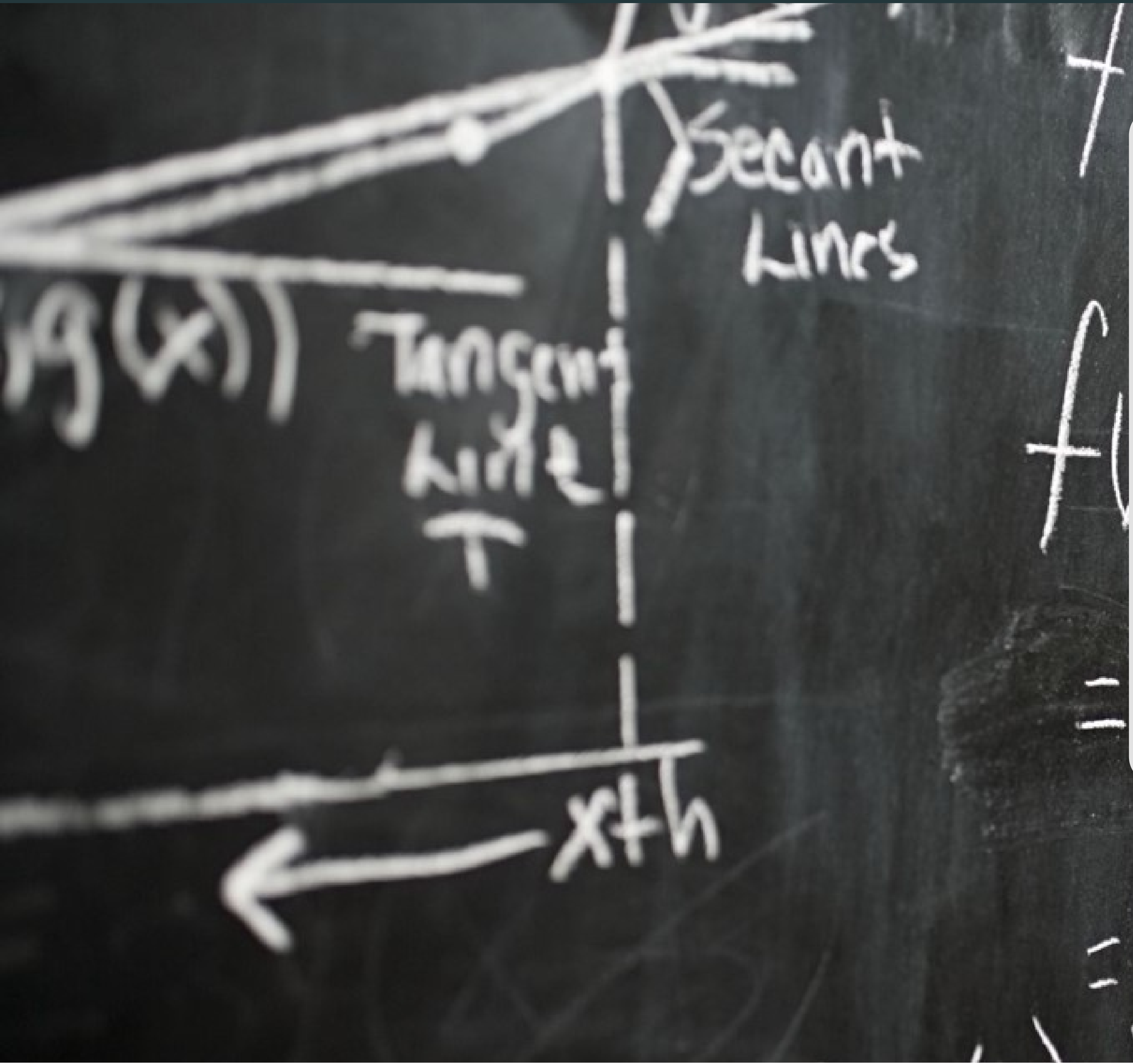
Estatística descritiva

Covariância e Correlação Linear

Fabio Cop (fabiocopf@gmail.com)
Instituto do Mar - UNIFESP
Última atualização em 25 de abril de 2022

Conteúdo da aula



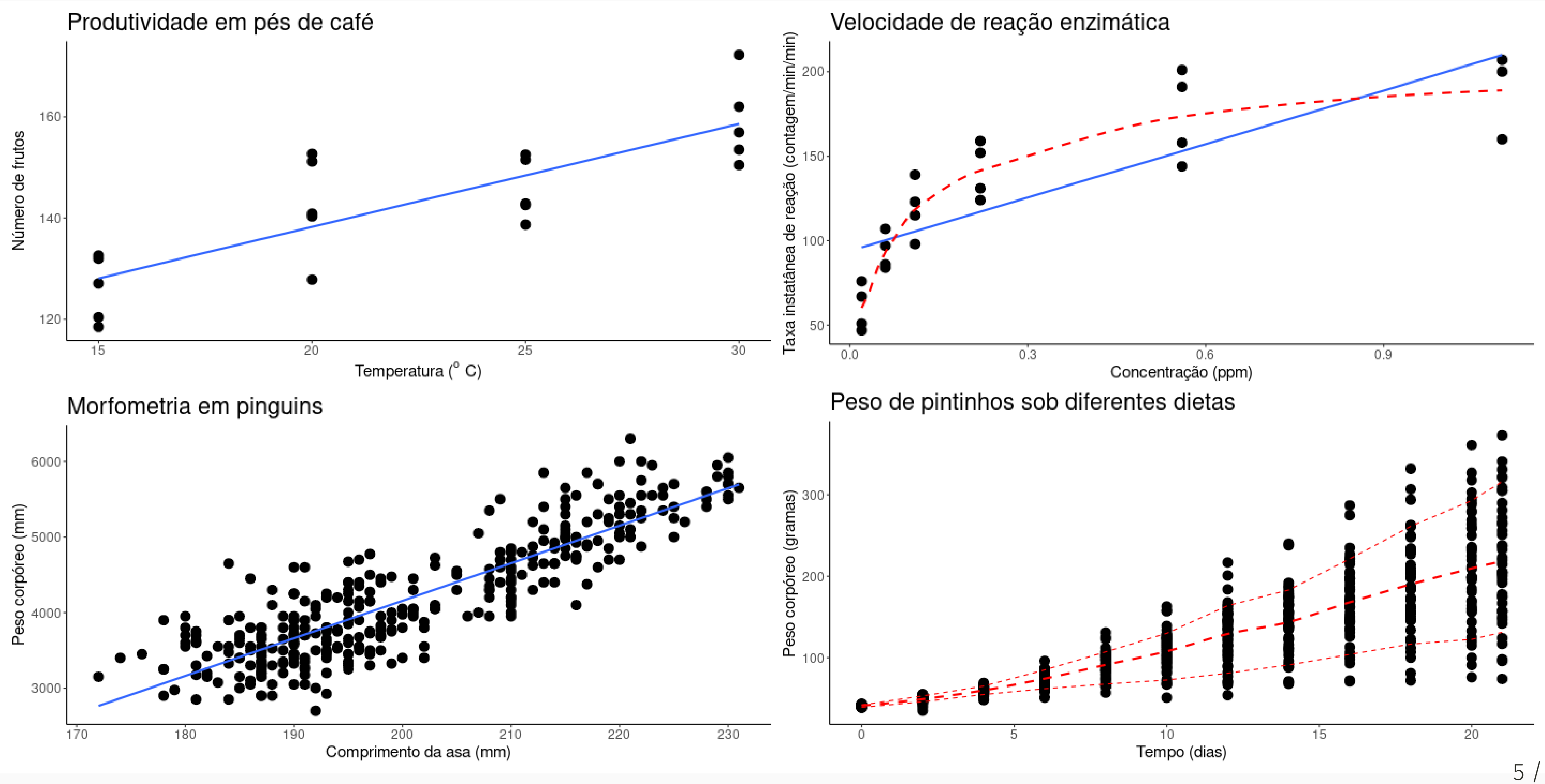
1. Medindo a intensidade de associações lineares
2. Soma dos Quadrados e Variância Amostral
3. Soma dos Produtos Cruzados e Covariância
4. O coeficiente de Correlaçao Linear de Pearson (r)
5. Associações Lineares e Causalidade
6. Os comandos em R

