**4**

UNIT 4 - Linking Exposure to

Disease

# 

# 

# Introduction

Welcome to Unit 4: you are now well into the module, and will have a good sense of descriptive epidemiology, descriptive statistics and the use of statistics in sampling. This unit presents the basic toolbox that epidemiologists use to explore the relationships between health outcomes and their risk factors.

Contents

Sessions 1-2 & Sessions 4-7 presents statistical methods which allow researchers to “quantify” uncertainty and evaluate the probability that the results of their studies are due to chance.

Session 3 introduces the basic concepts and methods of Analytical Epidemiology, i.e. the part of the epidemiological investigation which, going beyond the simple description of the data, aims at analysing what causes a given disease (or health outcome).

Unit 4 – Session 1: Sampling Methods

## Introduction

Given a study population, there are many different ways to extract a sample from this population in order to use it to infer (using the methods from the inferential statistics) the characteristics of the population. Each method has advantages and disadvantages, and most often the choice among them is a trade-off between “quality” of the sample (i.e. “how well” it represents the population) and practical applicability and cost of the sampling procedure.

## Timing

This session should take you about an hour. In this Session there is one reading, a task and feedback to the task. Please attempt the task before looking at the feedback.

## Learning outcomes

* Understand the difference between probability and nonprobability samples;
* Define sampling frame;
* Know the sampling methods most commonly used in epidemiology.

Sampling methods

The most important characteristic of a sample is its “representativeness”: a sample is *representative* of a population if the characteristics of the individuals in the sample are *similar* to the characteristics of the individual in the population. We study samples because we cannot study populations, but the aim of a study is to draw conclusions about the population: the sample is only a means to an end, but the interest lies in the generalization of what we see in the sample.

It is beyond the scope of this course to present the theory underlying a correct sampling and to define formally what in Statistics is considered a “good” and a “bad” sample, but it is quite intuitive that, if the characteristics of the individuals in the sample are different from the characteristics of the individuals in the population, the sample is not “good” and generalization is not warranted.

With the objective of representativeness in mind, different methods of sampling are available, and they can be classified in two categories*: Probability samples* *and Nonprobability samples*.

A *probability* (or *random*) *sample* is a sample in which the probability of being included in the sample is known for each subject in the population, while in a *nonprobability sample* these probabilities are unknown and may reflect selection biases of the person doing the study.

Although probability samples can be drawn in many different ways, among them a special importance is given to *simple random samples*, in which the probability of being included in the sample are known and *are the same for all individuals.* Most statistical methods are based on this assumption. A practical way to select a simple random sample from a population for which a list of individuals is available (this list is called the *sampling frame*) is using a table of random numbers or a computer- generated list of random numbers.

Statistical methods for inference always assume that the probabilities of inclusion in a sample are known, and, therefore, they assume that some form of random sampling has been done. For this reason, in the following sections of this module,when we will study estimation methods, the term “sample” will be considered as synonymous with probability sample,and, more precisely, as synonymous with a simple random sample).

The reading by Bruce *et al*. (2009) describes the main concepts underlying sampling and summarises the sampling methods most commonly used in epidemiology.

**Reading**

Bruce, N., Pope, D., & Stanistreet, D. (2008). Quantitative Methods for Health Research: A Practical Interactive Guide to Epidemiology and Statistics. Chichester: John Wiley & Sons. p. 133-141.

**Task 1 – Working with samples**

Consider the samples described below:

*Sample 1*: In a school with 1000 students, each student is given a different identification code 1 to 1000. Using a computer program, a set of 100 random numbers between 1 and 1000 is generated and the students whose identification codes match one of the random numbers is selected to be part of the sample.

*Sample 2*: In the same school, each student whose identification code ends with 0, is selected to be part of the sample.

*Sample 3*: Using a random number generator, 10 out of the 50 classes of the school are randomly selected. In each class, 10 students are again randomly selected using, within the class, a procedure as in Sample 1.

*Sample 4*: Each of the 100 teachers in the school is asked to give the names of 10 students to be part of the sample.

For each of these samples, using the classifications provided by Bruce *et al*., answer the following questions:

a) Is it a probability or nonprobability sample?

b) More, precisely, which kind of probability (or nonprobability) sample is it?

c) Do you think we can use the sample to draw reliable statistical inference about the population?

**Task feedback**

*Sample 1* is a **simple random sample**, and it is the basic type of sample used to draw reliable inference about the whole school population.

*Sample 2* is a **systematic sample**. It could be seen as similar to a probability sample, provided that the codes are assigned to the students without any specific rule (i.e. the codes are randomly assigned to the students). In this case a systematic sample can be used to draw reliable inferences. This is not the case if codes are assigned with some rule. For example, if male students are assigned even identification codes and female students odd codes, this procedure would create a non probability sample, because female students would have zero probability of being selected.

*Sample 3* is a **cluster random sample** (the classes are clusters). It is a probability sample and can be used to draw reliable inference, provided that appropriate methods of analysis are applied.

*Sample 4* is a **nonprobability** sample: we cannot calculate the probability of selection of each student, because it depends on the choice of the teachers. More precisely it is a **convenience sample**.

The representativeness of these kinds of samples is questionable and therefore, in general, we cannot draw reliable inference on the school population.

# Unit 4 - Session 2: Sample Size Calculation

# Introduction

Researchers must know how large a sample needs to be before beginning their research because they may not otherwise be able to determine significance when it occurs. The power of a study (as defined in session 3.4) depends on the research question but also on the sample size, and it increases with the sample size. A study which is underpowered is a waste of time and financial resources, because it cannot answer the research question.

To avoid this unethical waste of resources, and - even more important - to avoid submitting subjects to unnecessary risks, in studies involving human or animal subjects ethical committees require a preliminary calculation of the power of the study and, consequently, of the minimum sample size.

Sample size calculations can be also useful for the readers of research reports, especially when the results of the study are not statistically significant (often called "negative" results).

In fact, comparing the minimum sample size needed to obtain a given power with the actual sample size used in the study, allows the reader to discriminate between "true" negative results and result which are negative just because the sample was not big enough.

