ST-502 Final R project

Group 16

Premal Patel

Anand Morlawar

Part 1

The Framingham Heart Study is a population-based, observational cohort study that was initiated by the United States Health Service in 1948 to investigate risk factors for cardiovascular disease. Our project is based on their dataset. Here, we try to determine if there are any significant cardiovascular differences between two groups- namely smokers and nonsmokers, as measured by systolic blood pressure. We are performing 2 different hypothesis tests. First with an assumption that the sample dataset for smokers and nonsmokers are created from distributions with equal variances and second with an assumption that the samples are created with unequal variances.

The dataset we used represents two independent samples of systolic blood pressure (sysBP) for nonsmokers and smokers. We will assume that the data are both random samples from normal distributions with parameters (μ_1, σ_1^2) and (μ_2, σ_2^2) respectively where μ is mean and σ is variance.

Now we perform two-sample t-tests on the data at a significance level α=0.05 for the scenario mentioned above with equal variance. The hypothesis was further required to be tested using 2 ways, the p-values and confidence intervals methods. We started the hypothesis testing; the hypothesis are as follows:

$H_0 \rightarrow \mu_1 = \mu_2$ vs $H_a \rightarrow \mu_1 \neq \mu_2$

Where H_0 is the null hypothesis that both means are equal and H_a is the alternate hypothesis that both means are not equal. The formula for t-test is as given below:

$$t - test = \left[mean_{nonsmokers\ sysBP} - mean_{smokers\ sysBP}\right] / \sqrt{variance_{pooled}\left(\frac{1}{a} + \frac{1}{b}\right)}$$

where a = number of nonsmokers, b = number of smokers and pooled variance is gives as:

$$variance_{pooled} = [(a-1) * \sigma_1^2 + (b-1) *, \sigma_2^2\ ] / (a+b-2)$$

Now we will do the process for the second scenario with unequal variance and the following is the t-test equation:

$$t - test = \left[mean_{nonsmokers\ sysBP} - mean_{smokers\ sysBP}\right] / \sqrt{(\frac{\sigma_1^2}{a} + \frac{\sigma_2^2}{b})}$$

Now considering the significance level 0.05, the degrees of freedom for equal variance scenario is a+b-2=298, and for unequal scenario it is v, formula given in part 2. we produced the results as given below.

	Test statistic	p-value	Rejection region	Conclusion
Equal variance	3.041308	0.00256	1.96	Reject $H_0 \rightarrow \mu_1 = \mu_2$
Unequal variance	3.414188	0.00081	1.97	Reject $H_0 \rightarrow \mu_1 = \mu_2$

We can go further with the unequal variance results, as theoretically both variances are different and then we plotted the QQ plot to find if the data is distributed normally or not. Here we concluded that the data is not distributed normally. So, we can't use t-test and must look for the other options. That's why we have used Mann-Whitney (Wilcox) rank test to test the hypothesis as this test does not require normality assumptions. The results given in this test shows that we fail to reject the null hypothesis so the means of smokers and nonsmokers sysBP is different.

Appendix

Part 1 Code

```
#### Installing required packages if not installed in the system ####
#list.of.packages = c("ggplot2","dplyr","plotly")
#new.packages = list.of.packages[!(list.of.packages %in% installed.packages()[
,"Package"])]
#if(length(new.packages)) install.packages(new.packages)

#### 1.1 ####

#Import the csv while keeping the data in the same file as the R code
data_s = read.csv('framingham_data.csv')

#### 1.2 ####
#Getting the classification on number of non-smokers and smokers.
smoker_no = sum(data_s$currentSmoker == 1)
non_smoker_no = sum(data_s$currentSmoker == 0)

#Indexes for non-smoker and smoker.
smoker_ind = (data_s$currentSmoker == 1)
non_smoker_ind = (data_s$currentSmoker == 0)

############################################################

#Defining dataset for smoker and non-smoker.
#Later, we used in mean and variance calculation.
smoker = data_s$sysBP[smoker_ind]
non_smoker = data_s$sysBP[non_smoker_ind]

#Setting alpha value equals to 0.05
alpha = 0.05

#Computing the variance of non-smoker and smoker.
non_smoker_var = var(non_smoker)
smoker_var = var(smoker)

#getting the total degrees of freedom
df = non_smoker_no+smoker_no-2

#Computing the mean of non-smoker and smoker.
non_smoker_mean = mean(non_smoker)
smoker_mean = mean(smoker)


#Getting the pooled variance
p_var = ((smoker_no-1)*smoker_var + (non_smoker_no-1)*non_smoker_var)/(non_smo
ker_no+smoker_no-2)

#t-stat calculation
```

```
t_stat = (non_smoker_mean - smoker_mean)/(sqrt(p_var)*sqrt(1/smoker_no + 1/non
_smoker_no))

#checking whether the value is less than alpha
p_value = 2*pt(t_stat,df,lower.tail = FALSE)
if (p_value > alpha){
  print("Here, we fail to reject the null hypothesis using p-value test while
assuming equal variance")
} else{
  print("Here, we reject the null hypothesis using p-value test while assuming
 equal variance")
}
```

```
## [1] "Here, we reject the null hypothesis using p-value test while assuming
equal variance"
```

