Analysis of the Difference in Means of Systolic Blood Pressure between Smokers and Non-Smokers

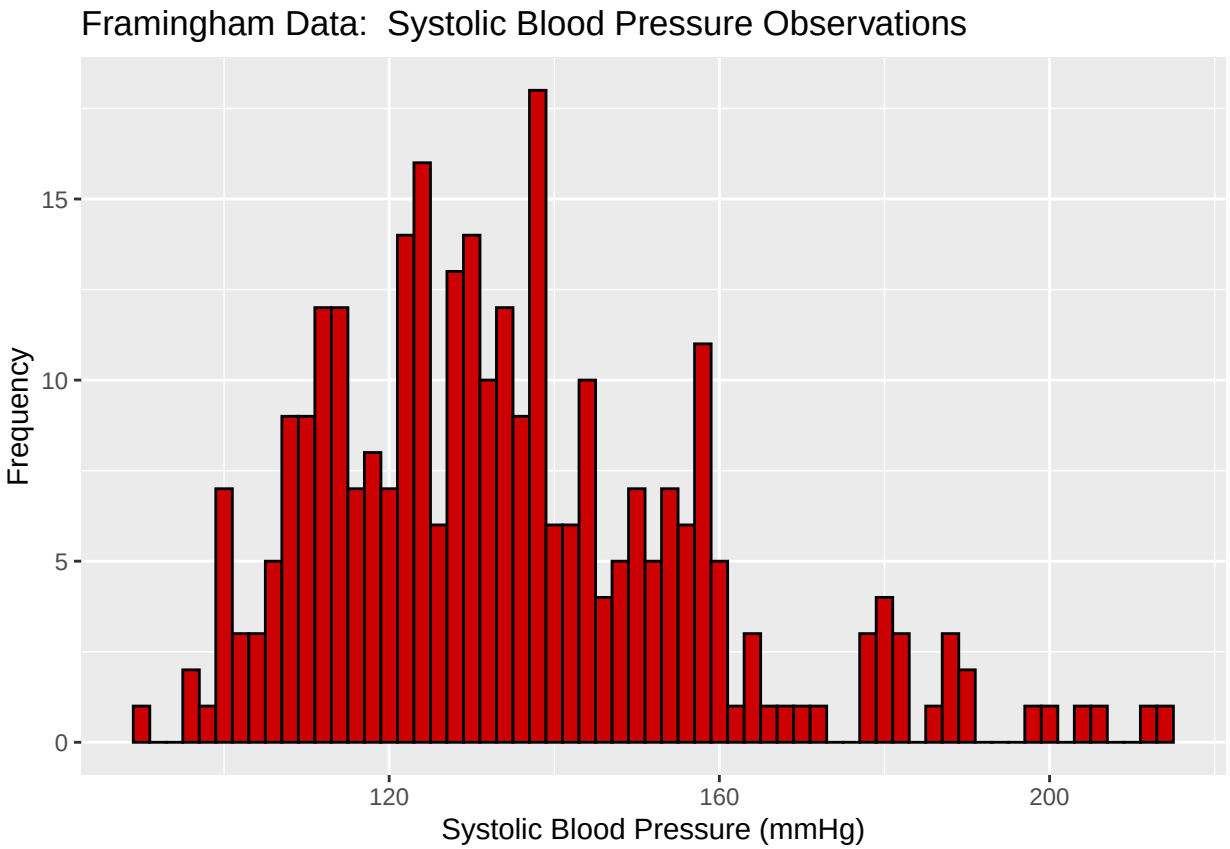
Brock Akerman and Hanan Ali

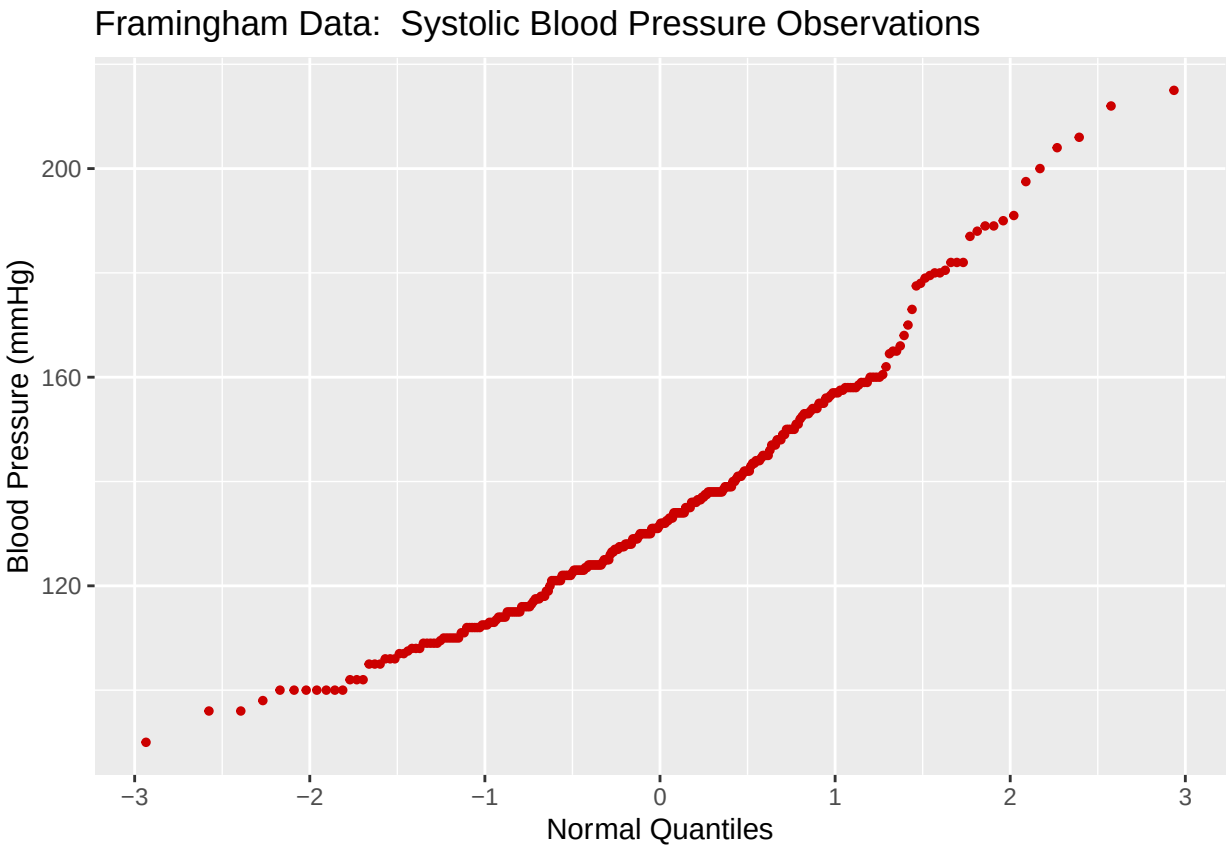
ST502 - Spring 2022

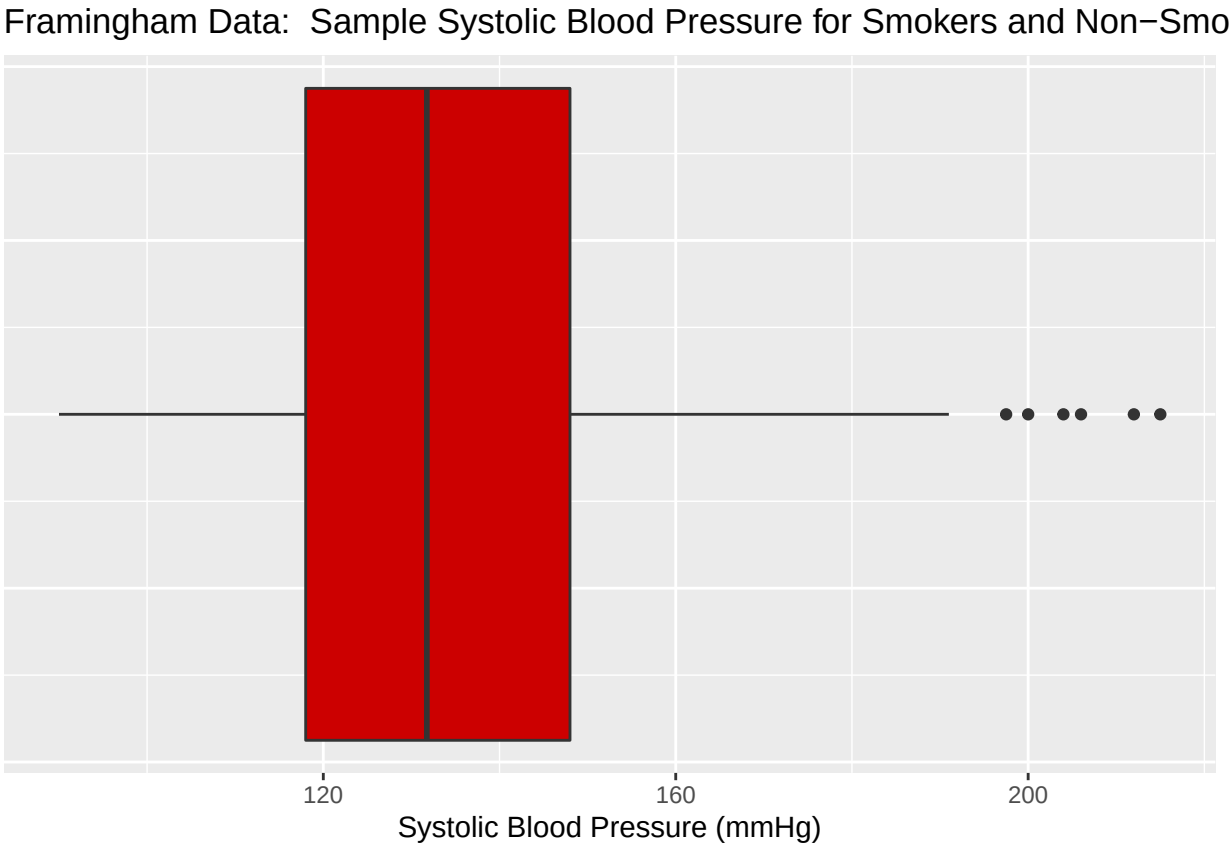
The t-Test

A sample of data is extracted from the Framingham Heart Study which includes the systolic blood pressure of participants who are identified as either smokers or non-smokers. Smokers in this study are defined as participants who have smoked cigarettes anytime during the year preceding the physical examination while non-smokers are those who have abstained from smoking during that same period of time (Magnani, 2017). The data contains a column of qualitative binary values characterizing the subjects' smoking habit and a second column pairing it with a single measurement of systolic blood pressure measured in millimeters of mercury (mmHg).

We are interested in making inference about the difference in blood pressure means between smokers and non-smokers. To test whether a difference between means exists, we will utilize the pooled t-test and the Welch-Satterthwaite t-test. Random sampling and a normal distribution are assumed for our samples; however, examining plots of the data before conducting any t-tests regardless of assumptions made is good habit. If assumptions about samples are made and there is any opportunity to strengthen the results of our test by reinforcing those assumptions, then we should do so.


Framingham Data: Systolic Blood Pressure Observations
Frequency
0
5
10
15
120
160
200
Systolic Blood Pressure (mmHg)

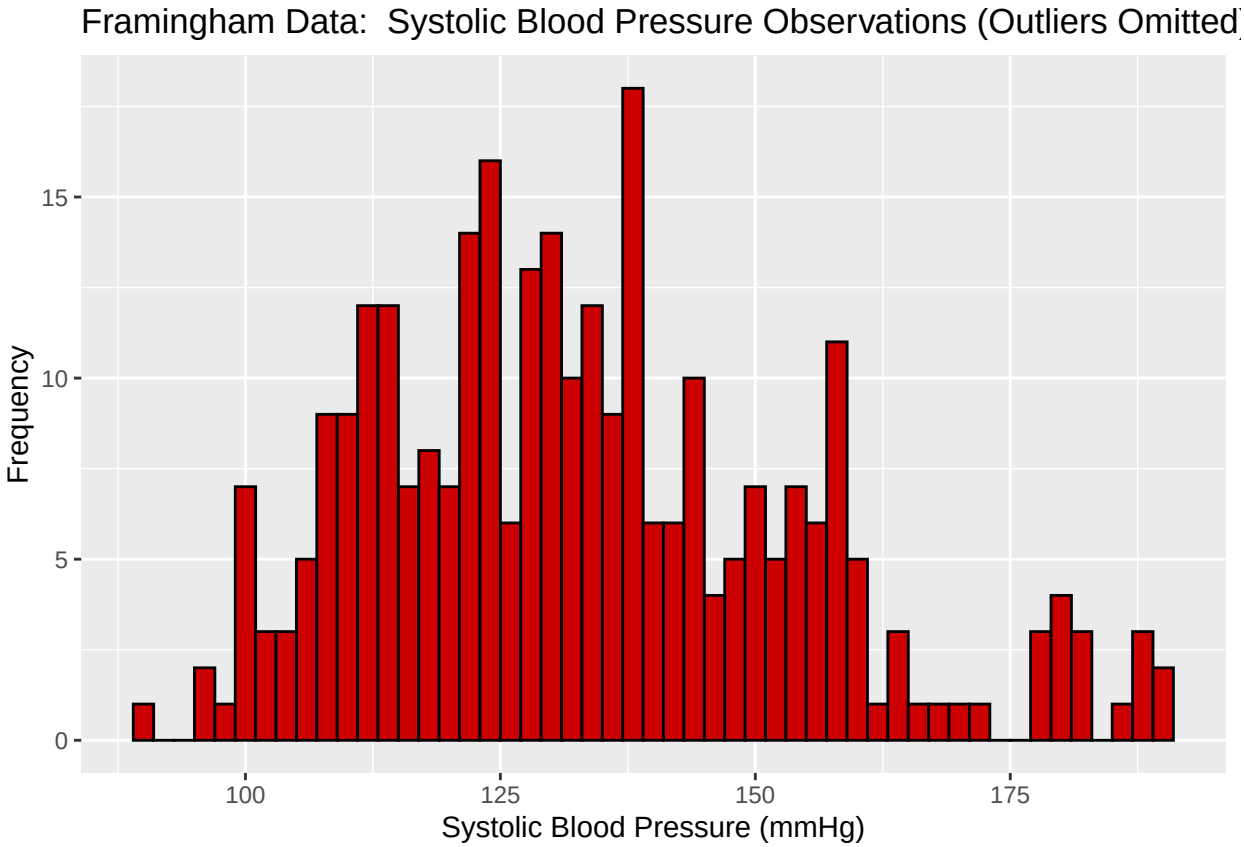

Framingham Data: Systolic Blood Pressure Observations
Blood Pressure (mmHg)
200
160
120
−3
−2
−1
0
1
2
3
Normal Quantiles

