Ensino de Estatística

Lupércio França Bessegato
Ronaldo Rocha Bastos
Departamento de Estatística/UFJF

Roteiro Geral

1. **Tratamento da informação**
2. Introdução ao R
3. Produção de dados
4. Probabilidade
5. Inferência estatística
6. Referências

Tratamento da Informação

Roteiro do Módulo

1. Tratamento da Informação:
   a) Conceitos básicos
   b) Resumos tabulares de dados
   c) Resumos numéricos de dados
   d) Visualização gráfica
   e) Análise exploratória univariada
   f) Análise exploratória bivariada

Questionário

Roteiro

1. Introdução
2. Tabelas de Frequência
3. Apresentação Gráfica
4. Medidas-resumo
5. Análise Exploratória Univariada

6

Introdução

O que é *Estatística*?

- Segundo *Magalhães e Lima (2005)*, ***Estatística*** é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

8

Organização e Representação de Dados

- Uma das formas de organizar e resumir a informação contida em dados observados é por meio de tabela de frequências e gráficos.
- *Tabela de frequência:* relaciona categorias (ou classes) de valores, juntamente com contagem (ou frequências) do número de valores que se enquadram em cada categoria ou classe.
- *Elementos gráficos:* ajudam na visualização das principais características dos dados.
- *Medidas resumo:* Medidas de posição, dispersão, assimetria e curtose.

9

Variáveis

- Qualquer característica associada a um elemento pertencente a uma população ou uma amostra
- Classificação de variáveis:

Qualitativa { Nominal — Sexo, cor dos olhos
Ordinal — Classe social, grau de instrução

Quantitativa { Discreta — Número de filhos, nº de carros
Contínua — Peso, altura, salário

10

Dados Brutos

- Obtidos diretamente de pesquisa
  - √ Ainda sem qualquer processo de síntese ou análise
- Incluídos em tabelas
  - √ Porém, não incluídos em publicações

11

Atividade nº 1

Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma	Tder	Exerc	Cine	OpCine	TV	OpTV
1	A	F	17	1,60	60,5	2	NAO	P	0	1	B	16	R
2	A	F	18	1,69	55,0	1	NAO	M	0	1	B	7	R
3	A	M	18	1,85	72,8	2	NAO	P	5	2	M	15	R
4	A	M	25	1,85	80,9	2	NAO	P	5	2	B	20	R
5	A	F	19	1,58	55,0	1	NAO	M	2	2	B	5	R
6	A	M	19	1,76	60,0	3	NAO	M	2	1	B	2	R
7	A	F	20	1,60	58,0	1	NAO	P	3	1	B	7	R
8	A	F	18	1,64	47,0	1	SIM	I	2	2	M	10	R
9	A	F	18	1,62	57,8	3	NAO	M	3	3	M	12	R
10	A	F	17	1,64	58,0	2	NAO	M	2	2	M	10	R
11	A	F	18	1,72	70,0	1	SIM	I	10	2	B	8	N
12	A	F	18	1,66	54,0	3	NAO	M	0	2	B	0	R
13	A	F	21	1,70	58,0	2	NAO	M	6	1	M	30	R
14	A	M	19	1,78	68,5	1	SIM	I	5	1	M	2	N
15	A	F	18	1,65	63,5	1	NAO	I	4	1	B	10	R
16	A	F	19	1,63	47,4	3	NAO	P	0	1	B	18	R
17	A	F	17	1,82	66,0	1	NAO	P	3	1	B	10	N
18	A	M	18	1,80	85,2	2	NAO	P	3	4	B	10	R
19	A	F	20	1,60	54,5	1	NAO	P	3	2	B	5	R
20	A	F	18	1,68	52,5	3	NAO	M	7	2	B	14	M
21	A	F	21	1,70	60,0	2	NAO	P	8	2	B	5	R
22	A	F	18	1,65	58,5	1	NAO	M	0	3	B	5	R
23	A	F	18	1,57	49,2	1	SIM	I	5	4	B	10	R
24	A	F	20	1,55	48,0	1	SIM	I	0	1	M	28	R
25	A	F	20	1,69	51,6	2	NAO	P	8	5	M	4	N
26	A	F	19	1,54	57,0	2	NAO	I	6	2	B	5	R
27	B	F	23	1,62	63,0	2	NAO	M	8	2	M	5	R
28	B	F	18	1,62	52,0	1	NAO	P	1	1	M	10	R
29	B	F	18	1,57	49,0	2	NAO	P	3	1	B	12	R

13

Tabelas de Frequência

Tabelas de Frequências

- Uso:
  - √ Variáveis Qualitativas ou Quantitativas Discretas.
- Contém valores da variável e suas respectivas contagens (frequências absolutas e relativas)
  - √ Frequência absoluta (n_i): contagem das ocorrências de cada valor da variável; seu total é n (o total da amostra);
  - √ Frequência relativa (f_i): proporção de ocorrência de cada valor ($f_i=n_i/n$); seu total é 1
    (útil para fazer comparações entre grupos).

15

Tabelas de Frequência - Exemplo

Tabela de Frequências		
Sexo	Freq. Absolutas	Freq. Relativas
F	37	0,74
M	13	0,26
Total	50	1

- Classe: contém, na base de dados, quantos alunos são do sexo Masculino e quantos são do sexo Feminino.

16

Tabelas de Frequência para Variáveis Ordenadas

- Quando existe uma ordenação das categorias de uma variável (qualitativa ordinal ou quantitativa), faz sentido inserirmos na tabela uma outra coluna, a da frequência acumulada (f_{ac}), que é a soma das frequências relativas, do menor valor até o atual.

17

Exemplo: Tabela de Frequência para a Variável 'Tolerância'

Toler	Freqüência Absoluta (n_i)	Freqüência Relativa (f_i)
M	19	38,0%
P	21	42,0%
I	10	20,0%
Total	**50**	**100,0%**

18

Exemplo: Tabela de Frequência para a Variável 'Nº de filhos'

Tabela de Frequências para a variável "Nº de Filhos"

Filhos	Freq. Absolutas	Freq. Acumuladas	Freq. Relativas	Freq. Acumuladas relativas
1	28	28	0,56	0,56
2	14	42	0,28	0,84
3	6	48	0,12	0,96
4	1	49	0,02	0,98
5	0	49	0	0,98
6	0	49	0	0,98
7	1	50	0,02	1
Total	50		1	

- % famílias que não têm filho único?
- % famílias com pelo menos 2 filhos?
- % famílias com mais de 3 filhos?

