frontiers

Analysis and Allocation of Cancer-Related Genes using Vague DNA Sequence Data

Muhammad Aslam[1*], Mohammad Albassam[1]

[1]King Abdulaziz University, Saudi Arabia

Submitted to Journal:
Frontiers in Genetics

Specialty Section:
Statistical Genetics and Methodology

Article type:
Original Research Article

Manuscript ID:
858005

Received on:
19 Jan 2022

Revised on:
16 Feb 2022

Journal website link:
www.frontiersin.org

Conflict of interest statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest

Author contribution statement

M.A and M.A.B wrote the paper.

Keywords

Multinomial distribution, chi-square test, Classical statistics, Neutrosophy, DNA data

Abstract

Word count: 113

To test the equality of several independent multinomial distributions, the Chi-square test for count data is applied. The existing test can be applied when complete information about the data is available. The complex process such as DNA count, the existing test under classical statistics may mislead. To overcome the issue, the modification of the Chi-square test for multinomial distribution under neutrosophic statistics is presented in this paper. The modified form of Chi-square test statistic under indeterminacy/uncertainty will be presented and applied using the DNA count data. From the DNA count data analysis, simulation and comparative studies, the proposed test is found to be informative, springy, and goodish as compared to the existing tests.

Contribution to the field

The Chi-square test for multinomial distribution available in the literature can be applied when full information about the data is given. In complex processes or processes under uncertainty do not possess the full information about the data or level of significance. Therefore, there is a gap in to design of the Chi-square test for multinomial distribution under neutrosophic statistics. Therefore, in this study, the Chi-square test for multinomial distribution using neutrosophic statistics will be introduced the first time according to the best of the author's knowledge. The application of the proposed test will be given with the aid of DNA cancer data. It is expected the proposed test will be competent than the existing tests in terms of springy, deftness and goodish.

Ethics statements

Studies involving animal subjects

Generated Statement: No animal studies are presented in this manuscript.

Studies involving human subjects

Generated Statement: No human studies are presented in this manuscript.

Inclusion of identifiable human data

Generated Statement: No potentially identifiable human images or data is presented in this study.

Data availability statement

Generated Statement: The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Analysis and Allocation of Cancer-Related Genes using Vague DNA Sequence Data

Muhammad Aslam[1*] and Mohammad Albassam[1]

Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah 21551, Saudi Arabia; Email: aslam_ravian@hotmail.com; malbassam@kau.edu.sa

Abstract

To test the equality of several independent multinomial distributions, the Chi-square test for count data is applied. The existing test can be applied when complete information about the data is available. The complex process such as DNA count, the existing test under classical statistics may mislead. To overcome the issue, the modification of the Chi-square test for multinomial distribution under neutrosophic statistics is presented in this paper. The modified form of Chi-square test statistic under indeterminacy/uncertainty will be presented and applied using the DNA count data. From the DNA count data analysis, simulation, and comparative studies, the proposed test is found to be informative, springy, and goodish as compared to the existing tests.

Keywords: Multinomial distribution; Chi-square test; classical statistics; neutrosophy; DNA data

1 Introduction

Without the statistical analysis, it is not possible to check the significance of the variables under study. For testing the significance of the variable, statistical tests have been applied in a variety of fields, (Ali & Bhaskar, 2016) and (Greenland et al., 2016). The Chi-square test for multinomial distribution is applied for testing whether the allocation of objects to different groups is equally likely or not. This test is applied for testing the null hypothesis that allocation of objects to different groups is equal vs. the alternative hypothesis that allocation of objects to different groups is unequal. The test statistic is computed from the data and the null hypothesis is accepted if the values of the statistic fall within the acceptance region. (Cohen, Kolassa, & Sackrowitz, 2006) used the test for equality of multinomial distributions. (Chafai & Concordet, 2009) studied confidence interval for multinomial distribution in the case of small samples. (Turner, Deng, & Houle, 2020) used the statistical tests for head and face data. (Shin, Yamamoto, Brady, Lee, & Haynes, 2019) and (Mollan et al., 2019) discussed the applications of the statistical tests.

