Course Title: Advanced Statistical Methods

Course code: MAT6001

Dr. G. Hannah Grace

Vellore Institute of Technology, Chennai

Module 1: Basic Statistical tools and Analysis

.FALL 2020

Prepared by Dr. G. Hannah Grace, Assistant Professor, Division of Mathematics, School of Advanced Sciences, VIT Chennai.
OFFERED TO : AI & ML , BIG DATA (M.Tech Engineering Program)

Contents

Course Content : Basic Statistical Tools for Analysis

Summary Statistics, Correlation and Regression, Concept of R^2 and Adjusted R^2 and Partial and Multiple Correlation, Fitting of simple and Multiple Linear regression, Explanation and Assumptions of Regression Diagnostics.

Summary Statistics:

Summary statistics summarizes and provides quick a simple description of the data. It tells us about the value in the data set. It includes the measures of central tendency, measures of dispersion and Graphs.

Measures of Central Tendency:

The measures of central tendency represent a central value for all observations. It is also called averages or measures of location. The different measures of central tendency are mean (also called the arithmetic mean or averages), median (the middle value of the data set) , Geometric mean (used for interest rates and other types of growth), mode (most repeated value of data) and harmonic mean (calculate the **average** of the ratios or rates)

Mean, Median, Mode

For raw data

$$Mean = \frac{Sum\ of\ all\ values}{total\ number\ of\ values} = \frac{\Sigma x}{N}$$ ✓

$$Median = the\ middle\ value\ (\ when\ the\ data\ is\ arranged\ in\ order)$$

$$Mode = the\ most\ common\ value$$

Example:

1. Consider the raw data 9,8,5,9,8,7,12,6,5,9. Calculate its mean median and mode.

$$Mean = \frac{\Sigma x}{N} = \frac{78}{10} = 7.8$$

$$\frac{9+8+5+9+8+7+12+6+5+9}{10}$$

To calculate median arrange the data in ascending order.

5,5,6,7,8,8,9,9,9,12

Median = mean of both middle values = 8+8/2 =8

Mode= 9 (since it appears 3 times)

mode?

1,2,1,2,1,2,3,7

For discrete data

Mean:

$$Mean = \frac{\Sigma fx}{\Sigma f}$$ ✓

$$\frac{\Sigma fx}{N}$$

Σf = sum of frequen.

where f is the corresponding frequency.

Median:

To find median: Calculate the cumulative frequency by adding the frequencies successively. Then find (N/2) where N=Σ*f*=total frequencies. Then identify the cumulative frequency just greater than (N)/2. The value of x corresponding to that cumulative frequency is the median.

Mode: The value corresponding to the maximum frequency

Problem:

Calculate the mean, median and mode for the following data

x f

x	Frequency
0	1
1	5
2	10
3	6
4	3

Cumulative freq?

N/2

f	Cf
1	1
5	6
10	16
6	22
3	25

$\Sigma f = N$

Solution: ✓

x	Frequency(f)	fx	Cummulative frequency (cf)
0	1	0	1 ✓
1	5	5	6
2	10	20	**16**
3	6	18	22
4	3	12	25
Total	25	55	

Median class
Modal class

med.

mode

N

N

$$Mean = \frac{\Sigma f x}{\Sigma f} = \frac{55}{25} = 2.2$$

Median= value of (N/2)th observation

= value of (25/2)th observation

=value of 12th observation = 2 (value 12 likes just above 16)

Mode = the frequency of observation which has maximum frequency

=2

For continuous data:

$$Mean = \frac{\Sigma f x}{\Sigma f}$$

$$Median = l + \frac{\frac{N}{2} - m}{f} \; c$$

Where

l = Lower limit of the median class

N = Total Numbers of frequencies

f = Frequency of the median class

m = Cumulative frequency of the class preceding the median class

c = the class interval of the median class.

above

$\frac{N}{2} =$

Note: Median class is, that class which correspond to the cumulative frequency just greater than N/2.

$$Mode = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \; c$$

where

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

c = width of the class limits

Note: Modal class is the class which has maximum frequency.

Problem:

1. The ages of students in a small primary school were recorded in the table below

Age $-x$	5-6	6-7	7-8
Frequency f	29	40	38

Estimate the mean, median and mode

To estimate the mean we use mid point of the interval

Class interval	Mid point(x)	Frequency(f)	fx
5-6	5.5	29	159.5
6-7	6.5	40	260
7-8	7.5	38	285
	Total	107	704.5

$$Mean = \frac{\Sigma fx}{\Sigma f} = \frac{704.5}{107} = 6.584$$

To calculate median find the cumulative frequency

Class interval	Frequency(f)	cf
5-6	29 f_0	29
6-7	**40** f_1	**69**
7-8	38 f_2	107

107

$\frac{107}{2}$ = 53.5

← median class → modal class.

