

Research design, measurement, and statistics

The CONSORT guidelines

- CONsensus Statement On Reporting of Trials
- they include:
 - all patients assessed for the trial should be accounted for, and that the report should be accompanied by a diagram that explains what happened to all the patients involved in the trial
 - the randomization process should be clearly specified
 - inclusion and exclusion criteria should be clearly stated
 - pre-study power calculations should be provided
 - the method of blinding should be specified
 - there should be an intention-to-treat analysis

Epidemiology in psychiatric research

Prevalence

- is a measure of disease frequency, and since it is a proportion it can be expressed as a percentage
- it has no units
- is determined by a cross-sectional study
- prevalence has to be related to a time period, for example weekly, monthly, or lifetime:
 - **point prevalence** – the proportion of a defined population that has a given disease at a given point in time
 - **period prevalence** – the proportion of a defined population that has a given disease during a given interval of time
 - **lifetime prevalence** – the proportion of a defined population that has or has had a given disease
 - **birth defect rate** – the proportion of live births that has a given disease

Incidence rates (rate ratios)

- is the rate of occurrence of new cases of the disease in a defined population over a given period of time
- incidence is not expressed as a proportion since the denominator is person-time
- *cumulative incidence* is the proportion of subjects who have who have developed a disease within a specified time period
 - if the disease is uncommon, and the loss to follow-up is small, the cumulative incidence will not introduce any major inaccuracy
 - it depends on the time period of follow-up and inevitably rises as the period of follow-up lengthens

Population at risk

- is the population of individuals free of a given disease, who have not already had the disease by the time of the commencement of a given period of time, who are at risk of becoming new cases of the disease

Mortality rate

**MR = number of deaths in a defined population during a given period
population size during that time period**

- **standardized mortality rate** – the mortality rate adjusted to compensate for a confounder such as:
 - gender
 - age
 - ethnicity
- **age-standardized mortality rate** – adjusted to compensate for age
- **standardized mortality ratio (SMR)** – the ratio of the observed standardized mortality rate (from the population being studied) to the expected standardized mortality rate (from a comparable population)

Morbidity rate

- is the rate of occurrence of new non-fatal cases of a given disease in a defined population at risk during a given period of time
- it has equivalents of the above type of mortality rates and ratio

Measures of association between disease and exposure

Relative risk

- the relative risk of a disease with respect to a given risk factor is the ratio of the incidence of a disease in people exposed to that risk factor to the incidence of the disease in people not exposed to the same risk factor
- gives an estimate of the 'aetiological force' of an exposure
- it can only be calculated from prospective studies
- a value of 1 for the relative risk implies no causation, and this is the null hypothesis in hypothesis testing
- the association is positive if the relative risk is greater than 1, and negative if it is less than one
- when the relative risk has a value that is statistically significantly different from 1, this is taken as evidence of an association between the disease and the relevant risk factor
- e.g. the relative risk of cigarette smoking and lung cancer is 10 – i.e. there is a 10 times increased risk of lung cancer if you smoke
- there are three measures of relative risk:
 1. risk ratios
 2. odds ratios
 3. incidence rate ratios, or rate ratios

Worked example

- in a study of those exposed to a risk factor, the results were:

	Disease	No disease
Risk factor present	a	b
Risk factor absent	c	d

Incidence in those exposed = $a/a+b$

Relative risk = $c(a+b)/a(c+d)$

Odds ratio = bc/ad

Attributable risk = $a(a+b) - c/(c+d)$

Odds ratios

- are also used as a measure of association
- easier to calculate than risk ratios
- logistic regression also provides odds ratios
- in case-control studies, odds ratios are the only measure of association that provide an unbiased estimate of the odds ratio in the population on which the case-control study was based
- odds ratios will approximate towards risk ratios and rate ratios as the incidence of the disease falls
 - the *rare disease assumption* means that for many studies the odds ratio can be interpreted as a rate ratio (i.e. an odds ratio can be calculated to check the relative risk)

Standardised mortality ratios

- the SMR gives the rate ratio in comparison with the England and Wales rates
- SMRs are usually given after being multiplied by 100

Attributable risk

- provides an estimate of the absolute difference in risk or rates between the exposed and non-exposed
- a value of 0 implies no causation
- it therefore provides an assessment of the individual change in risk associated with the exposure

Population attributable fraction (PAF)

- the PAF aims to estimate the proportion of cases that can be attributed to an exposure within a population
- it is dependent on:
 - frequency of exposure
 - size of the relative risk
- if there is no bias in the relative risk estimate, the PAF gives the percentage reduction in the number of cases in that population which would be expected if the exposure were eliminated

Experiments

- are situations in which the researcher manipulates one variable and then observes the effect of that manipulation on another variable
- the variable manipulated by the observer is the **independent variable** (e.g. drug levels)
- the variable to be observed is called the **dependent variable**, because it *depends* on the independent variable (e.g. recovery rates, scores on rating scales)
- the group that receives the experimental treatment is called the **experimental group**, while the group receiving no treatment, or some other treatment, is called the **control group**
- any factor that might have affected the dependent variable, along with or instead of the independent variable, is called a **confounding variable** - sources of confounding variables are:
 1. random variables
 2. age or sex imbalances
 3. placebo effects / subjects expectations:
 4. experimenter bias
- **controlled variable**- a variable whose influence is kept constant by the experimenter (e.g. other medications)
- **uncontrolled variable**- a variable which is not manipulated or held constant, though it may be measured (e.g. life events)

Types of data

- **Qualitative:**
 - refer to attributes that can be categorized such that the categories do not have a numerical relationship with each other, e.g. eye colour
 - **nominal/ categorical** – where categories bear no relation to each other; the categories are mutually exclusive and bear no relation to each other (e.g. marital status, eye colour, type of psychiatric disorder)
- **Quantitative:**
 - refer to numerically represented data
 - data are either:
 - **continuous** (e.g. height, weight, etc.)
 - or **discrete**
 - discrete data can be either:
 - **ordinal** – where order is inherent, but not quantified; the numbers assigned to the categories indicate the amount of characteristic possessed (e.g. social class I-V, score on the CGI)
 - **interval** – differences between points on the scale reflect equal differences in the characteristic being measured; the zero point is just another point in the scale (e.g. scores on most scales)
 - **ratio scales** – (a type of interval scale) has a true zero point which indicates the beginning of the scale; the figures obtained can be further manipulated to calculate means, etc.

