Métodos Empíricos de Pesquisa I

- Análise Bidimensional

Aula de hoje

- Temas
  - Associação entre variáveis
    - Qualitativas e Quantitativas
  - Covariância: conceitos e propriedades
  - Coeficiente de correlação
  - Observações e análise
- Bibliografia básica
  - Bussab, W. e Morettin, P. Estatística básica. 5. ed. São Paulo: Saraiva, 2005. Cap. 4
  - Lapponi, J. Estatística usando Excel 5 e 7. Rio de Janeiro: Elsevier, 2005. 7ª reimpressão Capítulo 6

Considerações preliminares

- Dada a classificação vista anteriormente, sabemos que, no caso do estudo com duas variáveis, três combinações são possíveis
  - duas variáveis qualitativas
  - duas variáveis quantitativas
  - uma variável qualitativa e a outra quantitativa

Distribuição conjunta das frequências

- Usando exemplo apresentado em Bussab-Morettin, p.71
- Variáveis grau de instrução (Y) e região de procedência (V)

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	4	5	2	11
Interior	3	7	2	12
Outro	5	6	2	13
Total	12	18	6	36

Frequência em análise bidimensional

- Como calcular a frequência em análise com duas variáveis? Qual o total utilizar? Da coluna? Da linha? Ou o total geral?
  - Depende da análise desejada. A divisão pelo total geral expressa a composição do grupo por ambas características.
  - A divisão pelo total da linha ou da coluna expressa um resultado *condicional* à observação da linha ou coluna.

Análise pelo total geral

- Frequência das observações em relação ao total da população (ou da amostra), em %

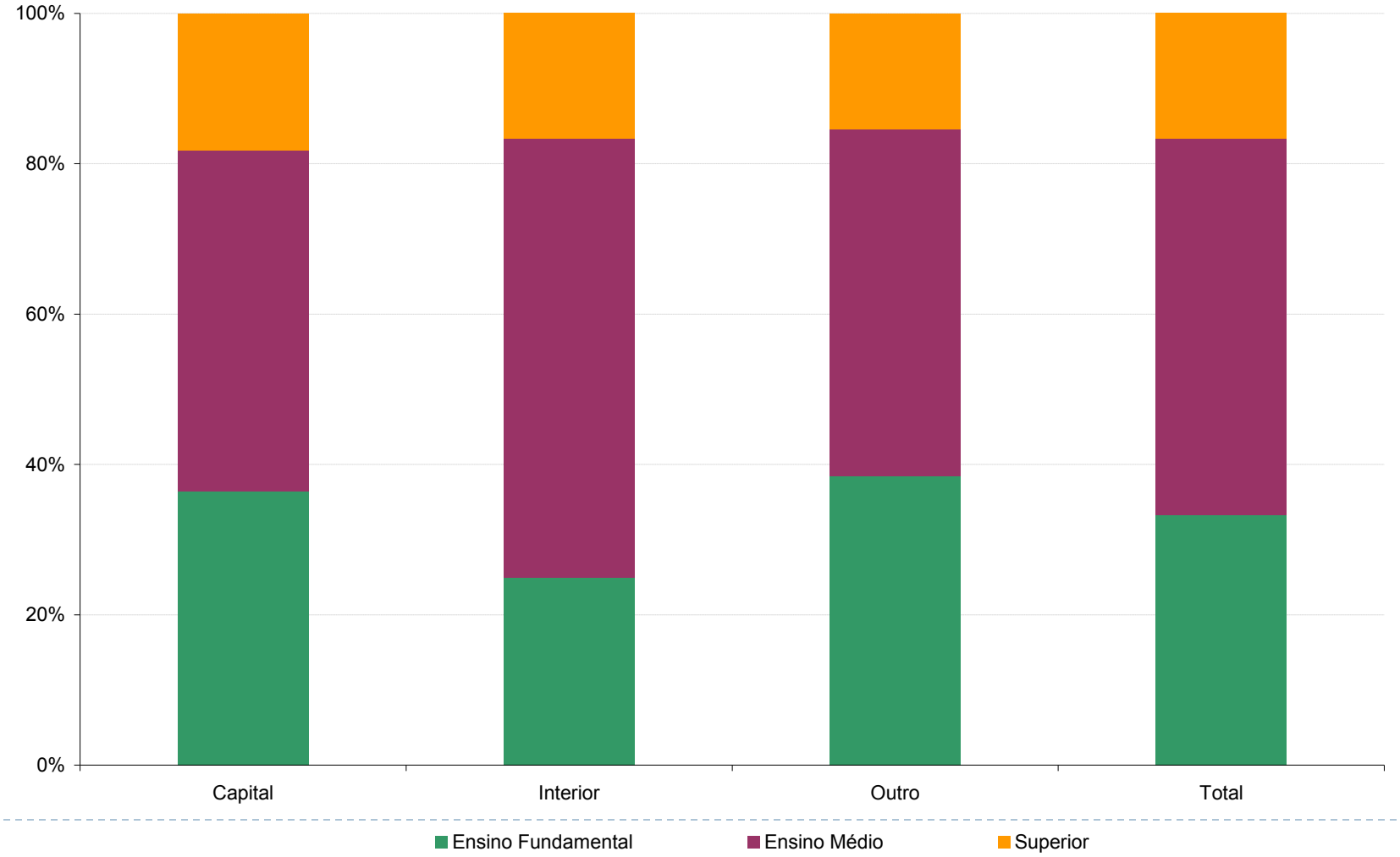
V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	11	14	6	31
Interior	8	19	6	33
Outro	14	17	6	36
Total	33	50	17	100

Análises pelo total da coluna ou da linha

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	33	28	33	31
Interior	25	39	33	33
Outro	42	33	33	36
Total	100	100	100	100

V \ Y	Ensino Fundamental	Ensino Médio	Superior	Total
Capital	36	45	18	100
Interior	25	58	17	100
Outro	38	46	15	100
Total	33	50	17	100

Distribuição do grau de instrução por região de procedência (em %)



O que dizem os dados?

