Chapter 5: Statistical Reasoning

Chapter 5 will familiarize you with the fundamentals of statistics, focusing on many common real-world uses. There are several main ideas that appear throughout the chapter, such as what a statistical study is, how information is gained from statistical studies, and how to represent the results of a study. However, statistics are often used to mislead. The student should learn what to look out for that may be misleading, including sample bias, graphs, charts, and causality.

Unit 5A: Fundamentals of Statistics

This section deals with the structure of statistical sampling. Most importantly, it discusses how the sampling method impacts the results of the study. Several main ideas are defined and explored.

How Statistics Work

Statistical studies seek to make conclusions about a group (or a population). The studies usually focus on a few population paramaters that the researchers wish to explore. However, it is often impossible to look at every member of the group in question. Therefore, a sample is chosen and data is taken from this sample. Then the statistics are analyzed in attempts to generalize the result from that sample to the entire population.

The Process of a Statistical Study

Statistical studies, as mentioned above, follow this basic outline:

1. State the goals of your study precisely. This includes identifying the population you wish to study, and exactly what you wish to learn about it.

2. Choose a representative sample from the population.

3. Collect raw data from the sample and summarize these by finding sample statistics of interest.

4. Use the sample statistics to infer the population paramaters.

5. Draw conclusions: Determine what you learned and whether you achieved your goal.

Ok, but how does all of this work? The first step is the easiest. You basically decide what you want to study. The second step is a bit more tricky. This brings us to our next section.

Choosing a Sample

This can be one of the most difficult parts to get right. You need a sample of the population that will likely be an accurate portrayal of the total population. There are several methods for doing this. They all have advantages and disadvantages. One of the main points here is that if your sample is not an accurate portrayal of the population, then your results can be flat out wrong! Furthermore, if you do get a good sample (whatever that means), it is still just a sample and you can not be sure that it accurately represents the population. This is one of the major drawbacks of statistical analysis: the statistics are just educated guesses.

Sample Type	Example	Advantages	Disadvantages
Simple Random Sample	You want to know the opinions of people in your area about the upcoming election. You have a computer pick 500 random addresses from the area.	It is likely that this sample is representative of the population	It can be difficult to get them randomly.
Systematic Sample	You want to know about the ages of people who go to the mall on Saturdays.p You have ask every 50th person who walks through the front door of the mall.	It is easier to select the sample.	It may not be representative depending on what door you are at, what time you pick the sample, etc.

Sample Type	Example	Advantages	Disadvantages
Convenienc Sample	You want to know what people think about Rock and Roll music. You ask people you run into at the bar, in class, and on the street what they think.	It is easy to pick the sample.	It is not likely that this is a representative sample since the people you ask share something in common with you (like the bar you go to, the class you are in, age, etc.).
Stratified Sample	You want to know which professor gives the highest grades in the math department. You get a simple random sample of students who have taken a class with each professor in question. You ask them what their grade was.	You get information about specific subgroups of the population, and you used a simple random sample which is likely to give you an accurate representation of each group.	Stratifying some studies may introduce bias.

Watching Out for Bias

Statistical bias is one of the cardinal sins of statistics. However, it may not be obvious that the study has bias built in. For example, lets say you want to determine whether or not women are better at english than men. You select students your english class and your math class as the sample. However, you are an English major and are taking Creative Writing for your english class and math 55 for your math class. By chance, your Creative writing is predominantly female and your math class is predominantly male. Then, your study is bias since you have more men in math 55 and more women in the upper level english class.

Systematic sampling can bring about sample bias. In particular, if you use systematic sampling on something that is not random, it will likely introduce bias into your study. For example, say you want to study the gender make-up of Temple University's freshman class. You decide to sample the first 5 rooms on every other floor of one of Johnson Hall. However, you did not realize that most of the floors are single gender floors. To your surprise, you find that 85 percent of the sample population is female! This is because most of the odd numbered floors are female only floors, and only two are coed.

