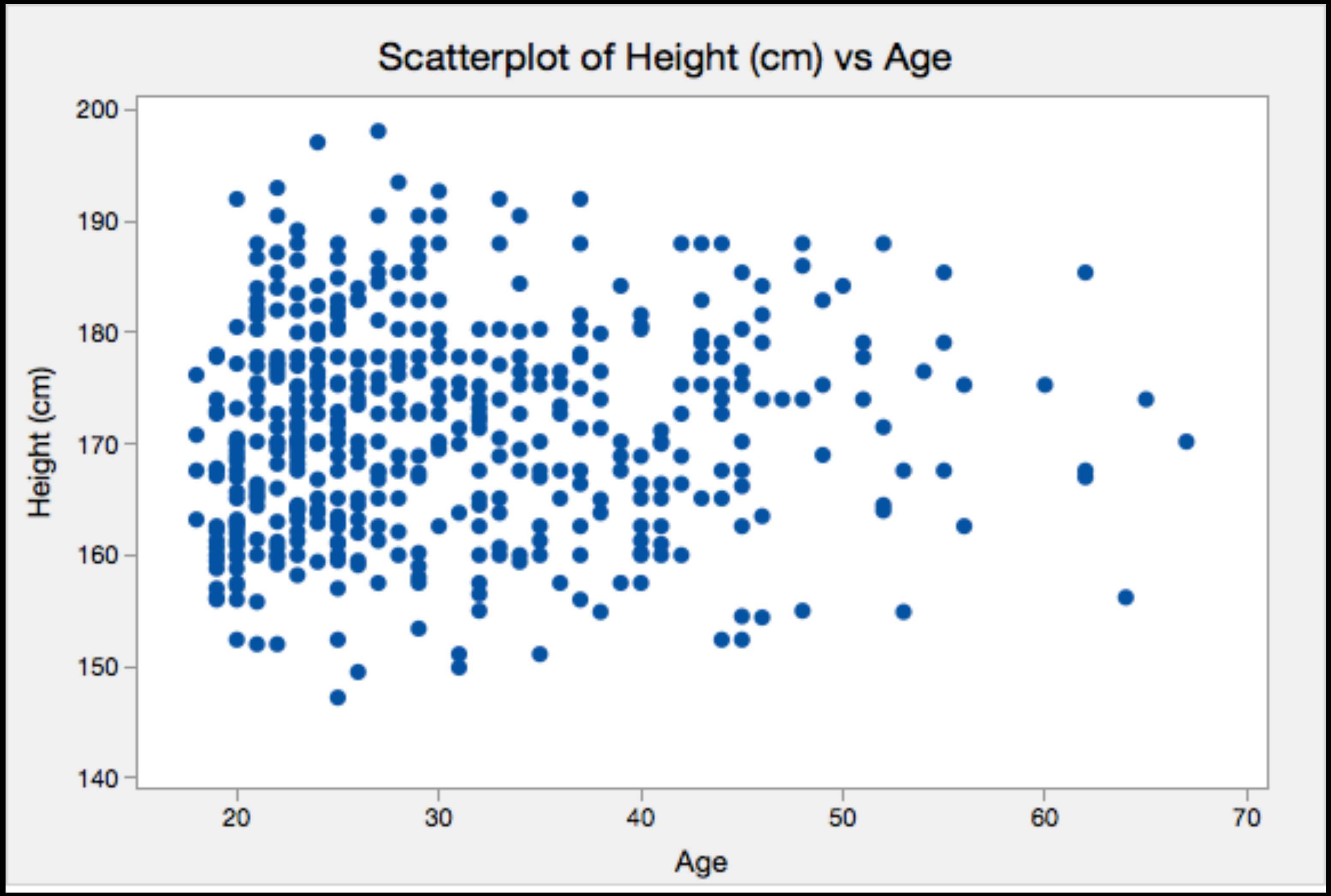
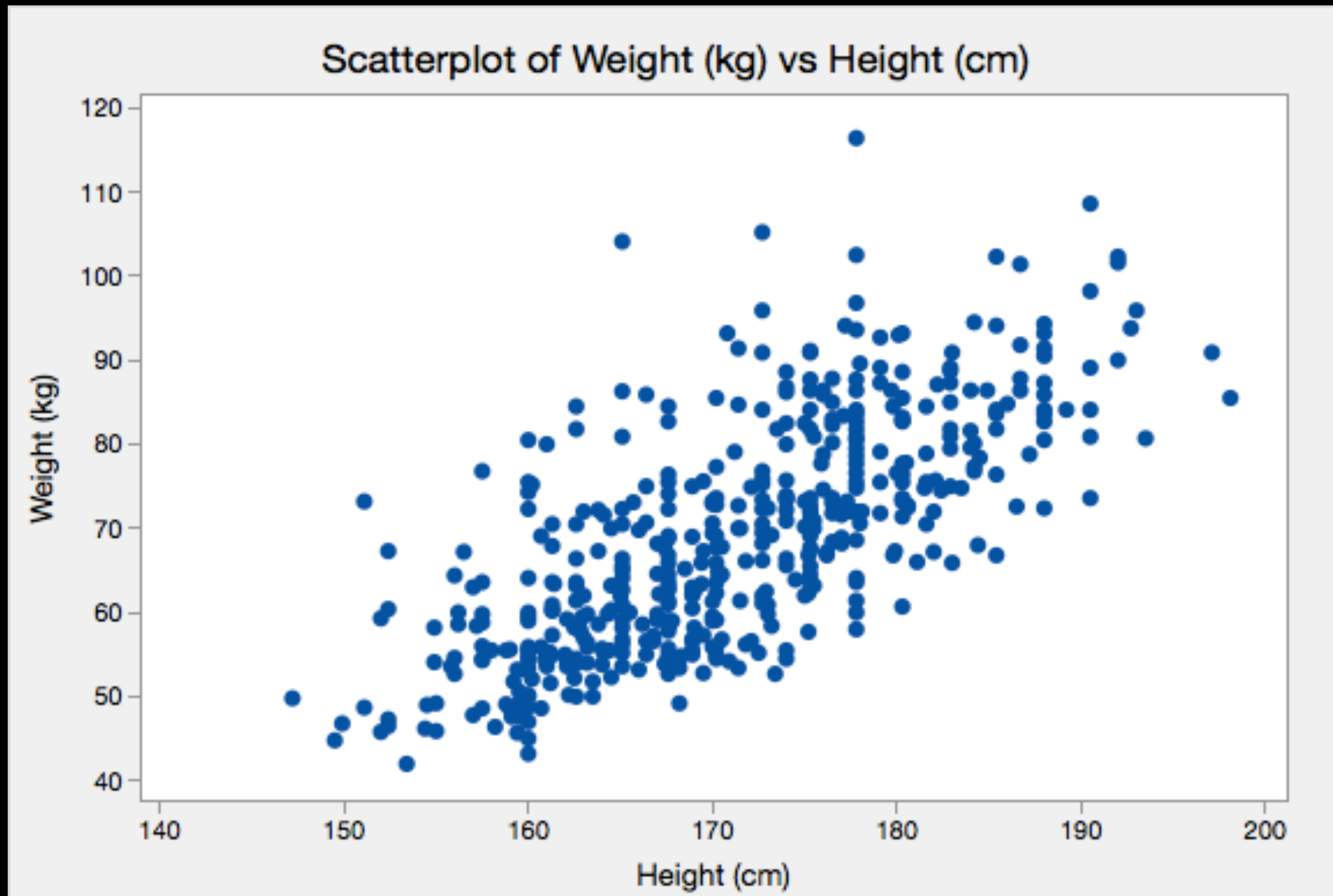
