DATA ANALYTICS

Unit 1:Data Exploration

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 1:Data Exploration

Mamatha H R

Department of Computer Science and Engineering

What is data exploration?

A preliminary exploration of the data to better understand its characteristics.

- Key motivations of data exploration include
  - Helping to select the right tool for preprocessing or analysis
  - Making use of humans’ abilities to recognize patterns
    - People can recognize patterns not captured by data analysis tools

Techniques Used In Data Exploration

- In EDA, as originally defined by Tukey
  - The focus was on visualization
  - Clustering and anomaly detection were viewed as exploratory techniques

- In our discussion of data exploration, we focus on
  - Summary statistics
  - Visualization

Summary Statistics

- Summary statistics are numbers that summarize properties of the data

  - Summarized properties include frequency, location and spread
    - Examples: location - mean
      spread - standard deviation

  - Most summary statistics can be calculated in a single pass through the data

Data Type

- Cross-Sectional Data: A data collected on many variables of interest at the same time or duration of time is called cross-sectional data.

- Time Series Data: A data collected for a single variable such as demand for smartphones collected over several time intervals (weekly, monthly, etc.) is called a time series data.

- Panel Data: Data collected on several variables (multiple dimensions) over several time intervals is called panel data (also known as longitudinal data).

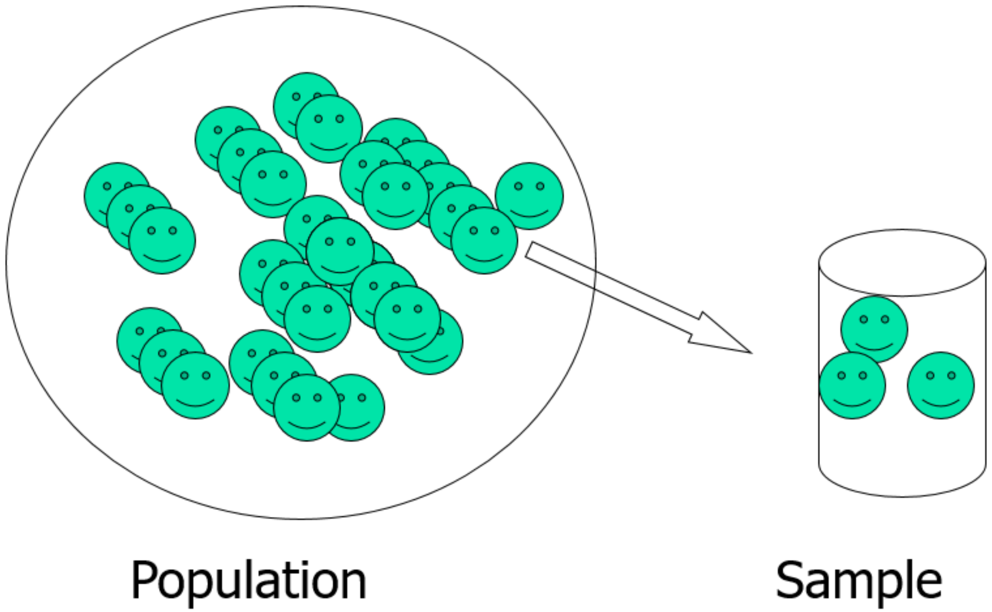
TYPES OF DATA MEASUREMENT SCALES

- Nominal scale refers to variables that are basically names (qualitative data) and also known as categorical variables.
- Ordinal scale is a variable in which the value of the data is captured from an ordered set, which is recorded in the order of magnitude.
- Interval scale corresponds to a variable in which the value is chosen from an interval set. Variable such as temperature measured in centigrade) or intelligence quotient (IQ) score are examples of interval scale
- Any variable for which the ratios can be computed and are meaningful is called ratio scale.

Population And Sample

- Population is the set of all possible observations (often called cases, records, subjects or data points) for a given context of the problem.

- Sample is the subset taken from a population.



An Illustration: Which Group is Smarter?

Class A--IQs of 13 Students		Class B--IQs of 13 Students	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

Each individual may be different. If you try to understand a group by remembering the qualities of each member, you become overwhelmed and fail to understand the group.

Which group is smarter now?

Class A--Average IQ	Class B--Average IQ
110.54	110.23

They're roughly the same!

With a summary descriptive statistic, it is much easier to answer our question.

