

**Lecture – 20**  
**Variation in Data – Mean Squared Error, Variance and Standard Deviation**

Welcome to the, next lecture on the course,

Refer Slide Time: (00:17)

# 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

Descriptive Statistics with R software. You may recall that, in the earlier lecture, we had considered the aspect of measuring the variation, by using the absolute deviation and absolute mean deviation, and if you recall, we had developed that tool, on the concept of deviations of the observations, around any arbitrary point or around some central tendency value like as mean, or say median like this. We also discussed that, whenever we want to develop these types of tools, we had to convert that deviations only into a positive value, that means, we needed to convert or we needed to consider only the magnitude of the deviations, and for that we have two options. You try to consider the absolute value of the deviation or second option is that, we can consider the square values of those deviations. So, now in this lecture, we are going to, to discuss the second aspect that, how to build up or measure of variation by considering, the magnitude of the deviations by squaring them. So, in this lecture, we are going to discuss the concept of mean squared error, variance and standard deviation, and we will try to see that, how to implement it on a R software. So, before we start, let us try to the fix our notations once again. Although, I had done it in the earlier lecture but, here I will be just doing it quickly.

Refer Slide Time: (02:00)

### Notations for Ungrouped (Discrete) Data

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

$X \rightarrow n$  observations

So, he may be considering here, two types of variable, one discrete variable on which we will have an group data, and in this case, the variable is going to be denoted by capital letter X, and on this variable, we are going to obtain the n observations and these observations are denoted here like small  $x_1$ , small  $x_2$ , small  $x_n$ .

Refer Slide Time:(02:25)

**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  $x_1, x_2, \dots, x_k$  which occur with frequencies  $f_1, f_2, \dots, f_k$  respectively and  $n = f_1 + f_2 + \dots + f_k$ .

Class intervals	Mid point ( $x_i$ )	Absolute frequency ( $f_i$ )
$e_1 - e_2$	$x_1 = (e_1 + e_2)/2$	$f_1$
$e_2 - e_3$	$x_2 = (e_2 + e_3)/2$	$f_2$
...	...	...
$e_{k-1} - e_k$	$x_k = (e_{k-1} + e_k)/2$	$f_k$

Similarly, when we are trying to consider, a continuous variable here, and we have grouped data on that variable, this means, we have a variable, continuous variable X on which we have obtained the observations and those observation have been tabulated in K classes, or say K class intervals, and the entire revelation has been presented in the form of a frequency table. For example, here you can see the frequency table in which all the observations have been converted into the groups, and these groups are the class intervals they are denoted by here even to  $e_2$ ,  $e_2$  to  $e_3$  and so on. So, we have here K class intervals, or say K groups and after this whatever is the midpoint of this interval, that is going to be denoted by here  $x_1$ . So,  $x_1$  is going to denote the, midpoint of the first-class interval,  $x_2$  is going to denote the, midpoint of the second-class interval and so on. And all these  $x_1, x_2, x_k$ . Here, now in the case of group data, they are going to denote the midpoint and not the value of the observation, as it happens in the case of one group data and the frequency of these intervals is, they converted by an  $f_1, f_2, f_k$ . So,  $f_1$  is going to denote the frequency of the first-class interval,  $f_2$  is going to denote the frequency of the second-class interval and so on. And the, some of all this frequency is denoted by here n. So, that is going to be our basic notations for grouped n and grouped data, So, as soon as I say that, we are going to define the

measures on grouped data I will be using these to know a notations and as soon as I say that, I am going to develop the tool for the ungrouped data, then I will be using the earlier symbols and notations, Right! Now, I come on the aspect of, developing a tool called as mean squared error and I will be following the lecture almost, on the same line as I did in the earlier lecture, you may recall that first, I define the absolute deviations and these deviations were defined around any arbitrary value A, and then I developed the measure, and then I replace A, by some measure of central tendency and we define the absolute mean deviation by replacing A, by the median, because median was the value, around which the absolute deviations were minimum. Now, similarly on the same lines, now instead of absolute deviation, I will be considering the squares of the deviations, Right! So, if you remember,

Refer Slide Time:(05:28)

**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.

$$\frac{(x_1 - A)^2 + (x_2 - A)^2 + \dots + (x_n - A)^2}{n}$$

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.

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

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

that in the case of absolute deviation, I have used the quantity absolute value of  $x_i$  minus A. Now, I will try to, consider the squared values of  $x_i$  minus A, and I will write down here the squares of these deviations as  $x_i$  minus A whole square. So, these deviations are, the squared deviation around any arbitrary point A. Now, I will try to obtain, this quantity for each and every observation, say  $x_1$  minus A whole square,  $x_2$  minus A whole square, and up to here  $x_n$  minus A whole square, and then, I will try to take the arithmetic mean of all these values, and once I try to do it, then in the case of ungrouped data, discrete variable, this quantity is denoted by here like this, you can see here this, is the same quantity what I have given here. This quantity, is denoted here as, S square A, which is called as mean squared error with respect to A. And similarly, whenever we have continuous variable or grouped data, then in that case, the mean squared error, with respect to any arbitrary value A is defined here like this. So, you can

see here that, now this summation is going over the number of classes, and here this is small n is here the sum of all the frequencies, and this quantity is as sort of weighted mean, where the weights have been given by the frequency  $f_i$ . So, this is how we try to define the mean squared error in case of grouped data and ungrouped data with respect to any arbitrary value A, Right!

Refer Slide Time:(07:46)

**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.,  $A = \bar{x}$ .  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

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

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

for discrete (ungrouped) data.