- continuous measures are normally distributed and can be analyzed with parametric statistics
- discrete data are usually not normally distributed and must be analyzed with non-parametric tests

Random variables

- **random variables** are uncontrolled, sometimes uncontrollable, factor such as differences among the subjects e.g. backgrounds, personalities, vulnerability to stress
- **random assignment** tends to distribute the impact of uncontrolled variables randomly (and probably evenly) across groups

Placebo effects

- improvement created by a subject's knowledge and expectations is called the **placebo effect**
- a **placebo** is a (non-active) treatment which produces benefits because a person *believes* it will be beneficial
 - green tablets best for anxiety
 - yellow tablets better for depression
 - very small, or very large tablets are most effective
- the placebo response may be increased by double-blind trials because both sets of tablets will be administered with equal faith and enthusiasm
- placebo responses tend to occur early and fluctuate
- extending the length of the trial can help to reduce the placebo response
- the following conditions should be met for placebo-controlled trials:
 1. there is genuine uncertainty as to whether the proposed treatment is more likely to help patients than placebo
 2. there is no agreed alternative treatment of more value than placebo
 3. it must be important to know whether the new treatment is better than placebo
 4. there must be a reasonable possibility that the proposed treatment is better than placebo
 5. the illness must not be so serious that any delay in receiving active treatment would be harmful or dangerous

Bias

- **ascertainment bias** results if two populations of cases/ controls are recruited in different ways and differ because of this, e.g. general population vs. hospital staff
- **selection bias** is introduced when an individual is recruited in such a way that their exposure or disease status is linked to their group status, e.g. unidentified depressed patients in primary care are likely to have a better prognosis
- **experimenter bias** is the unintentional effect that experimenters may exert on results
 - **observation biases** include recall, interviewer, misclassification
 - **response biases** include selection, social desirability
- to prevent experimenter bias from confounding results, experimenters often use a **double-blind design**

Quasi-experiments

- **quasi-experiments** are studies whose designs approximate the control of a true experiment (e.g. do certain drugs cause birth defects in pregnancy? - you cannot randomly allocate subjects to take drugs)

The Quality of Tests

- a **test** is a systematic procedure for observing behaviour in a standard situation and describing it with the help of a numerical scale or system of categories
- a test has three advantages over other means of evaluation:
 1. the administration, scoring, and interpretation are *standardized*
 2. tests summarize the test taker's performance in *quantifiable* terms, which allows testers to calculate *norms*
 3. tests are *economical* and *efficient*

Reliability

- **reliability** indicates that the results are repeatable, and stable
- the higher the reliability, the less susceptible its scores are to being affected by changes in the conditions that the test is completed under
- to estimate reliability, researchers compute the *correlation coefficient* between two sets of scores from the same person on the same test
 - a high, positive correlation indicates the test is reliable
- the two sets of scores can be obtained in various ways:
 1. **test-retest reliability** - high correlation between scores on the same test given on different occasions
 2. **alternate-form reliability** - two forms of the same test can be used
 3. in the **split-half** method, a correlation coefficient is calculated between a person's scores on two comparable halves of the test
 - *Cronbach's alpha* gives a measure of the average correlation between all the items when assessing split-half reliability – it therefore indicates the internal consistency of the test
 4. **inter-rater reliability** - high correlation between results of two or more raters of the same material at roughly the same time; measured using *intraclass correlation coefficient* (ICC - range 0-1; ICC of ≥ 0.7 is acceptable)
 5. **intra-rater reliability** – the level of agreement between assessments made by two or more assessors of the same material presented at two or more times

Validity

- the **validity** of a test is the degree to which it measures what it is supposed to measure
- validity depends on how the test is used - the same test can be valid for one purpose but invalid for another
- most measures of validity are correlation coefficients between test scores and another variable
- types of validity include:
 1. **face/ content validity** - the degree to which the test's content is related to what the test is supposed to measure, i.e. are the criteria used overtly related to the diagnosis
 2. **criterion validity** - the extent to which test scores correlate with another direct and independent measure (*criterion*) of what the test is supposed to measure
 3. **concurrent validity** – compares the measure being assessed with an external valid yardstick at the same time
 4. **incremental validity** – indicates whether the test is superior to other measurements in approaching true validity
 5. **cross-validity** – determines whether after establishing criterion validity for one sample, it maintains criterion validity when applied to another sample
 6. **predictive validity** – can the diagnosis be used to accurately predict outcome or other patient-related events?
 7. **convergent validity** – established when measures expected to be correlated are indeed found to be associated
 8. **divergent validity** – when measures discriminate successfully between other measures of unrelated constructs
 9. **construct validity** - the extent to which the scores on a test are in accordance with one's theory about what is being tested and in terms of future consequences
 - relies on both convergent and divergent validity being established
 - most psychiatric diagnoses have this type of validity
- the **reliability paradox** - a very reliable test may have low validity precisely because its results do not change, i.e. it does not measure true changes

Rating Scales

- records qualitatively but measures quantitatively

Thurstone scale

- dichotomous scale indicating agreement/ disagreement with a statement
- disadvantages:
 - different response patterns may result in the same mean score
 - setup is unwieldy
 - ranking may be biased

Likert scale

- 5-point equal interval scale indicating level of agreement
- advantages:

- increased sensitivity
- easily administered
- disadvantages:
 - different response patterns may result in the same mean score

Semantic Differential scale

- a bipolar visual analogue scale
- advantages:
 - ease of use
 - good test-retest reliability
- disadvantages:
 - positional response bias
 - no consistent meaning to midpoint mark

Sources of error

- **Response set** - subject always tends either to agree or to disagree with questions; it measures defensiveness
- **Extreme responding** – the tendency always to agree or to disagree with the questions being asked
- **Bias towards centre** - subject tends to choose the middle response and shun extremes
- **Social acceptability/ desirability** - subject chooses the answers that they believe the interviewer wants to hear (occur consciously and unconsciously)
 - occurs more commonly in self-related questionnaires
 - reduced by:
 - using the forced-choice technique
 - including a lie-scale
- **Halo effect** - an observer error, which arises in data collection when there is carry-over from one judgment to another. E.g. if you believe someone is a good student, you tend to mark them favourably.
- **Hawthorne effect** - researchers alter the situation by their presence; it can affect most types of observation, including naturalistic ones. Named after the Hawthorne car factory, where experimenters found that simply being there altered the outcome.

