Math 185 Homework 4

Lucien Chen

2024-05-18

Homework 4

```
library(readr)
library(dplyr)
library(ggplot2)
df <- read.csv("cars.csv")
head(df)
```

```
##   Dimensions.Height Dimensions.Length Dimensions.Width Engine.Information.Driveline               Engine.Infor
mation.Engine.Type
## 1               140               143              202              All-wheel drive         Audi 3.2L 6 cylind
er 250hp 236ft-lbs
## 2               140               143              202            Front-wheel drive Audi 2.0L 4 cylinder 200 h
p 207 ft-lbs Turbo
## 3               140               143              202            Front-wheel drive Audi 2.0L 4 cylinder 200 h
p 207 ft-lbs Turbo
## 4               140               143              202              All-wheel drive Audi 2.0L 4 cylinder 200 h
p 207 ft-lbs Turbo
## 5               140               143              202              All-wheel drive Audi 2.0L 4 cylinder 200 h
p 207 ft-lbs Turbo
## 6                91                17               62              All-wheel drive        Audi 3.2L 6 cylinde
r 265hp 243 ft-lbs
##   Engine.Information.Hybrid Engine.Information.Number.of.Forward.Gears Engine.Information.Transmission Fuel.In
formation.City.mpg
## 1                      True                                          6  6 Speed Automatic Select Shift
18
## 2                      True                                          6  6 Speed Automatic Select Shift
22
## 3                      True                                          6                  6 Speed Manual
21
## 4                      True                                          6  6 Speed Automatic Select Shift
21
## 5                      True                                          6  6 Speed Automatic Select Shift
21
## 6                      True                                          6                  6 Speed Manual
16
##   Fuel.Information.Fuel.Type Fuel.Information.Highway.mpg Identification.Classification          Identificatio
n.ID
## 1                   Gasoline                           25        Automatic transmission           2009 Audi A3
3.2
## 2                   Gasoline                           28        Automatic transmission      2009 Audi A3 2.0
T AT
## 3                   Gasoline                           30           Manual transmission        2009 Audi A3
2.0 T
## 4                   Gasoline                           28        Automatic transmission 2009 Audi A3 2.0 T Qua
ttro
## 5                   Gasoline                           28        Automatic transmission 2009 Audi A3 2.0 T Qua
ttro
## 6                   Gasoline                           27           Manual transmission           2009 Audi A5
3.2
##   Identification.Make Identification.Model.Year Identification.Year Engine.Information.Engine.Statistics.Horse
```

```
power
## 1                    Audi              2009 Audi A3                 2009
250
## 2                    Audi              2009 Audi A3                 2009
200
## 3                    Audi              2009 Audi A3                 2009
200
## 4                    Audi              2009 Audi A3                 2009
200
## 5                    Audi              2009 Audi A3                 2009
200
## 6                    Audi              2009 Audi A5                 2009
265
##   Engine.Information.Engine.Statistics.Torque
## 1                                         236
## 2                                         207
## 3                                         207
## 4                                         207
## 5                                         207
## 6                                         243
```

Question 1

a. We want want investigate whether the drive train affects the city mpg that a car gets or not. We can formulate the hypotheses as follows:

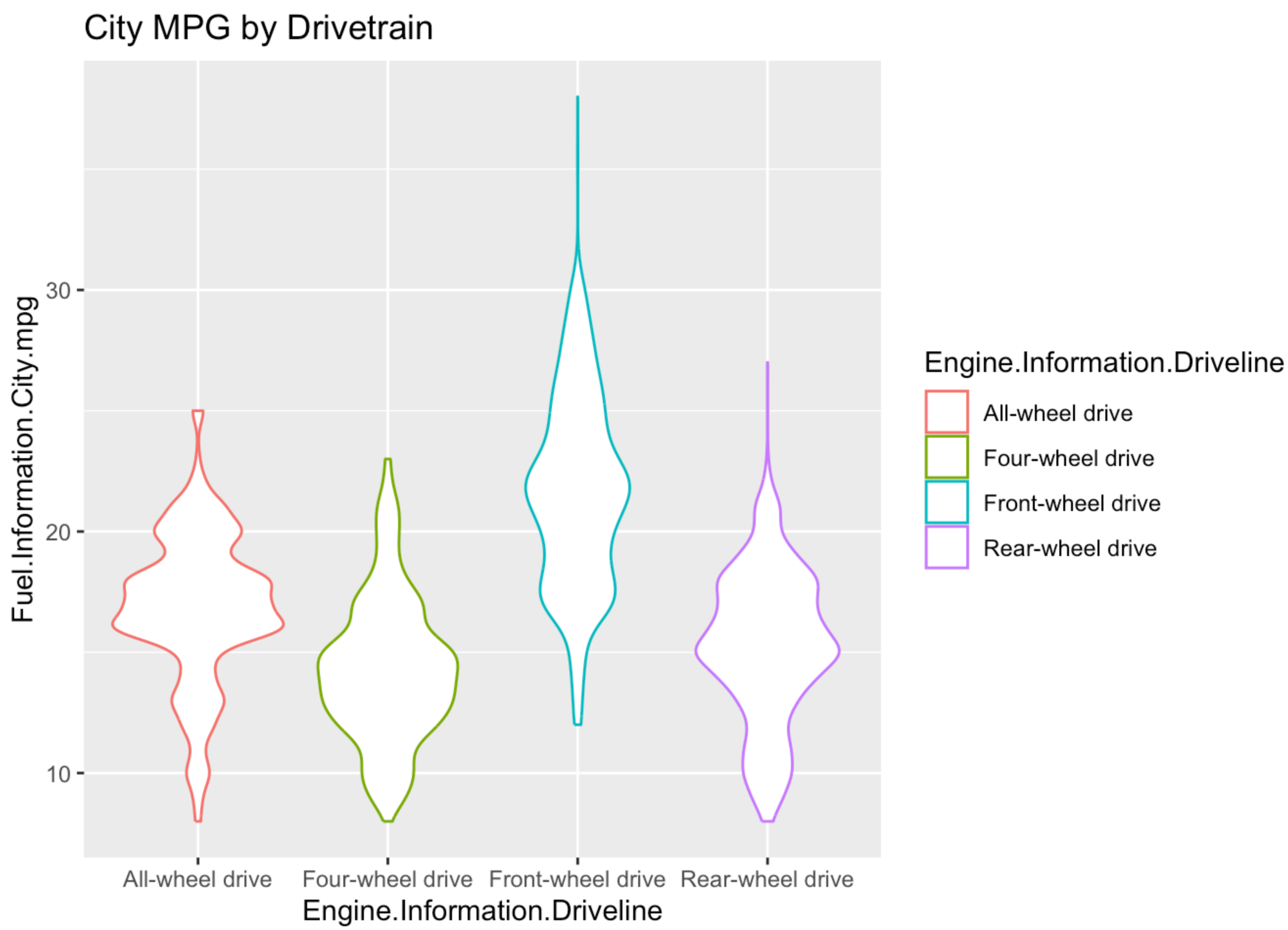
Let μ_i denote the average city mpg for a drive train i, then we can formulate the hypotheses as follows:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_n \text{ versus } H_1 : \exists\, \mu_i \neq \mu_j, i \neq j$$