19

Atividade nº 2

Nº	Estado Civil	Grau de Instrução	No de filhos	Salário (X Sal. Mín)	Idade anos meses	Região de procedência
1	Solteiro	1º grau	-	4,00	26 03	Interior
2	Casado	1º grau	1	4,56	32 10	Capital
3	Casado	1º grau	2	5,25	36 05	Capital
4	Solteiro	2º grau	-	5,73	20 10	Outro
5	Solteiro	1º grau	-	6,26	40 07	Outro
6	Casado	1º grau	0	6,66	28 00	Interior
7	Solteiro	1º grau	-	6,86	41 00	Interior
8	Solteiro	1º grau	-	7,39	43 04	Capital
9	Casado	2º grau	1	7,59	34 10	Capital
10	Solteiro	2º grau	-	7,44	23 06	Outro
11	Casado	2º grau	2	8,12	33 06	Interior
12	Solteiro	1º grau	-	8,46	27 11	Capital
13	Solteiro	2º grau	-	8,74	37 05	Outro
14	Casado	1º grau	3	8,95	44 02	Outro
15	Casado	2º grau	0	9,13	30 05	Interior
16	Solteiro	2º grau	-	9,35	38 08	Outro
17	Casado	2º grau	1	9,77	31 07	Capital
18	Casado	1º grau	2	9,80	39 07	Outro
19	Solteiro	Superior	-	10,53	25 08	Interior
20	Solteiro	2º grau	-	10,76	37 04	Interior
21	Casado	2º grau	1	11,06	30 09	Outro
22	Solteiro	2º grau	-	11,59	34 02	Capital
23	Solteiro	1º grau	-	12,00	41 00	Outro
24	Casado	Superior	0	12,79	26 01	Outro
25	Casado	2º grau	2	13,23	32 05	Interior
26	Casado	2º grau	2	13,60	35 00	Outro
27	Solteiro	1º grau	-	13,85	46 07	Outro
28	Casado	2º grau	0	14,69	29 08	Interior
29	Casado	2º grau	5	14,71	40 06	Interior
30	Casado	2º grau	2	15,99	35 10	Capital
31	Solteiro	Superior	-	16,22	31 05	Outro
32	Casado	2º grau	1	16,61	36 04	Interior
33	Casado	Superior	3	17,26	43 07	Capital
34	Solteiro	Superior	-	18,75	33 07	Capital
35	Casado	2º grau	2	19,40	48 11	Capital
36	Casado	Superior	3	23,30	42 02	Interior

21

Apresentação Gráfica

Gráficos

- Objetivo:
  - √ Identificação da forma do conjunto de dados
  - √ Resumo e identificação
  - √ Padrão dos dados
- Em geral, facilita a visualização de informações contida em tabelas
- Construção simplificada atualmente por programas computacionais

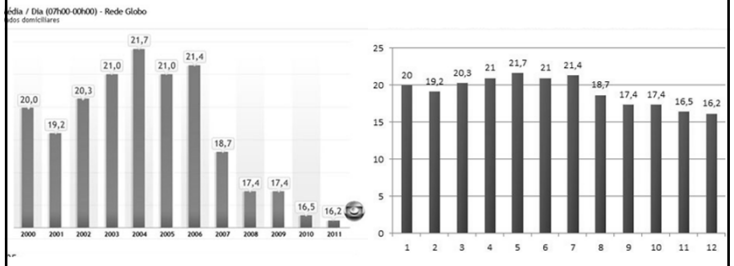
23

Cuidados

- Gráfico com medidas desproporcionais pode:
  - √ Dar falsa impressão de desempenho
  - √ Conduzir a conclusões equivocadas

24

Exemplo: Audiência



Tipos Básicos

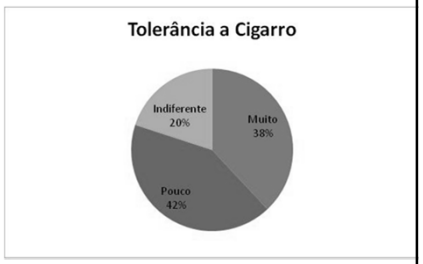
- Gráfico de setores (disco, pizza)
  - √ adapta-se muito bem às variáveis qualitativas nominais
- Gráfico de barras
  - √ adapta-se melhor às variáveis quantitativas discretas ou às variáveis qualitativas ordinais
- Histograma
  - √ utilizado com variáveis quantitativas contínuas

Gráfico de Setores

- Adapta-se muito bem às variáveis qualitativas nominais
- Repartição de disco em setores circulares correspondentes às frequências relativas de cada valor da variável

Exemplo: Tolerância a Cigarro

Toler	n_i	f_i
M	19	38,0%
P	21	42,0%
I	10	20,0%
Total	**50**	**100,0%**



- Importante:
  - √ Use com variáveis com até no máximo 6 níveis
  - √ Os valores não devem ser muito próximos

Gráfico de Setores – Comentários

- O gráfico de setores não é uma forma boa de visualizar informações!
  - √ O olho é bom para julgar medidas lineares e ruim em julgar áreas relativas.
- Um gráfico de barras ou um diagrama de pontos são formas preferíveis de dispor este tipo de dado.

Cleveland (1985): *"Dados que podem ser mostrados por um gráfico de setores sempre podem ser mostrados por um gráfico de barras ou um diagrama de pontos. Isto significa que julgamentos da posição em meio a uma escala comum podem ser feitos em vez de julgamentos menos acurados via ângulos dos setores."*

29

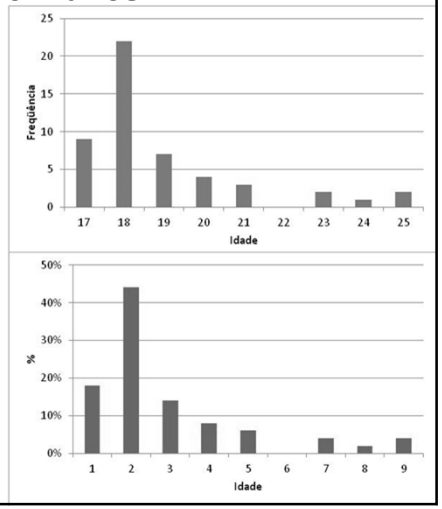
Gráfico de Barras

- Para cada valor da variável desenha-se uma barra com altura correspondente à sua frequência (absoluta ou relativa)
  - ✓Eixo das abscissas (x): valores da variável
  - ✓Eixo das ordenadas (y): frequências absolutas ou relativas
- Adapta-se melhor às variáveis quantitativas discretas ou qualitativas ordinais

30

Exemplo: Idade de Alunos

Idade	n_i	f_i
17	9	18,0%
18	22	44,0%
19	7	14,0%
20	4	8,0%
21	3	6,0%
22	0	0,0%
23	2	4,0%
24	1	2,0%
25	2	4,0%
Total	**50**	**100,0%**



31

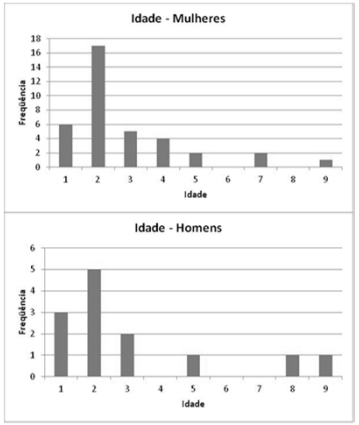
Recomendações

- Colunas sempre com mesma largura
- Distância entre colunas deve ser constante
- Para comparar diferentes amostras:
  - √ Utilizar frequências relativas
  - √ Uniformizar as escalas de ambos os eixos

