The Basic Application of Biostatistics to Biomedical Science Using R Programming

Ogbolu Melvin Omone[1]
BioTech Research Center, EKIK,
Óbuda University,
Budapest, Hungary,
ogbolu.melvin@biotech.uni-obuda.hu

Levente Kovács (PhD, habil)[2]
PhysCon Research Center, EKIK,
Óbuda University,
Budapest, Hungary
kovacs@uni-obuda.hu

Miklós Kozlovszky (PhD, habil)[3]
BioTech Research Center, EKIK,
Óbuda University,
Budapest, Hungary,
kozlovszky.miklos@nik.uni-obuda.hu

2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI) | 978-1-7281-8053-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/SAMI50585.2021.9378648

***Abstract*— Statistics is defined as the combination of numerous mathematical methods and logic models that are used for appropriate decisions making or judgments that involves uncertainty. However, the expected result could be certain/uncertain. Hence, it can be applied in many areas of life, which includes problem-solving purposes, investigations, and for making scientific conclusions. In biomedicine, the study of statistics is known as biostatistics. It involves the applications of control methods and pathophysiological modeling for data interpretation and presentation. This paper reveals the concept of data analysis as related to biostatistics and its methods which are useful for Statisticians with the use of R programming language. When raw data cannot be interpreted, it is then processed using fundamental statistical tools/methods. Therefore, this paper highlights the statistical methods that are required for a well-processed data and how the methods are applied. The purpose of this paper is to prove the ability to choose the statistical method and test which is suitable for a specific investigation, how to apply the test, and interpret the results using tables and graphs.**

Keywords*— *Biomedicine, Biostatistics, Control methods, Data analysis, Statistics, Statistical tools

I. INTRODUCTION

The application of biostatistics is essential in scientific investigations as it facilitates achieving meaningful results [1]. We would be right if we say that biomedicine is not complete without the help of biostatistics in the interpretation and presentation of it results [2]. Therefore, biostatistics is defined as the application of the mathematical tools, methods, and software applications which provides result for subjects related to the statistical fields of biological sciences and medical sciences [3], [4], [5]. It is a fast growing field which is related to many areas of science (especially biology); such as biotechnology, epidemiological studies, medical sciences, health sciences, bioeconomics, nutritional science, genetics, educational analysis, environmental biology, etc. [6]. The following areas of biostatistics are discussed in this paper;

i. Applications of biostatistics, and
ii. The fundamental methods and techniques of biostatistics.

A. Applications of Biostatistics

Biostatistics cannot be discussed without talking about statistics and the application of statistics. Basically, statistics is the collection, organization, summarization, and analysis of raw data (mostly numerical data) in large quantities [7], [8] with the application of mathematical formulas, tools/methods, and software applications [9]. Raw data are in form of data types.

1) What are data types?

In order to appropriately apply the statistical measurements which are accurate for making conclusions about certain assumptions, the basic concept of data types must be understood for data processing [10], [11]. There are mainly four (4) types of data type (also called measurement scales). They are;

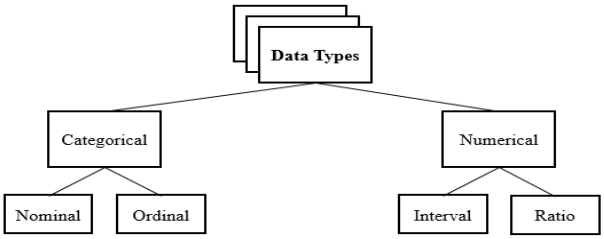

Fig 1: Hierarchical structure of Data types

a) ***Nominal data:*** Nominal data is used for labeling data variables; thus, the rows/columns of the contained raw data are labelled by categories (names [12]), with or lacking any quantitative values represented as counts/numbers. An example of nominal data type is a case whereby we say a patient may die of cancer or survive from cancer. For example, the question below from a survey contains nominal data:

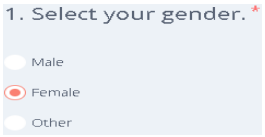

Fig 2: "Gender-related" question, which is a categorical question

b) ***Ordinal data:*** Ordinal data is used for grouping/ordering values contained in raw data [13]. It can measure non-numeric features such as the level of customer satisfaction, financial status, degree of happiness, level of experience, degree of discomfort, etc. It can also be used in ranking a particular feature as show in the example in Figure 3 below (where % ranges from low to high/moderate is the same as the ranges in number % as it increases "5% to 41% and above").

