An efficient and exact approach for detecting trends with binary endpoints

Guogen Shan[1], Changxing Ma[1], Alan D. Hutson[1], and Gregory E. Wilding[1] *

[1] *Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY 14214*

April 28, 2011

Abstract

An exact testing approach was developed by Lloyd [21] to eliminate nuisance parameters which was shown to be advantageous in testing for differences between two population proportions. We utilizes this approach to obtain unconditional tests for trends in $2 \times K$ contingency tables. The unconditional procedure is compared to other unconditional and conditional approaches based on the well know Cochran Armitage test statistic. An example is given to illustrate the approach and a comparison between the methods with regards to Type I error and power is provided. The proposed procedure is preferable due to it being less conservative and its superior power properties.

Keywords: Cochran Armitage Test, Contingency Tables, Exact Tests, E+M p-value, Nuisance Parameters, Unconditional test.

*Corresponding author. Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY, 14214, USA.

E-mail address:******@buffalo.edu

1 Introduction

The problem of statistically testing the equality of K binomial populations against an ordered alternative exists in a wide variety of research settings. In clinical trials, for example, it is often of interest to investigate the relationship between the dosage and the effect of the drug under study with regards to a binary outcome. Several testing procedures have been proposed for analyzing the data from such studies which can be organized in a $2 \times K$ contingency tables. The Cochran-Armitage (CA) test (Cochran [14], Armitage [3]) is the most frequently used test in this setting. The test is based upon the estimated slope coefficient from the weighted regression of the observed proportions of success on fixed scores corresponding to the grouping variable. The test can be highly sensitive to linearity between the true proportions and the fixed scores and will have higher power than the Pearson chi-square test which ignores the intrinsic ordering of the grouping variable. The asymptotic null distribution of the square of the CA test statistic, which is that of a chi-square random variable with one degree of freedom, has been recommended for use in practice. The properties of the test when in use with the asymptotic null distribution have been investigated by Agresti and Yang [1], who showed the Type I error rate to be inflated thereby implying that use of the asymptotic null distribution is questionable (see, Chen et al. [12], Corcoran et al. [16]). Kang and Lee [19] studied the case of large sample sizes, and theoretically proved that the actual significance level of the CA trend test is always greater than or equal to the nominal level. The non-null asymptotic distribution of the test statistic, which was derived by Nam [26], is of the form $a\Phi(b, a)$, where a and b are functions of the proportions under the alternative and the dose scores, and $\Phi(b, a)$ is the CDF of a normal random variable with mean b and standard deviation a. The power of the CA test based on the asymptotic non-null distribution can be over estimated [24], which makes the use of the asymptotic approach in the context of study planning misleading. As an alternative to the CA test, Neuh$\ddot{a}$user [28] proposed a nonparametric trend test among binomial proportions based on a modification of the Baumgarter-Weiß-Schindler statistic.

In light of the problems of Type I error control, an exact CA test may be considered in order to preserve the nominal level of the test (Agresti et al. [2], Tang et al. [30]). The difficulty in the development of an exact procedure is the reliance of the null distribution on the unknown common probability of success across all groups. Mehta et al. [24] proposed an exact conditional test where both the row and column totals are treated as fixed.

Corcoran et al. [16] studied the exact power of the asymptotic CA test, the exact conditional CA test, and the procedure given by Cohen and Sackrowitz [15]. The commercial softwares StatXact [25] and SAS [29] are available to compute both the exact conditional p-value for a given data set, and within StatXact, the exact power of the conditional CA testing procedure is available for the purpose of study planning.

An alternative for nuisance parameter elimination in the unconditional framework is the maximization method (Basu [5], and Bickel and Doksum [9]). In short, the p-value is calculated by maximizing the null likelihood, defined as the tail probability determined by the test statistic, over the complete domain of the unknown nuisance parameter. Barnard [4] was the first to propose the unconditional approach for testing independence in the 2×2 contingency tables and further claimed that the unconditional test can be more powerful than the conditional version provided by Fisher [17]. In addition to the mechanics, there exist philosophical differences in the conditional and unconditional approaches. The debate in the context of testing for differences between two independent proportions has been summarized by Mehrotra et al. [23]. An empirical investigation of exact conditional and unconditional testing approaches in conjunction with the CA test statistic was given by Tang et al. [31], where the actual significance level and exact power were provided using a recursive polynomial multiplication algorithm. They demonstrated that the exact unconditional CA test not only possesses an actual size closer to the nominal level, but also is associated with a considerable power gain compared with the conditional version.

A modification of the unconditional maximization method has been suggested by Berger and Boos [8] in the 2×2 table setting, who noticed that the full maximization method can be more conservative than the conditional approach in some scenarios. They proposed a new method of computing the so called confidence interval p-value, which is defined as the maximum of the null likelihood over a $100(1-\beta)\%$ confidence region C_β for the nuisance parameter, then adding a penality term β. For use in practice, the value of β is rather small, such as 0.001 or 0.0001. Berger [7] showed that the proposed confidence interval method reduces the conservatism of the unconditional approach. Freidlin and Gastwirth [18] extended the confidence interval p-value approach to obtain unconditional tests for trends in $2 \times K$ tables, and showed that the unconditional test is more powerful than the conditional test when the expected number of responses is small. Although less conservative, the confidence interval p-value approach to statistical problems suffers from a number of weakness as discussed by Lloyd [21].

In order to address the weakness he proposed use of the E+M (Estimation + Maximization) p-value [21]. The estimated p-value is first obtained by replacing the unknown nuisance parameter of the null distribution with its maximum likelihood estimate (MLE); the E+M p-value is then obtained by maximizing the null likelihood using the estimated p-value as a test statistic. The E+M approach has been successfully applied to the likelihood ratio statistic for testing for differences between two independent proportions [20, 22], and has been recommended for use in practice due to its power compared to other unconditional testing procedures. No work has been done thus far in extending the E+M approach to testing for monotone trends among more than two independent proportions.

