Available online at www.dergipark.org.tr/en

INTERNATIONAL ADVANCED RESEARCHES
and
ENGINEERING JOURNAL

Journal homepage: www.dergipark.org.tr/en/pub/iarej

e-ISSN: 2618-575X

International Open Access

Volume 08
Issue 03

December, 2024

Research Article

A research on the importance of testing the normality assumption in microbiological data

Murat Çimen [a,*]

[a] *CMN İstatistik ve Bilimsel Araştırma Merkezi, Çorum, Türkiye*

ARTICLE INFO

Article history:
Received 13 August 2024
Accepted 02 December 2024
Published 20 December 2024

Keywords:
Data
Microbiology
Normality test

ABSTRACT

The most used distribution in statistical analysis is the normal distribution. Parametric tests (e.g. one sample t-test) require that the data are normally distributed. In this study, milk somatic cell count data (SCC) used to test the normal distribution was obtained from a farm for the first and second month of lactation. According to the findings of the present study, SCC data of the first month showed normal distribution. However, the SCC data of the second month did not show normal distribution. Since the first month data showed a normal distribution, one-sample t-test, which is one of the parametric test methods, was applied for comparison with a specific reference value; since the second month data did not show a normal distribution, the Wilcoxon One-Sample Signed Rank Test, which is the non-parametric equivalent of the one-sample t-test, was applied. When the parametric test was applied to the second month data that did not show a normal distribution, results that did not comply with the standards in terms of SCC were found. When the same data was analyzed with the nonparametric test method, results that complied with the standards were obtained. It is noteworthy that different results are obtained in both analyses. As can be seen from the research results, since existing data sets in the field of microbiology may tend to show large variations, it should be tested whether the data show normal distribution before determining the statistical analysis method. According to the research results, the normality test must be applied in the statistical analysis of microbiological data showing large variations.

1. Introduction

Before applying statistical analysis, it is of great importance to determine the analysis method appropriate to the data sets at hand. Determining the appropriate analysis method is essential for the error-free completion of scientific research design, implementation, results acquisition and interpretation stages [1]. The results obtained from the statistical analysis output will affect the research hypothesis and will direct the inferential results of the research [2]. For this reason, classification of data and detection of data that disrupts normality are also important [3]. After detecting data that violates normal distribution, two options should be considered. The first of these is to remove absurd data from the data set. However, this is not possible in every case depending on the research hypothesis. The second method is to prefer nonparametric methods instead of applying parametric methods in the analysis of data that does not conform to normal distribution [4]. This second method is a more commonly used method. Before determining the most appropriate statistical method for the current data set, it is important to know whether the data are normally distributed. A statistical method chosen without applying the normal distribution test will lead to erroneous results and erroneous interpretations [5]. This will negatively affect the reliability of the results of scientific research. The first factor that impairs the reliability of results in scientific research is sometimes the interventions in the data as a result of manipulations made by researchers intentionally to direct the hypothesis, but sometimes it is also effective in unintentionally incorrect statistical analyses [6]. Any wrong application, whether intentionally or unintentionally, can cause information pollution. This is a situation that is not welcomed in the scientific community. It is noteworthy that in recent years, articles that create information pollution have been referred to ethics committees. Authors are primarily responsible for any information pollution that may occur

* Corresponding author. Tel.: +90-*530-226-9672*..
E-mail addresses: mcimen19@gmail.com
ORCID: *0000-0003-4290-2718*

DOI: 10.35860/iarej.1532064

as a result of the application of incorrect statistical methods. Insufficient statistical knowledge does not constitute a justifiable justification for the information pollution resulting from incorrect results [7]. Researchers must have sufficient statistical knowledge to publish articles. Therefore, researchers must test whether the data is normally distributed before applying statistical analysis to their data set. [8]. According to the results, researchers should prefer parametric methods if the data fits the normal distribution and non-parametric statistical methods if they do not [9]. The methods to be determined according to these results cannot be arbitrary and random, and determining the most appropriate method is absolutely essential. There are also problems in publishing research that does not comply with this rule in academic journals [10].

