

Article

Homogeneity Test of the First-Order Agreement Coefficient in a Stratified Design

Mingrui Xu, Zhiming Li , Keyi Mou and Kalakani Mohammad Shuaib

College of Mathematics and System Science, Xinjiang University, Urumqi 830046, China

* Correspondence: zmli@xju.edu.cn

Abstract: Gwet's first-order agreement coefficient (AC_1) is widely used to assess the agreement between raters. This paper proposes several asymptotic statistics for a homogeneity test of stratified AC_1 in large sample sizes. These statistics may have unsatisfactory performance, especially for small samples and a high value of AC_1 . Furthermore, we propose three exact methods for small pieces. A likelihood ratio statistic is recommended in large sample sizes based on the numerical results. The exact E approaches under likelihood ratio and score statistics are more robust in the case of small sample scenarios. Moreover, the exact E method is effective to a high value of AC_1 . We apply two real examples to illustrate the proposed methods.

Keywords: Gwet's AC_1 ; homogeneity test; asymptotic statistic; exact method

1. Introduction

In the medical field, it is necessary to judge the accuracy and the interchangeability of different diagnostics. Inter-rater agreement is widely used to quantify the closeness of ratings for subjects by two raters. The recommendation of an efficient and economical method should guarantee a high degree of agreement between its result and the gold-standard method. A simple example is that independent raters A and B assess each subject with binary outcomes (e.g., $+/ -$, and Yes/No). Let $n_{ij}(i, j = 1, 2)$ be the numbers of independent subjects judged by two raters as $(+, +)$, $(-, +)$, $(+, -)$, and $(-, -)$, and $P_{ij}(i, j = 1, 2)$ be the corresponding probabilities, respectively. Denote $n = n_{1+} + n_{2+} = n_{+1} + n_{+2} = \sum_{i=1}^2 \sum_{j=1}^2 n_{ij}$. The data can be arranged into a 2×2 original table (Table 1).

Table 1. A 2×2 original table.

Rater		Rater A		Total
		+	−	
Rater B	+	$n_{11}(P_{11})$	$n_{12}(P_{12})$	$n_{1+}(P_B)$
	−	$n_{21}(P_{21})$	$n_{22}(P_{22})$	$n_{2+}(1 - P_B)$
Total		$n_{+1}(P_A)$	$n_{+2}(1 - P_A)$	n

Researchers have developed several indices by which to measure the degree of agreement between raters on a nominal scale category, where the unordered categories are independent, mutually exclusive, and exhaustive. Denote $P_i = \Pr(\text{the probability is classified as "+" by Rater } i)$, $i = A, B$. We call $P_i(i = A, B)$ as the marginal probability. Cohen [1] showed that the χ^2 test was indefensible because of the null hypothesis with independence, not agreement. Furthermore, he presented the kappa coefficient to compute the extent of agreement between raters. For the problem of nominal scale agreement between raters A and B in Table 1, there are only two relevant quantities: the overall agreement probability p_a , and the chance-agreement probability p_c . Cohen's kappa coefficient is defined by $\kappa = (p_a - p_c)/(1 - p_c)$, where $p_a = P_{11} + P_{22}$, $p_c = P_A P_B + (1 - P_A)(1 - P_B)$. That is to say, the coefficient κ is the



Citation: Xu, M.; Li, Z.; Mou, K.; Shuaib, K.M. Homogeneity Test of the First-Order Agreement

Coefficient in a Stratified Design.

Entropy **2023**, *25*, 536. [https://](https://doi.org/10.3390/e25030536)

doi.org/10.3390/e25030536

Academic Editor: Richard D. Gill

Received: 22 January 2023

Revised: 13 March 2023

Accepted: 16 March 2023

Published: 20 March 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

proportion of agreement after the removal of chance agreement. Suppose that the distribution of proportions over the categories for the population is known and is taken to be equal for the judges. Thus, Scott [2] proposed π coefficient $\pi = (p_a - p_\pi)/(1 - p_\pi)$, where p_π is the percent agreement to be expected on the basis of chance, and $p_\pi = (\frac{P_A + P_B}{2})^2 + (1 - \frac{P_A + P_B}{2})^2$. Despite the wide range of applications, the limitations of these coefficients have two main aspects: (i) it highly depends on marginal probabilities [3], and (ii) it is often affected by the composition of the population for subjects easy or difficult to agree upon [4]. For example, Cicchetti and Feinstein [3] illustrated one of the limitations by an example: $n_{11} = 118$, $n_{12} = 5$, $n_{21} = 2$, and $n_{22} = 0$. Through simple calculation, $p_a = 118/125 + 0/125 = 0.944$, $p_c = (123/125) \times (120/125) + (5/125) \times (2/125) = 0.9453$. Thus, the estimator $\hat{\kappa} = (0.944 - 0.9456)/(1 - 0.9456) = -0.0234 < 0$. For Scott's π coefficient, we have $\hat{\pi} = -0.0288$ because $p_\pi = 0.9456$. It is unreasonable that a high agreement has low κ and π coefficients. To solve the problem, some alternative indices have been derived to measure the consistency, such as Holley and Guilford's G index [5], Aickin's α agreement parameter [6], Andrés and Marzo's delta measure [7]. Gwet [8] revealed the origin of these limitations and proposed the first-order agreement coefficient (AC₁) as an alternative index. The definition of this coefficient is based on two premises: (a) chance agreement occurs when at least one rater rates an individual randomly, and (b) only an unknown portion of observed ratings is subjected to randomness. Define two events

$$G = \{\text{both raters agree}\}, \quad R = \{\text{at least one rater performs random rating}\}.$$

Thus, the probability of agreement expected by chance can be defined by $p_e \triangleq P(G \cap R) = P(R)P(G|R)$. Generally, a random rating may classify an individual into either category with the same probability of $1/2$. Since agreement may occur in either type, we have $P(G|R) = 2 \times (1/2)^2 = 1/2$. As for the probability of random rating $P(R)$, a normalized measure of randomness (Ψ) is used to approximate it as follows,

$$P(R) \approx \Psi = \frac{\pi_+(1 - \pi_+)}{1/2(1 - 1/2)} = 4\pi_+(1 - \pi_+),$$

where π_+ represents the probability that a random rater classifies a randomly chosen individual into the "+" category. That is to say, p_e can be quantified by $p_e^* = P(G|R)\Psi = 2\pi_+(1 - \pi_+)$. Then, the AC₁ coefficient can be expressed as $\gamma = (p_a - p_e^*)/(1 - p_e^*)$,

where p_a denotes the agreement probability. In the above example, $\hat{\pi}_+ = (123/125 + 120/125)/2 = 0.9720$ and $p_e^* = 0.0544$. By the definition of AC₁, we have $\gamma = 0.9408$. Thus, the AC₁ coefficient is more consistent with the observed extent of agreement than Cohen's κ and Scott's π coefficients. There have been quite a few pieces of literature about agreement coefficients [9–11].

As with Scott's π coefficient, Ohyama [12] assumed that two raters have a common marginal probability, that is, $P_A = P_B \triangleq \pi_+$. Thus, $P_{12} = P_{21}$, and Table 1 can be simplified as Table 2. Define

$$X_{ij} = \begin{cases} 1, & \text{if rater } i \text{ classifies the subject } j \text{ into category } +, \\ 0, & \text{otherwise} \end{cases}$$

for $i = A, B, j = 1, 2, \dots, n$. Suppose that the underlying probability of classifying a subject depends not on raters but subjects, which is $P(X_{ij} = 1|j) = p_j$. We can obtain the overall agreement probability (p_a) based on the idea of Vanbelle and Albert [13]. The agreement can occur in "(+, +)" and "(-, -)", and the corresponding probabilities for j th subject are $P_{11j} = p_j^2$ and $P_{22j} = (1 - p_j)^2$, respectively. Thus, the agreement probability of two raters for the j th subject is $p_{aj} = p_j^2 + (1 - p_j)^2$. We denote the mean of positive classification probability as $E(p_j) = \sum_{j=1}^n p_j/n \triangleq \pi_+$, and the corresponding variance as $Var(p_j) = \sum_{j=1}^n (p_j - \pi_+)^2/n \triangleq \sigma^2$, where n is the size of the population. Then, the probabilities of "(+, +)" and "(-, -)" ratings over the population are $P_{11} = E(p_j^2) =$

$Var(p_j) + (E(p_j))^2 = \sigma^2 + \pi_+^2$ and $P_{22} = E((1 - p_j)^2) = 1 - 2\pi_+ + \sigma^2 + \pi_+^2$. Finally, the agreement probability over the population is $p_a = P_{11} + P_{22} = 1 + 2(\sigma^2 - \pi_+(1 - \pi_+))$. The AC_1 coefficient (γ) for a binary outcome judged by two raters is rewritten by

$$\gamma = \frac{p_a - p_e^*}{1 - p_e^*} = \frac{1 + 2\sigma^2 - 4\pi_+(1 - \pi_+)}{1 - 2\pi_+(1 - \pi_+)}. \quad (1)$$

Up until now, the application [14,15] and the statistical inference [12] of the AC_1 have been concentrated at the situation without stratification. However, the ignorance of confounding variables or covariates may lead to a biased conclusion. Researchers often stratify the data into multiple strata to control the influence of these factors. A stratified analysis is applied to evaluate the relationship between the nontreatment factors of a clinical trial (age, gender, or severity of disease, etc.) and agreement. A test of homogeneity is the first step of the stratified analysis. It is essential to analyze the factors that lead to heterogeneity when we reject the homogeneity hypothesis. Suppose K levels of the subject covariates are introduced into Table 2 for two raters with binary outcome, and the data can be arranged in a $3 \times K$ table of observed cell counts. Generally speaking, a sample can be classified as a large or small sample by the sample size. Hannah et al. [16] analyzed the data about the alcohol-drinking status of twins. A subject is categorised as nondrinker if he/she consumes less than 30 gm alcohol per week, and otherwise is a drinker. Thus, the binary outcome is the drinking status (drinker or nondrinker). A number of same-sex twins are stratified by zygosity, including monozygotic (MZ), and dizygotic (DZ). Nam [17] used the kappa index to investigate the agreement of alcohol-drinking status between twins. The data structure of male twins is shown in Table 3. The large-sample inference has been performed for the data type, including score, likelihood ratio, and Wald-type statistics [18]. Honda and Ohyama [19] proposed score and goodness-of-fit tests for the homogeneity test of stratified AC_1 . Unfortunately, both tests performed poorly due to the conservative or liberal type I error rates, especially for small sample sizes. Meanwhile, a high AC_1 may lead to conservative type I error rates for small and moderate sample sizes.

Table 2. A 2×2 table under $P_A = P_B$.

Category	Ratings	Frequency	Probability
1	(+, +)	n_{11}	P_{11}
2	(+, −) or (−, +)	$n_{12} + n_{21}$	$2P_{12}$
3	(−, −)	n_{22}	P_{22}
Total		n	1

Table 3. Agreement of alcohol drinking status between male twins stratified by zygosity.

Alcohol Drinking	Zygosity	
	MZ	DZ
Both	19	8
One	14	16
Neither	19	7
Total	52	31

Note: Both: (twin 1, twin 2) = (drinker, drinker); One: (twin 1, twin 2) = (drinker, nondrinker) or (nondrinker, drinker); neither: (twin 1, twin 2) = (nondrinker, nondrinker).

In practice, we often encounter small sample cases of agreement data, for example, a clinical trial about coronavirus disease 2019 (COVID-19) [20]. In this trial, the enzyme-linked immunosorbent assay (ELISA) and gold-standard methods are used to detect the novel coronavirus IgG and IgM antibodies, classifying each of them as either positive (+) or negative (−). ELISA positive criterion is that the sample's optical density (OD) value is greater than or equal to the critical value. The positive criterion of the gold-standard

method is the appearance of two colored bands. Table 4 lists the data stratified by the IgG and the IgM antibodies ($K = 2$), 17 patients in each group. Similar to Table 3, “One” entry corresponds to the number of “(+, −)” and “(−, +)”.

Table 4. Agreement between ELISA and gold-standard methods stratified by antibody type.

Number of Agreement	Antibody Type	
	IgG	IgM
Both	9	7
One	3	7
Neither	5	3
Total	17	17

Note: Both: (ELISA, gold-standard) = (+, +); One: (ELISA, gold-standard) = (+, −) or (−, +); neither: (ELISA, gold-standard) = (−, −).

Unfortunately, asymptotic test statistics do not apply to small data. Exact approaches are effective for small samples, such as Fisher’s exact test [21–23], and its extensions [24–26]. A conservative performance of Fisher’s exact method supported the appearance of other exact approaches. We note that there exist nuisance parameters in the model of AC_1 coefficient. Significant progress has been achieved in the elimination of nuisance parameters for decades [27–31]. By fixing the marginal totals in the contingency table, Mehta [27] extensively used the conditional test (referred to as the C approach) to analyze various classical categorical data. Liddell [28] derived a test based on the exact distribution of the difference in sample proportions. As an alternative, Storer and Kim [29] modified Liddell’s exact test, abbreviated as the E approach. Basu [30] provided a new procedure by maximizing the tail probability over the whole range of parameters, called the M approach. The global maximum is a challenge when the parameter space is not finite. Lloyd [31] pointed out the weakness of the M approach, and he suggested a so-called E+M approach by defining the tail area with the E approach and maximizing the tail probability over the parameter space. Generally, E, M, and E+M approaches are called unconditional tests. Tang et al. [32] showed that the exact conditional approach was generally inferior to the exact unconditional approach for small samples. Shan and Wilding [33] compared asymptotic and exact procedures for the kappa coefficient in a 2×2 table. However, little work has been carried out in extending the exact approaches to test the homogeneity of the AC_1 coefficients across several independent strata.

This paper aims to propose asymptotic and exact methods for the homogeneity test of stratified AC_1 . The novelty and contribution are shown by three main aspects as follows. (i) For large sample sizes, we propose two asymptotic statistics, including likelihood ratio and Wald-type tests, to extend the study of homogeneity test in Honda and Ohyama [19] under large sample sizes. Our results show that the likelihood ratio test is more robust than other tests regarding type I error rates. The powers of these tests are close to each other. Thus, we recommend the likelihood ratio test for large samples’ homogeneity test of stratified AC_1 . (ii) Based on the asymptotic statistics, we derive three exact approaches (E, M, and E+M methods) to investigate the small sample cases ($n = 10, 25$). These exact methods can effectively improve the performance of the homogeneity test concerning type I error rates. Among these methods, the exact E approaches based on likelihood ratio and score tests are more robust in small samples. (iii) We investigate the strengths and weaknesses of asymptotic and exact methods through plentiful numerical analyses, respectively. Some beneficial conclusions are obtained from the analyses of actual examples. The rest of this paper is organized as follows. In Section 2, we review the AC_1 coefficient in a stratified condition and establish a probability model. The maximum likelihood method and iterative algorithm are used to estimate the unknown parameters. We further review the score statistic and derive two asymptotic test statistics for large samples in Section 3. Based on these statistics, several exact methods are used for small sample sizes in Section 4. In Section 5, we conduct numerical studies to investigate the performance of

all the derived methods regarding type I error rates and powers. In Section 6, we study the aforementioned real examples of large and small samples to illustrate these methods. Finally, a brief conclusion is given in Section 7.

2. A Probability Model and Homogeneity Test

Following Ohyama [12], we introduce K covariates into Table 2 and establish a probability model. Suppose that N subjects are divided into K independent strata. In the k th ($k = 1, 2, \dots, K$) stratum, there are n_{1k} , n_{2k} , and n_{3k} subjects in the three categories. Denote $n_k = \sum_{l=1}^3 n_{lk}$ as the total number of subjects in the k th stratum. Table 5 shows the data structure across the strata.

Table 5. Frequencies and probabilities of ratings in K strata.

Category	Ratings	Frequency of Subjects				Total
		Stratum 1	Stratum 2	...	Stratum K	
1	(+, +)	$n_{11} (P_{11}(\gamma_1, \pi_1))$	$n_{12} (P_{12}(\gamma_2, \pi_2))$	...	$n_{1K} (P_{1K}(\gamma_K, \pi_K))$	S_1
2	(+, -) or (-, +)	$n_{21} (P_{21}(\gamma_1, \pi_1))$	$n_{22} (P_{22}(\gamma_2, \pi_2))$	...	$n_{2K} (P_{2K}(\gamma_K, \pi_K))$	S_2
3	(-, -)	$n_{31} (P_{31}(\gamma_1, \pi_1))$	$n_{32} (P_{32}(\gamma_2, \pi_2))$	...	$n_{3K} (P_{3K}(\gamma_K, \pi_K))$	S_3
Total		n_1	n_2	...	n_K	N

For the stratified analysis, we need to construct AC_1 for each stratum. Let X_{kij} be an indicator of the i th ($i = 1, 2$) rater's judgement for the j th ($j = 1, 2, \dots, n_K$) subject in the k th ($k = 1, 2, \dots, K$) stratum. If there is a positive "(+)" classification, then $X_{kij} = 1$, and otherwise 0. Ohyama [12] assumed that the underlying probability of classifying a subject does not depend on raters but on subjects; that is, $Pr(X_{kij} = 1|j) = p_{kj}$. The N subjects are classified into K strata based on covariates, and every stratum has different subjects. Thus, the data of every stratum is independent of each other. Denote $E(p_{kj}) = \sum_{j=1}^{n_k} p_{kj}/n_k \triangleq \pi_k$, and $Var(p_{kj}) = \sum_{j=1}^{n_k} (p_{kj} - \pi_k)^2/n_k \triangleq \sigma_k^2$. Then, AC_1 of the k th stratum is

$$\gamma_k = \frac{1 + 2[\sigma_k^2 - 2\pi_k(1 - \pi_k)]}{1 - 2\pi_k(1 - \pi_k)}, \quad k = 1, 2, \dots, K.$$

Suppose that $P_{1k}(\gamma_k, \pi_k)$, $P_{2k}(\gamma_k, \pi_k)$, and $P_{3k}(\gamma_k, \pi_k)$ are the corresponding probabilities in the k th stratum, where π_k and γ_k are the common positive classification probability and the AC_1 coefficient, respectively. As the AC_1 coefficient in the k th stratum, γ_k includes the information of π_k and σ_k . It is obvious that there is no one-to-one correspondence between γ_k and π_k . Denote $\mathbf{n}_k = (n_{1k}, n_{2k}, n_{3k})^T$, and $\mathbf{P}_k = (P_{1k}(\gamma_k, \pi_k), P_{2k}(\gamma_k, \pi_k), P_{3k}(\gamma_k, \pi_k))^T$. For the k th stratum, $\mathbf{n}_k$ ($k = 1, 2, \dots, K$) follows a trinomial distribution. Thus, the probability density of $\mathbf{n}_k$ is expressed as follows:

$$f(\mathbf{P}_k|\mathbf{n}_k) = \frac{n_k!}{n_{1k}!n_{2k}!n_{3k}!} P_{1k}^{n_{1k}} P_{2k}^{n_{2k}} P_{3k}^{n_{3k}}. \quad (2)$$

Through calculation, the probabilities P_{lk} ($l = 1, 2, 3, k = 1, 2, \dots, K$) are obtained by

$$\begin{cases} P_{1k}(\gamma_k, \pi_k) = \pi_k(2 - \pi_k) - 1/2 + \gamma_k(1 - 2\pi_k(1 - \pi_k))/2, \\ P_{2k}(\gamma_k, \pi_k) = (1 - 2\pi_k(1 - \pi_k))(1 - \gamma_k), \\ P_{3k}(\gamma_k, \pi_k) = (1 - \pi_k)(1 + \pi_k) - 1/2 + \gamma_k(1 - 2\pi_k(1 - \pi_k))/2, \end{cases} \quad (3)$$

where $0 \leq P_{ik}(\gamma_k, \pi_k) \leq 1$ and $\sum_{i=1}^3 P_{ik}(\gamma_k, \pi_k) = 1$. Figure 1 shows the admissible range of γ_k , satisfying

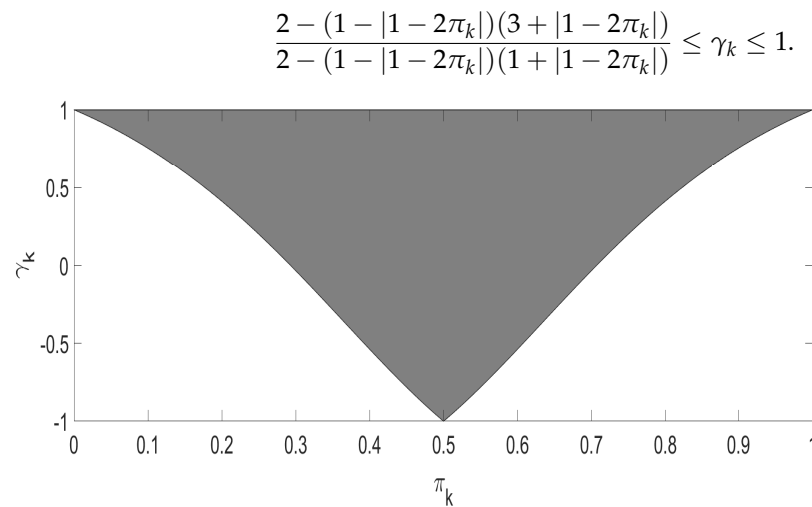


Figure 1. The range of γ_k .

