



Topic materials:

Dr Raquel Iniesta

Department of Biostatistics and
Health Informatics



Narration and contribution:

Zahra Abdula

Improvements:

Nick Beckley-Hoelscher

Kim Goldsmith

Sabine Landau

Institute of Psychiatry, Psychology and Neuroscience

03/08/2020

Module Title: Introduction to Statistics

Session Title: Correlation

Topic title: Correlation and Linear Regression



Learning Outcomes

- Understand the features of a Pearson's correlation coefficient and when to use it
- Understand the features of a Spearman's Rank correlation coefficient and when to use it
- Understand 'Spurious' correlation

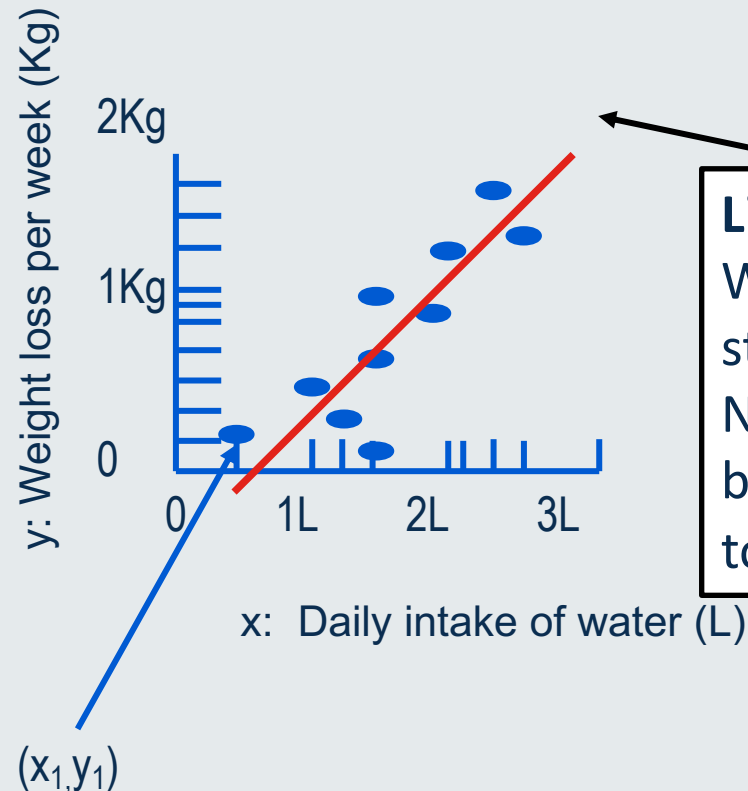
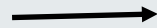


Previously on 'Introduction to Statistics'

10 people were studied for the Hypothesis 'The higher the intake of water, the higher the weight loss'.

- Plotting the data is essential to understand and visually assess the relationship between pairs of continuous variables
- The plot of data points (x,y) with x and y being continuous is called a **scatterplot**

	x	y
(x_1, y_1)	0.5	0.10
(x_2, y_2)	1.0	0.30
(x_3, y_3)	1.2	0.40
...



Linear relationship:

We can draw a straight line.
Not perfect fit, but the line is "close" to the points

Correlation

We need an objective measure of strength of a linear relationship.

Correlation ' r ' is a statistical concept that refers to how close two variables are to having a linear relationship with each other, or in other words, the strength of their linear relationship. Correlation ' r ' is a method to quantify the **Direction** and **Magnitude**, of linear association between two continuous variables.

' r ' belongs to the range $[-1,1]$

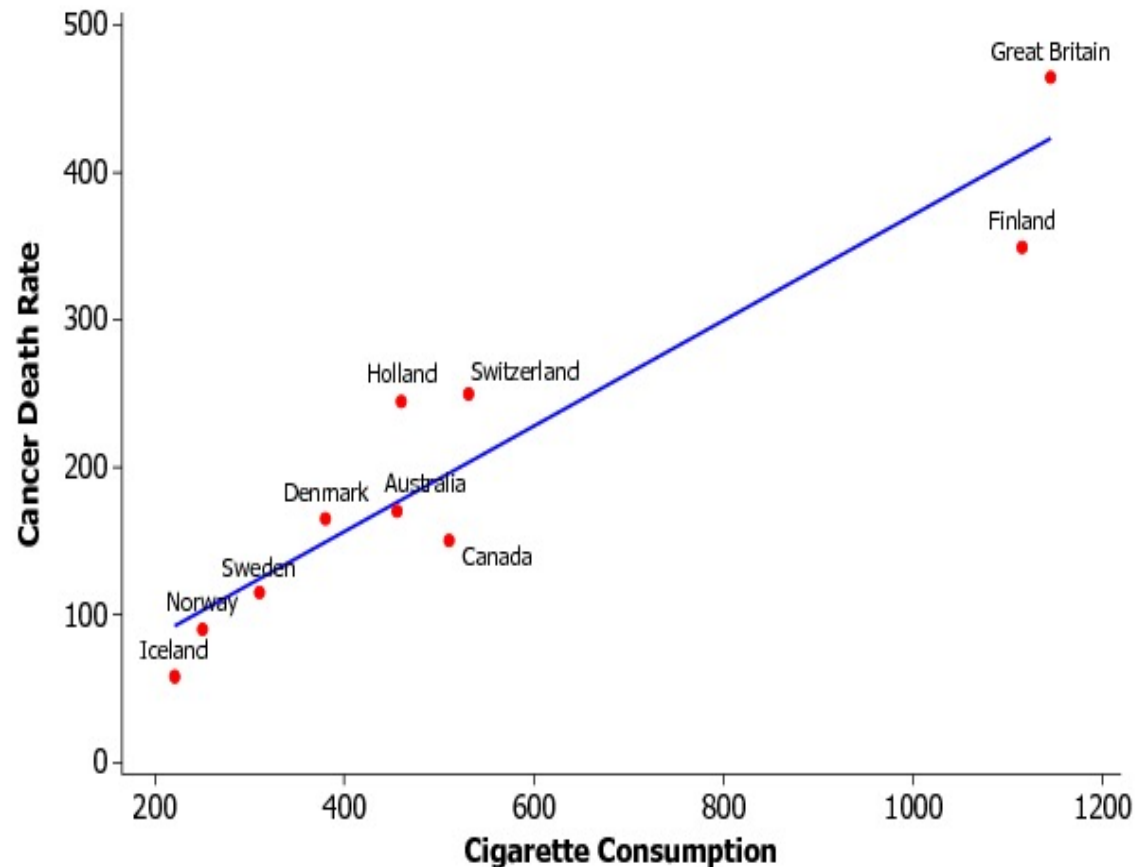
There are two types of correlation coefficients:

- Pearson's Correlation Coefficient (Parametric approach)

- Spearman's Correlation Coefficient (non-Parametric approach)

Example

Cigarette consumption (in 1930, average number of packets per year), Lung cancer death rate (in 1950)



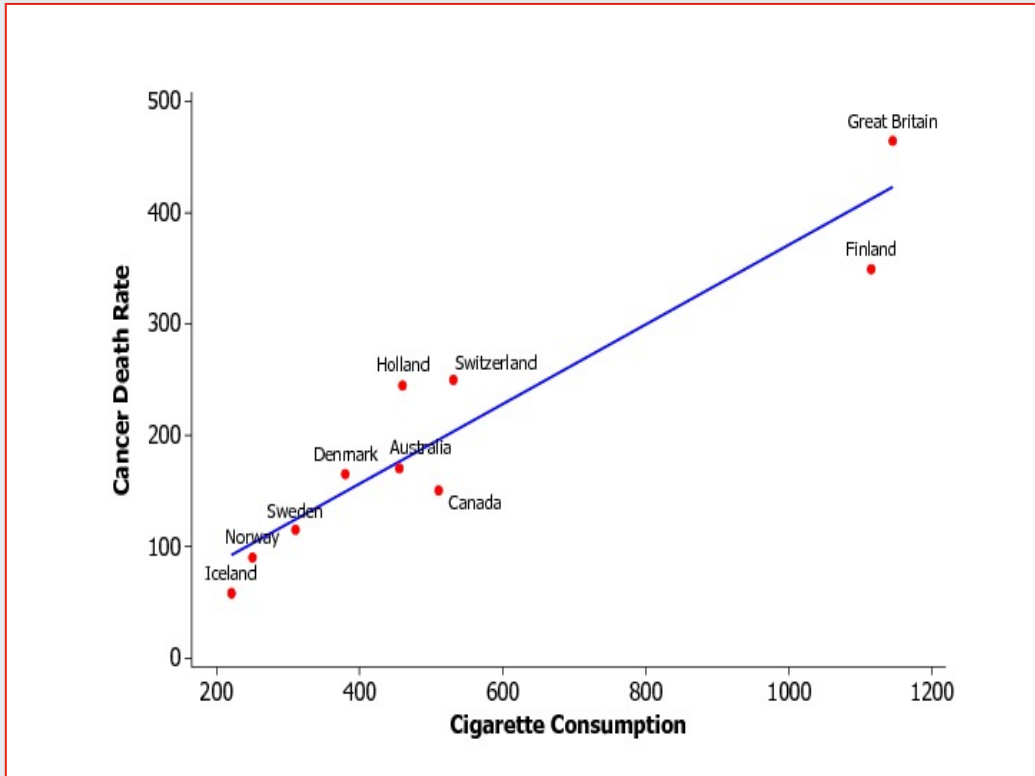
Here we show a scatterplot between country level: Cigarette Consumption and Cancer Death Rate

In addition to the scatterplot we fit a line of best fit through the centre of the data

The line indicates an increase in Cigarette Consumption is associated with an increase in Cancer Death Rate

The line and points are close, but we need a measure of the magnitude of this linear relationship.

Example



Direction of effect

The co-efficient is **positive**, thus a country with increased cigarette smoking is associated with an increased Cancer Death Rate

Magnitude of effect

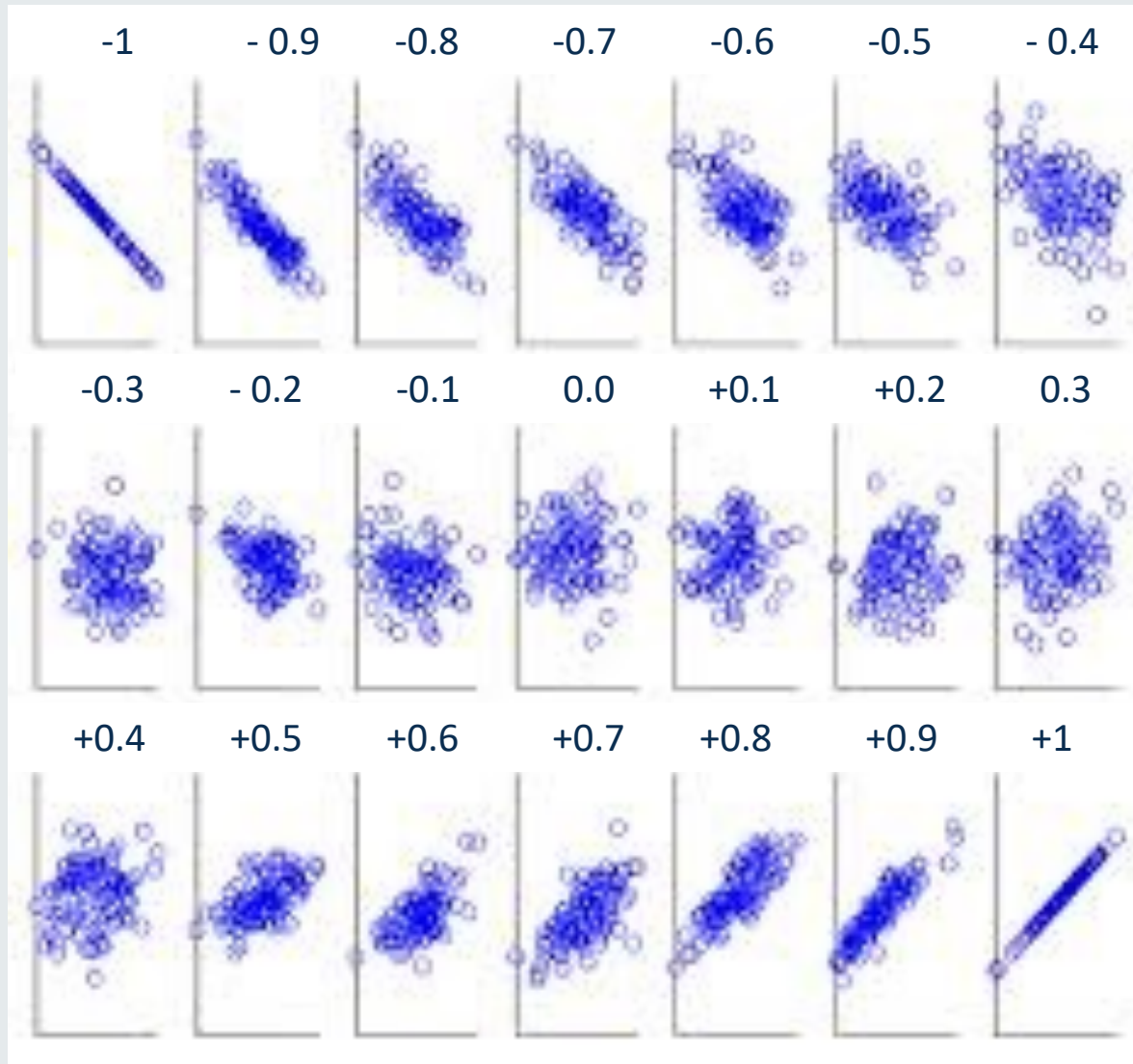
The **magnitude** of the correlation coefficient is **0.73**, thus, there is **strong** correlation.