The statistical methods have been widely used in analyzing and testing the significance of DNA data. A rich literature of statistical methods analyzing the DNA data is available. (Goldman, 1993a) applied the statistical tests using DNA data. (Buldyrev et al., 1998) and (Kugiumtzis & Provata, 2004) analyzed the DNA data using statistical physics. (Yoshida, Kobayashi, Futagami, & Fujikoshi, 1999) used the statistical analysis for the DNA data. (Pai, Mathew, & Anindya, 2021) worked on the prediction using the DNA data. (Yao, Jin, & Lee, 2018) improved the statistical analysis for genetic data. (Gunasekaran et al., 2021) analyzed the DNA data using hybrid models. (Halla-aho & Lähdesmäki, 2021) used statistical analysis for DNA cancer data. More applications of

the statistical techniques for DNA data can be seen in (Goldman, 1993b), (Keinduangjun, Piamsa-nga, & Poovorawan, 2005) (Rodriguez et al., 2012) and (Pai et al., 2021).

Fuzzy-based statistical tests are applied when the data in hand has vague or incomplete information. (Viertl, 2006) mentioned that "statistical data are frequently not precise numbers but more or less non-precise also called fuzzy. Measurements of continuous variables are always fuzzy to a certain degree". Several studies using the fuzzy-based multinomial distribution are available in the literature. (Amirzadeh, Mashinchi, & Yaghoobi, 2008) studied the multinomial distribution using fuzzy logic. (Mashuri & Ahsan, 2018) worked on a fuzz-based chart using multinomial distribution. More information for fuzzy-based multinomial distribution can be seen in (Amirzadeh et al., 2008) and (Hrafnkelsson, Oddsson, & Unnthorsson, 2016).

(Florentin Smarandache, 2013) discussed that the neutrosophic logic is more efficient than interval-based analysis and fuzzy-based analysis. Neutrosophic statistics is applied to analyze the data having neutrosophic numbers, see (F Smarandache, 2014). The interval statistics used the interval data to capture the data in the interval only and silent about the measure of indeterminacy. On the other hand, the fuzzy-based analysis only gives information about the measure of truth and the measure of falseness. Neutrosophic statistics become classical statistics when no indeterminate information is found in the data. (Chen, Ye, & Du, 2017) and (Chen, Ye, Du, & Yong, 2017) introduced the methods to deal with the neutrosophic data. Later on, (Sherwani et al., 2021), (Aslam, 2021) and (Albassam, Khan, & Aslam, 2021) introduced statistical tests under neutrosophic statistics.

The Chi-square test for multinomial distribution available in the literature can be applied when full information about the data is given. Complex processes or processes under uncertainty do not possess the full information about the data or level of significance. Therefore, there is a gap into the design of the Chi-square test for multinomial distribution under neutrosophic statistics. Therefore, in this study, the Chi-square test for multinomial distribution using neutrosophic statistics will be introduced the first time according to the best of the author's knowledge. The application of the proposed test will be given with the aid of DNA cancer data. It is expected the proposed test will be more competent than the existing tests in terms of springy, deftness, and goodish.

2 Method

The existing test for testing the equality of multinomial distribution can only be utilized when no vague information is presented. To overcome this issue, the modification of the existing test is necessary. In this section, the modification of the existing test under classical statistics will be presented under the neutrosophic statistics. With the expectation that the proposed test for testing the equality of multinomial distribution will perform better for testing the null hypothesis under an uncertain environment. The main objective of the paper is to introduce the test for the equality of h_N independent neutrosophic multinomial distributions. Let $Y_{1jN}, Y_{2jN} \dots Y_{kjN} (j = 1,2, \dots, h_N)$ present the neutrosophic frequencies for the neutrosophic events $A_{1N}, A_{2N} \dots A_{kN}$. Let $p_{ijN} = P(A_{iN});\ i_N = 1,2, \dots, k_N; j_N = 1,2, \dots, h_N$. The neutrosophic form of $p_{ijN} \epsilon \left[p_{ijL}, p_{ijU}\right]$ is expressed as