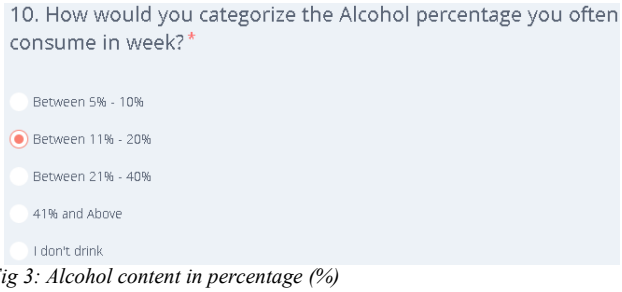

Fig 3: Alcohol content in percentage (%)

c) ***Interval data:*** Interval data is usually used to find intervals/differences between ordered values. It is either in continuous/discrete form [13]. For example, to find the interval between 20°C and 80°C, we have; 20°C – 80°C = -60°C. It is important to note that interval data do not have **an absolute value** or **true zero** i.e. in the example afore mentioned, result cannot equal **0°C**. Thus, possible values are

either negative/positive (i.e. -20°C, -30°C, -40°C, +80°C, +90°C, etc., but -0°C or +0°C does not exist in this case).

d) **Ratio data:** The ratio data type includes a few characteristics as the interval data type, but it does not have **an absolute value** or **true zero**. Hence, it intervals being definite are equal [14]. For example, ratio data type can be measured in values for height measurements, pulse rate, flow rate, weight measurements, length measurements, etc.

B. *The fundamental Methods and Techniques of Biostatistics*

Statistics is a scientific field of study that deals with a branch of applied mathematics. With the use of highly developed set of methods/tools; data collection, data analysis, data interpretation, data explanation, data clarification, and data presentation are made possible in all its applications [15], [16]. Statistics has two (2) major methods used for problem solving and for carrying out statistical tests; they are known as descriptive method and inferential method [17]. These methods use certain mathematical tools for data analysis. For better understanding of these tools [18], we shall perform some analysis on a dataset embedded in RStudio ("the MASS package" [19]) with the use of **R programming language** (results are shown below).

1) *Descriptive Statistics*

Descriptive statistics represent results in meaningful way as quantitative form or in visual. This method explains basic features and the distribution of population measurements, whereas data types are investigated using certain statistical methods, such as estimates of central tendency (i.e. mean, mode, and median), correlation coefficient, and measures of variability (i.e. standard deviation and correlation coefficient) [9]. In a nutshell, descriptive statistics is the enumeration, organization, and graphical representation of data from a given sample or an entire population which involves numerical variable summaries and measures [20].

2) *Inferential Statistics*

Inferential statistics is known to provide qualitative results. It is used to express the level of certainty about estimates about a population which includes hypothesis testing, standard error of mean, regression analysis, and level of significance [21]. It helps to draw conclusions from theoretical information or a population [22].

II. Tools For Descriptive Statistics

Descriptive statistical tools are the tools/methods used for the measurements and estimation of raw data that are descriptive in nature [23]. These tools are mainly used for raw data processing, information compressing (i.e. eradication of missing data), and for extracting useful information from a given sample or an entire population [22].

A. *Measures or Estimates of Central Tendency*

1) ***Mean (M):*** It is commonly used for measuring central tendencies and for the calculation of averages [24]. For example: Let's find the mean value of birthwt$bwt, we have;

```
> mean(birthwt$bwt) or with(birthwt, mean(bwt)) #we use this command to find the mean value of birthwt and bwt (mean of all numeric variables)
Output:
2944.587
```

2) ***Mode (Mo):*** It reveals the most frequent/appearing value in a set of observation or a population [24]. It is also called the point of maximum concentration. The values of mode are either discrete or rounded off [6]. For example: Let's find the group of children in the **race group** which appears the most, we have;

```
> table(birthwt$race) #we use this command to show the race column in a tabular form
Output:
Caucasian        Afro-American        Other
96               26                   67
> max(table(birthwt$race)) #we use this command to find the mode by frequency
Output:
96
> names(sort(-table(birthwt$race)))[1] #we use this command to print the name of the mode
Output:
"Caucasian"
```

3) ***Median (Me):*** It is an average value (the 50th percentile) which separates the higher half of a sample from the lower half. It obtains the **middle** value(s) of a set of data arranged from lowest to the highest [24]. For exmaple: Let's find the median value of birthwt$bwt, we have;

```
> median( birthwt$bwt ) #we use this command to find the median value of birthwt and bwt
Output:
2977
```

B. *Measures of Variability*

1) ***Standard Deviation (SD):*** It depicts the variability of the examination of the Mean (M) [24]. It is imperative to know that to find the value of SD, its square called variance must first be calculated. Thus, variance is the average square deviation around the mean and is calculated as; $Variance = \Sigma(x- \bar{x})\ 2/n\ Or\ \Sigma(x- \bar{x})\ 2/n-1$. Thus, $SD = \sqrt{variance}$.

