MATH 109 Lecture Notes

1 Introduction to Statistics

1.1 An Overview of Statistics

Definition 1: Data is information coming from observations, counts, measurements, or responses.

A list of numbers or words is NOT data. Without context, words and numbers will mean nothing to us.

Definition 2: Statistics is the science of collecting, organizing, summarizing, analyzing, and interpreting data to draw conclusions, answer questions, and make decisions.

Some professionals consider statistics to be a specific field within mathematics. Others consider statistics to be a completely separate field from mathematics. Some universities even have separate statistics and mathematics departments. What should this tell you? While we will be performing computations and using some mathematical techniques in this course, it is also probably very unlike most other math classes you've ever taken in high school or college.

Our most common form of expressing or gathering data will be in a data table. Each row in a data table is a case, individual or record. (In other specific situations, a case may be known as a respondent, a subject, a participant, or an experimental unit. For our purposes, we will almost always use the general terms.) The collection of all cases is the population. Often our data table will only include data for SOME cases of the total population. When this happens, our data cases are only a sample of the total population.

Example 1.1: Below is an example of a data table, based on some of the information collected in the first-day-of-class survey for a previous class.

Number	Gender	States Visited	Area Code	Left-handed?	Hours sleep	Siblings
1	M	7	214	No	8	3
2	F	5	469	No	6	1
3	F	11	832	No	8	0
4	F	30	972	Yes	8	1
5	M	4	325	Yes	7	5
6	F	4	817	No	8	1
7	F	5	210	No	6	1
8	M	12	903	No	6	1
9	M	3	972	No	5	5
10	M	12	979	No	8	2

Is this table showing data related to a population or a sample?

A statistic is a numerical summary of a sample. A parameter is a numerical summary of the population. Descriptive statistics is a specialty in the field of statistics that involves organizing and summarizing the data from samples, usually using numerical summaries, tables, and graphs. Inferential statistics is a specialty in the field of statistics that uses methods to extend the result of a sample to draw a conclusion about the population and measure the reliability of that result. In other words, inferential statistics uses statistical theory to use descriptive statistics of samples to try to draw conclusions about the population parameters.

Example 1.2: In each scenario below, determine the population and sample. Then determine if each number given is a parameter or a statistic.

A survey of 1015 US adults found that 42% have put off car repairs in the past 10 years.

The average annual salary of 20 of a company's 60 janitors is $29,000.

Fifty-four current US Senators are from the Republican Party.

1.2 Data Classification

There are two major types of data:

- Qualitative data is data that can be divided into named categories. Usually the values will be words (rather than numbers), but this is *not* always the case.
- Quantitative data is measured data WITH UNITS. If your data value is not a number and/or does not include units, it is almost definitely a categorical variable.

There are various ways to measure data values:

- Data at the nominal level of measurement are categorized using names, labels, or qualities. This level of measurement is for qualitative data only, and no mathematical computations can be made at this level.
- Data at the ordinal level of measurement can be arranged in order, but differences between data entries are not meaningful. Data at this level of measurement may be qualitative or quantitative. For example, on a course evaluation survey, students may be prompted with the statement "I am interested in this subject." and must respond with "Strongly agree," "Agree," "Disagree," or "Strongly Disagree." These responses may be stored as 4, 3, 2, and 1, respectively.
- Data at the interval level of measurement are numerical, can be ordered, and addition and subtraction have meaning. For example, temperature may be considered to have interval level of measurement because subtracting one temperature from another has meaning in comparing those two values.
- Data at the ratio level of measurement are numerical and multiplication and division have meaning. In this case, unlike with the interval level of measurement, zero means the absence of quantity. For example, height may be considered to have ratio level of measurement because someone 60 inches tall is twice as tall as someone who is 30 inches tall (i.e., division makes sense in this context).

Example 1.3: Why is temperature at the interval level of measurement and not the ratio level of measurement?

Definition 3: A variable is a characteristic that is recorded for each case of data.