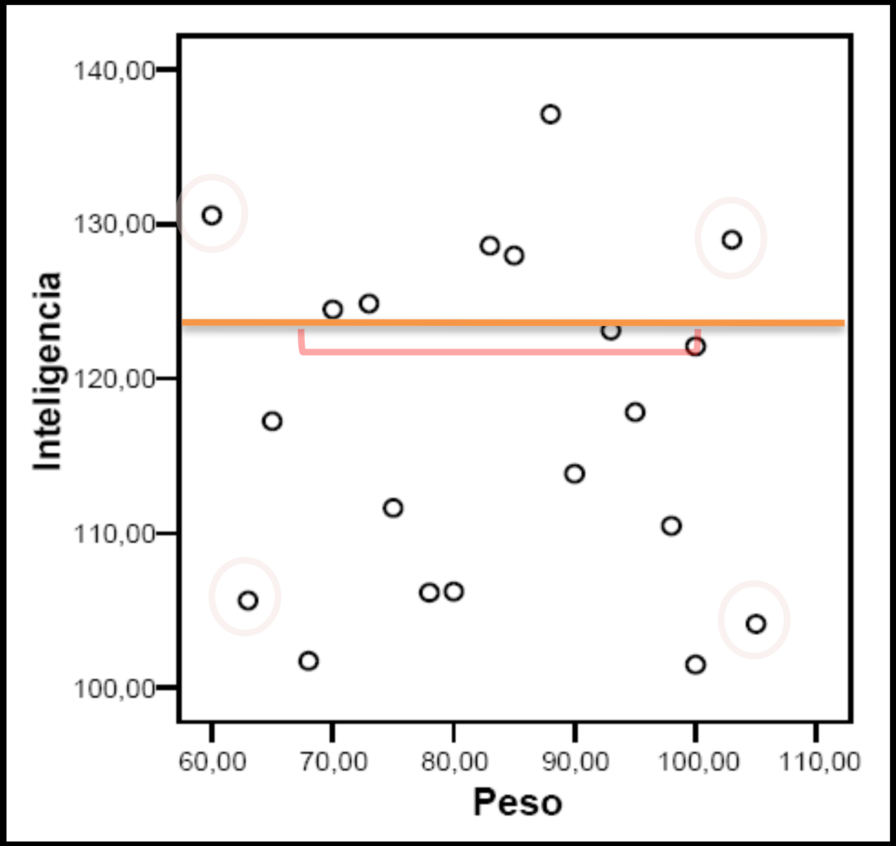
Measures of association- continuos variables

