Pearson Product-Moment Correlation or Pearson correlation coefficient

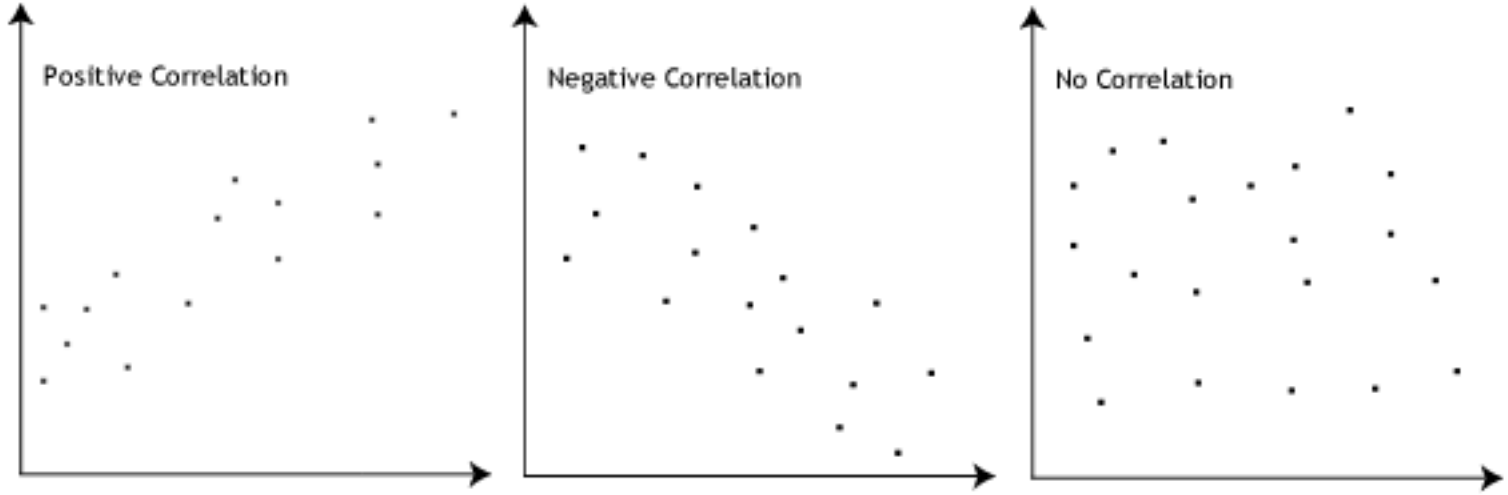
- The Pearson product-moment correlation coefficient is a measure of the strength of a linear association between two variables and is denoted by r.

- A Pearson product-moment correlation attempts to draw a line of best fit through the data of two variables, and the Pearson correlation coefficient, r, indicates how far away all these data points are to this line of best fit

    - how well the data points fit this new model/line of best fit.

What values can the Pearson correlation coefficient take?

- The Pearson correlation coefficient, r, can take a range of values from +1 to -1.
  - A value of 0 indicates that there is no association between the two variables.
  - A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable.
  - A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases.
- This is shown in the diagram below:



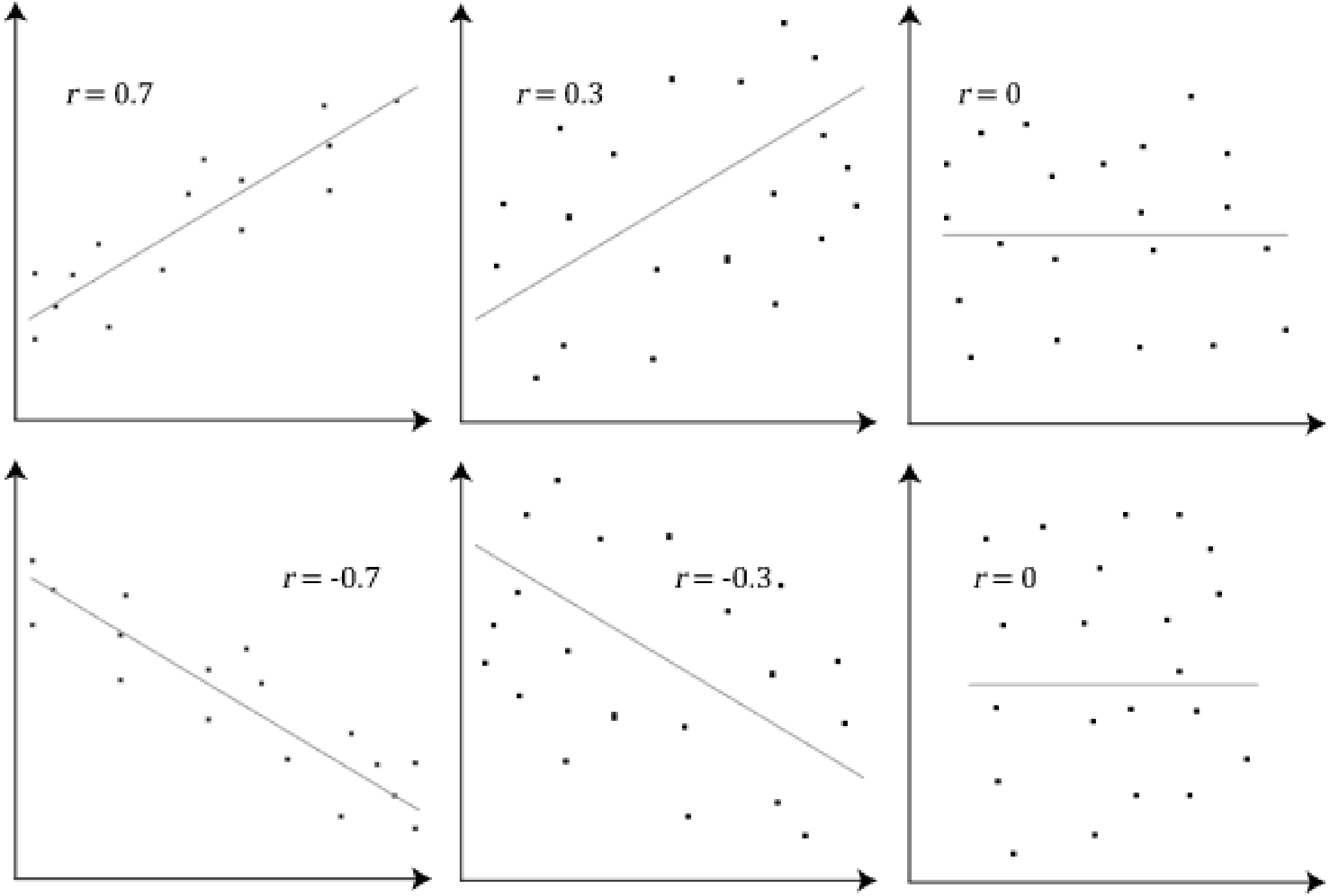
How can we determine the strength of association based on the Pearson correlation coefficient?

- The stronger the association of the two variables, the closer the Pearson correlation coefficient, r, to either +1 (relationship is positive) or -1 (relationship is negative).

- Achieving a value of +1 or -1 means that all your data points are included on the line of best fit – there are no data points that show any variation away from this line.

- Values for r between +1 and -1 (for example, r = 0.8 or -0.4) indicate that there is variation around the line of best fit.

- The closer the value of r to 0 the greater the variation around the line of best fit.


r = 0.7
r = 0.3
r = 0
r = -0.7
r = -0.3
r = 0

Are there guidelines to interpreting Pearson's correlation coefficient?

The following guidelines have been proposed:

	Coefficient, *r*	
Strength of Association	Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to -1.0

