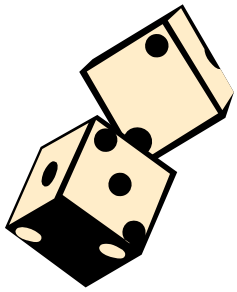


Biostatistics in Oral Health Care

Core Curriculum 2001 Dental Faculty University of Oslo

Asbjørn Jokstad
University of Oslo

What is statistics?

Examples

1. Political polls
2. Birth weights
3. What's the chance of ...?
4. How many ... ?
5. Are there any differences between ...?

The term "statistics" may have several meanings:
1. Statistics is the term used for summary values derived from raw data.
2. Statistics is the use of information from a **sample of individuals** to draw **inferences** about a wider **population** of like individuals.
3. Statistics is the tool we use to demonstrate real differences or effects of whatever we are testing. Often we do that by "disproving" (or challenge or contradict) the so-called null hypothesis of no difference.
4. Statistical tests are commonly used to

- describe the likely values of some measurement
- compare variation in two or more groups to detect differences

Statistical analyses never prove anything, but allows us to put limits to our uncertainty. Since statistical analyses rarely lead to definite answers, we should always indicate a degree of uncertainty in our answers.

Two basic approaches in statistical analyses are **estimation** and **hypothesis testing**.

A prerequisite for correct statistical analyses of collected data is a proper study design. A careful planning of an optimal study design must precede any study. Unfortunately, this is often not the case, and statistical errors are discovered at the moment the investigator approaches a statistician to obtain help with their data.

A problem one wish to avoid in clinical trials is **bias.** Several types of bias can be introduced: selection- or sampling bias, recall bias, examination bias, etc. Sampling bias occurs usually because the individuals in the study samples are not chosen at **random**.

Furthermore, study reports sometimes lack a complete description of the prerequisites to carry out correct statistical tests. Usually, this is due to poor reporting, but occasionally clear statistical misuse can be identified (an estimate is approx. 5% of all reports in medicine). In general, statistical errors may be placed into six categories:

Errors in design	Errors in execution
Errors in analysis	Errors in presentation
Errors in omission	Errors in interpretation

Detailed descriptions of the different types of errors are described in chapter 16 of the textbook, and should be read in detail!

Lack of understanding statistics may lead to unexpected, hilareous and sometimes even dramatic headlines in newspapers.

However, there are other more ethical implications of misusing or misinterpreting statistic:

- **misuse of patients by exposing them to unjustified risk and inconvenience**
- **misuse of resources, including the researchers time, which could be better employed on more valuable activities**
- **publishing misleading results with consequences such as:**
    - **the carrying out of unnecessary further work**
    - **it may prove impossible to get ethics committee approval to carry out further research because a published study has found the experimental treatment beneficial, even though the study was flawed**
    - **leading other scientists to follow false lines of investigation**
    - **future patients may receive an inferior treatment, either as a direct consequence of the results of the study or possibly by the delay in the introduction of a truly effective treatment**
    - **if the results go unchallenged the researchers may use the same inferior statistical methods in future research, and others may copy them.**

Data may be either categorical or numerical

I. CATEGORICAL or QUALITATIVE DATA

May consist of two or multiple categories.

Two categories.
Often described as *dichotomous*, *binary*, *attribute*, *yes-no* or *0-1* data or variables
Examples from dentistry:
caries: yes/no
dentist: male/female

Multiple categories, are on a

Nominal measurement level – (nominal variable) i.e., the categories have no ordering
Examples from dentistry:
Deciduous dentition/ mixed /permanent dentition
Angle class I/II/III
Cavity class I, II, III, IV, V
Type of filling: amalgam, composite, gold, etc.

Ordinal measurement level (ordinal variable) i.e., the categories have a natural or defined order (sometimes the data are reduced to two categories)
Examples from dentistry:
Caries: enamel (1), dentoenamel border (2), outer 1/2 of dentin (3), inner 1/2 dentin (4).
Pain: small, moderate, severe, unbearable
Tooth mobility or tooth fracture O-I-II-III-IV
Wear teeth: none, in enamel, dentin, secondary dentin, pulp
Plaque index (Løe & Silness)
Gingival index (Silness & Løe)
Bleeding index
Tooth prognosis: good, fair, bad
CPITN index

II. NUMERICAL or QUANTITATIVE DATA

May be on an **interval measurement scale** or on a **ratio measurement scale.**
In both scales, the differences between the values are constant in an interval scale. In interval scales the zero is arbitrarily set. e.g. temperature. The data or variables may be discrete or continuous

Discrete, only certain numerical values can be used. Most typical examples are different types of counts of events.
Examples from dentistry:
Visits to the dentist or # Carious lesions /year
DMFT
Dentists' work experience

Continuous the observations are not restricted to certain values, except by the accuracy of the measuring method or instrument. Some times, continuous variables are discretely recorded due to rounding off of digits, or e.g., 5-year age groups.
Examples from dentistry:
Strep.Mutans count/tooth
Lactobacillus/saliva sample
Dentin bonding agent strength
Dental materials - physical properties

III. OTHER TYPES OF DATA

Ranks e.g. in sports competitions or exam results. Ranks are rarely used in medical statistics. Athough the concept of ranking is incorporated in so-called non-paramatric tests..
Examples in dentistry may be:
Preference of impression material.

Percentages often the ratio of two quantities. Cautiousness is needed when presenting and analysing such data.
Examples in dentistry may be:
Carious lesions / number of remaining teeth
number of amalgam fillings / number of total fillings
percent reduction of VSC

Rates and ratios are more common in epidemiology.
Examples in dentistry may be:
number of amelogenesis imp. /1000 patients,
number of cavities/year

Scores when using scores, it becomes necessary to include the inter- and intra-examiner agreement, as well as detailed description of criteria consensus and weighting.
Examples in dentistry may be:
USPHS (US Public health system) criteria
Pain dysfunction scores.

VAS-scales: (Visual analogue scales)
Examples in dentistry may be.
Pain description
Course evaluations

Odds: Are defined as the number of events/number of non-events. E.g. when the

distribution of new-born boys and girls are 51% & 49% the odds of having a boy becomes 51/49 = 1.04. Odds are always used in case-control studies where the disease prevalence is unknown.

A <u>variable:</u> ***is a term used to denote anything that varies within a set of data.***

The type of data will determine which statistical tests are appropriate. Several statistical tests may be applied for some data, while others may require specific tests. The main difference between tests are based on whether the variables are continuous or categorical.

The term "**censored data**" is used when a measure is below the detection limit of a test or apparatus, although probably not zero. Another use of the term is in survival statistics, when at the end of a trial the term is used on the data or event that is unchanged.

One key concept underlying the subject of statistics is that of **variability**.
Variability of measurements is due to known causes, or unexplained. I.e. **random variation**.

Exercise

1. Identify the variables in two articles/reports related to your area of interest in dentistry, and categorise the variables into correct type of data.

Descriptive statistics

Data can be presented by different methods, depending on the type of data.

I. Qualitative data

Number, frequency or percentage of categories relative to the total number of patients are often presented in a table or graphically shown in a **bar diagram.**

- <u>A table should always be self-explaining.</u>
- In the bar diagram, the adjacent bars should not touch, to indicate that the variable is not continuous.
- **The vertical axis of a bar diagram should always start at zero.**
- Under each bar, an appropriate text is given.
- A typical bar chart simply shows the number of subjects per category with % added.

II. Quantitative data

1. <u>Frequency distributions</u>

The number or percentages of observations are often presented as the **frequency distribution.** In the **relative frequency distribution**, all frequencies are converted into percentages.

If the relative frequencies are summed at each level, we find the **cumulative relative frequency distribution.**

Frequency distributions can be shown graphically shown **histograms. (**An exception is that if the data are discrete with few possible values, a bar diagram should be made.)

