ETHIOPIAN PUBLIC HEALTH ASSOCIATION

LEADERSHIP IN STRATEGIC INFORMATION TRAINING PROGRAM

MODULE 2

(*ANALYTIC EPIDEMIOLOGY,* INFERENTIAL STATISTICS)

PARTICIPANT MANUAL

June, 2014
Addis Ababa, Ethiopia

Approval of the Training Material

The Federal Ministry of health of Ethiopia has been working towards standardization and institutionalization of in-service (IST) trainings at national level. As part of this initiative the ministry developed a national in-service training directive and implementation guide for the health sector. The directive requires all in-service training materials fulfill the standards set in the implementation Guide. Accordingly, the ministry reviews and approves existing training materials based on the IST standardization checklist annexed on the IST implementation guide.

All in-service training materials shall to be reviewed and approved by the ministry accordingly; as part of the national IST standardization process, this **Leadership in Strategic Information** IST material has been reviewed based on the standardization checklist and approved by the ministry in January 2014.

Dr Wendemagegn Enbiale Yeshaneh

Human Resources Development & Administration Directorate

Federal Ministry of Health, Ethiopia

Acknowledgment

The Ethiopian Public Health Association would like to acknowledge and pass its deep appreciation to the following professional contributors for developing LSI training Participant module.

Contributors

Addis Ababa University, School of Public Health Staff

Dr. Fikre E/Silassie, Dr. MEsfinAdisse and Dr. AbabiZergaw, Prof. MisganawFantahun , Dr. NegussieDeyessa and Dr. AlemayehuWorku

Gondar University, Institute of Public Health Staff

Dr. GashawAndarge, Dr. BerihunMegabiw, Prof. YigzawKebede, Mr. Tadesse Awoke, Dr. AbebawGebeyehu and Mr. Solomon Meseret

Jimma University, Institute of Health Science Research Staff

Dr. FesehayeAlemseged, Mr. FasilTesema, Mr. YibeltalKiflie, Mr. NegalignBerhanu and Dr. MirkuzieWoldie

EPHA Staff

Mr. BekeleBelayihun

Independent Consultant

Dr. HailuYeneneh

Module 2: Analytic epidemiology and biostastics, SPSS and Epi-Info or EpiData statistical software

Overview

Descriptive epidemiology and biostatistics has been discussed in the previous phase of of module 1. Do you remember the purpose of descriptive epidemiology and biostastics? How about types of descriptive designs? Descriptive epidemiology is concerned with study of frequency and distribution of diseases and other health related events in a population. The types of descriptive designs include case-study, case-series, cross-sectional and ecological designs. Descriptive designs also help to generate hypothesis about causal factors which are tested using analytic designs. Each session begins with brief introduction, and a list of learning objectives that you should be able to achieve when the session is completed. Following the objectives, session contents are listed, and the main reference material is then mentioned, whereas list of all references is indicated at the end of the session. Then, the session contents are discussed. Mean while, questions are raised and exercises given for practice in order to help you understand key concepts. Towards the end of each session, summary of the main points is presented. Inferencial biostatistics deals with the process of making inferences regarding population characteristics through the application of estimation and hypothesis theories. In the sessions and sub-sessions that follow, the concepts of inference statistical. In addition, correlation, linear regression and logistic regression concepts and methods are included.

Goal of the module

To train trainees with the capacity of public health practiciner to use Epidemiology and inferencialbiostastics with the application of epidemiological and stastical software in health related information analysis for applying making admission.

General Objectives of the Module

To trainpublic health sector workers with the capacity of analytical epidemiology and biostatistics with the application of statistical software for making right decision in public health sector leaders in the nation.

After completing this module, participants will be able to;

- ✓ Apply epidemiological methods in public health practice
- ✓ Apply statistical thequencs and managing and analysis health and health realteddata
- ✓ Use data management and statistical software

Contents of the Module

The Module is organized in two parts. The first part is analytic epidemiology and the second part is inferencialto biostatistics. The two parts are given for a given period of two weeks, each lasting one week.

Part 1: ANALYTIC EPIDEMIOLOGY

1. Introduction to analytic epidemiology
2. Analytic epidemiologic studies
   - Analytic ecological and cross-sectional studies
   - Case-control studies
   - Cohort studies
3. Measures of association and impact
4. Chance, Bias and confounding
5. Establishing causation in epidemiology
6. Validity and precision

Part2: Inferential statistics

1. Estimation and hypothesis testing
2. Correlation and Regression

Statistical software

- SPSS statistical software
- EpiInfo or EpiData epidemiological software

Acronym/ Abbreviations

AIDS	Acquired Immunodeficiency Virus
AR	Attributable Risk
ART	Antiretroviral Therapy
CHD	Coronary Heart Disease
CI	Confidence Interval
EPA	Environmental Protection Agency
HIV	Human Immunodeficiency Virus
HR	Hazard Ratio
LSI	Leadership Strategic Information
MDR TB	Multidrug Resistance Tuberculosis
MH	Mantel Haenszel
MLE	Maximum Likelihood Estimation
MoA	Measure of Association
MTCT	Mother to Child Transmission
MVE	Minimum Variance Estimate
NA	Not Applicable
OR	Odds Ratio
PAR	Population Attributable Risk
RD	Risk Difference
RR	Relative Risk
SD	Standard Deviation
SE	Standard Error
SPSS	Statistical Package for Social Science
SSE	Sum of Squares Error
SSR	Sum of Squares Regression
TB	Tuberculosis
VLCD	Very-Low-Calorie Diet

Part 1:ANALYTICAL EPIDEMIOLOGY

1. Introduction to analytic epidemiology
2. Analytic epidemiologic studies
    - Analytic ecological and cross-sectional studies
    - Case-control studies
    - Cohort studies
3. Measures of association and impact
4. Chance, Bias and confounding
5. Establishing causation in epidemiology
6. Validity and precision

Session 1: Introduction to Analytic Epidemiology

Session overview

This session gives overview of the purposes of analytic epidemiology, the common features of analytic studies and the types of analytic studies. This session will enable you to answer the questions: why are analytic designs needed? How do they work and what are their types?

Learning Objectives

At the end of this session, you should be able to:

- Describe the purposes of analytic epidemiologic designs
- Explain analytic concepts in identifying determinants
- Mention the types of analytic designs

1.1. Purposes of analytic epidemiology

Warm-up Question: What is the concern of analytic epidemiology?

The concern of analytic epidemiology is, in general, the identification of determinants of heath related outcomes and its application to the prevention and control of diseases.

Identifying determinants is mandatory for selection and implementation of effective health interventions. Some of the specific purposes of analytical epidemiology include:

- To identify causal factors in the prevention and control of a disease
- To measure degree of contribution of risk factors for prioritizing interventions
- To determine causal and contributing factors during epidemic investigation and control
- To establish prognostic factors that would help in clinical management of patients
- To find out reasons for success or failure of programs during program evaluation
- To detect factors related with better health for health promotion

Determinants are identified with the help of analytic epidemiologic studies. The way how analytic studies help in identifying determinants is discussed in the next session.

1.2. Basic methods in analytic studies

A determinant is a factor which affects occurrence of a heath outcome. The distribution of the health outcome in the presence of the determinant factor, therefore, differs from the distribution in its absence, and vice versa. This condition is referred as association between the determinant and outcome. Identification of determinant involves detection of association between the hypothesized determinant or exposure and the health outcome. The association between health outcome and an exposure is assessed using analytical studies.

Warm-up question: How do you think an association between exposure and a heath outcome can be assessed?

Making comparative analysis is the basic tool for identifying association. Analytic studies compare different values of the exposure or the outcome with regard to occurrence of the outcome or the exposure respectively.Association is said to be present between the outcome and exposure variable when the distribution of one differs along values of the other.

Example: In a building, 30 of 100 residents became ill with gastroenteritis. What made these 30 individuals ill?

- The possible hypothesized causes for the illness could be contaminated common water source or contaminated fish in cafeteria commonly used by the residents
- In order to identify the real cause of the illness, making comparison is the key step. Comparing the 30 sick to the 70 healthy ones with regard to proportion exposed to hypothesized causes, is one method. Another method is to compare those exposed with non-exposed with regard to proportion who become ill.

Though detecting association is important, it is not sufficient evidence to establish determinants. Additional evidences, as discussed in the last session, are also considered in identifying a determinant.

1.3. Types of analytic epidemiologic designs

Analytic epidemiologic designs are broadly categorized into two. The two broad categories of analytic designs are:

- **Observational**– where researcher simply observes events. Their purpose is to identify possible determinants of health outcomes, commonly causes of diseases.
- **Interventional** – the researcher allocates an intervention to one or more of comparison groups. The purpose is to evaluate performance of interventions aimed at preventing a disease or improving its outcome.

In this phase of the training, we will deal with only the observational analytic designs. The interventional designs will be dealt in the third phase under the session about evaluation designs. The main types of observational analytic designs are case-control and cohort studies. But under some conditions, ecological and cross-sectional designs could also serve analytic purpose.

The specific types of analytic observational designs will be discussed in the next sessions. The sessions are arranged sequentially to proceed from the weakest to strongest type of designs.

Summary

In this session you have learned that:

- The purpose of analytic epidemiology is to identify determinants of health outcomes.
- Analytic designs help in identifying determinants through detecting association between exposure and outcome by making comparative analysis.
- Case-control and cohort studies are main types of observational analytic designs.

References

Carr S, Unwin N and Mulloli PT. An Introduction to Public Health and Epidemiology, 2nd Ed. England: Open University Press, 2007.

Bonita R. Beaglehole R and Kjellstom T. Basic Epidemiology.WHO, Geneva, 2000.

Session 2: Analytic Epidemiologic Studies: Analytic Ecological and Cross-sectional Studies

Session overview

Though analytic ecological and cross-sectional designs are the weakest in establishing determinants, they are still commonly conducted mainly due to their feasibility. One of their uses is for screening hypotheses that can be further tested using stronger analytic designs. The design of analytic ecological and cross-sectional differs from the descriptive ones. In this session you will learn briefly about their design and conduct.

Learning Objectives

At the end of this session, you should be able to:

- Describe the design of analytic ecological and cross-sectional studies
- Describe the limitations of analytic ecological and cross-sectional studies
- Recognize applications of analytic cross-sectional design

2.1. Analytic Ecological Studies

In an ecological study, data is collected on whole group of people or populations rather than individuals. They can have descriptive and analytic purposes. For analytic purpose, aggregate data on disease and exposure rate is collected from different populations. The data about disease rate may be obtained from the same or different sources, as the data regarding exposure rate in the same population. The population level rates of exposure and rates of disease are compared to assess their association (correlation). The finding of correlation suggests possible causal relationship. For instance, ecologic data can be used to examine the association between the prevalence of dental carries and the concentration of fluoride in water supply. The main advantage of ecological studies is that they tend to be conducted using routinely collected data and therefore they can usually be done quickly and inexpensively.

The major disadvantage of ecological studies is that they are based on data on groups of people and not on individuals. In the above example, it is possible there are many other differences that could explain association between fluoride level and dental caries. An association found at the group level may not exist at the individual level, or conversely there may be no association found at the group level when in fact it exists at the individual level. In either case the wrong conclusion would be drawn from the ecological study. This type of misleading result is called an ‘ecological fallacy’. Therefore ecological studies are best seen as useful means of generating hypotheses on the possible determinants of health states, hypotheses which can be tested in more detailed studies.

2.2. Analytic Cross-sectional Studies

Cross-sectional designs are primarily used to assess the prevalence of health conditions. Cross-sectional studies may also be used to identify associations between exposures and health outcomes. The design, conduct and analysis of analytic (comparative) cross-sectional studies differ from the descriptive ones. Unlike for descriptive designs, sample size is determined using two population proportion estimation formula. During their conduct, data is collected simultaneously both about exposure and outcome status.

During analysis, association between exposure and outcome is assessed using prevalence ratio. But for relative simplicity statistical analysis methods, the odds ratio (OR) is commonly used instead.

Analytic cross-sectional designs have serious limitations in identifying determinants. Since exposure and outcome statuses are assessed simultaneously, it may be difficult to assure that the exposure precedes the outcome. If the outcome precedes the exposure the error that results is termed as temporal bias. Moreover, cross-sectional designs use prevalent cases rather than incident cases which could result in distortion of the true association between outcome and exposure due to the following reasons:

- The person might have changed or have difficulty to recall (recall bias) the behaviour that resulted in the disease
- Prevalent cases represent survivors who may be atypical with respect to exposure status (survival bias)
- As exposure and outcome have already happened it is difficult to fully control the effect other determinants.

For the above reasons, analytic cross-sectional studies in general provide weak evidence of causation. Due to their weaknesses their applications is limited. They are best used for screening hypotheses to be further tested using stronger designs. However, they can provide better evidence for factors that remain unaltered overtime like sex, race and blood group and variables that do not influence survival. For instance, if a study finds association between type of blood group and peptic ulcer diseases, one can be sure that the association could not be due to temporal bias though other possible explanations should be ruled out.

Despite their limitations they have the advantage of giving results quickly and being cheap. Hence, they are the preferred design options when there is shortage of time and other resources required to conduct stronger analytic studies.

Summary

In this session you have learned that:

- Analytic ecological and cross-sectional studies provide weak evidence of causation
- The design of analytic cross-sectional studies differs from descriptive ones
- Analytic cross-sectional designs may be the only available option in resource limited settings

References

Carr S, Unwin N and Mulloli PT. An Introduction to Public Health and Epidemiology, 2nd Ed. England: Open University Press, 2007.

Varkevisser C, Pathmanathan I and Ann Brownlee A. Designing and Conducting Health Systems Research Projects. Amsterdam: KIT Publishers and International Development Research Centre in association with WHO Regional Office for Africa, 2003.

Session 3: Analytic Epidemiologic Studies: Case-control Studies

Session Overview

Case-control studies in comparison to cohort studies are conducted more frequently due to their relative feasibility to conduct. This section introduces you to basic concepts, limitations and applications of case-control studies.

Learning Objectives

At the end of this session, you should be able to:

- Outline the design of case-control studies
- Describe the limitations of case-control studies
- Identify applications of case-control designs

3.1. Definition, Overview of Design and Types of Case-Control Studies

3.1.1. Definition

Case-control studies aim to identify possibly causal associations between exposure and health outcome by comparing likelihood of past exposure among individuals having a health outcome with those not having the condition. People having the health outcome of interest are termed as cases and those without it are termed controls. The term case refers to individuals having any health outcome which is of interest to the study. But the outcome is commonly a disease.

A typical example of a case-control study is epidemiologists' investigation of the association between cigarette smoking and lung cancer. In this study, people with lung cancer were the cases and those without lung cancer were controls. Comparison was made between the cases and controls with regard to their ratio of smokers to non-smokers. The finding of a higher ratio of smokers to non-smokers among cases in comparison to controls suggested that smoking is possibly a cause of lung cancer.

3.1.2. Overview of Design

Case-control studies start with selection of cases and controls from a source population. Then, past exposure status of all individuals is assessed and the ratio of exposed to non-exposed is compared between the cases and the controls.

The direction of inquiry about exposure is always backwards in time, but the timing of actual data collection can be carried out in either retrospective or prospective (Figure 3.1.).

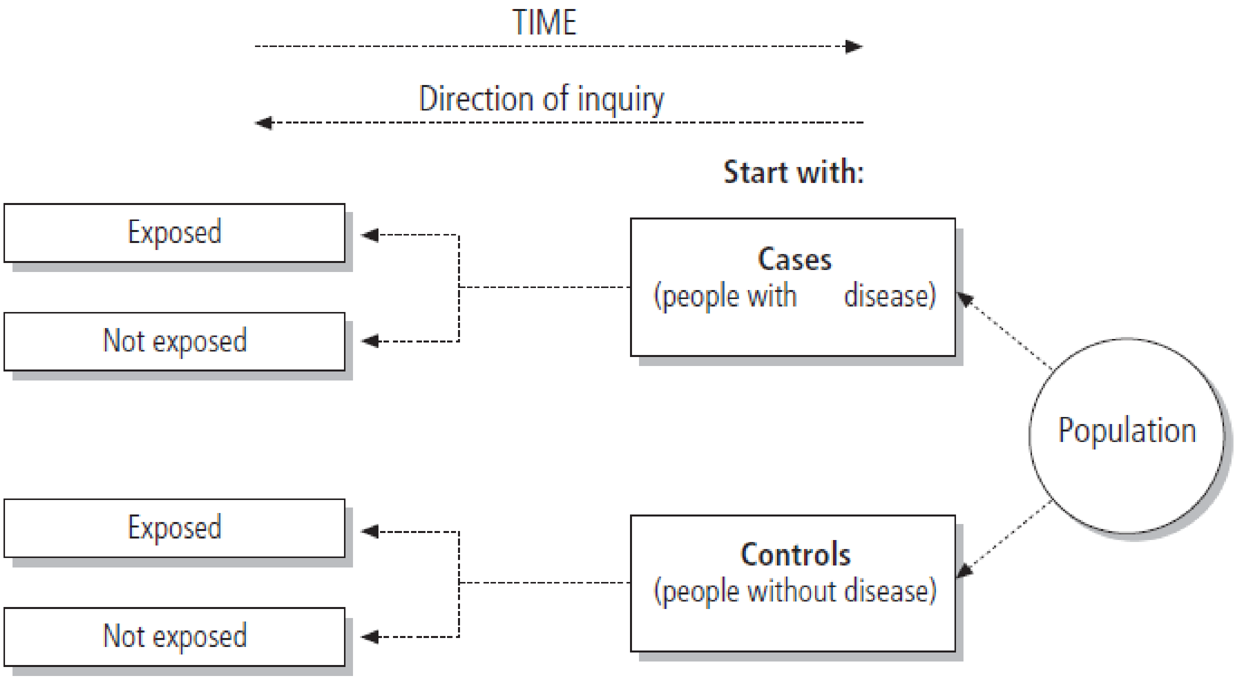


Figure 3.1.: overview of design of case-control studies.

Question for self-practice:

The following research objectives can be met through the conduct of case-control studies. For each objective identify who will be the cases, controls, exposed and not exposed?

- Assessment of determinants of neonatal mortality
- Assessment of determinants of Goiter
- Assessment of determinants of MDR TB
- Assessment of determinants of Epilepsy

- Assessment of factors associated with delayed care seeking among HIV infected individuals

3.1.3. Types of Case-Control Studies

The cases in case-control studies could be new (incident) cases or existing (prevalent) cases. Case-control studies based on incident cases are preferable as they avoid survival bias and are better in assuring the exposure precedes the outcome. However, they are more expensive due expenses for follow-up to recruit the new cases. On the other hand, case-control studies based on prevalent cases, despite their liability to bias, are more feasible. When records are available they can be conducted using secondary data alone. Due to their feasibility, they are the more commonly used ones.

Considering the whole aim of LSI training, which is to enhance utilization of secondary data for decision-making, the focus of this session will be on the latter type. Hereafter, unless specified, the term case-control would refer to a case-control study which uses prevalent cases.

Question for self-practice:

Which of the research objectives listed in the previous question do you think can be achieved using secondary data? Can you think of other research areas that can be carried out using secondary data available in your setting?

3.2. The design and conduct of case-control studies

The main activities during the design and conduct of case-control studies include: identifying source population, selection of cases, selection of controls, sampling, assessment of outcome and exposure status and analysis of data. Each of those activities is discussed below in some details.

3.2.1. Identifying source population

At the initial stages during design of a study, researchers need to decide who will be the population of interest for the study. The population of interest is the target population i.e. the population about which conclusions will be made. Ideally the target should serve as a source population for selecting representative samples. But practically all members of the target population may not be accessible and the source population may be smaller. Nevertheless, it is from the same source population that both cases and controls need to be selected.

3.2.2. Selection of cases

Selection of cases requires defining who would constitute cases and identifying sources for the cases. Definition of cases involves setting criteria to decide who will be considered as cases for the study. Based on the definition, cases should be selected from appropriate sources. The potential sources include hospital or clinic records, general population or records of employers. Which sources to use depends on whether the selected cases represent all cases in the target population.

3.2.3. Selection of controls

Selection of controls also requires defining who would constitute as controls and identifying sources for the controls. Definition of controls involves establishing criteria to decide who will be considered as controls for the study. Using the definition, controls should be selected from appropriate sources. The potential sources of controls include general population, neighbors of cases, friends of cases and patients hospitalized with other diseases. Which sources to use depends on whether the selected controls are representative of the source population in terms of probability of exposure to the risk factor.

3.2.4. Sample size determination

A representative sample size of cases and controls should be determined either manually using appropriate sample size formula or preferably using software like Epi Info and OpenEpi. The parameters whose values need to be known for sample size determination are: confidence level, power, ratio of controls to cases, percent of controls exposed and odds ratio (OR). Confidence level and power are conventionally taken to be 95% and 80%.Ratio of controls to cases is preferably taken to be 1:1 but could be increased depending on scarcity of cases. The percent of controls exposed is taken from other similar studies and value of OR is decided considering minimally important difference that needs to be detected. Alternatively instead of OR, the percent of cases exposed can be used in determining the sample size.

3.2.5. Assessment of outcome and exposure status

The required sample of cases and controls are selected using their respective definitions from the identified sources. For deciding eligibility to be a case or control data is collected regarding value variables used as criteria in the definition. The data can be collected directly from individuals and/or obtained from secondary sources.

For each of the sampled cases and controls, their past exposure status is to be assessed. In order to assess exposure status, what constitutes an exposure should be defined first. All necessary data about exposure can then be obtained through interview and/or review of medical records. As much as possible the data collectors should be blinded to outcome status of each study unit in order to minimize bias that could result from intensively looking for history of exposure among cases.

3.2.6. Analysis and interpretation of data

The Odds Ratio (OR) is used in case-control studies to measure association between exposure and outcome. OR is computed as: the odds of exposure in cases divided by odds of exposure in controls. Further details about calculation of OR and its interpretation is discussed under the session about measures of association.

3.3. Sources of error in case-control studies

In case-control studies, errors could occur during selection of study participants or during measurement of exposure and outcome. The effect of confounders is also another problem.

Samples selected for the study should be representative of the source population. Error occurs when the study participants are not representative of the source population. The resulting non-representativeness is termed as selection bias. Case-control studies are liable to selection bias. For instance, selection of hospital cases to represent all cases in a population may result in selection bias. Hospital cases tend to be those with severe disease and bias could occur if causes differ in severe and mild forms of the disease. Selection bias may also occur from selection of controls which do not represent the source population.

During measurement of variables, errors could occur from defects in the data source or data collection methods. Due to the retrospective nature of the design, case-control studies are particularly susceptible to the effects of recall bias and inaccuracy and incompleteness of records regarding exposure and outcome variables. In addition, there may be differential reporting of exposure information between cases and controls based on their outcome status. Cases and controls may recall past exposure differently, if knowledge of being a case affects whether the individual remembers a certain exposure. Similarly, the efforts to retrieve exposure data may vary depending on the investigator’s knowledge of an individual's outcome status (interviewer/reviewer bias).

In addition to exposure and outcome status, measurement of the timing of occurrence of exposure and outcome is also important. This is because, in order to establish causation, the exposure must have happened before occurrence of the outcome. But in case-control studies, both outcome and exposure have already happened and retrieving the exact time of occurrence of exposure and outcome may be difficult. When the exposure had actually happened after the outcome, it results in temporal bias (also known as reverse causality).

Confounding could also occur in case-control studies. Confounding, which will be discussed in detail in the latter sessions, implies distortion of association between exposure and outcome variables due to presence of another variable/s. Collection of data regarding confounder variables is important for controlling their effect through analytic methods (the analytic methods are discussed in later sessions). But in case-control studies, data regarding confounders may not be available or accurately recorded, and respondents may have difficulty in recalling. This may result in distortion of the association between exposure and outcome.

Question for self-practice:

Suppose you have developed a hypothesis that states ‘HIV infection among TB patients is associated with development of MDR TB’. Then you decide to conduct a case-control study using secondary data. How would you define a case and a control for the study? Where would you find the cases and controls? What do you think would be the possible sources of error in the study?

3.4. Merits, demerits and applications of case-control studies

Case-control designs are, in general, considered to be stronger than comparative cross-sectional designs but weaker than cohort designs for assessing cause and effect relationship. This is because generally different sources of error are commoner and more serious in cross-sectional than case-control and in case-control than cohort designs. But in terms of feasibility the order is reversed with cohort studies being least feasible.

The merits of case-control studies over cohort studies include:

- They require relatively small sample size due to efficient nature of the design
- The results are obtainable relatively quickly
- They are less expensive
- They are suitable for identifying determinants of rare outcomes like rare disease

The demerits of case-control studies include:

- They are highly affected by selection bias, information biasand confounding
- The OR, which is the measure of association obtained from case-control studies, doesn't give direct estimate of relative risk, rather it exaggerates risk.
- They are not efficient for rare exposures.

The application of case-control studies depends on their merits and demerits. Accordingly, they are applicable in the following conditions:

- To test hypothesis generated from descriptive studies particularly when resources are limited as case-control studies are less expensive and results can be obtained more quickly than cohort studies. Their efficiency is generally due to requirement for smaller sample size, absence of follow-up and the possibility to use secondary data.
- For screening factors to be further studied using cohort
- When the outcome of interest is rare, case-control studies are preferred for statistical efficiency.

Summary

In this session you have learned that:

- Case-control studies compare cases and controls with regard to their likelihood of exposure.
- The conduct of case-control studies involves mainly identifying source population, selection of cases and controls, determining sample size, assessment of outcome and exposure status and analysis of data.
- Case-control studies are affected by selection bias, information bias and confounding.
- Case-control studies are more applicable when resources are limited

References

Farmer R, Lawrenson R. Lecture Notes: Epidemiology and Public Health Medicine. USA: Blackwell publishing Ltd, 2004.

Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

Bonita R. Beaglehole R and Kjellstom T. Basic Epidemiology. Geneva: WHO, 2000.

