ST502 Project

Halid Kopanski and Justin Feathers

Part I

Framingham Heart Study

The Framingham data set contains the systolic blood pressure of 300 smokers and nonsmokers. For this study we will assume the data are random samples from normal distributions (see Appendix I for more details) where μ_1 and σ_1^2 are the mean and variance for nonsmokers and μ_2 and σ_2^2 are the mean and variance for smokers. The null hypothesis, H_0: μ_1 - $\mu_2 = 0$, states there is no significant difference in blood pressure between smokers and nonsmokers. We wish to prove the alternative hypothesis, H_A: μ_1 - $\mu_2 \neq 0$, which states there is a significant difference in the means. We will accomplish this through hypothesis testing to determine if enough evidence does, indeed, exist to reject the null hypothesis.

Methods

We conducted the analysis under two conditions: the first assumed equal population variance (using pooled sample variance, S_p^2) and the second assumed unequal population variance. The results will be presented and any discrepancies between the two outcomes will be discussed.

Analysis

To begin, we first found the difference in mean blood pressure between smokers and nonsmokers in this dataset; this was calculated to be -9.16. Despite this value being nonzero, we cannot safely assume the difference occurred other than by random chance. For this part of the analysis we assumed equal variances, thereby allowing us to use pooled sample variance (see Appendix I: Equation 1) which was calculated to be 510. Under these conditions, the p value (probability of obtaining the observed results) was calculated to be 0.0041, using 298 degrees of freedom, which is smaller than (probability of committing a Type I error) of 0.05. In terms of t values, our observed t value of -3.04 is smaller than the t value, -1.97 for a two sided α of 0.05.

Additionally, the 95% confidence interval was calculated under the pooled variance assumption. The observed interval is -15.08 to -3.23. One can see that zero does not fall into this interval. Both calculations (p-value and 95% confidence interval) provides enough evidence to allow for the rejection of the null hypothesis. In other words, at an level of 0.05, there is significant evidence to suggest this group of smokers and nonsmokers have a different range of systolic blood pressure when using pooled variance.

Repeating the same approach as before but now assuming unequal population variances. In this case, the computed sample variances of the two groups were 352.2 for smokers and 562.1 for nonsmokers. Using equation 3 from Appendix 1, an observed t value of -2.9 was obtained. Comparing that the two-sided α of -1.98 with 158 degrees of freedom (as computed using the Satterthwaite Approximation; see Appendix I: Equation 4), the observed t value was less than the chosen α value of 0.05, as in the case of pooled sample variance.

Furthermore, the 95% confidence interval in the non pooled case was found to be -15.4 to -2.91. It can be noted that zero is absent from this range, as in the case of the pooled sample variance. In this second scenario, there is sufficient evidence to reject the null hypothesis and to support the alternative at an α level of 0.05.

We have shown that there is sufficient evidence to support the systolic blood pressure between smokers and nonsmokers is different. This statement holds true in all cases described in this paper. It should be noted, that while the observed systolic blood pressure values vary between smokers and non smokers, we do not know the underlying cause based on the provided data.

Test Preference

Here, both tests allow us to reject the null hypothesis, so either one could be useful. In this case, we must allow the computed statistics to dictate which test we ought to prefer. The standard deviation (pooled variance) was calculated from the denominator of Equation 2 (see Appendix I) which returned a value of 3.01. Similarly, the standard deviation (unequal variance) was calculated from the denominator of Equation 3 which returned a value of 9.99. We want to have data that is as consistent as possible, in other words, with the smallest spread, i.e., variance; therefore, we prefer the test with pooled variance.

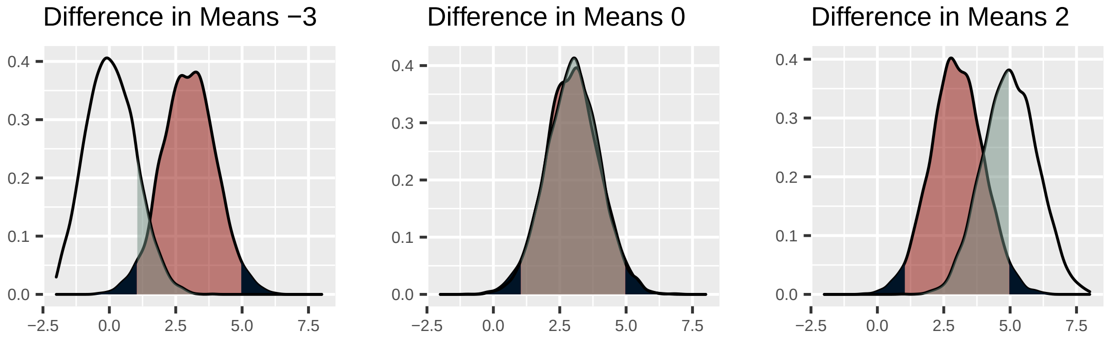
Part II

Introduction

In Part I, we were able to see how hypothesis testing allowed us to produce enough evidence to challenge the status quo argument that “smoking has no effect on the blood pressure of a person.” We found, through a number of robust statistical methods, the difference in means (calculated to be -9.16) was significant enough to not have arisen merely by chance. Within an α of 5%, we are confident that the two groups do exhibit measurable differences. In the following sections, we will explore, through simulated data, the concept of randomness in data and what steps we can take to mitigate the effect of randomness. All simulated data will be generated using the built-in rnorm() function in R.

Power of Hypothesis Testing

It is important to briefly discuss the concept of power, and by extension error types, in hypothesis testing. Power describes the probability of not committing a Type II error, which is to not reject the null hypothesis when there was in actuality enough evidence to do so. The probability of a Type II error is represented by β. Power is the complement to that probability (1 - β) and can be thought of as test sensitivity. The more powerful the test the more sensitive it is to differences between distributions. The value of β is the portion of the alternate distribution that is within the null hypothesis non rejection region limits. Below are a number of plots to help depict this concept. In each of the plots the shaded red is the null distribution and its location stays constant (mean = 3). The dark tails represent the rejection region of the null distribution and correspond to the size of α (in this case 2.5% on each side). α also indicates the probability of committing a Type I error, which is rejecting the null hypothesis incorrectly. The outline distribution represents the “true” alternative distribution. Within that distribution is the shaded light blue region which is equal to β. It can be seen that as the difference between the two distributions shrinks the value of β increases and the power shrinks. Power reaches minimum and β reaches maximum when the two distributions are the same. It should be noted that unless the null and alternative distributions are exact copies of one another, there is always a non zero probability of committing a Type I error.



Simulation Study

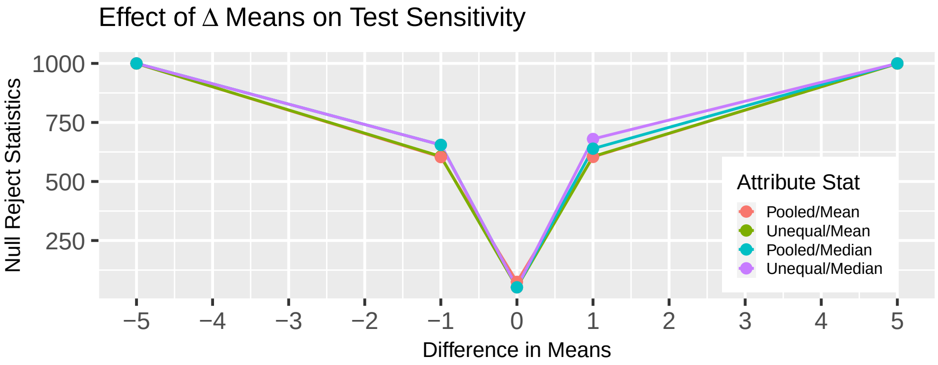
Multiple scenarios were simulated by creating 2 normally distributed data sets of various mean, variances, and sample sizes. Each scenario was repeated a thousand times. For each of those repeats a hypothesis test was conducted and the number of times the null hypothesis was rejected was recorded (see Appendix III for results). All hypothesis testing was using pooled and unequal variance, regardless of the actual values of variance.