$$p_{ijN} = p_{ijL} + p_{ijU} I_{p_{ijN}}; I_{p_{ijN}} \epsilon \left[I_{p_{ijL}}, I_{p_{ijU}}\right] \quad (1)$$

where p_{ijL} presents the determined part, and $p_{ijU}I_{p_{ijN}}$ presents the indeterminate part and $I_{p_{ijN}}\epsilon\left[I_{p_{ijL}}, I_{p_{ijU}}\right]$is the measure of indeterminacy. The alternative expression of Eq. (1) can be given as

$$p_{ijN} = \left(1 + I_{p_{ijN}}\right)p_{ij}; I_{p_{ijN}}\epsilon\left[I_{p_{ijL}}, I_{p_{ijU}}\right] \quad (2)$$

The j_{th} experiment is carried out n_{jN} times under the assumption that n_{jN} instances are independent. The modified form of the test statistic $Q_N\epsilon[Q_L, Q_U]$ is expressed as follows

$$Q_N = Q_L + Q_U I_{Q_N}; I_{Q_N}\epsilon\left[I_{Q_L}, I_{Q_U}\right] \quad (3)$$

where

$$Q_N = \sum_{j=1}^{h_N}\sum_{i=1}^{k_N}\frac{\left(Y_{ijN} - n_{jN}p_{ijN}\right)^2}{n_{jN}p_{ijN}}$$

The proposed statistic $Q_N\epsilon[Q_L, Q_U]$ can be written as

$$Q_N = \sum_{j=1}^{h_L}\sum_{i=1}^{k_L}\frac{\left(Y_{ijL} - n_{jL}p_{ijL}\right)^2}{n_{jL}p_{ijL}} + \sum_{j=1}^{h_U}\sum_{i=1}^{k_U}\frac{\left(Y_{ijU} - n_{jU}p_{ijU}\right)^2}{n_{jU}p_{ijU}}I_{Q_N}; I_{Q_N}\epsilon\left[I_{Q_L}, I_{Q_U}\right] \quad (4)$$

The simplified form of statistic can be written as

$$Q_N = \left(1 + I_{Q_N}\right)\sum_{j=1}^{h_N}\sum_{i=1}^{k_N}\frac{\left(Y_{ijN} - n_{jN}p_{ijN}\right)^2}{n_{jN}p_{ijN}}; I_{Q_N}\epsilon\left[I_{Q_L}, I_{Q_U}\right] \quad (5)$$

Note that the proposed test $Q_N\epsilon[Q_L, Q_U]$ is a generalization of the test under classical statistics. The proposed test $Q_N\epsilon[Q_L, Q_U]$ reduces to the classical test under classical statistics when I_{Q_L}=0. The proposed test is also a generalization of the tests under interval-statistics and fuzzy-based logic. The proposed test $Q_N\epsilon[Q_L, Q_U]$ follows the

neutrosophic Chi-square distribution with $h_N(k_N - 1)$ degree of freedom. The proposed test $Q_N \epsilon [Q_L, Q_U]$ will be applied to test the following null hypothesis

$$H_{0N}: p_{i1} = p_{i2} = \cdots = p_{ih_N} = p_{iN}, \;\; i = 1,2,3, \ldots, k_N \tag{6}$$

Under the null hypothesis, we estimate $k_N - 1$ probabilities from

$$\hat{p}_{iN} = \frac{\sum_{j=1}^{h_L} Y_{ijL}}{\sum_{j=1}^{h_L} n_{jL}} + \frac{\sum_{j=1}^{h_U} Y_{ijU}}{\sum_{j=1}^{h_U} n_{jU}} I_{\hat{p}_{iN}}; I_{\hat{p}_{iN}} \epsilon \left[I_{\hat{p}_{iL}}, I_{\hat{p}_{iU}} \right] \tag{7}$$

