CPEQS

AULA 06
Correlação

Ernesto F. L. Amaral

04 de outubro de 2013

Centro de Pesquisas Quantitativas em Ciências Sociais (CPEQS)
Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:
Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 10 (pp.408-428).

VISÃO GERAL

– Nas próximas aulas, vamos falar de métodos para:
  – Fazer inferências sobre a relação (**correlação**) entre duas variáveis.
  – Elaborar uma equação que possa ser usada para prever o valor de uma variável dado o valor de outra (**regressão**).

– Serão considerados dados amostrais que vêm em pares.
  – No capítulo anterior, as inferências se referiam à **média das diferenças** entre pares de valores.
  – Neste capítulo, as inferências têm objetivo de verificar **relação** entre duas variáveis.

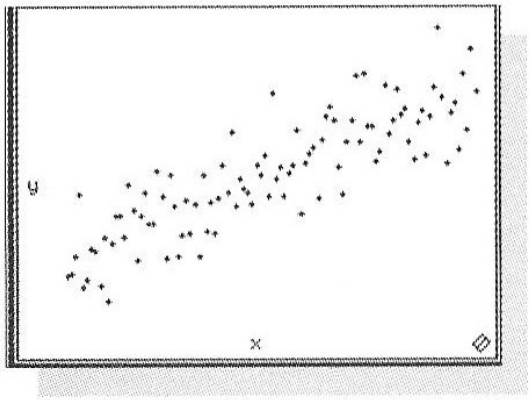
CORRELAÇÃO

CONCEITOS BÁSICOS

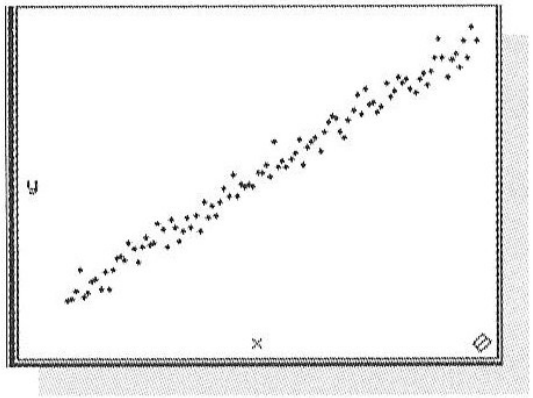
– Existe uma correlação entre duas variáveis quando uma delas está relacionada com a outra de alguma maneira.

– Antes de tudo é importante explorar os dados:
  – Diagrama de dispersão entre duas variáveis.
  – Há tendência?
  – Crescente ou decrescente?
  – *Outliers*?

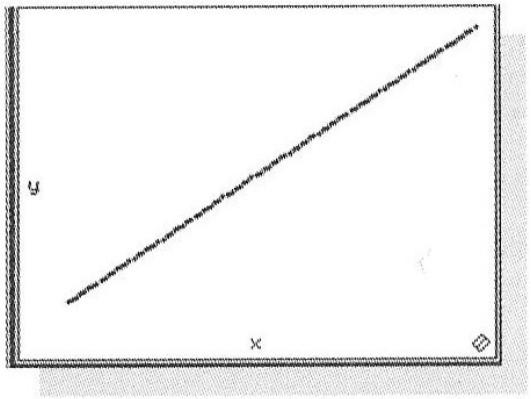
DIAGRAMAS DE DISPERSÃO
(correlação linear)



