Chapter– 19

DEATH RATE OF INDIA USING THE DATA SCIENCE APPLICATION

Ch.Id:-ASU/GRF/EB/RAPOET/2022/Ch-19
DOI: https://doi.org/10.52458/9789391842765.2022.eb.grf.asu.ch-19

[1]VISHAL RAJPUT
[1]Apeejay Stya University, Sohna, Gurugram

ABSTRACT

In this chapter firstly, introduction of what the statistics is and how to the find the mean, median, mode and some other statistical tools are discussed. And some of the applications of the statistics are discussed and the most important application for the real life which is in data science it is discussed with the help of an example of the India's death rate In which we have the data from 1995 to 2022(till now) which have been discussed. Because the deaths are the most important concern in today's life on an individual

INTRODUCTION

Statistics is the mathematical science of collecting, organizing, and analysing data so that meaningful conclusions can be drawn. It is a form of mathematical analysis in which various quantitative models and analysis are used to obtain a set of experimental data or real-life studies. In statistics the valid conclusions are made and we make the reasonable decisions on the basis of the analysis or by calculation by different formulas.

During calculation or analysing the statistics data we can see the data in two form which are mentioned below:

1) Grouped data: grouped data in statistics means that in which the data is arranged in the form of the class intervals and are grouped. Example: 0 to10, 10 to 20.

2) Ungrouped data: it is also known as raw data. It is not sorted into categories or groups or in class intervals. Ungrouped data are the list of numbers or the data in which numbers are randomly placed and not sorted. Example: 1, 4, 5, 3, 8.

Two foremost statistical techniques are utilized in facts evaluation: descriptive statistics, which summarize facts from a pattern the use of indexes which includes the mean or standard deviation, and inferential statistical record, which draw conclusions from data which can be problem to random variation (e.g., observational errors, sampling variation). Descriptive statistical records are most usually involved with two sets of properties of a distribution (sample or population) and central tendency (or location) seeks to signify the distribution's central or usual value, whilst dispersion (or variability) characterizes the quantity in which the individuals of the distribution leave from its middle and every other. Inferences on mathematical statistical records are done under the framework of possibility theory, which offers with the evaluation of random phenomena.

A standard statistical method involved the check of the connection among two statistical statistics sets, or a statistics set taken or seen from an idealized model. A speculation in statistics called as hypothesis is proposed for the statistical relationship among the two statistics sets, and that is as compared as an another method to an idealized null hypothesis of no relationship among statistics set.

Some of the statistical measures are mentioned below the following:

1. Measures Of Central Tendency
2. Measures Of Variations
3. Skewness
4. Moments
5. Kurtosis

Measures of central tendency: Central tendency is defined as the measure of the statistical data which represents the single/average value of the entire data. It is used to provide and to solve an accurate description of the data.

The measure of central tendency of the statistical data can be found using three important measures:

1. Mean
2. Median
3. Mode

Mean: The mean is one of the most important measures of central tendency, other than the median and mode. Mean is however the common/average of the given set of values.

It denotes the same or we can say that the equal distribution of values for a given statistics data set.

For ungrouped data:

For calculating the mean for ungrouped data, we are needed the whole (add the total values of the given set) values given in a data sheet and then divide the sum by total number of the value. Simply, the Mean means the sum of total observations divided by the total number of observations.

$$\overline{x} = \frac{\sum x_i}{N}$$

Where, X_1, X2-------X_n is the sum of the given observations and n = total number of observations.

Example: In a class there are 8 students and they have secured a percentage of 88, 82, 88, 85, 84, 80, 81, 82. Calculate the mean percentage obtained by the class.

Answer:

Mean = (Sum of observations)/total number of observations

= (88 + 82 + 88 + 85 + 84 + 80 + 81 + 82)/8

=670/8=83.75

Mean percentage obtained by the class is 83.75.

For Grouped data:

For calculating the mean for grouped data, we need to multiply the frequencies with the data values and get the sum then divide the sum.

In grouped data mean we will have the data values Xi, where i = 1, 2, 3, ---n.

And also have the frequencies for every data values which is denoted by fi, where i = 1, 2, ------ n. And the formula is given below:

$$\overline{x} = \frac{\Sigma f_i x_i}{\Sigma f_i}$$

Where, fi in the numerator it is the total sum of all the frequencies given

Median: The median is the middle value of the statistical data in this the data is to be arranged in the lowest to highest order of the values. The median is less affected by skewed data.

