Chapter 1

1.1 Definitions of statistics, Probability, and Key Terms

Statistics is a field of study.
It is the study of methods for planning studies and experiments, collecting data, and then organizing, analyzing, presenting, interpreting, and drawing conclusions and making predictions based on the data.

We will study two main areas of Statistics:

Descriptive Statistics
Methods for organizing, displaying, and summarizing data.

Inferential Statistics
Methods for drawing conclusions from data.
We use probability to determine how confident we are that our conclusions are correct.

A population is the collection of ___________ elements to be studied.

The population of interest is also often called the ____________________________.

A sample is a __________________________________ of elements selected from a population.

A survey is a collection of information from the elements in a population or sample.

A census is a survey that includes __________ element of the target population.

A sample survey is a survey conducted on a ________________. 

A census is rarely taken since it usually is ____________________________________ .

We can use sample data to form conclusions about populations, if we obtain sample data that are representative of the data from the target population. This is called a representative sample. (More about samples in 1.2)

A variable is a characteristic that we are studying and which takes on different values for different elements. A variable is usually notated by a capital letter such as X or Y.
Information collected by sampling is called a **data set**. Information is collected on **experimental units** (individuals or objects being measured). The actual values of a variable are called **data**. Below is an example of a data set.

<table>
<thead>
<tr>
<th>Student</th>
<th>Grade</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maria</td>
<td>B</td>
<td>22</td>
</tr>
<tr>
<td>Bob</td>
<td>A</td>
<td>45</td>
</tr>
<tr>
<td>Juan</td>
<td>B</td>
<td>21</td>
</tr>
<tr>
<td>Theresa</td>
<td>C</td>
<td>34</td>
</tr>
</tbody>
</table>

In above example the experimental units are ________________________________

In above example the variables are __________________________________________

Circle the data in above example.

**ex.** A fitness center is interested in the mean amount of time a client exercise in the center each week.

**Identify:**

The population ____________________________________________
The sample ____________________________________________
The parameter ____________________________________________
The statistics ____________________________________________
The variable ____________________________________________
The data ____________________________________________

1.2 Data, Sampling, and Variation in Data and Sampling

A **quantitative variable** can be measured numerically. Its data are called **quantitative data**.

**ex.**

A **qualitative variable**, also called categorical variable, cannot be measured numerically, but separate individuals into different categories distinguished by some characteristic. Its data is called **qualitative data**.

**ex.**

There are two types of **quantitative** variables: discrete and continuous.
A **discrete variable** has values that are countable.

*ex.*

A **continuous variable** can take on any numerical value over a certain interval(s). Continuous data are results of measuring rather than counting.

*ex.*

**Sampling**

A sample should have the same characteristics as the population it is representing.

In a **random sample** elements from the population are selected in such a way that each individual element has an equal chance of being selected.

In order to get a representative sample we usually use simple random sampling whenever possible. In a **simple random sample**, all samples of the same size selected from a population have equal chance of being selected.

*ex.*

In a **systematic sampling**, the population is ordered, and after randomly deciding a starting place one will sample every kth individual.

*ex.*

**In stratified sampling**, the population is divided into groups, strata, where the members of each stratum are similar in some way. Then a simple random sample is drawn from each stratum.

In **cluster sampling**, the population is divided into clusters (groups) and then some of these clusters are randomly selected, and all members from each of the selected clusters will be included in the sample.

Some Non-Random Sampling Techniques:

**In convenience sampling** the most accessible individuals/elements of the population are selected.

*ex.*

A sample of convenience usually does NOT give us a representative sample but may be acceptable when it isn’t possible to draw a simple random sample and if it is reasonable to believe that there is no systematic difference between the sample and the population.

**In a voluntary response sampling**, a survey is published in a magazine or similar and people are invited to respond to the survey.

*more ex.*

Voluntary response samples are NOT reliable, as people who go through the hassle of volunteering their opinion tend to have ....................... opinions than the typical population.
From now on, unless otherwise stated, we will assume that any sample that we work with is a *simple random sample*.

*ex. Identify the kind of sample that is described below*

a) *Police at a sobriety checkpoint pull over every fifth car to determine whether the driver is sober.*

b) *A cell phone company wants to draw a sample of 600 customers to gather opinions about potential new features on upcoming phone models. The company draws a random sample of 200 from customers with Blackberry phones, a random sample of 100 from customers with Samsung phones, and a random sample of 200 from customers with other phones.*

c) *Viewers of a television show are asked to vote for their favorite performer by sending a text message to the show.*

d) *The manager of a restaurant walks around and asks selected customers about the service they have received.*

e) *A research group randomly selects 20 of Cabrillo’s classes and includes all the students from these classes in their study.*

Each time we select an element from the population to be included in our sample, we can put it back in the population before we select the next element, called __________________________________________ (note that the same element could get picked again),

or we don't put it back, called _______________________________________________________.

