# Analysis of Systolic Blood Pressure of Smokers and Non-Smokers Using Statistical Techniques

**ST 502**

**Spring 2022**

**Barde, Karan**

**Shah, Siddhant**

**Part 1: Introduction:**

As part of our project, we seek to make an inference on the difference in mean between 2 datasets related to systolic blood pressure, one pertaining to smokers and the other pertaining to non-smokers. The data for both samples is normally distributed as $(\mu_1, \sigma_1^2)$ and $(\mu_2, \sigma_2^2)$. As we will explain further below, we use 2 tests: the two-sample t-test for equal variance and the two-sample t-test for unequal variance. As part of these tests, we apply the p-value and confidence interval techniques in our analyses. Lastly, the signicance level, α, used is 0.05. We now test the hypthoses where the null hypothesis indicates that the difference in mean is 0 and the alternative hypothesis indicates a non-zero difference in means.

**Testing and Results:**

We start by performing the 2-sample t-test for equal variance. As depicted in our code snippet for this portion, we inititiate the process by calculating the mean and variance of each of the two datasets, the smokers and non-smokers. Having done so, we proceed to calculate the pooled variance via the below formula:

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

To test our hypothesis: $H_o : \mu1 - \mu2 = 0$ $H_A : \mu1 - \mu2 \neq 0$ where μ1 is the mean sysBP of non-smokers and μ2 is the mean sysBP of smokers.
We then calculate the t-statistic, degrees of freedom, p-value and the confidence interval using the formulas depicted in the code. The p-value is found to be 0.000256 and the confidence interval is found to be (3.232, 15.084).

We perform a similar order of operations for the 2-sample t-test for unequal variance and retreive a p-value of 0.000812 and confidence interval of (3.860, 14.455). A key point to highlight is the Satterthwaite approximation we made use of to calculate the degrees of freedom as displayed below:

$$\nu = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

**Normality in Data:**

Having received these results, we then proceeded to plot our data to check for normality. We do so by using the QQ plots that produce a linear graph (angled at approximately 45 degrees) to indicate normality. Upon observing the QQ plots for both our datasets, we come to the conclusion that our data is in fact not normal. Hence, we proceed to use the Mann-Whitney test as this test does not assume normality of data.

**Conclusion:** In terms of our conclusion from the observations, we reject the null hypothesis in both cases since the difference in means between the 2 samples is not zero as we see the p-value is less than 0.05 in both cases. The t-test that makes use of unequal variance is found to a better test as we see a larger difference in the means between the smoker and non-smoker datasets, and hence this test would be more representative of our data.

