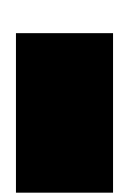
Unit

1

# Measuring Health and Disease Revision



**1**

**1 Introduction**

In this unit of the Quantitative Methods module the following sessions will be covered:

* An approach to quantitative research – mainly revision
* The **Seven Steps Approach** to conducting an epidemiologic study, which is the approach we will be using to help you learn about quantitative research methods. The similarity between this approach and the quantitative research study protocol will be highlighted.
* The concepts covered in Steps 1 – 3 that are covered in the Measuring Health and Disease module you have already completed are summarised – with some additional points – and include:
  + - 1. Define the population of interest;
    - 2. Conceptualise and create measures of exposures and health indicators, and how these can be measured; and
    - 3. Take a sample of the population. A section on Sources of Data is also included.
  1. **An approach to quantitative research**

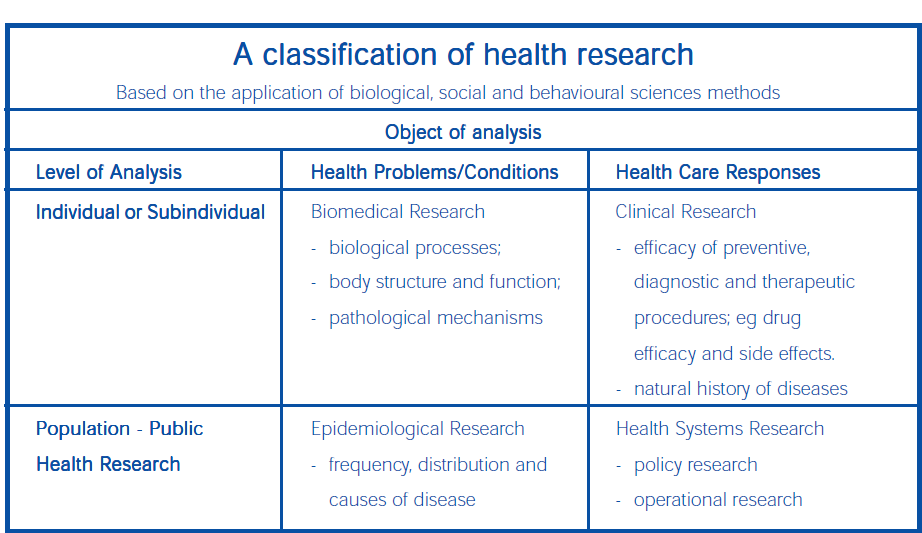
Research is the systematic enquiry into nature and society leading to the development of new knowledge. It involves the systematic collection, processing, analysis and interpretation of data to answer a certain question or solve a problem.

* + - It demands a clear statement of the problem.
    - It requires clear objectives and a plan (it is not aimlessly looking for something in the hopes that you will come across a solution).
    - It builds on existing data, using both positive and negative findings.
* **Basic research** is necessary to generate new knowledge and technologies to deal with major unresolved health problems. It may not necessarily have immediate practical application. Basic biomedical research is often laboratory-based, and focusses on how the body works. The biological processes, structures, functions and mechanisms within an organism. Clinical research focuses on the response of the body to various preventative, diagnostic or therapeutic interventions.
* **Applied research** is necessary to identify priority problems and to design and evaluate policies and programmes that will deliver the greatest health benefits, making optimal use of available resources. Sometimes the problem has already been identified in society whether by the researcher, industry, the public or the government. Epidemiology is an applied science.

Epidemiological principles, which form the basis of most health research can be applied in three different ways:

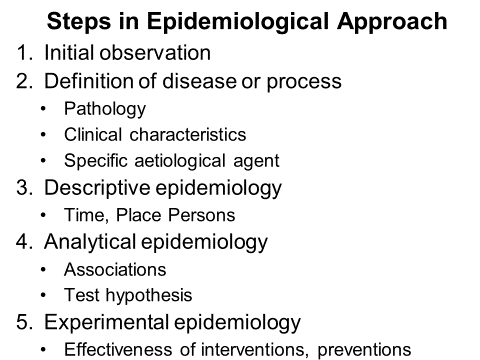
1. Epidemiology can be applied in clinical medicine at an individual patient level – usually called clinical epidemiology - looks at the efficacy of preventive diagnostic and therapeutic procedures and natural history of disease.

Public health research focuses on groups of people (populations). It has two main components:

1. **Epidemiological research**, which considers the frequency, distribution and causes of ill health; and
2. Health systems research, which focuses on the organised response to health and disease.

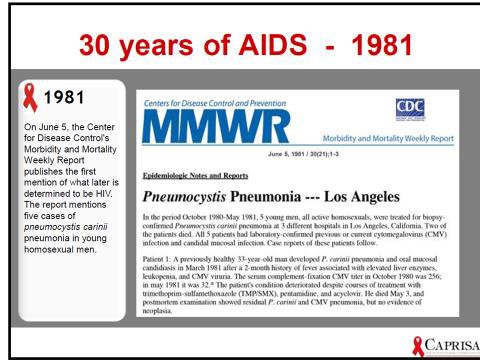
|  |
| --- |
| ***Task 1 - Think about these questions***   * *What are some of the questions that need to be answered in order to understand (have new knowledge) about HIV prevention and male medical circumcision?*   + *What are the different things that you need to know about MMC & HIV?*   + *How would you find things out about MMC & HIV?*   + *What are the different types of research needed to investigate MMC & HIV?* |

So what are the steps involved in the epidemiological or applied research approach. Think of HIV infection and AIDS disease. How was new knowledge of HIV and AIDS discovered?



The first step used an observational descriptive case series study design and was based on 5 cases of pneumocystis pneumonia and published in 1981.

.



*Task 2:*

*What are some of the other key steps in understanding the epidemiology of HIV and AIDS?*

## Protocol Template

The QRM module is an applied module, with the goal of you being able to produce a quantitative research study protocol. We will thus try and link all the content covered in this module back to the knowledge and skills you need to complete a protocol which would be at an acceptable level for a Master of public Health research project and be acceptable to the university Research Ethics Committee.

There are a number of manuals and formats that one can use in your approach to research. The International Development Research Centre uses simple format for development of the health research proposal ([www.idrc.ca/openbooks/069\_1.](http://www.idrc.ca/openbooks/069_1) This is an excellent manual. You really don’t need a textbook if you use this manual which can be freely downloaded from the Internet.

Although the IDRC is a manual aimed at developing a health systems research study – it would apply to any quantitative research study protocol.

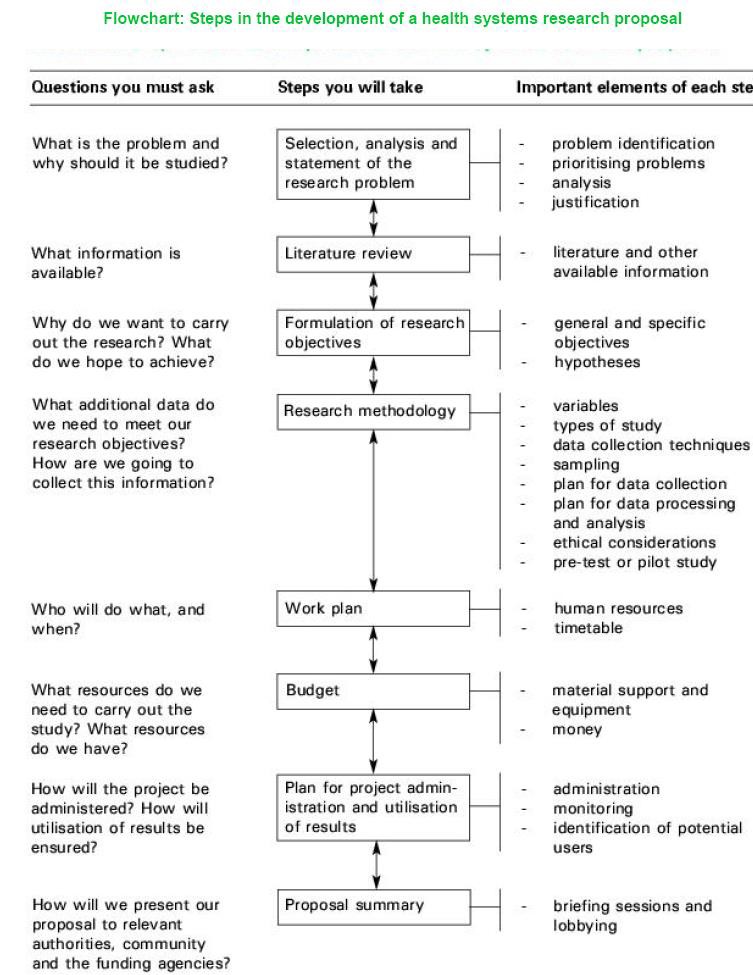
Attached is another protocol template for you to consider using (Quantitative Research Study Protocol template 2018).

For any protocol a staged approach is ideal. These are the IDRC 8 stages to include in a protocol:

1. Selection, analysis and statement of the research problem.

The question that needs to be asked is: ‘What is the problem and why should it be studied?’

1. Literature review, and should answer the question – ‘What information is available?’
2. Formulation of the research objectives and should answer the question - ‘Why do we carry out this research? & What do we hope to achieve from it?’
3. Describe the research methods in detail. The Methods section is like a recipe – it should provide sufficient information for somebody else to conduct the research, process the data and write up the findings without your additional inputs. Most of the rest of this module covers the methods. ‘What additional data do we need to meet out research objectives? How are we going to collect this data?
4. Work plan: ‘Who will do what and when?’
5. Budget: What resources are needed to carry out the study? What resources do we have?’
6. Plan for project administration and utilisation of results. How will the project be administered? How will the utilisation of the results be ensured?’
7. Proposal summary: How will we present out proposal to relevant authorities, community and the funding agencies?



BUT now, back to some basic principles and revision.

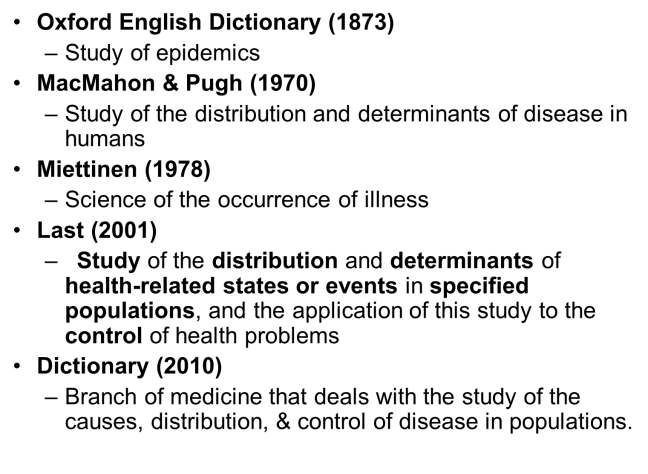
## Evolution of epidemiology (revision):

Epidemiology is the science of understanding the causes and distribution of population health so that we may intervene to prevent disease and promote health.



**Table 1: Some definitions of epidemiology**

This definition has changed very little in the last 150 years (Table 1)



The two core functions of epidemiology are:

1. To identify causes of population health; in order
2. To intervene (Galea, 2013)

It combines a **scientific enquiry** PLUS a **call to action.**

Identifying the causes of population health requires an understanding of the nature of population; the conditions that shape them over time and place and the policies, politics and practices that create these conditions.

