Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 30

6-1-2020

A Note on Inferences About the Probability of Success

Rand Wilcox
University of Southern California, rwilcox@usc.edu

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP3296
doi: 10.22237/jmasm/1556670420

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

A Note on Inferences About the Probability of Success

Rand Wilcox
University of Southern California
Los Angeles, CA

There is an extensive literature dealing with inferences about the probability of success. A minor goal in this note is to point out when certain recommended methods can be unsatisfactory when the sample size is small. The main goal is to report results on the two-sample case. Extant results suggest using one of four methods. The results indicate when computing a 0.95 confidence interval, two of these methods can be more satisfactory when dealing with small sample sizes.

Keywords: binary data, binomial distribution, categorical data, linear contrasts

Introduction

Let p denote the probability of success associated with a binomial distribution. Schilling and Doi (2014) derived a method for computing a confidence interval for p such that the actual probability coverage is greater than or equal to the specified level. Moreover, their method is optimal in the sense that the shortest possible confidence interval is computed that guarantees that the probability coverage is at least $1 - \alpha$. However, a practical limitation is execution time quickly becomes prohibitive as the sample size increases. There are several alternative methods that might be used, which are reviewed in the next section of this paper. A minor goal in this note is to point out situations where these methods can perform poorly when dealing with the common goal of computing a 0.95 confidence interval and the sample size is less than 35. The main goal is to report results on the two-sample case. Extant results suggest using one of four techniques. The results indicate that when testing at the 0.05 level, two of these methods have a practical advantage.

doi: 10.22237/jmasm/1556670420 | Accepted: August 13, 2019; Published: June 1, 2020.
Correspondence: Rand Wilcox, rwilcox@usc.edu

Brief comments on linear contrasts, when there are more than two groups, are included.

Review of Extant Techniques for the One-Sample Case

Consider the one-sample case where the goal is to compute a $1 - \alpha$ confidence interval for p. Let w denoted the number of successes among n trials. A classic approach was derived by Clopper and Pearson (1934). The lower and upper ends of their $1 - \alpha$ confidence interval are $B(\alpha/2; w, n - w + 1)$ and $B(1 - \alpha/2; w + 1, n - w)$, respectively, where $B(q; u, v)$ is the qth quantile of a beta distribution with shape parameters u and v. It guarantees that the actual coverage probability is at least $1 - \alpha$, but in general does not give the shortest-length confidence interval.

Brown et al. (2002) compared various techniques and concluded that the Agresti–Coull method, which stems from Agresti and Coull (1998), performs relatively well. Let

$$\hat{p} = \frac{w}{n},$$

be the proportion of successes among the n observations and let c denote the $1 - \alpha/2$ quantile of a standard normal distribution. Let

$$\tilde{n} = n + c^2,$$

$$\tilde{w} = w + \frac{c^2}{2},$$

and

$$\tilde{p} = \frac{\tilde{w}}{\tilde{n}}.$$

The Agresti-Coull $1 - \alpha$ confidence interval for the probability of success, p, is

$$\tilde{p} \pm c\sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{\tilde{n}}}.$$

The Agresti-Coull method is, in essence, a simple approximation of the score method derived by Wilson (1927). Wilson's confidence interval is

$$\left(\hat{p}+c^2/\left(2n\right) \pm c\sqrt{\left[\hat{p}\left(1-\hat{p}\right)+c^2/\left(4n\right)\right]/n}\right)/\left(1+c^2/n\right).$$

Zou et al. (2009) review the literature in support of Wilson's method.

