Journal of Biostatistics and Biometric Applications
Volume 6 | Issue 1
ISSN:2455-765X

Research Article Open Access

Consistent Confidence Limits, P Values, and Power of the Non-Conservative, Size – α Modified Fisher Exact Test

van der Meulen EA[*1], Raymond K[2] and van der Meulen PJ[3]

[1]Ferring Pharmaceuticals A/S, Copenhagen, Denmark
[2]GenMab A/S, Copenhagen, Denmark
[3]Edinburgh, Scotland

[*]**Corresponding author:** van der Meulen EA, Ferring Pharmaceuticals A/S, Copenhagen, Denmark, Tel: +45 28787424, E-mail: evdm@ferring.com

Citation: van der Meulen EA, Raymond K, van der Meulen PJ (2021) Consistent Confidence Limits, P Values, and Power of the Non-Conservative, Size – α Modified Fisher Exact Test. J Biostat Biometric App 6(1):102

Received Date: March 22, 2021 **Accepted Date:** April 16, 2021 **Published Date:** April 19, 2021

Abstract

The classical Fisher exact test [1], which is unconditionally the uniformly most powerful unbiased (UMPU) test, requires randomization at the critical value(s) to be of size α. Obviously, one needs a non-randomized version of this. Rejecting the null only if the test-statistic's outcome is more extreme than the critical values, reduces the actual size considerably. The modified Fisher exact test introduced in [2] additionally rejects the null when the test attains the critical value *c(t)* and the randomization probability γ(t) (that depend on the total number of successes *T*) exceeds a threshold γ_0, which is determined such that, for all values of the nuisance parameter, the size of the unconditional modified test is smaller, but as close as possible to α. This greatly improves the size and power of the test as compared to, for example, the conservative nonrandomized Fisher exact test, while controlling the Type 1 error rate.

Without corresponding *p*-values and confidence intervals, however, the test is of limited practical use. This paper aims to address this deficiency, providing associated agreeing test-based two-sided *p*-values for H_0: $\theta=\theta_0$ and two-sided confidence intervals for θ_0 (the log odds-ratio) as well as the power of this, non-conservatively sized test. A SAS IML macro is provided in the appendix, and an R-package will shortly be released. Woolf's asymptotic test closely resembles this test but does not strictly control the Type 1 error rate. Currently used exact tests, are potentially disagreeing, may not control the Type 1 error or are overly conservative in size, an example of which is the "exact" test implemented in SAS Proc FREQ, where the *p*-value is constructed as the sum of less or equally likely conditional outcomes. Besides, this exact test is less powerful than the proposed modified Fisher exact test.

Keywords: Modified Fisher Exact; Test-based *p*-values; Test-based Confidence Intervals; Power; Neyman-structure; Uniformly most powerful unbiased tests

Introduction

The Fisher exact test [1] for testing equality of two-binomial success probabilities is one of the most applied tests in biostatistical sciences. While some seem to think it is just a conditional test, it is the uniformly most powerful test among the unbiased tests [2], a property first proven by Tocher [3], and well described by Lehmann [4]. A test is unbiased if it is of size α, and has power greater or equal than α. Although not a compelling property, it is natural to require that the probability of a false positive should be smaller than that of a true positive. Suissa and Shuster [5] argue that a test can be more powerful once you permit it to have a power less than the nominal level of significance locally for alternatives that are not of interest, i.e., are not clinically relevant treatment differences. However, one could then easily counter argue that, if this is the case, the null hypothesis is ill-defined, and should be defined as the odds ratio being smaller than the minimum clinically relevant odds ratio – at which point unbiasedness is again in play.

The downside of the Fisher exact test in this theoretical context is that it is of limited practical use, primarily since it requires randomization at the critical values. As such, it does not have a natural accompanying *p*-value and confidence interval. Many suggestions to construct a non-randomized version of the Fisher exact test and corresponding *p*-values and confidence intervals have been made. Boschloo [6] suggested to reject based on conditional *p*-values using raised conditional nominal levels, such that the overall level is as close to (but smaller or equal to) α as possible. McDonald *et al.* [7] suggested raising the overall unconditional size, such that the size of the non-randomized Fisher exact test is a close to (but smaller or equal to) α as possible. A practical and much advocated alternative is the use of mid *p*-values, introduced by Lancaster [8] and first applied to 2x2 contingency tables by Haber [9]. Upton [10] and Lydersen and Laake [11] also both advocate the mid-*p*-value associated with the Fisher exact test. The recognized drawback of the mid-*p*-value is, like with Pearson's Chi-square test [12] although to a much lesser extent, that the size is not guaranteed to be below the nominal level of significance α.

The contemporary and much used *p*-values and confidence intervals are those implemented in software packages, such as SAS Proc FREQ. This SAS procedure has an exact option where the two-sided *p*-value is constructed as the sum of the probabilities of outcomes that are less or equally likely than the observed response, conditionally on the total number of responses. The confidence interval for the odds ratio is, however, not constructed utilizing this *p*-value but instead uses one-sided equal tail *p*-values to extract confidence limits. Consequently, its *p*-value and confidence interval are potentially disagreeing. It generally also offers the well-known asymptotic version of *p*-values and confidence intervals introduced by Woolf [13]. The latter, while easily constructed as it is based on a simple closed formula, is not controlling the Type 1 error rate, although *p*-values and confidence intervals are by definition agreeing.

