Open Access

Analysis of COVID-19 data using neutrosophic Kruskal Wallis H test

Rehan Ahmad Khan Sherwani[1*], Huma Shakeel[1], Wajiha Batool Awan[1], Maham Faheem[1] and Muhammad Aslam[2*] ⊕

Abstract

Background: Kruskal-Wallis H test from the bank of classical statistics tests is a well-known nonparametric alternative to a one-way analysis of variance. The test is extensively used in decision-making problems where one has to compare the equality of several means when the observations are in exact form. The test is helpless when the data is in an interval form and has some indeterminacy.

Methods: The interval-valued data often contain uncertainty and imprecision and often arise from situations that contain vagueness and ambiguity. In this research, a modified form of the Kruskal-Wallis H test has been proposed for indeterminacy data. A comprehensive theoretical methodology with an application and implementation of the test has been proposed in the research.

Results: The proposed test is applied on a Covid-19 data set for application purposes. The study results suggested that the proposed modified Kruskal-Wallis H test is more suitable in interval-valued data situations. The application of this new neutrosophic Kruskal-Wallis test on the Covid-19 data set showed that the proposed test provides more relevant and adequate results. The data representing the daily ICU occupancy by the Covid-19 patients were recorded for both determinate and indeterminate parts. The existing nonparametric Kruskal-Wallis H test under Classical Statistics would have given misleading results. The proposed test showed that at a 1% level of significance, there is a statistically significant difference among the average daily ICU occupancy by corona-positive patients of different age groups.

Conclusions: The findings of the results suggested that our proposed modified form of the Kruskal-Wallis is appropriate in place of the classical form of the test in the presence of the neutrosophic environment.

Keywords: Neutrosophic, Fuzzy logic, Kruskal Wallis H test, Covid-19

Introduction

Hypothesis testing is a scientific process used to investigate the acceptance or rejection of a proposition under consideration. Two approaches are used in statistics to verify a hypothesis: 1) Parametric approach 2) Nonparametric approach. The most important aspect of the parametric approach is the satisfaction of the assumption

about data's normality, and a few tests require the equality of population variances [1]. In most situations, the distributional assumption under a parametric test hardly satisfy, and the use of nonparametric or distribution-free tests is a common practice [2, 3]. However, all such nonparametric tests apply to data containing determined observations. In real life, there are various scenarios where we have non-precise data, and in such cases, the existing hypothesis testing approach based on classical test statistics cannot be implemented. Recent studies have suggested nonparametric tests based on interval-valued data and fuzzy logic [4]. Smarandache

*Correspondence: rehan.stat@pu.edu.pk; aslam_ravian@hotmail.com
[1] College of Statistical and Actuarial Sciences, University of the Punjab Lahore, Lahore, Pakistan
[2] Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah 21551, Saudi Arabia

Sherwani *et al. BMC Med Res Methodol*   (2021) 21:215

Page 2 of 7

[5] generalized the fuzzy logic in the neutrosophic sense by considering the interval-valued data and the measure of indeterminacy or falseness. Smarandache introduced Neutrosophic Statistics as a generalization of classical statistics applied when the data under consideration is in neutrosophic numbers [6]. Smarandache and Khalid [6] verified the efficiency of neutrosophic logic. Several authors have implemented neutrosophic logic for data containing uncertainty and vagueness; see ref [7–11].

Furthermore, several authors have developed statistical tests to analyze fuzzy data; see, for example, refs [12–15]. Also, in fuzzy logic and neutrosophic statistics, several research works have been contributed by introducing decision-making analysis for the data set containing uncertainty and vagueness [16–18]. Recently, Aslam introduced different statistical tests using Neutrosophic Statistics, including the tests of homogeneity of variance for uncertainty environment, the goodness of fit test in the presence of uncertain parameters, and the Kolmogorov-Smirnov tests under uncertainty [19–21].

