Estatística Descritiva (III)

Associação entre Variáveis

Associação entre variáveis qualitativas

Tabelas de Contingência

Exemplo: Suponha que queiramos analisar o comportamento conjunto das variáveis Y: grau de instrução e V: região de procedência dos funcionários de uma empresa. A distribuição de frequências é representada por uma tabela de dupla entrada e está na Tabela 4.2.

Tabela 4.2 Distribuição conjunta das frequências das 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
Outra	5	6	2	13
Total	12	18	6	36

Fonte: Tabela 2.1.

Em vez de trabalharmos com as frequências absolutas, podemos construir tabelas com **as frequências relativas (proporções)**, como foi feito no caso unidimensional. Existem três possibilidades de expressarmos a proporção de cada casela:

(a) em relação ao total geral;

(b) em relação ao total de cada linha;

(c) ou em relação ao total de cada coluna.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

Tabela 4.3 Distribuição conjunta das proporções (em porcentagem) em relação ao total geral das variáveis Y e V definidas no texto.

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%

Fonte: Tabela 4.2.

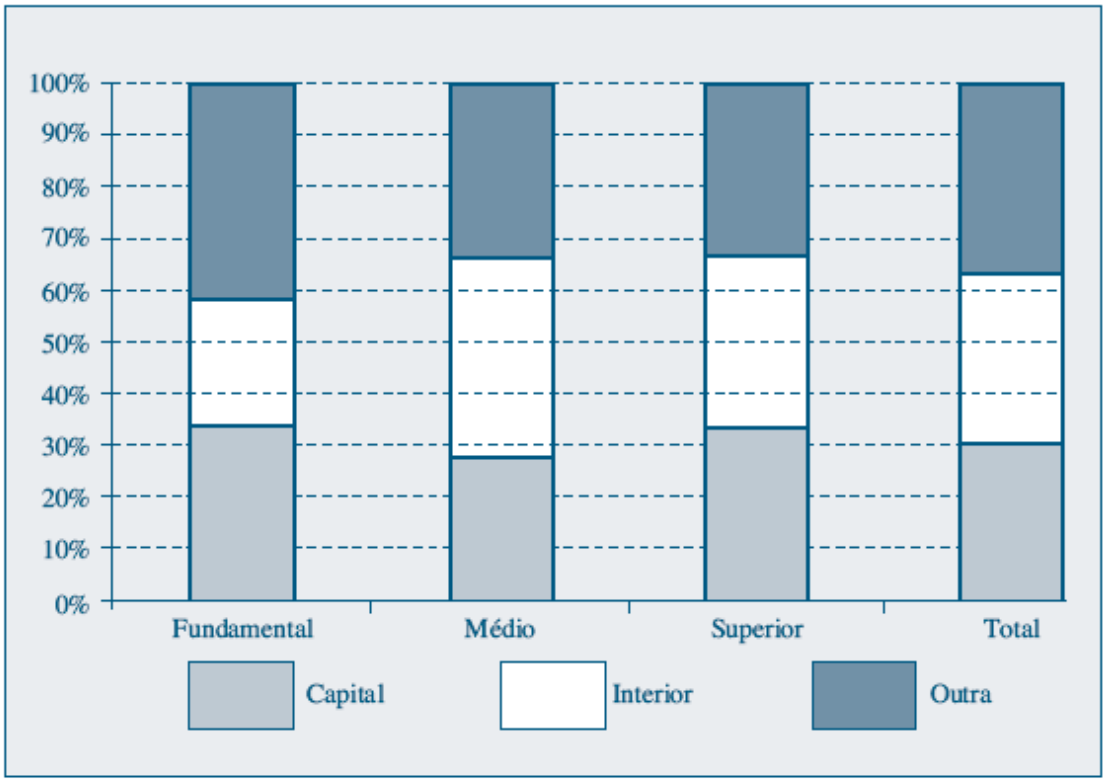
Tabela 4.4 Distribuição conjunta das proporções (em porcentagem) em relação aos totais de cada coluna das variáveis Y e V definidas no texto.

V \ Y	Fundamental	Médio	Superior	Total
Capital	33%	28%	33%	31%
Interior	25%	39%	33%	33%
Outra	42%	33%	34%	36%
Total	100%	100%	100%	100%

Fonte: Tabela 4.2.

De acordo com o objetivo do problema em estudo, uma delas será a mais conveniente.

Figura 4.1 Distribuição da região de procedência por grau de instrução.



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.

Exemplo 4.2 Queremos verificar se existe ou não associação entre o sexo e a carreira escolhida por 200 alunos de Economia e Administração. Esses dados estão na Tabela 4.5.

Tabela 4.5 Distribuição conjunta de alunos segundo o sexo (X) e o curso escolhido (Y).

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

Fonte: Dados hipotéticos.

Tabela 4.6 Distribuição conjunta das proporções (em porcentagem) de alunos segundo o sexo (X) e o curso escolhido (Y).

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

Fonte: Tabela 4.5.

A partir dessa tabela podemos observar que, **independentemente** do sexo, 60% das pessoas preferem Economia e 40% preferem Administração (observe na coluna de total).

Não havendo dependência entre as variáveis, esperaríamos essas mesmas proporções para cada sexo.

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.

Tabela 4.7 Distribuição conjunta das frequências e proporções (em porcentagem), segundo o sexo (X) e o curso escolhido (Y).

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%)

Fonte: Dados hipotéticos.

Nesse caso, as variáveis sexo e curso escolhido parecem ser associadas.

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.

Duas medidas:

- coeficiente de contingência, devido a K. Pearson e uma modificação desse.

Exemplo 4.3 Queremos verificar se a criação de determinado tipo de cooperativa está associada com algum fator regional. Coletados os dados relevantes, obtemos a Tabela 4.8.

Tabela 4.8 Cooperativas autorizadas a funcionar por tipo e estado, junho de 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 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%)

Fonte: Sinopse Estatística da Brasil — IBGE, 1977.

Tabela 4.9 Valores esperados na Tabela 4.8 assumindo a independência entre as duas variáveis.

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%)

Fonte: Tabela 4.8.

Tabela 4.10 Desvios entre observados e 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)

Fonte: Tabelas 4.8 e 4.9.

construir, para cada casela, a medida

$$\frac{(o_i - e_i)^2}{e_i} \qquad (4.1)$$

no qual o_i é o valor observado e e_i é o valor esperado.

Uma medida do afastamento global pode ser dada pela soma de todas as medidas (4.1). Essa medida é denominada χ^2 (qui-quadrado) de Pearson, e no nosso exemplo teríamos

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

Um valor grande de χ^2 indica associação entre as variáveis, o que parece ser o caso.

Tabela 4.11 Notação para tabelas de contingência.

X \ Y	B_1	B_2	...	B_j	...	B_s	Total
A_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1s}	$n_{1.}$
A_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2s}	$n_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{is}	$n_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.s}$	$n..$

Suponha que temos duas variáveis qualitativas X e Y, classificadas em r categorias $A_1, A_2, ..., A_r$ para X e s categorias $B_1, B_2, ..., B_s$, para Y.

Na tabela, temos:

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

$n_{i.} = \sum_{j=1}^{s} n_{ij}$ = número de elementos da i-ésima 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 de Pearson pode ser escrito

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_{ij}^*)^2}{n_{ij}^*}, \qquad (4.4)$$

$$n_{ij}^* = \frac{n_{i.} n_{.j}}{n}, \; i = 1, \; ..., r, \; j = 1, \; ..., s. \qquad (4.3)$$

Se a hipótese de não associação for verdadeira, o valor calculado de (4.4) deve estar próximo de zero. Se as variáveis forem associadas, o valor de χ^2 deve ser grande.

Pearson definiu uma medida de associação, baseada em (4.4), chamada *coeficiente de contingência*, dado por

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \qquad (4.5)$$

Contudo, o coeficiente acima não varia entre 0 e 1. O valor máximo de C depende de r e s. Para evitar esse inconveniente, costuma-se definir um outro coeficiente, dado por

$$T = \sqrt{\frac{\chi^2/n}{(r-1)(s-1)}}, \qquad (4.6)$$

que atinge o máximo igual a 1 se $r = s$.

No exemplo, C = 0,32 e T = 0,14

Associação entre variáveis quantitativas

Correlação e Regressão

Objetivo

Estudar a relação entre duas variáveis quantitativas.

