REVSTAT – Statistical Journal
Volume 21, Number 2, April 2023, 219–233
https://doi.org/10.57805/revstat.v21i2.403

Performance Comparison of Independence Tests in Two-Way Contingency Tables

Authors: Ebru Ozturk
– Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey
ebru.ozturk3@hacettepe.edu.tr

Merve Basol ✉
– Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
merve.basol@erciyes.edu.tr

Dincer Goksuluk
– Department of Biostatistics, Faculty of Medicine, Erciyes University, Kayseri, Turkey
dincergoksuluk@erciyes.edu.tr

Sevilay Karahan
– Department of Biostatistics, Faculty of Medicine, Hacettepe University, Ankara, Turkey
sevilaykarahan@gmail.com

Received: March 2021 Revised: November 2021 Accepted: November 2021

Abstract:

- Several test statistics are available for testing the independence of categorical variables from two-way contingency tables. A vast majority of published articles used the Pearson's chi-squared test for such purposes; however, this test statistic may lead to biased conclusions under certain conditions. Therefore, we aimed to compare the performance of test statistics via a comprehensive simulation study considering several factors in contingency tables. We also evaluated the performance of each test statistic on a real-life dataset. This study contributes to the literature guiding researchers to select an appropriate test statistic under different conditions.

Keywords:

- *contingency table; power; power divergence; simulation study; type one error.*

AMS Subject Classification:

- 62H17, 62E10.

✉ Corresponding author.

1. INTRODUCTION

The data type of measured variables is important to determine the statistical methods for summarizing and testing the relationship or independence between variables [9]. Analyzing categorical data is generally less tractable and may require much effort for selecting appropriate statistical methods, such as log-linear models, logistic regression, and chi-square tests. The contingency table approach is one of the frequently used methods to summarize the joint distribution of two categorical variables. An example of r-by-c contingency table showing the joint distribution of categorical variables X and Y is given in Table 1. Here, n_{ij} $(i = 1, 2, ..., r$ and $j = 1, 2, ..., c)$ represents the frequencies of joint occurrences, $n_{i+} = \sum_{j=1}^{c} n_{ij}$ and $n_{+j} = \sum_{i=1}^{r} n_{ij}$ are row and column totals (i.e., row/column marginals), and $n = \sum_{i=1}^{r} n_{i+} = \sum_{j=1}^{c} n_{+j} = \sum_{j=1}^{c} \sum_{i=1}^{r} n_{ij}$ is the grand total of contingency table that also refers to sample size.

Table 1: An example of r-by-c contingency table.

	Y_1	Y_2	$\cdots$	Y_c	Total
X_1	n_{11}	n_{12}	$\cdots$	n_{1c}	n_{1+}
X_2	n_{21}	n_{21}	$\cdots$	n_{2c}	n_{2+}
$\cdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	n_{r+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+c}	n

Specification of the joint probability distribution of Table 1 is crucial since it plays a key role in the type of statistical analysis used. The distribution of a contingency table may be one of multinomial, product multinomial, hypergeometric, and Poisson based on the cell counts that are fixed such that row/column marginals or totals. The inference about the independence between categorical variables can be evaluated using the appropriate sampling distribution and statistical hypotheses. The hypotheses for testing the independence of categorical variables in Table 1 is defined as

$$\begin{aligned} &H_0: \ \pi_{1j} = \pi_{2j} = ... = \pi_{rj}, \\ &H_1: \ \pi_{ij} \neq \pi_{kj} \quad \text{at least one } i, j, k, \quad i \neq k, \end{aligned} \tag{1.1}$$

where π_{ij} is the hypothesized cell probability of the i-th row and the j-th column, and $\hat{\pi}_{ij}$ is the estimated cell probability from sampling distribution. There are several methods for estimating cell probabilities, i.e., π_{ij}, and testing a hypothesis (1.1) depending on the joint distributions [13, 8].

Pearson's chi-square test statistic is widely used for testing the hypothesis (1.1). However, it is not a gold standard and may not be appropriate for small samples [1]. There exist various test statistics proposed to test the independence, where each performs better under certain conditions, such as sample size, number of rows and columns, sampling methods, etc. In this study, we used the most common of these methods, which are:

(i) Pearson's chi-square test;

(ii) likelihood ratio test;

(iii) Freeman–Tukey test;

(iv) Cressie–Read test;

(v) Fisher–Freeman–Halton's exact test.

A hypothesis established from a contingency table, considering the purpose of the study, could be tested using different statistical test procedures. The results of the hypothesis tests might be in the opposite direction for the variety of hypothesis tests. It is a crucial issue since it may mislead the researcher in their studies. Therefore, it is essential to choose appropriate statistical tests or methods to achieve correct and unbiased conclusions. In this study, we aimed to compare different test procedures and related test statistics under various scenarios for the power $(1-\beta)$ and the type-I error rate (α) of the test statistic. We conducted a comprehensive simulation study using the combinations of sample size, effect size, and sampling design. Furthermore, we applied each method to a real-life dataset for making a fair comparison between simulation and real-life data results. This study contributed to the literature by considering each test procedure under several conditions and comparing the performances of each test statistic via a comprehensive simulation study. Furthermore, the current study compared the simulation results with a real-life dataset and showed the concordance (or discordance) between the simulation study and the real-life example. All the analyses were performed on the R programming language (`https://cran.r-project.org/`) through self-written codes available upon request to the correspondent author.

The plan of this study is as follows. The methods, statistical background, simulation scenarios, and real datasets are explained in detail in the Material and Methods section. The results of simulated and real datasets are presented in the Results section, and finally, we discussed the results in the Discussion section with conclusions and future work.

