Statistics and Interpretation Of Clinical Trials

Syllabus

This subject should be studied with the aid of the textbooks indicated below and the following points noted:

Clinical Trials

- Advantages and disadvantages.
- Retrospective and prospective studies.
- Controlled and uncontrolled trials.
- Historical and concurrent controls.
- Blind and double-blind studies.
- Phase I, II and III trials.
- Ethics (Helsinki declaration).

Planning a Trial

- Establishing objectives, short-term and long-term.
- Determining the appropriate criteria.
- Establishing grounds for inclusion and exclusion of patients.
- Deciding how many treatment schedules are to be compared.
- Determining the treatment schedules and any appropriate modifications.
- Determining the method of allocation of treatments: the allocation ratio and the method and timing of randomization.
- Determining what measures are to be taken, how they will be taken, who will take them, at what time(s) and where they will be recorded.
- Designing the appropriate forms for documentation.
- Determining the proposed duration of the trial, either in terms of a fixed closing date, or the entry of a pre-determined number of patients.
- Establishing conditions under which the trial may be terminated earlier than planned and procedures for detecting these conditions.
- Establishing responsibilities.
- Re-assessing the proposed trial in terms of ethics, appropriateness to the short and long term objectives, feasibility and the availability of resources.
- Writing the protocol.
- Running a pilot study.

Interpretation of Results

- Type 1 and Type 2 errors.
- Estimation and tests of significance.
- The interpretation of a P-value.
- Handling post-randomization exclusions, withdrawals, losses and treatment deviations.
- The life table.
- The logrank test.
- Identifying and assessing prognostic factors, hence refining treatment comparisons.

1 Peto R., Pike MC, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG. *Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient -*
   I. *Introduction and Design*. Br. J. Cancer 1976, 34, 585-612.
   II. *Analysis and Examples*. Br. J. Cancer 1977, 35, 1-39.

2 Mainland D. The Clinical Trial: Some Difficulties and Suggestions. *J. Chron. Dis*. 11, 1960, 484-496.

3 *Clinical Trials : Design and Analysis*. Seminars in Oncology 1981, 8, No. 4 (December).

Additional References

Mainland D. *Elementary Medical Statistics*, 2nd Edition. Saunders, 1963.

You may wish to refer to notes provided by Dr Gail Ryan on methods of critical appraisal, which are reproduced here:

CRITICAL APPRAISAL FOR RADIATION ONCOLOGY TRAINEES

One of the more daunting tasks facing any doctor is "keeping up with the literature". There are more journals around than one can hope to read in the limited time the average Radiation Oncologist has available. Worse - all the articles may seem to be written in a language unknown to you. But if you feel vaguely nauseated at the sight of the Red Journal or your eyes glaze over as you flick through the contents of JCO, take heart. A P-value need not bring on an acute panic reaction! With some basic epidemiological and statistical skills, and a systematic approach, critical appraisal becomes very simple and you may be surprised how easy it is to distinguish the few pearls from the vast array of not very informative literature.

This guide is divided into four sections, each one building on the concepts of preceding sections. For this reason, it is best to avoid the temptation to read them out of order, no matter how unexciting a section may sound to you!

SECTION A: TERMINOLOGY

Research Question

This is the essential question the study has been set up to answer. Most studies are concerned with determining one of the following:

i) the magnitude of a health problem,
ii) the natural history (clinical course and prognosis) of a disease,
iii) the etiology or causation of a disease (the relationship between a set of factors and the outcome of interest), or
iv) the efficacy of an intervention.

In oncology journals, most of the published articles are of the type iv), e.g., is new treatment "A" better than the standard treatment "B"?

Hypothesis

This is a proposal about the relationship between two or more variables, e.g., new treatment A is better than standard treatment B. For the hypothesis to be testable, all components need to be clearly defined in complete form (what exactly is the population to which the treatment has been applied, what is the measurement of "better", etc.). Hypotheses may not always be stated in papers but are implicit. Hypotheses can be formulated in the null form e.g. Not all patients with X respond to treatment Y, so as to allow them to be refuted. Not all studies test hypotheses - some descriptive studies can be hypothesis-generating.

Study Factor

This is the exposure or variable of interest that is hypothesized as related to the outcome of interest. It is also known as the independent variable.

Example: In a study of asbestos exposure and the development of mesothelioma, the study factor is asbestos exposure.

The study factor must be quantifiable but may be assessed by a variety of means.

Outcome Factor

This is the outcome of interest (event or occurrence) that is hypothesized as occurring as a result of the operation of the study factor. It may also be called the dependent variable. In the example above, the outcome factor is the development of mesothelioma. The outcome factor may also be measured by a variety of means.

Sampling

In setting up a study it is frequently impossible or impractical to include all members of the reference population (i.e., the group of people to which the results of a study, if valid, may be applied). The researcher will use a subset of the reference population, known as the source population. The researcher then focuses on a smaller group which will represent the source population. From this sample frame, a study sample is collected.

Sampling is the process for selecting a study sample from the sample frame under such circumstances. The usually preferred method is random sampling whereby inclusion in the study sample is decided by chance.

The hierarchy of sampling is illustrated in the following example.
I wish to study the use of sunscreens by teenage girls.

Reference population - the abstract concept of people to whom the results would apply in this case, Australian girls aged 13 - 19.

Source population - the broad group of people from whom the subjects will be drawn, e.g., teenage girls attending school in the Melbourne metropolitan region.

Sampling frame - the list of potential subjects from which a sample will be drawn, e.g., class lists of girls' schools in the Melbourne metropolitan region.

Sample - the subjects who are selected to take part in the study, e.g., a random selection of 5 classes in each of the 10 schools which have been randomly selected from the 50 in the Melbourne metropolitan region.

Sample subjects - these provide the data, e.g., of the chosen study sample, the 95% who attended school on the day on which the study was carried out.

Bias

Bias is any effect at any stage of investigation or inference tending to produce results that depart systematically (i.e. consistently in a particular direction, not due to chance alone) from the true values. Many varieties of bias occur, e.g., selection bias, confounding bias, measurement bias.

Confounding

This is a form of bias. A confounding variable is a factor that distorts the apparent magnitude of the effect of a study factor on the outcome factor. The confounder is a determinant of the outcome of interest and is unequally distributed amongst the subjects who are exposed or not exposed to the study factor. When an observed relationship is described as confounded then not only does its magnitude come into question but its very existence is subject to doubt.

A common situation is that in which a disease has a number of causes, e.g., exposure to asbestos is a cause of lung cancer but so too is cigarette smoking. If the distribution of cigarette smokers is different in two comparison groups, then cigarette smoking would confound a demonstrated relationship between asbestos exposure and lung cancer. In oncology studies, so-called prognostic factors are frequently confounders.

Modification

This occurs when the relationship between the study factor and the outcome factor is influenced by varying levels of a third variable, e.g., the development of second malignancy following radiotherapy may be modified by smoking status.Risk

Risk is the probability that an event will occur, e.g., that a patient with ulcerative colitis will develop bowel cancer.

Risk Factor

Risk factor is an attribute or exposure that increases the probability of the occurrence of disease or a specified outcome, e.g., positive family history is a risk factor for development of breast cancer.

Relative Risk

This is one measure of the strength of the relationship between exposure (the study factor) and the outcome factor. It is the ratio of the risk of disease or death amongst those exposed to the study factor to the risk of disease or death amongst those not exposed to the study factor. To calculate relative risk it is necessary to use incidence figures for disease or death. Some study designs preclude this possibility and an approximation to relative risk, called the odds ratio, is calculated instead.

Odds Ratio

The risk of the person with the outcome of interest having been exposed to the study factor is known as the odds ratio. It is a good approximation of relative risk, provided that the incidence of the disease in the community is relatively low (< 1:100)

Attributable Risk

Relative risk does not give any idea of the size of a health problem in absolute terms. To do this, "attributable risk" is defined. This is the difference between the rate in the exposed population and the rate in the non-exposed population and it measures the amount of disease that might reasonably be attributed to the exposure in question.

Incidence

The incidence of a disease is the number of new cases of that disease occurring in a defined population within a specified period of time. The incidence rate has the number of new cases of a disease in a defined period of time as the numerator and the number of persons in the stated population in which the disease occurred as the denominator.