In 1952, Kruskal and Wallis [12] provided a robust rank-based test for the *k* sample problem as an alternative to the parametric approaches, such as the one-way analysis of variance (ANOVA). Kruskal-Wallis H test has been used for analysis purposes in various manners; for example, see refs [13–18]. In the classical k sample problem, data are determined and do not contain any ambiguity or vagueness. However, in many current scientific studies, the observations are not necessarily relentless, and indeterminate parts quantitatively express the uncertainties in a sample. The existing Kruskal-Wallis H test cannot be used to investigate the data which is measured in the neutrosophic intervals. A detailed literature review has given a shred of clear evidence that no such test is available that can be useful as a nonparametric alternate for ANOVA under an indeterminate environment. The unavailability of a method for the said purpose is a source of motivation for the current research. The goal is to develop a test that compares several sample observations or group(s); the proposed test is easy to apply and understandable. The proposed modified Kruskal-Wallis test results in the interval-valued form and is preferable for data containing vagueness and uncertainty. The objectives of this article are (1) to introduce the modified neutrosophic Kruskal Wallis test; (2) to define the methodology of the neutrosophic Kruskal Wallis test; and (3) to compare the performance of the existing Kruskal Wallis test with the proposed test through an application on Covid-19 data set under Neutrosophy. More information about the application of neutrosophic statistics can be seen in [22–24].

The article is planned as follows. Section 2 presents the computational method for the application of the neutrosophic Kruskal Wallis test. In section 3, the modified Kruskal Wallis test has been demonstrated with an eloquent example of the Covid-19 data set for scrutinizing its efficiency and competence. It is anticipated that the modified nonparametric Kruskal Wallis test will proficiently analyze the data in the presence of uncertainty and vagueness as compared to the existing Kruskal Wallis test under classical statistics. Finally, the results are discussed and generalized with some conclusive remarks.

Computational method of the modified Kruskal Wallis test under uncertainty

In Classical Statistics, nonparametric tests are methods of statistical analysis that do not require a distribution to meet the assumptions necessary to be analyzed. These tests apply to non-normal data sets. Due to this reason, they are sometimes referred to as distribution-free tests. The basic purpose of suggesting the Kruskal Wallis test is to scrutinize that all independent samples containing neutrosophic observations come from neutrosophic populations with equal means implying that the populations under uncertainty are identical. The proposed nonparametric test is applicable for data where the measure of uncertainty or the measure of falseness has been recorded. Suppose $X_N = a_N + b_N I_N$; $X_N \in [X_L, X_U]$ is a neutrosophic number where the first part represents the measure of determinacy and the second part represents the measure of vagueness or uncertainty. For $I_N \in [I_L, I_U] = 0$, the neutrosophic number reduces to a random variable under classical statistics. The neutrosophic variable X_N represents the neutrosophic sample obtained from the population containing imprecise, uncertain, and indeterminate observations; for detail, see ref [5].

Modified Kruskal Wallis H test

Under Classical Statistics, the Kruskal-Wallis H test is used to test the null hypothesis that all k independent samples come from populations having equal means against the alternative hypothesis that at least one population varies. The existing nonparametric test is a generalization of the two-sample Mann-Whitney U test. It is an extremely useful test when the assumptions of normality do not hold, or the population variances are not equal, but helpless for data under uncertainty. The modified Kruskal Wallis test under uncertainty will be applicable under the following assumptions:

1. The data consists of uncertain, imprecise, and indeterminate values.
2. The neutrosophic samples must be random.

3. The two neutrosophic samples must be mutually independent.
4. The test is generally considered robust to ties, but if there are ties in the data set, they shouldn't be concentrated together in one part of the sample.

Suppose we have k_N independent neutrosophic samples of sizes n_{1N}, n_{2N}, ..., n_{kN} $(\sum n_{iN} = n_N)$. Let X_{iN} $(X_{i1N}, X_{i2N}, X_{i3N}, ..., X_{inN})$ represents the neutrosophic observations of the *ith* sample. To perform this test under uncertainty, arrange all the n_N observations containing uncertainty of the k_N samples combined in ascending order of magnitude and assign the ranks to them. In the case of ties, assign the average of the ranks. To distinguish the neutrosophic sample observations, let the letters A_N, B_N, $C_{N, ...}$ represent the sample observations of the first, second, and third neutrosophic samples, respectively. The observations of the neutrosophic samples are replaced with their corresponding ranks. Add these ranks for each sample and denote the sums by R_{1N}, R_{2N}, ..., R_{nk}. Now compute