Bailey L, Vardulaki K, Langham J and Chandramohan D. Introduction to Epidemiology. England: Open University Press, 2005.

Session 4. Analytic Epidemiologic Studies: Cohort Studies

Session Overview

Cohort studies are a form of longitudinal study designs that flow from the exposure to outcome. This section introduces you to basic concepts, limitations and applications of cohort studies.

Learning Objectives

At the end of this session, you should be able to:

- Describe the design of cohort studies
- Describe the limitations of cohort studies
- Identify applications of cohort designs

4.1. Definition, Overview of Design and Types of Cohort Studies

4.1.1. Definition

Cohort studies are epidemiological studies that start by identifying individuals with and without exposure to a factor/s who are followed over time to determine the rates of development of a health outcome/s. Cohort studies aim to identify possibly causal associations between exposure and outcome by comparing rates of the health outcome in exposed with non-exposed individuals.

A typical example of cohort study is epidemiologists' investigation of association between cigarette smoking and coronary heart disease (CHD). In this study, people who smoke cigarettes are considered as exposed and those who do not smoke as non-exposed. Both groups are followed for a period of time and compared with regard to frequency of development of CHD. The finding of higher frequency of CHD in smokers as compared to non-smokers would suggest that smoking is possibly a cause of CHD.

4.1.2. Overview of design

Cohort studies typically start with selection of exposed and non-exposed individuals from a given population. Then, the subsequent development of outcome assessed and the rate of outcome is compared between the exposed and non-exposed. The direction of inquiry about outcome is always forwards in time (Figure 4.1.). However, the actual data collection can be carried out in either retrospective or prospective manner as described in the next section which is about types of cohort studies.

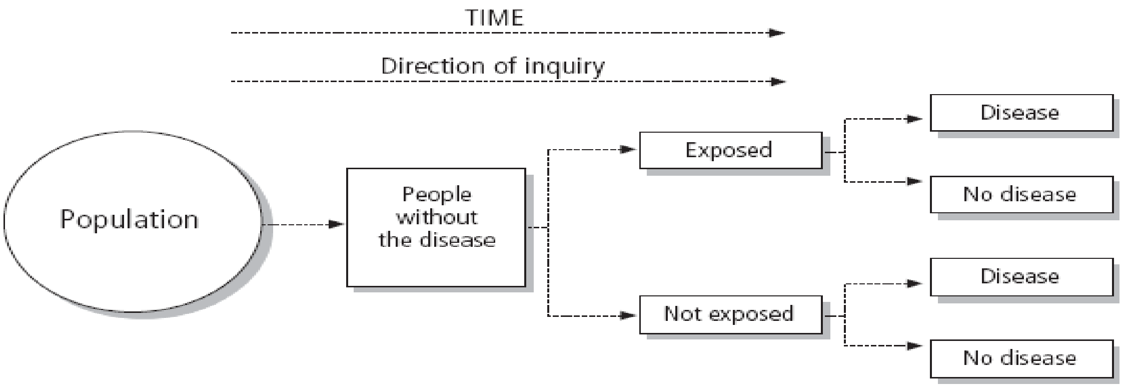


Figure 4.1.: overview of design of cohort studies.

Question for self-practice:

The following research objectives can be met through the conduct of cohort studies. For each objective identify who will be the exposed, non-exposed, those with the outcome and those without the outcome.

- Assessment of determinants of neonatal mortality
- Assessment of determinants of Goiter
- Assessment of determinants of MDR TB
- Assessment of association between HIV and development of MDR TB
- Assessment of factors associated with AIDS mortality

4.1.3. Types of cohort studies

The number of exposures assessed in a cohort study could be one or multiple factors. In either case, data about outcome may be collected pro- or retrospectively. Prospective cohort studies start with assessment of exposure status of outcome-free study participants who are then followed for development of outcome (Figure 4.2.).

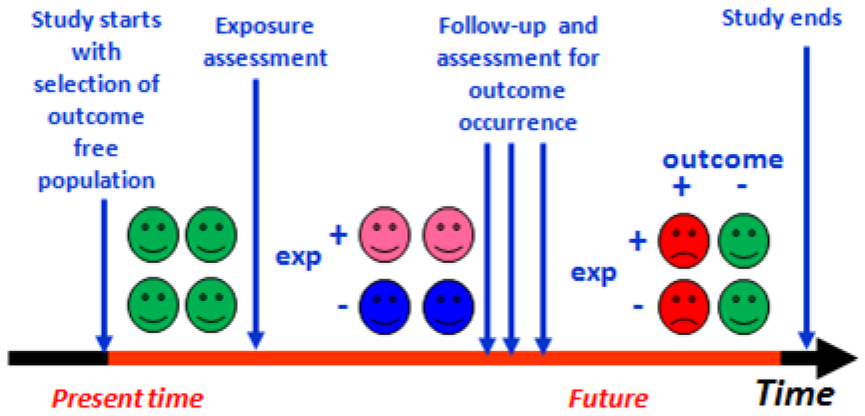

Study starts with selection of outcome free population
Exposure assessment
Follow-up and assessment for outcome occurrence
Study ends
outcome
+
-
exp
+
-
exp
+
-
Present time
Future
Time

Fig 4.2.: Design of prospective cohort study

In retrospective cohort studies, the outcome has happened before start of the study and data about past exposure status and subsequent development of outcome is reviewed from records (Figure 4.3.). These studies start with identification of outcome-free population and review of their exposure status unlike case-control studies which start with identification of cases and controls.

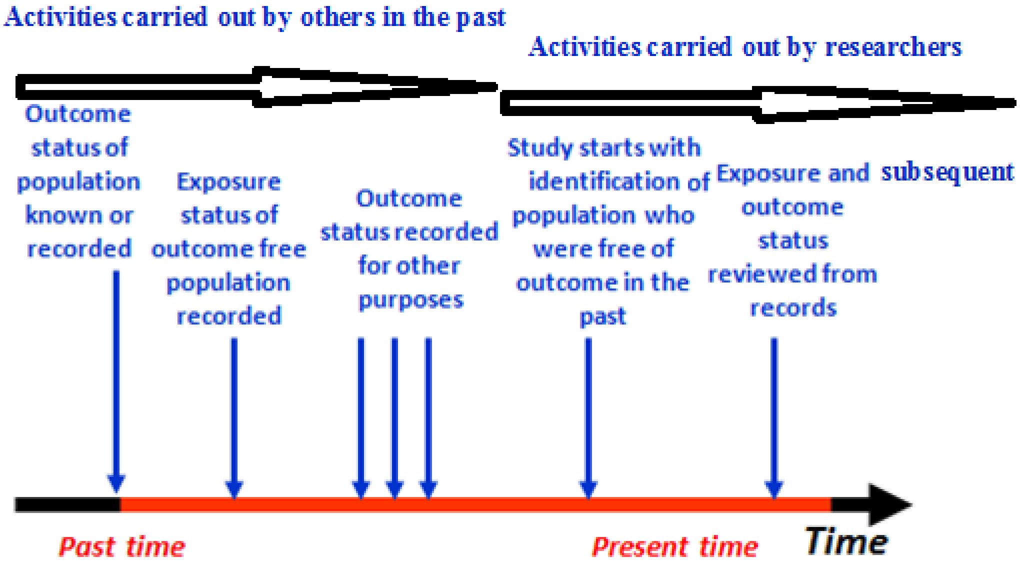

Activities carried out by others in the past
Activities carried out by researchers
Outcome status of population known or recorded
Exposure status of outcome free population recorded
Outcome status recorded for other purposes
Study starts with identification of population who were free of outcome in the past
Exposure and subsequent outcome status reviewed from records
Past time
Present time
Time

Fig 4.3.: Design of retrospective cohort study

In prospective cohort studies participants are followed by researchers whereas in retrospective cohort, follow-up time has elapsed and data is collected from secondary sources. The prospective design enables to get more accurate data as compared to retrospective ones where risk of bias and confounding is higher. But the retrospective design is much cheaper.

Question for self-practice:

Which of the research objectives listed in the previous question do you think can be achieved using retrospective cohort design? Can you think of other research areas that can be carried out using secondary data available in your setting?

4.2. The design and conduct of cohort studies

The major activities in the design and conduct of cohort studies are: Identifying source population, selection of exposed, selection of non-exposed, sampling, assessment of exposure and outcome status, and analysis of data. Each of those activities is discussed below in some details.

4.2.1. Identifying of source population

The researchers should first decide who will be their population of interest for the study. Ideally the source population should be representative of the target population i.e. the population about which conclusions will be made.

The source population should be free of the outcome and serve as a base for selecting exposed and non-exposed individuals.

4.2.2. Selection of exposed

Selection of exposed requires defining who would constitute exposed and identifying appropriate sources. Definition of exposed involves setting criteria to decide who will be considered as exposed. The exposed could be selected from different sources. The potential sources of exposed include the general population, occupational groups, hospital or clinic records. The selected exposed should represent all exposed in the source population.

4.2.3. Selection of non-exposed

Selection of non-exposed requires defining who would constitute as non-exposed and identifying the source of non-exposed. Definition of non-exposed involves establishing criteria to decide who will be considered as non-exposed for the study. The non-exposed could be selected from different sources. The potential sources of non-exposed include general population, hospital records or occupational groups. The selected non-exposed should be representative of the source population in terms of probability of developing the outcome of interest.

4.2.4. Sample size determination in cohort studies

A representative sample size of exposed and non-exposed should be determined either manually using appropriate sample size formula or preferably using softwares like Epi Info and OpenEpi. The parameters whose values need to be known for sample size determination are: confidence level, power, ratio of exposed to non-exposed, percent of unexposed with the outcome and relative risk (RR). Confidence level and power are conventionally taken to be 95% and 80%.Ratio of non-exposed to exposed individuals in the sample is preferably taken to be 1:1 but could be increased depending on scarcity of exposed individuals. The percent of exposed with the outcome is taken from other similar studies and the value of RR is decided considering minimally important difference that need to be detected. Alternatively instead of RR, the percent of non-exposed with outcome can be used in determining the sample size.

4.2.5. Assessment of exposure and outcome status

Exposure status should be measured using the established definition. Exposure data may be obtained using different techniques like review of records, interviews or observation. In the retrospective type of cohort study data is obtained through review of records like medical charts, or service registries.

For both the exposed and non-exposed, subsequent development of outcome is assessed. In prospective cohort the researchers follow study participants to measure outcome, whereas in retrospective cohort the follow-up time has elapsed and outcome status is obtained from records.

Data about outcome measures may be collected directly from the participants or obtained from records. Note that the method used to measure outcome must be identical in both exposed and unexposed groups to avoid the risk of bias. Whenever possible it is valuable to record the time interval between exposure and development of outcome.

4.2.6. Analysis and interpretation of data

The Relative Risk (RR) is used in cohort studies to measure association between exposure and outcome. RR is computed as: the incidence of outcome in exposed divided by incidence of outcome in non-exposed. Further details about calculation of RR and its interpretation are discussed later under the session about measures of association. When data about time to development outcome are available it is better to compute Hazard Ratio (HR) whose numerator and denominator are incidence density of outcome in exposed and non-exposed, respectively.

4.3. Sources of error in cohort studies

The major source of error in cohort studies is loss to follow-up. Error could also occur from other sources but less commonly than in case-control studies. In general, error in cohort studies could occur during selection and follow-up of study participants, and during measurement of exposure, outcome and confounders.

Selection of study participants should be carried out in a representative way. Findings from study participants selected from institutions may not be generalizable to the general population.

Loss to follow up is a major source of bias in cohort studies. Members of the study population may die, migrate, change jobs or refuse to continue to participate in the study. Loss to follow-up reduces the sample size and could also result in distorted measure of association between exposure and outcome. The error becomes more serious if losses to follow up are related to the exposure, to the outcome or to both. For example, in a cohort study assessing relationship between smoking and lung cancer, error results if smokers are more likely to be lost from follow-up when they develop lung cancer than nonsmokers who develop the disease.

During measurement of variables, errors could occur due to defects in the data source or data collectors. In retrospective cohort, inaccuracy and incompleteness of records regarding exposure, outcome and confounder variables is a major source of error. In prospective type, though less severe, errors could occur during measurement. One typical example is overdiagnosis of outcome among exposed than non-exposed. This occurs due to exertion of more effort during measurement of outcome among exposed than non-exposed by biased observers. In order to avoid such error, the methods used to measure outcome must be identical in both exposed and unexposed groups and, if possible, observers should be blinded to exposure status.

Question for self-practice:

Suppose you wanted to identify factors associated with MTCT of HIV among infants born to HIV infected mothers using a retrospective cohort design.Who will be your source population, exposed and non-exposed? What are the possible sources of error in the study?

4.4. Merits, demerits and applications of cohort studies

Cohort designs, in general are considered to be the strongest among observational designs for assessing cause and effect relationship. This is because different sources of error can be minimized. But application of the design is less common than other analytic designs mainly due to limited feasibility of the prospective type and lack of accurate data for the retrospective type.

The merits of cohort over case-control designs include:

- The design enables to get more valid findings as several of the sources of error can be reduced. The design, particularly the prospective type, assures the exposure precedes the outcome avoiding temporal bias. On top of this, the risk of error during sample selection and collection of information is lower especially for the prospective type.

- The design allows computing RR which is a direct measure of risk.

- The design enables to identify more than one possible outcomes of an exposure.

- For rare exposures, the design allows recruiting as many exposed individuals as available, and also increasing the size of non-exposed individuals to compensate for the few number of exposed thus improving study efficiency for rare exposures.

The demerits of cohort studies include:

- They are highly liable to bias from loss of study participants to follow-up.
- The prospective type of cohort is particularly expensive. The expense is mainly due to cost related to follow-up and repeated measurements. Besides, cohort studies require relatively higher sample size which increases the cost.

- The results of the prospective type of cohort are not obtainable quickly due to the time taken for follow-up especially for outcomes which need long time to develop like cancers.
- They are efficient for rare exposures like radiation exposure but not for rare outcomes.

The application of cohort studies depends on their merits and demerits in comparison to the other analytic studies. But even within cohort, the relative merits of the prospective versus the retrospective type also vary. The prospective type is much more expensive but gives more valid findings. The retrospective on the other hand though cheap requires availability of accurate records. Accordingly, the conditions for their application may differ. Overall cohort studies are more applicable in the following conditions:

- To test hypothesis generated from descriptive studies, retrospective cohort can be used whenever recorded data is available. The prospective cohort however, due to its huge expense, is reserved for issues of major public health importance and outcomes that develop in a short period after exposure.

- When the exposures are screened with other analytic designs.

- In investigating epidemics that are localized.

- When interested to identify multiple outcomes.

- When the exposure is rare like exposure to radiation or chemicals at workplace.

Question for self-practice:

Suppose you wanted to assess factors associated with HIV sero-discordance among couples visiting ART clinic, using secondary data. Which analytic design would you prefer to use? Justify your answers.

Summary

In this session you have learned that:

- Cohort studies compare exposed and non-exposed with regard to their rate of outcome development.

- The conduct of cohort studies involves selection of source population, selection of exposed and non-exposed, determining sample size, assessment of exposure and outcome status and analysis of data.
- Cohort designs are in general considered to be the strongest among observational designs for assessing cause and effect relationship.
- The prospective cohort design is stronger but expensive. Its application is limited to testing of exposures screened by other analytic studies and to causal study of major public health issues.
- The retrospective on the other hand though cheap requires availability of accurate records about exposure and outcome status.

References

Farmer R, Lawrenson R. Lecture Notes: Epidemiology and Public Health Medicine. USA: Blackwell publishing Ltd, 2004.

Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

Bonita R. Beaglehole R and Kjellstom T. Basic Epidemiology.WHO, Geneva, 2000.

Bailey L, Vardulaki K, Langham J and Chandramohan D. Introduction to Epidemiology. England: Open University Press, 2005.

Session 5. Measures of Association

Session overview

Descriptive epidemiologic designs help in generating hypothesis about determinants of heath. Those hypotheses are tested using analytic designs. Analytic studies identify determinants through assessing their association with the outcome. The association is indicated using measures of association. In this session you will learn about the measures of association.

There are different types of measures of association. Do you remember any measure of association that has been mentioned in the previous sessions? As you would say, RR and OR are the measures of association for cohort and case-control studies respectively. The RR and OR, being the commonly used ones, will be the focus of discussion in this session. But bear in mind that there are also other types of measures of association like regression coefficient and hazard ratio.

The RR and OR are used when both the outcome and exposure variables have categorical values. For other type of variables, different types of measures of association are used. For instance the regression coefficient is used for continuous variables and the hazard ratio is used for time dependent outcome variables.

This session covers the basic types of measures of association, starting with the RR and then followed by the OR.

Learning Objectives

At the end of this session, you should be able to:

- Define the measures of association
- Explain applications of the measures of association
- Compute and interpret values of measures of association

5.1. Relative Risk (RR)

5.1.1. Definition and application

RR is a ratio that compares incidence of outcome among exposed with incidence among non-exposed. Itis computed using the following formula:

RR=Incidence in exposed/Incidence in non-exposed

RR indicates whether the probability (risk) of developing the outcome varies between the exposed and non-exposed. It also tells how much higher or lower the risk is in exposed as compared to non-exposed.

RR is used in cohort studies which are the appropriate designs for obtaining incidence rates.

5.1.2. Calculation and interpretation of RR

As indicated in the formula above, the numerator and denominator in the calculation of RR are incidence in exposed and incidence in non-exposed, respectively.

In order to compute RR, it is helpful to present the findings of a cohort study in a 2x2 table as follows:

Table 5.1: a 2 by 2 table indicating findings of a cohort study.

Exposure status	Outcome status		
	Developed outcome	Didn’t develop outcome	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Warm-up question: Which parameters in the two-by-two table are the numerator and denominator of RR?

From the above table the RR is calculated as:

RR = [a/a+b] / [c/c+d]

The interpretation of RR varies depending on its value. A RR of 1 implies that the probability of developing outcome is the same in exposed and unexposed individuals. It indicates the absence of association between the exposure and outcome. A value greater than 1, let’s say x, implies that exposed individuals are x times highly likely to develop the outcome as compared to non-exposed. This indicates the exposure is positively associated with the outcome. A value below 1, let’s say x, implies that exposed individuals have (1-x)100% lower probability of developing the outcome than the non-exposed.

Example: Let’s calculate RR from findings of a hypothetical cohort study conducted to assess association between malaria during pregnancy and low birth weight (Table 5.2.).

Table 5.2.Hypothetical cohort study of association between malaria during pregnancy and low birth weight.

Malaria during pregnancy	Low birth weight		
	Yes	No	Total
Yes	50	50	100
No	100	800	900
Total	150	850	1000

RR= [50/100] / [100/900]

RR= 4.5

The RR calculated from the hypothetical data indicates that women who had malaria during pregnancy were 4.5 times more likely to deliver a low weight baby than those who didn't have malaria during pregnancy.

N.B: Statistical significance of the association is assessed using a corresponding p-value or confidence interval (CI).

5.2. Odds Ratio (OR)

5.2.1. Definition and application

Odds Ratio is a ratio that compares odds of exposure among cases with odds of exposure among controls. It is computed using the following formula:

- OR=odds of exposure among cases/odds of exposure among controls

OR indicates whether the likelihood (odds) of being exposed varies between cases and controls. It also tells how much higher or lower, the odds of exposure is in cases as compared to controls.

OR is the appropriate measure of association for case-control studies. It is also used for analytic cross-sectional studies.

5.2.2. Calculation and interpretation of OR

As indicated in the formula above, the numerator and denominator in the calculation of OR are odds of exposure in cases and odds of exposure in controls, respectively.

In order to compute OR, it is advisable to present findings of the study in a 2 by 2 table as follows.

Table 5.3: 2 by 2 table indicating findings of case-control studies.

Exposure status	Outcome status		
	Cases	Controls	Total
Exposed	a	b	a+b
Non-exposed	c	d	c+d
Total	a+c	b+d	a+b+c+d

Warm-up question: From the 2 by 2 table above, identify the parameters to be used as numerator and denominator of OR.

From the above table the OR is calculated as: $OR = \frac{a/c}{b/d} = \frac{ad}{bc}$

The interpretation of OR varies depending on its value. An OR of 1 implies that the likelihood of having been exposed is the same in cases and controls. It indicates the absence of association between the exposure and outcome. A value greater than 1, let's say x, implies that the odds of exposure is x times higher among cases than controls. This indicates the exposure is positively associated with the outcome. A value below 1, let's say x, implies that the odds of exposure is by (1-x)100% lower in cases than controls.

The OR unlike the RR is based on comparison of odds which exaggerates probability. Hence, the OR tells exaggerated risk. For rare outcomes, the OR approximates to and can be interpreted as RR.

Example: Consider a hypothetical case-control study conducted to assess association between low birth weight and neonatal mortality among 100 cases and 400 controls. Let's calculate OR from findings of the study indicted in the table below.

Table 5.4.Hypothetical case-control study of association between low birth weight and neonatal mortality.

Birth weight	Neonatal death		
	Yes	No	Total
Low	60	100	160
Normal	40	300	340
Total	100	400	500

$$OR = \frac{60 \times 300}{100 \times 40}$$

$$\mathbf{OR = 4.5}$$

The OR calculated from the hypothetical data indicates that the odds of dying at neonatal age for those with low birth weight was 4.5 times than for those with normal birth weight.

N.B: Statistical significance of the association is assessed using a corresponding p-value or confidence interval (CI).

Summary

In this session you have learned that:

- RR compares incidence of outcome among exposed with incidence among non-exposed. On the other hand, OR compares odds of exposure among cases with odds of exposure among controls.
- RR is used in cohort studies and OR in case-control studies.
- RR and OR values greater than 1 indicate positive association, values less than 1 indicate negative association and a value of 1 indicates absence of association between the exposure and the outcome.

Exercise

Suppose you are interested to test whether sexually transmitted infection increases risk of HIV infection. You then conducted a study among individuals tested for HIV and managed to review last 5 years data from registers in HIV counseling and testing unit of a hospital and come-up with the findings indicated in the table below.

Had history of symptoms of sexually transmitted infection	HIV test results		
	Positive	Negative	Total
Yes	40	100	140
No	60	1800	1860
Total	100	1900	2000

- Use the information given above to answer the following questions.
- a) Which analytic design was, most likely, employed in the study?
- b) Which measure of association would you calculate from the data?
- c) Calculate and interpret the measure of association.
- d) What are the possible sources of error in the study?

References

Spasoff R. Epidemiologic Methods for Health Policy. New York: Oxford University Press, 1999.

Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

Bailey L, Vardulaki K, Langham J and Chandramohan D. Introduction to Epidemiology. England: Open University Press, 2005.

Session 6. Measures of Impact

Session overview

Measures of association are helpful for identifying determinants of disease and other health outcomes. Identifying determinants by itself is not sufficient for implementing public health measures. The health impact of the determinants needs to be measured for prioritization and resource allocation.

Health impact of determinants is assessed using different types of measures. In this session you will learn about the basic types of measures of impact specifically attributable risk, attributable risk percent, population attributable risk and population attributable risk percent.

Learning Objectives

At the end of this session, you should be able to:

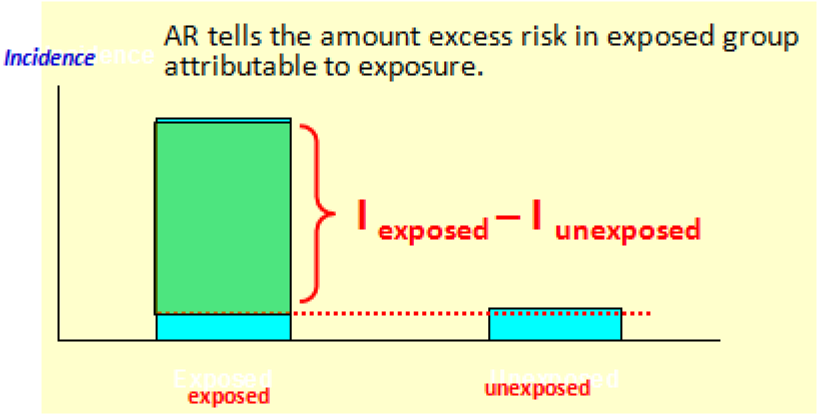
- Define the measures of impact
- Explain applications of the measures of impact
- Compute and interpret values of measures of impact

6.1. Attributable Risk (AR)

AR is a measure of the amount of risk of disease among exposed individuals that is attributable to the exposure.

AR is computed as the difference between incidence of disease among exposed and incidence of disease among non-exposed. This implies that not the entire disease incidence among exposed is due to exposure since even some of non-exposed individuals develop the disease. The formula for AR is:

- AR = Incidence in exposed – Incidence in non-exposed (Fig 6.1.)



AR tells the amount excess risk in exposed group attributable to exposure.

Fig 6.1.Pictorial illustration of AR calculation.

AR is obtained from cohort studies as its calculation needs incidence rates.

Its use in public health is that it indicates the amount of risk of disease that can be prevented by avoiding the exposure.

Example: Consider the hypothetical cohort study conducted to assess association between malaria during pregnancy and low birth weight. Let's calculate AR from findings of the study indicated in the table below.

Table 6.1.Hypothetical cohort study of association between malaria during pregnancy and low birth weight.

Malaria during pregnancy	Low birth weight		
	Yes	No	Total
Yes	50	50	100
No	100	800	900
Total	150	850	1000

AR= [50/100] - [100/900] = 0.39 or 39%; the value of AR implies that 39% low birth weight deliveries that occur among women who had malaria during pregnancy are attributable to malaria.

Attributable Risk Percent (AR%)

AR% is an AR expressed as a percentage of risk in exposed. It is a measure of the proportion of disease risk among the exposed that is attributed to the exposure.

