Understanding correlation

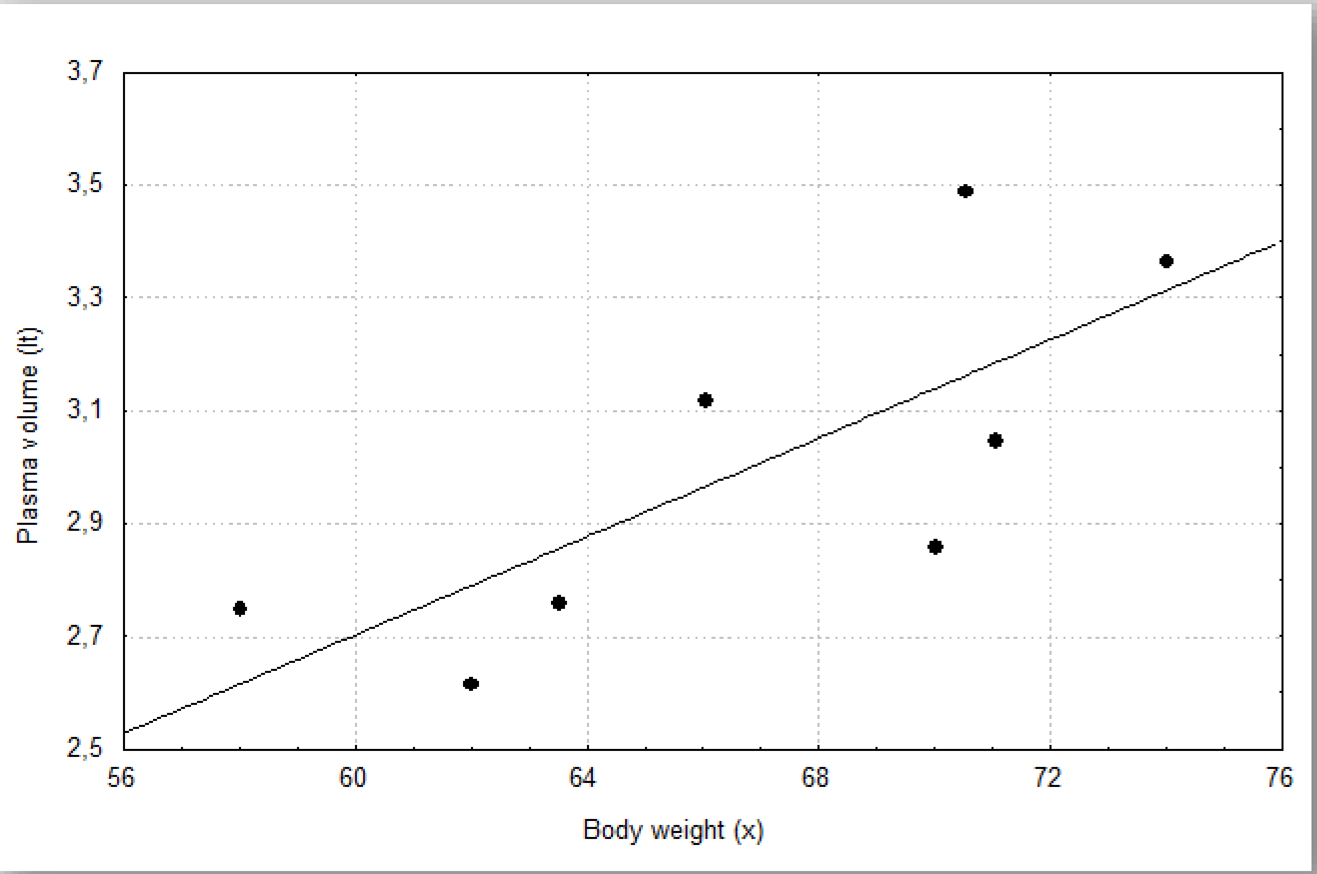
Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

