UNIT 2 HW

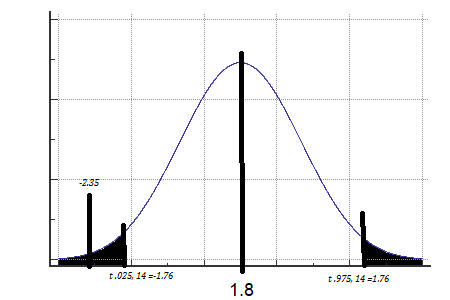
1. The world’s smallest mammal is the bumblebee bat, also known as the Kitti’s hog nosed bat. Such bats are roughly the size of a large bumblebee! Listed below are weights (in grams) from a sample of these bats. Test the claim that these bats come from the same population having a mean weight equal to 1.8 g.  *(Beware: This data is not the same as in the lecture slides!)*

Sample: 1.7 1.6 1.5 2.0 2.3 1.6 1.6 1.8 1.5 1.7 1.2 1.4 1.6 1.6 1.6

* 1. Perform a complete analysis using SAS. Use the six step hypothesis test with a conclusion that includes a statistical conclusion, a confidence interval and a scope of inference (as best as can be done with the information above … there are many correct answers given the vagueness of the description of the sampling mechanism.)

***1. H0: Ha:***

***2. Critical values = -1.76131, 1.76131***



***3. Test statistic =*** 

***4. p-value =*** 

***5. Because our t statistic is in the critical value range and p value =.0324 we can reject H0***

***6. The evidence (from the sample provided) suggests that the true mean weight of Bumblebee Bats is not equal to 18 (p-value=0.0342)***

***95% Confidence Interval = ***

* 1. Inspect and run this R Code and compare the results (t statistic, p-value and confidence interval) to those you found in SAS. To run the code, simply copy and paste the below code into R.

*sample = c(1.7, 1.6, 1.5, 2.0, 2.3, 1.6, 1.6, 1.8, 1.5, 1.7, 1.2, 1.4, 1.6, 1.6, 1.6)*

*t.test(x=sample, mu = 1.8, conf.int = "TRUE", alternative = "two.sided" )*

***One Sample t-test***

***data: sample***

***t = -2.3457, df = 14, p-value = 0.03424***

***alternative hypothesis: true mean is not equal to 1.8***

***95 percent confidence interval:***

***1.506466 1.786868***

***sample estimates:***

***mean of x***

***1.646667***

1. In the United States, it is illegal to discriminate against people based on various attributes. One example is age. An active lawsuit, filed August 30, 2011, in the Los Angeles District Office is a case against the American Samoa Government for systematic age discrimination by preferentially firing older workers. Though the data and details are currently sealed, suppose that a random sample of the ages of fired and not fired people in the American Samoa Government are listed below:

**Fired**

34 37 37 38 41 42 43 44 44 45 45 45 46 48 49 53 53 54 54 55 56

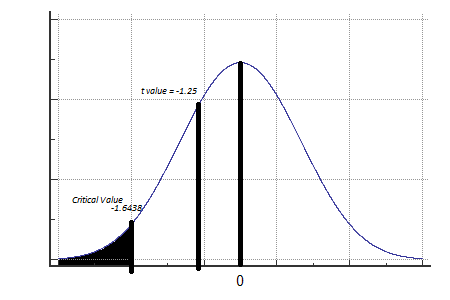
**Not fired**

27 33 36 37 38 38 39 42 42 43 43 44 44 44 45 45 45 45 46 46 47 47 48 48 49 49 51 51 52 54

* + - * 1. Perform a permutation test to test the claim that there is age discrimination. Provide the Ho and Ha, the p-value, and full statistical conclusion, including the scope. Note: this was an example in Live Session 1. You may start from scratch or use the sample code and PowerPoints from Live Session 1.

