Introductory statistics for researchers

Marike Cockeran
2017

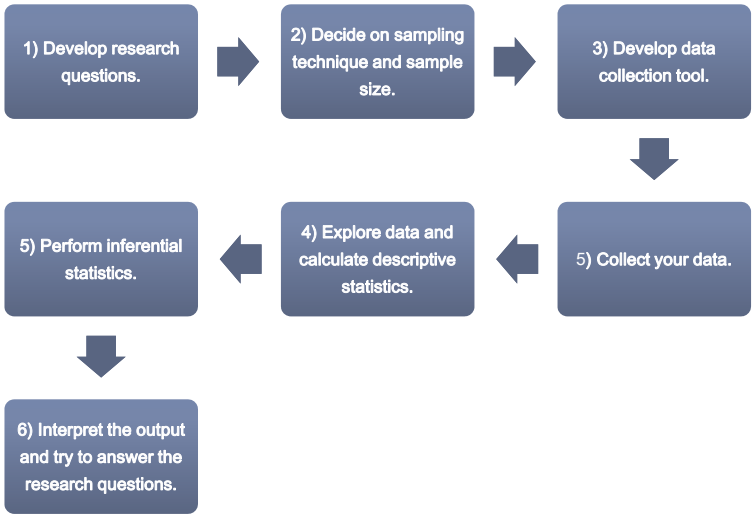
It all starts here ®

NORTH-WEST UNIVERSITY ®
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

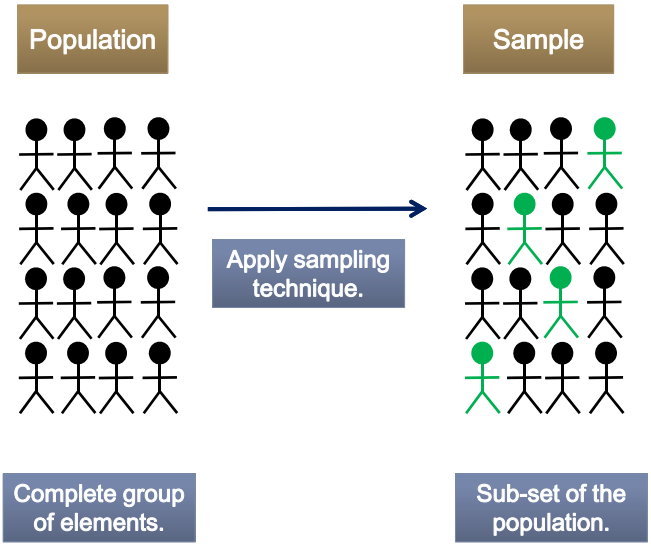
Introduction

- Every quantitative study yields a set of data.
- A complete set of data will not necessarily provide a researcher with information that can easily be interpreted.
- Between the raw data and the reported results of the study lies some intelligent and imaginative manipulation of the numbers, carried out using statistics.
- Statistics explores the
  - collection
  - organisation
  - analysis
  - and interpretation of numerical data.

Role of statistics in the research process



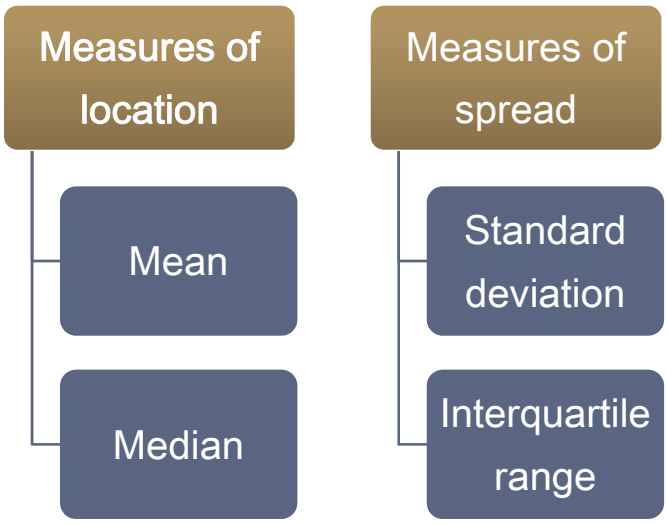
Difference between population and sample