- No exemplo, a distribuição pelo total da população mostra que, por exemplo, 36% dos funcionários da empresa que vieram da capital, terminaram o ensino fundamental
- Por outro lado, no exemplo da divisão pelos totais das colunas, temos que *entre os funcionários com ensino médio,* 39% vieram do interior

Associação entre variáveis

- O objetivo de estabelecer a distribuição conjunta de duas variáveis é o de compreender a existência de alguma associação entre elas, ou o grau de dependência entre elas

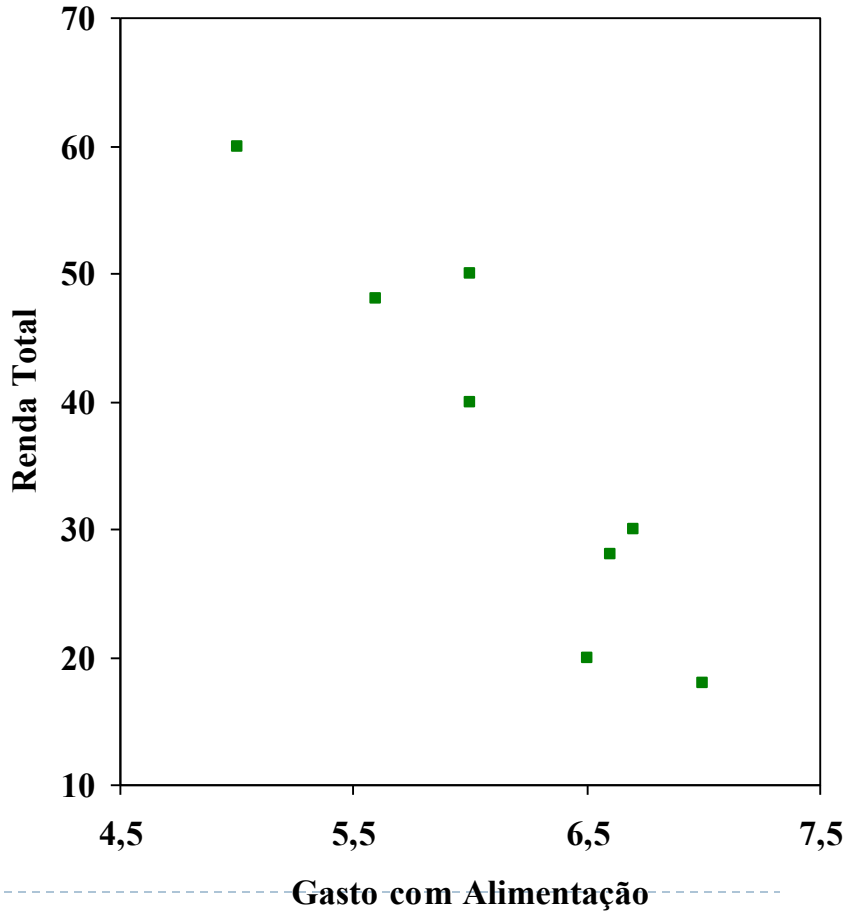
Associação entre variáveis quantitativas

- O procedimento de cálculo de frequências entre tabelas pode ser feito normalmente no caso de variáveis quantitativas
- Mas, em alguns casos, devemos atribuir intervalos que formem as categorias de análise - os chamados intervalos de classes
- Uma ferramenta importante na análise de variáveis quantitativas é o *gráfico de dispersão*

Exemplo: Renda familiar e gastos com alimentação (em % da renda)

- Como esperado, à medida em que aumenta a renda familiar, diminui o percentual da renda destinado à alimentação

Família	Renda Total	Gasto em Alimentação
A	12	7,2
B	16	7,4
C	18	7,0
D	20	6,5
E	28	6,6
F	30	6,7
G	40	6,0
H	48	5,6
I	50	6,0
L	60	5,0



Exemplo livro (Bussab-Morettin), p.81

- Consideremos as duas variáveis abaixo

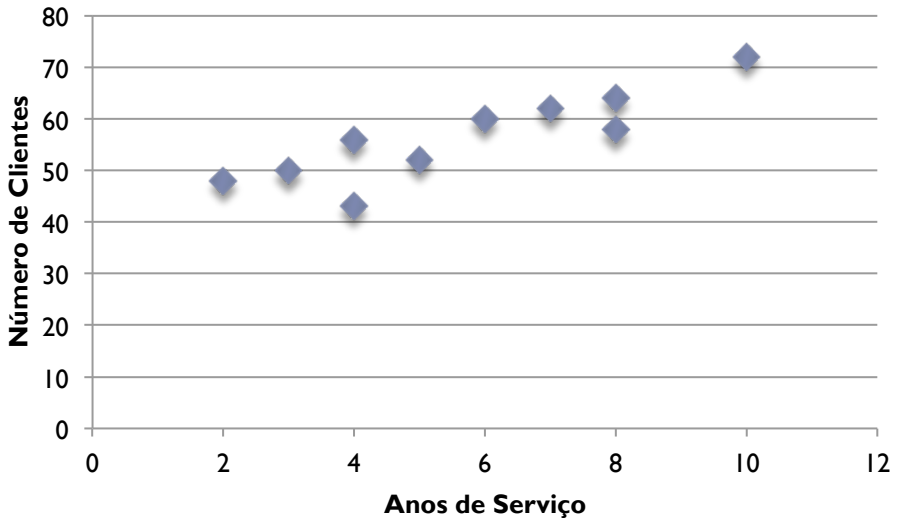
  Número de anos de serviço (X) por número de clientes de agentes de uma cia de seguros

Agente	Anos de serviço (X)	Número de clientes
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