The rest of this article is organized as follows. In Section 2, we briefly review exact CA tests for detecting linear trend and propose an exact unconditional version based on the E+M approach. In Section 3, we compare the performance of the competing tests, studying the size and power of the procedures under a wide range of conditions. Section 4 is given to conclusions.

2 Existing methods and the proposed approach

Suppose that K ordered doses are used in a dose-response study, $d_1, d_2, \cdots, d_K$ with $d_1 < d_2 < \cdots < d_K$, and say a fixed number of subjects, n_i, are assigned to the $i-$th group where all subjects receive dose d_i. Letting p_i be the probability of response at dose d_i, we wish to test the null hypothesis

$$H_0 : p_1 = p_2 = \cdots = p_K =: p,$$

against an ordered alternative of the form

$$H_a : p_1 \leq p_2 \leq \cdots \leq p_K \quad \text{and} \quad p_1 < p_K. \tag{1}$$

An efficient test of the null hypothesis against the ordering alternative is the CA test, which is based on the test statistic

$$T_{CA}(\tilde{x}) = \frac{\sum_{i=1}^K x_i d_i - \hat{p} \sum_{i=1}^K n_i d_i}{\sqrt{\hat{p}(1-\hat{p}) \sum_{i=1}^K n_i (d_i - \bar{d})/n^2}},$$

where $\hat{p} = \sum_{i=1}^{K} x_i/n$, $n = \sum_{i=1}^{K} n_i$, $\bar{d} = \sum_{n=1}^{K} n_i d_i/n$, and x_i is the observed number of responders among the n_i subjects at dose level d_i. Larger values of the test statistic would denote evidence against the null. Although we focus on an alternative of the form (1), the CA test statistic is also suitable for the ordered alternative $p_1 \geq p_2 \geq \cdots \geq p_K$ and $p_1 > p_K$, and two-sided tests may be based on use of the statistic T_{CA}^2. Letting $s = \sum_{i=1}^{K} x_i$ and $t = \sum_{i=1}^{K} x_i d_i$, the test statistic T_{CA} can be alternatively expressed as

$$T_{CA}(s,t) = \frac{t - \bar{d}s}{\sqrt{s(n-s)\sum_{i=1}^{K} n_i(d_i - \bar{d})/n^2}}.$$

The remainder of this section is given to the formal description of existing methods for obtaining p-values based on the above test statistic, an outline of the proposed approach, and an example illustrating use of the procedures.

2.1 Existing methods

A commonly implemented approach to eliminate nuisance parameters in categorical analysis is conditioning on observed marginal totals. Mehta et al. [24] suggested computing the observed significance of the test statistic T_{CA} by conditioning on the total number of responses s. Let $\tilde{x} = (x_1, x_2, \cdots, x_K)$ be a vector of the number of responders for the K populations, and $\tilde{x_0}$ be the observed vector for a given data set. The exact conditional p-value for testing linear trends is defined as [24]

$$P_{Cond}(\tilde{x_0}) = Pr(T_{CA}(\tilde{x}) \geq T_{CA}(\tilde{x_0}) | \sum \tilde{x} = \sum \tilde{x_0}, H_0) = \sum_{\tilde{x} \in \Omega_{Cond}(\tilde{x_0})} \frac{\prod_{i=1}^{K} \binom{n_i}{x_i}}{\binom{n}{s}},$$

where $s = \sum \tilde{x_0}$ and the reject region $\Omega_{Cond}(\tilde{x_0}) = \{\tilde{x} : T_{CA}(\tilde{x}) \geq T_{CA}(\tilde{x_0})$ and $\sum \tilde{x} = \sum \tilde{x_0}\}$. We note that the conditional null distribution consists of a small set of unique values of the test statistic when sample sizes are small, thereby resulting in a testing procedure which performs in a conservative manner. A straightforward unconditional scheme to obtaining the p-value is to define the p-value (referred to as the M p-value) to be the supremum of the null likelihood over the whole range of the nuisance parameter so to eliminate the dependence of the test on the quantity. The M p-value [30] is given as

$$P_M(\tilde{x_0}) = \sup_{p \in [0,1]} \{ \sum_{\tilde{x} \in \Omega_M(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \},$$

where $\Omega_M(\tilde{x_0}) = \{\tilde{x} : T_{CA}(\tilde{x}) \geq T_{CA}(\tilde{x_0})\}$ is the reject region.

Mehta et al. [24] evaluated the exact Type I error rate and power of the conditional and unconditional M approaches. For various choices of p and n, they found that the actual Type I error rate is generally more conservative using the conditional strategy compared to the unconditional test, although the difference between the tests diminishes with increasing sample sizes. Both tests have nearly identical power when the sample sizes are large, but for small sample sizes the unconditional test is generally more powerful due to the less discrete null distribution. In some scenario, the conditional test was shown to be more powerful than the unconditional test.

Berger and Boos [8] developed another unconditional approach to eliminate the nuisance parameter based on the following two observations: If the parameter space is not bounded, then it can be difficult to find the global maximum of the null likelihood over the parameter space of the nuisance parameters. In addition to this computational problem, there is additional information contained in the observed data, specifically the MLE and the corresponding confidence interval of the parameters. Therefore, they proposed the p-value to be the supremum of the null likelihood taken over the confidence interval rather than over the whole range. Freidlin and Gastwirth [18] extended this approach for testing homogeneity in the $2 \times K$ table. The confidence interval p-value (referred to as the CI p-value), based on the CA test statistic, is defined by

$$P_{CI}(\tilde{x_0}) = \sup_{p \in C(\tilde{x_0})} \{ \sum_{\tilde{x} \in \Omega_{CI}(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \} + \beta,$$

where $\Omega_{CI}(\tilde{x_0}) = \Omega_M(\tilde{x_0})$, and $C(\tilde{x_0})$ is the Clopper-Pearson $100(1-\beta)\%$ interval [13], and $\beta = 0.001$ as per Berger and Boos' [8] recommendation. Freidlin and Gastwirth [18] showed a power gain by using the CI approach over competitors which decreased as the sample size increased. Lloyd [20] pointed out that the CI p-value proposed by Berger and Boos is not exact and can be highly conservative in some cases, and a similar observation can be made for the test given by Freidlin and Gastwirth [18].

