MSDS 6371 SPRING 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!**

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 not enough evidence to suggest that the mean salary of DISD teachers is different than** $45,832 when the actual mean is $49,000.

**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 not enough evidence to suggest that the mean salary of DISD teachers is different than** $45,832 when the actual mean is $49,000.

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

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

**C. 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 is true.**

**D. 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.**

**E. A and C are True**

**F. B and D are True**

**G. None 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) A 95% confidence interval of the difference of means  is found to be [22.3, 25.6]. Which is a correct interpretation of this confidence interval?

Select one:

**A. There is a 95% chance that both** and  **are each between 22.3 and 25.6.**

**B. There is a 95% chance that both** or  **is between 22.3 and 25.6.**

**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 22.3 and 25.6.**

**F. The sample mean***x̄***is likely between 22.3 and 25.6. The procedure used gives a confidence interval containing the sample mean***x̄***for 95% of samples.**

Question 5

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

Question 6

(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 7

(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 8

(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 determine skewness.**

Question 9

(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 10

(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 11

(10 pts) With respect to the last problem, let’s say that Ver 2, Ver 3 and Ver 4 used video content while Ver 1 did not. You would thus like to compare the average of the click rates of Version 2,3 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 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 between the *lsmeans* call and the r*un* call unde the comment.

**proc** **glm** data = WebsiteTest;

class WebsiteType;

model ClickRate = WebsiteType;

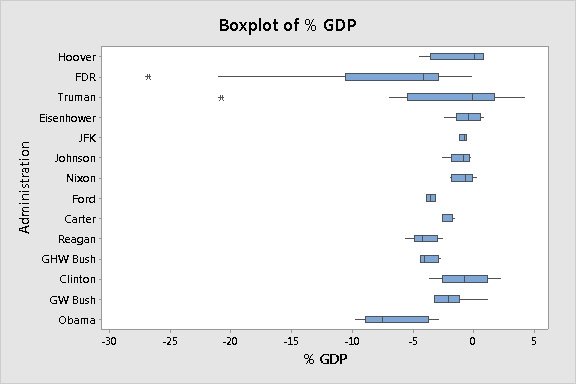
lsmeans WebsiteType / pdiff;

**/\* Place contrast code below. \*/**

estimate “Original versus Average of Ver 2,3 and 4” WebsiteType 3 0 -1 -1 -1 / Divisor = 3;

**run**;

Question 12



(3pts) The box plots for the % GDP under different presidents are displayed above. The box plot for FDR is consistent with what type of skew?

**A. Left Skewed Data**

**B. Symmetric Data**

**C. Uniformly Distributed Data**

**D. Right Skewed Data**

Question 13

(15 pts) Assume a study was conducted in which there were 4 groups and 10 observations were conducted per group. When each of the four groups are allowed to have their own mean (the separate means model), the sum of squared residuals between each groups 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. We didn’t talk much about actually finding the pvalue so you may leave that blank or fill it in for 1 extra point.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **df** | **SS** | **MS** | **F** | **Pvalue** |
| **Model** | **3** | **45** | **15** | **5** | **.00532** |
| **Error** | **36** | **108** | **3** |  |  |
| **Total** | **39** | **153** |  |  |  |

The critical value for this test was 2.85. If your test statistic is greater than this critical value you should reject the null hypothesis. With this information, write a conclusion describing the results of your ANOVA.

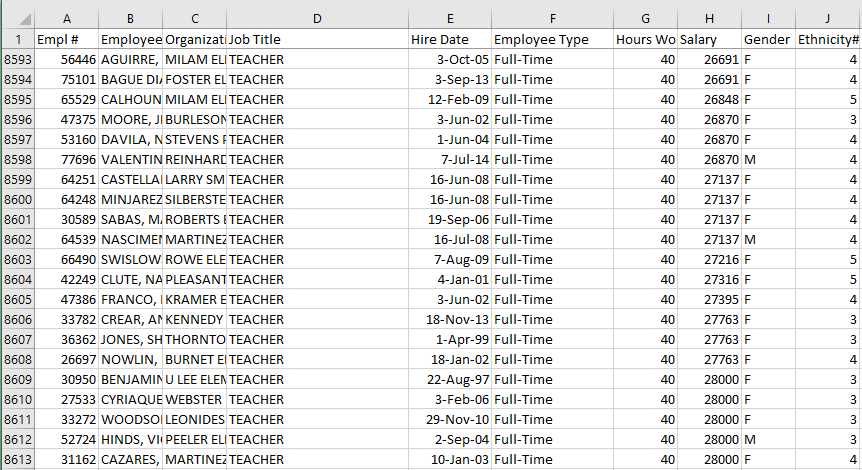


There is strong evidence to suggest that the means between at least 1 pair of groups are different (pvalue = .00532).