Sampling methods

Simple random sampling

- is one chosen from a given population such that every possible sample of the same size has the same probability of being chosen

Systematic sampling

- **Periodic sampling** – every n^{th} member of the population is chosen
 - may not always lead to a random choice because of an unforeseen underlying pattern
- **Using random numbers** – is a better method than periodic sampling to ensure random choice
 - uses computers or a random number table
- **Stratified random sampling** – a given population is stratified before random samples are chosen from each stratum
 - can be useful when studying a disease that varies with respect to age or sex

Types of Study

- Simple designs may be:
 1. **Cross-sectional**
 2. **Longitudinal (cohort)**
 3. **Case-control**
- Designs may also be:
 1. **Cross-over** - patients are used as their own controls
 2. **Double-blind** - observer and patient are unaware of independent variables when assessing dependent variables
 3. **Latin square** - variables are allocated randomly to subjects and longitudinally in time
- Research may be:
 1. **Normothetic**- information derived from classifying a number of events and experiences
 2. **Ideographic** - information derived from a single event or individual
- **Operational definition** - precise definition of terms used, in measurable quantities or attributes

Cross-sectional study

- compares data collected simultaneously from people of different ages
- they are used for descriptive purposes in estimating the prevalence of a particular disorder
- selection bias is unlikely due to the random sampling
- confounding variables:
 - **cohort effect** - because older people were born in a different year, they may have had different educational, cultural, nutritional, and medical experiences

Longitudinal study

- a group of people are repeatedly tested as they grow old
- allows changes associated with age to be measured
- types of longitudinal designs:
 - **retrospective** - observations in the present to elicit information about past events; may be methodologically unsatisfactory due to *information bias*
 - **prospective (cohort study)** - ongoing information gathering after the event studied; more difficult and expensive
 - **mixed** - retrospective study to elicit desired group of subjects, then prospective follow up
- confounding variables:
 - **mortality effect** - fewer and fewer people can be tested over time as death and illness reduce the sample size

- **history effect** - some event, such as a reduction in healthcare benefits for senior citizens, may have an effect that could be mistakenly attributed to age
- **testing effects** - participants may improve over time due to learning during repeated testing
- longitudinal studies may **underestimate** the degree to which abilities decline with age since those remaining in the study are likely to be the healthiest in the group

Case-control studies

- subjects with the disease are compared in terms of the frequency of exposure with a comparison group
- the comparison group is to provide an estimate of the frequency of exposure in the population from which the cases are drawn
- they are quick and simple to carry out, and can study a variety of exposures for a single disease
- however, they are prone to:
 1. *selection bias* – the controls give a biased estimate of the frequency of exposure in the population
 2. *recall bias* – occurs when the measurement of exposure is biased by the presence of disease; for example when sick individuals try to explain their illness (‘effort after meaning’)

	Case-control	Cohort
Strengths	<ul style="list-style-type: none"> • suitable for rare diseases • better for studying causes in the distant past • can examine many risk factors for a single disease • quick and cheap 	<ul style="list-style-type: none"> • suitable for rare exposures • can examine many outcomes for a single exposure • can calculate incidence rates
Weaknesses	<ul style="list-style-type: none"> • selection bias • recall bias • reverse causality • unsuitable for rare exposures 	<ul style="list-style-type: none"> • unsuitable for rare diseases • expensive • long delay before availability of results • losses to follow-up can affect validity

Descriptive statistics

- the numbers that researchers use to describe and present a set of data

Discrete probability distributions

Events

- *independence*: the occurrence of one event does not in any way influence the probability with which the other can occur
 - when events are independent, the probability of both occurring is equal to the product of their individual probabilities
- *mutually exclusive*: the occurrence of one event means that for all practical purposes, the other cannot occur
 - when events are mutually exclusive, probabilities of one or the other occurring is the sum of their individual probabilities

Bernoulli trial

- is a trial or experiment having two and only two alternative outcomes

Bernoulli distribution

- is the probability distribution for a discrete binary variable (range =0,1) which is a special case of the binomial distribution $B(1,p)$, where p is the probability of 'success':

$$\text{mean} = p$$

$$\text{variance} = p(1-p)$$

Binomial distribution

- the binomial distribution, $B(n,p)$, is the probability distribution for a discrete finite variable (range = 0,1,2,...,n):

$$\text{mean} = np$$

$$\text{variance} = np(1-p)$$

Poisson distribution

- a type of binomial distribution
- the Poisson distribution, Poisson (μ), is the probability distribution for a discrete infinite variable (range = 0,1,2,...) where $\mu = np$

$$\text{mean} = \mu$$

$$\text{variance} = \mu$$

- can be used in situations when the following criteria apply:
 - events occur randomly in time or space
 - the events are independent
 - two or more events cannot take place simultaneously
 - the mean number of events per given unit of time or space is constant

Continuous probability distributions

The Normal Distribution

- the normal distribution, $N(\mu, \sigma^2)$, is the probability distribution for a continuous variable (range = R)
 - mean = μ
 - variance = σ^2
- the data representing most biological variables (e.g. IQ, height, etc.) form a bell-shaped curve known as the **normal distribution**, or *normal curve*
- it has the following properties:
 - it is unimodal
 - it is continuous
 - it is symmetrical about its mean
 - the mean, median, and mode are all equal
 - the area under the curve is one
 - the curve tends to zero as the variable moves in either direction from the mean
- **Skewness** - measures deviation from normal distribution curve. The normal distribution is symmetric and has a skewness of zero. A distribution with a positive skewness has a long right tail, and a negative skewness has a long left tail.
- **Kurtosis** - measures peakedness or flatness of curve. It is a measure of the extent to which observations cluster around a central point. For a normal distribution, the value of the kurtosis statistic is zero. Positive kurtosis indicates that the observations cluster more and have longer tails.

