Elementos de Estatística

Lupércio F. Bessegato & Marcel T. Vieira

UFJF – Departamento de Estatística
2013

Gráficos & Tabelas

Descrição Tabular e Gráfica

- Tabelas:
  - √ Tipos de variáveis e tabelas
  - √ Frequências absolutas, relativas e acumuladas
  - √ Tabelas de dupla entrada
  - √ Representação tabular brasileira
- Gráficos
  - √ de setores e de barras
  - √ histogramas
  - √ de Pareto
  - √ etc.

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).

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.

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.

Exemplo: Tabela de Frequência para a Variável 'Tolerância'

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

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?

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	20 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	-	6,25	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	25 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

Apresentação Gráfica

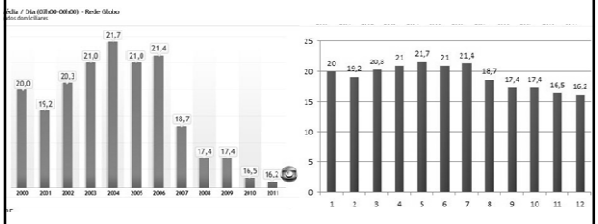
Gráficos

- 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

Cuidados

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

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.

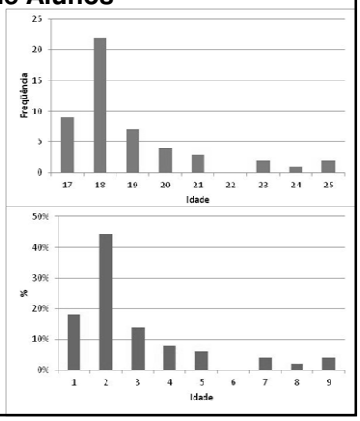
Cleveland (1985): *"Dados que podem ser mostrados por um gráfico de setores sempre podem ser mostrados por um gráfico de barras ou um diagrama de pontos. Isto significa que julgamentos da posição em meio a uma escala comum podem ser feitos em vez de julgamentos menos acurados via ângulos dos setores."*

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

Exemplo: Idade de Alunos

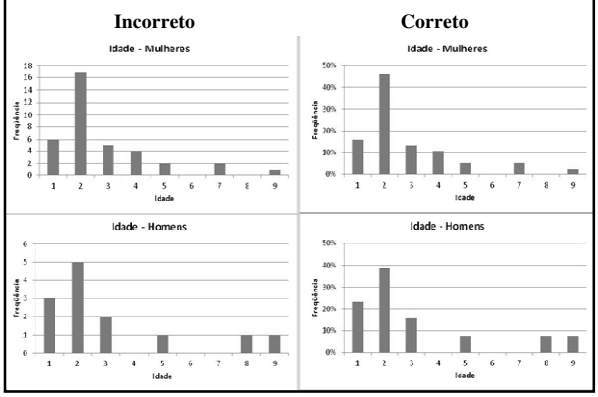
Idade	n_i	f_i
17	9	18,0%
18	22	44,0%
19	7	14,0%
20	4	8,0%
21	3	6,0%
22	0	0,0%
23	2	4,0%
24	1	2,0%
25	2	4,0%
Total	50	100,0%



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

Comparação Idade *vs.* Sexo



Comparação Idade *vs.* Sexo

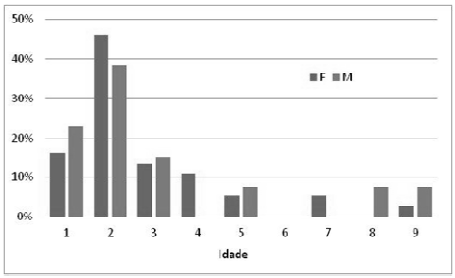
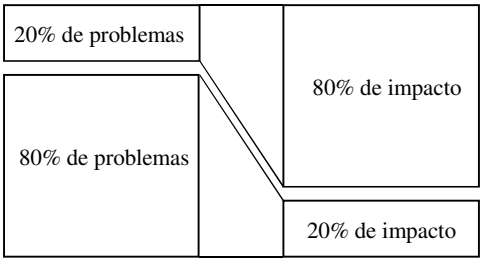


Gráfico de *Pareto*

- 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**.

