Correlação

Frases

"Uma probabilidade razoável é a única certeza"

Samuel Howe

"A experiência não permite nunca atingir a certeza absoluta. Não devemos procurar obter mais que uma probabilidade ."

Bertrand Russel

Roteiro

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

Coeficiente de Correlação

Dados Emparelhados

- Há uma relação?
- Se há, qual é a equação?
- Usar a equação para predição

Correlação

- Entre duas variáveis, existe correlação quando uma delas está, de alguma forma, relacionada com a outra.

Suposições

- A amostra de dados emparelhados *(X, Y)* é uma amostra aleatória.
- Os pares de dados *(X, Y)* tem distribuição normal bivariada.

Diagrama de Dispersão

- Gráfico de dados amostrais emparelhados *(x, y)* com o eixo das abcissas (eixo *x*) e o eixo das ordenadas (eixo *y*).
- Cada par individual *(x, y)* é plotado como um ponto.

Exemplo

Dados de algumas regiões metropolitanas:

√ Porcentagem da população economicamente ativa empregada no setor primário

√ Índice de analfabetismo

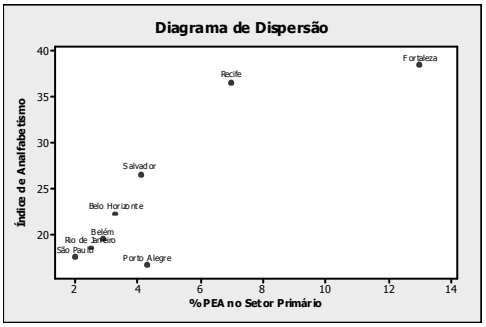
Planilha: *analfabetismo*

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

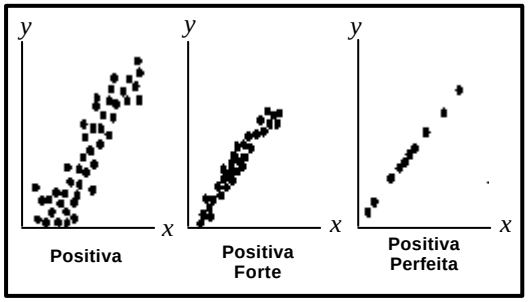
Região	Setor Primário	Índice Analfabetismo
São Paulo	2,0	17,5
Rio de Janeiro	2,5	18,5
Belém	2,9	19,5
Belo Horizonte	3,3	22,2
Salvador	4,1	26,5
Porto Alegre	4,3	16,6
Recife	7,0	36,6
Fortaleza	13,0	38,4

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

Diagrama de Dispersão

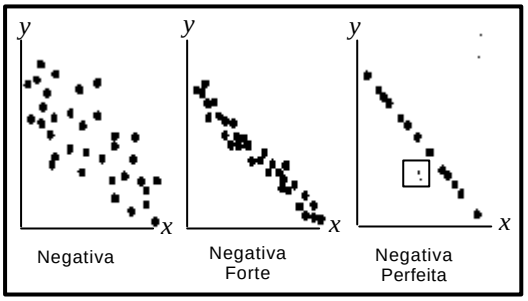


Correlação Linear Positiva



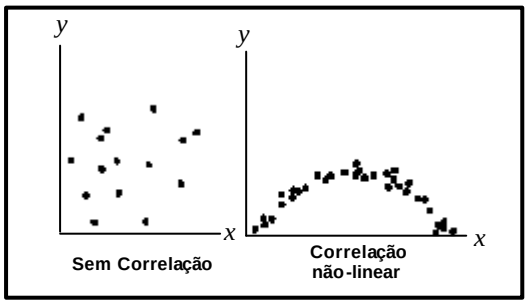
Diagramas de Dispersão

Correlação Linear Negativa



Diagramas de Dispersão

Sem Correlação Linear



Diagramas de Dispersão

Notação

x_i : i-ésimo valor observado da variável x

y_i : i-ésimo valor observado da variável y

$\bar{x}$: média dos valores observados da variável x (média amostral)

$\bar{y}$: média dos valores observados da variável y (média amostral)

Soma de Quadrados – Notação

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

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

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

Coeficiente de Correlação Linear Amostral

- Mede o grau de relacionamento linear entre os valores emparelhados *x* e *y* em uma <u>amostra.</u>

