Hypothesis Testing

1)One and two sample t tests

One sample t test

Sample example: t.test(x, y = NULL, alternative = c("two.sided", "less", "greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95)

1)We can pass numeric vector data values for x and y. If we include y then it performs two sample test and otherwise it will perform one sample t test.
2)mu indicates true value of mean
3)The option alternative specifies the alternative hypothesis, and must be one of the following: "two.sided" (default), "greater" or "less".

4)The option paired indicates whether you want a paired t-test or not. Default(False)

5)The option var.equal indicates whether or not to assume the two variances as being equal when performing a two-sample t-test. If True pooled variance is used, otherwise Welch approximation to the degrees of freedom is used. Default(False)

Ex. An outbreak of Salmonella-related illness was attributed to ice cream produced at a certain factory. Scientists measured the level of Salmonella in 9 randomly sampled batches of ice cream. The levels (in MPN/g) were:
0.593 0.142 0.329 0.691 0.231 0.793 0.519 0.392 0.418 Is there evidence that the mean level of Salmonella in the ice cream is greater than 0.3 MPN/g?

The hypothesis of interest can be expressed as:
H0: m= 0.3
Ha: m> 0.3
we include the options alternative="greater", mu=0.3.

```
ex1 = c(0.593, 0.142, 0.329, 0.691, 0.231, 0.793, 0.519, 0.392, 0.418)

t.test(ex1, alternative="greater", mu=0.3)
```

```
##
##  One Sample t-test
##
## data:  ex1
## t = 2.2051, df = 8, p-value = 0.02927
## alternative hypothesis: true mean is greater than 0.3
## 95 percent confidence interval:
##  0.3245133      Inf
## sample estimates:
## mean of x
## 0.4564444
```

p value = 0.029. Hence we need to **reject** the null hypothesis.
Hence, there is moderately strong evidence that the mean Salmonella level in the ice cream is above 0.3 MPN/g.

Two sample t test

Two sample t test

Example:6 subjects were given a drug (treatment group) and an additional 6 subjects a placebo (control group). Their reaction time to a stimulus was measured (in ms).
We want to perform a two-sample t-test for comparing the means of the treatment and control groups.
Let 1 be the mean of the population taking medicine and 2 the mean of the untreated population.
Here the hypothesis of interest can be expressed as:
H0: m1- m2=0
Ha: m1- m2<0
Here we will need to include the data for the treatment group in x and the data for the control group in y. We will also need to include the options alternative="less", mu=0.
Finally, we need to decide whether or not the standard deviations are the same in both groups.

```
Control = c(91, 87, 99, 77, 88, 91)
Treat = c(101, 110, 103, 93, 99, 104)

t.test(Control,Treat,alternative="less", var.equal=TRUE)
```

```
##
##   Two Sample t-test
##
## data:  Control and Treat
## t = -3.4456, df = 10, p-value = 0.003136
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.082744
## sample estimates:
## mean of x mean of y
##  88.83333 101.66667
```

As p = 0.003137 we need to reject the null hypothesis and there is a strong evidence that the drug is effective.

```
t.test(Control,Treat,alternative="less")
```

```
##
##   Welch Two Sample t-test
##
## data:  Control and Treat
## t = -3.4456, df = 9.4797, p-value = 0.003391
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -6.044949
## sample estimates:
## mean of x mean of y
##  88.83333 101.66667
```

Here the pooled t-test and the Welsh t-test give roughly the same results
As p = 0.003391 we need to reject the null hypothesis and there is a strong evidence that the drug is effective.

Paired t test

This can be done using the option paired =TRUE.

Ex. A study was performed to test whether cars get better mileage on premium gas than on regular gas. Each of 10 cars was first filled with either regular or premium gas, decided by a coin toss, and the mileage for that tank was recorded. The mileage was recorded again for the same cars using the other kind of gasoline. We use a paired ttest to determine whether cars get significantly better mileage with premium gas.

```
reg = c(16, 20, 21, 22, 23, 22, 27, 25, 27, 28)
prem = c(19, 22, 24, 24, 25, 25, 26, 26, 28, 32)
t.test(prem,reg,alternative="greater", paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  prem and reg
## t = 4.4721, df = 9, p-value = 0.0007749
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.180207       Inf
## sample estimates:
## mean of the differences
##                       2
```

The results show that the t-statistic is equal to 4.47 and the p-value is 0.00075. Since the p-value is very low, we reject the null hypothesis. There is strong evidence of a mean increase in gas mileage between regular and premium gasoline.

