**2**

Measuring the Distribution of

Health and Disease

# Introduction

Unit 2 introduces concepts, terms and methods to describe how health and disease statuses (*health outcomes*) are distributed in a population, and how this information can be effectively summarised using numbers and graphs.

Session 1 explains how epidemiologists describe the distribution of health outcomes in terms of *person*, *place* and *time*, and introduces the most common *measures of disease occurrence*.

Session 2 presents the most important indices (*demographic indices*) used to characterise and compare populations.

Sessions 3 to 5 explain how a set of techniques drawn from the discipline of Statistics (Statistics applied to the health sciences, or *Biostatistics*) are used (1) to summarise the distribution of health outcomes in a population; and (2) to obtain reliable estimates of a population’s characteristics, by only looking at a small portion (*sample*) of the population.

### Unit 2 – Session 1: Descriptive Epidemiology

## Introduction

## Descriptive epidemiology is primarily concerned with quantifying the presence of a disease (or, more generally, of any health outcome), and describing it in terms of the personal characteristics and behaviour of those at risk, and the place and timing of occurrence.

## Beside its immediate public health utility (e.g. for allocating effectively health system resources), knowing the distribution of health statuses in a population also offer insights on their possible determinants (which study is the purpose of analytic epidemiology), based on the fundamental assumption in epidemiology that health and disease do not occur at random.

This session introduces the categories used in epidemiology to characterise the distribution of health outcomes and the fundamental concept of population at risk; and explains how this distribution is summarised using a set of indicators known as *measures of disease occurrence*.

## Timing

In this session you will have to do seven readings, and one task involving answering questions about one of the readings. It will probably take you about three to four hours to complete, as the readings are not very long.

## Learning outcomes

* Understand aims and scope of descriptive epidemiology;
* Know how data about health outcome distributions are organised in epidemiology;
* Define counts, proportions, ratios, rates, person-time, risk, odds;
* Distinguish between incidence and prevalence and know their relationship;
* Calculate measures of disease distribution and understand their relationships.

Contents

1. Populations, health outcomes and populations at risk
2. Person, time and place
3. Basic concepts: counts, proportions, ratios, rates
4. Measures of occurrence: incidence, prevalence and their interrelationship

1 Populations, health outcomes and populations at risk

The first reading in this session, by Aschengrau and Seage (2013) introduces and explains what epidemiologists mean by the term *population*. Essential to epidemiology and public health is the concept of a population. In epidemiology the word population is used rather differently to its everyday use. In everyday use population refers to all the people in a country or a fixed geographical area. In epidemiological use a p*opulation* is a group of people (or other living beings) who share some common characteristic. This characteristic could literally be anything such as being the same age, being of the same gender, attend the same school, play the same sport, live in the same neighbourhood, have the same disease, attend the same hospital, etc. Epidemiology then seeks to measure the health status of any defined population and to assess which factors affecting health impact on that population (causing good health or causing bad health and what types of bad health) and what health outcomes those factors affect.

A *health outcome* is any health-related status or event. Traditionally, health outcomes of interest in epidemiology are represented by diseases. As a consequence, measures of occurrence of health outcomes are indicated as measures of *disease* occurrence. In some of the readings and examples in this course this terminology is used, and the student is advised to interpret the word *disease* in the broad sense of health outcome.

From the point of view of descriptive epidemiology, what is fundamental is that health outcomes must be defined precisely and in *operational terms*, i.e. epidemiological study must give unambiguous *rules to distinguish between subjects who have the health outcome and subject who don’t have it*.

Given an unambiguous definition of health outcome (or disease), we can finally introduce the concept of population at risk. *A population at risk* is a population whose individuals are susceptible to a particular health outcome: the population of non-immunized children in a country is an example of population at risk for measles.

The concept of population at risk is a relative concept, and depends on the health event considered, which must be specified, because a population at risk for a health outcome is not necessarily at risk for another. A precise definition of *population at risk* is necessary to calculate unambiguously (and correctly) the measures of diseases occurrence.

A discussion of the concepts of population, health outcomes and ways of measuring health outcomes is provided in the reading by Aschengrau and Seage (2013). **Go through that reading now from page 33 to 39** and immerse yourself in and familiarise yourself with these concepts. An excellent grasp of these concepts is essential in order for you to understand epidemiology.

**Reading**

Aschengrau, A. &Seage, G. R. (2013).Measures of disease frequency. In *Essentials of Epidemiology in Public Health* (3rd edition). Burlington, MA: Jones & Bartlett Learning. **p 33 -39.**

### 2 Person, time and place

The reading by Friis (2009) introduces the field of descriptive epidemiology, its scope and aims:

**Reading**

Friis, R. H. (2009). Descriptive epidemiology: patterns of disease - person, place, time. In *Epidemiology 101*. Burlington, MA: Jones & Bartlett Publishers. p. 65-67

In descriptive epidemiology, data are organised and summarised according to *time*, *place*, and *person*. These three characteristics are sometimes called *epidemiologic variables*:

**Time**

The distribution of a disease, and of health outcomes in general in a population, changes over time. Some of these changes occur regularly and can be predicted. For example, the seasonal increase of influenza cases with the onset of cold weather is a pattern familiar to nearly everyone. By knowing when flu outbreaks will occur, health care professionals can time flu shot campaigns effectively. Other disease rates change unpredictably. By examining events that precede an increase or decrease in disease rates, we may identify causes and appropriate actions to control or prevent further occurrences of the disease.

Depending on what condition we are describing, we may be interested in a period of years or decades, or we may limit the period to days, weeks, or months when the number of cases reported is greater than normal. For many chronic diseases, for example, we are interested in long-term changes in the number of cases or in the rate of the condition. For other conditions, we may find it more revealing to look at the occurrence of the condition by season, month, day of the week, or even time of day. For a newly recognized problem, we need to assess the occurrence of the problem over time in a variety of ways until we discover the most appropriate and revealing time period to use.

**Place**

We describe a health event by ‘place’ to gain insight into the geographical distribution of a problem. For place, we may use place of residence, birthplace, place of employment, school district, hospital unit, etc., depending on which may be related to the occurrence of the health event. Similarly, we may use large or small geographic units: country, state, district, census tract, street address, map coordinates, or some other standard geographical designation. Sometimes we may find it useful to analyse data according to place categories such as urban or rural, domestic or foreign, and institutional or non-institutional.

**Person**

In descriptive epidemiology, when we organize or analyse data by ‘person’ there are several categories available. We may use inherent characteristics of people (for example, age, race, sex), their acquired characteristics (immune or marital status), their activities (occupation, leisure activities, use of medications/tobacco/drugs), or conditions under which they live (socioeconomic status, access to medical care). These categories determine to a large degree who are at greatest risk of experiencing some undesirable health condition, such as becoming infected with a particular disease organism.

The next reading by Friis (2009) further explains these fundamental concepts, and gives various illustrative examples:

**Reading**

Friis, R. H. (2009). Descriptive epidemiology: patterns of disease - person, place, time. In *Epidemiology 101*. Burlington, MA: Jones & Bartlett Publishers. p. 69-84

### 3 Basic concepts: counts, proportions, ratios, rates

When epidemiologists measure disease occurrence, in most cases they do so in terms of *fractions* or *ratios*, i.e. numbers obtained dividing a quantity (*numerator*) by another (*denominator*). In ratios used in epidemiology, the numerator is very often a count of number of events/people (e.g. the number of people who have flu), while the denominator is in some way related to the population at risk.

Read the following excerpt from Aschengrau and Seage (2013) **from page 39 starting at the section “Measuring Disease Occurrence” to page 41.**

**Reading**

Aschengrau, A., Seage, G. R. (2013).Measures of disease frequency. In *Essentials of Epidemiology in Public Health* (3rd edition). Burlington, MA: Jones & Bartlett Learning. **p 39-41.**

The reading explains why fractions are so common, and presents the different types of ratios we can encounter in epidemiology, which are:

*Proportions*: ratios in which the numerator is included in the denominator (e.g. the number of low birth weight out of total number of newborns at Hospital X in 2012). Proportion can only assume values between 0 and 1, and are often expressed in percentage (0-100%);

*Rates*: ratios in which the denominator includes some time component (e.g. the number of patients admitted in a hospital *per month*). Ratios express *how quickly* something happens, such as development of disease. It is important to notice that, despite this precise definition, for historical reasons some measures of disease occurrence that are not technically rates, are, nevertheless, classified in this way (a classical example is the Infant Mortality Rate).

### 4 Measures of occurrence: incidence, prevalence and their

### interrelationship

**Measures of frequency**

There are two main measures of disease frequency:

a. **Prevalence -** measures *existing* cases of disease at a specific point or period in time:

* *Point prevalence* measures the frequency of existing disease in a defined population at a single point in time.

