MEASURES OF CENTRAL TENDENCY

Odisha Social Science
Research & Consultancy

CONTENTS

- Descriptive Measures
- Measure of Central Tendency (CT)
  - Concept and Definition
  - Mean
  - Median
  - Mode
- Uses of Different Measures of CT
- Advantages and Disadvantages of Different Measures of CT

Enabling Objectives

- To equip the trainee with skills to manipulate the data in the form of numbers that they encounter as a health sciences professional.
- The better able they are to manipulate such data, better understanding they will have of the environment and forces that generate these data.

Descriptive Statistics

- Classification and Tabulation of Data
- Presentation of Data
- Measure of Central Tendency (CT)
- Measure of Shape
- Measure of Dispersion

Measures of Central Tendency

Concept

- Data, in nature, has a tendency to cluster around a central Value.
- That central value condenses the large mass of data into a single representative figure.
- The Central Value can be obtained from sample values (Called Statistics) and Population observations (Called Parameters)

Measures of Central Tendency

Definition

A measure of central tendency is a typical value around which other figures congregate.

Simpson and Kofka

(Statistician)

Average is an attempt to find an single figure to describe a group of figures.

Clark and Schakade

(Statistician)

Different Measure of Central Tendency (MCT)

1. **Mathematical Average**
   - Arithmetic Mean Simply Mean
   - Geometric Mean
   - Harmonic Mean
2. **Positional Average**
   - Median
   - Mode
3. **Mean, Median and Mode are the Most Commonly used MCT in Health Science**

Characteristics of an Ideal MCT

1. It should be rigidly defined so that different persons may not interpret it differently.
2. It should be easy to understand and easy to calculate.
3. It should be based on all the observations of the data.
4. It should be easily subjected to further mathematical treatment.
5. It should be least affected by the sampling fluctuation .
6. It should not be unduly affected by the extreme values.

Characteristics of Mean

- Most representative figure for the entire mass of data
- Tells the point about which items have a tendency to cluster
- Unduly affected by extreme items (very sensitive to extreme values)
- The positive deviations must balance the negative deviations (The sum of the deviations of individual values of observation from the mean will always add up to zero)
- The sum of squares of the deviations about the mean is minimum

Arithmetic Mean (AM) or Mean

AM is obtained by summing up all the observations and dividing the total by the number of observation.

Calculation of Mean

The mean is calculated by different methods in three types of series.

I. Individual Data Series
II. Discrete Data Series
III. Grouped Data

Individual data: $X_1, X_2, X_3, X_4, \ldots\ldots\ldots X_n$
(n = Total No. of observation)

$$\text{Mean} = \frac{X_1 + X_2 + X_3 + \ldots..}{n}$$

Example:

Find the mean size of Tuberculin test of 10 boys measured in millimeters.

3,5,7,7,8,8,9,10,11,12.

$$\text{Mean} = \frac{3 + 5 + 7 + 7 + 8 + 8 + 9 + 10 + 11 + 12}{10}$$

$= 80 / 10 = 8$ mm

Discrete frequency distribution

Let $x_1, x_2, x_3 \ldots\ldots x_n$ be the variate values and $f_1, f_2, f_3 \ldots f_n$ be their corresponding frequencies, then their mean is given by

$$\text{Mean} = \frac{x_1.f_1 + x_2.f_2 + x_3.f_3 + \ldots\ldots + x_n.f_n}{f_1 + f_2 + f_3 + \ldots + f_n}$$

Example:

Find mean days of Confinement after delivery in the following series.

Days of Confinement (X)	No. of patients (f)	Total days of confinement of each group. Xf
6	5	30
7	4	28
8	4	32
9	3	27
10	2	20
Total	**18**	**137**

Ans. Mean days of confinement = 137 / 18 = 7.61

Grouped frequency distribution

Let $m_1, m_2, \ldots\ldots m_n$ be the mid-point of the class and $f_1, f_2, \ldots\ldots\ldots f_n$ be the corresponding frequency

Direct method:

$$\text{Mean: } \bar{X} = \frac{\Sigma f_i m_i}{\Sigma f_i}$$

, Where m_i refers to the mid point of the i^{th} class f_i frequency of the i^{th} class and $\Sigma f_i = N$

Shortcut method:

$$\text{Mean} = A + \frac{\Sigma f_i d_i}{N}$$

Where, A is the assumed mean and $d_i = m_i - A$.