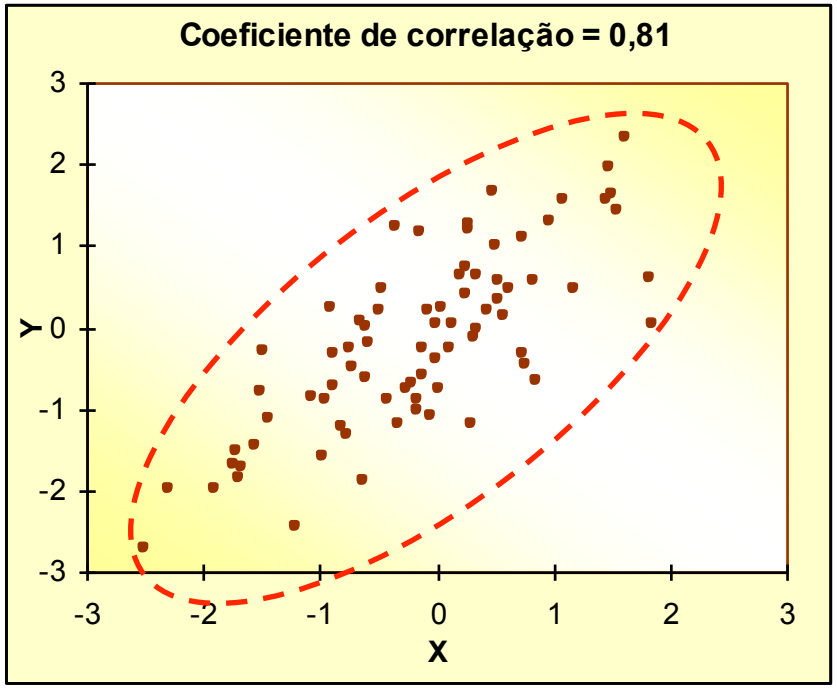
- Dados hipotéticos

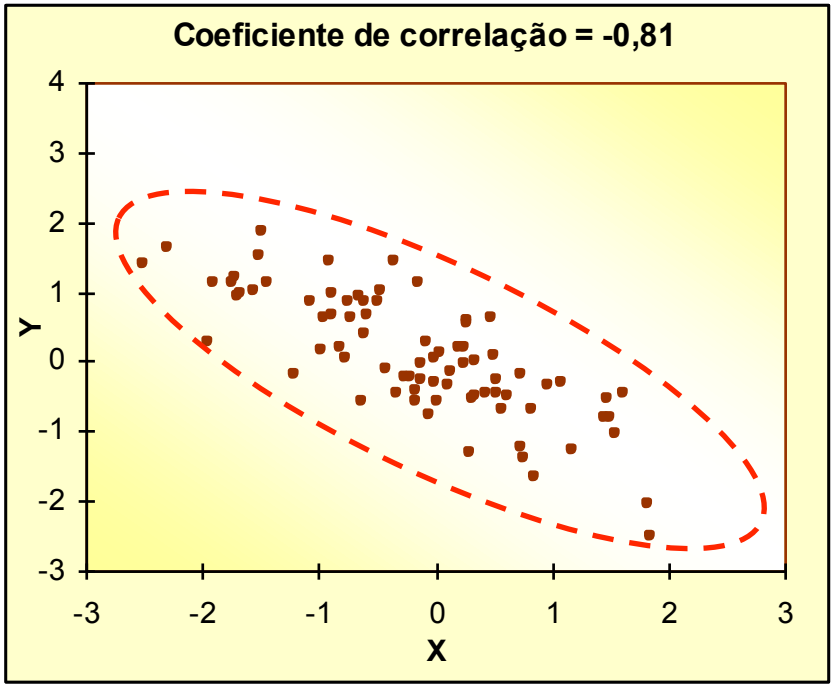
Exemplo livro (Bussab-Morettin), p.81

- Gráfico de Dispersão



- Dados hipotéticos





O gráfico de dispersão da esquerda mostra uma relação direta ou positiva entre as variáveis *X* e *Y*, tendência destacada pela declividade positiva da elipse tracejada. Enquanto o gráfico de dispersão da direita mostra uma relação inversa ou negativa, tendência também destacada pela declividade negativa da elipse tracejada.

Covariância

- Dados n pares de valores $(x_1, y_1)..., (x_n, y_n)$, chamaremos de covariância entre as variáveis X e Y, consideradas como população:

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$$

- É a média dos produtos dos valores centrados das variáveis
- Tendo esta definição, podemos escrever o coeficiente de correlação como:

$$corr(X,Y) = \frac{\mathrm{cov}(X,Y)}{dp(X).dp(Y)}$$

Covariância

Usando, agora, a notação de Lapponi
(lembrem-se que é a mesma coisa...)

- A *covariância* σ_{XY} das variáveis $X = X_1, X_2, \cdots, X_N$ e $Y = Y_1, Y_2, \cdots, Y_N$, consideradas como população é[4]:

$$\sigma_{XY} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu_X)\times(Y_i - \mu_Y)$$

- A *covariância* S_{XY} das variáveis $X = X_1, X_2, \cdots, X_n$ e $Y = Y_1, Y_2, \cdots, Y_n$, consideradas como amostra é:

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})\times(Y_i - \bar{Y})$$

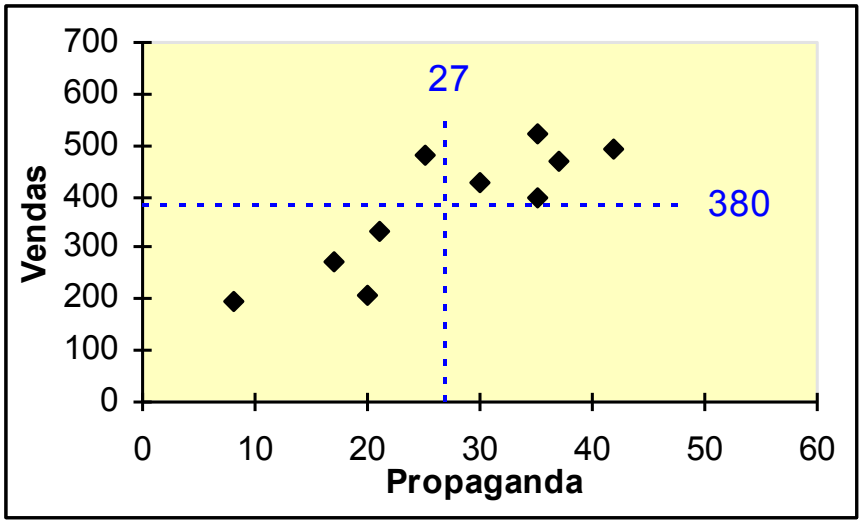
Características da covariância

- Das expressões da covariância, população e amostra:
  - As duas variáveis devem ter o mesmo número de dados.
  - Os pares de dados ocorrem ao mesmo tempo, são pares casados. Embora possa parecer redundante, é importante observar que não se pode mudar a ordem de uma única variável; a mudança de ordem deverá ser realizada nas duas amostras sem descasar os pares de dados.

