Assignment 5 (MC 205 Lab)
Aneesh Panchal 2K20/MC/21

ANOVA TEST

- **Code for One Way Test:**

```
dt<-data.frame(id=c(1:10), name=c('a','b','c','d','e','f','g','h','i','j'),
        income=c(100,150,125,200,230,260,250,330,230,400),
        age=c(23,25,24,26,27,29,32,21,30,31),
        stringsAsFactors=FALSE)

one.way<-aov(income~age, data=dt)
one.way
summary(one.way)
```

Results:

```
> dt<-data.frame(id=c(1:10), name=c('a','b','c','d','e','f','g','h','i','j'),
+         income=c(100,150,125,200,230,260,250,330,230,400),
+         age=c(23,25,24,26,27,29,32,21,30,31),
+         stringsAsFactors=FALSE)

> one.way<-aov(income~age, data=dt)

> one.way
Call:
  aov(formula = income ~ age, data = dt)

Terms:
                      age     Residuals
Sum of Squares  13274.25  62088.25
Deg. of Freedom       1         8

Residual standard error: 88.09672
Estimated effects may be unbalanced

> summary(one.way)
```

```
             Df  Sum     Sq Mean Sq F value  Pr(>F)
age          1   13274   13274    1.71          0.227
Residuals    8   62088    7761
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Code for Two Way Test:**

```
dt<-data.frame(id=c(1:10), name=c('a','b','c','d','e','f','g','h','i','j'),
               income=c(100,150,125,200,230,260,250,330,230,400),
               age=c(23,25,24,26,27,29,32,21,30,31),
               stringsAsFactors=FALSE)

two.way<-aov(dt$id~dt$income+dt$age, data=dt)
two.way

summary(two.way)

two.way2<-aov(dt$id~dt$income*dt$age, data=dt)
two.way2<-aov(dt$id~dt$income+dt$age + income:age, data=dt)

summary(two.way2)
```

Results:

```
> dt<-data.frame(id=c(1:10), name=c('a','b','c','d','e','f','g','h','i','j'),
+              income=c(100,150,125,200,230,260,250,330,230,400),
+              age=c(23,25,24,26,27,29,32,21,30,31),
+              stringsAsFactors=FALSE)

> two.way<-aov(dt$id~dt$income+dt$age, data=dt)

> two.way

Call:
  aov(formula = dt$id ~ dt$income + dt$age, data = dt)

Terms:
```

```
                    dt$income   dt$age   Residuals
Sum of Squares    66.13443    3.53923  12.82634
Deg. of Freedom        1           1          7

Residual standard error: 1.353637
Estimated effects may be unbalanced

> summary(two.way)


                Df  Sum    Sq Mean  Sq F value   Pr(>F)
dt$income    1   66.13  66.13      36.093       0.000538 ***
dt$age       1   3.54   3.54       1.932        0.207186
Residuals    7   12.83  1.83
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> two.way2<-aov(dt$id~dt$income*dt$age, data=dt)

> two.way2<-aov(dt$id~dt$income+dt$age + income:age, data=dt)

> summary(two.way2)
                Df   Sum    Sq Mean   Sq F value   Pr(>F)
dt$income    1    66.13    66.13     37.427       0.000871 ***
dt$age       1    3.54     3.54      2.003        0.206740
income:age   1    2.22     2.22      1.259        0.304762
Residuals    6    10.60    1.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Analysis:

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyse the differences among group means in a sample.
The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal.

One-way or two-way refers to the number of independent variables (IVs) in your Analysis of Variance test.

