MSDS 6371 HW 2

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

a. 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.)

State the null and alternative hypothesis:
The null hypothesis is that the mean weight of the bats sampled is 1.8 g. The alternative hypothesis is that the mean weight of the bats sampled is not 1.8 g.

$$H_o: \mu = 1.8\ g$$
$$H_A: \mu \neq 1.8\ g$$

Calculate

The UNIVARIATE Procedure
Variable: size

Moments			
N	15	Sum Weights	15
Mean	1.64666667	Sum Observations	24.7
Std Deviation	0.25317037	Variance	0.06409524
Skewness	1.09630411	Kurtosis	2.72573796
Uncorrected SS	41.57	Corrected SS	0.89733333
Coeff Variation	15.374719	Std Error Mean	0.06536831

Basic Statistical Measures			
Location		Variability	
Mean	1.646667	Std Deviation	0.25317
Median	1.600000	Variance	0.06410
Mode	1.600000	Range	1.10000
		Interquartile Range	0.20000

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	25.1906	Pr > |t|	<.0001
Sign	M	7.5	Pr >= |M|	<.0001
Signed Rank	S	60	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	2.3
99%	2.3
95%	2.3
90%	2.0
75% Q3	1.7
50% Median	1.6
25% Q1	1.5
10%	1.4
5%	1.2
1%	1.2
0% Min	1.2

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
1.2	11	1.7	1
1.4	12	1.7	10
1.5	9	1.8	8
1.5	3	2.0	4
1.6	15	2.3	5

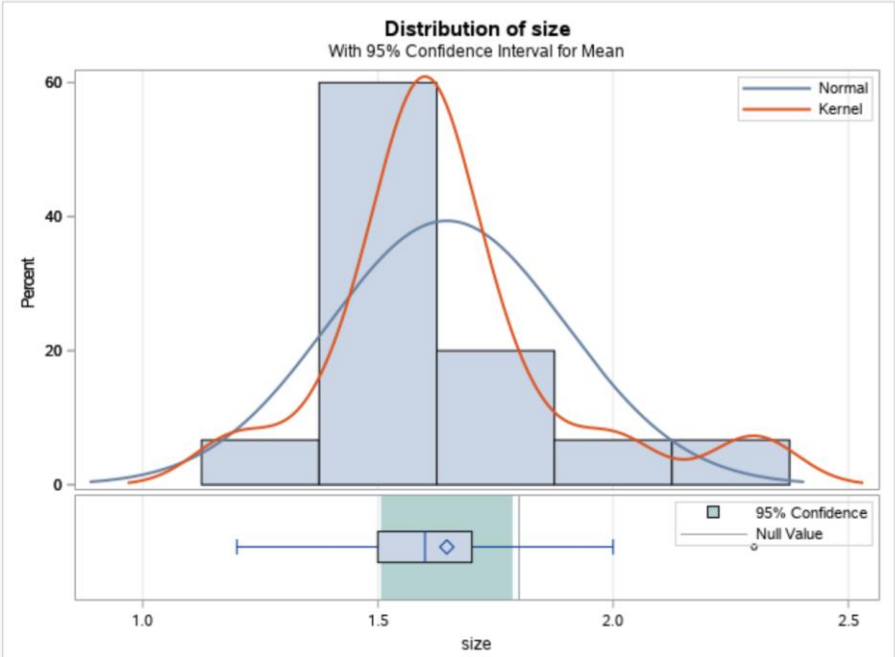
The TTEST Procedure

Variable: size

N	Mean	Std Dev	Std Err	Minimum	Maximum
15	1.6467	0.2532	0.0654	1.2000	2.3000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1.6467	1.5065	1.7869	0.2532	0.1854	0.3993

DF	t Value	Pr > |t|
14	-2.35	0.0342



Statistical conclusion – The evidence suggests the mean weight of the sample does not equal 1.8 g (two-sided p-value 0.0342). The mean weight of the bats sample was 1.65 g, which is 0.15g less than the population mean.

Confidence interval – A 95% confidence interval for the difference in means is 1.51g and 1.79g. This means that we are 95% confident the difference in means of the two parties lies in this range. Since the range doesn't include 1.8g, this would support our final decision.

