Unit 2

1) Introduction to Single variable:

* Exploratory data analysis is classified into two types they are graphical or non-graphical. and then each method is either univariate, bivariate, multivariate.

* univariate analysis is simplest analysis of statistical data.

* The term univariate analysis refers to the analysis of one variable.

* It doesn't ~~deal~~ deal with causes or relationships and it's major purpose is to describe data.

*univariate analysis can be performed in two types of statistics. they are

* Descriptive statistics
* Inferential statistics

*Descriptive statistics.

* Descriptive statistics are used to describe data.
* The statistics are commonly used to summary statistics
* Descriptive statics can be used for calculating things like missing value proportions, upper and Lower limits for outliers, level of variance through coefficient of variance.

Inferential Statistics:

∝ The data is dealing with a subset of the complete data.

Univariate testing:

1. Z Test – Used for numerical data
   – The sample size is greater than 30
   – Standard deviation is known.

2. One-Sample-t-test – Used for numerical data
   – ~~Used~~ Sample size is greater than 30
   – Standard deviation is known.

3) Chi-Square Test – Used with ordinal categorical data.

4) Kolmogorov-Smirnov Test – Used with nominal categorical data.

methods for performing univariate.

* The common methods used to perform univariate analysis are

1) Summary statistics
2) Frequency distributions
3) charts
4) univariate tables.

1) Summary statistics:

* It is most common way to perform the univariate analysis.

measures of tendency:

* These values describe the middle value of the dataset
* It is used to locate middle value.
* The mean, median mode are the examples for measure of central tendency.

* Dispersion measures:

* These numbers describe how evenly the values are distributed in dataset.

* The range, standard deviation and variance are some examples.

* Measure of Shape:

* It describes the shape of distribution.

2) Frequency distributions,

* A frequency distribution describes how frequently different values occur in a dataset.

* This is a another way to Perform Univariate analysis.

3) Charts:

Histogram
Boxplot
Density curve
Bar chart
Pie chart

4) univariate table:

Frequency table:

* Each unique value and it's respective frequency in data are shown through a table.

Grouped table:

* Count of each unique values are binned or grouped

percentage:

Rather than showing frequency of unique value of group - wise show propotation of table.

Cummaltive table:

* It is smilar to propotion table. It is used with binned data.

Univariate example:

* Salary of employees.
* Heigh of ten students
* weigh of 20 cats.
* average height of imen in country

2) Distribution and variables:

It is of two types they are categorical or numerical.

Categorical Data:

* Categorical data classify items into groups. This type of data can be further broken down into nominal, ordinal and binary values.

Nominal:

values have no set order

eg: gender alignment

Ordinal data:

* Values have a set order.

eg: Ranking low to high.

Binary data:

* Binary data has only two values

* This could be represented as true / false or 0/1

Numerical data:

* Numerical data are values that can perform mathematical operations.

* This is of two types they are descrete and continous.

Discrete: No of students in class

Continous: Height and weight

* Numerical data can be visulaized with histogram

eg: Age, income, debt

3. Numerical Summaries of Level and Spread:

* A numerical summary is a number used to describe the characteristics of data set

Center - mean, median, mode.

Quantiles - percentiles, five number sum

Spread : Standard deviation, variance
interquartile range

Outliers

Shape - Skewness, Kurtosis

Co Cordance - Correlation, quantile pl

mean:

* Adds all number or data values and divides by total.

* It is represented by 'X-bar'

$$\bar{x} = \frac{\sum_{i}^{n} x_i}{n}$$

median:

middle point of data values.

$$\text{median} = \frac{x_{(n/2)} + x_{(n+2+1)}}{2}$$

* when n is even there is no middle point

* when n is odd, the middle data point is median.

mode:

It returns the most commonly occurring value in data value.

percentile:

It describe the percentage of data value.

* Five number summary.

1). Minimum

2) 25th percentile (lower quartile)

3) 50th percentile (median)

4) 75th percentile (upper quartile)

5) Maximum.

* Standard deviation.

* The idea is use the mean as a reference point and observation point

** It is used to calculate distance between two points.

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

Range:

* The difference between the maximum and minimum values.

$$x_{(n)} - (x)_{(1)}$$

Variance:

* It is similar to standard deviation

* It represented by S Square

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - x)^2}{n-1}$$

4) Standardization and Normalization:

$$x_{norm} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

* finding normalization using the above formula.

* All the value will be in range of 0 to 1