Variables will correspond to the column headings of a data table, and pertain to *what* has been measured. We are not using "variable" to mean an unknown quantity, like in algebra. For us, a "variable" is not unknown; it is a specific piece of information we already know about each piece of data.

Example 1.4: The following questions appeared on our first-day-of-class survey. Each question corresponds to a variable. Determine the type and level of measurement of each variable.

Gender	Hours of sleep each night	Number semesters at Mercyhurst
Major	State/Country of Birth	Number of US States Visited
Area Code	Years between parents' ages	Number siblings
Height	Left-Handedness	Color-blindness
Age	Number of E-Mail Addresses	Favorite Movie
Shoe Size	Number musical instruments played	Favorite Sport
Eye Color	Favorite TV Show	Favorite Genre Music

Example 1.5: In June 2003 *Consumer Reports* published an article on some sport-utility vehicles they had tested recently. They reported some basic information about each of the vehicles and the results of some tests conducted by their staff. Among other things, the article told the brand of each vehicle, its price and whether it had a standard or automatic transmission. They reported the vehicle's fuel economy, it acceleration (number of seconds to go from zero to 60 mph) and its braking distance to stop from 60 mph. The article also rated each vehicle's reliability as much better than average, better than average, average, works or much worse than average.

(a) Identify the variables.

(b) Determine the type and level of measurement of each variable.

1.3 Data Collection and Experimental Design

In this section, we will discuss the answers to two questions: "Where does data come from?" and "What steps can I take to get *good* data?"

Example 1.6: A February 2015 survey of 1,504 adults in the US found that 64% of the population believes "most people who want to get ahead can make it if they're willing to work hard." [Source: Click here in PDF for source.]

What is the population of interest? What is the sample? Is the number given a parameter or a statistic?

The latest (July 2014) US Census Bureau estimate of the US population is approximately 318.857 million people (245.273 million adults). How can a survey of less than 2,000 people possibly tell us ANYTHING about the population of interest? This is the question we hope to answer as we explore inferential statistics later this semester (Chapters 6, 7, 9). However, before we can build that theoretical framework (Chapters 3, 4, 5), we will discuss how to obtain data (this section) and how to describe and summarize that data both verbally and numerically (descriptive statistics, in Chapter 2).

There are several types of statistical studies which can be used to gather data:

- A survey that obtains responses from *every* member of the population is called a census. While this seems like the best way to obtain the most accurate data, it is rare that a census will be accurate or even possible. A census is extremely difficult to perform for large populations, a complete census is more susceptible to under-counting and over-counting than good sample, and a census can take so long to complete that the population changes while you're doing it (people are born and die, opinions change, etc.). A survey that obtains responses from a portion of the population (a sample) is called a sample survey.
- In an observational study, researchers observe the choices and/or conditions of individuals in a sample, then record and analyze the data.
  - A retrospective study is an observational study that looks at choices and conditions that *have already happened*. Retrospective studies can involve looking at historical data, or asking subjects about what has happened to them in the past. These are often called case-control studies because researchers can choose which cases to study that fit the criteria of their research objectives.
  - A prospective study is an observational study that observes the choices and conditions *as they are happening*. This may involve following the subjects for a period of time, or checking in with subjects periodically to record data. These are often called cohort studies because a cohort of people are chosen and then observed over a period of time.
  - Observational studies can describe the association between variables, but cannot determine cause-and-effect relationships.
- When the researchers identify subjects in advance and assign them to the conditions of interest and then observe the results, this is an experiment. In an observational study, researchers have no control over which subjects are under which conditions. In an experiment, the research imposes conditions upon the subjects. Unlike with an observational study (which never indicates causal relationships), a well-designed experiment actually **can** determine if one condition *causes* another. The specific values of the response variable(s) the experimenter assigns to one particular subject is called the treatment for that subject.