In this paper we revisit the modified Fisher exact test introduced in [2] and propose associated *p*-values and confidence intervals that are agreeing by definition, as they are both fully test-based. The test is derived from the randomized Fisher exact test by only rejecting the null if, in case the test-statistic attains the critical value, the randomization probability exceeds a certain limit, γ_0, which is determined such that the maximum size is as large as possible but smaller than the required nominal level α. As *p*-values and confidence intervals must be, by regulatory guidance, two-sided, the focus will only be on the two-sided testing problem H_0: $\theta=\theta_0$ versus $H_1 : \theta\neq\theta_0$, where θ is the log odds ratio. A SAS/IML macro is provided in the appendix and a R-package will shortly be available that both will provide associated *p*-values for any null (not just $H_0 : \theta=1$), two-sided 1-α confidence intervals for the odds ratio (e^{θ}), a plot of the size of the test as a function of the nuisance parameter which can be used as a diagnostic tool, and the power of the test. This test will be compared to Woolf's asymptotic test as well as the popular "exact" test implemented in SAS Proc FREQ and other software packages.

In Section (The Two-Sided Fisher Exact Test) we will describe the two-sided Fisher exact test and how to obtain its critical values and associated randomization probabilities. Next, in Section (The Modified Two-Sided Fisher Exact Test), we re-introduce the modified Fisher exact test [2] and explain how to derive γ_0 numerically. In subsequent sections we compare its size (Section - Size Comparisons to Other Contemporary Tests), test-based confidence limits and *p*-values (Section *p*-Values and Confidence Limits), and its power (Section Power Comparison) against some if it's contemporary popular alternatives. Discussion and conclusions are presented in Section (Discussion and Conclusion) and a SAS IML macro is provided in the Appendix.

The Two-Sided Fisher Exact Test

Tocher [3] proved that the Fisher exact test is uniformly most powerful among all unbiased test of nominal level α. The general line of reasoning, for ease of discussion restricted to two-binomials (but extendable to the contingency table approach), is as follows.

Let, respectively, U and V be independent binomial variables, with U~Bin(m, π_1) and V~Bin(n, π_2), and let $T=U+V$ be the total number of successes. For constructing a confidence limit, we are interested in testing the two-sided hypothesis H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$, where θ is the log odds ratio, i.e., $\theta=\log(\pi_1/(1-\pi_1))/(\pi_2/(1-\pi_2))$. As the power of any test φ in the case of a family of exponential distributions must be continuous, any unbiased test must be α-similar, i.e. $E_\pi(\varphi) = \alpha$ for any $\pi=(\pi_1, \pi_2)$ on the boundary $\Theta_B=\{(\pi_1, \pi_2)$ with $\pi_2=\pi_1/(\pi_1+(1-\pi_1)e^{\theta_0})\}$. Since T is a complete-sufficient statistic for $\pi \in \Theta_B$, any α-similar test must, in its turn, have Neyman-structure, i.e. must be conditional of size α, as $E_\pi\varphi(U,T) = E_\pi(E\,\varphi(U,T) \mid T) = \alpha$ implies $E(\varphi(U,T)|T)=\alpha$ ($\forall \pi \in \Theta_B$). Using the Neyman-Pearson Fundamental Lemma, conditionally on T, the Fisher exact test φ* rejecting for extreme values of U given T, i.e.

$$\phi^*(u,t;\theta_0) = \begin{cases} 1 & if\ u < c_1(t;\theta_0)\ or\ u > c_2(t;\theta_0) \\ \gamma_1(t;\theta_0) & if\ u = c_1(t;\theta_0) \\ \gamma_2(t;\theta_0) & if\ u = c_2(t;\theta_0) \\ 0 & if\ c_1(t;\theta_0) < u < c_2(t;\theta_0) \end{cases} \quad (1)$$

is, uniformly most powerful among the unbiased tests. But, with $\pi \in [0, 1]^2$ and $\theta=\log\{\pi_1/(1-\pi_1)\}/\{\pi_2/(1-\pi_2)\}$ denoting the log odds ratio, we have

$$E_\pi\varphi^*(U,T) = E_\pi(E_\theta\varphi^*(U,T) \mid T) \geq E_\pi(E_\theta\varphi(U,T) \mid T) = E_\pi\varphi(U,T),$$

i.e., the Fisher exact is also unconditionally more powerful than any other unbiased test. Please note that the distribution of U given T is the non-central hypergeometric with θ as canonical parameter, i.e.

$$P_\theta(U = u|T = t) = c(\theta)e^{\theta u}\ . \quad (2)$$

where

$$c(\theta) = \frac{\binom{m}{u}\binom{n}{t-u}/\binom{m+n}{t}}{\sum_{i=0}^{m\wedge t}\binom{m}{i}\binom{n}{t-i}/\binom{m+n}{t}e^{\theta i}}$$

Note that

$$\frac{d}{d\theta}c(\theta) = -\,c(\theta)E_\theta(U) \quad (3)$$

The randomizations at the critical values and the critical values themselves need to be determined such

$(a)\ E_{\theta_0}(\phi^*(U,T)|\,T) = \alpha$, i.e conditional of size α, and

$(b)\ \frac{d}{d\theta}\,E_{\theta_0}(\phi^*(U,T)|\,T) = 0$, for the test to be unbiased, i.e. $E_\theta(\phi^*(U,T)|\,T) \geq \alpha$ for $\theta \neq \theta_0$

By interchanging integration and differentiation, equation (b), using (3), leads to

$$(c)\ E_{\theta_0}(U\,\phi^*(U,T)\big|T) \;-\; \alpha\,E_{\theta_0}(U|T) = 0 \quad (4)$$

which, in its turn, leads to

(1) $\sum_{u<c_1 \text{ and } u>c_2} p_{\theta_0}(U=u|T=t) + \gamma_1 p_{\theta_0}(U=c_1|T=t) + \gamma_2 p_{\theta_0}(U=c_2|T=t) = \alpha$, and

