Math 437 - Midterm Exam 1

Your Name Here

Due March 13, 2023

Exam Rules

1. Solutions to this exam must be uploaded to Canvas by 11:59 PM on Monday, March 13. **NO EXTENSIONS WILL BE GRANTED**. Do not ask for them.
2. You should submit *both* this .Rmd file with your solutions included *and* the knitted pdf file. If you are having trouble knitting, knit to HTML and print/save the html file to pdf.
3. **Exams must be done individually**. Do not confer with anyone other than me about the exam.
4. You may consult any reference materials available to you, including the required and recommended textbooks, your course notes/code you have written, and anything on Canvas (such as my solutions). You are allowed to search the Internet or the library, but provide citations for outside references used.
5. Approach the exam from the point of view of a statistician, not a student. There may be multiple valid approaches on a problem, and any valid approach will be accepted so long as it is accompanied with appropriate justification and/or explanation.
6. I will provide a three-tiered hint system by e-mail:

- You may ask me to clarify something in a problem for no point penalty.
- You may give me pseudocode or buggy code, and ask me to implement/fix the code, for a partial point penalty (I will reply with how many points you will lose, then you confirm you want me to do it).
- You may ask me for my solution code, and I will provide it for a full point penalty (you will lose all Code points for that part). This may still be a good deal if you need the code to do the Explanation part or another part of the problem.

```r
# Packages you probably need
library(ggplot2)
library(dplyr)

# Please load any other packages you need either all in this chunk
# or in the chunk you need it for

# Data you need for the applied problem
SIS <- readr::read_csv("SIS.csv")
```

This exam is worth a total of **105** points, but is graded out of **100** points.

Conceptual Problem (16 points total)

Consider testing $H_0 : \mu_1 - \mu_2 = 0$. Make the following assumptions:

1. The response variable has been scaled such that $\sigma = 1$ in both groups.
2. We are performing a two-sided hypothesis test using $\alpha = 0.05$.

3. We are using the difference in sample means $\bar{x}_1 - \bar{x}_2$ as our test statistic, NOT a standardized t or z statistic.
4. The sample size in each group is large enough to assume that the distribution of $\bar{x}$ is normal in both groups.

Part a (2 pts)

Write a sentence explaining what the p-value represents in the context of this test, using the assumptions (1)-(4) above. You should start your sentence, "The p-value is the probability…".

Part b (2 pts)

Suppose that we obtain a sample of 32 observations in each group and find that $\bar{x}_1 - \bar{x}_2 = 0.8$. In the chunk below, write code to compute (and print/output) the p-value.

The following hints may be useful:

1. The function for computing areas under the normal distribution curve is `pnorm`.
2. The first argument to `pnorm` should indicate the boundary value (e.g., test statistic value) that you are converting to an area under curve. The second and third arguments are the `mean` and `sd` of the normal distribution. The argument `lower.tail` takes the value `TRUE` if computing the area to the left of the boundary value or `FALSE` if computing the area to the right of the value.
3. Correctly explaining *what* you want R to do in order to compute the p-value will earn at least 1 pt even if you code it completely wrong.
4. If you get stuck, I would recommend reviewing assumptions (1)-(4) above, as well as your Math 338 notes.

Part c (2 pts)

Based on the p-value you computed in Part (b), is it possible to make a Type I Error? Is it possible to make a Type II Error? Explain your reasoning.

The following hints may be useful:

1. You can still earn full credit even if you computed the wrong p-value in Part (b).
2. If your code in Part (b) does not run, just pick an arbitrary p-value (tell me what it is!) and answer the question assuming that p-value is the right one.

Part d (2 pts)

Suppose we obtained a second sample from the same population with 32 observations in each group, but this time $\bar{x}_1 - \bar{x}_2 = 0.3$. Would the power of our test applied to the second sample be lower, higher, or the same compared to the original sample with $\bar{x}_1 - \bar{x}_2 = 0.8$? Explain your reasoning.

Part e (2 pts)

Suppose that Group 1 and Group 2 were instead two of six groups we wish to compare the population means between. Using a Bonferroni correction to control the FWER, what significance level should we compare the p-value to and why?

Part f (2 pts)

Under the conditions from part (e), suggest a better alternative to the Bonferroni correction for controlling the FWER and explain why it is better.

Part g (2 pts)

Suppose that we only obtained 8 observations in each group, so that assumption 4 - the normal distribution of $\bar{x}$ - may or may not hold. What would you do instead to get an accurate p-value for testing $H_0 : \mu_1 = \mu_2 = 0$? (You may assume we only have two groups again.) What additional assumptions (if any) would you have to make?

Part h (2 pts)

Under the conditions from part (g), a t-based confidence interval may not be appropriate. How would you obtain a 95% bootstrap percentile confidence interval for $\mu_1 - \mu_2$? You do *not* have to write any code to obtain the interval, but pseudocode may be useful. Think very carefully about the fact that the data consists of 8 observations in each group!

Simulation Problem (16 points total)

Part a (Code: 5 pts)

Simulate a single sample under $H_a : \mu_1 - \mu_2 = 0.25$; that is, the sample should contain 32 observations from $N(0.25, 1)$ in Group 1 and 32 observations from $N(0, 1)$ in Group 2. You can either use separate variables `x1` and `x2` for the values from the two groups or create a 64×2 data frame to store the values and groups in.

Obtain the (simulated) value of $\bar{x}_1 - \bar{x}_2$ and the corresponding p-value for testing $H_0 : \mu_1 - \mu_2 = 0$ against a two-sided alternative.

You should make the same assumptions as in the Conceptual Problem:

1. The response variable has been scaled such that $\sigma = 1$ in both groups.
2. We are performing a two-sided hypothesis test using $\alpha = 0.05$.
3. We are using the difference in sample means $\bar{x}_1 - \bar{x}_2$ as our test statistic, NOT a standardized t or z statistic. *Do not* use the `t.test` function to get a p-value.
4. The sample size in each group is large enough to assume that the distribution of $\bar{x}$ is normal in both groups.

The following hints may be useful:

1. Think carefully about the distribution of $\bar{x}_1 - \bar{x}_2$ under H_0. Your solution to Conceptual Problem part (b) may be informative.
2. The `abs` function takes the absolute value of each element in a vector. You don't necessarily *need* to use it, but it may be helpful.
3. Correctly explaining what you want R to do in order to get the p-value will earn at least 1 point even if you code it completely wrong.

Part b (Pseudocode and Code: 3 pts)

Using a `for` loop and your code from part (a), obtain 1000 samples of x1 and x2 under H_a and the corresponding test statistics and p-values. Using $\alpha = 0.05$, estimate the power based on these 1000 samples.

The following hints may be useful:

1. Boolean operations in R include `==` ($a = b$), `!=` ($a \neq b$), `<` ($a < b$), `>` ($a > b$), `<=` ($a \leq b$), and `>=` ($a \geq b$).
2. Remember what the `sum` and `mean` functions do with Boolean (TRUE/FALSE) variables.
3. Correctly explaining what you want R to do in order to estimate the power after you do the simulation will earn at least 1 point even if you code it completely wrong.

Part c (Pseudocode and Code: 3 pts)

Perform another simulation as in part (b), but this time simulate 1000 samples under H_0; that is, each sample should contain 32 observations from $N(0, 1)$ in Group 1 and 32 observations from $N(0, 1)$ in Group 2. Estimate the Type I Error Rate in this simulation.

The following hints may be useful:

1. In the next part you will need to use both simulations, so do not overwrite the vectors containing the test statistics and p-values from part (b).
2. This simulation is almost the exact same as in part (b). If you did that simulation correctly, you should be able to copy the code and make minor changes.
3. Correctly explaining what you want R to do in order to estimate the Type I Error Rate after you do the simulation will earn at least 1 point even if you code it completely wrong.

Part d (Code: 1 pt; Explanation: 2 pts)

Combine the 2000 p-values from parts (b) and (c) into a single vector. Using this vector, produce a table like the ones in Lab 13.6.1 and estimate the false discovery rate at $\alpha = 0.05$ for this set of 2000 tests.

Part e (Code: 1 pt; Explanation: 1 pt)

Adjust your vector of 2000 p-values from part (d) using the Benjamini-Hochberg method. Comment on how the new false discovery proportion when controlling $q = 0.05$ compares to the false discovery proportion from part (d).

Applied Problem (63 points total)

In this problem you will be replicating and expanding on the results of Bissig and DeCarli (2019). Bissig and DeCarli recorded a number of variables on 100 patients who visited the emergency department at University of California, Davis hospital and who required a neurology consultation. This data can be found in the `SIS` file on Canvas.

Please see the *SIS_dictionary* file for a description of each of the variables in the dataset.

Part a (Code: 3 points)

The researchers defined their primary numerical response variable of interest as the total number of tests ordered and their primary explanatory variable of interest as whether the patient passed (combined score of 4 or higher on the Orientation and Free Recall tasks) or failed (combined score of 3 or lower) the SIS. Unfortunately, neither of these variables is present in the dataset.

Using the `dplyr` package, add two new variables to the dataset, `Quantity_Total_Orders` and `Pass`, that represent the two variables of interest. `Pass` may be a numerical, character, factor, or logical variable.

Part b (Code: 8 points; Explanation: 12 points)

Perform an exploratory data analysis on the two new variables you created as well as `Admission` and `Duration_of_Admission`. In particular, I will be grading on how well you can answer the following questions and support your answer with numerical, tabular, and/or graphical summaries:

- What data is missing? Is there an obvious explanation for any of the missingness?
- What proportion of patients passed the SIS? What proportion of patients were admitted to the hospital after consultation? Does there appear to be a relationship between these two variables?
- What is the distribution of `Quantity_Total_Orders`? Does there appear to be a relationship between this response variable and any of the other three variables (`Pass`, `Admission`, `Duration of Admission`) under investigation? If so, describe the relationship.

- Does the relationship between `Quantity_Total_Orders` and `Duration_of_Admission` appear to change depending on whether the patient passed the SIS or not?

Part c (Code: 4 pts; Explanation: 9 pts)

The authors report the following about a least-squares regression of the natural log of `Quantity_Total_Orders` ("ln[studies]" in the paper) against the natural log of `Duration_of_Admission` ("ln[LOS]" in the paper):

- The least-squares regression equation for the "good performers" (those who passed the SIS) is ln[studies] = 0.36 * ln[LOS] + 3.0
- The least-squares regression equation for the "poor performers" (those who did not pass the SIS) is ln[studies] = 0.65 * ln[LOS] + 3.0
- "The full model of ln(LOS) + SIS performance + the interaction explains 55% of the variance in ln(studies)"
- The p-value for testing H_0: in the population model, the interaction has no effect on the number of total orders against H_a: in the population model, the interaction does affect the number of total orders was 0.012
- "Based on the best-fit lines, about 40% more studies will be ordered on a poor performer than on a good performer during a 3-day hospitalization."

Using your `Quantity_Total_Orders`, `Duration_of_Admission`, and `Pass` variables, fit the full model indicated by the authors (note that the R function to take natural logarithms is `log`; you may do this transformation either before fitting the model or in the model-fitting code itself). Determine whether each statement above is correct or incorrect, and justify your answer based on the model output.

Part d (Explanation: 2 pts)

Would it be appropriate to remove the variable `Pass` from this model, thus leaving only `log(Duration_of_Admission)` and the interaction term? Why or why not?

Part e (Code: 2 pts; Explanation: 2 pts)

Create a *new* linear regression model that predicts `Quantity_Total_Orders` from `Duration_of_Admission`, `Pass`, and the interaction term (i.e., no log-transformations). Obtain the residual plot and q-q plot for both models (i.e., with and without the log-transformations). Why do you think the researchers chose to use a log-transformation?

Part f (Explanation: 3 pts)

Write a sentence to interpret the slope corresponding to the variable `Pass` in the model from part (e) (i.e., the one without any log-transformations).

Part g (Code: 1 pt; Explanation: 4 pts)

The "False-Positive Psychology" paper refers to *researcher degrees of freedom* - small, seemingly insignificant decisions that can have large impacts on the ultimate inferential results.

In this paper, the researchers made several decisions about how they would represent the number of tests ordered (response), the duration of hospital stay (explanatory), and the performance on the SIS (explanatory) in their model.

Other than the log transformations, identify one alternative choice the researchers *could have* made but didn't when representing these variables in the model. Then, modify your linear regression model from part (c) to incorporate this change (you may need to create a new variable in the dataset before fitting the model). How did this change affect the slope estimates and their corresponding p-values, compared to the model you fit in part (c)?

Part h (Pseudocode and/or Explanation: 8 pts)

How would you conduct a permutation test to determine whether patients who can correctly draw a clock (`Clock_Draw` in the `SIS` dataset) have higher SIS scores, on average, than patients who cannot?

Be explicit about the null and alternative hypothesis (in context) and the steps you would take to obtain a p-value. *You do not need to actually write any code.*

Part i (Code: 5 pts)

Obtain a 99% confidence interval for the proportion of emergency room patients undergoing neurological consult that have an ischemic etiology (i.e., the proportion of patients for which `Ischemic == Y`). Use the bootstrap method of your choice with 10,000 bootstrap resamples.

Oral Exam (10 points)

Sign up on Canvas for a 15-minute window in which we will talk for about 10 minutes about your exam solutions.

You will receive 5 points for showing up and up to 5 points of extra credit based on the clarity of your explanations.