where n represents the number of different types of drive trains, and each $i \in \{1, 2, \ldots, n\}$ corresponds to a drive train.

b.

```
p <- df %>% ggplot() + aes(Engine.Information.Driveline, Fuel.Information.City.mpg, color=Engine.Information.Driv
eline)
p + geom_violin() + labs(title="City MPG by Drivetrain")
```



Here from the violin plot, we can observe that on average, front-wheel drive cars have the highest city fuel consumption whereas four-wheel drive cars have the lowest.

c. Since we are comparing multiple group means, I will carry out a test using analysis of variance (anova). The assumptions of such a test are that the group distributions are assumed to have homogenous variance (although there is a variant which does not make such an assumption), independence between the groups, and normality. The p-value is obtained by a F-statistic calculated by the ANOVA test which is a ratio of variances between the groups means and the variance within the groups.

```
test <- aov(Fuel.Information.City.mpg ~ Engine.Information.Driveline, data=df)
summary(test)
```

```
##                                Df Sum Sq Mean Sq F value Pr(>F)
## Engine.Information.Driveline    3  44973   14991    1337 <2e-16 ***
## Residuals                    5072  56861      11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

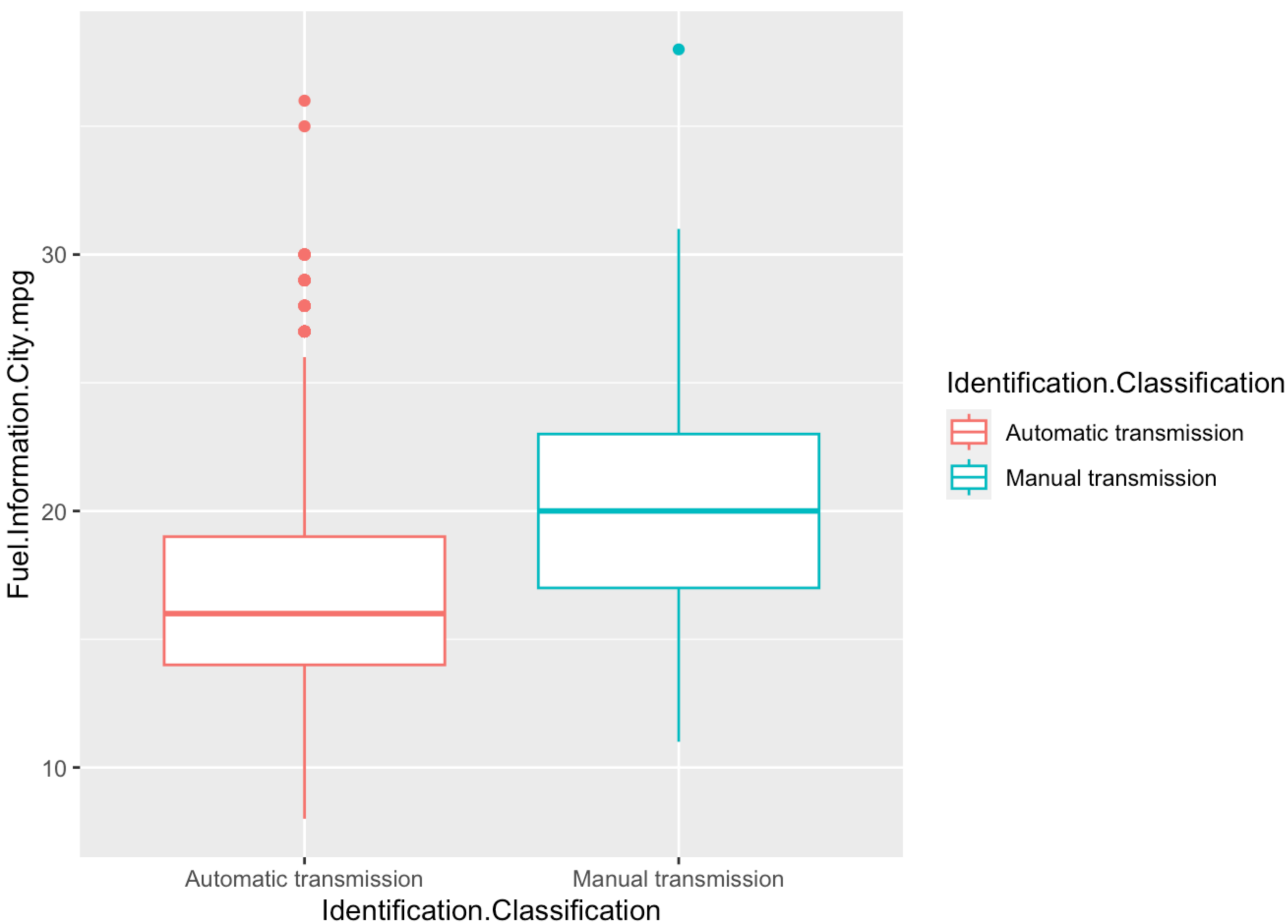
As we can see from the results of the test above, the p-value obtained from the test above is incredibly small and we would reject the null hypothesis at most levels of significance. It seems that the drivetrain of the car has an impact on its city fuel consumption.

Question 2

a. For this question, I will investigate whether the type of transmission a car has an affect on the city fuel consumption or not. This results in a two-way partition of the data, automatic and manual where we will let automatic be the main categorical variable and manual be the secondary that I am controlling for. I define the hypotheses as follows:

$$H_0 : \mu_A = \mu_M$$
$$\text{versus}$$
$$H_1 : \mu_A \neq \mu_M$$