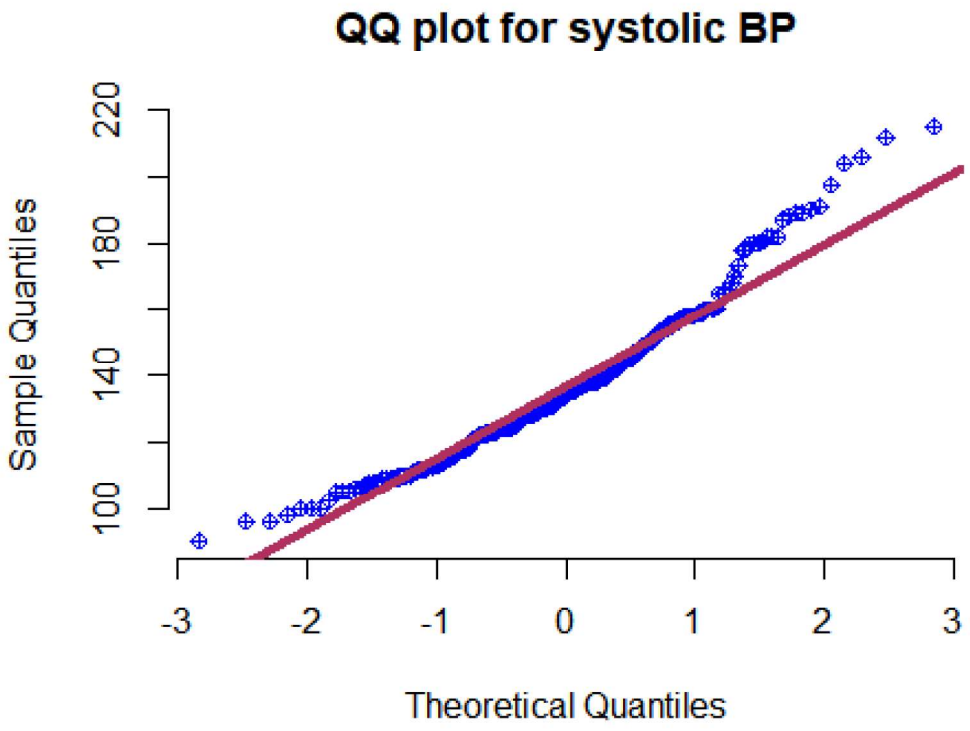
```
#We compute standard deviation and t-critical value & use that in confidence i
nterval
p_sd = sqrt(p_var*(1/smoker_no + 1/non_smoker_no))
t_critical = qt(alpha/2,df,lower.tail = FALSE)
upper = (non_smoker_mean - smoker_mean) + t_critical*p_sd
lower = (non_smoker_mean - smoker_mean) - t_critical*p_sd
if (0>lower & 0<upper){
  print("we fail to reject null hypothesis using confidence interval test whil
e assuming equal variance")

} else{
  print("we reject null hypothesis using confidence interval test while assumi
ng equal variance")
}
```

```
## [1] "we reject null hypothesis using confidence interval test while assumin
g equal variance"
```

```
#Now, we calculate assuming unequal variance (with Satterthwaite approximation
 for the degrees of freedom)
#Calculating Satterthwaite approximation degree of freedom
df_sw = floor((smoker_var/smoker_no + non_smoker_var/non_smoker_no)**2)/(((smo
ker_var/smoker_no)**2)/(smoker_no-1) +  ((non_smoker_var/non_smoker_no)**2)/(n
on_smoker_no-1))

#We first calculate the t value and use it in p-value test
t_stat = (non_smoker_mean - smoker_mean)/(sqrt(smoker_var/smoker_no + non_smok
er_var/non_smoker_no))
p_value = 2*pt(t_stat,df_sw,lower.tail = FALSE)
if (p_value < alpha){
  print("we reject the null hypothesis using p-value test while assuming unequ
al variance")
} else{
  print("we fail to reject the null hypothesis using p-value test while assumi
ng unequal variance")
}
```