The following table includes the values of each parameter used in the simulation:

Parameter	Possible Values
μ_1	0, 4, 5, 6, 10
σ_1^2	1, 4, 9
n_1	10, 30, 70
μ_2	5
σ_2^2	1
n_2	10, 30, 70

Observations

It was noted that the smaller difference between the alternative and null mean values were, the less likely the null hypothesis was rejected (see plot below and for more detail Appendix IV Item A). Larger values of variance also reduced the number of rejected null hypothesis (see Appendix IV Item B). In both cases, the more overlap between the two distributions, the less powerful the test. It was observed that increasing the sample size increased power (see Appendix IV Item C, D, and E). For a specific example, we can look at test cases #9 and #129 where $\mu_1 = 6$, $\mu_2 = 5$, $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$. The two cases differed in sample size where $n_9 = 10$ and $n_{129} = 70$. This increase in sample size resulted in almost a 4 fold increase in power (271 vs 951 (pooled) or 265 vs 948 (unequal) rejected nulls). It stands to reason that large variances and small deltas in mean can be mitigated by increasing the sample size accordingly.



Type I errors can be clearly observed in the simulated data. All test cases where $\Delta\mu$ was zero, roughly 50 out of the 1000 simulations resulted in rejected null hypotheses (see Appendix IV Item A table). These erroneous results, in roughly 5% of the cases, are due to the difference in variance between the distributions. That percentage value closely matches that of α.

One final observation, for each of the test cases (1 - β) was calculated under the assumption that the observed $\Delta\mu$ was the true distribution. It was found that (1 - β) values followed a similar pattern to that of the number of rejected null hypotheses, indicating the relationship between power and sensitivity. Below are the calculations for test cases 9 and 129:

	Test #9 (pooled)	Test #9 (unequal)	Test #129 (pooled)	Test #129 (unequal)
1st Quartile	0.1169414	0.1211982	0.8610788	0.8635657
Median	0.2979222	0.3113339	0.9652580	0.9587586
Mean	0.3793437	0.3775096	0.8953985	0.8888423
3rd Quartile	0.6114560	0.5889342	0.9947153	0.9937438

Conclusion

In Part I, we sought to reject the null hypothesis and determine which testing procedure was preferable by performing t-tests in terms of both p-values and confidence intervals with pooled and unequal variances. The results showed that there is a measurable difference in systolic blood pressure and a test with lower variance is preferable in order to reject the null hypothesis. These results, however, were the end-product of two tests run in two scenarios. To help us better understand the meaning behind the results, it was important to run a numerical simulation study to explore how various factors can effect testing outcomes.

After performing 1,000 trials for each of 135 different scenarios, the data showed that the probability of rejecting the null hypothesis increases with large differences in means or low variances. The data also showed that test sensitivity can be improved by increasing sample sizes. The greatest take away from this simulation study is that increasing the sample size greatly increases the power, or probability of rejecting the null hypothesis when it should, in fact, be rejected. We have also seen that the possibility of committing a Type I error is always present, even for similar distributions.

Appendix I: Equations and Figures

Normality of Data

Plotting the density of total data (Figure 1A), we can see that it closely follows a normal distribution. Density plots of the split data exhibit similar forms (Figure 1B and 1C). We calculated the kurtosis and the skew of the data and found them to 3.812, 0.88 for kurtosis and skew respectively. These values are quite close to those typically found in normally distributed data. Additionally, the calculated values of sample standard deviation, $\frac{IQR^*}{1.35}$, and $\frac{MAD^*}{0.675}$ are 22.89, 22.22, and 21.48. The similarities between the three values indicate that the data is not heavily influenced by outliers and the Q-Q plot confirms this (See Figure 2). Therefore, we will assume the data and the subsequent split data to be normal.

*Interquartile Range (IQR), Median Absolute Deviation (MAD)

Figure 1: Data plots

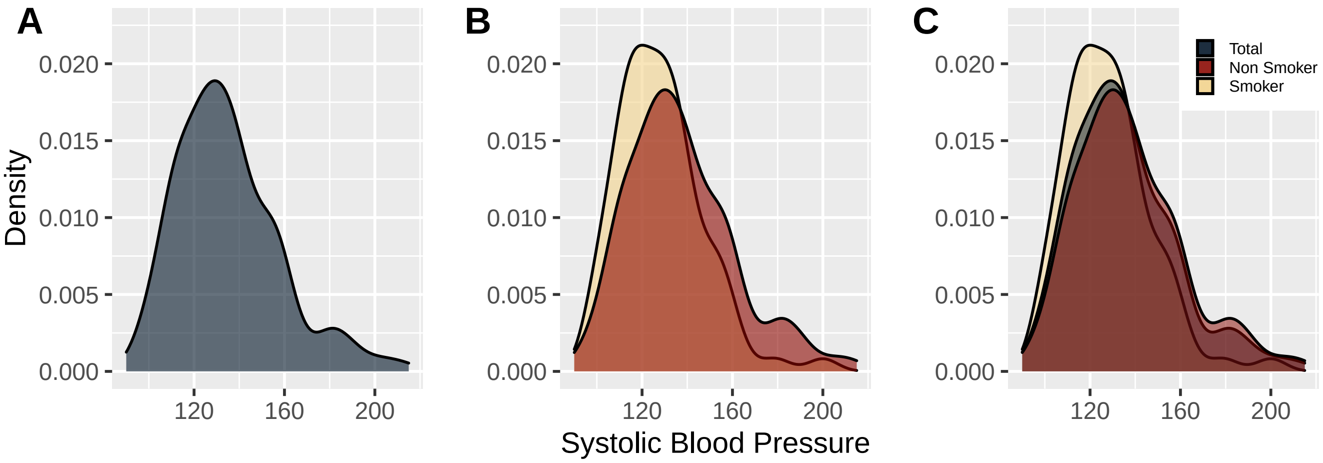
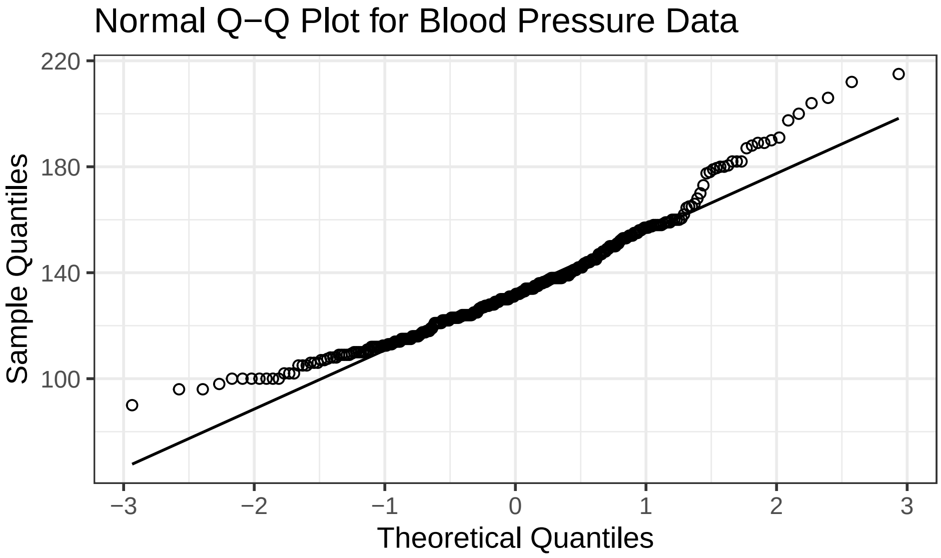


Figure 2: Q-Q Plot



Equations:

Equation 1: Pooled Sample Variance:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Equation 2: Observed T Statistic (pooled variance):

$$t_{obs} = \frac{\mu_1 - \mu_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Equation 3: Observed T Statistic (distinct variance):

$$t_{obs} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Equation 4: Satterthwaite Approximation:

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

Appendix II: Part I Results

	Attribute Value
Data Mean	134.935000
Data Sample Variance	524.085226
Smoker Mean	128.066667
Smoker Variance	352.211712
Nonsmoker Mean	137.224444
Nonsmoker Variance	562.144712
Difference in Mean	-9.157778
Pooled Sample Varance	510.013699
T-Statistic Std Deviation (Pooled)	3.011131
T-Statistic Std Deviation (Unequal)	9.993684
CI Pooled Lower	-15.080000
CI Pooled Upper	-3.230000
CI Nonpooled Lower	-15.400000
CI Nonpooled Upper	-2.910000

Appendix III: Part II Results

Simulation Data

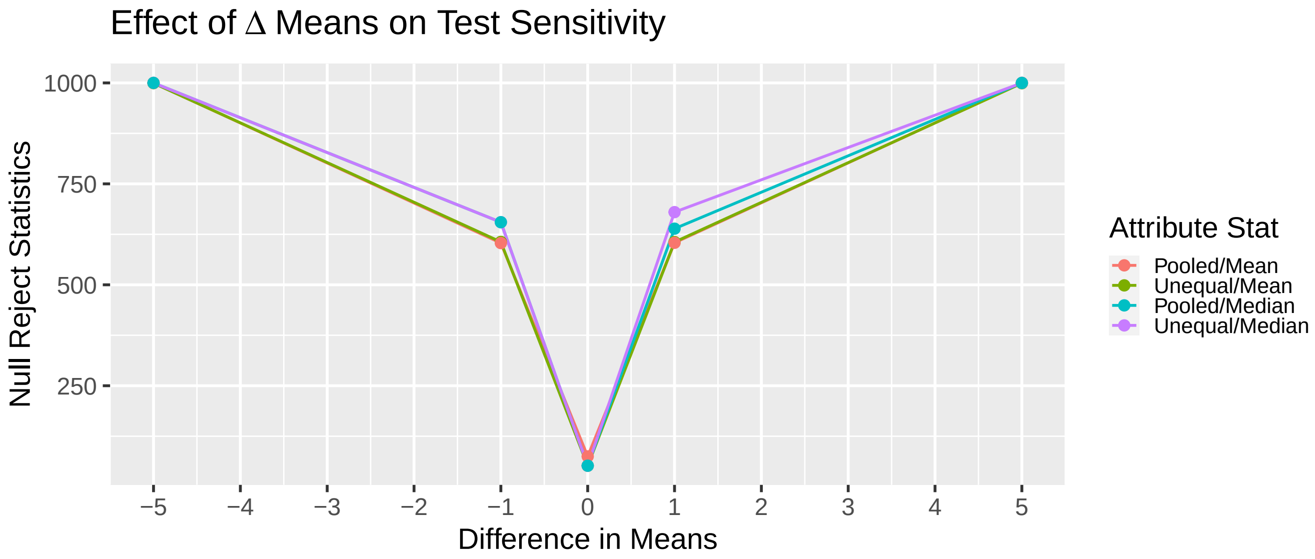
Test Case	μ_1	σ_1^2	n_1	μ_2	σ_2^2	n_2	# of Rejected Null (Unequal σ^2)	# of Rejected Null (Pooled σ^2)	Difference in Mean
1	0	1	10	5	1	10	1000	1000	-5
2	4	1	10	5	1	10	578	562	-1
3	5	1	10	5	1	10	54	39	0
4	6	1	10	5	1	10	527	546	1
5	10	1	10	5	1	10	1000	1000	5
6	0	4	10	5	1	10	1000	1000	-5
7	4	4	10	5	1	10	268	259	-1
8	5	4	10	5	1	10	43	55	0
9	6	4	10	5	1	10	271	265	1
10	10	4	10	5	1	10	1000	1000	5
11	0	9	10	5	1	10	992	995	-5
12	4	9	10	5	1	10	126	156	-1
13	5	9	10	5	1	10	53	50	0
14	6	9	10	5	1	10	133	200	1
15	10	9	10	5	1	10	992	999	5
16	0	1	30	5	1	10	1000	1000	-5
17	4	1	30	5	1	10	723	776	-1
18	5	1	30	5	1	10	49	52	0
19	6	1	30	5	1	10	731	740	1
20	10	1	30	5	1	10	1000	1000	5
21	0	4	30	5	1	10	1000	1000	-5
22	4	4	30	5	1	10	523	260	-1
23	5	4	30	5	1	10	43	6	0
24	6	4	30	5	1	10	549	281	1
25	10	4	30	5	1	10	1000	1000	5
26	0	9	30	5	1	10	1000	1000	-5
27	4	9	30	5	1	10	338	82	-1
28	5	9	30	5	1	10	57	3	0
29	6	9	30	5	1	10	349	87	1
30	10	9	30	5	1	10	1000	1000	5
31	0	1	70	5	1	10	1000	1000	-5
32	4	1	70	5	1	10	784	812	-1
33	5	1	70	5	1	10	55	54	0
34	6	1	70	5	1	10	776	842	1
35	10	1	70	5	1	10	1000	1000	5
36	0	4	70	5	1	10	1000	1000	-5
37	4	4	70	5	1	10	648	242	-1
38	5	4	70	5	1	10	70	4	0
39	6	4	70	5	1	10	668	244	1
40	10	4	70	5	1	10	1000	1000	5
41	0	9	70	5	1	10	1000	1000	-5
42	4	9	70	5	1	10	526	29	-1
43	5	9	70	5	1	10	44	1	0
44	6	9	70	5	1	10	530	31	1
45	10	9	70	5	1	10	1000	1000	5
46	0	1	10	5	1	30	1000	1000	-5
47	4	1	10	5	1	30	728	734	-1
48	5	1	10	5	1	30	47	53	0
49	6	1	10	5	1	30	729	753	1
50	10	1	10	5	1	30	1000	1000	5
51	0	4	10	5	1	30	1000	1000	-5

Test Case	μ_1	σ_1^2	n_1	μ_2	σ_2^2	n_2	# of Rejected Null (Unequal σ^2)	# of Rejected Null (Pooled σ^2)	Difference in Mean
52	4	4	10	5	1	30	301	534	-1
53	5	4	10	5	1	30	50	161	0
54	6	4	10	5	1	30	273	523	1
55	10	4	10	5	1	30	1000	1000	5
56	0	9	10	5	1	30	996	1000	-5
57	4	9	10	5	1	30	165	417	-1
58	5	9	10	5	1	30	53	203	0
59	6	9	10	5	1	30	158	404	1
60	10	9	10	5	1	30	995	1000	5
61	0	1	30	5	1	30	1000	1000	-5
62	4	1	30	5	1	30	967	968	-1
63	5	1	30	5	1	30	62	46	0
64	6	1	30	5	1	30	967	963	1
65	10	1	30	5	1	30	1000	1000	5
66	0	4	30	5	1	30	1000	1000	-5
67	4	4	30	5	1	30	655	696	-1
68	5	4	30	5	1	30	62	65	0
69	6	4	30	5	1	30	680	665	1
70	10	4	30	5	1	30	1000	1000	5
71	0	9	30	5	1	30	1000	1000	-5
72	4	9	30	5	1	30	391	383	-1
73	5	9	30	5	1	30	54	56	0
74	6	9	30	5	1	30	374	397	1
75	10	9	30	5	1	30	1000	1000	5
76	0	1	70	5	1	30	1000	1000	-5
77	4	1	70	5	1	30	997	997	-1
78	5	1	70	5	1	30	45	42	0
79	6	1	70	5	1	30	997	994	1
80	10	1	70	5	1	30	1000	1000	5
81	0	4	70	5	1	30	1000	1000	-5
82	4	4	70	5	1	30	902	763	-1
83	5	4	70	5	1	30	48	8	0
84	6	4	70	5	1	30	910	764	1
85	10	4	70	5	1	30	1000	1000	5
86	0	9	70	5	1	30	1000	1000	-5
87	4	9	70	5	1	30	700	396	-1
88	5	9	70	5	1	30	46	7	0
89	6	9	70	5	1	30	681	422	1
90	10	9	70	5	1	30	1000	1000	5
91	0	1	10	5	1	70	1000	1000	-5
92	4	1	10	5	1	70	788	825	-1
93	5	1	10	5	1	70	52	47	0
94	6	1	10	5	1	70	756	838	1
95	10	1	10	5	1	70	1000	1000	5
96	0	4	10	5	1	70	1000	1000	-5
97	4	4	10	5	1	70	285	628	-1
98	5	4	10	5	1	70	57	242	0
99	6	4	10	5	1	70	297	623	1
100	10	4	10	5	1	70	1000	1000	5
101	0	9	10	5	1	70	996	1000	-5
102	4	9	10	5	1	70	163	544	-1
103	5	9	10	5	1	70	60	323	0
104	6	9	10	5	1	70	160	544	1
105	10	9	10	5	1	70	992	1000	5
106	0	1	30	5	1	70	1000	1000	-5
107	4	1	30	5	1	70	992	993	-1

