Análise Bidimensional

Roteiro

1. Coeficiente de Correlação
2. Interpretação de *r*
3. Análise de Correlação
4. Aplicação Computacional
5. Referências

Coeficiente de Correlação

Objetivos

Análise de duas variáveis quantitativas:

- traçar diagramas de dispersão, para avaliar possíveis relações entre as duas variáveis;
- calcular o coeficiente de correlação entre as duas variáveis;
- obter uma reta que se ajuste aos dados segundo o critério de mínimos quadrados.

Exemplo 1 - Diagramas de Dispersão e Correlação

- Dados de algumas regiões metropolitanas:
  - ✓Porcentagem da população economicamente ativa empregada no setor primário
  - ✓Índice de analfabetismo

  Planilha: *analfabetismo*

  Fonte: Indicadores Sociais para Áreas Urbanas, IBGE – 1977 (Bussab)

Região	Setor Primário	Índice Analfabetismo
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

Fonte: Indicadores Sociais para Áreas Urbanas - IBGE - 1977.

Problema

- Existe alguma relação entre a porcentagem da população economicamente ativa no setor primário e o índice de analfabetismo?
- Em caso afirmativo, como quantificá-la?

- Obter o diagrama de dispersão dos dados:
  Graph > Scatter Plot > Simple

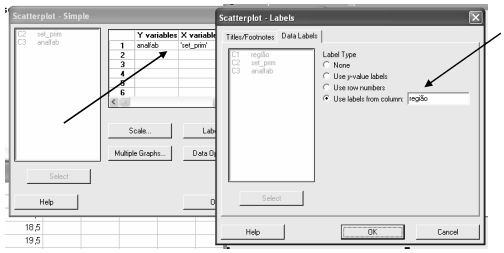
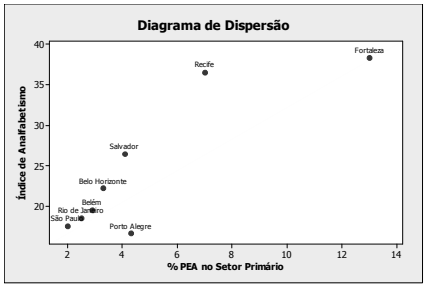


Diagrama de Dispersão



Há dependência linear entre as variáveis?

Coeficiente de Correlação

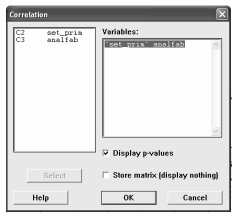
$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

- Cálculo do Coeficiente de Correlação
  - ✓Em *Session, Editor > Enable Commands*.

```
MTB > Correlation 'set_prim' 'analfab'

Correlations: set_prim; analfab

Pearson correlation of set_prim and analfab = 0,867
P-Value = 0,005
```

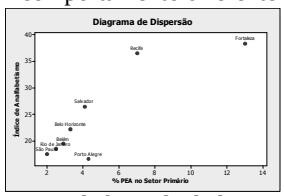
Ou através de:

Stat > Basic Statistics > Correlation



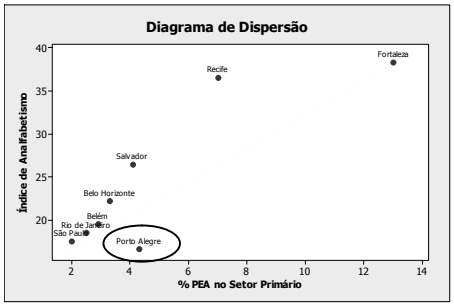
Correlação

- Há alguma região com comportamento diferente das demais?



- Em caso afirmativo, retire-a da base de dados e recalcule a correlação.

dados



Porto Alegre

- Correlação sem dados da região metropolitana de Porto Alegre
  (linha 6 da base de dados).

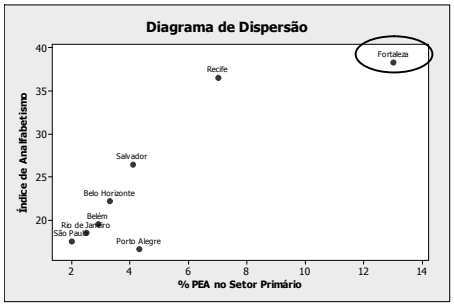
```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 6.

Correlations: set_prim; analfab

Excluding specified rows: 6
1 rows excluded

Pearson correlation of set_prim and analfab = 0,908
P-Value = 0,005
```

Porcentagem de Variação

$$100 \times \left| \frac{r_{(i)} - r}{r} \right|$$

- *r*: correlação calculada com todas as observações
- *r(i)*:correlação calculada sem a i-ésima observação.