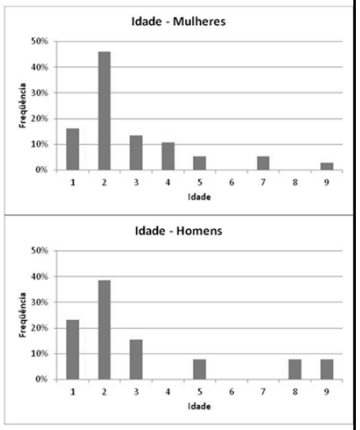
32

Comparação Idade *vs.* Sexo

Incorreto



Correto



Comparação Idade *vs.* Sexo

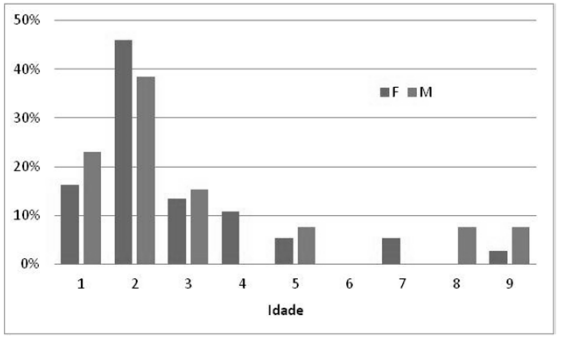
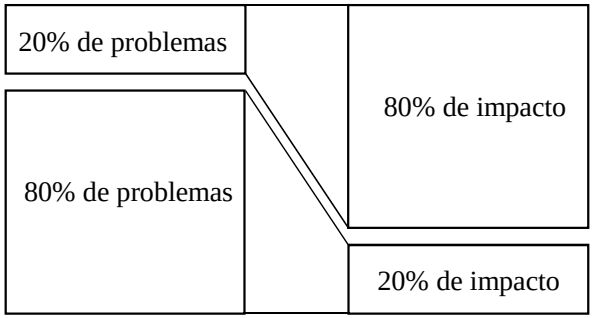


Gráfico de *Pareto*

- É essencialmente um gráfico de barras com os itens ordenados por tamanho
- Objetivo:
  - √ Ordenar tipo de problemas por tamanho
  - √ Foco na gestão dos problemas mais importantes

Princípio de *Pareto*

- Técnica que busca separar os problemas vitais (poucos) dos triviais (muitos)



Problemas

- "Poucos e vitais":
  - √ Representam um **pequeno número de problemas** que, no entanto, resultam em **grandes perdas**.
- "Muitos e triviais":
  - √ São um **grande número de problemas** que resultam em **perdas pouco significativas**.

37

Objetivo

- Identificar as causas dos "poucos problemas vitais";
- Focar na solução dessas causas;
- Eliminar uma parcela importante dos problemas com um pequeno número de ações.

38

Diagrama de *Pareto*

- Distribuição de frequências de dados organizados por categorias:
  - √ Marca-se a frequência total de ocorrência de cada defeito *vs.* o tipo de defeito
  - √ Uma escala para frequência absoluta e outra para a frequência relativa acumulada.

39

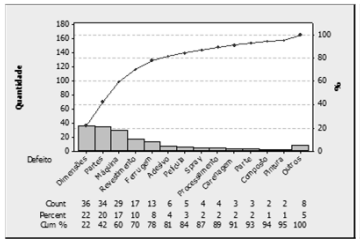
Diagrama de *Pareto*

- Identifica-se rapidamente os problemas que ocorrem com maior frequência
- Os problemas mais frequentes não são necessariamente os mais importantes.

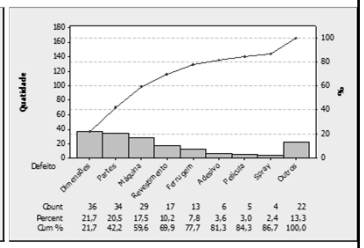
40

Exemplo

- Gráfico Pareto



Outros: 5%



Outros: 15%

Procedimento

- Categorizar os quesitos (problemas) do processo
- Coletar a frequência de cada um deles durante um período
- Ordenar do mais frequente para o menos frequente
- Construir um gráfico de barras
- Adicionar um gráfico de frequências acumuladas

Exemplo

√ Problemas em empréstimos de livros em biblioteca escolar

Problema	n_i	f_i	f_{ac}
Emprestados	135	32,9%	32,9%
Em uso no recinto	108	26,3%	59,3%
Pedido não localizado	57	13,9%	73,2%
Na encadernação	43	10,5%	83,7%
Fora de lugar	12	2,9%	86,6%
Em processamento	12	2,9%	89,5%
Classificação errada	7	1,7%	91,2%
Outros	36	8,8%	100,0%
Total	**410**	**100,0%**	

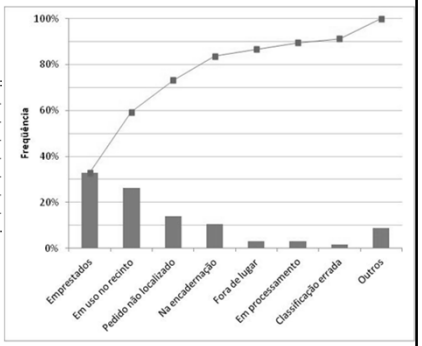
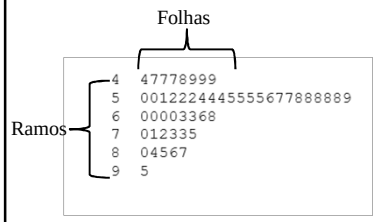
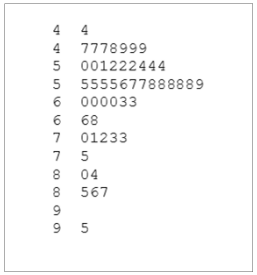


Gráfico Ramo-e-Folhas

- Dados são agrupados preservando quase toda a informação numérica
- Adequado para representação de conjunto de dados de 15 a 150 valores, aproximadamente

Exemplo: Peso





cada linha: folhas 0, 1, 2, ..., 9

1ª. linha: folhas 0, 1, 2, 3, 4

2ª. linha: folhas 5, 6, 7, 8, 9

- folha representa um único dígito
  - √ 60,5 kg ➔ 6 | 0

- Representar os valores:

  220 214 222 218 223 210 223 210 227 225 212
- Suponha que queremos dividir cada número após o 2º. dígito:

  220 = 22 | 0
- Procedimento:
  - √ Ramos do gráfico
  - √ Adicione 220 ao gráfico
  - √ Adicione 214 ao gráfico
  - √ Adicione demais números
  - √ Ordene as folhas

21	**0 0 2 4 8**
22	**0 2 3 3 5 7**

- No exemplo: intervalo de classes = 10

Expandindo o Gráfico

- Folhas 0, 1, 2, 3, 4 em uma linha
- Folhas 5, 6, 7, 8, 9 na seguinte
- Valores:

  220 214 222 218 223 210 223 210 227 225 212

21	**0 0 2 4**
21	**8**
22	**0 2 3 3**
22	**5 7**

Informe de Unidades

- Unidades 8 | 3 = 83.000

  9 | 7 = 97.000
- Unidades 8 | 3 = 0,083

  9 | 7 = 0,097

Comentários

- Um gráfico ramo-e-folhas com menos de 5 ramos ativos é altamente não informativo
- Em geral, não se usa mais que 10 a 15 ramos ativos
- Regras práticas e definitivas são improdutivas
  - √ gráficos de comprimentos diferentes podem transmitir informações diferentes