For Ungrouped data:

If there is an odd number of total values then the median value is the middle value which have same number of upper and lower values.

Median = $\left(\frac{n+1}{2}\right)^{th}$**obs**

And if there is an even number of the total values then the median will be the sum of the middle pair and then divided by two.

Median = $\frac{\frac{n}{2}obs+(\frac{n+1}{2})obs}{2}$

Example: In a data set of {3, 13, 2, 34, 11, 26, 47}, find median.

ANSWER

Step 1: sort the elements {2, 3, 11, 13. 26, 34, 47}

Step 2: check that the total number of elements is even or odd. So, in this case the total number of elements is 7 which is an odd number so the median is the middle number which is 13

Step 3: median = 13

For Grouped data:

In grouped data the data is to be continuous and in the frequency distribution form. The formula for median for grouped data is given below:

Median = l + $[\frac{\frac{n}{2}-c}{f}]\times$**h**

Where, l = lower limit of the median class

c = cumulative frequency of the class preceding the median class

f = frequency of the median class

h = size of class

Mode: In statistics, mode in a statistical data which refers to the number that occur most frequently/more repeatedly. We can also say that the number in a data set, which has a high frequency or is also called mode or modal value. Sometimes, it is also possible that there cannot be anyS mode or no modal value.

For Ungrouped data:

In ungrouped data the number or the integer which occurs the most in the frequency is termed or assigned as mode.

Mode = maximum occurrence of the number/integer in frequency

Example: Given the marks of the science test of 10 students as 10,11,11,14,15,11,17,19,19,13. Find the mode.

ANSWER: the mode of the given data is 11 as it is occurring 3 times.

For Grouped data:

In grouped data we have continuous data in class form. So, in this which class have the higher frequency holds the value of mode. And mode for the grouped data can be find by the given formula:

Mode = l + $(\frac{f1-f0}{2f1-f0-f2})\times$h

Where l = lower limit of the modal class, h = size of class interval

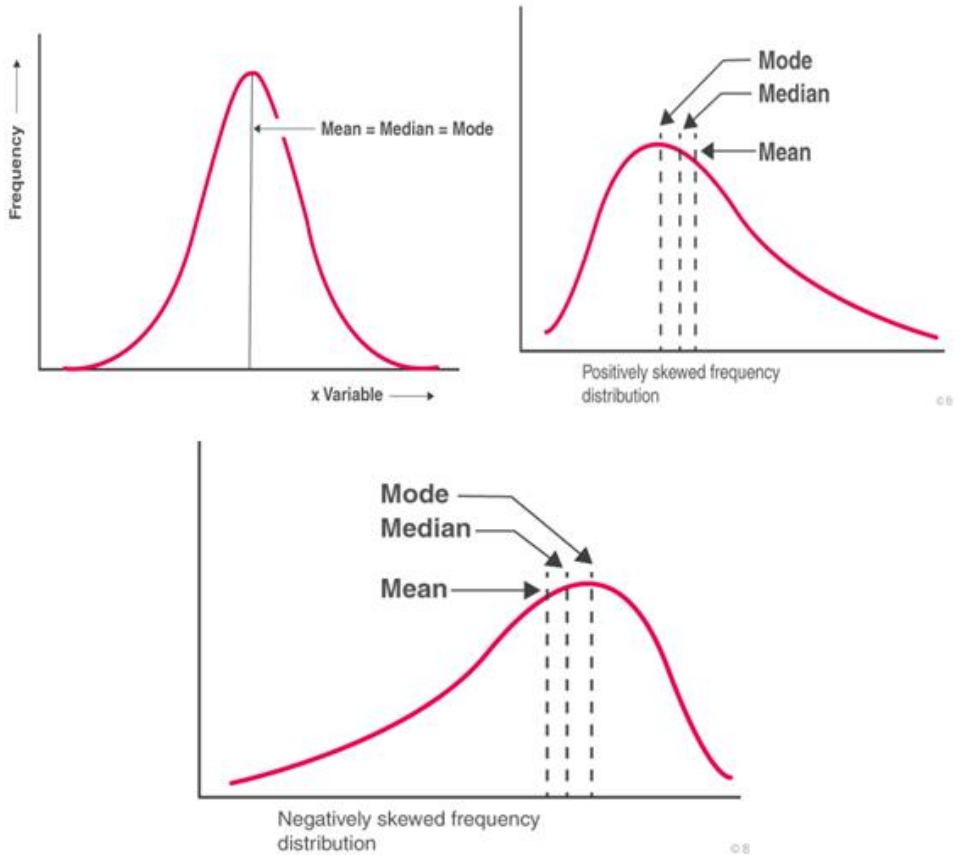
F1= frequency of the modal class, f0 = frequency of the class preceding the modal class, f2 = frequency of the class succeeding the modal class.

