |
| 26 | 1.315 | 1.706 | 2.056 | 2.779 | 3.707 |
| 27 | 1.314 | 1.703 | 2.052 | 2.771 | 3.690 |
| 28 | 1.313 | 1.701 | 2.048 | 2.763 | 3.674 |
| 29 | 1.311 | 1.699 | 2.045 | 2.756 | 3.659 |
| 30 | 1.310 | 1.697 | 2.042 | 2.750 | 3.646 |
| 40 | 1.303 | 1.684 | 2.021 | 2.704 | 3.551 |
| 50 | 1.299 | 1.676 | 2.009 | 2.678 | 3.496 |
| 60 | 1.296 | 1.671 | 2.000 | 2.660 | 3.460 |
| 70 | 1.294 | 1.667 | 1.994 | 2.648 | 3.435 |
| 80 | 1.292 | 1.664 | 1.990 | 2.639 | 3.416 |
| 90 | 1.291 | 1.662 | 1.987 | 2.632 | 3.402 |
| 100 | 1.290 | 1.660 | 1.984 | 2.626 | 3.390 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.576 | 3.291 |

Exemplos

Identificação de bactérias em hemoculturas

Um método padrão para identificação de bactérias em hemoculturas vem sendo utilizado há muito tempo, e seu tempo médio de execução (desde a etapa de preparo das amostras até a identificação do gênero e espécie) é de 40,5 horas. Um microbiologista propôs uma nova técnica afirmando que o tempo de execução deste novo processo é menor que o do método padrão.

Os dados abaixo (em horas) são resultantes da aplicação desta nova técnica.

41 38 38 42 39 40 40 38 36 35 43 40 40 41 40,5 40 39 39

$n=18$, $\bar{x}=39,42$ horas e $s=1,96$ horas

Vamos construir o intervalo de confiança de 95% para o verdadeiro tempo médio de execução deste novo processo.

O erro padrão é portanto:

$$SE = \frac{s}{\sqrt{n}} = \frac{1,96}{\sqrt{18}} = 0,462.$$

Temos uma amostra de tamanho $n = 18$, então da tabela da distribuição t com $18-1=17$ gl e $p=0,05$, temos que $t = 2,110$.

Então o intervalo de confiança de 95% para a média populacional é

$$\bar{x} \pm t \times SE = 39,42 \pm 2,110 \times 0,462 = (38,44; 40,39)$$

Portanto estamos 95% confiantes de que o tempo médio de execução do novo processo está entre 38,44 e 40,39 horas e concluímos que existem evidências amostrais de que o novo método para identificação de bactérias tem tempo médio de execução menor que o método padrão.

Exercícios:

1. Os pulsos em repouso de 920 pessoas sadias foram tomados, e uma média de 72.9 batidas por minuto (bpm) e um desvio padrão de 11.0 bpm foram obtidos. Construa um intervalo de confiança de 95% para a pulsação média em repouso de pessoas sadias com base nesses dados.
2. Os QIs de 20 meninos com idades entre 6-7 anos de Curitiba foram medidos. O QI médio foi 108.08, e o desvio padrão foi 14.38.
 - Calcule um intervalo de confiança de 95% para o QI médio populacional dos meninos entre 6-7 anos de idade em Curitiba usando estes dados.
 - Interprete o intervalo de confiança com palavras.
 - Foi necessário assumir que os QIs têm distribuição normal neste caso? Por quê?

7.1.4 Intervalos de confiança para uma proporção

Da mesma forma que um conjunto de médias amostrais são distribuídas nas proximidades da média populacional, as proporções amostrais \hat{p} são distribuídas ao redor da verdadeira proporção populacional p .

Devido ao Teorema Central do Limite, para n grande e p não muito próximo de 0 ou 1, a distribuição de \hat{p} será aproximadamente normalmente distribuída com média p e um desvio padrão dado por

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

Chamamos $SE = \sqrt{\frac{p(1-p)}{n}}$ de erro padrão da proporção amostral. Podemos usar isto na construção de um intervalo de confiança para a verdadeira proporção p .

Um intervalo de confiança de aproximadamente 95% para p é portanto

$$(\hat{p} - 1.96 \times SE, \hat{p} + 1.96 \times SE)$$

em que

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Note que não sabemos o verdadeiro valor de p , e portanto usamos \hat{p} na fórmula acima para estimar SE.

Uma regra geral é que este intervalo de confiança é válido quando quando temos ambos $n\hat{p}$ e $n(1-\hat{p})$ maiores do que digamos 10.

Exemplo:

Um ensaio clínico foi realizado para determinar a preferência entre dois analgésicos, A e B, contra dor de cabeça. Cem pacientes que sofrem de dor de cabeça crônica receberam em dois tempos diferentes o analgésico A e o analgésico B.

A ordem na qual os pacientes receberam os analgésicos foi determinada ao acaso. Os pacientes desconheciam esta ordem.

Ao final do estudo foi perguntado a cada paciente qual analgésico lhe proporcionou maior alívio: o primeiro ou o segundo. Dos 100 pacientes, 45 preferiram A e 55 preferiram B.

Baseado nestas informações podemos dizer que há preferência por algum dos analgésicos?

Dizemos que não há preferência por um dos analgésicos quando a proporção dos que preferem A (p_A), é igual a proporção dos que preferem B (p_B). Como temos dois resultados possíveis, p_A e p_B são iguais quando $p_A = p_B = 0,5$.

Um intervalo de 95% de confiança para a verdadeira proporção de pacientes que preferem o analgésico A é:

$$\left(0,45 \pm 1,96 \sqrt{\frac{0,45 \times 0,55}{100}} \right) = (0,35; 0,55)$$

Então com 95% de confiança, a verdadeira proporção de pacientes que preferem o analgésico A está entre 0,35 e 0,55. Observe que este intervalo contém o valor 0,5 então concluímos que não existem evidências amostrais de preferência por um dos analgésicos.

7.1.5 Comparação de intervalos de confiança

Suponha que tenhamos dois ou mais grupos separados, por exemplo, machos e fêmeas. Podemos construir um intervalo de confiança de 95% para a média para cada um dos grupos, e então construir um gráfico com esses intervalos contra um eixo comum para verificar se existe uma interseção (i.e. se existem alguns valores em comum). Se os intervalos não se sobrepõem, então temos (pelo menos) 95% de confiança de que as verdadeiras médias não são iguais.

Embora estes gráficos sejam úteis para visualização, utilizaremos uma abordagem mais formal para construir um intervalo de confiança para a diferença entre duas médias ou a diferença entre duas proporções.

Exemplo:

Considere os dados de um estudo investigando a existência de um balanço entre a proporção de peixes machos e fêmeas de uma certa espécie em dois lagos distintos.

A proporção observada de machos capturados no primeiro lago foi 74.4% dentre 43 capturados e no segundo foi 60% dentre 50.

Podemos agora construir intervalos de confiança para as percentagens correspondente nas populações dos dois lagos.

7.1.6 Dimensionamento de amostras

Vimos neste capítulo como construir intervalos para alguns parâmetros populacionais. Em todos os casos, fixamos o nível de confiança dos intervalos de acordo com a probabilidade de acerto que desejamos ter na estimação por intervalo.

Sendo conveniente, o nível de confiança pode ser aumentado até tão próximo de 100% quanto se queira, mas isso resultará em intervalos de amplitude cada vez maiores, o que significa perda de precisão na estimação.

Seria desejável termos intervalos com alto nível de confiança e grande precisão. Isso porém requer uma amostra suficientemente grande, pois, para n fixo, a confiança e a precisão variam em sentidos opostos.

Veremos a seguir como determinar o tamanho das amostras necessárias nos casos de estimação da média ou de uma proporção populacional.

Vimos que o intervalo de confiança de 95% para a média μ da população quando σ é conhecido tem semi-amplitude (ou precisão) d dada pela expressão

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

em que $z = 1.96$ para uma confiança de 95%.

Ora, o problema então resolvido foi, fixados o nível de confiança ($1 - \alpha = 0.95$) e n , determinar d . Mas, é evidente dessa expressão que podemos resolver outro problema.

Fixados, d (ou seja, fixada a precisão) e o nível de confiança, determinar n , que é o

problema da determinação do tamanho de amostra necessário para se realizar a estimação por intervalo com a confiança e a precisão desejadas.

Vemos imediatamente que

$$n = \left(\frac{z\sigma}{d} \right)^2.$$

Essa será a expressão usada se σ for conhecido.

Não conhecendo o desvio-padrão da população, deveríamos substituí-lo por sua estimativa s e usar t de Student na expressão acima.

Ocorre porém que não tendo ainda sido retirada a amostra, não dispomos em geral do valor de s . Se não conhecemos nem ao menos um limite superior para σ , a única solução será colher uma amostra-piloto de n_0 elementos para, com base nela obtermos uma estimativa de s , empregando a seguir a expressão

$$n = \left(\frac{t_{(n_0-1, 0.05)} s}{d} \right)^2.$$

Se $n \leq n_0$, a amostra-piloto já terá sido suficiente para a estimação. Caso contrário, deveremos retirar, ainda, da população os elementos necessários à complementação do tamanho mínimo de amostra.