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.

Diagrama de *Pareto*

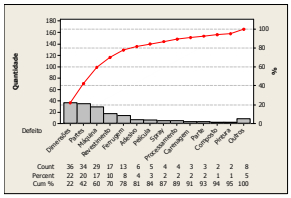
- Distribuição de frequências de dados organizados por categorias:
  - √ Marca-se a frequência total de ocorrência de cada defeito *vs.* o tipo de defeito
  - √ Uma escala para frequência absoluta e outra para a frequência relativa acumulada.

Diagrama de *Pareto*

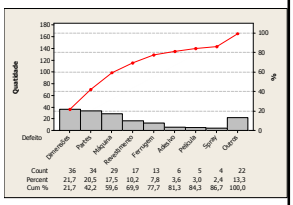
- Identifica-se rapidamente os problemas que ocorrem com maior frequência
- Os problemas mais frequentes não são necessariamente os mais importantes.

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	**413**	**100,0%**	

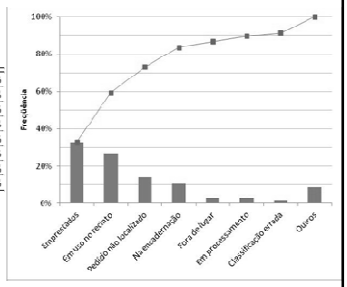
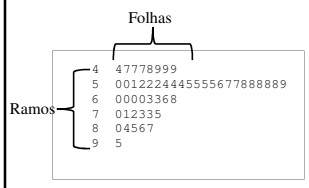


Gráfico Ramo-e-Folhas

- 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

```
4  4
4  7778999
5  001222444
5  55556778888889
6  000033
6  68
7  01233
7  5
8  04
8  567
9
9  5
```

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
- No exemplo: intervalo de classes = 10

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

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

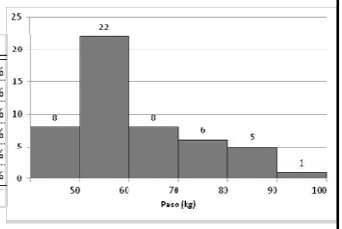
Atividade nº 3

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

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

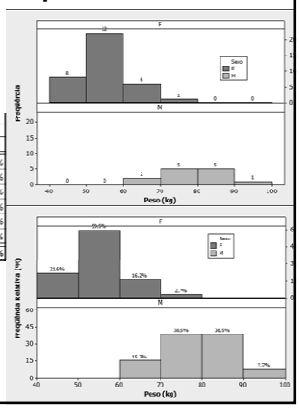
Ponto médio = (Ls –Li)/2

Histograma – Comparações

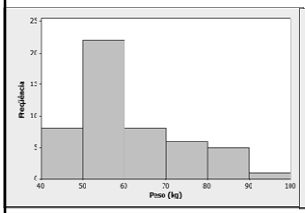
- Histograma de frequência relativa:
  - √ Altura do retângulo = frequência relativa do intervalo
  - √ Conveniente para comparar histogramas baseados em amostras de tamanhos diferentes
- Motivo: aspectos principais captados no histograma: formato geral e área dos retângulos
  - √ Se intervalos de classe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

Exemplo: Peso por Sexo

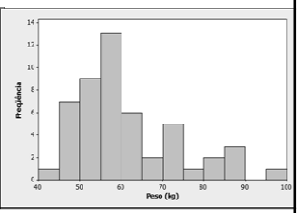
Sexo	F		M		Total	
Peso	n_1	f_1	n_1	f_1	n_1	f_1
40 ⊢ 50	8	21,6%	0	0,0%	8	16,0%
50 ⊢ 60	22	59,5%	0	0,0%	22	44,0%
60 ⊢ 70	6	16,2%	2	15,4%	8	16,0%
70 ⊢ 80	1	2,7%	5	38,5%	6	12,0%
80 ⊢ 90	0	0,0%	5	38,5%	5	10,0%
90 ⊢ 100	0	0,0%	1	7,7%	1	2,0%
Total	37	100,0%	13	100,0%	50	100,0%



