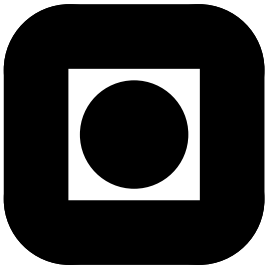
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Statistical hypothesis testing for categorical data using enumeration in the presence of nuisance parameters

by

Clara-Cecilie Günther, Øyvind Bakke, Håvard Rue and
Mette Langaas

PREPRINT STATISTICS NO. 4/2009
DEPARTMENT OF MATHEMATICAL SCIENCES



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL
http://www.math.ntnu.no/preprint/statistics/2009/S4-2009.pdf
Mette Langaas has homepage: http://www.math.ntnu.no/∼mettela
E-mail: Mette.Langaas@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

STATISTICAL HYPOTHESIS TESTING FOR CATEGORICAL DATA USING ENUMERATION IN THE PRESENCE OF NUISANCE PARAMETERS

CLARA-CECILIE GÜNTHER, ØYVIND BAKKE, HÅVARD RUE AND METTE LANGAAS
Department of Mathematical Sciences.
The Norwegian University of Science and Technology,
NO-7491 Trondheim, Norway.

MARCH 2009

SUMMARY

The existing asymptotic tests for comparing positive predictive values of two diagnostic tests do not preserve the test size when the sample is small. As an exact approach we suggest using enumeration for small sample spaces, i.e. to utilize the exact distribution of the test statistic by adding probabilities of each outcome. In the problem of comparing positive predictive values, there are nuisance parameters present which must be handled. We discuss different solutions, e.g. estimation, maximization, integration and combinations thereof. The methods presented in this report are general and can be applied to different discrete finite distributions. Further insight into the mechanisms behind the different approaches are given and the performance of various test statistics and p-values are compared systematically with respect to test size and power, both in the setting of positive predictive values and in an example from literature comparing independent binomial proportions. We find in general that a combination of estimation and maximization yields the highest test size and power among the valid p-values, and when comparing the positive predictive values, the test statistics involving maximum likelihood estimates under the null hypothesis perform the best in terms of test size and power.

1 INTRODUCTION

In many hypothesis testing problems, tests statistics with a known asymptotic distribution are available. When the sample size is small, however, the asymptotic distribution may approximate the exact distribution poorly and the exact distribution of the test statistics can be challenging or impossible to derive. For discrete models, one solution is to use enumeration, i.e. to find p-values by adding probabilities under the null hypothesis of all possible outcomes having a more extreme value of the test statistic than the observed outcome. If there are nuisance parameters in the model, this is however not straight forward, the unknown parameters must be handled appropriately.

We consider different approaches, in particular estimation, maximization and integration. Our main focus will be on the problem of comparing positive predictive values from two diagnostic tests where a multinomial distribution is assumed, but the methods are general and can be applied to other null hypotheses for other finite discrete distributions.

We start by defining important properties for p-values and different ways to handle the problem of nuisance parameters in Section 2. A trinomial situation is used as an example to explain how to calculate the various p-values. As a stepping stone to our main problem, comparing positive predictive values, in Section 3 we go through a fictitious example discussed and analyzed by Berger and Boos (1994) and by Lloyd (2008) that concerns testing independence in a 2×2 contingency table. We suggest alternative test statistics and compare their performance in terms of test size and power to the test statistics used by Lloyd (2008). In Section 4 we present the problem of comparing positive predictive values for two diagnostic tests, and evaluate a variety of test statistics and p-values for this problem. Some computational details are given in Section 5, we discuss further aspects of the presented problems in Section 6 and summarize the conclusions in Section 7.

2 THEORY

Before applying the methods, the general framework should be set. We present the necessary notation, definitions and properties of p-values and explain different approaches on how to calculate p-values by enumeration in the presence of nuisance parameters.

2.1 NULL HYPOTHESIS

In the general outline we assume that the random variables $Y_1, \ldots, Y_n$ are multinomially distributed with parameters $\boldsymbol{p} = (p_1, \ldots, p_n)$ and N, but other discrete distributions are possible (see e.g. Section 3). Let $\boldsymbol{Y}$ denote the vector of the random variables, i.e. $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, and let $\mathcal{Y}$ be the sample space or reference set of $\boldsymbol{Y}$.

Our null hypothesis is that a function f of some or all the parameters $p_1, \ldots, p_n$ equals 0, i.e.

$$H_0 : \ f(\boldsymbol{p}) = 0. \tag{1}$$

The alternative hypothesis is

$$H_1 : \ f(\boldsymbol{p}) \neq 0.$$

Let $\mathcal{P}$ be the parameter space for $\boldsymbol{p}$ and $\mathcal{P}_0$ the subspace of $\mathcal{P}$ for which the null hypothesis (1) is satisfied, i.e $\mathcal{P}_0 = \{\boldsymbol{p} : f(\boldsymbol{p}) = 0\}$. For illustrative purposes, an example from the trinomial distribution will be studied throughout this section.

Trinomial example As an illustrative example we will use the trinomial model where $\boldsymbol{Y} = (Y_1, Y_2, Y_3)$ are multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, p_3)$ and N, or alternatively $\boldsymbol{Y} = (Y_1, Y_2, N - Y_1 - Y_2)$, are multinomially distributed with parameters $\boldsymbol{p} = (p_1, p_2, 1 - p_1 - p_2)$. The joint probability function of $\boldsymbol{Y}$ is

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} p_1^{y_1} p_2^{y_2} (1 - p_1 - p_2)^{N - y_1 - y_2}.$$

We consider the null hypothesis,

$$H_0 : \ f(\boldsymbol{p}) = p_1 - p_2 = 0, \tag{2}$$

that is $\mathcal{P}_0 = \{(\phi, \phi, 1-2\phi) : 0 \leq \phi \leq 1/2\}$. So $p_1 = p_2 = \phi$ under the null hypothesis, which can be considered an unknown nuisance parameter. The probability function of $\boldsymbol{Y}$ simplifies to

$$\Pr(Y_1 = y_1, Y_2 = y_2) = \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} \phi^{y_1 + y_2} (1 - 2\phi)^{N - y_1 - y_2} \tag{3}$$

under the null hypothesis. □

2.2 PROPERTIES OF p-VALUES

When testing whether a null hypothesis is true, one usually calculates a p-value and if this p-value is less than or equal to some chosen significance level α the null hypothesis is rejected.

A p-value may initially be defined as the probability of what has been observed or something more extreme, given that the null hypothesis is true. A p-value can also be considered a test statistic in its own right. We let $P(\boldsymbol{Y})$ denote our p-value statistic which is a function of the random variables $\boldsymbol{Y}$. For continuous models without nuisance parameters and for simple null hypotheses, i.e. when the parameter space under H_0 consists of only one point, the p-values are uniformly distributed under the null hypothesis and the test size of a test that rejects H_0 when $P(\boldsymbol{Y}) \leq \alpha$ is exactly equal to α, Bickel and Doksum (2001). Our sample space is discrete which means that not all p-values can possibly be obtained. Instead, is is usually demanded that the p-value is *valid*, i.e. the probability of rejecting the null hypothesis when it is true is less than or equal to the significance level α,

$$\Pr(P(\boldsymbol{Y}) \leq \alpha; \boldsymbol{p}) \leq \alpha$$

for all $\boldsymbol{p}$ in $\mathcal{P}_0$ and all α, $0 \leq \alpha \leq 1$, Casella and Berger (2002). The valid p-values yield a valid test for any chosen significance level, although they are often conservative. If a p-value satisfy

$$\sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(P(\boldsymbol{Y}) \leq P(\boldsymbol{y}); \boldsymbol{p}) = P(\boldsymbol{y}),$$

for all $\boldsymbol{y}$ in $\mathcal{Y}$, then Lloyd (2008) call it *exact*.

In general, p-values are found by means of a test statistic $T(\boldsymbol{Y})$ having the property that for all $\boldsymbol{y}$ in $\mathcal{Y}$ and for all $\boldsymbol{p}$ in $\mathcal{P}_0$, $\Pr(P(\boldsymbol{Y}) \leq P(\boldsymbol{y}); \boldsymbol{p}) = \Pr(T(\boldsymbol{Y} \geq T(\boldsymbol{y}); \boldsymbol{p})$, assuming without loss of generality that the null hypothesis is rejected for larges values of $T(\boldsymbol{y})$. We define the tail set of an outcome $\boldsymbol{y}_{\text{obs}}$ to be the set of all $\boldsymbol{y}$ for which $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$, i.e. the critical region for a significance level given $T(\boldsymbol{y}_{\text{obs}})$ as a critical value. For an observed outcome $\boldsymbol{y}_{\text{obs}}$, the reference set $\mathcal{Y}$ can be partitioned into the tail set $R(\boldsymbol{y}_{\text{obs}})$ of the observed outcome and the complement of the tail set $R^C(\boldsymbol{y}_{\text{obs}})$, so that $\mathcal{Y} = R \cup R^C$ where $R(\boldsymbol{y}_{\text{obs}}) = \{\boldsymbol{y} : T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})\}$ and $R^C(\boldsymbol{y}_{\text{obs}}) = \{\boldsymbol{y} : T(\boldsymbol{y}) < T(\boldsymbol{y}_{\text{obs}})\}$.

Trinomial example We set $N = 3$, and then the reference set is $\mathcal{Y} = \{(0,0,3), (0,1,2), (0,2,1), (0,3,0), (1,0,2), (1,1,1), (1,2,0), (2,0,1), (2,1,0), (3,0,0)\}$. One possible test statistic is

$$T(\boldsymbol{Y}) = |Y_1/N - Y_2/N|. \tag{4}$$

Table 1 shows the calculated test statistic for all the outcomes in the reference set. For example, $T(0,2,1) = 2/3$ and $R(0,2,1) = \{(0,2,1), (0,3,0), (2,0,1), (3,0,0)\}$. □

The test statistic $T(\boldsymbol{Y})$ used to define the tail set can be an ordinary test statistic like the likelihood ratio test statistic, a p-value originating from another test statistic, or even the multinomial probabilities of the outcomes themselves. If the tail sets are defined by the probabilities of the outcomes, they will

Outcome $\boldsymbol{y}$	y_1/N	y_2/N	$T(\boldsymbol{y})$
(0,0,3)	0	0	0
(0,1,2)	0	1/3	1/3
(0,2,1)	0	2/3	2/3
(0,3,0)	0	1	1
(1,0,2)	1/3	0	1/3
(1,1,1)	1/3	1/3	0
(1,2,0)	1/3	2/3	1/3
(2,0,1)	2/3	0	2/3
(2,1,0)	2/3	1/3	1/3
(3,0,0)	1	0	1

TABLE 1: The reference set in the trinomial example with associated test statistic, $T(\boldsymbol{y})$ given in (4).

depend on $\boldsymbol{p}$. This is not so if the tail sets are defined by either a p-value or some test statistic that does not depend on $\boldsymbol{p}$. A practical detail, when the multinomial probabilities or a p-value are used as the test statistic, actually the negative of the probabilities and the p-values will be applied since only the outcomes with probabilities or p-values smaller than or equal to the probability or p-value of the observed outcome will be in the tail set.

2.3 CALCULATING p-VALUES BY ENUMERATION

Let $\pi(\boldsymbol{y};\boldsymbol{p}) = \Pr(\boldsymbol{Y} = \boldsymbol{y};\boldsymbol{p})$ be the probability of an outcome $\boldsymbol{y}$. If $\pi(\boldsymbol{y};\boldsymbol{p})$ is known, the p-value for the observed outcome can be calculated using the following algorithm which is motivated by Fisher's exact test for 2×2 tables, Fisher (1935):

1. Generate all possible outcomes in the reference set $\mathcal{Y}$.
2. Calculate the probability of observing each outcome under the null hypothesis.
3. The p-value of an observed outcome is the sum of the probabilities of all outcomes that are in the tail set of the observed outcome.

Zelterman, Chan and Mielke (1995) tested mutual independence of all the three factors of a 2^3 contingency table using a multinomial distribution with eight parameters. Any outcome given N will then correspond to a specific table where the entries sum to N and the reference set $\mathcal{Y}$ will be all possible tables with grand total N. By conditioning on the set of one-way marginal totals, $\boldsymbol{M}$, the probability $\pi(\boldsymbol{y}|\boldsymbol{M})$ under H_0 can be derived. It does not depend on nuisance parameters, and therefore the second step in the algorithm is easily performed once the tables are generated.

With other null hypotheses it might be impossible to get rid of the nuisance parameters and conditioning only reduces the number of possible outcomes or the number of nuisance parameters. In this case, we must find a way to deal with the (remaining) nuisance parameters to be able to calculate the probability of each outcome. There are several ways to do this.

ESTIMATION The simplest approach to deal with nuisance parameters is to insert e.g. the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ under H_0 for $\boldsymbol{p}$. This is called the plug-in p-value by Bayarri and Berger (2000)

and the estimation (E) p-value by Lloyd (2008). For an observed outcome $\boldsymbol{y}_{\text{obs}}$ we insert $\tilde{\boldsymbol{p}}_{\text{obs}}$ for $\boldsymbol{p}$ and the p-value is given as

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y}; \tilde{\boldsymbol{p}}_{\text{obs}}).$$

This p-value, however, is not valid as we will see numerically in Section 4.2.

Trinomial example Under H_0, given the outcome $\boldsymbol{y}_{\text{obs}} = (y_{1,\text{obs}}, y_{2,\text{obs}}, y_{3,\text{obs}})$ the maximum likelihood estimate of ϕ is $\widetilde{\phi}_{\text{obs}} = \frac{y_{1,\text{obs}}+y_{2,\text{obs}}}{2N}$. If we insert this estimate in the multinomial probability function, we obtain the estimation p-value

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y}; \widetilde{\phi}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} \widetilde{\phi}_{\text{obs}}^{y_1+y_2} (1 - 2\widetilde{\phi}_{\text{obs}})^{N-y_1-y_2}.$$

The third column of Table 2 shows the estimation p-values for all outcomes in the reference set when $N = 3$. To explain how the p-values are calculated, we consider the outcome $\boldsymbol{y} = (0, 2, 1)$. The maximum likelihood estimate under H_0 is $\widetilde{\phi} = 1/3$. We then calculate the probability for each outcome from (3) with $\widetilde{\phi}$ inserted for ϕ. The tail set consists of the four outcomes $\boldsymbol{y}$ of Table 1 for which $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$, where $T(\boldsymbol{y})$ is given in (4), and the estimation p-value of $\boldsymbol{y}_{\text{obs}}$ is the sum 0.30, of the four probabilities. □

Outcome $\boldsymbol{y}$	$\widetilde{\phi}$	$P_{\text{E}}(\boldsymbol{y})$
(0,0,3)	0	1.00
(0,1,2)	1/6	0.56
(0,2,1)	1/3	0.30
(0,3,0)	1/2	0.25
(1,0,2)	1/6	0.56
(1,1,1)	1/3	1.00
(1,2,0)	1/2	1.00
(2,0,1)	1/3	0.30
(2,1,0)	1/2	1.00
(3,0,0)	1/2	0.25

TABLE 2: P-values for the trinomial example when substituting $\widetilde{\phi}$ for ϕ.