N/2= 107/2=53.5 lies the interval 6-7 which is the median class

$$Median = l + \frac{\frac{N}{2} - m}{f}\ c$$

Σf

Here L=6, f=40, N=107, m=29, c=1

$$Median = 6 + \frac{53.5 - 29}{40}\ (1) = 6.6125$$

To find Modal class- max freq is 40 so modal class is 6-7

L=6, f1=40,f0=29, f2=38, c=1

$$Mode = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right)\ c$$

$$Mode = 6 + \left(\frac{40 - 29}{2(40) - 29 - 38}\right)\ (1) = 6.8462$$

Range : max - min

20/ 9/21

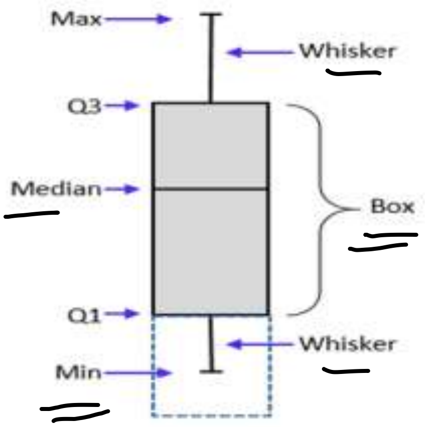
The Measure of dispersion:

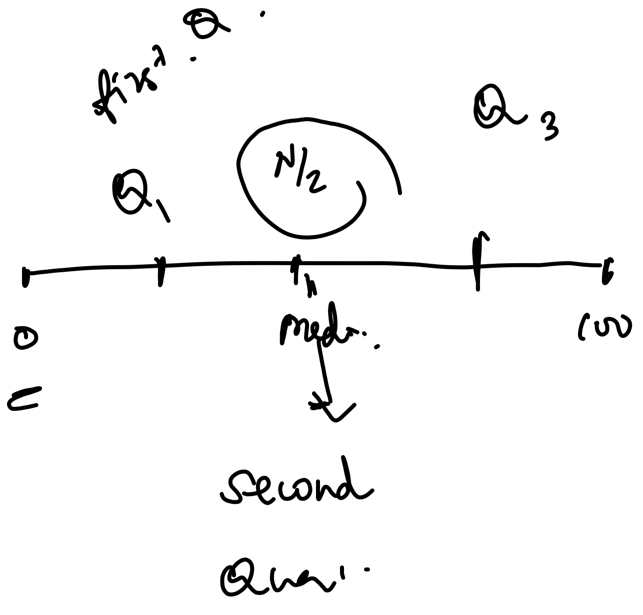
The measure of dispersion tells how spread out or varied your data set is. This can given important information. It is also called the measure of spread of data. The various measures are Range (how spread out the data is), Interquartile range (tells us where the middle fifty percent of your data lies), quartiles (boundaries for lowest, middle and upper quarter of data, skewness (does the data have low or high values) and kurtosis (a measure of how flat or peaked data appears)

Quartiles:

Quartiles divide the data into quarters(four segments) according to where the numbers fall on the number line. The four quarters that divide a data set into quartiles are:

1. First Quartile : the lowest 25% of numbers
2. Second quartile: between 26% and 50% (up to the median)
3. Third quartile : 51% to 75% (above the median)
4. Fourth Quartile: The highest 25% of numbers





$$Q_1 = l_1 + \left(\frac{\frac{N}{4} - m_1}{f_1}\right) c_1$$

$$Q_3 = l_3 + \left(\frac{3\frac{N}{4} - m_3}{f_3}\right) c_3$$

l_1 *and* l_3 are the lower limit of the first and third quartile class respetively

m_1 *and* m_3 are the cummulative frequency preceeding the first and third quartile class respectively

f_1 *and* f_3 are the frequencies of the quartile class respectively

c_1 *and* c_3 are the class intervals

N is the total of all the frequencies

Coefficient of Quartile Deviation:

$$Inter\ Quartile\ Range(IQR) = Q_3 - Q_1$$

Quartile Deviation is defined as, half of the distance between Q_1 and Q_3. It is also called as semi-inter quartile range.

$$Quartile\ Deviation = \frac{Q_3 - Q_1}{2}$$

The coefficient of Quartile deviation is the relative measure corresponding to Quartile Deviation.

$$Coefficient\ of\ Quartile\ deviation = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Standard deviation

$\boldsymbol{\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}}$ **, where d=x-A (for raw data)**

$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}$, where d=x-A (for discrete data)

Where, *f* = frequency of each class interval

N = total number of observation (or elements) in the population

c = width of class interval

d= mid-value of each class interval where *A* is an assumed A.M.

Note: The square root of the variance is known as **standard deviation**.

Problem:

Find the Quartiles Q1, Q2, Interquartile range, coefficient of quartile deviation and standard deviation for the ~~above~~ problem.

Class interval	Frequency(f)	cf
5-6	**29**	**29**
6-7	40	69
7-8	**38**	**107**

first Quartile class.

3rd Quartile class

107/4th

first Quartile class.