AR% is computed using the following formula:

AR% = (Incidence in exposed – Incidence in non-exposed) / (Incidence in exposed)

Alternatively AR% can be obtained using the formula AR% = (RR – 1)/RR (Fig 6.2.). This formula is derived from the previous one by dividing each of the parameters in the formula with 'incidence in non-exposed'. It can be seen from the later formula that AR% can be obtained from a given RR alone.

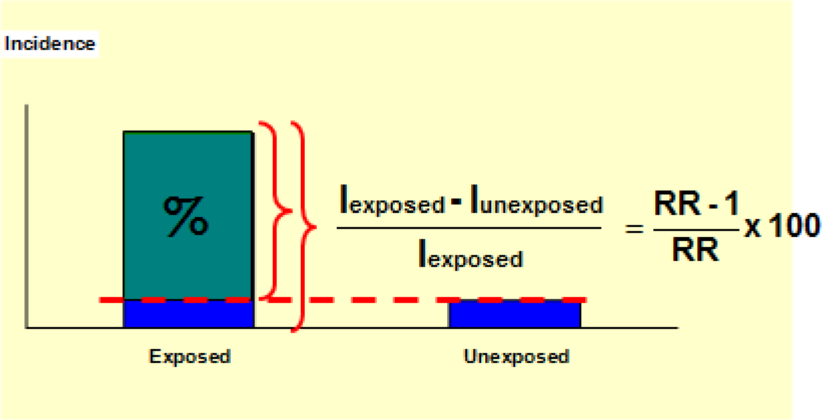

Incidence
%
$\frac{I_{exposed} - I_{unexposed}}{I_{exposed}} = \frac{RR - 1}{RR} \times 100$
Exposed
Unexposed

Fig 6.2. Pictorial illustration of how to compute AR%

The measure is sometimes termed as etiologic fraction as it indicates relative contribution of an exposure in causing the disease. For exposures that are protective the AR% is termed as prevented fraction.

Its public health implication is that it indicates the proportion of risk of a disease that can be prevented by avoiding the exposure. As such, this helps for prioritizing exposures to be avoided in prevention of occurrence of a disease. It is also useful for patient diagnosis.

Example: Let's calculate AR% for the previous cohort study conducted to assess association between malaria during pregnancy and low birth weight.

AR% = [39%] / 50% = 78%; the value indicates that 78% of low birth weight deliveries among pregnant women who had malaria is attributable to malaria.

Population Attributable Risk (PAR)

AR and AR% measure the impact of exposure on exposed individuals.

Warm-up question: Do you think the amount of disease risk in the population attributable to specific exposure would be equal, higher or lower that the AR?

If the whole population is exposed, the amount would the same as the AR. On the contrary, if none of population members are exposed the amount would be zero. The usual condition is however, for some proportion of the population to be exposed. In such situations, the amount is measured using population attributable risk (PAR).

PAR measures amount of disease risk in the total population attributable to a specific exposure. PAR is computed as the difference between incidence of disease among exposed and incidence of disease among non-exposed. This implies that not the entire disease incidence in the population is due to the specific exposure since even some of non-exposed individuals develop the disease. The formula for PAR is:

- PAR = Incidence in population – Incidence in non-exposed (Fig 6.3.)

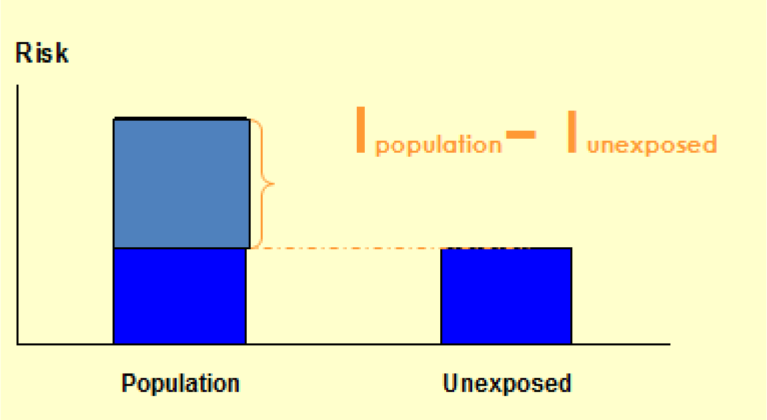


Fig 6.3.Pictoral illustration of PAR calculation.

PAR tells the amount of excess risk in the population due to the exposure. In other words it indicates reduction in risk that can be achieved if population was entirely unexposed. It helps for resource allocation against existing exposures in the population.

Warm-up question:

Is it possible to calculate PAR from the data in the previous example of a cohort study conducted to assess association between malaria during pregnancy and low birth weight?

In order to calculate PAR, incidence rate of delivery to low birth weight baby among all pregnant women in the population is required. If the sampled 1000 pregnant women are assumed to represent all pregnant women in the population, the incidence (50 in exposed + 100 in non-exposed) will be (150/1000) 15%. The PAR would then be: 15% - 11% = 4%. This implies that 4% low birth weight deliveries among the population of pregnant women are due to malaria infection during pregnancy.

Population Attributable Risk Percent (PAR%)

PAR% is a PAR expressed as a percentage of risk in population. It is a measure of proportion of disease risk in population that is attributable to the exposure. It is computed using the following formula:

> PAR% = (Incidence in population – Incidence in non-exposed) / (Incidence in population) (Fig 6.4.)

When population prevalence of the exposure is known, PAR% can be computed using the alternative following formula:

$$\text{PAR\%} = \frac{P\ (RR - 1)}{P\ (RR - 1) + 1} \times 100$$

where P = proportion of population who are exposed

The public health implication of PAR% is that it indicates the proportion of risk of a disease in the population that can be averted by eliminating the exposure. PAR% is useful to prioritize exposures for intervention during control of a disease.

Example: Let’s calculate PAR% for the cohort study in the previous example of a cohort study conducted to assess association between malaria during pregnancy and low birth weight.

PAR% = [15% - 11%] / 15% = 26.7%; the value indicates that 26.7% of low birth weight deliveries among pregnant women is attributable to malaria.

Summary

In this session you have learned that:

- AR and AR% measure the impact of exposure on exposed individuals whereas PAR and PAR% measure the impact of exposure on the population.
- AR is a measure of the amount of disease risk among exposed individuals that is attributable to the exposure.
- AR% is a measure of proportion of disease risk in exposed individuals that is attributable to the exposure.
- PAR measures amount of disease risk in the total population attributable to specific exposure.
- PAR% is a measure of the proportion of disease risk in the total population attributable to the exposure.

References

Spasoff R. Epidemiologic Methods for Health Policy. New York: Oxford University Press, 1999.

Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

Bailey L, Vardulaki K, Langham J and Chandramohan D. Introduction to Epidemiology. England: Open University Press, 2005.

Session 7: Chance, Bias, and Confounding

Session overview :

Epidemiologic research studies possible associations between exposures and outcomes. While the results of an epidemiological study may reflect the true effect of an exposure(s) on the development of the outcome under investigation, it should always be considered that the findings may in fact be due to an alternative explanation.

Such alternative explanations may be due to the effects of chance (random error), bias or confounding, which may produce spurious results, leading us to conclude the existence of a valid statistical association when one does not exist, or alternatively the absence of an association when one is truly present.

Observational studies are particularly susceptible to the effects of chance, bias and confounding, and these need to be considered at both the design and analysis stage of an epidemiological study so that their effects can be minimized.

False findings can arise due to lack of study power, chance, bias, confounding, or a combination of these errors. Unfortunately, for any given study, it is impossible to know how much these factors play a role in the results found since we don't know the truth when the study is started

Learning Objectives

After completing this lesson, you will be able to:

- Explain how chance (random error) might explain an association between exposure and disease.
- Distinguish between selection bias, information bias, and other types of biases in observational studies.
- Discuss the impact of confounding on observational studies.
- Identify ways to minimize confounding.

7.1. CHANCE

Chance is **random error**.

If the investigator conducts the same study in two distinct groups of patients, the results of the two studies may differ simply because of chance variation between the samples. Inadequate sample size increases the risk of chance findings. Larger sample size minimizes random error and makes it less influential on study results. P-values provide information about the probability that chance is responsible for the study result. More information on p-values is presented in the *Principles of Clinical Research* course of this program series. Chance can be detected/measured by confidence intervals, p-values, hypothesis testing or generating, clear prior hypothesis or post hoc.

Table 7.1: Validity of association between exposure and outcome

	True relation between exposure and outcome		
		No	Yes
Study finds relation between exposure and outcome	No	Correct finding	False negative finding
	Yes	False positive finding	Correct finding

Example

In 1997, the Swedish Medical Birth Registry identified six cases of hypospadias in a small number of pregnancies exposed to loratadine. This "signal" was monitored continuously for 7-8 years. The updated analysis of 4,450 exposed pregnancies showed there was no association between hypospadias and loratadine exposure. A number of other studies also showed no association. The conclusion is that the initial signal was likely a chance finding.

7.2. BIAS

Bias is a **systematic error** (an error that is not introduced by chance but by inaccuracy of methods or measurements) in the design, conduct, or analysis of study that results in an incorrect estimate of an association between exposure and outcome.

Bias can occur during any stage of a study:

- literature review of the study question
- selection of the study sample

- measurement of exposure and outcome
- analysis of data
- interpretation of the analysis
- publication of the results

Key points about bias:

- Bias threatens internal validity, and makes the study's validity or accuracy questionable.
- Distortion of the association cannot be corrected by statistical manipulation.
- Bias can affect all types of study designs.
- Larger sample size does not eliminate bias; a larger sample size may simply yield a more precise estimate of biased results if the biases are strong

Types of bias

Selection bias

This is distortion stemming from the procedures used to select subjects and from factors that influence subject participation. Selection bias occurs when the two groups being compared differ systematically. That is, there are differences in the characteristics between those who are selected for a study and those who are not selected, and where those characteristics are related to either the exposure or outcome under investigation.

Minimizing selection bias

- In a case-control study, define criteria for selection of diseased and non-diseased participants independent of exposures.
- In a cohort study, define criteria for selection of exposed and non-exposed participants independent of disease outcomes.
- When using electronic health record database, include participants who are equally eligible to receive all treatments under study

Information Bias

This is distortion stemming from the way data are collected about disease or exposure from the study groups or from the study investigators and interviewers.

Information bias results from systematic differences in the way data on exposure or outcome are obtained from the various study groups.

Types of information bias include:

Observer bias

Observer bias occurs when there are systematic differences in the way information is collected for the groups being studied. Observer bias may occur as a result of the investigator's prior knowledge of the hypothesis under investigation or knowledge of an individual's exposure or disease status. Such information may result in differences in the way information is collected, measured or interpreted by the investigator for each of the study groups.

Minimizing observer bias

- Where possible, observers should be blinded to the exposure and disease status of the individual.
- Blind observers to the hypothesis under investigation.
- In a randomized controlled trial, blind investigators and participants to treatment and control group (double blind randomized controlled trial).
- Development of a protocol for the collection, measurement and interpretation of information.
- Use of standardized questionnaires.
- Training of interviewers.

Loss to follow-up

Loss to follow-up is a particular problem associated with cohort studies. Bias may be introduced if the individuals lost to follow-up differ with respect to the exposure and outcome from those persons who remain in the study.

Recall bias

In a case-control study, data on exposure are collected retrospectively. The quality of the data, therefore, is determined to a large extent by the patient's ability to accurately recall past exposure(s). Recall bias may occur when the information provided on exposure is different between the cases and controls. For example, individuals with the outcome under investigation (case) may report their exposure experience differently than individuals without the outcome (control) under investigation. That is, cases may tend to have a better recall on past exposures than controls.

Recall bias may result in either an underestimate or overestimate of the association between exposure and outcome.

Methods to minimize recall bias include the collection of exposure data from work or medical records or to blind the study participants as to the hypothesis under investigation.

7.3. Confounding and Effect Modification

Confounding involves the possibility that an observed association is due, totally or in part, to the effects of differences between the study groups (other than the exposure under investigation) that could affect their risk of developing the outcome being studied.

Confounding occurs when the effects of two associated exposures have not been separated, resulting in the interpretation that the effect is due to one variable rather than the other. The consequence of confounding is that the estimated association is not the same as the true effect.

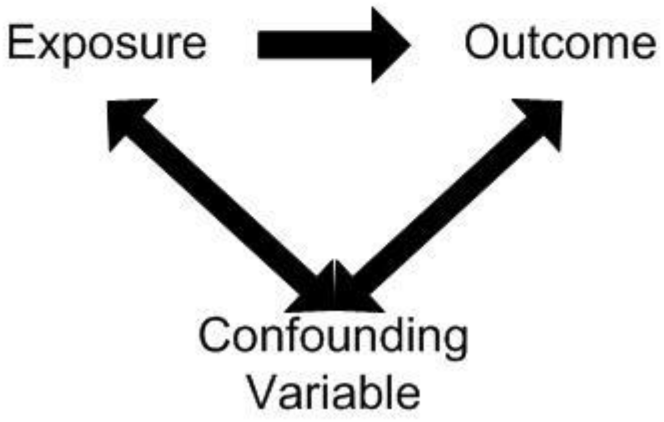
Confounding is introduced when at least one extraneous variable interferes with the observed association between an exposure (the predictor variable) and an outcome.

A confounder is a variable that is:

• Associated with the exposure

• Also a risk factor for the outcome

• Not part of the causal link between the exposure and the disease.

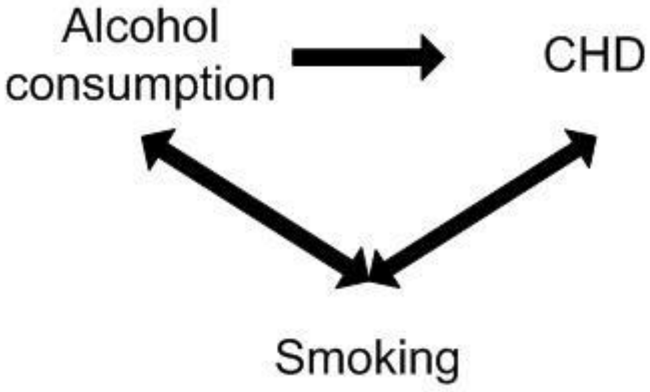


In order for a variable to be considered as a confounder:

1. The variable must be independently associated with the outcome (i.e. be a risk factor).
2. The variable must be associated with the exposure under study in the source population.
3. It should not lie on the causal pathway between exposure and disease.

Examples of confounding

A study found alcohol consumption to be associated with the risk of Coronary Heart Disease. However, smoking may have confounded the association between alcohol and CHD. For example smoking is independently associated with CHD (is a risk factor) and is also associated with alcohol consumption (smokers tend to drink more than non-smokers).



Controlling for the potential confounding effect of smoking may in fact show no association between alcohol consumption and CHD.

Effects of confounding

Confounding factors, if not controlled for, cause bias in the estimate of the impact of the exposure being studied.

The effects of confounding can result in:

* An observed difference between study populations when no real difference exists.
* No observed difference between study populations when a true association does exist.
* An underestimate of an effect.
* An overestimate of an effect.

Controlling Confounding

The validity of the result in an epidemiology study depends, in part, on the **ability to control and account for bias and confounding** in the study design or in the analysis of data.

To the extent that this is not accomplished, there can be "residual confounding" of the study result.

Because of missing or unmeasured data, inaccurate data, unrecognized sources of bias, and confounding, it can be assumed there is almost always some residual confounding of epidemiologic study result.

A major advantage of randomized clinical trials is that the **randomization balances groups** across measured and unmeasured confounders.

Ways of controlling confounding

There are several ways of controlling confounding from design to…..

1. **Design Stage**

   a) **Randomization:**

Patients are randomly assigned to treatment or control group as in randomized clinical trials.

Advantage

•Balances treatment and control groups on known and unknown factors, and measured and unmeasured factors

Disadvantages

- Can be unethical in certain circumstances or impractical due to the rarity of the outcome of interest.
- Patients are not always willing to be randomized.
- Population may be selected and not reflective of the population that will take a product.
- Patients may not adhere to the treatment to which they were randomized.

b) Restriction:

Study participation is limited to individuals who fall within a specified level of the confounder; (e.g., only include patients over the age of 65 in the study) or with certain disease characteristics.

Advantage

- Straightforward and inexpensive

Disadvantages

- Narrow restriction range can limit recruitment and ability to generalize results.
- Broad restriction may result in residual confounding that is the remaining effect when a confounding variable is measured imperfectly and the adjustment using this imperfect measure does not completely remove the effect of confounding.
- Can only examine association between exposure and disease in the specified level (???)

c) Matching:

Subjects are selected in a way such that potential confounders are matched equally among the study groups.

- Individual match: Each case is individually matched with one or more controls on confounding variables
- Frequency match: Controls are matched to cases as a group to have a similar distribution on confounding variables

Advantage

- Very powerful method to control for confounders

Disadvantages

- Time-consuming and expensive.
- Not feasible when there are many different confounders.

•Cannot study the variable on which matching was performed.

2. Analysis Stage

The effect of confounders can also be minimized at the stage of analysis using the following techniques..

a) Stratification:

This method allows evaluation of exposure-disease association within “homogeneous” strata of the confounder.

Advantage

•Very powerful method to control for confounders

Disadvantages

•Not feasible when controlling for (many) multiple confounders.

•Depending on how the strata are defined, there may be residual confounding.

•Sample size may be too small for certain strata.

b) Multivariable analysis:

It is a statistical technique that adjusts for multiple variables simultaneously.

•Multivariate Cox regression; linear, Poisson and logistic regression, Propensity scores

Advantages

•Can control for many confounders at the same time

•Powerful method whether outcome is categorical or continuous or count

Disadvantages

•Sample size must be large to accommodate control for many confounders.

•Depending on the statistical technique, the data must satisfy certain assumptions.

•As other approaches, experts agree that this model adjustment is n not absolute in controlling for confounding

Summary: Chance, Bias, and Confounding

Chance is a random error and can be minimized by increasing sample size.

Bias is a systematic error and cannot be fixed by statistical manipulation or increasing sample size. Rather, following strict procedures during planning, data collection and other process of the research is essential.

Confounding occurs due to a variable that distorts the association between exposure and disease. However, the effect can sometimes be controlled for when confounders are measured adequately and adjusted in statistical analyses.

To conclude that an association between exposure and disease outcome exists, the study must:

- Have adequate sample size,
- Be free of bias, and
- Be adjusted for possible confounders.

Table 1: Prevention of Selection Bias in Study Designs

Type of Selection Bias	Cross Sectional	Case-Control	Retrospective Cohort	Prospective Cohort
Berkson's	Avoid selecting subjects from hospitals	Use population based case and population based control	NA	NA
Prevalence/incidence	Include non-surviving subject in the study through proxy interviews		NA	NA
		Use incident cases		
Detection	NA	Case and controls should be restricted to patients who have under gone	Exposed and unexposed subjects should be under identical disease detection	

		identical detection manoeuvres	
Membership	Difficult to prevent in these four designs		
			Use multiple comparison cohorts
Healthy worker effect	NA	NA	1. Use working cohorts for comparison 2. Use multiple comparison cohorts
Volunteer	1. Use repeated contacts or questionnaire to achieve response rate of at least 80% 2. Compare respondents with a sample of non respondents		
Loss to follow-up	NA	NA	Maintain a high follow-up rate

Table 2. Prevention of Information Bias in Basic Study Design

Type of information Bias	Cross sectional	Case-Control	Retrospective Cohort	Prospective Cohort
Interview	1. "Blinding" of the interviewer with respect to the study hypothesis 2. Use a trained and experienced interviewer			
Inter-interviewer	1. Use only one interviewer in the study 2. Train interviewers according to standard protocols 3. Use the same interviewer for study and comparison groups 4. Discard data from incompetent interviewers			
Questionnaire	1. Careful wording to avoid leading questions 2. Pretest questionnaire several times 3. Use dummy question to conceal hypothesis 4. Offer categorized values for subjects to select instead of requesting specific values			
Recall	Difficult to prevent. May be measured by asking questions whose answers may be checked		NA	NA

	against records			
Diagnostic suspicion	Difficult to prevent Case C?		Both exposed and non-exposed groups should be observed using comparable methods	
Exposure suspicion	Difficult to prevent	Both cases and controls should be observed using comparable methods	NA	NA

Table 3; Prevention of confounding Bias in Cross-Sectional, and Matched and Unmatched Case-Control, and Cohort Studies

Cross-sectional	Case- Control		Cohort	
	Matched	Unmatched	Matched	Unmatched
Although control covariates ensures unbiasedness, unnecessary control for non-confounding covariates always reduces power of the study			Crude estimate is always unbiased	Same as case control unmatched studies
To maximize both validity and power, an investigator should always perform analyses controlling (adjusted estimates) and not controlling (crude estimates) for the covariate(s) If both estimates are similar, then the crude estimate is unbiased and should be adopted on power considerations. If both estimates are not similar then the adjusted estimate, which is the only unbiased one should be used.				

Effect modification

Effect modification means that an exposure's effect on a health outcome differs in different subgroups, or more simply, there are different effects in different groups.

What is the efficacy of measles vaccine? Depends on age at vaccination.Different effect (efficacy) in different age groups.

Who is at greater risk for hip fracture, men or women? (Let class respond.) Depends on age. At younger ages, men are at greater risk, from MVA's, occupational injury, etc. Among elderly, women are at greater risk, because of osteoporosis. Different effect in different groups.

Effect modification is sometimes confused with confounding, because both involve a third variable. But really, effect modication is quite different.

What an epidemiologist calls effect modification a statistician would probably call interaction.

- Conduct crude analysis (simple 2-by-2 table)
- Stratify data by third variable
- Calculate a measure of association for each stratum
- Determine whether association is consistent across strata
    - If not, STOP!
- If so, can calculate a *weighted average* (e.g., Mantel-Haensel)

Example: Diarrhea and Breastfeeding: Stratified by Age of Infant

Overall

Breast fed	Cases	Controls
Yes	120	136
N0	50	204

Odds Ratio =3.6, 95% CI = 2.4-5.5 p < 0.0001

The investigators then stratified the data by age of the infant, and computed a stratum-specific odds ratio for infants less than 1 month of age and infants greater than or equal to one month of age.

Stratum 1

Breast fed	Cases	Controls
Yes	10	3
N0	7	68

OR = 32.4, 95% CI = 6-203, p<0.0001

Stratum 2

Breast fed	Cases	Controls
Yes	110	133
N0	43	136

OR = 2.6, 95% CI = 1.7-4.1, p<0.0001

What do you conclude from these data? (It appears that lack of breastfeeding is a huge problem in the infants younger than 1 month, less so for infants 1 month or older.)

Would you be content to present the summary, or would you present the stratum-specific effects?

How to Tell if Effect Modification is Present

There are two approaches to assessing the presence of effect modification.

The first is to use judgment — are the two effects really that different from a clinical or public health point of view? If not, combine. If so, leave separate.

The second method involves statistical tests such as tests for interaction, homogeneity, or heterogeneity. However, remember that statistical differences do not necessarily mean important differences from a public health or clinical point of view.

Class activity:

1. Let’s say you have four odds ratios -- crude OR, stratum 1 OR, stratum 2 OR, and an adjusted (Mantel-Haenszel) OR.

A. Which two odds ratios do you compare to look for confounding? (crude and adjusted)

B. Which two odds ratios do you compare to look for effect modification? (stratum 1 and stratum 2)

2. For each variable in the table below, indicate whether you think there’s confounding, effect modification, both, neither, calculation error, or you can’t tell from the data provided.

Here’s a hint: look for effect modification first -- compare stratum 1 and stratum 2. If they are pretty close, compare them with the adjusted (which should be the weighted average) just to make sure that it looks correct, then compare the crude and adjusted.

Variable A - confounding (stratum-specific OR’s are very close)

Variable B - effect modification

Variable C - both (not that you would want to summarize)

Variable D - confounding (stratum-specific OR’s are very close).

	Var. A	Var. B	Var. C	Var. D
Crude	4.0	4.0	4.0	4.0
Stratum 1	5.1	1.0	1.0	2.9
Stratum 2	4.9	6.0	6.0	3.1
Adjusted (MH)	5.0	4.0	2.1	Not calculated
Example of:	____	____	____	____

Summary (Taking in to account when we have a third factor)

Before the study

1. Think of potential confounding factors
2. Collect accurate data on them

During Analysis

3. Conduct crude analysis
4. Stratify

5. Look for effect modification (Are the RR's or PR's different from each other?)
6. If effect modification – report, do not adjust
7. If confounding – do adjustment for controlling confounding factors

Exercise

The following are documents that must be distributed to the participants at the beginning of the exercise for their reference during discussion.

Food Consumption Histories, Foodborne Outbreak 1

Subject				Subject			
Number	Case?	Soup	Dumpling	Number	Case?	Soup	Dumpling
1	Y	Y	Y	25	N	Y	Y
2	Y	Y	Y	26	N	Y	Y
3	Y	Y		27	N	Y	Y
4	Y	Y	Y	28	N	Y	Y
5	Y	Y	Y	29	N	Y	Y
6	Y	Y	Y	30	N	Y	Y
7	Y	Y	Y	31	N	Y	Y
8	Y	Y	Y	32	N	Y	Y
9	Y	Y	Y	33	N	Y	Y
10	Y	Y	Y	34	N	Y	Y
11	Y	Y	Y	35	N	Y	Y
12	Y	Y	Y	36	N	Y	Y
13	Y	Y	Y	37	N	Y	Y
14	Y	Y	Y	38	N	Y	Y
15	Y	Y	Y	39	N	Y	Y
16	Y	Y	Y	40	N	Y	Y
17	Y	Y	Y	41	N	Y	Y
18	Y	Y	Y	42	N	Y	Y
19	Y	Y	Y	43	N	Y	Y

20	Y	Y	Y	44	N	Y	Y
21	Y	Y	N	45	N	Y	N
22	Y	N	Y	46	N	Y	N
23	Y	N	N	47	N	Y	N
24	Y	N	N	48	N	Y	N

Food Consumption Histories, Foodborne Outbreak 1

Participant Answer Sheet

Question 1. Analyze whether each food is associated with illness using the following table shells.

