Análise Univariada

Frases

"Torture os dados por um tempo suficiente, e eles contam tudo!"

fonte: mcrsoft@aimnet.com (Barry Fetter)

"Um homem com um relógio sabe a hora certa. Um homem com dois relógios só sabe a média."

Anônimo

Roteiro

Introdução

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

Classificação

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

Variáveis Qualitativas

Exemplo 1 – Tipos de Sangue

- Registro do tipo sangüíneo de 40 doadores voluntários de sangue em um dia.

  √ Os dados estão na planilha tipo_sangue.

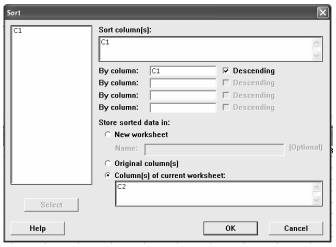
- Problema:

  Descrever estes dados numa tabela de freqüências e representá-los graficamente.

Comando *Sort*

- Os dados podem ser ordenados pelo comando *sort*, neste caso usando a ordem alfabética como chave. Pode-se ainda escolher a ordem de crescimento.

Data > Sort ➔

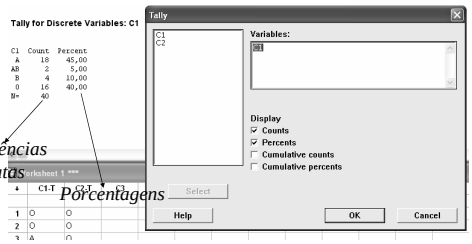


- Poderíamos contar manualmente os casos de cada tipo de sangue

Contagem de Respostas

- O comando *Tally* realiza esta tarefa para conjuntos de dados de qualquer tamanho.

Stat > Tables > Tally Individual Variables



Freqüências absolutas

Porcentagens

Gráfico de Setores

- O comando *Pie Chart* produzirá um gráfico de setores caracterizando a distribuição de freqüências das respostas em dados

Graph > Pie Chart ➔

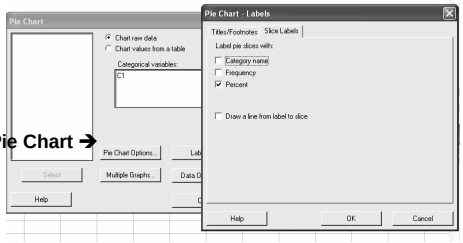
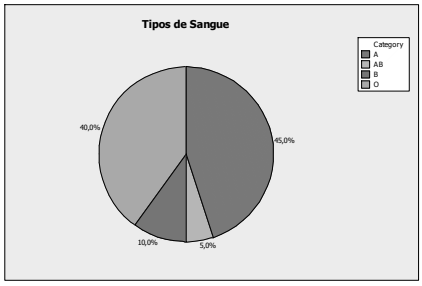
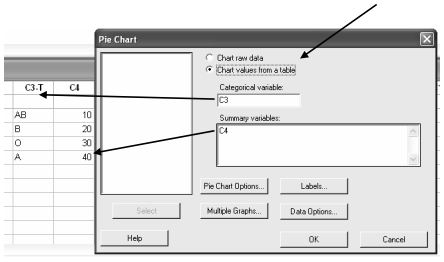


Gráfico de Setores (2)



Detalhes do Comando *Pie Chart* (1)

- Podemos criar um gráfico de setores a partir de uma tabela de freqüências especificada:



Detalhes do Comando *Pie Chart* (2)

- Podemos configurar detalhes no gráfico (cores, títulos, legendas, etc.) selecionando e editando o elemento correspondente. Exemplo, clique duas vezes no maior setor do gráfico e altere sua cor

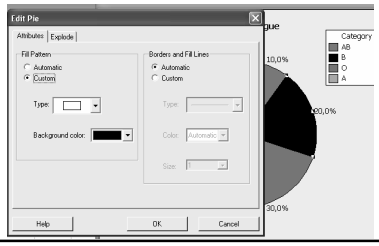


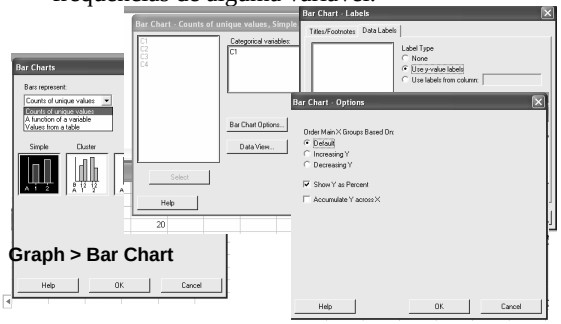
Gráfico de Setores – Comentários

- O gráfico de setores não é uma forma boa de dispor 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

- Produz um gráfico de barras da distribuição de freqüências de alguma variável.



Graph > Bar Chart

Exemplo 1 – Gráfico de Barras

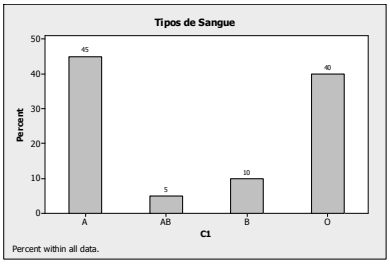
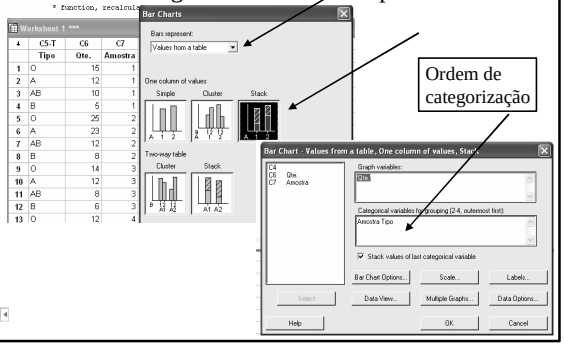


Gráfico de Barras (2)

- Suponha que em vez de uma única amostra, observamos a variável tipo sangüíneo em 4 amostras de diferentes regiões, obtendo para os sangues tipo O, A, AB e B respectivamente as seguintes freqüências:
  - √ amostra 1: 15, 12, 10, 5
  - √ amostra 2: 25, 23, 12, 8
  - √ amostra 3: 14, 12, 8, 6
  - √ amostra 4: 12, 20, 5, 10

Gráfico de Barras Empilhado

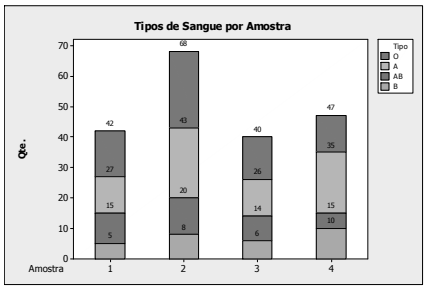
- Produz um gráfico de barras empilhado.



	C5-T	C6	C7
	Tipo	Qte.	Amostra
1	O	15	1
2	A	12	1
3	AB	10	1
4	B	5	1
5	O	25	2
6	A	23	2
7	AB	12	2
8	B	8	2
9	O	14	3
10	A	12	3
11	AB	8	3
12	B	6	3
13	O	12	4

Ordem de categorização

Gráfico de Barras (3)



Uniformização Escalas

- Cada amostra conta com um número diferente de observações e se o objetivo é comparar as diferentes amostras, o melhor é utilizar freqüências relativas para uniformizar a escala.

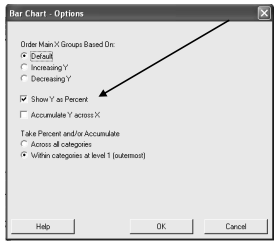


Gráfico de Barras Uniformizado

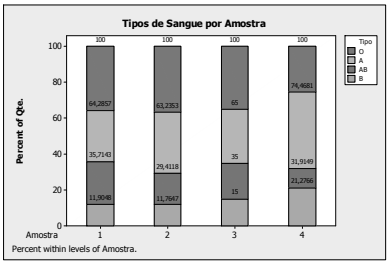


Gráfico de Barras por Categoria

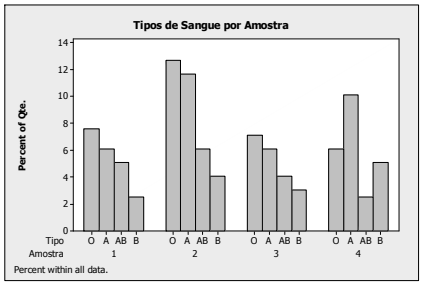
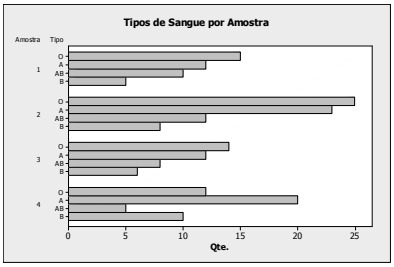


Gráfico de Barras Horizontais



Bar Chart – Scale > Axes and Ticks > Transpose value

Atividade 1 – Classes Sociais

"*Pesquisa mostra que o Brasil é classe C*".

- "Mapeamento inédito sobre o poder de compra das classes sociais brasileiras, feito pela Fundação Getúlio Vargas, mostra que 5,8 milhões das famílias estão na classe C e ganham entre R$1157 e R$2039."
- "Outras 4,6 milhões são consideradas classe D."

Classes Sociais no Brasil

- “Duas a cada três famílias nos 83 maiores municípios do país estão nas faixas de renda média e baixa.”
- “O topo da pirâmide tem apenas 27mil domicílios.”

Renda Familiar Média

Classe	renda média (R$)	Brasil	Rio de Janeiro	São Paulo	Brasília
A	26.827	0,2	0,7	0,6	0,2
A1	11.293	1,0	2,1	1,9	2,8
A2	8.313	5,0	6,8	9,6	9,0
B1	5.066	8,5	12,2	12,5	10,0
B2	3.047	13,4	22,1	16,3	12,0
C1	2.039	18,0	22,9	20,3	16,9
C2	1.157	18,0	17,3	18,8	17,0
D	621	30,7	15,1	18,4	28,1
E	282	5,3	0,8	1,6	4,0

Fonte: FGV, março/2005

- Dados armazenados na planilha classes_sociais

- Distribuição percentual das famílias por classes sociais segundo a região

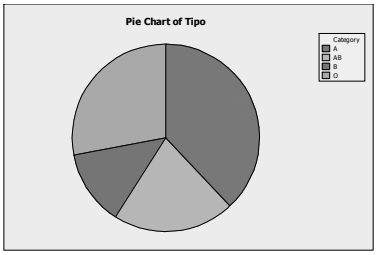
- Reagrupamento de classes do gráfico da distribuição de freqüências das classes sociais:
  - √ Para melhorar a visualização da distribuição de classes, reduzindo-las de 9 para 4 classes. Classes A, A1, A2 = A, B1 e B2 = B e C1, C2 e D = C/D.