2)Test Of Equal Or Given Proportions

prop.test can be used for testing the null that the proportions (probabilities of success) in several groups are the same, or that they equal certain given values.

```
heads <- rbinom(1, size = 100, prob = .5)
prop.test(heads, 100)              # continuity correction TRUE by default
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  heads out of 100, null probability 0.5
## X-squared = 1.21, df = 1, p-value = 0.2713
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3420219 0.5426412
## sample estimates:
##    p
## 0.44
```

```
prop.test(heads, 100, correct = FALSE)
```

```
##
##  1-sample proportions test without continuity correction
##
## data:  heads out of 100, null probability 0.5
## X-squared = 1.44, df = 1, p-value = 0.2301
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3467203 0.5377190
## sample estimates:
##    p
## 0.44
```

smokers <- c(83, 90, 129, 70)
patients <- c(86, 93, 136, 82)

Data from Fleiss (1981), p. 139.
H0: The null hypothesis is that the four populations from which the patients were drawn have the same true proportion of smokers.
HA: The alternative is that this proportion is different in at least one of the populations.

```
smokers  <- c( 83, 90, 129, 70 )
patients <- c( 86, 93, 136, 82 )
prop.test(smokers, patients)
```

```
##
##  4-sample test for equality of proportions without continuity
##  correction
##
## data:  smokers out of patients
## X-squared = 12.6, df = 3, p-value = 0.005585
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.9651163 0.9677419 0.9485294 0.8536585
```

**Here p-value = 0.005585 hence we need to reject the null hypothesis and there is a strong evidence that the proportion is different in atleast one of the populations.

3)F-Test: Compare Two Variances in R

R function
The R function var.test() can be used to compare two variances as follow:

Method 1
var.test(values ~ groups, data, alternative = "two.sided") Method 2
var.test(x, y, alternative = "two.sided")

x,y: numeric vectors
alternative: the alternative hypothesis. Allowed value is one of "two.sided" (default), "greater" or "less".

Example:

```r
my_data <- ToothGrowth

library("dplyr")
```

```
## 
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
## 
##     filter, lag
```
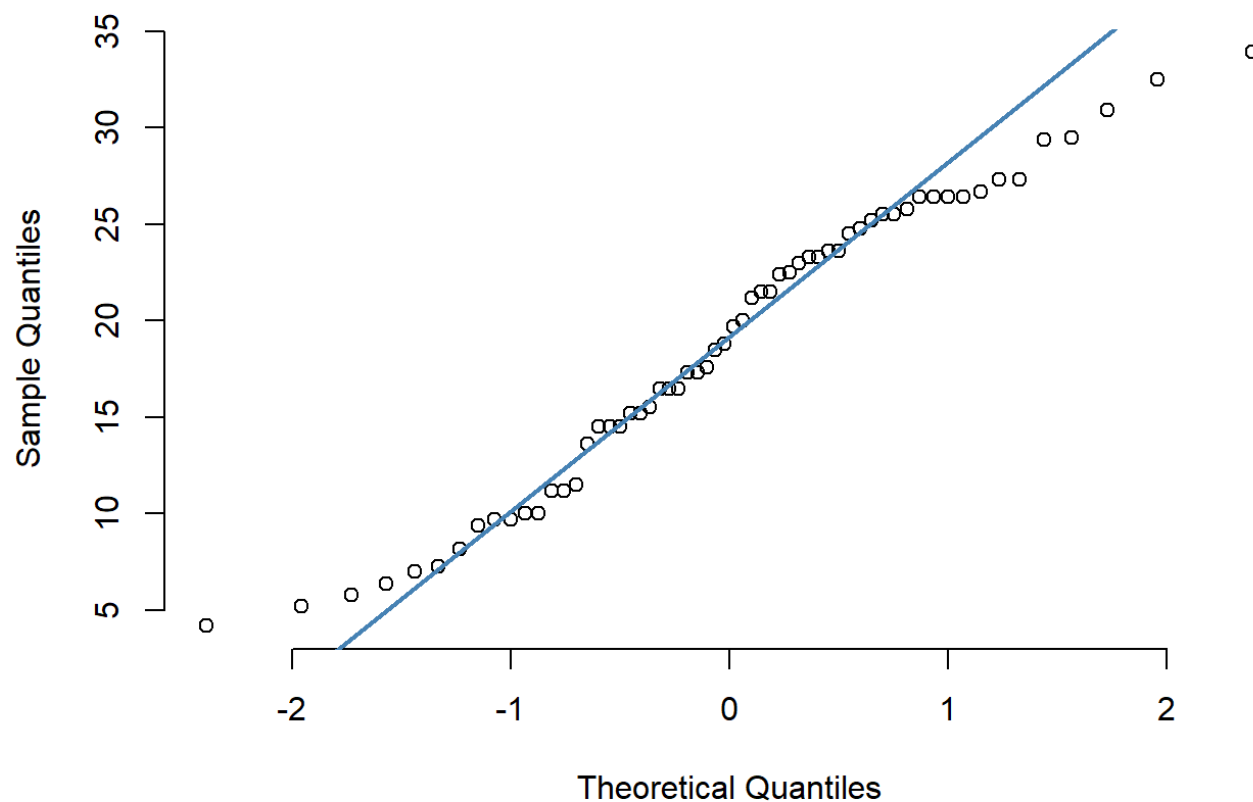
```
## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union
```

```r
sample_n(my_data, 10)
```

```
##     len supp dose
## 13 15.2   VC  1.0
## 54 24.5   OJ  2.0
## 9   5.2   VC  0.5
## 37  8.2   OJ  0.5
## 14 17.3   VC  1.0
## 44 26.4   OJ  1.0
## 55 24.8   OJ  2.0
## 5   6.4   VC  0.5
## 49 14.5   OJ  1.0
## 20 15.5   VC  1.0
```

```r
qqnorm(my_data$len, pch = 1, frame = FALSE)
qqline(my_data$len, col = "steelblue", lwd = 2)
```

