Chapter 5

Producing Data
Planning and Conducting a Study

- Understand the nature of the problem.
  - We must know the goal of the research and what questions we hope to answer.

- Deciding what to measure and how to measure it.
  - We must decide what information is needed to answer the questions of interest

- Data Collection
  - We first must decide if there is existing data or whether new data must be collected.
Planning and Conducting a Study

- Data summarization and preliminary analysis.
  - Summarization the data graphically and numerically.

- Formal data analysis
  - Select and apply the appropriate statistical methods

- Interpretation of results
  - What conclusions can be drawn from the analysis?
  - How do the results of the analysis inform us about the stated research problem or question?
  - How can our result guide future research?
Lesson 5-1, Part 1

Designing Samples
Random Samples
Gathering Data

- **Observational Study**
  - Observes individuals and measures variables of interest but does not attempt to influence the response

- **Experiment**
  - Deliberately imposes some treatment on individuals in order to observe their response.
Observational Studies

- To learn characteristics of a population.
- To determine whether there is an association between two or more variables.
- After-the-fact study because the value of the variable of interest has already been established.
Experiment

- Identifies certain causes and effect relationships among the variables in the study.
- Are used whenever control of certain variables are desired.
Population and Sample

- **Population**
  - The entire group of individuals (not necessarily people) we want information about.

- **Sample**
  - Part of the population in the study.
Collecting Data

- **Sampling**
  - Studies a part in order to gain information about the whole

- **Census**
  - Attempts to contact every individual in the entire population.
The method we use to select the sample is called the **sample design**.

The design of the sample is very important.

- If the design is poor, the sample will not accurately represent the population.
Types of Sample Designs

- **Voluntary Response Sample**
  - People choose themselves to be in the sample by responding to a general appeal.
  
  *Example*: We post an advertisement in the newspaper asking SHS students to respond.

  - *Problem*: People with strong opinions (often strong negative opinions) tend to reply, so they are overrepresented.
Types of Sample Designs

- **Convenience Sample**
  - Individuals who are easiest to reach are chosen for the sample.
  
  *Example:* We use students in this class as our sample.
  
  *Problem:* This group may not be diverse enough to accurately represent all students at SHS.
Both Voluntary Response Samples and Convenience Samples result in a sample that is not representative of the population.

These are biased samples because they favor certain outcomes.

Random selection eliminates bias from sample choice.
Identify the population as exactly as possible. That is, say what kind of individuals the population consists of and say exactly which individuals fall in the population. If the information given is not complete, complete the description of the population in a reasonable way.

A) Each week, the Gallup Poll questions a sample of about 1500 adult U.S. residents to determine national opinion on a wide variety of issues.

An individual is a person; the population is all adult US residents.
B) The 2000 census tried to gather basic information from every household in the United States. But a long form requesting much additional information was to sample of about 17% of households.

An individual is a household; the population is all adult US households.
C) A machinery manufacturer purchases voltage regulators from a supplier. There are reports that variation in the output voltage of the regulators is affecting the performance of the finish products. To assess the quality of supplier’s production, the manufacture sends a sample of 5 regulators from the last shipment to a laboratory for study.

An individual is a voltage regulator; the population is all regulators in the last shipment.
The National Halothane Study was a major investigation of the safety of anesthetics used in surgery. Records of over 850,000 operations performed in 34 major hospitals showed the following death rates for four common anesthetics.

<table>
<thead>
<tr>
<th>Anesthetics</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death Rate</td>
<td>1.7%</td>
<td>1.7%</td>
<td>3.4%</td>
<td>1.9%</td>
</tr>
</tbody>
</table>

There is a clear association between anesthetic used and the death rate of patients. Anesthetic C appears to be dangerous.
A) Explain why we call the National Halothane Study an observational study rather than experiment, even though it compared the results of using different anesthetics in actual surgery.

The data was collected after the anesthetics was administered.
B) When the study looked at other variables that are confounded with a doctor’s choice of anesthetic, it found that Anesthetic C was not causing extra deaths. Suggest variables that are mixed up with what anesthetic a patient receives.

Type of surgery, patient allergy to certain anesthetic, how healthy the patient was before the surgery.
Simple Random Sample (SRS)

- Individuals are selected so that all possible combinations of individuals are equally likely to be in the sample.
  - *Example:* Generate a list of students ID numbers for all students SHS; then randomly select student ID numbers and choose those students for the sample.
Choosing a Random Sample

- Table of Random Digits
  - Table B – back of book

- TI Calculator
  - Math/PRB/5:randInt(lower, upper, [numtrials])
A firm wants to understand the attitudes of its minority managers toward its system for assessing management performance. Below is a list of all the firm’s managers who are members of minority groups. Use the Table B at line 139 to choose 6 to be interviewed in detail about the performance appraisal system.