Procedemos de forma análoga se desejamos estimar uma proporção populacional com determinada confiança e dada precisão. No caso de população suposta infinita, da expressão

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

podemos obter

$$n = \left(\frac{z}{d} \right)^2 p(1 - p).$$

O obstáculo à determinação do tamanho de amostra por meio da expressão acima está em desconhecermos p .

Essa dificuldade pode ser resolvida através de uma amostra-piloto, analogamente ao caso descrito para a estimação de μ , ou analisando-se o comportamento do fator $p(1 - p)$ para $0 \leq p \leq 1$.

Vê-se da Figura 36 a seguir que $p(1 - p)$ é a expressão de uma parábola cujo ponto de máximo é $p = 1/2$.

Se substituirmos, $p(1 - p)$ por seu valor máximo, $1/4$, seguramente o tamanho de amostra obtido será suficiente para a estimação de qualquer que seja p . Isso equivale a considerar

$$n = \left(\frac{z}{d} \right)^2 \frac{1}{4} = \left(\frac{z}{2d} \right)^2.$$

Evidentemente, usando-se essa expressão corre-se o risco de se superdimensionar a amostra. Isso ocorrerá se p for na realidade próximo de 0 ou 1. Se o custo envolvido for elevado e proporcional ao tamanho de amostra, é mais prudente a tomada de uma amostra-piloto.

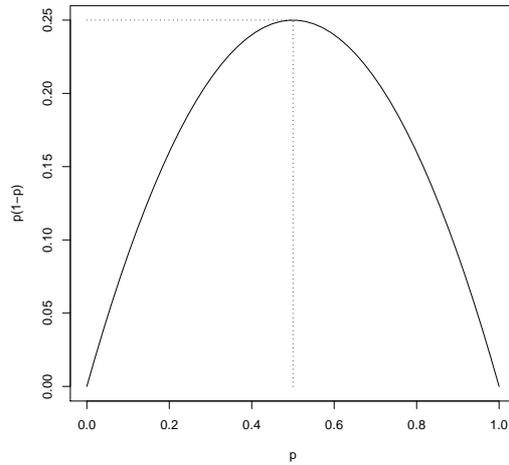


Figura 36: Gráfico da função $p(1-p)$.

Exemplos

1. Qual o tamanho de amostra necessário para se estimar a média de uma população infinita cujo desvio-padrão é igual a 4, com 98% de confiança e precisão de 0.5?
2. Qual o tamanho de amostra suficiente para estimarmos a proporção da área com solo contaminado que precisa de tratamento, com precisão de 0,02 e 95% de confiança, sabendo que essa proporção seguramente não é superior a 0.2?

7.2 Testes de Hipóteses

Em geral, intervalos de confiança são a forma mais informativa de apresentar os achados principais de um estudo.

Contudo, algumas vezes existe um particular interesse em decidir sobre a verdade ou não de uma hipótese específica (se dois grupos têm a mesma média ou não, ou se o parâmetro populacional tem um valor em particular ou não).

Os **Testes de hipóteses** fornecem-nos uma estrutura para que façamos isto. Veremos que intervalos de confiança e testes de hipóteses estão intimamente relacionados.

Exemplo:

Um pesquisador deseja responder a seguinte pergunta:

Os pássaros migratórios engordam antes de migrar?

Considere os dados coletados por um ornitologista sobre o uso de um determinado lugar para engorda por pássaros de uma certa espécie.

Pode-se perguntar se em média estes pássaros engordam entre Agosto e Setembro.

Somente 10 pássaros foram capturados e seu peso médio nas duas ocasiões foram 11.47 e 12.35 então o peso médio aumentou para esta amostra em particular. (Note que o mesmo conjunto de pássaros foram medidos ambas as vezes.)

Podemos generalizar para o resto dos pássaros que não foram capturados? Será que esta diferença poderia ser devida simplesmente ao acaso?

Em termos estatísticos queremos **testar** a **hipótese nula ou de nulidade** (H_0) de que, em média, não existe mudança no peso dos pássaros.

Assumiremos que os 10 pássaros foram uma amostra aleatória de todos os pássaros migradores daquela espécie e usaremos primeiramente o que aprendemos sobre intervalos de confiança para responder nossas perguntas.

Primeiro vamos calcular as mudanças de peso (Setembro-Agosto):

1.9 0.7 2.2 - 0.1 2.0 1.0 - 0.8 - 0.2 1.8 0.3

Seja μ a mudança média de peso na população. Então nossa hipótese nula H_0 e a **hipótese alternativa** H_1 podem ser escritas como segue:

$$H_0 : \mu = 0, \quad H_1 : \mu \neq 0.$$

Um procedimento útil é calcular um intervalo de confiança para a média populacional μ , e verificar se o intervalo inclui 0 como um valor plausível.

Alternativamente, pode-se proceder da seguinte forma:

Denotando por x as diferenças de peso e $n = 10$ tem-se que $\bar{x} = 0.88$ e $s = 1.065$, então o erro padrão da diferença de peso média é

$$SE = s/\sqrt{n} = 1.065/\sqrt{10} = 0.337,$$

e um valor- t de 2.262 é obtido da coluna $P = 0.05$ e linha $r = n - 1 = 9$.

Um intervalo de confiança de 95% para μ é portanto

$$(0.88 - 2.262 \times 0.337, 0.88 + 2.262 \times 0.337) = (0.12, 1.64).$$

O intervalo não contém o valor 0, fornecendo evidências contra a hipótese nula.

Podemos dizer que existem evidências significativas ($P < 0.05$) de que, em média, os pássaros da espécie estudada mudam de peso de Agosto para Setembro; ou que estamos 95% confiantes de que em média os pesos aumentam por um montante entre 0.12 e 1.64 gramas.

Mas e o intervalo de 99%? Será que ele conteria o valor 0? Este intervalo seria mais amplo e então é mais provável que ele contenha 0. Se ele não incluir 0, isto indicaria uma evidência ainda mais forte contra H_0 .

Calculando o intervalo de confiança exatamente da mesma forma, exceto que desta vez precisamos olhar na coluna $P = 0.01$ para obter $t = 3.250$:

$$(0.88 - 3.250 \times 0.337, 0.88 + 3.250 \times 0.337) = (-0.21, 1.97).$$

Como esperado, este é mais amplo, e agora inclui o valor 0.

Podemos agora dizer: “não existem evidências significativas ao nível de 1% de que, em média, os pássaros da espécie estudada mudam de peso de Agosto para Setembro.”

O que nós acabamos de fazer foi conduzir um teste perfeitamente válido para a hipótese nula usando intervalos de confiança. Podemos fazer o teste mais rapidamente e obter exatamente as mesmas conclusões pelo seguinte procedimento:

- Calcule $t = (\bar{x} - 0)/SE = 0.88/0.337 = 2.61$ (o número de erros padrão que \bar{x} dista de 0).
- Compare este valor de t com aqueles na linha $r = n - 1 = 9$ da tabela.
- Para este exemplo, $t = 2.61$ está entre os valores nas colunas $p = 0.01$ e $p = 0.05$. Então nosso valor deve corresponder a um p entre estes e portanto devemos ter $0.01 < p < 0.05$.

O valor de p é interpretado como a probabilidade de observar um valor de t mais extremo do que o observado quando $\mu = 0$. É uma medida análoga à proporção de pessoas sadias que são erroneamente diagnosticadas como doentes num exame de laboratório, ou seja, uma medida de falsos positivos.

7.2.1 Procedimento geral de teste

1. Estabeleça a **hipótese nula**, H_0 e a hipótese alternativa H_1 .
2. Decida qual o **teste** a ser usado, checando se este é válido para o seu problema.
3. Calcule a **estatística de teste**, T .
4. Encontre a probabilidade (**p -valor**) de observar um valor tão extremo ou maior do que T se a hipótese nula é de fato verdadeira. Você precisará se referir aos *valores críticos* nas tabelas estatísticas as quais fornecem p -valores correspondendo aos valores das estatística de teste.
5. Avalie a força da evidência contra H_0 . (Quanto menor p -valor, tanto mais evidência contra a hipótese nula.) Se necessário, decida se esta é evidência suficiente para **rejeitar** (ou **não rejeitar**) a hipótese nula.
6. Estabeleça as **conclusões e interpretação** dos resultados.

O p -valor é a probabilidade de observar dados tão extremos quanto os obtidos caso a hipótese nula seja verdadeira.

Note as seguintes interpretações de p -valores:

| | |
|---------------|---|
| $P \geq 0.10$ | Não existe evidência contra H_0 |
| $P < 0.10$ | Fraca evidência contra H_0 |
| $P < 0.05$ | Evidência significativa ... |
| $P < 0.01$ | Evidência altamente significativa ... |
| $P < 0.001$ | Evidência muito altamente significativa ... |

Esteja ciente da diferença entre significância estatística e significância prática.

Um efeito pode ser estatisticamente significante mas não ter qualquer importância prática e vice-versa.