BUT, always in a way to lead to healthier populations. This is what Galea (2014) refers to as an “epidemiology of consequence”.

*Task 3:*

*What are the 4 principles of bio-ethics?*

*Which of these principles is addressed in the protocol to cover the ‘intervention’?*

South Africa is incredibly dynamic both epidemiologically and demographically. Some of these health status indicators are shown in Table 2.

## Table 2: Some health status indicators for South Africa.

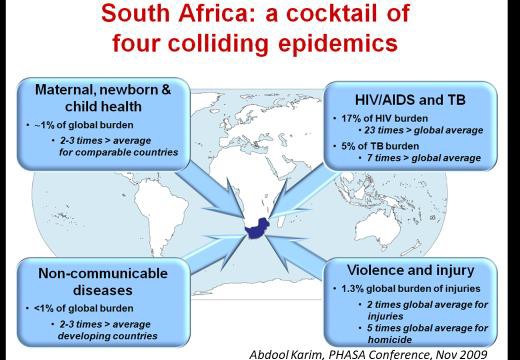
***Source: District Health Barometer- Health System Trust*** *Task 4:*

*How do you define the indicator ‘Life Expectancy’? How is it calculated?*

*Why has life expectancy changed so rapidly in SA? How is IMR defined? Why is it a useful indicator?*

*How is HIV prevalence and HIV incidence determined? Why is HIV prevalence increasing?*

South Africa also has a ‘quadruple burden of disease’ (Figure 1).



## Figure 1: Quadruple Burden of Disease in South Africa

*Task 5:*

*How would you describe the burden of disease in: High Income Countries; Middle Income countries; and Low Income Countries.?*

*When does one categorise a country as ‘Low Income’ and when as ‘Under-developed’?*

In addition to the dynamic epidemiological and demographic transition other global dynamics – digital technology, urbanisation, rapid transport and other pressing health challenges, - there is a dire need for well-trained epidemiologist and public health professionals.

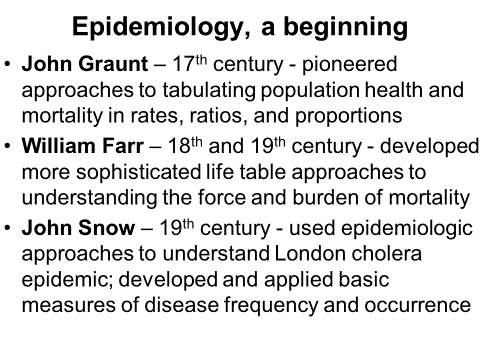
*Task 6:*

*What is demographic transition? What is epidemiological transition?*

*How are these two ‘transitions’ related?*

For more detail of the history of public health, see SoPH module, *Measuring Health and Disease*, Unit 1.

## Table 3: Epidemiology - a beginning

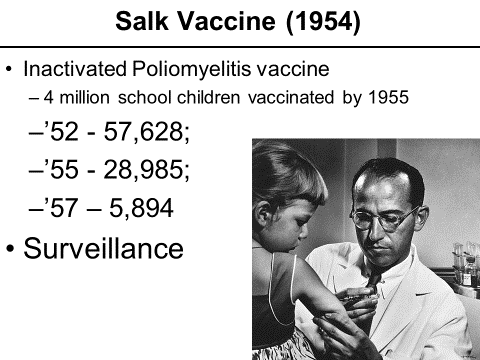


The practice of conducting epidemiologic studies is not new; ‘counting’ health and disease goes back centuries.

*Task 7:*

*What was the contribution of Hippocrates and Lind to epidemiology?*

Epidemiology though, is relatively new as a formal scientific discipline



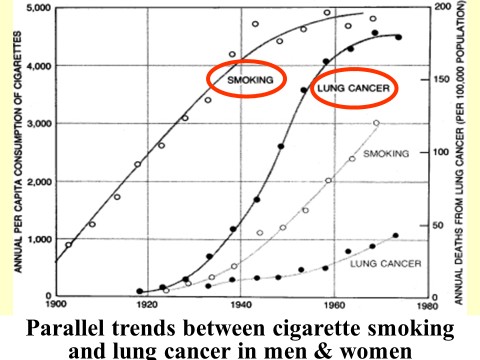
*Task 8*

*What is surveillance?*

*How was surveillance used to measure the effectiveness of the inactivated poliomyelitis vaccine?*

*If this polio vaccine were to be introduced in 2018, how would it’s efficacy be determined? Why is data ‘dumb’? How can data (from whichever source) be transformed to ‘useful information for management’?*

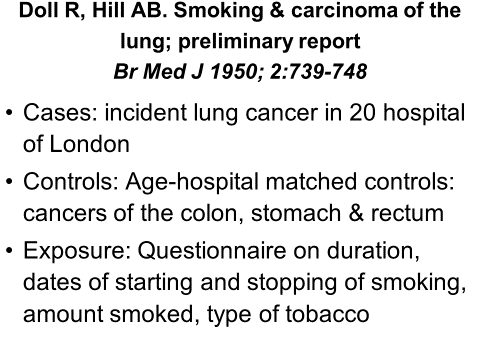
## Table 4: Parallel trends between smoking and cancer



*Task 9:*

*Which study design was used to acquire this data?*

*Why was this data not sufficient to prove that tobacco smoking causes lung cancer? What was the next step in research that needed to be undertaken in answering this question?*



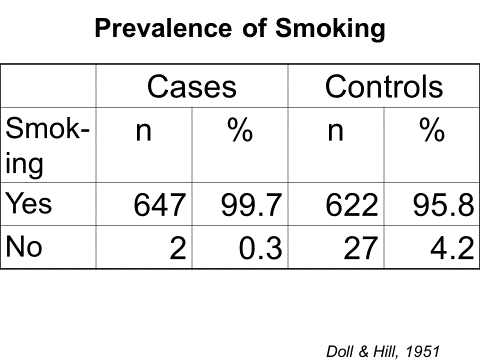
*Task 10:*

*What study design is described in Doll & Hill’s 1950 paper? What measure of association did they calculate? Why did the authorities not ban smoking as a result of this study?*

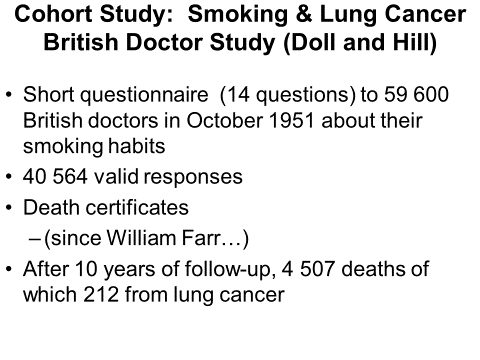
*Why did Doll and Hill implement the ‘Doctors Study’?*

*When was this study published? Why did it take some long to complete the study? What was the outcome measure in this study?*

*What measure of association could you calculate from the data?*



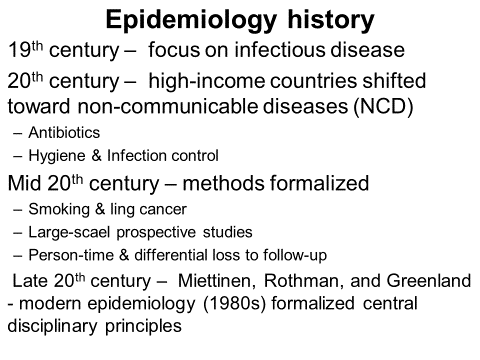
*How would you interpret this finding?*



*Task 10:*

*Which Bradford-Hill criteria were shown by the results of the British Doctors Study?*

Many design & analytic techniques used today arose in response to health concerns during 19th and 20th century



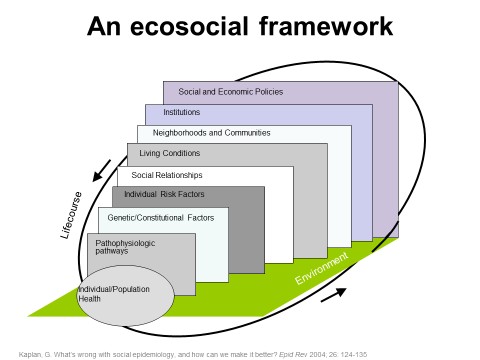
The current conceptual movements in epidemiology include:

## Eco-social perspective on population health

* + Suggests policies, institutions, and characteristics of context contribute to the shaping of health

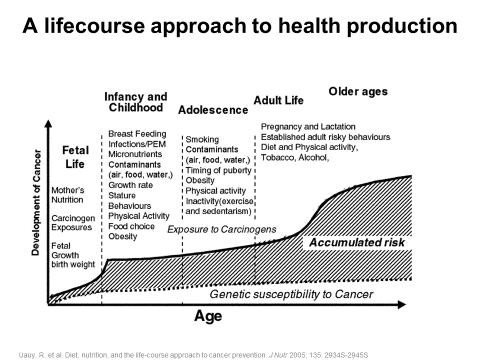
## Life course perspective

* + Determinants of health are distributed across the life course and even before conception



There is a growing consensus that health is not produced cross-sectionally, but throughout life. Pre, peri, and post-natal influences all may contribute to adult health.

Early life exposures may matter from a critical period perspective or cumulatively Epidemiology understands **causes** of population health across **levels of influence** - from cells to society - and **across life course**.



*Task 11:*

*What other conceptual models have been used in the evolution of epidemiology?*

Epidemiology understands **causes** of population health across **levels of influence** - from cells to society - and **across the life course**.

Epidemiology is not always easy. Some of the reasons why it is difficult to measure health and disease include:

* We depend on disease incidence, which may not occur that frequently in person time experience
* Epidemiological research takes time and effort and often requires the cooperation of a number of people to make observations
* In observational epidemiology, one has no control over experimental settings
* Money for research is limited
* The ethics involved with measuring health and disease who in human subjects are stringently control in order to preserve the privacy of subjects
* Identifying ‘cause’ is difficult when considering its complexity – e.g. Life-course and Eco-social models.

## Contents

Session 2: What is a population? What is population health?

Session 3: What is an exposure? What is a disease? How do we measure them (measures of disease occurrence and frequency)?

Session 4: What is a sample? What are other sources of data?

## Timing

This unit has several readings and tasks and it should therefore take you 8 hours to complete.

## Learning outcomes

Appreciate the value of descriptive epidemiology in preparing a quantitative research study protocol in public health

## Readings

This module guide, as with all the module guides provided by the UWC School of Public Health, contains explanatory text, which is bolstered by an array of readings. The explanatory text usually introduces the topic/s, provides relevant background information, attempts to present a contextualised overview of the topic/s and links it/them to what you have already covered in the modules as well as what you will be covering later on in the module. The readings then provide the details or the meat of the topic/s and hence it is crucial for you to work through them carefully in order to fully comprehend the topic/s they cover. The readings that will be used in this unit are listed below. They are also displayed at the points in the unit where it is indicated that it is appropriate to proceed to work through them.

## Tasks

With every topic covered there are several tasks which are designed to assess whether you have assimilated the knowledge required to successfully complete the task and whether you can apply that knowledge to solve the problems posed in the task. It is therefore crucial that you attempt every task. Once you have attempted it you should then look at the feedback provided to check to what extent you have been able to successfully apply the knowledge

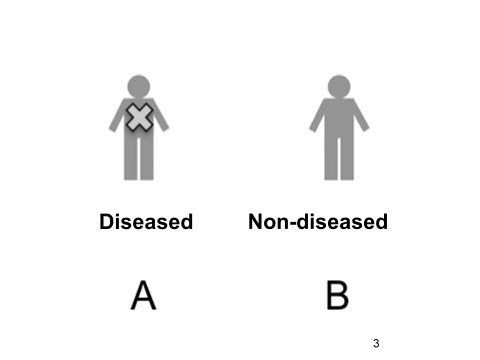
learnt to solve the problem. If there are parts of the problem that you did not successfully attend to, then you will be able to see where you were deficient and you should go back over those sections of the module text and readers that covers those areas, in order to solidify your understanding. Feedback on the tasks is provided at the end of each study session.