- The adjacent bars should touch, to indicate that the variable is continuous.
- The y-scale may be either on an arithmetic (equal distance means equal absolute distance) or logarithmic (equal distance means equal proportional distance) scale.
- Some sort of grouping must take place, but the intervals should not go beyond the precision of the data.
- Open intervals must be dealt with, either by closing them or take extreme values out and present them separately.
- Vertical axes should always begin at zero, and there should be no breaks in the scale.
- Three-dimensional effects should be avoided.
- The width of the intervals may vary, therefore, remember that it is the <u>area of each bar</u> that is proportional to the frequency, not the height.

A **frequency polygon** is a plot where the mid-tops of all the vertical bars are joined by straight lines, to smooth out random variation. By doing so guesswork on the underlying data structure in

larger populations are done. Frequency polygons or curves are often more practical than histograms when displaying cumulative frequency distributions to show the **cumulated frequency polygon.**

A modification of the histograms is the **stem and leaf diagram.**

2. Measures of central tendency

		Advantage	Disadvantage
arithmetic mean,	$\overline{X} = \sum_{i=1}^{n} \frac{x_i}{n}$	Can compare groups	Affected by outliers and skewness
median	$\tilde{X}$ = middle observation	Not affected by outliers	Poor use of information
mode	largest frequency of observations		

3. Measures of variability

1. Range $X_{max} - X_{min}$ Very sensitive to outliers.

2. Variance: is a measure of the average of the individual deviations from the mean.

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$$

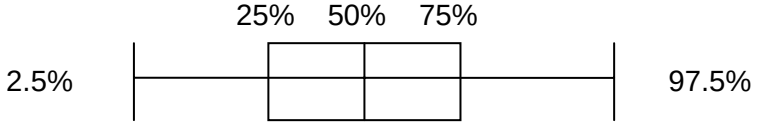
It is calculated by summing the square of the distances between the observations divided by n-1. The advantage of using variance is that all measurement values are used, and that it has an interpretation in the normal distribution. A disadvantage is that the formula is complicated, and there is no clear interpretation if the distribution of the data is asymmetric, i.e. **normal**.

When the distribution is not normal it is described as **skewed. Positive skewed** data are fairly common. In these cases, the logarithms of the values are often used to obtain a more symmetric distribution. Alternatively, the median and centiles should be used. It is probable that the distribution is skewed if the standard deviation is more than half the mean value. (Another way to say this is that the mean is less than twice the SD)

Since the variance is not in the same units as the raw data, the square root is used, which is called the **standard deviation** abbreviated SD, sd or s. This is the most used measure of variability for practical purposes today.

$$S = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$$

3. Centiles, e.g., 5th and 95th centile . The area between these two values indicate also the 90% **central range**, other centiles are the **(inter-) quartiles**, and the difference between the 25th and 75th centiles (the quartiles) is the **inter-quartile range**, these values are often presented in a **box-whisker plot**. Where in addition, the 2.5 and 97.5% values are indicated by the "whiskers".



4. Coefficient of variation (CV)
CV is defined as the standard deviation divided by the mean value x 100.

$$CV = \frac{S}{\bar{X}} * 100\%$$

CV is often used when groups are compared with respect to variability, e.g., in quality control of laboratory blood sample measurements.

Other diagrams

1. Curve diagrams
Are often used to show how much a variable is changed as a function of another variable. A time series measurement is a typical example of a curve diagram.

2. Scatter diagrams
Are often used to study if two or more variables covariates with another. The units may be continuous or discrete numbers, percentages, area measurements, etc. At least one of the axes must be ordered.

Exercises

1. The following 35 measurements of bond strength (Mpa) were made of a new dentin bonding agent.

3	6	9	3	5	3	9
3	1	2	1	6	5	1
2	1	3	1	5	4	2
4	6	5	9	3	6	2
8	6	1	6	1	2	6

Report the following measures: range, mean, median, mode, variance, standard deviation, 25 and 75 percentile values and draw a box-plot of the relevant values.

2. In a test for strength of an alloy for partial dentures the following values were recorded (N).

2	5	6	7	7	1
9	6	12	2	6	11
1	5	12	11	8	4
8	2	10	9	8	8
4	9	1	8	5	9

1. Make a frequency distribution table and a histogram with intervals of 2N.
2. What is the mean, median and variance?
3. Would you find this alloy suitable for prostheses? (Mean value for vitallium is 8N (SD.5)).

3. In group A the mean systolic blood pressure is 145 mmHg (SD = 15). In group B the mean is also 145 mmHG with SD = 10 mmHg. The distribution of values in both groups is unimodal and normal. Are the following statements correct?

a) The variation regarding the level of the blood pressure is larger in group A than in group B.

b) There are more people with a blood pressure over 160 mmHG in group A than in group B.

c) There are more people with a blood pressure over 130 mmHG in group A than in group B.

d) Half of the people in group A have a systolic blood pressure between the upper and lower quartile values

An essential concept in the application of statistical methods is that of **probability**.

1. The **probability** - P- of some specific outcome in an experiment is the proportion of times that that outcome would occur if we repeated an experiment or observation a large number of times.

   Thus, the frequency of a certain outcome A resembles P (A) when n increases:

   $$\text{freq.}(A) = \frac{n_A}{n} \rightarrow P(A)$$

2. The probability of an outcome A is always between zero and one:

   $$0 <= P(A) <= 1$$

3. The probability of a specific outcome A + the probability of the negative of outcome A is always 1.

   $$\frac{n_A}{n} + \frac{n_{\hat{A}}}{n} = n \rightarrow \frac{n_A}{n} = 1 - \frac{n_{\hat{A}}}{n} \rightarrow P(A) = 1 - P(\hat{A})$$

Probability calculations are commonly used in biomedicine to estimate the risk or probability of e.g. adverse effects from using drugs or smoking, the probability of cancer from x-ray exposure, etc. More advanced use of such probability calculations are also used, e.g., in computer expert systems, to calculate probability of certain diagnoses on basis of specific findings. Another use in medicine is calculations of the chances of giving birth to a baby with inherited defects, given certain information such as genetic predisposition or specific illness of one or both parents, etc.

The typical examples of observation outcomes used in statistical textbooks are:

- Coin tossing: head or tail? We "know" that each outcome will be P(tail) = P(head) = ½ → P(tail)=.5
- Dice rolling: values 1-2-3-4-5-6 . We "know" that each outcome will be P(1)=P(2)…=P(6) = 1/6 → = P =.17. We could also say that the probability of odds and even numbers are P = .5
- Deck of cards: the outcome of each card value is P = 1/13, each colour P = .5 and each type P =.25 etc.

The P values in these examples are something we "know" intuitively, but can also be described by using complicated mathematical formulas. These formulas will not be described in the present course of statistics.

The same mathematical formulas can also be used to calculate how many times e.g., we would pick e.g. a card value above 10 with a red colour, or the probability of rolling a dice three times in a row with the value 6.

Theoretical probability distributions

The same line of thoughts - and mathematical formulas- can be applied when the number of observation outcomes increases. Ideas of theoretical **probability distributions** are, therefore, important in this context. These distributions have been defined mathematically. The probabilities of different values- or more correct the area within a range of values- within these distributions can, therefore, be calculated.

Examples of theoretical probability distributions with continuous values are the **Normal or Gaussian** and the **Lognormal** distribution. Examples of theoretical distributions of discrete values are the **Binomial** and the **Poisson** distributions.

Other theoretical distributions are the ***t*-**distribution, the **F-**distribution and the **Chi-square** distribution.

I. Normal distribution

The normal distribution is unimodal and symmetric with no upper and lower limits. It is completely described by two parameters, the mean and the standard deviation. The height of the frequency curve, called the **probability density,** is of no practical use. The force of the normal distribution is determined by the fact that the total area under the curve is always taken to be 1. The probability distribution is calculated by considering the area corresponding to a particular restricted range of values, e.g., the area to the left of the mean is 0.5.