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

- Em geral, calculadoras financeiras calculam o valor de *r.*

Exemplo

Região	*Setor Primário*	*Índice Analfabetismo*			
	X	**Y**	**X^2**	**Y^2**	**XY**
São Paulo	2	17,5	4,00	306,25	35,00
Rio de Janeiro	2,5	18,5	6,25	342,25	46,25
Belém	2,9	19,5	8,41	380,25	56,55
Belo Horizonte	3,3	22,2	10,89	492,84	73,26
Salvador	4,1	26,5	16,81	702,25	108,65
Porto Alegre	4,3	16,6	18,49	275,56	71,38
Recife	7	36,6	49,00	1.339,56	256,20
Fortaleza	13	38,4	169,00	1.474,56	499,20
Total	*39,10*	*195,80*	*282,85*	*5.313,52*	*1.146,49*

$\bar{x} = 4{,}89$ $\bar{y} = 24{,}48$ $S_{xx} = 282{,}85 - 8(4{,}89)^2 = 91{,}75$

$S_{yy} = 5.313{,}52 - 8(24{,}48)^2 = 519{,}37$

$S_{xy} = 1.146{,}49 - 8(4{,}89)(24{,}48) = 188{,}83$

$$r = \frac{188{,}83}{\sqrt{(91{,}75)(519{,}37)}} = 0{,}865$$ ⇦

HP 12C – Cálculo do Coeficiente de Correlação

Para dados pareados:

- √ Digite o valor y_1
- √ Pressione ***ENTER***
- √ Digite o valor x_1
- √ Pressione $\Sigma+$
- √ Repita a operação para todos os pares
- √ Pressione ***g*** $\hat{x}$***, r*** ou ***g*** $\hat{y}$***, r***
- √ Pressione ***x? y*** e leia no visor o valor de *r*

HP 12C – Correção de Estatísticas Acumuladas

- Caso tenha errado na entrada do último par de dados:
  - √ Pressione ***g LST x*** e ***g*** $\Sigma-$
- Caso tenha errado algum par anterior ao último:
  - √ Digite novamente o par e pressione ***g*** $\Sigma-$

Armazenamento das Estatísticas Acumuladas

Registro	***Estatística***
R_1 (e visor)	*n* : # pares acumulados
R_2	*? x* : soma valores de *x*
R_3	*? x²* : soma valores de x^2
R_4	*? y* : soma valores de *y*
R_5	*? y²* : soma valores de y^2
R_6	*? xy* : soma valores de *xy*