To obtain maximum benefit from the module guide and readers you should obviously work through all of the tasks before looking at the feedback. The tasks will also be extremely helpful to you when answering questions set in the assignments, as the questions set in the assignments will have the same structure and format as the tasks.

# Unit 1 – Session 2: What is a population? What is population health?

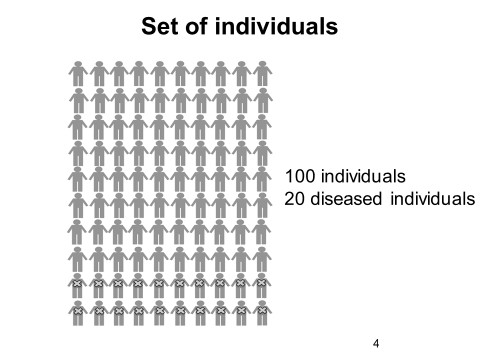
In the QRM module, we will use the graphical notation used by **Epidemiology Matters: a new introduction to methodological foundations** (*www.epidemiologymatters.org)*. 5

This notation is used to conceptualise, define and operationalise what we mean by individual, population and population health. Health indicators in Public Health ultimately occur in individuals.



## Figure 2: Individuals

For clinicians the individual client/patient is key. Measuring health and disease in a population starts with measuring health and disease in the individual (Figure 2).



## Figure 3: Set of individuals - population

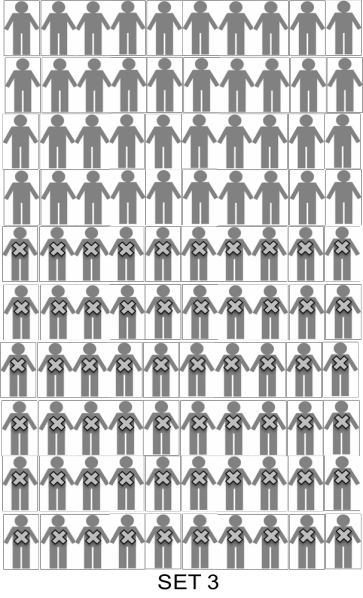
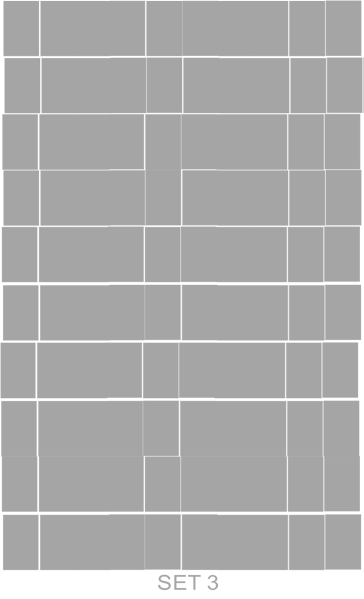
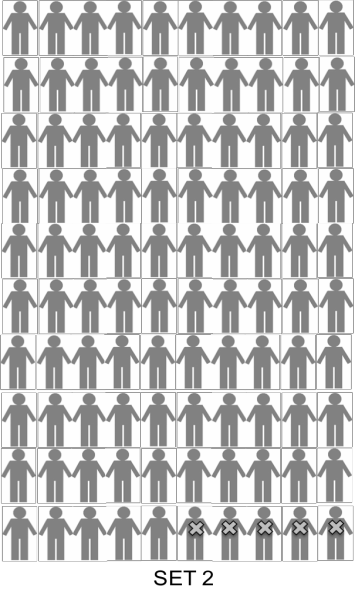
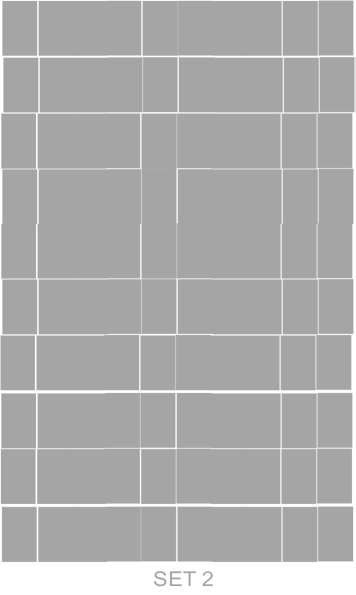
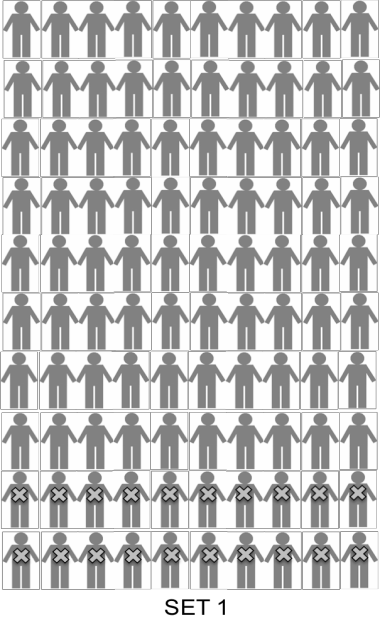
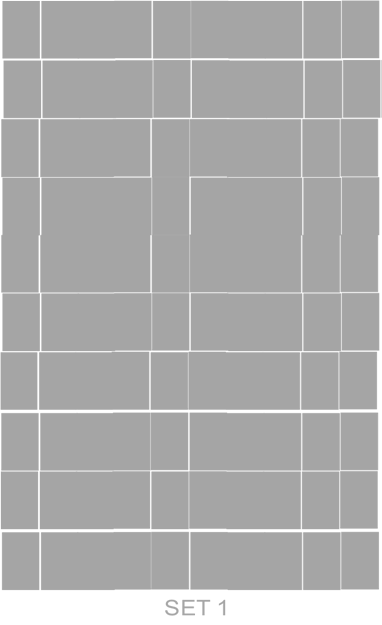
A population consists of a set of individuals (Figure 3).

5 Keyes KM, Galea S. Epidemiology Matters: A New Introduction to Methodological Foundations. New York, NY: Oxford University Press; 2014.

*Task 1:*

*In figure 3, what proportion of individuals is affected? Is this disease a threat to health? Are there factors common to the 20 diseased cf. non-diseased?*

*If so, can we intervene to prevent disease in future?*



**Figure 4: Multiple sets of individuals - populations**

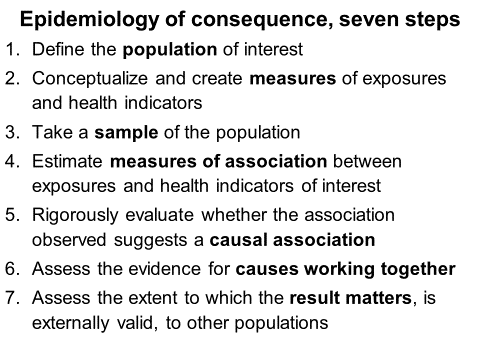
*Task 2:*

*What is different in these populations? Why is it different?*

Descriptive epidemiology asks:

‘What is the distribution of health and illness across sets of people – populations?’ Analytic epidemiology asks:

‘Why is there a difference or how did the difference occur?’



Defining the population of interest is the first in the **Seven Steps Approach.**

## Defining a population of interest

A population is a collection of individuals, at moments in time, defined by at least one organizing characteristic.

The definition of a population has implications for **analysis**, **interpretation**, and

**generalisability** of results in epidemiologic studies.

There a t**wo axes of population definitions:**

## Population defined by eligibility criteria

* + - * Geographic area and time period of interest
      * Characteristics of persons, events, or exposures for which health-related factors are of interest
      * Factors that promote successful study completion.

*Task 3:*

*Give an example of study populations defined by characteristic, event and exposure*.

## Populations, defined by study considerations

The Study Population of interest may be selected group collected specifically to achieve

## successful study:

* **Good responders** to our surveys
* Likely to **continue to attend** the health facility for follow-up visits
* Individuals who are **healthy** so that they are unlikely to die during the course of the study outside scope of inquiry

## An example

The Women’s Health Initiative is a study that assesses health benefits and risks of hormone replacement therapy after menopause. Eligibility is limited to women who had **undergone menopause**.

The study **excludes:**

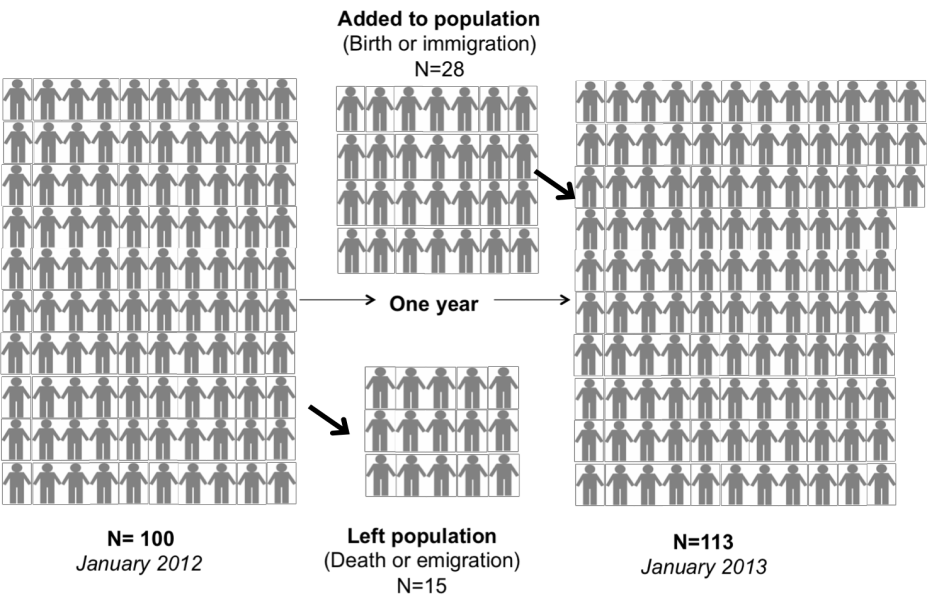
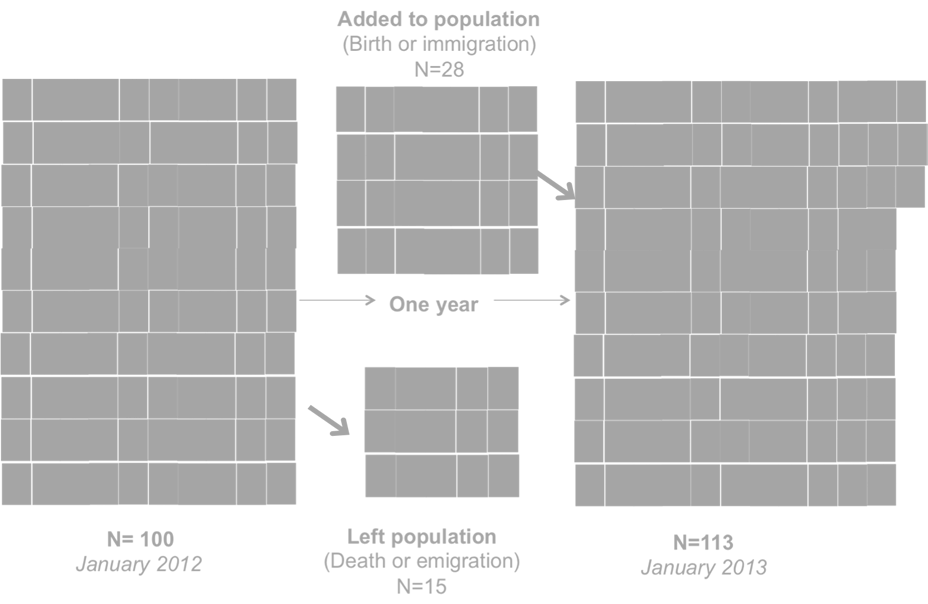
1. Women with a history of certain **medical conditions**
2. Women who were unlikely to be **adherent** to study protocol
3. Women who failed to meet **other criteria**

*Task 4:*

*Which study design/s often selects a study population, which is chosen and is defined by study considerations?*

## Populations defined by whether individuals move in and out of eligibility; i.e., populations can be dynamic or stationary

* + - * Populations can be dynamic
      * Stationary populations



**Figure 5: Dynamic population in Pholeleni, 2012**

The population depicted in Figure 1 is from a defined area (Pholeleni6) and comprised 100 individuals in January 2012. By January 2013, the population was 113. During the year, 28 individuals were added to the population but 15 left. This is a dynamic population.

