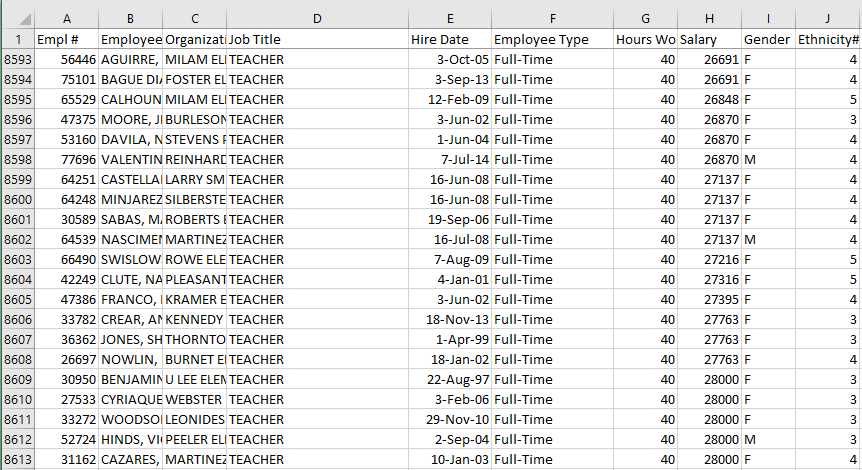
MSDS 6371 SPRING 2020 Midterm AQ Solution

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:

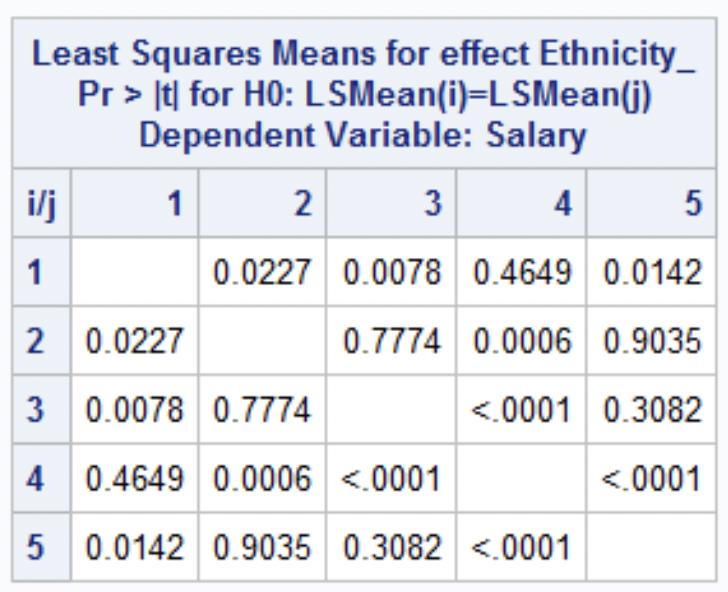
t = -.01

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 (pvalue = .0682). 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**; |
|  |
|  |