Linear association

The points follow the line of best fit in a linear manner, thus there is **linear association**

There is strong, positive, linear association between country level cigarette consumption (in 1930) and Cancer Death Rate (in 1950) ($r = 0.73$).

Direction and Strength of 'r'



Range of correlation coefficients	Degree of Correlation
0.80 to 1.00	Very strong positive
0.60 to 0.79	Strong positive
0.40 to 0.59	Moderate positive
0.20 to 0.39	Weak positive
0.00 to 0.19	Very weak positive - none
-0.19 to 0.00	Very weak negative - none
-0.39 to -0.20	Weak negative
-0.59 to -0.40	Moderate negative
-0.79 to -0.60	Strong negative
-1.00 to -0.80	Very strong negative

Direction of effect

The co-efficient is **positive** or **negative**

Magnitude of effect

The **magnitude** of the correlation coefficient ranges from -1 to 1, the closer to ± 1 the stronger the effect

Pearson's Correlation Coefficient 'r'

When to use it

- To check the magnitude and direction of a linear relationship between two variables.

Hypotheses:

- H_0 : the correlation in the population equals to 0
- H_a : the correlation in the population does not equal to 0

Assumptions:

- Variables should be approximately normally distributed.
- Each variable should be continuous.
- Each participant or observation should have a pair of values
- No significant outliers in either variable
- Linearity, a “straight line” relationship between the variable should be formed

Pearson's Correlation Coefficient 'r'

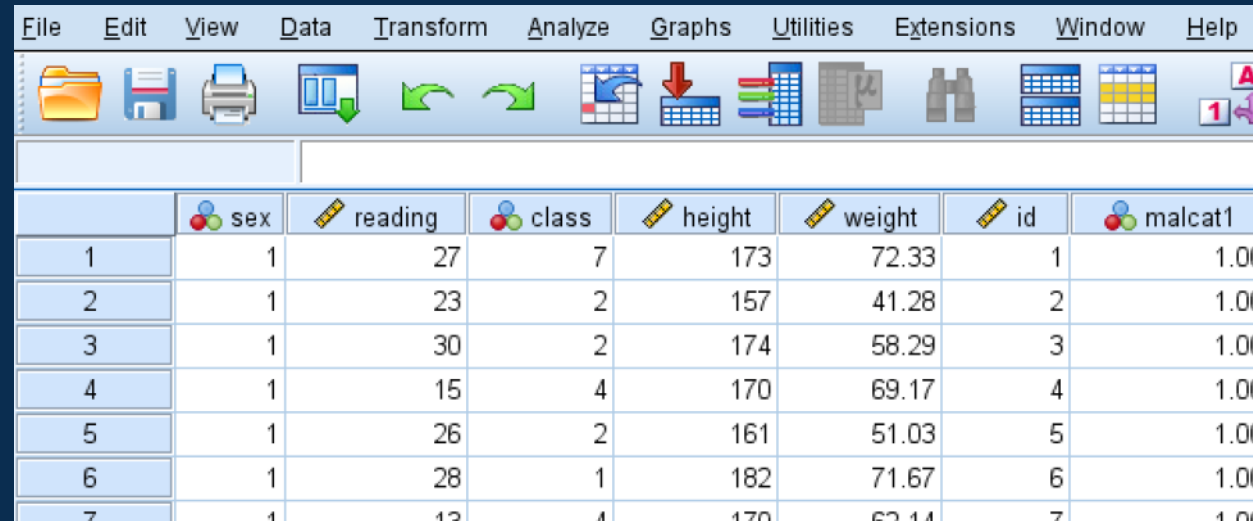
Correlation can be measured using the **Pearson's correlation coefficient 'r'**, defined as

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

where s_x is the st. dev, \bar{x} is the mean of x_i
and similarly for s_y and \bar{y}

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_6a_data.sav**.



	sex	reading	class	height	weight	id	malcat1
1	1	27	7	173	72.33	1	1.00
2	1	23	2	157	41.28	2	1.00
3	1	30	2	174	58.29	3	1.00
4	1	15	4	170	69.17	4	1.00
5	1	26	2	161	51.03	5	1.00
6	1	28	1	182	71.67	6	1.00
7	1	13	4	170	62.14	7	1.00

