Descriptive Statistics With R Software

Variation in Data

::

Mean Squared Error, Variance and Standard Deviation

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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

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

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

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

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.

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.

Variance vs. Absolute Median Deviation

Since in the presence of outliers, median is less affected and arithmetic mean is more affected, so absolute median deviation is preferred over variance (or standard deviation).

Variance has its own advantages.

Variance

Difference between standard deviation and standard error.

Statistic: A function of random variables $X_1, X_2, \ldots, X_n$ is called as statistic. For example, mean of $X_1, X_2, \ldots, X_n$, denoted as $\overline{X}$, is a random variable.

Standard error: When we find the standard deviation of a statistic, it is called as standard error.

Variance

Difference between standard deviation and standard error

Ideally, standard deviation (sd) is a function of unknown parameter.

Let μ be the parameter representing the population mean, which is usually unknown, then the standard deviation is defined as

$$sd = +\sqrt{\text{var}(x)} = \sqrt{\sigma^2} = \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Difference between standard deviation and standard error:

Since μ is unknown, σ^2 can not be found.

So we can estimate μ by the mean of given sample observations.

Replace μ by sample mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Then the standard error is defined as

$$se = +\sqrt{\text{var}(x)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Difference between standard deviation and standard error:

Then, the variance $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2$ **becomes**

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

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

Variance

R command: Ungrouped data

Data vector: `x`

R command for variance

```
var(x)
```

R command `var(x)` gives the variance with divisor (*n* – 1) as

$$\text{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

R command to get the variance with divisor *n* as $s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$

`((n - 1)/n)*var(x)` where `n = length(x)`

Variance

R command: Grouped data

Data vector: x

Frequency vector: f

Variance of x

```
sum(f * (x - xmean)^2)/sum(f)
```

Variance

R command: Ungrouped data and missing values

If data vector `x` has missing values as `NA`, say `xna`, then R command is

```
var(xna, na.rm = TRUE)
```

Standard Deviation

R command: Ungrouped data

Data vector: `x`

R command for standard deviation based on the variance with divisor (*n* – 1) is

```
sqrt(var(x))
```

R command for standard deviation based on the variance with divisor *n* is

```
sqrt(((n - 1)/n)*var(x))
```

where `n = length(x)`

Standard Deviation

R command: Grouped data

Data vector: `x`

Frequency vector: `f`

Standard deviation of `x` is

```
sqrt(sum(f * (x - xmean)^2)/sum(f))
```

Variance and Standard Deviation

Example: Ungrouped data

Following are the time taken (in seconds) by 20 participants in a race: 32, 35, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time = c(32, 35, 45, 83, 74, 55, 68, 38, 35,
55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58)
```

```
> var(time)
```
 # variance with divisor (n-1)

```
[1] 283.3684
```

```
> sqrt(var(time))
```
 # standard deviation with divisor (n-1)

```
[1] 16.83355
```

Variance and Standard Deviation

Example: Ungrouped data

```
> ((length(time) - 1)/length(time))*var(time)
   [1] 269.2                    # variance with divisor n

> sqrt(((length(time) -  1)/length(time))*var
(time))                  # standard deviation with divisor n
[1] 16.40732
```

Variance and Standard Deviation

Example: Ungrouped data

R Console

```
> time
 [1] 32 35 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
>
> var(time) # variance with divisor (n-1)
[1] 283.3684
>
> sqrt(var(time)) # standard deviation with divisor (n-1)
[1] 16.83355
>
> ((length(time) - 1)/length(time))*var(time) # variance with divisor n
[1] 269.2
>
> sqrt(((length(time) - 1)/length(time))*var(time)) # sd with divisor (n-1)
[1] 16.40732
>
```

Variance and Standard Deviation

Example: Grouped data

Considering the data as grouped data, we can present the data as

Class intervals	Mid point	Absolute frequency (or frequency)
31 – 40	35.5	5
41 – 50	45.5	3
51 – 60	55.5	3
61 – 70	65.5	5
71 – 80	75.5	2
81 - 90	85.5	2
	Total	20

We need to find the frequency vector and mean.

Variance and Standard Deviation

Example: Grouped data

Using the following commands, we get finally the frequency vector:

```
> breaks = seq(30, 90, by=10)
> time.cut = cut(time,breaks,right=FALSE)
> table(time.cut) # Frequency distribution


> f=as.numeric(table(time.cut)) # Extract frequencies
> f
[1] 5 3 3 5 2 2

> x = c(35,45,55,65,75,85) #Mid points from  frequency table
> x
[1] 35 45 55 65 75 85
```

Variance and Standard Deviation

Example: Grouped data

Data vector: `x`

Frequency vector: `f`

Mean of `x` is

