Sean Kennedy Homework 2 – Statistical Foundations for Data Science

Question 1:

Step 1 Ho: =1.8g

Ha:  ≠1.8g

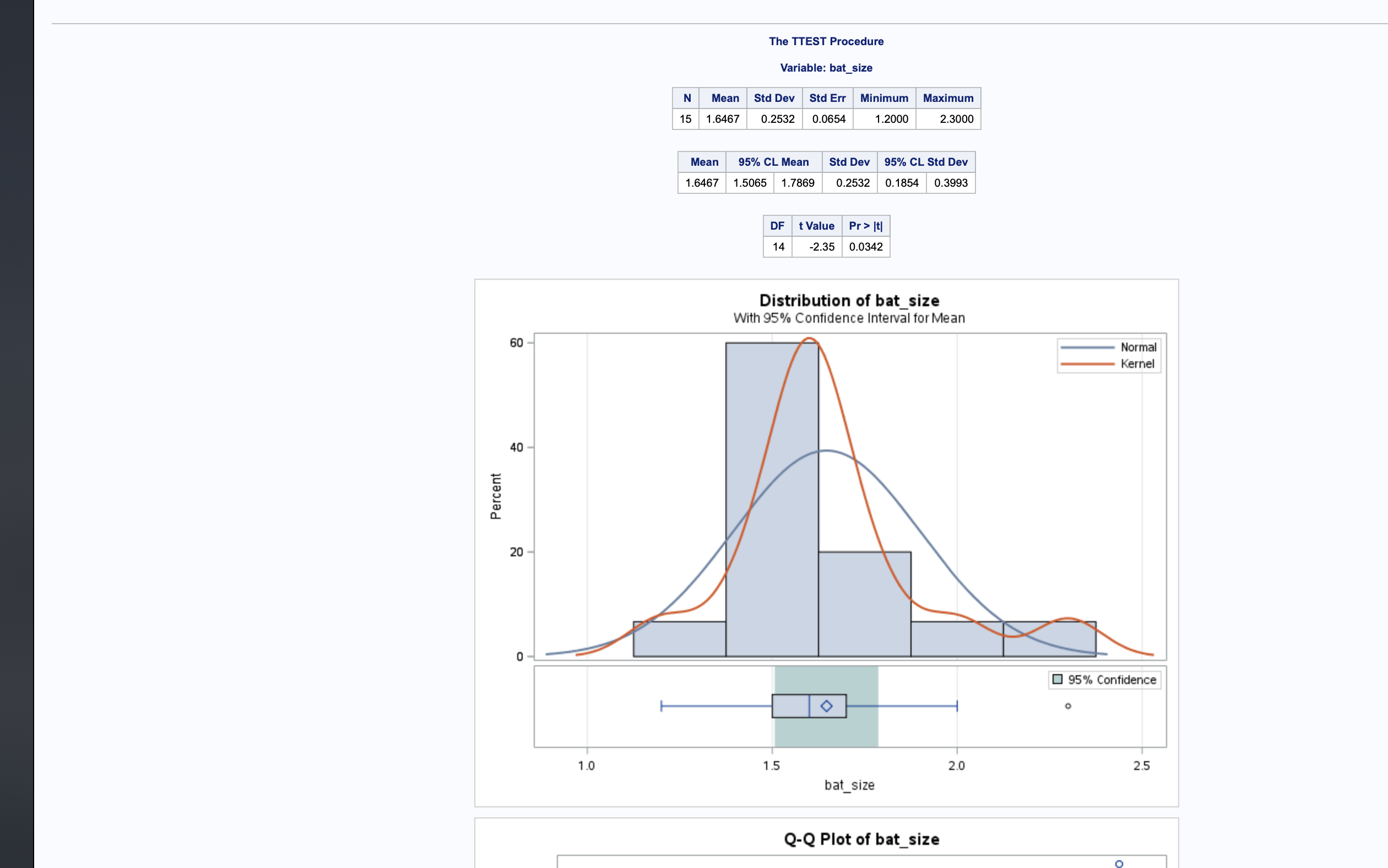
Steps 2/3:

Performing a 2-Sided T-Test in SAS

At 95% CL (p 2.14479)

