Correlação e Covariância

Prof. Dr. Evandro Marcos Saidel Ribeiro
FEA-RP
Universidade de São Paulo

RAD1408 - Estatística Aplicada à Administração: Explorar relações entre duas variáveis

É muito frequente, em estatística, explorar possíveis relações entre duas variáveis numéricas. Numa primeira etapa utiliza-se um gráfico de dispersão para visualizar essas possíveis relações.

Gráfico de dispersão: Um gráfico com uma variável numérica no eixo-X e outra variável numérica no eixo-Y. Cada ponto no gráfico corresponde a medições de X e Y para um determinado indivíduo, ou item.

Por exemplo, a figura ao lado apresenta Receitas e Avaliações (ambas em US$ milhões) relativas a times profissionais NBA (tabela completa na próxima página).

Para explorar a possível relação entre as receitas geradas por um time e a avaliação correspondente a esse time, podemos criar um diagrama de dispersão.

Forbes.com — Home Page for the World's Business Leaders

U.S. EUROPE ASIA

Home | Business | Investing | Technology | Entrepreneurs | Op/Ed | Lea

NBA Team Valuations

12.03.08, 06:00 PM EST

RANK	TEAM	CURRENT VALUE[1] ($MIL)	1-YR VALUE CHANGE (%)	DEBT/VALUE[3] (%)	REVENUE[4] ($MIL)	OPERATING INCOME[5] ($MIL)
21	Atlanta Hawks	306	7	23	102	6.7
9	Boston Celtics	447	14	40	149	20.1
29	Charlotte Bobcats	284	-1	53	95	-4.9
3	Chicago Bulls	504	1	11	165	55.4
5	Cleveland Cavaliers	477	5	42	159	13.1
7	Dallas Mavericks	466	1	26	153	-13.6
19	Denver Nuggets	329	3	14	112	-26.3
4	Detroit Pistons	480	1	0	160	40.4
18	Golden State Warriors	335	8	22	112	14.2
6	Houston Rockets	469	1	15	156	31.2
22	Indiana Pacers	303	-9	16	101	-6.5
25	Los Angeles Clippers	297	1	0	99	10.7
2	Los Angeles Lakers	584	4	22	191	47.9
27	Memphis Grizzlies	294	-3	51	95	-3.2
12	Miami Heat	393	-6	43	131	-1.1
30	Milwaukee Bucks	278	5	20	94	5.4
23	Minnesota Timberwolves	301	-2	17	100	-5.7
26	New Jersey Nets	295	-13	71	98	-0.9
28	New Orleans Hornets	285	5	35	95	3.2
1	New York Knicks	613	1	0	208	29.6

https://www.forbes.com/lists/2008/32/nba08_NBA-Team-Valuations_MetroArea.html

Gráfico ou diagrama de dispersão é usado para explorar a possível relação entre duas variáveis. Vamos analisar: Avaliação (Y) e Receita (X) para 30 times da NBA em 2008:

Tabela 1. Avaliações e Receitas para times da NBA (US$ milhões)

Time	Valor	Receita	Time	Valor	Receita
Atlanta	306	102	Milwaukee	278	94
Boston	447	149	Minnesota	301	100
Charlotte	284	95	New Jersey	295	98
Chicago	504	165	New Orleans	285	95
Cleveland	477	159	New York	613	208
Dallas	466	153	Orlando	300	82
Denver	329	112	Philadelphia	349	100
Detroit	480	160	Phoenix	360	116
Golden State	335	112	Portland	452	148
Houston	469	156	Sacramento	307	114
Indiana	303	101	San Antonio	350	117
Los Angeles Clippers	297	99	Seattle	415	138
Los Angeles Lakers	584	191	Toronto	400	138
Memphis	294	95	Utah	358	119
Miami	393	131	Washington	353	118

Fonte: www.forbes.com.lists/2008/32/nba08_NBA-Team-Valuations_MetroArea.html

Relação linear?