Test Case	μ_1	σ_1^2	n_1	μ_2	σ_2^2	n_2	# of Rejected Null (Unequal σ^2)	# of Rejected Null (Pooled σ^2)	Difference in Mean
108	5	1	30	5	1	70	40	47	0
109	6	1	30	5	1	70	995	991	1
110	10	1	30	5	1	70	1000	1000	5
111	0	4	30	5	1	70	1000	1000	-5
112	4	4	30	5	1	70	728	872	-1
113	5	4	30	5	1	70	45	121	0
114	6	4	30	5	1	70	714	872	1
115	10	4	30	5	1	70	1000	1000	5
116	0	9	30	5	1	70	1000	1000	-5
117	4	9	30	5	1	70	391	655	-1
118	5	9	30	5	1	70	60	186	0
119	6	9	30	5	1	70	423	639	1
120	10	9	30	5	1	70	1000	1000	5
121	0	1	70	5	1	70	1000	1000	-5
122	4	1	70	5	1	70	1000	1000	-1
123	5	1	70	5	1	70	58	45	0
124	6	1	70	5	1	70	999	1000	1
125	10	1	70	5	1	70	1000	1000	5
126	0	4	70	5	1	70	1000	1000	-5
127	4	4	70	5	1	70	958	969	-1
128	5	4	70	5	1	70	45	53	0
129	6	4	70	5	1	70	951	948	1
130	10	4	70	5	1	70	1000	1000	5
131	0	9	70	5	1	70	1000	1000	-5
132	4	9	70	5	1	70	742	738	-1
133	5	9	70	5	1	70	50	57	0
134	6	9	70	5	1	70	760	732	1
135	10	9	70	5	1	70	1000	1000	5

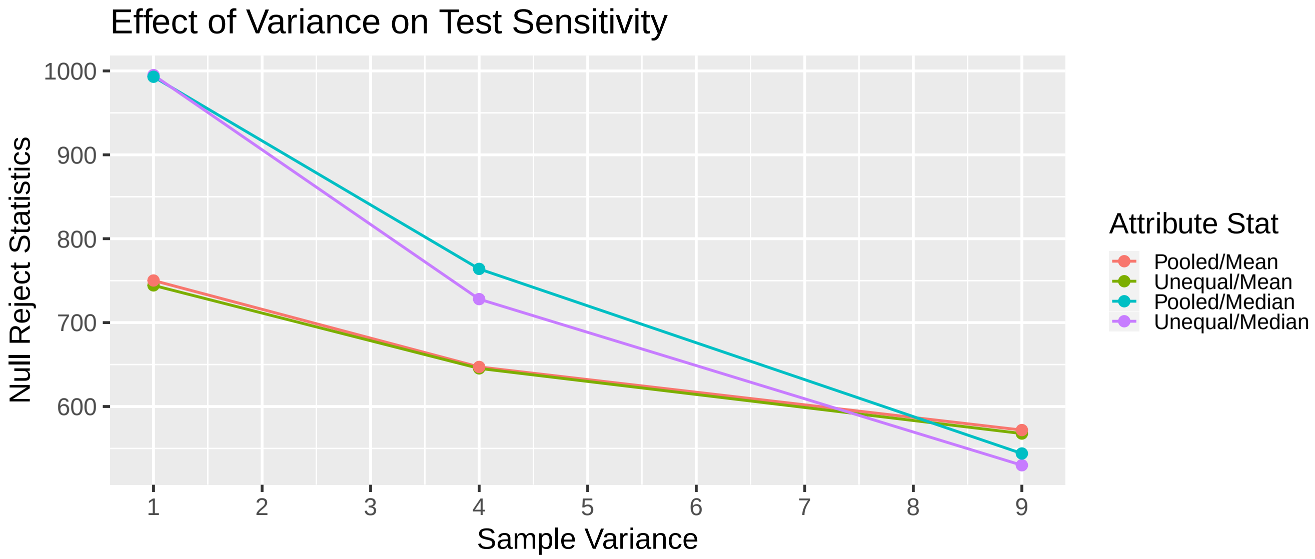
Appendix IV: Part II Summary

Item A



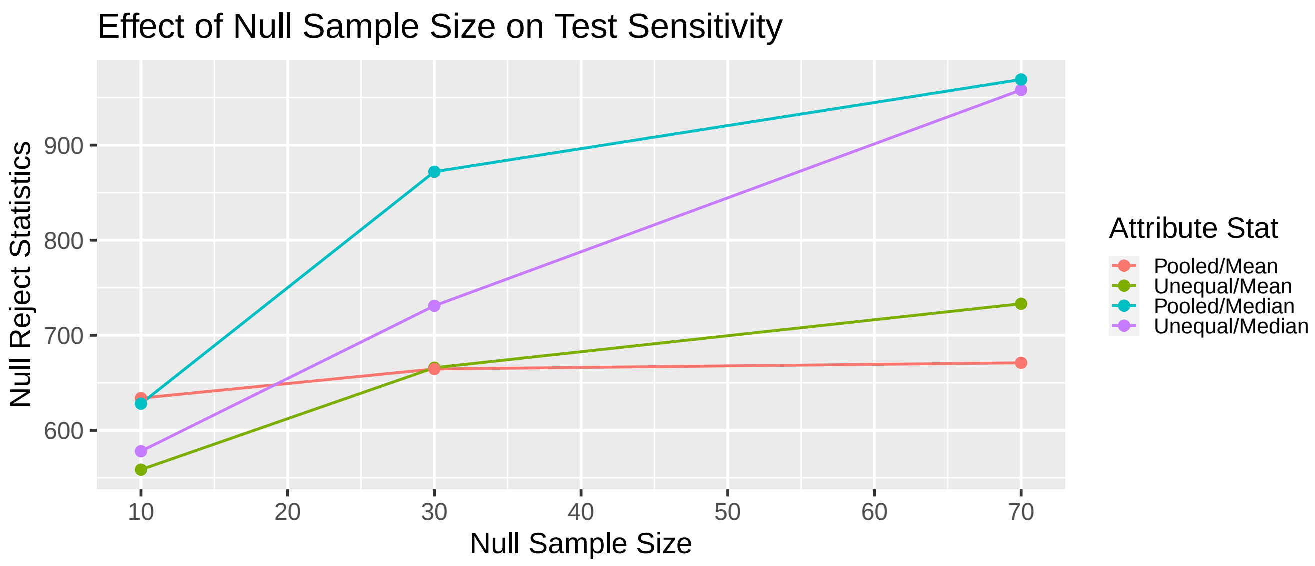
Δ Means	Average # Rejects (Unequal)	Median # Rejects (Unequal)	Average # Rejects (Pooled)	Median # Rejects (Pooled)
-5	999.40741	1000	999.81481	1000
-1	606.18519	655	603.33333	655
0	51.92593	52	75.03704	52
1	605.85185	680	604.00000	639
5	999.22222	1000	999.96296	1000

