Section 1: Introduction to Statistics and Data Collection

The Science of Statistics

How do you define **statistics**?

What do you think of when you hear the term **statistics**?

Can you think of examples where you see **statistics** utilized?

Statistics are commonly used in most fields of study and are regularly seen in newspapers, on television, and in professional work.

Finding examples pertinent to your interests should be quite easy. Below are a few of the many possible examples:

- Sports
    - Free throw percentage in Basketball
    - Quarterback rating in Football
    - RBI's in Baseball
    - Assists in Soccer
- Medical
    - The success rates and likelihood of side effects with a weight loss drug
    - The percentage of people who go into remission with a type of cancer
    - The effect of a drug on blood pressure
- Education
    - Writing down material helps in the comprehension
    - Reading test scores in 1^{st} graders are different for males and females
    - Depression is more likely in a student with low academic achievement
- Politics
    - Prediction that the incumbent president will be re-elected
    - Approval rating of senate
    - Percentage of people in Kentucky who support a bill that just passed

A formal definition of statistics is as follows:
Statistics - study of how to collect, organize, analyze, and interpret information

The advantage of statistics is that it gives a process for estimation and decision making when faced with uncertainties without prejudice (bias). This ideal will be discussed throughout the material this semester.

Fundamental Elements of Statistics

In order to understand the statistical process, we will begin with some basic statistical terminology.

Population - the collection of individuals or items of interest

Examples
1. All residents of the United States
2. All admits to hospitals in Lexington, KY

Census - measurements are taken from the entire population

When do you see a Census utilized?

A common response: every 10 years the U.S. census bureau conducts a census.
- They attempt to reach every resident in the United States
    - Do you think this is possible? What problems do you think they face?
        - One of the biggest problems is the homeless in the U.S. It is difficult to contact or even know how many are in this group.
        - The census bureau unlike most organizations, have legal authority which obligates people to fill out the census. Therefore, they do come very close to having a true census.

Often, it is not feasible to study the entire population. The main reasons are as follows:
- A census is rarely attempted because of the cost. It is very expensive to conduct a census unless you are dealing with a very small population. (a census may be reasonable with a very small population)
- In statistical practice, it is not usually possible to get everyone (or even close) to take part in a given study. Participants must agree to take part in the study and obviously everyone will not agree.

So if we cannot take measurements from the entire population, what should we do?

We often take measurements from a subset or smaller group of individuals instead of from the population.

Sample - the subset of the population on which we make measurements

Data – the measurements we collect

Variable - any characteristic of an individual that can take different values for different individuals

Statistic – a numerical characteristic of a sample
This value is known when we take our sample, but it will change from sample to sample.

Parameter – a numerical characteristic of a population
This value is a fixed number, but we could only know its value if we did a perfect census.

Since the population is often not available, we use statistics to estimate parameters.

Example
Suppose we are interested in estimating the average income for all residents of the state of Kentucky and we select 1000 people from the state and collect their incomes. Answer the following questions based on this scenario.

1. What is the population?
2. What is the sample?
3. What is the data?
4. What is the variable?
5. What is the parameter?
6. What is the statistic?
7. If we had taken measurements from all residents of Kentucky, this would be called what?

Answers
1. All residents of the state of Kentucky
2. The 1000 people selected from the state
3. The list of the 1000 incomes collected from the sample
4. Income (it is different for each individual)
5. The average income for all resident of the state of Kentucky
6. The average income for the 1000 people selected from the state
7. A census

Branches of Statistical Applications

We are now ready to look at the two main parts or branches in the science of statistics.

1) **Descriptive Statistics** - methods of summarizing a set of data (There is no error here because you are summarizing the data you have.)

2) **Inferential Statistics** - methods of making inference about a population based on the information in a sample (There is error here because you do not have all of your data; you have a subset of the population and cannot expect to be exactly correct. We can control and calculate this error with probability. Therefore, probability will be covered before inferential statistics.)

Examples

Identify each of the following as a descriptive or inferential statistic:

1. A basketball players free throw percentage based their performance up to the midpoint of the season
2. The average GPA for all current students based on complete data from the records office
3. The presidents current approval rating based on a group of 5,000 U.S. residents
4. A graph showing all sales this year for all major vehicle manufacturers
5. Suppose I collect data from this class and use it to calculate the percentage of students from the class who live in Lexington
6. Suppose I collect data from this class and use it to estimate the percentage of all BCTC students who live in Lexington

Answers

1. Descriptive statistic (there is no error since we know exactly how many free throws the player has shot and made at this point in the season)
2. Descriptive statistic (there is no error since we have the GPA of all current students)
3. Inferential statistic (there is error here because we do not have data from the entire group of interest; we are using a sample of 5,000 U.S. residents to estimate the approval rating for the population of all U.S. residents)
4. Descriptive statistic (there is no error here because we have all sales figures for the major vehicle manufacturers; almost all graphs and charts are descriptive)
5. Descriptive statistic (there is no error because we have data from the entire group)
6. Inferential statistic (there is error because we are estimating the population of all students at BCTC with data from a sample, the students in this class)

Notice that 5 and 6 above yield the same data. However, the question is different which leads to the analysis being descriptive in one case and inferential in the other. The group of interest is key to identifying if there will be error.