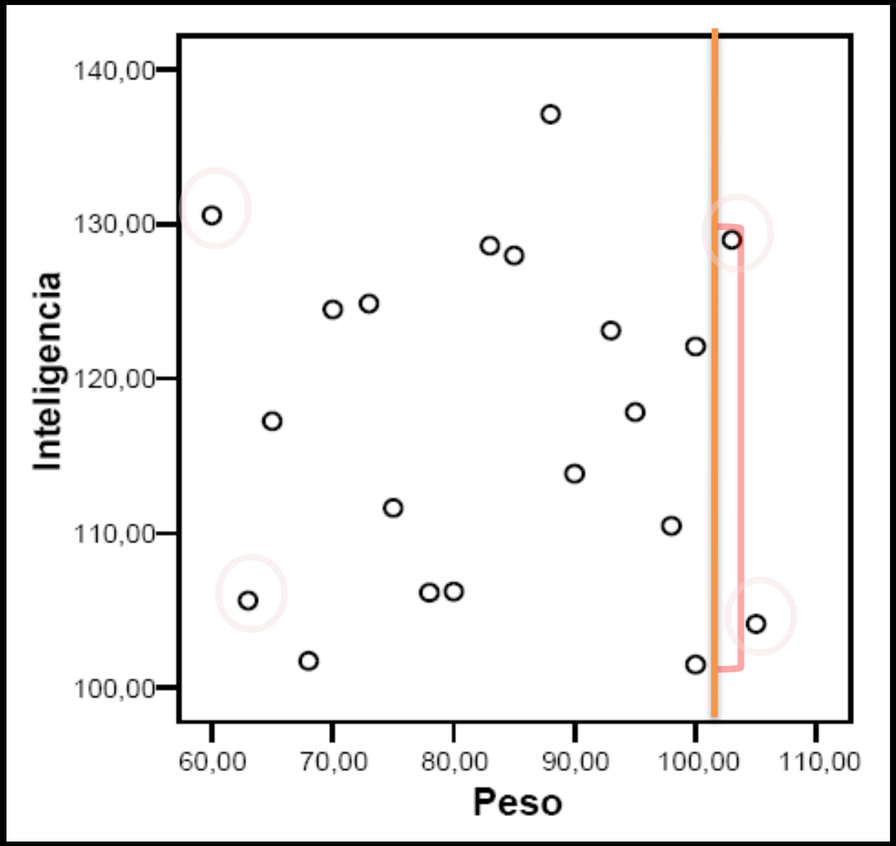

Scatterplot of Height (cm) vs Age
Height (cm)
200
190
180
170
160
150
140
20
30
40
50
60
70
Age


Scatterplot of Weight (kg) vs Height (cm)
Weight (kg)
Height (cm)
120
110
100
90
80
70
60
50
40
140
150
160
170
180
190
200

Let’s develop a bit the intuition….