In this study, normal distribution test was applied to a sample data set obtained in the field of microbiology. Although a data set in the field of microbiology is shown as an example in the research, people working in other fields of Science and Health sciences must apply normality test to the parametric data set they have, as shown in the example. This test method is very important for their research to yield reliable results [11]. This study aims to be an exemplary research in terms of showing the importance of applying the normal distribution test to the data obtained in the field of microbiology with the help of SPSS in determining the statistical method.. In this study, it was aimed to determine whether data showing large changes, especially in the field of microbiology, conform to normal distribution, and also to reveal the negative results that would be caused by incorrect statistical methods applied to data that do not conform to normal distribution. Thus, researchers will be able to determine the method according to the results to be obtained by performing a normality test on the data they have before determining a statistical method. The meticulous application of the normality test will also prevent information pollution [12]. The current study creates significant awareness in terms of both showing the importance of applying the normal distribution test and showing the possible negative effects of choosing the right statistical method on research results.

2. Material and Methods

The material and methods section consists of obtaining milk data, milk analysis and statistical analysis sections.

2.1. Obtaining data

Milk data was obtained from Çağdaş Pertek Agriculture and Livestock Tourism Construction Food Production Marketing Industry Trade Ltd. Co. To determine somatic cell counts (SCC), 10 milk samples were collected from the first and second months of the enterprise is given in Table 1.

2.2. Milk analysis

Milk samples were collected in 100-ml sterile containers to determine somatic cell counts. The somatic cell counts within taken raw milk samples were detected by the standard analysis (Microscopic count) method.

2.3 Statistical analysis

Normality test was applied to determine whether the values shown in Table 1 show a normal distribution. Shapiro Wilk test was used to determine the normal distribution of the data. If the significance level found as a result of Shapiro Wilk analysis is less than 0.05, the data is considered to comply with the normal distribution, otherwise it is considered not to comply [13]. If the data are suitable for normal distribution, parametric tests should be applied, if not, non-parametric tests should be applied [14]. The normality test of the data, one simple t test and Wilcoxon One-Sample Sign Rank Test were calculated using the SPSS 18.0 package program produced by IBM Company.

Entering the data into the SPSS program and applying the Shapiro test are shown below. In the research, monthly data were subjected to normal distribution test. The stages of entering the data into the SPSS program and analyzing them are shown in Figure 1-3.

Figure 1 shows the data being entered into the SPSS program.

Then, the Plots button is pressed and in the window that opens, the fields marked none and the fields marked Normality plots with tests are marked (Figure 3).

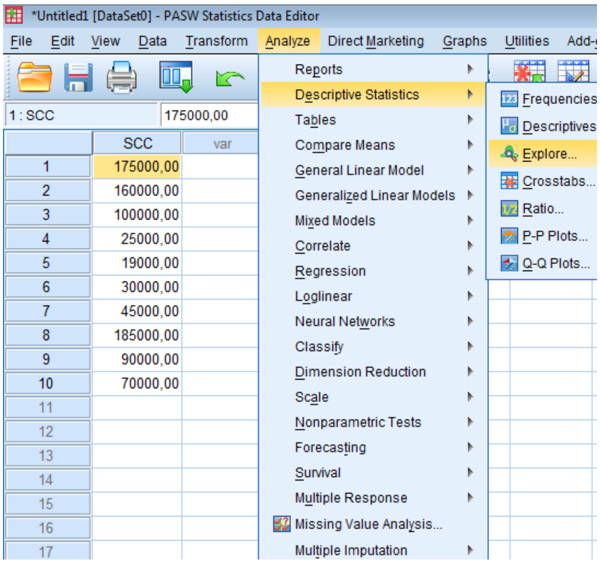
When you press Continue and then the OK button in the rear window, the analysis result is obtained.

3. Results and Discussion

It would be useful to examine the scatter plots before examining the normal distribution analysis of the data. As can be seen from Figure 4, the data show a homogeneous distribution on a line. There is no excessively concentrated distribution to the right and left of the line. Before performing the analysis, an idea arises that such a distribution can be expected to show a normal distribution.