```
p <- df %>% ggplot() + aes(Identification.Classification, Fuel.Information.City.mpg, color=Identification.Classif
ication) + geom_boxplot()
p
```



From the visualization above, we can see that manual cars have better city fuel consumption that automatic cars on average.

c. Now I will apply the two-sample t-test to compare the means between the two groups and perform permutation to obtain the p-value.

```
compute_t <- function(data) {
  t <- t.test(Fuel.Information.City.mpg ~ Identification.Classification, data=data)$statistic
  return(t)
}
obs_t <- compute_t(df)
n <- 1000
permutations <- numeric(n)
for (i in 1:n) {
  permutation_data <- df
  permutation_data$Fuel.Information.City.mpg <- sample(df$Fuel.Information.City.mpg)
  permutations[i] <- compute_t(permutation_data)
}
p_value <- mean(abs(permutations) >= abs(obs_t))
p_value
```

```
## [1] 0
```

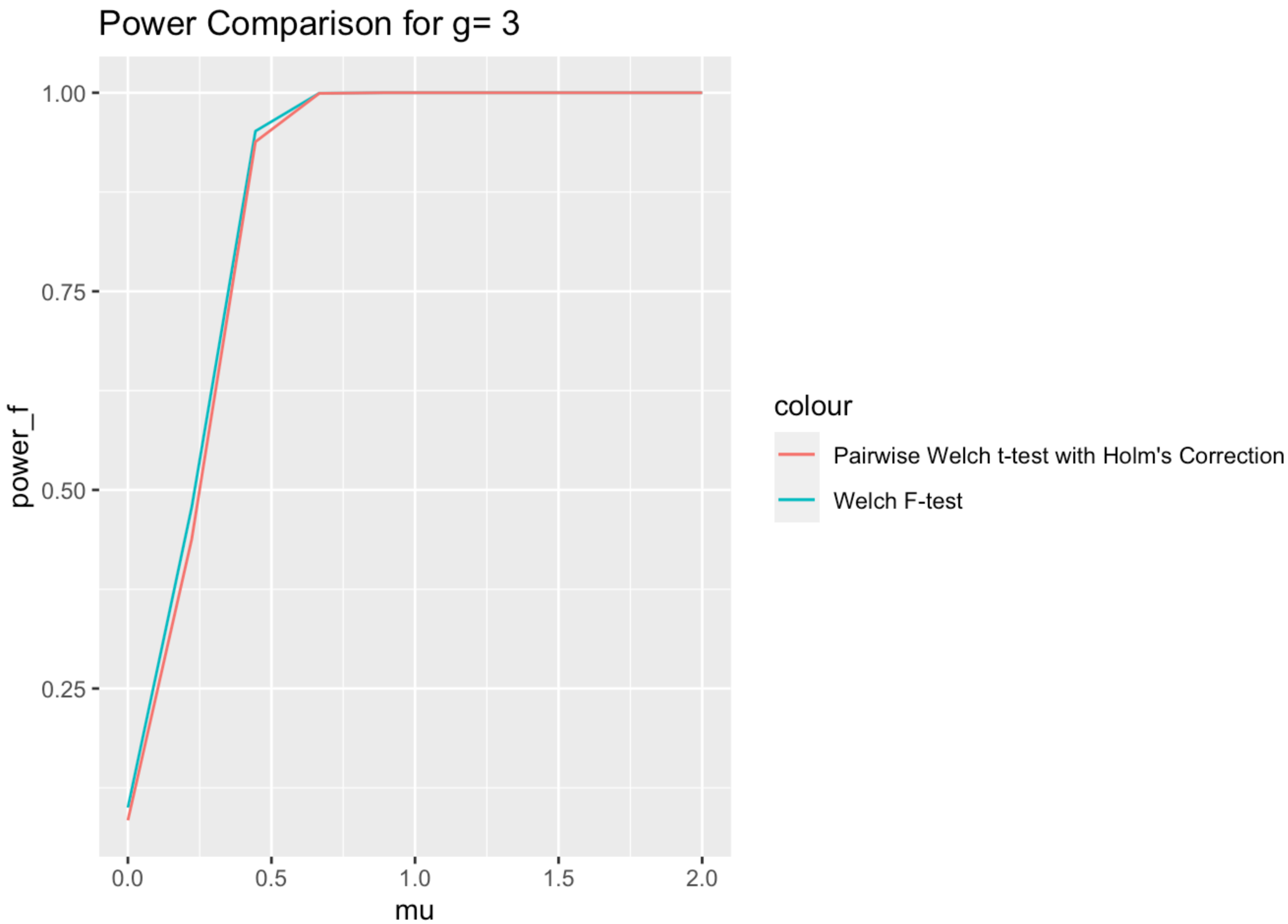
From our permutation test, we obtain a p_value of 0. We reject the null, it appears that the average city fuel consumption differs between automatic and manual cars.