2. MATERIAL AND METHODS

The statistical methods proposed to test the hypothesis (1.1) are detailed in subsection 2.1. These methods use the observed (n_{ij}) and expected (E_{ij}) frequencies to compute test statistics. All test statistics are asymptotically chi-square distributed with degrees of freedom $(r-1)(c-1)$.

2.1. Test Statistics

The most common test statistic proposed to test independence between categorical variables is the Pearson's chi-square statistic [1],which takes the difference between observed and expected frequencies into account. The test statistic (χ^2) is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-E_{ij})^2}{E_{ij}}. \tag{2.1}$$

The likelihood ratio test statistic is another approach to test independence [1]. Unlike Pearson's chi-square statistic, it is based on the ratio of the observed and expected frequencies. The test statistic is

$$G^2 = 2 \times \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \times \log\left(\frac{n_{ij}}{E_{ij}}\right). \tag{2.2}$$

The Freeman and Tukey test statistic aims to approximate Binomial or Poisson distribution to normal distribution by stabilizing the variance [7, 2]. It is based on the differences between the square root of observed and expected frequencies. The test statistic is

$$\mathrm{FT}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \left(\sqrt{n_{ij}} + \sqrt{n_{ij}+1} - \sqrt{4 \times E_{ij} + 1}\right)^2. \tag{2.3}$$

Cressie and Read [4] proposed the power divergence family as a generalization of goodness-of-fit test. It is flexible and converges to other well-known test statistics based on the choice of tuning parameter λ. The family of power divergence test statistic is

$$PD = \frac{2}{\lambda \times (\lambda+1)} \times \sum_{i=1}^{r} \sum_{j=1}^{c} \pi_{ij} \times \left[\left(\frac{n_{ij}}{E_{ij}}\right)^{\lambda} - 1\right]. \tag{2.4}$$

The power divergence test statistic converges to Pearson's chi-square, likelihood ratio, and Freeman–Tukey statistics when λ equals 1, 0 and 0.5, respectively. They [4] suggested taking λ as 2/3, called the Cressie–Read test statistic, as being an excellent compromise between Pearson's chi-square and likelihood ratio test statistics [4]. The test statistic is

$$\mathrm{CR} = \frac{9}{5} \times \sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij} \times \left[\left(\frac{n_{ij}}{E_{ij}}\right)^{2/3} - 1\right]. \tag{2.5}$$

In addition to the above-mentioned test statistics, we evaluated the Fisher–Freeman Halton (FFH) exact test statistic [6], which is the extension of Fisher's exact test to r-by-c tables. The Fisher–Freeman–Halton test statistic gives the exact p-value, which is calculated from sequentially generated contingency tables until one of the cells in the given margin is equal 0. This method becomes computationally intensive as the sample size increases. To overcome this problem, the Monte Carlo approach that selects samples randomly from the contingency tables is recommended [1]. In this study, we used large sample sizes. However, we benefited from the Monte-Carlo approach to decrease the computation time of the FFH test statistic.

2.2. Simulation Scenarios

We conducted a comprehensive simulation study using the R language environment [12]. We considered several factors such as sample size (n), effect size (w), and sampling design in the simulation. We used two different contingency tables, with dimensions of 5-by-5 and

5-by-2, in all simulation scenarios. Simulation scenarios consist of all possible combinations of:

- Sample size (n): $\{100, 200, 500\}$ for the 5-by-5 table and $\{40, 80, 200\}$ for the 5-by-2 table as *small*, *medium* and *large*, respectively,
- Effect size: (w): $\{0.10, 0.30, 0.50\}$ as *small*, *medium*, and *large* [3],
- Sampling design: balanced $(0.20, 0.20, 0.20, 0.20, 0.20)$,
  almost balanced $(0.15, 0.15, 0.20, 0.25, 0.25)$ and
  imbalanced row margins $(0.05, 0.05, 0.30, 0.30, 0.30)$,

where different sample sizes were used for 5-by-5 and 5-by-2 contingency tables while effect sizes and sampling designs were similar. The sample sizes were chosen so that the contingency tables were not sparse. Furthermore, the effect sizes were specified as in the literature [3]. Data were generated under product multinomial distribution via an `R` package `rTableICC` [5] by setting row marginal and total sample size fixed. Cell probabilities were specified according to changing effect size and sampling design. We compared each method using type I error rate and power. Each simulation scenario was repeated $10,000$ times. Each generated contingency table was tested with the Pearson's chi-square test, likelihood ratio test, Freeman–Tukey test, Cressie–Read test, and Fisher–Freeman–Halton's exact test. The type-I error rate of each test statistic was calculated as the proportion of false rejection obtained from $10,000$ replications when the null hypothesis was true, i.e., the effect size is $w = 0$. The power of each test, on the other hand, was calculated as the proportion of rejection obtained from $10,000$ replications assuming that the null hypothesis was false, i.e., the effect size is $w \neq 0$. The power and type-I error rate of the Pearson's chi-square test, likelihood ratio test, Freeman–Tukey test, and Cressie–Read tests statistics were obtained using the underlying Chi-square distribution. The comparison for the result of the Fisher–Freeman–Halton's exact test was evaluated using a p-value against the level of statistical significance. The statistical significance was taken as $p < 0.05$ in all simulation scenarios.

2.3. Real-life datasets

In addition to the simulation study, we evaluated the selected methods on real datasets. The first of the datasets is related to suicides. Suicides adversely affect not only the person who committed suicide, but also the people around the person, communities, and countries. According to the World Health Organization [17], suicide leads to a serious public health issue. Therefore, we decided to examine the specific causes of suicide within education level in Turkey in the year 2018. The datasets were provided by the Turkish Statistical Institute [15] and are represented in Table 2.

