INTRODUCTION

Prof. dr. Siswanto Agus Wilopo, M.Sc., Sc.D.  
Department of Biostatistics, Epidemiology and  
Population Health  
Faculty of Medicine  
Universitas Gadjah Mada
Learning Objectives

In this lecture you learn:

- Identify the contents of course syllabus,
- Identify the laboratory exercise activities,
- Record on how course evaluates students’ activities
- List the use of biostatistics
- Assess the difference between nominal, ordinal, and metric discrete and metric continuous variables.
- Distinguish the type of a variable.
- Identify the non-numeric nature of ordinal data.
- Illustrate examples of dependent, independent and confounding variables
Module I:

Part 1: Health Research Data
Basic Concepts of Biostatistics

Biostatistics is concerned with:
- Designing and planning of the study
- Processing and analyzing data
- Collecting, presenting, and transforming data to assist decision makers
The Big Picture

The four-step process that encompasses statistics (as it is presented in this course):

1. Producing Data — Choosing a sample from the population of interest and collecting data.
2. Exploratory Data Analysis (EDA) {Descriptive Statistics} — Summarizing the data we’ve collected.
3. Probability and Inference
4. Drawing conclusions about the entire population based on the data collected from the sample.

Even though in practice it is the second step in the process, we are going to look at Exploratory Data Analysis (EDA) first.
What percentage support the death penalty?

1. Producing Data

2. Exploratory Data Analysis

3. Probability

4. Inference

Conclusion: We can be 95% confident that the population percentage is within 3% of 65% (i.e. between 62% and 68%)
Key Definitions

- A population (universe) is the collection of all members of a group.
- A sample is a portion of the population selected for analysis.
- A parameter is a numerical measure that describes a characteristic of a population.
- A statistic is a numerical measure that describes a characteristic of a sample.
Population vs. Sample

Population

a b c d
ef gh i j k l m n
o p q r s t u v w
x y z

Measures used to describe a population are called parameters

Sample

b c
g i n
o r u
y

Measures computed from sample data are called statistics
Two Common of Biostatistical approach

- Descriptive statistics
  - Collecting, summarizing, and presenting data

- Inferential statistics
  - Drawing conclusions about a population based only on sample data
Descriptive Statistics

- Collect data
  - e.g., Survey

- Present data
  - e.g., Tables and graphs

- Characterize data
  - e.g., Sample mean = \( \frac{\sum X_i}{n} \)
Inferential Statistics

- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight

- **Hypothesis testing**
  - e.g., Test the claim that the population mean weight is 60 kilogram

Drawing conclusions about a population based on sample results.
Collecting Data

**Primary Sources**
- Data Collection
  - Observation
  - Experimentation
- Survey

**Secondary Sources**
- Data Compilation
  - Print or Electronic
Types of Data

- **Categorical**
  - Examples:
    - Marital Status
    - Political Party
    - Eye Color

- **Numerical**
  - **Discrete**
    - Examples:
      - Number of Children
      - Defects per hour
  - **Continuous**
    - Examples:
      - Weight
      - Voltage
Broad classification of the different types of data with examples

**Categorical**
- **Dichotomy**
  - Categories are dichotomous (binary) two levels
  - Examples: Dead/alive, male/female, Case/control

- **Nominal**
  - Categories are mutually Exclusive and unordered
  - Examples: Gender, blood group, eye color, minimal status

- **Ordinal**
  - Categories are mutually Exclusive and ordered
  - Examples: Disease stage, social class, education level

**Numerical**
- **Counts**
  - Integer Values
  - Examples: Days sick per year, number of pregnancies

- **Measured**
  - As: continuous --- interval
  - Takes any value in a range of values
  - Examples: Weight in kg, height in meter, age in years, hours, minute

**Qualitative**
Basic Causal Model

Exposure: FLOURIDISATION

Outcome: CANCER MORTALITY

CONFOUNDING FACTORS:
AGE, SEX, ETHNICITY
Module I:

Part 2:
Study Design in Health Research
Key Concepts

1. Study designs in public health fall into two categories: studies in which subjects are observed, and studies in which the effect of an intervention is observed.

