Correlation

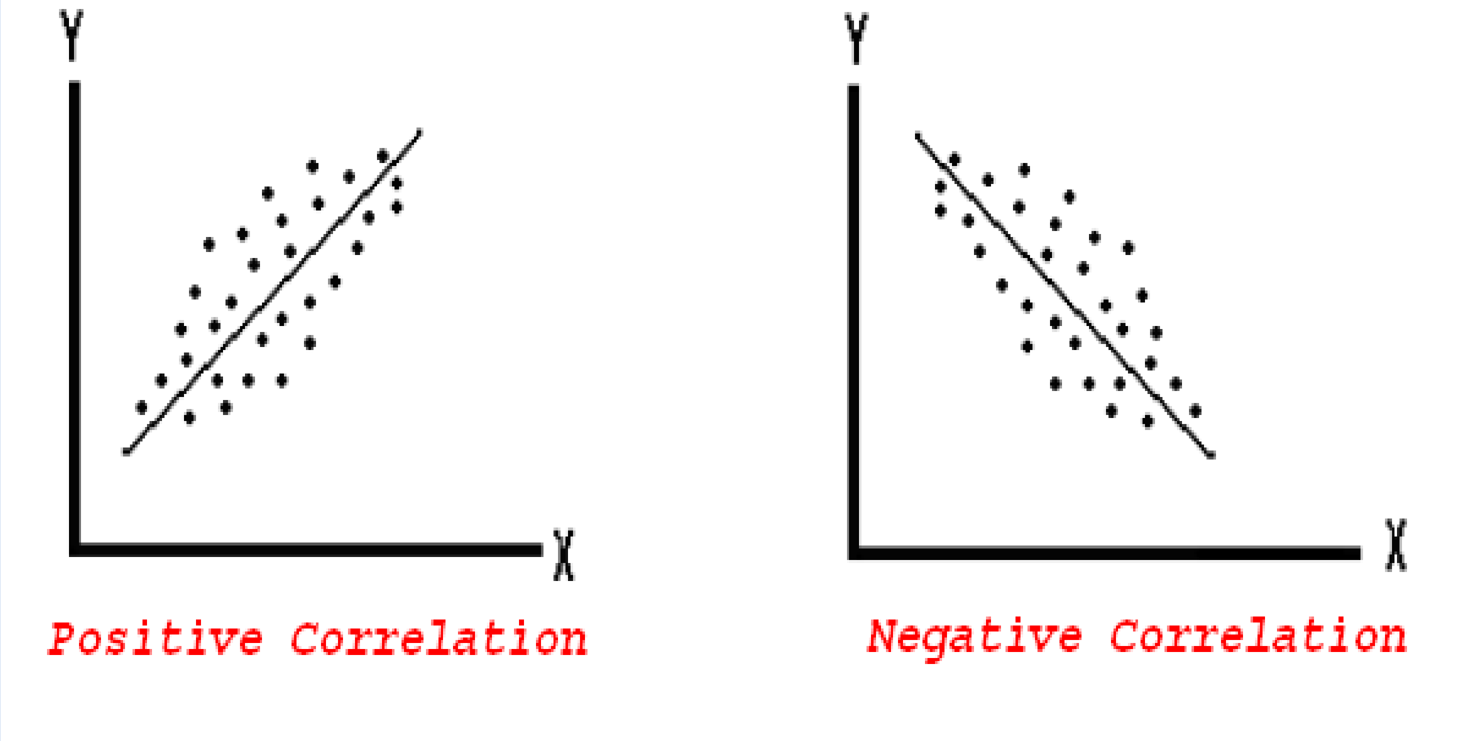
Correlation

Correlation is observed when two Variables either increase together or decrease together in a roughly linear pattern

