The Impact and Importance of Statistics in Data Science

Pallavi Gupta
Department of Computer Science and Engineering,
CSMSS Chh Shahu College of Engineering
Aurangabad, Maharashtra, India

Nitin V. Tawar
Department of Computer Science and Engineering,
International Centre of Excellence in Engineering
and Management (ICEEM),
Aurangabad, Maharashtra, India

ABSTRACT

With the massive amount of data pouring in, the data science has become one of the most challenging yet promising field to deal with such tremendous quantity of data and bring out the quality information out for strategic business decisions. The way to data science begins with collection of huge amount of data which should be managed enough to start processing on it to analyze it. The statistics plays a vital role from molding data into the required format to final presentation of results to make it easy for the operations to be carried out on data almost in every step of data science.

In this paper, we give a manifestation of how important the statistics is to provide the necessary tools and methods to handle data to provide deep insights into the data and how useful statistics is for quantification and analysis of data. We will discuss various tools and techniques of statistics used in data science beginning from measures of dispersion to advanced tools for visualization of results to be able to understand the role and importance of statistical approaches in data processing and analysis.

General Terms

Data Science, Statistics, Algorithms, Hypotheses Testing

Keywords

Inferential Analysis, Mean, Median, Mode, Null hypotheses, p-value

1. INTRODUCTION

The data science is an advanced branch of science and engineering which combines the areas of mathematics, statistics, computer science, informatics, management and research. In 1996, for the first time, the term Data Science was included in the title of a statistical conference (International Federation of Classification Societies (IFCS) “Data Science, classification, and related methods”)[2]. The data science term was coined by statisticians but the branches of computer science and informatics are given more importance in this world of increasing data. In 1977, the International Association for Statistical Computing (IASC) was established whose objective was to combine traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge. [4] In 1989, Gregory Piatetsky-Shapiro organized and chaired the first Knowledge Discovery in Databases (KDD) workshop.

In 1995, it became the annual ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).[2] Statistics provides the tools and techniques to not only provide mathematical results but also provides the deeper insights into the unstructured complex data. As Josh Wills once said, “Data Scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”[8] As the need of statistics was realized in dealing with data and uncertainty, statistical learning was evolved. We understand the crucial role of statistics in the basic to advanced concepts of data science. This paper aims at stating the importance of statistics in data science.

2. BASIC STATISTICS AND TERMINOLOGY

Statistics is used to deal with collecting, pre-processing analyzing, interpreting and visualizing data. Any field of study which includes data will involve statistics in use to some or more extent. The data science is no exception. It deals with an enormous amount of data so statistics plays a key role in giving a proper form to the data before being fed to the algorithms before further analysis. It also helps in getting detailed insight into the data.

2.1 Terminology in Statistics

Population is the total set of data of a specified type or group under consideration.

A **Sample** is a subset of the Population taken to reduce the data quantity keeping the quality.

A **Variable** is any feature, characteristics, number, or quantity/ quality value that can be measured or counted.

A **statistical Parameter or population parameter** is a characteristic of the population or it is a quantity that is used to index a class of probability distributions. For example, the mean, median, etc of a population.

2.2 Types of Data

The data can be classified into two types:

1. Qualitative or Categorical Data

2. Quantitative or Numerical Data

In categorical data, there are two types. The first is nominal data which classifies data into distinct unordered classes or categories such as gender which can take two values i.e. M for male and F for female.[8]

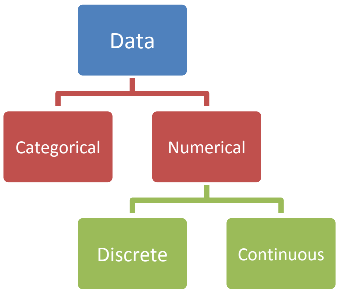

Fig. 1 Classification of Data

The second one under categorical data is ordinal data which takes ordered values based on certain ranking, for example, Product Satisfaction-Satisfied, Neutral, Unsatisfied.