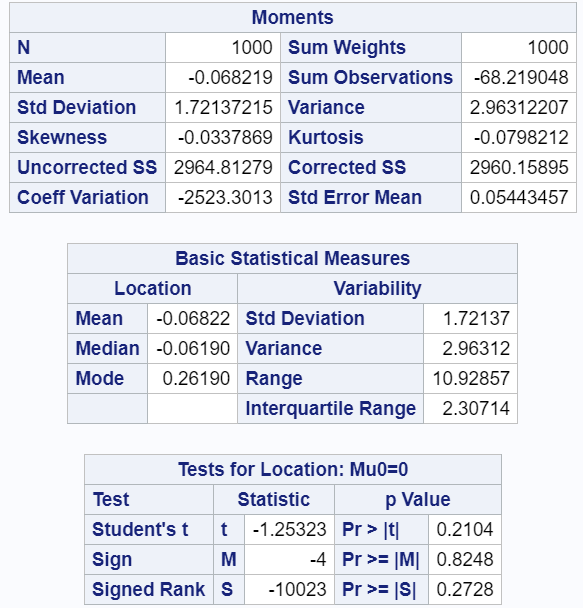
***1. Permutation test H0: Ha:***

***2. Critical Value a = .05, df = 998, t = -1.6438***

******

***3. t statistic = -1.25323***

***4. p value = 0.2104***

******

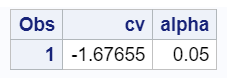
***5. Because the t statistic is not in the critical value range and p value = 0.2104***

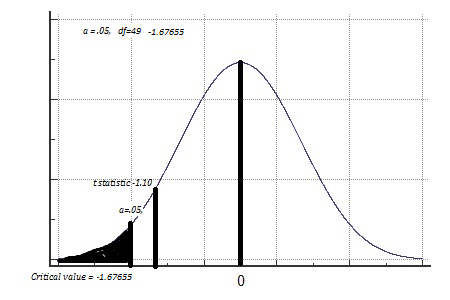
***we fail to reject H0.***

***6. There is not enough evidence to suggest that the mean age of terminated employees is higher than retained employees. (p-value=0.2104)***

* + - * 1. Now run a two-sample t-test appropriate for this scientific problem. (Use SAS.) *(Note: we may not have talked much about a two-sided versus a one-sided test. If you would like to read the discussion on pg. 44 (Statistical Sleuth****), you can run a one-sided test if it seems appropriate****. Otherwise, just run a two-sided test as in class. There are also examples in the Statistics Bridge Course).* Be sure to include all six steps, a statistical conclusion and scope of inference.

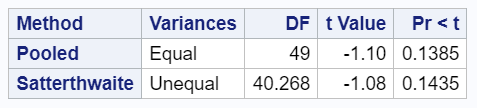
***1. H0: Ha:***

***2. Critical Value a = .05, df = 49, ***

******

***3. t statistic = -1.10***

***4. p value = .1385***

******

***5. Because the t statistic is not in the critical value range and p value = .1385 we fail to reject H0.***

***6. There is not enough evidence to suggest that the mean age of terminated employees is higher than the mean age retained employees. (p-value=0.1385)***

* + - * 1. Compare the two pvalues generated in a and b.

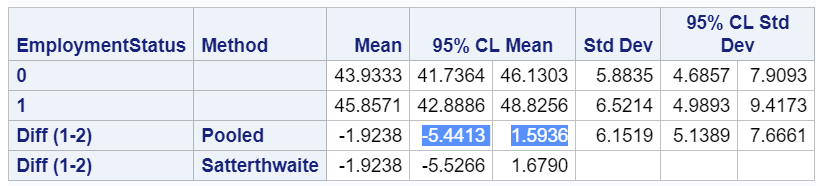
***The permutation test p-value is larger than the and the non-permutated test though they yield the conclusion (fail to reject Ho).***

***However, I did notice the p value for the permutation test varies significantly every time I run the same code while the non-permutated does not.***

The jury wants to see a range of plausible values for the difference in means between the fired and not fired groups. Provide them with a confidence interval for the difference of means and an interpretation.

***It is plausible that the average difference in the age of Retained employees - Fired Employees can range from negative 5.4 years to positive 1.6 years.***

***We are 95% confident that the difference in the average ages of Retained Employees - Fired Employees can range from -5.4 years to 1.6 years***

******

* + - * 1. Given the sample standard deviations from SAS, calculate by hand

Pooled standard deviation (sp)

The standard error of ()

* + - * 1. Inspect and run this R Code and compare the results (t statistic, p-value, and confidence interval) to those you found in SAS. To run the code, simply copy and paste the code below into R.