Variáveis Quantitativas

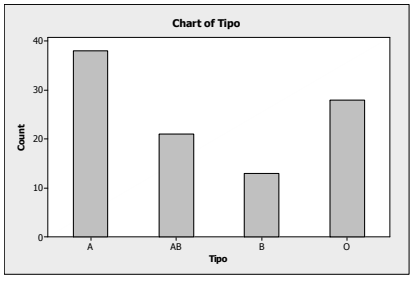
Exemplo 2

- Base de dados contendo informações biométricas de 100 indivíduos sobre tipo sangüíneo, peso (kg) e altura (cm).
- Banco de dados na planilha: biometria

Variável Qualitativa – Gráfico de Setores

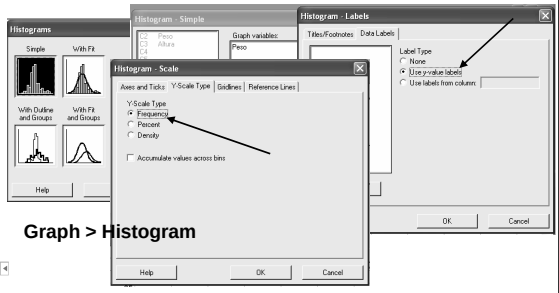


Variável Qualitativa – Gráfico de Barras



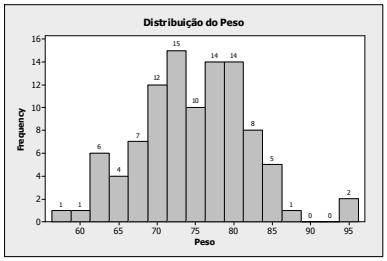
Histograma

- É importante indicador da distribuição. Usado para examinar a forma (multimodalidade, simetria, etc.) e a dispersão dos dados



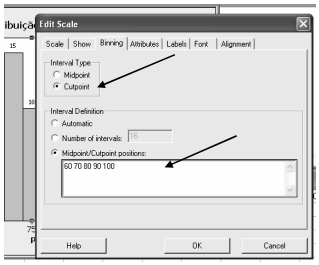
Graph > Histogram

Histograma – Peso

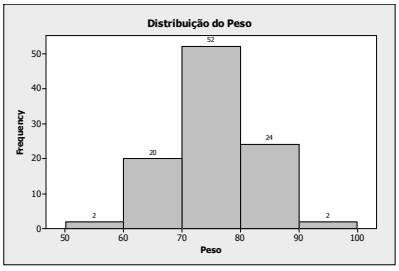


- Pode-se estabelecer os extremos das classes
  - √ Clique duas vezes no gráfico e:

Edit X-Scale > Binning



- Histograma com as amplitudes de classes escolhidas

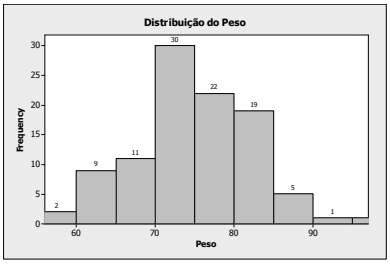


- Escolha da quantidade de intervalos de classe e escala do eixo x. Clicar 2 vezes no eixo x:

Edit X-Scale > Scale

Edit X-Scale > Binning

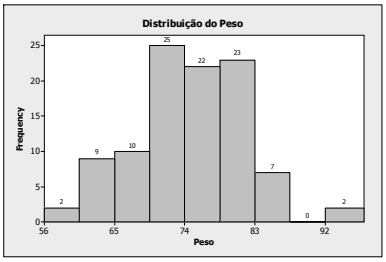
- Histograma com 9 classes



Cálculo Intervalos de Classe

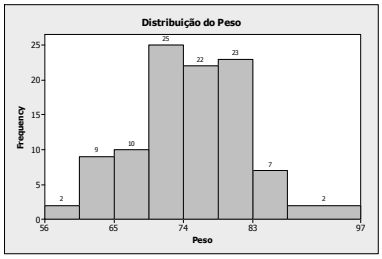
- Amplitude amostral: ˜ *96 – 57 = 39*
- Para 9 intervalos, amplitude de classe: *39/9 ˜ 4,5*
- Amplitude: *9 x 4,5 = 40,5* (˜ 2 a mais)
- Para distribuir o excesso, pode-se iniciar em *56* e concluir em *97*.
- Limites de classe:
  56; 60,5; 65; 69,5; 74; 78,5; 83; 87,5; 92,0; 96,5

- Histograma com classes calculadas



Para cutpoints (56; 60,5; 65; 69,5; 74; 78,5; 83; 87,5; 92,0; 96,5)

- Numa distribuição de freqüências não convém haver classes intermediárias vazias
  - √ Pode-se diminuir os intervalo ou agrupar as duas classes finais



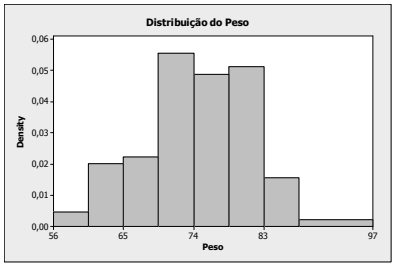
Para cutpoints (56; 60,5; 65; 69,5; 74; 78,5; 83; 87,5; 92,0; 96,5)

Densidade de Freqüência

- Pode-se representar a distribuição dos dados na escala de densidade de freqüência, definida como a razão entre a freqüência relativa e a amplitude de classe
- Não distorce a representação da distribuição quando as amplitudes de classe são desiguais
- Quando a quantidade de classes torna-se muito grande, o histograma de densidade se aproxima da função de densidade de probabilidade

- Clique duas vezes no eixo Y e

Edit Y-Scale > Type > Density

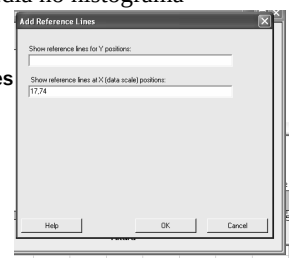


Atividade 2

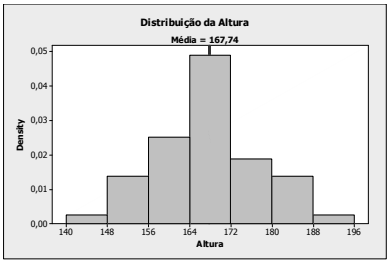
1. Construir histograma das alturas com 7 intervalos de classe
2. Calcular a média das alturas, localizando-a no histograma

- Cálculo da média:
  - √ Em Session, Editor > Enable Comando.
    No prompt do Minitab digitar: mean 'altura'
    Mean of Altura = 167,74
- Localização da média no histograma
  - √ No gráfico

Add > Reference Lines

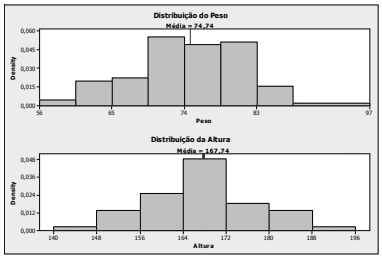


- Histograma Altura conforme solicitado:



- Gráfico com histograma Peso e Altura
  √ Selecionar gráfico e:

Editor > Layout Tool



Rows: 2
Columns: 1

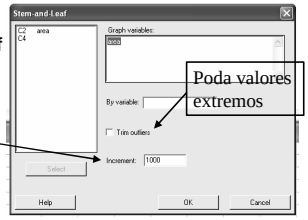
Exemplo 3 - Islands

- Banco de dados do R com áreas das maiores massas de terra do mundo (mais de 10.000 milhas quadradas), em milhares de milhas quadradas.
- Fonte: *The World Almanac and Book of Facts*, 1975, page 406.
- Problema: *Deve a Austrália ser considerada como uma ilha ou como um continente?*

Ramo e Folhas

- Usado para examinar a forma e a dispersão dos dados

Graph > Stem-and-Leaf



Poda valores extremos

Controla escala vertical (incremento entre linhas)

Ramo e Folhas – Saída

Stem-and-Leaf Display: area

```
Stem-and-leaf of area  N  = 48
Leaf Unit = 100

(41)  0   00000000000000000000000000000000000112238
 7    1
 7    2   9
 6    3   7
 5    4
 5    5   5
 4    6   7
 3    7
 3    8
 3    9   3
 2    10
 2    11  5
 1    12
 1    13
 1    14
 1    15
 1    16  9
```

Qte. na linha que contém a mediana

2.968 (Sumatra)

Qte acumulada da linha e de todas abaixo dela

Ramo e Folha de *Islands* – Saída

Stem-and-Leaf Display: area

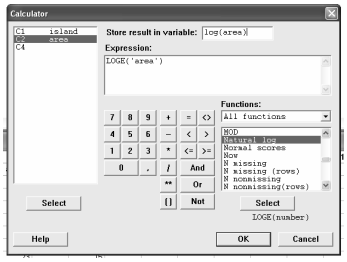
```
Stem-and-leaf of area  N  = 48
Leaf Unit = 100


(41)  0   00000000000000000000000000000000000112238
 7    1
 7    2   9
 6    3   7
 5    4
 5    5   5
 4    6   7
 3    7
 3    8
 3    9   3
 2    10
 2    11  5
 1    12
 1    13
 1    14
 1    15
 1    16  9
```

- Os valores são díspares. Em torno de 85% de valores no primeiro ramo
- Uma alternativa é trabalhar em outra escala

Mudança Escala

- Calcular *log(areas)*

Calc > Calculator



Ramo e Folha de *log(areas)* – Saída

Incremento: 1

Stem-and-Leaf Display: log(area)

```
Stem-and-leaf of log(area)  N  = 48
Leaf Unit = 0,10

 12   2  455566777779
(17)  3  01223344445677778
 19   4  024444
 13   5  22467
 8    6  7
 7    7  9
 6    8  268
 3    9  137
```

Ramo e Folha de *log(areas)* – Saída

Incremento: 0,5

Stem-and-Leaf Display: log(area)

```
Stem-and-leaf of log(area)  N  = 48
Leaf Unit = 0,10

 1   2  4
 12  2  55566777779
 22  3  0122334444
(7)  3  5677778
 19  4  024444
 13  4
 13  5  224
 10  5  67
 8   6
 8   6  7
 7   7
 7   7  9
 6   8  2
 5   8  68
 3   9  13
 1   9  7
```

Diagrama de Pontos

- Desenha um gráfico de pontos
- Para o banco de dados *islands:*

Graph > Individual Value Plot

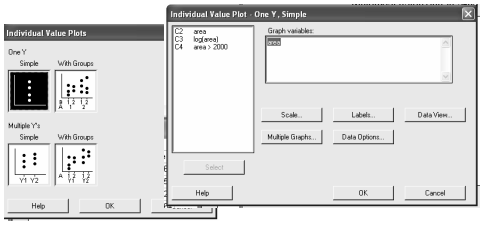
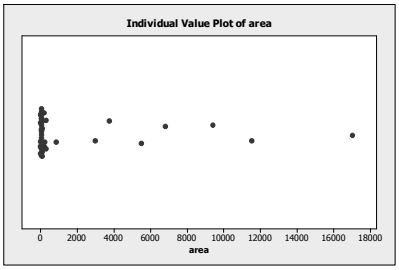


Diagrama de Pontos de *Islands*



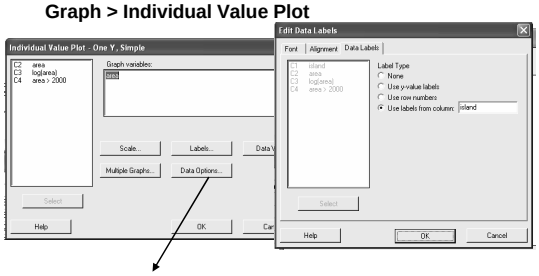
Quais são as *"Islands"* que se destacam em área?

Quais são as Areas > 2.000?

- Criar a coluna 'Areas > 2000'
- Calc > Calculator:
  - √ Store result in: 'Areas > 2.000'
  - √ Expression:
    - 'area' > 2000
- Resultado na coluna 'Area > 2.000':
  - √ 1: '*island*' com area > 2000
  - √ 0: caso contrário

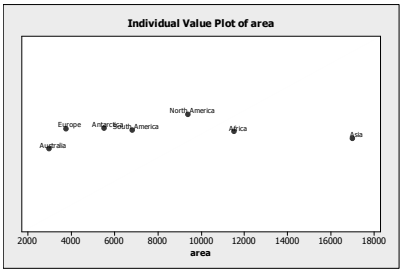
Diagrama de Pontos – Areas Maiores

Graph > Individual Value Plot



Rows that match: 'Area > 2000 = 1'

Diagrama de Pontos – Areas Maiores

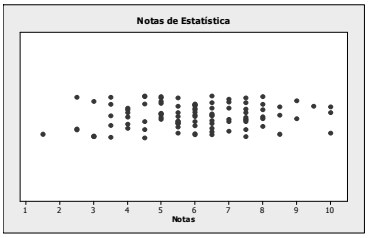


Exemplo 4 – Notas em Estatística

- O diagrama de pontos é útil também para uma avaliação de existência de algum tipo de estrutura no processo de observação dos dados
- Notas em Estatística em turma de 100 alunos
  √ Banco: notas_est

Data > Stack > Columns

Diagrama de Pontos – Notas Estatística



Empiricamente, não se percebe um padrão no registro das notas. Elas aparentam distribuir-se ao acaso.

Exemplo 5 – Temperaturas Médias

- Temperaturas médias mensais, em graus Celsius, de janeiro 1996 a dezembro 2005, em Cananéia e Ubatuba
- O termo Série Temporal refere-se a dados de uma variável quantitativa coletados ao longo do tempo

  Fonte: Boletim Climatológico, 1989 (Dados em Bussab)

Time Series – Comandos

Graph > Time Series Plot

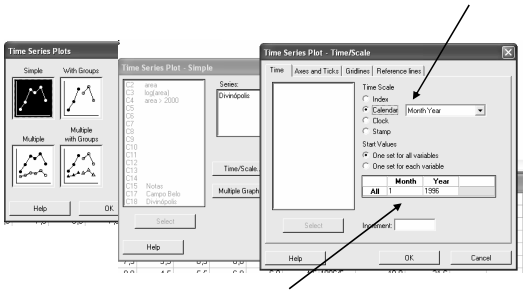
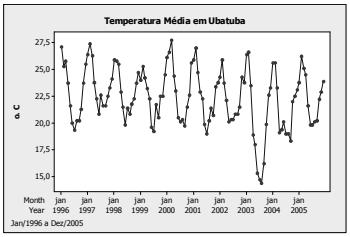


Gráfico Temperaturas



Observa-se certo padrão de comportamento das temperaturas: elas são mais altas no início dos anos, diminuem até o meio do ano, e sobem novamente.