Example:

Calculate overall fatality rate in smallpox from the age wise fatality rate given below.

Age Gr in yr	0-1	2-4	5-9	Above 9
No. of Cases	150	304	421	170
Fatality rate	35.33	21.38	16.86	14.17

Solution :

$$\text{Mean fatality rate} = \frac{\sum fx}{\sum f} = \frac{21305.98}{1045} = 20.39$$

Age Gr. in year	No. of Cases (f)	Fatality Rate (x)	fx
0-1	150	35.33	5299.5
2-4	304	21.38	6499.52
5-9	421	16.86	7098.06
Above 9	170	14.17	2408.9
Total	1045	87.74	21305.98

Problem:

Find the mean weight of 470 infants born in a hospital in one year from following table.

Weight in Kg	2.0-2.4	2.5-2.9	3.0-3.4	3.5-3.9	4.0-4.4	4.5+
No. of infant	17	97	187	135	28	6

Solution :

Weight in Kg. X	Continuous CI in kg.	Mid Value	$d_i = m_i - A$	f	fd	Sum of fd
2.0-2.4	1.95-2.45	2.2	-1.0	17	-17	**-65.5**
2.5-2.9	2.45-2.95	2.7	-0.5	97	-48.5	
3.0-3.4	2.95-3.45	3.2	0	187	0	
3.5-3.9	3.45-3.85	3.7	0.5	135	67.5	**+104.5**
4.0-4.4	3.95-4.45	4.2	1	28	28	
4.5+	4.45+	4.7	1.5	6	9	
Total				**470**		**+39**

Mean Weight = $w + (\Sigma fx / \Sigma f)$

Mean Weight = $3.2 + (39 / 470) = 3.28$ Kg

Step deviation method

Ex:- Calculation of the mean from a frequency distribution of weights of 265 male students at the university of Washington

CI-Wt)	Cont. CI	f	d	fd
90 - 99	89.5 – 99.5	1	-5	-5
100 – 109	99.5 – 109.5	1	-4	-4
110 – 119	109.5 – 119.5	9	-3	-27
120 – 129	119.5 – 129.5	30	-2	-60
130 – 139	129.5 – 139.5	42	-1	-42
140 – 149	*139.5 – 149.5*	*66*	*0*	*0*
150 – 159	149.5 – 159.5	47	1	47
160 – 169	159.5 – 169.5	39	2	78
170 – 179	169.5 – 179.5	15	3	45
180 – 189	179.5 – 189.5	11	4	44
190 -199	189.5 -199.5	1	5	5
200 – 209	199.5 – 209.5	3	6	18
Total		265		99

N=265

$$d = \frac{X-A}{h}$$

$$\bar{X} = A + \frac{\Sigma fd}{N} \times h$$

$$\bar{X} = 145 + \frac{99}{265} \times 10$$

$$= 145 + (0.3736)(10)$$

$$= 145 + 0 \quad 3.74$$

$$= 148.74$$

Merits, Demerits and Uses of Mean

Merits of Mean:

1. It can be easily calculated.
2. Its calculation is based on all the observations.
3. It is easy to understand.
4. It is rigidly defined by the mathematical formula.
5. It is least affected by sampling fluctuations.
6. It is the best measure to compare two or more series of data.
7. It does not depend upon any position.

Demerits of Mean :

1. It may not be represented in actual data so it is theoretical.
2. It is affected by extreme values.
3. It can not be calculated if all the observations are not known.
4. It can not be used for qualitative data i.e. love, beauty , honesty, etc.
5. It may lead to fallacious conditions in the absence of original observations.

Uses of Mean :

1. It is extremely used in medical statistics.
2. Estimates are always obtained by mean.

Median

Definition:

Median is defined as the middle most or the central value of the variable in a set of observations, when the observations are arranged either in ascending or in descending order of their magnitudes.