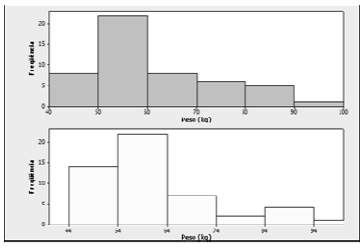
- Formato do histograma depende:
  - √ largura escolhida para os intervalos de classe
  - √ posicionamento dos extremos dos intervalos de classe



Histograma original
(largura do intervalo = 10)



Largura de intervalo modificada
(largura do intervalo = 5)



Mesmas larguras, limites diferentes
(largura do intervalo = 5)

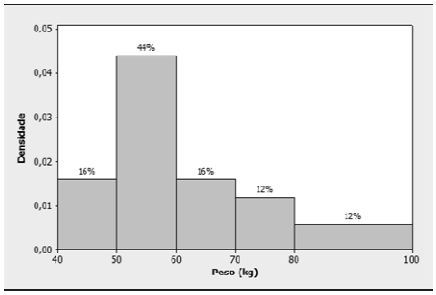
Histograma de Densidade

- 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



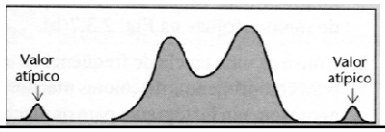
- Evitou distorção do intervalo entre 80 e 100!

Interpretação de Gráficos de Ramo-e-Folhas & Histograma

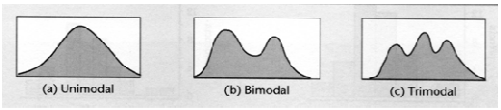
- Em uma análise gráfica procuramos identificar:
  - √ PADRÃO GLOBAL nos dados
  - √ Desvios acentuados em relação ao mesmo
- Importante:
  - √ Não perceberemos padrões nos dados se houver um número muito pequeno ou muito grande de intervalos de classe
- Procuramos uma impressão geral suavizada
  (não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

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?



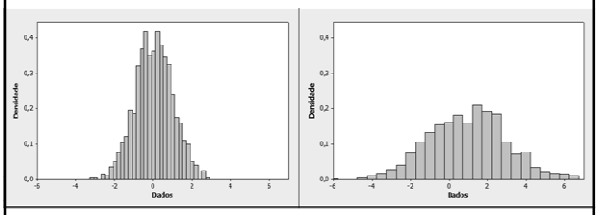
Existência de Mais de Um Pico



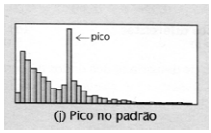
- √ Picos são chamados Modas
- √ Quando há apenas um pico, a moda representa o valor mais popular (ou classe)
- √ Presença de diversas modas é indicador de diversos grupos distintos de dados
- √ Em geral, deve-se investigar os motivos de multimodalidade

Valores Centrais e Dispersão

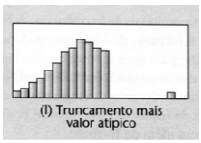
- Observar:
  - √ Onde os dados parecem estar centrados
  - √ Quão espalhados estão os dados
  - √ Posição das modas (caso de multimodalidade)



Mudanças Abruptas



(j) Pico no padrão

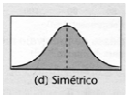

(l) Truncamento mais valor atípico

√ Suspeite de mudanças abruptas

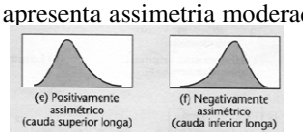
√ Tente estabelecer suas causas

Forma da Distribuição

- O gráfico parece ser aproximadamente simétrico?