To find Q1 class

class with (N/4)th observation = (107/4)th observation

= (26.75)th value of the observation

This means Q1 class lies in 5-6 of the class interval

Here L=5, f=29, m=0, c=1

$$Q_1 = l_1 + \left(\frac{\frac{N}{4}-m_1}{f_1}\right) c_1 = 5 + \left(\frac{\frac{107}{4}-0}{29}\right)(1) = 5.9224$$

To find Q3 class

class with (3n/4)th observation = (3*107/4)the observation

= 80.25th observation

This means Q3 class is 7-8

Q_1=

Q_3

here L=7,m=69,f =38, c=1, N= 107

$$Q_3 = l_3 + \left(\frac{3\,\frac{N}{4}-m_3}{f_3}\right) c_3 = 7 + \left(\frac{3\,\frac{(107)}{4}-69}{38}\right)(1) = 7.2961$$

$$Inter\ Quartile\ Range = Q_3 - Q_{1=1.3737}$$

$$Quartile\ Deviation = \frac{Q_3-Q_1}{2} = 0.6869$$

$$Coefficient\ of\ Quartile\ deviation = \frac{Q_3-Q_1}{Q_3+Q_1} = 0.1039$$

To find the standard deviation:

Class interval	Mid point(d)	Frequency(f)	fd^2
5-6	5.5	29	877.25
6-7	6.5	40	1690
7-8	7.5	38	2137.5
Total		107	4704.75

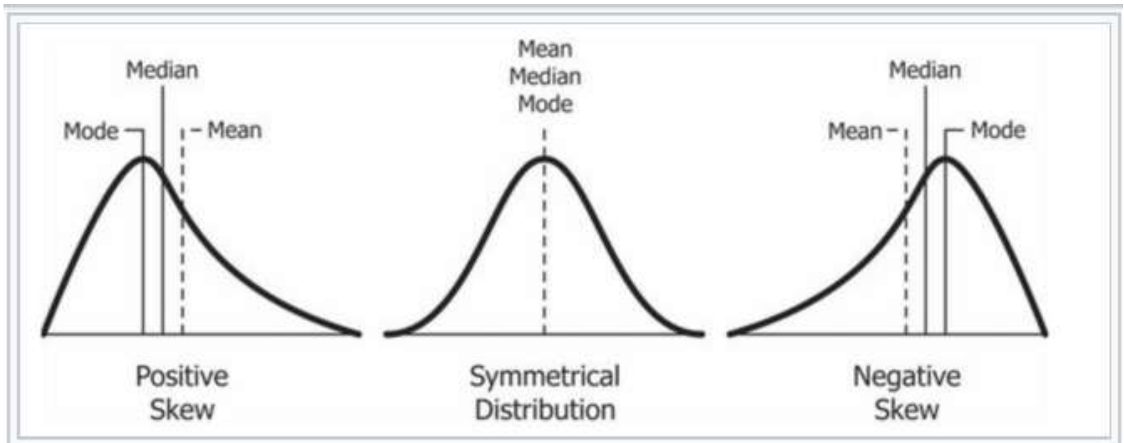
fd

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{0.6191} = 0.7869$$

Skewness:

skewness is the lack of symmetry.

$$\boldsymbol{skewness(\beta_1)} = \frac{\boldsymbol{\Sigma}(x-\bar{X})^3}{\boldsymbol{\sigma^3 N}}$$



Kurtosis:

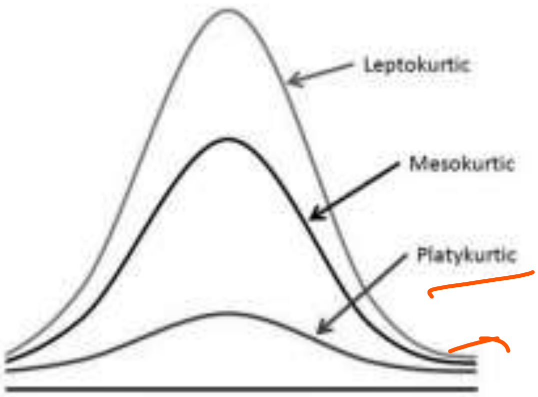
Kurtosis is the degree of peakedness of the curve. It identifies whether the tails of a given distribution contains extreme values. The kurtosis of a distribution can be calculated as:

$$\boldsymbol{kurtosis(\beta_2)} = \frac{\boldsymbol{\Sigma}(x-\bar{X})^4}{\boldsymbol{\sigma^4 N}}$$

An excess Kurtosis is a metric that compares the kurtosis of a distribution against the kurtosis of a normal distribution. The kurtosis of a normal distribution equals to 3. Hence

Excess kurtosis= kurtosis-3

A normal curve has kurtosis 3, If $\beta_2 > 3$ then it is leptokurtic and if $\beta_2 < 3$ then it is platykurtic.



flat

Problems for practice.

1) Calculate the mean, median and mode for the foll.

marks.	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90
no. of students	4	12	40	41	27	13	9	4

2) Find the missing frequency if mean is 38

marks.	10	20	30	40	50	60	70
no of students.	8	11	20	25	—	10	3

3) For the data below find the quartile deviation

x	351-500	501-650	651-800	801-950	951-1100
F	48	189	88	47	28

Also find Interquartile range and coefficient of Quartile Deviation