ANOVA Analysis

ANOVA also known as Analysis of variance is used to investigate relations between categorical variables and continuous variables.

Sometimes, if we have a categorical variable with values like Yes/No or Male/Female etc.

Example

Consider the R built in data set mtcars. In it we observer that the field "am" represents the type of transmission (auto or manual). It is a categorical variable with values 0 and 1. The miles per gallon value(mpg) of a car can also depend on it besides the value of horse power("hp").

We study the effect of the value of "am" on the regression between "mpg" and "hp". It is done by using the **aov()** function.

Input Data

Create a data frame containing the fields "mpg", "hp" and "am" from the data set mtcars. Here we take "mpg" as the response variable, "hp" as the predictor variable and "am" as the categorical variable.

```
input <- mtcars[,c("am","mpg","hp")]

print(head(input))
```

When we execute the above code, it produces the following result −

```
                      am   mpg   hp
Mazda RX4              1   21.0  110
Mazda RX4 Wag          1   21.0  110
Datsun 710             1   22.8   93
Hornet 4 Drive         0   21.4  110
Hornet Sportabout      0   18.7  175
Valiant                0   18.1  105
```

ANOVA Analysis For:

We create a regression model taking "hp" as the predictor variable and "mpg" as the response variable taking into account the interaction between "am" and "hp".

We create a regression model taking "hp" as the predictor variable and "mpg" as the response variable taking into account the interaction between "am" and "hp".

Model with interaction between categorical variable and predictor variable

```
# Get the dataset.
input <- mtcars


# Create the regression model.
result <- aov(mpg~hp*am,data = input)(direct
related)
print(summary(result))
```

When we execute the above code, it produces the following result −

```
            Df      Sum Sq      Mean Sq      F value      Pr(>F)
hp           1        678.4       678.4        77.391     1.50e-09 ***
am           1        202.2       202.2       23.072      4.75e-05 ***
hp:am        1          0.0       0.0            0.001    0.981
Residuals   28  245.4     8.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This result shows that both horse power and transmission type has significant effect on miles per gallon as the p value in both cases is less than 0.05. But the interaction between these two variables is not significant as the p-value is more than 0.05.

Model without interaction between categorical variable and predictor variable

```
# Get the dataset.
input <- mtcars

# Create the regression model.
result <- aov(mpg~hp+am,data = input)(indirectly related)
print(summary(result))
```

When we execute the above code, it produces the following result −

```
            Df  Sum Sq  Mean Sq   F value   Pr(>F)
hp           1  678.4   678.4   80.15 7.63e-10 ***
am           1  202.2   202.2   23.89 3.46e-05 ***
Residuals   29  245.4     8.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This result shows that both horse power and transmission type has significant effect on miles per gallon as the p value in both cases is less than 0.05.

Z - score

- The z-score is a measure that shows how much away (below or above) of the mean is a specific value (individual) in a given dataset.

- Z-score is the distance of the **raw score value** from the mean in terms of standard deviation.

scale(a,center = TRUE,scale=TRUE)

center,scale both are optional

Default both are TRUE

If center = FALSE
IF scale = FALSE # no mean or SD printed and calculated

A z-score less than 0 represents an element less than the mean.

A z-score greater than 0 represents an element greater than the mean.

A z-score equal to 0 represents an element equal to the mean.

center: if TRUE, the objects column means are subtracted

from the values in those columns (ignoring NAs);

if FALSE, centering is not performed

scale: if TRUE, the centered column values are divided by

the columns standard deviation

if FALSE, scaling is not performed

Example: Method 1

```
# create vector

a <- c(9, 10, 12, 14, 5, 8, 9)
# find mean
mean(a)

# find standard deviation
sd(a)

# calculate z
a.z <- (a - mean(a)) / sd(a)

print(a.z)
```

Method 2

```
# create vector

a <- c(9, 10, 12, 14, 5, 8, 9)

# calculate z
```

T-test

- Used to compare the mean of two independent groups

- Example:

Suppose a businessman with two sweet shops in a town wants to check if the average number of sweets sold in a day in both the stores is the same or not. So, the businessman takes the average number of sweets sold to 15 random people in the respective shops. He found out that the first shop sold 30 sweets on average whereas the second shop sold 40. So, from the owner’s point of view, the second shop was doing better business than the former. But the thing to notice is that the data-set is based on a mere number of random people and they cannot represent all the customers. **This is where T-testing comes into play it helps us to understand that the difference between the two means is real or simply by chance.**

- Classification of T-tests

1. One Sample T-test: used to test the statistical difference between a sample mean and a known or assumed value of the mean

2. Two sample T-test: used to help us to understand that the difference between the two means is real or simply by chance.

3. Paired sample T-test: statistical procedure that is used to determine whether the mean difference between two sets of observations is zero

Example:

```
> male.weight=c(90,91,110,150,152,112,80,90,142,115)
> female.weight=c(110,150,152,142,112,115,80,95,103,163)
```

1. One Sample T-test:

```
> t.test(male,mu=110)
```

2. Two sample T-test:

```
> t.test(male,female,var.equal = TRUE)
```

3. Paired sample T-test:

```
> t.test(male,female,paired=TRUE)
```

F-test

F-test is used to assess whether the variances of two populations (A and B) are equal

variances of populations :it indicates how data points are spread out in the population.

p-value will be used to check the output.
Compare the p-value with 0.05

Function: var.test()
Var.test(vector1, vector2)