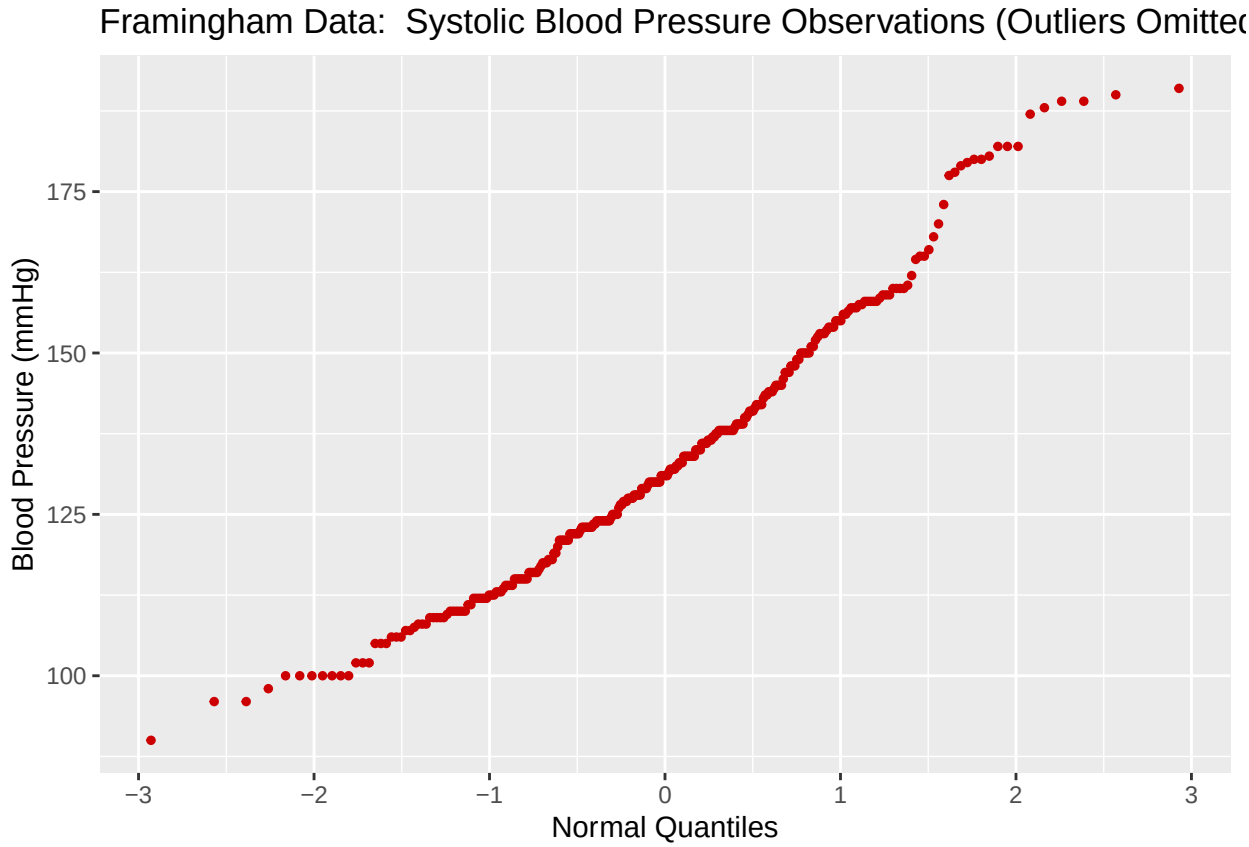


Visually summarizing the histogram above shows a skew toward the right as we can see several values inside the skew that might not be a good representation of the rest of the data points. The concerning values we think may be outliers are systolic blood pressure observations where the subject had a 180 mmHg reading or higher. A pressure observed above 180mmHg is consider a medical emergency and requires urgent care and hospitalization (Heart.org, 2017).

The Quantile-Quantile plot exhibits a linear tendency though variability in observations in the right tail may be distorting linearity. It is not enough to reject the normality assumption on this fact alone though purging outliers would serve to strengthen the argument in favor of assuming normality.

We produced a box-plot to examine the distribution and highlight outliers. We found six values that were outside of the interquartile multiplier range–five observations from the non-smokers sample and one observation from the smokers sample. We will report both t-tests with outliers and without outliers for comparison.


Framingham Data: Systolic Blood Pressure Observations (Outliers Omitted)
Frequency
0
5
10
15
100
125
150
175
Systolic Blood Pressure (mmHg)



After removing the outliers we can observe that the histogram looks less skewed and more normalized in comparison to the histogram with outliers. The skew is gentler and better representative of the data overall. Likewise, the Q-Q plot appears slightly more linear than before with the tail in the upper end showing less variability, visually. The preliminary data analysis is complete and we are ready to use our t-tests to make inference on our sampled means.

Pooled T-test (outliers included)

P-Value
$H_0 : \mu_1 - \mu_2 = 0$, There is no difference between means of smokers and non-smokers.
$H_0 : \mu_1 - \mu_2 \neq 0$, There is a difference between means of smokers and non-smokers.

$\bar{X}_1 = \frac{1}{225} \sum(x_1, x_2, ..., x_{225}) = 137.22444$
$s_1^2 = \frac{1}{225^2} \sum_{i=1}^{225}(x_i - \bar{x_1}) = 562.1447$
$n_1 = 225$

$\bar{X}_2 = \frac{1}{75} \sum(x_1, x_2, ..., x_{75}) = 128.06667$
$s_2^2 = \frac{1}{75^2} \sum_{j=1}^{75}(x_j - \bar{x_2}) = 352.2117$

$n_2 = 75$

$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(225-1)562.1447+(75-1)352.2117}{225+75-2} = 510.01370$

$S_p = \sqrt{(S_p^2)} = \sqrt{(510.01370)} = 22.58350$

$T = \frac{\bar{X_1}-\bar{X_2}}{S_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{137.22444-128.06667}{22.5835\sqrt{(\frac{1}{225}+\frac{1}{75})}} = 3.0413$

Under the null hypothesis, $H_0 : T \sim t_{298}$. The rejection region for the test $H_a : \mu_1 - \mu_2 \neq D_0 :$

Rejection Region: $t_{obs} : |t_{obs}| > t_{\frac{\alpha}{2},df}$
Rejection Region: $t_{obs} : |3.04131| > 1.96796$

P-Value (two-tailed):
$2 * (1 - pt(T = 3.0413, df = n_1 + n_2 - 2)) = 2 * (1 - 0.9987) = 0.00258$

Conclusion:

Reject H_0. At the 0.05 significance level, there is sufficient evidence to support the claim that mean systolic blood pressure between smokers and non-smokers is different.

Confidence Interval
Lower_Bound
$(\bar{X_1}-\bar{X_2})-t_{(\frac{\alpha}{2},df)}(S_p)(\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}) = (137.22444-128.06667)-(1.96796)(22.58350)(\sqrt{(\frac{1}{225}+\frac{1}{75}}) = 3.23194$

Upper_Bound
$(\bar{X_1}-\bar{X_2})+t_{(\frac{\alpha}{2},df)}(S_p)(\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}) = (137.22444-128.06667)+(1.96796)(22.58350)(\sqrt{(\frac{1}{225}+\frac{1}{75}}) = 15.0835$

Conclusion:

We are 95% confident that the mean difference in systolic blood pressure between smokers and non-smokers is between 3.23194 mmHg and 15.0835 mmHg.

Pooled T-test (outliers omitted)

P-Value

$H_0 : \mu_1 - \mu_2 = 0$, There is no difference between means of smokers and non-smokers.
$H_0 : \mu_1 - \mu_2 \neq 0$, There is a difference between means of smokers and non-smokers.

$\bar{X_1} = \frac{1}{220}\sum(x_1, x_2, ..., x_{220}) = 135.6409$

$s_1^2 = \frac{1}{220^2}\sum_{i=1}^{220}(x_i - \bar{x_1}) = 460.7586$
$n_1 = 220$

$\bar{X}_2 = \frac{1}{74}\sum(x_1, x_2, ..., x_{74}) = 127.0946$
$s_2^2 = \frac{1}{74^2}\sum_{j=1}^{74}(x_j - \bar{x_2}) = 285.1964$
$n_2 = 74$

$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(22401)460.7586+(74-1)285.1964}{220+74-2} = 416.868$

$S_p = \sqrt{(S_p^2)} = \sqrt{(416.868)} = 20.41735$

$T = \frac{\bar{X_1}-\bar{X_2}}{S_p\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}} = \frac{135.6409-127.0946}{20.41735\sqrt{(\frac{1}{220}+\frac{1}{74})}} = 3.1148$

Under the null hypothesis, $H_0 : T \sim t_{292}$. The rejection region for the test $H_a : \mu_1 - \mu_2 \neq D_0 :$

Rejection Region: $t_{obs} : |t_{obs}| > t_{\frac{\alpha}{2},df}$
Rejection Region: $t_{obs} : |3.1148| > 1.968121$

P-Value (two-tailed):
$2 * (1 - pt(T = 3.1148, df = 220 + 74 - 2)) = 2 * (1 - 0.99898) = 0.00204$

Conclusion:

Reject H_0. At the 0.05 significance level, there is sufficient evidence to support the claim that mean systolic blood pressure between smokers and non-smokers is different.

Confidence Interval
Lower_Bound
$(\bar{X_1} - \bar{X_2}) - t_{(\frac{\alpha}{2},df)}(S_p)(\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}) = (135.6409 - 127.0946) - (1.96796)(20.41735)(\sqrt{(\frac{1}{220}+\frac{1}{74}}) = 3.14669$

Upper_Bound
$(\bar{X_1} - \bar{X_2}) + t_{(\frac{\alpha}{2},df)}(S_p)(\sqrt{(\frac{1}{n_1}+\frac{1}{n_2})}) = (135.6409 - 127.0946) + (1.96796)(20.41735)(\sqrt{(\frac{1}{220}+\frac{1}{74}}) = 13.9459$

Conclusion:

We are 95% confident that the mean difference in systolic blood pressure between smokers and non-smokers is between 3.14669 mmHg and 13.9459 mmHg.

Welch-Satterthwaite T-test (outliers included)

P-Value For the Welch-Satterthwaite t-test we will test using the same hypothesis as in the t-test pooled; however, with the assumption about the variance removed, the test statistic formula.

$H_0 : \mu_1 - \mu_2 = 0$, There is no difference between means of smokers and non-smokers.
$H_0 : \mu_1 - \mu_2 \neq 0$, There is a difference between means of smokers and non-smokers.

$\bar{X_1} = \frac{1}{225}\sum(x_1, x_2, ..., x_{225}) = 137.22444$
$s_1^2 = \frac{1}{225^2}\sum_{i=1}^{225}(x_i - \bar{x_1}) = 562.1447$
$n_1 = 225$

$\bar{X_2} = \frac{1}{75}\sum(x_1, x_2, ..., x_{75}) = 128.06667$
$s_2^2 = \frac{1}{75^2}\sum_{j=1}^{75}(x_j - \bar{x_2}) = 352.2117$
$n_2 = 75$

$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})}{n_2 - 1}} = \frac{(\frac{562.1447}{225} + \frac{352.2117}{75})^2}{\frac{(\frac{562.1447}{225})^2}{225-1} + \frac{(\frac{352.2117}{75})^2}{75-1}} = 158.8316$
$T = \frac{\bar{X_1} - \bar{X_2} - D_0}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = \frac{137.22444 - 128.06667}{\sqrt{(\frac{562.1447}{225} + \frac{352.2117}{75})}} = 3.4142$
Under the null hypothesis, $H_0 : T \sim t_{v, \frac{\alpha}{2}} \sim t_{158.8316, 0.025}$. The rejection region for the test $H_a : \mu_1 - \mu_2 \neq D_0$:

Rejection Region: $t_{obs} : |t_{obs}| > t_{0.025,158}$

Rejection Region: $t_{obs} : |3.4142| > 1.9751$

P-Value (two-tailed):
$2 * (1 - pt(T = 3.4142, df = 225 + 75 - 2)) = 2 * (1 - 0.9996358) = 0.0007284$

Conclusion:

Reject H_0. At the 0.05 significance level, there is sufficient evidence to support the claim that mean systolic blood pressure between smokers and non-smokers is different.

Confidence Interval
Lower_Bound
$(\bar{X_1} - \bar{X_2}) - t_{(\frac{\alpha}{2}, v)}(\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}) = (137.22444 - 128.06667) - (1.975092)(\sqrt{(\frac{562.1447}{225} + \frac{352.2117}{75}}) = 3.86004$

Upper_Bound
$(\bar{X_1} - \bar{X_2}) + t_{(\frac{\alpha}{2}, v)}(\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}) = (137.22444 - 128.06667) + (1.975092)(\sqrt{(\frac{562.1447}{225} + \frac{352.2117}{75}}) = 14.4555$

Conclusion:

We are 95% confident that the mean difference in systolic blood pressure between smokers and non-smokers is between 3.23194 mmHg and 15.0835 mmHg.

