Homework 3

Statistical Inference, Fall 2021

1- For each following statement, determine if there is any problem fix it.

a. In the CLT we are looking for a confidence interval for the sample mean.

b. With the same null hypothesis, it's always possible that we reject 2 side tests while 1 side test doesn't reject.

c. The central limit theorem states that the sampling distribution of sample means will closely resemble the normal distribution regardless of the sample size.

d. For a positively skewed distribution, the mean usually has a larger value than either the median or the mode.

e. For a given standard error, lower confidence levels produce wider confidence intervals.

f. The statement, "the 95% confidence interval for the population mean is (350, 400)", is equivalent to the statement, "there is a 95% probability that the population mean is between 350 and 400".

g. A 95% confidence interval obtained from a random sample of 1000 people has a better chance of containing the population percentage than a 95% confidence interval obtained from a random sample of 500 people.

h. Suppose we have constructed the following confidence interval about the mean age of freshmen college students in a State: $18.4 < \mu < 21.5$ The proper interpretation is that we are 95% confident that the sample mean is in the range 18.4 and 21.5 years.

i. A confidence interval is an estimate for which there is a specified degree of certainty that the population parameter will be in the interval.

j. The larger the difference between the actual value of the population mean and the value hypothesized, the larger beta will be.

k. A researcher cannot commit type I and type II errors at the same time, on the same hypothesis test.

l. Even though you rejected the null hypothesis, it may still be true.

m. The first type of error in using a "corona thermometer test" at the public place entrance is more important than the second type of error. with the ground assumption that everyone is healthy and we are looking for people who are suspected of having the coronavirus.

2- A student visits her doctor with concerns about being infected by covid-19. If the forehead temperature exceeds 38.2°C, the person is suspected that infected with the coronavirus. A patient's forehead temperature measure often has a considerable variation during a given day. Suppose an adult's forehead temperature measures during a given day has a normal distribution with a mean of 37.1°C and a standard deviation of 0.5°C.

a. What is the probability that a single forehead temperature measurement will fail to detect that the patient suspected coronavirus infection?

b. If five forehead temperature measurements are taken at various times during the day, what is the probability that the average of the five measurements will be less than 36.0°C and hence fail to indicate that the patient suspected coronavirus infection?

c. **(R)** How many measurements would be required in a given day so that there is at most 1% probability of failing to detect that the patient suspected coronavirus infection?

3- A game of rolling 6-sided dice will be held which every person could roll one 10 times. X is the total value of all 10 dice and is equal to $X_1 + X_2 + \cdots + X_{10}$. For appointing the winner, calculate the sum of each person's total value of all 10 dice. If the person's total value is in $X \leq 25$ or $X \geq 45$, she/he will win. Use the central limit theorem to calculate the probability that each person wins.
Recall that $E[X_i] = 3.5$ and $Var(X_i) = 35$.

4- London's mean high temperature in July is 72˚F with the standard deviation is 6˚F, while in January high temperatures average is 38˚F with a standard deviation of 12˚F. In which month is it more unusual to have a day with a high temperature of 45˚F? Explain.

5- You have a set of water bottles whose volume is known to follow a N(μ, 25) distribution. For having a 98% confidence interval for μ with width 2, how many bottles do you have to sample for their volumes?

6- It has been reported that the time which people spend searching, finding, and completing a purchase order at an online shopping follows a normal distribution. A research has been done with a random sample of 35 of the purchase process that gives a sample mean of 14 minutes with a sample standard deviation of 5 minutes.

a. Give a 95% confidence interval for the true average time per purchase process.
b. What is the probability that the estimate of 14 minutes differs from the true average by 2 minutes or more?
c. With the current level of communication resources for this online shopping and their projected growth over the next 8 months, the shopping store will be able to provide satisfactory service if the average purchase process time per customer is no more than 12.5 minutes. Based on the random sample of 45 purchase processes would you recommend that the shopping store upgrades its communication resources? (Perform a one-sided test at a 5% significance level.)

Homework 3

Statistical Inference, Fall 2021

7- Remy Ratatouille (master chef in the "Ratatouille" animation) wishes to estimate the mean number of rejected meals reported per weekend at Gusteau's restaurant. From a sample of 50 weekends, Remy found 2 weekends with 7 occurrences, 18 weekends with 9, 16 weekends with 10, and the remainder with 11 reported occurrences. Remy wants to check whether the average number of rejected meals per weekend has changed due to the new "meal in the menu (Ratatouille)" he helped initiate last year. The average number of rejected meals per weekend before the change on the menu was 11.5.

   a. Based on the information collected from Anton Ego (a cutthroat food critic), a star was added to the restaurant stars because of a change in menu. Do you agree with this rising? (Answer this question using a one-sided test at the 5% significance level – you will need to calculate the sample standard deviation of your estimate.)
   b. What is the p-value for your test?

8- An insurance company wants to audit health insurance claims in its very large database of transactions. In a quick attempt to assess the level of overstatement of this database, the insurance company selects at random 400 items from the database (each item represents a dollar amount). Suppose that the population mean overstatement of the entire database is \$8, with a population standard deviation of \$20.

   a. Find the probability that the sample mean of the 400 would be less than \$6.50.
   b. The population from where the sample of 400 was selected does not follow the normal distribution. Why?
   c. Why can we use the normal distribution in obtaining an answer to part (a)?
   d. For what value of ω can we say that $P(\mu - \omega < \overline{X} < \mu + \omega)$ is equal to 80%?
   e. Let T be the total overstatement for the 400 randomly selected items. Find the number b so that $P(T > b) = 0.975$.

9- **(R)** According to the patient image dataset[1], answer the following questions:

   a. Choose two variables (X as numeric & Y as categorical) and construct a 95% confidence interval for the mean of the X column between categories of the Y column.
   Is there a statistically significant between the means, change X, Y?
   b. Choose two variables (X & Y as categorical) and construct a 99% confidence interval for the proportion of the X column between categories of the Y column.
   Is there a statistically significant between the proportions, change X, Y?

10- **(R)** According to Exercise 7, we are ambiguous about the actual value between 9,10,11,12,13, but we know σ=0.7. Answer the following questions by computing power.

   a. When do we use the power curve? What is the ability of that? (explain)
   b. Draw the power curve of different μ_a.
   c. Choose the most reliable μ_a and plot distributions of "$\bar{x}\ under\ \mu_0$" versus "$\bar{x}\ under\ \mu_a$". Specify the acceptance region, $\alpha, \beta, power$ on the plot.

[1] https://www.kaggle.com/bachrr/covid-chest-xray

Homework 3

Statistical Inference, Fall 2021

11- **(Bonus)** Answer the following question:

a. Explain “what-if study” and “Double what-if study” with an example.

b. List Hill’s nine items and briefly explain each of them.

c. According to Hill’s nine items, answer the following statements about evidence of cause and effect correspond.

i. On November 9, 2020, Pfizer reported early results from a randomized controlled trial of their coronavirus vaccine based on messenger RNA technology. The report stated that the vaccine appears to be at least 90% effective. To which of Hill’s nine items does this evidence of cause and effect correspond?

ii. At the same time, Moderna was developing a coronavirus vaccine also based on messenger RNA technology. Which of Hill’s nine items suggested that the Moderna vaccine would also be effective?

iii. Lee and Skerrett (2001) found that higher levels of physical activity are more strongly protective against all-cause mortality. Which of Hill’s nine items does this evidence of causation address?