Agrawal  Fernandez  Liao  Santiago
Anderson  Fleming  Mourning  Shen
Baxter  Gates  Naber  Vega
Bonds  Goel  Peters  Wang
Bowman  Gomez  Pliego
Castillo  Hernandez  Puri
Cross  Huang  Richards
Dewald  Kim  Rodriguez
Lesson 5-1, Part 2

Designing Samples

Other Random Samples
Systematic Random Sample

- The first individual is chosen at random, then system or rule is used to choose all other individuals.
  - *Example:* Obtain an alphabetized list of students at SHS. Choose every 6th person on the list.
Stratified Random Sample

- Divide the population into groups similar individuals; choose a SRS in each group to form the full sample
  - *Example:* Divide all of the students at SHS into four groups: freshmen, sophomores, juniors and seniors; then choose a SRS from each grade level
Select several groups; within each group, select a subgroup; within each subgroup select individuals for the sample.

*Example:* Select several departments within the school (Math, English, Art) Within each of those departments, select several teachers. Choose several student within each class.
Cluster Samples

- Select several groups; within each group, select several subgroups; with each subgroup select ALL individuals for the sample.
  - *Example:* Select several departments within school (Math, English, Art). Within each of those departments, select several teachers. Choose ALL students in each class.
Accountants often use stratified samples during audits to verify a company’s records of such things as accounts receivable. The stratification is based on the dollar amount of the item and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 100 are amounts over $50,000; 500 are amounts between $1000 and $50,000; and the remaining 4400 are amounts under $1000. Using these groups as strata, you decide to verify all of the largest accounts and to sample 5% of the midsize accounts and 1% of the small accounts. How would you label the two strata from which you will sample? Use Table B, starting at line 115, to select only the first 5 accounts from each strata.
Assign 500 midsize accounts: 001 to 500 and assign

Line 115
61041  77684  94322  24709  73698  14526  31893  32592
417
494
322
247
097
Assign 4400 small accounts: 0001 to 4400

Line 115
61041 77684 94322 24709 73698 14526 31893 32592
14459 26056 31424 80371 65103 62253 50490 61181
3698
1452
2605
2480
3716
Potential Problems

- **Undercoverage**
  - Some groups are left out of the process of choosing the sample
  - *Example:* Students in SCTI, early release, on suspension, or absent may be left out of the sample.

- **Nonresponse**
  - An individual chosen for the sample cannot be contact or refuses to cooperate.
  - *Example:* A student chosen for the sample may refuse to divulge information or may be absent.

- **Response Bias**
  - The behavior of the individual or interviewer may influence the accuracy of the response.
  - *Example:* Students may lie about drug or alcohol use.
Wording of Questions

Confusing or leading questions influence response; poorly worded questions will not yield accurate responses.

- *Example 1*: “In a recent study, students in an Algebra 1 course were given a 25 question basic skills test. On average, students used a graphing calculator to answer 21 out of 25 questions. Do you think graphing calculators are overused?

- *Example 2*: By using a graphing calculator, students in Algebra 1 course were able to make visual connection between equations and their graphs, reinforcing difficult concepts. Do you think graphing calculators are overused?
Inferences About the Population

- Sample results are only estimates about the population.
- Results from samples survey usually come with a margin of error.
- Large random samples gives more accurate results than smaller samples.
A newspaper article about an opinion poll says that “43% of Americans approve of the president’s overall job performance.” Toward the end of the article, your read: “The poll is based on telephone interviews with 1210 adults from around the United States, excluding Alaska and Hawaii.” What variable did this poll measure? What population do you think the newspaper wants information about? What was the sample? Are there any sources of bias in the sampling method used?

Variable: Approval of president’s performance

Population: Adults citizens of the US or registered voters

Sample: 1210 adults interviewed

Possible sources of bias: Only adults with phones were contacted. Alaska and Hawaii were omitted.
Comment on each of the following as a potential sample survey question. Is the question clear? Is it slanted toward a desired response?

A). “Some cell phone users have developed brain cancer. Should all cell phones come with a warning label explaining the danger of using cell phones?”

The wording is clear. The question is somewhat slanted in favor of warning labels.
B). “Do you agree that a national system of health insurance should be favored because it would provide health insurance for everyone and would reduce administrative costs?”

