Comparative Analysis of Mcnemar's Test and Liddell's Exact Test for Assessing Marginal Homogeneity

[1]Aremu, Zainab O. and [1,2]Akanni, John O.
[1]Department of Mathematical and Computing Sciences, KolaDaisi University, Ibadan, Oyo State, Nigeria., [2]Department of Mathematics, Universitas Airlangga, Kampus C Mulyorejo Surabaya 60115, Indonesia.

***Corresponding author:** Akanni, John O. (*zainabolaide0808@gmail.com*)
DOI: https://doi.org/10.5281/zenodo.17438475

Abstract
In this study, we aim to identify the preferred test for assessing marginal homogeneity, compare McNemar's and Liddell's exact tests, and determine the conditions under which each test is most appropriate. A Monte Carlo simulation approach was employed to develop and analyze both tests. The power and Type I error rates were computed for various sample sizes, effect sizes, and significance levels (α). Random samples ranging from 20 to 1000 were analyzed over 500 iterations. Three different hypothetical scenarios were used to evaluate the performance of McNemar's test and Liddell's exact test, considering different P_{12} and P_{21} values and using significance levels of 1% and 10%. The results indicate that Liddell's exact test is generally preferable, mainly when the effect sizes are moderate to large across nearly all sample sizes. McNemar's test is not recommended for sample sizes of 20 or fewer. Liddell's exact test is more advantageous when the proportions are small and close together, especially with larger sample sizes.

Keywords: Comparative analysis; McNemar's test; Liddell's exact test; Marginal homogeneity; Monte carlo simulation

Introduction

In statistical analysis, variable classification and the choice of appropriate tests are crucial for accurate data interpretation. This research focuses on the comparative evaluation of McNemar's test and Liddell's exact test; both used to assess marginal homogeneity in paired categorical data. Williams' (1946) classification of variables into nominal, ordinal, interval, and ratio scales has long been a standard, but its application in selecting the right test for marginal homogeneity remains a topic of discussion.

The effectiveness of statistical tests in evaluating marginal homogeneity in paired categorical data is a significant concern in research. This study undertakes a comparative analysis of McNemar's test and Liddell's exact test, using Monte-Carlo simulations to assess their performance under various conditions. This approach provides insights into which test is more suitable based on sample and effect sizes.

Williams (1946) developed a measurement scale hierarchy with four categories: nominal, ordinal, interval, and ratio scales, each with prescribed appropriate statistical analyses. The nominal scale is the lowest, and the ratio scale is the highest. However, this typology has faced criticism from scholars such as Lord (1965), Guttman (1968), Tukey (1961), and Velleman and Wilkinson (1993), particularly regarding the justification of statistical methods based on these scale types. Velleman and Wilkinson (1993) highlighted situations where Stevens' categorization failed, leading to alternative taxonomies like the one proposed by Mosteller and Tukey (1977), which includes grades, ranks, counted fractions, counts, amounts, and balances.

A categorical variable consists of a set of non-numerical categories and comes in two types: nominal and ordinal. Nominal variables have unordered, mutually exclusive categories identified by numerals, letters, or colors, such as gender, marital status, party affiliation, race, and religious affiliation. The frequency count is the number of occurrences in each category. Nominal variables remain unchanged under transformations that preserve the relationship between subjects and their identifiers as long as categories are not combined, and their statistical analysis is invariant under the permutation of categories.

A 2x2 contingency table displays the joint frequency distribution of two dichotomous

classificatory variables, A and B. This table provides four category combinations and is summarized as follows:

Table 1.1: Observed Frequencies

A	B		Total
	1	2	
1	$n_{11}n_{12}$		$n_{1.}$
2	$n_{21}n_{22}$		$n_{2.}$
Total	$n_{.1}n_{.2}$		N

In this formulation, the marginal number of subjects in the i-th level of A is denoted by $n_{i.}$ and the marginal total number of subjects in the j-th level of B is denoted by $n_{.j}$. The total sample size is N. If the row margin is assumed to be fixed, these totals are denoted by n_i, and if the column margin is assumed to be fixed, these totals are denoted by $n_{.j}$. Each entry in the table's body refers to a cell of the table.

McNemar's test is a nonparametric method for assessing marginal homogeneity in a 2×2 table, valid for paired data analysis. For larger N×N tables, it can be adapted to test for rater bias or equality of category thresholds. The Stuart-Maxwell test complements it by providing an overall significance value across all categories. These tests are easy to use, practical, and require minimal assumptions. McNemar's test statistic is $X^2 = \frac{(n_{12} - n_{21})^2}{n_{12}} + n_{21}$ with one degree of freedom.

Liddell's exact test compares paired proportions, which is especially useful when McNemar's assumptions are unmet. It treats the data as a binomial variable and calculates the probability that the ratio R' = $\frac{n_{12}}{n_{21}+1}$ equals 1, testing the null hypothesis. This test is ideal for scenarios where the same subjects are exposed to different conditions, such as comparing consumer preferences for two commercials.

Method

The power of a statistical test, as defined by Brown (2011), is the probability of rejecting the null hypothesis when it is false, thus confirming the alternative hypothesis. It is inversely related to the probability of a Type II error (false negative) and is influenced by the effect size, sample size, and significance criterion. Power analysis helps determine the minimum sample size needed to detect a given effect size and compares statistical tests. It is essential in hypothesis testing to ensure that the test can detect differences in the population (Brown, 2011).

Several factors influence the power of a test, including the statistical significance criterion, the effect size, and the sample size. Increasing the significance criterion (e.g., from 0.05 to 0.10) can boost power and increase the risk of Type I errors (false positives). Larger effect sizes and sample sizes generally enhance power. Additionally, the precision of data measurements and the design of the experiment can impact power. For instance, balanced sample sizes in two-sample comparisons and optimized values in regression analysis can improve power (Brown, 2011).

Typically, a power of 0.80 is considered adequate, implying a 4-to-1 trade-off between Type II (β) and Type I (α) errors. However, in contexts like medicine, higher power is often desired to avoid false negatives, even at the cost of more false positives. Power analysis is crucial for correct hypothesis rejection and determining necessary sample sizes to achieve precise estimates of population effect sizes (Brown, 2011).

Power analysis can be performed before (a priori) or after (post hoc) data collection. A priori analysis estimates the required sample size to achieve adequate power, while post hoc analysis assesses the power of a completed study. However, post hoc power analysis is controversial and can be misleading, as it tends to reflect the p-value rather than providing meaningful insights (Brown, 2011).

