Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 5

7-17-2020

Identifying Which of *J* Independent Binomial Distributions Has the Largest Probability of Success

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

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Wilcox, R. (2019). Identifying which of J independent binomial distributions has the largest probability of success. Journal of Modern Applied Statistical Methods, 18(2), eP3359. doi: 10.22237/jmasm/1604190960

Journal of Modern Applied Statistical Methods
November 2019, Vol. 18, No. 2, eP3359.
doi: 10.22237/jmasm/1604190960

בס"ד
Copyright © 2020 JMASM, Inc.
ISSN 1538 − 9472

INVITED ARTICLE

Identifying Which of *J* Independent Binomial Distributions Has the Largest Probability of Success

Rand Wilcox
University of Southern California
Los Angeles, CA

Let $p_1,\ldots, p_J$ denote the probability of a success for J independent random variables having a binomial distribution and let $p_{(1)} \leq \ldots \leq p_{(J)}$ denote these probabilities written in ascending order. The goal is to make a decision about which group has the largest probability of a success, $p_{(J)}$. Let $\hat{p}_1,\ldots, \hat{p}_J$ denote estimates of $p_1,\ldots,p_J$, respectively. The strategy is to test $J-1$ hypotheses comparing the group with the largest estimate to each of the $J-1$ remaining groups. For each of these $J-1$ hypotheses that are rejected, decide that the group corresponding to the largest estimate has the larger probability of success. This approach has a power advantage over simply performing all pairwise comparisons. However, the more obvious methods for controlling the probability of one more Type I errors perform poorly for the situation at hand. A method for dealing with this is described and illustrated.

Keywords: Binary data, binomial distribution, rank and selection, multiple comparisons

Introduction

Consider J independent groups and let $\theta_1,\ldots, \theta_J$ denote a parameter of interest. Let $\theta_{(1)} \leq \ldots \leq \theta_{(J)}$ in ascending order. There is interest determining which group corresponds to $\theta_{(J)}$. For example, such as which group has the largest median. Let $\hat{\theta}_j$ ($j = 1,\ldots, J$) denote an estimate of θ_j based on a random sample of size n_j and denote the estimates written ascending order by $\hat{\theta}_{(1)} \leq \ldots \leq \hat{\theta}_{(J)}$. The objective of ranking and selection methods is to determine the sample size needed to ensure that

doi: 10.22237/jmasm/1604190960 | Accepted: December 1, 2019; Published: July 17, 2020.
Correspondence: Rand Wilcox, rwilcox@usc.edu

$\hat{\theta}_{(J)}$ corresponds to $\theta_{(J)}$. Focus on the situation where $\theta_{(1)} = \ldots = \theta_{(J-1)}$ and $\theta_{(J)} - \theta_{(J\text{-}1)} = \delta$, where δ is a constant. This is the indifference zone approach. Bechhofer (1954) addressed this issue, assuming that observations are randomly sampled from normal distributions having a common known variance σ^2 and that the goal is to identify the group with the largest population mean. For a variety of situations, the sample size can be determined so that the probability of a correct decision (PCD), meaning the probability that the group corresponding to $\theta_{(J)}$ has the largest estimate $\hat{\theta}_{(J)}$, is equal to some specified value, β (e.g., Bechhofer, Dunnett, & Sobel, 1954; Bechhofer, Kiefer, & Sobel, 1968; Dudewicz & Dalal, 1975; Rinott, 1978; Gibbons et al., 1987; Gupta & Panchapakesan, 1987; Mukhopadhyay & Solanky, 1994). When dealing with means, and the variances are unknown, two-stage procedures are used. The goal of the first stage is to get an estimate of the variances, which can be used to determine the required sample size.

Consider the special case of identifying which of J independent variables, each having a binomial distribution, has the largest probability of success. The goal is to suggest a method that does not require the specification of an indifference zone. The approach has obvious similarities to comparing groups to a control group. In particular, compare the group with the largest estimate to each of the remaining groups. A seemingly simple approach to controlling the familywise error rate (FWE), meaning the probability of one or more Type I errors, is to use the Bonferroni method. That is, perform each of the $J - 1$ tests at the $\alpha / (J - 1)$ so that FWE will be at most α. But preliminary simulations clearly indicated that the actual level can be substantially higher than the nominal level. This is the case when $J = 4$ and $n = 40$. For $J = 8$, this approach can be unsatisfactory even with $n = 100$. Evidently, the difficulty is controlling the Type I error probability of the individual tests when $\alpha / (J - 1)$ gets too close to zero. Using improvements on the Bonferroni method (e.g. Hochberg, 1988; Hommel, 1988) does not eliminate this concern.

