PROBABILIY AND STATISTICS I

LECTURE SEVEN

Measures of spread

Lecturer: Dr. Emily Roche

INTRODUCTION

This lecture will focus on the various absolute measures of dispersion together with their corresponding relative measures.

Intended learning outcomes

At the end of this lecture, you will be able to evaluate measures of dispersion and apply results to interpret statistical data.

References

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

MEASURES OF DISPERSION OR SCATTEREDNESS

Measures of central tendency alone cannot adequately describe a set of observations unless all the observations are exactly the same. Two or more sets of observations may have the same measure of central value but be very different in terms of the disparities in individual observations. It is therefore necessary to describe the variability or dispersion of the observations.

- A measure of variability is a number that describes the dispersion or variation in a set of observations.
- A measure of dispersion may be either absolute or relative.

- Absolute measures of dispersion are expressed in the same statistical unit as the original data.
- A measure of relative dispersion is the ratio of a measure of absolute dispersion to an appropriate average. It is also known as the coefficient of dispersion, because it is independent of the unit of measurement.
- Each absolute measure of dispersion can be converted into its relative measure.

Absolute measures of dispersion include:

- The range
- The interquartile range and quartile deviation
- The mean deviation or average deviation
- The standard deviation

The corresponding relative measures are:

- Co-efficient of Range
- Co-efficient of Quartile Deviation
- Co-efficient of mean Deviation
- Co-efficient of Variation.

Measures of variation are basically needed for the following four purposes

1. To determine the reliability of an average
2. To serve as a basis for control of variability
3. To compare two or more sets of observations with regard to their variability
4. To facilitate the use of other statistical measures

The properties of a good measure of variation are:

1. Be simple to understand
2. Be easy to compute
3. Be rigidly defined

4. Be based on each and every item of the distribution
5. Be capable of further algebraic treatment
6. Have sampling stability
7. Not be unduly affected by extreme items

The range:

The range is the difference between the value of the largest item of data and the smallest item of data in the set of observations.

$$\text{Range} = \text{Largest item (L)} - \text{Smallest item (S)}$$

For data grouped into classes or categories, the range can either be found by finding the difference between the upper limit of the highest class and the lower limit of the lowest class, or by finding the difference between the midpoint of the highest class and the midpoint of the lowest class.

Its corresponding relative measure is known as the coefficient of range and is obtained by applying the formula

$$\text{Coefficient of range} = \frac{L-S}{L+S}$$

Merits of the range:

This is the simplest measure to understand and compute compared to any other measures of dispersion.

Limitations:

1. It computation is not based on every item of the distribution.
2. It has no sampling stability because it is subject to fluctuations of considerable magnitude from sample to sample.
3. It cannot tell us anything about the distribution within the two extreme observations.
4. It cannot be computed for open ended distributions without making assumptions which will in turn result to in errors.

Uses of the range

1. Quality control to ensure the difference between the largest and smallest of mass produced items does not exceed a certain value
2. Fluctuations in share prices
3. Weather forecasts e.g difference between maximum and minimum temperature

Interquartile range:

This is the difference between the upper quartile and the lower quartile in a distribution, thus, the interquartile range $Q_3 - Q_1$

It is reduced to the form of the semi-interquartile range or quartile deviation by dividing it by 2 that is:

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

The quartile deviation gives the average amount by which the two quartiles differ from the median. In a symmetrical distribution the two quartiles are equidistant from the median.

A very small value of quartile deviation indicates that the variation of the middle 50% items is small, likewise a high quartile deviation means variation of the middle 50% items is large.

The relative measure corresponding to quartile deviation is the coefficient of quartile deviation.

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

Coefficient of quartile deviation can be used to compare the degree of variation of different distributions.

Merits of Quartile Deviation

1. In certain respects it is superior to the range as a measure of dispersion.
2. Can be used in open ended distributions.
3. Useful in erratic or badly skewed data

Demerits of Quartile deviation

1. It does not depend on each and every item in the data set. It ignores the top 25% and the bottom 25% of items in a distribution.
2. It is not capable of mathematical manipulation
3. Its value is affected by sampling fluctuations
4. It is more of a positional average rather than a measure of dispersion as it does not show scatter around an average.

Average deviation or Mean deviation

This is the average of the absolute values of deviations from the mean given as

$$\text{Mean deviation} = \frac{\sum|x - \bar{x}|}{n}$$

Example:

The number of cappuccinos sold at a cafeteria for a sample of five days is 20,40,50,60 and 80. The mean is number of cappuccinos sold $\bar{x} = 50$

No of cappuccinos(x)	$x - \bar{x}$	$\lvert x - \bar{x}\rvert$
20	-30	30
40	-10	10
50	0	0
60	10	10
80	30	30
		Total $= 80$

$$\text{Mean deviation} = \frac{80}{5} = 16$$

The number of cappuccinos sold deviates on average by 16 from the mean of 50 per day.

The relative measure corresponding to mean deviation is the coefficient of mean deviation computed as:

$$\text{Coefficient of Mean deviation} = \frac{\text{Mean deviation}}{\text{Mean}}$$

Merits

1. It is simple to understand and easy to compute
2. It is based on each and every item of the data
3. It is less affected by extreme values

Limitations

1. Ignoring of the signs makes the method non algebraic
2. It is not capable of further algebraic treatment.
3. It may not give very accurate results because mean deviation gives best results when deviations are taken from the median, but the median is not a satisfactory measure when variability is high in the data set.

Variance and Standard Deviation

Variance is the arithmetic mean of the square of the deviations from the mean.

$$\text{Population variance } (\sigma^2) = \frac{\sum f(x-\mu)^2}{n}$$

$$\text{Sample variance } (s^2) = \frac{\sum f(x-\bar{x})^2}{n-1}$$

The primary use of the sample statistics like (s^2), is to estimate the population parameters like (σ^2). $n-1$ is used in the denominator for sample variance because using n would under estimate the population variance. Dividing by $n-1$ gives a slightly larger value and an unbiased estimate of the population variance.
Variance is difficult to interpret because the units are squared. For example if original data was heights of trees in meters, the units of variance would be meters squared. The standard deviation is the square root of variance

$$\text{Population standard deviation } (\sigma) = \sqrt{\frac{\sum f(x-\mu)^2}{n}}$$

$$\text{Sample standard deviation } (s) = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}}$$

The respective computational formulae for population and sample variance are

$$\sigma^2 = \frac{\sum fx^2}{n} - \left(\frac{\sum fx}{n}\right)^2$$

And

$$s^2 = \frac{\sum fx^2 - \left(\frac{(\sum fx)^2}{n}\right)}{n-1}$$

For grouped data midpoints of classes are taken as values of x

Standard deviation is an absolute measure of variation. The corresponding relative measure of variation is the coefficient of variation computed as:

$$\text{Coefficient of variation } (\text{C.V}) = \frac{s}{x} \times 100\%$$

It is used when making comparisons of variability of two or more data sets. The data set with a higher coefficient of variation has more variation and that with a lower value has less variation or is more consistent or uniform or homogeneous.

Merits of Standard Deviation

1. It is the best measure of variation. It is based on every item of the data set
2. It lends itself to further algebraic treatment. It is possible to calculate the combined standard deviation of two or more groups.
3. It is used in further statistical work for example in computing skewness and correlation.

Limitations of Standard deviation

1. It is difficult to compute
2. It gives more weight to extreme items and less to those which are near the mean.

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.