MSDS 6371 Fall 2020 Midterm

**Important: By taking this exam, the student is agreeing to not discuss the midterm or anything that might help with the midterm *with anyone* over any media (in person, slack, email, text, etc.). You may use your textbook, powerpoints, notes, or any other material except for another human being or super advanced AI that be indistinguishable from a human. With that said, have a blast showing off your knowledge… go for it!**

**Ohh… and you can simply type you answers on this test or you may print the test out and write directly on the paper or on additional sheets of papers and scan it to a pdf and submit the pdf. Please submit the final .docx or .pdf file to 2DS.**

Question 1

(3pts) Suppose we are testing the claim that the average income of Dallas Independent School District (DISD) teacher is $45,832 and the following null and alternative hypothesis identified:

H0: *μ*= $45,832

Ha: *μ*≠ $45,832

Which of the following is a true statement?

Select one:

**A. A Type I Error is concluding that there is enough evidence to suggest that the mean salary of DISD teachers is different than** $45,832 when the actual mean is $45,832.

**B. A Type I Error is is concluding that there is enough evidence to suggest that the mean salary of DISD teachers is different than** $45,832 when the actual mean is $49,000.

**C. A Type II Error is concluding that there is enough evidence to suggest that the mean salary of DISD teachers is different than** $45,832 when the actual mean is $45,832.

**D. A Type II Error is is concluding that there is enough evidence to suggest that the mean salary of DISD teacher is different than** $45,832 when the actual mean is $45,832.

Question 2

(3pts) Suppose a hypothesis test is a performed and p-value of 0.523 is obtained. Which of the following is the correct interpretation of the p-value?

Select one:

**A. If we were to repeat the experiment many, many times there is a 52.3% chance that one would get a test statistic as extreme or more extreme than the observed value by chance alone if the null hypothesis is true.**

**B. There is a 52.3% chance that one would get a test statistic as extreme or more extreme than the observed value by chance alone if the alternative is true.**

**C. There is a 52.3% chance that the null hypothesis is true.**

**D. There is a 52.3% chance that the alternative hypothesis is true.**

**E. A and C are True**

**F. B and D are True**

**G. None are True**

**H. All are True**

Question 3

(3pts) Assume we conducted a study in which we wanted to see if listening to music helped subjects hold their breath longer. A study was set up in which subjects were asked to hold their breath as long as they can once while listening to music and again without listening to music. Two times were recorded for each subject, once while they were listening to music and once while they a were not listening to music. Which type of study would be most appropriate for this data?

1. **Two Sample T Test with pooled standard deviation**
2. **Welch’s T Test for the difference of means**
3. **Rank Sum Test**
4. **ANOVA**
5. **Kruskal-Wallis test**
6. **Signed-Ranked Test**
7. **None of the above are appropriate**

Question 4

(3pts) Assume this time we conducted a similar study although this time we only took one measurement per subject. In this experiment we had a “Listening to Music” group and a “Not Listening to Music” group. Subjects were randomly sampled from the population and then randomly assigned to be in one of the two groups. Each subject was then asked to hold their breath as long as they could and their time was recorded. However, the timing device could only measure up to a minute and thus everyone who held their breath longer than a minute was recorded as a “1 minute”. Which type of study would be most appropriate for this type of study?

1. **Two Sample T Test with pooled standard deviation**
2. **Welch’s T Test for the difference of means**
3. **Rank Sum Test**
4. **ANOVA**
5. **Kruskal-Wallis test**
6. **Signed-Ranked Test**

Question 5

(3pts) A 95% confidence interval of the difference of means  is found to be [2.3, 2.5]. Which is a correct interpretation of this confidence interval?

Select one:

**A. There is a 95% chance that both** and  **are each between 2.3 and 2.5.**

**B. There is a 95% chance that** or  **is between 2.3 and 2.5.**

**C. We are at least 95% confident that** is larger than

**D**. **We are at least 95% confident that** is larger than

**E. There is a 95% chance that** *x̄1 - x̄2*  **is between 2.3 and 2.5.**

**F. The sample mean***x̄***is likely between 2.3 and 2.5. The procedure used gives a confidence interval containing the sample mean***x̄***for 95% of samples.**

Question 6

(3pts) A 95% confidence interval of the difference of means  is found to be [22.3, 25.6]. Which is a correct statement about the corresponding hypothesis test?

Select one:

1. **A two-sided hypothesis test with alpha = .05 would reject Ho:**  = 0 and conclude that there is enough evidence to support **Ha:**  ≠ 0.
2. **A two- sided hypothesis test with alpha = .05 would fail to reject Ho:**  = 0 and conclude that there is not enough evidence to support **Ha:**  ≠ 0.

Question 7

(3pts) All else held constant, which increases the power of a one sample t-test?

Select one (the most appropriate answer):

**A. Increasing the Effect Size only**

**B. Increasing the Sample Size only**

**C. Increasing the significance level only**

**D. A and B will both increase the power of the test.**

**E. A and C will both increase the power of the test.**

**F. B and C will both increase the power of the test.**

**G. All will increase the power of the test.**

**H. None will increase the power of the test.**

Question 8

(3pts) As long as subjects have been randomly assigned to one of the treatment groups, inference can be generalized to the population that the sample was taken from.

Select one:

**A. True**

**B. False**

Question 9

(3pts) Suppose a data set of continuous numbers consists of most of the numbers clustered together with a few outliers much higher than the others (and no other outliers). Choose the best answer that describes the skewness of the data.

Select one:

**A. The data set is skewed to the right.**

**B. The data set is skewed to the left.**

**C. The data set is not skewed. (It is symmetric.)**

**D. There is not enough information to suggest skewness.**

Question 10

(3pts) A researcher for Car and Driver magazine was interested in if there was a difference between the MPGs (miles per gallon) of hybrid cars and their manufacturer. In order to test this, the magazine gained access to 3 Toyota Corolla Hybrid, 4 Ford Fusions and 5 Chevy Malibu Hybrids and recorded the MPGs from each of these cars. From a previous study, there is reason to believe that the distributions of mpgs from these cars are very right skewed and that the standard deviations are similar. What is the best test to test for a difference in the centers between any pair of these distributions?

Select one:

**A. Signed Rank Test**

**B. Rank Sum Test**

**C. Welch's T Test**

**D. Kruskal Wallis Test**

**E. Brown and Forsythe Test**

**F. Pooled T Test**

**G. 1-way ANOVA**

Question 11

(3pts) Your company is trying out a new website to try and generate more business. Assume your boss has asked you to compare the mean daily mouse click traffic on the company’s 5 different versions of its new website: Original Version, Ver 1, Ver 2, Ver 3, Ver 4. You were asked to compare the mean click rate between each new website (Ver 1, Ver 2, Ver 3 and Ver 4) and the original website (Original Website). You of course want to do the appropriate multiple comparison correction. Which correction is most appropriate here?

Select one:

1. **Dunnett**
2. **Tukey-Kramer**
3. **Bonferroni**
4. **Sheffee**

Question 12

(8 pts) With respect to the last problem, let’s say that Ver 3 and Ver 4 used video content while Ver 1 and Ver 2 did not. You would thus like to compare the average of the click rates of Version 3 and 4 with the Original but you are assuming the standard deviations to all be the same so you would like to include Ver 1 and Ver 2 in the estimation of the standard deviation. What are the contrast weights that would be used to test the hypothesis:

Assume the order of the groups is alphabetical: “Original Version”, “Ver 1”, “Ver 2”, “Ver 3”, “Ver 4”

Assume the data is contained in a dataset called ***WebsiteTest***, the groups are in a variable called ***WebsiteType*** and the click rates are in a variable called ***ClickRate***. Simply finish the code below to perform the desired contrast. Place your additional code in place of the green placeholders.

**proc** **glm** data = **WebsiteTest ORDER=data**; <-keeps data in order it came in so control is first and can be assigned 0

class **WebsiteType**;

model **ClickRate=WebsiteType**

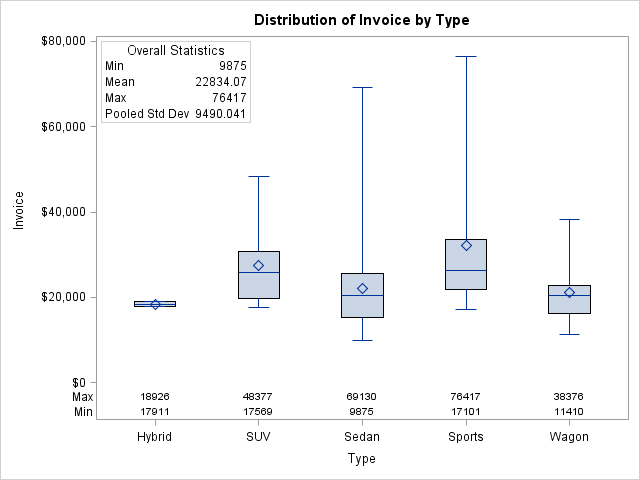
lsmeans WebsiteType / pdiff;

estimate **‘Avg. V3 & V4 vs Avg V3 & V4’ WebsiteType 0,0,0,1,-1**;

estimate **‘Avg. V3 & V4 vs Avg V3 & V4’ WebsiteType 0,0,0,-1,1**;

**run**;

Question 13



(3pts) The box plots for the Invoice amount for different car types is above. The box plot for Sedans is consistent with what type of skew?

**A. Left Skewed Data**

**B. Symmetric Data**

**C. Uniformly Distributed Data**

**D. Right Skewed Data**

Question 14

(3pts) With respect to the last question, assume that the sample size was 500 cars so that each car type had 100 observations. Note that an estimate of the pooled standard deviation is given in the upper left-hand corner of the plot. Do you feel it is an appropriate estimate for the standard deviation of the invoice price for each type of car? Why or why not? (1 to 3 sentences should be enough to answer this question.)

I think that it is an appropriate estimate for the standard deviation of the invoice price for each type of car. This is because pooled standard deviation is the weighted average of the standard deviations for two or more groups, and since the individual average standard deviations are weighted, we have more weight given to larger samples. And since we have such a large sample of cars, the CLT comes into play. Since it’s such a large sample for each, I feel confidence about the SD for invoice price for each car.

Question 15

1. Assume a study was conducted in which there were 4 groups (A, B, C and D) and 10 observations were conducted per group. When each of the four groups (variable: Group) are allowed to have their own mean (the separate means model), the sum of squared residuals between each group’s observations and its group sample mean is 108. In contrast, when only a single mean is fit to the 40 observations (the equal means model), the sum of the squared residuals between the 40 observations and the overall sample mean was 153. With this information, fill out the ANOVA table below (5 pts).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **SS** | **MS** | **F** | **Pvalue** |
| **Model** | **3** | **45** | **45** | **2.85** | **0.95** |
| **Error** | **37** | **108** |  |  |  |
| **Total** | **40** | **153** |  |  |  |

1. The critical value for this test was 2.85. Perform a 6-step hypothesis test to test the claim that at least one pair of the means for groups A, B, C and D are different. (6 pts).

Step 1: QOI, we are interested in seeing is at least one pair of the group means for A,B,C and D are different

Step 2: Assumptions normally distributed data, equal variances, and independent data. We will assume all are met

Step 3: F-statistic: 2.85

Step 4: P-value = 0.95

Step 5: Decision: Fail to reject H\_0

Step 6: Conclusion, there isn’t strong evidence to suggest at least one pair of the group means for A,B,C and D are different.

Step 7: Scope, based on the problem statement, we cannot determine if this study will apply to the boarder population.

1. We know that one of the assumptions of the ANOVA is that the standard deviations are the same for all 4 groups (A,B,C and D). Given this data, what is the estimate of this common standard deviation? (2 pts)
2. Find and interpret R2.

R^2= SSmodel/SStot= 0.29 based on the data I calculated

1. Assume now that the question of interest is to compare the mean of the A group versus that of the C group. Explain which code below would conduct the more powerful test and why. Include in your answer the degrees of freedom for estimating the error of each test.(3pts)

