General Biostatistics Concepts

Dongmei Li
Department of Public Health Sciences
Office of Public Health Studies
University of Hawai’i at Mānoa
Outline

1. What is Biostatistics?
2. Types of Measurements
3. Organization of Data
4. Surveys
5. Comparative Studies
1. Biostatistics

- A discipline concerned with the treatment and analysis of numerical data derived from public health, biomedical and biological studies.
  - Design of experiment
  - Collection and organization of data
  - Summarization of results
  - Interpretation of findings
Biostatisticians are:

- **Data detectives**
  - who uncover patterns and clues
  - This involves *exploratory data analysis (EDA)* and *descriptive statistics*

- **Data judges**
  - who judge and confirm clues
  - This involves *statistical inference*
2. Types of measurements

- Measurement (defined): the assigning of numbers and codes according to prior-set rules (Stevens, 1946).

- There are three broad types of measurements:
  - Categorical
  - Ordinal
  - Quantitative
**Measurement Scales**

- **Categorical** - classify observations into named categories,
  - e.g., HIV status classified as “positive” or “negative”

- **Ordinal** - categories that can be put in rank order
  - e.g., Stage of cancer classified as stage I, stage II, stage III, stage IV

- **Quantitative** – true numerical values that can be put on a number line
  - e.g., age (years)
  - e.g., Serum cholesterol (mg/dL)
Illustrative Example:
Weight Change and Heart Disease

This study sought to determine the effect of weight change on coronary heart disease risk. It studied 115,818 women 30- to 55-years of age, free of CHD over 14 years. Measurements included:

- Body mass index (BMI) at study entry
- BMI at age 18
- CHD case onset (yes or no)

Source: Willett et al., 1995
Illustrative Example (cont.)

Examples of Variables

- **Categorical**
  - Smoker (current, former, no)
  - CHD onset (yes or no)
  - Family history of CHD (yes or no)

- **Ordinal**
  - Non-smoker, light-smoker, moderate smoker, heavy smoker

- **Quantitative**
  - BMI (kgs/m$^3$)
  - Age (years)
  - Weight presently
  - Weight at age 18
Exercise

Variable types. Classify each of the measurements listed here as quantitative, ordinal, or categorical.

- White blood cells per deciliter of whole blood
- Presence of type II diabetes mellitus (yes or no)
- Body temperature (degrees Fahrenheit)
- Grade in a course coded: A, B, C, D, or F
- Movie review rating: 1 star, 2 star, 3 star and 4 star
Variable, Value, Observation

- **Observation** $\equiv$ the unit upon which measurements are made, can be an individual or aggregate
- **Variable** $\equiv$ the generic thing we measure
  - e.g., AGE of a person
  - e.g., HIV status of a person
- **Value** $\equiv$ a realized measurement
  - e.g., “27”
  - e.g., “positive”
3. Organization of Data

Data Collection Form

Var1 (ID) 1
Var2 (AGE) 27
Var3 (SEX) F
Var4 (HIV) Y
Var5 (KaposiSarc) Y
Var6 (REPORTDATE) 4/25/89
Var7 (OPPORTUNIS) N

On this form, each questionnaire contains an observation

Each question corresponds to a variable
U.S. Census Form

Start Here

Please use a black or blue pen.

1. How many people were living or staying in this house, apartment, or mobile home on April 1, 2000?

   | Number of people

INCLUDE in this number:
- foster children, roomers, or housemates
- people staying here on April 1, 2000 who have no other permanent place to stay
- people living here most of the time while working, even if they have another place to live

DO NOT INCLUDE in this number:
- college students living away while attending college
- people in a correctional facility, nursing home, or mental hospital on April 1, 2000
- Armed Forces personnel living somewhere else
- people who live or stay at another place most of the time

4. What is Person 1’s telephone number? We may call this person if we don’t understand an answer.

   Area Code + Number
   _______ - _______ - _______

5. What is Person 1’s sex? Mark ONE box.
   - Male
   - Female

6. What is Person 1’s age and what is Person 1’s date of birth?
   Age on April 1, 2000
   ______

   Print numbers in boxes
   Month   Day   Year of birth
   ______   ______    ______

→ NOTE: Please answer BOTH Questions 7 and 8.
## Data Table

<table>
<thead>
<tr>
<th>AGE</th>
<th>SEX</th>
<th>HIV</th>
<th>ONSET</th>
<th>INFECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>M</td>
<td>Y</td>
<td>12-OCT-07</td>
<td>Y</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>N</td>
<td>30-MAY-05</td>
<td>Y</td>
</tr>
<tr>
<td>32</td>
<td>F</td>
<td>N</td>
<td>11-NOV-06</td>
<td>N</td>
</tr>
</tbody>
</table>

- Each **row** corresponds to an **observation**
- Each **column** contains information on a **variable**
- Each **cell** in the table contains a **value**
Illustrative Example: Cigarette Consumption and Lung Cancer

<table>
<thead>
<tr>
<th>county</th>
<th>cig1930</th>
<th>mortalit</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 USA</td>
<td>1300</td>
<td>20</td>
</tr>
<tr>
<td>2 GrBrit</td>
<td>1100</td>
<td>46</td>
</tr>
<tr>
<td>3 Finland</td>
<td>1100</td>
<td>35</td>
</tr>
<tr>
<td>4 Switzerlan</td>
<td>510</td>
<td>25</td>
</tr>
<tr>
<td>5 Canada</td>
<td>500</td>
<td>15</td>
</tr>
<tr>
<td>6 Holland</td>
<td>490</td>
<td>24</td>
</tr>
<tr>
<td>7 Australia</td>
<td>480</td>
<td>18</td>
</tr>
<tr>
<td>8 Denmark</td>
<td>380</td>
<td>17</td>
</tr>
<tr>
<td>9 Sweden</td>
<td>300</td>
<td>11</td>
</tr>
<tr>
<td>10 Norway</td>
<td>250</td>
<td>9</td>
</tr>
<tr>
<td>11 Iceland</td>
<td>230</td>
<td>6</td>
</tr>
</tbody>
</table>

cig1930 = per capita cigarette use in 1930
mortality = lung cancer mortality per 100,000 in 1950

