Ann. Inst. Statist. Math.
30 (1978), Part A, 421-427

TESTING FOR THE EQUALITY OF TWO BINOMIAL PROPORTIONS

NOEL CRESSIE

(Received June 12, 1978)

Abstract

There are two statistics one might choose when testing whether two binomial probabilities are the same. This note provides a large sample answer to Robbins' question of which is preferable.

1. The question

Robbins [4] has asked the following " fundamental question of practical statistics ": Suppose in a set of m Bernoulli trials governed by p_1, we observe x successes, and in another independent set of n trials governed by p_2, we observe Y successes. To test the hypothesis H_0: $p_1=p_2$, one of two statistics is usually chosen.

Now,

$$V\equiv\frac{X/m-Y/n}{\{(X/m)(1-X/m)/m+(Y/n)(1-Y/n)/n\}^{1/2}}\,, \tag{1.1}$$

will be (for m and n large) approximately $N(0, 1)$ under H_0, and so the standard normal tables can be used to obtain an *approximate* level α test. But,

$$U\equiv\frac{X/m-Y/n}{[\{(X+Y)/(m+n)\}\{1-(X+Y)/(m+n)\}\{1/m+1/n\}]^{1/2}}\,, \tag{1.2}$$

is also (for m and n large) approximately $N(0, 1)$ under H_0. (The quantity U^2 is in fact the goodness of fit statistic used for testing homogeneity in a 2×2 table.)

Which of these two procedures is better with respect to power, against the various possible alternatives to H_0?

Eberhardt and Fligner [1] considered two sided alternatives H_1: $p_1\neq p_2$, and found conditions on (p_1, p_2) and $m/(m+n)$ for which the denominator of (1.1) is less than the denominator of (1.2); see their Figure A for a summary. Their large sample techniques compare the

Key words: Binomial distribution, Pitman alternatives, power, test statistic.

two test statistics via approximate Bahadur efficiency, however our approach will be through Pitman efficiencies, and a sequence of local alternatives. We will concentrate on one sided alternatives, $H_1 : p_1 > p_2$, or $H_1 : p_1 < p_2$. Section 2 looks at the large sample comparison of U and V and makes the recommendations displayed in Table 1.

A criticism that could be made of the work of Section 2 and that of Eberhardt and Fligner is that although the two test statistics are compared for power, there has been no effort to initially *match* their Type I errors. Hence we may detect V to be more powerful than U under certain alternatives, but this could be due to the Type I error of V being higher than that of U. We could reason that regardless of which test is used, in practice, the critical region is found from the standard normal tables, and hence it makes sense to compare the two tests with identical critical regions. To be fair, Eberhardt and Fligner did do some numerical work that compared differences in power to differences in Type I error, however they presented no mathematical arguments that allowed one to do it in general. Section 3 does this for large samples, and no different conclusions to those of Section 2 are reached.

2. An answer

Because we know that both U and V are asymptotically $N(0, 1)$, we will try to compare their rates of approach to normality. Now if $p_1 - p_2$ is fixed, and non zero, then the power of both tests can be made as close to 1 as desired, by choosing m and n large enough. Pitman's remedy (see Fraser [2], p. 108) was to choose a sequence of alternatives approaching the null hypothesis in such a way that the limiting power approached a limit lying somewhere between the Type I error α, and 1. Two tests can then be compared via their respective limits.

In our case, the sequence of alternatives to use is:

$$H_m : p_1 = p\ , \qquad p_2 = p + (\Delta / m^{1/2})\ . \tag{2.1}$$

Now let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution, and define k_α by: $1 - \Phi(k_\alpha) = \alpha$. Also assume,

$$m = \lambda^2 n\ , \qquad \lambda^2 \text{ fixed}\ . \tag{2.2}$$

Then it is easy to show that for $\Delta < 0$, *both* the test based on V and the test based on U, have limiting power,

$$\begin{aligned}\lim_{m\to\infty} \Pr\{V \geqq k_\alpha \mid H_m\} &= \lim_{m\to\infty} \Pr\{U \geqq k_\alpha \mid H_m\} \\ &= 1 - \Phi(k_\alpha + \Delta(1+\lambda^2)^{-1/2} p^{-1/2} q^{-1/2})\ ;\end{aligned}$$

i.e. Pitman's approach is not sensitive enough to distinguish between the two tests. In what follows, we will stay with the sequence of alternatives $\{H_m\}$, but will look for *higher order terms* in expansions of V and U.