Como medir?

Numa segunda etapa da análise da relação entre duas variáveis, podemos utilizar a covariância. A covariância mede a força de uma relação linear entre duas variáveis numéricas (X e Y, por exemplo). A expressão a seguir define a covariância da amostra:

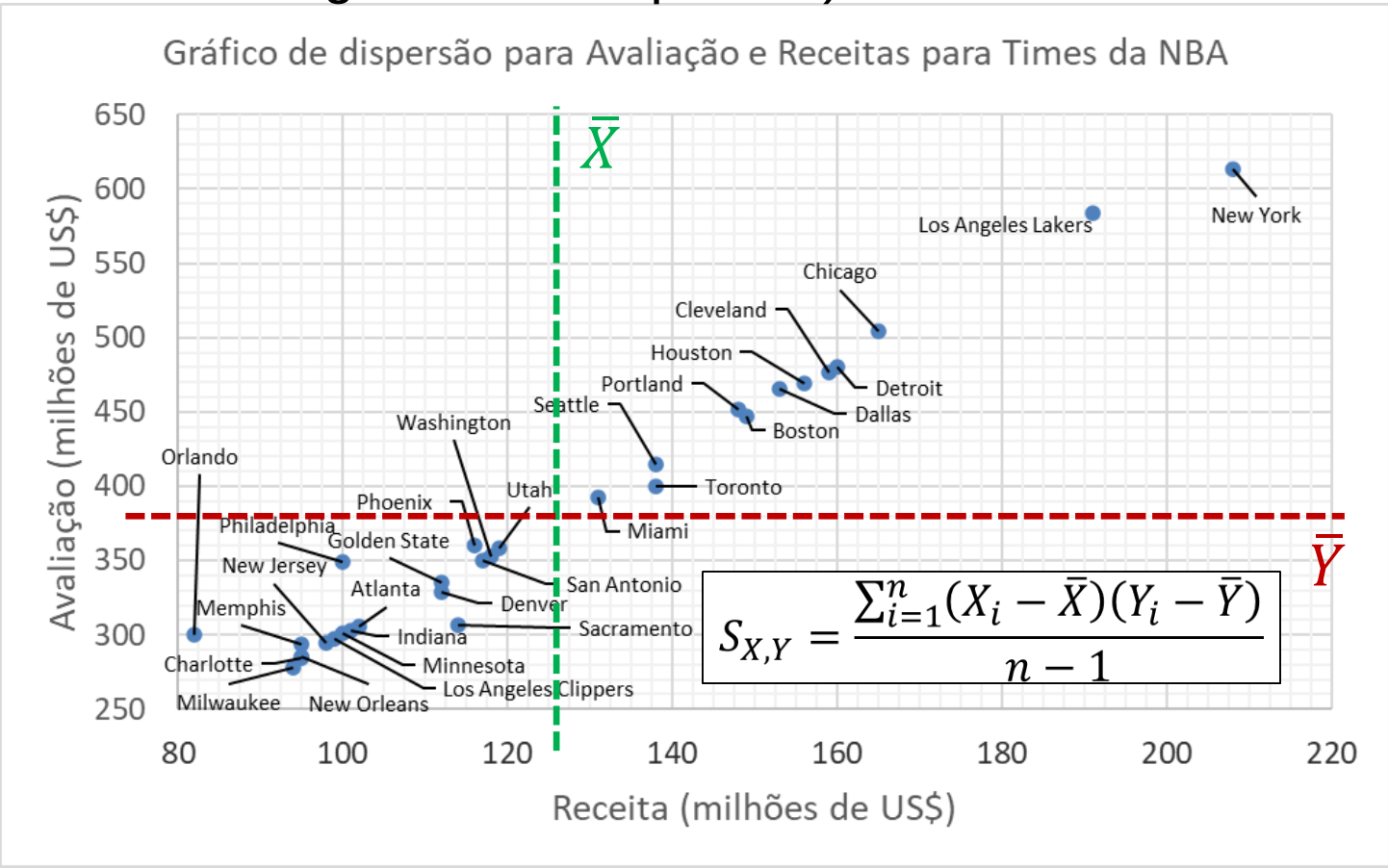
$$cov(X,Y) = S_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