Prevalence

Prevalence is the total number of all individuals who have the disease or factor of interest at a particular time or during a particular period (the numerator) divided by the population at risk of having the disease at that time (the denominator).

Study Validity

“The degree to which the inference drawn from a study, especially generalizations beyond the study sample, are warranted when account is taken of the study methods, the representativeness of the study sample and the nature of the population from which it is drawn. Two varieties of study validity are distinguished." (Last)

1. Internal Validity

   The groups being compared (treated vs. untreated, exposed vs. unexposed or cases vs. controls) have been selected and measurements made in such a way that the results can be considered a good approximation to truth.

2. External Validity

   The subjects in the study are selected and described in such a way that the results, given internal validity, can be applied or generalized outside of the study sample.

Threats to internal validity come from systematic error or bias.
Threats to external validity come from the sampling and selection of subjects for a study.
Internal validity always has precedence over external validity.

Measurement Issues

There are two issues of concern with all measurements:

1. Measurement validity - the degree to which a measurement measures what it purports to measure, e.g., an open biopsy has more validity than an aspiration cytology specimen.

2. Reliability - the degree of stability exhibited when the measurement is repeated under identical conditions. Most biological measurements show some variation when repeated.

SECTION B: RESEARCH METHODS

Much of epidemiology concerns methods used to obtain valid data to answer a research question or to refute a well-refined hypothesis. Given that there are restrictions on human research, for ethical and practical reasons, epidemiologists have developed methods and rules for obtaining evidence about a research question. There is a hierarchy of evidence ranging from a short description of what happened to a controlled experiment of the laboratory type. From an epidemiological point of view, the type of research study is very important since there are a number of advantages and disadvantages to various types and, more particularly, there are traps.

Research studies can be broadly grouped as experimental or non-experimental.

Types of experimental studies are:

i) Clinical trials - patients as subjects
ii) Field trials - with healthy subjects
iii) Community intervention trials - with the intervention assigned to groups of healthy subjects

Types of non-experimental studies are:

i) Longitudinal studies - subjects are selected with reference to their exposure status
ii) Case-control studies - subjects are selected in reference to their disease status
iii) Cross-sectional studies
iv) Ecologic studies - unit of observation is a group of people

A more detailed description of types of study that may be carried out in the oncological setting follows.

Randomized Controlled Trials

A randomized controlled trial is an investigation in which similar groups of individuals are allocated at random, by the investigator, to receive or not to receive a particular therapeutic or preventive intervention. Without random allocation, the results of an experimental study are at best suggestive.

The steps in conducting a valid clinical trial are:

i) Define the purpose of the trial - state the specific hypothesis
ii) Design the trial - a written protocol is necessary
iii) Conduct the trial - requires organization
iv) Analyse the data - descriptive studies, tests of hypotheses, etc
v) Draw conclusions - publish results.

Issues

1. Selection of the Study Population.

The two groups should be equal in all respects other than the introduction of a preventive or therapeutic measure to one group. This requires that clear eligibility criteria be specified. Note that if the selection criteria are too narrow the study population may not in fact be representative of the reference population to whom the investigators wish the results to be applicable.

2. Methods of Allocation.

A number of different randomization schemas can be used and a brief description should be included with any published study. To allow for a number of subsidiary factors known to influence outcome or response, it may be advantageous to match patients allocated to the different treatment groups. This can be achieved using a modified method of randomization - stratification. Matching beyond three parameters creates difficulty in obtaining a balanced study owing to the small numbers in some of the subcategories.

3. Sample Size.

Will be discussed in Section C

4. Standardization of the Intervention.

This is essential and can be a stumbling point in multicentre trials.

5. Assessment of Outcome.

Problems of standardization, loss to follow up, observer bias.

6. Ethical Considerations.

The study must have the potential to produce valuable scientific information (sample size calculations may be vital to this consideration). There must be predetermined rules for deciding if a trial needs to be stopped before it is due to end, e.g. a clear difference in favour of one treatment or the other is demonstrated by interim or sequential analysis.

Advantages of the Randomized Controlled Trial

1. With randomization, comparability of groups is very likely for both known and unknown confounders, since the groups are derived from the same source population and should only differ by chance.
2. Experiments provide the best chance of obtaining strong evidence of a cause and effect.
3. Allows standardization of eligibility criteria, the intervention and outcome assessments.
4. Allows use of statistical methods which have few inbuilt assumptions.

Disadvantages of the Randomized Controlled Trial

1. May be expensive in terms of time, money and people.
2. Many research questions are not suitable due to ethical considerations, likely co- operation or rarity of outcome.
3. New research may supersede the premise on which the trial is based, making it irrelevant.
4. To a greater or lesser extent, a randomized controlled trial tends to be an artificial situation.
   (a) Patients who volunteer for a trial may differ from those to whom the results would be applied.
   (b) Standardized interventions may be different to common practice - this is the difference between efficacy (does the treatment work under ideal conditions?) and effectiveness (does the treatment work in the "real world", where factors such as non-compliance, less selected patients, less selective clinicians, cost and impracticability may be operative?). A related issue concerns whether data should be analysed by "treatment to which randomized" or "treatment actually received" in the case of non-compliers. Analysis by "intention to treat" provides a more valid assessment of treatment effectiveness, but the investigator may be more interested in efficacy. It is preferable to decide whether efficacy or effectiveness is being studied prior to commencing a study.

Some Appropriate Statistical Tests (not exhaustive, more detail in Section C)

Comparison of proportions of responders, with confidence intervals.
Survival analysis.
Analysis of variance (more than two groups).
Multivariate analysis for prognostic factors.

Example

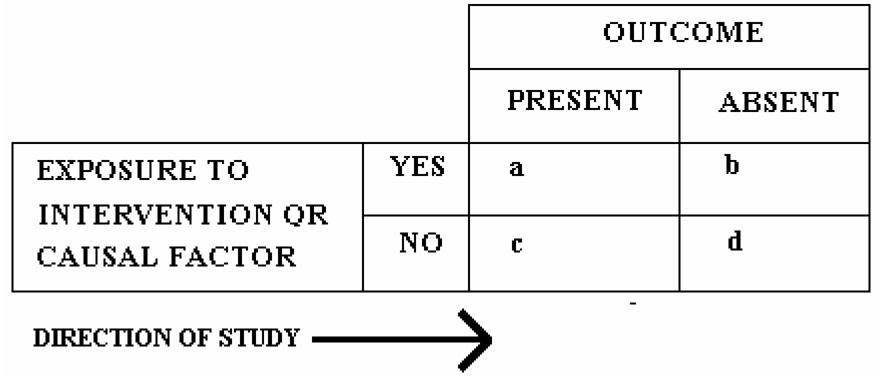
Radiotherapy alone vs radiotherapy + 5 fluorouracil in the palliative treatment of rectal cancer.

Longitudinal Studies

A longitudinal study is one in which people who are free of the disease of interest (outcome) but differ on a certain exposure (study factor) are followed and the incidence of disease measured. A study in which a group of patients with a disease is followed to ascertain prognosis is also called a longitudinal study. The non-exposure or comparison groups may come from within the initial group of subjects who were measured to exclude disease and define exposure status. Alternatively the control group may be a separate population altogether, e.g., the national population.

The start point for the study may be defined as some time in the past and the experience of the population followed up to the present - a historical or non-concurrent cohort or retrospective study. Alternatively, the start point of observation may be now and the population followed into the future for some period of time, with analysis being performed at intervals along the way - this is a prospective or concurrent cohort study.

Cohort studies are best suited to situations where the exposure is relatively rare and the outcome (disease) is relatively common.



		OUTCOME	
		PRESENT	ABSENT
EXPOSURE TO INTERVENTION QR CAUSAL FACTOR	YES	a	b
	NO	c	d

DIRECTION OF STUDY →

Issues

1. Definition of the Study Population.

The population is defined in terms of the study factor, the best population being one in which there is a range of exposure to the study factor. However, sometimes the exposure is a natural dichotomy. In a prospective study there is a need to define a study population in which an adequate response rate can be anticipated or ensured by various techniques and non-responders can be identified, e.g., by occupational group or by electoral roll. In a retrospective study the population is defined by pre-existing records and the investigator is limited by what information was collected at that time. In this situation it is difficult to be sure of the completeness, or at least unbiased nature, of the recorded information. It is unlikely that information about possible confounders was collected, particularly lifestyle factors.