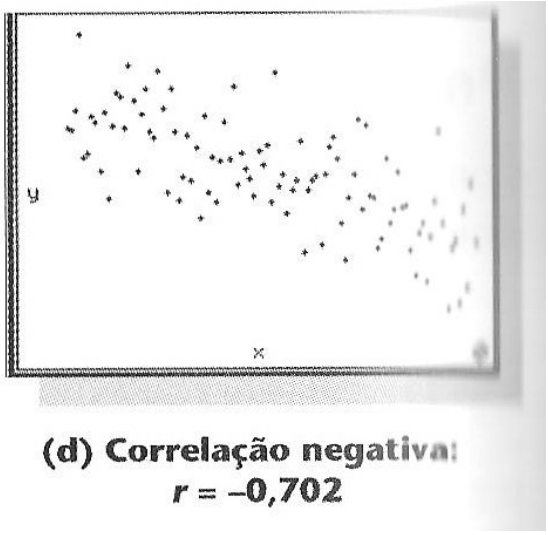
(a) Correlação positiva:
$r = 0{,}851$



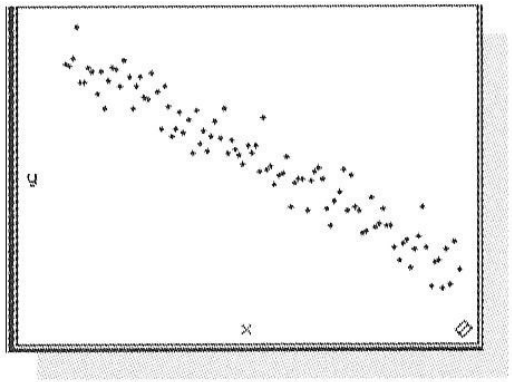
(b) Correlação positiva:
$r = 0{,}991$



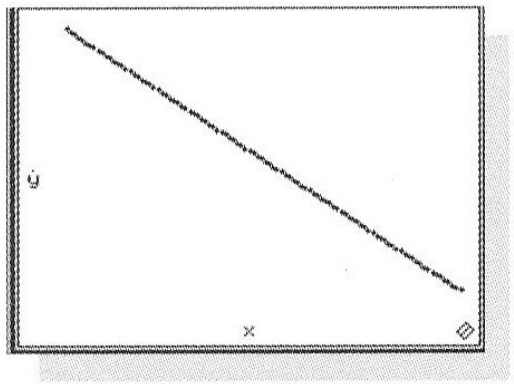
(c) Correlação positiva perfeita:
$r = 1$



