Conditional versus Unconditional Exact Tests for Comparing Two Binomials

Cyrus R. Mehta and Pralay Senchaudhuri

4 September 2003

Introduction

There are two fundamentally different exact tests for comparing the equality of two binomial probabilities – Fisher's exact test (Fisher, 1925), and Barnard's exact test (Barnard, 1945). Fisher's exact test (Fisher, 1925) is the more popular of the two. In fact, Fisher was bitterly critical of Barnard's proposal for esoteric reasons that we will not go into here. Notwithstanding Fisher's reservations, both tests are available in StatXact.

Consider the following example of a vaccine efficacy study (Chan, 1998). In a randomized clinical trial of 30 subjects, 15 were innoculated with a recombinant DNA influenza vaccine and the 15 were innoculated with a placebo. Twelve of the 15 subjects in the placebo group (80%) eventually became infected with influenza whereas for the vaccine group, only 7 of the 15 subjects (47%) became infected. The data are tabulated as a 2×2 table (see Table 1). Suppose π_e is the probability of infection for the vaccine (or experimental) group and

Table 1: Results of Vaccine Efficacy Study

Infection	Treatment		Combined
Status	Vaccine	Placebo	Response
Yes	7 (47%)	12 (80%)	19
No	8 (53%)	3 (20%)	11
Totals	15	15	30

π_c is the probability of infection for the placebo (or control) group. What is the exact one-sided p-value for testing the null hypothesis

$$H_0\colon \pi_e = \pi_c \, .$$

The answer depends on the type of exact test adopted. Fisher's exact test produces an exact p-value of 0.0641. On the other hand Barnard's exact test produces an exact p-value of 0.0341. The smaller, statistically significant, exact p-value produced by Barnard's method is no accident. For 2×2 tables, Barnard's test is more powerful than Fisher's, as Barnard noted in his 1945 paper, much to Fisher's chagrin. The purpose of this short article is to explain why Barnard's method is more powerful in this setting and to discuss the limitations of the method.

What's the Difference Between the Two Methods?

In order to appreciate the difference between Fisher's and Barnard's exact tests, let us consider a generic 2×2 contingency table obtained by innoculating 15 subjects with vaccine and 15 subjects with placebo. Let us

Table 2: A Generic 2×2 Table Generated by the Vaccine Efficacy Trial

Infection	Treatment		Combined
Status	Vaccine	Placebo	Response
Yes	x_e	x_c	$x_c + x_e$
No	$15 - x_e$	$15 - x_c$	$30 - x_c - x_e$
Totals	15	15	30

denote this generic 2×2 table (Table 2) by $\boldsymbol{X}$ and let us denote the table that was actually observed (see Table 1) by $\boldsymbol{X}_0$. Suppose that the common probability of infection under the null hypothesis is $\pi_e = \pi_c = \pi$. Then the probability of observing any generic table $\boldsymbol{X}$ is a product of two binomials:

$$P(\boldsymbol{X}|\pi) = \binom{15}{x_c}\binom{15}{x_e}\pi^{x_c+x_e}(1-\pi)^{30-x_c-x_e} \; . \tag{1}$$

The exact p-value is then the sum of such probabilities for all tables, $\boldsymbol{X}$, that could have been observed which are at least as extreme as the observed table, $\boldsymbol{X}_0$, under the null hypothesis. Specifically, suppose $T(\boldsymbol{X})$ is a "discrepancy measure" or test statistic that measures how discrepant any table $\boldsymbol{X}$ is relative to the type of table one would expect under the null hypothesis. Then, for any given π, the exact p-value of the observed table $\boldsymbol{X}_0$ is

$$p(\pi) = \sum_{T(\boldsymbol{X}) \geq T(\boldsymbol{X}_0)} P(\boldsymbol{X}|\pi) \; . \tag{2}$$