Types of variables

- Discrete variables
  - A variable is said to be discrete if the possible values that it can take on are clearly distinguishable and disconnected from one another.
  - Also called categorical variables.
  - Examples:
    - Gender, race, education level.
- Continuous variables
  - A variable is said to be continuous if the possible values it can take are not clearly distinguishable, i.e., for any two possible values of the variable it is always possible to find another that lies between them.
  - Examples:
    - Height, weight, cholesterol levels.
- Note
  - A continuous variable can be categorised.

Descriptive statistics



Measures of location: Arithmetic mean

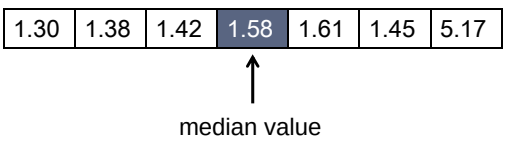
- The mean is calculated by summing all the observations in a set of data and dividing by the total number of measurements.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Properties of the mean:
  - The mean takes into consideration the magnitude of every observation in a set of data.
  - This causes the mean to be extremely sensitive to unusual values.

Measures of location: Median

- If a list of observations is ranked from smallest to largest, half the values are greater than or equal to the median, whereas the other half are less than or equal to it.
- The median is the middle value of an ordered dataset.
- Properties of the median:
  - The median takes into consideration only the ordering and relative magnitude of observations.
  - The median is less sensitive to unusual data points.



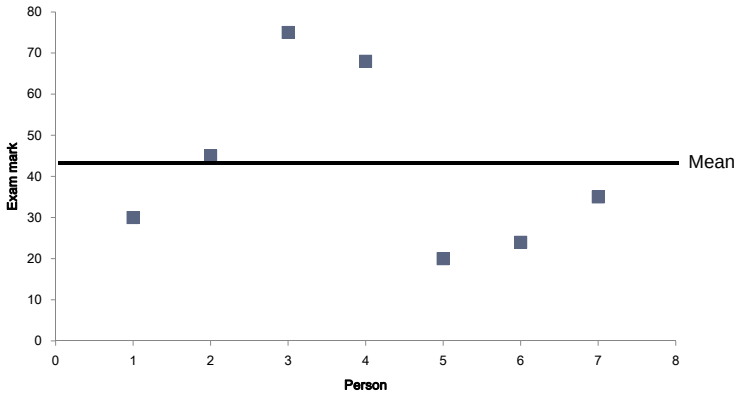
Comparison of the mean and median

- HDL cholesterol values of 7 women:

Dataset 1	Dataset 2
1.30	1.30
1.38	1.38
1.42	1.42
1.45	1.45
1.57	1.58
1.58	1.61
1.61	5.17

- Average value for Dataset 1: $\bar{X} = 1.47$ Median value for Dataset 1: $m = 1.45$
- Average value for Dataset 2: $\bar{X} = 1.99$ Median value for Dataset 2: $m = 1.45$

Measures of spread



Standard deviation

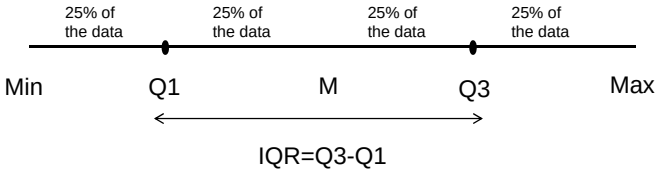
- The standard deviation of a set of observations is the square root of the variance.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