2.2 E+M approach based on CA test

An unconditional approach based on estimation and maximization, applicable to many statistical problems, has been proposed by Lloyd [21]. Suppose there is a test statistic Q, and without loss of generality, say large values of Q support the alternative H_a. The estimated p-value (E p-value) for a given observed data $\tilde{x_0}$ is

$$P_E(\tilde{x_0}) = \sum_{\tilde{x} \in R_E(\tilde{x_0})} \prod Pr(\tilde{x} | \hat{\boldsymbol{\psi}}),$$

where $\hat{\boldsymbol{\psi}}$ and $R_E(\tilde{x_0}) = \{\tilde{x} : Q(\tilde{x}) \geq Q(\tilde{x_0})\}$ are the MLE for the nuisance parameter $\boldsymbol{\psi}$ and the reject region for given data $\tilde{x_0}$, respectively. This E p-value, although convenient for use in practice, is not an exact p-value for the test [20, 22]. An exact testing procedure is obtained by using the E p-value as a test statistic, fully maximizing the null likelihood over the whole range of the nuisance parameters, thereby defining the E+M p-value as

$$P_{E+M}(\tilde{x_0}) = \sup_{\boldsymbol{\psi} \in \omega} \{ \sum_{\tilde{x} \in R_{E+M}(\tilde{x_0})} \prod Pr(\tilde{x}|\boldsymbol{\psi})\},$$

where ω is the parameter space for $\boldsymbol{\psi}$, and $R_{E+M}(\tilde{x_0}) = \{\tilde{x} : P_E(\tilde{x}) \leq P_E(\tilde{x_0})\}$ is the reject region.

We extend the E+M approach (Lloyd [21]) to obtain an alternative unconditional version of the CA test. For given data $\tilde{x_0}$, the reject region for the E p-value is

$$R_E(\tilde{x_0}) = \{\tilde{x} : T_{CA}(\tilde{x}) \geq T_{CA}(\tilde{x_0})\},$$

and the calculated p-value is

$$P_E(\tilde{x_0}) = \sum_{\tilde{x} \in R_E(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} \hat{p}^{x_i} (1-\hat{p})^{n_i - x_i},$$

where $\hat{p} = \sum_{i=1}^{K} x_i/n$ is the value of maximum likelihood estimate of p under the null hypothesis. The corresponding reject region for E+M p-value for the CA test is

$$\Omega_{E+M}(\tilde{x_0}) = \{\tilde{x} : P_E(\tilde{x}) \leq P_E(\tilde{x_0})\},$$

and the E+M p-value is

$$P_{E+M}(\tilde{x_0}) = \sup_{p \in [0,1]} \{ \sum_{\tilde{x} \in \Omega_{E+M}(\tilde{x_0})\}} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \}.$$

Although not considered here, further estimation steps may be performed before maximization in order to obtain a null likelihood which is even less reliant on the nuisance parameter. We now applied the theoretical results from Lloyd [21] to our newly defined test.

Theorem 2.1. *The E+M p-value $P_{E+M}(\tilde{x_0})$ is a non-decreasing function of $P_E(\tilde{x_0})$*

Proof: Suppose $P_E(\tilde{x_1}) < P_E(\tilde{x_2})$, where $\tilde{x_1}$ and $\tilde{x_2}$ are the vectors from two observed tables. For any $\tilde{x} \in \Omega_{E+M}(\tilde{x_1})$, from the the definition of the reject region for the E+M p-value, we know that $P_E(\tilde{x}) \leq P_E(\tilde{x_1})$. Then obviously $P_E(\tilde{x}) < P_E(\tilde{x_2})$ when $P_E(\tilde{x_1}) < P_E(\tilde{x_2})$ is ture. Thus, $\tilde{x} \in \Omega_{E+M}(\tilde{x_2})$, and so $\Omega_{E+M}(\tilde{x_1}) \subseteq \Omega_{E+M}(\tilde{x_2})$. For any given p,

$$\sum_{\tilde{x} \in \Omega_{E+M}(\tilde{x_1})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i} \leq \sum_{\tilde{x} \in \Omega_{E+M}(\tilde{x_2})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i}(1-p)^{n_i - x_i},$$

therefore the E+M p-value $P_{E+M}(\tilde{x_0})$ is a non-decreasing function of $P_E(\tilde{x_0})$.

Theorem 2.1 is also true for the M p-value, but not always true for the CI p-value due to Freidlin and Gastwirth [18]. The CI p-value is calculated based on the partial maximization, and the bounds of the CI are dependent on the observed data. For these reasons, the CI p-value is not a non-decreasing function of the test statistic.

The traditional definition of validity of a p-value, $P(\tilde{x})$, is basic [11], which only require a guarantee of the nominal level of the test: $Pr(P(\tilde{x}) \leq \alpha | \theta) \leq \alpha$, where θ is a nuisance parameter, is satisfied for any θ in the parameter space and $0 \leq \alpha \leq 1$. For discrete data, the Type I error rate will be less than the nominal level unless $P(\tilde{x}) = \alpha$ for some $\tilde{x}$, therefore resulting in a conservative test for any nominal size [21]. Lloyd [21] defined the exact p-value in a more stringent way: If $\sup_{\theta} Pr(P(\tilde{x}) \leq P(\tilde{x_0}) | \theta)$, is exactly equals to $P(\tilde{x_0})$ for any $\tilde{x_0}$, then the p-value $P(\tilde{x_0})$ is exact.