Exemplos:

Idade e altura das crianças

Tempo de prática de esportes e ritmo cardíaco

Tempo de estudo e nota na prova

Taxa de desemprego e taxa de criminalidade

Expectativa de vida e taxa de analfabetismo

Investigaremos a presença ou ausência de relação linear sob dois pontos de vista:

a) Quantificando a força dessa relação: correlação.

b) Explicitando a forma dessa relação: regressão.

Representação gráfica de duas variáveis quantitativas:

Diagrama de dispersão

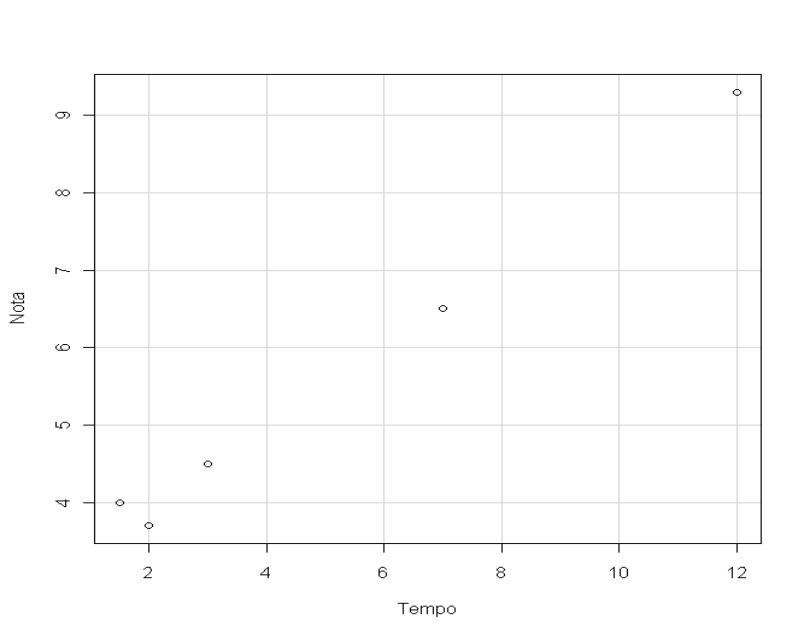
Exemplo 1: nota da prova e tempo de estudo

X : tempo de estudo (em horas)

Y : nota da prova

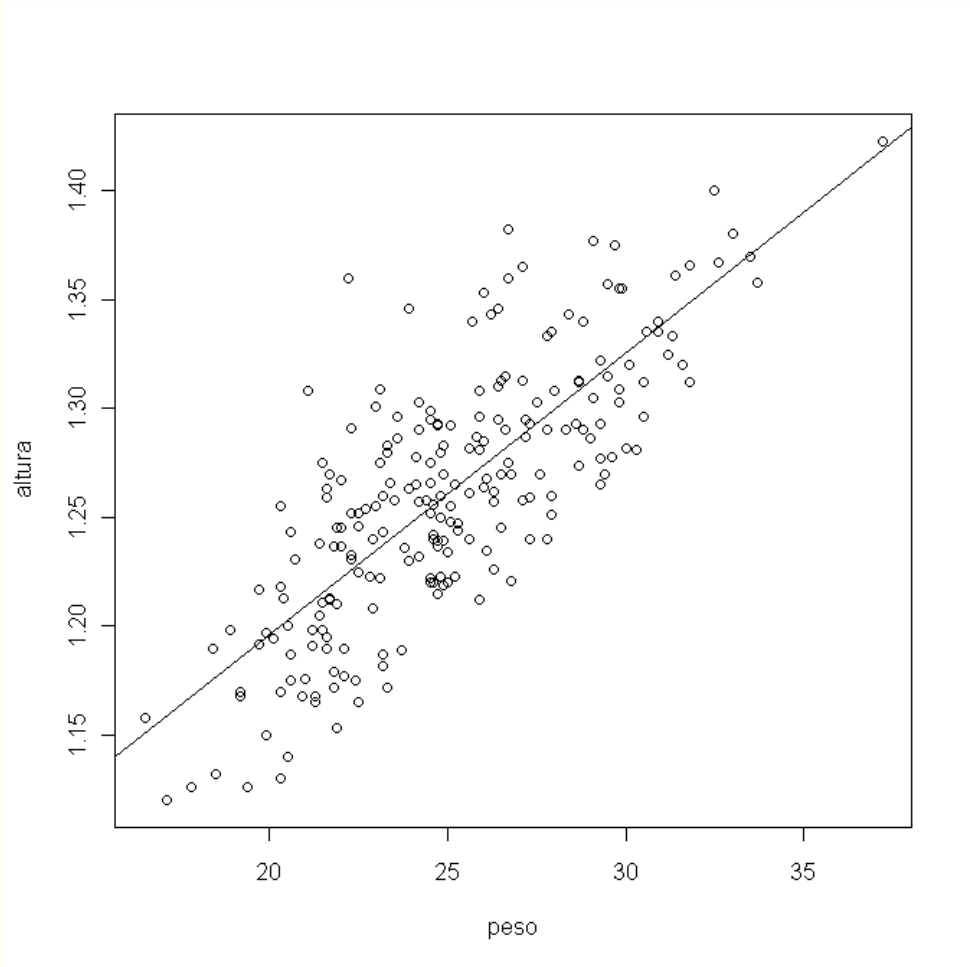
Pares de observações (X_i , Y_i) para cada estudante

Tempo(X)	Nota(Y)
3,0	4,5
7,0	6,5
2,0	3,7
1,5	4,0
12,0	9,3



Coeficiente de correlação linear

É uma medida que avalia o quanto a “nuvem de pontos” no diagrama de dispersão aproxima-se de uma reta.



O **coeficiente de correlação linear de *Pearson*** é calculado por:

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

sendo que

$\overline{X}$ e $\overline{Y}$ são as médias amostrais de X e Y, respectivamente,

S_X e S_Y são os desvios padrão de X e Y, respectivamente.

Fórmula alternativa para o coeficiente de correlação:

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

sendo

$$S_X^2 = \frac{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}}{n-1}.$$

Voltando ao Exemplo 1:

Tempo (X)	Nota (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
3,0	4,5	-2,1	-1,1	2,31
7,0	6,5	1,9	0,9	1,71
2,0	3,7	-3,1	-1,9	5,89
1,5	4,0	-3,6	-1,6	5,76
12,0	9,3	6,9	3,7	25,53
25,5	28,0	0	0	41,2

$\bar{X} = 5,1$ $\bar{Y} = 5,6$

$$S_x^2 = \frac{(-2,1)^2 + \ldots + (6,9)^2}{4} = \frac{78,2}{4} = 19,55 \Rightarrow S_x = 4,42$$

$$S_y^2 = \frac{(-1,1)^2 + \ldots + (3,7)^2}{4} = \frac{21,9}{4} = 5,47 \Rightarrow S_y = 2,34$$

Então,

$$r = \frac{41,2}{4 \,.\, 4,42 \,.\, 2,34} = 0,9959$$

Propriedade: $-1 \leq r \leq 1$

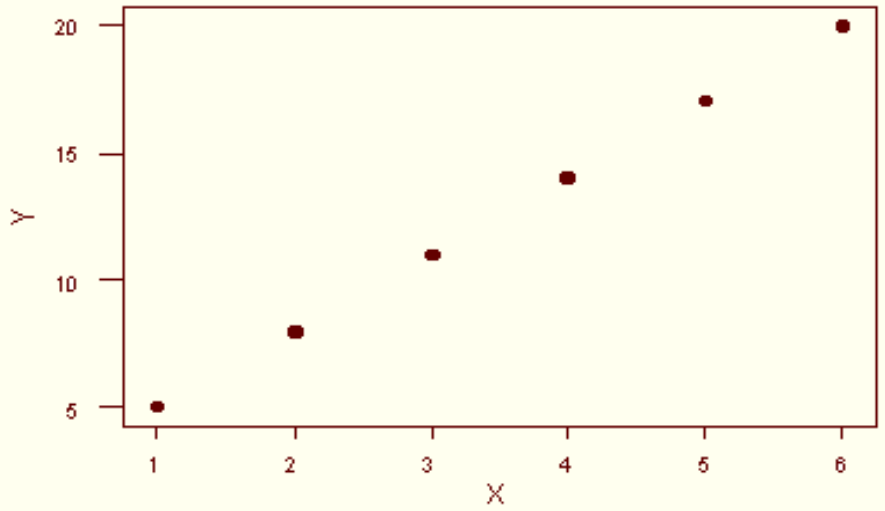
Casos particulares:

$r = 1 \Rightarrow$ correlação linear positiva e perfeita

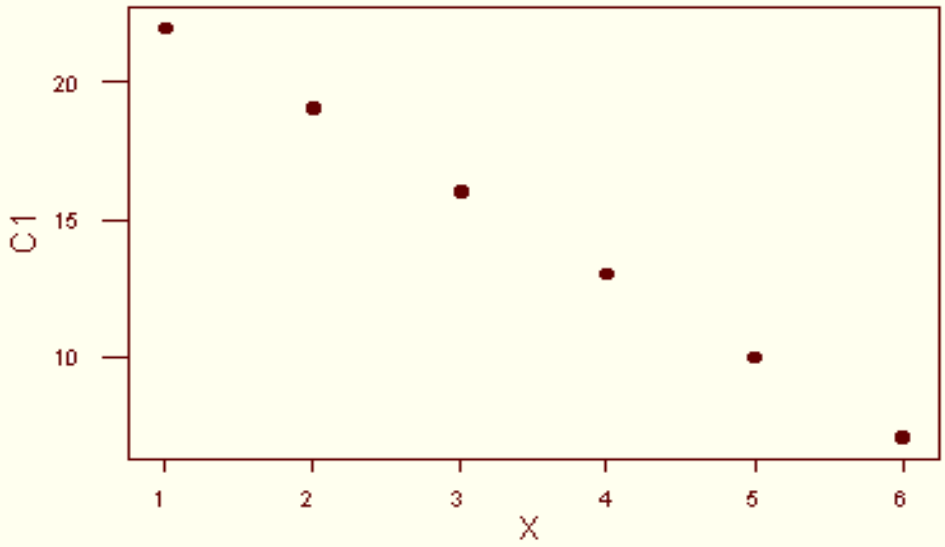
$r = -1 \Rightarrow$ correlação linear negativa e perfeita

$r = 0 \Rightarrow$ inexistência de correlação linear

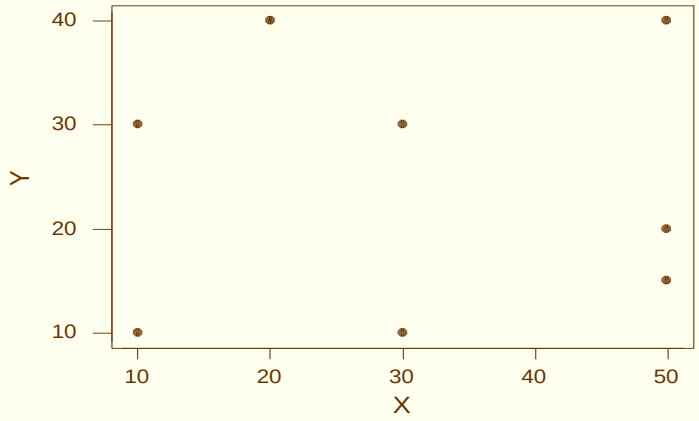
r = 1, correlação linear positiva e perfeita

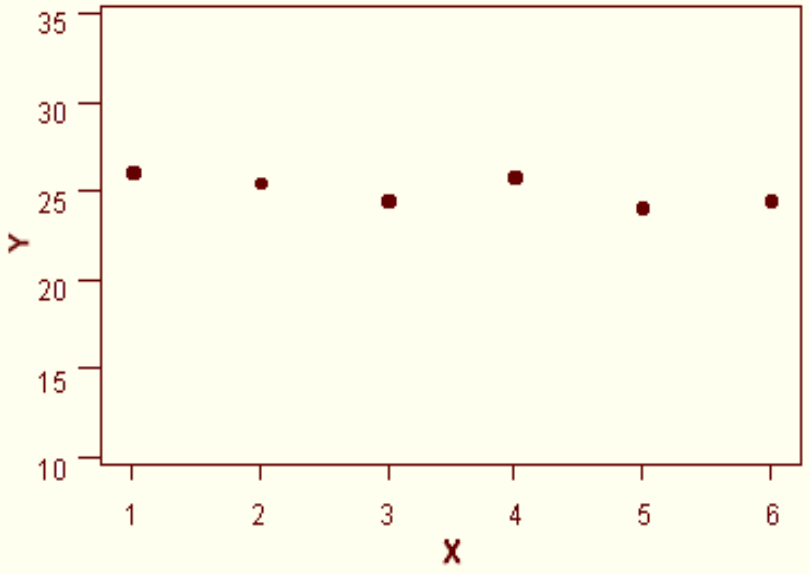


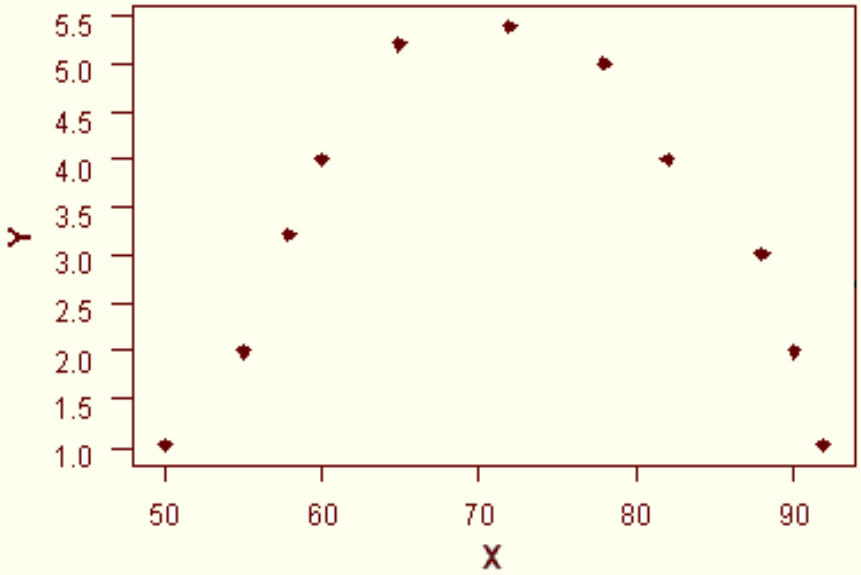
r = -1, correlação linear negativa e perfeita

