FIAP

MBA+

Artificial Intelligence
&
Machine Learning

MBA+

Conceitos Estatísticos para IA

Prof . André Silva de Carvalho
Email: asdc@uol.com.br
www.linkedin.com/in/andresilvadecarvalho
http://lattes.cnpq.br/6876528572507972
2019

Artificial Intelligence & Machine Learning

FIAP

Correlação

Coeficiente de Correlação

Correlação indica a força e a direção do relacionamento linear entre duas **variáveis aleatórias**. No uso estatístico geral, correlação se refere à medida da relação entre duas variáveis, embora correlação não implique **causalidade**. Neste sentido geral, existem vários coeficientes medindo o grau de correlação, adaptados à natureza dos dados.

Vários coeficientes são utilizados para situações diferentes. O mais conhecido é o **coeficiente de correlação de Pearson**, o qual é obtido dividindo a **covariância** de duas variáveis pelo produto de seus **desvios padrão**. Apesar do nome, ela foi apresentada inicialmente por Francis Galton, em meados do século XVII.

Coeficiente de correlação de Pearson, em geral é expresso por (R ou ρ).

Covariância

Em teoria da probabilidade e na estatística, a **covariância**, ou **variância conjunta**, é uma **medida do grau de interdependência** (ou inter-relação) numérica **entre duas variáveis aleatórias**. Assim, **variáveis independentes têm covariância zero**.

A covariância é por vezes chamada de medida de dependência linear entre as duas variáveis aleatórias

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

Covariância

A covariância será **positiva** se as duas variáveis tendem a variar no mesmo sentido, isto é, valores de X acima da sua média estão associados a valores de Y acima de sua média, o mesmo ocorrendo para valores de ambos inferiores à média.

A covariância será **negativa** se valores acima da média de uma variável estão associados a valores inferiores à média da outra

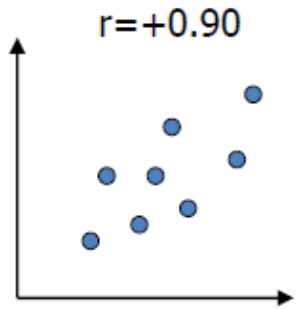
Se X e Y são variáveis aleatórias independentes Cov(X, Y) = 0

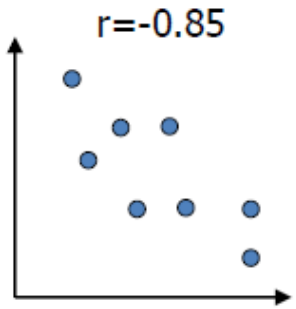
Análise Exploratória de Dados

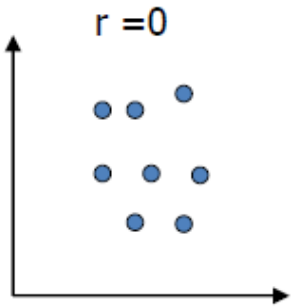
Correlação Linear

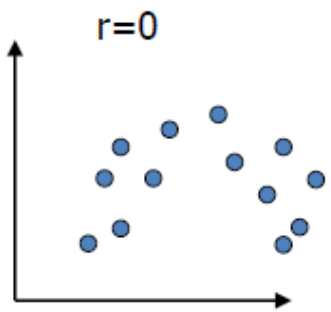
coeficiente de correlação (r) indica a força e a relação linear entre duas variáveis

