hw4

2024-05-18

Problem 1

A. Main question:

Do different types of drivetrain, measure by Engine Information.Driveline affect the gas consumption in the city, measure by Fuel Information.City mpg?

Hypothesis testing:

Null: H0: The gas consumption in the city doesn't affect by different types of drivetrain.

H0: The distributions of gas consumption for all drivetrains are the same.

Alternative: H1: The gas consumption can be influenced by different types of drivetrain.

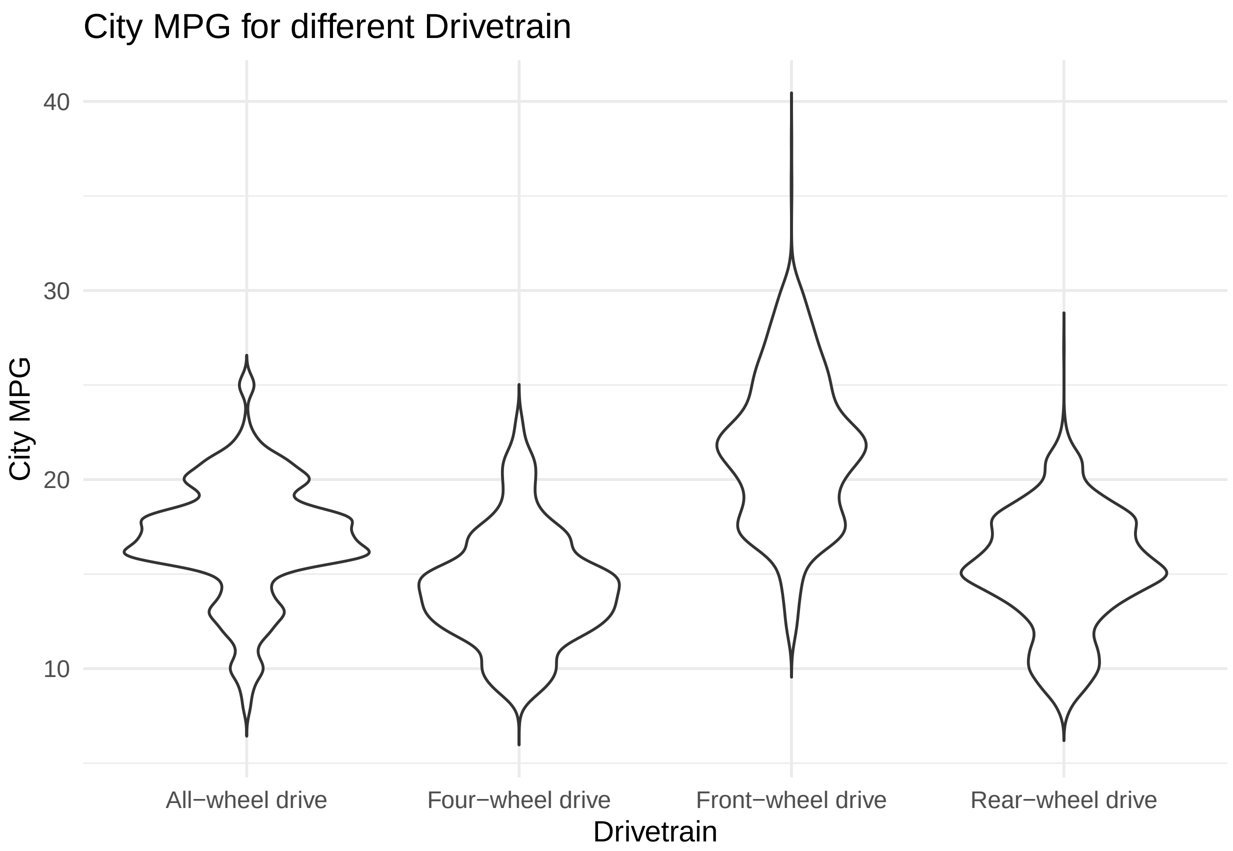
H1: At least one drivetrain has a different distribution of gas consumption in the city compared to the others

B.

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
data <- read.csv("cars.csv")
drivetrain <- data$Engine.Information.Driveline
gas <- data$Fuel.Information.City.mpg

ggplot(data, aes(x = drivetrain, y = gas)) +
  geom_violin(trim = FALSE) +
  labs(title = "City MPG for different Drivetrain",
       x = "Drivetrain",
       y = "City MPG") +
  theme_minimal()
```



C. I will apply the Kruskal-Wallis test to test whether the gas consumption in the city follows the same distribution.