Example: Of 10000 female residents in town A on the first of March 2012, 2000 have hypertension. The prevalence of hypertension among women in town A on this date is calculated as:

* *Period prevalence* is the number of individuals identified as cases during a specified period of time, divided by the total number of people in that population.

The point or period in time that prevalence refers to should always be clearly stated. *Prevalence is a proportion, so has no units*.

Prevalence is a useful measure to quantify the burden of disease in a population. This is of particular use when planning health services. However, prevalence is not a useful measure for establishing the determinants of disease in a population.

b. **Incidence -** measures *new* cases of disease in a specific period. Prevalence measures the frequency of existing cases of disease in a population. In contrast, incidence is a measure of the number of new cases of a disease or other health outcome that develop in a population of individuals at risk, during a specified time period.

The most commonly used measures of incidence are:

* *Cumulative Incidence* (CI): is related to the population at risk at the beginning of the study period. This is also known as risk. It is defined as the proportion of individuals in a population initially free of disease who develop the disease within a specified time interval. Incidence risk is expressed as a percentage (or, if small, as per 1000 persons).

The denominator is the total number of people who were free of disease at the start of the study period, i.e. the *population at risk* at that point in time. The incidence risk assumes that the entire population at risk at the beginning of the study period has been followed for the specified time period for the development of the outcome under investigation. This is called a closed population.

When people enter and leave a population (which is always the case in real life, if the period of study is long enough: some may develop the outcome of interest, or be lost during follow-up for a variety of other reasons, including death), the population is called a *dynamic* population, and the analysis of the study data must take into account this aspect.

* *Incidence Rate (IR)*: is related to a more precise measure of the population at risk during the study period and is measured in person time units. This measurement that seeks to account for varying time periods of follow up, which may occur for the reasons outlined above. Incidence rates also measure the frequency of new cases of disease in a population, but take into account the sum of the time that each participant remained under observation and at risk of developing the outcome under investigation.

*Calculating person-time at risk -* In a dynamic population, individuals in the group may have been at risk for different lengths of time, so instead of counting the total number of individuals in the population at the start of the study, the time each individual spends in the study before developing the outcome of interest needs to be calculated.

The denominator in an incidence rate is the sum of each individual's time at risk (i.e. the length of time they were followed up in the study) and is commonly expressed as *person years at risk*. The incidence rate is the rate of contracting the disease among those still at risk. When a study subject develops the disease, dies or leaves the study, they are no longer at risk and will no longer contribute person-time units at risk.

Note that for most rare diseases, risks and rates are numerically similar because the number at risk of developing the disease will approximately equal the total population at all times.

* *Odds*: are related to the population at risk at the end of the study period.

Another method of measuring incidence is to calculate the odds of disease. Instead of the number of individuals who are disease-free at the start of the study, odds are calculated using the number disease-free at the end of the time period.

**The relationship between prevalence and incidence**

The proportion of the population that has a disease at a point in time (prevalence) and the rate of occurrence of new disease during a period of time (incidence) are closely related.

Prevalence (P) depends on:

1. The incidence rate (IR)
2. The duration of disease(D)

For example, if the incidence of a disease is low but the duration of disease (i.e. time until recovery or death) is long, the prevalence will be high relative to the incidence. An example of this would be diabetes.

Conversely, if the incidence of a disease is high and the duration of the disease is short, the prevalence will be low relative to the incidence. An example of this would be influenza.

A change in the duration of a disease, for example the development of a new treatment that prevents death but does not result in a cure, will lead to an increase in prevalence. Fatal diseases or diseases from which a rapid recovery is common have a low prevalence, whereas diseases with a low incidence may have a high prevalence if they are incurable but rarely fatal and have a long duration.

The relationship between incidence and prevalence, in a population in steady state in which the incidence rate does not change over time, can be expressed as;

When the disease is rare (P<5-10%), the formula above simplifies as follows:

i.e. the prevalence of disease approximates to the product of Incidence Rate and duration of the disease.

All these concepts – and examples of their application in the epidemiological and public health practice – are presented in the reading by Aschengrau and Seage (2013) **from page 42 to page 54.**

**Reading**

Aschengrau, A. &Seage, G. R. (2013). Measures of disease frequency. In *Essentials of Epidemiology in Public Health* (3rd edition). Burlington, MA: Jones & Bartlett Learning. **p 42-54.**

Read the Methods and Results sections of the article by Gerritsen *et al.* (2008), then complete Task 1 below.

**Reading**

Gerritsen, A., Kruger, P., van der Loeff, M. & Grobusch, M. (2008).Malaria incidence in Limpopo Province, South Africa, 1998-2007. *Malaria Journal*, 7(1), 162.

|  |
| --- |
| **Task 1 – Answer questions about a reading**  After reading the text, consider and answer these questions   1. What health outcome/disease is this study about? 2. What is the population at risk? 3. Is this disease described in operational terms? Can you explain how the Authors decided whether or not a subject had the disease? 4. How do the authors describe the distribution of the disease of interest? 5. Which measures of occurrence do they use? 6. How do they obtain the number of individuals in the population at risk? |

## Task feedback

1. The health outcome of interest is malaria: the study aims at describing “*the malaria incidence and mortality in Limpopo Province for the seasons 1998–1999 to 2006–2007 and to detect trends over time and place*”.
2. The population at risk is the population of Limpopo Province between 1998 and 2007. Notice that the authors, correctly, define the population ‘north’ in geographic and temporal terms: the population of Limpopo is not a static population, and, therefore, we need to indicate at which period or point in time we are looking.
3. Yes. The first paragraphs of the Methods section describe in details how the surveillance system works and how an individual is defined as having malaria (i.e. how an individual becomes a “case”).
4. They describe the distribution in terms of place (look, for example, at Figure 1), time (Table 1) and person (Table 3, for example, shows malaria incidence by age of the individuals).
5. Incidence rates (and death rates, which are a specific form of incidence rate in which the health outcome is “death by malaria”, rather than malaria disease itself).
6. From the South African Census (adjusted to take into account population dynamics). This is a very common way to obtain denominators for calculation rates in large populations.

Unit 2- Session 2: Describing populations

## Introduction

Epidemiology is about analysing distribution and determinants of health and disease in populations. This session will present methods and summary measures to describe and compare characteristics of populations as a whole.

The concept of age structure of a population and its graphical representation as population pyramid is presented first, followed by a list of the most widespread numerical indicators used to convey various other characteristics of populations and its health status. Finally, the problem of comparing morbidity and mortality indices across populations with different age structure is presented and methods of standardization are introduced as a possible solution.

Timing

You are asked to do readings of three texts and a website, as well as two tasks in this session. One task involves analysing two data sets and using them to construct a diagram, which could take you about an hour in itself. The readings are not particularly long, so you should be able to complete the session in about four hours.

## Learning outcomes

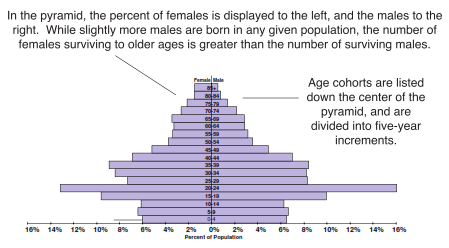
* Define and interpret the most important demographic indices and summary measures of population health;
* Understand concept and methods of direct and indirect adjustment of rates

Contents

1. Age structure and population pyramids
2. Most important demographical indices and summary measures of population health
3. Crude, Adjusted and Specific rates

### 1 Age structure and population pyramids

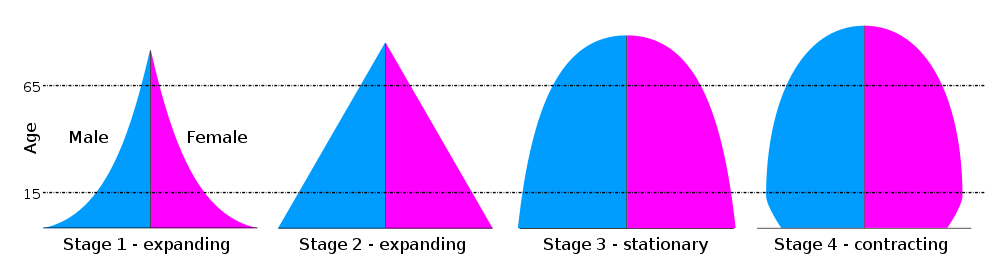
One of the tools that demographers use to understand population is the *age structure diagram*, commonly known as *population (or age) pyramid* despite the fact that it is not always pyramidal in shape. This diagram shows the distribution by ages of females and males within a certain population in graphic form, and it is described in the following figure:



Source: M. Richmond, Population Pyramids. In: Willamette River Basin Atlas, 2nd Edition. Available at: <http://www.fsl.orst.edu/pnwerc/wrb/Atlas_web_compressed/5.Human_Populations/5h.pyramids_web.pdf> [Accessed 2 November 2013]

Age structure diagrams can be easily constructed form data on age and sex distribution for a population: as indicated in the figure, the length of each bar is proportional to the fraction of the population belonging to the specific age class.

