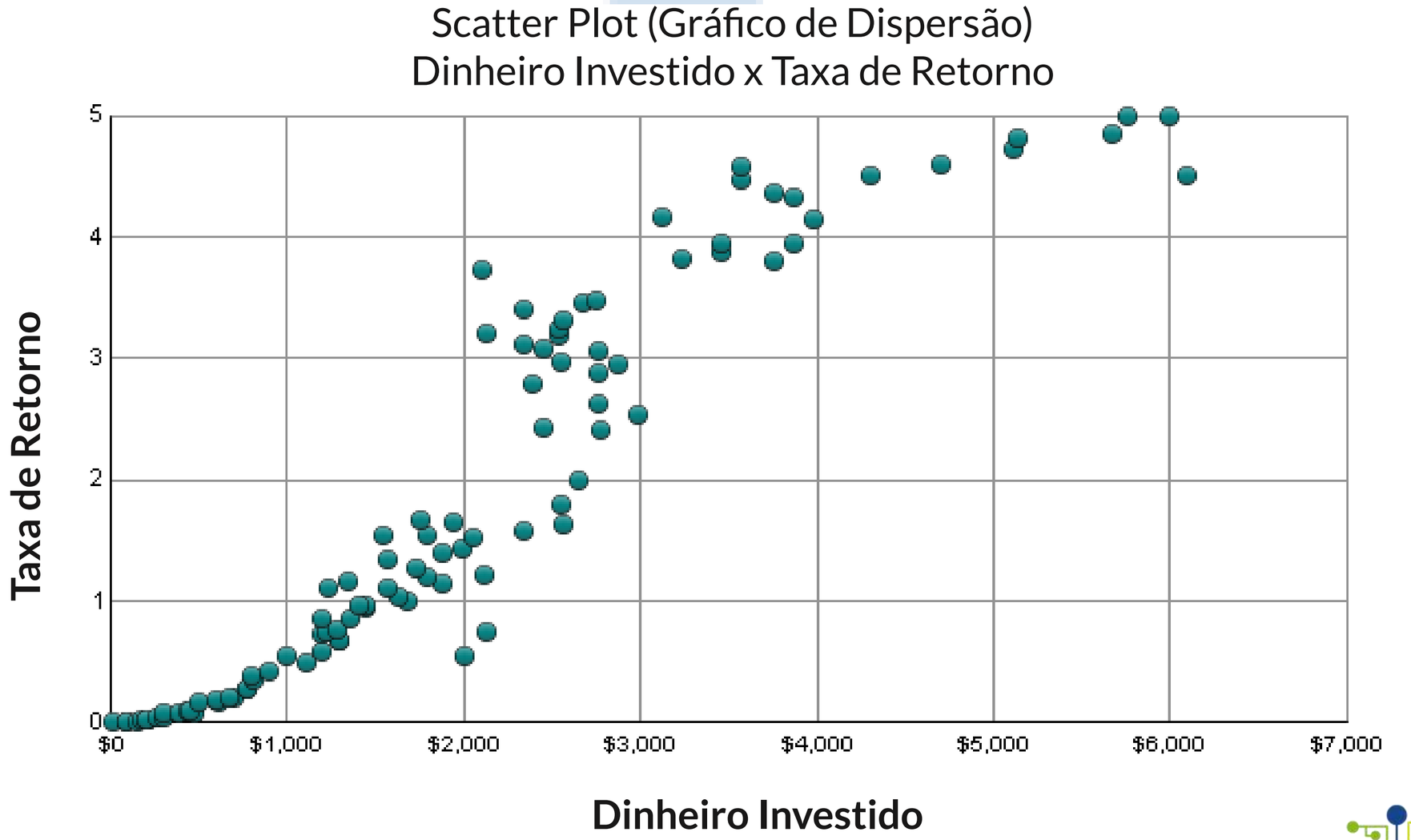
Análise Estatística para Data Science I com R e SAS

Análise Estatística para Data Science I com R e SAS

Associações e Correlações

Data Science Academy

Associações e Correlações



Análise Estatística para Data Science I com R e SAS

Análise Bidimensional

Data Science Academy

Análise Bidimensional

Na Análise Bidimensional estamos interessados em estudar e analisar o comportamento conjunto de duas ou mais variáveis aleatórias.

Os dados aparecem na forma de uma matriz, com as colunas representando as variáveis e as linhas representando os indivíduos (ou elementos, ou observações, ou registros).

Indivíduo	Variável					
	X_1	X_2	...	X_j	...	X_p
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
⋮	⋮	⋮		⋮		⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
⋮	⋮	⋮		⋮		⋮
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{np}

Análise Bidimensional

Quando consideramos duas variáveis podemos ter três situações:

As duas variáveis são qualitativas.

As duas variáveis são quantitativas.

Uma variável é qualitativa e a outra é quantitativa.

Análise Estatística para Data Science I com R e SAS

Associação Entre Variáveis Qualitativas

Data Science Academy

Associação Entre Variáveis Qualitativas

Suponha que queiramos analisar o comportamento conjunto das variáveis:
Y = Grau de Instrução
V = Região de Procedência

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

A linha dos totais fornece a distribuição da variável Y, ao passo que a coluna dos totais fornece a distribuição da variável V. As distribuições assim obtidas são chamadas tecnicamente de **distribuições marginais**, enquanto a tabela constitui a **distribuição conjunta** de Y e V.

Associação Entre Variáveis Qualitativas

Suponha que queiramos analisar o comportamento conjunto das variáveis:

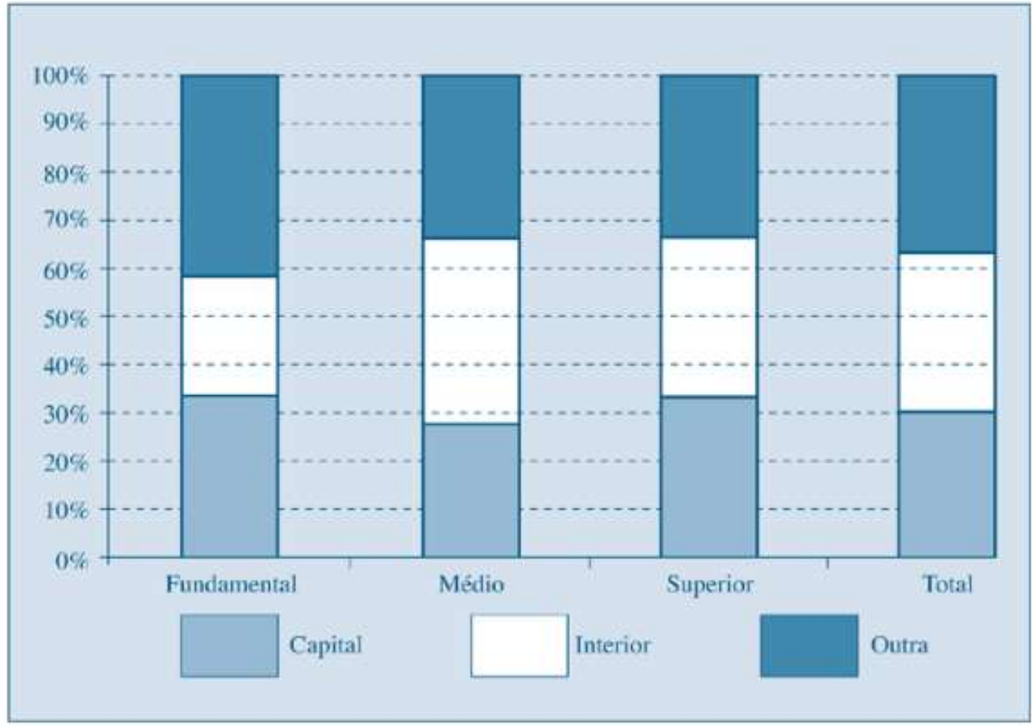
Y = Grau de Instrução

V = Região de Procedência

V \ Y	Fundamental	Médio	Superior	Total
Capital	11%	14%	6%	31%
Interior	8%	19%	6%	33%
Outra	14%	17%	5%	36%
Total	33%	50%	17%	100%

Associação Entre Variáveis Qualitativas

Suponha que queiramos analisar o comportamento conjunto das variáveis:
Y = Grau de Instrução
V = Região de Procedência



Stacked Bar Chart

Associação Entre Variáveis Qualitativas

Um dos principais objetivos de se construir uma distribuição conjunta de duas **variáveis qualitativas** é descrever a associação entre elas, isto é, queremos conhecer o grau de dependência entre elas, de modo que possamos prever melhor o resultado de uma delas quando conhecermos a realização da outra.