Stratified sampling can also introduce bias. Consider districting in Texas for Senate and House elections. This is stratifying the population of Texas by districts. Just last year, the Republicans were pushing to redistrict as to give themselves the advantage since the majority in each district is what matters. Many of the Democrats on the Texas legislature absconded to Oklahoma so no vote could be held. This is common in politics since it is easier to manipulate strata rather than the entire population. In other words, the Republicans were attempting to take advantage of the stratification of the population to have a better chance of getting a Republican into office. They were trying to introduce bias into the stratification by redefining each district.

Even the nature of questions on a survey can introduce bias. For example, if the questions are personal and someone knows which form *you* fill out, you may well lie to avoid embarrassment.

There are many ways for bias to sneak into a study. Many of these stem from the sampling technique. So, whenever you see the results of a survey or study, ask yourslef if it was likely to be biased or if the sample is likely to be representative of the population.

Types of Statistical Study

There are two major types of statistical studies: observational studies an experiments. The main difference is that in an observational study, the people conducting the study are not acting on the subjects whereas in an experiment they are. Experiments are largely used in scientific research such as medicine where a treatment is administered. Observational studies incorporate studies such as polls, surveys and examining behavior.

The Placebo Effect and Blinding

In experiments where the study requires action by those doing the study, there are several elements that may impact the results. The first of these is the placebo effect. This is when people have a specific reaction due to the belief that they are being given a treatment. For example, if you want to know if a pill relieves cold symptoms, you may give half of the subjects the miracle pill, and the other half a sugar pill (the placebo). The placebo will probably make some of the subjects feel better, and some may have "side effects" of the placebo pill. This tells the researchers how many of those getting the miracle pill are likely to feel better just because they are getting some kind of treatement. It also tells them how many will have some side effects just because they are getting some kind of treatment.

For these reasons, it is important that the subjects do not know whether they are part of the control group getting the placebo or the test group getting the miracle pill.

Another potential problem is if the people who design the study are also taking the data. Say they want to prove that their pill really works, and they know which participants are the control and which are the treatment group. When they ask how the subjects feel, they may word their questions as to encourage the subjects to respond how they want. For example, the administrator of the test may ask: "How much better do you feel now than you did 5 days ago", instead of: "How do you feel now compared to 5 days ago." This would introduce bias since the desired response is encouraged. Therefore, it is a good idea for those taking the data to be unaware of whether the person is in the control group or the treatment group. This is the motivation for double blind experiments.

Case-Control Studies

Case-control studies are observational studies where it is unreasonable to conduct an experiment. For example, if it is unethical or impractical to conduct the experiment, it is better to observe groups that do and do not have the given characteristic. One example that comes to mind is the drug Valtrex for genetal herpes. The comercials always say that it may reduce the risk of transmitting genetal herpes if taken however often you are supposed to take it. It would not be ethical for them to give Valtrex to one group and a placebo to another and encourage them to engage in unsafe sex. Therefore, they provide the medication to whoever wants it (or whoever bought it, I guess). Then, they track those people and their partners. They also track people who do not take any medication and their partners. Next, they look at how many people who had no medication transmitted it to their partners and the same for those who took Valtrex. This way, they get information from people who voulentarily engage in unsafe sex when infected with genital herpes. From this, they can make conclusions as to whether it is effective or not at lowering the risk of transmission of herpes.

Surveys and Opinion Polls

Surveys and opinion polls are the most common types of observational studies. Unfortunately, they are also the most frequently misrepresented or poorly conducted. Any survey or poll should come with a margin of error. What this means is that the sample statistic they got is likely between sample statistic − margin of error and sample statistic + margin of error. This range is the confidence interval. Typically, it means they are 95 % confident that the actual value for the population (population paramater) is somewhere in this interval. Always be cautious if a survey or poll does not give a margin of error or a confidence interval. Also, be very careful when it is given but ignored by the commentary on it. For example, in the polls before an election, it is not uncommon for it to be 47% for candidate A and 53% for candidate B with a margin of error 5%. This means that they CAN NOT SAY WITH CONFIDENCE THAT ONE CANDIDATE HAS THE LEAD since they are only confident that candidate A has between 42% and 52% support, which means candidate A MAY BE IN THE LEAD! Unfortunately, news casters like to say that candidate B is ahead, contrary to the statistics.