Por exemplo, um estudo muito grande pode estimar a diferença entre a média de peso de plantas como sendo 0.0001 gramas e concluir que a diferença é estatisticamente significativa ($p < 0.05$). Contudo, na prática, esta diferença é negligível e provavelmente de pouca importância prática.

7.2.2 Teste para uma média

No início deste capítulo conduzimos, através de um exemplo, o chamado **teste-t** para uma única média. Os passos principais de tal teste-t para uma amostra aleatória x_1, x_2, \dots, x_n de uma população com média μ são dados a seguir:

1. Estabeleça a hipótese nula, $H_0 : \mu = \mu_0$, e a hipótese alternativa $H_1 : \mu \neq \mu_0$.
2. Calcule a média amostral $\hat{\mu} = \bar{x}$ e o desvio padrão amostral s .
3. Calcule o erro padrão, $SE = s/\sqrt{n}$.
4. Calcule a estatística de teste $t = (\hat{\mu} - \mu_0)/SE$. Este é o número de erros padrão que $\hat{\mu}$ dista do valor de hipótese μ_0 .
5. Encontre o p -valor da distribuição t , com $r = n - 1$ graus de liberdade, da tabela usando os valores absolutos da estatística de teste.
6. Estabeleça conclusões e interprete os resultados.

7.2.3 Teste para uma proporção

Agora suponha que tenhamos um valor hipotético p_0 para uma proporção. Podemos realizar um teste de $H_0 : p = p_0$ praticamente da mesma forma que o teste-t acima. A dualidade com intervalos de confiança segue exatamente da mesma forma.

Suponha que tenhamos uma amostra aleatória de tamanho n de uma população de interesse onde a verdadeira proporção de membros numa categoria em particular é p . A hipótese nula é $H_0 : p = p_0$. Se o número observado na categoria de interesse é x , então um teste da hipótese é como segue:

1. Estabeleça a hipótese nula, $H_0 : p = p_0$, e a hipótese alternativa $H_1 : p \neq p_0$.
2. Calcule a proporção amostral $\hat{p} = x/n$.
3. Calcule o erro padrão, $SE = \sqrt{\hat{p}(1 - \hat{p})/n}$.
4. Calcule $t = (\hat{p} - p_0)/SE$, o número de erros padrão que \hat{p} dista do valor de hipótese p_0 .
5. Encontre o p -valor usando o valor absoluto da estatística de teste da tabela da distribuição normal (ou equivalentemente da t com $r = \infty$ graus de liberdade).

Uma regra geral é que este teste é válido quando quando temos ambos $n\hat{p}$ e $n(1 - \hat{p})$ maiores do que digamos 10.

Exemplo:

Suponha que alguém tenha sugerido de experiências passadas que 60% das larvas de mosquito num certo lago deveriam ser da espécie *Aedes detritus*. Foram encontrados 60 desse tipo de uma amostra de 80. Os dados suportam esta hipótese?

Exercício

1. Um amigo sugere que você lance uma moeda para ajudar você a tomar uma decisão muito importante, o resultado também o afetará. Seu amigo sugere que você escolha cara para tomar a decisão A, e coroa para tomar a decisão B a qual é a preferida por ele. O único problema é que seu amigo insiste que você use uma moeda “da sorte” dele. Você fica um pouco suspeito e decide fazer um experimento enquanto seu amigo não está olhando. Você lança a moeda 40 vezes e cara aparece somente 13 vezes. Realize um teste estatístico para ajudá-lo na decisão se você deve ou não acreditar que a moeda é balanceada. Qual a sua conclusão?
2. Suponha que estejamos interessados em estimar a proporção de todos os motoristas que excedem o limite máximo de velocidade num trecho da rodovia entre Curitiba-São Paulo. Quão grande deve ser a amostra para que estejamos pelo menos 99% confiantes de que o erro de nossa estimativa, a proporção amostral, seja no máximo 0.04?
3. Refaça o exercício anterior, sabendo que temos boas razões para acreditar que a proporção que estamos tentando estimar é no mínimo 0.65.

8 Comparando dois grupos

Uma questão importante que surge no trabalho de pesquisa na área biológica é a comparação de drogas, de métodos cirúrgicos, de condições experimentais, de procedimentos de laboratórios, de dietas ou, em geral, de tratamentos.

Um caso especial que ocorre frequentemente é o da comparação de **dois** tratamentos. O objetivo pode ser o de se estabelecer a superioridade de um tratamento ou a equivalência entre eles.

A escolha entre dois tratamentos é menos simples do que em princípio parece.

Isto porque os seres vivos geralmente reagem de forma diferente a um tratamento. O resultado de um tratamento pode variar enormemente de indivíduo para indivíduo. Como não se conhece a priori a reação de cada indivíduo, em geral, considera-se como tratamento mais eficiente aquele que *na média* fornece os melhores resultados.

Em outras palavras, a situação ideal da escolha do melhor tratamento para cada indivíduo não é possível na prática.

Consequentemente, considera-se como o melhor tratamento aquele que produz bons resultados para a grande maioria da população em estudo.

8.1 Diferença entre médias de dois grupos

No capítulo anterior vimos como construir um intervalo de confiança para a média populacional μ , de uma amostra aleatória de tamanho n .

Lembre-se que este intervalo de confiança era da forma $\bar{x} \pm t \times SE$ or $(\bar{x} - t \times SE, \bar{x} + t \times SE)$.

Agora consideremos a comparação das médias de duas populações através da estimação da **diferença das médias** e calculando intervalos de confiança e testes de hipóteses para estas diferenças.

8.2 Amostras pareadas

Num estudo pareado, temos duas amostras mas cada observação da primeira amostra é pareada com uma observação da segunda amostra.

Tal delineamento ocorre, por exemplo, num estudo de medidas feitas antes e depois no mesmo indivíduo ou num estudo de gêmeos (em que cada conjunto de gêmeos forma um dado pareado).

Como esperado, as duas observações do mesmo indivíduo (ou de um conjunto de gêmeos) são mais prováveis de serem similares, e portanto não são considerados estatisticamente independentes.

Com dados pareados, podemos usar a seguinte notação:

$$\begin{aligned}x_{1i} &= \text{medida 1 do par } i, \\x_{2i} &= \text{medida 2 do par } i\end{aligned}$$

a então escrevemos as diferenças nas medidas de cada par como

$$d_i = x_{2i} - x_{1i}.$$

Agora temos **uma amostra de diferenças** d_i , e podemos usar os métodos para uma única amostra que já estamos familiares.

Podemos calcular um intervalo de confiança para a **diferença média** e testar se a diferença média é igual a zero ou não.

Nos referimos a tal teste como um **t-test pareado** ao contrário do **test-t para duas amostras independentes** que veremos a seguir.

Note que neste caso estamos interessados na **diferença média** enquanto que quando temos duas amostras independentes, estaremos interessados na **diferença das médias**.

Ainda que numericamente estas quantidades sejam as mesmas, conceitualmente elas são diferentes.

Exemplo: A mudança nos valores de imc de indivíduos do início ao final de seis meses tratamento foram:

$$-1.5 \quad -0.6 \quad -0.3 \quad 0.2 \quad -2.0 \quad -1.2$$

A média e o desvio padrão são -0.9 e 0.81 , respectivamente. Então o erro padrão é $0.81/\sqrt{6} = 0.33$.

Podemos agora realizar um test- t pareado para testar a hipótese nula de que a perda média de imc é 0. Para isso calculamos

$$t = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{-0.9}{0.33} = -2.73.$$

Note que este valor é negativo (porque a mudança média observada foi a redução no imc — um valor positivo seria um aumento no imc).

Observamos o valor absoluto da estatística de teste (2.73) na tabela, usando a linha com $n - 1 = 5$ graus de liberdade.

A quinta linha da tabela mostra que $0.01 < p < 0.05$ (porque o valor 2.73 está entre os valores tabelados 2.571 e 4.032). Então, rejeitamos a hipótese nula ao nível de 5%.

Podemos concluir que existem evidências ao nível de 5% de que há uma redução média de imc durante o período de seis meses em indivíduos sujeitos ao tratamento.

Podemos adicionar à nossa conclusão o intervalo de confiança de 95% para a redução média no imc:

$$-0.9 \pm 2.57 \times 0.33 = -0.9 \pm 0.85 = (-1.75, -0.05)$$

Estamos 95% confiantes que a redução média de imc está entre 0.05 e 1.75.

Suposições feitas: a distribuição das mudanças de imc não é muito diferente de uma Normal.

8.3 Amostras independentes

Quando temos **amostras independentes** de cada uma de duas populações, podemos sumariá-las pelas suas médias, desvios padrão e tamanhos amostrais.

Denote estas medidas por \bar{x}_1, s_1, n_1 para a amostra 1 e \bar{x}_2, s_2, n_2 para a amostra 2.

Denote as correspondentes médias populacionais e desvios padrão μ_1, μ_2, σ_1 e σ_2 respectivamente.

Para os dados de alturas dos estudantes da UFPR, vamos comparar a altura média dos estudantes do sexo masculino com as dos sexo feminino.

Seja o grupo dos homens a amostra 1, e o grupo das mulheres a amostra 2.