CONDITIONING ON A SUFFICIENT STATISTIC Another solution to the problem of nuisance parameters is to condition on a sufficient statistic X for $\boldsymbol{p}$, Casella and Berger (2002), then the probability of the observed outcome given H_0 and the sufficient statistic can be calculated and the p-value is given by

$$P_{\text{suff}}(\boldsymbol{y}_{\text{obs}}) = \Pr(T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}}) \mid X; \boldsymbol{p} \in \mathcal{P}_0) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \pi(\boldsymbol{y} \mid X; \boldsymbol{p} \in \mathcal{P}_0).$$

Trinomial example Under H_0, $X = Y_1 + Y_2$ is a sufficient statistic for ϕ. The conditional probability distribution of (Y_1, Y_2) given X is

$$\Pr(Y_1 = y, Y_2 = y \mid X = x) = \frac{x!}{y_1! y_2!} \left(\frac{1}{2}\right)^x$$

and the p-value is then the sum of these probabilities over the outcomes in the tail set,

$$P_{\text{suff}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} \frac{x!}{y_1! y_2!} \left(\frac{1}{2}\right)^x.$$

The p-value for outcome $\boldsymbol{y}_{\text{obs}} = (0, 2, 1)$ is found by considering only the outcomes with $X = 2$. They are (0,2,1), (1,1,1) and (2,0,1). Looking back at Table 1, we see that $T(0, 2, 1) = 2/3,\ T(1, 1, 1) = 0$ and $T(2, 0, 1) = 2/3$. The p-value for $(0, 2, 1)$ is then the sum of the probabilities $\Pr(\boldsymbol{Y} = \boldsymbol{y} | X = 2)$ for outcome $\boldsymbol{y} = (0, 2, 1)$ and $\boldsymbol{y} = (2, 0, 1)$ which are both 0.25, so the p-value is 0.50. The conditional probabilities and p-values for all the outcomes are given in Table 3. □

Outcome $\boldsymbol{y}$	x	$P(\boldsymbol{Y} = \boldsymbol{y}; X = x)$	P_{suff}
(0,0,3)	0	1	1
(0,1,2)	1	0.5	1
(0,2,1)	2	0.25	0.5
(0,3,0)	3	0.125	0.25
(1,0,2)	1	0.5	0.25
(1,1,1)	2	0.5	1
(1,2,0)	3	0.375	1
(2,0,1)	2	0.25	0.5
(2,1,0)	3	0.375	1
(3,0,0)	3	0.125	0.25

TABLE 3: P-values obtained for the trinomial example by conditioning on the sufficient statistic $X = Y_1 + Y_2$.

However, an appropriate sufficient statistic does not always exist. Instead of conditioning on a sufficient statistic, we may condition on an ancillary statistic, Berger and Boos (1994). We will not pursue this approach here.

FULL MAXIMIZATION Another approach to deal with nuisance parameters is to maximize over the set of unknown parameters, Casella and Berger (2002). In this approach, called full maximization by Lloyd (2008), the p-value is calculated as the supremum of the probability of the tail set over the parameter space of $\boldsymbol{p}$ under H_0, i.e. over $\mathcal{P}_0$. This p-value is valid and exact (as we will explain later in this section) and is given as

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \boldsymbol{p}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \boldsymbol{p})} \pi(\boldsymbol{y}; \boldsymbol{p}).$$

Trinomial example For each outcome we calculate the full maximization p-value by maximizing the sum of multinomial probabilities for the outcomes in the tail set over all values of $\phi,\ 0 \leq \phi \leq 1/2$. Thus, the full maximization p-value for each outcome is the maximum of the sums of multinomial probabilities,

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\phi \in [0, 0.5]} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \phi)} \frac{N!}{y_1! y_2 (N - y_1 - y_2)!} \phi^{y_1 + y_2} (1 - 2\phi)^{N - y_1 - y_2}.$$

In this example, numerically we used a grid for ϕ of 5001 points, $\{0, 0.0001, 0.0002, \ldots, 0.5000\}$ in the maximization. For the outcome $\boldsymbol{y}_{\text{obs}} = (0, 2, 1)$, in each grid point, the multinomial probabilities are calculated for the outcomes in the tail set defined by $T(\boldsymbol{y}) \geq T(\boldsymbol{y}_{\text{obs}})$ where $T(\boldsymbol{y})$ is given in (4), i.e. for the outcomes (0,2,1), (0,3,0), (2,0,1), (3,0,0), and added. Then the maximum of those sums is the full maximization p-value. For this outcome, the maximum p-value is obtained when $\phi = 0.4$, then $\pi((0, 2, 1); \phi = 0.4) = \pi((2, 0, 1); \phi = 0.4) = 0.096$ and $\pi((0, 3, 0); \phi = 0.4) = \pi((3, 0, 0); \phi = 0.4) = 0.064$. Adding these probabilities yields the p-value 0.32. The p-values for the other outcomes are given in the third column of Table 4 with the value of ϕ for which the maximum p-value is obtained in the second column. □

PARTIAL MAXIMIZATION Not all values of $\boldsymbol{p}$ are equally likely under the null hypothesis, therefore it might not be desirable to maximize over all possible values of $\boldsymbol{p}$. The set over which the supremum is found can be restricted to a confidence set for $\boldsymbol{p}$ as suggested by Berger and Boos (1994). This partial maximization p-value is valid when a penalty ζ is added, Berger and Boos (1994), but it is not exact by the definition of Lloyd (2008). It is given by

$$P_{\text{PM}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in C_\zeta} \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); f(\boldsymbol{p}) = 0) + \zeta = \sup_{\boldsymbol{p} \in C_\zeta} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \boldsymbol{p})} \pi(\boldsymbol{y}; \boldsymbol{p}) + \zeta,$$

where C_ζ is the $1 - \zeta$ confidence region for $\boldsymbol{p}$ under the null hypothesis.

Trinomial example Under H_0, $Y_1 + Y_2$ is binomially distributed with parameters 2ϕ and N. We will use the Clopper–Pearson confidence interval which is an exact confidence interval for binomial proportions, Agresti (2002). The $1 - \zeta$ confidence interval for ϕ is given by its lower limit C_{L} and upper limit C_{U},

$$C_{\text{L}} = \frac{1}{2}\left(1 + \frac{N - y_1 - y_2 + 1}{(y_1 + y_2) F_{2(y_1+y_2), 2(N-y_1-y_2+1)}(1 - \zeta/2)}\right)^{-1}$$

$$C_{\text{U}} = \frac{1}{2}\left(1 + \frac{N - y_1 - y_2}{(y_1 + y_2) F_{2(y_1+y_2+1), 2(N-y_1-y_2)}(\zeta/2)}\right)^{-1}$$

where $F_{\nu_1, \nu_2}(c)$ denotes the $1 - c$ quantile from the F distribution with ν_1 and ν_2 degrees of freedom. When $y_1 + y_2 = 0$, the lower limit is 0 and when $y_1 + y_2 = N = 3$, the upper limit is 0.50. We choose $\zeta = 0.001$ as suggested by Berger and Boos (1994) and calculate the p-values from the formula

$$P_{\text{PM}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\phi \in [C_{\text{L}}, C_{\text{U}}]} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}}; \phi)} \frac{N!}{y_1! y_2 (N - y_1 - y_2)!} \phi^{y_1 + y_2} (1 - 2\phi)^{N - y_1 - y_2} + \zeta.$$

The p-values are calculated the same way as the full maximization p-values except that the maximization is now done over the possible values of ϕ in the confidence interval. The tail set is unchanged. Table 4, column 4 and 5, show the partial maximization p-values and the value of ϕ for which the maximum p-value is found. We see that the partial maximization p-values are the same as the full maximization p-values, except for the outcomes (0,1,2) and (1,0,2) where the p-values are smaller because the value of $\boldsymbol{p}$ that maximizes the full maximization p-value is outside the confidence interval for ϕ used by the partial maximization p-value.□

ESTIMATION AND MAXIMIZATION Lloyd (2008) proposed the estimation followed by maximization (E+M) p-value, where the negative of the estimation p-values serve as the values of a new test statistic, followed by a full maximization step. In this way valid and exact p-values are obtained.

Since the test statistic is the estimation (E) p-values, performing the estimation step results in a different ordering of the outcomes before performing the maximization step than the ordering defined by the original test statistic. Thus the tail sets are changed and the E+M p-values may differ from the full maximization (M) p-values. The estimation step can be done more than once with or without a final maximization step, each time resulting in a different ordering of the outcomes. If two estimation steps are performed before a maximization step, the p-values are called $\mathrm{E}^2\mathrm{M}$ p-values, again yielding valid p-values.

Another way to look at the difference and similarity between the E and M p-values, is that for the E p-values, first the probability of the observed outcome is maximized through the maximum likelihood estimate of $\boldsymbol{p}$ under H_0, and then the probability of the tail set is calculated. For the M p-values, the tail sets are defined first, and then the probability of the tail set is maximized over $\boldsymbol{p}$ in $\mathcal{P}_0$. That is,

$$P_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\mathrm{obs}})} \sup_{\boldsymbol{p} \in \mathcal{P}_0} \pi(\boldsymbol{y}; \boldsymbol{p}),$$

and

$$P_{\mathrm{M}}(\boldsymbol{y}_{\mathrm{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\mathrm{obs}})} \pi(\boldsymbol{y}; \boldsymbol{p}).$$

Performing more than one maximization step in a sequence has no effect on the p-values. The reason is that the tail sets remain the same. Assume some chosen test statistic defines the tail set to be used in the first maximization step. The outcome having the largest value of this test statistic will have the smallest M p-value, the outcome having the second largest value of the test statistic will have the second smallest M p-value and so on. In the second maximization step, the negative of the M p-values are the values of the test statistic that defines the new tail set. The outcome having the largest negative M p-value is the outcome that had the largest value of the first test statistic, the outcome having the second largest negative M p-value is the outcome that had the second largest value of the first test statistic and so on. Since the p-value is the maximum sum of probabilities of the outcomes in the tail set, maximized over $\boldsymbol{p}$, and the tail set is the same, the p-values are unchanged.

Regardless of the choice of test statistic the M p-values are always valid. Assume that the chosen significance level is α. We would then reject the null hypothesis for all outcomes for which the test statistic $T(\boldsymbol{Y})$ is greater than or equal to some critical value k, where k is chosen so that all values of the test statistic that are greater than or equal to k yield a p-value less than or equal to α. The probability that a random outcome $\boldsymbol{y}_{\mathrm{obs}}$ is rejected under the null hypothesis is the probability that the test statistic $T(\boldsymbol{y}_{\mathrm{obs}})$ is greater than or equal to the critical value and this probability is less than or equal to the p-value for an outcome $\boldsymbol{y}$ for which $T(\boldsymbol{y}) = k$, which is less than or equal α. Also exactness of the M p-value follows by construction.

The M p-values are often conservative, see Bayarri and Berger (2000), whereas the E p-values are not valid and thus generally smaller than or equal to the M p-values. It is desired when comparing different p-values to obtain p-values as small as possible while still valid.

Trinomial example We first find the E p-values for all the outcomes where each p-value is calculated by inserting the maximum likelihood estimate $\tilde{\boldsymbol{p}}$ for $\boldsymbol{p}$ for that particular outcome as described previously and given in Table 2. The E p-values are used to define the tail set of the observed outcome

which is the set of outcomes for which $P_{\mathrm{E}}(\boldsymbol{y}) \leq P_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}})$. For the observed outcome we then calculate the E+M p-value as the sum of multinomial probabilities of the outcomes in the tail set maximized over ϕ. Let $R_{\mathrm{E}}(\boldsymbol{y}_{\mathrm{obs}})$ be the tail set of the observed outcome defined by the E p-values smaller than or equal to the E p-value of the observed outcome. The E+M p-value is then given by

$$P_{\text{E+M}}(\boldsymbol{y}) = \sup_{\phi \in [0,0.5]} \sum_{\boldsymbol{y} \in R_{\mathrm{E}}(\boldsymbol{y})} \frac{N!}{y_1! y_2! (N - y_1 - y_2)!} \phi^{y_1 + y_2} (1 - 2\phi)^{N - y_1 - y_2}.$$

The E p-value of outcome (0,2,1) is 0.30. The tail set $R^E(0, 2, 1) = \{(0, 2, 1), (0, 3, 0), (2, 0, 1), (3, 0, 0)\}$. This is the same tail set as when we used the test statistic $T(\boldsymbol{Y}) = |Y_1/N - Y_2/N|$ and therefore the maximization over ϕ here yields the same maximum p-value as the full maximization approach. The E+M p-values for all the outcomes are given in Table 4, column 7, with the value of ϕ for which the maximum p-value is found, ϕ_M, in column 6. The p-values are maximized with respect to ϕ over the same grid as in the full maximization approach. We see that for all the outcomes, except (0,1,2) and (1,0,2), the E+M p-values are the same as the full maximization p-values. For those two outcomes, the p-values are reduced from 1 to 0.5982 if we use the E+M approach, and they are the same outcomes for which the p-values were reduced when performing partial maximization instead of full maximization. □

Outcome $\boldsymbol{y}$	ϕ_{M}	P_{M}	ϕ_{PM}	P_{PM}	$\phi_{\mathrm{E+M}}$	$P_{\mathrm{E+M}}$
(0,0,3)	0.4535	1	0.4591	1	0.4535	1
(0,1,2)	0.50	1	0.4935	0.982	0.2265	0.5982
(0,2,1)	0.40	0.32	0.40	0.32	0.40	0.32
(0,3,0)	0.50	0.25	0.50	0.25	0.50	0.25
(1,0,2)	0.50	1	0.4935	0.982	0.50	0.5982
(1,1,1)	0.4535	1	0.4726	1	0.4535	1
(1,2,0)	0.50	1	0.50	1	0.4535	1
(2,0,1)	0.40	0.32	0.40	0.32	0.40	0.32
(2,1,0)	0.50	1	0.50	1	0.4535	1
(3,0,0)	0.50	0.25	0.50	0.25	0.50	0.25

TABLE 4: P-values obtained for the trinomial example by full maximization, partial maximization and estimation plus maximization with values of ϕ for which the p-values are maximized.

INTEGRATION In the partial maximization approach, the points in $\mathcal{P}_0$, the parameter space for $\boldsymbol{p}$ under H_0, are given weights 0 or 1. If $\boldsymbol{p}$ lie within their confidence interval, they are given weight 1 and 0 otherwise. Instead of weighing the probabilities with 0 and 1, we want to apply Bayesian methodology and weigh the points in $\mathcal{P}_0$ according to a prior distribution $\pi(\boldsymbol{p})$. We integrate out $\boldsymbol{p}$ in order to be able to calculate $\pi(\boldsymbol{y} \mid H_0)$. Bayarri and Berger (2000) review several Bayesian p-values as well as suggesting two new p-values. First there is the prior predictive p-value where a prior $\pi(\boldsymbol{p} \mid H_0)$ is chosen, so that the probability of an outcome under the null hypothesis is

$$\pi(\boldsymbol{y} \mid H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p}) \pi(\boldsymbol{p} \mid H_0) \mathrm{d}\boldsymbol{p}.$$

The prior predictive (PP) p-value of an observed outcome $\boldsymbol{y}_{\mathrm{obs}}$ is the sum of the probabilities

$\pi(\boldsymbol{y} \mid H_0)$ that are less than or equal to the probability $\pi(\boldsymbol{y}_{\text{obs}} \mid H_0)$. As we will see numerically in Section 4.2, these PP p-values are not valid.

