Project 1

A780 - Astrostatistics and Scientific Computing

Justin A. Kader and Robert E. Butler, III

1 Introduction

In this project we carried out hypothesis testing using various data files either prepared for us in advance, or using data we created ourselves. Some of the data consisted of lists of ones and zeros, and in these cases we used the z-distribution and z-statistic to evaluate various hypotheses concerning the proportion of ones and zeros in these data. We also had data consisting of lists of random numbers ranging between zero and one; in cases where we created these data ourselves they were drawn randomly from either uniform or normal distributions. For the lists of random numbers, we used the Student's t-test to evaluate hypotheses about the relationship between the means of pairs of lists.

We used `Python` and the statistical computing language `R` to conduct the hypothesis testing on the data. Results of our tests are presented in this report in the form of tables and graphical output. In section 2 we describe some of the theory of proportion tests using z-scores, and how we used `Python` and `R` to carry out these tests, as well as presenting our results. In section 3, we give a description of the theory of hypothesis testing using Student's t-test as well as backgroun information about type I & II errors an statistical power. In section 4, we describe our methods and present our results of the Student's t-test. The appendix includes additional tests and power analyses, as well as graphical output from our `Python` and `R` codes.

2 Theoretical Background of Hypothesis Testing

The most common mode of scientific investigation relies on the process of using data analysis to deduce properties of an underlying probability distribution. For the subset of cases where the investigator has a well-defined hypothesis they want to test, there exist powerful hypothesis evaluation methods falling under the heading of "statistical hypothesis tests".

In statistical hypothesis tests, the investigator typically compares two statistical data sets. A hypothesis is proposed for the relationship between the two sets, and this hypothesis is compared as an *alternative* to an idealized *null* hypothesis that proposes no relationship (**difference?**) between the two data sets.

The alternative hypothesis is deemed *statistically significant* if the null hypothesis can be rejected according to a comparison between a specified test statistic and chosen significance level. For data that are assumed to be drawn from a normal distribution, which is the usual assumption in statistical inference, and if the data samples are small, then it is common to use the **t-test** to evaluate the hypothesis.

The Student's t-test makes use of the t-distribution, which can be defined as the distribution of locations of a sample mean relative to a "true", or population mean, divided by the sample standard deviation, after multiplying by a standardizing term $\sqrt{n}$. Being a continuous probability density

distribution, any integrated area under the curve may be interpreted as a probability. The shape of the curve resembles a normal, which follows from the central limit theorem: the distribution of means computed from many realizations of the sample will be normal (c.f. Figure 1). The functional form of the t-distribution is

$$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{1}$$

Where t is the eponymous test statistic, and ν is the number of degrees of freedom, computed as $\nu = n - 1$ where n is the sample size (assumed to be equal for the two samples). When using the t-test to evaluate the difference between the means of two samples, which is the typical usage, the first task is to specify a "significance level" α, which is *defined* as the probability of the investigator rejecting the null hypothesis given that the null hypothesis were in fact true. Usually this is set at the 5% level. In other words, if the study found that their t-statistic fell beyond the significance level, then the study rejects the null hypothesis in favor of the alternative, *but accepts* a 5% probability that the null hypothesis was in actual fact *true* – this is known as a type I error.

The t-test can be one- or two-sided. If for example the null hypothesis we are testing is that the means of two samples, $\bar{X}_1$ and $\bar{X}_2$ are equal, and the alternative hypothesis is that they are different, then we have a two-sided t-test, i.e. we are willing to reject the null hypothesis if the t-statistic falls far enough away from $t = 0$ on either side. If instead we have a single sample and we are testing the hypothesis that the sample mean $\bar{X}$ is equal to some value μ, and the alternative hypothesis is that $\bar{X}$ is *larger* than μ, then this is a one-sided t-test, since we are only willing to reject the null hypothesis if the t-statistic falls far enough away from μ in the positive direction. It is the significance level t^* that determines the distance that t can fall from the center of the distribution before the null hypothesis is rejected.

In a two-sided t-test, the significance level, say 5% probability or $\alpha = 0.05$, is the sum of the area under the distribution for $t \leq -t^*$ (2.5%) and the area under the distribution for $t \geq t^*$ (2.5%), which is also shown in Figure 1 as the shaded red areas. In a one-sided t-test, the 5% significance level would belong on just one side of the distribution. Some controversy surrounds usage of the one-sided t-test, since for the same significance level, t^* would be located in fact closer to the center of the distribution, and therefore the investigator would be able to claim rejection of the null hypothesis for less deviant results while still accepting the same 5% probability of making a type I error. The t-statistics for the one-sample and two-sample t-tests are

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p\sqrt{\frac{2}{N}}} \qquad \text{(Two sample t-statistic)} \tag{2}$$

$$t = \frac{\bar{X} - \mu}{S/\sqrt{N}} \qquad \text{(One sample t-statistic)} \tag{3}$$

Where $\bar{X}_i$ are the sample means, s_p is a "pooled standard deviation" for a pair of two independent samples, μ is a hypothesized population mean, S is a sample standard deviation, and N is the sample size. An important and closely related quantity is the "p-value" of an observation: for a calculated t, the p-value is the area under the t-distribution beyond t (in the direction away from the peak of the distribution). The p-value represents the probability of the observation (i.e. computed $\bar{X}$), or a more extreme value, assuming the null hypothesis were true. If this probability is smaller than the significance level α, then the observation is statistically significant and the null hypothesis is rejected. The p-value can be computed as

$$p = \int_{-\infty}^{-t} \mathrm{PDF}_t(t)\ dt \quad \text{or} \quad p = \int_{t}^{\infty} \mathrm{PDF}_t(t)\ dt \qquad \text{(One-tailed t-test)} \tag{4}$$

$$p = \int_{-\infty}^{-t} \mathrm{PDF}_t(t)\ dt + \int_{t}^{\infty} \mathrm{PDF}_t(t)\ dt \qquad \text{(Two-tailed t-test)} \tag{5}$$

Another important concept for hypothesis testing is that of a "confidence interval". For a two-tailed t-test with significance level of 5% ($\alpha = 0.05$, $\pm t^* = \pm t_{0.025}$), the confidence level would be 95% and the confidence interval for a given sample mean is computed as

$$\bar{X} \pm \left(\frac{S}{\sqrt{N}}\right) t_{0.025}. \tag{6}$$

Were the process of computing a sample mean repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter (μ) would tend toward 95%, given a 95% confidence level. In other words, $\pm(S/\sqrt{N})$ $t_{0.025}$ is the distance (in either direction), in numbers of standard errors, from the true population parameter where the area under the t-distribution in this range corresponds to a 95% probability. A confidence interval is therefore of the same length around any computed sample mean, and this length is set simply by the confidence level chosen.

