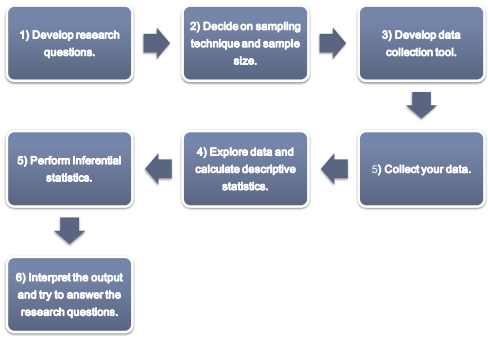
Introductory statistics for researchers

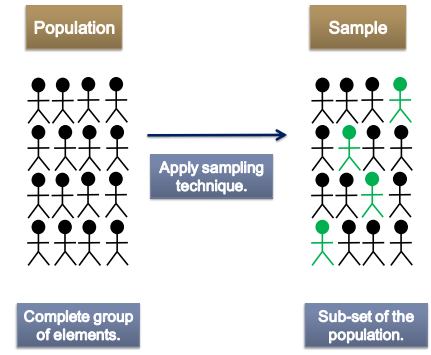
Introduction

- 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



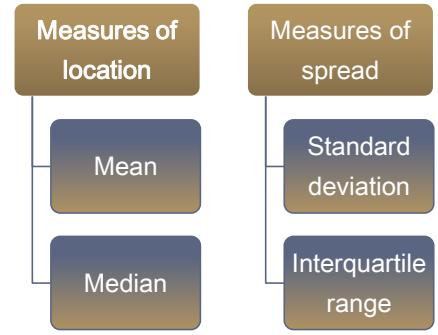
Difference between population and sample



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

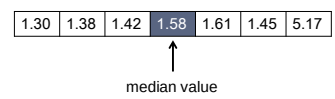
- The mean is calculated by summing all the observations in a set of data and dividing by the total number of measurements.

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

- 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.



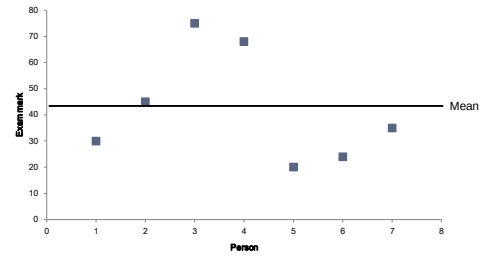
Comparison of the mean and median

- 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

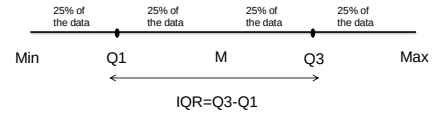
- The standard deviation of a set of observations is the square root of the variance.

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

- 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.



IQR=Q3-Q1

Reporting descriptive statistics (1)

- The descriptive statistics summarise data from a sample, for example, demographic profiles.
- Whenever there are a number of groups, it is useful to provide the descriptive statistics by group and for the overall sample.
- Report total sample and group sizes for each analysis.
- Report numerators and denominators for all percentages.

Reporting descriptive statistics (2)

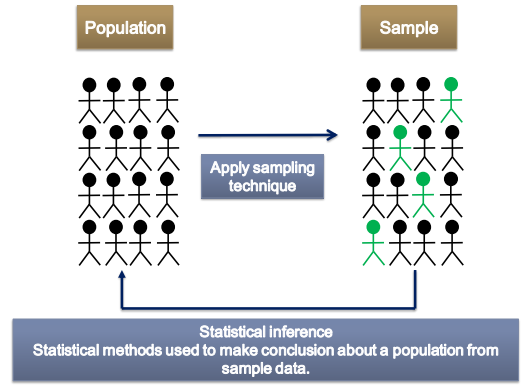
- Summarize data that are approximately normally distributed with means and standard deviations (SD).
  - Use the form: mean (SD).
- Summarize data that are not normally distributed with medians and inter-percentile ranges. Report the upper and lower boundaries of inter-percentile ranges.
- This gives a visual impression of the comparability of the groups in term of their baseline characteristics. It is not necessary to report statistical tests and *p*-values in such a summary because the main concern is the comparability of the participants (which reflects the sampling), not the populations.

Reporting descriptive statistics (3)

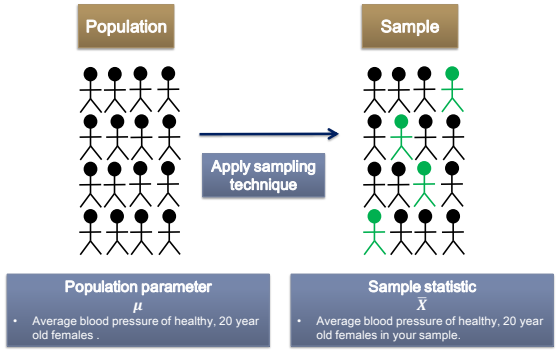
Table 1: Patient demographics (n = 95)

