Exploratory Statistical Data Analysis With R Software (ESDAR)

Swayam Prabha

Lecture 28

Variance, Standard Error and Their Computations in R

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 with mid points of the intervals as $x_1, x_2, \ldots, x_k$ 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

Variance: Divisor *n*

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

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

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

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

Variance vs. Absolute Mean Deviation

Since in the presence of outliers, median is less affected and arithmetic mean is more affected, so absolute mean 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

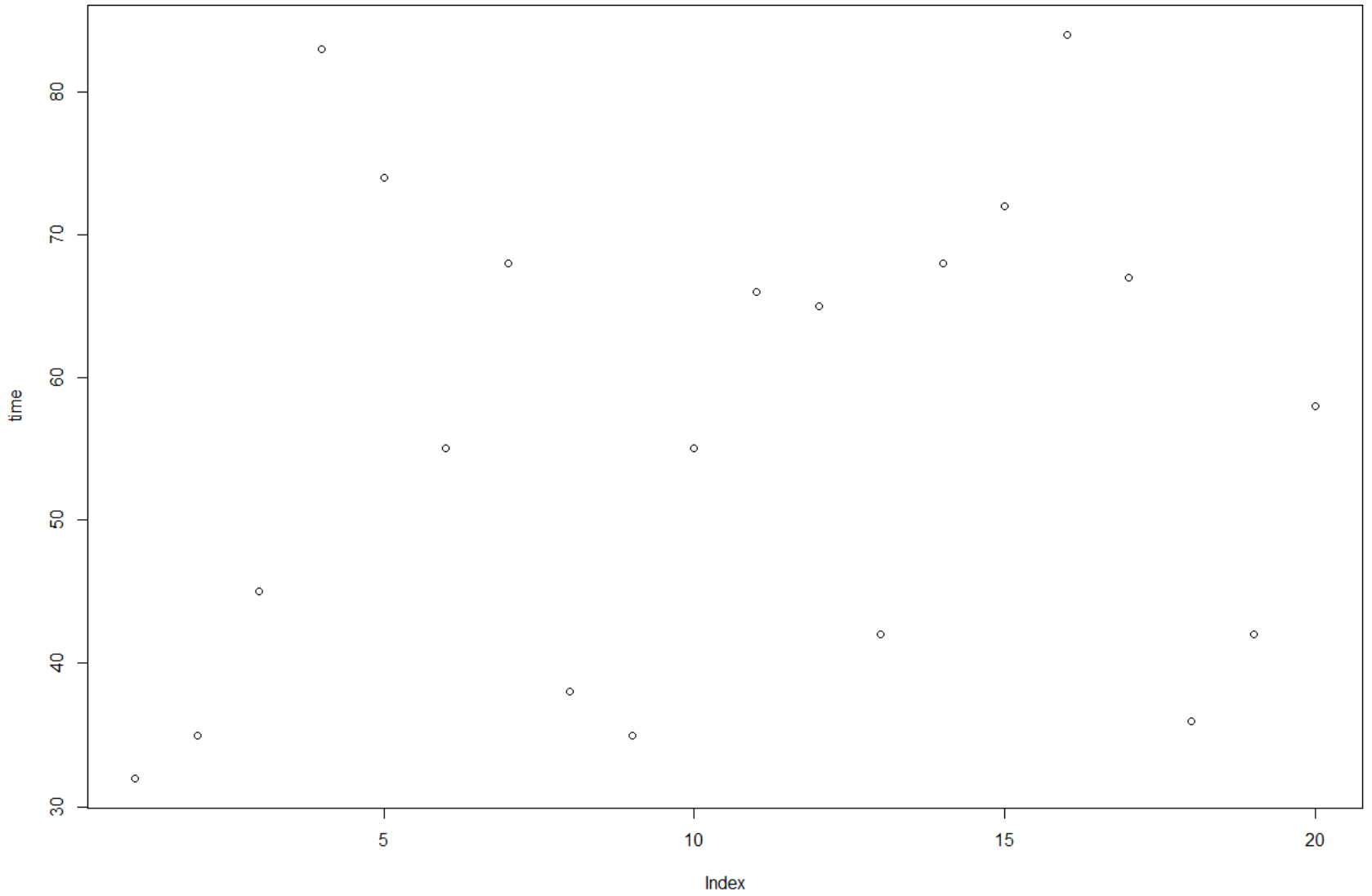
Example: Ungrouped data

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)
```

Variance and Standard Deviation