Hypothesis testing can also involve proportions of one independent variable with respect to another independent variable, such as the proportion of spiral to elliptical galaxies in a sample of galaxies (e.g. a galaxy cluster). In these cases, the hypothesis testing is operationally the same as described above, except instead of a Student's t-distribution, we use a binomial distribution and compute a z-statistic (z-score) in order to get our p-value:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \tag{7}$$

where $\hat{p}$ is the observed proportion, p_0 is the null hypothesized proportion, and n is the sample size.

2.1 Type I/II Error and Statistical Power

In statistical hypothesis testing, results are not always perfect reflections of the true case. In particular, there are two distinct error modes when one considers a null hypothesis H_0 and its alternative.

Type I error refers to the case when the null hypothesis is true, but is incorrectly rejected. In other words, the hypothesis test will have flagged a result as statistically significant when in reality it is not. This type of error can be called a "false positive." The type I error rate is simply the chosen significance level for the hypothesis test (α). If $\alpha = 0.05$ (a commonly-used significance level), the tester is asserting that a 5% probability of incorrect rejection of the null is acceptable.

Type II error refers to the opposite case from Type I error; that is, the null hypothesis is false (and should be rejected), but is *not* rejected. Here, the test fails to flag a result as statistically significant when in reality it is. This type of error can be called a "false negative"—simply the inversion of a false positive. The type 2 error rate (β) is a bit more complex than its type I counterpart. In general, we are accustomed to setting the value of α, but not explicitly setting the value of β.

The statistical **power** of a test is defined as $1-\beta$, or just one minus the probability of type II error. In words, it can be defined as the probability that the test correctly rejects the null hypothesis when a *specific* alternative hypothesis is true. To find this, one must set the alternative hypothesis at a particular value as opposed to just one-sided or two-sided.

3 Proportion Tests Using Z-Scores

We were tasked with carrying out one- and two-proportion tests on two different files, each with two columns of zeroes and ones. The first has 30 observations in each column, and the second has 400. We needed to carry out a Z-test to test the null hypotheses provided in the homework. Tables 1 and 2 show the results of these tests. I've grouped them in terms of test type as opposed to by file.

From the p-values in the tables and using $\alpha = 0.05$, we can clearly and obviously reject the null hypothesis in both test cases for file 2. We cannot reject the null in either case for file 1.

Across different methods in R and Python, we came up with slightly different p-values for each test; however, none of the differences was significant enough to change whether we reject the null hypothesis. We wrote a manual function in Python to do the tests, as well as using the suggested `statsmodels` package. In R, we used the `prop.test`. I won't report the specific differences here, but they were on the order of a few to ten/twenty percent. In addition, within the prop test in R, there is an optional Yates continuity correction, which is automatically applied unless turned off. We turned it off since we have enough samples in both files to eliminate the need for it.

It was not at all clear why we have differences between the methods, especially using the manual function, since the formula is quite simple.

Table 1: Null hypothesis is that the fraction of ones in the first column of a file is equal to 0.6; alternative is two-sided (fraction not equal to 0.6). The rows represent files 1 and 2, respectively.

p-value	conf_low	conf_hi	f1
0.1360	0.3023	0.6385	0.46667
9.633e-07	0.4314	0.5289	0.48

Table 2: Null hypothesis is that the fraction of ones in the first column of a file is equal the fraction in the second column; alternative is that they are not equal. The rows represent files 1 and 2, respectively.

p-value	conf_low	conf_hi	f1	f2
0.3006	-0.3835	0.1169	0.4667	0.6
6.447e-09	-0.2694	-0.1356	0.48	0.6825

4 Mean Tests Using Student's T-test

We were tasked with performing one- and two-sample means hypothesis testing using Student's t-test for a pair of files each containing two samples of random numbers. We also performed two-sample mean hypothesis tests for an additional 500 files, each with a pair of samples. First we discuss our method and results for the means testing for the two initial files (meansfile1 and meansfiles2).

The null hypothesis, H_0, that we are testing in each data file is that the two samples have equivalent means. We will use a two-sample t-test on the unpaired (independent and identically distributed) samples. The t-statistic used to test whether the means are different is computed as per Equation 2.

We used the function `ttest_ind()` from the Scipy.stats package in Python to generate the t-statistic and p-value for each test of the null hypothesis that the two samples in each data file had equivalent mean values. By setting a significance level of 95%, we were able to reject the null hypothesis in those cases where the p-value was less than 0.05.

For the two meanfiles, we got the following results. For the first file we tested the hypothesis that the mean of the first column was 0.8. We got a t-statistic of 1.1988, and a p-value of 0.233. With this p-value and a significance level of 0.95%, we fail to reject the hypothesis. For file 2, we find t-statistic of 4.6143, and a p-value of 5.022, so we also fail to reject the hypothesis for this sample.

Both meanfiles had two columns and we tested the hypothesis that the means of the columns were the same. For **file 1 we found a t-statistic of 0.498 and a p-value of 0.6191, so we fail to reject the hypothesis. For file 2 we got a t-statistic of 3.778 and a p-value of 0.00017, which is far less than $\alpha = 0.05$, so we *do* reject the hypothesis that the means are the same for this file.**

We used the `Python` function `scipy.stats.ttest_ind` to conduct a t-test on each of the 500 sample pairs. For each test we recorded a t-statistic, and a p-value. With our significance level of $\alpha = 0.05$, we were able to identify statistically significant results using the p-values we computed in each case. The number of statistically significant results (times we rejected the null hypothesis) is recorded in Table 3. For each data pair we also computed a confidence interval and conducted a power analysis using the `Python` function `statsmodels.tt_ind_solve_power`, where we input an effect size of 0.2, our significance level of 0.05, and ratio of the sample sizes in each case to get a power, 1 - β. For each set of samples (A through E), we plotted the mean difference ($\bar{X}_1 - \bar{X}_2$), plus or minus