$$S_{kN}^2 = \sum_{i=1}^{k_N} \frac{R_{iN}^2}{n_{iN}}; R_N \in [R_L, R_U]; n_N \in [n_L, n_U]; k_N \in [k_L, k_U]$$
(1)

and

$$S_{rN}^2 = \sum_{i,j} r_{ijN}^2;$$
(2)

where r_{ijN} is the rank assigned to neutrosophic observation X_{ijN}; $X_{ijN} \in [X_{ijL}, X_{ijU}]$. If there are no ties, then

$$S_{rN}^2 = \frac{n_N(n_N+1)(2n_N+1)}{6}$$
(3)

The modified Kruskal-Wallis statistic H_N; $H_N \in [H_L, H_U]$ is given by

$$H_N = \frac{(n_N - 1)\left(S_{kN}^2 - C_N\right)}{\left(S_{rN}^2 - C_N\right)}$$
(4)

where C_N; $C_N \in [C_L, C_U]$ denotes the appropriate correction term given by

$$C_N = \frac{n_N(n_N+1)^2}{4}$$
(5)

In case of no ties, the neutrosophic statistic H_N becomes

$$H_N = \frac{12S_{kN}^2}{n_N(n_N+1)} - 3(n_N+1)$$
(6)

The neutrosophi form of the proposed test $H_N \in [H_L, H_U]$ can be expressed as follows

$$H_N = H_L + H_U I_{NH}; I_{NH} \in [I_{LH}, I_{UH}]$$
(7)

Note here that the proposed statistic $H_N \in [H_L, H_U]$ is a generalization of the existing test under classical statistics. The first part H_L shows the determined part, $H_U I_{NH}$ denoted the indeterminate part and $I_{NH} \in [I_{LH}, I_{UH}]$ is the measure of indeterminacy/uncertainty. The proposed test reduces to the existing test when $I_{LH} = 0$.

The neutrosophic Kruskal Wallis H_N test is used to test the null hypothesis that all k_N populations have identical distributions. For a large value of the test statistic under uncertainty is rejected. For example, only three samples have five or fewer neutrosophic observations; the significance of this test statistic is determined by using Kruskal and Wallis' Table [19] having critical values for all combinations of sample sizes up to 5,5,5. In case one of the neutrosophic samples contains more than five observations, or there are more than five observations in each neutrosophic sample and the null hypothesis is true, the neutrosophic test statistic H_N follows a chi-square distribution with (k-1) degrees of freedom.

Application of the proposed modified Kruskal Wallis H test

For applying the proposed neutrosophic Kruskal Wallis test, data representing the daily ICU occupancy by Corona-positive patients have been considered, which was recorded specifically from Pakistan. The hypothesis under investigation for the research is to test a statistically significant the difference in the daily ICU occupancy of Covid-19 patients based on their age groups. Neutrosophy or uncertainty is introduced in the data for a better illustration. The neutrosophic Kruskal Wallis test is applied to test the null hypothesis that there is no difference in the daily ICU occupancy of Covid-19 patients among different age groups in Pakistan during December 2020. Daily ICU occupancy of Covid-19 patients aged 55 and above are shown in Fig. 1, Daily ICU occupancy of Covid-19 patients aged 35–55 are shown in Fig. 2 and Daily ICU occupancy of Covid-19 patients aged 35 and below are shown in Fig. 3.

