A Festschrift for Herman Rubin
Institute of Mathematical Statistics
Lecture Notes – Monograph Series
Vol. 45 (2004) 250–254

Conservative bounds on extreme P-values for testing the equality of two probabilities based on very large sample sizes

Herman Chernoff[1]

Harvard University

Abstract: With very large sample sizes the conventional calculations for tests of the equality of two probabilities can lead to very small P-values. In those cases, the large deviation effects make inappropriate the asymptotic normality approximations on which those calculations are based. While reasonable interpretations of the data would tend to reject the hypothesis in those cases, it is desireable to have conservative estimates which don't underestimate the P-value. The calculation of such estimates is presented here.

1. Introduction

There are several excellent alternatives for testing the hypothesis that $p_1 = p_2$ where p_1 and p_2 are probabilities governing two binomial samples. These include the Yates continuity correction and the Fisher Exact test and several others based on the asymptotic normality of the observed proportions. All these test procedures have the desireable property that the calculated P-value does not depend on the unknown common probability under the hypothesis. There is a slight problem with the Fisher exact test, i.e., it is not strictly appropriate for the problem because the calculated probability is conditional on the values of the margins, which are not fixed in advance. The problem is considered slight because the information in the margins is quite small Chernoff (2004).

In a legal case the problem arose where there were 7 successes out of 16 trials for one sample and 24 successes out of 246 in the second sample. It is clear that the hypothesis is not plausible in the light of these data. Since the various alternative tests provide substantially different calculated P-values, all very small, it was considered wise to present a very conservative P-value. While one sample size was substantial, the other was quite modest. Neither was so large that modern computers would be frustrated by calculating the exact P-value rather than relying on asymptotic theory. One consequence of such an approach is that the P-value is no longer independent of the unknown value of the nuisance parameter, the common value of the probabiities under the hypothesis. This problem is dealt with in several publications (Berger and Boos (1994), Chernoff (2003)). A crucial aspect of the difficulty in using asymptotic theory is that in extreme cases where the P-values are very very small, we are in the tails of the distribution and asymptotic normality no longer fits in these large deviation cases.

A new problem recently came to my attention, where both sample sizes are enormous, i.e. $n_1 = 19,479$ and $n_2 = 285,422$, Here agains there are several cases

[1] Department of Statistics, Harvard University, Cambridge, MA 02138, USA. e-mail: chernoff@stat.harvard.edu

Keywords and phrases: P-values, Chernoff bound, Yates continuity correction, Poisson.

AMS 2000 subject classifications: primary 62F03; secondary 62B10.

where we have a large deviation problem, and asymptotic normality is not appropriate, and probably not conservative. How should we deal with this problem in this example where ordinary high speed computers may find it difficult to provide exact calculations such as were feasible in the previous case? The Chernoff bound, originally derived by H. Rubin, provides a method of deriving an upper bound on the desired probability which is convenient to calculate.

2. The Poisson approximation

While the normal approximation is unreliable, the Poisson approximation may be better. In any case, it is to be used here merely to provide an initial approximation for the quantities required for the binomial calculation. We outline the analysis which provides a solution assuming the Poisson approximation fits.

The main tool to deliver a conservative bound on the P-value is the Chernoff bound, first derived by Herman Rubin, using a Chebyshev type of inequality, that states that if $d \geq E(X)$,

$$P(X \geq d) \leq E\big(e^{t(X-d)}\big)$$

for all t. The right hand side attains its minimum for $t \geq 0$.

Let X_1 and X_2 be the number of successes in n_1 and n_2 independent trials with common probability p, and let

$$D = \frac{X_1}{n_1} - \frac{X_2}{n_2},$$

Using the Poisson approximation to the binomial distribution, we shall derive the curve in the (p,d) space for which the bound on $\log(P(D \geq d))$,

$$q = \log\big(\inf_t E\big(e^{t(D-d)}\big)\big)$$

attains a given value, for $d > 0$. Under the assumption that the number of successes in each trial has a Poisson distribution, we have

$$Q(t,d) = \log\big(E\big(e^{t(D-d)}\big)\big) = -dt + n_1 p\big(e^{t/n_1} - 1\big) + n_2 p\big(e^{-t/n_2} - 1\big).$$

Differentiating with respect to t, the value of t which minimizes Q satisfies

$$e^{t/n_1} - e^{-t/n_2} = d/p = a$$

while

$$Q(t,d) = pr(t,a)$$

where

$$r(t,a) = -at + n_1\big(e^{t/n_1} - 1\big) + n_2\big(e^{t/n_2} - 1\big).$$

For each value of t, there is a corresponding value of a for which t is optimal and a corresponding value of r. Let $p = q/r$ and $d = ap$. As t varies these values of p and d trace out the (p,d) curve corresponding to the given value of $q \geq \log(P)$.