- A simulation is the use of a mathematical or physical model to reproduce the conditions of a situation or process. Simulations do not produce "real" data, but allow researchers to guess outcomes, create reasonable estimates, and avoid potentially dangerous situations. For example, when studying the effects of safety features during crashes, an observational study or simulation is much safer than an experiment. (An experiment would require the researchers to assign some experimental subjects to crash on purpose!)

The Basic Building Blocks of a Sample Survey

1. Clearly identify and state the information you are interested in obtaining, and the population you desire to have this information about. Determine if sampling is helpful and/or necessary.
2. Select an appropriate sample size.
3. Identify the sampling frame.
4. Choose a random sample. (This obtains a representative sample that avoids bias.)

Sample Size. Our intuition may be to assume that the higher percentage of the population we use in our sample, the more precise our results will be. However, this is actually NOT the case! A good sample size does not depend on the fraction (or percentage) of the total population, but on the size of the sample itself. We won't get into the technicalities of choosing a "good" sample size in this chapter (that will come in Chapter 6). For now it is enough to know that it is the number of people in the sample, not the percentage of population in the sample, that affects how accurately the statistics of the sample predict the characteristics of the population.

Selecting the sample. Once you know how many people (or things) you intend to survey, you must select which individual people (or things) will actually be surveyed. The most important thing is to be sure that your sampling method is random, that is, that each member of the population has an equal chance of being selected in the survey.

The sampling frame is the list of individuals from which we select our sample of the population. The sampling frame is where the sample was chosen or generated from. If, for example, we want to take a sample of all Mercyhurst University students, our sampling frame may be a list of all currently enrolled students that is generated by the registrar. The sampling frame for election polls may be the voter registration records that list all registered voters.

Once you have identified your sampling frame, you should choose your sample from that list. Below are several common and effective ways of choosing your random sample.

- Simple Random Sample (SRS)
  - First choose your sampling frame. Then assign a random number to each member in the sampling frame. Then select your sample from the random numbers that satisfy some rule.
  - One possibility is to assign numbers randomly in a large range (larger than the size of the sampling frame), then sort the list by putting the random numbers in increasing order. Choose the first n entries to get a sample of size n. For example, if we want to randomly select 100 students from the approximately 2,680 undergraduate students at Mercyhurst [and we've obtained a list of all the students], we can assign a random number between 1 and 50,000 to each student in the list, then use the students who have the 100 lowest numbers as our sample.
  - Another possibility is to assign numbers randomly in a small range (smaller than the size of the sampling frame), then choose only those entries that have a specified value.

For example, if we want to randomly select around 250 students from the approximately 2,680 undergraduate students at Mercyhurst, we can assign a random number between 1 and 11 to each student in the list, then only use the students who have been assigned the number 4 as our sample.

- Stratified Sampling
  - First choose your sampling frame. Then consider homogeneous pieces separately using SRS on each piece.
  - For example, decide how many men and how many women you want in the sample, then randomly select from the men and the women separately. Or if you want a certain age group distribution, you can randomly sample separately from each age group. The groups are called strata (plural), or each individual group is called a stratum.
- Cluster Sampling
  - Break the population up into *representative clusters*. Then randomly select one or more clusters. Then for each chosen cluster, perform a SRS or census.
  - When is this useful? Let's say our sampling frame is all people listed in a certain phone book. We may cluster the population into pages, then randomly select 10 pages, and then either randomly select individuals from each page, or survey all people on those 10 pages.
  - What is the difference between strata and clusters? Strata are homogeneous, meaning all people in a stratum are "the same" in some way. A cluster should be representative, meaning it should be a fair reflection of the randomness qualities of the entire population.
- Systematic Sampling
  - A systematic sample chooses every x people from a list or queue to survey.
  - For example, instead of using a SRS to select 250 students from all Mercyhurst undergraduate students, we may choose to use the alphabetical list and survey every 11th student. However, to better model randomness, we should start counting from a random point in the list of students (instead of always starting with the first name).
- Multistage Sampling
  - A multi-stage sampling method combines two or more of the above methods. For example, you may used stratified sampling to ensure a 50-50 distribution of men and women in the sample, but then use systematic sampling to choose the men and the women.