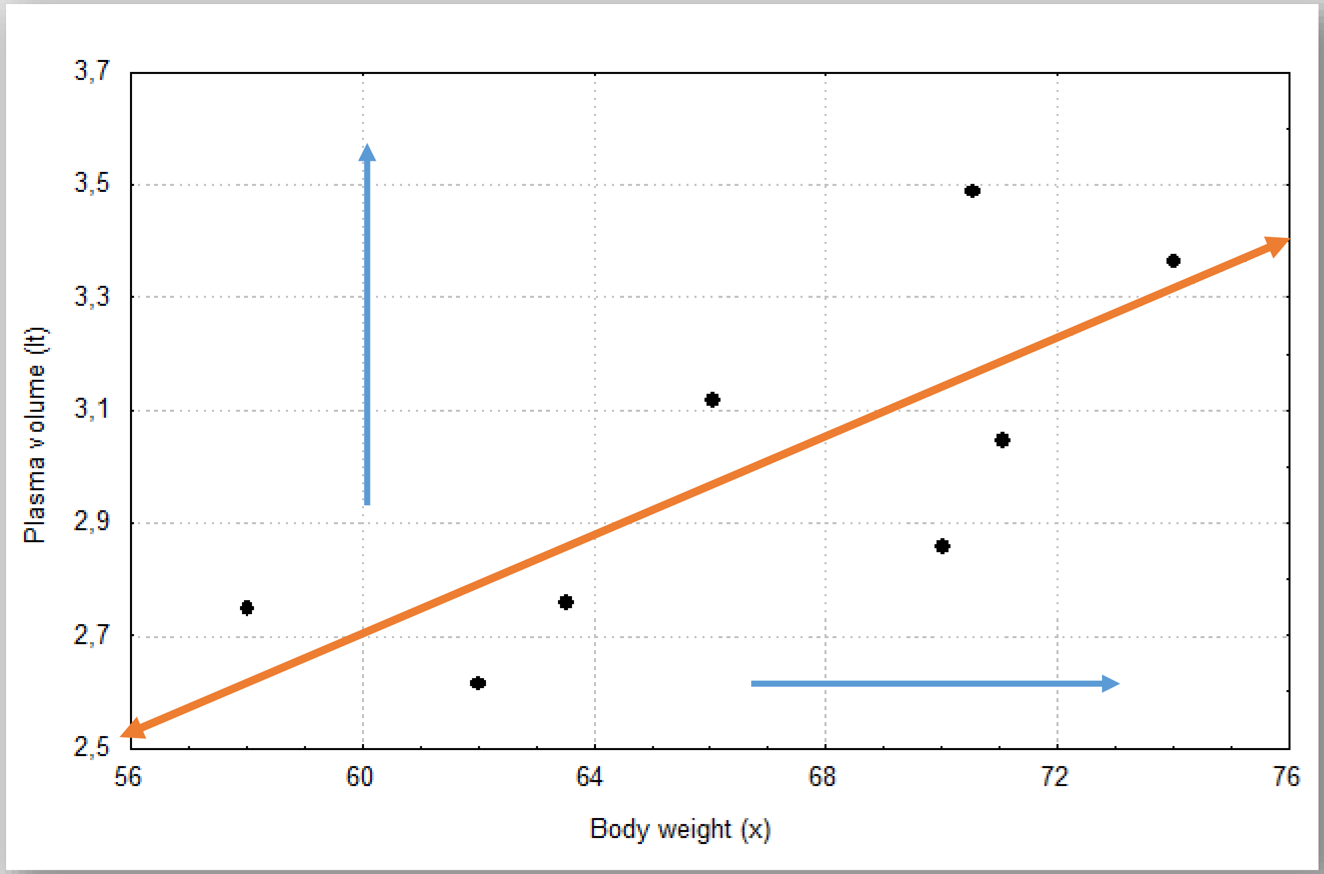
Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

The body weight and plasma volume of 8 men



Subject	Weight in Kg (x)	Plasma volume in lt (y)
1	58.0	2.75
2	70.0	2.86
3	74.0	3.37
4	63.5	2.76
5	62.0	2.62
6	70.5	3.49
7	71.0	3.05
8	66.0	3.12

The body weight and plasma volume of 8 men



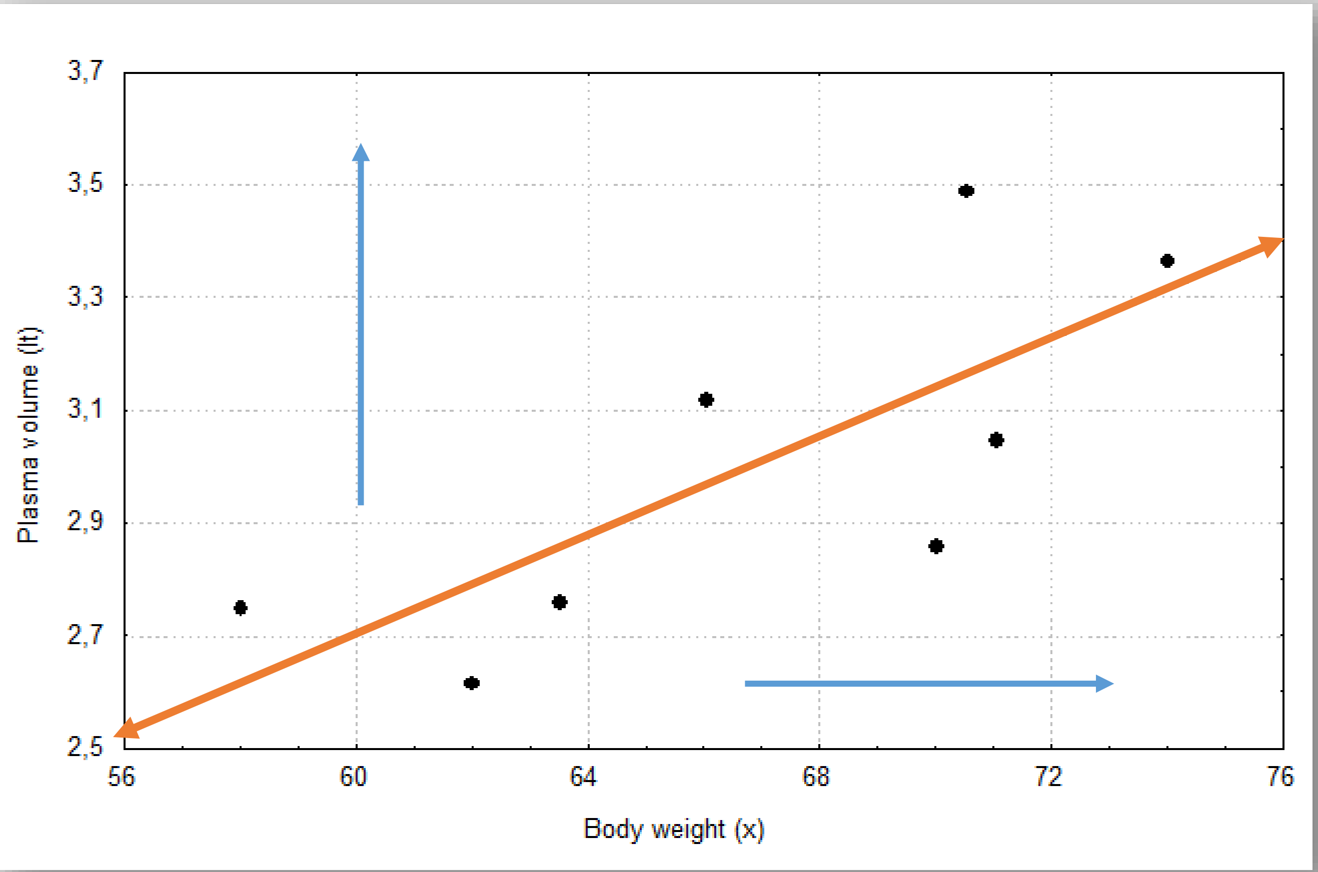
How would you describe the shape or pattern of the data points?