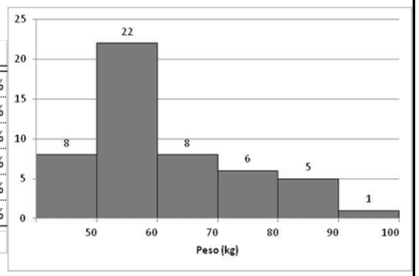
Atividade nº 3

Histograma

- Características da forma do histograma:
  - √ número, largura e altura dos retângulos
- Retângulos contíguos:
  - √ eixo abscissas (x): base correspondente ao intervalo de classe
  - √ eixo das ordenadas (y): altura correspondente à frequência (ou porcentagem) do intervalo de classe
- Usado para representação gráfica da distribuição de variáveis contínuas
  - √ São parecidos com os gráficos de ramo-e-folhas

Exemplo: Peso

Peso	n_i	f_i	f_{ac}
40 ⊢ 50	8	16,0%	16,0%
50 ⊢ 60	22	44,0%	60,0%
60 ⊢ 70	8	16,0%	76,0%
70 ⊢ 80	6	12,0%	88,0%
80 ⊢ 90	5	10,0%	98,0%
90 ⊢ 100	1	2,0%	100,0%
Total	50	100,0%	



- Em geral, utilizam-se de 5 a 8 faixas com mesma amplitude (preferencialmente)

Histograma – Construção

- Determinam-se o máximo e o mínimo dos dados
- Divide-se a amplitude dos dados em um número conveniente de intervalos de classe de tamanhos iguais
- Contam-se a quantidade de observações que caem em cada um desses intervalos (frequência)
- Altura do retângulo acima de um intervalo de classe é igual à frequência

53

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 → 150 ├ 154

Li → 150

Ls → 154

h = Ls-Li

AT = Ls max-Li min

Ponto médio = $(L_s + L_i)/2$

54

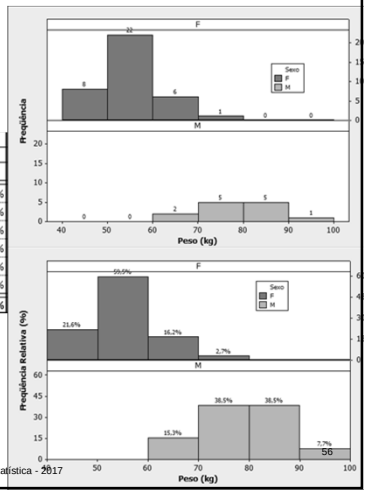
Histograma – Comparações

- Histograma de frequência relativa:
  - √ Altura do retângulo = frequência relativa do intervalo
  - √ Conveniente para comparar histogramas baseados em amostras de tamanhos diferentes
- Motivo: aspectos principais captados no histograma: formato geral e área dos retângulos
  - √ Se intervalos de classe são iguais essas áreas são proporcionais às frequências

55

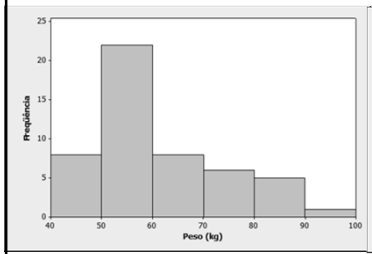
Exemplo: Peso por Sexo

	Freqüências					
Sexo	F		M		Total	
Peso	n_i	f_i	n_i	f_i	n_i	f_i
40 ├ 50	8	21,6%	0	0,0%	8	16,0%
50 ├ 60	22	59,5%	0	0,0%	22	44,0%
60 ├ 70	6	16,2%	2	15,4%	8	16,0%
70 ├ 80	1	2,7%	5	38,5%	6	12,0%
80 ├ 90	0	0,0%	5	38,5%	5	10,0%
90 ├ 100	0	0,0%	1	7,7%	1	2,0%
Total	37	100,0%	13	100,0%	50	100,0%

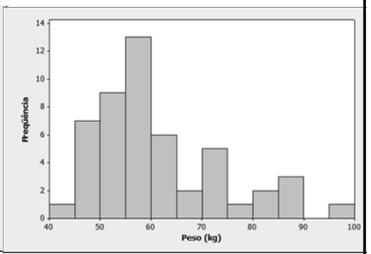


56

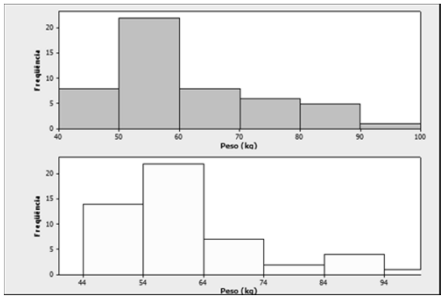
- Formato do histograma depende:
  - √ largura escolhida para os intervalos de classe
  - √ posicionamento dos extremos dos intervalos de classe



Histograma original
(largura do intervalo = 10)



Largura de intervalo modificada
(largura do intervalo = 5)



Mesmas larguras, limites diferentes
(largura do intervalo = 10)

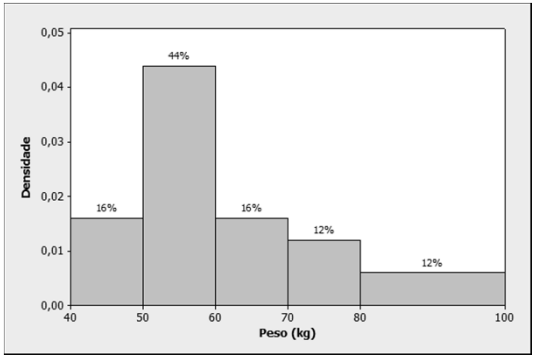
Histograma de Densidade

- Área de cada retângulo representa a frequência relativa do intervalo de classe correspondente
  - √ Soma das áreas de todos os retângulos = 1 (100%)
- Densidade de frequência: altura do retângulo

$$\text{densidade} = \frac{\text{área retângulo}}{\text{amplitude intervalo}}$$

- O histograma de densidade não fica distorcido quando ele é construído com intervalos de amplitudes diferente

Exemplo: Peso de Estudantes



- Evitou distorção do intervalo entre 80 e 100!

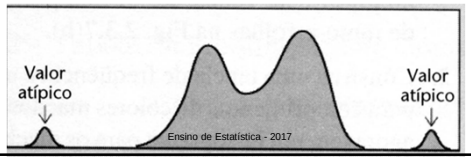
Interpretação de Gráficos de Ramo-e-Folhas & Histograma

- Em uma análise gráfica procuramos identificar:
  - √ PADRÃO GLOBAL nos dados
  - √ Desvios acentuados em relação ao mesmo
- Importante:
  - √ Não perceberemos padrões nos dados se houver um número muito pequeno ou muito grande de intervalos de classe
- Procuramos uma impressão geral suavizada
  (não reagimos a pequenas subidas ou descidas)

61

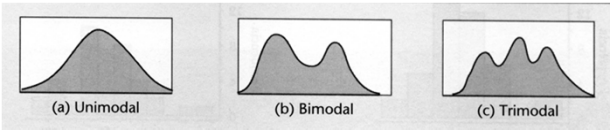
Valores Atípicos (*Outliers*)

- Procuramos por observações que estejam bem afastadas da maioria dos dados
  - √ Observações discrepantes (*outliers*)
- Analisar estas observações com mais cuidado
  - √ Porque razão são tão diferentes?
  - √ Está ocorrendo algo incomum ou interessante?
  - √ São erros?