Note: A convenience sample is NOT random and almost certainly NOT representative, and therefore little or no confidence can be put in the results of a survey that comes from a convenience sample. A convenience sample consists of members of the population that are easy to survey and/or consists only of volunteers. For example, survey results from voluntary online surveys are notoriously biased.

Example 1.7: Suppose ESPN wants to schedule a TV double header for Sunday afternoon/evening. Usually, of course, they look at which teams are most marketable to bring in the maximum number of viewers. However, suppose when they designed the schedule, they wanted to mix it up for this week's games and choose the teams randomly.

1. Perform an SRS to randomly choose the four teams that will play.
2. Suppose ESPN wants to show one game between two American League teams and one game between two National League teams. Randomly choose the teams that will play. What sampling method did you use?

Example 1.8: Now suppose the MLB wants to survey players at the end of this season to determine if they are happy with the way contracts are handled. A rough estimate for the number of major league players at the end of the season, including those on the 40-man roster and the 60-day disabled list, is 1500. This is comprised of approximately 20% starting pitchers, 25% relief pitchers, 10% catchers, 25% infielders, and 20% outfielders. Several sampling strategies are suggested below for choosing 100 players to survey. Determine the sampling frame and sampling method used in each scenario.

- A list of each player's name, phone number, and position is obtained from all teams. The list is divided up by position, and 20 starting pictures, 25 relief pitchers, 10 catchers, 25 infielders, and 20 outfielders are randomly chosen.

- A list of each player's name and phone number is obtained from all teams. The list is arranged alphabetically by last name. Starting at a random point in the list, every 15th player is surveyed.

- A list of each player's name and e-mail address is obtained from all teams. An online survey is e-mailed to every player and the first 100 responses are used.

- A list of each player's name and phone number is obtained from all teams. A digit between 0 and 9 is randomly chosen. Every player that has a phone number ending in that digit is surveyed.

- A list of each player's name and phone number is obtained from all teams. A random number between 0 and 999,999 is assigned to every player. The players with the lowest 100 numbers are used.

- A list of each player's name and phone number is obtained from all teams. Five letters of the alphabet are randomly chosen and every player whose last name begins with those letters is surveyed.

- A list of each player's name and phone number is obtained from all teams. Four players are randomly chosen from each team.

Example 1.9: Suppose you want to conduct an experiment to investigate the relationship between hours of sleep per night and grades for Mercyhurst students. Assume you have a group of students to study.

- If you simply ask each participant what their GPA is and how many hours of sleep they sleep each night, this is a **retrospective observational study**.
- If you ask each participant to keep a log of how many hours they sleep each night throughout the semester, then collect this information and their GPA at the end of the semester, this is a **prospective observational study**.
- If you tell each participant that they must sleep a specified number of hours each night (for example, you tell some students to sleep 4 hours per night, some to sleep 6 hours per night, and some to sleep 8 hours per night), and then record their grades at the end of the experimental period (for example, the semester), this is an **experiment**.

    1. What are the explanatory variables for this experiment?

    2. What is the response variable?

    3. What are the treatments that are assigned?

Example 1.10: Researchers want to know if parents who have a certain medical condition tend to pass that condition on to their children. What is the explanatory variable in this study? What is the response variable in this study? For each of the observational studies described below, identify whether it is retrospective or prospective.

1. The researchers identify subjects who have the condition and are expecting a child. They monitor the child after it is born for signs of developing the condition.
2. The researchers gain access to the medical records of people with the condition, and the medical records of their children. They record whether the children also have the condition.