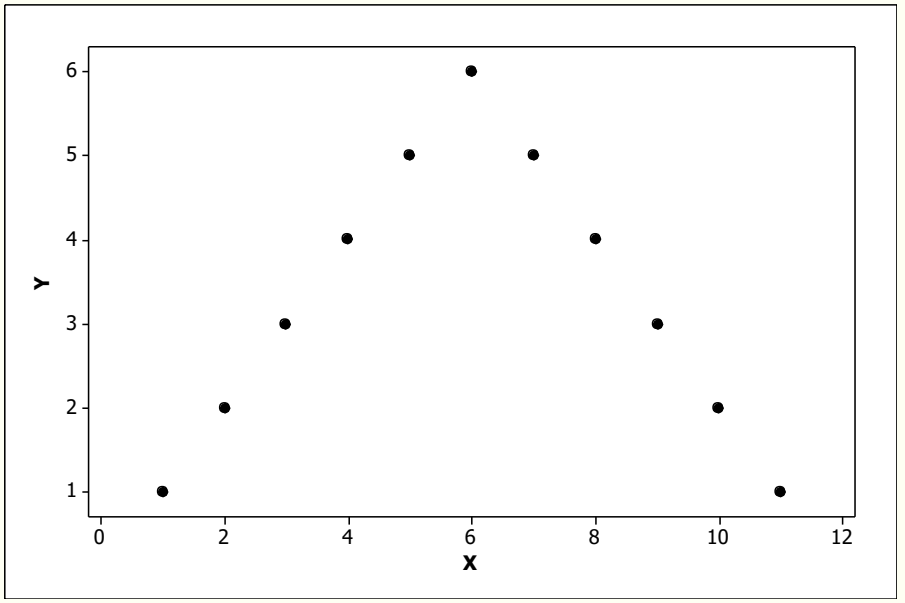


$r \cong 0$

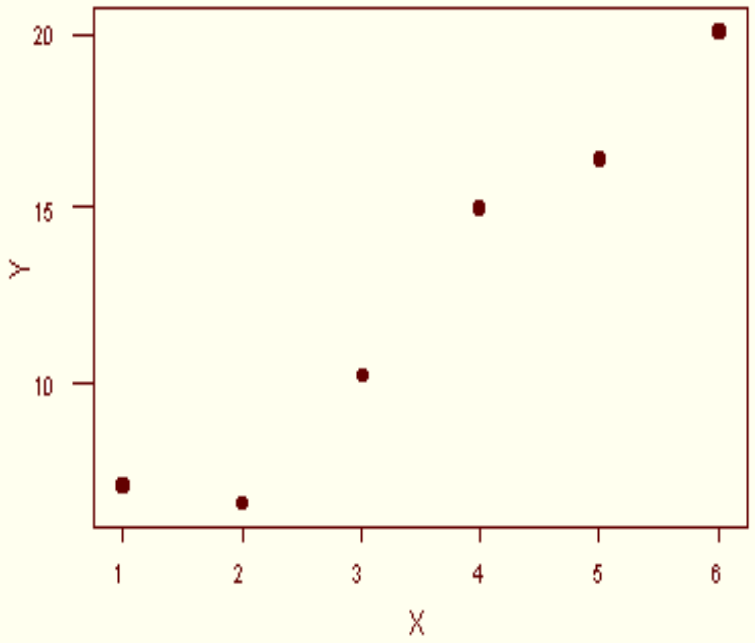

40
30
20
10
Y
10
20
30
40
50
X


35
30
25
20
15
10
Y
1
2
3
4
5
6
X

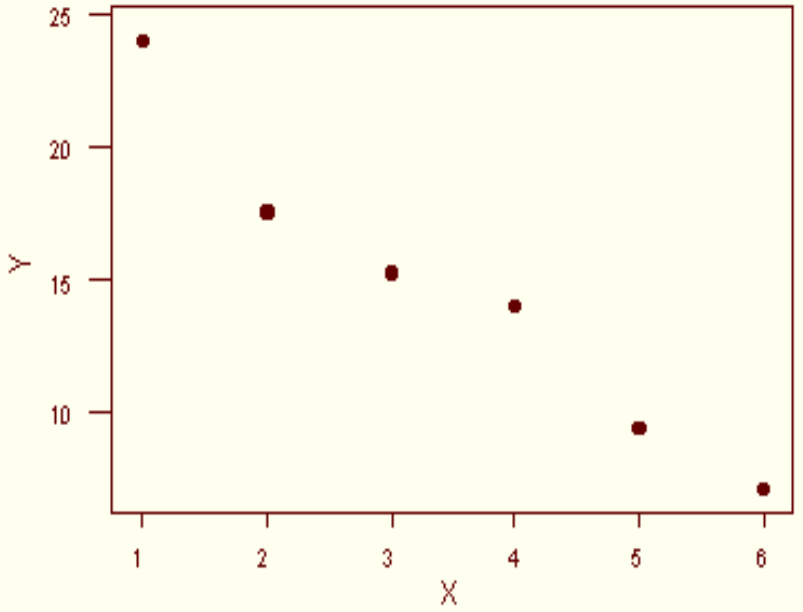

5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Y
50
60
70
80
90
X


6
5
4
3
2
1
Y
0
2
4
6
8
10
12
X

$r \cong 1$



$r \cong -1$



Exemplo 3: criminalidade e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: taxa de criminalidade

X: taxa de analfabetismo

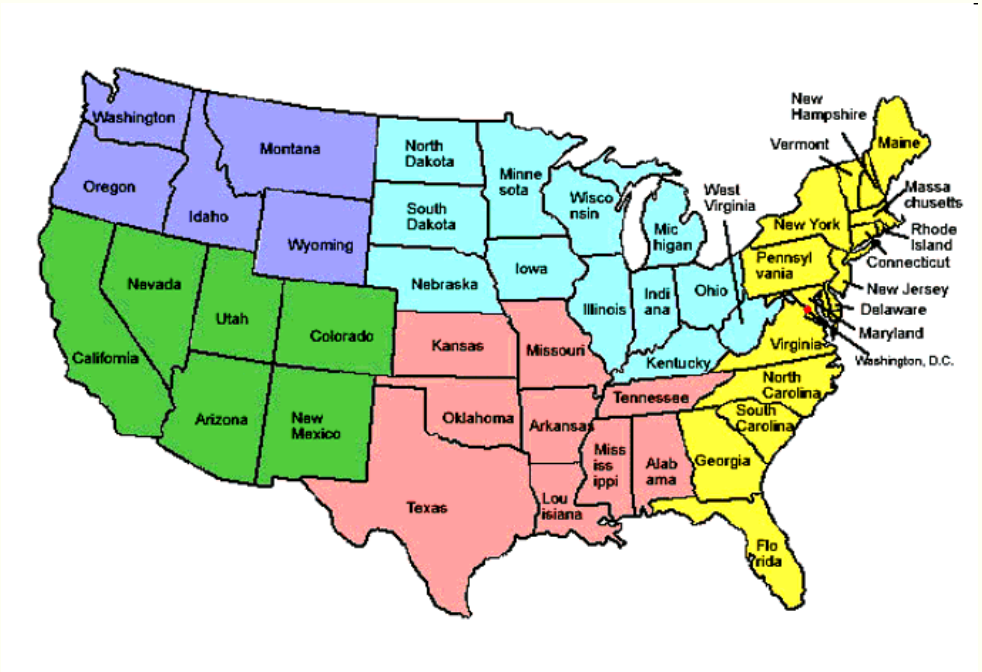
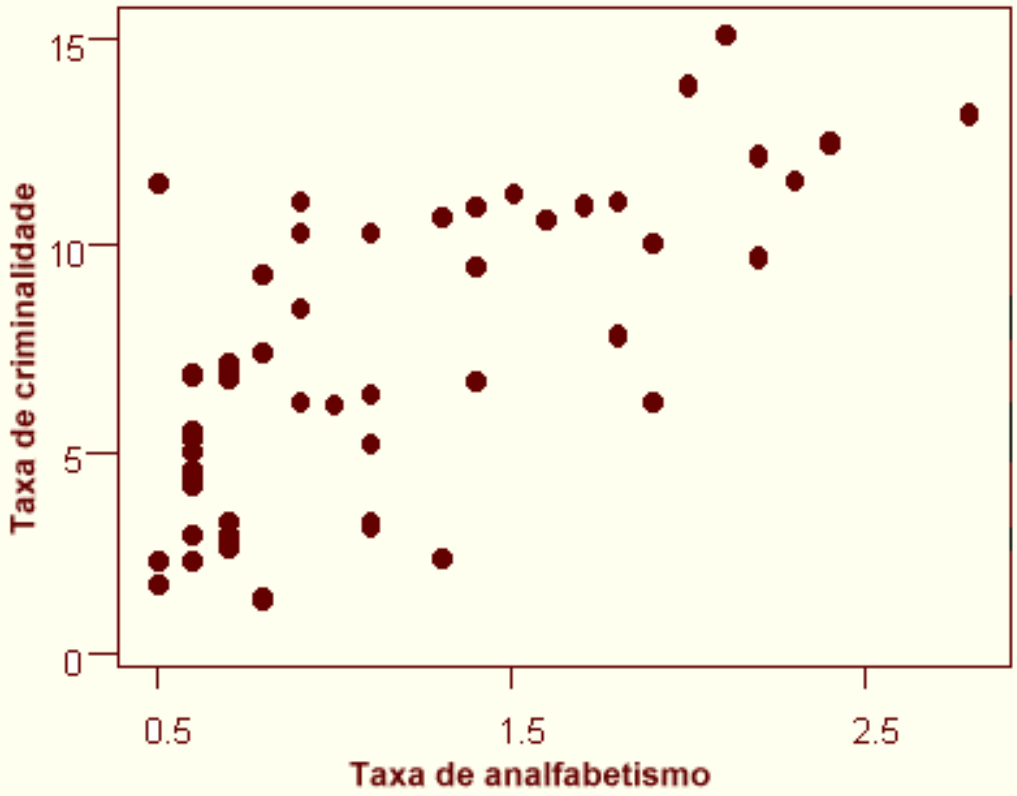


Diagrama de dispersão



Pode-se notar que, conforme aumenta a taxa de analfabetismo (X), a taxa de criminalidade (Y) tende a aumentar. Nota-se também uma tendência linear.

