Chapter 1
Introduction to Statistics and Analysis

Objectives

After completing this chapter, you will be able to:

- To perform simple statistical calculations
- Understand the limitations of formulas used in statistical analysis
- Know when more complex statistical methods are required
- Summary statistics for continuous and discrete data
- Understand the meaning of Statistics and Analysis
- Different types of data, distributions and structure within data

Statistical Analysis Defined

What is statistical analysis? It's the science of collecting, exploring and presenting large amounts of data to discover underlying patterns and trends. Statistics are applied every day – in research, industry and government – to become more scientific about decisions that need to be made. For example:

- Manufacturers use statistics to weave quality into beautiful fabrics, to bring lift to the airline industry and to help guitarists make beautiful music.
- Researchers keep children healthy by using statistics to analyze data from the production of viral vaccines, which ensures consistency and safety.
- Communication companies use statistics to optimize network resources, improve service and reduce customer churn by gaining greater insight into subscriber requirements.
- Government agencies around the world rely on statistics for a clear understanding of their countries, their businesses and their people.
- Look around you. From the tube of toothpaste in your bathroom to the planes flying overhead, you see hundreds of products and processes every day that have been improved through the use of statistics.

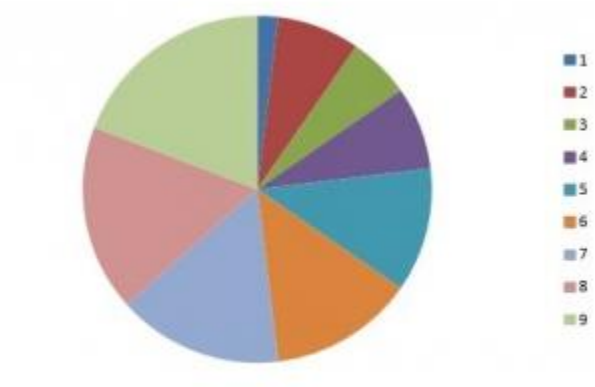
Statistical Analysis

Statistical analysis is the collection and interpretation of data, to uncover patterns and trends. It is a component of data analytics.
In the context of business intelligence (BI), statistical analysis involves collecting and scrutinizing every data sample in a set of items from which samples can be drawn. A sample, in statistics, is a representative selection drawn from a total population.

The goal of statistical analysis is to identify trends. A retail business, for example, might use statistical analysis to find patterns in unstructured and semi-structured customer data that can be used to create a more positive customer experience and increase sales.

Statistical analysis can be broken down into five discrete steps, as follows:

- Describe the nature of the data to be analyzed.
- Explore the relation of the data to the underlying population.
- Create a model to summarize understanding of how the data relates to the underlying population.
- Prove (or disprove) the validity of the model.
- Employ predictive analytics to run scenarios that will help guide future actions.



A pie chart is one way to display data.

Statistical analysis is the science of collecting data and uncovering patterns and trends. It's really just another way of saying "statistics." After collecting data you can *analyze* it to:

- **Summarize the data.** For example, make a pie chart.
- **Find key measures of location.** For example, the mean tells you what the average (or "middling") number is in a set of data.
- **Calculate measures of spread:** these tell you if your data is tightly clustered or more spread out. The standard deviation is one of the more commonly used measures of spread; it tells you how spread out your data is about the mean.
- **Make future predictions based on past behavior.** This is especially useful in retail, manufacturing, banking, sports or for any organization where knowing future trends would be a benefit.
- **Test an experiment's hypothesis.** Collecting data from an experiment only tells a story when you analyze the data. This part of statistical analysis is more formally called "Hypothesis Testing," where the null hypothesis (the commonly accepted theory) is either proved or disproved.

Statistical Analysis and the Scientific Method

Statistical analysis is used extensively in science, from physics to the social sciences. As well as testing hypotheses, statistics can provide an approximation for an unknown that is difficult or impossible to measure. For example, the field of quantum

field theory, while providing success in the theoretical side of things, has proved challenging for empirical experimentation and measurement. Some social science topics, like the study of consciousness or choice, are practically impossible to measure; statistical analysis can shed light on what would be the most likely or the least likely scenario.

When Statistics Lie

While statistics can sound like a solid base to draw conclusions and present "facts," be wary of the pitfalls of statistical analysis. They include deliberate and accidental manipulation of results. However, sometimes statistics are just plain wrong. A famous example of "plain wrong" statistics is Simpson's Paradox, which shows us that even the best statistics can be completely useless. In a classic case of Simpson's, averages from University of Berkeley admissions (correctly) showed their average admission rate was higher for women than men, when in fact it was the other way around.