McNemar's test is often required by funding agencies and review panels to ensure that studies are adequately powered. While underpowered studies in frequentist statistics are unlikely to effectively distinguish between hypotheses, Bayesian statistics focus on updating prior beliefs based on data. Despite this, power remains a useful measure to gauge the expected impact of an experiment on refining beliefs (Brown, 2011).

Marginal Homogeneity

Marginal homogeneity refers to the equality (or lack of significant difference) between the row marginal proportions and the corresponding column proportions. This concept is crucial in analyzing rater agreement, as differences in raters' marginal rates can be formally assessed using statistical tests of marginal homogeneity (Barlow,

1998; Bishop et al., 1975). Testing marginal homogeneity is straightforward when different raters rate different cases using a chi-squared test but becomes complex when different raters rate the same cases due to statistical dependence. Approaches to this problem include nonparametric tests, bootstrap methods, loglinear models, and latent trait models.

McNemar's Test

McNemar's test, introduced by Quinn McNemar in 1947, determines whether the row and column marginal frequencies are equal in 2x2 contingency tables with matched pairs. This test is standard in studies involving matched pairs, such as case-control studies, and measurements at two different time points. The test focuses on the off-diagonal elements of the table, as shown below:

Table 1.2: Observed Frequency for Matched Pairs Data

Response 2	Response 1		Total
	Yes	No	
Yes	n_{11}	n_{12}	$n_{1.}$
No	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	N

McNemar's test statistic is calculated as: $X^2 = \frac{(|n_{12} - n_{21}| - 1)^2}{n_{12} + n_{21}}$

Liddell's Exact Test

Liddell's exact test, an alternative to McNemar's test, is used for paired proportions and is a particular case of the sign test (Journal of Epidemiology and Community Health, 1983). This test treats the n12 count as a binomial variable from the sample $n_{12} + n_{21}$ and uses the ratio R'= n_{12} / n_{21} to calculate a two-sided probability and confidence limits for relative risk. It is applied when comparing two laboratory methods or assessing the effect of a risk factor on a matched sample.

Table 1.3: Corresponding Probability for Matched Pairs Data

Response 2	Response 1		Total
	Yes	No	
Yes	p_{11}	p_{12}	$p_{1.}$
No	p_{21}	p_{22}	$p_{2.}$
Total	$p_{.1}$	$p_{.2}$	

Both McNemar's test and Liddell's exact test provide methods to assess marginal homogeneity, which is crucial for understanding rater agreement and paired proportion comparisons in various research contexts.

Simulation Study

In this section, the simulation of this study was presented:

3.1 First Scheme When α = 0.01

(i) H_0: P_{12}-P_{21}=δ versus H_1: P_{12}- P_{21}= δ, where δ=0 under H_0, when P_{12}=0.2 and P_{21}=0.6 with effect size δ=0.4.

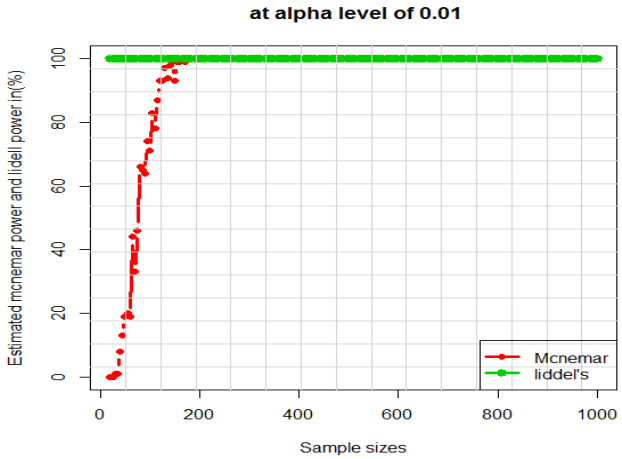

Fig. 3.1: A graphical display of the result for the power of both tests when P_{12} =0.2 and P_{21} = 0.6 at α=0.01

(ii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.3$ and $P_{21}=0.6$ with effect size $\delta=0.3$.

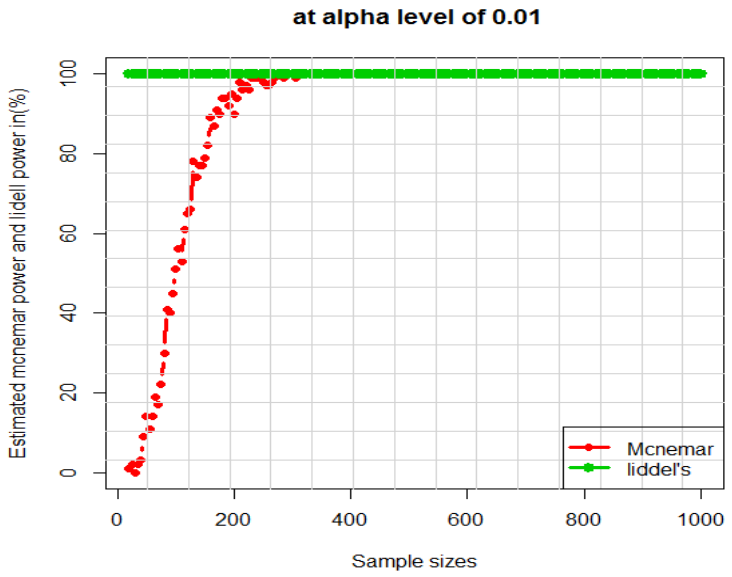

Fig. 3.2: A graphical display of the result for the power of both tests when P_{12} =0.3 and P_{21} = 0.6 at α=0.01

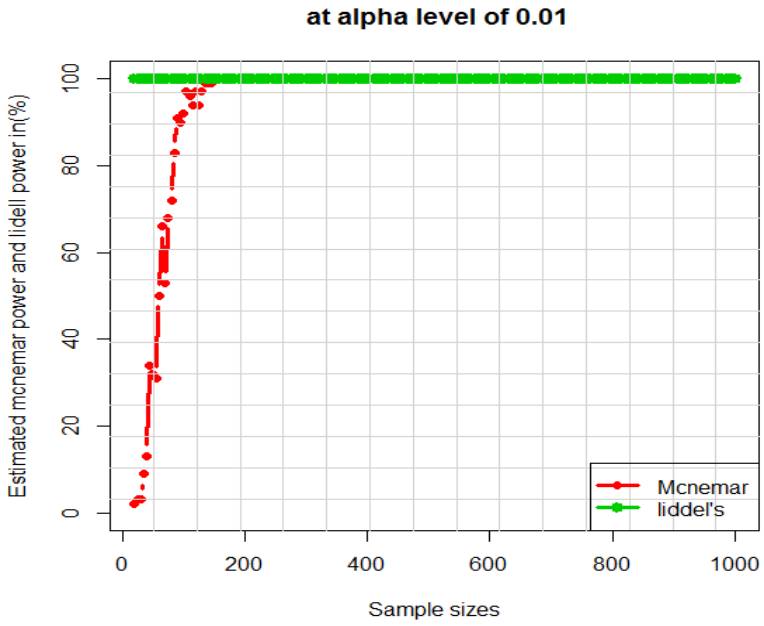

Fig. 3.3: A graphical display of the result for the power of both tests when P_{12} =0.4 and P_{21} = 0.6 at α=0.01

It can be seen that Liddell's exact test has 100% power for almost the sample sizes, while McNemar's test power is low until it reaches 150 sample sizes before it becomes stable.