Valores de r e suas implicações



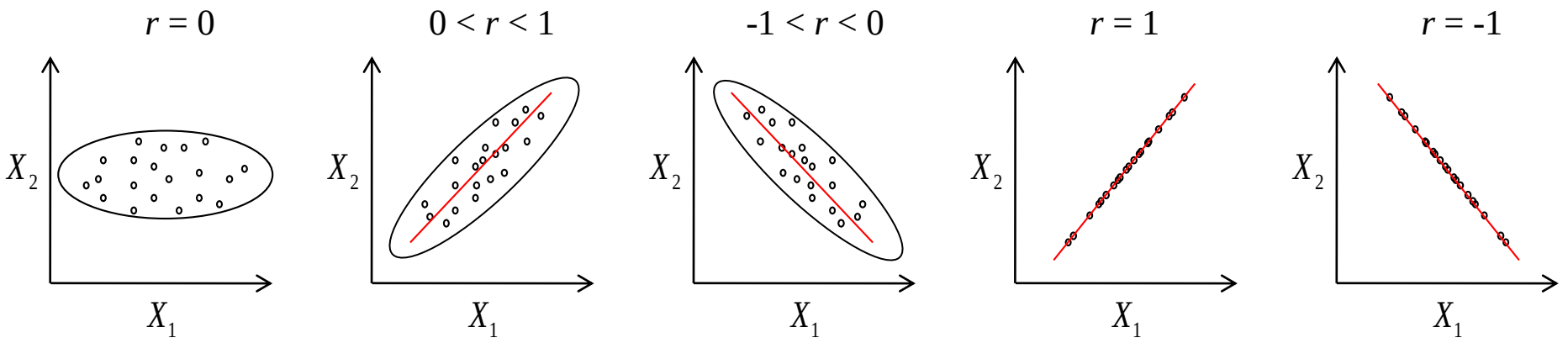






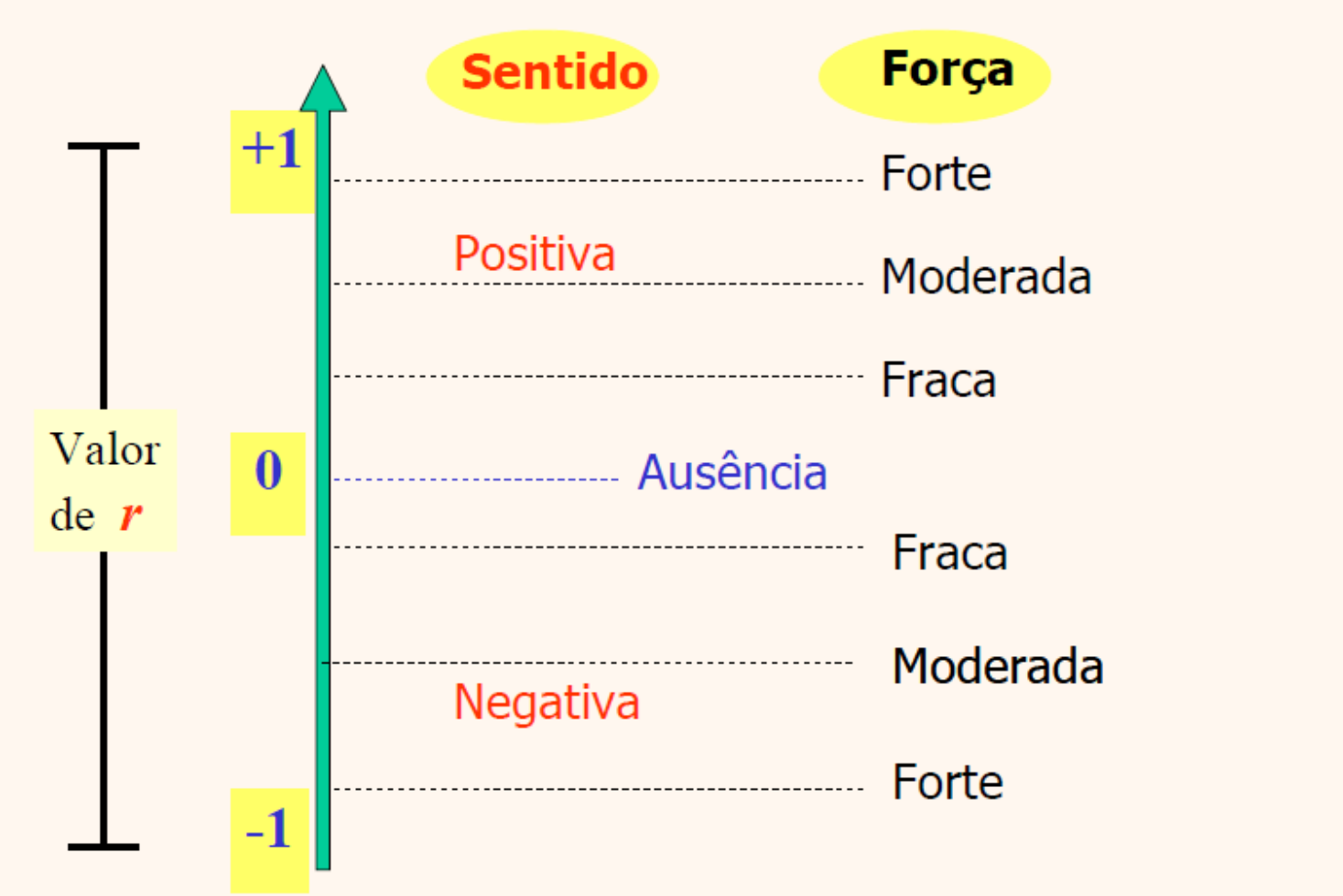
Para avaliar-se a correlação entre variáveis, é importante conhecer a magnitude ou força tanto quanto a significância da correlação.