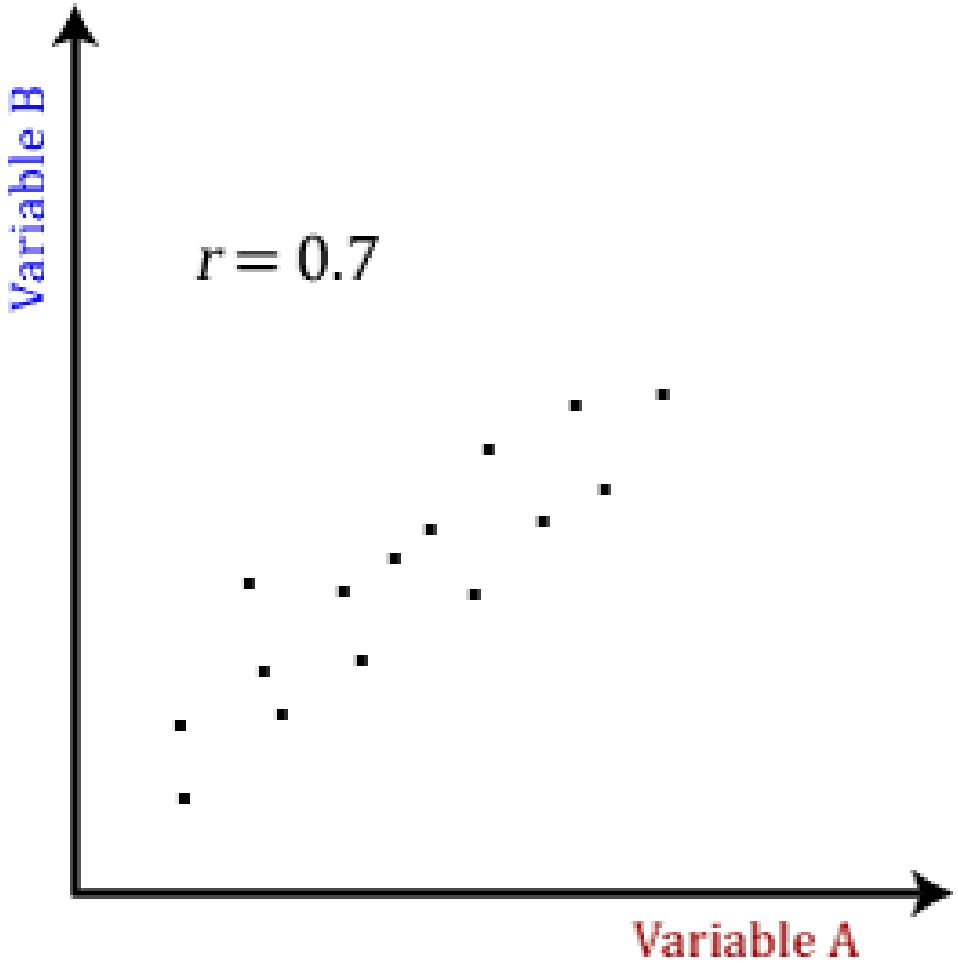
Do the two variables have to be measured in the same units?

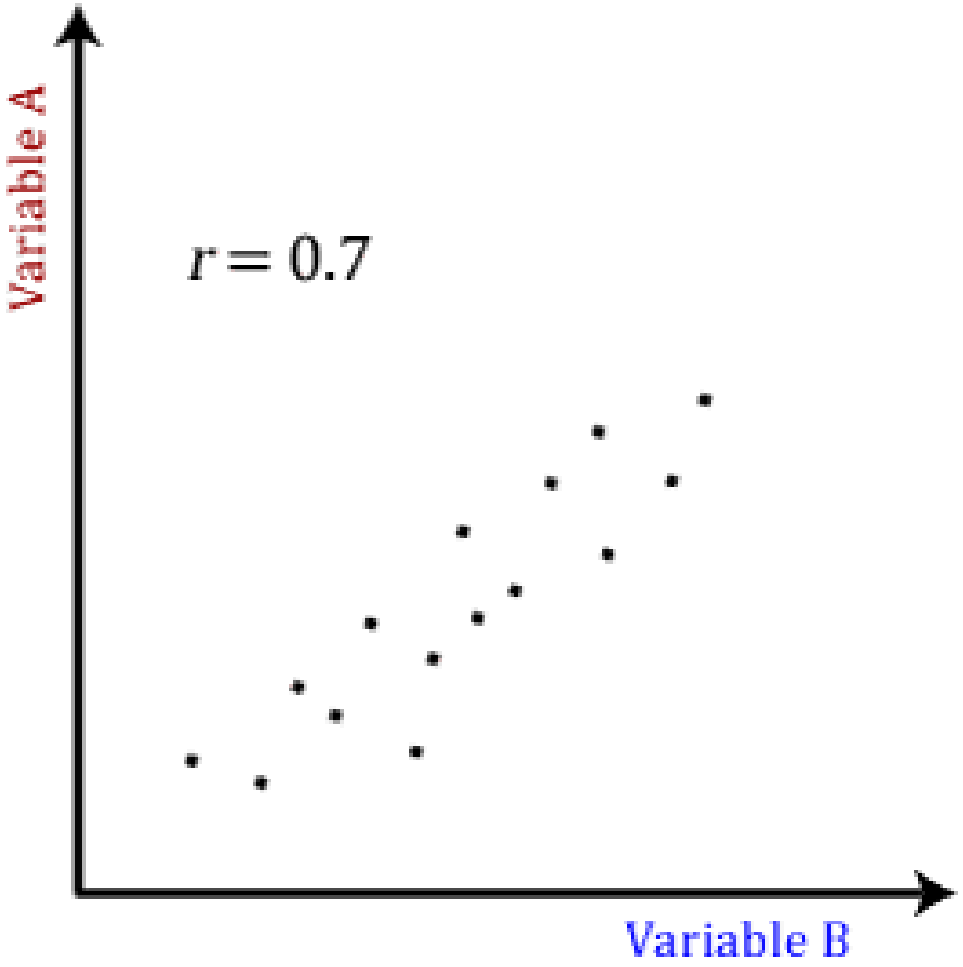
- The two variables can be measured in entirely different units.
- For example, you could correlate a person's age with their blood sugar levels.
- Here, the units are completely different; age is measured in years and blood sugar level measured in mmol/L (a measure of concentration).

