Lab 1: The binomial test

This first lab focuses on the binomial test, but you should also use it as an opportunity to (re-)familiarize yourself with `R`. If you have any questions about using `R` as you work through the lab, be sure to ask!

A few notes on working with `R`

One tip to make working with `R` easier is to use a script file. A scriptfile is a file where you type your `R` commands as you work through an analysis. To make a script file for this lab, open a notepad file and save it to your computer as `Lab1Script.txt`. Type your commands into the script file instead of directly into `R`. Then, when you want to run a command (or group of commands), simply copy and paste into `R`. This way you can easily make changes to your code and re-run it without have to retype the whole thing.

The second thing you will want to do is make a results file for this lab. For each numbered question below, answer in your results file. Write in complete sentences and keep your results file neat! Also, be aware there is a "Lab Summary" at the end of this lab. Complete this section as well. **AT THE END OF YOUR DOCUMENT, IN AN APPENDIX, INCLUDE YOUR R CODE AND ANY ASSOCIATED FIGURES/CHARTS/GRAPHS. DO NOT INCLUDE YOUR CODE FOR EACH PROBLEM WITH YOUR RESPONSE TO EACH PROBLEM.**

The binomial test

The first part of this lab is to work through the binomial test by hand.

Problem: An anti-smoking organization is investigating a claim that smokers start smoking before the age of 18. To investigate this claim, they take a random sample of smokers and ask them the age they started to smoke. They gather the following data:

18 19 30 16 17 15 14 14 17 12 14 13 19 19 17 13 20 12 17 15

They have hired you to do their data analysis:

1. Perform a one-sample t-test (by hand) to test the claim. Be sure to include all the steps in your test and clearly state a conclusion about the result.

   Hint: In order to easily calculate the mean/sd of the data, you can enter data into `R` using the `c()` function:

   ```
   smokeage = c(18,19,30,16,17,15,14,14,17,12,14,13,19,19,17,13,20,12,17,15)
   ```

2. Check the assumption of normality by making a histogram of `smokeage`. Describe the distribution of the age when people start to smoke and comment on whether you think the normality assumption is met.

3. Perform a binomial test (by hand) to test the claim. Be sure to include all the steps in your test and clearly state a conclusion about the result.

4. Comment on the difference in results between the t-test and the binomial test. Which test do you feel more comfortable presenting to the organization? Why?

The binomial test in R

We can let R do the calculations for the binomial test for us. The function that does this is the `sign.test()` function from the `BSDA` (Basic Statistics and Data Analysis) library. In order to use a function from a different library, we must first load the library into R. The first time we want to use the library, we will also have to install the package. After the package is installed, we will only have to load the library to use it. (You can use this same procedure to install the `BSDA` package on your personal computer.)

Install the `BSDA` package:

- Under the **Packages** menu option, select **Install Package(s)**.
- You'll be asked to select a mirror for downloading. I usually choose **USA(MD)**.
- Next you select the package you want to install. We want **BSDA**.
- R should now automatically install the package.

After the `BSDA` package has been installed, load the package by typing `library(BSDA)` at the `>` prompt. Now you are ready to use `sign.test()`.

5. Perform a binomial test on the smoking data using the following command:

   ```
   sign.test(smokeage, md=18, alternative="less")
   ```

   Verify that R gives you the same p-value and conclusion that you calculated by hand.

6. Look closely at the output from this test. You will see that R gives you three different confidence "intervals". Since you performed a one-sided test, you are really getting confidence bounds, rather than confidence intervals. Explain what it means to have a 95% upper bound of 17.21.

7. It we want to change these confidence bounds to confidence intervals, we need to change the alternative to `"two.sided"`. We can also change the confidence level to 99% by adding the argument `conf.level = .99`. Make these two changes and report and interpret the 99% confidence interval for the median.

The calculations for these confidence intervals are complicated to do manually. You'll notice that of the three confidence intervals given, only one has 99%confidence. The others are slightly higher and slightly lower than the 99% you requested. This is because the binomial distribution is discrete, and with a discrete distribution it isn't always possible to calculate your confidence level exactly. The confidence interval with exactly 99% confidence is actually a non-linear interpolation of the two non-exact intervals. When reporting a confidence interval, you may choose to report either of the three intervals, so long as you link it with the appropriate level of confidence.

8. Before moving to the next section, play with the `sign.test()` function a little bit more. Check out the help function with `?sign.test`. Then use `R` to repeat the test three times using a different alternative (`"less"`, `"greater"`, `"two.sided"`) each time. How does your conclusion change for each of these three hypotheses? Do the results reconcile with each other?

Comparing the binomial test to the t-test

You are going to compare the two tests on two datasets – data.symm and data.skew. You can find these two data sets on Collab. Note that both data sets have 50 observations, which means that the assumptions for the t-test are met.

9. For the `data.symm`, you should do the following:
   - Make a histogram of the data. Comment on the shape of the distribution. Specifically, comment on the relationship between the mean and the median for a population with that shape.
   - Perform a t-test of H_0: $\mu = 8$ versus H_1: $\mu > 8$. State your p-value and a conclusion. [Hint: You can use the `t.test()` function in `R` to perform your t-test. The `t.test()` function is standard in `R` – no need to load an additional library. It works much the same as `sign.test()`, for example you can confirm your t-test analysis in 1 with the command `t.test(smokeage, mu = 18, alternative="less")`.]
   - Perform a binomial test of H_0: $\theta_{.5} = 8$ versus H_1: $\theta_{.5} > 8$. State your p-value and a conclusion.
   - Compare the results of the two tests.
10. Repeat the steps in 9 for `data.skew`.

The main thing to keep in mind is that although both tests are one-sample tests, they are NOT testing the same thing. The t-test is a test of the mean, while the binomial test is a test of the median. In a skewed distribution, the mean and median are not equal, so a test might find evidence about the mean but not find the same evidence about the median. This is why if we are really going to compare the behavior of these two tests, we need to compare them on populations that are symmetric for the comparisons to make sense.

Lab Summary

Write a paragraph summarizing your findings about the age at which people start to smoke. Write this paragraph as though your audience is the anti-smoking organization that hired you. Present the results to both of your analyses (the t-test and the binomial test) and explain why the binomial test has more trustworthy results. Although you may use the term p-value in your summary, be sure that you are also explaining your findings so that your non-statistically oriented employer can understand your results.