(2) $$\sum_{u<c_1 \text{ and } u>c_2} u\, p_{\theta_0}(U=u|T=t) + \gamma_1 c_1 p_{\theta_0}(U=c_1|T=t) + \gamma_2 c_2 p_{\theta_0}(U=c_2|T=t) = \alpha E_{\theta_0}(U|T)$$

This yields as a solution for $\gamma = (\gamma_1, \gamma_2)$ given (c_1, c_2)

$$\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = \frac{1}{c_2 - c_1}\begin{pmatrix}1/p_{\theta_0}(U=c_1|T=t) & 0\\ 0 & 1/p_{\theta_0}(U=c_2|T=t)\end{pmatrix}\begin{pmatrix}c_2 & -1\\ -c_1 & 1\end{pmatrix}\begin{pmatrix}\alpha - \sum_{u<c_1 \text{ and } u>c_2} p_{\theta_0}(U=u|T=t)\\ \alpha E_{\theta_0}(U|T) - \sum_{u<c_1 \text{ and } u>c_2} u\, p_{\theta_0}(U=u|T=t)\end{pmatrix}$$

For t=0, $\gamma_1 = \gamma_2 = \alpha/2$, (and $c_1 = c_2 = 0$), and for $t=1$ straightforward calculation lead to $\gamma_1 = \gamma_2 = \alpha$ and $c_1 = 0$ and $c_2 = 1$. Similarly, for $t=n+m$ we have $\gamma_1 = \gamma_2 = \alpha/2$ (and $c_1 = c_2 = m$), and for $t=n+m-1$ we have $\gamma_1 = \gamma_2 = \alpha$ and $c_1 = m-1$ and $c_2 = m$. A proof for the existence of an UMPU test of the form (a) has been provided by [14].

That the test is *unique* can be seen from equation (a) and (c) as follows. A different test would need to have different c_1 and c_2 as (γ_1, γ_2) are uniquely determined by c_1 and c_2. Increasing the acceptance region on both ends would lead to a size < α, while decreasing it on both ends would lead to a size > α. So, the only option remaining is to shift the acceptance region to the left or to the right. This, however, would have to be done by shifting equal probability mass from one side to the other to maintain the size α. But this would lead to a decrease or increase $E_{\theta_0}(U\,\phi(U,T))$ so that equation (*c*) is no longer valid.

To numerically identify c_i and γ_i (i=1,2) we search for a combination of c_1 and c_2 for which $\gamma = (\gamma_1, \gamma_2)$ is in [0,1]x[0,1], starting at the α/2 and 1- α/2 percentiles of the non-central hypergeometric distribution and then "circling" around that in ever larger getting squares until the solution is identified. An example of these critical values and associated randomization probabilities is given in Table 1.

T	c_1	γ_1	c_2	γ_2
0	0	0.025	0	0.025
1	0	0.05	1	0.05
2	0	0.15	2	0.09
3	0	0.6	3	0.18
4	1	0.178	4	0.349
5	2	0.005	4	0.005
6	2	0.349	5	0.178
7	3	0.18	6	0.6
8	4	0.09	6	0.15
9	5	0.05	6	0.05
10	6	0.025	6	0.025

Note: the table in [2] was incorrect.

Table 1: Critical values and associated randomization probabilities for the two-sided hypothesis testing problem with n=4 and m=6

The Modified Two-Sided Fisher Exact Test

The modified Fisher exact test differs from the randomized Fisher exact test in that at the critical values the hypothesis is only rejected if the probabilities $\gamma_i(t;\theta_0)$ (i=1,2) exceed a certain threshold γ_0, i.e.

$$\varphi_{mFE}(u,t;\theta_0) = \begin{cases} 1 & if\ u < c_1(t;\ \theta_0)\ or\ u > c_2(t;\ \theta_0) \\ \Omega(\gamma_1(t;\theta_0) > \gamma_0) & if\ u = c_1(t;\ \theta_0) \\ \Omega(\gamma_2(t;\theta_0) > \gamma_0) & if\ u = c_2(t;\ \theta_0) \\ 0 & if\ \ c_1(t;\ \theta_0) < u < c_2(t;\ \theta_0) \end{cases}$$

where $\boldsymbol{\Omega}(.)$ is the indicator-function. The value of γ_0 is chosen such that the size of the test is as large as possible but less than or equal to α. The size of this test for testing H_0: $\theta = \theta_0$ is given by

$$E_{\pi_1}\left(\varphi_{mFE}(u,t;\theta_0)\right) = \sum_{i,j}^{m,n} \varphi_{mFE}(i,i+j;\ \theta_0)\binom{m}{i}\binom{n}{j}\frac{\pi_1^{i+j}(1-\pi_1)^{n+m-i-j}e^{(n-j)\theta_0}}{(\pi_1+(1-\pi_1)e^{\theta_0})^n}$$

which is a function of the nuisance parameter π_1. To find the maximum size as a function π_1, 3 numerical approaches can be adopted

(1) Using a selected set of different starting values, apply SAS IML procedure NLPTR [15] for each and then take the maximum of these local extrema. The NLPTR procedure is a trust region method to identify the optimum of a function. This procedure is slow.

(2) Using a selected number of points (e.g. 0.1 to 0.9 by 0.1) identify which has the maximum size and use this as the starting value for the NLPTR subroutine.