Types of descriptive statistics:

- Organize Data
  - Tables
  - Graphs

- Summarize Data
  - Central Tendency
  - Variation

Types of descriptive statistics:

- Organize Data
  - Tables
    - Frequency Distributions
    - Relative Frequency Distributions
  - Graphs
    - Bar Chart or Histogram
    - Stem and Leaf Plot
    - Frequency Polygon

Summarizing Data:

- Central Tendency (or Groups' "Middle Values")
  - Mean
  - Median
  - Mode

- Variation (or Summary of Differences Within Groups)
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation

- Mean (or Average) Value

Mean is the arithmetical average value of the data and is one of the most frequently used measures of central tendency.

$$\text{Mean}=\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Mean

Symbol $\bar{x}$ is frequently used to represent the estimated value of the mean from a sample. If the entire population is available and if we calculate mean based on the entire population, then we have the population mean which is denoted by μ (population mean).

Property of Mean

An important property of mean is that the summation of deviation of observations from the mean is zero, that is

$$\sum_{i=1}^{n}\left(X_i - \bar{X}\right) = 0$$

Median (or Mid) Value

- Median is the value that divides the data into two equal parts, that is, the proportion of observations below median and above median will be 50%.

- Easiest way to find the median value is by arranging the data in the increasing order and the median is the value at position $(n + 1)/2$ when n is odd. When n is even, the median is the average value of $(n/2)^{th}$ and $(n + 2)/2^{th}$ observation after arranging the data in the increasing order.

Mode

- Mode is the most frequently occurring value in the dataset

- Mode is the only measure of central tendency which is valid for qualitative (nominal) data since the mean and median for nominal data are meaningless.

- For example, assume that a customer data with a retailer has the marital status of customer, namely, (a) Married, (b) Unmarried, (c) Divorced Male, and (d) Divorced Female. Mean and median are meaningless when we try to use them on a qualitative data such as marital status. On the other hand, mode will capture the customer type in terms of marital status that occurs most frequently in the database

Percentile

- Percentile, decile and quartile are frequently used to identify the position of the observation in the dataset.
- Percentile, denoted as P_x, is the value of the data at which x percentage of the data lie below that value

$$\text{Position corresponding to } P_x \approx x\,(n+1)/100$$

- P_x is the position in the data calculated , where n is the number of observations in the data.

Decile and Quartile

- Decile corresponds to special values of percentile that divide the data into 10 equal parts. First decile contains first 10% of the data and second decile contains first 20% of the data and so on.

- Quartile divides the data into 4 equal parts. The first quartile (Q_1) contains first 25% of the data, Q_2 contains 50% of the data and is also the median. Quartile 3 (Q_3) accounts for 75% of the data

Example

Time between failures of wire-cut (in hours)

2	**22**	**32**	**39**	**46**	**56**	**76**	**79**	**88**	**93**
3	24	33	44	46	66	77	79	89	99
5	24	34	45	47	67	77	86	89	99
9	26	37	45	55	67	78	86	89	99
21	31	39	46	56	75	78	87	90	102

1. Calculate the mean, median, and mode of time between failures of wire-cuts
2. The company would like to know by what time 10% (ten percentile or P_{10}) and 90% (ninety percentile or P_{90}) of the wire-cuts will fail?
3. Calculate the values of P_{25} and P_{75}.

Solution

1) Mean = 57.64, median = 56, and mode = 46

2) Note that the data in Table is arranged in increasing order in columns. The position of $P_{10} = 10 \times (51)/100 = 5.1$. We can round off 5.1 to its nearest integer which is 5. The corresponding value from table is 21 (10 percentage of observations in Table have a value of less than or equal to 21). That is, by 21 hours, 10% of the wire-cuts will fail. In asset management (and reliability theory), this value is called P_{10} life.

Contd..

Instead of rounding the value obtained from Eq, we can use the following approximation:

$$P_{10} = 10 \times (51)/100 = 5.1$$

Value at 5th position is 21. Value at position 5.1 is approximated as 21 + 0.1 × (value at 6th position – value at 5th position) = 21 + 0.1(1) = 21.1

$$P_{90} = 90 \times 51/100 = 45.9$$

The value at position 45 is 90 and at position 45.9 is

$$90 + 0.9 \times (3) = 92.7$$

That is, 90% of the wire-cuts will fail by 92.7 hours

Contd..