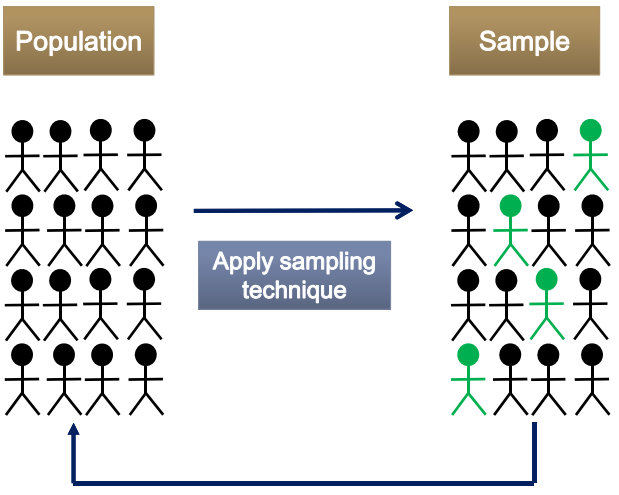
- The standard deviation has the same units of measurement as the mean, rather than squared units.
- In a comparison of two groups of data, the group with the smaller standard deviation has the more homogeneous observations and the group with the larger standard deviation exhibits a greater amount of variability.
- The magnitude of the standard deviation depends on the values in the dataset – what is large for one group of data may be small for another.
- Since the standard deviation has units of measurement, it is meaningless to compare standard deviations for two unrelated quantities.

Interquartile range

- The interquartile range is calculated by subtracting the 25th percentile of the data from the 75th percentile.
- The interquartile range includes the middle 50% of the observations.

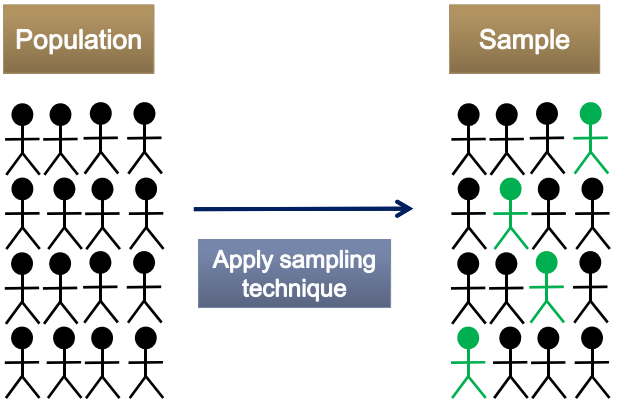


Statistical inference



Statistical inference
Statistical methods used to make conclusion about a population from sample data.

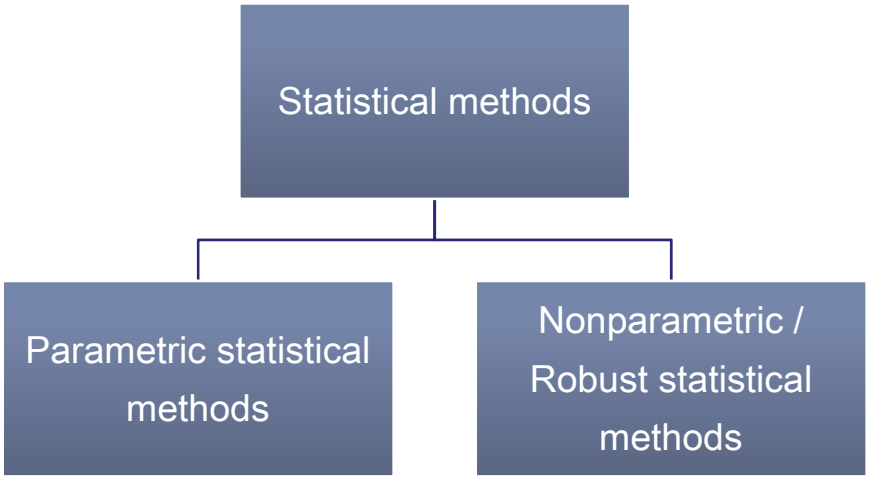
Statistical inference



Population parameter
μ
- Average blood pressure of healthy, 20 year old females .
- Average hours studied by first year NWU students.

Sample statistic
$\bar{X}$
- Average blood pressure of healthy, 20 year old females in your sample.
- Average hours studied by first year NWU students in your sample.