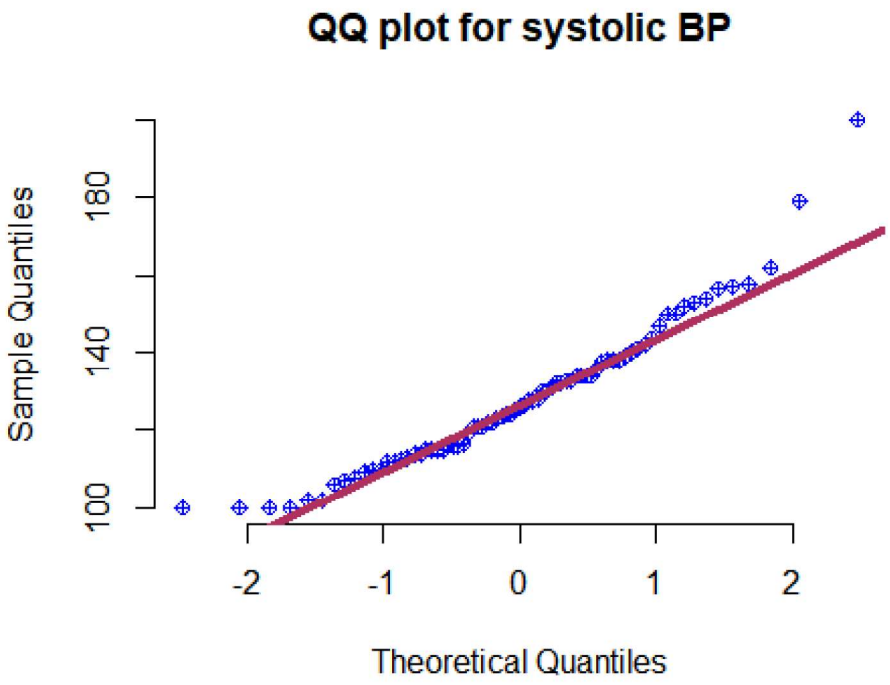
```
## [1] "we reject the null hypothesis using p-value test while assuming unequa
l variance"

#First, we calculate t-critical and standard deviation.
#Later, We use it in upper bound and lower bound calculation
t_critical = qt(alpha/2,df_sw,lower.tail = FALSE)
p_sd = sqrt(smoker_var/smoker_no + non_smoker_var/non_smoker_no)
upper = (non_smoker_mean - smoker_mean) + t_critical*p_sd
lower = (non_smoker_mean - smoker_mean) - t_critical*p_sd


if (0>lower & 0<upper){
  print("we fail to reject null hypothesis using confidence interval test whil
e assuming unequal variance")
} else{
  print("we reject null hypothesis using confidence interval test while assumi
ng unequal variance")
}

## [1] "we reject null hypothesis using confidence interval test while assumin
g unequal variance"

#### 1.3 ####
#QQ-plot for non-smoker
qqnorm(data_s$sysBP[non_smoker_ind], pch = 10, col = "blue", frame = FALSE, ma
in="QQ plot for systolic BP")
qqline(data_s$sysBP[non_smoker_ind],lwd = 4, col = "maroon")
```

1.1 QQ-Plot: Non-Smoking Quantiles vs Normal Quantiles




```
#QQ-plot for smoker
qqnorm(data_s$sysBP[smoker_ind], pch = 10, col = "blue", frame = FALSE, main="
QQ plot for systolic BP")
qqline(data_s$sysBP[smoker_ind],lwd = 4, col = "maroon")
```

1.2 QQ-Plot: Smoking Quantiles vs Normal Quantiles




```
#### 1.4 ####
#From QQ-plot, we conclude that the data is not normal.
#Hence, we use Mann-Whitney test.
#First, we pool the data in the order of smoker and non smoker.
#As smoker data is minority, we rank the smoker data and used it in wilcox tes
t
df = c(smoker, non_smoker)
rank_smoker = rank(df)[1:smoker_no]
pwilcox(sum(rank_smoker), smoker_no, non_smoker_no)

## [1] 0.9170233

#Since this value is above 0.05, we do not reject the null hypothesis. This is
 different from
# the previous tests assuming normality.
```

Part 2

We started the part 1 of the project with a goal to infer the differences in the means of systolic blood pressure (sysBP) of smokers and nonsmokers. Earlier we performed hypothesis tests for that and here we performed a complete simulation study. In this simulation study we compare the two hypothesis testing procedures from part 1 where we did two sample t-test with the assumption of equal variances and another two-sample t-test with the assumption of unequal variances. To compare the effectiveness of the 2 hypothesis tests, we analyze how both control the type 1 error rate (α) and the power (1-β) (primary criteria for testing the effectiveness.)

$H_0 \rightarrow \mu_1 = \mu_2$ vs $H_a \rightarrow \mu_1 \neq \mu_2$

We take the same hypotheses as we took in part 1. Then we take the 2 samples with mean and variance parameters (μ_1, σ_1^2) and (μ_2, σ_2^2). Then we simulate each sample for 1000 times for each of the following combinations of true variance, sample size, and true means.

1. True variance: consider $\sigma_1^2 = 1,4,9$ and $\sigma_2^2 = 1$.
2. Sample size: consider $n_1 = 10,30,70$ and $n_2 = 10,30,70$.
3. True means equal ($\mu_1 - \mu_2 = 0$) **vs** true mean difference ($\mu_1 - \mu_2$) of -5, -1,1,5.