The Proposed Approach

Let X_j denote a random variable having a binomial distribution with probability of success p_j. That is, for J independent groups, X_j denotes the number of successes associated with the j^{th} group based on be a random sample of size n_j. First consider the basic problem of comparing two independent binomial distributions. Numerous methods were proposed. Based on results reported by Wilcox (2020), a method derived by Kulinskaya et al. (2010) is used here, which will be called method KMS henceforth. Their confidence interval for $p_1 - p_2$ is given by

$$\frac{\hat{w}}{u}\sin\left(\arcsin\left[\frac{u\hat{\Delta}+\hat{v}}{\hat{w}}\right]\pm c\sqrt{\frac{u}{2n_1n_2/N}}\right)-\frac{\hat{v}}{u}, \tag{1}$$

where c is the $1-(\alpha/2)$ quantile of a standard normal distribution, $0\leq A\leq 1$ is chosen by the user,

$$u=2\left((1-A)^2\frac{n_2}{N}+A^2\frac{n_1}{N}\right), \hat{\Delta}=\frac{X_1+0.5}{n_1+1}-\frac{X_2+0.5}{n_2+1},$$
$$\hat{\psi}=A\frac{X_1+0.5}{n_1+1}+\frac{(1-A)(X_2+0.5)}{n_2+1}, \hat{v}=(1-2\hat{\psi})\left(A-\frac{n_2}{N}\right),$$
$$\hat{w}=\sqrt{2u\hat{\psi}(1-\hat{\psi})+\hat{v}^2},$$

and $N = n_1 + n_2$. Here, following the suggestion made by Kulinskaya et al. (2010), $A = 0.5$ is used.

For the situations considered by Wilcox (2020), a method derived by Storer and Kim (1990) was found to have a power advantage over the method derived by Kulinskaya et al. (2010) at the expense of no confidence interval. For the situation at hand, however, no advantage was found using the Storer and Kim method so for brevity no details are provided.

Consider the goal of making a decision about which group has the largest probability of success. Let $\hat{p}_j = X_j / n_j$ ($j = 1,\ldots, J$). Put these estimates in ascending order yielding $\hat{p}_{(1)} \leq \ldots \leq \hat{p}_{(J)}$. Let $p_{\pi(j)}$ denote the probability of success associated with $\hat{p}_{(j)}$. The basic idea is to test

$$\mathrm{H}_0 : p_{\pi(j)} = p_{\pi(J)} \tag{2}$$

for each j, $j = 1,\ldots, J-1$. For each j for which (2) is rejected, decide that the group corresponding to $\hat{p}_{(J)}$ has a higher probability of success. If all $J-1$ hypotheses are rejected, decide that the group corresponding to $\hat{p}_{(J)}$ is the group with the largest probability of success, $p_{(J)}$. Otherwise no decision is made.

Consider $p_1 = \ldots = p_J = p$ and testing (2) for each $j < J$. Let P_j denote the p-value based on the KMS method and suppose the j^{th} hypothesis is rejected if $P_j \leq c_j$. Consider the goal of choosing c_j so that the probability of a Type I error is 0.05. A simulation based on 5000 replications was used to determine c_j when $J = 4$, $p = 0.5$, and $n = 35$. The result was $(c_3, c_2, c_1) = (0.094, 0.0345, 0.006)$. So, in particular,

when comparing the group with the largest estimate to the group with the second largest estimate, to achieve a Type I error probability equal to 0.05, reject when the p-value is less than or equal to 0.094. When comparing the group with the largest estimate to the group with the third largest estimate, reject if the p-value is less than or equal to 0.0345. For the same situation except now $p = 0.1$, the result was $(c_3, c_2, c_1) = (0.122, 0.051, 0.011)$.

The difficulty, of course, is p is not known and there is the additional issue of controlling FWE. An outline of the proposed strategy is as follows. First, estimate p assuming $p_1 = \ldots = p_J = p$. Based on this estimate of p and the sample sizes, use a simulation to estimate the critical p-values $(c_{J-1},\ldots, c_1)$ so that the Type I error probability for each individual test is α. Finally, replace the critical p-values with $(d_{J-1},\ldots, d_1) = f(c_{J-1},\ldots, c_1)$, where the constant f is chosen so that FWE is equal to α. That is, reject the j^{th} hypothesis if the corresponding p-value is less than or equal d_j. A simple choice for f is $f = 1 / (J - 1)$. That is, use the Bonferroni method. Here, however, a refinement of this approach is used.