Associação Entre Variáveis Qualitativas

Y \ X	Masculino	Feminino	Total
Economia	85	35	120
Administração	55	25	80
Total	140	60	200

Y \ X	Masculino	Feminino	Total
Economia	61%	[illegible]	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Associação Entre Variáveis Qualitativas

Y \ X	Masculino	Feminino	Total
Economia	61%	58%	60%
Administração	39%	42%	40%
Total	100%	100%	100%

Observando a tabela, vemos que as proporções do sexo masculino (61% e 39%) e do sexo feminino (58% e 42%) são próximas das marginais (60% e 40%). Esses resultados parecem indicar não haver dependência entre as duas variáveis, para o conjunto de alunos considerado. Concluímos então que, neste caso, as variáveis sexo e escolha do curso parecem ser não associadas.

Associação Entre Variáveis Qualitativas

Y \ X	Masculino	Feminino	Total
Física	100 (71%)	20 (33%)	120 (60%)
Ciências Sociais	40 (29%)	40 (67%)	80 (40%)
Total	140 (100%)	60 (100%)	200 (100%)

Comparando agora a distribuição das proporções pelos cursos, independentemente do sexo (coluna de totais), com as distribuições diferenciadas por sexo (colunas de masculino e feminino), observamos uma disparidade bem acentuada nas proporções. Parece, pois, haver maior concentração de homens no curso de Física e de mulheres no de Ciências Sociais. Portanto, nesse caso, as variáveis sexo e curso escolhido parecem ser associadas.

Análise Estatística para Data Science I com R e SAS

Medidas de Associação Entre Variáveis Qualitativas

Data Science Academy

Medidas de Associação Entre Variáveis Qualitativas

De modo geral, a quantificação do grau de associação entre duas variáveis é feita pelos chamados **coeficientes de associação ou correlação**. Essas são medidas que descrevem, por meio de um único número, a associação (ou dependência) entre duas variáveis.

Para facilitar a compreensão, esses coeficientes usualmente variam entre 0 e 1, ou entre −1 e +1, e a proximidade de zero indica falta de associação.

Medidas de Associação Entre Variáveis Qualitativas

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 G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

A análise da tabela mostra a existência de certa dependência entre as variáveis. Caso não houvesse associação, esperaríamos que em cada estado tivéssemos 24% de cooperativas de consumidores, 42% de cooperativas de produtores, 22% de escolas e 12% de outros tipos. Então, por exemplo, o número esperado de cooperativas de consumidores no Estado de São Paulo seria 648 × 0,24 = 157 e no Paraná seria 301 × 0,24 = 73

Medidas de Associação Entre Variáveis Qualitativas

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 G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores observados.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	−32 (3,81)	−65 (29,55)	40 (20,25)
Paraná	−22 (6,63)	−22 (3,90)	59 (51,96)	−15 (6,08)
Rio G. do Sul	−35 (8,39)	54 (11,66)	6 (0,27)	−25 (8,56)

Desvios entre observados e esperados.

Medidas de Associação Entre Variáveis Qualitativas

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 G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores observados.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	−32 (3,81)	−65 (29,55)	40 (20,25)
Paraná	−22 (6,63)	−22 (3,90)	59 (51,96)	−15 (6,08)
Rio G. do Sul	−35 (8,39)	54 (11,66)	6 (0,27)	−25 (8,56)

Desvios entre observados e esperados.

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

Medidas de Associação Entre Variáveis Qualitativas

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 G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores observados.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores esperados.

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

Aplicando a fórmula para Escola-São Paulo obtemos (−65)^2/143 = 29,55 e para Escola-Paraná obtemos (59)^2/67 = 51,96, o que é uma indicação de que o desvio devido a essa última combinação é “maior” do que aquele da primeira.

Uma medida do afastamento global pode ser dada pela soma de todas as medidas. Essa medida é denominada **χ2 (qui-quadrado) de Pearson**.

Medidas de Associação Entre Variáveis Qualitativas

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 G. do Sul	111 (18%)	304 (51%)	139 (23%)	48 (8%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores observados.

Estado	Tipo de Cooperativa				Total
	Consumidor	Produtor	Escola	Outras	
São Paulo	157 (24%)	269 (42%)	143 (22%)	79 (12%)	648 (100%)
Paraná	73 (24%)	124 (42%)	67 (22%)	37 (12%)	301 (100%)
Rio G. do Sul	146 (24%)	250 (42%)	133 (22%)	73 (12%)	602 (100%)
Total	376 (24%)	643 (42%)	343 (22%)	189 (12%)	1.551 (100%)

Valores esperados.

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	−32 (3,81)	−65 (29,55)	40 (20,25)
Paraná	−22 (6,63)	−22 (3,90)	59 (51,96)	−15 (6,08)
Rio G. do Sul	−35 (8,39)	54 (11,66)	6 (0,27)	−25 (8,56)

Desvios entre observados e esperados.

$$\chi^2 = 20{,}69 + 6{,}63 + \ldots + 8{,}56 = 171{,}76$$

Medidas de Associação Entre Variáveis Qualitativas

Estado	Tipo de Cooperativa			
	Consumidor	Produtor	Escola	Outras
São Paulo	57 (20,69)	−32 (3,81)	−65 (29,55)	40 (20,25)
Paraná	−22 (6,63)	−22 (3,90)	59 (51,96)	−15 (6,08)
Rio G. do Sul	−35 (8,39)	54 (11,66)	6 (0,27)	−25 (8,56)

Desvios entre observados e esperados.

$$\chi^2 = 20,69 + 6,63 + \ldots + 8,56 = 171,76$$

Um valor grande de χ2 indica associação entre as variáveis, o que parece ser o caso.

Medidas de Associação Entre Variáveis Qualitativas

Um valor grande de χ2 (qui-quadrado) indica associação entre as variáveis.

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

O Coeficiente de Contingência é uma medida de associação entre variáveis qualitativas.

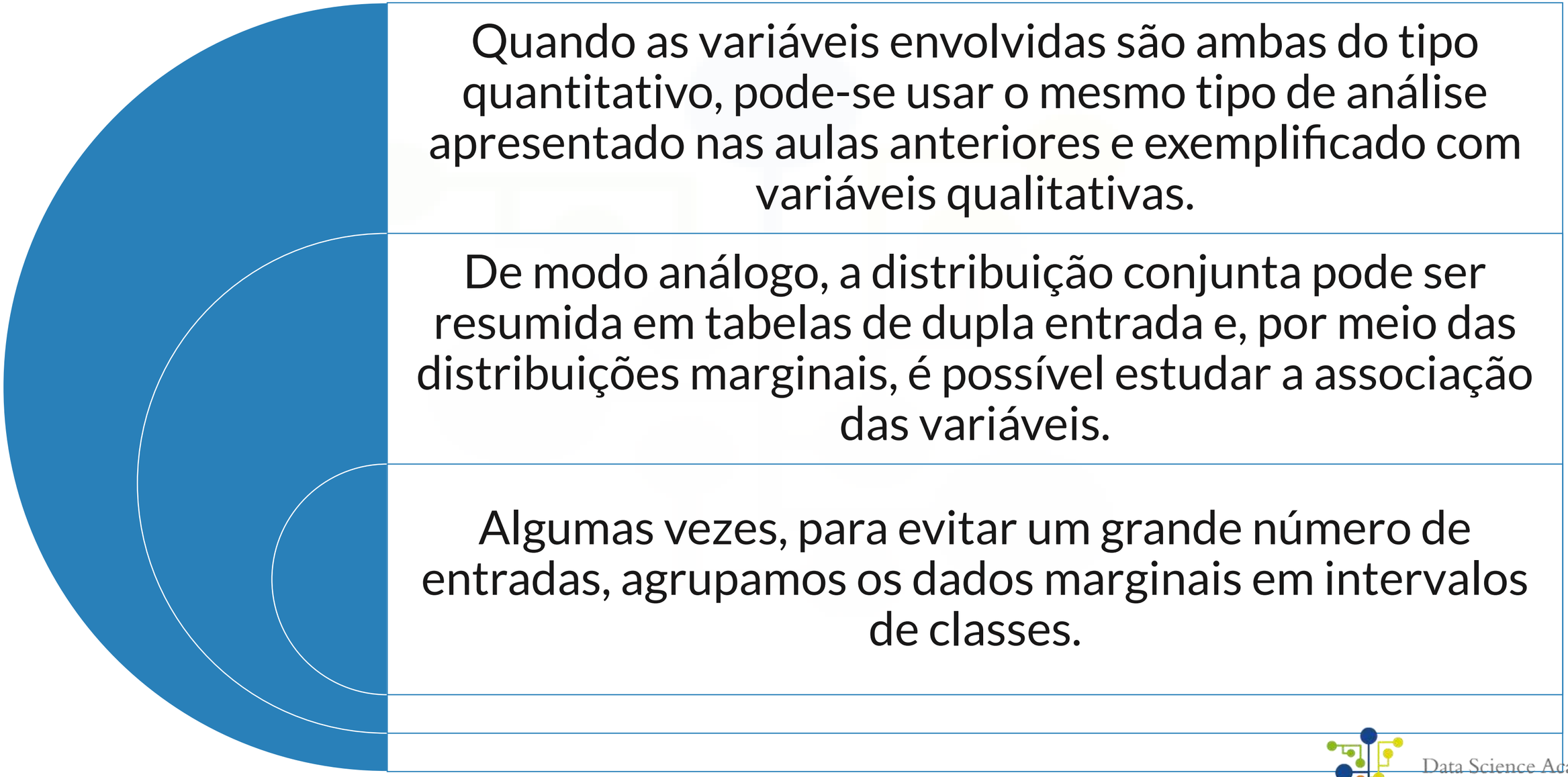
$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \quad \rightarrow \quad T = \sqrt{\frac{\chi^2 / n}{(r-1)(s-1)}}$$