(d) Correlação negativa:
$r = -0{,}702$

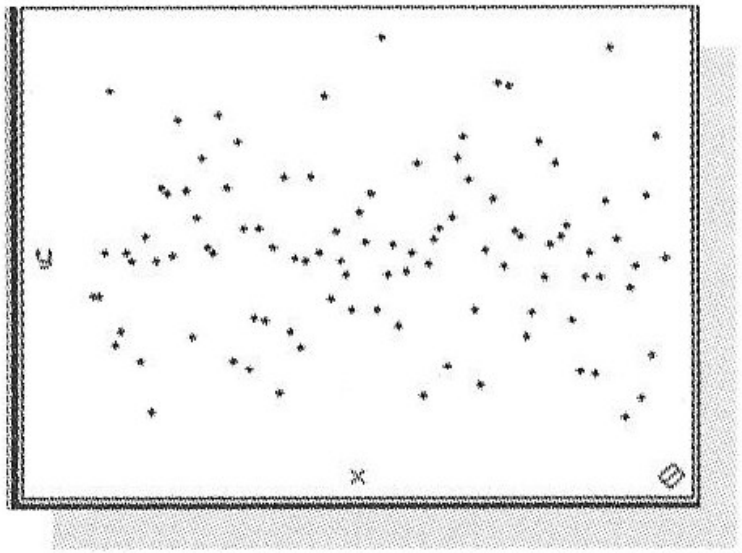


(e) Correlação negativa:
$r = -0{,}965$

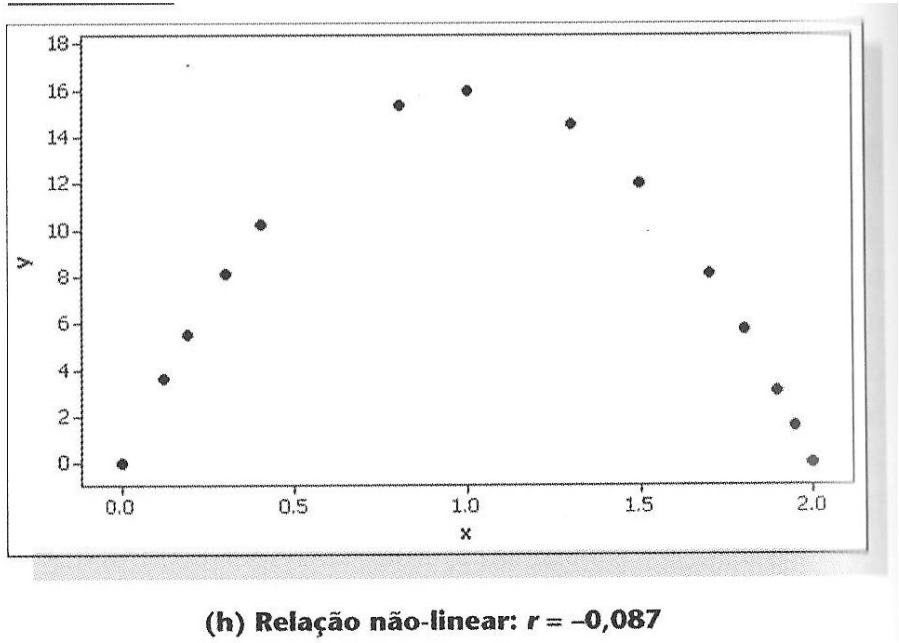


(f) Correlação negativa perfeita:
$r = -1$

DIAGRAMAS DE DISPERSÃO
(não há correlação linear)



(g) Nenhuma correlação: $r = 0$



(h) Relação não-linear: $r = -0,087$

CORRELAÇÃO

– O coeficiente de correlação linear (*r*):

– Medida numérica da força da relação entre duas variáveis que representam dados quantitativos.

– Mede intensidade da relação linear entre os valores quantitativos emparelhados *x* e *y* em uma amostra.

– É chamado de coeficiente de correlação do produto de momentos de Pearson.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

OBSERVAÇÕES IMPORTANTES

– Usando dados amostrais emparelhados (dados bivariados), estimamos valor de *r* para concluir se há ou não relação entre duas variáveis.

– Serão tratadas relações lineares, em que pontos no gráfico (x, y) se aproximam do padrão de uma reta.

– É importante entender os conceitos e não os cálculos aritméticos.

– *r* é calculado com dados amostrais. Se tivéssemos todos pares de valores populacionais *x* e *y*, teríamos um parâmetro populacional (ρ).

REQUISITOS

– Os seguintes requisitos devem ser satisfeitos ao se testarem hipóteses ou ao se fazerem outras inferências sobre *r* :

- – Amostra de dados emparelhados (*x*, *y*) é uma **amostra aleatória** de dados quantitativos independentes.
  - – Não pode ter sido utilizado, por exemplo, amostra de resposta voluntária.
- – Exame visual do diagrama de dispersão deve confirmar que pontos se aproximam do **padrão de uma reta**.
- – **Valores extremos** (*outliers*) devem ser removidos se forem erros.
  - – Efeitos de outros *outliers* devem ser considerados com estimação de *r* com e sem estes *outliers*.

VALORES CRÍTICOS DO COEFICIENTE DE CORRELAÇÃO DE PEARSON (*r*)

n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
4	0,950	0,999
5	0,878	0,959
6	0,811	0,917
7	0,754	0,875
8	0,707	0,834
9	0,666	0,798
10	0,632	0,765
11	0,602	0,735
12	0,576	0,708
13	0,553	0,684
14	0,532	0,661
15	0,514	0,641
16	0,497	0,623
17	0,482	0,606
18	0,468	0,590
19	0,456	0,575
20	0,444	0,561
25	0,396	0,505
30	0,361	0,463
35	0,335	0,430
40	0,312	0,402
45	0,294	0,378
50	0,279	0,361
60	0,254	0,330
70	0,236	0,305
80	0,220	0,286
90	0,207	0,269
100	0,196	0,256

NOTA: Para testar H_0: $\rho = 0$ *versus* H_1: $\rho \neq 0$, rejeite H_0 se o valor absoluto de *r* for maior que o valor crítico na tabela.

– **Arredonde** o coeficiente de correlação linear *r* para três casas decimais, permitindo comparação com esta tabela.

– Interpretação: com 4 pares de dados e **nenhuma correlação** linear entre *x* e *y*, há chance de 5% de que valor absoluto de *r* exceda 0,950.

INTERPRETANDO *r*

– O valor de *r* deve sempre estar entre –1 e +1.