So here we check for the hypothesis that the means of two samples are equal vs means of two samples are unequal and for different combinations given above for 1000 times. Now we store the number of times we fail to reject the null hypothesis, given the difference of true means is zero than this stored number is type 1 error rate α. When the difference of two means is nonzero, given the difference of true means is zero then it is power (1-β).

The formula for t-test is as given below:

$$t - test = \left[mean_{nonsmokers\ sysBP} - mean_{smokers\ sysBP}\right] / \sqrt{variance_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$variance_{pooled} = \left[(n_1 - 1) * \sigma_1 + (n_2 - 1) * \sigma_2\right] / (n_1 + n_2 - 2)$$

The rejection region (RR) is given as:

$$RR = \{T: |T| > t_{\frac{0.05}{2}}, n_1 + n_2 - 2\}$$

After that we perform t-test for unequal variance as given below:

$$t - test = \left[mean_{nonsmokers\ sysBP} - mean_{smokers\ sysBP}\right] / \sqrt{\left(\frac{\sigma_1}{n_1} + \frac{\sigma_2}{n_2}\right)}$$

The degrees of freedom for unequal variance scenario are as follows:

$$v = \frac{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}{\left(\frac{\left(\frac{\sigma_1^2}{n_1}\right)^2}{n_1 - 1}\right) + \left(\frac{\left(\frac{\sigma_2^2}{n_2}\right)^2}{n_2 - 1}\right)}$$

And the rejection region is $RR = \{T: |T| > t_{\frac{0.05}{2},v}\}$

The produced result of the simulation study is given in the appendix along with the code. We have plotted three graphs with different σ1 values of 1, 2, 3. For each plot, we have equal variance test on left side and unequal variance test on the right side. The different shapes indicate whether it is an alpha value, or the power of test value and different colors indicate different mean differences. Here, we can see that it gives type I error or power on y-axis and (n_1, n_2) on x-axis. Both the tests performed well in controlling the value of alpha and have also obtained higher value of power of test when sample size (n1, n2) has increased. But, as the difference in variance increases, the alpha value for equal variance test increases and power of test value decreases. From this simulated study, we can say that hypothesis test with unequal variance gives a much preferable result by devising a lower value for α (type 1 error) and higher value of power compared to the hypothesis test on samples with equal variance.

Therefore, we can conclude that the numerical simulation study has provided us a clear distinction of t-test with equal variance and t-test with unequal variance. Hence, the result that the appropriate test to infer the differences of mean of systolic blood pressure of smokers and nonsmokers is two sample t-test that assumes unequal variance.

Part 2 Code

```
#### Installing required packages if not installed in the system ####
#list.of.packages = c("ggplot2","dpylr","plotly")
#new.packages = list.of.packages[!(list.of.packages %in% installed.packages()[
,"Package"])]
#if(length(new.packages)) install.packages(new.packages)

#Importing required libraries
library(ggplot2)
library(dplyr)

#defining equal variance test function
Variance_eq_Testing <- function(a,b){
  #Threshold level
  threshold_alpha = 0.05
  a_len = length(a)
  b_len = length(b)
  a_variance = var(a)
  b_variance = var(b)
  mean_a = mean(a)
  mean_b = mean(b)

  #getting degrees of freedom
  degrees_of_freedom = a_len+b_len-2
  #pooled variance
  pool_var = ((a_len-1)*a_variance + (b_len-1)*b_variance)/(degrees_of_freedom
)
  #defining t-statistic
  t_stat = (mean_a-mean_b)/(sqrt(pool_var)*sqrt(1/a_len+1/b_len))
  #function returns TRUE if null hypothesis is rejected & FALSE otherwise
  return(abs(t_stat)>qt(threshold_alpha/2, degrees_of_freedom,  lower.tail = F
ALSE))
}

#defining unequal variance test function
Variance_uneq_Testing <- function(p,q){
  threshold_alpha = 0.05
  mean_p = mean(p)
  mean_q = mean(q)
  p_len = length(p)
  q_len = length(q)
  p_variance = var(p)
  q_variance = var(q)

  #calculating degrees of freedom
  degrees_of_freedom = ((((p_variance/p_len) + (q_variance/q_len))^2) / (((p_v
ariance/p_len)^2/(p_len-1)) + (q_variance/q_len)^2/(q_len-1)))
```

```
  #getting t-stat value
  t_stat = (mean_p - mean_q) / sqrt(p_variance/p_len + q_variance/q_len)

  #this function returns TRUE if null hypothesis is rejected. If not, it retur
ns FALSE
  return(abs(t_stat) > qt(threshold_alpha/2, degrees_of_freedom, lower.tail =
FALSE))
}


#simulate 10000 instances and perform t-test after defining simulation functio
n.
simulation <- function(sample_size1, sample_size2, mean, std_dev, t_test_value
){
  #taking 10000 simulations
  iterations = 10000
  #list for storing status of each simulation(accept or reject)
  list_res = c()
  #iterating over the 1000 instances
  for (i in 1:iterations){
    #simulating a sample from normal distribution with mean difference "mean"
and standard deviation "std_dev"
    sample1 = rnorm(sample_size1, mean, std_dev)
    #simulating a sample from normal distribution with mean 0 and standard dev
iation 1
    sample2 = rnorm(sample_size2, 0, 1)
    #selecting test according to input by user (equal variance vs unequal vari
ance)
    if (t_test_value == "equal"){
      #returning list holding TRUE or FALSE values
      list_res[i] = Variance_eq_Testing(sample1, sample2)
    } else {
      #returning list holding TRUE or FALSE values
      list_res[i] = Variance_uneq_Testing(sample1, sample2)
    }
  }
  #rejection of null hypothesis is indicated by true and FALSE indicates other
wise
  result = length(list_res[list_res == TRUE])/iterations
  #returns the fraction of TRUE values
  return(result)
}

#initializing dataframe to hold the return of simulation function for each com
bination
output <- as.data.frame(matrix(0,1,6))

difference_in_mean <- c(0,-5,5,-1,1)
std_dev <- c(1,2,3)
n <- c(10,30,70)
```

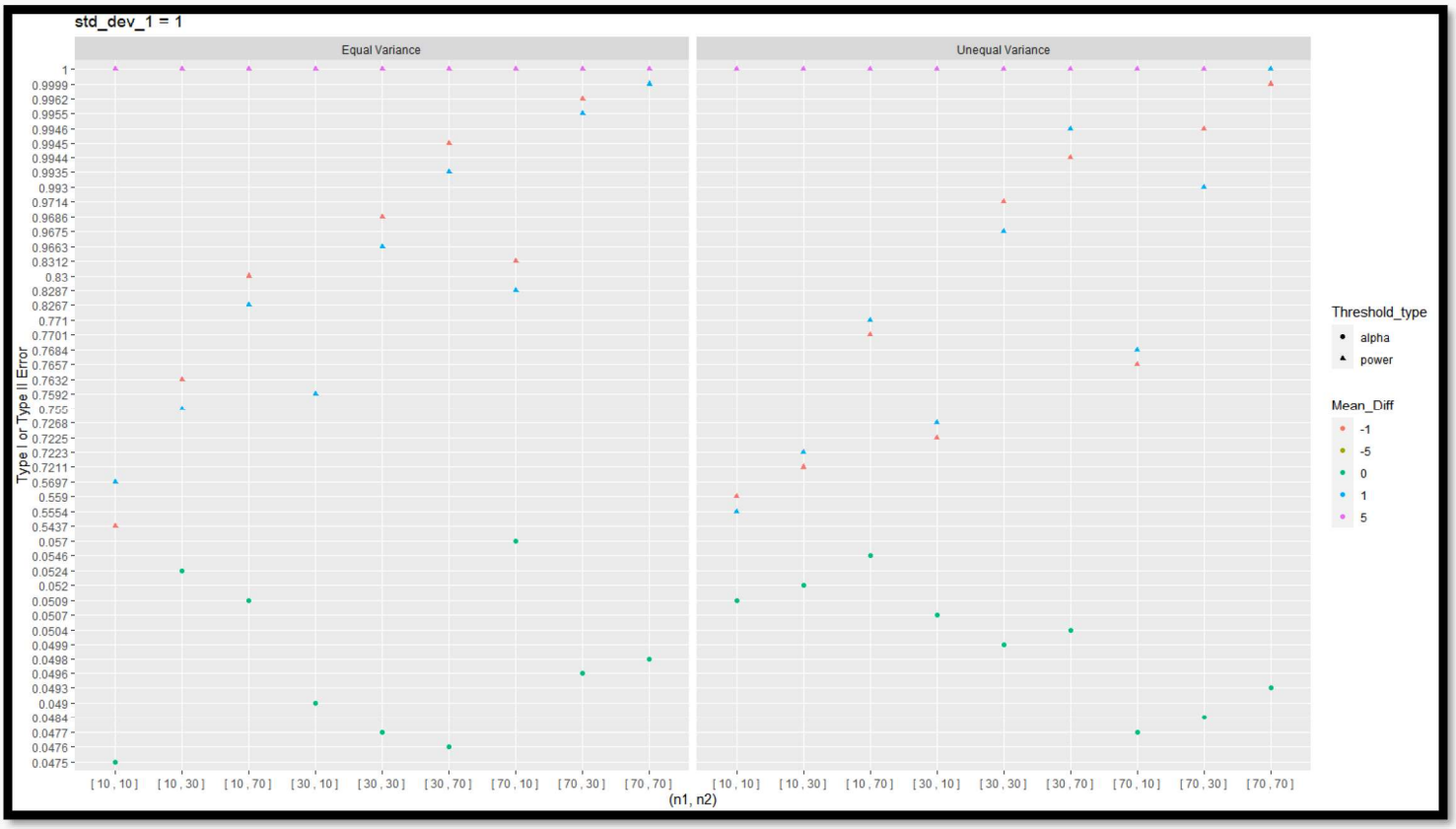
```r
#naming the columns of the output dataframe accordingly
names(output) <- c("type_t_test", "sample_size", "std_dev", "Mean_Diff","Thres
hold_type","threshold_value")

#iterating over all possible permutations and combinations
for (p in std_dev){
  for (q in n){
    for (r in n){
      for (s in difference_in_mean){
        std_dev_val <- paste("[",p,",",1,"]")
        size <- paste("[",q,",",r,"]")
        #performing equal variance test & storing data in dataframe
        output[nrow(output)+1,] <- c("Equal Variance",size, std_dev_val, s, if
else(s == 0, "alpha", "power"), simulation(q,r,s,p,"equal"))
        #performing unequal variance test & storing data in dataframe
        output[nrow(output)+1,] <- c("Unequal Variance", size, std_dev_val, s,
 ifelse(s == 0, "alpha", "power"), simulation(q,r,s,p,"unequal"))
      }
    }
  }
}
#displaying the final output
##removing the initial dummy row
output <- output[-c(1),]
head(output, n = 30)
```