We need to be critical as we evaluate statistical studies. A study conducted by a procedure that produces the correct result on the average, is said to be **unbiased**.

A study conducted by a procedure that tends to overestimate or underestimate the true value is said to be **biased**. We want to design studies to have as little bias as possible.

Below are some common types of bias and problems to be aware of:

- **Sample size issues** – Small samples may be unreliable. The larger the sample the better.

- **Voluntary response bias** - People with strong opinions are more likely to participate in such surveys.

- **Self interest bias**- People who have an interest in the outcome of an experiment have an incentive to use biased methods.

- **Social acceptability bias** - People are reluctant to admit to behavior that may reflect negatively on them.
- **Leading question bias** - The way a question is worded can influence the response. Some interviewers may deliberately use subtle leads to obtain the answers they desire, but often neither the interviewer nor respondent is aware of the extent to which the wording of the question can influence the response.

  *ex. In a study, 40 people were asked about headaches. Those who were asked 'Do you get headaches frequently and, if so, how often?' reported an average of 2.2 headaches per week, whereas those who were asked 'Do you get headaches occasionally and, if so, how often?' reported only 0.7 per week.*

- **Nonresponse bias** - People who refuse to participate in a study are called nonresponders, and their opinions tend to differ from the people who do respond. So a study with many nonresponders are often biased.

- **Sampling bias** - This happens when some members of the population are more likely to be included in the sample than others. We can minimize this by using ......................................................, but we might not be able to avoid it completely.

  If you use a biased method to draw a sample, you cannot make up for it by drawing a big sample. A big nonrepresentative sample does not describe the population any better than a small nonrepresentative sample.

- **Misleading use of data** – Improperly displayed graphs, incomplete data, etc.

- **Confounding** – Multiple factors can contribute to certain outcomes and it may be impossible to know what variable causes another to occur.
  
  *ex.*

---

**1.3 Frequency, Frequency Tables, and Levels of Measurement**

Data recorded in the sequence in which they are collected and before they are processed or ranked are called raw data. Raw data is often difficult to make sense out of (especially if you have very big data sets) so we prefer to organize and summarize our data.

When **rounding off** our answers, we usually let our final answer carry one more decimal place than was present in the original data. Round off only the final answer and not any intermediate results, if possible. If you have to round off in an intermediate step, include at least twice as many decimals as the final answer.

**Frequency** is the number of times a value of the data occurs.

A **frequency distribution** is a table that lists all categories and the number of elements that belong to each of the categories.

**Reasons for Constructing Frequency Distributions:**

1. Large data sets can be summarized.
2. We can gain some insight into the nature of data.
3. We have a basis for constructing important graphs.
The relative frequency of a category is the proportion of items in the category and can be written as a fraction, percent, or decimal.

Relative frequency of a category \( = \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}} \)

A relative frequency distribution is a table that represents the relative frequency of each category (often times we just add another column with this information in the same table as our frequency distribution).

Cumulative relative frequency is the accumulation of the previous relative frequencies.

Practice:
Suppose we ask 30 people if they ever suffer from insomnia. The responses are classified as Never (N), Sometimes (S), Often (O), Always (A), and are recorded below.

<table>
<thead>
<tr>
<th>S</th>
<th>O</th>
<th>S</th>
<th>A</th>
<th>N</th>
<th>N</th>
<th>S</th>
<th>O</th>
<th>A</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>O</td>
<td>O</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>O</td>
<td>N</td>
<td>S</td>
</tr>
<tr>
<td>S</td>
<td>N</td>
<td>A</td>
<td>O</td>
<td>S</td>
<td>S</td>
<td>O</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

1. Construct a frequency distribution table for the data above. Compute the relative frequencies and percentages for all categories.

<table>
<thead>
<tr>
<th>Insomnia Occurrence</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Cumulative Relative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \text{Sum} = \text{Sum} = \text{Sum} = \)

2. What proportion of the people in this sample suffer from insomnia Often or Always?

A graph made of bars whose heights represent the frequencies or relative frequencies of respective categories is called a bar graph. The bars should be of equal width and not touch each other. Make sure to always label both the horizontal and vertical axis.
3. Draw a bar graph for the frequency distribution of the Insomnia Occurrence data.