SD helps us to predict how far a given value is away from the mean which is being calculated [25]. Results are usually more accurate when the data under investigation are normally distributed. Hence, if observations are clustered around the population sample **Mean (M)** and uniformly scattered around it, the SD helps to calculate a range that will include a given percentage (%) of observations [24]. However, if these observations are discrete and the central values are less descriptive of the given data, the variance is taken [1]. For example*:* Let's find variance before standard deviation;

```
> var(birthwt$bwt) #we use this command to find the variance of bwt
Output:
531753.5
> sd(birthwt$bwt) #we use this command to find the standard deviation of bwt
Output:
729.2143
```

2) ***Correlation Coefficient:*** Correlation can be define as the existing relationship between two variables. It is used to measure the degree (how strong) of linear relationship between two continuous variables. Thus, a linear relationship is said to exist if there is a direct relationship or a common inherent factors between two variables; it is expressed as a coefficient [1], [24]. The existing correlation coefficient values are always between **-1 and +1**. However, if the variables are not correlated, then the correlation coefficient is **0** (zero) and it is denoted by *"r"* (wherefore, *r* shows how closely data in a scatterplot falls along a straight line). The closer the absolute value of *r* and where a straight/linear line can be achieved in the scatterplot, the better the data are described by a linear equation. **If *r* = 1 or *r* = -1** then the given data is perfectly aligned as positive (+) or negative (-) results, respectively. Samples with values of *r* resulting to **0** reveals

slightly or no linear relationship [1], [26]. The mathematical formula for correlation coefficient is;

$$r = \frac{(\sum xy) - \frac{\sum x \sum y}{N}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right]\left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}} \ldots\ldots\ldots (1)$$

For example, the results for correlation is given below;

```
> cor(birthwt$lwt, birthwt$age) #Finding correlation between Low Weight and Age
Output:
0.1800732
> ggscatter(birthwt, x = "age", y = "lwt", add = "reg.line", conf.int = TRUE, cor.coef = TRUE, cor.method = "pearson", xlab = "Age of Children", ylab = "Children with Low Weight") #Graphical view
Output:
```

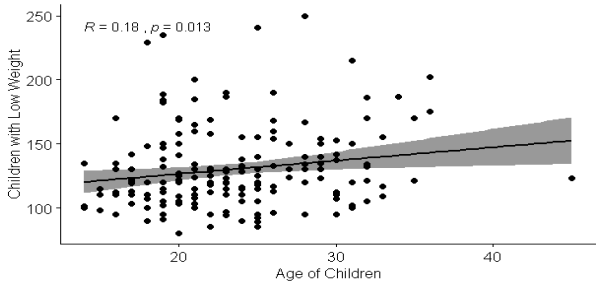

Fig 4: GGScatter for Correlation Coefficient between Age and Lwt using Pearson

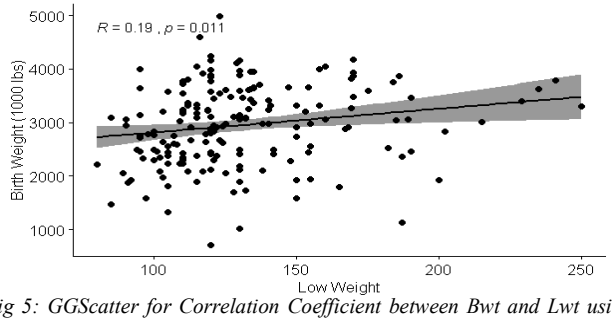

Fig 5: GGScatter for Correlation Coefficient between Bwt and Lwt using Pearson

In Figure 4 above, it shows a positive correlation between the population ages and their low weight. With the use of Pearson correlation, it reveals that there is a linear relationship between the two sets of data and a line graph is drawn to represent the two data. Thus, *R = 0.18*, with a *P-value, P = 0.013*. Also, in Figure 5 above, there is also a positive correlation between the birthweight and low weight. With the use of Pearson correlation, there is a linear relationship between the two sets of data and a line graph is drawn to represent the two data. Thus, *R = 0.19*, with a *P-value, P = 0.011*.

III. Tools For Inferential Statistics

Inferential statistical tools are used to make inferences from a given set of raw data/group/population [21], [27], [28]. Hence, before exploring the tools used for inferential statistics there are certain concept of distributions in descriptive statistics that must first be understood.

A. Types of Distribution

1) ***Gaussian or Normal Distribution****:* When data is symmetrically distributed on both sides of the **Mean (M)** and forms a bell-shaped (i.e. the standard normal distribution curve) [29] in the frequency distribution plot, such distribution is called normal/Gaussian distribution [1], [30]. This curve is usually used to describe the ideal distribution of continuous values or actual random variables. These values could be physiological measurements (e.g. human blood pressure, heart rate, blood sugar level, hemoglobin percentage (%) level, etc.) and size of living tissues.