Any position along the horizontal axis can be expressed as a distance of a number of standard deviations.
The distance is known as the **standard normal deviate, z** or **Normal score**. The area to the left of such z-values (Table B1), or within or outside such z-values (Table B2) has been defined.
Any symmetrical unimodal distribution of data can be converted into a standard normal distribution by subtracting the mean and dividing by the standard deviation.

$$z = \frac{X - \bar{X}}{SD}$$

$$z = \frac{75-55}{5} = 3 \rightarrow P = .9987$$

Thus, in a population where e.g. the mean age is 55 years and SD = 5 years the age 75 years correspond to

In a normal distribution the intervals:
mean±1 SD includes approx. 68% of the area under the curve (P = .6827)
mean±2 SD includes approx. 95% " (P = .9545)
mean±3 SD includes approx. 99% " (P = .9973)
(This can be read directly from table B2 in the textbook)

Exercises

1. A person is 185 cm in a population where the mean height is 180 cm (SD 5 cm). (The values of measured heights is normally distributed).
a) can this person - from a statistical point of view- be considered as tall ?- why?
b) how many are shorter than this person in the population?
c) how many are taller ?

2. The average distance from the incisal edge to the apex of the upper lateral incisor is 22 mm (SD = 4). The values of measured distance is normally distributed. In a large sample of incisors:
a. What is the probability of finding incisors longer than 26 mm?
b. What is the probability of finding incisors shorter than 18 mm?
c What is the probability of finding incisors in the area 14 to 30 mm?
d What is the length of the longest 5% of the incisors ?

3. An orthodontist has the option of two treatment choices for moving the incisors in a child patient. Previous studies have shown that an expensive technique A moves the teeth 2,0 mm (SD =0,5) in 2 months, while a slower technique and more inexpensive technique B moves the teeth 1,8 mm (SD= 0,2) in 4 months. A movement of 1,5 mm is considered a failure.
a. What is the chance of failure of technique A?
b. What is the chance of failure of technique B?

4. 400 Norwegian dentists worked in average 7.5 saturdays in the year 1995 (SD 1.5). The distribution of answers was unimodal and symmetric, i.e., normal distributed.
a. In this sample, how many dentists worked more than 9 saturdays?
b. In this sample, how many dentists worked less than 4.5 saturdays?
c. In this sample, how many saturdays did 95% of the dentists work?
d. Can these observations be generalized to represent all Norwegian dentists?

5. Assume that the bodylength of the dental students is normally distributed with mean=179,5 cm (SD = 6,44 cm). Let X be the length of a randomly chosen dental student. Find the following probablities:
a. P (175.0 < X < 185.0)
b. P(X>200.0)
c. P(185.0<X<190.0)

6. Stanford Binets IQ-scale is defined so that IQ=100 is the average in the population (SD = 15). Normal intelligence is in the range of 85-115. Those with intelligence between 70 and 55 are often called debile, those between 55 and 40 imbeciles, and those under 40 as idiotes. Find the following probabilities for a randonly schosen person to have:
a. normal IQ b. debile IQ c. imbecile IQ d. idiot IQ

II. Binomial distributions

The probabilities in a binomial distribution can be calculated using a general formula. It is important to remember that the outcome of each event or case tells us nothing of the other events or cases. As the sample size increases the binomial distribution becomes more symmetric and resembles the Normal distribution.

The binomial probability of r events is:

$$\binom{n}{r} p^r (1-p)^{n-r}$$

$\binom{n}{r}$ is the number of ways of choosing r items from n, and must be calculated

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$ n! is pronounced factorial, and means 1 * 2 * 3 * … * n

If the true proportion of events of interest is p, then in a sample of size n,
the mean of the binomial distribution is np
the standard deviation is: $\sqrt{np(1-p)}$

Exercise:

Dr. Feelgood claims that he can, to some extent, categorise patients into either a high or a low-medium caries risk group just by judging their family history and food consumption. This claim is to be tested in the cariologic department, where the history of 5 patients is revealed to him.
p = the probability of a correct guess in each of the 5 trials
r = the number of correct guesses

The distribution of a binomial variable with n=5 and p= .05 is given by:

r	0	1	2	3	4	5
Prob:	.031	.156	.313	.313	.156	.031

a) What is the value of p for an ordinary dentist without Dr Feelgood "abilities".

b) What is the distribution of variable r called?

c) In the trials, Dr. Feelgood got 4 out of the 5 guesses correct. Give an estimate of p.

d) Suppose now that Dr. Feelgood actually cannot diagnose correctly. What is then the probability of getting 4 or more correct diagnoses in 5 tries?

e) Does this experiment prove that Dr. Feelgood's p is greater than 0.5? why/why not

IV. The **Lognormal distribution is** often transformed to a normal distribution before analysis.

V. The **Poisson distribution** is appropriate for studying rare events. It is not used greatly in medical research.

Designing research & preparing for analysis

Empirical studies are observations of the reality with the aim of elucidating a scientific problem. Studies must have high external validity and internal validity in order to have scientific value. The formulation of a **hypothesis** is the first step in the design of a study. The importance of this detail cannot be underemphasized because it will subsequently influence the choice of study method. Empirical studies are based on two processes, the design phase and the analytic phase. Both phases must be planned and carried out according to a predefined plan.

Careful study design is the foundation of quality clinical research.

Study design includes two main components:

1. Choice of the population, sample size and sampling method

- restriction: qualitative criteria

- sampling: sampling variation added to random variation

- random, limited by place, time, or other criterion's
- cluster or blocks, stratified samples
- confounding and bias

- sample size power calculation

2 Choice of observation method

- Active manipulation, versus passive observation
- Random experiment- confounding - external validity
- Causal relationship - nature's own experiment
- Observational vs. experimental studies
- Time dimension related to observations

- inference about cause-effects
- Cross-sectional, longitudinal study, cohorts

- scale and measurement precision: information value - blinding, replication
- association direction: exposition: cohort, follow-up, experiments

- situation: case-control (case-referent), survey

- observation and analytic unit: individuals- groups
- data accumulation
- ad hoc data, prospective , retrospective: antecedent data
- manipulation

yes experimental study: random allocation yes controlled experiment
no -quasi-experimental
no : non-experimental study

- sampling according to exposition characteristics: follow-up study
- sampling according to effect characteristics: case-control study

Preparing for data analysis

- Data checking Categories, range
- logical checks incompatible variables
- outliers careful treatment
- missing data why?
- data screening normal plot , skewness, kurtosis,
- data transformation parametric tests, logit
- digit preference hidden time effects

Critical appraisal of scientific reports

Introduction- reason for starting the study

- Previous studies have been
  i) undersized or
  ii) have conflicting results or
  iii) demonstrate a difference which needs clarification
- A clinical trial is a planned experiment on human beings.
- The objective is to evaluate the effectiveness of one or more forms of treatment.

Material & methods

Subjects- clarity with which selected subjects are characterised

Adequate description includes:

- source of subjects: Dental school / Private practice patients, Dental students,School children,Other
- demographic data, distribution of age, sex, no.teeth, assessment of (oral) health status,
- description of diagnostic workup performed to determine health/disease status.
- the diagnostic criteria for entry into the trial, i.e., symptoms- disorder severity- disorder duration. The criteria should ensure that the patients have the condition being studied, could potentially benefit from the intervention and are willing and able to give informed consent.

