Deep Data Stat - A Survey Analysis on Impact of Statistics in Data Science for Students

A. Mohammed Harun Babu R[1], B. Shebana M [2,]
[1]*Data Science & AI, iNurture Education Solutions, Bangalore*
[2] *Data Science, iNurture Education Solutions, Bangalore*

***Abstract*— Data science is the most developing technology in recent years. The need of Data science is most important thing for the development of institutions. It is the process of analyzing, interpreting and decision making of data. There are various methods are included in the analysis of data science. Among those components, Statistics plays an important role. Without the help of statistics, the data cannot be analyzed. The arrangement and visualization of the data are also done with the use of Statistics. This paper explains the basic statistical methods used in the process of analyzing the data in Data science. As the basic terminologies are explained in the beginning, the advanced tools such as Hypothesis testing, Analysis of variance, t test, F test and Chi square tests are discussed. Then, the interconnection between the Data science and Statistics are explained with the calculations of two tests such as Tukey test and Dunnet test. Finally, the future development and the impact of Statistics in Data science have been explained.**

***Keywords*—Statistics, Hypothesis, ANOVA, Prediction analysis, Test Statistics**

I. INTRODUCTION

As the technology develops rapidly, machines will replace the men in everywhere. People can do anything from wherever they are. The comfort level of people is the basic seed for the development of business. The institutions should know the people's interest and do the business according to their interest. The running of a business is not an easy task. There are a lot of ups and downs in the money earning sector of the business. To raise a successful business, businessmen and employers of that institution should know the two important categories. Those are interpretation understood by the past profit and the involvement of identifying the future prediction [1].

These two components can be identified only with the help of analyzing the data using some technique. The technology which is created to solve this type of problems are called Data science. In simpler words, the process of analyzing, interpreting and decision making is called Data Science [2]. As started to using it in minor games, its level reached into higher medical research, space investigation etc. The researcher should analyze the data given by the people and refers the products according to their need. The future prediction of profit in the companies is also based on Data Science [4].

In the analysis of data, the result cannot be identified without the use of Statistics. It plays an important role in Data science. Statistics is a mathematical science that includes the compilation, analysis and interpretation of data and it is also a mathematical discipline to collect and summarize information [1]. Statistics is a collection of procedures and rules used to reduce large amounts of data to manageable Proportions and to allow people to draw conclusions from those data. There are many statistical method are used to calculate the final result. Some of those statistical methods has been explained in this paper.

The explanation work of the statistical methods has been arranged as follows. Section II contains the background of the paper which describes the basic terminologies of statistics, Section III starts with the methodology of the paper which consists of methods and calculations of Probability, Hypothesis, Analysis of variance, T-test, F-test, Chi square test, Section IV explains the bridge between the data science and statistics which describes the importance of Data science and Statistics and the advanced techniques such as Tukey test and Dunnet test and finally Section V ends with the conclusion of the paper which gives the future development of the impact of statistics in data science.

II. BACKGROUND

In the process of doing Data science of the population, there are some important steps involved in it. They are domain specific software systems, Information storage and management, Data performance improvement, Data modeling and Representation, Advanced analysis, Training and development of the model, Simulation and test design, high efficiency computing and monitoring, Networking, Communication, Intent to decision and behavior [21].

For conducting all these methods, an important seed is needed and that is Data. Without data, nothing can be

Manuscript received May 28, 2020; accepted December 16, 2020.
*Corresponding author: contact.harishharunn@gmail.com

processed. By having the data as the root of the analysis, researcher can conduct any methods or calculations with such data. In the beginning of the calculation of Data science, Researcher asks himself so many questions. The solution for all those questions can be identified with the help of statistics [22].

Before entering into the statistical methods used in the field of statistics, there are some important terminologies has to be understood. For that purpose, Table I shows the notations, equations and descriptions of important terms in statistics.