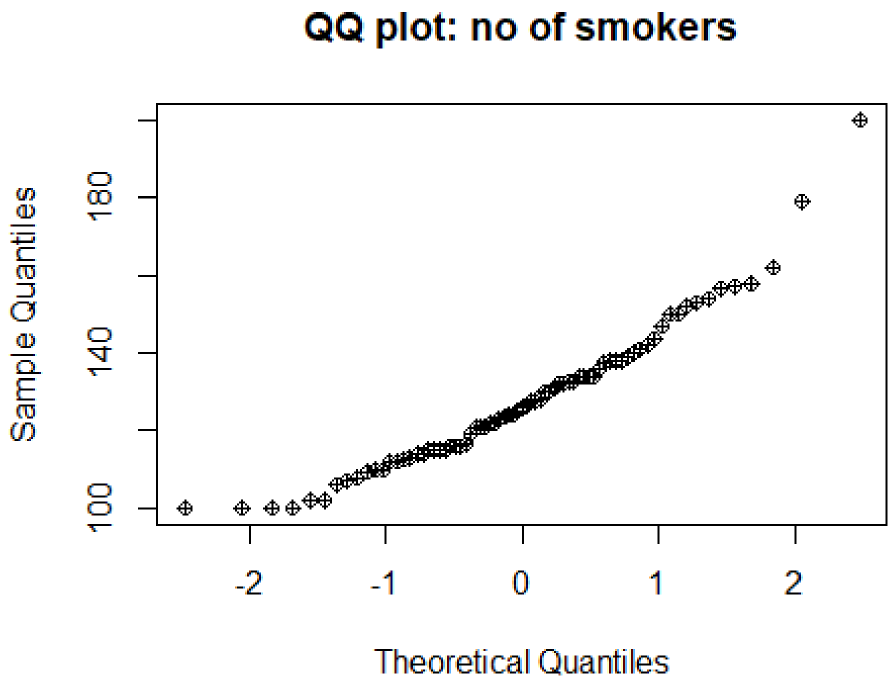
# R code- Part 1

```
getwd()

## [1] "C:/Users/LENOVO/Downloads"

#reading the data into the R studio (answer to part 1 question 1)
data <- read.csv("framingham_data.csv", header = TRUE , sep=",")
head(data)

##   currentSmoker sysBP
## 1             0 157.5
## 2             0 130.0
## 3             0 122.0
## 4             0 136.5
## 5             1 100.0
## 6             1 134.0

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.2

#answer part 1 question 2
#number of smokers and Non smokers
no_of_smoker <- sum(data$currentSmoker == 1)
no_of_non_smoker <- sum(data$currentSmoker == 0)
no_of_smoker

## [1] 75

no_of_non_smoker

## [1] 225

# indexing sysBP which is then used to calculate mean and varaince
index_no_Of_smoker <- (data$currentSmoker == 1)
index_no_of_non_smoker <- (data$currentSmoker == 0)

#Variance of number of smokers and Non smokers samples
var_no_Of_smoker <- var(data$sysBP[index_no_Of_smoker])
var_no_of_non_smoker <- var(data$sysBP[index_no_of_non_smoker])
var_no_Of_smoker

## [1] 352.2117

var_no_of_non_smoker

## [1] 562.1447

#Mean of both samples
mu_no_Of_smoker <- mean(data$sysBP[index_no_Of_smoker])
```

