CORE CURRICULUM 2002
RESEARCH METHODOLOGY

Coordinators: Asbjørn Jokstad and Dorthe Holst

Videreutdanningen, Geitmyrsveien 69, 4th floor

Tuesdag 27.8. 15.00 – 16.30 Research methodology. Introduction

Monday 2.9 15.00 - 16.30 Research protocol: From ideas and questions to answers

Monday 9.9. 15.00 – 16.30 Units of measurements, variables, values

Monday 16.9. 15.00 – 16.30 Reliability and validity

Monday 23.9. 15.00 – 16.30 Causality

Monday 30.9. 15.00 – 16.30 Study design

Monday 7.10. 15.00 – 16.30 Epidemiologic research

Monday 15.10. 15.00 – 16.30 Clinical and experimental research

Monday 21.10. 15.00 – 16.30 Study design, practical training

Monday 28.10. 15.00 – 16.30 Preliminary protocol presentations

Monday 4.11. 15.00 – 16.30 Preliminary protocol presentations

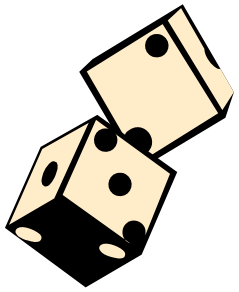
Monday 11.11. 15.00 – 16.30 Analytical approaches

Monday 18.11. 12.00 – 16.30 Project work

Monday 25.11 12.00 – 16.30 Exam. Final presentations

Monday 2.12. 12.00 – 16.30 Exam. Final presentations

Recommended literature:
L Gordis. Epidemiology. WB Saunders 1996
JH Abramson. Survey methods in community medicine. Churchill Livingstone 1990
Articles and handouts distributed during the course



Research methodology – some notes

Core Curriculum 2002
Dental Faculty
University of Oslo

Asbjørn Jokstad
University of Oslo

Why learn study design methodology?

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 errors are discovered at the moment the data are to be analysed.

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 individuals in study samples are not chosen at **random**.

Study reports sometimes lack a complete description of the prerequisites to carry out an appraisal of the study validity. Usually, this is due to poor reporting, but clearly inadequate study designs can occasionally be identified (an estimate is approx. 5% of all reports in medicine).

In general, errors are in six categories:

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

Lack of understanding study methodology may lead to unexpected, hilarious and sometimes even dramatic headlines in newspapers. However, there are other more ethical implications of misusing or misinterpreting research :

- **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.**

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**.

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 some special 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. 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 $\pm$ 1.96 x SE.
A 99% confidence interval is mean $\pm$ 2.58 x SE.
The interval between mean $\pm$2SE will be a 95.4% confidence interval and between mean $\pm$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.

Standard error of the difference between two sample means
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.

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.**

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 useful 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.

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.

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/two-way ANOVA
k, independent	Chi-square test	Kruskal-Wallis test	F test One/two-way ANOVA

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 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, Fisher'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,

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

Designing your research project & 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

Statistical issues 2

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.

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

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.

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.

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.

Cohen 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 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. 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.

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

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