Trinomial example We choose the uniform Dirichlet prior as the joint distribution of p_1 and p_2, thus $\pi(\boldsymbol{p}) = 2$. Let $z = p_1 - p_2$. Under H_0, $z = 0$, so that $\pi(\boldsymbol{p} \mid H_0) = \pi(\boldsymbol{p} \mid z = 0)$. The joint density of the transformed variables is $\pi(p_1, z) = 2$. Then $\pi(\boldsymbol{p} \mid z = 0) = \pi(p_1, z = 0)/\pi(z = 0)$. The density of z can be found by integrating out p_1 from $\pi(p_1, z)$, giving $\pi(z = 0) = 1$, after having identified the triangular region to which (p_1, z) belongs. The probability of the trinomial outcome given H_0 is

$$\pi(\boldsymbol{y} \mid H_0) = \int_0^{1/2} \frac{N!}{y_1!y_2!} p_1^{y_1+y_2}(1-2p_1)^{N-y_1-y_2} 2\mathrm{d}p_1 = \frac{(y_1+y_2)!}{y_1!y_2!(N+1)} \left(\frac{1}{2}\right)^{y_1+y_2}.$$

The p-value of the observed outcome is the sum of the probabilities $\pi(\boldsymbol{y} \mid H_0)$ that are less than or equal to the probability of the observed outcome. Table 5 shows the calculated probabilities as well as the p-values for the possible outcomes in the trinomial example. □

Outcome $\boldsymbol{y}$	$\pi(\boldsymbol{p} \mid H_0)$	P_{PP}
(0,0,3)	0.25	1
(0,1,2)	0.125	0.625
(0,2,1)	0.0625	0.1875
(0,3,0)	0.03125	0.0625
(1,0,2)	0.125	0.625
(1,1,1)	0.125	0.625
(1,2,0)	0.09375	0.375
(2,0,1)	0.0625	0.1875
(2,1,0)	0.09375	0.375
(3,0,0)	0.03125	0.0625

TABLE 5: P-values for the trinomial example using the Bayesian approach and a uniform Dirichlet prior on $\boldsymbol{p}$.

One challenge with the prior predictive approach is that the resulting p-values depend on the prior. To make them less dependent on the choice of prior and more dependent on the data, one can use the posterior predictive p-value, Bayarri and Berger (2000), where the probability of the observed outcome is given in terms of the posterior probability,

$$\pi(\boldsymbol{y} \mid H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p}) \pi(\boldsymbol{p} \mid \boldsymbol{y}_{\text{obs}}) \mathrm{d}\boldsymbol{p}.$$

To calculate this probability, improper priors can be used, and the probability will be less influenced by the choice of prior. However, the data are used twice since it first is needed to determine the posterior distribution and then in computing the tail set.

As improvements to the posterior predictive p-value, Bayarri and Berger (2000) also suggested the partial posterior predictive p-value and the conditional predictive p-value. In this work, we will only consider the prior predictive p-values.

TEST SIZE AND POWER. The test size and test power are common evaluation measures on the performance of a statistical hypothesis test. The test size is the probability of making a type I error,

i.e. to reject the null hypothesis, when it is true. The power is the probability of rejecting the null hypothesis when it is not true, which is one minus the probability of making a type II error, i.e. to not reject the null hypothesis when it is not true. Given the chosen significance level α, the test size for a test for which we reject the null hypothesis when the p-value $P(\boldsymbol{y})$ is less than α is

$$\Pr(P(\boldsymbol{Y}) \leq \alpha; \boldsymbol{p}) = \sum_{\boldsymbol{y};\, P(\boldsymbol{y}) \leq \alpha} \pi(\boldsymbol{y}; \boldsymbol{p}) \tag{5}$$

for a parameter $\boldsymbol{p}$ in $\mathcal{P}_0$.

The test power is

$$\Pr(P(\boldsymbol{y}) \leq \alpha; \boldsymbol{p}) = \sum_{\boldsymbol{y};\, P(\boldsymbol{y}) \leq \alpha} \pi(\boldsymbol{y}; \boldsymbol{p}) \tag{6}$$

for a parameter $\boldsymbol{p}$ in $\mathcal{P}$.

3 INDEPENDENT BINOMIAL PROPORTIONS

Before focusing on our main problem of comparing positive predictive values, we go through a fictitious example analyzed in Berger and Boos (1994) and Lloyd (2008), and present alternative test statistics and a more elaborate analysis of the p-values.

3.1 PRESENTATION OF THE PROBLEM

There are $n = 330$ subjects in a clinical trial of which $n_1 = 47$ subjects receive treatment and $n_2 = 283$ subjects receive placebo. Let X_1 be the number of subjects that survive among those who received treatment and let X_2 be the number of subjects that survive among those who received placebo. If p_1 is the survival probability for the treatment group and p_2 is the survival probability for the placebo group, we assume that X_1 is binomially distributed with parameters n_1 and p_1 and X_2 is binomially distributed with parameters n_2 and p_2. Let $\boldsymbol{X} = (X_1, X_2)$ and $\boldsymbol{p} = (p_1, p_2)$. The two-sided null hypothesis is that the survival probabilities in the two groups are equal, i.e.

$$H_0 : f(\boldsymbol{p}) = p_1 - p_2 = 0 \tag{7}$$

versus the alternative that they are not equal,

$$H_1 : f(\boldsymbol{p}) = p_1 - p_2 \neq 0.$$

Lloyd (2008) also considered the one-sided null hypothesis that the survival probability of the treatment group is no better than the survival probability of the placebo group, i.e.

$$H_0 : f(\boldsymbol{p}) = p_1 - p_2 \leq 0 \tag{8}$$

versus the alternative that the survival probability of the treatment group is better than the probability of the placebo group,

$$H_1 : f(\boldsymbol{p}) = p_1 - p_2 > 0.$$

Assuming independence between the treatment and placebo group, the joint distribution of X_1 and X_2 is the product of the two binomial distributions,

$$\Pr(X_1 = x_1, X_2 = x_2) = \binom{n_1}{x_1} p_1^{x_1}(1-p_1)^{n_1-x_1} \cdot \binom{n_2}{x_2} p_2^{x_2}(1-p_2)^{n_2-x_2}$$

In this situation, the reference set is all possible outcomes (x_1, x_2) given $n_1 = 47$ and $n_2 = 283$ which is a set of 13 682 outcomes. When calculating p-values various test statistics can be used to define the tail set. One of the test statistics used by Berger and Boos (1994) and Lloyd (2008) is

$$T_{\text{T}}(x_1, x_2) = \frac{x_1/n_1 - x_2/n_2}{\sqrt{(x_1+x_2)(n-x_1-x_2)/(nn_1n_2)}}.$$

When testing the null hypothesis (7) the tail set of an observed outcome $(x_{1,\text{obs}}, x_{2,\text{obs}})$ is $R(x_{1,\text{obs}}, x_{2,\text{obs}}) = \{(x_1, x_2) : |T_{\text{T}}(x_1, x_2)| \geq |T_{\text{T}}(x_{1,\text{obs}}, x_{2,\text{obs}})|\}$, and when testing the null hypothesis (8) the tail set is $R(x_{1,\text{obs}}, x_{2,\text{obs}}) = \{(x_1, x_2) : T_{\text{T}}(x_1, x_2) \geq T_{\text{T}}(x_{1,\text{obs}}, x_{2,\text{obs}})\}$.

Lloyd (2008) also uses the likelihood ratio test statistic

$$T_{\text{LR}} = 2\sum_{i=1}^{2}\left(x_i \text{log}\frac{\hat{p}_i}{\tilde{p}_i} + (n_i - x_i)\text{log}\frac{1-\hat{p}_i}{1-\tilde{p}_i}\right)$$

where $\hat{p}_i = x_i/n_i$ is the general maximum likelihood estimate for p_i, $i = 1, 2$ and $\tilde{p}_i$ is the maximum likelihood estimate for p_i, $i = 1, 2$ under the null hypothesis. If we are testing the two-sided null hypothesis $\tilde{p}_1 = \tilde{p}_2 = (x_1 + x_2)/(2n)$, and if we are testing the one-sided null hypothesis, $\tilde{p}_1 = \tilde{p}_2 = (x_1 + x_2)/(2n)$ when $x_1/n_1 \geq x_2/n_2$ and $\tilde{p}_i = x_i/n_i,\ i = 1, 2$, when $x_1/n_2 < x_2/n_2$. These estimates were also used for the E step. For the maximization in the M step, 1001 equally spaced values of $p_1 = p_2$ in [0,1] were used for the two-sided test and 5151 equally spaced points in a rectangular grid in the triangular region $0 \leq p_1 \leq 1$, $p_1 \leq p_2 \leq 1$, were used for the one-sided test.

In addition to T_{T} and T_{LR} we propose three additional test statistics. Let $\pi(\boldsymbol{x}; \boldsymbol{p})$ denote $\Pr(X_1 = x_1, X_2 = x_2; \boldsymbol{p})$. First, we define a simplified version of the likelihood ratio test statistic,

$$T_{\pi_{\text{e}}}(\boldsymbol{x}_{\text{obs}}) = \pi(\boldsymbol{x}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}),$$

which is simply the probability of the observed outcome $\boldsymbol{x}_{\text{obs}} = (x_{1,\text{obs}}, x_{2,\text{obs}})$ with the maximum likelihood estimate of $\boldsymbol{p}$ under H_0 for this outcome, $\tilde{\boldsymbol{p}}_{\text{obs}}$, inserted for $\boldsymbol{p}$.

In our second and third additional test statistic, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$, we let the probability $\pi(\boldsymbol{x}; \boldsymbol{p})$ of an outcome $\boldsymbol{x}$ play the role of a test statistic. It is of course dependent on the unknown parameters, and thus not a test statistic in the ordinary sense. It still makes sense to apply an E or M step to it, yielding the π_{E} p-value

$$T_{\pi_{\text{E}}}(\boldsymbol{x}_{\text{obs}}) = \Pr(\pi(\boldsymbol{X}; \tilde{\boldsymbol{p}}_{\text{obs}}) \leq \pi(\boldsymbol{x}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) = \sum_{\boldsymbol{x} \in R^*(\boldsymbol{x}_{\text{obs}})} \pi(\boldsymbol{x}; \tilde{\boldsymbol{p}}_{\text{obs}})$$

where $R^*(\boldsymbol{x}_{\text{obs}})$ consists of those $\boldsymbol{x}$ for which $\pi(\boldsymbol{x}; \tilde{\boldsymbol{p}}_{\text{obs}}) \leq \pi(\boldsymbol{x}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}})$, and the π_{M} p-value

$$T_{\pi_{\text{M}}}(\boldsymbol{x}_{\text{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(\pi(\boldsymbol{X}; \boldsymbol{p}) \leq \pi(\boldsymbol{x}_{\text{obs}}; \boldsymbol{p}); \boldsymbol{p}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{\boldsymbol{x} \in R^*(\boldsymbol{x}_{\text{obs}})} \pi(\boldsymbol{x}; \boldsymbol{p}),$$

where $R^*(\boldsymbol{x}_{\text{obs}})$ consists of those $\boldsymbol{x}$ for which $\pi(\boldsymbol{x}; \boldsymbol{p}) \leq \pi(\boldsymbol{x}_{\text{obs}}; \boldsymbol{p})$. Note that the sets $R^*(\boldsymbol{x})$ for these two statistics are dependent on the parameter $\boldsymbol{p}$, as opposed to the tail sets defined by an ordinary statistic. Although $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$ are constructed in a similar manner as p-values constructed by an E and an M step, respectively, their use will be as test statistics, rather than p-values.

3.2 COMPARISON OF TEST STATISTICS

Lloyd (2008) recommends using the E+M p-values in the problem of comparing independent binomial proportions and we want to compare the test size and power of the T_{T}, T_{LR}, $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$ test statistics for these p-values when testing both the one-sided and two-sided null hypotheses given in (7) and (8). To calculate the test size, i.e. the probability of rejecting the null hypothesis given that the null hypothesis is true, we generated 10001 equally spaced values of $p_1 = p_2$ in $[0, 1]$ and calculated the test size for each value $\boldsymbol{p}$ according to (5) by adding the probabilities of the outcomes that had a p-value less than or equal 0.05, which was the chosen significance level. To assess power, we used 9001 equally spaced points on the line $p_1 = p_2 + 0.1$, for which we calculated the power by adding the probabilities of the outcomes that had p-values less than or equal to 0.05, i.e. using (6).

Table 6 shows the mean test size and power for the five test statistics followed by an E and an M step. For the two-sided hypothesis, T_{T}, $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$ have similar mean test size, of which $T_{\pi_{\mathrm{e}}}$ has the greatest. T_{LR} yields a smaller test size than the other test statistics. When testing the one-sided hypothesis, the test size of $T_{\pi_{\mathrm{e}}}$ is 0.0434 which is greater than the test size for the other test statistics which ranges from 0.0382 to 0.0387. We see that the test statistics have similar power, lower for the two-sided test than for the one-sided test, and for the two-sided test, T_{LR} has the smallest power and $T_{\pi_{\mathrm{e}}}$ has the largest power. For the one-sided test, $T_{\pi_{\mathrm{E}}}$ has the lowest power and $T_{\pi_{\mathrm{e}}}$ has the largest power. This indicates that for this problem, the $T_{\pi_{\mathrm{e}}}$ statistic performs best and should be considered an alternative to T_{T} and T_{LR}.

Table 7 shows the E+M p-values for the observed outcome $(x_1, x_2) = (14, 48)$ which for T_{T} and T_{LR} agree with Lloyd (2008). For the two-sided test, the p-value for outcome (14,48), which is used as a test case by Berger and Boos (1994) and Lloyd (2008), is less than 0.05 for all the test statistics except T_{LR} so the null hypothesis would be rejected on a 5% significance level for four of the test statistics. T_{T} yields the smallest p-value. For T_{LR} we reject the one-sided null hypothesis. All test statistics yield the p-value 0.025 for the one-sided test and thus reject the null hypothesis.

Hypothesis	T_{T}	T_{LR}	$T_{\pi_{\mathrm{e}}}$	$T_{\pi_{\mathrm{E}}}$	$T_{\pi_{\mathrm{M}}}$
Two-sided	0.0472	0.0435	0.0479	0.0478	0.0472
One-sided	0.0386	0.0384	0.0434	0.0382	0.0387
Two-sided	0.3629	0.3475	0.3649	0.3644	0.3622
One-sided	0.4421	0.4410	0.4639	0.4388	0.4427

TABLE 6: Mean test size in the two upper rows and mean test power in the two lower rows for the two-sided and one-sided hypothesis for the E+M p-values using different test statistics.