The statistic $Q_N \epsilon [Q_L, Q_U]$ based on $\hat{p}_{iN} \epsilon [\hat{p}_{iL}, \hat{p}_{iU}]$ is expressed as

$$Q_N = \sum_{j=1}^{h_L} \sum_{i=1}^{k_L} \frac{\left(Y_{ijL} - n_{jL}\hat{p}_{ijL}\right)^2}{n_{jL}\hat{p}_{ijL}} + \sum_{j=1}^{h_U} \sum_{i=1}^{k_U} \frac{\left(Y_{ijU} - n_{jU}\hat{p}_{ijU}\right)^2}{n_{jU}\hat{p}_{ijU}} I_{Q_N}; I_{Q_N} \epsilon \left[I_{Q_L}, I_{Q_U} \right] \tag{8}$$

The simplified form of statistic can be written as

$$Q_N = \left(1 + I_{Q_N}\right) \sum_{j=1}^{h_N} \sum_{i=1}^{k_N} \frac{\left(Y_{ijN} - n_{jN}\hat{p}_{ijN}\right)^2}{n_{jN}\hat{p}_{ijN}}; I_{Q_N} \epsilon \left[I_{Q_L}, I_{Q_U} \right] \tag{9}$$

Note that $Q_N \epsilon [Q_L, Q_U]$ based on $\hat{p}_{iN} \epsilon [\hat{p}_{iL}, \hat{p}_{iU}]$ follows the neutrosophic Chi-square distribution with $(h_N - 1)(k_N - 1)$ degree of freedom.

3 Application

In this section, the application of the proposed test will be given using the DNA sequence data. The data is related to the cancer-related gene BRCA 2. According to https://medlineplus.gov/genetics/gene/brca2/#:~:text=Mutations%20in%20the%20BRCA2%20gene,one%20generation%20to%20the%20next "Mutations in the BRCA2 gene are associated with an increased risk of breast cancer in both men and women, as well as several other types of cancer. These mutations are present in every cell in the body and can be passed from one generation to the next". By following

https://www.math.mcgill.ca/~dstephens/OldCourses/204-2007/Handouts/Math204-ChiSquareWithResults.pdf, the counts of nucleotide (A, C, G, T) having two counting groups are reported in Table 1. Note here that in Table 1, the data is given in the "Count Group 1" is selected from the given reference, and the data is given in the "Count Group 2" is generated by simulation. The DNA sequence is a complex process and there may be uncertainty/indeterminacy in counts, see (Yurov, Vorsanova, & Iourov, 2011). In the presence of uncertainty/indeterminacy in counts, the proposed test can be applied more effectively than the existing test under classical statistics. Suppose that there is 5% uncertainty/indeterminacy in counts of the numbers of nucleotides (A, C, G, T) in the DNA sequence of the cancer-related gene BRCA 2. Based on the information and data is given in Table 1, the proposed test statistic will be calculated as follows

$\sum_{j=1}^{4}\sum_{i=1}^{4}\frac{(Y_{ijL}-n_{jL}\hat{p}_{ijL})^2}{n_{jL}\hat{p}_{ijL}}$ =0.000365921+0.002051303+…+0.000748132=0.00664

The statistic $Q_N\epsilon[Q_L, Q_U]$ in neutrosophic form can be expressed as follows

$$Q_N = 0.00664 + 0.00664I_{Q_N}; I_{Q_N}\epsilon[0,0.05]$$

The simplified form of statistic can be written as

$$Q_N = (1 + 0.05)0.00664 = 0.00697; I_{Q_N}\epsilon[0,0.05]$$

The proposed test DNA count data will be implemented in the following steps

Step-1: State the null hypothesis H_0: The allocation of DNA count is equally likely vs. the alternative hypothesis H_1: The allocation of DNA count is unequal.

Step-2: The level of significance α=0.05 and the tabulated value from (Kanji, 2006) is 9.35.

Step-3: Compute the value of statistic Q_N=0.00697 and compare it with the tabulated value.