Item B



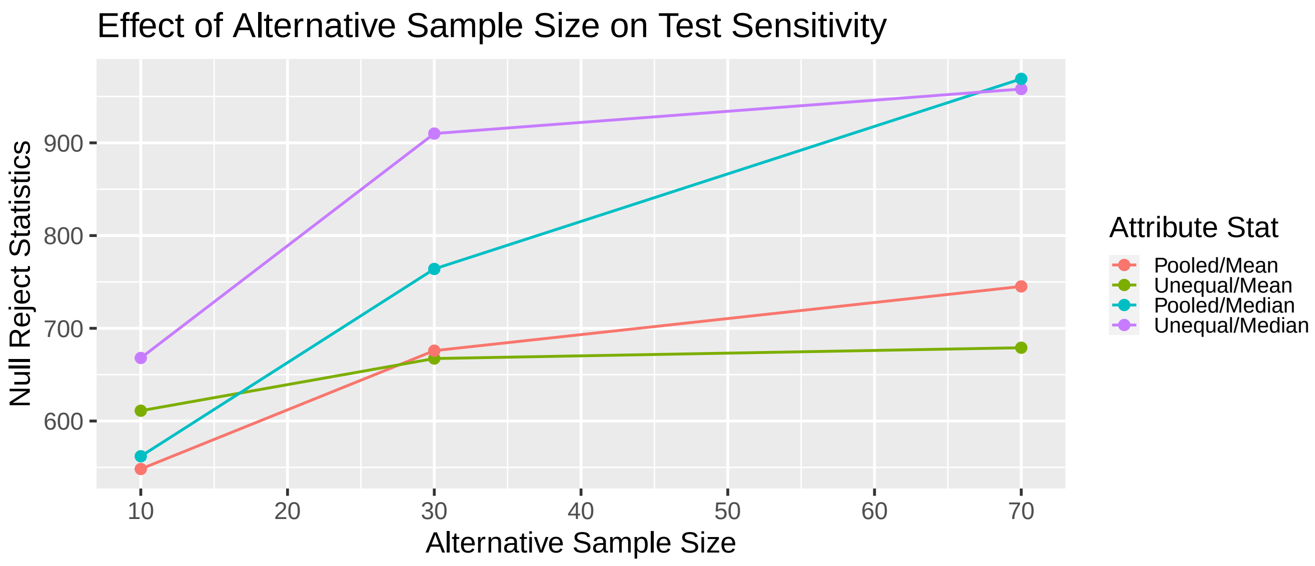
Sample Variance	Average # Rejects (Unequal)	Median # Rejects (Unequal)	Average # Rejects (Pooled)	Median # Rejects (Pooled)
1	744.3556	995	750.2000	993
4	645.4222	728	647.1778	764
9	567.7778	530	571.9111	544

Item C



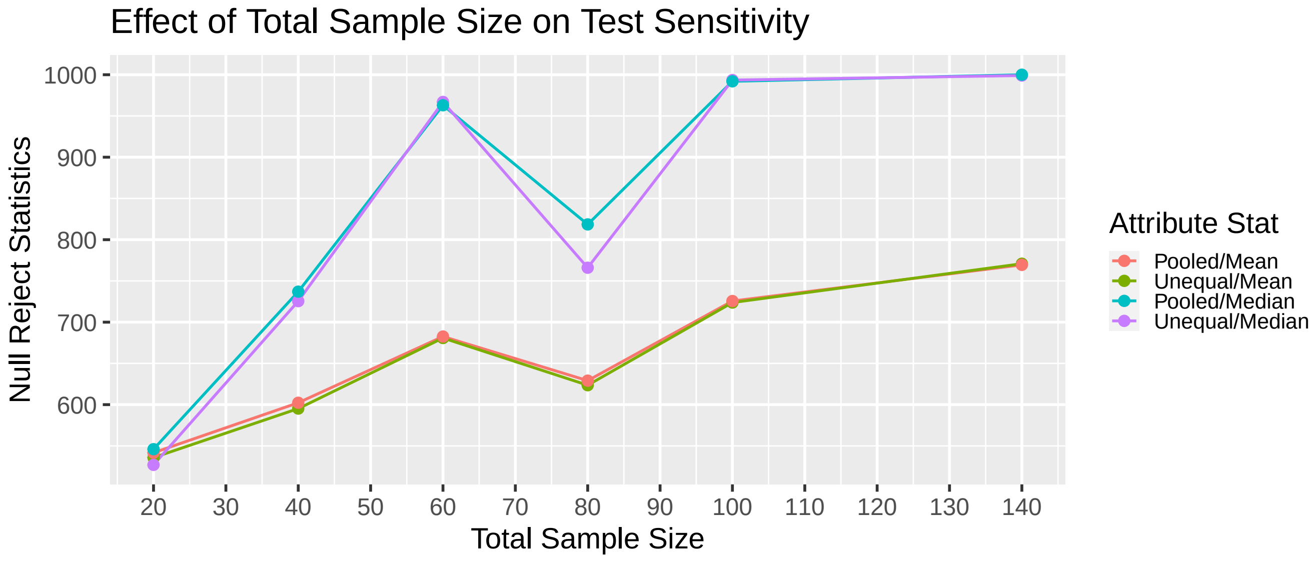
Null Sample Size	Average # Rejects (Unequal)	Median # Rejects (Unequal)	Average # Rejects (Pooled)	Median # Rejects (Pooled)
10	558.6222	578	633.8222	628
30	665.8222	731	664.4889	872
70	733.1111	958	670.9778	969

Item D



Alternative Sample Size	Average # Rejects (Unequal)	Median # Rejects (Unequal)	Average # Rejects (Pooled)	Median # Rejects (Pooled)
10	611.1111	668	548.2667	562
30	667.4000	910	675.8667	764
70	679.0444	958	745.1556	969

Item E



Combined Sample Size	Average # Rejects (Unequal)	Median # Rejects (Unequal)	Average # Rejects (Pooled)	Median # Rejects (Pooled)
20	535.8000	527.0	541.7333	546.0
40	595.2333	725.5	602.3000	737.0
60	680.8000	967.0	682.6000	963.0
80	623.5667	766.0	629.1000	818.5
100	723.8000	993.5	725.6333	992.0
140	770.8667	999.0	769.4667	1000.0