		Ill	Well	Total	Risk	Measure of Association
Ate Soup?	Yes			_______	_______	_______
	No			_______	_______	
	Total	_______	_______	_______		

		Ill	Well	Total	Risk	Measure of Association
Ate Dumpling?	Yes			_______	_______	_______
	No			_______	_______	
	Total	_______	_______	_______		

Question 2a. Stratify the analysis of dumplings and illness by soup. Calculate the stratum-specific measures of association.

Stratum 1 = ________________

		Ill	Well	Total	Risk	MoA	
Ate	Yes			______	______		aH_0/T = ______
						______	
______?	No			______	______		cH_1/T = ______
	Total	______	______	______			

Stratum 2 = ________________

		Ill	Well	Total	Risk	MoA	
Ate	Yes			______	______		aH_0/T = ______
						______	
______?	No			______	______		cH_1/T = ______
	Total	______	______	______			

MH MoA numerator = __________

MH MoA denominator = __________

MH MoA = __________

Crude MoA = __________

Question 2b. Calculate the Mantel-Haenszel measure of association.

Question 2c. Is there evidence of confounding?

Question 3a. Stratify the analysis of soup and illness by dumplings. Calculate the stratum-specific measures of association.

Stratum 1 = ________________

		Ill	Well	Total	Risk	MoA	
Ate	Yes			______	______	______	aH_0/T = ______

______?	No			______	______		cH_1/T = ______
	Total	______	______	______			

Stratum 2 = ________________

		Ill	Well	Total	Risk	MoA	
Ate	Yes			______	______	______	aH_0/T = ______
______?	No			______	______		cH_1/T = ______
	Total	______	______	______			

MH MoA numerator = __________

MH MoA denominator = __________

MH MoA = __________

Crude MoA = __________

Question 3b. Calculate the Mantel-Haenszelsummary measure of association.

Question 3c. Is there evidence of confounding?

Question 4a. Tabulate the data into the following 2-by-4 table. For the top three rows, calculate risks (attack rates) and measures of association (each row's risk compared with the bottom row's risk).

Soup?	Dumpling?	Ill	Well	Total	Risk	Measure of Association
Yes	Yes			______	______	______
Yes	No			______	______	______
No	Yes			______	______	______
No	No			______	______	Reference

Total ______ ______ ______

Question 4b. Is this outbreak consistent with only one food being the likely vehicle? If so, which one? If not, what is the evidence that both foods are likely vehicles?

ESTABLISHING A CAUSAL ASSOCIATION

Judgments of causality must first consider whether, for any individual study, the observed association is valid (i.e. whether the finding reflect the true relation ship between exposure and disease or may be explained by chance, bias , or confounding) and second , whether the accumulated evidence supports a cause effect relation ship. The validity of an observed association is established by eliminating alternative explanations of the association.

Association can be:

1. Artificial (spurious) associations (due to chance or bias)
2. Non causal (indirect) associations
   a. reverse causation (associated factor is an effect rather than a cause)
   b. Reciprocal causation (both cause and an effect) E.g. Vitamin A deficiency can cause diarrhea or diarrhea can cause Vitamin A deficiency.
   c. The association is due to a confounding effect by a third variable
2. Causal associations, which can be established only when other potential explanations of the association can be ruled out.

In the absence of experimental evidence the **Bradford - Hill** criteria are used to assess the strength of evidence for a cause and effect relationship.

The criteria's include:

1. Strength of the association

The stronger the association, the more likely that it is causal

2. Dose-response relationship

The risk of disease often increases with increasing exposure

with the causal agent

3. **Consistency of the relationship**

The same association should be demonstrated in studies with different methods, conducted by different investigators, and in different populations

4. **Temporal relationship:** the exposure to the factor must precede the onset of the disease

5. **Specificity of the association**

The association is more likely causal if a single exposure is linked to a single disease

6. **Biological plausibility (coherence)**

The finding of the study should be coherent with what is known about the biology & the descriptive epidemiology of the disease

7. **Prevention**

Eliminating the exposure should be followed by a decrease in the incidence rate of the disease

References

1. Hennekens CH, Buring JE. Epidemiology in Medicine, Lippincott Williams & Wilkins, 1987.

2. Breslow NE & Day NE. Statistical Methods in Cancer Research. Vol. 1: The Analysis of case control studies, IARC, 1980.

Session 8: VALIDITY AND RELIABILITY

Session overview

Measurement issues differ in the social sciences in that they are related to the quantification of abstract, intangible and unobservable constructs. In many instances, then, the meaning of quantities is only inferred.

Let us begin by a general description of the paradigm that we are dealing with. Most concepts in the behavioral sciences have meaning within the context of the theory that they are a part of. Each concept, thus, has an operational definition which is governed by the overarching theory. If a concept is involved in the testing of hypothesis to support the theory it has to be measured. So the first decision that the research is faced with is "how shall the concept be measured?" That is, the type of measure. At a very broad level, the type of measure can be observational, self-report, interview, etc. These types ultimately take shape of a more specific form like observation of ongoing activity, observing video-taped events, self-report measures like questionnaires that can be open-ended or close-ended, Likert-type scales, interviews that are structured, semi-structured or unstructured and open-ended or close-ended. Needless to say, each type of measure has specific types of issues that need to be addressed to make the measurement meaningful, accurate, and efficient.

Another important feature is the population for which the measure is intended. This decision is not entirely dependent on the theoretical paradigm but more on the immediate research question at hand.

A third point that needs mentioning is the purpose of the scale or measure. What is it that the researcher wants to do with the measure? Is it developed for a specific study or is it developed with the anticipation of extensive use with similar populations?

Once some of these decisions are made and a measure is developed, which is a careful and tedious process, the relevant questions to raise are "how do we know that we are indeed measuring what we want to measure?" since the construct that we are measuring is abstract, and "can we be sure that if we repeated the measurement we will get the same result?". The first question is related to validity and the second to reliability. Validity and reliability are two important characteristics of behavioral measure and are referred to as psychometric properties.

It is important to bear in mind that validity and reliability are not an all or none issue but a matter of degree.

Learning Objectives

After completing this lesson, you will be able to:

- Explain Validity and Reliability
- Discuss different types of Validity and Reliability
- Identify ways to maximize validity and reliability on measurements

8.1. Validity:

Very simply, validity is the extent to which a test measures what it is supposed to measure. The question of validity is raised in the context of the three points made above, the form of the test, the purpose of the test and the population for whom it is intended. Therefore, we cannot ask the general question "Is this a valid test?" The question to ask is "how valid is this test for the decision that I need to make?" or "how valid is the interpretation I propose for the test?" We can divide the types of validity into logical and empirical.

Content Validity:

When we want to find out if the entire content of the behavior/construct/area is represented in the test we compare the test task with the content of the behavior. This is a logical method, not an empirical one. Example: if we want to test knowledge on American Geography it is not fair to have most questions limited to the geography of New England.

Face Validity:

Basically, face validity refers to the degree to which a test appears to measure what it purports to measure.

Criterion-Oriented or Predictive Validity:

When you are expecting a future performance based on the scores obtained currently by the measure, correlate the scores obtained with the performance. The later performance is called the criterion and the current score is the prediction. This is an empirical check on the value of the test – a criterion-oriented or predictive validation.

Concurrent Validity:

Concurrent validity is the degree to which the scores on a test are related to the scores on another, already established, test administered at the same time, or to some other valid criterion available at the same time. Example: a new simple test is to be used in place of an old cumbersome one, which is considered useful; measurements are obtained on both at the same time. Logically, predictive and concurrent validation are the same, since (??) the term concurrent validation is used to indicate that no time elapsed between measures.

Construct Validity:

Construct validity is the degree to which a test measures an intended hypothetical construct. Many times psychologists assess/measure abstract attributes or constructs. The process of validating the interpretations about that construct as indicated by the test score is construct validation. This can be done experimentally, e.g., if we want to validate a measure of anxiety. We have a hypothesis that anxiety increases when subjects are under the threat of an electric shock, then the threat of an electric shock should increase anxiety scores (note: not all construct validation is this dramatic!)

A correlation coefficient is a statistical summary of the relation between two variables. It is the most common way of reporting the answer to such questions as the following: Does this test predict performance on the job? Do these two tests measure the same thing? Do the ranks of these people today agree with their ranks a year ago? (Rank correlation and product-moment correlation)

According to Cronbach, to the question “what is a good validity coefficient?” the only sensible answer is “the best you can get”, and it is unusual for a validity coefficient to rise above 0.60, though that is far from perfect prediction.

All in all we need to always keep in mind the contextual questions: what is the test going to be used for? How expensive is it in terms of time, energy and money? What implications are we intending to draw from test scores?

8.2. Reliability:

Research requires dependable measurement (Nunnally). Measurements are reliable to the extent that they are repeatable and that any random influence which tends to make measurements different from occasion to occasion or circumstance to circumstance is a source of measurement error (Gay). Reliability is the degree to which a test consistently measures whatever it measures.

Errors of measurement that affect reliability are random errors and errors of measurement that affect validity are systematic or constant errors.

Test-retest, equivalent forms and split-half reliability are all determined through correlation.

Test-retest Reliability:

Test-retest reliability is the degree to which scores are consistent over time. It indicates score variation that occurs from testing session to testing session as a result of errors of measurement. Problems: Memory, Maturation, Learning.

Equivalent-Forms or Alternate-Forms Reliability:

Two tests that are identical in every other way except for the actual items included. Used when it is likely that test takers will recall responses made during the first session and when alternate forms are available. [Correlate the two scores]. The obtained coefficient is called the coefficient of stability or coefficient of equivalence. Problem: Difficulty of constructing two forms that are essentially equivalent.

Both of the above require two administrations.

Internal Consistency Reliability:

Determining how all items on the test relate to all other items. Kudser-Richardson is an estimate of reliability that is essentially equivalent to the average of the split-half reliabilities computed for all possible halves.

Split-Half Reliability:

This requires only one administration, especially appropriate when the test is very long. The most commonly used method to split the test into two is using the odd-even strategy. Since longer tests tend to be more reliable, and since split-half reliability represents the reliability of a test only half as long as the actual test, a correction formula must be applied to the coefficient: Spearman-Brown prophecy formula.

Split-half reliability is a form of internal consistency reliability.

Rationale Equivalence Reliability:

Rationale equivalence reliability is not established through correlation but rather estimates internal consistency by determining how all items on a test relate to all other items and to the total test.

Standard Error of Measurement:

Reliability can also be expressed in terms of the standard error of measurement. It is an estimate of how often you can expect errors of a given size.

Summary

Assessment - Reliability and Validity

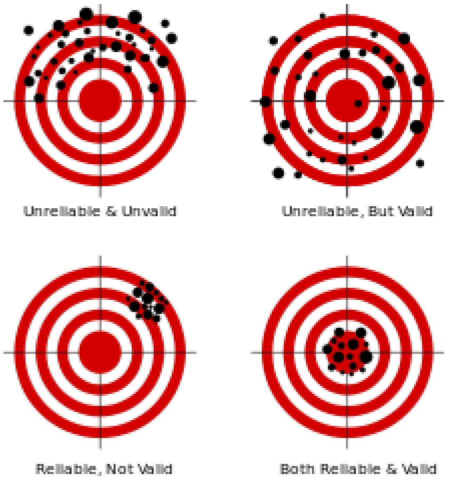


Fig Validity and reliability

Validity or face validity is defined as the degree to which the instrument measures what it's supposed to measure. If an instrument is not reliable over time, it cannot be valid, as results can vary depending upon when it is administered. An instrument can be: a)neither reliable nor valid, b) reliable and not valid, or c) both reliable and valid. However, an instrument must be reliable in order to be valid.

Although three primary approaches to test validity are reported by Mason and Bramble (1989), Patton (2002) details the associated sub-categorical types of measurement validity:

1. **Content validity**: Warrants that an overall sample of the content being measured is represented. Identification of the content must be accurately represented by the test items. A panel or grouping of content experts is typically consulted to identify a broad spectrum of content.
2. **Criterion Validity**: Targets the accuracy of a measure itself. Examining criterion validity is demonstrated by comparing the selected measure with another valid measure.
1. **Predictive validity**: Predicts a recognized association between the identified construct and something else. Typically, one measure occurs at an earlier time and is used to predict a later measure.
2. **Concurrent validity**: Exists when the identified measure positively correlates with a measure that has been previously found to be valid. The two measures could be for the same or different constructs that are related.
3. **Construct validity**: Ensures that the assessment measures the construct it claims to measure. Construct validity can be determined by demonstration of comparative test performance results (differential-groups study) or pre and post-testing of implementation of the construct (intervention study). This type of validity can also show how the measure relates to other measures as defined in the construct.

Discriminant Validity: Illustrates that measures that should not be related. A lack of correlation is expected to establish discriminant validity

Reliability determines how consistently a measurement of skill or knowledge yields similar results under varying conditions. If a measure has high reliability, it yields consistent results.

There are four principal ways to estimate the reliability of a measure:

1. **Inter-observer**: Is determined by the extent to which different observers or evaluators examine the same presentation, demonstration, project, paper, or other performance and agree on the overall rating on one or more dimensions.
2. **Test-retest**: Is determined by the extent to which the same test items or kind of performance evaluated at two different times yields similar results.

3. **Parallel-forms**: Is determined by examining the extent to which two different measurements of knowledge or skill yield comparable results.
4. **Internal Consistency Reliability:** Used to assess the consistency of results across items within a test
5. **Split-half reliability**: Is determined by comparing half of a set of test items with the other half and determining the extent to which they yield similar results.

Session 9: Screening Program and Evaluation

Session overview

Screening is the search for unrecognized disease or defect by means of rapidly applied tests, examinations or other procedures in apparently healthy individuals.

Learning objective

At the end of the chapter, the trainee will be able to

- Define screening, validity and reliability
- Calculate and interpret sensitivity and specificity of screening tests
- Calculate and interpret reliability measurements of screening tests
- Identify approaches of evaluating screening programs

Primary requirements for screening:

1) Early detection of disease leads to a more favorable prognosis due to early treatment, as compared to delayed treatment.

2) Pre-clinical disease left untreated typically progresses to clinically-evident disease (e.g. no spontaneous regression).

3) The disease should be serious (relates to cost effectiveness, ethics, and prognosis).

4) Prevalence of pre-clinical disease should be relatively high among those screened.

An ideal screening test should be

a. Inexpensive,

b. Easy to administer,

c. impose minimal discomfort on the patients,

d. Valid and reliable.

Validity of a screening test

How good is the screening test compared with the confirmatory diagnostic test?

Validity of a test is the ability to differentiate accurately between those who have the disease and those who do not.

Sensitivity and Specificity of a screening test

A. Sensitivity - is the ability of a test to identify correctly those who have the disease. The test will actually classify a diseased person as likely to have the condition.

B. Specificity - is the ability of a test to identify correctly those who do not have the disease.The test will actually classify a non-diseased person as unlikely to have the condition.

		Definitive diagnosis		
		Diseased	Non diseased	Total
Test result	Positive	TP (a)	FP (b)	TP+FP
	Negative	FN (c)	TN (d)	TN+FN
	Total	TP+FN	TN+FP	TP+FP+TN+FN

Sensitivity: The probability of testing positive if the disease is truly present

$$\text{Sensitivity} = a / (a + c)$$

$$= \frac{TP \quad X\ 100}{TP+FN}$$

Specificity: The probability of screening negative if the disease is truly absent

$$\text{Specificity} = d / (b + d)$$

$$= \frac{TN}{TN + FP} X100$$

Accuracy= a+b/a+b+c+d

Relationship between sensitivity & specificity:

1. Lowering the criterion of positivity results in increased sensitivity, but at the expense of decreased specificity.
2. Making the criterion of positivity more stringent increases the specificity, but at the expense of decreased sensitivity.
3. The goal is to have both high sensitivity and high specificity, but this is often not possible or feasible.
4. The decision for the cutpoint involves weighing the consequences of leaving cases undetected (false negatives) against erroneously classifying healthy persons as diseased (false positives).
5. In general, specificity must be at least 98%to be effective --- because misclassifying 2% of the population will create as many false positives as the sensitivity of the test will actually detect.
6. Sensitivity should be increased when the penalty associated with missing a case is high (e.g. minimize false negatives)
    - when the disease can be spread
    - when subsequent diagnostic evaluations are associated with minimal cost and risk
7. Specificity should be increased when the costs or risks associated with further diagnostic techniques are substantial (minimize false positives – e.g. positive screen requires that a biopsy be performed).

Predictive Value of a Screening Test

Predictive value is the ability of a test to predict the presence or absence of disease from test results.

1. **Predictive Value of a Positive Test** (PVPT) or Positive Predictive Value. PVPT shows the probability that the person tested positive by this specific test truly has the disease.

$$\text{PVPT} = \frac{\text{TP}}{\text{TP + FP}} \text{ X } 100\%$$

2. Predictive Value of a Negative Test (PVNT) or Negative Predictive Value. PVNT Shows the degree of confidence the disease can be ruled out by using this specific test.

$$\text{PVNT} = \frac{\text{TN}}{\text{TN+FN}} \text{ X } 100\%$$

Predictive value of a test is determined by **Sensitivity**, **Specificity** and the **Prevalence** of the disease.

- ✓ The higher the prevalence, the more likely it is that a positive test is predictive of the diseases i.e. PVPT will be high.
- ✓ The more sensitive a test, the less likely it is that an individual with a negative test will have the disease and thus the greater the predictive value negative.
- ✓ The more specific the test, the less likely an individual with a positive test will be to be free from the disease and the greater the predictive value positive.

Sensitivity and Specificity in Multiple Tests

- Multiple tests are commonly done in medical practice
- Choices of tests depend on cost, invasiveness, volume of test, presence and capability of lab infrastructure, urgency, etc.
- The tests can be done **sequentially** or **simultaneously**

1. Sequential Testing (Two-Stage Screening)

After the first (screening) test was conducted, those who tested **positive** were brought back for the second test to further reduce false positives. Consequently, the overall process will increase specificity but with reduced sensitivity.

Two-Stage Screening: Re-Screen the Positives from the First Test

Subject is disease positive when test positive in both tests

Subject is disease negative when test negative in either test

Net sensitivity = Sensitivity 1 x Sensitivity 2

Net specificity = Spec1 + Spec2 – (Spec1 x Spec2)

2. Simultaneous Testing

When two (or more) tests are conducted in parallel. The goal is to maximize the probability that subjects with the disease (true positives) are identified (increase sensitivity). Consequently, more false positives are also identified (decrease specificity).

When two tests are used simultaneously, disease positives are defined as those who test positive by either one test or by both tests

Net sensitivity = sens 1 + sens 2 – (sens 1 x sens 2)

When two tests are used simultaneously, disease negatives are defined as those who test negative by both tests

Net specificity = specificity test 1 x specificity test 2

Exercise:

- In a population of 1000, the prevalence of disease is 20%
- Two tests (A and B) are used at the same time
- Test A has sensitivity of 80% and specificity of 60%
- Test B has sensitivity of 90% and specificity of 90%
- Calculate net sensitivity and net specificity from using Test A and Test B simultaneously

Net Gain and Net Loss

- In simultaneous testing, there is a net gain in sensitivity but a net loss in specificity, when compared to either of the tests used

- In sequential testing when positives from the first test are retested, there is a net loss in sensitivity but a net gain in specificity, compared to either of the tests used

Reliability (Precision) of Screening Test

A reliable screening test is one that gives consistent results when the test is performed more than once on the same individual under the same conditions.

Two major factors affect consistency of results: the variation inherent in the method and observer variation (observer error).

1. The variability of a method- depends on such factors as the stability of the reagents used and fluctuation in the substance being measured (e.g in relation to meals, diurnal variation).
2. Observer variation- can stem from differences among observers (interobserver variation) and also from variation in readings by the same observer on separate occasions (intraobserver variation).

These variations can usually be reduced by:

1. Careful standardization of procedures

2. An intensive training period for all observers (or interviewers)

3. Periodic checks on their work

4. The use of two or more observers making independent observations.

Calculating agreement between two observers (or two observations)

Table: Agreement between two observers

		Observer 1		
		Positive	Negative	Total
Observer 2	Positive	a	b	a + b
	Negative	c	d	c + d
	Total	a + c	b + d	a+b+c+d

A perfect agreement occurs when b = 0 and c= 0

Overall Percent Agreement = a + d x 100/a + b + c + d

Percent Positive Agreement = a x 100/a + b + c

Kappa (Measures of reliability)

The range of Kappa is between -1 (perfect disagreement) through 0 (agreement by chance) and +1 (perfect agreement).

Kappa	Interpretation
> 0.80	Almost perfect
0.61 - 0.80	Substantial
0.41 – 0.60	Moderate
0.21 – 0.40	Fair
0.00 – 0.20	Slight
< 0.00	Poor

Evaluation of screening program

Evaluation of a screening program involves consideration of two issues:

1. Whether the program is feasible, and
2. Whether it is effective.

Both must be considered carefully. No matter how effective a screening procedure is in reducing subsequent morbidity and mortality, it will not be accepted if it can not be conducted efficiently, with minimal inconvenience and discomfort, and at a reasonable cost. Conversely, the implementation of a screening program, no matter how cost-effective, will not be warranted if it does not accomplish its goal of reducing morbidity and mortality.

Feasibility

The feasibility of a screening program is determined by a number of factors related to program performance, which measure the acceptability of the program to the potential screenees, cost-effectiveness, the subsequent diagnosis and treatment of individuals who test positive, and the yield of cases. The acceptability of the program can be measured by factors such as the number of persons examined and the proportion of the target population that is screened. The costs of the screening program must be considered in terms of total costs as well as with regard to resources expended per detected case of the disease. The successful screening program must also include provision for follow up of persons whose screening tests are positive. This can be measured by considering the proportions of those with positive tests who are followed, diagnosed, and treated.

The yield of the screening program should be high. Yield is the number of cases detected by the screening program.

$$\text{Yield} = \frac{\text{Persons with the disease detected by the test}}{\text{Total screened}} \text{X } 100$$

$$\text{Yield} = \frac{TP}{TP + FN + TN + FP} \text{X } 100$$

With respect to the yield, one measure that is commonly considered is the predictive value of a screening test. The predictive value of a screening test is determined not only by factors that determine validity of the test itself (i.e. sensitivity and specificity), but also by the prevalence of preclinical disease.

Effectiveness

The evaluation of the effectiveness of a screening program must be based on measures that reflect the impact of a program on the course of a disease. An effective screening program should result in reduction of morbidity, mortality and disability.

REFERENCES

Berk, R., 1979. Generalizability of Behavioral Observations: A Clarification of Interobserver Agreement and Interobserver Reliability. American Journal of Mental Deficiency, Vol. 83, No. 5, p. 460-472.

Cronbach, L., 1990. Essentials of psychological testing.Harper& Row, New York.

Carmines, E., and Zeller, R., 1979.Reliability and Validity Assessment.Sage Publications, Beverly Hills, California.

Gay, L., 1987. Eductional research: competencies for analysis and application. Merrill Pub. Co., Columbus.

Guilford, J., 1954. Psychometric Methods. McGraw-Hill, New York.

Nunnally, J., 1978. Psychometric Theory. McGraw-Hill, New York.

Winer, B., Brown, D., and Michels, K., 1991.Statistical Principles in Experimental Design, Third Edition. McGraw-Hill, New York

Roberts B. Biographical research [Internet]. Open University Press Buckingham; 2002 [cited 2013 Nov 13].

Miller ML. Reliability and validity in qualitative research [Internet]. Sage; 1986 [cited 2013 Nov 13].

Baskerville RL, Wood-Harper AT.A critical perspective on action research as a method for information systems research.Journal of Information Technology. 1996;11(3):235–46.

Bryman A. Social research methods [Internet]. Oxford university press; 2012 [cited 2013 Nov 13].

Part 2:

INFERENTIAL BIOSTATISTICS

1. Estimation and hypothesis testing
2. Correlation and Regression

Session 1: Statistical Inference

Session overview

This session gives overview of the inferencialstatisticsandtesting hypothesis. Differenciat point and interval estimation. How to use estimation in public health research

Learning Objectives

Completing this section, participants should be able to

- Define terms in inferential statistics
- Explain the difference between sample statistic and population parameter,
- Differentiate the symbols and abbreviations used in inference,
- Describe the sampling distribution of a statistic and define the standard error of a statistic,
- Compare and contrast point estimation and interval estimation,
- Construct confidence intervals to population values
- List and explain the steps in hypothesis testing,
- State how to decrease the probability of Type I and Type II errors,
- Understand the purpose of statistical hypothesis testing,
- Explain the difference between statistical significance and practical importance,
- Statestatisticalhypothesisand conduct testing

1.1.Sampling Distributions of Statistics

Population (or process) is the object of interest for which we would like to make inference. Due to limited resource and time, it is usually impossible to know every aspect of the population. Instead, we obtain a (random) sample from the population and base our inference on the sample results. When we take a sample of size n from a population and calculate summary statistics like the sample mean $(\overline{X})$, the sample median (Med), the sample variance (s^2), the sample standard deviation (s), or the sample proportion $(\hat{p})$, we must realize that these quantities will differ and hence are themselves random variables.

Any random variable in statistics has a probability distribution. We have been talking about two common probability distributions in probability section of this module. When X = # of “successes” in n independent trials, we used the binomial distribution to talk about X probabilistically and when X was continuous and had an approximate bell-shaped distribution we used the normal distribution to calculate probabilities and quintiles associated with X. Because the summary statistics discussed above are random variables they also have a probability distribution that determines the likelihood of certain values of these statistics being obtained. The distribution of a summary statistic, e.g. the sample mean $(\bar{X})$, is called the **sampling distribution of statistic**.