Step-4: As the computed value of Q_N is less than 9.35, therefore H_0 will be accepted.

Based on the analysis, it can be concluded that there is no evidence to suspect unequal allocation of counts of nucleotide (A, C, G, T).

Table 1 is around here

4 Simulation Study

A simulation study is performed to assess the effect of indeterminacy I_{Q_N} in counts of the numbers of nucleotides (A, C, G, T) in the DNA sequence of the cancer-related gene BRCA 2 on the statistic Q_N. To see the effect of I_{Q_N} on the statistic Q_N, various values of I_{Q_N} are considered. Using the neutrosophic form obtained for the DNA count data, the values of statistic Q_N are shown in Table 2. From Table 2, it can be noted that as the values indeterminacy I_{Q_N} increases, the values of Q_N also increase. The decision about H_0 at various values of I_{Q_N} is also shown in Table 2. From Table 2, although, the values of statistic Q_N increase as I_{Q_N} increases but it does not change the decision about the acceptance H_0.

Table 2 is around here

5 Comparative Studies

The springy, deftness and goodish of the proposed test over the tests under interval-statistic, fuzzy-based approach and classical statistics is shown in this section. The efficiency of the proposed test will be shown in terms of the measure of indeterminacy, springy, deftness, and goodish. The neutrosophic form of statistic $Q_N \epsilon [Q_L, Q_U]$ is expressed as follows

$$Q_N = 0.00664 + 0.00664 I_{Q_N}; I_{Q_N} \epsilon [0,0.05]$$

The above-mentioned neutrosophic form is based on two types of information. The first part 0.00664 gives information about the determinate part and the second part $0.00664 I_{Q_N}$ gives the information about the indeterminate part. The proposed statistic $Q_N \epsilon [Q_L, Q_U]$ reduces to the test under classical statistics when I_{Q_L}=0. Therefore, it can be analyzed that the existing test under classical statistics gives only information about the determinate part. On the other hand, the proposed test gives information about the indeterminacy additionally as compared to the test using classical statistics. Therefore, the proposed test is more bendable than the existing test under classical statistics. The interval-statistics only utilizes the information is given in the interval. In simple words, the interval-statistics captured the information between intervals. Now comparing the results of the proposed test under test statistic under interval-statistics, it can be seen that the proposed test is more explanatory than the test using interval-statistics as the earlier did not give any information about the measure of indeterminacy. Therefore, the proposed test is also more efficient than the test using the interval-based statistic. The test

statistic using fuzzy logic can be considered measures of truth and falseness. The neutrosophic statistics uses the set analysis and can be used for any type of set. The proposed statistic $Q_N \epsilon [Q_L, Q_U]$ gives three types of information. The proposed test states that the chance of accepting H_0 is 0.95 (a measure of truth), the chance of committing a type-I error is 0.05 (a measure of falseness) and the measure of indeterminacy associated with the test is 0.05. From the study, it is concluded that the proposed test is also a generalization of the test using fuzzy-logic. Therefore, the proposed test is more informative than the three existing tests.

6 Concluding Remarks

The modification of the existing test for testing the equality of multinomial distribution under neutrosophic statistics was introduced in the paper. The proposed test was the generalizations of several existing tests under interval-statistics, fuzzy-based and classical statistics. The modification of the test statistic was presented in the presence of indeterminacy. The simulation study and comparative study have shown that the proposed test was adequate and effective to apply in the presence of uncertainty. The application of the proposed test for DAN count data also showed its efficiency. The proposed test can be applied for testing allocation of count is equally likely or not in medical science, engineering and political science. More properties of the proposed test can be studied in future research. The proposed test using a double sampling scheme is another fruitful area for future research.

Acknowledgements: The authors are deeply thankful to the editor and reviewers for their valuable comments to improve the quality of the paper.

Funding: The paper was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia. The authors, therefore thanks DSR for their financial and technical support.

