Bioestatística

Lupércio F. Bessegato & Marcel T. Vieira

UFJF – Departamento de Estatística
2010

Roteiro

1. Análise Exploratória de Dados
2. Análise Univariada
3. Medidas-resumo
4. Análise Bivariada

Noções Básicas de Estatística

Dados

- Em geral, nas pesquisas da área coletam-se dados sobre:
  - √ Pessoas
  - √ Animais experimentais
  - √ Fenômenos físicos e químicos

Variabilidade

- Porque existe variabilidades nos fenômenos naturais?
- Fontes de variação:
  - √ Natural
  - √ Temporal
  - √ Erros de medida

Variabilidade Natural

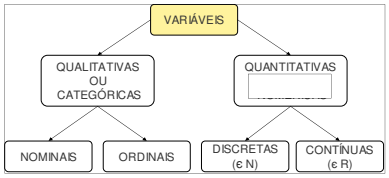
- Como ocorre?
  - √ Entre indivíduos
- Ocorre devido a diferenças de:
  - √ Idade
  - √ Sexo
  - √ Genética
  - √ Outros fatores que afetam a característica medida

Variabilidade Temporal

- Como ocorre?
  - √ No mesmo indivíduo
- Ocorre devido a diferenças:
  - √ Estado emocional
  - √ Idade
  - √ Clima
  - √ Outros fatores que mudam com o tempo a característica medida

Variável

- Característica de interesse qualquer, cujo valor pode variar em cada "indivíduo"
- Classificação:



Variável Categórica Nominal

- Valores distribuídos em categorias mutuamente exclusivos
- Exemplos:
  - √ Sexo
  - √ Causa da morte
  - √ Grupo sanguíneo

Variável Categórica Ordinal

- Valores distribuídos em categorias mutuamente exclusivas que possuem uma ordenação natural
- Exemplos:
  - √ Grau de instrução
  - √ Classe sócio-econômica

Variáveis Quantitativas Discretas

- Dado um valor, é possível estabelecer seu sucessor
- Em geral, são resultados de contagens
- Exemplos:
  - Qte. de bactérias em volume de urina
  - Qte. de batimentos cardíacos

Variáveis Quantitativas Contínuas

- Podem assumir qualquer valor em um intervalo contínuo
- Em geral, são resultados de medições:
  - √ Pressão sanguínea
  - √ Peso
  - √ Altura

Apuração dos Dados

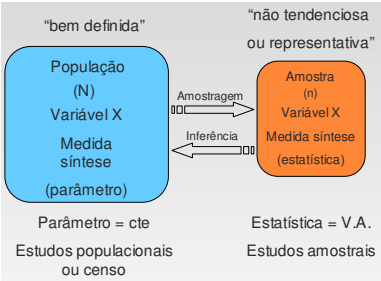
- Arquivo de dados (banco de dados):

Nº do Prontuário	Peso ao Nascer (Kg)	...
10525	3,25	...
10526	2,00	...
10527	3,20	...
⋮	⋮	⋮

- Apuração a partir de pacotes estatísticos

População e Amostra

- População:
  - √ Conjunto de elementos que apresentam pelo menos uma característica em comum
- População Alvo:
  - √ População de interesse da pesquisa
- Amostra:
  - √ Qualquer subconjunto não vazio da população



População e Amostra

- Recenseamento:
  - √ Procedimento de coleta de informações de toda a população
- Censo:
  - √ Conjunto de dados obtidos a partir do recenseamento
- Amostragem:
  - √ Procedimento de coleta de dados de uma amostra

Vantagens

- Principais vantagens dos levantamentos amostrais:
  - √ Menor custo
  - √ Maior velocidade

Comentários

1. Populações muito grandes podem ser estudadas apenas por amostras:
   - √ Número de doentes com HIV existentes no mundo
   - √ Pesquisas que avaliam novos medicamentos só podem ser feitas por amostragem
   - √ Nenhum investigador dispõe de todos os doentes do mundo!

2. Estudo cuidadoso de uma amostra tem mais **valor científico** do que o estudo superficial de toda uma população

Censo ou Amostragem

- Em igualdade de condições, o censo produz **resultados mais precisos**
- Com restrição orçamentária, uma amostra pode produzir resultados **mais informativos**

- Censo é recomendado quando:
  - √ População é pequena
  - √ Informações são baratas
  - √ É alto o custo de se tomar uma decisão errada
- Uma amostra deve ser sempre utilizada quando a população é grande e/ou os custos são altos

Técnicas de Amostragem

- Procedimento a ser adotado na seleção dos elementos da amostra
- O principal objetivo central é obter uma amostra representativa
  - √ Amostra que representa toda a população da melhor maneira possível
- A representatividade depende de:
  - √ Metodologia adotada para seleção da amostra
  - √ Tamanho da amostra

Análise Exploratória de Dados

O que é Análise Exploratória de Dados?

- Uma filosofia/abordagem para análise de dados
- Emprega uma variedade de técnicas (a maioria gráficas)…trabalharemos com alguns deles:
  - √ Diagrama de dispersão
  - √ Ramo e folhas (p/ conhecer)
  - √ Boxplot
  - √ Individual Plot

Técnicas que buscam:

- maximizar o "insight" do conjunto de dados;
- perceber a estrutura subjacente;
- extrair variáveis importantes;
- detectar valores atípicos (extremos) e anomalias;
- testar hipóteses fundamentais;
- desenvolver modelos parcimoniosos; e
- determinar conjunto ótimo de fatores

Idéia Básica

- Modelo = Suave + Irregular (tosco)
- Técnicas visuais podem frequentemente separar mais o "suave" do "irregular" ("ruído")

Clássica vs Exploratória

- Seqüência Clássica:
  √ Problema > Dados > Modelo> Análise > Conclusões
- Exploratória:
  √ Problema > Dados > Análise > Modelo > Conclusões

Tratamento de Dados

- Clássica:
  - √ Média e desvio padrão = estimativas pontuais
  - √ Medida de variabilidade explicada – r de Pearson
- Exploratória
  - √ Resumo Numérico (5): Min, Q1, Median, Q3, Max
  - √ todos (maioria) dados=resumos visuais
  - √ Dispersão
  - √ Histograma
  - √ boxplot

Análise Descritiva

- Inicia-se quase sempre pela verificação dos tipos disponíveis de variáveis
- Elas podem ser resumidas por tabelas, gráficos e/ou medidas

Objetivos

- Familiarização com os dados
- Detecção de estruturas interessantes
- Presença de valores atípicos (*outliers*)

Classificação

- Qualitativas (Categóricas)
  - √ Nominais:
  - √ Ordinais
- Quantitativas:
  - √ Discretas
  - √ Contínuas

Análise Univariada – Gráficos e Tabelas

Dados Brutos

- Obtidos diretamente de pesquisa
  - √ Sem qualquer processo de síntese ou análise
- Incluídos em tabelas
  - √ Não incluídos em publicações

Exemplo 3.1

- Teor de gordura fecal em crianças
  - √ Tamanho amostral: *43 crianças*

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

Tabela de Freqüências

- Considera-se a freqüência de ocorrência das observações
  - √ Caso discreto (ou categórico)
    - Conta-se a quantidade de vezes em que cada valor da variável ocorre

Exemplo 3.5

- Distribuição de profissões entre pacientes potencialmente suicidas, 2002

Profissão	Freqüência	Proporção
Serviços gerais *	75	
Doméstica **	55	
Do lar	53	
Indeterminada	29	
Emprego especializado ***	23	
Menor	20	
Desempregado	15	
Estudante	14	
Lavrador	12	
Autônomo	4	
Aposentado	2	
Total	302	

Exemplo – Vieira, 1998

- Internações em estabelecimentos de saúde, por espécie clínica

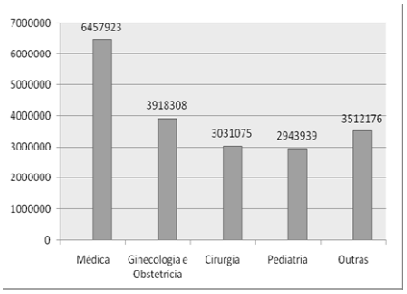
ESPÉCIE DE CLÍNICA	Freqüência Absoluta	Freqüência Relativa
Médica	6.457.923	32,51
Ginecologia/Obstetrícia	3.918.308	19,73
Cirurgia	3.031.075	15,26
Pediatria	2.943.939	14,82
Outras	3.512.176	17,69

Gráficos

- Objetivo:
  - √ Identificação da forma do conjunto de dados
  - √ Resumo e identificação
  - √ Padrão dos dados