Análise Estatística para Data Science I com R e SAS

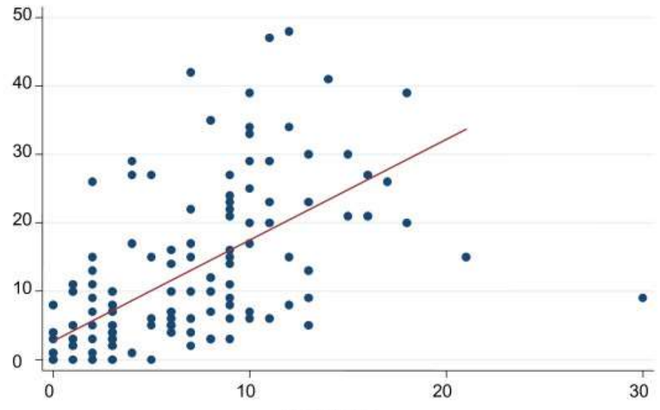
Associação Entre Variáveis Quantitativas

Data Science Academy

Associação Entre Variáveis Quantitativas



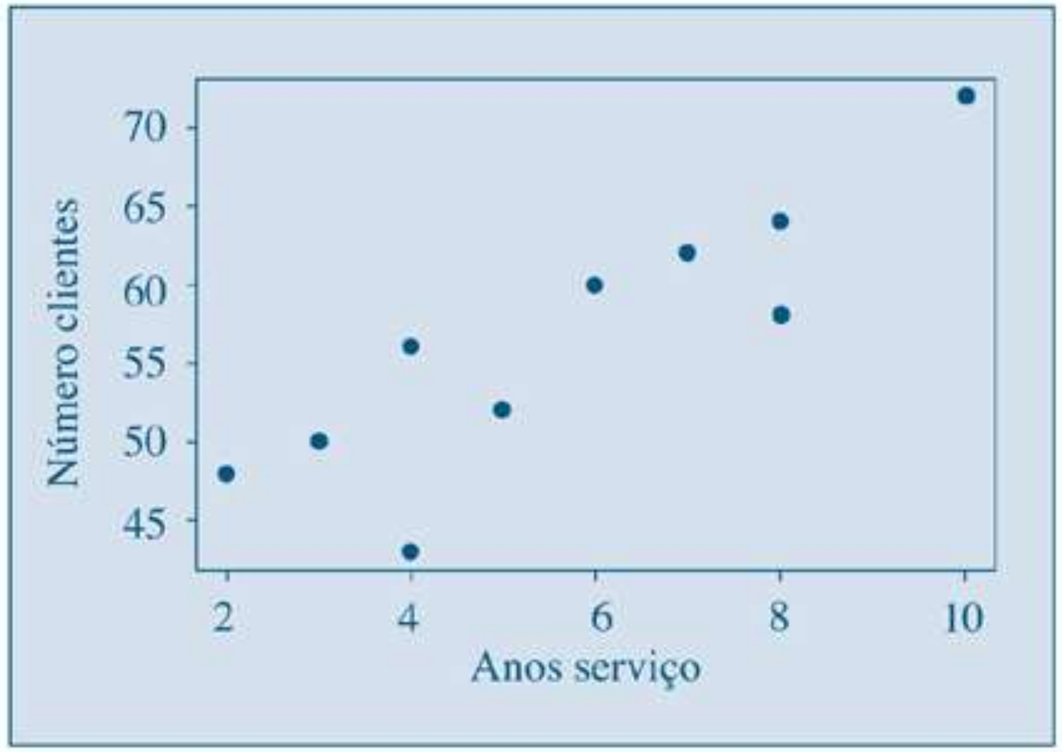
Associação Entre Variáveis Quantitativas



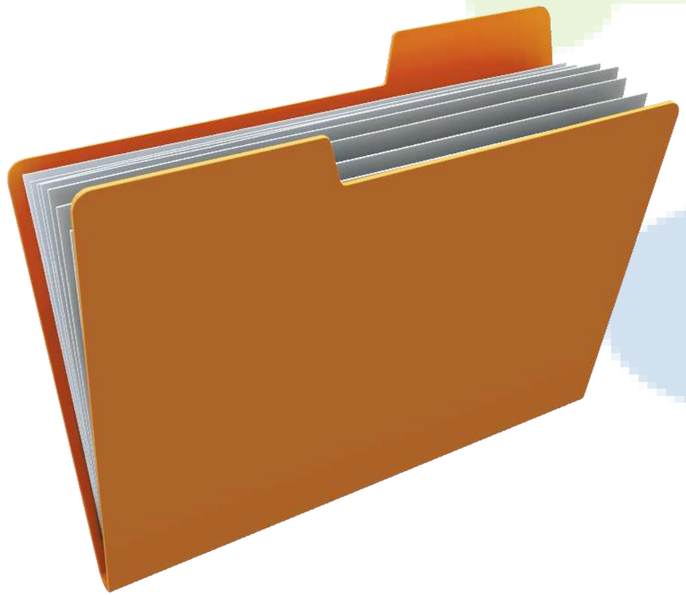
Mas, além desse tipo de análise, as variáveis quantitativas são passíveis de procedimentos analíticos e gráficos mais refinados. Um dispositivo bastante útil para se verificar a associação entre duas variáveis quantitativas é o gráfico de dispersão (ou Scatter Plot).

Associação Entre Variáveis Quantitativas

Agente	Anos de serviço (X)	Número de 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



Associação Entre Variáveis Quantitativas

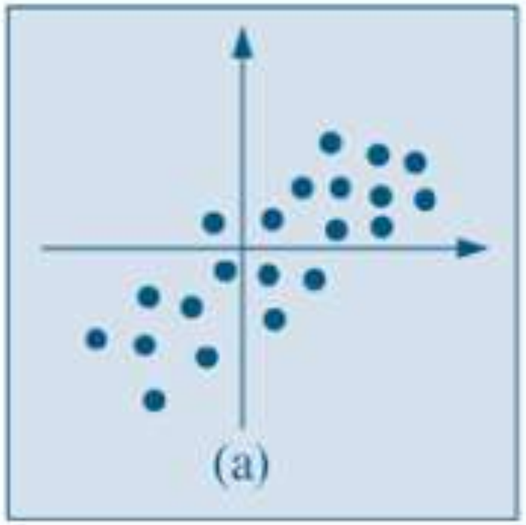


A representação gráfica das variáveis quantitativas ajuda muito a compreender o comportamento conjunto das duas variáveis quanto à existência ou não de associação entre elas.

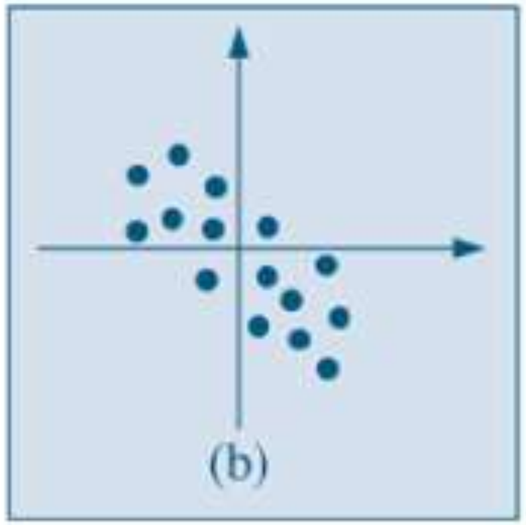
Contudo, é muito útil quantificar esta associação. Existem muitos tipos de associações possíveis, e aqui iremos apresentar o tipo de relação mais simples, que é a linear. Isto é, iremos definir uma medida que avalia o quanto a nuvem de pontos no gráfico de dispersão aproxima-se de uma reta.