Relationship between Mean, Median, Mode

The relationship between mean, median, mode is known as empirical relationship. **Which is given by**

Mode = 3Median – 2Mean

Relationship between Mean, Median, Mode when doing with frequency distribution:



Measures of Variations: Measures of Variations is the extent in which we check or we got to know how our observations of our data vary from the average value. And in the measures of the central tendency we only gets one singled value that is average and which is not sufficient to represent the whole data by only one observation. So, it is required to check the or to get the variations between the observations so that anyone can easily understand.

Variations are of two types that are relative and absolute variation.

The measure of variations of the statistical data can be found using four important measures:

1. Range
2. Interquartile range and quartile deviation
3. Average distribution
4. Standard deviation

- **Range:** The range is defined as the difference between the highest observation and the lower observation. It is one of the simplest method for studying the variations because in this we only have to subtract the lowest observation value from the highest observation value.

  Range = L-S

  Where, L = Largest observation value

  S = Smallest observation value

  The relative measure of range is also known as coefficient of range. To get the coefficient of range we have the following formula:

  Coefficient of range = (L-S)/(L+S)

  Example: Shubham drove through 8 states on his summer vacation.

  Fuel prices varied from state to state he travelled. Calculate the range of fuel prices given? Rs. 4.79, Rs. 0.61, Rs. 2.96, Rs. 3.09, Rs. 2.64, Rs. 2.25, Rs. 3.73, Rs. 1.97.

ANSWER

Range = L-S

L = 4.79, S = 0.61

Range = 4.79 – 0.61 = 4.18

- **Interquartile range or Quartile deviation:** Interquartile range is defined as the difference between the third quartile and the first quartile. Interquartile range tells

us about the middle 50% value of the data when arranged from lower to highest. It is given by:

Interquartile range = Q3-Q1

Where Q3 = Third quartile and formula for Q3 is

$$\mathbf{Q_3 = L + \frac{\frac{3N}{4} - p.c.f}{f} \times i}$$

And Q1 = First quartile and formula for Q1 is

$$\mathbf{Q_1 = L + \frac{\frac{N}{4} - p.c.f}{f} \times i}$$

Quartile deviation is defined as the average of the interquartile range. It is also known as semi-interquartile range and it is given by:

$$\mathbf{Q.D = \frac{Q3 - Q1}{2}}$$

And the Quartile deviation is an absolute measure of variation then the relative measure is known as coefficient of Quartile deviation and denoted by:

$$\mathbf{Coeff\ of\ Q.D = \frac{Q3 - Q1}{Q3 + Q1}}$$

- **Average deviation:** Average deviation is calculated because to calculate the scatterness around the average. Average deviation can be calculated using mean and median. For calculating the average deviation first we have to find mean and median from which we have to calculate then averaging these deviations by their arithmetic mean. Formula for calculating the Average deviation is given below:

  When average deviation is taken from median

  $$\mathbf{A.D_{(Med)} = \frac{\Sigma |X - Med|}{N}}$$

  When Average deviation is taken from mean

  $$\mathbf{A.D_{(\bar{x})} = \frac{\Sigma |X - \bar{x}|}{N}}$$

  And the relative measure of average deviation is known as the coefficient of average deviation. And it is given by:

  $$\mathbf{Coeff\ of\ A.D_{(med)} = \frac{A.D}{Median}}$$

- **Standard deviation:** Standard deviation is denoted by σ (this symbol called as Greek letter sigma) and also as S.D. It gives us the information about how much the value is deviated. If the value of SD is deviated the less then it is said it is close to the mean and if the value of SD is deviated the more then it is said to be far from the mean. In simple language it talks about the deviation about the mean. And square root of SD is known as variance.