(3) As 2, but don't use the NLPTR subroutine. Instead evaluate the size around the point with the maximum size in ever smaller intervals and update the maximum point at each round. We refer to this simple approach as the "zoom-in" approach. In practise, it is generally the quickest approach. The number of evaluations can be chosen, the default we use is 19 (starting at 0.05, 0.1, 0.15, …, 0.95).

The best way to be sure we have not identified a local optimum, is to plot the size of the test, particularly at the lower and upper bound of the 95% confidence interval (see Section - Size Comparisons to Other Contemporary Tests).

Having obtained the size of the test, i.e. $\max_{\pi_1} E_{\pi_1}\left(\varphi_{mFE}(u,t;\theta_0)\right)$ we need to identify that γ_0 that yields the largest size smaller than α. This is accomplished by sorting the $2^*(m+n+1)$ critical values $c_i(t;\ \theta_0)$ (i=1,2 and t=0, …, $n+m$) by their randomization probabilities $\gamma_i(t;\theta_0)$ and using a bisection method over these critical values to identify the optimum γ_0. Note that γ_0 is not unique but is an left-closed, right open interval between two ordered critical values.

Size Comparisons to Other Contemporary Tests

We shall compare the size of the modified Fisher exact test as a function of the nuisance parameter π_1 (where $\pi_2 = \pi_1/(\pi_1+(1-\pi_1)\theta_0)$ to the following tests

1. The conservative Fisher exact test, rejecting only if the number of responses exceed the critical value, i.e.

$$\varphi(u,t;\theta_0) = \begin{cases} 1 & if\ u < c_1(t;\ \theta_0)\ or\ u > c_2(t;\ \theta_0) \\ 0 & otherwise \end{cases} \qquad (5)$$

Confidence limits will be test-based (see Section *p*-Values and Confidence Limits)

2. The asymptotic test by Woolf [13], i.e. reject based on the Wald statistic with the OR estimated by $\widehat{OR} = \frac{u/(m-u)}{v/(n-v)}$ and the SE (log $(\widehat{OR})) = \sqrt{1/u + 1/(m-u) + 1/v + 1/(n-v)}$ and reject if $|\log(\widehat{OR})/\mathrm{SE}(\log(\widehat{OR}))| \geq z_{1-\alpha/2}$. Rather than adding 0.5 for all 4 terms, *u, v, m-u* and *n-v*, as suggested by Hadane [16], we only replace zero values of *u, v, m-u* or *n-v* by 0.5. Confidence limits are then given by

$$e^{\log(\widehat{OR}) \pm z_{1-\alpha/2}\, SE(\log(\widehat{OR}))}$$

3. The test based on the *p*-value constructed as the sum of all conditional probabilities (*1-b*) of outcomes less likely than or equally likely as the one observed, that is reject if

$$p\text{-value} = \sum_{i \in C} P_\theta(U = i | T = t) \leq \alpha,$$

where, with u being the number of responses observed,

$$C = \{i\,;\, P_\theta(U = i|T = t) \leq P_\theta(U = u|T = t)\} \qquad (6)$$

which is the exact *p*-value SAS Proc FREQ is providing. Confidence limits are, however, constructed using equal-tail one sided *p*-values, i.e., the confidence limit for the upper and lower 1- α confidence limits are the solution of, respectively,

$$\sum_{i=0}^{u} P_\theta(U = i|T = t) = \alpha/2 \quad \text{and} \quad \sum_{i=u}^{m} P_\theta(U = i|T = t) = \alpha/2$$

Some examples of the size as a function of the nuisance parameters and for $\exp(\theta_0)$=1 and $\exp(\theta_0)$=2 are given in Figures 1,2 and 3. Of note is that (1) the modified Fisher exact generally outperforms the other tests, (2) the test based on the exact *p*-value provided by SAS Proc FREQ is fairly conservative, and (3) the Woolf's asymptotic test for larger samples is most similar to the modified Fisher exact test, but does not strictly control the Type 1 error as witnessed by Figure 3a.

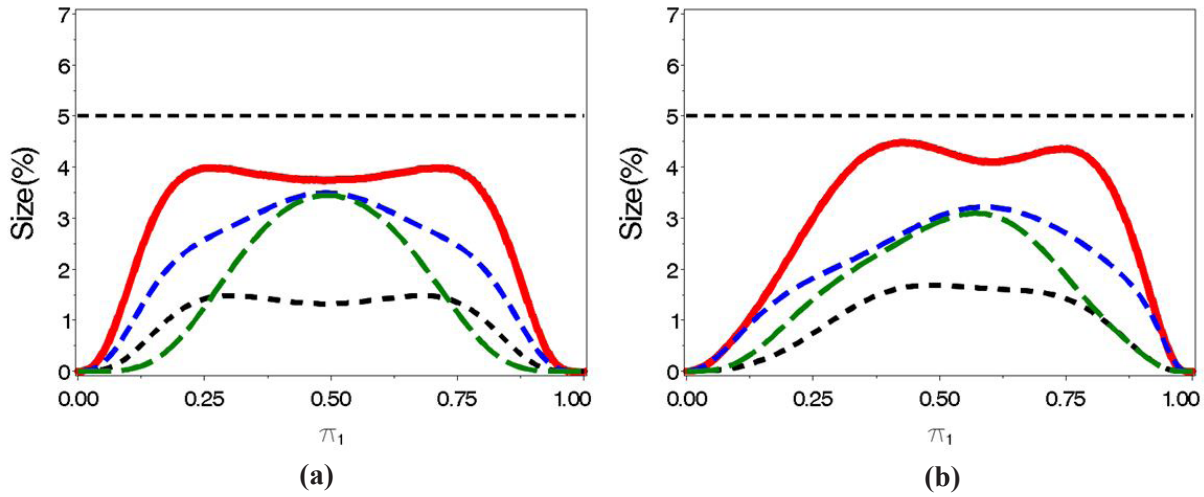

Figure 1: The size of the modified FE test (solid red line), the asymptotic test (green, large dash), the SAS Proc FREQ based "exact" *p*-value based test (blue, medium dash) and the conservative FE test (black, small dash), for *m*=12, and *n*=11 for testing **(a)** H_0: $e^{\theta_0} = 1$ (OR = 1) and **(b)** H_0: $e^{\theta_0} = 2$ (OR = 2)