Minitab – Cálculo do Coeficiente de Correlação

√ Em *Session*, *Editor* > *Enable Comando*.

```
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

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

Outra expressão para Cálculo de *r*

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

- s_x: desvio padrão amostral de x
- s_y: desvio padrão amostral de y

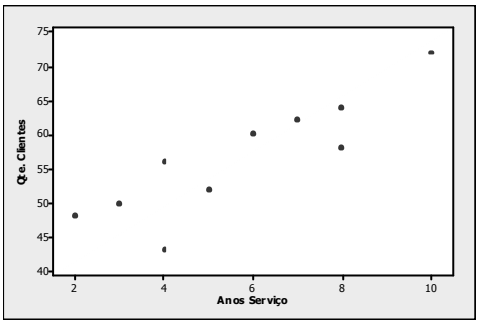
Interpretação de *r*

Exemplo

- Número de anos de serviço *(X)* por número de clientes *(Y)* de uma seguradora:

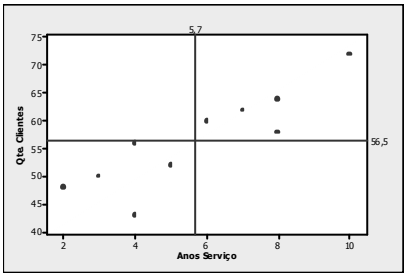
Agente	Anos Serviço (X)	Qte. Clientes (Y)
A	2	48
B	3	50
C	4	56
D	5	52
E	4	43
F	6	60
G	7	62
H	8	58
I	8	64
J	10	72

Diagrama de Dispersão dos Dados



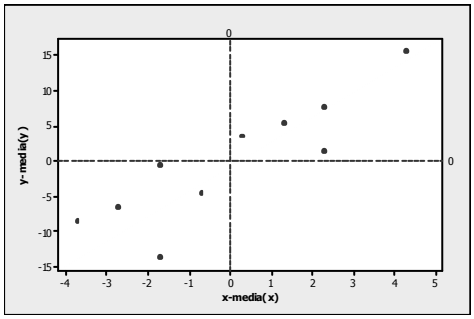
O coeficiente de correlação linear é também uma medida da proximidade dos dados a uma reta

- Considere a origem passando no centróide dos dados



- Maioria dos pontos está no 1º e no 3º quadrantes
- Nesses quadrantes, o produto das coordenadas será sempre positivo (soma dos produtos será positiva).

- Para se obter esta visão transfere-se a origem para o centro da nuvem de dados



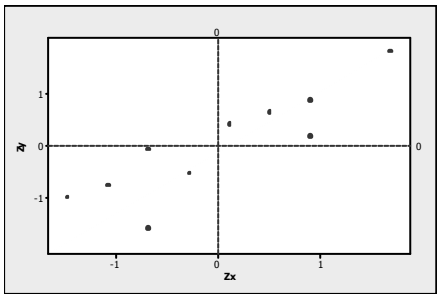
- Outro problema relevante é quanto à escala dos dados
- *Y* tem variabilidade muito maior que *X* e o produto ficaria muito mais afetado pelos valores de *Y* do que pelos de *X*
- Podemos reduzir as duas variáveis a uma mesma escala, dividindo-se os desvios pelos respectivos desvios padrões

Agente	X	Y	Z_x	Z_y	Z_xZ_y
A	2	48	-1,46	-0,99	1,45
B	3	50	-1,06	-0,76	0,81
C	4	56	-0,67	-0,06	0,04
D	5	52	-0,28	-0,53	0,14
E	4	43	-0,67	-1,58	1,06
F	6	60	0,12	0,41	0,05
G	7	62	0,51	0,64	0,33
H	8	58	0,91	0,18	0,16
I	8	64	0,91	0,88	0,79
J	10	72	1,69	1,81	3,07
Total	57	565	0,00	0,00	7,891
Média	5,70	56,50			
D.Padrão	2,54	8,55			

$$Z_{x_i} = \frac{x_i - \bar{x}}{s_x}$$

Como esperado, a soma é positiva

- Mudança das escala dos eixos



- A soma dos produtos das coordenadas depende (muito) do número de pontos
- Para facilitar a comparação usa-se a média da soma dos produtos das coordenadas
- Por razões técnicas, divide-se por *(n-1)*

$$r = \frac{7{,}891}{9} = 0{,}877$$

Grau de associação linear quantificado por 87,7%

Expressão Final para *r*

$$r = \frac{1}{n-1}\sum_i \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

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

Covariância Amostral

- Dados *n* pares (x_1, y_1), ..., (x_n, y_n), a covariância amostral entre as variáveis *X* e *Y* é dada por:

$$\text{cov}(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)} = \frac{S_{xy}}{(n-1)}$$

- A covariância pode ser entendida como uma média de produtos centrados das variáveis

Coeficiente de Correlação Amostral

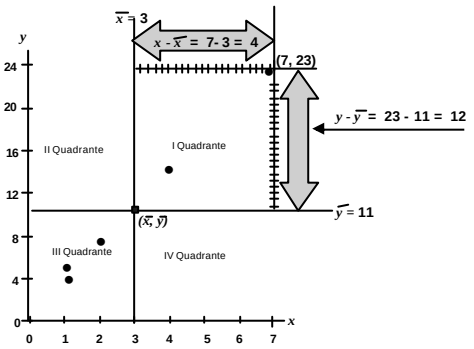
- Pode-se usar a covariância para calcular *r*:

$$r = \frac{\text{cov}(X,Y)}{s_x s_y}$$

Justificação para a Fórmula de *r*

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

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



Explicação da Variação

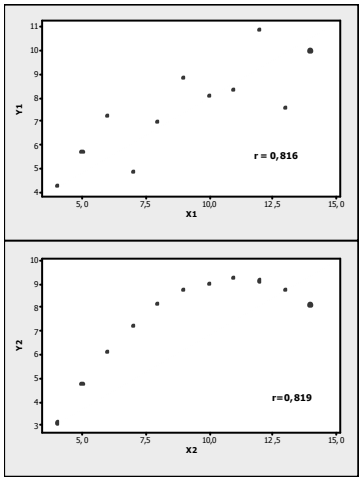
- A quantidade $100r^2$ pode ser entendida como a porcentagem de variação total dos y's que é explicada por sua relação com x (ou vice-versa)
