CORRELATION

Pearson & Spearman Correlation

Dr Asrul Anuar Zulkifli

M.D (UKM), MscMPH (MSU),

Micro credential Digital Health (Monash)

SCAN ME

Course Materials

01 Overview

02 Pearson Correlation

- ❑ Definition & Key Concepts
- ❑ Assumptions
- ❑ Pearson Correlation Coefficient, r
- ❑ Calculation – Manual & SPSS
- ❑ Examples

03 Spearman Rank Correlation

- ❑ Definition & Key Concepts
- ❑ Applications
- ❑ Spearman Rank Correlation Coefficient
- ❑ Calculation – Manual & SPSS

Overview

Correlation: Correlation describes the **relationship** between **two variables**.

<u>Methods of studying correlation:</u>

(depending on data type, assumptions, and objective)

1. Scatter Diagram (Visual Method)
2. **Pearson's Correlation Coefficient (r)**
3. **Spearman's Rank Correlation (ρ or r_s)**
4. Kendall’s Tau
5. Concurrent Deviation Method (Less common)

1. Scatter Diagram (Visual Method)

- A graphical representation of data pairs on an XY plot.
- **Visual insight** into the type and strength of relationship.
- Pros: Simple and intuitive.
- Cons: No precise value for correlation.

2. Pearson's Correlation Coefficient (r)

•Measures **linear correlation** between two continuous variables.
•Assumes normal distribution, linearity, no outliers.
•Value between -1 and +1.
•Formula-based method.

3. Spearman's Rank Correlation (ρ or r_s)

•Measures **monotonic relationship** using ranks.
•Useful when data are **ordinal** or not normally distributed.
•Less sensitive to outliers.

4. Kendall's Tau

•Measures association based on the **concordance** of paired ranks.
•Often used in small sample sizes or when tied ranks are present.
•More robust than Spearman in some cases.

5. Concurrent Deviation Method *(Less common)*

•Based on the direction (+ or –) of deviations from the mean.
•Simple, but provides only a rough idea of correlation.

Method	Data Type	Type of Relationship	Notes
Scatter Plot	Any	Visual/Any	Exploratory
Pearson Correlation (r)	Continuous	Linear	Requires normality & linearity
Spearman Correlation (ρ)	Ordinal/Nonparam.	Monotonic	Non-parametric, based on ranks
Kendall’s Tau	Ordinal	Monotonic	Robust, good for small samples
Concurrent Deviation	Any	Direction only	Simplified method

Pearson Correlation (r)

OBJECTIVES

01

Explain the concepts and principles of Pearson Correlation

02

Discuss the assumptions

03

Perform Pearson correlation test manually and using SPSS software, interpret and report the results.

Pearson Correlation (r)

- A statistical measure that assesses the **strength** and **direction** of a **linear relationship** between two **continuous** variables.
- Also called the **Pearson Product-Moment Correlation Coefficient**.
- Symbol: **r**

Historical Background

Developed by **Karl Pearson** in the early 20th century.

Used widely in fields like psychology, medicine, economics, and social sciences.

Assumptions of Pearson Correlation

1. Variables must be continuous (interval or ratio)
2. Linear relationship between variables
3. Normal distribution
4. Absence of Outliers
5. Homoscedasticity (equal variances)

Points 1 to 4 are a core assumption for Pearson correlation.

1. Continuous (Interval or Ratio) Variables

- Both variables should be measured at the interval or ratio level (i.e., they are continuous).
- Examples of variables that meet this criterion 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.
- Ordinal or categorical variables are not appropriate for Pearson correlation. If the data is ordinal, consider using Spearman's rank correlation.

📘 Examples of Interval Variables

These have equal intervals but **no true zero**:

Variable	Description
Temperature (°C or °F)	0°C does not mean "no temperature"
IQ Score	An IQ of 0 doesn't mean no intelligence
Dates (Calendar Years)	0 AD is arbitrary, not a true zero
Standardized Test Scores	e.g., SAT, GRE

📗 Examples of Ratio Variables

These have a **true zero** and allow full range of mathematical operations (including ratios):

Variable	Description
Height (cm, m)	0 cm = absence of height
Weight (kg, lbs)	0 kg = no weight
Age (in years)	0 years = newborn
Income ($)	0 = no income
Distance (km, miles)	0 = no distance
Time (seconds)	0 = no time elapsed

2. Linearity

Definition: The relationship between the two variables should be linear — i.e., changes in one variable are associated with proportional changes in the other.

Why it matters: Pearson correlation only captures linear relationships. A nonlinear pattern may result in a low r-value, even when a strong relationship exists.

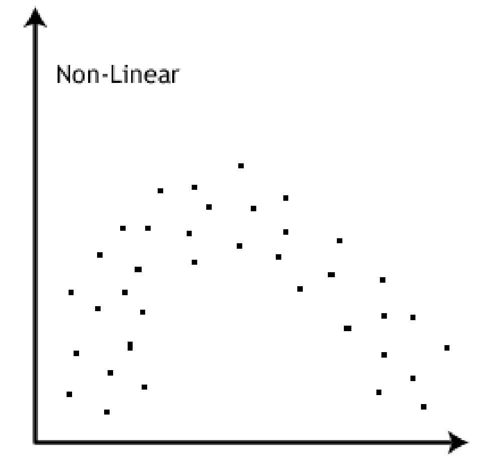
How to check: Use a scatterplot to visually inspect if the data follows a straight-line trend.

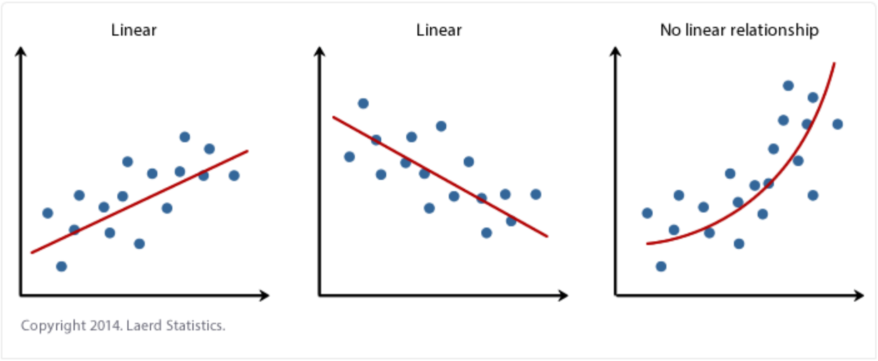
Scatterplot using SPSS Statistics

Quick Steps

1. Click **Graphs** -> **Legacy Dialogs** -> **Scatter/Dot** *(in older versions of SPSS)* OR
   Click **Graphs** > **Scatter/Dot** *(in newer versions of SPSS)*
2. Select **Simple Scatter**
3. Click **Define**
4. Click **Reset** (recommended)
5. Select the predictor/independent variable and move it into the **X Axis** box
6. Select the criterion/dependent variable and move it into the **Y Axis** box
7. Select **Titles** to add a title (recommended), then click **Continue**
8. Select **OK**
9. Your scatter plot will appear in SPSS Output Viewer

Your scatterplot may look something like one of the following:

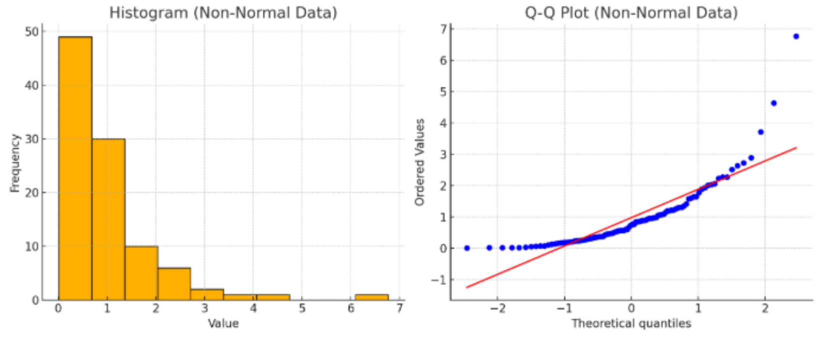


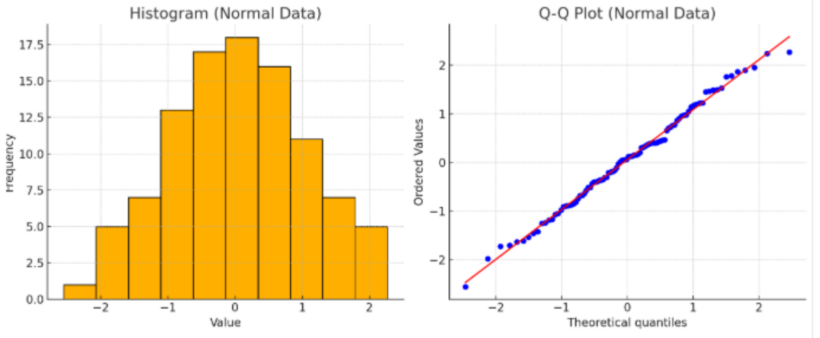


3. Normal Distribution (Bivariate Normality)

- Each variable should be approximately **normally distributed**.
- The joint distribution of both variables should be bivariate normal.
  - But this assumption is difficult to assess, so a simpler method is more commonly used.
  - This simpler method involves determining the normality of each variable separately.
- Especially important when testing significance (p-values).
- How to check:
  - Histogram or Q-Q plot for each variable
  - Use normality tests like Shapiro-Wilk or Kolmogorov-Smirnov.
    - ✓ If the $p > 0.05$, the data is normal.

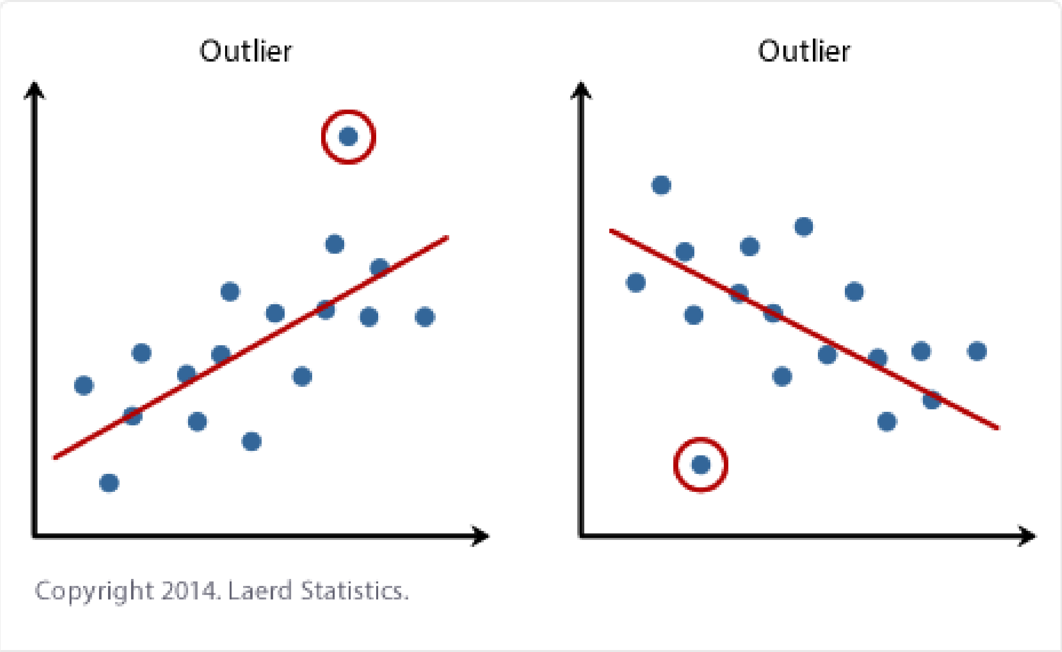
1. **Open SPSS:** Start the SPSS software.
2. **Navigate to Explore:** Go to the SPSS menu: Analyze > Descriptive Statistics > Explore.
3. **Select Variables:** In the "Explore" window, transfer the variable(s) you wish to test for normality to the "Dependent List".
4. **Customize Plots:** Click on the "Plots..." button.
5. **Check Normality Plots:** In the "Plots" window, select the "Normality plots with tests" option.
6. **Run the Test:** Click "Continue" to return to the main "Explore" window and then click "OK" to run the normality test.
7. **Interpret Results:** SPSS will generate output, including descriptive statistics, histograms, and the results of the Kolmogorov-Smirnov and Shapiro-Wilk tests of normality. The output will show the p-values for these tests. If the p-value is greater than 0.05 for both tests, the data is considered approximately normally distributed.