Scope of Inference – This was an observational study, in which the sampling procedure was unclear. It is uncertain if the bats acquired were a convenience sample or an actual randomized sample. Regardless, the evidence suggest we must reject the null hypothesis. The mean of the bats' weight was slightly lower than that of the population, as demonstrated in the confidence interval. However, this begs the question if a larger sample size would echo this conclusion. Considering the sample size can affect the outcome of the result, it would be wise to conduct additional studies with a larger sample size to truly determine whether the mean weight is 1.8g.

b. 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")

The outcome of R supports the decision rendered and produced the same results.

```
> 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 = "t
wo.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
```

2. 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

a. 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 (inference on population and causal inference). 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.

Hypothesis:
The null hypothesis is that there isn't a difference in age mean regarding those fired and currently employed. The alternative hypothesis is that the mean of both groups is different.

$$H_o: \mu_1 = \mu_2 \qquad H_A: \mu_1 \neq \mu_2$$

Permutation test and calculations – 1000 observations were made and the resulting p value was 0.3583 with a t statistic of 0.919111.

Statistical conclusion – Based on the results of the permutation test (p value 0.3583), there is overwhelming evidence in favor of the null hypothesis. The mean from the test was -0.05073333, which is extremely close to 0. This would indicate the difference in means was minor, thus supporting the null hypothesis once again. As a result, we are unable to reject the null hypothesis from the results of the permutation test.

Scope of Inference – Based on the results, the evidence suggests there is no statistically significant difference in the mean different of the fired employees and those still working. This study was an observational study, which would indicate we are unable to draw causal inference as a result of this study. There are a multitude of other variables that were not included in the study, which may or may not have had an impact in the results.

b. 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.

Hypothesis:
The null hypothesis is that there isn't a difference in age mean regarding those fired and currently employed. The alternative hypothesis is that the mean of both groups is different.

$$H_o: \mu_1 = \mu_2 \qquad H_A: \mu_1 \neq \mu_2$$

Two-sample t Test (calculations)

The TTEST Procedure

Variable: Age

Status	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		21	45.8571	6.5214	1.4231	34.0000	56.0000
1		30	43.9333	5.8835	1.0742	27.0000	54.0000
Diff (1-2)	Pooled		1.9238	6.1519	1.7503		
Diff (1-2)	Satterthwaite		1.9238		1.7830		

Status	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		45.8571	42.8886	48.8256	6.5214	4.9893	9.4173
1		43.9333	41.7364	46.1303	5.8835	4.6857	7.9093
Diff (1-2)	Pooled	1.9238	-1.5936	5.4413	6.1519	5.1389	7.6661
Diff (1-2)	Satterthwaite	1.9238	-1.6790	5.5266			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	49	1.10	0.2771
Satterthwaite	Unequal	40.268	1.08	0.2870

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	20	29	1.23	0.6005

Mean age for fired employees was 46 and mean age of currently working employees was 44. The difference in means was about 2 years. The p value was 0.2771 and the t statistic was 1.10.

Statistical conclusion
Based on the results of the t test (2 side p value 0.277), there is strong evidence to support the null hypothesis. The mean difference was under 2 years, which is minor and creates the potential for overlap in the distribution of the samples.

Scope of Inference
The evidence suggests the difference in ages is not statistically significant and may be attributed to chance. While it is true that the mean of those employees fired is 2 years higher than those still working, that number is likely attributed to chance or other variables not measured in this study. Considering this was an observational study, it is difficult to attribute cause-and-effect to the study. Nonetheless, we are able to support the notion that the employees were likely not discriminated based on age when it came to their termination.

c. Compare this p-value to the randomized p-value found in the previous sub-question.

The p value for the permutation test was higher then the p value for the two sample t test. Although they both support the null hypothesis, it begs the question if a larger sample would then yield a much higher p value as well.

d. 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.