```
xmean = sum(f * x)/sum(f)
```

```
> xmean = sum(f * x)/sum(f)
> xmean
[1] 56
```

Variance and Standard Deviation

Example: Grouped data

Variance of `x`

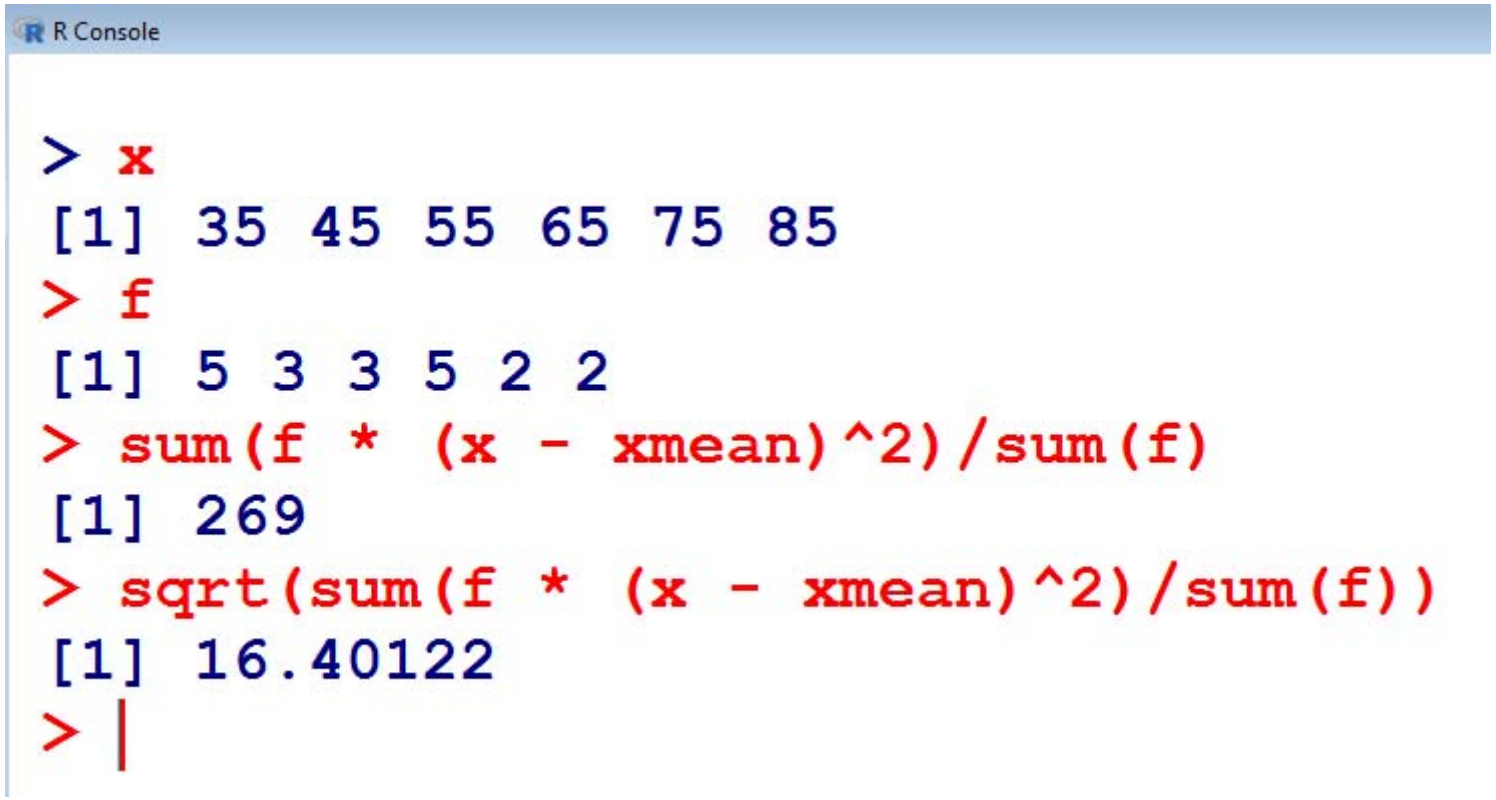
```
> sum(f * (x - xmean)^2)/sum(f)
[1] 269
```

Standard deviation of `x`

```
> sqrt(sum(f * (x - mean(x)))^2/sum(f))
[1] 16.40122
```

Variance and Standard Deviation

Example: Grouped data



Variance and Standard Deviation

Example: Handling missing values

Suppose two data points are missing in the earlier example where the time taken (in seconds) by 20 participants in a race. They are recorded as NA

NA, NA, 45, 83, 74, 55, 68, 38, 35, 55, 66, 65, 42, 68, 72, 84, 67, 36, 42, 58.

```
> time.na
 [1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68
72 84 67 36 42 58

> var(time.na, na.rm=TRUE) # variance
[1] 250.2647

> sqrt(var(time.na, na.rm=TRUE)) # standard deviation
[1] 15.81976
```

Variance and Standard Deviation

Example: Handling missing values