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

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

Handwritten derivation:

$$s^2(\bar{x}) \equiv s^2$$

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

$$= \frac{1}{n} \sum_{i=1}^n x_i^2 + \frac{n}{n} \bar{x}^2 - 2\bar{x} \left( \frac{\sum_{i=1}^n x_i}{n} \right)$$

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

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

Now, one can choose any value of A, so this can be shown mathematically that, the mean squared error is going to assume, the minimum value, when A, chooses the value arithmetic mean or in simple words, I can say, mean squared error takes the minimum value, when the deviations are measured around the arithmetic mean. So, now what I will do? I will try to replace capital A by x bar, which is the sample mean, or the mean of the observations, and when I try to do it, then what happens, that I simply try to replace A by here x bar, x bar here is 1 over n, summation i goes from 1 to n,  $x_i$ , that is the arithmetic mean, and then I try to define here the deviations,  $x_1$  minus x bar whole square,  $x_2$  minus x bar whole square, and up to here  $x_n$  minus x bar whole square. And then I try to find out, the average of these things simply arithmetic mean and this quantity is denoted here, in the case of ungrouped data or the discrete variable, that these values are the values which are given here, in case of a ungrouped data or a data on the discrete variable. So, you can see here, this is the same thing and this quantity which is essentially here, s s square, x bar which is in general, we denote by here, s s square, this is called the variance, and in case, if you try to simplify this expression, so you can write down here, this same thing here 1 upon n, i goes from 1 to here n,  $x_i$  square plus x bar square, minus twice of  $x_i$  in to x bar. So, this comes out to be 1 over n, summation, n  $x_i$  square, plus this will be come here, n upon n x bar square, minus twice of x bar,

summation  $x_i$  goes from 1 to n upon n. Now, this can be further simplified to i goes from 1 to n, summation  $x_i$  square, divided by n, plus  $\bar{x}$  square, minus this quantity is simply the  $\bar{x}$ . So, this becomes here, twice of  $\bar{x}$  square. So, this quantity becomes a 1 upon n, summation i goes from 1 to n,  $x_i$  square minus  $\bar{x}$  square. So, the alternative expression for the variance is here given by this thing, which is the same thing. So, actually you can use any of this expression to compute the value of variance in case of ungrouped data.

Refer Slide Time:(11:05)

**Variance**

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

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

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

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

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

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

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

And similarly, when we have grouped data on a continuous variable, then in this case, the variance is defined like this. So, this is 1 upon n summation i goes from 1 to K, where K is the number of classes, and then multiplied by  $f_i$  into  $x_i$  minus  $\bar{x}$  whole square like this, and where  $\bar{x}$  is defined like this, and this is small n, is defined as the sum of total frequencies. Similarly, if you want to simplify this expression as we did in the earlier case, you can see here, this will also come out to be i goes from 1 to K,  $f_i$ ,  $x_i$  squared, plus  $\bar{x}$  square, minus twice of  $x_i$ ,  $\bar{x}$ , and if you try to see here, this is 1 upon n summation i goes from 1 to K,  $f_i$ ,  $x_i$  square, plus  $\bar{x}$  square, summation  $f_i$ , i goes from 1 to K upon n, minus twice of  $\bar{x}$ , summation i goes from 1 to K,  $f_i$ ,  $x_i$  upon here n. So, this quantity here, becomes here same as  $\bar{x}$ . So, I can write down here 1 upon n, i goes wrong 1 to K,  $f_i$ ,  $x_i$  square, plus  $\bar{x}$  square. Why because, the numerator summation  $f_i$  becomes here n upon n, minus here, twice of here  $\bar{x}$ , into  $\bar{x}$ , so this quantity comes here, 1 upon n, summation i goes from 1 to K,  $f_i$ ,  $x_i$  square, minus  $\bar{x}$  whole square, this is the same quantity given over here. So, any of this expression can be used to compute the variance in case of, grouped data.

Refer Slide Time:(13:10)

**Another form of variance: Divisor  $n - 1$**

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

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

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

Now, after giving the definition of variance, I would like to address here, one more aspect, you have seen that, in the definition of variance, I am trying to take the average of  $n$  value. So, this is defined as 1 upon  $n$ , summation  $i$  goes from 1 to  $n$ ,  $x_i$  minus  $\bar{x}$  whole square. In statistics, there is another expression of variance that is quite popular, and this expression is given by this, in case of ungrouped or discrete data, this is given by here this quantity, you can see here that this expression is like, the earlier one, the only difference here is that, instead of here  $n$ , I have here 1 upon  $n$  minus 1, and earlier we had only here  $n$ . So, this is what you have to keep in mind, and similarly in case of continuous data or a group data, the divisor that was earlier, and now this becomes here 1 over  $n$  minus 1, now an obvious question comes that, why I am using this expression? Actually, the properties of this expression, when we have divisor  $n$ , or when we have divisor  $n$  minus 1, they are different. If you have the idea of an unbiased estimator, in the context of statistical inference, then when I say or when I use the division  $n$  minus 1, then this form of the variance is an unbiased estimator of the population variance. Whereas, when I am using the divisor  $n$ , then 1 upon  $n$  summation  $x_i$  minus  $\bar{x}$  whole square, is not an unbiased estimator of the population variance. So, that is why many times the software uses this definition. For example, in the R software, this definition is used where the divisor is  $n$  minus 1. So, that is the reason, I would like to inform you here that, whenever you are using any software, please try to look into the manuals of the software and see what they are trying to do. Well, in case if the data is very large, then the value of the variances, may not differ much but, in case if the data is small, can the values of the variances computed by using the division  $n$  or say  $n$  minus 1, they may have difference. So, you should know, what you are obtaining, it should not happen



that, you are assuming that, the divisor is n, and your divisor is actually inside the software is n minus 1. Okay? So, just be careful.

