Determine Randomness of the Data based on the Size and Number of Runs for Large Sample

Ranjan Kumawat[1] Mr. Rahul Pawar[2]
[1]M Tech Student [2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]L.N.C.T. (Bhopal) Indore Campus Indore, India

Abstract— Wald-Wolfowitz Run test of randomness is a statistical test that is used to know the randomness in data. Run test of randomness assumes that the mean and variance are constant and the probability is independent. The Run Test is actually one of the most interesting statistical tests ever and it is so easy to understand, or even the easiest. There is a risk of selection bias when collecting a sample using consecutive sampling for an ordered population. No assurance that the sequence of data generated will be random because there is no mechanism to ensure random selection. In the proposed work we have taken more 1000 populations which have been vaccinated with corona vaccine for Covid-19 from 10 different vaccination center data from 10 different. We used run test to check the randomness between male and female. We applied properties of run test to check the randomness of data .First we taken 500 person's data in with 330 Male and 170 female and the numbers of runs are 237. Similarly for second population we have taken 1000 persons data in which 453 Male and 547 Female and number of runs are 512. We calculate z statistics we found in both the cases the calculated z statistics value with run test we found in both the cases the calculated value is less than the critical z value of tabulated value.so can say that both the sample have taken have randomness.

Keywords: Run, Randomness Population, Mean, Deviation

I. INTRODUCTION

Wald-Wolfowitz Run test of randomness is a statistical test that is used to know the randomness in data. Run test is a statistical test used to determine of the data obtained from a sample is random. That is why it is called Run Test for Randomness. Run test of randomness is basically based on the run. Run is basically a sequence of one symbol such as + or -. Run test of randomness assumes that the mean and variance are constant and the probability is independent. The Run Test is actually one of the most interesting statistical tests ever and it is so easy to understand, or even the easiest. There is a risk of selection bias when collecting a sample using consecutive sampling for an ordered population. No assurance that the sequence of data generated will be random because there is no mechanism to ensure random selection. It is difficult to describe research study in concrete terms rather than theoretical terms. We need sufficient evidence to prove the validity of the research. A random sample of data will be collected if the probability sampling method has been adopted. It is still possible for non-probability sampling to generate a random data. The methods of calculation for both test statistics and critical values are different between a small sample and a large sample[13,14].

II. WHAT IS A RUN

A run is a sequence of similar or like events, items or symbols that is preceded by and followed by an event, item or symbol of a different type, or by none at all. Randomness of the series is unlikely when there appear to be either too many or two few runs. In this case, a run test needs to be carried out to determine the randomness. The Run Test when performed helps us to decide whether a sequence of events, items or symbol is the result of a random process[3,5].

Example of Runs

A data scientist carrying out a research interviewed 10 persons during a survey. We denote the genders of the people by M for men and W for women. Assuming the respondents were chosen as follows:

1) Scenario 1

M M M M M F F F F F

Scenario 1 has only 2 runs and therefore the scenario cannot be considered random because there are too few runs

2) Scenario 2

F M F M F M F M F M

Scenario 2 has too many runs, 5 runs. And therefore would not be considered as random

3) Scenario 3

F F F M M F M M F F

Scenario 3 has 5 runs and therefore we need to perform a test to determine the randomness of the data.

III. ASSUMPTIONS IN RUN TEST OF RANDOMNESS

A. Data Level

In run of randomness it is assumed that the data is recorded in order and not in a group .In the data is not in order, then we have to assign the mean, maiden or mode value to the data

B. Data Scale

In run of randomness it is assumed that the data is in numeric from. This condition is compulsory in run test of randomness, because in numeric data it is easy to assign run to the numeric value

C. Distribution

In run of randomness is a non-parametric test so it does not assume any assumption about the distribution

D. Independent Probability

In run of randomness the probability of run is independent

IV. LITERATURE SURVEY

In 2010 Jean-Baptist du et al proposed "Choosing Statistical Tests". They discuss about frequently used statistical tests and their correct application. Methods: The most commonly used statistical tests were identified through a selective search

on the methodology of medical research publications. They acquainted not just with descriptive methods, but also with Pearson's chi-square test, Fisher's exact test, and Student's t test will be able to interpret a large proportion of medical research articles. Criteria are presented for choosing the proper statistical test to be used out of the most frequently applied tests. [1].

