Unit I : Introduction to Statistics

Analysis of Quantitative Data

Sumit Kr. Choubey

Table of contents

1. Measures of Central Tendency

2. Measures of Dispersion

3. Moments

4. Skewness

5. Kurtosis

Measures of Central Tendency

Measures of Central Tendency

Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.

Statistics is concerned with the use of data in the context of uncertainty and decision making in the face of uncertainty.

Statistics

Statistics deals with variability. You're different from everybody else. Today differs from both yesterday and tomorrow. In an experiment designed to detect whether psychotherapy improves self-esteem, self-esteem scores will differ among subjects in the experiment, whether or not psychotherapy improves self-esteem.

1. **Statistics** is the study of how to collect, organize, analyze, and interpret numerical information from data.

2. Statistics is both the science of uncertainty and the technology of extracting information from data.

Types of Statistics

Descriptive statistics consists of methods for organizing and summarizing information.

Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

Descriptive Statistics

Statistics exists because of the prevalence of variability in the real world. In its simplest form, known as **descriptive statistics**, statistics provides us with tools [*tables, graphs, averages, ranges, correlations*] for organizing and summarizing the inevitable variability in collections of actual observations or scores. Examples are:

1. A tabular listing, ranked from most to least, of the total number of romantic affairs during college reported anonymously by each member of your stat class

2. A graph showing the annual change in global temperature during the last 30 years

3. A report that describes the average difference in grade point average (GPA) between college students who regularly drink alcoholic beverages and those who don't

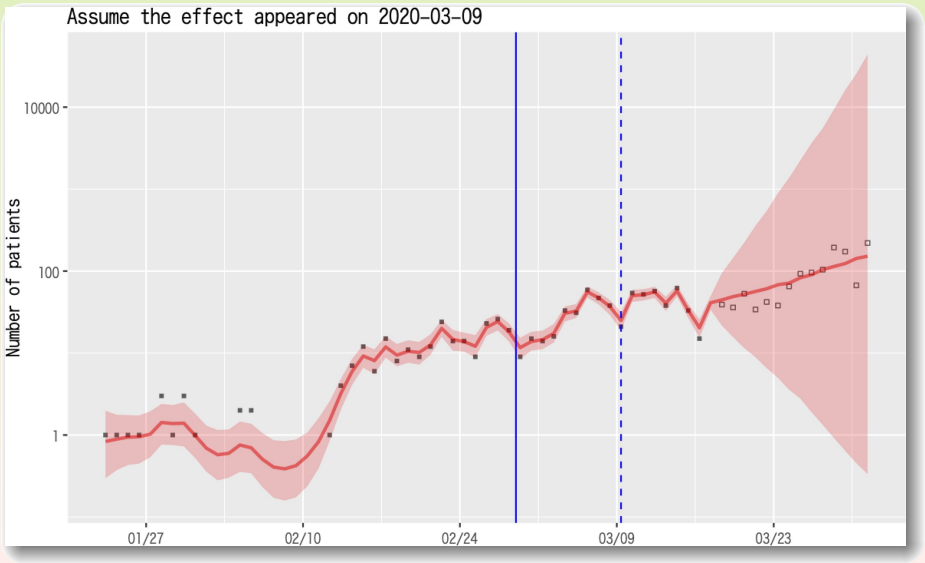
Descriptive Statistics

Inferential Statistics

Statistics also provides tools [*a variety of tests and estimates*] for generalizing beyond collections of actual observations. This more advanced area is known as inferential statistics. Tools from inferential statistics permit us to use a relatively small collection of actual observations to evaluate, for example:

1. A pollster's claim that a majority of all U.S. voters favor stronger gun control laws
2. A researcher's hypothesis that, on average, meditators report fewer headaches than do nonmeditators
3. An assertion about the relationship between job satisfaction and overall happiness

Inferential Statistics



Statistics

Question

Indicate whether each of the following statements typifies descrip- tive statistics (because it describes sets of actual observations) or inferential statistics (because it generalizes beyond sets of actual observations).

1. Students in my statistics class are, on average, 23 years old.
2. The population of the world exceeds 7 billion (that is, 7,000,000,000 or 1 million multiplied by 7000).
3. Either four or eight years have been the most frequent terms of office actually served by U.S. presidents.
4. Sixty-four percent of all college students favor citizenship amendment law.