Question 3

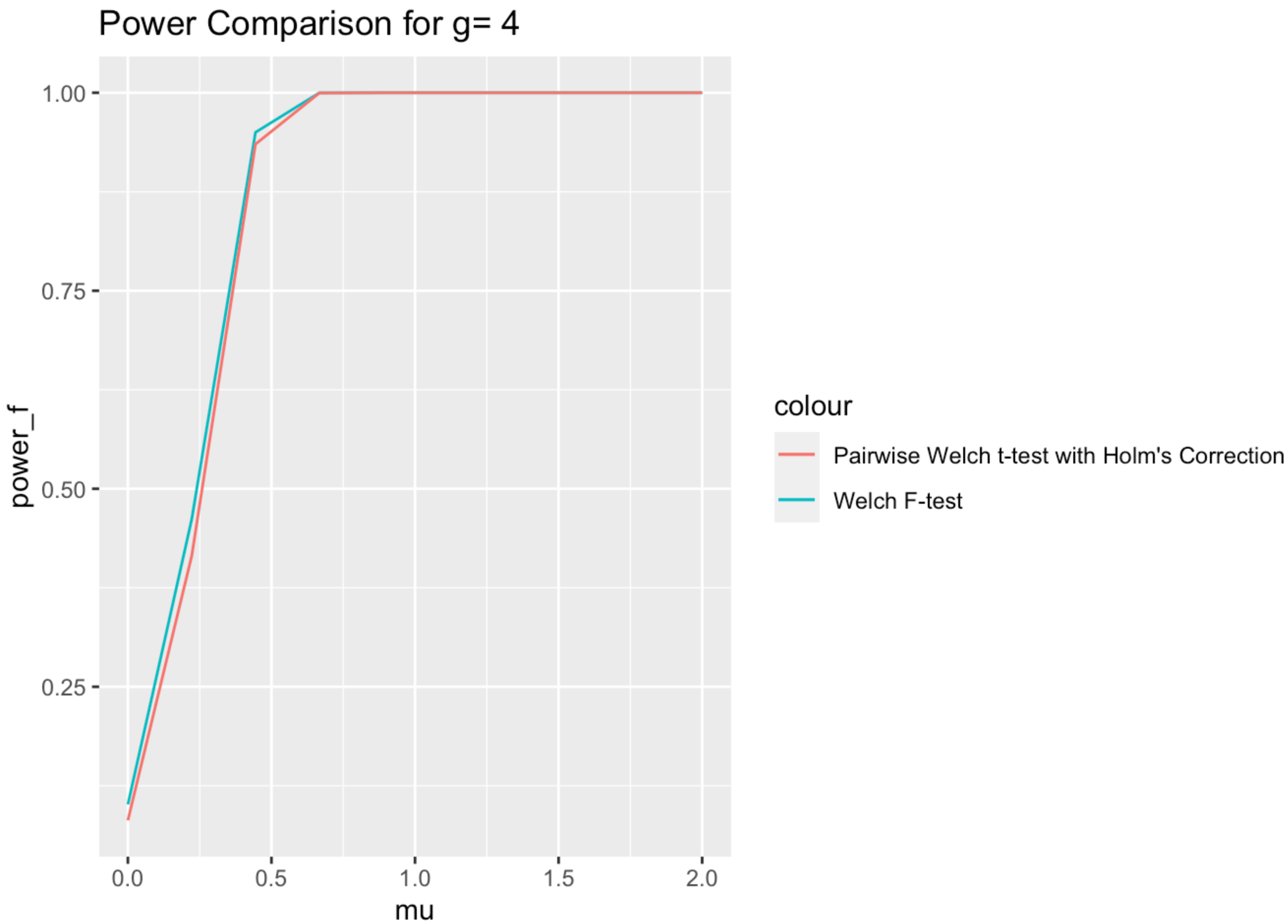
```r
M <- 1e4
n <- 100
simulate_power <- function(mu, alpha) {
  f_rejects <- numeric(M)
  pair_rejects <- numeric(M)
  for (i in 1:M) {
    data <- data.frame(
      group = rep(1:g, each = n),
      value = c(rnorm(n, mu, 1), replicate(g-1, rnorm(n, 0, 1)))
    )
    f_test <- oneway.test(value ~ group, data=data, var.equal=F)
    pair_test <- pairwise.t.test(data$value, data$group, p.adjust.method="holm", pool.sd=F)
    f_rejects[i] = f_test$p.value < alpha
    pair_rejects[i] = any(pair_test$p.value < alpha, na.rm=T)
  }
  return(list(power_f = mean(f_rejects), power_pair = mean(pair_rejects)))
}
mus <- seq(0, 2, length.out=10)
G <- 3:10
res <- vector(mode="list")
for (g in G) {
  powers <- lapply(mus, function(mu) {
      simulate_power(mu, alpha=0.10)
    }
  )
  res[[as.character(g)]] <- data.frame(
    mu = mus,
    power_f = sapply(powers, function(x) {x$power_f}),
    power_pair = sapply(powers, function(x) {x$power_pair})
  )
}