O coeficiente de correlação de Pearson

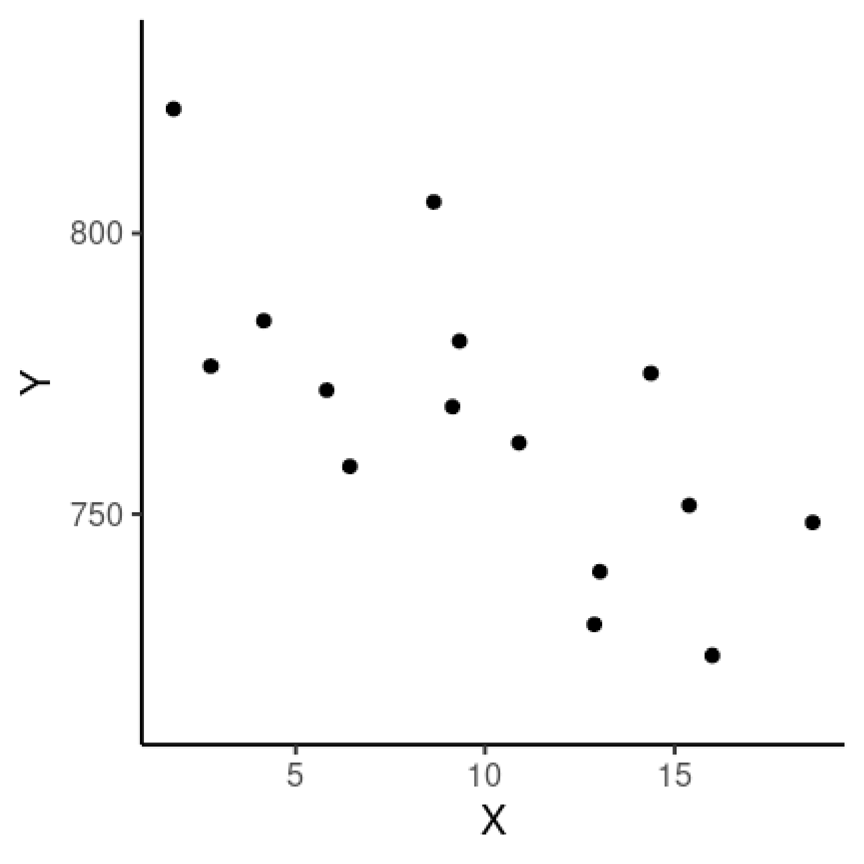
Um pouco de história

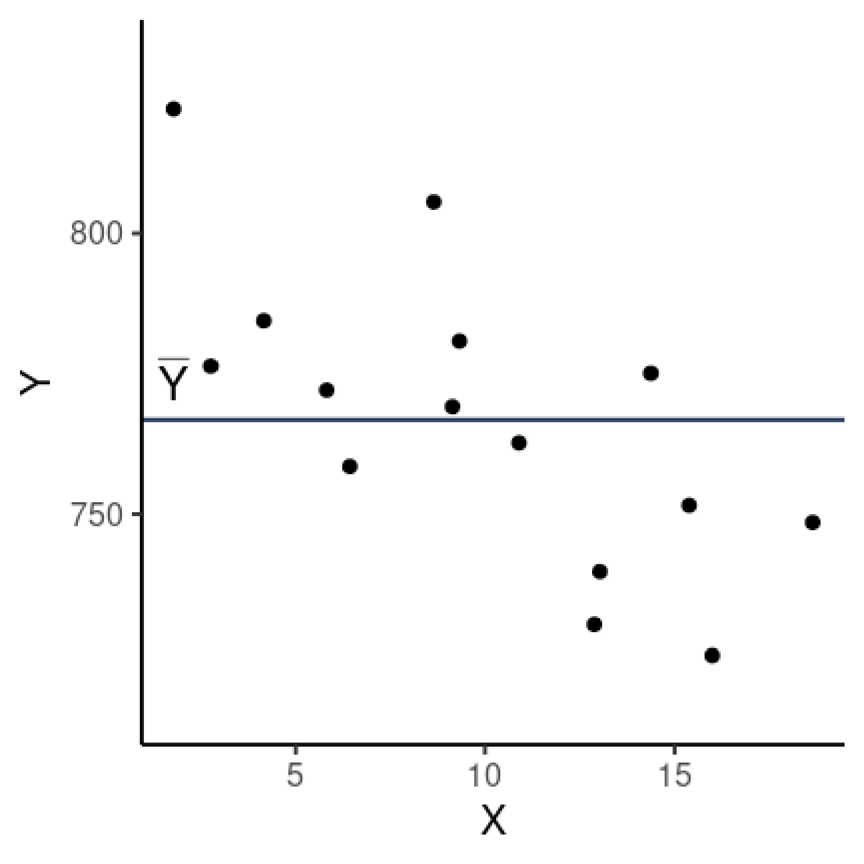
Na década de 1890, Karl Pearson foi apresentado a Francis Galton pelo zoólogo Walter Weldon. Juntos fundaram a revista Biometrika, com o objetivo de desenvolver teoria em estatística. Galton (primo de Charles Darwin) e Pearson trabalharam juntos em vários problemas relacionados à teoria da evolução, genética, biometria e estatística. Galton trouxe as primeiras ideias sobre a medida de associação entre duas variáveis quantitativas no contexto da hereditariedade e propôs o coeficiente de correlação linear para medir esta associação. Suas idéias foram estendidas por Karl Pearson e Udny Yule para um contexto estatístico mais geral. Pearson trouxe ainda muitas outras contribuições á estatística como o coeficiente de χ^2 e a ideia de *graus de liberdade*. O termo distribuição normal para variáveis com dustribuição Gaussiana também surgiu como fruto e seu trabalho (veja em: Karl Pearson).