Welch-Satterthwaite T-test (outliers omitted)

P-Value

$H_0 : \mu_1 - \mu_2 = 0$, There is no difference between means of smokers and non-smokers.
$H_0 : \mu_1 - \mu_2 \neq 0$, There is a difference between means of smokers and non-smokers.

$\bar{X}_1 = \frac{1}{220} \sum(x_1, x_2, ..., x_{220}) = 135.6409$
$s_1^2 = \frac{1}{220^2} \sum_{i=1}^{220}(x_i - \bar{x_1}) = 460.7586$
$n_1 = 220$

$\bar{X}_2 = \frac{1}{74} \sum(x_1, x_2, ..., x_{74}) = 127.0946$
$s_2^2 = \frac{1}{74^2} \sum_{j=1}^{74}(x_j - \bar{x_2}) = 285.1964$
$n_2 = 74$

$v = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})}{n_2 - 1}} = \frac{(\frac{460.7586}{220} + \frac{285.1964}{74})^2}{\frac{(\frac{460.7586}{220})^2}{220-1} + \frac{(\frac{285.1964}{74})^2}{74-1}} = 158.314$
$T = \frac{\bar{X}_1 - \bar{X}_2 - D_0}{\sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}} = \frac{135.6409 - 127.0946}{\sqrt{(\frac{460.7586}{220} + \frac{285.1964}{74})}} = 3.50412$
Under the null hypothesis, $H_0 : T \sim t_{v, \frac{\alpha}{2}} \sim t_{158.8316, 0.025}$. The rejection region for the test $H_a : \mu_1 - \mu_2 \neq D_0 :$

Rejection Region: $t_{obs} : |t_{obs}| > t_{0.025,158}$

Rejection Region: $t_{obs} : |3.50412| > 1.9751$

P-Value (two-tailed):
$2 * (1 - pt(T = 3.50412, df = 220 + 74 - 2)) = 2 * (1 - 0.999735) = 0.00053$

Conclusion:

Reject H_0. At the 0.05 significance level, there is sufficient evidence to support the claim that mean systolic blood pressure between smokers and non-smokers is different.

Confidence Interval
Lower_Bound

$$(\bar{X}_1 - \bar{X}_2) - t_{(\frac{\alpha}{2},v)}(\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}) = (135.6409 - 127.0946) - (1.975092)(\sqrt{(\frac{460.7586}{220} + \frac{285.1964}{74}}) = 3.7292$$

Upper_Bound
$$(\bar{X}_1 - \bar{X}_2) + t_{(\frac{\alpha}{2},v)}(\sqrt{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})}) = (135.6409 - 127.0946) + (1.975092)(\sqrt{(\frac{460.7586}{220} + \frac{285.1964}{74}}) = 13.3634$$

Conclusion:

We are 95% confident that the mean difference in systolic blood pressure between smokers and non-smokers is between 3.23194 mmHg and 15.0835 mmHg.

Discussion Are the mean systolic blood pressures different between smokers and non-smokers? All four t-test concluded with the same result; a rejection of the null hypothesis. We found evidence in all four tests of evidence supporting a difference between systolic blood pressure means of participants who smoke versus those who do not.

An important step at the beginning of the analysis of our data was the identification of outliers. We were able to tighten-up the distribution by removing outliers through a box-plot. Once we saw several observations that extended beyond the whiskers of the plot, we calculated the interquartile range and calculated the minimum and maximum values of the whiskers. Any value exceeding the maximum were found and removed. We now have reorganized our data so that it became more representative of the population of Framingham, Massachusetts. The histogram return less skewed while the QQ plot returned more linear.

From the summary below we observe that the p-values for the Pooled t-tests would result in rejecting the null hypothesis but were closer to the significance alpha of 0.05 than those from the Satterthwaite t-tests. Those tests with outliers were closer to the significance alpha of 0.05 than those p-values from the same tests without the outliers. What we are finding is that if we were to take many samples, on average we would find that samples would result in rejecting the null hypothesis more frequently in Satterthwaite t-tests and pooled t-tests. We would would also reject more often with many samples where outliers were omitted than if they were to remain.

Test Type	P-Value
t-Test pooled with outliers	0.00258
t-Test pooled without outliers	0.00204
t-Test Satterthwaite with outliers	0.00073
t-Test Satterthwaite without outliers	0.00053

Confidence levels listed below contain the range of the difference between each of the means. The t-Test pooled with outliers result had the widest confidence interval range at 11.85156 mmHg. Our smallest range was the Satterthwaite t-Test without outliers with a range of 9.6342 mmHg. The pooled t-test without outliers and the Satterthwaite t-test with outliers had a very similar range with a difference of approximately 0.11 mmHg.

Test Type	Lower Bound	Upper Bound
t-Test pooled with outliers	3.23194	15.0835
t-Test pooled without outliers	3.14669	13.9459
t-Test Satterthwaite with outliers	3.86004	14.4555
t-Test Satterthwaite without outliers	3.7292	13.3634

With consideration to all the data and summary statistics, I would recommend using the Satterthwaite t-Test with outliers removed. When presenting results, particularly when they have implications on health consultation and treatment or government policy making, it would be safer to err on the side of being more conservative with our data. The Satterthwaite follows this principle by not making assumptions about the equality of variance. By using non-pooled testing results, we were able to work with our data in such a way that it produced output more raw and natural than synthetically assuming a characteristic of sample data that may be true but cannot be verified. Likewise, the removal of outliers creates more representative collection of observations about the population. The data initially contained observations of blood pressures above 180 mmHg which would have required urgent emergency medical care and likely hospitalization. I would argue that elevated blood pressure that high is not a normal occurrence in any population and thus we would not expect that type of data to be suited for a normal distribution of blood pressures in our Framingham population. We trimmed those data points off which produced a distribution that was better suited for manipulation.

I would have chosen the Satterthwaite t-Test with outliers as my next option. The assumption of equal variance has a enormous impact on the testing results and it manifested in our work as higher p-values and wider confidence interval ranges.

The confidence interval of the difference in means is significant because of the medical implications. According to the Mayo Clinic, blood pressures can be compartmentalized into several categories of increasing severity; Normal ($<$ 120 mmHg), Elevated (120-129 mmHg), Stage One (130-139 mmHg), Stage Two ($>$ 140 mmHg), and Hypertensive Crisis ($>$ 180 mmHg) (Mayo Clinic, 2018). Both of our observed means sit on the upper limit of their respective range. A confidence interval disparity could bring the non-smoker mean from the stage two tier into hypertensive crisis or the smokers from stage one to stage two hypertension.

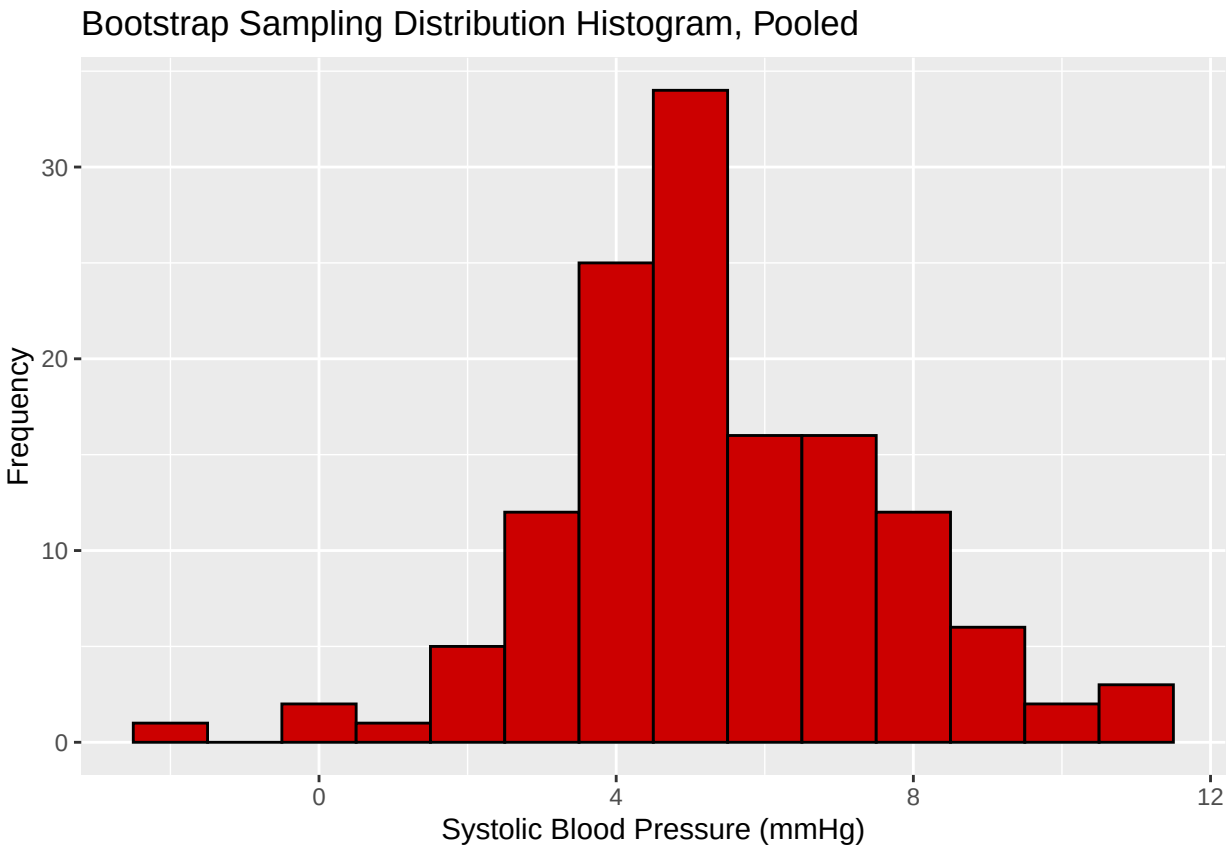
The sample mean for smokers falls within the elevated risk category while the sample mean for non-smokers is on the high-end of the stage one hypertension category. We are certain a difference in the mean systolic blood pressure exists. We can interpret these results as suggesting smoking cigarette has a suppression affect on blood pressure; however, blood pressure suppression through smoking may not be the best means of lowering blood pressure (Centers for Disease Control and Prevention, 2017). There are other means of controlling blood pressure without harming the body through a smoking habit. Diet, exercise, and blood pressure medications such as beta-blockers, alpha blockers and central or receptor agonists can be prescribed to lower pressure without the risks associated with smoking.

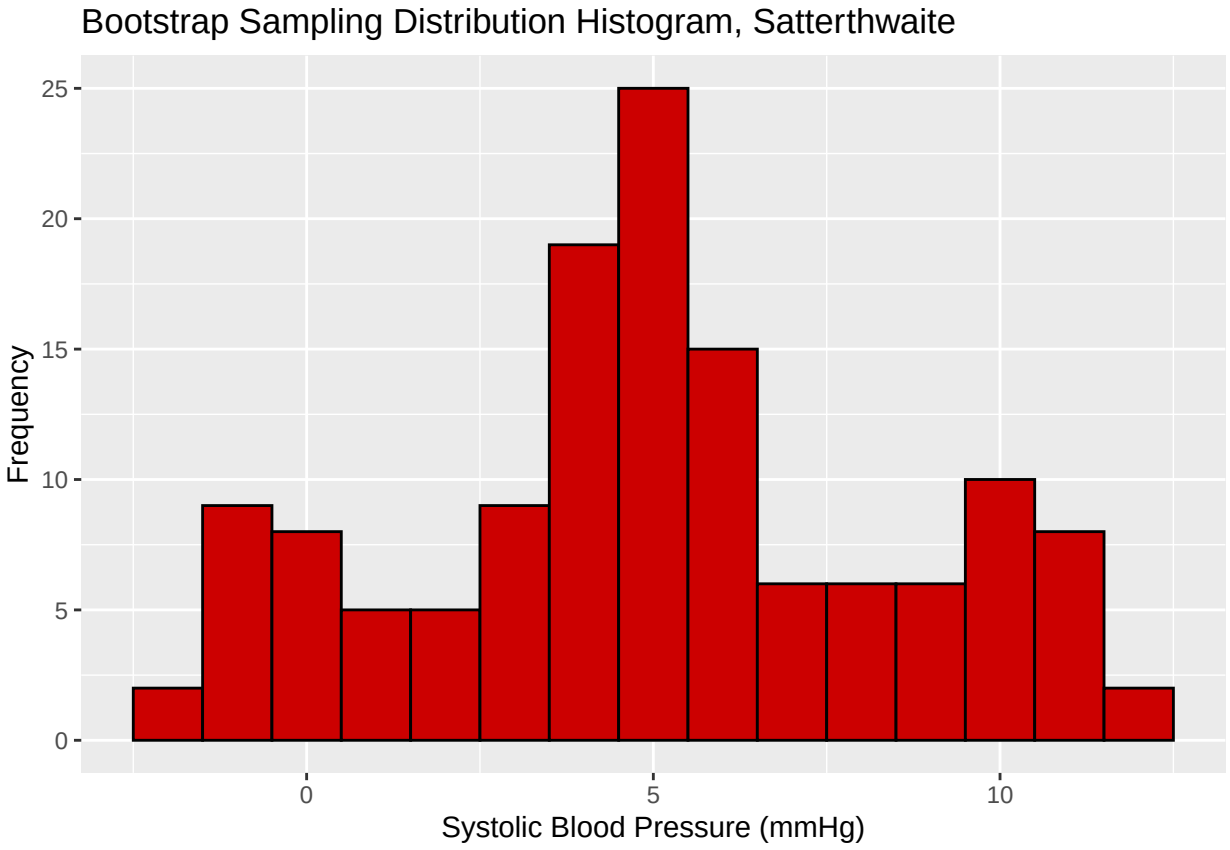
Bootstrapping

Another effective means of estimating a population parameter is the bootstrapping method. To model the Framingham Heart Study data and their means using this method, we will draw samples from the study and make inference based on those samples. Computer software does the heavy lifting for this exercise as the iterations require randomization of normality and many calculations.

