ANALYSIS OF VARIANCE - ANOVA

Mohan M J

ANOVA

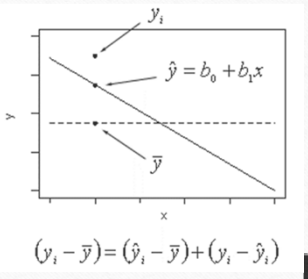
$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

SSTO — Total sum of squares

SSR — Regression sum of squares

SSE — Error sum of squares

$$SSTO = SSR + SSE$$



$$(y_i-\bar{y})=(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)$$

Break down the total variation in *y* ("**total sum of squares**") into two components:

- a component that is "due to" the change in *x* ("**regression sum of squares**")
- a component that is just due to random error ("**error sum of squares**")

If the regression sum of squares is a "large" component of the total sum of squares, it suggests that there *is* a linear association between the predictor *x* and the response *y*

ANOVA

- Analysis of variance is the test of means for two or more populations
- Partitions the total variability in the variable under study to different components
- H0: $Mean_1 = Mean_2 = \ldots\ldots = Mean_k$
- Reject H0 if p value < 0.05
- Use F Distribution

- Example: To study locations of shelf on sales revenue

EXERCISE

An electronics and home appliance chain suspect the location of shelves where television sets are kept will influence the sales revenue. The data on sales revenue in lakhs from the television sets when they are kept at different location inside the store are given in sales revenue data file. The location is denoted as 1: front, 2: middle & 3: rear. Verify the doubt?

- Data is given in Sales_Revenue_Anova.csv
  - Factor : Location
  - Levels : front, middle, rear
  - Response : Sales revenue

PYTHON CODE

```
import pandas as mypd
from scipy import stats as mystats
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm
myData=mypd.read_csv('.\datasets\Sales_Revenue_Anova.csv')
myData
sales=myData.Sales_Revenue
location=myData.Location
```

```
#computing ANOVA table
mymodel=ols('sales ~ C(location)',myData).fit()
anova_table=anova_lm(mymodel)
anova_table
```

	df	sum_sq	mean_sq	F	PR(>F)
C(location)	2.0	11.082715	5.541358	20.099493	0.000057
Residual	15.0	4.135446	0.275696	NaN	NaN

#conclusion is that <0.05 means on an average the revenue changes with location==> location has significant effect on sales revenue

ANOVA Table

- **mean square error (*MSE*)**
- **regression mean square (*MSR*)**

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}.$$

$$MSR = \frac{\sum(\hat{y}_i - \bar{y})^2}{1} = \frac{SSR}{1}.$$

Source of Variation	DF	SS	MS	F
Regression	1	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	$MSR = \frac{SSR}{1}$	$F^* = \frac{MSR}{MSE}$
Residual error	n-2	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$MSE = \frac{SSE}{n-2}$	
Total	n-1	$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2$		

CROSS TABULATION

CROSS TABULATION

- An approach to summarize and identify the relation between two or more variables or parameters
- Describes two variables simultaneously
- Expressed as two way table
- Variables need to be categorical or grouped

EXERCISE 1:

A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1,2,3 representing light, medium and heavy usage respectively. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7point scale (1: unfavorable to 7: very favorable). The data is given in Apparel_Data.csv file

- Does male and female differ in their usage?
- Does male and female differ in their awareness of the brand?
- Does male and female differ in their preference?
- Does higher awareness means higher preference?

PYTHON CODE

```
import pandas as pd
mydata= pd.read_csv('.\datasets\Apparel_Data.csv')
usage = mydata.Usage
gender = mydata.Gender
mytable = pd.crosstab(gender, usage)
```

CHI SQUARE TEST

- Objective:
  - To test whether two variables are related or not
  - To check whether a metric depends on another metric
- Usage:
  - When both the variables (x, y) need to be categorical
- H0: Relation between x & y = 0 or x and y are independent
- H1: Relation between x & y ≠ 0 or x and y are not independent
- If p value < 0.05, then H0 is rejected

EXERCISE:

A branded apparel manufacturing company has collected the data from 50 customers on usage, gender, awareness of brand and preference of the brand. Usage has been coded as 1,2,3 representing light, medium and heavy usage respectively. The gender has been coded as 1 for female and 2 for male users. The attitude and preference are measured on a 7point scale (1: unfavorable to 7: very favorable). The data is given in Apparel_Data.csv file

- Does male and female differ in their usage?
- Does male and female differ in their awareness of the brand?
- Does male and female differ in their preference?
- Does higher awareness means higher preference?

PYTHON CODE

```
import pandas as mypanda
from scipy import stats as mystats
myData=mypanda.read_csv('.\datasets\Apparel_Data.csv')
myData
usage=myData.Usage
gender=myData.Gender
awareness=myData.Awareness
preference=myData.Preference
myTable=mypanda.crosstab(gender,usage)
myTable
# looking at this table relation between usage and gender can be inferred.
#There is a relationship
```

```
mystats.chi2_contingency(myTable) #first table
Out[]:      (6.6701680672268902, 0.035611593514746492, 2,
             array[[ 10.92, 6.24,8.84], [ 10.08,  5.76,  8.16]]))
#here gender and usage is related since p<0.05
```

CORRELATION

CORRELATION

- Correlation analysis is a technique to identify the relationship between two variables
- Type and degree of relationship between two variables

- Statistical Learning – Parametric Methods
- Machine Learning - Non parametric Methods
- Semi Parametric Methods

Correlation Usage:

- Explore the relationship between output characteristic and input process variable
- Output Variable : y - Dependent variable
- Input/Process variable : x - Independent variable
  - Scatter Plot
  - Correlation Coefficient

- Positive Correlation : y increases as x increases and vice versa
- Negative Correlation : y decreases as x increases and vice versa
- No correlation : Random distribution of points

Measure of Correlation: Coefficient of Correlation

- Symbol : r
- Range : -1 to 1
- Sign : Type of Correlation
- Value: Degree of Correlation
- Examples:
  - r = 0.6, 60% positive correlation
  - r = -0.82, 82% negative correlation
  - r = 0, No correlation

EQATIONS

- Sxy – Sum of product of (x - Mean x) and (y - Mean y)
- Sxy = $\sum$ (x - Mean x) * (y - Mean y)
- Sxx = $\sum$ (x - Mean x) ^{2}
- Syy = $\sum$ (y - Mean y) ^{2}
- Correlation Coefficient, r = Sxy / $\sqrt{}$(Sxx * Syy)

EXERCISE 1:

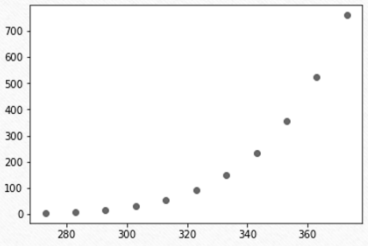
The data on vapor pressure of water at various temperatures are given in Correlation.csv file

- Construct scatter plot and interpret
- Compute the correlation coefficient

PYTHON CODE

```
import pandas as mypandas
import numpy as mynp
import matplotlib.pyplot as myplot
myData=mypandas.read_csv('.\datasets\Correlation.csv')
myData

temperature=myData.Temperature
vaporPressure=myData.Vapor_Pressure
myplot.scatter(temperature,vaporPressure)
myplot.show()
mynp.corrcoef(temperature,vaporPressure)
```



THANK YOU