* + 1. **Dynamic vs stationary populations Dynamic** eligibility criteria for the population
* Allow movement in and/or out of the population

E.g. People may move out of geographic region and no longer be considered part of the population of interest.

**Stationary** eligibility criteria for the population:

* Do not allow movement in and/or out of the population.

E.g. population may include all those born on a particular date only and follow them forward in time, never losing any nor adding others

6 Pholeleni is the name of the population used in the examples used in this module.

*Task 5:*

*Populations may be defined by a variety of criteria, such as geographic location, time period, chronological age, stage of life, exposure to a particular event, or considerations relevant to the completion of a successful study. What eligibility criteria define the following populations?*

* *a. Baby boomers*
* *b. Inner-city teens*
* *c. Women with ruptured breast implants willing to undergo an experimental implant removal technique*
* *d. People who were bartenders in Cape Town prior to the ban on smoking in restaurants and bars*
* *e. People whose homes were flooded during the winter floods on the Cape Flats in 2011*
* *f. People subject to strict gun laws*
* *g. Indian women who work in nail salons*
* *h. Children with asthma whose parents will follow instructions to administer a new medication*
* *i. Survivors of the Marikana Massacre*
* *j. Men who have been prisoners in Pollsmoor Prison*

## 2.2 In summary:

* **Epidemiology** is concerned with understanding the **health of populations**
* Populations of interest can be defined by geography, space, time or by characteristics of participants or of events of interest
* Regardless of eligibility criteria, population may be **dynamic or stationary**
* **Careful definition** of the **population base** from which we conduct an epidemiologic study underlies many of the core methods in epidemiology

*Task 6:*

*Look at the UWC SoPH Proposal Checklist (and the protocol templates provided). You will see that an important box that needs to be ticked is ‘Population described’. This is where you need to have defined the Study Population from which the Study Sample will be selected.*

*The Study Population should have clear and unambiguous Inclusion and Exclusion Criteria. Give theoretical inclusion and exclusion criteria of Age, Gender and socio-economics for a theoretical population.*

*When should the exclusion criteria be the inverse or opposite of the inclusion criteria? What is the difference between the study context and the study population?*

*What is generalisability of a study? What is external validity?*

*What do you call the population to which the results of a study could be generalised?*

# Unit 1 – Session 3: What is an exposure? What is a disease? How do we measure them (measures of disease occurrence and frequency)?

Once the population has been defined the measures that will be used to quantify the population both exposures and outcomes needs to be conceptualized and defined.

*Task 1:*

*Where does this step fit into the Seven Step Approach?*

## What is a variable?

Any measured characteristic of individuals that differs across individuals Exposures

Health indicators Health-related events

That differ across individuals & populations

## 3.1.1 Variable examples

* Age
* Sex
* Place of birth
* Occupation
* Education
* Ethnicity
* Cigarette smoking
* Diet
* Alcohol consumption
* Blood pressure
* Gun ownership
* Diabetes
* Pancreatic cancer
* Depression
* HIV

## Defining variables

* + 1. Binary
    2. Ordinal
    3. Continuous

## Binary health variables

* Variable that takes on two values
  + Health outcomes: present or absent
* Examples
  + Individual has diabetes
  + Individual does not have cancer
  + Individual has Alzheimer’s disease
  + Individual does not have HIV

## Ordinal health variables

Variable that has multiple (>2) graded values Individual health rating

How would you rate your health?

Response options: **Excellent, Good, Fair, or Poor**

Symptom frequency

How often do you experience night sweats?

Response options: **Always, Often, Rarely, or Never**

Ability to perform health-related activity

How difficult is it to climb a flight of stairs?

Response options: **Very difficult, Somewhat difficult, or Not difficult**

* + 1. **Continuous health variables** Variable with continuous response options Examples
* Age
* Weeks of pregnancy
* Diastolic and systolic blood pressure
* Cholesterol level
* Viral load
* Cancer stage

*Task 2:*

*What is a continuous numeric variable and what is a discrete numeric variable? How else could a discrete numeric variable be categorised?*

*Why is it important to categorise each of the variables for which data is collected?*

In analytic epidemiological studies variables can be:

## Exposure variables

Any measurable variable that **affects** or is **associated** with health (outcomes) An exposure variable (may be harmful)

From macro / social environment TO Molecular level

## Policies and laws

Areas with higher taxes on alcohol have lower alcohol consumption rates

## Biological sex

Men die, on average, younger than women

## Outcome variables

* Population health is often defined by ‘absence of health’ or the occurrence of disease
* Health outcomes are typically measures of the occurrence of infections, syndromes, symptoms, & biological or subclinical markers
* Health outcomes can be measured over the life course
  + E.g.
  + Measures disability associated with adverse health states,
  + potential years of life lost due to an illness
* Health outcomes can also be positive

 E.g.,

* + well-being
* **‘Outcomes’** are sometimes called ‘indicators’7

## Types of exposures

1. Innate
2. Acute
3. Chronic or stable
4. Time-varying
   * 1. **Innate exposures** Born with them E.g.
        + Biological sex
        + Race & ethnicity
        + DNA sequence

These could be seen as chronic exposure

## Acute exposures

Occur for a relatively short duration Do not repeat

E.g.

* + - Natural disasters
    - Motor vehicle collisions

## Chronic exposures

Stable over time

May be present at birth e.g.

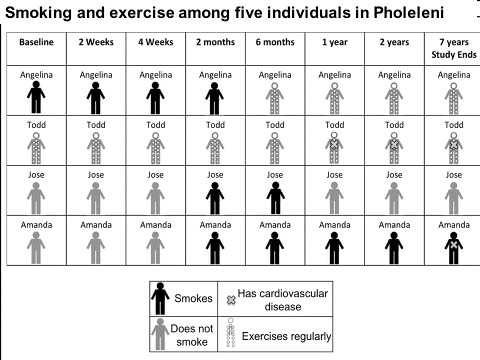
* + - * Pollution
      * Poverty
      * Policies and laws can change with time
      * E.g. family poverty
      * Time varying or dynamic exposures

## Time-varying exposures

Vary across the life course of an individual e.g.

* + - * Diet
      * Exercise
      * Smoking
      * Alcohol consumption

Epidemiologists capture variation over time with different measures of exposure.



Here is an example of 4 individuals with 1 outcome variable (CVS disease) and 2 exposure

/risk factors (Smokes & exercises regularly) The study continued for 7 years.

*Task 3:*

*What study design is used in this example?*

*What are the characteristics of this type of study design? What is the outcome measure when the study ends?*

*How would you quantify the exposure and outcome variables? Which of these is a dependent and which an independent variable?*

## Summary: exposure

There are a wide range of potential variables that individuals are ‘exposed to’ e.g.*:* Age Sex

Education

Water consumption

*Task 4*

*What type of exposures are these?*

*Task 5:*

*Consider the following variables. State whether each is binary, ordinal, or continuous.*

1. *Income level*
2. *Native English speaker*
3. *Blood lead concentration*
4. *Poverty*
5. *Minutes of exercise per day*
6. *Sex*
7. *BMI category*
8. *Number of cigarettes smoked per day*
9. *Percent living below the poverty line*
10. *Legalization of marijuana in state of residence*

*Classify the following exposures as acute, chronic, intrinsic or time-varying.*

1. *Hours of sunlight exposure per day*
2. *Exposure to particulate matter from the collapse of the World Trade Centre*
3. *Dietary supplement use over past 10 years*
4. *Death of a spouse*
5. *Adult height*
6. *Folic acid deprivation in early gestation*
7. *Exposure to electromagnetic fields from living next to high-voltage power lines*
8. *Country of birth*
9. *Premature birth*
10. *Exposure to fluoride in drinking water*

## Characterizing exposures

* + 1. Duration of exposure
    2. Latency and critical windows

## Exposure duration

Duration that individual is exposed matters for production of adverse health for certain exposures

## Exposure duration, examples Smoking

Smoking a cigarette **is unlikely** to have long-term health consequences

Smoking > a pack of cigarettes per day for 40 years **is likely** to have long-term health consequences

## Trans fat

One trans-fat and calorie laden meal **is unlikely** to affect health

Years of unhealthy eating **is likely** to accumulate to adversely impact health

**Exposure timing**

Timing of the exposure across the life course may also be important for the production of health

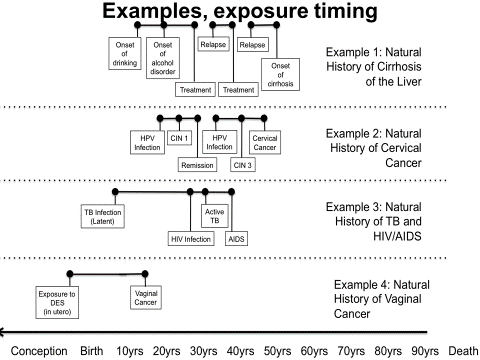
Core concepts: Latency and critical windows

## Exposure timing examples: Latency

Low birth weight associated with the development of chronic diseases in adulthood

## Critical window

Extreme caloric restriction during first trimester of fetal development associated with schizophrenia development in adulthood



*Task 6:*

*What is the difference between infection and disease? Explain Example 4 above*

## Measuring exposure and disease

We have conceptualized the exposures and health indicators of interest Now we are interested in measuring these factors

Good measurement of variables is critical for epidemiologists

## 3.8.1 Measurement example Research question

* Are individuals who have depression more likely to be overweight than individuals without depression?

