BAX-400 Homework 3

Mehul Rangwala

Summer 2022

Assignment Weight: 12.5% of your grade

Due Date: Sunday, August 28, 2022 11:59 PM

Instructions

1. There are 12 questions, some having multiple parts.
2. All questions should be completed using `R`.
3. Please complete each question for full or partial credit.
4. Submit an RMD file knitted as HTML. The knitted file should show the code and the result (output) below it.
5. For hypothesis testing questions, the hypotheses should be written formally (with symbols and signs and NOT words.) For example, you should **NOT** write the hypotheses as:

$$H0 : mu >= \$340$$
$$H1 : mu < \$340$$

   but write them as:

$$H_0 : \mu \geq \$340$$
$$H_1 : \mu < \$340$$

6. For confidence interval questions, you might be asked to provide insights and interpretations. They are not the same. Here is an example of the difference between insights and interpretations.

Example Interpretations and Insights

Example of an interpretation: We are 95% confident that the interval between \$425.50 and \$632.40 contains the true population mean amount of commissions earned per day.

Examples of insights:

- We see from the confidence intervals that the drivers 1 and 2 have the lowest delivery times.
- Drivers 1 and 3 have the lowest delivery speeds.
- We see from the confidence intervals that the majority of the applicants who enroll have low median scores.

7. Email me if you have any questions. mrangwala@ucdavis.edu

Question 1 *(5 points)*

A special industrial battery must have a life of at least 400 hours. A hypothesis test is to be conducted with a 0.02 significance level. If the batteries from a particular production run have an actual mean use life of 385 hours, the production manager wants a sampling procedure that only 10% of the time would show erroneously that the batch is acceptable. What sample size is recommended for the hypothesis test? Assume 30 hours as an estimate of the standard deviation.

Question 2 *(5 points)*

The values listed in the CSV file *WaitingTimes* are waiting times (in minutes) of customers at the Bank A, where customers enter a single waiting line that feeds three teller windows. Construct a 95% confidence interval for the population standard deviation σ.

Also available on the same CSV file are the waiting times (in minutes) of customers at the Bank B, where customers may enter any one of three different lines that have formed at three teller windows. Construct a 95% confidence interval for the population standard deviation σ.

Interpret the results found. Do the confidence intervals suggest a difference in the variation among waiting times? Which arrangement seems better: the single-line system or the multiple-line system?

Assume that each sample is a simple random sample obtained from a population with a normal distribution.

Question 3 *(5 points)*

You wish to estimate, with 95% confidence, the population proportion of U.S. adults who think that the president can do a lot about the price of gasoline. Your estimate must be accurate within 4% of the true population proportion.

a) No preliminary estimate is available. Find the minimum sample size needed.

b) Find the minimum sample size needed, using a prior study that found that 35% of U.S. adults think the president can do a lot about the price of gasoline. (Source: CBS News/New York Times Poll)

c) Compare the results from parts a) and b) above.

Question 4 *(30 points)*

Williamson University is a liberal arts university in the Southeastern U.S. that attempts to attract the highest-quality students, especially from its region of the country. It has gathered data on 178 applicants who were accepted by Williamson (a random sample from all acceptable applicants over the past several years). The data are in the CSV file *Admissions*. The variables are as follows:

- Accepted: whether the applicant accepts Williamson's offer to enroll
- MainRival: whether the applicant enrolls at Williamson's main rival university
- HSClubs: number of high school clubs applicant served as an officer
- HSSports: number of varsity letters applicant earned
- HSGPA: applicant's high school GPA
- HSPctile: applicant's percentile (in terms of GPA) in his or her graduating class
- HSSize: number of students in applicant's graduating class
- SAT: applicant's combined SAT score
- Combined Score: a combined score for the applicant used by Williamson to rank applicants

The derivation of the combined score is a closely kept secret by Williamson, but it is basically a weighted average of the various components of high school performance and SAT. Williamson is concerned that it is not getting enough of the best students, and worse yet, that many of these best students are going to Williamson's main rival. Analyze the following parts and then, based on your analysis, comment on whether Williamson appears to have a legitimate concern.