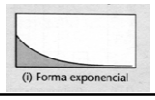

(d) Simétrico

- O gráfico apresenta assimetria moderada?

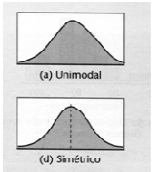

(e) Positivamente assimétrico (cauda superior longa)

(f) Negativamente assimétrico (cauda inferior longa)

- O gráfico apresenta assimetria extrema?

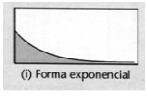

(i) Forma exponencial

- A envoltória do gráfico tem aproximadamente forma de sino?

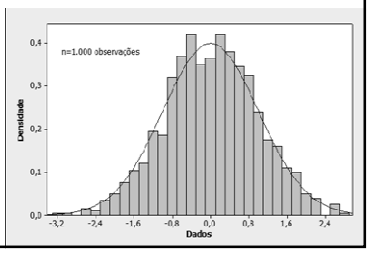

(a) Unimodal

(d) Simétrico

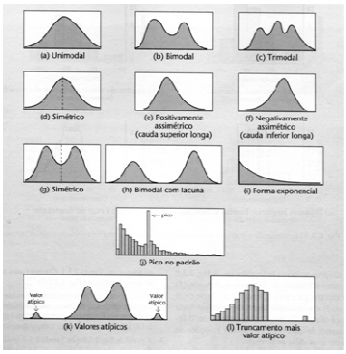
- ou tem forma exponencial?


(i) 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



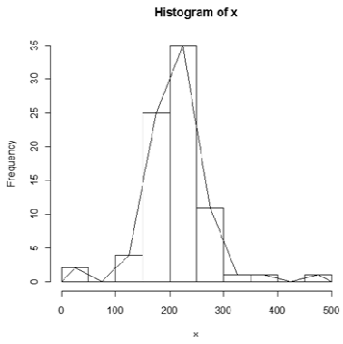
- Características a serem procuradas nos histogramas:



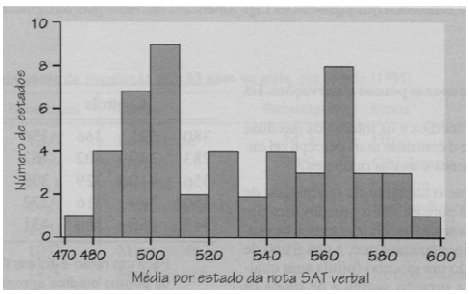
Fonte: Wild, C.J & Seber, G.A *Encontros com o Acaso, LTC, 2000*

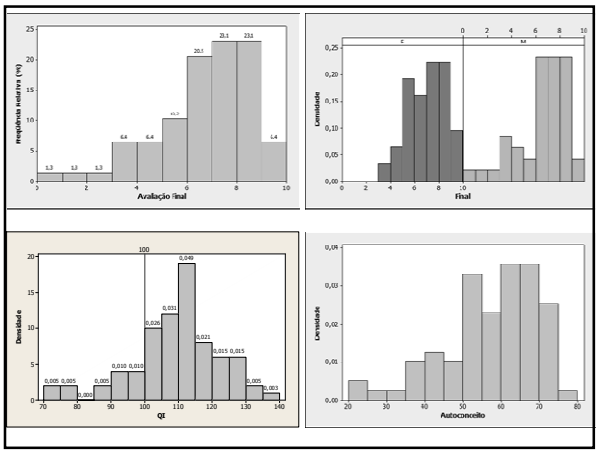
Polígono de Frequências

- 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



Atividade nº 4





Referências

Bibliografia

- Magalhães, M.N. e Lima, A.C.P.L. (Edusp) *Noções de Probabilidade e Estatística*
- Wild, C.J. e Seber, G.A.F. (LTC) *Encontros com o Acaso: um Primeiro Curso de Análise de Dados e Inferência*
- Agresti, A. e Agresti, B.F. (Dellen Pub.) *Statistical Methods for the Social Sciences*