P_{25} (1st Quartile or Q_1) = 25 × 51/100 = 12.75 , Value at 12th position is 33, so

P_{25} = 33 + 0.75 (value at 13th position – value at 12th position) = 33 + 0.75 (1) = 33.75

P_{75} (3rd Quartile or Q_3) = 75 × 51/100 = 38.25

Value at 38th position is 86, so

P_{75} = 86 + 0.25 (value at 39th position – value at 38th position) = 86 + 0.25 (0) = 86

Measures of Variation

- Predictive analytics techniques such as regression attempt to explain variation in the outcome variable (Y) using predictor variables (X)
- Variability in the data is measured using the following measures:
  - Range
  - Inter-Quartile Distance (IQD)
  - Variance
  - Standard Deviation

Range, IQD and Variance

- Range is the difference between maximum and minimum value of the data. It captures the data spread.
- Inter-quartile distance (IQD), also called inter-quartile range (IQR) is a measure of the distance between Quartile 1 (Q_1) and Quartile 3 (Q_3)
- Variance is a measure of variability in the data from the mean value. Variance for population, σ^2, is calculated using

$$\text{Variance} = \sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}$$

Sample Variance

- In case of a sample, the Sample Variance (S^2) is calculated using

$$S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$$

- While calculating sample variance S2, the sum of squared
- deviation $\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ is divided by (n-1), this is known as
- Bessel's correction.

Standard Deviation

- The population standard deviation (σ) and sample standard deviation (S) are given by

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{n}} \qquad S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}}$$

Degrees of Freedom

- Degrees of freedom is equal to the number of independent variables in the model (Trochim, 2005). For example, we can create any sample of size *n* with mean value of $\bar{X}$ by randomly selecting (n – 1) values. We need to fix just one out of *n* values. Thus the number of independent variables in this case is (n – 1)

- Degrees of freedom is defined as the difference between the number of observations in the sample and number of parameters estimated (Walker 1940, Toothaker and Miller, 1996). If there are *n* observations in the sample and *k* parameters are estimated from the sample , then the degrees of freedom is (n – k).

Chebyshev's Theorem

- Chebyshev's theorem (also known as Chebyshev's inequality) is an empirical rule that allows us to predict proportion of observations that is likely to lie between an interval defined using mean and standard deviation. Probability of finding a randomly selected value in an interval defined by $\mu \pm k\sigma$ is $1-\frac{1}{k^2}$ that is

$$P(\mu - k\sigma \le X \le \mu + k\sigma) \ge 1 - \frac{1}{k^2}$$

Example

- Amount spent per month by a segment of credit card users of a bank has a mean value of 12000 and standard deviation of 2000. Calculate the proportion of customers who are spending between 8000 and 16000?

- **Solution:**

$$P(8000 \le X \le 16000) = P(\mu - 2\sigma \le X \le \mu + 2\sigma) \ge 1 - \frac{1}{2^2} = 0.75$$

That is, the proportion of customers spending between 8000 and 16000 is at least 0.75 (or 75%)

Measures of Shape – Skewness and Kurtosis

- Skewness is a measure of symmetry or lack of symmetry. A dataset is symmetrical when the proportion of data at equal distance (measured in terms of standard deviation) from mean (or median) is equal. That is, the proportion of data between μ and μ - $k\sigma$ is same as μ and μ+ $k\sigma$, where k is some positive constant.
- **Pearson's moment coefficient of skewness** for a dataset with *n* observations is given by

$$g_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3 / n}{\sigma^3}$$

- The value of g_1 will be close to 0 when the data is symmetrical. A positive value of g_1 indicates a positive skewness and a negative value indicates **negative skewness.**