Timing

This session should take you about 2 hours to do as the reading. In this session, there is one reading and a short task based on the reading. The feedback on the task has been provided to you but you are encouraged to attempt the task before having a look at the feedback.

## Learning outcomes

* Understand the reasoning underlying the calculation of sample sizes and the preliminary information needed;
* Calculate the minimum sample size for simple study design and sampling schemes.

## Sample size calculation

Larger sample sizes generally lead to increased precision when estimating unknown parameters of a population. For example, if we wish to know the proportion of people with diabetes in a geographical area, we would generally have a more accurate estimate of this proportion if we sampled and examined 200 rather than 100 subjects.

Several fundamental facts of mathematical statistics describe this phenomenon, including the law of large numbers and the central limit theorem. However, it is important to bear in mind that the relationship between sample size and precision is a complex one, and, in general, doubling the sample size does not double the precision of the estimates. Moreover, increasing sample size cannot do anything to reduce systematic error: a biased study produces biased results, regardless the sample size.

The actual calculation of the minimum sample size needed for a specific study can be done using a variety of formulae and graphical tools, general purpose or specialised statistical software, and also one of the (many) sample size calculators available online. To apply any of these methods the researchers need to specify (1) the *characteristics of the study*; (2) the *required precision*; and (3) and the *hypothesised characteristics of the population*:

1. The first piece of information we need in order to calculate the minimum sample size needed is the **design of the study**. Holding everything else constant, the sample size is different if our study is a cross-sectional study or a pre-post study; it changes is we are comparing two samples which are matched rather than independent; and so on.
2. The second decision we have to make is about the required **precision**. In general, it is obvious that the researchers try to reach the highest precision possible, but, because this comes to a cost (large, and therefore costly samples), a trade-off is necessary.   
   To set the desired precision we need to decide the **significance level** for the statistical tests (usually α=5%), and either the **power** for studies comparing two or more groups (1-β, usually 80% or 90%) or the **absolute precision** of the estimate for prevalence studies (d, in % of the estimate).
3. Finally, we need a **preliminary idea about what we are trying to estimate**. This last requirement often seems counterintuitive, because it requires a previous knowledge about what is the result of the study: for example, to calculate the minimum sample size for a prevalence study, we need a preliminary idea about the prevalence we are going to find after the study is done. In practice, these preliminary estimates (that need not to be extremely precise) come often from previous studies, from extrapolation from similar populations, or, when no previous literature is available, from a pilot study with a small sample.

A note of caution is that most formulae commonly used (and also most algorithms implemented in statistical software) assume that the researchers will implement a **simple random sampling**, and must be modified when this is not the case.

The sample size needed will vary for other forms of probability sampling (remember that all these considerations apply **only to probability samples**), and, in particular, they tend to increase when cluster sampling is used. To overcome this problem, it is common practice to calculate the sample size in the hypothesis of simple random sampling, and then multiply the obtained results by a coefficient (the **design effect**, *deff*) which takes into account the actual sampling scheme.

Again, the design effect needs to be known in advance and, when possible, it is drawn from previous studies **in the same population**. When these studies are not available, a common “default” design effect is deff=2, i.e. the minimum sample size calculated in the hypothesis of simple random sampling is doubled.

The reading by Machin *et al.* (2007) presents a more exhaustive discussion of the basic principles outlined above, and a selection of methods and formulae to calculate the minimum sample size required for a variety of study designs:

**Reading**

Machin, D., Campbell, M. J. & Walters, S.J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 262-274.

**Task 1 – Calculating minimum sample size**

*adapted from* Machin *et al.(2007)*

Using the formulae provided in the suggested reading, calculate the minimum sample size required for the following studies. When appropriate, consider a design effect deff=2.

*Study 1***:** Estimation of the prevalence of nurses dissatisfied with their job in Hospital “X”.

We think it will be about 20%, and we would like to estimate this with an absolute precision of ± 5%. We have a complete sampling frame and we are going to draw a simple random sample.

*Study 2***:** A survey of the results of various types of cholesterol-lowering drugs in general practice, comparing a sample of patients using traditional drugs with patients using a new drug recently introduced in the market. A previous study found that the average serum cholesterol level for the patients in treatment in the same area was 4.99 mmol/l (standard deviation σ = 1.02 mmol/l).

Assume a difference of 0.5 mmol/l in serum cholesterol is worthwhile, and that the researcher wants to perform a two-side test with significance α=5% and power 1-β=80%. Simple random sampling is assumed.

*Study 3***:** An exercise trial in 50–74-year-old men identified as at high risk of a heart attack, to see if a daily exercise regime for a year will lead to a reduction in the number of heart attacks. One group will be given the daily exercise regime and the other control group will receive no help. On the basis of published evidence we expect that in the control group 20% of the men will have suffered a heart attack within the year. We would be interested in detecting a reduction of heart attacks to 15% in the exercise group. How many patients would be needed for a two-sided significance of 5% and 80% power? The study is a multi-site study, and cluster sampling will be used.

**Task feedback**

*Study 1*

Study design Cross-sectional (prevalence) study

Precision d=±0.05 (5%)

Level of significance α=0.05 (5%)

Anticipated prevalence πplan= 0.20(20%)

From Table 14.2 of the reading by Matchin *et al*, we can read the required sample size n=246 (in correspondence of πplan=0.20 and d==±0.05).

*Study 2*

Study design Cross-sectional (prevalence) study

Level of significance α=0.05 (5%)

Anticipated effect size δ=± 0.5 mmol/l

Standard deviation σ = 1.02 mmol/l

Power 1-β=80%.

Using the approximate formula at page 274 and Table 14.3 of the reading by Matchin *et al*, the required sample size (in **each** group) is:

*Study 3*

Study design Cross-sectional (prevalence) study

Level of significance α=0.05 (5%)

Anticipated prevalence group 1 π1= 0.20(20%)

Anticipated prevalence group 2 π2= 0.15(15%)

Power 1-β=80%.