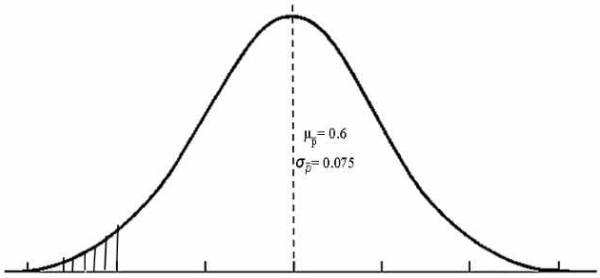

Fig 6: Normal Distribution Curve for One-sample [11]

In a normal distribution curve, the values of the **Mean (M), Mode (Mo),** and **Median (Me)** are usually the same within the population under investigation (i.e. the mean, mode, and median are all equal). With the use of a normal distribution curve, there are certain statistical tests for analysis which are used for making assumptions about normality. These statistical tests includes analysis of variance (ANOVA) [1], t-test, z-test, etc. [31]. A normal distribution curve is symmetric at the center, wherefore half of the values are to the left of center and the other half are pushed to the right (i.e. evenly divided), and the total area value under the curve is usually constant (i.e. **1**) [32], [33].

For example, the results for normal distribution is given below;

```
> hist(birthwt$bwt) #Plots a histogram for birthwt$bwt
Output:
```

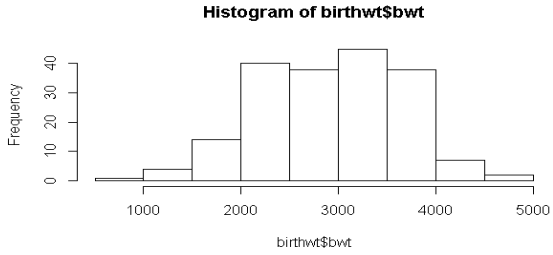

Fig 7: Normal Distribution for birthwt$bwt using Histogram

In Figure 7 above, the mass of the bwt seems to be equally distributed in the data. After plotting the histogram, it seems a normal distribution should fit well.

2) ***Non-Gaussian (non-normal) Distribution:*** In this case, if the data is skewed on one side of the graph, then the distribution is called a non-normal distribution [34].

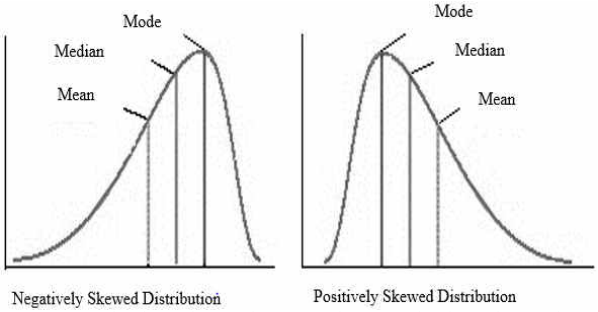

Fig 8: Negatively-skewed Distribution and Positively-skewed Distribution

The non-gaussian distribution is also known as the binominal distribution. In binominal distribution, event(s) can only have one of two possible outcomes such as yes/no,

positive/negative, survival/death, smokers/non-smokers, etc. It is a distribution with an asymmetric curve of the variables about its **Mean (M)** [35]. As shown in Figure 8 above, in a negatively skewed distribution, the mass of the distribution is concentrated on the right of the figure leading to a longer left tail, while in a positively skewed distribution, the mass of the distribution is concentrated on the left of the figure leading to a longer right tail [36].

For example: Let's check the normality and distribution of birthwt$lwt and birthwt$bwt using a histogram;

```
> hist(birthwt$lwt) #Plots a histogram for birthwt$lwt
Output:
```

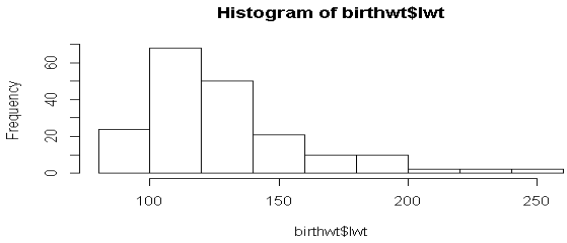

Fig 9: Non-Normal Distribution for birthwt$lwt using Histogram

In Figure 9 above, the mass of the *lwt* seems to be densely distributed in the data on the right, which signifies a positively skewed distribution.

B. Hypothesis Testing

When data is being analyzed using inferential statistics, the collected data is used to make inferences for a larger population, which is to help test for a hypothesis. By so doing, we are trying to find out whether our findings produced by a known question under investigation is true and not a guess work or an assumption [24]. Hence, hypothesis can be defined as the description (i.e. a possible answer or a predictive statement) for a phenomenon while hypothesis tests are the procedures used for making coherent conclusions about some findings which are usually accompanied by a test of statistical significance (i.e. methods that are scientifically testable and measurable) [1], [37].