- patient expectation for improvement is also frequently an important information

Number of eligible population not accepted for participation
Adequate description includes:
- total number of subjects, specifying potentially eligible and actually included
- the number of subjects excluded before randomisation along with relevant reasons for exclusions. Typical are subjects who have contraindications to the procedures, are unlikely to comply with the protocol or follow-up, or whom randomisation would be unethical as well as extraneous conditions.
- the outcome should ideally be compared to outcome of rejected subjects to obtain information about potential bias in selection

Therapeutic regimes- must be described in detail
The treatments should be defined by:
- A complete description of procedures followed instead of a name
- Information about the extent and frequency of treatment
- If applicable, placebo appearance and/or taste should be controlled to be identical to experimental agent, and the control adequately described.
- The follow up -schedule described in detail, including time, procedures performed and evaluations
- The treatment groups were studied concurrently
- The delay from allocation to commencement of treatment was... (short=acceptable)

Blinding- designed to eliminate bias
- The potential degree of blindness was used during the trial.
- Include the specific dates of the beginning and end of randomisation and of enrolling subjects

The randomisation can be based on:
- centralised office +++
- centralized pharmacy +++
- tables of random numbers +
- sealed envelopes +/-
- flip of coin (unbalanced group sizes) -
- ID-number -
- alternate patients -

Blinding- treatment allocation
- The mechanism of treatment allocation was...(e.g. sealed envelopes)

Blinding observers
- Blinding to therapy - if possible and
- Blinding to ongoing results
- Report the effectiveness of these blinding mechanisms.
- Test the success of randomisation- specific information about (pre-treatment/baseline) differences of prognostic factors, symptoms, and other characteristics in the groups
- Results of randomisation analyses- either statistical tests for significant differences or when data analysis take into account unbalanced randomisation

Stopping rules
- Define the criteria for non-adequate response and describe the alternate treatment for these.
- Include a statement about how decisions will be made to stop the study.
- Describe the number of subjects affected.

Treatment outcomes/error measurements
- All physical equipment used and sequence during the examination as well as duration should be specified. Report the uniformity of such standardised examinations.
- The criteria for outcome measures are... (satisfactory stated).

All clinically important outcome measures should be considered, including the appropriateness for using these outcomes
- Measure the intra-examiner error of the criteria used for defining health/disease/outcome before the trial and during the trial. If more than one examiner also the inter-examiner variation.
- The duration of post-treatment follow-up should be described

Size of the study
- Include the criteria for sample size calculation.

The level of the difference of clinical interest and structure of the outcome measure defines this.
- Preferably do a pre-study calculation of sample size based on considerations of statistical power

Statistics
- A statement adequately describing or referencing all the statistical procedures used.
- Why were the statistical methods used appropriate for the data?
- Were the statistical methods used correctly?

Results

Statistics

- Both test statistic and its significance levels are included
- Were no statistic differences the possibility of type 2 error should be mentioned and an estimate of the probability for this should be computed.
- Confidence intervals or SE of differences must be included
- The number of subjects evaluated at each time & variable values should preferably be given in table.
- Life table should be used when appropriate
- Regression or correlation analysis should be performed to allow for variables in prognostic factors
- Repeat measurements of outcomes of interest should be included.
- Account for multiple outcome measures and mult. statistical tests can lead to erroneous conclusions

Adherence to treatment/drop-out

- Describe the proportion of subjects who followed up each visit
- " " who completed the treatment.
- Reasons for dropouts are described separately for each treatment group.
- More than 15% loss of subjects is in general unacceptable
- Present results of the assigned group with and without withdrawals in the analysis.

Side effects

- The frequency of and type of side effects of treatment are described separately for each group.

Retrospective analysis

- Should be done for a number of prognostic factors - e.g. initial state

Discussion

- Discuss if the likely benefits are worth the potential harm and costs.

Conclusions

- Check that the conclusions drawn from the statistical analyses are justified

Principles of statistical analysis, sampling distributions, estimation

The relationship between **sample** and **population** is subject to uncertainty, and we use ideas of probability to indicate the uncertainty. This is called the theory of **statistical induction**.

One basic idea in statistical is that mean and standard deviations calculated in samples are used as **estimates** for what is true in the relevant population. This is because if we take a large number of samples from a specific population the distribution of the e.g. means of these samples, the so-called **sampling distribution,** will have the certain characteristics. The variability of sample variable values, e.g. the sample means will:

- be less among the means of large samples than among small samples
- be less than the variability of the individual observations in the population
- will increase with greater variability among the individual values in the population

Mathematically, it can be shown that the means of random samples has these properties:

1. The **expected value** of the **mean of the distribution of the sample means** is the same as the population mean. Further the expected value of the variance of a sample is the variance of the population.

2. The expected value of the standard deviation of the means of several samples is $\frac{\delta}{\sqrt{n}}$

δ is the standard deviation of the variable in the population and
n is the size of each sample. The quantity is known as the **standard error, or standard error of the mean, SEM**.

We can **estimate the standard error** from a single sample using the observed standard deviation, SD, in that sample.

3. The distribution of the sample means will be nearly normal whatever the distribution of the variable in the population as long as the samples are large enough.

This is the **central limit theorem**. The central limit theorem applies equally to sums and means.

These properties mean that we use the methods based on the Normal distribution to indicate the uncertainty of a sample mean as **an estimate** of the population mean.

Estimation

Quantification of the results by simple estimates is an essential part of the analysis of study data. One-value estimates of unknown real values are called **point estimates**. Examples of point estimates are mean, mean difference, correlation coefficient r, regression coefficient b, odds ratio, percent reduction etc.

The **standard error,** SE, is an indication of the variability among many sample point estimates.

The **confidence interval estimate**, CI, is another estimate via a range of values. The confidence interval is calculated on the basis of the SE-values.

Confidence interval

The confidence interval is an estimate of the underlying true value of the population mean. The confidence interval extends either side of the mean by a multiple of the standard error.
A 95% confidence interval is defined as the mean ± 1.96 x SE.
A 99% confidence interval is mean ± 2.58 x SE.

The interval between mean ±2SE will be a 95.4% confidence interval and between mean ±3SE the 99.7% confidence interval. Thus, we can say that we are respectively 95.4% or 99.7% confident that the values between the two confidence limits calculated from sample data include the unknown real values of the mean in the population.

Other standard errors that can be computed are:

Standard error of the difference between two sample means

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{SD_1^{\,2}}{n_1} + \frac{SD_2^{\,2}}{n_2}}$$

Standard error of a sample proportion

$$SE(P) = \sqrt{\frac{p(100-p)}{n}}$$

Standard error of the difference between two proportions

$$SE(P_1 - P_2) = \sqrt{\frac{p_1(100-p_1)}{n_1} + \frac{p_2(100-p_2)}{n_2}}$$

These standard errors can be used to construct confidence intervals for the difference in proportions in two independent samples, and for the difference in the means of values of a continuous variable, **as long as the samples are large**. For small samples, a slightly different approach is used, i.e., the ***t*-distribution,** is used for constructing confidence intervals. When the sample sizes decreases we use the *t*-values relevant to the correct sample size.

We will come back to this issue in a subsequent chapter.

Principles of analysis, hypothesis testing

The null hypothesis is often the negation of the research hypothesis. Having set up the null hypothesis we then evaluate the probability that we could have obtained the observed data, or data that were more extreme, if the null hypothesis was true.

Note that we never "prove" any research hypotheses in research.

The probability of obtaining our data if the null hypothesis is true is by calculating a **test statistic**- a value that we can compare with the known distribution of what we expect when the null hypothesis is true.

Test statistic =

$$\frac{\text{observed value - hypothesised value}}{\text{standard error of observed value}}$$

Usually, the hypothesised value is zero, so that the test statistic becomes the ratio of the observed quantity of interest to its standard error.

When analysing data we choose between statistical methods that make distribution assumptions called **parametric** methods, and those that make no assumptions about distributions, called **distribution-free** or **non-parametric** methods. These are sometimes termed **rank methods**. Most statistical methods are specific to a certain type of data. The major difference is that between continuous and categorical variables. Further, for continuous or ordered categorical variables there is also the possibility of using rank methods, which are of a much wider applicability.

The test statistic, whether it be t, F, chi-square, etc., calculated from our data will lead us to two and only two possible conclusions; that is either our data deviate significantly from zero or no difference; or do not deviate significantly from the null hypothesis of no difference. The decision is based on pre-determined cut-off points in the percentage of our probability distribution of the test statistic that we use. Cut off points are referred to as the critical values of the test statistics. The critical values are arbitrary and have no specific importance.

Significance level

The P-value is **the probability of having observed our data (or more extreme data) when the null hypothesis is true.**

Another way of expressing this is:
The p-value is the probability of making an error in concluding a difference when none really exists. We are saying in essence that we know our magnitude of error when we conclude a difference.