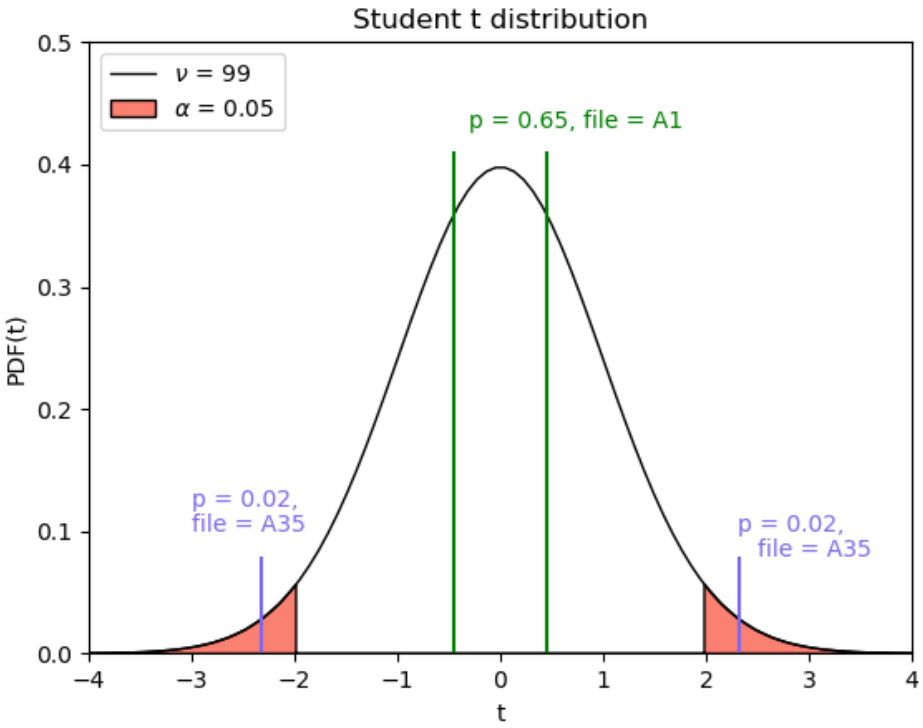
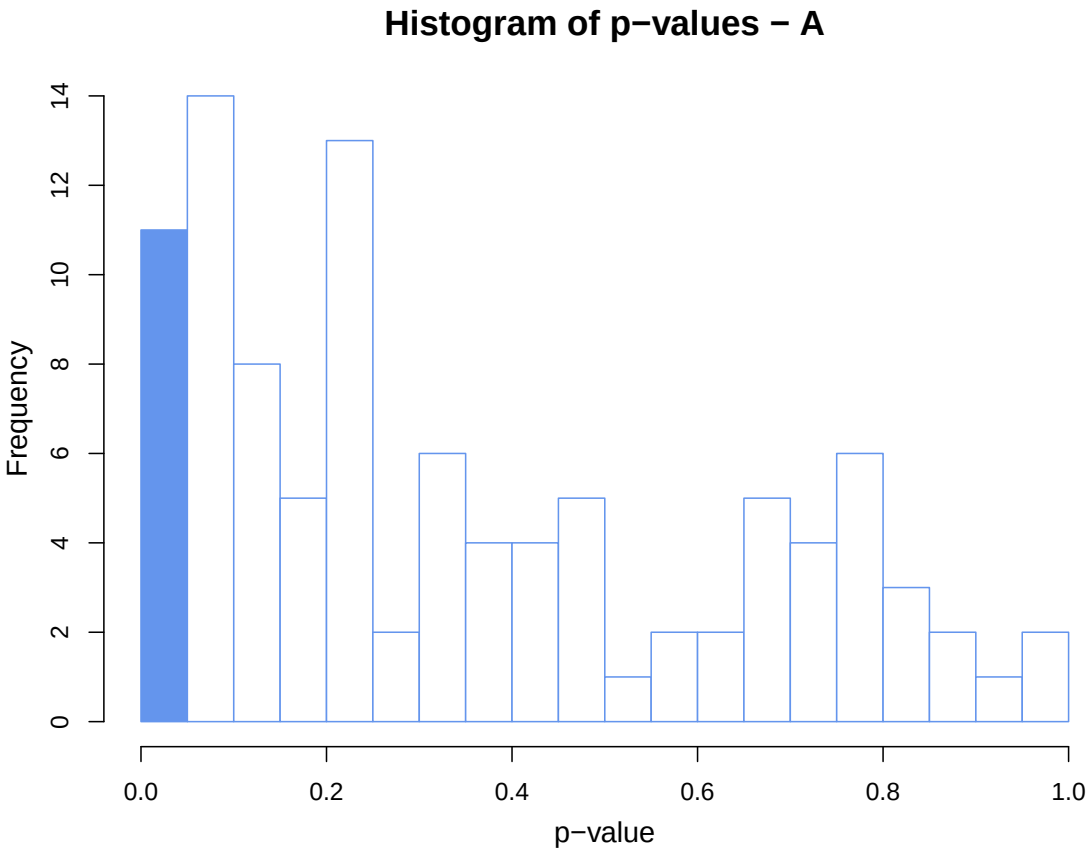


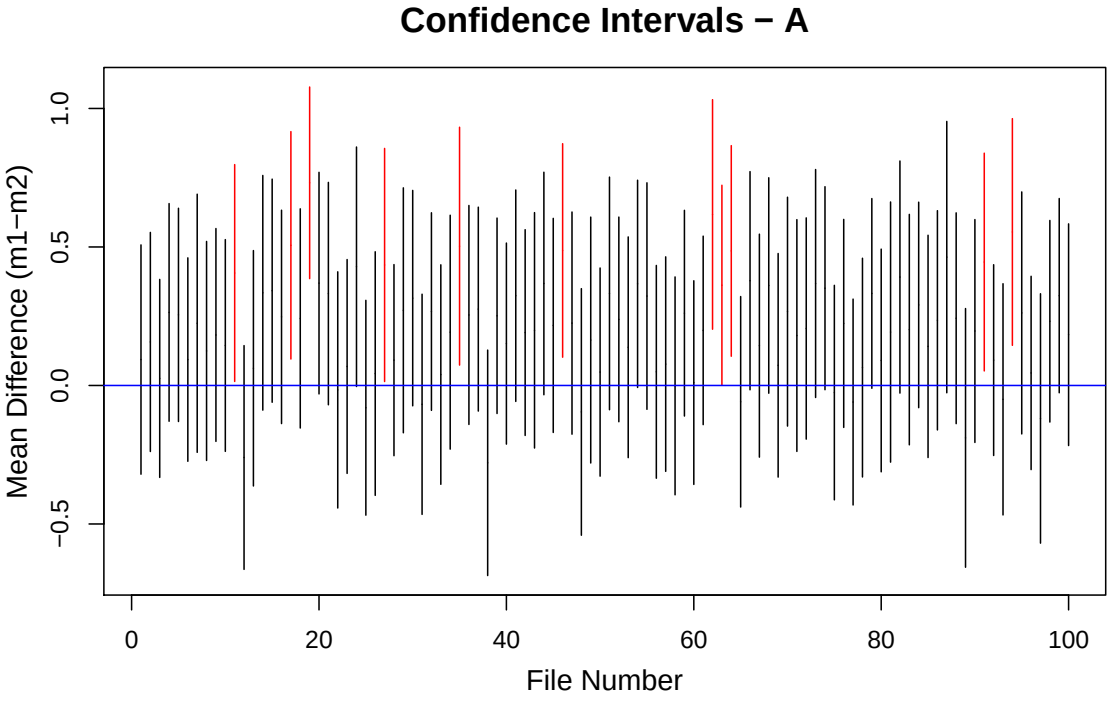
Figure 1: Student's t-distribution for $\nu = 99$. Areas outside t* are shaded in red, where t* is the t-statistical corresponding to significance level $\alpha = 95\%$. The t-statistics resulting from two hypothesis tests are marked as green lines for the case of a high p-value where we fail to reject the null hypothesis in that case, and blue vertical lines for the case where we do choose to reject the null hypothesis based on our significance level owhere we've shown the p-value computed for those two tests.