Gráfico de Barras

- Para representação de variáveis **qualitativas**



- Caso variáveis contínuas (ou quantidade muito grande de valores de respostas)
  - √ Dados são agrupados em intervalos de valores das respostas (classes)

Exemplo 3.5 – Continuação

- Distribuição de Idade de Pacientes Potencialmente Suicidas, 2002

Idade	Freqüência		Freqüência Acumulada	
(anos)	Absoluta	Relativa	Absoluta	Relativa
10 ⊢ 20	57	18,87	57	18,87
20 ⊢ 30	113	37,42	170	56,29
30 ⊢ 40	59	19,54	229	75,83
40 ⊢ 50	32	10,60	261	86,42
50 ⊢ 60	19	6,29	280	92,72
60 ⊢ 70	7	2,32	287	95,03
≥ 70	2	0,66	289	95,70
Indeterminada	13	4,30	302	100,00
Total	302	100,00		

Construção de Tabelas

- Classes de um histograma :
  - √ Amplitude (h): comprimento do intervalo
  - √ Preferencialmente de mesmo tamanho
  - √ União dos intervalos incluem todos os dados (exaustivo)
  - √ Intervalos com intersecção vazia (exclusicos)
  - √ Extremos conhecidos como *limites de classe*

- Determinação quantidade de classes:
  - √ Não há uma quantidade 'ótima' de classes
  - √ Critério de Sturges: $k = \lceil 1 + \log_2 n \rceil$
  - √ Critério da raiz quadrada: $k = \lceil \sqrt{n} \rceil$
  - √ Outros critérios: Scott, Freedman–Diaconis
  - √ Em geral, *k* está entre 5 e 20 classes

- Determina-se o máximo e o mínimo dos dados
- Cálculo do intervalo de classes (*h*):

$$h = \frac{\max x_i - \min x_i}{k}$$

  - √ *h* pode ser modificado para facilitar construção e interpretação da tabela
  - √ Limite inferior da primeira classe deve ser menor que o mínimo valor dos dados
  - √ Limite superior da última classe deve ser maior que o máximo valor dos dados

- Freqüência absoluta (f_i), $i = 1, ..., k$:
  - √ Quantidade de elementos de cada classe
  - √ $\sum_{i=1}^{k} f_i = n$
- Freqüência relativa: $fr_i = \frac{f_i}{n}$
- Freqüência acumulada: $F_j = \sum_{i=1}^{j} f_i$
- Freqüência acumulada relativa: $Fr_j = \frac{F_j}{n}$

ESTATURA DAS MENINAS DESTA SALA - 2009

ESTATURAS (cm)	FREQUÊNCIA
150 ├ 154	4
154 ├ 158	9
158 ├ 162	11
162 ├ 166	8
166 ├ 170	5
170 ├ 174	3
TOTAL	40

FONTE: Novaes, 2009.

CLASSE → Li, Ls

h = Ls-Li

AT = Ls max-Li min

Ponto médio = (Ls –Li)/2

Exemplo

- Taxa de colesterol (mg/dL) – Tabela 3.2
  - √ Máximo: 479
  - √ Mínimo: 9 (não compareceram)
    Considerar este fato quando analisar-se a forma
- Quantidade de classes:

$$k = \lceil 1 + \log_2 n \rceil = \lceil 1 + \log_2 80 \rceil = \lceil 6{,}321928 \rceil = 7$$

$$k = \lceil \sqrt{n} \rceil = \lceil \sqrt{80} \rceil = \lceil 8.944272 \rceil = 9$$

- Adotado: $k = 9$

- Intervalo de classe (h):

$$h = \frac{\max x_i - \min x_i}{k} = \frac{479 - 9}{9} = 52{,}22222 \approx 50$$

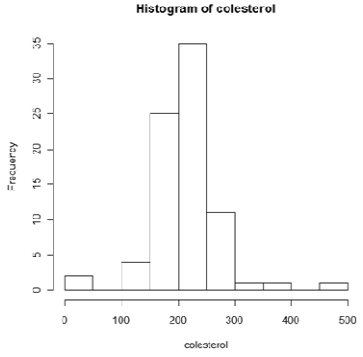
O valor adotado (50) leva a 10 classes

- Tabela de Freqüências

Colesterol	Freqüência		Freqüência Acumulada	
(mg/dL)	Absoluta	Relativa	Absoluta	Relativa
0 ⊢ 50	2	2,50	2	2,50
50 ⊢ 100	0	0,00	2	2,50
100 ⊢ 150	4	5,00	6	7,50
150 ⊢ 200	25	31,25	31	38,75
200 ⊢ 250	35	43,75	66	82,50
250 ⊢ 300	11	13,75	77	96,25
300 ⊢ 350	1	1,25	78	97,50
350 ⊢ 400	1	1,25	79	98,75
400 ⊢ 450	0	0,00	79	98,75
450 ⊢ 500	1	1,25	80	100,00
Total	80	100,00		

Histograma

- Gráfico de barras justapostas
  - √ Eixo horizontal:
    - Largura das barras (geralmente de mesmo tamanho)
    - Divisão em classes da variável
    - As barras são centradas no ponto médio das classes
  - √ Eixo vertical:
    - Altura da barra
    - Valor da freqüência absoluta (ou relativa)
- Utilizado apenas com variáveis contínuas



- Dados com assimetria
  - √ (algumas ocorrências de valores elevados)

Ramo-e-Folhas

- Gráfico simplificado da distribuição dos dados
- Cada valor da variável consiste em no mínimo dois dígitos
- Números são divididos em duas partes:
  - √ Ramo: Dígito inicial (um ou mais)
  - √ Folha: Dígitos restantes (resumindo em um único)

Ramo-e-Folhas – Colesterol

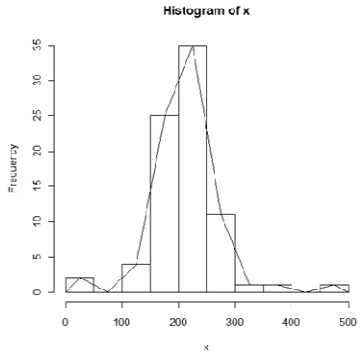
```
> stem(colesterol)

      The decimal point is 2 digit(s) to the right of the |

        0 | 11
        0 |
        1 | 2
        1 | 55567777778888888899999
        2 | 00000001111111112222223333333444444
        2 | 55555666678888899
        3 | 2
        3 | 6
        4 |
        4 | 8
```

- Escolher poucos ramos em comparação com a quantidade de observações
- Após definição do conjunto de ramos, os mesmos são listados na margem esquerda do diagrama
- Listam-se todas as folhas correspondentes a cada ramo
- Valores podem ser arrendondados (facilita interpretação

- Para comparar duas distribuições
  - √ Aproximadamente mesmo número de observações
  - √ Gráfico duplo em torno de ramo vertical comum

Polígono de Freqüências

- Construído a partir do histograma
- Segmentos de retas unindo as ordenadas dos pontos médios de cada classe
- Assim como o histograma, serve para visualização da forma da distribuição de freqüências da variável

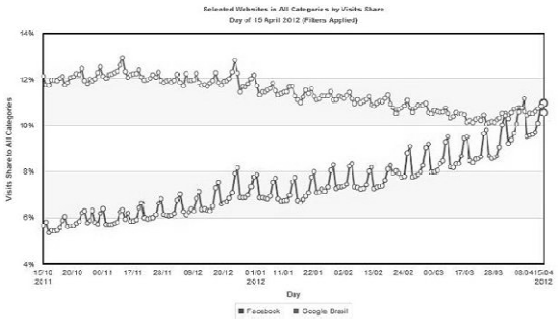


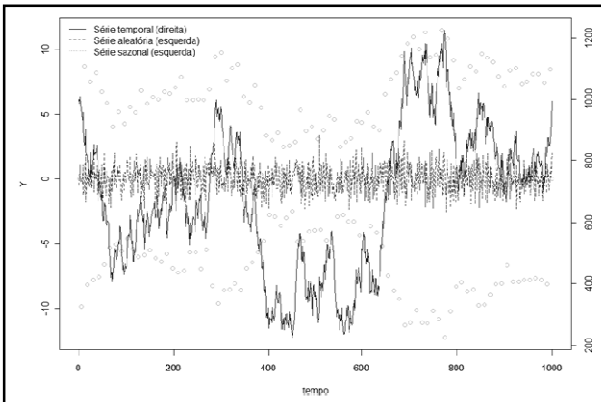
Séries Temporais

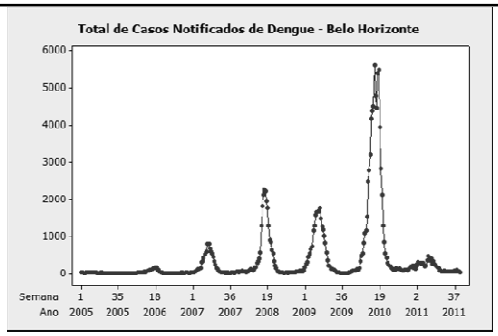
- Coleção de observações feitas sequencialmente ao longo do tempo
  - √ Em séries temporais a ordem dos dados é fundamental.
- Característica importante:
  - √ Observações vizinhas são dependentes
- Interesse: analisar e modelar esta dependência