a) Calculate and interpret a 95% confidence interval for the proportion of all acceptable applicants who accept Williamson's invitation to enroll. Do the same for all acceptable applicants with a combined score less than 330, with a combined score between 330 and 375, and then with a combined score greater than 375.

b) Calculate and interpret a 95% confidence interval for the proportion of all acceptable students with a combined score less than the median who choose Williamson's rival over Williamson. Do the same for those with a combined score greater than or equal to the median.

c) Calculate and interpret 95% confidence intervals for the mean combined score, the mean high school GPA, and the mean SAT score of all acceptable students who accept Williamson's invitation to enroll. Do the same for all acceptable students who choose to enroll elsewhere. Then calculate and interpret 95% confidence intervals for the differences between these means, where each difference is a mean for students enrolling at Williamson minus the similar mean for students enrolling elsewhere. Assume equal variances between the groups.

d) Williamson is interested (as are most schools) in getting students who are involved in extracurricular activities (clubs and sports). Does it appear to be doing so? Calculate and interpret a 95% confidence interval for the proportion of all students who decide to enroll at Williamson who have been officers of at least two clubs. Calculate and interpret a similar confidence interval for those who have earned at least four varsity letters in sports.

e) The combined score Williamson calculates for each student gives some advantage to students who rank highly in a *large* high school relative to those who rank highly in a small high school. Therefore, Williamson wonders whether it is relatively more successful in attracting students from large high schools than from small high schools. Calculate one or more confidence intervals for relevant parameters to shed some light on this issue. For this part, you are free to use your own judgement and criteria to decide what classifies a high school as large and small.

Question 5 *(30 points)*

The Yǒuhǎo Restaurant is a Chinese carryout/delivery restaurant. Most of Yǒuhǎo's deliveries are within a 10-mile radius, but it occasionally delivers to customers more than 10 miles away. Yǒuhǎo employs a number of delivery people, four of whom are relatively new hires. The restaurant has recently been receiving customer complaints about excessively long delivery times. Therefore, Yǒuhǎo has collected data on a random sample of deliveries by its four new delivery people during the peak dinner time. The data are in the CSV file *DeliveryTimes*. The variables are as follows:

- Deliverer: which person made the delivery
- Prep Time: time (in minutes) from when order was placed until delivery person started driving it to the customer
- Travel Time: time (in minutes) to drive from Yǒuhǎo to customer
- Distance: distance (miles) from Yǒuhǎo to customer

Analyze the following parts and then based on your analysis write the interpretations and recommendations.

a) Yǒuhǎo is concerned that one or more of the new delivery people might be slower than others.

   (i) Let μ_{Di} and μ_{Ti} be the mean delivery time and mean total time for delivery person i, where the total time is the sum of the delivery and prep times. Calculate and interpret 95% confidence intervals for each of these means for each delivery person. Although these might be interesting, give two reasons why they are not really fair measures for comparing the efficiency of the delivery people.

   (ii) Responding to the criticisms in part (i), calculate and interpret a 95% confidence interval for the mean speed of delivery for each delivery person, where speed is measured as miles per hour during the trip from Yǒuhǎo to the customer. Then calculate and interpret 95% confidence intervals for the mean difference in speed between each pair of delivery people. Assume equal variances between the groups.

b) Yǒuhǎo would like to advertise that it can achieve a total delivery time of no more than M minutes for all customers within a 10-mile radius. On all orders that take more than M minutes, Yǒuhǎo will give the customers a $10 certificate on their next purchase.

   (i) Assuming for now that the delivery people in the sample are representative of all of Yǒuhǎo's delivery people, calculate and interpret a 95% confidence interval for the proportion of deliveries (within the 10-mile limit) that will be on time if $M = 25$ minutes; if $M = 30$ minutes; if $M = 35$ minutes.

