INDIAN INSTITUTE OF TECHNOLOGY MADRAS

सिद्धिर्भवति कर्मजा

IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1

Prof. Usha Mohan

Department of Management Studies

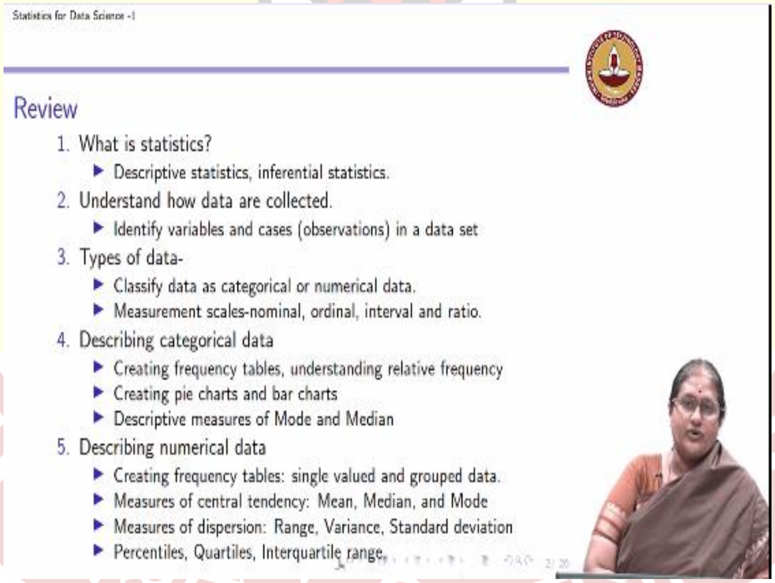
Indian Institute of Technology, Madras

Lecture No. 4.1

Association between Two Variables - Review of Course

Welcome. This is the week 4 of your online statistics for data science 1 course. In this week, we will understand about association between two variables. So, what are the key things you are going to learn here?

(Refer Slide Time: 00:29)



Statistics for Data Science -1

Review

1. What is statistics?
   - Descriptive statistics, inferential statistics.
2. Understand how data are collected.
   - Identify variables and cases (observations) in a data set
3. Types of data-
   - Classify data as categorical or numerical data.
   - Measurement scales-nominal, ordinal, interval and ratio.
4. Describing categorical data
   - Creating frequency tables, understanding relative frequency
   - Creating pie charts and bar charts
   - Descriptive measures of Mode and Median
5. Describing numerical data
   - Creating frequency tables: single valued and grouped data.
   - Measures of central tendency: Mean, Median, and Mode
   - Measures of dispersion: Range, Variance, Standard deviation
   - Percentiles, Quartiles, Interquartile range.

Before we understand about what we can expect from this week we will just take a quick look at where we stand now. So, we started with understanding what is statistics. Here we actually said that there are two main branches of statistics namely the descriptive statistics and the inferential statistics part and where our course actually focuses on is in the descriptive statistics and lays the foundation for inferential statistics by introducing probability.

Then we went on to know how data is collected and how data is tabulated and presented. At this point you should know that my columns the way I have constructed a dataset, my columns represent the variables and my rows represent the cases. This is what we should be aware of. Then we went on to classify data mainly when you look at data we can classify it as categorical and numerical. With a numerical you have discrete and continuous.

So, then we saw examples of what we mean by categorical data and what we mean by numerical data. Then we spend some time to understand what were the scales of measurement of variables. Here we introduced 4 prominent scales namely the nominal, ordinal, interval, and ratio scales. The nominal and ordinal scales are used for categorical data. Whereas the interval and ratio scales are used for numerical data.

And during our discussion about scales of measurement we also said what are the arithmetic operations possible for each scale of measurement. Then we went on to describe categorical data. Again we focused on describing categorical data where there is one variable. We started by introducing the notion of frequency table and we introduced the concept of what is relative frequency.

When it came to the graphical measures we talked about creating pie charts and bar charts. Pie charts basically are used when you want to tell the story about what is the share of a particular category in the overall picture whereas bar charts are useful when you want to represent counts. Then we went on to look at descriptive measures of mode and median. While mode can be applied to nominal data also. When you want to talk about a median you want a data to be ordinal or there should be an order in your data.

We then went on to describe numerical data. When we describe numerical we again started by how to create frequency tables. Here we borrowed from our categorical table and we again we talked about a single valued and group data how you create frequency table. Then we introduced the measures of central tendency. These were the numerical measures we introduced a new measure called mean.

When we talk about mean, we want a data to be only interval or ratio. I cannot talk about a mean for a categorical data. We also introduced the notion of median and mode for my numerical data. One important thing which we introduced when we talked about numerical data was measures of dispersion or variation. Here, we started with the range we introduced a measure of variance and standard deviation.

Then we went on to introduce what was percentile and from here we talked about the interquartile range. When we looked at the graphical summaries we focused on two graphical summaries; the histogram and the stem leaf plot. So, at this point of time in the course you should be able to categorize your data as categorical data or numerical data.

You should be able to identify what is the scale of measurement associated with your variable. If you have a single variable you should be able to summarize that variable both using graphical methods and descriptive measures as an where it is applicable. This is what you should be knowing at this point of time.