62

Existência de Mais de Um Pico

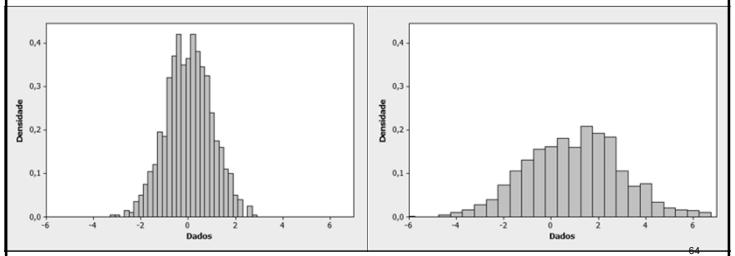


- √ Picos são chamados Modas
- √ Quando há apenas um pico, a moda representa o valor mais popular (ou classe)
- √ Presença de diversas modas é indicador de diversos grupos distintos de dados
- √ Em geral, deve-se investigar os motivos de multimodalidade

63

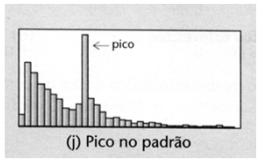
Valores Centrais e Dispersão

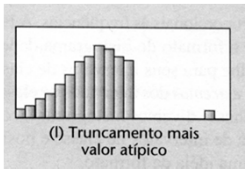
- Observar:
  - √ Onde os dados parecem estar centrados
  - √ Quão espalhados estão os dados
  - √ Posição das modas (caso de multimodalidade)



64

Mudanças Abruptas



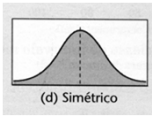


√ Suspeite de mudanças abruptas

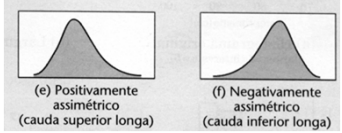
√ Tente estabelecer suas causas

Forma da Distribuição

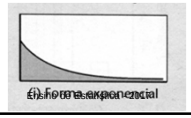
- O gráfico parece ser aproximadamente simétrico?



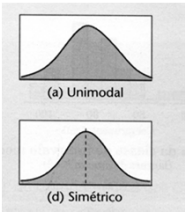
- O gráfico apresenta assimetria moderada?



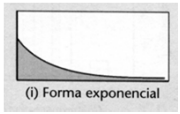
- O gráfico apresenta assimetria extrema?



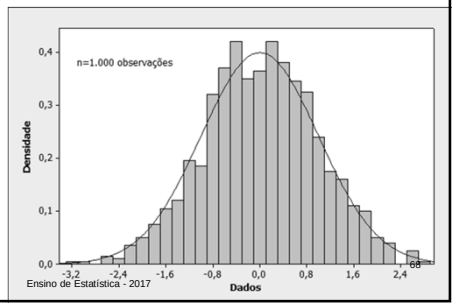
- A envoltória do gráfico tem aproximadamente forma de sino?



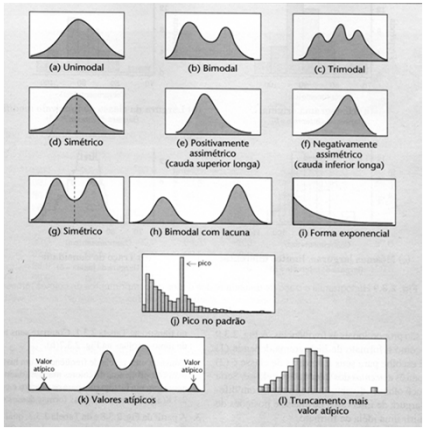
- ou tem forma exponencial?



- Usualmente, técnicas estatísticas formais preferem trabalhar com um histograma simétrico com forma de sino
- A forma do histograma pode sugerir uma função matemática cuja curva se ajusta bem ao histograma



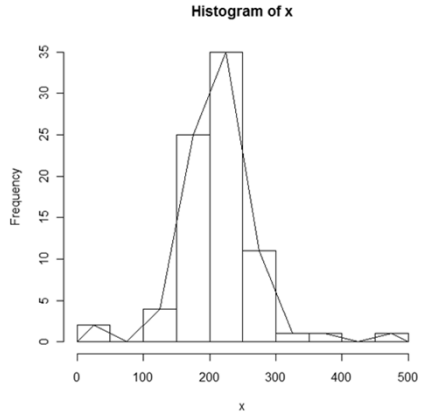
- Características a serem procuradas nos histogramas:



Fonte: Wild, C.J & Seber, G.A *Encontros com o Acaso, LTC, 2000*

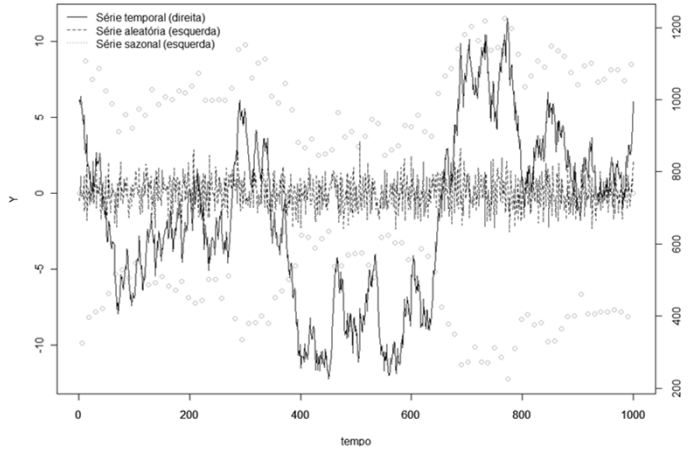
Polígono de Frequências

- 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 frequências da variável

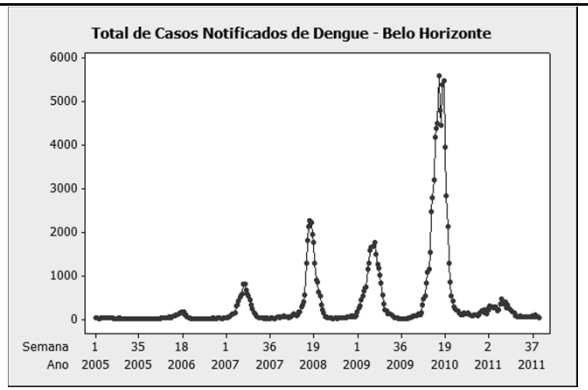


Séries Temporais

- 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

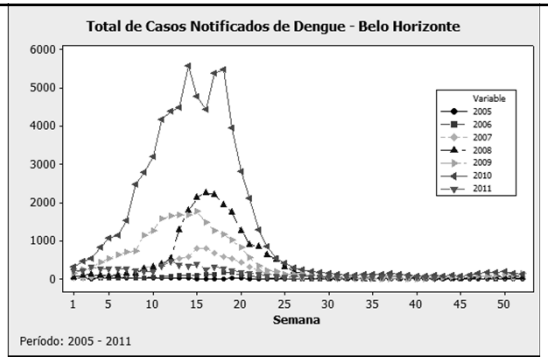


73



- Tendência
- Sazonalidade
- Estacionariedade

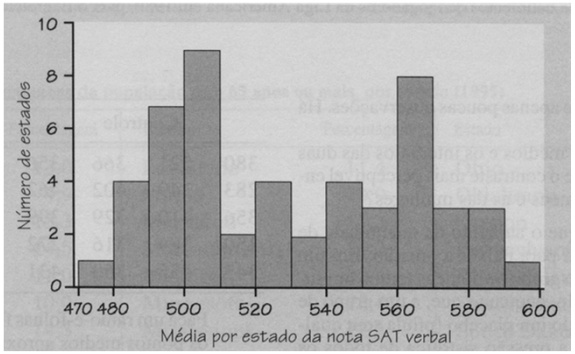
74



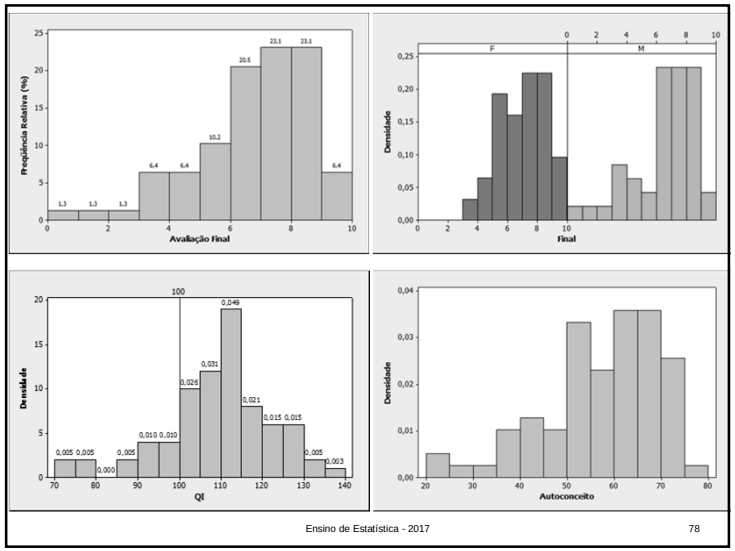
- Sazonalidade
- Variáveis que podem ajudar a explicar total de casos

75

Atividade nº 4



77



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

80

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 selecionada de uma população
- Definição:

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

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

Exemplo – Peso

- Peso (kg)
- *n = 50* indivíduos
- Média amostral

$$\bar{x} = \frac{3.046,4}{50} = 60,93\,\text{kg}$$

85

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

86

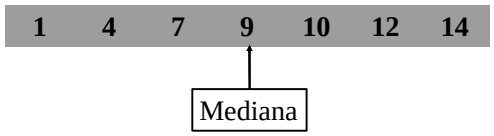
Mediana

- Valor que divide a distribuição dos dados em duas partes de igual tamanho

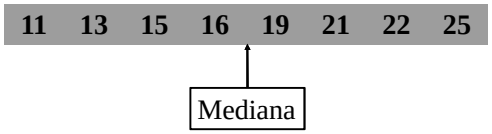
- 50% das observações ficam acima da mediana e 50%, abaixo

87

- Determinação da mediana:
  - √ Quantidade ímpar de observações:



  - √ Quantidade par de observações



88

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 $\frac{n}{2}+1$

89

Exemplo – Peso (kg)

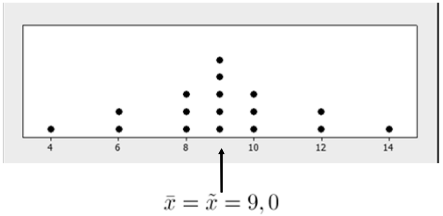
- Peso (kg)
- $n = 50$ indivíduos
- Valor médio entre o 25º e o 26º valores ordenados

  $x_{(25)} = 58$; $x_{(26)} = 58$
- Mediana

$$\tilde{x} = \frac{58+58}{2} = 58\,\text{kg}$$

90

Média & Mediana



91

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 que a média
- A mediana é uma medida mais robusta (menos afetada pro valores atípicos)

92

Média *vs.* Mediana

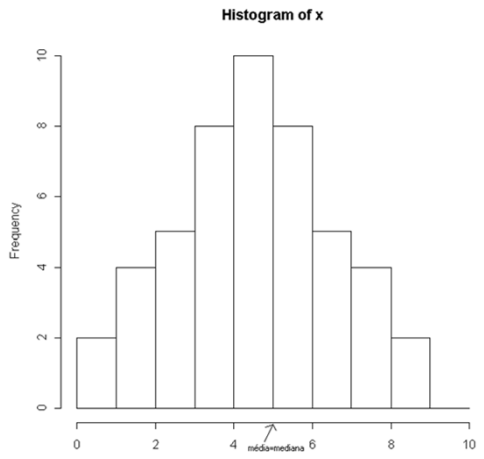
Média	Mediana
• fácil de ser manipulada algebricamente;	• difícil de ser manipulada algebricamente;
• representa o "centro de massa" dos dados (ponto de equilíbrio no histograma).	• valor da posição central dos dados ordenados;
• afetada grandemente por valores extremos .	• não é afetada por valores extremos.

93

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.

94



95

Média – Dados em Tabelas de Frequência

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

Ponto Médio	Freqüência
x_1	f_1
x_2	f_2
⋮	⋮
x_k	f_k

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

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

96

- Exemplo - Tabela de Frequências – Peso

Peso (kg)	Ponto Médio (x_i)	Freq. Absoluta (f_i)	$x_i \cdot f_i$
40 ⊢ 50	45	8	360
50 ⊢ 60	55	22	1210
60 ⊢ 70	65	8	520
70 ⊢ 80	75	6	450
80 ⊢ 90	85	5	425
90 ⊢ 100	95	1	95
Total			3060

$$\bar{x}_{tab.} = \frac{3060}{50} = 61,20\,\text{kg}$$

$$\bar{x}_{exata} = 60,93\,\text{kg}$$

97

Moda

- É o valor mais frequente da distribuição.
- No histograma, ou na tabela de frequências, a classe modal é a classe de maior frequência e a moda são aproximadas pelo ponto médio da classe.

98

Exemplo: Peso

- Classe Modal: [50; 60)
  √ Maior frequência = 22 observações
- Moda: 55 kg

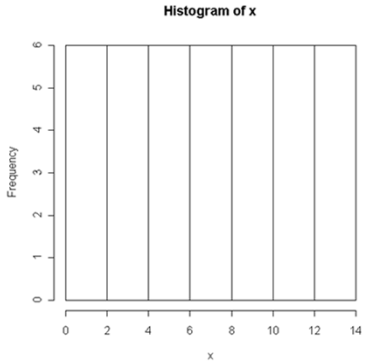
99

Moda (2)

- Uma distribuição pode não possuir moda (amodal – distribuição “achatada”).
- Uma distribuição pode possuir mais de uma moda (multimodal).
- Uma distribuição pode possuir apenas uma moda (unimodal).