   (ii) Suppose Yǒuhǎo makes 1000 deliveries within the 10-mile limit. For each of the values of M in part (i), calculate the range for the dollar amount of certificates it will have to pay for being late.

c) The policy in the previous part is simple to state and simple to administer. However, it is somewhat unfair to customers who live close to Yǒuhǎo — they will never get $10 certificates. A fairer policy is the following. Yǒuhǎo first analyzes the data and finds that total delivery times can be predicted fairly well with the regression equation

$$Predicted\ Delivery\ Time = 14.8 + 2.06 Distance$$

   Also, most of these predictions are within 5 minutes of the actual delivery times. Therefore, whenever Yǒuhǎo receives an order over the phone, it looks up the customer's address in its computerized geographical database to find distance, calculates the predicted delivery time based on this equation, rounds this to the nearest minute, adds 5 minutes, and guarantees this delivery time or else a $10 certificate. It does this for all customers, even those beyond the 10-mile limit.

(i) Assuming again that the delivery people in the sample are representative of all of Yǒuhǎo's delivery people, calculate and interpret a 95% confidence interval for the proportion of all deliveries that will be within the guaranteed total delivery time.

(ii) Suppose Yǒuhǎo makes 1000 deliveries. Calculate the range for the dollar amount of certificates it will have to pay for being late.

Question 6 *(10 points)*

A large courier company sends invoices to customers requesting payment within 30 days. Each bill lists an address, and customers are expected to use their own envelopes to return their payments. Currently, the mean amount of time taken to pay bills are 24 days. The chief financial officer (CFO) believes that including a stamped self-addressed (SSA) envelope would decrease the amount of time. She calculates that the improved cash flow from a 2-day decrease in the payment period would pay for the costs of the envelopes and stamps. Any further decrease in the payment period would generate a profit. To test her belief, she randomly selects 220 customers and includes an SSA envelope with their invoices. The numbers of days until payment is received were recorded on the CSV file *SSA*. Can the CFO conclude that the plan will be profitable? Make a decision using an appropriate level of significance. Choose a significance level by performing the Type I and Type II error analysis. Also please state your null and alternative hypotheses.

Question 7 *(5 points)*

When an election for political office takes place, the television networks cancel regular programming and instead provide election coverage. When the ballots are counted, the results are reported. However, for important offices such as president or senator in large states, the networks actively compete to see which will be the first to predict a winner. This is done through exit polls[1], wherein a random sample of voters who exit the polling booth is asked for whom they voted. From the data, the sample proportion of voters supporting the candidates is computed. A statistical technique is applied to determine whether there is enough evidence to infer that the leading candidate will garner enough votes to win. Suppose that in the exit poll from the state of Florida during the 2000 year elections, the pollsters recorded only the votes of the two candidates who had any chance of winning, Democrat Albert Gore (code = 1) and Republican

[1] Warren Mitofsky is generally credited for creating the election day exit poll in 1967 when he worked for CBS News. Mitofsky claimed to have correctly predicted 2,500 elections and only six wrong. Exit polls are considered so accurate that when the exit poll and the actual election result differ, some newspaper and television reporters claim that the election result is wrong! In the 2004 presidential election, exit polls showed John Kerry leading. However, when the ballots were counted, George Bush won the state of Ohio. Conspiracy theorists now believe that the Ohio election was stolen by the Republicans using the exit poll as their "proof." However, Mitofsky's own analysis found that the exit poll was improperly conducted, resulting in many Republican voters refusing to participate in the poll. Blame was placed on poorly trained interviewers (Source: *Amstat* News, December 2006).

George W. Bush (code = 2). The polls close at 8:00 p.m. Can the networks conclude from these data that the Republican candidate will win the state? Should the network announce at 8:01 p.m. that the Republican candidate will win? Use a 10% significance level. Also please state your null and alternative hypotheses. Write your final interpretation. The dataset is available on the CSV file *Elections*.

Question 8 *(10 points)*

A study is performed in a large town to determine whether the average amount spent on food per four-person family in the town is significantly different from the national average. A random sample of the weekly grocery bills of four-person families in this town is given in the file *GroceryBills*. Assume the national average amount spent on food for a four-person family is $150.

a) *5 points* Identify the null and alternative hypotheses for this situation. Is the sample evidence statistically significant? If so, at what significance levels can you reject the null hypothesis?

b) *5 points* For which values of the test statistic, sample mean (i.e., average weekly grocery bill), would you reject the null hypothesis at the 1% significance level? For which values of the test statistic, sample mean, would you reject the null hypothesis at the 10% level?