(iii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.5$ and $P_{21}=0.6$ with effect size $\delta=0.1$.

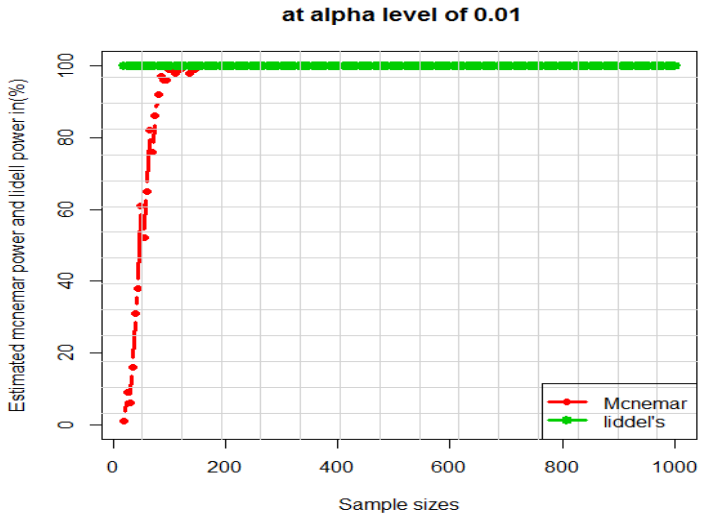

Fig. 3.4: A graphical display of the result for the power of both tests when $P_{12}=0.5$ and $P_{21}=0.6$ at $\alpha=0.05$

It can be seen that Liddell's exact test has 100% power for almost the sample sizes, while McNemar's test power is low until it reaches 150 sample sizes before it becomes stable.

3.2 Second Scheme When α = 0.01

(i) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.9$ and $P_{21}=0.1$ with effect size $\delta=0.8$.at $\alpha= 0.01$

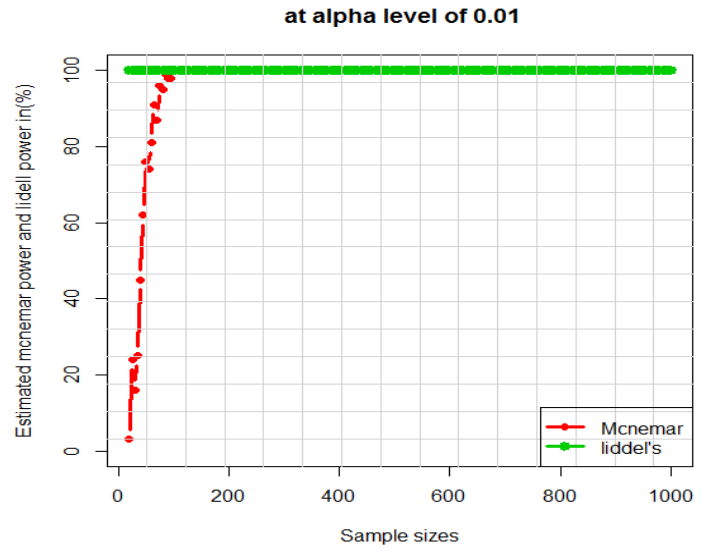

Fig. 3.5: A graphical display of the result for the power of both tests when $P_{12}=0.9$ and $P_{21}=0.1$ at $\alpha=0.01$

It can be seen that Liddell's exact test has 100% power for all the sample sizes, while McNemar's test power is only for lower sample sizes.

(ii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.8$ and $P_{21}=0.2$ with effect size $\delta=0.6$.

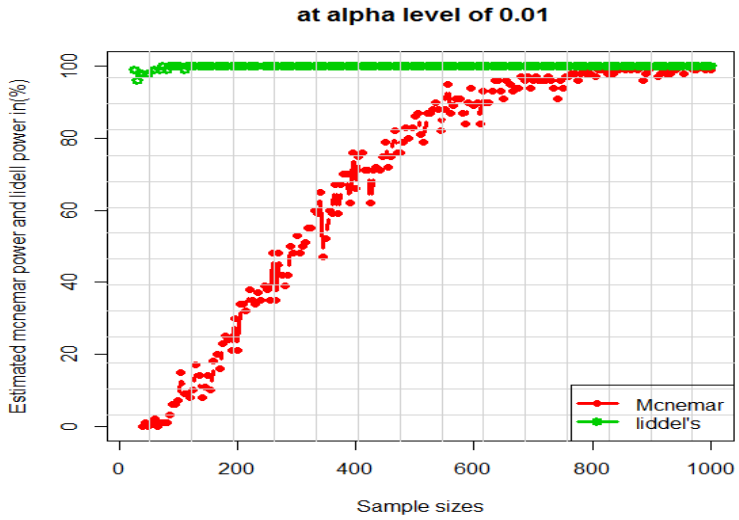

Fig. 3.6: A graphical display of the result for the power of both tests when $P_{12}=0.8$ and $P_{21}=0.2$ at $\alpha=0.01$

Both of our very poor for low sample sizes (20), but Liddell's exact test has 100% power for the remaining sample sizes, but McNemar's test power increases as the sample size increases.

(iii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.7$ and $P_{21}=0.3$ with effect size $\delta=0.4$.

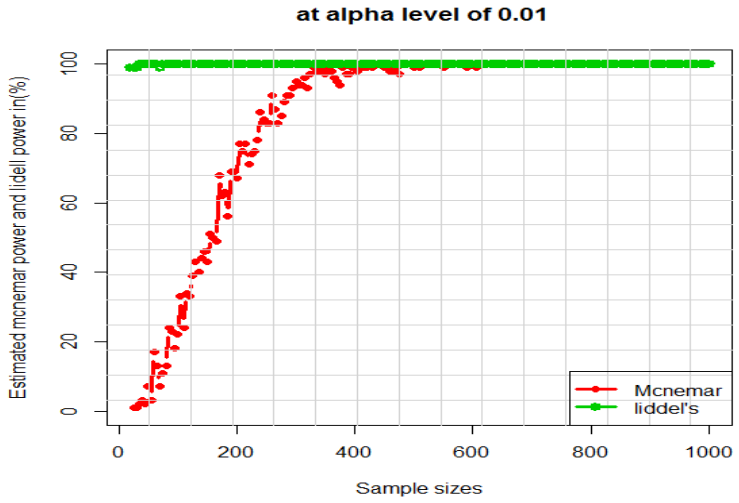

Fig. 3.7: A graphical display of the result for the power of both tests when $P_{12}=0.7$ and $P_{21}=0.3$ at $\alpha=0.05$

It can be seen that Liddell's exact test has 100% power for almost the sample sizes while McNemar's test power increases and the sample size increasing.