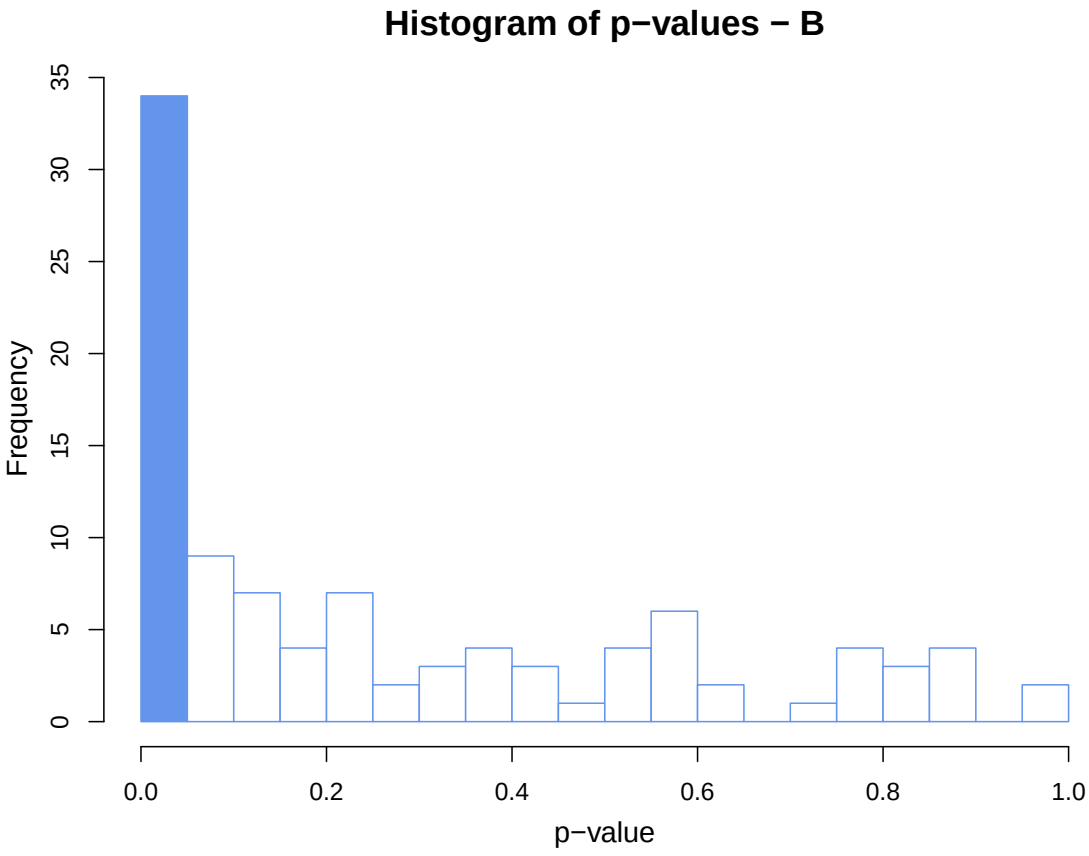
the confidence interval, versus file number. We noticed that depending on which set we plotted, there was a different proportion of confidence intervals containing a mean difference of zero, i.e., a different proportion of statistically significant results, ranging from six significant results in set D, all the way up to a proportion of 50% for set C. In addition to the confidence interval plots, we also provide histograms of p-values for each set of statistical samples. In each histogram we highlight the smallest bin (0 - 0.05) of p-values, which is the bin containing all statistically significant results. The histograms also illustrate the level to which the sample sets have results which deviate from the null hypothesis.

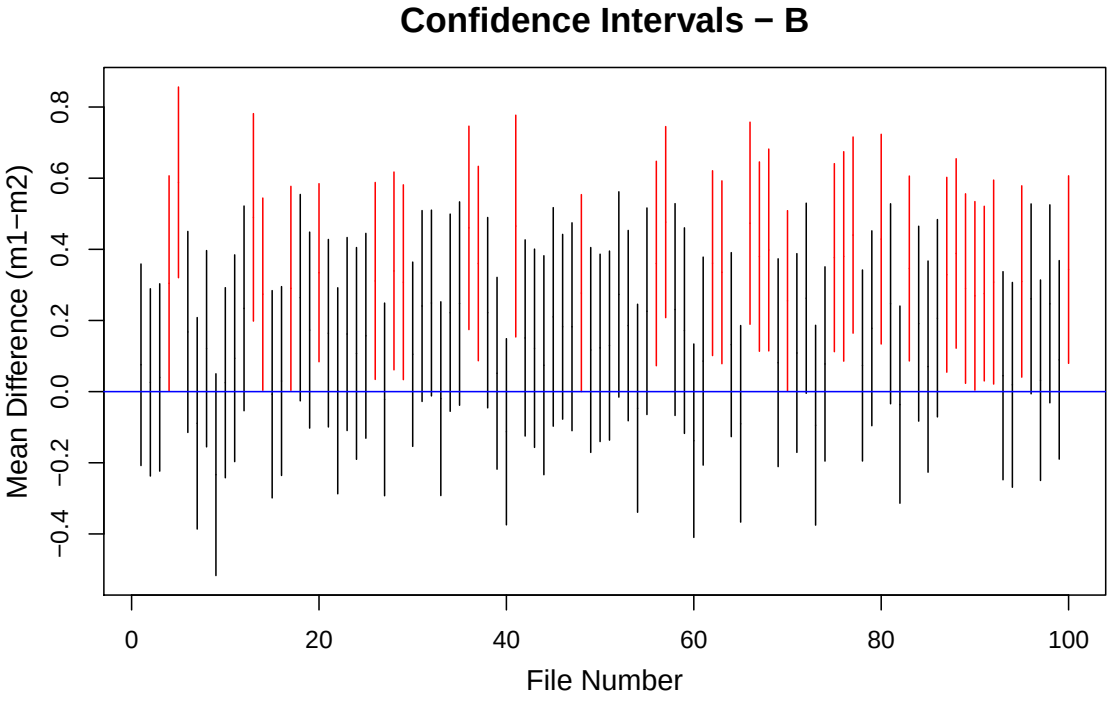
Table 3: Record of statistically significant results from t-tests of the 500 data pairs.