Hypothesis	T_{T}	T_{LR}	$T_{\pi_{\mathrm{e}}}$	$T_{\pi_{\mathrm{E}}}$	$T_{\pi_{\mathrm{M}}}$
Two-sided	0.037	0.057	0.040	0.041	0.040
One-sided	0.025	0.025	0.025	0.025	0.025

TABLE 7: E+M p-values from the two-sided and one-sided tests for outcome (14,48).

Figure 1 shows the E+M p-values for $T_{\pi_{\mathrm{e}}}$ test plotted against the E+M p-values for T_{T} for p-values less than or equal to 0.11. The plot shows that $T_{\pi_{\mathrm{e}}}$ overall yields smaller p-values than T_{T} and as we want p-values that are as small as possible provided that they are valid, $T_{\pi_{\mathrm{e}}}$ seems to be preferable

over T_{T}.

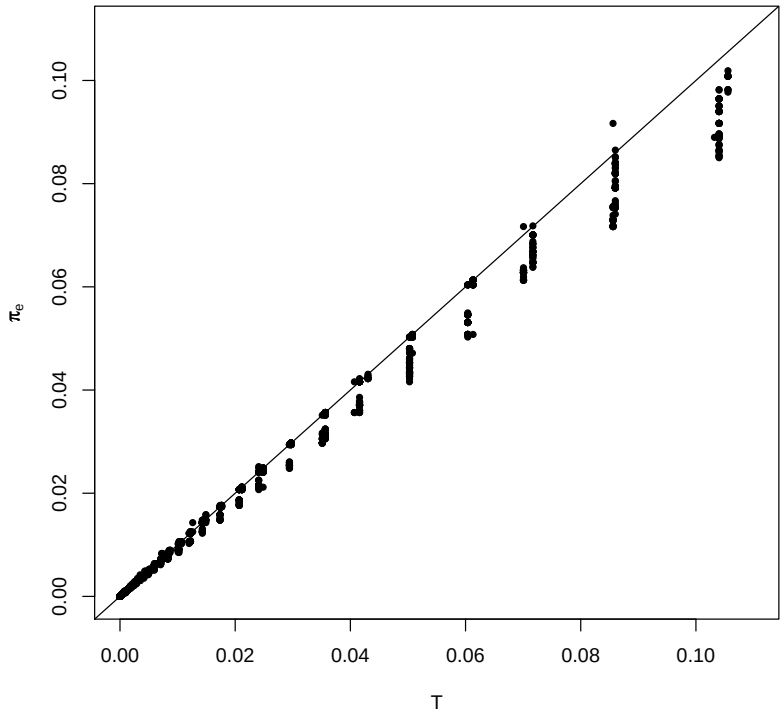


FIGURE 1: E+M p-values for $T_{\pi_{\mathrm{e}}}$ plotted against E+M p-values for T_{T}.

To explain what happens in the E and M steps, we look into the results for three particular outcomes, (16,52), (1,0) and (30,126), which are consecutive decreasing outcomes when ordering by T_{T}. Table 8 shows the value of T_{T} for these outcomes in the third column and the probabilities $\pi(\boldsymbol{x};\boldsymbol{p})$ of the outcomes inserted the maximum likelihood estimates under the null hypothesis in the second column. Note how much larger $\pi((1,0);\tilde{\boldsymbol{p}}_{(1,0)})$ is than $\pi((16,52);\tilde{\boldsymbol{p}}_{(16,52)})$ and $\pi((30,126);\tilde{\boldsymbol{p}}_{(30,126)})$. When performing the E step, this larger probability has a significant effect, as it is included in the sum of probabilities that yields the p-value for outcome (1,0). The fourth column of Table 8 shows the E p-values for the three outcomes. The p-value for outcome (1,0) is 0.05970 which is much greater than the two other p-values and H_0 is rejected on a 5% significance level, the other two p-values are both less than 0.01 and H_0 will not be rejected. Thus, the decision of whether to reject H_0 differ for these three outcomes, even though the values of T_{T} are almost the same. The effect of the large probability of outcome (1,0) also shows in the M p-values in the fifth column in Table 8. We note a large increase in the M p-value for outcome (30,126) as compared to the E p-value, the reason being that the M p-value is at least as large as the E p-value by construction, in particular for (1,0), and next, that the the M p-value is at least as large as the M p-value of (1,0), since (1,0) has a larger test statistic value. According to the M p-values, we would not reject H_0 for any of those outcomes as opposite to the E p-values where H_0 is rejected for outcome (30,126). To avoid the effect of outcome (1,0), we need a different ordering of the outcomes, in which (1,0) is placed further down on the list where the outcomes are sorted by decreasing value of a test statistic. This is obtained by treating the E p-value as

$\boldsymbol{x}$	$\pi(\boldsymbol{x};\tilde{\boldsymbol{p}})$	$T_{\mathrm{T}}(x_1,x_2)$	P_{E}	P_{M}	$P_{\mathrm{E+M}}$
(16,52)	0.00049	2.45928	0.00968	0.02458	0.01029
(1,0)	0.05247	2.45756	0.05970	0.06112	0.07229
(30,126)	0.00028	2.45512	0.00722	0.06112	0.00746

TABLE 8: The test statistic $T_{\mathrm{T}}(x_1,x_2)$ and corresponding E, M and E+M p-values for outcomes (16,52), (1,0) and (30,126).

a test statistic and then applying the M step. The outcomes that had unusually large p-values after the E step compared to their neighbours, e.g. as outcome (1,0) had, is then moved down on the list when sorting the outcomes by decreasing negative E p-values. The sixth column of Table 8 shows these E+M p-values and we see that outcome (30,126) now has a p-value of 0.00746 and is thus unaffected by the outcome (1,0). This example indicates why it is beneficial to perform an E step prior to the M step. The M step is necessary to obtain valid p-values and therefore the E step alone is not sufficient.

The M p-value for the outcome (14,48) is 0.06114, see Lloyd (2008), which is significantly greater than the E+M p-value of Table 7. Our investigation showed that the value 0.06114 also arises from outcome (1,0), because the value of T_{T} is less for outcome (14,48) than for outcome (1,0) which is thus in the tail set of (14,48). The E+M p-value is smaller (0.025) because the E step changed the ordering of the outcomes and (1,0) was placed behind (14,48) so that after performing the E step, (1,0) is no longer in the tail set of (14,48). It should also be noted that $\pi((14,48);\phi)$, as a function of $\phi = p_1 = p_2$ has a prominent and narrow peak near $\phi = 0$ (but $\phi > 0$), which explains the large value of $\pi((1,0);\tilde{\boldsymbol{p}})$ and thus of $P_{\mathrm{E}}(1,0)$. Partial maximization avoids this peak, explaining that the PM p-value is reasonable as reported by Berger and Boos (1994), though not as small as the E+M p-values as reported by Lloyd (2008).

4 COMPARING POSITIVE PREDICTIVE VALUES

We now present the main problem of comparing the positive predictive values of two diagnostic tests. The performance of various test statistics and p-values are compared in terms of test size and power.

4.1 PRESENTATION OF THE PROBLEM

Suppose that two diagnostic tests are available for a particular disease of interest. We want to compare the prediction abilities of the two tests, which can be quantified by the positive and negative predictive values. The positive predictive value is defined as the probability that a subject has the disease given that the test is positive and the negative predictive value is the probability that a subject does not have the disease given that the test is negative. Without loss of generality, in this work we will only consider the positive predictive values, as the tests for the negative predictive values can easily be derived along the same lines. We want to test whether the positive predictive value of test A is equal to the positive predictive value of test B against the alternative that they are not equal;

$$H_0\colon \mathrm{PPV_A} = \mathrm{PPV_B} \quad \text{vs} \quad H_1\colon \mathrm{PPV}_A \neq \mathrm{PPV}_B.$$

In this situation we define six random variables that are given in Table 9. Let $\boldsymbol{Y}$ be the vector of these

random variables, i.e. $\boldsymbol{Y} = (Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)$. We assume that $\boldsymbol{Y}$ is multinomially distributed with parameters N and $\boldsymbol{p} = (p_1, p_2, p_3, p_4, p_5, p_6)$. Thus, the probability function of $\boldsymbol{Y}$ is

$$\Pr(\boldsymbol{Y} = \boldsymbol{y}) = N! \prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!}.$$

Variable	Description
Y_1	Number of non-diseased subjects with positive test A and B.
Y_2	Number of non-diseased subjects with positive test A and negative test B.
Y_3	Number of non-diseased subjects with negative test A and positive test B.
Y_4	Number of diseased subjects with positive test A and B.
Y_5	Number of diseased subjects with positive test A and negative test B.
Y_6	Number of diseased subjects with negative test A and positive test B.

TABLE 9: Definition of the random variables $Y_1, \ldots, Y_6$.

The positive predictive value of test A is

$$\text{PPV}_A = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5}$$

and the positive predictive value of test B is

$$\text{PPV}_B = \frac{p_4 + p_6}{p_1 + p_3 + p_4 + p_6}.$$

The null hypothesis is then

$$H_0 : \ f_{\text{PPV}}(\boldsymbol{p}) = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5} - \frac{p_4 + p_6}{p_1 + p_3 + p_4 + p_6} = 0. \tag{9}$$

The parameters $\boldsymbol{p}$ are not completely determined by the null hypothesis, we only know that $f_{\text{PPV}}(\boldsymbol{p}) = 0$ and that $\sum_{i=1}^{6} p_i = 1$. Thus, from these two constraints, two of the parameters can be expressed in terms of the four other remaining parameters, but these four parameters will be unknown nuisance parameters.

In order to test the null hypothesis (9) we will calculate p-values by enumeration as described by the algorithm in Section 2.3. The first step is to find the reference set.

4.1.1 FINDING THE REFERENCE SET

In the first step in the algorithm for calculating p-values, we enumerate using five nested for-loops to find all possible outcomes having the value of N, which is the number of observations. It can be distributed among six non-negative integer random variables having sum N, and the number of possible outcomes is $\binom{N+5}{5}$. This is a general result for the number of distinct unordered selections of N elements from six elements drawn with replacement. Figure 2 shows the number of outcomes on a log10 scale plotted against N. The number of outcomes grows very quickly when N increases. When $N = 25$ there are 142506 possible outcomes and when $N = 50$, there are nearly 3.5 million possible outcomes.

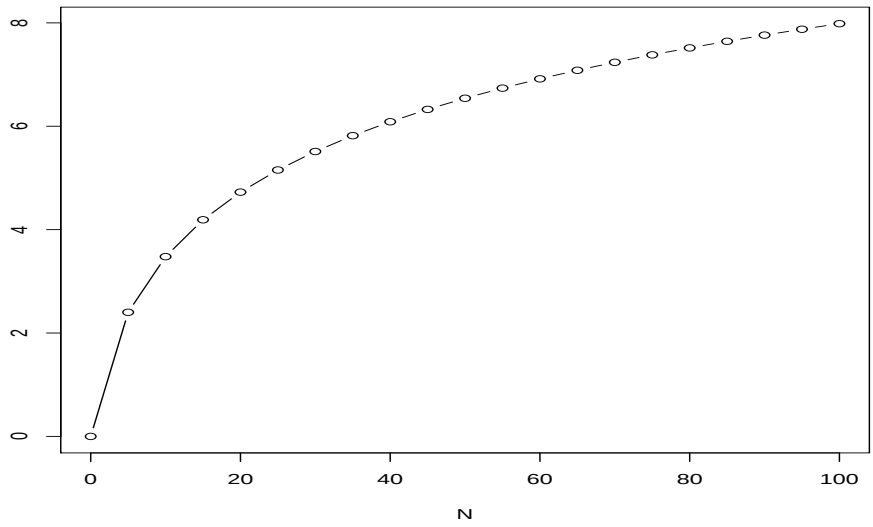


FIGURE 2: Number of possible outcomes on $\log_{10}$ scale as a function of N .

4.1.2 CALCULATING THE PROBABILITY OF THE OBSERVED OUTCOME

In the setting of comparing positive predictive values we will focus on calculating the probability of an outcome either by substituting maximum likelihood estimates for $\boldsymbol{p}$, maximize over the parameter space for $\boldsymbol{p}$ or integrate out $\boldsymbol{p}$ by a Bayesian approach. This results in the estimation (E), maximization (M) or combinations of these like the estimation and maximization (E+M) p-values, and the Bayesian prior predictive p-values. As far as we know, there is no sufficient or ancillary statistic for $\boldsymbol{p}$ in this problem. To calculate the p-values, a test statistic $T(\boldsymbol{Y})$ must be chosen. There are several possible test statistics for this problem, and they will be presented in Section 4.1.3.

ESTIMATION AND MAXIMIZATION If we substitute the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ for $\boldsymbol{p}$ under H_0, the E p-value for an outcome $\boldsymbol{y}_{\text{obs}}$ is

$$P_{\text{E}}(\boldsymbol{y}_{\text{obs}}) = \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) = \sum_{\boldsymbol{y} \in R(\boldsymbol{y}_{\text{obs}})} N! \prod_{i=1}^{6} \frac{\tilde{p}_{i,\text{obs}}^{y_i}}{y_i!} \tag{10}$$

where the tail set $R(\boldsymbol{y})$ is defined by the chosen test statistic, $T(\boldsymbol{Y})$, and $\tilde{p}_{i,\text{obs}}$ is the maximum likelihood estimate under H_0 for p_i, $i = 1, \ldots, 6$ for the outcome $\boldsymbol{y}_{\text{obs}}$.

By maximizing the probability of the outcome $\boldsymbol{y}_{\text{obs}}$ over the parameter space $\mathcal{P}_0$ where $f_{\text{PPV}}(\boldsymbol{p}) = 0$, the M p-value is given by

$$P_{\text{M}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(T(\boldsymbol{Y}) \geq T(\boldsymbol{y}_{\text{obs}}); \boldsymbol{p}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \sum_{y \in R(\boldsymbol{y}_{\text{obs}})} N! \prod_{i=1}^{6} \frac{p_{i,\text{obs}}^{y_i}}{y_i!}, \tag{11}$$

where $R(\boldsymbol{y})$ is defined by the chosen test statistic. When we calculate the E+M p-value, the expression is the same as in (11), but the tail set is then defined by the p-values calculated from (10). We will also

consider the double estimation (E^2) p-values, where we first calculate p-values from (10) and then use these p-values as test statistics to define the tail set when calculating p-values from (10) once more. Finally we will maximize the E^2 p-values by using these as test statistics to define the tail set in (11), which results in E^2M p-values.

INTEGRATION We also consider the Bayesian prior predictive p-values, which requires a different approach. The starting point is still that the probability of an outcome under the null hypothesis is unknown because the parameters $\boldsymbol{p}$ are not completely specified. Instead of estimating $\boldsymbol{p}$ or maximizing the p-values over $\boldsymbol{p}$, we weigh them according to how likely they are under the null hypothesis.