What about dependent and independent variables?

- The Pearson product-moment correlation does not take into consideration whether a variable has been classified as a dependent or independent variable.
- **For example**, you might want to find out whether basketball performance is correlated to a person's height.
  - You might, therefore, plot a graph of performance against height and calculate the Pearson correlation coefficient.
  - Lets say, for example, that r = .67. That is, as height increases so does basketball performance. This makes sense.
  - However, if we plot a graph of height against performance to determine whether a person's height was determined by their basketball performance (which makes no sense), we would still get r = .67.
  - This is because the Pearson correlation coefficient makes no account of any theory behind why you chose the two variables to compare.

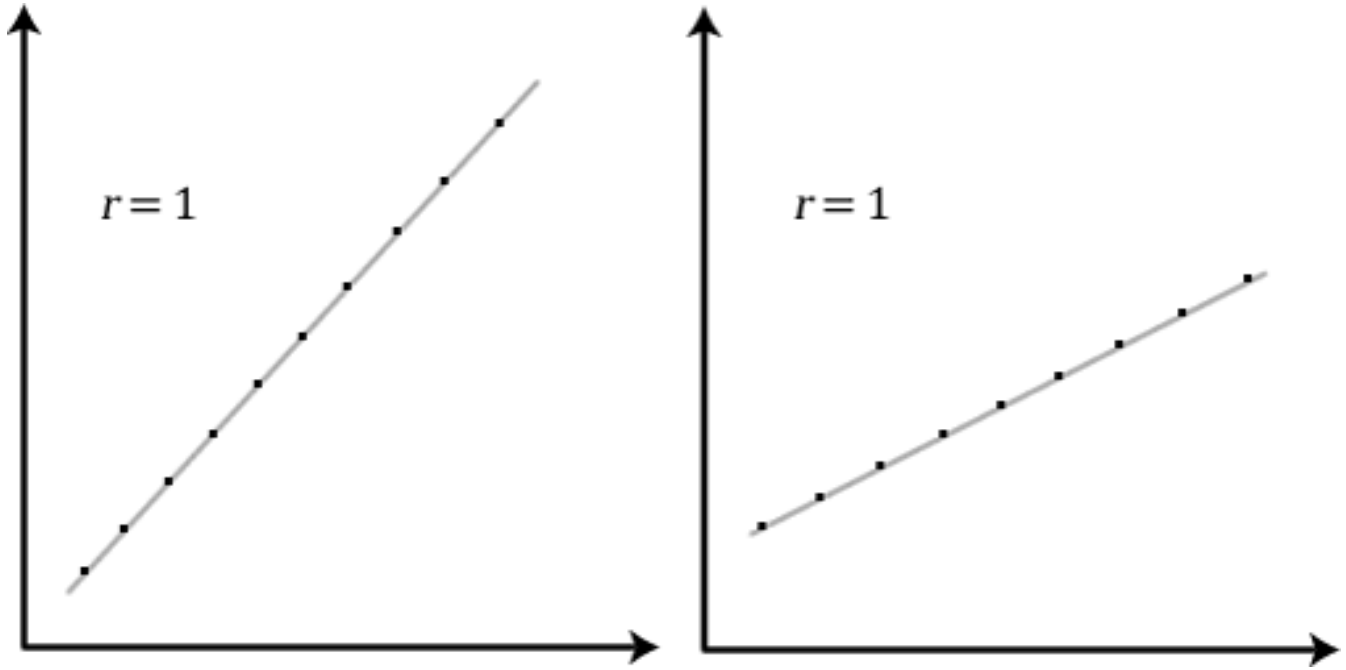
What about dependent and independent variables?





