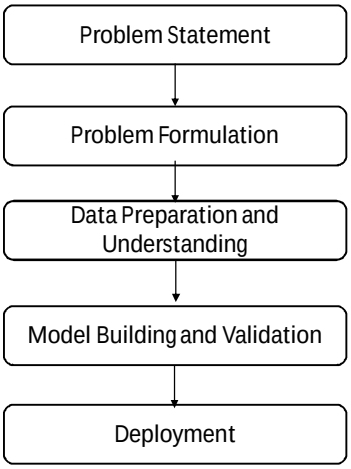
PREDICTIVE MODELING

MOHAN M J

PREDICTIVE MODELING

- Set of methods to arrive at quantitative solutions to problems of business interest
- Predictive modeling is a process that uses data mining and probability to forecast outcomes
- Part of Data Science or Statistical Learning
- High importance in recent past because data availability rising exponentially
- Examples
  - Preventive Maintenance: Automotive manufacturer want to predict the occurrence of fault or failure (or classify the condition of vehicles) through the sensor data captured
  - Insurance company want to classify drivers as very risky, risky, safe, very safe etc. on the basis of captured driving habits so that insurance premium can be intelligently fixed

PREDICTIVE MODELING PROCESS



SUPERVISED LEARNING

- Understanding the behavior of a target variable(y) as a set of input variable (x1, x2....) change
- Develop a function or model to estimate the target
- Machine learning task of inferring a function from labeled training data
- Algorithm analyzes the training data and produces an inferred function, which can be used for mapping new examples
- Examples
  - Predict the price of stock 6months or 1year from now on the basis of company performance and economic data
  - Identify the impact of price, relative brand position, economic condition, competition level on the demand of a particular product during a given period

UNSUPERVISED LEARNING

- Machine learning technique for finding hidden patterns or intrinsic structures in data
- Type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses
- Most common unsupervised learning method is clustering
- Examples
  - Identification of typical profile of employees who quit quickly
  - Identification of products sold together
  - Grouping of cities with respect to their characteristics
  - Develop a scale to measure brand position

PREDICTIVE MODELING TASKS

✓ Hypothesis Testing

✓ Classification and Class Probability Estimation

✓ Value Estimation

HYPOTHESIS TESTING

- Hypothesis are statements about a given problem
- Hypothesis testing consists of determining the plausibility of the statements on the basis of data
- Examples
  1. Increasing number of years of education increases earning potential
  2. Design A produces a lower defect rate compared to design B
  3. A particular design of a web page leads to more conversion compared to another

CLASSIFICATION AND CLASS PROBABILITY ESTIMATION

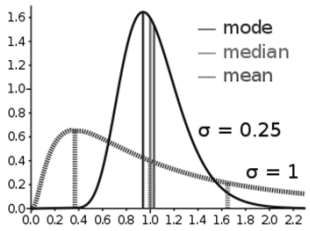
- Used in situation where the target is to be classified
- The problem is to allocate the target variable to one of the classes based on the value of some explanatory variables
- Allocation to a particular class is made on the basis of the estimated probabilities
- Examples
  1. Classification of credit card transaction as fraudulent or not
  2. Prediction of whether a customer will renew contract or not
  3. Whether a sales bid will be won, lost or abandoned by the customer
  4. Classification of a loan application as low, high or medium risk

VALUE ESTIMATION

- Used to estimate or predict the value of a target variable rather than classifying the same
- Value estimated based on explanatory variables
- Examples
  1. Finding the lifetime value of a customer
  2. Estimating the effort required to complete a Software Project
  3. Finding the total number of cheques that may arrive for processing

DESCRIPTIVE STATISTICS

MEAN, MEDIAN, MODE

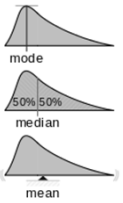


- Mean - average of all values and is sometimes called the arithmetic mean
- Median - statistical median is the 'middle' number in a sequence of numbers
- Mode - the number that occurs most often within a set of numbers
- Range - the difference between the highest and lowest values within a set of numbers

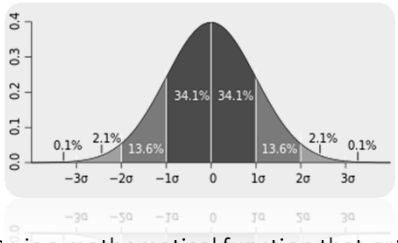
$$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

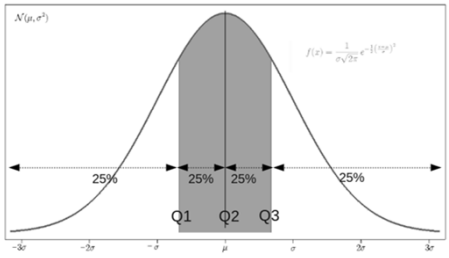
1, 3, 3, **6**, 7, 8, 9
Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9
Median = (4 + 5) ÷ 2
= 4.5



QUANTILE



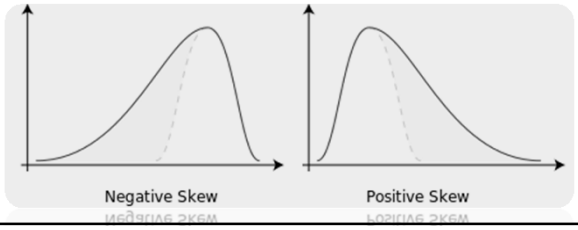


- Probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment
- A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
- For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.
- Quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities.
- The area below the red curve is the same in the intervals (-∞,*Q1*), (*Q1*,*Q2*), (*Q2*,*Q3*), and (*Q3*,+∞)