(iv) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.5$ and $P_{21}=0.4$ with effect size $\delta=0.1$.

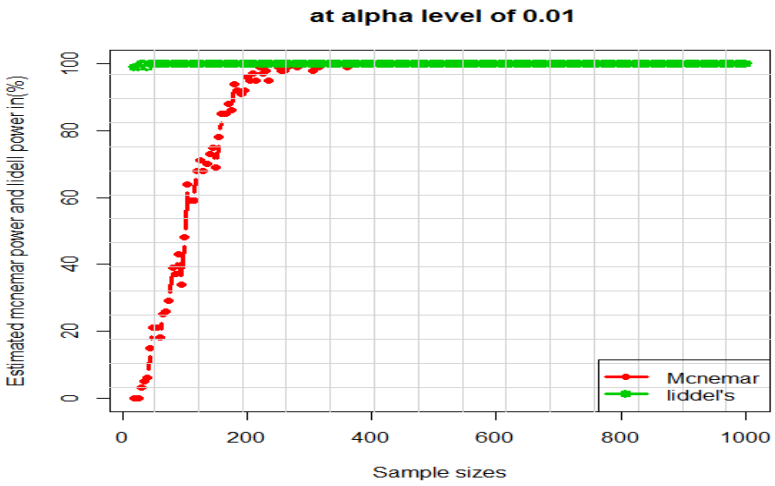

Fig. 3.8: A graphical display of the result for the power of both tests when $P_{12}=0.5$ and $P_{21}=0.4$ at $\alpha=0.01$

It can be seen that Liddell's exact test has 100% power for almost all sample sizes, while McNemar's test power increases as the sample size increases.

3.3 Third Scheme When α = 0.01

(i) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.2$ with effect size $\delta=0$

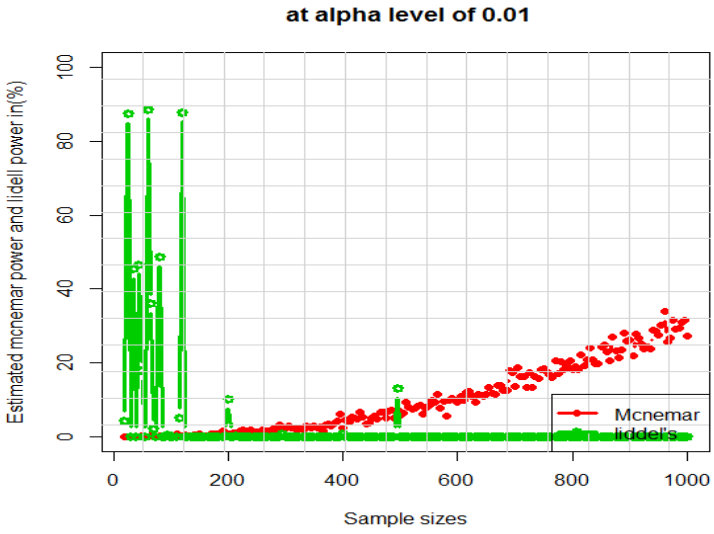

Fig. 3.9: A graphical display of the result for the type I error committed by both tests when $P_{12}=0.2$ and $P_{21}=0.2$ at $\alpha=0.05$

From the graph, it can be seen that the type I error committed by McNemar's test was very low in the sample size (≤ 75) but increased as the sample sizes increased. In contrast, the type I error committed by Lindell's exact test is only high for the sample sizes (≤ 50) and very low for the remaining sample sizes.

(ii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.21$ with effect size $\delta=0.01$.

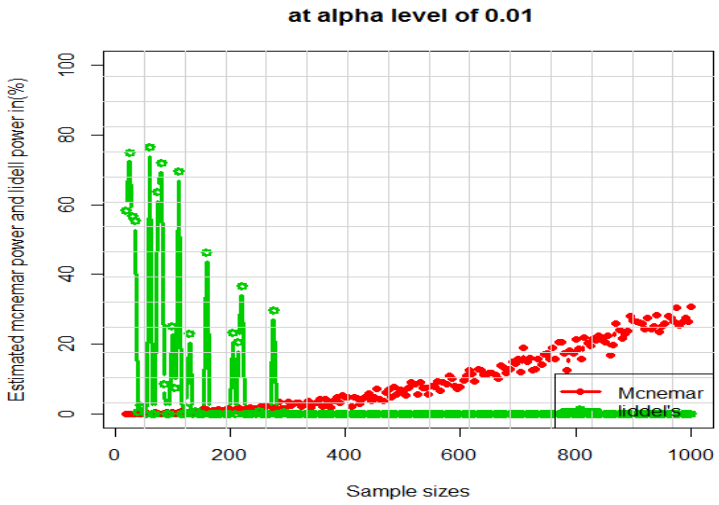

Fig. 3.10: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.21 at α=0.05

From the graph, it has been shown that I error committed by McNemar's test was very low for a low sample size ($\leq$ 50) but increased as the sample sizes. Type I error committed by Lindell's exact test is unstable throughout the sample sizes considered.

(iii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.22$ with effect size $\delta=0.02$.

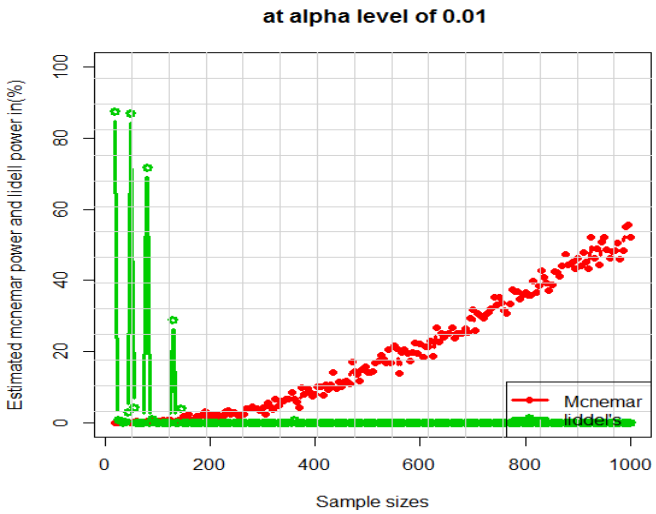

Fig. 3.11: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.22 at α=0.05

From the graph, it has been shown that I error committed by McNemar's test was deficient for a low sample size ($\leq$ 150) but increased as the sample sizes. The type I error committed by Liddell's exact test is unstable throughout the sample sizes.