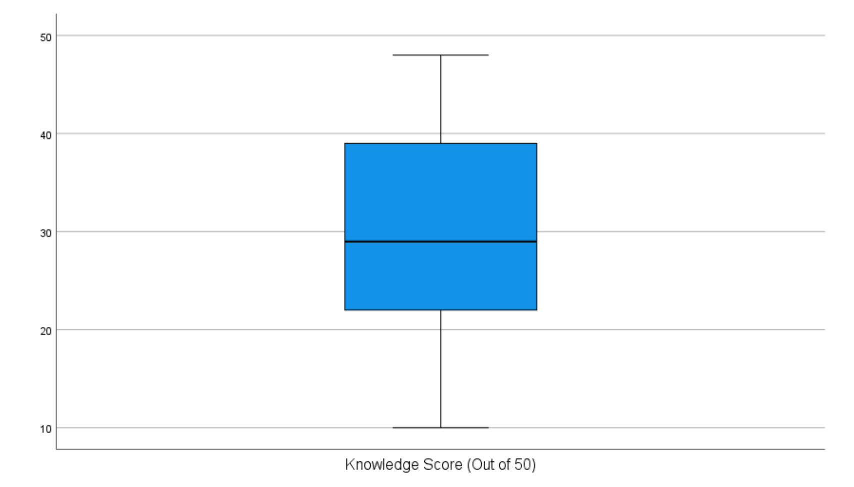
4. Absence of Outliers

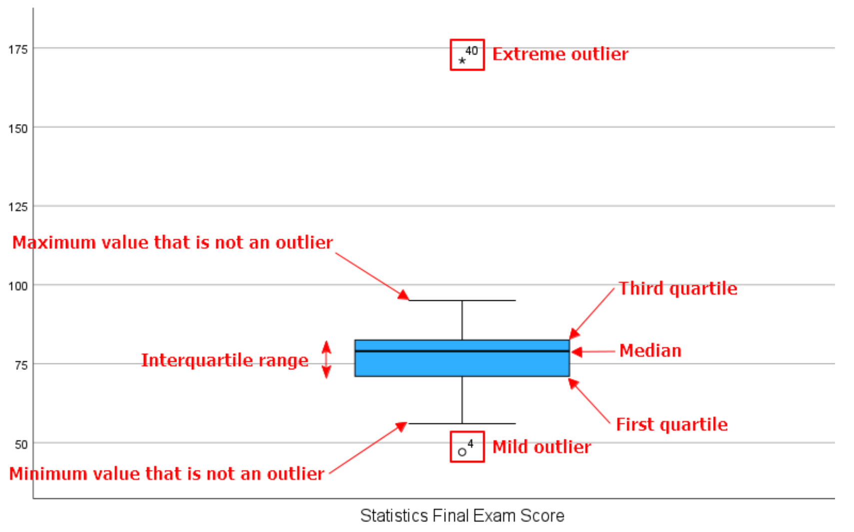
- Pearson correlation is very sensitive to outliers.
- Even a single extreme value can dramatically skew the correlation coefficient.
- How to check:
  - ✓Boxplots
  - ✓Explore -> Using outlier identification
  - ✓Z-scores (values > ±3 considered outliers)



1. Using Box Plots:

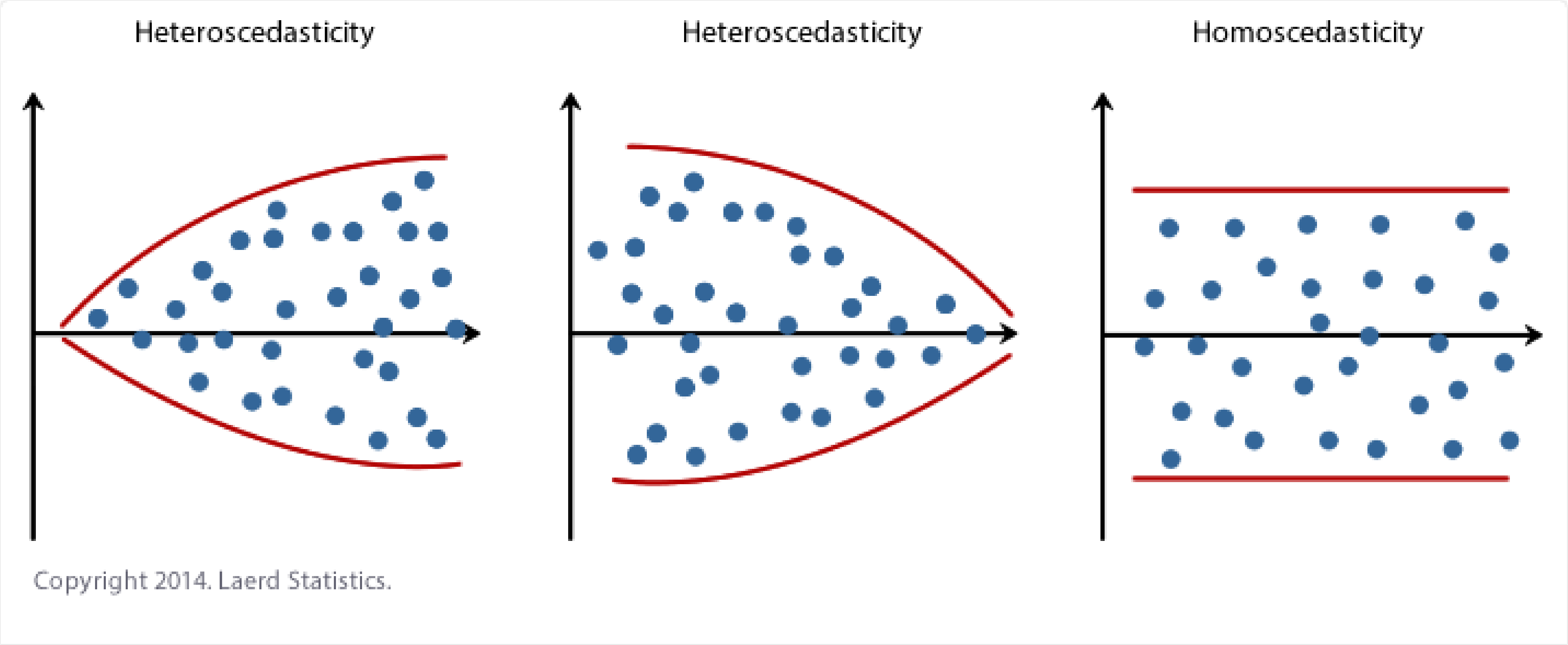
- **Open SPSS:** Launch the SPSS software and open your dataset.
- **Navigate to Graphs:** Go to "Graphs" > "Legacy Dialogs" > "Boxplot".
- **Select Variable:** Choose the variable you want to analyze for outliers and move it to the "Dependent" box.
- **Define Categories (Optional):** If you have grouping variables, move them to the "Categorical" box to create separate box plots for each group.
- **Generate the Plot:** Click "OK" to create the box plot.
- **Examine Outliers:** Look for points that fall outside the "whiskers" of the box plot. These points are potential outliers.





5. Homoscedasticity (Equal Variances)

- Homoscedasticity refers to the assumption that the **variance of the errors (or residuals) is constant** across all levels of the independent variable.
- Also called **homogeneity of variances**.
- Homoscedasticity is **not a core assumption** for Pearson correlation itself.
- Violations lead to heteroscedasticity, which can distort the strength of the correlation.
- How to check:
  - ✓ Plot residuals vs. fitted values
  - ✓ Look for a “fan shape” pattern — indicates heteroscedasticity



Heteroscedasticity
Heteroscedasticity
Homoscedasticity
Copyright 2014. Laerd Statistics.

Summary Table

Assumption	Description	How to Check
1. Continuous Variables	Both variables must be interval or ratio scale	Know your data scale
2. Linearity	Straight-line relationship	Scatterplot
3. Normality	Each variable is normally distributed	Histogram, Q-Q Plot, Normality Tests
4. No Outliers	No extreme values affecting results	Boxplot, Z-scores
5. Homoscedasticity	Equal spread of data along regression line	Residual plots

Pearson Correlation Coefficient, r

The calculations are based on the differences between the observations x_i, y_i and their means $\bar{x}$ *and* $\bar{y}$ as shown in the following formula:

Formula

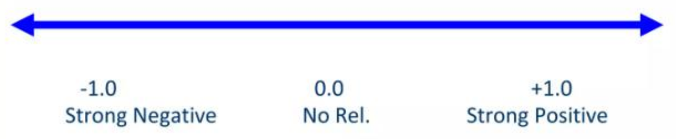
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- X_i, Y_i are individual data points
- $\bar{X}, \bar{Y}$ are the mean values of X and Y

where x_i, y_i are values of the *n* pairs of two variables.

Interpretation of r



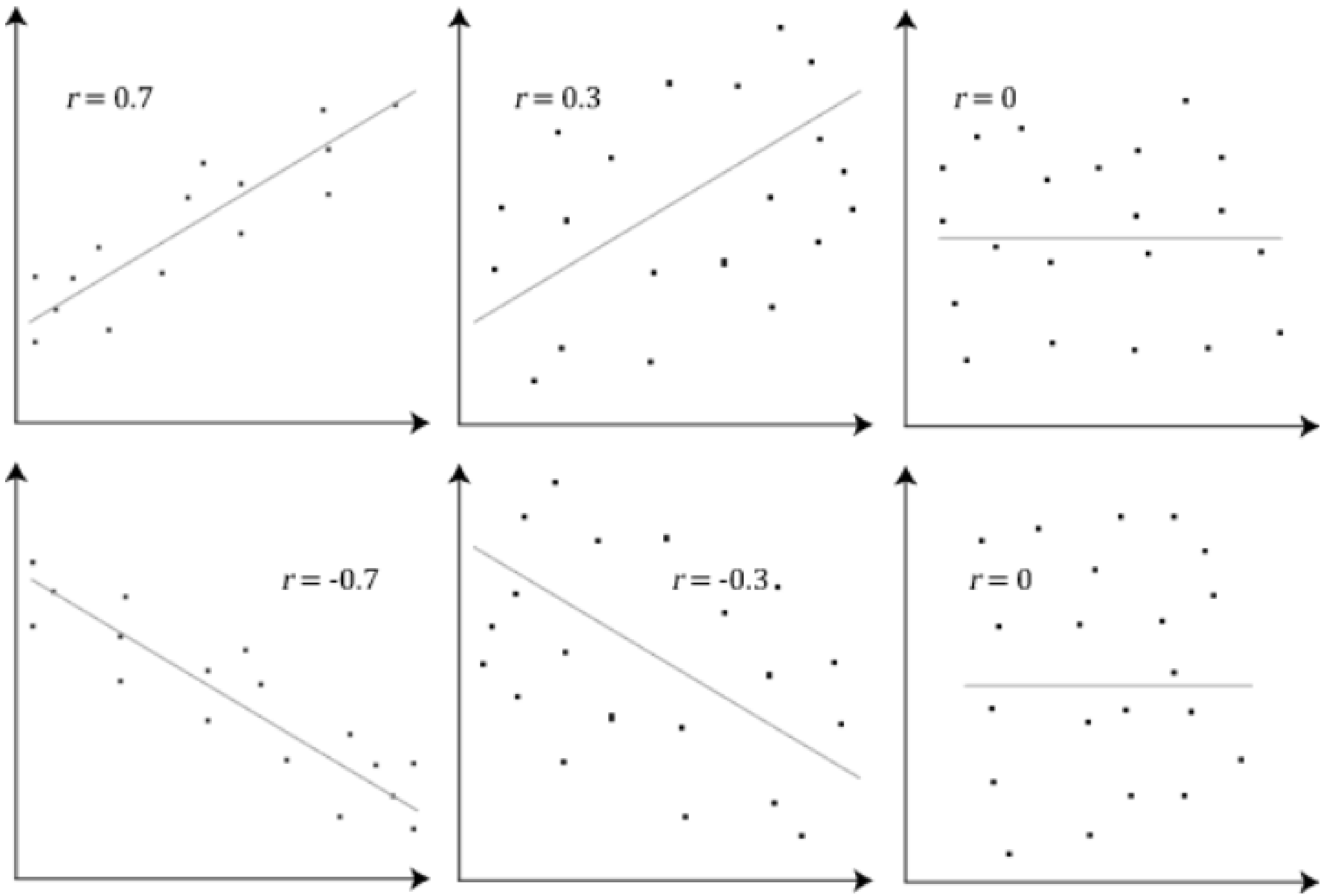
- R tells us how close is the linear relationship between the 2 variables
- R lies between **-1 and +1**
- **Negative** values indicate a **negative linear relationship**
  - ✓ 1 variable increases, the other decreases
- **Positive** values indicate a **positive linear relationship**
  - ✓ 1 variable increases, so does the other
- **R = 0** indicates **no linear relationship**
  - ✓ The values of each variable are independent of each other

Interpreting Pearson Correlation Coefficient

Range of Correlation Coefficient Values	Level of Correlation	Range of Correlation Coefficient Values	Level of Correlation
0.80 to 1.00	Very Strong Positive	-1.00 to -0.80	Very Strong Negative
0.60 to 0.79	Strong Positive	-0.79 to -0.60	Strong Negative
0.40 to 0.59	Moderate Positive	-0.59 to -0.40	Moderate Negative
0.20 to 0.39	Weak Positive	-0.39 to -0.20	Weak Negative
0.00 to 0.19	Very Weak Positive	-0.19 to -0.01	Very Weak Negative

Meghanathan, N. (2016). Assortativity Analysis of Real-World Network Graphs based on Centrality Metrics. Computer and Information Science, 9(3), p7. https://doi.org/10.5539/CIS.V9N3P7

Different relationships and their correlation coefficients are shown in the diagram below:




TIME FOR
A BREAK!