In 2011 Olaf Verschuren et al proposed "Reliability of a shuttle run test for children with cerebral palsy who are classified at Gross Motor Function Classification System level III". They analyzed data using SPSS 15.0 and MS Excel 2007 for Windows. They analyzed, the 'precision' indicates how well the methods agree for an individual. By multiplying the precision by 1.96, the 'limits of agreement' are calculated. The SDC is an estimate of the smallest change in score that can be detected objectively for a participant. Level III before the actual shuttle run test [2].

In 2012 M Bati Et al proposed "A new development cycle of the Statistical Toolkit". The Statistical Toolkit is an open source system specialized in the statistical comparison of distributions. It addresses requirements common to different experimental domains, such as simulation validation, regression testing in the course of the software development process, and detector performance monitoring. The new tests extend the Statistical Toolkit capabilities with tests for randomness and tests for categorical data analysis. The new user layer component makes it possible to use the Toolkit with many spreadsheet applications that allow exporting data directly to comma separated list of values [3].

In 2013 Otuken Senger proposed "Comparison of Type I Error Probabilities of Wald Wolfowitz and Mann Whitney Tests for Large, Small and Equal Sample Sizes". They presented a comparison of type I error probabilities of Wald Wolfowitz and Mann Whitney tests were done. They found small and equal size samples, all probabilities of type I error of Wald Wolfowitz test was below the significance level. In small and equal sample sizes, Wald Wolfowitz test is defined as liberal while Mann Whitney test is defined as conservative. Probabilities of type I error of nonparametric tests are not affected both by variance of kurtosis coefficient at a fixed skewness coefficient [4].

In 2014 I. V. Veretel nikova et al proposed "The analytical review of tests for randomness and the absence of a trend". They proposed a analytical review for randomness and the absence of a trend To ensure the correctness of statistical conclusions using tests considered in situations when the standard assumptions, are violated providing the legitimacy of using the classical results, or no information about the "true" distribution of the statistic of a used tests (under specific conditions and at a particular sample size), an interactive mode of research of statistic distributions with the following usage of the resulting distribution when deciding on the results of testing the hypothesis for calculation value is released [5].

In 2015 Daniel Mayorga-Vega et al proposed "Criterion-Related Validity of the 20-M Shuttle Run Test for Estimating Cardi-orespiratory Fitness: A Meta-Analysis". They showed that the criterion-related validity of Léger's protocol was statistically higher for adults than for children. When an individual's maximum oxygen uptake attained during a laboratory-based test is not feasible, the 20-m shuttle run test seems to be a useful alternative for estimating cardiorespiratory fitness. In adults the performance score only seems to be a strong estimator of cardiorespiratory fitness, in contrast among children the performance score should be combined with other variables[6].

In 2016 Slobodan B. Gadzuric*a* et al proposed "Multivariate Chemometrics with Regression and Classification Analyses in Heroin Profiling Based on the Chromatographic Data". They used gas chromatography analysis of heroin samples seized from three different locations in Serbia. Since the correlation was extremely good, our mathematical models can be used to predict geographic origin of seized heroin samples in Serbia, using the GC results[7].

In 2017 Mohamad Adam et al proposed "An Application of the Runs Test to Test for Randomness of Observations Obtained from a Clinical Survey in an Ordered Population". They applied to test the randomness of data in a survey that collect data from an ordered population. They performed a runs test and explained the rationale for performing it by providing some examples of how this test can be applied. Their aim of this article was to describe on ways to use the runs test in a clinical survey from an ordered population to determine the degree of randomness in the sequence of subjects who are recruited within a sample obtained from the whole population [8].