The neutrosophic null and alternate hypotheses for neutrosophic data given in Table 1 are: The average daily ICU occupancy of Covid-19 patients for all three age groups are equal against the alternative hypothesis that the average daily ICU occupancy of Covid-19 patients for at least two of the three age groups are not equal. Table 1 contains data on daily ICU occupancy of Corona-positive patients by three different age groups. By combining and
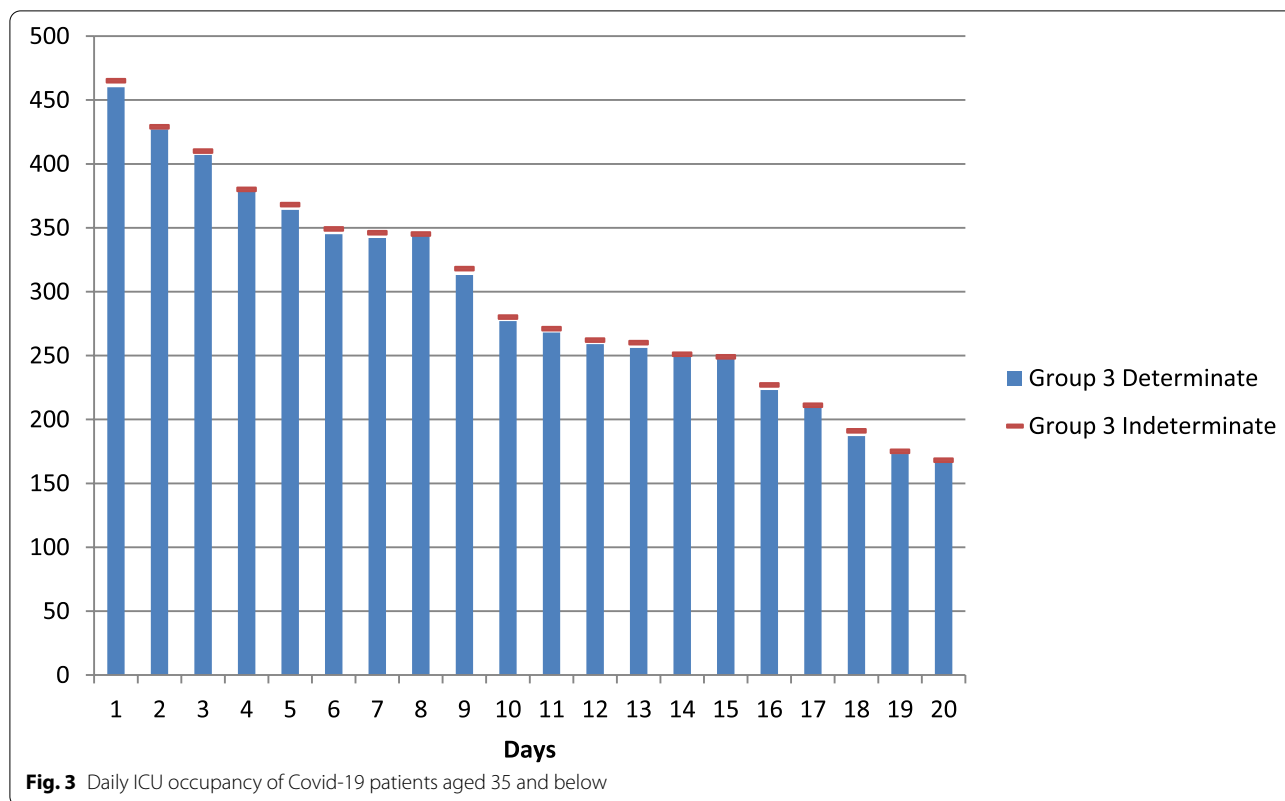
Fig. 1 Daily ICU occupancy of Covid-19 patients aged 55 and above