The shape of the diagram gives an immediate visual indication on the expected population trends: for example, if the diagram shows a pyramidal shape, then one can expect a rapid rise in population; if the diagram shows a generally straight up and down shape except for the older age groups, a stable population is thus revealed; if the diagram shows a top-heavy shape, then a decline is forecast for that population. The following figure summarises these concepts:



<http://en.wikipedia.org/wiki/File:DTM_Pyramids.svg>

**Task 1**

Using the data in the table below construct the population pyramid for Canada and Bolivia in 2010. Which kind of population trends do these pyramid show?

**Canada, 2010**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Male Population** | **Female Population** | **Male Percent** | **Female Percent** |
| 0-4 | 883748 | 840839 | 5.3 | 4.9 |
| 5-9 | 892543 | 851811 | 5.3 | 5 |
| 10-14 | 970935 | 920293 | 5.8 | 5.4 |
| 15-19 | 1145267 | 1083618 | 6.8 | 6.4 |
| 20-24 | 1164117 | 1101324 | 6.9 | 6.5 |
| 25-29 | 1169343 | 1113625 | 7 | 6.5 |
| 30-34 | 1132104 | 1085771 | 6.8 | 6.4 |
| 35-39 | 1114527 | 1080367 | 6.7 | 6.4 |
| 40-44 | 1172387 | 1168062 | 7 | 6.9 |
| 45-49 | 1371767 | 1353600 | 8.2 | 8 |
| 50-54 | 1308918 | 1303784 | 7.8 | 7.7 |
| 55-59 | 1136699 | 1157779 | 6.8 | 6.8 |
| 60-64 | 997109 | 1008582 | 6 | 5.9 |
| 65-69 | 732555 | 770146 | 4.4 | 4.5 |
| 70-74 | 554429 | 631541 | 3.3 | 3.7 |
| 75-79 | 444091 | 562087 | 2.7 | 3.3 |
| 80-84 | 307948 | 460528 | 1.8 | 2.7 |
| 85-89 | 161849 | 295874 | 1 | 1.7 |
| 90-94 | 68302 | 149561 | 0.4 | 0.9 |
| 95-99 | 19471 | 55365 | 0.1 | 0.3 |
| 100+ | 3283 | 13793 | 0 | 0.1 |

**Bolivia, 2010**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Age** | **Male Population** | **Female Population** | **Male Percent** | **Female Percent** |
| 0-4 | 614845 | 590312 | 12.5 | 11.7 |
| 5-9 | 601952 | 579465 | 12.2 | 11.5 |
| 10-14 | 559682 | 540794 | 11.4 | 10.8 |
| 15-19 | 533519 | 517781 | 10.8 | 10.3 |
| 20-24 | 495749 | 488030 | 10.1 | 9.7 |
| 25-29 | 433864 | 433785 | 8.8 | 8.6 |
| 30-34 | 354823 | 362967 | 7.2 | 7.2 |
| 35-39 | 285258 | 306276 | 5.8 | 6.1 |
| 40-44 | 224303 | 255070 | 4.6 | 5.1 |
| 45-49 | 196734 | 223518 | 4 | 4.4 |
| 50-54 | 171693 | 193513 | 3.5 | 3.9 |
| 55-59 | 140256 | 157422 | 2.8 | 3.1 |
| 60-64 | 110192 | 123050 | 2.2 | 2.4 |
| 65-69 | 77966 | 90126 | 1.6 | 1.8 |
| 70-74 | 50064 | 61516 | 1 | 1.2 |
| 75-79 | 36109 | 46719 | 0.7 | 0.9 |
| 80-84 | 23165 | 32295 | 0.5 | 0.6 |
| 85-89 | 9801 | 15527 | 0.2 | 0.3 |
| 90-94 | 2636 | 4973 | 0.1 | 0.1 |
| 95-99 | 460 | 1052 | 0 | 0 |
| 100+ | 41 | 115 | 0 | 0 |

(Source: U.S. Census Bureau, International Data Base)

**Task Feedback**

Use the website below to check your results. It provides an interactive tool to draw population pyramids for all countries in the world from 1950 onwards, with estimates up until 2100. (You can also use it to compare population pyramids for different countries and periods and “get a feel” of what is happening across the different regions of the world.)

populationpyramid.net [*http://populationpyramid.net*/](http://populationpyramid.net/) [accessed 04/10/13]

Comparing the two pyramids you can see that Bolivia is clearly a country in expansion, while Canada can be possibly classified at the boundary between a stationary and a contractist stage in its population dynamics.

Irregularities and/or sex discrepancies in the shape of the diagram are also useful indicators of transient demographic phenomena, for example change in population structure due to epidemics. See, for example, the following excerpt from an article by Oramasionwu *et al.* (2011) on the demographic effects of the AIDS epidemics in sub-Saharan Africa:

**Reading**

Oramasionwu, C. U., Daniels, K. R., Labreche, M. J., & Frei, C. R. (2011). The environmental and social influences of HIV/AIDS in sub-Saharan Africa: a focus on rural communities. *International journal of environmental research and public health.* 8(7), 2971–73.

### 2 Most important demographical indices and summary measures of

### population health

A set of numerical indices are commonly used to describe population dynamics and characteristics and to summarise its global health status.

The reading by Bonita *et al*. (2006) provides definitions and examples for the most widespread of these, namely: birth and fertility rate; mortality; morbidity; life expectancy; disability; years of potential life lost (YPLL); quality-adjusted life years (QALY); and disability-adjusted life years (DALY). Further indices of epidemiological importance related to specific disease are also defined: case fatality ratio; 5-year survival; survival rates; and median survival time.

**Reading**

Bonita, R., Beaglehole, R. & Kjellstrom, T. (2006). Measuring health and disease. In *Basic Epidemiology*. Geneva: WHO. p. 23-33

**(This is your prescribed textbook. You were given an electronic copy of it.)**

### 3 Crude, Adjusted and Specific rates

Often it is of epidemiological interest to compare measures of disease occurrence such as rates or ratios across different populations. However, comparing *crude* rates (i.e. rates calculated for each population with the usual formulæ) is often not correct, because it doesn't take into account possible confounding factors.

A confounding factor is basically another risk factor for the outcome of interest that is also unequally distributed among the populations being compared. This is often the case of age, which is a risk factors for almost any disease (i.e. the risk of disease is different in different age groups), when the age structure of the populations is different.

Take the example of comparing cancer mortality between population A and population B, where the crude mortality ratio for cancer in population A is greater that the crude mortality ratio for cancer in population B. Does it mean that the risk of death for cancer for a person in A is greater than the risk for a person in B? Not necessarily! We know that cancer mortality is related to age (it increases with age), so if the proportion of older people in population A is greater than in population B, it might be that the risk of death due to cancer in population A only appears to be greater simply because it has a greater percentage of old people who have an inherently greater risk of dying. In other words, the crude mortality rate for population A might be higher just because it is weighted more heavily with old people.

To overcome this fallacy, various techniques are available to adjust a crude rate to take into account the confounding effect of age (or, similarly, of other confounders) making cross-population comparison meaningful. Among these, standardisation is one of the most common techniques. Standardisation consists in calculating, for a population A, an *adjusted rate* which represents the rate of disease that A would have if its age structure were the same as the age structure of a *reference population* R. In doing that, the rates are no longer dependent on the specific age structure of the population, and can be compared meaningfully.

Different methods (*direct* and *indirect* standardisation) for calculating adjusted rates as a weighted average of *age-specific rates* (i.e. rated for a specific age group) are detailed in the next reading, from the Epidemiological Bulletin (2002):

**Reading**

Epidemiological Bulletin. (2002). Standardization: A Classic Epidemiological Method for the Comparison of Rates. Analysis Group of PAHO’s Special Program for Health Analysis. *Epidemiological Bulletin / PAHO*, 23(3), 9–12.

## Task 2

Repeat the calculations in Box 2 in the *Epidemiological Bulletin* (2002) reading (age-standardised mortality rates for the US and Mexico), using the reference population from the table below.