plots <- lapply(names(res), function(g) {
  data <- res[[g]]
  ggplot(data, aes(x=mu)) + geom_line(aes(y=power_f, color="Welch F-test")) + geom_line(aes(y=power_pair, color
="Pairwise Welch t-test with Holm's Correction")) + labs(title = paste("Power Comparison for g=", g))
})
plots
```

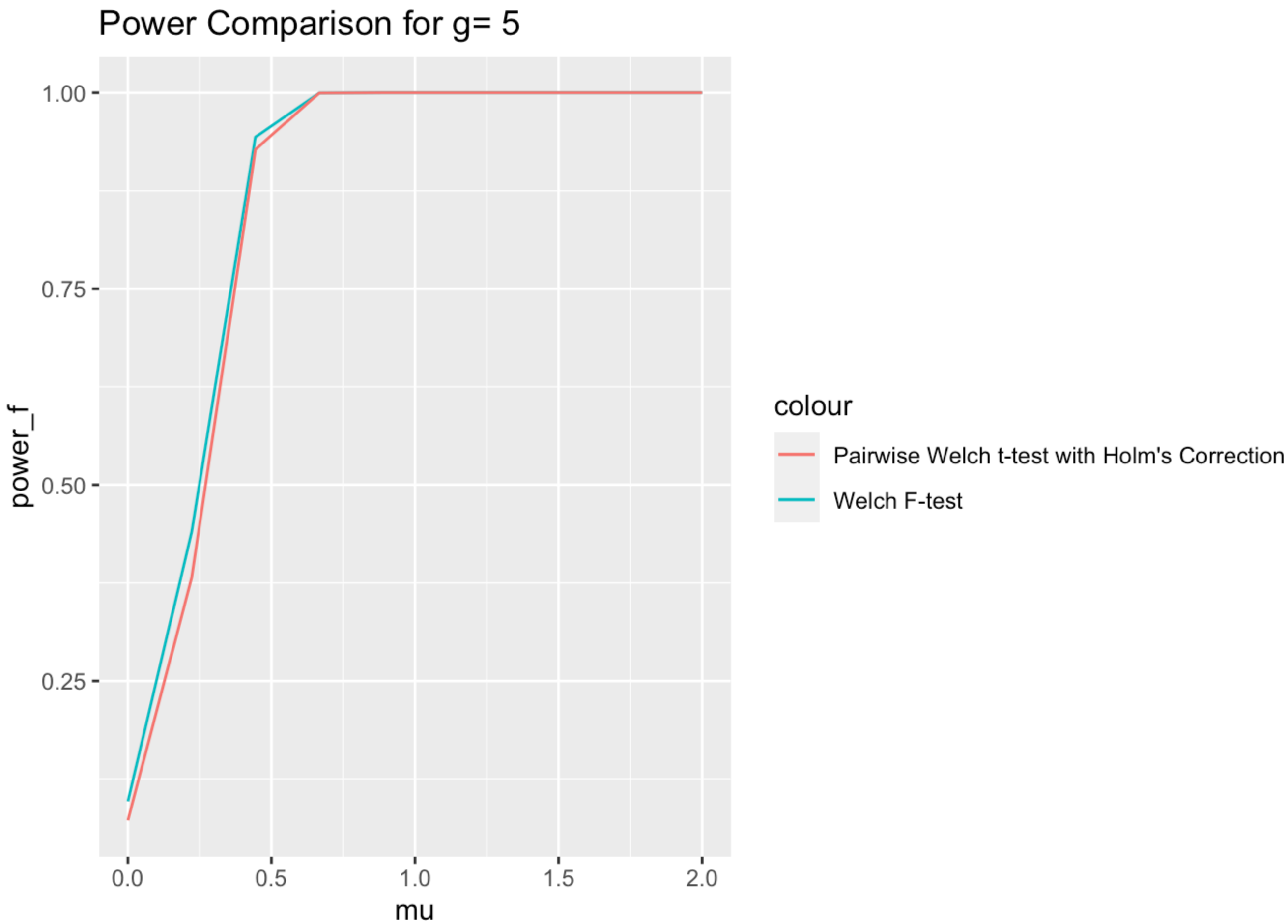
```
## [[1]]
```