Fig. 2 Daily ICU occupancy of Covid-19 patients aged 35–55

Sherwani *et al. BMC Med Res Methodol*   (2021) 21:215

Page 5 of 7

Fig. 3 Daily ICU occupancy of Covid-19 patients aged 35 and below
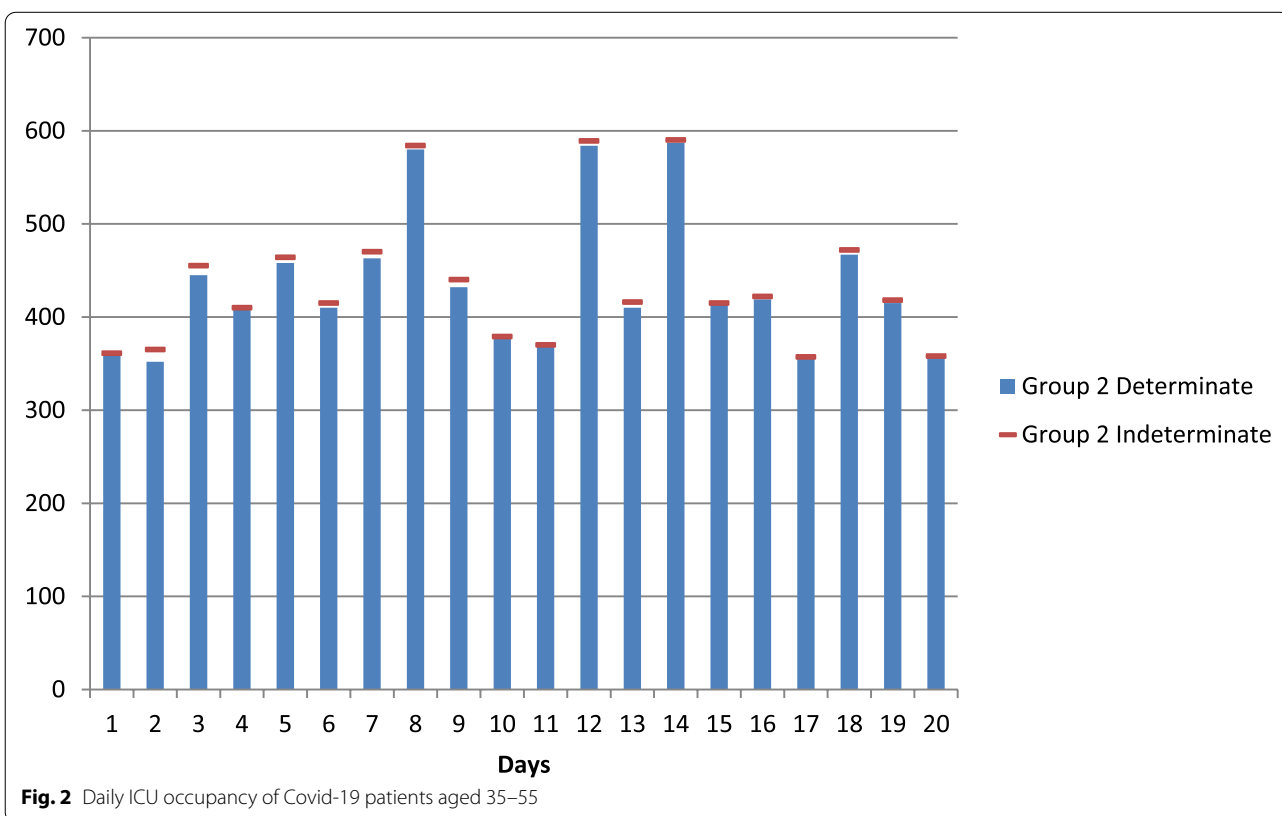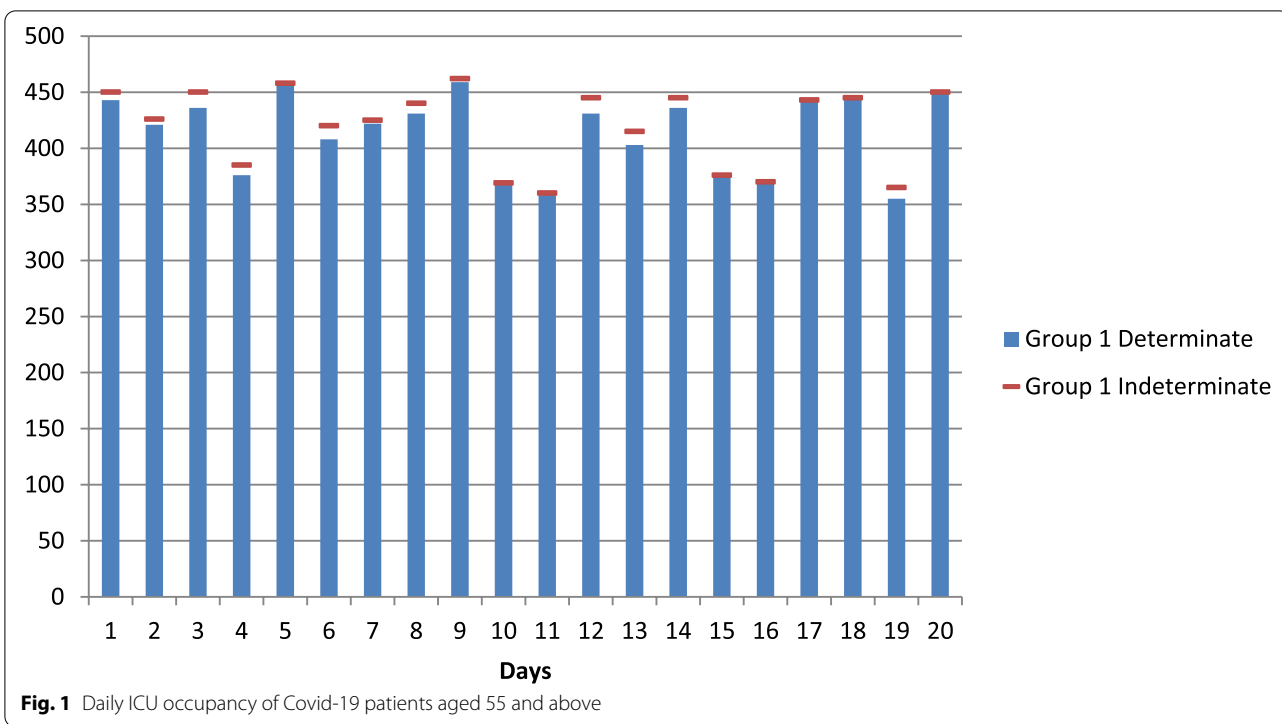
Table 1 Daily ICU occupancy of Covid-19 patients in Pakistan in December 2020