Exemplo

Compare com a expressão da variância de uma variável (variância de X, por exemplo):

$$var(X) = S_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

Vamos calcular a covariância entre Avaliação (Y) e Receita (X), para 30 times da NBA em 2008 (exemplo visto no diagrama de dispersão).



Avaliação	Receita	Y - Ym	X - Xm	$(Y-Y_m)(X-X_m)$
306	102	-73,47	-23,50	1.726,467
447	149	67,53	23,50	1.587,033
284	95	-95,47	-30,50	2.911,733
504	165	124,53	39,50	4.919,067
477	159	97,53	33,50	3.267,367
466	153	86,53	27,50	2.379,667
329	112	-50,47	-13,50	681,300
480	160	100,53	34,50	3.468,400
335	112	-44,47	-13,50	600,300
469	156	89,53	30,50	2.730,767
303	101	-76,47	-24,50	1.873,433
297	99	-82,47	-26,50	2.185,367
584	191	204,53	65,50	13.396,933
294	95	-85,47	-30,50	2.606,733
393	131	13,53	5,50	74,433
278	94	-101,47	-31,50	3.196,200
301	100	-78,47	-25,50	2.000,900
295	98	-84,47	-27,50	2.322,833
285	95	-94,47	-30,50	2.881,233
613	208	233,53	82,50	19.266,500
300	82	-79,47	-43,50	3.456,800
349	100	-30,47	-25,50	776,900
360	116	-19,47	-9,50	184,933
452	148	72,53	22,50	1.632,000
307	114	-72,47	-11,50	833,367
350	117	-29,47	-8,50	250,467
415	138	35,53	12,50	444,167
400	138	20,53	12,50	256,667
358	119	-21,47	-6,50	139,533
353	118	-26,47	-7,50	198,500

Ym	Xm
379,47	125,50

Soma
82.250,0

n =30

Covariância
2.836,2069

RAD1408 - Estatística Aplicada à Administração: Covariância

Resumindo: covariância entre Avaliação (Y) e Receita (X), para 30 times da NBA em 2008:

$$S_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n}(X_i - 125{,}5)(Y_i - 379{,}47)}{29}$$

$$S_{X,Y} = 2.836{,}2069$$

Indica uma relação linear forte ou fraca?

Tem unidade:
Neste caso, [US$ mi]^2

Deficiência da covariância:
Não é uma medida relativa.

Avaliação	Receita	Y - Ym	X - Xm	$(Y-Y_m)(X-X_m)$
306	102	-73,47	-23,50	1.726,467
447	149	67,53	23,50	1.587,033
284	95	-95,47	-30,50	2.911,733
504	165	124,53	39,50	4.919,067
477	159	97,53	33,50	3.267,367
466	153	86,53	27,50	2.379,667
329	112	-50,47	-13,50	681,300
480	160	100,53	34,50	3.468,400
335	112	-44,47	-13,50	600,300
469	156	89,53	30,50	2.730,767
303	101	-76,47	-24,50	1.873,433
297	99	-82,47	-26,50	2.185,367
584	191	204,53	65,50	13.396,933
294	95	-85,47	-30,50	2.606,733
393	131	13,53	5,50	74,433
278	94	-101,47	-31,50	3.196,200
301	100	-78,47	-25,50	2.000,900
295	98	-84,47	-27,50	2.322,833
285	95	-94,47	-30,50	2.881,233
613	208	233,53	82,50	19.266,500
300	82	-79,47	-43,50	3.456,800
349	100	-30,47	-25,50	776,900
360	116	-19,47	-9,50	184,933
452	148	72,53	22,50	1.632,000
307	114	-72,47	-11,50	833,367
350	117	-29,47	-8,50	250,467
415	138	35,53	12,50	444,167
400	138	20,53	12,50	256,667
358	119	-21,47	-6,50	139,533
353	118	-26,47	-7,50	198,500

O coeficiente de correlação de Pearson mede a força relativa de uma relação linear.
É uma medida que se estende de -1 (para correlação negativa perfeita) até +1 (para uma correlação positiva perfeita.