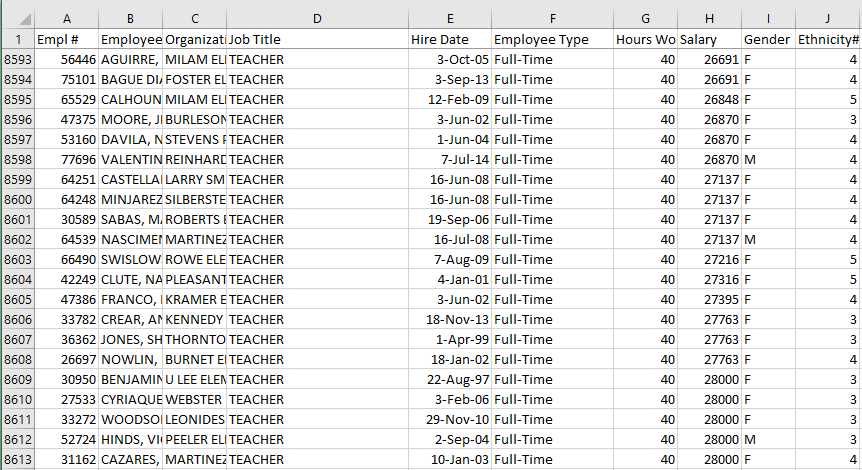
|  |  |
| --- | --- |
| **Code Snippet 1** | **Code Snippet 2** |
| **proc** **glm** data = Question13;  class Group;  var Score;  estimate “A versus C” Group / 1,0,-1,0;  **run**; | **proc** **ttest** data = Question13;  where Group == A | Group == C;  class Group;  var Score;  **run**; |

I believe code snippet 1 will be more powerful, since we are using contrasts here. We can isolate the groups of interest and study those.

Analysis Question



These questions involve real public data pulled from the Dallas Independent School District (DISD) and is a compilation of the salaries of all full-time teachers in DISD for 2015. Below is a snippet of the dataset and show, among other variables, the variables of interest in our study: Job Title, Hours Worked, Salary, Ethnicity# (Ethnicity Number), and Gender.



Our overall question of interest is to test the claim that there is gender and / or ethnicity discrimination ***in Texas*** (of which this data is a sample from). Assume you have been hired as an expert consultant by a lawyer for the State of Texas to investigate if there is evidence of a significant difference in the mean or median salary of the five ethnicity groups identified by the “Ethnicity#” variable and / or significant evidence of a difference in the mean or median salary between males and females identified by the “Gender” column.

You have actually taken over from the state’s last expert consultant who had to leave suddenly due to an undisclosed issue and have thus been left a series of code and output that can be found below the questions. Use this code and output to answer the following questions. You may not need all the code and / or output to answer the questions… you will have to choose what is relevant. The dataset was read into SAS as the dataset: ***DISD40T***

1. (20 pts) Perform a complete analysis in order to test the claim that men have a higher average mean / median salary than women. A complete analysis includes the following:
   1. State the problem
   2. Address the assumptions
   3. Perform the most appropriate test with a 6 step hypothesis test. You may skip step 2. (But be sure and include a confidence interval to quantify your uncertainty.)
   4. Provide a scope of inference.
2. Problem statement, we are interested in testing the claim that men have a higher average mean/median salary than women. We will use statistical tests to determine if this question requires further research.
3. Assumptions, in a pooled two sample t-test we have three assumptions. The first is the normality of data assumption, the second is that the variances of the these normal distributions are equal and the data are independent within each group.

Since this is salary data it is quite skewed, we can address this by performing a log transform, although this dataset is quite large so applying the CLT, it may be robust to this assumption. Based on the data and the output below and fail to reject a difference in variances, and conclude there isn’t evidence against equal variances. and finally we will assume the data is intendent.

1. Step 1: QOI: Problem statement, we are interested in testing the claim that men have a higher average mean/median salary than women. We will use statistical tests to determine if this question requires further research.

Step 2: Assumptions, Assumptions, in t-test we have three assumptions. The first is the normality of data assumption, the second is that the variances of the these normal distributions are equal and the data are independent within each group.

Since this is salary data it is quite skewed, we can address this by performing a log transform, although this dataset is quite large so applying the CLT, it may be robust to this assumption. Based on the data and the output below and fail to reject a difference in variances, and conclude there isn’t evidence against equal variances. finally we will assume the data is intendent. I believe a log transform is appropriate, and thus we will make inference on median.