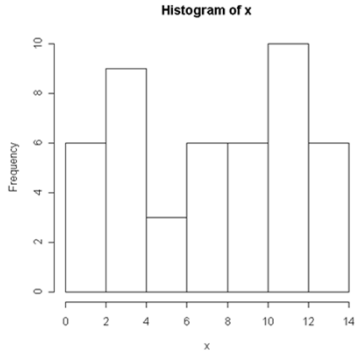
100

Distribuição "Achatada"



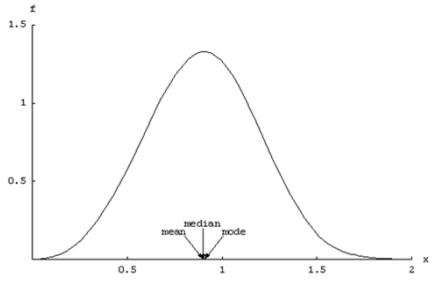
101

Distribuição Multimodal



102

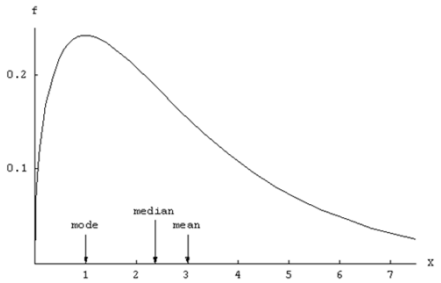
Medidas de Posição – Distribuições Simétricas



média = mediana = moda

103

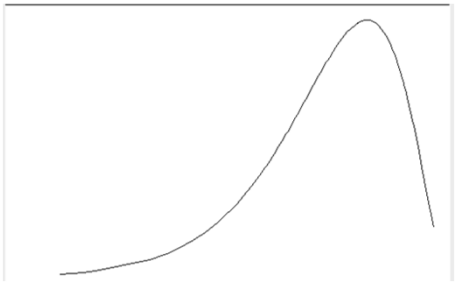
Medidas de Posição – Distribuições Assimétricas à Direita



média > mediana > moda

104

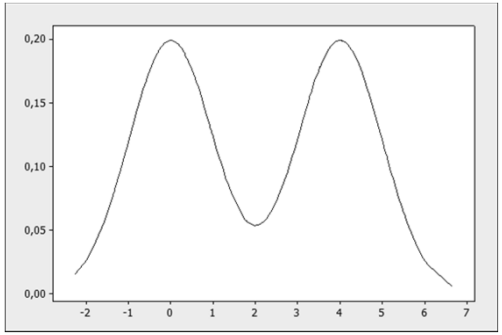
Medidas de Posição – Distribuições Assimétricas à Esquerda