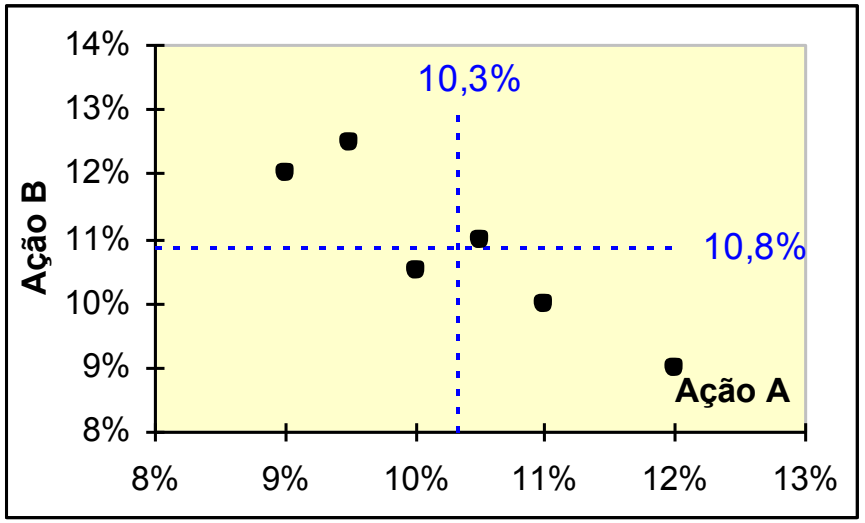
Características da covariância

- A covariância é:
  - No caso de população, a soma dos produtos dos desvios de duas variáveis dividida pela quantidade de dados das variáveis.
  - No caso de amostra, a soma dos produtos dos desvios de duas variáveis dividida pela quantidade de dados das variáveis menos um.
- Os numeradores das expressões da covariância para população e para amostra são iguais, o resultado da soma dos produtos dos desvios.

- A covariância pode ser nula, negativa ou positiva.
- A covariância é a medida do afastamento simultâneo das respectivas médias.
- Se as ambas variáveis aleatórias tendem a estar simultaneamente acima, ou abaixo, de suas respectivas médias, então a covariância tenderá a ser positiva e nos outros casos poderá ser negativa, como mostram os gráficos abaixo.



A maioria dos pares de valores tem os dois valores acima de sua média correspondente, provocando covariância positiva.



A maioria dos pares de valores tem um valor acima de sua média e outro abaixo da média correspondente, provocando covariância negativa.

Características da covariância

- A covariância de uma variável e ela mesma é a própria variância da variável, seja no caso de população ou amostra. Como *Y* = *X*,

$$\sigma_{XX} = \frac{\sum_{i=1}^{N}(X_i - \mu_X) \times (X_i - \mu_X)}{N} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N} = \sigma_X^2$$

- A permutação das variáveis não altera o resultado da covariância, se os pares de valores não forem alterados

$$\sigma_{XY} = \sigma_{YX}$$

Características da covariância

- Da mesma forma que a variância, a covariância é afetada pelos valores extremos da variável, ela não é uma medida resistente.
- A unidade de medida é o resultado do produto das unidades dos valores das variáveis.

Coeficiente de correlação

- Para facilitar o entendimento da relação entre duas variáveis e evitar a unidade de medida da covariância, foi definido o coeficiente de correlação $r_{XY.}$
- Os valores de r_{XY} estão limitados entre os valores -1 e +1, e sem nenhuma unidade de medida

Coeficiente de correlação

- O coeficiente de correlação busca auferir a direção da relação entre as variáveis, dentro de um intervalo determinado entre -1 e 1
- O objetivo do intervalo é discriminar a direção e a intensidade da relação:
  - valores próximos de zero indicam ausência de relação entre as variáveis
  - valores próximos de 1 indicam forte relação positiva
  - valores próximos de -1 indicam forte relação negativa

Coeficiente de correlação

- O coeficiente de correlação é a medida do grau de associação linear entre duas variáveis
- Fórmula do coeficiente de correlação:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

Cálculo do coeficiente de correlação

Agente	Anos de serviço (X)	Número de clientes	$x - \bar{x}$	$y - \bar{y}$	$\frac{x - \bar{x}}{dp(X)} = z_x$	$\frac{y - \bar{y}}{dp(Y)} = z_y$	zx.zy
A	2	48	-3,7	-8,5	-1,54	-1,05	1,608
B	3	50	-2,7	-6,5	-1,12	-0,80	0,897
C	4	56	-1,7	-0,5	-0,71	-0,06	0,043
D	5	52	-0,7	-4,5	-0,29	-0,55	0,161
E	4	43	-1,7	-13,5	-0,71	-1,66	1,173
F	6	60	0,3	3,5	0,12	0,43	0,054
G	7	62	1,3	5,5	0,54	0,68	0,366
H	8	58	2,3	1,5	0,95	0,18	0,176
I	8	64	2,3	7,5	0,95	0,92	0,882
J	10	72	4,3	15,5	1,78	1,91	3,407

Para calcular o coeficiente de correlação, devemos dividir o somatório dos valores da última coluna (8,77) pelo número de observações (n=10)

Então: Corr(X,Y) = 8,77/10=0,877

Coeficiente de correlação

- O coeficiente de correlação pode ser escrito da seguinte forma:

$$corr(X,Y) = \frac{1}{n}\sum\left(\frac{x_i - \bar{x}}{dp(X)}\right)\left(\frac{y_i - \bar{y}}{dp(Y)}\right)$$

$$corr(X,Y) = \frac{\sum x_i y_i - n\overline{xy}}{\sqrt{\left(x_i^2 - n\bar{x}^2\right)\left(y_i^2 - n\bar{y}^2\right)}}$$

Sendo que -1≤ corr(X,Y) ≤1

- Lembremos da **variância**, que usamos para observar a dispersão de uma só variável

$$\text{var}(X) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$$

Coeficiente de correlação

Na notação usada por Lapponi

Coeficiente de correlação r_{XY} das variáveis X e Y é um valor único calculado com a seguinte fórmula:

- Se os dados referem-se à população: $r_{XY} = \frac{\sigma_{XY}}{\sigma_X \times \sigma_Y}$
- Se os dados referem-se à amostra: $r_{XY} = \frac{S_{XY}}{S_X \times S_Y}$

Voltando ao coeficiente de correlação

- Da fórmula do coeficiente de correlação pode-se obter também a covariância das mesmas variáveis quando conhecidos os desvios padrões correspondentes:

$$\sigma_{XY} = r_{XY} \times \sigma_X \times \sigma_Y$$

Características de r

- Se a variável *Y* é a mesma variável *X*, então o coeficiente de correlação é igual a 1:

$$r_{XX} = \frac{\sigma_{XX}}{\sigma_X \times \sigma_X} = \frac{\sigma_X^2}{\sigma_X^2} = 1$$

- A permutação das variáveis não altera o resultado do coeficiente de correlação, se os mesmos pares de valores forem mantidos.