	type_t_test	sample_size	std_dev	Mean_Diff	Threshold_type	threshold_value
2	Equal Variance	[10 , 10]	[1 , 1]	0	alpha	0.0475
3	Unequal Variance	[10 , 10]	[1 , 1]	0	alpha	0.0509
4	Equal Variance	[10 , 10]	[1 , 1]	-5	power	1
5	Unequal Variance	[10 , 10]	[1 , 1]	-5	power	1
6	Equal Variance	[10 , 10]	[1 , 1]	5	power	1
7	Unequal Variance	[10 , 10]	[1 , 1]	5	power	1
8	Equal Variance	[10 , 10]	[1 , 1]	-1	power	0.5437
9	Unequal Variance	[10 , 10]	[1 , 1]	-1	power	0.559
10	Equal Variance	[10 , 10]	[1 , 1]	1	power	0.5697
11	Unequal Variance	[10 , 10]	[1 , 1]	1	power	0.5554
12	Equal Variance	[10 , 30]	[1 , 1]	0	alpha	0.0524
13	Unequal Variance	[10 , 30]	[1 , 1]	0	alpha	0.052
14	Equal Variance	[10 , 30]	[1 , 1]	-5	power	1
15	Unequal Variance	[10 , 30]	[1 , 1]	-5	power	1
16	Equal Variance	[10 , 30]	[1 , 1]	5	power	1
17	Unequal Variance	[10 , 30]	[1 , 1]	5	power	1
18	Equal Variance	[10 , 30]	[1 , 1]	-1	power	0.7632
19	Unequal Variance	[10 , 30]	[1 , 1]	-1	power	0.7211
20	Equal Variance	[10 , 30]	[1 , 1]	1	power	0.755
21	Unequal Variance	[10 , 30]	[1 , 1]	1	power	0.7223
22	Equal Variance	[10 , 70]	[1 , 1]	0	alpha	0.0509
23	Unequal Variance	[10 , 70]	[1 , 1]	0	alpha	0.0546
24	Equal Variance	[10 , 70]	[1 , 1]	-5	power	1
25	Unequal Variance	[10 , 70]	[1 , 1]	-5	power	1
26	Equal Variance	[10 , 70]	[1 , 1]	5	power	1
27	Unequal Variance	[10 , 70]	[1 , 1]	5	power	1
28	Equal Variance	[10 , 70]	[1 , 1]	-1	power	0.83
29	Unequal Variance	[10 , 70]	[1 , 1]	-1	power	0.7701
30	Equal Variance	[10 , 70]	[1 , 1]	1	power	0.8267

```r
#dividing data based on standard deviation of samples
data_one <- output[output$std_dev == "[ 1 , 1 ]",]
data_two <- output[output$std_dev == "[ 2 , 1 ]",]
data_three <- output[output$std_dev == "[ 3 , 1 ]",]
```