Associação das variáveis



Quanto maior a variância, maior é a variabilidade e portanto maior a informação contida na variável. Num caso extremo, se a variância é zero, a variável não apresenta nenhuma informação a respeito do fenômeno por ela representada

Interpretação



$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}}$$

Exemplo de Correlação

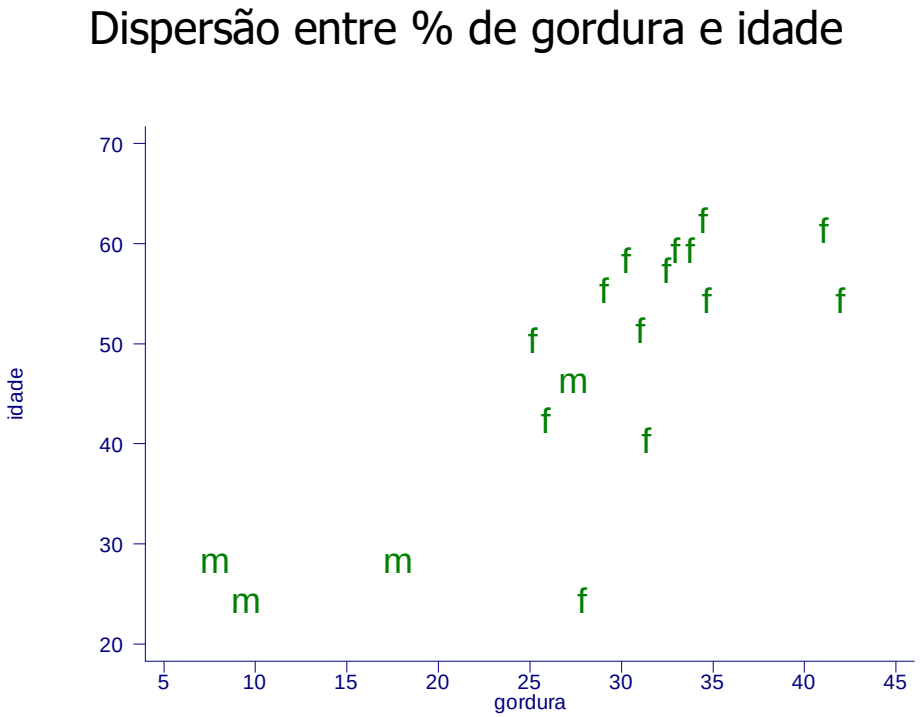
Os dados a seguir são provenientes de um estudo que investiga a composição corporal e fornece o percentual de gordura corporal (%), idade e sexo para 18 adultos com idades entre 23 e 61 anos.

Idade	% gordura	sexo	Idade	% gordura	sexo
23	9,5	M	53	34,7	F
23	27,9	F	53	42,0	F
27	7,8	M	54	29,1	F
27	17,8	M	56	32,5	F
39	31,4	F	57	30,3	F
41	25,9	F	58	33,0	F
45	27,4	M	58	33,8	F
49	25,2	F	60	41,1	F
50	31,1	F	61	34,5	F

