PROBABILIY AND STATISTICS I

LECTURE FIVE

Measures of central tendency (1)

Lecturer: Dr. Emily Roche

INTRODUCTION

This lecture will focus on mean and median as measures of central tendency.

Intended learning outcomes

At the end of this lecture, you will be able to describe reasons for measuring central value and compute mean and median for both simple and grouped data.

References

These lecture notes should be supplemented with relevant topics from the book listed in the Bibliography at the end of the lecture.

MEASURES OFCENTRAL TENDENCY

A measure of central tendency is a number that locates the approximate center of a distribution of data. The purpose of a measure of central tendency is to locate the "average" or "typical" case in a distribution of cases. The term 'average' in statistics is defined as that value of a distribution which is considered as the most representative or typical value for a group. The most commonly used measures of central tendency or averages are the mean, mode and median. Other types of averages include the weighted arithmetic mean, trimmed mean, geometric mean and harmonic mean.

There are two main objectives of averaging

1. To get a single value that describes the characteristics of the entire group

2. To facilitate comparison between groups.

Properties of a Good (Average) Measure of Central tendency

1. It should be easy to understand. It should be readily understood otherwise its use will be limited.

2. It should be simple to compute. It is important to note though that ease of computation should not be at the expense of other advantages.

3. It computation should be based on all items. It should depend on each and every item of the data set so that if any item is dropped its value is altered.

4. It should not be unduly affected by extreme items of data. If one or two very small or very large items unduly affect the average then it cannot be typical of the entire data set. Extremes may distort its value and reduce its usefulness

5. It should be rigidly defined. An average should be properly defined preferably by an algebraic formula so that it has only one interpretation. Different people computing it from the same figures should get the same answer.

6. It should be capable of further algebraic treatment. It can be used for further statistical computations to enhance its usefulness.

7. It should have sampling stability. If different samples are picked from a population and the average computed for each of them, we should expect to get approximately the same value.

ARITHMETIC MEAN

This is computed by summing up all the data items in a distribution and then dividing by the total number of the data items. It is suitable for ordinal or nominal data.

$\bar{x} = \frac{\sum x}{n}$ where $\bar{x}$ is the sample mean, n is the total number of data items in the sample.

$\mu = \frac{\sum x}{N}$ where μ is the population mean, N is the total number of data items in the population.

When some items of data recur, then:

$$\bar{x} = \frac{\sum fx}{n} \qquad \text{or} \qquad \mu = \frac{\sum fx}{N}$$

Example

1. Find the mean for the following raw data; x; 3, 5, 6, 10, 23, 12, 30

Solution

$$\bar{x} = \frac{\sum fx}{n} = \frac{3+5+6+10+23+12+30}{7} = \frac{89}{7} = 12.71$$

2. Calculate the arithmetic mean for the following data: 10,9,9,11,10, 10, 12, 11, 10, 11, 11, 11,

Solution

The data can be represented in a frequency distribution table as follows:

$\boldsymbol{X}$	$\boldsymbol{f}$	$\boldsymbol{Fx}$
9	2	18
10	4	40
11	5	55
12	1	12
	$\sum f = \mathbf{11}$	$\sum fx = \mathbf{125}$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{125}{11} = 11.364$$

For a grouped distribution, the midpoint of each interval is used to represent all the scores within that interval. It is assumed that the scores in the interval are evenly distributed.

Example

The data in the table is for the number of kilometers run during one week for a sample of 20 runners.

No of km	No of runners(f)	Midpoint (x)	f. x
$5.5 - 10.5$	1	8	8
$10.5 - 15.5$	2	13	26
$15.5 - 20.5$	3	18	54
$20.5 - 25.5$	5	23	115
$25.5 - 30.5$	4	28	112
$30.5 - 35.5$	3	33	99
$35.5 - 40.5$	2	38	76
	$\sum f = 20$		$\sum fx = 490$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{490}{20} = 24.5$$

<u>Properties of the mean</u>

1. It is the point in a distribution of scores about which the sum of the deviations is equal to zero.

For example, the mean of $2, 3, 5, 7$ and 8 is 5.

$$\sum (x - \bar{x}) = (2 - 5) + (3 - 5) + (5 - 5) + (7 - 5) + (8 - 5) = 0$$

The mean is therefore the value that balances all scores above and below it, hence, a point of balance.

2. The sum of the square of the deviations of the data items from the arithmetic mean is minimum, that is, it is less than the sum of squared deviations of the mean from any other items of data.

For instance, taking the sum of the squared deviations from the mean for the data set 2, 3, 5, 7 and 8 for each of the data items as below:

x	$(x-2)^2$	$(x-3)^2$	$(x-\bar{x})^2$ $\bar{x}=5$	$(x-7)^2$	$(x-8)^2$
2	0	1	9	25	36
3	1	0	4	16	25
5	9	4	0	4	9
7	25	16	4	0	1
8	36	25	9	1	0
	$\sum(x-5)^2=71$	$\sum(x-3)^2=46$	$\sum(x-\bar{x})^2=26$	$\sum(x-7)^2=46$	$\sum(x-8)^2=71$

It is clearly evident that the sum of squared deviations from the mean is the least.