They seem to follow a linear pattern

When the body weight increases, what happens to plasma value?

It also increases

The body weight and plasma volume of 8 men

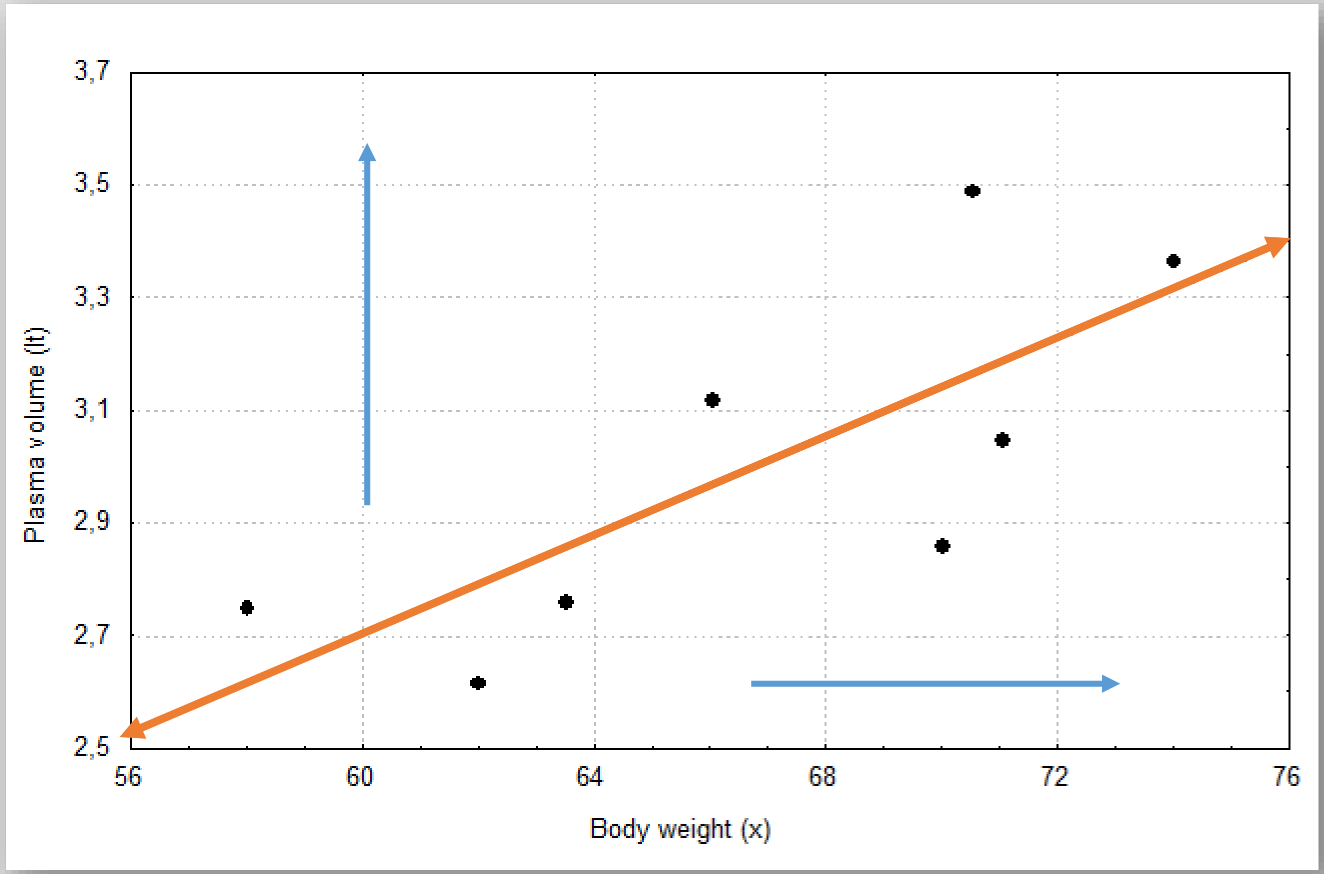


They follow a linear relationship!

