International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-09, Issue-02, May 2023

A Novel Approach to Understanding of Statistics in Various Engineering Fields

Abhinav Chandra, Student, Sandip University, Nashik, India, rajmitrachandra@gmail.com

Neetu Sharma, Professor, Sandip University, Nashik, India, neetu.sharma@sandipuniversity.edu.in

Abstract - Engineering is a field that heavily relies on statistical analysis and probability theory to make informed decisions. The use of statistics and probability in engineering can be seen in a wide variety of applications such as quality control, reliability analysis, experimental design, and risk assessment. In this paper, we will discuss the use of statistics in various engineering applications. Along with the use of statistics in various fields we will be analyzing a given dataset and using basic statistical methods to find a favorable answer from the given dataset.

Keywords —Statistics, Regression, Python, Mean, Visualization, ML

I. INTRODUCTION

Engineering is the application of scientific and mathematical principles to design and develop systems, structures, machines, and processes that improve human life. Statistical analysis and probability theory are fundamental tools in engineering that help engineers make informed decisions based on data. Statistical analysis involves the collection, analysis, interpretation, and presentation of data.

The word "Statistics" is derived from a German word "Statistik" which basically means 'description of a state, a country'. Statistics is concerned with the collection, organization, analysis, interpretation and presentation of data. A statistical model usually helps with the applications of statistics to an engineering, scientific, medical or a social situation where large amounts of data has to be worked with. Linear algebra, differential and integral calculus and probability theory are the major branches that are relied on while using statistics. Quality control, Design of experiments, reliability analysis, process optimization and statistical modelling are some of the major applications of statistics in engineering.

II. MEAN AND STANDARD DEVIATION

Mean and Standard Deviation form the basis on statistics. In essence, it is where statistics begins from. Being the easiest and simplest concept in Statistics, it is taught in schools and college alike from a very early period. For efficiency of the task at hand and overall time usage, a software such as Excel or any spreadsheet software is used. A particular example that helps students understand how these can be used in engineering applications was illustrated with test data from a faculty research project involving wireless communication (Zhan and Goulart 2009). Six different tests where conducted under four different test conditions. The basic task was to calculate the signal-to-noise ratio (SNR) for bandwidth of wireless communication using the given formula: [1]

$$SNR = \frac{mean}{stdev}$$

Table 1. SNR for bandwidth

Test No.	Condition 1	Condition 2	Condition 3	Condition 4
1	4.0	2.0	16.7	14.3
2	3.3	2.2	25.0	14.3
3	2.5	2.5	20.0	12.5
4	1.9	2.3	20.0	12.5
5	2.3	2.0	14.3	12.5
6	3.1	1.9	25.0	12.5
mean	2.85	2.15	20.17	13.10
stdev	0.76	0.23	4.32	0.93
SNR	3.73	9.52	4.67	14.09

If statistical analysis is not used, one might conclude that Condition 3 is the best condition but Condition 4 is the best condition. All of this is possible by the use of the SNR formula defined above. Using SNR it is also found out that condition 2 was better than condition 3. [1]

III. REGRESSION ANALYSIS

1. Machine Learning Regression

Finding the relationship between independent variables or features and a dependent variable or outcome is called Machine Learning Regression. It is used a method for predictive modelling in machine learning algorithms. In a forecasting or predictive model, regression analysis is an integral part. Regression is a common use for supervised machine learning models. Input and output data is a required label for this type of approach towards training models. Machine Learning regression models need to understand the relationship between features and outcome variables. [2]

The most common regression techniques in machine learning can be grouped into Simple Linear Regression, Multiple Linear Regression and Logistic Regression. [2]

 DOI : 10.35291/2454-9150.2023.0098

Regression is useful for engineers as it can help them identify the relationship between variables based on test data. [1]

Most Popular Statistical Regularization methods may include the following:

Ridge regression is a statistical technique used when analysing data or models of regression suffering from ill-conditioning or multicollinearity. Variances are usually large and far from the true value although the estimates may be unbiased. [3]

The Least Absolute Shrinkage and Selection Operator (LASSO), in this technique prediction accuracies are enhanced within a model thanks to both variable selection and regularization. Variable shrinkage is the process of choosing or selecting variables within statistical model results. [3]

2. **Use of Regression Analysis to study load parameters of mine excavator equipment:**

The presence of numerous operating factors that are challenging to describe using mathematical formulas, many of which are random, poses a significant problem in setting the actual loads in the working equipment of mining excavators. As a result, studying loads in working equipment while managing a mining excavator often proves difficult with traditional methods of analysis and modelling. Typically, when analyzing scientific research results, a scenario arises in which the quantitative variation of the studied quantity, i.e., voltage in the handle (σ), depends on multiple factors, including but not limited to density (x1), dust (x2), experience (x3), illumination (x4), learnability (x5), maneuverability (x6), mass (x7), noise level (x8), serviceability (x9), technological efficiency (x10), vibration (x11), and speed (x12). [4]