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.csv***

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

**We would like to test the claim that men have a higher average mean / median salary than women.**

* 1. Address the assumptions

Assumptions of T-Test:

1. **Normality:** Through inspection of the histograms and qq plots, there is strong evidence against the distribution of both groups salary being normal. As is usually the case with salary data, there is strong evidence of the distribution being right skewed. However, there is very large sample size for both men and women thus the central limit theorem will ensure that the sampling distribution of the sample mean is normally distributed and thus the t-test will be robust to this assumption in this case.
2. **Equal Standard deviation:** Visually, the data look consistent with data that have distributions with equal standard deviations. The F-Test of homoskedasticity is not appropriate to use in this case since the data appear to be heavily right skewed (no normal) although again, visually and intuitively the data appear to come from distributions with the same standard deviations.
3. **Independence:** We will assume the data are independent both within the gender groups and between the groups.
   1. 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.)

1. Critical Values: The sample sizes are so large here that are basically dealing with normally distributed sample means thus we will use critical values of +/- 1.96.
2. Test Statistic:

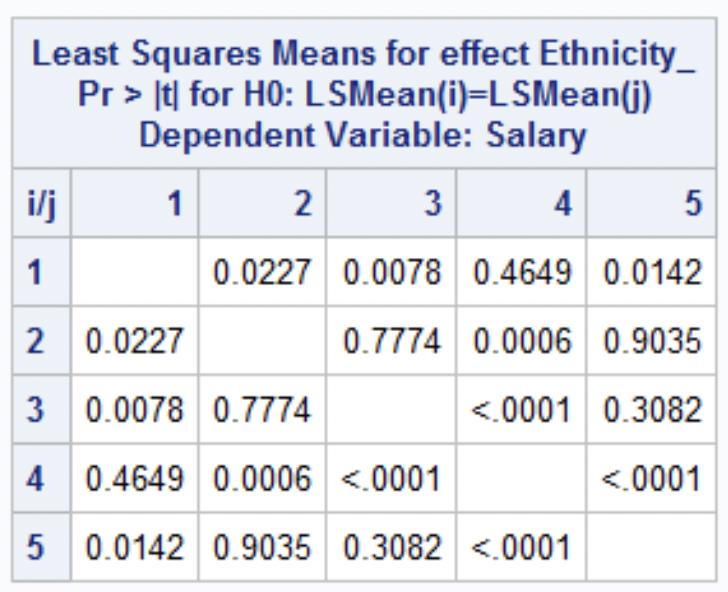
t = -/01

1. P-Value = .9946
2. Fail To Reject Ho
3. The is not sufficient evidence to suggest that the mean salary of males is different from that of females (pvalue = .9946). A 95% confidence interval for the difference between the mean and female mean salaries is ($275.80, $273.90).
   1. Provide a scope of inference.

Gender cannot be assigned randomly to each subject thus this is an observational study. Since we failed to reject, the question of causation v. association is a moot point. The sample, however, was not a random sample from the population thus this conclusion cannot be generalized to the population of Texas educational ISD employees although the results are still compelling.

***NOTE: A Rank Sum Test could have been run here as well if the analyst wanted inference on the median rather than the mean. We would not have to run the rank sum test because of assumptions; the motivation would be to get inference on the median.***

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 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.



The Bonferroni correction her is simple as we only need to multiply the pvalues of interest by 3 (the number of comparisons we were initially interested in). Therefore, at the alpha = .05 level of significance:

We do not see a statistically significant difference in means between groups 1 and 2 (Bonferroni adjusted pvalue = .066).

We do see a statistically significant difference in the mean salaries between groups 1 and 5 (pvalue = .0426).

We see strong evidence of a difference in mean salaries between groups 2 and 4 (pvalue = .0018).

1. (10 pts) Finally, the state’s lawyer was interested in investigating if there was a significant difference in mean or median 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 or median salaries. Please also include a 95% confidence interval in your answer and show your work in calculating this interval.

Version 2 is the correct output to use here in order to compare the average. Using this result, we see that there is not sufficient evidence to suggest that the mean salary of the 5th ethnic group is different than the mean salary of the 1st, 2nd and 3rd ethnic groups together. We are 95% confident that the true difference in mean salaries is contained in the interval (-$1,146.88, $40.88). Since $0 is in the interval, it is plausible given the data that the mean salary of the 5th ethnic group is the same as the average of ethnic groups 1,2 and 3.

-$553 +/- 1.96\*$303 =(-$1,146.88, $40.88)

That’s all folks!

I hope you have a great weekend!

It is certainly well deserved!

(Except the code is Below!)

Output

|  |  |
| --- | --- |
| **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**; |
|  |
|  |