Table 1. SCC for the first month and second month

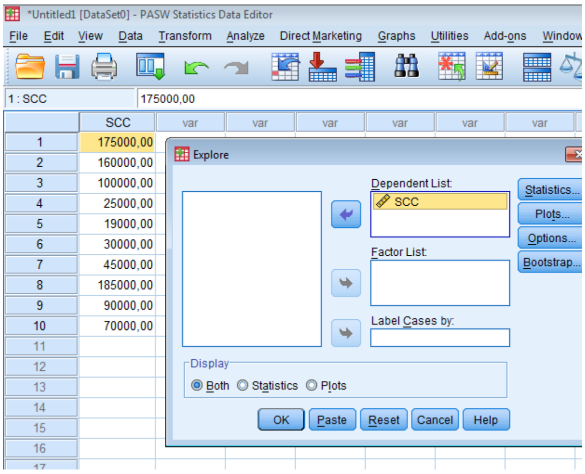
SCC for the first month	SCC for the second month
175000	150000
160000	180000
100000	10000
25000	85000
19000	13000
30000	10000
45000	15000
185000	190000
90000	170000
70000	185000



1 : SCC 175000,00

	SCC	var
1	175000,00	
2	160000,00	
3	100000,00	
4	25000,00	
5	19000,00	
6	30000,00	
7	45000,00	
8	185000,00	
9	90000,00	
10	70000,00	
11		
12		
13		
14		
15		
16		
17		

Figure 1. Entering the data into the Spss program


*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window
Explore
Statistics...
Plots...
Options...
Bootstrap...
OK Paste Reset Cancel Help

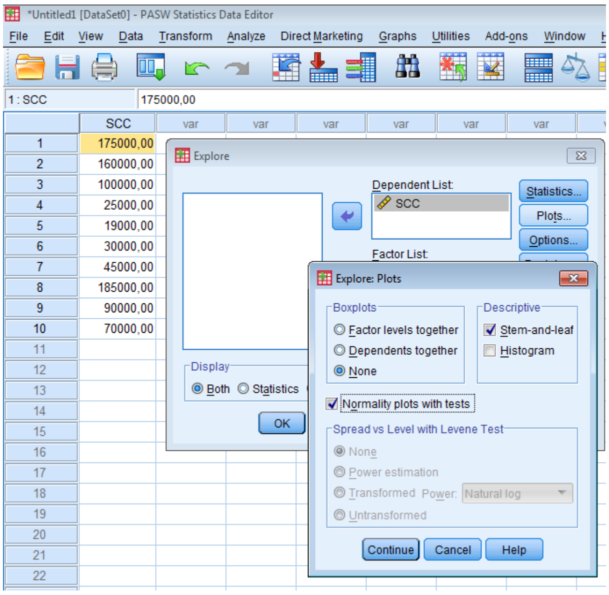
Dependent List: SCC

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Figure 2. Moving the dependent variable to the dependent list field


*Untitled1 [DataSet0] - PASW Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window
Explore
Statistics...
Plots...
Options...
OK
Explore: Plots
Continue Cancel Help

Dependent List: SCC

Factor List:

Display: ● Both ○ Statistics

Boxplots: ○ Factor levels together ○ Dependents together ● None

Descriptive: ☑ Stem-and-leaf ☐ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test: ● None ○ Power estimation ○ Transformed Power: Natural log ○ Untransformed

Figure 3. Using the plots window for normality testing

In the Explore window, the text SCC is moved to the section labeled Dependent List (Figure 2).

When Figure 5 is examined, larger variations in the distributions are noted compared to the previous figure. The idea that the large variation seen in Figure 5 may negatively affect the normal distribution for the second month data arises. However, the definitive result will emerge with the analysis to be performed. The benefit of scatter plots is that they enable visual detection of data that may affect the normal distribution in the data set and enable seeing where absurd data falls around the line [15].