In 2018 Sasa Tokic et al proposed "Testing efficient market hypothesis in developing Eastern European countries". They suggest that all analyzed indices, except Serbia, confirm weak form of efficient market hypothesis, while the results on the index are mixed and it can be concluded that it does not follow weak form of efficient market hypothesis. Various researches during the last couple of decades tried to confirm or reject efficiency of the markets using different statistical analysis tests. They believed that newer markets with lower market capitalization are often inefficient and at least partially reject efficient market hypothesis in its weak form, which suggests that it is possible to achieve above market returns if the transaction costs and slippage allow exploiting these inefficiencies [9].

In 2019 Bhaswar B. Bhattacharya proposed "A General Asymptotic Framework For Distribution-Free Graph-Based Two-Sample Tests". They introduced a general notion of distribution-free graph-based two-sample tests, and provided a unified framework for analyzing and comparing their asymptotic properties. The results show how the combinatorial properties of the underlying graph affect the performance of the associated two-sample test, and can be used to validate and decide which tests to use in practice. Applications of the results are illustrated both on synthetic and real datasets[10].

In 2020 Valdir Adilson Steinke et al "Trend Analysis of Air Temperature in the Federal District of Brazil: 1980–2010"**.** They designed the work to identify trends in maximum, minimum, and average air temperatures in the Federal District of Brazil from 1980 to 2010, measured at five weather stations. Three statistical tests Wald–Wolfowitz, Cox–Stuart, and Mann–Kendall were tested for their applicability for this purpose, and the ones found to be most suitable for the data series were validated. This investigation was primarily concerned with assessing whether trend tests

are capable of confirming or refuting the existence of trends in air temperature data in the Federal District of Brazil between 1980 and 2010[11].

In 2021 C. Sahatsathatsana et al proposed "Error estimation for non-overlapping success runs with length k via Stein-Chen method". They are interested in studying the problem of the number of non-overlapping success runs of length k ($1 \leq k \leq n$) to provide an error estimation of this problem through the use of the Stein-Chen coupling method. They showed the distribution of the number of non-overlapping occurrence of the success run of length k in n trials. They defined the most important and frequently used statistics of success runs associated with non-identical independent Bernoulli trials X1,X2, . . . ,Xn with success probabilities p and failure probabilities $q = 1 - p$, for n and k ($1 \leq k \leq n$), in case when Wn,k is the number of non-overlapping success runs of length k[12].

V. PROPOSED APPROACH

We would use the following 7 steps to find the z-statistic.
Step 1: State the null and alternate hypothesis
Step 2: Determine the number of runs
Step 3: Calculate the mean runs
Step 4: Calculate the Standard deviation
Step 5: Calculate the z-Statistic
Step 6: Determine the Critical value
Step 7: Draw a conclusion

VI. ILLUSTRATE WITH EXAMPLE

The following arrangement of men, M, and women, W, lined up for covid-19 vaccination center:

M	**W**	**M**	**W**	**M**	**M**	**M**	**W**	**M**	**W**	**M**	**M**
M	**W**	**W**	**M**	**M**	**M**	**M**	**W**	**W**	**M**	**W**	**M**
M	**M**	**W**	**M**	**M**	**M**	**W**	**W**	**W**	**M**	**W**	**M**
M	**M**	**W**	**M**	**W**	**M**	**M**	**M**	**M**	**W**	**W**	**M**

Table 1: Man and women, W, lined up for covid-19 vaccination center

Test for randomness at the $\alpha = 0.05$ significance level

1) Solution Steps

In this case, we see that the sample size is fairly large, so we are going to use a slightly different method in this case. We are going to calculate the mean runs and the standard deviation.

2) Step 2: Determine the Number of Runs

This means that you need to mark each of the categories so you can easily count them Here I mark each run alternatively with red for M and black for W. The outcome is given below.