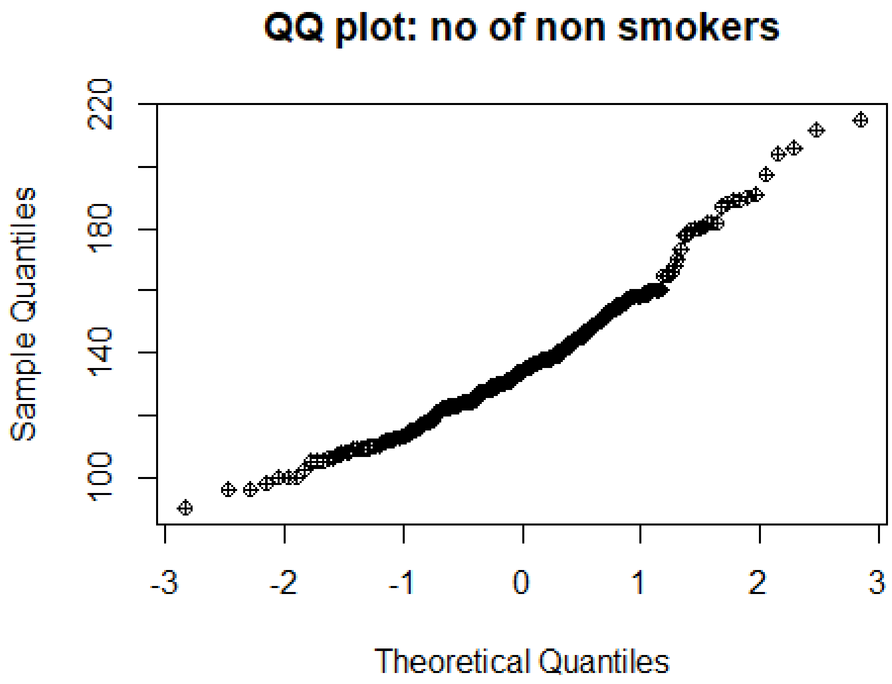
```
mu_no_of_non_smoker <- mean(data$sysBP[index_no_of_non_smoker])
mu_no_Of_smoker

## [1] 128.0667

mu_no_of_non_smoker

## [1] 137.2244

#calculated the pooled variance  and standard deviation
pooled_var <- ((no_of_smoker-1)*var_no_Of_smoker + (no_of_non_smoker-1)*var_n
o_of_non_smoker)/(no_of_non_smoker+no_of_smoker-2)
pooled_var

## [1] 510.0137

pooled_std <- sqrt(pooled_var)
pooled_std

## [1] 22.58348

#The test statistic
t_test <- (mu_no_of_non_smoker - mu_no_Of_smoker)/(pooled_std*sqrt(1/no_of_sm
oker + 1/no_of_non_smoker))
t_test

## [1] 3.041308

#degrees of freedom
df <- no_of_non_smoker + no_of_smoker - 2
df

## [1] 298

## Method 1 : Using p value for alpha(a)=0.05
p_val <- 2*pt(t_test,df,lower.tail = FALSE)
p_val

## [1] 0.002564943

# As p value = 0.0025 < alpha , we reject null hypothesis that the mean of BP
of smokers and non smokers is equal

##Method 2 : Using Confidence interval
p_sigma <- sqrt(pooled_var*(1/no_of_smoker + 1/no_of_non_smoker))
p_sigma

## [1] 3.011131

a = 0.05

t_crit <- qt(a/2,df,lower.tail = FALSE)
t_crit
```

```
## [1] 1.967957

upper_bound <- (mu_no_of_non_smoker - mu_no_Of_smoker) + t_crit*p_sigma
lower_bound <- (mu_no_of_non_smoker- mu_no_Of_smoker) - t_crit*p_sigma

c(lower_bound,upper_bound)

## [1]  3.232003 15.083553

##for 95% C.I, the difference in mean Blood pressure between smoker and non s
mokers is between 3.23 to 15.08

# Finding the Satterthwaite approximation ( This is for Unequal Variance wher
e degree of freedom will change compared to equal variance)
df_v <- ((var_no_Of_smoker/no_of_smoker + var_no_of_non_smoker/no_of_non_smok
er)**2)/(((var_no_Of_smoker/no_of_smoker)**2)/(no_of_smoker-1) +  ((var_no_of
_non_smoker/no_of_non_smoker)**2)/(no_of_non_smoker-1))
df_v

## [1] 158.8316

#calculating the test statistic
t_test1 <- (mu_no_of_non_smoker - mu_no_Of_smoker)/(sqrt(var_no_Of_smoker/no_
of_smoker + var_no_of_non_smoker/no_of_non_smoker))
t_test1

## [1] 3.414188

#Method 1 : Using p value
p_value1 <- 2*pt(t_test1,df_v,lower.tail = FALSE)
p_value1

## [1] 0.0008119864

##As p value is 0.0008 which is less than significance level 0f 0.05 , we rej
ect null hypothesis that mean of Blood pressure of smokers and non smoker is
equal

# Method 2 : Using confidence interval
t_crit1 <- qt(a/2,df_v,lower.tail = FALSE)
t_crit1

## [1] 1.975012

p_sigma1 <- sqrt(var_no_Of_smoker/no_of_smoker + var_no_of_non_smoker/no_of_n
on_smoker)
p_sigma1

## [1] 2.682271
```

```
upper_bound_1 <- (mu_no_of_non_smoker - mu_no_Of_smoker) + t_crit1*p_sigma1
lower_bound_1 <- (mu_no_of_non_smoker - mu_no_Of_smoker) - t_crit1*p_sigma1
c(lower_bound_1,upper_bound_1)

## [1]  3.86026 14.45530

## For 95% C.I, he difference in mean Blood pressure between smoker and non s
mokers is between 3.86 to 14.45



wilcox.test(sysBP~ currentSmoker, data=data)

## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  sysBP by currentSmoker
## W = 10386, p-value = 0.002749
## alternative hypothesis: true location shift is not equal to 0

qqnorm(data$sysBP[index_no_Of_smoker], pch=10, main="QQ plot: no of smokers")
```



```
qqnorm(data$sysBP[index_no_of_non_smoker], pch=10, main="QQ plot: no of non s
mokers")
```


QQ plot: no of non smokers
Sample Quantiles
100
140
180
220
-3
-2
-1
0
1
2
3
Theoretical Quantiles

**PART 2:**

**Introduction:**

In part II, we conduct a simulation study to compare the two t-tests we used in Part I on the null and alternative hypotheses. The first test assumed equal variance and the second test assumed unequal variance. The effectiveness of test is determined by the Type I error rate alpha and the power (1-β).