=0.05

t= -2.35

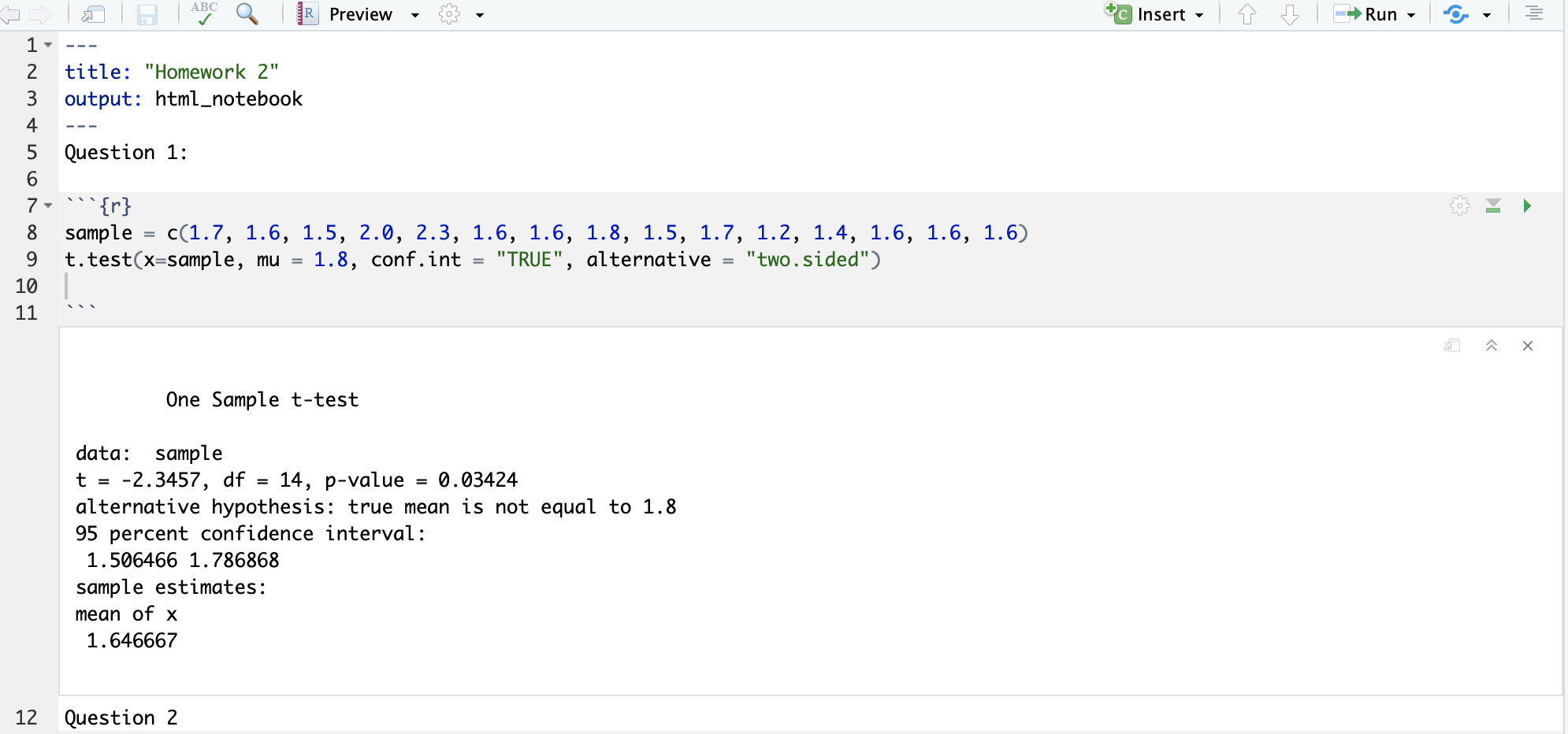


Step 4: p-value (statistical likelihood) = 0.0342

Step 5: p-value is less than alpha, therefore we reject Ho. Mean of 1.8 lies outside the 95%CL (upper bound 1.7869)

Step 6:

The t-test results demonstrated that the assumed mean of 1.8g was not accurate. Study suffers from small sample size relative to population and no indication was made that the sampling was done randomly – hence the result of this study cannot be extended to the bumblebee bat population.



Question 2:

1. Use SAS for permutation test

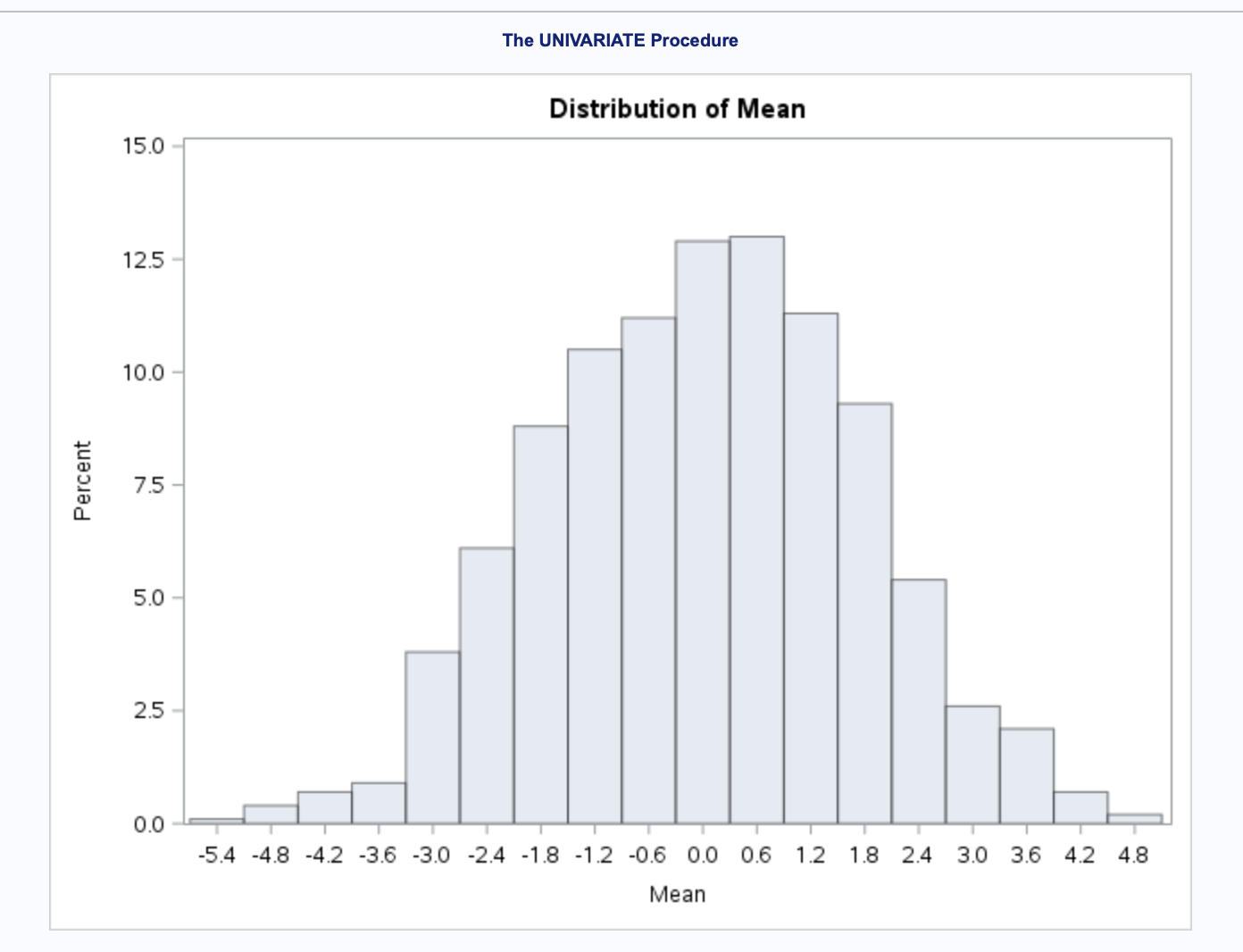
Step 1 Ho: m1=m2 (sample means equal)

Ha: m1 ≠ m2 (sample means not equal)

Steps 2/3:

Performing permutation test in SAS:





Step 4: p-value (statistical likelihood) = 0.9985

Step 5: p-value is greater than alpha, therefore we accept Ho. The means of the two populations are equal.

Step 6:

The t-test results demonstrated that there is not a significant difference between the mean age of employees that were fired versus those that were not. The high p-value (0.9985) and expected mean difference calculated in the random permutation test (-0.0014) indicate a strong causal relationship though it is impossible to make causal inferences due to the nature of observational studies. Though statistically significant – the scope of this analysis is limited to a population on a remote island in the pacific and hence difficult to extend to a larger population.