*Fired = c(34, 37, 37, 38, 41, 42, 43, 44, 44, 45, 45, 45, 46, 48, 49, 53, 53, 54, 54, 55, 56)*

*Not\_fired = c(27, 33, 36, 37, 38, 38, 39, 42, 42, 43, 43, 44, 44, 44, 45, 45, 45, 45, 46, 46, 47, 47, 48, 48, 49, 49, 51, 51, 52, 54)*

*t.test(x = Fired, y = Not\_fired, conf.int = .95, var.equal = TRUE, alternative = "two.sided")*

***data: Fired and Not\_fired***

***t = 1.0991, df = 49, p-value = 0.2771***

***alternative hypothesis: true difference in means is not equal to 0***

***95 percent confidence interval:***

***-1.593635 5.441254***

***sample estimates:***

***mean of x mean of y***

***45.85714 43.93333***

***Because this Rcode has x = Fired, y = Not\_fired the ci= -1.593635, 5.441254***

***My SAS code has x = Not\_fired and y = Fired. The resulting range is the same size but the values are opposite the R code with ci -5.4 years, 1.6 years.***

***Although the t statistic and p-value are the same. (exception is that R code is 2 tailed)***

In the last homework, it was mentioned that a Business Stats professor here at SMU polled his class and asked students them how much money (cash) they had in their pockets at that very moment. The idea was that we wanted to see if there was evidence that those in charge of the vending machines should include the expensive bill / coin acceptor or if it should just have the credit card reader. However, another professor from Seattle University was asked to poll her class with the same question. Below are the results of our polls.

**SMU**

34, 1200, 23, 50, 60, 50, 0, 0, 30, 89, 0, 300, 400, 20, 10, 0

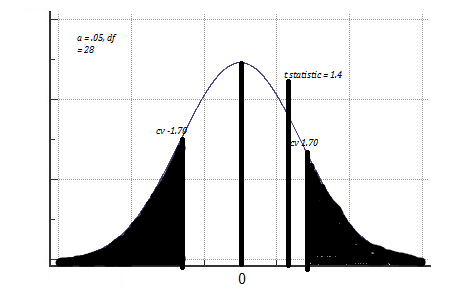
**Seattle U**

20, 10, 5, 0, 30, 50, 0, 100, 110, 0, 40, 10, 3, 0

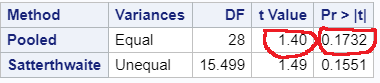
* + - * 1. Run a two sample t-test to test if the mean amount of pocket cash from students at SMU is different than that of students from Seattle University. Write up a complete analysis: all 6 steps including a statistical conclusion and scope of inference (similar to the one from the PowerPoint). (This should include identifying the Ho and Ha as well as the p-value.) Also include the appropriate confidence interval. **FUTURE DATA SCIENTIST’S CHOICE!: YOU MAY USE SAS *OR* R TO DO THIS PROBLEM!**

***1. H0: Ha:***

***2. Critical Value a = .05, df = 28 ***

******

***3. t statistic = 1.4***

***4. p value = .1732 ***

***5. Because the t statistic is not in the critical value range and p value = .1732 we fail to reject H0.***

***6. There is not enough evidence to conclude that there is a difference between the cash carried SMU students is different than the Seattle students with 95% confidence interval -53.3711, 282.6. (p .1732). Inference cannot be drawn of confounding variables. We cannot determine if there is a difference in population means because this data is only from stats students. Also, the data from Seattle is stale; changes may have occurred since last year; SMU and Seattle may have different demographics (confounding variables). Students are volunteering the information; it doesn’t seem that we are asking them to show the cash but tell us how much they have.***

* + - * 1. Compare the p-value from this test with the one you found from the permutation test from last week. Provide a short 2 to 3 sentence discussion on your thoughts as to why they are the same or different.

***The p value from last week’s permutation test was .149 and this week’s non-permutation test .1732. These are different because the permutation test randomly chooses samples of the samples different combinations (1000 in this case). The p-value for the permutation test is lower because the number of samples size is larger (though the real sample size is static). It is assumed that the permutation number could be more accurate however simply re-sampling to gain a lower p-value doesn’t necessarily mean that it is practical.***

1. A. Calculate the estimate of the pooled standard deviation from the Samoan discrimination problem. Use this estimate to build a power curve. Assume we would like to be able to detect effect sizes between 0.5 and 2 and we would like to calculate the sample size required to have a test that has a power of .8. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE’s lecture. Instead of using **groupstddevs**, use **stddev** since we are using the pooled estimate.