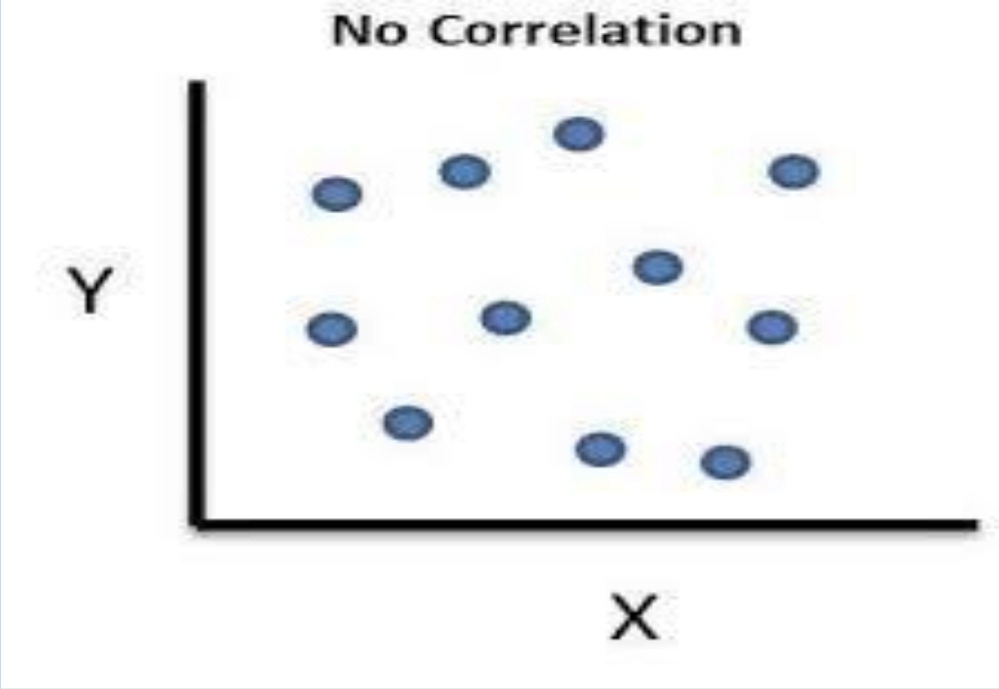
- Correlation is Negative when larger values for one Variable are paired with smaller numbers of the other. Positive Correlation is the opposite – the values of both Variables grow together

Correlation

Scatter



Correlation



Correlation

Covariance

- Covariance is a measure of how changes in one variable are associated with changes in a second variable

- The covariance measures the degree to which two variables are linearly associated

Correlation

For a single Variable, x, Variance is a measure of Variation of the values of x in the data about their Mean, (symbol $\bar{X}$ for a Sample, or μ for a Population.

Covariance is a measure of Variation of the values of (**2-PAIRED Variable**) data points (x, y)'s about the point made up of the Means of x and y – the point $(\bar{x}, \bar{y})$.

So, we can think of Covariance as a 2-Variable counterpart to the Variance

Correlation

Variance (1 Variable) Formulas

$$\text{Sample: } s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \qquad \text{Population: } \sigma^2 = \frac{\sum (x - \mu_x)^2}{N}$$

where n and N are the Sample Size and Population Size, respectively

Correlation

Covariance (2 Variable) Formulas

$$\underline{\text{Sample}}: \text{Cov}(x, y) = s_{xy} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1}$$

$$\underline{\text{Population}}: \text{Cov}(x, y) = \sigma_{xy} = \frac{\sum(x - \mu_x)(y - \mu_y)}{N}$$

Correlation

Covariance of Height (inches) and Weight (pounds)							
Individual	Height(x)	Weight(y)		x-Mean(x)	y-Mean(y)		Product
#1	70	180		2.3	21		48.3
#2	65	125		−2.7	−34		91.8
#3	67	140		−0.7	−19		13.3
#4	71	195		3.3	36		118.8
#5	62	105		−5.7	−54		307.8
#6	73	210		5.3	51		270.3
#7	68	190		0.3	31		9.3
#8	65	110		−2.7	−49		132.3
#9	70	200		2.3	41		94.3
#10	66	135		−1.7	−24		40.8
Total	677	1590					
Means	67.7	159.0		Sum of Products:			1127.0
Divide the Sum by $n-1=9$ to get the Covariance: 125.2 inch-pounds							