Define,

$$(2.3) \qquad Z_1 \equiv (X - mp_1)/(mp_1q_1)^{1/2}, \qquad Z_2 \equiv (Y - np_2)/(np_2q_2)^{1/2},$$

where $q_1 \equiv 1 - p_1$, $q_2 \equiv 1 - p_2$. Then writing V in terms of Z_1 and Z_2 we get,

$$\begin{aligned} V = \{&(p_1q_1/m)^{1/2}Z_1 - (p_2q_2/n)^{1/2}Z_2 - (p_2 - p_1)\} \\ /\{&m^{-1}(p_1 + Z_1(p_1q_1/m)^{1/2})(q_1 - Z_1(p_1q_1/m)^{1/2}) \\ &+ n^{-1}(p_2 + Z_2(p_2q_2/n)^{1/2})(q_2 - Z_2(p_2q_2/n)^{1/2})\}^{1/2}. \end{aligned}$$

Now under H_m,

$$p_1q_1 = pq, \qquad p_2q_2 = pq + (1 - 2p)\Delta/m^{1/2} - \Delta^2/m.$$

Ignoring terms of $O(m^{-1})$ or higher,

$$\begin{aligned} (2.4) \qquad V = \{&Z_1 - \lambda Z_2 - \lambda Z_2\Delta(1/2 - p)m^{-1/2}p^{-1}q^{-1} - \Delta(pq)^{-1/2}\}(1 + \lambda^2)^{-1/2} \\ \cdot \{&1 + (1 - 2p)Z_1(mpq)^{-1/2}(1 + \lambda^2)^{-1} + \lambda^2(1 - 2p)\Delta m^{-1/2}p^{-1}q^{-1} \\ &\cdot (1 + \lambda^2)^{-1} + \lambda^3(1 - 2p)Z_2(mpq)^{-1/2}(1 + \lambda^2)^{-1}\}^{-1/2}. \end{aligned}$$

Also, to the same order of magnitude,

$$\begin{aligned} (2.5) \qquad U = \{&Z_1 - \lambda Z_2 - \lambda Z_2\Delta(1/2 - p)m^{-1/2}p^{-1}q^{-1} - \Delta(pq)^{-1/2}\}(1 + \lambda^2)^{-1/2} \\ \cdot \{&1 + \lambda^2(1 - 2p)Z_1(mpq)^{-1/2}(1 + \lambda^2)^{-1} + (1 - 2p)\Delta m^{-1/2}p^{-1}q^{-1} \\ &\cdot (1 + \lambda^2)^{-1} + \lambda(1 - 2p)Z_2(mpq)^{-1/2}(1 + \lambda^2)^{-1}\}^{-1/2}. \end{aligned}$$

Therefore, collecting together terms of $O(m^{-1/2})$,

$$\begin{aligned} (2.6) \qquad U - V &= \{Z_1 - \lambda Z_2 - \Delta(pq)^{-1/2}\}\{(1/2 - p)(1 - \lambda^2)(1 + \lambda^2)^{-1} \\ &\quad \cdot (mpq)^{-1/2}(Z_1 - \lambda Z_2 - \Delta(pq)^{-1/2})\}/(1 + \lambda^2)^{1/2} \\ &= \frac{(1/2 - p)(1 - \lambda^2)(Z_1 - \lambda Z_2 - \Delta(pq)^{-1/2})^2}{(1 + \lambda^2)^{3/2}(mpq)^{1/2}}. \end{aligned}$$

Suppose the alternative hypothesis is $p_1 > p_2$; i.e. $\Delta < 0$. Then since rejection of H_0 is caused by large values of the test statistic, to the stated order of magnitude, (2.6) gives *the test based on V to be more powerful than the test based on U if either $p < 1/2$, $m > n$ or $p > 1/2$, $m < n$, and U to be more powerful than V if either $p < 1/2$, $m < n$ or $p > 1/2$, $m > n$.* When $\Delta > 0$, the roles of U and V in the above are to be reversed. This can either be seen directly from (2.6), or by appealing to the symmetry present in the problem. If we instead look at $X' \equiv m - X$, $Y' \equiv n - Y$, the number of *failures* in m, n Bernoulli trials governed by p_1, p_2 respectively, then X' Y' can be considered as the number of successes in m, n Bernoulli trials governed by $1-p_1$, $1-p_2$ respectively.