## Measuring depression

* + Constellation of symptoms
  + Condition characterized by disabling feelings of hopelessness, sadness, and loss of interest in activities

## Measuring overweight

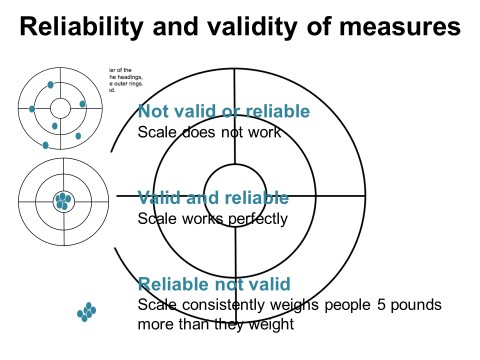
* + Obesity = Body Mass Index (BMI) ≥ 30

## Measurement

* + 1. Be clear about the construct being measured
    2. Assess the reliability of the measures
    3. Assess the validity of the measures

## Measurement example, clarity

1. Be clear about the construct being measured
   * Depression: validated scale
   * Obesity: BMI ≥ 30
2. If measurements include respondent answered questions, make sure questions are easily interpretable, short, clear, and precise.
   * Instead of “Are you depressed?”
   * Try “In the past week have you felt happy most of the time?”

**3.9**

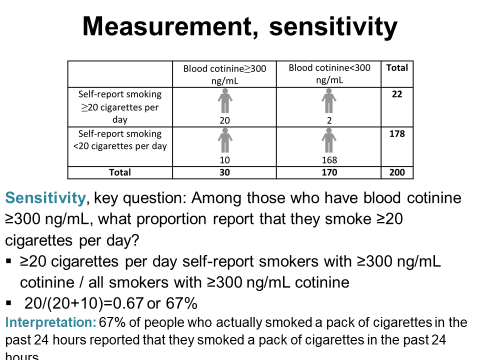
**3.9.1 Dimensions of reliability**

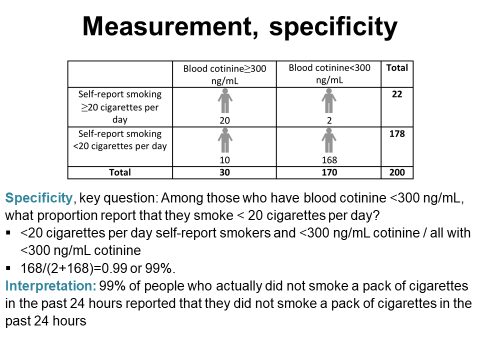
* + - Test-retest reliability:
      * Would the respondent answer the question similarly if asked at ≥ 2 time points?
    - Internal consistency:
      * Are all the items used to assess the construct indicative of that construct?
    - Inter-rater reliability:
      * Would ≥ 2 independent raters all rate the response the same?

## 3.9.2 Measurement validity

Questions to consider when assessing validity What is the gold standard?

What are the sensitivity and specificity of our measure as compared to the gold standard?





## 3.10 Summary: sensitivity and specificity

Provides an assessment of the validity of our measures Sensitivity:

Proportion who are accurately identified as positive on the measure Specificity:

Proportion who are accurately identified as negative on the measure Requires a gold standard

## 3.10.1 Measurement, validity

Questions to consider when assessing validity What is the gold standard?

What are the sensitivity and specificity of our measure as compared to the gold standard?

*Task 7:*

*In the preparation of your research protocol once you have defined the study population you will need to identify the variables you will use to measure the characteristics of the individuals that comprise the population (descriptive study design) or populations (analytic study design) in order to answer your research question.*

*What are the possible sources of data available to quantify populations?*

## 3.11 In summary

Conceptualization and measurement of health in populations is critical to improving population health. Health indicators are presence of disease, symptoms, syndromes, disability, wellness, quality of life, and other health-related states

Exposures are potential influences on these health-related exposures & can be acute or chronic, long or short in duration, have impact only at a critical point in human development or accumulate

# Unit 1 – Session 4: Measures of disease occurrence and frequency

This unit is a **revision of descriptive biostatistics** you covered in MHD. You should be able to complete the tasks easily.

## Seven measure of disease of occurrence and frequency

* + 1. Counts
    2. Prevalence
    3. Incidence/risk
    4. Mean/variance
    5. Median
    6. Mode
    7. Rates

## Tuberculosis in Pholeleni – an example

Tuberculosis disease is a reportable / notifiable medical condition. All diagnosed cases must be reported to the Department of Health.

*Task 1:*

*What is the legislation which governs notification of disease in your country? How is it decided which conditions should be notifiable?*

*Which epidemiological principle is applied in notification?*

In 2015, there were 689 new cases of tuberculosis in Pholeleni.

## Counts

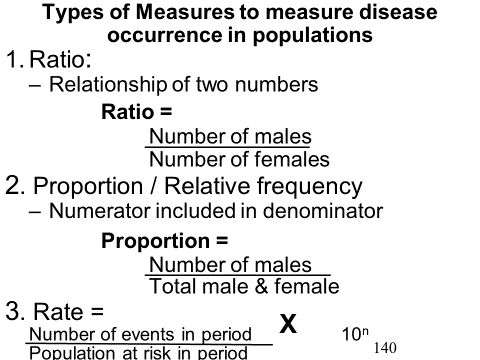
* Provide an absolute number of the burden of disease.
* BUT - counts have limited utility.
* Burden of disease based on the number of people with TB in the population is very different if the population size is 100 000 versus 1 000 000.
* With some infectious diseases some of those affected are not at risk for developing a new onset of the disease.

*Task 2:*

*Are those with TB i. Infection or ii. Disease at risk of developing a new episode of TB? Explain. Why does this happen with some infectious diseases?*

*How is this dealt with in a quantitative research study?*

There are different measures used in epidemiology.



## Remember:

A ratio is the relation of two numbers

A proportion or relative frequency includes the numerator and the denominator

A rate includes a number of events the period in the population at risk in the same period

## To calculate these you need to know

* Size of the total population &
* Size of the total population **at risk**

## Incidence and prevalence

Two measures overcome many of the limitations of a simple count of cases

## Prevalence

* + - Proportion of cases among the total population at any given time

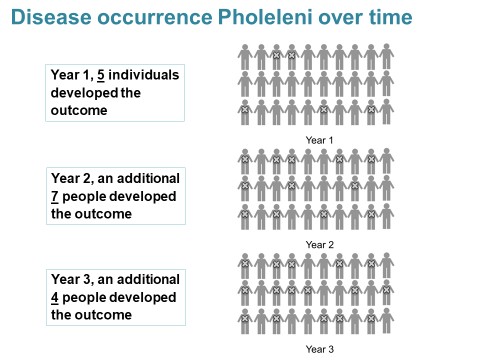
## Incidence

* + 1. Probability of a new onset of disease among those at risk for developing the illnessPrevalence
* Proportion

People who have the disease (existing cases plus new cases)

## Over / divided by

Total population for a given time period

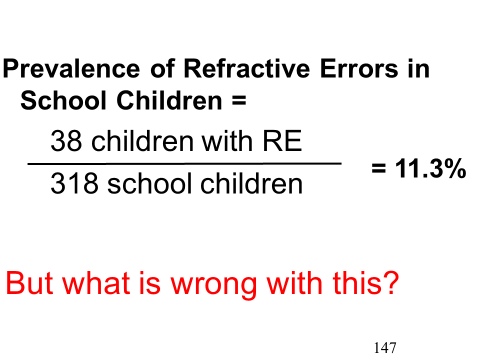


*Task 3:*

*What is the prevalence of disease in Year 2? What is the prevalence of disease in Year 3?*

## Summary: Prevalence

* Need a numerator (number of existing cases), and denominator (total sample size), and a time period of interest
* Time period should be specified as much as possible
* E.g. when we say “in Year 2” we mean over the duration of time that spanned up to Year 2



**Incidence / Risk**

* Most widely used tool in epidemiology
* Alternative name
  + “risk”
  + “incidence proportion”
* **Numerator** = number of **new** cases
* **Denominator** = population **at risk** of becoming a new case
* Specified over a **specific time period**

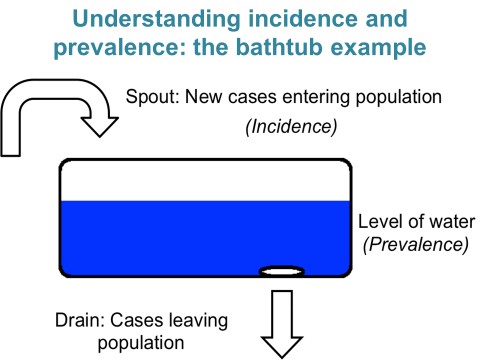
*Task 4:*

*What is incidence of disease in Year 2? What is incidence of disease in Years 2 & 3?*

## Summary: Incidence

* Need a **numerato**r (number of new cases), and **denominator** (total sample size at risk), and a **time period** of interest
* NB: Time period should be specified

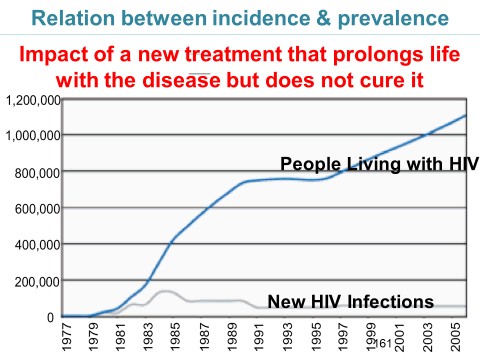
## Relation between incidence and prevalence



* **High incidence, steady prevalence**
  + E.g. contagious infectious disease with very short duration or a high case-fatality

## Low incidence, high prevalence

* + E.g. diseases with long duration such as arthritis, diabetes, and other chronic illnesses



**Summary, incidence, prevalence**

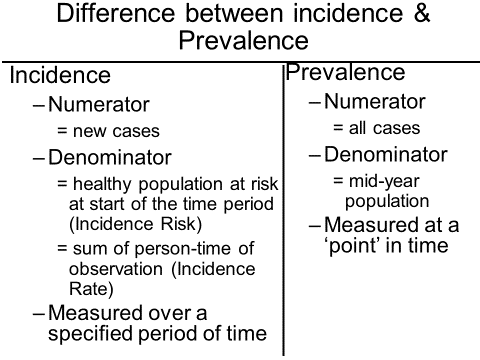
* Prevalence affected by **incidence & duration**
* If a disease has **short duration**
  + Prevalence ~= incidence\*

## \* Assumes that incidence is constant over time

* If a disease has long duration
  + Prevalence > incidence

*Task 5:*

*When can one assume that incidence is constant over time?*

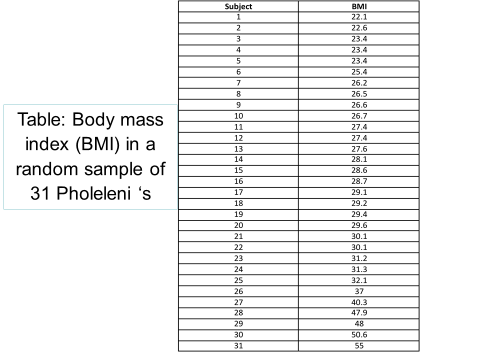


## Mean, variance, median, mode

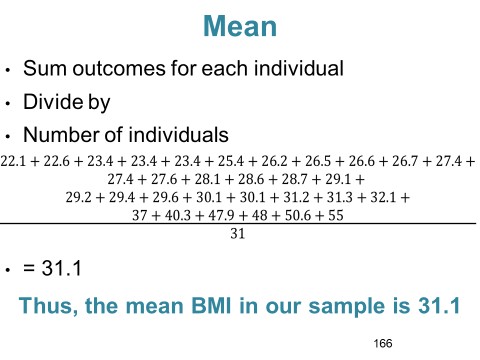
* Health outcomes are sometimes not measured by presence or absence, but rather as a continuous measure
  + E.g. Body Mass Index, blood pressure, cholesterol, birth weight, lung function, number of depression or anxiety symptoms
* In these cases, we need measures of **centrality** and **spread** to characterize occurrence and frequency