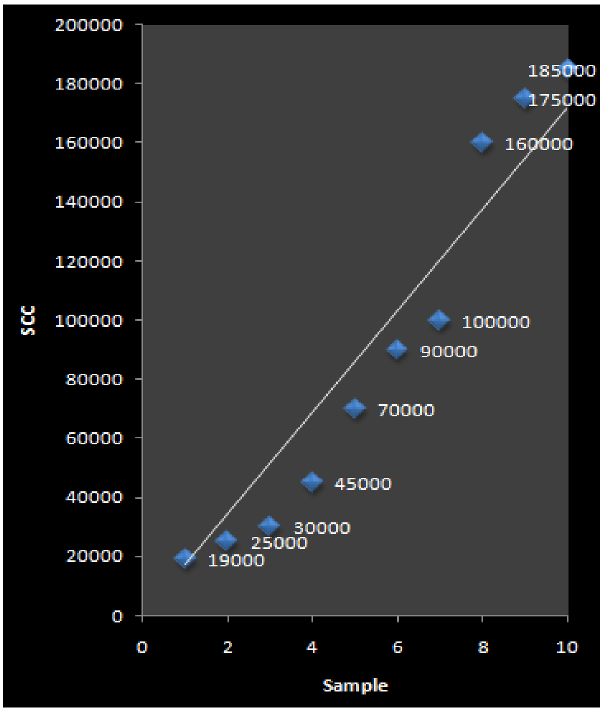


Figure 4. Scatter plot for SCC in first month data

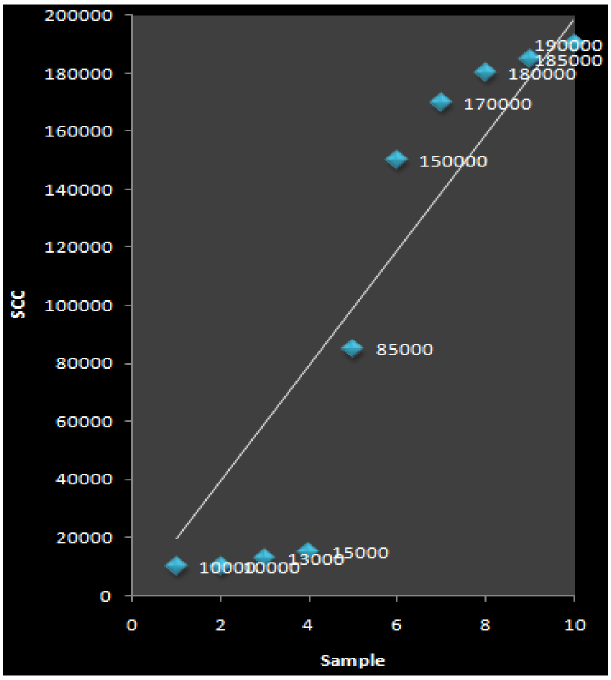


Figure 5. Scatter plot for SCC in second month data

It is noteworthy that the data in the first month data were distributed at narrower angles on the line. However, as seen in Figure 5, the data in the second month were distributed at wider angles on the line. This distribution can be expected to negatively affect the normal distribution. Because when the data is not arranged at equal distances and in equal numbers on the right and left of the line, this can negatively affect the normal distribution. However, as stated, in order to say this with certainty, the normal distribution test should be performed and comments should be made [16].

Although it is possible to bring the data closer to a normal distribution by removing the data that is farthest from the truth from the data set, this may not always be possible. Because while removing absurd data from some data sets does not negatively affect the structure of the data set, this is not possible in some data sets [17]. The removed data may have a feature that will disrupt the purpose of the hypothesis [18]. For example, in cases where all the examples of a population must be in the data set, such a removal is not possible. When the data is removed, the integrity of the population may be disrupted [19, 20]

Table 2 shows the normal distribution analysis results for the SCC values of the first month.

If less than 2000 data are used in the data set, the Sig value in the section labeled Shapiro Wilk should be taken into consideration. Since the current data set consists of 10 data in total, the Sig value in the section labeled Shapiro Wilk will be taken into account [21]. Since the sig value (0.156) found as a result of the research is above the value of 0.05, known as the significance level limit (0.05<0.156), it can be assumed that the data for the first month shows a normal distribution. The normality test was applied to the second month data in the same way.