References

Albassam, M., Khan, N., & Aslam, M. (2021). Neutrosophic D'Agostino test of normality: An application to water data. *Journal of Mathematics, 2021*, DOI: https://doi.org/10.1155/2021/5582102

Ali, Z., & Bhaskar, S. B. (2016). Basic statistical tools in research and data analysis. *Indian Journal of Anaesthesia, 60*(9), 662.

Amirzadeh, V., Mashinchi, M., & Yaghoobi, M. (2008). Construction of control charts using fuzzy multinomial quality. *Journal of Mathematics and Statistics*, *4*(1), 26.

Aslam, M. (2021). Neutrosophic statistical test for counts in climatology. *Scientific Reports*, *11*(1), 1-5.

Buldyrev, S. V., Dokholyan, N. V., Goldberger, A. L., Havlin, S., Peng, C. K., Stanley, H. E., & Viswanathan, G. M. (1998). Analysis of DNA sequences using methods of statistical physics. *Physica A: Statistical Mechanics and its Applications*, *249*(1-4), 430-438.

Chafai, D., & Concordet, D. (2009). Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association, 104*(487), 1071-1079.

Chen, J., Ye, J., & Du, S. (2017). Scale effect and anisotropy analyzed for neutrosophic numbers of rock joint roughness coefficient based on neutrosophic statistics. *Symmetry*, *9*(10), 208.

Chen, J., Ye, J., Du, S., & Yong, R. (2017). Expressions of rock joint roughness coefficient using neutrosophic interval statistical numbers. *Symmetry*, *9*(7), 123.

Cohen, A., Kolassa, J., & Sackrowitz, H. (2006). A test for equality of multinomial distributions vs increasing convex order *Recent Developments in Nonparametric Inference and Probability* (pp. 156-163): Institute of Mathematical Statistics.

Goldman, N. (1993a). Simple diagnostic statistical tests of models for DNA substitution. *Journal of Molecular Evolution, 37*(6), 650-661.

Goldman, N. (1993b). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution*, *36*(2), 182-198.

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, *31*(4), 337-350.

Gunasekaran, H., Ramalakshmi, K., Rex Macedo Arokiaraj, A., Deepa Kanmani, S., Venkatesan, C., Suresh Gnana Dhas, C. J. C., & Medicine, M. M. i. (2021). Analysis of DNA Sequence Classification Using CNN and Hybrid Models. *2021*.

Halla-aho, V., & Lähdesmäki, H. J. b. (2021). Probabilistic modeling methods for cell-free DNA methylation based cancer classification (bioRxiv preprint). https://doi.org/10.1101/2021.06.18.444402

Hrafnkelsson, B., Oddsson, G. V., & Unnthorsson, R. J. E. (2016). A method for estimating annual energy production using Monte Carlo wind speed simulation. *Energies*, *9*(4), 286.

Kanji, G. K. (2006). *100 statistical tests*: Sage.

Keinduangjun, J., Piamsa-nga, P., & Poovorawan, Y. (2005). *DNA sequence identification by statistics-based models.* Paper presented at the International Conference on Fuzzy Systems and Knowledge Discovery.

Kugiumtzis, D., & Provata, A. (2004). Statistical analysis of gene and intergenic DNA sequences. *Physica A: Statistical Mechanics and its Applications*, *342*(3-4), 623-638.

Mashuri, M., & Ahsan, M. (2018). *Perfomance fuzzy multinomial control chart.* Paper presented at the Journal of Physics: Conference Series.

Mollan, K. R., Trumble, I. M., Reifeis, S. A., Ferrer, O., Bay, C. P., Baldoni, P. L., & Hudgens, M. G. (2019). Exact Power of the Rank-Sum Test for a Continuous Variable. *arXiv preprint arXiv:1901.04597*. https://arxiv.org/abs/1901.04597

Pai, S. S., Mathew, A. R., & Anindya, R. J. b. (2021). A comparative analysis of computational tools for the prediction of epigenetic DNA methylation from long-read sequencing data, DOI**:** https://doi.org/10.1101/2021.04.24.441281

Rodriguez, B. A., Frankhouser, D., Murphy, M., Trimarchi, M., Tam, H.-H., Curfman, J., . . . Parikh, D. J. B. g. (2012). Methods for high-throughput MethylCap-Seq data analysis. *BMC Genomics*, *13*(6), 1-11.