Refer Slide Time:(16:15)

**Standard Deviation**

$s^2$  : (Sample) Variance → Variance calculated on the basis of given set of data or given sample of data

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

$\sigma^2$  : (Population) Variance.

$\sigma$  : (Population) standard deviation.

More popular notation among practitioners

Now, after the variance, I come to another concepts, that is called standard deviation. You have used possibly two types of terminologies; one is standard deviation and second is standard error. In general, people do not differentiate between these two names but here, I would try to explain you that, what is the difference between the two, but to start with, I will try to use the common terminology, and I will use here the standard deviation, Right! So, when I say that s square is going to denote my variance, then it is actually the sample variance, sample variance means, that the variance calculated on the basis of given set of data, or given set of, or given sample of data, Right! So essentially, we are trying to compute the sample variance, but we always call it without loss of generality as variance, when I try to take the positive square root of s square, then this is called as sample standard deviation. So, once again you can see here, I am writing here the sample word, inside the bracket just to indicate you, that the common language is simply the standard deviation, but here actually we are trying to compute the standard deviation on the basis of given sample of data, Right! Once I'm saying that the sample variance, or the variance, or the standard deviation has been computed, on the basis of given sample of data. What does this mean? If you try to see, in statistics what happens, that you are usually trying to collect a sample of data, and on the basis of sample of data your objective is to compute the population counterpart, you may recall that in the beginning of the lectures, we have discussed the concept of population and sample. So, suppose means, I would like to compute the total number of people, who are eating, say more rice in the country like India, which is a very huge country, ve very big country. Now, if I want to find out the



arithmetic mean, the average number of people, who are eating more rice than wheat, then what I have to do? I have to compute the number of such persons all over the country, which is very difficult to compute, unless and until you, you execute a census. So, we try to take a sample of data that means, I will try to choose a small number of observation and based on that, I will try to find out the mean, and that will be called a sample mean. Similarly, if I want to find out the variance of the data that I have collected inside the sample, then that will be called as sample variance, but there will always be a counterpart like as population mean or population variance. Population mean, means the arithmetic mean of all the population, simply the population variance means, the variance of all the units inside the population. So, what happened that when we are trying to compute the positive square root of the population variance, then this is called as standard deviation. But the problem is this we do not have the entire population in our hand. So, we always work on the basis of a sample of data, and that is why in a common language people, do not differentiate much between the two definition, a standard error and standard deviation. But, once you are trying to do a course on statistics as a student you must know, it. So, and that is my idea to explain this concept here in the next couple of slides. Okay?

Refer Slide Time:(20:50)

**Standard Deviation**

$s^2$  : (Sample) Variance → Variance calculated on the basis of given set of data or given samples of data

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

$\sigma^2$  : (Population) Variance.

$\sigma$  : (Population) standard deviation.

More popular notation among practitioners

So, Now, I will try to denote Sigma square as the population variance, and the positive square root of Sigma square that will be called as standard deviation or this is actually the population standard deviation. But, as I said, they say using Sigma to denote the standard deviation and using Sigma square to denote the population variance is a more popular notation among the practitioners.

Refer Slide Time:(21:16)

## 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  $\text{meter}^2$  which is not so convenient to interpret.

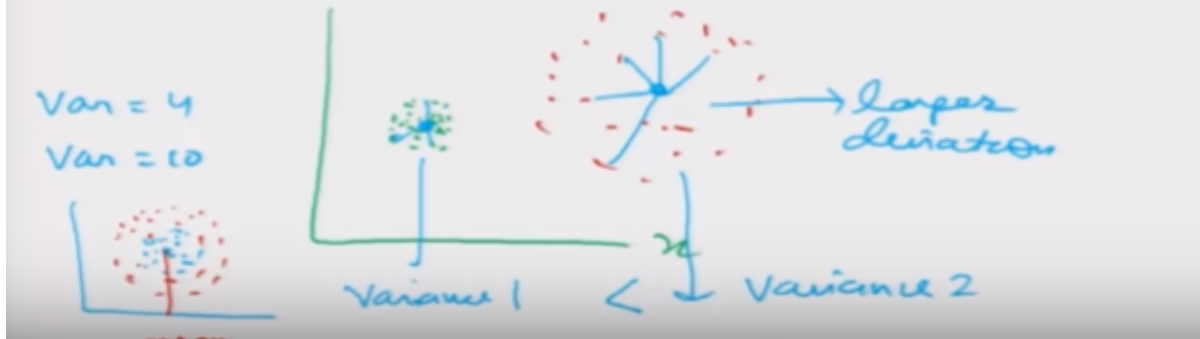
On the other hand, if  $x$  is in meter, then  $s$  is in meter which is more convenient to interpret.

Now, next question comes. What is the advantage of having a standard deviation or say standard error in place of variance? Now, if you try to see suppose I am trying to collect the data on the height of some children, say in meters. So, the arithmetic mean will be in meters. But, the variance will be in meter square .So, if I try to see that say that there are two data set whose variances are say 16 and 36 .Then, it is more convenient to compare the two values if they are in the same units as of the original data set. So, what we try to do we try to find out, the positive a square root of 16 and 36 which are 4 and 6 respectively .So, Now, once I say that, I have got a data set in which I have measured the heights of the children in meter, and then they have got the arithmetic mean of say of 1.2 meter and standard deviation of say 0.5 meter .Then, it is more easy to understand ,and similarly if I say that I have got two data sets in which the standard deviations are 4 and 6 .Then it is more easy to understand that the standard deviation of the second data set is higher than the standard deviation of the first data set that is the only reason actually. So, this the standard deviation or standard error has an advantage, that it has the same unit as of the original data. So, that is easy to compare. For example if I have a variable in which I have taken the observations denoted as ,say small  $x$  then if this is small  $x$  have been obtained in the unit meter ,then variance  $s$  square will be in meter square ,which is not so convenient to interpret also. on the other hand, if I have obtained the observation  $x$  in meter, then the standard deviation  $s$  will also be in meter, and which is more convenient to use ,more convenient to interpret .That is the reason that why people prefer to use the tool of standard deviation or standard error.

