LEARN BY DOING: Checking Conditions for Hypothesis Testing for the Population Mean

The purpose of this activity is to discuss how in some cases exploratory data analysis can help you determine whether the conditions that allow us to use the *t*-test for the population mean (μ) are met.

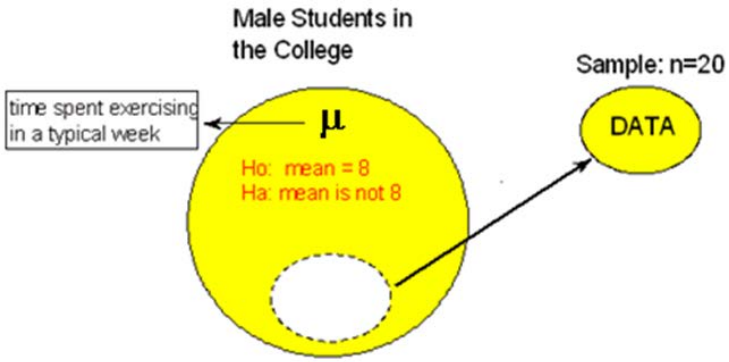
Background:

- In the Exploratory Data Analysis unit, we stressed that in general, it is always a good idea to **look at your data** (if the actual data are given).
- Moreover, related to our discussion now, looking at the data can be very helpful when trying to determine whether you can reliably use the test.
- Often in courses, data summaries (sample size, sample mean) are given rather than the raw data. We must trust what we are told in the problem. In practice, you often work with the raw data.

Now imagine the following situation:

- A health educator at a small college wants to determine whether the exercise habits of male students in the college are similar to the exercise habits of male college students in general.
- The educator chooses a random sample of 20 male students and records the time they spend exercising in a typical week.
- Do the data provide evidence that the mean time male students in the college spend exercising in a typical week differs from the mean time for male college students in general (which is 8 hours)?

NOTE: Whether σ is known or not is really not relevant to this activity.



Here is a situation in which we do not have any information about whether the variable of interest, "time" (time spent exercising in a typical week) varies normally or not, **and** the sample size (n = 20) is not really large enough for us to be certain that the Central Limit Theorem applies.

Recall from our discussion on the Central Limit Theorem that unless the distribution of "time" is extremely skewed and/or has extreme outliers, a sample of size 20 should be fine.

However, how can we be sure that is, indeed, the case?

- **If only the data summaries are given, there is really not a lot that can be done. You can say something like: "I'll proceed with the test assuming that the distribution of the variable "time" is not extremely skewed and does not have extreme outliers."**

- **If the actual data are given, you can make a more informed decision by looking at the data using a histogram. Even though the histogram of a sample of size 20 will not paint the exact picture of how the variable is distributed in the population, it could give a rough idea.**

Now we will look at a few different samples representing data in this scenario.

The data are available if you wish to create the graphs yourself but we also provide the graphs directly if you wish to skip analyzing the data yourself for this activity.

- **DATA:** EXCEL format, CSV format
- **OUTPUT:** GRAPHS (PDF)

For each sample, view the histogram and QQ-plot (or create them yourself). Comment on whether you think it would be safe to proceed with the test had those been the actual data in the problem above.

CHECK ANSWER