**Standard Population**

|  |  |
| --- | --- |
| <1 | 2400 |
| 1-4 | 7600 |
| 5-14 | 15000 |
| 15-24 | 17000 |
| 25-44 | 26000 |
| 45-64 | 19000 |
| 65+ | 14000 |
| Total | 100000 |

**Task Feedback**

The standardised mortality rates become higher in both countries, because they are now referred to a much older standard population (the proportion of people 65+ has doubled), and the specific mortality rates are higher for higher age groups (the new values are for 9.8 deaths per 1000 pop for Mexico and 9.3 deaths per 1000 pop for the US).

Unit 2 – Session 3: Descriptive Statistics

## Introduction

The first step of *any* data analysis deals with the *description of the data* itself, i.e. with giving an overall “picture” of the data which have been collected in a sample of a population. When data are collected on the whole population and the aim of the study is purely descriptive, this is the only required step, but when – as in the majority of cases – the objective is to use the sample to infer characteristics of the population, a preliminary description of the sample is necessary to choose the appropriate inferential techniques.

Considering that in real-life studies the quantity of data usually available is large, their description must involve some form of “summary” that make their interpretation possible. As every time we “summarise” something we are losing information, describing data involves a trade-off between interpretability (little information can be interpreted more easily than a large quantity of data) and quantity of information (discarding details in order to summarise lead to a loss of information).

The example in the figure below may clarify the concept:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Individual #** | **SBP (mmHg)** | | 1 | 80 | | 2 | 81 | | .... |  | | 100 | 95 |  |  |  | | --- | --- | | **Blood pressure range** | **# of individuals** | | ≤ 80 mmHg | 70 | | >80mmHg; <90mmHg | 10 | | ≥ 90 mmHg | 20 |   **Mean blood pressure of the sample:** 75 mmHg |
| INTERPRETABILITY |
| INFORMATION CONTENT |

If our original data is systolic blood pressure (SBP) of 100 individuals in a sample, a description of these data can be:

(a) a list of all blood pressure readings;

(b) a shorter list constituted by the number of individual whose blood pressure is below 80 *mmHg*, between 80 and 90 *mmHg*, and above 90 *mmHg* (i.e. normotensive, pre-hypertensive and hypertensive according to the usual classification);

(3) a single number representing the mean blood pressure of the individual of the sample.

It is clear that description (a) is the most complete (i.e. it carries the whole set of information we collected), while description (c) discards most of the information (e.g., we don’t know anymore the blood pressure of individual #1), but it is also clear that to interpret a list of 100 numbers is much more difficult than interpreting (i.e. understanding the implications of)a list of three numbers as in (b) or the mean as in (c).

*Descriptive statistics* (one of the two main subfield of statistics, the other being inferential statistics) offers a set of techniques to practically realise the “summary” of data in a meaningful way, i.e. minimising the loss of information.

This session will introduce the main descriptive methods (both *numerical* and *graphic*) and a basic guide on how to “get a feel” for a set of data we don’t know (more formally, how to perform a basic *exploratory data analysis*).

Timing

This is a long session. There is only one reading, but there are eight tasks that involve practising summarising numerical and categorical variables using numbers and tables, and analysing datasets. You will probably need at least six hours to complete the session.

## Learning outcomes

* Understand the purpose of descriptive statistics;
* Summarise categorical and numerical data using numbers and graphs;
* Perform a basic exploratory data analysis.

## Contents

1. Summarising categorical and numerical data with numbers

### Graphs and charts in presentation of categorical and numerical variables

1. Exploratory data analysis

### 1 Summarising categorical and numerical data with numbers

Numerical data (and, partially, ordinal data) can be described using summary measures (*descriptive statistics*) which indicate the relative position of the bulk of values and how much they are spread. *Measures of central tendency* (mean, median and mode) accomplish the first objective – to establish *position*- while *measures of dispersion* (variance, standard deviation, range and interquartile range) accomplish the second one– to establish *spread*.

Categorical data in a nominal scale cannot in general be described using measures of central tendency and dispersion, but they are summarised giving their *frequency distribution*, i.e. a list of the number of observations falling in each category. For ordinal variables, it is also possible (and meaningful) to include in the table the *cumulative frequencies*, i.e. the sum of the number of observations whose value is equal to or lower than a specific value, as for example in the table below, representing the age distribution of the 450 respondents to a population survey:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Age of respondent | Frequency | Frequency (%) | Cumulative Frequency | Cumulative Frequency (%) |
|  |  |  |  |  |
| 15-24 | 120 | 26.6% | 120 | 26.6% |
| 25-44 | 201 | 44.7% | 120+201= 321 | 26.6+44.7=71.3% |
| 45-64 | 102 | 22.7% | 120+201+102=423 | 26.6+44.7+22.7=94% |
| 65+ | 27 | 6.0% | 120+201+102+27=450 | 26.6+44.7+22.7+6=100% |
|  |  |  |  |  |
| Total | 450 | 100% |  |  |

The techniques of producing frequency distribution tables are sometimes called *data reduction* techniques.

### 2 Graphs and charts in presentation of categorical and numerical variables

Depending of the scale of measurement (numerical, ordinal or categorical) and the purpose of the summary, different types of graphical representation are available. Among these, *histograms*, *bar and pie charts*, *box-and-whisker plots*, *scatter plots* and *line graphs*, are commonly used in epidemiology.

The reading below by Dawson and Trapp (2004) presents a formal definition of the descriptive measures mentioned above and explains how all these measures can be calculated. It also explains how to build and interpret *histograms*, *bar and pie charts*, *box-and-whisker plots*, *scatter plots* and *line graphs*, and gives indications about “good practices” in graphical representation to avoid conveying misleading information.

**Reading**

Dawson, B., &Trapp, R. G. (2004).*Basic & clinical biostatistics* (4th ed.). New York: McGraw-Hill. p 27-42.

The following tasks will allow you to practice summarising numerical and categorical variables using numbers and tables:

##### TASK 1 – Calculate measures of central tendency

Clarify the concepts *Measures of central tendency*, which include *mean, median* and *mode*; refer to the readings noted above. Use the index to find these concepts and then answer these questions in the tables below:

a) What type of variables are these, and

b) Calculate the mean, median and mode for the age, distance to work and income variables in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Record No.** | **Age (yrs)** | **Distance to Work (km)** | **Income (R) per month** |
|  |  |  |  |
| 1 | 23 | 13 | 1 500 |
| 2 | 31 | 19 | 500 |
| 3 | 55 | 15 | 3 500 |
| 4 | 43 | 35 | 3 500 |
| 5 | 55 | 76 | 6 000 |
| 6 | 19 | 44 | 700 |
| 7 | 17 | 23 | 400 |
| 8 | 44 | 14 | 1 500 |
| 9 | 43 | 6 | 9 000 |
| 10 | 37 | 55 | 700 |

**Task Feedback**

###### The table is completed as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Age (yrs)** | **Distance to Work (km)** | **Income (R) per month** |
| Type/scale | Numeric interval | Numeric interval | Numeric interval |
| Mean | 36.7 | 30 | 2730 |
| Median | 40 | 21 | 1500 |
| Mode | 43 & 55 | None | 700, 1500 & 3500 |

###### **Type:** All variables in this table are numeric interval data – quantitative, continuous with equal units.

###### **Mean:** This is calculated by summing the numbers in that column and then dividing by the number of measurements, for example the sum of the ten values of age is 367 and 367 divided by the numbers of observations, i.e. 10, equals 36.7.

###### **Median:** To calculate the median, you first order the data from lowest to highest and then select the middle value. In this example there are an even number of observations, so the average between the two middle values are the median. For example, when you order the ages 17, 19, 23, 31, 37, 43, 43, 44, 55, 55, numbers 37 and 43 are the 5th and 6th values (the two middle values). The average of 37 and 43 is 40. This is the median age.

###### **Mode:** This is the most frequently occurring value. In this example, age has two variables appearing twice, i.e. the most frequently occurring; distance has no values that appear more than once; and income has three values that appear twice.

You will need to be able to use these concepts when analysing statistical information. Work with them, developing your own examples until you feel that they are part of your vocabulary.

Note that the most commonly used of these statistics is the mean, also called the average. This is the preferred measure unless there are extreme values in your data, for example the distance to work for the 55 km value was instead 250 km. This would increase the mean from 30 to 49.5. However, note the median would remain 21, as the middle values did not change. In a case where you have extreme measures that are very different from all or most of the other values, it is generally considered better to use the median, instead of the mean, because the median is not changed by these unusual people/values.

### Task 2 – Determine the degree of dispersion

Calculate the range, IQR, variance and standard deviation for the age, distance to work and income variable presented in the table in the previous section.