***proc power;***

***twosamplemeans***

***meandiff = 1.9238***

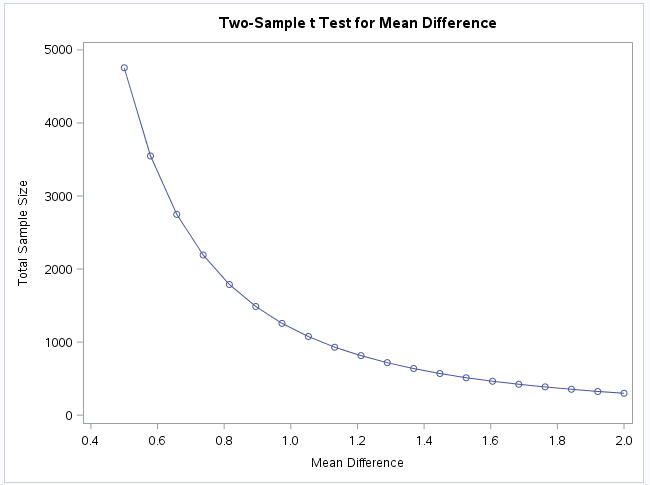
***stddev = 6.1519***

***ntotal = .***

***power = .8;***

***plot x = effect min =.5 max = 2;***

***run;***



B. Now suppose we decided that we may be able to live with slightly less power if it means savings in sample size. Provide the same plot as above but this time calculate curves of sample size (y-axis) vs. effect size (.5 to 2) (x axis) for power = 0.8, 0.7, and 0.6. There should be three plots on your final plot. Simply cut and paste your power curve and SAS code. HINT: USE THE CODE FROM DR. McGEE’s lecture. Instead of using **groupstddevs**, use **stddev** since we are using the pooled estimate.

***proc power;***

***twosamplemeans***

***meandiff = 1.9238***

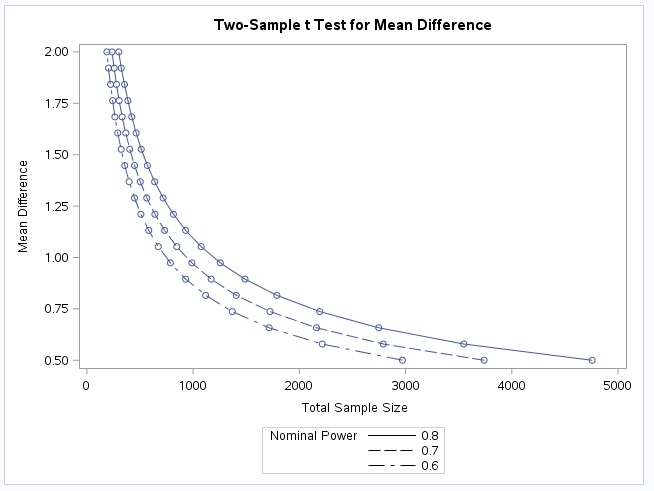
***stddev = 6.1519***

***ntotal = .***

***power = .8 .7 .6;***

***plot y = effect min =.5 max = 2;***

***run;***



C. Using similar code, estimate the savings in sample size from a test aimed at detecting an effect size of 0.8 with a power of 80% versus a power of 60%.

***proc power;***

***twosamplemeans***

***meandiff = 1.9238***

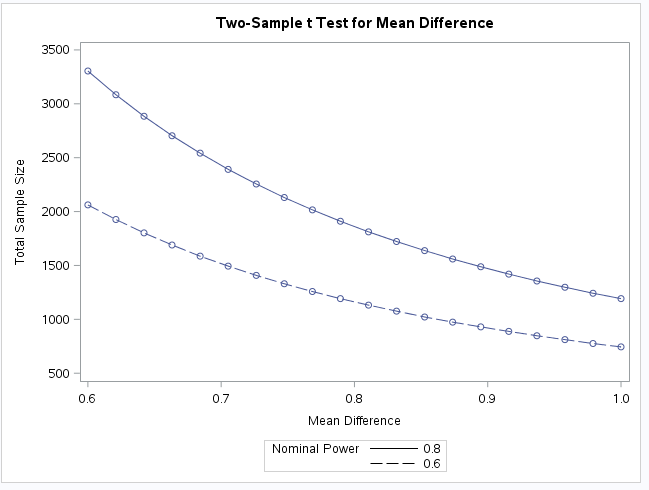
***stddev = 6.1519***

***ntotal = .***

***power = .8 .6;***

***plot x = effect min = .6 max =1;***

***run;***

******

Note: You will learn how to do this in R in a future HW!