Inferential statistics (1)

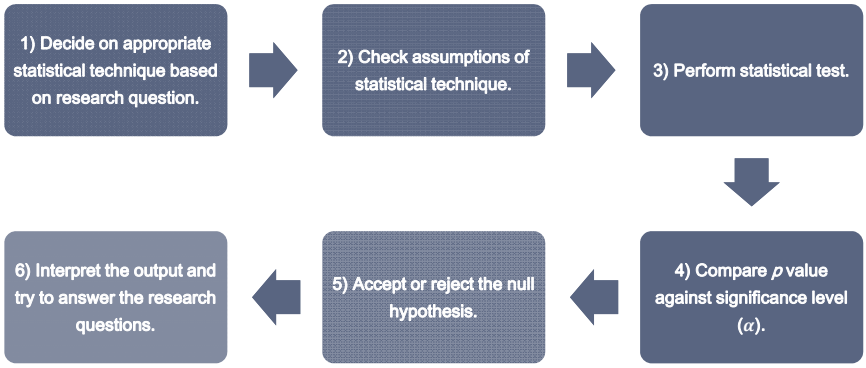


Inferential statistics (2)

Question	Parametric test	Nonparametric test
Is there a difference between two unrelated groups?	Independent t-test	Mann-Whitney test/
Is there a difference between two related groups?	Dependent t-test	Wilcoxon signed-rank test
Is there a difference between several unrelated groups?	One-way ANOVA	Kruskal-Wallis test
Is there a relationship between two continuous variables?	Pearson Correlation	Spearman correlation
Can I predict one variable from another?	Multiple linear regression analysis	
Is there an association between two categorical variables?	Pearson's chi-square test of independence	
Can I combine questions to form a latent variable?	Factor and reliability analysis	

The focus is on parametric statistical techniques.

Steps in inferential statistics



Hypothesis testing

- A hypothesis is a claim (assertion) about a population parameter:
- Population mean
  - The mean monthly cell phone bill for a NWU student is $\mu = \mathrm{R}\,360.00$
- Population proportion
  - The proportion of left handed math's lecturers is $\pi = 0.63$.
- Two independent populations
  - Male students perform better than female students in first year mathematics modules.

Null hypothesis and alternative hypothesis

- Null hypothesis
  - H_0 : It is usually a statement concerning the value of a population parameter under investigation. The value stated in the null hypothesis is usually the value presently accepted to be correct and is accepted as such until proven to be false.
  - Predicted effect does not exist.
- Alternative hypothesis
  - H_a: Usually contains a possible value, or a set of possible values, for the parameter not specified by the null hypothesis.
  - Predicted effect does exist.

P values and significance level

- *P* value = probability of test statistic (result) occurring by chance, given that null hypothesis is true.
- Small *p* value → reject null hypothesis.
- Significant result ($p < 0.05$) does not necessarily imply that null hypothesis is false.
- Non-significant result ($p > 0.05$) does not imply that null hypothesis is true. In fact: Null hypothesis can never concluded to be true.
- Statistical significance ($p<0.05$) does not necessarily imply that the effect is important in practice.

Effect sizes

- Importance of statistical significant effect in practice.
- Effect size is an objective and standardised measure of magnitude of effect.
- Can compare it over different studies, different variables, different scales of measurement.
- Effect sizes are independent of measuring units.
- Many different types of effect sizes:
  - Cohen's *d*
  - Cramer's *v*
  - Pearson's correlation coefficient (r).

Independent t-test

- The independent samples t-test compares the means between two unrelated groups on the same continuous, dependent variable.
- For example, you could use an independent t-test to compare blood pressure levels between smokers and non-smokers.

Give an example of where you would use an independent t-test.

Independent t-test: Hypothesis

- Null hypothesis:

$$H_0: \mu_1 = \mu_2$$

- Alternative hypothesis:

$$H_A: \mu_1 \neq \mu_2$$

- If the sample means differ a lot, H_0 is rejected. The researcher can conclude that the two population means differ.
- If the sample means do not differ a lot, H_0 is not rejected. The researcher cannot conclude that the two population means differ.