Appendix V: Code

```
knitr::opts_chunk$set(echo = TRUE)

#Use required packages
library(tidyverse) #for plots and data manipulation
library(cowplot) #aligning plots
library(gridExtra)
library(scales)

df_data <- read_csv("framingham_data.csv") # Read in data
df_data$index <- seq(nrow(df_data)) # Add an index column

#df_data %>% summary # Summarize Data

# Split data into smoker and nonsmoker
df_smoker <- df_data %>% filter(currentSmoker == 1)
df_nonsmoker <- df_data %>% filter(currentSmoker == 0)

#Create a sample variance function to ensure proper calculation
sample_variance <- function(x, sampling = TRUE){
    if (sampling == TRUE){
        sum((x - mean(x))^2) / (length(x) - 1)
    } else if(sampling == FALSE) {
        sum((x - mean(x))^2) / (length(x))
    }
}
#Create pooled sample variance function
f_pooled_variance <- function(x, y){
    ((length(x) - 1) * sample_variance(x) +
     (length(y) - 1) * sample_variance(y)) /
    (length(x) + length(y) - 2)
}

# Skewness function
skew_function <- function(x){
    mean((x - mean(x))^3) / sqrt(sample_variance(x))^3
}

# kurtosis function
kurt_function <- function(x){
    mean((x - mean(x))^4) / sqrt(sample_variance(x))^4
}

# Create a Satterthawaite Approximation Function

satterth <- function(s1, s2, n1, n2){
    term1 <- s1/n1
    term2 <- s2/n2
    nu <- (term1 + term2)^2 / ((term1^2/(n1 - 1)) + (term2^2/(n2 - 1)))
    return(floor(nu))
}


#Plot and compare split data

#options(repr.plot.width = 6, repr.plot.height = 4, repr.plot.res = 150)
```

```r
plot_colors <- c("#001427","#708d81","#f4d58d","#bf0603","#8d0801")
y_limits <- c(0, 0.0225)

total_data <- ggplot(df_data) + geom_density(aes(sysBP),
                                             fill = plot_colors[1],
                                             alpha = 0.6) +
                       ylim(y_limits) + ylab("Density") + xlab("")

sep_data <- ggplot() + geom_density(data = df_smoker, aes(sysBP),
                                    fill = plot_colors[3], alpha = 0.6) +
                       geom_density(data = df_nonsmoker, aes(sysBP),
                                    fill = plot_colors[5], alpha = 0.6) +
                       ylim(y_limits) + ylab("") + xlab("Systolic Blood Pressure")

plot_3 <- ggplot() + geom_density(data = df_smoker, aes(sysBP,
                                    fill = plot_colors[3]), alpha = 0.5) +
                       geom_density(data = df_data, aes(sysBP,
                                    fill = plot_colors[1]), alpha = 0.5) +
                       geom_density(data = df_nonsmoker, aes(sysBP,
                                    fill = plot_colors[5]), alpha = 0.5) +
                       ylim(y_limits) + ylab("") + xlab("") +
                       scale_fill_manual("",
                                         values = plot_colors[c(1, 5, 3)],
                                         labels = c("Total", "Non Smoker", "Smoker")) +
                       theme(legend.position = c(0.8, 0.9),
                             legend.text = element_text(size = 6),
                             legend.key.height = unit(0.25, 'cm'),
                             legend.key.width = unit(0.25, 'cm'))

#plot_grid(total_data, sep_data, plot_3, align = 'vh',
          #hjust = -1,nrow = 2, ncol = 2)

data_kurtosis <- kurt_function(df_data$sysBP)
data_skew <- skew_function(df_data$sysBP)
data_IQR <- as.numeric(quantile(df_data$sysBP, probs = 0.75)) -
       as.numeric(quantile(df_data$sysBP, probs = 0.25))
data_MAD <- median(abs(df_data$sysBP - median(df_data$sysBP)))
data_samVar <- sample_variance(df_data$sysBP)

eIQR <- data_IQR / 1.35
eMAD <- data_MAD / 0.675

# Q-Q Plot

data_qqplot <-
ggplot(df_data, aes(sample = sysBP)) +
stat_qq(shape = 1) + stat_qq_line() +
ggtitle("Normal Q-Q Plot for Blood Pressure Data") +
xlab("Theoretical Quantiles") +
ylab("Sample Quantiles")

# Common values for analysis

alpha <- 0.05

mu_smoker <- mean(df_smoker$sysBP)
var_smoker <- sample_variance(df_smoker$sysBP)
```

```
n_smoker <- length(df_smoker$sysBP)

mu_nonsmoker <- mean(df_nonsmoker$sysBP)
var_nonsmoker <- sample_variance(df_nonsmoker$sysBP)
n_nonsmoker <- length(df_nonsmoker$sysBP)

# Two Sample T-test - Pooled Sample Variance - P-value

dof_1 <- (n_smoker + n_nonsmoker - 2)

p_sample_var_1 <- f_pooled_variance(df_smoker$sysBP,
                                    df_nonsmoker$sysBP)
p_sample_var_w <- sqrt(p_sample_var_1/n_smoker + p_sample_var_1/n_nonsmoker)

t_obs_1 <- (mu_smoker - mu_nonsmoker) / (p_sample_var_w )

t_stat_1 <- qt(alpha / 2, dof_1)

p_value_obs_1 <- dt(t_obs_1, dof_1)

#Two Sample T-test - Difference Variance Sample Variance - P-value

dof_2 <- satterth(var_smoker, var_nonsmoker, n_smoker, n_nonsmoker)

np_sample_var_2 <- (var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker)

t_obs_2 <- (mu_smoker - mu_nonsmoker) / (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker))

t_stat_2 <- qt(alpha / 2, dof_2)

p_value_obs_2 <- dt(t_obs_2, dof_2)

# Confidence Limits

diff_mu <- mu_smoker - mu_nonsmoker

#Pooled Sample variance

CL_pooled <- t_stat_1 * p_sample_var_w

#Non pooled Sample variance

CL_nonpooled <- t_stat_2 * (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker))

CI_pooled <- round(c(diff_mu + CL_pooled, diff_mu - CL_pooled), 2)

CI_nonpooled <-round(c(diff_mu + CL_nonpooled, diff_mu - CL_nonpooled), 2)

#Power Calculation assuming delta means is the true delta

cv_lo_p <- qnorm(alpha / 2, 0, sqrt(p_sample_var_1/n_smoker +
                                      p_sample_var_1/n_nonsmoker))
cv_hi_p <- qnorm(1 - alpha / 2, 0, sqrt(p_sample_var_1/n_smoker +
                                          p_sample_var_1/n_nonsmoker))

power1 <- pnorm(cv_lo_p, (mu_smoker - mu_nonsmoker),
                sqrt(p_sample_var_1/n_smoker + p_sample_var_1/n_nonsmoker))
power2 <- 1 - pnorm(cv_hi_p, (mu_smoker - mu_nonsmoker),
                    sqrt(p_sample_var_1/n_smoker + p_sample_var_1/n_nonsmoker))

power_pooled <- sum(power1, power2)

cv_lo_non <- qnorm(alpha / 2, 0,
               (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker)))
cv_hi_non <- qnorm(1 - alpha / 2, 0,
               (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker)))

power1 <- pnorm(cv_lo_non, (mu_smoker - mu_nonsmoker),
                (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker)))
power2 <- 1 - pnorm(cv_hi_non, (mu_smoker - mu_nonsmoker),
                    (sqrt(var_nonsmoker/n_smoker + var_nonsmoker/n_nonsmoker)))

power_nonpooled <- sum(power1, power2)


#Part II
#Introduction

#options(repr.plot.width = 12, repr.plot.height = 3, repr.plot.res = 150)
set.seed(100)

null_mean <- 3
alt_means <- c(0, 3, 5)
plot_list <- list()

#plot_colors <- c("#072ac8","#1e96fc","#a2d6f9","#fcf300","#ffc600")

for(i in 1:length(alt_means)){

    sim1 <- rnorm(5000, null_mean, sqrt(1))
    sim2 <- rnorm(5000, alt_means[i], sqrt(1))

    alpha1 <- qnorm(0.025, null_mean, sqrt(1))
    alpha2 <- qnorm(0.975, null_mean, sqrt(1))

    df_set <- tibble("H0" = sim1, "HA" = sim2)

    title_string <- sprintf("Difference in Means %i", (alt_means[i] - null_mean))

    plot_list[[i]]  <-
    ggplot(data = df_set) + geom_density(aes(H0), alpha = 0.5, fill = plot_colors[5]) +
                            geom_area(
                                aes(x = stage(H0, after_scale = oob_censor(x, c(-Inf, alpha1)
                                                                          )
                                             )
                                   ),
                                stat = "density", fill = plot_colors[1]
                                 ) +
                            geom_area(
                                aes(x = stage(H0, after_scale = oob_censor(x, c(alpha2, Inf)
                                                                          )
                                             )
                                   ),
                                stat = "density", fill = plot_colors[1]
                                     ) +
                            geom_density(aes(HA), alpha = 0.5) +
                            geom_area(
                                aes(x = stage(HA, after_scale = oob_censor(x, c(alpha1, alpha2)
                                                                          )
                                             )
                                   ),
                                stat = "density", fill = plot_colors[2], alpha = 0.5
                                     ) +
                            xlim(-2, 8) + xlab("") + ylab("") + ggtitle(title_string) +
                            theme(text = element_text(size = 8))

                                }

do.call(grid.arrange, c(plot_list, ncol = 3, heights = 0.5))
#Part II
set.seed(1)

alpha <- 0.05