M	**W**	**M**	**W**	**M**	**M**	**M**	**W**	**M**	**W**	**M**	**M**
M	**W**	**W**	**M**	**M**	**M**	**M**	**W**	**W**	**M**	**W**	**M**
M	**M**	**W**	**M**	**M**	**M**	**W**	**W**	**W**	**M**	**W**	**M**
M	**M**	**W**	**M**	**W**	**M**	**M**	**M**	**M**	**W**	**W**	**M**

Table 2 Man and Women with marks Red and Green

The number of runs is given by R = 27 the number of Men, n1 = 30 The number of Women, n2 = 18

3) Step 3: Calculate the Mean runs

The mean is given by the formula

$$\mu_R = \frac{2n_1n_2}{n_1 + n_2} + 1$$

We can go ahead to substitute the value of $n_1 = 30$ and $n_2 = 18$

$$\mu_R = \frac{2 \times 30 \times 18}{30 + 18} + 1$$
$$\mu_R = \frac{1080}{48} + 1$$
$$\mu_R = 22.5 + 1$$
$$\mu_R = 23.5$$

4) Step 4: Calculate the Standard Deviation

You can find the standard deviation for a runs test using the formula

$$\sigma_R^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

We can go ahead to substitute the value of n1 = 30 and n2 = 18

$$\sigma_R^2 = \frac{2 \times 30 \times 18(2 \times 30 \times 18 - 30 - 18)}{(30 + 18)^2(10 + 18 - 1)}$$
$$\sigma_R^2 = \frac{1080(1080 - 30 - 18)}{(48)^2(47)}$$
$$\sigma_R^2 = \frac{1080(1032)}{2304(47)}$$
$$\sigma_R^2 = \frac{1114560}{108288}$$

Population size	Calculated z-statistic value	critical value (standard z value)
500	1.2	1.645

$$\sigma_R^2 = 10.2926$$
$$\sigma_R = \sqrt{10.2926}$$
$$\sigma_R = 3.2083$$

5) Step 5: Calculate the z-Statistic

The z-Statistic can be calculated using the formula

$$Z = \frac{R - \mu_R}{\sigma_R}$$

Then we can substitute the value of

R = 27

μ_R= 23.5

σ_R^2= 3.2083

$$Z = \frac{27 - 23.5}{3.2083}$$
$$Z = \frac{3.5}{3.2083}$$
$$Z = 1.091$$

6) Step 6: Determine the Critical Value

Look up the value of the critical value from statistical table of normal distribution.

We get a critical value for of 1.96

7) Step 7: State the Decision

Since the calculated value of z = 1.0909 is within the accept region(less than the critical value of 1.96), we therefore accept (fail to reject) the null hypothesis and conclude that there is not real evidence that the arrangement is not random.

VII. COMPARATIVE ANALYSIS

We compare the calculated z-Statistic value with standard z value for 500 population in which number of male is 330 and number of female if 170 and number of runs 237

Fig. 1: Comparison z-statistic value with critical value 500 populations

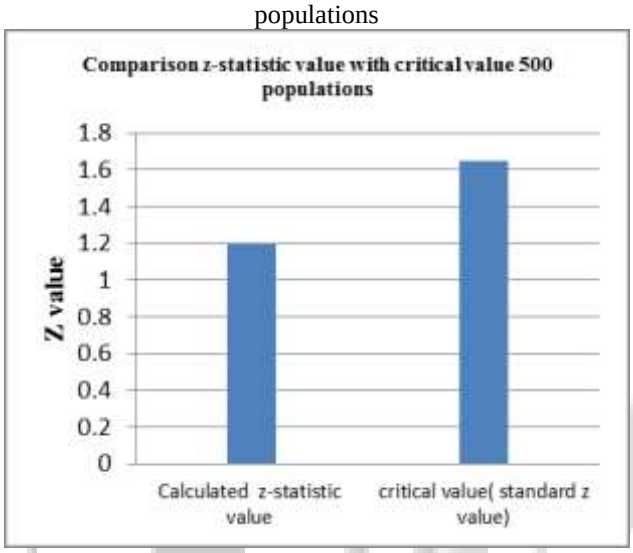

Chart 1: Comparison z-statistic value with critical value 500 populations

VIII. CONCLUSION

In this paper we want to check the randomness in Coivd-19 vaccination data form male and female. We collected data form more the Covid-19 vaccination center. The size of the data is more than 1000 (male and female). In the proposed work we apply Wald-Wolfowitz Run test to check the randomness in given data. We calculate z statistics value by using Wald-Wolfowitz Run and compare this value with the critical or standard value from z table. By the experimental analysis we found that for given population has random in nature because the calculated z statistics value is less than the critical value.