In fact, in the 2000 presidential elections, just about all of the major news networks declared Gore to be the winner when the statistics were not confident that he won. They were all just anxious to be one of the first to announce the winner. Unfortunately for them, they then had to race to be the first to make a retraction! Similarly, Fox went out on a limb in the 2004 election and delcared Bush the winner before statistics allowed them to. The difference this time was that no other network jumped on the band wagon and they did not report Bush as the victor for several more hours, when the statistics were sufficient to make this declaration.

5B: Should You Believe a Statistical Study

A word of advise is to never take a statistical study at face value. There are many aspects that can cause the study to be flawed and give inaccurate results. Furthermore, they can be difficult to identify since most reports on the news, etc. do not give enough information to determine whether or not it can be believed! There are eight guidelines for evaluating a statistical study:

1. Identify the goal of the study, the population considered and the type of study.

2. Consider the source, particularly with regard to whether the researchers may be biased.

3. Look for bias that may prevent the sample from being representative of the population.

4. Look for problems defining or measuring the variables of interest, which can make it difficult to interpret results.

5. Watch out for confounding fariables that can invalidate the conclusion of a study.

6. Consider the setting and the wording of questions in any survey, looking for anything that might tend to produce inaccurate or dishonest responses.

7. Check that results are presented fairly in graphs and concluding statements, since both researchers and the media often create misleading graphics or jump to conclusions that the results do not support.

8. Stand back and consider the conclusions. Did the study achieve its goals? Do the conclusions make sense? Do the results have any practical significance?

5C: Statistical Tables and Graphs

It is common for statistical studies to come equipped with graphs and quick summaries of the data. Both of these are ways of presenting data in a compact way for easy reference. Graphs and charts are common since they provied a pretty picture to represent the data. However, beware of charts and graphs since they can be very misleading. The most common ways to mislead with graphs and charts is by manipulating the axes.

The book has many examples and illustrations, so I will refer you to that for more details.

5E: Correlation and Causality

The purpose of many statistical studies is to determine whether one factor causes another. A correlation is an indicator that there is some relationship between two variables. Therefore, determining whether or not a correlation exists between two variables helps determine whether is it reasonable or not to say one may cause the other. However, be forewarned: CORRELATION DOES NOT IMPLY CAUSATION!!!

Seeking Correlation

Since a correlation is a relationship between two variables, it is common to use a scatter diagram (or scatter plot) to compare the two variables. If the points seem to look like a line, there is some kind of correlation between the variables. If the data points look like a blob (i.e. there is no apparent relationship between the two variables), then there is no correlation between the variables. If both variables tend to increase (or decrease) together, it is said to be a positive correlation. If one variable increases when the other decreases, it is said to be a negative correlation. However, the existance of a correlation is not very telling about how well one variable predicts the other. Therefore, we have a concept of the strength of a correlation. The more closely the points follow a general trend, the stronger the correlation is. That is, if all of the points lie on a single strait line, it is a perfect correlation. If the points are randomly scattered, there is no correlation. The closer the points are to a strait line, the stronger the correlation.

However, a strong correlation does not imply that one variable causes the other. There are several explainations for correlated data. First of all, it may just be sheer coincidence. Second, there may be another factor that causes the response in the two variables being examined. Or, it may be the case that one actually causes the other, or at least one variable influences the other.

Establishing Causality

There are six guidelines for establishing causality. However, just because all seven guidelines are met, and there is strong correlation, you can only say it is likely that one causes the other.

1. Look for situations in which the effect is correlated wtih the suspected cause even when other factors vary.

2. Among groups that differ only in the presence or absence of the suspected cause, check that the effect is similarly present or absent.

3. Look for evidence that larger amounts of the suspected cause lead to larger amounts of the effect.

4. If the effect my be produced by other potential causes, make sure that the effect still remains after accounting for these other potential causes.

5. If possible, test the suspected cause with an experiment. If the experiment can not be performed withi humans for ethical reasons, consider doing the experiment with animals, cell cultures or computer models.

6. Try to determine the physical mechanism by which the suspected cause produces the effect.

Confidence in Causality

If a correlation is present, all that can be said is that it is possible that there is causailty between the variables. If several of the guidelines for establishing causality have been met successfully, it is probable that the correlation involves a cause. If many of the guidelines for establishing causality are met and a physical model that successfully explains how one thing causes the other and it is unreasonable to doubt the causality, then it is just very likely that the correlation is based on a cause. You can NOT say for CERTAIN that the correlation is based on a cause.

