Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 27

Mean Squared Error, Variance Standard Deviation and Standard Error

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

Slides can be downloaded from

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

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

Mean Squared Error

We considered the absolute deviation values $|x_i - A|$ in absolute deviation. Instead of this, consider squared values of deviations $(x_i - A)^2$ around any point A.

Then the mean squared error (MSE) with respect to A is defined as

- $s^2(A) = \frac{1}{n}\sum_{i=1}^{n}(x_i - A)^2$ **for discrete (ungrouped) data.**

- $s^2(A) = \frac{1}{n}\sum_{i=1}^{K} f_i (x_i - A)^2$ **for continuous (grouped) data.**

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

Variance

$s^2(A)$: mean squared error (MSE) with respect to A is minimum when A is the arithmetic mean of the data, i.e., $A = \bar{x}$.

In this case, $s^2(\bar{x})$ is called as variance and is defined as

- $$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \quad ,$$ **for discrete (ungrouped) data.**

  where $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  $$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Variance

- $s^2 = \frac{1}{n}\sum_{i=1}^{K} f_i \left(x_i - \bar{x}\right)^2$, **for continuous (grouped) data.**

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{K} f_i x_i$, $n = \sum_{i=1}^{K} f_i$

$$s^2 = \frac{1}{n}\sum_{i=1}^{K} f_i x_i^2 - \bar{x}^2$$

Another form of variance: Divisor *n* - 1

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$ **for discrete (ungrouped) data.**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{K} f_i\left(x_i - \overline{x}\right)^2$$ **for continuous (grouped) data.**

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

Standard Deviation

s^2 : (Sample) Variance

s : Positive square root of s^2 is called as (sample) standard deviation (sd).

σ^2 : (Population) Variance.

σ : (Population) standard deviation.

More popular notation among practitioners

Standard Deviation

Standard deviation (or standard error) has an advantage that it has the same units as of data, so easy to compare. .

For example, if x is in meter, then s^2 is in meter^2 which is not so convenient to interpret.

On the other hand, if x is in meter, then s is in meter which is more convenient to interpret.

Variance

Variance (or standard deviation) measures how much the observations vary or how the data is concentrated around the arithmetic mean.

Variance

Decision Making

Lower value of variance (or standard deviation, standard error) indicates that the data is highly concentrated or less scattered around the mean.

Higher value of variance (or standard deviation, standard error) indicates that the data is less concentrated or highly scattered around the mean.

Same is followed for mean squared error (MSE).

Variance

Decision Making

The data set having higher value of variance (or standard deviation) has more variability.

The data set with lower value of variance (or standard deviation) is preferable.

If we have two data sets and suppose their variances are Var_1 and Var_2.

If $Var_1 > Var_2$ then the data in Var_1 is said to have more variability (or less concentration around mean) than the data in Var_2.