Power Comparison for g= 3
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

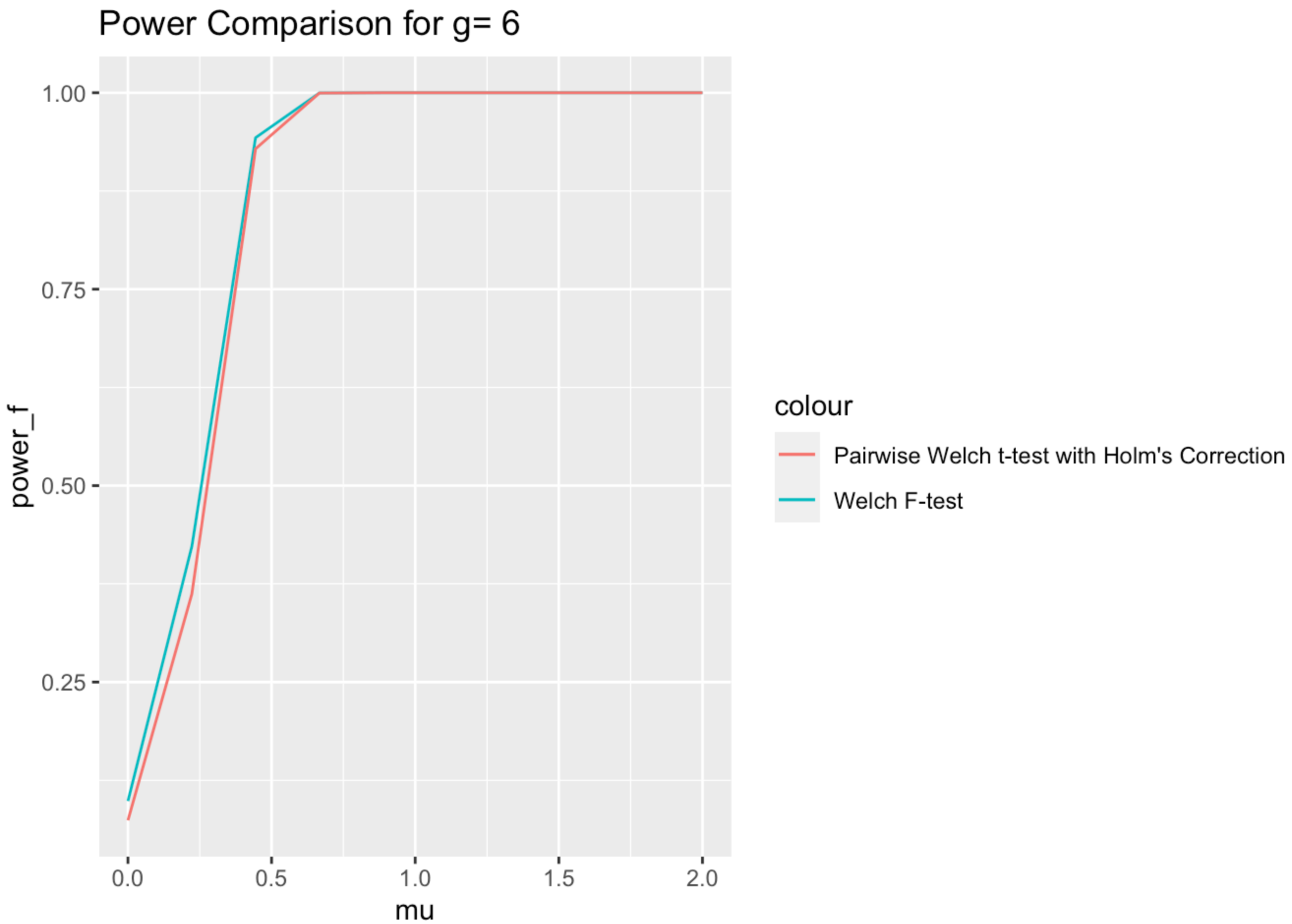
```
##
## [[2]]
```


Power Comparison for g= 4
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

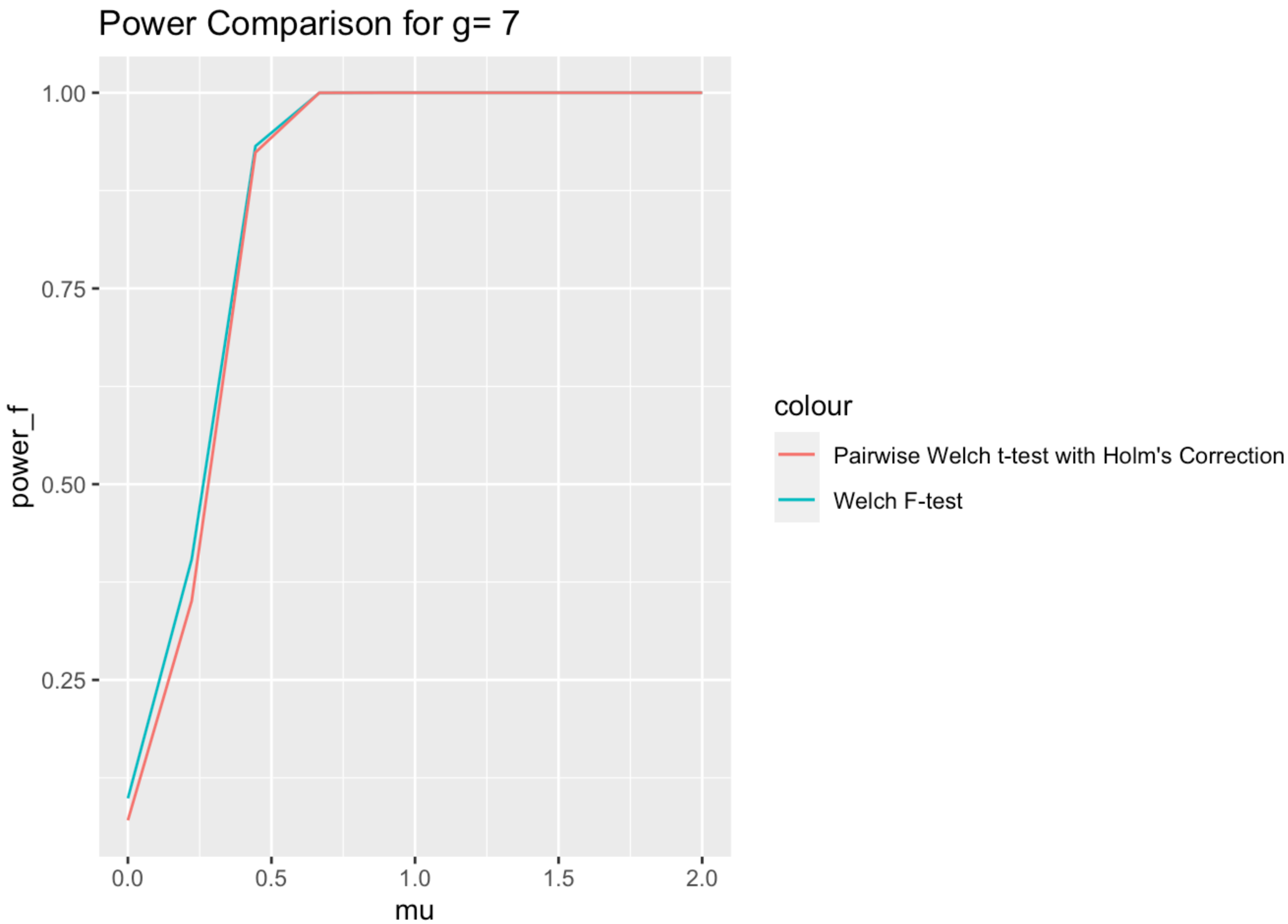
```
##
## [[3]]
```


Power Comparison for g= 5