Summary statistics – measures of location

Measures of central tendency

- the **mean** is the *arithmetic average*
 - the mean reflects the actual value of all the scores, whereas the median gives each score equal weight, whatever its size
 - the mean is the preferred measure of central tendency
 - ‘trimmed’ means that the value excludes one or more extreme scores that might distort the value of the conventional mean, e.g. ‘5% trimmed mean’
 - it is suitable for use with data measured on at least an interval scale
- the **median** is the halfway point in a set of data
 - the median is suitable for use with data measured on at least an ordinal scale
 - it gives a better measure of central tendency than the mean for skewed data
- the **mode** is the value or score that occurs most frequently in a data set
 - its value depends on the class intervals used to construct the frequency distribution
 - it can be used with all measurement scales
- **positive skew:**
 - mean > median > mode
- **negative skew:**
 - mode > median > mean

Quantiles

- quantiles are cut-off points that split a continuous distribution into equal groups
- they include:
 - the median – splits a distribution into two equal parts
 - tertiles (2) – these split the distribution into three equal parts
 - quartiles (3) – split the distribution into four equal parts
 - the first quartile is the value below which 25% of the observations lie
 - the **interquartile range** is the name given to the spread of the middle 50% of the scores, which is thus the difference between the upper quartile (75th percentile) and the lower quartile (25th percentile)
 - quintiles (4) – split the distribution into five equal parts
 - deciles (9) – split the distribution into 10 equal parts
 - percentiles (99) – split the distribution into 100 equal parts
 - a **percentile score** indicates the percentage of people or observations that fall below a given score in a normal distribution:
 - +1 S.D. = 84th percentile
 - +2 S.D. = 97.5 percentile

Range

- is the difference between the smallest and largest values in a distribution
- it can be measured with data that are measured on at least an interval scale

Standard Deviation

- the **Standard deviation**, or **S.D.** measures the average difference between each score and the mean of the data set
 - *S.D.* = square root of the variance
 - the more variable data are, the higher the standard deviation
- calculating the standard deviation:
 1. compute the mean
 2. calculate the difference, or *deviation* (D) of each score
 3. find the average of these deviations:
 - the deviations are first squared (to remove any negative values)
 - the squared deviations are then summed, divided by N (the number of observations in the data set), and the square root is taken
- 68% of scores fall between 1 SD above and 1 SD below the mean; 95% of scores fall between 1.96 SD above and 1.96 SD below; 99% of scores fall between 2.58 SD above and 2.58 SD below
- scores may also be expressed in terms of their distance in standard deviations from the mean, producing **standard (Z) scores** - a standard score of 1.5 is 1.5 standard deviations from the mean
 - a Z score of 2 on a depression scale (range 0-30) and a Z score of 0.75 on an anxiety scale (range 0-100) would indicate that the patient has a more extreme depression score than anxiety score
- **Standard error** - an estimate of the discrepancy between sample mean and true population mean
 - an SE allows you to judge the reliability of your sample data
 - *Standard error* = *S.D.* / square root of cases
- the **Sten** ('**Standard Ten**') and **Stanine** ('**Standard Nine**') scales use 0.5 SD size intervals - they have a SD of 2.0 (i.e. two intervals per z-score)

Variance

- *variance* = the average squared deviation from the mean
- it is the square of the standard deviation
- its units are the square of those of the observations
- V = the sum of all the differences between values and the mean, squared (so that negative and positive deviations do not cancel each other out), divided by the total number of observations minus one

Confidence Intervals (CIs)

- we have seen that 95% of normally distributed values will be between 1.96 SD above and below the mean
- it follows therefore that there is a 95% chance that for a normally distributed sample, there is a 95% chance that the mean will lie in the range between 1.96 SE above and 1.96 SE below the true mean
- this is the 95% confidence interval
- CIs can decrease type II errors
- a *t* distribution can be used to calculate confidence intervals

- CIs should be given for every comparison

Graphs

Drawing graphs

- a properly drawn graph should have the following properties:
 - a clearly labelled axis
 - the independent variable is usually represented on the horizontal axis
 - the units for both axes are clearly stated
 - the scales for both axes are given; these may be:
 - linear
 - logarithmic
 - broken

Linear relationship

$$y = mx + c$$

- m is the gradient of the line
- c is the intercept of the line on the y-axis

Power Law relationship

- $y = \log Y$
- $x = \log X$

$$Y = CX^m$$

- m is the gradient of the line
- $\log C$ is the intercept of the line on the y-axis

Exponential relationship

- $y = \ln Y$
- $x = X$

$$Y = Ce^{mx}$$

- m is the gradient of the line
- $\ln C$ is the intercept of the line on the y-axis

Outliers

- are extreme values
- exert an extreme effect on the mean:
 - especially when the total number of values is small
 - the median is less affected
- they exert an extreme effect on the range:
 - measures relating to quantiles are less affected
 - the SD may be less affected by outliers
- affect correlation and linear regression

Stem-and-leaf plots

- used to represent a continuous variable
- they allow the representation of all the individual data
- the stems consist of a vertical column of numbers on the left-hand side of the plot
- the leaves are numbers to the right of the stems, which may represent tenths
- the overall shape of the plot indicates the shape of the distribution
- for example, the distribution 13.5, 13.7, 14.5, 14.6, 14.6, 14.7, 15.2, 15.9, and 16.4 can be represented as:

```
13    5 7
14    5 6 6 7
15    2 9
16    4
```