$$100 \times \left| \frac{0{,}908 - 0{,}867}{0{,}867} \right| = 4{,}7\%$$

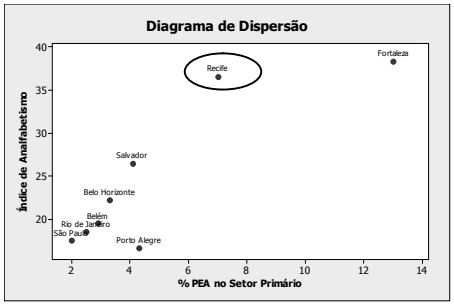


Fortaleza

- Correlação sem dados da região metropolitana de Fortaleza
  (linha 8 da base de dados).

```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 8.

Correlations: set_prim; analfab

Excluding specified rows: 8
1 rows excluded


Pearson correlation of set_prim and analfab = 0,858
P-Value = 0,013
```

porcentagem de variação em relação à correlação inicial: $100 \times \left| \frac{0{,}858 - 0{,}867}{0{,}867} \right| = 1{,}0\%$



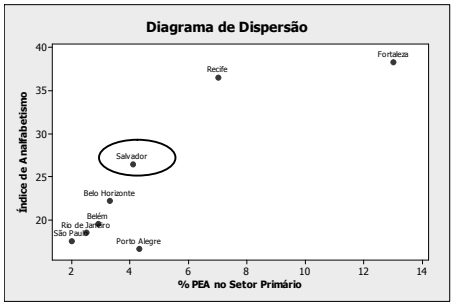
Recife

- Correlação sem dados da região metropolitana de Recife

  (linha 7 da base de dados).

```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 7.

Correlations: set_prim; analfab

Excluding specified rows: 7
1 rows excluded


Pearson correlation of set_prim and analfab = 0,916
P-Value = 0,004
```

porcentagem de variação em relação à correlação inicial: $100 \times \left| \frac{0{,}916 - 0{,}867}{0{,}867} \right| = 5{,}7\%$



Salvador

- Correlação sem dados da região metropolitana de Salvador

  (linha 5 da base de dados).

```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 5.

Correlations: set_prim; analfab

Excluding specified rows: 5
1 rows excluded


Pearson correlation of set_prim and analfab = 0,882
P-Value = 0,009
```

porcentagem de variação em relação à correlação inicial: $100 \times \left| \frac{0{,}882 - 0{,}867}{0{,}867} \right| = 1{,}7\%$

Resumo

Região Retirada	*Variação (%)*
Porto Alegre	4,8
Fortaleza	1,0
Salvador	1,7
Recife	5,7

Comentários (1)

- As regiões metropolitanas mais influentes no valor da correlação são Porto Alegre e Recife.
- Porto Alegre tem um comportamento diferente, pois sua taxa de analfabetismo é pequena comparada à sua PEA em relação às demais regiões.

Comentários (2)

- Recife tem uma taxa de analfabetismo alta comparada sua PEA com as demais regiões.
- Apesar de ser um ponto afastado dos demais, Fortaleza mantém o padrão da maioria das regiões.

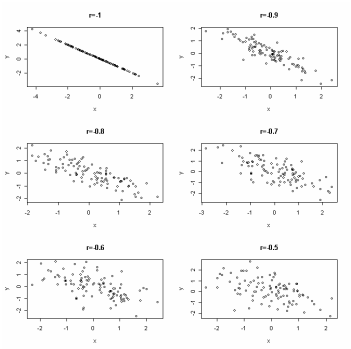
Propriedades de *r*

- Mede a intensidade de relacionamento linear
- *r* é adimensional e $-1 = r = 1$
- A conversão da escala de qualquer das variáveis não altera o valor de *r*.
- O valor de *r* não é afetado pela escolha de x ou y.

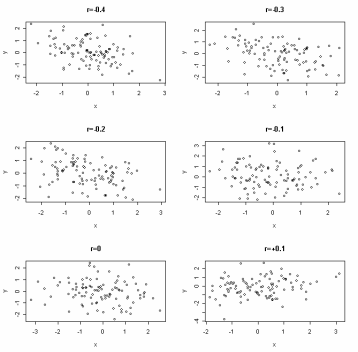
Propriedades de *r*

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

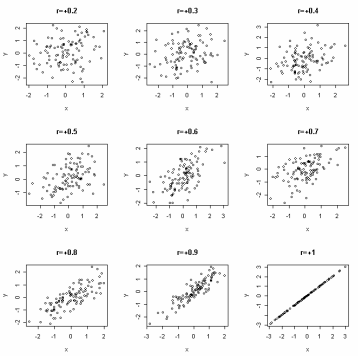
Diagramas de Dispersão (1)



Diagramas de Dispersão (2)



Diagramas de Dispersão (3)



Exemplo 2 – Relação Determinística

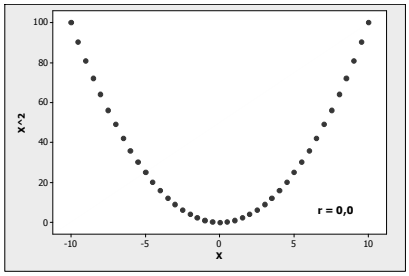
- Calcular o coeficiente de correlação entre *Y* e *X*, para $Y = X^2$, para $-10 = x = 10$
- Construção da coluna *X*:

Calc > Make Patterned Data > Simple Set of Numbers➔

First value: -10
Last value: 10
In steps of: 0,5

- Construção da coluna *Y*:
  √ Em *Session, Editor > Enable Commands.*
  *Let 'X^2' = X**2*
- Diagrama de dispersão entre *X^2* e *Y*
  √ Em *Session, Editor > Enable Commands.*
  *Plot 'X' * 'X^2'*
- Coeficiente de correlação entre *X^2* e *X*

```
MTB > correlation 'X' 'X^2'

Correlations: X; X^2

Pearson correlation of X and X^2 = -0,000
P-Value = 1,000
```



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

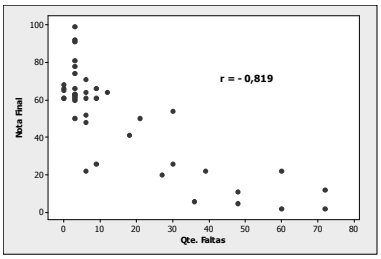
Exemplo 3 – Desempenho Acadêmico

- Dados acadêmicos da disciplina Probabilidade I
  ✓Nota final
  ✓Total de faltas
- Calcular a correlação entre elas

  Planilha: *prob1*

Correlations: Final; Faltas

```
Pearson correlation of Final and Faltas = -0,819
P-Value = 0,000
```



Correlação: Cuidados na interpretação

- Uma correlação amostral entre duas variáveis próxima de 1 ou -1 pode só indicar que as variáveis crescem no mesmo sentido (ou em sentidos contrários), e não que, aumentos sucessivos em uma, acarretarão aumentos sucessivos (ou diminuições sucessivas) na outra.

Exemplo 4 – Hábito de Fumar

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

  Planilha: *fumo*

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

Exemplo 4 – Hábito de Fumar

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

Exemplo 4 – Hábito de Fumar

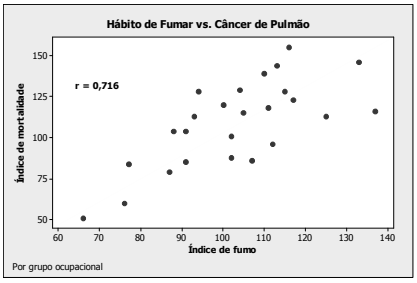
- imorte: razão da taxa de mortes sobre a taxa global de mortes (por câncer de pulmão).
  - ✓Base: 100
  - ✓imorte = 100: número médio de mortes por câncer de pulmão para o grupo é igual ao número médio global de mortes por câncer de pulmão
  - ✓imorte > 100: grupo com incidência de mortes por câncer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

Fumo vs Câncer

- Construa o diagrama de dispersão e calcule a correlação;
- Analise os dados e avalie se há relação entre os índices.

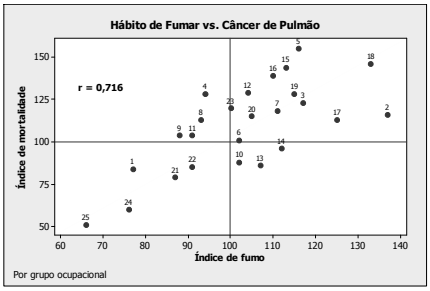
- Diagrama de dispersão entre *fumo* e *mortalidade*
  - ✓Em *Session, Editor > Enable Commands.*
    *Plot 'imorte' * 'ifumo'*
- Correlação entre *fumo* e *mortalidade*

```
MTB > Correlation imorte ifumo

Correlations: imorte; ifumo

Pearson correlation of imorte and ifumo = 0,716
P-Value = 0,000
```



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

- Identificação dos grupos de ocupação e sua posição em relação à média global:
  - ✓Selecione a figura e clique com o botão direito do mouse:
    - *Add > Data Labels > Use row numbers*
    - *Add > Reference Lines*
      - *Y positions: 100*
      - *X positions: 100*



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

Exemplo 5 – Dados de Anscombe

- O conjuntos de dados preparados para uso didático em aulas sobre correlação.
- Conjuntos 1, 2, 3 e 4 de variáveis X e Y
- Calcule a média, o desvio-padrão e o coeficiente de correlação para cada conjunto de dados

Planilha: *anscombe*

```
MTB > Correlation x1 y1 x2 y2 x3 y3 x4 y4;
SUBC> nopvalues.
```