In theory, the regression model has the ability to account for any number of factors, but in practice, this is often unnecessary. The selection of factors is based on a qualitative analysis, but theoretical analysis alone may not provide a clear answer on the quantitative relationship between the considered features and the inclusion of a factor in the model. Therefore, factor selection is typically conducted in two stages: first, factors are selected based on the nature of the problem, and second, student statistics are used to determine regression parameters based on a matrix of correlation indices. [4]

If $f(x1, x2, \dots x12)$ represents a linear combination of factors, then the regression can be expressed linearly as:

$$\hat{y} = b_0 + \sum_{j=1}^{12} b_j x_j,$$

IV. MACHINE LEARNING IN SOFTWARE ENGINEERING

Software Engineering (SE) involves the creation, upkeep, and supervision of software systems with high quality standards in a manner that is both efficient and predictable. The field of SE research explores real-world scenarios by focusing on the development of new software systems, the alteration of existing ones, and the technology, tools, methods, techniques, or languages that support SE activities. The discourse on the significance of statistical analysis in Experimental Software Engineering (ESE) points to the insufficient use of statistical power in interpreting the results and casts doubt on the validity of the findings. Software Engineering typically employs scientific methods to assess the advantages of new software-related techniques, theories or methods. This methodology has been effectively utilized in other fields of science, particularly in the social sciences, which share similarities with Software Engineering as they both recognize the growing significance of the human element in software. Unlike physics or mathematics, it is challenging to establish laws of nature in these fields. To perform an experiment, it is crucial to choose a variable to observe and analyze its effects, which is known as the "factor." The various classifications of this factor are termed as "treatments." Generally, the aim of conducting an experimental study is to compare these treatments and ascertain whether they produce similar effects on a measured characteristic or if there is a noticeable distinction among them. [5]

Version 17 of the Minitab tool was utilized to generate the analysis presented below. This information can be generated in Minitab through the menu “Stat –> Basic Statistics –> Display Descriptive Statistics" by selecting the variable “Difference (Expected x Held)". Table 3 provides important insights into the data presented in Figure 3, including the minimum and maximum values for the variable “Difference (Expected x Held)", the range between the lowest and highest values, the first and third quartiles calculated based on the comparison with the variable “Time", and the standard deviation representing the difference between the median of each moment. [5]

To evaluate the distribution of the empirical data, a boxplot is used, which is formed by the first and third quartiles and the median. The figures associated with the moments before and after the implementation of the plugin can be observed in relation to the median drawn by analyzing the boxplot. To generate this graph, Minitab offers the menu option “Graph Box plot" with the selection of “One Y / With Groups" for the “Difference (Expected x Held)" variable and the “Moment" variable for the category. [5]

Outlier analysis is another method for checking the presented data. This analysis refers to observations in the samples that are either very far from the others or inconsistent when compared to them. These observations

 DOI : 10.35291/2454-9150.2023.0098

are also known as abnormal, contaminant, strange, extreme, or discrepant. To treat outliers correctly, it is important to determine the reasons that lead to their appearance. The possible reasons for the appearance of outliers include measurement errors, data running, or inherent variability of the elements of the population. Outliers resulting from collection or measurement errors should be discarded. However, outliers resulting from possible observed values should not necessarily be discarded. [5]

Year/Month	Held Hours	Expected Hours	Number of Cases	Cases Size	Difference (Expected - Held)	Moment
2013/12	259.878	100.000	36	M	-159.878	Before
2014/01	749.272	580.000	84	L	-169.272	Before
2014/02	570.343	480.000	74	L	-90.343	Before
2014/03	535.014	480.000	74	L	-55.014	Before
2014/04	311.262	90.000	33	S	-221.262	Before
2014/05	285.988	80.000	28	S	-205.988	Before
2014/06	279.633	80.000	28	S	-199.633	Before
2014/07	256.495	480.000	52	M	223.505	Before
2014/08	437.427	680.000	52	M	242.573	After
2014/09	450.845	395.367	58	M	-55.478	After
2014/10	225.472	517.222	75	L	291.750	After
2014/11	602.305	791.996	95	L	189.691	After
2014/12	450.147	452.305	62	M	2.158	After
2015/01	327.089	516.024	70	L	188.935	After
2015/02	258.536	503.461	65	L	244.925	After
2015/03	310.315	620.772	80	L	310.457	After

Table 2: Planning Data (Expected and Held Values).

TREND MEASUREMENTS	VALUES - Difference (Expected - Held)
AVERAGE	33.6
MEDIAN	-26.4
MODE	* (NUMBER OF MODES 0)
TRACK	531.7
MINIMUM	-221.3
MAXIMUM	310.5
1st QUARTILE	-166.9
3rd QUARTILES	237.8
VARIANCE	40244.0
STANDARD DEVIATION	200.6

Table 3: Analysis of measurements.