##### Task feedback

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Age (yrs)** | **Distance to Work (km)** | **Income (R per Month)** |
| RANGE | 17-55 | 6-76 | 400-9000 |
| IQR | 23 | 20 | 2600 |
| VARIANCE | 192 | 498 | 8,095,670 |
| STANDARD DEVIATION | 14 | 22 | 2845 |

###### **Range:** This is the lowest and highest values in the series. These are easily obtained from the ordered (sequenced) list you did above to obtain the median.

###### **IQR:** is the 75th percentile minus the 25th percentile. If you use the ordered list for age that we used before, the value with 25% of the measures below is 23 (3rd value) and the value with 75% below is 44 (8th value). 44 minus 23 equals 23.

###### **Variance and Standard Deviation:** are calculated using the formula above. It is often easier to create a table as follows to get the necessary values:

|  |  |  |
| --- | --- | --- |
| **Observation** | **AGE (xi)** | **AGE Squared (xi2)** |
| 1 | 23 | 529 |
| 2 | 31 | 961 |
| 3 | 55 | 3025 |
| 4 | 43 | 1849 |
| 5 | 55 | 3025 |
| 6 | 19 | 361 |
| 7 | 17 | 289 |
| 8 | 44 | 1936 |
| 9 | 43 | 1849 |
| 10 | 37 | 1369 |
| SUM (Σ) | **(367)2=134689 ÷ 10 = 13468.9** | **15193** |

###### Therefore S2 = (15193 – 13468.9)/(10-1) = (1724.1) / (9) = 191.567 is the variance and the square root of 191.567 = 13.841 is the standard deviation.

##### Task 3 – Reduction of fluoridation data

###### The Fluoride Concentration of Municipal Water Supplies in South Africa is given as:

###### 1.2, 0.03, 0.05, 0.7, 0.5, 0.03, 0.08, 0.7, 0.9, 1.2

Order the measurements, then count the frequency for each concentration and calculate the relative and cumulative frequency distribution. These all can be represented in a single table, also known as a 1-way frequency table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Fl** | **f** | **Rf** | **Cf** |
| 0.03 | 2 | 2/10 = 0.2 or 20% | 2/10 = 0.2 or 20% |
| 0.05 | 1 | 1/10 = 0.1 or 10% | 2+1/10 = 0.3 or 30% |
| 0.08 | 1 | 1/10 = 0.1 or 10% | 2+1+1/10 = 0.4 or 40% |
| 0.5 | 1 | 1/10 = 0.1 or 10% | 2+1+1+1/10 = 0.5 or 50% |
| 0.7 | 2 | 2/10 = 0.2 or 20% | 2+1+1+1+2/10 = 0.7 or 70% |
| 0.9 | 1 | 1/10 = 0.1 or 10% | 2+1+1+1+2+1 = 0.8 or 80% |
| 1.2 | 2 | 2/10 = 0.2 or 20% | 2+1+1+1+2+1+2/10 = 1 or 100% |

**Task 4 – Work out the frequency distribution**

The following are course marks (%) for a class of postgraduate students in 1999.

65, 60, 73, 50, 65, 55, 70, 45, 65, 75, 50, 60, 70, 55, 75, 70, 55, 50, 65, 55, 65, 60, 73, 50, 65, 55, 70, 45, 65, 75, 50, 60, 70, 55, 75, 70, 55, 50, 65, 55

Draw a frequency distribution table of the marks using the same format as the table above.

|  |  |  |  |
| --- | --- | --- | --- |
| Observation | F | Rf | Cf |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

##### Task 5 – Define different forms of data reduction

Define the terms: *Data Reduction, Frequency Distributions, Relative Frequency, Cumulative Frequency and Class Intervals.*

##### Task feedback

###### The concept of *Data Reduction* includes the calculation of *Frequency Distributions*, *Relative Frequencies* and *Cumulative Frequencies*.

###### A *Frequency Distribution* is a set of ordered measurements and their corresponding frequencies. The frequency (f) is the number of times a measurement occurs in a data set. The *Relative Frequency (Rf) Distribution* is the proportion or percentage of frequency of the measurement in the data set. This is calculated by dividing the frequency of the measurement by the total number of observations (f/N).

###### If the data set is very large and become cumbersome to work with, one often has to group or condense the data set into class intervals. The formula used to calculate the number of class intervals is called the Sturges’ Rule (k=1+3.322[log N]), where k is the number of class intervals and N the total number of measurements. The formula used to calculate the width of a class interval is as follows: W=R/k, where W is the width of the class interval, R the highest to lowest measurements and k the number of class intervals. These formulae are used as a guideline for the number and width of class intervals. Statisticians suggest that the number of class intervals should range from 7 to 20. The *Cumulative Frequency (Cf) Distribution* is the successive addition of the number of frequencies.

##### Task 6 – Analyse a set of data

###### The table below emerged from a study of the determinants of heart disease that was carried out.

###### a) Indicate the qualitative/quantitative nature of the variables.

###### b) Calculate the mean, median, mode and standard deviation of the income variable.

###### c) Draw a frequency distribution, Rf and Cf of the heart disease and level of stress variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Income per Month (ZAR)** | **Heart Disease**  **(YES/NO)** | **Marital Status** | **Level of Stress** |
| 3000 | N | 2 | 2 |
| 4500 | Y | 1 | 2 |
| 1500 | N | 1 | 2 |
| 750 | Y | 1 | 1 |
| 12000 | Y | 3 | 1 |
| 450 | Y | 4 | 1 |
| 750 | Y | 4 | 2 |
| 900 | N | 1 | 2 |
| 1100 | N | 1 | 1 |
| 5000 | N | 4 | 2 |
| 350 | N | 1 | 1 |
| 3500 | Y | 3 | 2 |
| 900 | N | 2 | 2 |
| 5500 | Y | 3 | 2 |
| 3000 | N | 1 | 1 |
| 3500 | Y | 2 | 2 |
| 1200 | N | 4 | 2 |
| 800 | Y | 2 | 1 |
| 1200 | N | 2 | 2 |
| 1500 | Y | 4 | 1 |

##### Task feedback

See previous tasks and study session readings.

Some exercises relating to graphic visualization techniques are included in the following tasks. In modern epidemiological practice, graphical representations are almost always produced using software packages, both general purpose (Microsoft Excel© is a good example of a non-statistical program with noteworthy graphical abilities) and specialised (EpiInfo© and all major statistical packages have tools for graphical representation of data). However, students are strongly encouraged to experiment with the task of producing some of these representations manually. Although this is a tedious task, it is extremely useful for learning.

Some exercises relating to graphic visualization techniques are included in the following tasks.

## Task 7

Twenty-five randomly selected students were asked the number of movies they watched the previous month. Look at the table of results below, then answer the questions that follow.

|  |  |  |
| --- | --- | --- |
| Number of movies | Frequency | Cumulative Frequency |
| 0 | 5 |  |
| 1 | 8 |  |
| 2 | 6 |  |
| 3 | 4 |  |
| 4 | 1 |  |
| 18 | 1 |  |
| Total | 25 |  |

1. Find the sample mean and standard deviation;
2. Construct a bar chart of the data;
3. Complete the columns of the table;
4. Find the first quartile;
5. Find the median;
6. Find the third quartile;
7. Construct a boxplot of the data;
8. What percentage of the students saw fewer than three movies?
9. Construct a stem plot of the data.
10. Construct a pie chart of the data.

## Task Feedback

Notice that the variable is numerically discrete, therefore the histogram should have a gap between bars. Using the common cut-off to distinguish outliers in the distributions of number of movies seen (1.5 times IQR above the 3rt quartile or below the first), the value 18 is an outlier, and should be represented separately in the box plot.

### 3 Exploratory data analysis

*Exploratory Data Analysis* – i.e. getting a sense of the data – *must* precede any kind of statistical analysis. The main purposes of exploratory data analysis are:

* Gaining intuition about the data and identify salient features;
* Checking for problems in the data (e.g., missing values, coding mistakes, duplicate entries) which can hinder further analyses;
* Finding outliers (i.e. extreme values);
* Identifying the distribution of the different variables in order to choose the appropriate technique to summarise, represent graphically and further analyse;
* Presenting a meaningful (and useful) summary of the data.

Exploratory data analysis involves the use of both graphical and non-graphical methods, and it is not a mechanical procedure which is always performed in the same way: its details depend on the objectives of the study and on the characteristics of the data itself.

However, a basic general-purpose procedure can be delineated as follows:

* Step 1: Identify scale of measurement of each variable;
* Step 2: Check for obvious coding mistakes and impossible values;
* Step 3: Analyse missing data and outliers (i.e. values that do not fit in the overall distribution);
* Step 4: Determine the distribution of each variable and produce appropriate summary statistics and graphical representations;
* Step 5: Analyse bivariate relationships between couples of variables (we will see what this means and how it can be done in Unit 4).