Nowadays, one of the major issues in the world, which is the infection of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2), also known as COVID-19, has led to the global pandemic. Therefore, another dataset, which is taken from Ozsurekci *et al.* [10], was chosen to be used in this study. The children who were infected with or exposed to COVID-19 might have developed multisystem inflammatory syndrome (MIS-C) due to the triggering of the immune system. They compared children with MIS-C ($n = 30$) and severe/critical cases with COVID-19 ($n = 22$) in terms of respiratory support systems. This information is given in Table 3.

Table 2: Contingency table between causes of suicide and education level.

Education Level	Causes				
	Marital Conflict	Financial Difficulty	Disease	Emotional	Other
Never received formal education	9 (7.14%)	4 (1.63%)	53 (7.91%)	4 (4.71%)	53 (6.31%)
Primary School	27 (21.43%)	53 (21.54%)	155 (23.13%)	10 (11.76%)	174 (20.71%)
Secondary School	60 (47.62%)	74 (30.08%)	197 (29.40%)	37 (43.57%)	269 (32.02%)
High School	22 (17.93%)	81 (32.93%)	170 (25.37%)	23 (27.06%)	205 (24.40%)
Graduate	8 (6.35%)	34 (13.82%)	95(14.18%)	11 (12.92%)	139 (16.55%)

Table 3: Contingency table between disease group and respiratory support system.

Respiratory Support	Group	
	Cases with MIS-C	Severe/Critical cases with COVID-19
None	14 (46.67%)	6 (27.27%)
Oxygen Only	7 (23.33%)	8 (36.36%)
High Flow Support	0 (0.00%)	2 (9.09%)
Non-invasive ventilation	6 (20.00%)	0 (0.00%)
Invasive mechanical ventilation	3 (10.00%)	6 (27.27%)

3. RESULTS

The performance of the test statistics was compared according to type-I error rate and power. The power of test statistics were presented in Figures 1 and 2 while the type-I error rates were presented in Figures 3 and 4[1]. In each figure, effect sizes and sampling designs were given in the rows and columns, respectively. The test statistics were given on the x-axis and the sample size was indicated using different line type within each figure. Although we graphically presented the power and type-I error rate results in Figures 1–4, it was not easy to read exact values from corresponding figures when the points and lines were overlapped or test statistics slightly differed. Therefore, we provided the findings of Figures 1–4 with supplementary tables in the Appendix section.

When the power results are examined in Figures 1 and 2 (Tables 5 and 6 in the Appendix) for 5-by-5 and 5-by-2 contingency tables, we observe that both the effect size and the sample size have a positive effect on power of test statistics. The statistical power of methods increases with the increasing sample size and effect size. However, the sampling design has no or a considerably small effect on power for each method. Among the methods considered, the likelihood ratio test has the highest power in almost all scenarios. The Pearson's chi-square and the Cressie–Read test statistics had less power in almost all designs when the sample size was small. The power of Freeman–Tukey test decreased as the sampling design became imbalanced. We also observed that the power of the Fisher–Freeman–Halton test was higher in the imbalanced design, except for the likelihood ratio test.

[1] Figures were generated using the `ggplot2` [16] package in the `R` programming language.

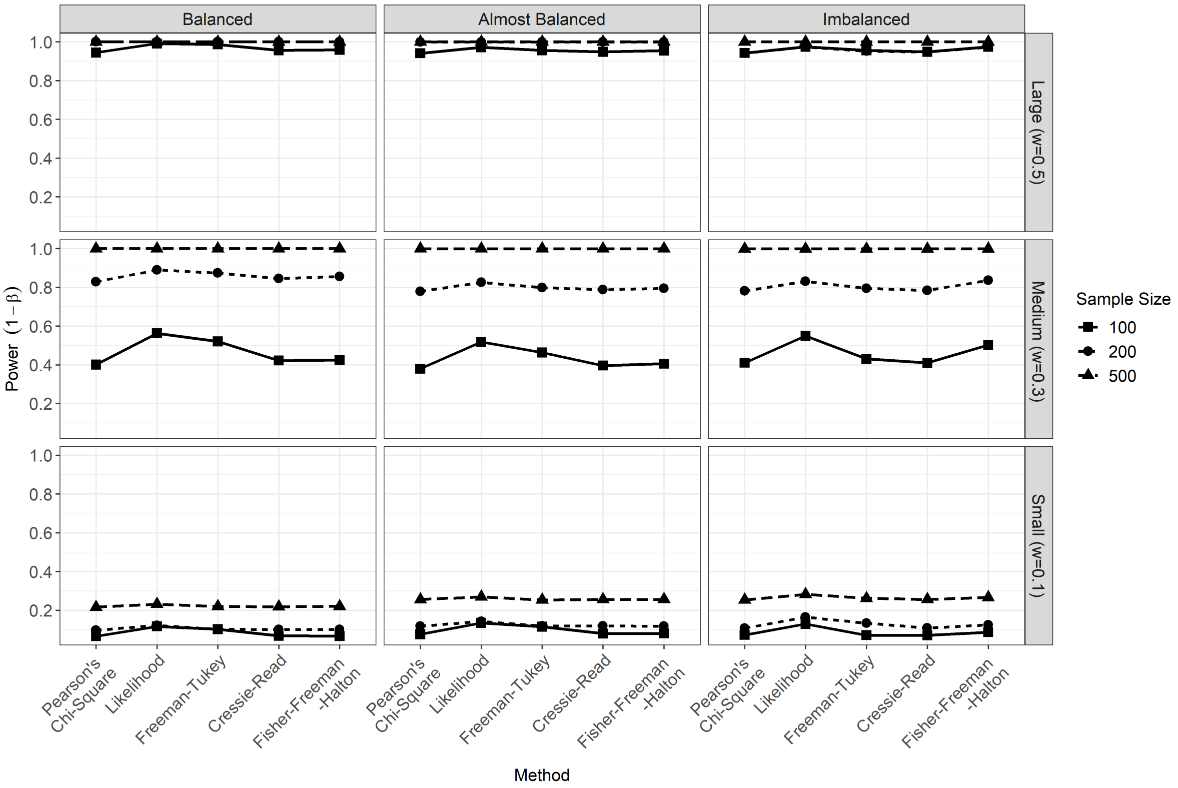

Figure 1: Simulation results – Power of tests in 5-by-5 contingency table.

