Data Analytics using SPSS   
7th Semester 2019

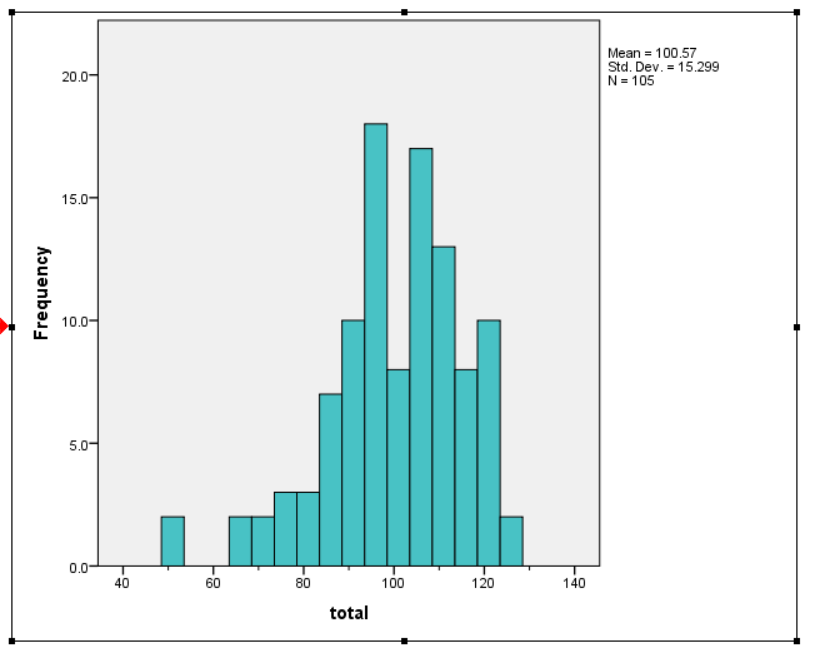
INSTRUCTOR: Mr Satish Kumar

horizontal line

# 10 July 2019

**Graphs**

## Histogram:

Histograms are Bar-plots with only one axis with a value.

This is used when we need to compare data with respect to one axis i.e. the X- axis.

## 

## 

## 

## 

## 

## 

## 

## 

## Box Plots:

## 

The Box plot has *Outliers*, and the *box* itself.

This plot is able to denote the variation of the x-axis variable along the y-axis variable.

## Scatter Plot:

Check the Internet.

# Probability

## Types of Probability:

1. **Subjective Probability :**

Probability got based on **less** or **different** information ( information is not unanimous through-out all the members calculating probability.

1. **Relative Probability :**

It is a statistical concept that measures the likelihood (probability) of something happening. In a classic sense, it means that every statistical experiment will contain elements that are equally likely to happen ( equal chances of occurrence of something ).

Eg: Tossing of a coin.

1. **Classical Probability:**

is the statistical concept that measures the likelihood (probability) of something happening. In a classic sense, it means that every statistical experiment will contain elements that are equally likely to happen (equal chances of occurrence of something).

## Additive and Multiplicative Rule of Probability

**Dependent and independent events**

**Compliment of probability**

## 

## 24/07/19

## Joint, Marginal and Conditional Probability

|  |  |  |
| --- | --- | --- |
|  | **Independent Events** | **Dependent Events** |
| Marginal Probability: ***Of A*** | **P(A)** | **P(AB) + P(AC) + P(AD)+...** |
| Joint Probability: ***Of A & B*** | **P(A) x P(B)** | **P(A) x P(B/A)** |
| Conditional Probability ***Of B given A*** | **P(B/A) = P(B)** |  |

## Bayes Theorem: ( P(B) is the summation of denominators in the equation in (2) )

## 

## Probability Distributions:

### How values are distributed from the middle value of the dataset.

### **Theoretical Frequency distribution**. *( because generally the frequency of values got is different from the frequencies of values got when the actual experiment is done. )*

### Helps in making inferences i.e.

## Types of Probability Distributions

### Discrete probability distribution *( dataset has discrete values in the dataset hence finding patterns is difficult )*