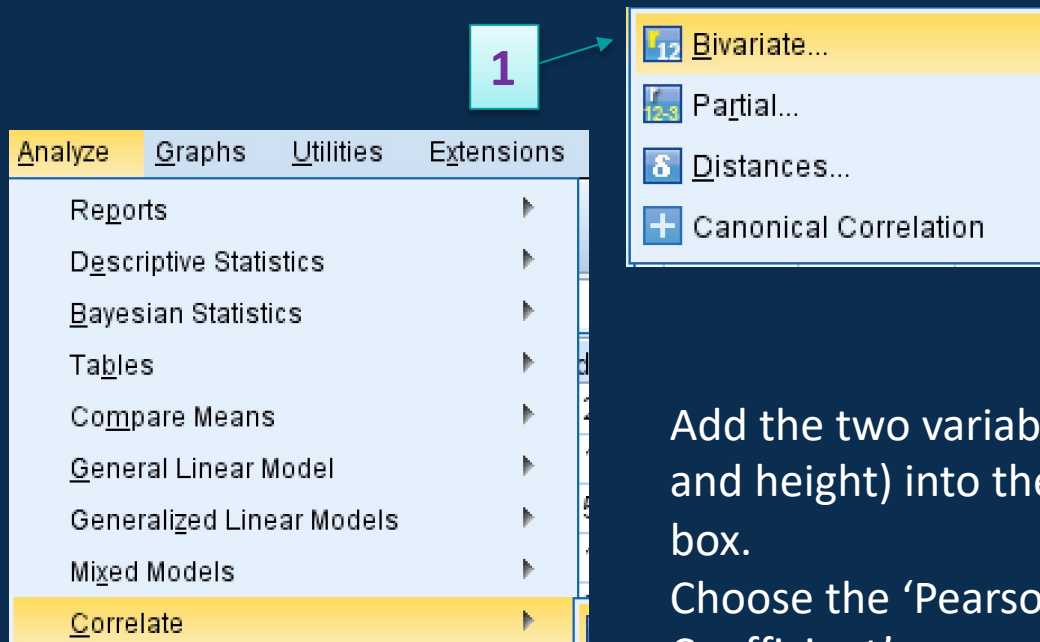
The dataset contains data from 1000 individuals, from the National Child Development Study (NCDS) with respect to their

- **sex**: gender of child (1=male, 2=female)
- **height** : height in cm at age 16
- **weight**: weight in kg at age 16
- **reading**: reading score
- **malcat1**: incidence of malaise at 22 years (0=yes, 1 = No)

SPSS Slide: 'how to'

According to the researchers, in the population from which our data came, they believe there is a relationship between weight and height of the 16 year old children and want to understand the direction and magnitude of the relationship.

Step 1: Calculate a correlation coefficient for variables 'height' and 'weight' from the data

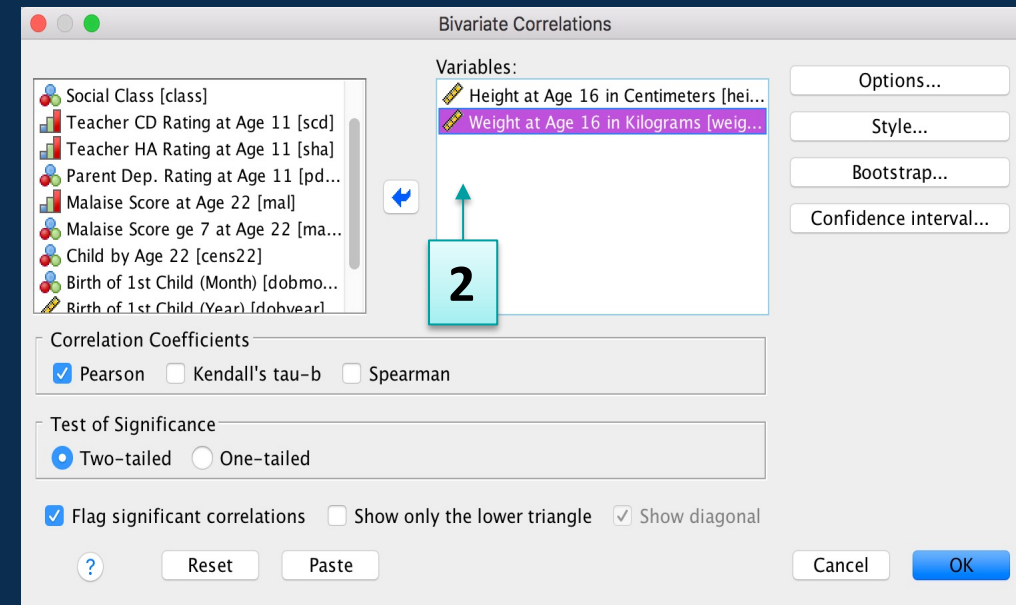


Use 'Analyze' -> 'Correlate' -> 'Bivariate'

Add the two variables (weight and height) into the 'Variables' box.

Choose the 'Pearson' 'Correlation Coefficient'

Click 'Ok'



Output and Interpretation Slide

		Correlations	
		Height at Age 16 in Centimeters	Weight at Age 16 in Kilograms
Height at Age 16 in Centimeters	Pearson Correlation	1	.520**
	Sig. (2-tailed)		.000
	N	1000	1000
Weight at Age 16 in Kilograms	Pearson Correlation	.520**	1
	Sig. (2-tailed)	.000	
	N	1000	1000

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

There is a positive moderate correlation ($r=0.52$) between the height and weight of children aged 16. The correlation coefficient is significantly different from 0 ($p<0.001$) so we can extrapolate the moderate linear relationship observed in the sample, to the whole population.



Spearman's Correlation Coefficient 'r_s'

When to use it?

When **one or both of the variables** are not **normally distributed**. This concept of correlation is less sensitive to extreme influential points, so it should be used in the case of non normality.

What it measures?

- The strength and direction of the **monotonic** relationship between two variables.
- **A monotonic** relationship is a relationship varying in such a way that when one variable decreases or increases the other variable also decreases or increases (but not necessarily at a constant rate, as it does a linear relationship for which we use the Pearson correlation)

Hypotheses:

- H₀: the correlation in the population equals to 0
- H_a: the correlation in the population does not equal to 0

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

Spearman's Correlation Coefficient 'r_s'

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

There are two methods to calculate Spearman's correlation depending on whether: (1) your data **does not have tied ranks** or (2) your data has **tied ranks**.

The formula for when there are no tied ranks is:

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

where d_i is the difference in paired ranks and n = number of cases

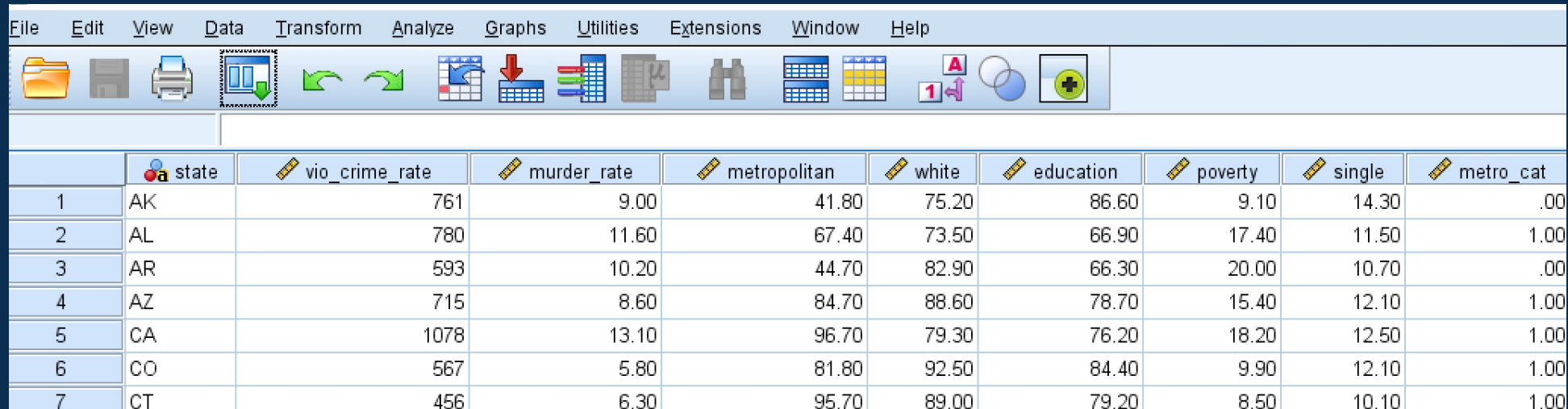
The formula for when there are tied ranks is:

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

where i is the paired score

SPSS Slide

Download the data that we are going to use during the lecture. The dataset is the **lecture_6b_data.sav**.



The screenshot shows the SPSS Data Editor window. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions, Window, and Help. The toolbar contains icons for file operations, data manipulation, and analysis. The data grid displays the following variables and values:

	state	vio_crime_rate	murder_rate	metropolitan	white	education	poverty	single	metro_cat
1	AK	761	9.00	41.80	75.20	86.60	9.10	14.30	.00
2	AL	780	11.60	67.40	73.50	66.90	17.40	11.50	1.00
3	AR	593	10.20	44.70	82.90	66.30	20.00	10.70	.00
4	AZ	715	8.60	84.70	88.60	78.70	15.40	12.10	1.00
5	CA	1078	13.10	96.70	79.30	76.20	18.20	12.50	1.00
6	CO	567	5.80	81.80	92.50	84.40	9.90	12.10	1.00
7	CT	456	6.30	95.70	89.00	79.20	8.50	10.10	1.00

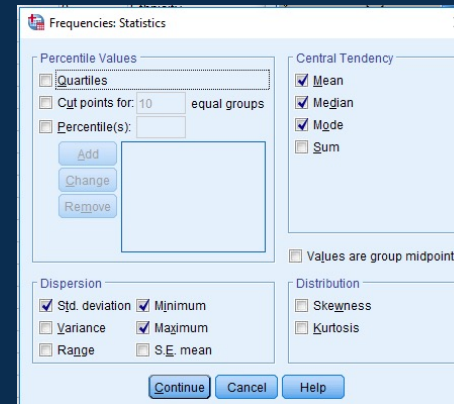
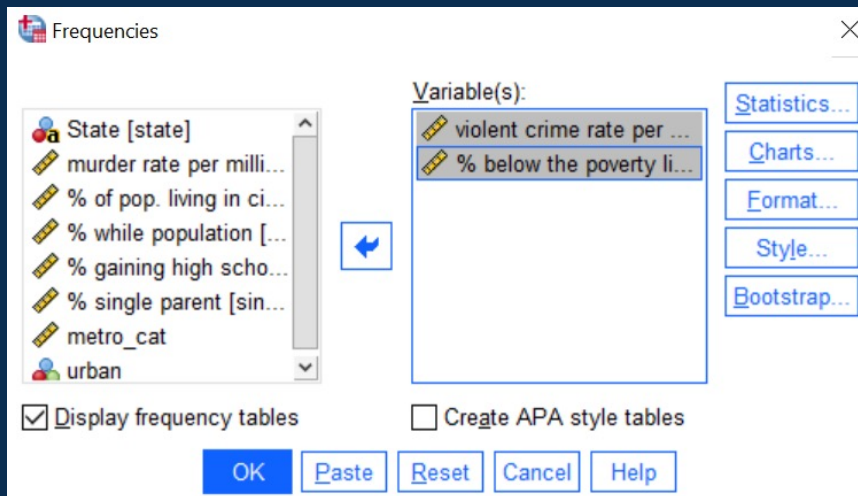
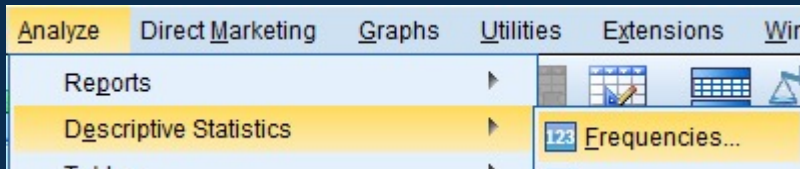
The dataset contains data from 51 US states, measuring the crime rates and background measures for each state with respect to their

- **violent crime:** per 100,000 population
- **murder:** per 100,000 population
- **poverty:** percent below the poverty line
- **single:** percentage of lone parents

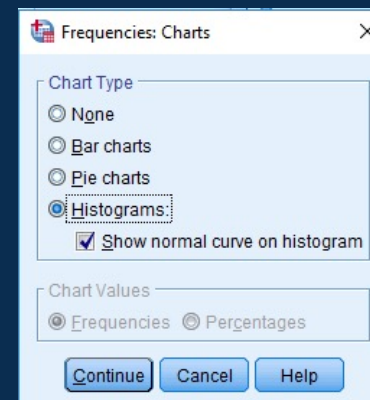
SPSS Slide: 'how to'

According to the researchers, in the population from which our data came, they believe there is a relationship between violent crime measured per 100,000 and the percentage of people below the poverty line per 100,000.

Step 1: Check the suitability of the data.