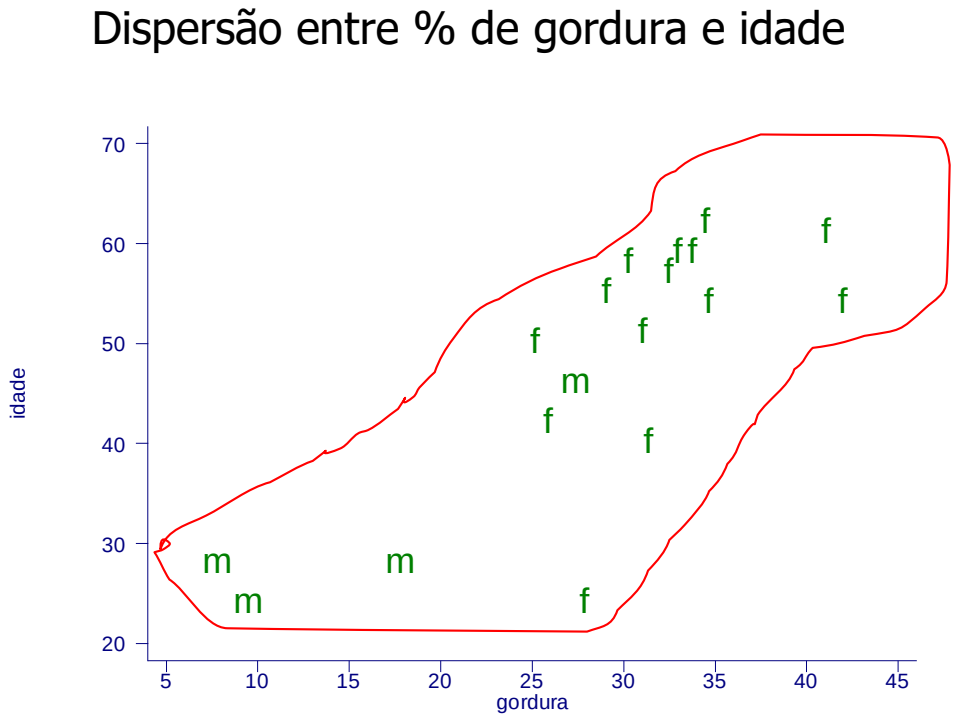
F = feminino

M = masculino

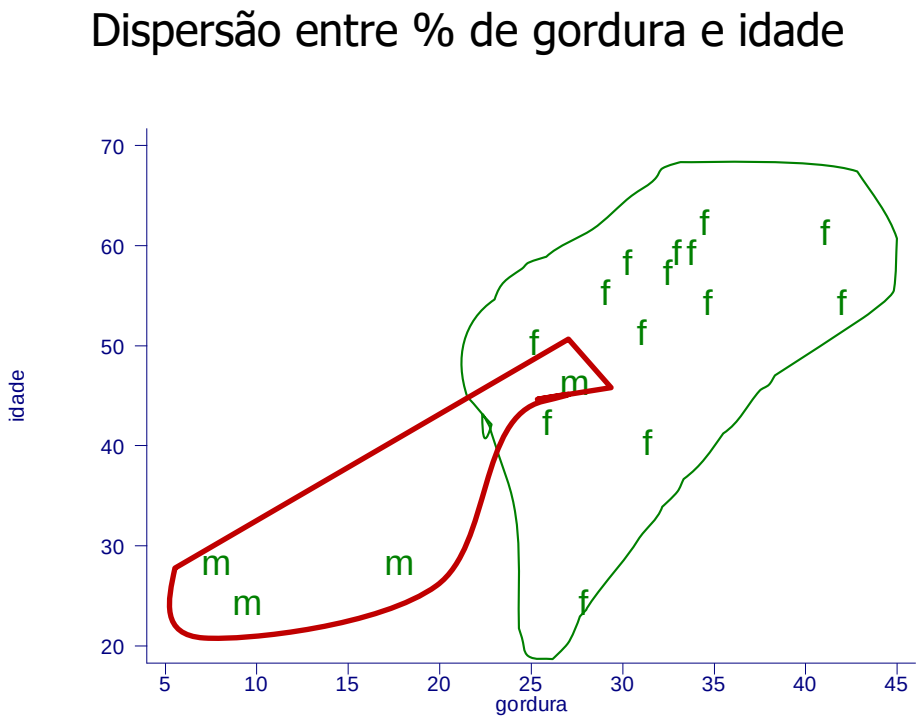
Exemplo de Correlação



Exemplo de Correlação



Exemplo de Correlação



Exemplo de Correlação

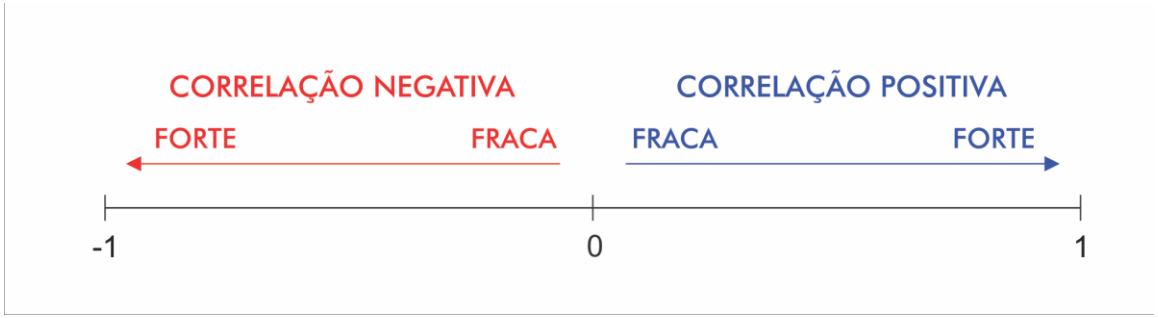
FIAP

Cálculo do coeficiente de correlação de Pearson

Sexo: masculino

Idade	% gordura	$(y-\bar{y})$	$(x-\bar{x})$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
23	9,5	-7,5	-6,13	45,94	56,25	37,52
27	7,8	-3,5	-7,83	27,39	12,25	61,23
27	17,8	-3,5	2,18	-7,61	12,25	4,73
45	27,4	14,5	11,78	170,74	210,25	138,65
$\bar{y}$ =30,5	$\bar{x}$ =15,63		**Total**	**236,45**	**291,00**	**242,13**