Let $\hat{p} = \Sigma X_j / \Sigma n_j$ be the estimate of p. The critical p-values $(c_{J-1},\ldots, c_1)$ are determined by first generating Y_j successes $(j = 1,\ldots, J)$ when the probability of success is $\hat{p}$ and the sample size is n_j. Next, compute a p-value for each j, and repeat this B times. This results in a B-by-$(J - 1)$ matrix of p-values, $\mathbf{P}$. The columns of $\mathbf{P}$ yield estimates of $c_{J-1},\ldots, c_1$. The value of c_j is estimated via some quantile estimator applied to the j^{th} column. That is, the estimated α quantile is the estimate of c_j. Moreover, this matrix of p-values can be used to determine f such that

$$P\left(p_1 \le d_1,\ldots, p_{J-1} \le d_{J-1}\right) = \alpha . \tag{3}$$

Critical p-values can estimated in a manner that takes into account their multivariate distribution.

Given a value for f, let C_i be equal to one if for the i^{th} row of $\mathbf{P}$ it is simultaneously the case that $P_{i,j} \le d_j$ for each $j = 1,\ldots, J - 1$; otherwise $C_i = 0$. Let $D = \Sigma C_i$, where the estimate of FWE is D / B. An approximate way of controlling FWE is to choose f such that D / B is equal to some specified value, α. Here, the R function optim was used to estimate f using the Brent method. This will be called method ECP henceforth.

There is a variation of method ECP that deserves consideration. Proceed as just described, but rather than estimate $(c_{J-1},\ldots, c_1)$ based on the matrix of p-values, simply set $c_1 = \ldots = c_{J-1} = \alpha$ and then determine f satisfying (3). This will be called method EQA.

It might seem the matrix **P** can be used to compute a type of *p*-value that quantifies the strength of the empirical evidence that a decision can be made about which group has the largest probability of success. Let $\mathbf{P}_i$ denote the i^{th} row of **P**. Define the indicator function $I(\mathbf{P}_i) = 1$ if $P_{ij} \leq p_j$ for each $j = 1,\ldots, J-1$; otherwise $I(\mathbf{P}_i) = 0$. Then a type of *p*-value is

$$\frac{1}{L}\sum I_i \tag{4}$$

However, in terms of controlling the probability of one or more Type I errors, simulations indicate that using this *p*-value is unsatisfactory. A more satisfactory approach is to compute the d_j values for $\alpha = 0.001(0.001)0.1(0.01)0.99$ and then determine the smallest α value for which all $J-1$ hypotheses are rejected.

Results

Simulations were used to study the small sample properties of methods ECP and EQA. Table 1 reports the estimate of one or more Type I errors when the goal is to have FWE equal to 0.05 and the *J* groups have a common probability of success, *p*. The choices for *p* were 0.10, 0.15, 0.20, 0.30, 0.40, and 0.50. Table 1 shows the results for a common sample size of $n = 20$ and 40, and $J = 4$ groups. Also shown are results for eight groups and $n = 20$. The estimates are based on 2000 replications. Although the seriousness of a Type I error can depend on the situation, Bradley (1978) suggests that as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. As can be seen, the estimates for ECP

Table 1. Estimates of the FWE rate when testing at the 0.05 level

J	*n*	*p*	ECP	EQA	*p*	ECP	EQA
4	20	0.10	0.051	0.038	0.3	0.046	0.055
	20	0.15	0.054	0.048	0.4	0.051	0.049
	20	0.20	0.061	0.053	0.5	0.053	0.042
	40	0.10	0.054	0.054	0.3	0.050	0.051
	40	0.15	0.052	0.048	0.4	0.046	0.049
	40	0.20	0.050	0.053	0.5	0.045	0.042
8	20	0.10	0.049	0.054	0.3	0.047	0.056
	20	0.15	0.057	0.051	0.4	0.048	0.049
	20	0.20	0.053	0.054	0.5	0.053	0.104

range between 0.045 and 0.061. For $n = 40$ and $J = 4$, the estimates ranged between 0.045 and 0.054. Even for $n = 20$ and $J = 8$, control over FWE is very good. That is, all indications are that for method ECP, Bradley's criterion is met. For method EQA and $J = 4$, no estimate exceeds 0.055 and the lowest estimate of 0.038. However, for $J = 8$, the estimate when $p = 0.5$ is 0.104, and it is 0.11 when using the Bonferroni method, both of which are unsatisfactory based on Bradley's criterion. Increasing the sample sizes to 40, the estimate was 0.086 using EQA. Some additional simulations were run with $n = 10$. The FWE decreases from those values in Table 1.