## Task 8 - Conduct an exploratory data analysis

The simulated dataset below refers to the effects of a new drug in reducing the risk of cardiac events high-risk adults. Using the dataset of 25 observations in the table, conduct an exploratory data analysis according to the steps summarised above (excluding step 5). Answer questions 1 – 6 below the table, to guide your analysis.

*The variables involved are:*

*AGE: age of patient in years*

*GENDER: 1=male; 0=female*

*EVENT: 1= patient experienced a cardiac event in the last year;*

*0= patient did not experience a cardiac event in the last year*

*NEWDRUG: 1=patient uses the new drug;*

*2= patient uses the new drug in addition to a traditional therapy;*

*0= patient is not on treatment;*

*EF: cardiac ejection fraction (a measure of the heart's pumping efficiency): it is expressed in percentage (physiological values vary between 55% and 70%. Values below 35% are life-threatening.)*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient # | AGE | GENDER | EVENT | NEWDRUG | EF |
| 1 | 66 | 10 | 0 | 0 | 65.0 |
| 2 | 60 | 0 | 0 | 1 | 61.3 |
| 3 | 74 | 0 | 0 | 2 | 57.2 |
| 4 | 85 | 0 | 0 | 0 | 72.0 |
| 5 | 65 | 0 | 0 | 0 | 55.0 |
| 6 | 69 | 0 | 0 | 0 | 67.0 |
| 7 | 67 | 1 | 0 | 1 | 57.9 |
| 8 | 59 | 1 | 0 | 1 | 60.0 |
| 9 | 67 | 1 | 1 | 2 | 65.0 |
| 10 | 75 | 1 | 1 | 2 | 53.0 |
| 11 | 70 | 0 | 0 | 1 | 62.0 |
| 12 | 73 | 0 | 0 | 1 | 65.1 |
| 13 | 79 | 1 | 1 | 2 | 55.0 |
| 14 | 65 | 1 | 0 | 0 | 63.0 |
| 15 | 79 | 1 | 1 | 0 | 77.9 |
| 16 | 90 | 0 | 0 | 0 | 51.0 |
| 17 | 43 | 0 | 0 | 1 | 65.0 |
| 18 | 64 | 0 | 0 | 1 | 62.0 |
| 19 | 88 | 1 | 0 | 1 | 65.0 |
| 20 | 76 | 1 | 0 | 1 | 55.0 |
| 21 | 88 | 1 | 0 | 2 | 32.4 |
| 22 | 73 | 0 | 0 | 2 | 50.0 |
| 23 | 75 | 1 | 1 | 1 | 60.0 |
| 24 | 63 | 1 | 0 | 1 | 53.0 |
| 25 | 82 | . | 0 | 2 | 55.8 |

|  |
| --- |
| Answer these questions:   1. Identify the scale of measurement of each variable; 2. Check for impossible or implausible values and exclude those observations from the dataset; 3. Identify possible outliers, but do not eliminate them from the dataset; 4. Produce appropriate graphical visualization(s) of the data (using histograms, bar charts, pie charts, box plots, depending of the type of data); 5. Calculate the appropriate summary statistics for each variable; 6. Write a paragraph for each variable, summarising what you learned about the data. |

## Task feedback

There is no single way to conduct exploratory data analysis. You may send your results to the lecturer, who will give you personal feedback on the results of your analysis.

Unit 2 – Session 4: Study Population and Sampling

## Introduction

Very often epidemiologists interested in some characteristics of a population don’t have the possibility to study *all the individuals* in the population. They only draw a *sample* of the population, and, analysing only the individuals in the sample, they derive (a technical term is *estimate*) the characteristics of the population. This session introduces the basic concepts that underlie and justify the statistical procedures used for *estimation*.

The session first briefly introduces the concepts of *probability*, *probability distribution* and *random variable*, then it presents some examples of *theoretical distributions* of particular importance (including the fundamental *Normal distribution* and its properties). Building on that, it defines the concepts of *sampling variability* and *sampling error,* and describes some important properties of the distribution of the latter when the characteristic of interest is the mean of some variable (these properties derive from the *Central Limit Theorem*).

Timing

In this session you are asked to do two fairly long readings, which are divided into shorter sections. There are also three tasks to complete. The session should take you about three to four hours.

## Learning outcomes

* Understand the concept of probability, probability distribution and random variable
* Know the most common theoretical distributions of categorical and continuous variables
* Know the Normal distribution and its properties
* Understand the concepts of sampling distribution, sampling error and the implication of the Central Limit Theorem

## Contents

1. Probability, probability distributions and random variables
2. Some important theoretical distributions
3. Sampling error, sampling distribution and the central limit theorem

### 1 Probability, probability distributions and random variables

**Probability** measures uncertainty, i.e. the chance of a given event occurring as a result of a set of circumstances (this set of circumstances is called a *random experiment* in the language of probability theory). It is a number that *takes a value from zero to one* (often expressed in a percentage, in which case it varies from 0 to 100). If it is equal to zero, the event cannot occur. If it is equal to one, then the event certainly occurs.

Probabilities can be calculated with different approaches. The approach we follow in this course (which is the most common) is called a *frequentist* approach, in which the probability of an event is defined as *the proportion of times the event would occur if we were to repeat the same random experiment a large number of times* (e.g. the number of times we would get “heads” if we tossed a fair coin 1000 times).

The reading by Petrie and Sabin (2005) briefly explains other approaches to calculating probabilities, and some rules we can use to calculate the probability of a *complex event*. This is an event which is the result of two or more events, e.g. the probability that, throwing two dice, we obtain 6 in the first throw and 4 in the second.

**Reading**

Petrie, A. & Sabin, C. (2005). *Medical Statistics at a Glance* (2nd ed.). Malden: Blackwell: **20**

Using the concept of probability just introduced, we can define two other fundamental concepts in statistics: the concepts of *random variable* and its associated *probability distribution.*

A **random variable** is a quantity that can take any of a set of mutually exclusive values, each of them with a certain probability. A random variable does not have *a single value*, but *a set of different possible values*, each of them has an associated probability. The probabilities associated with each possible value are described by a *probability distribution*, i.e. a table, a graph or a mathematical function which “link” each value to the relative probability. For example, considering the random variable which contains the number we can read on the upper face of a fair dice tossed on a table, its probability distribution can be represented by the following Table 1 or, graphically, by the bar graph in Figure 2.

Table 1

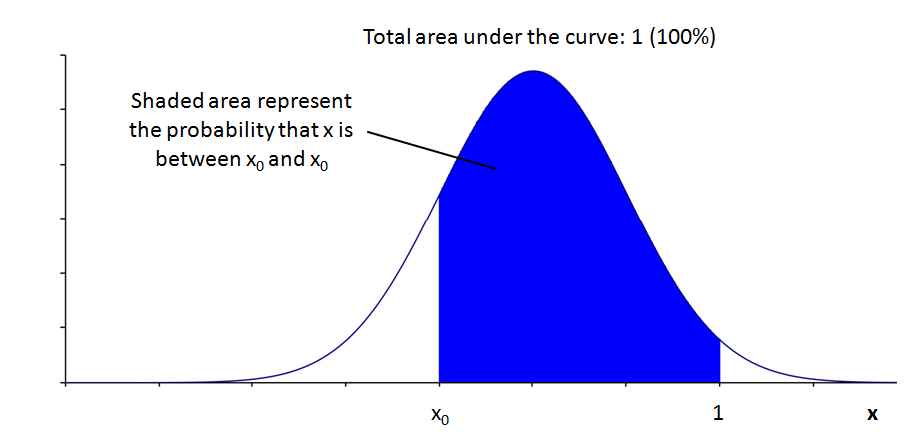
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Value** | **Probability** | | 1 | 1/6=0.166 | | 2 | 1/6=0.166 | | 3 | 1/6=0.166 | | 4 | 1/6=0.166 | | 5 | 1/6=0.166 | | 6 | 1/6=0.166 |   C:\Users\Annibale\Desktop\aa.png |

Fig. 2

Probability distributions are *theoretical distributions* (as opposed to the *empirical distributions*, which represent the observed frequencies of the different values of a variable across the individuals in a sample and which we summarise using tables, graphs or summary measures, as explained in the first sessions of this unit).

Each probability distribution is defined by certain parameters which are summary measures (e.g. mean, variance) characterizing that distribution (i.e. knowledge of them allows the distribution to be fully described). Depending on whether the random variable is discrete or continuous, the probability distribution can be either discrete or continuous. For discrete probability distributions we can derive the probability of each possible value, and the sum of these probabilities is one.