Coeficiente de correlação (idade,%gordura) masculino: $r = \frac{236,45}{\sqrt{291x242,13}} = 0,89$

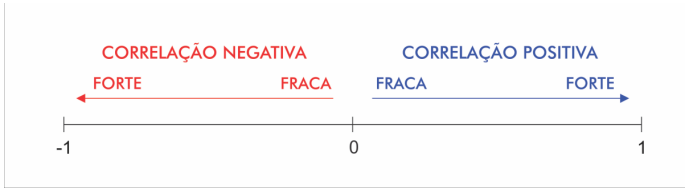


Exemplo de Correlação

Sexo: feminino

Idade	% gordura	$(y-\bar{y})$	$(x-\bar{x})$	$(x-\bar{x})(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})^2$
23	27,9	-27,86	-4,42	123,17	776,02	19,55
39	31,4	-11,86	-0,92	10,93	140,59	0,85
41	25,9	-9,86	-6,42	63,30	97,16	41,23
49	25,2	-1,86	-7,12	13,23	3,45	50,71
50	31,1	-0,86	-1,22	1,05	0,73	1,49
53	34,7	2,14	2,38	5,10	4,59	5,66
53	42	2,14	9,68	20,74	4,59	93,67
54	29,1	3,14	-3,22	-10,12	9,88	10,38
56	32,5	5,14	0,18	0,92	26,45	0,03
57	30,3	6,14	-2,02	-12,42	37,73	4,09
58	33	7,14	0,68	4,85	51,02	0,46
58	33,8	7,14	1,48	10,56	51,02	2,19
60	41,1	9,14	8,78	80,26	83,59	77,06
61	34,5	10,14	2,18	22,10	102,88	4,75
$\bar{y}$ = 50,86	$\bar{x}$ 32,32		**Total**	**333,64**	**1389,71**	**312,12**

Coeficiente de correlação (idade,%gordura) feminino: $r = \frac{333{,}64}{\sqrt{1389{,}71 x 312{,}12}} = 0{,}51$



Exemplo de Correlação

Coeficiente de correlação considerando o grupo todo (homens e mulheres)

Idade (X)	% gordura (Y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$	$(y-\bar{y})^2$
23	9,5	-23,33	-19,11	445,93	544,44	365,23
27	7,8	-19,33	-20,81	402,35	373,78	433,10
27	17,8	-19,33	-10,81	209,01	373,78	116,88
45	27,4	-1,33	-1,21	1,61	1,78	1,47
23	27,9	-23,33	-0,71	16,59	544,44	0,51
39	31,4	-7,33	2,79	-20,45	53,78	7,78
41	25,9	-5,33	-2,71	14,46	28,44	7,35
49	25,2	2,67	-3,41	-9,10	7,11	11,64
50	31,1	3,67	2,49	9,13	13,44	6,19
53	34,7	6,67	6,09	40,59	44,44	37,07
53	42	6,67	13,39	89,26	44,44	179,26
54	29,1	7,67	0,49	3,75	58,78	0,24
56	32,5	9,67	3,89	37,59	93,44	15,12
57	30,3	10,67	1,69	18,01	113,78	2,85
58	33	11,67	4,39	51,20	136,11	19,26
58	33,8	11,67	5,19	60,54	136,11	26,92
60	41,1	13,67	12,49	170,68	186,78	155,97
61	34,5	14,67	5,89	86,37	215,11	34,68
			Soma	1627,53	2970,00	1421,54

0,79

$\bar{x} = 46{,}33$; $\bar{y} = 28{,}61$

Correlação Postos de Sperarman

Exige que ambas as variáveis se apresentem em escala de mensuração, de modo que os objetos ou indivíduos em estudo possam dispor-se por postos em duas séries ordenadas.

A medida a ser usada são as di´s, onde: $d_i = x_i - y_i$

$$rs = 1 - \frac{6\sum di^2}{n * \left(n^2 - 1\right)}$$

Correlação Postos de Sperarman

Exemplo

Escores referentes a autoritarismo e aspirações de status social

ESTUDANTE	SCORE AUTORITARISMO	SCORE ASPIRAÇÃO
A	82	42
B	98	46
C	87	39
D	40	37
E	116	65
F	113	88
G	111	86
H	83	56
I	85	62
J	126	92
K	106	54
L	107	81

