MSIT 431 Probability and Statistical Methods

Chapter 3 Producing Data

Dongning Guo

Fall 2017

Chapter 3 Producing Data

Introduction

3.1 Sources of Data

3.2 Design of Experiments

3.3 Sampling Design

3.4 Ethics

2

3.1 Sources of Data

- Anecdotal data
- Available data
- Sample surveys and experiments
- Observation vs. experiment

3

Obtaining data

Available data (e.g., from library or the Internet).

Anecdotal data (may or may not be representative, sometimes outliers).

Can also **design** observational or experimental studies to collect data.

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. The purpose is to describe some group or situation.

An **experiment** deliberately imposes some treatment on individuals to measure their responses. The purpose is to study whether the treatment causes a change in the response.

A frequently used method is to do sample surveys.

4

3.2 Design of Experiments

- Experimental units, subjects, treatments
- Comparative experiments
- Bias
- Principles of experimental design
- Statistical significance
- Matched pairs design
- Block design

5

Individuals, Factors, Treatments

- An **experimental unit** is the smallest entity to which a treatment is applied. When the units are human beings, they are often called **subjects.**
- A specific condition applied to the individuals in an experiment is called a **treatment.**
- The explanatory variables in an experiment are often called **factors.**

6

Confounding

A **lurking variable** is a variable that is not among the explanatory or response variables in a study but that may influence the response variable.

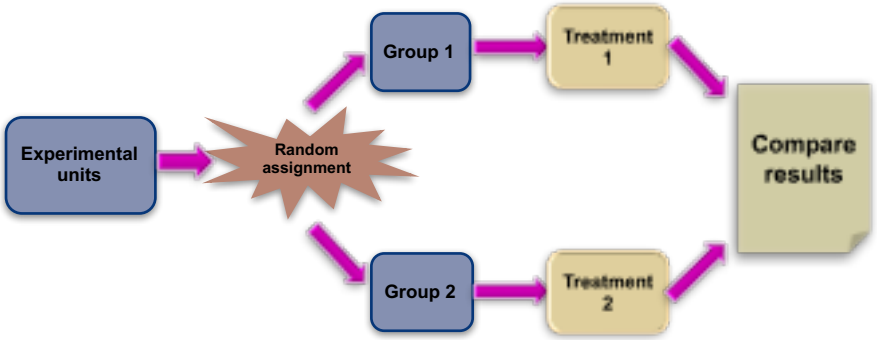
Confounding occurs when two variables are associated in such a way that their effects on a response variable cannot be distinguished from each other.

Well-designed experiments take steps to avoid confounding.

7

Randomized comparative experiments

The remedy for confounding is to perform a **comparative experiment**. In a **completely randomized design,** the treatments are assigned to all the experimental units completely by chance.



Some experiments may include a **control group** that receives an inactive treatment or an existing baseline treatment.

8

Randomization: chose *n* from a group

- Label each of the *N* individuals with a number (say from 1 to *N*).
- Imagine writing the whole numbers from 1 to *N* on separate pieces of paper. Now put all the numbers in a hat.
- Mix up the numbers and randomly select one.
- Mix up the remaining $N - 1$ numbers and randomly select one of them.
- Continue in this way until we have our sample of *n* numbers. Statistical software can do this for you, so you don't actually need a hat!

9

Principles of Experimental Design

Randomized comparative experiments are designed to give good evidence that differences in the treatments actually cause the differences we see in the responses.

Principles of Experimental Design

1. **Control** for lurking variables that might affect the response, most simply by comparing two or more treatments.
2. **Randomize:** Use chance to assign experimental units to treatments.
3. **Replication:** Use enough experimental units in each group to reduce chance variation in the results.

An observed effect so large that it would rarely occur by chance is called **statistically significant.**
A statistically significant association in data from a well-designed experiment does imply causation.

10

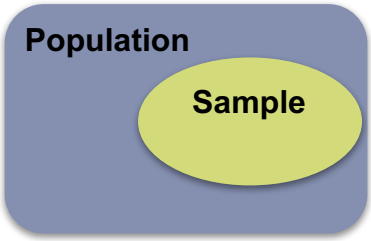
3.3 Sampling Design

- Population and sample
- Voluntary response sample
- Simple random sample
- Stratified samples
- Undercoverage and nonresponse

11

Population and Sample

- The **population** is the entire group of individuals.
- A **sample** is the part of the population from which we actually collect information.
- **Inference:** To use information from a sample to draw conclusions about the entire population.