Question 9 *(5 points)*

One important factor in inventory control is the variance of the daily demand for the product. An operations research analyst has developed the optimal order quantity and reorder point, assuming that the variance is equal to 250. Recently, the company has experienced some inventory problems, which induced the operations manager to doubt the assumption. To examine the problem, the manager took a sample of 25 days and recorded the demand. The data are provided on the *Demand* CSV file. Do these data provide sufficient evidence at the 5% significance level to infer that the operations research analyst's assumption about the variance is wrong? Also please state your null and alternative hypotheses. Write your final interpretation.

Question 10 *(3 points)*

Approaching San Francisco International Airport (SFO) for landing, a British Airways flight from London Heathrow Airport (LHR) has been on hold for 45 minutes due to dense fog and inclement weather. The flight crew could declare an emergency and land immediately, but an FAA investigation will be launched and other flights might be endangered. The flight crew believes that there is enough fuel to stay aloft for 15 more minutes. Define Type I and Type II errors and identify the consequences of each type of error.

Question 11 *(30 points)*

Demand for systems analysts in the consulting industry is greater than ever. Graduates with a combination of business and software knowledge — some even from liberal arts programs — are getting great offers from consulting companies. Once these people are hired, they frequently switch from one company to another as competing companies lure them away with even better offers. One consulting company, BAY, has collected data on a sample of systems analysts with undergraduate degrees they hired several years ago. The data are in the file *BAY*. The variables are as follows:

- Starting Salary: employee's starting salary at BAY
- On Road Pct: percentage of time employee has spent on the road with clients
- State Univ: whether the employee graduated from State University (BAY's principal source of recruits)
- CIS Degree: whether the employee majored in Computer Information Systems (CIS) or a similar computer-related area
- Stayed 3 Years: whether the employee stayed at least three years with BAY
- Tenure: tenure of employee at BAY (months) if he or she moved before three years

BAY is trying to learn everything it can about retention of these valuable employees. You can help by answering the following questions for BAY. Clearly write your interpretations and insights from the results.

a) Although starting salaries are in a fairly narrow band, BAY wonders whether they have anything to do with retention.

   i Calculate a 95% confidence interval for the mean starting salary of all employees who stay at least three years with BAY. Calculate a 95% confidence interval for the mean starting salary of all employees who leave before three years. Then calculate a 95% confidence interval for the difference between these means. Assume equal variances.

   ii Among all employees whose starting salary is less than or equal to the median ($37,750$), calculate a 95% confidence interval for the proportion who stay with BAY for at least three years. Among all employees with starting salaries above the median, calculate a 95% confidence interval for the proportion who stay with BAY for at least three years. Then calculate a 95% confidence interval for the difference between these proportions.

b) BAY wonders whether the percentage of time on the road might influence who stays for at least 3 years and who leaves before 3 years. Repeat part a) i and a) ii, but now do the analysis in terms of (for the variable) percentage of time on the road rather than for the variable starting salary.

c) Find a 95% confidence interval for the mean tenure (in months) of all employees who leave BAY within three years of being hired. Why is it not possible with the given data to calculate a confidence interval for the mean tenure at BAY among all systems analysts who leave BAY after three years?

d) State University's students, particularly those in its nationally acclaimed CIS area, have traditionally been among the best of BAY's recruits. But are they relatively hard to retain? Find the 95% confidence interval for proportion who stayed 3 years, broken down by whether they went to State University or not. What do you learn from these confidence intervals? Find the 95% confidence interval for proportion who stayed 3 years, broken down by whether they got a CIS degree or not. What do you learn from these confidence intervals?

Question 12 *(5 points)*

A study involving hypothesis testing is conducted in a major hospital to understand the proportion of mothers who stay for less than two days in the hospital after delivery. A sample of 50 births is taken to test whether more than half of all mothers have a length of stay (LOS) of less than two days. Use $\alpha = 0.10$. If the true proportion of mothers who have an LOS of two days is 0.60, what is the power of the test? What sample size is needed to attain an 80% power?