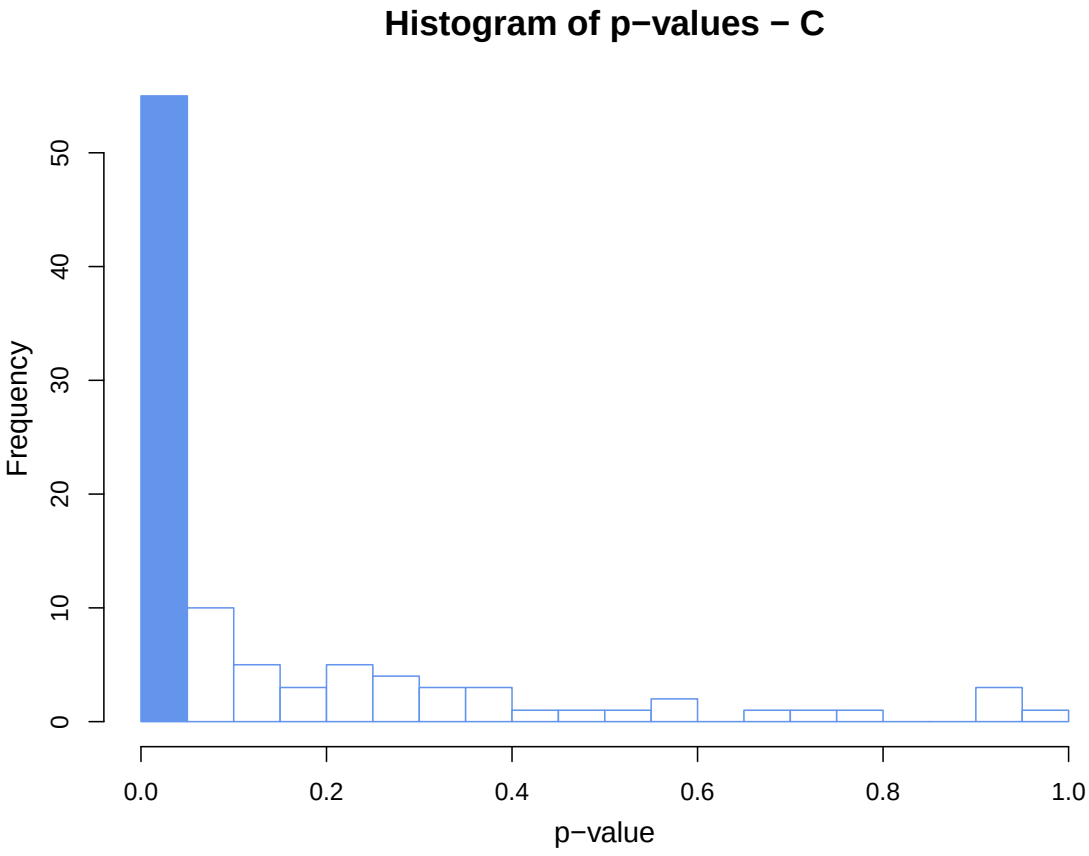
Data Set	Rejections in Python	Rejections in R
A	11	11
B	34	34
C	55	55
D	6	6
E	22	22

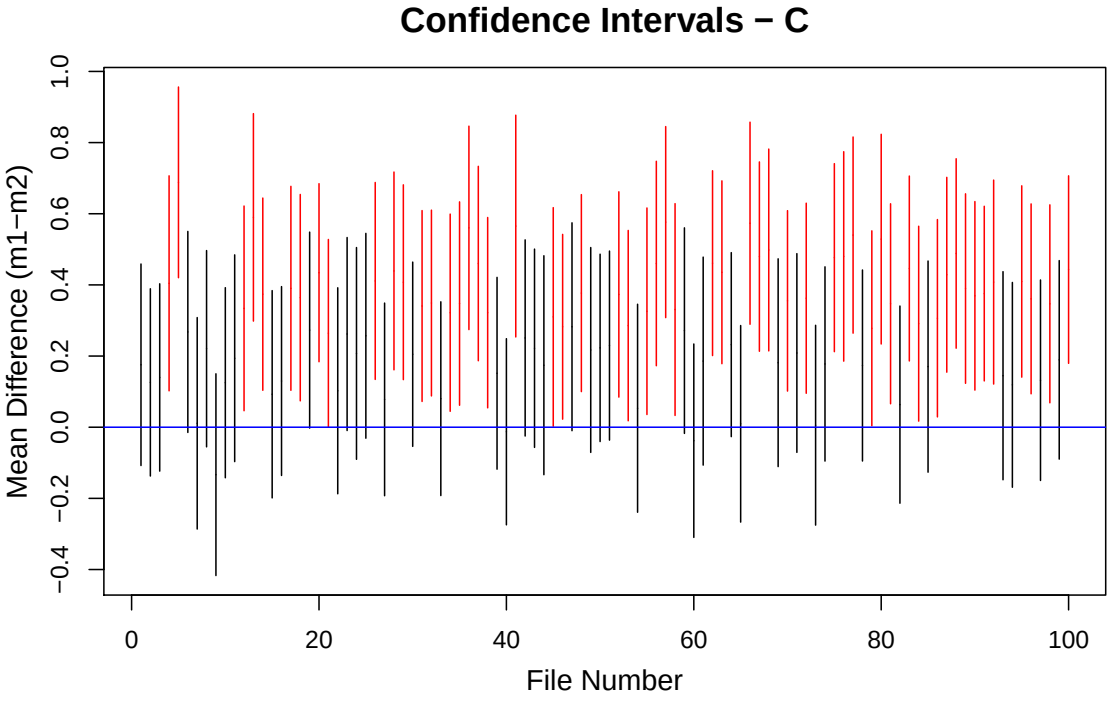

Histogram of p−values − A
Frequency
0
2
4
6
8
10
12
14
0.0
0.2
0.4
0.6
0.8
1.0
p−value