From Table 14.1 of the reading by Machin *et al*, we can read the required sample size n=906 (in correspondence of π1=0.20 and π1=0.15

# Unit 4 – Session 3: Analytical Epidemiology

# Introduction

You have already covered in Unit 2 how to perform basic descriptive epidemiological measurements and basic biostatistical tests. To be able to conduct and critique epidemiological research, basic proficiency in measurement of the occurrence of events and the assessment of the possible association between events is required, as this is the basis of analysis and interpretation of epidemiological research data. This Section therefore provides a relatively in-depth overview of Measurements of Association.

Timing

This is a relatively long section but you should easily be able to work your way through the various readings and tasks within 3-4 hours. There are 2 readings including the reading in your prescribed textbook by Bonita and Beaglehole.

Learning Outcomes

* Understand measures of association
* Understand the distinction between epidemiologic measurements that measure frequency of occurrence of disease *versus* those that measure the association between risk factors and disease

Having gone through the basic biostatistical exploration of data obtained from an epidemiological study and the epidemiological measurements of frequency we now move on to cover epidemiological measurements of association. As an introduction to this we will review measurements of disease frequency and then proceed to introduce epidemiological measurements of association.

Epidemiological measurements are split into those that measure the frequency of occurrence of diseases, or health status, or determinants of diseases (risk factors for or causes of disease or good health) and those that measure the association between risk factors for disease (or good health) and the development of that disease (or the continuance of good health). In more general terms measurements of disease frequency are used to measure the frequency of occurrence (how much, how often) of health exposures and the frequency of health outcomes.

Frequencies are measured as a simple count (simply the number of things; how many), a ratio (the number of things compared to the number of other things), a proportion or percentage (the number of things within a particular group stated as a fraction of that group with the thing) and a rate (the speed at which a thing occurs). These frequencies are divided into incidence frequencies and prevalence frequencies. Incidence refers to new cases of disease or new events. Incidence therefore measures how rapidly a disease is spreading amongst a population. Prevalence refers to existing cases of disease. Prevalence therefore measures the size of the disease problem (or burden of disease) amongst a population and is often used to determine what and how much health services are required. Prevalence, since it measures currently present cases of disease would miss those with the disease who had died, or gotten better, or left the area. This is not necessarily a bad thing as cases falling into the categories of got better, died, or left would clearly not require health services. However because prevalence of the disease is mixed up with duration of living with the disease and duration of staying within the population, measures of association (see below) need to be interpreted with caution when using prevalence in measures of association.

These frequencies of occurrence of exposures and outcomes are then compared to see if and how they are associated with each other. **Associations could be absolute ones or relative ones.**

A **relative association** is when the frequency of disease in one group is compared to another group to determine if relative to the other group the first group has more or less disease. This is done by dividing the frequency in one group by the frequency in another group. Typically the frequency of disease in a group exposed to some potential cause of the disease (“the exposed”) is compared to (divided by) the frequency of disease in a group not exposed to that potential cause of the disease (“the unexposed”). This provides an answer as to how many times more likely (or less likely) a disease occurs amongst the exposed compared to the unexposed. If it is more likely (greater than one), then the exposure potentially causes the disease. If it is less likely (less than one), then the exposure potentially prevents the disease. If disease frequencies are the same in the two groups, then the relative measure would equal (or be very close to) one and hence the exposure neither causes nor prevents the disease, but rather has no effect on the disease.

An **absolute association** is when the actual (real) difference in disease frequency in one group versus another group is measured. This is done by subtracting the frequency of disease in a group from the frequency of disease in another group. Typically the frequency of disease in those unexposed to some potential cause of the disease (“the unexposed”) is subtracted from the frequency of disease in those exposed to some potential cause of the disease (“the exposed”). If the frequency of disease is greater amongst the exposed than the unexposed then that amount more is the extra amount of disease which could be due to (caused by) the exposure. It is the absolute extra amount of disease which might be caused by the exposure. Similarly if the frequency of disease is less amongst the exposed than the unexposed then that amount less is the amount of disease which could be prevented due to the exposure. It is the absolute amount of disease which might be prevented by the exposure. If the frequency of disease is the same (or very similar) amongst the exposed and the unexposed then the exposure has no effect on the disease. The difference between the exposed and the unexposed would then be zero (as the frequency is the same). Hence if the absolute measure (the difference) is zero (or close to zero) then this indicates no association between the exposure and the disease.

Both the above absolute and relative associations compare the exposed group to the unexposed group considering only one exposure. However amongst groups there are usually more than one exposure which could cause (or prevent) disease. These other exposures could either **confound** the effect of the exposure being studied (the effect of the exposure is not purely identified but rather mixed in with the effect of some other exposure) or they could **interact** with the exposure by having a greater causative effect (or greater preventative effect) when the other exposure is present. To then unravel the true pure effect of the exposure when it is being confounded, one would have to adjust for the confounder by doing stratified analysis (see later section for details). This then gives the **adjusted measure of association**. The previous measure of association mentioned in the paragraphs above are then referred to as **crude measures of association** as they do not consider the possible effect of confounders. The interaction of one exposure with another by either increasing or decreasing the effect of the exposure is however not something that can be unraveled as it is a real extra effect that happens when the two exposures are experienced together, rather than the effect of one exposure mixed up with another as in confounding. So this interaction between two exposures actually changes the effect of each exposure (making its effect either greater or lesser) and hence this is called effect modification. Since this **effect modification (or interaction)** is real it should simply be reported as being present (see later sessions for details on confounders and effect modifiers).

Another problem which could occur when assessing an effect of an exposure by comparing an exposed group to an unexposed group is that the groups might not be similar. For example the ages in the groups could be different with one group having more elderly people and the other group having more young people. This makes the comparison awkward and unfair. This could easily be avoided by always comparing similar groups. However sometimes it is not possible to get completely similar groups to compare and then one could address the problem in two ways. One could split each of the groups into smaller sub-groups so that the sub-groups of the exposed and the unexposed could then be compared as these sub-groups are more likely to be similar to each other. For example if the sub-group of only young people amongst the exposed group were compared to the sub-group of only young people amongst the unexposed group, then the comparison would be fair. Similarly if one compared only the sub-groups of elderly people amongst the exposed and unexposed groups. This is referred to as **category specific associations,** with in this example us using age categories. Alternatively one could standardize the two groups (exposed group and unexposed group) and then compare them. This **standardisation i**s very similar to adjusting for confounding as it shows the pure effect of the exposure. However in this case the exposure is not mixed up with another exposure (as in confounding) but its effect is rather not clearly shown because the groups are not similar (see later sessions for details instructions on how and when to implement standardisation).

