Chapter 1 - Overview

1.1 - What is Statistics

**Statistics** is defined as the science of collecting, organizing, summarizing, and analyzing data to draw conclusions and answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

As you will learn during this course, in statistics we are not always 100% certain about our results, we will say things like we are providing results to "95% certainty."

This class will provide an introduction to statistics and include examples and exercises from many different fields.

A statistical study or analysis generally begins with a question. For example:

- What percentage of men wear a tie to work?
- What is the average age of a Santa Monica College student?
- Does a certain medication provide relief for headaches?
- Is Friday the 13th unlucky?

If you wanted to determine the average age of a Santa Monica College student, how would you start?

There are really two possibilities, you could ask every SMC student their age or you could ask a smaller subset of the students and use that smaller subset to draw a conclusion about all SMC students. This second possibility is really what this course is about.

Why would you not use the entire SMC student body to answer this question?

Definitions:

1) population

2) sample

3) parameter

4) statistic
5) census

If you use a sample to draw a conclusion about the population, the results are not always going to be 100% accurate. For example, if we ask 100 SMC students how old they are and find the average age of this group, we would not expect the results to be exactly equal to the value we would get if the entire population of SMC students were used. Hopefully using good statistical techniques our answer would be close enough to the actual average age to give us useful results. **A statistic is used to estimate a population parameter, the value of the statistic can change with the choice of sample, the value of the population parameter being estimated never changes.**

Much of the first part of this course will focus on **descriptive statistics**. Descriptive Statistics is the organizing and summarizing of data, often using tables, charts and graphs.

The second part of the course will focus on **inferential statistics** or taking the results from a sample and extending it to the population and measuring the reliability of the results.

For example:

A sample of 100 SMC students had a mean age of 20.8 years, based on this I am 95% certain that the average age of all SMC students is between 19.7 and 21.9 year old.

**note that we are estimating the population parameter of the average age of all SMC students using the interval and giving the reader information about the confidence we have in the result.**

Definition: Each person or object that participates in a study is called an **individual or element**. The above study had 100 individuals (or elements).

Once the questions for a study are defined, a list of information or characteristics we want from the individuals in the study must be defined. Each of these characteristics is called a **variable**.

Example:

Suppose we ask 25 high school seniors applying to college their age in years, number of siblings, proposed major and social security number and the top choice for college.

What is the population of interest?

What is the sample?

What are the variables?
Variables come in two types, qualitative or quantitative.

Definitions:

1) **Qualitative Variable** -

2) **Quantitative Variable** -

Classify each of the variables in the example.

Other examples of each type:

Qualitative -

Quantitative -

We will work with both types of variables this semester, but mostly we work with quantitative measures.

Quantitative variables can be categorized as discrete or continuous.

1) **Discrete** -

2) **Continuous** -

We will frequently use the terms discrete data or continuous data referring to these definitions.

**Levels of Measurement**

**Qualitative** data can be further classified as nominal or ordinal data.

1) **nominal data**- values that are names, labels or categories with no natural or inherent ordering or ranking of the labels.

   Examples - Favorite color, gender, phone number

2) **ordinal data** - has the properties of nominal data, but the words or labels used allow for the ranking or ordering of the values.
Examples - rating a restaurant as poor, fair, good or excellent. Course grades: A, B, C, D, F. Movie ratings of one to five stars.

Quantitative data can be further classified as interval or ratio data.

3) Interval Data - It has the properties of ordinal data, but the differences between the values are well defined and meaningful. Further, the zero value or start value of the data is chosen and does not mean the absence of the quantity.

   Examples: temperature, the years

4) Ratio data - has the properties of interval data and ratios of the values have meaning. A value of zero means the complete absence of the quantity.

   Examples: income, weight, number of books in a bag.

Section 1.2

Definitions:

1) response variable - the question being asked in a statistical study

2) observational study - a study where the value of the response variable is being measured without any attempt to influence the value of the variable.

3) exploratory (or explanatory) variable - a variable that is used to look at the difference in the response variable.

4) a designed study - A study where the researcher assigns the individuals to groups, intentionally changes the value of the exploratory variable and records the value of the response variable.

Examples:

1) A researcher is interested in blood pressure in women between the ages of 50 - 59. The blood pressure of 200 women is recorded.

2) A researcher is interested in the blood pressure of women between the ages of 50-59. The 200 individuals are divided into 4 groups, those that exercise less than 30 minute per week, those that exercise 30 - 89 minutes, those that exercise 90 - 179 minutes and those that exercise 180 minutes or more.
3) We are interested in vending machine soda sales on campus. The first 15 days of the month we make the price $1, the second 15 days we make the price $1.50.

Confounding occurs when the effects of two or more explanatory variables cannot be separated. The relationship between the response and exploratory variables may be due to some other variable or variables not accounted for in the study. These variables are called lurking variables.

Consider the soda experiment above, what is a possible lurking variable?