TABLE I.REPRESENTATION OF EQUATIONS AND DESCRIPTION OF TERMINOLOGIES IN STATISTICS

TERMINOLOGY	NOTATION OR EQUATION	DESCRIPTION
MEAN	$X=\Sigma X_i/N$	X=mean X_i=individual values N=number of elements
MEDIAN	ODD=N/2 EVEN=(N+1)/2	N=number of elements
MODE	MODE=X_i<1	X_i=individual value
STANDARD DEVIATION	$\sigma=\sqrt{(\Sigma X_i-\mu^2)/N}$	X_i=individual value μ^2=Population mean N=number of elements σ=Population standard deviation
VARIANCE	$SV=\Sigma(X_i-X)^2/N-1$	X=mean X_i=individual values N=number of elements
COVARIANCE	$PC=\Sigma(X_i-X)/N$ $SC=\Sigma(Y_i-Y)/N-1$	X,Y=mean values SC=Sample covariance N=number of elements Xi,Yi=individual values PC=Population covariance
COEFFICIENT OF VARIATION	PCV=(σ/100%)*μ SCV=(s/100%)*μ	μ=Population mean s=sample standard deviation σ=Population standard deviation SCV=sample coefficient of variance PCV=population coefficient of variation
SIGNIFICANT DIFFERENCE	σ=0.01,0.095,0.05	σ=Alpha value
UNION	U	U=notation of union
INTERSECTION	∩	∩=notation of intersection

III. METHODOLOGY

A. Probability

The topic of probability is seen in many facets of the modern world [1]. From its origin as a method of studying games, probability has involved in a powerful and widely applicable branch of mathematics. The uses of the probability range from the determination of life insurance premium, to the prediction of electron outcomes, the description of the behaviour of molecules in a gas. Our entire world is filled with uncertainty. People make decisions affected by uncertainty virtually every day. In order to measure uncertainty, they turn to a branch of mathematics called theory of probability. Probability is a measure of likeliness that an event will occur [7].

There are three main important types in probability. They are Marginal probability, Joint probability and Conditional probability. The probability which deals with the actually occurring event is called Marginal probability. The probability which deals with the intersection of two or more events is called Join probability. The probability which deals with already occurred events are called Conditional probability [20].

As knowing the concepts of the basic statistical tool probability, there are many statistical tools will be explained in Fig 1.

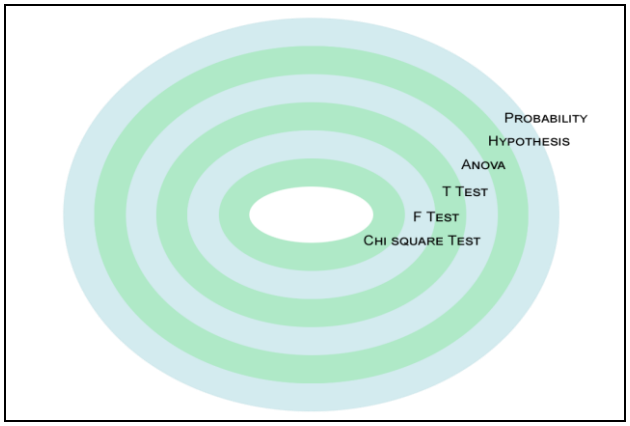

Fig. 1.Representation of inter-relationship between various statistical methods.

B. Hypothesis

Hypothesis is a theory or a claim that can be tested by doing some experiments to find the desired output [3] For the analysis of data, hypothesis should be declared. In data science, the analysis of given data cannot be done without the help of hypothesis. After declaring the hypothesis, that claim must be tested. The process of doing experimentation in the hypothesis is called hypothesis testing. It is a systematic way to test the assumptions or theories about a group or population. Hypothesis testing is also called as sense testing where it can be calculated with the help of population parameter and collected data [1]. The hypothesis testing is very important in the analysis where it can only check whether the primary idea of a researcher is accurate or not. Hypothesis testing can be done with the data of many groups. There are two types of hypothesis in analysis. They are Null hypothesis and Alternative hypothesis [9].

Null hypothesis is a basic assumption of a researcher about the specific data. It is a currently accepted parameter. Null hypothesis was denoted as H_0 as in equation 1.

$$H_0 : \mu_x = \mu_y = \mu_z \qquad \ldots\ldots(1)$$

This is the equation of null hypothesis. It denotes that there is no statistical significant relationship between all groups [6].

Alternative hypothesis is also called as research hypothesis where it was an accurate result of the experiment. It always be as the negotiation of null hypothesis.

Alternative hypothesis can be denoted as H_1 or H_A [11] as an equation 2.

$$H_A : \mu_x \neq \mu_y \neq \mu_z \quad \ldots\ldots(2)$$

This is the equation of alternative hypothesis. It describes that means of all groups was not same and there is a statistical significant relationship in the given data [6]

TABLE II.REPRESENTATION OF TYPE 1 ERROR AND TYPE 2 ERROR.