Correlations: X1; Y1; X2; Y2; X3; Y3; X4; Y4

```
       X1      Y1      X2      Y2      X3      Y3      X4
Y1   0,816
X2   1,000   0,816
Y2   0,819   0,752   0,819
X3   1,000   0,816   1,000   0,819
Y3   0,816   0,469   0,816   0,591   0,816
X4  -0,500  -0,529  -0,500  -0,720  -0,500  -0,345
Y4  -0,314  -0,489  -0,314  -0,478  -0,314  -0,155   0,817

Cell Contents: Pearson correlation
```

```
MTB > Describe 'X1' 'Y1' 'X2' 'Y2' 'X3' 'Y3' 'X4' 'Y4';
SUBC>   Mean;
SUBC>   StDeviation.
```

Descriptive Statistics: X1; Y1; X2; Y2; X3; Y3; X4; Y4

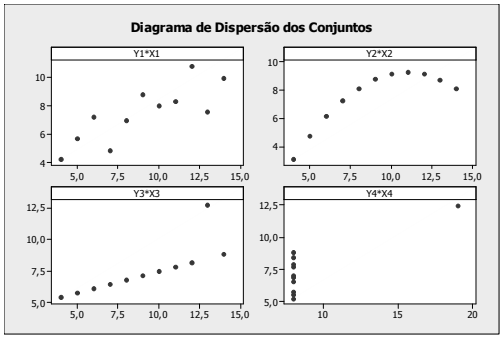
```
Variable   Mean  StDev
X1         9,00   3,32
Y1        7,501  2,032
X2         9,00   3,32
Y2        7,488  2,021
X3         9,00   3,32
Y3        7,500  2,030
X4         9,00   3,32
Y4        7,501  2,031
```

Resumo dos Dados

	X		*Y*		*r*
	$\bar{x}$	*s*	$\bar{y}$	*s*	
1	9,00	3,32	7,501	2,032	0,816
2	9,00	3,32	7,488	2,021	0,819
3	9,00	3,32	7,500	2,030	0,816
4	9,00	3,32	7,501	2,031	0,817

- Construir o diagrama de dispersão dos quatro conjuntos de dados, dispondo-os separadamente em painéis de um mesmo gráfico

Diagramas de Dispersão

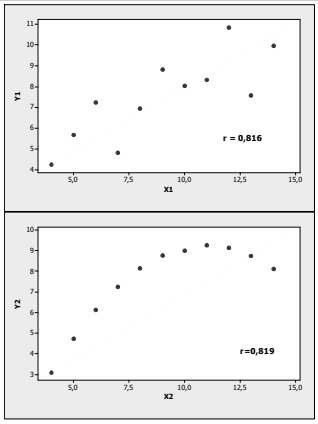


Correlação – Erros Comuns

- Linearidade:

  r mede apenas a intensidade de relações lineares

  Pode haver alguma relação entre x e y mesmo quando não há correlação linear significativa.



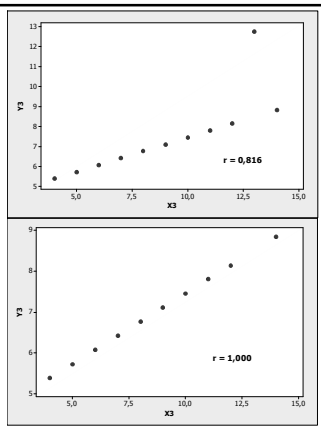
Outliers

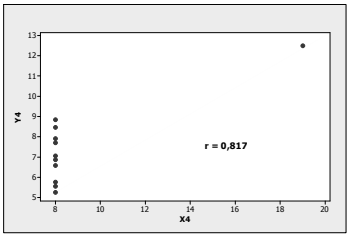
- São observações muito extremas do conjunto de dados;

  Tal como a média e o desvio-padrão, a correlação não é robusta, sendo fortemente afetadas por *outliers*

  Não devem ser descartados, a não ser que exista razão sólida;

  Utilize a correlação com cautela quando houver *outliers*: a melhor estratégia é relatar ambos os valores de *r* (com e sem o outlier)





Sem o *outlier*, não há variação em *x* e o coeficiente de correlação não pode ser calculado

Correlação – Erros Comuns

- Causalidade:

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

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

Correlação e Causalidade

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

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

  Ex.: Relação entre gastos com propaganda e vendas

  É razoável concluir que mais propaganda resulta mais vendas

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

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

  É bastante improvável que as variáveis estejam diretamente relacionadas.

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

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

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

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

Correlação e Causalidade

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

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

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

Variável Qualitativa e Quantitativa

Variável Qualitativa *vs.* Quantitativa

Objetivo:

- representar graficamente as duas variáveis combinadas;
- definir e calcular uma medida de associação entre as variáveis.