2. Observational studies may be forward-looking (cohort), backward-looking (case-control), or looking at simultaneous events (cross-sectional). Cohort studies generally provide stronger evidence than the other two designs.

3. Studies that examine patient outcomes are increasingly published in the literature; they focus on specific topics, such as resource utilization, functional status, quality of life, patient satisfaction, and cost-effectiveness.

4. Studies with interventions are called experiments or clinical trials. They provide stronger evidence than observational studies.
Key Concepts

5. The single best way to minimize bias is to randomly select subjects in observational studies or randomly assign subjects to different treatment arms in clinical trials.

6. Bias occurs when the way a study is designed or carried out causes an error in the results and conclusions. Bias can be due to the manner in which subjects are selected or data are collected and analyzed.

7. Clinical or intervention trials without controls (subjects who do not receive the intervention) are difficult to interpret and do not provide strong evidence.

8. Each study design has specific advantages and disadvantages.
I. Observational studies

- **Descriptive or case-series**
  - Causes and incidence of disease
  - Identification of risk factors

- **Case-control studies (retrospective)**
  - Causes and incidence of disease
  - Identification of risk factors

- **Cross-sectional studies, surveys (prevalence)**
  - Disease description
  - Diagnosis and staging
  - Disease processes, mechanisms

- **Cohort studies (prospective)**
  - Causes and incidence of disease
  - Natural history, prognosis
  - Identification of risk factors

- **Historical cohort studies**
II. Experimental studies

- Controlled trials
  - Parallel or concurrent controls
    - Randomized
    - Not randomized
  - Sequential controls
    - Self-controlled
    - Crossover
  - External controls (including historical)

- Studies with no controls

III. Meta-analyses
CLASSIFICATION OF STUDY DESIGNS
Schematic diagram of case–control study design

Exposed

Unexposed

Cases

Exposed

Unexposed

Onset of study

Time

Direction of inquiry

Question: "What happened?"

Source: Dawson B, Trapp RG: Basic & Clinical Biostatistics, 4th Edition:
http://www.accessmedicine.com

Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of cross-sectional study design

Subjects selected for the study

With outcome

Without outcome

Onset of study

Time

No direction of inquiry

Question: "What is happening?"


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of cohort study design


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of historical cohort study design

Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of the time relationship among different observational study designs


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of randomized controlled trial design


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of trial with crossover


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Schematic diagram of trial with external controls.


Copyright © The McGraw-Hill Companies, Inc. All rights reserved.
Module 01:

Part 3: Producing Data
Producing Data:

Observational Study
Data Quality

- An analysis is only as good as its data
- GIGO ≡ garbage in, garbage out
- Does a variable measure what it purports to?
  - Validity = freedom from systematic error
  - Objectivity = seeing things as they are without making it conform to a worldview
- Consider how the wording of a question can influence validity and objectivity
## From Exploration to Inference

<table>
<thead>
<tr>
<th>Exploratory Data Analysis</th>
<th>Statistical Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose: search and describe patterns</td>
<td>Purpose: answer specific question</td>
</tr>
<tr>
<td>Conclusions apply only to data and specific circumstance</td>
<td>Conclusions apply beyond data and broad circumstance</td>
</tr>
<tr>
<td>Conclusion are informal</td>
<td>Conclusions are formal and apply probabilities</td>
</tr>
</tbody>
</table>
Surveys

- Goal: to describe population characteristics
- Studies a subset (sample) of the population
- Uses sample to make inferences about population
- Sampling:
  - Saves time
  - Saves money
  - Allows resources to be devoted to greater scope and accuracy
Illustrative Example: *Youth Risk Behavior Surveillance (YRBS).* The Youth Risk Behavior Surveillance System monitors health behaviors in youth and young adults in the United States. Six categories of health-risk behaviors are monitored. These include: (1) behaviors that contribute to unintentional injuries and violence; (2) tobacco use; (3) alcohol and drug use; (4) sexual behaviors; (5) unhealthy dietary behaviors; and (6) physical activity levels and body weight. The 2003 report used information from 15,240 questionnaires completed at 158 schools to infer health-risk behaviors for the public and private school student populations of the United States and District of Columbia.\(^a\) The 15,240 students who completed the questionnaires comprise the sample. This information is used to infer the characteristics of the several million public and private school students in the United States for the period in question.  