– Se *r* estiver muito próximo de 0, concluímos que não há correlação linear significativa entre *x* e *y*.

– Se *r* estiver próximo de –1 ou +1, concluímos que há uma relação linear significativa entre *x* e *y*.

– Mais objetivamente:

  – Usando a tabela anterior, se valor absoluto de *r* excede o valor da tabela, há correlação linear.

  – Usando programa de computador, se valor *P* é menor do que nível de significância, há correlação linear.

PROPRIEDADES DE *r*

– Valor de *r* está entre: $-1 \leq r \leq +1$

– Valor de *r* não muda se todos valores de qualquer das variáveis forem convertidos para uma escala diferente.

– Valor de *r* não é afetado pela inversão de *x* ou *y*. Ou seja, mudar os valores de *x* pelos valores de *y* e vice-versa não modificará *r*.

– *r* mede intensidade de relação linear, não sendo planejado para medir intensidade de relação que não seja linear.

– O valor de r^2 é a proporção da variação em *y* que é explicada pela relação linear entre *x* e *y*.

ERROS DE INTERPRETAÇÃO

– Erro comum é concluir que correlação implica **causalidade**:
  – A causa pode ser uma variável oculta.
  – Uma variável oculta é uma variável que afeta as variáveis em estudo, mas que não está incluída no banco.

– Erro surge de dados que se baseiam em **médias**:
  – Médias suprimem variação individual e podem aumentar coeficiente de correlação.

– Erro decorrente da propriedade de **linearidade**:
  – Pode existir relação entre *x* e *y* mesmo quando não haja correlação linear (relação quadrática, por exemplo).

TESTE DE HIPÓTESE FORMAL PARA CORRELAÇÃO

– É possível realizar um teste de hipótese formal para determinar se há ou não relação linear significativa entre duas variáveis.

– Critério de decisão é rejeitar a hipótese nula (ρ=0) se o valor absoluto da estatística de teste exceder os valores críticos.

– A rejeição de (ρ=0) significa que há evidência suficiente para apoiar a afirmativa de uma correlação linear entre as duas variáveis.

– Se o valor absoluto da estatística de teste não exceder os valores críticos (ou seja, o valor *P* for grande), deixamos de rejeitar ρ=0.

H_0: ρ=0 (não há correlação linear)

H_1: $\rho \neq 0$ (há correlação linear)

MÉTODO 1: ESTATÍSTICA DE TESTE É *t*

– Estatística de teste representa o valor do desvio padrão amostral dos valores de *r* :

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

– Valores críticos e valor *P*: use tabela A-3 com *n*–2 graus de liberdade.

– Conclusão:

  – Se $|t|$ > valor crítico da Tabela A-3, rejeite H_0 e conclua que há correlação linear.

  – Se $|t| \leq$ valor crítico da Tabela A-3, deixe de rejeitar H_0 e conclua que não há evidência suficiente para concluir que haja correlação linear.