2. Study Factor.

The study factor needs careful and precise definition to avoid bias. The way in which the research question is formulated can change its meaning and interpretation.

3. Outcome Measures.

These must also be precisely defined.

Advantages of Longitudinal Studies

1. There is a logical process of exposure to outcome, i.e., the exposure has definitely preceded the outcome.
2. Absolute risk (incidence) can be established.
3. The natural history of a disease may be demonstrated.
4. In prospective studies, exposure can be measured without bias because, at that time, the outcomes are not known.
5. Retrospective studies allow rapid and economical testing of a hypothesis.

Disadvantages of Longitudinal Studies

1. Prospective Studies

   (i) require large number of subjects to be studied over long periods; loss to follow up can be a serious source of bias in such a study.

   (ii) are expensive because of the resources needed to follow up people over a long time.

   (iii) do not produce results for a long period, by which time the problem may no longer have relevance.

   (iv) are usually restricted to exposures which are measured at the start of the study; increasing knowledge may reveal far more important exposures which were not known at the start of the study.

   (v) may actually result in participants automatically modifying their behaviour - the Hawthorne effect - this influence may not affect all of the potential respondents to the same extent, thus introducing further bias.

2. Retrospective Studies

   These are prone to bias because of the retrospective collection of data - the longer the time lag between the initiating event and the development of disease, the greater the chance of error and bias occurring, e.g., recall bias, survivor bias.

Some Appropriate Statistical Tests

i) Relative and attributable risk.
ii) Incremental risk for varying levels of exposure (Mantel-Haenszel techniques).
iii) Survival analysis.
iv) Two or more aetiological factors - multivariate analysis.
v) Chi-square tests for categorical data.

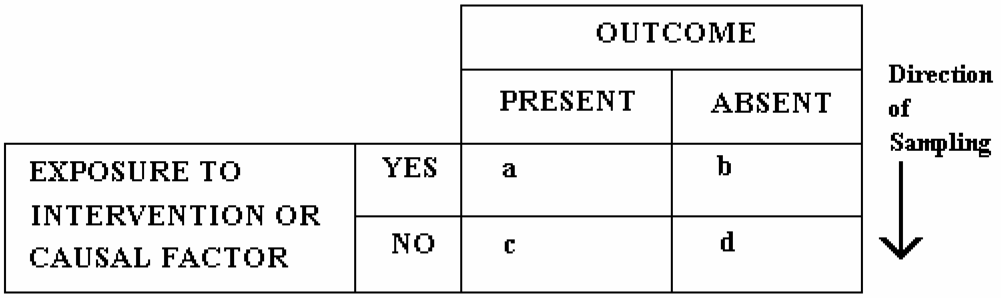
Examples

i) Studies of Hiroshima survivors ("one-off" cohort).
ii) Studies on cigarette smoking and the incidence of lung cancer.

Case Control Studies

A case-control study is an investigation in which the impact of either an exposure or an intervention is investigated after the outcome has occurred. It is a method of obtaining similar information to a longitudinal study, but much more efficiently and in a shorter period of time. The fundamental difference with a longitudinal study is that subjects are defined by disease status and not by exposure status.

Case-control methodology is particularly suitable for the study of rare diseases, especially if the exposure of interest is common.



		OUTCOME	
		PRESENT	ABSENT
EXPOSURE TO INTERVENTION OR CAUSAL FACTOR	YES	a	b
	NO	c	d

Direction of Sampling ↓

Case-control studies have been the subject of much criticism and it is true that a number have been poorly designed, implemented and analysed. However, "the poor reputation suffered by case-control studies stems more from their inept conduct than from an inherent weakness in the conceptual approach." (Rothman)

Issues

The potential for substantial bias is the major problem of case-control studies. The sources of bias are:

i) Selection of cases and controls.
ii) Confounding.
iii) Measurement.

i) Selection of Cases and Controls

The central condition for conducting valid case-control studies is that controls be selected independently of exposure status. If this is so, then the odds ratio provides an unbiased estimate of the incidence ratio.

An appropriate control is someone who would have been a case if he or she had developed the disease. It may be that cases and controls have different susceptibility to disease outside of the exposure of interest or be selected in a way that is not independent of exposure - "Berkson's Paradox". Hospital-based control series are especially prone to such biases.

In order to minimise selection bias, the source population may be a restricted group, e.g., single town or industrial community. This may limit external validity, but internal validity always has precedence.

An excellent method of obtaining cases and controls from a common source is a nested case - control study within a cohort study.

The division between selection and confounding bias is blurred in case-control studies.

ii) Confounding

Confounding is a mixing of the effect of the exposure under study and one or more other factors which influence the outcome. The potential confounder is predictive of the disease independent of the exposure (although it need not be causal e.g., age or sex), and is unequally distributed amongst cases and controls.

There are a number of methods to control, or adjust for, confounding:

a) Random selection of controls from a community-based sample has the potential for minimizing the effect of unknown confounders compared with hospital-based or neighborhood samples.

b) Restriction - limiting the study to a group of subjects (e.g., males aged 40-65 years) reduces potential confounding, in this case cause by sex and extremes of age.

c) Matching - the selection of a control series that is identical, or nearly so, to the case series with respect to one or more potentially confounding factors. Matching can be on a subject by subject basis (individual matching) or for groups of subjects (frequency matching). Frequency matching may be done in the form of stratification such as by age group (e.g. 5-year strata) containing equal numbers in each case and control groups. Over matching can be a problem - if the control group is too like the case group, some of the important determinants of the disease may be excluded from the analysis by the process of matching.

The use of matching is controversial, since it does not prevent confounding and must be accompanied by a matched analysis, but has the advantage of increasing efficiency. Some epidemiologists never use matching as they believe it introduces rather than prevents confounding in such studies.

iii) Measurement

Problems can arise where cases have been collected from medical or other records without strict adherence to a diagnostic regimen and where controls have had little or no attempt to exclude "disease" - ideally cases and controls should have had a similar extent of investigation. Being a case can also lead to more intense interrogation or better record keeping. Another source of information bias is where different sources of information about exposure have been used for cases and controls (relative of case vs. control him/herself).

Different recall may also occur in those with disease present compared with those without disease - recall bias.

Advantages of Case-Control Studies

Valuable for studying rare conditions.
Short duration.
Relatively inexpensive.
Relatively small number of subjects required.
Yield odds ratio (usually a good approximation of relative risk).

Disadvantages of Case-Control Studies

Limited to one outcome variable.
Do not establish sequence of events.
Potential bias from selection of cases and controls.
Potential bias in measuring exposure.
Potential survivor bias
Do not yield absolute risk estimates.

Some Appropriate Statistical Tests

Chi-square test
McNemar's test for matched samples
Logistic regression
Odds ratio

Example

Relationship of oestrogen usage and endometrial cancer.

Cross-Sectional Studies

A cross-sectional study is one in which all subjects in a population are investigated for outcome and/or exposure. It is the least complicated form of observational study and is usually a simple description of the prevalence of characteristics or disease with the objective of estimating the magnitude of a problem in a defined population. Occasionally, the study can be analytic, examining the association between an exposure and an outcome factor as they exist at one time. This is usually done by selecting the source population without regard to the study or outcome factors which are then measured simultaneously in each member of the population.

The important aspect of a cross-sectional study is that data from subjects is obtained only once. Cross-sectional studies may generate etiologic hypotheses but more complex studies (cohort or case-control) will need to be performed later to test such hypotheses.

Case series or reports often describe the relative prevalence of symptoms, signs or investigative results in a disease which may then be generalized to all cases of the disease or problem. Such studies are especially valuable for rare or newly described illnesses. However, case series and reports can only be used to generate hypotheses, not to test them.

Issues

The main issue in cross-sectional studies is the extent to which results can be generalized. Thus sample selection is critical. The choice of study population will be a balance of suitability, feasibility and availability. The sample needs to be representative of the larger population and have been taken in an unbiased manner using a proper sampling technique - a truly random sample of adequate size will allow the results of the survey to be considered and conclusions made about the population as a whole.