Condition	H_0 is true	H_0 is false
Reject H_0	**Type I error**	Correct decision
Accept H_0	Correct decision	**Type II error**

Table II shows the possible outcomes of type 1 and type 2 errors.

As everything has a negative side, sometimes hypothesis test also has an negative output where it meets some errors. The main two errors in hypothesis testing was Type 1 error and Type 2 error [7]. Type 1 error occurs when analyst rejects the null hypothesis when it was actually true. The probability of committing Type 1 error is denoted by α as in equation 3 and 4 .

$$\alpha = P \text{ (type 1 error)} \quad \ldots\ldots (3)$$

$$= P \text{ (rejecting } H_0 \text{ \& } H_0 \text{ is true)} \quad \ldots\ldots (4)$$

Type 2 error occurs if the null hypothesis H_0 is accepted, when it is actually false .The probability of committing type 2 error is denoted by β as in equation 5 and 6.

$$\beta = P \text{ (type 2 error)} \quad \ldots\ldots (5)$$

$$= P \text{ (accepting } H_0 \mid H_0 \text{ is false)} \quad \ldots\ldots (6)$$

To find the amount or a level of error occurred in a hypothesis, the power of the test should be calculated. The equation to define the power of hypothesis test has been shown in equation 7.

$$\text{Power of the test} = 1 - \beta \quad \ldots\ldots (7)$$

C. ANOVA

In data science, the most important technique used for the analysis of data is ANOVA. It refers to Analysis of variance [8]. This method is used to identify the significant relationship between the groups. The data which contains two or more independent variables can be able to processing ANOVA. It has been done by comparing the variable response means at various factor scales. In simple words, ANOVA is a hypothesis test which compares the means of two or more population. It is a statistical tool which deals with quantitative data [3].

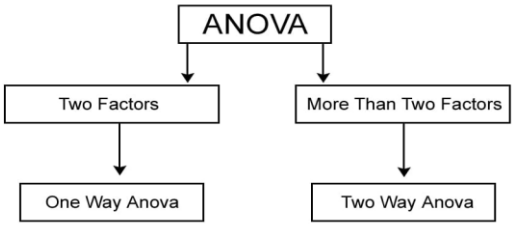

Fig 2. Representation of types of ANOVA

Fig 2 represents the types of Analysis of Variance. When the data contains two factors with one independent variable, it comes under One way ANOVA. When the data contains more than two factors with one independent variable, it comes under Two way ANOVA [9].

One-way analysis of variance involves only one categorical variable, or a single factor. Hypothesis in One way ANOVA can be denoted as,

$$H_0: \mu_1 = \mu_2 \quad \ldots\ldots (8)$$

$$H_A: \mu_1 \neq \mu_2 \quad \ldots\ldots (9)$$

Here, the equation 8 represents that the mean values in two groups were same and there is no statistically significant relationship between the groups. The equation 9 represents that the mean values in two groups were unequal and there is a statistical significant relationship between the groups [12].

TABLE III .REPRESENTATION OF METHODS AND EQUATIONS INCLUDED IN THE CALCULATION OF ONE WAY ANOVA.

Source variation	Sum of squares	Degrees of freedom	Mean squares	F-value	P-value	F-critical value
Between the groups	$\Sigma N_i(X_i-X)^2$	N-1	SS_G/DF_G	MS_G/MS_E	Tail Area above F	Value of F for α
Within the groups	$\Sigma(Ni-1)S_i^2$	N-1	SS_E/DF_E	-	-	-
Total	$SS_T=SS_G+SS_E$ $\Sigma(X_0-X)^2$	N-1	-	-	-	-
Co efficient of Determination $R^2=SS_G/SS_T$			Pooled standard deviation $MS_E=S_P$			

Here, the statistical tool which is used to find the relationship between the data which has two factors has been explained in table III. But if the data has more than two factors, another statistical tool has to be used and that is Two way Analysis of Variance. It is a statistical technique which deals with large number of data with an equal number of observations in each group. The primary purpose of Two way ANOVA is to understand whether there is an interaction between the two independent variables [13].