Theorem 2.2. *The E+M p-value $P_{E+M}(\tilde{x_0})$ is exact.*

The proof follows from Theorem 1 in Lloyd [21]. Note that the CI p-value $P_{CI}(\tilde{x_0})$ and the E p-value $P_E(\tilde{x_0})$ are not exact [21, 20].

2.3 An example

We consider an example and apply the five different approaches discussed in this article: (1) the asymptotic approach; (2) the conditional approach; (3) the M approach; (4) the CI approach; (5) the E+M approach. We analyzed the follicular cell adenomas data from the study reported in Bickis and Krewski [10] shown in the Table 1. Figure 1 shows the one sided exact null likelihood of the tests as a function of p for the data set. The exact p-values can be seen to be 0.090 for the conditional approach, 0.074 for the M approach, 0.047 for the CI approach with the CI (0.0052, 0.2064), and 0.036 for the E+M approach. At the 0.05 significance level, the CI and E+M methodologies support the alternative hypothesis.

As seen in the Figure 1, obviously the null likelihood for the asymptotic and conditional approaches does not depend on p. There is a spike in the curve for the M approach near the right end of the range of the nuisance parameter, contributing to the larger p-value for this approach. The CI p-value is smaller in comparison since the confidence intervals does not contain the spike. The curve for the E+M approach is much flatter than that of the M approach, and importantly, the spike in the curve for the M approach is no longer present for the E+M approach demonstrating the weaker dependence of the test statistic on the nuisance parameter. The curve for the E+M approach is uniformly lower than that of other approaches, therefore the E+M p-value is the smallest among them.

3 Results

To evaluate the performance of the proposed procedure with competitors, the exact Type I error rate and power at the 0.05 nominal significance level were compared. Figure 2 shows the Type I error rates of the tests as a function of the true value of the nuisance parameter. The balanced case with sample sizes 20 and 50 per group were considered. Plots (a) and (b) of Figure 2 show the Type I error rates of the five tests with $K = 3$, and doses $(d_1, d_2, d_3) = (0, 1, 3)$, and plots (c) and (d) of Figure 2 illustrate the corresponding quantities with $K = 4$, and doses $(d_1, d_2, d_3, d_4) = (0, 1, 3, 4)$. The sample size for each group in plots (a), (b), (c), and (d) are 20, 50, 20, and 50, respectively. Maximization was performed through calculation of probabilities using jumps of 0.01 in the range of p. As can be seen, the asymptotic test violates the nominal significance level for most values of p. Tests based on the M approach, E+M approach, and conditional approach guarantee the nominal significance level of the test. The conditional test can be seen to be conservative, even for the larger sample sizes, but the procedure can be less conservative than the unconditional M or CI approaches for some scenarios. The E+M exact unconditional test has a flat significance level curve, which is less conservative than the conditional and other unconditional tests in most cases.

Having examined the true Type I error of each test, we now compare the procedures with respect to power. For a given alternative $p_i, i = 1, 2, \cdots, K$, the exact power of the exact unconditional tests is

$$Power_A(\tilde{x_0}) = \sum_{\tilde{x} \in \Omega_A(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p_i^{x_i} (1 - p_i)^{n_i - x_i},$$

where $A = \{M, CI, E + M\}$. The power for the unconditional approach may be written as a weighted sum of conditional powers, and the detailed expression for the power calculation can be found in Mehta et al. [24]. The asymptotic approach will not be considered due to the substantial Type I error violation, thereby invalidating the procedure. The power of four exact tests for $K = 4$, $n_i = 20, i = 1, 2, 3, 4$, are shown in Figure 3, 4, and 5. Dose scores for Figure 3, 4, and 5 are (0,1,2,3), (0,1,2,4), and (0,2,3,4), respectively. Each of these three figures contains four plots, plot (a)–(d), with corresponding four values of p_1: 0.05, 0.1, 0.2, 0.45. Each curve is plotted as a function of the parameter γ from the logistic regression model [31] given by

$$p_i = \frac{e^{log(p_1/(1-p_1))+\gamma d_i}}{1+e^{log(p_1/(1-p_1))+\gamma d_i}}, \quad i = 2, 3, 4.$$

The difference in power between the conditional and unconditional tests is substantial. In all 12 cases, the E+M approach is uniformly more powerful across γ than the conditional approach, and the difference between points in the curves can be as high as 5.6%. The M approach is not as powerful as the E+M approach, although they have close to equivalent power in some cases. In 1 out of 12 cases, the power of CI approach at some γ values is 1.1% greater than that of the E+M approach. However, in other cases, the E+M approach can be 4.4% higher power than the CI approach.

4 Conclusion

In this article, we extended Lloyd's [21] approach to obtain unconditional versions of the procedure developed by Cochran [14] and Armitage [3] for testing the equality of the response rates in K binomial populations against an ordered alternative. By using the estimated p-value as a test statistic, and then maximizing the null likelihood over the range of the nuisance parameter, a procedure is obtained which not only guarantees the nominal level of the test, but is also associated with a power gain compared to competitors. Our investigation in this article is based on the CA trend test. This test was chosen because of its popularity in dose response data analysis. There are several other tests may be alternatively used, one of them being the rank based test statistic proposed by Neuh*ä*user [27], which is a modified Baumgartner-Weiß-Schindler statistic [6]. An investigation of the test using a similar approach as considered in this note is currently underway.

References

[1] A. Agresti and M. Yang. An empirical investigation of some effects of sparseness in contingency tables. *Computational Statistics and Data Analysis*, 5(1):9–21, 1987.

[2] A. Agresti, C. R. Mehta, and N. R. Patel. Exact Inference for Contingency Tables with Ordered Categories. *Journal of the American Statistical Association*, 85(410):453–458, 1990.

[3] P. Armitage. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, 11(3):375–386, 1955.

[4] G. A. Barnard. A new test for 2 × 2 tables. *Nature*, 156:177, 1945.