Esse comportamento é denominado sazonal.

- Pode-se representar simultaneamente as duas séries

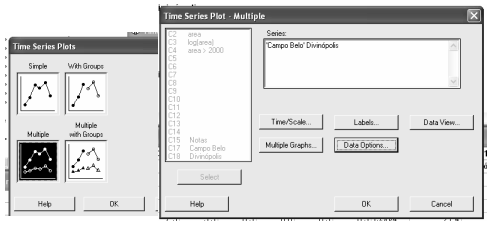
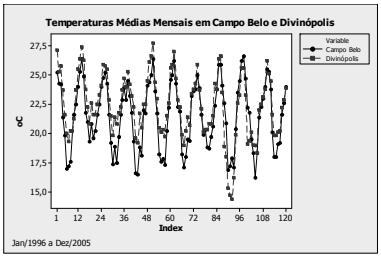
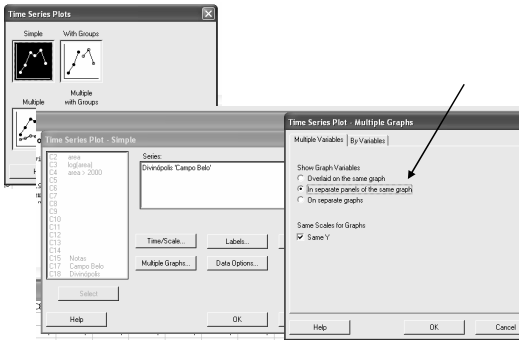


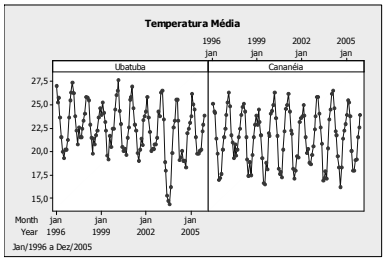
Gráfico das duas Séries



- Ou representá-las simultaneamente em painéis no mesmo gráfico

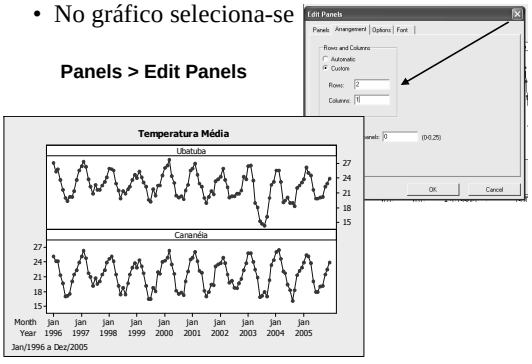


Painéis com as Séries



- Pode-se trocar a posição dos painéis:
- No gráfico seleciona-se

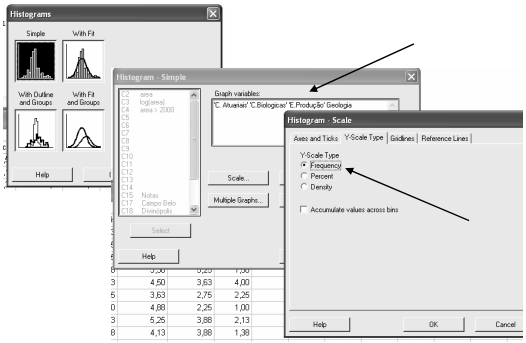
Panels > Edit Panels



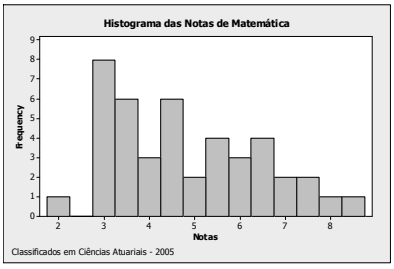
Exemplo 6 – Notas de Matemática

- Os dados estão em quatro planilhas nas quais a primeira coluna indica o gênero (masculino ou feminino), a segunda indica o ano de nascimento e, a terceira, a nota obtida em matemática.
- Os nomes das planilhas são: mat_atuaria, mat_biologia, mat_producao e mat_geo.
- Carregue-as no Minitab em uma única planilha acrescentando a coluna curso.

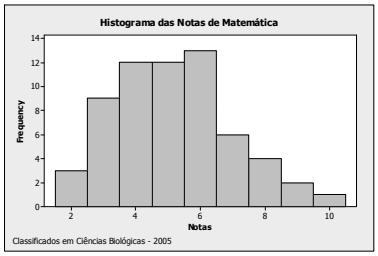
- Construção simultânea dos gráficos (separados)



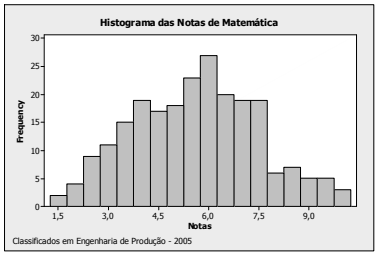
Histograma – Ciências Atuariais



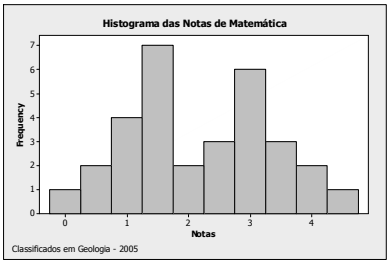
Histograma – Ciências Biológicas



Histograma – Engenharia de Produção

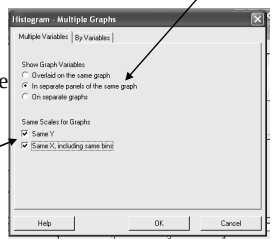


Histograma – Geologia

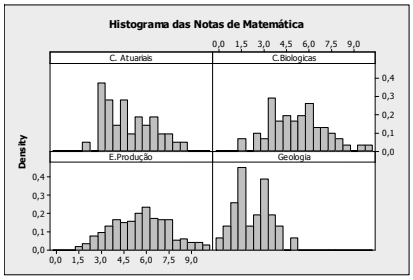


Uniformização das Escalas

- Para comparação de diversos histogramas é necessário trabalhar com as mesmas escalas
- Uniformização dos histogramas:
  - √ Escala de notas de 0 a 10
  - √ Densidade de freqüência
  - √ Mesma escala de densidade



Histograma – Painel dos Cursos



Comparação Histogramas – Comentários

- Geologia: mostraram um desempenho inferior, em comparação com os outros cursos.
- Demais cursos: variação das notas semelhante. Maior concentração ocorre em C. Atuariais.
- Produção: O comportamento das notas é aproximadamente simétrico, situação que não ocorre nos demais cursos.

Medidas de Tendência Central

Média

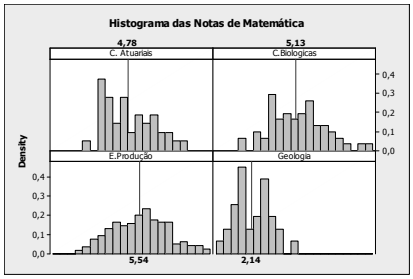
- A média é a soma dos valores observados sobre o número de observações (média aritmética).
- Pode-se obter a média na janela Session:
  √ Editor > Enable Commands
  √ mean ‘nome da coluna’

Média Candidatos- Saídas

Mean of C. Atuariais; C.Biologicas; E.Produção; Geologia

```
Mean of C. Atuariais = 4,78186
Mean of C.Biologicas = 5,12548
Mean of E.Produção = 5,53707
Mean of Geologia = 2,1387096
```

- No histograma ela representa o ponto de equilíbrio.



Média Aparada

- Calcula a média aparada, removendo 5% dos menores valores e 5% dos maiores valores
- Tenta evitar a influência de valores extremos
- Pode ser obtida através da janela Session:

  √ Editor > Enable Commands

  describe 'nome da coluna';

  trmean.

Notas Candidatos - Saída

```
MTB > describe c21 - c24;
SUBC> trmean.
```

Descriptive Statistics: C. Atuariais; C.Biologicas; E.Produção; Geologia

Variable	TrMean
C. Atuariais	4,731
C.Biologicas	5,078
E.Produção	5,517
Geologia	2,108

Notas - Médias

Variável	Média Aritmética	Média Aparada
C. Atuariais	4,78	4,73
C. Biológicas	5,13	5,08
E. Produção	5,54	5,52
Geologia	2,14	2,11

Mediana

- A mediana de uma distribuição de valores é o valor que ocupa a posição central quando os dados estão ordenados.
- Exemplo: considere o conjunto cujos valores são 11,23,14,15,16,20 e 21.
- Valores ordenados: 11,14,15,16,20,21,23

Mediana (2)

11,14,15,16,20,21,23

Valor que ocupa a posição central

Logo, a mediana deste conjunto é 16.

Mediana

- Pode-se obter a mediana pela janela Session:
  √ Editor > Enable Commands
  median 'nome da coluna'

Notas Candidatos – Medianas

```
MTB > describe c21 - c24;
SUBC> median.
```

Descriptive Statistics: C. Atuariais; C.Biologicas; E.Produção; Geologia

Variable	Median
C. Atuariais	4,500
C.Biologicas	4,940
E.Produção	5,500
Geologia	2,130

Notas – Medidas de Posição

Variável	Média Aritmética	Média Aparada	Mediana
C. Atuariais	4,78	4,73	4,50
C. Biológicas	5,13	5,08	4,94
E. Produção	5,54	5,52	5,50
Geologia	2,14	2,11	2,13

Ramo e Folha de *Islands*

Stem-and-Leaf Display: area

```
Stem-and-leaf of area  N  = 48
Leaf Unit = 100

(41)  0   00000000000000000000000000000000000000112238
 7    1
 7    2   9
 6    3   7
 5    4
 5    5   5
 4    6   7
 3    7
 3    8
 3    9   3
 2    10
 2    11  5
 1    12
 1    13
 1    14
 1    15
 1    16  9
```

Islands – Medidas de Posição

Média	1.253
Média Aparada	718
Mediana	41

Qual a melhor destas medidas para indicar a posição do conjunto de dados ?

Média e Mediana – Influência Extremos

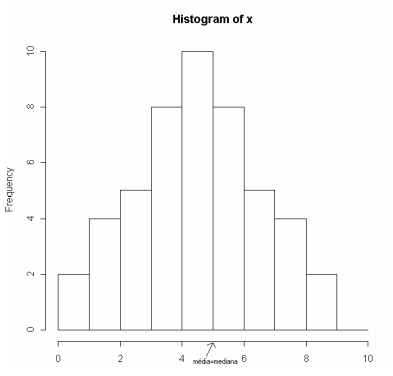
	Média	Médiana
Todos dados (48)	1.253	41
Exceto maior valor (47)	918	40
Exceto 2 maiores valores (46)	688	38
Exceto 3 maiores valores (45)	494	36

Média vs Mediana

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



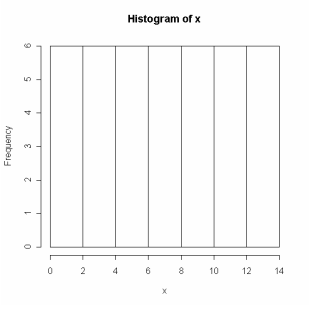
Moda

- 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)

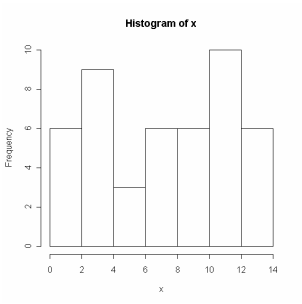
- Uma distribuição pode não possuir moda ("achatada").
- Uma distribuição pode possuir mais de uma moda (multimodal).
- Uma distribuição pode possuir apenas uma moda (unimodal).