## Mean / variance

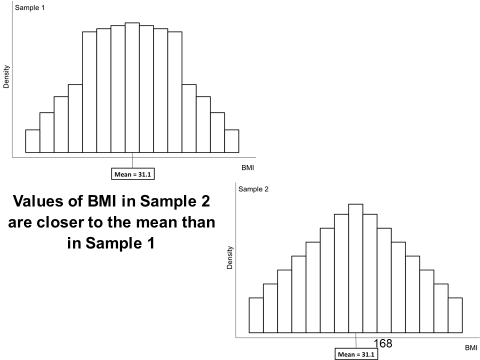
* Estimated by **summing** the outcomes for each individual and dividing that summed score by the number of individuals



* E.g. suppose we measured BMI in a sample of 31 individuals

**Variance**

* In addition to estimating the mean of a continuous variable, it is important to estimate how close all of the individual values are to that mean
* E.g. suppose we sampled two populations, and obtained the following histograms of their risk of disease



*Task 6*

*Which sample has a lower variance?*

## Variance

* Spread of individual values around the mean is a measure of the variance of the data
* Size of the variance gives important information about the distribution of the variable of interest within the sample

*Task 7:*

*How is the variance calculated?*

## Large variance

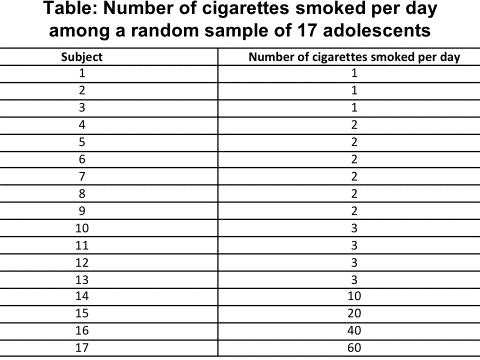
* + Mean may be 31.1, there is a wide range of total values across the whole sample (and, if a representative sample, underlying population)

## Small variance

* + Little variability in the sample (and, if a representative sample, underlying population) with respect to the variable of interest

## Mean and variance: limitations

* Mean can be influenced by **extremes** in the data
* If our data had one recorded miscoded as a BMI of 550 instead of 55, the mean would be 47.1 rather than 31.1
* In general, when the outcomes are not evenly distributed across a full range of potential values and instead are aggregated at the low end or the high end, the mean may not be the most informative measure of centrality
* E.g. Suppose we would like to measure the mean number of cigarettes smoked per day among a sample of adolescents



Mean would be **9.24**

However most of the values are between 1 and 3, thus reporting an average of 9.24 cigarettes smoked in the sample is not very informative

## Median

The median of a variable is the:

Numerical value that falls in the exact middle of the range of values

– Value for which 50% of the remaining values are above and 50% are below

## 5 7

The median value is 5

## 3 5 7 9 9 11

The median value of this variable is 7

## 1 3 4 7 9

There are six observations in this set, so that there is no single value that falls directly in the middle

We take the mean of the two values most centered. Since 3 and 4 are the most centered values (2 observations fall below, and 2 observations fall above), the median of this set is the mean of 3 and 4: (3+4)/2=3.5

Example: Considering our smoking variable, the median value would be 2

* + There are eight observations that fall below 2 in this string of values, and eight that fall above 2 Thus, whereas the mean number of cigarettes smoked was 9.24, the median was 2
    1. This signals that the distribution is quite skewed by a few heavy smokers
    2. Mode

The mode is a measure of centrality

It is the most frequently observed value E.g.

* + - * 3 students reported smoking 1 cigarette per day
      * 6 students reported 2 cigarettes per day
      * 4 reported 3 cigarettes per day
      * 1 student reported 10 per day
      * 1 student reported 20 per day
      * 1 student reported 40 per day
      * 1 reported 60 per day

Modal value is the value that is most frequent

* + - * 6 students reported 2 cigarettes per day
      * Modal value would be 2

## 4.1.7. Incidence rates

* “Incidence” or “risk” is the number of new cases over the population at risk of becoming a new case
* Incidence is an accurate representation of a sample experience of health and disease when we have **complete follow-up of a sample**
  + Each individual is observed at every measurement time point from the beginning of the study to the end

## Alcohol consumption & liver cirrhosis

A study to estimate the association between heavy alcohol consumption & liver cirrhosis

* + - Follow 20 people over time
    - 10 are heavy alcohol consumers
* Assumption
  + Complete follow-up data on all people in the study



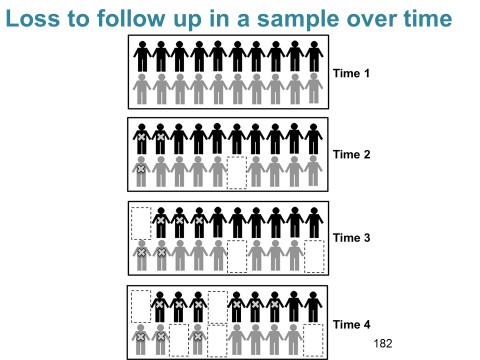
Let us imagine that this is the sample followed forward in time with complete follow up. Let us imagine that people in black are exposed and grey are unexposed – ignore this for now, we will return to black and grey when we learn about measures of association.

*Task 8:*

*What is the incidence over 4 time points?*

Now, let us imagine that we lost some people over time

Thus, we do not know whether these individuals became diseased or not



## Incidence when there is loss to follow-up

* True incidence is **65%**
* If we only analyzed the data based on who was present at the end of the study, we would estimate incidence as 9/15 = 0.60 or 60%
* If we assumed that individuals who dropped out did not become diseased we would get 9/20 = 0.45 or 45%
* If we assumed that individuals who dropped out did become diseased we would get 14/20 = 0.70 or 70%
* There is one more option: **a rate**

## Incidence rates

Used in prospective studies in which some people are lost over time To estimate a rate over the time frame of the study

* + Need to know how much total time each person contributed to the study follow-up before they either developed the outcome or dropped out

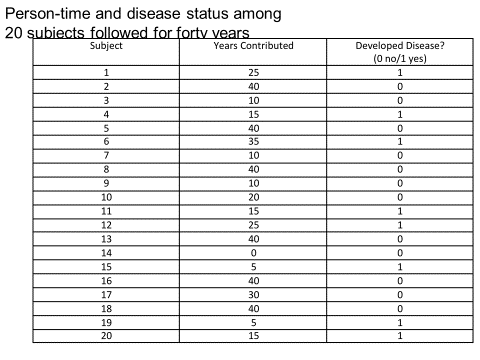
Total time that each person contributed = person**-time Calculating the incidence rate**

The **numerator** is the number of cases The **denominator** is the total person-time

## An example

*Task 9*

*What is the incidence rate of developing the outcome* ***X?***



## Calculating the incidence rate

Interpreted as the number of expected cases in every set of 1,000 person years If we were to observe 1,000 people for 1 year => expect 18 cases

If we were to observe 500 people for 2 years => expect 18 cases Assumption

* + Incidence rate is constant over time, so for every year in which 1,000 person years are observed an additional 18 cases will be expected

Incidence rate = average number of cases per a specified set of person time

## Rate versus proportion: what’s the difference?

* A proportion can range from 0 to 100, and the numerator is contained in the denominator
* A rate can range from 0 to infinity and the numerator is the number of cases whereas the denominator is the person-time at risk
* Incidence rates can be conceptualized as the speed at which disease is occurring in cases per person year
* When we have complete follow-up of a sample or a population, the rate can

**approximate** the proportion of disease or the risk

## Risks and rates, an example, part 1

* We have 10 people who are disease free at the start of follow-up, each followed for 1 year
  + Three of these individuals develop the disease.
  + All individuals are followed for the entirety of the study period
* The **risk (incidence)** of disease will be 3 out of 10, or 0.3
* Assuming these individuals developed the disease just as the year was ending, and the **rate** would be 3 per 10 person years or 0.3 (equivalent to the risk)

## Rate versus proportion, an example, part 2

* Now suppose that those who developed the disease did so halfway through the year
  + 7 people were followed and did not develop the disease, i.e., 1 person year for each totaling 7 person years
  + 3 people developed the disease, i.e., we assign each of them 0.5 person years for the midpoint of the time interval for a total of 1.5 person years

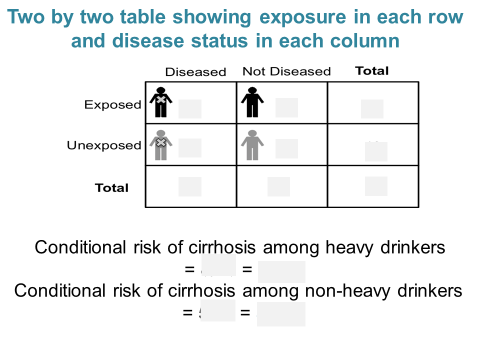
Thus, the **incidence rate** would be 3 per 8.5 person years, or 0.35

## Incidence vs. incidence rate: what’s the difference?

* Because measures of incidence are so central to epidemiological investigation, the term “incidence” can be used in various contexts
* The **incidence** refers to the number of new cases divided by the population at risk. It is also called the **incidence proportion**, or the **risk**
* When we refer to “incidence”, we mean the incidence proportion, also known as the risk
* The **incidence rate** refers to the number of new cases divided by the person-time at risk contributed by members of the study
  1. When we refer to “incidence rate”, we specifically refer to a measure in which the denominator is the person-time at risk contributed by members of the study.
  2. **Conditional risks**
     + We can “condition” risk estimate by other factors to begin to examine whether certain factors are associated with increased or decreased risk
     + Let us return to our earlier example of alcohol consumption an liver cirrhosis
     + In order to estimate whether heavy drinkers have a different incidence of cirrhosis compared with non-heavy drinkers, we can use a measure of the conditional incidence

*Task 10:*

*Complete the contingency (2X2) table for alcohol and cirrhosis? What is the conditional risk in heavy and non-heavy drinkers?*



## Conditional risks

* + - SO: heavy drinkers have a higher incidence of cirrhosis compared with non-heavy drinkers
    - Building these 2 x 2 tables crossing exposure with disease and using **these 2 x 2 tables** to estimate associations will become a **building block of epidemiology**

**NB:** Your protocol must include ‘dummy’ 2X2 tables for all the variables you aim to use in your study

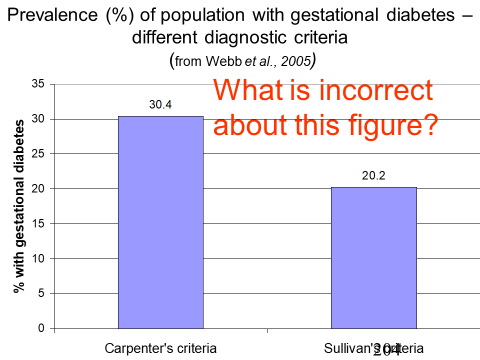
*Task 11*

*What is the application of this Session for preparing your protocol?*

## Summary

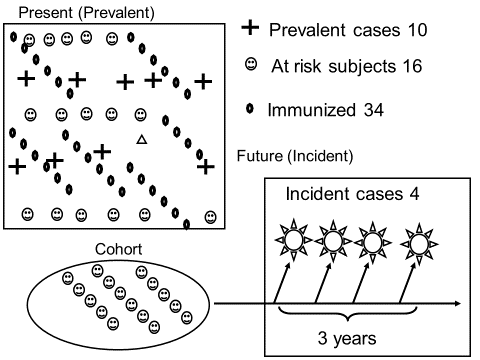
* + - Measures of disease occurrence and frequency in epidemiology are the cornerstone of how we build the science of population health
    - Key measures are: incidence/risk, prevalence, mean, median, mode, incidence rates, and conditional risks
    - Incidence rates are more appropriate than incidence when there are losses to follow-up
    - *Task 12:*