A few simulations were conducted comparing the power of ECP versus and EQA. By power is meant the probability that all $J - 1$ hypotheses are rejected when $p_{(J)} > p_{(J-1)}$. First consider $J = 4$, $n = 20$, $p_1 = p_2 = p_3 = 0.2$, and $p_4 = 0.5$. The power for method ECP was estimated to be 0.605 and for method EQA it was 0.502. For $p_1 = 0.5$, $p_2 = 0.6$, $p_3 = 0.7$, and $p_4 = 0.8$, methods ECP and EQA have power 0.368 and 0.387, respectively. ECP does not dominate, but all indications are that generally ECP is better than EQA.

An Illustration

Methods ECP and EQA are illustrated using data from the Well Elderly II study (Clark et al., 2012), which was generally aimed at improving the physical and mental well-being of older adults. A portion of this study measured depressive symptoms (CESD) before intervention. Here, the focus is on CESD measures for five groups based on education: less than high school, high school graduate, some college or technical school, four years of college completed and post-graduate study. The sample sizes are 136, 89, 158, 48, and 29, respectively. CESD scores greater than 15 are considered an indication of mild depression or worse. The focus here is on the probability of mild depression or worse. The estimates for these five groups were 0.485, 0.326, 0.297, 0.271, and 0.241, respectively. So, the highest estimate occurred for the first group. Using method ECP, the results indicated that group 1 has a higher probability than groups 2, 3 and 4 when testing at the 0.05 level. The p-values comparing group 1 to groups 2-5 were 0.018, 0.002, 0.009, and 0.014, and the corresponding critical p-values were 0.0808, 0.0337, 0.0139, and 0.0040, respectively. So, no decision can be made regarding group 5. The *p*-value associated with making a decision about which group has the highest probability is 0.0776. EQA rejects for group 3 only.

Conclusion

All indications are that method ECP performs relatively well. Method EQA might seem like the more natural way to proceed, generally it competes reasonably well with method ECP, but situations were found where ECP offers a clear advantage and no situation was found where the reverse is true. Of course, in some situations, all pairwise comparisons might be more relevant rather than determining which group has the highest probability. But if making a decision about which has the highest probability is the main goal, it is evident that method ECP offers an advantage in power simply because FWE is being controlled for a smaller number of hypotheses.

The ranking and selection literature deals with a range of issues related to this paper. For example, determine which of J dependent groups has the largest mean, or which has the smallest variance. Which cell of a multinomial distribution has the largest probability? The notion of an indifference plays a crucial role in these classic techniques. Thanks to modern computing power, it might be possible to address these issues in new and interesting ways.

Finally, the R function bin.best.PV applies method ECP and is stored in the file Rallfun-v37, which can be downloaded from https://dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm. Included is a p-value based on the strategy of computing computes the d_j values for $\alpha = 0.001(0.001)0.1(0.01)0.99$ and then determining the smallest α value for which all $J - 1$ hypotheses are rejected. This quantifies the strength of decision about which group has the largest trimmed mean. However, it does not reflect the probability that that a correct decision was made.

References

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics, 25*(1), 16-39. doi: 10.1214/aoms/1177728845

Bechhofer, R. E., Dunnett, C. W., & Sobel, M. (1954). A two-sample multiple decision procedure for ranking means of normal populations with a common unknown variance. *Biometrika, 41*(1-2), 170-176. doi: 10.1093/biomet/41.1-2.170

Bechhofer, R. E., Kiefer, J., & Sobel, M. (1968). *Sequential identification and ranking procedures*. Chicago, IL: University of Chicago Press.

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

Dudewicz, E. J., & Dalal, S. R. (1975). Allocation of observations in ranking and selection with unequal variances. *Sankhyā: The Indian Journal of Statistics, Series B, 37*(1), 27-78.

Gibbons, J., Olkin, I., & Sobel, M. (1987). *Selecting and ordering populations: A new statistical methodology*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Gupta, S. S., & Panchapakesan, S. (1987). *Multiple decision procedures: Theory and methodology of selecting and ranking populations*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800-802. doi: 10.1093/biomet/75.4.800

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383-386. doi: 10.1093/biomet/75.2.383

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

Mukhopadhyay, N., & Solanky, T. (1994). *Multistage selection and ranking procedures: Second order asymptotics*. New York: Marcel Dekker.

Rinott, Y. (1978). On two-stage selection procedures and related probability inequalities. *Communications in Statistics – Theory and Methods, 7*(8), 799-811. doi: 10.1080/03610927808827671

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. doi: 10.1080/01621459.1990.10475318

Wilcox, R. R. (2020). A note on inferences about the probability of success. *Journal of Modern Applied Statistical Methods, 18*(1), eP3296. doi: 10.22237/jmasm/1556670420