Performing Pearson Correlation (Manual Calculation)

Calculate a Pearson Correlation Coefficient (r) for the following dataset:

X	Y
6	45
12	47
13	39
17	58
22	68
25	76
27	75
29	74
30	78
32	81

Step 1: Calculate the Mean of X and Y

	X	Y
	6	45
	12	47
	13	39
	17	58
	22	68
	25	76
	27	75
	29	74
	30	78
	32	81
Mean		

Step 2: Calculate the Difference between Means

	x	y	$x - x_{mean}$	$y - y_{mean}$
	6	45		
	12	47		
	13	39		
	17	58		
	22	68		
	25	76		
	27	75		
	29	74		
	30	78		
	32	81		
Mean	**21.3**	**64.1**		

Step 3: Calculate the Remaining Values

	x	y	$x - x_{mean}$	$y - y_{mean}$	$(x - x_{mean})*(y - y_{mean})$	$(x - x_{mean})^2$	$(y - y_{mean})^2$
	6	45	-15.3	-19.1	[redacted]	[redacted]	[redacted]
	12	47	-9.3	-17.1	[redacted]	[redacted]	[redacted]
	13	39	-8.3	-25.1	[redacted]	[redacted]	[redacted]
	17	58	-4.3	-6.1	[redacted]	[redacted]	[redacted]
	22	68	0.7	3.9	[redacted]	[redacted]	[redacted]
	25	76	3.7	11.9	[redacted]	[redacted]	[redacted]
	27	75	5.7	10.9	[redacted]	[redacted]	[redacted]
	29	74	7.7	9.9	[redacted]	[redacted]	[redacted]
	30	78	8.7	13.9	[redacted]	[redacted]	[redacted]
	32	81	10.7	16.9	[redacted]	[redacted]	[redacted]
Mean	**21.3**	**64.1**					

Step 4: Calculate the Sums

	x	y	$x - x_{mean}$	$y - y_{mean}$	$(x - x_{mean})*(y - y_{mean})$	$(x - x_{mean})^2$	$(y - y_{mean})^2$
	6	45	-15.3	-19.1	292.23	234.09	364.81
	12	47	-9.3	-17.1	159.03	86.49	292.41
	13	39	-8.3	-25.1	208.33	68.89	630.01
	17	58	-4.3	-6.1	26.23	18.49	37.21
	22	68	0.7	3.9	2.73	0.49	15.21
	25	76	3.7	11.9	44.03	13.69	141.61
	27	75	5.7	10.9	62.13	32.49	118.81
	29	74	7.7	9.9	76.23	59.29	98.01
	30	78	8.7	13.9	120.93	75.69	193.21
	32	81	10.7	16.9	180.83	114.49	285.61
Mean	**21.3**	**64.1**		**Sum**			

Step 5: Calculate the Pearson Correlation Coefficient

Formula

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

	x	y	$x - x_{mean}$	$y - y_{mean}$	$(x - x_{mean})*(y - y_{mean})$	$(x - x_{mean})^2$	$(y - y_{mean})^2$
	6	45	-15.3	-19.1	292.23	234.09	364.81
	12	47	-9.3	-17.1	159.03	86.49	292.41
	13	39	-8.3	-25.1	208.33	68.89	630.01
	17	58	-4.3	-6.1	26.23	18.49	37.21
	22	68	0.7	3.9	2.73	0.49	15.21
	25	76	3.7	11.9	44.03	13.69	141.61
	27	75	5.7	10.9	62.13	32.49	118.81
	29	74	7.7	9.9	76.23	59.29	98.01
	30	78	8.7	13.9	120.93	75.69	193.21
	32	81	10.7	16.9	180.83	114.49	285.61
Mean	**21.3**	**64.1**		**Sum**	1172.7	704.1	2176.9

$r = 1172.7 / \sqrt{(704.1)*(2176.9)} = \mathbf{0.947}$

The Pearson Correlation Coefficient, r = 0.947.

Since this value is close to 1, this is an indication that X and Y are strongly positively correlated.

Step 6: Determine p-value Using Pearson Correlation Table

Step 1: Calculate r manually

✓ $r=0.947$

Step 2: Determine your sample size n

✓ $n=10$

Step 3: Calculate the degrees of freedom

- $df = n-2$
- $df = 10 - 2 = 8$

Step 4: Choose your significance level (α)

Common levels:

- α=0.05\alpha = 0.05α=0.05 (95% confidence)
- α=0.01\alpha = 0.01α=0.01 (99% confidence)

Step 5: Refer to a Pearson critical value table

- Look up the **critical r values** for your **df = n - 2** at the desired significance level.

df\α	0.2	0.1	0.05	0.02	0.01	0.001
1	0.951057	0.987688	0.996917	0.999507	0.999877	0.999999
2	0.800000	0.900000	0.950000	0.980000	0.990000	0.999000
3	0.687049	0.805384	0.878339	0.934333	0.958735	0.991139
4	0.608400	0.729299	0.811401	0.882194	0.917200	0.974068
5	0.550863	0.669439	0.754492	0.832874	0.874526	0.950883
6	0.506727	0.621489	0.706734	0.788720	0.834342	0.924904
7	0.471589	0.582206	0.666384	0.749776	0.797681	0.898260
8	0.442796	0.549357	0.631897	0.715459	0.764592	0.872115
9	0.418662	0.521404	0.602069	0.685095	0.734786	0.847047
10	0.398062	0.497265	0.575983	0.658070	0.707888	0.823305
11	0.380216	0.476156	0.552943	0.633863	0.683528	0.800962
12	0.364562	0.457500	0.532413	0.612047	0.661376	0.779998
13	0.350688	0.440861	0.513977	0.592270	0.641145	0.760351
14	0.338282	0.425902	0.497309	0.574245	0.622591	0.741934
15	0.327101	0.412360	0.482146	0.557737	0.605506	0.724657
16	0.316958	0.400027	0.468277	0.542548	0.589714	0.708429
17	0.307702	0.388733	0.455531	0.528517	0.575067	0.693163
18	0.299210	0.378341	0.443763	0.515505	0.561435	0.678781
19	0.291384	0.368737	0.432858	0.503397	0.548711	0.665208
20	0.284140	0.359827	0.422714	0.492094	0.536800	0.652378
21	0.277411	0.351531	0.413247	0.481512	0.525620	0.640230
22	0.271137	0.343783	0.404386	0.471579	0.515101	0.628710
23	0.265270	0.336524	0.396070	0.462231	0.505182	0.617768
24	0.259768	0.329705	0.388244	0.453413	0.495808	0.607360
25	0.254594	0.323283	0.380863	0.445078	0.486932	0.597446
26	0.249717	0.317223	0.373886	0.437184	0.478511	0.587988
27	0.245110	0.311490	0.367278	0.429693	0.470509	0.578956
28	0.240749	0.306057	0.361007	0.422572	0.462892	0.570317
29	0.236612	0.300898	0.355046	0.415792	0.455631	0.562047
30	0.232681	0.295991	0.349370	0.409327	0.448699	0.554119
35	0.215598	0.274611	0.324573	0.380976	0.418211	0.518898
40	0.201796	0.257278	0.304396	0.357787	0.393174	0.489570
45	0.190345	0.242859	0.287563	0.338367	0.372142	0.464673
50	0.180644	0.230620	0.273243	0.321796	0.354153	0.443201
60	0.164997	0.210832	0.250035	0.294846	0.324818	0.407865
70	0.152818	0.195394	0.231883	0.273695	0.301734	0.379799
80	0.142990	0.182916	0.217185	0.256525	0.282958	0.356816
90	0.134844	0.172558	0.204968	0.242227	0.267298	0.337549
100	0.127947	0.163782	0.194604	0.230079	0.253979	0.321095
125	0.114477	0.146617	0.174308	0.206245	0.227807	0.288602
150	0.104525	0.133919	0.159273	0.188552	0.208349	0.264316
175	0.096787	0.124036	0.147558	0.174749	0.193153	0.245280
200	0.090546	0.116060	0.138098	0.163592	0.180860	0.229840
250	0.081000	0.103852	0.123607	0.146483	0.161994	0.206079
300	0.073951	0.094831	0.112891	0.133819	0.148019	0.188431
350	0.068470	0.087814	0.104552	0.123957	0.137131	0.174657
400	0.064052	0.082155	0.097824	0.115997	0.128339	0.163520
450	0.060391	0.077466	0.092248	0.109397	0.121046	0.154273
500	0.057294	0.073497	0.087528	0.103808	0.114870	0.146436
600	0.052305	0.067103	0.079920	0.094798	0.104911	0.133787
700	0.048427	0.062132	0.074004	0.087789	0.097161	0.123935
800	0.045301	0.058123	0.069234	0.082135	0.090909	0.115981
900	0.042711	0.054802	0.065281	0.077450	0.085727	0.109385
1000	0.040520	0.051993	0.061935	0.073484	0.081340	0.103800
1500	0.033086	0.042458	0.050582	0.060022	0.066445	0.084822
2000	0.028654	0.036772	0.043811	0.051990	0.057557	0.073488
3000	0.023397	0.030027	0.035775	0.042457	0.047006	0.060027
4000	0.020262	0.026005	0.030984	0.036773	0.040713	0.051996
5000	0.018123	0.023260	0.027714	0.032892	0.036417	0.046512

Step 6: Compare your r to the critical values

- If **r > r critical**, the result is **statistically significant**.
- If **r < r critical**, the result is **not statistically significant**.
- Your r=0.947 is **greater than all critical values**.
  - ✅ 0.947 > 0.632 → **Significant at α = 0.05**
  - ✅ 0.947 > 0.872 → **Significant at α = 0.001**

Conclusion
Your correlation is very strong and highly statistically significant — $p < 0.001$.
This means there is very strong evidence that the relationship between your two variables is not due to chance.