No association

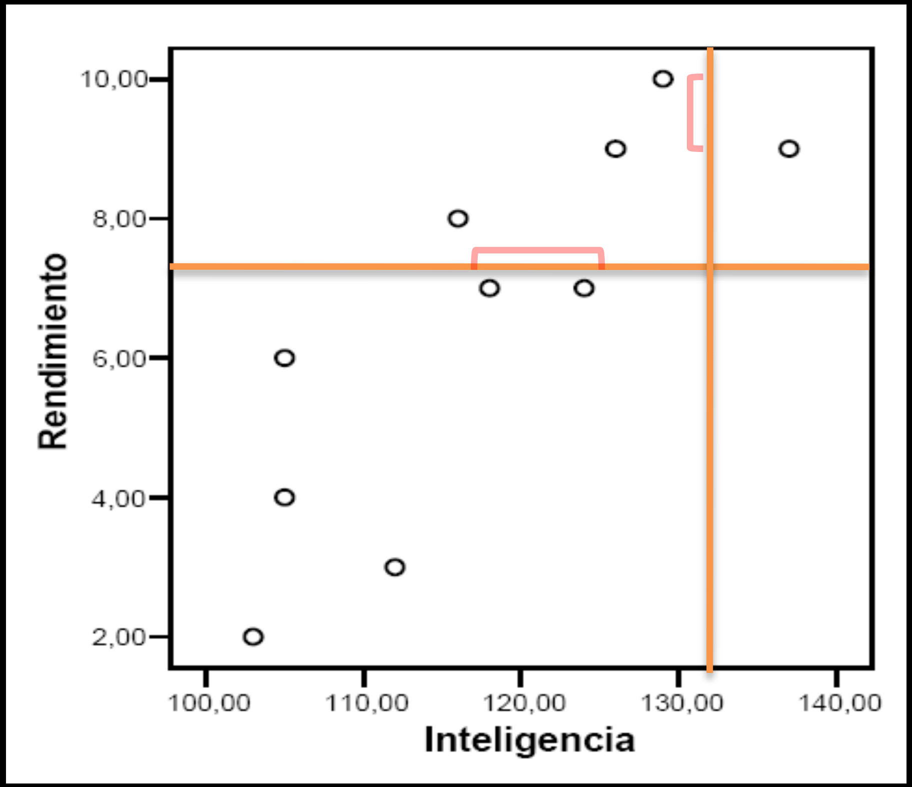


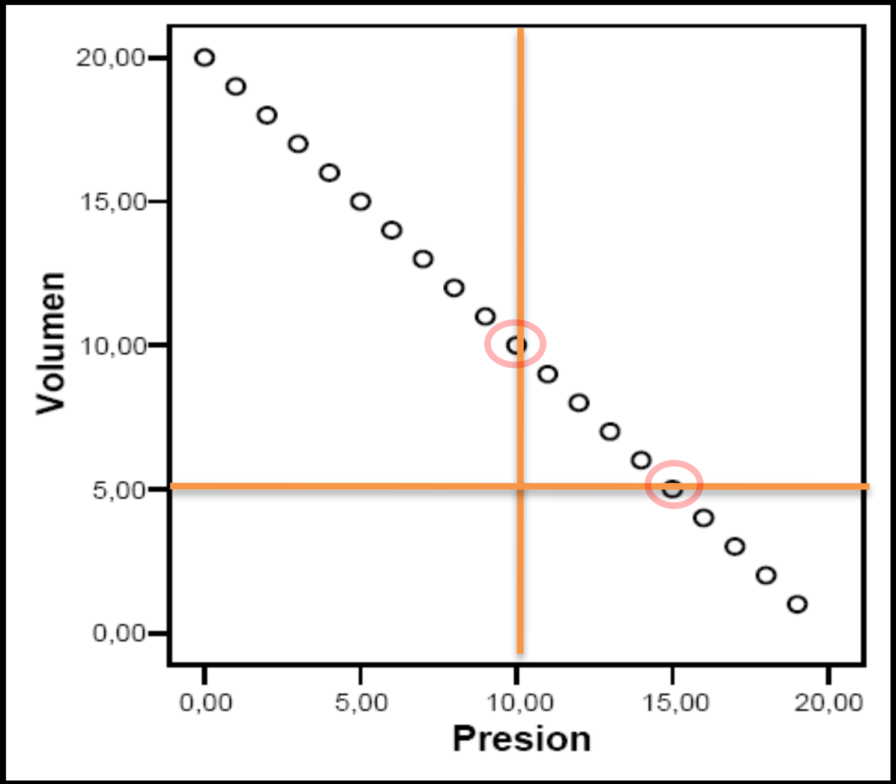


For a given value of one variable, the values of the other vary widely

In other words, large conditional variance

Associated variables

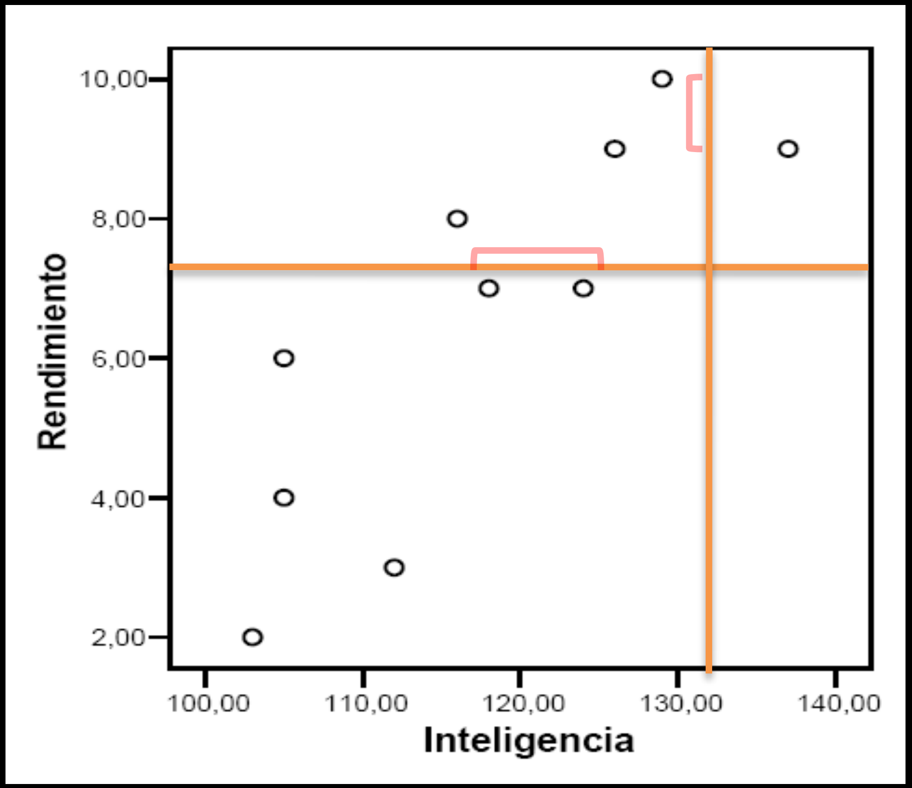


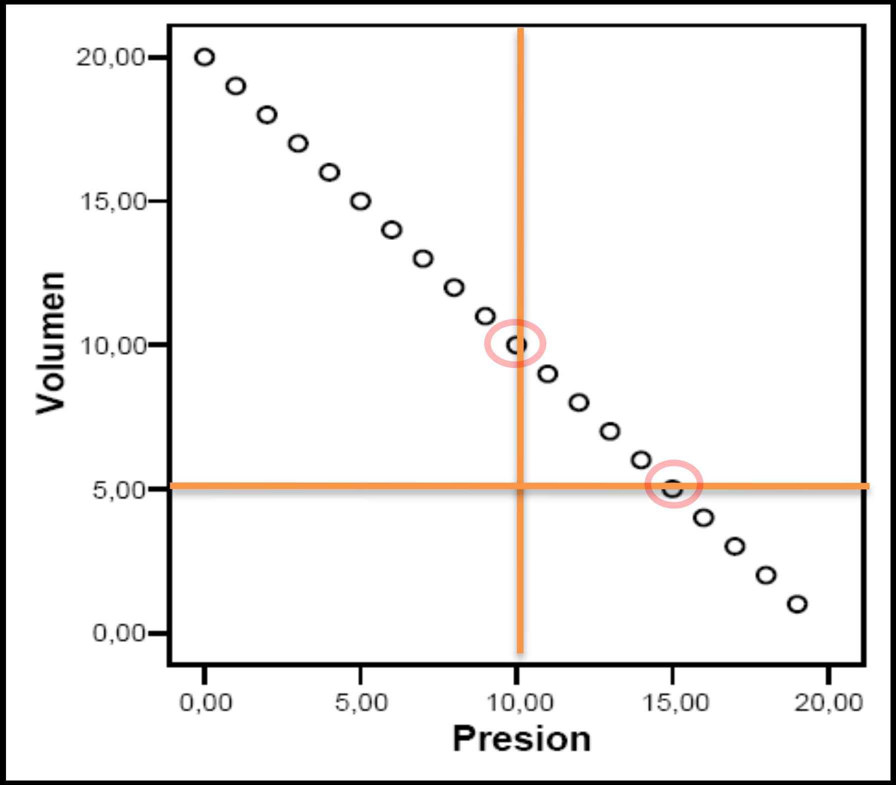


For a given value of one variable, the values of the other vary little

