

**Lecture – 10**  
**Measurement of study variables**

Welcome to the session of Health Research Fundamentals. We often **hear** statements like I have data on 100,000 Leprosy patients, I have data on climatic change in Chennai city, I have data on road traffic accidents and so on. In all these statements, there is a word “data”. In this session, we are going to rather see what this data means, what are the different types of data? And how to convert these data into pieces of information?

(Refer Slide Time: 00:49)

**Types of Data**

- Qualitative
  - Nominal
    - Eg. Color of Eyes
  - Ordinal
    - Eg. Stages of disease condition
- Quantitative
  - Discrete
    - Eg. Family size
  - Continuous
    - Eg. Height / Weight

Nicer 101  
HEALTH RESEARCH FUNDAMENTALS  
National Institute of Epidemiology  
Chennai  
nik.gov.in

Data can broadly be classified into Qualitative and Quantitative data. Qualitative data as the name **suggests**, we cannot quantify them, it is on some sort of a quality. Again, this qualitative data could be a Nominal data or an Ordinary data. The nominal data, the examples are the color of eyes, the different regions of a city and so on and the ordinary data are the data which can be arranged in a sort of an order like examples you know stages of disease condition. Quantitative data again are of two categories, one is a discrete data which essentially is a full number, a number of siblings, family size,

etcetera and the other one is a continuous data, where is a continuous measurement like height and weight. So, these are all different types of data which requires different type of an analytical skill.

(Refer Slide Time: 02:03)

**Describe - Central Value**

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode

NHSR 101  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
National Institute of Epidemiology  
nie.gov.in

Now, our aim is to get some information out of data. A large set of data, it is very essential but still looking at just the data, you can't rather get any information. So, we need to summarize them. One of the ways of summarizing the data is to get a value of, a sort of an average. Now, average you mean, the first average that comes to our mind is the Mean.

(Refer Slide Time: 02:37)

## Arithmetic Mean (AM)

- Most commonly used; Simply called MEAN
- Add all the observed values ( $\text{Sum} = \sum X_i$ )
- Mean = Sum / n
- Sample Mean is denoted by  $\bar{x}$
- Population Mean is denoted by  $\mu$

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

A mean which is also called an Arithmetic Mean, this is a most commonly used and simply it is called Mean. You add all the observed values, we call that as sum which is  $\sum X_i$  in a mathematical notation and mean is nothing but divide this sum by the number of observations you have used in your calculations, which is n. The sample mean is denoted by an  $\bar{x}$ , line on top of x and population mean is denoted by  $\mu$ .

(Refer Slide Time: 03:13)

## Example

- Age of 10 Pregnant women  
26, 31, 25, 21, 26, 26, 27, 25, 27, and 26

Sum = (26+31+25+21+26+26+27+25+27+26) = 260

n = 10

Mean = sum / n = 260/10 = 26 years

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Let us rather see an example. Suppose, there are 10 pregnant patients who had visited an ANC clinic and their ages are 26, 31, 25 and so on and what is a mean age of these pregnant woman? The mean is got by summing up all the ages which comes to 260. There are 10 observations, so divided it by 10. It is 260 by 10 which is equal to 26. We say the mean age of pregnant woman who visited the ANC clinic is 26 years.

(Refer Slide Time: 03:50)

**Describe - Central Value**

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode

National Institute of Epidemiology  
Chennai  
NICER 101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Now, one of the problem with this average mean is, some extreme values either big or small even one or two, if they are present in your data set, that could influence on their average because you are adding all and one big value, if we add the whole mean becomes an over estimation. In order to control this or in order to avoid this, we have another measure which is called Median.

(Refer Slide Time: 04:22)

## Median

**The Median** describes literally the middle value of the distribution

Divides the distribution **exactly into two halves**  
(i.e. 50% of the data will fall on either side)

Useful when there are extreme values

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

The median is, literally the middle value of the distribution. It divides the distribution exactly into two halves that is 50 percent of the data will fall on either side. This is a very useful measure, especially when you have extreme values.