For example, we naturally use sample mean $\bar{X}$ to estimate the population mean $(\mu_{\bar{x}})$. Since $\bar{x}$ was obtained from a sample , we are not guaranteed to get the same value for $\bar{x}$ if we conduct the same experiment (to obtain the data) again. So $\bar{x}$ can be viewed as a random variable , thus has a distribution. This distribution is called the sampling distribution of $\bar{x}$.

Hence, sampling distribution of a statistic is the distribution of all possible values of the statistic, computed from samples of the same size randomly drawn from the same population. It is the probability distribution of any particular sampling statistic.

Probability computation and statistical inference regarding the statistic under consideration are based on the sampling distributions of that statistic. Often, researchers base their inference regarding population parameters and probabilities of obtaining the sample results using a single sample. This requires theoretical understanding regarding the sample statistic when repeated samples of same size are drawn from the same population for the given statistic. Ultimately the information from the distribution of statistics will be used to estimate the precision of estimates associated with taking a specific sample from the given population under consideration.

Sampling distribution of a statistic can be obtained using the following procedure:

- From a population of size N, randomly draw all possible samples of size n.
- Compute the statistic of interest for each sample.
- Create a frequency distribution of the statistic.

Properties of sampling distributions

If you continue to take samples of data and compute every possible combination of samples (i.e. all permutations or combinations) of size n, then the sample statistics/point estimators can have their own distribution. As it might have been remembered, we have computed the mean, standard deviation and drawn the shape of binomial and normal distributions. Similarly, each sample statistic/point estimator will have its own distributions with its own mean, variance, and standard deviation. Hence, we will also discuss the three elements of the sampling distribution of the statistic under consideration as it is a frequency distribution by itself. We are interested in the mean, standard deviation and shape of the graph of a sampling distribution.

Generally, the sampling distribution of a statistic has the following properties:

1. The mean of the sampling distribution of the sample mean is equal to the population parameter, **unbiased estimator**
2. The standard deviation of the sampling distribution of the statistic is called standard error of the statistic. This is equivalent to the population standard deviation divided by the sample size.
3. The shape of the sampling distribution of the statistic normal provided the size of the sample is large due to the central limit theorem.

In this module, we will look at four types of sampling distributions

a. Distribution of the sample mean

b. Distribution of the difference between two means

c. Distribution of the sample proportion

d. Distribution of the difference between two proportions

A. Sampling distribution of sample mean ($\bar{x}$)

Given a finite population with mean (μ) and variance (σ^2), when sampling from a normally distributed population, it can be shown that the distribution of the sample mean will have the following properties.

1. The distribution of $\bar{x}$ will be normal.
2. The mean, $\mu_{\bar{x}}$, of the distribution of the values of $\bar{x}$, will be the same as the mean of the population from which the samples were drawn; $\mu_{\bar{x}} = \mu$.

3. The variance, $\sigma^2_{\overline{x}}$, of the distribution of $\overline{x}$, will be equal to the variance of the population divided by the sample size; $\sigma^2_{\overline{x}} = \sigma^2/n$.

Standard error

The square root of the variance of the sampling distribution of $\overline{x}$ is called the standard error of the mean; $SE(\overline{X}) = \sqrt{\sigma^2{}_{\overline{x}}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$

Sampling from non-normally distributed populations: When the sampling is done from a population which is not normally distributed, the central limit theorem is used.

The central limit theorem: Given a population which is not normally distributed with mean, μ, and variance, σ^2, the sampling distribution of $\overline{x}$, computed from samples of size n from this population will have mean, μ, and variance, σ^2/n, and will be approximately normally distributed when the sample is large (30 or higher). In general, whatever the distribution of the population from which the samples are drawn, the sampling distribution a statistic is normal provided that the sample size is large.

The Sampling Distribution of a Sample Mean (σ unknown)

The distribution of a sample mean when the population standard deviation σ is not known can be represented by a **t-distribution** (sometimes called the **Student's t-distribution**) with mean μ, standard deviation $s/\sqrt{n}$, and degrees of freedom n – 1 where n is the sample size. Then the t-statistic is given by: $t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$ (df = n – 1)

B. Sampling distribution of the difference between two means ($\overline{x}_1$ - $\overline{x}_2$)

Researchers may be interested to compare two population means. Knowledge of the sampling distribution of the difference between two means is useful in studies of this type as there is a need for computing the standard error of the difference between sample means in the subsequent estimation and hypothesis testing.

Given two normally distributed populations with means, $\mu_1$and μ_2, and variances, σ_1^2 and σ_2^2, respectively, the sampling distribution of the difference, $\overline{x}_1 - \overline{x}_2$, between the means of independent samples of size $n_1$and n_2 drawn from these populations is normally distributed with mean, $\mu_1 - \mu_2$, and variance, $\left(\sigma_1^2 / n_1\right) + \left(\sigma_2^2 / n_2\right)$. It is generally assumed that the two populations are normally distributed. This procedure is valid even when the population variances are different or when the sample sizes are different.

Plotting sample differences against frequency gives a normal distribution with mean equal to $\mu_1 - \mu_2$ which is the difference between the two population means.

Standard error of the sampling distribution of $\bar{x}_1$ - $\bar{x}_2$

The variance of the distribution of the sample differences is equal to $(\sigma_1^2 / n_1) + (\sigma_2^2 / n_2)$.

Therefore, the standard error of the differences between two means would be equal to $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1}}$.

C. Sampling distribution of the sample proportion $(\hat{p})$

While statistics such as the sample mean are derived from numeric or quantitative variables, the sample proportion is derived from categorical variables that are summarized using counts, proportions or frequency distributions. We will represent the sample proportion by $(\hat{p})$ and the population proportion by P. If we take repeated samples of size n from a variable that follows the Binomial distribution (i.e. the outcome is 0 or 1), and calculate $\hat{p}$=m/n for each of the samples (m=total count of successes), if n is large enough, then $\hat{p}$ will follow a norma l distribution (by the central limit theorem)

Procedures for obtaining sampling distribution of sample proportion

Construction of the sampling distribution of the sample proportion is done in a manner similar to that of the mean and the difference between two means.

Properties of the sampling distribution of sample proportion $(\hat{p})$ **: mean and variance and shape**

The mean of the distribution, $\mu_{\hat{p}}$, will be equal to the true population proportion, π, and the variance of the distribution, $\sigma_{\hat{p}}^2$, will be equal to $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. When the sample size is large, the distribution of the sample proportion is approximately normally distributed because of the central limit theorem.

D. Sampling distribution of the difference between two sample proportions $(\hat{p}_1 - \hat{p}_2)$

This is for situations with two population proportions. The appropriate distribution is the distribution of the difference between two sample proportions. This is a situation where researchers are interested to assess the probability associated with a difference in proportions computed from samples drawn from each of these populations.

Procedure to obtain sampling distribution of the difference between two sample proportions $(\hat{p}_1 - \hat{p}_2)$: The sampling distribution of the difference between two sample proportions is constructed in a manner similar to the difference between two means. Independent random samples of size n_1 and $n_2$are drawn from two populations of binary variables where the proportions of observations with the character of interest in the two populations are $\hat{p}_1$ and $\hat{p}_2$, respectively.

Properties of the sampling distribution of the difference between two sample proportions: mean, variance and shape

The mean, $\mu_{\hat{p}_1 - \hat{p}_2} = \hat{p}_1 - \hat{p}_2$, and variance, $\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$. These are true when $n_1$and $n_2$are large. The distribution of the difference between two sample proportions, $(\hat{p}_1 - \hat{p}_2)$, is approximately normal.

Estimation and Hypothesis Testing

Estimation and hypothesis testing are applications of sampling distribution of a statistic. Before embarking on performing the actual inferential process (a hypothesis testing or construction of confidence interval), one should entertain the following questions:

1) Are the data categorical or quantitative?
2) How many samples?
3) Are the data independent or dependent?
4) What **exactly** are we trying to learn?
5) About which populations do we wish to make an inference?

6) What are the conditions (assumptions) and how do we check them? The later part of this question may not be addressed in this module.

Estimation

Estimates and Estimators

An estimator is a description of the procedure on the how to determine a numerical value from any sample to estimate a certain population parameter, whereas an estimate is the actual numerical value obtained from a particular sample also known as statistic. Note that the word ‘estimate’ is often used in both senses, and there is no harm in this provided you understand the difference.

Types of estimates

There are two types of estimates such as point and interval estimates.

A. **Point Estimation** – this is using the data from the sample to compute the value of a sample statistic. This is what serves as an estimate for the population parameter.

Point estimator – the estimate of the population parameter; an example is $\bar{x}$ a point estimator for μ.

Sampling error – is the absolute value of the difference between the point estimate and the actual population parameter.

Formula: sampling error = | point estimate – population parameter |

Properties of estimates

There are obviously many possible estimators to choose from to estimate a certain population parameter. For instance, we usually estimate the population mean by the sample mean; and we often estimate the population variance by the sample variance. But, why not we, for example, estimate the population mean by the average of the smallest and largest values in our sample? Why do we choose the first procedure? The choice among these estimators is based on those that have certain ‘good’ properties. In this module, we will see the three important properties of estimates such as unbiased, relatively efficient and consistency properties.

Unbiasedness

In general we say that a statistic is unbiased if its expected value converges to the population value (parameter). If sample mean $(\bar{x})$ approaches the population mean (μ), then it is an unbiased estimator. If the sample standard variance (s^2) approaches the population standard variance (σ^2), then it is also an unbiased estimator.

In general, when we have a random sample, the sample mean is an unbiased estimator of the population mean and the sample variance is an unbiased estimator of the population variance; but the sample standard deviation is not an unbiased estimator of the population standard deviation.

Relatively efficient: Minimum Variance Estimate (MVE)

We should like our estimate to be close to the parameter being estimated as often as possible. It is not very helpful for an estimate to be correct 'on average' if it fluctuates widely from sample to sample. Thus, for an estimator to be good, its distribution should be concentrated fairly closely about the true value of the parameter of interest; in other words, if an estimator is unbiased the variance of the estimator's distribution should be small. If we have a choice among several estimators that are competing, we might proceed by eliminating any that are biased and then, from among those that are unbiased, choose the one with the smallest variance. Such an estimator is called a minimum variance unbiased estimator, or an efficient estimator. The estimate with the minimum variance is called minimum variance estimate (MVE).

Since the sampling distribution of $\bar{x}$ produces the smallest variances estimate of all possible other values that could estimate the mean (like median, mode, or any estimator). So we way it is MVE or the minimum variance estimator.

It can be shown mathematically that if the underlying population is normally distributed, the sample mean and the sample variance are minimum variance unbiased estimators of the population mean and the population variance, respectively. In other situations, however, the sample descriptive statistic may not have this property.

Consistency

As the sample size increases, the sample estimator approaches to the population parameter.

Law of Large Numbers – if we draw observations from a population with a finite mean μ at random, as we increase the number of observations we draw the value of the sample mean ($\bar{x}$) gets closer and closer to the population mean. Note that this makes sense because as you increase the size of your sample it gets closer to the size of the population. So it begins to look more and more like the population itself. For this reason the mean should approach the population mean.

Other properties are also of interest, but the important thing is to realize that criteria exist for evaluating estimators. One estimator may be preferred in one situation and a second, competing estimator may be preferred in another situation. There is a tendency to use estimators whose good properties depend on the data following a normal distribution, even though the data are not normally distributed. This is unfortunate because more appropriate methods of analysis are available. You cannot hope to learn all of the considerations that must be made in choosing estimators that are appropriate for each situation that might arise, but you should be aware of the necessity of examining such issues.

B. Interval estimate or confidence interval

Interval estimate is an estimate with a certain level of confidence. Level of confidenceis the probability of obtaining the population parameter within the error margin. It is denoted as **$(1-\alpha)$ x 100%** and can never be 100%! The general formula for an interval estimate is: interval estimate or confidence interval = Point Estimator ± Margin of Error.

Estimate ± Reliability Coefficient x SE of estimate

The product of the reliability coefficient and SE of estimate is called **margin of error**. It is a measure of precision and it is obtained as a product of reliability coefficient (table value from probability distributions) corresponding to yourconfidence and standard error of the estimate (obtained from its sampling distribution).

Confidence intervals interpretation

- The probability that the interval contains the true population parameter is $(1-\alpha)100\%$
- If we were to select 100 random samples from the population and calculate confidence intervals for each, approximately 95 of them would include the true population mean μ (and 5 would not)

Factors Affecting Confidence Interval

Sample size and level of confidence affects the width precision of the confidence interval.

Decreased confidence level is associated with narrow confidence interval.

Confidence level	**α**	**$Z_{\alpha/2}$**
99%	.01	2.58
95%	.05	1.96
90%	.10	1.64

Narrower confidence interval implies large sample size and less sampling error and hence highly precise estimate.

Estimation for the Population Mean (μ)

Confidence interval for mean

The (1-α) x 100% Confidence Interval for the population mean (μ) is:

$$\underbrace{\bar{x} - Zx\frac{\sigma}{\sqrt{n}}}_{a} \le \mu \le \underbrace{\bar{x} + Zx\frac{\sigma}{\sqrt{n}}}_{b} = (a \le \mu \le \mathrm{b})$$

If your confidence level is 95%, then the confidence interval for the population mean becomes:

$$(\bar{x} - 1.96x\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96x\frac{\sigma}{\sqrt{n}}) = (a, \mathrm{b})$$

Interpretation: Pr (a ≤ μ≤ b) = 0.95

Confidence interval for mean, σ unkown

The (1-α) x 100% confidence interval for the population mean is:

$$\bar{x} - t_{critical} \cdot \frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{critical} \cdot \frac{s}{\sqrt{n}}$$

Example 1: Assume a researcher wants to estimate the mean Hb level of adult male population in Gondar. He studied 100 adult males. Previous studies shows that mean Hb level of Gondar population is 12gm%. Average Hb level in 100 adult males is 11.5gm% with SD 1gm%.

Calculate 95% CI of average Hb level of adult male population in Gondar and interpret the results.

Solution 1:

$$95\% = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}} = 11.5 \pm 1.96 \times \frac{1}{\sqrt{100}} = 11.5 \pm 1.96 \times 0.1 = 11.5 \pm 0.196 = (11.30, 11.69)$$

95% CI =11.30-11.69, we are 95% confident that average Hb level of adult male population lie in the range of 11.30-11.69 gm%. Also since this range does not contain 12gm% therefore mean Hb level of Gondar is different from 12gm%).

Example 2: A random sample of 100 subjects with family history of diabetes was selected and their mean fasting blood sugar is 100mg/dl with S.D. of 0.06mg/dl. It is known from previous study that mean fasting blood sugar level in persons without history of diabetes is 95 mg/l. Calculate 95% CI for mean fasting blood sugar with history of diabetes and interpret the results.

Solution.2:

$$95\% = \overline{X} \pm Z \times \frac{\sigma}{\sqrt{n}} = 100 \pm 1.96 \times \frac{0.06}{\sqrt{100}} = 100 \pm 1.96 \times 0.006 = 100 \pm 0.012 = (99.998, 100.012)$$

Interpretation

The probability of obtaining the population mean fasting blood sugar for persons without history of diabetes within 99.998 to 100.012 is 95%.

Example 3: Suppose we are trying to estimate the birth weight of infants born to women who smoke during pregnancy. A sample of n = 73 women who smoked during pregnancy and the birth weight of their baby was obtained yielding a sample mean of $\overline{X} = 6.08\,\text{lbs}$. Construct a 95% plausible interval for the population mean birth weight of infants born to women who smoke during pregnancy.

Solution 3: The central limit theorem states that if our sample size (n) is sufficiently large, then $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ which also says by standardizing that $Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

This means that when we collect our data the probability our observed sample mean will fall within two standard errors of the mean is approximately 0.95 or a 95% chance, or more precisely.

$$P(-2 < Z < 2) = P(-2 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < 2) = P(-2 \cdot \frac{\sigma}{\sqrt{n}} < \overline{X} - \mu < 2 \cdot \frac{\sigma}{\sqrt{n}})$$

$$P(\mu - 2\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 2\frac{\sigma}{\sqrt{n}}) = .9544$$

To make this 95% exactly, we simply use 1.96 in place of 2.00 in the expression above, because Pr (-1.96 < Z < 1.96) = 0.9500. For 99% confidence we use 2.57 and for 90% we use 1.64 in place of 1.96.

Starting with the statement, $P(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = .9500$, we can perform similar algebraic manipulations to those above to isolate the population mean μ in the middle of the inequality instead. By doing this we will obtain an interval that has an approximate 95% chance of covering the true population mean (μ).

This says that the interval from $\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ up to $\bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$ has a 95% chance of covering the true population mean μ. This interval is simply the sample mean plus or minus roughly two standard errors. However, this interval cannot be calculated in practice! WHY?

The problem is that the distribution of $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is not a standard normal, i.e. N(0,1)!!!

If the population we are sampling from is approximately normal then

$\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has a **t-distribution** with **degrees of freedom**df = n – 1.

Example 4:

The following Data represent body weights of 18 diabetics expressed as a percentage of ideal. Thus, a value of 100 represents ideal body weight, a value of 120 represents 120% of ideal body weight (i.e., 20% overweight), and so on (Pagano &Gauvreau, 1993, p. 208). Data are: {107, 119, 99, 114, 120, 104, 88, 114, 124, 116, 101, 121, 152, 100, 125, 114, 95, 117}.

Solution 4: The sample mean is the point estimator of expected value μ.
A (1- α) x100% confidence interval for μ is calculated with the formula : $\bar{x} \pm (t_{(n-1,1-a/2)})$(Std Err).
Where, $t_{(n-1,1-a/2)}$ represents the (1 - a/2) percentile of a t distribution with n - 1 degrees.
The 95% CI = 112.778 ± ($t_{17,1-.05/2}$)(14.424/sqrt(18)) = 112.778 ± (2.11)(3.400)
= 112.778 ± 7.174 = **(105.6, 120.0).**

Confidence Intervals for the difference between two population means: (μ_1- μ_2)

Large Sample (n ≥ 30) and known Variances

i. Point Estimation of (μ_1- μ_2) is: ($\bar{x}_1$ - $\bar{x}_2$)
ii. A (1- α) 100% confidence interval for (μ_1- μ_2)

$$(1-\alpha)100\%\ \text{CI for}\ (\mu_1 - \mu_2) = (\bar{x}_1 - \bar{x}_2) \pm \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For small sample (n < 30) and Unknown Variances (σ_1^2 and σ_2^2)

If population variances are unknown, they can be approximated by the sample variances: s_1^2 and s_2^2. Accordingly, the standard error of the sample mean difference will be:

$$\mathrm{SE}(\overline{x_1}-\overline{x_2})=\sigma_{(\overline{x_1}-\overline{x_2})}=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

With degrees of freedom given by: $df'=\dfrac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{(s_1^2/n_1)^2/(n_1-1)+(s_2^2/n_2)^2/(n_2-1)}$

Pooled variance (S^2)

As to the question which s to use, we pool the information from two samples and get the so-called "pooled estimator of population variance"

$$s_p^2=\frac{\sum_{i=1}^{n_1}(x_{1i}-\overline{x_1})^2+\sum_{i=1}^{n_2}(x_{2i}-\overline{x_2})^2}{n_1+n_2-2}=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$$

S_p^2 is called pooled estimator of population standard deviation

Hence, equality of variance is assumed, then the SE become:

$$\mathrm{SE}(\overline{x_1}-\overline{x_2})=\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}=S_p x\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Example 5:

Researchers wish to know if the data they have collected provide sufficient evidence to indicate a difference in mean serum uric acid levels between normal individuals and individuals with mongolism. The data consist of serum uric acid readings on 12 mongoloid individuals and 15 normal individuals . The means are $\bar{x}_1$ = 4.5 mg/100 ml and $\bar{x}_2$ = 3.4 mg/100 ml. The data constitute two independent simple random samples each drawn from a normally distributed population with a variance equal to 1 mg/100 ml. Construct a 95% CI for the difference in mean serum uric acid levels between the two populations.

Solution 5:

i. Point estimate: $\overline{x_1}-\overline{x_2}$ = 4.5-3.4 = 1.1

ii. 95% confidence interval for $\mu_1-\mu_2$

$$(\overline{x_1}-\overline{x_2})\pm z\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}=(4.5-3.4)\pm 1.96\sqrt{\frac{1}{12}+\frac{1}{15}}=1.1\pm 0.8=(0.3,1.9)$$

Example 6:

A research team collected serum amylase data from a sample of healthy subjects and from a sample of hospitalized subjects. They wish to know if they would be justified in concluding that the population means are different. The data consist of serum amylase determinations on n_2=15 healthy subjects and n_1=22 hospitalized subjects . The sample means and standard deviations are as follows : $\bar{x}_1$ = 120 units/ml, s_1 = 40 units/ml, $\bar{x}_2$= 96 units/ml, s_2 = 35 units/ml. Construct a 95% CI for the difference between the two population mean serum amylase.

Solution 6:

$$s_p^2 = \frac{(22-1)(40^2)+(15-1))(35^2)}{22+15-2} = 1450$$

$$95\%\ \text{CI for}\ \mu_1-\mu_2 = (\overline{x_1}-\overline{x_2}) \pm t_{(\alpha,\, n_1+n_2-2)}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}} = (\overline{x_1}-\overline{x_2}) \pm t_{(\alpha,\, n_1+n_2-2)} \times s_p \times \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= (120-96) \pm 2.0301 \times 1450 \times \sqrt{\frac{1}{15}+\frac{1}{22}} = 24 \pm 2.0301 \times 12.75 = 24 \pm 25.884 = (-1.884, 49.884)$$

Estimation about a Population Proportion (P)

Proportion – this is the percentage that our population takes on a certain characteristic.

$\hat{p}$ = number of successes / total individuals

$\hat{p}$ = this is the sample proportion and is designated at p-hat. It is an actual calculated value.

Many of the techniques and statistics that we have used in estimating mean will be used again. The central limit theorem for proportions states that if our sample size (n) is sufficiently large, then $\hat{p} \sim N(P, \sqrt{\frac{P(1-P)}{n}})$.

This means that when we take our sample and find our sample proportion, $\hat{p}$, the probability our observed sample proportion will fall within approximately two standard errors of the population proportion is roughly 95%, or more precisely

$$P(P-1.96\cdot\sqrt{\frac{P(1-P)}{n}} < \hat{p} < P+1.96\cdot\sqrt{\frac{P(1-P)}{n}}) = .9500 \leftarrow \text{Recall: } P(-1.96 \le Z \le 1.96) = .9500$$

Starting with this statement we can perform some algebraic manipulations to isolate the population proportion, P, in the middle of the inequality above. By doing this we will see that the resulting interval will have a 95% chance of covering the true population proportion (P).

$$\underbrace{\hat{p} - Zx\sqrt{\frac{P(1-P)}{n}}}_{a} \le P \le \underbrace{\hat{p} + Zx\sqrt{\frac{P(1-P)}{n}}}_{b} = (a \le \pi \le \mathrm{b})$$

This says that the interval from $\hat{p} - 1.96 \cdot \sqrt{\frac{P(1-P)}{n}}$ up to $\hat{p} + 1.96 \cdot \sqrt{\frac{P(1-P)}{n}}$ has a 95% chance of covering the true population proportion π. This interval is simply the sample proportion plus or minus roughly two standard errors, i.e. $\hat{p} \pm 1.96 \cdot SE(\hat{p})$. However, this interval cannot be calculated in practice! WHY?

The problem is in reality the population proportion P is NOT known. If π were known we would not be conducting a study in first place!

$$\text{Margin of Error} = Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} \pm (\mathrm{Re}\,liability\ \text{Coeffivient}) \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \text{or} \qquad \hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example 7: Consider the problem of estimating the prevalence of malignant melanoma in 45- to 54-year-old women in the United States. Suppose that a random sample of n = 5000 women is selected from this age group and X = 500 women are found to have the disease.

A. Calculate the point estimator of the prevalence of malignant melanoma for the population of women in this age range.
B. Construct a 95% CI for the population prevalence of malignant melanoma for the population of women in this age range.

Solution 7:

A. point estimate for the prevalence of this disease is:
   p̂ = 500/5000 = 0.1 = 10%
B. 95% CI for the population prevalence of malignant melanoma for the population of women in this age range:

$$95\%\ \mathrm{CI} = \hat{\mathrm{p}} \pm \mathrm{Z} \times \sqrt{\frac{p(1-p)}{n}} = \hat{\mathrm{p}} \pm \mathrm{Z} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.1 \pm 1.96 \times \sqrt{\frac{0.1(1-0.1)}{5000}}$$

$$= 0.1 \pm 1.96 \times 0.0042 = 0.1 \pm 0.0083 = (0.0917, 0.1083)$$

Example 8:

The lead level in a child’s body is considered to be dangerously high if it exceeds 30 micrograms per deciliter. A random sample of 1000 of 20,000 children living in public housing projects in a particular city revealed that 200 of them had dangerously high lead levels in their bodies.

Construct a 99% confidence interval for the true population proportion of children who had high lead levels in their bodies.

Solution 8:

Point estimator (p̂) = m/n=200/1000= 0.2

$$99\%\,\mathrm{CI} = \hat{p} \pm Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.2 \pm 2.58\sqrt{\frac{0.2 \times 0.8}{1000}} = 0.2 \pm 2.58 \times 0.013 = 0.2 \pm 0.033 = (0.167, 0.233)$$

Confidence interval for the difference between two proportions: $\pi_1 - \pi_2$

Assumption: It is assumed that the sampling distribution of P_1 - P_2 is approximately normally distributed.

Standard error (SE): $\mathrm{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

95% CI for the difference between two proportions (RD)

1) Compute the risk difference RD, $(\hat{p}_{\mathrm{Exposed}} - \hat{p}_{\mathrm{Unexposed}})$.