Cálculo da correlação

$\overline{Y} = 7{,}38$ (média de Y) e $S_Y = 3{,}692$ (desvio padrão de Y)

$\overline{X} = 1{,}17$ (média de X) e $S_X = 0{,}609$ (desvio padrão de X)

$\Sigma X_i Y_i = 509{,}12$

Correlação entre X e Y:

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

$$r = \frac{509{,}12 - 50 \times 7{,}38 \times 1{,}17}{49 \times 3{,}692 \times 0{,}609} = \frac{77{,}39}{110{,}17} = 0{,}702$$

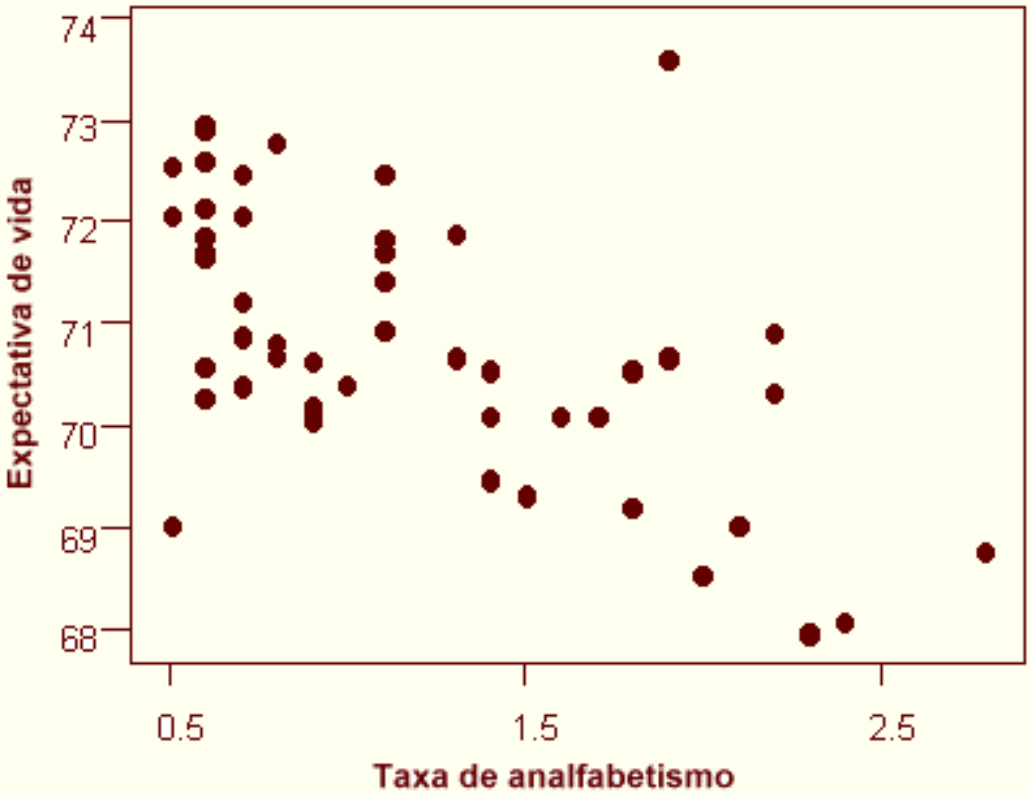
Exemplo 4: expectativa de vida e analfabetismo

Considere as duas variáveis observadas em 50 estados norte-americanos.

Y: expectativa de vida

X: taxa de analfabetismo

Diagrama de dispersão



Pode-se notar que, conforme aumenta a taxa de analfabetismo (X), a expectativa de vida (Y) tende a diminuir. Nota-se também uma tendência linear.

Cálculo da correlação

$\overline{Y} = 70{,}88$ (média de Y) e $S_Y = 1{,}342$ (desvio padrão de Y)

$\overline{X} = 1{,}17$ (média de X) e $S_X = 0{,}609$ (desvio padrão de X)

$\Sigma X_i Y_i = 4122{,}8$

Correlação entre X e Y:

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

$$r = \frac{4122{,}8 - 50 \times 70{,}88 \times 1{,}17}{49 \times 1{,}342 \times 0{,}609} = 40{,}047 = -0{,}59$$

Comentário:

- Na interpretação do coeficiente de correlação é importante visualizar o diagrama de dispersão.

Considere o seguinte exemplo: 6 variáveis são medidas em 11 indivíduos.

	X	$Y1$	$Y2$	$Y3$	$X4$	$Y4$
1	10	8,04	9,14	7,46	8	6,58
2	8	6,95	8,14	6,77	8	5,76
3	13	7,58	8,74	12,74	8	7,71
4	9	8,81	8,77	7,11	8	8,84
5	11	8,33	9,26	7,81	8	8,47
6	14	9,96	8,10	8,84	8	7,04
7	6	7,24	6,13	6,08	8	5,25
8	4	4,26	3,10	5,39	19	12,50
9	12	10,84	9,13	8,15	8	5,56
10	7	4,82	7,26	6,42	8	7,91
11	5	5,68	4,74	5,73	8	6,89

correlação linear de X e $Y1 = 0{,}816$

correlação linear de X e $Y2 = 0{,}816$

correlação linear de X e $Y3 = 0{,}816$

correlação linear de $X4$ e $Y4 = 0{,}817$

$\Rightarrow$ Mesmos valores de correlação.

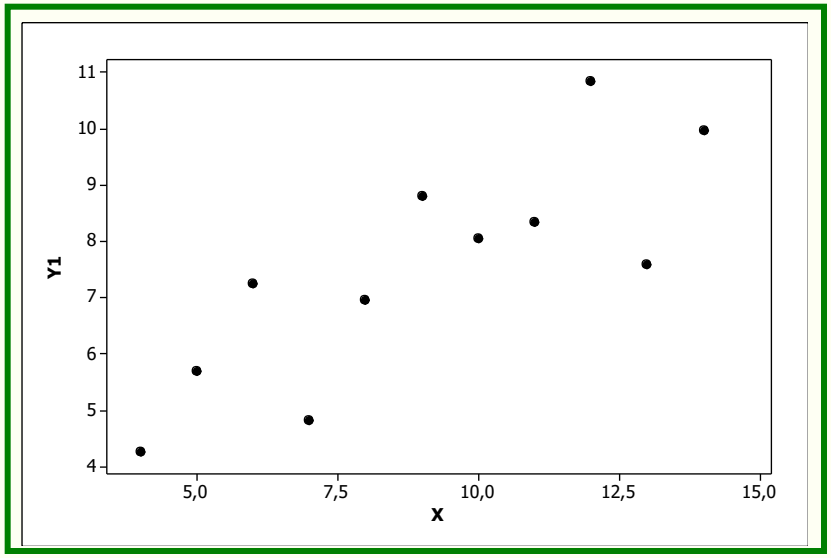
$\Rightarrow$ Qual é a forma esperada da dispersão conjunta destas variáveis?