The first thought that comes to mind when considering the use of iterative functions was the creation of a "for" loop. We wanted to process our data efficiently and creating a loop helped cut down on much of the manual formula-making. There are predefined parameter assumptions to consider which resulted in the addition of a nesting feature within our "for" loop. We considered variances of σ_1^2 = 1,4,9 and $\sigma_2^2 = 1$, samples sizes of n_1 = 10, 30, 70 and n_2 = 10, 30, 70, and fixed mean $\mu_1 = 5$ and the mean differences $\mu_1 - \mu_2$ = -5, -1, 1, and 5 with μ_2 = c(10, 6, 4, 0). There ended up being 135 unique combinations of parameters to evaluate. Before running the for loop, we created several trivial matrices that we could fill in later with data from our iterations.

We randomly generated 100 data sets each for the pooled and the Satterthwaite t-tests. The histograms below represents the sample distribution. We can observe that the samples follow Central Limit Theorem property regarding a convergence to a normal distribution with an increase in sample size.


Bootstrap Sampling Distribution Histogram, Pooled
Frequency
30
20
10
0
0
4
8
12
Systolic Blood Pressure (mmHg)



Our loop generated many simulated samples of smokers and non-smokers datasets and then processed each group under the conditions of a pooled t-test. A second iteration captured the Satterthwaite variant. We plotted the changes in power when the variance changed, the sample size changed and when the difference in means changed. We found that as the variance increase the power decreased. We also observed an increase in power as N1/N2 increased. When the difference between means grew larger, so did the power.

t-Test Pooled Facet grids containing power variability.

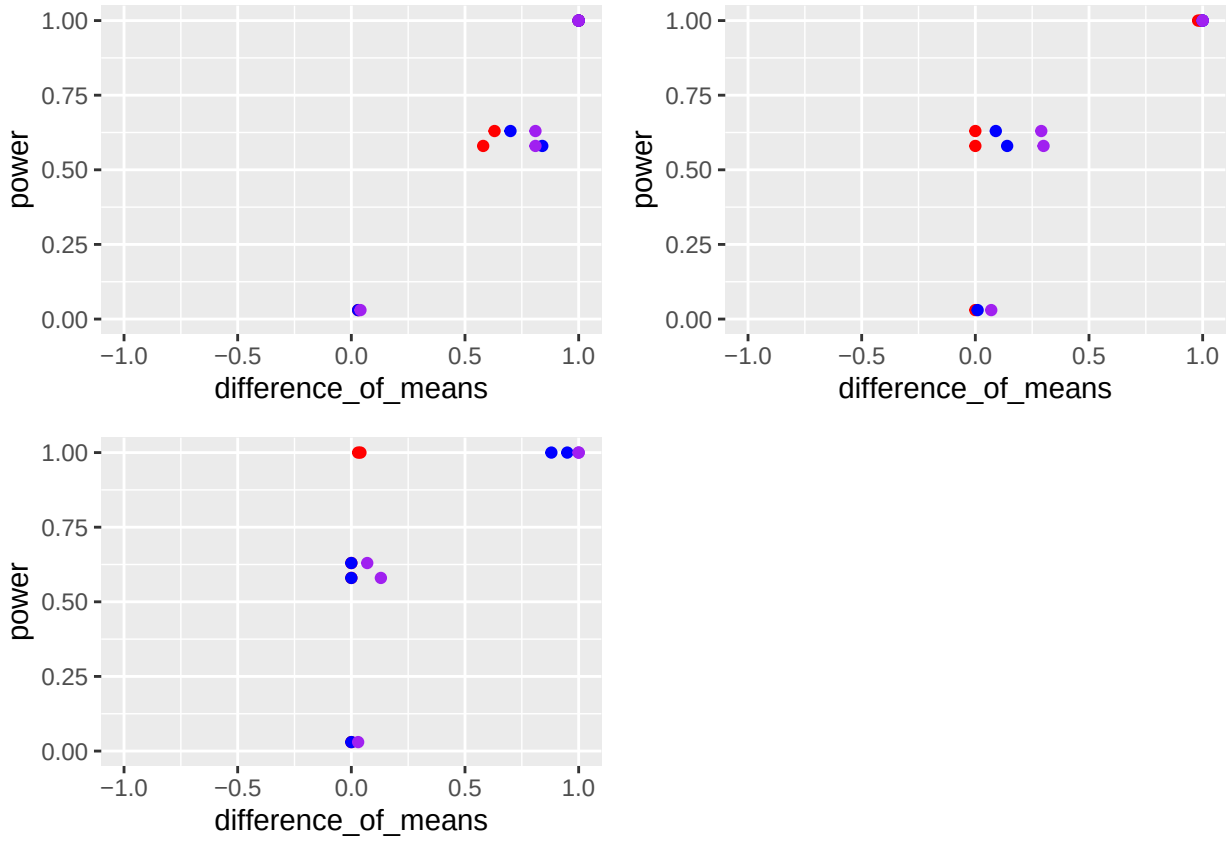


Figure 1: Power Plot when n2 = 10 and all other values held constant

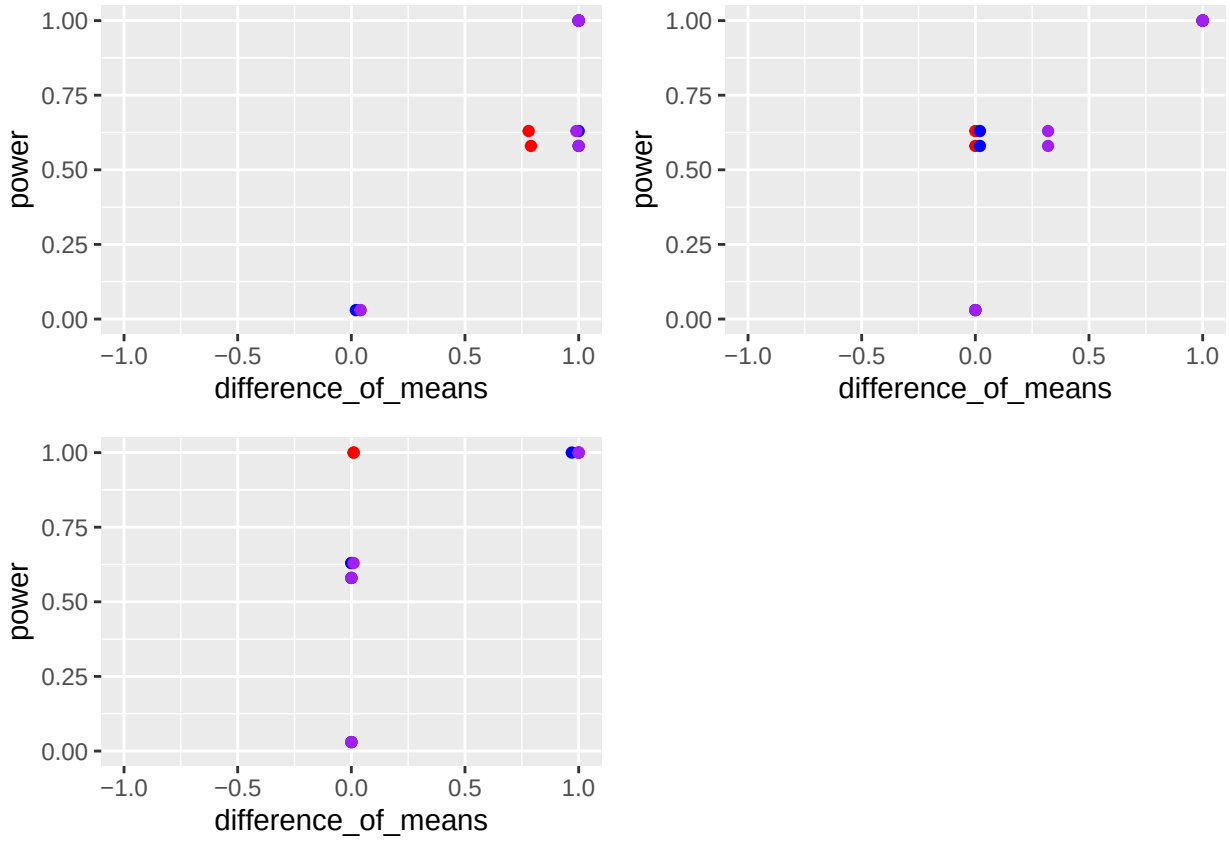


Figure 2: Power Plot when n2 = 30 and all other values held constant

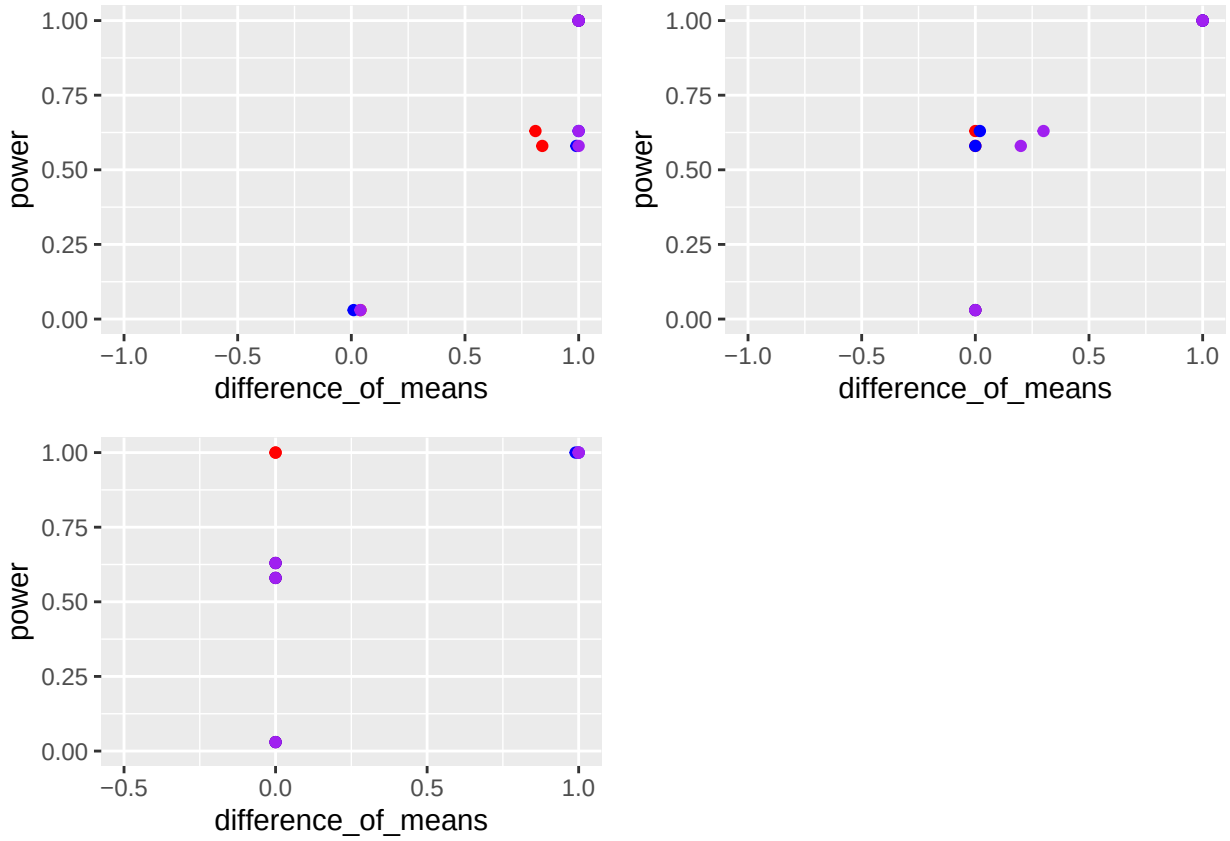


Figure 3: Power Plot when n2 = 70 and all other values held constant

We can see the power increase in rate as we consider each increasing N2 value. The data points migrate to the center and one of the points reaches capacity.