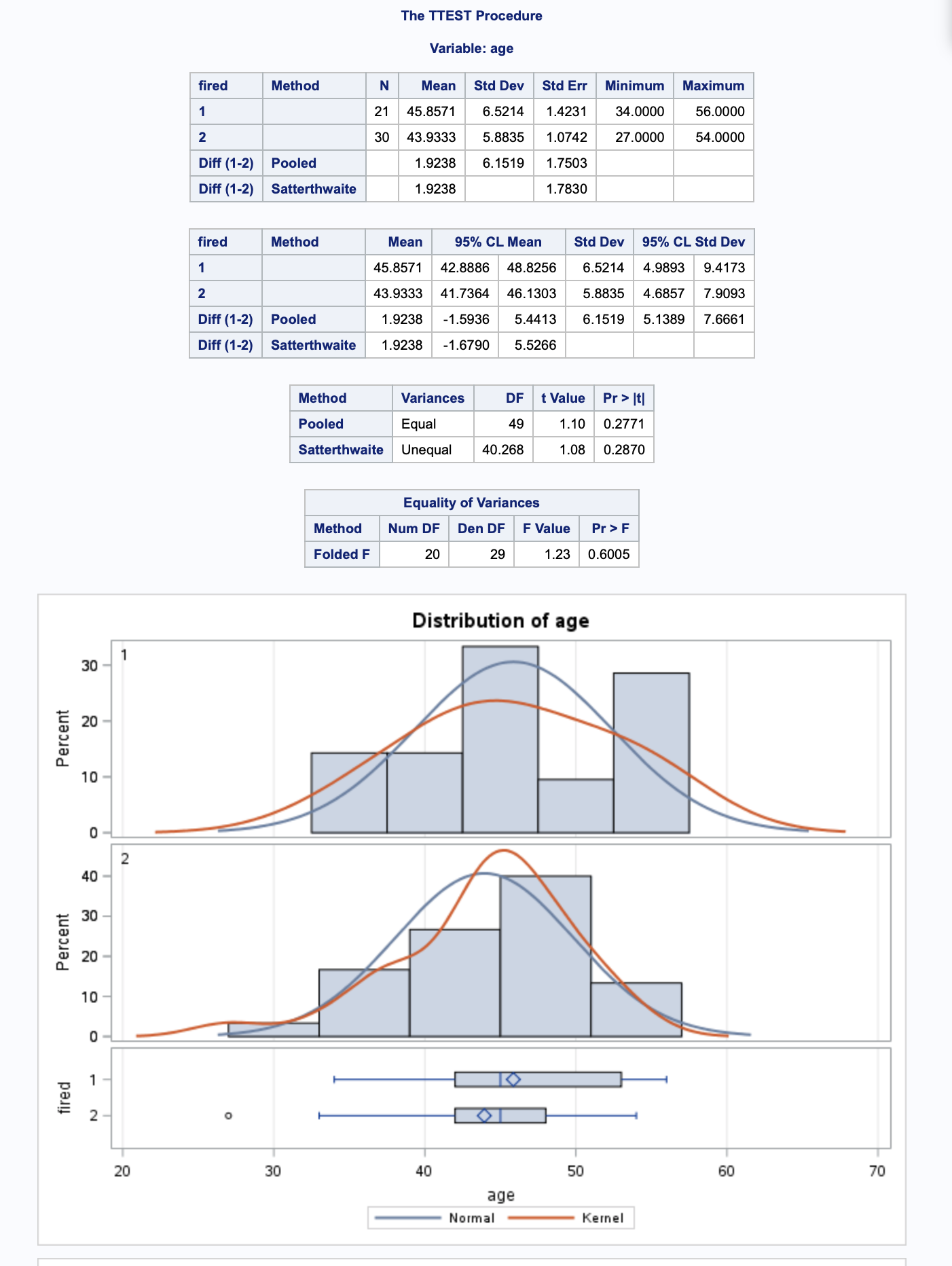
1. Use SAS for 2 sided t-test

Step 1 Ho: m1=m2 (sample means equal)

Ha: m1 ≠ m2 (sample means not equal)

Steps 2/3:

Performing t-test in SAS



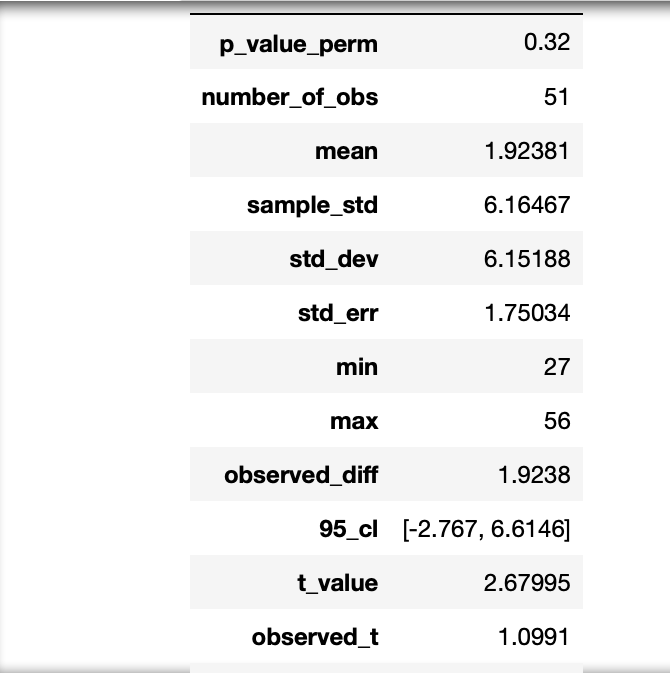
Step 4: p-value (statistical likelihood) = 0.2771