When $m=n$ or $p=1/2$, (2.6) tells us that $U-V=0+O(m^{-1})$. Hence for these cases we might turn to the higher order terms. When $m=n$, (denominator of $V)^2=$(denominator of $U)^2-(X-Y)^2/2m^3$, as pointed out by Robbins [4]. (In fact it is straightforward to show that, $U-V=-(1/(8\sqrt{2}\,m))\{Z_1-Z_2-(pq)^{-1/2}\}^3+O(m^{-3/2})$.) *Therefore for $m=n$, $\Delta<0$ or $\Delta>0$, V is more powerful than U.*

When $p=1/2$, we have, upon using the definitions (2.1), (2.2), (2.3),

$$U-V=\frac{(Z_1-\lambda Z_2-2\Delta)^2}{2m(1+\lambda^2)^{5/2}}\{(\lambda^4-1)(Z_1+\lambda Z_2+2\Delta)-\lambda^2(Z_1-\lambda Z_2-2\Delta)\}$$
$$+O(m^{-3/2})\ .$$

The term in braces can take either positive or negative values, but has expectation, $2\Delta(\lambda^4+\lambda^2-1)=2\Delta(\lambda^2+1/2+\sqrt{5}/2)(\lambda^2+1/2-\sqrt{5}/2)$. Hence by modifying our criterion, we could say that *for $\Delta<0$ and $\Delta>0$, V is to be preferred to U* ("on the average") *when $\lambda^2>\sqrt{5}/2-1/2=.618$. Conversely, U is to be preferred to V* ("on the average") *for $\lambda^2<.618$.*

The results of this section are summarized in Table 1. In general, we have shown that for two identical critical values, the power of one test always dominates the power of the other, even when $H_0:\Delta=0$ is true. Which test this is, depends upon conditions on p, m/n, and the direction of the alternative Δ.

Table 1. Entries show when to use V or U

Sample sizes	$m<n$	$m=n$	$m>n$
Alternatives			
$p_2<p_1$ $(p_1<1/2)$	U	V	V
$p_2<p_1$ $(p_1=1/2)$	V*	V	V
$p_2<p_1$ $(p_1>1/2)$	V	V	U
$p_1<p_2$ $(p_1<1/2)$	V	V	U
$p_1<p_2$ $(p_1=1/2)$	V*	V	V
$p_1<p_2$ $(p_1>1/2)$	U	V	V

* Provided also $m>(.618)n$.

3. Power comparisons, matched for size

The previous section has compared powers in absolute terms. However when power comparisons are made on tests now matched so that the powers at $\Delta=0$ (i.e. Type I error) are the same, it is possible that different conclusions might be reached. This section shows that no modification to Table 1 is necessary. All comparisons will be on powers of order up to and including $m^{-1/2}$.

If U is expanded up to and including terms of $O(m^{-1/2})$, then from (2.5),

$$\begin{aligned} U-\mathrm{E}(U)=\Big\{ & Z_1-\lambda Z_2-(Z_1^2-1)\frac{\lambda^2}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}}+(Z_2^2-1)\frac{\lambda^2}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}} \\ & -Z_1Z_2\frac{\lambda(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}}-Z_1\frac{(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)\varDelta}{m^{1/2}pq} \\ & +Z_2\frac{\lambda(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)\varDelta}{m^{1/2}pq}\Big\}(1+\lambda^2)^{-1/2} . \end{aligned}$$

Similarly,

$$\begin{aligned} V-\mathrm{E}(V)=\Big\{ & Z_1-\lambda Z_2-(Z_1^2-1)\frac{1}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}}+\frac{(Z_2^2-1)\lambda^4}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}} \\ & +Z_1Z_2\frac{\lambda(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)}{(mpq)^{1/2}}+Z_1\frac{(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)\varDelta}{m^{1/2}pq} \\ & -Z_2\frac{\lambda(1-\lambda^2)}{1+\lambda^2}\,\frac{(1/2-p)\varDelta}{m^{1/2}pq}\Big\}(1+\lambda^2)^{-1/2} . \end{aligned}$$