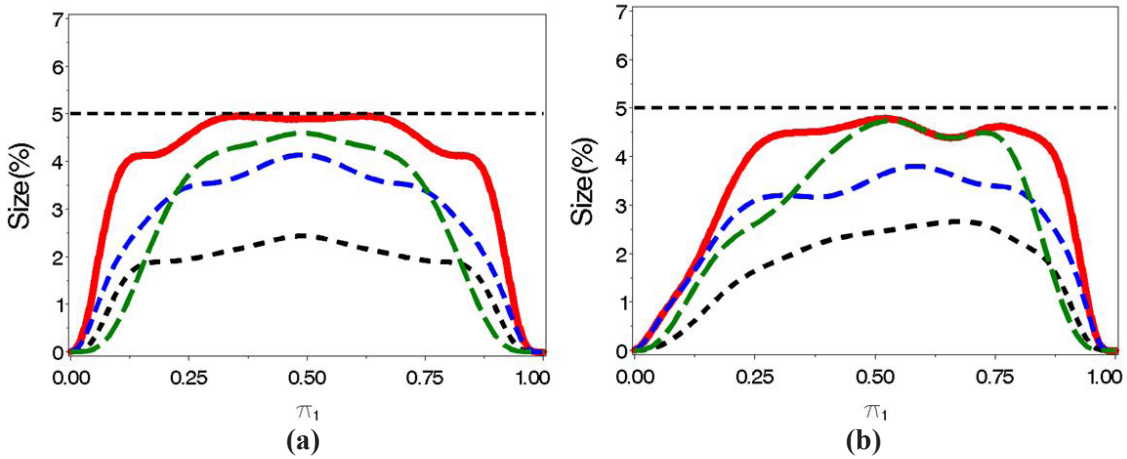

Figure 2: The size of the modified FE test (solid red line), the asymptotic test (green, large dash), the SAS Proc FREQ based "exact" *p*-value based test (blue, medium dash) and the conservative FE test (black, small dash), for *m*=27, and *n*=21 for (**a**) H_0: e^{θ_0}=1 (OR=1) and (**b**) H_0: e^{θ_0}=2 (OR=2)

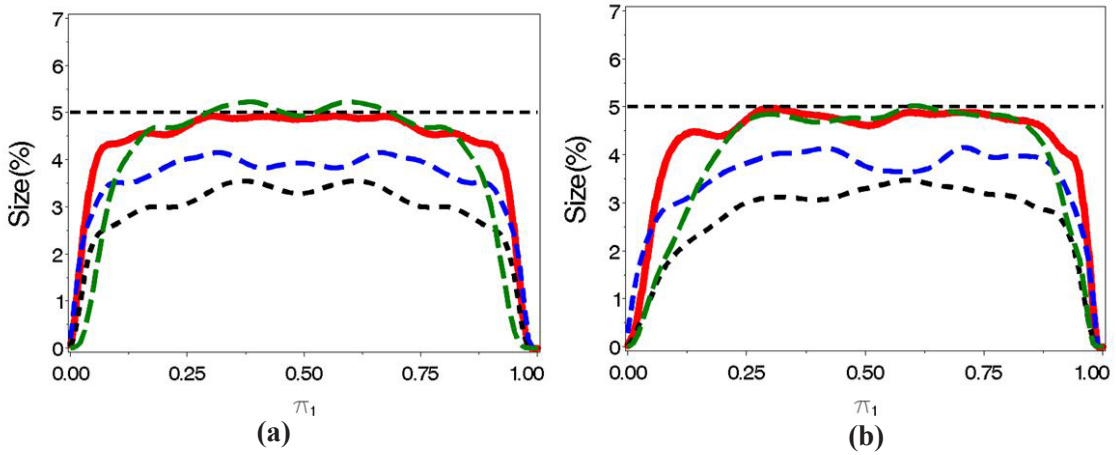

Figure 3: The size of the modified FE test (solid red line), the asymptotic test (green, large dash), the SAS Proc FREQ based "exact" *p*-value based test (blue, medium dash) and the conservative FE test (black, small dash), for *m*=65, and *n*=71 for testing (**a**) H_0: e^{θ_0}=1 (OR=1) and (**b**) H_0: e^{θ_0}=2 (OR=2)