1) Types of Hypothesis

a) ***Null Hypothesis (statistical hypothesis):*** The null hypothesis is denoted by H_0 ('H-naught' or 'H-null'). Its interpretation implies that there is no relationship/difference between the existing variables of the population under investigation. Also, the findings of H_0 could be an assumption that there exist no relationship/differences between two groups that are under an investigation [38].

b) ***Alternative hypothesis (research hypothesis)***: The alternative hypothesis is denoted by H_1 ('H-one' or H_a 'H-a'). Its interpretation means that a statement (finding) about two variables/groups under investigation is expected to be true. Also, it means that there is a relationship/difference between two variables/groups that are being investigated [24].

For example: Hypothesis testing example for the outcome of Bio-fertilizer 'x' on plant growth. We assume that a researcher has a new formulation Bio-fertilizer and wants to test the outcome on plant growth. According to Table I below, at the end of the experiment, we will either accept H_1 or H_0. Accepting H_0 does not mean that the investigation was/is a failure, it suggests that H_1 can be re-examined/modified to get a better result for the question asked under investigation or a better solution for the problem.

TABLE I. Experimental differences between Alternative Hypothesis (H_1) and Null Hypothesis (H_0)

Alternative Hypothesis	Null Hypothesis
It is the opposite of the null hypothesis.	It is the opposite of the alternative hypothesis.
A researcher wants to prove that the hypothesis is true.	A researcher wants to disprove or nullify that the hypothesis is true.
A researcher applies Bio-fertilizer 'x' and predicts that it improves plant growth.	A researcher applies Bio-fertilizer 'x' and predicts that it does not improve plant growth in any way.
This is for H_1: H_1 means that the application of Bio-fertilizer 'x' increases plant growth.	This is for H_0: H_0 means that the application of Bio-fertilizer 'x' does not increase plant growth.
The prediction states that there exist a statistical significance or relationship between the variables under investigation.	The prediction states that there does not exist a statistical significance or relationship between the variables under investigation.
If H_1 is accepted, it proves that the researcher's prediction is true.	If H_0 is accepted, it proves that the researcher's prediction needs to be re-examined.
Independent variable is Bio-fertilizer 'x'.	Independent variable is Bio-fertilizer 'x'.
Dependent variable is the plant. The plant is affected by the application of the independent variable 'x' which results into an increase in plant growth, increase in number of leaves, and increase in number of fruits. This means that the plant is dependent on Bio-fertilizer 'x'.	Dependent variable is the plant. The plant is not affected by the application of the independent variable 'x'. Hence, no result is expected. No increase in plant growth, no increase in number of leaves, and no increase in number of fruits. This means that the plant is not dependent on Bio-fertilizer 'x'.
The conclusion is that the independent variable can affect the dependent variable. Hence, H_1 predicts that there is a relationship between variable 'x' and the plant.	The conclusion is that the independent variable cannot affect the dependent variable Hence, H_0 predicts that there is no relationship between variable 'x' and the plant.

2) Types of Errors

When testing for statistical significance, there are two (2) types of errors that could occur during analysis. They are [4];

a) ***Type I Error (false positive):*** The type I error (denoted by α) occurs when H_0 is rejected because there is not a true conclusion. Thus, it occurs when the null hypothesis is rejected instead of being accepted/retained. The probability of it occurring has a threshold that is set at 0.05 (i.e. the significance level) [39]. A threshold of 0.05 means that we accept there is a 5% likelihood/probability of identifying an outcome when truly there isn't one existing [40].

b) ***Type II Error (false negative):*** The type II error (denoted by β) occurs if we fail to reject H_0 when there is a true conclusion to prove the investigation. Thereby yielding into a false negative result. Therefore, our conclusion is that there is no significant outcome when actually there really is one [40].

The difference between these two errors is that; type I error is falsely detects an outcome that is not existing, while a type II error is the failure to detect an outcome that exist.

3) Level of Significance

Level of significance involves confidence level, p-value, and the one-tailed and two-tailed test. It is the probability of rejecting a null hypothesis by a conforming test when it is true and it is also denoted by α, i.e. P *(type I error)* $= \alpha$ [24].

a) ***Confidence level (CI):*** CI gives a predictable range of values where there is an undetermined population parameter calculated from a determined population sample. The possibility of these parameters lies within a stated range of values (e.g. 0 – 10). The CI is denoted as c and given in

percentage (%) and is associated with the level of significance [1]. Therefore, the relationship that exist between level of significance and CI is written as: $c = 1 - \alpha$ [37], [41]. However, the level of significance and it conforming CI can be differentiated by the statements below;

i. When the level of significance is 0.10, it is associated with 90% CI.
ii. When the level of significance is 0.05, it is associated with 95% CI.
iii. When the level of significance is 0.01, it is associated with 99% CI.