Therefore,

$$\mathrm{Var}(U)=1-\frac{2\varDelta}{1+\lambda^2}\,\frac{(1/2-p)(1-\lambda^2)}{m^{1/2}pq} ,$$

$$\mathrm{Var}(V)=1+\frac{2\varDelta}{1+\lambda^2}\,\frac{(1/2-p)(1-\lambda^2)}{m^{1/2}pq} .$$

Also,

$$\mathrm{E}(U-\mathrm{E}(U))^3=\frac{1}{(1+\lambda^2)^{5/2}}\,\frac{(1/2-p)(1-\lambda^2)}{(mpq)^{1/4}}\{2\lambda^4+13\lambda^2+2\}$$

$$\mathrm{E}(V-\mathrm{E}(V))^3=\frac{1}{(1+\lambda^2)^{5/2}}\,\frac{(1/2-p)(1-\lambda^2)}{(mpq)^{1/2}}\{5\lambda^4+\lambda^2+5\} ,$$

independent of $\varDelta$.

We will now use an Edgeworth expansion; see Johnson and Kotz [3], Chapter 12; to give the distribution functions (accurate up to and including terms of $O(m^{-1/2})$) $G(x)$, $H(x)$ of U, V respectively.

$$\begin{aligned} \text{(3.1)}\qquad G(x)=\Phi(y)-\phi(y)\Big\{ & \frac{\varDelta^2(1/2-p)}{(1+\lambda^2)m^{1/2}(pq)^{3/2}}-\frac{\varDelta(1/2-p)(1-\lambda^2)}{(1+\lambda^2)m^{1/2}pq}y \\ & +\frac{(1/2-p)(1-\lambda^2)}{6(1+\lambda^2)^{5/2}m^{1/2}(pq)^{1/2}}(2+13\lambda^2+2\lambda^4)(y^2-1)\Big\} , \end{aligned}$$

$$(3.2) \qquad H(x)=\Phi(y)-\phi(y)\left\{\frac{\Delta^2(1/2-p)\lambda^2}{(1+\lambda^2)m^{1/2}(pq)^{3/2}}+\frac{\Delta(1/2-p)(1-\lambda^2)}{(1+\lambda^2)m^{1/2}pq}y\right.$$
$$-\frac{(1/2-p)(1-\lambda^2)}{m^{1/2}(pq)^{1/2}}+\frac{(1/2-p)(1-\lambda^2)}{6(1+\lambda^2)^{5/2}m^{1/2}(pq)^{1/2}}$$
$$\left.\cdot(5+\lambda^2+5\lambda^4)(y^2-1)\right\},$$

where $y\equiv x+\Delta(1+\lambda^2)^{-1/2}p^{-1/2}q^{-1/2}$, and $\Phi(x)$, $\phi(x)$ are respectively the distribution, density functions of the standard normal. Now under H_0: $\Delta=0$, we wish to solve for c_α and d_α, such that $G(c_\alpha)=H(d_\alpha)$, and *then* compare $G(c_\alpha)$ and $H(d_\alpha)$, when $\Delta\neq 0$. But because the coefficients of (y^2-1) do not depend on Δ, this is easily done.

The special cases of $p=1/2$ and $m=n$, show the two powers to be indistinguishable to the order of magnitude considered, and so will not be considered further. Now amongst the cases left, we will single out one, do the analysis that will show the conclusions of Section 2 to be unchanged, and leave the others for the interested reader to verify. It matters little, but suppose we choose the case where $\Delta>0$ under the alternative, and $(1/2-p)(1-\lambda^2)>0$. Then from (3.1) and (3.2) the differences of the two powers is,

$$P_U-P_V=G(c_\alpha)-H(d_\alpha)$$
$$=\frac{\Delta(1/2-p)(1-\lambda^2)}{(1+\lambda^2)m^{1/2}pq}\{\phi(c'_\alpha)c'_\alpha+\phi(d'_\alpha)d'_\alpha\}$$
$$-\frac{\Delta^2(1/2-p)}{(1+\lambda^2)m^{1/2}(pq)^{3/2}}\{\phi(c'_\alpha)-\lambda^2\phi(d'_\alpha)\}+O(1/m)$$
$$=\frac{(1/2-p)(1-\lambda^2)}{(1+\lambda^2)m^{1/2}pq}\{\phi(c'_\alpha)c'_\alpha\Delta+\phi(d'_\alpha)d'_\alpha\Delta-\phi(c'_\alpha)\Delta^2/(pq)^{1/2}\}$$
$$+O(1/m),$$