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

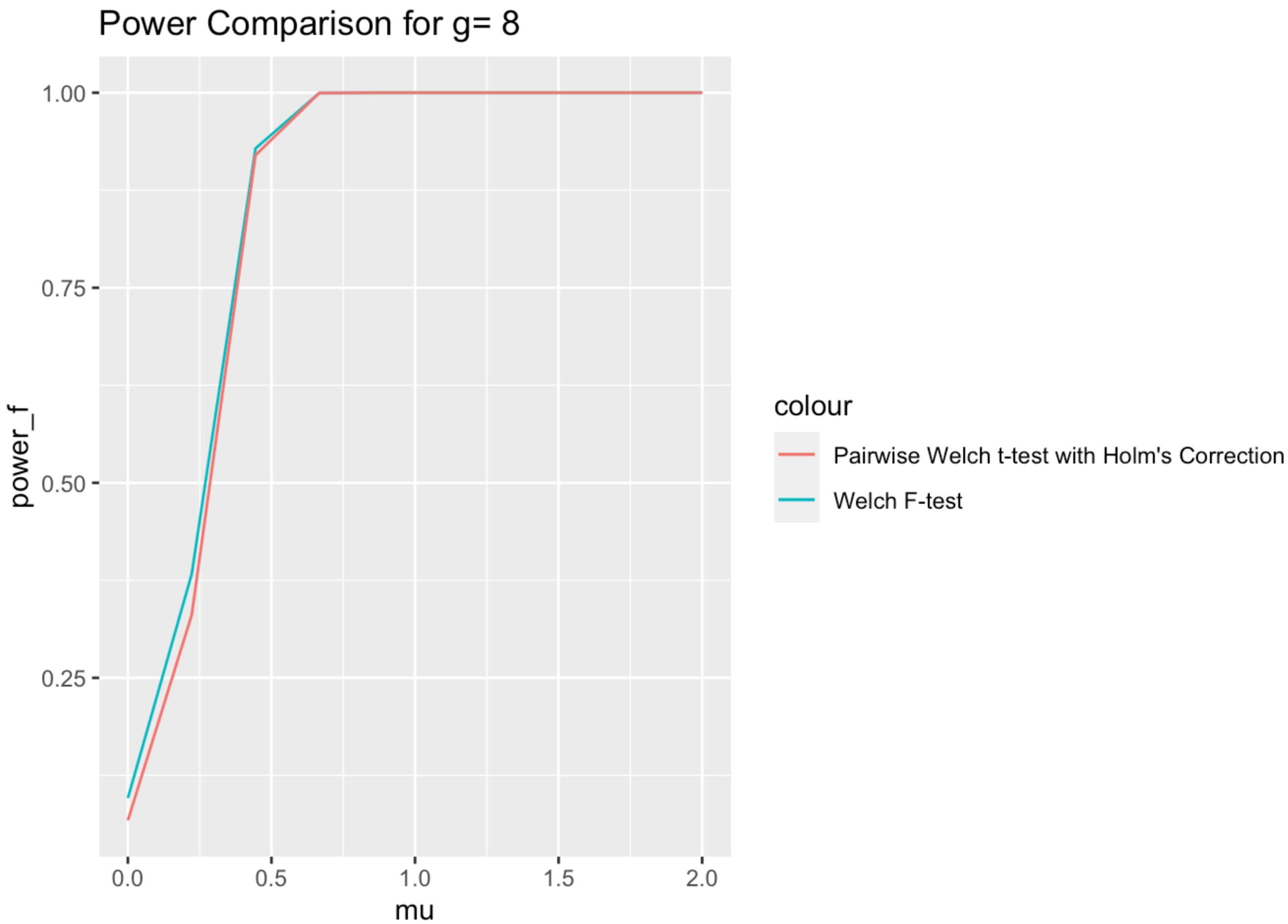
```
##
## [[4]]
```


Power Comparison for g= 6
power_f
1.00
0.75
0.50
0.25
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

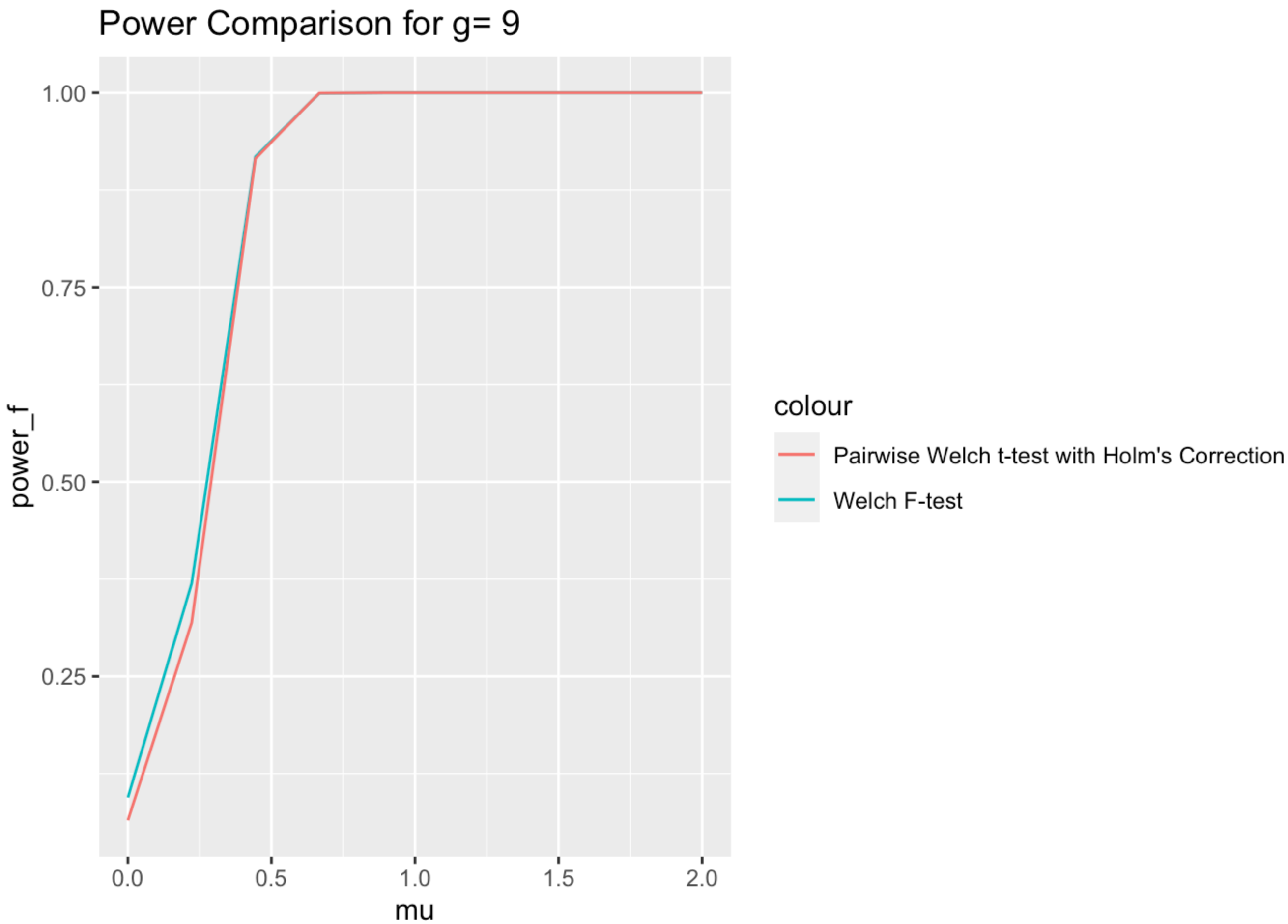
```
##
## [[5]]
```


Power Comparison for g= 7
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