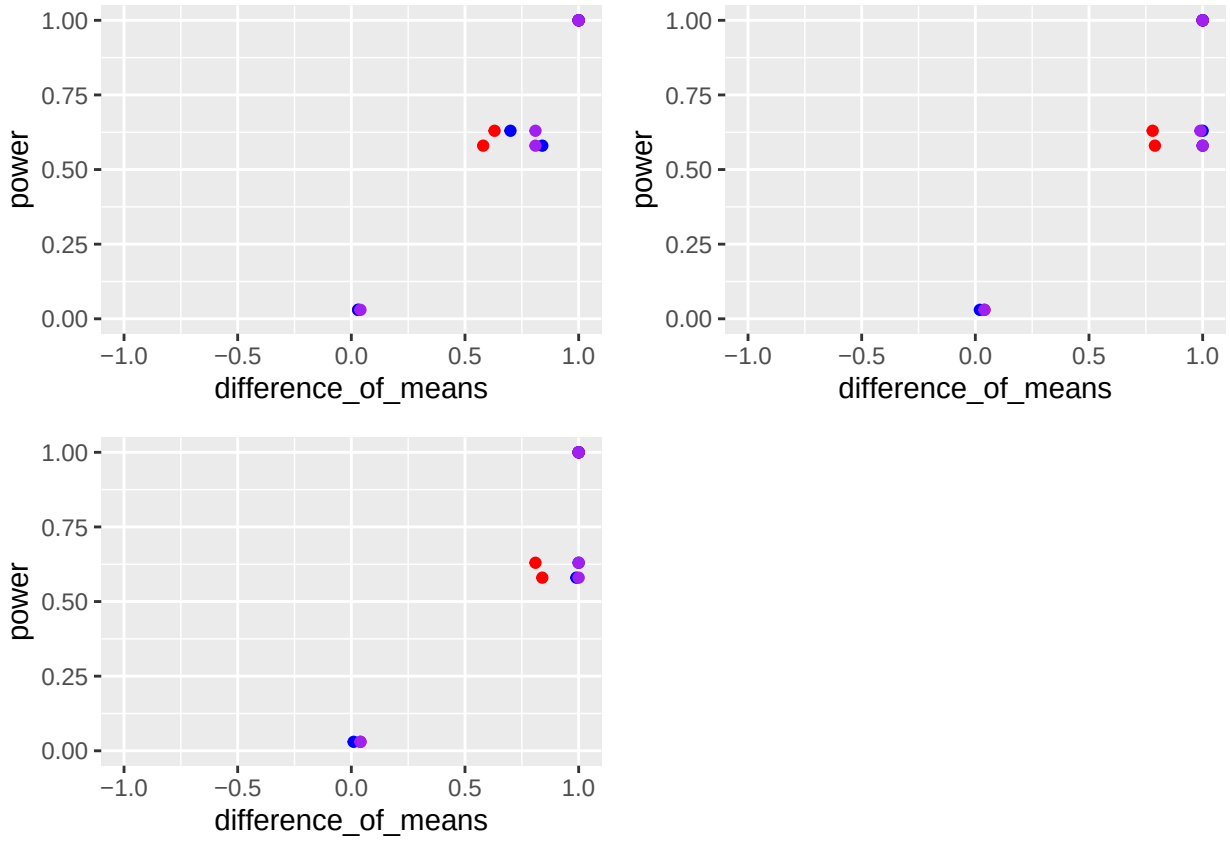


Figure 4: Power Plot when S1 = 1 and all other values held constant

Satterthwaite Facet grids containing power variability.