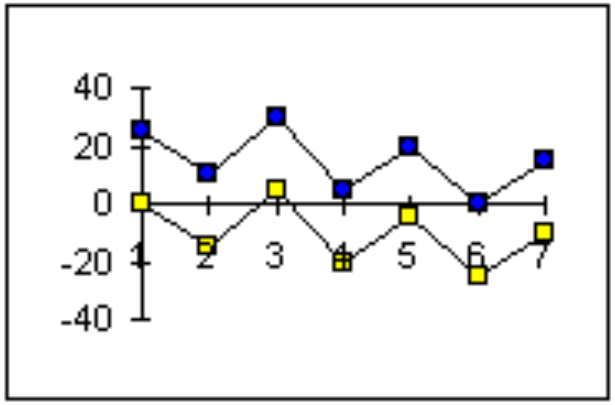
$$r_{XY} = r_{YX}$$

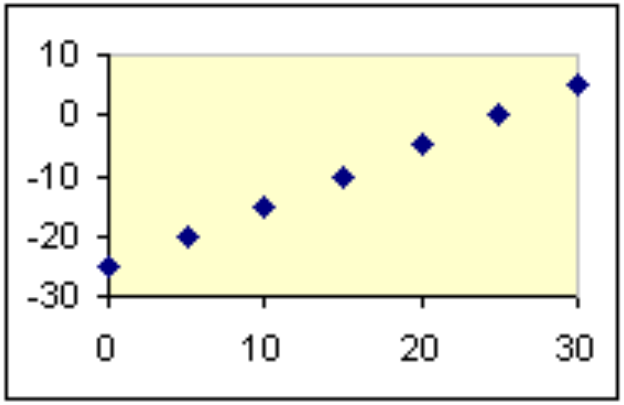
$r = +1$

Amostras perfeitamente correlacionadas em sentido positivo

X	Y
25	0
10	-15
30	5
5	-20
20	-5
0	-25
15	-10

r	1,00



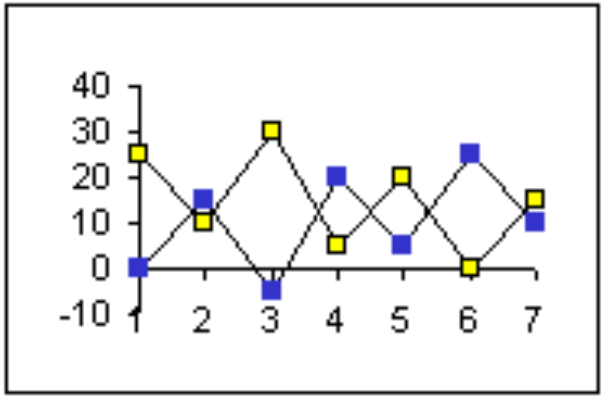


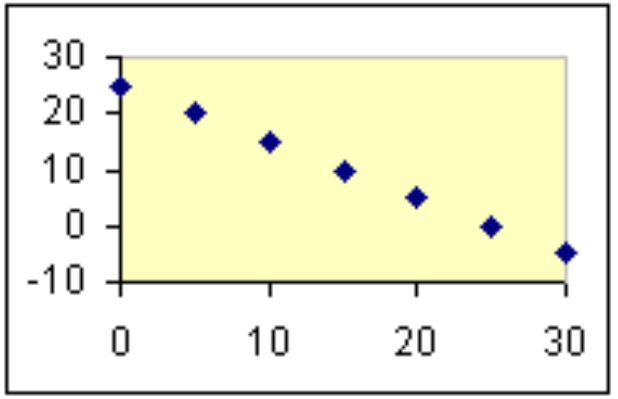
$r = -1$

Amostras perfeitamente correlacionadas em sentido negativo

X	Y
25	0
10	15
30	-5
5	20
20	5
0	25
15	10

r	-1,00



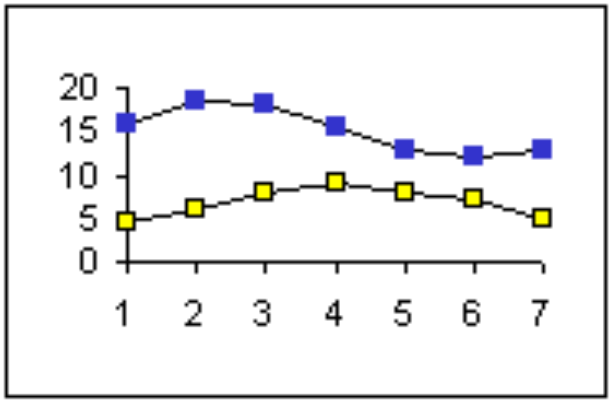


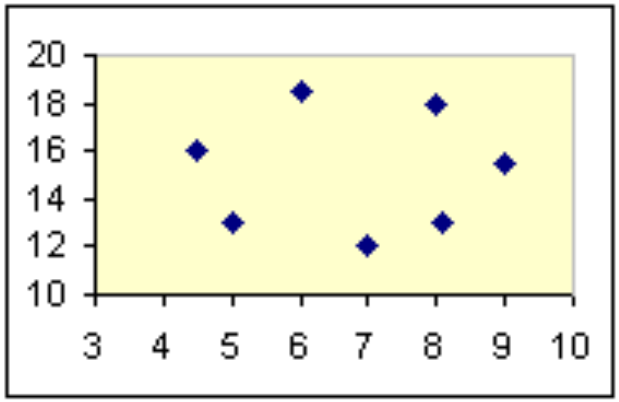
$r = 0$

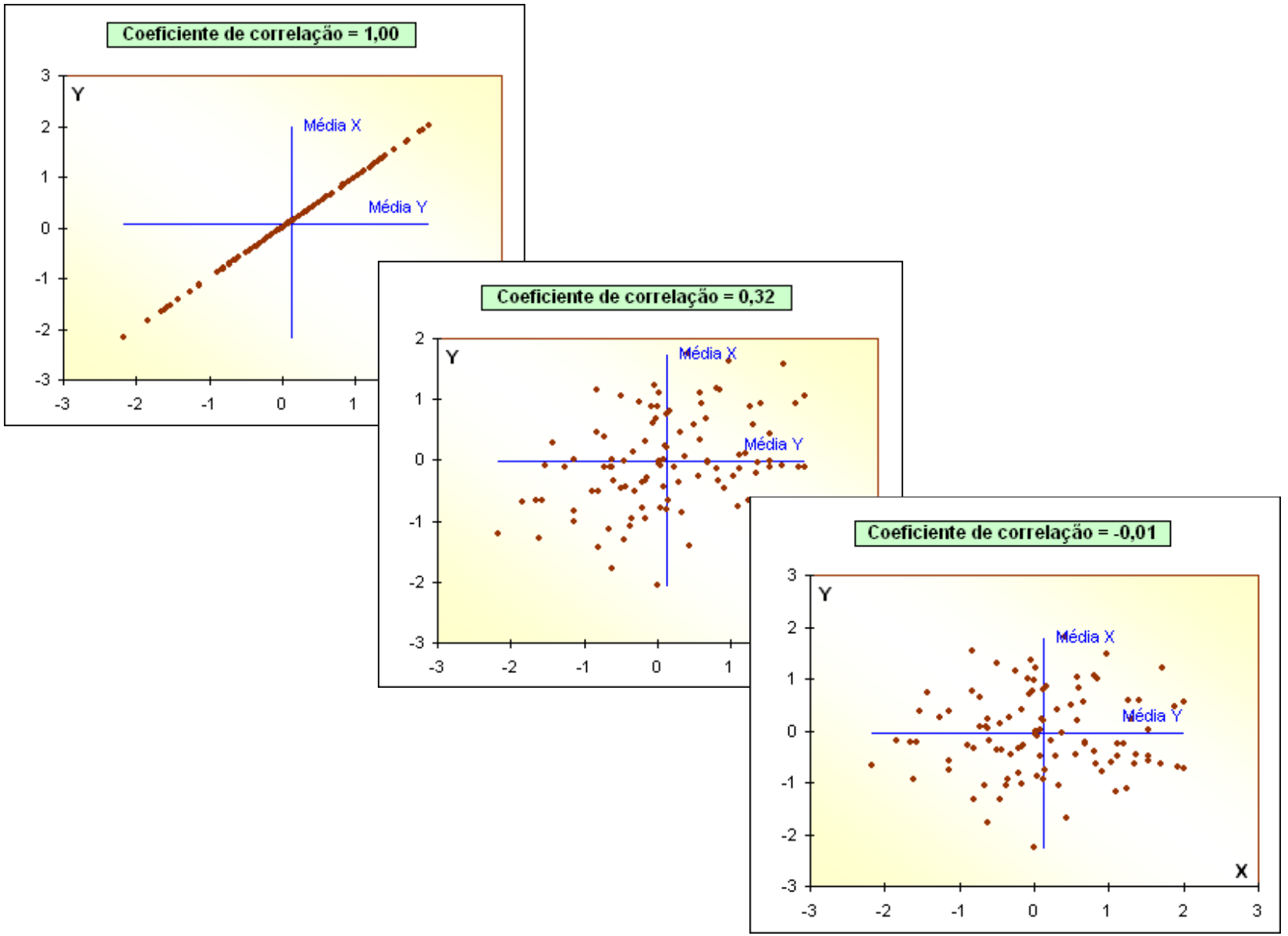