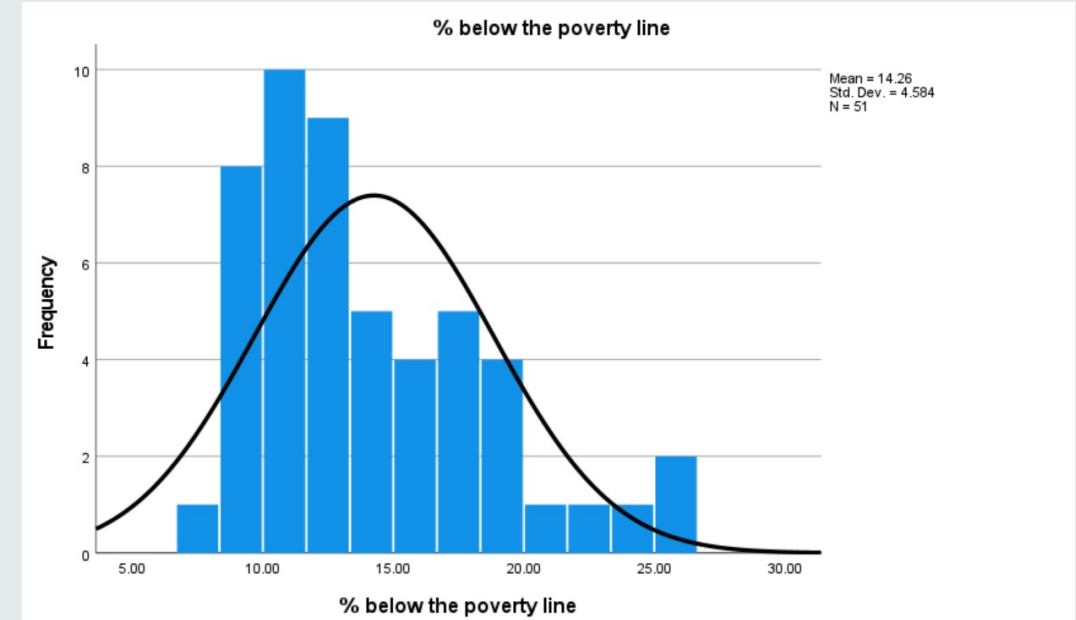
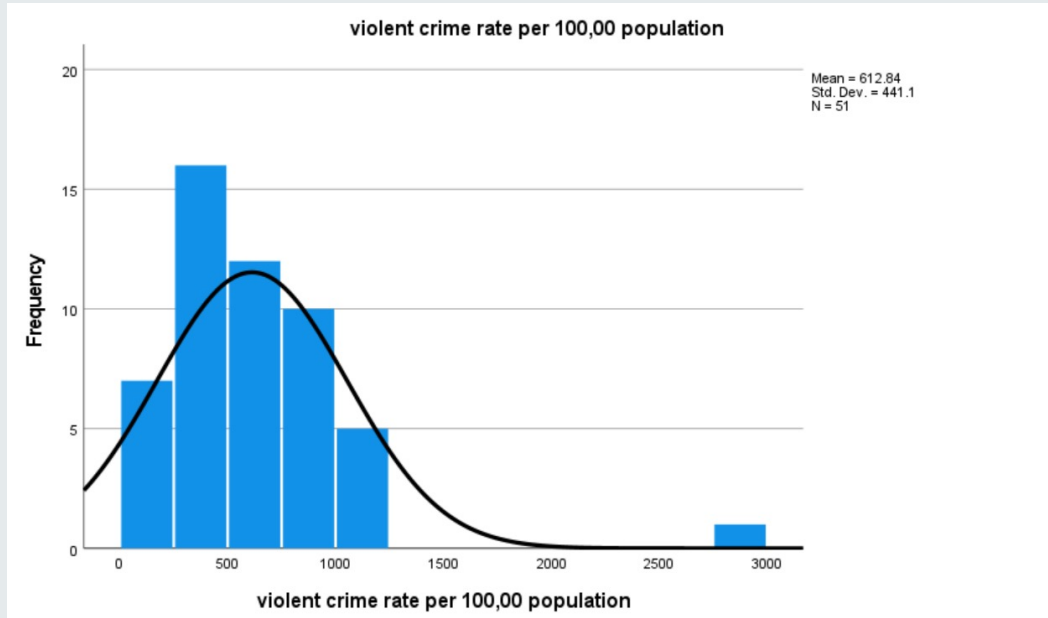
In 'Statistics' ask for descriptive statistics



In 'Charts' ask for a Histogram



Output and Interpretation Slide



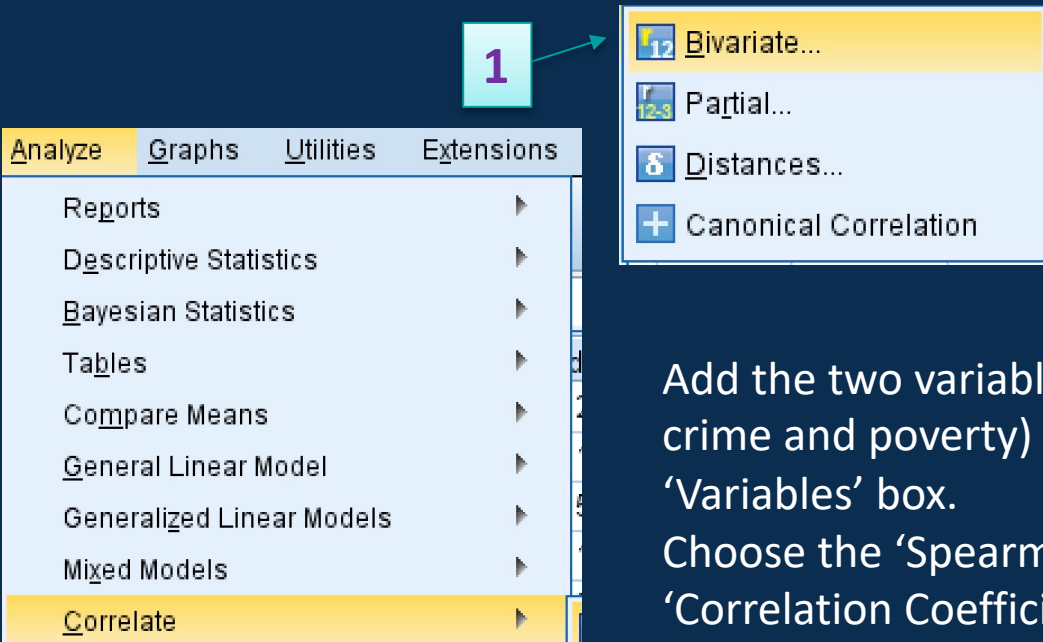
‘Violent Crime’ is a positively skewed variable. ‘Poverty’ is a positively skewed variable. Pearson’s product moment correlation coefficient is unsuitable for this data. Use Spearman’s correlation coefficient instead.



SPSS Slide: 'how to'

According to the researchers, in the population from which our data came, they believe there is a relationship between Violent crime measured per 100,000 and the percentage of people below the poverty line per 100,000.