2) Compute SE(RD) = $SE(\hat{p}_{\mathrm{Exposed}} - \hat{p}_{\mathrm{Unexposed}}) = \sqrt{\frac{\hat{p}_{\mathrm{Exposed}}(1-\hat{p}_{\mathrm{Exposed}})}{n_{Exposed}} + \frac{\hat{p}_{\mathrm{Unexposed}}(1-\hat{p}_{Unexposed})}{\mathrm{n}_{\mathrm{Unexposed}}}}$

3) Find $(\hat{p}_{\mathrm{Exposed}} - \hat{p}_{\mathrm{Unexposed}}) \pm 1.96 \cdot SE(\hat{p}_{\mathrm{Exposed}} - \hat{p}_{\mathrm{Unexposed}})$ to obtain (LCL, UCL)

4) Interpret the CI

Example 9: consider a randomized study with n=50 in each group.

	Exposed	Unexposed	Total
Diseased	19	10	29
Not diseased	31	40	71
Total	50	50	100

$$\mathrm{RD} = \hat{p}_{\mathrm{Exposed}} - \hat{p}_{Unexposed} = \frac{a}{a+c} - \frac{b}{b+d} = \frac{19}{50} - \frac{10}{50} = 0.38 - 0.20 = 0.18$$

$$\mathrm{SE(RD)} = \mathrm{SE}(\hat{p}_{\mathrm{Exposed}} - \hat{p}_{\mathrm{Unexposed}}) = \sqrt{\frac{\hat{p}_{\mathrm{Exposed}}(1-\hat{p}_{\mathrm{Exposed}})}{\mathrm{n}_{\mathrm{Exposed}}} + \frac{\hat{p}_{\mathrm{Unexposed}}(1-\hat{p}_{\mathrm{Unexposed}})}{\mathrm{n}_{\mathrm{Unexposed}}}}$$

$$\sqrt{\frac{\mathrm{ac}}{(\mathrm{a}+\mathrm{c})^3} + \frac{\mathrm{bd}}{(\mathrm{b}+\mathrm{d})^3}} = \sqrt{\frac{19*31}{(50)^3} + \frac{10*40}{(50)^3}} = 0.08895$$

Exercise

1. We wish to estimate the mean serum indirect bilirubin level of 4-day-old infants. The mean for a sample of 16 infants was found to be 5.98 mg/dl. Assuming bilirubin levels in 4-day-old infants are approximately normally distributed with a standard deviation of 3.5 mg/dl find:

a) The 90% confidence interval for μ
b) The 95% confidence interval for μ
c) The 99% confidence interval for μ

2. In a study of preeclampsia, Kaminski and Rechberger found the mean systolic blood pressure of 10 healthy, nonpregnant women to be 119 with a standard deviation of 2.1.(Preeclampsia: Development of hypertension, albuminuria, or edema between the 20th week of pregnancy and the first week postpartum. Eclampsia: Coma and/or convulsive seizures in the same time period, without other etiology.)

a) What is the estimated standard error of the mean?

b) Construct the 99% confidence interval for the mean of the population from which the 10 subjects may be presumed to be a random sample.

c) What is the precision of the estimate?

d) What assumptions are necessary for the validity of the confidence interval you constructed?

3. A research study obtained data regarding sexual behavior from a sample of unmarried men and women between the ages of 20 and 44 residing in geographic areas characterized by high rates of sexually transmitted diseases and admission to drug programs. Fifty percent of 1229 respondents reported that they never used a condom. Construct a 95 percent confidence interval for the population proportion never using a condom.

4. A study of teenage suicide included a sample of 96 boys and 123 girls between ages of 12 and 16 years selected scientifically from admissions records to a private psychiatric hospital. Suicide attempts were reported by 18 of the boys and 60 of the girls. We assume that the girls constitute a simple random sample from a population of similar girls and likewise for the boys. Construct a 99 percent confidence interval for the difference between the two proportions.

5. Early-Stage Breast Cancer Treatment Method and Age

In a sample of n = 658 women who underwent a partial mastectomy and subsequent radiation therapy contains 292 women under 55, which is a sample percentage of 44.4%. Find a 95% CI for the true proportion of women under 55 in this population.

6. In a sample of n = 1580 women who received a modified radical mastectomy 397 women were under 55, which is a sample percentage of 25.1%. Find a 95% CI for the true proportion of women under 55 in this population.Do these intervals suggest that the proportion of women under the age of 55 differs significantly for these two courses of treatment of early-stage breast cancer?

Solutions on exercises on 1:

a) At 90% confidence level (z = 1.645)

5.98 ± 1.645 (.875)= 5.98-1.439375, 5.98+1.439375 = (4.5408, 7.4129)

b) At 95% confidence level (z = 1.96)

5.98 ± 1.96 (.875) = (4.265, 7.695)

c) At 99% confidence level (z = 2.575)

5.98 ± 2.575 (.875)= (3.7261, 8.2339)

d) A higher percent confidence level gives a wider band.

2.

a) $\mathrm{SE}(\overline{\mathrm{x}}) = \frac{\sigma}{\sqrt{10}} = \frac{s}{\sqrt{10}} = \frac{2.1}{\sqrt{10}} = .6640783086$

b) 119 ± 3.2498 (.66407) = (116.84, 121.16

c) Precision = 3.2498 (.66407...) = 2.158121687

d) Assumptions:The population is normally distributed. The 10 subjects represent a random sample from this population.

3.

$$\hat{p} \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.5 \pm 1.96\sqrt{\frac{0.5(1-0.5)}{1229}} = (0.4725, 0.5284)$$

4.

$$(\hat{p}_1 - \hat{p}_2) \pm z_{(1-\frac{\alpha}{2})}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = (0.3003) \pm 2.575\sqrt{\frac{0.4878(0.5122)}{123} + \frac{(0.1875)(0.8125)}{96}}$$

$$= 0.3003 \pm 2.575(0.0602) = (0.1453, 0.4553)$$

2. Hypothesis testing or significance test

Hypothesis testing comprises a set of procedures for assessing presence or absence of effect, relationship and difference. Hypothesis is an opinion regarding unknown true population value called parameter. For instance, you might have, I belief that mean diastolic blood pressure of the adult population in Addis is 90 mmHg. A pharmaceutical company might have produced a new drug A that will improve survival time of cancer patients than the standard treatment. A physician might be interested to check the effect of a certain diet – exercise program which may lower cholesterol with the purpose of recommending diet and exercise to his patients.

Example: consider the following research question: Do elderly men and women (≥70 years) randomized to get a resveratrol supplement have a lower mortality rate than those who get a placebo?

Research hypothesis: Men and women > age 70 years randomized to get a resveratrol supplement have a lower mortality rate than those who get a placebo.

The Null Hypothesis: Men and women > age 70 years randomized to receive a resveratrol supplement do not have lower mortality rate than those who receive placebo.

The 'alternative' hypothesis: Men and women > age 70 years randomized to get a resveratrol supplement have a lower mortality rate than those who get a placebo.

Types of Hypothesis

Null hypothesis (H_o) – it is astatement regarding the value(s) of unknown parameter(s). Typically, it will mean no association between explanatory and response variables in our applications, no treatment effect, no difference

Alternative hypothesis(H_A) – it is a statement contradictory to the null hypothesis (will always contain an inequality)

Goal of Hypothesis Testing

The main goal of hypothesis testing is making statement(s) regarding unknown population parameter values based on sample data by evaluating the role of chance for the observed result.

Elements of a hypothesis test:

Test statistic: a quantity based on sample data and null hypothesis used to test between null and alternative hypotheses.

Rejection region: values of the test statistic for which we reject the null in favor of the alternative hypothesis.

P – Value: A p value is a probability that the result is as extreme or more extreme than the observed value if the null hypothesis is true. If the p value is less than or equal to a, we reject the null hypothesis, otherwise we do not reject the null hypothesis

Decision in hypothesis testing

The ultimate goal of any hypothesis testing is to reject the null hypothesis by providing adequate evidence against it. A certain hypothesis is rejected if the probability of obtaining the sample result (also called observed value/or test statistic/or calculated value) by chance is very small (close to zero). This probability is called p-value. P-value is also known as (aka) observed significance level.**P-value** – is a measure of the strength of evidence the sample data provides against the null hypothesis:P(Evidence this strong or stronger against H_0 | H_0 is true).

Mathematically, p – value is computed as $p - value : p = P(| Z | \geq | z_{obs} |)$

Outcomes in Hypothesis Testing

Null Hypothesis	Decision	
	Reject	Accept
True	Type I Error (α)	Correct
False	Correct	Type II Error (ß)

$\alpha = P(Type\ I\ Error)$ $\beta = P(Type\ II\ Error)$, and keep α and β reasonably small!

Significance Level (α)

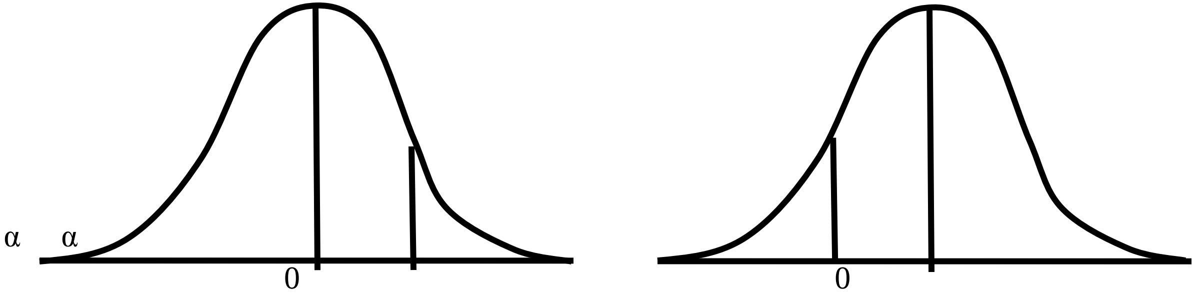
A particular hypothesis testing is significant if we reject the null hypothesis. Researcher decides the maximum risk (called significance level) s/he is ready to take. Usual significance level is 5%.

Types of Hypothesis Testing

One tailed test

In 1-tailed test we know, beforehand, that only deviations to one direction are possible or interesting. Alternative hypothesis takes the form of “less than” or “greater than”.

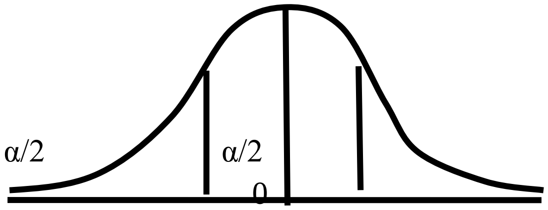
In a one tail test, the rejection region is at one end of the distribution or the other.



Two tailed test

In the 2-tailed test the aalternative hypothesis takes the form “different than or different from.”

In a two tail test, the rejection region is split between the two tails. Which one is used depends on the way the null hypothesis is written.



Note. The alternative hypothesis informs whether one tailed or two tailed

Steps in hypothesis testing

1. Set the null hypothesis and the alternative hypothesis.
2. Choose criteria: α%
3. Choose and calculate appropriate test statistic ($Z_{statistic}$, $t_{Statistic}$, $\chi_{statistic}$, $F_{statistic}$)

$$\text{Test Statistic} = \frac{\text{Estimator - Parameter}}{\text{Standard Error of Estimator}}$$

4. Compute the the p-value.
5. Decision rule: reject or fail to reject Ho based on the p-value (or critical value)
   If the p-value is less than α% then reject the null hypothesis otherwise the null hypothesis remains valid.
6. Write a conclusion in terms of the problem

Testing a population mean (μ)

Null hypothesis: mean equals μ_0

Alternative hypothesis (2-tailed): mean is different from μ_0

Alternative hypothesis (1-tailed): mean is less than μ_0

Alternative hypothesis (1-tailed): mean is bigger than μ_0

Steps for testing a population mean (with σ known)

1. State the null and alternative hypothesis: $\mathrm{H}_0 : \mu = \mu_0$ vs $\mathrm{H}_\mathrm{A} : \mu \neq \mu_0$) or $\mathrm{H}_\mathrm{A} : \mu < \mu_0$) or $\mathrm{H}_\mathrm{A} : \mu > \mu_0$)

2. State the level of significance (Assume a = 0.05 unless otherwise stated)

3. Calculate the test statistic: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

4. Find the P-value:

For a two-sided test: $P\text{-value} = \Pr(Z \geq |z| \text{ or } Z \leq -|z|) = 2\Pr(Z \geq |z|)$

For a one-sided test (right tailed): $P\text{-value} = \Pr(Z \geq z)$

For a one-sided test (left tailed): $P\text{-value} = \Pr(Z \leq z)$

Testing a population mean (μ): σ unknown

- Calculate standardized sample mean: $t = \dfrac{\bar{x} - \mu}{s / \sqrt{n}}$
- Calculate the p-value that indicates, how likely it is to get this kind of value if we assume that null hypothesis is true

Example1:

Researchers claim that the mean age of population having a certain disease 'A' is 35 years. To prove their claim, a researcher collected information from a random sample of 20 individuals drawn from population of interest. Population variance is known and is equal to $\sigma^2 = 25$ and the study found that the mean age of 20 individuals is as 29.Test whether the mean age of population having disease 'A' is 35 years.

Solution 1:

1. $\mathrm{H}_0 : \mu = 35$ against $\mathrm{H}_\mathrm{A} : \mu \neq 35$
2. $\alpha = 5\%$
3. **Test Statistic**:
4. P – value = $2P(Z_{calc} \leq -2.68) = 2P(Z_{calc} \geq 2.68)$
   P – value = 2x0.0037 = 0.0074
5. **Decision**: Reject H_0 since p – value is less than α.
6. **Conclusion**: We conclude that the mean age of the population with a specific disease 'A' is not equal to 35 years ($p<0.05$).

Example 2:

We know from our background knowledge that the mean fasting blood sugar level of non pregnant young adult women is 88 mg/dl. With this background, we conducted a study on a sample of 25 ladies in 2nd / 3rd trimester of pregnancy, attending the obstetric department.

We found that the mean fasting blood sugar of this sample of 25 ladies was 95 mg/dl with a standard deviation (SD) of 14. Apparently, our sample shows that the fasting blood sugar, on an average is higher by (95-88) = 7 mg/dl among pregnant ladies, as compared to non pregnant ladies. We now want to see, statistically, whether this is a significant finding or simply a matter of chance, i.e, simply due to random (sample to sample) variations.

Solution 2:

1. $H_0 : \mu = 88$ against $H_A : \mu \neq 88$
2. $\alpha = 5\%$
3. **Test Statistic**:
4. P – value = $P(t_{calc} \geq 2.50)$
   With 24 df, $0.01 \leq P$ – value ≤ 0.02,
5. **Decision**: Reject H_0 since p – value is less than α.

The higher average fasting blood sugar that we have seen among our sample of pregnant ladies is not likely to have come up simply because of “chance.”

6. **Conclusion**: We finally conclude, clinically, that pregnancy definitely leads to a rise in fasting blood sugar level

Example 3: A factory that discharges waste water into the sewage system is required to monitor the arsenic levels in its waste water and report the results to the Environmental Protection Agency (EPA) at regular intervals. Sixty beakers of waste water from the discharge are obtained at randomly chosen times during a certain month. The measurement of arsenic is in nanograms per liter for each beaker of water obtained.”Suppose the EPA wants to test if the average arsenic level exceeds 30 nanograms per liter at the 0.05 level of significance.

Information given: sample size: n = 60, $\bar{x} = 30.83$ and assume it is known that $\sigma = 34$

37.6	56.7	5.1	3.7	3.5	15.7	24.1	36.2
20.7	81.3	37.5	15.4	10.6	8.3	25.6	33.6
23.2	9.5	7.9	21.1	40.6	35	48.9	16.5
19.4	38.8	20.9	8.6	59.2	6.2	12.2	9.9
24	33.8	21.6	15.3	6.6	87.7	24.1	33.2
4.8	10.7	182.2	17.6	15.3	37.6	14.5	30
152	63.5	46.9	17.4	17.4	26.1	21.5	3.2
45.2	12	128.5	23.5				

Solution 3:

1. State the null and alternative hypothesis: $\mathrm{H}_0 : \mu \pm 30$ or $\mathrm{H}_0 : \mu \le 30$ vs $\mathrm{H}_\mathrm{A} : \mu > 30$
2. State the level of significance: a = 0.05
3. Calculate the test statistic: $z = \dfrac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \dfrac{30.83 - 30}{34 / \sqrt{60}} = \dfrac{0.83}{4.39} = 0.19$
4. Find the P-value:

   $P\text{ - value} = \Pr(Z \ge z) = \Pr(Z \ge 0.19) = 1 - \Pr(Z < 0.19) = 1 - 0.5753 = 0.4247$
5. Decision: Therefore, we fail to reject H_0
6. "There is no significant statistical evidence that the average arsenic level exceeds 30 nanograms per liter at the 0.05 level of significance."

Comparing two group means

Step 1: H_0:μ_1-μ_2=0 vsH_A: μ_1-μ_2>0

H_0: μ_1-μ_2 = 0 (No difference in population means)

H_A: μ_1-μ_2>0 (Population Mean 1 >PopulationMean 2)

Step2: Determine the level of significance (α, 5%)

Step3: Test statistic

$$z_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Step4: Decision: Reject H_0 if p – value is less than α

Step 5: Conclusion

Example 4- Efficacy test for new drug

A drug company has new drug, wishes to compare it with current standard treatment. Federal regulators tell company that they must demonstrate that new drug is better than current treatment to receive approval. Firm runs clinical trial where some patients receive new drug, and others receive standard treatment. Numeric response of therapeutic effect is obtained (higher scores are better). Parameter of interest: μ_{New} - μ_{Std}.

Null hypothesis - New drug is no better than standard trt

$$H_0 : \mu_{New} - \mu_{Std} \le 0 \qquad (\mu_{New} - \mu_{Std} = 0)$$

Alternative hypothesis - New drug is better than standard trt

$$H_A : \mu_{New} - \mu_{Std} > 0$$

Experimental (Sample) data:

$$\bar{x}_{New},\ s_{New},\ n_{New},\ \bar{x}_{Std},\ s_{Std},\ n_{Std}$$

In large samples, the difference in two sample means is approximately normally distributed:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2\ ,\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

Under the null hypothesis, μ_1-μ_2=0 and:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

If σ_1^2 and σ_2^2 **are unknown, estimate them by** s_1^2 and s_2^2

Discussion - Efficacy Test for New drug

Type I error - Concluding that the new drug is better than the standard (H_A) when in fact it is no better (H_0). Ineffective drug is deemed better. Traditionally α = Pr(Type I error) = 0.05

Type II error - Failing to conclude that the new drug is better (H_A) when in fact it is. Effective drug is deemed to be no better. Traditionally a clinically important difference (D) is assigned and sample sizes chosen so that. β = Pr (Type II error | μ_1-μ_2 = D) ≤ .20).

Large-sample test

Step 1: H_0:μ_1-μ_2=0 vsH_A: μ_1-μ_2>0

H_0: μ_1-μ_2 = 0 (No difference in population means)

H_A: μ_1-μ_2> 0 (Population Mean 1 >PopulationMean 2)

Step2: Determine the level of significance (α= 5%)

Step3: Test statistic

$$z_{obs} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Step4: Decision: Reject H_0 if p – value is less than α

Step 5: Conclusion

Example 5: Botox for Cervical Dystonia

Patients - Individuals suffering from cervical dystonia

Response - Tsui score of severity of cervical dystonia (higher scores are more severe) at week 8 of treatment.

Research (alternative) hypothesis - Botox A decreases mean Tsui score more than placebo

Groups - Placebo (Group 1) and Botox A (Group 2)

Experimental (sample) results:

$$\overline{x}_1 = 10.1 \quad s_1 = 3.6 \quad n_1 = 33; \quad \overline{x}_2 = 7.7 \quad s_2 = 3.4 \quad n_2 = 35$$

Test whether Botox A produces lower mean Tsui scores than placebo (α= 0.05)

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_A : \mu_1 - \mu_2 > 0$$
$$z_{obs} = \frac{10.1 - 7.7}{\sqrt{\frac{(3.6)^2}{33} + \frac{(3.4)^2}{35}}} = \frac{2.4}{0.85} = 2.82$$
$$z_{obs} \geq z_{\alpha} = z_{.05} = 1.645, \; P-val: P(Z \geq 2.82) = .0024$$

Conclusion: Botox A produces lower mean Tsui scores than placebo (since 2.82 > 1.645 and P-value =0.0024)

Comparing two group means: selecting appropriate t-test

If we have an experiment, in which observations are paired (e.g. group1: Obese women cholesterol level before exercise and group2: same obese women cholesterol level after exercise), then we should use **paired sample t-test**.If we compare two independent groups with equal variances then we should use **independent samples t-test** for equal variances.If we compare two independent groups with unequal variances then we should use **independent samples t-test** for unequal variances.

Assumptions of the independent t Test

Test statistics using t distribution assume data are:

1. groups and individuals within groups are independent,
2. the sampling distribution of the mean difference is normal, and
3. variances in the two populations are equal (homoscedasticity).

Example 6: independent samples t-test

We consider cholesterol levels (mg/dl) of men with Type A and Type B behaviors. Data are:

Type A: 233, 291, 312, 250, 246, 197, 268, 224, 239, 239, 254, 276, 234, 181, 248, 252, 202, 218, 212, 325

Type B: 344, 185, 263, 246, 224, 212, 188, 250, 148, 169, 226, 175, 242, 252, 153, 183, 137, 202, 194, 213.

T test for paired comparisons

Variable of interest in this test is the difference between individual pairs of observations

Assumption: the observed differences constitute a simple random sample from a normally distributed population of differences that could be generated under the same circumstances

Test statistic is: $t = \dfrac{\bar{d} - \mu_d}{s_{\bar{d}}}$ and standard error: $s_{\bar{d}} = \dfrac{s_d}{\sqrt{n}}$

Example 7: paired t test

Twelve subjects participated in an experiment to study the effectiveness of a certain diet, combined with a program of exercise, in reducing serum cholesterol levels. Do the data provide sufficient evidence to conclude that the diet exercise program is effective in reducing serum cholesterol levels? Let α = .05

Serum Cholesterol Levels for 12 Subjects Before and After Diet-Exercise Program

Subject	Serum Cholesterol		Difference (after–before)
	Before (x_1)	After (x_2)	
1	201	200	-1
2	231	236	+5
3	221	216	-5
4	260	233	-27
5	228	224	-4
6	237	216	-21
7	326	296	-30
8	235	195	-40
9	240	207	-33
10	267	247	-20
11	284	210	-74
12	201	209	+8

Solution 7:

Step 1: State the hypothesis

Ho: The mean difference between before and after diet-exercise program is = 0 ($\mu_B – \mu_A$)

H_A: The mean difference between before and after diet-exercise program is = 0 ($\mu_B > \mu_A$)

2. Select the level of significance = .05

3. Select the appropriate test statistic: $t = \frac{\bar{d} - \mu_d}{s_{\bar{d}}}$, where

$$\bar{d} = \frac{\sum d_i}{n} = \frac{(-1)+(5)+\ldots+(8)}{12} = \frac{-242}{12} = -20.17 \text{ and}$$

$$s_d^2 = \frac{\sum\left(d_i - \bar{d}\right)^2}{n-1} = \frac{n\sum d_i^2 - \left(\sum d_i\right)^2}{n(n-1)} = \frac{12(10766)-(-242)^2}{12(11)} = 535.06$$

Hence, $t_{\text{statistic}} = \frac{-20.17-0}{\sqrt{535.06/12}} = \frac{-20.17}{6.68} = -3.02$

And the corresponding p value is 0.012.

Testing Population Proportion (P)

P is a value between 0 and 1

Step 1: State the null and alternative hypothesis

Step 2: Determine the level of significance (α, 5%)

Step 3: Calculate test statistics

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}}$$

Step 4: Calculate the p-value that indicates, how likely itis to get this kind of value if we assume that null hypothesis is true

Step 5: Decision

Step 6: Conclude

Example 8: A researcher wants to study prevalence of anemia among rural pregnant ladies. Previous studies have shown prevalence to be 50%. She took 100 pregnant ladies from rural areas and took their hemoglobin readings and found 40% as anemic.

(i) Which statistical procedure will be used in this method?

(ii) What would be the null hypothesis and the alternative hypothesis?

(iii) Can researcher conclude that prevalence of anemia in her study is different from previous studies?

Solution 8:

(i) Which statistical procedure will be used in this method? Z – test

(ii) What would be the null hypothesis and the alternative hypothesis?

H_0 : P= P_0 (50%) against H_A : P≠ P_0(50%)

(iii) Can researcher conclude that prevalence of anemia in her study is different from previous studies? Yes

$$z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.4 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = \frac{-0.1}{\sqrt{\frac{0.5^2}{100}}} = -2.00$$

P – value = P ($Z_{calc} \leq$ -2.00) = P ($Z_{calc} \geq$ 2.00) = 0.5 – 0.4772 = 0.0228

Example 9: Immunization Shots

The superintendent of a large school district wants to know if the proportion of first graders in her district that have received their immunization shots is different from last year. Last year, 74% of the first grade children had received their immunization shots. The superintendent random selects 100 first grade students and 77 of them have received their immunization shots.

Solution 9: Immunization Shots

Information given: sample size: n = 100, m = 77 and hence $\hat{p} = \frac{m}{n} = \frac{77}{100} = 0.77$

1. State the null and alternative hypothesis: $\mathrm{H_0} : P = 0.74 \text{ vs } \mathrm{H_A} : P \neq 0.74$
2. State the level of significance: assume a = 0.05
3. Calculate the test statistic: $z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.77 - 0.74}{\sqrt{\frac{0.74(1-0.74)}{100}}} = \frac{0.03}{0.04386} = 0.68$
4. Find the P-value:

$$P\text{-value} = 2*\Pr(Z \geq |z|) = 2*\Pr(Z \geq |0.68|) = 2*\Pr(Z \geq 0.68) = 2*(1\text{-}\Pr(Z < 0.68)) = 2*(1\text{-}0.7517)$$
$$= 2*(0.2483) = 0.4966$$

5. Decision:Therefore, we fail to reject H_0
6. Conclusion: “There is statistically significant evidence that this year’s true proportion of first grade students with immunization shots is different from last year’s proportion of 0.74.”