REFERENCES

[1] Jean-Baptist du Prel, Bernd Rohrig "Choosing Statistical Tests" Cite this as: Dtsch Arztebl Int 2010; 107(19): 343–8 DOI: 10.3238/arztebl.2010.0343 Part 12 of a Series on Evaluation of Scientific Publications.

[2] Olaf Verschuren Liesbeth Bosma "Reliability of a shuttle run test for children with cerebral palsy who are classified at Gross Motor Function Classification System level III" Accepted for publication 1st December 2010. Published online 11 February 2011. Developmental Medicine & Child Neurology.

[3] M Bati A. M. Paganoni, A. Pfeiffer, "A new development cycle of the Statistical Toolkit" arXiv: 1209.5999 26 Sep 2012. Dept. of Mathematics, Politecnico di Milano, Milano, Italy.

[4] Otuken Senger "Comparison of type I Error Probabilities of Wald Wolfowitz and Mann Whitney Tests for Large, Small and Equal Sample Sizes". International Journal of Academic Research Part B; 2013; 5(4), 188-195. DOI: 10.7813/2075-4124.2013/5-4/B.28.

[5] I. V. Veretel nikova "The analytical review of tests for randomness and the absence of a trend" 978-1-4799-6019-4/14/ © 2014 IEEE Novosibirsk State Technical University.

[6] Daniel Mayorga-Vega, Pablo Aguilar-Soto and Jesús Viciana "Criterion-Related Validity of the 20-M Shuttle Run Test for Estimating Cardi-orespiratory Fitness: A Meta-Analysis" Journal of Sports Science and Medicine (2015) 14, 536-547 http://www.jssm.org.

[7] Slobodan B. Gadzurica, Sanja O. Podunavac Kuzmanovic "Multivariate Chemometrics with Regression and Classification Analyses in Heroin Profiling Based on the Chromatographic Data" Iranian Journal of Pharmaceutical Research (2016), 15 (4): 725-734.

[8] Mohamad Adam Bujang, Fatin Ellisya Sapri "An Application of the Runs Test to Test for Randomness of Observations Obtained from a Clinical Survey in an Ordered Populatio" Malays J Med Sci. Jul–Aug 2018; 25(4): 146–151 www.mjms.usm.my © Penerbit Universiti Sains Malaysia, 2018.

[9] Sasa Tokic , Berislav Bolfek , Anita Radman Pesa "Testing efficient market hypothesis in developing eastern European countries". Investment Management and Financial Innovations, Volume 15, Issue 2, 2018.

[10] Bhaswar B. Bhattacharya "A General Asymptotic Framework For Distribution-Free Graph-Based Two-Sample Tests" arXiv:1508.07530v5 16 Apr 2019.

[11] Valdir Adilson Steink1, Luis Alberto Martins Palhares de Melo "Trend Analysis of Air Temperature in the Federal District of Brazil: 1980–2010" Climate 2020, 8, 89; doi: 10.3390/cli8080089 www.mdpi.com/journal/climate.

[12] C. Sahatsathatsana, S. Sahatsathatsana, W. Pimpasalee Error estimation for non-overlapping success runs with length k via Stein-Chen method International Journal of Mathematics and Computer Science, 16(2021), no. 4, 1771–1781.

[13] Anna Borucka, Edward Kozłowski, Piotr Oleszczuk, and Andrzej Świderski "Predictive analysis of the impact of the time of day on road accidents in Poland" Open Access. © 2021 A. Borucka et al., published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 License.

[14] Welber Ferreira Alves, Henrique Roig" Analyzing Trends in Rainfall and Their Impacts in Water Management in a Cerrado Region in Brazil" September 23rd, 2021 DOI: https://doi.org/10.21203/rs.3.rs-830788/v1.