We say that two variables showing this kind of pattern have a **positive linear relationship**

When one variable moves in a certain direction, the other moves in the same direction

The body weight and plasma volume of 8 men



They follow a linear relationship!

This is called the:

COVARIANCE

CO-vary

How they change together

Linear relationships

- Covariance is one of a family of statistical measures used to analyze the linear relationship between two variables
- How this two variables behave as a PAIR?

Correlation

Covariance

Linear Regression

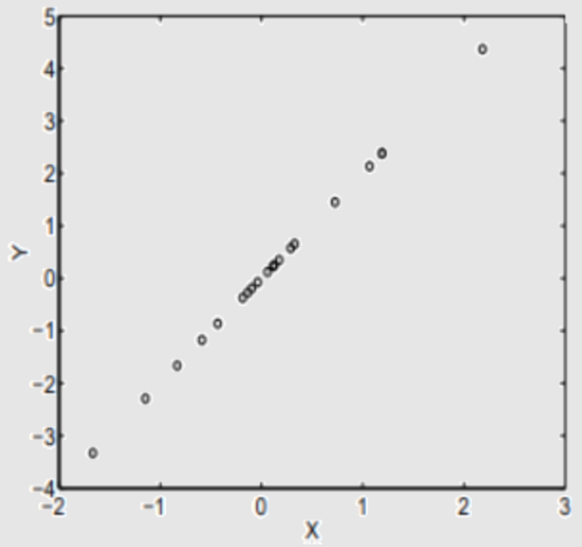
Covariance vs Correlation

- **Covariance** provides the DIRECTION (positive, negative, near zero) of the linear relationship between two variables
  - While **correlation** provides DIRECTION and STRENGTH
- **Covariance** has no upper or lower limit and its size is dependent on the measure of the variables
  - While **correlation** is always between -1 and $+1$ and its scale is independent of the scale of the variables themselves
- **Covariance** is not standardized
  - While correlation is standardized (think of $z -$ scores)

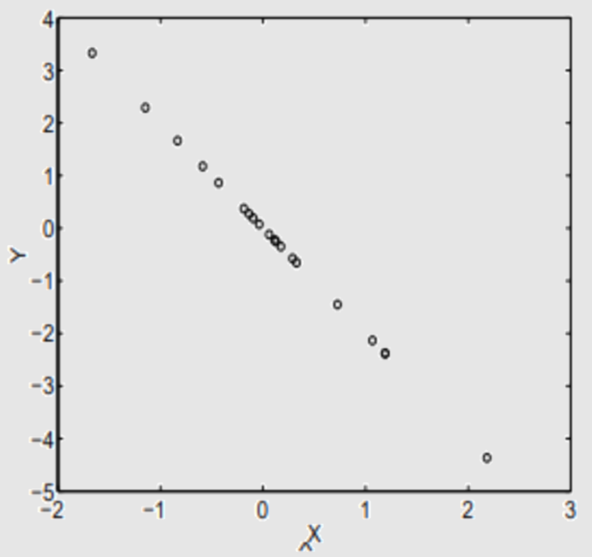
Limitations of correlation

1. Before going computing correlations look at a **scatterplot** of your data. What pattern does it show?
2. **Correlation** is only applicable to **LINEAR relationships**. There are other types of relationships that can exist between two variables
3. **Correlation** is NOT causation
    1. Correlation cannot be used to infer causation between variables
       # For example, the correlation between mood and health in individuals is less causally transparent: Does improved mood lead to better health, or does good health lead to a better mood, or both? Or is there some other factor underlying both?
4. Correlation strength does not necessarily mean the correlation is statistically significant