For Ungrouped data:

For the ungrouped data the SD is given by:

$$\boldsymbol{\sigma = \sqrt{\frac{\Sigma(X - \bar{x})^2}{N}}}$$

Where $\bar{x}$ = mean of the data observations

N = total number of observations

Example:

Some of the data set observations are given as 8, 1, 5, 6. Find the SD.

ANSWER:

Step 1) calculate the mean = (8+1+5+6)/4 = 20/4 = 5

Step 2) sum the square the difference of the mean from the observation = $(8-5)^2+(1-5)^2+(5-5)^2+(6-5)^2 = 9 + 16 + 1 = 26$

Step 3) calculate the variance = (sum of the square difference)/N

= 26/4 = 6.5

Step 4) find out the square root of the variance which is known as SD

= $\sqrt{6.5}$ = 2.54

For Grouped data:

- For the grouped data when SD is taken from the actual mean then it is given by:

$$\boldsymbol{\sigma = \sqrt{\frac{\Sigma(X - \bar{x})^2}{N}}}$$

Where, $\bar{x}$ = mean of the data

N = sum of all frequencies given of the data (Σfi),

i = 1, 2, 3, ---------, n.

- For the grouped data when SD is taken from the assumed mean then the SD is given by:

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

Where d = X-A, A= assumed mean

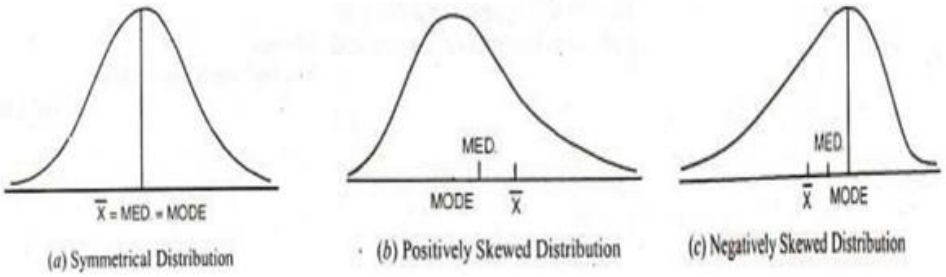
i = class interval difference

N = sum of all frequencies given of the data (Σfi),

i = 1, 2, 3, ---------, n.

Skewness: Measure of central tendencies and variations are not sufficient to describe the nature of the observations. So, for that we have Skewness and skewness means lack of symmetry.

The following graphs will give you the proper understanding of the skewness.



(*a*) Symmetrical Distribution (*b*) Positively Skewed Distribution (*c*) Negatively Skewed Distribution

In graph (a) it is called as symmetrical distribution because the value of mean = median = mode.

In graph (b) it is called as positively skewed distribution because the value of mean is greater than the mode and the median lies between the mean and mode.

In graph (c) it is called as negatively skewed distribution because the value of mean is less than the mode and the median is in between mean and mode.

Following three are the methods of finding and measuring the skewness:

1. Karl Pearson's coefficient of skewness: It is denoted by SK_p. it is used most for calculating the skewness. The value of Karl Pearson's coefficient varies between ±1. Its formula is:

$$\mathbf{SK_p} = \frac{Mean - Mode}{\sigma}$$

2. Bowley's coefficient of skewness: It is denoted by SK_B. Bowley's coefficient is bases on quartiles. The value of Bowley's coefficient also varies between ±1. Its formula is:

$$\mathbf{SK_B} = \frac{Q3 + Q1 - 2Med}{Q3 - Q1}$$

3. Kelly's coefficient of skewness: It is denoted by SK_k. It is based on both percentiles and deciles. Its formula is:

   Kelly's coefficient of skewness when based on percentiles:

$$\mathbf{SK_K} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

   Kelly's coefficient of skewness when based on deciles:

$$\mathbf{SK_K} = \frac{D_9 - 2D_5 + D_1}{D_9 - D_1}$$

Moments: It is denoted by μ(a Greek letter read as mu). It helps in describing the characteristics of the observations.