1. Medindo a intensidade de associações lineares

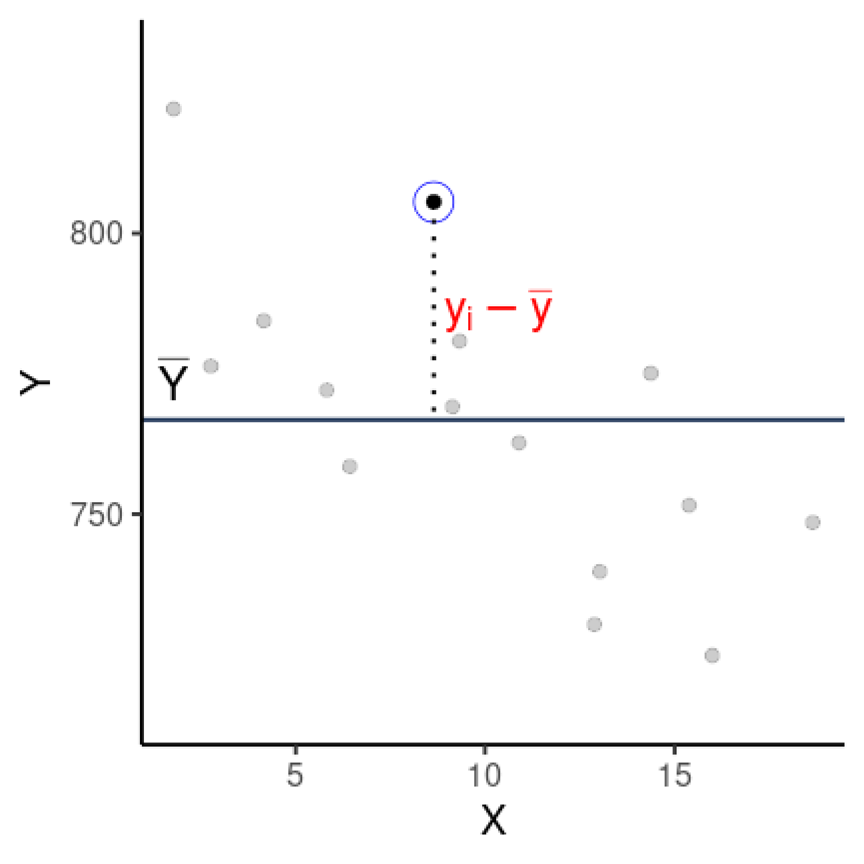


2. Soma dos Quadrados e Variância Amostral

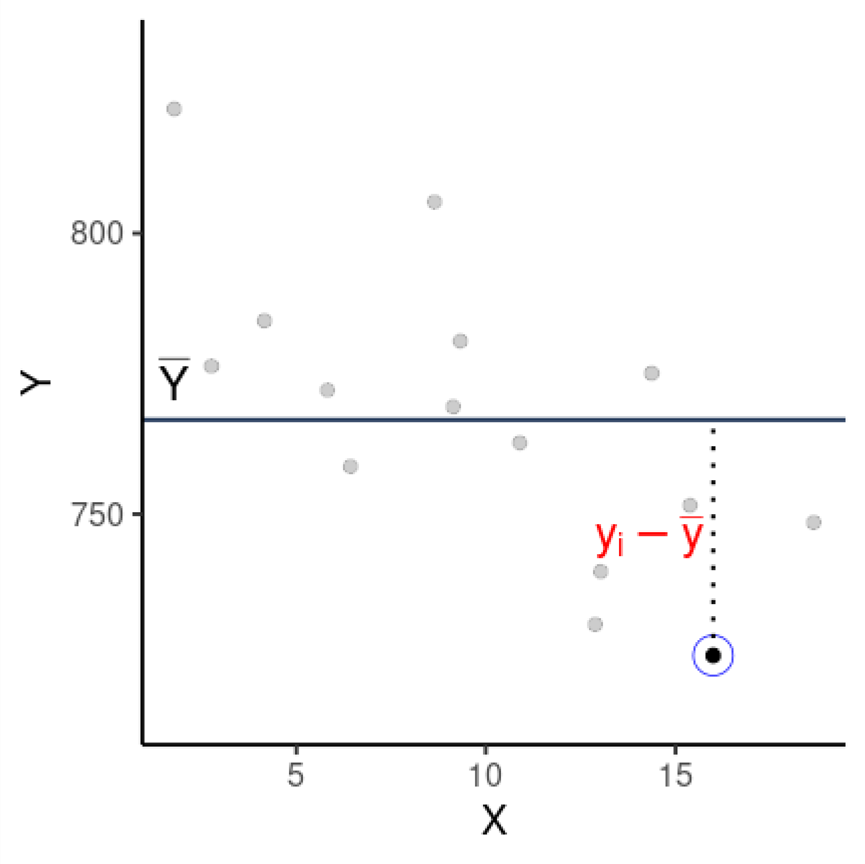


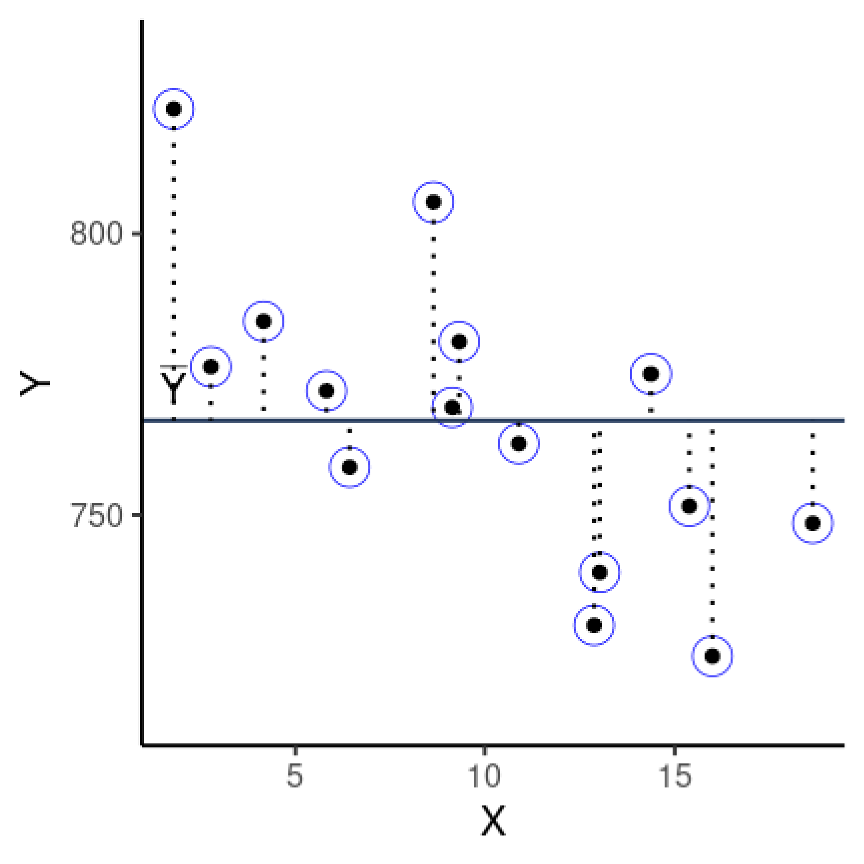

800
750
Y
$\overline{Y}$
5
10
15
X

2. Soma dos Quadrados e Variância Amostral



2. Soma dos Quadrados e Variância Amostral



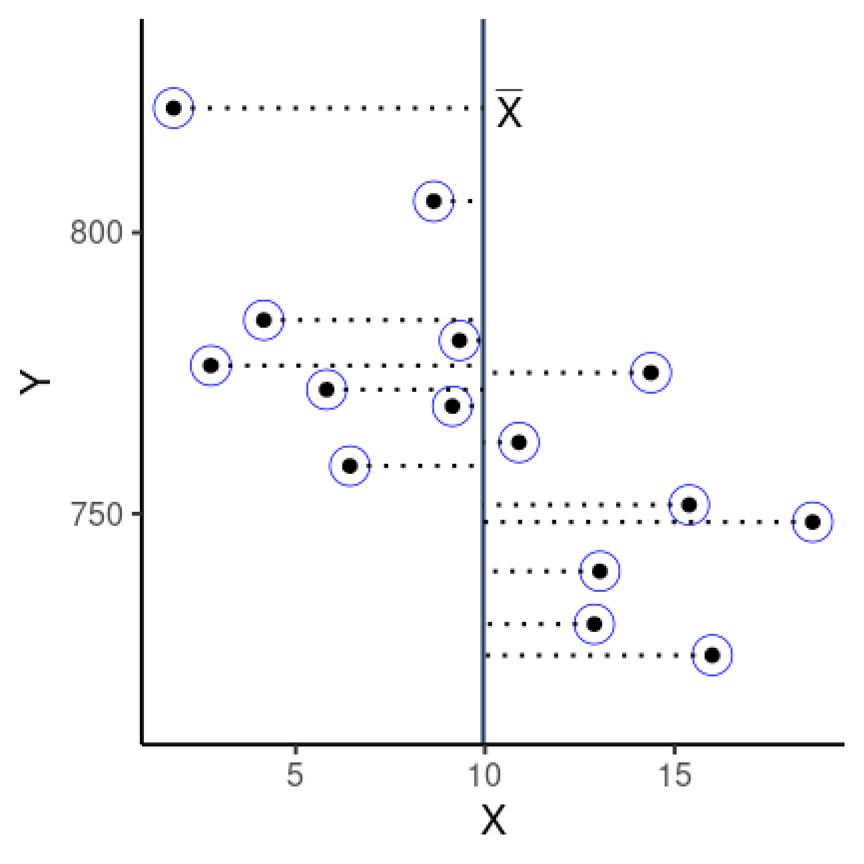


Soma dos Quadrados de Y

$$SQ_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \bar{y})(y_i - \bar{y})$$