It divides the series into two equal parts. It is a positional average, whereas the mean is the calculated average.

Characteristic of Median

- A positional value of the variable which divides the distribution into two equal parts, i.e., the median is a value that divides the set of observations into two halves so that one half of observations are less than or equal to it and the other half are greater than or equal to it.
- Extreme items do not affect median *and is specially useful in open ended frequencies.*
- For discrete data, mean and median do not change if all the measurements are multiplied by the same positive number and the result divided later by the same

Computation of Median

For Individual data series:

- Arrange the values of X in ascending or descending order.
- When the number of observations N, is odd, the middle most value –i.e. the $(N+1)/2^{th}$ value in the arrangement will be the median.
- When N is even, the A.M of $N/2^{th}$ and $(N+1)/2^{th}$ values of the variable will give the median.

Example:

ESRs of 7 subjects are 3,4,5,6,4,7,5. Find the median.

Ans: Let us arrange the values in ascending order.

3,4,4,5,5,6,7. The 4^{th} observation i.e. 5 is the median in this series.

Example:

ESRs of 8 subjects are 3,4,5,6,4,7,6,7. Find the median.

Ans: Let us arrange the values in ascending order.

3,4,4,5,6,6,7,7. In this series the median is the a.m of 4^{th} and 5^{th} Observations

Discrete Data Series

- Arrange the value of the variable in ascending or descending order of magnitude. Find out the cumulative frequencies (c.f.). Since median is the size of $(N+1)/2^{th}$. Item, look at the cumulative frequency column find that c.f. which is either equal to $(N+1)/2$ or next higher to that and value of the variable corresponding to it is the median.

Grouped frequency distribution

- N/2 is used as the rank of the median instead of (N+1)/2. The formula for calculating median is given as:
- $$\text{Median} = L + [(N/2 - c.f)/f] \times I$$

Where L = Lower limit of the median class i.e. the class in which the middle item of the distribution lies.

c.f = Cumulative frequency of the class preceding the media n class

f = Frequency of the median class. I= class interval of the median class.

Median for grouped or interval data

Example : Calculation of the median (x). data represent weights of 265 male students at the university of Washington

Class – interval (Weight)	f	Cumulative frequency f "less than"
90 - 99	1	1
100 – 109	1	2
110 – 119	9	11
120 – 129	30	41
130 – 139	42	83
140 – 149	*66*	*149*
150 – 159	47	196
160 – 169	39	235
170 – 179	15	250
180 – 189	11	261
190 -199	1	262
200 - 209	3	265
N = 265 N/2=132.5		

$$X = L + \frac{N/2 - Cf}{f} \times I$$

$$X = 140 + \frac{132.5 - 83}{66} \times 10$$

$$= 140 + \frac{49.5}{66} \times 10$$

$$= 140 + (0.750)(10)$$

$$= 140 + 7.50$$

$$= 147.5$$

Mode

Definition:

Mode is defined as the most frequently occurring measurement in a set of observations, or a measurement of relatively great concentration, for some frequency distributions may have more than one such point of concentration, even though these concentration might not contain precisely the same frequencies.

Characteristics of Mode

- Mode of a categorical or a discrete numerical variable is that value of the variable which occurs maximum number of times and for a continuous variable it is the value around which the series has maximum concentration.
- The mode does not necessarily describe the 'most'(for example, more than 50 %) of the cases
- Like median, mode is also a positional average. Hence mode is useful in eliminating the effect of extreme variations and to study popular (highest occurring) case (used in qualitative data)
- The mode is usually a good indicator of the centre of the data only if there is one dominating frequency.
- Mode not amenable for algebraic treatment (like median or mean)
- Median lies between mean & mode.
- For normal distribution, mean, median and mode are equal (one and the same)

Discrete Data Series

- In case of discrete series, quite often the mode can be determined by closely looking at the data.

Example. Find the mode -

Quantity of glucose (mg%) in blood of 25 students				
70	88	95	101	106
79	93	96	101	107
83	93	97	103	108
86	95	97	103	112
87	95	98	106	115

First we arrange this data set in the ascending order and find the frequency.