Unit of observation in these data are individual regions, not individual people.
Types of Studies

**Surveys:** describe population characteristics (e.g., a study of the prevalence of hypertension in a population)

**Comparative studies:** determine relationships between variables (e.g., a study to address whether weight gain causes hypertension)

```
Surveys
/   
Studies        Experimental
\   /     
Comparative studies     Nonexperimental ("observational")
\   
Nonexperimental ("observational")
```
4. 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. 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.
Simple Random Samples (SRS)

The reason that we use SRS:

- To generalize the result from the samples to the entire population we are interested.

The idea of SRS is sampling independence:

- Each population member has the same probability of being selected into the sample.
- The selection of any individual into the sample does not influence the likelihood of selecting any other individual.
Simple Random Sampling Method

Example of randomly choose 20 subjects from 1000 subjects:

1. Number population members 1, 2, . . . , 1000
2. Alternatively, use a random number generator (e.g., www.random.org) to generate 20 random numbers between 1 and 1000.
3. Use function in software such as the EXCEL Data Analysis ToolPak
Simple Random Sampling Method

- Install the Data Analysis ToolPak in Microsoft Excel
  - Click the **Microsoft Office Button**, and then click **Excel Options**.
  - Click **Add-Ins**, and then in the **Manage** box, select **Excel Add-ins**.
  - Click **Go**.
  - In the **Add-Ins available** box, select the **Analysis ToolPak** check box, and then click **OK**.
Simple Random Sampling Method using Excel
Simple Random Sampling Method using Excel

<p>| | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td>E</td>
<td>F</td>
<td>G</td>
<td>H</td>
</tr>
<tr>
<td>1</td>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>925</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>614</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>931</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>358</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>860</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>748</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>176</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>290</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>232</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>632</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>785</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>108</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>574</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>509</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>302</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>840</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>852</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>316</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Random Number Generation**

- **Number of Variables:** 1
- **Number of Random Numbers:** 20
- **Distribution:** Uniform
- **Parameters:**
  - Between 0 and 1000
- **Random Seed:** 121

**Output options**
- Output Range:
- New Worksheet Ply:
- New Workbook
Cautions when Sampling

- **Undercoverage**: groups in the source population are left out or underrepresented in the population list used to select the sample.
  - EX: Choose SRS from phone list.

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

- **Nonresponse bias**: occurs when a large percentage of selected individuals refuse to participate or cannot be contacted.
  - EX: Sensitive topics.
Other Types of Random Samples

- **Stratified random samples**
  - Draws independent SRSs from within relatively **homogeneous** groups or "strata".

- **Cluster samples**
  - Randomly select **large units** (clusters) consisting of smaller subunits.

- **Multistage sampling**
  - Large-scale units are selected at random.
  - Subunits are sampled in successive stages.
5. Comparative Studies

Comparative designs study the relationship between an explanatory variable and response variable.

Comparative studies may be experimental or non-experimental.

In experimental designs, the investigator assigns the subjects to groups according to the explanatory variable (e.g., exposed and unexposed groups).

In nonexperimental designs, the investigator does not assign subjects into groups; individuals are merely classified as “exposed” or “non-exposed.”
Study Design Outlines

**Experimental**

- Random assignment
- Group 1: $n_1$ subjects
- Treatment 1
- Compare outcomes
- Group 2: $n_2$ subjects
- Treatment 2

**Nonexperimental**

- Classify individuals
- Group 1 (no intervention): $n_1$ exposed individual
- Compare outcomes
- Group 2 (no intervention): $n_2$ non-exposed individual

**FIGURE 2.1** Experimental and nonexperimental study designs.
Example of an Experimental Design

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

Subjects were followed for \( \sim 5 \) 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 nonexperimental (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 nonexperimental design, the study subjects (or their physicians) decided on whether or not subjects were exposed.
Excercise

Determine whether the following studies are experimental or nonexperimental and identify the explanatory variables and response variables.

- A study of cell phone use and primary brain cancer suggested that cell phone use was not associated with an elevated risk of brain cancer.

- Records of more than three-quarters of a million surgical procedures conducted at 34 different hospitals were monitored for anesthetics safety. The study found a mortality rate of 3.4% for one particular anesthetic. No other major anesthetics was associated with mortality greater than 1.9%.
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** = 120 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></td>
<td>Medication B</td>
</tr>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Placebo</td>
</tr>
<tr>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Passive</td>
<td>Medication A</td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Medication B</td>
</tr>
<tr>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Placebo</td>
</tr>
<tr>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>
Definitions in design of experiment

- **Explanatory variable (independent variable)**
  - A variable which is used in a relationship to explain or to predict changes in the values of response variable.

- **Response variable (dependent variable)**
  - Outcome or response being investigated.

- **Lurking variable (confounding factor, confounder)**
  - A variable that has an important effect on the response variable in a study but is not included among the explanatory variables studied.

- **Confounding effect (effect of lurking variable)**
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)
Randomization – Example

- Number subjects 01,…,60
- Use Excel to select 30 random numbers between 01 and 60
- Keep selecting random numbers until you identify 30 unique individuals
- The remaining subjects are assigned to the control group
Blinding

- *Blinding* is the third principle of experimentation

- Blinding: an experimental technique in which individuals involved in the study are kept unaware of treatment assignments.

- 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
Questions ?