Normal Q-Q Plot

We can see the data is normally distributed.

```
res.ftest <- var.test(len ~ supp, data = my_data)
res.ftest
```

```
##
##  F test to compare two variances
##
## data:  len by supp
## F = 0.6386, num df = 29, denom df = 29, p-value = 0.2331
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.3039488 1.3416857
## sample estimates:
## ratio of variances
##           0.6385951
```

```
res.ftest$estimate
```

```
## ratio of variances
##           0.6385951
```

```
res.ftest$p.value
```

```
## [1] 0.2331433
```

The p-value of F-test is p = 0.2331433 which is greater than the significance level 0.05. In conclusion, there is no significant difference between the two variances.

4) Use one of the real world examples (data sets) and at least one of the above tests and R functions to conduct a hypothesis test then report your findings and give proper conclusion(s).

We will use PlantGrowth Dataset to perform Hypothesis testing
Let us do some manipulations on the dataset

```
df1 <- PlantGrowth
df1$ctrl <- df1[1:10,1]
df1$trt1 <- df1[11:20,1]
df1$trt2 <- df1[21:30,1]
Plant <- df1[1:10,c(-1,-2)]
Plant
```

```
##     ctrl trt1 trt2
## 1   4.17 4.81 6.31
## 2   5.58 4.17 5.12
## 3   5.18 4.41 5.54
## 4   6.11 3.59 5.50
## 5   4.50 5.87 5.37
## 6   4.61 3.83 5.29
## 7   5.17 6.03 4.92
## 8   4.53 4.89 6.15
## 9   5.33 4.32 5.80
## 10  5.14 4.69 5.26
```

The experimental hypothesis would be that the treatment affects the plant growth.

H0: The null hypothesis is treatment does not affect plant growth.
HA: The alternative is treatment affects plant growth.

Ploting density distribution for treatment and plant growth

```
library(ggplot2)

t.test(Plant$ctrl,Plant$trt2,var.equal = FALSE)
```

```
##
##   Welch Two Sample t-test
##
## data:  Plant$ctrl and Plant$trt2
## t = -2.134, df = 16.786, p-value = 0.0479
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   -0.98287213 -0.00512787
## sample estimates:
## mean of x mean of y
##      5.032     5.526
```

The p value = 0.047 hence we need to reject the null hypothesis. Hence we can see treatment affects plant growth.

```
denx<-dnorm(Plant$ctrl,mean(Plant$ctrl),sd(Plant$ctrl))
deny<-dnorm(Plant$trt2,mean(Plant$trt2),sd(Plant$trt2))


plot(Plant$ctrl,denx,ylab="Density", type="p",lwd=2, main="Distribution of Plant gr
owth and treatment", ylim = range(0,1.5),xlab = "Plant Treatment 1");par(new=TRUE)

plot(Plant$trt2,deny,ylab="Density", type="p",lwd=2,col = "red", xlab = " ",  ylim
= range(0,1))
```


Distribution of Plant growth and treatment