For continuous probabilities distributions, we can only derive the probability that the random variable takes values in certain ranges (because there are infinitely many values of x). In these cases, the probability distribution is represented, rather than by a table as is common for discrete distributions, by a curve (called the *density function*) in which the horizontal axis represents the values of x and the area under the curve between two values x0 and x1 represents the probability that the random variables take values between x0 and x1, as shown in the figure below:



The total area under the curve is one, because this area represents the probability of all possible events.

### 2 Some important theoretical distributions

Some probability distributions have a special importance in statistics, either because they have important mathematical properties, or because they represent well the probability of occurrence of some real-life experiments, or both. For example, the *Binomial distribution* represents well the probability of having a given number of “heads” when throwing a coin N times. The *Poisson distribution* is a good approximation of the probability of having a certain number of admissions in a hospital in a given day, when we know the average rate over the year.

Both distributions (which are *discrete*) are described in more detail in the reading by Machin *et al.* (2007):

**Reading**

Machin, D., Campbell, M. J.& Walters, S. J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 64-67

Among the continuous theoretical distributions, a special place is reserved for the *Normal distribution* (or *Gaussian*, after the German mathematician Johann Carl Friedrich Gauss). The normal distribution has a central place in statistics both because the values of many real-life variables are approximately normally distributed (including many physiological variables in large populations, like blood pressure, weight, height, BMI); and because of its mathematical properties.

Among the features of a normal distribution are:

1. it is completely defined by its mean μ and its standard deviation σ;
2. it is unimodal and symmetrical around its mean;
3. its mean, median and mode coincide;
4. the probabilities that a value is included in a range of μ ± σ, μ ± 2σ and μ ± 3σ are fixed and approximately equal 68%, 95% and 99% respectively as shown in the figure below:

|  |
| --- |
| C:\Users\Annibale\Desktop\distribution_normal.jpg |

The next reading by Machin *et al*. (2007) presents in details the characteristics of the Normal distribution and its properties:

**Reading**

Machin, D., Campbell, M. J. & Walters, S. J. (2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 68-73

The reading below by Petrie and Sabin (2005) briefly introduces a few more continuous distributions of widespread use in biostatistics (namely the *Student’s t distribution*, the *F distribution* and the *χ2 distribution*:

**Reading**

Petrie, A. & Sabin, C. (2005). *Medical Statistics at a Glance* (2nd ed.). Malden: Blackwell: 22

Note: the theory underlying probability and probability distributions is complex and understanding the details would require more than basic algebra in your “mathematical toolbox”. But this is a course of applied statistics and you are not required to master the maths behind probability distributions, but only to understand the basic ideas, the terminology, and know how to apply these concepts to your practical problems.

The most common practical thing you have to do with probability distribution is calculating the probability of a specific value of a discrete distribution, of for an interval of values for a continuous distribution. This can be done using statistical tables, which are widely available for all common theoretical distributions: they report the probabilities of values or intervals as a function of some parameters.

## Task 1- Calculating proportions

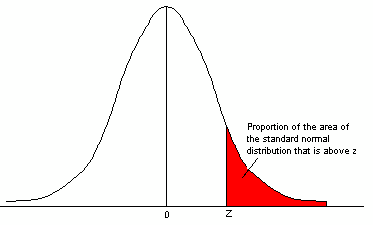
The time required to complete a college achievement test was found to be normally distributed, with a mean of 110 minutes and a standard deviation of 20 minutes.

Following the procedure in the illustrative example on p. 72 in the reading by Machin *et al*. above (birthweight), and using the Standard Normal distribution table below, answer these questions:

a) What proportion of the students will finish in 2 hours?

b) What proportion of students will finish in one hour?

c) When should the test be terminated to allow just enough time for 90% of the students to complete the test?



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **z** | p>z |  | **z** | p>z |  | **z** | p>z |  | **z** | p>z |
|  |  |  |  |  |  |  |  |  |  |  |
| **0.0** | 0.5000 |  | **1.0** | 0.1587 |  | **2.0** | 0.02275 |  | **3.0** | 0.00135 |
| **0.1** | 0.4602 |  | **1.1** | 0.1357 |  | **2.1** | 0.01786 |  | **3.1** | 0.00097 |
| **0.2** | 0.4207 |  | **1.2** | 0.1151 |  | **2.2** | 0.01390 |  | **3.2** | 0.00069 |
| **0.3** | 0.3821 |  | **1.3** | 0.0968 |  | **2.3** | 0.01072 |  | **3.3** | 0.00048 |
| **0.4** | 0.3446 |  | **1.4** | 0.0808 |  | **2.4** | 0.00820 |  | **3.4** | 0.00034 |
| **0.5** | 0.3085 |  | **1.5** | 0.0668 |  | **2.5** | 0.00621 |  | **3.5** | 0.00023 |
| **0.6** | 0.2743 |  | **1.6** | 0.0548 |  | **2.6** | 0.00466 |  |  |  |
| **0.7** | 0.2420 |  | **1.7** | 0.0446 |  | **2.7** | 0.00347 |  |  |  |
| **0.8** | 0.2119 |  | **1.8** | 0.0359 |  | **2.8** | 0.00256 |  |  |  |
| **0.9** | 0.1841 |  | **1.9** | 0.0287 |  | **2.9** | 0.00187 |  |  |  |

## Task feedback

a)

From the table, this value corresponds to a probability of 0.3085 of **not** having finished the test. Therefore the proportion of students who finished the test is 1-0.3086 69%

b)

We are now in the left part of the distribution (z<0) and the values are not reported in the table. This is very common, because the normal distribution is symmetrical and, therefore the probability values corresponding to –z is the same of the probability value associate with z. In this case, we can conclude that the proportion of students who finished in one hour (i.e the probability associated with a z-score lower than -2.5) is 0.0062=0.6%. In words, this means that less than 1% of the students finished in one hour or less.

c) In this case we are using the table in the opposite way: from the probabilities to the z-values. We want to know the z-value corresponding to a probability of 90% of having finished the test, i.e. the z-value corresponding to a probability of 1-0.9=10% of not having finished.From the table, we can see that the z-value of 1.3 corresponds to a probability of 0.09680.1=10%

Going back from the z-value to the corresponding time:

### 3 Sampling error, sampling distribution and the central limit theorem

The aim of inferential statistics is to use statistics calculated from a sample of a population to infer the values of the parameter of the population. For example, we may be interested in the mean age (μ) of the individuals in the population P. With this objective we extract a random sample (s) of size n, and we calculate the mean age of the individuals in the sample, which we indicate with ȳ. We then use the value of the statistics ȳ as an estimate of the “true” population parameter μ.

Unfortunately, there is no guarantee that the value of ȳ is equal to the value of μ, and in general the difference between ȳ and μ is different from 0. We call this difference between the value derived from the sample (statistic) and the true population value (parameter) *sampling error*:

Sampling Error = ȳ – μ

Sampling error is due to chance associated with the process of sampling, and it reflects the variability in the population. It does not mean that a mistake has been made in the process of sampling. The only way to eliminate sampling error is to enumerate the entire population (unless all the individuals in the population have the same age, i.e. there is no variation in the values of the characteristics of interest).

If we were to take repeated samples of the same size n from P and calculate the statistics ȳ, it is unlikely that the values of ȳ would be exactly the same each time. What we would obtain is a set of different values (a sampling distribution), some of them are more frequent and others less frequent: we indicate these values with ȳi and we can represent graphically their distribution with a histogram, as in the figure below:

|  |
| --- |
|  |

In other words, the statistics calculated from random samples of a population in general are neither equal to the true population parameter, nor constant across every sample. It is, instead, a random variable, i.e. a set of different values in which each of them has an associated probability.

Random variables are generated in epidemiological studies every time we calculate some kind of statistics (e.g. the sample mean for the subjects’ age) from a random sample of the population: because the subjects are randomly selected from the population, we cannot expect that every time we draw a sample, the mean value of their age would be the same (unless we are in the unlikely situation in which all subjects have the same age). Fortunately, when our aim is to estimate the mean value of a variable, the distribution of our statistics (which is this case is called the *sampling distribution of the mean*) have some interesting properties:

1. If the sample size is reasonably large, the estimates of the mean follows a Normal distribution, whatever the distribution of the original data in the population;
2. lf the sample size is small, the estimates of the mean follow also a Normal distribution, but only if the variable in the original population follows a Normal distribution;
3. The mean of the estimates across a large number of samples equals the true population mean (in the example above, this means that the mean of all ȳi is equal μ);
4. The standard deviation of the estimates across a large number of samples (denoted as standard error of the mean) equals the standard error of the original population divided by the square root of the sample size.

These properties of the sampling distribution of the mean are known as the *Central Limit Theorem*.