Population vs Sample

The population includes all objects of interest whereas the sample is only a portion of the population. Parameters are associated with populations and statistics with samples. Parameters are usually denoted using Greek letters (mu, sigma) while statistics are usually denoted using Roman letters (x, s).
There are several reasons why we don't work with populations. They are usually large, and it is often impossible to get data for every object we're studying. Sampling does not usually occur without cost, and the more items surveyed, the larger the cost.
We compute statistics, and use them to estimate parameters. The computation is the first part of the statistics course (Descriptive Statistics) and the estimation is the second part (Inferential Statistics)

Discrete vs Continuous

Discrete variables are usually obtained by counting. There are a finite or countable number of choices available with discrete data. You can't have 2.63 people in the room.
Continuous variables are usually obtained by measuring. Length, weight, and time are all examples of continous variables. Since continuous variables are real numbers, we usually round them. This implies a boundary depending on the number of decimal places. For example: 64 is really anything $63.5 <= x < 64.5$. Likewise, if there are two decimal places, then 64.03 is really anything $63.025 <= x < 63.035$. Boundaries always have one more decimal place than the data and end in a 5.

Levels of Measurement

There are four levels of measurement: Nominal, Ordinal, Interval, and Ratio. These go from lowest level to highest level. Data is classified according to the highest level which it fits. Each additional level adds something the previous level didn't have.

- Nominal is the lowest level. Only names are meaningful here.
- Ordinal adds an order to the names.
- Interval adds meaningful differences
- Ratio adds a zero so that ratios are meaningful.

Types of Sampling

There are five types of sampling: Random, Systematic, Convenience, Cluster, and Stratified.

- Random sampling is analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occuring. While this is the preferred way of sampling, it is often difficult to do. It requires that a complete list of every element in the population be obtained. Computer generated lists are often used with random sampling. You can generate random numbers using the TI82 calculator.
- Systematic sampling is easier to do than random sampling. In systematic sampling, the list of elements is "counted off". That is, every *k*th element is taken. This is similar to lining everyone up and numbering off "1,2,3,4; 1,2,3,4; etc". When done numbering, all people numbered 4 would be used.
- Convenience sampling is very easy to do, but it's probably the worst technique to use. In convenience sampling, readily available data is used. That is, the first people the surveyor runs into.
- Cluster sampling is accomplished by dividing the population into groups -- usually geographically. These groups are called clusters or blocks. The clusters are randomly selected, and each element in the selected clusters are used.
- Stratified sampling also divides the population into groups called strata. However, this time it is by some characteristic, not geographically. For instance, the population might be separated into males and females. A sample is taken from each of these strata using either random, systematic, or convenience sampling.

Lecture Notes

Below are listed the terms that usually used in Statistics and Analysis and definition.

Definitions

Statistics

Collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions.

Variable

Characteristic or attribute that can assume different values

Random Variable

A variable whose values are determined by chance.

Population

All subjects possessing a common characteristic that is being studied.

Sample

A subgroup or subset of the population.

Parameter

Characteristic or measure obtained from a population.

Statistic (not to be confused with Statistics)

Characteristic or measure obtained from a sample.

Descriptive Statistics

Collection, organization, summarization, and presentation of data.

Inferential Statistics
: Generalizing from samples to populations using probabilities. Performing hypothesis testing, determining relationships between variables, and making predictions.

Qualitative Variables
: Variables which assume non-numerical values.

Quantitative Variables
: Variables which assume numerical values.

Discrete Variables
: Variables which assume a finite or countable number of possible values. Usually obtained by counting.

Continuous Variables
: Variables which assume an infinite number of possible values. Usually obtained by measurement.

Nominal Level
: Level of measurement which classifies data into mutually exclusive, all inclusive categories in which no order or ranking can be imposed on the data.

Ordinal Level
: Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist.

Interval Level
: Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.

Ratio Level
: Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.

Random Sampling
: Sampling in which the data is collected using chance methods or random numbers.

Systematic Sampling
: Sampling in which data is obtained by selecting every *k*th object.

Convenience Sampling
: Sampling in which data is which is readily available is used.

Stratified Sampling
: Sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.

Cluster Sampling
: Sampling in which the population is divided into groups (usually geographically). Some of these groups are randomly selected, and then all of the elements in those groups are selected.

Video Links:

Introduction to Statistics and Analysis

- https://www.youtube.com/watch?v=4M-cx8jV2aI
- https://www.youtube.com/watch?v=XbHeCL_8UhA

References

- https://www.statisticshowto.com/statistical-analysis/#:~:text=Statistical%20analysis%20is%20the%20science,example%2C%20make%20a%20pie%20chart.
- https://www.sas.com/en_us/insights/analytics/statistical-analysis.html
- https://whatis.techtarget.com/definition/statistical-analysis
- https://people.richland.edu/james/lecture/m170/ch01-not.html
- https://people.richland.edu/james/lecture/m170/ch01-def.html