\(^a\)
Simple Random Sampling

- Probability samples entail chance in the selection of individuals
- This allows for generalizations to population
- The most fundamental type of probability sample is the simple random sample (SRS)
- SRS (defined): an SRS of size $n$ is selected so that all possible combinations of $n$ individuals from the population are equally likely to comprise the sample, SRSs demonstrate sampling independence
Cautions when Sampling

- **Under-coverage:**
  - groups in the source population are left out or underrepresented in the population list used to select the sample

- **Volunteer bias:**
  - occurs when self-selected participants are atypical of the source population

- **Non-response bias:**
  - occurs when a large percentage of selected individuals refuse to participate or cannot be contacted
Other Types of Probability Samples

- Stratified random samples
- Cluster samples
- Multistage sampling

These are advanced techniques not generally covered in introductory courses.
Illustrative Example of an Observational Study (Weight Gain & CHD)

- **Purpose of study:**
  - to understand the relationship between body mass and coronary heart disease (CHD)

- Started in 1976 with 115,818 women, 30 to 55 years of age
- Current weight and weight at age 18 determined
- Participants followed for 14 years
- Number of fatal and nonfatal CHD cases counted (1292 cases)
- Results mathematically *adjusted* for lurking factors such as smoking and family history of CHD

Illustrative Example: Results

Compared those who gained less than 11 pounds:

- 11 to 17 lbs gained: 25% more likely to develop CHD
- 17 to 24 lbs gained: 64% more likely
- 24 to 44 lbs gained: 92% more likely
- 44+ lbs gained: 165% more likely
Illustrative Example (Questions)

- What is the population in this study?
- What is the sample in this study?
- Is this study experiment or observational?
- Can we say weight gain is associated with CHD?
- Can we say weight gain caused CHD?
Sample Quality

- Sampling bias ≡ a sampling techniques that favor a certain outcome. This produces a poor quality samples and misleading results

- Examples of sampling bias
  - Voluntary response sampling: Allows individuals to choose to be in the study, e.g., call-in polls
  - Convenience sampling: individuals that are easiest to reach are selected, e.g., Interviewing at the mall
Voluntary Response Bias

- To prepare for her book *Women and Love*, Shere Hite sent questionnaires to 100,000 women asking about love and sexual relationships.
- Only 4.5% responded
- Respondents
  - “were fed up with men and eager to fight them…”
- Selection bias:
  - “angry women [were] are more likely” to respond.
- These women are not typical of the female population.
Convenience Sample

- A lab study was conducted to see if drug affected physical activity in caged lab animals.
- The lab assistant reached into the cage to select the mice for study.
- Which mice will likely be chosen? Mice that are less active will be selected (sampling bias).
Simple Random Sample (SRS)

- To avoid sampling bias, we use \textit{chance mechanisms} to select study subjects
- The Simple Random Sample (SRS) is the most basic chance selection technique
- Characteristics of SRSs
  1. Each individual in population has the same chance of being selected
  2. Every possible sample has an the chance to be studied
Selecting a SRS

- Physical, e.g., pick numbers from a hat
- Random number generators (computers)
- Table of random digits (e.g., Table B)
  - Each entry is equally likely to be 0 through 9
  - Entries are *independent* (knowledge of one entry gives no information about any other entries)
Choosing a Simple Random Sample (SRS)

STEP 1: Label each individual in the population with a identification number

STEP 2: Use Table B to select numbers at random (we will always use Table B; enter Table B at a different location each time or as specified)
Picking a SRS (Illustration)

- Population consists of 30 individuals
- Labeled the individuals 01 – 30
- Select a row in table at random
- For illustrative purposes, let us start in row 106
- Row 106 with lines to indicate pairs
  68 | 41 | 7  3 | 50 | 13 | 15 | 52 | 9
- First two individuals relevant numbers are 13 and 15
Remainder