Pearson Correlation Coefficient using SPSS

SPSS Statistics

Example

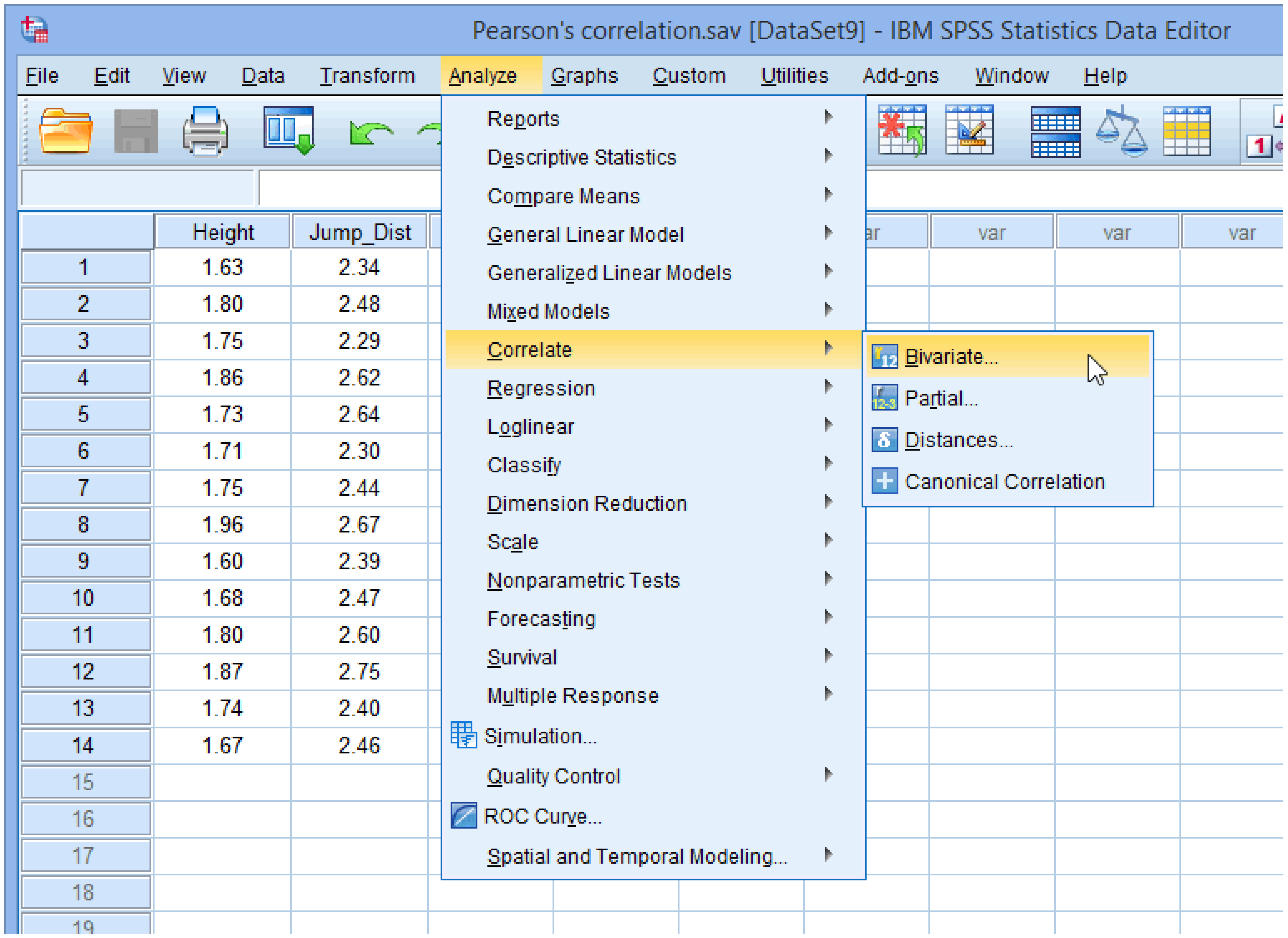
A researcher wants to know whether a person's height is related to how well they perform in a long jump. The researcher recruited untrained individuals from the general population, measured their height and had them perform a long jump. The researcher then investigated whether there was an association between height and long jump performance by running a Pearson's correlation.

Dataset:
In SPSS Statistics, we created two variables so that we could enter our data: Height (i.e., participants' height) and Jump_Dist (i.e., distance jumped in a long jump).

- **Dataset:** Height_Jump_Pearson.xlsx

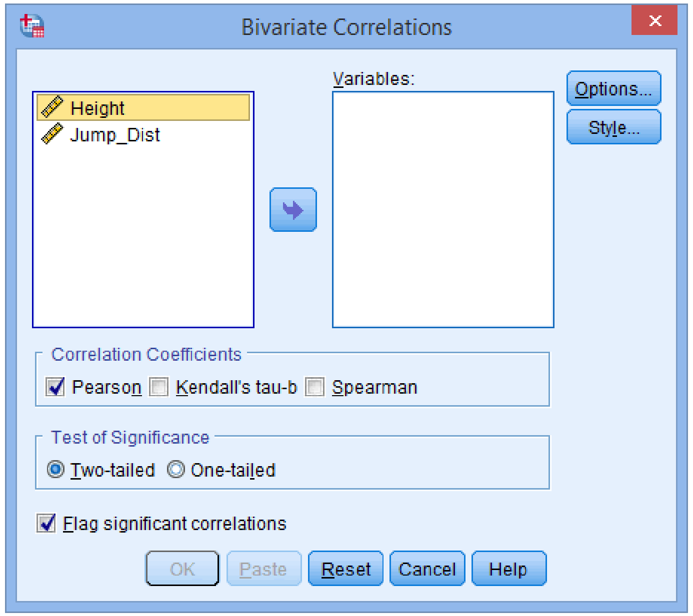
SCAN ME

1 Click **Analyze > Correlate > Bivariate...** on the main menu, as shown below:



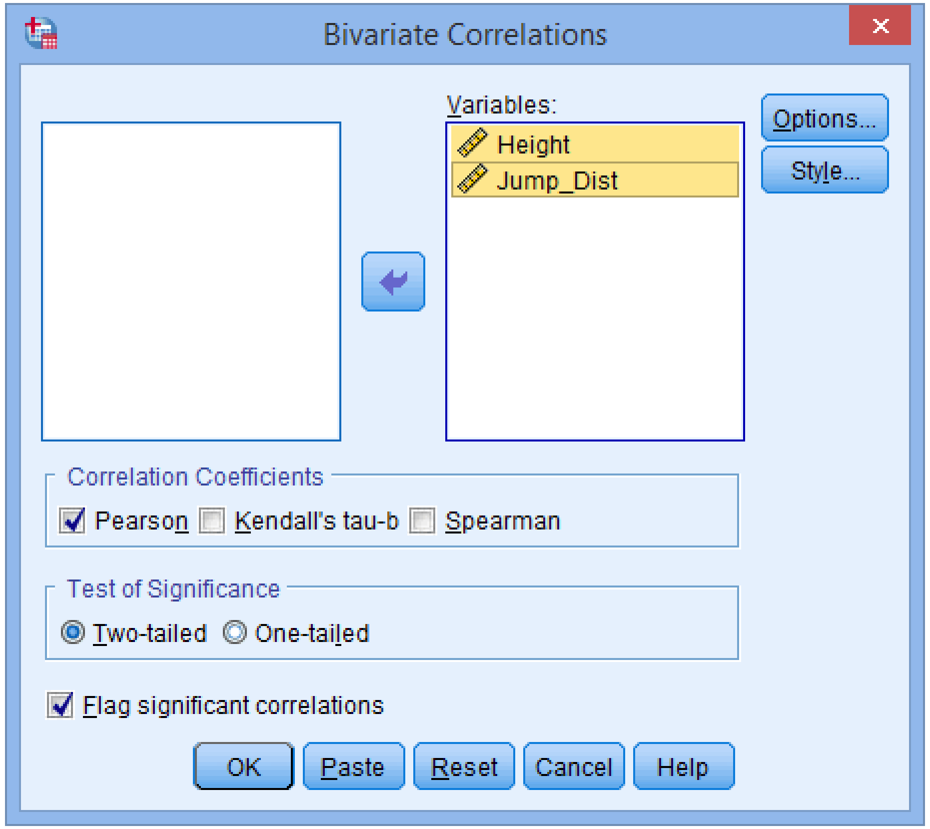
Published with written permission from SPSS Statistics, IBM Corporation.

You will be presented with the **Bivariate Correlations** dialogue box:



Published with written permission from SPSS Statistics, IBM Corporation.

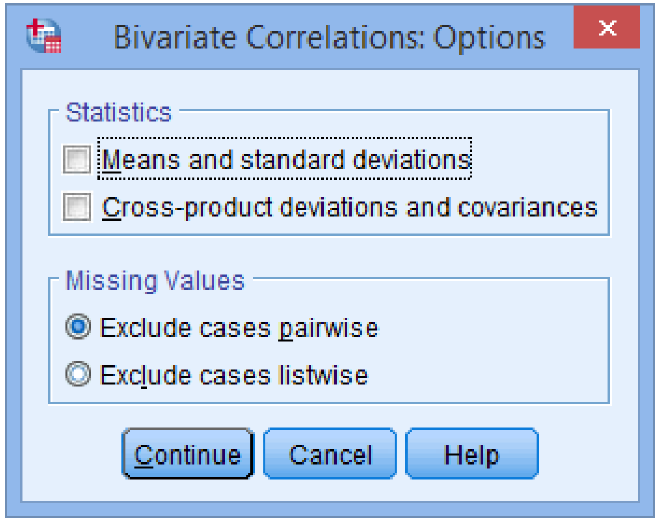
2. Transfer the variables Height and Jump_Dist into the Variables: box by dragging-and-dropping them or by clicking on them and then clicking on the button. You will end up with a screen similar to the one below:



Published with written permission from SPSS Statistics, IBM Corporation.

3. Make sure that the Pearson checkbox is selected under the –Correlation Coefficients– area (although it is selected by default in SPSS Statistics).

4 Click on the Options... button and you will be presented with the **Bivariate Correlations: Options** dialogue box. If you wish to generate some descriptives, you can do it here by clicking on the relevant checkbox in the –Statistics– area.



Published with written permission from SPSS Statistics, IBM Corporation.

---

5 Click on the Continue button. You will be returned to the **Bivariate Correlations** dialogue box.

---

6 Click on the OK button. This will generate the results of Pearson's correlation.

Output for Pearson's correlation

Correlations

		Height	Jump_Dist
Height	Pearson Correlation	1	.706**
	Sig. (2-tailed)		.005
	N	14	14
Jump_Dist	Pearson Correlation	.706**	1
	Sig. (2-tailed)	.005	
	N	14	14

**. Correlation is significant at the 0.01 level (2-tailed).

Reporting the Output

A Pearson product-moment correlation was run to determine the relationship between height and distance jumped in a long jump.
There was a strong, positive correlation between height and distance jumped, which was statistically significant ($r = .706$, $n = 14$, $p = .005$).

In this example, we can see that the Pearson correlation coefficient, r, is 0.706, and that it is statistically significant ($p = 0.005$).

Example 1

Topic:

Relationship between Physical Activity and Body Mass Index (BMI) among Adults: A Pearson Correlation Analysis

Background:

Public health studies often examine how lifestyle behaviors, like physical activity, relate to health outcomes such as obesity (measured by BMI).
Physical activity is known to reduce the risk of overweight and obesity. In this study, we randomly surveyed **40 adults** to record:

- Their **physical activity** (total minutes of exercise per week).
- Their **BMI** (Body Mass Index, kg/m²).

Dataset: PA_BMI.xlsx

SCAN ME

Hypothesis:

- **Null hypothesis (H_0):** There is no correlation between physical activity and BMI.
- **Alternative hypothesis (H_1):** There is a significant correlation between physical activity and BMI.

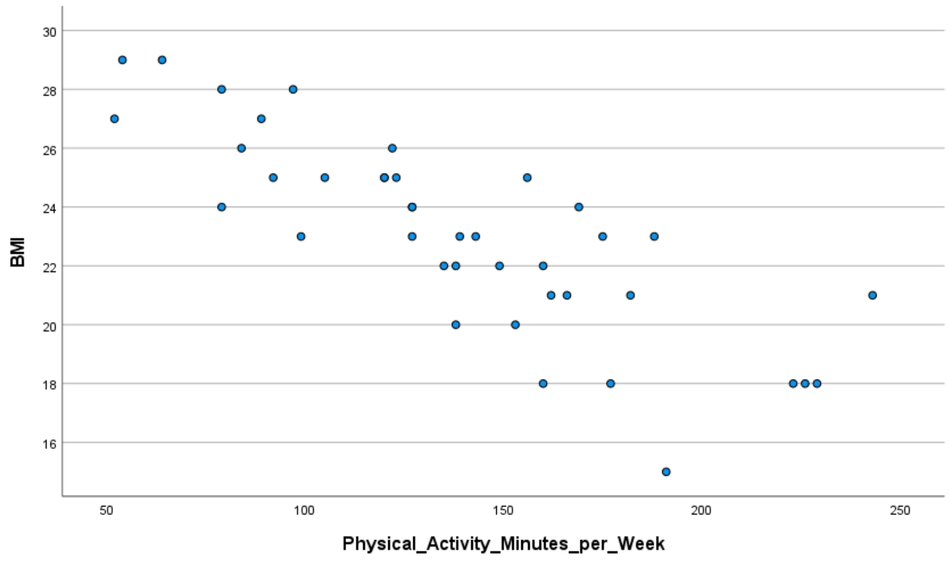
Assumptions

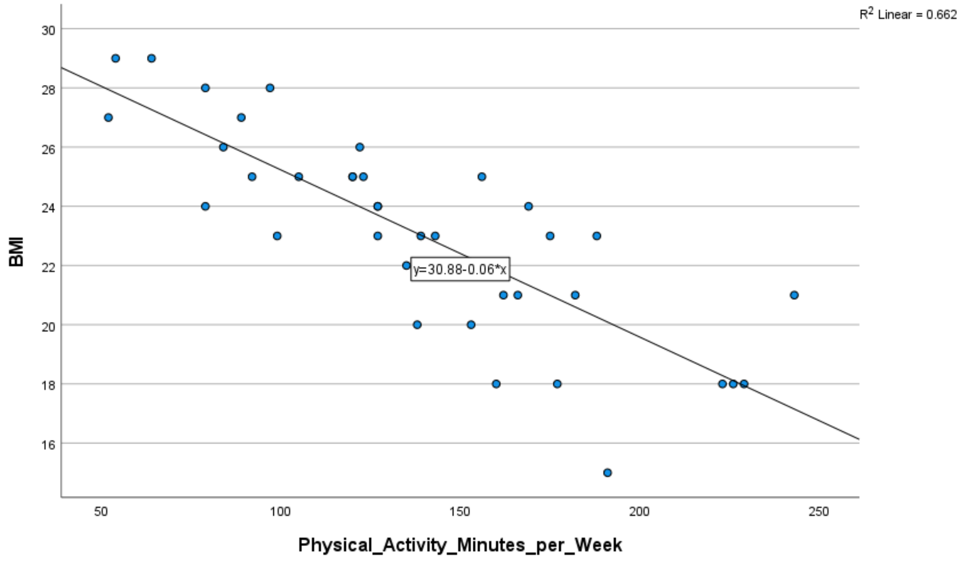
1. Variables?
2. Linearity?
3. Normality?
4. Outliers?
5. Homoscedasticity (optional)