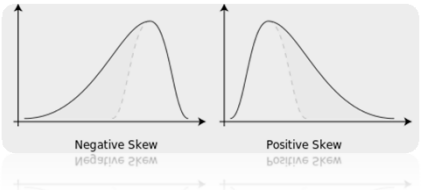
SKEWNESS, VARIANCE, SD

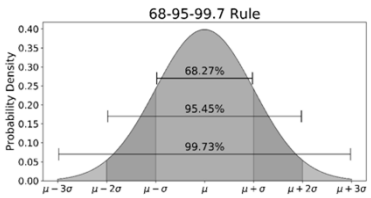
$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- Skewness - measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined
- skewness differentiates extreme values in one versus the other tail
- Kurtosis measures extreme values in either tail. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution
- Variance (σ^2) - the expectation of the squared deviation of a random variable from its mean
- Standard Deviation (σ)- measure that is used to quantify the amount of variation or dispersion of a set of data values



DISTRIBUTION





- The standard normal distribution has two parameters: the mean and the standard deviation - 68% of the observations are within +/- one SD of the mean, 95% are within +/- two SD, and 99.7% are within +- three SD
- The skewness and kurtosis coefficients measure how different a given distribution is from a normal distribution.
- Skewness of normal distribution is zero. If the distribution has a negative skewness, then the left tail of the distribution is longer than the right tail. Positive skewness implies that the right tail of the distribution is longer than the left.
- For investors, high kurtosis of the return distribution implies that the investor will experience occasional extreme returns (either positive or negative), more extreme than the usual + or - three standard deviations from the mean that is predicted by the normal distribution of returns. This phenomenon is known as *kurtosis risk*.

EXERCISE 1:

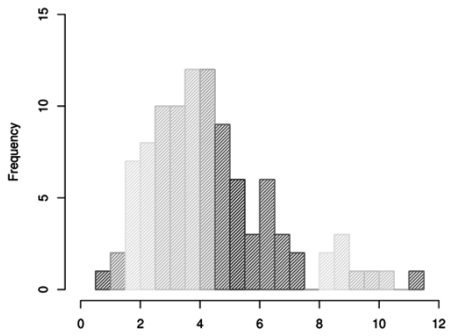
The monthly credit card expense of an individual in 1000 rupees is given in the file 'Credit_Card_Expenses.csv'

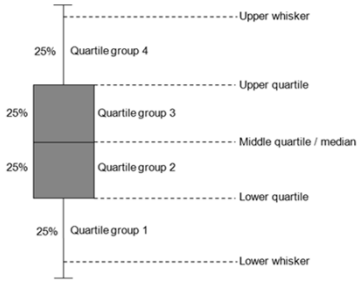
1. Read the dataset
2. Compute mean, median, minimum, maximum, range, variance, standard deviation, skewness and quantiles of CC Expenses
3. Draw a Histogram and Boxplot of CC Expenses

SOLUTION

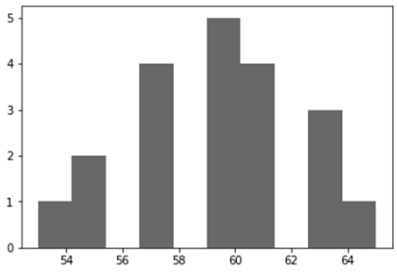
```
import pandas as mypd
myData=mypd.read_csv(".\datasets\Credit_Card_Expenses.csv")
myData
cc= myData.CC_Expenses
Cc
cc.mean()
cc.median()
cc.mode()
cc.std()
cc.var()
cc.min()
cc.max()
cc.quantile(0.9)
cc.skew()
cc.describe()
```

HISTOGRAM AND BOX PLOT

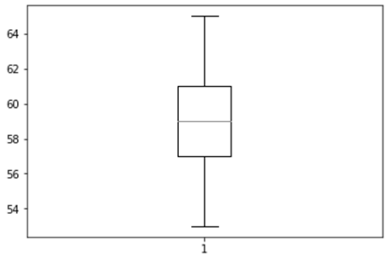





```
import matplotlib.pyplot as myplot
myplot.hist(cc)
myplot.show()
```




```
#Boxplot
myplot.boxplot(cc)
myplot.show()
```



EXERCISE 2:

Data of 30 customers on credit card usage in INR 1000, sex (1-male, 2-female) and whether they have done shopping or banking (1:yes, 2:no) with credit card are given in file 'Credit_Card_Exercise.csv'

1. Read the dataset
2. Compute mean, median, minimum, maximum, range, variance, standard deviation, skewness and quantiles of CC Expenses
3. Check whether average usage varies with sex?
4. Check whether average usage varies with those who do shopping with credit card and those who don't do shopping?
5. Check whether average usage varies with those who do banking with credit card and those who don't do banking?
6. Compute the aggregate average of usage with sex and shopping?
7. Compute the aggregate average of usage with all three factors
8. Draw a Histogram of CC Expenses

SOLUTION

```
import pandas as mypd
myData=mypd.read_csv(".\datasets\CC_Expenses_Exercise.csv")
myData
cc=myData.Credit_Card_usage
cc.describe()
gender=myData.Sex
cc.groupby(gender).mean()
shopping =myData.Shopping
banking=myData.Banking
cc.groupby(shopping).mean()
cc.groupby(banking).mean()
```

SOLUTION

```
import matplotlib.pyplot as myplot
myplot.hist(cc)
myplot.show()
myData.boxplot(column='Credit_Card_usage',by='Sex')
myplot.show()
cc.groupby([gender, banking]).mean()
cc.groupby([gender, shopping, banking]).mean()
cc.groupby([gender,shopping,banking]).describe()
```

THANK YOU