As alturas foram medidas em centímetros e as medidas sumárias foram como segue:

$$\begin{aligned}\bar{x}_1 &= 178.85, & s_1 &= 7.734, & n_1 &= 20, \\ \bar{x}_2 &= 164.09, & s_2 &= 9.750, & n_2 &= 17.\end{aligned}$$

Agora claramente uma estimativa natural da diferença entre médias na população, $\mu_1 - \mu_2$, é dada pela **diferença nas médias amostrais**:

$$\bar{x}_1 - \bar{x}_2,$$

e para nossos dados esta é $178.85 - 164.09 = 14.76$.

Agora o que precisamos é um erro padrão para esta estimativa para que possamos construir um intervalo de confiança ou realizar um teste da hipótese nula $H_0: \mu_1 - \mu_2 = 0$ versus $H_1: \mu_1 - \mu_2 \neq 0$.

O cálculo do erro padrão de $\bar{X}_1 - \bar{X}_2$ depende da suposição feita a respeito dos desvios padrão de cada grupo de comparação.

Uma regra prática é assumir que os desvios padrão populacionais σ_1 e σ_2 são iguais se a *razão* do maior desvio padrão amostral para o menor for menor do que 2 ou 3.

Além disso a suposição de variâncias iguais pode ser grosseiramente avaliada através de histogramas dos dados.

Testes formais estão disponíveis se necessário. Um deles é o teste F para igualdade de variâncias de Levene cuja hipótese nula é a de que $\sigma_1 = \sigma_2$.

8.3.1 Erro padrão - assumindo desvios padrão iguais

Primeiramente, assumimos que os desvios padrão populacionais são os mesmos em cada grupo, i.e. $\sigma_1 = \sigma_2 = \sigma$.

Podemos combinar os dois desvios padrões amostrais para formar uma estimativa combinada do desvio padrão.

Atribuímos mais peso às amostras maiores. Este **desvio padrão combinado** s_p é a raiz quadrada da variância combinada s_p^2 dada por

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Para nossos dados temos:

$$s_p^2 = (19 \times 7.734^2 + 16 \times 9.750^2)/35 = 75.92801$$

então $s_p = \sqrt{75.92801} = 8.71$.

Note que s_p está entre s_1 e s_2 como esperado. *Se você obtiver um valor que não está entre estes valores então seus cálculos estão errados!*

Agora podemos calcular o **erro padrão das diferenças nas médias** como

$$SE = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

a qual para nossos dados é $8.71 \times \sqrt{(1/20 + 1/17)} = 2.87$.

8.3.2 I.C. para a diferença entre médias assumindo desvios padrão iguais

Um intervalo de confiança para $\mu_1 - \mu_2$ é dado por

$$((\bar{x}_1 - \bar{x}_2) - t \times SE, (\bar{x}_1 - \bar{x}_2) + t \times SE),$$

em que t é escolhido apropriadamente.

Quando os tamanhos amostrais são grandes um intervalo de confiança aproximado de 95% é obtido usando $t = 1.96$.

Se os tamanhos amostrais não forem tão grandes então um intervalo exato de 95% de confiança deveria de ser calculado selecionando o valor de t da tabela da distribuição t , com $n_1 + n_2 - 2$ graus de liberdade e coluna $p = 0.05$.

Para um intervalo de 99% de confiança deveríamos selecionar o valor na coluna $p = 0.01$.

Exemplo: Para os dados de altura, temos $n_1 + n_2 - 2 = 20 + 17 - 2 = 35$, resultando $t = 2.03$ para um intervalo de confiança de 95% (através de interpolação entre a linha 30 e 40). Um intervalo de confiança de 95% para a diferença nas médias é dado por:

$$(14.76 - 2.03 \times 2.87, 14.76 + 2.03 \times 2.87) = (8.93, 20.59).$$

Estamos 95% confiantes que, em média, estudantes do sexo masculino são entre 9cm e 21cm mais altos do que as estudantes do sexo feminino.

8.3.3 Teste para a diferença das médias

Um teste para a diferença entre médias corresponde a um teste de $H_0: \mu_1 - \mu_2 = 0$. Seguindo o mesmo tipo de procedimento visto para uma única amostra.

Nosso **teste estatístico** é:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE},$$

que é a estimativa de $\mu_1 - \mu_2$ menos o valor hipotético (zero neste caso) e tudo dividido pelo erro padrão.

Sob a hipótese nula, este segue uma distribuição t com $n_1 + n_2 - 2$ g.l.

O valor obtido para t (ignorando seu sinal) é comparado com os valores tabelados com os graus de liberdade apropriados, para obter um p -valor.

Para os nossos dados, temos $t = (14.76 - 0)/2.87 = 5.14$, e comparando este à linha 30 e 40 da tabela, vemos que devemos ter $p < 0.001$.

Assumindo que nossas amostras foram amostras aleatórias de todos os estudantes, temos fortes evidências de que a altura média dos estudantes do sexo masculino é diferente da altura média dos estudantes do sexo feminino.

Suposições feitas: alturas dos estudantes tem uma distribuição razoavelmente simétrica, não muito diferente de uma Normal em cada grupo, e que os desvios padrão das duas distribuições são iguais.

8.3.4 I.C. para diferença de médias - desvios padrão diferentes

Se os desvios padrão populacionais *não puderem ser assumidos iguais*, usamos uma outra fórmula para o erro padrão de $\bar{x}_1 - \bar{x}_2$, dado por

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Note que esta abordagem é usada somente para **grandes amostras**.

A estatística de teste usando este SE *não* segue uma distribuição t sob a hipótese nula. Contudo, para tamanhos amostrais razoavelmente grandes (digamos ambos maiores do que 30), podemos comparar a estatística de teste acima com uma distribuição Normal padrão (última linha da tabela t).

Em nosso exemplo, calculamos um erro padrão de 2.87 sob a suposição de igualdade de desvios padrão populacionais para ambos os grupos.

A fórmula alternativa (a qual não assume desvios padrão populacionais iguais) resulta em

$$SE = \sqrt{\frac{(7.734)^2}{20} + \frac{(9.750)^2}{17}} = 2.93 \text{ kg}$$

que praticamente não difere do valor prévio.

Então o intervalo de confiança e o resultado de teste de hipótese seriam virtualmente os mesmos usando este erro padrão.

8.4 Comparando proporções

Um estudo investigando a existência de uma igualdade na proporção de machos de uma certa espécie em dois lagos distintos resultou em proporções observadas de machos de 74.4% dentre 43 peixes capturados no primeiro lago e 60% dentre os 50 do segundo.

Se construirmos intervalos de confiança para os percentuais correspondentes de machos na população (peixes da mesma espécie naqueles dois lagos), encontraríamos que podemos estar 95% confiantes de que o percentual está entre 61.4% e 87.4% no primeiro lago, e entre 46.4% e 73.6% no segundo.

Contudo, neste tipo de experimento a idéia principal é **comparar diretamente** os dois lagos. Portanto gostaríamos de calcular um **intervalo de confiança de 95%** para a **diferença em proporções**.

Note contudo que isto é apropriado somente para grandes amostras, e desse modo quando a amostra é pequena devemos ser cautelosos para não super valorizar os resultados.

8.4.1 Intervalo de confiança para a diferença em proporções

Seja p_1 a verdadeira proporção populacional no grupo 1 (lago 1), se seja p_2 a proporção no grupo 2 (lago 2).

Estamos interessados na diferença em proporções,

$$p_2 - p_1.$$

Estimativas de p_1 e p_2 são dadas por

$$\hat{p}_1 = 0.744 \quad , \quad \hat{p}_2 = 0.600,$$

então uma estimativa da diferença em proporções é

$$\hat{p}_2 - \hat{p}_1 = 0.744 - 0.600 = 0.144$$

O **erro padrão** desta diferença é

$$SE = \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

Com isso podemos construir um intervalo de confiança da forma usual, ou seja

$$(\hat{p}_2 - \hat{p}_1) \pm 1.96 \times SE.$$

Então para os nossos dados temos

$$SE = \sqrt{\frac{0.744 \times (1 - 0.744)}{43} + \frac{0.600 \times (1 - 0.600)}{50}} = 0.096.$$

Portanto um intervalo de confiança aproximado de 95% para a diferença em proporções é dado por

$$0.144 \pm 1.96 \times 0.096, \text{ o qual é } (-0.044, 0.332), \text{ ou } (-4.4\%, 33.2\%).$$

Estamos 95% confiantes que a verdadeira diferença percentual entre as proporções de peixes machos nos dois lagos está entre -4.4% e 33.2%.

Note que de acordo com este intervalo o valor zero é um valor plausível para as diferenças nos percentuais, e portanto não existem evidências estatísticas de que o percentual de peixes do sexo masculino diferem nos dois lagos.

8.4.2 Teste para a diferença de duas proporções

Podemos testar a hipótese nula $H_0: p_2 - p_1 = 0$ versus a alternativa $H_1: p_2 - p_1 \neq 0$ usando a estatística

$$t = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{SE}$$

e comparando este valor com a tabela t com ∞ graus de liberdade.