Results reported by Blyth (1986) suggest proceeding as follows when w is equal to 0, 1, $n-1$ or n. If $w = 0$,

$$c_U = 1 - \alpha^{1/n}$$

$$c_L = 0.$$

If $w = 1$,

$$c_L = 1-\left(1-\frac{\alpha}{2}\right)^{1/n}$$

$$c_U = 1-\left(\frac{\alpha}{2}\right)^{1/n}.$$

If $w = n - 1$,

$$c_L = \left(\frac{\alpha}{2}\right)^{1/n}$$

$$c_U = \left(1-\frac{\alpha}{2}\right)^{1/n}.$$

If w = n,

$$c_L = \alpha^{1/n}$$

and

$$c_U = 1.$$

The cases $w = 0$ and $w = n$ can be shown to be the Clopper-Pearson confidence intervals. Otherwise, Blyth recommends a method stemming from Pratt (1968), which is computed as follows. Let

$$A = \left(\frac{w+1}{n-2} \right)^2$$

$$B = 81(w+1)(n-w) - 9n - 8$$

$$C = -3c\sqrt{9(w+1)(n-w)(9n+5-c^2) + n + 1}$$

$$D = 81(w+1)^2 - 9(w+1)(2+c^2) + 1$$

$$E = 1 + A\left(\frac{B+C}{D} \right)^3$$

in which case the upper end of the confidence interval is

$$c_U = \frac{1}{E}.$$

As for the lower end, now let

$$A = \left(\frac{w}{n-w-1} \right)^2$$

$$B = 81(w)(n-w-1) - 9n - 8$$

$$C = -3c\sqrt{9x\left(n-w-1\right)\left(9n+5-c^2\right)+n+1}$$

$$D = 81w^2 - 9w\left(2+c^2\right)+1$$

$$E = 1+ A\left(\frac{B+C}{D}\right)^3$$

The lower end of the confidence interval is

$$c_L = \frac{1}{E}.$$

Kulinskaya et al. (2008, p. 140) derived yet another method for computing a confidence interval. Let $\breve{p} = (w + 3)/(n + 3/4)$,

$$A = \sin\left(\arcsin\left(\sqrt{\breve{p}}\right) - \frac{c}{2\sqrt{n}}\right)$$

and

$$B = \sin\left(\arcsin\left(\sqrt{\breve{p}}\right) + \frac{c}{2\sqrt{n}}\right)$$

where again c is the $1 - \alpha/2$ quantile of a standard normal distribution. Their $1 - \alpha$ confidence interval is (A^2,B^2). Evidently there are no published results on how this method compares to the other methods listed here.

Finally, there is the Schilling and Doi (2014) method, but for brevity the involved computational details are not described. But an R function that computes their confidence interval is described in the final section of this paper.

Review of Methods for the Two-Sample Case

Consider the two sample case where p_1 and p_2 are the probability of success associated with two independent groups and the goal is to test H_0: $p_1 = p_2$. The first

method described here was derived by Storer and Kim (1990). For the *j*th group, let r_j be the number of successes among n_j trials. The possible number of successes in the first group is any integer, x, between 0 and n_1, and for the second group it is any integer, y, between 0 and n_2. For any x and y, set

$$a_{xy} = 1$$

if

$$\left|\frac{x}{n_1} - \frac{y}{n_2}\right| \geq \left|\frac{r_1}{n_1} - \frac{r_2}{n_2}\right|;$$

otherwise

$$a_{xy} = 0.$$

Let

$$\hat{p} = \frac{r_1 + r_2}{n_1 + n_2}.$$

The test statistic is

$$T = \sum_{x=0}^{n_1}\sum_{y=0}^{n_2} a_{xy} b\left(x, n_1, \hat{p}\right) b\left(y, n_2, \hat{p}\right),$$

where

$$b\left(x, n_1, \hat{p}\right) = \left(\begin{array}{c} n_1 \\ x \end{array}\right) \hat{p}^x \left(1 - \hat{p}\right)^{n_1 - x},$$

and $b(y, n_2, \hat{p})$ is defined in an analogous fashion. The null hypothesis is rejected if

$$T \leq \alpha.$$

T is the p-value. The Storer-Kim method does not provide a confidence interval, but it seems it might offer a bit more power than other methods.