#function to take two datasets and run a hypothesis test on them. This outputs the test results.

test_function <- function (x, y, pooled = FALSE){

    mu_1 <- mean(x)
    var_1 <- sample_variance(x, sampling = TRUE)

    mu_2 <- mean(y)
    var_2 <- sample_variance(y, sampling = TRUE)

    #Calculate the pooled sample variance
    pooled_sample <- ((length(x) - 1) * var_1 +
                          (length(y) - 1) * var_2) / (length(x) + length(y) - 2)

    #calculate the observed t statistic
    if (pooled == TRUE){

        cal_sigma <- (sqrt(pooled_sample/length(x) + pooled_sample/length(y)))

        ttest <- (mu_1 - mu_2) / cal_sigma
        dof <- length(x) + length(y) - 2 #Determine degrees of freedom

        } else {

        cal_sigma <- (sqrt(var_1/length(x) + var_2/length(y)))

        ttest <- (mu_1 - mu_2) / cal_sigma
        dof <- satterth(var_1, var_2, length(x), length(y))
    }

    # Determine whether or not the null hypothesis
    # can be rejected (1 = rejected, 0 = not rejected)
    verdict <- !between(ttest, qt(alpha / 2, dof), qt(1 - alpha / 2, dof))

    #Power calculation assuming calculated difference in means is Ha
    cv_lo <- qnorm(alpha / 2, 0, cal_sigma)
    cv_hi <- qnorm(1 - alpha / 2, 0, cal_sigma)

    power1 <- pnorm(cv_lo, (mu_1 - mu_2), cal_sigma)
    power2 <- 1 - pnorm(cv_hi, (mu_1 - mu_2), cal_sigma)
    power <- sum(power1, power2)

    #Return calculated values
    return(c(mu_1, var_1, mu_2, var_2, ttest, cal_sigma, dof, verdict, power))
}


# Simulation study parameters
mu1 <- c(0, 4, 5, 6, 10)
var1 <- c(1, 4, 9)
n1 <- c(10, 30, 70)

mu2 <- 5
var2 <- 1
n2 <- c(10, 30, 70)

sim_test <- function(x_mu, x_var, x_n, y_mu, y_var, y_n, pooled){

    sim_data_results <- matrix(rep(0, 9), ncol = 9)

    for (i in 1:1000){

        #create simulation test sets
        sim_set1 <- rnorm(x_n, x_mu, sqrt(x_var))
        sim_set2 <- rnorm(y_n, y_mu, sqrt(y_var))

        sim_data_results <- rbind(sim_data_results,
                                  test_function(sim_set1, sim_set2, pooled))

        #print(sim_data_results)
    }

    df_sim_data <- data.frame(sim_data_results[2 : nrow(sim_data_results),])
    colnames(df_sim_data) = c("Null Mean", "Null Variance", "Alternate Mean",
                              "Alternate Variance", "T statistic",
                              "Calculated Variance", "DoF", "Null Reject",
                              "Power")
    return(df_sim_data)
}

# All possible combinations of study conditions
df_combo <- expand.grid(mu1, var1, n1, mu2, var2, n2)

# Final table of conditions and results
df_combo2 <- tibble(cbind(1:nrow(df_combo), df_combo,
                          matrix(rep(0, 2 * nrow(df_combo)), ncol = 2)))
colnames(df_combo2) <- c("Test_Case", "mu1", "var1", "n1", "mu2", "var2",
                         "n2", "Test_Results_up", "Test_Results_po")

test_results <- list() # Pooled assumption raw data
test_results2 <- list() # unequal variance assumption raw data

# Filling out raw data and final table with test results
for (i in 1:nrow(df_combo)){
    test_results[[i]] <- do.call(sim_test,
                                 as.list(as.numeric(c(df_combo[i,],
                                                    pooled = FALSE))))
    test_results2[[i]] <- do.call(sim_test,
                                  as.list(as.numeric(c(df_combo[i,],
                                                       pooled = TRUE))))
    df_combo2[i, 8] <- sum(as.data.frame(test_results[i])[,8])
    df_combo2[i, 9] <- sum(as.data.frame(test_results2[i])[,8])
    }

df_combo2 <- df_combo2 %>% mutate(diff = mu1 - mu2)
#df_combo2 %>%  head()

# Some summary statitics to look for relationships in the attributes
av_med_stats_by_diff <-
    df_combo2 %>% group_by(diff) %>% summarise(avg_up = mean(Test_Results_up),
                                               med_up = median(Test_Results_up),
                                               avg_po = mean(Test_Results_po),
                                               med_po = median(Test_Results_po))


av_med_stats_by_var <-
    df_combo2 %>% group_by(var1) %>% summarise(avg_up = mean(Test_Results_up),
                                               med_up = median(Test_Results_up),
                                               avg_po = mean(Test_Results_po),
                                               med_po = median(Test_Results_po))

av_med_stats_by_nulln <-
    df_combo2 %>% group_by(n1) %>% summarise(avg_up = mean(Test_Results_up),
                                             med_up = median(Test_Results_up),
                                             avg_po = mean(Test_Results_po),
                                             med_po = median(Test_Results_po))

av_med_stats_by_altn <-
    df_combo2 %>% group_by(n2) %>% summarise(avg_up = mean(Test_Results_up),
                                             med_up = median(Test_Results_up),
                                             avg_po = mean(Test_Results_po),
                                             med_po = median(Test_Results_po))

av_med_stats_by_comn <-
    df_combo2 %>% mutate(comn = n1 + n2) %>% group_by(comn) %>%
    summarise(avg_up = mean(Test_Results_up),
              med_up = median(Test_Results_up),
              avg_po = mean(Test_Results_po),
              med_po = median(Test_Results_po))

av_med_stats_by_diff %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                         ggplot() + geom_line(aes(x = diff, y = calc, color = stat)) +
                                    geom_point(aes(x = diff, y = calc, color = stat)) +
                                    xlab("Difference in Means") +
                                    ylab("Null Reject Statistics") +
                                    scale_x_continuous(breaks = seq(-5, 5, 1)) +
                                    ggtitle("Effect of" ~ Delta ~ "Means on Test Sensitivity") +
                                    scale_color_discrete(name = "Attribute Stat",
                                                         labels = c("Pooled/Mean",
                                                                    "Unequal/Mean",
                                                                    "Pooled/Median",
                                                                    "Unequal/Median")) +
                                    theme(legend.position = c(0.85, 0.3),
                                          title = element_text(size = 8),
                                          legend.text = element_text(size = 6),
                                          legend.key.height = unit(0.25, 'cm'),
                                          legend.key.width = unit(0.25, 'cm'))

pow_tab_9a <- test_results[[9]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)
pow_tab_9b <- test_results2[[9]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)
pow_tab_129a <- test_results[[129]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)
pow_tab_129b <- test_results2[[129]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)

d <- cbind(pow_tab_9a, pow_tab_9b, pow_tab_129a, pow_tab_129b)[2:5,]

colnames(d) = c("Test #9 (pooled)", "Test #9 (unequal)", "Test #129 (pooled)", "Test #129 (unequal)")
rownames(d) = c("1st Quartile", "Median", "Mean", "3rd Quartile")

knitr::kable(d)

# Plotting
options(repr.plot.width = 6, repr.plot.height = 4, repr.plot.res = 150)

plot_grid(total_data, sep_data, plot_3, align = 'vh',
          hjust = -1,nrow = 1, ncol = 3, labels = c("A", "B", "C"))
data_qqplot + theme_bw()
d1 <- matrix(c(mean(df_data$sysBP), sample_variance(df_data$sysBP),
              mu_smoker, var_smoker,
              mu_nonsmoker, var_nonsmoker,
              diff_mu, p_sample_var_1, p_sample_var_w,
              np_sample_var_2, CI_pooled, CI_nonpooled), ncol = 1)


rownames(d1) <- c("Data Mean", "Data Sample Variance", "Smoker Mean",
                 "Smoker Variance",
                 "Nonsmoker Mean", "Nonsmoker Variance",
                 "Difference in Mean", "Pooled Sample Varance",
                 "T-Statistic Std Deviation (Pooled)",
                 "T-Statistic Std Deviation (Unequal)",
                 "CI Pooled Lower", "CI Pooled Upper",
                 "CI Nonpooled Lower", "CI Nonpooled Upper")

d1 %>% knitr::kable(col.names = c("Attribute Value"))
knitr::kable(df_combo2, col.names = c("Test Case",
                                      "$\\mu_1$",
                                      "$\\sigma_1^2$",
                                      "$n_1$",
                                      "$\\mu_2$",
                                      "$\\sigma_2^2$",
                                      "$n_2$",
                                      "# of Rejected Null (Unequal $\\sigma^{2}$)",
                                      "# of Rejected Null (Pooled $\\sigma^{2}$)",