Our work is interested in testing whether the AC_1 coefficients $\gamma_k (k = 1, 2, \dots, K)$ are homogeneous among the K independent strata, that is,

$$H_0 : \gamma_1 = \dots = \gamma_K \triangleq \gamma \text{ vs. } H_a : \gamma_k \text{ is not all the same,}$$

Denote $\pi = (\pi_1, \dots, \pi_K)$ and $\gamma = (\gamma_1, \dots, \gamma_K)$. First, we calculate the unknown parameters under the alternative hypothesis H_a . The corresponding log-likelihood function of the observed data $N = (n_1, n_2, \dots, n_K)$ is

$$\begin{aligned} l(\gamma, \pi | N) &= \log\left(\prod_{k=1}^K f(P_k | n_k)\right) = \log\left(\prod_{k=1}^K \frac{n_k!}{n_{1k}! n_{2k}! n_{3k}!} P_{1k}^{n_{1k}} P_{2k}^{n_{2k}} P_{3k}^{n_{3k}}\right) \\ &= \sum_{k=1}^K [n_{1k} \log P_{1k}(\gamma_k, \pi_k) + n_{2k} \log P_{2k}(\gamma_k, \pi_k) + n_{3k} \log P_{3k}(\gamma_k, \pi_k)] + C \\ &\triangleq \sum_{k=1}^K l_k(\gamma_k, \pi_k | n_k) + C, \end{aligned}$$

where $C = \log\left(\prod_{k=1}^K \frac{n_k!}{n_{1k}! n_{2k}! n_{3k}!}\right)$ is a constant, and $l_k(\gamma_k, \pi_k | n_k)$ is the log-likelihood function of the k th stratum under H_a . Let $\hat{\gamma}_k$ and $\hat{\pi}_k$ ($k = 1, 2, \dots, K$) be the unconstrained maximum likelihood estimates (MLEs) of γ_k and π_k under H_a . By solving the following equations,

$$\begin{aligned} \frac{\partial l_k}{\partial \pi_k} &= \frac{2n_{2k}(2\pi_k - 1)}{2\pi_k^2 - 2\pi_k + 1} - \frac{2n_{3k}(\gamma_k + 2\pi_k - 2\pi_k\gamma_k)}{1 - 2\pi_k^2 + \gamma_k(2\pi_k^2 - 2\pi_k + 1)} + \frac{2n_{1k}(2 - 2\pi_k - \gamma_k + 2\pi_k\gamma_k)}{4\pi_k + \gamma_k(2\pi_k^2 - 2\pi_k + 1) - 2\pi_k^2 - 1} = 0, \\ \frac{\partial l_k}{\partial \gamma_k} &= \frac{n_{2k}}{\gamma_k - 1} + \frac{n_{1k}(2\pi_k^2 - 2\pi_k + 1)}{4\pi_k + \gamma_k(2\pi_k^2 - 2\pi_k + 1) - 2\pi_k^2 - 1} + \frac{n_{3k}(2\pi_k^2 - 2\pi_k + 1)}{1 - 2\pi_k^2 + \gamma_k(2\pi_k^2 - 2\pi_k + 1)} = 0, \end{aligned}$$

we have

$$\hat{\pi}_k = \frac{2n_{1k} + n_{2k}}{2n_k}, \quad \hat{\gamma}_k = 1 - \frac{2n_k n_{2k}}{n_k^2 + (n_{1k} - n_{3k})^2}, \quad k = 1, 2, \dots, K.$$

Next, we estimate the parameters γ and $\pi = (\pi_1, \dots, \pi_K)$ under the null hypothesis $H_0 : \gamma_1 = \gamma_2 = \dots = \gamma_K \triangleq \gamma$. The log-likelihood function is rewritten by

$$\begin{aligned}
l_0(\gamma, \boldsymbol{\pi}|N) &= \sum_{k=1}^K \left\{ n_{1k} \log \left[\pi_k(2 - \pi_k) - \frac{1}{2} + \frac{\gamma}{2}(1 - 2\pi_k(1 - \pi_k)) \right] + n_{2k} \log \left[(1 - 2\pi_k(1 - \pi_k))(1 - \gamma) \right] \right. \\
&\quad \left. + n_{3k} \log \left[(1 - \pi_k)(1 + \pi_k) - \frac{1}{2} + \frac{\gamma}{2}(1 - 2\pi_k(1 - \pi_k)) \right] \right\} + C \\
&\triangleq \sum_{k=1}^K l_{0k}(\gamma, \pi_k|\mathbf{n}_k) + C,
\end{aligned}$$

where $l_{0k}(\gamma, \pi_k|\mathbf{n}_k)$ is the log-likelihood function of the k th stratum under H_0 . Let $\tilde{\gamma}$ and $\tilde{\pi}_k$ ($k = 1, 2, \dots, K$) be the constrained MLEs of γ and π_k under H_0 . Similarly, we can differentiate $l_0(\gamma, \boldsymbol{\pi}|N)$ to γ and π_k , and set them to zero as follows:

$$\begin{aligned}
\frac{\partial l_0}{\partial \pi_k} &= \frac{2n_{2k}(2\pi_k - 1)}{2\pi_k^2 - 2\pi_k + 1} - \frac{2n_{3k}(\gamma + 2\pi_k - 2\pi_k\gamma)}{1 - 2\pi_k^2 + \gamma(2\pi_k^2 - 2\pi_k + 1)} + \frac{2n_{1k}(2 - 2\pi_k - \gamma + 2\pi_k\gamma)}{4\pi_k + \gamma(2\pi_k^2 - 2\pi_k + 1) - 2\pi_k^2 - 1} = 0, \\
\frac{\partial l_0}{\partial \gamma} &= \sum_{k=1}^K \left[\frac{n_{2k}}{\gamma - 1} + \frac{n_{1k}(2\pi_k^2 - 2\pi_k + 1)}{4\pi_k + \gamma(2\pi_k^2 - 2\pi_k + 1) - 2\pi_k^2 - 1} + \frac{n_{3k}(2\pi_k^2 - 2\pi_k + 1)}{1 - 2\pi_k^2 + \gamma(2\pi_k^2 - 2\pi_k + 1)} \right] = 0.
\end{aligned}$$

However, there are no closed-form solutions for the above equations. The Fisher scoring algorithm is used to obtain the constrained MLEs. Three steps describe the iteration process as follows.

- (i) Given the initial values $\gamma^{(0)} = 0.5$, and $\pi_k^{(0)} = (2n_{1k} + n_{2k}) / (2n_k)$ in the k th stratum.
- (ii) The $(t + 1)$ -th approximates of γ and $\boldsymbol{\pi}$ can be updated by

$$\begin{bmatrix} \gamma^{(t+1)} \\ \boldsymbol{\pi}^{(t+1)} \end{bmatrix} = \begin{bmatrix} \gamma^{(t)} \\ \boldsymbol{\pi}^{(t)} \end{bmatrix} + \mathbf{I}_1^{-1}(\gamma, \boldsymbol{\pi}) \times \begin{bmatrix} \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \boldsymbol{\pi}} \end{bmatrix} \bigg|_{\gamma=\gamma^{(t)}, \boldsymbol{\pi}=\boldsymbol{\pi}^{(t)}},$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)^T$, $\frac{\partial l}{\partial \boldsymbol{\pi}} = (\frac{\partial l}{\partial \pi_1}, \dots, \frac{\partial l}{\partial \pi_K})^T$, and $\mathbf{I}_1$ is the $(K + 1) \times (K + 1)$ Fisher information matrix (Appendix A.1).

- (iii) Repeat the processes (i)–(ii) until the results converge.

3. Asymptotic Methods

3.1. Likelihood Ratio Statistic T_L

The unconstrained and constrained MLEs construct a likelihood ratio test statistic. It is defined by

$$T_L = 2[l(\hat{\gamma}, \hat{\boldsymbol{\pi}}|N) - l_0(\tilde{\gamma}, \tilde{\boldsymbol{\pi}}|N)] = 2 \sum_{k=1}^K [l_k(\hat{\gamma}_k, \hat{\pi}_k|\mathbf{n}_k) - l_{0k}(\tilde{\gamma}, \tilde{\pi}_k|\mathbf{n}_k)],$$

where $\mathbf{N} = (\mathbf{n}_1, \mathbf{n}_2, \dots, \mathbf{n}_K)$ is the observed data, $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \hat{\gamma}_2, \dots, \hat{\gamma}_K)$ and $\hat{\boldsymbol{\pi}} = (\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_K)$ are the unconstrained MLEs, $\tilde{\gamma}$ and $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_K)$ are the constrained MLEs.

3.2. Score Statistic T_{SC}

Honda and Ohayama [19] proposed the score statistic. Denote

$$\mathbf{U} = \left(\frac{\partial l_1}{\partial \gamma_1}, \frac{\partial l_2}{\partial \gamma_2}, \dots, \frac{\partial l_K}{\partial \gamma_K}, 0, 0, \dots, 0 \right)_{1 \times 2K}.$$

Under H_0 , the score test statistic can be represented as

$$T_{SC} = \mathbf{U} \mathbf{I}_2^{-1} \mathbf{U}^T|_{\gamma_1=\gamma_2=\dots=\gamma_K=\tilde{\gamma}, \boldsymbol{\pi}=\tilde{\boldsymbol{\pi}}},$$

where $\tilde{\gamma}$ and $\tilde{\pi} = (\tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_K)^T$ are the constrained MLEs. The $2K \times 2K$ Fisher information matrix I_2 is given in Appendix A.2. Through calculation, its simplified form is

$$T_{SC} = \sum_{k=1}^K \frac{r_k^2 d_k}{n_k (b_k d_k - c_k^2)} \Big|_{\gamma_k = \tilde{\gamma}, \pi_k = \tilde{\pi}_k},$$

where

$$\begin{aligned} r_k &= \frac{n_{1k}}{P_{1k}(\gamma_k, \pi_k)} - \frac{2n_{2k}}{P_{2k}(\gamma_k, \pi_k)} + \frac{n_{3k}}{P_{3k}(\gamma_k, \pi_k)}, \\ b_k &= \frac{1}{P_{1k}(\gamma_k, \pi_k)} + \frac{4}{P_{2k}(\gamma_k, \pi_k)} + \frac{1}{P_{3k}(\gamma_k, \pi_k)}, \\ c_k &= \frac{1}{P_{1k}(\gamma_k, \pi_k)} - \frac{1}{P_{3k}(\gamma_k, \pi_k)} + (1 - \gamma_k)(1 - 2\pi_k)b_k, \\ d_k &= \frac{1}{P_{1k}(\gamma_k, \pi_k)} + \frac{1}{P_{3k}(\gamma_k, \pi_k)} + (1 - \gamma_k)(1 - 2\pi_k) \left(\frac{1}{P_{1k}(\gamma_k, \pi_k)} - \frac{1}{P_{3k}(\gamma_k, \pi_k)} + c_k \right) \end{aligned}$$

for $k = 1, 2, \dots, K$.

3.3. Wald-Type Statistic T_W

Denote $\beta = (\gamma_1, \gamma_2, \dots, \gamma_K, \pi_1, \pi_2, \dots, \pi_K)_{1 \times 2K}$, and

$$C = \begin{bmatrix} 1 & -1 & 0 & \dots & \dots & 0 & 0 & \dots & \dots & 0 \\ 0 & 1 & -1 & \dots & \dots & 0 & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \dots & 0 & 1 & -1 & 0 & \dots & \dots & 0 \end{bmatrix}_{(K-1) \times 2K}.$$

The null hypothesis H_0 is equivalent to $C\beta^T = \mathbf{0}$, where $\mathbf{0}$ is a zero vector. Thus, we define the Wald-type statistic as

$$T_W = (\beta C^T)(CI_3^{-1}C^T)^{-1}(C\beta^T)|_{\gamma_k = \hat{\gamma}_k, \pi_k = \hat{\pi}_k},$$

where $\hat{\gamma}_k$ and $\hat{\pi}_k$ are the unconstrained MLEs. The Fisher information matrix I_3 is the same as that of the score test. We obtain the simplified form of T_W as

$$T_W = \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} (\hat{\gamma}_i - \hat{\gamma}_{i+1})(\hat{\gamma}_j - \hat{\gamma}_{j+1})(CI_3^{-1}C^T)^{-1}_{ij}.$$

Appendix A.3 provides the detailed process.

Under H_0 , these three statistics T_L , T_{SC} , and T_W are asymptotically distributed as a chi-square distribution with $K - 1$ degrees of freedom [34]. Given a significance level α , H_0 would be rejected if $T_\theta \geq \chi_{(K-1), (1-\alpha)}^2$, $\theta = L, SC, W$, where $\chi_{(K-1), (1-\alpha)}^2$ is the 100(1 - α) percentile of the chi-square distribution with $K - 1$ degrees of freedom. For a special observed data $N^* = (n_1, n_2, \dots, n_K)$, the p -values of these statistics are defined as

$$p_\theta^A(N^*) = P(\chi_{(K-1), (1-\alpha)}^2 > T_\theta(N^*)), \quad \theta = L, SC, W,$$

where $T_\theta(N^*)$ is the value of the statistic for the observed data N^* . For convenience, p_L^A , p_{SC}^A , and p_W^A are called asymptotic (A) approaches. Generally, asymptotic tests work well for large sample scenarios. However, they are conservative or liberal in the case of small sample sizes. Thus, we propose several exact methods based on the above statistics.

4. Exact Methods

Researchers often use the p -value to summarise the evidence against a null hypothesis. Thus, the key to the exact method is the calculation of the exact p -value. We uniformly

denote the aforementioned test statistics T_L , T_{SC} , and T_W as T_θ ($\theta = L, SC, W$). Instead of relying on the chi-square distribution, the exact test can use the true sampling distribution of T_θ and compute an exact p -value. The calculation process is as follows. First, we need to generate all possible tables. For a given observed data N^* , the column margins $n_1, n_2, \dots, n_K$ are fixed. We enumerate all possible tables by varying the cell values. The detailed process is described as follows.

(i) Produce all possible values of each stratum, which is formed by all combinations (n_{1k}, n_{2k}, n_{3k}) such that $n_{1k} + n_{2k} + n_{3k} = n_k$ ($k = 1, \dots, K$), and n_k is fixed. We take $K = 2$ and $n_1 = n_2 = 2$ as an example. There are six combinations in the k th stratum, including $(0, 0, 2), (0, 1, 1), (0, 2, 0), (1, 0, 1), (1, 1, 0), (2, 0, 0)$.

(ii) Enumerate all possible tables determined by the combination of all strata. For $K = 2$ and $n_1 = n_2 = 2$, we can obtain 36 possible tables in Table 6.

Note that each column corresponds to a categorical table with K strata.

Through steps (i)–(ii), we can enumerate all possible tables for any observed data $N^* = (n_1, n_2, \dots, n_K)$. Then we identify the tail area from this reference set. The tail area includes all the tables whose statistic values equal or exceed the statistics of the observed data N^* . Finally, the exact p -value is calculated by summing the probabilities of all the tables in the tail area. The calculation of the exact p -value needs to eliminate the unknown parameters shown in the previous section. The following exact methods use different ways for the elimination of the unknown parameters.

4.1. E Approach

The E approach eliminates the unknown parameters by replacing them with the constrained MLEs. We first generate all possible tables. Define the tail area $\Omega_E(N^*) = \{N : T_\theta(N) \geq T_\theta(N^*)\}$ based on the test statistic T_θ . The exact p -value of the observed data N^* is expressed by

$$p_\theta^E(N^*) = P(T_\theta(N) \geq T_\theta(N^*) | \tilde{\gamma}^*, \tilde{\pi}^*) = \sum_{N \in \Omega_E(N^*)} L(\tilde{\gamma}^*, \tilde{\pi}^* | N), \quad \theta = L, SC, W,$$

where $\tilde{\gamma}^*$ and $\tilde{\pi}^* = (\tilde{\pi}_1^*, \tilde{\pi}_2^*, \dots, \tilde{\pi}_K^*)$ are the constrained MLEs of γ and $\pi = (\pi_1, \pi_2, \dots, \pi_K)$. Meanwhile, the probability of a table in the tail area is $L(\tilde{\gamma}^*, \tilde{\pi}^* | N) = \exp\{l_0(\tilde{\gamma}^*, \tilde{\pi}^* | N)\}$, which is the likelihood function under the null hypothesis. For convenience, p_L^E , p_{SC}^E , and p_W^E are collectively called the E approach.

4.2. M Approach

In Basu [30], the size of a test is always understood as the maximum probability of the type I error rate. Thus, the elimination of the unknown parameters for the M approach is to find the values of parameters over the whole range of γ and π , which can maximize the sum of probabilities of all the tables in the tail area. This maximum is the p -value of the M approach. Denote $\Theta = \{\pi : \pi_k \in [0, 1], k = 1, 2, \dots, K\}$ and

$$\Lambda = \left\{ \gamma : \frac{2 - (1 - |1 - 2\pi_k|)(3 + |1 - 2\pi_k|)}{2 - (1 - |1 - 2\pi_k|)(1 + |1 - 2\pi_k|)} \leq \gamma_k \leq 1 \right\},$$

where $\pi = (\pi_1, \pi_2, \dots, \pi_K)$ and $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_K)$. Similar to the E approach, the tail area can be calculated by $\Omega_M(N^*) = \{N : T_\theta(N) \geq T_\theta(N^*)\}$. Under these conditions, the exact p -value of the M approach can be defined as

$$p_\theta^M(N^*) = \sup_{\gamma \in \Lambda, \pi \in \Theta} \left\{ \sum_{N \in \Omega_M(N^*)} L(\gamma, \pi | N) \right\}, \quad \theta = L, SC, W,$$

where $L(\gamma, \pi | N) = \exp\{l(\gamma, \pi | N)\}$ is the likelihood function under H_a . M approaches based on the three statistics are denoted as p_L^M , p_{SC}^M , and p_W^M .