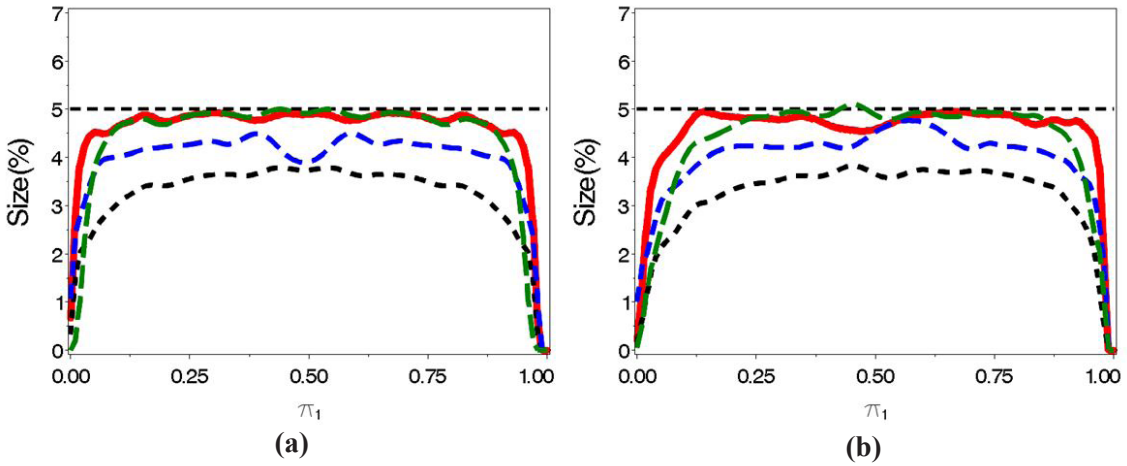

Figure 4: The size of the modified FE test (solid red line), the asymptotic test (green, large dash), the SAS Proc FREQ based "exact" p-value based test (blue, medium dash) and the conservative FE test (black, small dash), for m=128, and n=142 for testing (**a**) H_0: e^{θ_0}=1 (OR=1) and (**b**) H_0: e^{θ_0}=2 (OR=2)

p-Values and Confidence Limits

A two-sided 1-α confidence interval for the modified Fisher exact test is constructed as all those θ_0 that are not rejected at the two-sided α level, i.e.

$$CI_{1-\alpha} = \{\theta_0;\ H: \theta = \theta_0 \text{ is not rejected using the two} - \text{sided } \varphi_{\mathbf{mFE}} \text{ test of size} - \alpha\}.$$

It is obtained by using the bisecting method starting with lower and upper limit of Woolf's asymptotic confidence limits as starting points for the bisecting variable θ_0, respectively, until the point of first rejection has been reached with a desired precision.

When evaluating the quality of a confidence interval, coverage can be considered, where ideally

$$P_\theta(\theta \in CI_{1-\alpha}) = 1 - \alpha \qquad \forall\theta$$

However, with finite discrete distributions, this is generally not feasible. There are only *(m+1)(n+1)* possible confidence intervals $C_{i,j}$ (for *U=i*, and *V=j*) and to require that the expectation

$$\sum_{i,j}^{m,n} \Omega(\theta \in C_{i,j})\, P_{\pi_1,\pi_2}(U = i, V = j)$$

is always exactly (1-α) is unrealistic. Perhaps for some values of θ and underlying values of the nuisance parameter, but surely not for all. What is possible though is to plot the coverage as a function of $(\pi_1, \pi_2) \in [0,1]x[0,1]$, to see how the confidence interval generally performs compared to others from a coverage perspective. Or, alternatively, as a function of π_1, for a selected set of values of the log odds ratio θ. As this exercise is computer-time intense, as all possible confidence intervals must be constructed, and as it will be challenging to judge from such plots which of two competitive tests has the better coverage, we do not pursue this in this paper.

The *p*-value is constructed as the smallest level of α for which the hypothesis H_0: $\theta = \theta_0$ is rejected, i.e.

Test	OR Estimate	p-value	95% CI of the OR
Example 1: 5/12 versus 7/11			
- Modified FE Test	**0.408**	**0.321**	**0.716 – 2.210**
- Woolf's Asymptotic Test	0.408	0.296	0.760 – 2.193
- Exact SAS Proc FREQ	0.408	0.414	0.550 – 2.871
Example 2: 13/41 versus 6/47			
- Modified FE Test	**3.173**	**0.034**	**1.082-10.25**
- Woolf's Asymptotic Test	3.173	0.036	1.077- 9.34
- Exact SAS Proc FREQ	3.173	0.039	0.969-11.31
Example 3: 37/65 versus 28/71			
- Modified FE Test	**2.029**	**0.0439**	**1.023 – 4.083**
- Woolf's Asymptotic Test	2.029	0.0425	1.024 – 4.021
- Exact SAS Proc FREQ	2.029	0.0584	0.970 – 4.258
Example 4: 72/128 versus 58/142			
- Modified FE Test	**1.804**	**0.0175**	**1.108 – 2.936**
- Woolf's Asymptotic Test	1.804	0.0167	1.113 – 2.925
- Exact SAS Proc FREQ	1.804	0.0203	1.082 – 3.018

Table 2: Comparison of two-sided *p*-values and 95% confidence intervals by examples

$$p-\text{value} = inf\{\alpha;\ \ H{:}\,\theta = \theta_0 \text{ is rejected using the two} - \text{sided } \varphi_{\text{mFE}} \text{ test of size } \alpha\}$$

It is obtained via the bisecting method based on bisecting α.

Some example of confidence intervals and *p*-values are provided in Table 2. That of the classical conservative Fisher exact test (5) is not shown as it is not used in practice.

Of note is the disagreement between the *p*-value and the 95% confidence interval in Example 2 of the SAS Proc FREQ based exact method.