If a defined group is used as the study population, such as those working in a particular occupation, one must always consider how representative the study population is of the general population. The problem of "survivor bias" exists (those experiencing certain outcomes may have died or left the industry). To check for bias in a proposed study population, information concerning sickness absences, staff turnover and mortality amongst employees/ex-employees must be sought from reliable sources.

Response rates can also threaten the appropriateness of generalizing the findings of a cross-sectional study. The general topic of the enquiry has an effect on the response rate and it has been shown that a higher proportion of non-responders are elderly, involved in manual work and/or of limited education. A response rate of less than 70% may introduce sufficient bias so as to cast doubts on the validity of the study. It is essential to obtain some information about non-responders, e.g., age, sex, ethnicity.

Advantages of Cross-Sectional Studies

May study several outcomes
Control over selection of subjects
Control over measurements
Relatively short duration
Good first step for a cohort study
Yield prevalence

Disadvantages of Cross-Sectional Studies

Do not establish sequence of events
Potential bias in measuring exposure
Potential survivor bias
Not feasible for rare conditions
Do not yield incidence or true relative risk

Some Appropriate Statistical Tests

Chi-square tests
Analysis of variance
Correlation coefficients
Means of different groups with confidence intervals

Example

The prevalence of a positive family history in those subjects in the population who have been treated for breast cancer.

Ecological Studies

An ecological study is one in which data are obtained from different groups or populations and a correlation coefficient is produced between an outcome (such as breast cancer) and an exposure (such as dietary fat). The problem is that the data for exposure and outcome are obtained from populations or groups and not from individuals. Such studies are potentially flawed (the so-called ecological fallacy) and data thus obtained should be interpreted very carefully.

Advantages of Ecological Studies

They use information already available
They are quick and inexpensive
They can be used as a first step in exploring a relationship between an exposure and a disease

Disadvantages of Ecological Studies

A link to individuals cannot be made
Potential confounding factors cannot be controlled
A lack of correlation may not mean a lack of association
Exposure data are averages and do not represent individual levels - dose/response relationships may be masked

Some Appropriate Statistical Tests

Regression
Correlation

Example

Relationship between hepatitis B surface antigen prevalence and liver cancer.

SECTION C: STATISTICS

Statistics have been called "the arithmetic of human welfare" (Lancelot Hogben). "They add precision to the assessment of significance of observations, the results of experiments and the meaning of relationships. They also aim to eliminate emotional pre-judgements, banish wishful thinking and allow results to be assessed with accuracy and checked by others." (Professor Sir David Smithers)

Statistics are estimates of population parameters based on samples. There are two branches of statistics:

i) Descriptive statistics, which covers the organizing and summarizing of data

ii) Inferential statistics, in which conclusions are drawn (or questions answered, or information obtained) about populations based on samples. When a sample of the population is measured and the population has been sampled properly (i.e., the sample is a random sample and therefore most probably is representative of the population as a whole) information from the sample data can be used to answer questions about the population.

The concept of statistical inference works because :

i) Certain assumptions can be made about the nature of a particular sample and the population and how they are related.
ii) Descriptive information can be calculated for a sample .

Some important estimates are:

- Mean = average of observations $\mu = {}^{1}/_{n}\, \Sigma x_i$
- Variance = estimate of variability in population $\sigma 2 = 1 / n - 1 \ (xi - \mu)2$
- Standard deviation $\sigma = \surd$ Variance
- Standard error = measure of accuracy of calculation of mean = $\sigma / \ n$

The Normal Distribution and Probability

When many independent random effects are summed, a frequency distribution is described. This distribution can be used to estimate the probability of obtaining a certain value. Much biological data happens to follow a distribution known as the Gaussian or normal distribution - the curve is bell-shaped and symmetric so the mean and the median fall at the centre of symmetry; the standard deviation describes the spread of the curve. It has been found that if a sample is large (>100) and if multiple such samples are taken from a population, the distribution of the means of the samples follows a normal distribution even if the distribution of the individual observations is not normal - this is the Central Limit Theorem.
In practice this enables the normal distribution to be used for probability calculations involving means of large samples. Probabilities are represented as areas under the curve; about 68% of probability falls within one standard deviation of the mean; about 95% within two standard deviations and 99.7% within three standard deviations.

For small samples (<20) taken from a normally distributed population, the distribution of the means follows a t-distribution, which has longer tails than a normal distribution - the smaller the sample size, the more spread out is the t-distribution (more probability is in the tails). The probabilities of the t-distribution vary with sample size and can be shown to be dependent on the quantity known as the "degrees of freedom" (n-1, where n = sample size). As the sample size increases, the t-distribution gradually assumes the form of a normal distribution.

For small samples that cannot be assumed to follow a normal distribution, the t-distribution should not be used; "distribution free" or non-parametric statistics are appropriate in this case.

There are other distributions which apply to some biological data, e.g., the binomial distribution, the Poisson distribution, but these are used relatively infrequently in oncological literature and so will not be discussed here.

Significance Testing

The purpose of significance testing is to assess how strong is the evidence for a true difference between two sets of measurements, or between one set of measurements and a standard. This strength of evidence is quantified in terms of probabilities, P-values, such that the smaller the value of P, the less likelihood there is of a difference having arisen by chance. However, a small P-value is not absolute proof of a true difference; if it is assumed that a P-value < 0.05 is the criterion for evidence of a true difference, it is then accepted that there are five chances in one hundred of finding a so-called difference when no such difference really exists.

The starting point in the use of a test of significance is the formulation of two opposing hypotheses about the population of interest. A hypothesis is a set of assumptions, expressed in a coherent manner, about observable phenomena. The null hypothesis, H_0, is that there is no difference between the two sets of measurements. Alternative hypotheses, of which there may be more than one, are denoted H_1, H_2, etc. The null hypothesis is then assumed to be true and the sample statistic calculated under the probability distribution of the null hypothesis. A one-sided test (allowing for difference only in one direction, i.e., x_0 only $> x_1$, or x_0 only $< x_1$) or a two-sided test (difference can be in favour of one measurement or the other) can be applied.

If the probability of obtaining a particular value for the sample statistic is small under the probability distribution, this could be evidence that the assumption of the null hypothesis, and thus the null probability distribution, is not correct. However, there are two other factors which may result in a small P-value:

1) Chance alone
2) Bias due to other factors - sample or study problems.

Significance tests enable one to test whether chance variation could reasonably explain the difference. Once a significant difference is found, the relevance of bias needs to be considered - all assumptions need to be checked very carefully before hypothesis test results are put forth as evidence in support of research conclusions. The level of significance, α, needs to be set by the investigator; it is commonly $\alpha < 0.05$.

If the null hypothesis is unable to be rejected at the specified level of significance, this does not prove that a difference does not exist - it means that if the null hypothesis is in fact false the experiment is not able to detect it at the specified level of significance.

One should also be aware that if the level of significance is set at $P = 0.05$, the difference between a sample statistic of $P = 0.04$ and $P = 0.06$ is not great; $P = 0.06$ may well be indicative of a true difference, but the available data provide insufficient evidence. Borderline P-values should be interpreted with care. It should also be realized that statistical significance is not the same as clinical importance.

Repeated or multiple hypothesis testing ("data dredging") can jeopardize the validity of significance tests. If excessive use is made of significance tests a certain number of false positives are bound to arise. Bonferroni's inequality implies that when k tests are performed, each with significance level P, the probability of one or more significant tests by chance alone is at most kP. Thus it is sometimes said that the significance level should be multiplied by the number of tests (probably an over-correction). One general way of overcoming the problem is to specify in advance a limited number of major analyses that are to be performed; any extra analyses derived after data inspection must then be viewed with considerable caution.

Multiple endpoint analysis similarly increases the risk of false positives.

Type I and II Errors

A Type I error occurs when the null hypothesis is rejected when it in fact is true.
A Type II error occurs when the null hypothesis cannot be rejected but is in fact false.

The level of significance, a, is the probability of making a Type I error.
The probability of a Type II error is denoted by β.
The quantity 1 - β is the probability of correctly rejecting the null hypothesis when the null hypothesis is false and is called the power of a statistical test, i.e., the probability that a study will find a statistically significant difference when a difference really exists. The levels of α and β are inter-related.