Step 2: Calculate a correlation coefficient for variables 'violent crime' and 'poverty' from the data



1

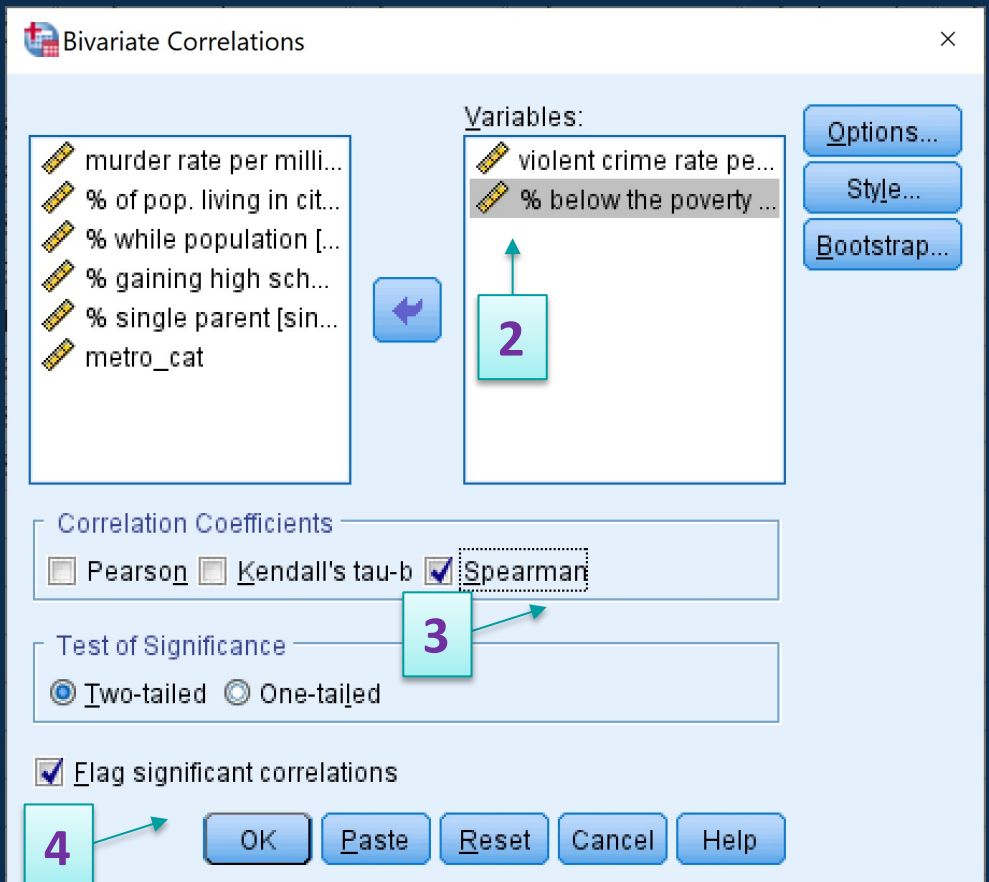
Analyze Graphs Utilities Extensions

- Reports
- Descriptive Statistics
- Bayesian Statistics
- Tables
- Compare Means
- General Linear Model
- Generalized Linear Models
- Mixed Models
- Correlate**

- Bivariate...**
- Partial...
- Distances...
- Canonical Correlation

Add the two variables (violent crime and poverty) into the 'Variables' box.
Choose the 'Spearman's' 'Correlation Coefficient'
Click 'Ok'

Use 'Analyze' -> 'Correlate' -> 'Bivariate'



Bivariate Correlations

Variables:

- murder rate per milli...
- % of pop. living in cit...
- % while population [...]
- % gaining high sch...
- % single parent [sin...
- metro_cat

2

3

Correlation Coefficients

☐ Pearson ☐ Kendall's tau-b ☒ **Spearman**

Test of Significance

☒ Two-tailed ☐ One-tailed

☒ Flag significant correlations

4

Options... Style... Bootstrap...

OK Paste Reset Cancel Help

Output and Interpretation Slide

Correlations

			violent crime rate per 100,00 population	% below the poverty line
Spearman's rho	violent crime rate per 100,00 population	Correlation Coefficient	1.000	.391**
		Sig. (2-tailed)		.005
		N	51	51
	% below the poverty line	Correlation Coefficient	.391**	1.000
		Sig. (2-tailed)	.005	.
		N	51	51

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

There was a weak positive ($r_s=0.39$) relationship between 'violent crime per 100,000' and 'percent below the poverty line per 100'000'. The correlation coefficient is significantly different from 0 ($p=0.005$) so we can extrapolate the weak linear relationship observed in the sample, to the whole population.



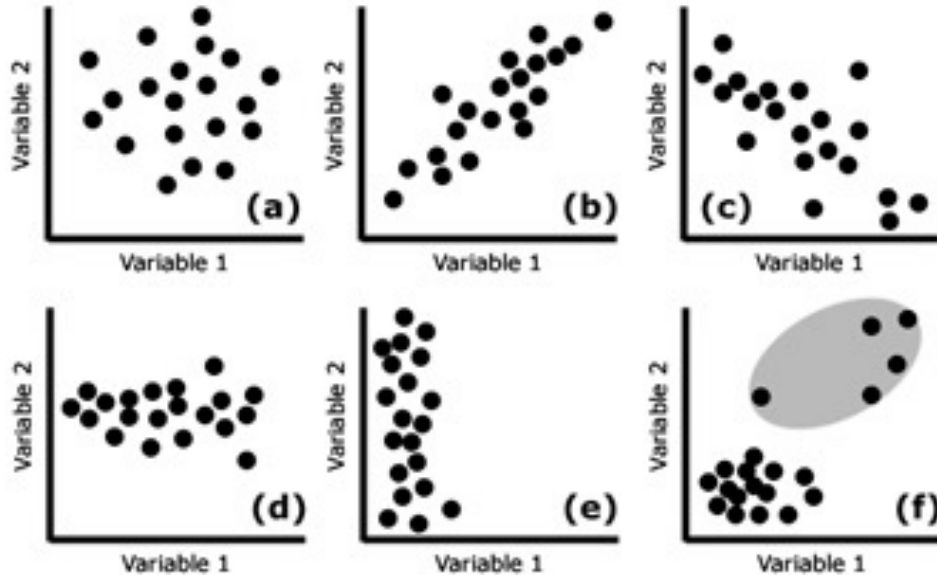
Spurious Correlation: A Word of Caution

Just because two variables are correlated, this doesn't mean there is a causal association between the variables.