The wording is clear, but it is clearly slanted in favor of national insurance by asserting it would reduce administrative costs.
C). “In view of escalating environmental degradation and incipient resources depletion, would you favor economic incentives for recycling of resource-intensive consumer goods?”

The question could be clearer by using simple language. It is slanted in favor of incentives by starting out discussing environmental degradation.
Lesson 5-2, Part 1

Designing Experiments
Designing Experiments

- If we want to observe individuals and record data without intervention, we conduct an **observational study**.
- If we want to examine cause and effect relationship we conduct an **experiment**.
The individuals on which the experiment is done are called **experimental units**.

If the units are people, they are called **subjects**.

The experimental condition we apply to the units is called the **treatment**.

The explanatory variables (causing a change in the other variables) are called **factors**.

The factors may be applied in different **levels**.
Identify the experimental unit or subjects, the factors, the treatments and the response variable.

A manufacturer of food products uses package liners that are sealed at the top by applying heated jaws after the package is filled. The customers peels the sealed pieces apart to open the package. What effect does the temperature of the jaws have on the force required to peel the liner? To answer this question, the engineers prepare 20 pairs of pieces of package liner. They sealed five pairs at each of 250°F, 275°F, 300°F, and 325°F. Then they measure the strength needed to peel each seal.

The liners are the experimental units. The heat applied to the liners is the factor, the levels are 250°F, 275°F, 300°F, and 325°F. The force required to open the factor is the response variable.
Control Groups

- When designing an experiment we want to minimize the effect of lurking variables so that our results are not biased.
- Because we may not be able to identify and eliminate all lurking variables, it is essential that we use a control group.
- The control group gets a fake treatment to counter the placebo effect and/or any other lurking variables present.
- Having a control group allows us to compare the results of the treatments.
Experimental Design

- Choose treatments
  - Identify factors and levels
  - Control group

- Assign the experimental units to the treatments
  - Matching
    - Place similar units in each treatment group
  - Randomization
    - Randomly assign units to each treatment group
Experimental Design

- Remember, if we want to examine a cause and effect relationship, we conduct an **experiment**.

- If an experiment is well-designed, a strong association in the data **does imply causation**, since any possible lurking variables are controlled.
Principles of Experimental Design

- Control the effects of lurking variables by comparing several treatments (include a control group if possible)
- Use randomization to assign subject/units to treatments.
- Replicate the experiment on many subjects/units to reduce chance of variation in the results
- An effect is called **statistically significant** if it is too great to be caused by simply chance.
Comparison Experiments

Random assignment

Group 1
15 rats

Treatment 1
New diet

Compare weight gain

Group 2
15 rats

Treatment 2
Standard diet
A). Use a diagram to describe a completely randomized experimental design for the package liner experiment of Exercise 5.32.

B). Use Table B, starting at line 120, to do the randomization required by your design.
Random Allocation

Group 1
5 Pairs

Group 2
5 Pairs

Group 3
5 Pairs

Group 4
5 Pairs

Treatment 1
250° F

Treatment 2
275° F

Treatment 3
300° F

Treatment 4
325° F

Measure of force required to open
Number the liners from 01 to 20, then take Group 1 to be 16, 04, 19, 07, and 10; Group 2 is 13, 15, 05, 09 and 08, Group 3 is 18, 03, 01, 06, and 11. The others are in Group 4.
Cautions about Experiments

- Even a well-designed experiment can contain hidden bias, so it is extremely important to handle the subjects/units in exactly the same way.

- One way to avoid hidden bias is to conduct a double-blinded experiment.
  - A double-blinded experiment, neither the subjects nor the people who have contact with them know which treatment a subject has received.
Lesson 5-2, Part 2

Experimental Design
Types of Experimental Design

- In a **completely randomized design**, all subjects are randomly assigned to treatment groups.
- In a **block design**, subjects are first split into groups called **blocks**.
  - Subjects within each block have some common characteristic (for example: gender, age, education, ethnicity, etc.)
  - Then, within each block subjects are randomly assigned to treatment groups.
- **Matched pair design**, there are only two treatments. In each block, there is either:
  - A single subject receiving both treatments, or
  - A pair of subjects, each receiving a different treatment
Fizz Laboratories, a pharmaceutical company, has developed a new pain-relief medication. Sixty patients suffering from arthritis and needing pain relief are available. Each patient will be treated and asked an hour later, “About what percentage of pain relief did you experience?