Power Calculations

If significance testing has failed to reject the null hypothesis, then the power of the study should be calculated, using the specified value of α, the sample size and the difference between the two sets of measurements that is required to be detected (e.g., a 10% difference in two proportions). If the power of the study is shown to be small, then one must question the validity of the study's conclusions. An acceptable level of Power (1 - β) is commonly set at 80% or 90%.

A study of 71 negative randomized trials by Freiman et al found that 67 (94%) had insufficient power to detect a reduction in morbidity or mortality of 25%, and 50 (70%) could not detect a 50% reduction, using a one-tailed significance level of 5%. Thus the power of a study is a practical clinical issue, not just a statistical nicety. In a case-control study where the number of cases may be limited, the power of the study can be increased by increasing the case:control ratio up to 1:4 or 1:5; it is not recommended that higher ratios be used, as there is little gain in statistical power above this level.

Examples

1) In a sample of 500 patients with lung cancer, 70 were found to have clubbing. Is this data consistent with an incidence rate of 10% for clubbing?

2) An experimental oncologist is treating artificially-induced cutaneous tumours in laboratory animals with two different drugs. Group 1 receives Notsonicin and Group 2 Flatinem, the response being recorded (as the decrease in the size of the tumour) two weeks later. The means, standard deviations and sample sizes for the groups are as follows:

Drug	Sample Size	Mean (cm)	Standard Deviation
Notsonicin	10	2.2	1.2
Flatinem	9	1.2	1.0

Do the drugs differ in effectiveness?

3) Eventually, Notsonicin makes it to a phase III study, where it is compared to the standard drug therapy for carcinoma of the big toe, Emesium. The complete response rates for the two groups are as follows:

Drug	Sample Size	Response Rate
Notsonicin	200	90%
Emesium	230	80%

Does the data support the null hypothesis of no difference between the two response rates? What is the power of the study to detect a 10% difference in the two response rates?

Answers (Method only)

1) This is a large sample significance test for a single proportion - a P-value based on the normal distribution should be calculated (presuming necessary assumptions have been met).

2) This is a two sample significance test comparing means on different groups with small numbers - a P-value is calculated under the t-distribution (note assumptions again).

3) This is a large sample significance test for the difference between two proportions - the P-value using the normal distribution is calculated (note assumptions again).
Power calculations based on sample size and chosen levels of α and β.

Confidence Intervals

Significance tests give the strength of evidence for one set of measurements being truly different to another, but they do not indicate the magnitude of this difference. To find out how different two sets of measurements are, confidence limits (or the confidence interval) are calculated. The term "95% confidence limits" implies that there is a 95% chance that the interval between these limits contains the true difference that would be found if the entire population were tested.

There is a link between confidence limits and significance tests - if the test is significant at the 5% level then the two 95% confidence limits will be in the same direction, (e.g., +0.8% and +22.6%). If the confidence interval contains zero, then the significance test will not show a statistically significant difference at the chosen level of significance.

People not used to the concept of confidence limits are often surprised at how wide the confidence interval can be, e.g., from 0.4% to 6.0% - the true difference in this case is likely to be around 3% but could be almost nothing or nearly double that. The larger the samples, the smaller the confidence interval and the more confident one can be about the precision of the estimate of the difference.

Example

In a clinical trial of patients with nastiomas, the percentage of responders to chemotherapy is 80% and the percentage of responders to radiotherapy is 95%. The 95% confidence interval for the difference is 5% to 40%. What are the implications of these values?

Answer

A true difference between the two treatments has been demonstrated; the confidence interval does not contain zero. The difference in favour of radiotherapy may be as high as 40% or as low as 5% but is likely to be around 20%. A corresponding P-value of 0.03 was established by significance testing.

Sample Size

Before undertaking any study, calculations should be done to determine the appropriate sample size. The method will depend on whether one is estimating a confidence interval or undertaking a significance test.

To calculate the sample size necessary to achieve a confidence interval of a specified width requires one to specify α, the width of the confidence interval (i.e., within 5% of the true value) and the standard deviation.

To calculate the sample necessary for comparison of two sets of measurements, one must specify α, β and the difference between the two sets of measurements that one would like to be able to detect should such a difference exist (e.g., 10% difference between the two proportions).

If in the planning stage of any study sample size calculations indicate the required sample size to be such that there is no chance of the study attaining this number, then the study should not go ahead or thought should be given to a multicentre trial.

Assumptions

All forms of statistical analysis operate under a number of assumptions. These assumptions vary with the type of statistical test being undertaken, but some common examples are:

1) The sample has been chosen randomly.
2) The population is normally distributed.
3) The two sets of measurements have been made independently.

For each test that is applied to data, one should be sure that all the assumptions necessary for a valid test have been satisfied. If this is not the case, or if there is any doubt, the results should be treated with suspicion, as they may not be meaningful. If a problem with assumptions is suspected, it is prudent to ask a statistician to check the data.

Non-Parametric Statistics

If there is any doubt that the assumptions necessary for a valid statistical test can be satisfied, then non-parametric statistics should be used, e.g. for a small sample where the variable of interest does not appear to be normally distributed. Non-parametric statistics are also appropriate when one or more variables are measured on an ordinal scale (observations are ordered, but differences between values do not have a precise meaning, e.g. a composite score for well-being calculated from a questionnaire) as the variables would not be expected to have a normal distribution.

The simplest form of non-parametric test is the sign test, where measurements are classified as positive or negative with reference to a median value and the signs summed. More sophisticated non-parametric tests take into account the magnitude of the differences as well as their direction. Below are listed some non-parametric tests and their parametric equivalent:

Wilcoxon matched pairs test	paired t-test
Mann-Whitney U test	two sample t-test
Spearman's or Kendall's rank correlation coefficient	product-moment correlation coefficient

Example

The following are the responses of subjects (measured as decrease in size in cm) with metastases to the big toe treated with either radiotherapy alone (Group 1) or with radiotherapy + flatinem (Group 2):

Group 1:	1.2	0.9	0.7	3.2	1.4	1.7	2.1	4.7	1.3	1.1
Group 2:	0.6	0.7	1.3	4.6	1.2	1.1	1.1	4.0	2.0	1.0

Is there evidence of a difference in treatment effect between the two groups?

Answer (Method Only)

These data look rather skewed, so it may not be appropriate to apply a two-sample t-test. The Mann-Whitney U test should be used instead. It can be informative to compare the two results, and there may be little difference. However, if the results show considerable variation, the Mann-Whitney U test should be taken as the more accurate estimate.

Analysis of Variance (ANOVA)

This procedure is an extension of hypothesis testing which is used to compare three or more means. As previously discussed, using multiple significance tests inflates the Type I error rate substantially. ANOVA avoids this problem by testing only one null hypothesis to determine if any of the populations differs from the rest.

Despite being a test to compare means, ANOVA examines variances. The test statistic is called the F-statistic, which is compared with the F-distribution tables to obtain significance levels. These levels depend on the total number of observations - all the sample sizes added together and the number of different populations from which the samples are drawn. If the null hypothesis is rejected then at least one of the population means differs from the rest. To find out which one is different, confidence intervals for each mean are estimated. These confidence intervals will overlap except for that of the different population mean. Alternatively, hypothesis testing can be used to compare selected pairs of means using a modified technique known as the "protected" t- procedure.

ANOVA is described as one-way or two-way, depending on the experimental design which has been employed. One-way ANOVA is used to test the means of three or more treatments which have been assigned completely at random (completely randomised design). Two-way ANOVA is used to analyse data from a randomized complete block design where patients are subdivided into groups ("blocks") such that the number of patients in a block is equal to the number of treatments being studied and each treatment is assigned at random to patients within each block. Two-way ANOVA is also used to analyse data in which there are two factors of interest and each group which is sampled is formed by combining a level (or value) for each of the two factors.

When the assumptions of the ANOVA procedures cannot be met, non-parametric statistics should be used:
e.g., Kruskal-Wallis test substitutes for One-way ANOVA;
Friedman test substitutes for Two-way ANOVA (for RCBD).

Example

A three-arm randomized control trial is being conducted comparing three different radiation schedules in the treatment of fascinoma. The responses of the first 30 patients (measured in cm) are as follows:

Schedule 1	Schedule 2	Schedule 3
6.4	9.6	5.8
8.8	9.0	7.4
7.2	8.0	6.6
8.0	8.0	6.0
7.9	7.9	8.2
7.1	8.8	7.5
7.5	8.7	7.6
6.0	8.2	9.0
8.0	9.1	8.2
6.5	7.3	7.4

Is there any evidence of a difference in treatment effect?

Answer (Method Only)

This is an example of one-way ANOVA.

Contingency Tables and Chi-Square Tests

Contingency tables are the frequency counts of a series of observations grouped according to category, with each category having two or more levels. The general format is:

		Outcome Factor		
		Yes	No	
Study	Yes	a	b	n3
Factor	No	c	d	n4
		n1	n2	N

These tables allow analysis of categorical data, i.e., data that does not have a numerical value, but is in the form of counts, e.g., nodal status vs presence or absence of metastatic disease.

The commonest form of contingency table is a 2 x 2 table as above (2 rows and 2 columns) but larger tables can also be analysed.
It can be very difficult to measure the strength of the association between two qualitative variables, but it is easy to test the null hypothesis that there is no relationship or association between the two variables. This is the basis of the Chi-square test.

For each outcome cell (a, b, c, d, above), the "expected" frequency under the null hypothesis is found by calculating the value:

$$\frac{\text{row total x column total}}{\text{grand total}}$$

Observed and expected frequencies must then be compared. If the two variables are not associated, the observed and expected frequencies should be close together, any discrepancy being due to random variation. For each outcome cell the following is calculated:

$$\frac{(\text{observed value - expected value})^2}{\text{expected value}}$$

These quantities are then summed over all of the cells to obtain a value for the chi-square statistic (c):

$$c^2 = \frac{\underline{S\ (O - E\ )^2}}{E}$$

The distribution of this test statistic when the null hypothesis is true and the sample is large enough is the chi-square distribution, with degree of freedom given by:
(number of rows - 1) x (number of columns - 1)
i.e., for a 2 x 2 table, there is one degree of freedom.

The calculated chi-square statistic can be measured against standard tables giving probability values for varying degrees of freedom, e.g., for 1 degree of freedom the probability of a c^2 value of > 3.84 is 0.05. The numeric value of the chi-square statistic is not an indication of the strength of the relationship but is dependent on the sample size. The strength of a relationship so established can be estimated by calculating relative risk, attributable risk, odds ratio and confidence intervals (exact calculation depending on the study design).

If a matched pairs design is used, e.g., a matched case-control study, only the discordant pairs are analysed.

		Outcome Control		
		Yes	No	
Outcome	Yes	a	**b**	n3
Case	No	c	d	n4
		n1	n2	N

McNemar's statistic is derived:

$$c^2 = \frac{(b - c)^2}{(b + c)}$$

If $^{b + c}/_{2} < 5$, the binomial distribution is used to calculate a P-value.
The odds ratio in this situation is estimated by $^{b}/_{c}$

The chi-square statistic may not be accurate for tables with expected values less than 5. In such situations, an alternative test of association of a 2 x 2 table is used. This is Fisher's exact test, which is rather difficult to calculate manually, but is performed by most statistical software packages.

The chi-square test for trend can be used if there are more than two categories which are on an ordered scale (e.g., non-smoker, smokes occasionally, smokes regularly). A trend may be significant even if the overall contingency table c^2 is not. This is because the test for trend has greater power for detecting trends than has the ordinary chi-square test.

Mantel-Haenszel Method of Combining 2 x 2 Tables

This method was developed to overcome the effect of confounding, which may be a problem when combining tables.

If one wants to combine information across contingency tables, (e.g., an identical study carried out in three different hospitals) one should first investigate each table separately to verify that the relationship between variables is similar in all tables. If this appears to be so, then the next step is to assess the average association across the tables using an appropriate method which is not affected by differences in the distribution of the variables across the tables. One such method is the Mantel-Haenszel method, which can be used to test the null hypothesis that on average there is no association, and also measures the average strength of any demonstrated association by calculating an odds ratio summed across all strata / tables with a confidence interval estimated.

Confounding may be identified by comparing the crude and adjusted measures of association ("crude" c^2 calculated by straight addition of values for a particular cell across all tables). If the crude and adjusted measures are different then confounding is present.

Before calculating the adjusted estimate of the odds ratio, one should check that the odds ratios for the individual tables are relatively constant (i.e., consistently elevated or reduced or approximately equal to 1). If this is not the case, then there is an interaction present between the exposure of interest and the stratifying variable. This third variable is called an effect modifier. It would be inappropriate to apply the Mantel-Haenszel procedure under these circumstances because the study exposure does not have a single overall effect.

Meta Analysis

Meta-analysis is a technique which combines evidence across a number of trials or studies which all address the same research question but which individually are too small to detect an effect. The statistical procedure used to test the null hypothesis and to obtain an overall estimate of the measure of effect (such as a relative risk or odds ratio) is based on the Mantel-Haenszel methods.

Meta-analysis aims to:

1) Increase statistical power to detect significant new treatments.
2) Demonstrate the importance of contextual effects.
3) Show how to optimise treatment by comparing different treatments.
4) Assess the stability and robustness of treatment effectiveness.
5) Indicate when research findings are sensitive to research design.

However, meta-analysis has a number of problems. As well as the usual potential sources of bias within each study, there are biases specific to the technique, e.g., publication bias, bias in the selection of studies for inclusion, data extraction bias. There are risks in combining data from different populations and there may be a loss of information by emphasising average outcome, e.g., when the treatment effect is not homogeneous throughout the studies examined. However, a carefully planned and executed meta-analysis can provide very useful information. Whenever a decision is made to utilise meta-analysis the aim should be to maximise the benefits and to minimise the pitfalls.

Example

a) Patients with nauseoma have been classified as node negative or node positive at presentation. Their rates of local recurrence following treatment are as follows:

	Local Recurrence	
	Yes	No
Node Positive	37	204
Node Negative	28	196

Is there any association between nodal status at presentation and development of local recurrence?

b) The nodal status was assessed clinically in some patients and confirmed surgically in others, as follows:

Surgical staging

	Local Recurrence	
	Yes	No
Node Positive	10	32
Node Negative	7	24

Clinical staging

	Local Recurrence	
	Yes	No
Node Positive	27	172
Node Negative	21	172

Test the association across the two strata. Is there any evidence that confounding by the method of staging occurs?

Answer (Method Only)

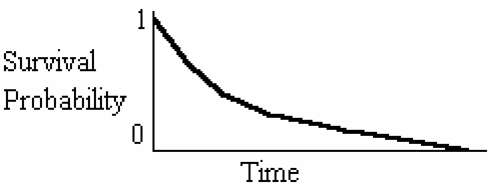
a) A c^2 statistic can be calculated to ascertain if there is any relationship. A "crude" odds ratio can also be calculated.

b) A c^2 statistic and an odds ratio should be calculated for each stratum. If the odds ratios are similar, it is reasonable to combine the strata using the Mantel-Haenszel method. If the adjusted odds ratio is different to the crude odds ratio then confounding as a result of the method of staging is occurring.

Survival Analysis

Methods have been devised for summarising and graphing the survival experience of one or more groups of patients using the time to death or the occurrence of some event, and then comparing this experience in two or more groups. These methods have been designed to accommodate different lengths of follow up due to patients being enrolled in a study at different times, death or occurrence of the event of interest before the end of the follow up period and loss to follow up. "Censored" observations are those in which individuals were observed for only a part of the follow up period, with death or the event of interest not occurring during that time interval. Such techniques are well suited to prospective studies and clinical trials.

The survival function represents the probability that an individual survives longer than a specified time, and follows the general form:

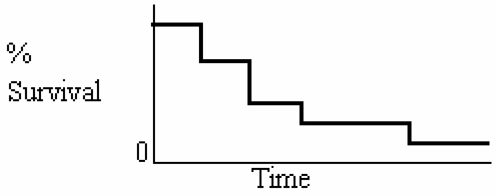


There are two methods of estimating this survival function from sample data in which there may be censored observations - one for ungrouped and the other for grouped data.