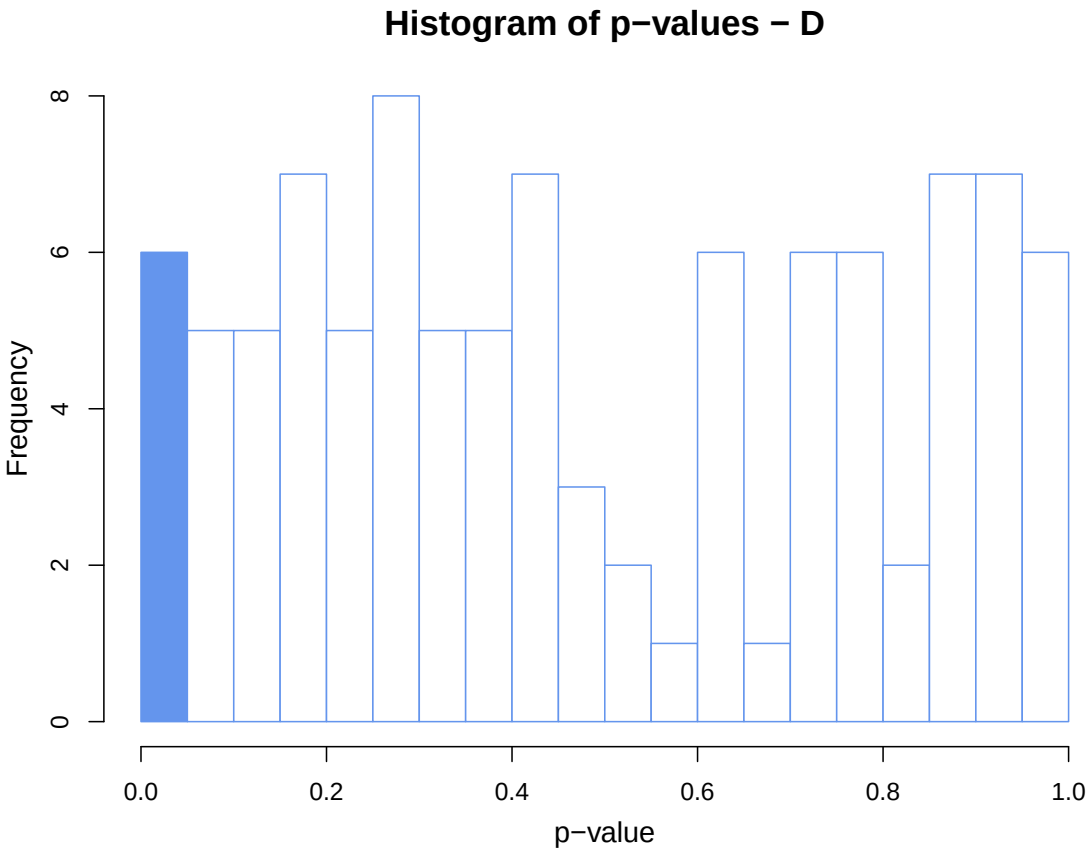

Confidence Intervals − A
Mean Difference (m1−m2)
−0.5
0.0
0.5
1.0
0
20
40
60
80
100
File Number

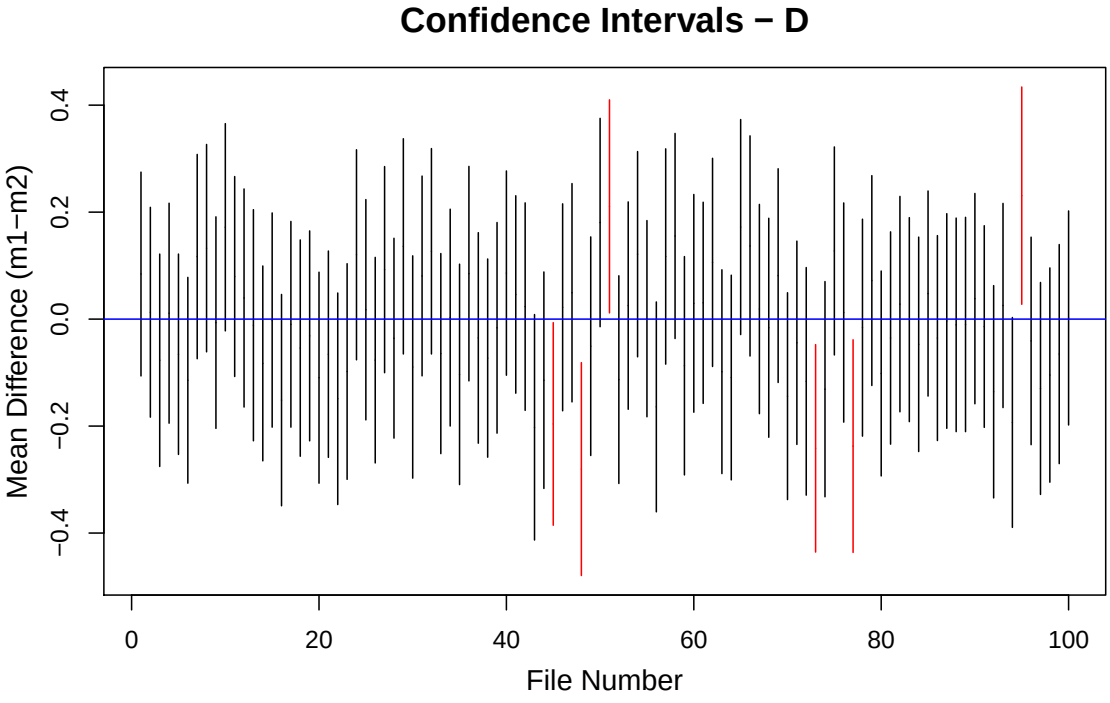

Histogram of p−values − B
Frequency
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
p−value