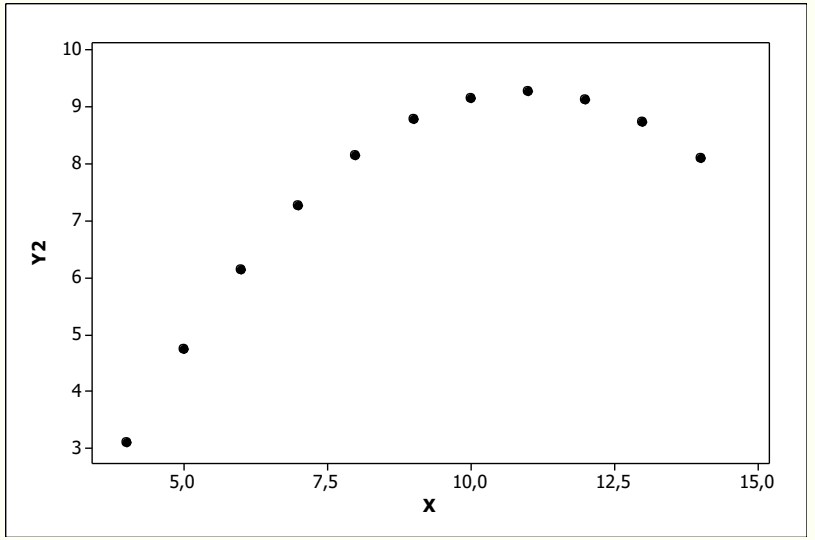
ARQUIVO *FA.xls*

Diagramas de dispersão e Coeficientes de Correlação

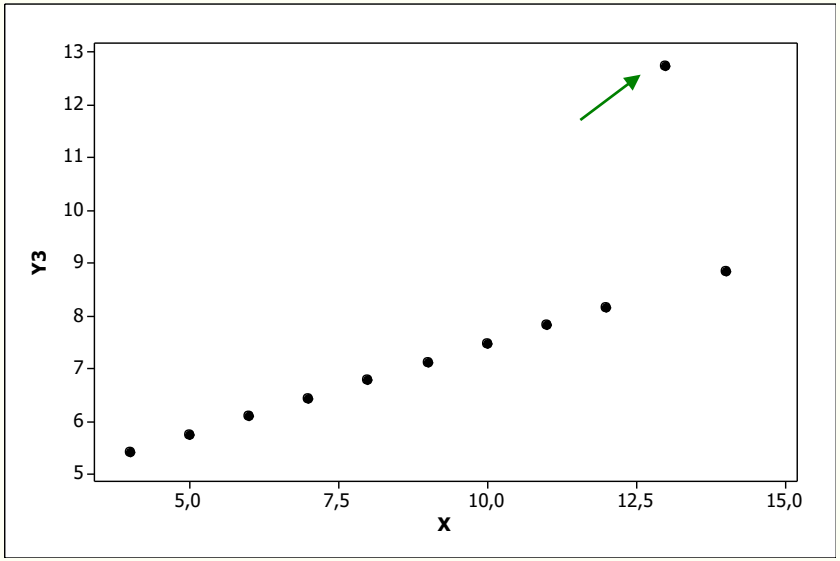
$r = 0{,}816$

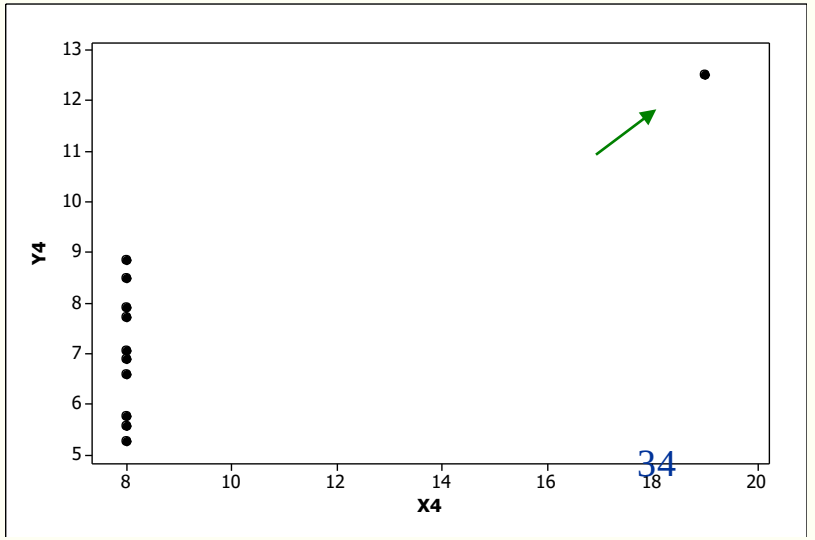
Dispersão esperada!





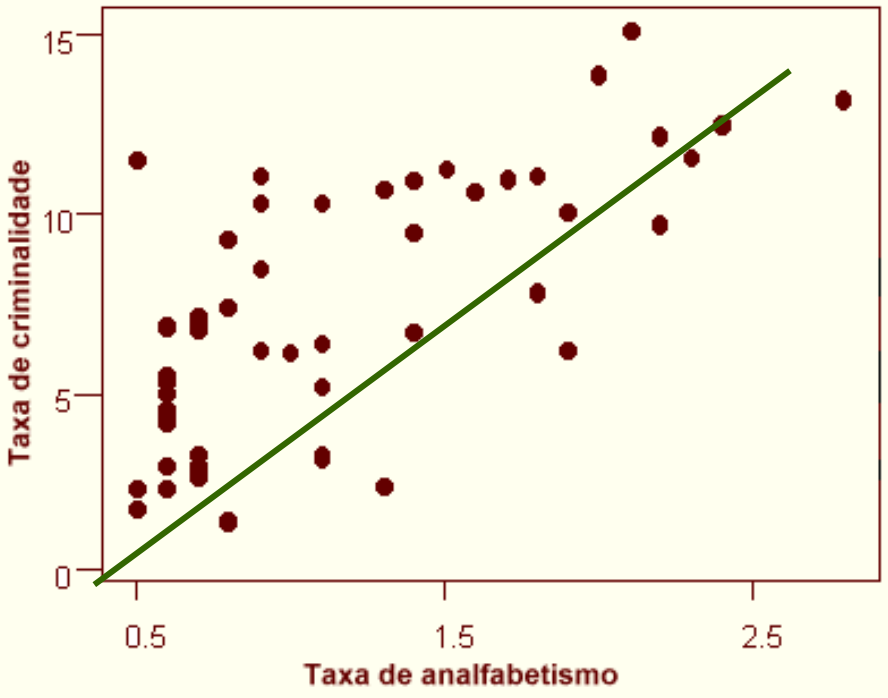
Pontos influentes!