The p-level for concluding a difference can be very small but never zero, because certainty is never absolute in scientific research.

The inverse 1- P value does not add up to unity. Thus, **we never state the probability of a real difference in statistical testing.**

Several variants to describe significance levels can be found in the dental literature:

- The significance level was set at P<0.05
- Statistical significance was set at the 0.05 probability level
- Significance between groups was determined by using P<0.05
- The difference shows statistical significance at a probability of 0.05
- The difference was significant at the 0.05 probability level
- The groups show a difference at the 0.05 significance level

The results of a statistical analysis may be incorrect. This may be due to a

Type 1 error - or **alpha error**- when we obtain a significant result and thus reject the null hypothesis- when the null hypothesis is in fact true. (A false positive result). I.e. we report e.g. that there was a difference ($p < .05$), when this p-value in fact is due to pure chance

Type 2 error - or **beta error** - when we do not obtain a significant result when the null hypothesis is not true. (A false negative finding). I.e. we report e.g. that a difference was insignificant ($p > .05$), when this p-value in fact is due to pure chance- and usually would have been significant if the sample had been larger.

A usefull way of remembering what is type I and type error is to think of them as “optimism” and “pessimism” errors.
Type 1= alpha = optimism error, i.e. a tendency to believe there is a difference, although there really is none.
Type 2 = beta = pessimism error, i.e. a tendency to believe there is no difference, although there really is one.

Power calculation
Beta errors can be avoided by estimating the **power** of a study. A wide confidence interval in a study is an indication of low power. Power calculations depends on the measurement variability, the relevant difference of clinical significance and choice of significance level. There is a dramatic lack of presenting power calculations in the medical and dental literature.

Significance levels and confidence intervals
The significance levels and confidence intervals may together give more informative results than either alone. Especially in cases where P>0.05 and near borderline, the confidence interval for mean difference gives helpful information that may be clinically or scientifically meaningful. Thus, the statement "P>0.05 not significant" is as informative as specifying the actual P levels

obtained and showing the confidence limits for the mean difference.
The p -value will only be significant if the confidence interval does not include zero since both methods are based on similar aspects of the theoretical distribution of the test statistic.

Non-parametric methods

Skewed data are commonly analysed by non-parametric methods, and methods using ranks are especially suitable for data that are scores rather than measurements. Rank methods tend to be more suited to hypothesis testing than estimation. The methods are mostly based on comparing sums of ranks, and the central limit theorem applies also to these rank sums.

Exercise

An experiment is carried out to find the mean concentration of retained clorhexidine 7 hours after rinsing with a solution. 8 women participated in the experiment, and the following values were measured:

0.37 0.42 0.18 0.57 0.51 0.21 0.29 0.32

a) Find an estimate of the mean value for all women

b) What assumptions do you have to make to construct a confidence interval for the population mean? How would you try to check if these assumptions are valid in this situation?

c) Let us now assume that the necessary assumptions are valid. SD = .14 . Find the 95% confidence interval for the mean.

Choosing an appropriate method of analysis.

A methodological approach to choose the appropriate statistical method is to recognise the following characteristics:

1. Number of sample groups
   One group
   Two group
   Several groups

2. Independent or dependent groups
   Independent, size may differ
   Paired, size equal

3. Data type
   Continuous (mean and SD usually presented)
   Categorical

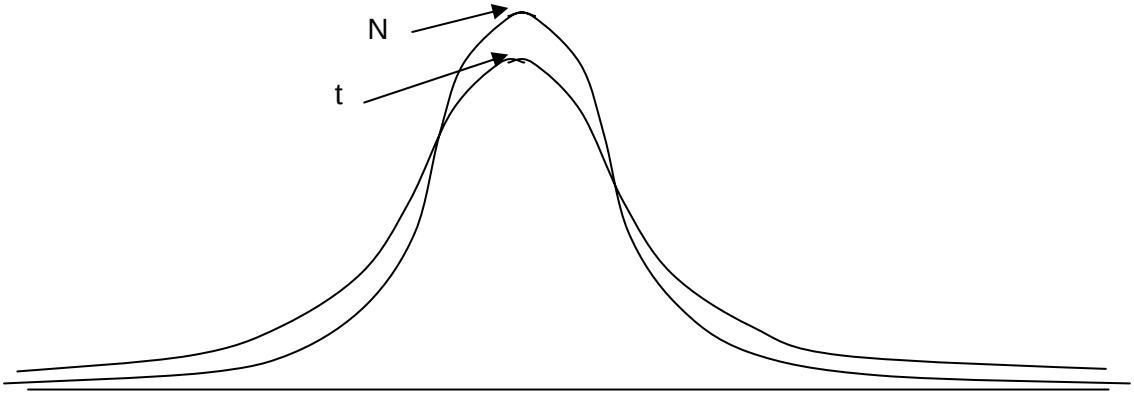
4. Data distribution
   Normal distribution equal variances: parametric tests
   Normal distribution, nonequal variance (tested with the F- or variance ratio test)
   Non Normal distribution, non parametric tests

It is not possible to give any general rule of how departures from Normal distribution affects the validity of the results. Very few samples of data show an exact Normal distribution - the principal assumption is not that it does, but rather that the sample comes from a population which does. When there are doubts about the validity, carry out a non-parametric test. This is likely to be the more reliable.

T-distribution

The mean of a sample from a Normal distribution population with an unknown variance has a distribution that is similar to, but not quite the same as a Normal distribution. This distribution is called the *t*-distribution. As the sample size increases, this distribution becomes closer to the Normal distribution. When the samples are larger than 100 the distribution of means can in practice be regarded as normally distributed. The distribution has one extra parameter in addition to mean and SD, which is the quantity **degrees of freedom**. The sample size minus 1 is termed degrees of freedom, i.e. df.

Different *t*-values can be found in table B4. If we wish to construct the 95% confidence interval we use the P .05 in table B4. For example, the t_{975}-value when we have sample size 11 = 10 df. is 2.228. For constructing a 99% CI we use P =0 .01 in table B4, i.e., t_{995}-value for sample size 22 = 21 df. is 2.831.



Common statistical tests appropriate to specific study design and level of measurement.

Samples	<- Categorical	-> <-	Continous ->
	Nominal	Ordinal Non-normal distribution	Normal distribution
One	chi-square test	One sample run test Sign test Wilcoxon signed rank (sum) test	One sample t test
two, paired	McNemar test	Wilcoxon matched pairs signed rank (sum) test	Paired t test
two, independent	Chi-square test	Mann-Whitney-Wilcoxon Median test	T test
k, paired	Cochran Q test	Friedman test	F test One-way ANOVA Two-way ANOVA
k, independent	Chi-square test	Kruskal-Wallis test	F test One-way ANOVA Two-way ANOVA

Comparing groups, continuous data, one observation group

Estimation

1. Normal distribution

Confidence interval for the mean

$$\bar{x} \pm t_{0.975} \times se(\bar{x})$$

2. Non-normal distribution

Confidence interval for the median

(Table B11 in textbook)

Hypothesis tests

One sample *t* - test

$$t = \frac{(\bar{x} - k)\sqrt{n}}{s}$$

k= hypothetical value, s= std.deviation
n = sample size, x = mean

Unsymmetric distribution : Sign test

$$z = \frac{|r - np| - \frac{1}{2}}{\sqrt{np(1-p)}}$$ r=observed count

p = .5

Symmetric distribution:
Wilcoxon signed rank (sum) test
Sum of positive or negative.
Sum of all ranks = n(n+1)/2
(Table B9 in textbook).

Exercise

1. A dentist believes that he in average use about 30 minutes to complete a composite restoration. One week the dentist decides to test if this feeling may not be correct. To make an investigation the time is recorded spent on the next 20 patients. The following results were recorded:

18 22 47 29 26 63 54 28 40 26 36 38 49 59 29 56 25 52 52 59

a. Why cannot an ordinary t-test be used in this case? (Make a histogram)
b. Use a sign test to see if the average time spent on each restoration is unequal to 30 minutes. Do you have to make any assumptions about the data distribution?
c. Carry out the same test using a Wilcoxon one-sample test. Do you have to make any assumptions about the data?