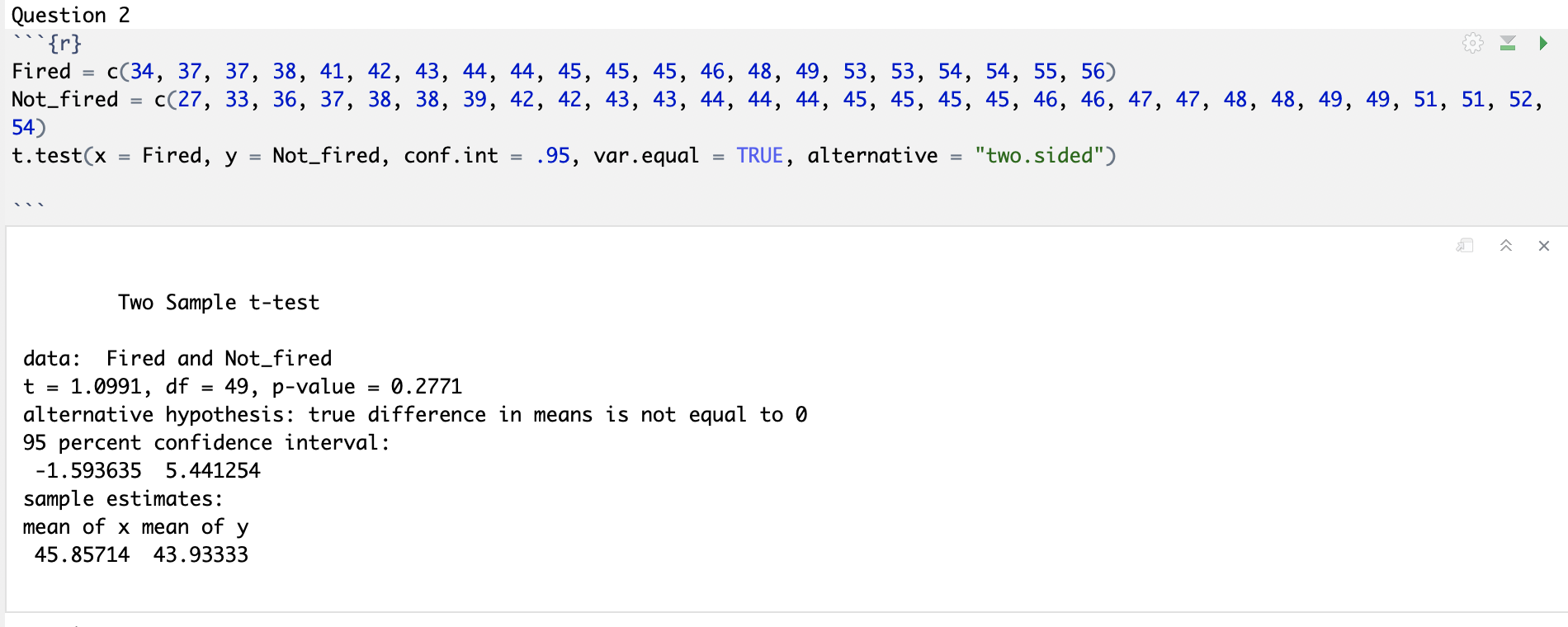
Step 5: p-value is greater than alpha, therefore we accept Ho. The means of the two populations are equal.

Step 6:

The t-test results demonstrated that there is not a significant difference between the mean age of employees that were fired versus those that were not. The pooled mean difference between the two samples (1.9238) falls within the 95% confidence interval. The relatively high p-value (0.0.2771) indicates a relationship though it is not possible to infer causality due to the nature of observational studies in general. Though statistically significant – the scope of this analysis is limited to a population on a remote island in the pacific and hence difficult to extend to a larger population.

1. The p-values produced in the random permutation test is significantly higher (0.9985 vs 0.2771) than the p-value in the 2-sided t-test.
2. From the two sided t-test, we can see that the expected mean difference between the two samples should fall between 5.1389 and 7.6661 if we are to assume the null hypothesis to be true. In this case – the observed pooled mean difference of 1.9238 is almost centered within the 95% interval.
3. See python code



1. 

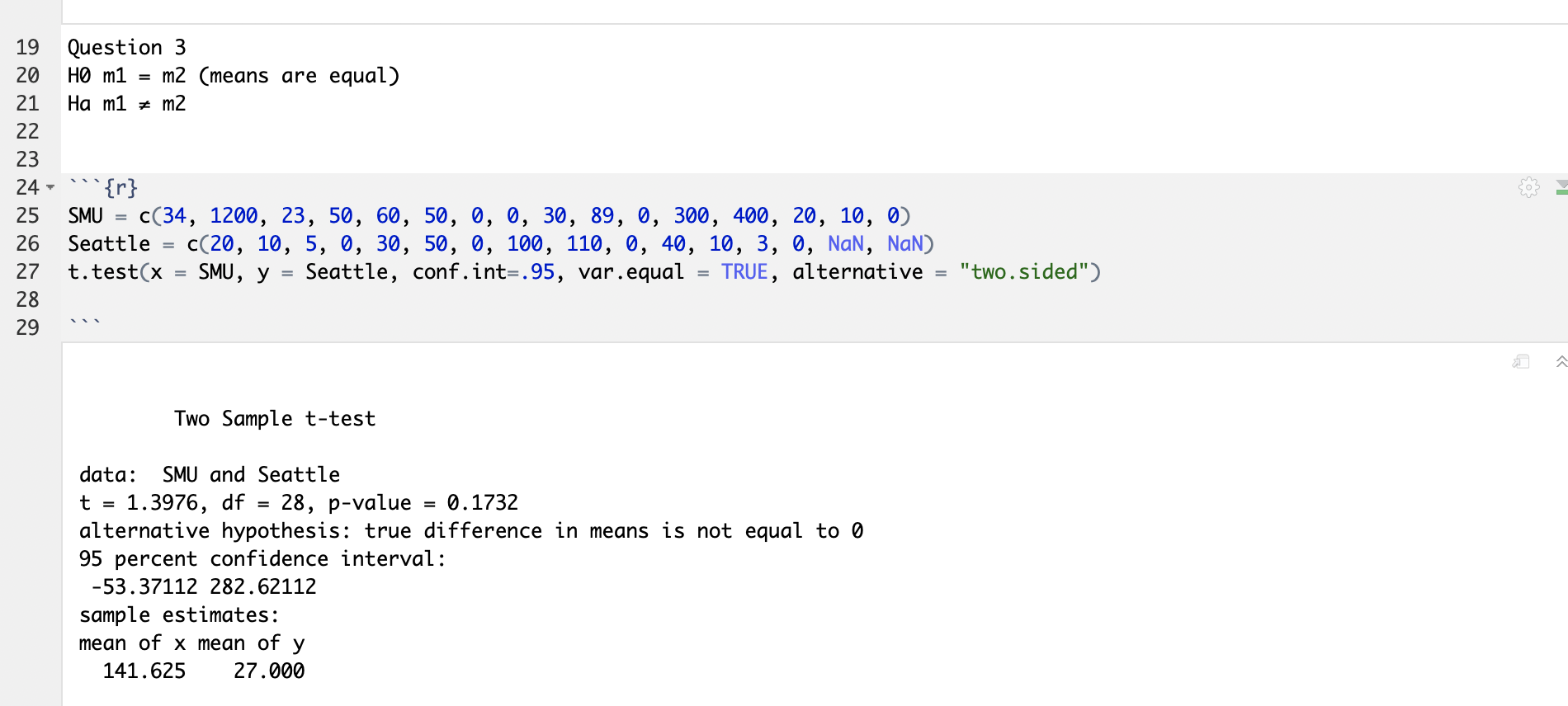
Question 3:

1. Use R for t-test

Step 1 Ho: m1=m2 (sample means equal)

Ha: m1 ≠ m2 (sample means not equal)

Step 2/3 T-Test in R



Step 4: p-value (statistical likelihood) = 0.1732

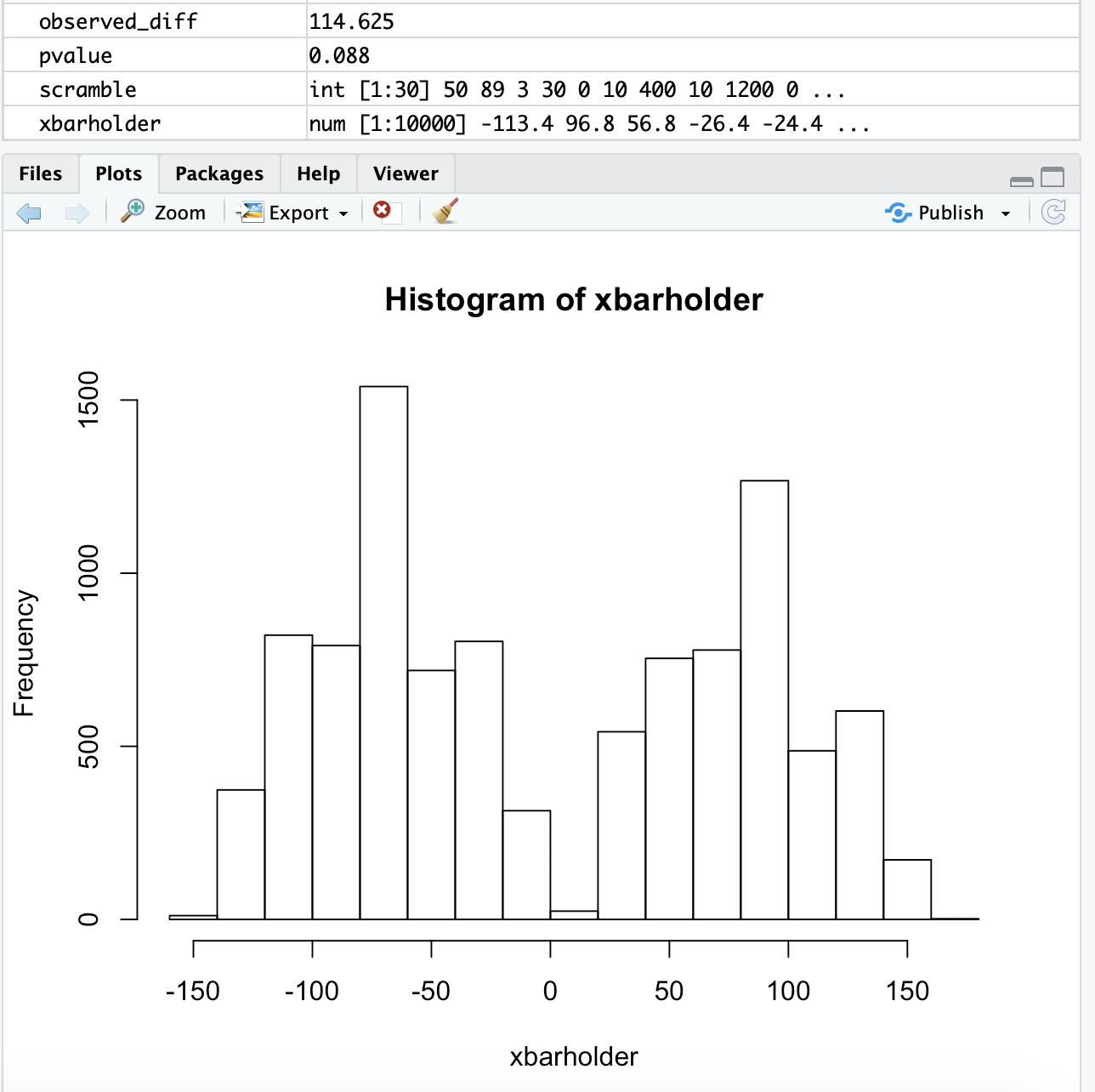
Step 5: p-value is greater than alpha, therefore we reject Ho. The means of the two populations are not equal.