Day/Age	55 and above	35–55	35 and below
1	[443, 450]	[359, 361]	[460, 465]
2	[421, 426]	[352, 365]	[427, 429]
3	[436, 450]	[445, 455]	[407, 410]
4	[376, 385]	410	[378, 380]
5	[458, 458]	[458, 464]	[364, 368]
6	[408, 420]	[410, 415]	[345, 349]
7	[422, 425]	[463, 470]	[342, 346]
8	[431, 440]	[580, 584]	345
9	[459, 462]	[432, 440]	[313, 318]
10	369	379	[277, 280]
11	360	370	[268, 271]
12	[431, 445]	[584, 589]	[259, 262]
13	[403, 415]	[410, 416]	[256, 260]
14	[436, 445]	[587, 590]	251
15	376	415	249
16	370	[419, 422]	[223, 227]
17	443	357	[209, 211]
18	445	[467, 472]	[187, 191]
19	[355, 365]	[415, 418]	[173, 175]
20	450	358	168

Source: Our World in Data COVID-19

arranging the data in ascending order and assigning ranks to them, it is found that ties exist in the neutrosophic data set for both determinate and indeterminate parts. Therefore, the neutrosophic statistic is given in (4) applies to this data set containing the measure of uncertainty. Here $R_{1N} = [757.5, 766.5]$, $R_{2N} = [775.5, 767]$, $R_{3N} = [297, 296.5]$, where R_{1N}, R_{2N} *and* R_{3N} represents the sum of ranks of age groups 1, 2, and 3, respectively.

From (1) and (2), we have

$$S_{kN}^2 = \sum_{i=1}^{k_N} \frac{R_{iN}^2}{n_{iN}} = [63170.78, 63186.18]$$

and

$$S_{rN}^2 = \sum_{i,j} r_{ijN}^2 = [73803.5, 73802]$$

From (4)

$$H_N = \frac{(n_N - 1)\left(S_{kN}^2 - C_N\right)}{\left(S_{rN}^2 - C_N\right)} = [24.12, 24.17]$$

$$p_N - value = [0.000011, 0.000011]$$

Sherwani *et al. BMC Med Res Methodol*   (2021) 21:215

Page 6 of 7

Assuming the level of significance to be 1%, the critical region is $H_N > \chi^2_{0.01,2} = 9.21$. Since the calculated value of test statistic based on neutrosophic observations lies in the critical region (*p*-value < α), we, therefore, reject the neutrosophic null hypothesis and conclude that daily ICU occupancy of different age groups of Covid-19 patients is not equal.

