Learning Goals:

- Given a statistical research question, identify the population(s) and the variable(s).
- Categorize variables as quantitative or categorical.
- Distinguish between observational studies and experiments.

Introduction:

The goal of Statistics is to draw a conclusion about a large population based on a small sample. This is called *statistical inference.*

Up to this point in the course, we have learned many of the building blocks for statistical inference. For example, we know the difference between categorical and quantitative variables. This will be very important in identifying which statistical inference method to use.

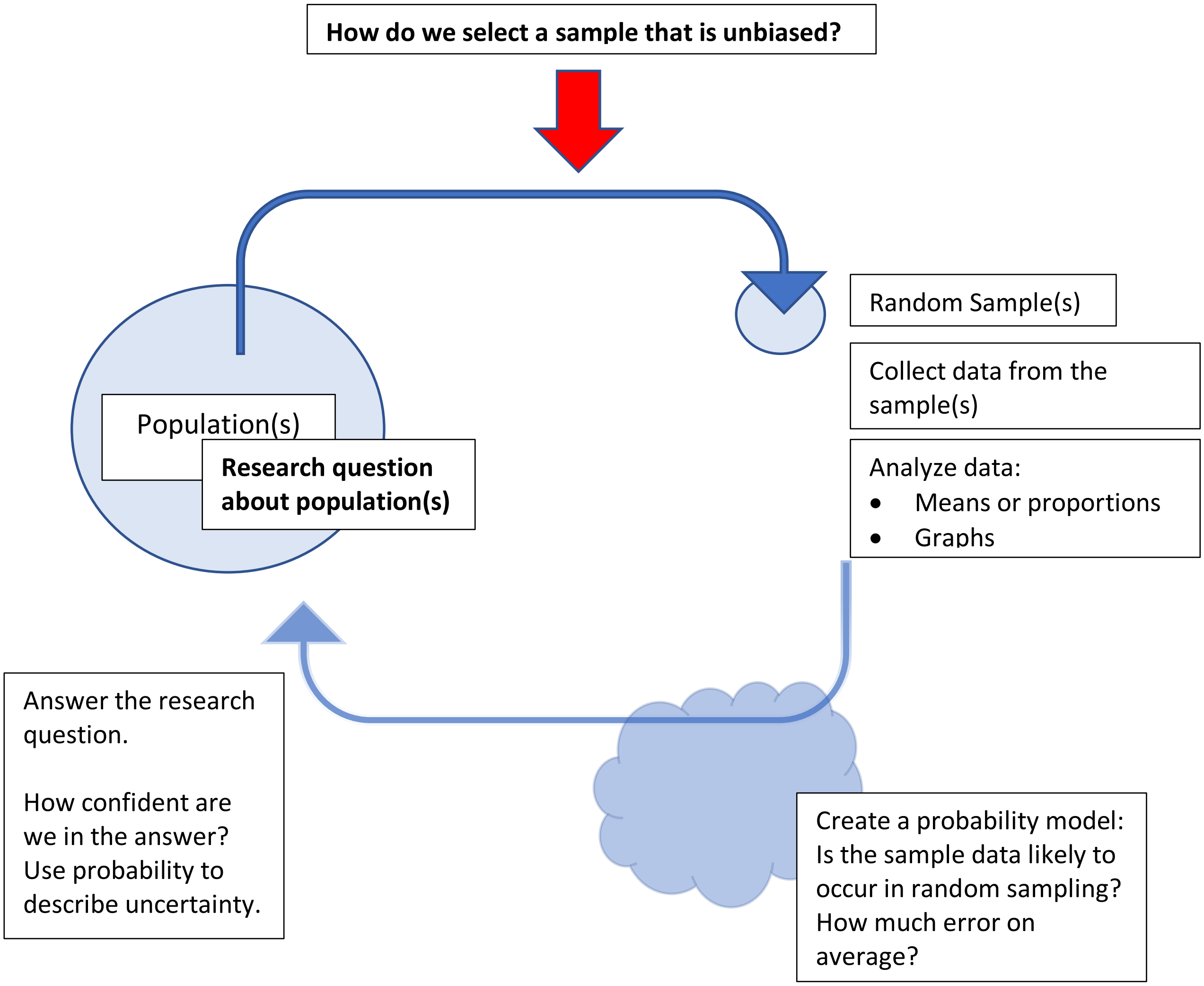
For each type of data (quantitative vs. categorical), we know how to summarize it using numerical measures (e.g. means and standard deviations for quantitative data or percentages for categorical data) and graphs (e.g. histograms or bar charts.) We know how to explore relationships between two variables using scatterplots, correlation, and linear regression if the variables are quantitative or using conditional percentages if the variables are categorical. We will continue to explore data and summarize it as part of statistical inference.

We have also learned about probability and probability distributions. Probability is based on relative frequency, which is the long-term frequency of an outcome in repeated random sampling. We recently finished learning about how to model probability distributions with a density curve to estimate probability with area under the curve. We will be using probability density curves with every statistical inference method.

In this Unit we tackle the last building block for statistical inference: data collection. In order for our conclusions about a population to be valid, we must have data that is not biased in any way. This Unit focuses on responsible data production.