From **1 way Anova test** Result we conclude:

- The p-value of age is 0.227, which indicates little or no real evidence against H_o

From **2 way Anova test** Result we conclude:

- The p-value of income is 0.000871, which indicates very strong evidence against H_o
- The p-value of age is 0.20674, which indicates little or no real evidence against H_o
- The p-value for the interaction between income*age is 0.304762, which indicates little or no real evidence against H_o

CHI SQUARE TEST

Chi Square Test for Independence

- **Code I:**

```
data_<- cbind(x=c(12,23,26,17,9,45),y=c(34,25,41,19,53,33))
data_ <- data.frame(data_)
data_ <- table(data_)
data_
chisq.test(data_)
```

Results:

```
> data_ <- cbind(x=c(12,23,26,17,9,45),y=c(34,25,41,19,53,33))
> data_ <- data.frame(data_)
> data_ <- table(data_)

> data_
   y
x    19 25 33 34 41 53
  9   0   0   0   0   0   1
 12   0   0   0   1   0   0
 17   1   0   0   0   0   0
 23   0   1   0   0   0   0
 26   0   0   0   0   1   0
 45   0   0   1   0   0   0

> chisq.test(data_)

        Pearson's Chi-squared test

data:  data_
X-squared = 30, df = 25, p-value = 0.2243
```

- **Code II:**

```
df <- data.frame(rnorm(10),rnorm(10))
df <- table(df)
df
chisq.test(df)
```

Results:

```
> df <- data.frame(rnorm(10),rnorm(10))
> df <- table(df)

> df
                rnorm.10..1
rnorm.10.          -1.02642090030678 -0.627906076039371 -0.600259587147127 -
0.235700359100477 0.238731735111441 0.54839695950807
 -0.370660031792409              0              0              0              0              1              0
 -0.325931585531227              0              0              0              1              0              0
 -0.220486561818751              0              0              0              0              0              0
 0.00576418589988693             0              0              0              0              0              0
 0.331781963915697               0              0              1              0              0              0
 0.38528040112633                0              0              0              0              0              1
 0.435181490833803               0              0              0              0              0              0
 0.644376548518833               0              1              0              0              0              0
 1.09683901314935                0              0              0              0              0              0
 1.14880761845109                1              0              0              0              0              0
                rnorm.10..1
rnorm.10.          0.993503855962119 1.36065244853001 1.53261062618519
2.18733299301658
 -0.370660031792409              0              0              0              0
 -0.325931585531227              0              0              0              0
 -0.220486561818751              0              1              0              0
 0.00576418589988693             1              0              0              0
 0.331781963915697               0              0              0              0
 0.38528040112633                0              0              0              0
 0.435181490833803               0              0              1              0
 0.644376548518833               0              0              0              0
 1.09683901314935                0              0              0              1
 1.14880761845109                0              0              0              0

> chisq.test(df)

        Pearson's Chi-squared test

data:  df
X-squared = 90, df = 81, p-value = 0.2313
```

Analysis:

A chi-squared test, is a statistical hypothesis test that is valid to perform when the test statistic is chi-squared distributed under the null hypothesis, specifically Pearson's chi-squared test and variants thereof.

Pearson's chi-squared test is used to determine whether there is a statistically significant difference between the expected frequencies and the observed frequencies in one or more categories of a contingency table.

In both of the above cases we can see that the p value is greater than 0.05

For **Code I** we have p value = 0.2243 > 0.05

For **Code II** we have p value = 0.2313 > 0.05

So, in both of the cases x value is independent of y values.

Hence, there is a weak or no correlation between the two variables.

T - TEST

- **One Sample T - Test Code**

```
df1 <- c(rnorm(50, mean = 140, sd = 5))
t.test(df1, mu = 150) # H_o: mu = 150
```

Results:

```
> df1 <- c(rnorm(50, mean = 140, sd = 5))

> t.test(df1, mu = 150) # H_o: mu = 150

        One Sample t-test

data:  df1
t = -16.109, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 150
95 percent confidence interval:
 137.3247 140.1363
sample estimates:
mean of x
138.7305
```

- **Two Sample T - Test Code**

```
df21 <- rnorm(50, mean = 140, sd = 4.5)
df22 <- rnorm(50, mean = 150, sd = 4)

t.test(df21, df22, var.equal = TRUE)
```