Exemplo

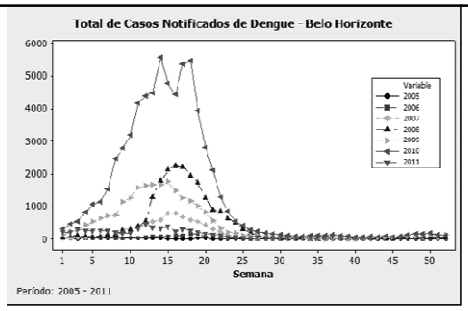
- Número de visitas ao Facebook e Google no período de 15/10/2011 a 15/04/2012







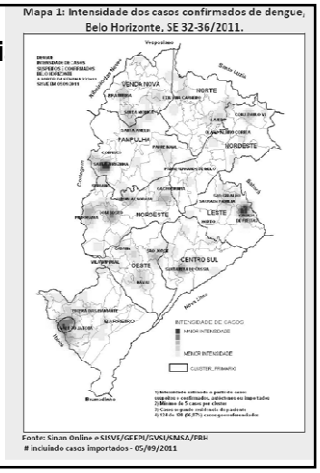
- Tendência
- Sazonalidade
- Estacionariedade



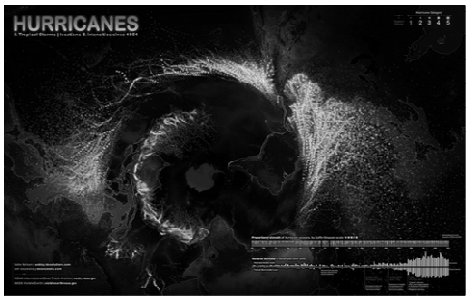
- Sazonalidade
- Variáveis que podem ajudar a explicar total de casos

- Observa-se algum padrão?

Representação Espaci



√ Um século e meio de furacões, tornados, ciclones e tufões

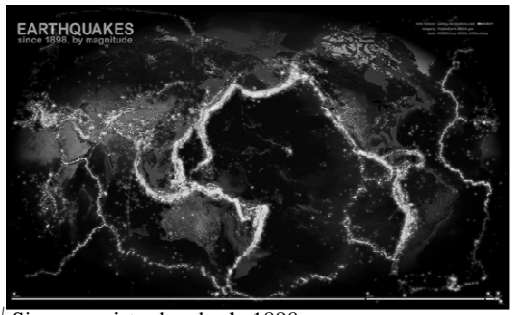


√ Centrado no Polo Sul

√ Dados entre 1851 e 2010

√ Trajetórias de furacões e tempestades tropicais

√ Todos os terremotos de pouco mais de um século



√ Sismos registrados desde 1898

√ Cada ponto luminoso é o epicentro.

- Quanto mais brilhante, mais tremores.
- Magnitudes separadas por cores (de 4 a 8 na escala Ritcher.)

Medidas Resumo

Medidas Resumo

- Medidas que sintetizam informações contidas nas variáveis em um único número
- Tipos:
  - √ Medidas de tendência central
  - √ Medidas de dispersão
  - √ Quartis, Decis e Percentis
  - √ Medidas de assimetria
  - √ Medidas de curtose

Medidas de Tendência Central

Medidas de Tendência Central

- Em geral, podem ser interpretadas como o ponto ao redor do qual os dados são distribuídos
- Algumas medidas de posição (tendência central):
  - √ Média
  - √ Mediana
  - √ Moda

Média

- Tendência central dos dados caracterizada pela média aritmética simples;
  - √ Média amostral
  - √ Média populacional

Média Amostral

- Os dados em geral são provenientes de uma amostra de observações de uma população
- Definição:

  Se n observações em uma amostra forem denotadas por x_1, x_2, ..., x_n, a média amostral será:

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

Exemplo – Colesterol

- Taxa de Colesterol (mg/dL)
- Tabela 3.2
- *n = 80* indivíduos
- Média amostral

$$\overline{x} = 215{,}9625 \approx 216{,}0$$

Média Populacional

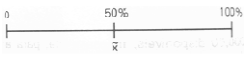
- Valor médio de todas as observações em uma população:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- A média amostral é um '***bom***' estimador da média populacional

Mediana

- Valor que divide a distribuição dos dados em duas partes iguais



- 50% das observações ficam acima da mediana e 50%, abaixo

Exemplo – Colesterol

- Taxa de Colesterol (mg/dL)
- Tabela 3.2
- *n = 80* indivíduos
- Valor médio entre o 40º e o 41º valor ordenado
- Mediana

$$\tilde{x} = 212{,}5$$

Procedimento

- Ordenar os dados
- Se n for ímpar:
  - √ A mediana é o valor do elemento central
  - √ Elemento de ordem $\frac{n+1}{2}$
- Se n for par:
  - √ A mediana é o valor médio entre os dois elementos centrais
  - √ Elementos de ordem $\frac{n}{2}$ e $\frac{n}{2}+1$

Média e Mediana

- Valores atípicos (muito grandes ou muito pequenos) causam grandes variações na média
- Em geral, a mediana não é afetada da mesma forma
- A mediana é uma medida mais robusta (menos afetada pro valores atípicos)

Média e Mediana – Influência Extremos

	Média	Mediana
Todos dados (80)	216,0	212,5
Exceto maior valor (79)	212,6	212
Exceto 2 maiores valores (78)	210,7	210,5
Exceto 3 maiores valores (77)	209,3	209,0

Média *vs* Mediana

Média

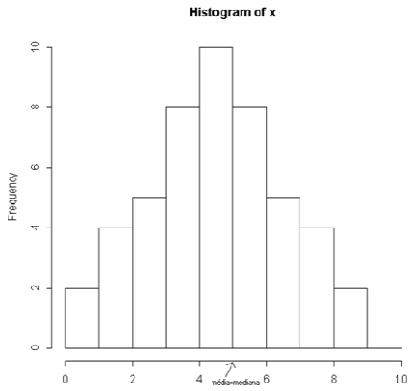
- fácil de ser manipulada algebricamente;
- representa o "centro de massa" dos dados (ponto de equilíbrio no histograma).
- afetada grandemente por valores extremos (ex.: islands).

Mediana

- difícil de ser manipulada algebricamente;
- valor da posição central dos dados ordenados;
- não é afetada por valores extremos.

Média *vs* Mediana (2)

- Para distribuições muito assimétricas, a mediana é uma medida mais apropriada para caracterizar um conjunto de dados.
- Se a distribuição é aproximadamente simétrica, então média e mediana são aproximadamente iguais.
  - √ Em distribuições perfeitamente simétricas média = mediana.



Média – Dados em Tabelas de Freqüência

- Para dados disponíveis apenas em tabela de freqüências
- Para calcular a média em tabela com *k* classes:

Ponto Médio	Freqüência
x_1	f_1
x_2	f_2
⋮	⋮
x_k	f_k

$$n = \sum_{i=1}^{k} f_i$$

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_k f_k}{n} = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i f_i$$

- Exemplo - Tabela de Freqüências – Colesterol

(mg/dL)	Ponto Médio (x_i)	Freq. Absoluta (f_i)	$x_i \cdot f_i$
0 ⊢ 50	25	2	50
50 ⊢ 100	75	0	0
100 ⊢ 150	125	4	500
150 ⊢ 200	175	25	4375
200 ⊢ 250	225	35	7875
250 ⊢ 300	275	11	3025
300 ⊢ 350	325	1	325
350 ⊢ 400	375	1	375
400 ⊢ 450	425	0	0
450 ⊢ 500	475	1	475
Total		80	17.000

$$\bar{x}_{Tab} = \frac{17.000}{80} = 212{,}5$$

$$\bar{x}_{Exata} = 215{,}9625 \approx 216{,}0$$

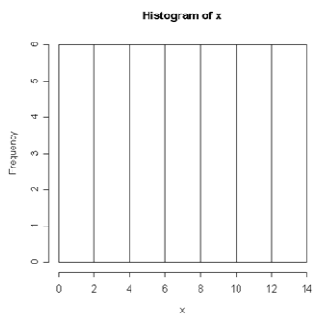
Moda

- É o valor mais freqüente da distribuição.
- No histograma, a classe modal é a classe de maior freqüência e a moda é aproximada pelo ponto médio da classe.