Measures of Dispersion

Measures of Dispersion

Data are characteristics or information, usually numerical, that are collected through observation. In a more technical sense, data are a set of values of qualitative or quantitative variables about one or more persons or objects

Data

"facts and statistics collected together for reference or analysis."

Qualitative

approximates and characterizes.

Quantitative

association of unique numerical value

Moments

Moments I

What is?

Looming behind any data, the level of measurement specifies the extent to which a number (word,letter) actually represents some attribute and, therefore, has implications for the appropriateness of various arithmetic operations and statistical procedures.

1. The **nominal level** of measurement applies to data that consist of names, labels, or categories. There are no implied criteria by which the data can be ordered from smallest to largest.
2. The **ordinal level** of measurement applies to data that can be arranged in order. However, differences between data values either cannot be determined or are meaningless.
3. The **interval level** of measurement applies to data that can be arranged in order. In addition, differences between data values are meaningful.

Moments II

4. The **ratio level** of measurement applies to data that can be arranged in order. In addition, both differences between data values and ratios of data values are meaningful. Data at the ratio level have a true zero.

Question

Identify the type of data.

1. Taos, Acoma, Zuni, and Cochiti are the names of four Native American pueblos from the population of names of all Native American pueblos in Arizona and New Mexico.
2. In a high school graduating class of 319 students, Jim ranked 25th, June ranked 19th, Walter ranked 10th, and Julia ranked 4th, where 1 is the highest rank.
3. Body temperatures (in degrees Celsius) of trout in the Yellowstone River.
4. Length of trout swimming in the Yellowstone River.

Basic Terminology I

Individuals

Individuals are the people or objects included in the study.

Variable

A variable is a characteristic of the individual to be measured or observed. The variables in a study may be quantitative or qualitative in nature.

- A quantitative variable has a value or numerical measurement for which operations such as addition or averaging make sense.
- A qualitative variable describes an individual by placing the individual into a category or group, such as male or female.

Basic Terminology II

Population & Sample

In statistics, a population refers to any complete collection of observations or potential observations, whereas a sample refers to any smaller collection of actual observations drawn from a population.

Population & Sample Data

In population data, the data are from every individual of interest. In sample data, the data are from only some of the individuals of interest.

Population & Sample Data

In population data, the data are from every individual of interest. In sample data, the data are from only some of the individuals of interest.

Basic Terminology III

Population parameter	A population parameter is a numerical measure that describes an aspect of a population.
Sample statistic	A sample statistic is a numerical measure that describes an aspect of a sample.

Basic Terminology IV

Question

How important is music education in school (K–12)? The Harris Poll did an online survey of 2286 adults (aged 18 and older) within the United States. Among the many questions, the survey asked if the respondents agreed or disagreed with the state- ment, “Learning and habits from music education equip people to be better team players in their careers.” In the most recent survey, 71% of the study participants agreed with the statement.

1. Identify the individuals of the study and the variable
2. Do the data comprise a sample? If so, what is the underlying population?
3. Is the variable qualitative or quantitative?
4. Identify a quantitative variable that might be of interest.

Kurtosis

Kurtosis

Data in statistics is sometimes classified according to how many variables are in a particular study. For example, “height” might be one variable and “weight” might be another variable. Depending on the number of variables being looked at, the data might be univariate, or it might be bivariate.

“How Many variables?”

Univariate Analysis

What is?

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

Univariate Descriptive Statistics

Some ways you can describe patterns found in univariate data include central tendency (mean, mode and median) and dispersion: range, variance, maximum, minimum, quartiles (including the interquartile range), and standard deviation.
You have several options for describing data with univariate data.

- Frequency Distribution Tables.
- Bar Charts & Histograms.
- Frequency Polygons & Pie Charts.

Bivariate Analysis

What is?

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

Common types of bivariate analysis include:

- Scatter plots
- Regression Analysis
- Correlation Coefficients

Multivariate Analysis

What is?

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle. This type of analysis is almost always performed with software, as working with even the smallest of data sets can be overwhelming by hand.

There are more than 20 different ways to perform multivariate analysis. Which one you choose depends upon the type of data you have and what your goals are. For example, if you have a single data set you have several choices:

- Additive trees, multidimensional scaling, cluster analysis are appropriate for when the rows and columns in your data table represent the same units and the measure is either a similarity or a distance.
- Principal component analysis (PCA) decomposes a data table with correlated measures into a new set of uncorrelated measures.
- Correspondence analysis is similar to PCA. However, it applies to contingency tables.