Table 6. All the possible tables for $K = 2$ and $n_1 = n_2 = 2$.

[illegible]

4.3. E+M Approach

The E approach is not always effective because of unsatisfactory type I error rates. Lloyd [31] used an additional maximization step to improve it, which is called the E+M approach. First, the p -value of the E approach is used as a test statistic to define the tail area. Then, we maximize the sum of probabilities of all the tables in the tail area as the exact p -value. Based on the above procedures, the tail area of the E+M approach is defined as $\Omega_{E+M}(N^*) = \{N : p_{\theta}^E(N) \leq p_{\theta}^E(N^*)\}$. The exact p -value of the E+M approach is expressed as

$$p_{\theta}^{EM}(N^*) = \sup_{\gamma \in \Lambda, \pi \in \Theta} \left\{ \sum_{N \in \Omega_{E+M}(N^*)} L(\gamma, \pi | N) \right\}, \quad \theta = L, SC, W,$$

where $L(\gamma, \pi | N)$ is the same as the likelihood function in the M approach. The E+M approach includes p_L^{EM} , p_{SC}^{EM} , and p_W^{EM} .

5. Numerical Simulation

This section investigates the performance of asymptotic and exact methods in terms of type I error rates and powers. Given a significance level of 0.05, the type I error rate is the probability of rejecting H_0 when H_0 is true. According to Tang et al. [35], a test is considered liberal when the type I error rate is larger than 0.06, and conservative if it is less than 0.04; otherwise, it is robust. In several tables of this paper, we put the robust region of type I error rate (0.04–0.06) in bold to illustrate the performance of statistics. The power is defined by

$$\text{power} = 1 - \beta = 1 - P(H_0 \text{ is accepted} | H_0 \text{ is false}).$$

A test is optimal if it is robust and has more significant power.

5.1. Simulations of Asymptotic Methods

In the simulation, we first compare the performance of test statistics T_L , T_{SC} , and T_W in terms of empirical type I error rates under different parameter settings. Under $H_0 : \gamma_1 = \gamma_2 = 0.1$, we take $K = 2$, $n_1 = n_2 = 50$, and $\pi_1 = \pi_2 = 0.3$ as an example to describe the detailed calculation process of empirical type I error rates.

- (i) Bring the given values of γ_k , n_k , and π_k ($k = 1, 2$) into (3). Let F be the cumulative distribution for the three types of ratings ($l = 1, 2, 3$) of the two strata. Through calculation, we have

$$\begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \\ P_{31} & P_{32} \end{pmatrix} = \begin{pmatrix} 0.0390 & 0.0390 \\ 0.5220 & 0.5220 \\ 0.4390 & 0.4390 \end{pmatrix}, F = (F_{lk}) = \begin{pmatrix} 0.0390 & 0.0390 \\ 0.5610 & 0.5610 \\ 1.0000 & 1.0000 \end{pmatrix}.$$

- (ii) We produce an $n_1 \times K$ pseudorandom matrix drawn from the standard uniform distribution on the open interval (0,1), denoted by $r = (r_{ik})$ ($i = 1, 2, \dots, n_1, k = 1, 2, \dots, K$). Define

$$n_{1k} = \sum_{i=1}^{n_1} I_{\{r_{ik} < F_{1k}\}}, \quad n_{3k} = \sum_{i=1}^{n_1} I_{\{r_{ik} > F_{2k}\}}, \quad k = 1, 2, \dots, K.$$

Then, $n_{2k} = n_1 - n_{1k} - n_{3k}$ for $k = 1, 2, \dots, K$. When $K = 2$, we can obtain a sample (or table) with two strata as follows:

$$\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \\ n_{31} & n_{32} \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 23 & 26 \\ 25 & 22 \end{pmatrix}.$$

When 10,000 pseudorandom matrices are given, 10,000 samples are randomly produced under the null hypothesis $H_0 : \gamma_1 = \gamma_2 = 0.1$.

- (iii) For each sample, we calculate the corresponding MLEs and construct three statistics T_L, T_{SC} , and T_W . Given a significance level α , H_0 would be rejected if $T_\theta \geq \chi^2_{(K-1), (1-\alpha)}$, $\theta = L, SC, W$, where $\chi^2_{(K-1), (1-\alpha)}$ is the $100(1 - \alpha)$ percentile of the chi-square distribution with $K - 1$ degrees of freedom.
- (iv) The empirical type I error rate is calculated by the proportion of rejecting H_0 , which is the number of rejections/10,000.

Through steps (i)–(iv), we can calculate the empirical type I error rates of asymptotic test statistics T_L, T_{SC} , and T_W under different parameter settings. In practice, the AC_1 coefficient is usually positive. Under $H_0 : \gamma_1 = \dots = \gamma_K \triangleq \gamma$, Table 7 shows the empirical type I error rates of asymptotic statistics for $K = 2$ under the balanced and unbalanced π settings. The corresponding results of $K = 3, 4$ are shown in Tables A1 and A2 of the Appendix A.4. The tables show that the type I error rates of all the statistics are closer to the significance level of 0.05 with the increasing sample size. When the sample sizes are relatively small, the type I error rates of the likelihood ratio and score statistics are smaller than 0.05. The Wald-type test has a few liberal type I error rates. These three test statistics have conservative type I error rates for small and moderate samples when γ is close to 1. As the number of strata increases, T_W becomes more liberal. For the unbalanced π , some type I error rates under $\gamma = 0.1$ are more significant than 0.06 under $K = 3$ and $K = 4$. Overall, T_L should be recommended because of the more robust type I error rates among the three statistics.

We use three-dimensional figures to investigate the asymptotic methods' type I error rates. For convenience, the sample sizes are given as $n_1 = n_2 = \dots = n_K \triangleq n = 10, 50, 100$. Parameters are selected from $\pi_k = \pi$ and $\gamma_k = \gamma$ ($k = 1, 2, \dots, K, K = 2, 3, 4$). For each sample size, π increases from 0.1 to 0.9 by 0.04, and γ increases from -0.9 to 0.9 by 0.04. Figure 2 shows the distribution surfaces of type I error rates for all tests under $K = 2$. Similarly, the cases of $K = 3, 4$ are displayed in Figures A1 and A2 of the Appendix A.4. From these figures, we observe that the type I error rates of these statistics are smaller than 0.05 when the value of γ is close to -1 or 1 . The type I error rates of Wald-type statistics tend to be larger for the same sample size with the increasing number of strata. Thus, it is more liberal than the other two statistics. The empirical type I error rates of the likelihood ratio and score statistics are closer to 0.05 as the sample size increases. For large sample scenarios, T_L, T_{SC} , and T_W are usually robust, the type I error rates of T_L are more concentrated at 0.05. Overall, the likelihood ratio statistic performs better under all configurations. However, when the sample sizes are small, most type I error rates of T_L and T_{SC} are smaller than 0.04, and those of T_W are greater than 0.06.

Next, we analyze the powers of the proposed tests, which are similar to the calculation of empirical type I error rate when the samples are generated from the alternative hypothesis H_a . Take $n_1 = n_2 = \dots = n_K \triangleq n = 10, 50, 100$. The following parameter settings are considered for each sample size: (i) $K = 2, \pi = (0.5, 0.5), \gamma_1 = 0.1, \gamma_2 = -0.9 : 0.05 : 0.95$, (ii) $K = 3, \pi = (0.5, 0.5, 0.5), \gamma_1 = \gamma_3 = 0.1, \gamma_2 = -0.9 : 0.05 : 0.95$, (iii) $K = 4, \pi = (0.5, 0.5, 0.5, 0.5), \gamma_1 = \gamma_3 = \gamma_4 = 0.1, \gamma_2 = -0.9 : 0.05 : 0.95$. Here, $a : b : c$ means increasing from a to c by b . Under the alternative hypothesis H_a , we randomly generate 10,000 samples for each design. The empirical power equals the proportion of rejecting H_0 in all samples. Figure 3 shows the empirical powers of the three asymptotic tests. The Wald-type test has higher empirical powers than the other two statistics, especially in small samples. The powers of test statistics become higher if there exists a more considerable difference between γ_2 and γ_1 (γ_3). Among these statistics, the values of powers are closer as the sample size becomes larger.

Above all, T_L is recommended for the homogeneity test of stratified AC_1 in large sample sizes because of the robust type I error rates and satisfactory powers.

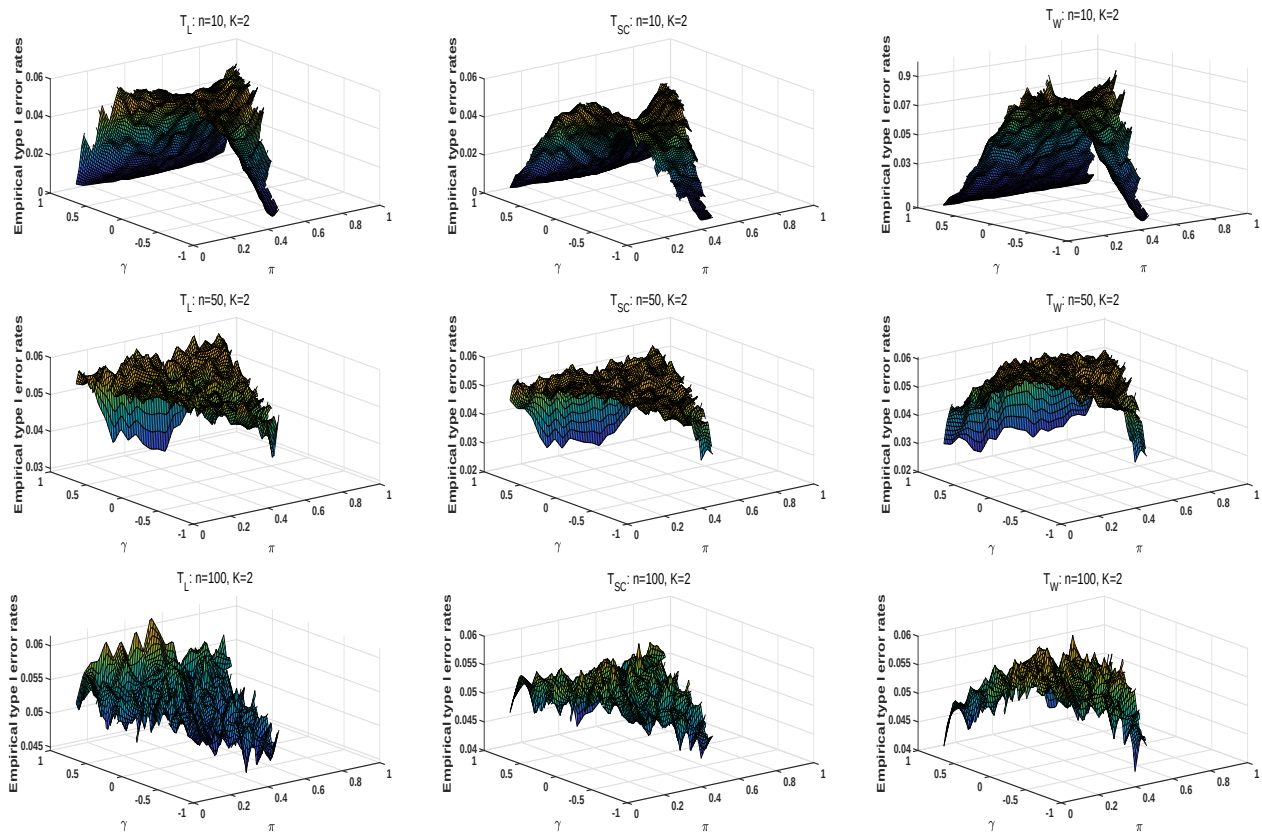


Figure 2. The surfaces of empirical type I error rates for asymptotic tests under $K = 2$.

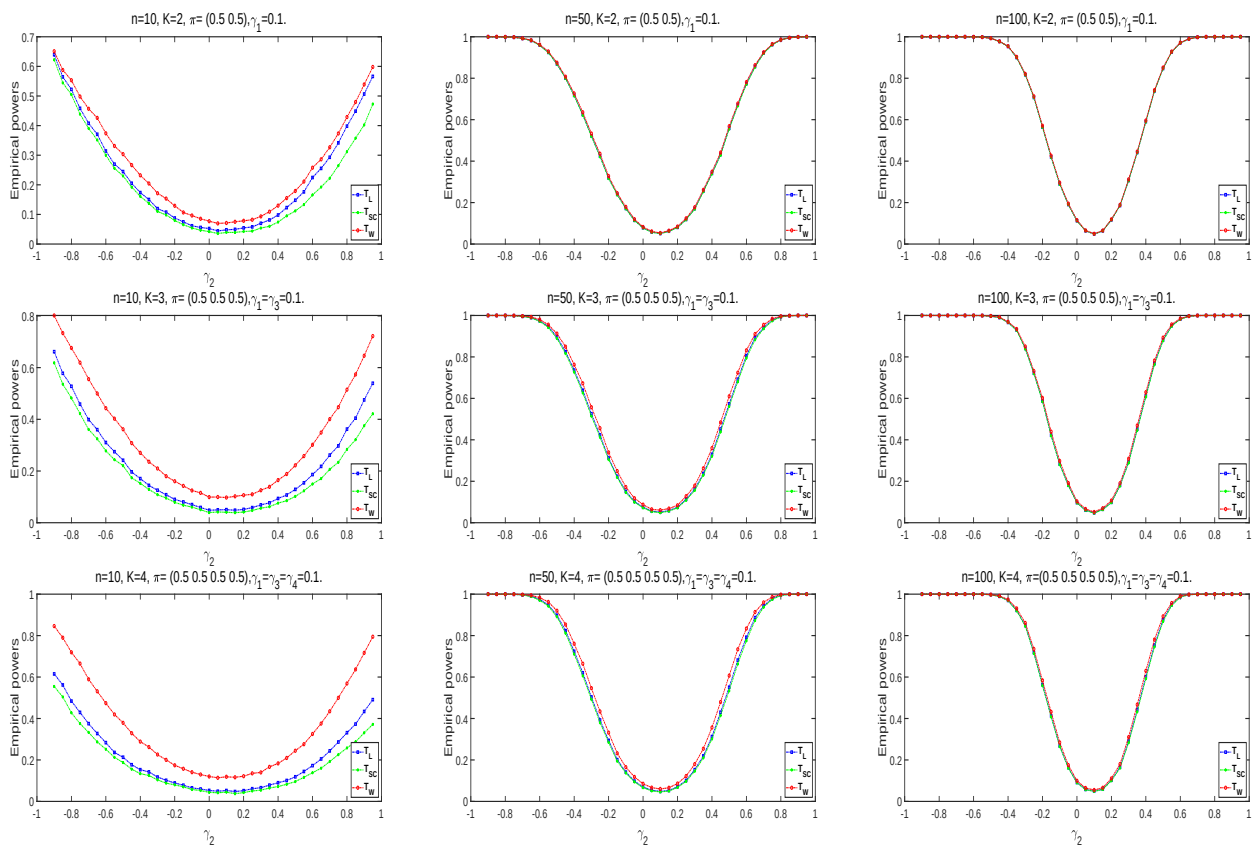


Figure 3. Empirical power curves of asymptotic tests for $K = 2, 3, 4$.

Table 7. Empirical type I error rates of asymptotic statistics for $K = 2$ (balanced sample sizes).

Balanced π Conditions						Unbalanced π Conditions						
$n_1 = n_2$	γ	$\pi_1 = \pi_2$	T_L	T_{SC}	T_W	$n_1 = n_2$	γ	π_1	π_2	T_L	T_{SC}	T_W
10	0.1	0.3	0.0536	0.0456	0.0789	10	0.1	0.3	0.5	0.0526	0.0425	0.0787
	0.3		0.0562	0.0483	0.0753		0.3			0.0506	0.0403	0.0702
	0.5		0.0434	0.0363	0.0540		0.5			0.0445	0.0326	0.0541
	0.7		0.0233	0.0186	0.0228		0.7			0.0192	0.0129	0.0225
	0.9		0.0018	0.0016	0.0018		0.9			0.0014	0.0007	0.0013
	0.1	0.6	0.0522	0.0452	0.0764		0.1	0.6	0.4	0.0521	0.0431	0.0750
	0.3		0.0515	0.0397	0.0697		0.3			0.0490	0.0362	0.0645
	0.5		0.0384	0.0264	0.0461		0.5			0.0440	0.0271	0.0513
	0.7		0.0189	0.0110	0.0227		0.7			0.0215	0.0127	0.0250
	0.9		0.0005	0.0002	0.0006		0.9			0.0009	0.0005	0.0013
	0.1	0.5	0.0483	0.0386	0.0727		0.1	0.5	0.4	0.0504	0.0422	0.0770
	0.3		0.0513	0.0335	0.0653		0.3			0.0504	0.0342	0.0670
	0.5		0.0424	0.0238	0.0496		0.5			0.0453	0.0276	0.0530
	0.7		0.0161	0.0080	0.0194		0.7			0.0140	0.0071	0.0173
	0.9		0.0006	0.0003	0.0012		0.9			0.0011	0.0005	0.0013
50	0.1	0.3	0.0496	0.0480	0.0558	50	0.1	0.3	0.5	0.0539	0.0526	0.0601
	0.3		0.0525	0.0515	0.0554		0.3			0.0544	0.0528	0.0576
	0.5		0.0524	0.0512	0.0542		0.5			0.0547	0.0537	0.0565
	0.7		0.0531	0.0493	0.0500		0.7			0.0529	0.0489	0.0515
	0.9		0.0394	0.0309	0.0280		0.9			0.0330	0.0219	0.0221
	0.1	0.6	0.0511	0.0504	0.0567		0.1	0.6	0.4	0.0498	0.0495	0.0549
	0.3		0.0480	0.0470	0.0521		0.3			0.0539	0.0525	0.0572
	0.5		0.0512	0.0490	0.0533		0.5			0.0518	0.0503	0.0544
	0.7		0.0552	0.0515	0.0545		0.7			0.0550	0.0505	0.0539
	0.9		0.0366	0.0269	0.0310		0.9			0.0375	0.0297	0.0330
	0.1	0.5	0.0493	0.0482	0.0517		0.1	0.5	0.4	0.0500	0.0491	0.0539
	0.3		0.0472	0.0464	0.0508		0.3			0.0516	0.0500	0.0540
	0.5		0.0506	0.0490	0.0532		0.5			0.0506	0.0487	0.0531
	0.7		0.0516	0.0471	0.0519		0.7			0.0535	0.0485	0.0531
	0.9		0.0326	0.0256	0.0310		0.9			0.0345	0.0248	0.0291
100	0.1	0.3	0.0493	0.0484	0.0520	100	0.1	0.3	0.5	0.0514	0.0508	0.0545
	0.3		0.0494	0.0492	0.0506		0.3			0.0473	0.0466	0.0490
	0.5		0.0516	0.0512	0.0524		0.5			0.0507	0.0499	0.0520
	0.7		0.0489	0.0476	0.0478		0.7			0.0527	0.0504	0.0523
	0.9		0.0547	0.0463	0.0440		0.9			0.0564	0.0493	0.0503
	0.1	0.6	0.0521	0.0518	0.0541		0.1	0.6	0.4	0.0483	0.0482	0.0514
	0.3		0.0493	0.0486	0.0518		0.3			0.0499	0.0495	0.0514
	0.5		0.0522	0.0514	0.0528		0.5			0.0509	0.0498	0.0520
	0.7		0.0530	0.0505	0.0522		0.7			0.0517	0.0496	0.0514
	0.9		0.0576	0.0463	0.0476		0.9			0.0568	0.0455	0.0466
	0.1	0.5	0.0473	0.0472	0.0489		0.1	0.5	0.4	0.0512	0.0510	0.0536
	0.3		0.0463	0.0461	0.0475		0.3			0.0552	0.0544	0.0568
	0.5		0.0523	0.0513	0.0533		0.5			0.0513	0.0503	0.0520
	0.7		0.0510	0.0489	0.0508		0.7			0.0568	0.0539	0.0568
	0.9		0.0579	0.0407	0.0457		0.9			0.0583	0.0478	0.0494

Note: The type I error rates of robust region (0.04–0.06) are shown in bold.