As the data for Two way ANOVA has more than two factors, the null and alternative hypothesis for such data can be denoted as,

$$H_0: \mu_1 = \mu_2 = \mu_3 \ldots\ldots = \mu_N \quad \ldots\ldots (10)$$

$$H_A: \mu_1 \neq \mu_2 \neq \mu_3 \ldots\ldots \neq \mu_N \quad \ldots\ldots (11)$$

Here, H_0 denotes the null hypothesis where it describes that there is no significant relationship between the n number of groups as equation 10 and H_A denotes the alternative hypothesis as in equation 11 where it describes that there is a statistical significant relationship between the groups.

D. T test

The identification of the statistical significant relationship between the various groups in the data will be very useful for the analysis. For further analysis, the consequential difference should be calculated for the given data. For that process, T-test can be used. It is a process of determine the consequential differences between the groups and also find out if the two data sets are different. It is a statistical tool to checks whether the mean values among various groups were accurately different or not [13].

In the calculation methods, there are two important components in T-test. They are One sample T-test and Two sample T-test and its displayed in Fig 3. The One sample T-test is used to test the null hypothesis that the average of the population is equal to an obvious value. The Two sample T-test is used to correlate the average values of both groups.

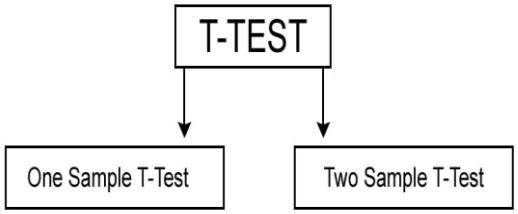

Fig 3. Representation of types of t test

For conducting T-test, T-value of the data should be calculated. The T-value is expressed as the ratio of the difference between the average of the two sample sets and the difference that exists within the average of two sample sets. If the calculated T-value is high, it denotes that there is a large difference among the various groups. If the calculated T-value is low, it denotes that there is a less difference among the various groups. The equation to find the T-value of given data is shown below as equation 12 where the Numerator denotes the difference of an average of two sample test. Denominator denotes the difference exists within the various groups [14].

T-value =Variance between groups/Variance within groups …… (12)

E. F test

While the given data set is statistically different, the correlation of variance and standard deviation should be calculated. This process can be done through F-test. It is a method to compares the variances and standard deviation of various groups but only when it is statistically significant. While conducting F-test, when all the given parameters are zero then it will be used to analyze the regression of such hypothesis. It is used in a data set to determine the model with the perfect fit and it can be performed only when the Statistical models are fitted by same underlying factors. This method mainly scopes on ratio of the variance which has quantitative values [10]. So, it is called as "Ratio Test". It is a test in which the test statistic has an F distribution under null hypothesis. The F distribution often arises when we are working with ratios of variances. An important aspect of F-test is the data's independent variable must have normally distributed and contains at least ordinal scales.

For the calculation of F-test, the first step is to determine the null and alternative hypothesis. After declaring the hypothesis, the F-value should be calculated. The equation to calculating F-value was shown in equation 13.

$$\text{F-value} = (SSE_1\text{-}SSE_2/m) / (SSE_2/n\text{-}k) \quad \text{….. (13)}$$

Here, Residual Sum of square of group 1 is denoted as SSE_1 and Residual Sum of square of group 2 is denoted as SSE_2, n represents total number of values, m represents number of restrictions and k represents number of independent variables [15].

The next step in the calculation of F-test was identifying the critical value of the data. Critical value can be find from F-distribution table. The equation to find the critical value is shown in equation 14.

Critical value = means between the group variance / Means within the group variance …… (14)

By comparing the F-value and the critical value, the researcher come to the conclusion whether to accept or reject the null hypothesis.

F. Chi Square Test

While collecting a data and making analysis in it, each and every component is very important. Here, our scientists discovered all statistical tools to analyze the data. In addition to values, variance and means, frequencies of the data also used to analyze the data. The statistical tool which deals with frequencies of the data is called Chi-Square test. This method is used to analyze a data when there is a large number of counts and frequencies in a data. This test assesses whether two categorical variables are related in one-way. A simple Chi-Square begins with the single independent variable with two levels. There is one important Chi square design called two cell design. The easiest form of this test is a two-cell experiment in which the result is that each event falls within one of the two cells. The Chi-Square test will be denoted as X^2. [16]