Esta medida será definida de modo a variar num intervalo finito, especificamente, de −1 a +1.

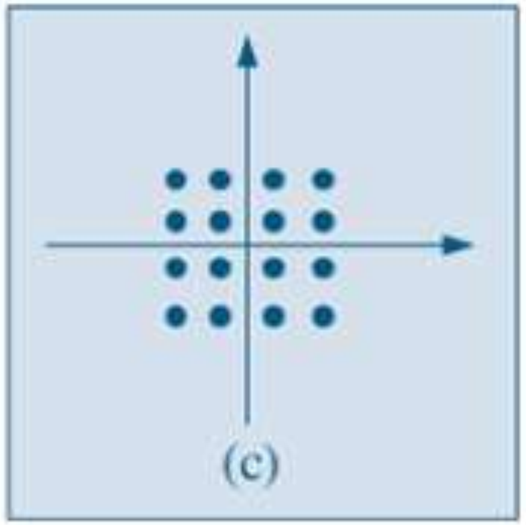
Associação Entre Variáveis Quantitativas



Associação Linear Positiva



Associação Linear Negativa



Sem Associação

Análise Estatística para Data Science I com R e SAS

Medidas de Associação Entre Variáveis Quantitativas

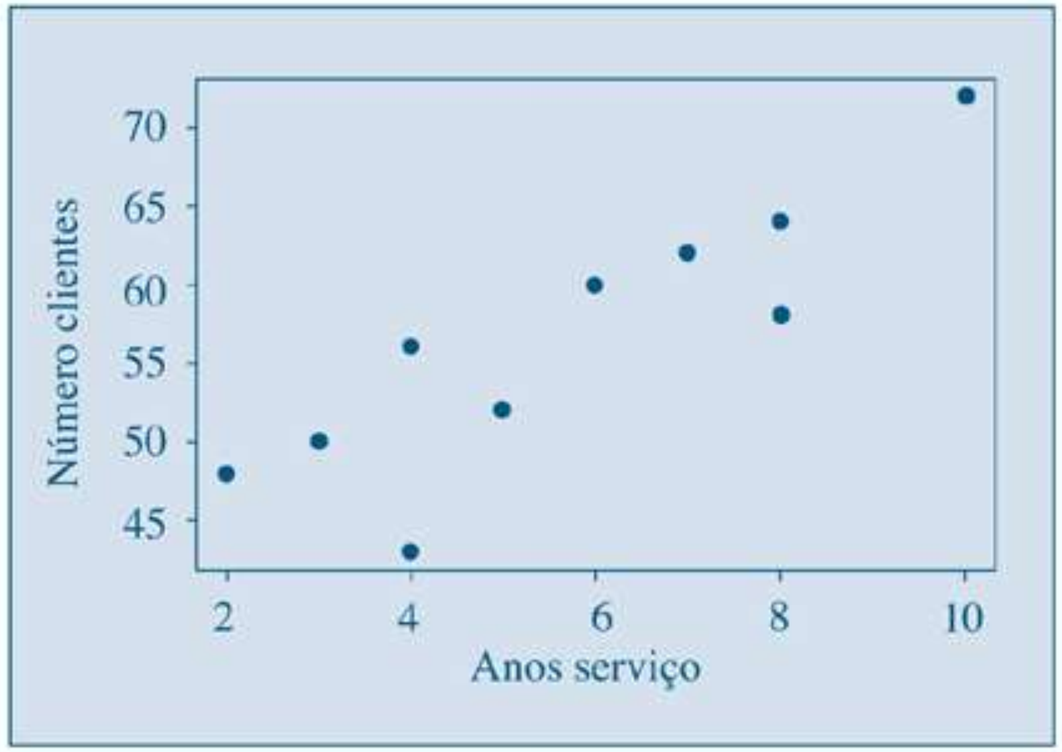
Data Science Academy

Medidas de Associação Entre Variáveis Quantitativas

O coeficiente de correlação (linear) entre duas variáveis é uma medida do grau de associação entre elas e também da proximidade dos dados a uma reta.

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos de serviço (X)	Número de 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



Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	-3,7	-8,5	-1,54	-1,05	1,617
B	3	50	-2,7	-6,5	-1,12	-0,80	0,846
C	4	56	-1,7	-0,5	-0,71	-0,06	0,043
D	5	52	-0,7	-4,5	-0,29	-0,55	0,160
E	4	43	-1,7	-13,5	-0,71	-1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7, \qquad dp(X) = 2{,}41, \qquad \bar{y} = 56{,}5, \qquad dp(Y) = 8{,}11$

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	-3,7	-8,5	-1,54	-1,05	1,617
B	3	50	-2,7	-6,5	-1,12	-0,80	0,846
C	4	56	-1,7	-0,5	-0,71	-0,06	0,043
D	5	52	-0,7	-4,5	-0,29	-0,55	0,160
E	4	43	-1,7	-13,5	-0,71	-1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5,7, \quad dp(X) = 2,41, \quad \bar{y} = 56,5, \quad dp(Y) = 8,11$

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$
A	2	48	-3,7	-8,5
B	3	50	-2,7	-6,5
C	4	56	-1,7	-0,5
D	5	52	-0,7	-4,5
E	4	43	-1,7	-13,5
F	6	60	0,3	3,5
G	7	62	1,3	5,5
H	8	58	2,3	1,5
I	8	64	2,3	7,5
J	10	72	4,3	15,5
Total	57	565	0	0

$\bar{x} = 5{,}7, \quad dp(X) = 2{,}41, \quad \bar{y} = 56{,}5, \quad dp(Y) = 8{,}11$

A soma dos produtos das coordenadas depende, e muito, do número de pontos. Considere o caso de associação positiva: a soma tende a aumentar com o número de pares (x, y) e ficaria difícil comparar essa medida para dois conjuntos com números diferentes de pontos. Por isso, costuma-se usar a média da soma dos produtos das coordenadas.

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	−3,7	−8,5	−1,54	−1,05	1,617
B	3	50	−2,7	−6,5	−1,12	−0,80	0,846
C	4	56	−1,7	−0,5	−0,71	−0,06	0,043
D	5	52	−0,7	−4,5	−0,29	−0,55	0,160
E	4	43	−1,7	−13,5	−0,71	−1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7, \quad dp(X) = 2{,}41, \quad \bar{y} = 56{,}5, \quad dp(Y) = 8{,}11$

Observando esses valores centrados, verificamos que ainda existe um problema quanto à escala usada.

A variável Y tem variabilidade muito maior do que X, e o produto ficaria muito mais afetado pelos resultados de Y do que pelos de X.

Para corrigirmos isso, podemos reduzir as duas variáveis a uma mesma escala, dividindo-se os desvios pelos respectivos desvios padrões.

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	−3,7	−8,5	−1,54	−1,05	1,617
B	3	50	−2,7	−6,5	−1,12	−0,80	0,846
C	4	56	−1,7	−0,5	−0,71	−0,06	0,043
D	5	52	−0,7	−4,5	−0,29	−0,55	0,160
E	4	43	−1,7	−13,5	−0,71	−1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7, \quad dp(X) = 2{,}41, \quad \bar{y} = 56{,}5, \quad dp(Y) = 8{,}11$

Esses novos valores estão nas colunas 6 e 7.

Finalmente, na coluna 8, indicamos os produtos das coordenadas reduzidas e sua soma, 8,769, que, como esperávamos, é positiva.

Para completar a definição dessa medida de associação, basta calcular a média dos produtos das coordenadas reduzidas, isto é, correlação (X, Y) = 8,769/10 = 0,877.