Quantity of glucose (mg%) in blood	Frequency	Quantity of glucose (mg%) in blood	Frequency
70	1	97	2
79	1	98	1
83	1	101	2
86	1	103	2
87	1	106	2
88	1	107	1
93	**3**	108	1
95	2	112	1
96	1	115	1

This data set contains 25 observations. We see that, the value of 93 is repeated most often. Therefore, the mode of the data set is 93.

Grouped frequency distribution

Calculation of Mode:

$$\text{Mode} = L + \frac{\{(f_1 - f_0)}{[(f_1 - f_0) + (f_1 - f_2)]} \times I$$

L = lower limit of the modal class i.e.- the class containing mode

f_1 = Frequency of modal class

f_0 = Frequency of the pre-modal class

f_2 =Frequency of the post modal class

I = Class length of the modal class

Example: Find the value of the mode from the data given below

Weight in Kg	Exclusive CI	No of students	Weight in Kg	No of students
93-97	92.5-97.5	2	113-117	14
98-102	97.5-102.5	5	118-122	6
103-107	102.5-107.5	12	123-127	3
108-112	107.5-112.5	17	128-131	1

Solution: By inspection mode lies in the class 108-112. But the real limit of this class is 107.5-112.5

$$\text{Mode} = L + \frac{\{(f_1 - f_0)}{[(f_1 - f_0) + (f_1 - f_x)]} \times I = 110.63$$

Where, $L = 107.5, f_1 = 17, f_0 = 12, f_2 = 14, I = 5$

Merits, demerits and use of mode

Merits:

- Mode is readily comprehensible and easy to calculate.
- Mode is not at all affected by extreme values,
- Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude,

Demerits:

- Mode is ill-defined. Not always possible to find a clearly defined mode.
- It is not based upon all the observations,
- It is not amenable to further mathematical treatment
- As compared with mean, mode is affected to a great extent by fluctuations of sampling,
- When data sets contain two, three, or many modes, they are difficult to interpret and compare.

Uses

Mode is useful for qualitative data.

Comparison of mean, median and mode

- The mean is rigidly defined. So is the median, and so is the mode except when there are more than one value with the highest frequency density.
- The computation of the three measures of central tendency involves equal amount of labour. But in case of continuous distribution, the calculation of exact mode is impossible when few observations are given. Even though a large number of observations grouped into a frequency distribution are available, the mode is difficult to determine. The present formula discussed here for computation of mode suffers from the drawback that it assumes uniform distribution of frequency in the modal class. As regards median, when the number of observations are even it is determined approximately as the midpoint of two middle items. The value of the mode as well as the median can be determined graphically where as the A.M cannot be determined graphically.
- The A.M is based upon all the observations. But the median and, so also, mode are not based on all the observations

- When the A.Ms for two or more series of the variables and the number of observations in them are available, the A.M of combined series can be computed directly from the A.M of the individual series. But the median or mode of the combined series cannot be computed from the median or mode of the individual series.
- The A.M median and mode are easily comprehensible.
- When the data contains few extreme values i.e. very high or small values, the A.M is not representative of the series. In such case the median is more appropriate.
- Of the three measures, the mean is generally the one that is least affected by sampling fluctuations, although in some particular situations the median or the mode may be superior in this respect.
- In case of open end distribution there is no difficulty in calculating the median or mode. But, mean cannot be computed in this case.
- Thus it is evident from the above comparison, by and large the A.M is the best measure of central tendency.

REFERENCE

- Applied Biostatistical Analysis, W.W. Daniel
- Biostatistical Analysis, J.S. Zar
- Mathematical Statistics and data analysis, John. A. Rice
- Fundamentals of Applied Statistics, S.C Gupta and V. K Kapoor.
- Fundamental Mathematical Statistics, S.C. Gupta, V.K. Kapoor

THANK YOU

Outlier

- An observation (or measurement) that is unusually large or small relative to the other values in a data set is called an **outlier**. Outliers typically are attributable to one of the following causes:
- The measurement is observed, recorded, or entered into the computer incorrectly.
- The measurements come from a different population.
- The measurement is correct, but represents a rare event.