[5] D. Basu. On the Elimination of Nuisance Parameters. *Journal of the American Statistical Association*, 72(358):355–366, 1977.

[6] W. Baumgartner, P. Weiß, and H. Schindler. A Nonparametric Test for the General Two-Sample Problem. *Biometrics*, 54(3):1129–1135, 1998.

[7] R. L. Berger. More Powerful Tests from Confidence Interval p Values. *The American Statistician*, 50(4):314–318, 1996.

[8] R. L. Berger and D. D. Boos. P Values Maximized Over a Confidence Set for the Nuisance Parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

[9] P. J. Bickel and K. A. Doksum. *Mathematical Statistics*. Holden-Day, Inc, 1977.

[10] M. Bickis and D. Krewski. Statistical issues in the analysis of the long-term carcinogenicity bioassay in small rodents: an empirical evaluation of statistical decision rules. *Fundamental and Applied Toxicology*, 12(2):202–221, 1989.

[11] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, second edition, 2001.

[12] J. J. Chen, R. L. Kodell, and B. A. Pearce. Significance Levels of Randomization Trend Tests in the Event of Rare Occurrences. *Biometrical Journal*, 39(3):327–337, 1997.

[13] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934.

[14] W. G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451, 1954.

[15] A. Cohen and H. B. Sackrowitz. An Evaluation of Some Tests of Trend in Contingency Tables. *Journal of the American Statistical Association*, 87(418):470–475, 1992.

[16] C. Corcoran, C. Mehta, and P. Senchaudhuri. Power comparisons for tests of trend in dose-response studies. *Statistics in Medicine*, 19(22):3037–3050, 2000.

[17] R. A. Fisher. *Statistical methods for research workers.* Oliver & Boyd, enlarged 14th edition, 1970.

[18] B. Freidlin and J. L. Gastwirth. Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables. *Biometrics*, 55(1):264–267, 1999.

[19] S. Kang and J. Lee. The size of the cochranarmitage trend test in 2 × c contingency tables. *Journal of Statistical Planning and Inference*, 137(6):1851–1861, 2007.

[20] C. J. Lloyd. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics*, 64(3):716–723, 2008.

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

[22] C. J. Lloyd and M. V. Moldovan. A more powerful exact test of noninferiority from binary matched-pairs data. *Statistics in Medicine*, 27(18):3540–3549, 2008.

[23] D. V. Mehrotra, I. S. F. Chan, and R. L. Berger. A Cautionary Note on Exact Unconditional Inference for a Difference between Two Independent Binomial Proportions. *Biometrics*, 59(2):441–450, 2003.

[24] C. R. Mehta, N. R. Patel, and P. Senchaudhuri. Exact Power and Sample-Size Computations for the Cochran-Armitage Trend Test. *Biometrics*, 54(4):1615–1621, 1998.

[25] C. R. Mehta, N. R. Patel, P. Senchaudhuri, and C. D. Corcoran. StatXact. 2005.

[26] J. M. Nam. A Simple Approximation for Calculating Sample Sizes for Detecting Linear Trend in Proportions. *Biometrics*, 43(3):701–705, 1987.

[27] M. Neuhäuser. One-side two-sample and trend tests based on a modified Baumgartner-Weiß-Schindler statistic. *Journal of Nonparametric Statistics*, 13:729–739, 2001.

[28] M. Neuhäuser. An exact test for trend among binomial proportions based on a modified Baumgartner-Weiß-Schindler statistic. *Journal of Applied Statistics*, 33(1):79–88, 2006.

[29] SAS Institute Inc. *SAS 9.1.3 Help and Documentation*, 2000-2004.

[30] M.-L. Tang, P.-S. Chan, and W. Chan. On Exact Unconditional Test for Linear Trend in Dose-Response Studies. *Biometrical Journal*, 42(7):795–806, 2000.

[31] M.-L. Tang, H. K. Ng, J. Guo, W. Chan, and B. P. Chan. Exact Cochran-Armitage trend tests: comparisons under different models. *Journal of Statistical Computation and Simulation*, 76(10):847–859, 2006.

Table 1: Follicular cell adenomas study

Response	Scores		
	0	0.5	1
Tumor	0	0	4
No Tumor	8	23	35

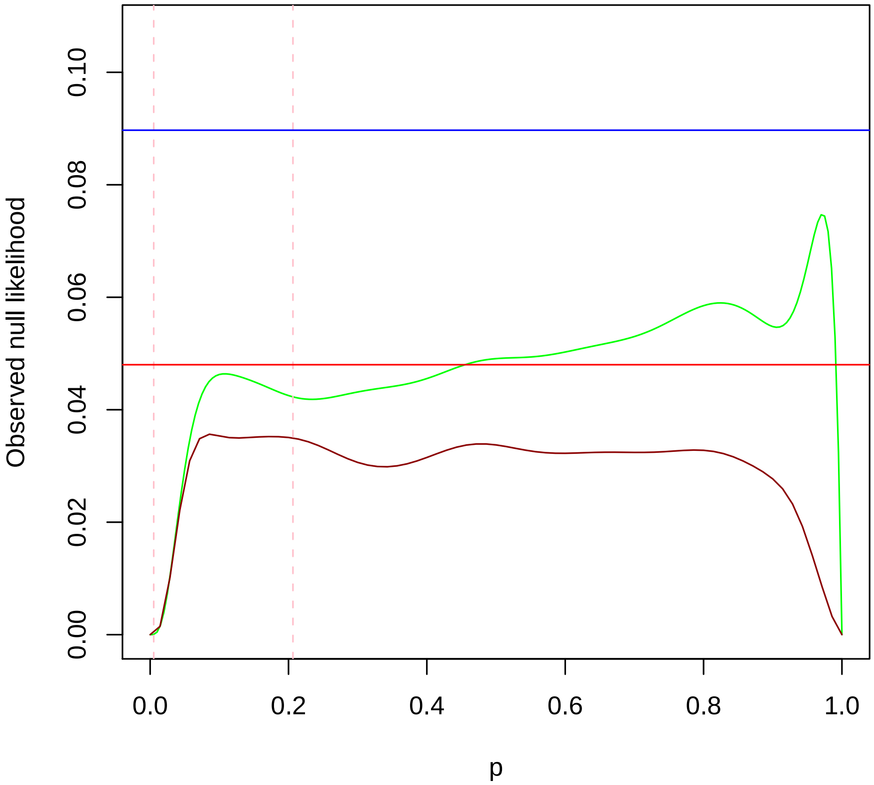

Figure 1: Observed null likelihood curves for different tests using the data from the follicular cell adenomas study: Asymptotic approach (red); Conditional approach (blue); M approach (green); E+M approach (darkred); Confidence interval lines for CI approach (pink).