For ungrouped data:

The moment about arithmetic mean of the given variables is given by:

$$\boldsymbol{\mu_r = \frac{1}{N}\ \Sigma(X - \bar{x})^r}$$

The moment about any arbitrary point of the given variables is given by:

$$\boldsymbol{\mu_r = \frac{1}{N}\ \Sigma(X - A)^r}$$

For grouped data:

The moment about the arithmetic mean of the given variables is given by:

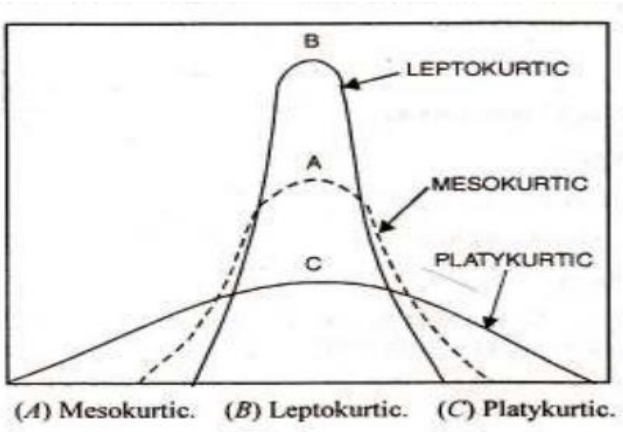
$$\boldsymbol{\mu_r = \frac{1}{N}\ \Sigma f(X - \bar{x})^r}$$

The moment about any arbitrary point of the given variables is given by:

$$\boldsymbol{\mu_r = \frac{1}{N}\ \Sigma f(X - A)^r}$$

Kurtosis: A measure of the peakness or convexity of a curve is known as Kurtosis. It also tells us about the degree of flatness in the region about the mode of the frequency.

Following curve will tells us about the three different scopes of the curve:



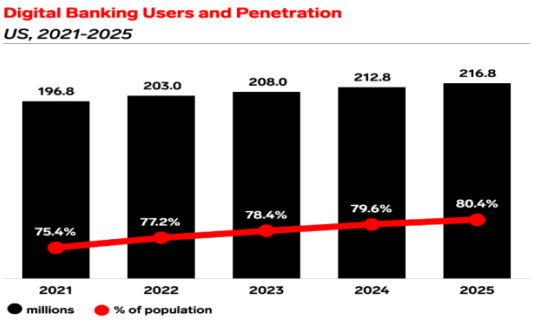
(*A*) Mesokurtic. (*B*) Leptokurtic. (*C*) Platykurtic.

1. Curve (a) which is Mesokurtic tells us that it is a normal curve.
2. Curve (b) which is Leptokurtic tells us about that the curve is more peaked than the Mesokurtic.
3. Curve (c) which is Platykurtic tells us about that the curve is more flattened than the Mesokurtic.

Kurtosis is measured by β_2 and its derivate by γ_2. Where $\beta_2 = \mu_4/\mu_2^2$ and $\gamma_2 = \beta_2 - 3$.

- **Applications of statistics:**

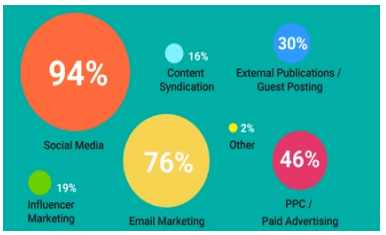
1) **Banking sector:** Statistics plays a important function withinside the financial industry, specially in banking and investment. Banks use information to lessen threat of their loans, examine the economic marketplace and are expecting the effect of the monetary crisis. The statistical equipment are utilized in banking region for creating a decision. How do a financial institution make investments their cash in capital marketplace or for lending to their customer, they ought to use statistics. It is likewise use in forcusting withinside the banking region.



2) **Data science:** A data scientist makes use of Statistics that is one of the beneficial measures for data scientists to get the applicable outputs of the pattern area and to gather, review, or will do the analyze, and then draw the conclusions from records, in addition to observe quantified mathematical models to suitable variables.

3) **Economics:** There are such a lot of principles of economics which can be absolutely depending on statistics. All the information accrued to discover the countrywide income, employment, and the concept of call for and supply, the connection among exports and imports etc., are interpreted thru it.

4) **Marketing:** Another large application of statistics is in advertising especially in social media and marketing. In social media, they analyse facts to growth quantity of fans and in advertising, they analyse facts to optimize campaign's performance.



Topic: India-Historical death rate

I have chosen topic of death rate of India in time period of (1995 to 2022). I will begin with introduction to my topic by providing a gist of how the death are occurring. What are the causes of the deaths which are increasing year by year. Then I will present my data and using statistical measures like mean, median, mode, standard deviation, skewness and kurtosis I will analyze the data and conclude.