*Standard deviation and standard error*

The standard error of the mean (often indicated with SE or SEM) is a measure of the variability of the sampling distribution of the mean, and depends on both the variability of the population (measured by its standard deviation, σ) and the sample size n. SE is a measure of the precision of our estimate (and can be increased as much desired, increasing the sample size), while σ is a characteristics of the population, and nothing can be done to modify it. The Central Limit Theorem ensures that:

Because the “true” standard deviation of the population is generally unknown, we normally estimate σ with the standard deviation of the sample(s), and therefore the formula above becomes:

*Sampling distribution of the proportion*

When our parameter of interest is not a population mean, but a population proportion (π) which we want to estimate with the corresponding sample proportion (p), the resulting sampling distribution of the proportion has similar properties to the sampling distribution of the mean. Specifically, for samples reasonably large, it follows a normal distribution with mean p= π and standard deviation SE (p), where:

*Interpreting standard errors*

A large standard error indicates that the estimate is imprecise, while a small standard error indicates that the estimate is precise. The standard error is reduced, i.e. we obtain a more precise estimate if: (1) the size of the sample is increased and/or (2) the data are less variable (i.e. the standard deviation of the population is small).

The concepts above are revised and accompanied by examples in the reading by Machin *et al.* (2007):

**Reading**

Machin, D., Campbell, M. J. & Walters, S. J.(2007). *Medical Statistics* (4th ed.). Chichester: John Wiley & Sons: 82-87

Now do Task 2 below, in which you are asked to practise calculating the mean and standard deviations in a given example:

## Task 2 – Work out mean and standard deviation

### The distribution of blood pressure in the adult population of country X is approximately normal with mean 124 mmHg and standard deviation of 44 mmHg.

### a) What are the mean and standard deviation of the sampling distribution of the mean blood pressure for n = 10?

### b) What are the mean and standard deviation of the sampling distribution of the mean blood pressure for n = 200?

c) What term is used to refer to the standard deviation of the sampling distribution?

## Task feedback

## a) mean = 124 mmHg; standard deviation = 13.9 mmHg

b) mean = 124 mmHg; standard deviation = 3.11 mmHg

c) Standard error of the mean (SEM)

## Task 3 – Calculating mean and standard error

In a city, 70% of the people have brown hair. Suppose 30 people from this city were sampled.

a) What is the mean p of the sampling distribution of the proportion of people with brown hair?

b) What is the standard deviation of p?

c) What is it called?

## Task feedback

a) p=0.70=70% (same as the population)

b) SE=0.038=3.8%

c) Standard error

# Unit 2 – Session 5: Confidence Intervals for descriptive measures

## Introduction

As explained in the previous session, the estimates of population parameters calculated using a sample are subject to random error, i.e. we are not *sure* that our estimates coincide with the *true* population parameters. Building on the concepts introduced in the last session, this session explains how we can complement our *uncertain* estimates with a *confidence interval*, i.e. a measure of the *trust* we put in these estimate (or, more technically, a measure of *precision* of the estimates). Actual formulæ for confidence intervals are presented for *population means and proportions*, when the size of the sample is relatively large.

## Timing

This is a short session, which includes one short reading and two tasks. You should be able to complete it in less than three hours.

## Learning outcomes

* Understand the concept of confidence interval
* Calculate confidence intervals for means and proportions using the t- and z-approximation

## Contents

1. Concept and uses of confidence intervals
2. Calculation of confidence intervals for means and proportions

### 1 Concept and uses of confidence intervals

Once we have taken a sample from our population, we can calculate statistics which represent the estimate of the parameter of interest in the population. For example, if we want to calculate the mean age in the population (μ), we can use the sample mean age (ȳ) as an estimate of μ.

The problem is that we are not sure that this estimate is equal to the true mean μ, and also if we were to draw another sample we would obtain probably different values. It would be useful to attach to this estimate something that indicates the degree of uncertainty of our estimate.

Ideally we would like to calculate an *interval* around our estimate (*point* estimate) which includes the true population parameter with a certain probability (usually 95%). In reality, usual methods do not allow calculating such intervals, but they are able to calculate a similar measure of uncertainty, the so called *(95%) confidence interval.*

Confidence intervals are measures of incertitude (or precision) of our estimates, and the wider they are, the less precise the estimates are. A formal definition of confidence intervals is that *if we were to draw repeated samples for a population and calculate the confidence intervals with the formulae shown in the next section, in the long run 95% of those intervals would include the true population parameter*.

### 2 Calculation of confidence intervals for means and proportions

We can obtain the confidence interval for a population mean using our knowledge about the sample distribution of the mean. In fact, we know from the Central Limit Theorem that the sample distribution of the mean is approximately normal with mean μ and standard deviation equal to the standard error SE. But we also know that 95% of values of a normal distribution are included in the interval of 2 times (more precisely 1.96 times) its standard deviation around the mean.

Therefore, we can say that if we have to repeat the sampling many times, the probability that the interval μ ± 1.96 SE would include the true population mean μ is 95%.

The following intervals called the 95% *confidence interval for the population mean*:

[μ ± 1.96 SE; μ ± 1.96 SE]

As explained above, the standard error equals the standard deviation of the original population divided by the square root of the sample size, and it is approximated by the standard deviation of the sample divided by the square root of the sample size.

The formula above is only valid when n is large (a rule of thumb is that n must be >30). When n is small, it can be shown that the sample mean follows a Student’s T-distribution with n-1 degrees of freedom instead of a normal distribution.

So, a more general formula for the confidence interval of a population mean, where tn-1 is called the *critical value* of a Student’s t distribution with n-1 *degrees of freedom* (and can be read from a t-table)is:

[μ ± 1.96 SE ; μ ± 1.96 SE] for large samples

and

[μ ± tn-1 SE ; μ ± tn-1 SE] for small samples

*Confidence interval for the proportion*

A confidence interval for the population mean can be obtained using similar reasoning.

In this case, it can be shown that the sampling distribution of a proportion follows a Binomial distribution. However, if the sample size n is reasonably large the distribution is approximately normal, and we can obtain a 95% *confidence interval for the population proportion* with the same formula as above, in which the sample mean is substituted by the sample proportion p, and the standard error of the mean is substituted by the standard error for a proportion, which formula is reported inSession 4:

[p ± 1.96 SE(p) ; p ± 1.96 SE(p)]

In this case the concept of “reasonably large” sample depends on the value of p. A rule of thumb is that the formula above gives a good approximation of the 95% confidence interval when both np and n(1-p) are greater than 5. With smaller samples, other more complex formulae must be used.

The concepts above are revised and accompanied by examples in the reading by Petrie and Sabin (2005):

**Reading**

Petrie, A. & Sabin, C. (2005).*Medical Statistics at a Glance* (2nd ed.). Malden: Blackwell: 28-29.

## Task 1 - Answer questions about confidence intervals

Answer the following questions in your own words:

a) Using a sample size n=10, a researcher calculated the 95% confidence interval for the mean weight of pupils in a school, and the results were 12 to 28kg. A new sample of 36 observations is going to be taken. You can't know in advance exactly what the confidence interval will be because it depends on the random sample. Even so, you should have some idea of what it will be. What do you think will happen?

b) Why is a 99% confidence interval wider than a 95% confidence interval?

c) When you construct a 95% confidence interval, what are you 95% confident about?

## Task feedback

a) The new confidence interval will be narrower than the previous one. Increasing the sample size decreases the standard error and this reduces the width of the confidence interval.

b) When calculating a 99% confidence interval, we want to decrease the risk that the “true” population mean does not belong to the interval. Therefore we need to consider a wider range of values than when calculating 95% intervals.

c) We are confident that, if we have to repeat the sampling procedure many times and calculate the confidence interval each time, 95% of those intervals will include the “true” value of the population parameters.

## Task 2 – Work with confidence intervals

A population is known to be normally distributed with a standard deviation of 4.8.

a) Compute the 95% confidence interval on the mean based on the following sample of nine observations: 8, 9, 10, 13, 14, 16, 17, 20, 21.

b) Compute the 95% confidence interval using the same sample, in a case where the standard deviation of the population is unknown (which is almost always the case).

c) Explain the differences between case (a) and case (b).

## Task feedback

a) [11.08 ; 17.36]

b) [11.05 ; 17.39]

c) The confidence interval in case (b) is slightly wider than in case (a). In case (b), when the standard deviation of the population is unknown and the sample size is small, the normal approximation (which allowed us to calculate the CI in case a) is not valid anymore. Therefore, in case (b), we used the standard deviation of the sample and the t-distribution for the calculation. The use of the t-distribution lead to wider confidence intervals in small samples (the difference becomes negligible with sample sizes greater than 30).