In other words, low conditional variance

Positively and negatively associated variables





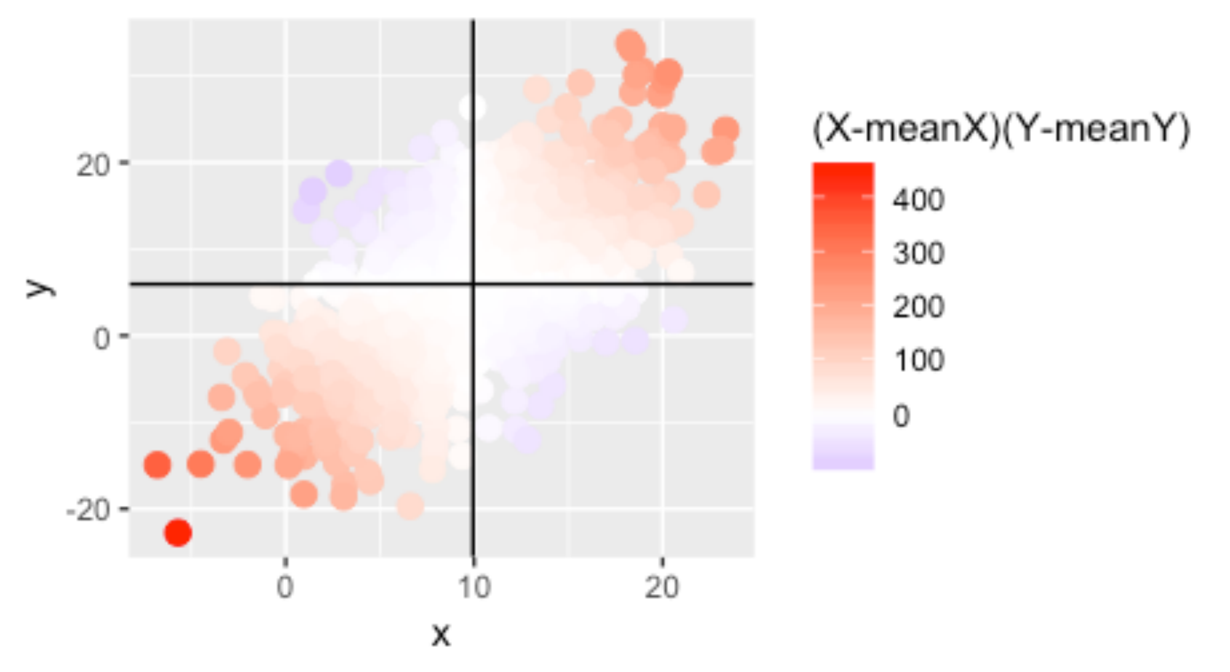
Positive association: when one variable take on a large value, the other does as well

Negative association: when one variable take on a large value, the other takes on a low value

Pearson correlation coefficient is the standard tool to measure association between continuous variables.

Built on the "covariance"

$$cov_{XY} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$



cov(xy) = 26

Problem: interpretability of covariance

- **Variables have potentially different scales (mean)**
- **Variables have potentially different dispersion (sd)**