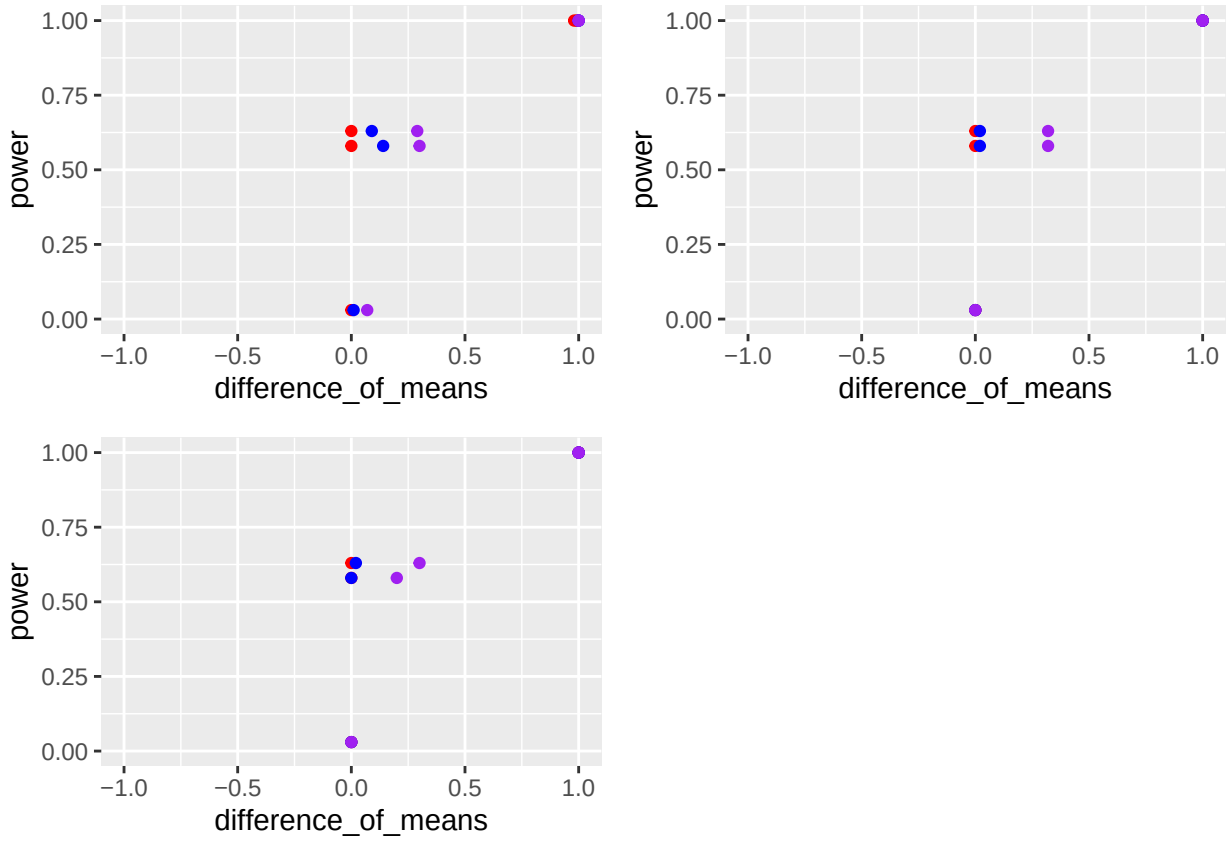


Figure 5: Power Plot when S1 = 4 and all other values held constant

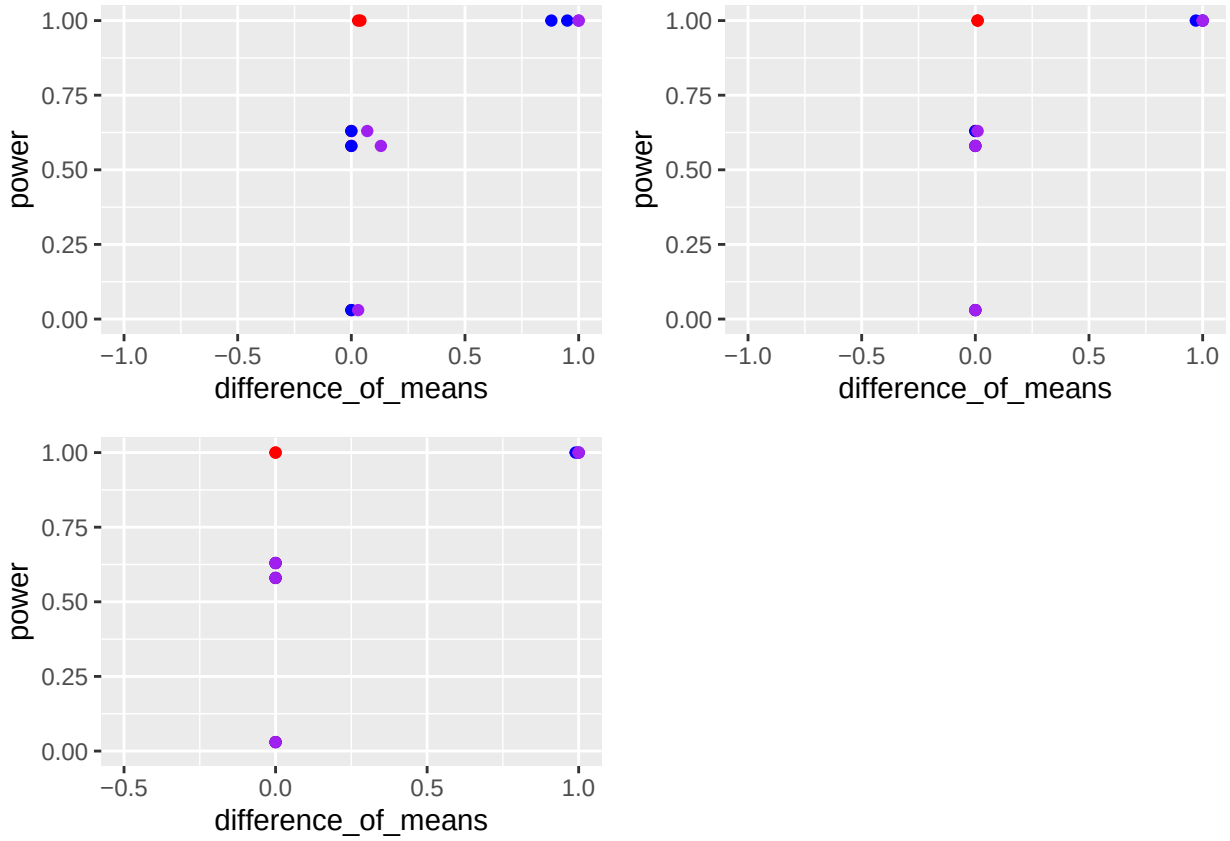


Figure 6: Power Plot when S1 = 9 and all other values held constant

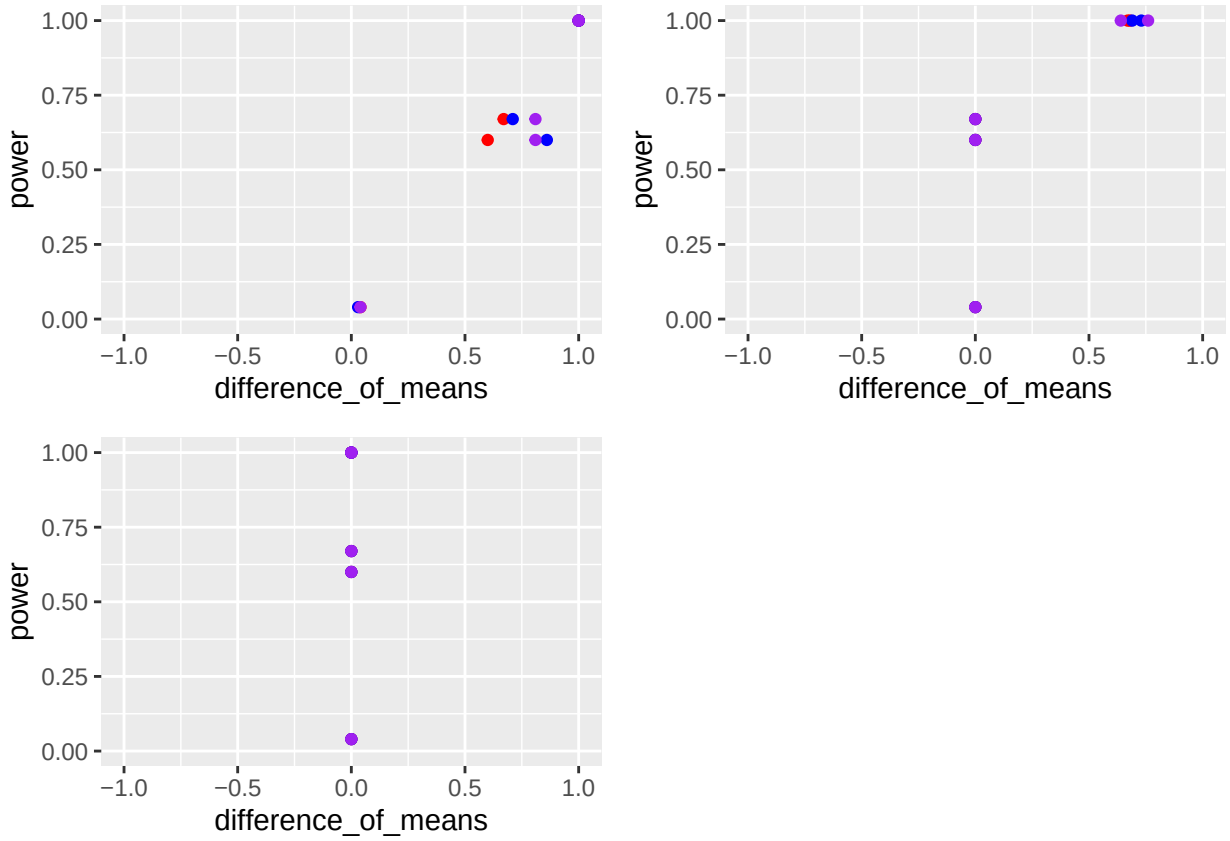


Figure 7: Power Plot when n2 = 10 and all other values held constant

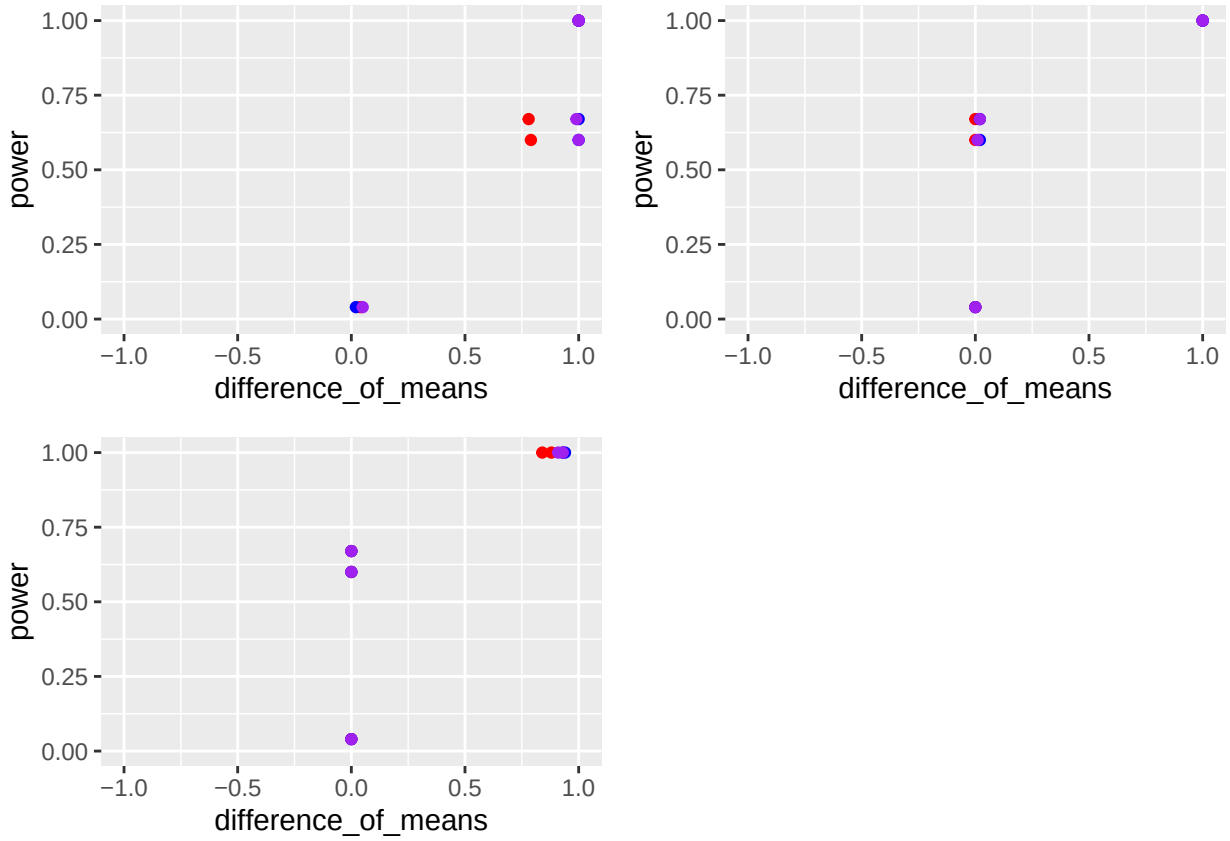


Figure 8: Power Plot when n2 = 30 and all other values held constant

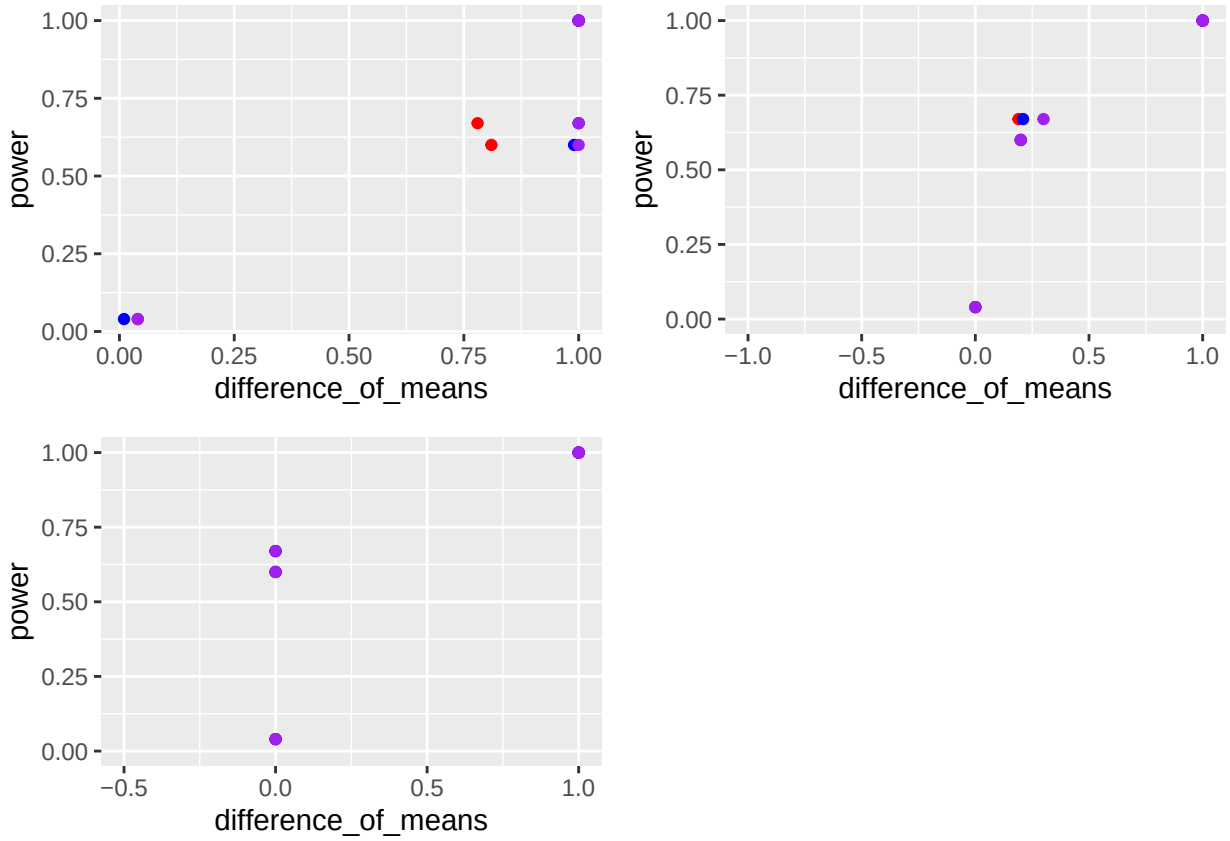


Figure 9: Power Plot when n2 = 70 and all other values held constant

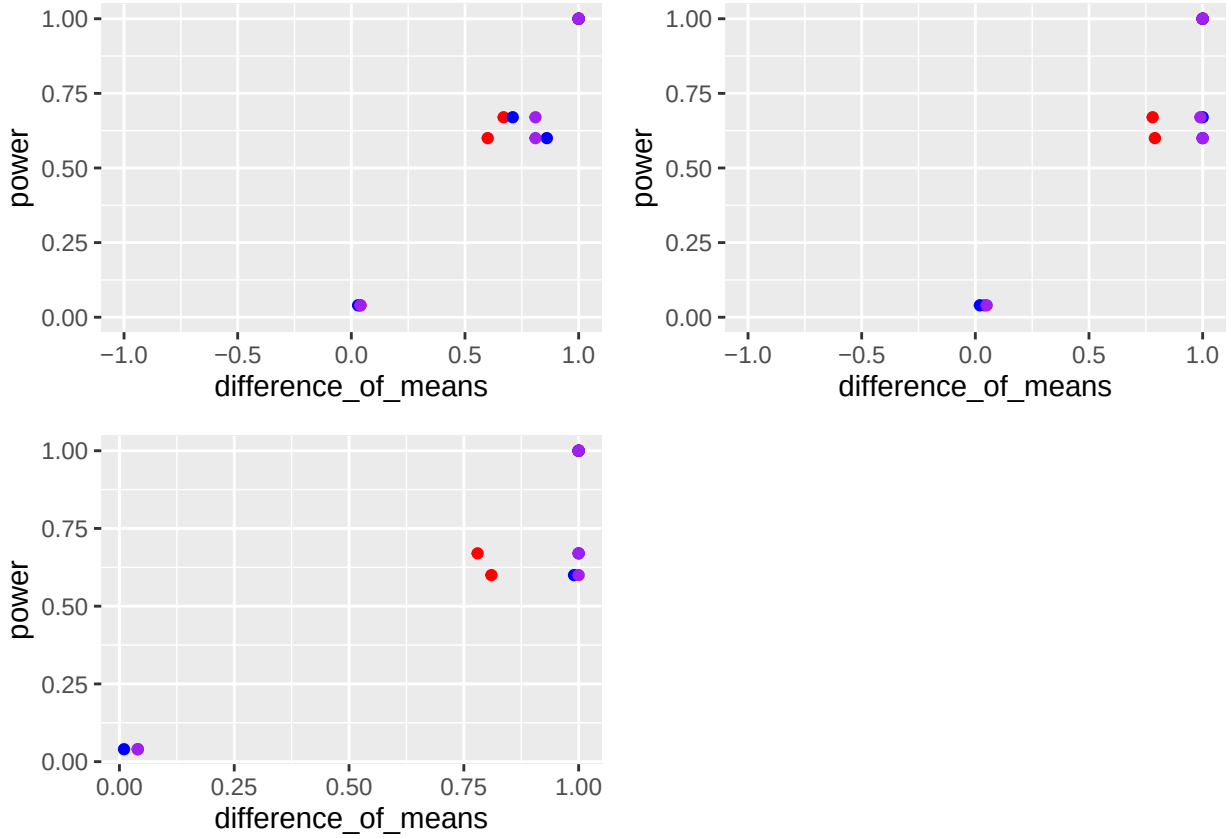


Figure 10: Power Plot when S1 = 1 and all other values held constant

Confidence Interval for Bootstrapping

Finally, we captured the confidence interval of our sampling distribution by finding the middle 95% of our distribution.

Simulation	Mean	Standard Deviation	Lowerbound	Upperbound
Smokers	5.2175	3.0735	-0.9296	11.3646

Simulation	Mean	Standard Deviation	Lowerbound	Upperbound
Non-Smokers	0.0266	0.9929	-1.9596	2.0119

We are 95% Confidence that the simulated mean of the smoker group is between -0.9296 and 11.3646.
We are 95% Confidence that the simulated mean of the non-smoker group is between -1.9596 and 2.0119.

Bibliography Burke GM, Genuardi M, Shappell H, D'Agostino RB Sr, Magnani JW. Temporal Associations Between Smoking and Cardiovascular Disease, 1971 to 2006 (from the Framingham Heart Study). Am J Cardiol. 2017;120(10):1787-1791. doi:10.1016/j.amjcard.2017.07.087

Hypertensive Crisis: When You Should Call 9-1-1 for High Blood Pressure. (2017). Www.heart.org. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings/hypertensive-crisis-when-you-should-call-911-for-high-blood-pressure

Mayo Clinic. (2018). High blood pressure (hypertension) - Diagnosis and treatment. Mayoclinic.org. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/diagnosis-treatment/drc-20373417

Centers for Disease Control and Prevention. (2017, February 9). Health Effects of Smoking and Tobacco Use. Centers for Disease Control and Prevention. https://www.cdc.gov/tobacco/basic_information/health_effects/index.htm#:~:text=Smoking%20causes%20cancer%2C%20heart%20disease

R CODE Plot code

```
#Histogram with outliers
ggplot(framingham_data,mapping = aes(`sysBP`)) +
  geom_histogram(binwidth = 2, fill="#CC0000", color="#000000") +
  labs(title="Framingham Data:  Systolic Blood Pressure Observations", y="Frequency", x="Systolic Blood

#QQ-plot with outliers
ggplot(framingham_data, mapping = aes(sample = framingham_data$sysBP)) +
  geom_qq(size = 1, fill="#CC0000", color="#CC0000") +
  labs(title="Framingham Data:  Systolic Blood Pressure Observations", y="Blood Pressure (mmHg)", x="No

#Boxplot with outliers
ggplot(framingham_data, mapping = aes(x=`sysBP`)) +
  geom_boxplot(fill="#CC0000") +
  labs(title="Framingham Data:  Sample Systolic Blood Pressure for Smokers and Non-Smokers", x="Systoli
  theme(axis.text.y=element_blank(),
        axis.ticks.y=element_blank())

#Historgram post outliers
```

