Chapter 2 Producing Data

2.1 Introduction

Scientific method

1. Formulate a theory
2. Collect data to test the theory
3. Analyze the results
4. Interpret the results and make decisions.

Interpret results & make decision (1, 9, 10, 11)

Formulate theories

Collect data (2, 3)

Summarize results (4, 5, 6)
2.2 Why Sample?

Census - a sample consisting of the entire population.

2.3 The Language of Sampling.

Population - the entire group of objects or individuals under study, about which information is wanted.

Unit - an individual object or person in the population.

Subject (people) - a part of the population that is actually used to get information.

Variable - a characteristic of interest to be measured for each unit in the sample.

Parameter - a numerical value that would be calculated using all of the values of the units in the population.

Statistic - a numerical value that is calculated using all of the values of the units in a sample.
Notation: 
\[ N \] - size of the population. 
\[ n \] - size of the sample.

\[ N = 13 \]
\[ n = 3 \]

Example: 60% of students own computers. 5290 enrolled. 1%
(a) \[ N = 5290 \] \[ n = 53 \]
(b) What is the parameter? (proposed) 60%
(c) What is the variable? Percentage of students own computer.
(d) What statistic will I obtain? 0% out of 53.

Important:
- Parameter is a fixed quantity.
- Statistic will vary from sample to sample.
2.4. Good Data?

Our goal - produce accurate estimates.
Our enemy - Bias

Bias: a systematic prejudice in one did

Convenience sample
Volunteer sample

Probability sampling method -

A sample method that gives each unit in the population a known nonzero chance of being selected. (good method)

Selection bias - systematic tendency on the part of the sampling procedure to exclude or include a certain type of unit.

Non-response bias
Response bias

Ex: Whether or not own computer?
only full-time students on my list.

Selection bias.
2.5 Simple Random Sampling

Simple random sample of size \( n \): a sample of \( n \) units selected in such a way that every possible sample of the given size \( n \) has the same chance of being selected as any other sample of size \( n \).

How to obtain a random sample:

1. Number the elements in the population.
2. Generate \( n \) random numbers using either a random number table (p. 100) or the calculator random number feature.

Ex: choose the first 10 students to be in the sample using the random number table. \( N = 60, n = 10. \)

1. Assign labels 01 to 60 for each student.
2. Read off random labels.

Row 2: 21468, 46573, 25595, 85393, 30995, 89198, 27982, 53162, 93965, 34095, 99, 55, 98, 27, 98, 39, 53, 68, 58, 73, 25, 59, 58, 93, 20...
2.6 Stratified random sampling
mutually exclusive

2.7 Systematic Sampling
1-in-k systematic sample

2.8 Cluster Sampling
divided into clusters
choose clusters

Chap 3. Observational Study & Experiments

3.3 The Language of Studies

Types of Studies:
1. Designed experiment: actively imposing treatments.
2. Observation study: simply observes.

Survey: observational study.

Types of Variables:
1. Explanatory variable (factor). A variable that is thought to explain the changes in the response variable (independent...
2. Response variable \( Y \) measures the outcome of the study and depends on the explanatory variable (dependent).

Explanatory \( \rightarrow \) Response

\( V_1, V_2 \):

\( V_1 \rightarrow V_2 \)
\( V_2 \rightarrow V_1 \).

**Ex:** \( V_1 \): weight of a package.
\( V_2 \): postage rate.

\( V_1 \rightarrow V_2 \) --- \( V_2 \) -- response

Levels of a factor are the possible values of the explanatory variable.

A treatment is a specific combination of the levels of the explanatory variables.

**Ex:** curvature of copper plates depends on temperature & copper content.

\( T \): 50\(^\circ\)C, 75\(^\circ\)C, 100\(^\circ\)C, 125\(^\circ\)C.

\( P \): 40\%, 60\%, 80\%.

Two observations for each treatment.
1. Response variable: amount of curr
2. Explanatory variables: 
   - Temperature: 50°C, 75°C, 100°C, 125°C
   - Copper content percentage: 40%, 60%, 80%
3. $3 \times 4 = 12$ treatment combinations
4. $2 \times 12 = 24$ observations
5. Experiment

Compounding variables - a variable whose effect on the response variable can't be separated from the effect of the explanatory variable.

3.4. Understanding Observational Study

- Retrospective study (past event) ←
- Prospective study (ongoing or future) →

Disadvantage: simply observe

- Can not control the explanatory
- Can not control compounding variable

Experiments:

Pros: no compounding variables
Cons: cost, feasibility, ethics
3.5 Understanding experiments. Design layout table.

Ex.

<table>
<thead>
<tr>
<th>Copper content</th>
<th>50°C</th>
<th>75°C</th>
<th>100°C</th>
<th>20°C</th>
</tr>
</thead>
<tbody>
<tr>
<td>40%</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>60%</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>80%</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Basic principles for design of experiment:

1. Control: two groups (compounding effects)
2. Blinding: Placebo
3. Randomization
4. Replication: at least 2 units for each treatment