Variância amostral de Y

$$s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

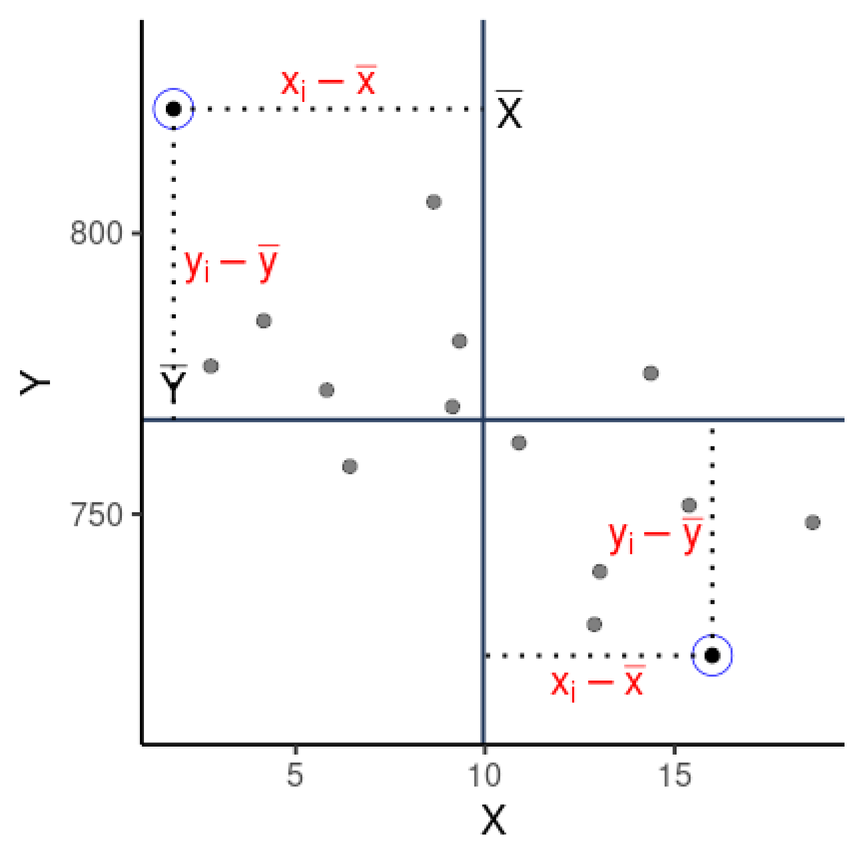


Soma dos Quadrados de X

$$SQ_X = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})$$

Variância amostral de X

$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$



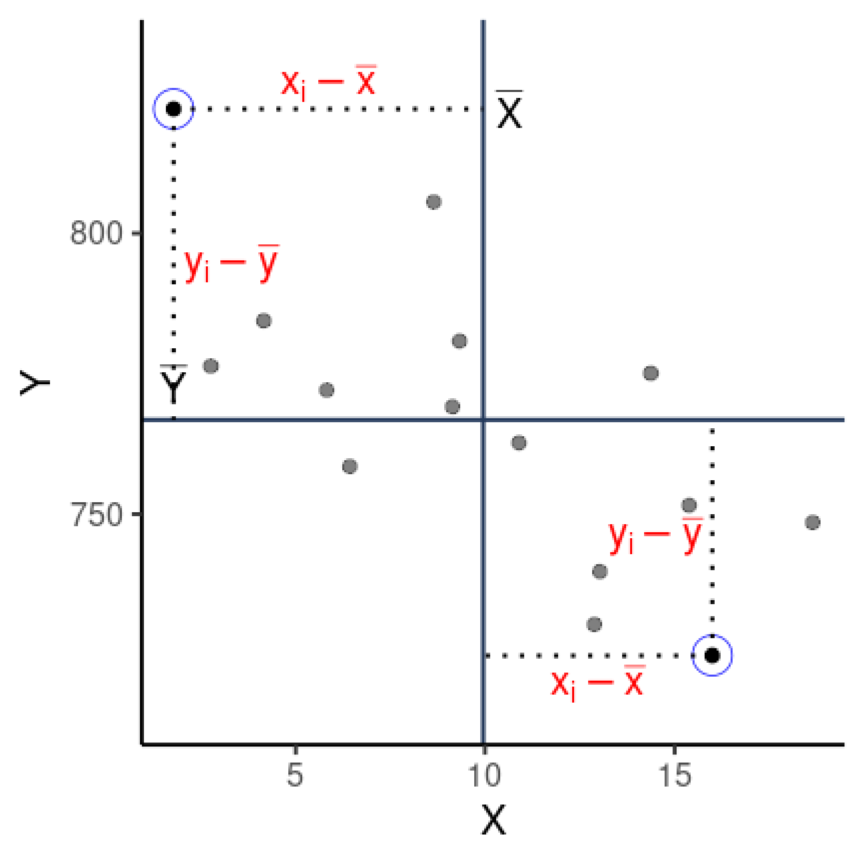
Soma dos produtos cruzados de Y e X

$$SQ_{YX} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Covariância amostral entre Y e X