Thus far we have only considered associations between an exposed group and an unexposed group. However in public health we are more concerned about the entire population rather than just groups within the entire population. Hence it is desirable to measure associations between the whole population and the exposed group to assess the degree of effect the exposure would have on the whole population. While this could be done for both relative associations and for absolute associations, it is mostly only done for absolute associations as absolute associations shows one the real size of the problem and hence one could determine the real size of the problem in the population as a whole. This is called **population attributable associations.**

It is hoped that the overview of epidemiological measures provided above has put into context the epidemiologic measurements of association. **Now go** **through the reading by Hennekens and Buring (measures of association) which provides a detailed overview of measurements of Association from page 73 – 96.**

**Reading**

Hennekens, C. H. and Buring, J. E. (1987) *Epidemiology in Medicine*. Boston: Lippincott Williams & Wilkins. Pages 73 – 98.

Having gone through the above reading and you should be able to apply your knowledge to tackle the tasks shown below.

Feedback on the tasks is shown at the end of this session.

Tasks

**Measurements of Association**

1. The strength of an association is one of the criteria for evaluating the cause and effect relationship between an exposure and outcome. Which of the following is a measure of the strength of association? (Choose one best answer).
   1. incidence rate among the exposed
   2. cumulative incidence among the exposed
   3. the ratio of odds of exposure among cases to the odds of exposure among the non-cases
   4. odds of disease among exposed relative to the prevalence of exposure in the source population
   5. none of the above
2. Incidence a disease is often referred to as direct measures of risk. Can the incidence be calculated from case-control studies? Briefly explain in 1-2 sentences why they can or cannot be calculated.
3. For each of the following epidemiological measures, indicate whether it is a rate, a proportion or that it is neither a rate nor a proportion. Circle the best answer.

|  |  |  |  |
| --- | --- | --- | --- |
| a. Population attributable risk | RATE | PROPORTION | NEITHER |
| b. Incidence density (ID) | RATE | PROPORTION | NEITHER |
| c. Prevalence | RATE | PROPORTION | NEITHER |
| d. Relative risk | RATE | PROPORTION | NEITHER |

|  |  |  |
| --- | --- | --- |
|  | | |
|  |  | 1. Indicate true or false next to each of the following. |
|  |  | 1. A risk ratio measure and a correlation coefficient are both measures of association. |
|  |  | 1. A population attributable risk proportion depends on the prevalence of exposure and is not directly related to the strength of an association. |

1. Ischaemic heart disease (IHD) is a leading cause of morbidity amongst people 65 years and older and is estimated to affect between 16 and 26% of people in this age group. In a recent study residents aged 60 to 85 years in the town of Polokwane were asked to participate in a study to determine whether cigarette smoking was related to IHD. At a baseline examination, participants were asked to report their lifetime smoking habits. After 5 years, participants had an examination to determine whether they had developed IHD. The following table presents the number of cases of IHD measured at the follow-up examination among the 1232 participants who did not have IHD at the baseline examination:

|  |  |  |
| --- | --- | --- |
| Smoking status | N | Cases of IHD |
| Never smokers | 368 | 26 |
| Ever smokers | 864 | 79 |

1. Create a 2 x 2 table where one axis is smoking status and the other is IHD.
2. Calculate the 5-year cumulative incidence of IHD in both ever smokers, and in never smokers.
3. Calculate the cumulative incidence ratio comparing the incidence of IHD in ever smokers with that in never smokers.
4. Assuming causality, what is the proportion of cases of IHD that could have been prevented in the population of aged 60-85 in Polokwane if the smokers had never smoked?
5. The evidence supporting obesity as a risk factor for colon cancer remains inconclusive, especially among women. A recent study (*Am J Epidemiol* 1999;150:390-398) reported the association between obesity (measured at baseline) and colon cancer morbidity as determined from a review of medical records and death certificates in a nationally representative cohort of men and women in the USA aged 25-74 years who participated in the First National Health and Nutrition Examination Survey from 1971 to 1975 and were subsequently followed up through 1992. The following table is from this study for men and women combined.

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline body  mass index\* | Number of incident cases of colon cancer | Person-years  of follow up | Crude incidence rate/100,000 PY |
| <22 | 28 | 53,475 |  |
| 22 - <24 | 41 | 38,919 |  |
| 24 - <26 | 36 | 36,610 |  |
| 26 - <28 | 40 | 32,635 |  |
| 28 - <30 | 35 | 21,122 |  |
| 30+ | 42 | 34,904 |  |

\* kg body weight per height in meters squared

1. Complete the table by calculating the crude body mass index-specific incidence rates of colon cancer.
2. Calculate the relative risk (RR) of colon cancer associated with a BMI of 28-<30. Use the lowest BMI category as referent. In one sentence interpret your answer.
3. Calculate the attributable risk of those in the 28-<30 BMI category. In one sentence interpret your answer.
4. Calculate the attributable risk percent of those in the 28-<30 BMI category. In one sentence interpret your answer.
5. Calculate the incidence rate of colon cancer for the whole study population.
6. Calculate the population attributable risk of those in the 28-<30 BMI category. In one sentence interpret your answer.
7. Calculate the population attributable risk percent of those in the 28-<30 BMI category. In one sentence interpret your answer.
8. Calculate the relative risk (RR) of colon cancer associated with a BMI of 30+. Use the lowest BMI category as referent. In one sentence interpret your answer.
9. Calculate the attributable risk percent of those in the 30+ BMI category. In one sentence interpret your answer.
10. Calculate the population attributable risk percent of those in the 30+ BMI category. In one sentence interpret your answer.
11. Compare the relative risk, the attributable risk and the population attributable risk amongst those in the 28-<30 BMI category with those in the 30+ BMI category. Interpret your findings on these comparisons.

**Feedback on Tasks**

Question 1

The “ratio of odds of exposure among cases to odds of exposure among non-cases” is the odds ratio, which is a measure of association.