Moda (2)

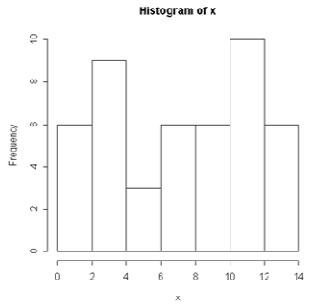
- Uma distribuição pode não possuir moda ("achatada").
- Uma distribuição pode possuir mais de uma moda (multimodal).
- Uma distribuição pode possuir apenas uma moda (unimodal).

Distribuição "Achatada"



voltar

Distribuição Multimodal



voltar

Medidas de Dispersão

Comparação entre Grupos de Dados

Stem-and-Leaf Display: grupo_1

```
Stem-and-leaf of grupo_1  N  = 10
Leaf Unit = 0,10

(10)  5  0000000000
```

Stem-and-Leaf Display: grupo_2

```
Stem-and-leaf of grupo_2  N  = 10
Leaf Unit = 0,10

 4  2  0000
 5  3  0
 5  4
 5  5
 5  6
 5  7  0
 4  8  0000
```

Stem-and-Leaf Display: grupo_3

```
Stem-and-leaf of grupo_3  N  = 10
Leaf Unit = 0,10

 3  4  000
(4)  5  0000
 3  6  000
```

Stem-and-Leaf Display: grupo_4

```
Stem-and-leaf of grupo_4  N  = 10
Leaf Unit = 0,10

 1  1  0
 2  2  0
 3  3  0
 4  4  0
(2)  5  00
 4  6  0
 3  7  0
 2  8  0
 1  9  0
```

Stem-and-Leaf Display: grupo_5

```
Stem-and-leaf of grupo_5  N  = 10
Leaf Unit = 0,10

 1  3  0
 3  4  00
(4)  5  0000
 3  6  00
 1  7  0
```

Média e Mediana

- Todos os conjuntos têm média e mediana iguais a 5
- Podemos afirmar que a distribuição dos dados é a mesma?

Comentários

- Há grandes diferenças entre os grupos;
  - √ Grupo 1: Todos os valores são iguais a 5.
  - √ Grupo 2: Nenhum valor igual a 5;
  - √ Grupo 3: Valores concentrados entre 4 e 6.
  - √ Grupo 4: Valores espalhados entre 1 e 9
  - √ Grupo 5: Valores dispersos entre 3 e 7
- Além da média e da mediana, é necessário outro tipo de medida para caracterizar os grupos

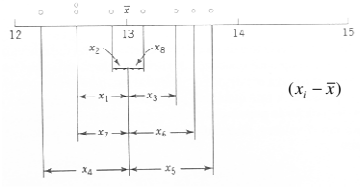
Medidas de Dispersão

- Informações importantes sobre os dados:
  - √ Valor em torno do qual os dados se **concentram**
  - √ Valor do grau de dispersão dos dados
- Medidas de dispersão mais comuns:
  - √ Amplitude amostral
  - √ Variância amostral (Desvio-padrão amostral)
  - √ Distância interquartílica

Amplitude Amostral - *r*

- É a mais simples das medidas de dispersão.
- É definida como: $r = \max(x_i) - \min(x_i)$
- Desvantagem:
  - ✓Omite toda a informação entre o mínimo e o máximo
  - ✓Em geral, quando $n < 10$, esta perda de informações não será muito séria

Construção de uma Medida de Dispersão



- Quanto maior a variabilidade dos dados, maior o valor absoluto de alguns desvios
- Valor absoluto complica o tratamento matemático
- A soma dos desvios é zero
- Uma solução: considerar o quadrado dos desvios

Variância Amostral

- É a média dos desvios quadráticos em relação à média.

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

- Tem unidade diferente dos dados.
- Por questões técnicas (Inferência), adota-se *n-1* no denominador da média.
  - ✓Torna-se o 'melhor' estimador

Desvio-padrão Amostral (*s*)

- É a raiz quadrada da variância amostral
  - √ A unidade de medida é a mesma dos dados

Coeficiente de Variação

- Medida relativa de dispersão: $cv = \frac{s}{\bar{x}} \cdot 100$
- Medida adimensional
- Fornece medida de homogeneidade dos dados
  - √ Quanto menor o *cv*, maior a homogeneidade
- Utilidades:
  - √ Comparação grau de concentração (dispersão) em torno da média
  - √ Comparação entre variáveis (ou grupos)

Exemplo – Colesterol

- Taxa de Colesterol (mg/dL) – Tabela 3.2
- *n = 80* indivíduos
- Variância: $s^2 = 3649{,}125$
- Desvio-padrão: $s = 60{,}40799$
- Média: $\bar{x} = 215{,}9625$
- Coeficiente de variação:

$$cv = \frac{s}{\bar{x}} = \frac{60{,}40799}{215{,}9625} = 27{,}97\%$$

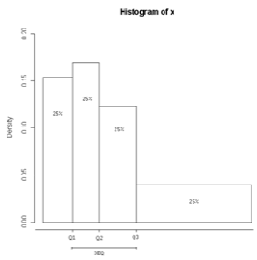
Média e Mediana – Influência Extremos

	Média	Mediana
Todos dados (80)	216,0	212,5
Exceto maior valor (79)	212,6	212
Exceto 2 maiores valores (78)	210,7	210,5
Exceto 3 maiores valores (77)	209,3	209,0

Quartis e Percentis

Quartis

- Dividem o conjunto de dados em 4 partes iguais



- 1º Quartil (Q_1):

  25% dos dados estão abaixo (75% acima)
- 3º Quartil (Q_3):

  75% dos dados estão abaixo (25% acima)
- 2º Quartil:

  É a mediana

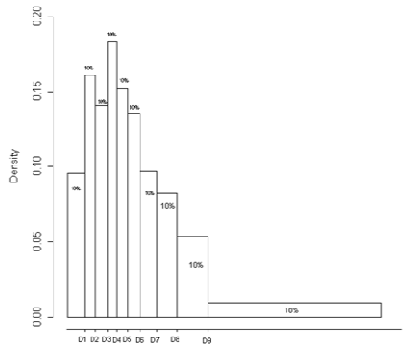
Exemplo – Colesterol

- Taxa de Colesterol (mg/dL) – Tabela 3.2

Mínimo	9,0	$x_{(1)} = 9$
Q_1	184,5	$x_{(20)}=184$ $x_{(21)}=185$
Q_2 = Mediana	212,5	$x_{(40)}=184$ $x_{(41)}=185$
Q_3	243,5	$x_{(60)}=243$ $x_{(61)}=244$
Máximo	479,0	$x_{(80)}=479$

Decis

- São 9 medidas que dividem a distribuição em 10 intervalos de mesma freqüência (10%):
  - √ D_1: primeiro decil → *q(0,10)*
  - √ D_2: segundo decil → *q(0,20)*
  - √ D_3: terceiro decil → *q(0,30)*
  - √ etc.



Exemplo – Colesterol

- Taxa de Colesterol (mg/dL) – Tabela 3.2

D_1	166,8	D_6	226,4
D_2	181,6	D_7	242,0
D_3	194,0	D_8	250,0
D_4	205,8	D_9	276,1
D_5	212,5		

Percentis

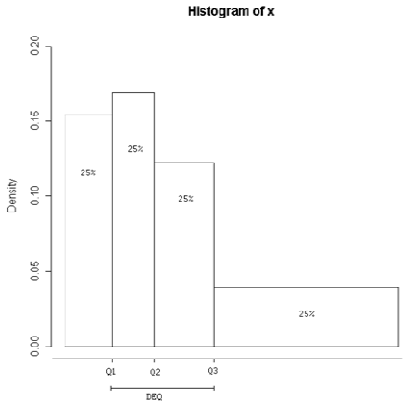
- São 99 medidas que dividem a distribuição em 100 intervalos de mesma freqüência (1%)
  - √ *q(0,01):* primeiro percentil;
  - √ *q(0,02):* segundo percentil;
  - √ *q(0,03):* terceiro percentil;
  - √ etc.