5.2. Exact Methods Results

Considering the unsatisfactory performance of asymptotic methods in small sample sizes, we introduce the exact E, M, and E+M methods to improve effectiveness. The type I error rates and powers of these three methods are compared with asymptotic approaches to investigate the advantages of exact methods. The algorithm for exact p -value is usually computationally intensive, time consuming, and sometimes exceeds the memory limits of the computer. Thus, running time is an important determinant for the appropriate numbers of strata and sample size in the numerical study of exact methods. For simplicity, we only

focus on $n_1 = \dots = n_K \triangleq n$, $\pi_k = 0.5$, and $\gamma_k = 0.1$ ($k = 1, 2, \dots, K$). The average of 100 running times is used as the running time of the exact p -value. We study the running times in the case of the following parameter settings: (i) $K = 2$, $n = 2, \dots, 11$, and (ii) $n = 3$, $K = 2, 3, 4, 5$. The running times of different methods are shown in Table 8. From the results, it is obvious that running time will increase exponentially for the growth of the number of strata K and the marginal numbers n . It is challenging for a computer's megabytes of storage and clock speed. Thus, the cases $n = 10$ and $K = 2, 3$ are considered in our work. Take $\pi_k = \pi$, $\gamma_k = \gamma$ ($k = 1, 2, \dots, K$), where $\pi = 0 : 0.02 : 1$, and $\gamma = -1 : 0.02 : 1$. Figure 4 shows the surfaces of type I error rates for $K = 2$. In the Appendix A.4, we provide the case of $K = 3$ in Figure A3. The small diagrams in the upper right corner reflect the curves of the type I error rates under $\pi = 0.5$ and $\gamma = -1 : 0.02 : 1$. For the large diagrams, p_L^A and p_W^A have liberal type I error rates, while the type I error rates of p_{SC}^A are smaller than 0.05. The M and E+M approaches produce conservative type I error rates. The E approaches under the likelihood ratio and score statistics are better than that under the Wald-type test when $K = 2$. For $K = 3$, the surfaces of E approaches under three statistics are closer to the significance level in the case of positive γ . From small diagrams, the curves of type I error rates have bimodal shapes. To reveal the reason, we consider the case of $K = 2$, $\gamma_k = \gamma$, and $\pi_k = 0.5$. As a part of tail probability,

$$LL = \exp \left(\sum_{k=1}^K (l_k(\gamma_k, \pi_k) | n_k) \right)$$

has the same value for each γ under the fixed sum of n_{2k} . Table 9 reflects the changes of LL as the increases of γ and $(n_{21} + n_{22})$. The table shows that the increase of γ can affect the peak change with the given sum of n_{2k} . The peak also occurs when the sum of n_{2k} increases for a given γ . Meanwhile, each method's tail area determines the bimodal shape's location and values. Next, we compare the type I error rates of exact and asymptotic methods under several parameters. Let $K = 2$ and $n_1 = n_2 = 10, 25$. From Table 10, the type I error rates of p_L^E and p_{SC}^E are closer to 0.05 than those of p_W^E . The E approach works better as the sample size increases in the range of small sample sizes. The type I error rates of E approaches are close to 0.05 when $n_1 = n_2 = 25$. It reveals that the E approach is more effective than asymptotic methods for high γ .

According to the relationship between γ_k and π_k , the parameter settings are considered under H_a as follows: (i) $n_k = n = 10$, $K = 2$, $\pi = (0.5, 0.5)$, $\gamma_1 = 0.1$, $\gamma_2 = -0.9 : 0.05 : 0.9$; (ii) $n_k = n = 10$, $K = 3$, $\pi = (0.5, 0.5, 0.5)$, $\gamma_1 = \gamma_3 = 0.1$, $\gamma_2 = -0.9 : 0.05 : 0.9$, ($k = 1, 2, \dots, K$). Figure 5 provides the power curves of the exact methods. For the A approach, p_W^A has higher powers under different parameter settings. The power becomes larger as the absolute value of $\gamma_2 - \gamma_1$ (γ_3) becomes larger. On the contrary, the powers of each method tend to be 0.05 as γ_2 becomes closer to 0.1. The power curves of the E+M approaches are close to each other. Then, we compare the powers of asymptotic and exact methods under different parameter settings. For $K = 2$ and $n_1 = n_2 = \dots = n_K = 10$, the values of γ_k are taken as (i) $\gamma_1 = 0.1$, $\gamma_2 = 0.5 : 0.1 : 0.9$; (ii) $\gamma_1 = 0.3$, $\gamma_2 = 0.6 : 0.1 : 0.9$; and (iii) $\gamma_1 = 0.5$, $\gamma_2 = 0.8, 0.9$. When $K = 3$, let $\gamma_1 = \gamma_3$ and other settings be the same. In Table 11, we provide the power comparisons of $K = 2$ under the balanced π conditions. Tables A3–A5 show the comparisons of these methods under other settings. The powers of exact methods are generally smaller than those of asymptotic methods. However, p_{SC}^E has higher powers than p_{SC}^A . For exact methods, the E approach has higher power than the other two methods, in which the E approach under the Wald-type statistic has higher power.

Table 8. Running times (seconds) for $\gamma = 0.1$ and $\pi = 0.5$.

Value	n										K			
	2	3	4	5	6	7	8	9	10	11	2	3	4	5
p_L^E	0.0088	0.0278	0.0540	0.1416	0.3046	0.6356	1.1910	2.2842	3.9241	7.6825	0.0278	1.0149	33.7988	3743.0157
p_{SC}^E	0.0115	0.0324	0.0692	0.1778	0.3651	0.7224	1.2988	2.4110	4.3842	8.4029	0.0324	0.9989	37.1695	3814.3481
p_W^E	0.0125	0.0364	0.0837	0.2092	0.4009	0.8020	1.4019	2.5217	4.5078	8.4676	0.0364	0.8794	39.2842	3797.6383
p_L^M	0.2946	0.4833	0.8561	1.7686	4.1932	8.1121	11.5237	16.2005	22.9323	34.0228	0.4833	5.9487	86.4241	1584.3279
p_{SC}^M	0.2867	0.5865	0.9794	2.1689	4.5669	7.8500	11.4646	17.1822	23.6388	42.0095	0.5865	5.3177	58.3290	1237.8588
p_W^M	0.2608	0.5477	0.9320	2.2620	4.1170	6.6121	10.8523	14.0822	24.4280	37.1586	0.5477	6.4484	109.0489	990.4258
p_L^{EM}	0.3457	0.7130	1.1282	2.3653	4.3037	8.1787	13.6343	18.7686	30.1321	49.2284	0.7130	7.9202	159.2139	5696.1438
p_{SC}^{EM}	0.3446	0.6777	1.0412	2.2872	4.3691	8.0231	12.8295	20.0811	30.1819	45.1356	0.6777	6.2942	131.3634	4930.5572
p_W^{EM}	0.3745	0.7553	1.1046	2.6653	4.6391	8.3511	12.5336	17.4634	29.6501	47.6976	0.7553	6.9710	186.0982	6903.8451

Table 9. The values of LL in different settings for $K = 2$.

$n_{21} + n_{22}$	γ																		
	−0.9	−0.8	−0.7	−0.6	−0.5	−0.4	−0.3	−0.2	−0.1	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
2	1.31×10^{-29}	3.09×10^{-24}	4.07×10^{-21}	6.40×10^{-19}	3.12×10^{-17}	7.24×10^{-16}	1.00×10^{-14}	9.44×10^{-14}	6.61×10^{-13}	3.64×10^{-12}	1.64×10^{-11}	6.20×10^{-11}	2.00×10^{-10}	5.59×10^{-10}	1.34×10^{-09}	2.75×10^{-09}	4.60×10^{-09}	5.73×10^{-09}	3.79×10^{-09}
3	4.99×10^{-28}	5.56×10^{-23}	4.62×10^{-20}	5.12×10^{-18}	1.87×10^{-16}	3.38×10^{-15}	3.72×10^{-14}	2.83×10^{-13}	1.62×10^{-12}	7.28×10^{-12}	2.68×10^{-11}	8.27×10^{-11}	2.16×10^{-10}	4.79×10^{-10}	8.96×10^{-10}	1.37×10^{-09}	1.63×10^{-09}	1.27×10^{-09}	3.99×10^{-10}
4	1.90×10^{-26}	1.00×10^{-21}	5.23×10^{-19}	4.10×10^{-17}	1.12×10^{-15}	1.58×10^{-14}	1.38×10^{-13}	8.49×10^{-13}	3.95×10^{-12}	1.46×10^{-11}	4.39×10^{-11}	1.10×10^{-10}	2.32×10^{-10}	4.11×10^{-10}	5.97×10^{-10}	6.87×10^{-10}	5.74×10^{-10}	2.83×10^{-10}	4.20×10^{-11}
5	7.21×10^{-25}	1.80×10^{-20}	5.93×10^{-18}	3.28×10^{-16}	6.74×10^{-15}	7.36×10^{-14}	5.13×10^{-13}	2.55×10^{-12}	9.65×10^{-12}	2.91×10^{-11}	7.18×10^{-11}	1.47×10^{-10}	2.50×10^{-10}	3.52×10^{-10}	3.98×10^{-10}	3.44×10^{-10}	2.02×10^{-10}	6.28×10^{-11}	4.42×10^{-12}
6	2.74×10^{-23}	3.24×10^{-19}	6.72×10^{-17}	2.62×10^{-15}	4.05×10^{-14}	3.43×10^{-13}	1.91×10^{-12}	7.64×10^{-12}	2.36×10^{-11}	5.82×10^{-11}	1.17×10^{-10}	1.96×10^{-10}	2.70×10^{-10}	3.02×10^{-10}	2.66×10^{-10}	1.72×10^{-10}	7.14×10^{-11}	1.40×10^{-11}	4.65×10^{-13}
7	1.04×10^{-21}	5.84×10^{-18}	7.62×10^{-16}	2.10×10^{-14}	2.43×10^{-13}	1.60×10^{-12}	7.08×10^{-12}	2.29×10^{-11}	5.77×10^{-11}	1.16×10^{-10}	1.92×10^{-10}	2.61×10^{-10}	2.90×10^{-10}	2.59×10^{-10}	1.77×10^{-10}	8.59×10^{-11}	2.52×10^{-11}	3.10×10^{-12}	4.90×10^{-14}
8	3.95×10^{-20}	1.05×10^{-16}	8.63×10^{-15}	1.68×10^{-13}	1.46×10^{-12}	7.48×10^{-12}	2.63×10^{-11}	6.88×10^{-11}	1.41×10^{-10}	2.33×10^{-10}	3.15×10^{-10}	3.48×10^{-10}	3.13×10^{-10}	2.22×10^{-10}	1.18×10^{-10}	4.30×10^{-11}	8.90×10^{-12}	6.90×10^{-13}	5.15×10^{-15}
9	1.50×10^{-18}	1.89×10^{-15}	9.78×10^{-14}	1.34×10^{-12}	8.74×10^{-12}	3.49×10^{-11}	9.76×10^{-11}	2.06×10^{-10}	3.45×10^{-10}	4.66×10^{-10}	5.15×10^{-10}	4.64×10^{-10}	3.37×10^{-10}	1.90×10^{-10}	7.87×10^{-11}	2.15×10^{-11}	3.14×10^{-12}	1.53×10^{-13}	5.42×10^{-16}
10	5.71×10^{-17}	3.41×10^{-14}	1.11×10^{-12}	1.07×10^{-11}	5.24×10^{-11}	1.63×10^{-10}	3.63×10^{-10}	6.19×10^{-10}	8.42×10^{-10}	9.31×10^{-10}	8.42×10^{-10}	6.19×10^{-10}	3.63×10^{-10}	1.63×10^{-10}	5.24×10^{-11}	1.07×10^{-11}	1.11×10^{-12}	3.41×10^{-14}	5.71×10^{-17}
11	2.17×10^{-15}	6.13×10^{-13}	1.26×10^{-11}	8.59×10^{-11}	3.15×10^{-10}	7.60×10^{-10}	1.35×10^{-09}	1.86×10^{-09}	2.06×10^{-09}	1.86×10^{-09}	1.38×10^{-09}	8.26×10^{-10}	3.91×10^{-10}	1.40×10^{-10}	3.50×10^{-11}	5.37×10^{-12}	3.91×10^{-13}	7.57×10^{-15}	6.01×10^{-18}
12	8.25×10^{-14}	1.10×10^{-11}	1.42×10^{-10}	6.87×10^{-10}	1.89×10^{-09}	3.55×10^{-09}	5.00×10^{-09}	5.57×10^{-09}	5.03×10^{-09}	3.73×10^{-09}	2.26×10^{-09}	1.10×10^{-09}	4.21×10^{-10}	1.20×10^{-10}	2.33×10^{-11}	2.68×10^{-12}	1.38×10^{-13}	1.68×10^{-15}	6.33×10^{-19}
13	3.13×10^{-12}	1.99×10^{-10}	1.61×10^{-09}	5.50×10^{-09}	1.13×10^{-08}	1.66×10^{-08}	1.86×10^{-08}	1.67×10^{-08}	1.23×10^{-08}	7.45×10^{-09}	3.69×10^{-09}	1.47×10^{-09}	4.53×10^{-10}	1.03×10^{-10}	1.55×10^{-11}	1.34×10^{-12}	4.87×10^{-14}	3.74×10^{-16}	6.66×10^{-20}
14	1.19×10^{-10}	3.57×10^{-09}	1.83×10^{-08}	4.40×10^{-08}	6.80×10^{-08}	7.73×10^{-08}	6.90×10^{-08}	5.02×10^{-08}	3.01×10^{-08}	1.49×10^{-08}	6.04×10^{-09}	1.96×10^{-09}	4.88×10^{-10}	8.79×10^{-11}	1.04×10^{-11}	6.71×10^{-13}	1.72×10^{-14}	8.30×10^{-17}	7.01×10^{-21}
15	4.52×10^{-09}	6.43×10^{-08}	2.07×10^{-07}	3.52×10^{-07}	4.08×10^{-07}	3.61×10^{-07}	2.56×10^{-07}	1.50×10^{-07}	7.35×10^{-08}	2.98×10^{-08}	9.88×10^{-09}	2.61×10^{-09}	5.25×10^{-10}	7.54×10^{-11}	6.91×10^{-12}	3.36×10^{-13}	6.07×10^{-15}	1.85×10^{-17}	7.38×10^{-22}
16	1.72×10^{-07}	1.16×10^{-06}	2.35×10^{-06}	2.81×10^{-06}	2.45×10^{-06}	1.68×10^{-06}	9.52×10^{-07}	4.51×10^{-07}	1.80×10^{-07}	5.96×10^{-08}	1.62×10^{-08}	3.48×10^{-09}	5.66×10^{-10}	6.46×10^{-11}	4.60×10^{-12}	1.68×10^{-13}	2.14×10^{-15}	4.10×10^{-18}	7.77×10^{-23}
17	6.53×10^{-06}	2.08×10^{-05}	2.66×10^{-05}	2.25×10^{-05}	1.47×10^{-05}	7.85×10^{-06}	3.54×10^{-06}	1.35×10^{-06}	4.39×10^{-07}	1.19×10^{-07}	2.65×10^{-08}	4.64×10^{-09}	6.09×10^{-10}	5.54×10^{-11}	3.07×10^{-12}	8.39×10^{-14}	7.56×10^{-16}	9.11×10^{-19}	8.18×10^{-24}

Table 10. Empirical type I error rates of exact and asymptotic methods for $K = 2$.