As a result of the analysis shown in Table 3, it is concluded that the data for the second month does not show a normal distribution, since the sig value of 0.011 in the Shapiro-Wilk household is below the value of 0.05 (0.011<0.05). Therefore, since the first month data show a normal distribution, if a comparison is to be made with a certain reference value, one-sample t-test, one of the parametric test methods, should be applied, whereas since the second-month data does not show a normal distribution, the Wilcoxon One-Sample Sign Rank Test, which is the non-parametric equivalent of the one-sample t-test, should be applied [22, 23]. As mentioned before, absurd data in the second month data can be eliminated and brought closer to normal distribution. When the data approaches normal distribution as a result of the analysis to be made after this, parametric test methods can be used. However, this situation is not always desired. In addition, even if this method is desired, using missing data in the data set may not always comply with the hypothesis. For such reasons, it would be more meaningful to apply non-parametric data analysis instead of eliminating absurd data from the data set. Sometimes, when a data that should not be removed from the data set is removed, incorrect results that do not comply with the research hypothesis can be encountered [24].

One is curious about the results that will be seen when the second month data, which does not show a normal distribution, is analyzed with one sample t test. For example, a dairy farm does not want the somatic cell count to be over 20000. It is desired to know whether the second month milk data exceeds the limit reference value of 20000. When we look at Table 4, we see that there is a statistically significant difference (p = 0.012) when we compare the mean value (100800) with reference value (20000). The dairy farm sees the unwanted level of SCC data in these results. However, it would be wrong to apply this method since the data does not show a normal distribution.

Now, Wilcoxon one simple sign rank test is applied to the same data. Looking at the analysis results given in Table 5, when the numbers forming the data set are compared on a median basis rather than an average, there is no statistical difference compared to the reference value (p = 0.074> 0.05). Since the normal distribution is disrupted because the operating data shows great variation, there is no negative effect compared to the reference value of 20000 in terms of somatic cell numbers. In other words, the dairy farm did not encounter any negative results in the 2nd month compared to the reference limit value (20000). As can be seen, very opposite results were obtained from the two analyses. Accordingly, an analysis performed without knowing whether the data complies with normal distribution may cause information pollution.

As can be seen from the research results, since existing data sets in the field of microbiology may tend to show large variations, it should be tested whether the data show normal distribution before determining the statistical analysis method [25, 26]. Non-parametric test methods should be applied to data that do not show normal distribution [27]. Interpreting according to rank values is a method used in nonparametric tests and these tests are used in some research [28, 29].

Table 2. Normality test result for SCC in the first month

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig
SCC	.164	10	.200*	.887	10	.156

Table 3. Normality test result for SCC in the second month

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
SCC	.253	10	.070	.790	10	.011

Table 4. One-sample t-test analyses results

	N	Mean	Std. Deviation	Std. Error Mean
SCC	10	100800.00	81879.0	25892.42

One-Sample Test					
	Test Value = 20000				
				95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
SCC	3.12	9	.012	80800.00	22227.2	139372.7

Table 5. Wilcoxon one-sample sign rank test analyses results

Ranks				
		N	Mean Rank	Sum of Ranks
Reference value - SCC	Negative Ranks	6[a]	7.50	45.00
	Positive Ranks	4[b]	2.50	10.00
	Ties	0[c]		
	Total	10		

a. Referencevalue < SCC
b. Referencevalue > SCC
c. Referencevalue = SCC

Test Statistics[b]

	Referencevalue-SCC
Z	-1.785[a]
Asymp. Sig. (2-tailed)	.074

a. Based on positive ranks.
b. Wilcoxon Signed Ranks Test

In this research, while the first month data set showed a normal distribution, the second month data set did not show a normal distribution. Therefore, it is seen from the research results that the statistical methods to be applied for each month should be different. Before using parametric statistical methods such as t-test and analysis of variance, which compare the means of treatment groups, it must be tested whether the data sets of the treatment groups show normal distribution [30-32]. In order to obtain reliable information, research must be planned correctly and the data obtained must be tested according to appropriate analysis methods [33-35]. But, it is noteworthy that in some studies there are major errors in the collection and analysis of data [36, 37]. To prevent information pollution, care should be taken during the planning and analysis stages [38, 39]. Otherwise, there may be problems in the publication phase of research articles. Even if such studies are published, they may cause negativities in the academic life of the researchers because they create information pollution [40].