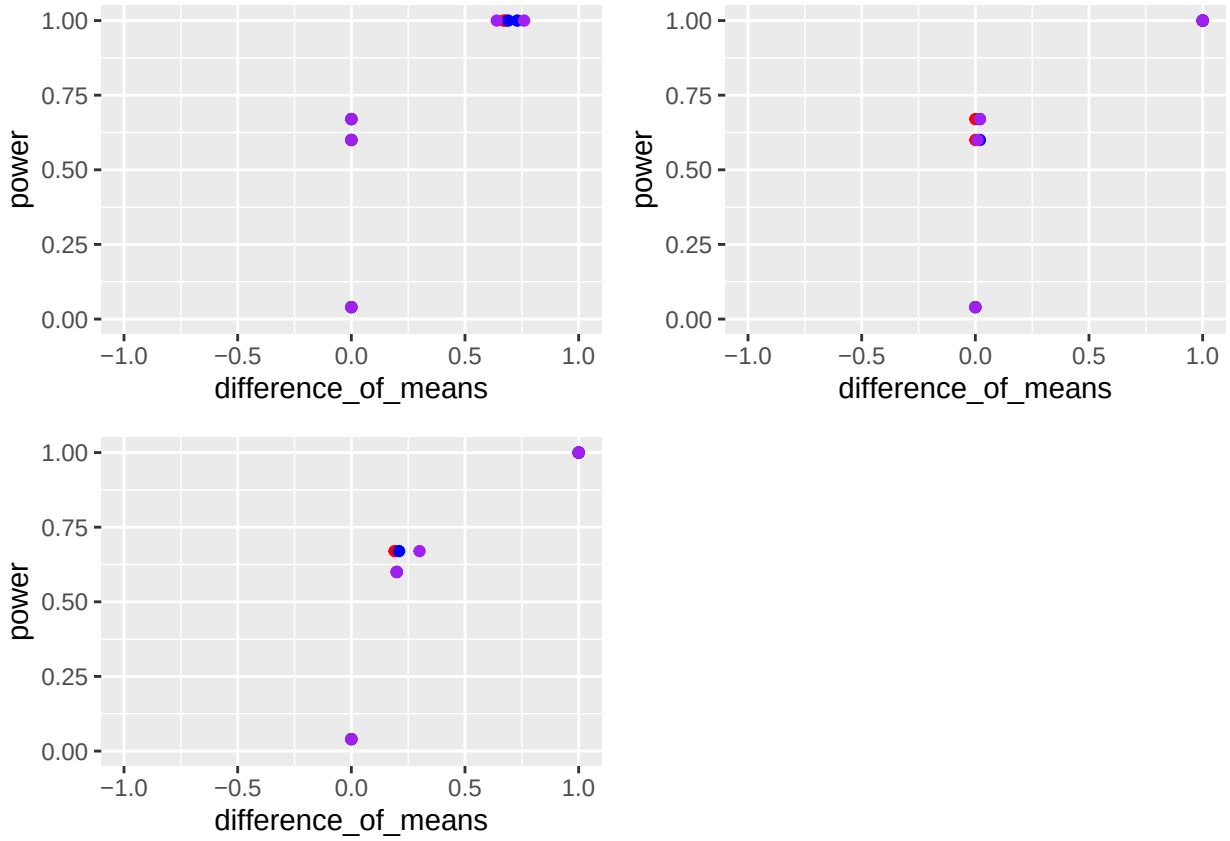


Figure 11: Power Plot when S1 = 4 and all other values held constant

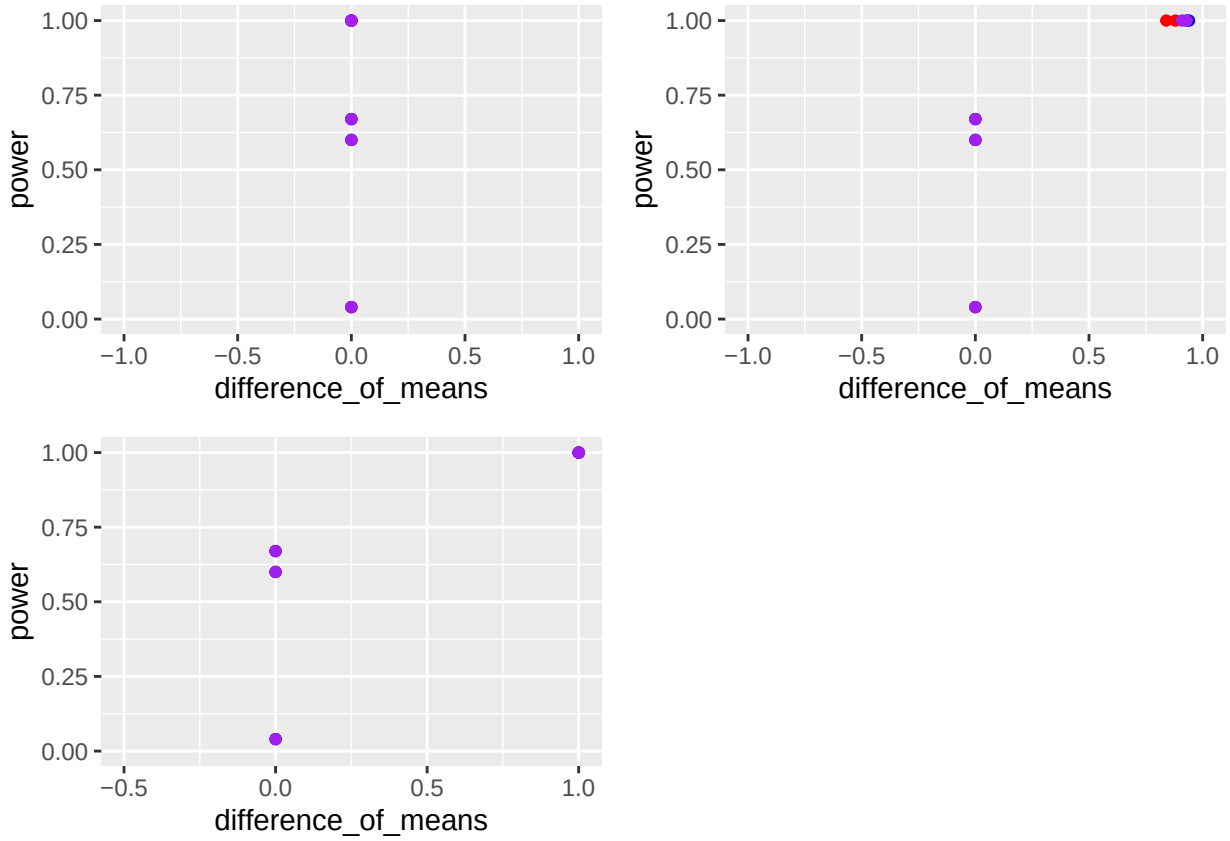


Figure 12: Power Plot when S1 = 9 and all other values held constant

```
ggplot(framingham_data_no_outlier,mapping = aes(`sysBP`)) +
  geom_histogram(binwidth = 2, fill="#CC0000", color="#000000") +
  labs(title="Framingham Data:  Systolic Blood Pressure Observations (Outliers Omitted)", y="Frequency"

#QQ-plot post outliers
ggplot(framingham_data_no_outlier, mapping = aes(sample = `sysBP`)) +
  geom_qq(size = 1, fill="#CC0000", color="#CC0000") +
  labs(title="Framingham Data:  Systolic Blood Pressure Observations (Outliers Omitted)", y="Blood Press
```

Pooled T-Test (Outliers Included)

```
#Variable Naming Convention:  <SampledGroup>_<content>_<specialcondition>
#Code until break contains t-test pooled values for p-value with outliers
NS_RawData_OL <- framingham_data[currentSmoker==0,2] #Grouped non-smokers together with outliers
S_RawData_OL <- framingham_data[currentSmoker==1,2] #Grouped smokers together with outliers

NS_RawData_OL <- unlist(framingham_data[currentSmoker==0,2])
NS_Mean_OL <- mean(NS_RawData_OL) #Mean systolic blood pressure of non-smokers with outliers

S_RawData_OL = unlist(framingham_data[currentSmoker==1,2])
S_Mean_OL = mean(S_RawData_OL) #Mean systolic blood pressure of smokers with outliers

NS_STDDEV_OL = sd(NS_RawData_OL) #Standard deviation systolic blood pressure for non-smokers with outli
S_STDDEV_OL = sd(S_RawData_OL) #Standard deviation systolic blood pressure for smokers with outliers

NS_OBS_OL = length(NS_RawData_OL) #number of non-smokers in the sample with outliers
S_OBS_OL = length(S_RawData_OL) #number of non-smokers in the sample with outliers

SNS_WgtAve_OL = ((NS_OBS_OL-1)*NS_STDDEV_OL^2 + (S_OBS_OL-1)*S_STDDEV_OL^2)/(NS_OBS_OL+S_OBS_OL-2) #Wei
SNS_TStat_OL = (NS_Mean_OL - S_Mean_OL)/(sqrt(SNS_WgtAve_OL)*sqrt(1/NS_OBS_OL + 1/S_OBS_OL)) #Test stat
SNS_PVAL_OL <- 2*(1-pt(3.04,298)) #P-Value for a two tailed t-test with degrees of freedom 3.04131 and




#Code until break contains t-test pooled values for Confidence Interval test method with outliers
SNS_Crit_OL <- qt(0.025,298)

Lower_Bound_OL <- (NS_Mean_OL - S_Mean_OL) - (SNS_Crit_OL)*(sqrt(SNS_WgtAve_OL)*sqrt(1/NS_OBS_OL + 1/S_

Upper_Bound_OL <- (NS_Mean_OL - S_Mean_OL) + (SNS_Crit_OL)*(sqrt(SNS_WgtAve_OL)*sqrt(1/NS_OBS_OL + 1/S_
```

Pooled T-Test (Outliers Omitted)

```
#Code until break contains t-test pooled values for p-value with no outliers
NS_RawData_NoOL <- framingham_data[currentSmoker==0,2] #Grouped non-smokers together with no outliers
S_RawData_NoOL <- framingham_data[currentSmoker==1,2] #Grouped smokers together with no outliers
```

```
NS_RawData_NoOL <- unlist(framingham_data[currentSmoker==0,2])
NS_Mean_NoOL <- mean(NS_RawData_NoOL) #Mean systolic blood pressure of non-smokers with no outliers

S_RawData_NoOL = unlist(framingham_data[currentSmoker==1,2])
S_Mean_NoOL = mean(S_RawData_NoOL) #Mean systolic blood pressure of smokers with no outliers

NS_STDDEV_NoOL = sd(NS_RawData_NoOL) #Standard deviation systolic blood pressure for non-smokers with n
S_STDDEV_NoOL = sd(S_RawData_NoOL) #Standard deviation systolic blood pressure for smokers with no outl

NS_OBS_NoOL = length(NS_RawData_NoOL) #number of non-smokers in the sample with no outliers
S_OBS_NoOL = length(S_RawData_NoOL) #number of non-smokers in the sample with no outliers

SNS_WgtAve_NoOL = ((NS_OBS_NoOL-1)*NS_STDDEV_NoOL^2 + (S_OBS_NoOL-1)*S_STDDEV_NoOL^2)/(NS_OBS_NoOL+S_OBS
SNS_TStat_NoOL = (NS_Mean_NoOL - S_Mean_NoOL)/(sqrt(SNS_WgtAve_NoOL)*sqrt(1/NS_OBS_NoOL + 1/S_OBS_NoOL))
SNS_PVAL_NoOL <- 2*(1-pt(3.04,298)) #P-Value for a two tailed t-test with degrees of freedom 3.04131 an




#Code until break contains t-test pooled values for Confidence Interval test method with no outliers
SNS_Crit_NoOL <- qt(0.025,298)

Lower_Bound_NoOL <- (NS_Mean_NoOL - S_Mean_NoOL) - (SNS_Crit_NoOL)*(sqrt(SNS_WgtAve_NoOL)*sqrt(1/NS_OBS

Upper_Bound_NoOL <- (NS_Mean_NoOL - S_Mean_NoOL) + (SNS_Crit_NoOL)*(sqrt(SNS_WgtAve_NoOL)*sqrt(1/NS_OBS
```

Welch-Satterthwaite T-test (Outliers Included)

```
SNS_TStat_OL_SW <- (NS_Mean_OL-S_Mean_OL)/sqrt((NS_STDDEV_OL/NS_OBS_OL)+(S_STDDEV_OL/S_OBS_OL)) #Test S
v <- (((NS_STDDEV_OL/NS_OBS_OL)+(S_STDDEV_OL/S_OBS_OL))^2)/(((NS_STDDEV_OL/NS_OBS_OL)^2/(NS_OBS_OL-1))+
```

Welch-Satterthwaite T-test (Outliers Omitted)

```
SNS_TStat_NoOL_SW <- (NS_Mean_NoOL-S_Mean_NoOL)/sqrt((NS_STDDEV_NoOL/NS_OBS_NoOL)+(S_STDDEV_NoOL/S_OBS_
v <- (((NS_STDDEV_NoOL/NS_OBS_NoOL)+(S_STDDEV_NoOL/S_OBS_NoOL))^2)/(((NS_STDDEV_NoOL/NS_OBS_NoOL)^2/(NS
```

Bootstrapping

```
#Setting Parameters
N1 <- c(10,30,70) #Sample Size from Sample_1
N2 <- c(10,30,70) #Sample Size from Sample_2
MU2 <- c(5, 10, 6, 4, 0) #mu1 fixed to a value of 5 such that the mean difference u1-u2 = c(0,-5, -1, 1
S1 <- c(1, 4, 9) #Variance of Sample_1
B = 100 #Replicating 100 times
alpha = 0.05 #Significance Level
result1 = matrix(0, nr=135, nc=B)  #setting matrix to record p-values for equal variance where there ar
```

```r
result2 = matrix(0, nr=135, nc=B)  #setting matrix to record p-values for unequal variance where there

#Dummy Matrix Creation
para = matrix(0, nr=135, nc=4)  #4 parameters c("N1", "N2", "S1", "Mu2")
pvalueresults1 = matrix(0, nr=135, nc=B) #pvalue1 capture matrix dim(135,100)
pvalueresults2 = matrix(0, nr=135, nc=B) #pvalue1 capture matrix dim(135,100)
Y1Results = matrix(0, nr=135) #rnorm_1 capture matrix dim(135,100)
Y2Results = matrix(0, nr=135) #rnorm_2 capture matrix dim(135,100)

#The "para" set are the combinatorics for each of the cases of paramaterization that we will evaluate.
para[1:5,1]=10;     para[1:5,2]=10;     para[1:5,3]=1;      para[1:5,4]=MU2
para[6:10,1]=10;    para[6:10,2]=30;    para[6:10,3]=1;     para[6:10,4]=MU2
para[11:15,1]=10;   para[11:15,2]=70;   para[11:15,3]=1;    para[11:15,4]=MU2
para[16:20,1]=10;   para[16:20,2]=10;   para[16:20,3]=4;    para[16:20,4]=MU2
para[21:25,1]=10;   para[21:25,2]=30;   para[21:25,3]=4;    para[21:25,4]=MU2
para[26:30,1]=10;   para[26:30,2]=70;   para[26:30,3]=4;    para[26:30,4]=MU2
para[31:35,1]=10;   para[31:35,2]=10;   para[31:35,3]=9;    para[31:35,4]=MU2
para[36:40,1]=10;   para[36:40,2]=30;   para[36:40,3]=9;    para[36:40,4]=MU2
para[41:45,1]=10;   para[41:45,2]=70;   para[41:45,3]=9;    para[41:45,4]=MU2
para[46:50,1]=30;   para[46:50,2]=10;   para[46:50,3]=1;    para[46:50,4]=MU2
para[51:55,1]=30;   para[51:55,2]=30;   para[51:55,3]=1;    para[51:55,4]=MU2
para[56:60,1]=30;   para[56:60,2]=70;   para[56:60,3]=1;    para[56:60,4]=MU2
para[61:65,1]=30;   para[61:65,2]=10;   para[61:65,3]=4;    para[61:65,4]=MU2
para[66:70,1]=30;   para[66:70,2]=30;   para[66:70,3]=4;    para[66:70,4]=MU2
para[71:75,1]=30;   para[71:75,2]=70;   para[71:75,3]=4;    para[71:75,4]=MU2
para[76:80,1]=30;   para[76:80,2]=10;   para[76:80,3]=9;    para[76:80,4]=MU2
para[81:85,1]=30;   para[81:85,2]=30;   para[81:85,3]=9;    para[81:85,4]=MU2
para[86:90,1]=30;   para[86:90,2]=70;   para[86:90,3]=9;    para[86:90,4]=MU2
para[91:95,1]=70;   para[91:95,2]=10;   para[91:95,3]=1;    para[91:95,4]=MU2
para[96:100,1]=70;  para[96:100,2]=30;  para[96:100,3]=1;   para[96:100,4]=MU2
para[101:105,1]=70; para[101:105,2]=70; para[101:105,3]=1;  para[101:105,4]=MU2
para[106:110,1]=70; para[106:110,2]=10; para[106:110,3]=4;  para[106:110,4]=MU2
para[111:115,1]=70; para[111:115,2]=30; para[111:115,3]=4;  para[111:115,4]=MU2
para[116:120,1]=70; para[116:120,2]=70; para[116:120,3]=4;  para[116:120,4]=MU2
para[121:125,1]=70; para[121:125,2]=10; para[121:125,3]=9;  para[121:125,4]=MU2
para[126:130,1]=70; para[126:130,2]=30; para[126:130,3]=9;  para[126:130,4]=MU2
para[131:135,1]=70; para[131:135,2]=70; para[131:135,3]=9;  para[131:135,4]=MU2

set.seed(0821)
for(i in 1:135){ #for loop to go through each case
  sigma1 = para[i,3]; #invokes the different evaluations of the variance terms
  sigma2 = 1; #variance_2 was said to be evaluated at a constant of 1.
  n1 = para[i,1]; #invokes the different evaluations of sample size of n_1
  n2 = para[i,2]; #invokes the different evaluations of sample size of n_1
  mu1 = 5; #u1 was said to evaluated at a constant of 5
  mu2 = para[i,4]; #invokes the different evaluations of mu2

  for(b in 1:B){
    Y1 = rnorm(n1, mu1, sqrt(sigma1)) #Creates a normal distribution using parameters outlined in assum
    Y1Results[i] <- Y1 #Save Y1 results to the Y1Results matrices
    Y2 = rnorm(n2, mu2, sqrt(sigma2)) #Creates a normal distribution using parameters outlined in assum
    Y2Results[i] <- Y2 #Save Y2 results to the Y2Results matrices
```

```
    #Evaluation of the Pooled
    meanY1 = mean(Y1) #Takes a mean of the normal distribution and saves it to meanY1
    meanY2 = mean(Y2) #Takes a mean of the normal distribution and saves it to meanY2
    Sp_2 = ((n1-1)*sigma1^2 + (n2-1)*sigma2^2)/(n1+n2-2) #Pooled Variance multiplier
    T_1 = abs(meanY1-meanY2)/(sqrt(Sp_2)*sqrt(1/n1 + 1/n2)) #equal variance test statistic (mean has to
    df_1 = n1 + n2 - 2 #Degrees of Freedom
    pvalue1 = 2*(1-pt(T_1, df_1)) #equal variance p-value
    pvalueresults1[i] <- pvalue1 #saves the p-value to one of our dummy matrices

    #Evaluation of the Satterthwaite
    T_2 = abs(meanY1-meanY2)/(sqrt((sigma1^2/n1)+(sigma2^2/n2))) #unequal variance t-test statistic
    df_2 = ((sigma1^2/n1)+(sigma2^2/n2)^2)/((((sigma1^2/n1)^2)/(n1-1) + (((sigma2^2/n2)^2)/(n2-1)))) #U
    pvalue2 = 2*(1-pt(T_2, df_2)) #Unequal variance p-value
    pvalueresults2[i] <- pvalue2 #saves the p-value to one of our dummy matrices


    if(pvalue1 <= alpha){  # checking if p-value for equal variance test is less than or equal to alpha
      result1[i, b] = 1    # recording number of rejections where ith row corresponds to ith situation/
    } else{                # if p-value is greater than alpha
      result1[i, b] = 0    # recording number of fail to rejections
    }
    if(pvalue2 <= alpha){  # testing if p-value for unequal variance test is less than or equal to alph
      result2[i, b] = 1    # recording number of rejections
    } else{                # if p-value is greater than alpha
      result2[i, b] = 0    # recording number of fail to rejections
    }
  }

}

Type1error_1 <- rowMeans(result1)  #this is type1error for the equal variance test and is the mean/prop
Type1error_2 <- rowMeans(result2)  #this is type1error for the unequal variance test and is the mean/pr

Y1Results <- as.data.frame(Y1Results) #Saves Y1Results as a data.frame to be used in ggplot2
Y2Results <- as.data.frame(Y2Results) #Saves Y2Results as a data.frame to be used in ggplot2

#Confidence Interval for the Y1
    Y1mean <- mean(Y1) #Mean of Y1
    Y1sd <- sd(Y1) #Standard Deviation of Y1
    Y1_lb <- Y1mean - 2*(Y1sd) #Lower Bound CI of Y1
    Y1_ub <- Y1mean + 2*(Y1sd) #Upper Bound CI of Y1

    Y2mean <- mean(Y2) #Mean of Y2
    Y2sd <- sd(Y2) #Standard Deviation of Y2
    Y2_lb <- Y2mean - 2*(Y2sd) #Lower Bound CI of Y2
    Y2_ub <- Y2mean + 2*(Y2sd) #Upper Bound CI of Y2
```

```
data1 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
one = ggplot(data=data1, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[1:5]), color='red') + #where n1=10, n2=10, s1=1
```

```
  geom_point(mapping=aes(Type1error_1[6:10]), color='blue') + #where n1=10, n2=30, s1=1
  geom_point(mapping=aes(Type1error_1[11:15]), color='purple') + #where n1=10, n2=70, s1=1
  #facet_grid(Type1error_1[1:5]) +
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  #main="Plot for case n1=10, n2=10,30,70 and s1=1" +
  xlab(expression(difference_of_means))
data2 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
two = ggplot(data=data2, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[16:20]), color="red") + #where n1=10, n2=10, s1=4
  geom_point(mapping=aes(Type1error_1[21:25]), color="blue") + #where n1=10, n2=30, s1=4
  geom_point(mapping=aes(Type1error_1[26:30]), color="purple") + #where n1=10, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data3 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
three = ggplot(data=data3, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[31:35]), color="red") + #where n1=10, n2=10, s1=9
  geom_point(mapping=aes(Type1error_1[36:40]), color="blue") + #where n1=10, n2=30, s1=9
  geom_point(mapping=aes(Type1error_1[41:45]), color="purple") + #where n1=10, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data4 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
four = ggplot(data=data4, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[46:50]), color="red") + #where n1=30, n2=10, s1=1
  geom_point(mapping=aes(Type1error_1[51:55]), color="blue") + #where n1=30, n2=30, s1=1
  geom_point(mapping=aes(Type1error_1[56:60]), color="purple") + #where n1=30, n2=70, s1=1
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data5 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
five = ggplot(data=data5, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[61:65]), color="red") + #where n1=30, n2=10, s1=4
  geom_point(mapping=aes(Type1error_1[66:70]), color="blue") + #where n1=30, n2=30, s1=4
  geom_point(mapping=aes(Type1error_1[71:75]), color="purple") + #where n1=30, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data6 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
six = ggplot(data=data6, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[76:80]), color="red") + #where n1=30, n2=10, s1=9
  geom_point(mapping=aes(Type1error_1[81:85]), color="blue") + #where n1=30, n2=30, s1=9
  geom_point(mapping=aes(Type1error_1[86:90]), color="purple") + #where n1=30, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data7 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
seven = ggplot(data=data7, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[91:95]), color="red") + #where n1=70, n2=10, s1=1
  geom_point(mapping=aes(Type1error_1[96:100]), color="blue") + #where n1=70, n2=30, s1=1
  geom_point(mapping=aes(Type1error_1[101:105]), color="purple") + #where n1=70, n2=70, s1=1
  ylim(c(0,1)) +
```

```
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data8 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
eight = ggplot(data=data8, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[106:110]), color="red") + #where n1=70, n2=10, s1=4
  geom_point(mapping=aes(Type1error_1[111:115]), color="blue") + #where n1=70, n2=30, s1=4
  geom_point(mapping=aes(Type1error_1[116:120]), color="purple") + #where n1=70, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data9 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_1[1:5]))
nine = ggplot(data=data9, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_1[121:125]), color="red") + #where n1=70, n2=10, s1=9
  geom_point(mapping=aes(Type1error_1[126:130]), color="blue") + #where n1=70, n2=30, s1=9
  geom_point(mapping=aes(Type1error_1[131:135]), color="purple") + #where n1=70, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-0.5 ,1)) +
  xlab(expression(difference_of_means))

#TYPE1
data_1 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
one2 = ggplot(data=data_1, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[1:5]), color="red") + #where n1=10, n2=10, s1=1
  geom_point(mapping=aes(Type1error_2[6:10]), color="blue") + #where n1=10, n2=30, s1=1
  geom_point(mapping=aes(Type1error_2[11:15]), color="purple") + #where n1=10, n2=70, s1=1
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_2 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
two2 = ggplot(data=data_2, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[16:20]), color="red") + #where n1=10, n2=10, s1=4
  geom_point(mapping=aes(Type1error_2[21:25]), color="blue") + #where n1=10, n2=30, s1=4
  geom_point(mapping=aes(Type1error_2[26:30]), color="purple") + #where n1=10, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_3 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
three2 = ggplot(data=data_3, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[31:35]), color="red") + #where n1=10, n2=10, s1=9
  geom_point(mapping=aes(Type1error_2[36:40]), color="blue") + #where n1=10, n2=30, s1=9
  geom_point(mapping=aes(Type1error_2[41:45]), color="purple") + #where n1=10, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_4 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
four2 = ggplot(data=data_4, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[46:50]), color="red") + #where n1=30, n2=10, s1=1
  geom_point(mapping=aes(Type1error_2[51:55]), color="blue") + #where n1=30, n2=30, s1=1
  geom_point(mapping=aes(Type1error_2[56:60]), color="purple") + #where n1=30, n2=70, s1=1
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_5 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
```

```
five2 = ggplot(data=data_5, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[61:65]), color="red") + #where n1=30, n2=10, s1=4
  geom_point(mapping=aes(Type1error_2[66:70]), color="blue") + #where n1=30, n2=30, s1=4
  geom_point(mapping=aes(Type1error_2[71:75]), color="purple") + #where n1=30, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_6 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
six2 = ggplot(data=data_6, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[76:80]), color="red") + #where n1=30, n2=10, s1=9
  geom_point(mapping=aes(Type1error_2[81:85]), color="blue") + #where n1=30, n2=30, s1=9
  geom_point(mapping=aes(Type1error_2[86:90]), color="purple") + #where n1=30, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_7 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
seven2 = ggplot(data=data_7, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[91:95]), color="red") + #where n1=70, n2=10, s1=1
  geom_point(mapping=aes(Type1error_2[96:100]), color="blue") + #where n1=70, n2=30, s1=1
  geom_point(mapping=aes(Type1error_2[101:105]), color="purple") + #where n1=70, n2=70, s1=1

  xlab(expression(difference_of_means))
data_8 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
eight2 = ggplot(data=data_8, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[106:110]), color="red") + #where n1=70, n2=10, s1=4
  geom_point(mapping=aes(Type1error_2[111:115]), color="blue") + #where n1=70, n2=30, s1=4
  geom_point(mapping=aes(Type1error_2[116:120]), color="purple") + #where n1=70, n2=70, s1=4
  ylim(c(0,1)) +
  xlim(c(-1,1)) +
  xlab(expression(difference_of_means))
data_9 <- data.frame(difference_of_means=c(0, -5, -1, 1, 5), power=c(Type1error_2[1:5]))
nine2 = ggplot(data=data_9, mapping=aes(x=difference_of_means, y=power)) +
  geom_point(mapping=aes(Type1error_2[121:125]), color="red") + #where n1=70, n2=10, s1=9
  geom_point(mapping=aes(Type1error_2[126:130]), color="blue") + #where n1=70, n2=30, s1=9
  geom_point(mapping=aes(Type1error_2[131:135]), color="purple") + #where n1=70, n2=70, s1=9
  ylim(c(0,1)) +
  xlim(c(-0.5 ,1)) +
  xlab(expression(difference_of_means))
```

Bootstrap Graphs

```
ggplot(Y1Results, aes(`V1`)) +
  geom_histogram(binwidth = 1, fill="#CC0000", color="#000000") +
  labs(title="Bootstrap Sampling Distribution Histogram, Pooled", y="Frequency", x="Systolic Blood Press

ggplot(Y2Results, aes(`V1`)) +
  geom_histogram(binwidth = 1, fill="#CC0000", color="#000000") +
  labs(title="Bootstrap Sampling Distribution Histogram, Satterthwaite", y="Frequency", x="Systolic Bloo

#Type1error_1 plots
#plots t-Test Facet grids of power variability as N2 changes.
plot_grid(one, two, three)
plot_grid(four, five, six)
```

```
plot_grid(seven, eight, nine)

#plots t-Test Facet grids of power variability as S1 changes.
plot_grid(one, four, seven)
plot_grid(two, five, eight)
plot_grid(three, six, nine)

#Type1error_2 plots
#plots Satterthwaite Facet grids of power variability as N2 changes.
plot_grid(one2, two2, three2)
plot_grid(four2, five2, six2)
plot_grid(seven2, eight2, nine2)

#plots t-Test Facet grids of power variability as S1 changes.
plot_grid(one2, four2, seven2)
plot_grid(two2, five2, eight2)
plot_grid(three2, six2, nine2)
```

Contributions Brock Akerman: Part I graphs, Part I/II dialogue, All RMarkdown file output.
Hanan Ali: Part II R Code, Part II graphs and plots,
Mutual Aid: Worked separately on Part I formulation and coding but arrived at the same answers and conclusions.