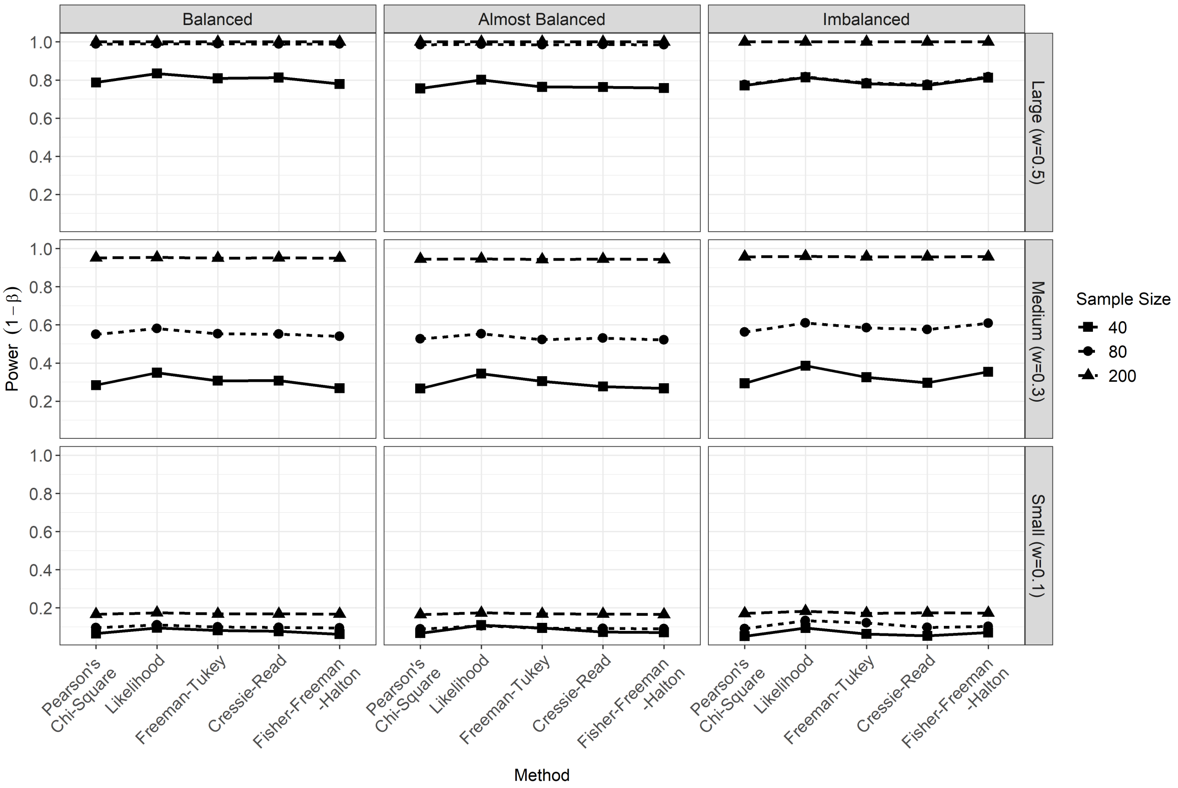

Figure 2: Simulation results – Power of tests in 5-by-2 contingency table.

The type-I error rate results of the 5-by-5 and 5-by-2 tables are given in Figures 3 and 4 (Tables 7 and 8 in the Appendix). According to the results, the likelihood ratio test was generally liberal generating type I error rates above the nominal level. Nevertheless, we observed that the type-I error rate of the likelihood ratio test was close to the nominal level as the sample size increased. In the balanced sampling design with the larger sample sizes, the type-I error rate of all test statistics, except for the likelihood ratio test statistic, was close to the nominal level. The Freeman–Tukey test statistic had a remarkably higher type-I error rate than the nominal level in small samples for balanced and almost balanced designs. However, it had the lowest type-I error rate below the nominal level in the imbalanced sampling design with a small sample size. In balanced and almost balanced designs, the Pearson's chi-square test, Cressie–Read test, and Fisher–Freeman–Halton test were better at controlling the type-I error rate at the nominal level in almost all sample sizes. However, in the imbalanced sampling design, Cressie–Reed and Pearson's chi-square test statistics were generally conservative for the small sample size and had type-I error rates closer to the nominal level as the sample size increased. Finally, the Fisher–Freeman–Halton test statistic had type-I error rates very close to the nominal level for the imbalanced sampling design.