```
kruskal.test(gas ~ drivetrain, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gas by drivetrain
## Kruskal-Wallis chi-squared = 2216.4, df = 3, p-value < 2.2e-16
```

P value: As the p value is very small and less than the threshold of 0.05, which means we will reject the null hypothesis and thus, the different type of drivetrain will influence the gas consumption as at least one drivetrain has a different distribution of gas consumption in the city compared to the others

How p value is obtained: The Kruskal-Wallis test ranks all the data values within each drivetrain, calculates the sum of ranks, and then computes the test statistic based on these ranks. Under the null hypothesis, the p-value is obtained from a chi-squared distribution.

Assumptions: the observations within and across each group of drivetrain are independent. It also assumes random sampling from the populations represented by each drivetrain group.

Problem 2

A. I will use 3 variables including Identification.Classification , Fuel Information.City mpg and Engine Information.Driveline.

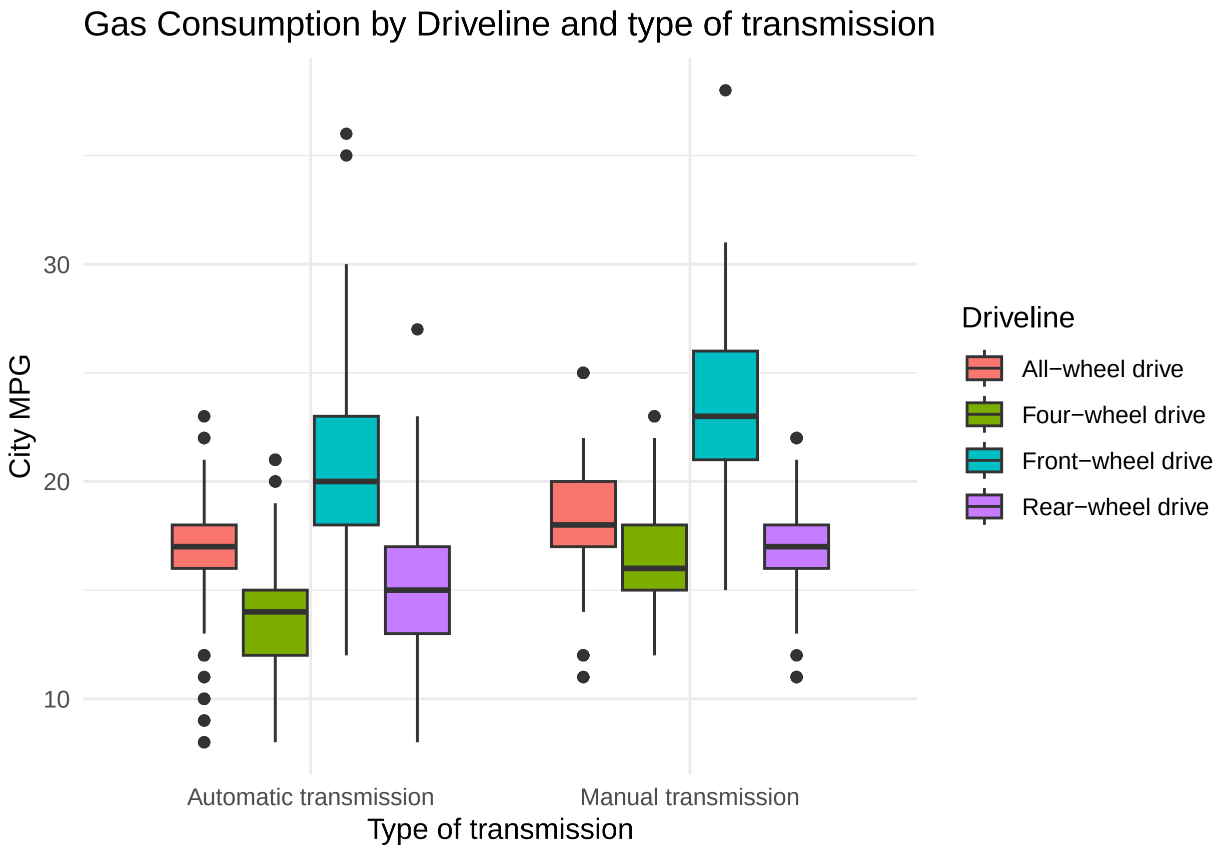
Problem: For the cars with the same type of transmission, does the different types of drivetrain influence the gas consumption?

Null hypothesis: There is no difference in gas consumption across drivetrain types within each category of transmission.

Alternative hypothesis: There is a difference in gas consumption across drivetrain types within at least one category of transmission.

B.

```r
# Plotting the data
make <- data$Identification.Classification
ggplot(data, aes(x = make, y = gas, fill = drivetrain)) +
  geom_boxplot() +
  labs(x = "Type of transmission", y = "City MPG", fill = "Driveline") +
  ggtitle("Gas Consumption by Driveline and type of transmission") +
  theme_minimal()
```



C. I will apply the Kruskal-Wallis test to test whether the gas consumption in the city, categorized by different drivetrains under the same type of transmission (automatic or manual), follows the same distribution.