Exemplo 6 – Dados de Empregados

- Dados sobre estado civil, grau de instrução, número de filhos, salário (fração SM), idade (anos e meses) e procedência funcionários de empresa

  Planilha: *ciaMB*

  Fonte: Bussab e Morettin, Cap. 2

Variáveis

- ecivil:

  níveis: solteiro ou casado (variável nominal)
- instrucao:

  níveis: F(Ensino Fundamental), M(Ensino Médio) e S(Ensino Superior) – (variável ordinal)
- nfilhos:

  número de filhos (apenas casados). Informação omitida para os solteiros

Variáveis (2)

- sal:

  salário expresso como fração do salário mínimo
- idadea:

  idade em anos completos
- idadem:

  meses
- rp: região de procedência

  níveis: interior, capital e outros

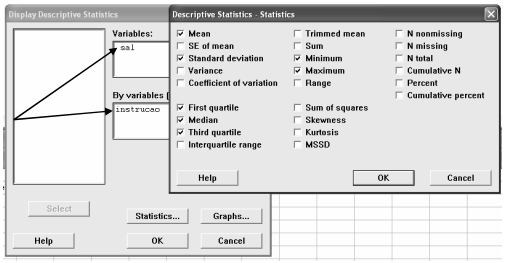
Análise de Salários

- Objetivo:

  Analisar o comportamento dos salários dentro de cada nível de instrução
  - ✓Análise de medidas resumo
  - ✓Análise gráfica

- Medidas Resumo por nível de instrução:

Stat > Basic Statistics > Display Descriptive Statistics



- Saída do Minitab

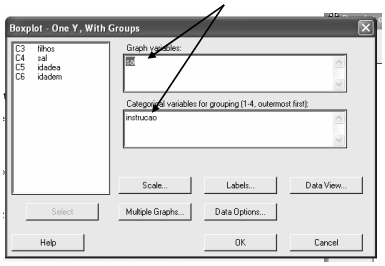
Descriptive Statistics: sal

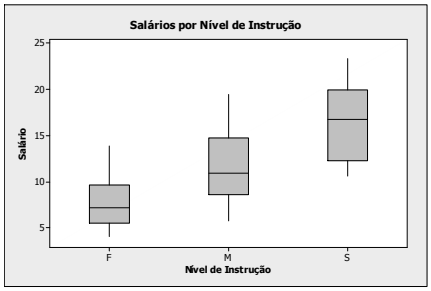
Variable	instrucao	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
sal	F	7,837	2,956	4,000	5,503	7,125	9,588	13,850
	M	11,528	3,715	5,730	8,585	10,910	14,695	19,400
	S	16,48	4,50	10,53	12,23	16,74	19,89	23,30

As medidas de posição crescem com o aumento do nível de instrução.

- Gráfico Salário por Nível de Instrução:

Graph > Box plot > With Groups





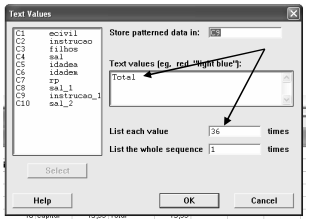
- Deseja-se inserir um box-plot para os salários globais

  1- Criar coluna dos *sal_1* duplicando todos os salários:

  ✓Em *Session, Editor > Enable Commands*.
  Stack 'sal' 'sal' 'sal_1'.

2. Criar coluna dos *inst_1* com os níveis de instrução e Total (no conjunto repetido de salários) :
   - ✓ Criar a classificação Total para um grupo de 36 salários

Calc > Make Patterned Data > Text Values

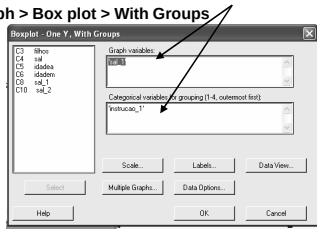


- ✓ Concluir coluna com os níveis de instrução de 36 salários e a classificação Total para um grupo repetido dos 36 salários

  Em *Session, Editor > Enable Commands*. *Stack 'instrucao' 'instrucao_1' 'instrucao_1'.*

3. Gráfico agregando box-plot Total:

Graph > Box plot > With Groups

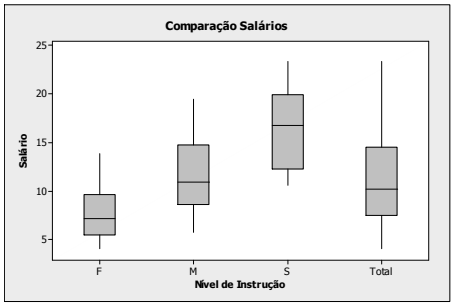


- Medidas resumo comparando-se com o Total

Descriptive Statistics: sal_1

Variable	instrucao_1	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
sal_1	F	7,837	2,956	4,000	5,503	7,125	9,588	13,850
	M	11,528	3,715	5,730	8,585	10,910	14,695	19,400
	S	16,48	4,50	10,53	12,23	16,74	19,89	23,30
	Total	11,122	4,587	4,000	7,478	10,165	14,480	23,300

- Box-plots para comparação global de salários



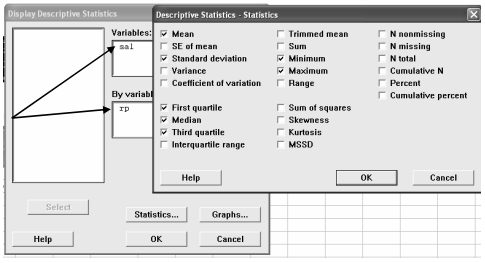
Comentário

- É possível perceber, a partir destes dados e gráficos, uma dependência entre salário e nível de instrução:

  o salário tende a ser maior quanto maior for o nível de escolaridade do empregado.