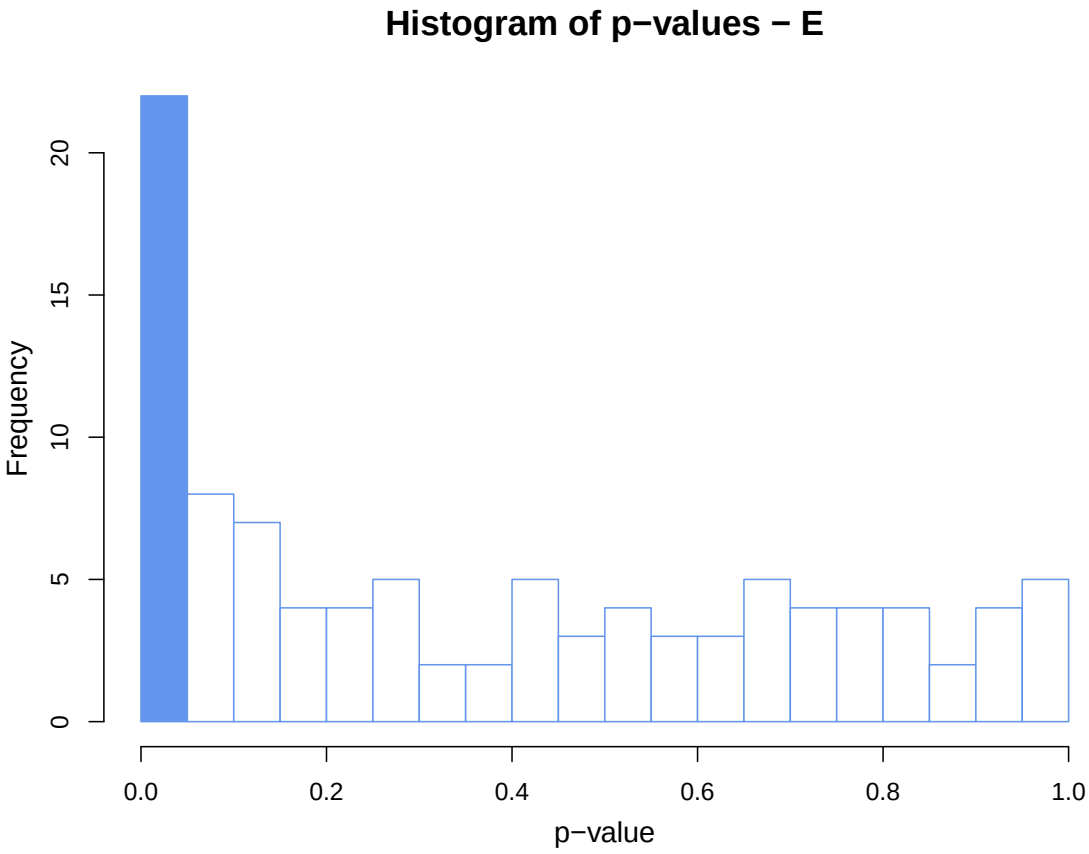

Confidence Intervals − B
Mean Difference (m1−m2)
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
File Number

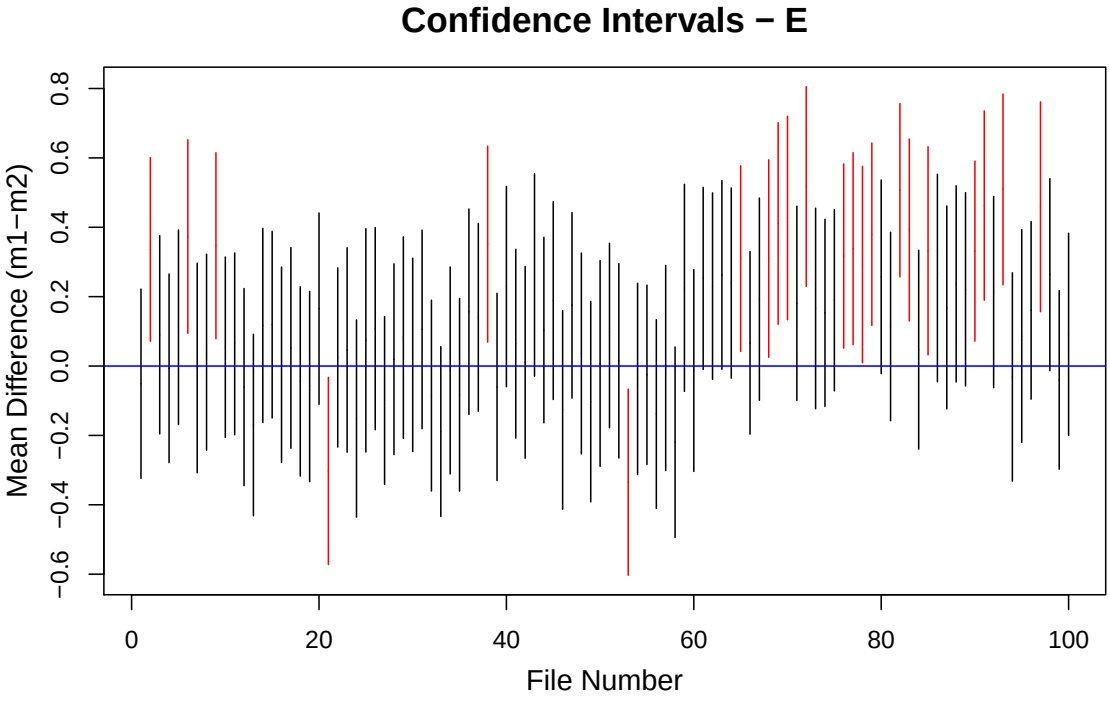

Histogram of p−values − C
Frequency
0
10
20
30
40
50
0.0
0.2
0.4
0.6
0.8
1.0
p−value


Confidence Intervals − C
Mean Difference (m1−m2)
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
File Number


Histogram of p−values − D
Frequency
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
p−value


Confidence Intervals − D
Mean Difference (m1−m2)
−0.4
−0.2
0.0
0.2
0.4
0
20
40
60
80
100
File Number


Histogram of p−values − E
Frequency
0
5
10
15
20
0.0
0.2
0.4
0.6
0.8
1.0
p−value


Confidence Intervals − E
Mean Difference (m1−m2)
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
0
20
40
60
80
100
File Number

5 Appendix

5.1 Part I

For the paired sample files found in tarfiles A and B, we conducted a power analysis in order to determine the likelihood of committing a type II error in the course of testing the null hypothesis (the means between each pair of samples were equal). In order to do so we used the Python package `statsmodels`, and in particular the function `stats.power.tt_ind_solve_power`. This function takes as input the so-called "standardized effect size", which is the difference between the two means divided by the standard deviation. We used $|\bar{X}_1 - \bar{X}_2| = 0.2$, and a standard deviation of unity. The function also takes the number of observations in sample 1 and in sample 2, (which were the same: 50 for sample pairs in tarfile A and 100 for sample pairs in tarfile B). Finally, we input to the function the significance level, in our case $\alpha = 0.05$.

The function outputs the "power", which is one minus the probability of a type II error (1 - β), i.e. the probability of *not* rejecting the null hypothesis when the hypothesis should have in fact been rejected. Thus, the power is a measure of how robust we are to type II errors. For the sample pairs in tarfile A, we find a power of $(1-\beta)_A = 0.168$ or about 17%. For the sample pairs in tarfile B, we find $(1-\beta)_B = 0.291$ or about 29%.