The confidence interval for $p_1 - p_2$, derived by Kulinskaya, Morgenthaler and Staudte (2010), is

$$\frac{\hat{w}}{u}\sin\left(\arcsin\left[\frac{u\hat{\Delta}+\hat{v}}{\hat{w}}\right]\pm c\sqrt{\frac{u}{2n_1n_2/N}}\right)-\frac{\hat{v}}{u},$$

where again c is the $1 - \alpha/2$ quantile of a standard normal distribution, r_1 and r_2 are the observed number of successes, $0 \le A \le 1$ is chosen by the user,

$$u = 2\left((1-A)^2\frac{n_2}{N}+A^2\frac{n_1}{N}\right),$$

$$\hat{\Delta} = (r_1+0.5)/(n_1+1)-(r_2+0.5)/(n_2+1),$$

$$\hat{\psi} = A(r_1+0.5)/(n_1+1)+(1-A)(r_2+0.5)/(n_2+1),$$

$$\hat{v} = (1-2\hat{\psi})\left(A-\frac{n_2}{N}\right), \text{ and}$$

$$\hat{w} = \sqrt{2u\hat{\psi}(1-\hat{\psi})+\hat{v}^2}.$$

Here, following the suggestion made by Kulinskaya et al. (2010), $A = 0.5$ is used.

The method derived by Zou et al. (2009) is applied as follows. Let (ℓ_j, u_j) be a $1 - \alpha$ confidence interval for p_j ($j = 1,2$). Following Zou et al., the confidence interval derived by Wilson (1927) is used. Then an approximate $1 - \alpha$ confidence interval for $p_1 - p_2$ is (L, U), where

$$L = \hat{p}_1 - \hat{p}_2 - \sqrt{(\hat{p}_1-\ell_1)^2+(u_2-\hat{p}_2)^2}$$

and

$$U = \hat{p}_1 - \hat{p}_2 + \sqrt{\left(u_1 - \hat{p}_1\right)^2 + \left(\hat{p}_2 - \ell_2\right)^2}.$$

Simulation

Consider the one-sample case and momentarily focus on the Clopper-Pearson (CP) and the Agresti-Coull (AC) methods. A simulation was performed for a sample size of $n = 25$ when computing a 0.95 confidence interval. This was done for $p = 0.05$ (0.01) 0.95, where for each value of p, 10,000 replications were used to estimate the actual value of α, the probability that the 0.95 confidence interval does not contain the p. Figure 1 shows a plot of the results, where the dashed line is for method AC and the solid line is method CP. Bradley (1978) has suggested that as a general guide, when $\alpha = 0.05$, the actual value should be between 0.025 and 0.075. The top horizontal line in Figure 1 is 0.075 and the bottom line is 0.025. As can be seen, AC performs reasonably well for $0.2 \leq p \leq 0.8$, but otherwise it can be highly unsatisfactory. Moreover, concerns persist when using Pratt's (P) method, the method derived by Kulinskaya et al. (KMS) as well as Wilson's (WIL) method. Consider, for example $p = 0.15$. Based on a simulation with 50,000 replications, the actual value of α is 0.112, 0.113, 0.024, 0.113, and 0.117 for methods AC, P, CP, KMS and WIL, respectively. As for the method derived by Schilling and Doi (SD), the actual level was estimated to be 0.044 (based on 5000 replications) consistent with results reported in their paper. Increasing the sample size to $n = 30$ the estimates for AC, P, CP, KMS and WIL are 0.073, 0.073 0.024, 0.073 and 0.077, respectively. For $n = 35$ the estimates are 0.053, 0.053, 0.037, 0.053 and 0.053.