Os casos extremos (-1 e +1), significam que, se desenharmos os pontos num gráfico de dispersão, todos os pontos serão interligados por uma linha reta.

O coeficiente de correlação é uma padronização da covariância. Considerando o desvio padrão da variável X como S_X, o desvio padrão da variável Y como S_Y, o coeficiente de correlação para uma amostra pode ser escrito por:

$$r = \frac{S_{X,Y}}{S_X S_Y}$$

$$r = \frac{S_{X,Y}}{S_X S_Y} = \frac{2.836{,}2069}{31{,}295 \times 91{,}932} = 0{,}9858$$

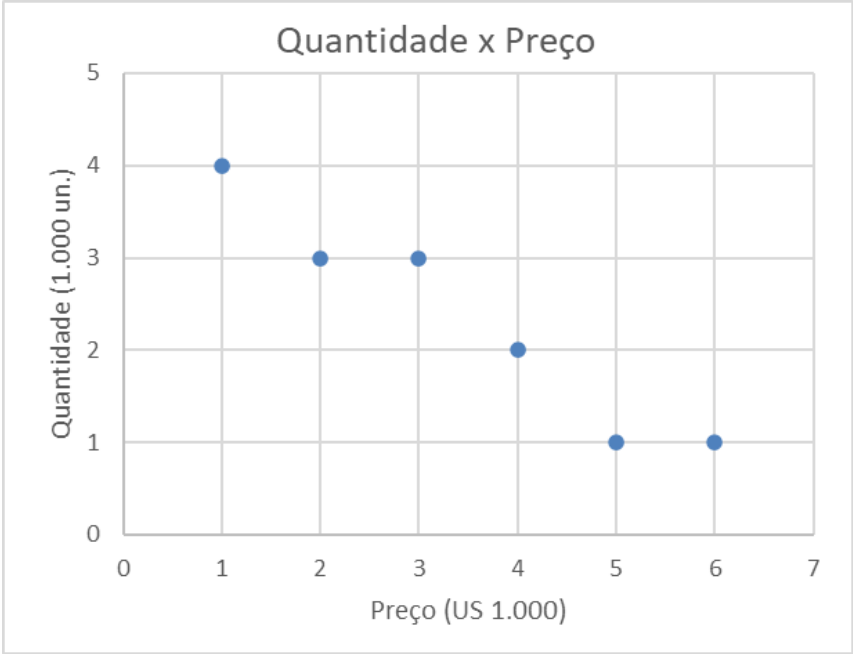
No exemplo de times da NBA:
Covariância: $S_{X,Y}$ = 2.836,2069
Desvio padrão da Receita: S_X = 31,295
Desvio padrão da Avaliação: S_Y = 91,932

A Receita e a Avaliação apresentam correlação amostral positiva muito forte.

Atenção: isto não implica causalidade!

Considerando um exemplo com amostra menor: relação entre Quantidade (Y) e Preço (X)

Quantidade (Q)	Preço (P)	
2	4	$n = 6$
1	6	
3	3	
1	5	
4	1	
3	2	
2,333	3,5	Média
1,211	1,871	Desvio Padrão



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

$$r_{X,Y} = \frac{(\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y}}{(n-1)S_X S_Y}$$

$S_{X,Y} = r_{X,Y} S_X S_Y$	$r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}$

Cálculo no Excel:

$r_{X,Y}$ = PEARSON(valoresX;valoresY)
$S_{X,Y}$ = COVARIAÇÃO.S(valoresX;valoresY)

Cálculo na HP-12C:

S_X = [g] [.], S_Y = [g] [.] [x↔y]
$r_{X,Y}$ = [g] [1] [x↔y]

Considerando um exemplo com amostra menor: relação entre Quantidade (Y) e Preço (X)

Quantidade (Q)	Preço (P)		$X_i\ Y_i$
2	4		8
1	6		6
3	3		9
1	5		5
4	1		4
3	2		6
2,333	3,5	Média	Soma 38
1,211	1,871	Desvio Padrão	n = 6
-0,971		r	
-2,2		$S_{P,Q}$	

$$S_{P,Q} = \frac{38 - 6 \times 3{,}5 \times 2{,}333}{6 - 1}$$

$$S_{P,Q} = \frac{-11}{5} = -2{,}2$$

$$r_{P,Q} = \frac{-2{,}2}{1{,}871 \times 1{,}211}$$

$$r_{P,Q} = -0{,}971$$

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

$$r_{X,Y} = \frac{(\sum_{i=1}^{n} X_i Y_i) - n\bar{X}\bar{Y}}{(n-1)S_X S_Y}$$

$$S_{X,Y} = r_{X,Y} S_X S_Y \qquad r_{X,Y} = \frac{S_{X,Y}}{S_X S_Y}$$

Cálculo no Excel:

$r_{X,Y}$ = PEARSON(valoresX;valoresY)
$S_{X,Y}$ = COVARIAÇÃO.S(valoresX;valoresY)

Cálculo na HP-12C:

S_X = [g] [.], S_Y = [g] [.] [x↔y]
$r_{X,Y}$ = [g] [1] [x↔y]

Quantidade (Q)	Preço (P)
2	4
1	6
3	3
1	5
4	1
3	2

Cálculo no Excel:

$r_{P,Q}$ = PEARSON(valoresP;valoresQ)
CORREL(valoresP;valoresQ)

$S_{P,Q}$ = COVARIAÇÃO.S(valoresP;valoresQ)

	A	B	C
1	Quantidade (Q)	Preço (P)	
2	2	4	
3	1	6	
4	3	3	
5	1	5	
6	4	1	
7	3	2	
8			
9	$r_{X,Y}$ =	-0,9710083	=PEARSON(A2:A7;B2:B7)
10	$r_{X,Y}$ =	-0,9710083	=CORREL(A2:A7;B2:B7)
11			
12	$S_{X,Y}$ =	-2,2	=COVARIAÇÃO.S(A2:A7;B2:B7)
13			

A ordem (X,Y) não importa, pois $S_{X,Y} = S_{Y,X}$

Quantidade (Q)	Preço (P)
2	4
1	6
3	3
1	5
4	1
3	2

Cálculo na HP-12C:

Insira valores aos pares:

Y	[enter]	*X*	[Σ+]
2	[enter]	4	[Σ+]
1	[enter]	6	[Σ+]
3	[enter]	3	[Σ+]
1	[enter]	5	[Σ+]
4	[enter]	1	[Σ+]
3	[enter]	2	[Σ+]

$\bar{P}$=[g] [0]

$\bar{Q}$=[g] [0] [x↔y]

S_P=[g] [.]
= 1,870829

S_Q=[g] [.] [x↔y]
= 1,211060

Correlação = -0,9710

$r_{P,Q}=$ 

Covariância = -2,2

$S_{P,Q} = r_{P,Q} S_P S_Q$

Bibliografia

LEVINE, David M.; STEPHAN, David F.; KREHBIEL, Thimothy C.; BERENSON, Mark L. *Estatística: Teoria e aplicações usando Microsoft® Excel em português*, **6ª ed. Rio de Janeiro: LTC, 2012.**

TRIOLA, M.F; *Introdução à Estatística*, 10ª ed. Rio de Janeiro: LTC, 2008.