Dataset

Variable	Type	Role
Physical Activity (minutes per week)	Continuous	Independent variable
Body Mass Index (BMI)	Continuous	Dependent variable

Physical_Activity_Minutes_per_Week	BMI
175	23
143	23
182	21
226	18
138	20
138	22
229	18
188	23
127	24
177	18
127	24
127	23
162	21
54	29
64	29
122	26
99	23
166	21
105	25
79	28
223	18
139	23
153	20
79	24
123	25
156	25
92	25
169	24
120	25
135	22
120	25
243	21
149	22
97	28
191	15
89	27
160	22
52	27
84	26
160	18


30
28
26
24
22
20
18
16
BMI
50
100
150
200
250
Physical_Activity_Minutes_per_Week

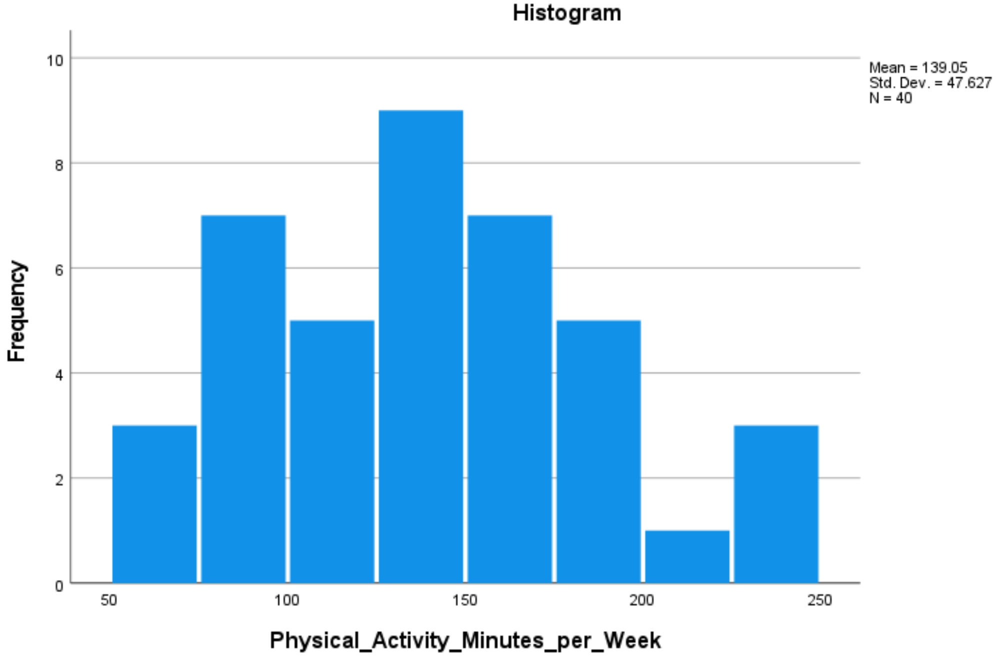

R2 Linear = 0.662
30
28
26
24
22
20
18
16
BMI
y=30.88-0.06*x
50
100
150
200
250
Physical_Activity_Minutes_per_Week

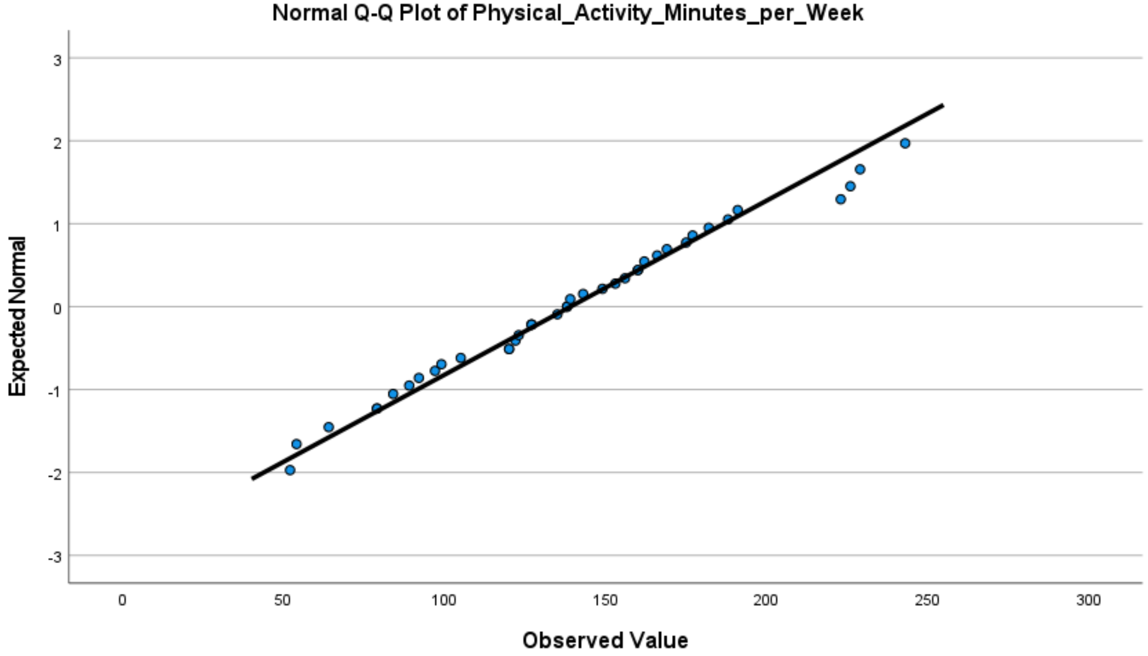
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Physical_Activity_Minutes_per_Week	.070	40	.200*	.979	40	.666

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



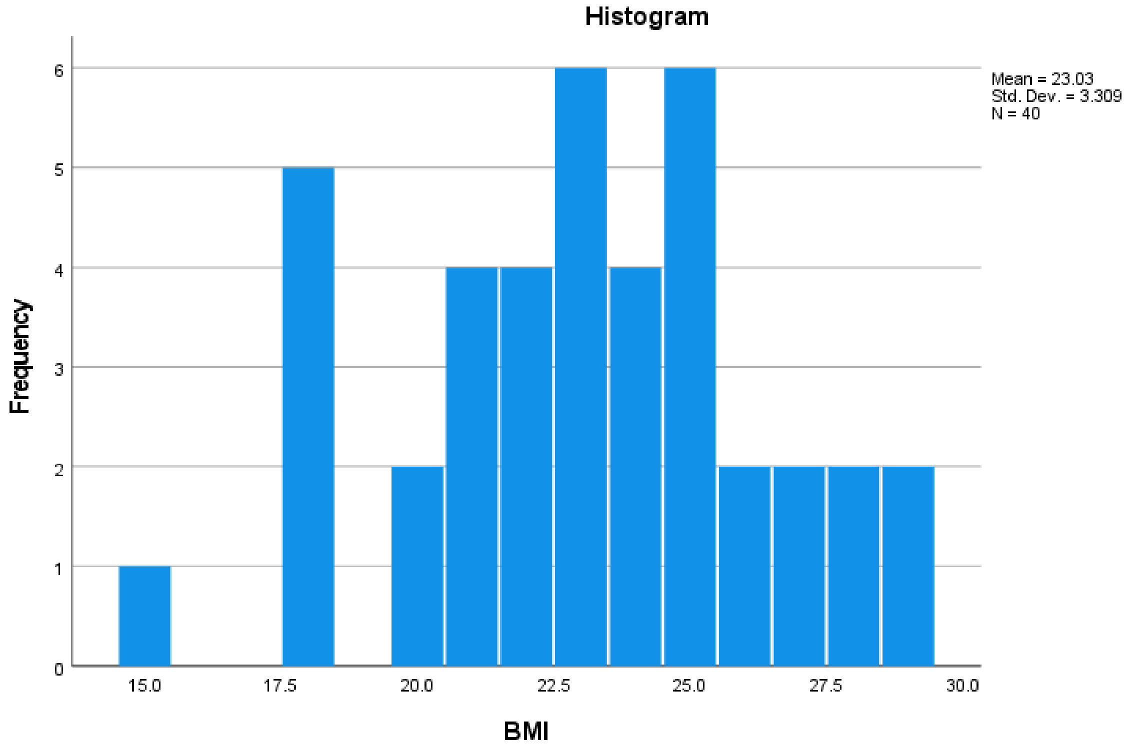


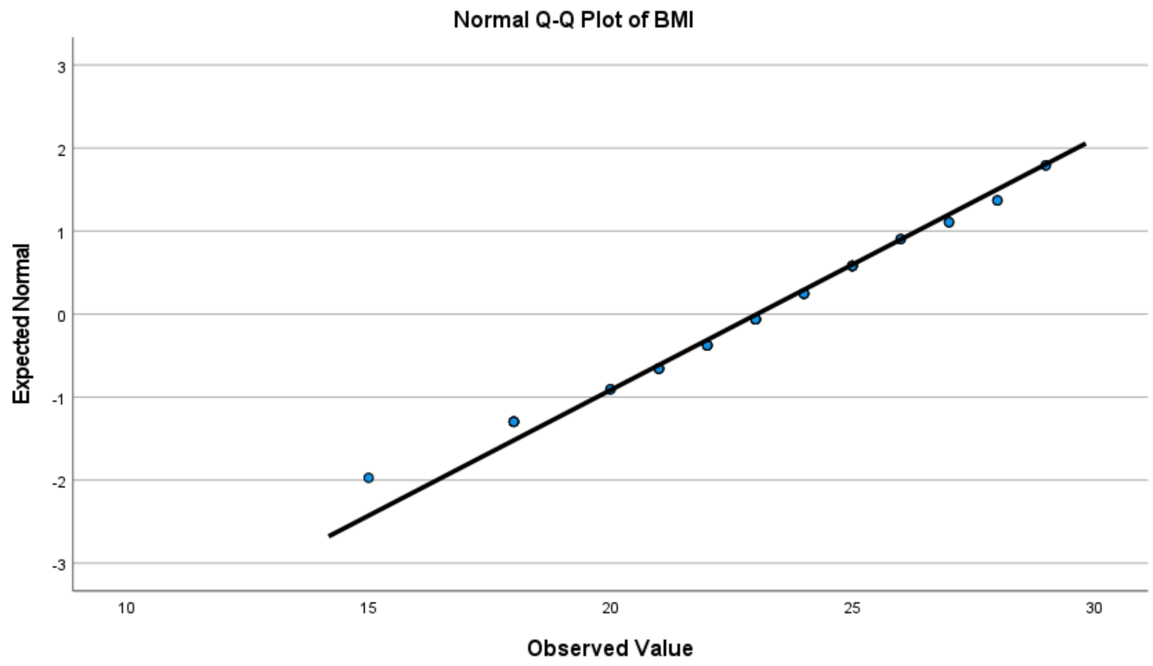
Tests of Normality

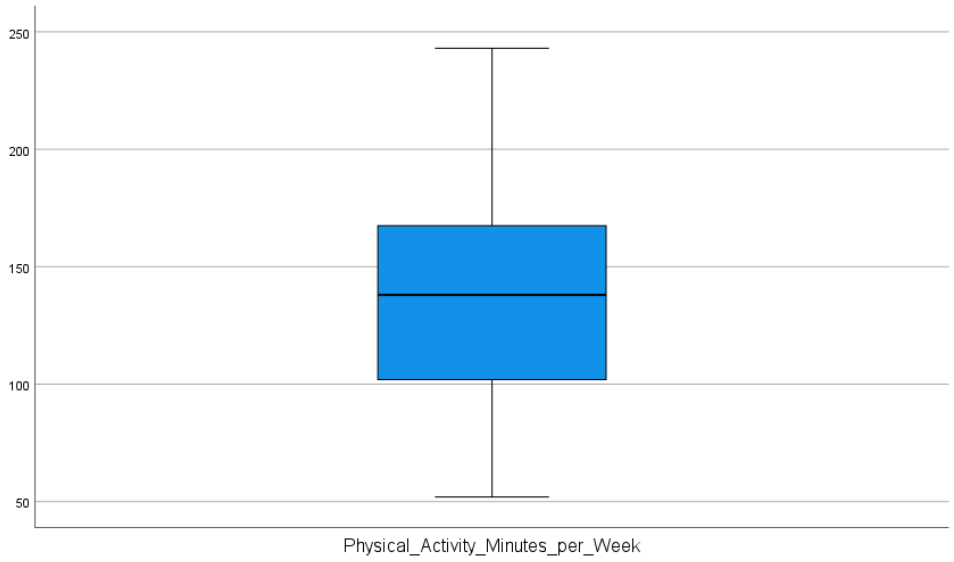
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
BMI	.097	40	.200*	.973	40	.441