(Refer Slide Time: 04:42)

## Example

**Duration (days) of hospital stay of 11 patients**  
1, 2, 3, 4, 5, 6, 7, 8, 8, 9, 77 (Arranged in ascending order)

**Median is the middle value (6<sup>th</sup> value) = 6**  
(Mean = 11.8)

**If n is even; then take average of middle two values.**

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Let us just see this example. Suppose, you have a data on the duration of stay in hospital

of 11 patients, the duration is 1 day, 2 days, 3 days, and 9 days for 10 patients and then the 11th patient, it is 77 days. Of course, I have arranged this data in an ascending order. The median is the middle value, which is a 6th value, the value you get:  $n + 1 / 2$ ,

$11 + 1 = 12 / 2 = 6$ . So, the 6th value is the value 6, which means the median is 6 here. Whereas, when we really compute the mean for this it comes out to be 11.8. As you could rather see 6 is more appropriate measure of average in this case rather than the mean 11.8. If n is even, then you take the average of middle 2 values.

(Refer Slide Time: 05:41)

Describe - Central Value

- Data is not information.
- Summarize
  - Average
    - Mean
    - Median
    - Mode

National Institute of Epidemiology  
Chennai  
NIECR-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Now, there is another measure which is called Mode.

(Refer Slide Time: 05:48)

## Mode

**The Mode** is the value that occurs most frequently

Mode is the only location statistics to be used – for **nominal data** - not measurable characteristic

Epidemiology – Describing an epidemic with respect to **TIME**

NISCR 101  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
National Institute of Epidemiology  
nie.gov.in

Mode is the value that occurs most frequently. In fact, mode is the only location statistics which we can use for nominal data, which are not measurable. In epidemiology, we do use more quite often. In an epidemic curve with respective time, we look for the modal class and then that gives an idea of the incubation period of the pathogen.

(Refer Slide Time: 06:17)

## Example

- Colour preference of people for their car

<u>Colour preference</u>	<u>No. of persons</u>
Green	354
Yellow	852
White	310
Red	474

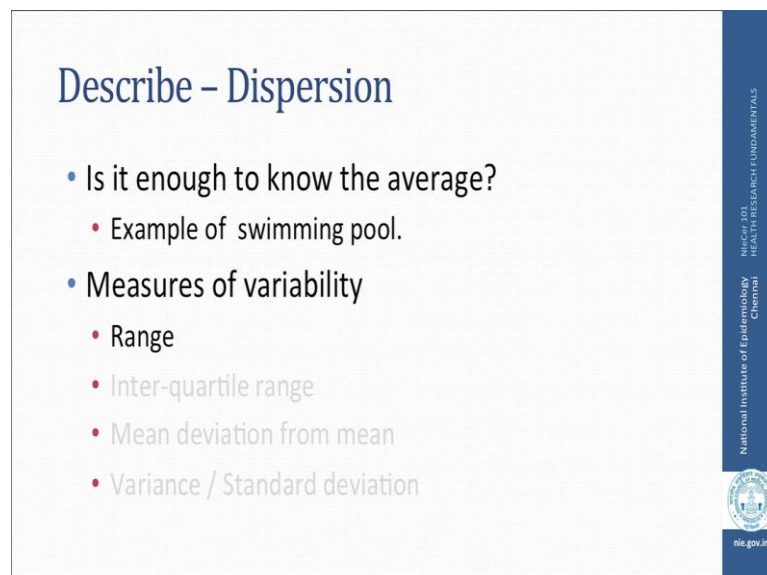
**Mode = Yellow**

NISCR 101  
HEALTH RESEARCH FUNDAMENTALS  
Chennai  
National Institute of Epidemiology  
nie.gov.in

The example for a mode is the color preference and the number of persons, 354 people they prefer green, 852 prefer yellow, 310 prefer white and 474 prefer red. So maximum number of people they prefer yellow, answer the mode of class is yellow. As you could rather see as in the respect of rather mode, there can be multiple modes, there cannot be a mode at all in a sequel. Suppose, if all the values are 354 here, then there is no mode. So, mode can exist, there can be multiple modes in a data set.

Now, we have seen Mean, Median, Mode are 3 good measures of summarizing your data to get an average value.