The 95% confidence interval ranges from -1.5936 to 5.4413. This means that the difference in mean ages between both groups is 95% likely to fall within this range. Considering the lower range ends at -1.5936,

that is indicative that in some instances, the fired employee may even be younger than the employee still working. This would also indicate we are 95% confident, the worst age difference we will see between the fired employees is 5.4413 years.

e. Given the sample standard deviations from SAS, calculate by hand

i. Pooled standard deviation (s_p)

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}} = \sqrt{\frac{(21-1)6.5214^2 + (30-1)5.8835^2}{(21+30-2)}} = 6.1519$$

ii. The standard error of $(\bar{X}_{FIRED} - \bar{X}_{Not\ Fired})$

$$SE(\bar{x}_{fired} - \bar{x}_{not\ fired}) = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 6.1519\sqrt{\frac{1}{21}+\frac{1}{30}} = 1.7503$$

f. 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")

The results from R are identical to those from SAS.

```
> 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 = T
RUE, alternative = "two.sided")

        Two Sample t-test

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
```

3. 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

a. 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!**

Hypothesis:

The null hypothesis is that there is no difference in the mean cash carried by students in SMU or Seattle U. The alternative hypothesis is that there is a difference in the mean cash carried by the students in either SMU or Seattle U.

$$H_o: \mu_1 = \mu_2 \qquad\qquad H_A: \mu_1 \neq \mu_2$$

Calculations:

The mean of cash held by SMU students is $141.60, while Seattle U students had a mean of cash amounting to $27.00. The difference in means is $114.60. Results produced a p value of 0.1732.

The TTEST Procedure

Variable: cash

school	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
0		16	141.6	304.3	76.0670	0	1200.0
1		14	27.0000	36.7193	9.8136	0	110.0
Diff (1-2)	Pooled		114.6	224.1	82.0131		
Diff (1-2)	Satterthwaite		114.6		76.6974		

school	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
0		141.6	-20.5079	303.8	304.3	224.8	470.9
1		27.0000	5.7989	48.2011	36.7193	26.6198	59.1564
Diff (1-2)	Pooled	114.6	-53.3711	282.6	224.1	177.8	303.1
Diff (1-2)	Satterthwaite	114.6	-48.3948	277.6			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	1.40	0.1732
Satterthwaite	Unequal	15.499	1.49	0.1551

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	15	13	68.66	<.0001

Statistical conclusion:

Based on the results, there is insufficient evidence to reject the null hypothesis (2 sided, p value 0.1732). With a t statistic of 1.40 and a wide range in the 95% confidence intervals (-$53.38 - $282.60), it appears the small sample size is being affected by the outliers in the sample.

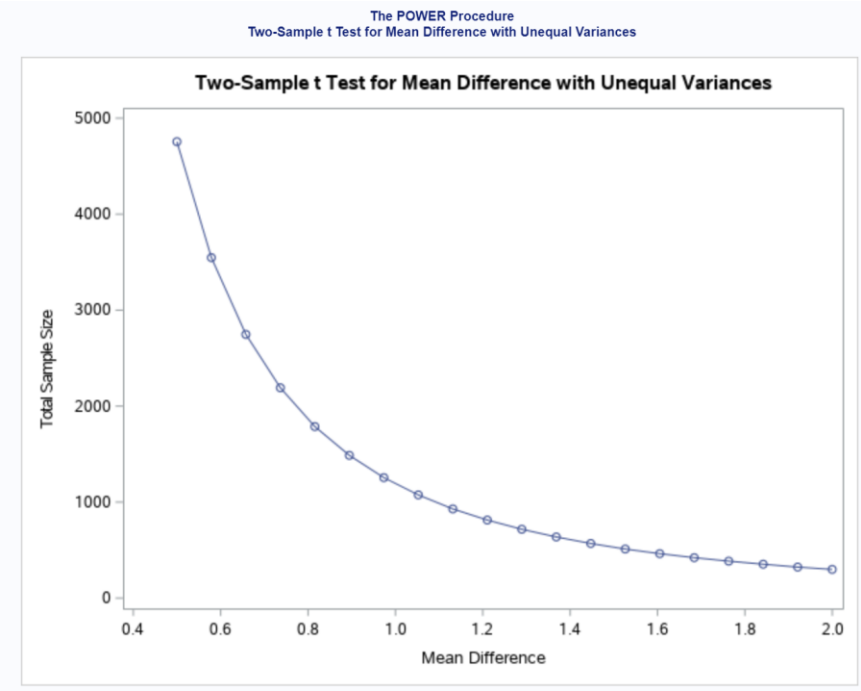
Scope of Inference:

Considering the study was an observational study, we are unable to draw any causal relationships from this data. As a result, the information is only relevant to the students that were part of the study. The sampling was a convenience sample, which also confirms it may not be a true indication of the student body population. Although the data suggests we are unable to reject the null, it appears other factors may have affected the results. In this instance, it would be wise to cast a wider net and increase the sample to the general student body population while also increasing the sample size.

b. 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 the permutation test was 0.49, whereas the p value from the two sample t test was 0.1732. This would indicate that there is a probability in that there may be other variables at play that were not taken into account. Based on these values, one of the conclusions would have resulted in a type 1 or type 2 error.

4. 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.

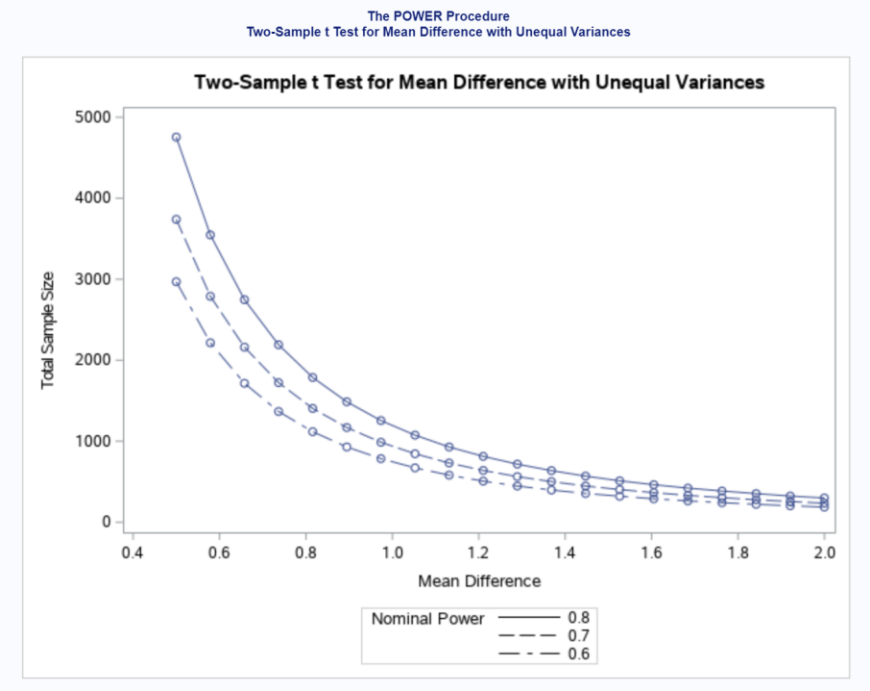



```
data Age_Discrimination;
input Status Age; *0 = Fired and 1 = Employed;
datalines;
0 34
0 37
```

0 37
0 38
0 41
0 42
0 43
0 44
0 44
0 45
0 45
0 45
0 46
0 48
0 49
0 53
0 53
0 54
0 54
0 55
0 56
1 27
1 33
1 36
1 37
1 38
1 38
1 39
1 42
1 42
1 43
1 43
1 44
1 44
1 44
1 45
1 45
1 45
1 45
1 46
1 46
1 47
1 47
1 48
1 48
1 49
1 49
1 51
1 51
1 52
1 54
;