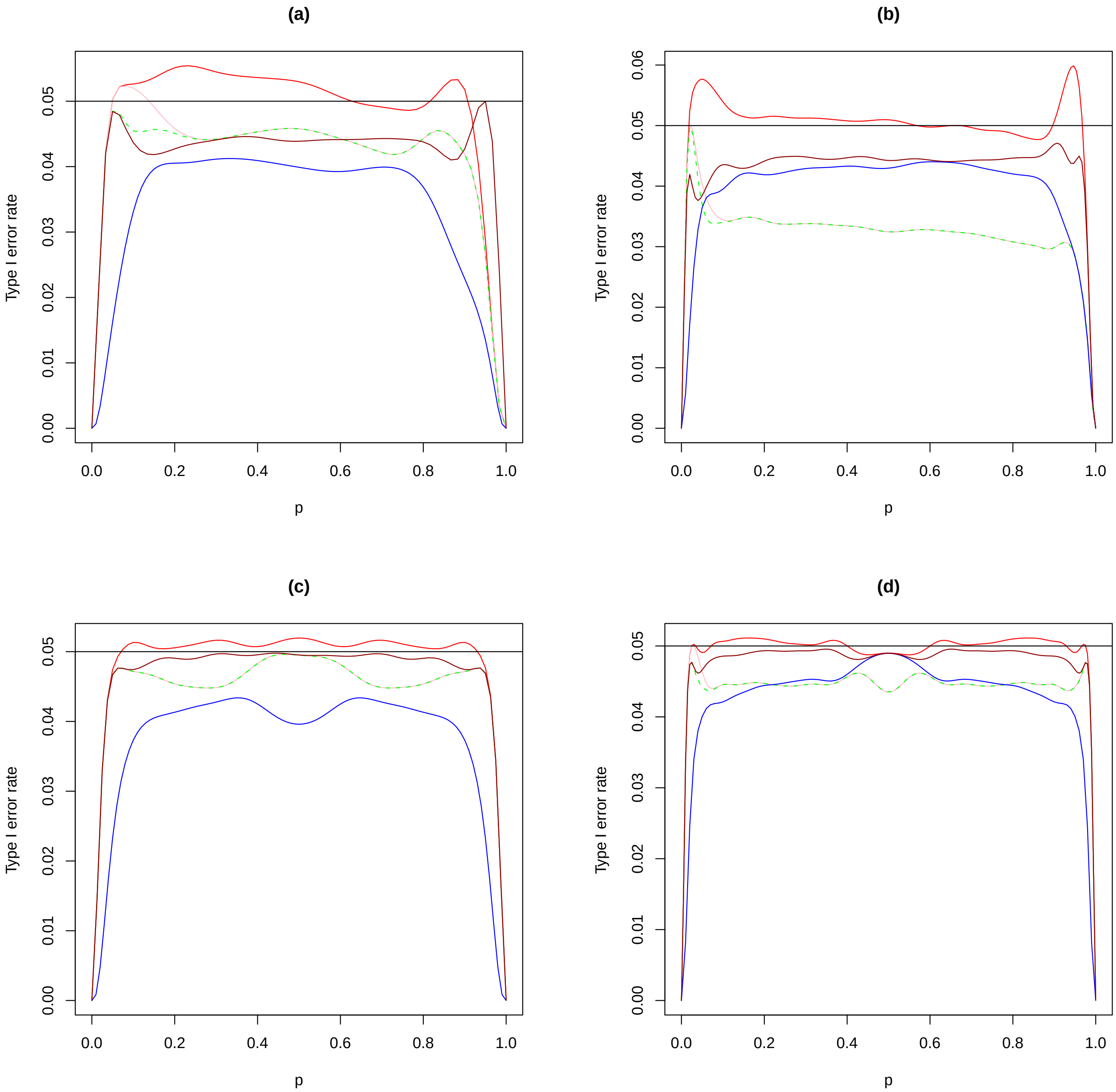

Figure 2: Type I error rates for: (a) K=3 n=20, (b) K=3 n=50, (c) K=4 n=20, and (d) K=4 n=50. Asymptotic approach (red); Conditional approach (blue); M approach (green); CI approach (pink); E+M approach (darkred).