### Continuous probability distribution *( used when dataset has continues data values)*

### **Bernoulii’s Distribution:**

Where

*n ---> Total number of trials*

*X ---> Number of success*

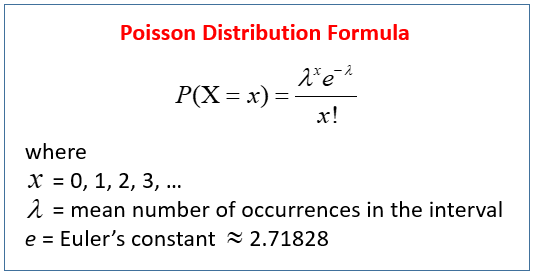
*P ---> Probability of success*

*Q ---> Probability of Failure*

Binomial Dristribution

##### Kindly refer to internet for study materials

Poison Distribution:



**Question ( Numerical )** The above problem is solved using **poisson distribution**.

## Normal Distribution:

* Used when data is continuous & symmetric ( 50% data is below average and 50 % above average, therefore the mean, mode are same )
* The Bell curve formed from such a distribution is always X ( data points) Vs Frequency ( Y-Axis )
* Area under the curve is the probability.
* Generally used when Probability has to be found when the units used are all different.

##### Normal Distribution can be identified by the ***Mean*** and ***Standard Deviation***

* **Z-Score** aka (***Standard Normal Variate****)* : is the number of Standard Deviations from the mean.

x → The data point

→ The mean

→ The Standard Deviation

On finding the z-Score value, calculate probability using the Z-Table( based on whether it is +ve or -ve ) in the following website: <http://www.z-table.com/>

**Empirical rule:**

If we have a normal distribution then The **empirical rule** states three parts: 68% of the data falls within the first standard deviation from the mean. 95% fall within two standard deviations. 99.7% fall within three standard deviations.

**Sampling Distribution:**

* Central limit Theorem is very important for Sampling distribution.

Here all the variables are wrt sample

→ is the standard error of the mean

Because the sampling distribution uses a sample that is biased therefore an interval / range is used because the sample chosen cannot reflect conclusions about the majority of the data from which the sample is received.

* **Population parameter** ( population mean, population variance, population Standard deviation, population proportion ) Sampling statistics are used here ( search the internet to find out how ? )

**Estimation:**

**Hypothesis Testing:**

Testing assumptions about population parameters

* **Significance level (** symbol: alpha **)** is required for hypothesis testing.
* **Statement of the hypothesis:** is the statement that depends on the significance level ( ) + / - hypothesis value
* Based on adding / subtracting the hypothesis value we get Alternate / Null Hypothesis.
* The **Null hypothesis** means that there is no error with respect to the hypothesis value ( Hence out hypothesis was correct. )
  + , *H → original Hypothesis value*
* **Alternate Hypothesis** means that there was some error with respect to the original hypothesis value OR that we got a different value in comparison to the hypothesis value.
* In hypothesis testing there are two types of errors that can be got
  + Type 1Error : ( aka alpha ) ( Suppliers’s Mistake wrt supplying antibiotics ) True
  + Type 2 Error: ( aka Beta ) ( Consumer's Mistake wrt supplying antibiotics ) False
* 2 tail test ( error can happen on both sides )
* 1 tail test ( error can happen only on one side )
* Critical values is got from the Z- table ( it helps to know the range where the is accepted or rejected.
* Test Statistics ( Z ) ( **used when sample size > 30** )
* → sample statistic
* **T- Test** is used for the below conditions:
  + Value taken from a **T-Table**
* Population sample size < 30
* Population standard deviation is not known
* **Z- test** is used when
  + Population sample > 30

**STEPS in HYPOTHESIS Testing**

* Set the Null and alternate hypothesis
* Decide on a significance level ( assume it )
* Calculate probability value using the Z- test
  + In SPSS you will get 2 tail test ( compare the significance value.. If calculated **Sig value > Significance value** then ACCEPT the null Hypothesis.
  + For 1 tail test ( you will divide both the calculated and tested significance value, then compare as before then either ACCEPT or Reject the sample test. )