Independent t-test: Effect size

- A measure of practical significance (effect size) is Cohen's d-value:

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{\max(s_1, s_2)}$$

- The effect size, Cohen's d-value is calculated manually.
- Guideline values for interpreting Cohen's d-value:

  $|d|$ ≈ 0.2 Small effect / No practically significant difference

  $|d|$ ≈ 0.5 Medium effect / Practically visible difference

  $|d|$ ≈ 0.8 Large effect / Practically significant difference

Independent t-test: Reporting of results

- **Practical significance** (effect size) guideline values:
  - |d| ≈ 0.2 small effect
  - |d| ≈ 0.5 medium effect
  - |d| ≈ 0.8 large effect

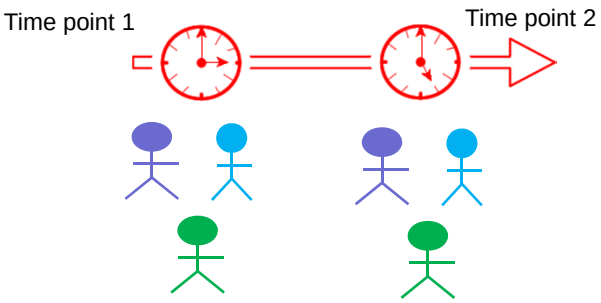
- **Statistical significance** guideline value:
  - Usually when a p-value is smaller than 0.05 the result is viewed as statistically significant.

- **Reporting of the results:**
  - An independent samples t-test was conducted to compare BMI values in males and females. There was a significant difference in the BMI values for males (M=24.12, SD=6.02) and females (M=27.47, SD=7.43), t(198)=2.89, p=0.02.

Dependent t-test: Introduction

- The dependent t-test (paired t-test) compares the means between two related groups on the same continuous dependent variable.

- Measurements are taken on a single subject at two distinct points in time.



Dependent t-test: Hypothesis

- Denote the difference in population means by:

$$\delta = \mu_1 - \mu_2$$

- Null hypothesis:

$$H_0 : \delta = 0$$

- Alternative hypothesis:

$$H_A : \delta \neq 0$$

Dependent t-test: Effect size

- A measure of practical significance (effect size) is Cohen's d-value:

$$d = \frac{|\bar{X}_1 - \bar{X}_2|}{s_1}$$

- The effect size, Cohen's d-value is calculated manually.
- Guideline values for interpreting Cohen's d-value:

  $|d|$ ≈ 0.2 Small effect / No practically significant difference

  $|d|$ ≈ 0.5 Medium effect / Practically visible difference

  $|d|$ ≈ 0.8 Large effect / Practically significant difference

Dependent t-test: Reporting of results

- **Practical significance** (effect size) guideline values:
  - |d| ≈ 0.2 small effect
  - |d| ≈ 0.5 medium effect
  - |d| ≈ 0.8 large effect
- **Statistical significance** guideline value:
  - Usually when a p-value is smaller than 0.05 the result is viewed as statistically significant.
- **Reporting of results:**
  - On average the triceps measurement was lower before the intervention (M=21.73, SD=9.39) compared to the triceps measurement after the intervention (M=21.47, SD=9.94). This difference was not statistically significant, $p = 0.409$.

Analysis of variance: Introduction

- In the previous section we looked at techniques for determining whether a difference exists between the means of two independent populations.
- In some situations we would like to test for differences among three or more independent means rather than just two.
- The extension of the independent two-sample t-test to three or more groups is known as the analysis of variance.

Give an example of where you would use ANOVA.

ANOVA: Hypothesis

- Null hypothesis

$$\mu_1 = \mu_2 = \cdots = \mu_k$$

- Alternative hypothesis
  - At least one of the population means differs from one of the others.
- The One-way ANOVA is an omnibus test and cannot tell you which specific groups were significantly different from each other.
- The omnibus test is referred to as the F-test.
- To determine which groups differ from each other you need to use post hoc tests.