We start out by conditioning on the parameter $\boldsymbol{p}$. Let $\pi(\boldsymbol{y}|H_0)$ be the probability of the outcome $\boldsymbol{y}$ under the null hypothesis (9). Then

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \pi(\boldsymbol{y}|\boldsymbol{p}) \cdot \pi(\boldsymbol{p}|H_0) \mathrm{d}\boldsymbol{p} \tag{12}$$

The first factor of the integrand, the probability of $\boldsymbol{y}$ given $\boldsymbol{p}$ is simply the multinomial distribution, i.e.,

$$\pi(\boldsymbol{y}|\boldsymbol{p}) = N! \prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!},$$

and $\pi(\boldsymbol{p}|H_0)$ is the probability density function for $\boldsymbol{p}$ under the null hypothesis.

Since $p_6 = 1 - \sum_{i=1}^{5} p_i$ we first reduce the problem to five unknown parameters. Let

$$z = \frac{p_4 + p_5}{p_1 + p_2 + p_4 + p_5} - \frac{1 - p_1 - p_2 - p_3 - p_5}{1 - p_2 - p_5}. \tag{13}$$

which is $f_{\text{PPV}}(\boldsymbol{p})$ (9) with $1 - \sum_{i=1}^{5} p_i$ inserted for p_6.

Under the null hypothesis $z = 0$ and from this an expression for p_4 can be derived, yielding four unknown parameters. We change variables from p_1, p_2, p_3, p_4, p_5 to p_1, p_2, p_3, z, p_5. The vector of the new parameters is denoted $\boldsymbol{p}^*$ in the following. Then $\pi(\boldsymbol{p}^*|z = 0) \propto \pi(\boldsymbol{p}^*, z = 0)$ so therefore we start by finding $\pi(\boldsymbol{p}^*, z)$. We use the formula for change of variables, $\pi(\boldsymbol{p}) = \pi(\boldsymbol{p}^*, z) \cdot |J|$, where J is the Jacobian determinant,

$$J = \begin{vmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \frac{\partial z}{\partial p_1} & \frac{\partial z}{\partial p_2} & \frac{\partial z}{\partial p_3} & \frac{\partial z}{\partial p_4} & \frac{\partial z}{\partial p_5} \\ 0 & 0 & 0 & 0 & 1 \end{vmatrix} = \frac{\partial z}{\partial p_4}$$

The absolute value $|J|$ is $\frac{p_1+p_2}{(p_1+p_2+p_4+p_5)^2}$. As a prior distribution for $\boldsymbol{p}$ we first apply the Dirichlet distribution with parameters $\alpha_1 = \alpha_2 = \ldots \alpha_5 = 1$, thus $\pi_1(\boldsymbol{p})$ is constant. Then

$$\pi(\boldsymbol{p}^*|z = 0) \propto \pi(\boldsymbol{p}^*, z = 0) = \pi_1(\boldsymbol{p}) \cdot |J|^{-1} \propto \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2},$$

and

$$\pi(\boldsymbol{p}|H_0) = \frac{1}{k_1} \cdot \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2},$$

where k_1 is a normalizing constant which can be found from $\int_{\mathcal{P}_0} \pi(\boldsymbol{p}|H_0) = 1$.

This leads to the following expression for the probability under H_0 of an outcome $\boldsymbol{y}$,

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_1} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{(p_1+p_2+p_4+p_5)^2}{p_1+p_2} \mathrm{d}\boldsymbol{p}. \tag{14}$$

For details on how to numerically compute the integral, see Section 5.

The p-value is the sum of these probabilities for outcomes in the tail set of the observed outcome. In this setting we use the probability of an outcome under H_0 as the test statistic and the tail set for an outcome $\boldsymbol{y}_{\text{obs}}$ is defined as the outcomes for which $\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)$, so the p-value is given as

$$\begin{aligned} P_{\text{PP},1}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_1} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{(p_1+p_2+p_4+p_5)^2}{p_1+p_2} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{15}$$

To assess the effect of the choice of prior, we choose the non-uniform prior $\pi_2(\boldsymbol{p}) \propto p_1$ as an alternative prior, which leads to

$$\pi_2(\boldsymbol{p}|H_0) = \frac{p_1(p_1+p_2+p_4+p_5)^2}{k_2(p_1+p_2)},$$

where k_2 is a normalizing constant and the probability under H_0 of an outcome $\boldsymbol{y}$ is

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_2} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{p_1(p_1+p_2+p_4+p_5)^2}{p_1+p_2} \mathrm{d}\boldsymbol{p}. \tag{16}$$

We see that the probability of the outcome depends on the chosen prior $\pi_2(\boldsymbol{p})$ as expected. The p-value, which is the sum of the probabilities in (16) for the outcomes that are in the tail set of the one observed, is denoted $P_{\text{PP},2}$ and given by

$$\begin{aligned} P_{\text{PP},2}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_2} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{p_1(p_1+p_2+p_4+p_5)^2}{p_1+p_2} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{17}$$

An alternative formulation of the null hypothesis (9) is

$$f^*_{\text{PPV}}(\boldsymbol{p}) = p_1(p_1+p_2+p_3+p_4+2p_5-1) - p_2(1-p_1-p_2-p_3-p_5) + p_3(p_4+p_5) = 0. \tag{18}$$

In this case the absolute value of the Jacobi determinant will be p_1+p_3 and if we assume the uniform Dirichlet prior $\pi_1(\boldsymbol{p})$,

$$\pi(\boldsymbol{p}|H_0) = \frac{1}{k_3} \cdot \frac{1}{p_1+p_3},$$

where k_3 is a normalizing constant and the probability under H_0 of an outcome is

$$\pi(\boldsymbol{y}|H_0) = \int_{\mathcal{P}_0} \frac{N!}{k_3} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \cdot \frac{1}{p_1+p_3} \mathrm{d}\boldsymbol{p}. \tag{19}$$

This probability is clearly not equal to the probability (14) and this is an example of Borel's paradox. The p-value is the sum of the probabilities in (19) for the outcomes in the tail set of the observed outcome. It is denoted $P_{\text{PP},3}$ and given by

$$\begin{aligned} P_{\text{PP},3}(\boldsymbol{y}_{\text{obs}}) &= \Pr(\pi(\boldsymbol{Y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)) \\ &= \sum_{\pi(\boldsymbol{y}|H_0)\leq\pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \frac{N!}{k_3} \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!}\right) \frac{1}{p_1+p_3} \mathrm{d}\boldsymbol{p}. \end{aligned} \tag{20}$$

4.1.3 Defining the tail set

The tail set of an outcome $\boldsymbol{y}$ is defined by a test statistic $T(\boldsymbol{y})$. To test the null hypothesis in (9) there are several tests available that are used for large samples, see Günther, Bakke, Lydersen and Langaas (2008) for a detailed description of four possible test statistics. In this work we will use these test statistics to define the tail set while ignoring their asymptotic distribution.

The first test statistic is the likelihood ratio test statistic which is the ratio between the maximum likelihood under the null hypothesis and the general maximum likelihood, of which by convenience the logarithm is taken and which is multiplied by -2, Casella and Berger (2002). In our multinomial situation, it is given as

$$T_{\text{LR}} = -2 \cdot \log \frac{\sup_{\boldsymbol{p}\in\mathcal{P}_0} L(\boldsymbol{p}|\boldsymbol{y})}{\sup_{\boldsymbol{p}\in\mathcal{P}} L(\boldsymbol{p}|\boldsymbol{y})} = -2 \sum_{i=1}^{6} y_i \cdot (\log \tilde{p}_i - \log \hat{p}_i), \tag{21}$$

where $\tilde{p}_i$ is the restricted maximum likelihood estimates of p_i, i.e. under H_0, $i = 1, \ldots, 6$, and $\hat{p}_i$ is the unrestricted general maximum likelihood estimates for the multinomial distribution, i.e., $\hat{p}_i = n_i/N$, $i = 1, \ldots, 6$, Johnson, Kotz and Balakrishan (1997). The maximum likelihood estimates under H_0, $\tilde{p}_i$, $i = 1, \ldots, 6$ cannot be written in closed form, but can be found analytically by solving a system of equations arising from the method of Lagrange multipliers, which we did using Maple 12. More details can be found in Günther et al. (2008), Section 3.1.2.

The difference test statistic is given by

$$T_g(\boldsymbol{y}) = \frac{(g(\boldsymbol{Y}) - g(\boldsymbol{\mu}))^2}{G^T(\boldsymbol{\mu})\boldsymbol{\Sigma}\, G(\boldsymbol{\mu})} \tag{22}$$

where $g(\boldsymbol{Y})$ is an estimator for the difference $f_{\text{PPV}}(\boldsymbol{p})$ in (9), i.e.

$$g(\boldsymbol{Y}) = \frac{Y_4 + Y_5}{Y_1 + Y_2 + Y_4 + Y_5} - \frac{Y_4 + Y_6}{Y_1 + Y_3 + Y_4 + Y_6},$$

and $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{Y}) = N \cdot \boldsymbol{p}$, $\Sigma = \mathrm{Cov}(\boldsymbol{Y}) = N(\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}^T\boldsymbol{p})$, G is a vector containing the first order partial derivatives of $g(\boldsymbol{Y})$ with respect to the components of $\boldsymbol{Y}$, G^T is the transpose of G, and $G(\boldsymbol{\mu})$ is G with $\boldsymbol{\mu}$ inserted for $\boldsymbol{Y}$. Under the null hypothesis $g(\boldsymbol{\mu}) = 0$. $G(\boldsymbol{\mu})$ and $\boldsymbol{\Sigma}$ depend on the unknown parameters $\boldsymbol{p}$ which must be estimated when calculating the test statistic. We can either insert the unrestricted maximum likelihood estimates for the multinomial distribution $\hat{\boldsymbol{p}}$ and then we refer to the test as the *unrestricted* difference test (uDT) and denote the test statistic T_{uDT}, or insert restricted maximum likelihood estimates under H_0, $\tilde{\boldsymbol{p}}$. Then the test is referred to as the *restricted* difference test (rDT) and the test statistic is denoted T_{rDT} .

Leisenring, Alonzo and Pepe (2000) presented a score test based on generalized estimating equations. We denote this test the LAP test. The test statistic can be written as

$$T_{LAP} = \frac{((Y_1 + Y_2 + Y_4 + Y_5)(Y_4 + Y_6) - (Y_1 + Y_3 + Y_4 + Y_6)(Y_4 + Y_5))^2}{h(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6)}, \tag{23}$$

where

$$\begin{aligned}
&h(Y_1, Y_2, Y_3, Y_4, Y_5, Y_6) \\
&\qquad = Y_1(Y_2 - Y_3 + Y_5 - Y_6)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\
&\qquad\quad + Y_2(Y_1 + Y_3 + Y_4 + Y_6)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\
&\qquad\quad + Y_3(Y_1 + Y_2 + Y_4 + Y_5)^2 \left(\frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\
&\qquad\quad + Y_4(Y_2 - Y_3 + Y_5 - Y_6)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\
&\qquad\quad + Y_5(Y_1 + Y_3 + Y_4 + Y_6)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 \\
&\qquad\quad + Y_6(Y_1 + Y_2 + Y_4 + Y_5)^2 \left(1 - \frac{2Y_4 + Y_5 + Y_6}{2Y_1 + Y_2 + Y_3 + 2Y_4 + Y_5 + Y_6} \right)^2 .
\end{aligned}$$

These four test statistics will be used to define the tail set for the E and M p-values. Other test statistics are possible and we suggest three, $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$, which are defined in the same way as they were for the independent binomial proportions (Section 3), but with the multinomial distribution with six parameters substituted for the joint distribution of two independent binomial distributions, that is,

$$T_{\pi_{\text{e}}}(\boldsymbol{y}_{\text{obs}}) = \pi(\boldsymbol{y}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}), \tag{24}$$

$$T_{\pi_{\text{E}}}(\boldsymbol{y}_{\text{obs}}) = \Pr(\pi(\boldsymbol{Y}; \tilde{\boldsymbol{p}}_{\text{obs}}) \leq \pi(\boldsymbol{y}_{\text{obs}}; \tilde{\boldsymbol{p}}_{\text{obs}}); \tilde{\boldsymbol{p}}_{\text{obs}}) \tag{25}$$

and

$$T_{\pi_{\text{M}}}(\boldsymbol{y}_{\text{obs}}) = \sup_{\boldsymbol{p} \in \mathcal{P}_0} \Pr(\pi(\boldsymbol{Y}; \boldsymbol{p}) \leq \pi(\boldsymbol{y}_{\text{obs}}; \boldsymbol{p}); \boldsymbol{p}). \tag{26}$$

Finally, we also consider the Bayesian prior predictive p-values, that in addition to being p-values in their own right, can be used as test statistics to define the critical region for the E and M p-values, and we denote them T_{PP} where

$$T_{\text{PP}}(\boldsymbol{y}_{\text{obs}}) = \sum_{\pi(\boldsymbol{y}|H_0) \leq \pi(\boldsymbol{y}_{\text{obs}}|H_0)} \int_{\mathcal{P}_0} \pi(\boldsymbol{y} \mid \boldsymbol{p}) \cdot \pi(\boldsymbol{p} \mid H_0) \mathrm{d}\boldsymbol{p}. \tag{27}$$

4.2 RESULTS

In the PPV setting, we have studied the performance of the different types of p-values with respect to test statistics and the parameters in the multinomial distribution. The performance will be evaluated in terms of test size and test power, which are calculated as given by (5) and (6). We choose the significance level $\alpha = 0.05$.

4.2.1 EVALUATION OF TEST SIZE

The test statistics considered were the LAP, likelihood ratio, unrestricted difference and restricted difference test statistics, T_{LAP}, T_{LR}, T_{uDT} and T_{rDT}. In addition we used the π_{e}-probabilities and the π_{E}, π_{M} and Bayesian p-values as test statistics, i.e. $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{M}}}$ and T_{PP}. For each of these test statistics and for the chosen values of N we calculated the E, M, E+M, E^2 and E^2M p-values. We also considered the performance of the Bayesian p-values as p-values in itself.

The performance of the test statistics can depend highly on the parameters $\boldsymbol{p}$ in the multinomial distribution. Both the overall mean performance as well as the performance for specific values are evaluated. For the mean performance a set of 10385 values of $\boldsymbol{p}$ in $\mathcal{P}_0$ is used, which are obtained by using a four dimensional grid for the four free parameters where each side in the grid is divided into 30 subintervals, and the 10385 values of $\boldsymbol{p}$ in the grid that belong to $\mathcal{P}_0$ are then the cases we consider. For this set of cases we calculate the mean test size, i.e. we calculate the test size from (5) for each case and then find the average for the 10385 cases. In addition, six specific cases of $\boldsymbol{p}$ in $\mathcal{P}_0$ are evaluated, see Table 10. These are the same cases as in Günther, Bakke and Langaas (2009) where the reasoning for choosing these values can be found.

Case	p_1	p_2	p_3	p_4	p_5	p_6
1	0.068	0.135	0.135	0.527	0.068	0.068
2	0.043	0.130	0.130	0.348	0.174	0.174
3	0.267	0.267	0.267	0.067	0.067	0.067
4	0.300	0.267	0.267	0.033	0.067	0.067
5	0.400	0.200	0.200	0.100	0.050	0.050
6	0.450	0.200	0.200	0.050	0.050	0.050

TABLE 10: Specification of multinomial parameters under H_0.