4. How would a bar graph on the relative frequency distribution of the Insomnia Occurrence data differ from the one you just constructed?

Bar graphs can be constructed vertically (more common) or horizontally. Sometimes we want to compare two or more bar graphs that have the same categories, so we can construct several bar graphs on the same axes, with corresponding bars next to each other. This is called a side-by-side bar graph. See example on the right.

A pie chart is a circle divided into sectors, each one representing a different category and its size proportional to the proportion size (relative frequency) of that category. Each sector should be labeled with its category and its relative frequency expressed as a percentage.

Angle = (Relative frequency) \cdot 360^\circ

In this class you don’t need to calculate and measure exact angles, but it’s acceptable to just approximate the size of each sector.

4. Construct a pie chart for the distribution of the Insomnia Occurrence data.
Since quantitative data doesn’t have natural categories, we divide the data into classes. The classes are intervals of equal width that cover all the values in the data set.

A Frequency Distribution for Quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called grouped data.

<table>
<thead>
<tr>
<th>Weekly Earnings (dollars)</th>
<th>Number of Employees</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>801 to 1000</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1001 to 1200</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td><strong>1201 to 1400</strong></td>
<td><strong>39</strong></td>
<td></td>
</tr>
<tr>
<td>1401 to 1600</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>1601 to 1800</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td><strong>1801 to 2000</strong></td>
<td><strong>6</strong></td>
<td></td>
</tr>
</tbody>
</table>

[1.4] Experimental Design and Ethics

Experimental units are the individuals/elements that are studied. When the experimental units are people, they are sometimes called subjects.

The treatments are the procedures applied to each experimental unit.

The outcome, or response, is what is measured on each experimental unit.

Randomized Experiment vs. Observational Study

A randomized experiment is a study in which the investigator assigns the treatments to the experimental units at random.

ex.

An observational study is one in which the assignment to treatment groups is not made by the investigator.

ex.

In general, we prefer ......................................... over ............................................ as it is easier to interpret these results.

Give an example on when a randomized experiment cannot be performed.

ex.
A **placebo** is a "medical" treatment that has no medical effect, used to deceive the recipients in the **control group** so that they won't know what treatment they are given.

**Placebo effect** is ..................................................................................................................................................

When a person involved in a research study is **blinded**, he does not know who is receiving the active treatment vs the placebo.

A **double-blind experiment** is an experiment in which ........................................................................
                                                                                   ..........................................................
                                                                                   ..................................................................

In a randomized experiment, if there are large differences in outcomes among the treatment groups, we can conclude that the differences are due to the treatments.

A **confounder**, also called lurking variable, is a factor other than the treatment that can cause the treatment groups to have different outcomes.

Problems with confounders are prevalent in observational studies, which is why we prefer to conduct
........................................................................................................ when possible.

**Practice:** Find the confounders for the examples below

* a) Those people who tend to carry boxes of matches in their pockets also have a higher risk of cancer.

* b) Taller people tend to have higher salaries.

* c) The more ice-cream that is being consumed, the more drownings.

* d) A study in the 1960s reported a significant tendency for Down’s syndrome to be more common in fourth-born or higher order children.

**ex.** Suppose you design a study to determine if heavy drinkers die at a younger age. You proceed to design a study, and set about gathering data. Your results indeed show that people who drink excessively are likely to die younger.

**Can you think of any confounding variables that one should consider in this example?**

**Watch out!** Many media outlets jump upon sensational results, but often don’t pay any regard to the possibility of confounding variables. Next time, YOU should think about what the confounding variables could be when you hear about new research results.
Some say that numbers don’t lie, but the people who use them sometimes do. An example of a famous case of statistical fraud is the now-retracted British study that linked autism to childhood vaccines. This particular false study has done long-lasting damage to public health as many parents decided to not vaccinate their children based on this study alone.

Statistical graphs, when properly used, are powerful forms of communication. Unfortunately, when graphs are improperly used, they can misrepresent the data, and lead people to draw incorrect conclusions.

Below are two common forms of misrepresentation:

- Incorrect position of the vertical scale
- Incorrect sizing of graphical images

**Incorrect Position of the Vertical Scale**

With charts or plots that represent how much or how many of something, it may be misleading if the baseline is not at zero.

**The Area Principle**

When amounts are compared by constructing an image for each amount, the areas of the images must be proportional to the amounts. For example, if one amount is twice as much as another, its image should have twice as much area as the other image. When the Area Principle is violated, the images give a misleading impression of the data.

*Which of the images below best represent the difference of Jet Fuel cost?*