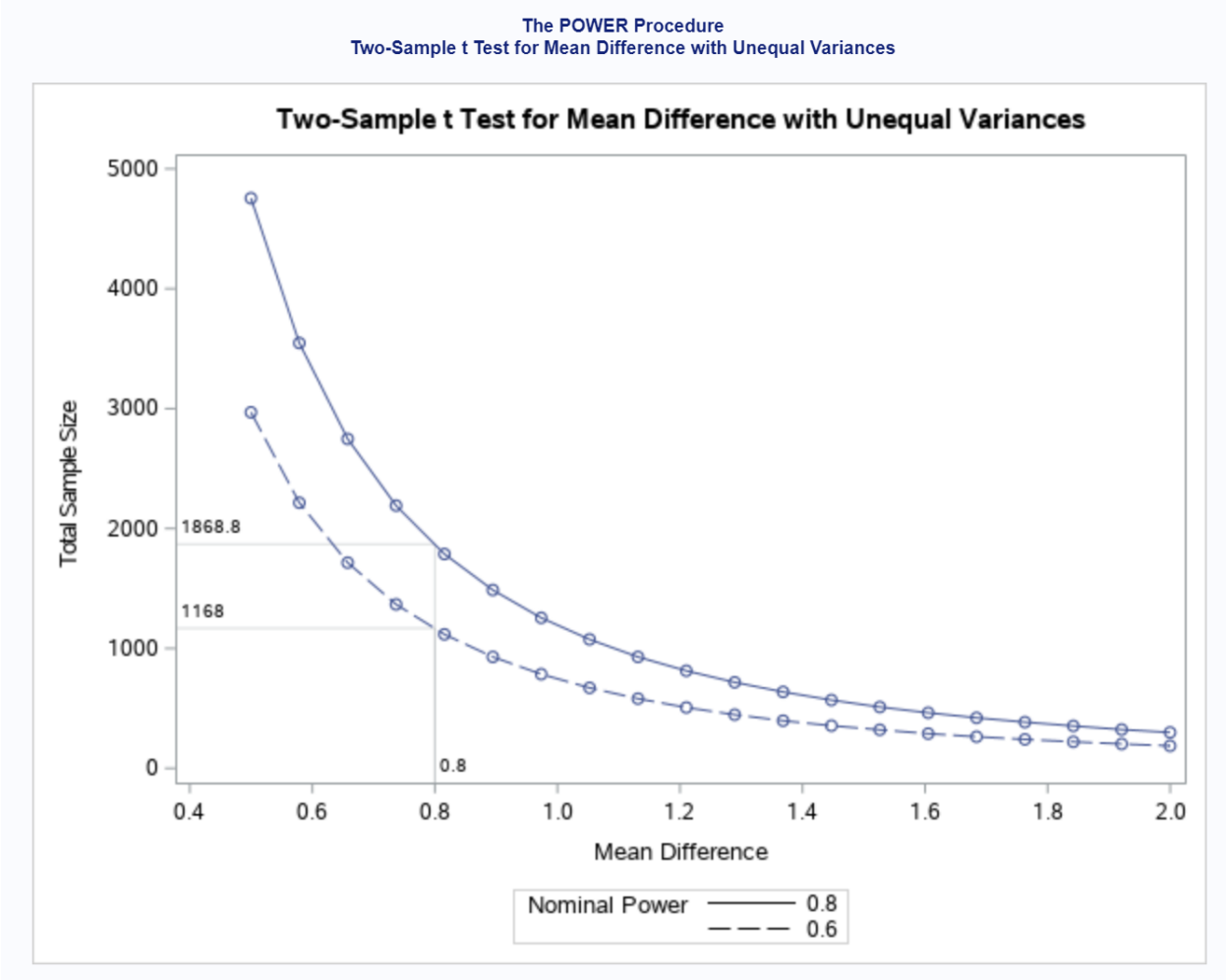
```
proc power;
twosamplemeans test=diff_satt
alpha=0.05
sides=2
meandiff= 0.5 2
stddev= 6.1519
power = 0.8
ntotal= .;
plot x=effect min=0.5 max=2;
run;
```

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. The effect size here refers to a difference in means, though there are many effect size metrics, such a Cohen's D.




```
Data age_discrimination
…
proc power;
twosamplemeans test=diff_satt
alpha=0.05
sides=2
meandiff= 0.5 2
stddev= 6.1519
power = 0.8 0.7 0.6
ntotal= .;
plot x=effect min=0.5 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%.




```
Data age_discrimination;
…
proc power;
twosamplemeans test=diff_satt
alpha=0.05
sides=2
meandiff= 0.5 2
stddev= 6.1519
power = 0.8 0.6
ntotal= .;
plot x=effect min=0.5 max=2 xopts=(crossref=YES ref=0.8);
run;
```

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