Refer Slide Time:(23:40)

## Variance

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



Now, the question comes what does this variance or standard deviation actually measure? The variance or equivalently standard deviations measure, how much the observations vary? or how the data is concentrated around the arithmetic mean? For example, if I try to take here a data set and suppose this data set is like this and suppose there is another data set which is here like this. So, you can see here in both the cases, the mean is somewhere here. But, these deviations, in the case of red dots, and in the case of green dots they are different, and the deviations in the case of red dots they are larger. So, in this case if I try to find out the value of the variance see here, variance 1, and here variance 2, then if I try to compute the value of variance 1 and variance 2 on the basis of given set of data, then we will find that variance 1, it is smaller than variance 2. So, whenever I have a value of variance, say variance equal to 4 and variance equal to here 10, then this obviously indicates that the data in the variance 1, is more concentrated, around the mean value, and the data with variance 10, suppose I'm denoting in the red dot that is more scattered around the mean value. Which is here like this. So, this is how we try to

take the interpretation of the value of the variance.

Refer Slide Time:(25:45)

## 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.

So, obviously when we want to make a decision on the basis of a given value of variance. Then the lower value of various are equivalently the standard error, standard deviation indicates, that the data is highly concentrated or less scattered around the automatic mean .Whereas the higher value of variance or the higher value of equivalently standard deviation or standard error indicates, that the data is less concentrated or highly scattered around the mean. So, this is the interpretation,

Refer Slide Time:(26:23)

## 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  $Var_2$ .

12

and obviously on the other hand if I have a data set which has got the higher value of variance or the standard error or standard deviation, then I can say simply that the data set has got, more variability. In statistics usually we always prefer to have a data set which has got the lower value of variance, or lower value of standard deviation, or standard error. So, in case if I have got two data sets, and suppose we compute the variances, and suppose these variances are coming out to be var 1 and var 2, then if var 1 is greater than var 2. Then we say, that the data in the variance 1 is having more variability, or less concentration, around the mean. Then the data in the dataset used for computing the variance 2.

Now, there is a very simple rule. We would always like to have a data which has got lower variance and if you remember in the initial slides I had discussed, that one of the basic objective in a statistical tool is that, we would always like to have a data, in which the variability is less. Okay?

Refer Slide Time:(27:41)

## **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.**

Now, in case if you try to compare the variance and absolute mean deviation, then you know, that when there are some outliers or some extreme observation inside the sample. Then the median is less affected than the arithmetic mean, and in this case means, if you the data has very high variability or the data has extreme observation or say outliers. Then in this case using the absolute mean deviation is a better option and it is preferred over the variance or standard deviation. Whereas variance has its own advantages. For example, if you are working with this statistical tool. Then the variance has some nice mathematical properties. So, from the statistics point of view, from the algebra point of view from the statistical analysis point of view, it is more easy to operate on the variance mathematically, algebraically, then on the absolute function, like absolute median deviation, or we call it a say, say here absolute mean deviation.

Refer Slide Time:(29:10)



## Variance

statistic

Difference between standard deviation and standard error.

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

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

← Random variables  
→ Statistic

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

Next, we try to understand what is the difference between standard deviation and standard error? To understand this thing, we have a concept which is called as a statistic. You see the spelling of the subject the statistics is s statistics. But we are not using here the last s, and this is called only as a statistic. As statistics is a function of random variables. So, if you have random variables  $X_1, X_2, X_n$ , then any function of  $X_1, X_2, X_n$  is called as a statistic. For example if I say you have random variables  $X_1, X_2, X_n$  and you try to find out the arithmetic mean like is here the capital  $\bar{X}$   $\frac{1}{n}$  summation I goes from 1 to n  $X_i$ . Then this is this  $\bar{X}$  is itself a random variable. So, this is called as the statistic.

Now, the concept is very simple. Whenever you are trying to find out the standard deviation of a statistic, then the outcome is again going to be a function of only the random variables, and this standard deviation is called as standard error. So, whenever we try to find out the standard deviation of a statistic it is called as standard error.

Refer Slide Time:(30:44)

## Variance

### Difference between standard deviation and standard error

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

Eg

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

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

population mean of  $x_i$ 's.  
Unknown

15

What does this mean actually? Actually, ideally what happens, that standard deviation is a function of some unknown parameters. For example, if you try to understand suppose  $\mu$  is representing the population mean. Right? The mean of all the units inside the population, which is very very large, and practically it is very difficult to find out the mean of entire population. So, usually it is unknown. Then ideally the standard deviation is defined as the positive square root of the variance of all the values in  $x$ . We which is denoted here that's a square root of Sigma square which is equal to here Sigma, and that is equal to a square root of 1 upon  $n$  summation  $X_i$  minus  $\mu$  whole square. So,  $\mu$  here is actually the population mean population mean of all the values  $x_i$ 's. But since this  $\mu$  is not known this is unknown. So, you cannot compute it. This value cannot be computed. Why because how you will get the value of  $\mu$ .

Refer Slide Time:(32:02)

## Difference between standard deviation and standard error:

Since  $\mu$  is unknown,  $\sigma^2$  can not be found.

So we can estimate  $\mu$  by the mean of given sample observations.