4. Conclusions

This study is of great importance in terms of emphasizing the importance of the normal distribution test and also showing how determining the wrong statistical method can negatively affect the research results. According to the results of the current study, in all studies where the parametric test method is applied, whether the data are suitable for normal distribution should be questioned before determining the current statistical method. Unfortunately, in many studies where parametric data analyses are used in the literature, it is not stated whether the normal distribution test is applied to the data. Studies where parametric data analyses are used and whether it is not known whether the data are suitable for normal distribution should not be accepted in journals. Authors should be asked whether they have applied the normal distribution test in their studies and whether they have determined the analysis method they have chosen according to the normal distribution test result. Studies where parametric data analyses are applied to data that are not known to meet the normal distribution condition should be viewed with suspicion and should never be accepted for publication in journals. This study will be a guide in determining the correct statistical method in future research.

Declaration

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article. The author also declared that this article is original, was prepared in accordance with international publication and research ethics, and ethical committee permission or any special permission is not required

Author Contributions

Murat Çimen: Obtaining data, analysis of data, article writing, literature review, editing

References

1. Al-Eideh, B.M., *Statistical methods for business data analysis using spss.* 2016, Scholars Press. ISBN-10: 9783639864892.
2. Flora, D.B., *Statistical methods for the social and behavioral sciences: a model based approach.* 2018, Pp. 472. Sage publications. ISBN-10: 1446269833.
3. Thalberg, M., *Statistical analysis: Methods and techniques for data interpretation.* 2024, Independently pub. ISBN-13979-8325377037:.
4. Padem, H., A. Göksu, and Z. Konaklı, A*raştırma*

yöntemleri. Spss uygulamalı. 2012, International Burch University. Sarajevo. ISBN: 978-9958-834-04-2.