The second type of data is quantitative data also called as numerical data. It may be discrete or continuous. A discrete variable can take values which can be counted for example number of pages in a book and a continuous variable can take values which are infinite for example temperature or height.

2.3 Types of Analysis

An analysis of data can be done in two ways:

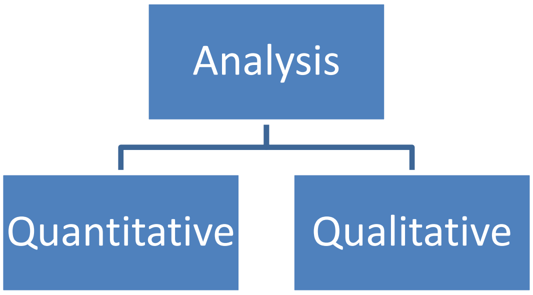

Fig. 2 Classification of Analysis

Quantitative Analysis: Quantitative Analysis or the Statistical Analysis is the method of collection and interpretation of data with numbers and graphs to identify patterns and trends in the data.

Qualitative Analysis: Qualitative or Non-Statistical Analysis gives general information and uses text, sound and other forms of media to do so.

For example, if I want a purchase a coffee from Cafe Coffee Day, it is available in different types. This is an example of Qualitative Analysis. But if a store sells 70 regular coffees a week, it is Quantitative Analysis because we have a number representing the coffees sold per week. [8]

2.4 Types of Statistics

There are two main categories in Statistics, namely:

1. Descriptive Statistics
2. Inferential Statistics

In descriptive analysis, when we represent data in the graphical form, using bar graphs, histogram, scatter plot, line graph, etc. The representation is based on central tendency. Mean, median or mode are the measure of central tendency of data and the variance and standard deviation are used as the measures of spread. [9]

Descriptive statistics give an insight into a specific group of data. For example, we could compute the mean and variance of the salaries for the 100 employees and this can give important information about this group of 100 employees. Any group of data under consideration is called a population. A population can be small or large.

However, many a times we do not have access to all the population under consideration. We can use only a part of that data known as sample. Properties of samples are not called parameters instead they are called as statistics.

Inferential statistics are techniques which use these samples to derive generalized conclusions about the populations under consideration from which the samples were drawn. It is necessary that the sample accurately and appropriately represents the population otherwise it introduces a sampling error and lead to improper results. The methods of inferential statistics are (1) the computation of parameter(s) and (2) statistical hypotheses testing.

3. DESCRIPTIVE ANALYSIS

Descriptive statistics are used to describe the basic properties of the data under study. They provide simple summaries about the sample data and the measures. Together with simple graphics analysis, they form the basis of every quantitative analysis of data and from which we can find the metrics of data which can help in future prediction. [11]

3.1 Measures of Central Tendency

The central tendency of a data distribution is a measure of the “center” of a distribution of values. There are three types of measures of central tendency:

i. Mean

ii. Median

iii. Mode

The mean is the most common measure of central tendency. The basic mean is computed by adding all the values in the distribution together and dividing the total by the number of values.

In case of ungrouped/grouped, data mean is computed by finding the sum of products of individual frequencies with corresponding value and dividing by the total frequency.

The Median is the value found at the exact middle i.e. at 50% of the set of values. One way to find the median is to arrange all the value in the ascending order and find the middle value. If there is an even number of values then we have to find the median by calculating the average of middle two values.

The Mode is the most frequently occurring value in the data distribution. To compute the mode, the data is arranged in ascending numerical order and then the values are counted. The value which occurs or appears the most number of times is called as Mode.

3.2 Measures of Dispersion

Dispersion gives a measure about how the data varies around the central tendency. It gives an insight into the outliers also.

3.2.1 Range

It is computed by finding the difference between the highest value and the lowest value.

3.2.2 Inter Quartile Range (IQR):

The quartiles are the values at particular percent portion of data like 25%, 50% and 75% of the sorted data distribution. We can know about the variability in data by dividing it into quartiles.