Amostras não correlacionadas

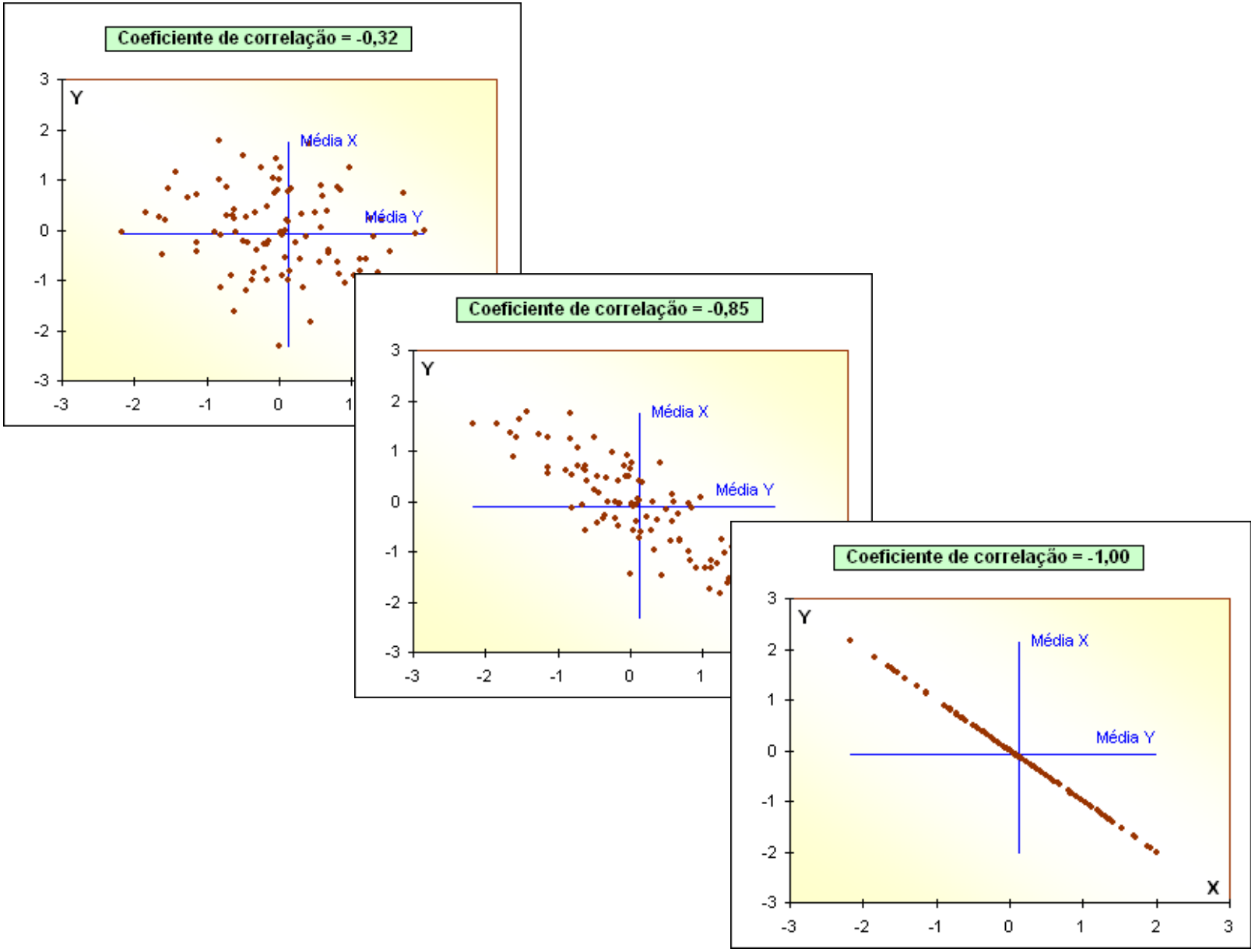
X	Y
4,5	16
6	18,5
8	18
9	15,5
8,1	13
7	12
5	13

r	0,00






Coeficiente de correlação = 1,00
Y
Média X
Média Y
Coeficiente de correlação = 0,32
Y
Média X
Média Y
Coeficiente de correlação = -0,01
Y
Média X
Média Y
X


Coeficiente de correlação = -0,32
Y
Média X
Média Y
Coeficiente de correlação = -0,85
Y
Média X
Média Y
Coeficiente de correlação = -1,00
Y
Média X
Média Y
X

Alguns cuidados

- O coeficiente de correlação **não mede a relação causa-efeito entre as variáveis, apesar de que essa relação possa estar presente**.