5. Borman, D., *Statistics 101: from data analysis and predictive modeling to measuring distribution and determining probability, your essential guide to statistics.* 2018, Simon and Schuster Pub. Pp 240. ISBN:1507208189, 9781507208182.
6. Sumpter, D., *Four ways of thinking: statistical, interactive, chaotic and complex.* 2023, Allen Lane pub. Pp. 336. ISBN-10:0241624169.
7. Jung, Y.M., *Data analysis in quantitative research*. 2018, In: Liamputtong, P. (eds) Handbook of research methods in health social sciences. Springer, Singapore. ISBN: 978-981-10-5251-4.
8. Ott, R. and M. Longnecker, *An introduction to statistical methods and data analysis.* 2021, Cengage Learning. 7 th Ed. Pp. 1296. ISBN-10: 0357670620.
9. Mahsin, M.. *Data analysis techniques for quantitative study.* 2022, In: Islam, M.R., Khan, N.A., Baikady, R. (eds) Principles of social research methodology. Springer, Singapore. ISBN: 978-981-19-5441-2.
10. Çimen, M., *Fen ve sağlık bilimleri alanlarında spss uygulamalı veri analizi.* 2015, Palme Yayıncılık, Yayın No: 905, ISBN: 978-605-355-366-3. Sıhhıye, Ankara.
11. Das, K., *A brief review of tests for normality.* American Journal of Theoretical and Applied Statistics, 2016. **5**(1): p.5-12.
12. Hatem, G., J. Zeidan, and M. Goessens, *Normality testing methods and the importance of skewness and kurtosis in statistical analysis*, BAU Journal Science and Technology, 2022. **3**(2): p. 1-7.
13. Huynh, K., *Getting started with spss: an introduction for beginners.* 2024. Independently pub. First Ed. P 174. ISBN-13:979-8878835145.
14. Jung, Y.M., *Data analysis in quantitative research*. 2019, In: Liamputtong, P. (eds) Handbook of research methods in health social sciences. Springer, Singapore. ISBN: 978-981-10-5251-4.
15. Rutheford, A., *Statistics for the rest us: mastering the art of understanding data without math skills (advanced thinking skills).* 2023, Independently pub. Pp 152. ISBN-13:979-8391345831.
16. Frost, J., *Introduction to statistics: an intuitive guide for analyzing data and unlocking discoveries.* 2020, Statistics by jim publishing. Pp 255. ISBN-101735431109:.
17. Theobald, O., *Statistics for Absolute beginners (Secon Edition) (Al data sciences, python&statistics for beginners.* 2020, Independently pub. Pp 157. ISBN-13:979-8654976123.
18. Rumsey, D.J., *Statistics al in on efor dummies.* 2022, For dummies pub.Pp 560. ISBN-101119902568: .
19. Honner, P., *Painless statistics (barron's painless).* 2022, Barrons educational services. Pp 320. ISBN-10: 1506281583.
20. Ural, A. and İ. Kılıç, *Bilimsel araştırma süreci ve spss ile veri analizi.* 2013, 4.Baskı. 296 Sayfa. Detay Yayınları. ISBN:978-975-8326-17-X.
21. Newbold, P., W. Carlson, and B. Thorne, *Statistics for business and economics.* 2012, Eighth Edition. Pearson Education. ISBN10 : 0273767062.
22. Can, A., *Spss ile bilimsel araştırma sürecinde nicel veri analizi.* 2014, 3. Baskı. 396 Sayfa. Pegem Yay. ISBN: 978-605-364-448-4.
23. Ganesan, R. and P.V. Sreinvasalah, *Textbook of statistics.* 2015, First Edition. Write And Print Publications. ISBN-10: 9789384649050.
24. Karagöz, Y., *Biyoistatistik.* 2014, Nobel Yay. No: 1075. 1. Basım. 733 sayfa. ISBN: 978-605-133-979-5.
25. Anonymous., *Digital literacy training. Spss advanced significance testing.* 2019, ANU Library. Pp.20. Available from: https://services.anu.edu.au/files/SPSSAdvancedSignificance Testing.pdf .
26. Rutherford, A. and J.H. Kim, *The art of statistical thinking: detect misinformation, understand the world deeper, and make better decisions. (Advanced Thinking Skills).* 2022, First Edition. ARB Publications. ISBN: 9798358180710.
27. Sheskin, D.J., *Handbook of parametric and non-parametric statistical procedures.* 2011, Fifth Edition. Taylor and Francis Group. 6000 Broken South Parkway NW Suite 300 Boca Raton. FI. 33487-2742. ISBN: 978-1-4398-5801-1.
28. Rezai, A. and S. Jalal, *Investigating the causes of delay and cost overrun in construction industry*. International Advanced Researches and Engineering Journal, 2018. **2**(2): p. 75-79.
29. Mamenko, 0. and S. Potiannyk, *Rank non-parametric correlation analysis of indicators of heavy metal transition from blood to cow's milk to assess its environmental safety.* Scientific Horizons, 2021. **24**(5): p.35-45.
30. Singh, J.P, *Statistical methods in public health*. 2022, In Gupta S.D. (eds) health care system managements. Pp. 85-127. Springer. ISBN: 978-981-19-3076-8.
31. Yuan, I., A.A. Topjian, C.D. Kurth, M.P. Kirschen, C.G. Ward, B. Zhang, and J.L. Mensinger, *Guide to the statistical analysis plan.* Pediatric Anesthesia, 2019. **29**(3): p.1-15.
32. Jebb, A.T., S. Parrigon, and S.E. Woo, *Exploratory data analysis as a foundation of inductive research.* Human Resource Management Review, 2017. **27**(2): p.265-276.
33. Mehta, S., *Statistics topics.* 2014, Kindle Store. Kindle Edition. p. 161. ASIN : B00KVPB8H8.
34. Dean, S. and B. Ilowsky, *Introductory statistics.* 2016, Open stax pub. p. 907. ISBN-10: 1938168208.
35. Hill, T. and P. Lewicki, *Electronic statistics textbook*. 2013, Stat soft inc. p. 800. ISBN-10: 1884233597.
36. Salcedo, J. and K. McCormick, *SPSS statistics workbook for dummies.* 2023, Wiley Pub. p. 336. ISBN: 97813941563 06,1394156308.
37. Salkind, N.J., *Statistics for people who (think they) hate statistics.* 2010, Sage Pub. p. 399. ISBN: 9781412971027.
38. Knap, H., *Introductory statistics using spss*. 2016, Sage Pub. p. 312. ISBN-10: 1506341004.
39. Schmuller, J., *Statistical analysis with R for dummies.* 2017, For dummies pub. First Ed. P 464. ISBN-10:9781119337065.
40. Gentle, E.J., *Theory of statistics*. 2013, Pub. L. No. 94-553, 90 Stat. 2541. p. 917. ISBN:n/a.