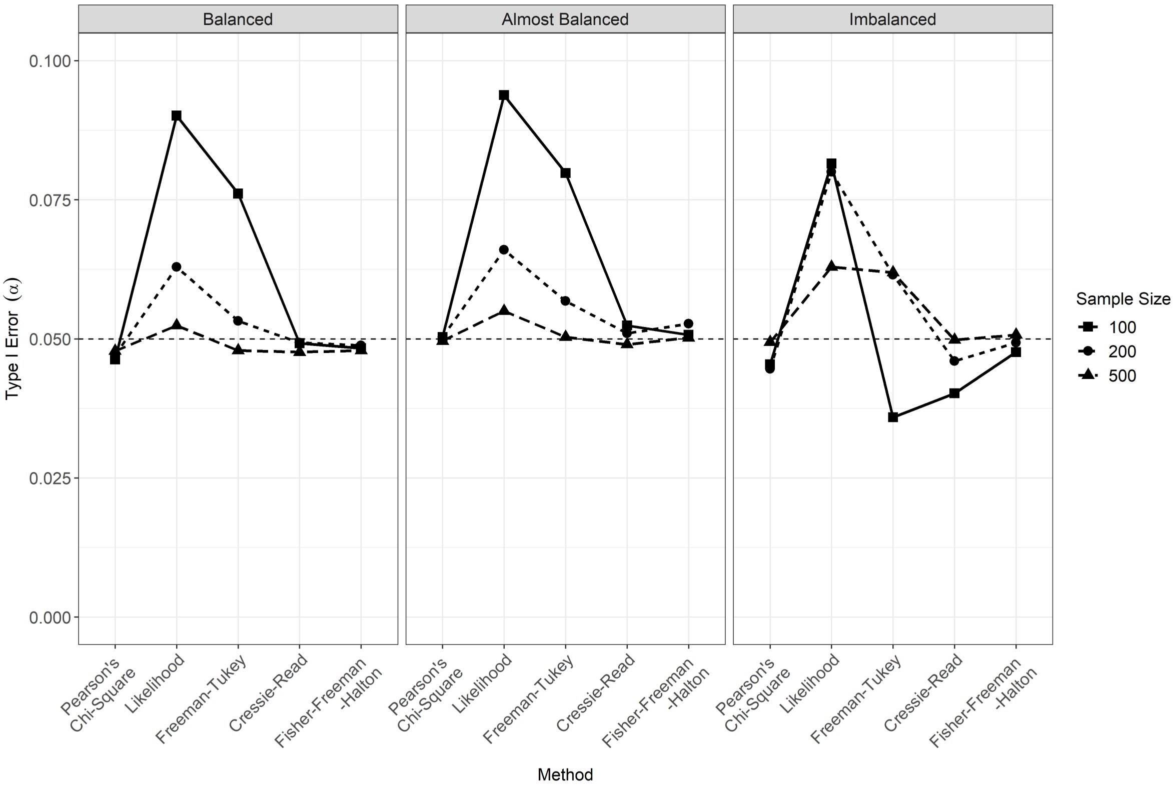

Figure 3: Simulation results – Type I error rates in 5-by-5 contingency table.

The results of real datasets are represented in Table 4. The suicide dataset (Table 2) had small effect size (i.e., $w = 0.16$), large sample size (i.e., $n = 1967$), and imbalanced design according to the row probabilities (i.e., $0.063, 40.2131, 0.324, 0.255$, and 0.146). Therefore, the suicide dataset corresponds to the simulation combination that was small effect size, large sample size, and imbalanced sampling design with the 5-by-5 table (bottom-right panel of Figure 1).

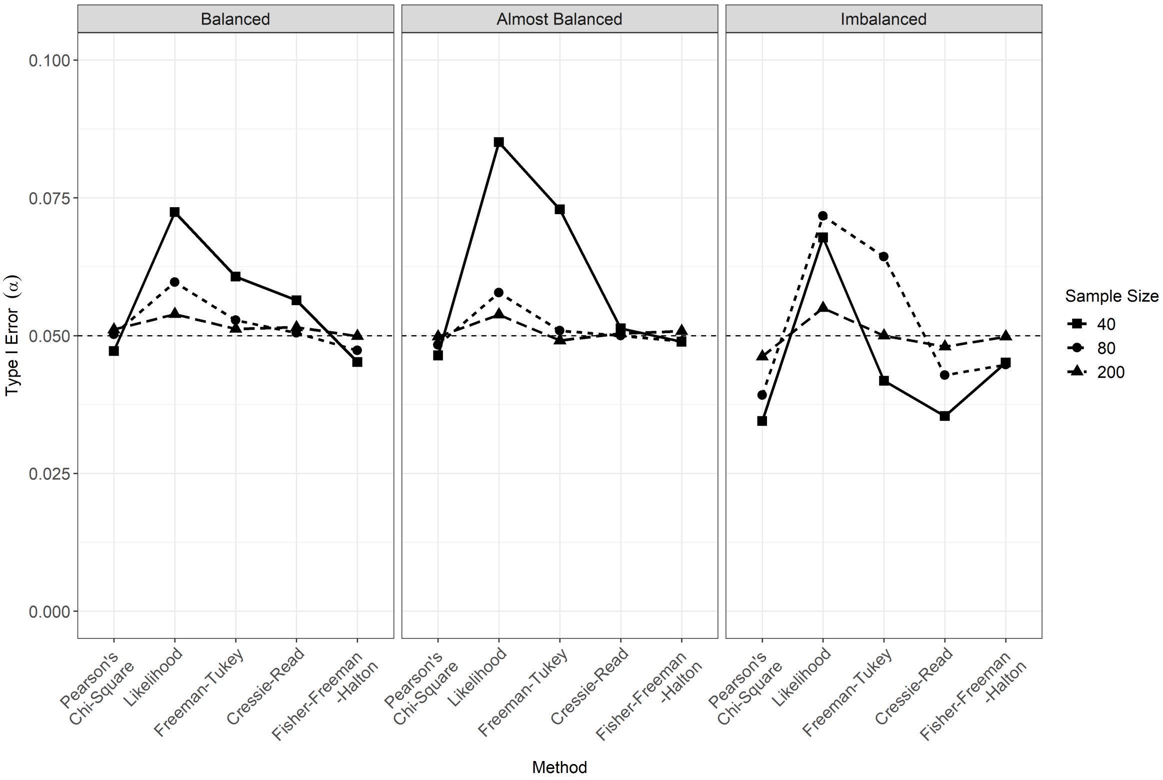

Figure 4: Simulation results – Type I error rates in 5-by-2 contingency table.