8.5 Aviso

Os métodos descritos neste capítulo e no anterior assumem que o tamanho de amostra é grande o suficiente para que a distribuição das médias amostrais seja aproximadamente normal. Em geral, por "grande" entenda-se 30 ou mais.

Se o tamanho da amostra for muito pequeno, digamos menor do que 30, e a distribuição for muito diferente da normal, pode-se considerar um teste não-paramétrico que será tratado a seguir.

8.6 Testes Não-paramétricos

Os métodos acima são válidos na maioria das ocasiões, mas algumas vezes métodos alternativos são necessários.

Note que para amostras pequenas é necessário assumir que a distribuição populacional não é muito diferente de uma Normal. Em geral, isso não é um problema, mas em alguns casos isso pode ser.

Exemplos em que os testes t não são apropriados são aqueles nos quais:

1. a natureza dos dados implica inevitavelmente numa distribuição extremamente assimétrica ou;
2. os dados não estão numa escala numérica e não faz sentido calcular uma média.

Por exemplo, se tivéssemos um escore de dor variando de 1 a 20, a mediana teria uma interpretação mais coerente do que a média.

Estes métodos não-paramétricos não fazem suposições acerca da distribuição de onde vieram os dados. Eles se baseiam na ordem (postos, ranks) dos dados.

Embora este procedimento possa parecer melhor, estes métodos são muito menos poderosos, e invariavelmente não fornecem intervalos de confiança.

Então um conselho é utilizá-los quando as suposições dos outros métodos realmente não parecerem razoáveis.

8.6.1 Amostras independentes

Um biólogo deseja comparar o número médio de besouros capturados numa amostra de 8 armadilhas montadas numa certa floresta, com o obtido numa amostra de 7 armadilhas colocadas numa outra floresta.

As contagens individuais estão listadas abaixo (em ordem numérica):

| | | | | | | | | |
|-----------|---|----|----|----|----|----|----|----|
| Amostra 1 | 8 | 12 | 15 | 21 | 25 | 44 | 44 | 60 |
| Amostra 2 | 2 | 4 | 5 | 9 | 12 | 17 | 19 | |

Contagens pequenas frequentemente têm distribuições assimétricas, principalmente porque elas devem ser maiores do que zero. Por esta razão, é aconselhável usar um teste não-paramétrico neste caso.

Para comparar dois grupos independentes (ou não pareados) como estes utiliza-se o **teste U de Mann-Whitney**.

Note que as medianas são bem diferentes, mas existe uma certa superposição dos dados, então não é óbvio se existe uma diferença real entre os dois grupos, ou se isto poderia ter ocorrido meramente por acaso.

O teste de Mann-Whitney primeiro ordena os dados, ou seja, assinala números de 1 a 15 por ordem de tamanho a cada observação, tratando todos os dados como uma grande e única amostra.

Ele então soma os postos de cada grupo e os compara (com auxílio de uma tabela).

Quanto maior a diferença nas somas, maior evidência de que existe uma diferença nos tamanhos das observações nos dois grupos.

Usando a tabela adequada para o teste U de Mann-Whitney vemos que neste caso o p -valor é de 0,024. Este p -valor é pequeno então podemos concluir que existe uma diferença estatisticamente significativa nos dois grupos ao nível de 5%.

Portanto, parece existir uma diferença nos números de besouros dependendo do tipo de floresta, e parece existir mais besouros no primeiro tipo de floresta.

8.6.2 Amostras pareadas

Em centros de tratamento de esgoto, amostras podem ser coletadas de duas formas: uma única amostra diária de 2 lts ou amostras pequenas retiradas em 24-horas.

A primeira refere-se a coleta de uma única amostra de 2 lts no mesmo horário diariamente e a segunda baseia-se num esquema de amostragem de 24 horas que retira 1 litro a cada hora.

Um experimento foi conduzido num período de 6 dias registrando-se o número de cistos de *Giardia* por litro do material.

É de interesse saber se os dados fornecem evidência de que os dois modos de amostragem diferem.

| Dia | 1 | 2 | 3 | 4 | 5 | 6 |
|--------------------|-----|----|-----|-----|-----|-----|
| amostras únicas 2L | 100 | 95 | 120 | 175 | 635 | 510 |
| amostras 24-horas | 145 | 60 | 215 | 670 | 350 | 130 |

Agora podemos usar o teste *t* pareado, mas como a amostra é muito pequena e os números em cada grupo parecem muito assimétricos, indicando que as diferenças não estarão próximas de uma Normal, um teste não-paramétrico pode ser mais apropriado.

O teste mais apropriado neste caso é o chamado **teste Wilcoxon para dados pareados**.

A forma como ele é feito consiste em primeiro calcular as diferenças das duas medidas em cada par, e então essencialmente testar a hipótese nula de que a diferença mediana é zero.

As diferenças em valor absoluto (ou em módulo) são ordenadas, ou seja, são assinalados postos às diferenças de 1 a 6. Os postos das observações com diferenças positivas são somados, e os postos das diferenças negativas são somadas.

Quanto maior for a diferença entre estas somas, maior a evidência de que existe uma diferença entre os métodos de amostragem.

O *p*-valor do teste para os nossos dados é 0,917 (obtido de tabela adequada), uma probabilidade muito grande. Isto significa que os dados são consistentes com a hipótese de que não existe diferença nos métodos de amostragem.

Contudo, devemos notar que com tão poucas observações não é de se esperar que existam fortes evidências de uma diferença.

8.7 Exercícios

1. Experimento sobre o efeito do álcool na habilidade motora.

Dez indivíduos são testados duas vezes, uma depois de ter tomado dois drinks e uma depois de tomado dois copos de água.

Os dois testes foram realizados em dois dias diferentes para evitar influência do efeito do álcool. Metade dos indivíduos tomou a bebida alcoólica primeiro e a outra metade água.

Os escores dos 10 indivíduos são mostrados abaixo. Escores mais altos refletem uma melhor performance.

Deseja-se testar se a bebida alcoólica teve um efeito significativo com um nível de significância de 1%.

| | indivíduo | | | | | | | | | |
|--------|-----------|----|----|----|----|----|----|----|----|----|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| água | 16 | 15 | 11 | 20 | 19 | 14 | 13 | 15 | 14 | 16 |
| álcool | 13 | 13 | 12 | 16 | 16 | 11 | 10 | 15 | 9 | 16 |

- Uma droga bastante utilizada para induzir anestesia geral é o Halotano, poderoso anestésico de inalação, não inflamável e não explosivo, com um odor relativamente agradável. Pode ser administrado ao paciente com o mesmo equipamento usado para sua oxigenação. Após a inalação, a substância chega ao pulmão tornando possível a passagem para o estado anestésico mais rapidamente do que seria possível com drogas administradas de forma intravenosa.

Os efeitos colaterais, no entanto, incluem a depressão do sistema respiratório e cardiovascular, sensibilização a arritmias produzidas por adrenalina e eventualmente o desenvolvimento de lesão hepática. Alguns anestesistas acreditam que esses efeitos podem causar complicações em pacientes com problemas cardíacos e sugerem o uso da Morfina como um agente anestésico nesses pacientes devido ao seu pequeno efeito na atividade cardíaca.

Conahan et al. (1973) compararam esses dois agentes anestésicos em um grande número de pacientes submetidos a uma cirurgia de rotina para reparo ou substituição da válvula cardíaca. Para obter duas amostras comparáveis, os pacientes foram alocados aleatoriamente a cada tipo de anestesia (experimento clínico controlado ou aleatorizado).

A fim de estudar o efeito desses dois tipos de anestesia, os pesquisadores registraram diversas variáveis hemodinâmicas, tais como pressão sanguínea antes da indução da anestesia, após a anestesia mas antes da incisão, e em outros períodos importantes durante a operação.

A tabela a seguir mostra a pressão sanguínea média observada desde o início da anestesia até o tempo de incisão para 122 pacientes.

| Anestesia | |
|-----------|---------|
| Halotano | Morfina |

| | | |
|---------------|------|------|
| média | 66,9 | 73,2 |
| desvio padrão | 12,2 | 14,4 |
| n | 61 | 61 |

As diferenças observadas entre esses dois grupos de pacientes são consistentes com a hipótese de que o efeito do Halotano e da Morfina na pressão sanguínea é o mesmo?

3. Agora vamos comparar a mortalidade dos dois grupos. Dos 61 pacientes anestesiados com Halotano 8 (13,1%) morreram e 10 dos 61 pacientes (16,4%) anestesiados com Morfina morreram.

9 Associação, correlação e regressão

Nesta seção consideramos diferentes formas de avaliar associação entre variáveis dependendo do tipo destas variáveis.

Usualmente, ou temos **dados categóricos**, os quais podem ser apresentados em tabelas de contagens, ou **dados numéricos** com os quais podemos traçar gráficos de dispersão, calcular correlações e ajustar modelos de regressão linear.

Ilustraremos a maioria destas idéias usando dados de pesos no nascimento.