General Correlation Patterns

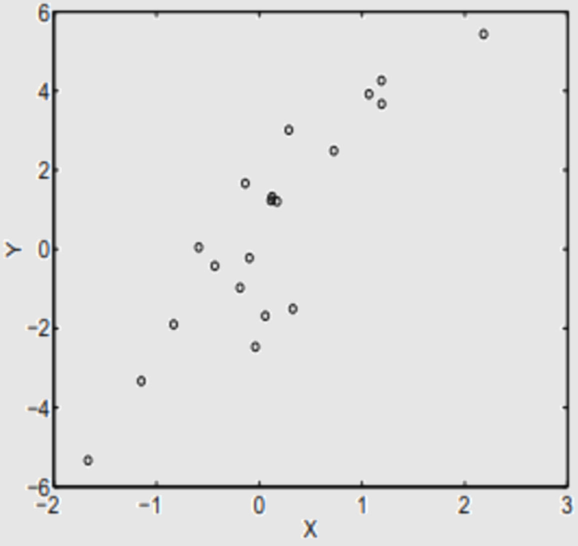


Perfect positive correlation, $+1$

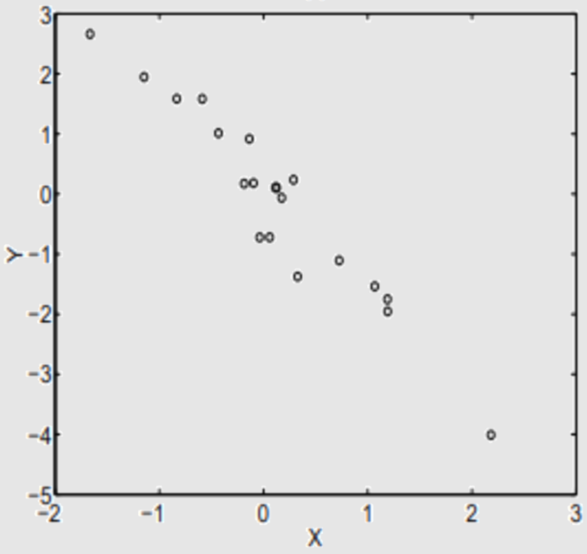


Perfect negative correlation, -1

General Correlation Patterns

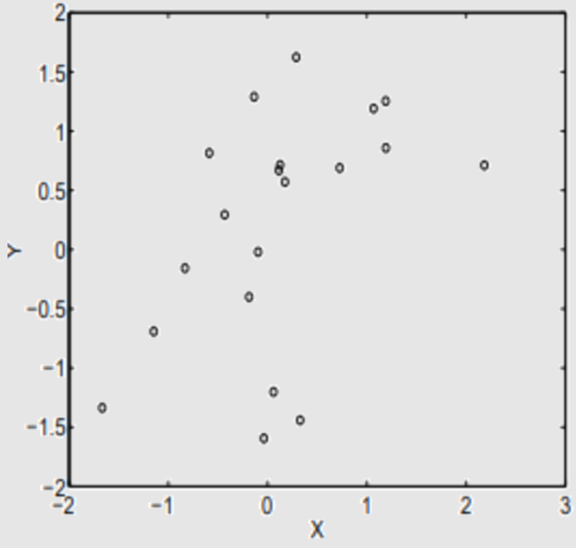


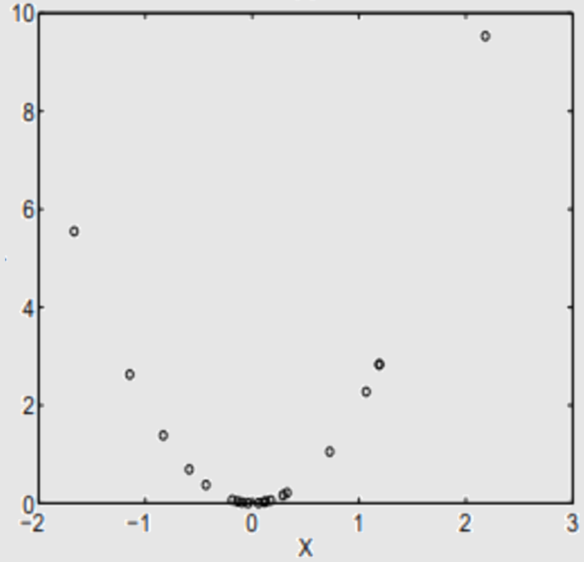
Positive correlation, near $+1$



Negative correlation, near -1

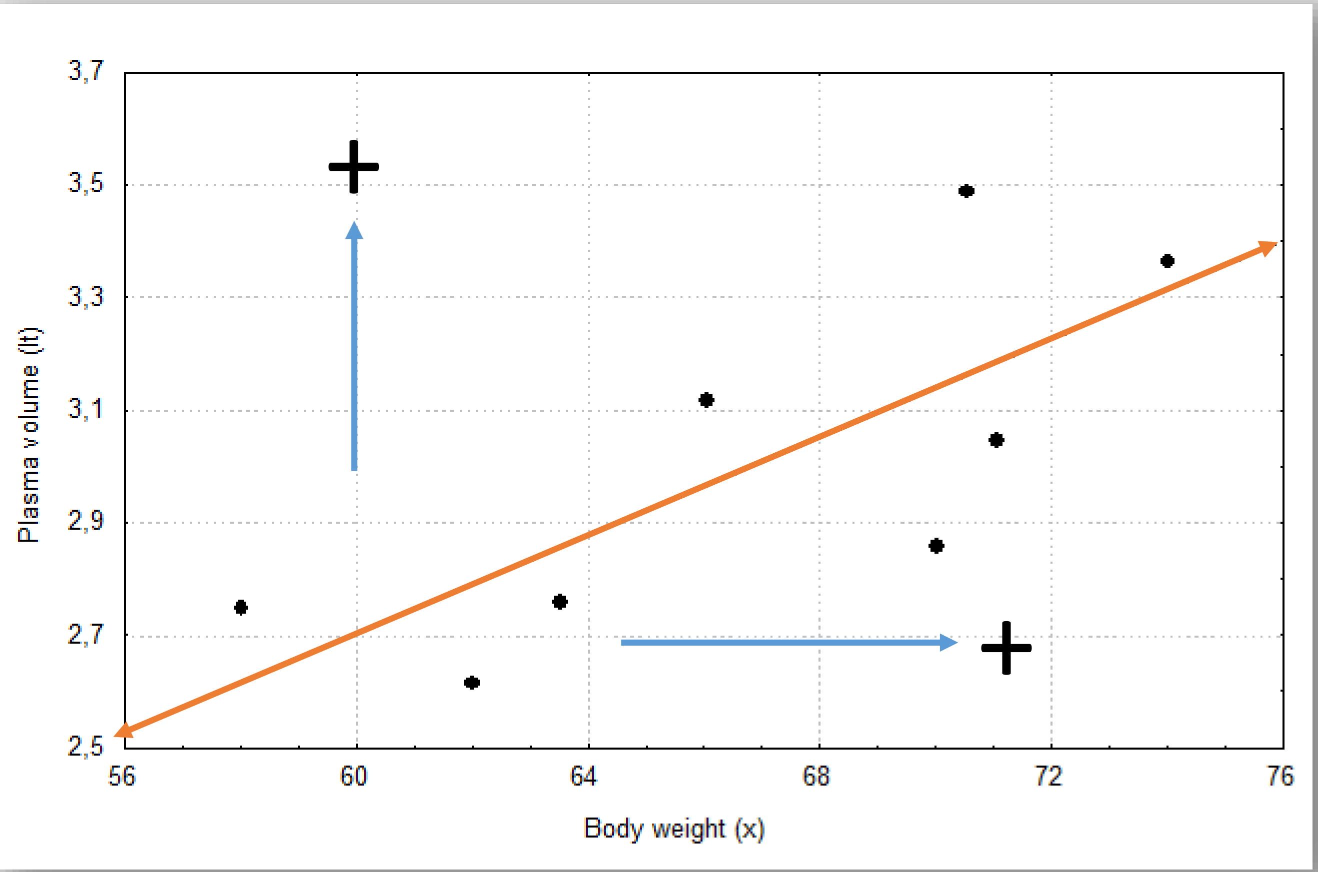
No linear relationships





Always look at a scatterplot of your data!!!

The body weight and plasma volume of 8 men



Correlations

		Body weight	Plasma volume (lt)
Body weight	Pearson Correlation	1	.759*
	Sig. (2-tailed)		0.029
	N	8	8
Plasma volume (lt)	Pearson Correlation	.759*	1
	Sig. (2-tailed)	0.029	
	N	8	8