Question 2

Incidence cannot be estimated from case-control studies without additional information. In the case-control design selection of subjects is based on disease status, so the number of cases is decided on by the researcher. Similarly the number of controls is determined by the researcher. Therefore the incidence cannot be calculated as the researcher decided on the numbers of cases and the numbers of people to compare them to, hence there is no population from which to determine an incidence. However if the investigator has access to ALL cases AND knows the size of the population from which they arise s/he can estimate incidence, but knowledge of the population size is not available from the case-control design and would have to be obtained in addition in order to calculate the incidence. This will become clearer after you have worked through the sessions on cohort studies and case control studies.

Question 3

* 1. Population attributable risk (PAR) is a PROPORTION.
  2. Incidence density (ID) is a RATE.
  3. Prevalence is a PROPORTION.
  4. Relative risk is NEITHER a rate nor a proportion. It is a ratio.

Question 4

Indicate true or false next to each of the following. (2 pt each)

TRUE – A risk ratio measure and a correlation coefficient are both measures of association.

FALSE – A population attributable risk proportion depends on the prevalence of exposure and is ALSO directly related to the strength of an association.

Question 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | a. Cigarette smoking status | | |
|  |  | Ever smokers | Never smokers | Total |
| Case | ARM cases | 79 | 26 | 105 |
| Status | Non-cases | 785 | 342 | 1127 |
|  | Total | 864 | 368 | 1232 |

b. CI in ever smokers = # new cases / population at risk = 79/864 = 0.091 in 5 years  
CI in never smokers = # new cases / population at risk = 26/368 = 0.071 in 5 years

c Cumulative incidence ratio (CIR) = CI in ever smokers / CI in never smokers  
= (79/864) / (26/368) = 1.29

d. PAR = (overall incidence – incidence in never smokers) / overall incidence of ARM  
= (0.0852 – 0.0707) / 0.0852 = 17%

Question 6

a.

|  |  |  |  |
| --- | --- | --- | --- |
| Baseline body mass index\* | Number of incident cases of colon cancer | Person-years of follow up | Incidence rate/100,000 PY |
| <22 | 28 | 53,475 | 52.4 |
| 22 - <24 | 41 | 38,919 | 105.3 |
| 24 - <26 | 36 | 36,610 | 98.3 |
| 26 - <28 | 40 | 32,635 | 122.6 |
| 28 - <30 | 35 | 21,122 | 165.7 |
| 30+ | 42 | 34,904 | 120.3 |

\* kg body weight per height in meters squared

b. RR of colon cancer for BMI 28-<30 kg/m2 vs. lowest = 165.7/52.4 = 3.16

The RR of 3.16 means that those with a BMI between 28 - <30 are 3.16 times more likely to develop colon cancer that those with a BMI <22.

**Send your answers on questions 6c to 6k for us to check for you.**

Unit 4 – Session 4: Hypothesis testing

## Introduction

This session presents the fundamental statistical concept of **hypothesis testing**, and the logic underlying a large family of procedures (**tests**) which allow us to use a random sample drawn from a population to decide if a given hypothesis about a population parameter is to be accepted or rejected. It also introduces the appropriate terminology and presents some examples of widely used statistical tests.

## Learning outcomes

* Understand the concept of hypothesis testing and p-value;
* Understand the concept of Type I and Type II errors and power of a test;
* Understand the difference between parametric and non parametric tests;
* Understand the relationship between hypothesis testing and confidence intervals.

## Contents

1. Hypothesis Testing
2. Errors and power of a test
3. Parametric and non-parametric tests

Timing

This is a relatively short section and you should easily be able to work your way through the various readings and tasks within 1 hour.

1 Hypothesis Testing

Hypothesis testing or significance testing is a set of methods for testing a claim or hypothesis about a parameter in a population, using data measured in a sample.

The conceptual steps of hypothesis testing are the following:

1. **State a hypothesis** about an unknown population parameter (e.g. “the mean of the variable X in the population is 0”);
2. **Drawn a random sample** from the population and calculate the value of an appropriate **statistic**;
3. Calculate the probability to observe that value in the hypothesis about the population parameter is true (the “**p−value**”);
4. Compare the probability calculated in the previous step with a pre-determined threshold probability (the **significance level**, usually indicate with α);
5. **Make a decision** about the initial hypothesis: if p−value ≤ α the hypothesis is rejected, if p−value > α we say that we don’t have enough evidence to reject the hypothesis.

Hypothesis testing is essentially a refutation method, i.e. it is used to “test” if a hypothesis on the population is compatible with what we observe in the sample.

Because sampling always introduces some form of uncertainty into our procedure, we accept that our decision is subject to a certain margin of error, which is defined by the confidence level α: the smaller α, the smaller is the risk to incorrectly reject the hypothesis.

However, in real-life, researchers are usually interested not in saying that something is wrong, but rather in “proving” (or at least supporting) a research hypothesis. Unfortunately, in hypothesis testing we can never “prove” that an hypothesis is true, but only “disprove” it, i.e. conclude that - a the level of significance α - it is unlikely that it stands.

This has an important practical consequence in the way we state our hypotheses. In fact, rather than stating an hypothesis, in most cases we are bound to state two hypotheses: the hypothesis we are interested to “prove” and another hypothesis which is “complementary” to it, i.e. such that disproving it is equivalent to “proving” the other.

In Statistics these two complementary hypotheses are called:

* H0 or Null Hypothesis (which is the hypothesis we want to disprove);
* HA or Alternative Hypothesis (which is the research hypothesis we aim to support).

The reading by Matchin et al. presents in greater detail these concepts and explains how in practice a statistical test is performed.

**Reading**

Machin, D., Campbell, M. J. & Walters, S.J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons:100-108.

2 Errors and power of a test

In Statistics the notion of statistical error is an integral part of hypothesis testing, to the point that when we presents the results of any statistical test, we always accompany them with a sentence like ”... at the level of significance α”, explicitly recognising that we are taking into account the possibility that our decision is not the correct one.