Distância Interquartílica

- Medida de variabilidade dada por .

$$DI = Q_3 \quad Q_1$$

- Menos sensível a valores extremos que a amplitude e a variância (desvio-padrão)
- É uma medida um pouco mais refinada que a amplitude amostral.

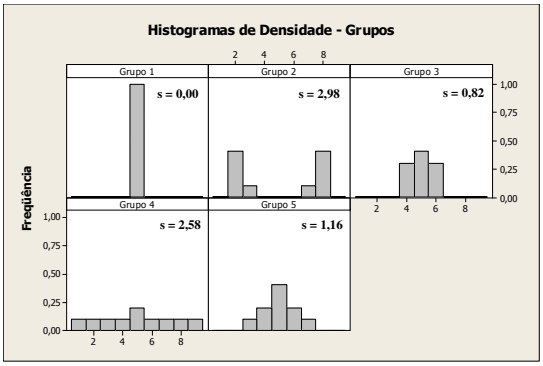


Grupos – Resumo

Grupo	R	DIQ	DMA	S^2	S
1	0,0	0,0	0,0	0,0	0,0
2	6,0	6,0 M	2,8 M	8,9	3,0 M
3	2,0 m	2,0 m	0,6 m	0,7	0,8 m
4	8,0 M	4,5	2,0	6,7	2,6
5	4,0	2,0 m	0,8	1,3	1,2

M **Máxima dispersão**

m **Mínima dispersão (excetuado grupo 1)**

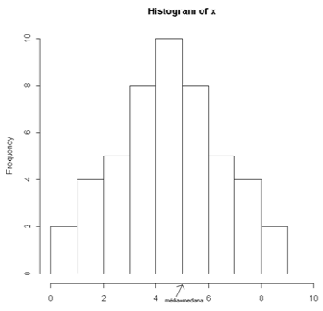


Simetria e Assimetria

Medida de Assimetria

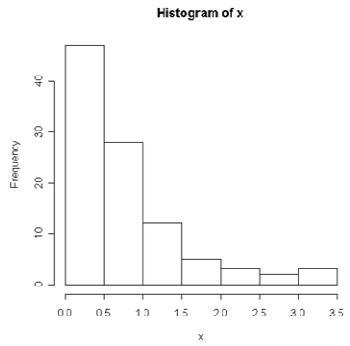
- Medida de assimetria indica se existem mais valores abaixo ou acima da média
- Se os valores se distribuem igualmente em torno da média:
  √ A distribuição é simétrica

Distribuição Unimodal – Simetria



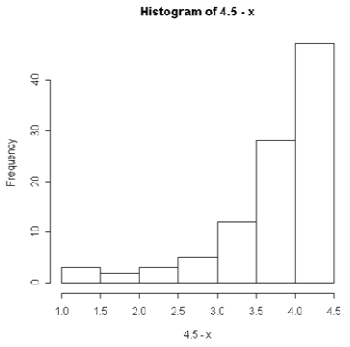
média = mediana = moda

Distribuição Unimodal – Assimetria Positiva



média > mediana > moda

Distribuição Unimodal – Assimetria Negativa



média < mediana < moda

Esquema dos 5 Números

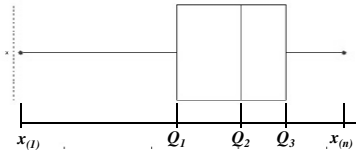
- São cinco valores importantes para se ter uma boa idéia da assimetria dos dados.
- São as seguintes medidas da distribuição:
- $x_{(1)}$, Q_1, Q_2, Q_3 e $x_{(n)}$.

Esquema dos 5 Números (2)

- Para uma aproximadamente simétrica, tem-se:
  - √ $Q_2 - x_{(1)} \cong x_{(n)} - Q_2$;
  - √ $Q_2 - Q_1 \cong Q_3 - Q_2$;
  - √ $Q_1 - x_{(1)} \cong x_{(n)} - Q_3$;
  - √ distâncias entre mediana e Q1, mediana e Q3 menores do que distâncias entre os extremos e Q1 e Q3.

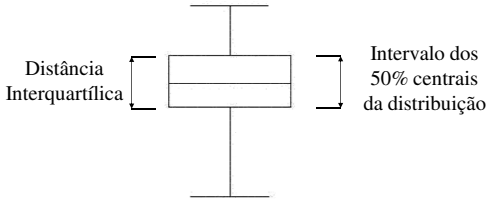
Box Plot

- A informação do esquema dos cinco números pode ser expressa num diagrama, conhecido como *box plot (gráfico-caixa).*
- Descreve várias características dos dados:
  - √ Centro, dispersão, simetria e valores atípicos



Box Plot (2)

- O retângulo é traçado de maneira que suas bases têm alturas correspondentes Q_1 e Q_3.
- Corta-se o retângulo por segmento paralelo às bases, na altura correspondente Q_2.
- O retângulo do *boxplot* correponde aos 50% valores centrais da distribuição.

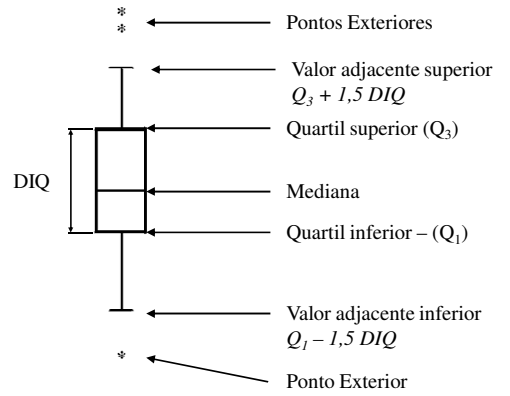


Região de Observações Típicas

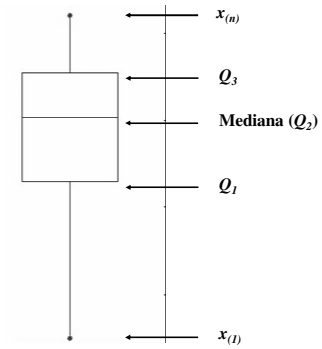
- Delimita-se a região que vai da base superior do retângulo até o maior valor observado que NÃO supere o valor de $Q_3+1,5 \times DIQ$.
- Procedimento similar para delimitar a região que vai da base inferior do retângulo, até o menor valor que NÃO é menor do que $Q_1-1,5 \times DIQ$.

Região de Observações Atípicas

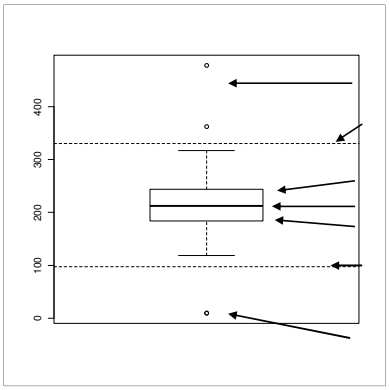
- Observações são representadas por asteríscos e situam-se:
  - √ ou, acima do Valor adjacente superior (*Q3 + 1,5 DIQ*)
  - √ ou, abaixo do Valor adjacente inferior (*Q1 – 1,5 DIQ*)
- Estes pontos exteriores são denominados *outliers* ou valores atípicos.



- Se não houver pontos exteriores:



Exemplo - Colesterol



DI = 58,5

Pontos Exteriores

Valor adjacente superior
$Q_3 + 1,5\ DIQ = 331$

Q3 = 243,25

Mediana = 212,5

Q_1 = 184,75

Valor adjacente inferior
$Q_1 - 1,5\ DIQ = 97$

Ponto Exterior

Análise Bivariada

Box-plot

- Pode ser utilizado para comparações entre diferentes grupos de dados
  - √ Variável quantitativa *vs.* variável categórica

Exemplo – Doenças Cardiovasculares

- Universo:
  - √ Homens doentes com idade entre 45 e 67 anos
- Amostra:
  - √ 100 casos coletados em 1969
- Variáveis
  - √ nível de glicose no sangue, em mg percentuais
  - √ atividade física em casa
    - 1 = sedentário; 2 = moderada; 3 = alta