The Kaplan Meier Survival Curve

This method (also known as product limit) is used for ungrouped data, i.e., follow up data for individual patients. It assumes that the reason for censored observations is independent of, or unrelated to, the cause of death. This situation may not always apply with censored observations.

The Kaplan-Meier technique calculates the probability of surviving one interval and multiplies this value by the probability of surviving the successive interval (the interval being defined by the occurrence of death or the event of interest) to produce a cumulative probability. A censored observation does not change the probability of surviving the interval in which it occurs but does have an impact on the probability of surviving the subsequent interval (by removing that patient from the calculation of survival probability in the subsequent interval). This process is continued serially through the entire follow up period. The survival function is thus a step function:

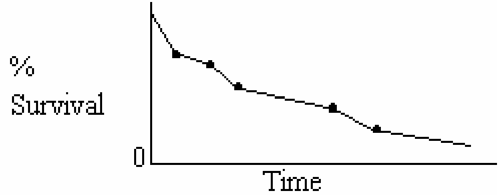


As more patients and survival times are observed the steps become smaller and the graph becomes more like a smooth curve. A median survival time can be estimated and is most reliable when nearly everyone experiences the event of interest, the data is extensive and the Kaplan-Meier curve falls rapidly between 70% and 30% surviving. If the Kaplan-Meier curve always exceeds 50% survival, then the median survival time cannot be estimated - all that can be stated is that the median is longer than the follow up period.

The Life Table Estimate

This method (also known as the actuarial method) is used where data is reported in categories of time or there is such a large number of patients that grouping them according to follow up time simplifies the computation and presentation of the data. In this method it is assumed that censored observations occur randomly throughout an interval and thus all are treated as having been observed for half of the time interval. It is also assumed that the probability of survival at one time period is independent of the probability of survival at other time periods.

The probability of surviving each interval is calculated as the number surviving, divided by the number of persons beginning the interval minus half the number censored. The cumulative survival probability is the product of the probabilities of surviving each interval up to the time of interest.



The Log Rank Test

The log rank test is used to detect a difference between survival curves resulting from the mortality rate in one group being consistently higher than the corresponding rate in another group, with the ratio of the two rates being constant over time (this is called proportional hazards).

The log rank test is primarily a test of significance, with the null hypothesis being that there is no difference in survival experience between the two groups. If the null hypothesis is true, then the number of observed and expected deaths in an interval should differ by chance only.

A c^2 statistic is used for comparing the two groups:

$$c^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

If the null hypothesis is true c^2 will be approximately distributed as a chi-square random variable. If c^2 is larger than the critical value obtained from a table of the chi-square distribution then the null hypothesis is rejected and it may be concluded that there is a statistically significant difference in survival for the two groups.

A stratified log-rank test can be performed to compare the survival experience of two groups adjusted for another variable that is a potential confounder.

The comparison of survival in two groups may be restricted to a specific interval of the observation period e.g. early or late survival. The generalised Wilcoxon test can also be used to detect differences in the early survival experience of two groups (more weight given to early deaths).

If too great a period of follow up occurs there is a great danger of over emphasising later treatment differences based on very few patients.

The log-rank test can be extended to more than two groups.

Proportional Hazards Model (See "Cox Regression" in "Regression and Correlation")

Competing Risks

Assessment of "Cure"

The chronicity of some diseases e.g. breast cancer makes the assessment of cure very difficult. Two techniques have been used to define "cure":

1) The observed survival of a group of patients is plotted. A second curve of the expected survival of a group with the same age and sex distribution as the treated group is also plotted and the two curves compared. Parallelism in the two curves indicates that the subsequent death rate in the two groups is identical - this can be equated to "cure".

2) The observed survival can be converted into an age-corrected survival by dividing the actual survival in each interval of follow up by the expected survival for members of the general population of the same age and composition. This age and sex corrected survival is then plotted, the point at which the curve becomes horizontal being referred to as the "point of definitive cure".

Example

In a small clinical study, the survival of patients following radiotherapy to a horridoma was documented. Data are given below:

Patient	Years Since Diagnosis	Death (1) or Censored (0)
1	0.5	1
2	0.9	0
3	1.0	0
4	1.7	1
5	2.3	1
6	3.4	0
7	5.2	1
8	2.6	0

The above patients were all male. Below are the figures for females:

Patient	Years Since Diagnosis	Death (1) or Censored (0)
1	0.8	1
2	1.4	1
3	1.4	0
4	1.9	1
5	3.7	1
6	5.2	1
7	6.3	0
8	7.2	0
9	8.8	1

a) What type of survival analysis is appropriate?
b) Is there any difference in the survival experience of the two groups?

Answers

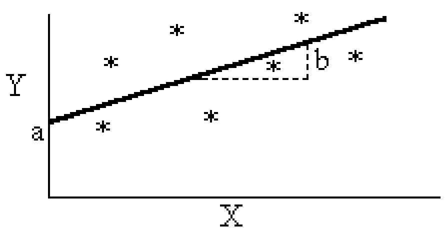
a) This data is ungrouped - a Kaplan-Meier survival curve should be carried out.
b) Calculate the log rank statistic to see if there is any significant difference in survival dependent on sex. A generalised Wilcoxon test may be interesting here.

Regression and Correlation

Regression and correlation are two methods of analysing the relationship between two quantitative variables.

Regression

Regression is a method which ascertains the probable form of a numerical relationship between two variables. Linear regression gives the equation of the straight line that describes how the outcome variable (Y) increases (or decreases) with an increase in the predictor variable (X):



$Y = a + b X + E$ where b = regression coefficient
E = error variable

a and b can be calculated using the method of least squares, which minimises the sum of squares of the deviations about the regression line.
The mean point X, Y always lies on the regression line.
The null hypothesis of no relationship can be tested by calculating a t-statistic:

$$t = {}^{b}/_{\text{standard error b}}$$

which has a t-distribution with n - 2 degrees of freedom when the null hypothesis is true.

The coefficient of determination R^2 can be calculated and is equal to the proportion of the variation in the outcome variable explained by the regression, e.g., $R^2 = 82\%$ means that 82% of the variation in the dependent variable can be explained by the regression.

The regression line can be used for predicting what value Y is likely to assume given a particular value of X, but should not be used to extrapolate beyond the data used to obtain the regression equation.

Note that there are a number of assumptions made about the data when performing linear regression and these should always be checked carefully.

Correlation

Correlation is a method which describes the strength of a relationship between two variables. It does not require one variable to be specified as dependent and the other as independent. In correlation X and Y are assumed to vary together in what is called a joint distribution. A correlation coefficient (r) can be estimated and is denoted the product moment (Pearson's) correlation coefficient; it is related to R^2, the coefficient of determination ($r^2 = R^2$).

When all points lie exactly on a straight line, r = 1 (or -1); and when there is no relationship at all, r = 0. Even if there is a perfect mathematical relationship, the correlation coefficient will not be exactly 1 unless the relationship is of the straight line form $Y = a + bX$.

The null hypothesis ($H_0 : r = 0$) can be tested using standard tables if at least one of the variables is normally distributed, or by a t-statistic with (n-2) degrees of freedom if both variables are normally distributed.

Multiple Regression

Multiple regression is an extension of linear regression to the situation where there are any number of independent variables to be considered. The general form of a regression model for k independent variables is given by

$$Y = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots\ldots + b_kX_k + E$$

where X_1, X_2, ...X_k are independent variables which may be continous or discrete
and a, b_1, b_2, ...b_k are the regression coefficients

Estimating these regression coefficients is very complicated manually, but there are many computer packages available which perform multiple regression.
As usual data should be checked prior to analysis to ensure that all assumptions are met.
The computer will test the null hypothesis that $b_i = 0$ (where b_i is the independent variable) by calculating a t-statistic and will estimate the coefficient of determination adjusted for the number of independent variables included in the regression equation. An analysis of variance table provides an overall summary of the multiple regression equation.

A process known as stepwise regression can be performed by computer. The purpose of stepwise regression is to find the model with the smallest number of independent variables. This is done by starting with some model involving one or more of the available independent variables and then adding or deleting variables one at a time to improve the model. When no appreciable gain is achieved by taking another step, the procedure stops. The resulting model may not be the best model but it is generally one of the best.