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

Then the standard error is defined as

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

S.E.: Will always be a function of observed values

$$\mu = \frac{1}{\text{popn size}} \sum_{i=1}^n x_i$$
 Population  
 unknown

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

So, then in that case the Sigma square cannot be found. So, one option is this, that we can replace the value of mu by its sample counterpart. So, when I say mu is here what this is the population mean? So, this is equal to 1 upon the total number of units in the population, which is population size, I goes from 1 to here. Population size and say here x i's. So, I try to replace it by the sample value So, this is for population and now for sample I try to replace by 1 upon see here n number of observation i goes from 1 to n. Say here x i and i denoted the automatic mean to be here x bar. So, now mu is unknown to us. But sample mean x-bar is known to us. So, what I can do? I can replace mu by the sample mean x-bar 1 upon n summation i goes from 1 to n xi and then in that case the standard error which is the positive square root of the variance becomes like this square root of 1 upon ends summation I goes from 1 to n xi minus x bar whole Square, and this quantity is called as standard error.

So, a standard error always remember, standard error will always be we are function of observed values. So, in simple language I can say that the standard error will always refer to a standard deviation which can always be computed on the basis of given sample of data .You have got the data you are asked to compute the, the standard deviation .You can compute it ,and in that case this will be called as standard error,

Refer Slide Time:(34:19)

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

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

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

*(sample variance)*

and in that case if you try to see here more specifically the population variance was defined here like this, and now this becomes s squared is equal to 1 upon n summation I goes from 1 to n xi minus x bar for the case when we have ungrouped data, and 1 upon n summation fi xi minus x bar whole square in case of we have group data. So, this is basically the definition of sample variance and in common language, we usually do not call again and again it to be sample. But we simply call it that find out the variance of the set of given data. Now, after this I, will come to the aspect that how are we going to compute the variance or standard deviation on the R software Right?

Refer Slide Time: (35:13)

## 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

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

*Handwritten notes:* A green arrow points from var(x) to the text above. A red circle highlights  $\frac{1}{n-1}$ . A red box highlights  $\frac{n-1}{n}$ . A red circle highlights  $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$  with the label  $\text{var}(x)$  written above it.

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

$((n - 1) / n) * \text{var}(x)$  where  $n = \text{length}(x)$

So I, will take the first case here the case of ungrouped data so, in this case the data vector is going to be denoted by say here a small x and the R command for computing the variance is here var and inside the argument you have to give the data vector. But remember one thing, this command var and inside the data vector x gives the variance with a divisor and minus 1 that is the this value so, in case if you want to obtain 1 upon n I goes from 1 to nxi minus x bar whole square quantity then what I can do I can multiply and divide it by the quantity n minus 1. So, now this I can write down here n minus 1 upon here n into 1 over n minus 1 summation I goes from 1 to nxi minus x bar whole square. So, in case if you are very particular in getting the divisor n in the variance then in that case I, would like to suggest that you try to use this command try to multiply the variance of (x) by the factor n minus 1 upon n. How? for example, you can see here now I'm writing in red color this quantity will give you variance of (x) and this quantity will be the factor by which if you try to multiply the variance you will get the variance with divisor n, and were n is the length of the (x)vector that with the number of observation present in the data set.

Refer Slide Time: (36:45)

## Variance

R command: **Grouped data**

Data vector: **x**

Frequency vector: **f**

Variance of **x**

$$\text{sum}(f * (x - \text{xmean})^2) / \text{sum}(f)$$
$$\frac{1}{n} \sum_{i=1}^k f_i (x_i - \underbrace{\bar{x}}_{x\text{mean}})^2$$
$$\frac{\text{sum}[f * (x - x\text{mean})^2]}{\text{sum}(f)}$$

Now, we are going to consider the case when we have the group data now in the case of group data you know that there is no built-in command in the base package of R so, I need to compute the mean value of the given data set separately along with the midpoints from the frequency table, and the frequencies from the frequency table. So, in this case we are going to compute the mean as say  $\bar{x}$  separately and in this case if you try to see your expression for variance was  $\frac{1}{n} \sum_{i=1}^k f_i (x_i - \bar{x})^2$  so, now this  $\bar{x}$  becomes  $x\text{mean}$  and then this quantity here I am trying to write down here  $x_i - \bar{x}$  whole square and then it has to be multiplied by the frequency vector here  $f$  and then this has to be summed means all the elements in this vector will be sum divided by the  $n$  which is the sum of all the elements in the frequency vector  $f$ . So, this is how this expression has been obtained to find out the variance in case of group data.

Refer Slide Time: (37:57)



## 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)`

And in case, if you have some missing data then in the case of ungrouped data if the data vector  $x$  has some missing values as NA, then we are going to denote this data vector as a here  $xna$ , and in that case the command remain the same but I have to give her the option `na dot rm is equal to true` Right?

Refer Slide Time: (38:21)