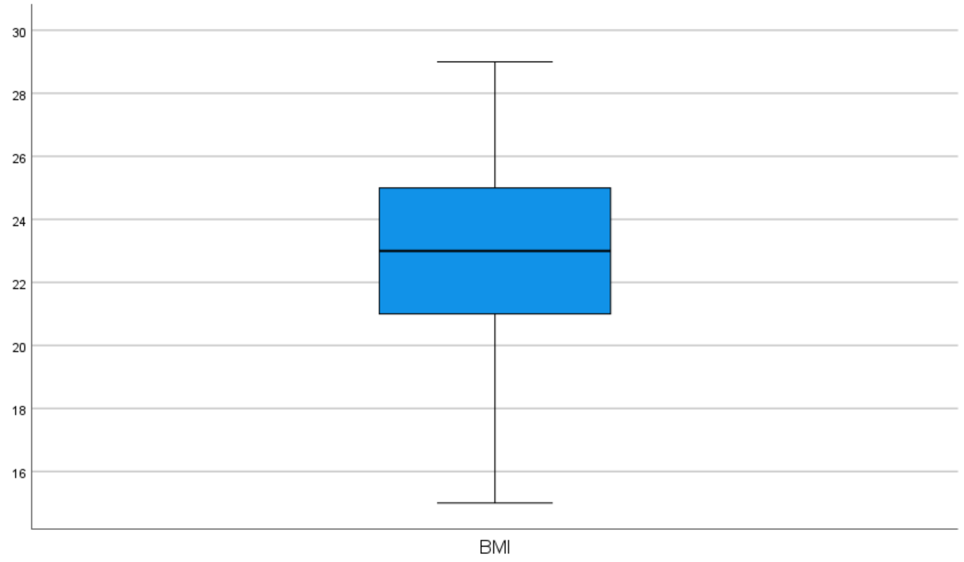
*. This is a lower bound of the true significance.

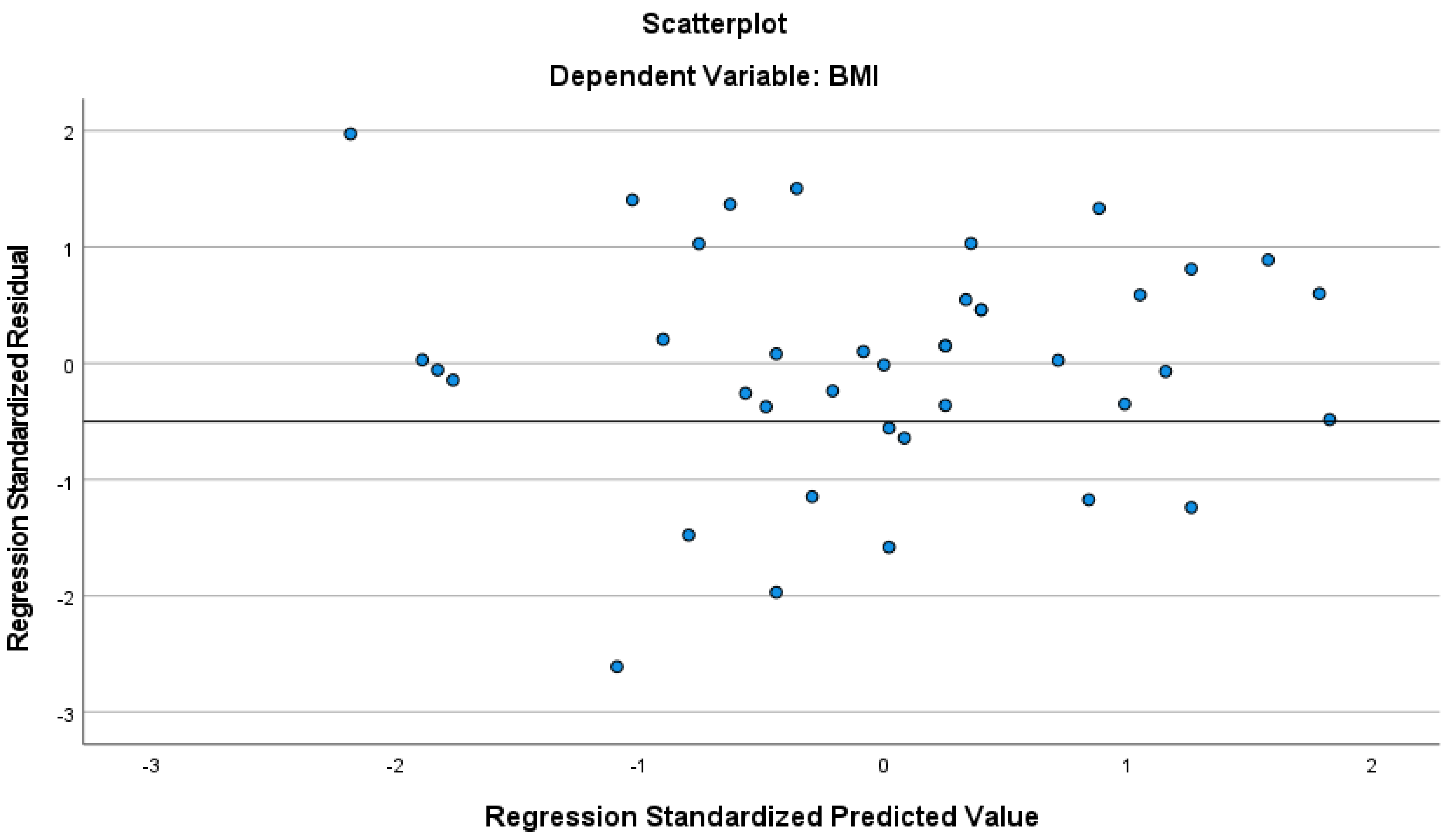
a. Lilliefors Significance Correction






250
200
150
100
50
Physical_Activity_Minutes_per_Week


30
28
26
24
22
20
18
16
BMI


Scatterplot
Dependent Variable: BMI
Regression Standardized Residual
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
Regression Standardized Predicted Value

The Pearson correlation test was conducted to assess the strength and direction of the relationship between physical activity and BMI.

Correlations

		Physical_Activity_Minutes_per_Week	BMI
Physical_Activity_Minutes_per_Week	Pearson Correlation	1	-.813**
	Sig. (2-tailed)		<.001
	N	40	40
BMI	Pearson Correlation	-.813**	1
	Sig. (2-tailed)	<.001	
	N	40	40

**. Correlation is significant at the 0.01 level (2-tailed).

✍ Interpretation:

- The **Pearson correlation coefficient (r = -0.813)** indicates a **strong negative correlation** between **Physical Activity Minutes per Week** and **BMI**.
- The **p-value (< 0.001)** shows that the correlation is **statistically significant** at the **0.01 level**.

Meaning:

➔ As physical activity increases, BMI tends to **decrease**.

➔ Adults who are more physically active tend to have a lower BMI.

Conclusion:

There is a **strong, statistically significant negative relationship** between physical activity levels and BMI among the 40 adults surveyed.

This supports public health strategies promoting regular physical activity to help manage or lower BMI.

Example 2

Title:
Relationship between Age and Systolic Blood Pressure among Hypertensive Patients: A Pearson Correlation Analysis.

Background:
Hypertension (high blood pressure) is a major risk factor for cardiovascular diseases.
Blood pressure tends to increase with age due to changes in blood vessel elasticity and other physiological factors.
This study aims to assess the relationship between **age** and **systolic blood pressure (SBP)** among patients diagnosed with hypertension.

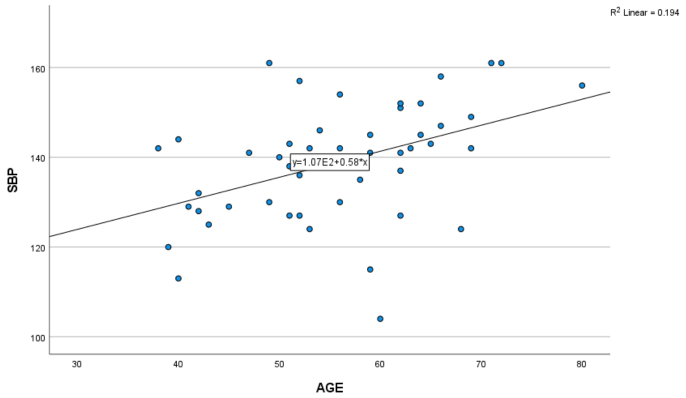
Dataset: AGE_BP.xlxs

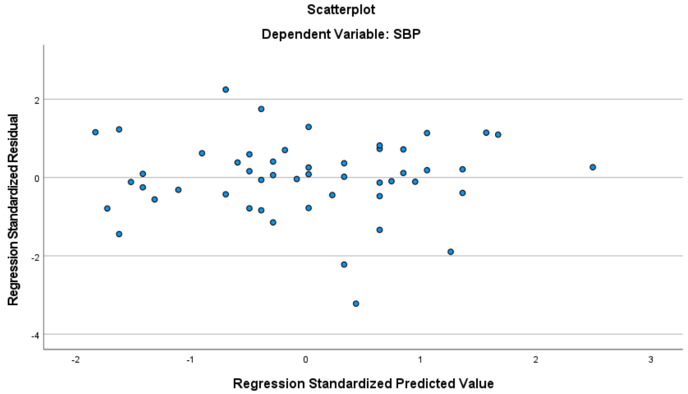
SCAN ME

Hypotheses:
•**Null Hypothesis (H_0):** There is no significant relationship between age and systolic blood pressure among hypertensive patients.
•**Alternative Hypothesis (H_1):** There is a significant relationship between age and systolic blood pressure among hypertensive patients.