```
plot(time)
```



Variance and Standard Deviation

Example: Ungrouped data

```
> 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

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)
```

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 median.

Variance and Standard Deviation

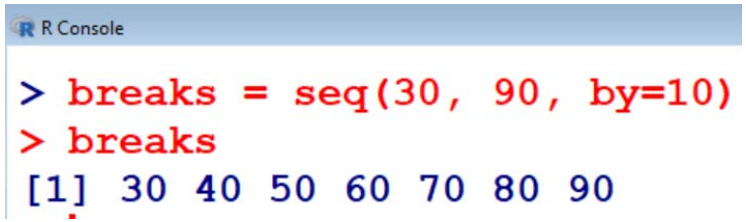
Example: Grouped data - Obtaining frequencies:

Create a sequence starting from 30 to 90 at an interval of 10 integers denoting the width.

```
breaks = seq(30, 90, by=10)
```

Sequence of 10 integers at interval of 10

```
> breaks = seq(30, 90, by=10)
> breaks
[1] 30 40 50 60 70 80 90
```

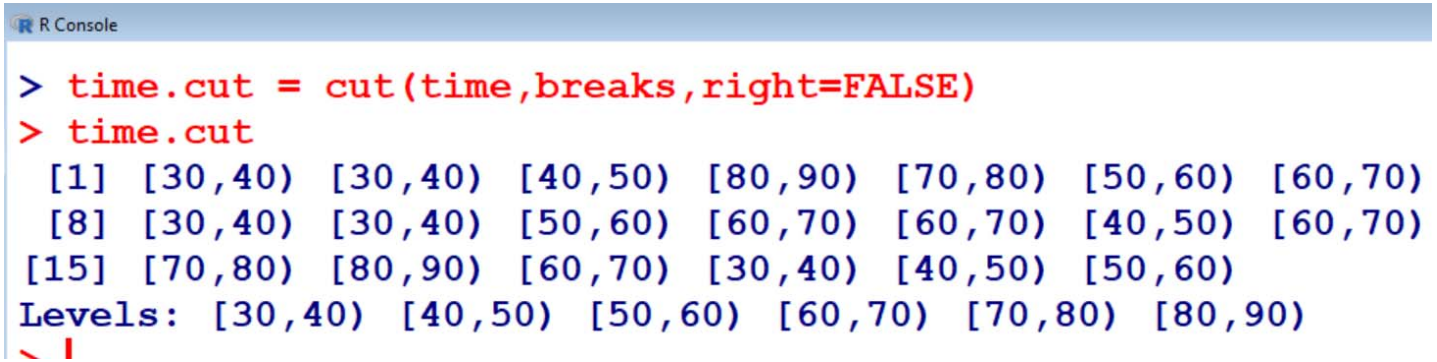


Variance and Standard Deviation

Example: Grouped data - Obtaining frequencies:

Now we classify the time data according to the width intervals with `cut`.

```
> time.cut = cut(time,breaks,right=FALSE)
> time.cut
 [1] [30,40) [30,40) [40,50) [80,90) [70,80) [50,60) [60,70)
 [8] [30,40) [30,40) [50,60) [60,70) [60,70) [40,50) [60,70)
[15] [70,80) [80,90) [60,70) [30,40) [40,50) [50,60)
Levels: [30,40) [40,50) [50,60) [60,70) [70,80) [80,90)
```



Variance and Standard Deviation

Example: Grouped data - Obtaining frequencies:

Frequency distribution

```
> table(time.cut)
time.cut
[30,40) [40,50) [50,60) [60,70) [70,80) [80,90)
      5       3       3       5       2       2
```

Extract frequencies from frequency table using command

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

Variance and Standard Deviation

Example: Grouped data - Obtaining mid points:

Mid points, as obtained from the frequency table, are

```
> x = c(35,45,55,65,75,85)

> x
[1] 35 45 55 65 75 85
```

Note that the mid points are obtained from the frequency table obtained from the R software