**Simulation Design:**

Two samples are generated, first with mean $\mu_1$ and variance $\sigma^2$, sample size of $n_1$ from normal distribution and the second with mean $\mu_2$ and variance $\sigma^2$, sample size of $n_2$ from normal distribution. 100 simulations are done for each pair of $\sigma_1^2$ , n1, n2, and d (difference of mean of two samples)

Hypothesis: $H_o: d = \mu_1 - \mu_2 = 0$ , $H_A: d = \mu_1 - \mu_2 \neq 0$

Alpha is the type 1 error rate which is rejecting the null hypothesis when null is actually true and Power is the probability of rejecting null hypothesis when alternate hypothesis is true.

Simulate the results in both Cases. Case 1: when the means of the two samples are the same. So here the null hypothesis is true. Case 2: the means are different which means the null hypothesis is not true. So in this we observe the number of times the null hypothesis is rejected for each case. So, calculate the Type I error rate. When means are different, there we calculate the power.

**Results:**

Graphing of $\sigma_1^2$= 1, 4 and 9 with $\sigma_2^2$ = 1 is done and results are also obtained for both equal variance and non-equal variance.

In graph 1, when the variance are equal both t test methods work good to keep in check the alpha and maximize power.

For graphs 2 and 3, when variances are not equal , the t test for equal variances is not able able to keep in check the alpha and power but the t test for unequal variance was able to control alpha and max power.

When it is unequal variances, mean difference increase when sample size increases but this is not the case with equal variance.

**Conclusions:**

So Once the Simulation was done and the results were out, we can conclude that the t test is good for performing hypothesis on difference between the mean of BP of smokers and non-smoker since it had superior result when mean were equal and when means were not equal.

**Distribution of work by each member**

Barde, Karan:

- Part I – Equal Variance Test, Mann-Whitney Test
- Part II – Sample generation, equal variance test function, discussion of our observations
- Writing of report
- Plot analysis and Explanation of Graphs
- Compared result of the test

Shah, Siddhant:

- Part I – Unequal Variance Test, QQ plots for normality check, and discussion of our observations
- Part II – Unequal Variance Test, Combinations of different parameters
- Writing of report
- Plot analysis
- Formatted structure of report