Correlação Postos de Sperarman

Exemplo

Escores referentes a autoritarismo e aspirações de status social

ESTUDANTE	SCORE AUTORITARISMO	SCORE ASPIRAÇÃO	POSTO AUTORITARISMO	POSTO ASPIRAÇÃO
A	82	42	2	3
B	98	46	6	4
C	87	39	5	2
D	40	37	1	1
E	116	65	10	8
F	113	88	9	11
G	111	86	8	10
H	83	56	3	6
I	85	62	4	7
J	126	92	12	12
K	106	54	7	5
L	107	81	11	9

Correlação Postos de Sperarman

FIAP

Exemplo

Escores referentes a autoritarismo e aspirações de status social

ESTUDANTE	SCORE AUTORITARISMO	SCORE ASPIRAÇÃO	POSTO AUTORITARISMO	POSTO ASPIRAÇÃO	di	di^2
A	82	42	2	3	-1	1
B	98	46	6	4	2	4
C	87	39	5	2	3	9
D	40	37	1	1	0	0
E	116	65	10	8	2	4
F	113	88	9	11	-2	4
G	111	86	8	10	-2	4
H	83	56	3	6	-3	9
I	85	62	4	7	-3	9
J	126	92	12	12	0	0
K	106	54	7	5	2	4
L	107	81	11	9	2	4

Correlação Postos de Sperarman

Exemplo

$$rs = 1 - \frac{6\sum di^2}{n*\left(n^2-1\right)}$$

$$rs = 1 - \left(\frac{6*52}{12*\left(12^2-1\right)}\right)$$

A correlação entre autoritarismo e o grau de aspiração a status social é: $rs = 0.82$

Associações Espúrias

FIAP

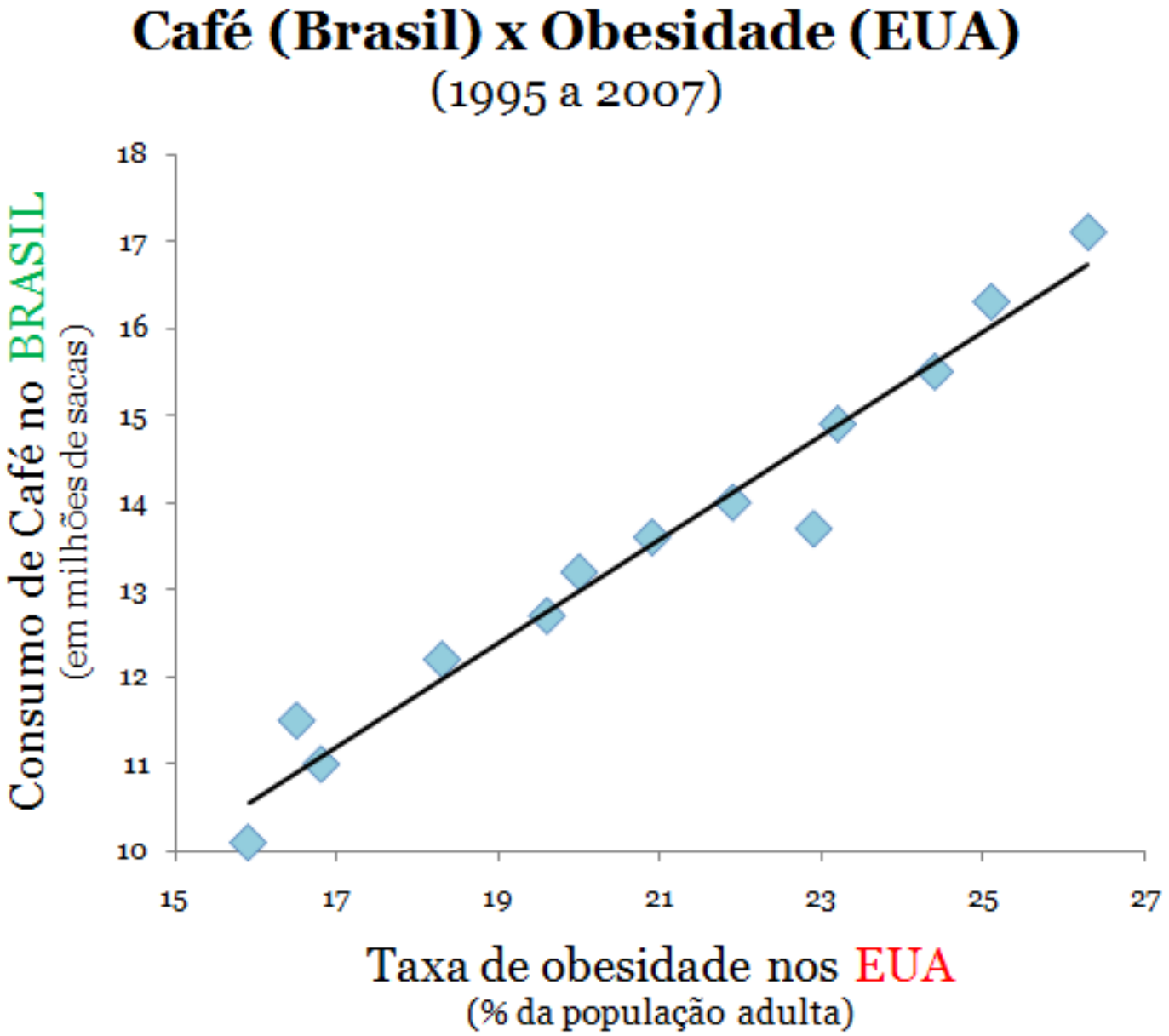
- Associação entre dois fatores e quando queremos saber se um **causa** o outro ?