Power Comparison

The (envelope) power of the modified Fisher exact was discussed in [2]. However, presently smaller experiments involving response rates of two groups are mostly based on the exact *p*-value as obtained via SAS Proc FREQ. It is worthwhile to see the impact on potential cost-savings for such experiments by adopting the less conservative, Type 1 error rate controlling, and modified Fisher exact test instead. We do this merely by example. The SAS/IML macro, however, does allow you to obtain the power for given

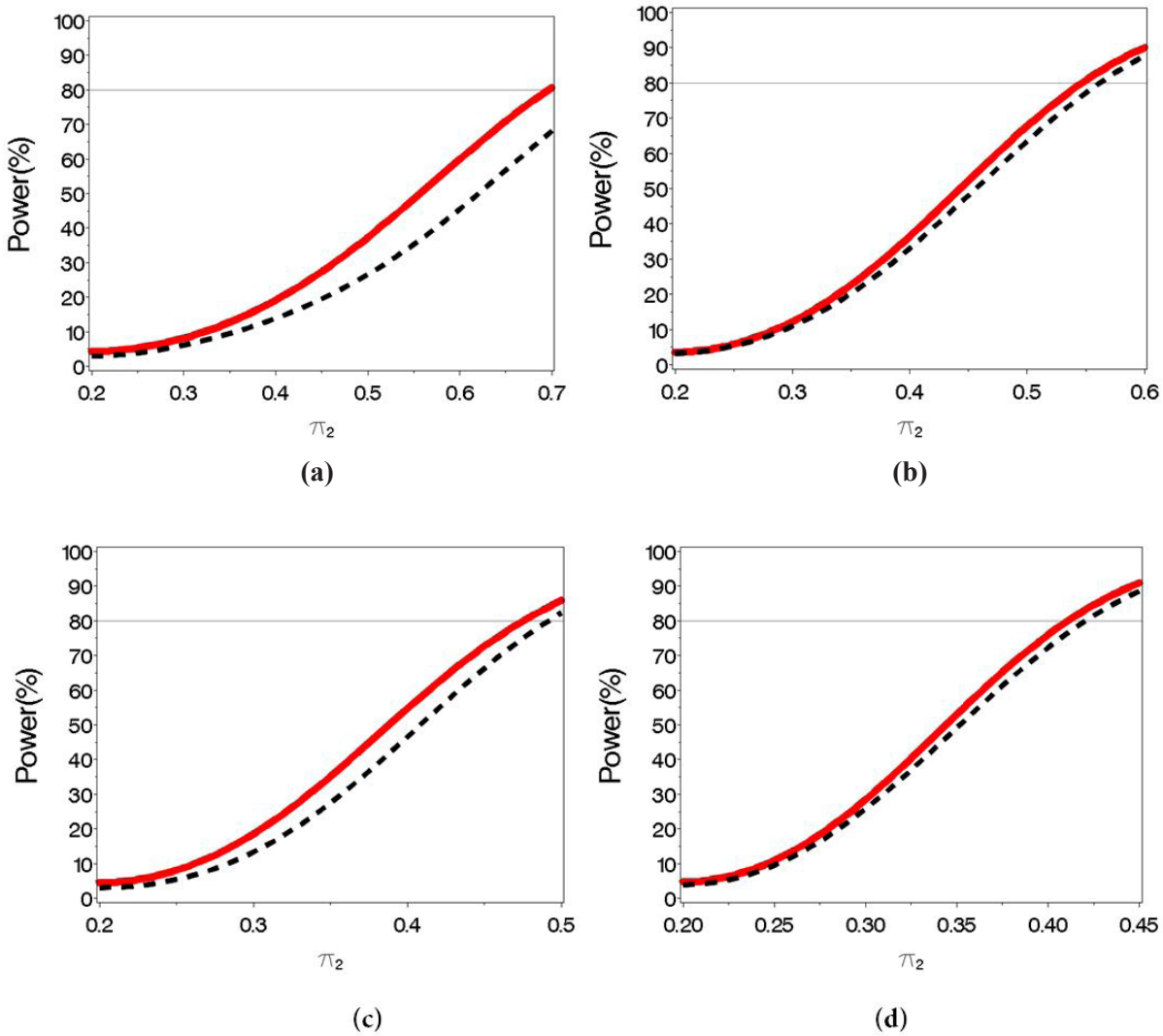

Figure 5: The power of the modified FE test (solid red line) compared to the "exact" *p*-value as provided by SAS Proc FREQ (dashed black line) for **(a)** n=m=15; **(b)** n=m=30; **(c)** n=m=45; and **(d)** n=m=75 and π_1=0.2

sample sizes n and m and response rates π_1 and π_2. It does that not only for testing equality but also for testing superiority. The latter, obtained by additionally requiring a positive observed effect-size, only has a negligibly smaller power for relevant effect sizes as compared to testing equality. The power for testing equality of the modified Fisher exact test and the exact test performed by SAS Proc FREQ exact test are displayed for various scenarios in Figure 5.

From these examples, particularly in very small experiments, the proposed modified Fisher exact test is more powerful than the contemporary exact test as implemented in SAS Proc FREQ. The difference in power is, however, not a strictly declining function of the sample size, but may fluctuate by configuration. Please note that the difference, for large sample sizes, may seem small but a few percentages difference, implies a fair increase in sample size. For example, in Figure 5d, the power at π_2=0.411 is 80.08% for the proposed modified Fisher exact test and 76.56% for the SAS Proc FREQ based "exact" p-value based test. To lift the latter to 80% requires an additional 10 subjects. This is a worthwhile cost-reduction.