Skewness

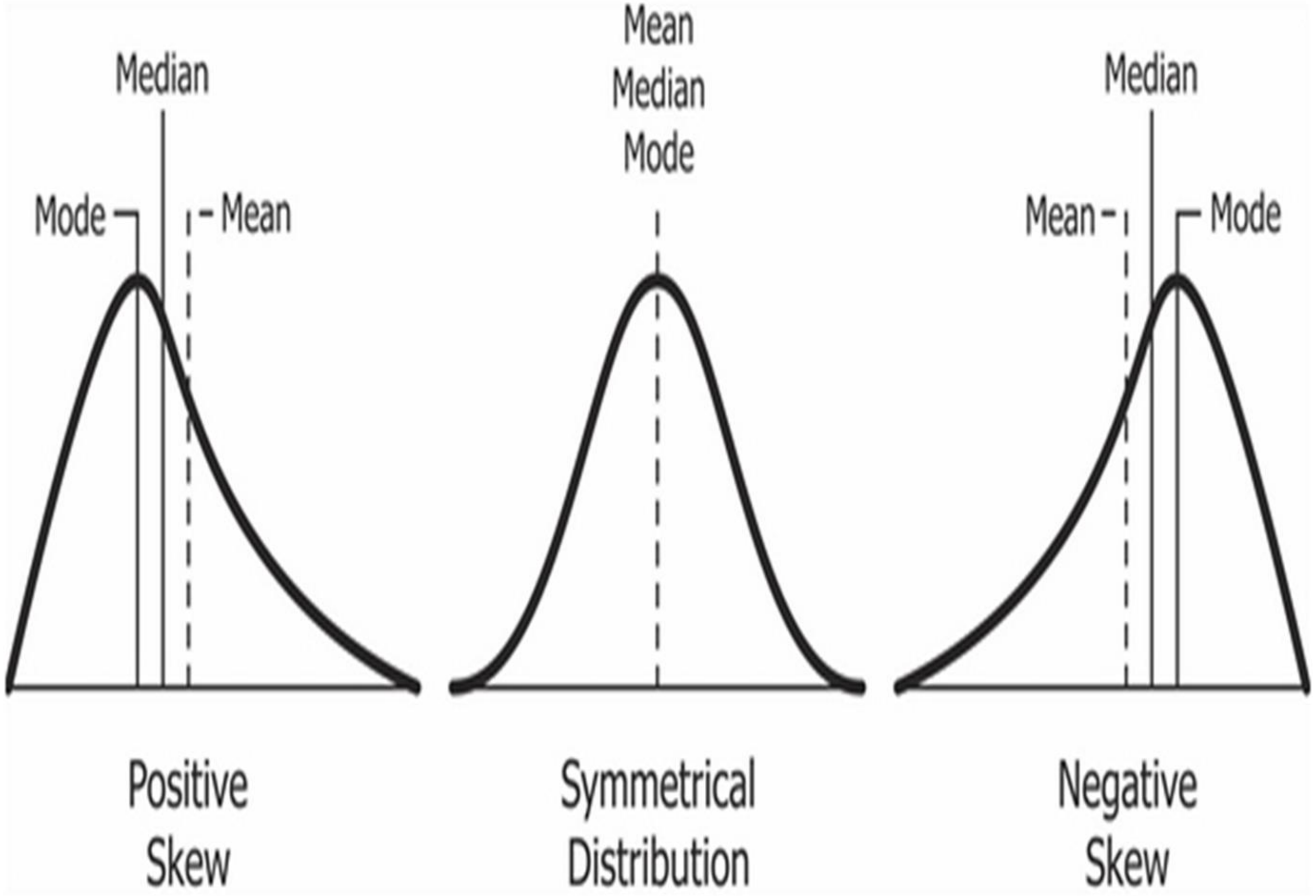
- The following formula is used usually for a sample with *n* observations (Joanes and Gill, 1998):

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

- The value of $\frac{\sqrt{n(n-1)}}{n-2}$ will converge to 1 as the value of *n* increases.