(iv) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.23$ with effect size $\delta=0.03$

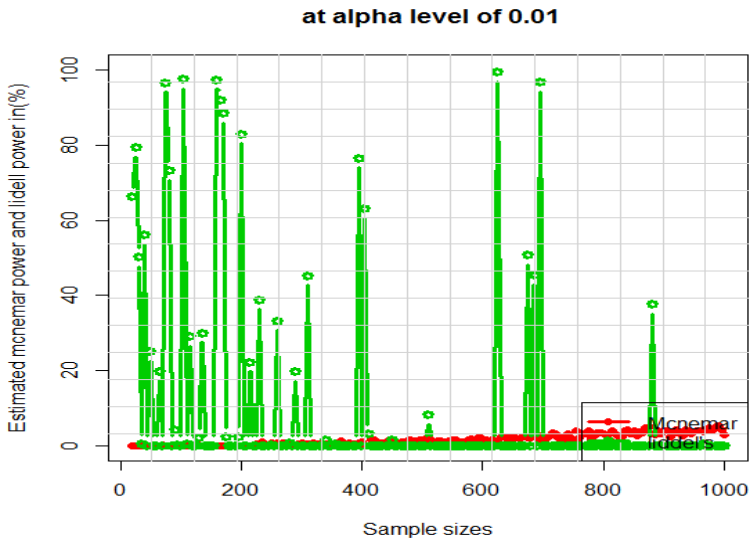

Fig. 3.12: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.23 at α=0.05

From the graph above, it can be deduced that the type I error committed by both tests is unstable throughout the sample sizes considered.

3.4 First Scheme When α = 0.1

(i) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.6$ with effect size $\delta=0.4$.

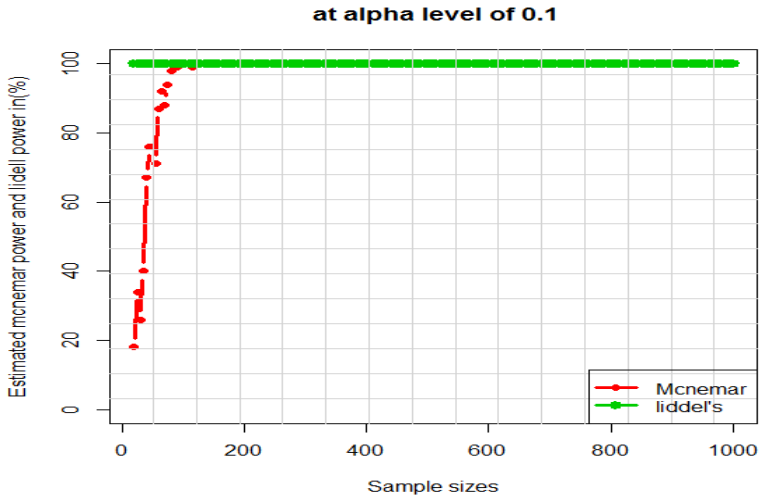

Fig. 3.13: A graphical display of the result for the power of both tests when P_{12} =0.2 and P_{21} = 0.6 at α=0.1

It can be seen that Liddell's exact test has 100% power for all the sample sizes, while McNemar's test power is only for lower sample sizes.

(ii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.3$ and $P_{21}=0.6$ with effect size $\delta=0.3$.

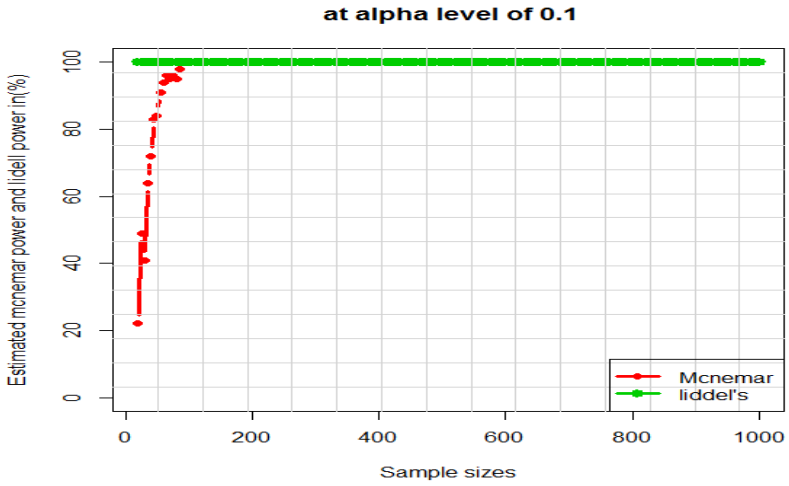

Fig. 3.14: A graphical display of the result for the power of both tests when P_{12} =0.3 and P_{21} = 0.6 at α=0.1

It can be seen that Liddell's exact test has 100% power for all the sample sizes, while McNemar's test power is only for lower sample sizes.

(iii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.4$ and $P_{21}=0.6$ with effect size $\delta=0.2$.

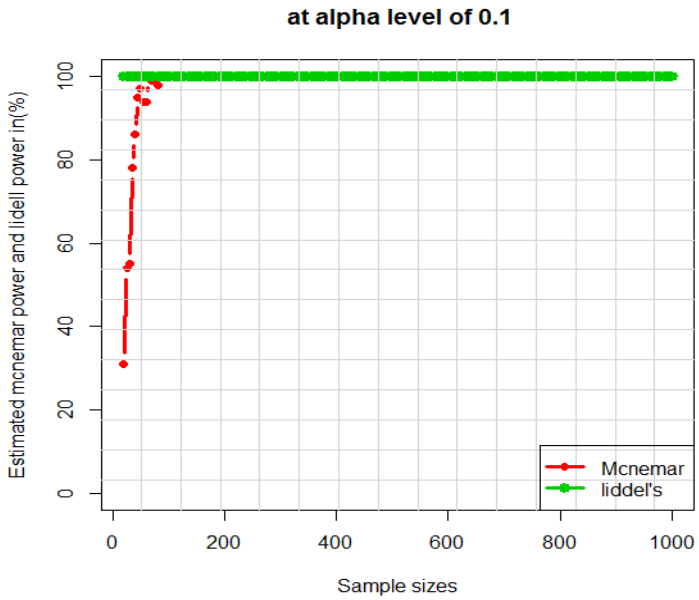

Fig. 3.15: A graphical display of the result for the power of both tests when P_{12} =0.4 and P_{21} = 0.6 at α=0.1

It can be seen that Liddell's exact test has 100% power for all the sample sizes, while McNemar's test power is only for lower sample sizes.