Distribuição "Achatada"



voltar

Distribuição Multimodal

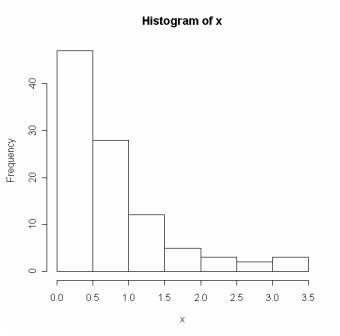


voltar

Distribuições Unimodais

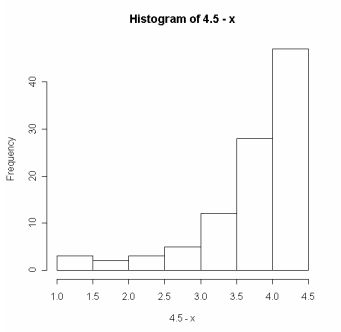
- Em distribuições unimodais tem-se sempre a mediana entre a média e a moda:
- Assimetria negativa:
  média = mediana = moda
- Assimetria positiva
  moda = mediana = média
- Perfeitamente simétricas
  média = moda = mediana.

Distribuição Unimodal – Assimetria Positiva



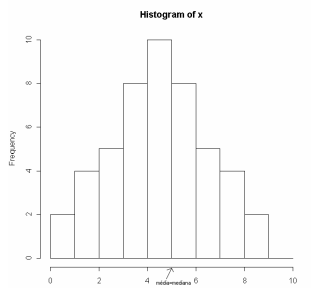
média > mediana > moda

Distribuição Unimodal – Assimetria Negativa



média < mediana < moda

Distribuição Unimodal – Simetria



média = mediana = moda

Medidas de Dispersão

Exemplo 7

- Suponha 5 conjuntos com valores variando de 0 a 10, cada um deles com 10 elementos.
- Os conjuntos estão na planilha *grupos*;
- Estes conjuntos são fictícios e têm objetivo didático.
- O objetivo é o uso de medidas para resumo de dados

Média e Mediana

- Calcule a média e a mediana de cada conjunto.

  Todos os conjuntos têm média e mediana iguais a 5

- Será que podemos afirmar que a distribuição dos dados é a mesma?

Ramo e Folhas

- Para responder a pergunta anterior, observar a variação dos dados nos diferentes conjuntos através de gráficos ramo-e-folhas.

Grupos – Ramo – e – Folhas

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
```

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

- É necessário caracterizar os grupos através de medidas que avaliem a variabilidade dos dados.
- Apresentamos as medidas de dispersão mais comuns:

Amplitude Amostral - *R*

- É a mais simples das medidas de dispersão.
- É definida como:
  Amplitude = máximo amostral – mínimo amostral
- Pode ser obtida pela janela Session:
  √ Editor > Enable Commands
  range ‘nome da coluna’.

Grupos – Amplitudes Amostrais

```
MTB > describe 'grupo_1' - 'grupo_5';
SUBC> range.
```

Descriptive Statistics: grupo_1; grupo_2; grupo_3; grupo_4; grupo_5

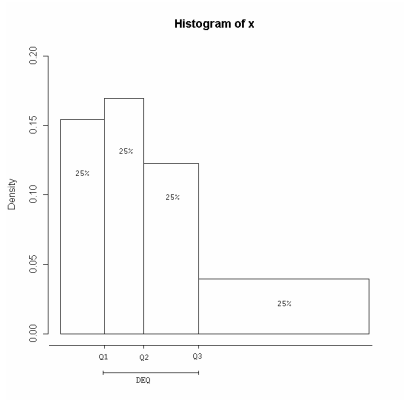
Variable	Range
grupo_1	0,000000000
grupo_2	6,000
grupo_3	2,000
grupo_4	8,000
grupo_5	4,000

Amplitude Amostral – Desvantagens

- Considera apenas os valores do mínimo e do máximo dos dados, sendo determinada por estes valores extremos.
- Ignora todo o restante da informação fornecida pela amostra.

Distância Interquartílica

- Ordena-se a amostra, dividindo-a em quatro partes com freqüência iguais.
- Tomam-se os valores do primeiro e do terceiro quartil (Q1 e Q3), os quais correspondem às freqüências relativa acumulada de ¼ e ¾
- É uma medida um pouco mais refinada que a amplitude amostral.



Minitab – *Interquartile Range*

- Pode ser obtida pela janela Session:
  √ Editor > Enable Commands
  describe ‘nome da coluna’*;*
  iqrange.

Grupos – Amplitudes Interquartílicas

Descriptive Statistics: grupo_1; grupo_2; grupo_3; grupo_4; grupo_5

```
Variable          IQR
grupo_1   0,000000000
grupo_2          6,000
grupo_3          2,000
grupo_4          4,500
grupo_5          2,000
```

Grupos - Comparação

Conjunto	Amplitude Amostral	Distância Interquartílica
Grupo 1	0,0	0,0
Grupo 2	6,0	6,0
Grupo 3	2,0	2,0
Grupo 4	8,0	4,5
Grupo 5	4,0	2,0

Distância Interquartílica – Desvantagem

- Esta medida, ainda tem a desvantagem de considerar apenas dois valores dos dados, ignorando o restante da informação fornecida pela amostra.

Desvio Médio

- É uma medida de dispersão que considera todos o conjunto de dados.
- Define-se desvio absoluto em relação à média:

$$\left| x_i - \overline{x} \right|$$

onde:

$x_1, x_2, \ldots, x_n$: valores observados

$\overline{x}$: média amostral

Desvio Médio Absoluto

- O desvio médio absoluto (DMA) é definido como a média aritmética dos desvios absolutos da média no vetor x, isto é,

$$DMA = \frac{1}{n} \sum_{i=1}^{n} | x_i - \overline{x} |$$

Cálculo do Desvio Médio Absoluto

Há várias maneiras para calculá-lo:

- Pode ser obtida pela janela Session, armazenando os resultados em constantes:

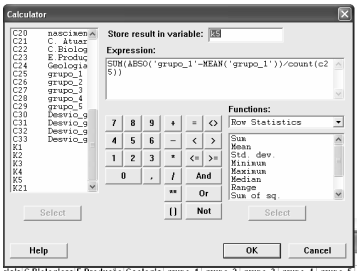
  √ Editor > Enable Commands

```
MTB > let k2=sum(ABSO('grupo_1'-MEAN('grupo_1')))/count('grupo_1')
MTB > let k21=sum(ABSO('grupo_1'-MEAN('grupo_1')))/count('grupo_1')
MTB > let k1=sum(ABSO('grupo_1'-MEAN('grupo_1')))/count('grupo_1')
MTB > let k2=sum(ABSO('grupo_2'-MEAN('grupo_2')))/count('grupo_2')
MTB > let k3=sum(ABSO('grupo_3'-MEAN('grupo_3')))/count('grupo_3')
MTB > let k4=sum(ABSO('grupo_4'-MEAN('grupo_4')))/count('grupo_4')
MTB > let k5=sum(ABSO('grupo_5'-MEAN('grupo_5')))/count('grupo_5')

MTB > print k1 k2 k3 k4 k5

Data Display

K1    0
K2    2,80000
K3    0,600000
K4    2,00000
K5    0,800000
```

Cálculo do Desvio Médio Absoluto (2)

- Armazenamento em constante através do Calculator:

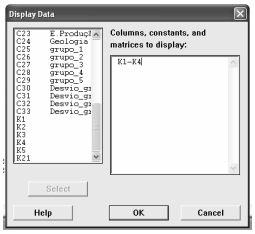
Calc > Calculator



Visualização dos Valores

- Os valores das constantes podem ser obtidos através de:

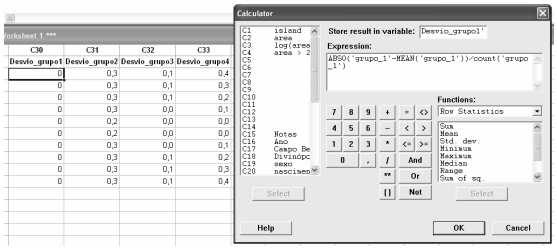
Data > Display Data



Cálculo do Desvio Médio Absoluto (3)

- Criando coluna para os desvios absolutos:

Calc > Calculator



Cálculo do Desvio Médio Absoluto (4)

- Em Session, soma-se cada coluna para se obter o desvio médio :

  √ Editor > Enable Commands :

```
MTB > sum 'Desvio_grupo1'
Sum of Desvio_grupo1
Sum of Desvio_grupo1 = 0
MTB > sum 'Desvio_grupo2'
Sum of Desvio_grupo2
Sum of Desvio_grupo2 = 2,8000
MTB > sum 'Desvio_grupo3'
Sum of Desvio_grupo3
Sum of Desvio_grupo3 = 0,6000
MTB > sum 'Desvio_grupo4'
Sum of Desvio_grupo4
Sum of Desvio_grupo4 = 2,0000
MTB > sum 'Desvio_grupo5'
Sum of Desvio_grupo5
Sum of Desvio_grupo5 = 0,8000
```

Variância Amostral

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

Variância Amostral (2)

- É a média dos desvios quadráticos em relação à média. Tem unidade diferente dos dados.
- Por questões técnicas (Inferência), adota-se *n-1* no denominador da média.
- Pode ser obtida pela janela Session:

  √ Editor > Enable Commands

  describe 'nome da coluna'*;*

  variance.

Grupos – Variância

```
MTB > describe c25 - c29;
SUBC> variance.
```

Descriptive Statistics: grupo_1; grupo_2; grupo_3; grupo_4; grupo_5

```
Variable     Variance
grupo_1   0,000000000
grupo_2         8,889
grupo_3         0,667
grupo_4         6,667
grupo_5         1,333
```

Desvio – Padrão

- É a raiz quadrada a variância.
- Pode ser obtida pela janela Session:
  - √ Editor > Enable Commands
    describe ‘nome da coluna’*;*
    stdeviation.
  - √ Para cálculo de uma coluna apenas, pode-se digitar:
    stde ‘nome da coluna

Grupos – Desvio-padrão

```
MTB > describe 'grupo_1' - 'grupo_5';
SUBC> stdeviation.
```

Descriptive Statistics: grupo_1; grupo_2; grupo_3; grupo_4; grupo_5

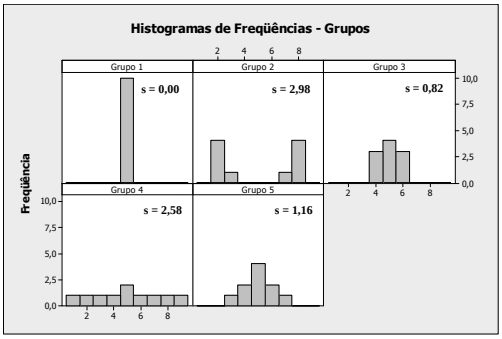
```
Variable        StDev
grupo_1   0,000000000
grupo_2         2,981
grupo_3         0,816
grupo_4         2,582
grupo_5         1,155
```

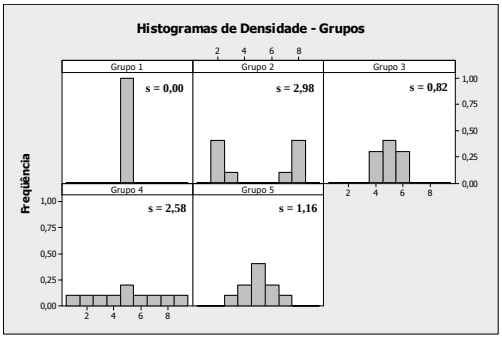
Grupos – Resumo

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)





Coeficiente de Variação Amostral

- Mede a variação relativa dos dados. É dado por:

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

onde:

desvio-padrão amostral

média amostral

- É adimensional. Em geral expresso em percentagens.
- Permite a comparação das variabilidades de diferentes conjuntos de dados.

Exemplo – Conjuntos

- Considere os seguintes conjuntos quaisquer de dados:

Conjunto 1	Conjunto 2
24	175
30	145
24	115
26	155
29	148

Disponível na planilha *conjuntos*

Conjunto – Cálculo

- Pode ser obtida pela janela Session:

√ Editor > Enable Commands

describe 'nome da coluna'*;*

cvariation.

```
MTB > Describe  'Conjunto 1' 'Conjunto 2';
SUBC>   Mean;
SUBC>   StDeviation;
SUBC>   CVariation.