Box-plot

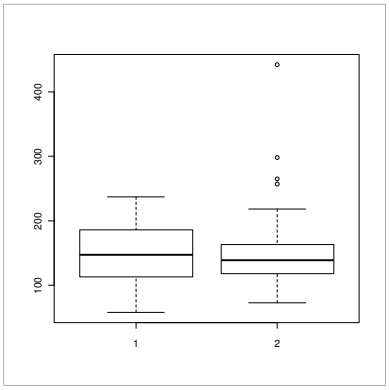


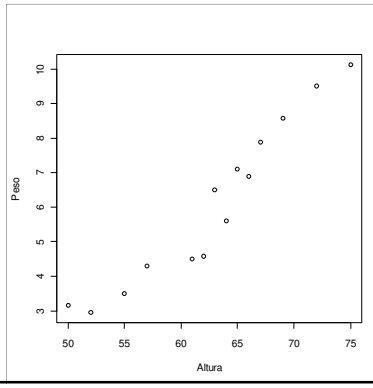
Diagrama de Dispersão

- Gráfico de pares ordenados por elementos da amostra (indivíduos)
- É a maneira mais simples de se estudar a relação entre duas variáveis quantitativas
- Objetivo:
  - ✓Ocorrência de tendências (lineares ou não)
  - ✓Agrupamentos de uma ou mais variáveis
  - ✓Mudanças de variabilidade de uma variável em relação à outra
  - ✓Ocorrência de valores atípicos (‘outliers’)

Exemplo

- Altura (cm) e peso (kg) de crianças até 1 ano

Altura	Peso
52	2,95
50	3,15
62	4,58
63	6,50
55	3,50
72	9,50
75	10,13
69	8,57
65	7,10
64	5,60
66	6,90
61	4,50
57	4,30
67	7,89

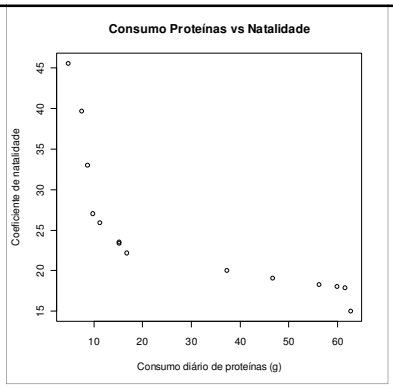


- Qual a relação entre o peso e a estatura das pessoas?
- Percebem-se 'clusters' no conjunto de dados?
- Há diferenças na variabilidade de uma variável, considerados os valores da outra?
- Há valores atípicos?

Relação entre consumo de proteínas e natalidade

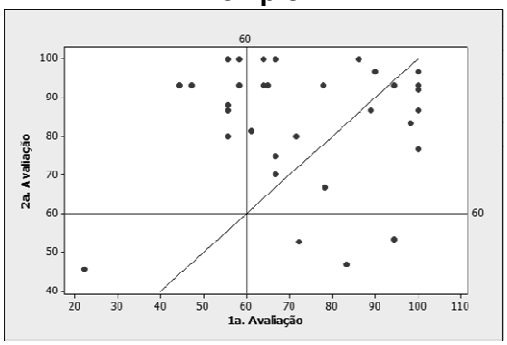
País	Consumo de Proteínas	Coeficiente de Natalidade
Formosa	4,7	45,6
Malásia	7,5	39,7
Índia	8,7	33,0
Japão	9,7	27,0
Iuguslávia	11,2	25,9
Grécia	15,2	23,5
Itália	15,2	23,4
Bulgária	16,8	22,2
Alemanha	37,3	20,0
Irlanda	46,7	19,1
Dinamarca	56,1	18,3
Austrália	59,9	18,0
Estados Unidos	61,4	17,9
Suécia	62,6	15,0

- Qual relação entre as variáveis?

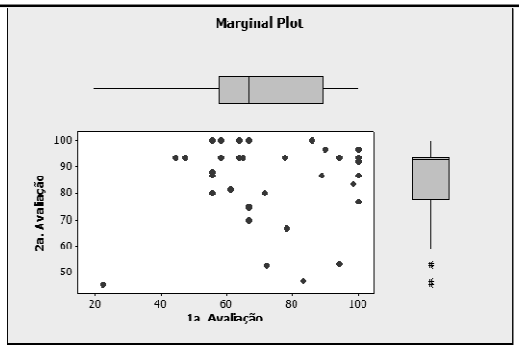


- Pode-se afirmar que há relação causal entre consumo de proteínas e natalidade ?
- Há indícios de clusters?

Exemplo



- Interpretação?



- Leituras gráficas

Correlação

- Correlação Positiva:
  - √ Se ambas as variáveis crescem no mesmo sentido
- Correlação Negativa:
  - √ Se as variáveis crescem em sentidos opostos
- Correlação significativa indica apenas associação entre as variáveis
  - √ NÃO INDICA RELAÇÃO DE CAUSALIDADE

Coeficiente de Correlação

- Como quantificar a correlação entre as variáveis?
  - √ Grau de associação

Coeficiente de Correlação de Pearson

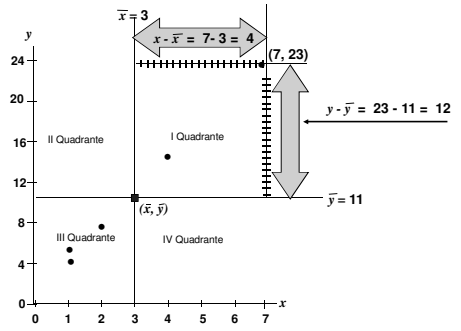
$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

- O numerador mede o total da concentração de pontos pelos quatro quadrantes
- Dá origem uma medida bastante usada

Justificação para a Fórmula de *r*

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$(\bar{x}, \bar{y})$ centróide da nuvem de dados



Notação

- x_i : i-ésimo valor observado da variável *x*
- y_i : i-ésimo valor observado da variável *y*
- $\bar{x}$: média dos valores observados da variável *x* (média amostral)
- $\bar{y}$: média dos valores observados da variável *y* (média amostral)

Soma de Quadrados – Notação

$$S_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2$$

$$S_{yy} = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n(\bar{y})^2$$

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n(\bar{x}.\bar{y})$$

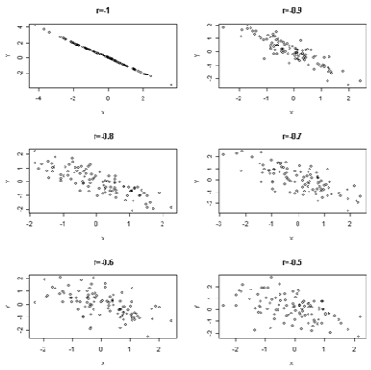
Propriedades de *r*

- Mede a intensidade de relacionamento <u>linear</u>
- *r* é adimensional e $-1 \leq r \leq 1$
  - √ r = 1 ou -1 ➔ correlação linear perfeita
  - √ r = 0 ➔ correlação linear nula
- O valor de *r* não é afetado pela escolha de x ou y.

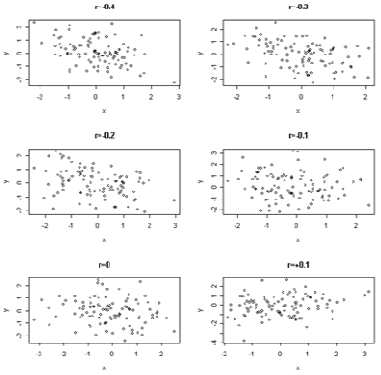
Propriedades de *r*

- A conversão da escala de qualquer das variáveis não altera o valor de *r*
- O valor de *r* não é alterado com a permutação de valores de x e y.
- Uma correlação baseada em médias de muitos elementos, em geral, é mais alta do que a correlação entre as mesmas variáveis baseada em dados para os elementos

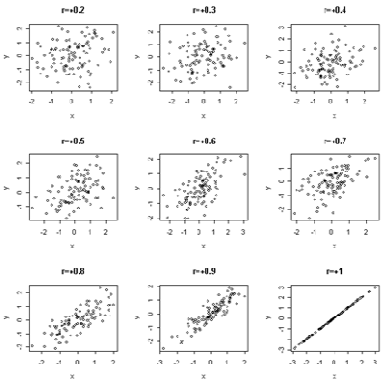
Diagramas de Dispersão (1)

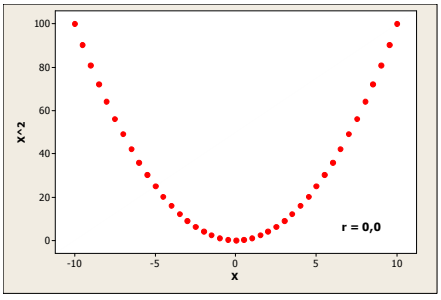


Diagramas de Dispersão (2)



Diagramas de Dispersão (3)





Existe uma relação de dependência NÃO –LINEAR entre as variáveis.

