BACS HW - Week 5

Question 1) *The following problem's data is real but the scenario is strictly imaginary*
The large American phone company Verizon had a monopoly on phone services in many areas of the US. The New York Public Utilities Commission (PUC) regularly monitors *repair times* with customers in New York to verify the quality of Verizon's services. The file `verizon.csv` has a recent sample of repair times collected by the PUC.

a. Imagine that Verizon claims that they take 7.6 minutes to repair phone services for its customers on average. The PUC seeks to verify this claim at 99% confidence (i.e., significance α = 1%) *using traditional statistical methods*.
   i. Visualize the distribution of Verizon's repair times, marking the mean with a vertical line
   ii. Given what the PUC wishes to test, how would you write the hypothesis? *(not graded)*
   iii. Estimate the population mean, and the 99% confidence interval (CI) of this estimate.
   iv. Find the t-statistic and p-value of the test
   v. Briefly describe how these values relate to the Null distribution of t *(not graded)*
   vi. What is your conclusion about the company's claim from this t-statistic, *and why*?
b. Let's re-examine Verizon's claim that they take no more than 7.6 minutes on average, but this time using bootstrapped testing:
   i. *Bootstrapped Percentile:* Estimate the bootstrapped 99% CI of the population mean
   ii. *Bootstrapped Difference of Means:*
   What is the 99% CI of the bootstrapped difference between the sample mean and the hypothesized mean?
   iii. Plot distribution the two bootstraps above on two separate plots.
   iv. Does the bootstrapped approach agree with the traditional t-test in part [a]?
c. Finally, imagine that Verizon notes that the distribution of repair times is highly skewed by outliers, and feel that testing the mean in not fair because the mean is sensitive to outliers. They claim that the median is a more fair test, and claim that the median repair time is no more than 3.5 minutes at 99% confidence (i.e., significance α = 1%).
   i. *Bootstrapped Percentile:* Estimate the bootstrapped 99% CI of the population median
   ii. *Bootstrapped Difference of Medians:*
   What is the 99% CI of the bootstrapped difference between the sample median and the hypothesized median?
   iii. Plot distribution the two bootstraps above on two separate plots.
   iv. What is your conclusion about Verizon's claim about the median, and *why*?

(continued on next page)

Question 2) Load the compstatslib package and run `interactive_t_test()`. You will see a simulation of null and alternative distributions of the t-statistic, along with significance and power. If you do not see interactive controls (slider bars), press the gears icon (⚙) on the top-left of the visualization.

Recall the form of t-tests, where $t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}}$

Let's see how hypothesis tests are are affected by various factors:
$diff$: the difference we wish to test $(\bar{x} - \mu_0)$
sd: the standard deviation of our sample data (s)
n: the number of cases in our sample data
$alpha$: the significance level of our test (e.g., alpha is 5% for a 95% confidence level)

Your colleague, a data analyst in your organization, is working on a hypothesis test where he has sampled product usage information from customers who are using a new smartwatch. He wishes to test whether the mean (x_i) usage time is higher than the usage time of the company's previous smartwatch released two years ago (μ_o):

> H_{null}: The mean usage time of the new smartwatch is the same or less than for the previous smartwatch.
> H_{alt}: The mean usage time is greater than that of our previous smartwatch.

After collecting data from just n=50 customers, he informs you that he has found diff=0.3 and sd=2.9.
Your colleague believes that we cannot reject the null hypothesis at alpha of 5%.

Use the slider bars of the simulation to the values your colleague found and confirm from the visualization that we cannot reject the null hypothesis. Consider the scenarios (a – d) independently using the simulation tool. For each scenario, start with the initial parameters above, then adjust them to answer the following questions:

i. Would this scenario create systematic or random error (or both or neither)?
ii. Which part of the t-statistic or significance (diff, sd, n, alpha) would be affected?
iii. Will it increase or decrease our *power* to reject the null hypothesis?
iv. Which kind of error (Type I or Type II) becomes more likely because of this scenario?

a. You discover that your colleague wanted to target the general population of Taiwanese users of the product. However, he only collected data from a pool of young consumers, and missed many older customers who you suspect might use the product *much less* every day.
b. You find that 20 of the respondents are reporting data from the wrong wearable device, so they should be removed from the data. These 20 people are just like the others in every other respect.
c. A very annoying professor visiting your company has criticized your colleague's "95% confidence" criteria, and has suggested relaxing it to just 90%.
d. Your colleague has measured usage times on five weekdays and taken a daily average. But you feel this will underreport usage for younger people who are very active on weekends, whereas it over-reports usage of older users.