- **Under-coverage:**
  - some population groups left out of sampling process

- **Non-response bias:**
  - some selected individuals can’t be contacted or refuse to participate

- **Even good quality samples may not be a perfect reflection of the population because of random error sampling error**
Producing Data: Experiments
Experimentation

- Recall this distinction:
- Experimental studies: investigator imposes an exposure or treatment to see its effect
- Observational studies: individuals are studied without intervention
Vocabulary

- Subjects = people studied in an experiment
- Factors = specific experimental conditions or interventions applied to subjects
- Treatment = a combination of a specific set of factors
Illustrative Example: Effects of Advertising

- Undergraduate students viewed a 40-minute television program that included ads for a digital camera
- Two explanatory variables ("factors"): message length, repetition
- Three response variables. After viewing, subjects answered questions about recall of the ads, attitude toward the camera, intention to purchase
Illustrative Example: Treatments

- **Factor A**: length of the commercial (2 levels): some subjects saw a 30-second commercial; others saw a 90-second version
- **Factor B**: Number of repetitions (3 levels) Commercial were shown either 1, 3, or 5 times during the program
- 6 combinations of factors → 6 treatments

<table>
<thead>
<tr>
<th>Factor A: Length</th>
<th>Factor B: Repetitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 seconds</td>
<td>1 time</td>
</tr>
<tr>
<td>90 seconds</td>
<td>3 times</td>
</tr>
</tbody>
</table>

Subjects assigned to treatment 3 see a 30-second advert. five times
Comparison

**Comparison** is first principle of experimentation: The effects of a treatment can be judged only in relation to what would happen in its absence.

You cannot assess the effects of a treatment without a comparison group because:

- Many factors contribute to a response
- Conditions change on their own over time
- Consider the Placebo effect as an example
Randomization

Randomization = use of chance mechanisms to assign treatments

Randomization balances lurking variables among treatments groups, mitigating confounding by lurking variables!
**Blinding**

Blinding is the third principle of experimentation

- Blinding = assessment of the response in subjects is made without knowledge of which treatment they are receiving
- Single blinding = subjects are unaware of treatment group
- Double blinding = subjects and investigators are blinded
Illustrative Example: Quitting Smoking with Nicotine Patches

- Explanatory variable: Nicotine patch / placebo patch
- 60 subjects, 30 assigned to each treatment group
- Response variable: Cessation of smoking (yes/no)
- Study design outline:

  Random Assignment
  
  Group 1
  30 smokers
  
  Treatment 1
  Nicotine Patch
  
  Group 2
  30 smokers
  
  Treatment 2
  Placebo Patch
  
  Compare Cessation rates

Source: *JAMA*, Feb. 23, 1994, pp. 595-600

- Subjects (30 undergraduate students) randomly assigned to one of three treatment groups
  - **Group 1**: Listen to Mozart
  - **Group 2**: Listen to relaxation tapes
  - **Group 3**: Silence
- Response variable: change in IQ score

Random Assignment → Group 1: 10 students → Treatment 1: Mozart → Compare Change in IQ score

Random Assignment → Group 2: 10 students → Treatment 2: Relaxation

Random Assignment → Group 3: 10 students → Treatment 3: Silence
The Logic of Randomization

- Randomization ensures that difference in the response are due to either
  - the treatment or
  - chance assignment of treatments
- Lurking variables will tend to distribute evenly among the treatment groups (preventing confounding)
- If an experiment finds a difference, we ask whether this difference is due to the treatment or chance.
- If the observed difference is larger than what would be expected just by chance, we say the results are statistically significant
Example of an Experimental Design

The Women's Health Initiative study randomly assigned about half its subjects to a group that received hormone replacement therapy (HRT).

Subjects were followed for ~5 years to ascertain various health outcomes, including heart attacks, strokes, the occurrence of breast cancer and so on.
Example of a Nonexperimental Design

The Nurse's Health study classified individuals according to whether they received HRT.