- Se $r = 0,80$ então $100\%(0,8)^2 = 64\%$ da variação total de uma variável é explicada pela outra variável
- Se r' = 0,40 teremos *16%* de explicação total entre as variáveis;
- A correlação *r* é 4 vezes mais forte que a correlação *r*.

Correlação – Erros Comuns

- Linearidade:

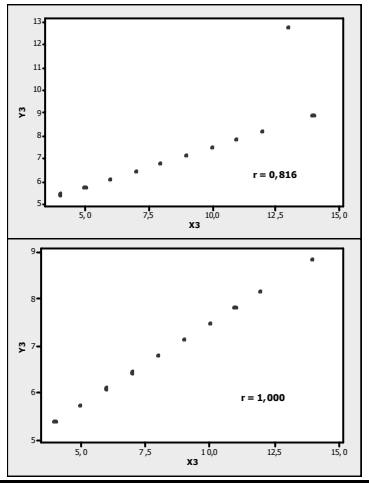
  r mede apenas a intensidade de relações lineares

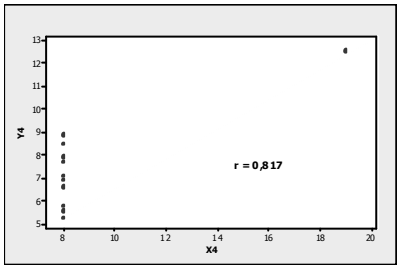
  Pode haver alguma relação entre x e y mesmo quando não há correlação linear significativa.



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;

Análise de Correlação

Coeficiente de Correlação Linear Populacional

- Mede o grau de associação de todos os dados emparelhados da população.

$$r = \frac{E(X - m_X)(Y - m_Y)}{\sqrt{Var(X)Var(Y)}}$$

$$r = \frac{\text{cov}(X,Y)}{DP(X)\,DP(Y)}$$

Inferência sobre *?*

- O coeficiente de correlação amostral *r* é apenas uma estimativa do parâmetro populacional *?*
- Hipóteses:

  H_0: ? = 0 *(não há correlação significativa)*

  vs

  H_1: ? ? 0 *(há correlação significativa)*
- Estatísticas de teste:
  - √ *t*
  - √ *r*

Método 1: Estatística de teste é *t*

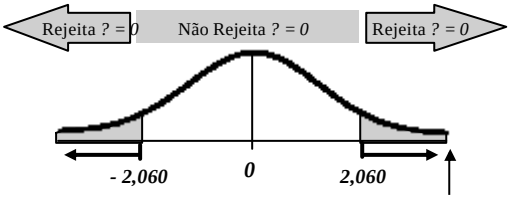
- Estatística de teste:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

- Distribuição da estatística:
  t-Student com *n – 2* graus de liberdade
- Suposição sobre a população:
  modelo normal bivariado

Exemplo

- Amostra: *n=27 e r = 0,82.* Usar *a = 0,05*



$t_{25;\ 0,025} = 2,060$

$t_{obs} = \frac{0,82\sqrt{25}}{\sqrt{1-(0,82)^2}} = 7,16$

Conclusão: A correlação é siginificativa

Método 2: Estatística de teste é *r*

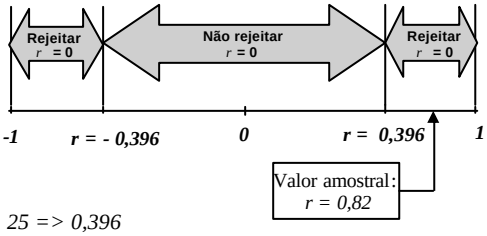
- Exige menos cálculos
- Valores críticos:
  - √ *Tabela de Valores Críticos do Coeficiente de Correlação de Pearson*

Coeficiente de Correlação – Valores Críticos

n	*a* = ,05	*a* = ,01
4	,950	,999
5	,878	,959
6	,811	,917
7	,754	,875
8	,707	,834
9	,666	,798
10	,632	,765
11	,602	,735
12	,576	,708
13	,553	,684
14	,532	,661
15	,514	,641
16	,497	,623
17	,482	,606
18	,468	,590
19	,456	,575
20	,444	,561
25	,396	,505
30	,361	,463
35	,335	,430
40	,312	,402
45	,294	,378
50	,279	,361
60	,254	,330
70	,236	,305
80	,220	,286
90	,207	,269
100	,196	,256

Exemplo

- Amostra: *n=27 e r = 0,82.* Usar *a = 0,05*



n = 25 => 0,396

Conclusão: A correlação é significativa

Aplicação Computacional

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 - 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

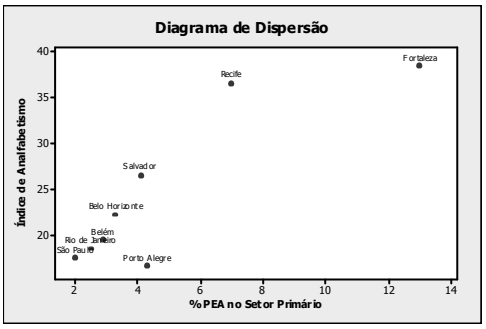
Existe alguma relação entre a porcentagem da população economicamente ativa no setor primário e o índice de analfabetismo?

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 Comando*.

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

Correlations: set_prim; analfab

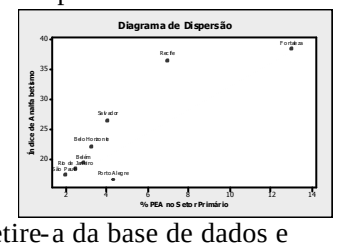
```
Pearson correlation of set_prim and analfab = 0,867
P-Value = 0,005
```

Ou através de:

Stat > Basic Statistics > Correlation

Correlação

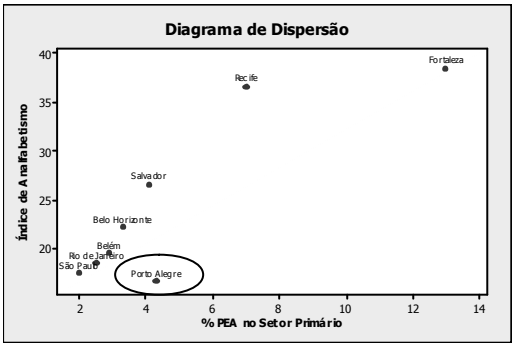
Há alguma região com comportamento diferente das demais?



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

dados

Diagrama de Dispersão



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
```

