In this lab, we'll explore sampling distributions using StatKey: www.lock5stat.com/statkey. We'll be using StatKey, rather than Minitab, for the next unit because it makes it very easy to simulate thousands of samples instantly while reinforcing conceptual understanding.

**Sampling Distribution for a Proportion: Red-Green Color Blindness**

Red-Green Color Blindness is when a person has difficulty distinguishing between red and green. It is a sex-linked trait, and is much more common in males than in females, because the gene associated with it is only on the X chromosome. Females have two X chromosomes, and are only color-blind if BOTH copies are defective, while males only have one X chromosome and so are red-green color-blind whenever their one X chromosome is defective.

Approximately 8% of all the associated alleles in Caucasians are defective, so approximately 8% of Caucasian men are red-green color-blind. (By the Hardy-Weinberg Principle, about 15% of Caucasian females are carriers, and only about 0.6% are red-green color-blind). We'll explore what kinds of sample proportions of red-green color blindness we might find for random samples of Caucasian males, given the true parameter of \( p = 0.08 \).

1. Go to StatKey and select Sampling Distribution for a Proportion.
2. Click on “Edit Proportion” to change the population proportion to 0.08, and click the box next to “Choose samples of size $n =$ “ and change the sample size to 100. This should update the original population proportion to show 0.08 on the right.

3. Click on Generate 1 Sample to draw one sample of size 100 from the population. You will see the results displayed under Sample, where it gives a frequency total and the sample proportion, $\hat{p}$. You will also see this sample proportion plotted with a dot on the big sampling distribution dotplot on the left. What proportion in your random sample of men happened to be red-green color-blind? Compare your answer with your neighbors – you likely got different results.

4. Repeat step c) a couple more times, generating one sample at a time, looking at the frequency table and sample proportion on the right, and then finding the corresponding dot on the sampling distribution. The point here is to become comfortable with the fact that each dot represents a **summary statistic from an**
**entire sample of size 100.** How much do statistics tend to be varying from sample to sample?

5. *Once you fully understand what is going on and what each dot in the sampling distribution represents* (discuss this more with your group or ask a TA/LA if you don’t fully understand – this is an important concept!), go ahead and click Generate 1000 Samples to automatically repeat the process 1000 times. This is simulating 1000 random samples, each of size 100, and plotting the sample proportion for each on the sampling distribution. What’s the farthest value you got from the truth?

6. Based on the sampling distribution, what do you think is a good distance such that it is rare for statistics to be farther than that from the parameter? (This will help you to know how far the parameter might be from a statistic, which is what you will more realistically want to know). (Answer this informally here, without specific formulas but just by looking at the picture)

7. The summary statistics for the sampling distribution are shown in the top right of the dotplot. This gives the number of samples simulated (samples), the mean of the sampling distribution, and the standard deviation of the sampling distribution.
   a. What is the mean of the distribution? Why?
   b. What is the *standard error* of the sample proportion?

8. Generate a few thousand more simulated samples. How does this change the mean? How does this change the standard error? (Hint: any changes should be only very minor!)

9. Change the sample size, *n*, from 100 to 1000. BEFORE generating new samples, predict what will change about the distribution. Start generating a few samples at a time, observe how much they are varying, and then generate 1000 samples.
a. What is the farthest from the truth?
b. What is the mean?
c. What is the standard error?
d. How do these numbers compare to the sampling distribution when \( n \) was 100? Discuss the effect of sample size on the sampling distribution with your group.

10. So far we’ve been looking only at Caucasians, but the prevalence is 5% in men of Asian origin and 4% in men of African origin. Edit the population proportion to reflect this new population, and recreate the sampling distribution for each population (you can leave the sample size at 1000). What is the mean for each? Why?

**Sampling Distribution for a Mean: Health Care Expenditure**

Here we’ll explore the sampling distribution for a mean by looking at health care expenditures for each country (as a percentage of all government expenditures). You can access the data (AllCountries) from [www.lock5stat.com/datapage.html](http://www.lock5stat.com/datapage.html) (.xls or .csv).

11. Click (or double click) on the downloaded data file to open it in Excel or any other spreadsheet program. Cut (highlight and CTRL+x) the Health column and insert it to the right of the country names (right click on the column where you want it and choose “Insert Cut Cells”). (Note: you don’t really have to do this every time, but it’s useful and more interesting to see the country names with the health expenditure values.) Highlight the columns titled “Country” and “Health” and copy them (to later be pasted into StatKey).

12. Go back to the StatKey homepage by clicking on the StatKey icon at the top, and switch to sampling distribution for a mean.

13. Click on “Edit Data,” highlight and delete all the existing data, and paste in the health data that you copied from your spreadsheet. Click on “Show Data Table” to see the data. What percent of government expenditures in the US (or your home country) are spent on health care?

14. The data is displayed with a dotplot in the top right (under Population), and you can see the corresponding relevant population parameters.
   a. How many countries are in this dataset?
   b. What is the mean health care expenditure?
   c. Describe the shape of the distribution.

---

1 This data includes some missing values, so is not really data on the entire population of all countries, but it is close, and for the purposes of this lab we’ll pretend it represents data on the entire population of all countries.
15. The default sample size is $n = 10$, and we can keep this. Now the process is the same as before – click Generate 1 Sample to draw a random sample of 10 countries from the population. Like before, the sample data will be displayed under Sample and the sample statistic (now the mean) will be plotted with a dot on the sampling distribution.

d. Click on “Show Data Table” next to Sample to see which countries you randomly selected to be in your sample.

e. What is the sample mean of your sample?

f. How do the summary statistics (mean, median, standard deviation) of the random sample compare to the parameters in the population?

16. Repeat step d several times. Try to get a feel for the following:

g. How much do sample means tend to be varying from sample to sample?

h. How do the sample statistics (random) relate to the population parameters (fixed)?

17. Generate 1000 samples to get a sampling distribution for the mean.

i. What is the mean of the sampling distribution? Why?

j. What is the standard error of the sample mean?

k. Describe the shape of the distribution.

l. What is the farthest value from the truth?

m. Try hovering your mouse over any of the dots to see the sample it came from. Remember – each dot is a **statistic from a sample**!

18. How does the standard deviation of the **statistic** (the standard error) compare to the standard deviation of the actual **data**? It’s important to keep these two separate in your minds: in the right panels each dot corresponds to a **single country**, while in the sampling distribution each dot is the mean for 10 countries. Although both distributions have a standard deviation, they are measuring standard deviation of two very different distributions/quantities. **Discuss this more with your group/TA/LA if this is unclear – this is an important concept.**

19. Select one last random sample, and suppose you only knew the sample mean from those 10 countries, not the population mean. That’s your best guess, but how much uncertainty lies in this estimate? How far might the parameter be from the statistic? Use statistic $\pm 2*SE$ to generate a 95% confidence interval and interpret this confidence interval in context. [Check your interpretation with a TA/LA if you have time.]