```r
# test in block
gas_auto <- data$Fuel.Information.City.mpg[data$Identification.Classification=="Automatic transmission"]
gas_manu <- data$Fuel.Information.City.mpg[data$Identification.Classification=="Manual transmission"]
drive_auto <-data$Engine.Information.Driveline[data$Identification.Classification=="Automatic transmiss
```

```
drive_manu <- data$Engine.Information.Driveline[data$Identification.Classification=="Manual transmissio
kruskal.test(gas_auto ~ drive_auto, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gas_auto by drive_auto
## Kruskal-Wallis chi-squared = 1552.7, df = 3, p-value < 2.2e-16
```

```
kruskal.test(gas_manu ~ drive_manu, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  gas_manu by drive_manu
## Kruskal-Wallis chi-squared = 657.21, df = 3, p-value < 2.2e-16
```

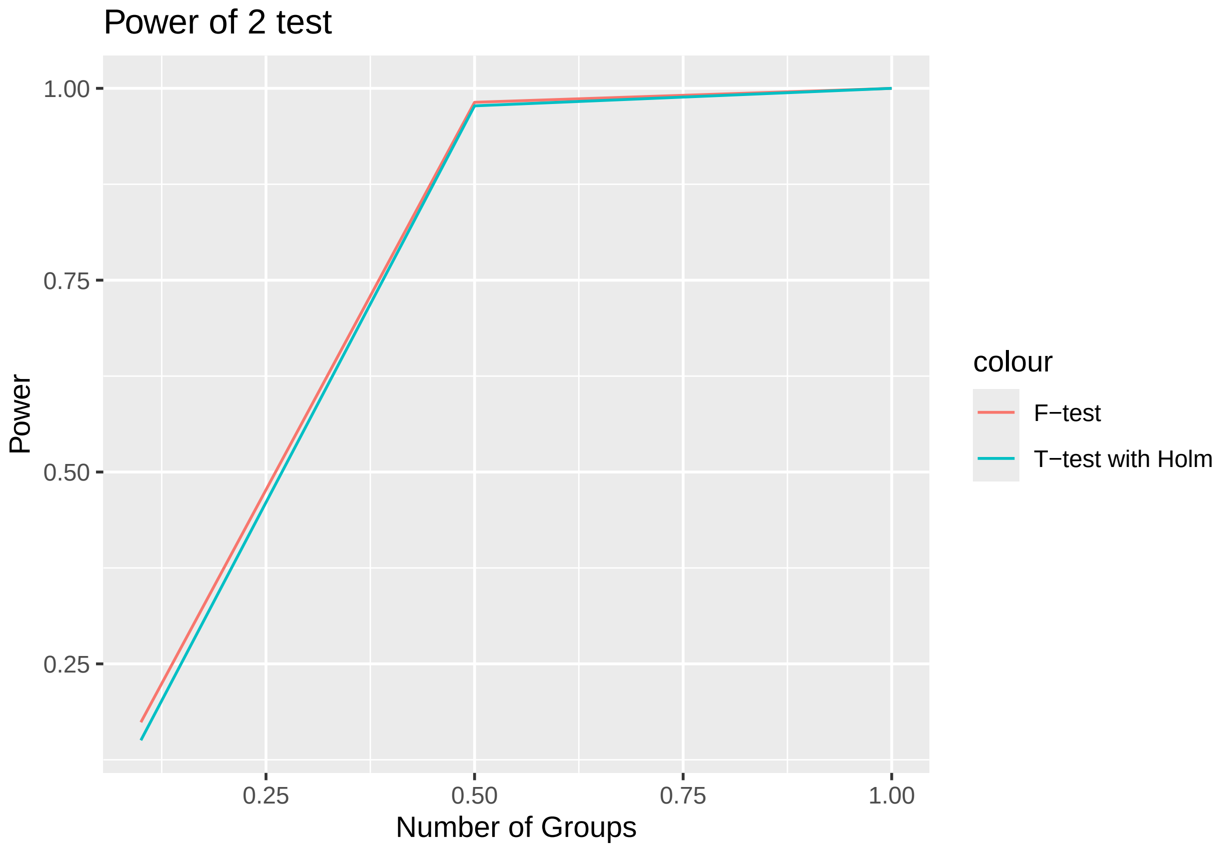
Hence, by the p value we got, we will reject the null hypothesis and thus, under the same type of transmission, the drivetrain can influence the gas consumption.

Problem 3

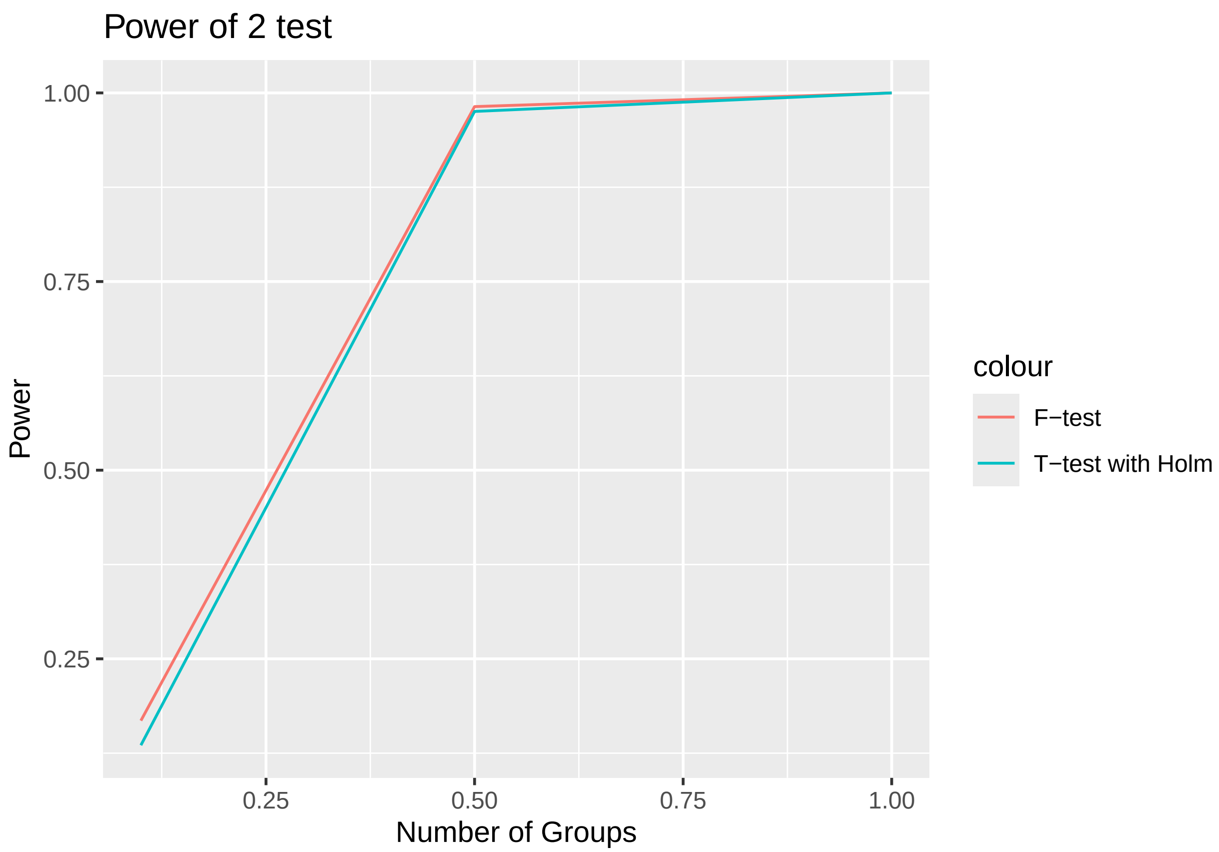
```
library(ggplot2)
M <- 10^4
n <- 100
alpha <- 0.1