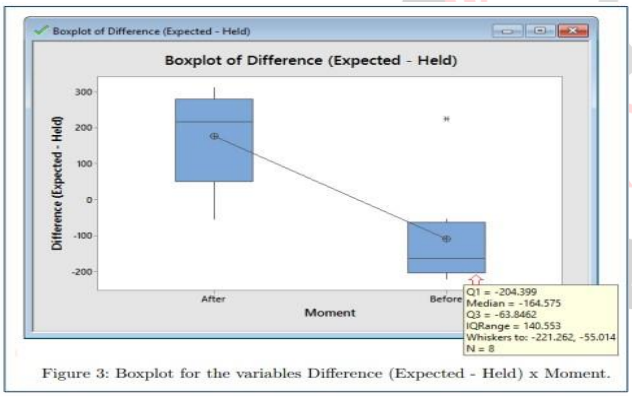


Figure 3: Boxplot for the variables Difference (Expected - Held) x Moment.

In order to identify outliers in a sample, it is necessary to calculate the median, lower quartile (Q1), and upper quartile (Q3) values. The difference between Q3 and Q1 is then calculated and stored as L. Values outside the range of Q3+1.5L to Q3+3L or Q1-1.5L to Q13L are deemed outliers and may be accepted in the population. However, values that exceed Q3+3L or are less than Q1-3L should be treated as outliers and investigated to determine their source of dispersion, as they are the most extreme points in the analysis. [5]

V. Use of Matplotlib in Python to Visualize a Dataset

We have acquired a dataset from a free use dataset website going by 'Kaggle' which displays the population of the various countries of the world. For our intents and purposes we have decided to use Python as a programming module for our Data Visualization so that we could have an idea of the population distribution of the various continents throughout the world. The pandas and matplotlib libraries/packages are used to plot these graphs so that we can have a visual idea of the population distribution of the continental regions of the world.

Raw Code:

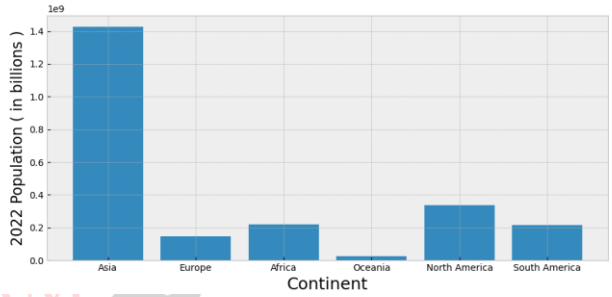
```
import matplotlib.pyplot as plt
import pandas as pd

plt.style.use('bmh')
df = pd.read_csv('world_population.csv')

x = df['Continent']
y = df['2022 Population']

plt.xlabel('Continent', fontsize = 18)
plt.ylabel('2022 Population', fontsize = 16)
plt.bar(x,y)
plt.show()
```

Graphical representation:



1e9 = 1,000,000,000

The X axis shows the continents of the world whereas the Y axis shows the increasing populations of the various regions of the world (in millions)

With just a few lines of code and statistical knowledge we can see that the population distribution throughout the world is not uniform. Asia having the largest population of more than a billion, reaching more than 2 billion, while areas such as Oceania have populations of less than 200 million.

VI. Conclusion

Thanks to the vast scale of Statistics, the application of statistics is not limited to only these methods. The methods discussed here are merely a tip of the iceberg in the vast ocean that is statistics. Mean and deviance, regression, adoption in machine learning are just some of the many concepts where Statistics can be applied. The future is bright as the adoption of Artificial Intelligence and its subset, Machine Learning is being adopted rapidly in nearly

 DOI : 10.35291/2454-9150.2023.0098

every application available. And the core foundation of these concepts is none other than Statistics.

REFERENCES

[1] Application Of Statistics In Engineering Technology Programs by Wei Zhan, Texas A&M University, USA Rainer Fink, Texas A&M University, USA Alex Fang, Texas A&M University, USA

[2] Seldon Blog Machine Learning Regression Explained

[3] Machine Learning – Regression by Haidara SALEH

[4] Practical application of regression analysis to study load parameters of mine excavator equipment by V S Velikanov, N V Dyorina, E I Rabina and T Yu Zalavina

[5] Application of Statistical Methods in Software Engineering: Theory and Practice by Tassio Sirqueira, Marcos Miguel, Humberto Dalpra, Marco Antônio Araújo, and José Maria David

[6] An alternative method to north-west corner method for solving transportation problem by Neetu M Sharma, Ashok P Bhadane

[7] Matplotlib 3.7.1 documentation by John D. Hunter

[8] Statistics by Robert S White and John S White

[9] World Population Dataset by Sourav Bannerjee

 DOI : 10.35291/2454-9150.2023.0098