Symmetric vs. Skewed Data

- Median, mean and mode of symmetric, positively and negatively skewed data



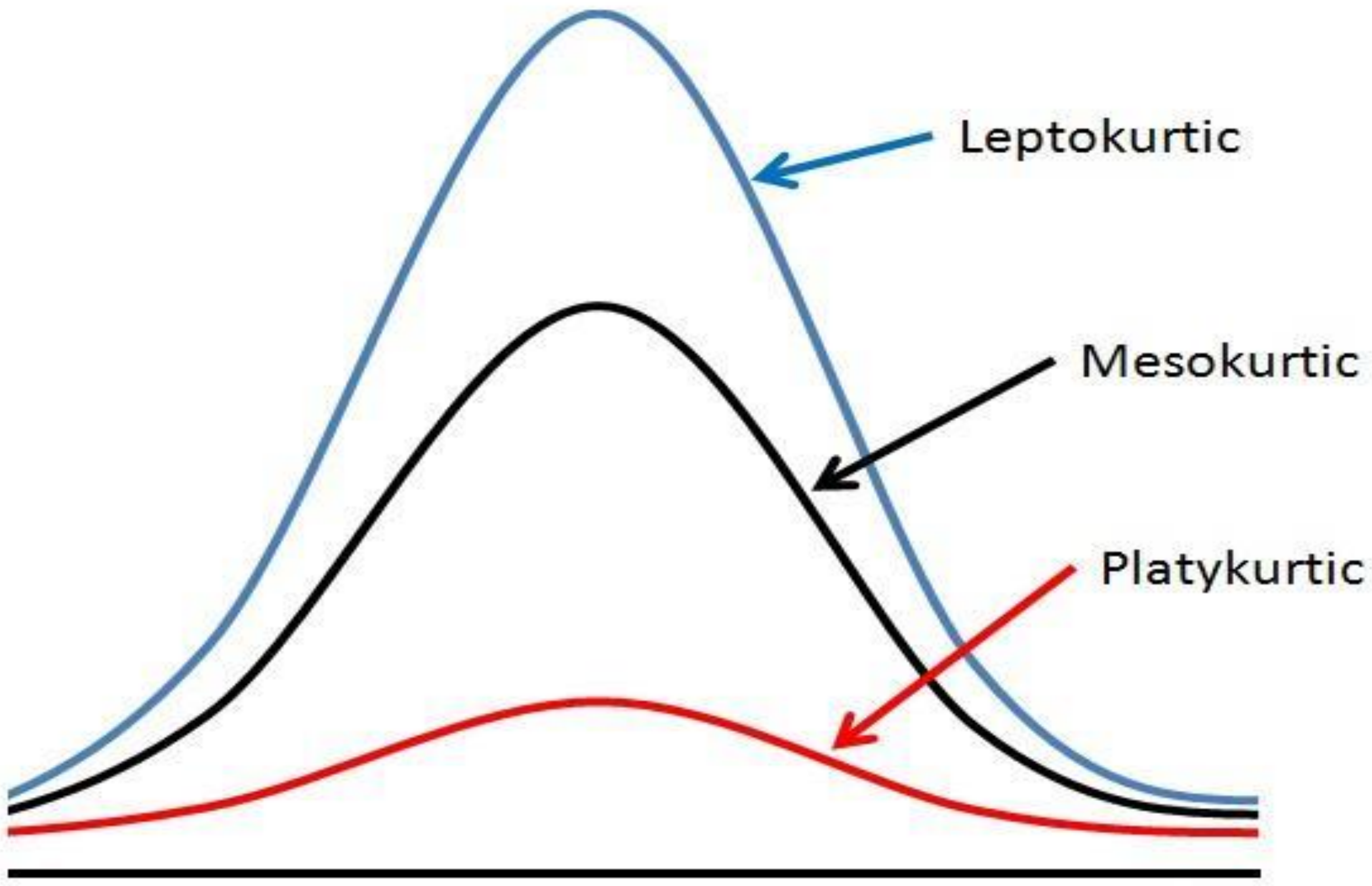
Kurtosis

- Kurtosis is another measure of shape, aimed at shape of the tail, that is, whether the tail of the data distribution is heavy or light. Kurtosis is measured using the following equation:

$$\text{Kurtosis} = \frac{\sum_{i=1}^{4}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4}$$

- Kurtosis value of less than 3 is called platykurtic distribution and greater than 3 is called leptokurtic distribution. The kurtosis value of 3 indicates standard normal distribution (also called **mesokurtic**)

Leptokurtic, mesokurtic, and platykurtic distributions



Excess Kurtosis

- The excess kurtosis is a measure that captures deviation from kurtosis of a normal distribution and is given by:

Excess Kurtosis= $$\frac{\sum_{i=1}^{4}\left(X_i - \bar{X}\right)^4 / n}{\sigma^4} - 3$$

Exercise

The daily football at a retail store in Bangalore over the last 30 days is shown in Table 1. calculate the Mean, Median , Mode and Standard Deviation.

Table 1.Footfall data

232	277	261	173	283	197	251	212	213	213
229	164	219	196	186	247	244	269	216	272
252	314	161	165	221	260	219	290	225	251

For the data in Table 1, calculate the skewness and kurtosis. what can you infer from the skewness and kurtosis of the football data?

For the data in Table 1, calculate the values of first quartile and third quartile. Are there any outliers in the data?

References

Text Book:

- "Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017
- Data Mining: Concepts and Techniques by Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems, 3rd Edition.
- Introduction to Data Mining , Tan, Steinbach, Kumar, 2nd Edition

THANK YOU

Dr.Mamatha H R
Professor,Department of Computer Science
mamathahr@pes.edu

+91 80 2672 1983 Extn 834