Step 6:

The t-test results demonstrated that there is a significant difference between the mean cash balance of students that in SMU vs those in Seattle. This can also be seen from a histogram of the data in each sample. The range of each sample size is drastically different due to the outlier of 1200 in the SMU sample. Causal inference cannot be established in this study as it is purely observational. Collecting data on age, gender and employment status would greatly enhance the efficacy of the study.



1. Use R for permutation test

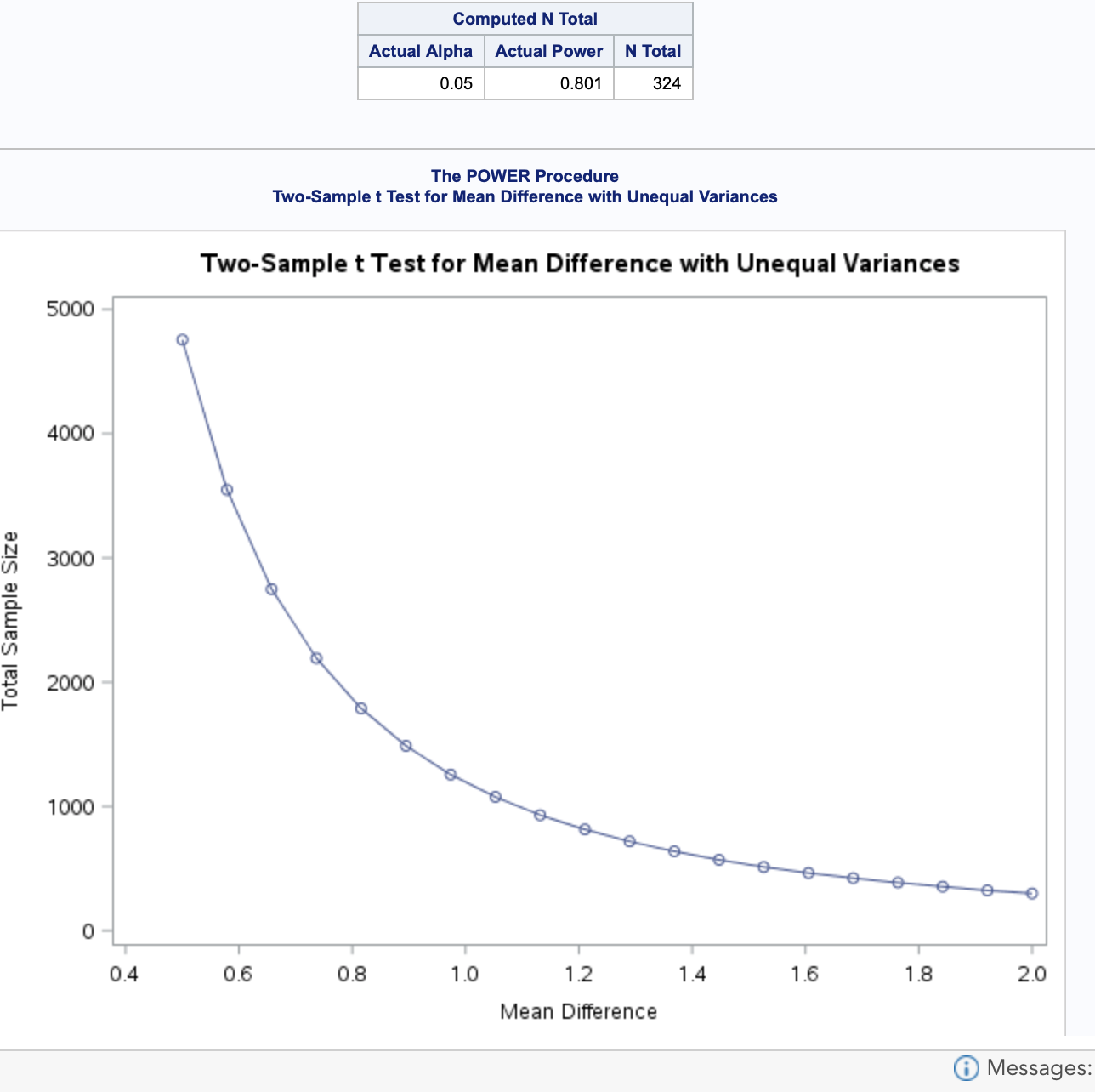


The p- value generated in the permutation test (0.088) is similar to the p-value calculated in the 95% CL two-sided t-test. Both values would lead to a rejection of the null hypothesis (the mean cash balances of the two sample sets are equal)