- Medidas Resumo por região de procedência:

Stat > Basic Statistics > Display Descriptive Statistics



- Saída do Minitab

Descriptive Statistics: sal

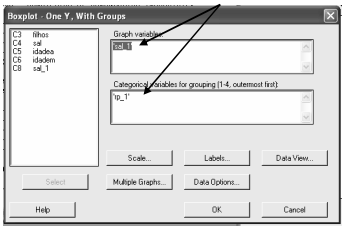
Variable	rp	Mean	StDev	Minimum	Q1	Median	Q3	Maximum
sal	capital	11,46	5,48	4,56	7,39	9,77	17,26	19,40
	interior	11,55	5,30	4,00	7,18	10,65	14,71	23,30
	outra	10,445	3,145	5,730	8,090	9,800	13,195	16,220

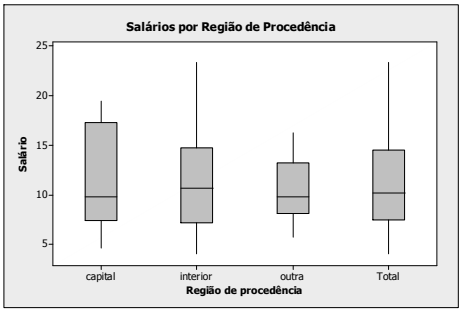
Parece não haver uma relação bem definida entre salário e região de procedência.

.

- Gráfico Salário por Região de Procedência, comparando com Total:
  - ✓Criar coluna *rp_1* com a região de procedência dos 36 empregados, repetindo a classificação Total

Graph > Box plot > With Groups





Comentários

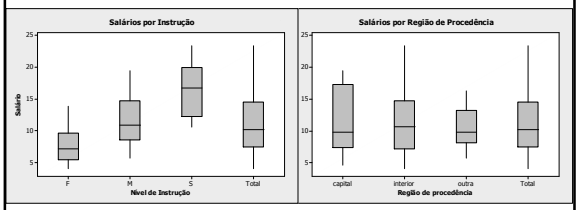
- Da análise percebe-se que não há uma relação bem definida entre salário e procedência.
- Os salários parecem estar mais relacionados com o nível de instrução do que com a região de procedência.

Quantificação de Dependência entre Variáveis

- Duas variáveis quantitativas:
  Correlação.
- Duas variáveis qualitativas:
  Qui-quadrado.
- E no caso de uma variável qualitativa e uma quantitiativa?

Medida Dependência: Quantitativa vs Qualitativa

- Definir uma medida de associação entre as variáveis usando as variâncias dentro de nível de resposta da variável qualitativa e a variância global;
- Caso a variância em cada nível de resposta for menor do que a global, então a variável qualitativa melhora a capacidade de previsão da variável quantitativa, existindo uma relação entre ambas variáveis.



Nível de instrução parece melhorar a capacidade de previsão de salário

Região de procedência não parece melhorar esta capacidade.

- Salários vs. Escolaridade:
  - ✓Cálculo das variâncias:

```
MTB > Describe  'sal_1';
SUBC>   By  'instrucao_1';
SUBC>   Variance.
```

Descriptive Statistics: sal_1

Variable	instrucao_1	Variance
sal_1	F	8,741
	M	13,802
	S	20,27
	Total	21,045

As variâncias DENTRO de cada nível são menores que a variância global

- Salários vs. Região de Procedência:
  - ✓Cálculo das variâncias:

```
MTB > Describe  'sal_1';
SUBC>   By  'rp_1';
SUBC>   Variance.
```

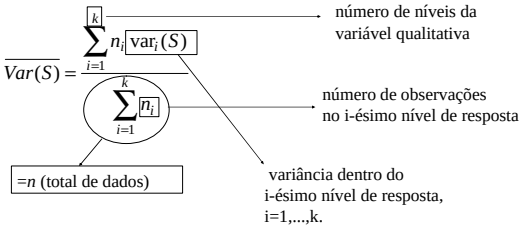
Descriptive Statistics: sal_1

Variable	rp_1	Variance
sal_1	capital	29,99
	interior	28,05
	outra	9,894
	Total	21,045

As variâncias DENTRO de cada nível não são menores que a variância global

Medida de Associação

- Utiliza-se a média das variâncias, ponderada pelo número de observações em cada nível:



Nos exemplos: k = 3
instrução (F,M,S) e região de procedência (capital,interior,outra).