Commonly Chi-Square deals with two groups where declaring the null hypothesis and alternative hypothesis as follows in equations 15 and 16:

$$H_0: P_A = P_B \quad \text{…… (15)}$$

$$H_1: P_A \neq P_B \quad \text{…… (16)}$$

Here, P_A denotes probability of group A and P_B represents probability of group B. In Chi-Square test, the observed and expected frequency of a data should be identified. The formula used for the calculation of expected cell frequency was shown in equation 17.

$$X^2 = (\text{observed value} - \text{expected value})^2 / \text{Expected value} \quad (17)$$

After identifying the expected cell frequency, the significance level should be fixed. Commonly used

significance level in statistical method was 0.05 and 0.01. As like F-test and T-test, identification of the critical value is important in Chi-Square test. For a particular significance level, the critical values needed for locating the rejection region depending upon the degrees of freedom as in equation 18.

Degrees of freedom = Number of Cells-1 …… (18)

The next step is to compare the calculated value with the critical value, if the critical value is greater than the calculated value, the null hypothesis can be accepted.

The table IV of observed and expected frequencies should be represented as Rows(r) and columns(c) in a contingency table or a two way frequency table.

TABLE IV. REPRESENTATION OF POSSIBLE OUTCOME OF DATA USED IN CHI SQUARE TEST

Variable	**Data type 1**	**Data type 2**	**Total**
Category 1	A	B	A+B
Category 2	C	D	C+D
Total	A+C	B+D	A+B+C+D

The columns symbolized by (C) represent the number of factors in the data. The rows symbolized by (R) represent the number of categories in the data. It is called a contingency table because we are trying to find if the two variables are dependent on each other or not.

IV. BRIDGE BETWEEN DATA SCIENCE AND STATISTICS

Data science is a well developing carrier for many youngsters. As the business gets develop, the need for a data science also getting higher. Data Science is very essential in both small shops and also multinational companies. Statistics plays an important role in the analysis of the quantitative data in data science field. The depth of the analysis and the statistical tools used for the calculation will get higher when the large number of data models is being analyzed. For the clear analysis of the survey, more number of statistical models has to be used [3].

With the use of Analysis of variance, the research can describes the statistical significant relationship between the groups only. But the location of the difference and the level of the means relationship have to be calculated. This process can be done with the help of post hoc tests. When the given data is statistically significant, the post hoc tests of the data can be identified. Post hoc is a Latin word which refers to 'after that'. It denotes that this type of tests can be done only after declaring the data as statistically significant and the calculation of F-test. Post hoc tests will give specific information on which means of three or more groups are differ significantly from one another [5].

There is a lot of post hoc tests are used in statistical models. They are Tukey test, Dunnet test, Fisher least significant difference, Duncan's new multiple range test.

A. Tukey Test

When the researcher needs to find the homogeneity of variances among the means in various groups, Tukey test can be used. It is a process which deals with the data of the groups which are independent in both between and within themselves. To calculate the variance, the mean of the each experiment should be normally distributed. Tukey test is also called as an honest significant difference because the real variance among the factors has been calculated [18].

To test all pair comparisons between means using the Tukey honest significant difference, the following equation 19 is used to calculate the variance level.

$$HSD = (M_I - M_J) / (\sqrt{MS_W / N_H}) \quad \ldots\ldots (19)$$

Here, M_I and M_J represent the mean values of different levels and $M_I - M_J$ denotes the difference between the pair of means. For this equation, the value of M_I should be greater than the value of M_J. MS_W denotes the value of Mean square within and N represents the number of values in a data.

B. Dunnett Test

After the process of Tukey test, the multiple comparisons of the data should be defined. I can be done with the help of Dunnett test. This method has a fixed control group where it can compare its value with all other samples. By doing this process, a deep analysis of multiple comparisons has been identified. Dunnett test compares averages from several observations groups against a control group average to see is there is a difference [19].

Dunnett's calculation is similar to the T-test because this test also deals with two groups. The equation for the calculation of Dunnett test is follows as equation 20:

$$D_{DUNNETT} = (T_{DUNNETT}\sqrt{2MS}) / N \quad \ldots\ldots (20)$$

Here, $T_{DUNNETT}$ represents the table value of mean squares and MS represents the value of Mean square and N denotes the total number of values in an experiment.