Descriptive Statistics: Conjunto 1; Conjunto 2

Variable      Mean  StDev  CoefVar
Conjunto 1  26,714  2,360     8,84
Conjunto 2  147,71  17,83    12,07
```

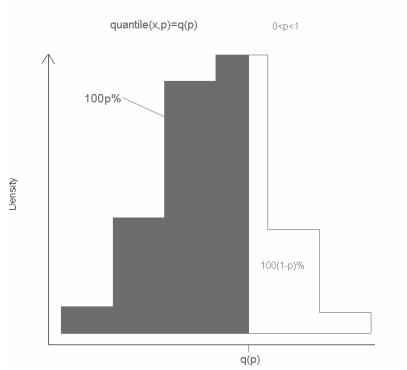
Medidas de Posição – Quantis

Quantis

- Em geral, a média e o desvio-padrão não representam completamente um conjunto de dados, pois:
  - √ são fortemente influenciados por valores extremos;
  - √ não oferecem uma idéia clara da simetria (ou assimetria) da distribuição dos dados.

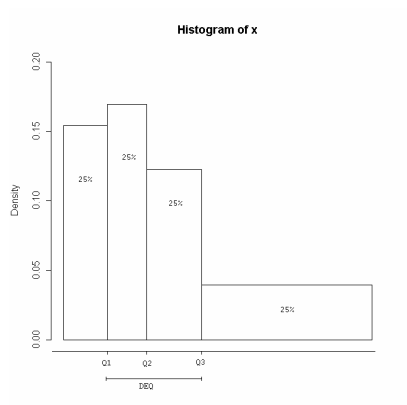
Quantis (2)

- Define-se uma medida chamada quantil de ordem *p*, com $0 < p < 1$, tal que $100 \times p\%$ das observações sejam menores do que o quantil de ordem p.
- Notação: $q(p)$



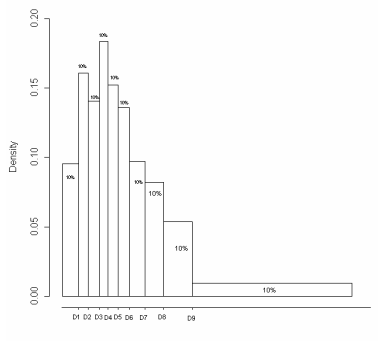
Quartis

- São três medidas (Q_1, Q_2 e Q_3) que dividem a distribuição em quatro intervalos de mesma freqüência (25%)
  - √ Q_1: primeiro quartil → *q(0,25)*
  - √ Q_2: segundo quartil ou mediana → *q(0,50)*
  - √ Q_3: terceiro quartil → *q(0,75)*



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.



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.

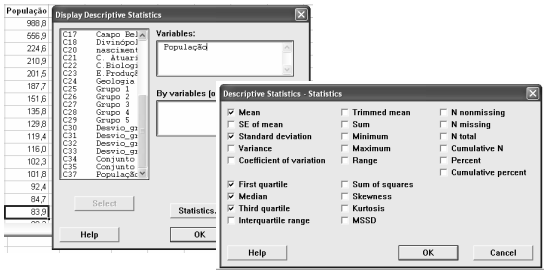
Exemplo 8 – População de Municípios

- Dados dos 30 municípios mais populosos do Brasil, em 1996

  Fonte: IBGE
- Determine os três quartis, a média e o desvio-padrão
- Dados estão na planilha *populações*

População - Comandos

Stat > Basic Statistics > Display Descriptive Statistics



População – Saída

```
MTB > Describe  'População';
SUBC>   Mean;
SUBC>   StDeviation;
SUBC>   QOne;
SUBC>   Median;
SUBC>   QThree.
```

Descriptive Statistics: População

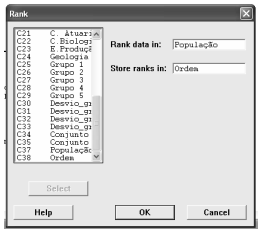
```
Variable    Mean  StDev    Q1  Median     Q3
População  145,4  186,6  63,5    84,3  139,8