```
##
## [[6]]
```

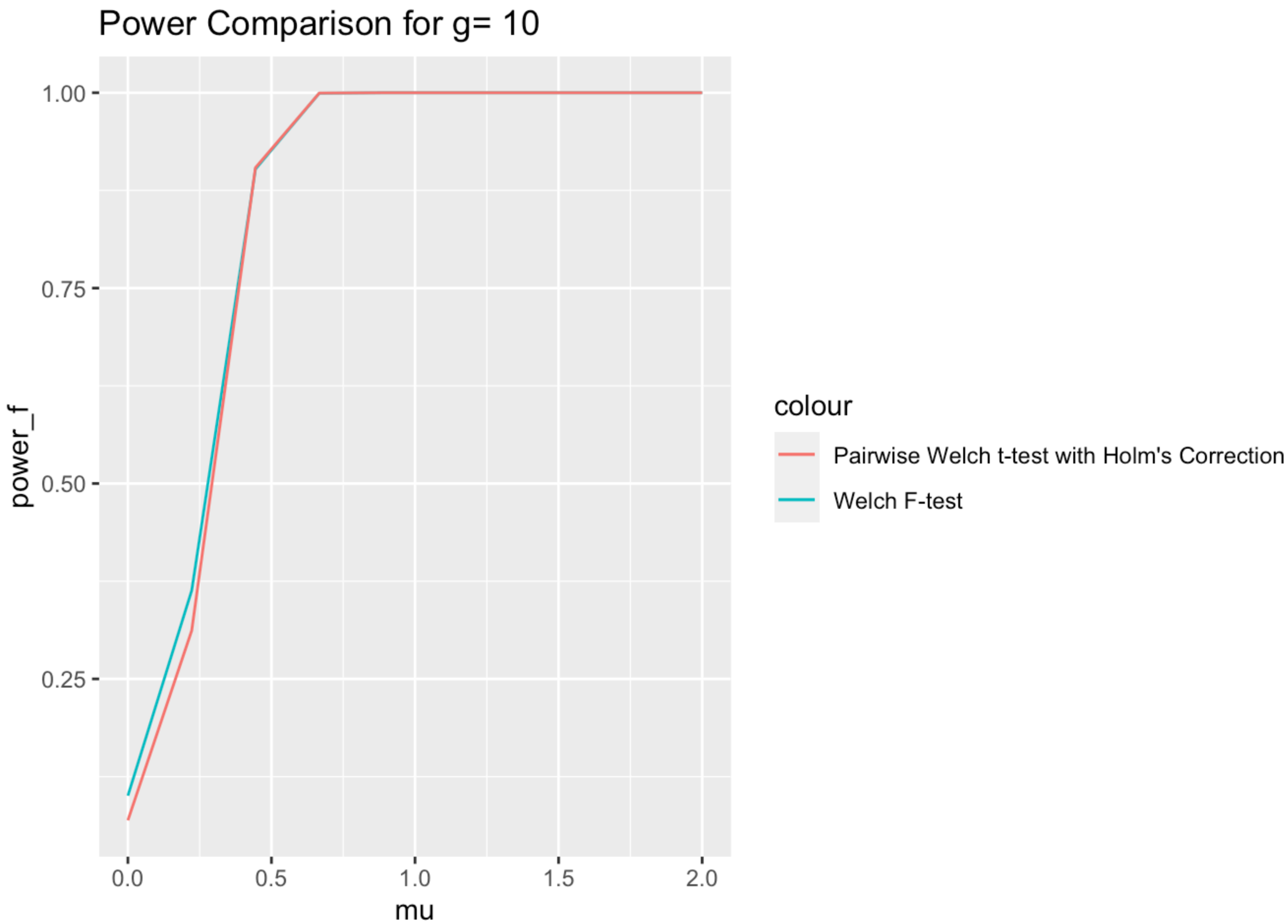

Power Comparison for g= 8
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test

```
##
## [[7]]
```



Power Comparison for g= 9
1.00
0.75
0.50
0.25
power_f
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test


```
##
## [[8]]
```



Power Comparison for g= 10
power_f
1.00
0.75
0.50
0.25
0.0
0.5
1.0
1.5
2.0
mu
colour
Pairwise Welch t-test with Holm's Correction
Welch F-test