(iv) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.5$ and $P_{21}=0.6$ with effect size $\delta=0.1$.

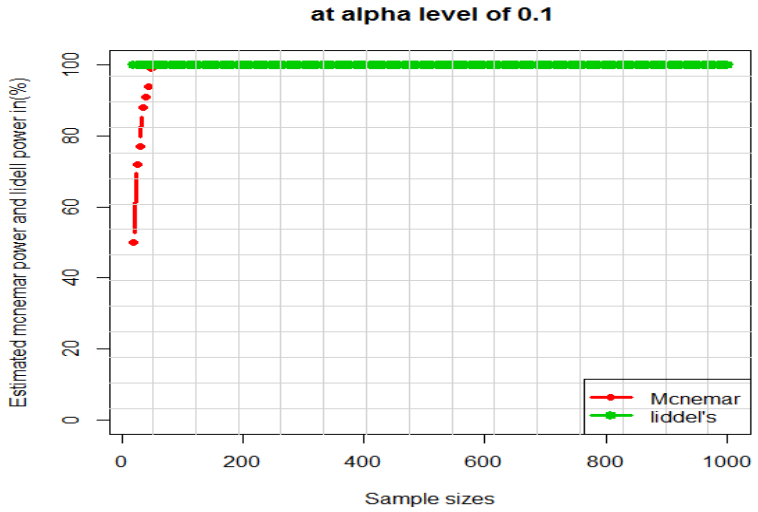

Fig. 3.16: A graphical display of the result for the power of both tests when P_{12} =0.5 and P_{21} = 0.6 at α=0.1

Liddell's exact test has 100% power for all sample sizes, while McNemar's test power is only fair for lower sample sizes.

3.5 Second Scheme When α = 0.1

(i) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.9$ and $P_{21}=0.1$ with effect size $\delta=0.8$.

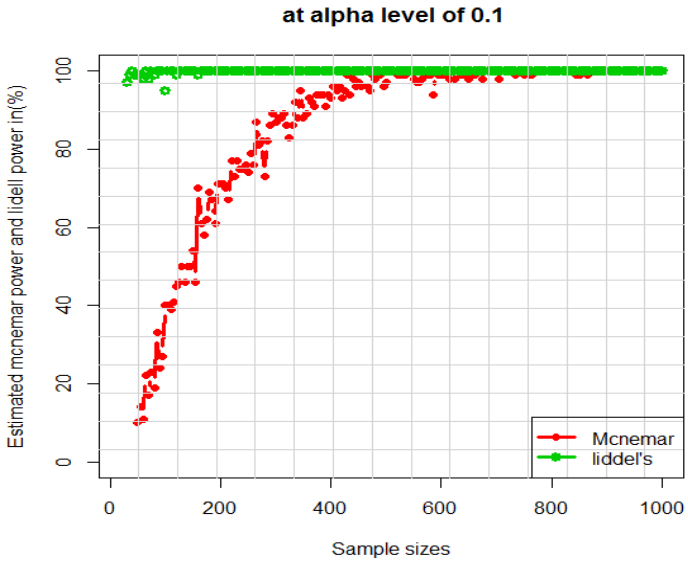

Fig. 3.17: A graphical display of the result for the power of both tests when P_{12} =0.9 and P_{21} = 0.1 at α=0.1

Both of our very poor for low sample sizes (≤ 20), but Liddell's exact test has 100% power for the remaining sample sizes. McNemar's test power increases as the sample size increases.

(ii) H_0: P_{12}-P_{21}=δ versus H_1: P_{12}- P_{21}= δ, where δ=0 under H_0, when P_{12}=0.8 and P_{21}=0.2 with effect size δ=0.6.

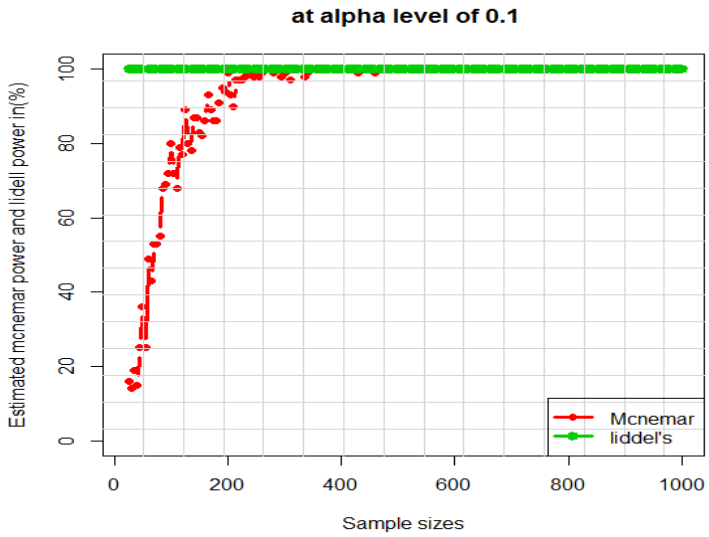

Fig. 3.18: A graphical display of the result for the power of both tests when P_{12} =0.8 and P_{21} = 0.2 at α=0.1

Both of our very poor for low sample sizes (≤ 20), but Liddell's exact test has 100% power for the remaining sample sizes. McNemar's test power increases as the sample size increases.

(iii) H_0: P_{12}-P_{21}=δ versus H_1: P_{12}- P_{21}= δ, where δ=0 under H_0, when P_{12}=0.7 and P_{21}=0.3 with effect size δ=0.4.