MTB >
```

Estatísticas de Ordem

- Seja uma amostra $x_1, x_2, ..., x_n$
- Ordene-a de tal forma que
  - $x_{(1)}$: menor valor da amostra
  - $x_{(2)}$: segundo menor valor da amostra
  - ...
  - $x_{(n)}$: maior valor da amostra
- $x_{(1)}, x_{(2)}, ..., x_{(n)}$ são chamadas estatísticas de ordem da amostra

- As estatísticas de ordem são obtidas por:

Data > Rank



	Município	População	Ordem
1	São Paulo(SP)	988,8	30,0
2	Rio de Janeiro(RJ)	566,9	29,0
3	Salvador(BA)	224,6	28,0
4	Belo Horizonte(MG)	210,9	27,0
5	Fortaleza(CE)	201,5	26,0
6	Brasília(DF)	167,7	25,0
7	Curitiba(PR)	151,6	24,0
8	Recife(PE)	136,8	23,0
9	Porto Alegre(RS)	129,8	22,0
10	Manaus(AM)	119,4	21,0
11	Belém(PA)	116,0	20,0
12	Goiânia(GO)	102,3	19,0
13	Guarulhos(SP)	101,8	18,0
14	Campinas(SP)	92,4	17,0
15	São Gonçalo(RJ)	84,7	16,0
16	Nova Iguaçu(RJ)	83,9	15,0

Quartil – Cálculo do Minitab

- Ordenando-se os dados:
- Q_1 está na posição $\frac{n+1}{4}$

  Se esta posição não é inteira, interpola-se
- No exemplo: *(30+1)/4 = 7,75*

$$Q_1 = x_{(7)} + 0{,}75(x_{(8)} - x_{(7)}) = 62{,}89 + 0{,}75(63{,}7 - 62{,}8)$$

$$Q_1 = 63{,}47$$

- Q_3 está na posição $\frac{3(n+1)}{4}$

Função de Distribuição Empírica

- Definição mais precisa da distribuição de freqüências acumuladas:

$$F_n(x) = \frac{N(x)}{n}$$

em que:

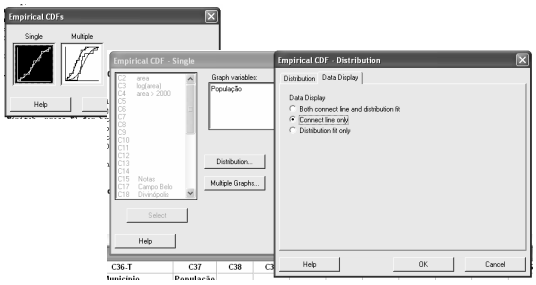
N(x): quantidade de observações = *x*

n: quantidade total de observações

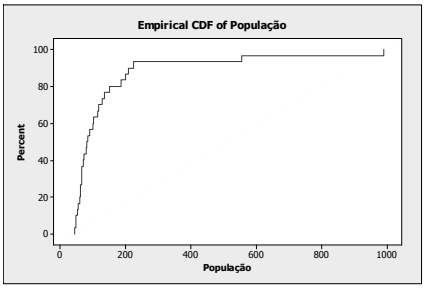
- Seu gráfico é descontínuo, com saltos de tamanho *1/n*.

Função Distribuição Empírica – Gráfico

Graph > Empirical CDF



Municípios – Função de Distribuição Empírica



Saltos de 1/30 nas ordenadas

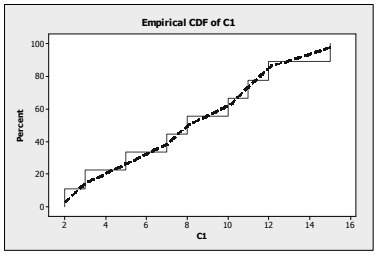
Função Distribuição Empírica “Alisada”

- Pode-se “alisar” ou suavizar uma função de modo a obter uma curva contínua
- A função empírica “alisada” é dada por:

$$\tilde{F}_n(x_{(i)}) = \frac{i-0{,}5}{n}$$

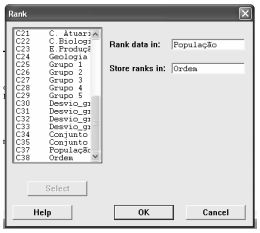
- Seu gráfico é formado pelos segmentos de reta que passam pelos pontos

Função Empírica e Alisada



Municípios – Empírica Alisada (1)

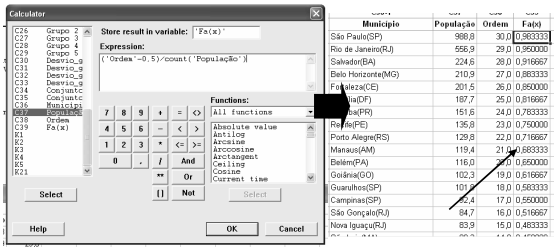
Data > Rank



	Município	População	Ordem
1	São Paulo(SP)	988,8	30,0
2	Rio de Janeiro(RJ)	566,9	29,0
3	Salvador(BA)	224,6	28,0
4	Belo Horizonte(MG)	210,9	27,0
5	Fortaleza(CE)	201,5	26,0
6	Brasília(DF)	187,7	25,0
7	Curitiba(PR)	151,6	24,0
8	Recife(PE)	135,8	23,0
9	Porto Alegre(RS)	129,8	22,0
10	Manaus(AM)	119,4	21,0
11	Belem(PA)	116,2	20,0
12	Goiânia(GO)	102,3	19,0
13	Guarulhos(SP)	101,8	18,0
14	Campinas(SP)	92,4	17,0
15	São Gonçalo(RJ)	84,7	16,0
16	Nova Iguaçu(RJ)	83,9	15,0

Municípios – Empírica Alisada (2)

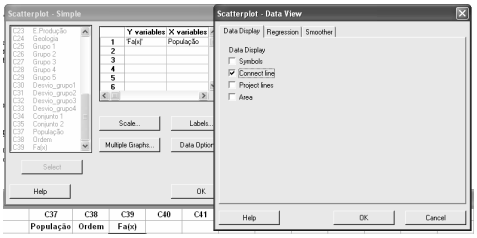
Calc > Calculator



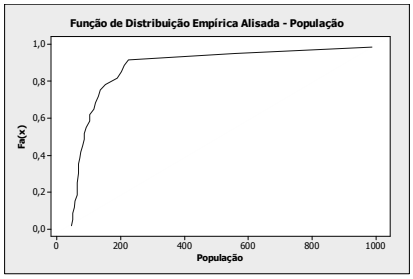
Município	População	Ordem	Fa(x)
São Paulo(SP)	988,8	30,0	0,983333
Rio de Janeiro(RJ)	556,9	29,0	0,950000
Salvador(BA)	224,6	28,0	0,916667
Belo Horizonte(MG)	210,9	27,0	0,883333
Fortaleza(CE)	201,5	26,0	0,850000
[illegible]ia(DF)	187,7	25,0	0,816667
[illegible]ba(PR)	151,6	24,0	0,783333
Recife(PE)	135,8	23,0	0,750000
Porto Alegre(RS)	129,8	22,0	0,716667
Manaus(AM)	119,4	21,0	0,683333
Belém(PA)	116,0	20,0	0,650000
Goiânia(GO)	102,3	19,0	0,616667
Guarulhos(SP)	101,9	18,0	0,583333
Campinas(SP)	82,4	17,0	0,550000
São Gonçalo(RJ)	84,7	16,0	0,516667
Nova Iguaçu(RJ)	83,9	15,0	0,483333

Municípios – Empírica Alisada (3)

Graph > Plot

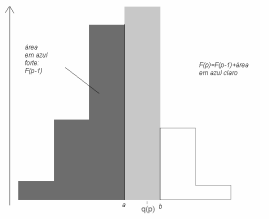


Municípios – Gráfico Empírica Alisada



Quantis para Dados Agrupados

- Passo 1: Encontre a classe que contém o p-quantil, com:
  $[a_p, a_{p+1})$: esse intervalo
  F_p: freqüência relativa acumulada desta classe.
- Passo 2: Encontre:
  c_p: comprimento desse intervalo
  f_p: freqüência relativa dessa classe
  F_{p-1}: freqüência relativa acumulada da classe anterior.
- Passo 3: Calcule *q(p)* como $q(p) = a_p + \frac{c_p}{f_p}(p - F_{p-1})$



- (a_p, a_{p+1}): classe que contém *q(p);*
- F_p: freqüência relativa acumulada dessa classe
- c_p: a amplitude dessa classe
- f_p: freqüência relativa dessa classe
- F_{p-1}: freqüência relativa acumulada da classe anterior

Quantis para Dados Não Agrupados

- Passo 1: Ordene a amostra e obtenha suas estatísticas de ordem:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Considere os pontos da forma:

$$(x_{(i)}, p_i), \text{com } p_i = \frac{i - 0.5}{n}, i = 1, \ldots, n$$

Exemplo

Se $n = 15$

$$p_i = \frac{i - 0.5}{n},\ i = 1,\ldots,n$$

i	*pi*
1	0,033
2	0,100
3	0,167
4	0,233
5	0,300
6	0,367
7	0,433
8	0,500
9	0,567
10	0,633
11	0,700
12	0,767
13	0,833
14	0,900
15	0,967

Quantis para Dados Não Agrupados (2)

- Passo 2: Determine *i* tal que: $p_i \leq p < p_{i+1}$
- Passo 3: Obtenha a reta que passa pelos pontos:

$$(x_{(i)}, p_i) \text{ e } (x_{(i+1)}, p_{i+1})$$

- Passo 4: Calcule a abscissa do ponto da reta obtida em 2, cuja ordenada é *p*.

$$q(p) \cong x_{(i)} + \frac{x_{(i+1)} - x_{(i)}}{p_{i+1} - p_i} \times (p - p_i)$$

- No exemplo dos Municípios, uma aproximação do 9º. Decil é dada por:

$$0{,}883 \cong \frac{27 - 0.5}{30} < 0.90 < \frac{28 - 0.5}{30} \cong 0{,}917 \text{ tal que } i = 27.$$

- Logo, consideramos:

  $x_{(27)} = 210{,}9$ e $x_{(28)} = 224{,}6$
- O quantil buscado é obtido por:

$$q(0{,}90) \cong 210{,}9 + \frac{224{,}6 - 210{,}9}{0{,}917 - 0{,}883} \times (0{,}90 - 0{,}883) \cong 217{,}8$$

9º. Decil – Outra Aproximação

- Ordenando-se os dados, D9 estará na posição

$$\frac{9}{10}(n+1) = \frac{9(30+1)}{10} = 27{,}9$$

- Como esta posição não é inteiro, interpola-se:

$$D_9 = x_{(27)} + 0{,}9(x_{(28)} - x_{(27)}) = 210{,}9 + 0{,}9(224{,}6 - 210{,}9)$$

$$D_9 = 223{,}23$$

- O Minitab efetua aproximações similares no cálculo dos quartis

Assimetria

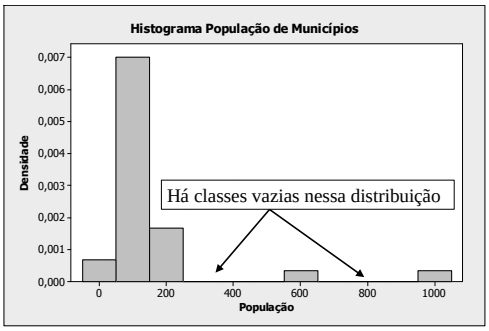
Exemplo 8 (cont.) – População de Municípios

- Dados dos 30 municípios mais populosos do Brasil, em 1996

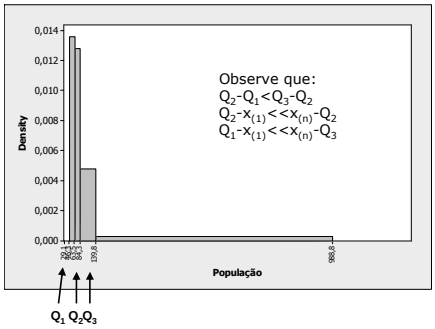
  Fonte: IBGE

Descriptive Statistics: População

Variable	Minimum	Q1	Median	Q3	Maximum
População	46,3	63,5	84,3	139,8	988,8



A distribuição dos dados apresenta forte assimetria positiva



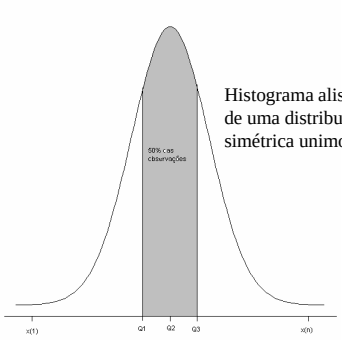
Esquema dos 5 Números

- 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)

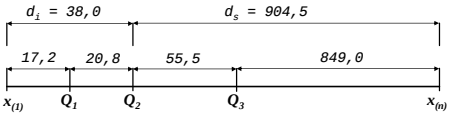
Para uma aproximadamente simétrica, tem-se:

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



Histograma alisado de uma distribuição simétrica unimodal

Municípios – Análise Complementar

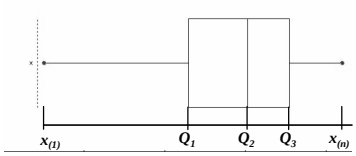


$d_i = Q_2 - x_{(1)}$: dispersão inferior

$d_s = x_{(n)} - Q_2$: dispersão superior

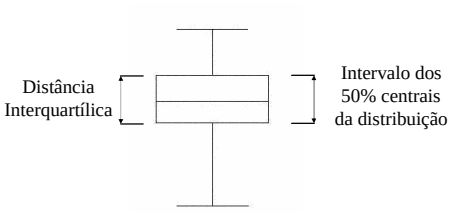
Box Plot

- A informação do esquema dos cinco números pode ser expressa num diagrama, conhecido como *box plot (gráfico-caixa).*



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.

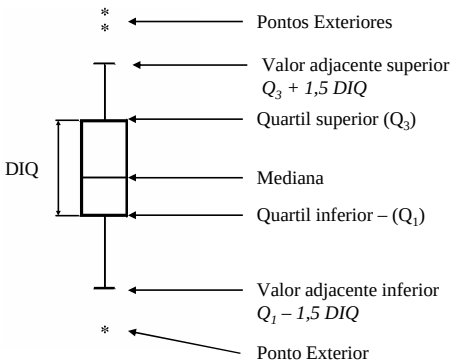


Região de Observações Típicas

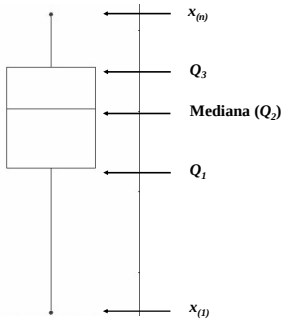
- Delimita-se a região que vai da base superior do retângulo até o maior valor observado que NÃO supere o valor de $Q_3+1,5 \times DIQ$.
- Procedimento similar para delimitar a região que vai da base inferior do retângulo, até o menor valor que NÃO é menor do que $Q_1-1,5 \times DIQ$.

Região de Observações Atípicas

- Observações são representadas por asterísticos 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.

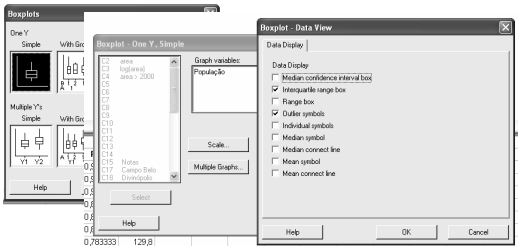


- Se não houver pontos exteriores:

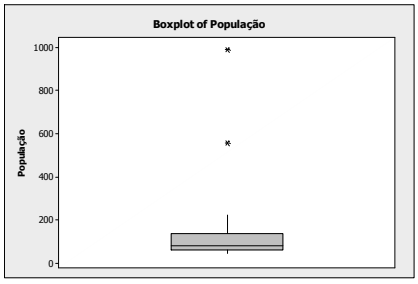


Box-plot no Minitab

Graph > Boxplot

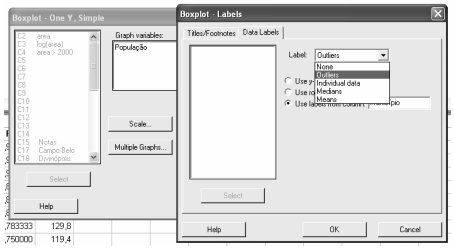


Municípios – *Box-plot*

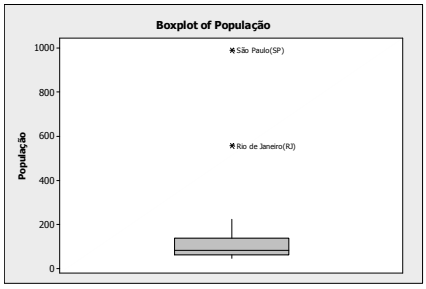


- Identificando os *outliers*

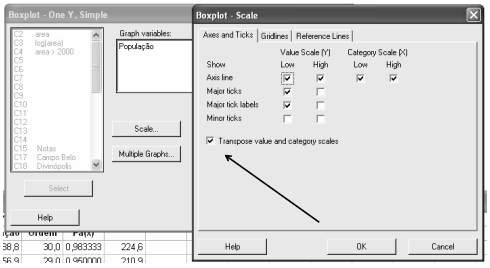
Graph > Boxplot

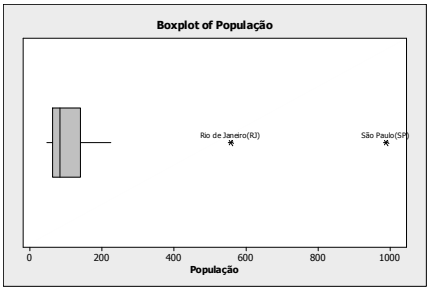


Municípios – *Outliers*



- *Box-plot* na posição horizontal





Exemplo 9 – Investimentos

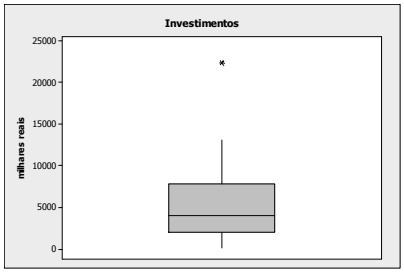
- Reportagem sobre o dinheiro da União disponível para investimentos nas prefeituras, em 2004.
- Pergunta:

  A distribuição foi justa?

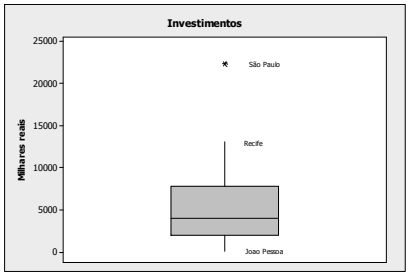
Banco de Dados

- Dados disponíveis na planilha *prefeituras*
- Variáveis:
  - √ Cidade: 25 capitais
  - √ partido (do prefeito)
  - √ hab1000: habitantes (em milhares)
  - √ invest1000: investimento (em milhares de $R)

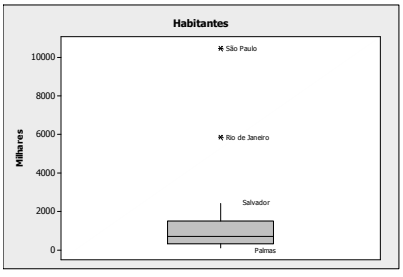
Investimentos – *Box-plot*



Detalhes dos Investimentos

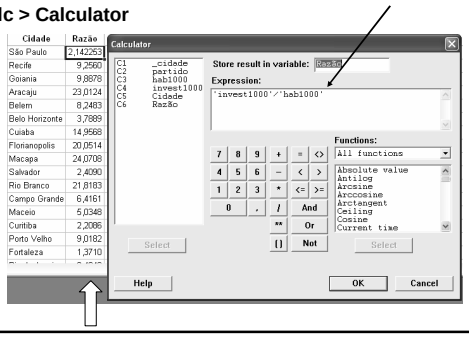


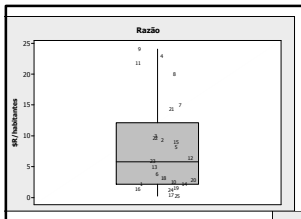
Habitantes – *Box-plot*



Razão de Investimento – Cálculo

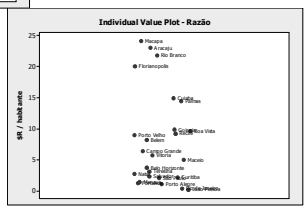
Calc > Calculator





Boxplot > Label: row number

Individual Value Plot > Label: Cidade



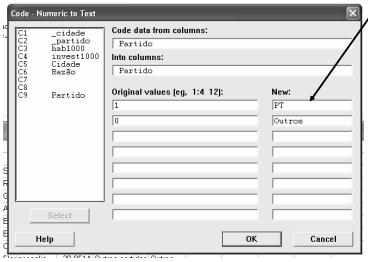
Distribuição dos Investimentos

- Criação de variável classificando partidos como: PT e Outros Partidos
  - √ Alternativa 1: Criar variável indicadora (0 e 1) através de operador lógico
    - *Editor > Enable Commands*

    Let 'Partido' = '_partido' = "PT"

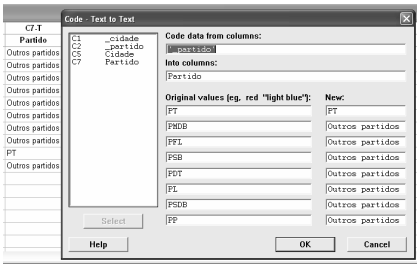
- Mudar os valores da inidcadora (0 e 1) para os valores desejados (Outros e PT)

Data > Code > Numeric to Text

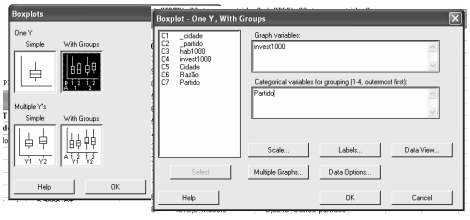


√ Alternativa 2: Criar a variável desejada modificando cada valor da variável _partido

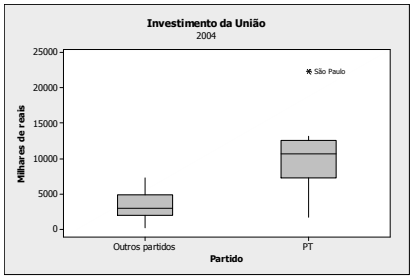
Data > Code > Text to Text



Graph > Boxplot

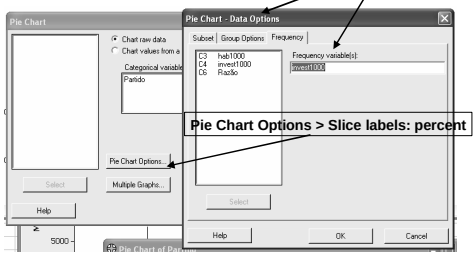


Comparação de Investimentos



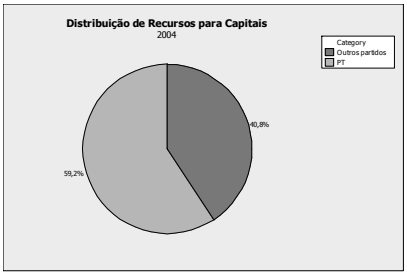
- Gráfico de percentual de variável contínua (investimento) separada por categoria (Partido)

Graph > Pie Chart



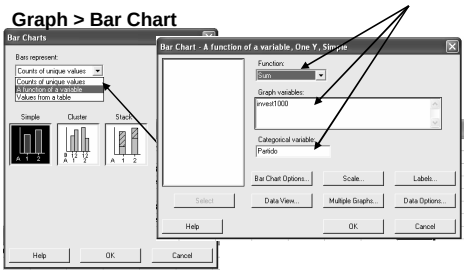
Pie Chart Options > Slice labels: percent

Comparação de Investimentos



- Gráfico de total de investimentos (variável contínua) separados por Partido(variável categórica)

Graph > Bar Chart



Comparação de Investimentos

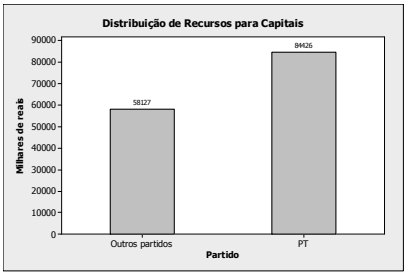


Gráfico de Quantis

- Representação gráfica dos quantis de distribuição de freqüências:
  - √ Eixo das abscissas: valores de *p*
  - √ Eixo das ordenadas: valores de *q(p)*.
- Os pontos obtidos são unidos por segmentos de retas, obtendo-se *q(p)* para todo *p*.

Exemplo 8 (cont.) – Populações Cidades

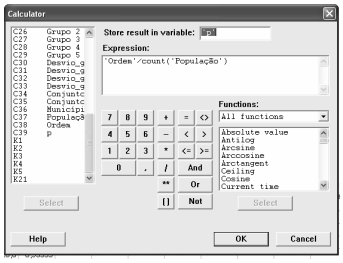
- Dados sobre os 30 municípios mais populosos do Brasil, em 1996
  √ Planilha *populações*

Construção do Gráfico de Quantis

- Determinação da ordem de cada população (efetuado anteriormente)

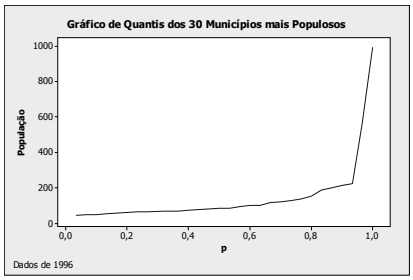
  Data > Rank

- Cálculo do percentual de municípios com população menor ou igual ao município em questão.



População	Ordem	p
988,8	30	1,000
556,9	29	0,967
224,6	28	0,933
210,9	27	0,900
201,5	26	0,867
187,7	25	0,833
151,6	24	0,800
135,8	23	0,767
129,8	22	0,733
119,4	21	0,700
116,0	20	0,667
102,3	19	0,633
101,8	18	0,600
92,4	17	0,567
84,7	16	0,533
83,9	15	0,500
80,2	14	0,467
74,7	13	0,433
72,7	12	0,400
68,4	11	0,367
66,8	9,5	0,317
66,8	9,5	0,317
63,7	8	0,267
62,8	7	0,233
61,9	6	0,200
56,2	5	0,167
54,1	4	0,133
50,3	3	0,100
49,7	2	0,067
46,3	1	0,033

Populações – Gráfico de Quantis



Graph > Scatterplot

Data View > Data Display: Connect line

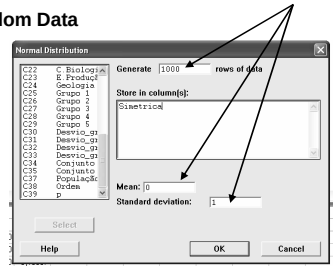
Gráfico de Quantis (cont.)

- Útil na verificação da simetria da distribuição dos dados;
- Em caso de simetria (ou aproximadamente), os pontos no topo superior direito do gráfico comportam-se de maneira semelhante aos pontos do canto inferior esquerdo.

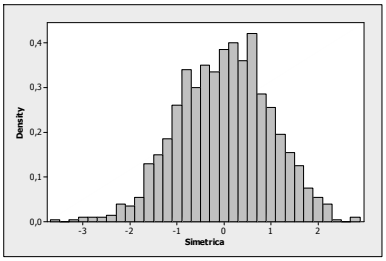
Exemplo – Gráfico de Quantis Simétrico

- Criaremos um exemplo, gerando 1.000 valores de uma normal padrão

Calc > Random Data



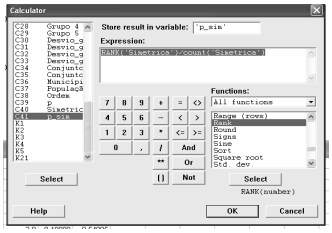
Histograma dos Números Gerados



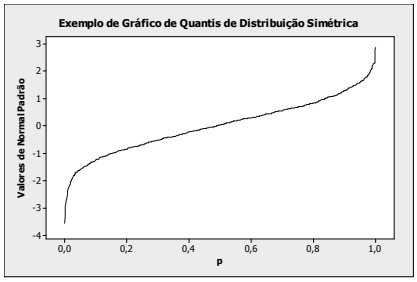
- Os resultados serão diferentes a menos que se estabeleça uma semente para a geração de números aleatórios em:

  Calc > Set Base

- Para cálculo de *p*:



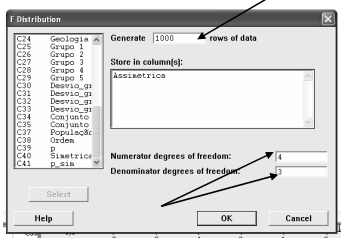
Distribuição Simétrica



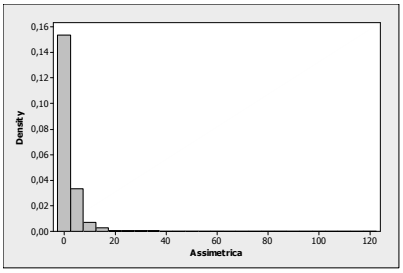
Exemplo – Gráfico de Quantis Assimétrico

- Criaremos um exemplo, gerando 1.000 valores de uma F com 4 graus de liberdade no numerador e 3 graus de liberdade no denominador

Calc > Random Data



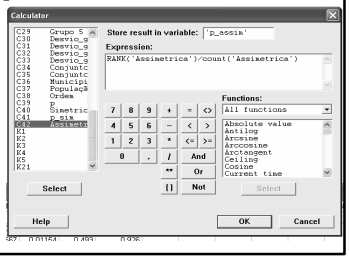
Histograma dos Números Gerados



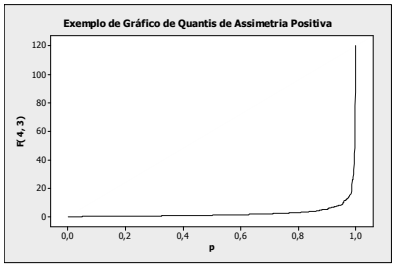
- Os resultados serão diferentes a menos que se estabeleça uma semente para a geração de números aleatórios em:

Calc > Set Base

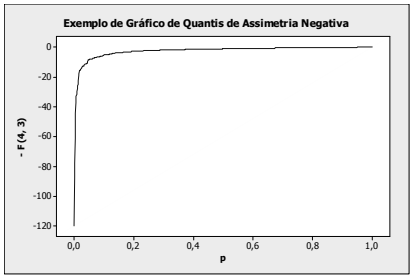
- Para cálculo de *p*:



Distribuição Assimétrica Positiva



Distribuição Assimétrica Negativa



Efetuado com o recíproco dos números gerados anteriormente

Medidas de Assimetria

- Coeficiente de assimetria de Pearson

$$sk_p = \frac{3(\bar{x} - \tilde{x})}{s}$$

em que:

: média

: mediana

s: desvio-padrão

Exemplo 8 (cont.) – Assimetria

- Dados das populações dos 30 maiores municípios brasileiros.
  √ planilha *populações*