## 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))` =  $sd$

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

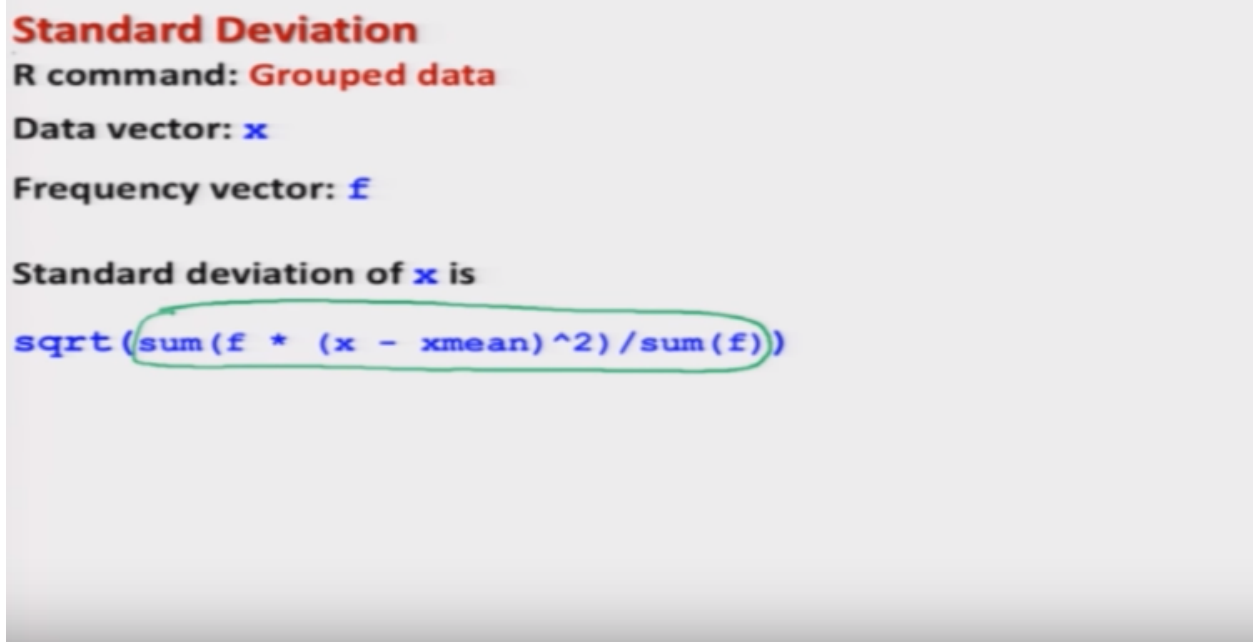
`sqrt(((n - 1)/n)*var(x))`

where  $n = \text{length}(x)$

And similarly if, you want to compute the standard deviation that is very simple simply try to find out the square root of the variance that you have obtained earlier so, if I try to find out that square root which is a function `sqrt` of the variance of  $(x)$  that we had earlier obtained then this is going to give you the standard

deviation or the standard error all right? but in this case you notice that the divisor is going to be n minus 1 in case if you want the divisor to be n then in that case simply try to take the square root of the earlier expression that we had obtained Right? So, finding out the standard deviation is simply equivalent to finding out the square root of the variance and considering its positive value.

Refer Slide Time: (39:08)



**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))
```

And, similarly in case of grouped data also whatever the expression you have obtained before the computing the variance just try to take the square root of the variance and this, is how you can compute the standard deviation in the case of group data.

Refer Slide Time: (39:23)

## 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
```

Now I, try to take the same example or the same data set that I, considered earlier where we have the data of 20 participants who participated in a race and their time taken to complete the race was recorded and this data is stored here in this data vector. Now I, try to find out the variance of this data vector and I use the command var and inside the argument the name of the data vector that is time and I get here this value 28 3 point 3684 and here you can see the divisor here is n minus 1 that you always have to keep in mind and I, will show you that how to find out the variance with divisor n in the next slide. Similarly, if you want to find out the standard deviation then you simply have to take the square root of the variance so whatever variants I, have obtained here I'm trying to operate the function sqrt, sqrt is a function to find out the square root and i get here the value 16.83355.

Refer Slide Time: (40:30)

## Variance and Standard Deviation

Example: Ungrouped data

$\frac{n-1}{n} \text{var}(n)$

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

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

Similarly, if you want to have the variance or standard deviation with the divisor n then in this case, we have learned that we need to multiply the variance (x) with a factor n minus 1 upon n where n is the length of the data vector. So, I simply try to multiply this n minus 1 upon n in the given various that we have obtained earlier and I will get here the value of variance with divisor n and similarly if I try to take the square root of this value then I will get the standard deviation which is based on the divisor n.

Refer Slide Time: (41:11)

## 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] 283.3694
>
> 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
> |
```

So, you can see that it is not difficult to operate or get the variance or standard deviation on the given set of data, and here is the the screenshot which I will try to show you. Now, next we try to find out the variance in the case of group data so, we consider the same example and we try to convert it into a group data we try to find out the frequency table and from there we will try to find out the frequency vector, vector of midpoints and the arithmetic mean. We, already had discuss this aspect that how to find out the frequency vector, vector of midpoints and how to create the frequency table so, I will not discuss here but I, will very briefly give you the background so, that you can look into the earlier lectures how to get it done.

Refer Slide Time: (41:54)

**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
	<b>Total</b>	<b>20</b>

~~We need to find the frequency vector and mean.~~

So, you see the, the given data has been classified in 6 class intervals and these are the midpoints, and these are the absolute frequencies that we already have obtained, and now we need to find out the frequency vector and mean from the given data.

Refer Slide Time: (42:08)

## 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
```

So, we had used the command breaks and then cut command and whatever the outcome of this cut command we, had created a frequency table using the table command and after that we had operated the as.numeric on the data obtained from the earlier command which has given us the frequency like this one, and we had created the vector of the midpoint from the given output from here like this Right? So now, we have obtained the x and f vector.

Refer Slide Time: (42:41)

## Variance and Standard Deviation

### Example: Grouped data

Data vector: x

Frequency vector: f

Mean of x is

$$\text{xmean} = \frac{\text{sum}(f * x)}{\text{sum}(f)}$$

$$\frac{1}{n} \sum_{i=1}^k f_i \cdot x_i$$

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



So, now I can say here that in case if, you want to find out here we already have obtained x we have already obtained f now I, need to find out the mean of x you can see here that in this case mean of s, is going to be defined by summation  $\sum_{i=1}^k x_i f_i$  goes from 1 to k Right? So you can see here, I'm trying to define here at x mean and this is sum of  $x_i f_i$  divided by n which is sum of f and if I, try to do it here I get here the value of here x mean to be 56.