Step 3: H0: median\_income\_men = median\_income\_women

Ha: median\_income\_men != median\_income\_women

Step 4: t-value: -0.1

Step 5: P-value: 0.9946

Step 6: Fail to reject the Null hypothesis

Step 7: Conclusion, since we failed to reject the null hypothesis we can report that there is insufficient evidence to support the claim of a difference in median salary for men and women. Scope of inference, we cant assume this data was sampled randomly, so we cannot apply this to a broader population, our p-value is 0.9946, and a 95% CI is:-275.8 and 273.9, since this CI contains zero, this supports our conclusion and we used a parametric pooled two sample t-test.

1. (12 pts) Next, the state’s lawyer wanted you to look into the question of ethnicity-based discrimination. She had no prior intuition as to if there was any discrimination and before she looked at the data, she indicated that she was only interested in investigating if there was a significant difference between ethnicities 1 and 2, 1 and 5, and 2 and 4. Assume all assumptions are met and find the appropriate code and output to address this question. Note that a multiple comparison correction procedure was not used so we will have to make this adjustment by hand. Apply a Bonferroni multiple comparison adjustment by hand before making any inference. After making the correction, please provide a sentence per comparison describing if there was significant evidence of a difference in mean or median salary and provide pvalues to support your answers.

Step 1: QOI: We are interested in determining if there is ethnicity-based discrimination between ethnicities 1 and 2, 1 and 5 and 2 and 4.

1. Step 2: Assumptions, Assumptions, in an ANOVA analysis and for multiple comparisons we have three assumptions. The first is the normality of data assumption, the second is that the variances of the these normal distributions are equal and the data are independent within each group.

Since this is salary data it is quite skewed, we can address this by performing a log transform, although this dataset is quite large so applying the CLT, it may be robust to this assumption. Based on the data and the output below and fail to reject a difference in variances, and conclude there isn’t evidence against equal variances. and finally we will assume the data is intendent.

Step 3: H0: 1=2=4=5

HA: at least one differs

Step 4: Bonferroni's adjustment:  
Lower the 0.05 to 0.0166667  
t-val for 1 sided testing: >= -3.07 test 1, -1.82 for test 2,3

Step 5:P-val: 0.0021 test 1, 0.0682 group 2,3

Step 6: Rejct H\_0

Step 7: Conclusion, we reject H\_0 and conclude that there is a difference between at least one of the groups, and we have determined that there are differences between 5 and 1, and 5 and 2, and 5 and 4. Scope of inference, we cant assume this data was sampled randomly, so we cannot apply this to a broader population, our p-values are 0.0021 and 0.0682 , this supports our conclusion and we used a Bonferroni correction. There is significant evidence of a difference in salary.

1. (10 pts) Finally, the state’s lawyer was interested in investigating if there was a significant difference in mean salary between the 5th ethnicity group and the average of the 1st, 2nd and , 3rd ethnic groups. Again, find the appropriate code and output below and provide a written conclusion as to if there is significant evidence of a difference in mean salaries. Please also include a 95% confidence interval in your answer and show your work in calculating this interval. Assume the assumptions of this test are either met or the test is sufficiently robust to any violations.

There was a significant difference between all three groups. I’m about out of time, but the CI’s are easy to produce with the PI the critical value and adding or subtracting.

**Output** (Note that there may be errors in some parts of the output. It is up to the student to identify and correctly interpret the relevant output / results.)

|  |  |
| --- | --- |
| **proc** **ttest** data = DISD40T;  class Gender;  var Salary;  **run**; | **proc** **npar1way** data = DISD40T Wilcoxon;  class Gender;  var Salary;  **run**; |
|  |  |
|  |  |
|  |  |

|  |
| --- |
| **proc** **glm** data = DISD40T;  class Ethnicity\_;  model Salary = Ethnicity\_;  lsmeans Ethnicity\_ / pdiff;  estimate "5 versus average of 1,2,3 Ver 1" Ethnicity\_ **1** **1** **1** **1** -**4** / divisor = **4**;  estimate "5 versus average of 1,2,3 Ver 2" Ethnicity\_ **1** **1** **1** **0** -**3** / divisor = **3**;  estimate "5 versus average of 1,2,3 Ver 3" Ethnicity\_ **.3** **.3** **.3** **0** -**3**;  estimate "5 versus average of 1,2,3 Ver 4" Ethnicity\_ **1** **1** **1 0** -**1**;  **run**; |
|  |
|  |

That’s all folks!

I hope you have a great weekend!

It is certainly well deserved!

(Except the code is Below!)