# code_Part2

```
#equal variance test function
sim_1 <- function(n1, n2, mean, sigma, t_test){
  #N=100 simulations
  N = 100
  # acceptance/rejection result stored in empty boolean list
  z = c()
  ##iterating N times
  for (i in 1:N){
    #sample 1 simulation from normal distribution
    p = rnorm(n1, mean, sigma)
    ##sample 2 simulation  from normal distribution
    q = rnorm(n2, 0, 1)
    if (t_test == "EV"){

      z[i] = equal_var_test(p, q)
    } else {

      z[i] = unequal_var_test(p, q)
    }
  }
  # If true, it means rejection of null hypothesis.
  return(length(z[z == TRUE])/100)
}


#equal variance test function
equal_variance <- function(m,n){
  mean_m = mean(m)
  mean_n = mean(n)
  variance_m = var(m)
  variance_n = var(n)
  len_m = length(m)
  len_n = length(n)

  #significance level
  a = 0.05

  #degree of freedom
  df = len_m+len_n-2
  #pooled variance
  pool_variance = ((len_m-1)*variance_m + (len_n-   1)*variance_n)/(df)
  # t statistic
  t_stat = (mean_m-mean_n)/(sqrt(pool_variance)*sqrt(1/len_m+1/len_n))
  #Returns True if null hypothesis is rejected
  return(abs(t_stat)>qt(a/2, df,  lower.tail = FALSE))
```

```r
}

# unequal variance test function
unequal_var_test <- function(x,y){
  mean_x = mean(x)
  mean_y = mean(y)
  len_x = length(x)
  len_y = length(y)
  variance_x = var(x)
  variance_y = var(y)
  a = 0.05
  df = ((((variance_x/len_x) + (variance_y/len_y))^2) / (((variance_x/len_x)^2/(len_x-1)) + (variance_y/len_y)^2/(len_y-1)))
  t_s = (mean_x - mean_y) / sqrt(variance_x/len_x + variance_y/len_y)
  #Returns True if null hypothesis is rejected
  return(abs(t_s) > qt(a/2, df, lower.tail = FALSE))
}

sigma <- c(1,2,3)
n <- c(10,30,70)
mean_diff <- c(0,-5,5,-1,1)
#dataframe will record the return of simulation function
result <- as.data.frame(matrix(0,1,6))

names(result) <- c("t_test", "sample_size", "sigma", "Mean_Dif","Alpha_Yes_No","Alpha_Power_value")
#looping over every combination
for (i in sigma){
  for (j in n){
    for (k in n){
      for (l in mean_diff){
        sigma_val <- paste("(",i,",",1,")")
        size <- paste("(",j,",",k,")")
        #performing equal variance test and storing the return of the simulation in the dataframe
        result[nrow(result)+1,] <- c("Equal_Var",size, sigma_val, l, ifelse(l == 0, "Yes", "No"), sim(j,k,l,i,"E_V"))
        #performing equal variance test and storing the return of the simulation in the dataframe
        result[nrow(result)+1,] <- c("Unequal_Var", size, sigma_val, l, ifelse(l == 0, "Yes", "No"), sim(j,k,l,i,"UnE_V"))
      }
    }
  }
}
#Result
result <- result[-c(1),]
head(result)
```

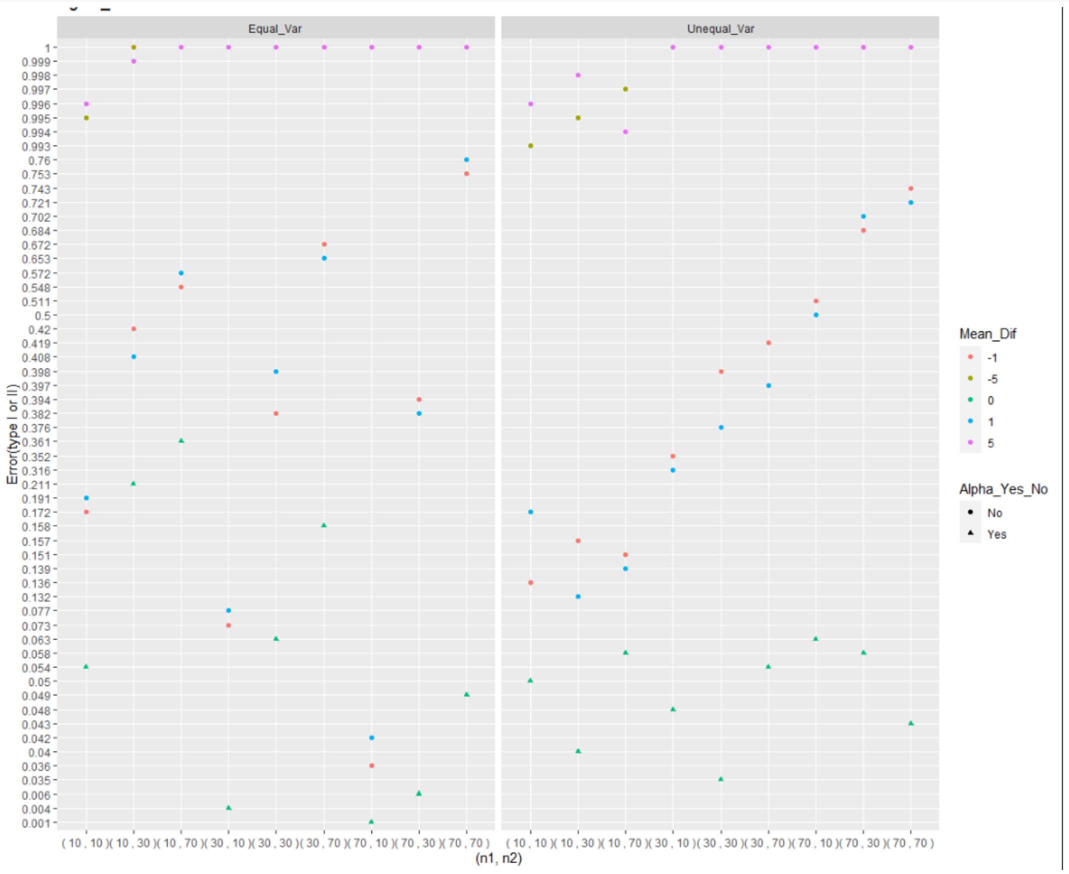
```
##      t_test sample_size     sigma Mean_Dif

2   Equal_Var ( 10 , 10 ) ( 1 , 1 )       0

3 Unequal_Var ( 10 , 10 ) ( 1 , 1 )       0

4   Equal_Var ( 10 , 10 ) ( 1 , 1 )      -5

5 Unequal_Var ( 10 , 10 ) ( 1 , 1 )      -5

6   Equal_Var ( 10 , 10 ) ( 1 , 1 )       5

7 Unequal_Var ( 10 , 10 ) ( 1 , 1 )       5

  Alpha_Yes_No Alpha_Power_value

2        Yes         0.058

3        Yes         0.047

4         No             1

5         No             1

6         No             1

7         No             1
```

```
# data divided based on std dev
df_1 <- result[result$sigma == "( 1 , 1 )",]
df_2 <- result[result$sigma == "( 2 , 1 )",]
df_3 <- result[result$sigma == "( 3 , 1 )",]
```

