UNIT - II Basic Data Analytic methods

2.1 Statistical methods for Evaluation.

- 2.1.1 Hypothesis Testing
- 2.1.2 Difference of means
- 2.1.3 Wilcoxon rank-sum test
- 2.1.4 type 1 type 2 errors
- 2.1.5 Power & sample size
- 2.1.6 ANNOVA

2.2 Advanced Analytical Theory & methods.

- 2.2.1 clustering - overview
- 2.2.2 K - means

Use cases
overview of methods
Determine no. of clusters
Diagnostics
Reasons to choose & Cautions.

2.1 ~~stastis~~
2.1 Statistical Methods for Evaluation -

- visualization is useful for data exploration & presentation, but statistics is crucial because it may exist throughout the entire Data Analytics Life cycle.
- statistical techniques are used during the initial data exploration & data preparation, model building, evaluation of final models & assessment of how the new models improve the situation when deployed in the field.
- There are some useful statistical tools that are ~~useful~~ listed below;

2.1.1 Hypothesis Testing -

- When comparing populations, such as testing or evaluating the difference of the means from two samples of data, a common technique to assess the difference or the significance of the difference is hypothesis Testing.

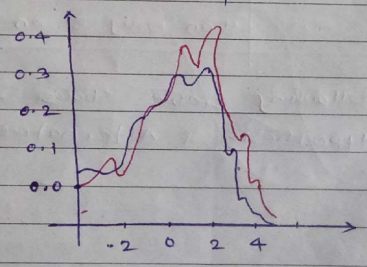


fig. Distribution of TWO samples of Data.

- The basic concept of hypothesis testing is to form an assertion & test it with data. When performing hypothesis tests, the common assumption is that there is no difference betn two samples. This assumption is used as the default position for building the test or conducting a scientific experiment, statisticians refer to this as the null hypothesis (H_0).
- The alternative hypothesis (H_A) is that there is a difference betn two samples. E.g. If the task is to identify the effect of drug A compared to drug B on patients, the null hypothesis & alternative hypothesis would be this.

H_0:	Drug A & Drug B have the same effect on patient
H_A:	Drug A has a greater effect than Drug B on patients.

- following table shows example Null Hypothesis & Alternative Hypotheses.

Application	Null Hypothesis	Alternative Hypothesis
i) Accuracy forecast	Model x does not predict better than the existing model.	Model x Predicts better than the existing model.
ii) Recommendation Engine	Algorithm Y does not Produce better recommendations than the current algo. being used.	Algorithm Y Produces better recommendations than the current algo. being used.
iii) Regression modeling	this variable does not affect the outcome because its coefficient is zero.	this variable affects outcome because it's coefficient is not zero.

2.1.2 Difference of Means -

- Hypothesis testing is a common approach to draw interferences on whether or not the two populations, denoted Pop1 & Pop2, are different from each other. This provides two hypothesis tests to compare the means of the respective Populations based on samples randomly drawn from each population.
- Specifically, the two hypothesis tests in this

point, consider the following null & alternative hypotheses.

- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$

The ~~mean~~ μ_1 & μ_2 denote the population means of POP1 & POP2 respectively.

- The basic testing approach is to compare the observed sample means, $\bar{x}_1$ & $\bar{x}_2$, corresponding to each population. If the values of $\bar{x}_1$ & $\bar{x}_2$ are approximately equal to each other, the distributions of $\bar{x}_1$ & $\bar{x}_2$ overlap substantially as shown in below fig. & the null hypothesis is supported. A large observed difference between the sample means indicates that the null hypothesis should be rejected.

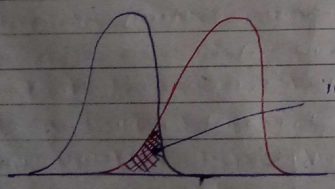


Fig. Overlap of the two distributions is large if $\bar{x}_1 \approx \bar{x}_2$

- formally, the difference in means can be tested using i) student's t-test and or ii) Welch's test.

i) student's t-test :

- Student's t-test assumes that distributions of the two populations have equal but unknown variances. suppose n_1 & n_2 samples are randomly & independently selected from two populations, POP1 & POP2, resp. If each population is normally distributed with the same mean ($\mu_1 = \mu_2$) & with the same variance, then T (t-static), given in equation ①, follows a t-distribution with $n_1 + n_2 - 2$ degree of freedom (df).

$$T = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

ii) Welch's t-test -

- When the equal population variance assumption is not justified in performing student's t-test for the difference of means, welch's t-test can be used based on T expressed in foll. eqn

$$T_{welch} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- where $\bar{x}_i$, s_i^2 & n_i corresponds to the i^{th} sample mean, sample variance & sample size.
- Welch's t-test uses the sample variance (s_i^2) for each population instead of the pooled sample variance.

