Analytics Foundations: Problem Set 2

1. You've been asked to explore differences between rates of death from Covid in the east vs. the west. You download some data (*covidPerCapita)* from the CDC that has Covid death rates per capita (**covidDeathsPerCapita**) by zip code (**zip**). Not all zip codes are reporting, and even handling the data from those that *do* report is difficult so you select a random sample of zip codes in the east and the west.

    a. Conduct a t-test for the difference of means in per capita deaths from Covid in the east **region** vs the west **region.** Don't forget to verify assumptions. Note anything of interest to you, including the results of any tests performed, violations of assumptions, etc.

The assumption of independence is reasonable and the deaths per capita appear to be normally distributed within the east and within the west. Looking to see if there is any significant difference in the average number of deaths within these two regions, we have a p-value less than 2.2 x10-16, which means we are able to reject the null hypothesis and conclude that there does indeed appear to be a significant difference in the average number of deaths per capita from the East to the West.

    b. Does there appear to be a difference?

Yes, see answer from a.

    c. Can you conclude from this analysis that individuals in one region have had higher death rates than individuals in the other region? Reflect carefully on the analysis we've done. Discuss any problems you find with this methodology and anything that you'd like to do differently.

The test that was conducted was a two tailed test (NOT directional!!). If the researcher was interested in directional, then this should have been specified and that would have been the test conducted (you are data snooping!! And would need to do another hypothesis test…we will talk about what happens to the type 1 error rate when you do multiple hypothesis tests a little later).