Correlation

Covariance of Height (meters) and Weight (kilograms)							
Individual	Height (x)	Weight (y)		x-Mean(x)	y-Mean(y)		Product
#1	1.8	81.7		0.1	9.5		0.6
#2	1.7	56.8		–0.1	–15.4		1.1
#3	1.7	63.6		0.0	–8.6		0.2
#4	1.8	88.5		0.1	16.3		1.4
#5	1.6	47.7		–0.1	–24.5		3.5
#6	1.9	95.3		0.1	23.2		3.1
#7	1.7	86.3		0.0	14.1		0.1
#8	1.7	49.9		–0.1	–22.2		1.5
#9	1.8	90.8		0.1	18.6		1.1
#10	1.7	61.3		0.0	–10.9		0.5
Total	17.2	721.9					
Mean	1.72	72.2		Sum of Product:			13.0
Divide the Sum by $n - 1 = 9$ to get the **Covariance: 1.4 meter-kilograms**							

Correlation

Covariance **cannot** tell us the strength of the Correlation

One thing we can say from both sets of measurements above is that there is a positive Correlation. That is, as height increases, weight also increases.

So, we can use the sign of these numbers (positive) to tell us the direction of Correlation (positive).

Correlation

But how good is this Correlation? How strong is it? We can’t use the values of the numbers, because the units are meaningless and we would have to make an arbitrary choice between whether the strength was 125.2 or 1.4

Correlation

The Covariance is a Statistic or a Parameter which can tell us the **DIRECTION** of a Correlation between two paired Variables, x and y, from data consisting of (x, y) pairs

Correlation

So the numerical values of the Covariance are not used. We only use the sign – positive or negative – of the Covariance to tell us the direction of the correlation

Correlation

REMEMBER

Correlation is not Causation

Correlation

When **normalized** or **standardized**, the Covariance becomes the Correlation Coefficient, a measure of the **direction** and **strength** of the Correlation

Correlation

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

r **is also known as "Pearson's r"**

or

The "Pearson product-moment correlation coefficient"

Correlation

- **r is a unit-less number**
- The Correlation Coefficient, r, ranges from −1 to +1.
- r = 0 indicates no Correlation.
- r = −1 and r = +1 indicate a perfect negative or positive Correlation, respectively. But perfection almost never happens

Correlation

Evidence of Correlation	e.g., Less Rigorous Standard	e.g., More Rigorous Standard
very strong	0.7 – 1.0	0.81 – 1.00
strong	0.5 – 0.7	0.61 – 0.80
moderate	0.3 – 0.5	0.41 – 0.60
weak	0.1 – 0.3	0.21 – 0.40
none	0.0 – 0.1	0.00 – 0.20

Correlation

Assumptions Pearson's correlation test:

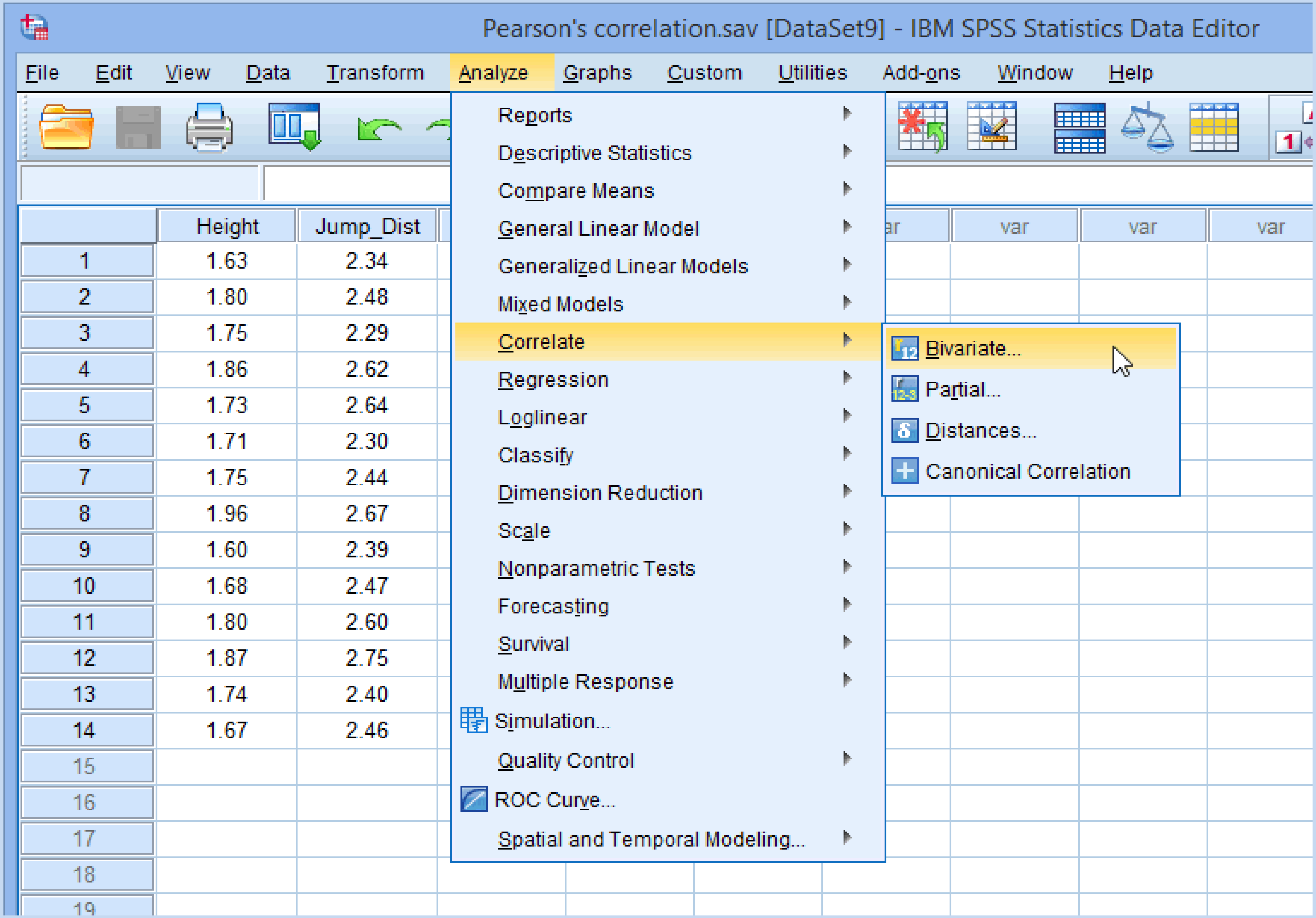
1- Correlation between 2 quantitative variables

2- Normally distributed

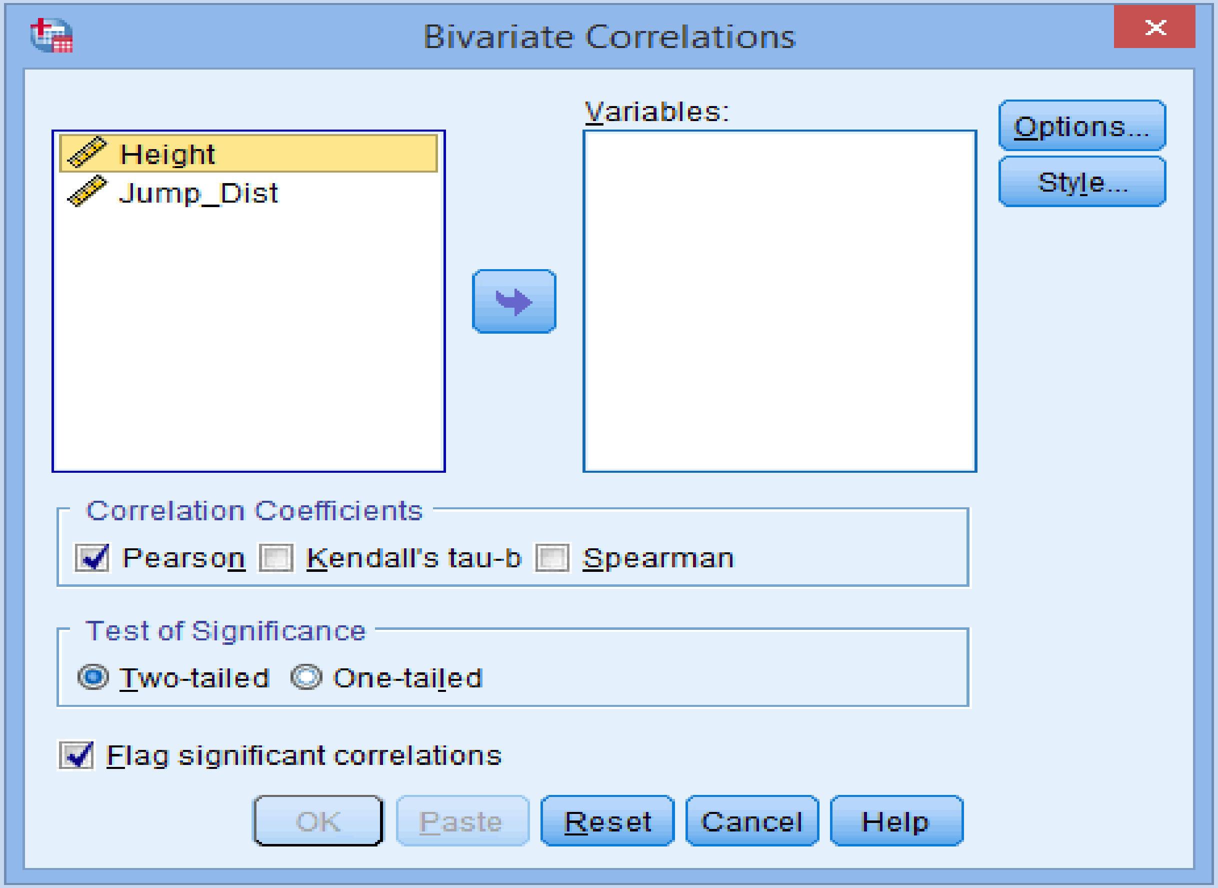
3- There is a linear relationship between the two variables