Boxplots (box-and-whisker plots)

- used to represent a continuous variable
- can be useful for comparing 2 or more sets of observations diagrammatically, before or in addition to more formal statistical analysis
- consists of a box whose longer sides are placed vertically, with vertical lines (whiskers) extending vertically
- it has the following features:
 - the upper boundary of the box is the upper (3rd) quartile
 - the lower boundary of the box is the lower (1st) quartile
 - the length of the box is the interquartile range
 - a thick horizontal line inside the box is the median (2nd quartile)
 - the lower whisker extends to the smallest observation
 - the upper whisker extends to the largest observation
 - outliers are indicated by the symbol O

Scattergrams (dot graphs)

- can be used to represent two continuous variables

Descriptive and Inferential statistics

Descriptive statistics

- are ways of organizing and describing data
- examples include:
 - diagrams
 - graphical representations
 - numerical representations
 - tables

Inferential statistics

- mathematical procedures used to draw conclusions from data and to make inferences about what they mean
- inferential statistics provide a measure of how likely it was that the results came about by chance

Hypothesis testing: The Null hypothesis

- a value or range of values for an unknown population parameter is hypothesized
- an experiment is then carried out and the value of the observed random variable is used to test whether or not the hypothesis should be rejected
- the hypothesis that there is no significant difference between comparison groups, or that any difference is only due to chance is termed the *null hypothesis*
- it represents no change:

$$H_0: \theta = \theta_0$$

- where θ is the unknown parameter and θ_0 is its hypothesized value

Alternate hypothesis

- the alternate hypothesis, H_1 , may be one of the following:
 - $H_1: \theta \neq \theta_0$
 - $H_1: \theta > \theta_0$
 - $H_1: \theta < \theta_0$
 - $H_1: \theta = \theta_1$

Simple hypothesis

- one involving a single value for the population parameter

Composite hypothesis

- one involving more than one value for the population parameter

One-sided significance test

- involves a composite alternative hypothesis of one of the following types:
 - $H_1: \theta > \theta_0$
 - $H_1: \theta < \theta_0$

Two-sided significance test

- involves a composite alternative hypothesis of the following types:
 - $H_1: \theta \neq \theta_0$

One- and two-tailed tests

- the 5% significance level is usually divided into two parts: 2.5% in the upper tail and 2.5% in the lower tail
- for example:
 - a drug trial comparing drugs X and Y
 - the upper tail values lead to the conclusion that X is better than Y, and the lower tail that Y is better than X
 - if you predict that X is the better drug, and maintain that whether Y is better is not relevant, you can add the probabilities in the two tails together and put in the upper tail
- one-tail tests make it more likely that a significant result is obtained
- if you are even vaguely interested in a result not consistent with your directional hypothesis, a two-tailed test should be used

Critical region

- is the region of the range of the random variable X such that if the observed value x falls in it the null hypothesis is rejected

Critical value

- is (are) the value(s) of the test statistic expected from the null hypothesis that define the boundaries of the critical region

Significance level

- the significance level, α , is the size of the critical region and represents the following probability:

$$\alpha = P(\text{type I error})$$

Type I and II errors

- **Type I error** – the null hypothesis is erroneously rejected when it is in fact true (i.e. a false positive)
 - the probability of making a type I error is equal to the p value and is expressed as α – for example an alpha of 0.05 means that there is a 5 % chance of erroneously rejecting the null hypothesis
 - multiple comparisons (e.g. throwing the dice) increase the risk of Type I error
 - the probability level can be reset to account for this using a *Bonferroni correction test*
- **Type II error** – the null hypothesis is erroneously accepted when it is in fact false (i.e. a false negative)
 - occur when the sample size is not large enough, or the variance is too large (e.g. from noise in measurement)
 - the probability of making a type II error is expressed as β
 - the conventional level of accepted risk is usually 20%

Levels of significance

- the risk of making a type I error is normally set at a maximum of 5%, i.e. 0.05
- therefore, significance is arbitrarily defined as being less than 0.05, or $P < 0.05$
- if the 95% CI for a mean difference contains the value zero, then the null hypothesis that the population mean difference was zero would not be rejected at the 5% level
- if the 95% CI for the true mean difference did *not* contain the value zero, then the null hypothesis of no difference *would* be rejected at the 5% level

Power calculations

- the **power** of a study is defined as the probability of rejecting the null hypothesis when a true difference exists, i.e. it is the sensitivity of the test to detect an effect
- it is calculated as follows:

$$\text{Power} = 1 - \beta$$

- since β is arbitrarily set at 0.2 – a typical study has a 0.8 chance, or 80% power, to detect a specified degree of difference (the **effect size**)
- power is related to:
 1. sample size
 2. size of effect
 3. reliability of your measures
 4. adopted significance level
- power can be increased by increasing 1. and 3.

Effect sizes

- effect sizes refer to the to the minimal size of effect that would be of clinical value
- they are calculated as the difference between the two means (*Cohen's d*) divided by the standard deviation in controls (*standardized difference*)
- numerically equivalent to 'z scores'
- the lower the significance level, the more subjects you will need for a given effect size
- two-tailed tests require larger samples than one-tailed tests

Comparing groups - Parametric and Non-parametric tests

- **Parametric statistics:**
 - assume a normal distribution
 - the scores within each sample must be quite independent of each other
- **Non-parametric statistics:**
 - do not assume a normal distribution
 - often based on ordered values or data that are arranged in ascending (or descending) order, and generally examine differences between groups in medians rather than means
 - non-parametric methods do not require precise values, only their relative ranks
 - less powerful
 - less dependent on sample size
 - sample size should be less than 50
 - increase the chance of type II error – for a given sample size, they are more likely to detect a statistically significant result than a parametric test