Advantages of the proposed test

In this section, the efficiency of the proposed test $H_N \in [H_L, H_U]$ will be compared with the existing test under classical statistics in terms of a measure of uncertainty. As mentioned earlier, the neutrosophic $H_N = H_L + H_U I_{NH}$; $I_{NH} \in [I_{LH}, I_{UH}]$ has consisted of determinate (the existing test) and indeterminate parts. The neutrosophic form of $H_N \in [H_L, H_U]$ for the real data is expressed as: $H_N = 24.12 + 24.17 I_{NH}$; $I_{NH} \in [0, 0.002]$; where the first value 24.12 shows the results of the existing test when $I_{LH} = 0$ and $24.17 I_{NH}$ is an indeterminate part. Note that the measure of indeterminacy associated with the test $H_N \in [H_L, H_U]$ is 0.002. From the study, it can be seen that the proposed test the result of the test statistic in the range of 24.12 to 24.17. On the other hand, the existing test provides only the determined/exact value of the test. In addition, the proposed test $H_N \in [H_L, H_U]$ gives information about the measure of uncertainty. Based on the information, the proposed test can be interpreted as follows: when the level of significance α =0.05, the chance of rejecting the null hypothesis when it is true is 0.05, the probability of accepting the null hypothesis is 0.95 with the chance of uncertainty of 0.002. From the comparisons, it can be concluded that the proposed test $H_N \in [H_L, H_U]$ gives more information about the test. In addition, the proposed test is flexible, adequate, and effective to be applied in uncertainty as compared to the existing test.

Conclusion and discussion

This article proposed the modified form of the rank-based nonparametric Kruskal Wallis H test for observations containing the measure of uncertainty or the measurement of falseness when comparing *k* samples. It is evident from Table 1 that uncertain data used for the illustration purpose reduces to the determined part under classical statistics if no observations of uncertainty are logged. For example, for sample one, the first observation 443 for the first group represents the determinate part of the indeterminate interval. The second value, which is 450, represents the indeterminate part of the interval. We can observe here that the modified Kruskal-Wallis test results in the indeterminacy interval rather than the determined values, and this implies that the proposed test provides

a good measure of uncertainty. Recent studies also show that the methods dealing with the interval-valued data are more suitable in the indeterminate environment than classical statistical techniques [25, 26]. The work was originally motivated by the extensive research work under the fuzzy logic and neutrosophic statistics used for the interval-valued data set. The proposed nonparametric test can be readily applied to compare k samples testing the hypothesis that they have equal means.

The application of this new neutrosophic Kruskal-Wallis test on the Covid-19 data set showed that the proposed test provides more relevant and adequate results. The data representing the daily ICU occupancy by the Covid-19 patients were recorded for both determinate and indeterminate parts. The existing nonparametric Kruskal Wallis H test under Classical Statistics would have given misleading results. The proposed test showed that at a 1% level of significance, there is a statistically significant difference among the average daily ICU occupancy by corona-positive patients of different age groups.

The modified Kruskal Wallis test can be used to compare the averages of several sample observations or group(s); the proposed test is easy to apply and understandable. The Neutrosophic Kruskal Wallis test results in an uncertain interval, which is ideal when the data is measured from the complex system. The application of the proposed test is recommended for different fields, including biomedical sciences, engineering, and many other statistical areas. On the other hand, applying nonparametric tests under classical statistics on the data containing vagueness can produce misleading results. In conclusion, the proposed neutrosophic nonparametric test provides an efficient tool to data analysts for analyzing k samples in the presence of uncertainty and indeterminacy. However, more properties of this modified Kruskal-Wallis test can be derived for future research. The evaluation of the proposed test using different measures can be studied as future research.

Sherwani *et al. BMC Med Res Methodol*   (2021) 21:215

Page 7 of 7

Declarations

Ethics approval and consent to participate
N/A

Consent for publication
N/A

Competing interests
none.