                                      "Difference in Mean"),
             escape = FALSE)
av_med_stats_by_diff %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                         ggplot() + geom_line(aes(x = diff, y = calc, color = stat)) +
                                    geom_point(aes(x = diff, y = calc, color = stat)) +
                                    xlab("Difference in Means") +
                                    ylab("Null Reject Statistics") +
                                    ggtitle("Effect of" ~ Delta ~ "Means on Test Sensitivity") +
                                    scale_x_continuous(breaks = seq(-5, 5, 1)) +
                                    scale_color_discrete(name = "Attribute Stat",
                                                         labels = c("Pooled/Mean",
                                                                    "Unequal/Mean",
                                                                    "Pooled/Median",
                                                                    "Unequal/Median")
                                                         ) +
                                    theme(#legend.position = c(0.85, 0.3),
                                          legend.text = element_text(size = 8),
                                          legend.key.height = unit(0.25, 'cm'),
                                          legend.key.width = unit(0.4, 'cm'))

knitr::kable(av_med_stats_by_diff, col.names = c("$\\Delta$ Means",
                                                 "Average # Rejects (Unequal)",
                                                 "Median # Rejects (Unequal)",
                                                 "Average # Rejects (Pooled)",
                                                 "Median # Rejects (Pooled)"),
             escape = FALSE)
av_med_stats_by_var %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                         ggplot() + geom_line(aes(x = var1, y = calc, color = stat)) +
                                    geom_point(aes(x = var1, y = calc, color = stat)) +
                                    xlab("Sample Variance") +
                                    ylab("Null Reject Statistics") +
                                    ggtitle("Effect of Variance on Test Sensitivity") +
                                    scale_x_continuous(breaks = seq(-5, 10, 1)) +
                                    theme(#legend.position = c(0.85, 0.8),
                                          legend.text = element_text(size = 8),
                                          legend.key.height = unit(0.25, 'cm'),
                                          legend.key.width = unit(0.4, 'cm')) +
                                    scale_color_discrete(name = "Attribute Stat",
                                                         labels = c("Pooled/Mean",
                                                                    "Unequal/Mean",
                                                                    "Pooled/Median",
                                                                    "Unequal/Median")
                                                         )

knitr::kable(av_med_stats_by_var, col.names = c("Sample Variance",
                                                 "Average # Rejects (Unequal)",
                                                 "Median # Rejects (Unequal)",
                                                 "Average # Rejects (Pooled)",
                                                 "Median # Rejects (Pooled)"),
             escape = FALSE)

av_med_stats_by_nulln %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                          ggplot() + geom_line(aes(x = n1, y = calc, color = stat)) +
                                     geom_point(aes(x = n1, y = calc, color = stat)) +
                                     xlab("Null Sample Size") +
                                     ylab("Null Reject Statistics") +
                                     ggtitle("Effect of Null Sample Size on Test Sensitivity") +
                                     scale_x_continuous(breaks = seq(10, 80, 10)) +
                                     theme(#legend.position = c(0.85, 0.65),
                                           legend.text = element_text(size = 8),
                                           legend.key.height = unit(0.25, 'cm'),
                                           legend.key.width = unit(0.4, 'cm')) +
                                     scale_color_discrete(name = "Attribute Stat",
                                                         labels = c("Pooled/Mean",
                                                                    "Unequal/Mean",
                                                                    "Pooled/Median",
                                                                    "Unequal/Median"))
```

```
knitr::kable(av_med_stats_by_nulln, col.names = c("Null Sample Size",
                                                 "Average # Rejects (Unequal)",
                                                 "Median # Rejects (Unequal)",
                                                 "Average # Rejects (Pooled)",
                                                 "Median # Rejects (Pooled)"),
             escape = FALSE)

av_med_stats_by_altn %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                          ggplot() + geom_line(aes(x = n2, y = calc, color = stat)) +
                                     geom_point(aes(x = n2, y = calc, color = stat)) +
                                     xlab("Alternative Sample Size") +
                                     ylab("Null Reject Statistics") +
                                     ggtitle("Effect of Alternative Sample Size on Test Sensitivity") +
                                     scale_x_continuous(breaks = seq(10, 80, 10)) +
                                     theme(#legend.position = c(0.85, 0.65),
                                           legend.text = element_text(size = 8),
                                           legend.key.height = unit(0.25, 'cm'),
                                           legend.key.width = unit(0.4, 'cm')) +
                                     scale_color_discrete(name = "Attribute Stat",
                                                          labels = c("Pooled/Mean",
                                                                     "Unequal/Mean",
                                                                     "Pooled/Median",
                                                                     "Unequal/Median"))


knitr::kable(av_med_stats_by_altn, col.names = c("Alternative Sample Size",
                                                 "Average # Rejects (Unequal)",
                                                 "Median # Rejects (Unequal)",
                                                 "Average # Rejects (Pooled)",
                                                 "Median # Rejects (Pooled)"),
             escape = FALSE)


av_med_stats_by_comn %>% gather(key = "stat", value = "calc", avg_up:med_po) %>%
                          ggplot() + geom_line(aes(x = comn, y = calc, color = stat)) +
                                     geom_point(aes(x = comn, y = calc, color = stat)) +
                                     xlab("Total Sample Size") +
                                     ylab("Null Reject Statistics") +
                                     ggtitle("Effect of Total Sample Size on Test Sensitivity") +
                                     scale_x_continuous(breaks = seq(10, 150, 10)) +
                                     theme(#legend.position = c(0.85, 0.7),
                                           legend.text = element_text(size = 8),
                                           legend.key.height = unit(0.25, 'cm'),
                                           legend.key.width = unit(0.4, 'cm')) +
                                     scale_color_discrete(name = "Attribute Stat",
                                                          labels = c("Pooled/Mean",
                                                                     "Unequal/Mean",
                                                                     "Pooled/Median",
                                                                     "Unequal/Median"))

knitr::kable(av_med_stats_by_comn, col.names = c("Combined Sample Size",
                                                 "Average # Rejects (Unequal)",
                                                 "Median # Rejects (Unequal)",
                                                 "Average # Rejects (Pooled)",
                                                 "Median # Rejects (Pooled)"),
             escape = FALSE)
```