The two complementary types of error which can be committed in statistical testing (rejecting the null hypothesis when it is true, or accepting the alternative hypothesis when it is false) are called **type I** and **type II** error, respectively.

The next reading explains how these two errors relate to each other and also defines an important characteristic which differentiates statistical tests, i.e. their **power**:

**Reading**

Machin, D., Campbell, M. J. & Walters, S.J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 108-110.

Finally, the last reading of this session clarifies that the concepts and procedures of hypothesis testing and calculation of confidence intervals are strictly related:

**Reading**

Machin, D., Campbell, M. J. & Walters, S.J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 110-113.

3 Parametric and non-parametric tests

Many different statistical tests are available to test many different hypotheses. And, for each hypothesis to be tested, various tests are available depending on the characteristics of the population. Each test, in fact, is applicable only when specified assumptions regarding the population are met, because otherwise its results are misleading.

Among others, many tests make assumption about the distribution of the variable of interest in the population (if the variable is continuous they often assume that it is normally distributed): these tests are called **parametric tests**. Conversely, test which do not make any assumption about the distribution of the variables involved.

Parametric tests, therefore, have stricter requirements than non-parametric tests (i.e. if a parametric test is applicable in a given case, a non parametric one is also applicable, but the opposite is not true). However, they usually have the advantage of a **greater power**. This is the reason why, when the assumptions for parametric tests are met, it is preferable to choose from this class rather than using a non-parametric procedure.

**Task 1**

Define the following terms related to statistical testing:

* Null & Alternative hypothesis;
* Type I & Type II error;
* Significance Test;
* p-value;
* level of significance;
* power of a test.

**Task Feedback**

The definitions are in the reading by Machin *et al*. above.

**Task 2**

Some researchers were interested in comparing the average weight between groups of children living in different areas of a city. To this end they drew a random sample of children in each area and calculated the confidence interval for the difference between their average weights. The results are summarised in the table below:

|  |  |  |
| --- | --- | --- |
|  | Comparison | 95% Confidence Interval [Kg] |
| 1 | Area A vs. Area B | [1 ; 2] |
| 2 | Area C vs. Area D | [-0.1; 3] |
| 3 | Area E vs. Area F | [-2 ; -1] |

What can we conclude (at the level of significance α=0.05) regarding the existence of a “true” difference between the average weights for each pair of areas?

**Task Feedback**

1. Because the 95% confidence interval does not include 0, we have enough statistical evidence to conclude that the average weight of children in area A is higher than the average weight of children in area B;
2. Because the 95% confidence interval includes 0, we do not have enough statistical evidence to conclude that the average weight of children in area C is different from the average weight of children in area D;
3. Because the 95% confidence interval does not include 0, we have enough statistical evidence to conclude that the average weight of children in area E is lower than the average weight of children in area F;

# Unit 4 - Session 5: Relationships between two variables

# Introduction

Analytical epidemiology is about searching for causes of health and disease, which is a long and complex process, involving various steps. The first of these steps is assessing if the hypothesised cause (*exposure* in epidemiological terms) and the hypothesised effect (disease occurrence or any other health outcome) are in some way associated in the individuals of the population of interest.

An exposure and an outcome are said to be associated if the knowledge of the value of the exposure for an individual “says something” about the value of the outcome.

Let’s consider for example age and height of children. If someone asked us the question: “Is Paul taller than James?”, and we don’t know Paul and James, we could not give a meaningful answer. But, if the question became “Is Paul, 5 year old, taller than James, 2 year old?”, our answer would be probably “yes”, even in absence of direct knowledge of Paul and James. This is because the age of children “says something” about their age.

In statistical terms, we say that age and height in children are *associated* (the term *correlated* is also used when the variables are numerical or ordinal). Notice that this association does not make us “sure” that *every* child aged 5 is taller than *every* child aged 2, but it tells us that most of the time this is true (i.e. this is true *in average*).

This session deals with statistical methods to establish whether an association exists between two variables (and exposure and an outcome) in a population, and, if an association exists, how to use this information to “predict” the outcome from the knowledge of the exposure.

More precisely, we will look at methods to decide if an association exists (and to make prediction) in a population only looking at a sample, taking into account sampling variability.

## Timing

This session should take you about 1 hour. Work your way through the various readings and tasks.

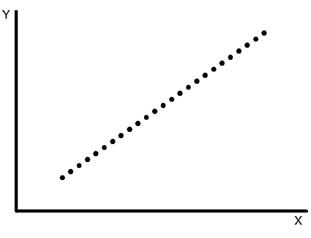
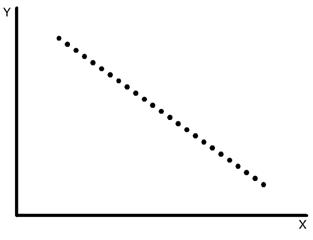
## Learning outcomes

* Assess the association between continuous variables with the Pearson’s correlation coefficient;
* Assess the association between ordinal variables with the Spearman’s correlation coefficient;
* Assess the association between nominal variables with the χ2 test and the contingency coefficient;

### 1 Relationships between continuous variables

When we want to assess if a correlation exists between two continuous variables, the most common method we use is the calculation of the *Pearson’s Correlation Coefficient*, usually indicated with the Greek letter ρ when referred to the whole population and with the Latin letter *r* when referred to a sample (remember that usually we use Greek letter for population parameters, and Latin letters for sample statistics: so *r* is *statistics* which we use to estimate ρ).

The Pearson’s Correlation Coefficient is a number between -1 and 1 which actually measures a special type of relationship between numerical variables: a *linear relationship*. When two variables X and Y are in a *perfect linear relationship*, a scatter plot of X as a function of Y looks like:

 or 

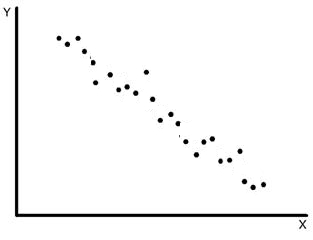
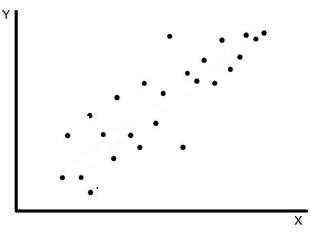
r = -1

r = +1

If we know the value of X we also know the value of Y (this is the meaning of *perfect* correlation: if we know X we can calculate *exactly* the values of Y, and vice-versa), and if X doubles its values, also Y doubles its values (this is the meaning of the word *linear*).