$n_1 = n_2$	γ	$\pi_1 = \pi_2$	Exact Methods						Asymptotic Methods					
			p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}	p_L^A	p_{SC}^A	p_W^A
10	0.1	0.3	0.0485	0.0497	0.0514	0.0366	0.0483	0.0404	0.0413	0.0327	0.0419	0.0569	0.0506	0.0855
	0.3		0.0482	0.0511	0.0528	0.0358	0.0491	0.0334	0.0405	0.0351	0.0425	0.0584	0.0514	0.0752
	0.5		0.0528	0.0544	0.0586	0.0428	0.0478	0.0267	0.0455	0.0394	0.0464	0.0686	0.0502	0.0632
	0.7		0.0580	0.0523	0.0546	0.0435	0.0354	0.0156	0.0481	0.0365	0.0414	0.0765	0.0383	0.0427
	0.8		0.0434	0.0371	0.0366	0.0290	0.0204	0.0073	0.0341	0.0243	0.0266	0.0582	0.0224	0.0256
	0.9		0.0149	0.0125	0.0106	0.0079	0.0048	0.0011	0.0105	0.0071	0.0072	0.0203	0.0055	0.0069
	0.1	0.5	0.0404	0.0425	0.0457	0.0340	0.0371	0.0354	0.0358	0.0332	0.0381	0.0485	0.0383	0.0749
	0.3		0.0445	0.0481	0.0559	0.0363	0.0379	0.0341	0.0383	0.0357	0.0431	0.0600	0.0389	0.0726
	0.5		0.0553	0.0550	0.0649	0.0399	0.0359	0.0301	0.0449	0.0371	0.0476	0.0751	0.0380	0.0694
	0.7		0.0549	0.0473	0.0542	0.0316	0.0222	0.0170	0.0404	0.0282	0.0372	0.0716	0.0243	0.0525
	0.8		0.0358	0.0292	0.0331	0.0177	0.0106	0.0072	0.0246	0.0157	0.0212	0.0469	0.0116	0.0314
	0.9		0.0099	0.0078	0.0086	0.0040	0.0019	0.0010	0.0062	0.0036	0.0050	0.0132	0.0021	0.0080
	0.1	0.6	0.0433	0.0451	0.0466	0.0334	0.0404	0.0366	0.0367	0.0327	0.0385	0.0506	0.0423	0.0783
	0.3		0.0458	0.0494	0.0537	0.0365	0.0414	0.0342	0.0387	0.0362	0.0427	0.0584	0.0426	0.0732
	0.5		0.0542	0.0556	0.0618	0.0420	0.0398	0.0295	0.0449	0.0387	0.0471	0.0727	0.0420	0.0671
	0.7		0.0547	0.0495	0.0528	0.0362	0.0261	0.0168	0.0418	0.0316	0.0379	0.0727	0.0285	0.0497
	0.8		0.0370	0.0319	0.0330	0.0214	0.0133	0.0074	0.0262	0.0186	0.0222	0.0496	0.0146	0.0300
	0.9		0.0109	0.0092	0.0088	0.0051	0.0026	0.0010	0.0068	0.0046	0.0054	0.0149	0.0029	0.0078
25	0.1	0.3	0.0493	0.0493	0.0493	0.0358	0.0358	0.0358	0.0409	0.0416	0.0393	0.0527	0.0516	0.0516
	0.3		0.0496	0.0496	0.0496	0.0344	0.0344	0.0344	0.0417	0.0430	0.0399	0.0523	0.0502	0.0501
	0.5		0.0494	0.0494	0.0494	0.0353	0.0353	0.0353	0.0408	0.0424	0.0395	0.0546	0.0514	0.0507
	0.7		0.0484	0.0483	0.0483	0.0400	0.0400	0.0400	0.0408	0.0442	0.0407	0.0599	0.0511	0.0477
	0.8		0.0562	0.0556	0.0556	0.0478	0.0478	0.0478	0.0483	0.0488	0.0440	0.0726	0.0481	0.0431
	0.9		0.0503	0.0497	0.0497	0.0370	0.0370	0.0370	0.0411	0.0378	0.0277	0.0702	0.0259	0.0207
	0.1	0.5	0.0495	0.0495	0.0495	0.0303	0.0303	0.0303	0.0389	0.0398	0.0358	0.0502	0.0477	0.0477
	0.3		0.0484	0.0484	0.0484	0.0312	0.0312	0.0312	0.0399	0.0412	0.0376	0.0496	0.0477	0.0477
	0.5		0.0486	0.0486	0.0486	0.0346	0.0346	0.0346	0.0392	0.0409	0.0386	0.0513	0.0492	0.0491
	0.7		0.0521	0.0521	0.0521	0.0401	0.0401	0.0401	0.0419	0.0451	0.0437	0.0612	0.0507	0.0507
	0.8		0.0593	0.0593	0.0593	0.0431	0.0431	0.0431	0.0467	0.0482	0.0487	0.0795	0.0457	0.0455
	0.9		0.0414	0.0414	0.0414	0.0223	0.0223	0.0223	0.0280	0.0276	0.0295	0.0690	0.0188	0.0186
	0.1	0.6	0.0493	0.0493	0.0493	0.0325	0.0325	0.0325	0.0404	0.0417	0.0386	0.0505	0.0481	0.0481
	0.3		0.0492	0.0492	0.0492	0.0328	0.0328	0.0328	0.0404	0.0417	0.0388	0.0509	0.0486	0.0486
	0.5		0.0486	0.0486	0.0486	0.0348	0.0348	0.0348	0.0399	0.0416	0.0392	0.0525	0.0498	0.0497
	0.7		0.0503	0.0502	0.0502	0.0408	0.0408	0.0408	0.0425	0.0448	0.0428	0.0607	0.0502	0.0496
	0.8		0.0584	0.0581	0.0581	0.0463	0.0463	0.0463	0.0495	0.0495	0.0473	0.0769	0.0454	0.0443
	0.9		0.0438	0.0436	0.0436	0.0280	0.0280	0.0280	0.0339	0.0324	0.0290	0.0685	0.0201	0.0187

Note: The type I error rates of robust region (0.04–0.06) are shown in bold.

Table 11. Powers of exact and asymptotic methods for $K = 2$ (balanced π conditions).

γ_1	γ_2	$\pi_1 = \pi_2$	Exact Methods									Asymptotic Methods			
			p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}	p_L^A	p_{SC}^A	p_W^A	
0.1	0.5	0.3	0.1403	0.1437	0.1549	0.1268	0.1368	0.1209	0.1323	0.1325	0.1334	0.1719	0.1593	0.1929	
	0.6		0.2108	0.2130	0.2281	0.1945	0.2041	0.1747	0.2006	0.2003	0.1996	0.2472	0.2304	0.2641	
	0.7		0.3114	0.3106	0.3317	0.2924	0.2990	0.2489	0.2979	0.2956	0.2943	0.3518	0.3276	0.3572	
	0.8		0.4509	0.4439	0.4723	0.4301	0.4283	0.3478	0.4336	0.4253	0.4265	0.4932	0.4556	0.4772	
	0.9		0.6408	0.6217	0.6570	0.6192	0.5991	0.4761	0.6201	0.5968	0.6076	0.6812	0.6180	0.6308	
0.3	0.6		0.1149	0.1187	0.1284	0.1019	0.1072	0.0781	0.1048	0.1042	0.1069	0.1389	0.1217	0.1378	
	0.7		0.1852	0.1869	0.2016	0.1683	0.1698	0.1208	0.1709	0.1671	0.1716	0.2173	0.1863	0.2040	
	0.8		0.2976	0.2937	0.3151	0.2764	0.2671	0.1859	0.2779	0.2655	0.2750	0.3409	0.2840	0.3029	
	0.9		0.4727	0.4555	0.4846	0.4470	0.4132	0.2824	0.4474	0.4142	0.4357	0.5307	0.4261	0.4461	
	0.8		0.1614	0.1570	0.1680	0.1456	0.1290	0.0703	0.1463	0.1333	0.1401	0.2000	0.1372	0.1444	
0.5	0.9		0.2818	0.2656	0.2828	0.2568	0.2176	0.1181	0.2581	0.2282	0.2423	0.3453	0.2255	0.2342	
0.1	0.5	0.5	0.1442	0.1533	0.1684	0.1186	0.1331	0.1142	0.1273	0.1310	0.1367	0.1671	0.1463	0.1852	
	0.6		0.2137	0.2265	0.2465	0.1830	0.1986	0.1711	0.1917	0.1967	0.2052	0.2455	0.2144	0.2647	
	0.7		0.3115	0.3277	0.3521	0.2761	0.2904	0.2511	0.2843	0.2899	0.3021	0.3532	0.3078	0.3696	
	0.8		0.4462	0.4631	0.4894	0.4071	0.4148	0.3607	0.4151	0.4186	0.4350	0.4969	0.4322	0.5015	
	0.9		0.6288	0.6382	0.6611	0.5876	0.5786	0.5076	0.5971	0.5922	0.6126	0.6835	0.5927	0.6587	
	0.3		0.6	0.1124	0.1210	0.1340	0.0946	0.0969	0.0777	0.0991	0.0999	0.1076	0.1383	0.1058	0.1413
			0.7	0.1797	0.1895	0.2065	0.1547	0.1533	0.1233	0.1609	0.1603	0.1717	0.2167	0.1643	0.2113
			0.8	0.2861	0.2942	0.3147	0.2507	0.2400	0.1936	0.2602	0.2554	0.2718	0.3376	0.2529	0.3119
			0.9	0.4501	0.4483	0.4700	0.3996	0.3690	0.2992	0.4153	0.4002	0.4229	0.5181	0.3823	0.4488
			0.8	0.1535	0.1508	0.1627	0.1232	0.1038	0.0754	0.1334	0.1249	0.1361	0.1950	0.1113	0.1476
0.5	0.9		0.2589	0.2471	0.2623	0.2089	0.1717	0.1252	0.2267	0.2100	0.2266	0.3219	0.1813	0.2295	
0.1	0.5	0.6	0.1478	0.1549	0.1678	0.1210	0.1378	0.1163	0.1307	0.1341	0.1413	0.1679	0.1531	0.1878	
	0.6		0.2183	0.2279	0.2453	0.1865	0.2048	0.1721	0.1960	0.2006	0.2109	0.2458	0.2227	0.2650	
	0.7		0.3169	0.3287	0.3504	0.2813	0.2983	0.2500	0.2896	0.2942	0.3088	0.3529	0.3177	0.3672	
	0.8		0.4516	0.4633	0.4876	0.4150	0.4244	0.3560	0.4212	0.4223	0.4425	0.4959	0.4430	0.4970	
	0.9		0.6323	0.6376	0.6602	0.5992	0.5891	0.4975	0.6040	0.5929	0.6197	0.6827	0.6029	0.6544	
	0.3		0.6	0.1139	0.1218	0.1337	0.0971	0.1014	0.0771	0.1008	0.1016	0.1114	0.1373	0.1119	0.1407
			0.7	0.1814	0.1905	0.2066	0.1595	0.1599	0.1217	0.1638	0.1626	0.1770	0.2158	0.1726	0.2101
			0.8	0.2881	0.2959	0.3161	0.2598	0.2500	0.1906	0.2653	0.2582	0.2792	0.3377	0.2642	0.3108
			0.9	0.4521	0.4515	0.4741	0.4160	0.3839	0.2940	0.4243	0.4033	0.4329	0.5212	0.3973	0.4495
			0.8	0.1536	0.1529	0.1648	0.1313	0.1116	0.0739	0.1373	0.1275	0.1410	0.1960	0.1199	0.1468
0.5	0.9		0.2608	0.2521	0.2681	0.2248	0.1853	0.1235	0.2357	0.2150	0.2357	0.3281	0.1952	0.2305	

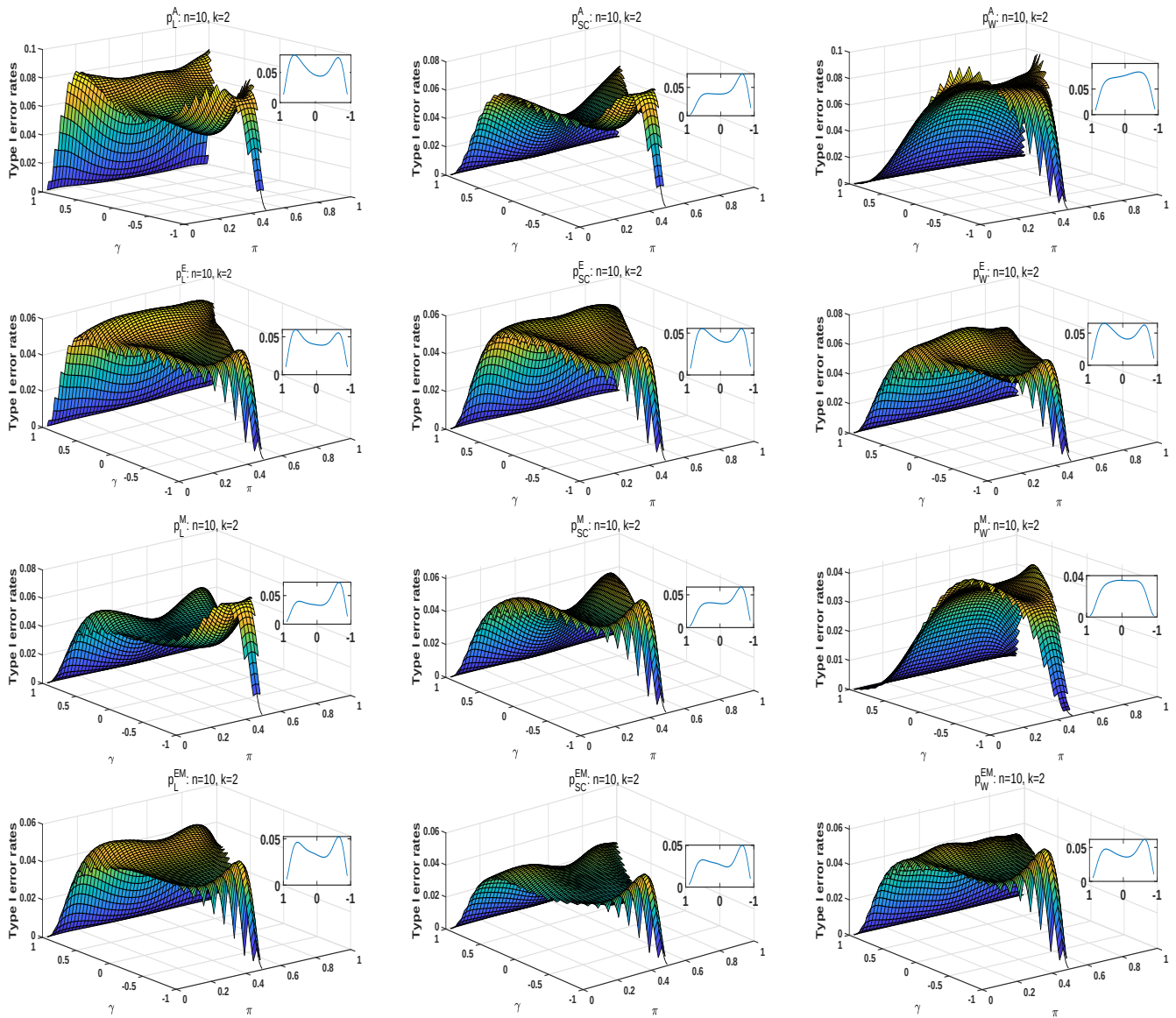


Figure 4. The surfaces and curves of type I error rates for exact methods under $K = 2$.

In summary, the exact E method can effectively improve the effectiveness of the homogeneity test of stratified AC_1 under small sample sizes and high γ . The E approaches under the likelihood ratio and score statistics perform better than that under the Wald-type statistic.

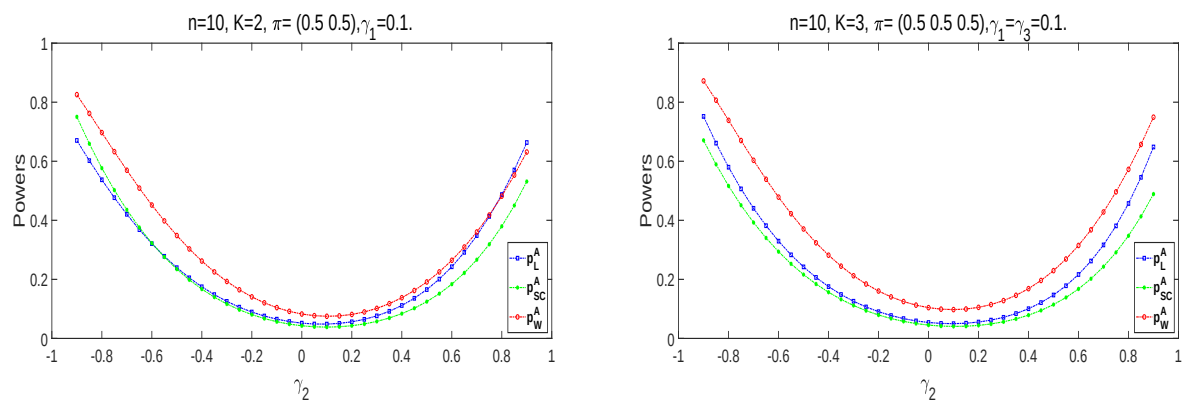


Figure 5. Cont.

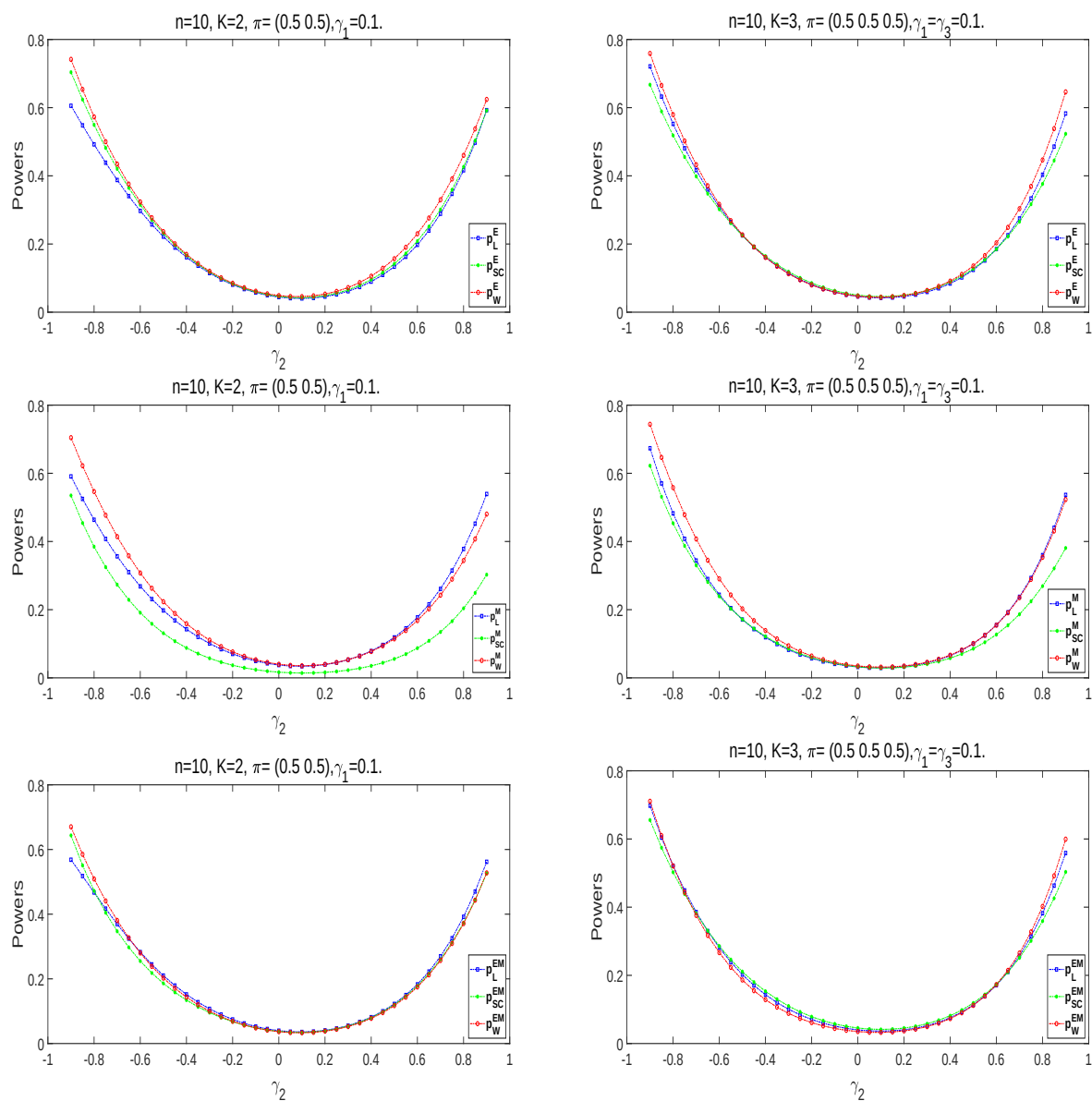


Figure 5. Power curves of exact methods for $K = 2, 3$.