# Welch's F-test
welch_f_test <- function(data) {
  f_test <- oneway.test(data ~ group, data = data, var.equal = FALSE)
  return(f_test$p.value)
}

# Welch t-tests with Holm correction
welch_t_test <- function(data) {
  pairwise_test <- pairwise.t.test(data$data, data$group, p.adjust.method = "holm", pool.sd = FALSE)
  return(pairwise_test$p.value)
}

get_plot <- function(g){
    final_data <- list()
    for (u in c(0.1, 0.5, 1)) {
        f <- 0
        t <- 0
        for (sim in 1:M) {
            generted <- vector("list", g)
            generted[[1]] <- rnorm(n, mean = u, sd = 1)
            for (i in 2:g) {
                generted[[i]] <- rnorm(n, mean = 0, sd = 1)
            }
```

```r
            data <- data.frame(data = unlist(generted), group = rep(1:g, each = n))
            # Welch's F-test
            p_value_f_test <- welch_f_test(data)
            f <- f + (p_value_f_test < alpha)

            # All-pairwise Welch t-test with Holm correction
            p_values <- welch_t_test(data)
            t <- t + any(p_values < alpha, na.rm = TRUE)
        }
        final_data[[length(final_data) + 1]] <- list(
            mu = u,
            F_ = f/M,
            T_ = t/M
        )
    }
  frame <- data.frame()
  # Convert to a data frame
  frame <- do.call(rbind, lapply(final_data, as.data.frame))