Goal, or question asked	Type of data		
	<i>Continuous measures (parametric)</i>	<i>Ranked measures (non-parametric)</i>	<i>Categorical measures (non-parametric)</i>
<i>Describing the data</i>			
Summarize the data in this group	Mean Standard deviation	Median Quartiles Percentiles	Proportions
<i>Looking for differences</i>			
Compare data from this group with hypothetical data	1-sample <i>t</i> test	1-sample Wilcoxon test	χ^2 -squared
Is there a difference between these two independent groups	Independent <i>t</i> test	Mann-Whitney U test	χ^2 -squared Fisher's exact test
Is there a difference between these two paired groups	Paired (dependent) <i>t</i> test	Wilcoxon signed rank test	McNemar's test
Is there a difference between these three or more independent groups	1-way ANOVA	Kruskal-Wallis test	χ^2 -squared
Is there a difference between these two or more independent groups, when they are categorized in several ways	Multiway ANOVA	Friedman test	Log-linear analysis
Is there a difference between these three or more matched groups	Repeated measures ANOVA	Friedman test	Cochrane Q test
<i>Associations between variables</i>			
Is there an association between two variables	Pearson product moment correlation	Spearman rank correlation	Contingency coefficient Phi coefficient Kappa

<i>Predicting</i>			
Can I predict one value from another	Univariate regression		Logistic regression Log-linear analysis
Can I predict one value from several others	Multiple regression		Multiple logistic regression Log-linear analysis

Parametric tests

Differences between means: The t-test

- allows the researcher to ask how likely it is that the *difference between two means* occurred by chance rather than as a function of the independent variable
- an *independent t-test* is used if the two groups are different subjects, a *paired t-test* if they are the same subjects at different time points
- conducting a *t* test of statistical significance requires the use of three descriptive statistics:
 1. the size of the observed effect - *the difference between the means*
 2. the *standard deviation* of scores in each group
 - the difference between groups is more likely to be significant when each group's S.D. is small
 - if variability is high enough that the scores overlap, the mean difference, though large, may not be statistically significant
 3. the *sample size*, *N*
 - the larger the number of subjects or observations, the more likely it is that a given difference between the means is significant
- the value of the *t* statistic is calculated as the difference between the means divided between the standard error (S.D. / \sqrt{N}) of the difference in means
- *t* increases if:
 - the differences between the means is larger
 - *N* increases
 - standard deviations are small
- to determine what a particular *t* means, we then have to use the value of *N* and the ***t* table**
- **degrees of freedom**, or **df** is $N_1 + N_2 - 2$
- if the *t* value is larger than the one in the appropriate row (governed by the df) in the 0.05 column, the difference between means is considered significant

Differences in means: the F statistic and ANOVA

- with continuous normally distributed data in **three or more groups**, the appropriate test is an **analysis of variance (ANOVA)**
- it is similar to the *t* test but is used for three or more groups
- *one-way ANOVA* is used for independent measures
- *repeated measures ANOVA* is used for several measures in the same individuals
- in any ANOVA, the variability of observations around the mean within a group is compared with the variability between the group means

- if the null hypothesis is true, the between-groups mean square and the within-groups mean square should be close to each other, and one divided by the other should be close to 1 – this is the **F statistic**
- the F statistic tells you if there are significant differences between the groups
- the F distribution is the ratio of variances derived from two samples of the same population

Non-parametric tests

Mann-Whitney U test

- this is the non-parametric equivalent of the independent *t* test
- it examines differences between medians rather than means
- it should be used, for example, if the data is very skewed

Differences in proportions: the Chi-squared (χ^2) test

- two proportions of discontinuous variables, from two independent samples, must be compared using the **chi-squared test**
- the data are in the form of frequency counts presented in a matrix or contingency table, for example:

Treatment	Improved	Not improved
Drug	a	b
Placebo	c	d

- Chi-squared tests estimate how far the pattern of the obtained data in the cells differs from what one might expect if the probabilities related to row and column categories are independent
- chi-squared values are very sensitive to sample size
- must be applied to original data, not to percentages

Fisher's exact test

- is an alternative to the Chi-squared for a 2x2 table
- it is not affected by low expected values

Wilcoxon matched-pairs signed ranks test

- this is the non-parametric equivalent of the paired *t* test – it can be used for matched pairs, or measures taken from the same subjects on two occasions

McNemar's test

- is a variant of the Chi-squared test used to assess change in subjects before and after an intervention, or when individual pair matching is used

Kruskal-Wallis one-way analysis of variance by ranks

- this examines differences in three or more independent groups when the data are ranked
- there should be at least 5 subjects per group

- it is non-parametric and therefore less powerful than ANOVA

Friedman two-way analysis of variance by ranks

- used for two-way analysis of variance on ranked scores

Correlation

Correlation and Correlation Coefficients

- a **positive correlation** means that two variables increase or decrease together
- a **negative correlation** means the variables move in opposite directions: when one increases, the other decreases
- the correlation between the variables is a measure of *test-retest reliability*
- the accuracy of predictions about one variable from knowledge of another depends on the **strength** of the correlation - the weaker the correlation, the less one variable can tell you about the other

Pearson product moment correlation

- to describe the strength of a correlation, a statistic called the **correlation coefficient** is used (also known as the '*Pearson product-moment correlation*')
 - it is given the symbol r , and can vary from +1.00 (strongly positive) to -1.00 (strongly negative); 0 implies no relationship
 - the *absolute value* of r indicates the strength of the relationship
- ideally, there should be at least 30 subjects per group to calculate an r
- r is not sensitive to changes in absolute magnitude
- it is the most commonly used statistic to measure the extent of agreement between two raters

Kappa values

- are a measure of correlation
- kappa values are better at measuring the extent of agreement between two raters than product moment correlation coefficient
- kappa statistics give a lower level of agreement than simple correlation measures

Spearman rank correlation (ρ)

- if either the dependent or independent variables are not normally distributed, a **Spearman rank** correlation coefficient is more appropriate – i.e. it is a non-parametric test

Regression

Multivariate analysis

- considers the relationship between three or more variables
- involves manipulation of matrix data
- in a multiple regression, as in any multivariate analysis, a measure of association is calculated taking a number of confounders in account simultaneously
 - the predictor variable his being predicted from a linear combination of outcome variables