MÉTODO 2: ESTATÍSTICA DE TESTE É *r*

– Estatística de teste: *r*

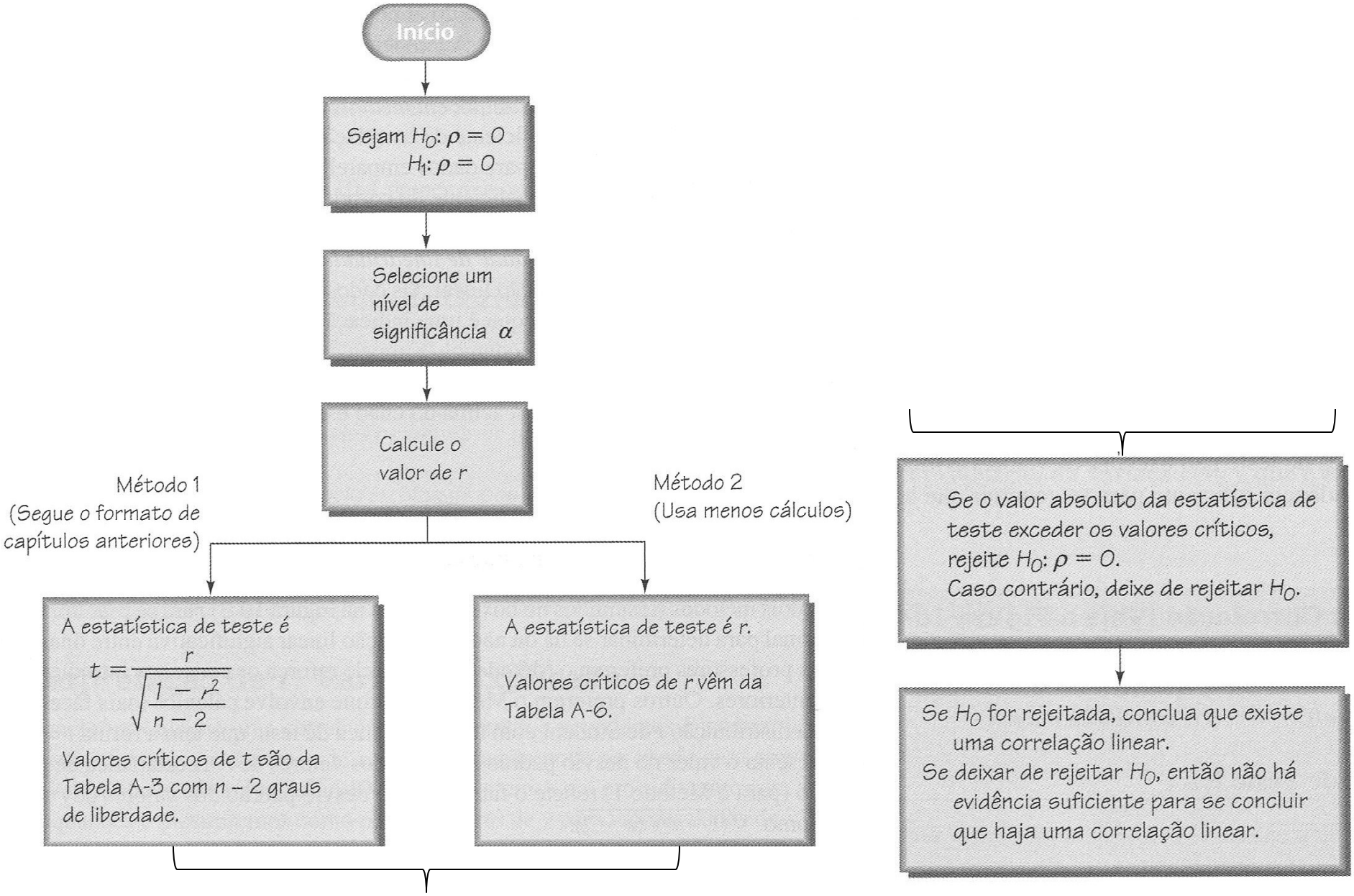
– Valores críticos: consulte Tabela A-6.

– Conclusão:

  – Se $|r|$ > valor crítico da Tabela A-6, rejeite H_0 e conclua que há correlação linear.

  – Se $|r|$ ≤ valor crítico da Tabela A-6, deixe de rejeitar H_0 e conclua que não há evidência suficiente para concluir que haja correlação linear.

TESTE DE HIPÓTESE PARA CORRELAÇÃO LINEAR



TESTES UNILATERAIS

– Os testes unilaterais podem ocorrer com uma afirmativa de uma correlação linear positiva ou uma afirmativa de uma correlação linear negativa.

– Afirmativa de correlação negativa (teste unilateral esquerdo):

$$H_0: \rho = 0$$

$$H_1: \rho < 0$$

– Afirmativa de correlação positiva (teste unilateral direito):

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

– Para isto, simplesmente utilize α=0,025 (ao invés de α=0,05) e α=0,005 (ao invés de α=0,01).