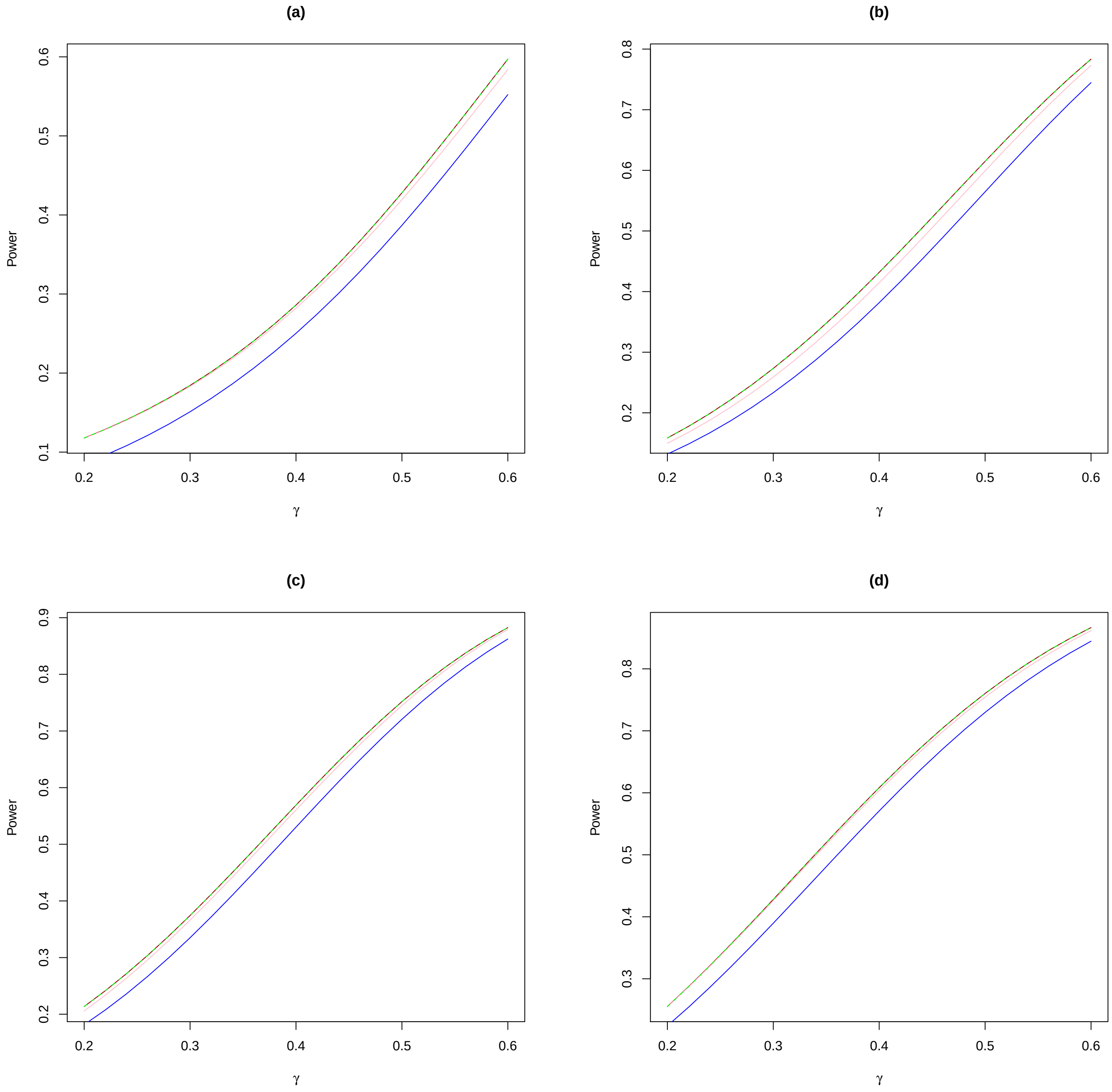

Figure 3: Power study with K=4, n=20, and the dose d=(0,1,2,3). Four difference alternatives: (a) p_1=0.01 (b) p_1=0.1 (c) p_1=0.2 (d) p_1=0.45. Four exact tests: Conditional approach (blue); M approach (green); CI approach (pink); E+M approach (darkred).