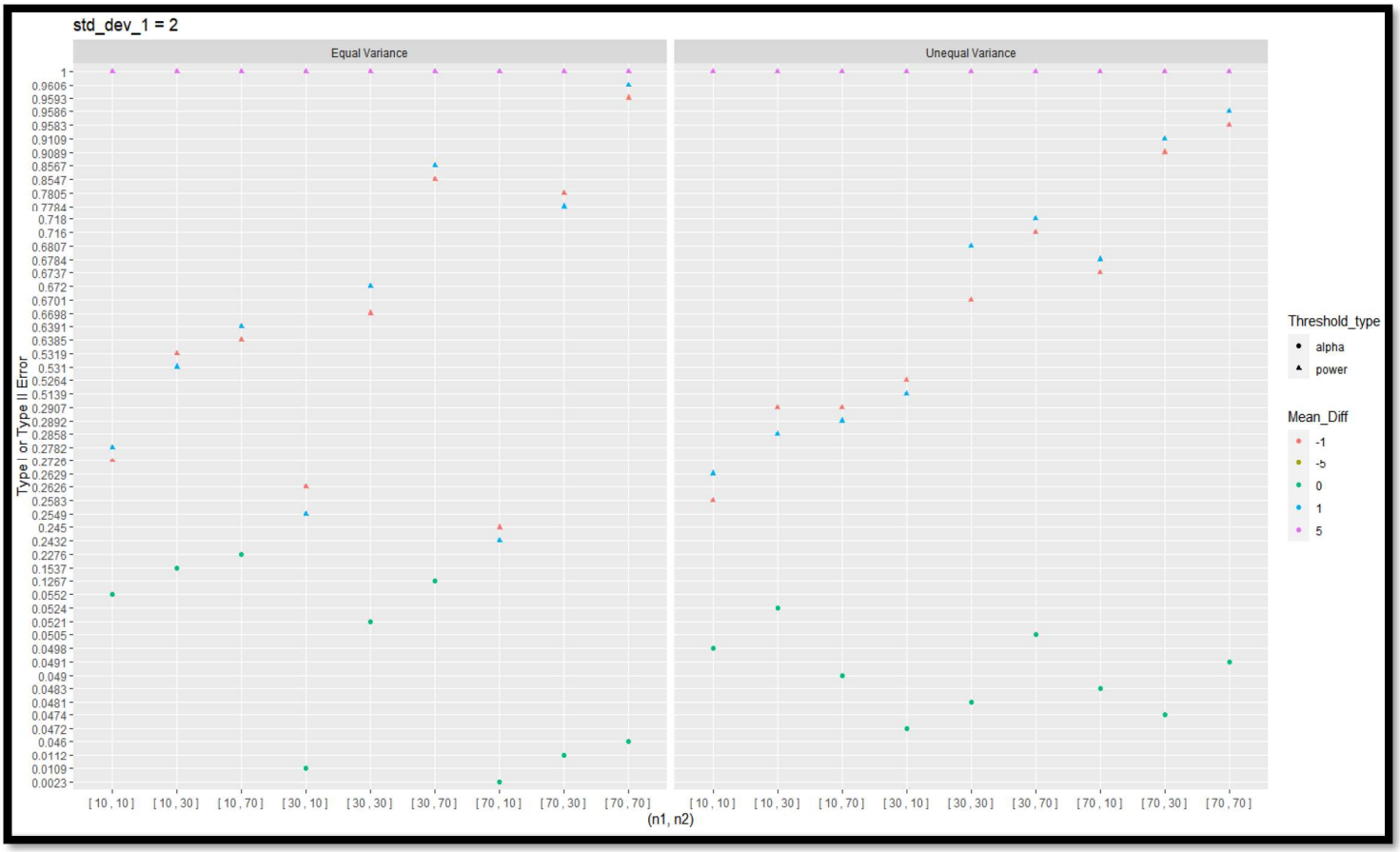
```
#plotting graph with the data_one, data_two & data_three
graph_1 <- ggplot(data.frame(data_one), aes(x=sample_size,y=threshold_value))+
geom_point(aes(shape = Threshold_type, color = Mean_Diff)) + xlab("(n1, n2)")
+ ylab("Type I or Type II Error") + ggtitle("std_dev_1 = 1")
graph_1 + facet_wrap(~type_t_test)
```