Medidas de Associação Entre Variáveis Quantitativas

Agente	Anos x	Clientes y	$x-\bar{x}$	$y-\bar{y}$	$\frac{x-\bar{x}}{dp(x)}=z_x$	$\frac{y-\bar{y}}{dp(y)}=z_y$	$z_x \cdot z_y$
A	2	48	−3,7	−8,5	−1,54	−1,05	1,617
B	3	50	−2,7	−6,5	−1,12	−0,80	0,846
C	4	56	−1,7	−0,5	−0,71	−0,06	0,043
D	5	52	−0,7	−4,5	−0,29	−0,55	0,160
E	4	43	−1,7	−13,5	−0,71	−1,66	1,179
F	6	60	0,3	3,5	0,12	0,43	0,052
G	7	62	1,3	5,5	0,54	0,68	0,367
H	8	58	2,3	1,5	0,95	0,19	0,181
I	8	64	2,3	7,5	0,95	0,92	0,874
J	10	72	4,3	15,5	1,78	1,91	3,400
Total	57	565	0	0			8,769

$\bar{x} = 5{,}7, \quad dp(X) = 2{,}41, \quad \bar{y} = 56{,}5, \quad dp(Y) = 8{,}11$

Esses novos valores estão nas colunas 6 e 7.

Finalmente, na coluna 8, indicamos os produtos das coordenadas reduzidas e sua soma, 8,769, que, como esperávamos, é positiva. Para completar a definição dessa medida de associação, basta calcular a média dos produtos das coordenadas reduzidas, isto é, correlação (X, Y) = 8,769/10 =

Portanto, para esse exemplo, o grau de associação linear está quantificado por 87,7%.

Medidas de Associação Entre Variáveis Quantitativas

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

$$\text{corr}(X,Y) = \frac{\text{cov}(X,Y)}{dp(X) \cdot dp(Y)}$$

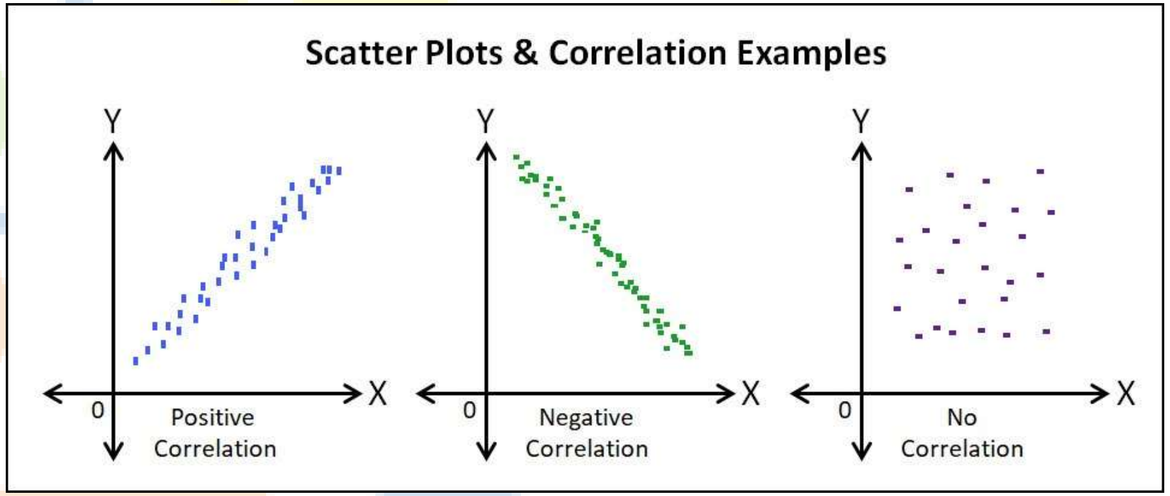
Análise Estatística para Data Science I com R e SAS

Analisando e Interpretando Scatter Plots

Data Science Academy

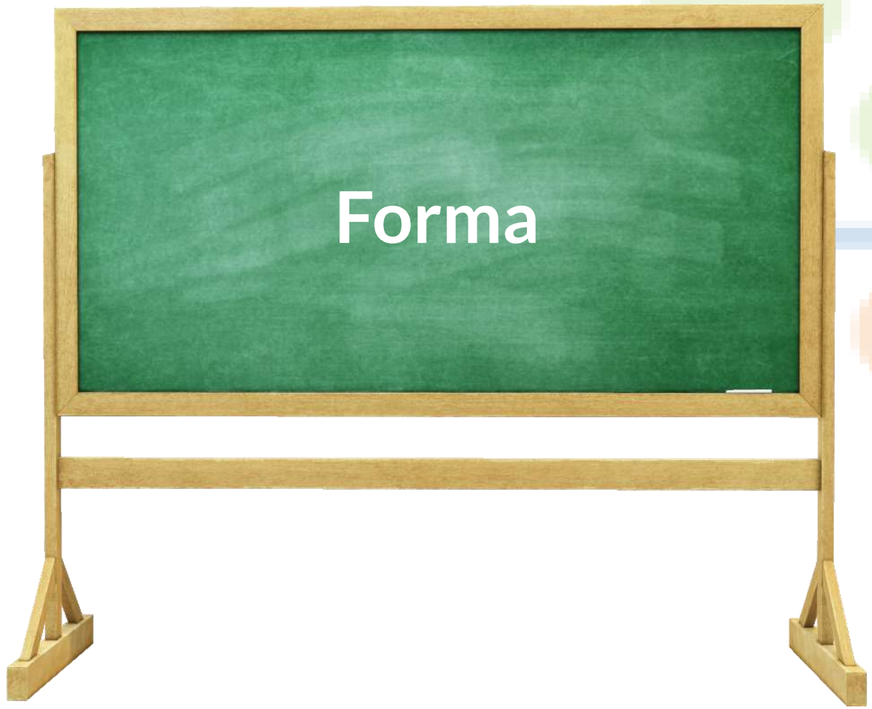
Analisando e Interpretando Scatterplots

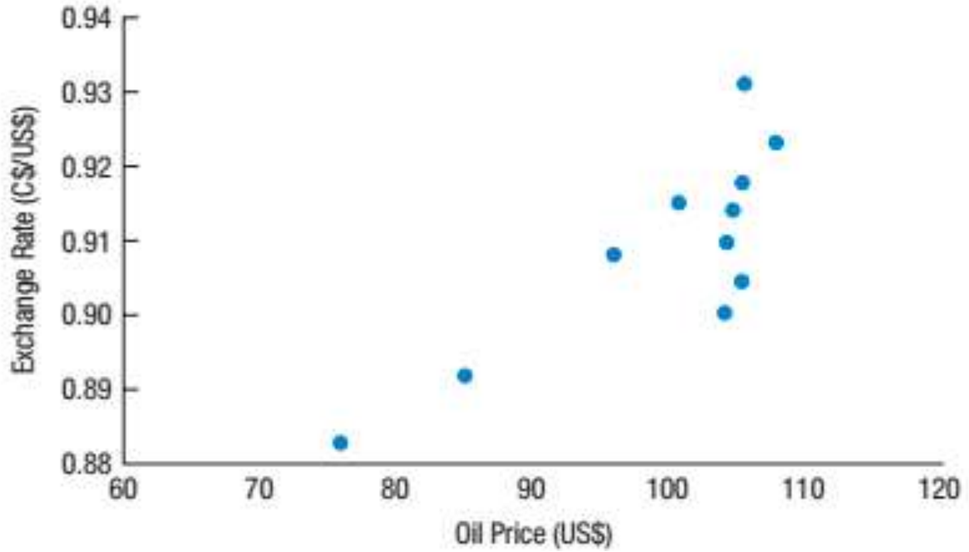
Ao analisar e interpretar Scatter Plots, devemos considerar 4 aspectos:



Analisando e Interpretando Scatterplots

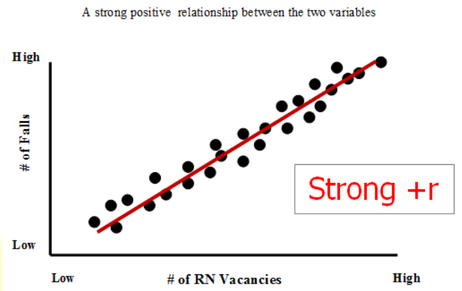
Ao analisar e interpretar Scatter Plots, devemos considerar 4 aspectos:

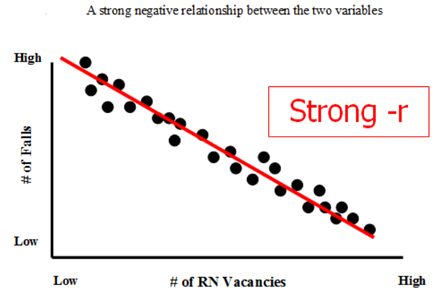


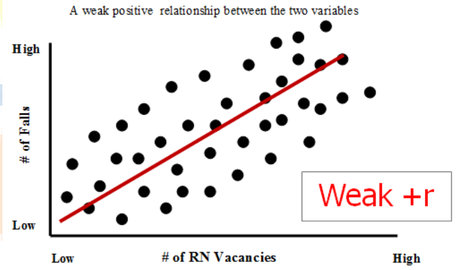


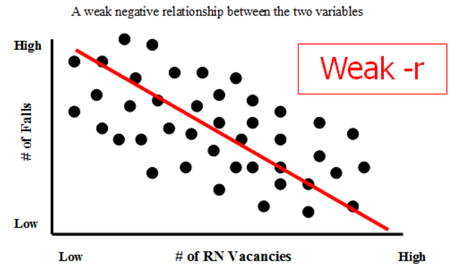
Analisando e Interpretando Scatterplots

Ao analisar e interpretar Scatter Plots, devemos considerar 4 aspectos:



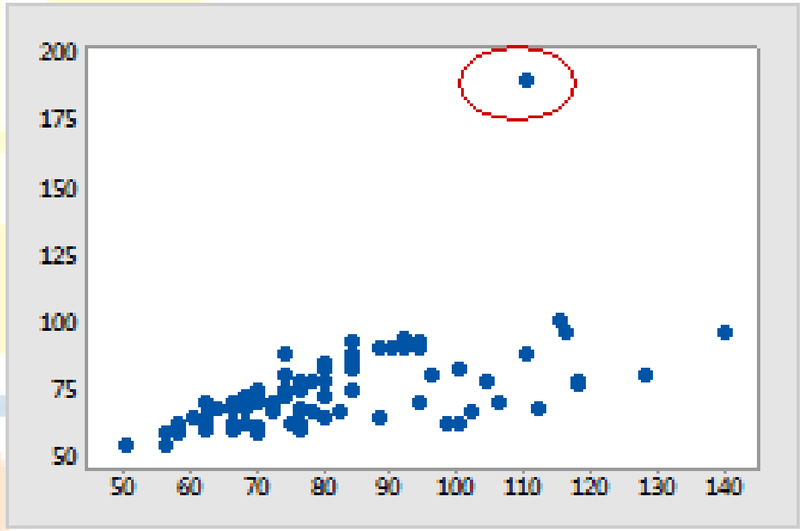






Analisando e Interpretando Scatterplots

Ao analisar e interpretar Scatter Plots, devemos considerar 4 aspectos:



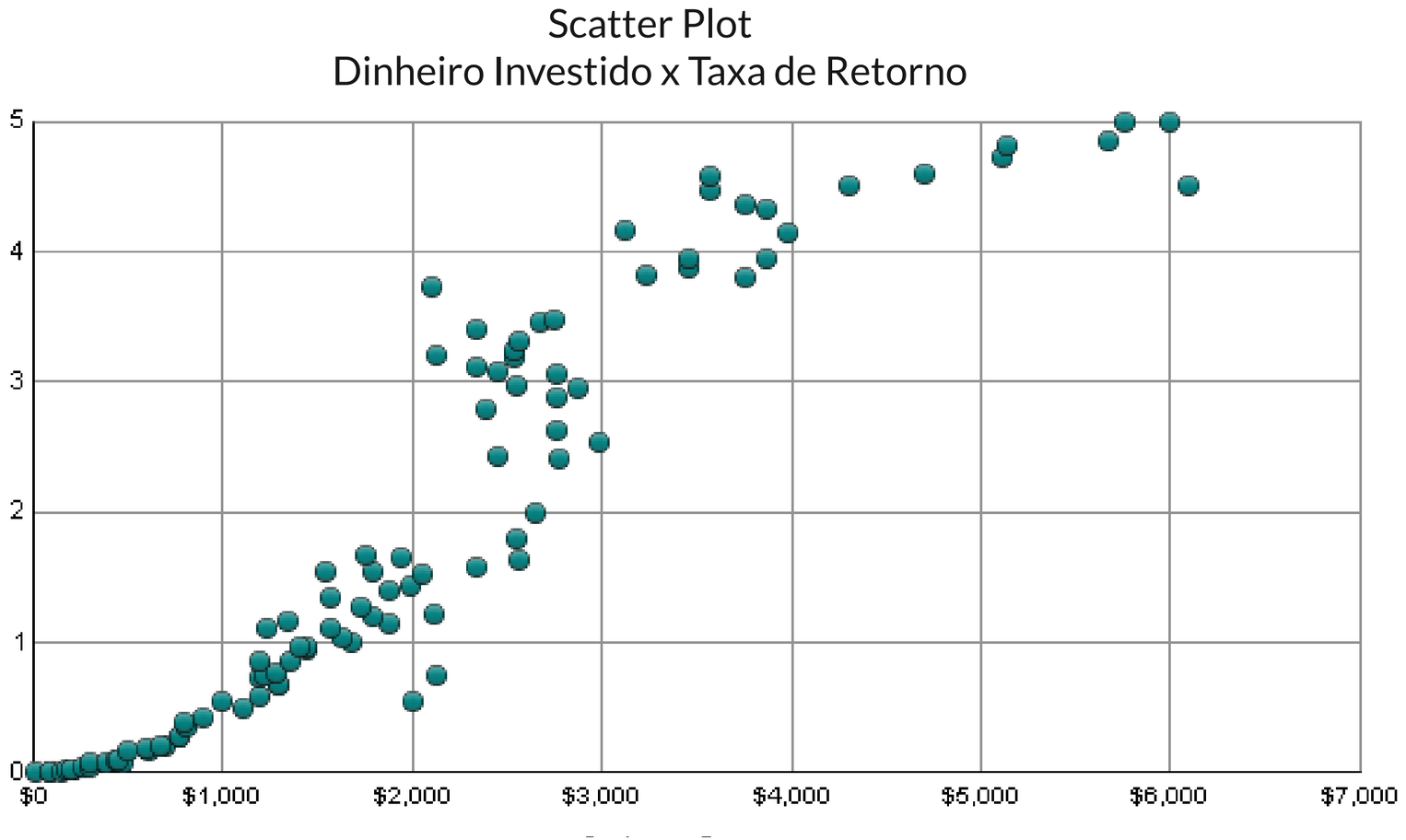
Análise Estatística para Data Science I com R e SAS