Chapter 5 Definitions

Statistics: the science of collecting, organizing and interpreting data.

Statistics: the data that describe or summarize something.

Population: the complete set of people or things being studied.

Sample: the subset of the population from which the data are obtained.

Population Parameters: specific characteristics of the population that a statistical study is desined to explore.

Sample Statistics: numbers or observations that summarize the raw data.

Representative Sample: a sample in which the relevant characteristics (i.e. the population parameters) of the sample members match those of the population.

Simple Random Sample: the sample of items is chosen in such a way that every sample of a given size has an equal chance of being selected.

Systematic Sample: the sample is chosen with a simple system, such as every 10th member of the population.

Convenience Sample: the sample is chosen at the convenience of the experimenters.

Stratified Sample: the sample is chosen by selecting a simple random sample from predetermined subgroups (or *strata*).

Bias: a tendency to favor certain results due to design or implimentation.

Observational Study: a study in which the experimenters observe or measure characteristics of the sample members but do not attempt to influence or modify these characteristics.

Experiment: a study in which researchers apply treatment to some or all of the sample members and then look to see whether the treatment has any effects.

Treatment Group: the members of the sample group who receive the treatment being tested.

Control Group: the members of the sample group who do not receive the treatment being tested.

Placebo: a treatment lacking the active ingredients being tested, but is identical in appearance to the active treatment.

Placebo Effect: the situation in which patients improve simply because they believe they are receiving a useful treatment.

Single Blind Experiment: an experiment in which the participants do not know whether they are in the treatment group or the control group, but the experimenters do.

Double Blind Experiment: an experiment in which neither the participants nor the experimenters know wich group the participants belong to (the treatment or control group).

Case-control Study: an obserational study that resembles an experiment because the sample naturally divides into groups. The participants who engage in the behavior under study form the **cases**, which make them like a treatement group in an experiment. The participants who do not engage in the behavior are the **controls**, making them like a control group in an experiment.

Margin of Error: describes how closely the data are expected to represent the actual population.

Confidence Interval: the amount that the sample statistic may vary to be able to say with confidence the sample statistic is accruate.

Selection Bias: the sample is selected in a way that tends to be unrepresentative of the population.

Participation Bias: the choice to participate (usually in a survey or a poll) typically has a larger percentage of people who feel very strongly about the issue.

Variable: any item or quantity that can vary or take on different values. The *variables of interest* in a statistical study are the items or quantities that the study seeks to measure.

Frequency Table: typically a two-column table with the first column containint the categories of data, and the second column containing the frequencey of each category (i.e. the number of times the category appears in the data set).

Relative Frequency: frequency expressed as a fraction or percentage of the total.

Cumulative Frequency: the sum of the frequencies for the given category and all previous categories.

Qualitative Data: nonnumerical categories such as hair color or gender.

Quantitative Data: numerical categories such as counts or measurements.

Histogram: a bar graph for quantitative data categories. The bars have a natural order and the bar widths have specific meaning.

Line Chart: the data value for each category is represented by a dot, and the dots are connected with strait lines. For each dot, the horizontal position is the center of the bin it represents and the vertical position is the data value for the bin.

Time-series Diagram: a histogram or line chart where the horizontal axis represents time.

Correlation: an indicator as to whether two variables are related. This is a measure how how well one variable indicates what the value of another variable will be. A correlation exists when higher values of one variable consistently go with higher values of another, or when higher values of one variable consistently go with lower values of another variable.

Scatter Diagram: a graph in which each point represents the values of two variables.

Chapter 5 Homework Problems

5A problems starting on pp. 313: 1-11, 14-23, 28, 29-32, 35-38, 41-44, 48-54, 56-58.
5B problems starting on pp. 325: 1-24, 26-30, 33-38, 41-43.
5C problems starting on pp. 339: 1-27, 29, 31-32, 36, 40-41, 46.
5E problems starting on pp. 371: 1-25, 29-30, 32-35, 38, 40.