3.2.3 Variance:

It describes how much a random variable differs from its expected value. It is computed from squares of deviations. Deviation is the difference of each class value from the mean. Population Variance is the average of squared deviations. Sample Variance is the average of squared differences from the mean.

3.2.4 Standard Deviation:

The Standard Deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range. It is the measure of the dispersion of a set of data from its mean. [7]

4. INFERENTIAL ANALYSIS

Inferential statistics gives qualitative information about general pattern or behaviour of data. Inferential Statistics involves hypotheses testing to determine if there is enough potential in the data sample to conclude or infer that a certain condition holds true for an entire population.

4.1 Hypotheses Testing

Hypothesis testing was defined by Ronald Fisher, Jerzy Neyman, Karl Pearson and Pearson's son, Egon Pearson. Hypothesis testing is a statistical technique that is used in making statistical decisions using experimental data. Hypothesis Testing is based on an assumption that we make about the population parameter.

To understand the properties of a general population, we take a random sample from the population and analyse the properties of the sample. We find the conclusion and we check whether or not the identified conclusion represents the population accurately or not and finally we interpret their results. The acceptance of hypothesis depends upon the percentage value that we get from the hypothesis. [10]

4.2 Key Terms & Concepts

4.2.1 Null Hypotheses:

Null hypothesis is a statistical hypothesis that is based on an assumption that the two groups of data are same. The result is same as that of assumption. It is denoted by H0.

4.2.2 Alternative hypothesis:

It is opposite of null hypotheses. Here, result is completely contrary to the assumption. It is denoted by H1.

4.2.3 Level of significance:

It refers to the degree of significance for accepting or rejecting a null hypothesis.

4.2.4 Type I error:

This type of error occurs when we reject a null hypothesis even if it was correct. It is denoted by alpha.

4.2.5 Type II errors:

It occurs when we accept the null hypothesis even if it is false. Type II error is denoted by beta.

In hypothesis testing, we have to make decisions regarding hypothesis whether we should accept the hypothesis as result or whether we should reject the hypothesis. Every test generates the significance value for that particular test. If the significance value exceeds the threshold value then we accept the null hypothesis. If the significance value is less than the threshold value then we reject the null hypothesis.

For example, if we want to see the degree relation between two variables and the significance value of the correlation coefficient is greater than the threshold, then we can accept the null hypothesis and conclude that there was no relation between the two variables.

When you are testing a hypothesis, we need to check for both the size of sample and also its variability.

4.3 Steps in Hypotheses Testing

There are five Steps in Hypothesis Testing:

1) Specify the Null Hypothesis
2) Specify the Alternative Hypothesis
3) Set the Threshold Significance Level (a)
4) Calculate the Test Statistic and Corresponding Probability
5) Drawing a Conclusion

4.3.1 Specify the Null Hypothesis:

The null hypothesis (H0) is a statistical hypothesis which poses no difference between two or more groups of data. In research studies, a researcher is always interested in proving the null hypothesis wrong to show that there is always a relation between the two variables in the hypothesis. Examples:

i. If one plant is watered club soda for one month and another plant is watered with plain water, there will be no significant difference in growth between the two plants. A null hypothesis may be of the form: There is no statistical significance between the type of water I use to feed the flowers and growth of the flowers.

ii. There is no association between injury type of a person and whether or not the patient received an IV in the primary treatment.

4.3.2 Specify the Alternative Hypothesis:

The alternative hypothesis H1 is a statement which depicts some relationship between the two variables. There is a difference between them. The researcher wants to prove this generally. The alternative hypothesis can be either one-sided (only provides one direction) or two-sided. We often employ two-sided tests because one-sided tests require more proofs against the null hypothesis to accept the alternative hypothesis.

Examples:

i. If one plant is watered club soda for one month and another plant is fed with plain water, the plant that is watered with club soda will grow faster and better than the plant that is fed plain water.

ii. There is an association between injury type of the person and whether or not the patient received an IV in the primary treatment **(two sided)**.