MANOVA

- multivariate analysis of variance
- it is a form of multivariate analysis
- the predictor variables are discrete, and outcome variables are continuous

Simple linear regression

- is used to predict the score on a dependent variable from knowledge of scores on independent variables
- if there is a linear relationship between two variables, then one (the dependent variable, y) can be calculated from another (the independent variable, x) from the following equation:

$$y = a + bx$$

- y = the dependent variable
- b = the **regression coefficient** (Pearson / Spearman)
- a = the **intercept** on the y axis (a constant)

- if the dependent variable is dichotomous, **logistic regression** is the technique of choice

- if subjects have not all been followed up for an equal length of time, the **Cox proportional hazards model** is appropriate

More complex statistical methods

Factor Analysis

- an attempt to express a set of multivariate data as a linear function of unobserved, underlying dimensions, or (common) factors together with error terms (specific factors)
- it is used to study interrelationships among a set of variables without reference to a criterion
- a matrix of correlations between every variable is created
- segregates data into the minimum number of dimensions that define a group, (e.g. used to generate positive, negative, and disorganization syndromes in schizophrenia)

- the *principal components analysis* uses linear combinations
- the principal components are independent of each other

Regression analysis

- determines the predictive power of each successive variable upon the outcome variable
 - uses multivariate statistical tests (e.g. MANOVA)

Cluster analysis

- applied to a multivariate dataset which derives homogeneous groups or clusters of cases based on their values for the variable set
- segregates data into groups, but with some overlap (e.g. Paykel's classification of depression)

Clinical trials

Classification

- Phase I clinical pharmacology and toxicity
- Phase II initial clinical investigation
- Phase III full-scale treatment evaluation
- Phase IV post-marketing surveillance

Treatment: Randomized Controlled Trials

Key characteristics

- **Blinding** – ensures that doctors and patients do not know which treatment they are on
 - reduces the patient's placebo response (in *single-blinding*)
 - reduces the doctor's over zealous desire to do good (in *double-blinding*)
- **Randomization** – aims to provide two more-or-less identical groups of patients so that any differences observed after the treatment can be ascribed to the differing treatment
 - usually requires an independent person to consult random number tables to decide which treatment consecutive trial recruits will receive
 - the results are written on cards which are sealed, and only opened by an independent investigator when an appropriate patient is recruited
 - a lack of adequate randomization and concealment typically overestimates the benefit of a new treatment by 30-40 %
- *Dichotomous outcomes* simplify the **intention-to-treat analysis** – drop-outs are regarded as treatment failures rather than ignored

Common problems

- failure of true randomization
- lack of concealment of the allocation
- a lack of blind treatment and/ or outcome assessment
- ignoring missing data
- using inappropriate and/ or too many outcome measures

Pragmatic trials

- include all available patients in a given location and therefore has generalizability, but the scientific quality of the trial is often compromised
- tries to counteract the fact that trial patients often differ from average patients (e.g. detained patients cannot ethically be included), and co-morbidity (e.g. drug misuse) is common

Definitions

Experimental event rate (EER)

- the rate of a dichotomous (yes/no) outcome in the group receiving the new intervention
- the event may be positive (e.g. recovery) or negative (e.g. side-effects, relapse)
- can be expressed as a percentage, or a fraction

Control event rate (CER)

- the rate of a dichotomous outcome or event in the group receiving the control, or standard intervention

Relative risk reduction (RRR)

- the proportional reduction in event rates between the treatment groups
- **$RRR = \frac{EER - CER}{CER}$**

Absolute risk reduction (ARR)

- the absolute difference in event rates in experimental and control groups, expressed as a proportion or a percentage
- **$ARR = EER - CER$**

Number needed to treat (NNT)

- the number of people one needs to treat with a specific intervention to achieve one additional favourable outcome
- **$NNT = \frac{1}{ARR}$**

Confidence intervals (95 % CI)

- the range of values within which we can be 95 % sure that the true value lies for the whole population of patients from whom the study patients were selected
- the 95 % CI of the NNT is the reciprocals of the CIs for the ARR

Question Checklist

1. Was the research valid?
 - a) Was the assignment of patients to treatments randomized?
 - b) Was the randomization list concealed?
 - c) Were all subjects who entered the trial accounted for at its conclusion?
 - d) Were they analyzed in the groups to which they were randomized?
 - e) Were subjects, clinicians and raters blind to which treatment was being received?
 - f) Aside from the experimental treatment, were the groups treated equally?
 - g) Were the groups similar at the start of the trial?
2. Is the research important?
 - a) Absolute risk reduction =?
 - b) Number needed to treat =?
3. Can I apply it to my patients?
 - a) Is this patient group so different from those in the trial?
 - b) How great would the benefit be for this particular patient group?
4. Is it consistent with my patients' values and preferences?
 - a) Do I have a clear assessment of the patients' values and preferences?
 - b) Are they met by this intervention and its potential consequences?

Worked example

- Treatment drop-outs with olanzapine as compared to haloperidol in the treatment of schizophrenia (Tollefson *et al.* 1997)

	OLANZAPINE	HALOPERIDOL	
Drop-out	446	337	739
Stay-in	866	299	1229
	1312	636	1948

1. $EER = 446 / 1312 = 0.34 = \mathbf{34\%}$
2. $CER = 337 / 636 = 0.53 = \mathbf{53\%}$
3. $ARR = 0.34 - 0.53 = -0.19 = \mathbf{19\%}$ (i.e. 19 drop outs are avoided for every 100 patients treated)
4. $NNT = 1 / 0.19 = \mathbf{5.3}$ (i.e. 6 patients need to be treated with olanzapine to avoid one treatment drop-out)

Diagnosis: Case-control studies

Question Checklist

1. Is this diagnosis study valid?
 - a) Was there an independent, blind comparison with a reference ('gold') standard diagnosis?
 - b) Was the diagnostic test evaluated in an appropriate spectrum of patients (i.e. those in whom it would be used in practice)?
 - c) Was the reference standard applied regardless of the diagnostic test results?
2. Can I apply it to my patients?
 - a) Is the diagnostic test available, affordable, accurate, and precise in your setting?
 - b) Can you generate a clinically sensible estimate of your patient's pre-test probability?
 - c) Will the resulting post-test probabilities affect your management and help your patient?