The size of the multinomial sample determines how many possible outcomes there are and is an interesting factor to consider. We want to investigate whether the performance of the test statistics depends on sample size, in particular for small sample sizes, so we use $N = 10, 15, 20, 25$.

When maximizing the p-values over $\boldsymbol{p}$ in $\mathcal{P}_0$ we used a four-dimensional grid since there are four free parameters, with 50 points on each side. In addition, the maximum likelihood estimates for all possible outcomes given N were included in the grid. The Bayesian p-values were calculated on the grid with 50 points in each side, for further details see Section 5.

Table 11 shows the mean test size for all the test statistics, values of N and type of p-values investigated. We first compare the performance of the different types of p-values, E, M, E+M, E^2 and E^2M.

The M p-values yield the smallest test size for all values of N for all the test statistics except for the likelihood ratio test when $N = 20$, there the E^2M p-values yields the smallest test size. In general the E and E^2 p-values result in larger test sizes than the E+M and E^2M p-values which we would expect since the E and E^2 p-values are not valid, whereas the E+M and E^2M p-values are. The exception is T_{uDT}, when $N = 10$, the E p-values yield smaller test size than the E+M and E^2M p-values, and when $N = 15$, the E p-values yield smaller test size than the E^2M p-values. The E^2 p-values yield larger test sizes than the E p-values, except for $T_{\pi_{\text{e}}}$.

Next we compare the performance of the different test statistics. First we consider the test statistics that originated from large samples where their asymptotic distributions were utilized, i.e. the LAP,

N	p-value	LRT	LAP	uDT	rDT	π_{M}	π_{E}	π_{e}	PP_1	PP_3
10	M	0.0164	0.0092	0.0019	0.0130	0.0177	0.0088	0.0145	0.0104	0.0190
10	E	0.0381	0.0282	0.0179	0.0352	0.0388	0.0359	0.0643	0.0330	0.0366
10	E+M	0.0285	0.0198	0.0181	0.0285	0.0286	0.0242	0.0207	0.0257	0.0297
10	E^2	0.0456	0.0420	0.0395	0.0439	0.0435	0.0441	0.0549	0.0441	0.0435
10	$\mathrm{E}^2\mathrm{M}$	0.0273	0.0247	0.0220	0.0268	0.0279	0.0213	0.0260	0.0297	0.0299
15	M	0.0256	0.0122	0.0006	0.0242	0.0227	0.0085	0.0138	0.0061	0.0150
15	E	0.0451	0.0395	0.0297	0.0447	0.0438	0.0376	0.0650	0.0395	0.0428
15	E+M	0.0354	0.0243	0.0234	0.0354	0.0364	0.0317	0.0274	0.0339	0.0360
15	E^2	0.0470	0.0466	0.0462	0.0470	0.0479	0.0453	0.0538	0.0472	0.0480
15	$\mathrm{E}^2\mathrm{M}$	0.0332	0.0303	0.0302	0.0350	0.0355	0.0305	0.0311	0.0374	0.0370
20	M	0.0333	0.0140	0.0002	0.0277	0.0239	0.0091	0.0130	0.0024	0.0088
20	E	0.0469	0.0430	0.0365	0.0475	0.0468	0.0394	0.0640	0.0433	0.0458
20	E+M	0.0395	0.0308	0.0283	0.0386	0.0392	0.0337	0.0317	0.0369	0.0378
20	E^2	0.0488	0.0477	0.0492	0.0482	0.0490	0.0465	0.0544	0.0491	0.0491
20	$\mathrm{E}^2\mathrm{M}$	0.0319	0.0331	0.0329	0.0339	0.0365	0.0341	0.0324	0.0381	0.0388
25	M	0.0335	0.0124	0.0001	0.0136	0.0214	0.0089	0.0131	0.0009	0.0032
25	E	0.0483	0.0449	0.0403	0.0486	0.0479	0.0405	0.0636	0.0449	0.0469
25	E+M	0.0385	0.0324	0.0313	0.0403	0.0403	0.0314	0.0343	0.0402	0.0398
25	E^2	0.0492	0.0479	0.0496	0.0490	0.0497	0.0482	0.0540	0.0494	0.0494
25	$\mathrm{E}^2\mathrm{M}$	0.0359	0.0353	0.0343	0.0360	0.0362	0.0374	0.0349	0.0381	0.0379

TABLE 11: Mean test size for the 10385 values of $\boldsymbol{p}$ in $\mathcal{P}_0$, for all the test statistics and M, E, E+M, E^2 and $\mathrm{E}^2\mathrm{M}$ p-values when the chosen significance level is $\alpha = 0.05$. The top row denotes the test statistics, LRT is given in (21), LAP in (23), uDT and rDT in (22) with unrestricted and restricted maximum likelihood estimates inserted for $\boldsymbol{p}$ respectively, π_{M} in (26), π_{E} in (25), π_{e} in (24), PP_1 in (15) and PP_3 in (20). N is the sample size.

likelihood ratio, unrestricted difference and restricted difference test statistics. In general, for all types of p-values, T_{LR} and T_{rDT} have the largest test size, T_{uDT} and T_{LAP} have the smallest test size and T_{uDT} has mostly smaller test size than T_{LAP}. T_{LR} has the largest test size for the M p-values. When $N = 20$, T_{uDT} has the largest test size for the E^2 p-values, a result for which we have found no apparent reason.

If we look into the performance of $T_{\pi_{\text{e}}}$, $T_{\pi_{\text{E}}}$ and $T_{\pi_{\text{M}}}$ we see that $T_{\pi_{\text{e}}}$ has the largest test size compared to all the other test statistics for the E and E^2 p-values for all N. $T_{\pi_{\text{M}}}$ has the largest test size for the M, E+M and E^2M p-values for $N = 10$, for the E+M and E^2M p-values when $N = 15$ and for the E^2M p-values when $N = 20$, whereas $T_{\pi_{\text{E}}}$ has the largest test size for the E^2M p-values when $N = 25$. We note that when N increases the likelihood ratio or restricted difference test performs better than the π_{M} statistic for the M, E+M and E^2M p-values, therefore the π_{M} test statistic is probably a better choice only when N is small.

What is worth noting, is that for the test statistics that are most conservative with respect to test size for the M p-values, the gain is greater when performing one or more E step(s) before the M step compared to the test statistics for which the test size for the M p-values is less conservative. This is particularly evident for T_{uDT}, T_{LAP} and $T_{\pi_{\text{E}}}$ compared to T_{LR}. The test size for T_{LR} increases less than the test size for the other three test statistics when comparing the M and E+M p-values. For T_{LR} the test size is also reduced if two E steps instead of one are applied before the M step, whereas for T_{LAP} and T_{uDT} the test size increases when two E steps are applied before the M step.

The mean test size increases when N increases. Comparing the test sizes for $N = 10$ to the test sizes for $N = 25$ reveals an increase for all test statistics and type of p-values except for some of the M p-values which have test size that is approximately 0. As an illustration, the mean test size for the M p-value for the likelihood ratio test statistic is 0.0164 when $N = 10$ and 0.0335 when $N = 25$.

In addition to the test statistics discussed so far, the Bayesian prior predictive p-values $P_{\text{PP},1}$ and $P_{\text{PP},3}$, originated from using the same prior $\pi_1(\boldsymbol{p})$, but different formulations of the null hypothesis, were used as test statistics to compute E, M, E+M, E^2 and E^2M p-values. When $N = 10$, the PP_3 test statistic yields larger test size than all the other test statistics for the M, E+M and E^2M p-values. Otherwise the test size of these two test statistics lies between the test size of the other test statistics, not following a clear pattern, except that the PP_3 yields larger test size than the PP_1 in general.

We also evaluated the performance of all the test statistics, values of N and types of p-values for the six multinomial cases of Table 10. Table 12 shows the test size for T_{LR} for $N = 10$ and $N = 25$. We see that the E and E^2 p-values yield a test size greater than 0.05 in case 1–5 for $N = 25$ and thus proves that these p-values are not valid. We also see that the test size is greater when $N = 25$ compared to $N = 10$. The results for the other test statistics and values of N are omitted in this report since the findings in respect to test statistics and p-values in the six specific multinomial cases were similar to the overall findings, however the test size for all test statistics was clearly dependent of the chosen multinomial cases, i.e. the parameter $\boldsymbol{p}$ in the multinomial distribution. In general, which can also be seen in Table 12, case 1 and 2 have larger test size than case 3–6. This trend was consistent through the different test statistics, types of p-values and N and indicates that when comparing test sizes the multinomial case chosen will have a large influence the test size, but it will not change the conclusions with respect to which test statistic or which p-value results in the largest or smallest test size.

Figure 3 shows histograms for the test size in the 10385 cases under H_0 for the M, E, E+M, E^2 and E^2M p-values for each of the test statistics T_{LAP}, T_{LR}, T_{uDT}, T_{rDT}, $T_{\pi_{\text{M}}}$, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{e}}}$ and $T_{\text{PP},1}$ for $N = 10$. We see that for T_{LR}, T_{rDT}, $T_{\pi_{\text{E}}}$, $T_{\pi_{\text{M}}}$,and $T_{\text{PP},1}$ the distribution of the test size for the E

N	p-value	case 1	case 2	case 3	case 4	case 5	case 6
10	M	0.0199	0.0193	0.0097	0.0071	0.0061	0.0035
10	E	0.0412	0.0426	0.0281	0.0218	0.0197	0.0122
10	EM	0.0281	0.0366	0.0242	0.0193	0.0170	0.0111
10	E^2	0.0475	0.0500	0.0379	0.0302	0.0333	0.0220
10	E^2M	0.0322	0.0294	0.0200	0.0157	0.0166	0.0101
25	M	0.0431	0.0428	0.0386	0.0391	0.0313	0.0287
25	E	0.0528	0.0529	0.0573	0.0563	0.0495	0.0441
25	EM	0.0431	0.0435	0.0458	0.0448	0.0389	0.0339
25	E^2	0.0510	0.0514	0.0573	0.0569	0.0523	0.0469
25	E^2M	0.0401	0.0387	0.0415	0.0404	0.0367	0.0322

TABLE 12: Test size for the likelihood ratio test statistic for the six multinomical cases.

p-values is skewed towards the right compared to the distribution for the M p-values, and we note that the test size is sometimes larger than 0.05, showing that the E p-values are not valid. The E+M p-values preserve the skewed distribution while shifting it to the left so that no test size is greater than 0.05. For the LAP and uDT test statistics, we note that the distribution of test size for the E p-values is not skewed in the same way, but the E^2 p-values are, so apparently it is necessary to do two E steps before maximization for the LAP and uDT statistics.

Figure 3 illustrates what happens under the E and M steps. To obtain an even better understanding of the effect of the E and M steps, we consider two possible outcomes when $N = 10$, $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$. Table 13 shows the p-values for these outcomes using the likelihood ratio and LAP test statistics.

Outcome	Test statistic	M	E	E+M	E^2	E^2M
$\boldsymbol{y}_1$	$T_{\text{LRT}} = 4.159$	0.1025	0.0940	0.1108	0.0770	0.1062
$\boldsymbol{y}_2$	$T_{\text{LRT}} = 4.077$	0.1025	0.0349	0.0450	0.0279	0.0408
$\boldsymbol{y}_1$	$T_{\text{LAP}} = 3.932$	0.1048	0.0805	0.1056	0.0768	0.1064
$\boldsymbol{y}_2$	$T_{\text{LAP}} = 2.492$	0.2297	0.0978	0.1329	0.0654	0.0855

TABLE 13: P-values for the likelihood ratio and LAP test statistics for the outcomes $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$.

Let us first consider the p-values for the likelihood ratio test statistic. We note that for $\boldsymbol{y}_2$ with a 5% significance level, we would reject the null hypothesis based on the E p-value and not reject it based on the M p-value. Since the likelihood ratio test statistic is greater for $\boldsymbol{y}_1$ than for $\boldsymbol{y}_2$, the M p-value for $\boldsymbol{y}_2$ will necessarily be greater than for $\boldsymbol{y}_1$, which will be greater than the E p-value for $\boldsymbol{y}_1$. Since the E p-value is less for $\boldsymbol{y}_2$ than for $\boldsymbol{y}_1$, $\boldsymbol{y}_1$ will not be in the tail set for $\boldsymbol{y}_2$ when performing the M step after the E step and the E+M p-value results in rejection of the null hypothesis on a 5% significance level for $\boldsymbol{y}_2$. Since the E and E^2 steps alone do not result in valid p-values, we should perform an M step afterwards. But as we see, the E step(s) are means to avoid certain outcomes having a large p-value because of other outcomes having greater test statistics and artifically large E and M p-values compared to other outcomes with similar magnitude of the test statistics.

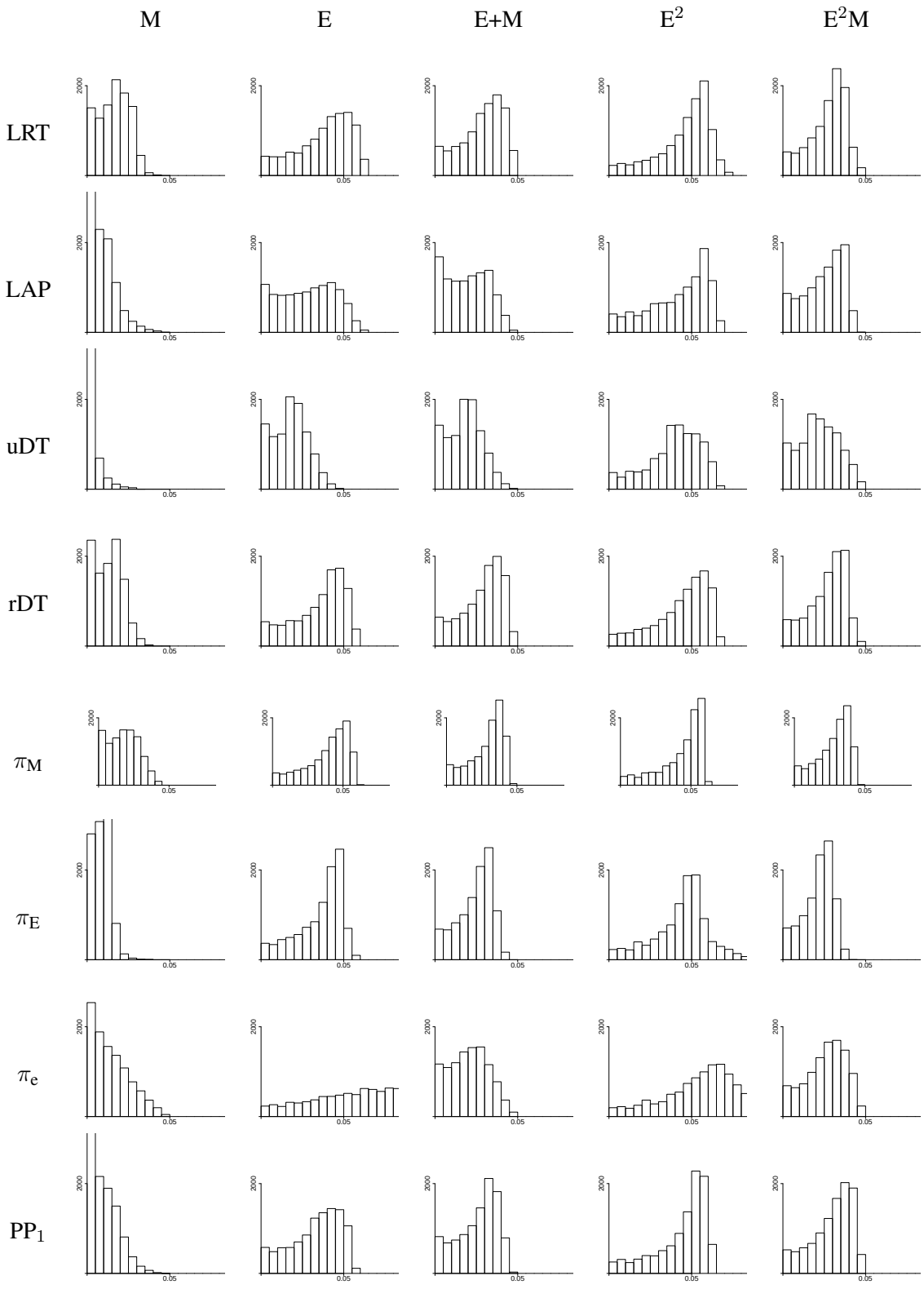


FIGURE 3: Distribution of test size for the various test statistics and p-values, $N = 10$, the x-axis is cut at 0.08 and the y-axis at 3000.