Refer Slide Time: (43:12)

```
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
```

Now I, try to use the command or syntax that we, had defined earlier using with a given set of data here and this gives me the value here 269 and if I, try to find out the standard deviation of this quantity this is here this, will give me the value of the standard deviation in this case.

Refer Slide Time: (43:35)

## Variance and Standard Deviation

### Example: Grouped data

```
R Console
> x
[1] 35 45 55 65 75 85
> f
[1] 5 3 3 5 2 2
> sum(f * (x - xmean)^2) / sum(f)
[1] 269
> sqrt(sum(f * (x - xmean)^2) / sum(f))
[1] 16.40122
> |
```

And, this is the screenshot whatever we have obtained here,

Refer Slide Time: (43:40)

## 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
```

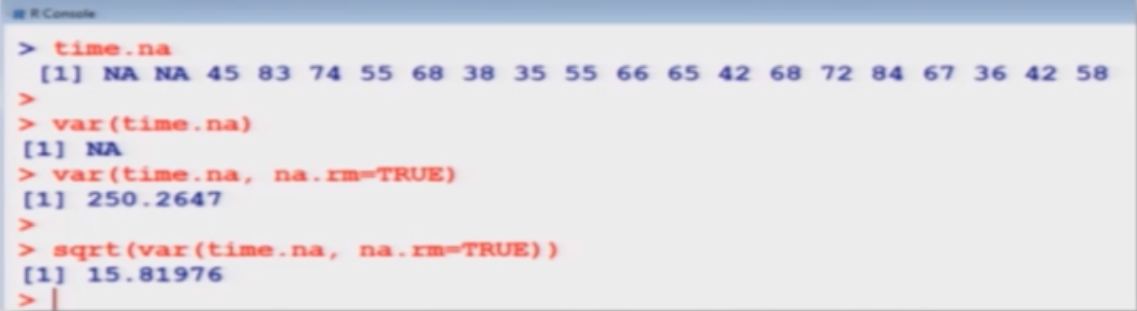
and similarly at the same time I, would try to show you that if you have some missing data so, then how to handle the situation how, to find out the variance and standard deviation so, in this earlier example I am trying to take that first two values to be missing and I am denoting the data inside new vector time dot na where two values are missing and if I, want to find out the variance it is simply the variance of time dot

any with the command `na.rm = TRUE` is equal to `TRUE` and this will give me the value of the variance as 250.2647 this will have the divisor and minus one and in case if, you want to convert it into the division  $n$  then you know how to get it done. And if I, try to take the square root of this value this will give me the standard deviation when, we have the missing values are present in the data Right?

Refer Slide Time: (44:32)

## Variance and Standard Deviation

### Example: Handling missing values

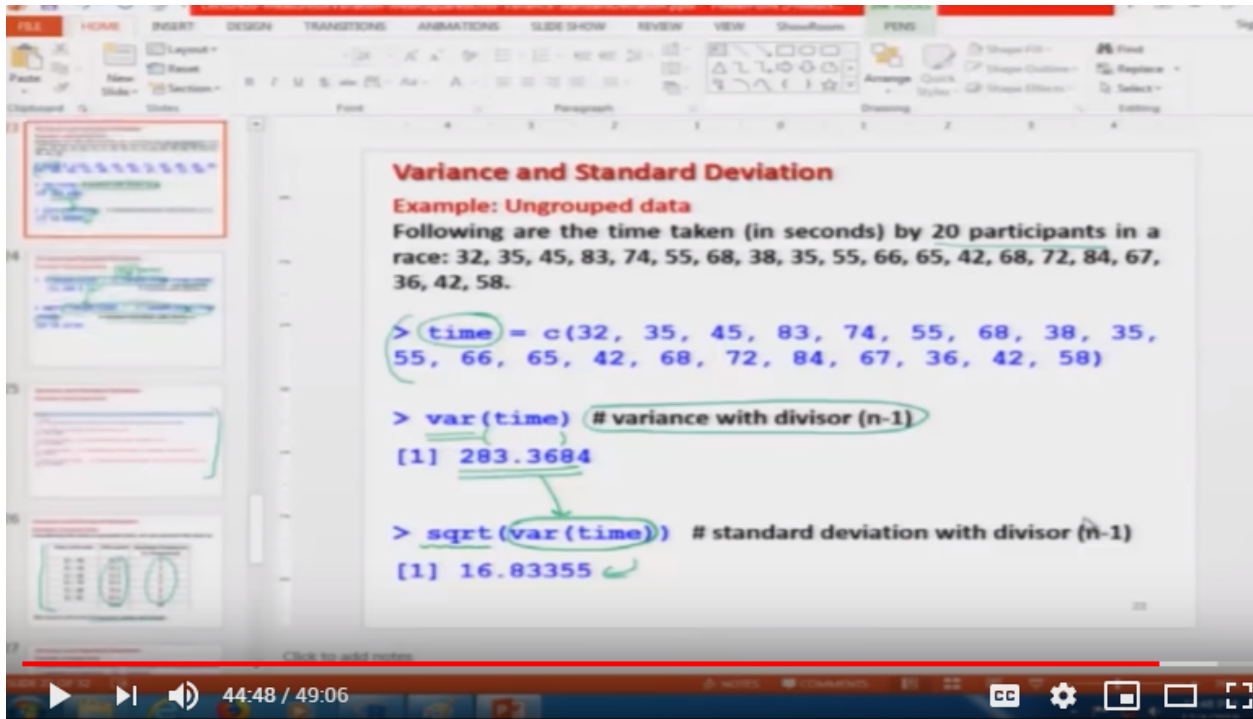