Exemplo 4 – Hábito de Fumar

- Dados sobre hábito de fumar entre homens e mortalidade por câncer de pulmão, na Inglaterra:
  - √ Dados distribuídos em 25 tipos de ocupação;
  - √ Variáveis:
    - Grupo: grupo de ocupação
    - Ifumo: índice de fumo
    - Imorte: índice de mortalidade

  Planilha: *fumo*

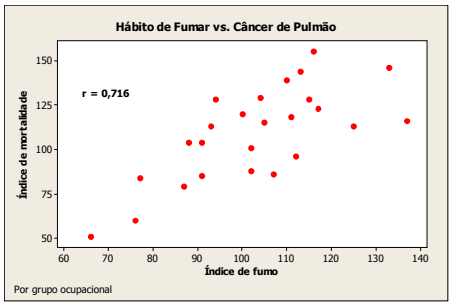
  Fonte: The Data and Story Library http://lib.stat.cmu.edu/DASL/

Exemplo 4 – Hábito de Fumar

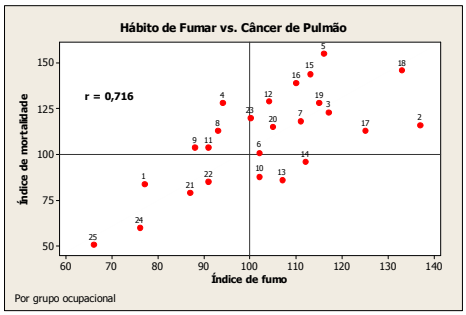
- ifumo: razão do número médio diário de cigarros fumados sobre a média global de cigarros.
  - √ Base: 100
  - √ ifumo = 100: número médio de cigarros por dia para o grupo é igual ao número médio global de cigarros fumados por dia
  - √ ifumo > 100: grupo fuma mais que o global
  - √ ifumo < 100: grupo fuma menos que o global

Exemplo 4 – Hábito de Fumar

- imorte: razão da taxa de mortes sobre a taxa global de mortes (por câncer de pulmão).
  - √ Base: 100
  - √ imorte = 100: número médio de mortes por câncer de pulmão para o grupo é igual ao número médio global de mortes por câncer de pulmão
  - √ imorte > 100: grupo com incidência de mortes por câncer de pulmão maior que o geral
  - √ imorte < 100: grupo com incidência de mortes por câncer de pulmão menor que o geral



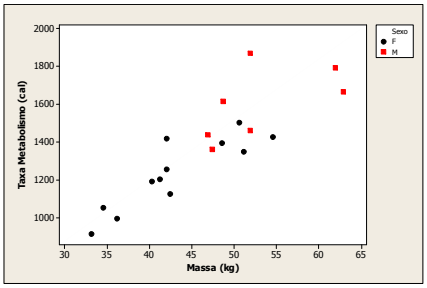
- Percebe-se uma correlação positiva entre as duas variáveis.



No contexto do exemplo faz sentido prever o índice de mortalidade por câncer de pulmão num particular grupo, dado o índice de fumo do grupo.

Exemplo

- Relação entre taxa de metabolismo e massa



- Evidências empíricas:
  - √ Associação linear e positiva
  - √ Associação mais forte entre a mulheres

- Cálculo correlações por grupos

```
MTB > corr c3 c4

Correlations: Massa; Taxa

Pearson correlation of Massa and Taxa = 0,865
P-Value = 0,000

MTB > corr c3 c4;
SUBC> by c2.

Correlations: Massa; Taxa

Results for Sexo = F

Pearson correlation of Massa and Taxa = 0,876
P-Value = 0,000

Results for Sexo = M

Pearson correlation of Massa and Taxa = 0,592
P-Value = 0,161
```

Não há evidências de correlação significativa entre os homens

- Valores médios dos grupos

```
MTB > describe c3 c4;
SUBC> by c2;
SUBC> stdev;
SUBC> mean.

Descriptive Statistics: Massa; Taxa

Variable  Sexo    Mean  StDev
Massa     F      43,03   6,87
          M      53,10   6,69

Taxa      F     1235,1  188,3
          M     1600,0  189,2
```

- Evidências empíricas:
  - ✓ Variabilidade semelhante entre os grupos;
  - ✓ Poucos homens com peso menor, poucas mulheres com peso maior
  - ✓ Possíveis influências na correlação:
    - Peso;
    - Sexo;
    - Variável não apresentada

Correlação – Erros Comuns

- Causalidade:

  Uma correlação forte (*r* vizinho de *+1* ou *–1*) não implica uma relação de causa e efeito.

  O fato de duas grandezas tenderem a variar no mesmo sentido não implica a presença de relacionamento causal entre elas.

Correlação e Causalidade

Perguntas pertinentes, no caso de correlação significante entre as variáveis:

- Há uma relação de causa e efeito entre as variáveis? (*x* causa *y*? ou vice-versa)

  Ex.: Relação entre gastos com propaganda e vendas

  É razoável concluir que mais propaganda resulta mais vendas

- É possível que a relação entre duas variáveis seja uma coincidência?

  Ex.: Obter uma correlação significante entre o número de espécies animais vivendo em determinada área e o número de pessoas com mais de 2 carros, não garante causalidade

  É bastante improvável que as variáveis estejam diretamente relacionadas.

- É possível que a relação das variáveis tenha sido causada por uma terceira variável (ou uma combinação de muitas outras variáveis)?

  Ex: Tempo dos vencedores das provas masculina e feminina dos 100 m rasos

  Os dados tem correlação linear positiva é duvidoso dizer que a diminuição no tempo masculino cause uma diminuição no tempo feminino;

  A relação deve depender de outras variáveis: técnica de treinamento, clima, etc.

Correlação e Causalidade

- A flutuação de uma 3ª variável faz com que *X* e *Y* variem no mesmo sentido;

  Esta 3ª variável é chamada variável intercorrente (não-conhecida);

  A falsa correlação originada pela 3ª variável é denominada correlação espúria;

Noções de Regressão

Regressão e Correlação

- Regressão:
- Usa variável(eis) explicativa(s) para explicar ou predizer comportamento de variável resposta (quando houver sentido).
- Correlação:
- Trata simetricamente duas variáveis

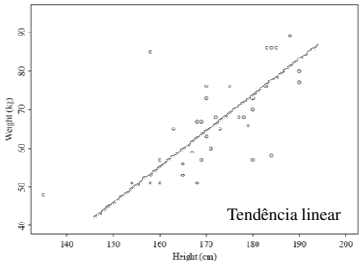
Regressão

- Variável resposta (Y):
- Variável resposta cujo comportamento se quer explicar
- Variável(eis) explicativa(s) (X_i):
- São de interesse caso ajudem a entender, explicar ou predizer o comportamento de Y.
- O enfoque da regressão é natural quando Y é aleatória e X_i é controlada ou não-aleatória.

x	Y
• Variável explicativa • Variável independente • Regressor • Preditor • Variável exógena • Variável de controle ou estímulos	• Variável explicada • Variável dependente • Regredido • Predito • Variável endógena • Variável resposta

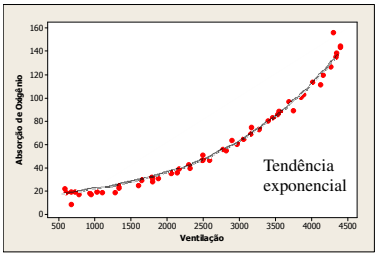
Exemplo 1 – Peso/Altura de Estudantes

- Variável resposta: Peso (kg)
- Variável explicativa: Altura (cm)



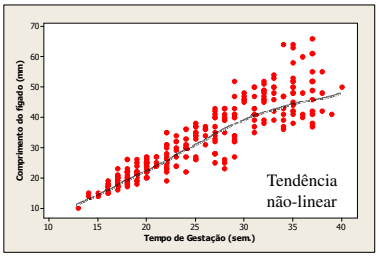
Exemplo 2 – Absorção de Oxigênio

- Variável resposta: Absorção de Oxigênio
- Variável explicativa: Ventilação

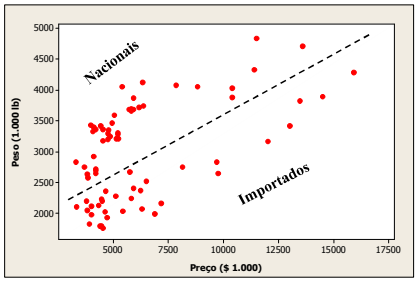


Exemplo 3 – Comprimentos de Fígados

- Variável resposta: Comprimento do fígado (mm)
- Variável explicativa: Tempo de gestação (sem.)