$$s_{YX} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{n - 1}$$

3. Soma dos Produtos Cruzados e Covariância



Se

$$(y_i - \bar{y}) > 0; (x_i - \bar{x}) < 0$$

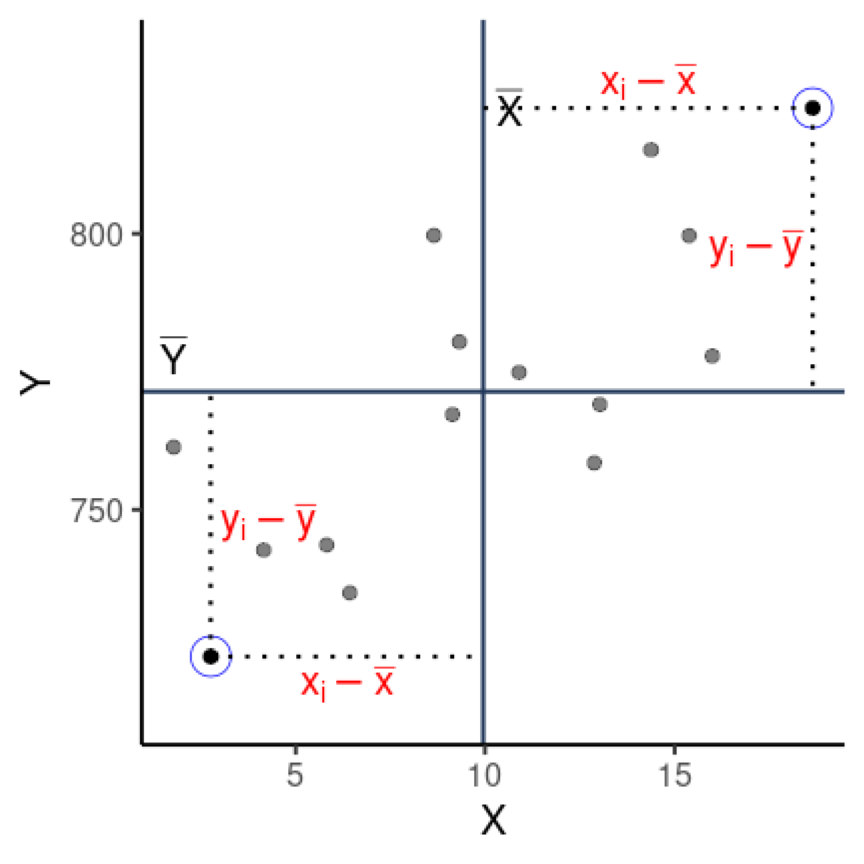
ou

$$(y_i - \bar{y}) < 0; (x_i - \bar{x}) > 0$$

temos

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} < 0$$

A covariância pode ser **NEGATIVA**



Se

$(y_i - \bar{y}) > 0; (x_i - \bar{x}) > 0$

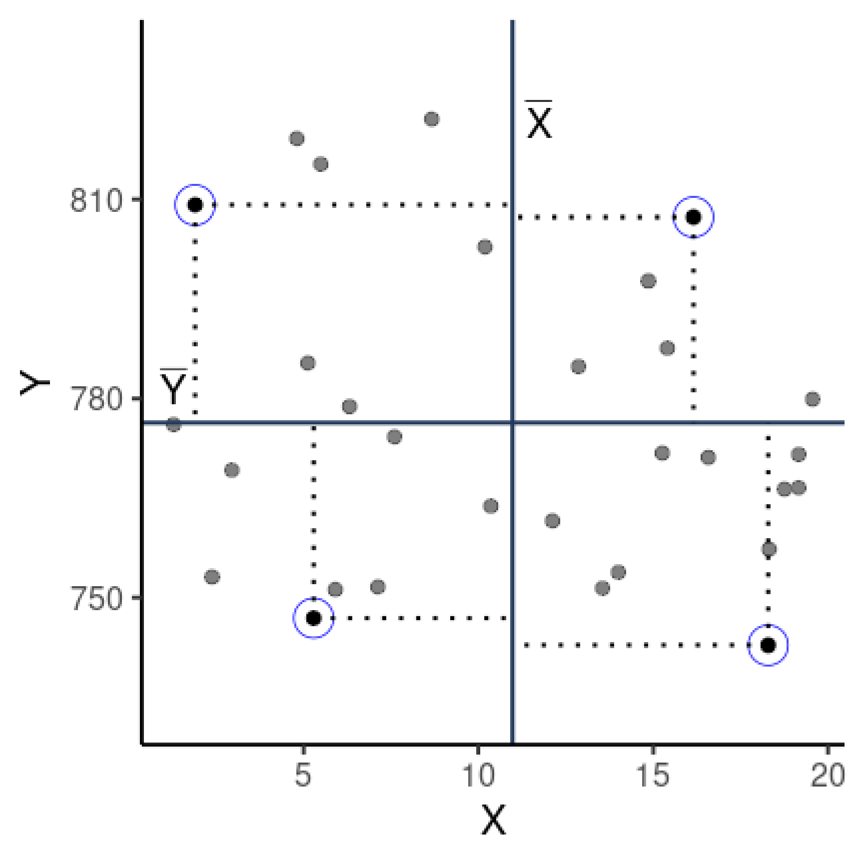
ou

$(y_i - \bar{y}) < 0; (x_i - \bar{x}) < 0$

temos

$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1} > 0$

A covariância pode ser **POSITIVA**



Se

$(y_i - \overline{y}) \approx 0; (x_i - \overline{x}) \approx 0$

ou

$(y_i - \overline{y}) \approx 0; (x_i - \overline{x}) \approx 0$

Temos

$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n-1} \approx 0$

A covariância pode ser **NULA**

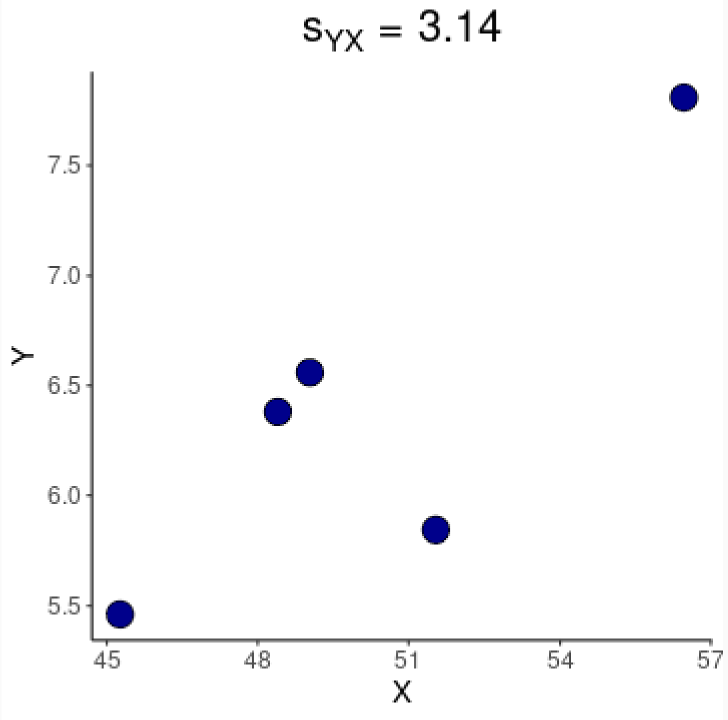
3. Soma dos Produtos Cruzados e Covariância

Cálculo da covariância entre Y e X

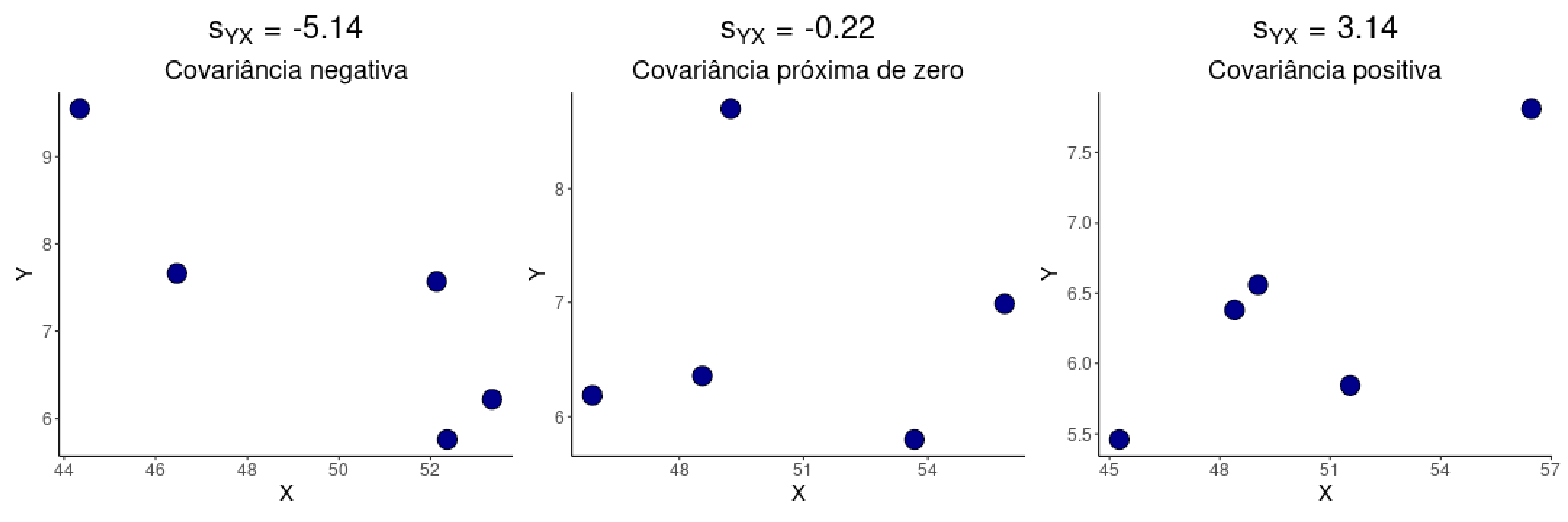
	Y	X	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})(x_i - \bar{x})$
1	45.26	5.46	-4.88	-0.95	4.63
2	49.04	6.56	-1.10	0.15	-0.16
3	51.54	5.84	1.40	-0.57	-0.79
4	56.46	7.81	6.32	1.40	8.84
5	48.40	6.38	-1.74	-0.03	0.05
$\sum$	**50.14**	**6.41**	**0.00**	**0.00**	**12.57**

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

$$s_{YX} = \frac{12.57}{5-1} = 3.14$$



Cenários possíeis



4. O coeficiente de correlaçao linear de Pearson

Covariância amostral entre Y e X

$$s_{YX} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{n-1}$$

Variância amostral de Y

$$s_Y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Variância amostral de X

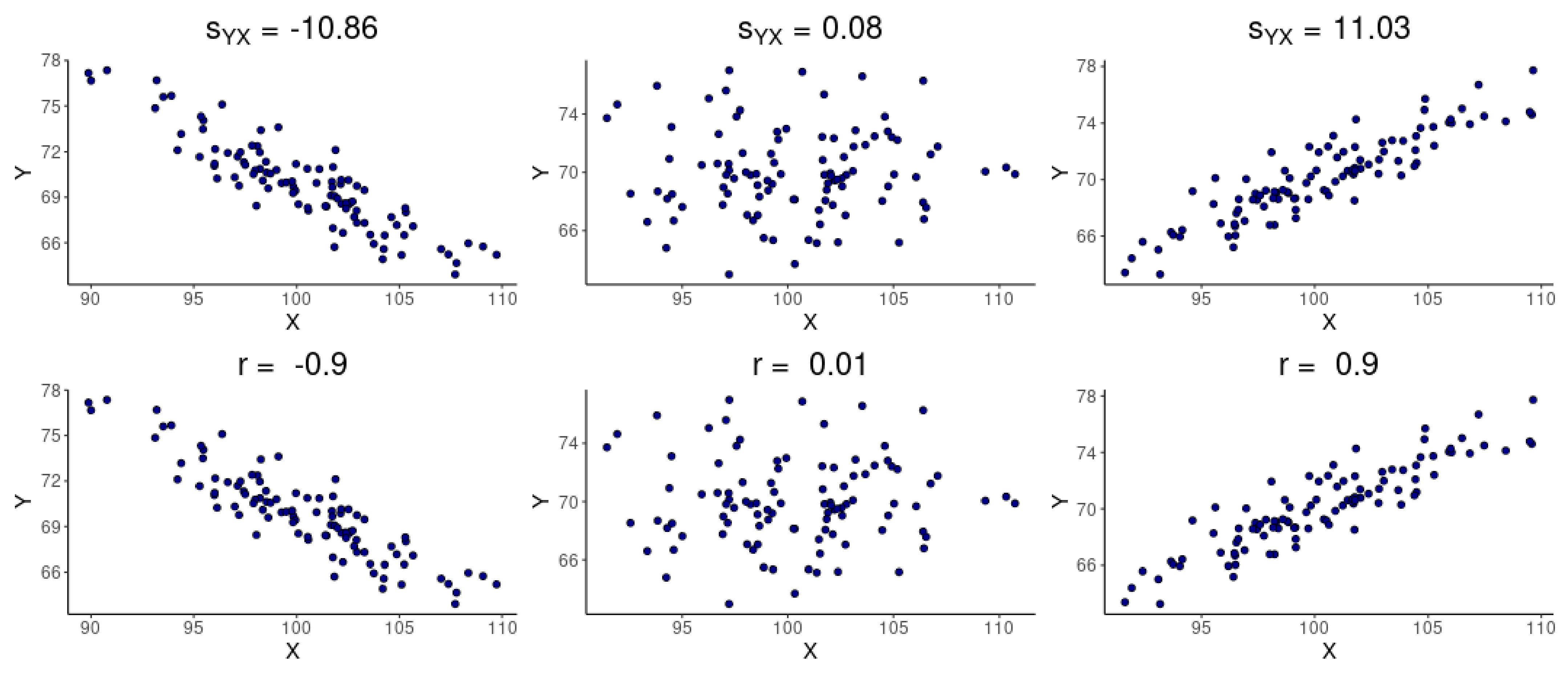
$$s_X^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