Results:

```
> df21 <- rnorm(50, mean = 140, sd = 4.5)

> df22 <- rnorm(50, mean = 150, sd = 4)

> t.test(df21, df22, var.equal = TRUE)

        Two Sample t-test
```

data: df21 and df22
t = -12.256, df = 98, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -11.376672 -8.205902
sample estimates:
mean of x mean of y
140.1746 149.9659

- **Paired Sample T - Test Code**

```
df31 <- c(rnorm(100, mean = 14, sd = 0.3))
df32 <- c(rnorm(100, mean = 13, sd = 0.2))

t.test(df31, df32, paired = TRUE)
```

Results:

```
> df31 <- c(rnorm(100, mean = 14, sd = 0.3))

> df32 <- c(rnorm(100, mean = 13, sd = 0.2))

> t.test(df31, df32, paired = TRUE)

        Paired t-test

data:  df31 and df32
t = 28.449, df = 99, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9783807 1.1250916
sample estimates:
mean of the differences
          1.051736
```

Analysis:

The t-test is any statistical hypothesis test in which the test statistic follows a student's t-distribution under the null hypothesis.

A t-test is the most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.

When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a student's t distribution. The t-test can be used, for example, to determine if the means of two sets of data are significantly different from each other.

- **In One Sample T - Test:**
  The One-Sample T-Test is used to test the statistical difference between a sample mean and a known or assumed/hypothesized value of the mean in the population.

  p-value < 2.2e-16 which indicates very strong evidence against H_o

- **In Two Sample T - Test:**
  It is used to help us to understand that the difference between the two means is real or simply by chance.

  p-value < 2.2e-16 which indicates very strong evidence against H_o

- **In Paired Sample T - Test:**
  This is a statistical procedure that is used to determine whether the mean difference between two sets of observations is zero. In a paired sample t-test, each subject is measured two times, resulting in pairs of observations.

  p-value < 2.2e-16 which indicates very strong evidence against H_o

F - TEST

Fisher's F-Test

- **Code I:**

```
x1 <- rnorm(249, mean = 20)
y1 <- rnorm(79, mean = 30)
var.test(x1, y1, alternative = "two.sided")
```

Results:

```
> x1 <- rnorm(249, mean = 20)

> y1 <- rnorm(79, mean = 30)

> var.test(x1, y1, alternative = "two.sided")

        F test to compare two variances

data:  x1 and y1
F = 0.87092, num df = 248, denom df = 78, p-value = 0.4282
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5960911 1.2293134
sample estimates:
ratio of variances
       0.8709239
```

- **Code II:**

```
x2 = c(25, 29, 35, 46, 58, 66, 68)
y2 = c(14, 16, 24, 28, 32, 35,
    37, 42, 43, 45, 47)
var.test(x2, y2)
```

Results:

```
> x2 = c(25, 29, 35, 46, 58, 66, 68)

> y2 = c(14, 16, 24, 28, 32, 35,
+      37, 42, 43, 45, 47)
```

```
> var.test(x2, y2)

        F test to compare two variances

data:  x2 and y2
F = 2.4081, num df = 6, denom df = 10, p-value = 0.2105
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5913612 13.1514157
sample estimates:
ratio of variances
          2.4081
```

Analysis:

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact "F-tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s.

- **For Code I:**
  Value of the F test statistic: 0.87092
  P value: 0.4282
  Ratio of the sample variances: 0.8709239

  The p-value of F-test is p = 0.4282 which is greater than the alpha level 0.05. In conclusion, there is no difference between the two sample.

- **For Code II:**
  Value of the F test statistic: 2.4081
  P value: 0.2105
  Ratio of the sample variances: 2.4081

  The p-value of F-test is p = 0.2105 which is greater than the alpha level 0.05. In conclusion, there is no difference between the two samples.