2.1.3 Wilcoxon Rank-sum Test -

- A t-test represents a parametric test in that it makes assumptions about the population distributions from which the samples are drawn. If the populations can't be assumed or transformed to follow a normal distribution, a nonparametric test can be applied or used.
- The Wilcoxon rank-sum test is nonparametric hypothesis test that checks whether two populations are identically distributed. Assuming the two populations are identically distributed, one would expect that the ordering of any sampled observations would be evenly intermixed among themselves.
- Ex:
  - Let the two populations again be pop 1 & pop 2, with independently random samples of size n_1 & n_2 respectively. The total no. of observations is then $N = n_1 + n_2$.
- The first step to this test is to rank
- the set of observations from the two groups as if they came from one large group. The smallest observation receives a rank of 1, the second smallest observation receives a rank of 2 & so on with the largest observation being assigned the rank of N. Ties among the observations receive a rank equal to the average of the ranks they span.
- After ranking all the observations, the assigned ranks are summed for least one population's sample. If the distribution of pop 1 is shifted to the right of the other distribution, the rank-sum corresponding to the pop 1's sample should be larger than

the rank-sum of pop2. The

The Wilcoxon rank-sum test determines the significance of the observed rank-sums

Because the Wilcoxon test does not assume anything about the population distribution, It is generally considered more robust than the t-test.

- In other words, there are fewer assumptions to violate. However, when it is reasonable to assume that the data is normally distributed, student's or Welch's t-test is an appropriate hypothesis test to consider.

2.1.8 Type I & Type II Errors -

- A hypothesis test may result in two types of errors, depending on whether the test accepts or rejects the null hypothesis. These two errors are know as type I & type II errors.

- A type I error -
  - It is the rejection of the null hypothesis when the null hypothesis is TRUE. The probability of the type I error is denoted by Greek letter α

- A type II errors :
  It is acceptance of null hypothesis when the null hypothesis is FALSE. The probability of the type II error is denoted by the Greek letter β.

Table -

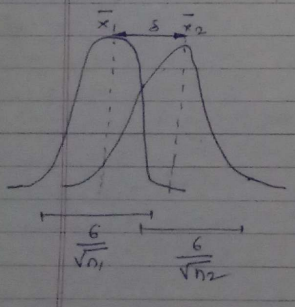
	H_0 is True	H_0 is False
H_0 is accepted	correct outcome	Type II Error
H_0 is rejected	Type I error	correct outcome

- for significance level such as $\alpha = 0.05$, If the null hypothesis ($\mu_1 = \mu_2$) is TRUE, there is a 5% chance that the observed T value based on the sample data will be large enough to reject the null hypothesis. By selecting an appropriate significance level, the probability of committing a type I error can be defined before any data is collected or analyzed.
- The probability of committing a Type II error is somewhat more difficult to determine. If two population means are truly not equal, the probability of committing a type II error will depend on how far apart the means truly are.
- To reduce the probability of a type II error to a reasonable level, it is often necessary to increase the sample size.

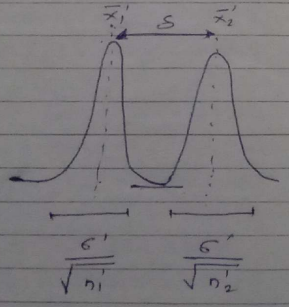
2.1.5 Power & Sample size -

- The power of a test is the probability of correctly rejecting the null hypothesis It is denoted by 1 - β, where β is the probability of a type II error. Because the power of a test improves as the sample size increases, power is used to determine the necessary sample size.
- In the difference of means, the power of a hypothesis test depends on the true difference of the population means. In other words, for a fixed significance level, a larger sample size is required to detect a smaller difference in the means. In general, the magnitude of the difference is known as the effect size. As the sample size becomes larger, it is easier to detect a given effect size δ.

Moderate sample size



Larger sample size



Sample

Fig. A Larger size better identifies a fixed effect size.

- With a larger enough sample size, almost any effect size can appear statistically significant. However, a very small effect size may be useless in a practical sense.

2.1.6 ANOVA -

- The hypothesis tests presented in the previous sections are good for analyzing means betn two populations, but what if there are more than two populations & consider an example of testing the impact of nutrition & exercise on 60 candidates betn age 18 & 50. The candidates are randomly split into six groups, each assigned with a different weight loss strategy & the goal is to determine which strategy is the most effective.
  - Group 1 : only eats junk food.
  - Group 2 : only eats healthy food.
  - Group 3 : eats junk food & does cardio exercise every other day.
  - Group 4 : eats healthy food & does cardio exercise every other day
  - Group 5 : eats junk food & does both cardio & strength training every other day.

- Group 6 eats healthy food & does both cardio & strength training every other day.

- multiple t-tests could be applied to each pair of weight loss strategies. In this example, the weight loss of group 1 is compared with the weight loss of group 2, 3, 4, 5 or 6. Similarly, the weight loss of group 2 is compared with that of the next 4 groups. Therefore 15 t tests would be performed.
- However, multiple t-tests may not perform well on several populations for two reasons. first because, the no. of t-tests increases as the no. of groups increases, analysis using the multiple t-tests becomes computationally more difficult. second, by doing a greater no. of analyses, the probability of committing at least one type I error somewhere in the analysis greatly increases.
- Analysis of variance (ANOVA) is designed to address these issues. ANOVA is generalization of the hypothesis testing of the difference of two population means, ANOVA tests if any of the population means differ from the other population means. The null hypothesis of ANOVA is that all the population means are equal. The alternative hypothesis is that at least one pair of the population means is not equal. In other words,
  - $H_0 : \mu_1 = \mu_2 = \cdots \mu_n$
  - $H_A : \mu_i \neq \mu_j$ for atleast one pair of i, j.

2.2 Advanced Analytical Theory & Methods.

2.2.1 Overview of clustering -

- In general, clustering is the use of unsupervised techniques for grouping similar objects. In machine learning, unsupervised refers to the problem of finding hidden structure within unlabeled data. clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.
- The structure of the data describes the objects of interest & determines how best to group the objects.
- Ex. based on customer's personal income, it is straightforward to divide the customers into three groups depending on arbitrarily selected values. The customers could be ~~selected~~ divided into three groups as follows:
  - Earn less than $10,000
  - Earn bet^n $10,000 & $99,000
  - Earn $100,001 or more

- In this case, the income levels were chosen somewhat subjectively based on easy-to-communicate points of delineation. However, such groupings do not indicate a natural affinity of the customers within each group.
- clustering is a method often used for exploratory analysis of the data. In clustering, there are no predictions made. Rather, clustering methods find the similarities betn objects according to the object attributes & groups the similar objects into clusters.

2.2.2 K-means -

- Given a collection of objects each with n measurable attributes, K-means is an analytical technique that, for a chosen value of K, identifies K clusters of objects based on the objects proximity to the center of the K groups.
- The center is determined as the arithmetic average (mean) of each cluster's n-dimensional vector of attributes.
- foll. fig. shows three clusters of objects with two attributes. Each object in the dataset is represented by a small dot color-coded to the closest large dot, the mean of the cluster.

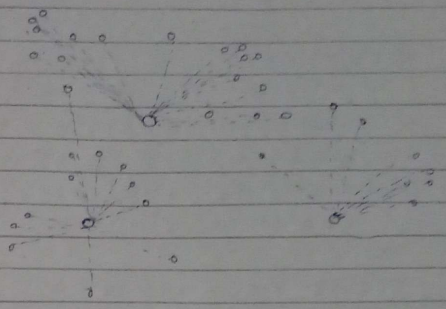


fig. Possible K-means clusters for K = 3.

2.2.2.1 Use Cases -

- clustering is often used as a lead-in to classification. Once the clusters are identified, labels can be applied to each cluster to classify each group based on its characteristics.
- clustering is primarily an exploratory technique to discover hidden structures of the data, possibly as a prelude to more focused analysis or decision processes.
- Applications of K-means -
  - ① Image processing -
    - video is one example of the growing volumes of unstructured data being collected. within each frame of a video, K-means analysis

can be used to identify

② Medical -
- patient attributes such as age, height, weight, systolic & diastolic blood pressures, cholesterol level & other attributes can identify naturally occurring clusters.

③ Customer segmentation -
- Marketing & sales groups use K-means to better identify customers who have similar behaviors & spending patterns.
- Ex. a wireless provider may look at the following customer attributes - monthly bill, no. of text messages, data volume consumed.

2.2.2.2 Overview of the method -
- To illustrate the method to find K clusters from a collection of M objects with n attributes, the two dimensional case (n = 2) is examined. Each object in this example has two attributes, it is useful to consider each object corresponding to the point (x_i, y_i), where x & y denote the two attributes & $i = 1, 2 \cdots M$. for given cluster of m points ($m \leq M$), the point that corresponds to the cluster's mean is called a centroid

- The K-means algo. to find K clusters can be described in the following four steps.

1) Choose the value of K & the K initial guesses for the centroids.
In this example, K = 3 & the initial centroids are indicated.

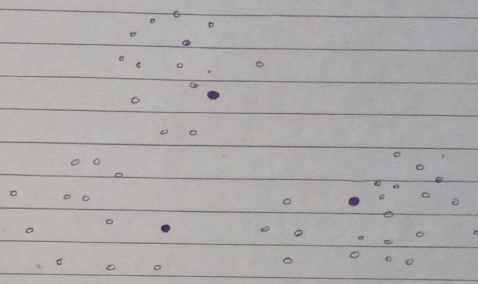


Fig. Initial starting points for the centroids.

2) Compute the distance from each data point (x_i, y_i) to each centroid. Assign each point to the closest centroid. This association defines the first K clusters.
- In two dimensions, the distance, d, betn any two points (x_1, y_1) & (x_2, y_2), in the cartesian plane is typically expressed by using the

Euclidean distance measure provided in eq. ①.

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \qquad \text{eq} - ①$$

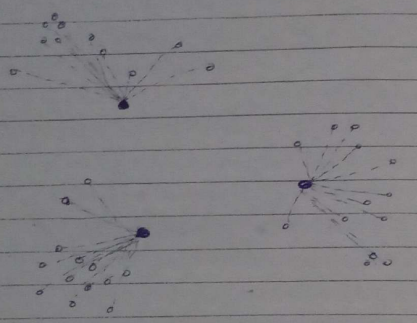


fig. points are assigned the closest centroid.

3) compute the centroid, the center of mass, of each newly defined cluster from step2.

- In foll. fig., the computed centroids in step 3 are lightly shaded points of the corresponding color. In two dimensions, the centroid (x_c, y_c) of the m points in a K-means cluster is calculated as follows eq. ②

$$(x_c, y_c) = \left[\frac{\sum_{i=1}^{m} x_i}{m}, \frac{\sum_{i=1}^{m} y_i}{m} \right]$$

eq - ②

Thus, (x_c, y_c) is the ordered pair of the arithmetic means of the coordinates of the m points in the cluster.

4) Repeat step 2 & 3 until the algorithm converges to an answer.
   a. assign each point to the closest centroid computed in step3
   b. compute the centroid of newly defined clusters.
   c. Repeat until the algorithm reaches the final answer.

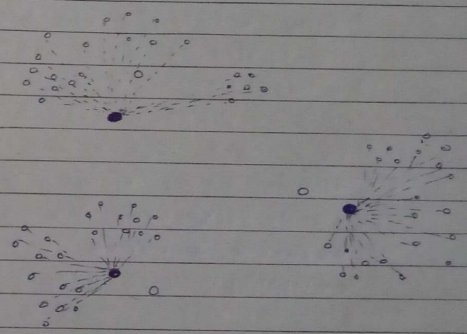


fig.. Compute the mean of each cluster

2.2.2.3 Determining the no of clusters.

- With the preceding algo, K clusters can be identified in a given dataset, but what value of K should be selected? The value of K can be chosen based on a reasonable guess or some predefined requirement.

- Using R to perform a K-means Analysis. To illustrate how to use the WSS to determine a appropriate number K of cluster the foll. example uses R to perform a K-means analysis. The task is to group 620 high school seniors based on their grades in three subject area: English, Mathematics & science. The grades are averaged over their high school career & assume values from 0 to 100. The foll. R code establishes the necessary R libraries & import the csv file containing the grades.

```
library (Plyr)
library (ggplot2)
library (cluster)
library (lattice)
library (graphics)
library (grid)
library (gridExtra)
```

WSS - Within sum of squre

```
#import the student grades
grade_input = as.data.frame (read.csv
("c:\\data\\grades-km_input.csv"))
```

- The following R code formats the grades for processing. The data file contains 4 (four) columns. The first column holds a student identification (ID) number & the other three columns are for the grades in the three subject areas. Because the student ID is not used in the clustering analysis, it is excluded from the K-means input matrix, kmdata

```
kmdata_orig = as.matrix (grade_input
[,c ("student", "English", "math", "science"
)])
kmdata ← kmdata_orig[,2:4]
kmdata [1:10,]
```

	English	math	science
[1,]	99	96	97
[2,]	99	96	97
[3,]	98	97	97
[4,]	95	100	95
[5,]	95	96	96
[6,]	96	97	96
[7,]	100	96	97
[8,]	95	98	98
[9,]	98	96	96
[10,]	99	99	95

2.2.2.4 Diagnostic -

- The heuristic using WSS can provide at least several possible K values to consider. When the no. of attributes is relatively small, a common approach to further refine the choice of K is to plot the data to determine how distinct the identified clusters are from each other.
- In general, the following questions should be considered

i) Are the clusters well separated from each other ?

ii) Do any of the clusters have only a few points ?

iii) Do any of the centroids appear to be too close to each other ?

- In first case, ideally the plot would look like the one shown in fig. 1, when n=2. The clusters are well defined, with considerable space bet? the four identified clusters. However, in other cases, such as fig. 2, the clusters may be close to each other, & the distinction may not be so obvious.

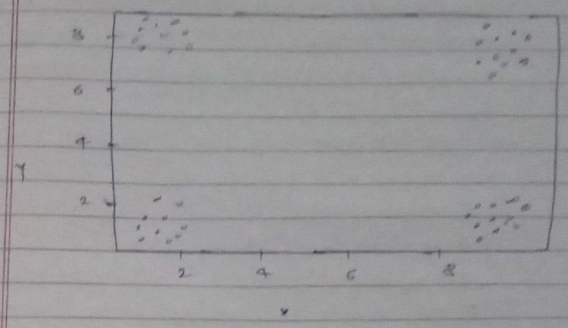


Fig. 1 Example of distinct cluster

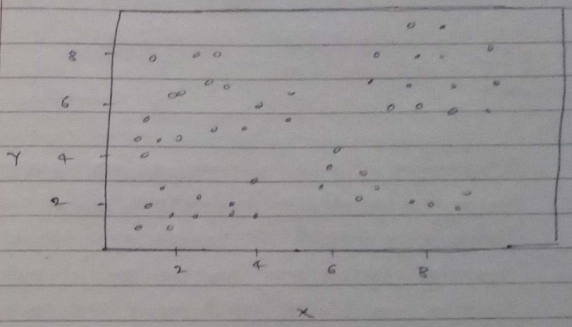


Fig. 2 Example of less obvious clusters

- In such cases, it is important to apply some judgment on whether anything different will result by using more clusters.

- Fig. 3, If using more clusters does not better distinguish the groups, it is almost certainly better to go with fewer clusters.

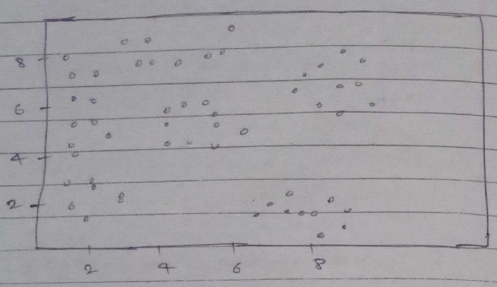


Fig.3 six clusters applied to the points

2.2.2.5 Reasons to choose & Cautions -

- K-means is a simple & straightforward method for defining clusters. Once clusters & their associated centroids are identified, it is easy to assign new objects to a cluster based on the object's distance from the closest centroid. Because the method is unsupervised.
- Although K-means is considered an unsupervised method, there are still several decisions that the practictioner must make:
  1) what object attributes should be included in the analysis?
  ii) what unit of measure should be used for each attribute?
  iii) Do the attributes need to be rescaled so that one attribute does not have disproportionate effect on the results?
  iv) what other considerations might apply?

•/object/ Attributes/-