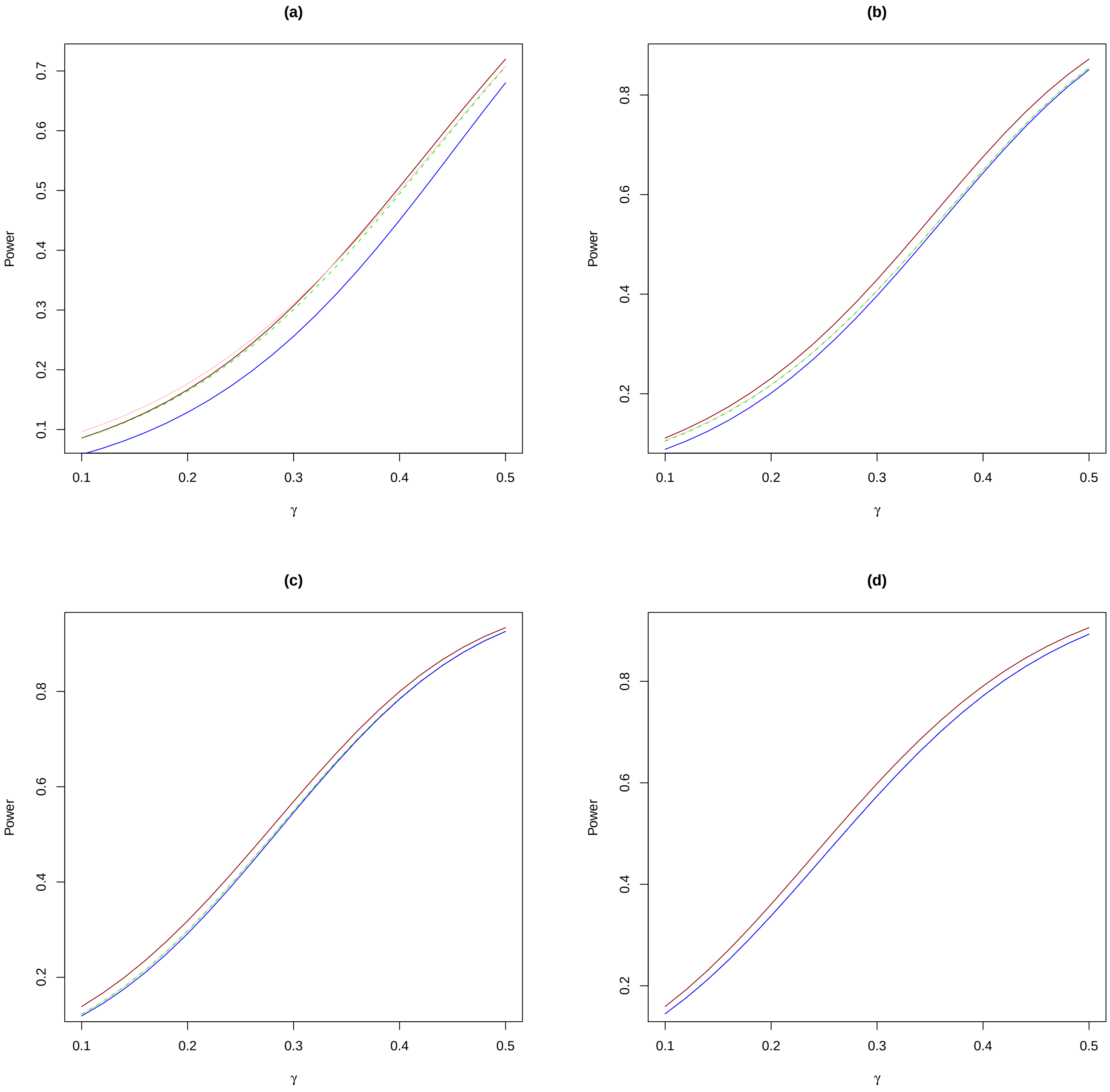

Figure 4: Power study with K=4, n=20, and the dose d=(0,1,2,4). Four difference alternatives: (a) p_1=0.01 (b) p_1=0.1 (c) p_1=0.2 (d) p_1=0.45. Four exact tests: Conditional approach (blue); M approach (green); CI approach (pink); E+M approach (darkred).

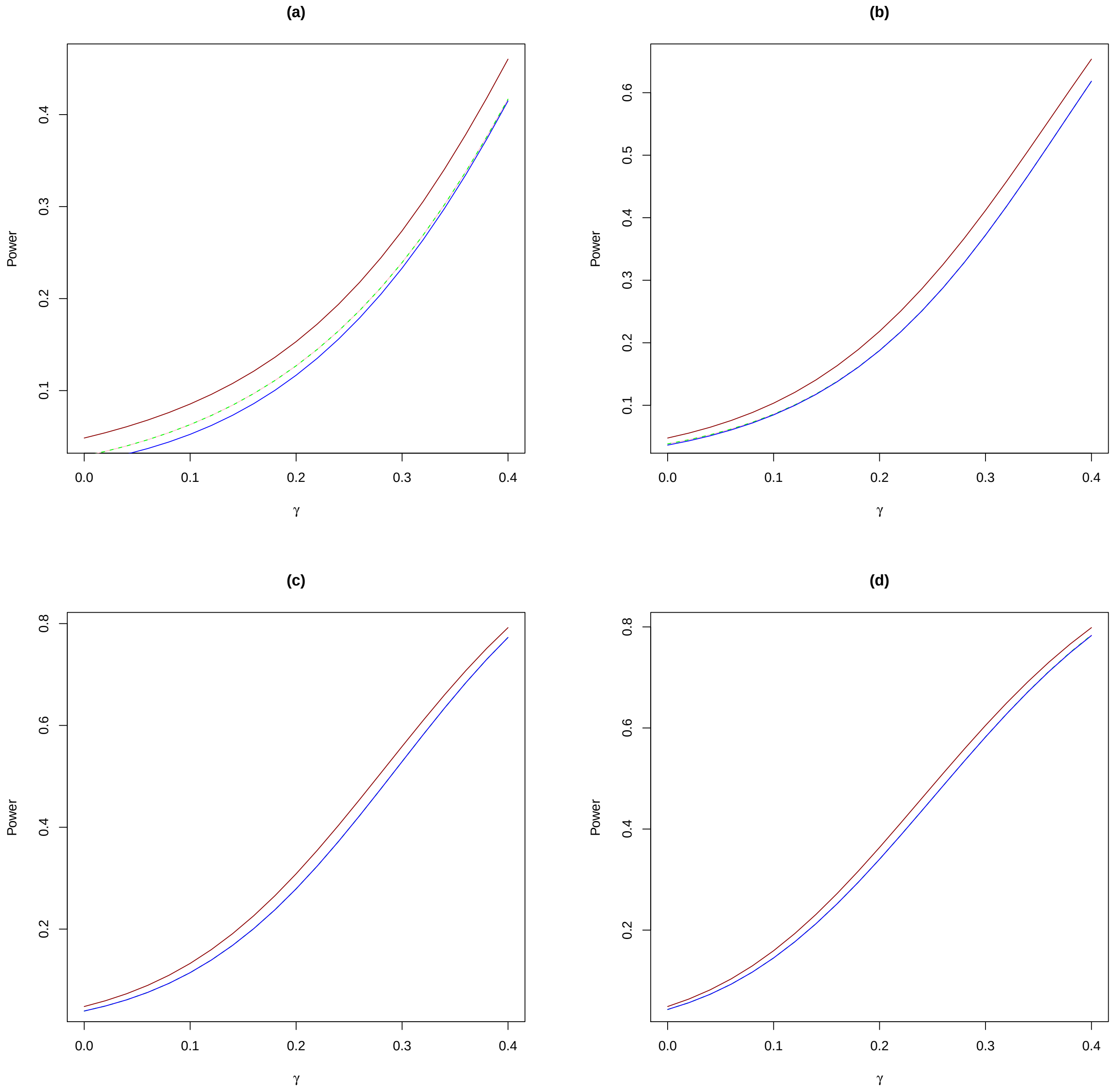

Figure 5: Power study with K=4, n=20, and the dose d=(0,2,3,4). Four difference alternatives: (a) p_1=0.01 (b) p_1=0.1 (c) p_1=0.2 (d) p_1=0.45. Four exact tests: Conditional approach (blue); M approach (green); CI approach (pink); E+M approach (darkred).