VI. CONCLUSION

Statistics is not only consists of data collection and presentation in graphs and tables, it is also used in the science basing inferences on observed data and making decision in the face of uncertainty [1]. Several years ago, Statistics is only concerned with economic, political and demographic characteristics. But nowadays, it has been expanded and includes many things and it is so important in day to day life that many people frequently use statistical analysis to make decisions without even realizing that. In addition to Data science, Statistics plays major role in Banking, Economics, Mathematics, Business, Sociology, Accounting etc [4].

In this paper the statistical methods which is used to analyze the data has been explained. But in the process of data science, still there are so many steps are involved. After using these techniques for determination of past development, the future prediction of the data has to be identified. Visualization and Mapping of such data will be used for the identification of accurate result [11].

Day to day, more number of systems such as Machine learning, Artificial Intelligence, Image processing, Biometric Authentications are developed with the help of Data Science. So, the inter connection between the Data science and Statistics plays an important role in the development of the business and enhances the comfort level of People.

ACKNOWLEDGMENT

This study was presented orally as abstract paper at the ICONDATA 2020 conference.

VII. REFERENCES

[1] Weihs C, Ickstadt K. Data Science: the impact of statistics. International Journal of Data Science and Analytics. 2018 Nov 1;6(3):189-94.

[2] De Veaux RD, Agarwal M, Averett M, Baumer BS, Bray A, Bressoud TC, Bryant L,Cheng LZ, Francis A, Gould R, Kim AY. Curriculum guidelines for undergraduate pro-grams in data science. Annual Review of Statistics and Its Application. 2017 Mar 7;4:15-30.

[3] Cleveland WS. Data science: an action plan for expanding the technical areas of the field of statistics. International statistical review. 2001 Apr;69(1):21-6.

[4] Donoho D. 50 years of data science. Journal of Computational and Graphical Statistics.2017 Oct 2;26(4):745-66.

[5] Ostertagova E, Ostertag O. Methodology and application of oneway ANOVA. American Journal of Mechanical Engineering. 2013 Nov;1(7):25661.

[6] Niedoba T, Pieta P. Applications of ANOVA in mineral processing. Mining Science.2016;23.

[7] Sha_er JP. Multiple hypothesis testing. Annual review of psychology. 1995 Feb;46(1):561-84.

[8] AJPAS A. A Feature Selection Based on One-Way-Anova for Microarray Data Classi_-cation. AJPAS JOURNAL. 2016;3:1-6.

[9] Sow MT. Using ANOVA to examine the relationship between safety security and human development. Journal of International Business and Economics. 2014 Dec;2(4):101-6.

[10] Waller MA, Fawcett SE. Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. Journal of Business Logistics. 2013 Jun;34(2):77-84.

[11] Savin NE. Multiple hypothesis testing. Handbook of econometrics. 1984 Jan 1;2:827-79.

[12] Stoline MR. The status of multiple comparisons: simultaneous estimation of all pair- wise comparisons in one-way ANOVA designs. The American Statistician. 1981 Aug 1;35(3):134-41.

[13] Park HM. Comparing group means: t-tests and one-way ANOVA using Stata, SAS, R, and SPSS.

[14] Kim TK. T test as a parametric statistic. Korean journal of anesthesiology. 2015 Dec;68(6):540.

[15] Moser BK, Stevens GR, Watts CL. The two-sample t test versus Satterthwaite's approxi- mate F test. Communications in Statistics-Theory and Methods. 1989 Jan 1;18(11):3963- 75.

[16] McHugh ML. The chi-square test of independence. Biochemia medica: Biochemia medica. 2013 Jun 15;23(2):143-9.

[17] Plackett RL. Karl Pearson and the chi-squared test. International Statistical Re- view/Revue Internationale de Statistique. 1983 Apr 1:59-72.

[18] Abdi H, Williams LJ. Newman-Keuls test and Tukey test. Encyclopedia of Research Design. Thousand Oaks, CA: Sage. 2010:1-1.

[19] Koenig F, Brannath W, Bretz F, Posch M. Adaptive Dunnett tests for treatment selection. Statistics in Medicine. 2008 May 10;27(10):1612-25.

[20] Billingsley P. Probability and measure. John Wiley Sons; 2008 Aug 4.

[21] Peck R, Olsen C, Devore JL. Introduction to statistics and data analysis. Cengage Learn- ing; 2015.

[22] Noether GE. Introduction to statistics: the nonparametric way. Springer Science Business Media; 2012 Dec 6.