References

1. Higgins JJ. An introduction to modern nonparametric statistics. CA: Brooks/Cole Pacific Grove; 2004.
2. Krzywinski M, Altman N. Nonparametric tests. Nat Methods. 2014;11:467–8. https://doi.org/10.1038/nmeth.2937.
3. Chan Y. Biostatistics 102: quantitative data–parametric & nonparametric tests. Blood Press. 2003;140:79.
4. Buckley JJ. Fuzzy statistics: hypothesis testing. Soft Comput. 2005;9:512–8. https://doi.org/10.1007/s00500-004-0368-5.
5. Smarandache F. Neutrosophic Logic-A Generalization of the Intuitionistic Fuzzy Logic. Multispace & Multistructure Neutrosophic Transdisciplinarity (100 Collected Papers of Science). 2010;4:396.
6. Smarandache, F., Khalid, H. E. & Essa, A. K. Neutrosophic Logic: the Revolutionary Logic in Science and Philosophy. (Infinite Study, 2018).
7. Nabeeh NA, Abdel-Basset M, El-Ghareeb HA, Aboelfetouh A. Neutrosophic multi-criteria decision making approach for iot-based enterprises. IEEE Access. 2019;7:59559–74.
8. Abdel-Basset M, Nabeeh NA, El-Ghareeb HA, Aboelfetouh A. Utilising neutrosophic theory to solve transition difficulties of IoT-based enterprises. Enterprise Inform Syst. 2020;14:1304–24.
9. Abdel-Baset M, Chang V, Gamal A. Evaluation of the green supply chain management practices: a novel neutrosophic approach. Comput Ind. 2019;108:210–20.
10. Abdel-Basset M, Atef A, Smarandache F. A hybrid Neutrosophic multiple criteria group decision making approach for project selection. Cogn Syst Res. 2019;57:216–27.
11. Broumi, S., Bakali, A., Talea, M. & Smarandache, F. Bipolar neutrosophic minimum spanning tree. (Infinite Study, 2018).
12. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47:583. https://doi.org/10.2307/2280779.
13. McKight, P. E. & Najab, J. Kruskal-wallis test. The corsini encyclopedia of psychology, 1–1 (2010).
14. Hecke TV. Power study of anova versus Kruskal-Wallis test. J Stat Manage Syst. 2012;15:241–7.
15. MacFarland, T. W. & Yates, J. M. in Introduction to nonparametric statistics for the biological sciences using R 177–211 (Springer, 2016).
16. Soltani N, Safajou F, Amouzeshi Z, Zameni E. The relationship between body image and mental health of students in Birjand in 2016 academic year: a short report. J Rafsanjan Univ Med Sci. 2017;16:479–86.
17. Lou, Y., Yuen, S. Y. & Chen, G. in Proceedings of the Genetic and Evolutionary Computation Conference Companion. 1337–1341.
18. Muremi, L. & Bokoro, P. in 2018 IEEE International Conference on Environment and Electrical Engineering and 2018 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I&CPS Europe). 1–4 (IEEE).
19. Aslam M. A new goodness of fit test in the presence of uncertain parameters. Complex Intell Syst. 2021;7:359–65. https://doi.org/10.1007/s40747-020-00214-8.
20. Aslam M. Introducing Kolmogorov–Smirnov tests under uncertainty: an application to radioactive data. ACS Omega. 2019;5(1):914–7.
21. Aslam M. Design of the Bartlett and Hartley tests for homogeneity of variances under indeterminacy environment. J Taibah Univ Sci. 2020;14(1):6–10.
22. Chen J, Ye J, Du S. Scale effect and anisotropy analyzed for neutrosophic numbers of rock joint roughness coefficient based on neutrosophic statistics. Symmetry. 2017;9:208.
23. Aslam M. Neutrosophic analysis of variance: application to university students. Complex Intelligent Syst. 2019;5(4):403–7.
24. Smarandache F. Introduction to neutrosophic statistics: infinite study. Columbus, OH, USA: Romania-Educational Publisher; 2014.
25. Meyer JP, Seaman MA. A comparison of the exact Kruskal-Wallis distribution to asymptotic approximations for all sample sizes up to 105. J Exp Educ. 2013;81:139–56.
26. Chen J, Ye J, Du S, Yong R. Expressions of rock joint roughness coefficient using neutrosophic interval statistical numbers. Symmetry. 2017;9:123.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.