- big data muitos resultados estatisticamente significativos que não fazem sentido causal

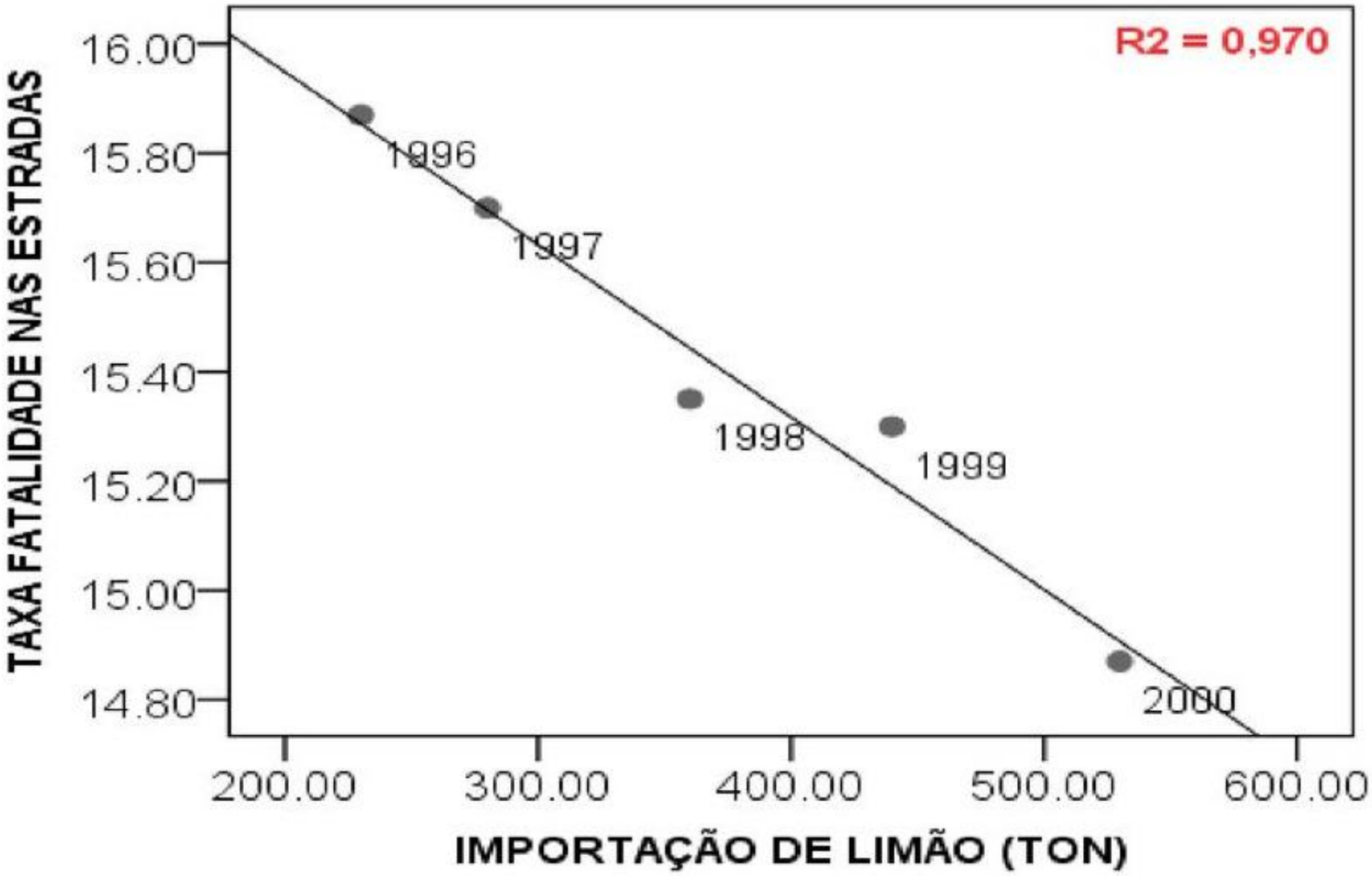
- variável de confusão quando há muitas variáveis na análise

Uma relação estatística existente entre duas variáveis, mas onde não existe nenhuma relação causa-efeito entre elas. Essa relação estatística pode ocorrer por pura coincidência ou por causa de uma terceira variável.

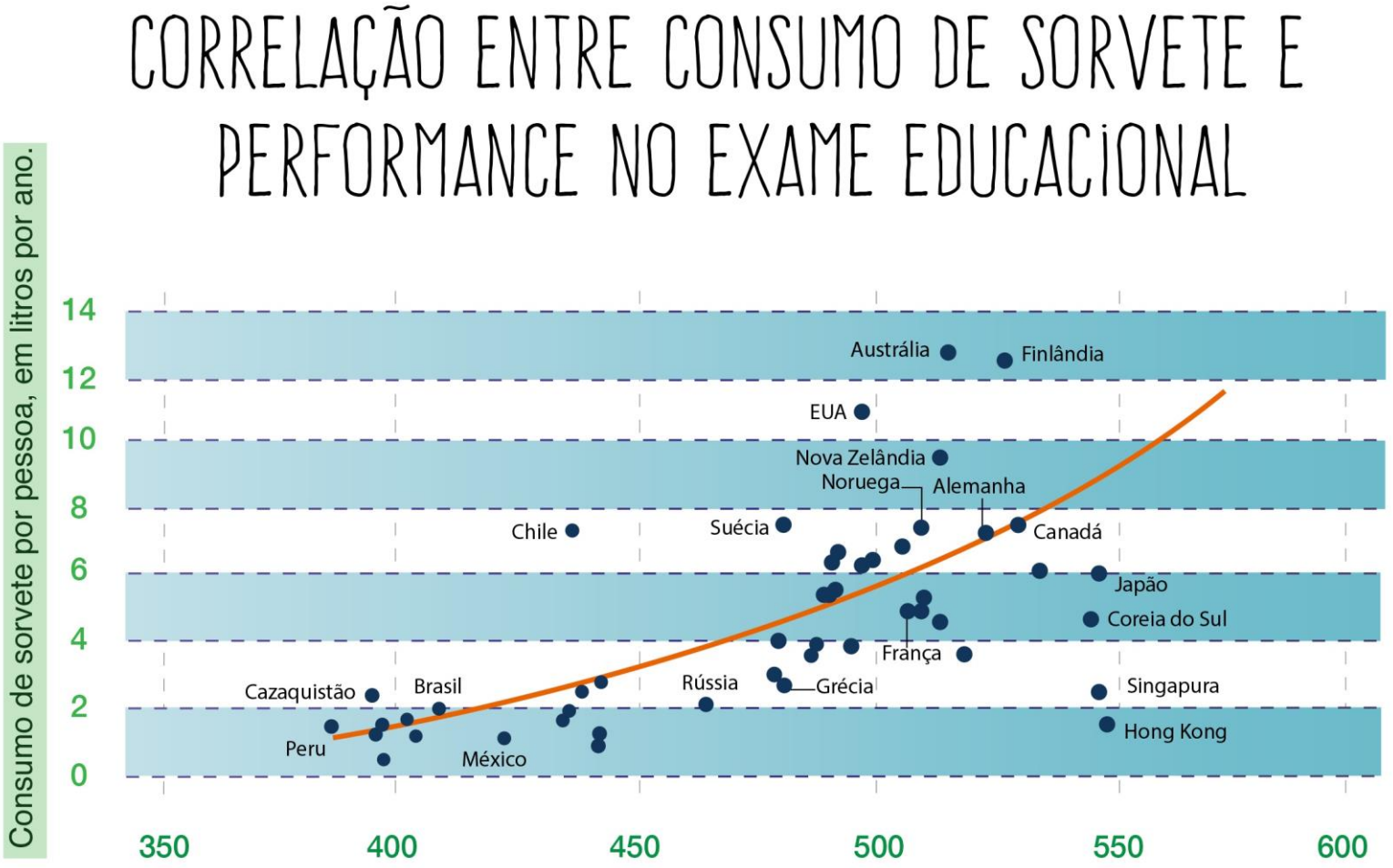
Correlações Espúrias



Correlações Espúrias



Correlações Espúrias



*Fontes: OCDE; Euromonitor; The Economist

Correlações Espúrias

Obrigado,
por enquanto !

FIAP

MBA+