4.3.3 Set the Significance Level(α):

The threshold significance level denoted by alpha *(α)* is generally set to the value 0.05. It means that you will that there is 5% probability that you will accept the alternative hypothesis even when your null hypothesis is actually true. The smaller the significance threshold value, the larger is the evidence needed in support of the alternative hypothesis.

4.3.4 Calculate the Test Statistic and Corresponding Probability:

Hypothesis testing is done using a test statistic that is responsible for comparing groups and examining associations between variables. A confidence interval is needed when we try to describe a single sample without establishing relationship among the variables of that sample. The p-value is the probability of obtaining a sample statistic as one or more extreme by chance alone if your null hypothesis is true. [8] Your conclusions about the acceptance or non-acceptance of hypothesis are based on your p-value and your threshold significance level. The p-value varies with varying sample size. If the sample size is large, it will have smaller p-value. Example:

i. P-value = 0.02 This will happen 2 in 100 times by pure chance if your null hypothesis is true. Not likely to happen strictly by chance.

ii. P-value = 0.80 This will happen 80 in 100 times by pure chance if your null hypothesis is true. Very likely to occur strictly by chance.

4.3.5 Drawing a Conclusion:

i. P-value <= threshold significance level (α) => Reject your null hypothesis in favour of your alternative hypothesis. Your result is statistically significant /important and it is different from the assumption or null hypothesis.

ii. P-value > significance level (α) => Accept null hypothesis. Your result is same as assumed.

5. STATISTICAL ANALYSIS TECHNIQUES

Finding structure in unstructured data and making predictions for facilitating decision making are the most important steps in Data Science. Here, in particular, statistical methods are necessary since they are capable of handling many different tasks. Important examples of statistical data analysis techniques are the following.

5.1 Classification

Classification methods are among the basic techniques for finding and predicting subpopulations from data. In the unsupervised learning situation, such subpopulations are found out from the data sets with having nay prior knowledge about the data set or the population. This is often called as clustering.[2]

In the supervised case, classification rulesare from a labelled training data set for the prediction of unknown labels when only influential factors areavailable. Nowadays, there is a plenty of methods for the unsupervised [2] as well for the supervised types [1]. A new approach is needed to learn and deploy the classical statistical methods since the computation efforts of complex analysis methods grows stronger than linear with the number of observations *n* or the number of features *p*.

5.2 Regression

Regression analysis is a predictive modelling technique which represents the relationship between a **dependent** (target) and **independent variable (s)** (predictor). For example, relationship between rash driving and number of road accidents by a driver can be best studied through regression.[4]

As mentioned above, regression analysis is used to estimate the relationship between two or more independent variables. Let’s understand this with an easy example:

Let’s say, we want to calculate growth in sales of a company based on current economic conditions. If we have a company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past structured information. [9]

There are multiple advantages of using regression analysis. They are as follows:

i. It indicates the significant relationships between dependent variable and independent variable.

ii. It indicates the strength of impact of multiple independent variables on a dependent variable.

When the output is numeric value, we learn a numeric function. If there is no noise, this task is called interpolation. In *polynomial interpolation*, $r^t = f(\boldsymbol{x}^t)$. Given *N* points, we find the *(N*-1*)* degree polynomial that we can use to predict the output for any $\boldsymbol{x}$. If x is outside the range of given data then the it is called extrapolation. If there is noise added, it becomes *regression.*

$r^t = f(\boldsymbol{x}^t) + \epsilon$

The explanation for noise is that there are extra *hidden* variables that we cannot observe $r^t = f*(\boldsymbol{x}^t, \boldsymbol{z}^t)$ where $\boldsymbol{z}^t$ denote those hidden variables.