The diagram on the next page ties all of these ideas together into the Big Picture of a Statistical Investigation.

Big Picture of a Statistical Investigation



How do we select a sample that is unbiased?

Random Sample(s)

Collect data from the sample(s)

Analyze data:

- Means or proportions
- Graphs

Create a probability model: Is the sample data likely to occur in random sampling? How much error on average?

Answer the research question.

How confident are we in the answer? Use probability to describe uncertainty.

Population(s)

Research question about population(s)

Example: Suppose that our research question is "What is the average amount of money that full-time LMC students spend on textbooks in a semester?"

In this example, the population is ___________________________________. To answer this question, we select a random sample from the population and ask each student in the

sample ____.

The data we collect is (circle one: categorical or quantitative) and we summarize it using (circle one: a mean or a percentage).

Eventually, we will learn to create a probability distribution and probability density curve that will help use determine if the sample we collected is likely to occur given our assumptions about the population. From this we will learn to draw a conclusion about the population.

Check your understanding:

1) Dissect the research questions to identify the population, variable and variable type.

Research question	**Population**	**Variable**	**Variable Type and Summary**
What is the average number of hours that community college students work each week?	Community college students	Number of hours a student works each week	Quantitative Use a mean
What proportion of all U.S. college students are enrolled in a community college?			
Is the average course load for a California community college student at least 12 units?			
Do the majority of LMC students qualify for federal student loans?			

2) The next set of research questions involve two variables. We use a categorical explanatory variable to create the comparison groups; we can think of these as separate populations. Individuals in the samples give information relevant to the response variable and this data is analyzed and compared. The response variable can be categorical or quantitative.

Research question	**Explanatory variable and comparison groups (two populations)**	**Response variable (data analyzed and compared); type and summary**
Are college athletes more likely to receive academic advising than students who are not athletes?	College Athlete? (Yes, No) Two populations: college athletes compared to non-athletes	Receive academic advising? (Yes, No) Categorical Compare proportions
In community colleges, do female students have a higher GPA than male students?		
Is chemotherapy or radiation a more effective treatment for shrinking the size of cancerous liver lesions?		
Are elementary school children who drink soda with their lunch more likely to be categorized as overweight?		

Observational Studies vs. Experiments

When the research question involves two variables, the goal of the study is to determine if the explanatory variable correlates with, or even causes, a change in the response variable.

In an *observational study*, the explanatory variable is some pre-existing characteristic that is used to divide the individuals into groups, e.g. gender or drinking soda at lunch. The researcher cannot control who is in each group. By contrast, in the *experiment*, the researcher assigns individuals to the groups and each group receives a different treatment, e.g. chemotherapy or radiation.

In both types of studies, the researcher compares the responses of the groups. In an observational study, researchers may take steps to reduce the influence of these other factors on the response, but it is difficult in an observational study to get rid of all the factors that may have an influence. For example, when examining medical records, researchers may remove people from a cancer study who have a family history of cancer, but there may be other factors affecting cancer rates, such as diet and exercise, that are not measured and cannot be accounted for or removed. For this reason, an observational study can provide evidence of an association between two variables, but it provides, at best, weak evidence of a cause-and-effect relationship. The observed association may be confounded by unmeasured variables.

Unlike an observational study, an experiment can provide evidence of a cause-and-effect relationship between the variables because the researchers can manipulate and control more of the *confounding variables* that might influence the response variable. In a well-designed experiment, they can conclude the differences in the response are due solely to the treatments they imposed.

Check your understanding:

3) In the 1980's doctors routinely prescribed hormone replacement therapy for women in menopause. A series of studies in the 1990's based on women's medical records showed that women taking hormone replacements also had a reduction in heart disease. But women who take hormones are different from other women. They tend to be richer and more educated, to have better nutrition, and to visit the doctor more frequently. These women have many habits and advantages that contribute to good health, so it is not surprising that they have fewer heart attacks.

    a) Are the 1990 studies observational studies or experiments? Why do you think so?

    b) In these studies what is the explanatory variable? What is the response variable? What are some of the confounding variables?

    c) Can we conclude from these studies that the hormones caused the reduction in heart attacks? Why or why not?

4) In 2002, the Women's Health Initiative sponsored a large-scale study to examine the health implications of hormone replacement therapy. In this study, researchers randomly assigned over 16,000 women to one of two treatments. One group took hormones. The other group took a *placebo*. A placebo is a pill with no active ingredients that looks like the hormone pill.

   The 2002 study was *double-blind*. *Blind* means that women did not know if they were receiving hormones or the placebo. *Double-blind* means that the information was coded, so researchers administering the pills did not know which treatment the women received.

   After 5 years, the group taking hormones had a *higher* incidence of heart disease and breast cancer. In fact, the differences were so significant that the researchers ended the study early. As a result, hormone replacement therapy is now rarely used.

   a) Is the 2002 study an observational study or an experiment? Why do you think so?

   b) What is the explanatory variable? What is the response variable?

   c) Explain how random assignment might control the effects of some of the confounding variables you identified in #3.

   d) What else did researchers do to control the effects of other variables on the response?

   e) Why did this study lead to the elimination of use of hormone replacement therapy despite the fact that the studies in the 1990's supported hormone replacement therapy?