2.1 Equal Variance vs Unequal Variance Testing Procedures: Equal Variances $\sigma_2^2 = \sigma_1^2 = 1$



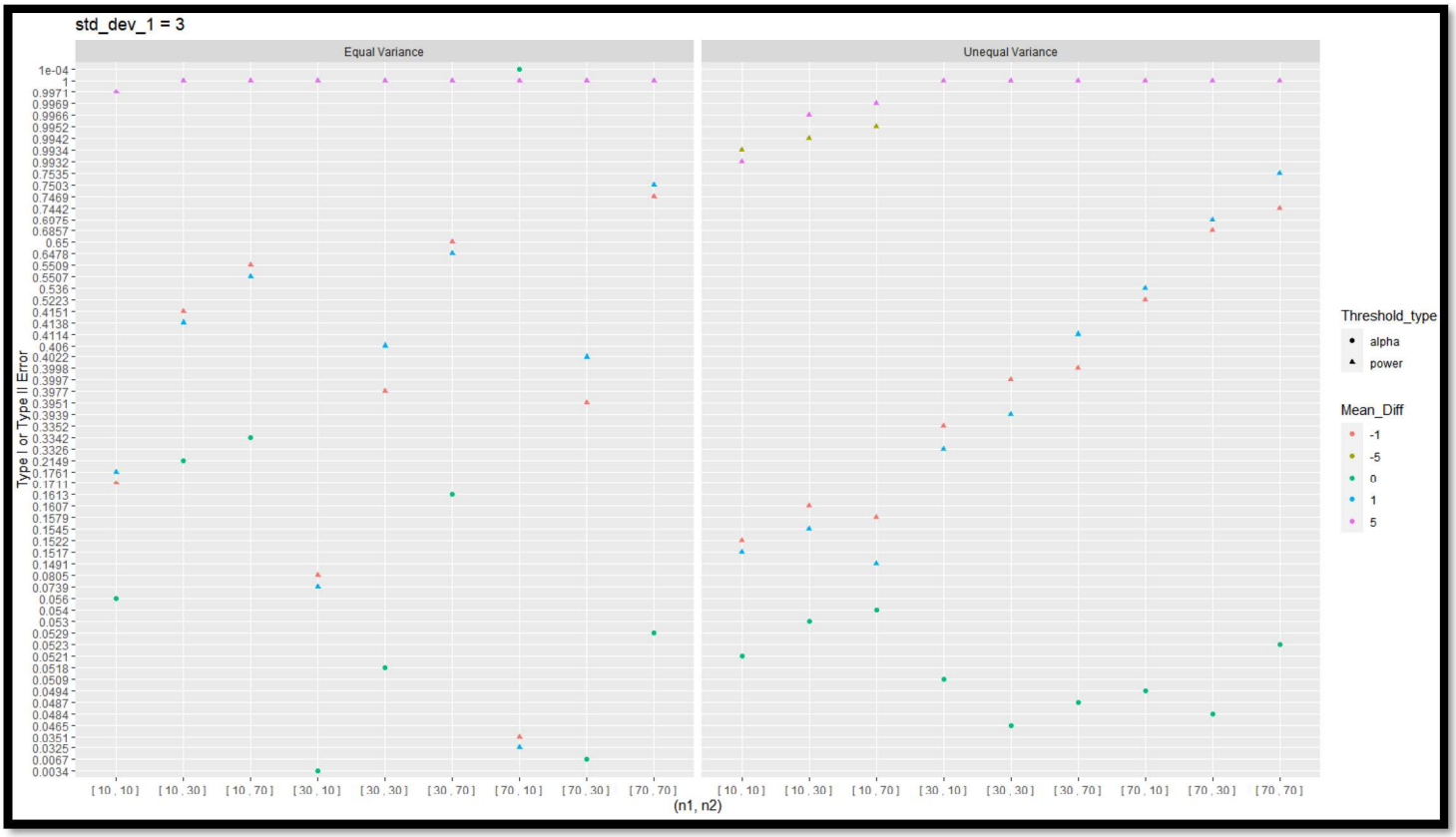
```
graph_2 <- ggplot(data.frame(data_two), aes(x=sample_size,y=threshold_value))+
geom_point(aes(shape = Threshold_type, color = Mean_Diff)) + xlab("(n1, n2)")
+ ylab("Type I or Type II Error") + ggtitle("std_dev_1 = 2")
graph_2 + facet_wrap(~type_t_test)
```

2.2 Equal Variance vs Unequal Variance Testing Procedures: Equal Variances $\sigma_2^2 = 1 \neq \sigma_1^2 = 4$



```
graph_3 <- ggplot(data.frame(data_three), aes(x=sample_size,y=threshold_value)
)+geom_point(aes(shape = Threshold_type, color = Mean_Diff)) + xlab("(n1, n2)"
) + ylab("Type I or Type II Error") + ggtitle("std_dev_1 = 3")
graph_3 + facet_wrap(~type_t_test)
```

2.3 Equal Variance vs Unequal Variance Testing Procedures: Equal Variances $\sigma_2^2 = 1 \neq \sigma_1^2 = 9$



Distribution of work

Anand Morlawar

- ✓ Performed the coding for Question 2
- ✓ Formatted the figures, plots, and legends for readability
- ✓ Added comments for question 1 and drafted the report for Question 1
- ✓ Made necessary adjustments for grammar and flow
- ✓ Compared result of the test

Premal Patel

- ✓ Drafted the first version of the report
- ✓ Draft the report for Question 2
- ✓ Coded Question 1 and Added comments to the Question 2 code
- ✓ Performed the analysis for the Plots
- ✓ Proofread final report

THE END

Bibliography

1) https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm
2) https://en.wikipedia.org/wiki/Framingham_Heart_Study
3) https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/bs704_nonparametric4.html
4) https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/paired-sample-t-test/