Does the Pearson correlation coefficient indicate the slope of the line?

- Pearson correlation coefficient, r, does not represent the slope of the line of best fit.
- Therefore, if you get a Pearson correlation coefficient of +1 this does not mean that for every unit increase in one variable there is a unit increase in another.
- It simply means that there is no variation between the data points and the line of best fit.



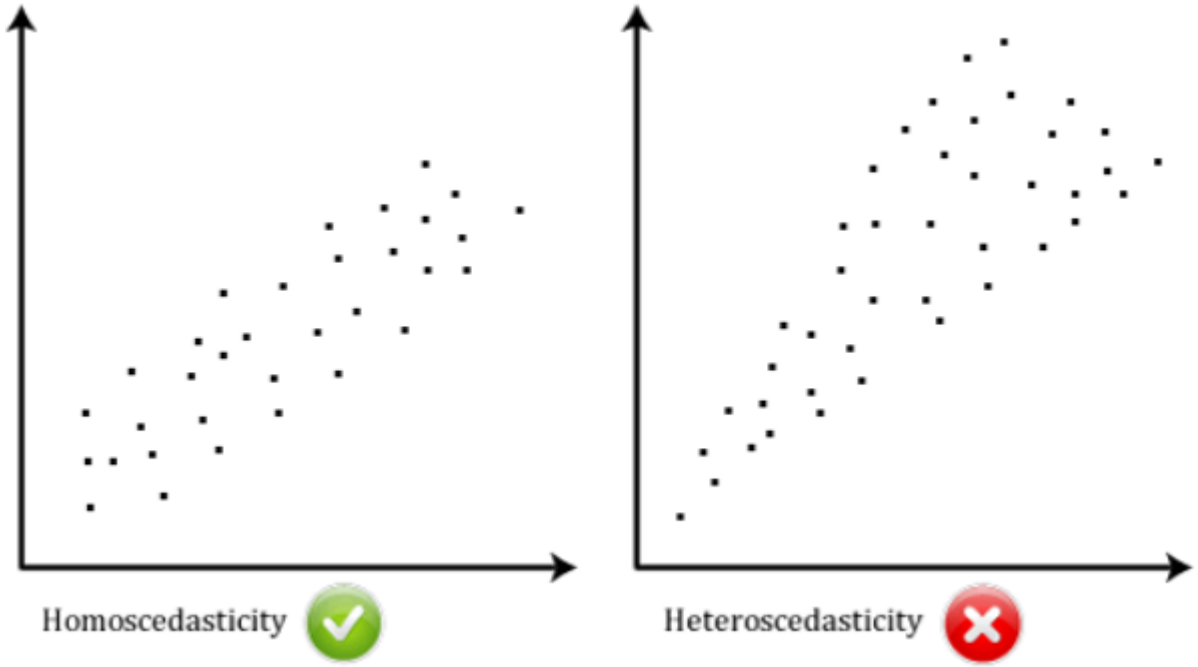
Pearson's correlation determines the degree to which a relationship is linear.

What assumptions does Pearson's correlation make?