  c <- ggplot(frame, aes(x = mu)) +
    geom_line(aes(y = F_, colour = "F-test")) +
    geom_line(aes(y = T_, colour = "T-test with Holm")) +
    labs(title = "Power of 2 test", y = "Power", x = "Number of Groups")
  print(c)
}
for (g in 3:10){
  print(g)
  get_plot(g)
}
```

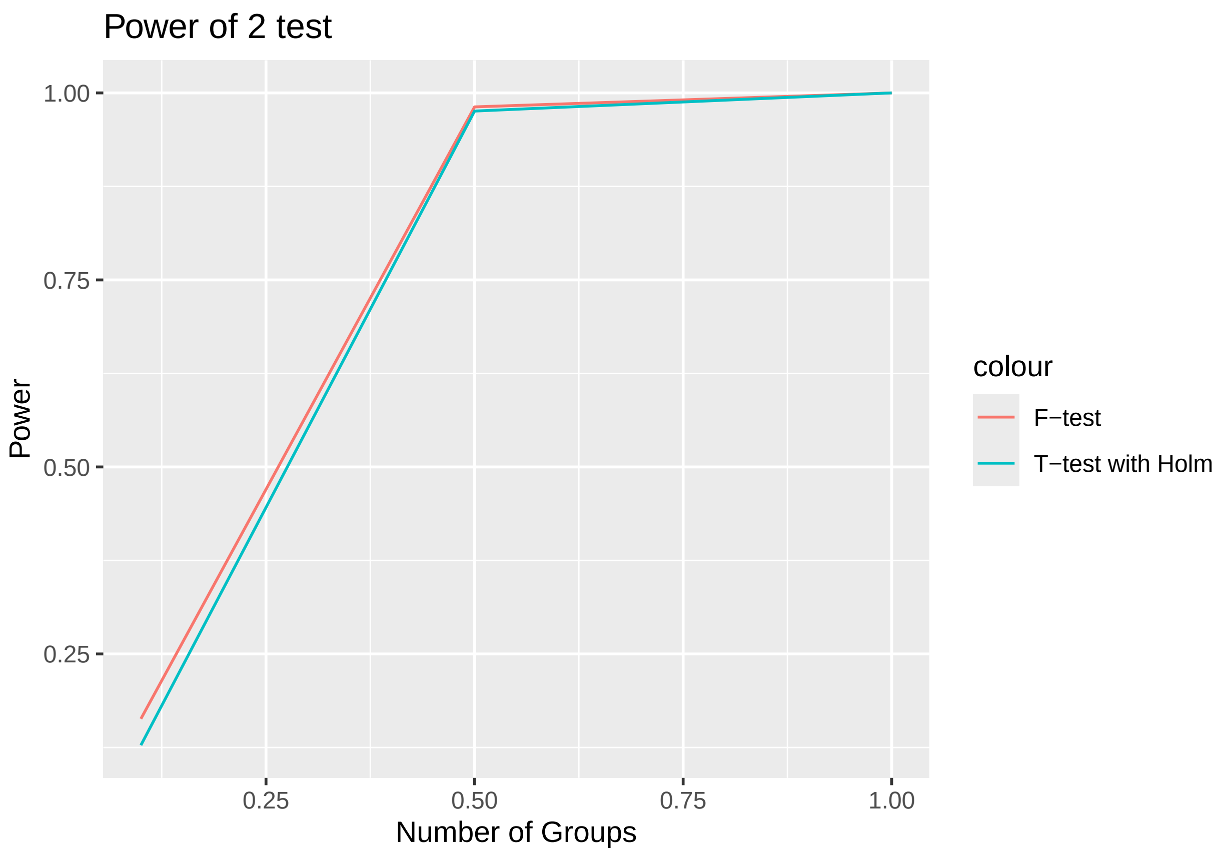
```
## [1] 3
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