Bonferroni multiple comparison test

- $H_0: \mu_1 = \mu_2 = \mu_3$
- F-test: A p-value of 0.03 is obtained. We can reject the null hypothesis. At least one of the population means differs from one of the others.
- $H_0: \mu_1 = \mu_2$ and $H_0: \mu_1 = \mu_3$ and $H_0: \mu_2 = \mu_3$
- You need to adjust the significance level used for each test to have an overall significance level of $\alpha = 0.05$.
- Bonferroni adjustment: $\frac{\alpha}{\#\ tests} = \frac{0.05}{3} = 0.0167$
- Bonferroni adjustment: p-value x # tests

ANOVA: Reporting of results

- **Statistical significance guideline value:**
  - Usually when a p-value is smaller than 0.05 the result is viewed as statistically significant.

- **Reporting of results:**
  - There was a statistically significant effect of educational level on nutrition knowledge, F (15.05, 2), $p < 0.001$. Tukey's test revealed that the nutrition knowledge of participants with no educational level differed statistically significantly from participants with any educational level, both $p < 0.001$. However, the nutrition knowledge of participants with a low educational level did not differ statistically significantly from participants with a medium educational level, $p = 0.413$.

Introduction: Data set

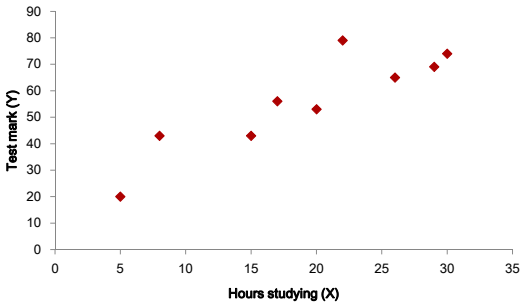
- Suppose we collect data on two variables:
  - hours spend studying (X).
  - test mark (Y).

- For each participant we now have a pair of observations (X_i, Y_i).

Person no	Hours spend studying (X)	Test mark (Y)
1	15	43
2	8	43
3	22	79
4	5	20
5	29	69
6	26	65
7	17	56
8	30	74
9	20	53

Introduction: Scatter plot

- We can plot these pair of observations (X_i, Y_i) on a scatter plot.

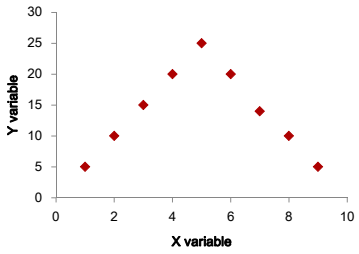


Is there a relationship between a student's hours studying (X) and test mark (Y) ?

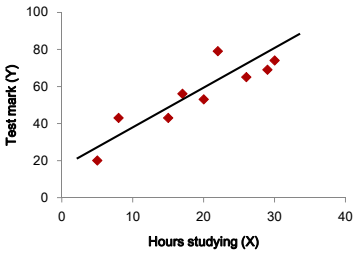
What is the form of the relationship between a student's hours studying (X) and test mark (Y) ?

Introduction: Linear relationship

Non-linear relationship



Linear relationship



We are investigating linear relationships between two continuous variables.

Introduction: Correlation coefficient

- Can we quantify the degree to which two continuous variables are related, provided that the relationship is linear?

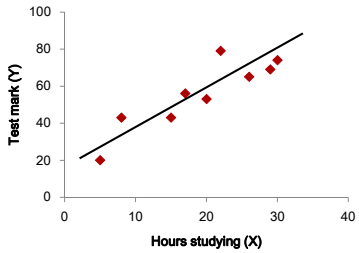
Pearson's correlation coefficient
It measures the strength of a linear relationship between two continuous variables.

Pearson's correlation coefficient: Properties

- The parametric estimator of the correlation between two variables is known as Pearson's correlation coefficient (r).
- The maximum value of r is 1; the minimum value of r is -1.
- If $r = 1$ or $r = -1$ then an exact linear relationship exists between the two variables.
- If $r = 0$ there is no linear relationship between the two variables and the variables are uncorrelated.
- If the values of the first variable increase as the values of the second variable increase, then the two variables are positively correlated.
- If the values of the first variable decrease as the values of the second variable increase, then the two variables are negatively correlated.

