iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VII **Month of publication:** July 2018

DOI: http://doi.org/10.22214/ijraset.2018.7058

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

Inferential Statistics for Data Science

Surapaneni Gopi Siva Sai Teja[1], P. Harika Reddy[2]
[1] *Student, Sreenidhi Institute Of Science and Technology, Hyderabad, India*

Abstract: Inferential Statistics is one of the key elemental skill required for the data science. This paper gives information about what is inferential statistics and how this statistics is back bone for data science, main areas of inferential statistics, Central Limit theorem, Sampling distribution which contains mean, range, standard deviation and variance. And also discusses sample proportion and sampling distribution of population. Confidence intervals consist of a range of potential values of the unknown population parameter. Hypotheses and their testing through observation are essential steps in the scientific process. There are two outcomes of statistical test, first a null hypothesis is rejected and alternative hypothesis is accepted, second null hypothesis is accepted, on the basis of the confirmation. In simple a null hypothesis is just opposite to alternative hypothesis.

I. INTRODUCTION

Inferential statistics is one of the two main branches of statistics. It uses a random sample of data taken from a population to describe and make inferences about the population. Inferential statistics are valuable when examination of each member of an entire population is not convenient or possible.

Example: Measuring the diameter of each nail that is manufactured in a mill is impractical. But a representative random sample of nails can be taken whose diameter can be measured. Information from the sample to make generalizations about the diameters of all of the nails.

Two main areas of inferential statistics

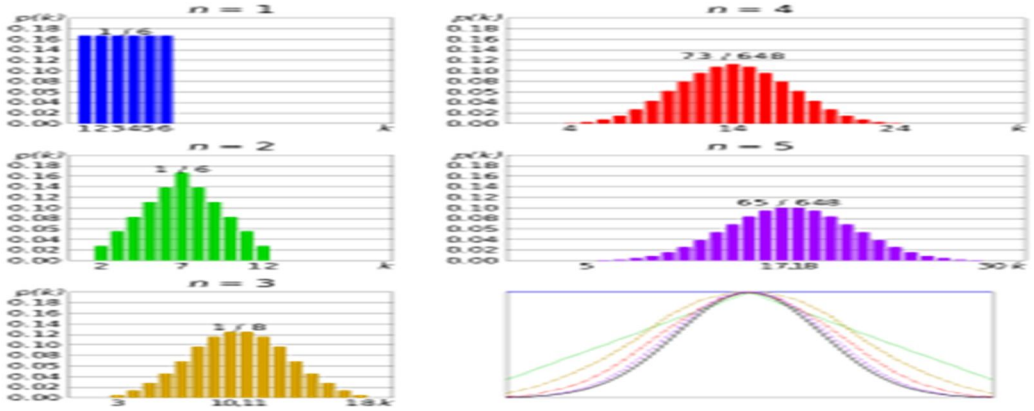
1) Estimating parameters: This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).
2) Hypothesis tests: This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

II. LITERATURE REVIEW

[1]Alien B Downey summarizes statistical analysis with the python. It focus completely on the understanding real life of statistics using different use cases. He also discusses about the Bayesian estimation. [2] Norm Matloff explained probability concepts and statistical measures and he also discusses about probabilistic models and choosing the best one for final evaluation.[3] William M Bolstad summarizes introduction to Bayesian estimation and introduces resources on statistics and scientific methods of data gathering.[4] Andy Field, Jeremy Miles and Zoe Field Discusses statistical concepts with R which are very useful and step by step understanding with examples.

III. CENTRAL LIMIT THEOREM

The Central Limit Theorem states that the sampling distribution of the sampling means approaches a normal distribution as the sample size gets larger no matter what the shape of the population distribution. It holds especially true for sample sizes over 30.For large samples graph of the sample means will look more like a normal distribution.



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

IV. SAMPLING DISTRIBUTION

A sampling distribution is a graph of a statistic for the sample data. A sampling distribution is where you a population (N), is taken and a statistic from that population is found.

Common statistics include:

1) Mean
2) Mean absolute value of the deviation from the mean
3) Range
4) Standard deviation of the sample
5) Unbiased estimate of variance
6) Variance of sample

V. SAMPLE PROPORTION

A sample proportion is where a random sample of objects *n* is taken from a population P; if x objects have a certain characteristic then the sample proportion "p" is: p = x/n.

Ex: 100 people are asked if they are democrat. If 40 people respond "yes" then the sample proportion is p = 40/100.

SAMPLING DISTRIBUTION OF POPULATION

The sampling distribution of a proportion is when you repeat your survey for all possible samples of the population.