Atribuindo Funções as Variáveis no Scatter Plot

Data Science Academy

Atribuindo Funções as Variáveis no Scatter Plot

Variável y
Resposta, Dependente

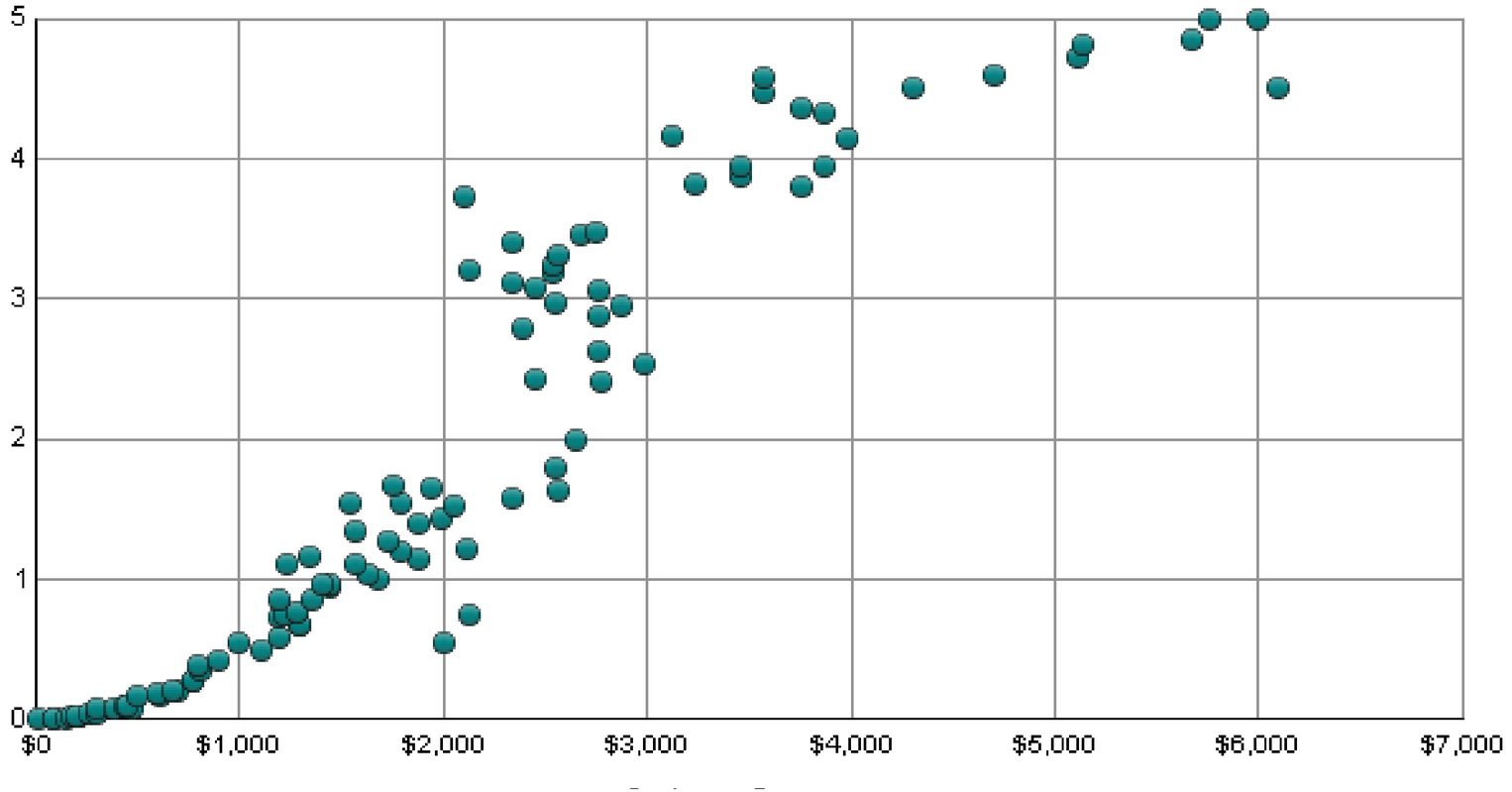


Variável x
Explanatória, Independente, Preditora

Atribuindo Funções as Variáveis no Scatter Plot

Variável y – Taxa de Retorno
Resposta, Dependente

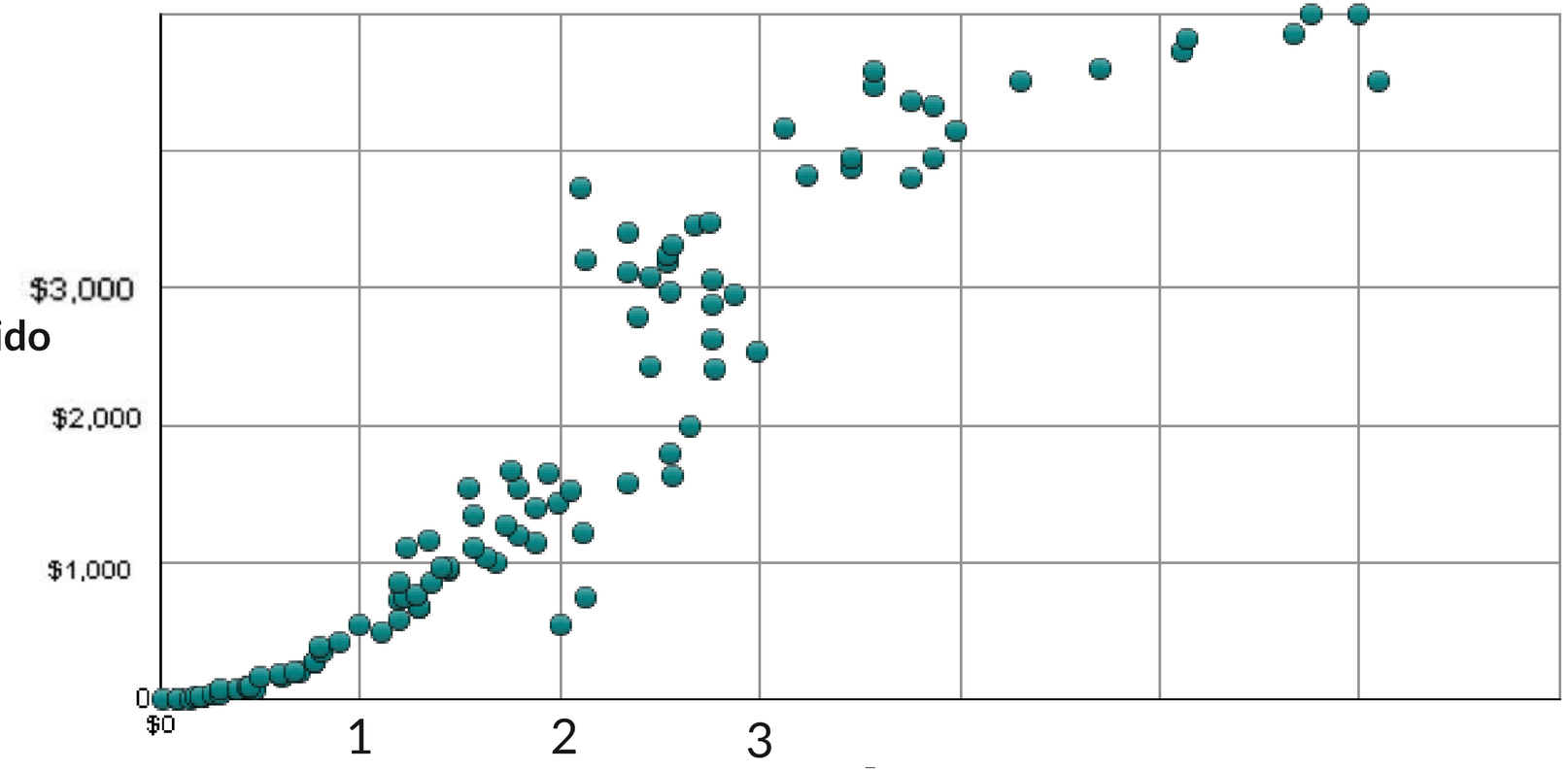
Scatter Plot
Dinheiro Investido x Taxa de Retorno



Variável x – Dinheiro Investido
Explanatória, Independente, Preditora

Atribuindo Funções as Variáveis no Scatter Plot

Scatter Plot
Dinheiro Investido x Taxa de Retorno



Variável y – Dinheiro Investido
Resposta, Dependente

Variável x – Taxa de Retorno
Explanatória, Independente, Preditora

Análise Estatística para Data Science I com R e SAS

Compreendendo o Que é Correlação

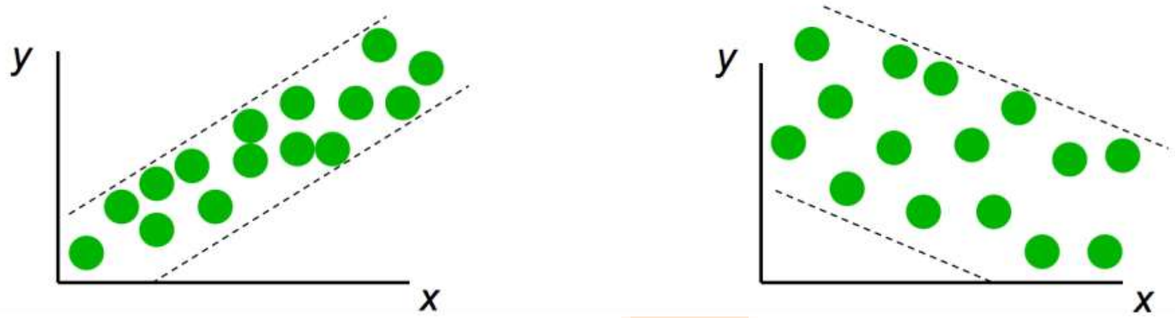
Data Science Academy

Compreendendo o Que é Correlação

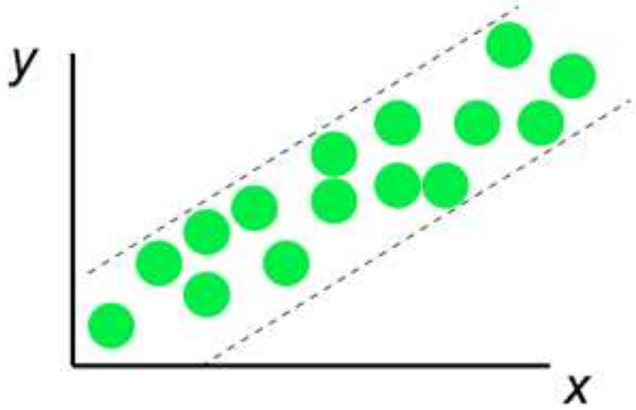
A correlação nos permite medir a **força** e **direção** de um relacionamento linear entre **duas** variáveis.