❖ **Introduction:**

Mortality records through motive of demise for the whole cross-phase of the populace withinside the country is critical for knowledgeable choice making withinside the Health sector. The data on reasons of demise to be had from the Medical Certification of Causes of Death (MCCD) below the Civil Registration System (CRS) be afflicted by the trouble of each the insurance and quality.

The pinnacle reasons of demise stratified through one-of-a-kind variables except analysing the mortality from unique situations which includes Tuberculosis, Malaria, Maternal Conditions and HIV/AIDS.

Overall non-communicable illnesses are the main reasons of demise in the country, constituting 42% of all deaths. Communicable, maternal, perinatal and dietary situations represent every other 38% of deaths. Injuries and ill-described reasons represent 10% of deaths each. However, majority of ill-described reasons are at older ages (70 or better years) and maximum of ill-described deaths are in all likelihood to be from non-communicable illnesses.

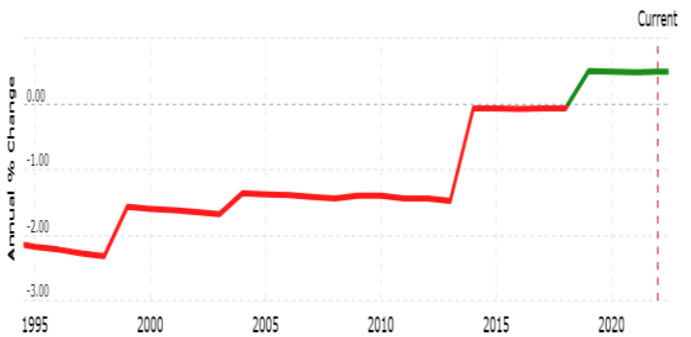
- **Sample data:** This is the data of number of deaths happened in India from the year 1995 to 2022.

Year	Death Rate
2022	7.380
2021	7.344
2020	7.309
2019	7.273
2018	7.237
2017	7.242
2016	7.247
2015	7.253
2014	7.258
2013	7.263
2012	7.372

Year	Death Rate
2011	7.480
2010	7.589
2009	7.697
2008	7.806
2007	7.920
2006	8.034
2005	8.147
2004	8.261
2003	8.375
2002	8.518
2001	8.661

Year	Death Rate
2000	8.804
1999	8.947
1998	9.090
1997	9.306
1996	9.523
1995	9.739

Graph study of the yearly deaths:



This graph is from the year 1995 to 2022(current), it is showing that how the increasing decreasing number of deaths had occurred in India. Now, I will analyse the

data using statistical tools like mean, median, range, standard deviation, skewness and kurtosis and give some conclusions based on it.

Calculation of data

1) **Mean:**

$\bar{x} = \frac{\sum x_i}{N}$

$\sum x_i$ = 9.73+ 9.52+ 9.30 + 9.09 + 8.94 + 8.80 + 8.66 + 8.51 + 8.37 + 8.26 + 8.14 +8.03 + 7.92 + 7.80 + 7.69 + 7.58 + 7.48 + 7.37 + 7.26 + 7.25 + 7.25 + 7.24 + 7.24 + 7.23 + 7.27 + 7.30 + 7.34 + 7.38 = 223.9

$N = 28$

$\bar{x} = \frac{223.9}{28}$ **= 7.99**

2) **Median:**

For median we have to first arrange the data in ascending order. And the arranged data is given below:

S no	Value	S no.	Value	S no.	Value	S no.	Value	S no.	Value
1	7.23	7	7.27	13	7.58	19	8.26	25	9.09
2	7.24	8	7.30	14	7.69	20	8.37	26	9.30
3	7.24	9	7.34	15	7.80	21	8.51	27	9.52
4	7.25	10	7.37	16	7.92	22	8.66	28	9.73
5	7.25	11	7.38	17	8.03	23	8.80		
6	7.26	12	7.48	18	8.14	24	8.94		

Median = $\frac{n}{2}$ ***obs***

Since, the number of observations is even so the formula for median is n/2 and (n/2)+1 = 28/2 = 14obs, 15obs

And in the table 14obs is 7.69 and 15obs is 7.80. so, the median is 7.69 and 7.80

3) **Mode:**

Mode = 3Median – 2Mean

=23.07 – 15.98

=7.09

4) **Range:**

Range = L – S

= 9.73 – 7.23

= 2.5

5) **Coefficient of range:**

(L-S)/(L+S)