Comparing groups, continuous data, two observation groups

Estimation

Hypothesis tests

1. PAIRED

a. Normal distribution of the differences between means

CI for the difference between means

Paired *t* - test

$$t = \frac{\bar{d} - 0}{se(\bar{d})}$$

b. non-normal distribution

Sign test

$$z = \frac{|r - np| - \frac{1}{2}}{\sqrt{np(1-p)}}$$ r=observed count

p = .5

Wilcoxon (matched pairs) signed rank sum test

2. INDEPENDENT

a. Normal distribution, variances similar. Calculated with the **F-test**, the test statistic is calculated by computing the square of the ratio of the largest variance divided by the smallest.

CI for the difference between means

The pooled variance, s^2, is calculated as:

$$\frac{(n_1-1)s_1^{\,2}+(n_2-1)s_2^{\,2}}{n_1+n_2-2}$$

The standard error is calculated by:

$$se(\bar{x}_1-\bar{x}_2)=s\times\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

The 95% confidence interval is calculated by:

$$\bar{x}_1-\bar{x}_2\pm t_{0.975}\times se(\bar{x}_1-\bar{x}_2)$$

Two sample *t* - test

$$t=\frac{\bar{x}_1-\bar{x}_2}{se(\bar{x}_1-\bar{x}_2)}$$

b. Non- normal distribution

CI for the difference between medians
(not used often - complicated)

Mann Whitney U & Wilcoxon T tests
T= sum of ranks in the smaller group
(Table B10 in textbook).

$$U=n_1n_2+\frac{1}{2}n_1(n_1+1)-T$$

c. Normal distribution, unequal variances

Calculated using the F-test (variance ratio)

$$F=\left(\frac{sd_{stor}}{sd_{liten}}\right)^2$$

t - test for unequal variance
(Welch test = complicated)

Mann-Whitney U/Wilcoxon T

(Table B6 in textbook)

Exercise

Exercise 9.1-9.4, 9.6-9.8 in textbook

1. A test to see if there exists a placebo effect with laser treatment was carried out on 22 patients with TMD pain. The maximum jaw opening was measured before and after the patients received irradiation with ordinary red low-powered light. The differences in jaw opening were now measured.
The changes in jaw opening were in millimetres:
11, 2, 18, 12, -7, -1, 10, -2, 14, 1, 1, -2, 8, 22, 50, 6, 12, -8, 3, 41, 40, -2

(a) Formulate the problem above as a hypothesis problem and use the sign test to try to state a conclusion.

(b) Why could you not use the paired *t*-test for hypothesis testing?

III. Three or more independent groups

Estimation | Hypothesis tests

a. Normal distribution, variances must be similar

(Can be tested with e.g. Bartlett-Box F or Cochran's C tests)

CI for the means

One-way analysis of variance is based on the assumption that the variance within each group is an estimate of the population variance. Therefore, the sample variances are pooled to get an estimate of the population variance. The pooled estimate of variance is used to calculate a confidence interval for the difference between any pair of means. Hypothesis tests are based

One way analysis of variance

on the null hypothesis that samples drawn have identical means and Std deviation. Comparisons take the form of the ratio of the variance between the groups and the variance within each group. The F-distribution is assumed. The residual variance is the sum of differences between the observed and fitted values calculated by the statistic.

b. Normal distribution, unequal variances

Kruskal-Wallis
R = average of all ranks

$$H = \frac{12 \sum n_i (\overline{R}_i - \overline{R})^2}{N(N+1)}$$

c. Non-normal distribution

Kruskal-Wallis

A significant F statistic indicates only that the population means are probably unequal. It does not pinpoint where the differences are. A variety of special techniques termed multiple comparisons tests are available to determine which population means are different from each other. Correction methods are needed to avoid false positive (Type I) errors.

E.g.:	Bonferroni,	(Student) Newman-Keul,	Duncan
	Scheffé,	Tukey	LSD

Exercise

Exercise 9.5 in textbook

Proportions are a way of expressing counts or frequencies when there are only two possible outcomes, such as the presence or absence of a symptom.

I. One proportion

Confidence intervals

$$se(p) = \sqrt{\frac{p(1-p)}{n}}$$

95% CI = p ± 1.96 * se(p)

Hypothesis test (Note: SE is different!)

$$Z = \frac{p - p_{exp}}{se(p)} \qquad se(p) = \sqrt{\frac{p_{exp}(1-p_{exp})}{n}}$$

Since the variable can only have integer values, a continuity correction should be made when samples are small.

$$Z = \frac{\left|p - p_{exp}\right| - \frac{1}{2n}}{se(p)}$$

II. Proportions, two independent groups

Confidence intervals

$$se(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$CI = (p_1 - p_2) \pm 1.96 \times se(p_1 - p_2)$$

Hypothesis test

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2}$$: estimation of true proportion

$$se(p_1 - p_2) = \sqrt{\frac{\hat{p}\ (1-\hat{p}\ )}{n_1} + \frac{\hat{p}(1-\hat{p}\ )}{n_2}}$$

$$Z = \frac{p_1 - p_2}{se(p_1 - p_2)}$$

Since the variables can only take integer values, a continuity correction should be made when samples are small.

$$Z = \frac{\left|p_1 - p_2\right| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{se(p_1 - p_2)}$$

III. Two paired proportions

Confidence intervals

Observation		
Group 1	Group 2	n (pairs)
yes	yes	a
yes	no	b
no	yes	c
no	no	d

$$p_1 - p_2 = \frac{a+b}{n} - \frac{a+c}{n} = \frac{b-c}{n}$$

$$se(p_1 - p_2) = \frac{1}{n}\sqrt{b+c-\frac{(b-c)^2}{n}}$$

$$CI = (p_1 - p_2) \pm 1.96 \times se(p_1 - p_2)$$

Hypothesis test

$$se(p_1 - p_2) = \frac{1}{n}\sqrt{\frac{b+c}{2} - \frac{b+c}{2}} = \frac{1}{n}\sqrt{b+c}$$

$$Z = \frac{p_1 - p_2}{se(p_1 - p_2)} = \frac{b-c}{\sqrt{b+c}}$$

With continuity correction:

$$Z = \frac{|p_1 - p_2| - \frac{1}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{se(p_1 - p_2)} = \frac{|b-c|-1}{\sqrt{b+c}}$$

Exercise

1. In the development of a new toothpaste, “Superfresh anti-all with fluoride, baking soda and powertaste”, the producer wish to compare the effect of their toothpaste to an ordinary one. A large experiment is carried out, where 982 individuals use Superfresh..etc. and 715 individuals use ordinary toothpaste for 6 months. After 6 months, the number of individuals without new demineralisation zones on their teeth is 951 in the Superfresh..etc-group , and 650 in the ordinary group.

Is the Superfresh more effective in reducing demineralization than the ordinary toothpaste?
Calculate p1, p2, p1-p2 and the confidence interval of p1-p2, as well as carry out a hypothesis test of no difference.

Comparing groups, categorical data, contingency tables

Frequency tables are also called **contingency tables**. Analysing frequency tables are largely based on hypothesis testing. The null hypothesis is that the variables are unrelated in the relevant population. This is measured by comparing the observed frequencies with what we expect if the null-hypothesis was true. The expected values are given by the r x c totals. Each r x c category is a so-called cell.

a	b	a+b
c	d	c+d
a+c	b+d	n

Expected $a = \frac{a+b}{n} x \frac{a+c}{n} xn$

The test statistic is the sum of the quantities square of (O-E)/E for all cells, were O and E denote the observed and expected frequencies. The formulae is written:

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

X is looked up in table B5. Degrees of freedom is (c-1) x (r-1)

Other types of frequency tables

The statistical method of choice varies according to:
1. The number of categories
2. Whether the categories are ordered or not (nominal versus ordinal data)
3. The number of independent groups of subjects
4. The nature of the question being asked.