3. The binomial case

We use the Poisson calculation to get a first approximation in the derivation of the $(p.d)$ curves for the binomial case. In the previous section we obtained values of p and d for each value of t. Here we will keep both p and q fixed, and starting with the value of t, we find

$$Q(t,d) = \log E\big(e^{t(D-d)}\big) = -td + n_1 \log\big(1 - p + pe^{t/n_1}\big) + n_2 \log\big(1 - p + pe^{-t/n_2}\big)$$

and the value of d for which Q is minimized by the given value of t is given by

$$d(t) = (1-p)\left(\frac{1}{1-p+pe^{-t/n_2}} - \frac{1}{1-p+pe^{t/n_1}}\right).$$

We note that

$$d'(t) = p(1-p)\left(\frac{e^{t/n_1}}{n_1(1-p+pe^{t/n_1})^2} + \frac{e^{-t/n_2}}{n_2(1-p+pe^{-t/n2})^2}\right).$$

Insofar as $Q(t, d(t))$ varies from the specified value of q, we apply the Newton iteration to modify t. This leads t to the new value $t + (q - Q(t, d(t))/Q'(t)$ where

$$Q'(t) = \partial Q/\partial t + d'(t)\partial Q/\partial d = -td'(t).$$

Thus t goes into $t - (q - Q)/td'(t)$.

If the new value of t and $d(t)$ do not provide $Q(t, d(t))$ close enough to the desired value q, one may iterate again. Finally we have for each initial value of t and the given value of q a new point (p, d) for the curve of specified $q \geq \log(P(D \geq d))$.

While the curves we have obtained of (p, d) values for a given value of q are useful, they don't resolve the inverse problem in which we may be interested. That is, how do we calculate a bound on the P-value for a given p and d? A series of curves provided above would be useful to get rough approximations for a set of cases with given n_1 and n_2, but do not provide a reasonable precise algorithm should that be desired. To obtain the bound on the P-value, we start with the estimate of p given by $p = (X_1 + X_2)/(n_1 + n_2)$. Assusming that value is fixed, we approximate t, assuming t is small compared to n_1 and n_2, by

$$t = \frac{dn_1n_2(1-p)}{(n_1+n_2)p}$$

This value of t together with the observed value of D yields $Q(t, D)$ and $d(t)$. Insofar as $d(t)$ differs from D, we modify t by the Newton method to $t + (D - d(t))/d'(t)$. With this new value of t, we recalculate Q and $d(t)$ and interate until $d(t)$ is approximately D. Then the bound on the P-value is given by e^Q assuming our estimate of p is accurate. Since the range of possible values of p is quite limited under the hypothesis, we can see how much the P-value changes by considering potential alternative values of p.

4. Summary

For the case of very large sample sizes, with data quite inconsistent with the hypothesis that two binomial distributions have the same value of p, we anticipate very small P-values. The usual calculations are unreliable because large deviation effects make the asymptotic normality on which these calculations depend unreliable. While it is clear in such cases that the hypothesis is false, it is often desireable to have a conservative bound on the P-value. The Chernoff bound provides such a result. We provide the basis for three algorithms. One provides the (p, d) values for which given bounds on the value of $\log(P)$ are attained assuming that a Poisson approximation to the binomial distribution is acceptable. This algorithm is used as a starting point in calculating the curve of (p, d) values for the binomial distribution. Finally we show how to calculate the conservative bound for the P-value in the binomial case.