For the LAP test statistic, we do not see the same effect, even though the p-value with the smallest test statistic, $\boldsymbol{y}_2$, has an M p-value greater than the M p-value for $\boldsymbol{y}_1$. Here the E p-value for $\boldsymbol{y}_2$ is also greater than the one for $\boldsymbol{y}_1$, and thus this ordering is preserved when performing an M step after the E step. The E^2 p-value however, is smaller for $\boldsymbol{y}_2$ than $\boldsymbol{y}_1$, and performing the M step afterwards does not change this.

Here $\boldsymbol{y}_1$ is an example of an outcome for which the decision of rejecting the null hypothesis does not only depend on the type of p-value, but also of the chosen test statistic. Using the likelihood ratio test statistic, we reject the null hypothesis on a 5% confidence level for all of the p-values E, E+M, E^2 and E^2M. If we use the LAP test statistic instead, we do not reject it for any of the p-values.

Table 14 shows the mean test size for the Bayesian prior predictive p-values for $N = 10, 15, 20, 25$ using a grid with 50 points in each direction for both formulations of H_0, i.e. $f_{\text{PPV}}(\boldsymbol{p})$ and $f^*_{\text{PPV}}(\boldsymbol{p})$ and both priors for $\boldsymbol{p}$. We see that the test size depends highly on the choice of prior and formulation of H_0. The test size is smallest using the uniform Dirichlet prior and $f^*_{\text{PPV}}(\boldsymbol{p}) = 0$ as H_0, it increases to around 0.055 with $f_{\text{PPV}}(\boldsymbol{p}) = 0$ as H_0 and if we choose the non-uniform Dirichlet prior $\pi_2(\boldsymbol{p})$, the test size becomes very high. Clearly, the non-uniform prior is not a good choice and it also indicates that the choice of prior has a larger effect than how we choose to formulate the null hypothesis. Comparing these results to the results when using the prior predictive p-values as test statistics to define the tail sets for the M, E, E+M and E^2M p-values in Table 11 shows that the M step reduces the test size in all cases for all values of N and for both formulations of H_0 as expected. The test size for the E^2 p-values is higher than for the Bayesian p-values in many of the cases and in some cases, e.g. case 5 for $N = 15, 20, 25$ for $H_0\colon f_{\text{PPV}}(\boldsymbol{p}) = 0$ and for $N = 20, 25$ for $H_0\colon f^*_{\text{PPV}}(\boldsymbol{p}) = 0$, the test size increases for all the p-values except the M p-values. Table 14 also shows that the Bayesian prior predictive p-values are not valid since the test size is larger than the significance level.

N	PP_1	PP_2	PP_3
10	0.0561	0.1272	0.0489
15	0.0557	0.1465	0.0491
20	0.0556	0.1533	0.0494
25	0.0552	0.1556	0.0494

TABLE 14: Test size for the Bayesian positive predictive p-values using different priors, formulation of H_0 and values of N, PP_1 is given in (15), PP_2 is given in (17) and PP_3 is given in (20).

The prior predictive p-values for the two outcomes $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$ are given in Table 15. We see that the three Bayesian p-values are quite different for both outcomes. For $\boldsymbol{y}_2$ we reject the null hypothesis, whereas for $\boldsymbol{y}_1$ we do not reject the null hypothesis. The two p-values both found from the model with uniform Dirichlet prior are similar for $\boldsymbol{y}_2$, but for $\boldsymbol{y}_1$ it is the two p-values that are based on the same formulation of H_0 that are similar. The null hypothesis is rejected for $\boldsymbol{y}_2$, but not for $\boldsymbol{y}_1$ for any of the p-values.

4.2.2 EVALUATION OF TEST POWER

We would like to compare the test power of T_{LR}, T_{LAP}, T_{uDT}, T_{rDT} and $T_{\pi_{\text{M}}}$. Since the results of the test size comparisons showed that the M p-values have the smallest test sizes and since the E and E^2 p-values are not valid, we consider only the E+M and E^2M p-values when comparing test power. We

Outcome	PP_1	PP_2	PP_3
$\boldsymbol{y}_1$	0.1086	0.1084	0.0506
$\boldsymbol{y}_2$	0.0382	0.0171	0.0323

TABLE 15: Bayesian prior predictive p-values for the outcomes $\boldsymbol{y}_1 = (1, 3, 0, 6, 0, 0)$ and $\boldsymbol{y}_2 = (1, 0, 1, 3, 5, 0)$.

expect that the power increases with N and we used $N = 10$ and $N = 25$ to investigate the magnitude of the increase. The power was calculated the same way as the test size except that the values of $\boldsymbol{p}$ are chosen so that $\boldsymbol{p}$ does not satisfy the null hypothesis (9).

We wanted to compare the test power in specific multinomial cases and we chose six sets of the parameters $\boldsymbol{p}$, these were denoted case 7–12 and are given in Table 16. They were chosen because of their decreasing distance from H_0 which is measured by the magnitude of $f_{\text{PPV}}(\boldsymbol{p})$. If $f_{\text{PPV}}(\boldsymbol{p})$ is close to 0, then $\boldsymbol{p}$ nearly satisfies H_0 while the greater $|f_{\text{PPV}}(\boldsymbol{p})|$ is, the further away from H_0 $\boldsymbol{p}$ is. Since the power in our chosen cases may not be representative for a randomly chosen case, we also generate 10385 random cases under H_1, by drawing 10385 vectors of length 6 from the uniform distribution and scaling each vector to sum to 1.

Case	p_1	p_2	p_3	p_4	p_5	p_6	$f_{\text{PPV}}(\boldsymbol{p})$
7	0.06	0.01	0.44	0.26	0.22	0.01	0.52
8	0.01	0.10	0.44	0.01	0.43	0.01	0.76
9	0.20	0.05	0.24	0.28	0.22	0.01	0.27
10	0.01	0.07	0.27	0.28	0.26	0.11	0.29
11	0.06	0.12	0.18	0.14	0.35	0.15	0.18
12	0.17	0.12	0.18	0.21	0.16	0.16	0.05

TABLE 16: Specification of cases under H_1.

Table 17 and 18 shows the test power for the chosen cases, test statistics and p-values when $N = 10$ and $N = 25$ respectively. As expected the power increases when N increases. The test statistics T_{uDT} and T_{LAP} have the smallest power except for the E^2M p-values in case 6 when $N = 25$. When $N = 25$ the $T_{\pi_{\text{M}}}$ statistic has the highest power except in case 6 for the E^2M p-values. For $N = 10$, the T_{LR} statistic has highest power for the E+M p-values in four of six cases, while only in one case for the E^2M p-values. The E+M p-values yields in general higher power than the E^2M p-values, except for the T_{LAP} and T_{uDT} statistics when $N = 10$. If we compare these results to the calculated mean power for all the power cases, given in the last column of Table 17 and 18, we see that T_{LAP} and T_{uDT} have smaller power for the E+M than the E^2M p-values when $N = 10$ and also when $N = 25$ for T_{uDT}. T_{LR} has the largest power for the E+M p-values when $N = 10$, otherwise the π_{M} has the largest test power.

When comparing the power in each of the six cases by considering the value of $f_{\text{PPV}}(\boldsymbol{p})$ in Table 16 we see that the power seems to decrease when $f_{\text{PPV}}(\boldsymbol{p})$ decreases which we would expect since in cases that are far from H_0 the test should have higher power than in cases closer to H_0. However, in case 7 and 8 $f_{\text{PPV}}(\boldsymbol{p})$ is 0.52 and 0.76 respectively and yet case 7 has the highest power, particularly

Test statistic	p-value	case 7	case 8	case 9	case 10	case 11	case 12	mean
T_{LRT}	E+M	0.8344	0.7210	0.4125	0.2332	0.1161	0.0547	0.1064
T_{LAP}	E+M	0.7165	0.7203	0.2909	0.1863	0.0801	0.0360	0.0810
T_{uDT}	E+M	0.6997	0.3815	0.3269	0.1446	0.0556	0.0345	0.0697
T_{rDT}	E+M	0.8372	0.7173	0.4096	0.2286	0.1108	0.0529	0.1044
$T_{\pi_{\text{M}}}$	E+M	0.8271	0.7568	0.3996	0.2119	0.1080	0.0472	0.1016
T_{LRT}	E^2M	0.8219	0.7159	0.4140	0.2020	0.1006	0.0463	0.0967
T_{LAP}	E^2M	0.7492	0.7044	0.3498	0.1857	0.0848	0.0416	0.0881
T_{uDT}	E^2M	0.7640	0.3903	0.3712	0.1923	0.0755	0.0442	0.0844
T_{rDT}	E^2M	0.8240	0.7160	0.4081	0.2162	0.1028	0.0476	0.0977
$T_{\pi_{\text{M}}}$	E^2M	0.8274	0.7500	0.4112	0.2031	0.1060	0.0467	0.0994

TABLE 17: Test power for the E and E^2M p-values in case 7–12 and mean over 10385 cases for $N = 10$.

Test statistic	p-value	case 7	case 8	case 9	case 10	case 11	case 12	mean
T_{LRT}	E+M	0.9979	0.9967	0.8014	0.4974	0.1936	0.0537	0.2163
T_{LAP}	E+M	0.9967	0.9935	0.7897	0.4713	0.1686	0.0509	0.2038
T_{uDT}	E+M	0.9970	0.9592	0.8056	0.4788	0.1699	0.0547	0.2075
T_{rDT}	E+M	0.9985	0.9976	0.8179	0.5275	0.2193	0.0585	0.2241
$T_{\pi_{\text{M}}}$	E+M	0.9987	0.9978	0.8296	0.5378	0.2257	0.0586	0.2242
T_{LRT}	E^2M	0.9976	0.9963	0.7908	0.4793	0.1835	0.0499	0.2080
T_{LAP}	E^2M	0.9961	0.9935	0.7866	0.4609	0.1691	0.0515	0.2063
T_{uDT}	E^2M	0.9969	0.9622	0.7874	0.4660	0.1704	0.0504	0.2068
T_{rDT}	E^2M	0.9980	0.9969	0.7938	0.5017	0.2005	0.0501	0.2087
$T_{\pi_{\text{M}}}$	E^2M	0.9983	0.9969	0.8091	0.5080	0.2070	0.0510	0.2101

TABLE 18: Test power for the E and E^2M p-values in case 7–12 and mean over 10385 cases for $N = 25$.

when $N = 10$. We see the same in case 9 and 10, $f_{\text{PPV}}(\boldsymbol{p})$ is then 0.27 and 0.29, and the power in case 9 is a lot higher than in case 10.

The mean value of $|f_{\text{PPV}}(\boldsymbol{p})|$ for the 10385 cases is 0.13, which explains the small overall power, since the mean value is not as far from H_0 as e.g. case 7 or 8. The mean power is comparable to case 11 where the distance from H_0 is 0.18.

It is not surprising that the likelihood ratio, restricted difference and π_M test statistics perform similarly, considering they are all functions of the maximum likelihood estimates for $\boldsymbol{p}$ under H_0, $\tilde{\boldsymbol{p}}$. The LAP and unrestricted difference test statistics however, do not depend on these estimates and this can be the reason their performance is poorer.

5 COMPUTATIONAL DETAILS

To compute the integral (14), we used the midpoint rule on a 4-dimensional grid. The four dimensions correspond to p_1, p_2, p_3 and p_5. Each side in the grid is divided into a number of subintervals of equal length, and the midpoint in each subinterval is calculated. For each point (p_1, p_2, p_3, p_5) in the grid, we set p_4,

$$p_4 = \frac{p_1(1 - p_1 - 2p_2 - p_3 - 2p_5) + p_2(1 - p_2 - p_3 - p_5) - p_3p_5}{p_1 + p_3}$$

which is derived from (13).

If $0 < p_4 < 1$, we set $p_6 = 1 - \sum_{i=1}^{5} p_i$ and if $0 < p_6 < 1$ then the value

$$N! \left(\prod_{i=1}^{6} \frac{p_i^{y_i}}{y_i!} \right) \frac{(p_1 + p_2 + p_4 + p_5)^2}{p_1 + p_2}, \tag{28}$$

which is $\pi(\boldsymbol{y}|\boldsymbol{p})$ multiplied with the non-normalized density of $\boldsymbol{p}$, is added to the present value of the integral. If either p_4 or p_6 are less than 0 or greater than 1, the current point in the grid is discarded. The total non-normalized integral is the sum of (28) over the $\boldsymbol{p}$'s satisfying the constraints for p_4 and p_6. The integrals (16) and (19) are computed similarly.

The number of points in the grid has to be chosen and in the results presented in this report, a grid where each side is divided into 50 subintervals was used. This resulted in 79876 points after discarding those with p_4 or p_6 outside [0,1]. Table 19 shows the test size for the Bayesian prior predictive values using the uniform Dirichlet prior and original formulation of H_0 (9) for the six values of $\boldsymbol{p}$ given in Table 10 and $N = 10$ when the number of subintervals, n_{int}, on each side in the grid is 30, 35, 40, 45 and 50. We see that the test size varies with the grid size to some extent, the largest difference is in case 1 between $n_{\text{int}} = 45$ and $n_{\text{int}} = 50$.

n_{int}	case 1	case 2	case 3	case 4	case 5	case 6
30	0.0395	0.0642	0.0401	0.0365	0.0186	0.0157
35	0.0389	0.0632	0.0405	0.0367	0.0191	0.0159
40	0.0384	0.0620	0.0405	0.0367	0.0191	0.0159
45	0.0384	0.0620	0.0405	0.0367	0.0191	0.0159
50	0.0407	0.0625	0.0408	0.0372	0.0193	0.0162

TABLE 19: Test size in case 1–6 for the Bayesian prior predictive p-value in (15) for $N = 10$ using different grid sizes, n_{int} is the number of sub intervals on each of the four sides in the grid.