6. Applications

We review the two real examples in the introduction. Table 3 shows the data structure among 83 twins. For the large sample size, the hypotheses of the homogeneity test for AC_1 across two strata are

$$H_0 : \gamma_1 = \gamma_2 \triangleq \gamma \text{ vs } H_a : \gamma_i \text{ is not all the same, } i = 1, 2.$$

By computation, the unconstrained MLEs of γ and π are $\hat{\gamma} = (0.4615, -0.0312)$ and $\hat{\pi} = (0.5000, 0.5161)$. The constrained MLEs of γ and π are $\tilde{\gamma} = 0.2788$ and $\tilde{\pi} = (0.5000, 0.5351)$. Table 12 provides the results of statistics and p -values. Given a significance level $\alpha = 0.05$, the values of test statistics T_L, T_{SC}, T_W are all larger than $\chi^2_{1,0.95} = 3.8415$. All the p -values of these three tests are smaller than 0.05. Thus, the null hypothesis H_0 is rejected at the significance level 0.05. There is a significant difference between two AC_1 coefficients, and we cannot merge the data of two strata to compute a common coefficient. Then, we need to investigate how zygosity affects the consistency of the alcohol-drinking status of male twins.

Table 12. Values of test statistics and p -value.

Value	Test Statistics		
	T_L	T_{SC}	T_W
Statistic value	5.0377	5.0762	5.1107
p -value	0.0248	0.0243	0.0238

Table 4 shows the small data structure for the clinical COVID-19 trial. The A, E, M, and E+M methods are applied to test $H_0 : \gamma_1 = \gamma_2 \triangleq \gamma$ vs $H_a : \gamma_1 \neq \gamma_2$. By calculation, the unconstrained MLEs of parameters γ and π are $\hat{\gamma} = (0.6656, 0.2197)$ and $\hat{\pi} = (0.6176, 0.6176)$. The constrained MLEs are $\tilde{\gamma} = 0.4537$ and $\tilde{\pi} = (0.5882, 0.6666)$. The values of statistics are $T_L(N^*) = 2.0150$, $T_{SC}(N^*) = 1.9674$, and $T_W(N^*) = 2.0805$. Table 13 shows the corresponding p -values of asymptotic and exact methods. The running time is about 8 min because of two strata and $n_1 = n_2 = 17$, referred to Table 8. There is no significant difference in stratified AC_1 for any approaches. Thus, we can merge the data in these two strata and estimate the value of common AC_1 by MLEs, which is 0.4537.

Table 13. Comparison of asymptotic and exact p -value.

Method	A Approach			E Approach			M Approach			E + M Approach		
	p_L^A	p_{SC}^A	p_W^A	p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}
p -value	0.1558	0.1607	0.1492	0.1953	0.1952	0.0854	0.2194	0.2076	0.2039	0.1989	0.1999	0.2127

7. Concluding Remarks

This article defines the stratified AC_1 coefficients as the object of study and constructs the likelihood function of the observed data. The primary purpose is to derive various statistics for testing the homogeneity of stratified AC_1 in the case of two raters with a binary outcome. We constructed asymptotic and exact methods for large and small sample sizes. Two asymptotic test statistics and their explicit expressions are derived for large sample sizes, including the likelihood ratio statistic (T_L) and Wald-type statistic (T_W). Meanwhile, the score statistic (T_{SC}) proposed by Honda and Ohyama [19] is also reviewed. Asymptotic p -values p_{θ}^A ($\theta = L, SC, W$) under the statistics mentioned above are denoted as the A approach. Three exact methods (E, M, and E+M) are proposed based on T_L , T_{SC} , and T_W for small sample sizes.

We conduct numerical studies to compare the performance of the above methods by type I error rates and powers. For large samples, the type I error rates of the likelihood ratio statistic are closer to the predetermined significance level of 0.05. The powers of statistics are better as the sample size increases. Overall, the likelihood ratio statistic is optimal among these three statistics in large sample sizes. However, asymptotic tests may generate unsatisfactory type I error rates at small sample scenarios and high γ . For small sample sizes in $K = 2$, p_L^A and p_W^A are liberal, and p_{SC}^A has the conservative type I error rates under different parameter settings. The type I error rates of the E approach are closer to the significance level of 0.05. The M and E+M approaches have conservative type I error rates. Moreover, p_L^E and p_{SC}^E are more robust than p_W^E . When $K = 3$, and γ is positive, the E approach has the robust type I error rates among three methods, in which p_L^E and p_{SC}^E perform better. The type I error rates of the exact E method are closer to 0.05 as the sample size increases. The E approach can improve the effectiveness of homogeneity tests at high γ . In the case of powers, the A approach has larger powers than these exact methods. Under the Wald-type statistic, each exact method has a higher power. Thus, the E approaches based on the likelihood ratio and score statistics have a robust performance for small sample sizes.

The proposed methods can be used not only in medical research but also in biometrics and psychological measurements. Meanwhile, exact methods can be applied to other data types, such as binary outcomes on multiple raters. This work focuses on constructing

parametric statistics through unconstrained and constrained MLEs. However, there are still many problems that need to be solved. For example, how to construct optimal tests? How to improve exact methods effectively for a larger K or sample size n_k ? How to simplify the heavy computations caused by the consideration of all the tables? More studies should be conducted on these problems and be extended in the future.

Author Contributions: Methodology, M.X.; software, M.X. and K.M.; writing—original draft preparation, M.X.; writing—review and editing, Z.L., K.M. and K.M.S.; supervision, Z.L.; visualization, M.X.; funding acquisition, Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No.: 12061070) and the Science and Technology Department of Xinjiang Uygur Autonomous Region (Grant No.: 2021D01E13).

Institutional Review Board Statement: Not applicable because of the study involved the development of statistical methods. Clinical examples retrospective data were published in Hannah et al. [16] and Hang et al. [20].

Informed Consent Statement: Not applicable.

Data Availability Statement: Clinical data referred to are from Hannah et al. [16] and Hang et al. [20].

Acknowledgments: We thank reviewers and editors for constructive and valuable advice for improving this article.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MLEs:	maximum likelihood estimates.	T_L :	likelihood ratio statistic.
T_{SC} :	score statistic.	T_W :	Wald-type statistic.
A:	asymptotic approach.	E:	E approach.
M:	M approach.	E+M:	E+M approach.
PVR:	proliferative vitreoretinopathy.		

Appendix A. Appendix

Appendix A.1. The Fisher Information Matrix of MLEs under H_0

Let I_1 be the $(K+1) \times (K+1)$ Fisher information matrix, and

$$\begin{aligned}
 I_1 = \begin{bmatrix} I_{111} & I_{112} \\ I_{112}^T & I_{122} \end{bmatrix} &= \begin{bmatrix} E\left(-\frac{\partial^2 l_0}{\partial \gamma^2}\right) & E\left(-\frac{\partial^2 l_0}{\partial \gamma \partial \pi_1}\right) & \cdots & E\left(-\frac{\partial^2 l_0}{\partial \gamma \partial \pi_K}\right) \\ E\left(-\frac{\partial^2 l_0}{\partial \gamma \partial \pi_1}\right) & E\left(-\frac{\partial^2 l_0}{\partial \pi_1^2}\right) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ E\left(-\frac{\partial^2 l_0}{\partial \gamma \partial \pi_K}\right) & \cdots & \cdots & E\left(-\frac{\partial^2 l_0}{\partial \pi_K^2}\right) \end{bmatrix} \\
 &= \begin{bmatrix} \sum_{k=1}^K \frac{n_k}{4} a_k^2 b_k & \frac{n_1 a_1 c_1}{2} & \cdots & \frac{n_K a_K c_K}{2} \\ \frac{n_1 a_1 c_1}{2} & n_1 d_1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \frac{n_K a_K c_K}{2} & \cdots & \cdots & n_K d_K \end{bmatrix},
 \end{aligned}$$

where

$$\begin{aligned} a_k &= 1 - 2\pi_k(1 - \pi_k), \\ b_k &= \frac{1}{P_{1k}(\gamma, \pi_k)} + \frac{4}{P_{2k}(\gamma, \pi_k)} + \frac{1}{P_{3k}(\gamma, \pi_k)}, \\ c_k &= \frac{1}{P_{1k}(\gamma, \pi_k)} - \frac{1}{P_{3k}(\gamma, \pi_k)} + (1 - \gamma)(1 - 2\pi_k)b_k, \\ d_k &= \frac{1}{P_{1k}(\gamma, \pi_k)} + \frac{1}{P_{3k}(\gamma, \pi_k)} + (1 - \gamma)(1 - 2\pi_k) \times \left(\frac{1}{P_{1k}(\gamma, \pi_k)} - \frac{1}{P_{3k}(\gamma, \pi_k)} + c_k \right) \end{aligned}$$

for $k = 1, 2, \dots, K$.

Appendix A.2. The Fisher Information Matrix I_2 of the Score Statistic

For the observed data $\mathbf{N} = (n_1, n_2, \dots, n_K)$, the Fisher information matrix of score statistic is

$$I_2 = \begin{bmatrix} I_{211} & I_{212} \\ I_{221} & I_{222} \end{bmatrix} = \begin{bmatrix} E\left(-\frac{\partial^2 l_1}{\partial \gamma_1^2}\right) & \cdots & \cdots & E\left(-\frac{\partial^2 l_1}{\partial \gamma_1 \partial \pi_1}\right) & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & E\left(-\frac{\partial^2 l_K}{\partial \gamma_K^2}\right) & \cdots & \cdots & E\left(-\frac{\partial^2 l_K}{\partial \gamma_K \partial \pi_K}\right) \\ E\left(-\frac{\partial^2 l_1}{\partial \gamma_1 \partial \pi_1}\right) & \cdots & \cdots & E\left(-\frac{\partial^2 l_1}{\partial \pi_1^2}\right) & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & E\left(-\frac{\partial^2 l_K}{\partial \gamma_K \partial \pi_K}\right) & \cdots & \cdots & E\left(-\frac{\partial^2 l_K}{\partial \pi_K^2}\right) \end{bmatrix},$$

Since $E(n_{lk}) = n_k P_{lk}(\gamma_k, \pi_k)$ ($l = 1, 2, 3$), we can obtain $E\left(-\frac{\partial^2 l_k}{\partial \gamma_k^2}\right) = \frac{n_k a_k^2 b_k}{4}$, $E\left(-\frac{\partial^2 l_k}{\partial \gamma_k \partial \pi_k}\right) = \frac{n_k a_k c_k}{2}$, and $E\left(-\frac{\partial^2 l_k}{\partial \pi_k^2}\right) = n_k d_k$. Thus, the Fisher information matrix I_2 is simplified as

$$I_2 = \frac{1}{4} \begin{bmatrix} n_1 a_1^2 b_1 & \cdots & \cdots & 2n_1 a_1 c_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & n_K a_K^2 b_K & \cdots & \cdots & 2n_K a_K c_K \\ 2n_1 a_1 c_1 & \cdots & \cdots & 4n_1 d_1 & \cdots & \cdots \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \cdots & \cdots & 2n_K a_K c_K & \cdots & \cdots & 4n_K d_K \end{bmatrix}.$$

Appendix A.3. The Simplification of T_W

To obtain the simplification of T_W , we calculate $CI_3^{-1}C^T$ and obtain

$$CI_3^{-1}C^T = \begin{bmatrix} g_1 + g_2 & -g_2 & 0 & 0 & \cdots & \cdots & 0 \\ -g_2 & g_2 + g_3 & -g_3 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \cdots & -g_{K-2} & g_{K-2} + g_{K-1} & -g_{K-1} \\ 0 & 0 & \cdots & \cdots & 0 & -g_{K-1} & g_{K-1} + g_K \end{bmatrix},$$

where $g_k = \frac{4d_k}{n_k a_k^2 (b_k d_k - c_k^2)}$. Obviously, $(CI_3^{-1}C^T)^{-1}$ is a symmetric matrix, in which

$$\begin{cases} (CI_3^{-1}C^T)_{i,j}^{-1} = g_{i+1} \cdots g_j \frac{h_{j+1} h_{K-1}}{s_i \cdots s_{K-1}}, & \forall j > i, \\ (CI_3^{-1}C^T)_{i,i}^{-1} = \frac{h_{i+1} h_{K-1}}{s_i \cdots s_{K-1}}, & \forall i, \end{cases}$$

where

$$\begin{cases} h_K = 1, h_{K-1} = g_{K-1} + g_K, \\ h_i = (g_i + g_{i+1} - \frac{g_{i+1}^2}{h_{i+1}}), i = 1, 2, \dots, K-2, \\ s_1 = g_1 + g_2, s_i = (g_i + g_{i+1}) - \frac{g_i^2}{s_{i-1}}, i = 2, \dots, K-1. \end{cases}$$

By computation, the simplified form of T_W is given as

$$T_W = \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} (\hat{\gamma}_i - \hat{\gamma}_{i+1})(\hat{\gamma}_j - \hat{\gamma}_{j+1})(CI_3^{-1}C^T)_{ij}^{-1}.$$

Appendix A.4. Other Tables and Figures

Table A1. Empirical type I error rates of asymptotic statistics for $K = 3$ (balanced sample sizes).

Balanced π Conditions						Unbalanced π Conditions							
$n_1 = n_2 = n_3$	γ	$\pi_1 = \pi_2 = \pi_3$	T_L	T_{SC}	T_W	$n_1 = n_2 = n_3$	γ	π_1	π_2	π_3	T_L	T_{SC}	T_W
10	0.1	0.3	0.0527	0.0454	0.1192	10	0.1	0.3	0.5	0.5	0.0515	0.0420	0.1030
	0.3		0.0576	0.0462	0.1069		0.3				0.0510	0.0393	0.0958
	0.5		0.0421	0.0337	0.0655		0.5				0.0338	0.0259	0.0627
	0.7		0.0152	0.0138	0.0167		0.7				0.0124	0.0114	0.0147
	0.9		0.0014	0.0014	0.0000		0.9				0.0005	0.0007	0.0001
	0.1	0.6	0.0512	0.0443	0.1015		0.1	0.6	0.4	0.2	0.0525	0.0439	0.1115
	0.3		0.0444	0.0341	0.0890		0.3				0.0529	0.0406	0.0974
	0.5		0.0307	0.0241	0.0574		0.5				0.0412	0.0315	0.0671
	0.7		0.0129	0.0118	0.0157		0.7				0.0161	0.0144	0.0204
	0.9		0.0006	0.0008	0.0001		0.9				0.0005	0.0008	0.0000
	0.1	0.5	0.0508	0.0413	0.0990		0.1	0.5	0.4	0.3	0.0578	0.0472	0.1078
	0.3		0.0437	0.0317	0.0857		0.3				0.0524	0.0416	0.0992
	0.5		0.0313	0.0223	0.0574		0.5				0.0379	0.0290	0.0662
	0.7		0.0086	0.0082	0.0127		0.7				0.0122	0.0115	0.0145
	0.9		0.0002	0.0003	0.0002		0.9				0.0004	0.0004	0.0001
50	0.1	0.3	0.0547	0.0528	0.0681	50	0.1	0.3	0.5	0.5	0.0494	0.0480	0.0600
	0.3		0.0498	0.0482	0.0600		0.3				0.0507	0.0477	0.0602
	0.5		0.0533	0.0514	0.0632		0.5				0.0513	0.0471	0.0586
	0.7		0.0548	0.0495	0.0606		0.7				0.0537	0.0479	0.0634
	0.9		0.0338	0.0289	0.0266		0.9				0.0272	0.0242	0.0221
	0.1	0.6	0.0480	0.0464	0.0583		0.1	0.6	0.4	0.2	0.0655	0.0646	0.0823
	0.3		0.0470	0.0450	0.0563		0.3				0.0563	0.0541	0.0671
	0.5		0.0517	0.0489	0.0590		0.5				0.0510	0.0480	0.0600
	0.7		0.0528	0.0457	0.0613		0.7				0.0577	0.0505	0.0628
	0.9		0.0237	0.0216	0.0211		0.9				0.0318	0.0286	0.0249
	0.1	0.5	0.0502	0.0476	0.0588		0.1	0.5	0.4	0.3	0.0485	0.0468	0.0598
	0.3		0.0483	0.0455	0.0586		0.3				0.0528	0.0499	0.0627
	0.5		0.0514	0.0477	0.0622		0.5				0.0531	0.0489	0.0601
	0.7		0.0526	0.0456	0.0588		0.7				0.0538	0.0470	0.0609
	0.9		0.0229	0.0220	0.0236		0.9				0.0281	0.0269	0.0248
100	0.1	0.3	0.0520	0.0515	0.0588	100	0.1	0.3	0.5	0.5	0.0509	0.0506	0.0562
	0.3		0.0518	0.0507	0.0571		0.3				0.0501	0.0491	0.0545
	0.5		0.0487	0.0480	0.0530		0.5				0.0495	0.0470	0.0531
	0.7		0.0523	0.0503	0.0542		0.7				0.0481	0.0455	0.0550
	0.9		0.0563	0.0473	0.0605		0.9				0.0524	0.0402	0.0618
	0.1	0.6	0.0492	0.0489	0.0543		0.1	0.6	0.4	0.2	0.0774	0.0772	0.0877
	0.3		0.0500	0.0484	0.0551		0.3				0.0542	0.0528	0.0619
	0.5		0.0579	0.0559	0.0631		0.5				0.0520	0.0509	0.0565
	0.7		0.0514	0.0499	0.0569		0.7				0.0529	0.0513	0.0539
	0.9		0.0529	0.0431	0.0610		0.9				0.0529	0.0430	0.0578
	0.1	0.5	0.0499	0.0487	0.0545		0.1	0.5	0.4	0.3	0.0497	0.0479	0.0546
	0.3		0.0475	0.0469	0.0519		0.3				0.0466	0.0461	0.0511
	0.5		0.0518	0.0496	0.0555		0.5				0.0478	0.0470	0.0513
	0.7		0.0525	0.0496	0.0581		0.7				0.0506	0.0478	0.0568
	0.9		0.0563	0.0428	0.0653		0.9				0.0555	0.0471	0.0617

Note: The type I error rates of robust region (0.04–0.06) are shown in bold.

Table A2. Empirical type I error rates of homogeneity tests for $K = 4$ (balanced sample sizes).