Sherwani, R. A. K., Shakeel, H., Saleem, M., Awan, W. B., Aslam, M., & Farooq, M. (2021). A new neutrosophic sign test: An application to COVID-19 data. *PloS one*, *16*(8), e0255671.

Shin, D., Yamamoto, Y., Brady, M. P., Lee, S., & Haynes, J. A. (2019). Modern data analytics approach to predict creep of high-temperature alloys. *Acta Materialia*, *168*, 321-330.

Smarandache, F. (2013). *Introduction to neutrosophic measure, neutrosophic integral, and neutrosophic probability*: Sitech – Education.

Smarandache, F. (2014). Introduction to Neutrosophic Statistics, Sitech and Education Publisher, Craiova. *Romania-Educational Publisher, Columbus, Ohio, USA*, *123*.

Turner, D. P., Deng, H., & Houle, T. T. (2020). Statistical hypothesis testing: Overview and application. *Headache: The Journal of Head and Face Pain*, *60*(2), 302-308.

Viertl, R. (2006). Univariate statistical analysis with fuzzy data. *Computational Statistics & Data Analysis*, *51*(1), 133-147.

Yao, Y., Jin, Z., & Lee, J. H. (2018). An improved statistical model for taxonomic assignment of metagenomics. *BMC genetics*, *19*(1), 1-11.

Yoshida, K., Kobayashi, M., Futagami, K., & Fujikoshi, Y. (1999). *Statistical analysis of DNA sequencing data (1): Accuracy test of DNA data by partial re-sequencing.* Paper presented at the Nucleic acids symposium series.

Yurov, Y. B., Vorsanova, S. G., & Iourov, I. Y. (2011). The DNA replication stress hypothesis of Alzheimer's disease. *The Scientific World Journal*, *11*, 2602-2612.

Table 1: The counts of nucleotide data

Category	1	2	3	4	Total
Nucleotide	A	C	G	T	
Count Group 1	38514	24631	25685	38249	127079
Count Group 2	38550	24635	25700	38288	127173

Table 2: The effect of Indeterminacy on Q_N

I_{Q_N}	Q_N	Decision about H_0	I_{Q_N}	Q_N	Decision about H_0
[0,0]	[0.00664, 0.00664]	Do not reject H_0	[0,0.1]	[0.00664,0.007304]	Do not reject H_0
[0,0.01]	[0.00664,0.006706]	Do not reject H_0	[0,0.2]	[0.00664,0.007968]	Do not reject H_0
[0,0.02]	[0.00664,0.006773]	Do not reject H_0	[0,0.3]	[0.00664,0.008632]	Do not reject H_0
[0,0.03]	[0.00664,0.006839]	Do not reject H_0	[0,0.4]	[0.00664,0.009296]	Do not reject H_0
[0,0.04]	[0.00664,0.006906]	Do not reject H_0	[0,0.5]	[0.00664,0.00996]	Do not reject H_0
[0,0.05]	[0.00664,0.006972]	Do not reject H_0	[0,0.6]	[0.00664,0.010624]	Do not reject H_0
[0,0.06]	[0.00664,0.007038]	Do not reject H_0	[0,0.7]	[0.00664,0.011288]	Do not reject H_0
[0,0.07]	[0.00664,0.007105]	Do not reject H_0	[0,0.8]	[0.00664,0.011952]	Do not reject H_0
[0,0.08]	[0.00664,0.007171]	Do not reject H_0	[0,0.9]	[0.00664,0.012616]	Do not reject H_0
[0,0.09]	[0.00664,0.007238]	Do not reject H_0	[0,1]	[0.00664,0.01328]	Do not reject H_0