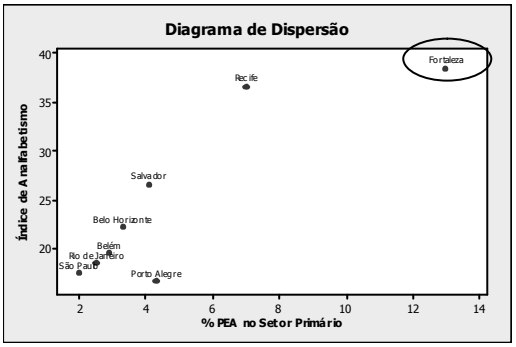
Porcentagem de Variação

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

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).

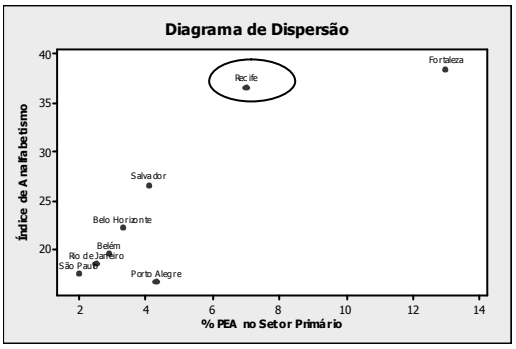
```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 8.

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).

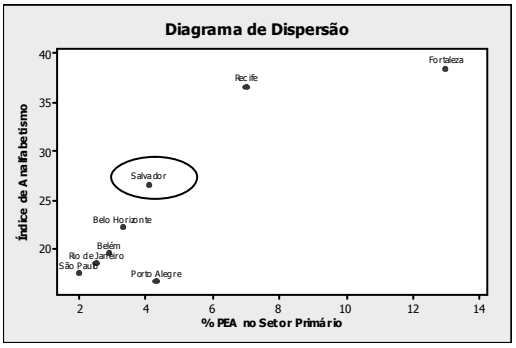
```
MTB > correlation 'set_prim' 'analfab';
SUBC> exclude;
SUBC> rows 7.

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.

Referências

Bibliografia Recomendada

- Freund, J. E. e Simon, G. A. (Artmed) *Estatística aplicada: economia, administração e contabilidade*
- Bussab, W. O. e Morettin, P. A. (Saraiva) *Estatística básica*