*. Correlation is significant at the 0.05 level (2-tailed).

What is the SPSS output telling us?

$$r = 0.759$$

Correlation formula

r is called the (Pearson) correlation coefficient

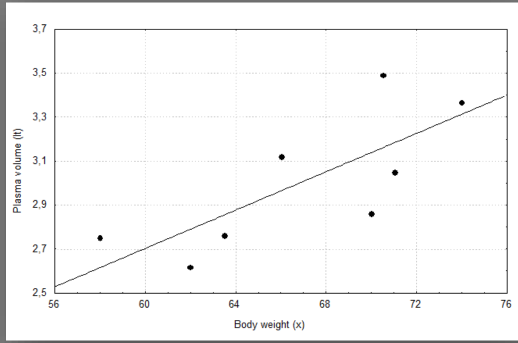
$$r = \frac{Covariance(x, y)}{Standard\ Deviation(x) \times Standard\ Deviation(y)}$$

$$r = \frac{Cov\ (x, y)}{s_x s_y}$$

1. Covariance between the two variables
2. Divided by the product of their standard deviations

Note: If you know the r and the standard deviations, you can find the covariance!

Example


Plasma volume (lt)
3,7
3,5
3,3
3,1
2,9
2,7
2,5
56
60
64
68
72
76
Body weight (x)

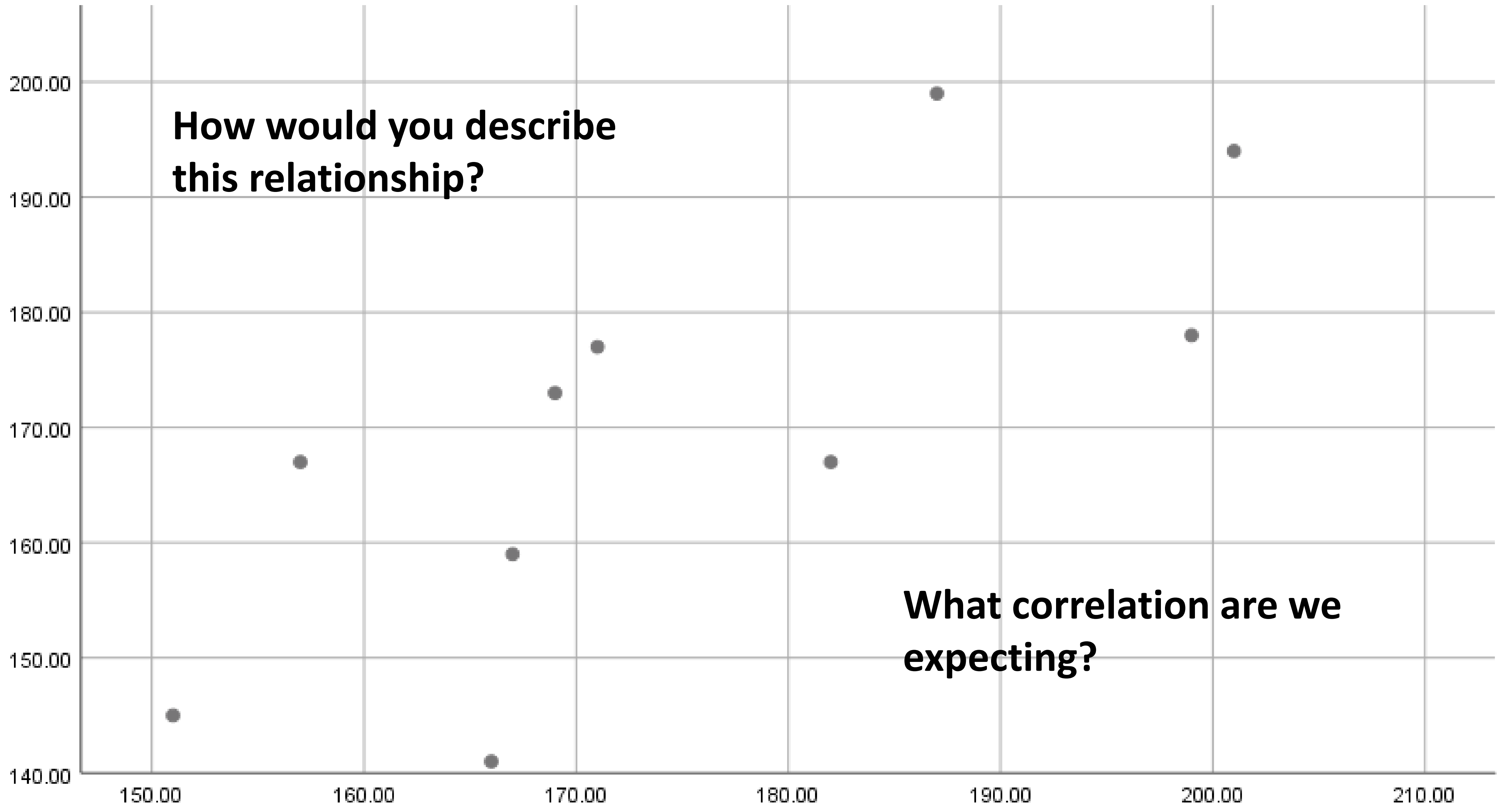
Assessing Correlation in Heights Among Married Couples