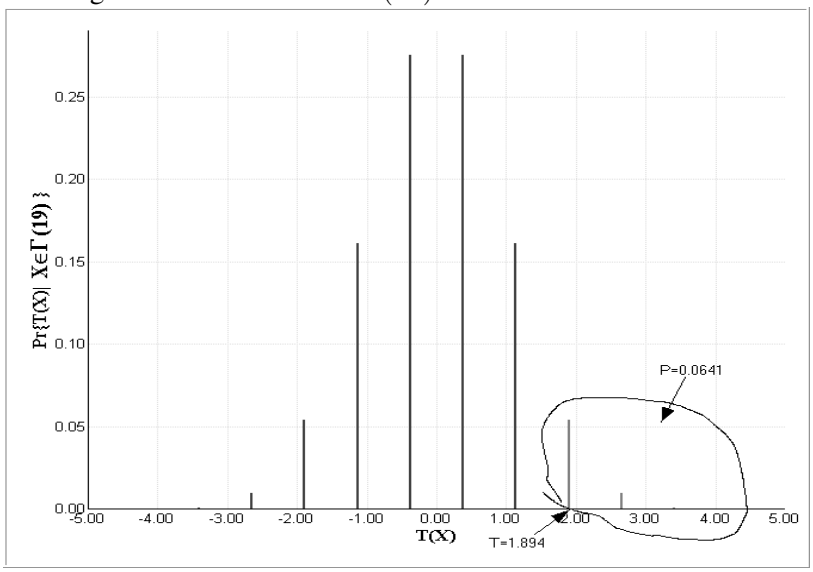
Now there exist many different test statistics for determining if a table $\boldsymbol{X}$ is more or less extreme than the observed table $\boldsymbol{X}_0$ under the null hypothesis. Since our purpose is to explain the difference between the two types of exact tests, we will not discuss the merits of these different test statistics. For this example we will use the popular Wald statistic

$$T(\boldsymbol{X}) = \frac{\hat{\pi}_c - \hat{\pi}_e}{\sqrt{\bar{\pi}(1-\bar{\pi})(\frac{1}{15}+\frac{1}{15})}} \; , \tag{3}$$

where $\hat{\pi}_e = x_e/15$, $\hat{\pi}_c = x_c/15$ and $\bar{\pi} = (x_c + x_e)/15$. However similar conclusions would be reached with other test statistics as well.

Although the Wald test statistic $T(\boldsymbol{X})$ is well defined, equation (2) is not of much practical use for computing an exact p-value because it depends on knowing the value of π, the common probability of an infection under the null hypothesis. One could of course use the data to estimate π and then substitute this estimate into equation (2). But then the p-value would no longer be exact. The main difference between the Fisher and Barnard tests is the manner in which they eliminate this nuisance parameter from (2) without sacrificing the exactness.

Figure 1: Distribution of $T(\boldsymbol{X})$ for Fisher's Exact Test



Fisher's Exact Test

Fisher's exact test gets rid of π by restricting attention only to generic 2×2 tables $\boldsymbol{X}$ in which $x_c + x_e = 19$, the sum of responses that were actually observed in the data. Accordingly define

$$\Gamma = \{\boldsymbol{X}\text{: } \boldsymbol{X} \text{ is a } 2 \times 2 \text{ table like the one captioned as Table 2}\}$$

and

$$\Gamma(19) = \{\boldsymbol{X}\text{: } \boldsymbol{X} \in \Gamma \text{ and } x_c + x_e = 19\} \ .$$

Fisher noted that, under the null hypothesis, provided we confine our attention only to tables $\boldsymbol{X} \in \Gamma(19)$, each table has a hypergeometric probability that no longer depends on the nuisance parameter π. That is,

$$P(\boldsymbol{X}|\boldsymbol{X} \in \Gamma(19)) = \frac{\binom{15}{x_c}\binom{15}{x_e}}{\binom{30}{19}} \ .$$

Thereby Fisher's exact one-sided p-value is readily evaluated as

$$P_{\mathcal{F}} = \sum_{T(\boldsymbol{X}) \geq T(\boldsymbol{X}_0)} P(\boldsymbol{X}|\boldsymbol{X} \in \Gamma(19)) \ .$$

Figure 1 displays the exact distribution of $T(\boldsymbol{X})$ conditional on $\boldsymbol{X} \in \Gamma(19)$.