For example; An example from one of our projects, where we calculated values for the CI of a **Mean (M)** in a graph. Results are shown in Figure 10 below;

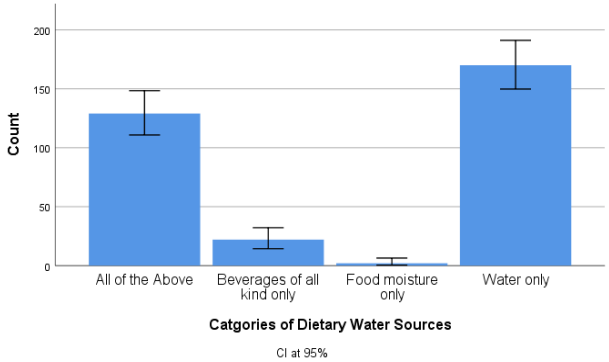

Fig 10: The participant's sources of TWI/dietary water

As shown in Figure 10 above, at 95% CI, there is no overlapping between participants who said only water is the source of dietary water and others who choose the rest of the category. Therefore, we conclude that a significant portion of the participants do not know the correct sources of dietary water. In other words, we can say majority of the participants (52.63%) belief that Total Water Intake (TWI) should come from plain water only, while only 39.94% of the population indicated correctly that TWI should come from water, food moisture, and beverages of all kind i.e. "All of the Above".

b) ***P-value:*** The p-value is defined as the probability that an event is likely to occur if H_0 is true. It is also known as the calculated probability [42]. Probability is the measure of the likelihood that an event will occur (i.e. the possibility of an outcome) and it is usually represented as a number; **0** or **1** (where 0 represents uncertainty/impossibility and 1 represents certainty/possibility). Therefore, **p-value** is a resulting number between **0** and **1** which can be used to interpret results where we decide if we want to reject/retain H_0 [43]. Therefore, to reject/retain the H_0, the following rejection rule should be considered;

i. If **p-value** $\leq$ level of significance, then the null hypothesis (H_0) is being rejected.
ii. If **p-value** $>$ level of significance, then the null hypothesis (H_0) is acceptable or not rejected.

The rejection region is the values of test statistic from the investigation for which H_0 is rejected, while the non-rejection region is the set of all possible values from the investigation for which H_0 is not rejected or cannot be rejected.

c) ***One-tailed and two-tailed test:*** The rejection region for two-tailed test is shown in Figure 11 below, while the rejection region for one-tailed test states that in the left-tailed test, the rejection region is shaded on the left side while in the right-tailed test, the rejection region is shaded on the right side [1].

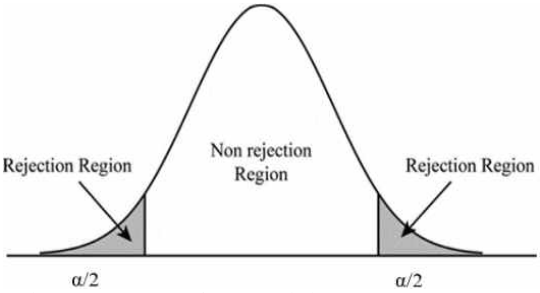

Fig 11: The rejection region for two-tailed test

d) ***Standard Error of Mean (SEM):*** Standard error measures the accuracy a sample population represents the **Mean (M)** value of an entire population; the sample mean is called Standard Error of the Mean **(SEM)**. However, in some cases, **SEM** and **SD** are used interchangeably to express variability, even though they can be used to measure different parameters during analysis. The difference between these two tools for variability measurement is that **SEM** quantifies uncertainty in the estimation of the **Mean (M),** while **SD** indicates dispersion of the data from **Mean (M)** [25], [43].

For example: Let us explain better what SEM and SD are using illustrations from one of our projects; We calculated the **Mean (M)** value and **SD** value to check if the entire population knows what European Food Safety Authority (EFSA) and World Health Organization (WHO) says about the sources of dietary water and what they personally think it should be. The findings of this sample are best described by two (2) parameters; **Mean (M)** and **SD** [25]. However, result shows that the participants are wrong about the sources of TWI and this can lead to overhydration. Most of the participants also could not indicate correctly how TWI has been recommended to be consumed by EFSA and WHO. Largely, 52.63% thinks water is the only main source of TWI and 51.08% thinks EFSA and WHO recommends that water is the only source of TWI. In Table II below, the **Mean (M)** of both results has no significant difference (i.e. the **Mean (M)** of "What do you think TWI should come from" is not significantly higher than "What does EFSA and WHO says TWI should come from") and indicates that both groups do not know the actual recommendation by EFSA and WHO.