Assumptions

1. **Variables?**
2. **Linearity?**
3. **Normality?**
4. **Outliers?**
5. **Homoscedasticity (optional)**



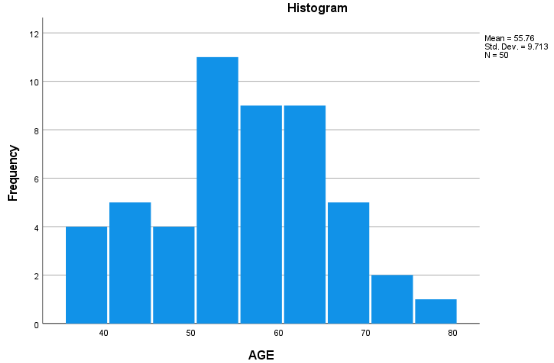


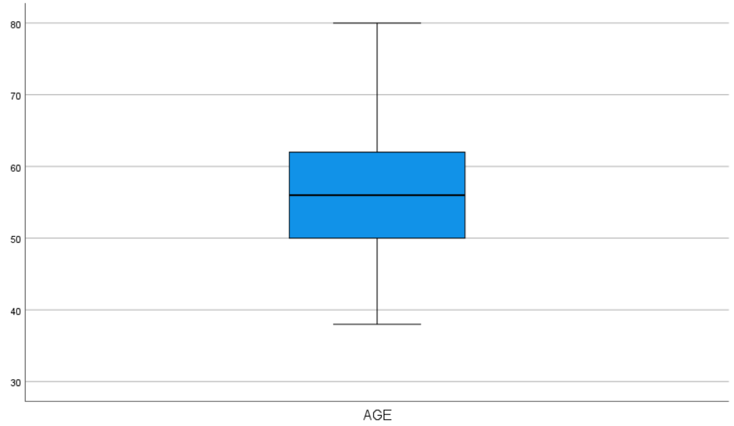
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AGE	.080	50	.200*	.981	50	.575

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction



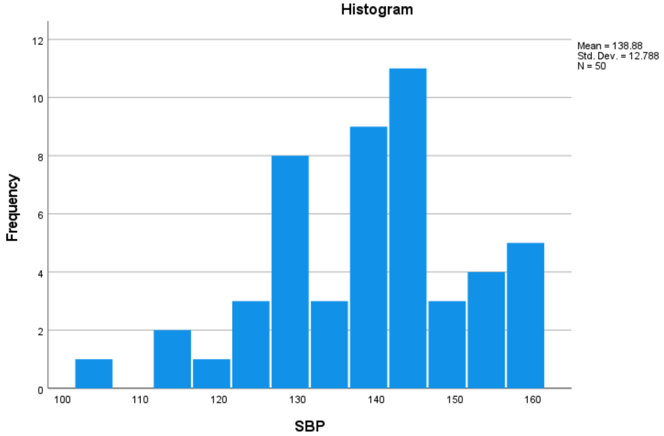


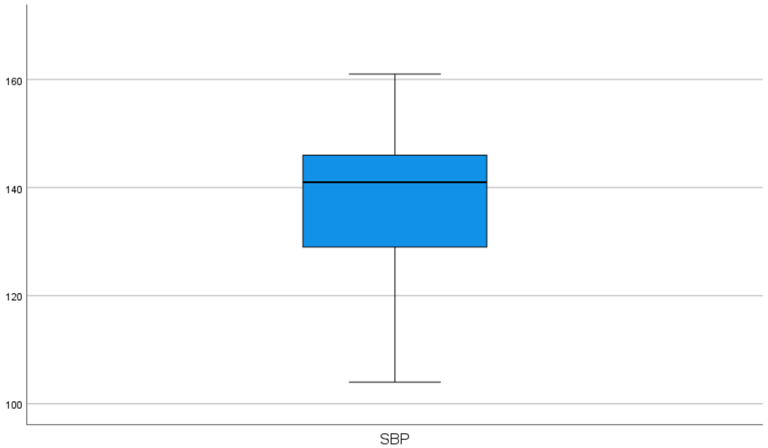
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SBP	.095	50	.200*	.975	50	.362

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction





Correlations			
		AGE	SBP
AGE	Pearson Correlation	1	.441**
	Sig. (2-tailed)		.001
	N	50	50
SBP	Pearson Correlation	.441**	1
	Sig. (2-tailed)	.001	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

✍ Interpretation:

- The **Pearson r = 0.441** indicates a **moderate positive correlation** between **Age** and **Systolic Blood Pressure**.
- The **p-value = 0.001** is **statistically significant** at the 0.01 level.

Meaning:

➔ As age increases, systolic blood pressure tends to increase as well among hypertensive patients.

📋 Conclusion (Public Health Context):

This study found a **moderate, statistically significant positive relationship** between **age** and **systolic blood pressure** among hypertensive patients.

From a public health perspective, these findings highlight the importance of **early screening, regular monitoring, and proactive blood pressure management**, especially in older adults.

References on Sample Size for Pearson Correlation

1. Bonett, D. G., & Wright, T. A. (2000).

Title: *Sample size requirements for estimating Pearson, Kendall, and Spearman correlations with fixed precision.*

Journal: *Psychological Methods, 5(2), 241–250.*

✅ This study discusses how **small samples (n ≈ 20–30)** can sometimes be acceptable for strong correlations (r > 0.5), but to **accurately estimate** correlations (especially moderate ones), **larger samples (n > 50)** are recommended.

- ❑ *Source: Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations with fixed precision. *Psychological Methods*, 5(2), 241–250. https://doi.org/10.1037/1082-989X.5.2.241*

2. Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D., & Newman, T. B. (2013).

Title: *Designing Clinical Research (4th ed.).*

Publisher: Lippincott Williams & Wilkins.

✅ A classic book on study design. They suggest that **for a correlation study**, **30 participants** is a **minimum rough rule of thumb**, but **larger samples** are needed if detecting **weak to moderate correlations**.

- ❑ *Source: Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2013). *Designing clinical research* (4th ed.). Lippincott Williams & Wilkins.*

3. Cohen, J. (1988).

Title: *Statistical Power Analysis for the Behavioral Sciences (2nd ed.).*

Publisher: Routledge Academic.

✅ Cohen provided guidelines for **effect sizes** and **sample size calculations**:

✓ To detect a **small correlation** ($r \approx 0.1$), you need about **783** samples.

✓ For a **medium correlation** ($r \approx 0.3$), you need about **84** samples.

✓ For a **large correlation** ($r \approx 0.5$), around **28 samples** are sufficient.

✓ So, if you're expecting a **large correlation**, your sample size of **n = 20** is **nearly acceptable**, but slightly underpowered for detecting medium to small correlations.

❑ *Source: Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.*

4. Mukaka, M. M. (2012).

Title: *A guide to appropriate use of Correlation coefficient in medical research.*

Journal: *Malawi Medical Journal, 24(3), 69–71.*

✅ Mukaka mentions that for reliable Pearson correlation studies, **larger sample sizes** are preferable but does not give a strict cutoff — just that **n ≥ 30** is commonly aimed for.

❑ *Source: Mukaka, M. M. (2012). A guide to appropriate use of correlation coefficient in medical research. *Malawi Medical Journal, 24*(3), 69–71. https://www.ajol.info/index.php/mmj/article/view/81588*

EXAMPLE SCIENTIFIC RESEARCH USING PEARSON CORRELATION

Healthscope 2021, Vol 4(1): 31-37

RESEARCH ARTICLE

Spatio-Temporal analysis of Dengue distribution pattern in Kuantan from 2015-2019

Mohamad Hizham Mohamad Hanapi, Farah Ayuni Shafie*

Centre of Environmental Health and Safety, Faculty of Health Sciences, Universiti Teknologi MARA Cawangan Selangor Kampus Puncak Alam, 42300 Bandar Puncak Alam, Selangor, Malaysia;

This study wants to prove the linear relationship between the high number of dengue cases to the high amount of rainfall and vice versa.

Table 3. Pearson correlation analysis of dengue cases per month and rainfall amount per month.

Pearson Correlation Results	**Dengue case per month**	**Rainfall amount per month**
R^2 linear	0.003	
Pearson Correlation, r	0.053	
Significance 2-Tailed	0.686	

Furthermore, the Pearson correlation (r) value in this study is not even close to 1. The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no relationship between the two variables. Values greater than 0 indicate a positive relationship; that is, as the value of one variable increases, so does the value of the other variable. These analyses prove that the amount of rainfall does not affect the increase in the number of dengue cases in Kuantan for 2015 to 2019.


TIME FOR
A BREAK!

SPEARMAN RANK CORRELATION

OBJECTIVES

- Explain the concepts and principles of Spearman rank correlation
- Discuss the requirements and application of Spearman rank correlation test
- Perform Spearman rank correlation test manually and using SPSS software, interpret and report the results.

What is Spearman Rank Correlation?

- Spearman Rank Correlation, is a **non-parametric** measure of the **strength and direction** of the association between **two ranked** variables.
- Often denoted as **ρ (rho)** or $\mathbf{r_s}$,
- Is the **Non-parametric** version of the Pearson correlation.

- Determines the strength and direction of the **monotonic relationship** between 2 variables.

- It does not assume a linear relationship or normally distributed data like Pearson correlation.

A monotonic relationship is a relationship that does one of the following:

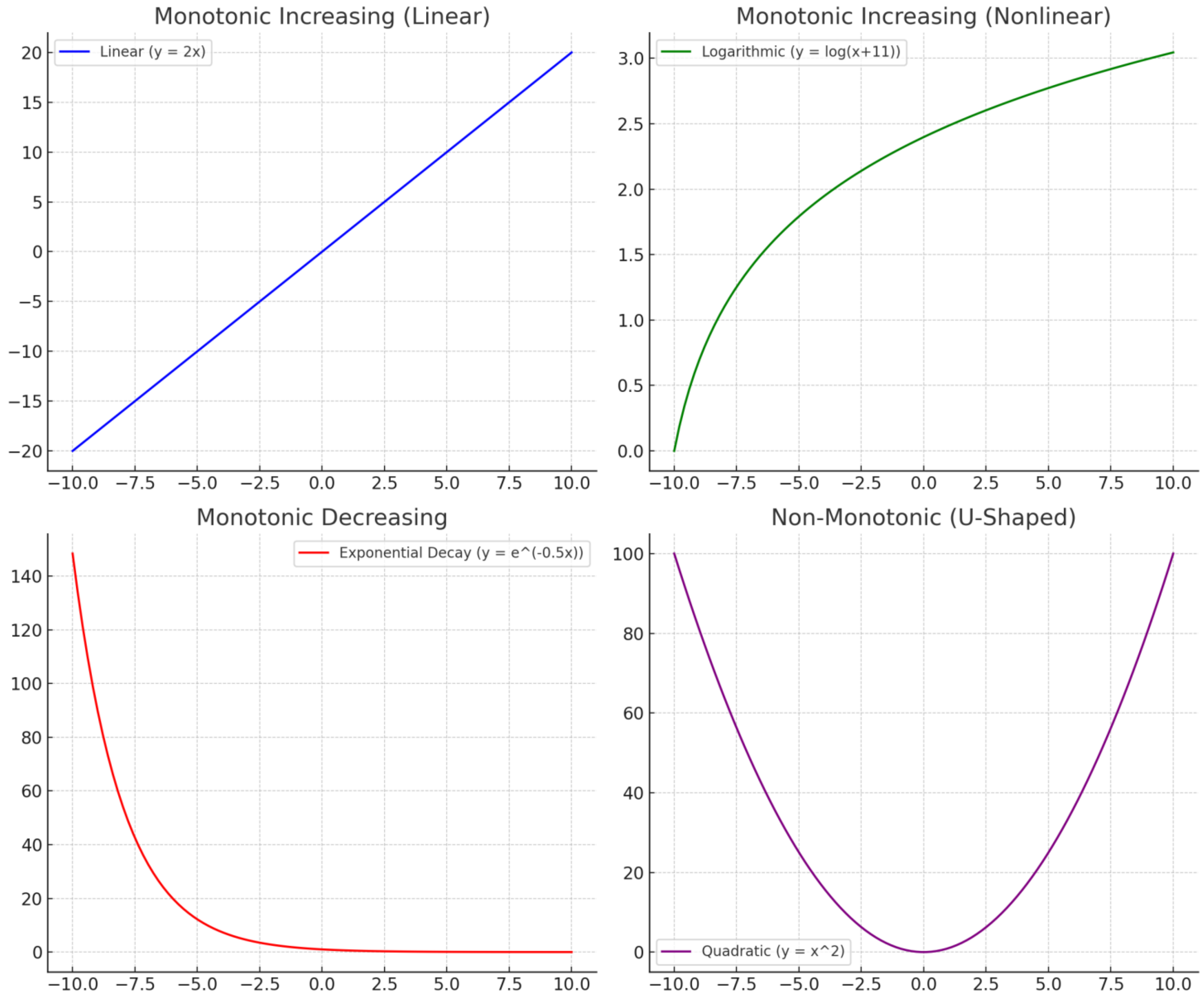
As the value of one variable increases, so does the value of the other variable; OR

As the value of one variable increases, the other variable value decreases.

Monotonicity is “less restrictive” than that of a linear relationship.

Key Point:

A monotonic relationship does **not require a straight-line (linear) relationship** but ensures that the direction of change is consistent.


Monotonic Increasing (Linear)
Linear (y = 2x)
Monotonic Increasing (Nonlinear)
Logarithmic (y = log(x+11))
Monotonic Decreasing
Exponential Decay (y = e^(-0.5x))
Non-Monotonic (U-Shaped)
Quadratic (y = x^2)

Key Principles for Spearman Rank Correlation

1. **Rank-based:** Rather than using the actual data values, Spearman works by ranking the data first.
2. **Monotonic relationship:**

If one variable increases, the other also increases (positive monotonic).

If one variable increases, the other decreases (negative monotonic).