where $c'_\alpha=c_\alpha+\Delta(1+\lambda^2)^{-1/2}p^{-1/2}q^{-1/2}$, and similarly for d'_α. But since $\{U\leqq c_\alpha\}$, $\{V\leqq d_\alpha\}$ are the rejection regions, then for $0<\alpha<1/2$, c'_α and d'_α are both negative. Hence,

$$P_U-P_V=a/m^{1/2}, \qquad \text{where } a<0;$$

i.e. *V is more powerful than U*, which is the same result as in Section 2. *Therefore Table* 1 *is*, apart from the special cases of $p=1/2$, and $m=n$, where the powers are indistinguishable, *unchanged*. There is some comparison possible between our results and those of Eberhardt and Fligner [1]. Looking at their Figure A, for (p_1, p_2) values very close to the diagonal $p_1=p_2$; i.e. for local alternatives; we see the same recommendations of Table 1. However our approach has been via Pitman efficiencies, and we have taken care to first match Type I errors of the two tests, and then compare powers.

As a final consideration, one might ask the question as to which test statistic is closer to normality under $H_0 : \Delta=0$; i.e. which of the two rejection regions $\{U \leqq 1.96\}$, $\{V \leqq 1.96\}$, gives the more accurate size .05 test? Under $H_0 : \Delta=0$, we have from (3.1), (3.2),

$$G(x)=\Phi(x)-\frac{\phi(x)(1/2-p)(1-\lambda^2)}{6(1+\lambda^2)^{5/2}(mpq)^{1/2}}(2+13\lambda^2+2\lambda^4)(x^2-1) ,$$

$$H(x)=\Phi(x)+\frac{\phi(x)(1/2-p)(1-\lambda^2)}{(mpq)^{1/2}}-\frac{\phi(x)(1/2-p)(1-\lambda^2)}{6(1+\lambda^2)^{5/2}(mpq)^{1/2}} \cdot (5+\lambda^2+5\lambda^4)(x^2-1) .$$

So the two terms that we have to compare are,

$$g(\lambda^2)=2+13\lambda^2+2\lambda^4 , \qquad h(\lambda^2)=(5+\lambda^2+5\lambda^4)-6(1+\lambda^2)^{5/2}/(k_\alpha^2-1) ;$$

the closer these values are to zero, the better is the normal approximation under H_0. For $\alpha=0.05$, $k_\alpha=1.96$; in this case Table 2 compares $g(\lambda^2)$ and $h(\lambda^2)$ for λ^2 $(=m/n)=0.0(0.2)3.0$. Clearly the test statistic V is in general, closer to the standard normal under $H_0 : \Delta=0$. This holds regardless of the signs of $(1/2-p)$ and $(1-\lambda^2)$.

Table 2. Comparison of two normal approximations

m/n	$g(m/n)$	$h(m/n)$	m/n	$g(m/n)$	$h(m/n)$
0.0	2.00	2.89	1.6	27.92	−3.62
0.2	4.68	2.07	1.8	31.88	−4.70
0.4	7.52	1.30	2.0	36.00	−5.91
0.6	10.52	0.56	2.2	40.28	−7.28
0.8	13.68	−0.18	2.4	44.72	−8.81
1.0	17.00	−0.94	2.6	49.32	−10.52
1.2	20.48	−1.76	2.8	54.08	−12.44
1.4	24.12	−2.64	3.0	59.00	−14.57

THE FLINDERS UNIVERSITY OF SOUTH AUSTRALIA

REFERENCES

[1] Eberhardt, K. R. and Fligner, M. A. (1977). A comparison of two tests for equality of two proportions, *Amer. Statistician*, **31**, 151-155.

[2] Fraser, D. A. S. (1957). *Non Parametric Methods in Statistics*, Wiley, N.Y.

[3] Johnson, N. L. and Kotz, S. (1970). *Continuous Univariate Distributions—1*, Houghton Mifflin, Boston.

[4] Robbins, H. (1977). A fundamental question of practical statistics, *Amer. Statistician*, **31**, 97.