Therefore, in this paper, **SEM** is simply the **SD** of the **Mean (M)** of certain random samples drawn from the original population [44] as shown in Table II below;

TABLE II. Mean and Standard Deviation values for TWI

	N	Mean	SD	SEM
What do you think TWI should come from	323	2.66	1.443	0.080
What does EFSA and WHO says TWI should come from	323	2.62	1.449	0.081

Table II above shows that the mean value for what the participants think TWI should come from is 0.04 more than the **Mean (M)** value of what the participants thinks EFSA and WHO recommends TWI should come from. Therefore, there is no significant difference, and we conclude that the population do not know the source of TWI.

IV. Conclusion

This paper has revealed that the most importance of biostatistics to biomedicine is the elucidation of raw data into knowledge. This paper may reveal the basic concept of biostatistics to researchers who are a novice in the use of R programming for statistical analysis.

According to the methods used for statistical analysis in this paper, conclusion is drawn that; descriptive statistics is most useful in describing the association between variables from a population sample. It provides a summary of the population in the form of **Correlation coefficient, Mean (M), Mode (Mo),** and **Median (Me)** while inferential statistics uses a random sample from the population to describe and make inferences about the whole population using **Null (H_0)** and **Alternative hypothesis (H_1), Confidence level (CI), Standard Deviation (SD), Standard Error of Mean (SEM)**, **Normal** and **Non-normal distribution** graphs and **P-value**. Hence, inference statistics is valuable when it is not possible to examine each variable in the whole population.

ACKNOWLEDGMENT

The authors thank the EFOP-3.6.1-16-2016-00010 project and the GINOP-2.2.1-15-2017-00073 named "Telemedicina alapú ellátási formák fenntartható megvalósítását támogató keretrendszer kialakítása és tesztelése, Hungary for their financial support.

REFERENCES

[1] G. Dakhale, S. Hiware, M. Mahatme, and A. Shinde, "Basic biostatistics for post-graduate students," *Indian J Pharmacol*, vol. 44, no. 4, p. 435, 2012.

[2] E. Cash and S. W. Boktor, "Understanding Biostatistics Interpretation," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.

[3] E. T. Liu, "Foundations for Systems Biomedicine: an Introduction," *AN INTRODUCTION*, p. 12.

[4] D. G. Altman, "Statistics in medical journals," *Stat Med*, vol. 1, no. 1, pp. 59–71, Mar. 1982.

[5] B. J. Strasser, *Biomedicine: Meanings, Assumptions, and Possible Futures*. Conseil suisse de la Science et de l'innovation CSSI, 2014.

[6] J. A. G. Iñiguez, "Basic Biostatistics for Clinical Research," Accessed: Sep. 25, 2019. [Online]. Available: https://www.academia.edu/37004352/Basic_Biostatistics_for_Clinical_Research.

[7] K. H. Ng and W. C. G. Peh, "Presenting the statistical results," *Singapore Med J*, vol. 50, no. 1, pp. 11–14, Jan. 2009.

[8] R. Romano and E. Gambale, "Statistics and Medicine: the Indispensable Know-How of the Researcher," *Transl Med UniSa*, vol. 5, pp. 28–31, Jan. 2013.

[9] S. H. Simpson, "Creating a Data Analysis Plan: What to Consider When Choosing Statistics for a Study," *Can J Hosp Pharm*, vol. 68, no. 4, pp. 311–317, 2015.

[10] "Types of data measurement scales: nominal, ordinal, interval, and ratio," *My Market Research Methods*, Nov. 29, 2012. https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/ (accessed Sep. 26, 2019).

[11] N. Donges, "Data Types in Statistics," *Medium*, Aug. 19, 2018. https://towardsdatascience.com/data-types-in-statistics-347e152e8bee (accessed Sep. 26, 2019).

[12] J. R. Dettori and D. C. Norvell, "The Anatomy of Data," *Global Spine Journal*, vol. 8, no. 3, pp. 311–313, May 2018.

[13] H. R. Marateb, M. Mansourian, P. Adibi, and D. Farina, "Manipulating measurement scales in medical statistical analysis and data mining: A review of methodologies," *J Res Med Sci*, vol. 19, no. 1, pp. 47–56, Jan. 2014.

[14] N. Lakshminarayan, "Know Your Data Before You Undertake Research," *J Indian Prosthodont Soc*, vol. 13, no. 3, pp. 384–386, Sep. 2013.

[15] T. Lang, "Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles," *Croat Med J*, p. 10.

[16] M. Petrove, "Statistical Errors in Biomedical Literature," p. 38.

[17] S. Al-Benna, Y. Al-Ajam, B. Way, and L. Steinstraesser, "Descriptive and inferential statistical methods used in burns research," *Burns*, vol. 36, no. 3, pp. 343–346, May 2010.

[18] C. G. Skinner, M. M. Patel, J. D. Thomas, and M. A. Miller, "Understanding common statistical methods, Part I: descriptive methods, probability, and continuous data," Mil Med, vol. 176, no. 1, pp. 99–102, Jan. 2011.