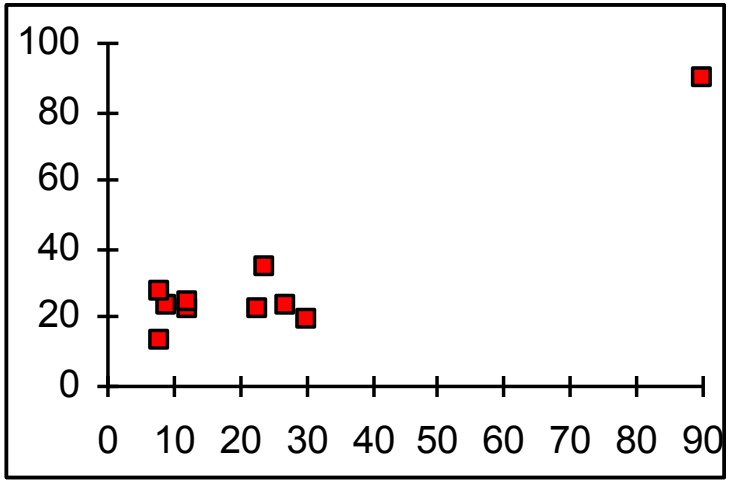
- Por exemplo, uma correlação fortemente positiva entre as variáveis *X* e *Y* não autoriza afirmar que variações da variável *X* provocam variações na variável *Y*, ou vice-versa.

- O coeficiente de correlação **sozinho não identifica a relação causa-efeito entre as duas variáveis**

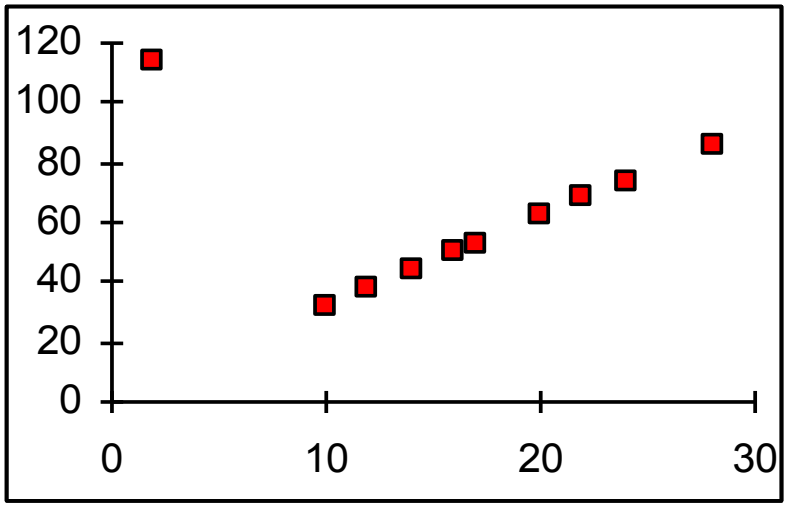
Alguns cuidados

- Em alguns casos, a relação causa-efeito pode ser provocada por um ou mais fatores ocultos, uma variável não considerada na análise.
  - Por exemplo, suponha que o número de vendas diárias de um jornal e a venda diária de ovos tenham uma forte correlação positiva.
  - Não se pode afirmar que o aumento da venda de ovos seja a causa do aumento do número de jornais vendidos, nem que o aumento do número de jornais vendidos resulte no aumento da venda de ovos!
  - Para compreender a forte e positiva correlação, devem procurar fatores ocultos, por exemplo, o aumento de riqueza da população que resulta em aumento de demanda dos dois produtos ao mesmo tempo, jornais e ovos.

Exemplo de anomalia com *r* próximo de +1



Exemplo de anomalias com *r* próximo de 0



Tabelas

- A covariância e o coeficiente de correlação sempre se referem a duas variáveis.
- Quando há mais de duas variáveis, é possível aplicar os conceitos estatísticos considerando as variáveis duas a duas. Nesse caso, as covariâncias e os coeficientes de correlação são registrados numa tabela ou matriz de tamanho definido pelo número de variáveis.
- Para as variáveis *A*, *B* e *C*, as possíveis covariâncias das três variáveis tomadas duas a duas estão registradas na tabela seguinte.

	A	B	C
A	$\sigma_{A,A}$	$\sigma_{A,B}$	$\sigma_{A,C}$
B	$\sigma_{B,A}$	$\sigma_{B,B}$	$\sigma_{B,C}$
C	$\sigma_{C,A}$	$\sigma_{C,B}$	$\sigma_{C,C}$

	A	B	C
A	σ_A^2		
B	$\sigma_{B,A}$	σ_B^2	
C	$\sigma_{C,A}$	$\sigma_{C,B}$	σ_C^2

	A	B	C
A	1		
B	$r_{B.A}$	1	
C	$r_{C.A}$	$r_{C.B}$	1

Exemplo

Evolução do PIB e do consumo da Alemanha entre 1999 e 2008, em milhões de euros correntes

	PIB	Consumo
1999	2012000	1175010
2000	2062500	1214160
2001	2113160	1258570
2002	2143180	1263460
2003	2163800	1284600
2004	2210900	1303090
2005	2243200	1324650
2006	2321500	1355140
2007	2422900	1373720
2008	2491400	1404570

Fonte: Eurostat

Cov= 9702861662
Corr(PIB,Cons)= 0,97603583

	GDP	*Consumption*
GDP	21427055684	
Consumption	9702861662	4612173761

	GDP	*Consumption*
GDP	1	
Consumption	0,976035828	1

Associação entre variáveis qualitativas

- O objetivo de estabelecer a distribuição conjunta de duas variáveis qualitativas é o de compreender a existência de alguma associação entre elas, ou o grau de dependência entre elas