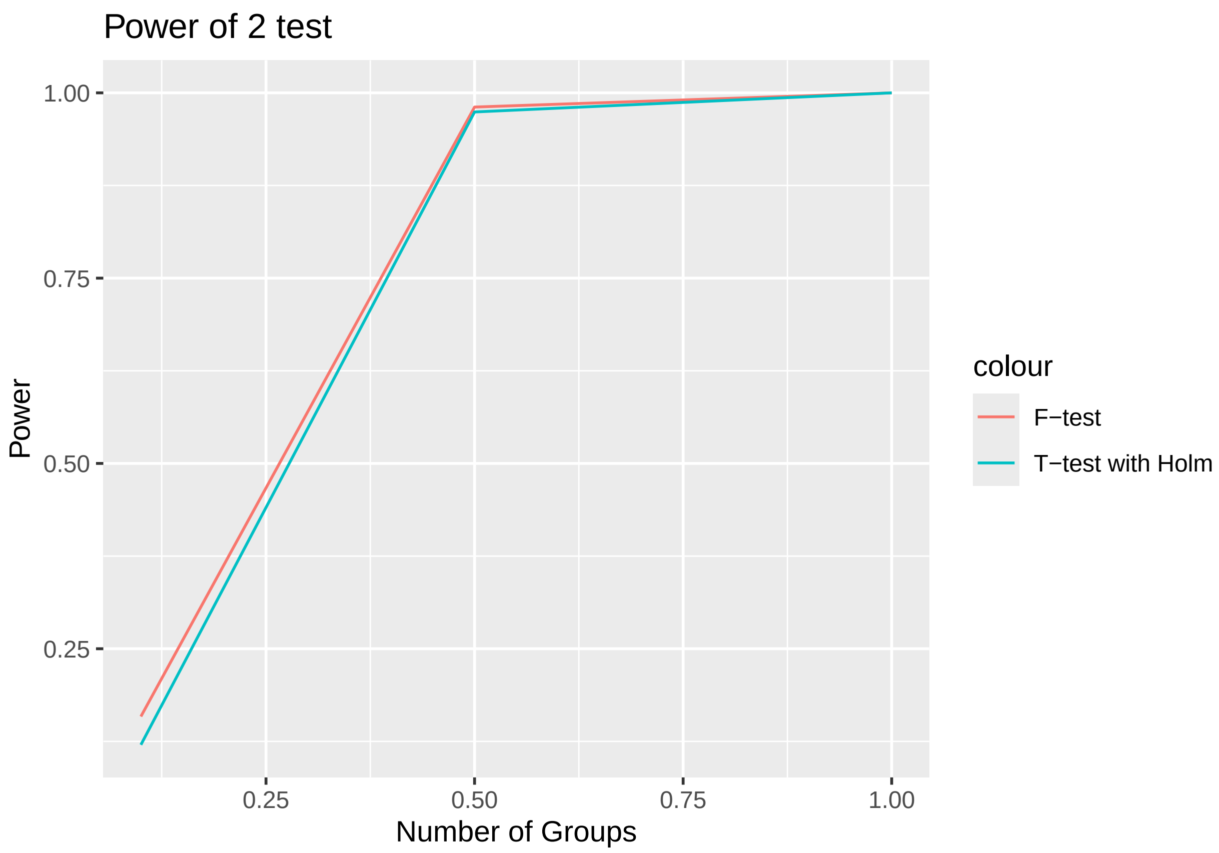
```
## [1] 4
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

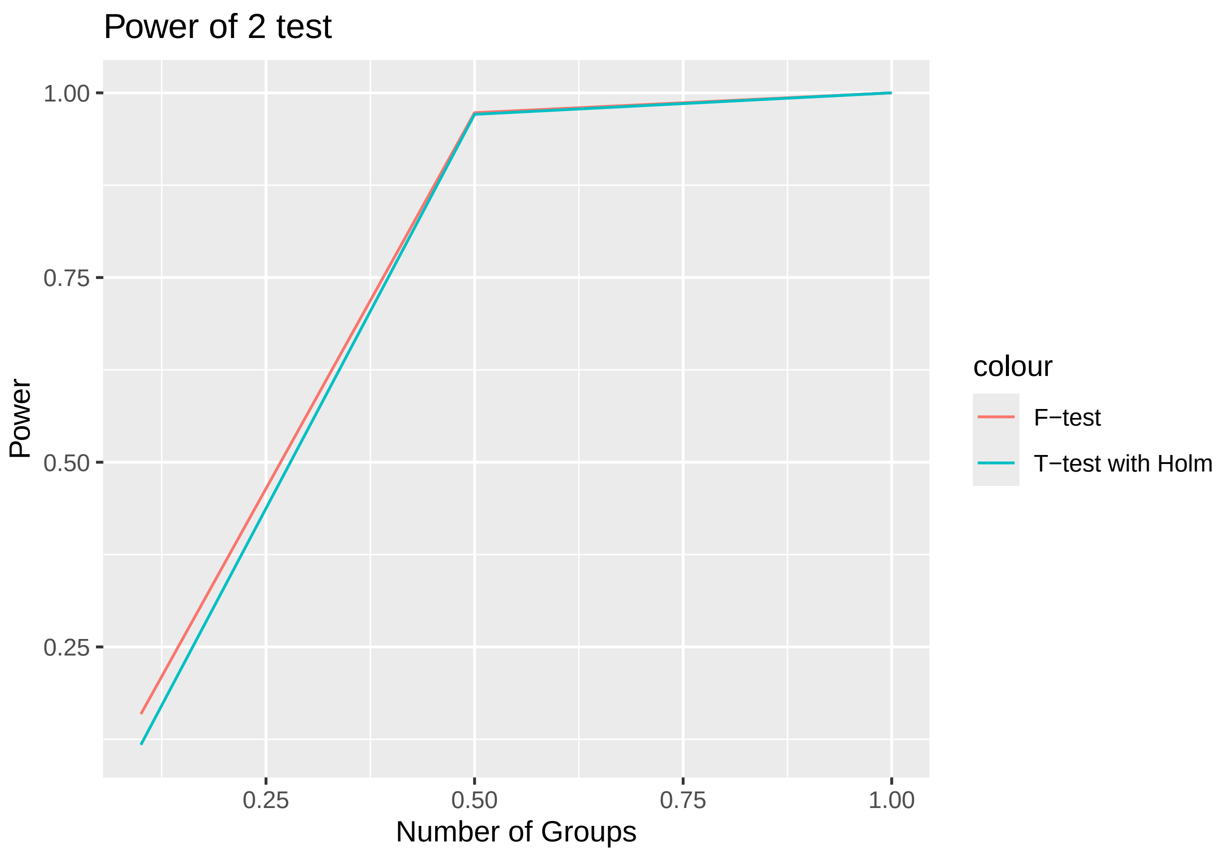
```
## [1] 5
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

