We have studied primarily *probability* in the first two-thirds of this course. For the remainder of the course we will study *statistics*. Some fundamental terms/ideas include:

- **The observational units** in a statistical study are the objects described by a set of data (people, animals, things).
- A **variable** is any characteristic of an observational unit.
  - A variable can take different values (i.e., can *vary*) for different observational units.
  - Some variables are **quantitative**, taking *numerical* values on which ordinary arithmetic operations make sense.
    - A quantitative variable can be *discrete* (taking on finite or countably many values) or *continuous* (taking on any value in an interval).
  - Other variables are **categorical**, taking *category* designations.
    - A **binary** categorical variable has only two categories.

Much of statistics involves making *inferences* from the observed data to some larger entity.

- A **population** is the *entire* group of observational units about which information is desired.
  - A **process** can be thought of as an infinite population.
  - A **parameter** is a number that describes a *population* (or process, or probability distribution).
    - The value of a parameter is typically fixed but unknown.
- A **sample** is a (typically small) *part* of the population from which data are gathered.
  - A **statistic** is a number that describes a *sample*.
    - The value of a statistic, which varies from sample to sample, can be used to estimate the unknown value of a parameter.

Much of statistics involves making inferences about parameters based on statistics.

- With **categorical** variables, the simplest relevant parameter/statistic is a *proportion*.
  - Common symbols are $\pi$ (or $p$) for a population proportion, $\hat{p}$ for a sample proportion.
- With **quantitative** variables, the simplest relevant parameter/statistic is a *mean*.
  - Common symbols are $\mu$ for a population mean, $\bar{x}$ for a sample mean; $\sigma$ for a population standard deviation, $s$ for a sample standard deviation.

Two primary kinds of statistical inference procedures are:

- **Confidence intervals**, which estimate the value of a population parameter with an interval of values based on a statistic;
- **Hypothesis tests**, which assess the plausibility of, or assess the strength of evidence against, a particular claim about the value of the parameter.

In order to study confidence intervals for a population proportion, we present a version of the Central Limit Theorem that describes the (approximate) sampling distribution of a sample proportion:
Central Limit Theorem (CLT) for Sample Proportion:
Suppose that the proportion of a population having some characteristic is denoted by $\pi$, and suppose that a random sample of size $n$ is taken from the population. Then the sampling distribution of the sample proportion $\hat{p}$ is approximately normal with mean $\pi$ and standard deviation $\sqrt{\frac{\pi(1-\pi)}{n}}$. This approximation is generally considered to be valid as long as $n\pi \geq 10$ and $n(1-\pi) \geq 10$.

Example 14-1: Teen Media
The Kaiser Family Foundation commissioned an extensive survey in 2004 that investigated the degree to which American children aged 8-18 have access to various forms of media. They distributed written questionnaires to a random sample of 2032 children in this age range. One of the questions asked whether the child has a television in their bedroom, and 68% answered yes.

a) What are the observational units in this study?

b) Identify the variable, and classify it as quantitative or categorical.

c) Is .68 a parameter or a statistic? What symbol would you use to represent it?

d) Describe in words the relevant parameter of interest in this study. What symbol would you use to represent it?

e) Does the Kaiser survey allow them to determine the exact value of the parameter? Explain.

f) Is the population proportion more likely to be close to the Kaiser survey’s sample proportion than to be far from it? Explain.

From the CLT for a Sample Proportion and the empirical rule, we know that there’s about a 95% chance that the sample proportion $\hat{p}$ would fall within $2\sqrt{\frac{\pi(1-\pi)}{n}}$ of the population proportion $\pi$.

g) What’s wrong with creating an interval estimate of $\pi$ by taking $\hat{p} \pm 2\sqrt{\frac{\pi(1-\pi)}{n}}$?
h) What’s reasonable to use as a sample-based approximation for \( \frac{\pi(1 - \pi)}{n} \)?

- The estimated standard deviation \( \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \) of the sample statistic \( \hat{p} \) is called the **standard error** of \( \hat{p} \).

i) Calculate the standard error of \( \hat{p} \) for the Kaiser survey about the proportion of American children who have a television in their bedroom.

j) Calculate \( \hat{p} \pm 2 \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \) to produce a reasonable interval estimate of \( \pi \).

k) Do you know for sure whether the actual value of the population proportion \( \pi \) is in this interval?

The multiplier 2 is just an approximation. To produce a 95% confidence interval for the population proportion \( \pi \), we first find the \( z \)-score with the property that the area between it and its negative is .9500.

l) Use the normal probability table to determine this \( z \)-score. (As always, start with a sketch.)

m) Calculate \( \hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \) to produce another reasonable interval estimate of \( \pi \). Also interpret what this interval means.
• This interval is called a 95% confidence interval (CI).
  o The value 95% is called the confidence level.
  o The number .020 is called the margin-of-error.
  o The number \( z^* \) is called a critical value.