Consider the two-sample case where the goal is to test H_0: $p_1 = p_2$ at the 0.05 level. Table 1 shows estimates of the actual Type I error probability for various sample sizes when $p_1 = p_2 = 0.05$, 0.10, 0.15, 0.25 and 0.50. As can be seen, when $n_1 = n_2 = 10$, the actual levels for all four methods are less than the nominal level. For $p = 0.05$ and 0.10, all four methods have levels less than 0.025. Overall, Beal's method is the least satisfactory. A feature of methods SK and KMS is that the actual level never exceeds the nominal level. There are situations where ZHZ performs better than the the other methods, but for $n_1 = n_2 = 20$ and $p_1 = p_2 = 0.15$ the actual level is 0.08. Increasing the sample sizes to $n_1 = n_2 = 30$, the level is estimated to be 0.065 and for $n_1 = n_2 = 35$ it is 0.054. For $p_1 = p_2 = 0.2$ and $n_1 = n_2 = 20$ the Type I error probability is 0.074. In general, ZHZ might perform reasonably well when the minimum sample size is less than 35, but the extent this is true depends on the unknown probabilities. If the goal is to keep the actual Type I error probability close to or less than the nominal level, SK and KMS are safer than ZHZ when either

sample size is small. As previously noted, method ZHZ is based on confidence intervals for p_1 and p_2 using Wilson's method. Replacing Wilson's confidence interval with the Agresti-Coull confidence interval does not improve matters. Replacing Wilson's confidence interval with the Schilling-Doi confidence interval does improve the ability of ZHZ to avoid Type I error probabilities greater than the nominal level, but now ZHZ has no practical advantage over SK and KMS.

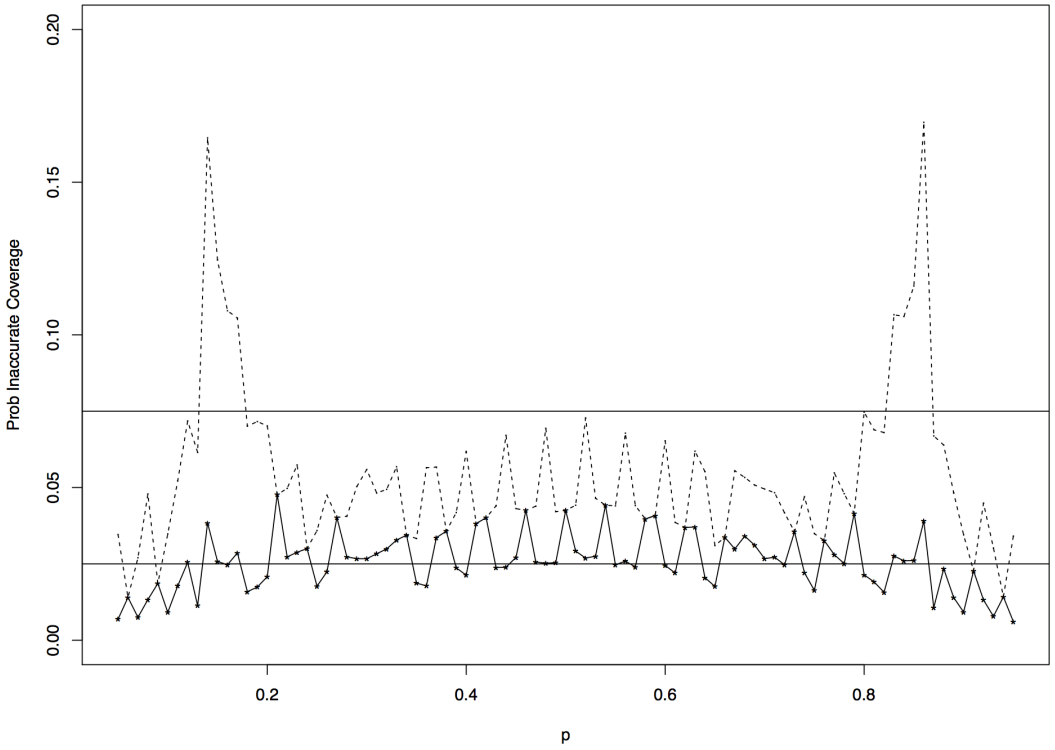

Figure 1. The probability of an inaccurate confidence interval when using the Agresti-Coull method (dotted line) and the Clopper-Pearson method (solid line).

Table 2 reports estimated power for various sample sizes and choices for p_1 and p_2. Note that ZHZ tends to have the highest power, even in situations where it avoids Type I errors well above the nominal level. Beal's method is the least satisfactory and SK tends to have about the same or higher power than KMS.

Table 1. Estimated Type I error probabilities when testing H_0: $p_1 = p_2$ at the 0.05 level.

	p	SK	Beal	KMS	ZHZ
$n_1 = 10, n_2 = 10$	0.05	0.001	0.000	0.001	0.001
	0.10	0.008	0.001	0.008	0.010
	0.15	0.023	0.006	0.023	0.029
	0.25	0.032	0.016	0.032	0.058
	0.50	0.041	0.040	0.041	0.053
$n_1 = 10, n_2 = 30$	0.05	0.024	0.003	0.003	0.023
	0.10	0.031	0.006	0.009	0.029
	0.15	0.037	0.016	0.023	0.039
	0.25	0.043	0.028	0.037	0.068
	0.50	0.043	0.027	0.047	0.059
$n_1 = 20, n_2 = 20$	0.05	0.002	0.002	0.002	0.014
	0.10	0.018	0.012	0.014	0.057
	0.15	0.037	0.023	0.025	0.080
	0.25	0.048	0.024	0.040	0.061
	0.50	0.042	0.021	0.040	0.044
$n_1 = 10, n_2 = 100$	0.05	0.051	0.006	0.006	0.031
	0.10	0.043	0.005	0.010	0.033
	0.15	0.042	0.011	0.019	0.039
	0.25	0.044	0.039	0.053	0.065
	0.50	0.048	0.026	0.049	0.046

Table 2. Estimated power.

	p_2	SK	Beal	KMS	ZHZ
$n_1 = 10, n_2 = 10, p_1 = 0.3$	0.05	0.215	0.096	0.215	0.260
	0.10	0.144	0.074	0.144	0.184
$n_1 = 10, n_2 = 10, p_1 = 0.5$	0.05	0.631	0.509	0.631	0.704
	0.30	0.128	0.124	0.128	0.167
$n_1 = 10, n_2 = 30, p_1 = 0.3$	0.05	0.522	0.309	0.361	0.535
	0.10	0.330	0.163	0.230	0.333
$n_1 = 10, n_2 = 30, p_1 = 0.5$	0.05	0.892	0.748	0.812	0.889
	0.30	0.201	0.116	0.162	0.224
$n_1 = 20, n_2 = 20, p_1 = 0.3$	0.05	0.535	0.444	0.470	0.593
	0.10	0.340	0.248	0.290	0.366
$n_1 = 20, n_2 = 20, p_1 = 0.5$	0.05	0.942	0.894	0.928	0.949
	0.30	0.218	0.164	0.218	0.244

Illustration

The results are illustrated using data from a study dealing with shoulder pain after surgery (Jorgensen et al., 1995). There were two independent groups. The first received an active treatment and the other was a control group. Shoulder pain was measured at three different times after surgery using integer values ranging between 1 (low pain) and 5. First focus on the control group and consider the issue of whether pain decreased between the first and third times pain was measured. This was tested using the sign test based on method SD. The sample size is 19, but after eliminating tied values, the sample size was 10. The estimate of p, the probability that pain is lower at time 1, was 0.7, the 0.95 confidence interval was (0.381, 0.913) and the p-value for H_0: $p = 0.5$ was 0.35. Using method AC instead, the 0.95 confidence interval was (0.392, 0.897) and the p-value was 0.22, the only point being that the choice of method can make a difference. For the treated group, the estimate of p was 0.09, the 0.95 confidence interval was (0.005, 0.404) and the p-value was 0.02 using SD. Now the sample size is 11. Using AC, the p-value was 0.001.

Consider the goal of comparing the two groups in terms of p_1, the probability that pain is rated 1 at time 3 for the treated group, and p_2, the probability that pain is rated 1 at time 3 for the control group. The estimates of p_1 and p_2 were 0.818 and 0.263, respectively. The p-value using method SK was 0.0003 versus 0.001 using KMS.

Conclusion

For small sample sizes, when dealing with the one-sample case, the Schilling-Doi method offers a distinct advantage. However, execution time becomes an issue as the sample size increases. A simple strategy is to use the Schilling-Doi method when $n < 35$. For $n \geq 35$, the choice of methods appears to make little or no difference when computing with a 0.95 confidence interval.

As for the two-sample case, there are situations where ZHZ performs well in terms of accurate probability coverage and offers a power advantage over the other methods considered here. However, when dealing small sample sizes, there are situations where it is unsatisfactory. If the goal is to avoid Type I errors well above the nominal level, when testing H_0: $p_1 = p_2$, the results suggest using SK or KMS. Method SK might offer a power advantage but at the expense of no confidence interval.

Zou et al. (2009) derived a generalization of ZHZ for testing linear contrasts for $J \geq 2$ groups. A few simulations were performed when $J = 3$ and $J = 4$ when the sample sizes are small. In contrast to $J = 2$, all indications are that it provides good control over the Type I error probability. That is, results reported here for the one-sample case suggest that ZHZ might be unsatisfactory when dealing with small sample sizes. This was found to be the case when for $J = 2$, but not when $J > 2$.

The R function `binom.conf` applies all of the methods considered here for the one-sample case and is available in the file `Rallfun-v36` at https://dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm. It defaults to the Schilling-Doi method if $n < 35$, otherwise the Agresti-Coull method is used. But the other methods can be used via the argument method. The R function `binom2g` deals with the two-sample case. It defaults to the KMS method, which provides a confidence interval. To use method SK, with the possibility of more power at the expense of no confidence interval, set the argument `method = 'SK'`. For linear contrasts, the R function `lincon.bin` can be used.

References

Agresti, A. & and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *American Statistician, 54*(4), 280-288. https://doi.org/10.1080/00031305.2000.10474560

Agresti, A. & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician, 52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Beal, S. L. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics, 43*(4), 941-950. https://doi.org/10.2307/2531547

Blyth, C. R. (1986). Approximate binomial confidence limits. Journal of the *American Statistical Association, 81*(395), 843–855. https://doi.org/10.1080/01621459.1986.10478343

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Brown, L. D., Cai, T. T. & DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Annals of Statistics, 30*(1), 160-201. https://doi.org/10.1214/aos/1015362189

Clopper, C. & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*(4), 404–413. https://doi.org/10.1093/biomet/26.4.404

Jorgensen, J. O., Gilles, R. B., Hunt, D. R., Caplehorn, J. R. M., & Lumley, T. (1995). A simple and effective way to reduce postoperative pain after laparoscopic cholecystectomy. *Australian and New Zealand Jouranl of Surgery, 65*(7), 466-469. https://doi.org/10.1111/j.1445-2197.1995.tb01787.x

Kulinskaya, E., Morgenthaler, S. & Staudte, R. G. (2008). *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. New York: Wiley. https://doi.org/10.1002/9780470985533

Kulinskaya, E., Morgenthaler, S. & Staudte, R. G. (2010). Variance stabilizing the difference of two binomial proportions. *American Statistician, 64*(4), 350-356. https://doi.org/10.1198/tast.2010.09080

Pratt, J. W. (1968). A normal approximation for binomial, F, beta, and other common, related tail probabilities, I. *Journal of the American Statistical Association, 63*(324), 1457–1483. https://doi.org/10.1080/01621459.1968.10480939

Schilling, M. & Doi, J. (2014). A coverage probability approach to finding an optimal binomial confidence procedure. *American Statistician, 68*(3), 133–145. https://doi.org/10.1080/00031305.2014.899274

Storer, B. E. & Kim, C. (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association, 85*(409), 146-155. https://doi.org/10.1080/01621459.1990.10475318

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association, 22*(158), 209-212. https://doi.org/10.1080/01621459.1927.10502953

Zou, G.Y., Huang, W. & Zhang, X. (2009). A note on confidence interval estimation for a linear function of binomial proportions. *Computational Statistics & Data Analysis, 53*(4), 1080-1085. https://doi.org/10.1016/j.csda.2008.09.033