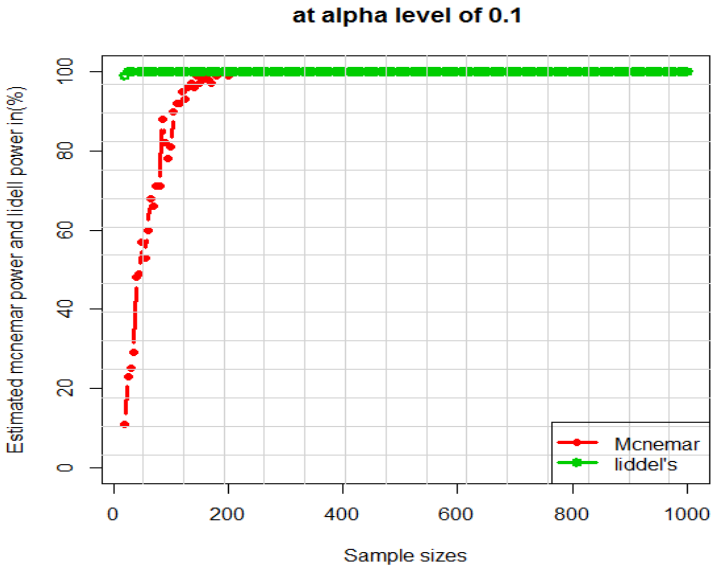

Fig. 3.19: A graphical display of the result for the power of both tests when P_{12} =0.7 and P_{21} = 0.3 at α=0.1

It can be seen that Liddell's exact test has 100% power for almost all the sample sizes, while McNemar's test power increases as the sample size increases.

(iv) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.5$ and $P_{21}=0.4$ with effect size $\delta=0.1$.

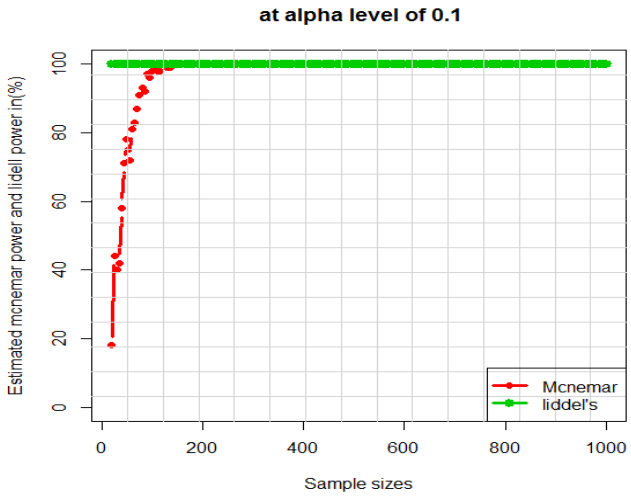

Fig. 3.20: A graphical display of the result for the power of both tests when P_{12} =0.5 and P_{21} = 0.4 at α=0.1

It can be seen that Liddell's exact test has 100% power for almost all sample sizes, while McNemar's test power increases as the sample size increases.

3.6 Third Scheme When α = 0.1

(i) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.2$ with effect size $\delta=0$

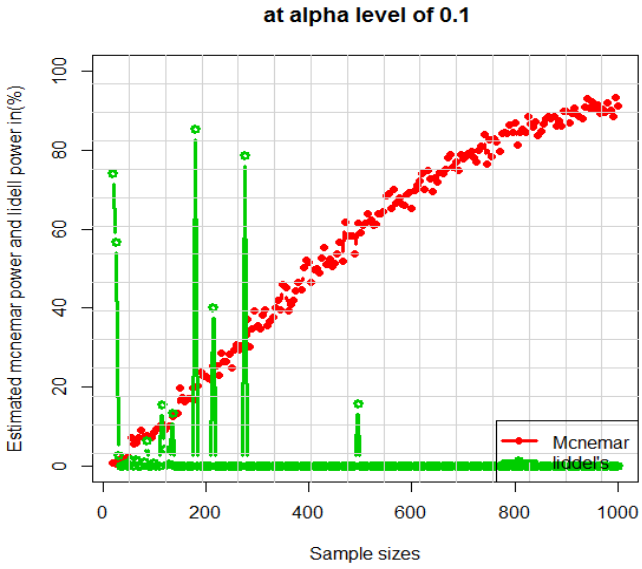

Fig. 3.21: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.2 at α=0.05

From the graph, it has been shown that the type I error committed by McNemar's test was low for a low sample size (≤250) but increased as the sample sizes increased, while the type I error committed by Liddell's exact test was low and unstable throughout the sample sizes considered.

(ii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.21$ with effect size $\delta=0.01$.

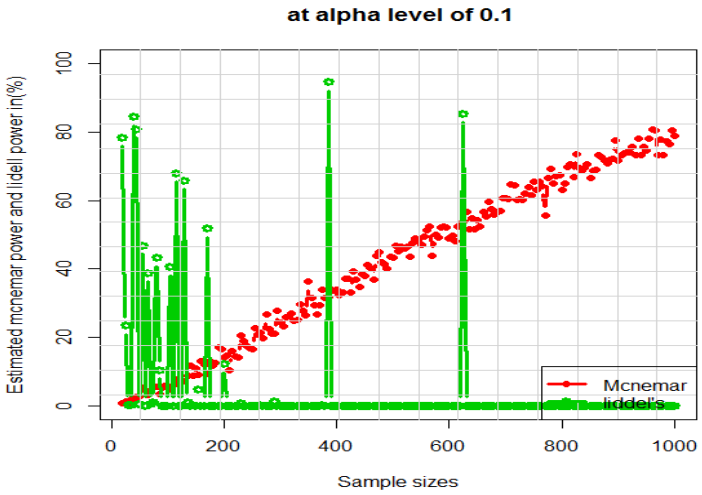

Fig. 3.22: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.21 at α=0.05

From the graph, it has been shown that the type I error committed by McNemar's test was low for a low sample size (≤500) but increases as the sample sizes increase. In contrast, the type I error committed by Liddell's exact test was high for the sample sizes ≤ 50, then low and unstable for the remaining sample sizes considered.

(iii) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.22$ with effect size $\delta=0.02$.

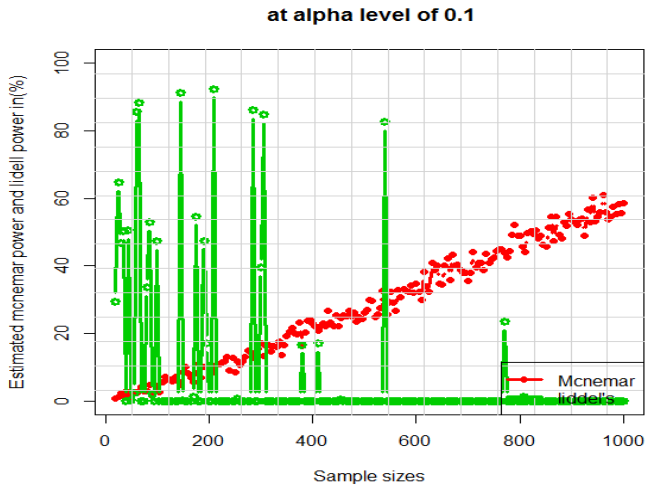

Fig. 3.23: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.22 at α=0.05

From the graph, it has been shown that the type I error committed by McNemar's test was low for a low sample size (≤750), while the type I error committed by Liddell's exact test was low and unstable for the sample sizes considered.