(Refer Slide Time: 07:08)



The slide is titled "Describe - Dispersion" and contains the following content:

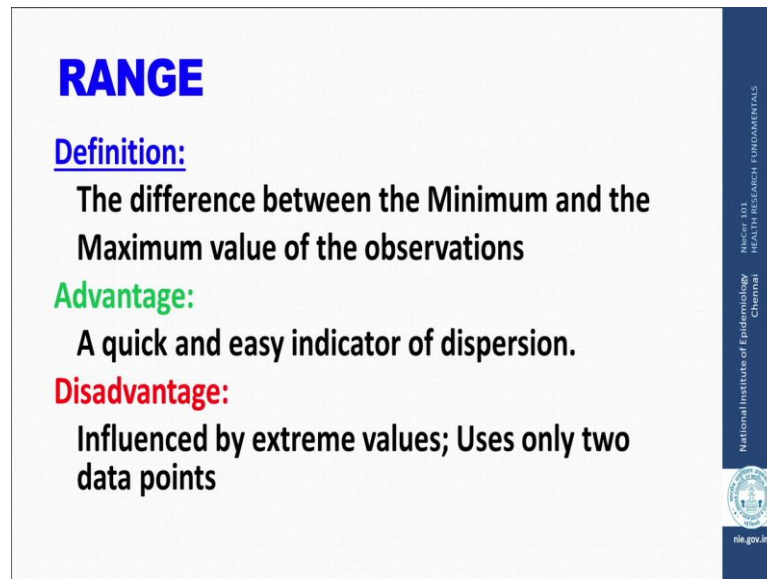
- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

On the right side of the slide, there is a vertical blue bar with the text "NICER 301 HEALTH RESEARCH FUNDAMENTALS" and "National Institute of Epidemiology Chennai". At the bottom right, there is a logo of the National Institute of Epidemiology and the website "nie.gov.in".

So it is not enough you just rather know the average value. Say for example, you go to a swimming pool and you do not know swimming and you are 5 feet 7 of inches and then, if the pool managers says the average depth of the swimming pool is 4 and a half feet you feel very comfortable and you jump, and suppose a place where you jump is 9 feet then you know the thing that you missed to ask is, yes the average is 4 and a half feet, but what is the variability? There maybe you know place where it is as shallow as 3 feet and as depth as 9 or 10 feet. So, you need to rather ask, what is the variability? One of the measures that comes to our mind is the Range.



(Refer Slide Time: 07:57)




**RANGE**

**Definition:**  
The difference between the Minimum and the Maximum value of the observations

**Advantage:**  
A quick and easy indicator of dispersion.

**Disadvantage:**  
Influenced by extreme values; Uses only two data points

National Institute of Epidemiology  
Chennai  
NICR-101  
HEALTH RESEARCH FUNDAMENTALS



nig.gov.in

The range is the difference between the minimum and then the maximum value of the observations. An advantage of this measure is it is very quick and easy indicator of dispersion. But, as I had said about the mean, the range also is influenced by extreme values and also we consider only two values, the first and then the last and in between we are not using the data at all and that is the great disadvantage of range. There is another value which is called Inter-quartile range.

(Refer Slide Time: 08:30)

## Describe – Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

National Institute of Epidemiology  
Chennai  
NIHER 101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

(Refer Slide Time: 08:34)

## INTER-QUARTILE RANGE

### Quartile Deviation

**Definition:**  
Defined as the interval between the value of the upper quartile (Q3) and the lower quartile (Q1)  
Inter Quartile Range =  $Q_3 - Q_1$