Suppose we want to study the relationship between the heights of spouses in 10 married couples.

To do this, we will take a sample of 10 measurements

$$x \;= \text{man's height}$$
$$y = \text{woman's height}$$

Male height versus female height



How would you describe this relationship?

What correlation are we expecting?

#Couple	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$		
1	151	145	-24	-25	576	625		
2	166	141	-9	-29	81	841		
3	167	159	-8	-11	64	121		
4	157	167	-18	-3	324	9		
5	187	199	12	29	144	841		
6	201	194	26	24	676	576		
7	199	178	24	8	576	64		
8	182	167	7	-3	49	9		
9	169	173	-6	3	36	9		
10	171	177	-4	7	16	49		
	$\bar{x} = 175$	$\bar{y} = 170$			$\sum = 2542$	$\sum = 3144$	$s_x = \sqrt{\frac{2542}{9}} = 16.81$	$s_y = \sqrt{\frac{3144}{9}} = 18.69$

#Couple	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	151	145	-24	-25	600
2	166	141	-9	-29	261
3	167	159	-8	-11	88
4	157	167	-18	-3	54
5	187	199	12	29	348
6	201	194	26	24	624
7	199	178	24	8	192
8	182	167	7	-3	-21
9	169	173	-6	3	-18
10	171	177	-4	7	-28
	$\bar{x} = 175$	$\bar{y} = 170$			$\sum = 2100$

$$Cov(x, y) = s_{xy} = \frac{2100}{n-1}$$

$$\frac{2100}{9}$$

$$Cov(x, y) = 233.33$$

$s_x = 16.81$ $\quad$ $s_y = 18.69$

Correlation calculation

$$r = \frac{Cov\,(x, y)}{s_x s_y}$$

$$r = \frac{s_{xy}}{s_x s_y}$$

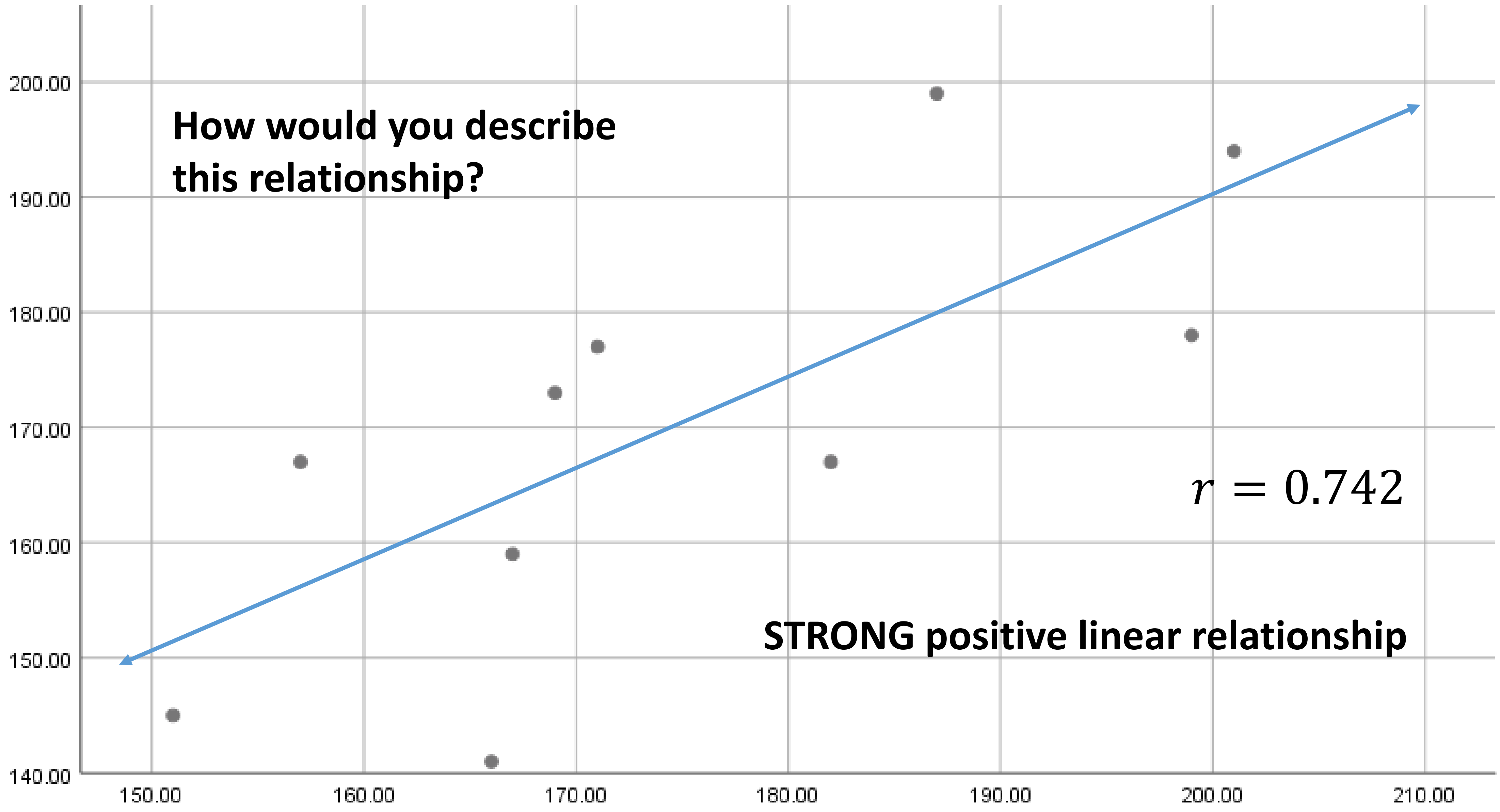
$$r = \frac{233.33}{16.81 \times 18.69}$$