Variables		Drug X (n = 45) n (%)	Placebo (n = 50) n (%)	Total n (%)
Age (years)[a]		45.3 (2.6)	47.8 (3.2)	46.5 (3.0)
Gender	Male	25 (55.6)	25 (50.0)	50 (52.6)
	Female	20 (44.4)	25 (50.0)	45 (47.4)
BMI groups	Underweight (BMI < 18.5)	10 (22.2)	11 (24.0)	21 (22.1)
	Normal (BMI 18.5 to 24.9)	12 (26.7)	13 (28.0)	25 (26.3)
	Overweight (BMI ≥ 25)	23 (51.1)	26 (48.0)	49 (51.6)

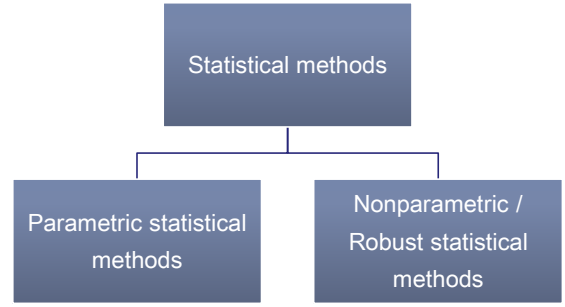
Statistical inference



Statistical inference



Inferential statistics (1)

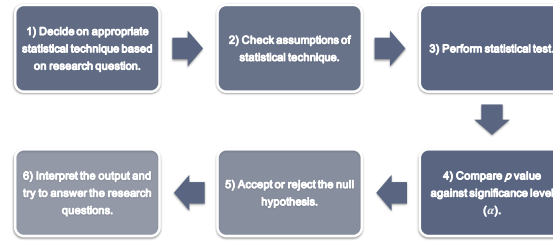


Inferential statistics (2)

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	

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 $\mu = \text{R } 360.00$
- 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.

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

Table 2: Comparison of systolic blood pressure between intervention and control groups.

Variable		Mean (SD)	Mean difference (95% CI)	t-statistic (df)	P-value[a]
SBP (mmHg)	Intervention n = 40	Control n = 40			
					< 0.001
	119.4 (5.06)	109.4 (6.34)	10.0 (7.5, 12.6)	7.83 (78)	

SBP = systolic blood pressure. [a] Independent t-test.

Independent t-test: Reporting of 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.

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

Table 2: Comparison of mean weight between the four diet plans.

Groups	n	Weight (kg) Mean (SD)	F-statistic (df1, df2)[a]	P-value[a]
Okinawa Diet	10	65.5 (9.98)		
Eastern Diet	10	75.4 (4.17)		
Western Diet	10	77.9 (5.70)	13.41 (3, 36)	< 0.001[b]
Fast food Diet	10	83.9 (5.07)		

[a] One-way ANOVA, [b] Post-hoc analysis with Bonferroni corrections shows significant difference between Okinawa diet and other diet plans (P < 0.001) and between Eastern diet and Western diet (P = 0.041).

ANOVA: Reporting of results

- **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.

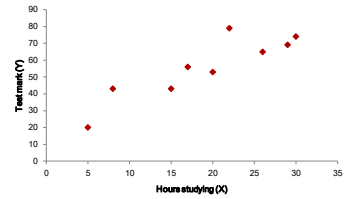
Introduction: Data set

- 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.

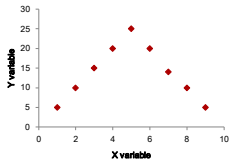


Is there a relationship between a student's hours studying (X) and test mark (Y) ?

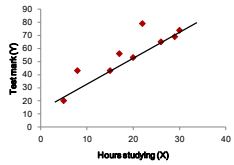
What is the form of the relationship between a student's hours studying (X) and test mark (Y) ?

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?

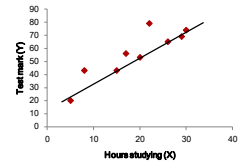
Pearson's correlation coefficient
It measures the strength of a linear relationship between two continuous variables.

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

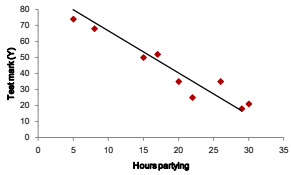
- If the values of the first variable increase as the values of the second variable increase, then the two variables are positively correlated.



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

Negative linear relationship

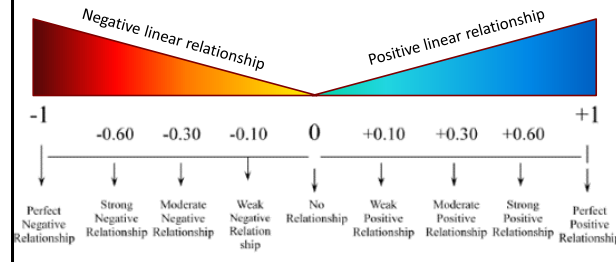
- If the values of the first variable decrease as the values of the second variable increase, then the two variables are negatively correlated.



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

Guess the correlation (1)

http://guessthecorrelation.com/



Correlation: Reporting of results

- Hours spent studying and GPA were strongly positively correlated, $r(123)=.61$, $p=.011$.
- Hours spent playing video games and GPA were moderately negatively correlated, $r(123)= -.32$, $p=.041$.