```
MTB > describe c37

Descriptive Statistics: População

Variable    N  N*   Mean  SE Mean  StDev  Minimum     Q1  Median     Q3  Maximum
População  30   0  145,4     34,1  186,6     46,3  63,5    84,3  139,8    988,8
```

```
MTB > Let K1 = 3*(MEAN(c37)-MEDI(c37))/STDEV(c37)
MTB > describe k1
MTB > print k1

Data Display

K1    0,982719
```

Medidas de Assimetria (2)

- Uma outra medida de assimetria é dada por:

$$sk = \frac{\frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x})^3}{s^3}$$

Exemplo 8 (cont.) – Assimetria

- Dados das populações dos 30 maiores municípios brasileiros.
  √ planilha *populações*

```
MTB > Let k2 = sum(('População'-
MEAN('População'))**3)/STDEV('População')**3/COUNT('População')
MTB > print k2

Data Display

K2    3,39369
```

Medidas de Assimetria

- Válida para as duas medidas (sk e sk_p):

Medidas de assimetria (3)

- O Minitab calcula como:

$$sk^* = \frac{n}{(n-1)(n-2)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^3$$

- Pode ser obtida pela janela Session:
  √ Editor > Enable Commands
  describe 'nome da coluna'*;*
  skewness.

- Com os dados de *população*:

```
MTB > describe c37;
SUBC> skew.