- A variância média será comparada com a variância global
- x_{ij}: salário do j-ésimo indivíduo do i-ésimo nível de instrução, i=1,2,3 e j=1,...,ni
- n_i: total de indivíduos no nível *i*,
- $Var_i(S)$: variância dentro do i-ésimo nível

$$Var_i(S) = \frac{1}{n_i} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$$

$\bar{x}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$ média de salário para o nível de escolaridade i.

Variância Global

$$Var(S) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$$

$n = \sum_{i=1}^{k} n_i$ número total de observações

$\bar{x} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$ média global.

Variância Ponderada

$$\overline{Var(S)} = \frac{1}{n}\sum_{i=1}^{k} n_i \overbrace{\sum_{j=1}^{n_i}\frac{1}{n_i}(x_{ij}-\bar{x}_i)^2}^{Var_i(S)} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2$$

Relação entre as Variâncias

$$Var(S) = \underbrace{\frac{1}{n}\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}_{\geq 0} + \overline{Var(S)}$$

tal que

$$\overline{Var(S)} \leq Var(S)$$

Decomposição de Somas de Quadrados

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2}_{\text{variação total}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}_{\text{variação devido aos grupos}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)^2}_{\text{variação residual}}$$

SQTotal | **SQExplicada** | **SQResiduos**

Medida de Associação

- O grau de associação entre as duas variáveis pode ser definido como o ganho relativo na variância, obtido com a variável qualitativa.
- A medida é baseada na decomposição de somas de quadrados.

$$\frac{\text{variação devida aos grupos}}{\text{variação total}} = 1 - \frac{\text{variação residual}}{\text{variação total}}$$

R^2

$$R^2 = \frac{\overbrace{Var(S) - \overline{Var(S)}}^{\text{variação devida aos grupos}}}{Var(S)} = 1 - \frac{\overline{Var(S)}}{Var(S)}$$

Variação residual

Variação total

- Média das variâncias próxima da variância global: grau de associação pequeno
- Média das variâncias bem menor que variância global: grau de associação grande.
- Quanto mais próximo de 1 for o valor de R2, maior a associação.

R^2

- $0 = R^2 = 1$
- O símbolo R^2 é usual em análise de variância e regressão, tópicos que vão ser abordados nas disciplinas Análise de Regressão e Planejamento de Experimentos.

Salários *vs* Instrução

- Saída do Minitab:

```
MTB > Describe  'sal_1';
SUBC>   By  'instrucao_1';
SUBC>   Variance.
```

Descriptive Statistics: sal_1

Variable	instrucao_1	Variance
sal_1	F	8,741
	M	13,802
	S	20,27
	Total	21,045

Salário x Escolaridade – Cálculo de R^2

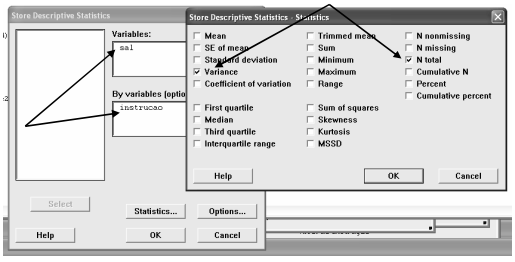
- Variância Global: $Var(S) = \frac{35(21,045)}{36} = 20,4602$
- Variância grupos:
  - ✓F $Var_F(S) = \frac{11(8,741)}{12} = 8,0123$
  - ✓M $Var_M(S) = \frac{17(13,802)}{18} = 13,0355$
  - ✓S $Var_S(S) = \frac{5(20,270)}{6} = 16,8933$
- Variância média: $\overline{Var(S)} = \frac{12(8,0123)+18(13,0355)+6(16,8933)}{36} = 12,0041$
- R^2: $R^2 = \frac{Var(S) - \overline{Var(S)}}{36} = \frac{20,4602 - 12,0041}{20,4602} = 0,4133$

Diz-se que 41,33% da variação total do salário é explicada pela variável instrução.

Usando o Minitab para o Cálculo

- Armazenamento das variâncias por grupo

Stat > Basic Statistics > Store Descriptive Statistics



Saídas Minitab

- Colunas Geradas:

```
MTB > print c11 - c13

Data Display

Row  ByVar1  Variance1  Count1
  1  F          8,7407      12
  2  M         13,8023      18
  3  S         20,2720       6
```

- Variância Média e Variância Global

$$\overline{Var(S)} = \frac{11\,Var_F(S) + 17\,Var_M(S) + 5\,Var_S(S)}{36}$$

```
MTB > let k1 = sum((Count1-1)*Variance1)/count(sal)
MTB > print k1

Data Display

K1     12,0041
```

Variância média

```
MTB > let k2 = ((count(sal)-1)/count(sal))*(stdev(sal)**2)
MTB > print k2

Data Display

K2     20,4602
```

Variância global

Tabela para Cálculo de R^2

	F	*M*	*S*	*Total*	
Soma simples	94,04	207,51	98,85	400,40	$\sum_i x_i$
Soma de quadrados simples	833,11	2.626,88	1.729,91	5.189,90	$\sum_i x_i^2$
n_i	12	18	6	36	
Média	7,84	11,53	16,48	11,12	$\overline{x}$
Soma de quadrados corrigida	96,15	234,64	10,36	736,57	$\sum_i (x_i - \overline{x})^2$