Comparing two proportions

Example 10: Research question: Are antidepressants arisk factor for suicide attempts in children and adolescents?

Researchers used Medicaid records to compare prescription histories between 263 children and teenagers (6-18 years) who had attempted suicide and 1241 controls who had never attempted suicide (all subjects suffered from depression).

Statistical question: Is a history of use of antidepressants more common among cases than controls?

Solution 10: What will we actually compare?

Proportion of cases that used antidepressants in the past vs. proportion of controls that did not used antidepressants.

Information: No (%) of cases (n=263)and of whom 120 who ever used any antidepressants , $\hat{p}_1$ = 46% and
No (%) of controls (n=1241) and whom 448 who ever used any antidepressants**,** $\hat{p}_2$ = 36%
Difference ($\hat{p}_1$- $\hat{p}_2$) = 10%

Is the difference statistically significant?
This 10% difference could reflect a true difference or it could be a fluke in this particular sample. The question: is 10% bigger or smaller than the expected sampling variability?
Solution 10:
Step 1: Assume the null hypothesis
Null hypothesis: There is no difference between antidepressant use and suicide attempts in the target population (= the difference is 0%).
Step 1: Determine the level of significance
Step 3: Predict the sampling variability assuming the null hypothesis is true:

The standard error of the difference in two proportions is:

$$= \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1\text{-}\bar{p})}{n_2}}, \text{ where } \bar{\text{p}} = \frac{120+448}{263+1241} = 0.377$$

$$= \sqrt{\frac{0.377(1-0.377)}{263} + \frac{0.377(1-0.377)}{1241}} = .033$$

$$Z = \frac{.10}{.033} = 3.0;$$

Step 4: Calculate a p-value

P-value=the probability of your data or something more extreme under the null hypothesis. $p = .003$

Step 5: Decision. We reject the null hypothesis.
Step 6: Conclusion. There is a statistically significant difference between antidepressant use and suicide in the target population.

Testing on association based on contingency table
The methods we will examine are:

- Chi square test for 2 ×2 (r ×c) Contingency Tables
- Fisher Exact test

Chi Square Test (χ^2)
Chi square test has the following ccharacteristics:

- Every χ^2 distribution extends indefinitely to the right from 0.
- Every χ^2 distribution has only one (right) tail.
- As df increases, the χ^2 curves get more bell shaped and approach the normal curve in appearance (but remember that a chi square curve starts at 0, not at - ∞)

Chi square test is used for nominal or ordinal explanatory and response variables. The variables can have any number of distinct levels, but mostly they appear to have two levels in epidemiologic researches the so called 2x2 contingency tables. Generally, for r levels of explanatory variable and c levels of response variable, the tables are called rxccontingency tables. For such classification tables, the chi square test is used to assess whether the distribution of the response variable is the same for each level of the explanatory variable (H_0: No association between the variables).

RxC contingency table classification:

		First criterion of classification level				**Total**
		1	**2**	**…**	**c**	
Second criterion of classification level	**1**	$\mathbf{O_{11}}$	$\mathbf{O_{12}}$	**…**	$\mathbf{O_{1c}}$	$\mathbf{r_1}$
	2	$\mathbf{O_{21}}$	$\mathbf{O_{22}}$	**…**	$\mathbf{O_{2c}}$	$\mathbf{r_2}$
	…	**…**	**…**	**…**	**…**	**…**
	r	$\mathbf{O_{r1}}$	$\mathbf{O_{r2}}$	**…**	$\mathbf{O_{rc}}$	$\mathbf{r_r}$
Total		$\mathbf{c_1}$	$\mathbf{c_2}$	**…**	$\mathbf{c_c}$	**n**

r = # of levels of explanatory variable and c = # of levels of response variable

Note. For contingency tables with more than 1 degree of freedom a minimum expectation of 1 is allowable if no more than 20 percent of the cells have expected frequencies of less than 5. Generally, for chi square test, the degree of freedom = (r-1)(c-1).

Hypothesis testing steps in chi square test

1. **State Hypotheses:**

Null hypothesis (H_o): The classification variables are independent

Alternative hypothesis (H_a): there is relationship between the variables

2. **Determine test criteria:** choose $\alpha = .05$

3. **Compute test statistic:**

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed frequency - expected frequency})^2}{\text{expected frequency}} = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

A statistic which measures the discrepancy between K observed frequencies O_1, O_2, ..., O_k and the corresponding expected frequencies e_1, e_2, ..., $e_{k.}$

If the null hypothesis is true the χ^2 test statistic follows a Chi-squared distribution with degrees of freedom df= $(r-1)\times(c-1)$. Here, r = # of rows and c = # of columns in the contingency table.

Notice there is a general pattern here, the expected value for frequency in the ith row and the jth column of the table is found by taking the row total for that row $(\mathrm{r_i})$ times the column total for that column $(\mathrm{c_j})$ and then dividing by the total sample size (n), i.e.

$$\text{Expected Frequency (E) for i}^{\text{th}}\text{ row and the j}^{\text{th}}\text{ column}\,(E_{ij})$$

$$=\frac{(row\,i\,total)\times(column\;j\,total)}{sample\,size}=\frac{\mathrm{r_i}c_j}{n}$$

$$r_i = row\,\mathrm{i\ total} \quad \mathrm{c_j} = column\;\mathrm{j\,total} \quad \mathrm{n = total\,sample\,size}$$

a. Calculate the Chi-Square statistic.

b. Find expected frequencies and put them in the contingency table beneath the observed frequencies in parentheses.

4. Compute p-value: the larger the χ^2 test statistic value, the smaller the p-value.

5. Decision: reject H_0 if p –value < α%

6. Conclusion:

Example 11: Age at First Pregnancy and Cervical Cancer (a Case-Control Study)

The following data come from a case-control study to examine the potential relationship between age at first pregnancy and cervical cancer. In a case-control study a random sample of cases (i.e. people with the disease in question) and controls (i.e. people similar to those in the case group, except they do not have the disease) and the proportion of people with some potential risk factor are compared across the two groups. Based on the data, can we conclude that there is association between age at first pregnancy and cervical cancer, i.e. age at first pregnancy and cervical cancer are independent. The data is presented in the table below:

	Age at First Pregnancy <= 25 (risk factor present)	**Age at First Pregnancy > 25** (risk factor absent)	Row Totals (fixed)
Cervical Cancer (Case)	**42**	**7**	49
Control	**203**	**114**	317
Column Totals (random)	245	121	n=366

Solution 11: In this study we will be comparing the proportion of women who had their first pregnancy at or before the ages of 25, because researchers suspected that an early age at first pregnancy leads to increased risk of developing cervical cancer.

1. **State Hypotheses**

H_o: Age at first pregnancy and cervical cancer are independent OR $\hat{p}_{case} = \hat{p}_{control}$

H_a: Age at first pregnancy and cervical cancer are associated OR $P_{case} \neq P_{control}$

2. **Determine Test Criteria:** choose $\alpha = .05$
3. **Compute Test Statistic**

a) Expected frequencies for the data

Disease Status	Age at 1st Pregnancy Age ≤ 25	Age at 1st Pregnancy Age > 25	Row Totals
Cervical Cancer (Case)	32.8	16.2	49
Healthy(Control)	212.2	104.8	317
ColumnTotals	245	121	366

$$E_{11} = \frac{r_1c_1}{n} = \frac{49 \cdot 245}{366} = 32.80 \qquad E_{12} = \frac{r_1c_2}{n} = \frac{49 \cdot 121}{366} = 16.20$$

$$E_{21} = \frac{r_2c_1}{n} = \frac{317 \cdot 245}{366} = 212.20 \qquad E_{22} = \frac{r_2c_2}{n} = \frac{317 \cdot 121}{366} = 104.80$$

b) Calculate the Chi-Square statistic.

$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E} = \frac{(42-32.8)^2}{32.8} + \frac{(7-16.20)^2}{16.20} + \frac{(203-212.20)^2}{212.20} + \frac{(114-104.80)^2}{104.80}$$

$\chi^2 = 9.011 \quad df = (2-1)\times(2-1) = 1$

4. **Compute p-value:** $p\text{-}value = .0027$

5. Decision: Since p – value is less than 0.05, we reject the null hypothesis

6. Conclusion: Age at first pregnancy and cervical cancer status are associated.

Exercises 2:

1. Body mass index

A simple random sample of 14 people from a certain population gives body mass indices as shown in Table below. Can we conclude that the BMI is not 35? Let α = .05.

Subject	BMI	Subject	BMI	Subject	BMI
1	23	6	21	11	23
2	25	7	23	12	26
3	21	8	24	13	31
4	37	9	32	14	45
5	39	10	57		

2. These data were obtained in a study comparing persons with disabilities with persons without disabilities. A scale known as the Barriers to Health Promotion Activities for Disabled Persons (BHADP) Scale gave the data. We wish to know if we may conclude, at the 99% confidence level, that persons with disabilities score higher than persons without disabilities.

3. Very-low-calorie diet (VLCD) Treatment

The Table below gives B (before) and A (after) treatment data for obese female patients in a weight-loss program. We calculate d_i = A-B for each pair of data resulting in negative values meaning that the participants lost weight.

We wish to know if we may conclude, at the 95% confidence level, that the treatment is effective in causing weight reduction in these people.

Participant	A	B
1	83.3	117.3
2	85.9	111.4
3	75.8	98.6
4	82.9	104.3
5	82.3	105.4
6	77.7	100.4
7	62.7	81.7
8	69.0	89.5
9	63.9	78.2

4. **Histological Type and Response to Treatment for Hodgkin's Patients** Is there a relationship between type of Hodgkin's and response to treatment? To answer this question, researchers randomly sampled medical records for 538 patients who had been classified as having some form of Hodgkin's disease and then looked at their response to treatment. The following table presents the data on the type of Hodgkin's and response to treatment.

	Response to Treatment			
Type of Hodgkin's	None	Partial	Positive	Row Totals (Random)
LD	44	10	18	72
LP	12	18	74	104
MC	58	54	154	266
NS	12	16	68	96
Column Totals (Random)	126	98	314	n = 538

Solutions on exercises 2

1. Given: n = 14, s = 10.64, $\bar{x}$ = 30.5 and α = .05

Assumptions: simple random sample, population of similar subjects and normally distributed

Hypotheses

H_0: $\mu = 35$

H_A: $\mu \neq 35$

Distribution: If the assumptions are correct and H_0 is true, the test statistic follows Student's t distribution with 13 degrees of freedom.

Test statistic: $t = \dfrac{\bar{x} - \mu}{\dfrac{s}{\sqrt{n}}} = \dfrac{30.5 - 35}{\dfrac{10.64}{\sqrt{14}}} = \dfrac{-4.5}{2.8434} = -1.58$

P – value: The value of p = .1375

Decision:Do not reject the null hypothesis as p=0.1375

Conclusion: Based on the data of the sample, it is possible thatm = 35.

2.

Disabled: $\bar{x}_1 = 31.83$ $n_1 = 132$ $s_1 = 7.93$

Nondisabled: $\bar{x}_2 = 25.07$ $n_2 = 137$ $s_2 = 4.80$

Assumption: independent random samples

Hypotheses:

H_0: $\mu_1 \leq \mu_2$

H_A: $\mu_1 > \mu_2$

Test statistic:

Because of the large samples, the central limit theorem permits calculation of the z score as opposed to using t. The z score is calculated using the given sample standard deviations.

If the assumptions are correct and H_0 is true, the test statistic is approximately normally distributed.

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = \frac{(31.83 - 25.07) - 0}{\sqrt{\dfrac{7.93^2}{132} + \dfrac{4.80^2}{137}}} = \frac{6.76}{0.8029} = 8.42$$

P – value: actual value = z = 8.42, p = 1.91×10^{-17}

Decision: We reject H_0, as p < 0.001

Conclusion: On the basis of these data, the average persons with disabilities score higher on the BHADP test than do the nondisabled persons.

3. Given, $d_i = A – B$, n = 9 and $\alpha = .05$

Assumption: the observed differences are a simple random sample from a normally distributed population of differences

Hypotheses:

H_0: $\mu_d = 0$

H_A: $\mu_d < 0$ (meaning that the patients lost weight)

Test statistic:

Distribution: the test statistic is distributed as Student's t with 8 degrees of freedom

The test statistic is t which is calculated as $t = \dfrac{\bar{d} - \mu_d}{s_{\bar{d}}}$

$$\bar{\mathrm{d}} = \frac{\sum_{i=1}^{n} \mathrm{d_i}}{\mathrm{n}} = \frac{-203.3}{9} = -22.5889$$

$$\mathrm{s}^2_{\bar{\mathrm{d}}} = 28.2961$$

$$\mathrm{t} = \frac{\bar{\mathrm{d}} - \mu_{\bar{\mathrm{d}}_0}}{\mathrm{s}_{\bar{\mathrm{d}}}} = \frac{-22.5899 - 0}{\sqrt{\frac{28.2961}{9}}} = -12.7395$$

P – value: $p = 6.79 \times 10^{-7}$

Decision: With $\alpha = .05$ and 8 df the critical value of t is -1.8595. We reject H_0 if $t < -1.8595$. or Reject H_0 because $-12.7395 < -1.8595$

Conclusion: On the basis of these data, we conclude that the diet program is effective.

References:

1. Michael J. Panik. Advanced Statistics from an Elementary Point of View. Elsevier Inc. 2005, USA
2. Alvan R. Feinstein. Principles of Medical Statistics. Chapman & Hall/CRC Boca, 2002.
3. Daniel, W. W. 1999. Biostatistics: a foundation for analysis in the health sciences. New York: John Wiley and Sons.
4. C.R.Cothari. Research Methodology: Methods and Techniques. 2nd ed. New Age International (P) Ltd, Publishers, New Delhi, 2004.
5. Daniel, W. W. 1999. Biostatistics: a foundation for analysis in the health sciences. New York: John Wiley and Sons.
6. Morton RF, Hebel JR, McCarter RJ: A Study Guide to Epidemiology and Biostatistics, 4th ed. Gaithersburg, Maryland, Aspen Publications, 1996.
7. Norman GR, Streiner DL: Biostatistics: The Bare Essentials, 2nd ed. Hamilton, Ontario, B.C. Decker, 2000.
8. Pagano M, Gauvreau K: Principles of Biostatistics, 2nd ed. Pacific Grove, CA, Duxbury Press, 2000.
9. BMJ. Statistics at Square One.
10. Kline et al. *Annals of Emergency Medicine* 2002; 39: 144-152.
11. Johnson R. *Just the Essentials of Statistics*. Duxbury Press, 1995.

Session 2: Correlation and Regression

Session overview

This session gives overview of the introduction to regression and correlation and the role of regression in public health research, and the types of regression based on the type of research out come in the application of public health research. Modeling the relationship between explanatory and response variables is a fundamental activity encountered in statistics. The simple and multiple linear regression methods are used to model the relationship between a quantitative response variable and one or more explanatory variables. A key assumption for these models is that the deviations from the model are normally distributed. In this session we describe similar methods that are used when the response variable has only two possible values (a binary response) e.g. alive or dead, positive or negative, success or failure and so on.

Learning Objectives

At the end of this session, the trainees are expected to:

- Identify and apply methods of analysis for continuous outcome variables
- Apply correlation and regression analysis
- Able to test the basic assumptions for linear regression
- Apply logistic regression and interpret the parameters

Correlation

How can we summarize a pair of variables measured on the same observational unit like for instance; percent of calories from saturated fat and cholesterol level, mother's weight gain during pregnancy and child's birth weight, % immunization of children and under 5 mortality and so on. How do we describe their joint behavior? This section explores the linear relationship between observed phenomena of continuous variables.

The first thing to do is construct a scatter plot, a graphical display of the data. A scatter- plot displays the form and direction of the relationship between two quantitative variables. Correlation is a statistical measurement of the relationship between two continuous variables.

The correlation coefficient (ρ) is a single number that measures the degree of linear relationship between two variables, X and Y. Correlation coefficient ranges from -1.00 to +1.00. The value of -1.00represents a perfect negative correlationwhile a value of +1.00 represents a perfect positive correlation. A value of 0.00 represents a lack of linear correlation.

Population correlation coefficient for variables X and Y is defined as:

$$\rho = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Where Cov(X, Y) represents covariance between X and Y and Var(X) represents variance of X. The covariance between X and Y can be defined as:

$$Cov(X,Y) = E(XY)-E(X)(Y)$$

Where E(X) and E(Y) are the expectations (population means) of the random variables X and Y, respectively.

Note that if X and Y are independent then E(XY) = E(X)E(Y), which makes identically zero. This implies that independent random variables have a correlation of zero. A correlation of zero does not imply independence. Because correlation is a measure of the linear relationship between X and Y, other non-linear relationships (e.g. $Y = X^2$) may result in a correlation of zero.

The population correlation coefficient is estimated by sample correlation coefficient r using:

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} = \frac{S_{XY}}{S_X S_Y}$$

which is called the sample correlation coefficient. This expression is sometimes referred to as Pearson's correlation coefficient.

Properties of correlation coefficient (r) are:

- r only measures the strength of a linear relationship. There are other kinds of relationships besides linear for which the correlation coefficient will be zero.

- r is always between -1 and 1 inclusive. -1 means perfect negative linear correlation and +1 means perfect positive linear correlation
- r does not change if the independent (x) and dependent (y) variables are inter- changed
- r does not change if the scale on either variable is changed. You may multi- ply, divide, add, or subtract a value from all the x-values or y-values without changing the value of r.

Hypothesis Testing

The claim we will be testing is "There is significant linear correlation between the two variables X and Y". The correlation at the population can be denoted by a parameter. Thus the hypothesis will be tested for the population parameter. The hypothesis can be stated as

The null hypothesis is stated as: H_0: $\rho = \rho_0$

The alternative hypothesis can be also stated as: H_1: $\rho < \rho_0$ or $\rho > \rho_0$

The standard error, SE(r) is computed as:

$$SE(r) = \sqrt{\frac{1-r^2}{n-2}}$$

The formula for the test statistic is based on t-test and defined as:

$$t_{cal} = \frac{r-\rho_0}{SE(r)} = \frac{r-\rho_0}{\sqrt{\frac{1-r^2}{n-2}}}$$

Based on the calculated value of t and its degree of freedom, determine the corresponding p-value to make decision as usual.

Example 1: Consider the variables under 5 mortality and percent of child immunization of 20 countries. The data is given below:

Table 1: Percent of child immunization against DPT3&under 5 mortality for 20 countries, 1992

Nation	% Immunized for DPT 3	< 5 Mortality Rate
Bolivia	77	118
Brazil	69	65
Cambodia	32	184

Canada	85	8
China	94	43
Chez Republic	99	12
Egypt	89	55
Ethiopia	13	208
Finland	95	7
France	95	9
Greece	54	9
India	89	124
Italy	95	10
Japan	87	6
Mexico	91	33
Poland	98	16
Russia	73	32
Senegal	47	145
Turkey	76	87
England	96	9

Before we conduct any statistical analysis, we should always create a scatter plot of the data. These data could be presented in a scatter plot as shown in Figure 1 below:

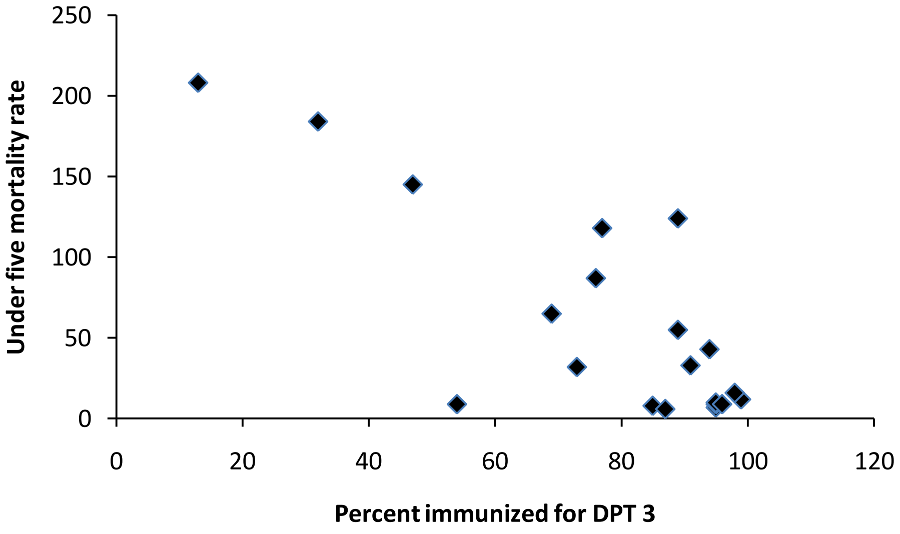


Fig. 1: Scatter plot for immunization and under 5 mortality

As can be seen from the scatter plot, it seems that the variable immunization and under 5 mortality are negatively related. As immunization increased, the number of child death decreased.

To check if the relationship is statistically significant, we need to apply hypothesis testing. The sample correlation between the two variables (children immunization against DPT and under 5 mortality) can be determined as:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}} = -0.79$$

The sign of the sample correlation is consistent with what has been observed from the scatter plot.

We can test the hypothesis whether the population correlation coefficient (ρ) is zero or not.

Null hypothesis: H_0: ρ=0 (there is no linear relationship between the two variables)

Alternative Hypothesis H_a:ρ≠0 (there is none zero linear relationship between the two variables)

The test statistics can be calculated as:

$$t_{cal} = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{n - 2}}} = \frac{-0.79 - 0}{\sqrt{\frac{1 - (-0.79)^2}{40 - 2}}} = -7.95$$

At tcal = -7.95 the p-value is < 0.001and therefore we reject the null hypothesis and concluded there is significant negative relationship between percent of DPT3 immunization and under 5 mortality.

Introduction to Simple Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable (usually denoted by X), and the other is considered to be a dependent variable (denoted by Y). For example, a researcher might want to relate the weights of individuals to their heights using a linear regression model. The dependent variable denoted by Y for linear regression should be numeric (continuous), while the explanatory can be any types of variables. Remember before attempting to fit a linear model to observed data, a researcher should first determine whether or not there is a relationship between the variables of interest. We can first use the exploratory analysis to check whether the two variables are expected to associate or not. There are different tools to see these including plots. A scatterplot can be a helpful tool in exploring relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables then fitting a linear regression model to the data probably will not provide a useful model. In general, the goal of linear regression is to find the line that best predicts Y from X. Linear regression does this by finding the line that minimizes the sum of the squares of the vertical distances of the points from the line.

How linear regression works?

The most common method for fitting a linear regression line is the method of least-squares. This method calculates the best-fitting line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line (if a point lies on the fitted line exactly, then its vertical deviation is 0). More precisely, the goal of regression is to minimize the sum of the squares of the vertical distances of the points from the line.

Mathematical equations describing these relationships are also called models, and they fall into two types: deterministic or probabilistic models.

- Deterministic Model: an equation or set of equations that allow us to fully determine the value of the dependent variable from the values of the independent variables.
- Probabilistic Model: a method used to capture the randomness that is part of a real-life process. To create a probabilistic model, we start with a deterministic model that approximates the relationship we want to model and add a random term that measures the error of the deterministic component.

Deterministic Model: Consider the model with yield (as response variable) and fertilizer (as dependent variable). We can display using plots as:

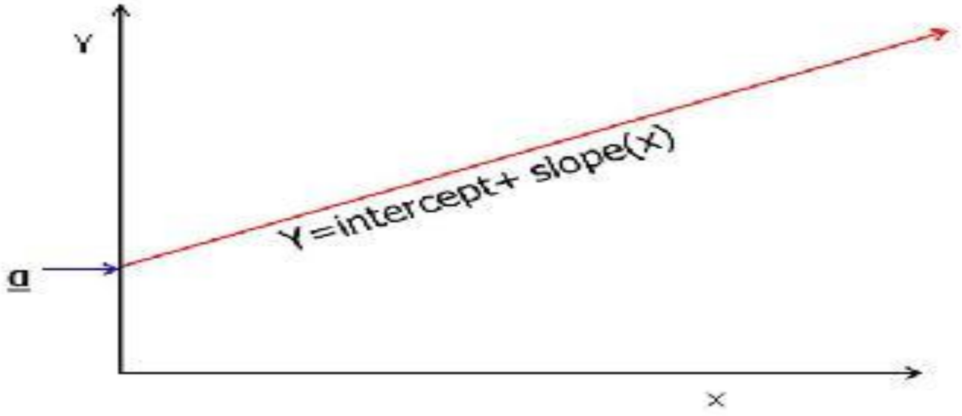


Figure 2: Deterministic regression model

In real life however, the response variable will vary even among the same size of the independent variable, thus the probabilistic model is given as: We start by recognizing that the response will vary even for constant values of the predictor, and model this fact by treating the responses (y_i) as realizations of random variables.

$$y_i \sim N(\mu; \sigma^2)$$

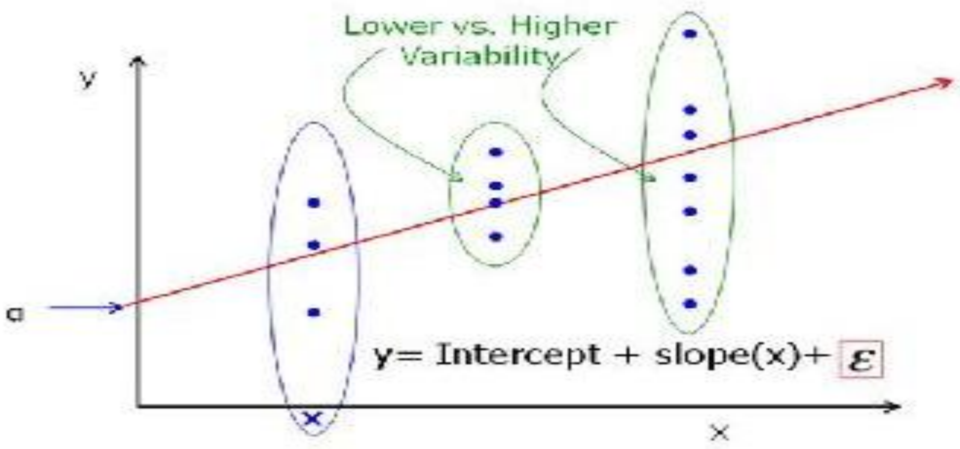


Figure 4.3: Probabilistic regression model

The simplest way to express the dependence of the expected response i on the predictor x_i is to assume that it is a linear function, say

$$\mu_i = \alpha + \beta x_i$$

The parameter is called the constant or intercept, and represents the expected response when x_i = 0. The parameter is called the slope, and represents the expected increment in the response per unit change in xi. Then the simple linear regression model written with an explicit error term as:

$$y = a + b(x) + e$$

Where e is a random component. It is the difference between the actual and the estimated dependent variable based on the size of the independent variable. The value of error (e) will vary from subject to subject, even if independent variable remains the same. Note that both α andβ are population parameters which are usually unknown and hence estimated from the data by a and b.