- There are five assumptions that are made with respect to Pearson's correlation:
  - The variables must be either interval or ratio measurements (difference between 20°C and 30°C is the same as 30°C to 40°C is an interval variable, where as a distance of 10 meters is twice the distance of 5 meters is ratio variable)
  - The variables must be approximately normally distributed
  - There is a linear relationship between the two variables
  - Outliers are either kept to a minimum or are removed entirely.
  - There is homoscedasticity of the data.

What is homoscedasticity?

- Homoscedasticity basically means that the variances along the line of best fit remain similar as you move along the line.
- It is required that your data show homoscedasticity for you to run a Pearson product-moment correlation.
- Homoscedasticity is most easily demonstrated diagrammatically as below:



Pearson’s correlation: Real Life Example

- Scientists in China wanted to know if there was a relationship between how weedy rice populations are different genetically.
- The goal was to find out the evolutionary potential of the rice.
- Pearson’s correlation between the two groups was analyzed.
- It showed a positive Pearson Product Moment correlation of between 0.783 and 0.895 for weedy rice populations.
- This figure is quite high, which suggested a fairly strong relationship.

Pearson's correlation: Real Life Example

Sample question: Find the value of the correlation coefficient from the following table:

Subject	Age x	Glucose Level y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Pearson's correlation: Real Life Example

1. Use the given data, and add three more columns: xy, x^2, and y^2.
2. Multiply x and y together to fill the xy column.
3. Take the square of the numbers in the x column, and put the result in the x^2 column.
4. Take the square of the numbers in the y column, and put the result in the y^2 column.
5. Add up all of the numbers in the columns and put the result at the bottom of the column.

Pearson's correlation: Real Life Example

Subject	Age x	Glucose Level y	xy	x^2	y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Use the following correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,][\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

= 6(20,485) – (247 × 486) /
[√[[6(11,409) – (2472)] × [6(40,022) – 4862]]]

The answer is: **2868 / 5413.27 = 0.529809**

which means the variables have a moderate positive correlation.

Spearman's rank-order correlation

When should you use the Spearman's rank-order correlation?

Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars).

- The Spearman's rank-order correlation is the nonparametric version of the Pearson product-moment correlation.
- Spearman's correlation coefficient, (ρ, also signified by rs) measures the strength and direction of association between two ranked variables.

Nonparametric statistics refer to a statistical method in which the data is not required to fit a normal distribution.

It uses data that is often ordinal, meaning it does not rely on numbers, but rather on a ranking or order of sorts.

For example, a survey conveying consumer preferences ranging from like to dislike would be considered ordinal data.

Parametric statistics includes parameters such as the mean, median, standard deviation, variance, etc.
They uses the observed data to estimate the parameters of the distribution.
Under parametric statistics, data is assumed to fit a normal distribution with unknown parameters μ (population mean) and σ2 (population variance), which are then estimated using the sample mean and sample variance.

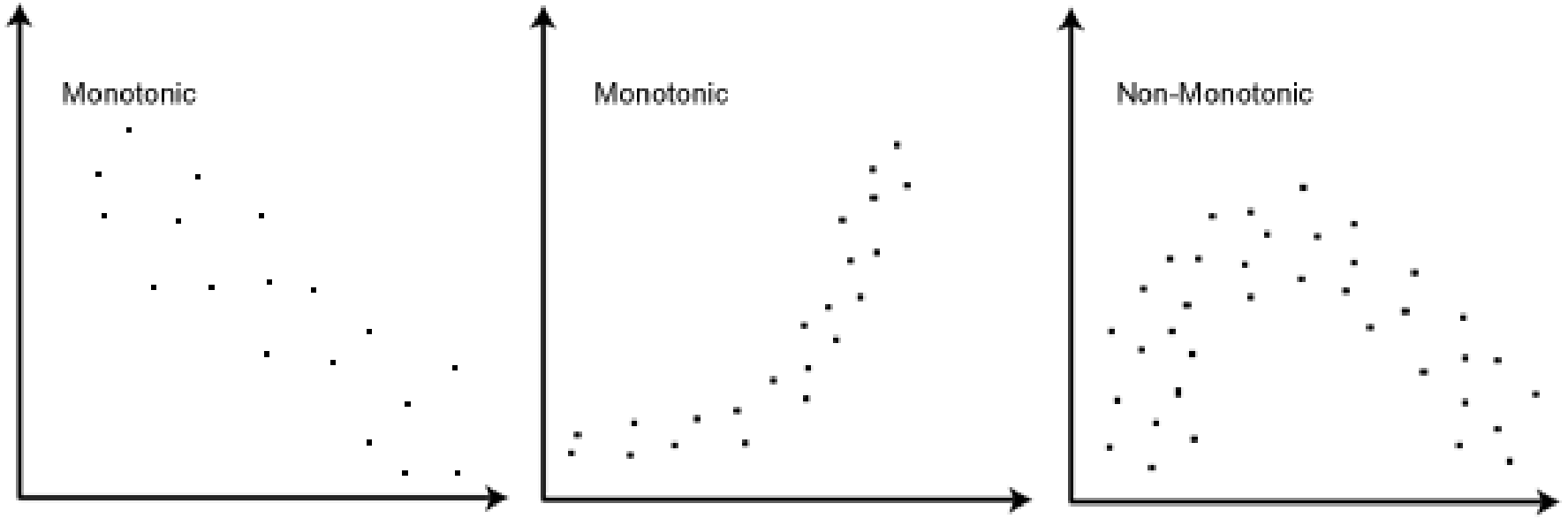
What are the assumptions of the test?

- You need two variables that are either ordinal, interval or ratio.

- Spearman's correlation determines the strength and direction of the monotonic relationship between the two variables

What is a monotonic relationship?

- A monotonic relationship is a relationship that does one of the following: (1) as the value of one variable increases, so does the value of the other variable; or (2) as the value of one variable increases, the other variable value decreases.



Monotonicity is "less restrictive" than that of a linear relationship.
For example, the middle image above shows a relationship that is monotonic, but not linear.

Why is a monotonic relationship important to Spearman's correlation?

- A monotonic relationship is not strictly an assumption of Spearman's correlation.
  - You can run a Spearman's correlation on a non-monotonic relationship to determine if there is a monotonic component to the association.

- If a scatterplot shows that the relationship between your two variables looks monotonic you would run a Spearman's correlation because this will then measure the strength and direction of this monotonic relationship.

- On the other hand if, for example, the relationship appears linear (assessed via scatterplot) you would run a Pearson's correlation because this will measure the strength and direction of any linear relationship.

Spearman's Rank-Order Correlation

- An example of calculating Spearman's correlation
- To calculate a Spearman rank-order correlation on data without any ties we will use the following data:

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

We then complete the following table:

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

where n = 10.
We have calculated a ρ (or r_s) of 0.67.
This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam.
That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Spearman's Rank-Order Correlation

- The formula to use when there are tied ranks is:
- Spearman Formula

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

- where i = paired score.

How do you report a Spearman's correlation?

- If you have simply run the Spearman correlation without any statistical significance tests, you are able to simple state the value of the coefficient as shown below:

$$\rho = coefficient \quad OR \quad r_s = coefficient$$

e.g.

$$\rho = 0.67 \quad OR \quad r_s = 0.67$$

- However, if you have also run statistical significance tests, you need to include some more information as shown below:

$$\rho(df) = \rho\ coefficient, P = P\ value$$

$$\rho(8) = 0.67, P = 0.033$$

where df = N – 2, where N = number of pairwise cases.

How do you express the null hypothesis for this test?

The general form of a null hypothesis for a Spearman correlation is:

- H_0: There is no [monotonic] association between the two variables [in the population].

A null hypothesis statement for the example used earlier in this guide would be:

- H_0: There is no [monotonic] association between maths and English marks.

How do I interpret a statistically significant Spearman correlation?

- The statistical significance testing of the Spearman correlation does not provide you with *any* information about the strength of the relationship.
- Achieving a value of $p = 0.001$, for example, does not mean that the relationship is stronger than if you achieved a value of $p = 0.04$.
- This is because the significance test is investigating whether you can reject or fail to reject the null hypothesis.

- If you set $\alpha = 0.05$, achieving a statistically significant Spearman rank-order correlation means that you can be sure that there is less than a 5% chance that the strength of the relationship you found (your ρ coefficient) happened by chance if the null hypothesis were true.

Kendall’s Tau?