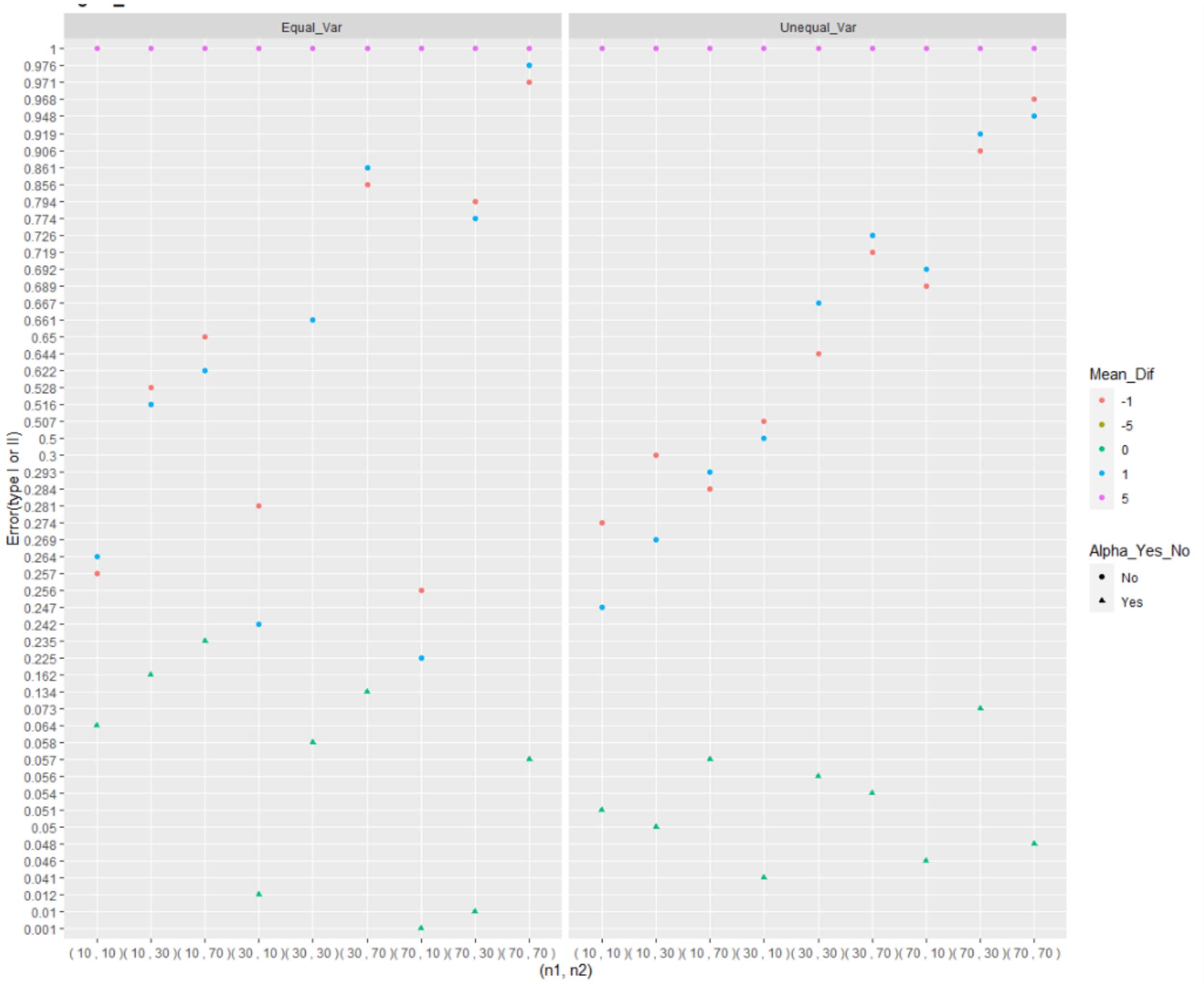
## Plots

```
library(ggplot2)
g_1 <- ggplot(data.frame(df_1), aes(x=sample_size,y=Alpha_Power_value))+geom_
point(aes(shape = Alpha_Yes_No) + xlab("(n1, n2)") + ylab("Error(type I or II
)")

g_1 + facet_wrap(~t_test)

g_2 <- ggplot(data.frame(df_2), aes(x=sample_size,y=Alpha_Power_value))+geom_
point(aes(shape = Alpha_Yes_No)) + xlab("(n1, n2)") + ylab("Error(type I or I
I)") +
graph_2 + facet_wrap(~t_test)

g_3 <- ggplot(data.frame(df_3), aes(x=sample_size,y=Alpha_Power_value))+geom_
point(aes(shape = Alpha_Yes_No) + xlab("(n1, n2)") + ylab("Error(type I or II
)") +
g_3 + facet_wrap(~t_test)
```