4- Outliers are either kept to a minimum or are removed entirely

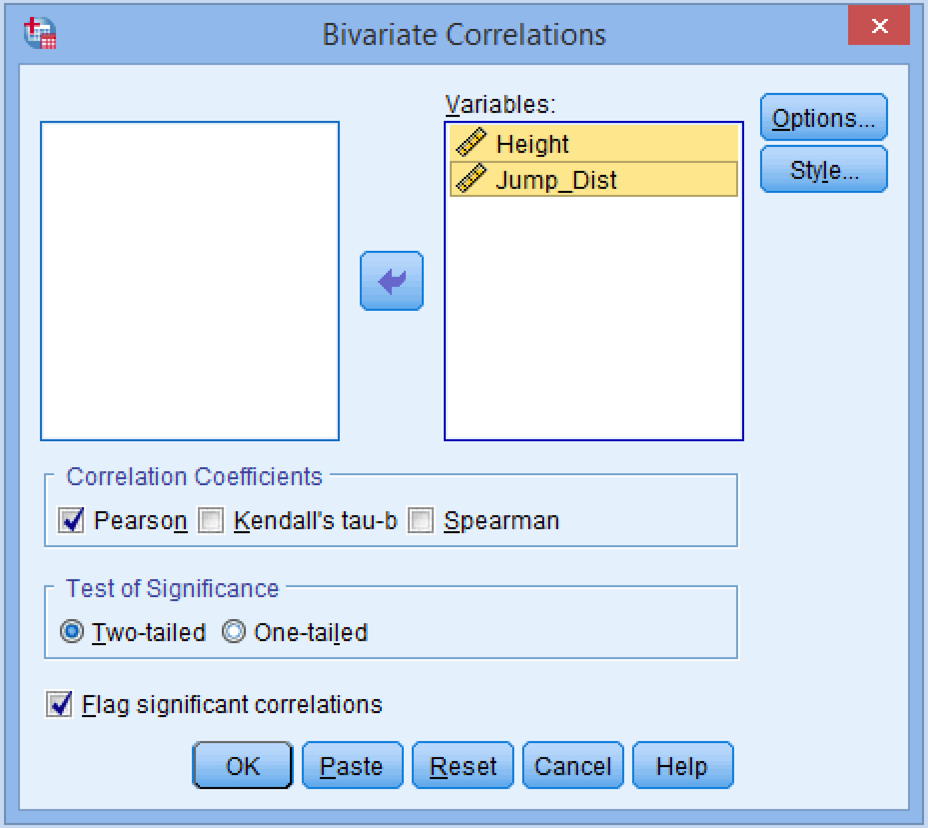
Correlation



Correlation



Correlation



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).

Correlation

The **Spearman's rank-order correlation** is the nonparametric version of the Pearson product-moment correlation.