- Kendall’s Tau is a non-parametric measure of relationships between columns of ranked data. The Tau correlation coefficient returns a value of 0 to 1, where:

- 0 is no relationship,
- 1 is a perfect relationship.
- Kendall's rank correlation provides a distribution free test of independence and a measure of the strength of dependence between two variables.

- Spearman's rank correlation is satisfactory for testing a null hypothesis of independence between two variables but it is difficult to interpret when the null hypothesis is rejected.
- Kendall's rank correlation improves upon this by reflecting the strength of the dependence between the variables being compared.

Kendall’s Tau?

- For example, you could use Kendall's tau-b to understand whether there is an association between exam grades (A, B, C, D, E and F) and time spent revising (less than 5 hours, 5-9 hours, 10-14 hours, 15-19 hours, and 20 hours or more).

- Alternately, you could use Kendall's tau-b to understand whether there is an association between customer satisfaction (i.e., where the level of agreement had five categories: strongly agree, agree, neither agree nor disagree, disagree and strongly disagree) and delivery time (i.e., where delivery time had four categories – next day, 2 working days, 3-5 working days, and more than 5 working days)

Assumptions in the Kendall's Tau

- Assumption #1: Your two variables should be measured on an ordinal or continuous scale.
- Examples of ordinal variables include Likert scales (e.g., a 7-point scale from strongly agree through to strongly disagree), amongst other ways of ranking categories (e.g., a 5-point scale explaining how much a customer liked a product, ranging from "Not very much" to "Yes, a lot").
- Examples of continuous variables (i.e., interval or ratio variables) include revision time (measured in hours), intelligence (measured using IQ score), exam performance (measured from 0 to 100), weight (measured in kg), and so forth.

- Assumption #2: Kendall's tau-b determines whether there is a monotonic relationship between your two variables.

Kendall's Tau

- Most statistical packages have Tau-B built in, but you can use the following formula to calculate it by hand:

- Kendall's Tau = (C – D / C + D)

Example Problem on Kendall's Tau

- Sample Question: Two interviewers ranked 12 candidates (A through L) for a position. The results from most preferred to least preferred are:

- Interviewer 1: ABCDEFGHIJKL.
- Interviewer 2: ABDCFEHGJILK.

- Calculate the Kendall Tau correlation.

Kendall's Tau

Make a table of rankings.

Candidate	Interviewer 1	Interviewer 2
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11

Count the number of concordant pairs:

Using the second column. Concordant pairs are how many larger ranks are below a certain rank. For example, the first rank in the second interviewer's column is a "1", so all 11 ranks below it are larger.

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	
B	2	2		
C	3	4		
D	4	3		
E	5	6		
F	6	5		
G	7	8		
H	8	7		
I	9	10		
J	10	9		
K	11	12		
L	12	11		

Kendall's Tau

When all concordant pairs have been counted, it looks like this:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	
B	2	2	10	
C	3	4	8	
D	4	3	8	
E	5	6	6	
F	6	5	6	
G	7	8	4	
H	8	7	4	
I	9	10	2	
J	10	9	2	
K	11	12	0	
L	12	11		

Kendall's Tau

- **Count the number of discordant pairs** and insert them into the next column. The number of discordant pairs is similar to Step 2, only you're looking for smaller ranks, not larger ones.

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	3	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11		

Kendall's Tau

- Sum the values in the two columns:

Candidate	Interviewer 1	Interviewer 2	Concordant	Discordant
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	3	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11		
		Totals	61	5

Insert the totals into the formula:
Kendall's Tau = (C – D / C + D)
= (61 – 5) / (61 + 5) = 56 / 66 = .85.

The Tau coefficient is .85, suggesting a strong relationship between the rankings.

Kendall's Tau

- If you want to calculate statistical significance for your result, use this formula to get a z-value:

$$z = \frac{3 * T\sqrt{N(N-1)}}{\sqrt{2(2N+5)}}$$

- Inserting the values from our results:

$$z = \frac{3 * 85\sqrt{12(12-1)}}{\sqrt{2(2(12)+5)}}$$

= 3 * .85 * 11.489 / 7.616

= 3.85.

- Finding the area for a z-score of 3.85 on a z-table gives an area of .0001 — a tiny probability value which tells you this result is statistically significant.