Ex: Instead of polling 100 people once to ask if they are democrat, you'll poll them multiple times to get a better estimate of your statistic.

VI. CONFIDENCE INTERVAL AND HYPOTHESIS

Confidence Interval is an experiment can take a random sample from a lot or population and compute a statistic such as mean from the data to help understand the mean of the population. However the challenge /issue can be like how well the computed sample statistic (i.e. sample mean) estimates the underlying population. 'Confidence interval' is the solution to address the issue as it provides a range of values which is likely to contain the population parameter of interest.

How to construct confidence interval?

A confidence interval is constructed at a confidence level (usually 95%), selected by the user.

i.e. if the same population is sampled on numerous occasions and interval estimates are made on each occasion, the resulting intervals would bracket the true population parameter in approximately 95 % of the cases.

Types of confidence intervals

Confidence intervals can be one or two-sided.

A two-sided confidence interval brackets the population parameter from above and below.

A one-sided confidence interval brackets the population parameter either from above or below and furnishes an upper or lower bound to its magnitude.

What is hypothesis?

A hypothesis is an educated guess about something in the world around us. An experiment conducted statistically to know the true relationship between an independent variable and dependent variable.

Ex: A new medicine might work better in treating headache.

A good hypothesis statement should include an "if" and "then" statement addressing the independent and dependent variables. It should be testable by experiment, survey or other scientifically sound technique. Should be based on information in prior research and have a design criteria.

VII. HYPOTHESIS TESTING

Aims at testing the results of a survey or experiment to see if the results are meaningful.i.e. Testing whether the results are valid by figuring out the results that have happened by chance. If results may have happened by chance, then the experiment won't be repeatable and has little use.

Hypothesis testing usually involves:

1) Figure out null hypothesis.
2) State null hypothesis.
3) Choose the kind of statistical tests to be performed.

Either support or reject the null hypothesis

What is Null Hypothesis?

Null hypothesis is always the accepted fact

Ex:

1) There are eight planets in the solar system (excluding Pluto).

2) DNA is a double shaped helix.

VIII. NULL AND ALTERNATIVE HYPOTHESES

Used in the context of statistical analysis. Null hypothesis is symbolized as H_0 and usually expresses the uniformity/equality among the items in a sample

Ex: In a sample of two methods if a researcher has an assumption where both methods are equally good, then such assumption is termed as the null hypotheses

In the sample of two methods if a researcher thinks that method A is superior or method B is inferior, then such assumption is termed as alternative hypotheses. In a research if a sample does not support the null hypotheses, then it can be concluded that null hypotheses can be rejected and the researcher looks for a set of alternatives that form an alternative hypotheses.

If μ (mean) of a population is equal to the hypothesized mean (μ_0) then it is evident that null hypothesis can be population mean is equal to hypothesized mean.

How to State Null Hypothesis?