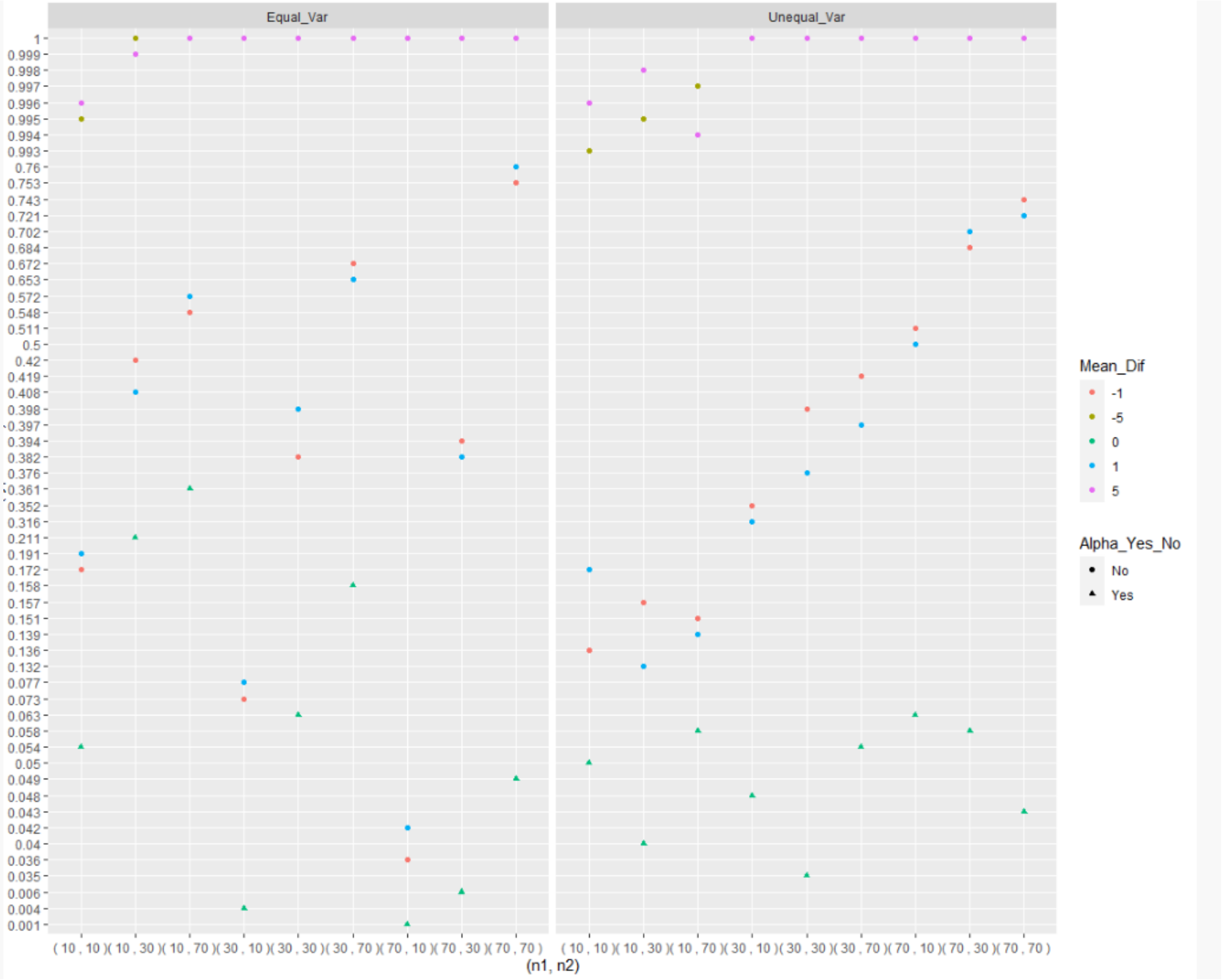
Graph 1: Comparison of Equal and unequal variance $\sigma_1^2 = 1$