**Left Tail: mew < = test\_value**

**Right Tail: mew > test\_value**

**TWO SAMPLE TEST:**

* Independent Sample Test
* Pair Sample Test : Two readings are taken in the same sample
  + One **before** intervention
  + One **after** intervention

This is done when we want to know performance in comparison to before and after a particular timestamp in the timeline ( where some meaningful work or event had occurred ). Eg: There is a training program where before joining you take a test ( just to know the person’s skills before joining the training program ). Then during the training program you take classes and teach the person all the required skills wrt to the test ( that would take place after the program ). Based on this we can conclude If there was an improvement after the training program or not.

**Binomial Test**

It is a

**Chi Squared Test**

* AKA Test of Homogeneity
* The test is applied to a single categorical variable from two or more different populations. It is used to determine whether frequency counts are distributed identically across different populations.
* Use only **categorical data**
  + In SPSS do a cross-tab and in that check the chi-Squared test button in statistics. Then fill the parameters that are similar to the hypothesis testing.

**F-Test**

Read from the internet / TextBook.

* Not to be done when Data Parameters are Categorical Data types.
* Done when there are multiple categories or populations, to find the relation between the different population and the dependent variable. ( Eg: If there are many ethnicities ( black, asian…) & total Marks then we can find out how much marks each ethnicity group population gets i.e are they getting the same average marks or not.

**Regression** ( defined as the relationship between a dependent variable

( Y ) and one or more independent variable.(x1, x2, .. ) )

Y = mx + c ( Regression Line Equation )

Y = mx1 + mx2 + ….. + c ( Multivariate Linear Regression )

M, C are called the parameters of the model

M → can also be the regression coefficient or the slope.

Error / Residual is the change / difference between the actual and predicted values on the regression.

* When doing a multivariate linear regression there is a need to standardise the regression coefficient which is represented by **𝜷**
* Hence the new equation being

**Assumptions made initially wrt regression**

* Data has to be normally distributed ( If data is not normal then it is transformed by taking log(datapoint), 1/ ( datapoint ) or so on. *( {****For a single Variable****} can be checked if normally distributed based on a histogram plotted for that variable ) {* ***for a multivariate*** *}**you check the error ( difference between the actual and the predicted value)*
* Values are co-linear or not.
* Residuals should be independent ( checked using the **Durbin Watson Test** )
* You cannot use a linear regression line for a data-set with data points having a curved graph when plotted on a graph.

Heteroscedasticity: Error variance is not constant ( like a funnel )

Homoscedasticity : error variance is constant

The error should be independent of each other

For that the **Durbin Watson Test** is done ( done to check if there is any **auto-corelation** { there should not be some pattern in the error values } ) the closer the value to 2 the more independent the errors are.

R^2 is the coefficient of Determination in the result of regression ( in SPSS ) { explained variance / total variance )

R → Corelation coefficient

* *\* Higher R^2 value then Higher prediction*

**Variance Inflation Factor** : It is a value based on which it is decided whether to remove or hold onto the datapoint.

**Methods to Determine the columns (variables) to consider for model prediction**

* **Forward method**: Starting from 1 variable and increasing variables and checking if model becomes accurate (
* **Backward Method**: starting with all variables and removing variables based on if they add to the accuracy of the model.
* **Step - Wise Method:** Similar to forward method but when new variables are added to the model then based on if the previous variables remain dependent / important model, the variables are retained or removed.

**Logistic Regression:** You know it

Types of Analysis

Factor Analysis, EDA, Principal Component analysis, Common Factor analysis

Steps involved

1. Start with Corelation matrix
2. Run Hypothesis test to extract factors with ( Kaiser- Meyer sampling OR Scree test
3. Decide on a number of factors based on Eigen-values or any of the other 3 methods
4. Component matrix table & Rotation matrix ( done for interpretation only )
5. Look at loadings and decide on the final list of factors.