Definitions

	Gold Standard +ve	Gold Standard -ve
Test +ve	<i>a</i>	<i>b</i>
Test -ve	<i>c</i>	<i>d</i>

a = true positive

b = false positive

c = false negative

d = true negative

Sensitivity

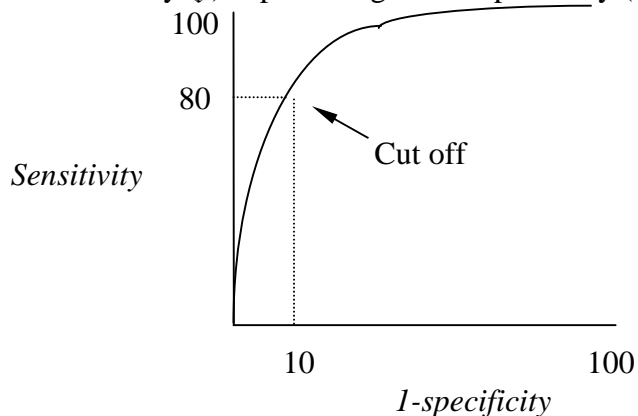
- the proportion of true cases correctly identified by the test
- sensitivity is an index for caseness
- increasing the caseness threshold decreases sensitivity
- $SENS = a / a+c$

Specificity

- the proportion of true negatives correctly identified by the test
- increasing caseness threshold increases specificity
- $SPEC = d / b+d$
- **SpIN: Specificity rules the diagnosis in**
- **SnOUT: Sensitivity rules diagnosis out**

Receiver Operating Characteristics (ROC) curve

- sensitivity (y) is plotted against 1-specificity (x)



- the graph allows one to calculate the best cut-off for your population
 - a higher cut-off means that one is less likely to get false positives
- the area under the graph can be compared to evaluate the tests
 - the greater the area, the better the test

Positive Predictive value

- indicates the value of a positive score, i.e. the proportion of positives that are true positives
- **PPV = $a / a + b$**

Negative Predictive value

- indicates the value of a negative score, i.e. the proportion of negative results that are true negatives
- **NPV = $d / c + d$**
- both the PPV and NPV vary with prevalence

Likelihood ratios (LR)

- the **likelihood ratio of a positive test** is: **LR+ = $\frac{\text{sensitivity}}{1 - \text{specificity}}$**
- the **likelihood ratio of a negative test** is the odds that a negative result will present in a patient with the target disorder compared to a patient without the disorder
- a LR- of 0.1 or less is useful
- **LR- = $\frac{1 - \text{sensitivity}}{\text{specificity}}$**

Pre- and Post-test Probability

- **pre-test probability** = prevalence

- **post-test probability** = probability that a patient scoring positive (or negative) on the test actually has the disorder

$$\text{PostTP} = \frac{\text{post-test odds}}{1 + \text{post-test odds}}$$

Pre- and Post-test Odds

- **Pre-test Odds** = $1 - \text{pre-test probability} / \text{pre-test probability}$
- **Post-test Odds** = LR x pre-test odds

Efficiency

- is measured by the proportion of all true results:

$$\text{Efficiency} = \frac{a}{a+b+c+d}$$

Odds ratios (OR)

- are a measure of the strength of a treatment effect or an aetiological association
- calculated by comparing outcome rates in exposed and non-exposed persons in a controlled trial
- an OR of 1.0 ('unity') reflects exactly the same outcome rates in both groups, i.e. no effect
- **OR = ad / bc**

Worked example

- the performance of the HADS in identifying depression in cancer patients was compared to a structured interview (the Gold standard)

	Gold standard +ve	Gold standard -ve	
HADS +ve	80	40	120
HADS -ve	20	360	380
	100	400	500

- **sensitivity** = $a / a+c = 80 / 100 = 0.8$ or 80 %
- **specificity** = $d / b+d = 360 / 400 = 0.9$ or 90 %
- **PPV** = $a / a+b = 80 / 120 = 0.67$ or 67 %
- **NPV** = $d / c+d = 360 / 380 = 0.95$ or 95 %
- **LR+** = $0.8 / 1-0.9 = 8$
- **LR-** = $1-0.8 / 0.9 = 0.22$
- **pre-test probability** = $100 / 500 = 0.2$ or 20 %
- **pre-test odds** = $1-0.2 / 0.2 = 4$
- **post-test odds** = $8 \times 4 = 32$
- **post-test probability** = $32 / 1+32 = 0.97$ or 97 %

Prognosis: Cohort studies

Relative risk

- calculated in the same way as odds ratios, but using data from cohort studies
- RR = ratio of odds of a positive outcome in the exposed ($a / a + b$) to the odds of a positive outcome in the non-exposed ($c / c + d$)
- **RR = $\frac{a / a + b}{c / c + d}$**

Confidence Intervals

- defined as the range within which the true magnitude of effect lies with a certain degree of assurance (usually 95 %)
- if an Odds Ratio includes 1.0 (unity) then the corresponding P value is by definition greater than 0.05 and hence not significant
- if the 95 % CI is narrow, we can have greater faith that the result is close to the actual value
- if the CI is wide, the data are compatible with a true increased/ decreased risk but the sample size was not sufficient to have adequate power to exclude chance as an explanation of the findings

Worked example

- Is RISPERIDONE generally 'better' in the treatment of schizophrenia than conventional neuroleptics?

	NOT improved	Improved	
RISPERIDONE	4	17	21
Other neuroleptics	8	12	20
	12	29	41

- OR = $4 \times 12 / 17 \times 8 = 48 / 136 = 0.35$
- i.e. patients on RISPERIDONE were 0.35 times more likely not to improve, or 2.86 (1/0.35) times more likely to improve