*What is wrong with this Figure?*



## Other measures

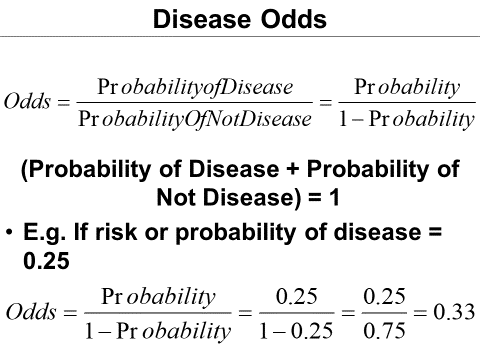
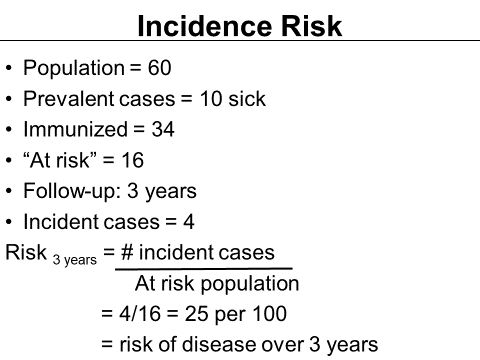
There are a number of other measures that are used in epidemiology / biostatistics which you should be aware of as well



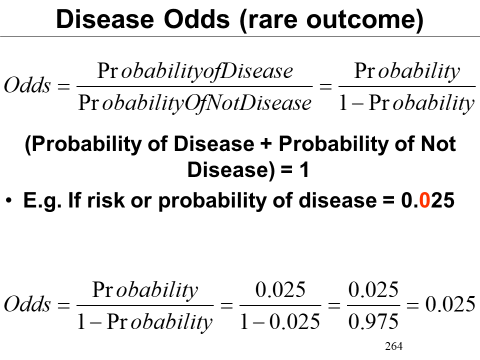
*Task 12:*

*What is the incidence risk or probability of the disease occurring in the schematic data shown above?*

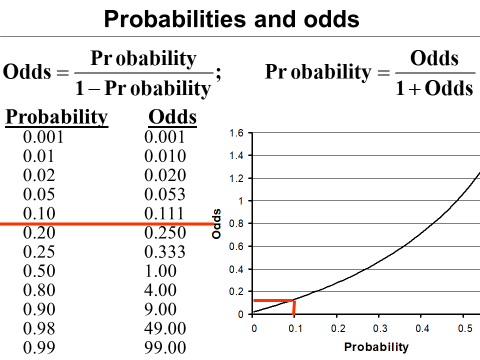
What is the disease odds in this example?

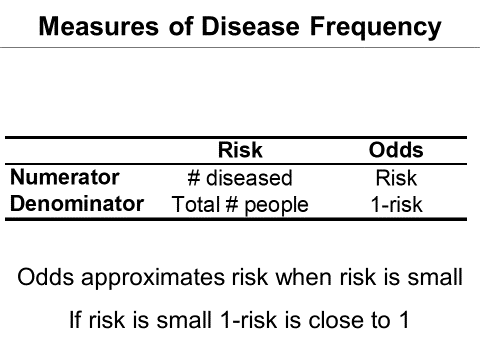


When the occurrence (probability) disease is rare, the odds of disease approximates the probability of disease measure



If the probability of disease is less than 10% then the disease odds can be used interchangeably with the probability





## Probabilities and odds

So in summary:

Either probabilities or Odds may be used to express frequency of disease Odds nearly equals probabilities the probability is small – less than 0.1

Odds can be calculated into in relation to any kind of proportion – prevalence or incidence risk

Because of its statistical properties odds and odds ratios are used frequently in epidemiology

Odds ratios are estimated in case-control studies Odds ratios on modelled in logistic regression

## Definitions of ratio, risk, rates & odds

* **Measure = a or b**
* **Ratio = a/b (where a & b could be distinct quantities)**
* **Risk = a / (a+b)**
* **Survival = (1-risk)**
* **Rate = Risk / time**
* **Odds = Risk / (1-risk)**

(Elandt-Johnson, *Am J Epidemiol,* 1975)

# Unit 1: Session 5 - What is a sample? What are other sources of data?

Having formulated the research question, prepared measureable objectives and defined the study population, one has to decide where the data is going to be collected to answer the research question.

## Sources of data

Existing or New data can be used.

There are two main sources of data - Existing or New

*Task 1:*

*What are some of the sources of existing data?*

|  |  |
| --- | --- |
| **EXISTING DATA** | **NEW DATA** |
| 1. Census 2. Population projections 3. Birth registration 4. Death certificates 5. Disease notifications 6. Routine health service statistics    Hospitals / Clinics / Laboratories / Pharmacies   Patient records   Admissions / Discharges / OPD   Diagnoses / Procedures   Personnel / Finances   1. Disease registries 2. Accident / Occupational registers 3. Published surveys / studies |  Collection tools   1. Questionnaires 2. Interviews 3. Direct measurement 4. Observation |

## 5.1.1 Measuring Disease Occurrence using Routine Data

* + - Routine surveillance data
      * Much data about disease occurrence
    - Crude data
      * Information about individuals
      * In a population (estimate from census)
    - ? Can we relate the occurrence of diseases to particular causes
      * High unemployment
      * High suicide
      * ?? Are they related
    - (ecological fallacy)
    - Mortality data
      * Not good for incidence of common conditions
        + Osteoarthritis, non-melanomatous skin cancers, psoriasis, rubella
    - Notifiable conditions
      * Under-reported
      * Political suppression
        + Cholera, HIV/AIDS, SARS

## Census:

A census is the procedure of systematically acquiring and recording data about the members of a given population. The essential features of population and housing censuses are "individual enumeration, universality within a defined territory, simultaneity and defined periodicity" In low income countries usually population censuses are taken every 10 years. In South Africa up to 80 data-variables are collected. Although population estimates remain an important function of a census, including exactly the geographic distribution of the population, statistics can be produced about combinations of attributes e.g. education by age and sex in different regions. Current administrative data systems allow for other approaches to enumeration with the same level of detail but raise concerns about privacy and the possibility of biasing estimates.

A census can be contrasted with sampling in which information is obtained only from a subset of a population. Typically the main population estimates are updated by such intercensal estimates. In South Africa, the Mid-year Population Estimates are used to update the census data.

Modern census data are commonly used for research, business marketing, and planning, and as a baseline for designing sample surveys by providing a sampling frame such as an address register. Census counts are necessary to adjust samples to be representative of a population by weighting them as is common in opinion polling. Similarly, stratification requires knowledge of the relative sizes of different population strata which can be derived from census enumerations. In many cases, a carefully chosen random sample can provide more accurate information than attempts to get a population census

A census is often construed as the opposite of a sample as its intent is to count everyone in a population rather than a fraction. However, population censuses rely on a [sampling](https://en.wikipedia.org/wiki/Sampling_frame) [frame](https://en.wikipedia.org/wiki/Sampling_frame) to count the population. This is the only way to be sure that everyone has been included as otherwise those not responding would not be followed up on and individuals could be missed. The fundamental premise of a census is that the population is not known and a new estimate is to be made by the analysis of primary data.

A census is one of the main sources of demographic data and measures (see MHD).

*Task 2:*

*What are the strengths and weakness of using census data to answer research questions?*

Other sources of demographic data include:

## Sample surveys

* + - * Demographic Health Survey
      * Demographic surveillance sites – Africa Centre

## Vital registration

* + - * Births & deaths

## Population registers

* + - * Civil registration
      * Voters role

*Task 3:*

*What is a demographic surveillance site?*

## Why take a sample?

To answer health-related research questions **efficiently**

* Complete **census** is the epidemiologic ideal

– Population too big Reasons not to take a census:

* + Lack of time
  + Lack of money
  + Waste of resources
* Can get adequate information from smaller subgroup

– BUT: introduce error / chance

* Target / Reference Population

– To which the study findings could be generalised / external validity

## To take a sample

1. Specify population of interest
2. Specify a research question of interest

## Specify population of interest

What are the **characteristics of the population** in which we want to understand health?

* + - * What is the prevalence of diabetes is in Langa? Western Cape? Natal?
      * What are the risk factors for Type 2 diabetes in this population?

NB: Population of interest has to be specified **before the sampling strategy** is defined

## Specifying a question

* Question of interest can help clarify appropriate way to sample population of interest
* Questions asked can include
* Estimating **population parameters**
* Estimating **causal effects** of **exposures** on **outcomes** (health indicators)

## Estimating population parameters

**Questions concerned with population parameters:**

* + What proportion of individuals in the population of interest has breast cancer?
  + What is the mean blood pressure in the population?
  + How many new cases of HIV are diagnosed in the population over three years?

## Population parameters include estimates of

* + Proportions
  + Means
  + Standard deviations

## Sample required

* + Representative sample

## Estimating causal effects of exposures on outcomes Questions for which these measures are needed are

* + - Does exposure to pollution cause lung cancer?
    - Does suffering abuse in childhood cause depression in adulthood?
    - Does a specific genetic marker cause Alzheimer’s disease?

## Parameter of interest

* + - Causal effect of an exposure on a health outcome

## Sampling concerns

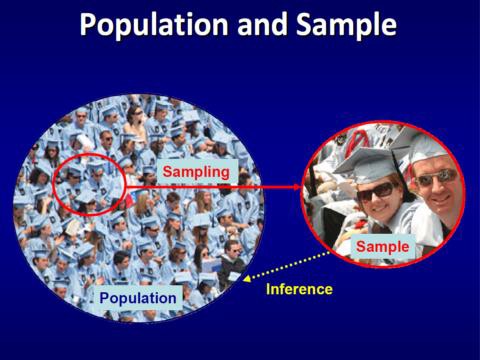
* + - Not representativeness (as in population parameters)
    - Whether individuals exposed to hypothesized cause of interest are **comparable** to individuals not exposed
    - Purposive sample sufficient

## Representative OR Purposive Representative sample

* + - The sample that is taken has **characteristics similar** to the overall population sample
    - **May or may not** include individuals who are comparable with respect to causal identification

## Purposive sample

* + - Selects from the population base on some criterion
    - **May or may not** be representative of a particular population of interest



## How to take a representative sample

Simple random sample

Each member of the population has an **equal probability** of being selected into the sample

A successful **simple random sample** should have the **same basic characteristics** as the original **population**

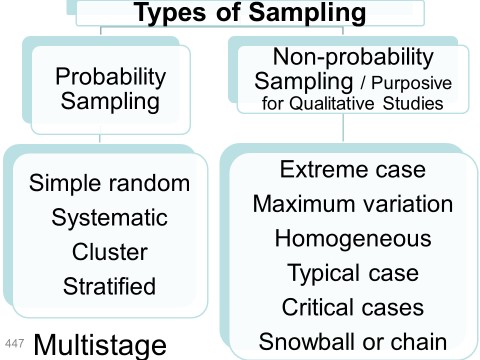
**Taking a simple random sample**

1. **Enumerate** all potential members of population of interest
2. **Assign** each member a probability of selection
3. **Ensure** selection of members are independent

## Populations and Samples

* Study Population
* Group of people from whom to gather data
* Defined by person, time, place
* Inclusion & Exclusion criteria to reduce confounding
* Study Sample / Sample Population

– **Representative** subgroup of total study population



There are two broad categories of sampling.