O coeficiente de correlaçao linear de Pearson r

$$r = \frac{s_{YX}}{\sqrt{s_Y^2} \times \sqrt{s_X^2}}$$

O r de Pearson é a covariância **padronizada** pelos desvios padrões de Y e X

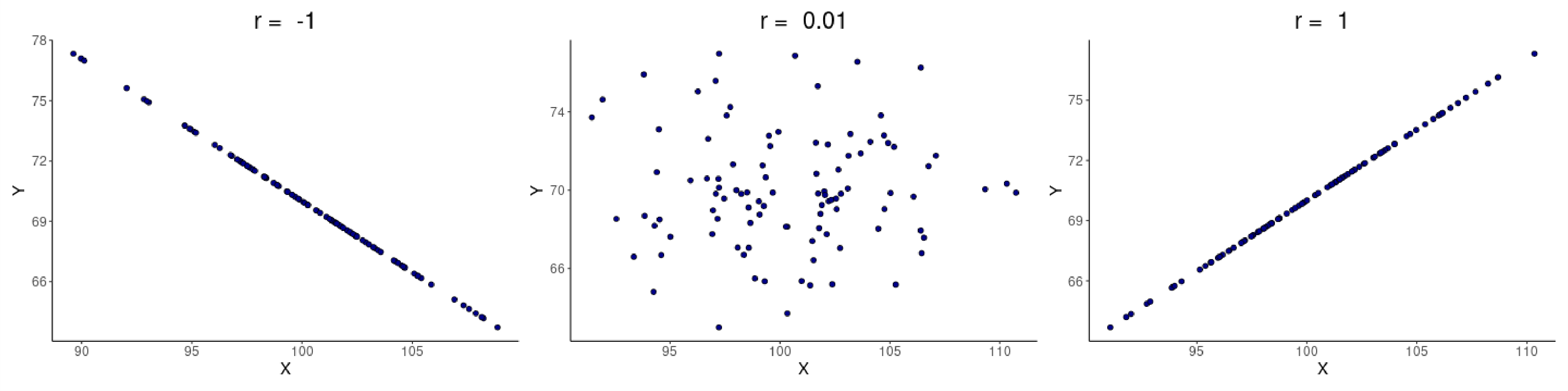
A covariância não tem limites negativos ou positivos. A escala depende das magnitudes de Y e de X.



O r de Pearson varia entre -1 e $+1$.

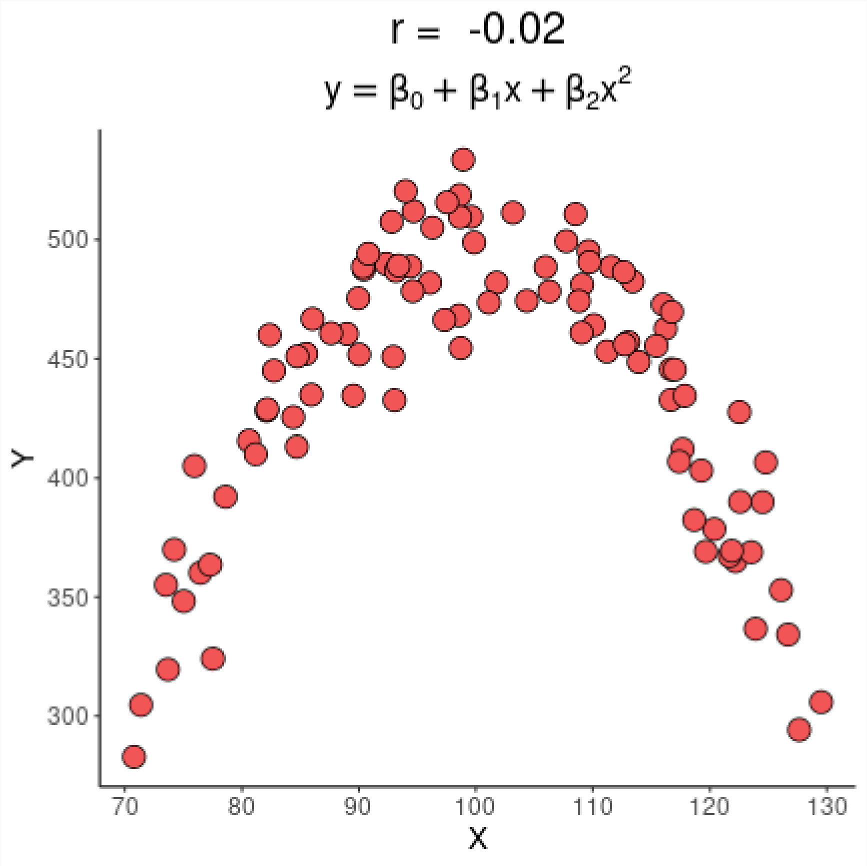
4. O coeficiente de correlaçao linear de Pearson

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

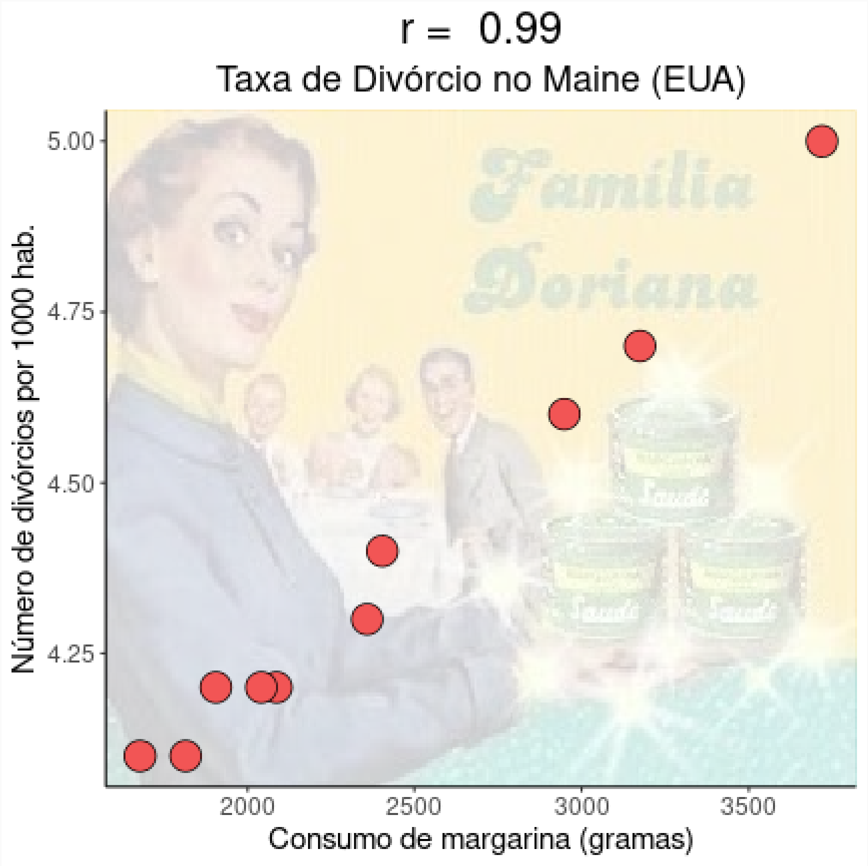


- $r = -1$ (Associção linear perfeitamente **negativa**)
- $r = 0$ (Associção linear inexistente)
- $r = 1$ (Associção linear perfeitamente **positiva**)

O r mede associações **lineares**



Correlação **não implica** causalidade



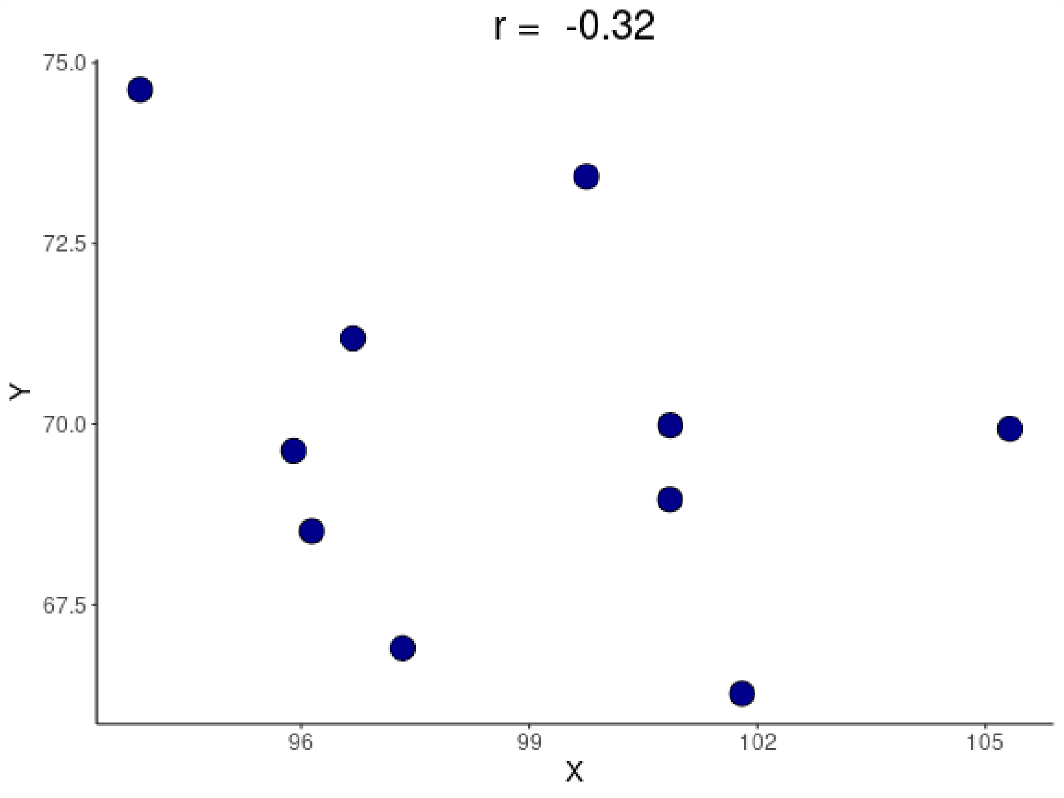
6. Teste de hipóteses sobre o r de Pearson

Dada uma **amostra** com n observações para os pares Y e X, a correlação entre Y e X na **população estatística** é diferente de zero?