Types of Mathematical Reasoning

Inductive reasoning – used when doing inferential statistics (reasoning by consistency; bottom-up reasoning)

Deductive reasoning (or probability theory) **–** is just the opposite, understanding properties of a sample from a known population (reasoning by certainty; top-down reasoning)

For this class, after introducing some basic terminology and concepts, the material will be covered in the following order: 1) Descriptive Statistics, 2) Deductive Reasoning (probability theory), and 3) Inferential Statistics (inductive reasoning).

Types of Data

Qualitative (Categorical) variable – places responses into categories and there is no numerical meaning to the outcomes (non-numerical data)

Quantitative variable – numeric values that have numerical meaning (numerical data)
- **Discrete variable** – things that can be counted
- **Continuous variable** – things that are measured

Examples
Identify the data type for each of the following:
1. Gender
2. Time to run a mile
3. Number of soccer goals scored
4. Religious preference
5. Distance to drive to work
6. Number of correct answers on a test
7. Social security number

Answers
1. Qualitative
2. Quantitative (Continuous)
3. Quantitative (Discrete)
4. Qualitative
5. Quantitative (Continuous)
6. Quantitative (Discrete)
7. Qualitative

Data Collection

Bias – a prejudice in one direction (this occurs when the sample is selected in such a way that favors the selection of elements with a particular characteristic)

In inferential statistics we use a statistic to estimate a parameter. If a statistic is biased it will tend to overestimate or underestimate the value we are trying to predict. We can eliminate (or at least minimize) bias by using proper sampling techniques.

For a sample to be useful when conducting inferential statistics, it needs to be **representative of the population**! This is the common terminology used in identifying a sample as unbiased and useful in extending the results to the population.

Random sample – A sample determined completely by chance. This is how to get a representative sample.

With a random sample, the sample will typically be similar to our population with respect to demographic and other variables. Also, we can control the probability of making a mistake (or probability of error).

Example

For illustration purposes, suppose that we are interested in estimating the average income for residents in the state of Kentucky. If we sample 1000 males from Lexington, would the sample be representative of the population? No, the sample would be biased and would tend to overestimate average income. Males tend to make more on average than females and people from Lexington tend to make more than what is average in the state. What we would want is if 50% of our population is male then 50% of our sample should be male. Also, if 10% of the population is from Lexington then 10% of our sample should be from Lexington. There are an uncountable number of other variables for which we would want the sample to be similar to the population. A random sample will take care of all these variables at once. Keep in mind this is not a guarantee. However, if 10% of the population is from Lexington and we take a random sample, then the probability is high that 10% of our sample will be from Lexington. In actuality, we know that the probability is high that the sample is similar to the population in terms of all demographic and other variables. Just as important, we can calculate this probability of getting a good sample as compared to a bad sample. We will look at these calculations later in the semester. Finally, it should be understood that we can make this probability higher by taking a larger sample size. It works kind of like flipping a coin. If you flip a coin 10 times what is the probability of getting tails about 50% of the time? If you flip a coin 50 times what is the probability of getting tails about 50% of the time? The important thing here is that if you flip the coin more times, you are more likely to get a percentage close to the population value of 50%.

A **simple random sample** of n measurements from a population is one selected in such a manner that every sample of size n from the population has equal probability of being selected. (n is standard notation for sample size)

There are multiple ways to take a random sample. Everything we do in this course will be based on a simple random sample. This is the most common method used and is preferred in most cases.

Example

Suppose that the students in this class are my population. For illustration purposes assume we are in a classroom and there are 5 rows and 10 students in each row. I want a sample of 10 students from the class so I randomly select one row as my sample. Is this a simple random sample? This is a situation where you must be careful about how you think of the probability. The mistake people tend to make is asking if each person has an equal chance of being selected. For this scenario, the answer is yes. Every person has a one out of five chance of being selected since 1 of the 5 rows is being selected. The question that should be asked is does every sample of size 10 have an equal chance of being selected. The answer here is no. You can only be in a sample with the people in your row. There is no probability of you being in a sample with someone from another row. Meaning every sample does not have equal probability of being selected and this is not a simple random sample. When you group individuals and select a group or groups, you no longer satisfy the definition of a simple random sample. In order to actually get a simple random sample something must be done equivalent to drawing names out of a hat. This will mean all samples have an equal probability of being selected.

It should be noted that the value of a statistic depends on which items are selected for the sample. When we take a random sample a statistic is a random variable. We are randomly selecting one value for the statistic out of all its possible values. This is what allows us to calculate probability. In this class we will look at the probability of a random variable and then extend that to a statistic.

Error in the Statistical Process

Error – the distance between a statistic and parameter in inferential statistics

Error comes from two primary sources, random sampling error and non-sampling error.

Random Sampling error – error that comes from using a random sample to estimate a population

Random sampling error is the only type of error that can be calculated.

Non-sampling error – errors that do not come from the sampling process

Since this is a methods class, we will focus on the idea of random sampling error. In fact, when the term error is used, really it is assumed that the data is representative and came from a sample with no issues. Therefore, any error would be random sampling error.