Descriptive Statistics: População

Variable   Skewness
População      3,76
```

Populações – Assimetria

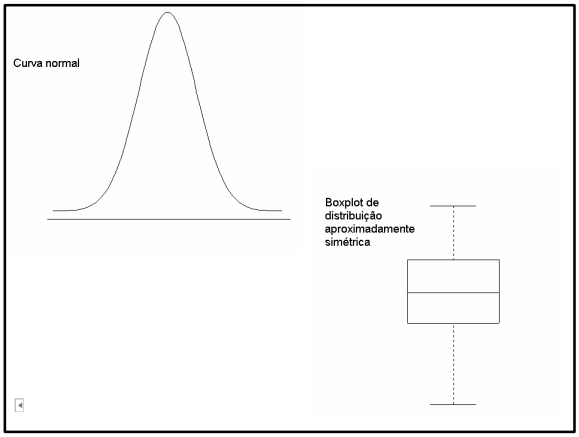
Medida	Valor
sk_p	0,98
sk	3,39
sk*	3,76

- Confirma a análise gráfica de assimétria à direita (todas as medidas são maiores que zero)

Transformações

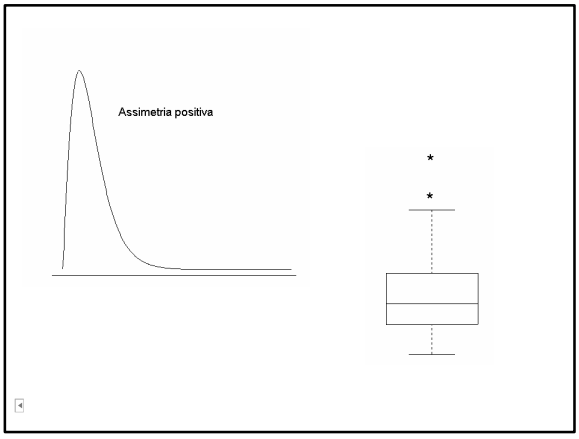
Transformações

- Muitas técnicas estatísticas baseiam-se na suposição de normalidade dos dados ou, pelo menos, de que a distribuição dos dados seja aproximadamente simétrica.



Transformações (2)

- Em muitas situações, os dados apresentam assimetria ou podem conter valores extremos (atípicos).



Transformações (3)

- Há metodologias desenvolvidas para dados não normais.
- Porém, pode-se transformar os dados no caso em que se deseja utilizar algum método para dados normais, quando os dados aparentam não ter esse comportamento;
- A transformação dos dados visa a simetrizar a distribuição.

Transformações (4)

- Uma família de transformações usada com freqüência é:

$$x^p = \begin{cases} x^p, se & p > 0 \\ \ln(x), se & p = 0 \\ -x^p, se & p < 0 \end{cases}$$

- Em geral, experimentam-se valores de p na seqüência:

  ...,-3,-2,-1,-1/2,-1/3,-1/4,0,1/4, 1/3, 1/2, 1, 2, 3, ...

Transformações (5)

- Para cada valor de p, constroem-se gráficos (histogramas, boxplots, quantis,...) para os dados originais e transformados, para escolha do valor apropriado de p.

Assimetria à Direita

- Para dados positivos, a distribuição é geralmente assimétrica à direita.
- Neste caso, sugere-se experimentar valores no intervalo $0<p<1$, pois os valores grandes de x decrescem mais em comparação com os valores menores.

Assimetria à Esquerda

- Para distribuições assimétricas à esquerda, sugerem-se valores de $p > 1$.

Exemplo 10 – Emissão

- A distribuição dos níveis de dióxido de carbono (planilha *emissão*) é assimétrica.

  √ A distribuição é assimétrica à direita

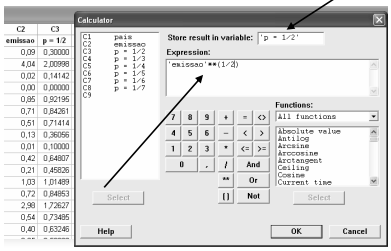
- Objetivo: Encontrar uma transformação que "simetrize" o conjunto de dados

Considerações

- Indica-se a busca de um valor de *p* entre 0 e 1.
- A transformação logarítmica não pode ser empregada pois há uma observação com emissão 0.
- Implementar transformações com os seguintes valores de *p*:
  - √ 1/2, 1/3, 1/4, 1/5, 1/6 e 1/7.

- Criar as colunas:
  - √ p = 1/2, p = 1/3, p = 1/4, p = 1/5, p = 1/6 e p = 1/7
- Calcular a transformação de potência para cada uma delas:

Calc > Calculator

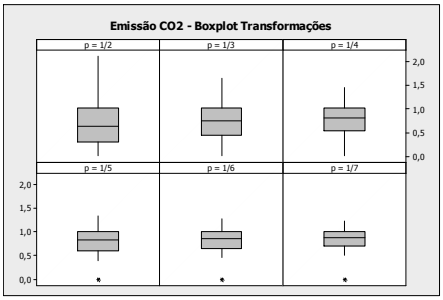


C2	C3
emissao	p = 1/2
0,09	0,30000
4,04	2,00998
0,02	0,14142
0,00	0,00000
0,85	0,92195
0,71	0,84261
0,51	0,71414
0,13	0,36056
0,01	0,10000
0,42	0,64807
0,21	0,45826
1,03	1,01489
0,72	0,84853
2,96	1,72627
0,54	0,73485
0,40	0,63246

Resultados

pais	emissao	p = 1/2	p = 1/3	p = 1/4	p = 1/5	p = 1/6	p = 1/7
Afeganistao	0,020	0,141	0,271	0,376	0,457	0,521	0,572
Albania	0,150	0,387	0,531	0,622	0,684	0,729	0,763
Algeria	0,890	0,943	0,962	0,971	0,977	0,981	0,983
Angola	0,120	0,346	0,493	0,589	0,654	0,702	0,739
Argentina	1,020	1,010	1,007	1,005	1,004	1,003	1,003
Armenia	0,270	0,520	0,646	0,721	0,770	0,804	0,829
Australia	4,430	2,105	1,642	1,451	1,347	1,282	1,237
Austria	2,010	1,418	1,262	1,191	1,150	1,123	1,105
Azerbaiao	1,540	1,241	1,155	1,114	1,090	1,075	1,064
Camboja	0,010	0,100	0,215	0,316	0,398	0,464	0,518
Camaroes	0,090	0,300	0,448	0,548	0,618	0,669	0,709
Canada	4,040	2,010	1,593	1,418	1,322	1,262	1,221
Africa Central	0,020	0,141	0,271	0,376	0,457	0,521	0,572
Chade	0,000	0,000	0,000	0,000	0,000	0,000	0,000
Chile	0,850	0,922	0,947	0,960	0,968	0,973	0,977
China	0,710	0,843	0,892	0,918	0,934	0,945	0,952
Colombia	0,510	0,714	0,799	0,845	0,874	0,894	0,908
Congo	0,130	0,361	0,507	0,600	0,665	0,712	0,747
Rep. Congo	0,010	0,100	0,215	0,316	0,398	0,464	0,518
Costa Rica	0,420	0,648	0,749	0,805	0,841	0,865	0,883
Costa Marfim	0,210	0,458	0,594	0,677	0,732	0,771	0,800
Croacia	1,030	1,015	1,010	1,007	1,006	1,005	1,004
Cuba	0,720	0,849	0,896	0,921	0,936	0,947	0,954
Rep.Theca	2,980	1,726	1,439	1,314	1,244	1,200	1,169
Equador	0,540	0,735	0,814	0,857	0,884	0,902	0,916
Egito	0,400	0,632	0,737	0,795	0,833	0,858	0,877
El Salvador	0,250	0,500	0,630	0,707	0,758	0,794	0,820
Estonia	3,020	1,738	1,445	1,318	1,247	1,202	1,171
Etiopia	0,060	0,245	0,391	0,495	0,570	0,626	0,669

Poluição – Resultados Transformação



Comentários

- Verifica-se que as transformações para *p* iguais a *1/5* e *1/4* resultaram uma distribuição aproximadamente simétrica.
- Poderíamos continuar a transformação, escolhendo um valor de *p* entre 1/5 e 1/4.
  Ex.: *p = (1/5+1/4)/2*

Exemplo 11 – Dados Brasil

- Dados sobre de superfície (km^2), população (urbana e rural) e densidade ($Hab./km^2$) das unidades federativas do Brasil, por região
- Banco de dados: planilha *brasil*
- Fonte: *IBGE, Contagem da População, 1996*.

Exemplo 11 – Dados Brasil (2)

- Objetivo:
  - √ Verifique a forma da distribuição da densidade demográfica
  - √ Proponha uma transformação buscando tornar a distribuição aproximadamente simétrica.

Transformação de Box-Cox

- Identifica automaticamente identifica uma transformação a partir de uma família de transformações potência
- A família de transformações é dada por:

Transformação de Box-Cox

- λ é um parâmetro a ser determinado a partir dos dados da amostra, através de procedimentos de máxima verossimilhança,
- Esta família inclui:

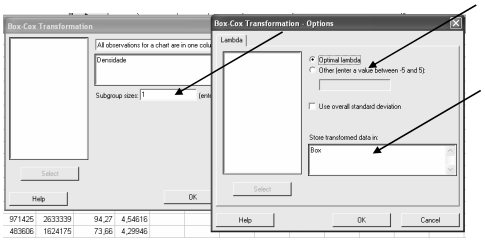
$l = 2 \Rightarrow Y' = Y^2$ $\quad l = 0{,}5 \Rightarrow Y' = \sqrt{Y}$

$l = -0{,}5 \Rightarrow Y' = \frac{1}{\sqrt{Y}}$

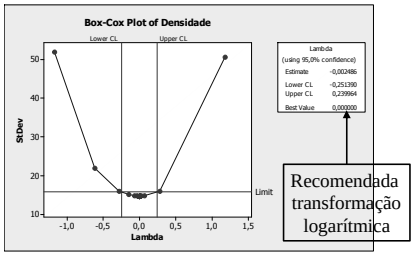
$l = 0 \Rightarrow Y' = \log_e Y$ *(por definição)* $\quad l = -1{,}0 \Rightarrow Y' = \frac{1}{Y}$

- Para uso desta transformação, no Minitab, é necessário que todos os dados sejam positivos (maiores que zero)
- Pode ser utilizado então em nosso

Stat > Control Charts > Box-Cox Transformation



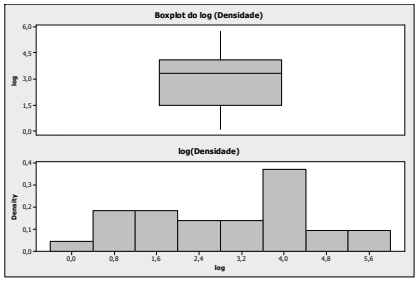
Densidade – Resultados



Recomendada transformação logarítmica

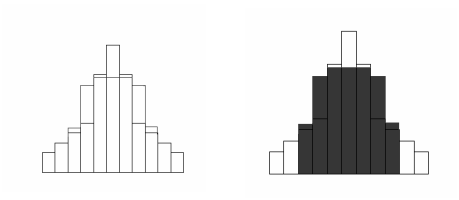
- No exemplo, a variável transformada está na coluna *Box:*

UF	Densidade	Box
RO	5,16	1,64
AC	3,16	1,15
AM	1,51	0,41
RR	1,10	0,10
PA	4,40	1,48
AP	2,65	0,97
TO	3,77	1,33
MA	15,67	2,75
PI	10,59	2,36
CE	46,53	3,84
RN	48,00	3,87
PB	58,42	4,07
PE	74,79	4,31
AL	94,27	4,55
SE	73,66	4,30
BA	22,11	3,10
MG	28,34	3,34
ES	60,69	4,11
RJ	305,32	5,72
SP	137,14	4,92
PR	45,08	3,81
SC	51,08	3,93
RS	34,17	3,53
MS	5,38	1,68
MT	2,47	0,90
GO	13,23	2,58
DF	312,94	5,75



Apesar da transformação, percebe-se forte assimetria

Curtose



- Distribuições simétricas, com mesma média e variância.
- Na vizinhança da média, apresentam densidades diferentes.

Achatamento ou Curtose

- Essas distribuições diferem quanto a um aspecto conhecido como “achatamento” ou curtose.

Medida de Curtose

- Uma medida do grau de achatamento:

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4}$$

em que:

: média

s: desvio-padrão

Curtose

- Se *k=3*, a distribuição é mesocúrtica.
- Se k<*3*, a distribuição é platicúrtica (mais achatada).
- Se *k>3*, a distribuição é leptocúrtica.

- A curtose pode ser interpretada como o quanto uma distribuição difere de uma normal.
- A curtose de uma distribuição normal é 3. Assim, para dados provenientes de uma distribuição normal a curtose deveria ser próxima a este valor.

Exemplo 10 (cont.) - Emissão

- Dados de emissão de dióxido de carbono

```
MTB > let k3 = sum(('emissao'-MEAN('emissao'))**4)/STDEV('emissao')**4/COUNT('emissao')
MTB > print k3

Data Display

K3     5,07497
```

- A distribuição é leptocúrtica ($k > 3$).

Medidas de Curtose (cont)

- O Minitab calcula como:

$$k^* = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_i \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

- Pode ser obtida pela janela Session:
  √ Editor > Enable Commands
  describe 'nome da coluna'*;*
  kurtosis.

Curtose – Minitab

Os valores terão signifcados diferentes

- Se $k = 0$, a distribuição é mesocúrtica.
- Se $k < 0$, a distribuição é platicúrtica (mais achatada).
- Se $k > 0$, a distribuição é leptocúrtica.

- Com os dados de *emissao*:

```
MTB > describe 'emissao';
SUBC> kurtosis.

Descriptive Statistics: emissao

Variable  Kurtosis
emissao       2,79
```

- A distribuição é leptocúrtica ($k > 0$).

Referências

Bibliografia Recomendada

- Bussab, W. O. e Morettin, P. A. (Saraiva) *Estatística básica*
- Montgomery, D. C. e Runger, G. C. (LTC) *Estatística aplicada e probabilidade para engenheiros*