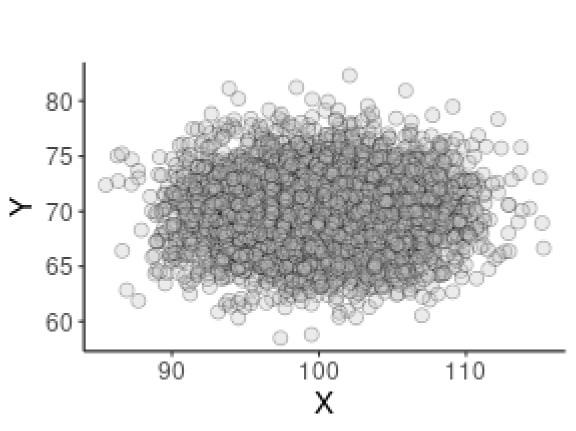
$H_0 : \rho = 0$

$H_a : \rho \neq 0$

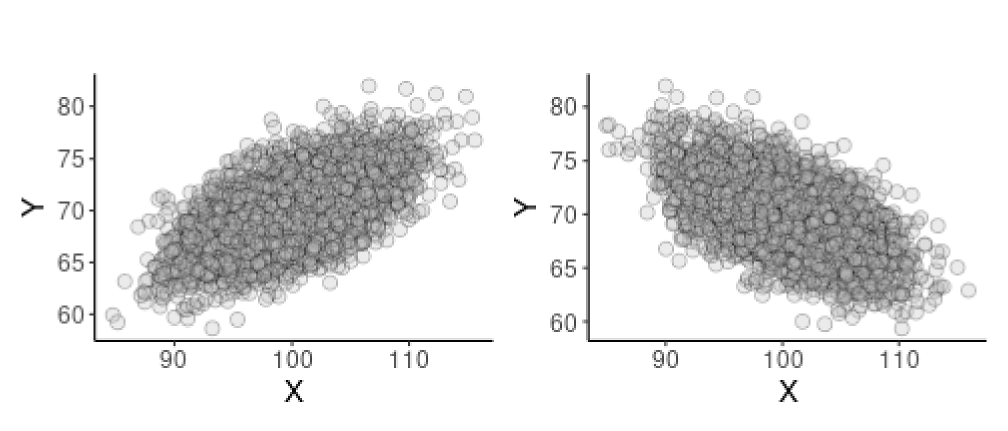
$n = 10$



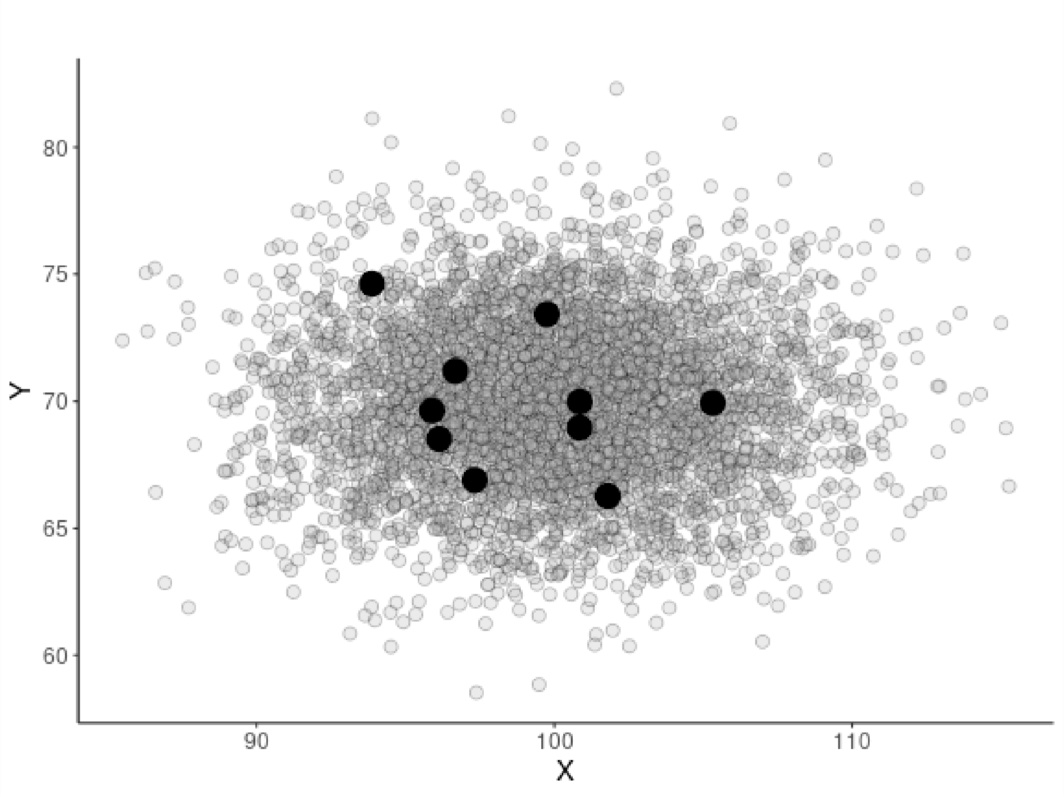
$H_0 : \rho = 0$



$H_a : \rho \neq 0$



Os dados segundo H_0



6. Teste de hipóteses sobre o r de Pearson

Assumimos que distribuição conjunta entre $f(Y, X)$ é Normal.

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

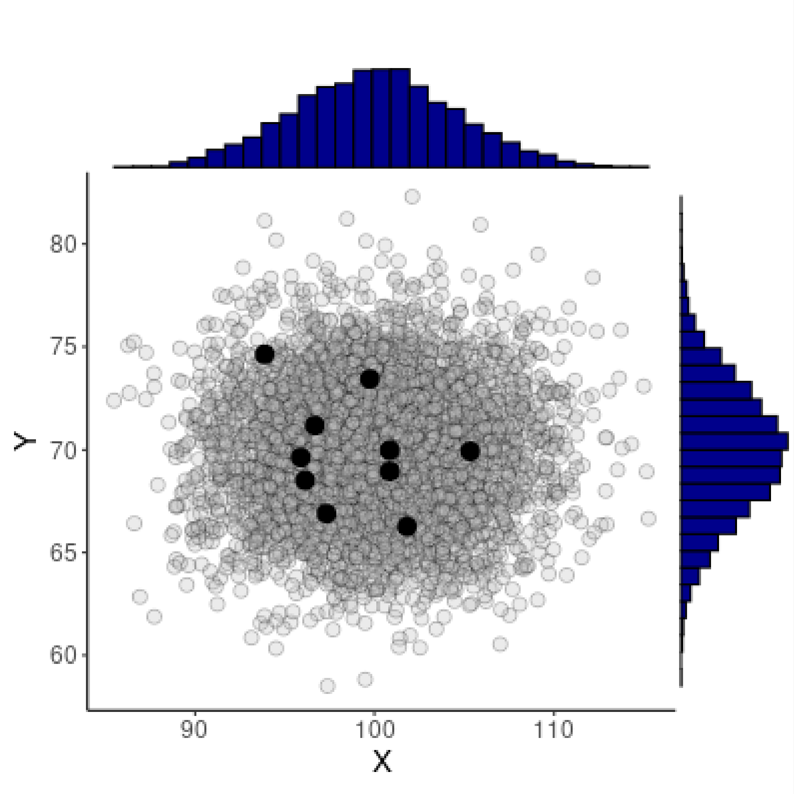
$n = 10$

$r = -0.32$

Estatística do teste - t

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Segundo H_0



6. Teste de hipóteses sobre o r de Pearson

Teste de hipótese sobre ρ

$\overline{Y} = 98.85$; $\overline{X} = 69.94$; $n = 10$

$r = -0.32$

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.32}{\sqrt{\frac{1-(-0.32)^2}{8}}} = -0.965$$

$p = 0.363$

Assumindo $\alpha = 0.05$, **Aceito** H_0:

> Não há evidências de correlação entre Y e X.