Discussion and Conclusion

There are many ways to extract two-sided p-values from the Fisher exact test as discussed by Agresti [17]. Another popular non-conservative one, is the mid p-value, constructed as

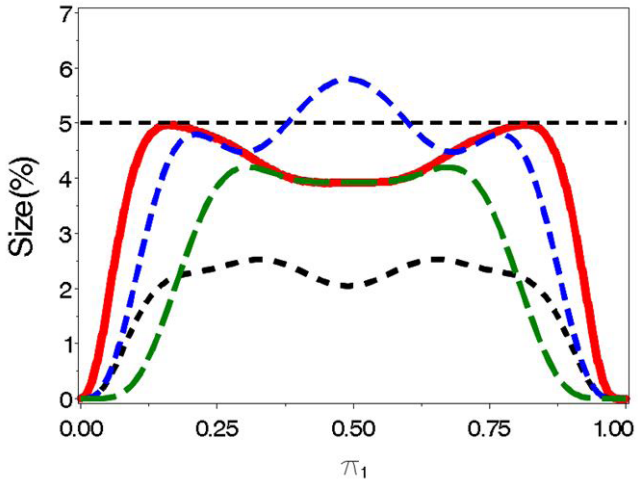

Figure 6: The size of the modified FE test (solid red line), the asymptotic test (green, large dash), the SAS Proc FREQ based "exact" mid p-value based test (blue, medium dash) and the conservative FE test (black, small dash), for m=20, and n=20 for testing H_0: e^{θ_0}=1 (OR=1)

$$p\text{-value} = \sum_{i \in C} P_\theta(U = i|T = t) + 0.5\, P_\theta(U = u|T = t)$$

where C is as in (6) but with a strict inequality. However, as earlier indicated, the Type 1 error rate is then no longer controlled as Figure 6 indicates.

The same applies for the Pearson Chi-square test and other tests. The proposed modification takes full control of the Type I error rate which many regulatory agencies so often mandate. It does that, however, in a non-conservative way, which gives a head-start when it comes to power, that seems to be sustained throughout the alternative space. The test-based approach by which associated p-values and confidence limits are constructed make them agree by definition.

The question is, can it be improved even further? One possible suggestion may be to have separate thresholds $\gamma_{0,i}$ (i=1,2) for $c_1(t)$ and $c_2(t)$, respectively, and then further maximize the area under the size. The direction for further research should, in our opinion, at least stay as close as possible to the randomized Fisher exact test, as this is the uniformly most powerful test among unbiased

tests. Possible criticisms to the proposed test are that the size is controlled not by a closed formula, but only by using numerical-analytic methods and that it requires a fair amount of computing time for larger sample sizes (CPU-times of 60 sec, 5 minutes, 30 minutes and 60 minutes , for N=50, 100, 200, and 300, respectively). The possibility to plot the size as a function of the nuisance parameter, largely takes away the first criticism. As computing power increases, the computational burden of the proposed method will diminish. Additionally, the potential to optimize the method through parallel programming could substantially improve the performance of the method. Moreover, for large sample sizes (and frankly also for small sample sizes) Woolf's asymptotic approach is a good alternative, particularly when strict control of the Type I error rate is not required, as for example when used in flagging significant adverse events for signal detection.

A SAS/IML macro is provided in the appendix. Realizing that not all have the possibility to utilize SAS/IML, an equivalent R-package will be developed and available shortly. Please contact Dr Kyle Raymond (E-mail: kyleraymond08@gmail.com) for further information in this respect.

Appendix

References

1. Fisher RA (1941) Statistical Methods for Research Worker, Oliver & Boyd: Edinburgh, UK.

2. Van der Meulen EA (2008) A Nonrandomized, Nonconservative Version of the Fisher Exact Test" Communications in Statistics. Theory and Methods 37: 699-708.

3. Tocher KD (1950) Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates. Biometrika 37: 130-44.

4. Lehmann EL (1997) Testing Statistical Hypotheses, Springer Verlag: New York, USA.

5. Suissa S, Shuster JJ (1984) Are Uniformly Most Powerful Unbiased Tests Really Best? The American Statistician 38: 204-6.

6. Boschloo RD (1970) Raised conditional level of significance for the 2x2 table when testing the equality of two probabilities. Statistica Neerlandica 24: 1-35.

7. McDonald LL, Davis BM, Milliken GA (1977) A Nonrandomized Unconditional Test for Comparing Proportions in 2x2 Contingency Tables. Technometrics 19: 145-50.

8. Lancaster HO (1961) Significance Tests in Discrete distributions. J Am Statist Ass 56: 223-34.

9. Haber MA (1986) A Modified Exact Test for 2x2 Contingency Tables. Biom J 4: 455-63.

10. Upton GJP (1992) Fisher's Exact Test. JR Statist Soc A 155: 395-402.

11. Lydersen S, Laake P (2003) Power comparison of two-sided exact tests for association in 2x2 contingency tables using standard, mid p, and randomized test versions. Statist Med 22: 3859-71.

12. Pearson ES (1949) The choice of statistical tests illustrated on their interpretation of data classed in a 2x2 table. Biometrika 34: 139-67.

13. Woolf B (1955) On Estimating the Relationship between Blood Group and Disease. Annals Human Genet 19: 251-3.

14. University of Washington, College of Arts & Sciences (2020) Statistics, Lecture note STAT 581: Advanced Theory of Statistical Inference, USA.

15. SAS (2004) SAS/IML User's Guide, USA.

16. Haldane JBS (1956) The Estimation and Significance of the Logarithm of a Ratio of Frequencies. Annals Human Genet 20: 309-11.

17. Agresti A (1992) A survey of Exact Inference for Contingency Tables. Statistical Sci 7: 131-77.

Submit your next manuscript to Annex Publishers and benefit from:

- Easy online submission process
- Rapid peer review process
- Online article availability soon after acceptance for Publication
- Open access: articles available free online
- More accessibility of the articles to the readers/researchers within the field
- Better discount on subsequent article submission

Submit your manuscript at
http://www.annexpublishers.com/paper-submission.php