Balanced π Conditions						Unbalanced π Conditions									
$n_1 = n_2 = n_3 = n_4$	γ	$\pi_1 = \pi_2 = \pi_3 = \pi_4$	T_L	T_{SC}	T_W	$n_1 = n_2 = n_3 = n_4$	γ	π_1	π_2	π_3	π_4	T_L	T_{SC}	T_W	
10	0.1	0.3	0.0559	0.0437	0.1519	10	0.1	0.3	0.5	0.5	0.2	0.0541	0.0429	0.1287	
	0.3		0.0595	0.0465	0.1349		0.3					0.0466	0.0349	0.1166	
	0.5		0.0408	0.0325	0.0857		0.5					0.0353	0.0277	0.0763	
	0.7		0.0126	0.0137	0.0159		0.7					0.0104	0.0114	0.0147	
	0.9		0.0005	0.0015	0.0001		0.9					0.0005	0.0016	0.0001	
	0.1	0.6	0.0543	0.0455	0.1281		0.1	0.6	0.4	0.2	0.4	0.0594	0.0482	0.1387	
	0.3		0.0462	0.0345	0.1206		0.3					0.0495	0.0376	0.1175	
	0.5		0.0265	0.0204	0.0671		0.5					0.0375	0.0290	0.0783	
	0.7		0.0084	0.0096	0.0120		0.7					0.0108	0.0122	0.0150	
	0.9		0.0002	0.0006	0.0001		0.9					0.0001	0.0005	0.0000	
	0.1	0.5	0.0491	0.0404	0.1158		0.1	0.5	0.4	0.3	0.3	0.0583	0.0475	0.1433	
	0.3		0.0436	0.0340	0.1049		0.3					0.0533	0.0396	0.1233	
	0.5		0.0234	0.0193	0.0572		0.5					0.0328	0.0250	0.0768	
	0.7		0.0073	0.0093	0.0109		0.7					0.0095	0.0112	0.0145	
	0.9		0.0002	0.0008	0.0000		0.9					0.0005	0.0011	0.0002	
50	0.1	0.3	0.0531	0.0515	0.0694	50	0.1	0.3	0.5	0.5	0.2	0.0638	0.0617	0.0831	
	0.3		0.0504	0.0485	0.0659		0.3					0.0525	0.0501	0.0671	
	0.5		0.0532	0.0499	0.0691		0.5					0.0508	0.0472	0.0633	
	0.7		0.0577	0.0497	0.0682		0.7					0.0537	0.0466	0.0705	
	0.9		0.0304	0.0301	0.0256		0.9					0.0264	0.0265	0.0239	
	0.1	0.6	0.0511	0.0477	0.0662		0.1	0.6	0.4	0.2	0.4	0.0621	0.0608	0.0810	
	0.3		0.0525	0.0495	0.0660		0.3					0.0523	0.0497	0.0692	
	0.5		0.0566	0.0513	0.0688		0.5					0.0525	0.0485	0.0658	
	0.7		0.0549	0.0475	0.0726		0.7					0.0532	0.0457	0.0674	
	0.9		0.0191	0.0208	0.0185		0.9					0.0263	0.0260	0.0236	
	0.1	0.5	0.0433	0.0416	0.0578		0.1	0.5	0.4	0.3	0.3	0.0486	0.0461	0.0633	
	0.3		0.0477	0.0444	0.0593		0.3					0.0499	0.0473	0.0651	
	0.5		0.0502	0.0458	0.0601		0.5					0.0519	0.0472	0.0631	
	0.7		0.0549	0.0452	0.0739		0.7					0.0549	0.0476	0.0688	
	0.9		0.0223	0.0244	0.0206		0.9					0.0287	0.0288	0.0256	
100	0.1	0.3	0.0508	0.0499	0.0585	100	0.1	0.3	0.5	0.5	0.2	0.0805	0.0794	0.0921	
	0.3		0.0494	0.0481	0.0571		0.3					0.0551	0.0551	0.0631	
	0.5		0.0498	0.0485	0.0568		0.5					0.0471	0.0457	0.0531	
	0.7		0.0504	0.0488	0.0556		0.7					0.0531	0.0505	0.0595	
	0.9		0.0568	0.0464	0.0712		0.9					0.0513	0.0408	0.0743	
	0.1	0.6	0.0462	0.0454	0.0523		0.1	0.6	0.4	0.2	0.4	0.0794	0.0789	0.0960	
	0.3		0.0484	0.0473	0.0547		0.3					0.0551	0.0535	0.0647	
	0.5		0.0493	0.0479	0.0556		0.5					0.0498	0.0473	0.0566	
	0.7		0.0498	0.0472	0.0599		0.7					0.0484	0.0465	0.0548	
	0.9		0.0494	0.0401	0.0737		0.9					0.0586	0.0482	0.0804	
	0.1	0.5	0.0476	0.0464	0.0531		0.1	0.5	0.4	0.3	0.3	0.0499	0.0491	0.0579	
	0.3		0.0493	0.0469	0.0561		0.3					0.0499	0.0486	0.0582	
	0.5		0.0484	0.0459	0.0571		0.5					0.0495	0.0476	0.0577	
	0.7		0.0536	0.0471	0.0620		0.7					0.0523	0.0493	0.0584	
	0.9		0.0482	0.0407	0.0717		0.9					0.0587	0.0467	0.0783	

Note: The type I error rates of robust region (0.04–0.06) are shown in bold.

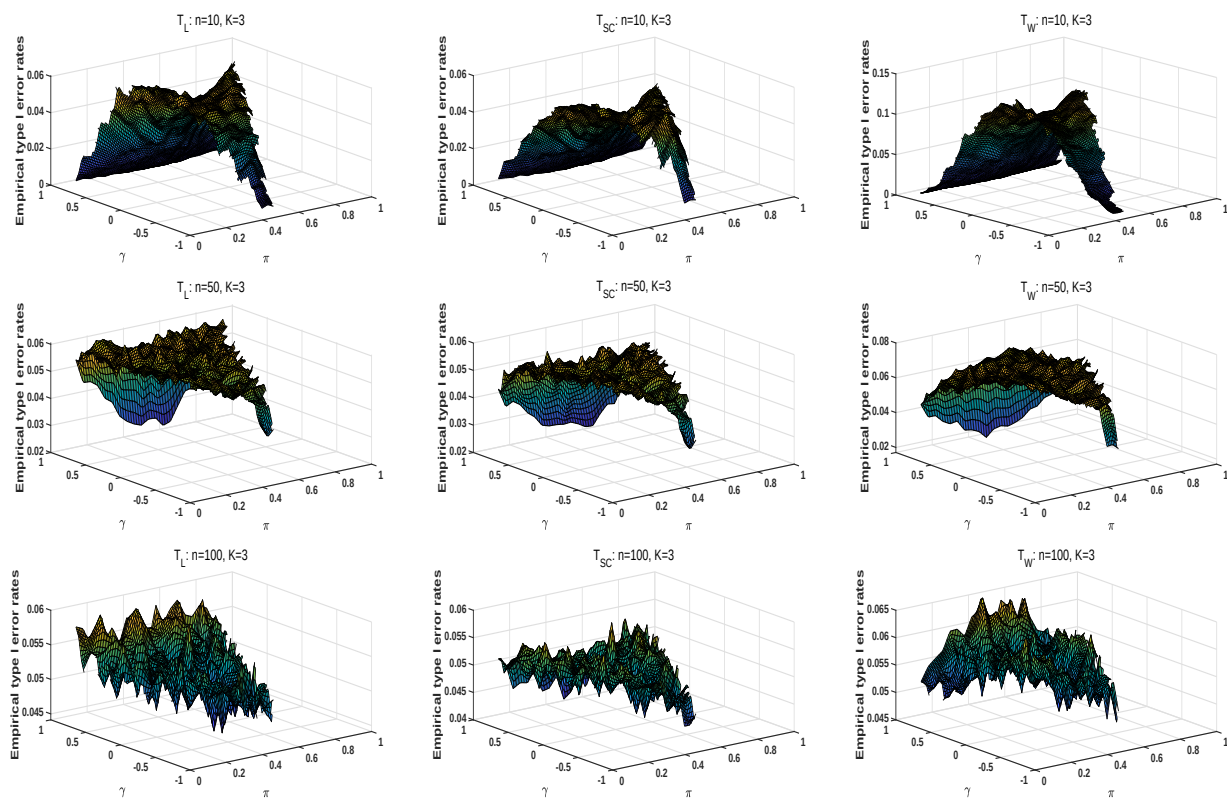


Figure A1. The surfaces of empirical type I error rates for asymptotic tests under $K = 3$.

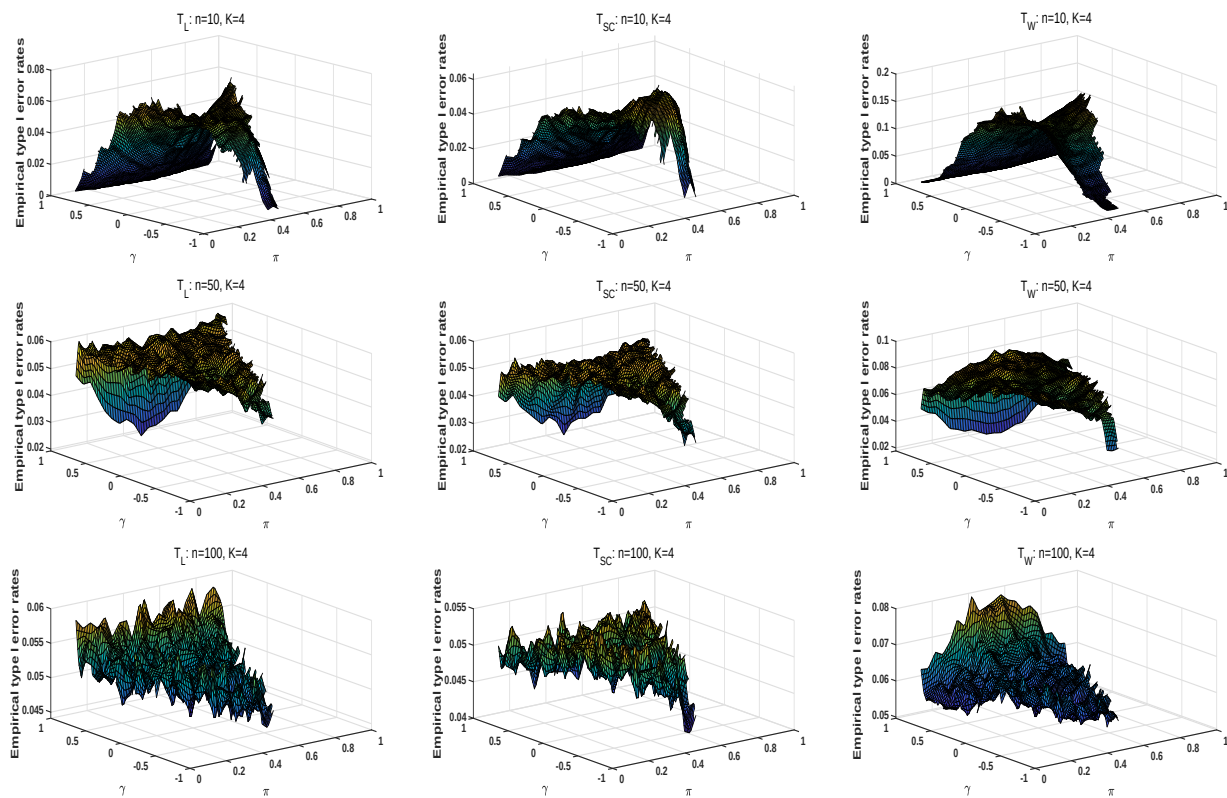


Figure A2. The surfaces of empirical type I error rates for asymptotic tests under $K = 4$.

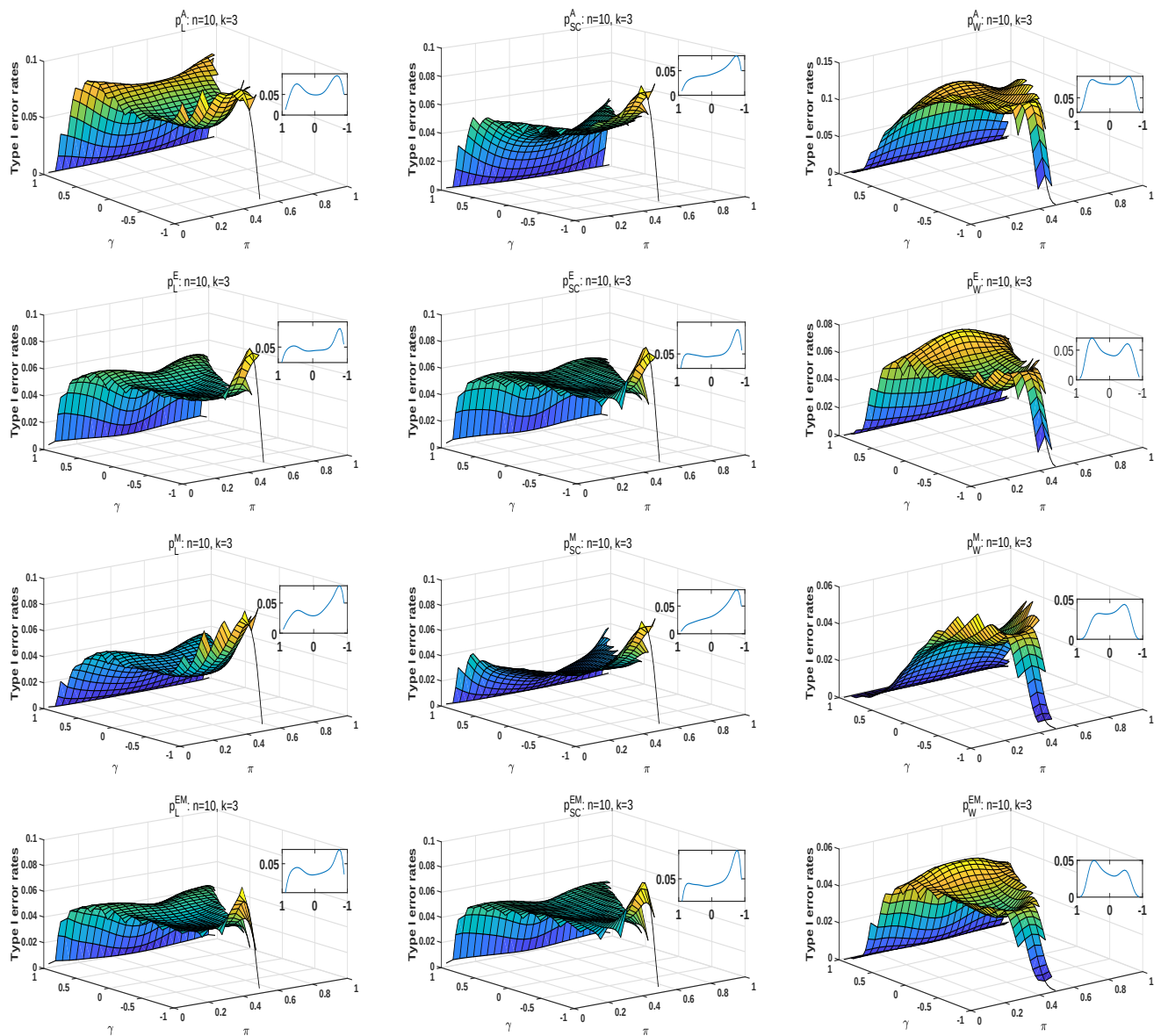


Figure A3. The surfaces and curves of type I error rates for exact methods under $K = 3$.

Table A3. Powers of exact and asymptotic methods for $K = 2$ (unbalanced π conditions).

γ_1	γ_2	π_1	π_2	Exact Methods						Asymptotic Methods					
				p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}	p_L^A	p_{SC}^A	p_W^A
0.1	0.5	0.3	0.5	0.1668	0.1656	0.1769	0.1466	0.1618	0.1341	0.1523	0.1383	0.1565	0.1808	0.1679	0.1945
	0.6			0.2400	0.2374	0.2520	0.2156	0.2318	0.1894	0.2208	0.2032	0.2263	0.2579	0.2386	0.2660
	0.7			0.3394	0.3341	0.3526	0.3110	0.3254	0.2629	0.3151	0.2937	0.3217	0.3619	0.3327	0.3599
	0.8			0.4720	0.4612	0.4844	0.4408	0.4469	0.3585	0.4431	0.4169	0.4493	0.4998	0.4540	0.4807
	0.9			0.6459	0.6253	0.6541	0.6146	0.6008	0.4816	0.6152	0.5807	0.6167	0.6807	0.6061	0.6341
0.3	0.6			0.1310	0.1310	0.1429	0.1112	0.1202	0.0850	0.1141	0.1038	0.1207	0.1496	0.1240	0.1483
	0.7			0.2027	0.2001	0.2161	0.1776	0.1819	0.1292	0.1799	0.1627	0.1879	0.2297	0.1862	0.2181
	0.8			0.3110	0.3030	0.3242	0.2806	0.2722	0.1946	0.2821	0.2528	0.2899	0.3508	0.2766	0.3186
	0.9			0.4707	0.4519	0.4792	0.4366	0.4001	0.2890	0.4374	0.3868	0.4402	0.5298	0.4038	0.4584
0.5	0.8	0.3	0.5	0.1636	0.1564	0.1692	0.1421	0.1228	0.0764	0.1440	0.1208	0.1462	0.2086	0.1255	0.1583
	0.9			0.2690	0.2518	0.2703	0.2378	0.1958	0.1233	0.2406	0.1987	0.2400	0.3427	0.1987	0.2456

Table A3. Cont.

γ_1	γ_2	π_1	π_2	Exact Methods						Asymptotic Methods					
				p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}	p_L^A	p_{SC}^A	p_W^A
0.1	0.5	0.5	0.4	0.1399	0.1501	0.1629	0.1143	0.1382	0.1110	0.1236	0.1209	0.1335	0.1612	0.1455	0.1823
	0.6			0.2089	0.2231	0.2406	0.1782	0.2056	0.1668	0.1875	0.1845	0.2013	0.2386	0.2140	0.2604
	0.7			0.3068	0.3248	0.3471	0.2718	0.2992	0.2462	0.2803	0.2763	0.2978	0.3462	0.3086	0.3645
	0.8			0.4428	0.4619	0.4868	0.4049	0.4249	0.3558	0.4125	0.4056	0.4312	0.4910	0.4349	0.4967
	0.9			0.6287	0.6403	0.6624	0.5899	0.5886	0.5042	0.5980	0.5836	0.6109	0.6803	0.5983	0.6557
0.3	0.6			0.1097	0.1194	0.1309	0.0927	0.1015	0.0747	0.0967	0.0943	0.1049	0.1332	0.1069	0.1373
	0.7			0.1768	0.1882	0.2036	0.1533	0.1594	0.1194	0.1585	0.1535	0.1682	0.2106	0.1664	0.2062
	0.8			0.2845	0.2947	0.3137	0.2517	0.2481	0.1895	0.2593	0.2485	0.2685	0.3316	0.2573	0.3066
	0.9			0.4527	0.4532	0.4734	0.4062	0.3795	0.2959	0.4189	0.3961	0.4218	0.5144	0.3908	0.4450
0.5	0.8			0.1542	0.1532	0.1633	0.1268	0.1094	0.0732	0.1348	0.1227	0.1344	0.1908	0.1160	0.1438
	0.9			0.2642	0.2540	0.2668	0.2182	0.1804	0.1232	0.2326	0.2092	0.2269	0.3195	0.1902	0.2265
0.1	0.5	0.6	0.4	0.1478	0.1552	0.1678	0.1210	0.1453	0.1163	0.1307	0.1255	0.1391	0.1679	0.1531	0.1878
	0.6			0.2184	0.2283	0.2453	0.1865	0.2141	0.1721	0.1961	0.1894	0.2076	0.2458	0.2227	0.2650
	0.7			0.3170	0.3294	0.3504	0.2813	0.3085	0.2500	0.2896	0.2807	0.3041	0.3529	0.3177	0.3672
	0.8			0.4517	0.4645	0.4876	0.4150	0.4338	0.3560	0.4212	0.4078	0.4364	0.4959	0.4430	0.4970
	0.9			0.6325	0.6398	0.6602	0.5992	0.5950	0.4975	0.6040	0.5808	0.6130	0.6827	0.6029	0.6544
0.3	0.6			0.1140	0.1225	0.1337	0.0971	0.1065	0.0771	0.1008	0.0969	0.1079	0.1373	0.1119	0.1407
	0.7			0.1816	0.1918	0.2066	0.1595	0.1659	0.1217	0.1639	0.1562	0.1718	0.2158	0.1726	0.2101
	0.8			0.2884	0.2982	0.3161	0.2598	0.2562	0.1906	0.2654	0.2502	0.2719	0.3377	0.2642	0.3108
	0.9			0.4527	0.4555	0.4741	0.4160	0.3885	0.2940	0.4245	0.3948	0.4236	0.5212	0.3973	0.4495
0.5	0.8			0.1541	0.1556	0.1648	0.1313	0.1142	0.0739	0.1374	0.1234	0.1349	0.1960	0.1199	0.1468
	0.9			0.2618	0.2568	0.2681	0.2248	0.1873	0.1235	0.2360	0.2089	0.2266	0.3281	0.1952	0.2305

Table A4. Powers of exact and asymptotic methods for $K = 3$ (balanced π conditions).