Cálculo do coeficiente de correlação

	Y	X	$\sum(y_i - \bar{y})^2$	$\sum(x_i - \bar{x})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	95.9	69.63	8.72	0.10	0.92
2	101.8	66.27	8.68	13.50	-10.83
3	100.85	69.98	4.01	0.00	0.08
4	99.75	73.43	0.81	12.13	3.14
5	93.88	74.63	24.69	21.95	-23.28
6	97.33	66.9	2.30	9.27	4.62
7	96.68	71.19	4.71	1.55	-2.70
8	100.85	68.96	3.99	0.97	-1.97
9	96.14	68.52	7.36	2.03	3.87
10	105.32	69.93	41.87	0.00	-0.06
$\sum$			**107.15**	**61.52**	**-26.21**

Aumentando o tamanho amostral

$H_0 : \rho = 0$

$H_a : \rho \neq 0$

$\alpha = 0.05$

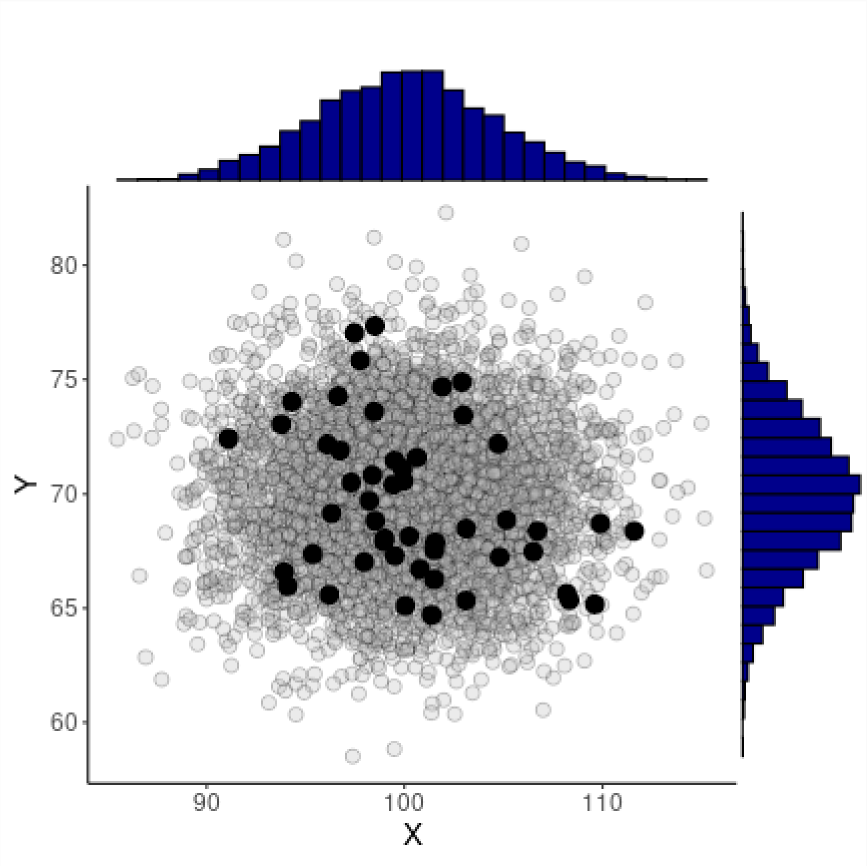
$n = 50$

$r = -0.32$

Estatística do teste - t

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Segundo H_0



Teste de hipótese sobre ρ

$\overline{Y} = 100.41$; $\overline{X} = 69.64$; $n = 50$

$r = -0.32$

$$t_{calculado} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{-0.32}{\sqrt{\frac{1-(-0.32)^2}{48}}} = -2.363$$

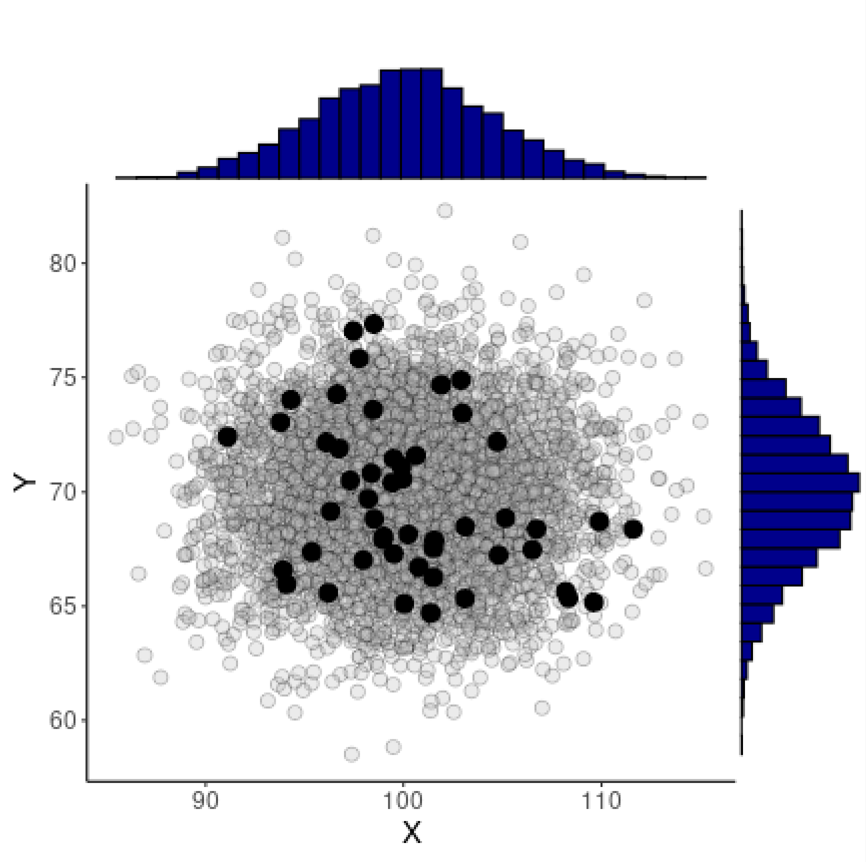
$p = 0.022$

Assumindo $\alpha = 0.05$, **Rejeito** H_0:

> Há evidências de correlação entre Y e X

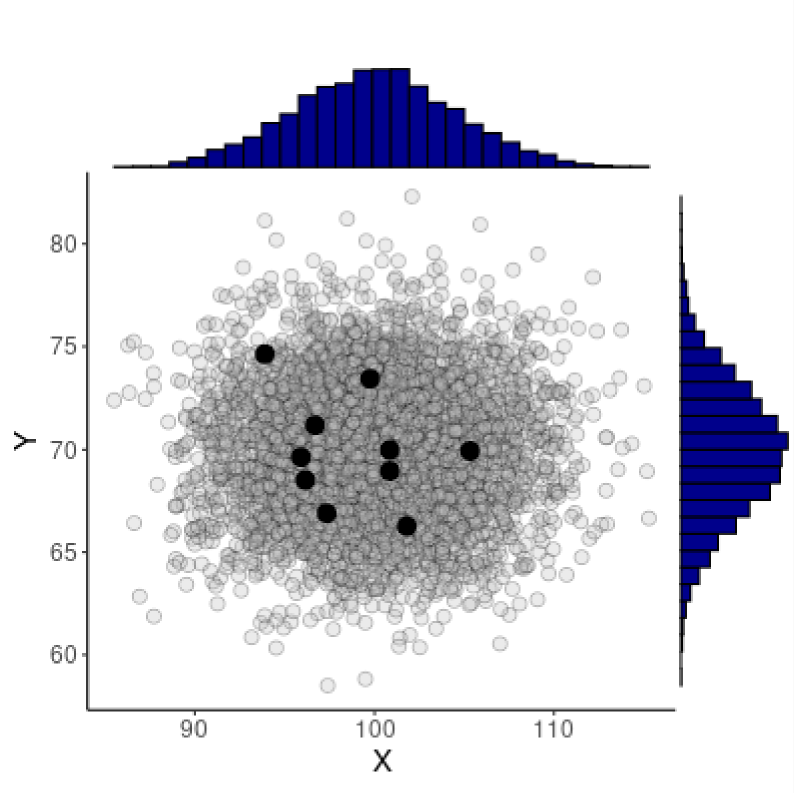
.

Segundo H_0



$r = -0.32$; $n = 10$

$t_{calculado} = -0.965$; $p = 0.363$



$r = -0.32$; $n = 50$

$t_{calculado} = -2.363$; $p = 0.022$