The correlation may be due the result of a third unknown variable.

Knowledge Test

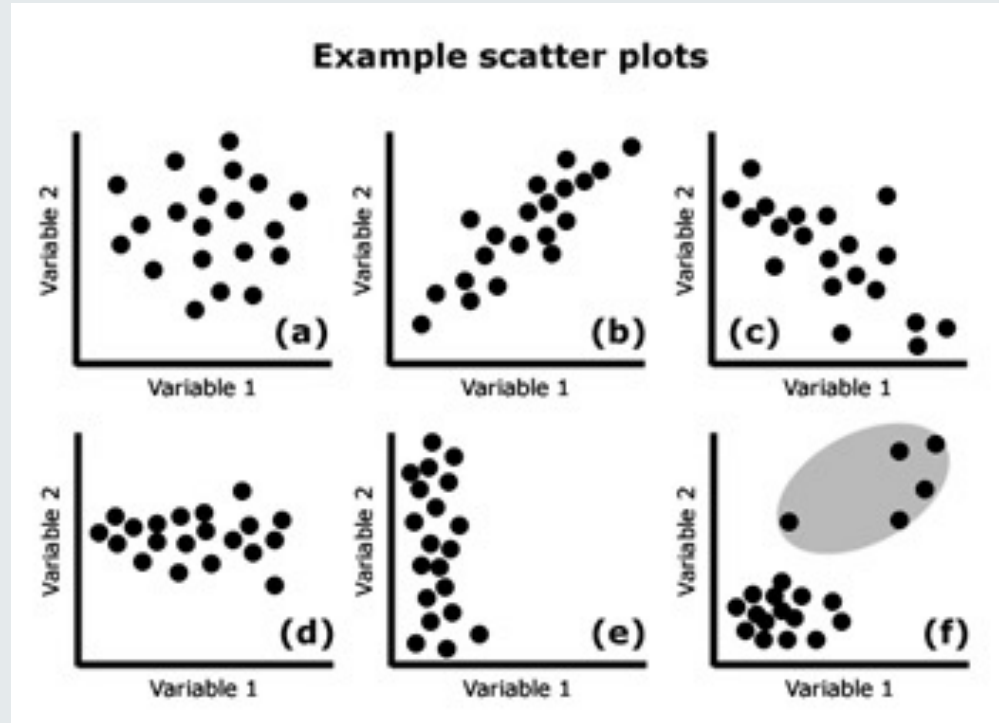
Example scatter plots



1. Quantifying linear relationships

Looking at the 6 figures to the left. For each figure how would you describe the linear relationship between variable 1 and variable 2?

Knowledge Test Solution



1. Quantifying linear relationships

Looking at the 6 figures to the left. For each figure how would you describe the linear relationship between variable 1 and variable 2?

- a) No linear relationship apparent
- b) Positive linear relationship
- c) Negative Linear relationship
- d) No linear relationship variable 1 is utterly immaterial to the variable 2.
- e) No linear relationship variable 1 is utterly immaterial to the variable 2.
- f) Two distinct clusters of data showing different relationships between variable 1 and 2. This may indicate there is some other variable modifying the relationship between these variables

Reflection

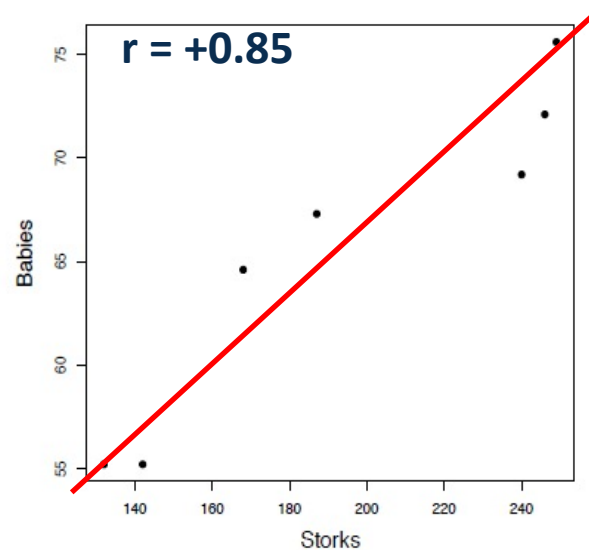
Biology: Stork Population vs. Births

If you examine the records of the city of Copenhagen for the ten or twelve years following World War II, you will find a strong positive correlation between (i) the annual number of storks nesting in the city, and (ii) the annual number of human babies born in the city. These data were researched by Dr. Gustav Fischer and subsequently published in

Ornithologische Monatsberichte, 44 No. 2, Jahrgang, 1936, Berlin *Ornithologische Monatsberichte*, 48 No. 1, Jahrgang, 1940, Berlin *Statistisches Jahrbuch Deutscher Gemeinden*, 27-33, Jahrgang, 1932-1938, Gustav Fischer, Jena.

Can we conclude that storks bring babies? Let's examine the correlation coefficient, just as Dr. Fischer has done in his published work. Visually inspecting the correlation coefficient between the dependent and independent variables, stork population and births, respectively, we can guess that the correlation coefficient is positive and near +0.85.

In this example what you have is a situation where two variables end up as correlated, not because one is influencing the other, but rather because both are influenced by a third variable, Z, that is not being taken into account. That is, the causal relationship here is not $X \rightarrow Y$ or $X \leftarrow Y$,



Read the following article:

What could be this third variable influencing these other variables?

Reflecting on your own field of study.

Write down an example from your research where it would be appropriate to investigate if there is a linear relationship between two continuous variables, what kind of direction and strength might you expect this relationship to have.

Reference List

- Agresti, A., & Finlay, B. (2009). Statistical Methods for the Social Sciences (4th ed., pp. 255-300) New Jersey, NJ: Pearson Hall.
- Field, A. (2005). Discovering Statistics using SPSS (2nd ed., pp. 116-204). London, England: Sage.



Thank you

Contact details/for more information:

Zahra Abdulla

Department of Biostatistics and Health Informatics (BHI)

IoPPN

+44 (0)20 7848 0847

Zahra.abdulla@kcl.ac.uk

www.kcl.ac.uk/xxxx

© 2020 King's College London. All rights reserved