A grid for $\boldsymbol{p}$ is also needed for the p-values that include a maximization step, i.e. the M, E+M and E^{2}M p-values. In the positive predictive value setting, we used the same grid as for the Bayesian prior predictive p-values with 50 possible subintervals for each of the four sides in the grid, but in addition we included the maximum likelihood estimates $\tilde{\boldsymbol{p}}$ of $\boldsymbol{p}$ under H_0 for all possible outcomes given N. Table 20 shows the number of possible outcomes in the positive predictive value situation for a given value of N and the size of the grid with the maximum likelihood estimates included. Thus, the size of this grid increased with N, for $N = 10$, the grid consisted of $3003 + 79876 = 82879$ points and when $N = 25$, it consisted of $142506 + 79876 = 222382$ points. Comparisons of the test size for different grid sizes showed that the grid did not have a great influence on the test size when the

N	Number of outcomes	Size of grid
10	3003	82879
15	15504	95380
20	53130	133006
25	142506	222382

TABLE 20: Number of possible outcomes given N and size of grid used when calculating p-values in the problem of comparing positive predictive values.

grid is used for maximization. We also investigated how often the maximum p-value was obtained in one of maximum likelihood points compared to the other points. The percentage increased with N and decreased with the size of the grid without maximum likelihood estimates. When N increases the number of maximum likelihood estimates increases, and it is not surprising that more of these points will give the maximum p-value and similarly, when the number of grid points in the grid without maximum likelihood estimates increases, more of the points that are not maximum likelihood estimates will give the maximum p-value.

The p-value computations for a sequence of E and M steps are quite computer intensive, as p-values for all outcomes (except in the last step), not only the one of interest in a specific study, must be computed for further use as a test statistic in the next step. The test statistic giving the original ordering of outcomes, e.g. the likelihood ratio test statistic, should be computed only once, as should the maximum likelihood estimates of $\boldsymbol{p}$ under the null hypothesis. The grid used for the numerical maximization in the M step and for calculation of the π_{M} statistic was also calculated in advance.

In both the E and the M step, the outcomes should be sorted according to the test statistic (original test statistic or negative output of a previous E or M step). In the E step, the p-values are then accumulated, starting with the probability of the outcome having the most extreme value of the test statistic, and the probabilities (with the maximum likelihood estimates of $\boldsymbol{p}$ under the null hypothesis of the outcome of interest as parameters) of the forthcoming outcomes successively being added until the outcome of interest is reached. Special care must be taken to include possible outcomes having an equal test statistic value ("draws"), and because of possible numerical inaccuracies also a threshold for when two values are counted as equal should be specified. In order to compute all possible p-values, this should be repeated for all outcomes – we have chosen to accumulate probabilities for all outcomes in parallel. Taking care when dealing with draws also applies to the M step and calculation of the π_{E} and π_{M} test statistics.

In the M step we accumulated probabilities given by the grid points as parameters in parallel while going through the sorted outcomes. As the number of grid points times the number of outcomes may be huge, only the accumulated probabilities for each outcome were saved, and for each outcome reached, the maximum of the accumulated probabilities were saved as the p-value of that outcome.

Calculation of the π_{E} and π_{M} test statistics, based on the probabilities of the outcomes themselves instead of on an external test statistic, are more computer intensive, as the ordering of the outcomes is specific for each outcome of interest, and not to a given test statistic. For π_{E}, the p-value for an outcome of interest is found by adding probabilities of all outcomes having a probability that is not greater, using the maximum likelihood estimate of $\boldsymbol{p}$ under the null hypothesis of the outcome of interest as parameters, thus the probability of each outcome has to be calculated for each outcome of interest. We found some gain in computation speed by sorting the outcomes before adding.

N	E	M	π_{e}	π_{E}	π_{M}
10	0m0.33s	0m17.69s	0m0.06s	0m2.84s	1m29.94s
25	16m17.13s	14m4.51s	0m0.96s	155m13.68s	101m20.51s
10	0m0.34s	0m18.65s	0m0.01s	0m2.86s	1m33.01s
25	15m51.13s	39m17.33s	0m0.97s	156m40.3s	262m6.08s

TABLE 21: Running time for E and M p-values and for calculating the test statistics $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi\mathrm{M}}$ in the positive predictive values setting for samples sizes $N = 10, 25$ (3003 and 142506 outcomes, respectively). The two upper rows show the time when using a grid with $n_{\text{int}} = 50$ without maximum likelihood estimates and the time in the two lower rows is the time when using the grid with $n_{\text{int}} = 50$ including maximum likelihood estimates (79876 points without estimates, 82879 including estimates for $N = 10$, and 222382 points including estimates for $N = 25$).

For π_{M}, the grid points rather than the outcomes were gone through in an outer loop. For each grid point, the probability of each outcome was calculated, the outcomes sorted accordingly, and probabilities accumulated from the smallest to the greatest. If the cumulative probability of an outcome was greater than some earlier maximum for that outcome, the maximum was replaced by the current sum.

In contrast, calculation of π_{e} is trivial, this is simply the probability of an outcome taking its maximum likelihood estimate of $\boldsymbol{p}$ under the null hypothesis as the parameter vector.

Power and size calculations for a given parameter vector are simply a matter of adding probabilities of outcomes having p-values not exceeding the significance level (in our case 0.05).

The code was written in C++, implemented in GCC and the calculations were performed with the Standard Template Library, using one of eight processors on a Dell PowerEdge 2950 with two Quad-core Xeon X5365 3.0 GHz processors, 4 MB cache, 16 GB RAM. The running time for calculating E and M p-values for any test statistic, along with the running time for calculating the π_{e}, π_{E} and π_{M} test statistics when comparing positive predictive values for $N = 10$ and $N = 25$ are given in Table 21 for the grid with $n_{\text{int}} = 50$, without and with the maximum likelihood estimates of $\boldsymbol{p}$ included. When $N = 10$, all the calculations are performed rather fast, except calculating the values of the π_{M} test statistic which takes one and a half minute. When N increases, the running time naturally increases severely since all calculations must be performed for all possible outcomes. We note that calculating the π_{E} test statistic takes longer than calculating the π_{M} test statistic when $N = 25$ for the grid without maximum likelihood estimates. This is because the number of possible outcomes is less than the number of grid points in this case. If the number of grid points is larger than the number of outcomes, as in the grid where the maximum likelihood estimates are included, calculating the π_{M} statistic takes much longer than calculating the π_{E} statistic.

6 DISCUSSION

The enumeration idea is not new as it goes back to Fisher (1935), but it has often been overlooked. We have demonstrated how to apply the idea for testing independent binomial proportions and comparing positive predictive values. Another recent application of the idea is in genome-wide association studies, in which single nucleotide polymorphisms (SNP) across the human genome are studied. When the mode of inheritance is unknown, the MAX test statistic, which is the maximum of the three Cochrane–

Armitage trend statistics for dominant, recessive and additive inheritance modes, see Freidlin, Zheng, Li and Gastwirth (2002), tests the association between the genotype and phenotype. The exact distribution of the MAX test statistic is unknown and calculating p-values based on proposed asymptotic distributions involves numerical integration. Another common approach is to use permutations tests, but both solutions leads to possible random errors in the calculated p-values. Moldovan, Langaas and Bahlo (2009) instead calculate exact p-values using the enumeration approach and thereby avoid this uncertainty.

When the sample size increases and enumeration will be too time consuming, the parametric bootstrapping approach can be used instead. Günther et al. (2009) used parametric bootstrapping to approximate the distribution of the likelihood ratio, LAP and restricted and unrestricted difference test statistics. The p-values obtained from this distribution are approximately the same as the E p-values we find by enumeration in this report, and the parametric bootstrap approach involving simulated outcomes is actually a numerical approximation that calculates the tail without using enumeration. This is seen if the test size for case 1–6 in Table 12 is compared to the test size for the small sample parametric bootstrap likelihood ratio test in Table 3 of Günther et al. (2009) – the values are almost the same. It may be of use for larger sample sizes when calculating maximum likelihood estimates and p-values for the bootstrap samples is less time consuming than calculating the maximum likelihood estimates and p-values for all possible outcomes. When using the formulas for calculating exact test size and power, i.e., (5) and (6), drawing outcomes from the multinomial distribution under H_0 or H_1 and estimating the test size or power by the proportion of these outcomes having p-values less than or equal to the significance level as was done in Günther et al. (2009) is not necessary, and therefore the uncertainty in the estimates are removed. This is however, only possible when the sample size is small enough so that the p-values for all possible outcomes can be calculated.

Another option when the sample size increases is to condition on sums of $N_i,\ i = 1, \ldots, 6$, which in a contingency table setting corresponds to conditioning on the marginals. This reduces the number of possible outcomes and makes it possible to use exact tests for higher values of N. The usability of this approach depends on the actual problem. In the example from Lloyd (2008), n_1 and n_2 are fixed as the number of subjects who receives treatment and placebo respectively. In the setting of positive predictive values, it is not clear which values that should be fixed. It could be the number of diseased and non-diseased subjects, if the disease status is decided before the two tests are applied, or it could be the number of subjects with positive test A, positive test B and positive tests A and B, but in practise, these numbers will usually not be fixed in advance.

As Table 12 showed, the test size of a test statistic for any p-value depends on the chosen value of $\boldsymbol{p}$, the parameter in the multinomial distribution. When the chosen significance level is 0.05, some cases have test size close to 0.05, whereas other cases have smaller test sizes. A further investigation reveals what the cases for which the test size is close to 0.05 have in common. Assume the outcomes are sorted by decreasing value of some chosen test statistic. The M step will result in rejection of the null hypothesis for outcomes that are above a certain limit, where the limit is the p-value closest to 0.05 (but not greater than 0.05). The null hypothesis is not rejected for any of the outcomes below the limit. Assume that the last outcome for which H_0 is rejected, $\boldsymbol{y}_0$ has a maximum tail probability $P_{\text{M},0}$, i.e. p-value, in the point $\boldsymbol{p}_0$. If the true value of $\boldsymbol{p}$ is in fact $\boldsymbol{p}_0$, then the probability of rejecting H_0 is the sum of the probabilities of this outcome and the outcomes above, which is $P_{\text{M},0}$. Thus a test size of almost 0.05 is always obtained for a particular $\boldsymbol{p}$, it is only the discreteness that prevents it from exactly being obtained for a specific value of $\boldsymbol{p}$. This value is the value of $\boldsymbol{p}$ that maximizes the p-value for the outcome that has the largest p-value less than or equal to 0.05. If one wants to report

the test size in a certain multinomial case, choosing this value of $\boldsymbol{p}$ will ensure that the test size is close to 0.05 unlike the six multinomial cases we chose.

7 CONCLUSIONS

In this work we have provided an in-depth effort of using enumeration and exact p-values to address the problem of comparing positive predictive values. The existing tests for this situation rely on asymptotic distributions and have previously been shown not to preserve the test size when the sample size was moderate. The test size and power of nine test statistics in combination with five types of p-values have been thoroughly evaluated for different sample sizes. As demonstrated, the M step yields valid p-values, although these are often conservative. The E step provides a reordering of the reference set in contrast to the M step and one or two E steps before the M step increases the test size while yielding valid p-values.

We have presented three new test statistics, $T_{\pi_{\mathrm{e}}}$, $T_{\pi_{\mathrm{E}}}$ and $T_{\pi_{\mathrm{M}}}$, that can be applied to any problem. In the problem of comparing binomial proportions, the π_{e} test statistic performed better than the test statistics analyzed by Lloyd (2008) in terms of test size and power for the E+M p-values.

For comparing the positive predictive values from two diagnostic tests, we recommend using either the likelihood ratio, restricted difference or π_{M} test statistic and to calculate the E+M p-values. These p-values are valid, and for these test statistics the results have indicated that there is no need to do more than one E step before the final M step. However, the importance of one or more E steps before maximization is greater for e.g. the LAP and unrestricted difference test than for the likelihood ratio test as it increases the test size more significantly, suggesting that the ordering provided by the LAP and unrestricted difference test is not optimal with respect to test size and power.

We do not recommend using the prior predictive p-values, as these are very sensitive to the choice of prior and on the null hypothesis formulation.

This report gives further general insight into the mechanisms behind the E, M and E+M p-values in general and in the example discussed by Lloyd (2008). We describe how the E p-value changes the ordering of outcomes and why this reduces the conservativeness of the M p-values if the E p-values are applied before the M step.

In further work, it would be of interest to find a test statistic that in some sense provides an optimal ordering of the outcomes with respect to test size and power and in particular, the π_{e}, π_{E} and π_{M} should be studied in greater detail and compared to other test statistics. We would also like to investigate if ordering of the outcomes converges after a certain number of E steps, and also the effect of performing two or more consecutive sequences of the form $\mathrm{E}^k\mathrm{M}$.

REFERENCES

Agresti, A. (2002). *Categorical data analysis*, second edn, John Wiley & Sons, Inc., Hoboken, NJ, chapter 1.4.4.

Bayarri, M. J. and Berger, J. O. (2000). P values for composite null models, *Journal of the American Statistical Association* 95(452): 1127–1142.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter, *Journal of the American Statistical Association* 89(427): 1012–1016.

Bickel, J. and Doksum, K. A. (2001). *Mathematical statistics*, second edn, Prentice Hall, Inc., chapter 4.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, second edn, Duxbury, chapter 8.

Fisher, R. A. (1935). The logic of inductive inference, *Journal of the Royal Statistical Society* 98: 39–82.

Freidlin, B., Zheng, G., Li, Z. and Gastwirth, J. L. (2002). Trend tests for case-control studies of genetic markers: power, sample size and robustness, *Human Heredity* 53: 146–152.

Günther, C.-C., Bakke, Ø. and Langaas, M. (2009). Comparing positive predictive values for small samples with application to gene ontology testing. Preprint Statistics No. 3, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Günther, C.-C., Bakke, Ø., Lydersen, S. and Langaas, M. (2008). Comparison of predictive values from two diagnostic tests in large samples. Preprint Statistics No. 9, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Johnson, N. L., Kotz, S. and Balakrishan, N. (1997). *Discrete multivariate distributions*, Wiley series in probability and statistics, chapter 35.

Leisenring, W., Alonzo, T. and Pepe, M. S. (2000). Comparisons of predictive values of binary medical diagnostic tests for paired designs, *Biometrics* 56: 345–351.

Lloyd, C. J. (2008). Exact p-values for discrete models obtained by estimation and maximization, *Australian & New Zealand Journal of Statistics* 50(4): 329–345.

Moldovan, M., Langaas, M. and Bahlo, M. (2009). Efficient error-free computation of MAX p-values with an application to genome-wide association studies. Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, submitted.

Zelterman, D., Chan, I. S.-F. and Mielke, P. W. (1995). Exact tests of significance in higher dimensional tables, *The American Statistician* 49(4): 357–361.