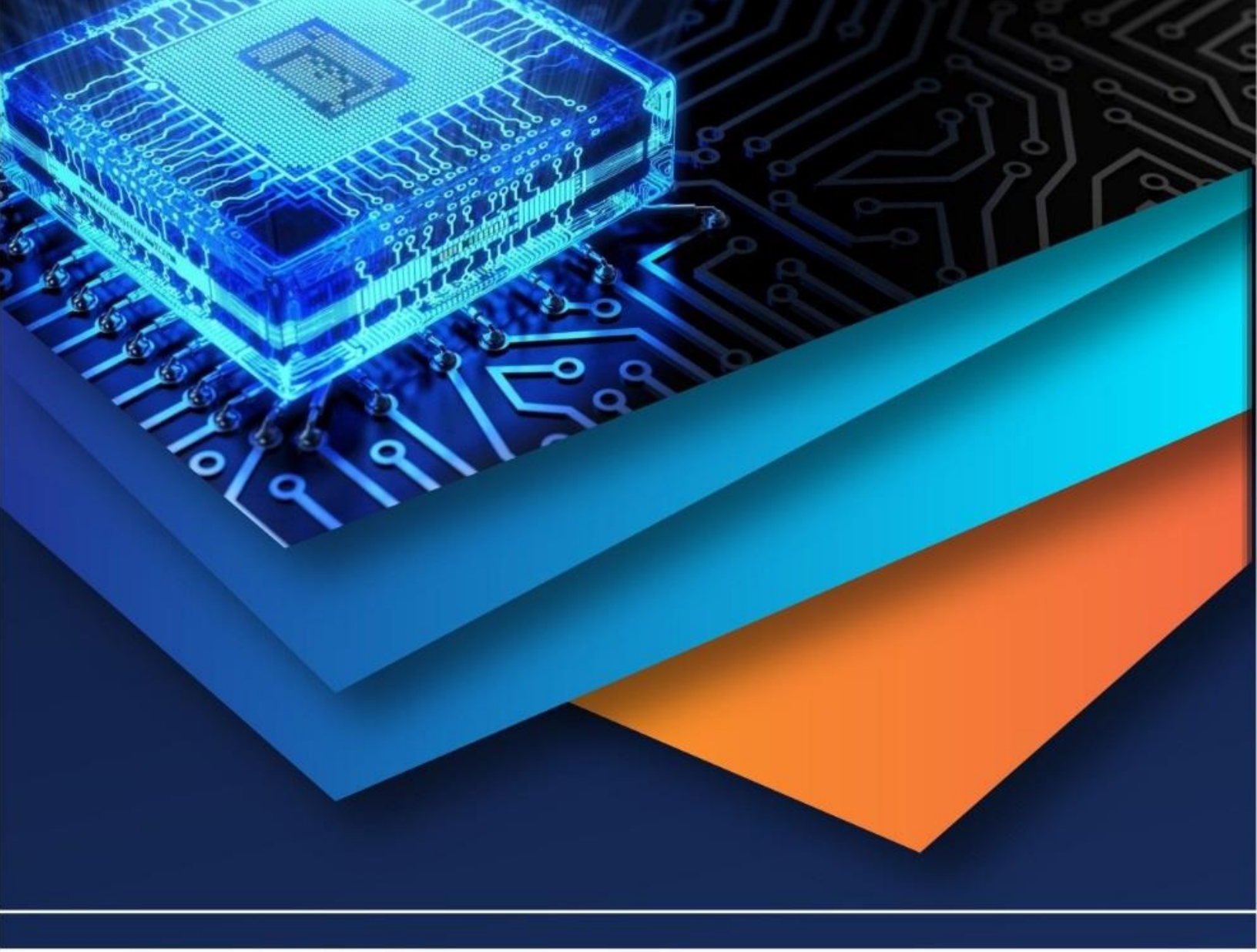
To state a null hypothesis, it would be sufficient if one can figure out the hypothesis from the world problems in problems of real life situations which can be a little trickier than just figuring out what the accepted fact

IX. CONCLUSIONS

This paper gives the complete view about the inferential statistics to make perception of the probability that an noticed difference between groups is a sturdy one or one that might have happened by chance in this study. Thus, we use inferential statistics to make interpretation from our data to general conditions, we use descriptive statistics to describe what's going on in our data.

REFERENCES

[1] Simplilearn. “Statistics for Data Science | Data Science Tutorial | Simplilearn.” YouTube, YouTube, 28 Aug. 2017, www.youtube.com/watch?v=Lv0xcdeXaGU.
[2] “How to Learn Statistics for Data Science, The Self-Starter Way.” EliteDataScience, 20 May 2018, elitedatascience.com/learn-statistics-for-data-science.
[3] “What Is Data Science vs. Statistics? - The Signal.” Mixpanel, 10 Dec. 2017, mixpanel.com/blog/2016/03/30/this-is-the-difference-between-statistics-and-data-science/.
[4] “What Is Data Science vs. Statistics? - The Signal.” Mixpanel, 10 Dec. 2017, mixpanel.com/blog/2016/03/30/this-is-the-difference-between-statistics-and-data-science/.
[5] Miller, James D. Statistics for Data Science. Packt Publishing, 2017.
[6] Hays, William L. Statistics. Wadsworth/Thomson Learning, 2008.
[7] Miller, James D. Statistics for Data Science. Packt Publishing, 2017.
[8] Ilersic, A. R. Statistics. H.F.L., 1979.
[9] “Data and Statistics about the U.S.” U.S. Data and Statistics | USAGov, www.usa.gov/statistics.
[10] Online Courses and Certificate Programs in Data Science. (n.d.). Retrieved from https://www.statistics.com/data-science
[11] S. (2017, August 28). Statistics For Data Science | Data Science Tutorial | Simplilearn. Retrieved from https://www.youtube.com/watch?v=Lv0xcdeXaGU
[12] How to Learn Statistics for Data Science, The Self-Starter Way. (2018, May 20). Retrieved from https://elitedatascience.com/learn-statistics-for-data-science
[13] Statistics for Data Science and Business Analysis. (2018, July 11). Retrieved from https://www.udemy.com/statistics-for-data-science-and-business-analysis/
[14] What is data science vs. statistics? - The Signal. (2017, December 10). Retrieved from https://mixpanel.com/blog/2016/03/30/this-is-the-difference-between-statistics-and-data-science/



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY