Research Article

On sensitivity of p-values in testing binomial proportion using asymptotic and exact methods

Sakthivel S[a], Subbiah M[b], Ramakrishnan R[c]

[1]Research Scholar, Department of Mathematics, L.N Government College, Ponneri, Tamil Nadu, INDIA.
[2]Department of Mathematics, L.N Government College, Ponneri, Tamil Nadu, INDIA.
[3]National Institute of Epidemiology (ICMR), Ayapakkam, Chennai, Tamil Nadu, INDIA.
Email: sakthiapj@gmail.com

Abstract

Statistical inference for single binomial proportion has witnessed active research in variety of studies; yet it provides ample scope to investigate its characteristics in classical approaches. The lack of consensus among applied and / or non-statisticians in choosing methods between exact, approximate tests with or without continuity corrections have augmented these activities. Objective of this paper is to evaluate the robustness of different classical methods using the following criteria i) Non-Polarized conclusion and ii) Continuous and bi-monotonic properties of P-value. Comparative analyses are aimed to recommend the appropriate methods based on the characteristics of binomial model. Simulation study has been performed using Monte Carlo method to substantiate the findings and recommendations.

Keywords: Approximation methods, Binomial proportion, Exact methods, Polarized conclusions, P-value.

[*]**Address for Correspondence:**
Dr. Sakthivel S, Research Scholar, Dept. Of Statistics, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu INDIA.
Email: sakthiapj@gmail.com

Received Date: 26/10/2014 Accepted Date: 02/11/2014

Access this article online

Quick Response Code:

Website: www.statperson.com

DOI: 08 November 2014

INTRODUCTION

Statistical research and literature on inference regarding the single binomial proportion is quite extensive. The underlying model based on the number of successes (y) isy ~ Binomial (n, θ) where θ is the proportion of success from n independent Bernoulli trials. Performance of many methods for point and interval estimation of θ has been routinely compared with or without specific applications (Tang et al. (2012), Vos et al. (2008), Pires et al. (2008), Zhou et al. (2008) and Williamson et al. (2004)).This includes the nature of n as large or small, and that of observed proportion which may be near the boundaries of θ in [0, 1]. Also, in the realm of null hypothesis and significance testing (NHST) includes approximate and exact test procedures involving research hypotheses about θ. However, not many discussions can be found in NHST about θ especially to address the careful choice of methods suitable to specific situations. Several studies including Joseph et al. (2005) discuss elaborately the genesis of testing for a binomial proportion from an application point of view. Yet the robustness of these methods in terms of consensual usage of large or small values of the parameters draws an active research attention still; secondly, usage of null proportion or sample proportion in the computation of standard errors, issues of continuity corrections (Agresti (2007), Kanji (2006)) make the situations more interesting to investigate the applicability of existing methods for different scenarios. Hitchcock (2009) has pointed out that the use of continuity correction is rather mixed in practice, with texts and software packages differing in their recommendations and that it would be debated for decades to come. Campbell (2007) has made similar observations in the comparison of two binomial parameters to point out that a consensus has not been reached even after two decades and these remarks are still applicable, perhaps more so with the increase use of statistical software by non-statisticians.

How to site this article: Sakthivel S, Subbiah M, Ramakrishnan R. On sensitivity of p-values in testing binomial proportion using asymptotic and exact methods. *International Journal of Statistika and Mathemtika* Nov 2014 to Jan 2015; 12(1): 69-76. http://www.statperson.com (accessed 06 November 2014).

Haber (1980) has also observed that Yates correction is inadequate when performing two sided test. Classical methods (Exact, Wald and Score test statistic with and without continuity correction and Likelihood ratio test) for testing single binomial proportion do have limitation in terms of polarized conclusion when different methods are applied that could be a reminiscence of Simpson's paradox. Also, another notable impediment for classical methods in computing standard errors is the presence of extreme observed values in the data (y= 0 or n). In view of this and noted by Pires et al. (2008),this work has identified the need for a comparative study to facilitate the applied statisticians in selecting an appropriate method; to implement the procedures directly in a statistical software; and for teachers of statistics who have to decide which method to include in a given course. The study has been organized to compare the existing procedures and also to evaluate the robustness of method based on polarization, continuity and monotonicity properties of P-values (Vos et al., (2008)). Three exact and five asymptotic methods are considered under sparse, moderate, extreme values of n and y. Illustrative data sets are extracted from many published application studies to resemble the parametric null values $\theta_0$and more details can be found from Table 1. This further helps in attempting a simulation approach for a reasonable recommendation on the choice of methods. Section 2 presents the materials and methods used for this work, section 3 details the results obtained from data analysis and section 4 provides the conclusion and recommendations.

2. METHODS AND MATERIALS

The inferential problem of single binomial proportion has two quantities; sample size or the number of Bernoulli trials (n) and the number of successes(y). The population proportion (θ) has to be estimated or tested for different plausible values or regions of parametric space [0, 1]. Eight combinations of classical methods have been considered for the comparative analysis; Clopper-Pearson (CPM), Mid-P (MPM) and Blaker test (BKM) are the three exact tests and five asymptotic methods are Wald with and without continuity correction (WCCM, WM), Score with and without continuity correction (SM, SCCM) and Likelihood ratio test (LRTM). Further, normal approximation methods with and without continuity correction of $\frac{1}{2n}$that use sample proportion $(\hat{\theta})$ or null value(θ_0); they are expressed as Z statistic using standard error of null value (SM), Z statistic using standard error of sample proportion (WM), Z statistic with continuity correction and standard error of null value (SCCM) and Z statistic with continuity correction and standard error of sample proportion (WCCM).

The corresponding test statistics are

1. CPM
   Two-sided P-value:
$$\mathrm{Pr}_{\mathrm{CPM}}(y,\theta) = \sum(\mathrm{Pr}(Y=y) \le \mathrm{Pr}(Y = y_{\mathrm{obs}}))$$
   One sided P-value:
   Left-tailed:
$$\mathrm{Pr}_{\mathrm{CPM}}(y,\theta) = \mathrm{Pr}(Y \le y)\ \mathrm{atobservedoutcome}$$
   Right-tailed:
$$\mathrm{Pr}_{\mathrm{CPM}}(y,\theta) = \mathrm{Pr}(Y \ge y)\ \mathrm{atobservedoutcome}$$
2. BKM
   Two-sided P-value:
$$\mathrm{Pr}_{\mathrm{BKM}}(y,\theta) = \sum(\mathrm{Pr}(Y=y) \le \mathrm{Pr}_{\min}(y,\theta))$$
$$\mathrm{where}\mathrm{Pr}_{\min}(Y=y) = \min[\mathrm{Pr}(Y \le y,\theta), \mathrm{Pr}(Y \ge y,\theta)]\ \mathrm{atobservedoutcome}$$
3. MPM
   One-sided P-value:
   Left-tailed:
$$\mathrm{Pr}_{\mathrm{MPM}}(Y \le y,\theta) = \sum_{\mathrm{Obs}=0}^{y-1} \mathrm{Pr}(Y=\mathrm{Obs}) + \frac{1}{2}\mathrm{Pr}(Y=y)$$
   Right-tailed:
$$\mathrm{Pr}_{\mathrm{MPM}}(Y \ge y,\theta) = \frac{1}{2}\mathrm{Pr}(Y=y) + \sum_{\mathrm{Obs}=y+1}^{n} \mathrm{Pr}(Y=\mathrm{Obs})$$
   Two-sided P-value:
$$\mathrm{Pr}_{\mathrm{MPM}}(y,\theta) = 2 * \min\left(\mathrm{Pr}_{\mathrm{MPM}}(Y \ge y,\theta), \mathrm{Pr}_{\mathrm{MPM}}(Y \le y,\theta)\right)$$

4. WM - To Compute the Z test statistic using the standard error of sample proportion $(\hat{\theta})$.

$$WM = \frac{\hat{\theta} - \theta_0}{s.e\,(\hat{\theta})}$$

$$where s.e(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

5. WCCM - To Compute the Z test statistic with continuity correction adjusted and using the standard error of sample proportion value $(\hat{\theta})$.

$$WCCM = \frac{(\hat{\theta} - \theta_0) + cc}{s.e\,(\hat{\theta})}$$

$$where s.e(\hat{\theta}) = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}$$

6. SM - To compute the Z test statistic using the standard error of null value (θ_0).

$$SM = \frac{\hat{\theta} - \theta_0}{s.e\,(\theta_0)}$$

$$where s.e(\theta_0) = \sqrt{\frac{\theta_0\,(1-\theta_0)}{n}}$$

7. SCC - To Compute the Z test statistic with continuity correction adjusted and using the standard error of null value(θ_0) .

$$SCCM = \frac{(\hat{\theta} - \theta_0) + cc}{s.e\,(\theta_0)}$$

$$where s.e(\theta_0) = \sqrt{\frac{\theta_0\,(1-\theta_0)}{n}}$$

$$cc = \begin{cases} \frac{-1}{2n} if \hat{\theta} > \theta_0 \\ \frac{1}{2n} if \hat{\theta} < \theta_0 \\ 0 \text{ if} |\hat{\theta} - \theta_0| < \frac{1}{2n} \end{cases}$$

The formula to determine P-value for approximation methods (WM, WCCM, SM, SCCM) are as follows:
One-sided P-value:
Left-tailed:$Pr_{LT}(Y \le y, \theta) = Pr_{LT}(Z \le z_\alpha)$
Right-tailed:$Pr_{RT}(Y \ge y, \theta) = Pr_{RT}(Z \ge z_\alpha)$
Two-sided:$Pr_{TS}(y, \theta) = 2 * \min(Pr_{LT}(y, \theta), Pr_{RT}(y, \theta))$

8. LRTM
Two-sided p-value:
Likelihood ratio statistic is

$$2\left[y\log\left(\frac{\hat{\theta}}{\theta_0}\right) + (n-y)\log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right)\right] \sim \chi_1^2$$

$$Pr_{LRTM}(y, \theta) will be computed from chi - squared distribution with 1 d.f.$$

For comparative analysis, twenty five data sets are considered that illustrate variety of combinations of values with respect to number of trials (n) and successes(y) in a binomial experiment. Data sets are extracted from more prominent sources for medical statistics such as (Zar (2010), Fleiss et al. (2003), Agresti (2002 and 2007) and Tang et al. (2012)). The selected data sets cover the following combinations of yandn.

i) Small n, y
ii) Medium n, small y
iii) Large n, small y
iv) Large n, y

v) y = 0 or y = n for small and large n.

Robustness of the methods in terms of sensitivity of P-value is evaluated using the following criteria

i) Non-Polarized conclusion.
ii) Continuous and bi-monotonic property of Two-sided P-value

Also the characteristics of $\hat{\theta}$ are as follows

i) Towards $\hat{\theta}$ boundary – when $\hat{\theta}$ lies in the regions of $(0 \leq \hat{\theta} \leq 0.1)or(0.9 \leq \hat{\theta} \leq 1)$.
ii) Towards $\hat{\theta}$ symmetry – when $\hat{\theta}$ lies in the region $(0.1 < \hat{\theta} < 0.5)$or $(0.5 < \hat{\theta} < 0.9)$.
iii) $\hat{\theta} = 0.5$.

Grid of varying equidistant value of θ_0 are chosen in such a way that when θ_0 is nearing boundary 10 percent of variation from the given θ_0 has been considered in either side of it, whereas 20 percent of variation from either side has been considered when θ_0 is exactly symmetry or towards symmetry. And also, a simulation study has been performed to choose various values between the lower and upper value of null as described above.

Simulation scheme for four different choices of θ_0

i) Fix n
ii) Randomly choose θ value from 0 to 1
iii) Simulate y from Binomial (n, θ) based on i) and ii)
iv) Select θ_o around θ- U(θ- 0.2* θ, θ+ 0.2* θ) or Select θ_o from $(0, \theta)$ or Select θ_o from $(\theta, 1)$or Select θ_o as 0.5. Values beyond boundary are trimmed to 0 or 1.
v) Repeat the step iii with iv for 2000 times

All these combinations are studied on the following five different composite hypotheses

Hypothesis I: $H_0: \theta = \theta_0 vs H_1: \theta \neq \theta_0$
Hypothesis II: $H_0: \theta = \theta_0 vs H_1: \theta > \theta_0$
Hypothesis III: $H_0: \theta = \theta_0 vs H_1: \theta < \theta_0$
Hypothesis IV: $H_0: \theta < \theta_0 vs H_1: \theta \geq \theta_0$
Hypothesis V: $H_0: \theta > \theta_0 vs H_1: \theta \leq \theta_0$

A significance level of α=0.05 is used throughout the study to make a decision regarding the null hypothesis. Considering few limitations of available methods in SAS (9.2 SAS Institute), suitable macro in SAS has been developed for the P-value determination from the eight methods that are discussed in this work.

3. RESULTS

The data analysis based on the illustrative data sets that include the possible values for yso as to reflect extreme, rare or equally likely events in practice. Hence, all the eight methods that are identified as CPM, BKM, MPM, WCCM, WM, SCCM, SM and LRTM and the twenty five datasets are tested for givenθ_0. Comparison has also been made based on a suitable simulation and grid of varying equidistant values of null valueθ_0. Such an attempt of varying values of θ_0 is considered to investigate the significant effect on the results of classical estimator. In addition, continuous and bi-monotonic properties of two-sided P-values are also tested for all methods by adding a success or failure to the observed data. Also it can be noted that P-values nearing zero are presented as < 0.001. Table 2 indicates the following observations when the datasets with the diverse characteristics of small, medium and large n, y, θ_0, $\hat{\theta}$ towards boundary and towards symmetry. When y =0 (dataset XXV) Wald and Likelihood ratio methods fail to determine the P-value. However, a consistent conclusion is observed among CPM, BKM, MPM, SCCM and SM. It can also be observed that both the CPM and BKM P-values are similar, whereas the MPM P-value is only half of the CPM. For small-sized n, the dataset XXIV shows an inconsistent conclusion among the exact methods due to BKM have resulted in reversal of conclusion with the ratio of BKM/CPM as 1.76. And for the dataset IX a consistent conclusion is evident across methods. It is also observed that BKM behaves differently when $\hat{\theta}$ approaches to left and right boundary. Dataset I demonstrates consistent conclusion across methods. For medium-sized n, datasets V, XV, XXIII reveals an inconsistent conclusion among the methods. Dataset V and XXIII show an inconsistent conclusion from the exact and asymptotic methods, whereas dataset XV shows an inconsistent conclusion only in asymptotic methods. Figure 1 depicts this situation. For large n, dataset VI and VII shows inconsistent conclusion among the exact methods due to BKM have resulted in reversal of conclusion with ratio of BKM/CP as 5.04 and 7.53 respectively, whereas consistent conclusion is observed among asymptotic methods. Datasets XIV exhibits consistent conclusion across methods. Table 3 provides the details for data set XXIII and also from Figure 1 undesirable properties can be observed; nevertheless the study has considered other data sets for the comparative analysis. Dataset XXIII with the characteristics

of mediumn, small y and $\hat{\theta}$ towards symmetry shows the reversal in conclusion for BKM, MPM, SCCM and LRTM. In this case, P-value decreases when an additional success is observed and increases when additional failure is observed. Further, it can be observed that dataset V with the characteristics of medium n, y and $\hat{\theta}$ towards symmetry exhibits the discontinuous and non-monotonic behavior of P-values which results in reversal of conclusion. In this case, except BKM, all other methods have resulted in an increase in P-value when an additional success is observed. However, P-value decreased when an additional failure is observed. Dataset XXII with the characteristics of smalln, y and $\hat{\theta}$ towards boundary also demonstrates the above mentioned behavior for CPM, BKM and SCCM. In this case, P-value decreases when an additional success is observed and increases when an additional failure is observed. Dataset XXIV with the characteristics of small n, y and $\hat{\theta}$ towards boundary demonstrates the undesirable properties of the P-value for the methods CPM, MPM and LRTM. Under this situation, P-value increases when an additional success is observed and decreases when an additional failure is observed. The data sets considered in the work provide a reasonable simulation study either on an equi-distant grid values of θ_o and from $U(\theta_o - 0.2*\theta_o, \theta_o + 0.2*\theta_o)$. The comparative results are quite similar to the observations made so far and this property prevails irrespective of different choices of n and y; FigureS3 and S4summarize these results illustrating three sizes of n. The simulation study is summarized in Figure 2 to obtain the necessary inferences and due to paucity of space only limited Tables and Figures are presented, other results of this study can be found from the below link https://sites.google.com/site/chennaibayes/publication

It can be observed that under small and medium size n, when $\hat{\theta}$ approaches left of right boundary then the choices of θ_0 around $\hat{\theta}$ and θ_0 from $(\hat{\theta}, 1)$ exhibit consistent and similar pattern in results among CPM, LRTM, MPM, WM and WM. However, WCCM and SCCM show uniform pattern in results but which is different from other methods. BKM is quite sensitive under these scenarios. Choice of θ_0 from $(0, \hat{\theta})$ yields consistent pattern in results across all methods except BKM and whenn is medium of size and $\hat{\theta}$ approaches left or right boundary. However, for small size of n LRTM, MPM, WM and SM are consistent with CPM whereas SCCM and WCCM show distinct patterns. BKM is quite sensitive. Under small and medium size n, when $\hat{\theta}$ towards symmetry then the choices of θ_0 around $\hat{\theta}$ and $\theta_0$towards symmetry exhibit consistent and similar pattern in results among CPM, LRTM, MPM, WM and WM. However, WCCM and SCCM show uniform pattern in results but which is different from other methods. BKM is quite sensitive under these scenarios. Choice of θ_0 from $(\hat{\theta}, 1)$ yields consistent pattern in results across all methods except BKM when n is small or medium of size and $\hat{\theta}$towards symmetry. Under small size n, when the choice of θ_0 from $(0, \hat{\theta})$ LRTM, MPM, SM and WM are consistent with CPM. Whereas, BKM, SCCM and WCCM exhibit similar pattern. However, the results do not support similar conclusion for medium size n, BKM and CPM exhibit similar pattern.

4. CONCLUSION

The hypotheses testing problems related to single binomial proportion has been investigated with all possible data models that represent physical phenomena. Especially, classical procedures might require additional knowledge to handle the models that occur in boundaries of proportion and sample size; non-consensus recommendations for continuity correction supplement the issue. The problem becomes more interesting because of the close association between null value and the parameter space of the parameter involved in the analysis. Any assumption regarding null value is to express the perceived or theoretical knowledge of population proportion. The general observations of this work provide a precautionary way of selecting appropriate methods for testing single binomial proportion:

- General recommendations:

1. Irrespective of sample size exact method can be recommended when conservatism is not the one of the goals; in particular MPM can be the choice among exact method
2. With the advent of computing facilities, exact methods can be included in any statistical courses
3. LRTM is the more better option if approximate methods are in the choice list and convenience
4. Choice of null should reflect the research problem pragmatically and an exploratory sensitivity analysis could supplement the findings
5. Except CPM method, all other exact and asymptotic methods which are considered in this paper results with the total probability of 1. (i.e sum of left and right tailed probability).
6. P-value from continuity corrected methods is relatively higher when compare to methods without correction in most of the situations.
7. When testing a null value of 0.5, under small, medium and large n and y there exists a relationship between asymptotic methods i.e $\Pr_{WM}(y, \theta) \leq \Pr_{LRTM}(y, \theta) \leq \Pr_{SM}(y, \theta)$

- For small n

1. For the datasets with characteristics of $\hat{\theta}$ towards boundary an absolute difference of 0.05 from $\hat{\theta}$ to $\theta_0$have resulted in failing to reject the null hypothesis across methods.
2. For the datasets with characteristic of $\hat{\theta}$ towards symmetry it is observed that an absolute difference of 0.1 from $\hat{\theta}$ to $\theta_0$have resulted in failing to reject the null hypothesis across methods, whereas an absolute difference of 0.3 from $\hat{\theta}$ to $\theta_0$have resulted in rejecting the hypothesis across methods.
3. And also, it is observed that P-values from SCCM are closer to CPM.

- For medium n:

1. For the datasets with characteristics of $\hat{\theta}$ towards boundary, it is observed that an absolute difference of 0.04 from $\hat{\theta}$ to $\theta_0$have resulted in rejecting the null hypothesis across methods.
2. For the datasets with characteristics of $\hat{\theta}$ towards symmetry a polarized conclusion is observed even between with and without continuity correction of both Wald and Score methods.

- For largen:

1. For the datasets with characteristics of $\hat{\theta}$ towards boundary, it is observed that a maximum absolute difference of 0.008 from $\hat{\theta}$ to $\theta_0$have resulted in failing to reject the null hypothesis across methods, whereas a minimum absolute difference of 0.02 from $\hat{\theta}$ to θ_0 more likely results in rejecting the null hypothesis across methods. And also, it is observed that P-values from SCCM are closer to CPM.
2. For the datasets with characteristics of $\hat{\theta}$ towards symmetry, it is observed that an absolute difference of 0.002 from $\hat{\theta}$ to $\theta_0$have resulted in failing to reject the null hypothesis across methods, whereas an absolute difference of 0.05 from $\hat{\theta}$ to $\theta_0$have resulted in rejecting the hypothesis across methods. And also, it is observed that P-values from WCCM and SCCM are closer to CPM.
3. For the datasets with characteristics of $\hat{\theta}$ symmetry, it is noticed that an absolute difference of atleast 0.04 from$\hat{\theta}$ to $\theta_0$will more likely results in rejecting the null hypothesis. And also, it is clear that P-values from SCCM and WCCM are closer to CPM. In addition, even the P-values from SCCM is closer to SM and WCCM is closer to WM.

Irrespective of the size of n, whether it is small or medium and position of the parameter values from the parameter space of [0,1] the choice of $\theta_0$around $\hat{\theta}$ demonstrates consistent pattern in results. In particular, LRTM, MPM, SM and WM are consistent with CPM, whereas SCCM and WCCM exhibit similar pattern which is different from other methods that might be the influence of continuity corrections which may require further investigation. Interestingly, WM, SM and LRTM exhibit consistent pattern with CPM even when n is small or medium and $\hat{\theta}$ nearing boundary that is value close to 0.1 or 0.9.Hence, because of the simplicity, availability and computational feasibility asymptotic methods can be equally considered to exact method. More importantly the study has indicated that increasing and decreasing behavior of P-value will result in reversal of conclusion when an additional success or failure is observed which is in contrast to what one would expect when sample size increases. Hence, it would be very important that one should choose the methods quite reasonably and the study can easily be accomplished through simulation as a part of sensitivity analysis.

REFERENCES

1. Agresti, A. (2002) Categorical data analysis, Wiley: New York.
2. Agresti, A. (2007) An Introduction to categorical data analysis, Wiley: New York.
3. Campbell, I. (2007) Chi-squared and Fisher-Irwin tests of two-by-two tables with small sample recommendations. Statistics in Medicine, 26, 3661-3675.
4. Fleiss, J.L., Levine, B. and Myunghee, M.P. (2003) Statistical methods for rates and proportions, John Wiley & Sons: New Jersey.
5. Haeber, M. A. (1980) Comparison of Some Continuity Corrections for the Chi-squared Test on 2 × 2 Tables. Journal of the American Statistical Association, 75, 510-515.
6. Hitchcock, D. (2009) Yates and Contingency Tables: 75 Years Later. Electronic Journal for History of Probability and Statistics, 5, 1-14.
7. Joseph, L., Reinhold, C. (2005) Statistical Inference for Proportions. American Journal of Roentgenology, 184, 1057-1064.
8. Kanji, G.K. (2006) 100 statistical tests, SAGE Publications: London.
9. Pires, A.M., Amado, C. (2008) Interval Estimators for a Binomial Proportion: Comparison of Twenty Methods. REVSTAT – Statistical Journal, 6, 165-197.
10. Tang, W., He, H. and Tu, X. M. (2012) Applied categorical and count data analysis, Chapman & Hall/CRC: Bocaraton.
11. Vos, P.W. and Hudson, S. (2008) Problems with binomial Two-sided tests and the associated confidence intervals. Aust. N. Z. J. Stat., 50(1), 81–89.
12. Williamson, P.P. (2004) Comparison of testing procedures utilizing P-values and Bayes factors in some common situations. Journal of Statistical Computation and Simulation, 74, 833-850.

13. Zar, J.H. (2010) Biostatistical analysis, Prentice Hall/Pearson: NJ.
14. Zhou, X. H., Li, C.M. and Yang, Z. (2008) Improving interval estimation of binomial proportions. Phil. Trans. R. Soc. A, 366, 2405–2418.

Source of Support: None Declared
Conflict of Interest: None Declared

Table 1: Description of data sets which are extracted from various published literature related to the analyses of single binomial proportion.

Data Set No.	Data source	n	y	θ_0	Hypothesis	Classification
I	Zar (2010)	30	18	0.50	I	S_nS_x
II	Kimberlee (2006)	128	10	0.025	I	M_nS_x
III	Fleiss (2003)	257	23	0.05	I	M_nS_x
IV	http://math.etsu.edu/stats/testing_proportion.pdf	100	68	0.75	I	M_nM_x
V	Kanji (2006)	100	40	0.50	I	M_nM_x
VI	http://lhvarsitymath.com/Bock - HT for 1 Sample Proportion.pdf	1000	100	0.08	I	L_nM_x
VII	Gupta et al (1994)	1000	540	0.50	I	L_nL_x
VIII	Agresti (2007)	893	400	0.50	I	L_nL_x
IX	Gupta et al (1994)	20	18	0.85	II	S_nS_x
X	Tang et al (2012)	93	48	0.40	II	M_nM_x
XI	http://www2.latech.edu/~schroder/slides/stat/hyptest_prop.pdf	500	50	0.05	II	L_nM_x
XII	http://www.ltcconline.net/greenl/courses/201/hyptest/hypprob.htm	1500	153	0.10	II	L_nM_x
XIII	http://www.ltcconline.net/greenl/courses/201/hyptest/hypprob.htm	1000	550	0.50	II	L_nL_x
XIV	http://lhvarsitymath.com/Bock - HT for 1 Sample Proportion.pdf	352	17	0.04	II	L_nS_x
XV	http://math.etsu.edu/stats/testing_proportion.pdf	122	103	0.90	III	M_nM_x
XVI	http://sites.stat.psu.edu/~ajw13/stat200/Fall06/10_hypoth/02_hypoth.htm	129	37	0.40	III	M_nM_x
XVII	Zar (2010)	12	10	0.50	IV	S_nS_x
XVIII	http://www.kean.edu/~fosborne/bstat/07CPMop.html	423	18	0.05	IV	L_nS_x
XIX	http://www.biostat.umn.edu/~dipankar/BKMtry711.11/lecture_02.pdf	5	3	0.4	III	S_nS_x
XX	http://www.biostat.umn.edu/~dipankar/BKMtry711.11/lecture_02.pdf	5	3	0.4	II	S_nS_x
XXI	http://www.biostat.umn.edu/~dipankar/BKMtry711.11/lecture_02.pdf	5	3	0.4	I	S_nS_x
XXII	Agresti (2007)	10	9	0.5	I	S_nS_x
XXIII	Vos et al (2008)	63	7	0.05	I	M_nS_x
XXIV	Vos et al (2008)	9	1	0.44	I	S_nS_x
XXV	Agresti (2002)	25	0	0.5	I	S_nS_x

Table 2: P-value from different methods for the illustrated datasets covering small, medium and large characteristics of sample size, number of successes with different parameter values

Size	Dataset	Exact and Asymptotic methods							
		CPM	BKM	MPM	WCCM	WM	SCCM	SM	LRTM
Small n	I	0.3616	1.0000	0.2810	0.3515	0.2636	0.3613	0.2733	0.2717
	XXIV	0.0498	0.0877	0.0467	0.0081	0.0015	0.0940	0.0444	0.0295
	XXV	< 0.001	< 0.001	< 0.001	NC	NC	< 0.001	< 0.001	NC
Medium n	IX	0.4049	NA	0.2902	0.3547	0.2280	0.3771	0.2656	NA
	V	0.0569	0.1332	0.0460	0.0525	0.0412	0.0574	0.0455	0.0448
	XXIII	0.0374	0.0374	0.0504	0.1793	0.1227	0.0528	0.0260	0.0534
	XV	0.0345	NA	0.0266	0.0579	0.0448	0.0286	0.0201	NA
Large n	VI	0.0229	0.1151	0.0236	0.0398	0.0350	0.0230	0.0197	0.0244
	VII	0.0124	0.0937	0.0114	0.0122	0.0112	0.0125	0.0114	0.0114
	XIV	0.2476	NA	0.2109	0.2737	0.2339	0.2552	0.2135	NA

NA - Not Applicable; NC - Not Computable

Table 3: Dataset with characteristics of Medium-sized n and small-sized $y$and the P-values across methods exhibiting the discontinuities and non-monotonic behavior of Two-sided P-value.

Dataset		n	X	θ_0	θ	CPM	BKM	MPM	WCCM	WM	SCCM	SM	LRTM
XXIII	Original sample (O)	63	7	0.05	0.11	0.0374	0.0374	0.0504	0.1793	0.1227	0.0528	0.0260	0.0534
	One additional success (a)	64	8	0.05	0.13	0.0142	0.0142	0.0187	0.1041	0.0696	0.0137	0.0059	0.0196
	One additional failure (b)	64	7	0.05	0.11	0.0403	0.0778	0.0545	0.1863	0.1280	0.0584	0.0293	0.0577

NA - Not Applicable; NC - Not Computable

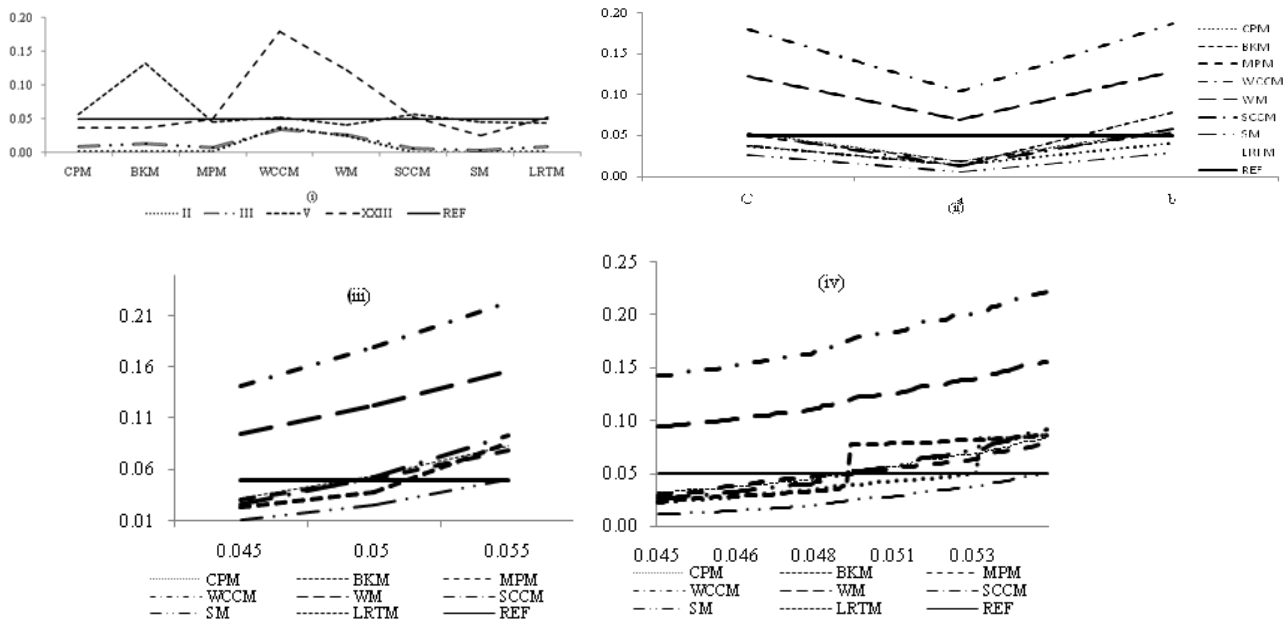

Figure 1: P-values from various methods for the dataset XXIII which exhibits (i) polarized conclusion, (ii) discontinuous and non-monotonic behavior with O represents the observed data, a- adding one success and b- adding one failure, (iii) portraying the P-value pattern using reasonable grid of varying equidistant values of $\theta_0$and (iv) using simulated values of θ_0.

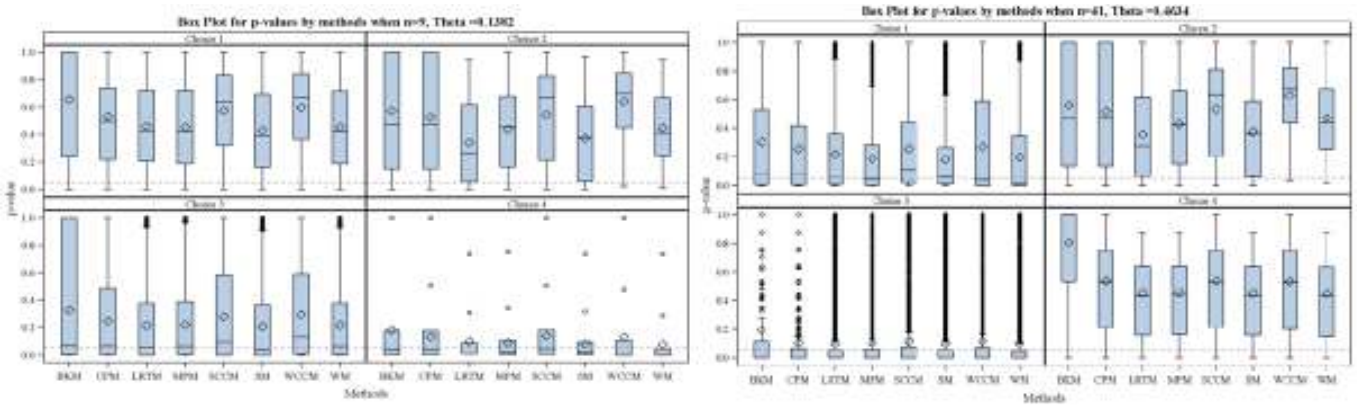

Figure 2: Box plot for p-values by methods under small and medium n with simulated values of θ and four different choices of θ_0, Choice 1- θ_0 around θ, Choice 2 - θ_0 from (0, θ), Choice 3 - θ_0 from (θ,1), Choice 4 - θ_0 towards symmetry.