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 following:

H0: 1 - 2 = 0

HA: 1 - 2 0

The null hypothesis being there is no difference in the blood pressure between smokers and non smokers. We believe that there is a difference and this paper will prove if we have enough evidence to reject the null hypothesis.

### Normality of Data

Plotting the density of total data (see Appendix I: Figure 1A), we can that it closely follows a normal distribution, Density plots of the split data exhibit similar forms (see Appendix I: Figure 1B). We can also calculate the kurtosis and the skew of the data, which are 3.812, 0.88 for kurtosis and skew respectively, and see that they are quite close to values typically found in normally distributed data. Additionally, the calculated values of sample standard deviation, , and are 22.89, 22.22, and 21.48. The similarities between the three values indicate that the data is not influenced by outliers. Therefore, we will assume the data and the subsequent split data to be normal.

### Statistical Analysis

The goal of this paper is to investigate, through hypothesis testing, whether or not there is enough evidence to reject the null hypothesis. We will conduct the analysis under two conditions; the first assuming equal population variance (using pooled sample variance, ) and the second assuming unequal population variance. The results will be presented and any discrepancies between the two outcomes will be discussed.

In order to begin we must first find the difference in mean blood pressure between smoker and non smokers in this dataset. This was calculated to be -9.1577778. Despite this being a non zero number we ar not sure if there is in fact a difference of that this value was due to chance.

In the first analysis we will be assuming equal variances, thereby allowing us to use pooled sample variance (see Appendix I: Equation 1) which was calculated to be 510 using a degree of freedom value 298. Under these conditions the p value was calculated to be 0.0041 which is smaller than of 0.05 (0.025 for each side of the distribution). In terms of t values, our observed t value of -3.04 is smaller than the t value, -1.97 for a two sided of 0.05.

Additionally, the 95% confidence limits were calculated for the pooled variance assumption. The observed 95% confidence intervals are -15.08 to -3.23 It can be seen that 0 does not fall into the 95% confidence interval in this case. Therefore the null hypothesis can be rejected.

Based on the aforementioned findings, we can reject the null hypothesis in favor of the alternative hypothesis at an level of 0.05. In other words, there is significant evidence to support this group of smokers and non smokers have a different range of systolic blood pressure when using pooled variance.

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

Furthermore, the 95% confidence limit 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 level of 0.05.

We have shown that there is sufficient evidence that the blood pressure between smokers and nonsmokers are 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 smoker and non smokers, we do not know the underlying cause based on the provided data.

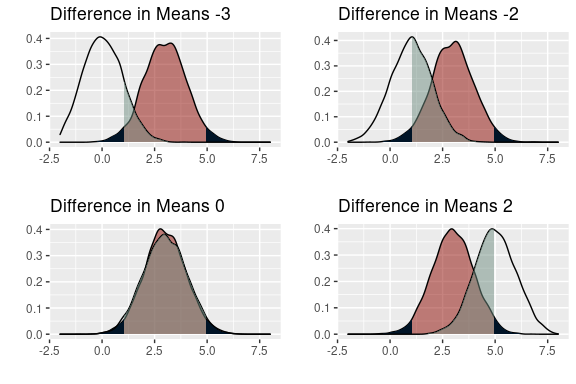
## Part II

### Introduction

In Part I we were able to see how hypothesis testing allowed to find enough evidence to challenge the “status quo” argument that smoking has no affect on the blood pressure of a person. What we found, through a number of robust statistical methods, that the difference in means, calculated to be 9.1577778, was significant to not have arisen by chance. That within an 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 that randomness. All simulated date 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 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 . Power is the complement to that probability (1 - ). The value of 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. The outline distribution represents the “true” alternative distribution. Within that distribution is the shaded light blue region which is equal to . It can be seen that as the difference between the two distributions shrinks the value of increases and the power shrinks. Power reaches minimum and reaches maximum when the two distributions are the same.



### Simulation Study

Below, various 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 unpooled variance, regardless of the actual values of variance.

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

|  |  |
| --- | --- |
| Parameter | Possible Values |
|  | 0, 4, 5, 6, 10 |
|  | 1, 4, 9 |
|  | 10, 30, 70 |
|  | 5 |
|  | 1 |
|  | 10, 30, 70 |

It was noted that the smaller difference between the alternative and null mean values were, the less likely the null hypothesis was rejected. Larger values of variance also reduced the number of rejected null hypothesis. In both cases, the more overlap between the two distributions, the less powerful the test. One way that can increase the power of the t-test is to increase the sample size. For a specific example, we can look at test cases #9 and #129. In both cases the distribution means were 6 and 5 for the null and alternative hypotheses respectively, along with the respective variances of 4 and 1. The only difference between the two cases was the sample size ( = 10 and = 70). This increase in sample size resulted in almost an 4 fold increase in power (286/952 (pooled) or 265/948 (unpooled) rejected nulls). Test cases 69 and 114 use different combinations of, but lower than 70, sample sizes than the aforementioned two. While they exhibited increased signs of power, they did not get as high as test case 129. It stands to reason that large variances and small deltas in mean can be mitigated by increasing the sample size accordingly.

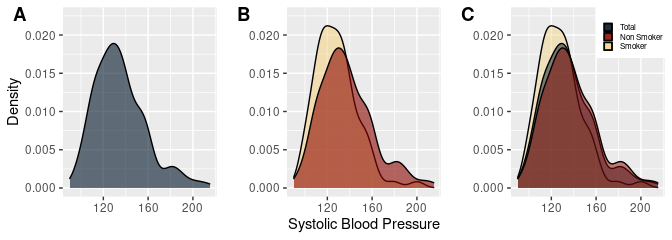
Below are summaries of (1 - ) calculations for test cases #9 and #129:

|  |  |  |
| --- | --- | --- |
|  | Test Case #9 | Test Case #129 |
| Min | 0.0500009 | 0.0707365 |
| 1st Quantile | 0.1169414 | 0.8610788 |
| Median | 0.2979222 | 0.9652580 |
| Mean | 0.3793437 | 0.8953985 |
| 3rd Quantile | 0.6114560 | 0.9947153 |
| Max | 0.9999994 | 0.9999993 |

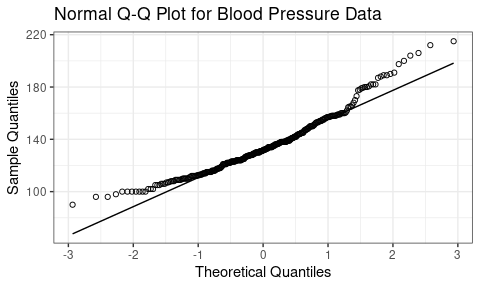
The pattern further…….

## Appendix I: Equations and Figures

## Figures



### Figure 1: Data plots



### Figure 2: Q-Q Plot

## Equations:

### Equation 1: Pooled Sample Variance:

### Equation 2: Observed T Statistic (pooled variance):

### Equation 3: Observed T Statistic (distinct variance):

### Equation 4: Satterthwaite Approximation:

## Appendix II: Part I Results

Data Mean: 134.935  
Data Sample Variance: 524.0852258

Smoker Mean: 128.0666667  
Smoker Variance: 352.2117117

Nonsmoker Mean: 137.2244444  
Nonsmoker Variance: 562.1447123

Difference in mean: -9.1577778

Pooled Sample Variance: 510.0136987  
Nonpooled Sample Variance: 9.9936838

CI Pooled: -15.08, -3.23  
CI Nonpooled: -15.4, -2.91

## Appendix III: Part II Results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Test Case |  |  |  |  |  |  | Test Results Unpooled | Test Results Pooled | 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 |
| 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 |
| 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: 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)  
  
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)  
  
  
t\_obs\_1 <- (mu\_smoker - mu\_nonsmoker) / (sqrt(p\_sample\_var\_1) \* sqrt(1/n\_smoker + 1/n\_nonsmoker))  
  
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 \* (sqrt(p\_sample\_var\_1/n\_smoker + p\_sample\_var\_1/n\_nonsmoker))  
  
#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 = 5, repr.plot.res = 150)  
set.seed(100)  
  
null\_mean <- 3  
alt\_means <- c(0, 1, 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)  
   
 }  
  
do.call(grid.arrange, plot\_list)  
#Part II  
set.seed(1)  
  
alpha <- 0.05  
  
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))  
}  
  
  
  
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){  
   
 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)  
}  
  
# HA: mean = 5, var = 1  
df\_combo <- expand.grid(mu1, var1, n1, mu2, var2, n2)  
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()  
test\_results2 <- list()  
  
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()  
  
pow\_tab\_9 <- test\_results[[9]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)  
pow\_tab\_129 <- test\_results[[129]] %>% pull(Power) %>% summary() %>% unname() %>% matrix(ncol = 1)  
  
d <- cbind(pow\_tab\_9, pow\_tab\_129)  
   
colnames(d) = c("Test Case #9","Test Case #129")  
rownames(d) = c("Min", "1st Quantile", "Median", "Mean", "3rd Quantile", "Max")  
  
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()  
knitr::kable(df\_combo2, col.names = c("Test Case",   
 "$\\mu\_1$",   
 "$\\sigma\_1^2$",   
 "$n\_1$",   
 "$\\mu\_2$",   
 "$\\sigma\_2^2$",   
 "$n\_2$",  
 "Test Results Unpooled",  
 "Test Results Pooled",  
 "Difference in Mean"),   
 escape = FALSE)