A). Why should Fizz not simply administer the new drug and record the patients responses?

Their responses will not be useful, because there will be nothing to compare them to.
B). Outline the design of an experiment to compare the drug’s effectiveness with that of aspirin and of a placebo.

Randomly assign 20 patients to each of the 3 groups: Group 1, the placebo; Group 2, the aspirin; Group 3, new treatment.

After treating the patients, ask them how much pain relief they feel, and then compare the average pain relief experienced by each group.
C). Should the patients be told which drug they are receiving? How would this knowledge probably affect their reactions?

The subjects should not be told what drug they are getting – a patient told that she is receiving the placebo will probably not expect any pain relief.
D). If patients are not told which treatment they are receiving, the experiment is single blind. Should this experiment be double blind also? Explain.

Yes, the interviews may subtly influence the subjects into giving responses that support the new treatment.
People who eat lots of fruits and vegetables have lower rates of colon cancer than those who eat little of these foods. Fruits and vegetables are rich in “antioxidants” such as vitamins A, B, and E. Will taking antioxidants help prevent colon cancer? A clinical trial studied 864 people who were at risk of colon cancer. The subjects were divided into four groups: daily beta carotene, daily vitamins C and E, all three vitamins every day, and daily placebo. After four years, the researchers were surprised to find no significant difference in colon cancer among the groups.

A.) What are the explanatory and response variables in this experiment?

Explanatory variable is the vitamins taken daily; the response variable is whether colon cancer develops.
B.) Outline the design experiment. Use your judgment in choosing the group sizes.

Random assignments

- Group 1: 216 Subjects → Treatment 1: Beta carotene
- Group 2: 216 Subjects → Treatment 2: Vitamins C & E
- Group 3: 216 Subjects → Treatment 3: All three
- Group 4: 216 Subjects → Treatment 4: Placebo

Observe the colon cancer
C.) Assign labels to the 864 subjects and use Table B, starting at line 118 to choose the first 5 subjects for the beta carotene group.

Using labels 001 through 864: 731, 253, 304, 470, and 296.
D.) The study was double-blinded. What does this mean?

Both the subject and those who work with the subjects do not know who is getting what treatment.

E). What does “no significance difference” mean in describing the outcome of the study?

The observed differences were no more than what might reasonably occur by chance even if there is no effect due to treatment.
F). Suggest some lurking variables that could explain why people who eat lots of fruits and vegetables have lower rates of colon cancer. The experiment suggests that these variables, rather than the antioxidants may be responsible for the observed benefits of fruits and vegetables.

Fruits and vegetables contain fiber; this could account for the benefits of those foods. Also people who eat lots of fruits and vegetables may have healthier diets overall.
Lesson 5-3, Part 1

Simulating Experiments
The imitation of chance behavior, based on a model that accurately reflects the experiment under consideration is called a simulation.
A basketball player makes 70% of her free throws in a long seasons. In a tournament game she shoots 5 free throws late in the game and miss 3 of them. The fans think that she is nervous, but the misses may simply be chance. You will shed some light by estimating the probability.

A). Describe how to simulate a single shot if the probability of making each shot is 0.7. Then describe how to simulate 5 independent shots.
Assign 1 – 7 “hit” and 8 -10 “misses” or 0 – 6 “hits” and 7 – 9 “misses”.

B). Simulate 50 repetitions of the 5 shots and record the number missed on each repetition. Use Table B starting at line 125. What is the approximate likelihood that the player will miss 3 or more shots.

96746  12149  37823  71868  18442  35119  62103  39244
3 hits  4 hits  3 hits  2 hits  4 hits  3 hits  5 hits  4 hits

Continued this until you have 50
The frequency counts are

<table>
<thead>
<tr>
<th>$X$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>18</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>

The relative frequency of 3 or more shots is $\frac{39}{50} = 0.78$
Same problem using the TI

```
randInt(1,10,5)
(8 3 8 8 5)
(5 9 5 7 6)
(6 8 5 4 4)
```
Amarillo Slim is a cardshark who likes to play the following game. Draw 2 cards from the deck of 52 cards. If at least one of the cards is a heart, then you win $1. If neither card is heart, then you lose $1.

A). Describe a correspondence between random numbers and possible outcomes in this game.

Let 1 – 13 “hearts” and 14 – 52 “another suit”
B). Simulate playing the game for 25 rounds. Do you think this is a “fair” game? That is, do both you and Slim have equal chance of winning?

Set the calculator 123→rand

123→rand
randInt(1,52,2)

No