In cases of perfect linear correlation, r assumes the value of +1 (perfect *positive* correlation, figure on the left), or -1 (perfect *negative* correlation, figure on the right).

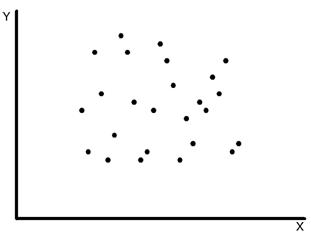
When things are less clear, i.e. when the values of X gives us some information about Y, but does not allow us to predict exactly its value, r assumes intermediate values, as in the examples below, in which there is a general tendency of the data points to follow a straight line, but they don’t lie exactly on that line:

r = +0.5

r = -0.7

The larger (in absolute value) *r* is, the more the data points follow a straight line. A Pearson’s coefficient equal to 0 means that the points don’t follow a straight line at all (no linear correlation), as in the example below.



r = +0

The reading by Muijs gives a precise definition and a way of calculating r. It also explains that we can use a statistical test to assess whether or not the value of *r* calculated in the sample is statistically significant, i.e. if we can expect that the value of the population parameter ρ is different from 0. Usually we don’t do the test manually, but rather use statistical software. It is important, however, to be able to interpret the results: a non-significant p-value (usually p> 0.05) means that we don’t have enough evidence to say that the two variables are associated in the population (i.e. it is possible that the value of ρ is 0 in the population).

**Reading**

Muijs, D. (2004). *Doing Quantitative Research in Education with SPSS*. London: Sage: 142-147.

### 2 Relationships between ordinal variables

The Pearson’s correlation coefficient can be meaningfully calculated only when both variables are numerical. When one (or both) variables are measured on an ordinal scale, a different measure of association is usually calculated: the *Spearman’s correlation coefficient* (ρs as a population parameter, and *r*s as statistics).

The interpretation of the Spearman’s correlation coefficient is similar to the Spearman’s coefficient, that is it ranges from -1 (perfect negative correlation) to +1 (perfect positive correlation), with 0 indicating no correlation. In this case, however, high values of the coefficients can be reached even when the correlation is not linear, provided that the relationship between the two variables is monotonic (i.e. the outcome either increases or decreases *constantly* with the values of the exposure).

The reading by Muijs explains how rs is calculated, its’ relationship with the Spearman’s *r,* and how the hypothesis that a correlation exists in the population (i.e. ρs is **not** equal to 0) can be tested formally.

**Reading**

Muijs, D (2004). *Doing Quantitative Research in Education with SPSS*. London: Sage: 151-153.

### 3 Relationships between nominal variables

Because the values of variables measured in a nominal scale cannot be ordered, there is no meaning in thinking about the *direction* (negative or positive) of their association, as we did for numerical and ordinal variables. Therefore, the measures of association used for nominal variables usually range from 0 (no association) to 1 (or values close to 1) in case of perfect association, i.e. when the value of one variable allows us to predict perfectly the values of the other.

One of these measures is the *contingency coefficient (c)*, which is derived from another well-known statistics, the χ2 statistics. An explanation of the χ2 statistics, its calculation and its application to measure the strength of the association between two nominal variables is given in the reading by Blaikie.

**Reading**

Blaikie, N. (2003). *Analyzing Quantitative Data: From Description to Explanation*. London: Sage: 97-99.

**Task 1 – Measuring correlations**

1. What does the Pearson’s correlation coefficient measure?
2. Can you use the Spearman’s correlation coefficient to measure the correlation between two numerical continuous variables?
3. Interpret the following data, referring to the relationship between two generic variables X and Y:

r=0.9

r=0.4 , 95%CI [-0.1; 0.6]

r=-0.3

rs=1

rs=0

c=0.9

**Task feedback**

1. It measures the degree of linear relationship between two numerical variables, i.e. the extent to which the knowledge of the value of a variable allow us to predict the value of the other variable.
2. Yes. The Spearman’s correlation coefficient can be calculated for any pair of variables, provided they are measured **at least** in an ordinal scale. Numerical variables can be always seen as a special case of ordinal variables. However, when both variables are continuous, it is more common to calculate the Pearson’s correlation coefficient, which has better statistical properties.
3. A positive Pearson’s correlation coefficient (*r*=0.6 or *r*=0.4) suggests a positive linear correlation between X and Y, i.e. the value of one variable goes up when the value of the other goes up. Because the values are <1 the correlation is imperfect, in a scatter plot the points would be scattered around a straight line. The dispersion around the line would be greater when *r*=0.4 (weak correlation) than *r*=0.9 (strong correlation).

The confidence interval attached to the second value of *r* includes 0: it means that there is a weak correlation in the sample (*r*>0), but we are not sure (at the 95% level of confidence) that a correlation exists in the population from which the sample has been drawn. In other words, we are not sure that ρ≠0.

The interpretation of the Spearman’s correlation coefficients *r*s is the same as the interpretation of *r*: higher values mean stronger correlation (in either direction).

A contingency coefficient c=0.9 means that the two categorical variables are quite strictly associated, i.e. there is a relationship between them (we don’t use the term “correlated” in this case, but the meaning is similar).

# Unit 4 – Session 6: Comparing two groups

## Introduction

The conceptual basis of analytical epidemiology lies in the possibility of comparing different groups of people in order to find associations between health outcomes and risk factors.

This session presents a series of statistical methods which can be used to this end in the special case in which only two groups are compared. (We will see in session 4.7 an extension of some of these methods allowing us to assess differences between more than two groups.)

Timing

This session should take you about an hour to go through.

## Learning outcomes

* Test hypotheses about difference between populations;
* Calculate confidence intervals for absolute measures of association;
* Calculate confidence intervals for relative measures of association.

1 Methods for comparing two populations

We learnt in the previous session how hypothesis testing and calculation of confidence intervals are two procedures strictly interlinked.

