Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 25

Variation Measures based on Absolute Deviations

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

http://home.iitk.ac.in/~shalab/sp

Deviation Based Measures of Variation

Range, interquartile range and quartile deviation are based on specific values and partitioning values.

Need a measure which can measure the deviation of every observation around any given value.

Deviation Based Measures of Variation

Deviation of any observation x_i from any value A is $d_i = (x_i - A)$.

If $x_i > A$, then such deviations d_i's are positive.

If $x_i < A$, then such deviations d_i's are negative.

If $x_i = A$, then such deviations d_i's are zero.

Deviation Based Measures of Variation

If we consider the average of these deviations d_i's, then the average value $\frac{1}{n}\sum_{i=1}^{n} d_i$ may be close to zero and will be reflecting that there is no variation or small variation, which may not be correct.

So we need not to consider the signs of the deviations.

We need to consider ONLY the magnitudes of the deviations.

Deviation Based Measures of Variation

Two options:

1. **Consider absolute values of these deviations.**

2. **Consider squared values of these deviations.**

Notations for Ungrouped (Discrete) Data

Observations on a variable X are obtained as $x_1, x_2, \ldots, x_n$.

Notations for Grouped (Continuous) data

Observations on a variable X are obtained and tabulated in K class intervals in a frequency table as follows. The mid points of the intervals are denoted by $x_1, x_2, \ldots, x_k$ which occur with frequencies $f_1, f_2, \ldots, f_K$ respectively and $n = f_1 + f_2 + \ldots + f_K$.

Class intervals	Mid point (x_i)	Absolute frequency (f_i)
$e_1 - e_2$	$x_1 = (e_1 + e_2)/2$	f_1
$e_2 - e_3$	$x_2 = (e_2 + e_3)/2$	f_2
…	…	…
$e_{K-1} - e_K$	$x_K = (e_{K-1} + e_K)/2$	f_K

Absolute Deviation

The absolute deviation of observations $x_1, x_2, \ldots, x_n$ around a value A

is defined as

- $$D(A) = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - A\right|$$ **for discrete (ungrouped) data.**

- $$D(A) = \frac{1}{n}\sum_{i=1}^{K} f_i \left|x_i - A\right|$$ **for continuous (grouped) data.**

  where $n = \sum_{i=1}^{K} f_i$

Absolute Mean Deviation

The absolute deviation of observations $x_1, x_2, \ldots, x_n$ is minimum when measured around median, i.e., *A* is the median of data.

In this case, the absolute deviation is termed as absolute mean deviation and is defined as

- $$D(\bar{x}_{med}) = \frac{1}{n}\sum_{i=1}^{n} \left| x_i - \bar{x}_{med} \right|$$ **for discrete (ungrouped) data.**

- $$D(\bar{x}_{med}) = \frac{1}{n}\sum_{i=1}^{K} f_i \left| x_i - \bar{x}_{med} \right|$$ **for continuous (grouped) data.**

  where $n = \sum_{i=1}^{K} f_i$

Absolute Mean Deviation

The absolute mean deviation measures the spread and scatterdness of data around, preferebly the median value, in terms of absolute deviations.

Absolute Deviation and Absolute Mean Deviation
Decision Making

The data set having higher value of absolute mean deviation (or absolute deviation) has more variability.

The data set with lower value of absolute mean deviation (or absolute deviation) is preferable.

If we have two data sets and suppose their absolute mean deviation are AMD_1 and AMD_2.

If $AMD_1 > AMD_2$ then the data in AMD_1 is said to have more variability than the data in AMD_2.