Number of categories			
variable 1	variable 2		test
2	2	independent	Proportions, Fischer's exact, Chi-square
2	2	paired	McNemar´s test
2	k unordered		Chi-square
2	k ordered		Chi-square for trend, Mann-Whitney
k not ordered	k not ordered		Chi-square
k ordered	k not ordered		Kruskal-Wallis
k ordered	k ordered	paired	Paired Wilcoxon,

Relation between two continuous variables

Analyses to study the relation between two variables in a sample may be carried out to:
1. Assess whether two variables are associated, e.g., **correlation analysis**
2. Enable the value of one variable to be predicted from any known value of the other, e.g., **linear or logistic regression analysis**
3. Assess the amount of agreement between the values of the two variables, e.g., duplicate measurements

Correlation

A possible association between two continuous variables may be described using the (Pearson) correlation coefficient, r. The r measure the scatter of the points around an underlying linear trend: the greater the spread of the points the lower the correlation. r can take any value between -1 and +1. A correlation of 0 indicates no linear association. Both confidence intervals and hypothesis tests of no association can be calculated.

At least one of the variables should be normally distributed. Both variables must be random and all observations must be independent.

Whenever correlation coefficients are calculated the data should always be plotted to see if there are any non-linear trends. A correlation may then be shown using a rank correlation coefficient. When the data deviate from an elliptical shape, or when a non-linear association is noted, a non-parametric rank correlation should be used instead of a linear correlation analysis, i.e., Spearman and Kendall.

In order to make valid **confidence intervals** for r, both variables should have a Normal distribution. Such data will display a rough elliptical pattern in a plot. In practice, therefore, it is preferable for both variables to have approximately normal distribution for any analysis of Pearson's r. The confidence intervals tend to be wide.

Apart from deviation from the distributional assumptions and adherence to observation independence, correlation statistics can be misused in several ways:
1. Two variables that correlate with time will always also be correlated.
2. Limiting the sample before performing a correlation computation is forbidden

3. Mixing of subgroups in the sample may confound the results
4. Assessing agreement between methods may be biased
5. Correlating changes over time to initial values is incorrect
6. Relating constituents with total amounts

- r is some times presented as r^2 to avoid unjustified conclusions about linear correlation. 100 * r^2 is the percentage of the variation of the data that is "explained" by the association between the two variables.
- Linear correlation **does not infer a cause-effect relationship.**
- Correlation is an often-overused analysis, especially when a large number of variables have been recorded. A correlation matrix of 10 variables yields 45 r coefficients alone. Recall that one in 20 will be significant when at the 5% level just by chance. Also the sample size will influence the magnitude of the correlation that is significant at the 5% level.
- Correlation analyses are often used when regression analyses should be preferred.

Regression

Regression analyses are used to **describe** the relation between two variables, and thus to predict the **dependent** (or response) variable from one or more **independent (or predictor)** variables**.**

One type of **regression line** may be constructed using the **least square** regression. This least square method produces the line that minimises the sum of the squares of the vertical distances, called residuals from this line**.** The values of included in the line are described as the **fitted** values. The **residual variance** is the sum of squares divided by the number of observations minus two.

The general equation of a regression of Y on X is:
Y = a + bX b is the **slope**,
a is the **intercept**, i.e. the fitted value of Y where the line crosses the axis.

1. Y should have a normal distribution for each value X.
2. the variability of Y should be the same for each value X
3. the relation between the two variables should be linear

- Unlike for correlation, the X values do not have to be normally distributed.
- For all regression lines a **confidence interval** as well as a much wider **predictor interval** of the slope can be estimated. While the former is a measure of the probability of including the true value within the interval, the latter is the limits of predicting the Y-values for 95% of future individuals correctly.
- A measure of the goodness of fit of the model is the proportion of the sum of squares explained by the regression as a percentage of the total sum of squares. The statistic R^2 represents the proportion of variation explained by the model.
- **Analysis of covariance** is an extension of regression, where the regression lines in two groups are compared and confidence intervals of differences or significance tests are carried out.
- **Non-linear** relationships may also exist. One not uncommon model is the **polynomial** regression.
- Correlation is a much over-used technique, with a significant correlation coefficient often wrongly interpreted as important and, even worse, as necessarily indicating a causal relationship. **Correlation tests should be used mainly to generate hypotheses rather than testing them.** Correlation reduces a set of data to a single number that bears no direct relation to the actual data.
- Regression is a much more useful method, with results that are clearly related to the measurements obtained. The strength of the relation is explicit, and uncertainty can be seen clearly from confidence intervals or prediction intervals.
- The predictive power of the function or the model is usually described by the R^2. The higher the value, the stronger the capacity to predict future Y- values.

Multivariate analyses, Survival statistics

ANOVA Analysis of variance

- Two-way ANOVA, i.e., the possible effect of an association between two independent variables on the response variable is assessed.
- Multiple ANOVA, i.e., ., the possible effect of an association between multiple independent variables on the response variable is assessed
- Multiple Classification analysis (MCA) has been used to adjust for possible relationships between the independent variables
- When there are more than two dependent/response variables a specific type of ANOVA is used: **MANOVA**, Multivariate analysis of variance

Multiple regression
When the SD is constant, and we have a normal variation of Y, regression models are linear
When the SD increases with Y, or is not normal we use log-linear models (multiplicative models)

Logistic regression
When the dependent (response) variable is dichotomous, logistic regression is used. Used frequently in epidemiology.

Factor analysis
The technique is used for examining a possible correlation structure among many variables

Principal component analysis
Technique for examining possible correlation structures among many variables

(CART) Classification and Regression Tree analysis
Classification technique, a function consisting of independent variables to best describe the different values of the response variable. CART functions are first determined, then validated next.

Cluster analysis
A method used for grouping units in samples based on specific variable combinations

(Linear) discriminant analysis
Discriminant functions include classification variables that minimise the within-group variability and maximise the between-group variability of a second group. The ratio between the between-group sum of squares and the within-group sum of squares are described by the so-called eigenvalues of the discriminant functions.

Common problems, examiner agreement

Frequently used indices for inter-examiner agreement are the percent agreement and the Pearson's correlation coefficient. These indices may be misleading. The kappa statistic is a measure of the proportion of agreement beyond chance that is actually achieved.

a	b	a+b
c	d	c+d
a+c	b+d	n

Kappa is calculated from the observed and expected frequencies on the diagonal of a square table of frequencies. If there are n observations in g categories, then the observed proportion agreement is Po.

$$p_o = \sum_{i=1}^{g} \frac{f_{ij}}{n} \qquad p_e = \sum_{i=1}^{g} \frac{r_i c_i}{n^2} \qquad \mathrm{K} = \frac{p_o - p_e}{1 - p_e} \qquad se(\mathrm{K}) = \sqrt{\frac{p_o - (1 - p_o)}{n(1 - p_e)^2}}$$

Cohen[4] described kappa (*k*) as a coefficient of agreement for nominal scales. It is a measure to help determine the extent to which judgements (categorisations) are reproducible[5] i.e. reliable.

The assessors would independently categorise a sample of responses (units) and determine the degree, significance and stability of their agreement. The following conditions must apply:

- The units must be independent
- The categories of the nominal scale are independent, mutually exclusive and exhaustive
- The judges operate independently

Other non-parametric tests (e.g. chi-squared test, correlation coefficient) are measures of association *not* agreement. It is possible to obtain a highly significant chi-squared result from analysis of such a contingency table, but this result may *not* be significant in the direction of agreement[6]. Cohen's coefficient (kappa) provides a measure of the degree of agreement in nominal scales and provides a means for hypothesis testing and deriving confidence intervals for the *k* coefficient. Kappa is the proportion of agreement *after* chance is removed from consideration. Kappa can take values from +1 (perfect agreement), though 0 (chance agreement) to –1 (perfect disagreement). Negative values can arise when agreement is less than chance, but as kappa is calculated as a measure of agreement, negative values are not very useful. Where the sample size is large (>100), the sampling distribution for kappa is (approximately) normal, so it is possible to calculate a standard error, derive confidence intervals, calculate probabilities and hence test hypotheses using kappa.

$$k = \frac{p_o - p_e}{1 - p_e}$$

p_o = proportion of units where judges agree
p_e = proportion of units where agreement is expected by chance
n = sample size

An approximate formula for the standard error of *k* is shown below. The inaccuracy falls as sample size increases.

$$SE\ k = \sqrt{\frac{p_o(1-p_o)}{n(1-p_e)^2}}$$

The 95% confidence limits are calculated as: *k* +- 1.96 * SE *k*

The marginal cells in the contingency table largely determine the maximum value for k. It can be a useful additional measure in reliability studies serving to indicate the "fuzziness" of category boundaries. Kappa has been developed further[6] to include measures of conditional agreement and agreement between more than two judges.