Exemplo: Formados no ensino superior, Argélia, 2007

- Distribuição conjunta de alunos segundo sexo (X) e área de formação

Y \ X	Feminino	Masculino	Total
Ciências humanas e artes	16397	5480	21877
Outras áreas	55045	43246	98291
Total	71442	48726	120168

Fonte: UNESCO

- Olhando assim, não podemos dizer muita coisa a priori

Fixando a distribuição das colunas

- Distribuição conjunta das proporções (em %) de formados segundo sexo (X) e área (Y)

Y \ X	Feminino	Masculino	Total
Ciências humanas e artes	23,0	11,2	18,2
Outras áreas	77,0	88,8	81,8
Total	100,0	100,0	100,0

- Vemos que, independentemente do sexo, cerca de 18% dos estudantes formados em 2007 escolheu a área de ciências humanas de artes
- Vemos, ainda, que a área de humanidades não é "tão" popular assim: embora ela seja mais escolhida entre as mulheres relativamente aos homens, a distribuição não é muito diferente da total
- As variáveis **parecem não serem associadas**

Tomando outro exemplo do livro, p. 77

- Cooperativas autorizadas a funcionar por estado, junho 1974

Estado	Tipo de cooperativa								Total	
	Consumidor		Produtor		Escola		Outras			
São Paulo	214	33%	237	37%	78	12%	119	18%	648	100%
Paraná	51	17%	102	34%	126	42%	22	7%	301	100%
Rio Gr. Sul	111	18%	304	50%	139	23%	48	8%	602	100%
Total	376	24%	643	41%	343	22%	189	12%	1551	100%

- Percebe-se certa dependência entre as variáveis
- Se não houvesse associação, seria de se esperar que em cada estado a distribuição das cooperativas por tipo fosse 24%, 42%, 22% e 12%, respectivamente

Exemplo das cooperativas

- O número esperado de cooperativas, se o padrão fosse o mesmo em todos os estados seria

Estado	Tipo de cooperativa								Total	
	Consumidor		Produtor		Escola		Outras			
São Paulo	157	24%	269	41%	143	22%	79	12%	648	100%
Paraná	73	24%	125	41%	67	22%	37	12%	301	100%
Rio Gr. Sul	146	24%	250	41%	133	22%	73	12%	602	100%
Total	376	24%	643	41%	343	22%	189	12%	1551	100%

- Há, portanto, um desvio entre os valores observados e os esperados

Exemplo das cooperativas

- Os desvios entre os valores esperados e observados podem ser chamados resíduos
- Para calcular os desvios relativos:

  (v observado$_i$-v esperado$_i$)^2/v esperado$_i$

  ou:

$$\frac{(o_i - e_i)^2}{e_i}$$

Ainda com as cooperativas

- O quadro abaixo mostra os desvios
- Os valores na coluna da direita em cada tipo de cooperativa é o cálculo dos desvios relativos

Estado	Tipo de cooperativa							
	Consumidor		Produtor		Escola		Outras	
São Paulo	57	20,62	-32	3,73	-65	29,76	40	20,30
Paraná	-22	6,61	-23	4,16	59	53,07	-15	5,87
Rio Gr. Sul	-35	8,36	54	11,87	6	0,26	-25	8,77

- Somando todos os valores dos desvios relativos, temos:
- Desvios: 20,62+6,61+...+8,77=173,38

Afastamento entre valores observados e esperados

- A soma de todas as medidas de afastamento é uma medida do afastamento global e é chamada qui-quadrado de Pearson e notada χ^2
- Um qui-quadrado grande indica associação entre as variáveis, o que é o caso no nosso exemplo

$$\chi^2 = 173,38$$

Notação

- Se tivermos duas variáveis qualitativas X e Y, classificadas em r categorias para X ($A_1, A_2, ..., A_r$) e s categorias para Y ($B_1, B_2, B_3, ..., B_s$)
- Temos

n_{ij} número de elementos pertencentes à iésima categoria de X e j-ésima categoria de Y

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ número de elementos da iésima categoria de X

$n_{.j} = \sum_{i=1}^{r} n_{ij}$ número de elementos da j-ésima categoria de Y

$n_{..} = n = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ número total de elementos

- O qui-quadrado pode ser escrito $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(n_{ij} - n_{ij}^*\right)^2}{n_{ij}^*}$

Qui-quadrado de Pearson

- Podemos reescrever o qui-quadrado de Pearson em termos de frequências relativas:

$$\chi^2 = n \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\left(f_{ij} - f_{ij}^*\right)^2}{f_{ij}^*}$$

Coeficiente de Contingência

- O coeficiente de contingência é uma medida de associação definida por Pearson do seguinte modo:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

- Contudo, o coeficiente o coeficiente não varia entre 0 e 1. O valor máximo depende de r e s.
- Por isso, foi definido o seguinte coeficiente:

$$T = \sqrt{\frac{\chi^2 / n}{(r-1)(s-1)}}$$

Associação entre variáveis qualitativas e quantitativas

- É comum, neste caso, ver o que ocorre com a variável quantitativa dentro de cada categoria da variável qualitativa
- Pode-se usar gráficos e tabelas para ver o que acontece
- Para verificar o grau de dependência entre as variáveis, precisamos de um indicador
- As variâncias das variáveis é um instrumento
  - A variância da var quantitativa mede a dispersão globalmente
  - Se a variância dentro de cada categoria for pequena e menor do que a global, significa que a var qualitativa melhora a capacidade de previsão da quantitativa e, portanto, existe uma relação entre as variáveis

Associação entre variáveis qualitativas e quantitativas

- Tomemos o exemplo do comportamento dos salários por grau de instrução (Bussab & Morettin, p.86)

Grau de instrução	n	Média	Variância
Fundamental	12	7.84	7.77
Médio	18	11.54	13.1
Superior	6	16.48	16.89
Todos	**36**	**11.12**	**20.46**

Associação entre variáveis qualitativas e quantitativas

- Definimos a média das variâncias, ponderada pelo número de observações em cada categoria:

$$var^*(X) = \sum\left(\frac{n_i \, \mathrm{var}_i(X)}{n}\right)$$

- O grau de associação entre as variáveis é dado pela redução relativa na variância da variável quantitativa através da introdução da variável qualitativa:

$$R^2 = \frac{\left(var(X) - var^*(X)\right)}{var(X)}$$