(iv) H_0: $P_{12}-P_{21}=\delta$ versus H_1: $P_{12}- P_{21}= \delta$, where $\delta=0$ under H_0, when $P_{12}=0.2$ and $P_{21}=0.23$ with effect size $\delta=0.03$.

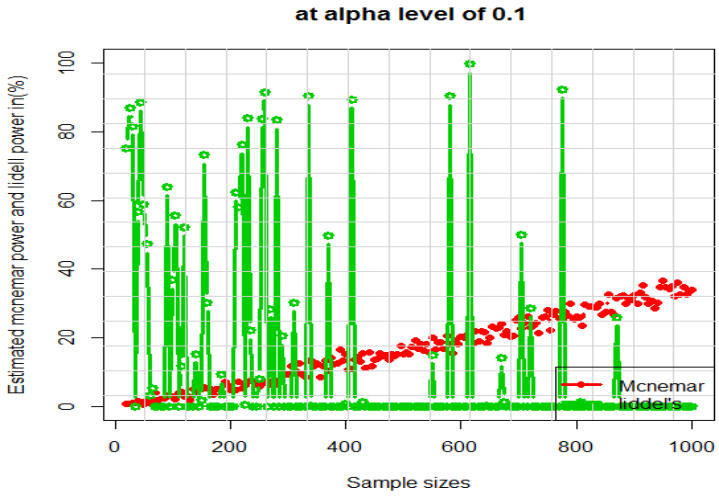

Fig. 3.24: A graphical display of the result for the type I error committed by both tests when P_{12} =0.2 and P_{21} = 0.23 at α=0.05

From the graph, it has been shown that the type I error committed by McNemar's test was low for all the sample sizes considered, while the type I error committed by Liddell's exact test was high for the sample sizes ≤ 50, then low and unstable for the remaining sample sizes considered.

Discussion

High Effect Sizes (0.6 and 0.8) at 0.01 Significance Level:

At a 0.01 level of significance, both tests exhibited low power for small sample sizes (≤ 20), particularly at an effect size of 0.6. As sample sizes increased, the power of McNemar's test improved, whereas Liddell's exact test consistently achieved 100% power across nearly all considered sample sizes.

Moderate Effect Sizes (0.1, 0.2, 0.3, 0.4) at 0.01 Significance Level:

For moderate effect sizes at the 0.01 significance level, both tests demonstrated poor power with small sample sizes (≤ 20). The power of McNemar's test increased with larger sample sizes, while Liddell's exact test maintained 100% power across nearly all sample sizes.

No and Low Effect Sizes (0, 0.01, 0.02, 0.03) at 0.01 Significance Level:

At a 0.01 level of significance with no or low effect sizes, McNemar's test exhibited shallow Type I error rates for small sample sizes (≤ 20), which increased with larger sample sizes. In contrast, Liddell's exact test showed high Type I error rates for small sample sizes (≤ 50), but these error rates decreased significantly for the larger sample sizes considered.

High Effect Sizes (0.6 and 0.8) at 0.1 Significance Level:

At a 0.1 level of significance, McNemar's test exhibited low power for small sample sizes (≤ 20), particularly at an effect size of 0.6. The power increased sharply, approaching 100% as the sample size increased. Liddell's exact test, however, consistently achieved 100% power across nearly all sample sizes.

Moderate Effect Sizes (0.1, 0.2, 0.3, 0.4) at 0.1 Significance Level:

For moderate effect sizes at the 0.1 significance level, Liddell's exact test demonstrated 100% power across nearly all sample sizes. In contrast, McNemar's test showed low power for small sample sizes (≤ 20), which increased rapidly, approaching 100% with larger sample sizes.

No and Low Effect Sizes (0, 0.01, 0.02, 0.03) at 0.1 Significance Level:

At a 0.1 level of significance with no or low effect sizes, McNemar's test showed shallow Type I

error rates for small sample sizes (≤ 20), which increased as sample sizes grew. Meanwhile, Liddell's exact test exhibited high Type I error rates for small sample sizes, which decreased and became less stable for larger sample sizes.

Conclusions

Based on the analysis conducted, several key conclusions can be drawn. Firstly, the power of both McNemar's and Liddell's exact tests is significantly influenced by sample size. Both tests exhibited improved performance with larger sample sizes compared to smaller ones. The level of significance also plays a critical role, as both tests demonstrated greater power at higher significance levels. Effect size is another critical factor; both tests were more potent at moderate to high effect sizes.

Furthermore, the preference between the two tests depends on specific conditions. Liddell's exact test is generally more suitable when dealing with large sample sizes and low effect sizes. However, McNemar's test may still be appropriate for situations involving smaller sample sizes with low effect sizes. Liddell's test is recommended when the proportions are small and close, especially when the sample size is large. Given these findings, it is crucial to carefully choose the appropriate test based on the specific characteristics of the data to achieve reliable results.

Conflict of Interest

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aberson, C. L. (2010). Applied power analysis for Behavioural Science.

Agresti, A. (2002). *Categorical data analysis*, Wiley, New York.

Barlow W. (1998). *Modelling of categorical agreement. The encyclopedia of biostatistics*, Wiley, New York.

Bayo A. Lawal (2012). Categorical Data Analysis with SAS and SPSS application.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. (2nd ed.) Lawrence Erlbaum Associates, New Jersey.

Ellis, P. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press. p. 52. ISBN 978-0521142465.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd edition), Wiley, New York.

Liddell, D. (1976). "Practical Tests of 2 x 2 Contingency Tables". *Journal of Royal Statistics Society* 25(4): 295 – 304

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika*, 30(3): 239-270.

McNemar, Q. (1947). "Note On the Sampling Error of the Difference between Correlated Proportions on Percentages". *Psychometrika* 12(2): 153 – 157

Sun, X.Z. (2008): "Generalized McNemar's test for Homogeneity of Marginal Distributions". SAS Global forum

Yates F. (1934). Contingency table involving small numbers and the X^2 test. Supplement to the Journal of the Royal Statistics Society 1(2): 217 – 235

Wikipedia (2011) Statistical power. The Free Encyclopedia of Biostatistics http:// en.m.wikipedia.org/wiki/statistical power

Williams, M. N. (2021). Levels of measurement and statistical analyses. *Meta-Psychology*, *5*.

HOW TO CITE

Akanni, J. O., & Aremu, Z. O. (2025). Comparative Analysis of Mcnemar's Test and Liddell's Exact Test for Assessing Marginal Homogeneity. KolaDaisi University Journal of Applied Sciences, 2, 87-102. https://doi.org/10.5281/zenodo.17438474