média < mediana < moda

105

Distribuições Bimodais



média = mediana ≠ moda

106

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
```

108

Média e Mediana

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

109

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!

110

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 (ou desvio interquartílico)

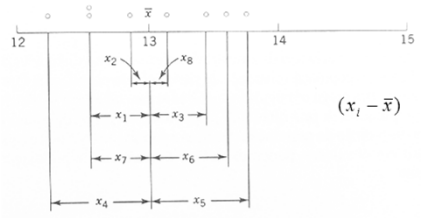
111

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

112

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!

- Conjunto de dados:

5 2 3 4 8

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_5 - \bar{x})^2}{5-1}$$
$$= \frac{(5-4,4)^2 + (2-4,4)^2 + (3-4,4)^2 + (4-4,4)^2 + (8-4,4)^2}{4}$$
$$= \frac{21,2}{4} = 5,3$$

$$s = \sqrt{s^2} = \sqrt{5,3} = 2,30$$

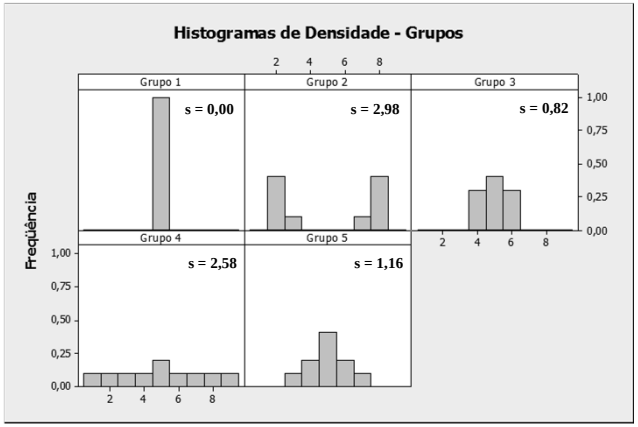
Cálculo Alternativo

- Variância: $s^2 = \frac{1}{n-1}\left[\sum x_i^2 - n(\bar{x})^2\right]$

x_i	x_i^2
5	25
2	4
3	9
4	16
8	64
22	118

$$s^2 = \frac{1}{5-1}\left[118 - 5(4,4)^2\right] = \frac{21,2}{4} = 5,3$$

$$s = \sqrt{5,3} = 2,30$$



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 – Peso

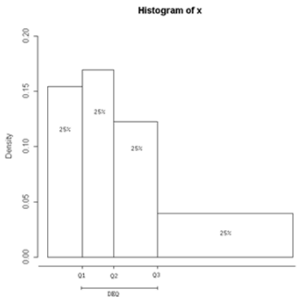
- Peso (kg)
- *n = 50* indivíduos
- Variância: $s^2 = 148,33$
- Desvio-padrão: $s = \sqrt{148,33} = 12,18$
- Média: $\bar{x} = 60,93$
- Coeficiente de variação: $cv = \frac{s}{\bar{x}} = \frac{12,18}{60,93} = 19,99\%$

Atividade nº 5

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!

123

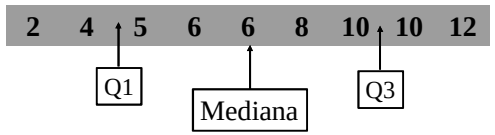
Procedimento para Determinação dos Quartis

- Várias definições são usadas na literatura e por diferentes pacotes computacionais
  - √ As diferentes definições dão respostas muito parecidas
- Regra que adotaremos:
  - √ O primeiro quartil (Q_1) é a mediana de todas as observações com posição estritamente abaixo da posição da mediana
  - √ O terceiro quartil (Q_3) é a mediana das observações que estão estritamente acima da posição da mediana.

124

- Determinação da mediana:

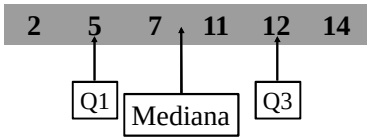
√ $n = 9$



$$Q_1 = \frac{4+5}{2} = 4,5$$

$$Q_3 = \frac{10+10}{2} = 10$$

√ $n = 6$



$$Q_1 = 5$$

$$Q_3 = 12$$

125

Exemplo – Peso

- Peso (kg)

Mínimo	44,0	$x_{(1)} = 44,0$
Q_1	52,0	$x_{(13)}=52,0$
Q_2 = Mediana	58,0	$x_{(25)}=58,0$ $x_{(26)}=58,0$
Q_3	68,5	$x_{(38)}=68,5$
Máximo	95,0	$x_{(50)}=95,0$

126

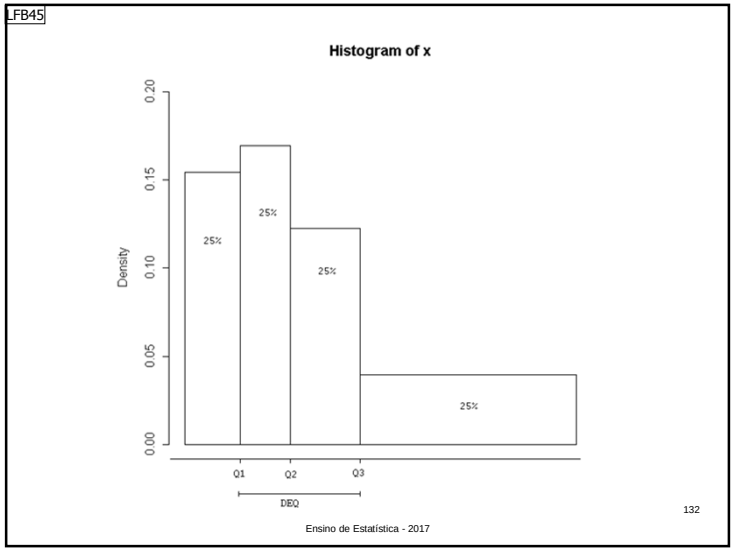
Distância Interquartílica

- Medida de variabilidade dada por .

$$DI = Q_3 - Q_1$$

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

131



132

LFB45 Refazer o gráfico em Minitab

Lupercio Bessegato; 25/09/2007

Exemplo: Peso

- Peso (kg)

Q_1	52,0	$x_{(13)}$=52,0
Q_2 = Mediana	58,0	$x_{(25)}$=58,0 $x_{(26)}$=58,0
Q_3	68,5	$x_{(38)}$=68,5
Distância Interquartílica	**16,50**	$Q_3 - Q_1$

133

Box-plot

Esquema dos 5 Números

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

135

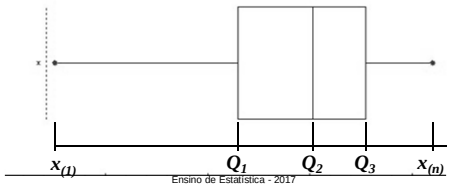
Esquema dos 5 Números (2)

- Para uma distribuição 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.

136

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

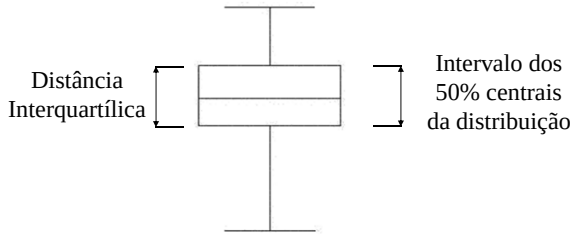


137

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* corresponde aos 50% valores centrais da distribuição.

138



139

Região de Observações Típicas

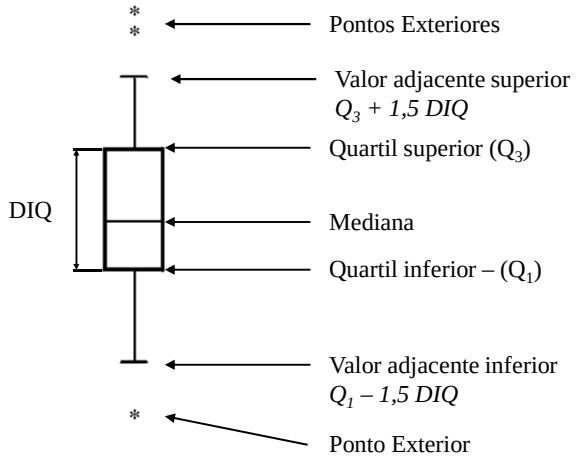
- Delimita-se a região que vai da base superior do retângulo até o maior valor observado que NÃO supere o valor de $Q_3+1{,}5 \times DIQ$.
- Procedimento similar para delimitar a região que vai da base inferior do retângulo, até o menor valor que NÃO é menor do que $Q_1-1{,}5 \times DIQ$.

140

Região de Observações Atípicas

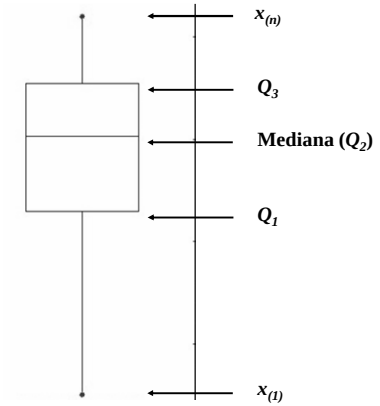
- Observações são representadas por asteriscos e situam-se:
  - √ ou, acima do Valor adjacente superior (*Q3 + 1,5 DIQ*)
  - √ ou, abaixo do Valor adjacente inferior (*Q1 – 1,5 DIQ*)
- Estes pontos exteriores são denominados *outliers* ou valores atípicos.

141



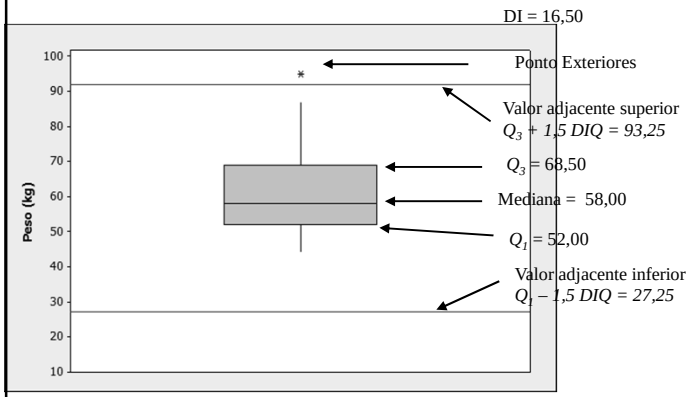
142

- Se não houver pontos exteriores:



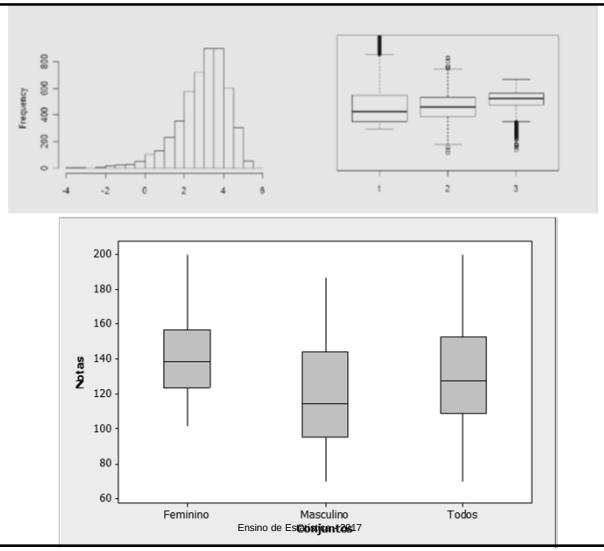
143

Exemplo - Peso



144

Atividade nº 6



Associação entre Variáveis

Atividade nº 7

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)…Neste curso, trabalhamos com alguns deles:
  - √ Diagrama de dispersão
  - √ **Ramo e folhas (p/ conhecer)**
  - √ **Boxplot**
  - √ Individual Plot

150

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

151

ideia Básica

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

152

Clássica *vs.* Exploratória

- Sequencia 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*)
- Todos estes aspectos foram tratados neste curso!

Referências

Bibliografia Recomendada

- AGRESTI, A.; FINLAY, B. *Métodos estatísticos para as ciências sociais.* Penso, 2012.
- MAGALHÃES, M.N.; LIMA, A.C.P.L. *Noções de Probabilidade e Estatística.* Edusp, 2011.
- MOORE, D. S.; MCCABE, G. P. *Introdução à prática da estatística*. LTC, 2002.
- WILD, C.J. E SEBER, G.A.F. *Encontros com o acaso: um primeiro curso de análise de dados e inferência*. LTC, 2000.

158

Ensino de Estatística - 2017