Graph2: Comparison of Equal and unequal variance $\sigma_1^2 = 2$



Graph 3: Comparison of Equal and unequal variance $\sigma_1^2 = 3$


Equal_Var
Unequal_Var
1
0.999
0.998
0.997
0.996
0.995
0.994
0.993
0.76
0.753
0.743
0.721
0.702
0.684
0.672
0.653
0.572
0.548
0.511
0.5
0.42
0.419
0.408
0.398
0.397
0.394
0.382
0.376
0.361
0.352
0.316
0.211
0.191
0.172
0.158
0.157
0.151
0.139
0.136
0.132
0.077
0.073
0.063
0.058
0.054
0.05
0.049
0.048
0.043
0.042
0.04
0.036
0.035
0.006
0.004
0.001
( 10 , 10 )( 10 , 30 )( 10 , 70 )( 30 , 10 )( 30 , 30 )( 30 , 70 )( 70 , 10 )( 70 , 30 )( 70 , 70 )
( 10 , 10 )( 10 , 30 )( 10 , 70 )( 30 , 10 )( 30 , 30 )( 30 , 70 )( 70 , 10 )( 70 , 30 )( 70 , 70 )
(n1, n2)
Mean_Dif
-1
-5
0
1
5
Alpha_Yes_No
No
Yes