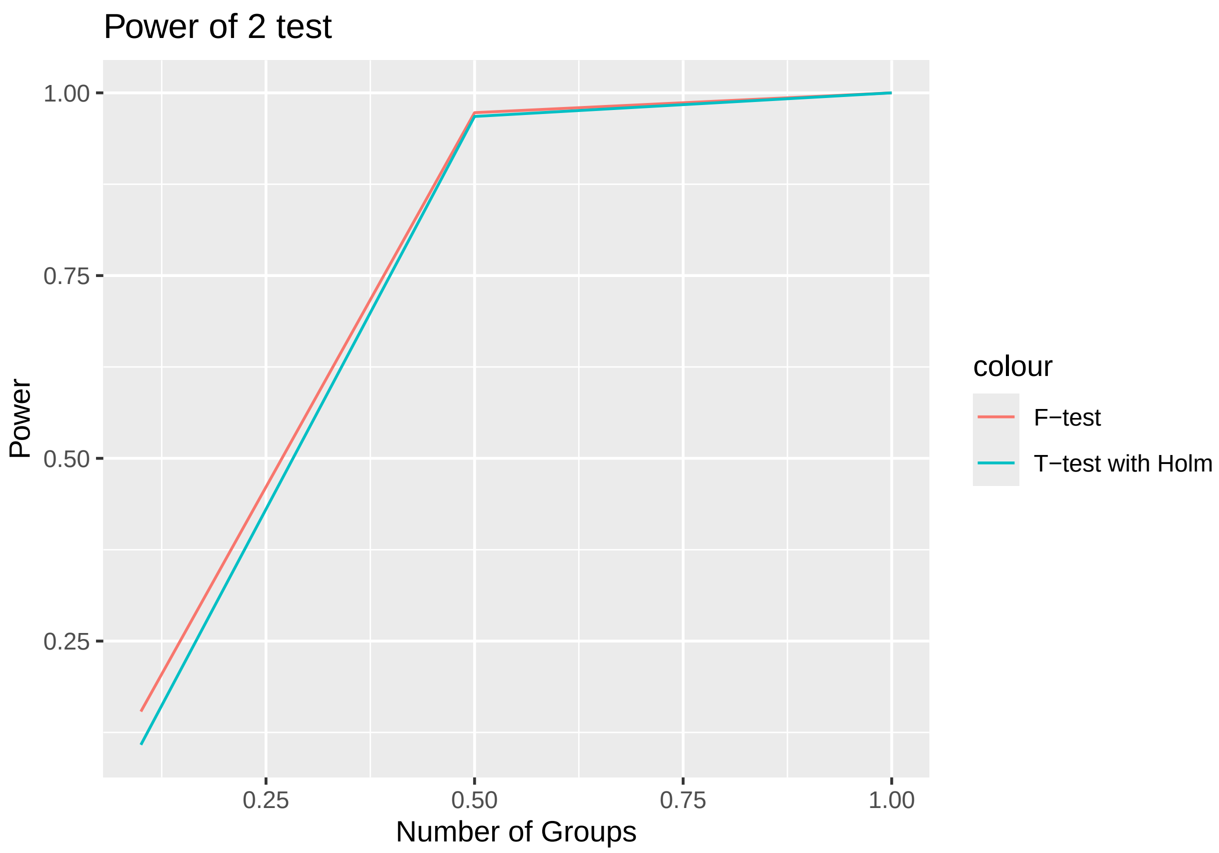
```
## [1] 6
```


Power of 2 test
1.00
0.75
0.50
0.25
Power
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

