Unit 6 Homework: Tests and and Confidence Intervals

w203: Statistics for Data Science

Low-Oxygen Statistics

The file `expeditions.csv` contains data about 10,000 climbing expeditions in the Himalayan Mountains of Nepal. The data was compiled by the Himalayan Database and published in csv format on Tidy Tuesday.

First, navigate to https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-09-22 to read some basic information about the data and examine the codebook.

The variable `highpoint_metres` represents the highest elevation reached by each expedition. Your task is to test whether the mean highest elevation is above 7400 meters.

a. Using the documentation about the data, your background knowledge, and the data itself, assess whether the assumptions underlying a valid t-test are met. If plots are useful to make this argument, include them; if numeric statements are useful to make this argument, use them.
b. Provide an argument for why you should conduct a two-tailed test in this case, even though your personal interest is primarily in whether the mean is higher than 7400.
c. Compute the t-statistic by plugging in the values from the data manually into the formula. A *great* solution would write a function (perhaps called `t_statistic`) that takes arguments and returns a value. However, writing a function isn't necessary for a full solution. Feel free to use functions `mean()`, `sd()`, and `sqrt()`.
d. Using `qt()`, compute the t-critical value for a two-tailed test.
e. Compute the p-value for your two-tailed test. You may use the `pt()` function.
f. Explain what your rejection decision should be in two ways.
g. Confirm that your work is correct, by running the `t.test` command.
h. Evaluate the practical significance of your result.

Did You Say T as in Zebra? :zebra:

You record the total amount spent on California Avocados for a sample of 15 grocery store shoppers at the Berkeley Bowl. You compute a 95% confidence interval for the mean, but using a normal distribution instead of a t-distribution. In other words, you compute,

$$\left(\left[\overline{X} - \Phi^{-1}(.975)\frac{s}{\sqrt{n}}\right], \left[\overline{X} + \Phi^{-1}(.975)\frac{s}{\sqrt{n}}\right]\right)$$

Where $\overline{X}$ is the sample mean, s is the sample standard deviation, and Φ^{-1} is the quantile function (`qnorm` in R) for the standard normal distribution. The result this process returns using the normal distribution is $[50, 70]$.

a. What is your sample mean and sample standard deviation? For sample standard deviation you can use the square-root of the *Unbiased Sample Variance* provided in **Definition 3.2.20**.

b. If you had correctly used a t-distribution, rather than the normal distribution, what would your confidence interval have been?

c. What is the true confidence level for the interval you computed with a normal distribution? In other words, what fraction of confidence intervals constructed in this way would include the mean?

Note: Maximum score on any homework is 100%