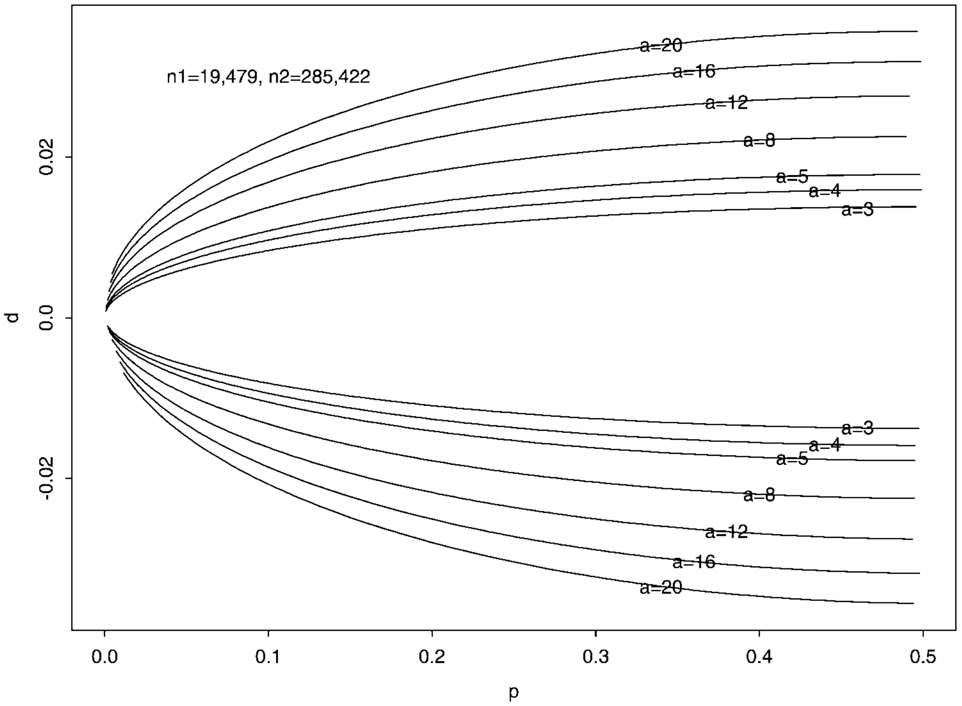

Figure 1:

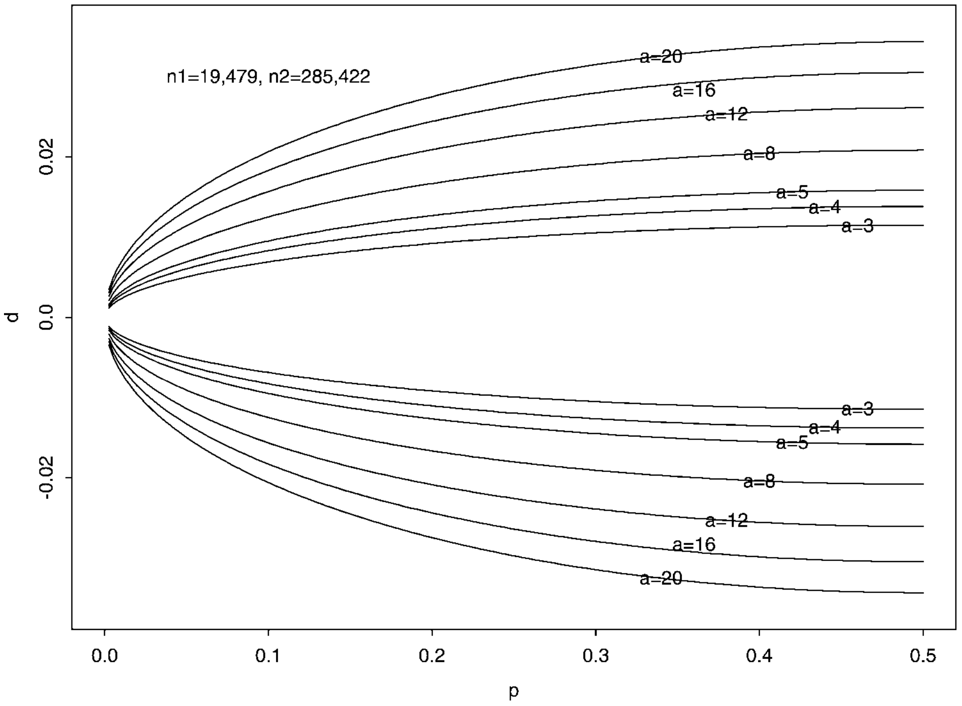

Figure 2:

We have in Figure 1, the (p, d) values for the case $P = 10^{-a}$ where a takes on the values 3,4,5,8,12,16, and 20, $n_1 = 19,479$, $n_2 = 285,422$, and we use the binomial distribution. In Figure 2 we use the calculation for the Yates continuity correction where p represents the estimate of the common probability.

In both of these cases we have calculated one sided P-values. The calculation for negative values of D can be obtained by interchanging n_1 and n_2 after replacing D by its absolute value.

References

[1] Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, **89**, 1012–1016. MR1294746

[2] Chernoff, H. (2003). Another View of the Classical Problem of Comparing Two Probabilities, *J. Iranian Statist. Soc.*, **1**, 35–54. MR1981752

[3] Chernoff, H. (2004). Information for testing the equality of two probabilities, from the margins of the 2×2 table. *J. Statist. Plann. Inference*, **121**, 209–214. MR2038817