9.1 Idéias básicas

A tabela abaixo mostra o número de mães que fumam e que não fumam para cada raça.

| | | Fumante? | | Percentual fumantes |
|------|--------|----------|-----|---------------------|
| | | Não | Sim | |
| Raça | Branca | 44 | 52 | 54% |
| | Negra | 16 | 10 | 38% |
| | Outra | 55 | 12 | 18% |

Existe evidência de uma relação entre raça e fumo das mães?

Parece que existe uma diferença entre raças, mas poderia esta ser devida simplesmente ao acaso?

Quão provável seria observar tais diferenças entre raças na amostra se de fato as proporções populacionais fossem as mesmas?

O gráfico abaixo mostra a relação entre peso da mãe e peso do bebê.

Existe alguma evidência de uma relação entre peso da mãe e peso de seu bebê? Se sim, assumindo que mães mais pesadas tendem a ter bebês mais pesados, quão mais pesados em média esperaríamos que fossem bebês de mães com peso 200lbs quando comparados bebês de mães pesando 100lbs?

Podemos construir um intervalo de 95% de confiança para o peso médio de bebês nascidos de mães pesando 200lbs?

Se a futura mamãe pesa 150lbs, qual seria nosso melhor palpite do peso do bebê?

Podemos construir um intervalo de predição de 95% para o qual estejamos 95% seguros de cobertura do peso ao nascer de um bebê de uma futura mamãe de 150lbs?

9.1.1 Associação não é causalidade

Se uma relação for encontrada entre duas variáveis, isto não significa que elas tem uma relação de causalidade.

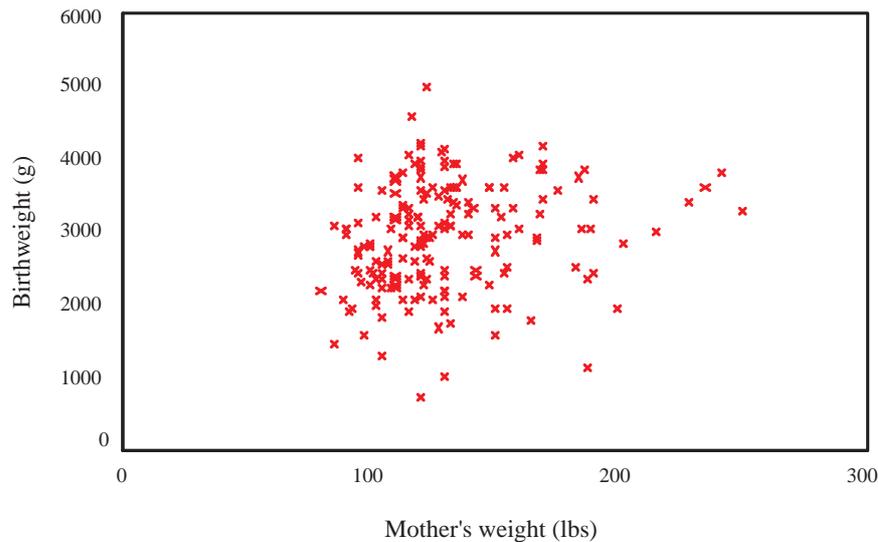


Figura 37: Gráfico de dispersão de peso da mãe e peso do bebê.

Por exemplo, uma relação razoavelmente forte é encontrada entre o número de doutores por pessoa na população e a expectativa média de vida.

É tentador pensar que isto pode ter uma relação de causalidade, mas na verdade a correlação é ainda maior entre expectativa de vida e número de aparelhos de TV por pessoa na população!

Um país com várias TVs é provavelmente um país próspero, com um padrão de vida razoavelmente alto e uma expectativa de vida mais longa.

Para estabelecer se uma variável tem um efeito causal sobre outra, necessitamos planejar um experimento, por exemplo, alocar aleatoriamente diferentes quantidades de fertilizante à plantas de tomate, e ver como a produção difere dependendo da quantidade de fertilizante.

9.1.2 Significância

Esteja atento ao fato de que associação/correlação/diferença **estatisticamente significativa** não implica necessariamente em **significância prática**.

Por exemplo, no gráfico de peso ao nascer versus peso da mãe, mesmo que pudessemos estar convencidos de que existe uma correlação estatisticamente diferente de zero, o gráfico indica que esta relação é muito fraca, e talvez de nenhuma significância prática.

Conhecer o peso da mãe ao nascer não nos permite de forma alguma prever o peso ao nascer do seu bebê de maneira precisa.

Quando existe uma grande quantidade de dados, é comum encontrar-se resultados altamente significantes, ou seja, p -valor quase zero, mesmo quando o desvio real da hipótese nula é muito pequena e de nenhuma importância prática.

Nestes casos, a construção de gráficos ou tabelas provavelmente dirão tudo o que se precisa saber.

9.2 Dados categóricos

Para verificar a significância estatística da aparente associação numa tabela de raça contra fumo, podemos conduzir o chamado **teste de associação de qui-quadrado** (χ^2).

A hipótese nula é de que não existe associação entre raça e fumo.

Quanta evidência existe contra esta hipótese em favor da alternativa de que existe uma associação?

Sabemos que o percentual total de fumantes é 39,2%.

Assumindo que a hipótese nula é correta então esperaríamos que o número de fumantes brancas seria 39,2% de 96, ou seja, 37,6.

Da mesma forma, podemos obter os números esperados para o resto da tabela:

| Esperado contagem | | Fumante? | | Total |
|----------------------|--------|----------|------|-------|
| | | Não | sim | |
| Raça | Branca | 58.4 | 37.6 | 96 |
| | Negra | 15.8 | 10.2 | 26 |
| | Outra | 40.8 | 26.2 | 67 |

A discrepância entre as contagens observadas e esperadas podem ser medidas com:

$$X^2 = \sum_k \frac{(O_k - E_k)^2}{E_k},$$

em que O_k é a contagem observada na casela k e E_k é a contagem esperada na casela k .

A soma é sobre todas as caselas na tabela. Valores grandes desta soma correspondem a maiores discrepâncias entre os valores observados e esperados, e portanto mais evidência contra a hipótese nula de não associação.

Para obter um p -valor, X^2 deveria ser comparada com a distribuição χ^2 com df graus de liberdade em que $df = (r - 1) \times (c - 1)$ com r o número de linhas na tabela, c o número de colunas na tabela. (Aqui $df = 2$.)

Neste caso, o p -valor é 0 com 3 casas decimais.

Concluimos que existe evidência estatística muito forte ($p < 0,001$) de uma associação entre raça e fumo.

A principal observação é que mulheres na categoria de raça **Outra** parecem ser muito menos prováveis de fumar durante a gravidez do que mães brancas ou negras.

Também parece que a proporção de mães brancas fumantes é maior do que a de mães negras.

9.2.1 Suposições

Por razões similares às daquelas dos testes t , necessitamos uma amostra grande o suficiente para que o teste χ^2 seja válido (ou seja, para resultar valores de p corretos).

Existem regras práticas para ajudar na decisão de o tamanho amostral é grande o suficiente:

1. 80% das contagens esperadas na tabela deveriam ser maiores do que 5 e;
2. todas as contagens esperadas devem ser maiores do que 1.

9.2.2 Tabelas 2x2

Frequentemente a tabela a ser analisada é uma simples tabela 2x2, ou seja, 2 linhas e 2 colunas.

Neste caso, o **teste exato de Fisher** também pode ser usado, o qual calcula um valor de p exato, baseado em todas as possíveis formas de alocação dos números numa tabela.

Isto não é uma tarefa fácil de executar manualmente, mas pode ser feita no computador.

Este teste não necessita de grandes contagens, sendo portanto útil para tabelas com contagens esperadas pequenas.

Uma correção chamada **correção de continuidade de Yates** deveria ser usada quando executando o teste χ^2 em tabelas 2x2. Isto implica em usar alternativamente:

$$X^2 = \sum_k \frac{(|O_k - E_k| - 0.5)^2}{E_k},$$

resultando num valor de X^2 menor do que a estatística sem a correção.

Esta correção, em geral, somente faz grande diferença na prática quando os valores esperados são pequenos, e neste caso o melhor mesmo é usar o teste exato de Fisher.

Nota: Em situações em que diversos testes são apropriados não é boa prática escolher o método que fornece o menor p -valor! É melhor definir antes qual teste será usado, e utilizar os resultados daquele teste. Se não houver uma escolha pré-definida e tem os resultados de vários, a opção mais segura é utilizar aquele que tiver o maior valor de p .

9.3 Correlação

Quando as duas variáveis são quantitativas, e podemos fazer um gráfico de dispersão, podemos medir associação calculando um **coeficiente de correlação**.

O mais comum é o **coeficiente de correlação de Pearson**, também conhecido como o coeficiente de correlação **produto de momentos**.

Uma versão alternativa não-paramétrica é o **coeficiente de correlação de postos de Spearman**, e discutiremos a seguir as circunstâncias nas quais este é preferível.

Quando o tipo de coeficiente de correlação não é especificado assume-se que é o de Pearson.

Ambos são denotados por r , e têm as seguintes propriedades:

- r varia entre -1 e $+1$
- $r = 0$ corresponde a não associação
- quanto maior o valor de $|r|$, mais forte a associação
- $r > 0$ corresponde a ambas variáveis crescendo juntas
- $r < 0$ corresponde a uma variável ficando menor à medida que a outra fica maior.