5.2 Part II

In the second page of instructions for the Appendix, we were tasked to generate 100 samples with n numbers from a normal distribution with $\mu_1 = 1$ and $\sigma_1 = 1$ (subsample 1) and n numbers from a normal distribution with $\mu_2 = 1.2$ and $\sigma_2 = 1$ (subsample 2). This was executed across a range of n from 10 to 1000 in steps of 10 (100 total steps).

We executed this section of the Appendix in R. For each of the 100 subsamples (with two subsamples), we run a `t.test` with H_0 that the means of the subsamples are equal, $\alpha = 0.05$, and alternative hypothesis that the means are different. We also assume equal variances (there is an input for this within `t.test`.

For each value of n, we come up with 100 `t.test` results. In R, it is relatively simple to extract the relevant values from these results—in this case, the t statistic, the p value, and the two means $m1$ and $m2$. Creating a dataframe with columns n, mean, power, and lengths achieves what we want. The mean column gives the mean $(m1 - m2)$ of all the statistically significant tests for a given n. The power column gives the power using a given n and effect size 0.2 (calculated using `pwr.t.test`). The lengths column gives the number (0 to 100) of statistically significant results for each n. This rises to 100 before too long.

The plotted results are given in Figures 2 and 3. Figure 2 shows the means of all statistically significant tests vs. the corresponding value of n. I've put a red line on the plot at mean$(m1-m2) = -0.2$, which is the actual difference in population means. We can see that the differences converge to this value as the samples get larger.

Figure 3 shows the other required plot: the means of all statistically significant tests vs. the power of the test. This part was a bit more confusing for us. We were able to calculate an effect size for each individual test using the `cohensD` function. However, it was unclear to us how to plot powers calculated using the "actual" effect sizes vs. the means of the differences because there are 100

different effect sizes for each point. In place of this, we've just plotted the powers calculated using the given effect size 0.2. You can still see the general trend that a higher-powered test gives more accurate results representing the real population difference in means.

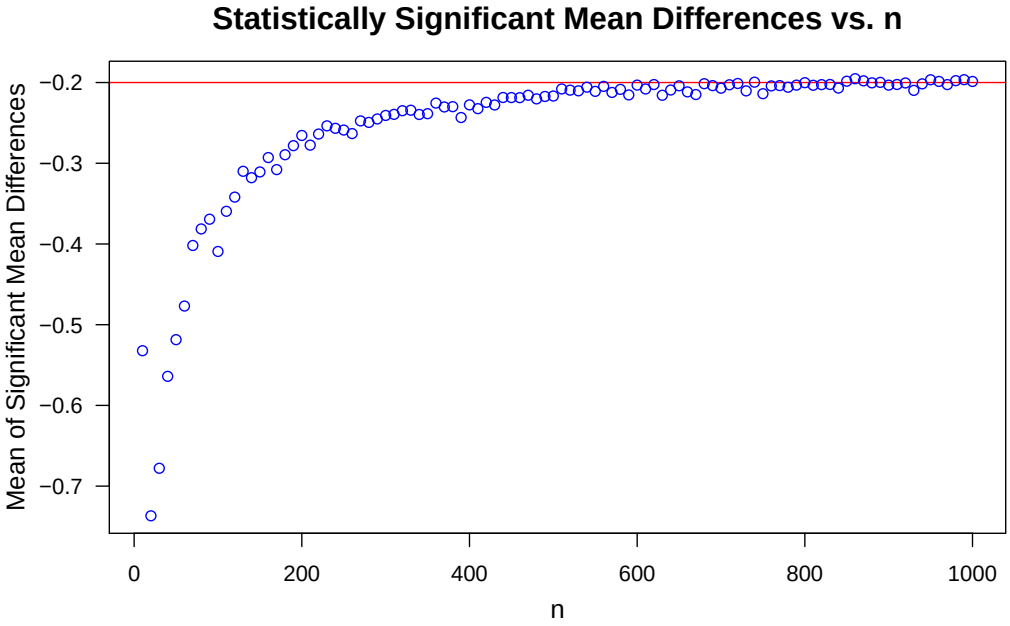


Figure 2: It's clear from this plot that as subsample size n increases, the mean of the statistically significant mean differences converges to the expected value based on the population means, -0.2.

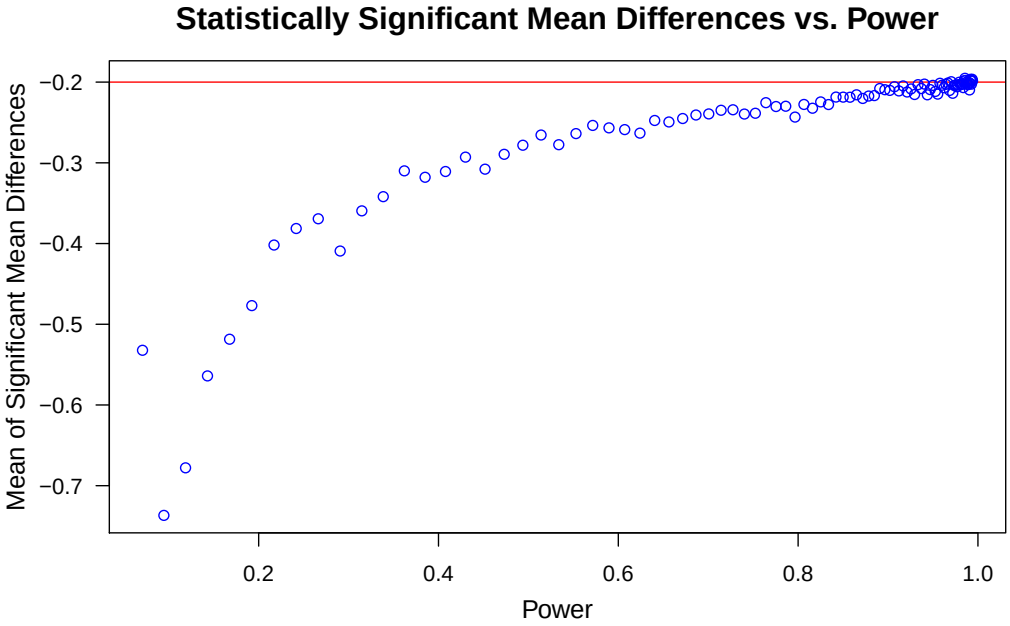


Figure 3: From this plot we can see that higher test power gives a more accurate representation of the true population mean difference.