Variação Residual → 432,15

Variação Total → 736,57

Cálculo Tabela pelo Minitab

- Cálculo parcial das linhas

```
MTB > Describe sal;
SUBC> By instrucao;
SUBC> Sums;
SUBC> SSQ;
SUBC> counts;
SUBC> mean.

Descriptive Statistics: sal

                    Total                          Sum of
Variable  instrucao  Count     Mean        Sum   Squares
sal       F            12    7,837     94,040   833,108
          M            18   11,528    207,510  2626,884
          S             6    16,48      98,85   1729,91
```

Cálculo Tabela pelo Minitab

- Cálculo da linha final:

$$\sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2$$

é a variância x por $(n_i - 1)$

```
MTB > Describe  'sal_1';
SUBC>   By  'instrucao_1';
SUBC>   Variance.
```

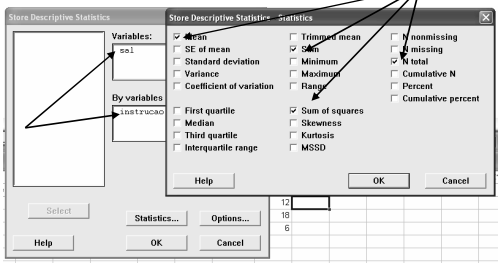
Descriptive Statistics: sal_1

Variable	instrucao_1	Variance
sal_1	F	8,741
	M	13,802
	S	20,27
	Total	21,045

Cálculo Tabela pelo Minitab

- Armazenamento dos valores:

Stat > Basic Statistics > Store Descriptive Statistics



- Cálculo da Soma dos Quadrados Corrigida

```
MTB > Let 'SQCorrigida' = 'SSQ2'-'Count2'*('Mean2'**2)
MTB > Sum 'SQCorrigida'
```

Sum of SQCorrigida

```
Sum of SQCorrigida = 432,146
```

- Resultados Armazenados

Média

Soma de Quadrados Simples

ByVar2	Mean2	Sum2	SSQ2	Count2	SQCorrigida
F	7,8367	94,04	833,11	12	96,147
M	11,5283	207,51	2626,88	18	234,639
S	16,475	98,85	1729,91	6	101,36

Soma Simples

n_i

Soma de Quadrados Corrigida

- Todos os resultados são rapidamente obtidos por Análise de Variância

Stat > Basic Statistics > Store Descriptive Statistics

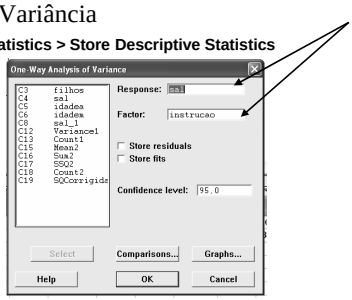


Tabela Anova - Saída

One-way ANOVA: sal versus instrucao

```
Source     DF     SS     MS      F      P
instrucao   2  304,4  152,2  11,62  0,000
Error      33  432,1   13,1
Total      35  736,6

S = 3,619   R-Sq = 41,33%   R-Sq(adj) = 37,77%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level   N    Mean  StDev  ----+---------+---------+---------+-----
F      12   7,837  2,956  (-----*-----)
M      18  11,528  3,715        (----*----)
S       6  16,475  4,502                  (--------*--------)
                          ----+---------+---------+---------+-----
                             7,0       10,5      14,0      17,5

Pooled StDev = 3,619
```

Soma dos Quadrados Residual

R^2

$$SQCorrigida_F = 11\,(2{,}956)^2 = 96{,}18$$

$$\overline{Var(S)} = \frac{35\,(3{,}619)^2}{36} = 12{,}733$$

Salários *vs* Região de Prodência

- Saída do Minitab:

```
MTB > Oneway sal rp

One-way ANOVA: sal versus rp

Source  DF     SS    MS     F      P
rp       2    9,4   4,7  0,21  0,810
Error   33  727,2  22,0
Total   35  736,6

S = 4,694   R-Sq = 1,27%   R-Sq(adj) = 0,00%

                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level      N    Mean  StDev  -+---------+---------+---------+--------
capital   11  11,455  5,477      (-------------*-------------)
interior  12  11,550  5,296      (-------------*-------------)
outra     13  10,445  3,145  (-----------*-----------)
                              -+---------+---------+---------+--------
                             8,0      10,0      12,0      14,0

Pooled StDev = 4,694
```

Soma dos Quadrados Residual

R^2

Comentário

- Comparando-se os valores de R^2 em cada associação estudada, verifica-se que há uma relação entre salário e instrução, não ocorrendo relação entre salário e região de procedência.

Referências

Bibliografia Recomendada

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