5.3 Time Series Analysis

Time series analysis focuses on understanding and predicting temporal structure [3]. Prediction for Time Series Data is one of the most important challenges. Typical application areas of time series analysis are the behavioral sciences and economics along with the natural sciences and engineering. As an example, consider signal analysis, e.g., speech or music data analysis. Here, statistical methods are based on the analysis of models in the time and frequency domains. The main task is to predict the future values of the time series itself or of its properties. For example, we can model the vibrato of an audio time series in order to predict real tone in the future. The fundamental frequency of a musical tone might be predicted by rules learned from elapsed time periods [6]. In econometrics, multiple time series are often analyzed [6].

6. REPRESENTATION AND REPORTING

Data visualization is the representation of structured and processed data in a pictorial or graphical format. It makes easy for the decision makers to see the results and better be able to derive conclusions from it by observing the visual patterns among the data.[5]

With interactive visualization, we can drill down in the charts and graphs for more minute details by interactively changing the data that is visible to see further processing or change the results based on requirements in real time.

Data Visualization can help in identifying areas that need improvement, it can be used to know the factors affecting customer behaviour or to learn customer patterns to improve sales or to predict sales.

There are a number of different ways to data visualization. The most common is the information representation, which typically includes statistical graphics. This approach comprises of following seven areas of concentration:

i. Displaying news

ii. Displaying data

iii. Displaying connections

iv. Displaying websites

v. Mind maps

vi. Articles and resources

vii. Tools and services

A more scientific approach to data visualization aims at computer science and may emphasize the following:

i. Visualization algorithms and techniques

ii. Volume visualization (2D and 3D renderings)

iii. Information visualization (visuals of abstract data)

iv. Modelling techniques (for business efficiency)

v. Multi-resolution methods (data modelling algorithms)

vi. Interaction techniques and architectures

Visualization to interpret found features, insights and storing of models in an easy-to-update and interactive form are very important tasks in statistical analytical methods and techniques to communicate the results and provide security to the data analysis deployment. Deployment is decisive for obtaining interpretable results in Data Science.

7. CONCLUSION

After studying the role of statistics in data analysis and data science, we can conclude by saying that the statistics is a very crucial part of data mining, analysis and data science but remains underestimated as compared Computer Science and Algorithms. Statistics provides the best of its methods to be used in data collection and pre-processing the data to advanced methods required for prediction and analysis.

As a data scientist play their role in designing models and algorithms for data science based systems, a statistician can also play a key role in this modern and emerging field of data science.

Thus we come to a conclusion that combining statistics with other branches of data science can lead to an advanced branch of science, mathematics and engineering to serve mankind.

8. ACKNOWLEDGEMENTS

We would like to thank all the authors and online content writers for giving us valuable information to gain an insight into the role of statistics and prepare this paper.

9. REFERENCES

[1] Aggarwal, C.C. (ed.): Data Classification: Algorithms and Applications. CRC Press, Boca Raton (2014)

[2] Claus Weihs, Katja Ickstadt, Data Science: the impact of statistics, International Journal of Data Science and Analytics, Springer, https://doi.org/10.1007/s41060-018-0102-5

[3] Aue, A., Horváth, L.: Structural breaks in time series. J. Time Ser. Anal. 34(1), 1–16 (2013)

[4] Brown, M.S.: Data Mining for Dummies. Wiley, London (2014)

[5] Cao, L.: Data science: a comprehensive overview. ACM Comput. Surv. (2017). https://doi.org/10.1145/3076253

[6] Lütkepohl, H.: New Introduction to Multiple Time Series Analysis. Springer, Berlin (2010)

[7] Wu, J.: Statistics = data science? http://www2.isye.gatech.edu/ ~jeffwu/presentations/datascience.pdf (1997)

[8] https://www.edureka.co/blog/math-and-statistics-for-data-science/

[9] https://link.springer.com/journal/11634

[10] https://onlinecourses.nptel.ac.in/noc20_cs46 /unit?unit=16&lesson=17

[11] https://statistics.laerd.com/statistical-guides /descriptive-inferential- statistics.php