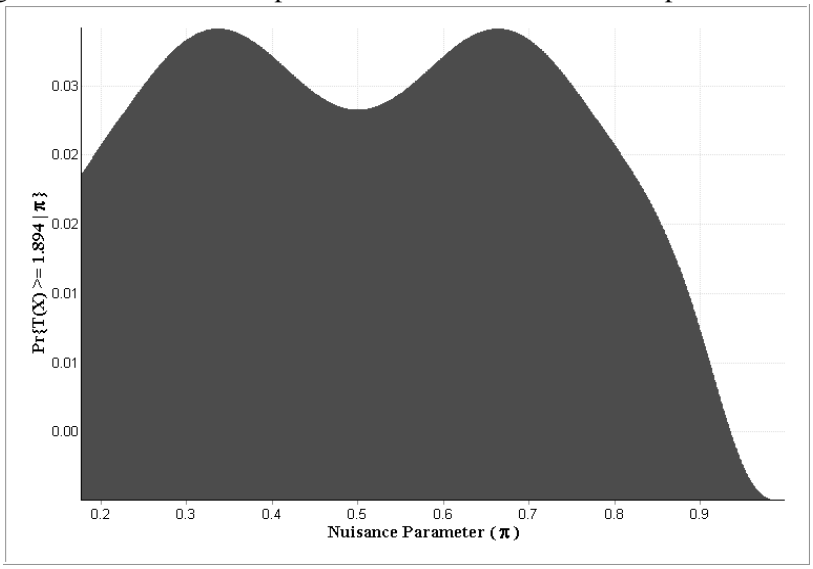
Substituting $x_c = 12$ and $x_e = 7$ into equation (3) we obtain $T(\boldsymbol{X}_0) = 1.894$. Fisher's exact p-value is thus the tail area to the right of 1.894 in Figure 1. This tail area is

$$P_{\mathcal{F}} = 0.0641 \ .$$

Barnard's Exact Test

Barnard's exact test is an unconditional test. It generates the exact distribution of $T(\boldsymbol{X})$ by considering all the tables $\boldsymbol{X} \in \Gamma$, and rather than just the tables $\boldsymbol{X} \in \Gamma(19)$. Since the observed Wald test statistic is

Figure 2: Barnard's exact p-value as a function of nuisance parameterπ



$T(\boldsymbol{X}_0) = 1.894$, the p-value given π is

$$p(\pi) = \sum_{T(\boldsymbol{X})\geq 1.894} \binom{15}{x_c}\binom{15}{x_e}\pi^{x_e+x_c}(1-\pi)^{30-x_e-x_c} \ . \tag{4}$$

What shall we do with the unknown nuisance parameter π in the above p-value? Barnard suggested that we calculate $p(\pi)$ for all possible values of $\pi \in (0,1)$ and choose the value, π^*, say, that maximizes $p(\pi)$. Specfically, Barnard's exact p-value is defined as

$$P_{\mathcal{B}} = \sup\{p(\pi)\colon\ \pi \in (0,1)\} \ .$$

To understand Barnard's approach, examine Figure 2.

This figure plots the nuisance parameter π against the corresponding p-value $p(\pi)$ for all $\pi \in (0,1)$. The function is maximized at $\pi^* = 0.3365$ where it attains the value $p(0.3365) = 0.0341$. Thus