Compreendendo o Que é Correlação

O relacionamento entre duas variáveis **é linear**, se o gráfico de dispersão entre elas tem o **padrão de uma linha reta.** Exemplos de relação linear:

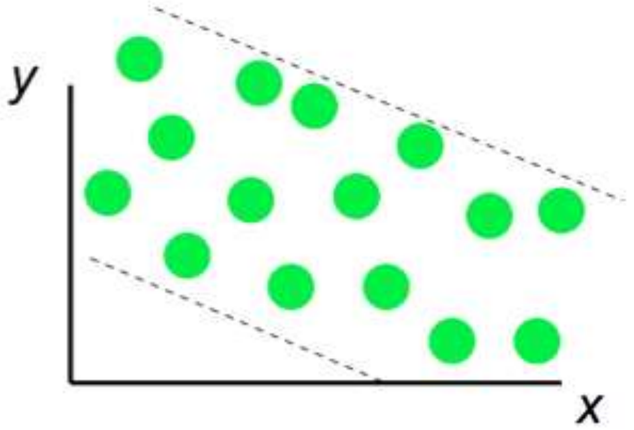


Compreendendo o Que é Correlação



Relacionamento **positivo**, inclinação se move para cima.

Relacionamento **negativo**, inclinação se move para baixo.



Compreendendo o Que é Correlação

O coeficiente de correlação **(r)** indica a força e direção de uma relação linear entre a variável independente e dependente.

Compreendendo o Que é Correlação

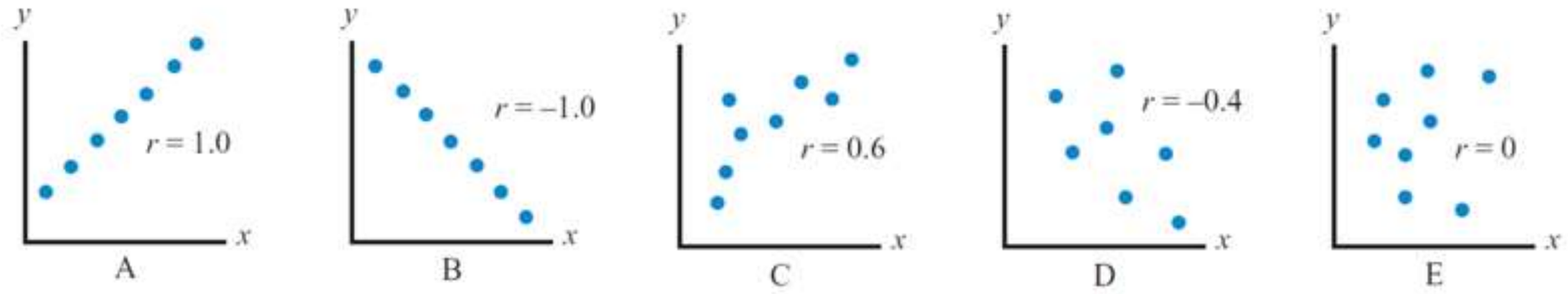


Gráfico A (*r* = 1.0): correlação positiva perfeita entre *x* e *y*

Gráfico B (*r* = -1.0): correlação negativa perfeita entre *x* e *y*

Gráfico C (*r* = 0.6): relação positiva moderada: y tende a aumentar se x aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

Gráfico D (*r* = -0.4): relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo: y tende a diminuir se x aumenta

Gráfico E (*r* = 0): Sem relação entre x e y

Os valores de **r** variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.

Compreendendo o Que é Correlação

A correlação, isto é, a ligação entre dois eventos, não implica necessariamente uma relação de causalidade, ou seja, que um dos eventos tenha causado a ocorrência do outro.

Compreendendo o Que é Correlação

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B).

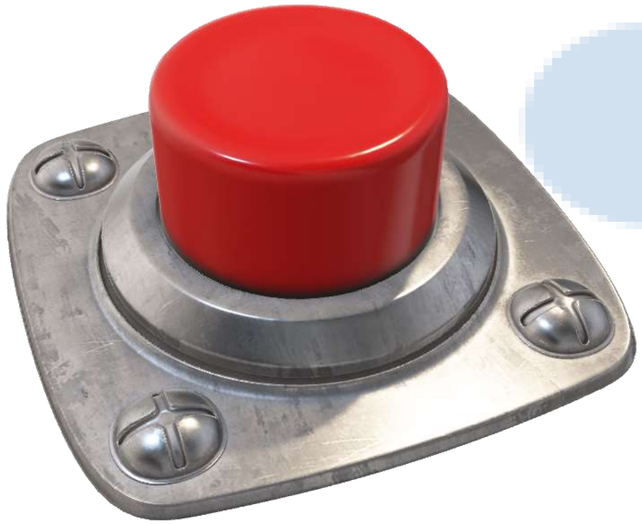
Análise Estatística para Data Science I com R e SAS

Condições Para Análise da Correlação

Data Science Academy

Condições Para Análise da Correlação

A correlação mede a força da associação linear entre duas variáveis quantitativas. Antes de usar a correlação, você deve verificar três condições:

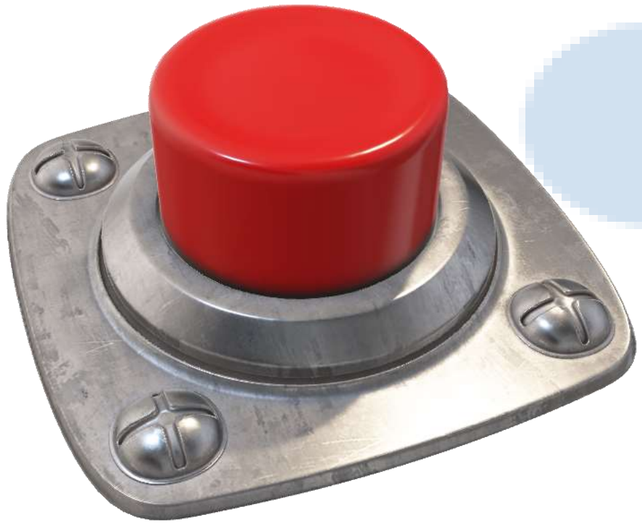


1. Variáveis Quantitativas

Condição: A correlação se aplica somente a variáveis quantitativas. Não aplique correlação a dados categóricos disfarçados de quantitativos. Verifique se você conhece as unidades das variáveis e o que elas medem.

Condições Para Análise da Correlação

A correlação mede a força da associação linear entre duas variáveis quantitativas. Antes de usar a correlação, você deve verificar três condições:

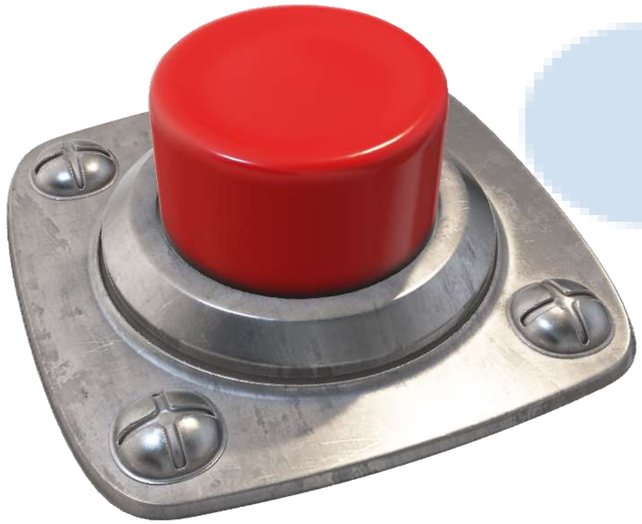


2. Linearidade

Condição: Você pode calcular um coeficiente de correlação para qualquer par de variáveis. Mas a correlação mede a força apenas da associação linear e será enganosa se o relacionamento não for direto o suficiente.

Condições Para Análise da Correlação

A correlação mede a força da associação linear entre duas variáveis quantitativas. Antes de usar a correlação, você deve verificar três condições:



3. Outliers

Condição: Observações incomuns podem distorcer a correlação e fazer com que uma correlação de outra forma pequena pareça grande ou, por outro lado, ocultar uma grande correlação. Quando você vê um ou mais valores discrepantes, geralmente é uma boa ideia relatar a correlação com e sem esses pontos.

É um prazer ter você aqui!

Muito Obrigado!

Pela Confiança em Nosso Trabalho.

Continue Trilhando Uma Excelente Jornada de Aprendizagem!

Data Science Academy