12

How to Sample Badly

- The design of a sample is **biased** if it systematically favors certain outcomes.

Choosing individuals simply because they are easy to reach results in a **convenience sample.**

A **voluntary response sample** consists of people who choose themselves by responding to a general appeal. Voluntary response samples often show bias because people with strong opinions (often in the same direction) may be more likely to respond.

13

Simple Random Samples

Random sampling, the use of chance to select a sample, is the central principle of statistical sampling.

A **simple random sample (SRS)** of size *n* consists of *n* individuals from the population chosen in such a way that every set of *n* individuals has an equal chance to be the sample actually selected.

In practice, people use random numbers generated by a computer or calculator to choose samples.

14

How to Choose an SRS

A **table of random digits** is a long string of the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 with these properties:

- Each entry in the table is equally likely to be any of the 10 digits 0–9.
- The entries are independent of one another. That is, knowledge of one part of the table gives no information about any other part.

How to Choose an SRS Using Table B

Step 1: Label. Give each member of the population a numerical label of the *same length.*
Step 2: Table. Read consecutive groups of digits of the appropriate length from Table B.

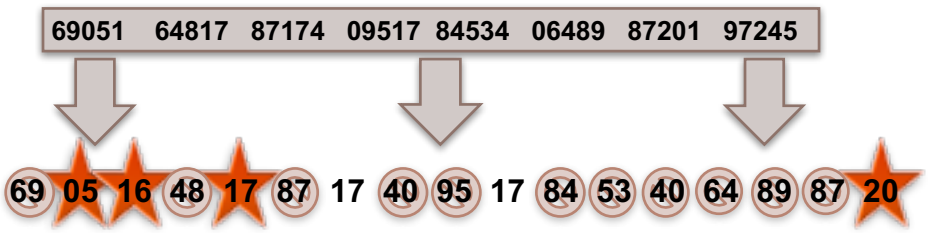
Your sample contains the individuals whose labels you find.

15

SRS Example

Use the random digits provided to select an SRS of four hotels.

01 Aloha Kai	08 Captiva	15 Palm Tree	22 Sea Shell
02 Anchor Down	09 Casa del Mar	16 Radisson	23 Silver Beach
03 Banana Bay	10 Coconuts	17 Ramada	24 Sunset Beach
04 Banyan Tree	11 Diplomat	18 Sandpiper	25 Tradewinds
05 Beach Castle	12 Holiday Inn	19 Sea Castle	26 Tropical Breeze
06 Best Western	13 Lime Tree	20 Sea Club	27 Tropical Shores
07 Cabana	14 Outrigger	21 Sea Grape	28 Veranda



Our SRS of four hotels for the editors to contact is: 05 Beach Castle, 16 Radisson, 17 Ramada, and 20 Sea Club.

16

Other Sampling Designs

The basic idea of sampling is straightforward: Take an SRS from the population and use your sample results to gain information about the population.

A **probability sample** is a sample chosen by chance. We must know what samples are possible and what chance, or probability, each possible sample has.

Sometimes, there are statistical advantages to using more complex sampling methods. One common alternative to an SRS involves sampling important groups (called strata) within the population separately. These "sub-samples" are combined to form one stratified random sample.

To select a **stratified random sample,** first classify the population into groups of similar individuals, called **strata.** Then choose a separate SRS in each stratum and combine these SRSs to form the full sample.

17

Cautions About Sample Surveys

Good sampling technique includes the art of reducing all sources of error.

Undercoverage occurs when some groups in the population are left out of the process of choosing the sample.

Nonresponse occurs when an individual chosen for the sample can't be contacted or refuses to participate.

A systematic pattern of incorrect responses in a sample survey leads to **response bias.**

The **wording of questions** is the most important influence on the answers given to a sample survey.

18

3.4 Ethics

Basic Data Ethics

The most complex issues of data ethics arise when we collect data from people.

Basic Data Ethics

The organization that carries out the study must have an **institutional review board** that reviews all planned studies in advance in order to protect the subjects from possible harm.

All individuals who are subjects in a study must give their **informed consent** before data are collected.

All individual data must be kept **confidential.** Only statistical summaries for groups of subjects may be made public.