```
[30,40) [40,50) [50,60) [60,70) [70,80) [80,90)
```

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`

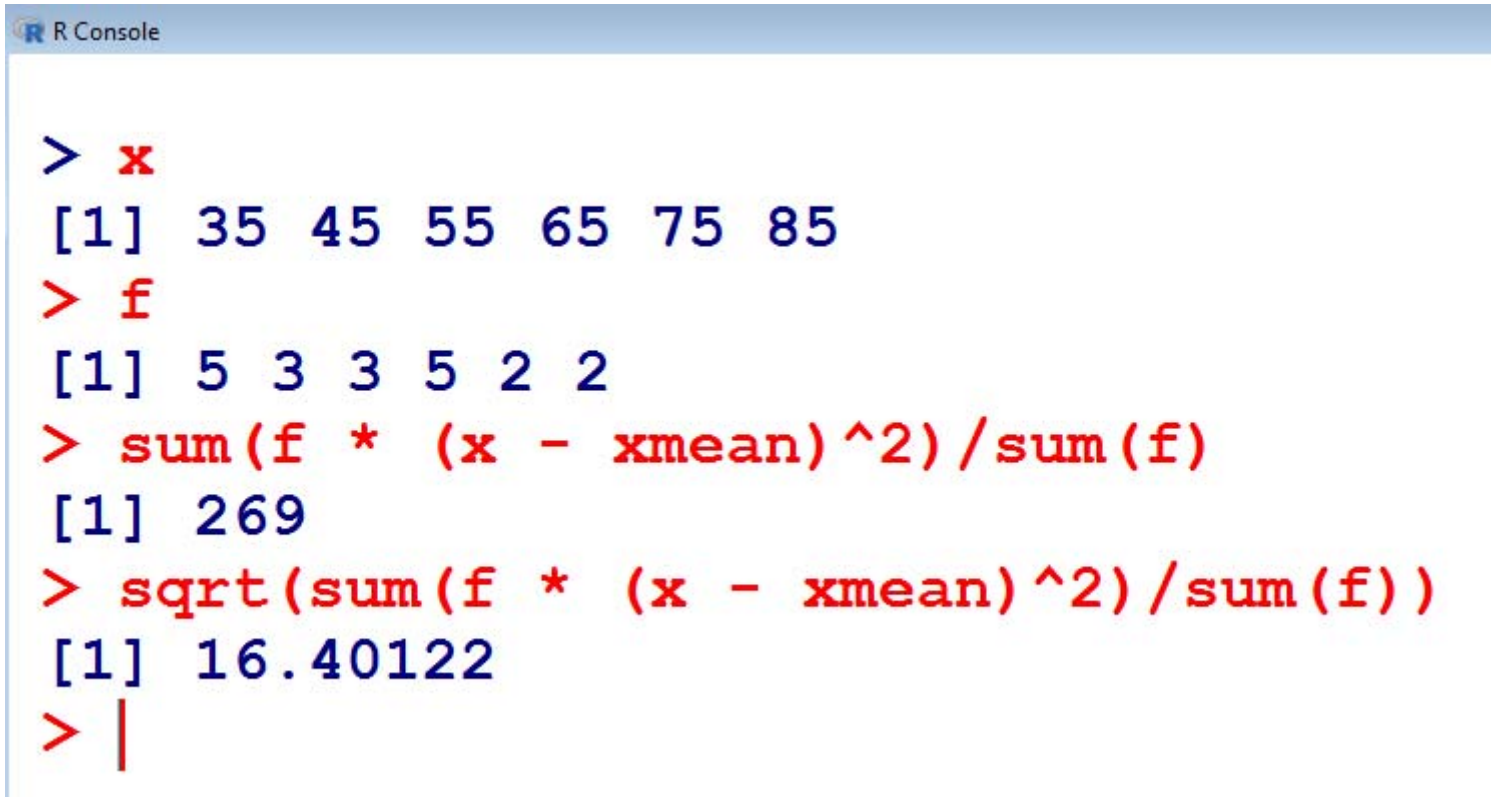
```
> sum(f * (x - xmean)^2)/sum(f)
[1] 269
```

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