Estimating the Coefficients

In much the same way we base estimates α by a, and β by b, intercept and slope respectively. The regression line is given by: $\hat{y} = a + bx$

This is an application of the least squares method and it produces a straight line that minimizes the sum of the squared differences between the points and the line. The coefficients a and b for the least squares line can be computed by:

$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ and } a = \bar{y} - b\bar{x}$$

Assumption for the error term, e

For the regression methods to be valid the following four conditions for the error term (e) must be met:

- The probability distribution of is normal.
- The mean of the distribution is 0, i.e., E(e) = 0.
- The standard deviation of e is constant regardless of the value of x.
- The value associated with any particular value of y is independent of any other value of y.

The least squares method will always produce a straight line, even if there is no relationship between the variables, or if the relationship is something other than linear.

The sum of square error (SSE) can be computed as:

$$SSE = (n-1)(S_y^2 - \frac{S_{xy}^2}{S_Y^2})$$

This is used in the calculation of the standard error of estimate as:

$$S(e) = \sqrt{\frac{SSE}{n-2}}$$

- If S(e) is zero then all the points lie on the regression line.
- If S(e) is small, then the fit is excellent and the linear model should be used for forecasting
- If S(e) large, then the model is poor.

Hypothesis Testing of the Slope

If no linear relationship exists between the two variables, we would expect the regression line to be horizontal, that is, to have a slope of zero. We want to see if the slope (1) is something other than zero.

The hypothesis can be stated as:

Null hypothesis: H0 : β = 0

Alternative hypothesis: H_1 : β≠0

The test statistic can be computed as:

$$t_{cal} = \frac{b-\beta}{S_{b_1}}$$

Where the standard error of b1 denoted as Sb1 can be obtained using the formula:

$$S(b) = \frac{S(e)}{\sqrt{(n-1)S_x^2}}$$

If the error variable (e) is normally distributed, the test statistic has a Student t-distribution with n-2 degrees of freedom. The rejection region depends on whether or not we're doing a one or two-tail test. We can also estimate interval for the slope parameter, β. The confidence interval estimator is given by:

$$(b \pm t_{\alpha/2,v} SE(b))$$

In which v stands for the degree of freedom determined by n-2

Coefficient of Determination

It is also useful to measure the amount of variation in y explained by variation in the independent variable x. This is done by calculating the coefficient of determination denoted by R^2. It can be computed using the formula:

$R^2 = \frac{S_{xy}^2}{S_x^2 S_y^2}$ or equivalently can be computed using: $R^2 = 1 - \frac{SSE}{\sum (y_i - \bar{y})^2}$

The coefficient of determination is the square of the coefficient of correlation (r). We can partition the variation in y into two parts as Variation in y = SSE + SSR. Sum of Squares Error (SSE) - measures the amount of variation in y that remains unexplained (i.e. due to error). Sum of Squares Regression (SSR) - measures the amount of variation in y explained by variation in the independent variable x.

Unlike the value of a test statistic, the coefficient of determination does not have a critical value that enables us to draw conclusions. In general the higher the value of R^2, the better the model fits the data. If $R^2 = 1$, it implies Perfect match between the line and the data points while if $R^2 = 0$ then it implies there are no linear relationship between x and y.

Assumptions of linear regression:

- Linearity - the relationship between x and y is linear.
- Independence of Error Terms - successive residuals are not correlated.
- Homoscedasticity - the variance of the error terms is constant for each value of x.
- Normally Distributed Error Terms - the error terms follow the normal distribution.

Regression Diagnostics

As we have seen before, there are three conditions that are required in order to perform a regression analysis. These are:

- The error variable must be normally distributed,
- The error variable must have a constant variance and
- The errors must be independent of each other.

How can we diagnose violations of these conditions?

- Residual Analysis: Residual Analysis, that is, examine the differences between the actual data points and those predicted by the linear equation. We can use these residuals to determine whether the error variable is non-normal, whether the error variance is constant, and whether the errors are independent

- Test of Non-normality: We can take the residuals and put them into a histogram to visually check for normality.

- Test of Heteroscedasticity: When the requirement of a constant variance is violated, we have a condition of heteroscedasticity. We can diagnose heteroscedasticity by plotting the residual against the predicted y.

- Test of Non independence of the Error Variable: When the data are time series, the errors often are correlated. Error terms that are correlated over time are said to be auto correlated or serially correlated. We can often detect autocorrelation by graphing the residuals against the time periods. If a pattern emerges, it is likely that the independence requirement is violated.

- Test of Outliers: An outlier is an observation that is unusually small or unusually large. Possible reasons for the existence of outliers include:
  - There was an error in recording the value
  - The point should not have been included in the sample
  - Perhaps the observation is indeed valid.

Outliers can be easily identified from a scatter plot. If the absolute value of the standard residual is greater than 2, we suspect the point may be an outlier and investigate further. Procedure for model dignosis:

- Develop a model that has a theoretical basis.
- Gather data for the two variables in the model.
- Draw the scatter diagram to determine whether a linear model appears to be appropriate.
- Determine the regression equation.
- Calculate the residuals and check the required conditions

- Assess the model's t.

If the model fits the data, use the regression equation to predict a particular value of the dependent variable and/or estimate its mean.

Multiple Linear Regression

Simple linear regression can be extended to multiple linear regression models by allowing the response variable to be a function of k explanatory variables x_1; x_2; : : : ; x_k. This relationship is straight-line and in its basic form it can be written as:

$$Y_i = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i + \varepsilon_i$$

Where the random errors i, for i = 1… n; are independent normally distributed random variables with zero mean and constant variance. The definition of a multiple linear regression model is that mean of the response variable,

$$E[Y_i] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

is a linear function of the regression parameters ; 1; 2, …, i. It is standard to assume normality in the definition of multiple linear regression models. In situations where the normality assumption is not satisfied, one might use a generalized linear model instead. In order to t the `best' regression line, we shall use the principle of least squares-the same principle we used in previous section. According to this principle, the best fitting model is the one that minimizes the sum of squared residuals, where the residuals are the deviations between the observed response variables and the values predicted by the fitted model. As in the simple case: the smaller the residuals, the closer the t. Note that the residuals i are given by:

$$\varepsilon_i = Y_i - (\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i)$$

Where i=1, 2, ...,n. It follows that the residual sum of squares, SSE, is given by

$$SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x)^2$$

We are interested in the values of; 1; 2, …, i which minimize this sum. This can be done using partial derivatives with respect to the i + 1 parameters.

Introduction to Logistic Regression

What's wrong with regressing against binary dependent variables?

- If you use linear regression, the predicted values will become greater than one and less than zero if you move far enough on the X-axis. Such values are theoretically inadmissible.
- Another more serious problem is that such an analysis violates many assumptions of linear regression. For example, the assumption of Homoscedasticity won't hold. Homoscedasticity means that the variance around the dependent variable is similar for all values of the independent variable. Variance for a distribution of a binary variable is npq where n is the sample size, p is the probability of a 1, and q is the probability of a 0.
- The significance testing of the b weights rest upon the assumption that errors of prediction (Y-Y') are normally distributed. Because Y only takes the values 0 and 1, this assumption is pretty hard to justify, even approximately. Therefore, the tests of the regression weights are suspect if you use linear regression with a binary, where Y is the observed value and Y' is the predicted value.

Logistic Regression Model

Logistic regression is part of a category of statistical models called generalized linear models. From a practical standpoint, logistic regression and least squares regression are almost identical. Both methods produce prediction equations. In both cases the regression coefficients measure the predictive capability of the independent variables. The response variable that characterizes logistic regression is what makes it special. With linear least squares regression, the response variable is a continuous variable, however with logistic regression, the response variable is an indicator of some characteristic, that is, a 0/1 variable.

The dependent variable can take the value 1 for the event of interest with a probability p, or the value 0 with probability of failure1-p. This type of variable is called a Bernoulli (binary) variable. The independent or predictor variables in logistic regression can take any form. That is, logistic regression makes no assumption about the distribution of the independent variables. They do not have to be normally distributed, linearly related or of equal variance within each group like linear regression do. Let us have a look the two models graphically as:

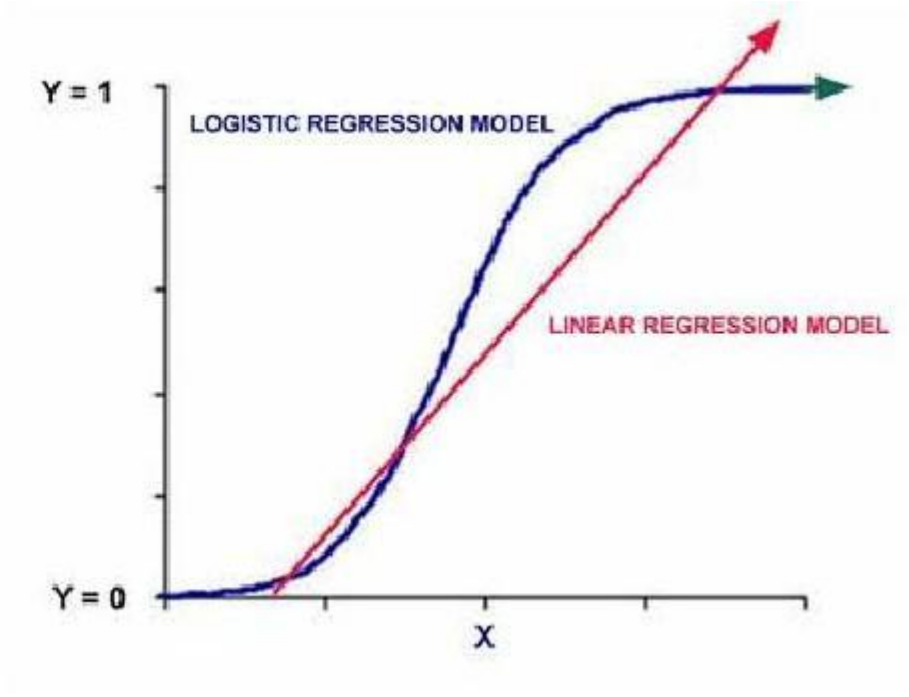


Figure 7.4: Comparison of logistic and linear regression models

Models for Binary Data

Logistic regression equation does not directly predict the probability that the indicator is equal to 1. It predicts the log odds that an observation will have an indicator equal to 1. The odd of an event is defined as the ratio of the probability that an event occurs to the probability that it fails to occur. The simplest function to define probability of an event given the exposure variables, i.e., P(Y=1/X) is defined by an exponential function as:

$$\Pr ob(Y = 1/X) = p(x) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}} \text{ called the logistic function}$$

But this function is not simple to estimate the coefficients therefore; we convert the probabilities to odds as:

$$\text{Odds} = \frac{\text{p(x)}}{\text{1 - p(x)}} = \frac{\exp(\alpha+\beta X)/(1+\exp(\alpha+\beta X))}{1-[\exp(\alpha+\beta X)/(1+\exp(\alpha+\beta X))]} = \exp((\alpha+\beta X))$$

The constraints at 0 and 1 make it impossible to construct a linear equation for predicting probabilities. With logistic regression we are interested in modeling the mean of the response variable p in terms of an explanatory variable x. We could try to relate p and x through the equation p(x) = α+ βx. Unfortunately, this is not a good model as long as β≠0, extreme values of x will give values of α+ βx that are inconsistent with the fact that 0 ≤p(x) ≤1.

The logistic regression solution to this difficulty is to transform the odds (p/(1 p)) using the natural logarithm (log(p(x)/(1-p(x)))). We use the term log odds for this transformation. Assuming the response variable has only one explanatory variable x, log odds is a linear function of the explanatory variable:

$$\log\left(\frac{\mathrm{p(x)}}{1-p(x)}\right)=\alpha+\beta x$$

The log odds called logitsliesbetween-∞ to +∞. In the model α is the intercept and β is the slope. In addition, exp(β) is called the odds ratio between two values of the exposure variable.

How do we estimate the parameters of this relationship? We need some method corresponding to the Least Squares method used for linear regression models which is the Maximum Likelihood estimation (MLE).

Example: Consider the data on Age and signs of coronary heart diseases (CHD) of a sample of 100 cases and sample of the first 33 cases is as follows.

Age	**CD**	**Age**	**CD**	**Age**	**CD**
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1

38	0	52	0	81	1

The scatter plot of age with cumulative percentage of disease can be drawn as follows:

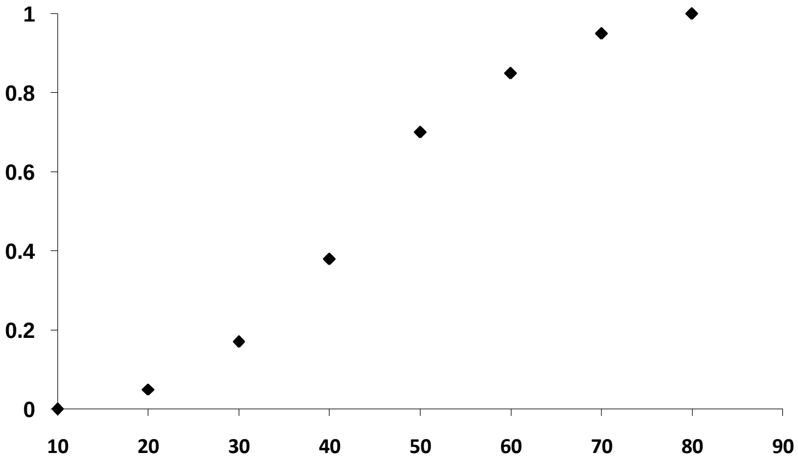


Fig. 4: Scatter plot of age with percentage of coronary heart diseases

As can be seen from the scatter plot, linear regression will not handle the data, but logistic curve does handle. The equation of the line is

$$\log it(p) = \log(\frac{p}{1-p}) = a + bX$$

Where p is the proportion of coronary heart diseases, x denotes age, s stands for intercept and b is the slope of the logistic curve. It measures the effect of increase in age on the prevalence of coronary heart diseases. Let us categorized the variable age in to two categories (≥55 and <55 years). As a result, the following contingency table can be constructed.

CD	**Age group**	
	55+	**<55**
Yes	21	22
No	6	51

Logistic regression analysis provided the following model output.

Variable	**Coefficient**	**SE**	**Wald**
Age	2.1	0.529	15.76
Constant	-0.84	0.255	11.29

Then the equation of logistic regression is given by:

$$\log it(p) = -0.841 + 2.1 \times Age$$

The odds ratio can be calculated as $e^{2.1}$=8.1. The odds ratio can be interpreted as being on age group 55+ years has 8.1 times risk to develop coronary heart diseases compared with age less than 55 years. The Wald test for coefficient 15.76 with 1 degree of freedom gives p-value < 0.001 and the 95% confidence interval for the odds ratio is given by:

$$e^{2.1 \pm 1.96 \times 0.529} = \mathrm{e}^{2.1 - 1.96 \times 0.529}, \mathrm{e}^{2.1 + 1.96 \times 0.529} = (2.9,\ 22.9)$$

Thus, since the 95% confidence interval doesn't include 1, it indicates that age has significant effect on coronary heart diseases.

Multiple Logistic Regression

When we include more than one explanatory variable for the binary response variable, then the model we need to fit is multiple logistic regression. The equation for k number of explanatory variables is given by:

$$\mathrm{logit}(\mathrm{p}(\mathrm{x})) = \log\left(\frac{\mathrm{p}(\mathrm{x})}{1 - p(x)}\right) = \alpha + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk$$

Again, the probability can be derived from multiple logistic regression function by taking exponent both sides as we did before:

$$\mathrm{p}(\mathrm{x}) = \frac{e^{\alpha + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk}}{1 + e^{\alpha + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk}}$$

Where α is the constant/intercept of the equation and, the bs are the coefficient of the predictor variables.

Model building methods

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious model. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available during model creation.

Enter Method: This method is the default method and enters all the variable in a single step. The method "enter" the variables into the model without specified order. This method is called the "simultaneous" method.

Forward stepwise: The variables (or interaction terms) that are specified on FSTEP are tested for entry into the model one by one, based on the significance level of the score statistic. The variable with the smallest significance less than PIN is entered into the model. After each entry, variables that are already in the model are tested for possible removal, based on the significance of the conditional statistic, the Wald statistic, or the likelihood-ratio criterion. The variable with the largest probability greater than the specified POUT value is removed, and the model is re-estimated. Variables in the model are then evaluated again for removal. When no more variables satisfy the removal criterion, covariates that are not in the model are evaluated for entry. Model building stops when no more variables meet entry or removal criteria or when the current model is the same as a previous model.

Backward stepwise: As a first step, the variables (or interaction terms) that are specified on BSTEP are entered into the model together and are tested for removal one by one. Stepwise removal and entry then follow the same process as described for FSTEP until no more variables meet entry or removal criteria or when the current model is the same as a previous model.

Significance test of regression coefficient

A Wald test is used to test the statistical significance of each coefficient (b) in the model. Thus we can apply the usual t-test to the parameter estimate for a test of whether the parameter is zero.

Having estimated the parameters of the model we would wish to test whether certain of these parameters might be zero.

Ho: β=0 (No relationship between event and exposure variable)

H_A: β≠0

The test is based on the Wald statistics defined as:

$$\mathrm{Wald} = \left[\frac{\mathrm{b}}{\mathrm{se(b)}}\right]^2 \approx \chi^2 \ \text{with 1 df}$$

The P-value for the significance test of the null hypothesis against the alternative HA is computed using the fact that when the null hypothesis is true is based on the chi-square test with 1 degree of freedom.

Confidence interval for β

The approximate (asymptotic) distribution of the parameter estimates is normal and furthermore we can find the approximate standard error of the estimate.

A (1-α)100% confidence interval for the slope β is given by:

$$b \pm Z_{1-\alpha/2} se(b)$$

The ratio of the odds for a value of the explanatory variable equal to x + 1 to the odds for a value of the explanatory variable equal to x is the odds ratio. A (1-a) 100% confidence interval for the odds ratio is obtained by taking the anti-log of the lower and upper limits of the confidence interval for β, i.e.:

$$\text{lower limit of the CI for odds ratio} = \exp(\mathrm{b} - \mathrm{Z}_{1\text{-}\alpha/2} se(b))$$
$$\text{upper limit of the CI for odds ratio} = \exp(\mathrm{b} + \mathrm{Z}_{1\text{-}\alpha/2} se(b))$$

Example: A study was conducted to identify factors for the probability of high systolic blood pressure in a community. Sample of 490 individuals were included and measurements of age, sex (0=Female, 1=Make), smoking status (1=Never, 2=Ex-smoker, 3=Current smoker) and BMI (kg/m^2) were measured together with systolic blood pressure. In this study high systolic blood pressure was defined as systolic BP >= 140 mm Hg. The model used was a binary logistic regression model and the two models with method ENTER (Model 1) and with method BACKWARD LR (Model 2) are as follows.

Model 1 – All exposure variables or covariates included

Variables in the model	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
							Lower	Upper
BMI	0.113	0.025	20.712	1	.000	1.119	1.066	1.175
Male	1.058	0.242	19.141	1	.000	2.880	1.793	4.625
Smoking			3.164	2	.206			
Ex-smokers	0.135	0.275	.243	1	.622	1.145	0.668	1.962
Current smoker	0.519	0.297	3.061	1	.080	1.681	0.940	3.006
Age in years	0.056	0.012	23.552	1	.000	1.058	1.034	1.082
Constant	-7.403	0.878	71.131	1	.000	.001		

As you can see, in model 1, all the variables are included irrespective of whether the variable showed statistically significant association with the probability of elevated SBP or not.

Model 2: Only significant exposure variables or covariates included

Variables in the model	B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
							Lower	Upper
BMI	0.114	.025	21.674	1	.000	1.121	1.068	1.176
Male	1.070	.241	19.760	1	.000	2.914	1.818	4.670
Age in years	0.056	.012	23.578	1	.000	1.058	1.034	1.082
Constant	-7.262	.862	71.060	1	.000	.001		

On the other hand, in model 2, the variables included are only those that showed statistically significant association with the probability of elevated SBP.

What do you observe form the two models with respect to the regression coefficients?

Likelihood Ratio Test: An alternative approach is to follow the Analysis of Variance method. The basis of this is to have a measure of model fit which measures the discrepancy between the model and the data. For normal models this is the Residual Sum of Squares. Testing a parameter then is based on how much this measure of discrepancy is reduced when this parameter is introduced into the model. For non-normal models this measure is called the Deviance and with count data is often called Chi-squared Goodness of Fit test. In general the Deviance is based on the value of the (maximized) Likelihood or a log transformation of it. The test is then based on the reduction in this measure of t. A good approximation to the distribution of this reduction is the Chi-squared distribution. Thus we can test the significance of parameters using a Chi-Squared test. This approximation is more reliable than the Normal approximation described earlier. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L1) over the maximized value of the likelihood function for the simpler model (L0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{Lo}{L_1}\right) = -2[\log(Lo) - \log(L_1) \cong \chi^2 \text{ with k - 1 degrees of freedom,}$$
$$\text{where k is the number of parameters in the model}$$

This log transformation of the likelihood functions yields a chi-squared statistic. This is the recommended test statistic to use when building a model through backward stepwise elimination. The higher the log likelihood value the best the model is.

Hosmer-Lemshow Goodness of Fit Test: The Hosmer-Lemshow statistic evaluates the goodness of fit by creating 10 ordered groups of subjects and then compares the number actually in each group (observed) to the number predicted by the logistic regression model (predicted). Thus, the test statistic is a chi- square statistic with a desirable outcome of non significance, indicating that the model prediction does not significantly differ from the observed.

References

1. Wayne W. Daniel. Biostatistics A Foundation for Analysis in the Health Sciences: eighth edition
2. Stephen C. Newman. Biostatistical Methods in Epidemiology: A Wiley-Inter science Publication
3. Michael R. Chernick& Robert H. Friis. Introductory Biostatistics for the Health Sciences: A John Wiley &Sons publication
4. Ann A. O’Connell (2006). Logistic Regression Models for Ordinal Response Variables: Sage Publications
5. R. B. D’Agostino.Tutorials in Biostatistics: Statistical Methods in Clinical Studies
6. Martin Bland. An introduction to Medical Statistics

7. Daniel W. Biostatistics a foundation for analysis in the Health Sciences
8. Kirkwood BR. Essentials of Medical Statistics
9. Knapp RG, Miller MC. Clinical epidemiology and Biostatistics. Baltimore Williams and Wilkins, 1992
10. P. Armitage& G. Berry. Statistical Methods in Medical Research

Sample Time schedule of Leadership strategic information program

On the second module

Week 1	**Monday**	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
08:30 – 09:00 am	Registration	Introduction to analytical epidemiology	Measures of Association	Causality	Chi-squer test And correlation
09:00 – 10:00 am	Introduction to course				
10:00 – 10:30 am	**Break**				
10:30 – 12:30 pm	Review projects Mentors	Analytic study design: Case-control study,	Measures of impact	Screening	Introduction Linear regression
12:30 – 02:00 pm	**Lunch**				
02:00 – 03:30 pm	Review projects mentors	Analytic study design: cohort study,	Chance, bias, confounding and effect modification	Inferencial statistics (estimation and Hypotesis testing)	Introduction Logistic regression
03:30 – 04:00 am	**Break**				
04:00 – 05:00 pm	Review projects mentors	Exercise Mentor	Exercise Mentor	Exercise Mentor	Exercise Mentor

* Note: the exercises in the afternoon are specifically for the information taught in the morning session of that day.

Week 2	Monday	Tuesday	Wednesday	Thursday	Friday
08:30 – 10:00 am	Introduction to Epi-info	Data cleaning/ Transformationta Merging, data comparision	Introduction to SPSS	Data Coding and decoding, cleaning	Protocol Presentation
10:00 – 10:30 am	Break				
10:30 – 12:30 pm	Make View (Develop templet)	Data analysis using EPI-Info	Data entry Difining variable	Data analysis using SPSS	Protocol Presentation
12:30 – 02:00 pm	Lunch				
02:00 – 03:30 pm	Data entry uing the developed templet (data entry)	Data manupolation and creat data base Data exercise	Data transformation from EPI info to SPSS. Data management	Data analysis(SPSS) and interpretation of the result	Protocol Presentation
03:30 – 04:00 am	Break				
04:00 – 05:00 pm	Study Protocol	Study Protocol	Study Protocol	Exercise for SPSS	Protocol Presentation

*Note: For the data exercises participants will use either their data or a dataset provided to them. Participants will work in Epi-info/ SPSS.