= 2.5/16.96

= 0.14

6) **Interquartile range:**

Q3 – Q1

Q3 = 3N/4 = 21obs

= 8.51

Q1 = N/4 = 7obs

= 7.27

Interquartile range = Q3 – Q1 = 8.51 – 7.27 = 1.24

7) **Quartile deviation:**

(Q3 – Q1)/2

= 1.24/2

= 0.62

8) **Coefficient of QD:**

(Q3-Q1)/(Q3+Q1)

= 1.24/9.75

= 0.12

9) **Standard deviation:**

$$\boldsymbol{\sigma} = \sqrt{\frac{\boldsymbol{\Sigma(X - \bar{x})^2}}{\boldsymbol{N}}}$$

Since, the data is so large so I am putting the direct values.

$=\sqrt{16.76/28}$

= 0.76

10) **Average deviation:**

A.D $(\bar{x})$ **=** $\frac{\Sigma|X-\bar{x}|}{N}$

In this also the data is too large so I am putting the direct values.

= 18.71/28

=0.66

11) Skewness:

$\mathbf{SK_p = \frac{Mean-Mode}{\sigma}}$

= (7.99-7.09)/0.76

=1.1 (It is positively skewed)

12) Kurtosis:

$\beta_2 = \mu_4/\mu_2^2$

$= 0.8/0.59^2$

= 2.35

STATISTICAL ANALYSIS OF DATA

After the calculation of data using the statistical measures I have analysed:

- Firstly, with the help of sample data provided I can tell that the number of deaths is always varying. First the deaths are more then it starts decreasing then after some years is slowly starts increasing.
- Mean is the average or the most common value in a collection of numbers. Here, mean of the data of the deaths from 1995 to 2022(till now) is 7.99 which means that every year at least 7.99 of people have died. The number of electors can be more or less than 7.99 also.
- Here, from the value of median we get to see that the deaths below the median value is decreasing and above that it is increasing.
- Range tells us that the minimum death rate between highest and lowest ranges between 2.5.
- Interquartile range is telling about middle 50% range of the death occurred.
- Value of SD is less than the mean so every deviation in deaths is very close to the mean deaths.
- The distribution is positively skewed distribution as the value of the mean is greater than the mode.

All the measures of statistics of the given data is given below:

Mean	**7.99**
Median	**7.69, 7.80**
Mode	**7.09**
Range	**2.5**
Coeff. of range	**0.14**

Interquartile range	**1.24**
Quartile deviation	**0.62**
Coeff. Of QD	**0.12**
Standard deviation	**0.76**
Average deviation	**0.66**
Skewness	**1.1**
Kurtosis	**2.35**

CONCLUSION

This statistical analysis is done to realize how the deaths are increasing and decreasing. If the deaths are increasing al should upon that what are the factors and if decreasing then what are the factors that are helping all for reducing the death rate in India. Like given is the graph of deaths occurring per 1000 people:

As we can see from this graph and after analysing the data which I had choose that in the year 1995 death rate is more and after that at some year it is seen that the death rate is minimum from which we can predict that the deaths are happening more in 19's year after that it is decreasing and that is a great thing that peoples are getting medical helps and different solutions as these solutions comes when some peoples starts analysing that what are the causes and how we can prevent them. Like in this graph which is showing deaths per 1000 people that firstly the deaths are more in number later after the precautions the death rate starts falling. After that it starts increasing but not that much like it is shown in early years. So, this is how statistics helps anyone for analysing the record or the data for studying and to get the correct and effective methods for the better result in the positive way in any situation like here the death rate in India.

BIBLIOGRAPHY

- *https://www.macrotrends.net/countries/IND/india/death-rate*
- *https://www.cghr.org/wordpress/wp-content/uploads/Causes_of_death_2001-03.pdf*
- *https://researchleap.com/applied-statistics-basic-principles-and-application/#:~:text=Applied%20statistics%20is%20the%20root,%2C%20accounting%2C%20business%2C%20etc.*
- *https://allthingsstatistics.com/descriptive-statistics/properties-uses-range-statistics/*
- *http://www.ascdegreecollege.ac.in/wp-content/uploads/2020/12/Business-Statistics-by-Gupta.pdf*
- *https://en.wikipedia.org/wiki/Mean*
- *https://byjus.com/maths/central-tendency/*