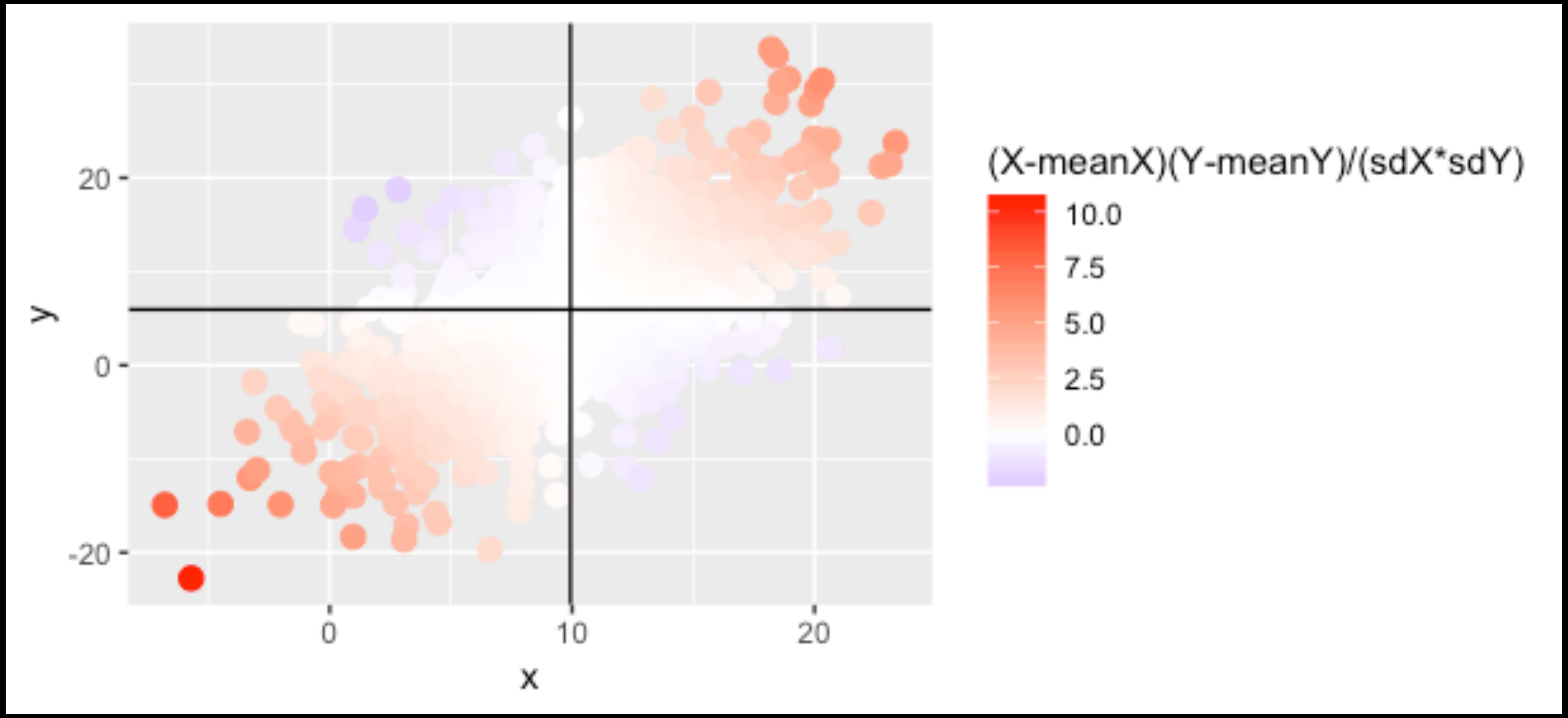
mean(x) = 10
sd(x) = 5

mean(y) = 6.1
sd(y) = 8.6

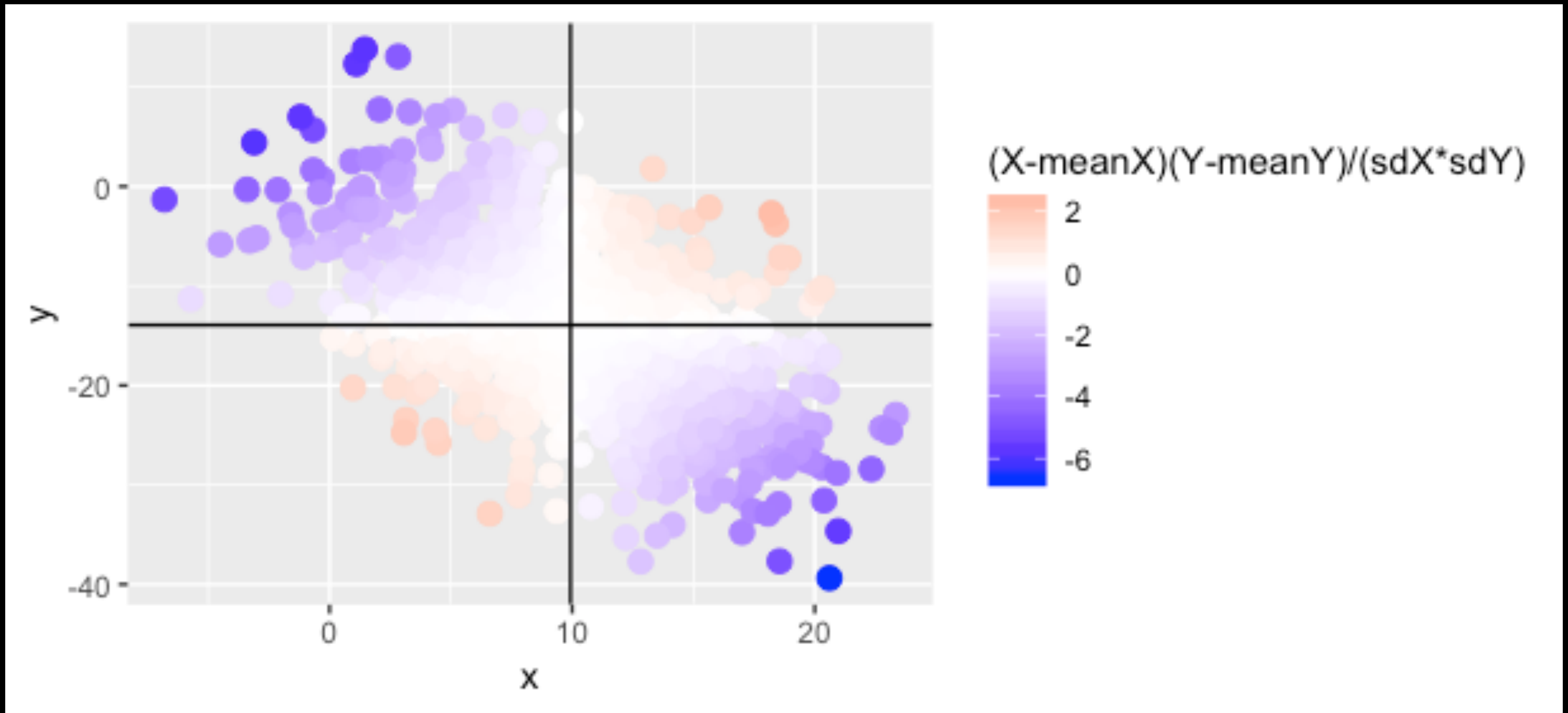
Pearson correlation coefficient is the standard tool to measure association between continuous variables