```
R Console
> time.na
[1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58
>
> var(time.na)
[1] NA
> var(time.na, na.rm=TRUE)
[1] 250.2647
>
> sqrt(var(time.na, na.rm=TRUE))
[1] 15.81976
> |
```

The screenshot shows an R console window with the following text: `> time.na` followed by a vector of 20 elements: `[1] NA NA 45 83 74 55 68 38 35 55 66 65 42 68 72 84 67 36 42 58`. Below this, the user enters `> var(time.na)` and receives `[1] NA`. Then, the user enters `> var(time.na, na.rm=TRUE)` and receives `[1] 250.2647`. Finally, the user enters `> sqrt(var(time.na, na.rm=TRUE))` and receives `[1] 15.81976`. A green bracket on the right side of the console window highlights the last two lines of output.

And, this is here the screenshot you can see here now I, will try to come on the our console and I, will try to show you that how you are going to obtain these values Okay?

Video start time (44:48)

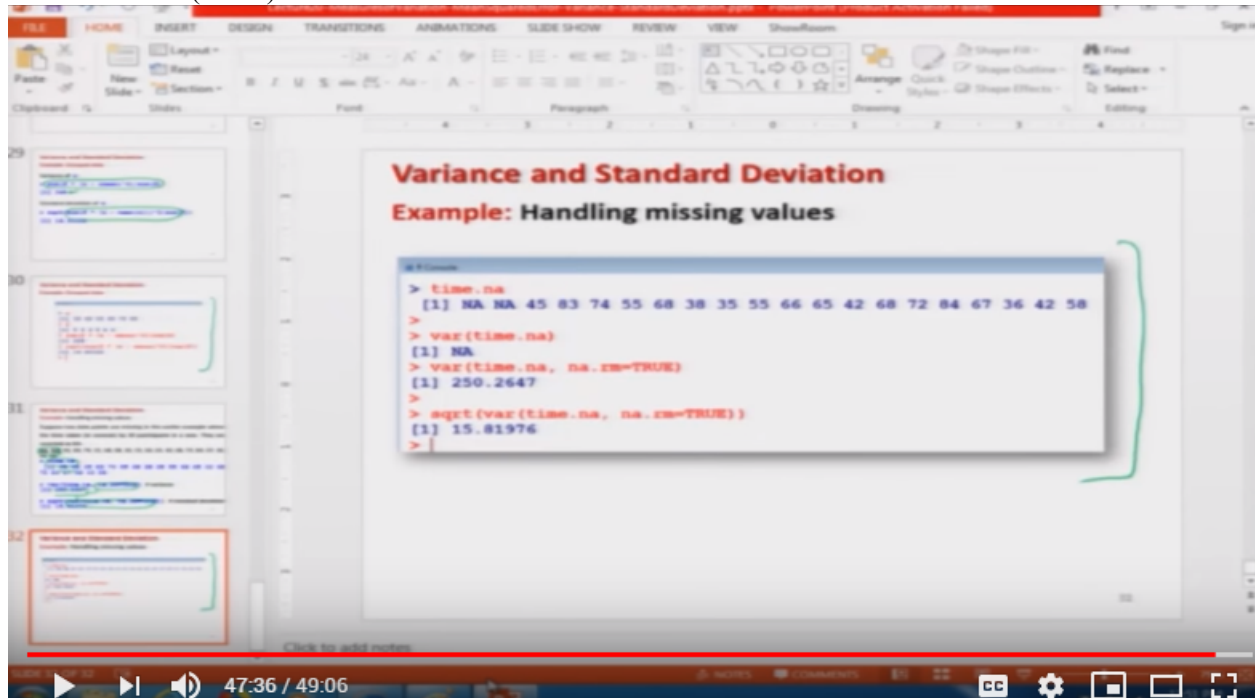


So I, try to first take it here the time so, you can see here I am copying the data on time I, will put it in the our console and I, try to find out here the variance of here (time) and you see I, get here this value if I, want to find out the square the standard deviation I, simply have to find out its square root and you can see here this is the value and similarly, these values have been obtained with the divisor say n minus 1 if, you want to have the divisor n in the variance then you just use the same command and you get here the outcome and if you want to find out here the standard deviation like the divisor n then you simply have to find out this square root of the earlier expression and this gives you the value sixteen point four zero seven three two and this is here the same screenshot which I have shown you here Right? and similarly if, you want to find it out in the case of of here this group data then first I, need to create here the data vector x and f which I, already have done.

So, you can see here I, can clear the slides x is here my midpoints f is here like this and if I, try to find out the mean in case of group data this is coming out to be 56 and now if I, try to find out the variance in this case this is going to be by the same command that we discuss here like this, this is 269, and if you want to find out the the standard deviation you simply have to find out its square root this has come out coming out to be like this and similarly if, you want to see that how the things are happening in case of a single use in the data so, I simply need to use this command here and you can see here that this data time dot na I already have copied here, and if you try to find out the variance when this na dot rm is equal to TRUE this give me this value, and if you want to find out the standard deviation in this case, then this is simply

the square root of this value that we have just opted and it is coming out to be here like this, and if you try to see here this is here the same screenshot that I have shown you here.

Video end time (47:36)



So now I, would like to stop in this lecture and I, would request you that we already have discuss the concept of variance which is one of the very popular tool to find out the variability in the data please try to understand it, please try to grasp it, and try to see that how this measure has been developed and please try to take some datasets and try to compute the variance with divisor n and with divisor n minus 1 try to compute the mean squared errors around any arbitrary value and get comfortable in computing the variance or a standard deviation or standard error using the R software. So, you practice, and we will see you in the next lecture. Till then, Good bye.