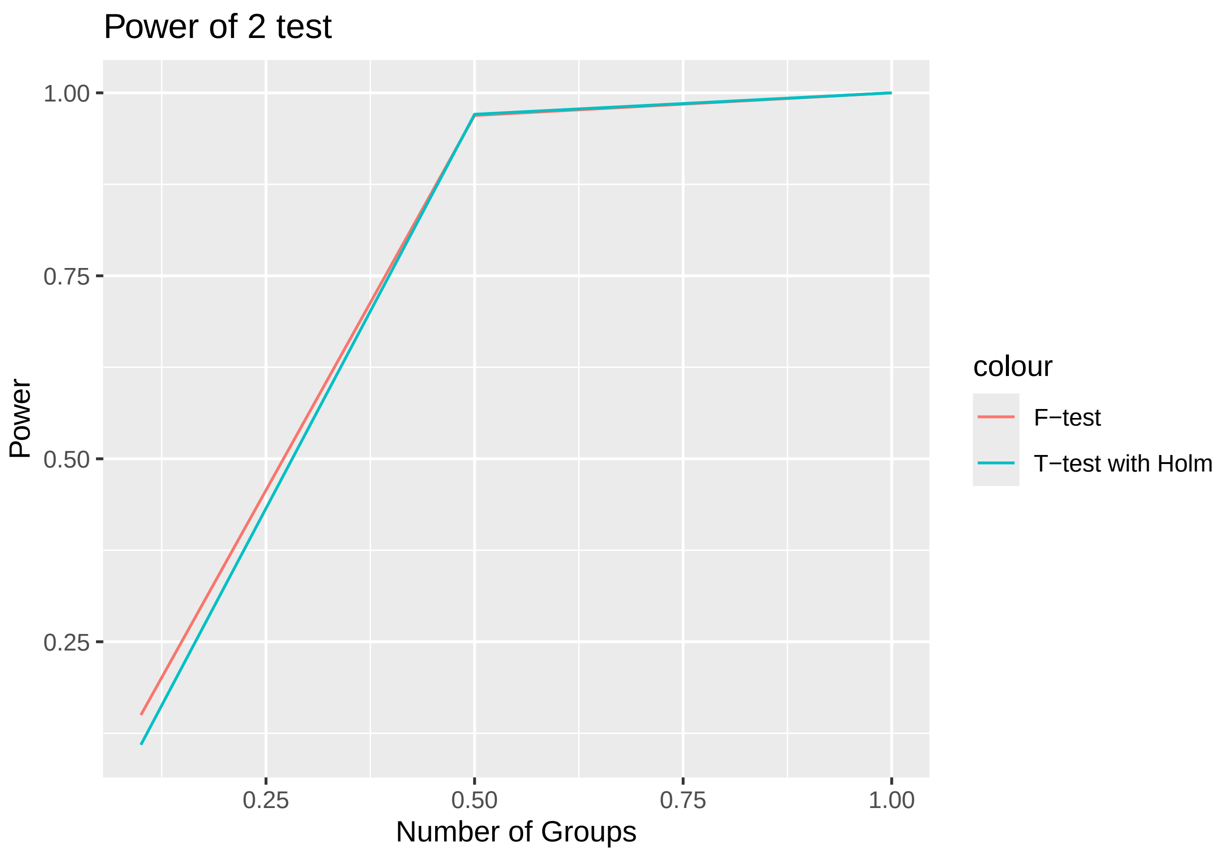
```
## [1] 7
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

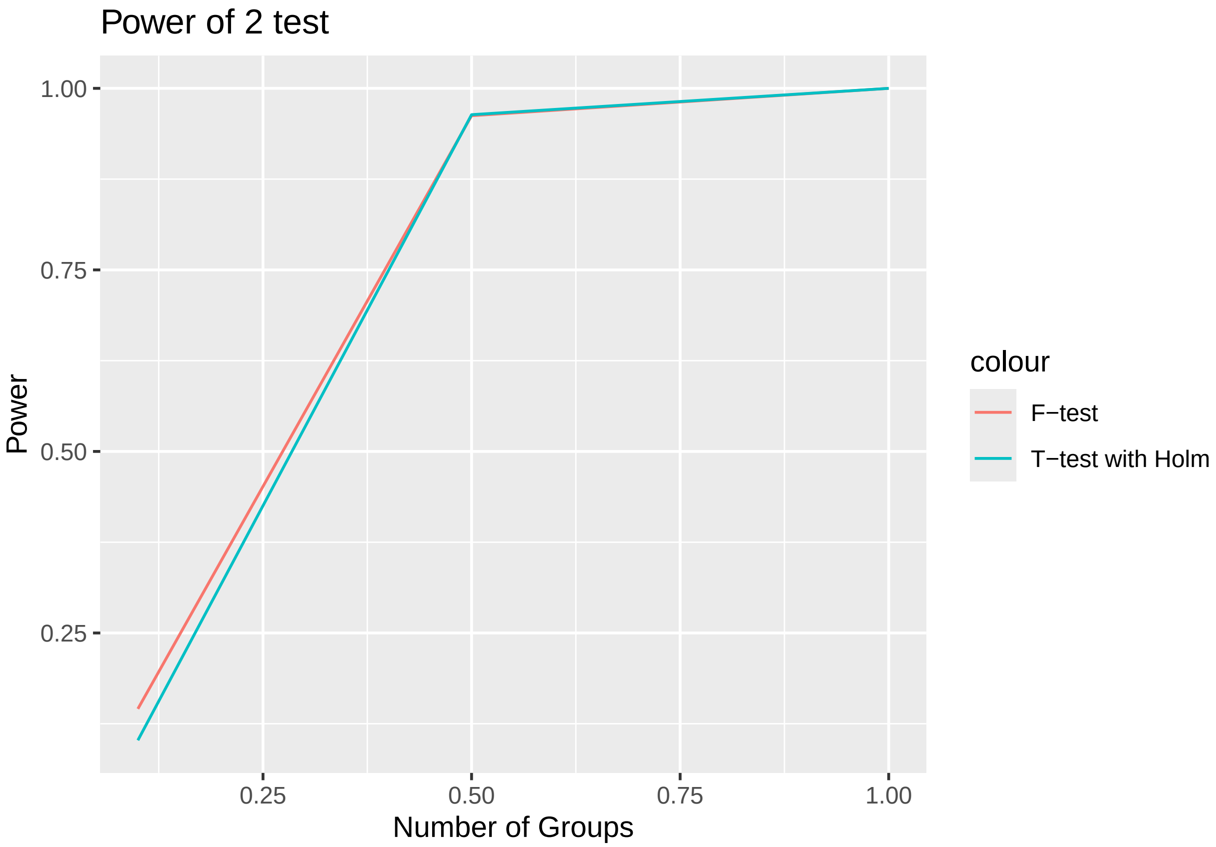
```
## [1] 8
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

```
## [1] 9
```


Power of 2 test
Power
1.00
0.75
0.50
0.25
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm

```
## [1] 10
```


Power of 2 test
1.00
0.75
0.50
0.25
Power
0.25
0.50
0.75
1.00
Number of Groups
colour
F−test
T−test with Holm