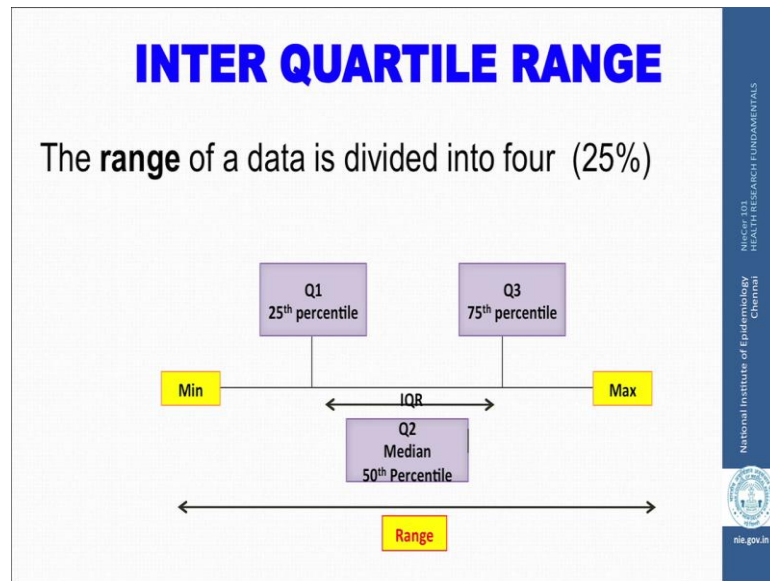
**Advantage:**  
Unaffected by the extreme values

**Disadvantage:**  
Covers only the middle 50% observations

National Institute of Epidemiology  
Chennai  
NIHER 101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

This to a large extent take care of this extreme values in the sense, we divide the data sets into 4 quarters.

(Refer Slide Time: 08:38)



And we try to remove the first quarter and then the last quarter and consider only the middle 50 percent of values and this inter-quartile range is the  $Q3 - Q1$  and a great advantage of this is, this value does not rather get affected by extreme values. But again the disadvantages is, it covers only the middle 50 percent of the value and then the same disadvantage that we had for range that uses only two values and in between values are not made use of and that is a great disadvantage of this value.

(Refer Slide Time: 09:15)

## Describe – Dispersion

- Is it enough to know the average?
  - Example of swimming pool.
- Measures of variability
  - Range
  - Inter-quartile range
  - Mean deviation from mean
  - Variance / Standard deviation

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

Another measure of variability is Mean deviation from mean. What you mean by that? Say for example, from your data set, every data point we try to subtract a mean and then we try to take average of these mean deviation, which is called Mean deviation from mean.

(Refer Slide Time: 09:40)

## MEAN DEVIATION

**Definition:** The mean deviation is the average of the absolute (ignoring the sign) deviations of the observations from the arithmetic mean.

**Advantage:** It is based on all the observations in the group. It is easy to grasp the meaning of the procedure.

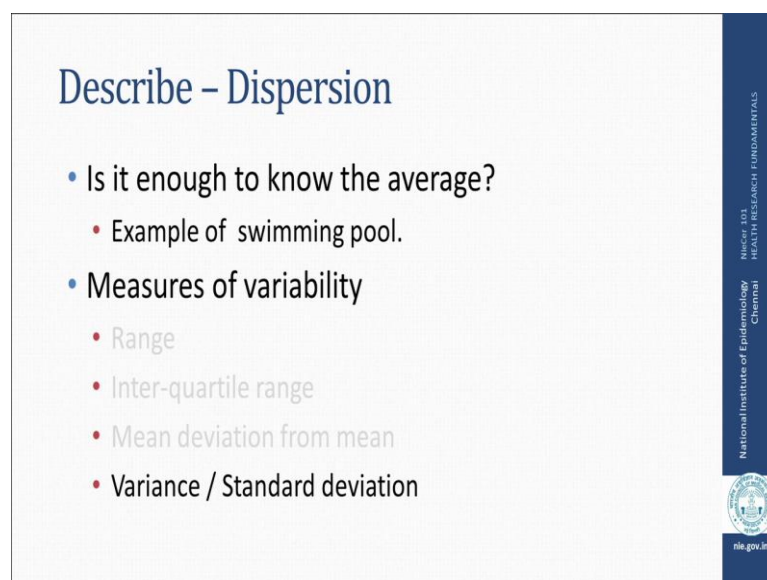
**Disadvantage:** It ignores the sign of the difference of the value of the observation and arithmetic mean.

It is not widely used because of the availability of a more advantageous measure.