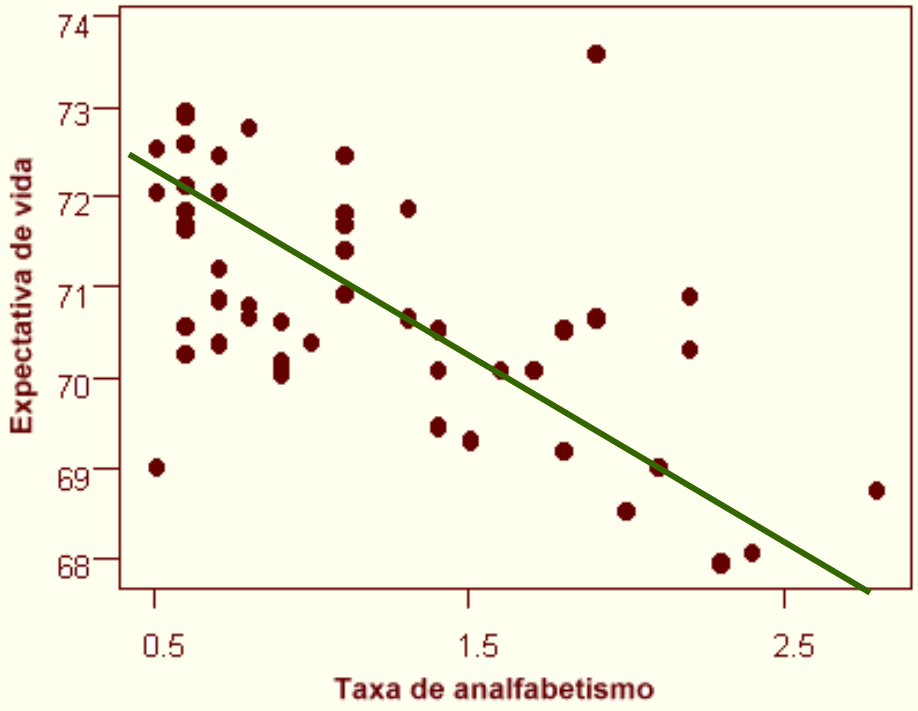




Análise de Regressão

Diagramas de Dispersão





⇒ Explicar a forma da relação por meio de uma função matemática: $Y = a + bX$

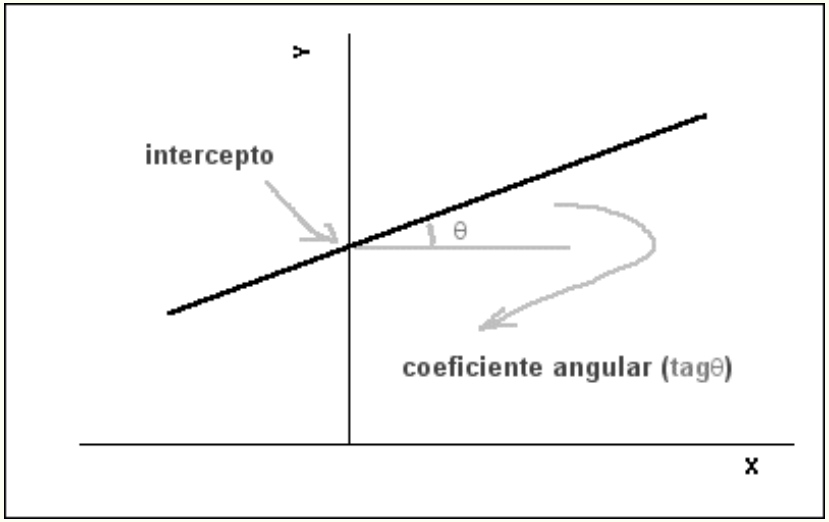
Análise de Regressão

Reta ajustada: $\hat{Y} = a + bX$

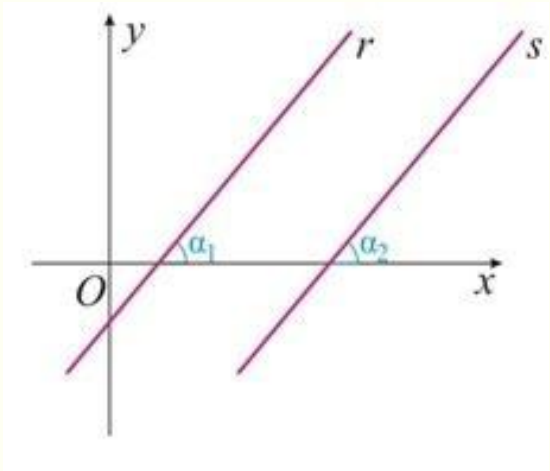
O que são a e b?

a : intercepto

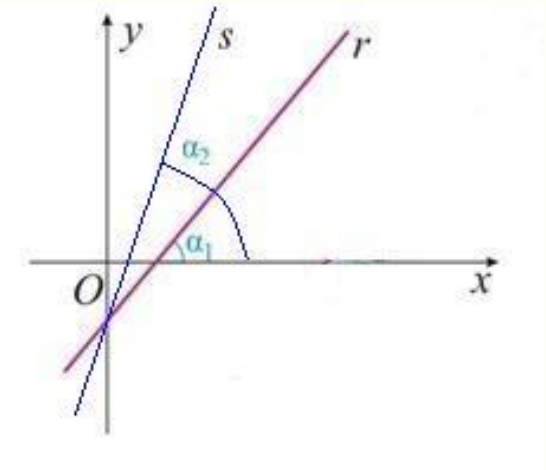
b : inclinação ou coeficiente angular



Análise de Regressão



- Iguais coeficientes angulares
- Diferentes interceptos



- Diferentes coeficientes angulares
- Iguais interceptos

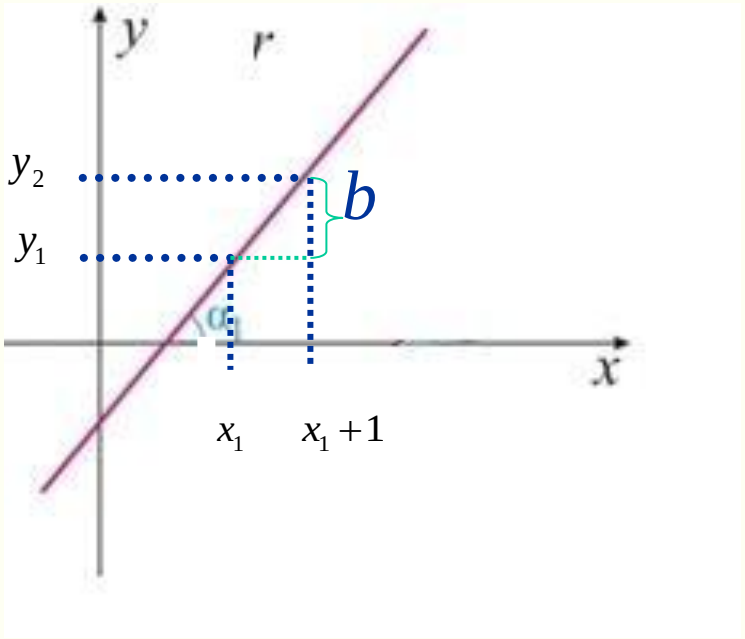
Reta ajustada: $\hat{Y} = a + bX$

Interpretação de b:

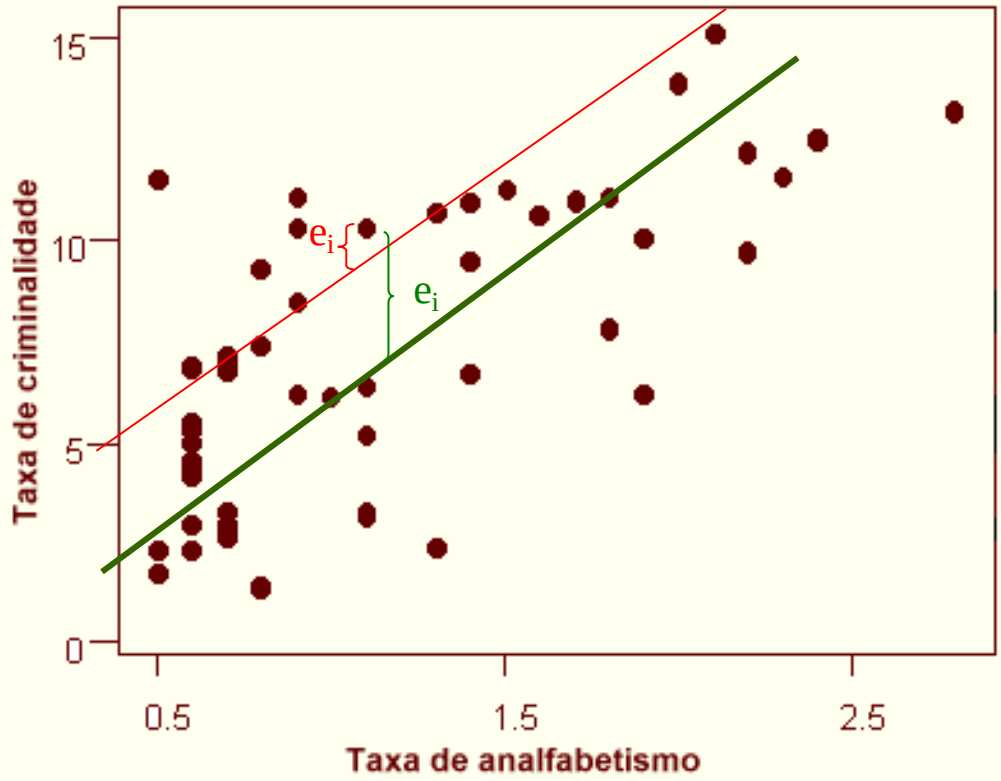
Para cada aumento de uma unidade em X, tem-se um aumento, em média, de b unidades em Y.

$$tag(\alpha) = \frac{y_2 - y_1}{x_2 - x_1} = \frac{y_2 - y_1}{x_1 + 1 - x_1}$$

$$= y_2 - y_1 = b$$



Reta ajustada (método de mínimos quadrados)



Reta ajustada (método de mínimos quadrados)

Os coeficientes a e b são calculados da seguinte maneira:

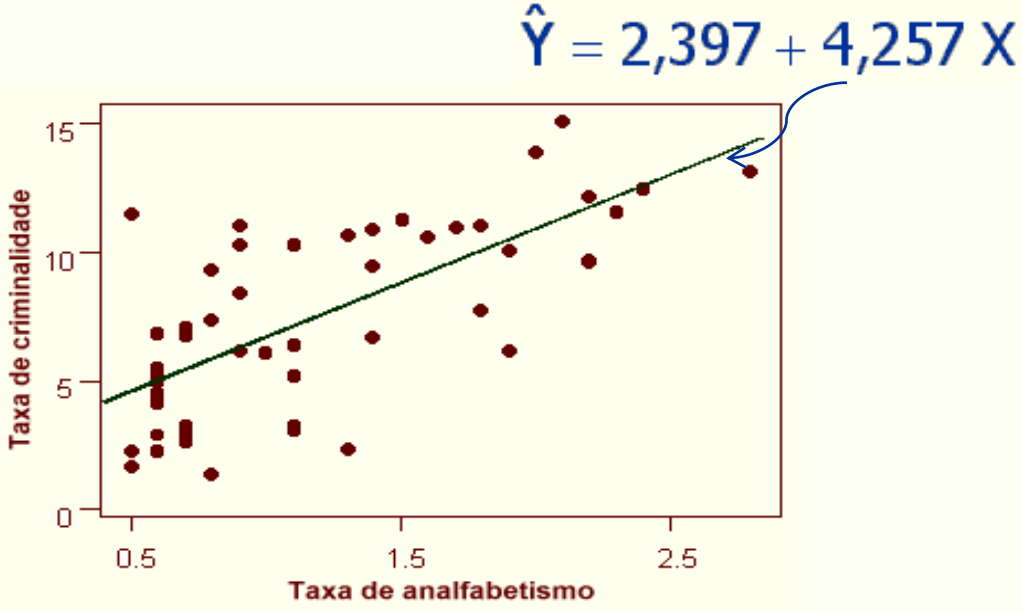
$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\,\overline{X}\,\overline{Y}}{(n-1)\,S_X^2}$$

$$a = \overline{Y} - b\,\overline{X}$$

No Exemplo 3,

A reta ajustada é:

$$\hat{Y} = 2{,}397 + 4{,}257\,X$$



$\hat{Y}$: valor predito para a taxa de criminalidade

X : taxa de analfabetismo

Interpretação de b:

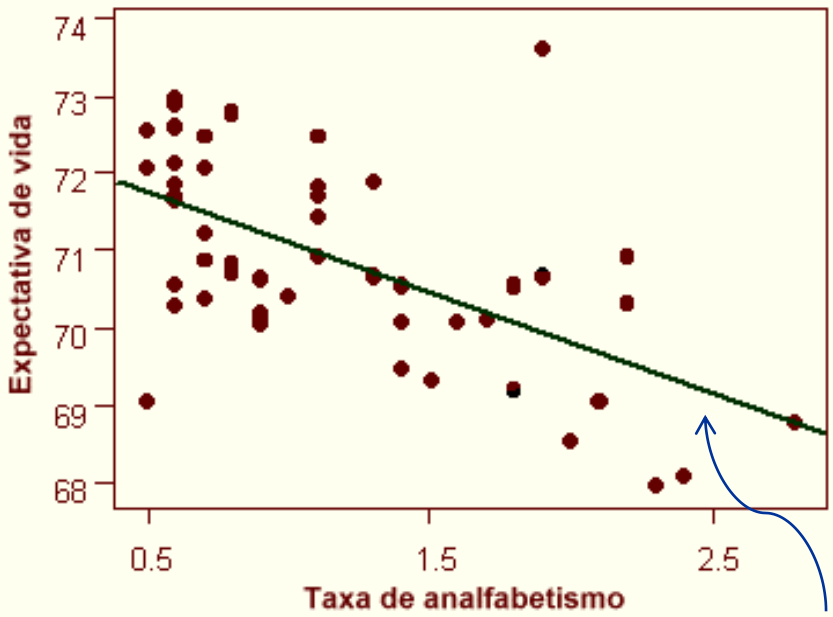
Para um aumento de uma unidade na taxa do analfabetismo (X), a taxa de criminalidade (Y) aumenta, em média, 4,257 unidades.

$\rightarrow$ Como desenhar a reta no gráfico?

No Exemplo 4,

A reta ajustada é:

$$\hat{Y} = 72{,}395 - 1{,}296\,X$$



$\hat{Y}$: valor predito para a expectativa de vida

X : taxa de analfabetismo

Interpretação de b:

Para um aumento de uma unidade na taxa do analfabetismo (X), a expectativa de vida (Y) diminui, em média, 1,296 anos.

Exemplo: consumo de cerveja e temperatura

Y: consumo de cerveja diário por mil habitantes, em litros.

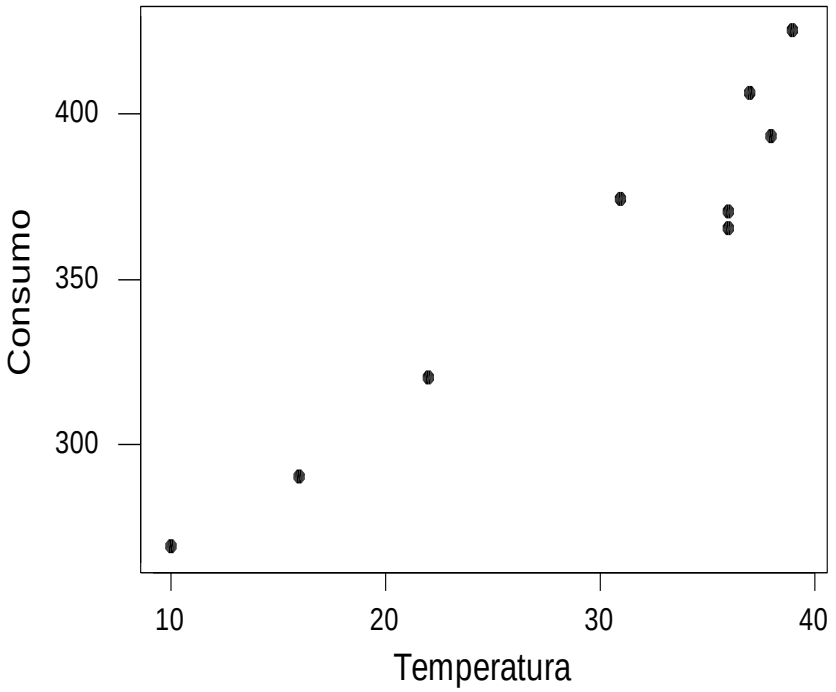
X: temperatura máxima (em ^{o}C).

As variáveis foram observadas em nove localidades com as mesmas características demográficas e socioeconômicas.

Dados:

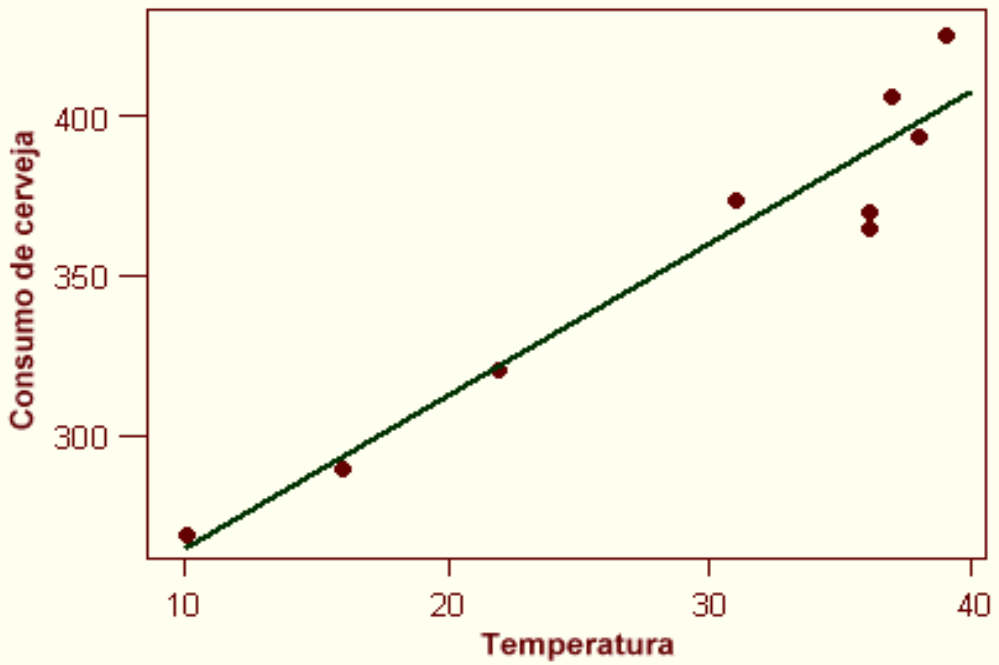
Localidade	Temperatura (X)	Consumo (Y)
1	16	290
2	31	374
3	38	393
4	39	425
5	37	406
6	36	370
7	36	365
8	22	320
9	10	269

Diagrama de dispersão



A correlação entre X e Y é $r = 0{,}962$.

A reta ajustada é: $\hat{Y} = 217{,}37 + 4{,}74\,X$



Qual é a interpretação de b?

Aumentando-se um grau na temperatura (X), o consumo de cerveja (Y) aumenta, em média, 4,74 litros por mil habitantes.

Qual é o consumo previsto para uma temperatura de $25^{o}C$?

$$\hat{Y} = 217{,}37 + 4{,}74 \times 25 = 335{,}87 \text{ litros/mil hab}$$