$$P_{\mathcal{B}} = 0.0341.$$

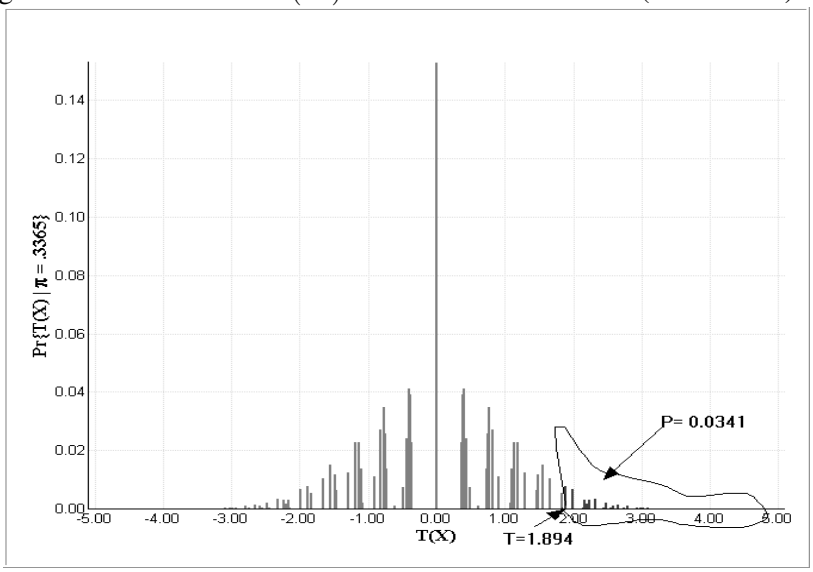
Notice that this p-value is only about half as large as that generated by Fisher's exact test, and suggests that Barnard's test is more powerful than Fisher's.

Is Barnard's Test Really More Powerful then Fisher's?

It is interesting to examine why $P_{\mathcal{B}}$ is so much smaller than $P_{\mathcal{F}}$. Figure 3 displays the exact probability distribution for Barnard's test; i.e., the exact probability distribution of the Wald test statistic $T(\boldsymbol{X})$, generated by assigning to each $\boldsymbol{X} \in \Gamma$ the probability $P(\boldsymbol{X}|\pi^*)$ given by equation (1), where $\pi^* = 0.3365$.

Notice that $T(\boldsymbol{X})$ has substantially more support points in Figure 3 than it does in Figure 1. In other words, the same test statistic is much more discrete for Fisher's exact test than it is for Barnard's exact test. This is a direct consequence of restricting Fisher's test to 2×2 tables belonging to the conditional reference reference $\Gamma(19)$ rather than the larger reference unconditional reference set Γ. By conditioning on $x_e + x_c = 19$ we have made

Figure 3: Distribution of $T(\boldsymbol{X})$ for Barnard's Exact Test ($\pi^* = .3365$)



the sample space for Fisher's exact test much more discrete than it is for Barnard's exact test. Consequently, the number of distinct p-values that one could obtain with Fisher's exact test is less than the corresponding number of distinct p-values that one could obtain with Barnard's exact test. This in turn implies that if we want to restrict the type-1 error to some upper limit, say 5%, Fisher's procedure will usually be more conservative than Barnard's, resulting in a loss of power. The power loss diminishes as the sample sizes get larger since the discreteness of the Fisher statistic is not as pronounced.

When comparing Fisher's and Barnard's exact tests, the loss of power due to the greater discreteness of the Fisher statistic is somewhat offset by the requirement that Barnard's exact test must maximize over all possible p-values, by choice of the nuisance parameter, π. For 2×2 tables the loss of power due to the discreteness dominates over the loss of power due to the maximization, resulting in greater power for Barnard's exact test. But as the number of rows and columns of the observed table increase, the maximizing factor will tend to dominate, and Fisher's exact test will achieve greater power than Barnard's. For details of this investigation see Mehta and Hilton (1993). For additional discussion of this topic see Appendix G of the StatXact-5 User Manual.

Reference

Barnard GA (1945). A new test for 2×2 tables. *Nature* 156:177.

Chan I (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine* 17, 1403-1413.

Fisher RA (1925). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh.

Mehta CR, Hilton JF (1993). Exact power of conditional and unconditional tests: going beyond the 2x2 contingency table. *American Statistician*, 47:91-98.