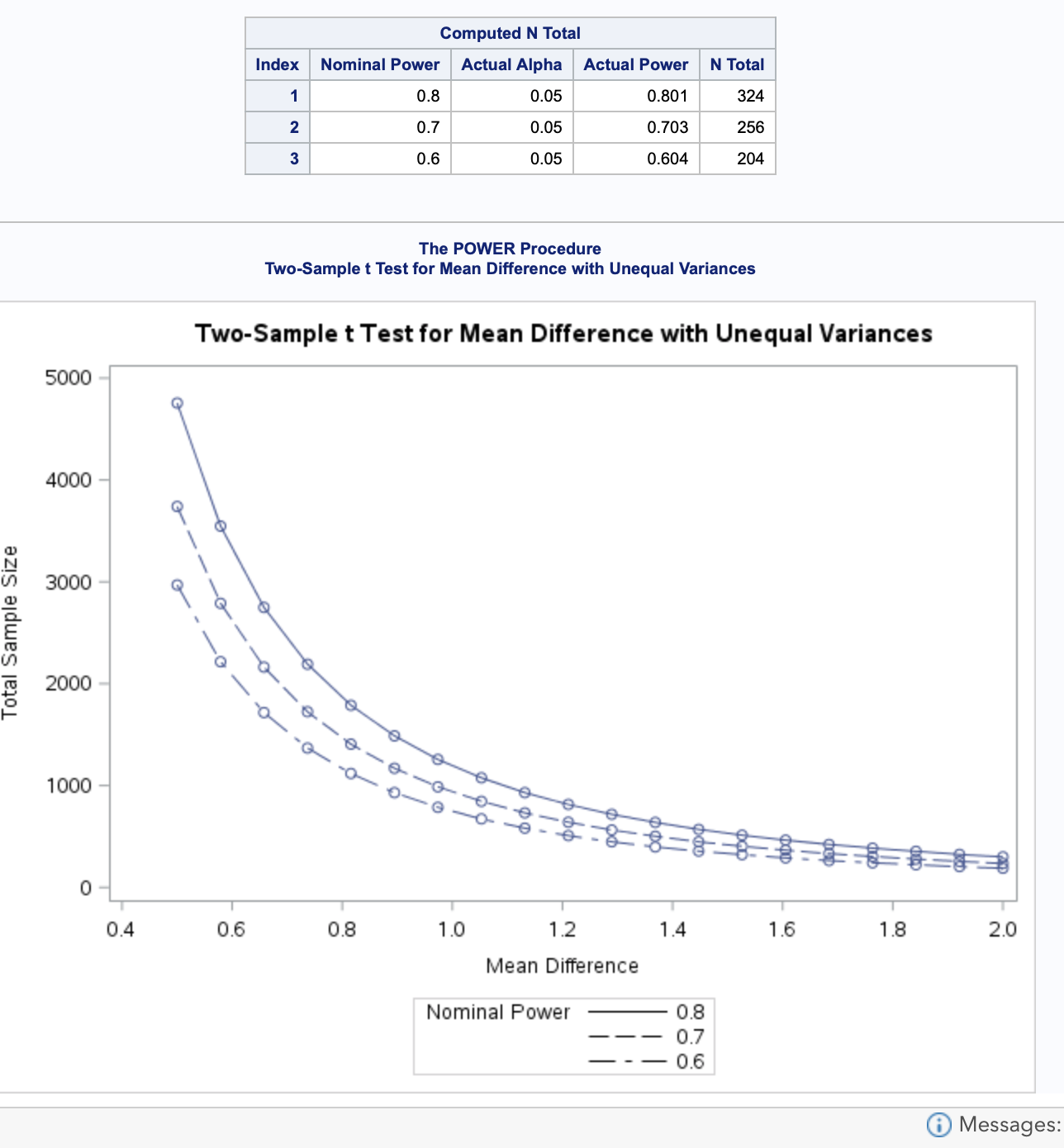
Question 4):

From question 2, the pooled SD of the population is 6.1519

\*PROBLEM 4;  
proc power;  
twosamplemeans test=diff\_satt  
stddev = 6.1519  
power=0.8  
meandiff=1.9238  
ntotal=.;  
plot x=effect min=0.5 max=2;  
run;



proc power;  
twosamplemeans test=diff\_satt  
stddev = 6.1519  
power=0.8 0.7 0.6  
meandiff=1.9238  
ntotal=.;  
plot x=effect min=0.5 max=2;  
run;



proc power;  
twosamplemeans test=diff\_satt  
stddev = 6.1519  
power=0.8 0.6  
meandiff=0.8  
ntotal=.;  
run;

Sample size diff = 698