• More generally, a CI for population proportion \( \pi \) is given by:
  \[
  \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
  \]

  o Technical conditions:
    ▪ Simple random sample from population of interest
    ▪ \( n \hat{p} \geq 10 \) and \( n(1-\hat{p}) \geq 10 \)
  o The midpoint of this interval is the sample proportion \( \hat{p} \).
  o The width of this interval is the difference in the endpoints.
    ▪ The interval’s width equals twice the margin-of-error.
  o The margin-of-error depends on three factors:
    ▪ confidence level \( C \), which determines \( z^* \)
    ▪ sample size \( n \)
    ▪ sample proportion \( \hat{p} \)
  o Values of \( z^* \) for some common confidence levels are:

<table>
<thead>
<tr>
<th>Confidence level</th>
<th>80%</th>
<th>90%</th>
<th>95%</th>
<th>99%</th>
<th>99.9%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical value ( z^* )</td>
<td>1.282</td>
<td>1.645</td>
<td>1.960</td>
<td>2.576</td>
<td>3.291</td>
</tr>
</tbody>
</table>

  o Minitab: Stat> Basic Statistics> 1 Proportion...
  o Applet: Theory-Based Inference

n) How would you expect a 99% confidence interval to differ from the 95% interval? First explain intuitively, and then explain based on the formula.

o) Determine and interpret a 99% confidence interval for the population proportion of American 8-18-year-olds who have a television in their bedroom.

p) How does the 99% interval compare with the 95% interval? Compare their midpoints as well as their half-widths.

• As you increase the confidence level, the confidence interval gets wider.
q) Based on these intervals, does it seem plausible to say that 50% of all American teens have a television set in their bedroom? Does 75% seem plausible? What about two-thirds? Explain.

50%:  75%:  Two-thirds:

The survey also provided sample results broken down separately for boys and girls. There were 1036 girls and 996 boys in the sample. Among the boys, 72% had a television in their bedroom, as compared to 64% of the girls.

r) Determine a 95% confidence interval for the proportion of American boys in this age group who have a television in their bedroom. Then do the same for the analogous proportion of American girls. Also interpret each interval.

Boys:

Girls:

s) How do these intervals compare? Do they seem to indicate that the proportion of boys who have a television in their room is different than that proportion for girls? Explain.

t) Report the half-widths (margins-of-error) of these intervals. Are these larger or smaller (or the same as) the half-width based on the entire sample? Explain why this makes intuitive sense.

- A larger sample size produces a narrower confidence interval (smaller margin-of-error), whenever other factors remain the same.
u) Now suppose that you want to do a follow-up survey for which the margin-of-error (with 95% confidence) will equal .01 or less. How many people would you need to sample to achieve this margin-of-error? [Hints: Based on the original study, assume that the sample proportion with a television in their bedroom will be about .68. To be conservative you could use .5 as the estimate. Set the margin-of-error expression equal to .01, and solve algebraically for the sample size $n$.]

v) How does this sample size compare to the original one? Explain why this makes sense.

w) How does the population size enter into these calculations? (Hints: Think about the population here. How many American children aged 8-18 are there? Does this number enter into your calculations?)

Example 14-2:
What does it mean to be “95% confident”?
- Interpreting confidence intervals correctly requires us to think about what would happen if we took random samples from the population over and over again, constructing a CI for the unknown population parameter from each sample.
- The confidence level (e.g., 95%) indicates the percentage of samples that would produce a CI that successfully contains the actual value of the population parameter.

We will turn to an applet called “Simulating Confidence Intervals” to illustrate this. First we will make sure that the method is set to “Proportions” and “Wald.” We’ll also set the population proportion to be .45, the sample size to be 75, and the confidence level to be 95%.

a) As we take new samples, what do you notice about the intervals? Are they all the same?

b) Does the value of the population proportion change as we take new samples?

c) As we take hundreds and then thousands of samples and construct their intervals, about what percentage seem to be successful at capturing the population proportion?

d) Sort the intervals, and comment on what the intervals that fail to capture the population proportion have in common.
e) Now change the confidence level to 80%. Before pressing the Recalculate button, what changes do you expect to see? Then press the button. What two things change about the intervals?

f) Now change the confidence level to 99%. What two things change about the intervals?

g) Now change the sample size to 300 (still with a confidence level of 99%). Does this produce a higher percentage of successful intervals? What does change about the intervals?

h) Is it desirable to have larger (e.g., 99%) or smaller (e.g., 80%) confidence levels? Explain.

i) Is it desirable to have wider or narrower confidence intervals? Explain.

j) What’s a drawback of using a very high confidence level such as 99.9%?

k) What would it take to achieve the best of both worlds: a very high confidence and a very narrow confidence interval? Why is this so difficult to achieve?

l) Unlike in this exercise, in most situations you do not know the value of the population proportion, and you can only afford to take one sample. In such cases, do you know whether or not your confidence interval succeeds in capturing the unknown value of the population proportion? In what sense can you be confident that your interval succeeds?

Example 14-3: Female Senators
Suppose that an alien lands on Earth, notices that there are two different sexes of the human species, and sets out to estimate the proportion of humans who are female. Fortunately, the alien had a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval. Suppose that the alien happened upon the members of the March 2017 U.S. Senate as its sample of human beings, so it finds 21 women and 79 men in its sample.
a) Use this sample information to produce a 95% confidence interval for the actual proportion of all humans who are female.

b) Is this confidence interval a reasonable estimate of the actual proportion of all humans who are female?

c) Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.

d) It clearly does not make sense to use the confidence interval in (a) to estimate the proportion of women on Earth, but does the interval make sense for estimating the proportion of women in the March 2016 U.S. Senate? Explain your answer.

- Confidence intervals depend on having taken a random sample from the population.
  - Without a random sample, the sample may not be representative of the population.
  - A biased sampling procedure is one that systematically under- or over-represents certain segments of the population.
- Confidence interval are not appropriate (and not necessary) when you have access to the entire population.