It is important to note that you can never draw cause and effect conclusions from an observational experiment. For example if we poll a sample and ask two questions,

1) Do you eat an apple everyday and 2) did you have the flu this year.

and then find that those that eat an apple every day had a lower incidence of the flu, we cannot say that eating an apple lowered the flu risk, (or that having the flu caused people to eat an apple everyday). We can only say there is an association between the two.

Section 1.3

Choosing the sample to be part of the study is extremely important. Ideally the sample will represent the population well. A "bad" sample leads to bad results and false information. For example, if I wanted to determine the percentage of SMC students who say they like math, polling the math club would not give good results.

Suppose I wanted to know the average age of an SMC student, do you think that this class would represent a "good" sample of the SMC population? Why or why not?
The best sampling method is **simple random sampling**.

**Simple random sampling** - a sample of size $n$ from a population of size $N$ is obtained through simple random sampling if every possible sample of size $n$ has an equally likely chance of being chosen.

Suppose that 12 people have volunteered for a committee and only three volunteers are needed. How can we choose the three using simple random sampling?

1) put all the names in a hat and choose 3.

2) list every possible group of three (there would be 220 possible groups of 3) put the lists of 3 in a hat and choose one of the lists.

Why is this method not always practical?

It is easy to see that the probability of any group being chosen is $1/220$ and that each group has the same chance of being chosen.

Suppose that desks in a classroom are set up in 6 rows of 8, for a total of 48 desks. A sample of 8 students is needed, so the teacher rolls a die and chooses the row corresponding to the number on the die. Is this simple random sampling? Why or why not?

A simple random sample can be obtained through a random number table or a random number generator on a calculator. Please read about this in the text, you should know how to use a random number table. We will discuss the random number generator on the calculator when we all have one!

**Section 1.4**

While simple random sampling is considered the most basic form of sampling, there are three other acceptable methods we will discuss.

1) **Systematic Sampling** - Each element in the population is numbered and the sample is obtained by taking every $k$th element.

   Example -
2) **Stratified Sampling** - The population is divided into subgroups or strata according to some characteristic and a simple random sample is taken from each subgroup.

   Example -

3) **Cluster Sample** - The population is divided into groups called clusters. Then a simple random sample of clusters are selected and all of the subjects of the chosen cluster are polled. The clusters are often geographic, but need not be.

   Example -

Note the similarity and differences between cluster and stratified sampling.

Another commonly used method of sampling is called **convenience sampling**. In a convenience sample the individuals in the sample of used because they are easily obtained and inclusion is not based on randomness. The results from such a sample are not considered useful.

1) I want to find the average age of an SMC student, so I use the students in this room.

2) A radio station asks listeners to call in and tell them how they rate their cell phone company.

The second example is called a **voluntary sample or a self-selected sample**. Who is likely to participate in such a sample?

While these are frequently used, they are not considered statistically sound and the results from these samples should be used with extreme caution.
Section 1.5

Not every sample is perfect, if the results of a sample are not representative of the population from which it is drawn, then the sample is said to be **biased or have a bias.** We will look at three types of bias, sampling bias, non-response bias and response bias.

1) **sampling bias** - this occurs when the methods used to obtain a sample favor one group in the sample over another. (Convenience sampling nearly always has this type of bias.)

Example - suppose that an organization wants to determine how a city will vote on a given issue. They choose three neighborhoods to poll and go door-to-door during the day. What is the possible sampling bias?

2) **non-response bias** - this occurs when individuals selected to be in a survey who refuse to respond to the questions have a different viewpoint than those that do. Nearly every survey has some element of non-response. This can be improved using various methods, including callbacks and rewards for participation.

3) **Response bias** - exists when the answers to survey questions do not accurately reflect the true nature or feelings of the respondent. Some types of this type of bias are summarized here, please see the for further examples.

   - interviewer error - the person recording the response does not obtain an accurate answer. Asking someone a sensitive question in an uncomfortable way, is an example of this type of error.
   - Misrepresented Answers - questions are asked in such a way that people tend to lie.
   - Wording of questions - vague questions and how a question is worded can lead to different answers. This is especially prevalent in political polling.
   - Ordering of questions or words - sometimes the order in which questions are asked can lead to different results.
   - Type of question - open questions vs. closed questions
   - Data Entry errors - errors that occur in recording a response

The error that results from the three types of bias above, is called **non-sampling error.** Any useful study will try to eliminate or minimize bias to reduce non-sampling error. There is also **sampling error,** this is error that occurs just because a sample or subset of the population was used. Every sample has this type of error, it is considered normal and is expected and accepted. This is why we can never be 100% certain of results obtained from a sample.
Section 1.6

An experiment is a controlled study conducted to determine the effect of varying one or more explanatory variables or factors has on a response variable. Any combination of the factors is called a treatment.

All experiments require a control group, a group that can be compared to other treatment groups.

Example - Suppose you are interested in seeing how well two new headache medications work. There could be one group that takes the currently accepted headache medication, and two more groups, each one taking one of the new medications. The first would be the control group.

Other terms and definitions:

1. Placebo -

2. Single Blind Experiment -

3. Double Blind Experiment -

Please finish reading the section through explaining the steps in designing an experiment.