FUNDAMENTOS

– Essas fórmulas são diferentes versões da mesma expressão:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{\sum\left[\frac{(x - \bar{x})}{s_x}\frac{(y - \bar{y})}{s_y}\right]}{n - 1}$$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y} \qquad r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

FUNDAMENTOS

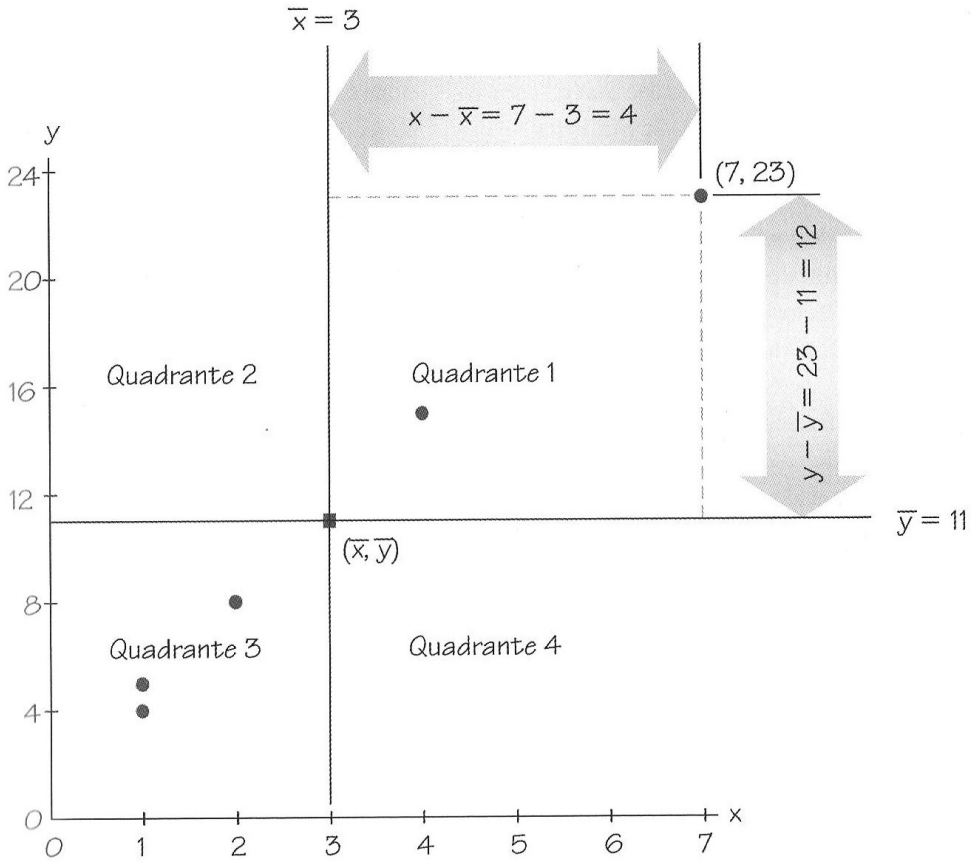
– Dada uma coleção de dados em pares (*x*,*y*), o ponto (*x*-barra, *y*-barra) é chamado de **centróide**.

– A estatística do produto dos momentos de Pearson (*r*) se baseia na soma dos produtos dos momentos:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Se pontos são reta ascendente, valores do produto estarão nos 1º e 3º quadrantes (soma positiva).

– Se é descendente, os pontos estarão nos 2º e 4º quadrantes (soma negativa).



OU SEJA...

– Podemos usar esta expressão para medir como pontos estão organizados:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Grande soma positiva sugere pontos predominantemente no primeiro e terceiro quadrantes (correlação linear positiva).

– Grande soma negativa sugere pontos predominantemente no segundo e quarto quadrantes (correlação linear negativa).

– Soma próxima de zero sugere pontos espalhados entre os quatro quadrantes (não há correlação linear).

PORÉM...

– Esta soma depende da magnitude dos números usados:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Para tornar *r* independente da escala utilizada, usamos a seguinte padronização:

  – Sendo s_x o desvio padrão dos valores amostrais *x*...

  – Sendo s_y o desvio padrão dos valores amostrais *y*...

  – Padronizamos cada desvio pela sua divisão por s_x...

$$\sum \left[\frac{(x - \bar{x})}{s_x} \frac{(y - \bar{y})}{s_y} \right]$$

– Usamos o divisor $n - 1$ para obter uma espécie de média:

$$r = \frac{\sum \left[\frac{(x - \bar{x})}{s_x} \frac{(y - \bar{y})}{s_y} \right]}{n - 1}$$