$$cov_{XY} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

$$\rho_{XY} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \equiv \frac{cov_{xy}}{\sigma_x \sigma_y}$$



$$\rho_{XY} = 0.61$$



$$\rho_{XY} = -0.54$$

$\boldsymbol{\rho}$ is bounded between **-1** and **1.**

Sign indicates the direction of the association:

- Si $\boldsymbol{\rho}$ **> 0** indicates positive association
- Si $\boldsymbol{\rho}$ **< 0** indicates negative association

It’s a measure of LINEAR association

Measures of association- discrete variables

Say 10 men and 15 women flip a coin once.
These are the results:

Contingency table:

		Y		
		Men	Women	
X	Head	3	8	11
	Tail	7	7	14
		10	15	

N=25

Is there an association between gender and the outcomes of the flipping coin game?

Contingency table:

Y

(i,j)	Men	Women	
Head	3/25= 0.12	0.32	11/25 = 0.44
Tail	0.28	0.28	0.56
	0.4	0.6	

X

N=25

The contingency table contains information about the "joint distribution" of X and Y

$P[X = Head, Y = Men] = 0.12$ .....

Contingency table:

Y

X

(i,j)	Men	Women	
Head	3/25= 0.12	0.32	11/25 = 0.44
Tail	0.28	0.28	0.56
	0.4	0.6	

N=25

Also contain the "marginal distributions"

$P[X = Head] = 0.44$ .....

$P[Y = Men] = 0.4$ .....

Contingency table:

	Y	
(i,j)	Men	Women
Head	3/25= 0.12	0.32
Tail	0.28	0.28

X

11/25 = 0.44

0.56

0.4 0.6

N=25

Also, the conditional distribution of X and Y

$P[X = x \mid Y = y]$ **And** $P[Y = \mid X = x]$

Example: $P[X = Head \mid Y = Men]$ **?**

Y

X		Men	Women	
	Head	3	8	11
	Tail	7	7	14
		10	15	

N=25

$$P[X \mid Y] = \frac{P[X, Y]}{P[Y]}$$

$$P[Head \mid Men] = \frac{3}{10}$$

$$P[Head \mid Men] = \frac{3/25}{10/25} = \frac{P[Head, Men]}{P[Men]}$$

Joint

Marginal

Conditional

Is there an association between gender and the outcomes of the flipping coin game?

If X and Y are independent (not associated), then:

$P[X|Y] = P[X]$ **<—>** $P[Y|X] = P[Y]$

$P[X|Y]$ **Y**

	Men	Women	$P[X]$
Head	0.3	0.53	**0.44**
Tail	0.7	0.47	**0.66**
	1	**1**	**1**

$P[Y|X] \neq P[Y]$

Also, If X and Y are independent (not associated), then:

$P[X, Y] = P[X]P[Y]$

If X ind of Y, joint should be:

X \ Y		Men	Women	
X	Head	0.18	0.26	0.44
	Tail	0.22	0.34	0.56
		0.4	0.6	