Principles of Experimental Design

1. **Control:** Our goal is to use values of factors in the experiment to explain changes in the response variable (this would be a causal relationship). A well-designed experiment should minimize (or eliminate) the effects of all variables except the factors of the experiment. In order to do this, we want to fix ("control") as many outside influences as we can.
2. **Randomize:** Whatever we can't control, we should randomize. This further "evens out" the effects of the variables that we aren't interested in tracking. We should always assign subjects to treatments randomly to minimize the effects of lurking variables.
3. **Replicate:** You should use more than one subject in an experiment. When possible, you should repeat the experiment more than once. And you should always make sure that your experiment CAN be repeated, so that other researchers can repeat and confirm your results if they want to. If your experiment can never be repeated, your results are not useful because they only record information about one snapshot in time and should not be extrapolated to other conditions.

Blinds

Definition 4: Sometimes it is helpful for a subject to not actually know what treatment they are receiving. When you purposefully hide treatment information, it is called blinding the experiment.

- We can blind subjects (so that individuals being experimented on do not know what treatment they are receiving), or we can blind people involved in administering the experiment (doctors evaluating the patients under different treatments, etc.). When ONE of these two groups are blinded, it is called a single-blind experiment. When BOTH groups are blinded, it is called a double-blind experiment.

Control Groups

Definition 5: A control group is a group of the subjects who are assigned the treatment of "no treatment." This is useful in a lot of experiments, especially those where you want to compare "doing something" to "doing nothing" (to determine if "doing something" helps or hurts or has no effect).

Definition 6: Sometimes a control group doesn't actually involve doing nothing. For example, in a blinded experiment, you may want to give an active drug to some participants and give no drug to other participants (a control group). However, if you don't give ANYTHING to the control group, they know they are the control group, and they may respond to the experiment by saying they feel worse, because they know they haven't taken any medications to fix their ailment. Usually, if such an issue is a concern for an experimenter, they will design a "fake" treatment for the control group (a sugar pill that doesn't actually do anything, etc.). The subjects will be unable to tell if what type of treatment they are receiving, or if they are receiving any treatment at all. This type of "fake" treatment is called a placebo. Often subjects who receive a placebo *believe* they are getting a real treatment, so their condition actually changes, even though nothing is being done by the experiment to change their condition. This is called the placebo effect.

Blocking

Definition 7: When there is a variable in play that we can neither control nor randomize, sometimes it can be helpful to divide the subjects up according to this variable, and then randomly assign the treatments within each group, rather than randomly assign the treatments to all subjects. This technique is called blocking. It is very similar to stratified sampling for sample surveys.

Example 1.11: Consider the medical condition that may be passed from parent to child from Example 10. What if we suspect that it matters whether it is the father or the mother with the condition? Then it may be helpful to block a study or experiment by gender of the parent, and then investigate the response variable.

In our experiment on Mercyhurst students regarding sleep and grades in Example 9, suppose we want to consider the sciences and the arts separately. How should we assign the treatments to the subjects if we want to use blocking for the experiment?

Definition 8: When the values of one explanatory variable are associated with the values of another explanatory variable, we say that these two variables are confounded. When we have confounded variables, it is not possible to determine which one is really affecting the results in the response variable. Confounding variables render the results of an experiment practically useless.

Example 1.12: Suppose a bank wants to know if customers prefer a low interest rate or a low annual fee for a credit card. They send out card offers to 10,000 people. Five thousand of the offers promise a low interest rate and no annual fee. The other 5,000 offers give a higher interest rate and a $50 annual fee. Nearly twice as many people accepted the first offer, compared to the number of people who accepted the second offer. What can we conclude from this study?

Example 1.13: An article in a local newspaper reported that dogs kept as pets tend to be overweight. Veterinarians suggest that diet and exercise will help these chubby dogs get in shape. They recommend two different diets (Diet A and Diet B) and two different exercise programs (Plan 1 and Plan 2). Sixty dog owners volunteer to take part in an experiment to determine which diet and exercise program best helps overweight dogs lose weight.

1. What are the subjects of this experiment?

2. What is/are the explanatory variable(s) in this experiment?

3. What are the treatments in this experiment?

4. Is the experiment single-blind, double-blind, or not blind at all?

5. What is the response variable in this experiment?

6. Describe the design of this experiment, including how you would assign treatments to the subjects.