Limitations of kappa

- Kappa summarises agreement, but misses patterns of agreement. It can therefore be useful to estimate *k* for each category in turn.
- The value of *k* depends on the expected proportion of agreement, which depends on the marginal proportions for each rater. These will depend in turn on the true prevalence (e.g. diagnosis) in the subjects being studied, so that a different case-mix would yield a different value for *k* even with the same pair of raters. This means that variations in *k* between studies can be hard to interpret.

Interpreting kappa[7,8]

Range	
< 0,20	poor
0,20-0,40	fair
0,41-0.60	moderate agreement
0,61 – 0,80	good agreement
0,81- 1	excellent agreement

Statistical options for more than two assessors

A version of kappa (weighted $k = k_w$) can also be used to measure agreement in ordinal data. The measure of choice for examining the strength of agreement in quantitative data is the intraclass correlation coefficient (r_1). Both *k* and r_1 can be modified to measure agreement between more than two observers. Where several observers are involved, *k* is effectively an average of the kappas for each pair of raters with each category examined separately[6]. This calculation of kappa (k_m) can estimate either the joint agreement of *m* observers categorising *n* items, or the agreement among *n m*-tuplets of observer types (e.g. mother, father, child) triplets, categorising a single item. This will provide more detailed analysis rather than just using kappa as a simple summary statistic. Kappa can be calculated in common statistical packages (e.g. *Stata*[9] and *SPSS*[10]) for the following :

- More than two raters, two ratings. The statistic is calculated as **kappa pos neg**, where **pos** records the number of raters assigning "positive", **neg** the number of raters assigning negative (*Stata*).

- More than two raters, more than two ratings, fixed number of raters. The statistic is calculated as **kappa cat1 cat2 cat3,** where **cat1** records the number of raters assigning category 1 etc. (*Stata*).
- In *SPSS* the calculation is performed on a variable list containing the ratings for each rater, with ratings coded as an integer list.

The statistic *G* can be used[6] to compare the observed agreements of each of the *m* respondents with a standard assignment (e.g. a teacher or clinician may wish to compare the assignments students make compared to his "correct" assignments).

Multiple raters, interval, ordinal and nominal measures can be grouped together and considered to be just special cases of multiresponse permutation procedures (MRPP), these include Spearman's rho, Pearson's R, Cohen's kappa, Cochran's Q, Kendall's coefficient of concordance and Spearman's footrule. For detailed analysis with multiple categories and multiple raters, the key step is to break the analysis down by pairs of raters (2x2 tables), otherwise seeing and showing the detail is very difficult. There are about two dozen distance measures listed in the *SPSS* Proximities covering various possibilities for simple 2x2 comparisons.

References

- .Cohen J. A coefficient for agreement for nominal scales. Educational and Psychological Measurement 1960;20 (1);37-46.
- Theodossi A. *et al*. Observer variation in assessment of liver biopsies including analysis by kappa statistics. Gastroenterology 1980;79 (232);232-41.
- Light RJ. Measures of response agreement for qualitative data: some generalisations and alternatives. Psychological Bulletin 1971;76 (5);365-77.
- Dixon RA, Munro JF, Silcocks PB. The Evidence Based Medicine Worktextbook – Critical appraisal for clinical problem solving. Butterworth-Heinemann 1997. ISBN 0750625902.
- Bandolier. http://www.jr2.ox.ac.uk/bandolier/band43/b43-2.html
- Stata. http://www.stata.com/
- SPSS http://www.spss.com/tech/macros/spss/Nskappa.html

Exercises

You are participating in a group that is planning a scientific project. During the discussion a number of statistical terms are used that one in the group do not understand. Since you recently have attended a statistical course you are asked to describe in your own words what the meaning of the highlighted terms in the following phases are:

*1. "..the study must have a **high external validity**"*
*2. " .. a **non-parametric test** should be preferred"*
*3. "I have seen that **the chi-square test** has been used before…"*
*4a,b. "how many **degrees of freedom** do you have in a **t-test** when n=28?"*
*5a,b. "I think **standard error** should be used instead of **standard deviation**…"*
*6. ".. so that the **p value** become statistically significant…"*

The same group now wishes to proceed to planning the study. Since you already have established yourself as the statistics guru you are left to decide specific details of the project:

1. Describe how you would proceed to make a representative population sample
2. You decide to compare two patient groups using measurement criteria based on continuous scales. Which hypothesis test is appropriate? When during the study do you know if this appropriate?

You are participating in a group that is planning a scientific project. During the discussion a number of statistical terms are used that one in the group do not understand. Since you recently have attended a statistical course you are asked to describe in your own words what the meaning of the highlighted terms in the following phases are:

*1. "… I think a **correlation test** is necessary"*
*2. "…why can't we use a **regression test** instead?"*
*3. "…we have a number of **independent variables**"*
*4 "…but what about the **inter-examiner agreement**?"*
*5 "… I think this researcher is **bias**ed…"*
*6. " the data distribution is **skewed** to the right"..*

The same group now wishes to proceed to planning the study. Since you already have established yourself as the statistics guru you are left to decide specific details of the project:

1. *How could you proceed to make the study double blind?*
2. *How would you calculate the confidence interval of the difference between two groups if the outcome measure was on a continuous scale?*
3. *Can you use the non-parametric test when the data distribution is highly skewed?*

You are an avid reader of articles in the Journal of Elusive Dentistry. Do you have any comments regarding the content and use of statistics in the following paraphrases? Please include as many comments as possible.

Article 1:
"*the mean values for the four samples were compared by using t-test. Significant differences were found between groups 1 and 2 , 1 and 4 and 3 and 4 (p < .05)…"*

Article 2:
"*the index-scores for all participants (n=45) varied between 1-10. Most values were in the range 1-2 (n=15). The index-scores were reordered so that 3 subgroups contained the same sizes. Thus the range 1-2 = 1, range 3-5 = 2, and range 6-10 = 3. The mean and std. deviations of the reordered scales within each subgroup (n = 3x15) were compared using ANOVA".*

Article 3:
"*a correlation matrix using the 28 variables showed that 4 of the variables correlated significantly, i.e. age and head ache, age and sex appeal, income and age, and age and years in practice. Thus, there seems to be a relationship between these variables."*

Article 4:
"*since the samples were fairly similar a paired t-test was used instead of the t-test for individual samples"*

Article 5:
"…*the mean values for the four samples were compared by using the Mann-Whitney U-test. Significant differences were found between groups 1 and 2 , 1 and 4 and 3 and 4 (p < .05)…"*

Article 6:
"..*the randomisation of the participants was made by selecting one patient from the hospital patient pool and one from the private practice with the same age"*

Article 7:
" … the inter-examiner agreement between the two examiners was established using Pearson's r. The correlation yielded r= .60, (p = .03) which was considered acceptable..."

Article 8:
"… the mean values for group A and B was .55 and .61 respectively. Thus, the values were fairly comparable…"

Article 9:
"… the difference between the control group and test group A was P = .03, and between the control group and test group B P= .001. Thus, a stronger effect was apparent in group B.."

Article 10:
"…although the participants in this study were obtained from the dental school patient pool we think that it is probable that the findings can be applied to the general patient population