In an automated program, the first variable to be chosen is that X_i with the highest correlation with Y. The second variable need not be that with the second highest correlation with Y because that variable may be too highly related to the first variable chosen and so be redundant. It is generally possible for the user to specify the criteria for adding and deleting variables in the model (the criteria relate to the strength of the relationship between the independent variable and the dependent variable). Forward selection and backward elimination procedures may lead to different final models for a data set because the level of significance of one independent variable depends on which other independent variables are also included in the model.

Multiple Logistic Regression

This is a technique devised for analysis of outcome variables which are dichotomous e.g. dead / alive, diseased / not diseased. Since the dependent variable Y can only take one of two values multiple linear regression techniques are not appropriate. A procedure called the logit transformation (Z) can be performed to transform the probability of success into a value between 0 and 1.

$$Z = \ln ({}^{P}/_{1 - P})$$

where ln = natural logarithm
and P = probability of success

This function is often used in epidemiology to model the risk (or probability) of disease development, death or recovery, or whatever the outcome of interest as a function of various independent variables. The following multiple logistic model defines how P (the probability of achieving a response) depends on X_i, the independent variables (or prognostic factors)

$$\ln ({}^{P}/_{1 - P}) = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots + b_kX_k$$

where a, b_1, b_2, ..., b_k are numerical constants called logistic coefficients

The quantity (${}^{P}/_{1 - P}$) is the odds and ln (${}^{P}/_{1 - P}$) is the log odds. The log odds is the most statistically manageable way of relating probabilities to explanatory variables. For a particular independent variable X_i, the value $X = 1$ indicates presence (exposure) and $X= 0$ indicates absence (no exposure), and the log odds for those with the variable compared to those without is given by:

$$\text{Odds ratio} = \exp (b_i)$$

Independent variables have a multiplicative effect on the odds ratio. A method called maximum likelihood is used by the computer to obtain estimates of the regression coefficients which maximise the likelihood function, i.e., the probability of observing the outcomes given an individual's value of the Xs. Once the maximum likelihood estimates have been obtained, the values of the regression coefficients can be used to make statistical inferences about the relationship between each of the independent variables and the outcome. The significance of the independent variables is tested by comparing models with and without the independent variable of interest using the likelihood ratio test or the Wald test.

To estimate directly the effect of a change in one or more variables on the risk of disease, one substitutes their values into the equation. The relative importance of each variable may be compared using standardised regression coefficients.

With the logistic model all that is required to adjust for confounding variables (e.g., age) is to include them in the equation. Interaction between variables can be dealt with by the inclusion of an interaction variable in the equation.

Cox Regression

A statistical method of analysing survival data which is equivalent to multiple regression for a quantitative response or the multiple logistic model for a qualitative response is the "proportional hazards" model (also known as Cox regression).
The "hazard function" l(t) for a patient alive up to time t is the risk of failure to survive a further unit of time. If treatment and prognostic variables are converted to numerical variables X_1, X_2, ...,X_k , the proportional hazard model is

$$\ln l\,(t) = c_o(t) + c_1X_1 + c_2X_2 + \ldots\ldots + c_kX_k$$

$c_o(t)$ represents the fact that the hazard function varies with time, whereas the constants c_i indicate the extent to which the risk of dying is affected by treatment and prognostic factors. The constants c_i , their standard errors and P-values may be obtained using maximum likelihood. A positive value for c_i indicates that the hazard function increases with X_i; that is, high values of X_i are associated with poorer survival.

Example

A study looking at two different dose schedules in the treatment of rectal cancer has been carried out. The outcome variables are local relapse, survival and occurrence of major treatment complications. Information on a number of possible prognostic factors has been obtained - age, stage of disease, extent of surgery etc. How can it be determined whether the dose of radiation is significantly related to the outcome variables?

Answer (Method)

"Major treatment complications" is a dichotomous outcome (yes / no). A logistic regression model will adjust for the possible prognostic factors and allow comparison of the effect of the two different dose schedules on this outcome. "Local relapse" and "survival" can be treated similarly (present / not present at X years, alive / dead at X years) using a logistic model, but if definite times for these events are available e.g. 3.4 years, then a proportional hazards model may allow a more precise comparison of the effect of the two different fractionation schemes.

SECTION D: A METHOD OF CRITICAL APPRAISAL

Critical appraisal is a process that allows the reader to assess the value of published research and determine the relevance of the results to the local situation. It requires the identification of key elements of the published study, as these features form the basis for assessment of the overall worth of the article.

Most people who read medical literature develop their skills in evaluating articles through experience, discussion and information acquired in other ways. Few individuals ever approach "the literature" in a systematic manner, and by so doing can miss important aspects influencing the overall value of a published study. Numerous systems have been proposed and used for evaluation of published articles. The following worksheet is one suitable method, developed by the Centre for Clinical Epidemiology and Biostatistics at the University of Newcastle, and incorporating all the skills you have hopefully acquired through reading the previous three sections.

One area of medical literature that is growing rapidly and which I have not covered in this guide is that pertaining to health economics e.g. cost-benefit analysis, cost-effectiveness analysis. This is another subject in itself and if you would like to know more a good introductory textbook is:

Drummond M.F., Stoddart G.L. and Torrance G.W. Methods for the Economic Evaluation of Health Care Programmes, 2nd Ed. Oxford University Press, 1997.

Good luck and good reading!

REFERENCES

Alderson, M. An Introduction to Epidemiology. MacMillan Press, 1976.

Bland, M. An Introduction to Medical Statistics.3rd Ed. Oxford University Press, Oxford, 2000.

Christie D, Gordon I, Heller R. Epidemiology. An Introductory Text for Medical and Other Health Science Students 2nd Ed. New South Wales University Press, 1997.

Freiman JA, Chalmers TC, Smith HJ, Keuhler RR. The importance of beta, the type II error and the sample size in the design and interpretation of the randomised control trial. N E J M 299 : 690-4, 1978.

Ingelfinger J A, Mosteller F, Thibodeau L A, Ware, JH. Biostatistics in Clinical Medicine. MacMillan Publishing Co, New York 2nd Edition, 1987.

Last JM. A Dictionary of Epidemiology. Oxford University Press, 4th Edition, 1997.

Mould RF. Introductory Medical Statistics. 3rd Edition, Institute of Physics Publishing, 1998.

Pocock SJ. Clinical Trials: A Practical Approach John Wiley and Sons Ltd, Chichester, 1984.

Rothman KJ. Modern Epidemiology. 2nd Edition, Little Brown & Co., Boston, 1998.

Various. Biostatistics I (distance learning): course modules. University of Newcastle, 1991.

Various. Epidemiology I (distance learning): course modules. University of Newcastle, 1991.

CRITICAL APPRAISAL WORKSHEET

	1	2	3
	Can you find this information in the paper?	Is the way this was done a problem?	Does this problem threaten the validity of the study?
1.	What is the research question and/or hypothesis?	Is it concerned with the impact of an intervention, causality, or determining the magnitude of a health problem?	
2.	What is the study type?	Is the study type appropriate to the research question?	If not, how useful are the results produced by this type of study?
3.	What are the sampling frame and the sampling method?	Is there selection bias?	Does this threaten the external validity of the study?
4.	In an experimental study, how were the subjects assigned to the groups? In a longitudinal study how many reached final follow-up?		Does this threaten the external validity of the study?
5.	What are the study factors and how are they measured?	Is there measurement error?	Is measurement error an important source of bias?
6.	What are the outcome factors and how are they measured?	a) Are all the relevant outcomes assessed? Is there measurement error?	How important are omitted outcomes? Is measurement error an important source of bias?
7.	What important potential confounders are considered?	Are potential confounders controlled?	Is confounding an important source of bias?
8.	Are statistical tests considered?	Were the tests appropriate for the data? Are confidence intervals given? Is the power given?	
9.	Are the results clinically and/or socially important?	Was the sample size sufficient to detect a clinically/socially significant result?	Is the study useful or is the result inconclusive?
10.	What conclusions did the authors reach about the research question? Did they generate new hypotheses? Do you agree with the conclusions?	Do the results apply to the population in which you might be interested?	Do you accept the results of this study?