9.3.1 Coeficiente de Pearson

Sejam x_1, x_2, \dots, x_n os valores de um conjunto de medidas em indivíduos $i = 1, \dots, n$. No exemplo dos pesos no nascimento $n = 189$ e x_i representam os pesos das mães.

Sejam y_1, y_2, \dots, y_n as outras medidas correspondentes, ou seja, pesos dos bebês. Então x_1 é o peso da primeira mãe, e y_1 é o peso ao nascer de seu bebê.

O coeficiente de correlação de Pearson é definido como:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}.$$

Ele quantifica a força de **associação linear** entre duas variáveis, e portanto descreve quão bem uma linha reta se ajustaria através de nuvem de pontos.

Se os pontos caem exatamente sobre uma linha crescente então $r = 1$, e se eles caem exatamente sobre uma linha decrescente, $r = -1$.

Para a Figura 37, a correlação é 0,189, bem próxima de zero como esperado mas positiva, o que também parece consistente com o gráfico e o bom senso.

Podemos também fazer um teste da hipótese nula de não associação. Aqui obtém-se um p -valor de 0,011. Temos portanto evidência estatisticamente significativa ao nível de 5% de uma associação entre peso da mãe e o peso de nascimento do bebê.

Este é um exemplo em que há significância estatística, mas não muita associação na prática.

Exercício:

Dados de peso de peixe e comprimento de otólitos.

| | | | | | | | | | | |
|-------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| Otolith length x (mm) | 6.6 | 6.9 | 7.3 | 7.5 | 8.2 | 8.3 | 9.1 | 9.2 | 9.4 | 10.2 |
| Fish mass y (g) | 86 | 92 | 71 | 74 | 185 | 85 | 201 | 283 | 255 | 222 |

Construa um gráfico de dispersão e calcule o coeficiente de correlação.

Parece existir uma relação entre as variáveis?

Podemos confiar que exista mesmo uma associação com somente 10 valores?

Uma relação linear parece ser apropriada?

Suposições: o teste assume que uma ou ambas as variáveis são aproximadamente normais. Se os dados não parecem formar uma nuvem aproximadamente em forma de elipse, isto é evidência de não-normalidade e o valor de p não deveria ser usado.

9.3.2 Coeficiente de correlação de postos de Spearman

Nos casos em que os dados não formam uma nuvem comportada, com alguns pontos bem distantes dos demais, ou em que parece existir uma relação crescente ou decrescente num formato de curva, o coeficiente de correlação por postos de Spearman é mais apropriado.

Ele também pode ser usado quando os dados não pertencem à uma escala de medida padrão, mas existe uma ordenação clara, por exemplo, escores numa escala de 1 a 20.

Este é um método não-paramétrico que usa somente os postos, e não faz quaisquer suposições. Essencialmente tudo o que faz é calcular o coeficiente de correlação de Pearson nos postos. Uma fórmula que é relativamente fácil de usar é:

$$r = 1 - \frac{6 \sum_i d_i^2}{(n^3 - n)},$$

em que n é o número de pares (x_i, y_i) e

$$d_i = (\text{posto de } x_i \text{ dentre os valores de } x) - (\text{posto de } y_i \text{ nos valores de } y).$$

Note que se os postos de x se são exatamente iguais aos postos de y , então todos os d_i serão zero e r será 1.

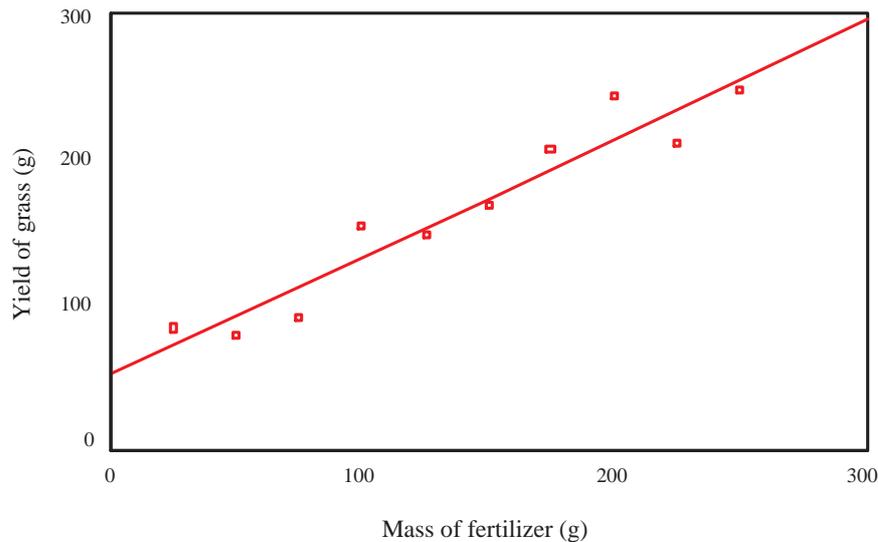
Os dados abaixo foram coletados tomando amostras de 13 nascentes de rios e é feita a contagem do número de ninfas de uma certa espécie de mosquito bem como medidas da dureza da água. Existe uma relação entre os dois?

| | | | | | | | | | | | | | |
|----------------|----|----|----|----|----|----|----|----|----|----|-----|-----|-----|
| dureza da água | 17 | 20 | 22 | 28 | 42 | 55 | 55 | 75 | 80 | 90 | 145 | 145 | 170 |
| No. de ninfas | 42 | 40 | 30 | 7 | 12 | 10 | 8 | 7 | 3 | 7 | 5 | 2 | 4 |

Um gráfico dos dados indica que existe uma relação negativa, mas uma linha curva descreveria melhor a relação do que uma reta.

O coeficiente de correlação de Pearson portanto não seria apropriado, e necessitamos usar o coeficiente de Spearman.

Encontre os postos manualmente e calcule as diferenças d_i . Calcula-se $\sum_i d_i^2 = 681$. Agora $n = 13$, a qual resulta no valor $r = -0.87$ para o coeficiente de correlação.



9.4 Regressão linear

Um pesquisador investiga o efeito de diferentes quantidades de fertilizante na produção de grama em solo calcário. Dez áreas de 1 m^2 foram escolhidos ao acaso, e diferentes quantidades do fertilizante foram aplicados em cada área.

Dois meses depois, as seguintes produções de grama foram obtidas:

| | | | | | | | | | | |
|---|----|----|----|-----|-----|-----|-----|-----|-----|-----|
| Massa de fertilizante (g/m^2) | 25 | 50 | 75 | 100 | 125 | 150 | 175 | 200 | 225 | 250 |
| Produção de grama (g/m^2) | 84 | 80 | 90 | 154 | 148 | 169 | 206 | 244 | 212 | 248 |

O gráfico de dispersão dos dados, com uma linha de melhor ajuste é mostrado abaixo.

Diferentemente dos dados de peso ao nascer vistos anteriormente, aqui se observa uma forte relação que segue claramente uma linha reta.

As questões que tínhamos acerca de predição para dados de peso ao nascer também são relevantes aqui.

Note que sempre colocamos a variável **resposta**, também chamada de variável **dependente**, aquilo que desejamos prever, no eixo vertical.

A variável **explanatória** ou variável **independente** vai no eixo horizontal.

Está claro que a linha ajusta-se bem, mas como ela foi escolhida?

A idéia básica é escolher a reta $y = a + bx$ que minimiza a soma de quadrados de desvios verticais dos pontos até a reta.

Denote os valores de fertilizante por x_1, x_2, \dots, x_n com $n = 10$, e os valores de produção de y_1, y_2, \dots, y_n . Se a e b são candidatos à intercepto e inclinação da reta, então $\hat{y}_i = a + bx_i$ é o valor ajustado para y_i dado por esta linha.

Queremos escolher a e b tais que \hat{y}_i é próximo de y_i para todo i , assim minimizamos a soma dos quadrados dos desvios:

$$\sum_i (y_i - \hat{y}_i)^2.$$

Existem fórmulas simples para as **estimativas de mínimos quadrados** de a e b que podem ser usadas para calculá-los manualmente, mas podemos obter os valores de programas estatísticos.

As estimativas do intercepto a e inclinação b da reta estão na tabela de **Coefficientes**. Neste caso encontramos que $\hat{a} = 51.93$ e $\hat{b} = 0.811$.

Informalmente podemos escrever:

$$\text{produção} = 51.93 + 0.811 \times \text{fert}$$

9.4.1 Testes e intervalos de confiança

Note que os erros padrão e os p -valores para os coeficientes também são mostrados nas saídas de programas estatísticos.

Os p -valores correspondem a testes da hipótese nula de que os valores verdadeiros de a e b na população são zero. O teste para o coeficiente de inclinação é em geral o único de interesse.

Aqui o p -valor é 0, a 3 casas decimais, então temos fortes evidências de um efeito de fertilizante na produção de grama. Podemos calcular um intervalo de confiança aproximado de 95% com sendo $(0.811 \pm 2 \times SE) = (0.811 \pm 2 \times 0.084) = (0.618; 1.004)$.