Although we found a statistically significant association between education level and suicide ($p < 0.001$ for all test statistics), the degree of association was not high ($w = 0.16$). Under this simulation scenario, the power of the Pearson and Cressie–Read test statistics was lower than the likelihood ratio test, which was similar to the real dataset results. On the other hand, the COVID-19 dataset had a large effect size (i.e., $w = 0.46$), small sample size (i.e., $n = 52$), and imbalanced sampling design according to the row probabilities (i.e., $0.385, 0.289, 0.039, 0.115$, and 0.173). This dataset corresponds to the simulation combination of large effect size, small sample size, and imbalanced sampling design with the 5-by-2 table (upper-right panel of Figure 2). In the COVID-19 dataset, all test statistics found a significant association between disease group and respiratory support system. According to the simulation results, there were slight differences between methods under a similar scenario in the COVID-19 dataset. Nonetheless, the power of likelihood ratio and Fisher–Freeman–Halton test statistics were higher than other methods. We observed results similar to simulation results in the COVID-19 dataset. The power of the likelihood ratio test statistic was the highest as compared to other methods. In addition, we saw that the Freeman–Tukey and Fisher–Freeman–Halton tests were almost similar to the likelihood ratio test.

Table 4: Results of real datasets.

Datasets	Methods								
	χ^2	p-value	G^2	p-value	FT^2	p-value	CR	p-value	FFH
Causes / Education level	48.66	<0.001	52.75	<0.001	54.01	<0.001	49.67	<0.001	0.001
Res. Support / Group	11.30	0.023	14.23	0.007	13.94	0.007	11.74	0.019	0.016

4. DISCUSSION

Previous studies in the literature evaluated the performance of various test statistics for r-by-c contingency tables. Rudas [14] compared the Pearson's chi-square, Cressie–Read, and likelihood ratio statistics for 2-by-2 and 3-by-3 tables. They reported that the Pearson's chi-square test statistic outperformed the likelihood ratio test when the sample size was small. Furthermore, they showed that the Cressie–Read and Pearson's chi-square test statistics had similar results. Parshall *et al.* [11] conducted a Monte Carlo simulation study to compare the type-I error rate and power of Pearson's chi-square, likelihood ratio, and Cressie–Read test statistics. They generated datasets from uniform distribution and found that the likelihood ratio test statistic failed to control the type I error rate at the nominal level. In addition to the previously published studies, this study considered the effects of sample size, effect size, and sampling design on the performance of various test statistics of contingency tables. A comprehensive simulation study were conducted and the findings showed that (Figures 1–4):

- The effect size and sample size were positively associated with the power of tests. The statistical power of each method increased as the number of samples or effect size increased.
- Sampling design did not affect the power of tests or slightly changed it.
- The likelihood ratio test had higher type-I error rates than the nominal level in almost all simulation scenarios. However, its statistical power was higher than other methods. We concluded that the likelihood ratio test was generally liberal, and the rejected null hypothesis should be validated using alternative methods.
- The Pearson's chi-square and Cressie–Read statistics had similar results in almost all scenarios. We mainly suggest these methods for balanced or almost balanced sampling designs when the sample size is large.
- The Fisher–Freeman–Halton (FFH) test had similar results with Pearson's chi-square and Cressie–Read tests in balanced sampling designs. However, results were promising and better than other methods in the imbalanced sampling designs. Hence, we suggest using the FFH test when the sampling design is imbalanced.
- The Freeman–Tukey (FT) test had decreased power as the sampling design became imbalanced. Even the type-I error rate was higher than the nominal level, except for the imbalanced sampling design with a small sample size, the FT test was better at controlling the type-I error rate than the likelihood ratio test.

To test the independence between variables in two-way contingency tables, one should be aware of the sampling design, the sample size, and the effect size. The power and type-I error rate are affected by those factors. The Pearson's chi-square test is a frequently used method for testing the independence in two-way contingency tables. However, we showed in our study that the Cressie–Read and Fisher–Freeman–Halton tests are efficient alternatives to the Pearson's chi-square test since they are good at controlling type-I error rates at the nominal level under certain conditions. Moreover, the power of these test statistics is as good as or better than the Pearson's chi-square test statistic. Therefore, researchers should consider the effect of the above-mentioned factors before selecting the appropriate test statistic for testing the independence in a contingency table.

Another significant issue in the analysis of the contingency tables is whether there are cells with zero observed frequencies and expected frequencies below 5. These cell frequencies affect the choice of the appropriate test statistic. In this study, we counted both the number of cells with zeros and the cells with an expected value of less than 5 for $10,000$ replication data in each simulation scenario. The average number of cells with zeros and the average number of cells with expected counts below 5 were calculated specifically in the small sample size and imbalanced design for both 5-by-5 and 5-by-2 tables. The average number of cells with zeros was 4 (16%) and the average number of cells with the expected value less than 5 was 14 (56%) in the 5-by-5 tables. For the 5-by-2 tables, these values were 1 (10%) and 5 (50%), respectively. The amount of cells with lower expected counts were in the majority as expected. However, the amount of zero inflation were slight to moderate in some of simulation scenarios. This study did not account for the effect of zero inflation since it was not severe in the generated datasets. However, the effect of zero inflation should carefully considered before selecting an appropriate test statistic in contingency tables. Lydersen [8] indicated that when no more than 20 percent of the cells have an expected value below 5, the Fisher's exact test was recommended. In this study, for the small sample size and imbalanced design, we also observed that the performance of the Fisher–Freeman–Halton test statistic was better than other test statistics according to the both type-I error level and power. Therefore, we observed that simulation results are concordant with the literature [8]. As a result, for a small sample size with an imbalanced sampling design, we could say that the Fisher–Freeman–Halton test statistic is more convenient for these conditions when considering the results.