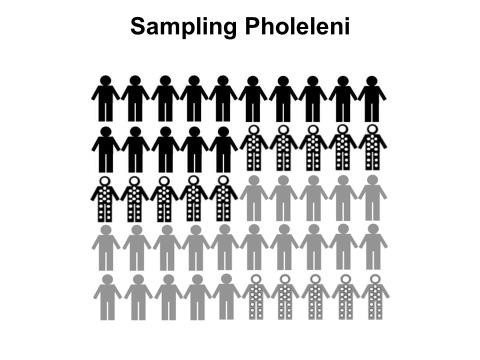
Probability or ‘lotto’ sampling, and nonprobability sampling. The gold standard of sampling is a simple random sample.

For this, the sampling frame needs to be clearly identifiable. If this is not possible systematic and cluster sampling is used.

In many instances, random sampling is part of a multistage and stratified sampling process Nonprobability or purposive sampling is used in studies where representativeness of the population is not required e.g. Experimental study designs, case-control studies and In studies collecting Qualitative data, a number of different ways to perform nonprobability sampling these include extreme case, maximum variation, homogenous, typical case, critical case and snowball or chain nonprobability sampling

## The perfect sample?

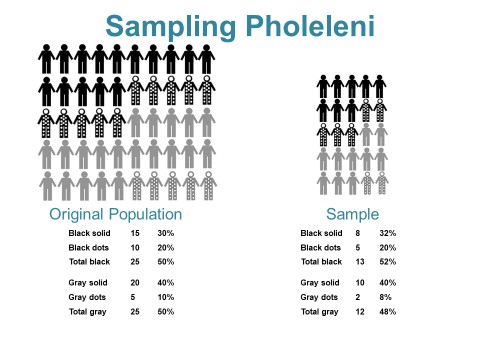
* There is no perfect sample
* The goal in epidemiology is to understand limitations of sampling methods and account for them



We want to collect our sample in such a way that the sample also has **50% exposed** and **30% dotted**.

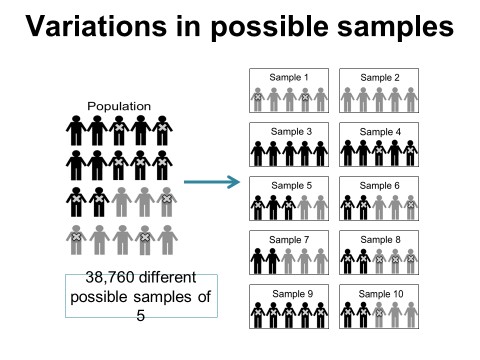
50 individuals– 25 of whom are exposed to a factor of interest (black), and 25 are not (grey). We also indicate in the figure that some are dotted and some are not. These dots could represent another factor of interest, such as age, or cigarette smoking, or living in poverty or not. In the population, 15 individuals have dots. Taken together, the population has the following characteristics: 50% Exposed 30% are dotted. The goal of any sampling Technique is to collect our sample in such a way that the sample also has 50% Exposed, and 30% Dotted

1. We can use a simple random sample
   * ½ the population (25)
   * Probability of selection 1/50 or 2%
   * Random number generator

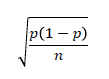


## Quantifying sampling variability

* Sampled population will not have the **exact** same population parameters as complete population census
* The ‘truth’, i.e., the population parameter of original population is called the **true population parameter**



## 5.8.1 Quantifying uncertainty - Central Limit Theorem (CLT)

1. Average proportion across all possible samples = true population proportion
   * Example:
     + 50% of true population has diabetes
     + Sample 1 has 100% diabetes
     + Sample 2 has 0% diabetes
     + Average of **all** samples will have 50% diabetes
2. Variance around average sample proportions (**standard error**)

p = sample proportion n = sample size

1. Large samples will have normally distributed samples

* Distribution is symmetrical
* Peak at average sample proportion
* How large is large?
  + > 30 people
  + No group < 5 people

Therefore the **principal drivers of uncertainty** are

1. Prevalence in the sample
2. Sample size

The larger the sample size, the smaller the amount of uncertainty in the sample estimate

## Purposive sample

Why?

* + - Cannot enumerate all members
    - Not measuring a population parameter

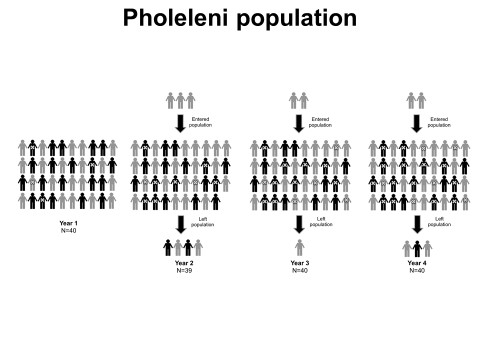
Eligibility criteria for study is the central design element;

Entry is based on exposure status, or sometimes on health outcome status.

**Study design**

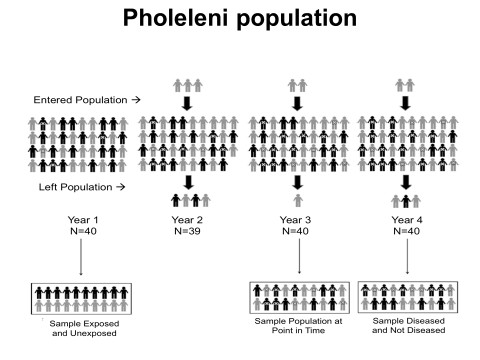
Study design considerations are similar for representative or purposive sample Study design reflects decisions made at one time point or over time

Timing of disease process can inform the study design



This is a dynamic population of people some exposed and unexposed, diseased and not diseased and followed up for 4 years

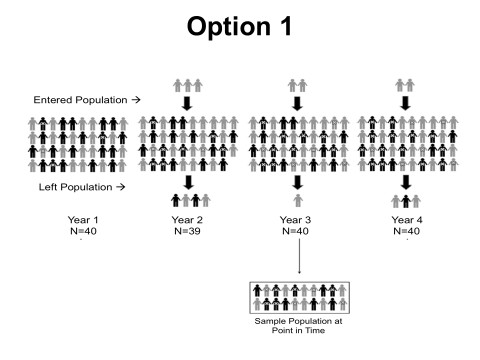
## Study design options



1. Sample one moment in time, irrespective of disease status, measure disease and potential cause simultaneously.
2. Sample over time, start with disease free individuals only, measure disease over time
3. Sample one moment in time, based on disease status

*Task:*

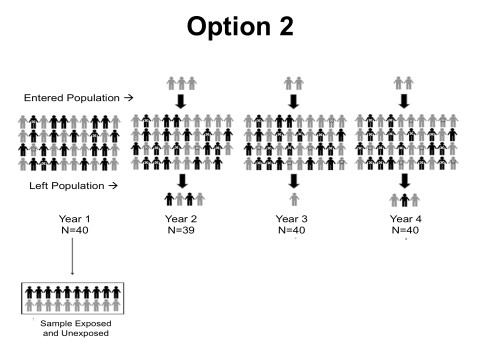
*Which study design is depicted as Option 1? What are the uses, advantages and disadvantages of this study design?*



*Task:*

*Which study design is depicted as Option 2?*

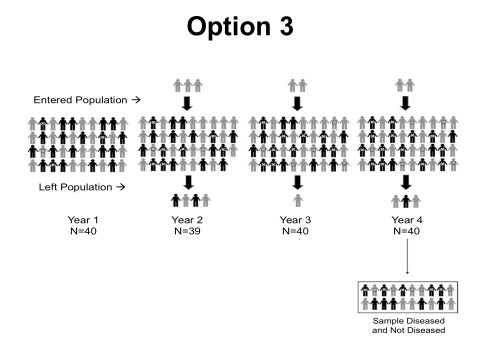
*What are the uses, advantages and disadvantages of this study design?*



*Task:*

*Which study design is depicted as Option 3?*

*What are the uses, advantages and disadvantages of this study design?*



**Summary**

1. Samples are efficient, representative or purposive
2. Representative sample; e.g., simple random sample
3. Sampling variability, standard error
4. Purposive sample, selection on exposure or disease status
5. Study designs can be cross-sectional, cohort, case-control

### Sampling Task

1. ***Although many research questions may be answered using either a representative or purposive sample, for the following studies, which kind of sample would be best suited and why?***
   1. *A study of the effect of sleep deprivation on glucose metabolism.*
   2. *A study comparing the average fertility rate in SA to the average fertility rate in Nigeria.*
   3. *A study to test whether a new cholesterol-lowering drug works among adults at high risk for cardiovascular disease.*
   4. *A study of predictors of sobriety among individuals attending a drug rehabilitation program.*
   5. *A study of the increasing prevalence of obesity in Western Cape between 1980 and 2010.*
   6. *A study of the neurocognitive effects of occupational exposure to pesticides among migrant farm workers in W Cape.*
   7. *A study of the association between smoking and lung cancer among British doctors in the 1950s.*
   8. *A study of the decline in the smoking rate among adult residents of eGoli from 2002 to 2012.*
   9. *A study of the mean age at first sex among Western Cape adolescents.*
   10. *A study of predictors of safe sex among men who have sex with men in Sea Point.*

### What characterizes a simple random sample?

1. ***Discuss the pros and cons of the following methods for selecting a simple random sample to gather data for answering the research question described.***
   1. *To assess mean reading level among fourth graders in Cape Town, you administer a reading test to one classroom of fourth graders in every even-numbered public school in the five boroughs.*
   2. *To determine the average wait time in a hospital emergency room, you spend an entire day interviewing every fifth patient treated between 6am and midnight in the emergency room at Pholeleni General Hospital.*
   3. *In a survey of hypertension among adults in Pholeleni Department of Health employees randomly selected 10% of the households in each census tract and went door to door between 9am and 5pm, Monday through Friday, taking blood pressure measurements.*

### Consider the research questions below and indicate whether the most efficient sample would be exposed and unexposed, the population at a single point in time, or diseased and not diseased.

* 1. *Are people with Lyme disease more or less likely to live in rural areas compared to people without Lyme disease?*
  2. *Is Type 2 diabetes less common among those who exercise more than 3 hours per week compared to those who exercise less than 1 hour per week?*
  3. *Are people who eat canned soup more likely to have elevated levels of bisphenol-A (BPA) compared to people who eat homemade soup?*
  4. *Is living in mixed-gender residences associated with prevalence of sexually transmitted infections in students?*
  5. *Are women with osteoporosis more likely to have had late onset of menstruation compared to women with healthy bone density?*
  6. *Is mood associated with sunlight exposure?*
  7. *Does eating 200 gm red meat per day lead to higher LDL cholesterol compared to eating no red meat?*
  8. *Are women who give birth prematurely more likely to have smoked compared to women who give birth at term?*
  9. *Are babies who survive car crashes more likely to have been sitting in rear-facing infant car seats compared to babies who do not survive?*
  10. *In women age 15 to 24 years, is alcohol consumption associated with unprotected sex?*