Positive linear relationship

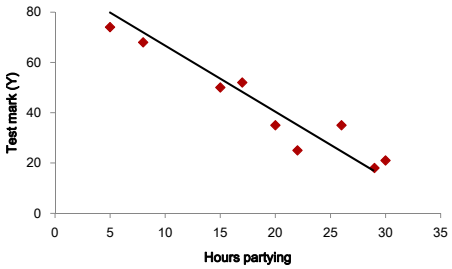
- If the values of the first variable increase as the values of the second variable increase, then the two variables are positively correlated.



Give an example of two continuous variables where you expect a positive linear relationship.

Negative linear relationship

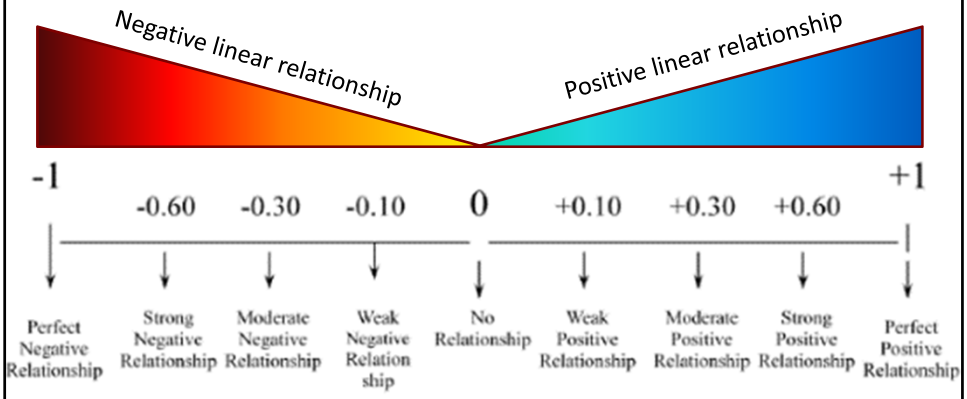
- If the values of the first variable decrease as the values of the second variable increase, then the two variables are negatively correlated.



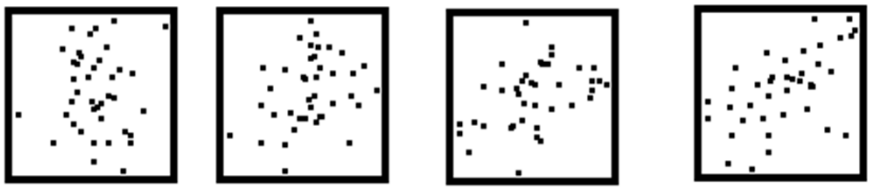
Give an example of two continuous variables where you expect a negative linear relationship.

Guess the correlation (1)

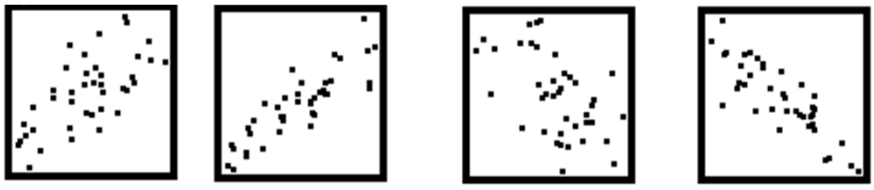
http://guessthecorrelation.com/



Guess the correlation (2)



r = 0 r = 0.28 r = 0.42 r = 0.55



r = 0.67 r = 0.86 r = -0.55 r = -0.85

Pearson's correlation coefficient as effect size

- The correlation coefficient can be interpreted as an effect size measure.
- It measure the strength or the magnitude of a linear relationship between the continuous variables.
- Guidelines for effect sizes:

  r = 0.1 → Small effect → No practically significant relationship

  r = 0.3 → Medium effect → Practically visible relationship

  r = 0.5 → Large effect → Practically significant relationship