This study considered two-way contingency tables with dimensions 5-by-5 and 5-by-2. In practice, researchers wish to work with contingency tables with lower dimensions due to simplicity and less sample size. However, one may be required to work with a contingency table having rows or columns above three. For example, in medical sciences, a binary response variable such as death versus alive or healthy versus diseased might be compared between five groups which can be summarized in a contingency table with dimensions 5-by-2. Furthermore, a response variable with five categories like a 5-point Likert scale or reasons of suicides as in Table 2 might be associated with another categorical variable with five categories such as the education level. Although high-dimensional contingency tables are not frequently used or preferred in researches, they may have to be used in some studies. Therefore, the performance of test statistics in high-dimensional contingency tables should be carefully considered for selecting an appropriate test statistic. Our study provided detailed results of test statistics in high-dimensional contingency tables. Furthermore, this study can be extended to a more general case by considering the dimension of contingency tables as a new factor in the simulation scenarios.

The problem of selecting the appropriate method for testing the independence in a contingency table is not a recent topic; however, it is an ongoing issue since the performance of each method is unclear for most of the scenarios. In this study, we conducted a comprehensive simulation study considering several factors, and compared the simulation results with real data examples. We aimed to provide comparative results and bring attention to other statistical methods than Pearson's chi-square test, which is the most common in practice. We highlighted that researchers should consider various factors such as sampling design, sample size, and effect size before selecting the statistical procedures to test the independence in contingency tables. Although we covered many scenarios in the simulation study, there still exist scenarios that are not covered and the performances are unclear. Our study was not able to reflect the performance of selected methods in sparse contingency tables. We leave this topic for further research.

APPENDIX

Table 5: Simulation results – Power of tests in 5-by-5 contingency table.

Effect Size	Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Low ($w = 0.1$)	Balanced	100	0.0658	0.1165	0.1018	0.0677	0.0667
		200	0.0968	0.1214	0.1024	0.1011	0.1002
		500	0.2161	0.2315	0.2193	0.2179	0.2200
	Almost Balanced	100	0.0757	0.1346	0.1151	0.0791	0.0801
		200	0.1172	0.1416	0.1189	0.1187	0.1169
		500	0.2550	0.2688	0.2531	0.2557	0.2561
	Imbalanced	100	0.0723	0.1289	0.0709	0.0703	0.0868
		200	0.1069	0.1638	0.1332	0.1084	0.1245
		500	0.2531	0.2820	0.2616	0.2559	0.2658
Medium ($w = 0.3$)	Balanced	100	0.4006	0.5628	0.5205	0.4221	0.4244
		200	0.8280	0.8898	0.8742	0.8449	0.8556
		500	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	100	0.3793	0.5174	0.4635	0.3958	0.4053
		200	0.7792	0.8260	0.7988	0.7874	0.7940
		500	0.9990	0.9991	0.9992	0.9990	0.9990
	Imbalanced	100	0.4104	0.5494	0.4300	0.4096	0.5020
		200	0.7810	0.8312	0.7939	0.7839	0.8366
		500	0.9985	0.9986	0.9985	0.9986	0.9988
Large ($w = 0.5$)	Balanced	100	0.9448	0.9910	0.9867	0.9566	0.9586
		200	1.0000	1.0000	1.0000	1.0000	1.0000
		500	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	100	0.9412	0.9713	0.9565	0.9487	0.9547
		200	0.9999	0.9999	0.9999	0.9999	0.9999
		500	1.0000	1.0000	1.000	1.0000	1.0000
	Imbalanced	100	0.9423	0.9745	0.9560	0.9482	0.9745
		200	0.9421	0.9734	0.9522	0.9477	0.9738
		500	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6: Simulation results – Power of tests in 5-by-2 contingency table.

Effect Size	Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Low $(w=0.1)$	Balanced	40	0.0656	0.0950	0.0819	0.0769	0.0619
		80	0.0964	0.1103	0.1002	0.0971	0.0946
		200	0.1661	0.1743	0.1683	0.1681	0.1680
	Almost Balanced	40	0.0673	0.1089	0.0948	0.0735	0.0713
		80	0.0899	0.1055	0.0936	0.0921	0.0906
		200	0.1653	0.1735	0.1684	0.1679	0.1656
	Imbalanced	40	0.0514	0.0949	0.0634	0.0531	0.0703
		80	0.0912	0.1333	0.1211	0.0966	0.1027
		200	0.1709	0.1826	0.1720	0.1737	0.1729
Medium $(w=0.3)$	Balanced	40	0.2837	0.3494	0.3071	0.3074	0.2674
		80	0.5500	0.5809	0.5527	0.5521	0.5403
		200	0.9513	0.9534	0.9504	0.9515	0.9502
	Almost Balanced	40	0.2663	0.3442	0.3044	0.2768	0.2672
		80	0.5260	0.5529	0.5231	0.5306	0.5216
		200	0.9449	0.9462	0.9430	0.9452	0.9429
	Imbalanced	40	0.2932	0.3855	0.3250	0.2962	0.3539
		80	0.5625	0.6094	0.5853	0.5755	0.609
		200	0.9567	0.9595	0.9571	0.9571	0.9575
Large $(w=0.5)$	Balanced	40	0.7870	0.8343	0.8088	0.8124	0.7790
		80	0.9890	0.9907	0.9894	0.9891	0.9888
		200	1.0000	1.0000	1.0000	1.0000	1.0000
	Almost Balanced	40	0.7558	0.8012	0.7636	0.7630	0.7579
		80	0.9852	0.9868	0.9849	0.9856	0.9848
		200	1.0000	1.0000	1.0000	1.0000	1.0000
	Imbalanced	40	0.7710	0.8137	0.7815	0.7723	0.8123
		80	0.7758	0.8171	0.7850	0.7766	0.8169
		200	1.0000	1.0000	1.0000	1.0000	1.0000

Table 7: Simulation results – Type I error rates in 5-by-5 contingency table.

Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Balanced	100	0.0463	0.0901	0.0761	0.0492	0.0483
	200	0.0471	0.0629	0.0532	0.0493	0.0488
	500	0.0478	0.0524	0.0479	0.0476	0.0479
Almost Balanced	100	0.0503	0.0938	0.0798	0.0524	0.0507
	200	0.0503	0.0660	0.0568	0.0510	0.0527
	500	0.0496	0.0550	0.0503	0.0490	0.0502
Imbalanced	100	0.0454	0.0815	0.0359	0.0402	0.0476
	200	0.0446	0.0800	0.0615	0.0460	0.0493
	500	0.0494	0.0629	0.0619	0.0498	0.0507

Table 8: Simulation results – Type I error rates in 5-by-2 contingency table.

Sampling Design	Sample Size	Pearson's Chi-Square	Likelihood	Freeman –Tukey	Cressie –Read	Fisher–Freeman –Halton
Balanced	40	0.0472	0.0724	0.0607	0.0564	0.0452
	80	0.0502	0.0597	0.0528	0.0505	0.0473
	200	0.0511	0.0539	0.0512	0.0515	0.0499
Almost Balanced	40	0.0464	0.0851	0.0729	0.0513	0.0489
	80	0.0483	0.0578	0.0509	0.0500	0.0490
	200	0.0498	0.0538	0.0491	0.0504	0.0508
Imbalanced	40	0.0345	0.0678	0.0418	0.0354	0.0451
	80	0.0392	0.0717	0.0643	0.0428	0.0447
	200	0.0462	0.0550	0.0500	0.0480	0.0498

ACKNOWLEDGMENTS

The authors are grateful and would like to thank the reviewers for their valuable comments that have increased the quality of our manuscript.

REFERENCES

[1] Agresti, B. (2002). *Categorical Data Analysis*, John Wiley & Sons, New Jersey.

[2] Bishop, Y.M.M. and Mosteller, F. (1969). *Smoothed contingency-table analysis.* In "National Halothane Study: a Study of the Possible Association Between Halothane Anesthesia and Postoperative Hepatic Necrosis" (J.P. Bunker; W.H. Forrest; F. Mosteller and L.D. Vandam, Eds.), National Research Council, Washington, DC, The National Academies Press, 237–286.

[3] Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, Routledge, United Kingdom.

[4] Cressie, N.A.C. and Read, T.R.C. (1984). Multinomial goodness-of-fit tests, *Journal of the Royal Statistical Society Series B (Methodological)*, **46**(3), 141–149.

[5] Demirhan, H. (2016). rTableICC: an R package for random generation of 2×2×K and R×C contingency tables, *The R Journal*, **8**(1), 48–63.

[6] Freeman, G.H. and Halton, J.H. (1951). Note on an exact treatment of contingency, goodness of fit and other problems of significance, *Biometrika*, **38**(1/2), 141–149.

[7] Freeman, M.F. and Tukey, J.W. (1950). Transformations related to the angular and the square root, *The Annals of Mathematical Statistics*, **21**(4), 607–611.

[8] Lydersen, P.S.; Senchaudhuri, P.V. and Laake, P. (2007). Choice of test for association in small sample unordered r × c tables, *Statistics in Medicine*, **26**(23), 4328–4343.

[9] Oyeyemi, G.M.; Adewara, A.A.; Adebola, F.B. and Salau, S.I. (2010). On the estimation of power and sample size in test of independence, *Asian Journal of Mathematics & Statistics*, **3**(3), 139–146.

[10] Ozsurekci, Y.; Gürlevik, S.; Kesici, S.; Akca, U.K.; Oygar, P.D.; Aykac, K.; Karacanoglu, D.; Saritas, N.O.; Ilbay, S.; Katlan, B.; Ertugrul, İ.; Cengiz, A.B.; Basaran, O.; Cura, Y.B.C.; Karakaya, J.; Bilginer, Y.; Bayrakci, B.; Ceyhan, M. and Ozen, S. (2021). Multisystem inflammatory syndrome in children during the COVID-19 pandemic in Turkey: first report from the Eastern Mediterranean, *Clin Rheumatol*, **12**, 1–11.

[11] Parshall, C.D.; Kromrey, J.D. and Dailey, R. (1999). Comparative performance of three statistical tests of homogeneity for sparse i × j contingency tables, *Communications in Statistics – Simulation and Computation*, **28**(1), 275–289.

[12] R Core Team. (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Austria.
http://www.R-project.org/

[13] Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag, New York.

[14] Rudas, T. (1986). A Monte Carlo comparison of the small sample behaviour of the Pearson, the likelihood ratio and the Cressie–Read statistics, *Journal of Statistical Computation and Simulation*, **24**(2), 107–120.

[15] TURKSTAT. (2019). *TURKSTAT: Distribution of selected causes of suicides with respect to gender and education level in 2018*, Turkish Statistical Institute.

[16] Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag, New York.

[17] World Health Organization. (2019). *World Health Organization: Suicide*, World Health Organization.