3. If each item of data in a series is replaced by the mean, then the sum of these substitutions will be equal to the sum of the individual items of data. For example, consider the data set 2, 3, 5, 7 and 8 whose mean is 5, then:

   $5+5+5+5+5=2+3+5+7+8=25$ Therefore, $n\bar{x}=\sum x$.

4. The combined arithmetic mean for the two or more data sets can be computed as:

$$\bar{x}_c=\frac{n_1\bar{x}_1+n_2\bar{x}_2+n_3\bar{x}_3+\cdots+n_n\bar{x}_n}{n_1+n_2+n_3+\cdots+n_n}$$

Where

$\bar{x}_c$ is the combined mean for the groups

$\bar{x}_i$ $(i = 1, 2, 3, \ldots, n)$ is the mean for the individual groups

n_i $(i = 1, 2, 3, \ldots, n)$ is the number of items in the individual groups

Merits of the Arithmetic mean

1. It is the simplest average to understand and easiest to compute.
2. Its computation is based on all items of data.
3. It is rigidly defined. Every one computing the arithmetic mean for the same data set will get the same answer.
4. It lends itself to further algebraic treatment.
5. It has sampling stability since it does not vary too much when repeated samples are taken from the same population.

Limitations of the Arithmetic mean

1. The mean is very sensitive to extreme items of data.
2. The mean cannot be computed in a distribution with open ended classes without making assumptions regarding the size of the class interval of the open ended classes which may lead to substantial errors.

THE MEDIAN

Median is the middlemost item of data when data is in order of magnitude. The median is the half way point in a data set. The median is a positional average and unlike the mean its computation is not based on all items.

For example, the following were marks obtained by 11 students;

$$140, 135,137,150,154,139,160,157,156,151,165.$$

Determine the median mark.

Solution

Arranging in ascending order we get:

$$135,137,139,140,150, \mathbf{151}, 154,156,157,160,165.$$

151 is the median since it's the middlemost item

Remark: when the items are even, the average of the two middlemost items is taken to be the median.

Example

The heights of six boys was recorded as follows; $115, 114, 116, 114,117, 118$. Determine the median height.

Solution

Arranging the data in ascending order to get: $114,114, \mathbf{115}, \mathbf{116}, 117,118$

$$\text{Median} = \frac{115 + 116}{2} = 115.5$$

Remark:

a. Median for an odd data set will be the item at the $\left(\frac{N+1}{2}\right)^{th}$ position..

b. The median for an even data set will be the average between items at $\left(\frac{N}{2}\right)^{th}$ and $\left(\frac{N}{2} + 1\right)^{th}$ positions.

Median for grouped data:

The calculation of an estimate for the median for grouped data involves interpolation. It is based on an assumption that the scores in each category are evenly distributed throughout the interval in question. For an even data set, $\frac{N}{2}$ is used to locate the position of the median, because in a grouped data all the frequencies lose their individuality. For an odd data set we try to increase the accuracy level by taking $\left(\frac{N+1}{2}\right)^{th}$ position since it is a single estimate.

For example, consider the following number of absences per year in different branches of a company

No of absences	$0-4$	$5-9$	$10-14$	$15-19$	$20-24$	$25-29$	$30-34$
No of branches	16	21	12	11	10	8	2
Cumulative frequency	16	37	49	60	70	78	80

The median absence is the $\frac{80}{2} = 40^{th}$ absence. This is in the class $10-14$. An assumption is that the number of absences are evenly distributed in this class. This then gives the difference between absence of one branch and the next as the class interval divided by the number of branches in the class, thus $\frac{5}{12}$.

The median is therefore estimated as:

Median $= 9.5 + \frac{5}{12} \times 3 = 10.75$

The general formula is then given as:

$$M_d = L_d + \left(\frac{\frac{N}{2} - C_a}{f_d}\right) i_d$$

Where

$M_d -$ is the estimated median

$L_d -$ is the lower boundary of the median class

$N -$ is the total frequency

$C_a -$ is the cumulative frequency of the class preceding the median class

$f_d -$ is the frequency of the median class

$i_d -$ is the class interval of the median class

Properties of the median

It is insensitive to extreme scores

Merits

1. Can be estimated for open ended classes.
2. Not influenced by extreme items of data.
3. It is most appropriate when dealing with qualitative data.
4. The median can be estimated graphically.

Limitations

1. Its computation is not based on all items of data.
2. It is not capable of further algebraic treatment.
3. It does not have sampling stability.

Bibliography

Gupta, SP (Dr.), (2014). *Statistical methods* (43rd Ed.). Sultan Chand & Sons.

S. C. Gupta and V. K. Kapoor, (2020). *Fundamentals of mathematical Statistics* (12th Ed). Sultan Chand & Sons.