Range:

+1 = perfect positive association

0 = no association

-1 = perfect negative association

Requirements

Aspect	Requirement
Type of Variables	Ordinal, Interval, or Ratio data (but must be converted to **ranks**)
Relationship Type	Monotonic (not necessarily linear)
Sample Size	Ideally $n \geq 10$; larger samples increase reliability
Assumptions	- Observations are independent - Variables should be at least ordinal - Monotonic relationship expected

When to Use Spearman Instead of Pearson?

✅ When your data is ordinal (ranked but not measured).

✅ When the relationship between variables is non-linear but monotonic.

✅ When your data violates normality assumptions (skewed distributions, outliers).

✅ When you want a robust correlation that's less sensitive to outliers.

- ❑Spearman = robust because it only cares about rank order, not actual size differences.

Applications

- ❑**Social Sciences:** Rank relationship between educational level and income satisfaction.
- ❑**Medical Research:** Rank association between severity of disease symptoms and quality of life.
- ❑**Environmental Studies**: Rank pollution levels and biodiversity loss.
- ❑**Marketing:** Rank customer satisfaction scores and product loyalty.

Spearman's rank correlation coefficient, r_s

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

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

Performing Spearman Rank Correlation (Manual Calculation)

- Let us consider the following example data regarding the marks achieved in a maths and English exam:

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

Dataset: english_maths_dataset_for_spearman.xlsx

SCAN ME

Step 1: Rank the English and Math marks separately.

- Higher marks = Higher rank (1 = highest)

English Marks	Rank (English)
80	1
76	2
75	3
71	4
64	5
61	6.5 (tie)
61	6.5 (tie)
58	8
56	9
45	10

Maths Marks	Rank (Maths)
77	1
70	2
67	3
66	4
65	5
63	6
60	7
59	8
56	9
40	10

Step 2: Create a Full Table

Student	English Marks	Rank E	Maths Marks	Rank M	d = (Rank E - Rank M)	d^2
1	56	9	66	4	5	25
2	75	3	70	2	1	1
3	45	10	40	10	0	0
4	71	4	60	7	-3	9
5	61	6.5	65	5	1.5	2.25
6	64	5	56	9	-4	16
7	58	8	59	8	0	0
8	80	1	77	1	0	0
9	76	2	67	3	-1	1
10	61	6.5	63	6	0.5	0.25

Step 3: Sum d^2

$$\sum d^2 = 25 + 1 + 0 + 9 + 2.25 + 16 + 0 + 0 + 1 + 0.25 = 54.5$$

Step 4: Apply the Spearman Rank Correlation Formula

The formula:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Where:

- $n = 10$ (number of students)

Substituting:

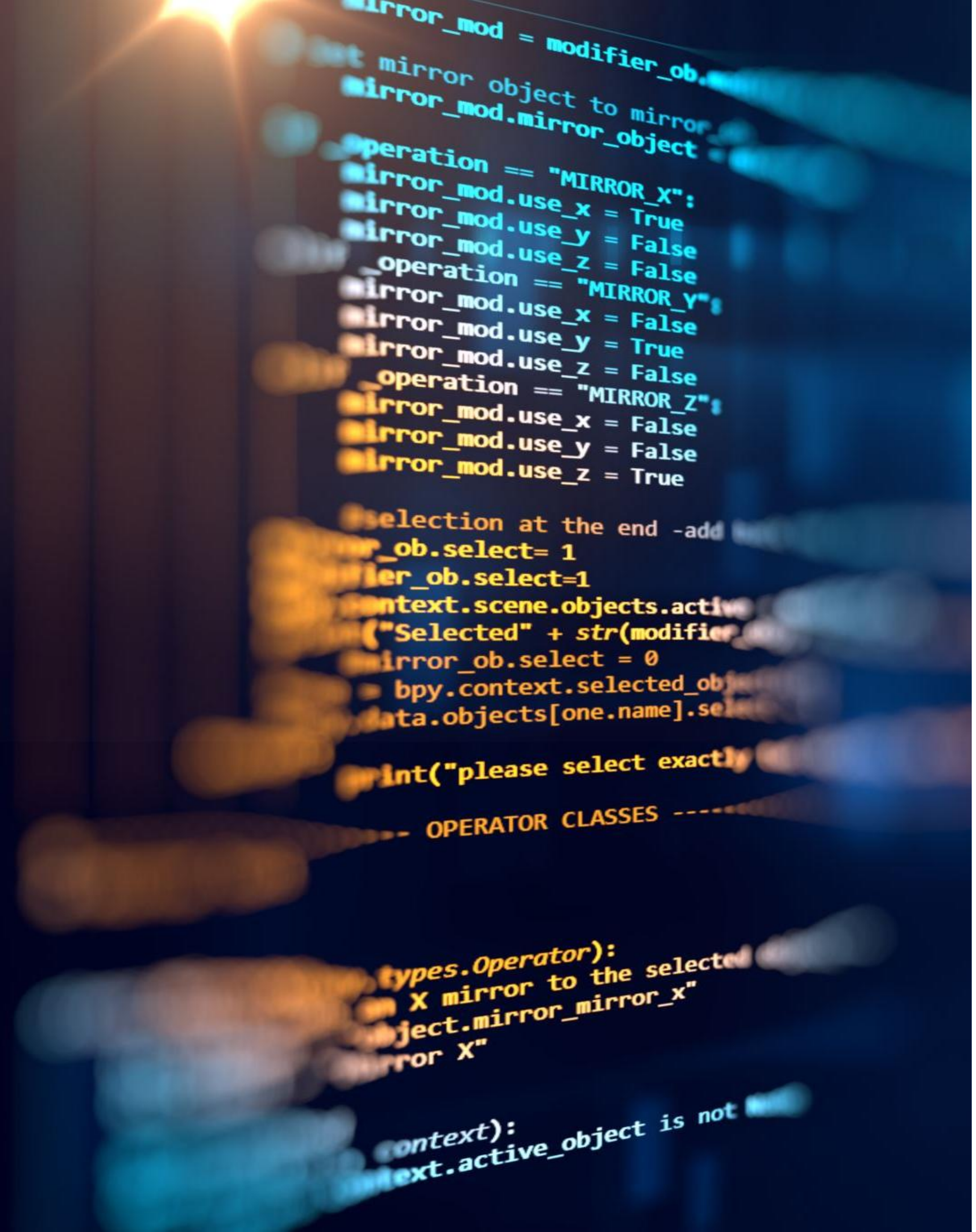
$$r_s = 1 - \frac{6(54.5)}{10(10^2 - 1)} = 1 - \frac{327}{10(99)} = 1 - \frac{327}{990} = 1 - 0.33 = 0.67$$

✅ Spearman's rank correlation coefficient $r_s = 0.67$

Interpretation:

- r_s=0.67 means a **moderate to strong positive** correlation between English and Math marks.
- Students who perform well in English tend to also perform well in Math, although not perfectly.

✅ Spearman rank correlation coefficient is r_s =0.67 , indicating a **moderate positive relationship** between Math and English exam marks.



Steps to Enter and Analyze in SPSS

Step 1: Open SPSS

- Open a new dataset.

Step 2: Define Variables

- In Variable View:
- First variable name: English (Type: Numeric)
- Second variable name: Math (Type: Numeric)

Step 3: Enter / Import Data

Step 4: Perform Spearman Correlation

1. Click **Analyze ➔ Correlate ➔ Bivariate**.

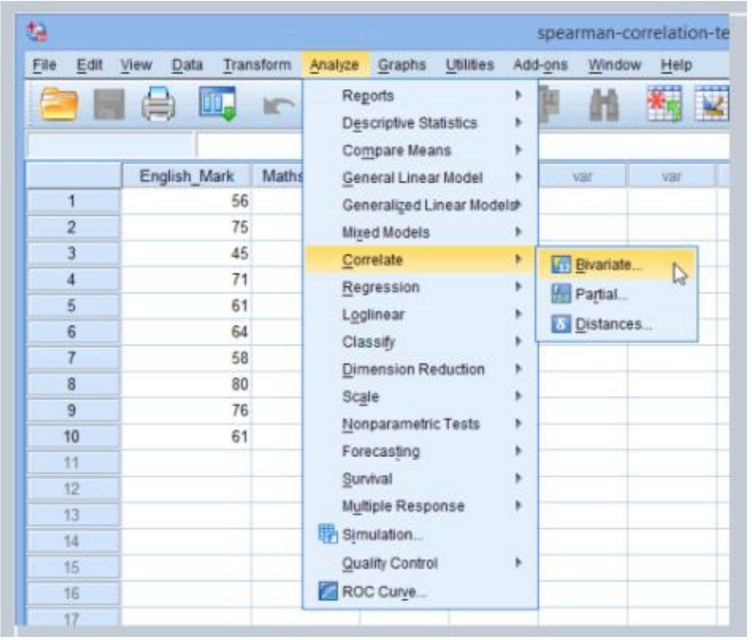
2. Move both variables (English and Math) into the Variables box.

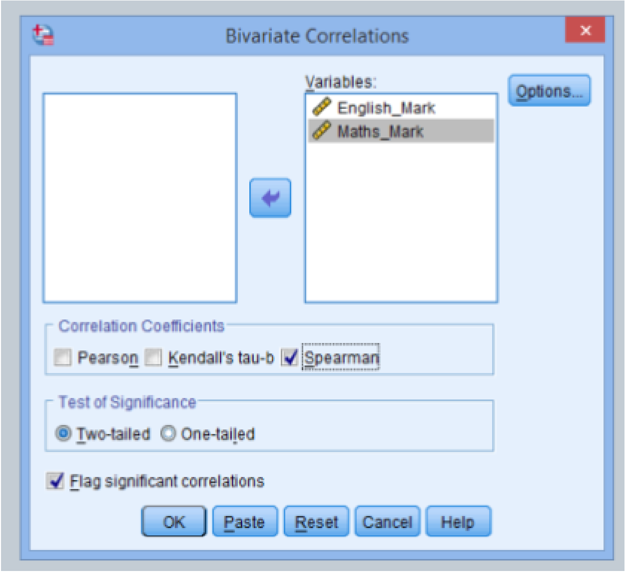
3. In the Correlation Coefficients section:
   - **Tick Spearman**.
   - Untick Pearson.

4. Ensure "Two-tailed" is selected under Test of Significance.

5. Click OK.

✅ SPSS will run the test and give you an output table.





SPSS Output

Correlations

			English	Maths
Spearman's rho	English	Correlation Coefficient	1.000	.669*
		Sig. (2-tailed)	.	.035
		N	10	10
	Maths	Correlation Coefficient	.669*	1.000
		Sig. (2-tailed)	.035	.
		N	10	10

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

r_s =0.670 (matches our manual calculation)
SPSS will also show the p-value
(e.g., if $p < 0.05$, the correlation is significant).

How to Report?

A Spearman rank-order correlation was conducted to assess the relationship between English and Math exam marks.

There was a **moderate positive correlation** between English and math marks, which was statistically significant, r_s=0.669, p<0.05.

Interpretation: The higher you ranked in math, the higher you ranked in English also, and vice versa.

SUMMARY

Pearson (r)

Measures the **strength and direction of a linear relationship** between **two continuous variables**.

🔁 **Direction**:

- Positive (↑↑): both variables increase together
- Negative (↑↓): one increases as the other decreases

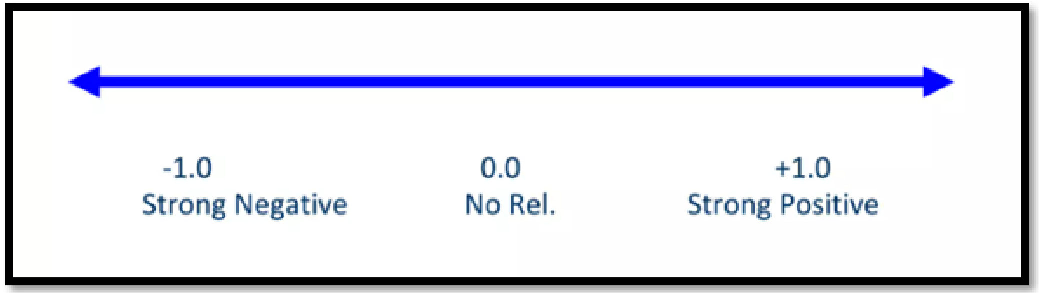
📐 Pearson Assumptions

✅ Continuous variables (interval/ratio scale)
✅ Linear relationship (check scatterplot)
✅ Normal distribution
✅ No extreme outliers
🔁 Homoscedasticity (equal spread of residuals)

Spearman (ρ or r_s)

Measures the **strength and direction of a monotonic relationship** using **ranked data**.

- ✅ Non-parametric, robust to outliers
- 📌 Suitable when data is **ordinal**, **non-linear**, or **non-normal**



THANK YOU