National Institute of Epidemiology  
Chennai  
NICER-101  
HEALTH RESEARCH FUNDAMENTALS  
nie.gov.in

One of the problems with this is, if you rather do with that what happen is there are some values which are less than the mean, some values which are more than the mean and if you do the summation of all these, you get a value 0. So, mean deviation from mean is always 0. In order to get over that, what we do it is, we ignore the sign and then we just take the difference and then we take the average. This is called Absolute mean deviation, an advantage it is based on all observations in the group it is easy to grasp the meaning of the whole procedure. But the disadvantage is, it ignores the signs of the difference of the value and it is mathematically it is not very rigorous to use this value.

(Refer Slide Time: 10:37)



The slide is titled "Describe - Dispersion" and contains a bulleted list of points. The first point is "Is it enough to know the average?" with a sub-bullet "Example of swimming pool." The second point is "Measures of variability" with sub-bullets: "Range", "Inter-quartile range", "Mean deviation from mean", and "Variance / Standard deviation". The slide has a blue vertical bar on the right side with the text "National Institute of Epidemiology Chennai" and "NICEER 101 HEALTH RESEARCH FUNDAMENTALS". There is also a logo and the website "nie.gov.in" at the bottom right.

- Describe - Dispersion
  - Is it enough to know the average?
    - Example of swimming pool.
  - Measures of variability
    - Range
    - Inter-quartile range
    - Mean deviation from mean
    - Variance / Standard deviation

So, in order to get over that we have another measure, what we do is? We do take the difference of each observations from mean and instead of ignoring the sign, we square them because square takes care of even the - and then the + everything becomes + and then we take an average of that, that value is called Variance.

(Refer Slide Time: 10:59)

## STANDARD DEVIATION - SD ( $\sigma$ )


**Definition:** The SD is the square root of the average of the squared deviations of the observations from the arithmetic mean

The square of the SD is called variance

**Advantage:** The SD is the most important measure of distribution. While the variance is in unit squared, the SD is expressed in the same units of measurement as the observation. It is suitable for further analysis

The SD together with arithmetic mean is useful for description of the data

National Institute of Epidemiology  
Chennai  
NIC-101  
HEALTH RESEARCH FUNDAMENTALS



nie.gov.in

And since variance is, we are squaring and then the measurement also squares we take a square root at the end and that is called Standard Deviation. The standard deviation, which is denoted as SD is a square root of the average of the squared deviations of the observations from the arithmetic mean. The square of the standard deviation is the variance. Advantage of standard deviation is most important measure of distribution, while the variance is in unit square, the standard deviation is expressed in the same units of the measurement and it is suitable for further analysis. So, standard deviation together with arithmetic mean is useful for describing the data and these two measures are extensively used for further treatment of your data set.

(Refer Slide Time: 11:51)

## Coefficient of Variation (CV)

Purpose: To compare the relative variability in different groups

Definition: The coefficient of variation is the SD expressed as a percentage of the arithmetic mean (AM).

$$CV = \left( \frac{SD}{AM} \right) \times 100$$



I am going to introduce to you one more measure which is called Coefficient of Variation. The purpose of this measure, suppose if you have a different groups, different data sets to compare and then you want to rather compare the relative variability in different groups. So, the coefficient of the variation is the standard deviation expressed as a percentage of arithmetic mean because the standard deviation by arithmetic mean, what happens is, they both are the same units of the measurements, so units of measurement get canceled, so what you get is a pure number and that number expressed in terms of percentage that is multiplied by 100, you get coefficient of variation.

(Refer Slide Time: 12:38)

## Summary

- Choose appropriate central / dispersion value
  - Mean / SD – if no extreme values
  - Median / IQR – if there are extreme values
  - Mode / Range – for qualitative variables/ time distribution in epidemic curve
- Mean and SD are used the most.

In summary, we have to choose an appropriate central or dispersion values. The Mean and Standard Deviation are the most appropriate central and dispersion values especially, if there are no extreme values. If there are extreme values, there are methods of still using mean and standard deviation using some transformations of your data that requires (Refer Time: 13:06) expert handling of your data. Otherwise, you go in for median and inter-quartile range and these two measures median, inter-quartile range do take care of extreme values. The mode and range is normally used for qualitative variables, time distributions in epidemic curve. The mean and standard deviation, as I said are the most used measures of variability and the summary statistics.

Thank you.