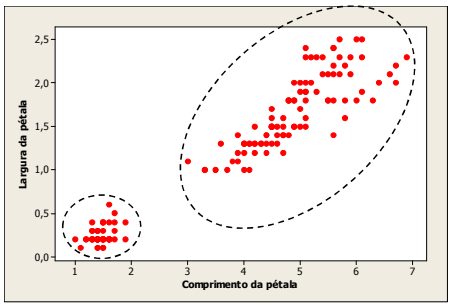


Outros Padrões (2)



Importante descobrir o que define os grupos

Outros Padrões (3)



Variedades diferentes de Flores

Modelo de Regressão

- Relação de regressão:
  - *Tendência + dispersão residual*
- Tendência:
  - √ Suavização dos dados
  - √ Explica a maior parte das diferenças de Y
- Valores atípicos:
- Observações muito diferente do restante dos dados

Relações Fortes e Fracas

- Relação Forte:
- A dispersão é pequena em relação à amplitude dos valores da curva de tendência
- Em dados observacionais, relações fortes não são necessariamente causais

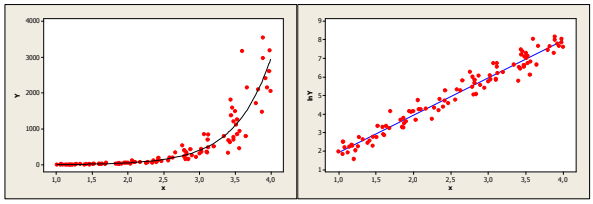
Resumo de Tendência – Abordagens

- Ajuste de funções matemáticas:
  - $Y = f(X)$
- Técnicas de suavização:
- *'Lowess'*, núcleo-estimador, *'spline'*

Ajuste de Funções

- Tendência linear: $Y = \beta_0 + \beta_1 X$
  - √ Para cada mudança de uma unidade em *X*, *Y* muda uma quantidade fixa.
- Tendência quadrática: $Y = \beta_0 + \beta_1 X + \beta_2 X^2$
  - √ Tendência levemente curva

- Tendência exponencial: $Y = \beta_0 e^{\beta_1 X}$



√ Cada mudança de uma unidade em *X*, *Y* muda uma % fixa

√ Se a tendência é exponencial, o gráfico de *log(Y)* vs *X* têm tendência linear

Tipos

- Simples:
  - √ Uma variável independente (explicativa)
- Múltipla:
  - √ Duas ou mais variáveis independentes

Objetivos

- Encontrar equação matemática que permita:
  - √ Descrever e compreender a relação entre 2 ou mais variáveis aleatórias
  - √ Projetar ou estimar uma nova observação
- Ajustar uma reta a partir dos dados amostrais

Utilidades

- Busca de relações de Causa e Efeito;
- Predição de valores;
- Estabelecer explicação sobre população a partir de uma amostra

Regressão Linear Simples

- Busca-se a equação de uma reta que permita:
  - √ Descrever e compreender a relação entre duas variáveis
  - √ Projetar e estimar uma das variáveis em função da outra.

Regressão Linear Simples (2)

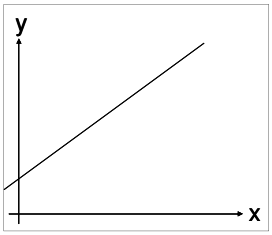
- A partir de valores observados de X e Y, modelar a tendência através de uma equação do tipo:

$$Y_i = \beta_0 + \beta_1 X_i$$

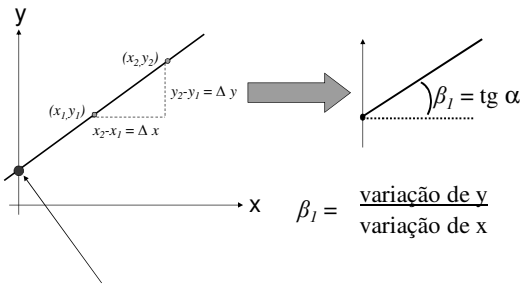
Função Linear

$$Y_i = \beta_0 + \beta_1 X_i$$

- f(x) se modifica a uma taxa constante em relação à sua variável independente
- β_0 e β_1 são constantes
- β_0 : intercepto-*y*
- β_1 : coeficiente angular



Intercepto e Coeficiente Angular



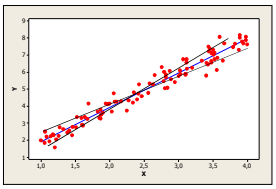
$\beta_1 =$ variação de y / variação de x

β_0: intersecção da reta com o eixo y

Interpretação dos Parâmetros

- β_1 : declividade da reta
- define o aumento ou diminuição da variável *Y* por unidade de variação de *X*
- β_0 = intercepto em y
- define o valor médio de *Y* sem a interferência de *X* (com *X=0*).

Ajuste da Reta



- Qual a reta que se ajusta melhor aos dados?
- ou seja quais os valores de β_0 e β_1?
- Escolher β_0 e β_0 de maneira a tornar mínima a distância entre a reta e os pontos

Método dos Mínimos Quadrados

- Critério:
- Valores dos parâmetros que minimizam a soma dos quadrados dos desvios

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

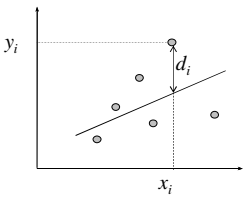
Método dos Mínimos Quadrados (2)

- Minimização em relação a β_0 e β_1 :

$$S = \sum d_i^2 = \sum \{Y_i - (\beta_0 + \beta_1 x_i)\}^2$$

$$\frac{\partial S}{\partial \beta_0} = 0$$

$$\frac{\partial S}{\partial \beta_1} = 0$$



Método dos Mínimos Quadrados (3)

- Resultados das derivadas parciais:

$$\hat{\beta}_1 = \frac{n.\sum(x_i y_i) - (\sum x_i).(\sum y_i)}{n.\sum x_i^2 - (\sum x_i)^2} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \frac{\sum y_i - b\sum x_i}{n} \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

- Calculando por medidas estatísticas :

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} = r_{XY}\frac{s_Y}{s_X}$$

Estimativa de Mínimos Quadros

- Dadas observações (X_1, Y_1), ... , (X_n, Y_n) os coeficientes da reta que melhor se ajusta aos dados são:

$$\beta_0 = \hat{\beta}_0 \quad \text{e} \quad \beta_1 = \hat{\beta}_1$$

- que são chamados estimativas de mínimos quadrados do intercepto e da declividade
- A reta de mínimos quadrados é dada por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Exemplo

- Hidrolisação de procaína no plasma humano em função do tempo decorrido após sua administração
- Variável resposta (Y)
  - √ Quantidade de procaína hidrolisada
- Variável explicativa:
  - √ Tempo decorrido após administração

Tempo (minutos)	Qte. Procaína Hidrolisada (10 moles/litro)
2	3,5
3	5,7
5	9,9
8	16,3
10	19,3
12	25,7
14	28,2
15	32,6

Qte. Procaína (10 moles/Litro)	Tempo (min.)			
Y	X	X^2	XY	Y^2
3,5	2	4,0	7,0	12,3
5,7	3	9,0	17,1	32,5
9,9	5	25,0	49,5	98,0
16,3	8	64,0	130,4	265,7
19,3	10	100,0	193,0	372,5
25,7	12	144,0	308,4	660,5
28,2	14	196,0	394,8	795,2
32,6	15	225,0	489,0	1.062,8
141,2	69	767,0	1.589,2	3.299,4

$\hat{\beta}_1 = \frac{371,82}{171,18} = 2,17$

$\hat{\beta}_0 = 17,65 - (2,17)(8,63) = -1,08$

$\hat{Y} = -1,08 + 2,17X$

$\overline{y} = 17,63 \overline{x} = 8,63$

$S_{xx} = 767 - 8(8,63)^2 = 171,18$

$S_{xy} = 1.589 - 8(8,63)(17,65) = 371,82$

$S_{yy} = 3.299,4 - 8(17,65)^2 = 807,22$

$r = \frac{371,82}{\sqrt{(171,18)(807,22)}} \approx 1$

Interpretação

$\hat{Y} = -1,08 + 2,17X$

- Inclinação:
- Taxa de hidrolisação de procaína por minuto
- *Quando o tempo aumenta 1 min, o aumento estimado na procaína hidrolisada é 21,7 moles por litro*
- As estimativas são válidas dentro da classe amostrada (tempo entre 2 e 15 minutos)
- Intercepto-*y*
- A reta indica –10,8 moles por litro de procaína hidrolisada no instante inicial
- Esta interpretação não é válida já que não há pontos amostrais próximos ao tempo igual a zero

Referências

Bibliografia

- Soares, F., Siqueira, A. (Coopmed) *Introdução à Estatística Médica*