The readings in this section show how the two methods can be applied to compare two populations, more precisely to answer two specific questions:

1. Given two populations P1 and P2 and a continuous variable X, is the mean of X different between P1 and P2 (hypothesis testing), and how much (confidence intervals)?
2. Given two populations P1 and P2 and a binary variable Y (with assumes values Y=y1 or Y=y2), is the proportion of individuals with Y=y1 different between P1 and P2 (hypothesis testing), and how by much (confidence intervals)?

The first reading by Dawson & Trapp presents two widely used tests to answer the first question: **the two sample t-test** and the **Wilcoxon sum rank test**. The difference between the two test is that the first one is a parametric test (which requires that X is normally distributed), while the second does not make any assumption about the distribution of X (non parametric).

The reading also presents the parallel formulae to calculate confidence interval for differences between the means of X in the two populations, which allows us not only to decide if the mean are statistically different, but also to quantify their difference.

**Reading**

Dawson, B., &Trapp, R. G. (2004).*Basic & clinical biostatistics* (4th ed.). New York: McGraw-Hill: 136-146.

The second reading by the same authors presents another set of statistical tests and formulae for the calculation of confidence intervals, this time with reference to binary variables. The tests presented in the reading are the **Z-test** and the **χ2 test for 2x2 tables** (which are alternative procedures to obtain the same results, and are only applicable in relatively large samples), and the **Fisher’s exact test** (which is applicable when the samples are small). Because the variables are binary and not continuous, the summary statistics of interest are not means but proportions, and, therefore, the procedures presented below answer the question whether the proportion of Y in the two populations are similar or different, and by how much they differ.

The **χ2 test** canbe easily extended to compare proportions in more than two groups or, more in general, to compare nominal variables in more than two groups. This will be shown in Session 4.7.

**Reading**

Dawson, B., &Trapp, R. G. (2004).*Basic & clinical biostatistics* (4th ed.). New York: McGraw-Hill:146-154.

**Task 1**

Consider the 2x2 table below, showing the number of individuals with and without lung diseases in a sample of miners and in a sample of non-miners.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lung Disease | No Lung Disease |  |
| Miners | 55 | 16 | 71 |
| Non Miners | 22 | 43 | 65 |
|  | 77 | 59 | 136 |

1. Test the hypothesis that the proportion of lung disease among miners is different from the proportion of lung disease among non-miners;
2. Interpret the results of the test and the p-value.

**Task feedback**

The variable “Lung Disease” is a binary variable (it can only assume two values: YES or NO), the groups we are comparing are independent and the sample size is large enough. Therefore we can test the hypothesis using alternatively the Z-test and the χ2 test for 2x2 tables (with identical results).

Using the χ2 test (and the “shortcut formula” at from the reading by Dawson & Trapp):

Because χ2 is greater that the critical value form the χ2 table (χ2α=5%=3.84), we can conclude that there is a statistically significant difference between the proportion of lung cancer among miners and non- miners.

Because χ2 > χ2α=5%, the p-value will be <0.05. we can interpret this saying that, if the hypothesis of equal proportion were true, the probability of finding a value of χ2 equal to 26.28 or greater than that is less than 5%. In practical terms it means that we can be confident that the proportions of lung cancer are truly different among miners and non- miners

# Unit 4 - Session 7: Comparing more than two Groups

# Introduction

In the previous section 4.6, we learned how to apply hypothesis testing to check whether or not two populations are similar in some specific characteristics.

In particular, we learned how to use sample statistics to decide (with a certain margin of error) whether the mean values of a variable in the populations from which the samples are drawn are the same or different. Also we learned how to compare proportion across two populations.

The same question can be generalised to the comparison of more than two groups, and becomes: given a set of samples drawn from three or more populations, are the mean values of a variable (statistically) different across the populations? Or, similarly, are the proportion of subject with a given characteristics different across the populations?

This section presents some basic techniques to answer this question.

Timing

This is a relatively short section and you should easily be able to work your way through the readings within 2 hours.

## Learning outcomes

* Understand the logic and interpret the results of the Analysis of Variance to compare population means;
* Understand the logic and interpret the results of the Kruskall-Wallis test to compare values of variables not normally distributed across populations.

### 1 One-way analysis of variance and Kruskal-Wallis test

The one-way analysis of variance (ANOVA) test is used to assess if the mean values of a variable are different across three or more groups. The reading by Blaikie explains the logic behind ANOVA and how the test result (usually calculated by statistical software) should be interpreted. It also explains the assumptions about the distribution of the variable in the two groups which must be met in order to ensure that the results are meaningful.

When these assumptions are not met, an alternative non-parametric test widely used is the Kruskal-Wallis test, which is an extension of the Wilcoxon rank-sum test applicable to compare two groups. The Kruskal-Wallis, being non-parametric, does not make any assumption regarding the distribution of the variables, and can also be applied with categorical variables measured in an ordinal scale (but not with variables measured in a nominal scale). The Kruskal-Wallis test is rarely performed manually.

**Reading**

Blaikie, N. (2003). *Analyzing Quantitative Data: From Description to Explanation*. London: Sage: 201-204.

### 2 Chi-squared test for nominal variables

When more than two groups need to be compared but the variables of interest is measured in a nominal scale, the most common alternative to the ANOVA test is the χ2 test.

The χ2 test is widely used (sometimes misused: as all other tests it has assumptions and when these are not met the results of the test are not interpretable!) for many purposes in statistics. In its more general form (see the reading of Petrie and Sabin) the χ2 tests the null hypothesis that two categorical variables (X and Y) are not associated.

If X has only two category and Y has r categories (r>2), we can also say that the χ2 tests the null hypothesis that the proportion of individual with characteristic X is the same across the different groups represented by the values of Y. This is the case we are considering here (testing differences between proportions across more than 3 groups).

A description of the χ2 test and a manual procedure to perform it is presented in the reading by Petrie and Sabin. Most commonly the χ2 test is calculated using statistical software.

**Reading**

Petrie, A. & Sabin, C. (2005). *Medical Statistics at a Glance* (2nd Ed.). Chichester: John Wiley & Sons: 64-65.