$\gamma_1 = \gamma_3$	γ_2	$\pi_1 = \pi_2 = \pi_3$	Exact Methods						Asymptotic Methods					
			p_L^E	p_{SC}^E	p_W^E	p_L^M	p_{SC}^M	p_W^M	p_L^{EM}	p_{SC}^{EM}	p_W^{EM}	p_L^A	p_{SC}^A	p_W^A
0.1	0.5	0.3	0.1351	0.1389	0.1421	0.1049	0.0929	0.1089	0.1216	0.1274	0.1150	0.1596	0.1318	0.2366
	0.6		0.1983	0.2002	0.2089	0.1606	0.1383	0.1583	0.1803	0.1845	0.1735	0.2316	0.1901	0.3167
	0.7		0.2905	0.2860	0.3075	0.2460	0.2047	0.2328	0.2677	0.2661	0.2635	0.3342	0.2729	0.4258
	0.8		0.4222	0.4028	0.4504	0.3737	0.2991	0.3433	0.3959	0.3799	0.3994	0.4762	0.3872	0.5682
	0.9		0.6067	0.5572	0.6533	0.5607	0.4296	0.5051	0.5807	0.5349	0.6011	0.6684	0.5407	0.7462
0.3	0.6	0.5	0.1055	0.1044	0.1183	0.0862	0.0662	0.0686	0.0951	0.0959	0.0949	0.1328	0.0967	0.1829
	0.7		0.1642	0.1548	0.1858	0.1391	0.1008	0.1081	0.1499	0.1436	0.1534	0.2035	0.1439	0.2632
	0.8		0.2602	0.2319	0.2989	0.2275	0.1543	0.1744	0.2404	0.2174	0.2531	0.3171	0.2166	0.3880
	0.9		0.4124	0.3459	0.4827	0.3707	0.2345	0.2822	0.3859	0.3282	0.4177	0.4943	0.3252	0.5742
0.5	0.8	0.6	0.1351	0.1169	0.1550	0.1087	0.0666	0.0565	0.1205	0.1085	0.1177	0.1799	0.1006	0.2119
	0.9		0.2198	0.1787	0.2595	0.1785	0.1027	0.0951	0.1970	0.1670	0.1991	0.2897	0.1538	0.3369
0.1	0.5	0.5	0.1241	0.1273	0.1355	0.1002	0.0855	0.1008	0.1131	0.1186	0.1119	0.1470	0.1146	0.2295
	0.6		0.1851	0.1847	0.2035	0.1549	0.1270	0.1544	0.1710	0.1735	0.1729	0.2164	0.1675	0.3148
	0.7		0.2748	0.2656	0.3035	0.2377	0.1864	0.2352	0.2573	0.2515	0.2655	0.3166	0.2430	0.4281
	0.8		0.4032	0.3762	0.4466	0.3600	0.2691	0.3535	0.3822	0.3590	0.4022	0.4569	0.3476	0.5725
	0.9		0.5829	0.5231	0.6462	0.5366	0.3809	0.5230	0.5588	0.5033	0.5995	0.6483	0.4888	0.7493
0.3	0.6	0.6	0.1004	0.0956	0.1207	0.0808	0.0535	0.0723	0.0904	0.0878	0.0981	0.1298	0.0808	0.1884
	0.7		0.1546	0.1408	0.1877	0.1274	0.0800	0.1138	0.1407	0.1302	0.1556	0.1969	0.1202	0.2698
	0.8		0.2406	0.2083	0.2960	0.2026	0.1198	0.1810	0.2210	0.1938	0.2498	0.3026	0.1798	0.3930
	0.9		0.3727	0.3057	0.4650	0.3201	0.1777	0.2861	0.3451	0.2866	0.3993	0.4634	0.2675	0.5732
0.5	0.8	0.6	0.1185	0.1056	0.1532	0.0862	0.0481	0.0577	0.1051	0.0954	0.1079	0.1634	0.0801	0.2225
	0.9		0.1841	0.1566	0.2393	0.1352	0.0722	0.0910	0.1643	0.1426	0.1710	0.2517	0.1202	0.3353
0.1	0.5	0.6	0.1278	0.1307	0.1389	0.1021	0.0892	0.1028	0.1156	0.1214	0.1128	0.1502	0.1200	0.2328
	0.6		0.1895	0.1891	0.2066	0.1580	0.1329	0.1560	0.1743	0.1771	0.1738	0.2201	0.1746	0.3180
	0.7		0.2802	0.2713	0.3063	0.2427	0.1958	0.2361	0.2615	0.2565	0.2664	0.3208	0.2523	0.4312
	0.8		0.4101	0.3837	0.4492	0.3678	0.2835	0.3534	0.3882	0.3662	0.4036	0.4616	0.3598	0.5753
	0.9		0.5917	0.5329	0.6489	0.5486	0.4025	0.5213	0.5676	0.5136	0.6019	0.6536	0.5043	0.7510
0.3	0.6	0.6	0.1023	0.0984	0.1197	0.0833	0.0583	0.0711	0.0922	0.0903	0.0973	0.1300	0.0853	0.1879
	0.7		0.1583	0.1452	0.1868	0.1324	0.0876	0.1126	0.1442	0.1345	0.1556	0.1983	0.1269	0.2695
	0.8		0.2478	0.2156	0.2964	0.2123	0.1320	0.1804	0.2280	0.2014	0.2523	0.3063	0.1901	0.3933
	0.9		0.3865	0.3180	0.4692	0.3381	0.1970	0.2873	0.3585	0.2995	0.4069	0.4721	0.2831	0.5745
0.5	0.8	0.6	0.1247	0.1088	0.1537	0.0936	0.0542	0.0580	0.1104	0.0996	0.1125	0.1677	0.0850	0.2203
	0.9		0.1958	0.1627	0.2446	0.1484	0.0819	0.0932	0.1744	0.1499	0.1813	0.2616	0.1280	0.3366

Table A5. Powers of exact and asymptotic methods for $K = 3$ (unbalanced π conditions).

$\gamma_1 = \gamma_3$	γ_2	π_1	π_2	π_3	Exact Methods						Asymptotic Methods					
					p_L^E	p_{Sc}^E	p_W^E	p_L^M	p_{Sc}^M	p_W^M	p_L^{EM}	p_{Sc}^{EM}	p_W^{EM}	p_L^A	p_{Sc}^A	p_W^A
0.1	0.5	0.3	0.5	0.5	0.1668	0.1656	0.1769	0.1466	0.1618	0.1341	0.1523	0.1383	0.1565	0.1620	0.1278	0.2409
	0.6				0.2400	0.2374	0.2520	0.2156	0.2318	0.1894	0.2208	0.2032	0.2263	0.2333	0.1832	0.3249
	0.7				0.3394	0.3341	0.3526	0.3110	0.3254	0.2629	0.3151	0.2937	0.3217	0.3341	0.2607	0.4357
	0.8				0.4720	0.4612	0.4844	0.4408	0.4469	0.3585	0.4431	0.4169	0.4493	0.4729	0.3662	0.5766
	0.9				0.6459	0.6253	0.6541	0.6146	0.6008	0.4816	0.6152	0.5807	0.6167	0.6598	0.5064	0.7498
	0.6				0.1310	0.1310	0.1429	0.1112	0.1202	0.0850	0.1141	0.1038	0.1207	0.1373	0.0892	0.1970
	0.7				0.2027	0.2001	0.2161	0.1776	0.1819	0.1292	0.1799	0.1627	0.1879	0.2067	0.1309	0.2800
	0.8				0.3110	0.3030	0.3242	0.2806	0.2722	0.1946	0.2821	0.2528	0.2899	0.3144	0.1931	0.4037
	0.9				0.4707	0.4519	0.4792	0.4366	0.4001	0.2890	0.4374	0.3868	0.4402	0.4767	0.2831	0.5823
0.5	0.8				0.1636	0.1564	0.1692	0.1421	0.1228	0.0764	0.1440	0.1208	0.1462	0.1712	0.0853	0.2283
	0.9				0.2690	0.2518	0.2703	0.2378	0.1958	0.1233	0.2406	0.1987	0.2400	0.2631	0.1266	0.3448
0.1	0.5	0.5	0.4	0.3	0.1399	0.1501	0.1629	0.1143	0.1382	0.1110	0.1236	0.1209	0.1335	0.1576	0.1252	0.2352
	0.6				0.2089	0.2231	0.2406	0.1782	0.2056	0.1668	0.1875	0.1845	0.2013	0.2282	0.1806	0.3183
	0.7				0.3068	0.3248	0.3471	0.2718	0.2992	0.2462	0.2803	0.2763	0.2978	0.3288	0.2588	0.4292
	0.8				0.4428	0.4619	0.4868	0.4049	0.4249	0.3558	0.4125	0.4056	0.4312	0.4686	0.3665	0.5713
	0.9				0.6287	0.6403	0.6624	0.5899	0.5886	0.5042	0.5980	0.5836	0.6109	0.6583	0.5107	0.7468
	0.6				0.1097	0.1194	0.1309	0.0927	0.1015	0.0747	0.0967	0.0943	0.1049	0.1337	0.0889	0.1911
	0.7				0.1768	0.1882	0.2036	0.1533	0.1594	0.1194	0.1585	0.1535	0.1682	0.2027	0.1314	0.2729
	0.8				0.2845	0.2947	0.3137	0.2517	0.2481	0.1895	0.2593	0.2485	0.2685	0.3113	0.1954	0.3966
	0.9				0.4527	0.4532	0.4734	0.4062	0.3795	0.2959	0.4189	0.3961	0.4218	0.4771	0.2891	0.5771
0.5	0.8				0.1542	0.1532	0.1633	0.1268	0.1094	0.0732	0.1348	0.1227	0.1344	0.1709	0.0872	0.2225
	0.9				0.2642	0.2540	0.2668	0.2182	0.1804	0.1232	0.2326	0.2092	0.2269	0.2660	0.1307	0.3400
0.1	0.5	0.6	0.4	0.3	0.1478	0.1552	0.1678	0.1210	0.1453	0.1163	0.1307	0.1255	0.1391	0.1616	0.1292	0.2398
	0.6				0.2184	0.2283	0.2453	0.1865	0.2141	0.1721	0.1961	0.1894	0.2076	0.2330	0.1853	0.3234
	0.7				0.3170	0.3294	0.3504	0.2813	0.3085	0.2500	0.2896	0.2807	0.3041	0.3342	0.2642	0.4344
	0.8				0.4517	0.4645	0.4876	0.4150	0.4338	0.3560	0.4212	0.4078	0.4364	0.4740	0.3722	0.5760
	0.9				0.6325	0.6398	0.6602	0.5992	0.5950	0.4975	0.6040	0.5808	0.6130	0.6624	0.5161	0.7501
	0.6				0.1140	0.1225	0.1337	0.0971	0.1065	0.0771	0.1008	0.0969	0.1079	0.1358	0.0914	0.1931
	0.7				0.1816	0.1918	0.2066	0.1595	0.1659	0.1217	0.1639	0.1562	0.1718	0.2057	0.1347	0.2757
	0.8				0.2884	0.2982	0.3161	0.2598	0.2562	0.1906	0.2654	0.2502	0.2719	0.3151	0.1998	0.4000
	0.9				0.4527	0.4555	0.4741	0.4160	0.3885	0.2940	0.4245	0.3948	0.4236	0.4817	0.2947	0.5806
0.5	0.8				0.1541	0.1556	0.1648	0.1313	0.1142	0.0739	0.1374	0.1234	0.1349	0.1737	0.0893	0.2241
	0.9				0.2618	0.2568	0.2681	0.2248	0.1873	0.1235	0.2360	0.2089	0.2266	0.2704	0.1335	0.3431

References

- Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [\[CrossRef\]](#)
- Scott, W.A. Reliability of content analysis: The case of nominal scale coding. *Public Opin. Q.* **1955**, *19*, 321–325. [\[CrossRef\]](#)
- Feinstein, A.R.; Cicchetti, D.V. High agreement but low Kappa: I. The problems of two paradoxes. *J. Clin. Epidemiol.* **1990**, *43*, 543–549. [\[CrossRef\]](#) [\[PubMed\]](#)
- Vach, W. The dependence of Cohen's Kappa on the prevalence does not matter. *J. Clin. Epidemiol.* **2005**, *58*, 655–661. [\[CrossRef\]](#)
- Holley, J.W.; Guilford, J.P. A note on the G index of agreement. *Educ. Psychol. Meas.* **1964**, *24*, 749–753. [\[CrossRef\]](#)
- Aickin, M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's Kappa. *Biometrics* **1990**, *46*, 293–302. [\[CrossRef\]](#)
- Andrés, A.; Marzo, P.F. Delta: A new measure of agreement between two raters. *Br. J. Math. Stat. Psychol.* **2004**, *57*, 1–19. [\[CrossRef\]](#)
- Gwet, K.L. Computing inter-rater reliability and its variance in the presence of high agreement. *Br. J. Math. Stat. Psychol.* **2008**, *61*, 29–48. [\[CrossRef\]](#)
- Byrt, T.; Bishop, J.; Carlin, J.B. Bias, prevalence and Kappa. *J. Clin. Epidemiol.* **1993**, *46*, 423–429. [\[CrossRef\]](#)
- Shankar, V.; Bangdiwala, S.I. Observer agreement paradoxes in 2×2 tables: Comparison of agreement measures. *BMC Med Res. Methodol.* **2014**, *14*, 1–9. [\[CrossRef\]](#)
- Wongpakaran, N.; Wongpakaran, T.; Wedding, D.; Gwet, K.L. A comparison of Cohen's Kappa and Gwet's AC₁ when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Med Res. Methodol.* **2013**, *13*. [\[CrossRef\]](#)
- Ohayama, T. Statistical inference of agreement coefficient between two raters with binary outcomes. *Commun. Stat. Theory Methods* **2019**, *49*, 2529–2539. [\[CrossRef\]](#)
- Vanbelle, S.; Albert, A. Agreement between two independent groups of raters. *Psychometrika* **2009**, *74*, 477–491. [\[CrossRef\]](#)
- Alencar, L.M.; Zangwill, L.M.; Weinreb, R.N.; Bowd, C.; Vizzeri, G.; Sample, P.A.; Jr, R.S.; Medeiros, F.A. Agreement for detecting glaucoma progression with the GDx guided progression analysis, automated perimetry, and optic disc photography. *Ophthalmology* **2010**, *117*, 462–470. [\[CrossRef\]](#) [\[PubMed\]](#)
- Henninger, B.; Kaser, V.; Ostermann, S.; Spicher, A.; Zegg, M.; Schmid, R.; Kremser, C.; Krappinger, D. Cervical disc and ligamentous injury in hyperextension trauma: MRI and intraoperative correlation. *J. Neuroimaging* **2020**, *30*, 104–109. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hannah, M.C.; Hopper, J.L.; Mathews, J.D. Twin concordance for a binary trait. II. Nested analysis of ever-smoking and ex-smoking traits and unnested analysis of a 'Committed-Smoking' trait. *Am. J. Hum. Genet.* **1985**, *37*, 153–165.

17. Nam, J.M. Assessment on homogeneity tests for kappa statistics under equal prevalence across studies in reliability. *Stat. Med.* **2010**, *25*, 1521–1531. [[CrossRef](#)]
18. Agresti, A. *Categorical Data Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2002.
19. Honda, C.; Ohyama, T. Homogeneity score test of AC_1 statistics and estimation of common AC_1 in multiple or stratified inter-rater agreement studies. *BMC Med Res. Methodol.* **2020**, *20*, 20. [[CrossRef](#)]
20. Huang, S.; Wu, Y.; Zhu, G.; Ji, J.; Lu, R. Clinical value of ELISA and gold-standard in the diagnosis of corona virus disease 2019. *Traffic Med.* **2020**, *34*, 119–120.
21. Fisher, R.A. *The Design of Experiments*; Oliver and Boyd: Edinburgh, Scotland, 1935.
22. Mehta, C.R.; Patel, N.R. A network algorithm for the exact treatment of the $2 \times k$ contingency table. *Commun. Stat. - Simul. Comput.* **1980**, *9*, 649–664. [[CrossRef](#)]
23. Upton, G.J.G. Fisher's exact test. *J. R. Stat. Soc. Ser. A (Stat. Soc.)*. **1992**, *155*, 395–402. [[CrossRef](#)]
24. Mehta, C.R.; Patel, N.R. A network algorithm for performing Fisher's exact test in $r \times c$ contingency tables. *Publ. Am. Stat. Assoc.* **1983**, *78*, 427–434. [[CrossRef](#)]
25. Mehta, C.R.; Patel, N.R. Algorithm 643 FEXACT: A FORTRAN subroutine for Fisher's exact test on unordered $r \times c$ contingency tables. *ACM Trans. Math. Softw.* **1986**, *12*, 154–161. [[CrossRef](#)]
26. Mehta, C.R.; Patel, N.R. A hybrid algorithm for Fisher's exact test in unordered $r \times c$ contingency tables. *Commun. Stat.-Theory Methods* **1986**, *15*, 387–404. [[CrossRef](#)]
27. Mehta, C.R.; Patel, N.R.; Senchaudhuri, P. Exact power and sample-size computations for the Cochran-Armitage trend test. *Biometrics* **1998**, *54*, 1615–1621. [[CrossRef](#)]
28. Liddell, D. Practical tests of 2×2 contingency tables. *J. R. Stat. Soc. Ser.* **1976**, *25*, 295–304. [[CrossRef](#)]
29. Storer, B.E.; Kim, C. Exact properties of some exact test statistics for comparing two binomial proportions. *J. Am. Stat. Assoc.* **1990**, *85*, 146–155. [[CrossRef](#)]
30. Basu, D. On the elimination of nuisance parameters. *J. Am. Stat. Assoc.* **1977**, *72*, 355–366. [[CrossRef](#)]
31. Lloyd, C.J. Exact p -value for discrete models obtained by estimation and maximization. *Aust. New Zealand J. Stat.* **2008**, *50*, 329–345. [[CrossRef](#)]
32. Tang, M.L.; Ng, H.K.T.; Guo, J.; Chan, W.; Chan, B.P.S. Exact Cochran-Armitage trend tests: Comparisons under different models. *J. Stat. Comput. Simul.* **2006**, *76*, 847–859. [[CrossRef](#)]
33. Shan, G.; Wilding, G.E. Powerful Exact Unconditional Tests for Agreement between Two Raters with Binary Endpoints. *PLoS ONE* **2014**, *9*, e97386. [[CrossRef](#)] [[PubMed](#)]
34. Engle, R.F. Chapter 13, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handb. Econom.* **1984**, *2*, 775–826.
35. Tang, M.L.; Tang, N.S.; Rosner, B. Statistical inference for correlated data in ophthalmologic studies. *Stat. Med.* **2006**, *25*, 2771–2783. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.