Spearman's correlation coefficient, (ρ, also signified by r_s) measures the **strength** and **direction** of association **between two ranked variables**

Correlation

- H_0: There is no association (correlation) between the two variables [in the population]

- H_A: There is an association (correlation) between the two variables

Correlation

- The prerequisite is that the 2 variables are ordinal

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

Correlation

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

Correlation

Correlation

Correlations

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

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

Correlation

- $\rho = 0.67$, $p = 0.033$

Correlation

Kendall's tau-b (T_b) correlation coefficient (Kendall's tau-b, for short) is a nonparametric measure of the strength and direction of association that exists between **two variables** measured on at least an **ordinal scale**

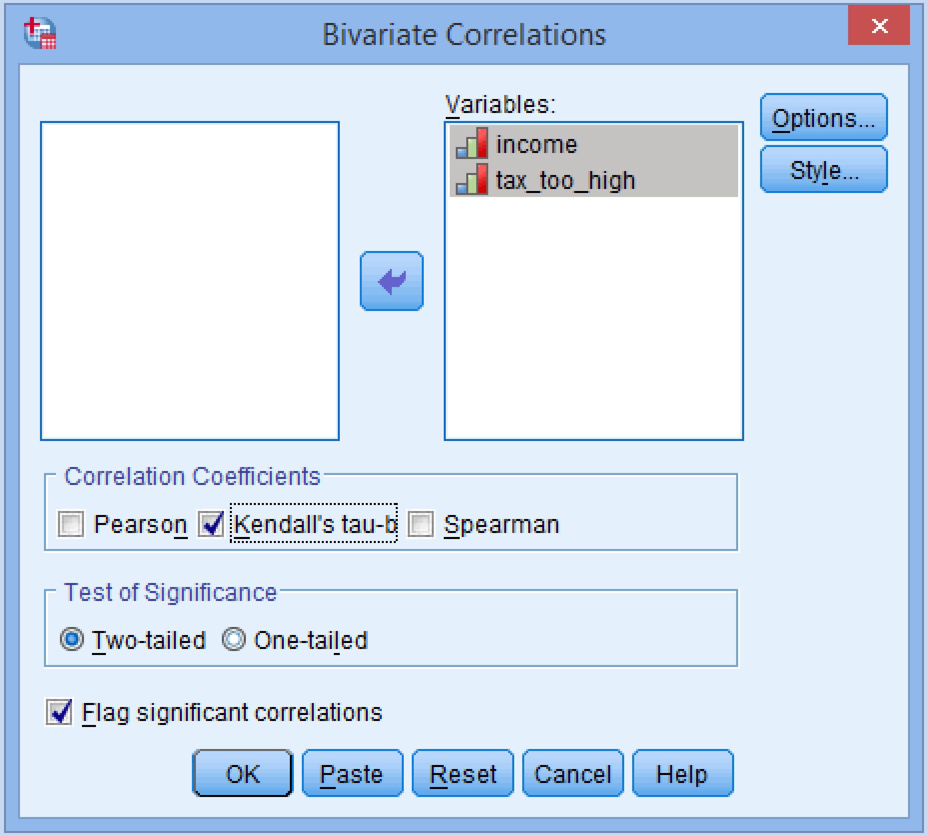
It is considered **a nonparametric alternative to the Pearson's product-moment correlation** when your data has failed one or more of the assumptions of this test

Correlation

EXAMPLE

- Exam grade and Time spent revising
- Exam grades – A, B, C, D, E and F – and
- Revision time was split into five categories: less than 5 hours, 5-9 hours, 10-14 hours, 15-19 hours, and 20 hours or more

Correlation



Correlation

Correlations

			income	tax_too_high
Kendall's tau_b	income	Correlation Coefficient	1.000	.535**
		Sig. (2-tailed)	.	.003
		N	24	24
	tax_too_high	Correlation Coefficient	.535**	1.000
		Sig. (2-tailed)	.003	.
		N	24	24

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

X	Y
72	45
73	38
75	41
76	35
77	31
78	40
79	25
80	32
80	36
81	29
82	34
83	38
84	26
85	32
86	28
88	27

Variable X

Exam grade: 70 – < 75 = C

75 - < 80 = B

80 - 90 = A

Variable B

Time spent studying :

25 hours - < 34 hours

35 hours - < 45 hours