$$r = \frac{233.33}{314.18}$$

$$r = 0.742$$

The result may differ slightly from the SPSS result because of rounding

Male height versus female height



Correlation coefficient r rule

- When $R = 1$, we have a **perfect positive correlation**.
- When $R = -1$, we have a **perfect negative correlation**.
- When $R = 0$, we have **no correlation at all**.
- When $0.7 < |r| < 1$, we have a **strong to very strong** correlation
- When $0.5 < |r| < 0.7$, we have **moderate to strong** correlation
- When $0.3 < |r| < 0.5$, we have a **weak to moderate** correlation
- When $R > 0$, we have a **positive correlation** between the two variables, i.e., x and y co-very move in the same direction
- When $R < 0$, we have a **negative correlation** between the two variables, i.e., x and y co-very move in opposite directions.
- The greater the absolute value of |R|, the stronger the linear relationship between the two variables

Relationship rule of thumb

How can we more objectively determine if there is a relationship between two variables?

Rule of thumb

$$\textbf{If } |r| \geq \frac{2}{\sqrt{n}}, \textbf{ then a relationship exists}$$

Relationship rule of thumb

So for our example

$$|r| \geq \frac{2}{\sqrt{10}}$$

$$0.742 \geq 0.632$$

The value $\mathbf{0.632}$ is the threshold for the rule

So, a relationship exists

Significance

To determine if the correlation coefficient r … rent from zero, we use the following $t - test$:

$$t = \frac{r\surd}{\sqrt{1}}$$

	Percentace points of the t distribution		
	p-value		
df(=n-1)	0.05	0.01	0.001
1	12.71	63.66	636.62
2	4.3	9.92	31.6
3	3.18	5.84	12.92
4	2.78	4.6	8.61
5	2.57	4.03	6.87
6	2.45	3.71	5.96
7	2.36	3.5	5.41
8	2.31	3.36	5.04
9	2.26	3.25	4.78
10	2.23	3.17	4.59
11	2.2	3.11	4.44
12	2.18	3.05	4.32
13	2.16	3.01	4.22
14	2.14	2.98	4.14
15	2.13	2.95	4.07
20	2.09	2.85	3.85
30	2.04	2.75	3.65
40	2.02	2.7	3.55
120	1.98	2.62	3.37
∞	1.96	2.58	3.29

If the absolute value of t is greater than the … th $n - 2$ degrees of freedom (where n is the numbe… t r is statistically significant

$t - test$ calculation

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

$$t = \frac{2.0984}{0.670}$$

$$t = \frac{0.742\sqrt{10-2}}{\sqrt{1-0.742^2}}$$

$$t = 3.139$$

$$t = \frac{0.742 \times 2.828}{\sqrt{0.449}}$$

The t value of 3.139 is greater than the 5% point of the $t-$distribution with $8\ (n-2)$ degrees of freedom, which is 2.31.

Thus, the correlation coefficient r is statistically significant.

Alternatively, there is a statistically significant correlation $(p < 0.05)$ between men's height and women's height.

Review

- The **Pearson correlation coefficient** detects the linear relationship between two quantitative variables, applicable to both continuous variables and discrete numerical values (e.g., families with one child, families with two children, families with three children)
- It is used to complete the **scatter plot**
- The correlation coefficient is **parametric**, meaning it assumes that the values come from normally distributed populations
- Normality can be assessed using a histogram
- If normality is not met, or if the variables have relatively few distinct values, it is better to calculate the **Spearman coefficient**, which is the non-parametric equivalent of the Pearson coefficient