Estamos 95% confiantes de que a produção de grama aumenta entre 0.618g e 1.004g para cada extra grama do fertilizante em 1 m² de área da plantação.

Note que isto significa que podemos também dizer que estamos 95% confiantes de que o efeito de adição de 10g a mais de fertilizante é um aumento na produção em algo entre 6.18g e 10.04g.

9.4.2 R-quadrado

Note que os programas estatísticos também retornam um valor $R = 0.960$ e um $R^2 = 0.9224$ na tabela de **resumo do modelo**.

Na verdade este R é a correlação entre **produção** e **fertilizante**. (Cheque isto calculando a correlação separadamente.) **R-square** é o valor quadrático deste coeficiente de correlação, e tem uma interpretação muito interessante.

Ele representa a proporção da variabilidade na variável resposta explicada pela variável preditora ou variável explanatória. Também conhecido como **coeficiente de determinação**.

Ele nos dá uma idéia de quão bem podemos prever a variável resposta a partir da(s) variável(eis) preditora(s).

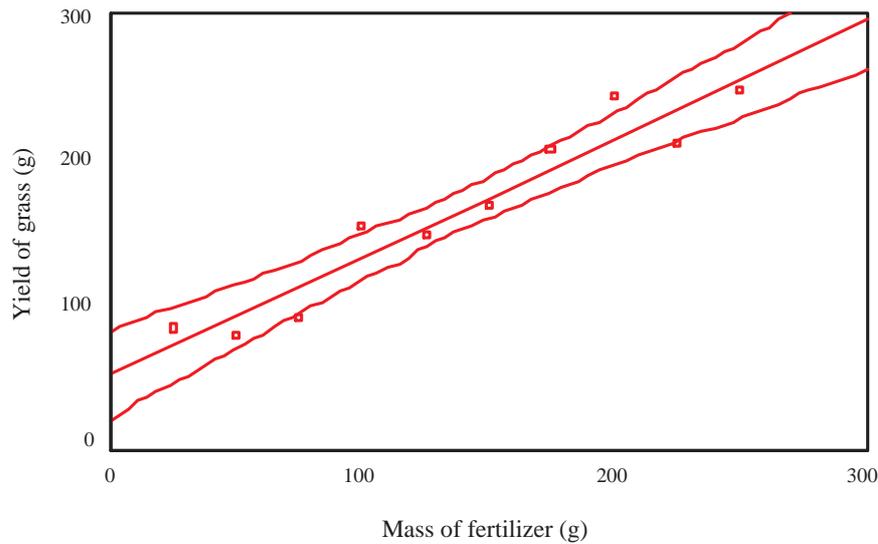


Figura 38: Intervalo de confiança de 95%.

Se os dados caem exatamente sobre a reta, $R^2 = 1$ e podemos prever a resposta exatamente.

9.4.3 Intervalos de confiança e de predição

Podemos obter intervalos de confiança para a média para qualquer quantidade de fertilizante.

As fórmulas para criar estes intervalos podem ser encontrados em diversos livros de Estatística, mas eles podem ser obtidos e adicionados aos gráficos usando qualquer programa estatístico.

Então para qualquer quantidade de fertilizante, podemos obter um intervalo de confiança de 95% para a produção média de grama (Figura 38).

Podemos também querer a produção de um novo talhão para o qual será aplicado 100g do fertilizante. Olhando o gráfico (Figura 39) podemos dizer que deverá algo entre 80g e 170g.

Podemos adicionar intervalos de predição de 95% para cada quantidade de fertilizante ao gráfico (Figura 39).

Note que os intervalos de predição são sempre mais amplos do que os intervalos de confiança para a média.

Ao obter mais dados, os intervalos de confiança para a média ficarão mais estreitos mas os intervalos de predição permanecerão próximos dos mesmos comprimentos.

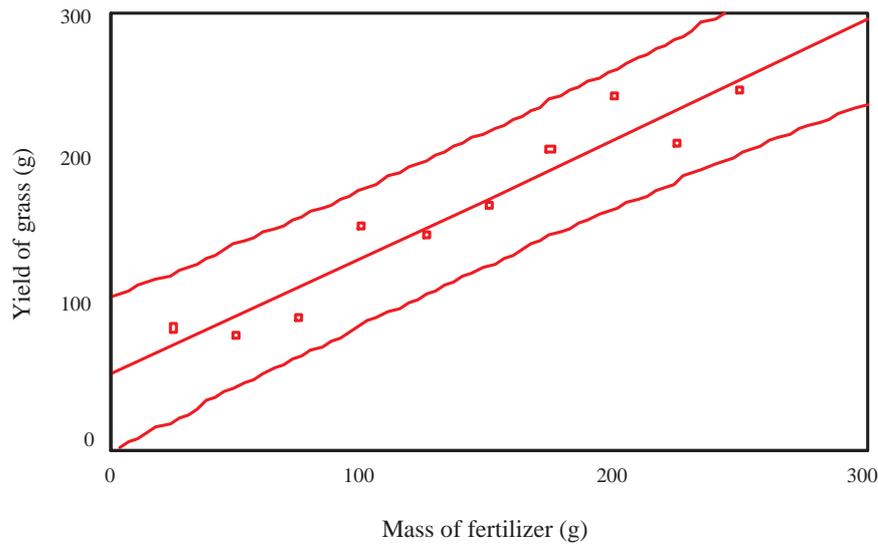


Figura 39: Intervalo de predição de 95%.

9.4.4 Suposições

Existem 3 suposições para a regressão, em ordem decrescente de importância:

- a relação é linear
- a **variabilidade** dos y valores é a mesma para todos os valores de x
- os valores de resposta são aproximadamente normalmente distribuídas para cada valor de x .

Quando algumas destas suposições não parecerem corresponder aos dados em mãos, então a regressão linear não é apropriada, no sentido de que os testes e intervalos de confiança não serão válidos.

Em algumas situações uma transformação dos dados pode ajudar, por exemplo, aplicar a transformação log a uma ou ambas variáveis.

O gráfico a seguir mostra um exemplo em que a transformação log ajuda.

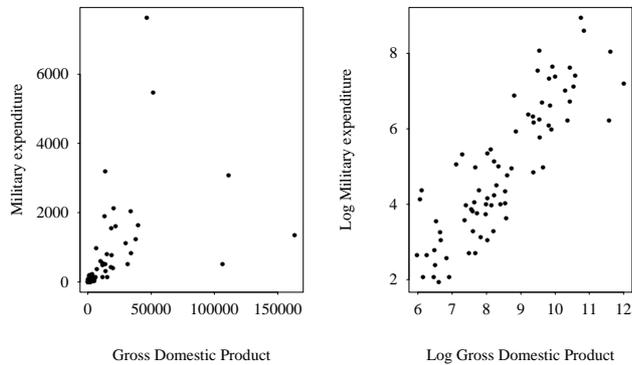
9.5 Regressão múltipla

Retornando aos dados de peso ao nascer, podemos ajustar um modelo de regressão linear que nos permita prever peso ao nascer a partir do peso da mãe.

Contudo, temos muito mais informações do que somente o peso da mãe.

Se nós realmente queremos prever peso ao nascer então seria sensato usar todos os dados que temos disponível.

Por exemplo, poderíamos tentar prever peso ao nascer usando a idade da mãe e seu estatus de fumo em adição a seu peso.



O procedimento é exatamente o mesmo de antes, exceto que agora o modelo ao invés de uma única variável explanatória `mwt`, ele terá `idade` bem como `fumo`.

O output fica parecido com o mostrado anteriormente, e obtemos a seguinte descrição informal:

$$\text{peso} = 2362.5 + 7.154 \times \text{idade} + 4.016 \times \text{mwt} - 269.3 \times \text{fumo}$$

Podemos também obter intervalos de confiança para os coeficientes da mesma forma como antes. O único problema é que porque existe mais do que uma variável preditora não é tão fácil de traçar gráficos dos dados.

A interpretação é que mães que fumam são mais prováveis de terem bebês pesando cerca de 269.3g a menos na média; o peso no nascimento parece aumentar cerca de 4.016g por lb de peso da mãe, e o peso no nascimento parece aumentar cerca de 7.154g por ano de idade da mãe. (*Repita estas conclusões usando intervalos de confiança.*)

Os testes e intervalos de confiança indicam que `idade` pode não ser uma variável preditora importante, e podemos ajustar o modelo novamente sem esta variável.

O **R-squared** tem a mesma interpretação como sendo a proporção da variância na resposta explicada pelas preditoras. (Aqui r é a correlação entre as respostas observadas e aquelas preditas pela equação do modelo.)

O valor de R-squared sempre aumenta à medida que mais variáveis explanatórias são acrescentadas no modelo, porque há sempre um ganho em poder de predição.

É importante ganhar um balanço entre ter um modelo complexo incluindo todas as possíveis preditoras, e um mais simples contendo somente as variáveis mais importantes. Na prática um modelo simples é frequentemente o melhor para predição.

Existem algumas técnicas para seleção de um subconjunto razoável de variáveis explanatórias, mas estas estão além do escopo deste curso.