[19] B. Ripley, B. Venables, D. M. Bates, K. H. (partial port ca 1998), A. G. (partial port ca 1998), and D. Firth, *MASS: Support Functions and Datasets for Venables and Ripley's MASS*. 2020.

[20] F. Kaliyadan and V. Kulkarni, "Types of Variables, Descriptive Statistics, and Sample Size," *Indian Dermatol Online J*, vol. 10, no. 1, pp. 82–86, 2019.

[21] S. Allua and C. B. Thompson, "Inferential statistics," *Air Med J*, vol. 28, no. 4, pp. 168–171, Aug. 2009.

[22] "Understanding Descriptive and Inferential Statistics." https://statistics.laerd.com/statistical-guides/descriptive-inferential-statistics.php (accessed Sep. 30, 2019).

[23] T. G. Nick, "Descriptive statistics," *Methods Mol Biol*, vol. 404, pp. 33–52, 2007.

[24] O. M. Omone, M. Takacs, and M. Kozlovszky, "Statistical Hypothesis Testing of Patients' Risk-Score Assessment Test For Human Papillomavirus (HPV)," in *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*, Subotica, Serbia, Sep. 2020, pp. 137–142.

[25] P. Barde and M. Barde, "What to use to express the variability of data: Standard deviation or standard error of mean?," *Perspect Clin Res*, vol. 3, no. 3, p. 113, 2012.

[26] "Correlation Coefficient Definition." https://www.investopedia.com/terms/c/correlationcoefficient.asp (accessed Sep. 29, 2019).

[27] "Social Research Methods - Knowledge Base - Inferential Statistics." https://socialresearchmethods.net/kb/statinf.php (accessed Sep. 30, 2019).

[28] A. Omair, "Understanding the process of statistical methods for effective data analysis," *J Health Spec*, vol. 2, no. 3, p. 100, 2014.

[29] E. Limpert and W. A. Stahel, "Problems with using the normal distribution--and ways to improve quality and efficiency of data analysis," *PLoS One*, vol. 6, no. 7, p. e21403, 2011.

[30] D. G. Altman and J. M. Bland, "Statistics notes: the normal distribution.," *BMJ*, vol. 310, no. 6975, p. 298, Feb. 1995.

[31] "Normal Distributions (Bell Curve): Definition, Word Problems - Statistics How To." https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/normal-distributions/ (accessed Sep. 30, 2019).

[32] M. Maltenfort, "Understanding a Normal Distribution of Data (Part 2)," *Clin Spine Surg*, vol. 29, no. 1, p. 30, Feb. 2016.

[33] R. Bono, M. J. Blanca, J. Arnau, and J. Gómez-Benito, "Non-normal Distributions Commonly Used in Health, Education, and Social Sciences: A Systematic Review," *Front Psychol*, vol. 8, Sep. 2017.

[34] B. Cundill and N. D. Alexander, "Sample size calculations for skewed distributions," *BMC Med Res Methodol*, vol. 15, Apr. 2015.

[35] P. Hougaard, "Fundamentals of survival data," *Biometrics*, vol. 55, no. 1, pp. 13–22, Mar. 1999.

[36] J. Shreffler and M. R. Huecker, "Hypothesis Testing, P Values, Confidence Intervals, and Significance," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2020.

[37] C. Pernet, "Null hypothesis significance testing: a short tutorial," *F1000Res*, vol. 4, Oct. 2016.

[38] K. J. Rothman, "Curbing type I and type II errors," *Eur J Epidemiol*, vol. 25, no. 4, pp. 223–224, Apr. 2010.

[39] A. Banerjee, U. B. Chitnis, S. L. Jadhav, J. S. Bhawalkar, and S. Chaudhury, "Hypothesis testing, type I and type II errors," *Ind Psychiatry J*, vol. 18, no. 2, pp. 127–131, 2009.

[40] S. Greenland *et al.*, "Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations," *Eur J Epidemiol*, vol. 31, pp. 337–350, 2016.

[41] E. Whitley and J. Ball, "Statistics review 3: Hypothesis testing and P values," *Crit Care*, vol. 6, no. 3, pp. 222–225, 2002.

[42] C. Andrade, "The P Value and Statistical Significance: Misunderstandings, Explanations, Challenges, and Alternatives," *Indian J Psychol Med*, vol. 41, no. 3, pp. 210–215, 2019.

[43] Investopedia, "Standard Error of the Mean vs. Standard Deviation: The Difference," *Investopedia*. https://www.investopedia.com/ask/answers/042415/what-difference-between-standard-error-means-and-standard-deviation.asp (accessed Oct. 29, 2019).

[44] D. G. Altman and J. M. Bland, "Standard deviations and standard errors," *BMJ*, vol. 331, no. 7521, p. 903, Oct. 2005.