Subjects were followed for ~5 years to ascertain the occurrence of various health outcomes.
Comparison of Experimental and Nonexperimental Designs

- In both the experimental (WHI) study and non-experimental (Nurse’s Health) study, the relationship between HRT (explanatory variable) and various health outcomes (response variables) was studied.

- In the experimental design, the investigators controlled who was and who was not exposed.

- In the non-experimental design, the study subjects (or their physicians) decided on whether or not subjects were exposed.
Let us focus on selected experimental design concepts and techniques

Experimental designs provides a paradigm for nonexperimental designs.
Jargon

- **A subject** ≡ an individual participating in the experiment
- **A factor** ≡ an explanatory variable being studied; experiments may address the effect of multiple factors
- **A treatment** ≡ a specific set of factors
Subjects, Factors, Treatments
(Illustration)

Illustrative Example: Hypertension trial. A trial looked at two explanatory factors in the treatment of hypertension. Factor A was a health-education program aimed at increasing physical activity, improving diet, and lowering body weight. This factor had two levels: active treatment or passive treatment. Factor B was pharmaceutical treatments at three levels: Medication A, Mediation B, and placebo. Because there were two levels of the health-education variable and three levels of pharmacological variable, the experiment evaluated six treatments, as shown in Table 2.2.

The response variable was “change in systolic blood pressure” after six months. One-hundred-twenty subjects were studied in total, with equal numbers assigned to each group. Figure 2.3 is a schematic of the study design.
Subjects, Factors, Treatments, Example, cont.

- **Subjects** = 100 individuals who participated in the study
- **Factor A** = Health education (active, passive)
- **Factor B** = Medication (Rx A, Rx B, or placebo)
- **Treatments** = the six specific combinations of factor A and factor B

<table>
<thead>
<tr>
<th>Factor A Health Education</th>
<th>Factor B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>Medication A</td>
</tr>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Passive</td>
<td>4</td>
</tr>
</tbody>
</table>
Schematic Outline of Study Design

Random assignment

Group 1
20 hypertensives

Group 2
20 hypertensives

Group 3
20 hypertensives

Group 4
20 hypertensives

Group 5
20 hypertensives

Group 6
20 hypertensives

Treatment 1
Active health ed. + med. A

Treatment 2
Active health ed. + med. B

Treatment 3
Active health ed. + placebo

Treatment 4
Passive health ed. + med. A

Treatment 5
Passive health ed. + med. B

Treatment 6
Passive health ed. + placebo

Compare change in blood pressure
Three Important Experimentation Principles:

- Controlled comparison
- Randomized
- Blinded
“Controlled” Trial

- The term “controlled” in this context means there is a non-exposed “control group”
- Having a control group is essential because the effects of a treatment can be judged only in relation to what would happen in its absence
- You cannot judge effects of a treatment without a control group because:
  - Many factors contribute to a response
  - Conditions change on their own over time
  - The placebo effect and other passive intervention effects are operative
Randomization

- *Randomization* is the second principle of experimentation
- Randomization refers to the use of chance mechanisms to assign treatments
- Randomization balances lurking variables among treatments groups, mitigating their potentially confounding effects
Randomization - Example

Consider this study (*JAMA* 1994;271: 595-600)

- Explanatory variable: Nicotine or placebo patch
- 60 subjects (30 each group)
- Response: Cessation of smoking (yes/no)

**Diagram:**

- Group 1: 30 smokers
  - Treatment 1: Nicotine Patch
  - Random Assignment
- Group 2: 30 smokers
  - Treatment 2: Placebo Patch
  - Compare Cessation rates
Blinding

- **Blinding** is the third principle of experimentation.
- Blinding refers to the measurement of the response of a response made without knowledge of treatment type.
- Blinding is necessary to prevent differential misclassification of the response.
- Blinding can occur at several levels of a study designs:
  - Single blinding - subjects are unaware of specific treatment they are receiving.
  - Double blinding - subjects and investigators are blinded.
Ethics

- Informed consent
- Beneficence
- Equipoise
- Independent (IRB) over-sight