The Exact Size of the χ^2 Test for Comparing Two Binomial Proportions

Roger L. Berger
North Carolina State University
and National Science Foundation

OUTLINE

1. Introduction to Problem/Notation
2. Computation of Exact Size for All $1 \leq m, n \leq 400$
3. Main Conclusions from Computations
   (a) Even for large sample sizes, exact size can be $\gg \alpha$
   (b) For large sample sizes, exact size depends only on n/m
4. Inflated Size a Problem?
5. Size of an Exact, Unconditional Test
6. Conclusions

Problem and Notation

$$X \sim \text{binomial}(m, p_1)$$
$$Y \sim \text{binomial}(n, p_2)$$

X and Y are independent.

To test "Homogeneity Hypothesis"

$$\mathrm{H}_0 : p_1 = p_2$$
$$\mathrm{H}_a : p_1 \neq p_2$$

Data often presented in a 2×2 table

	Cured		
	Yes	No	
Drug 1	$X = O_{11}$	$m - X = O_{12}$	$m = O_{1\cdot}$
Drug 2	$Y = O_{21}$	$n - Y = O_{22}$	$n = O_{2\cdot}$
	$R = O_{\cdot 1}$	$N - R = O_{\cdot 2}$	N

The χ^2 Test

For large sample sizes, the most commonly used test of H_0 versus H_a is reject H_0 if $\chi^2 \geq \chi^2_{\alpha,1}$ where

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

and

$$E_{ij} = \frac{O_{i\cdot} O_{\cdot j}}{N}.$$

The χ^2 statistic can also be written as $\chi^2 = Z^2$ where

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \frac{\hat{p}(1-\hat{p})}{n}}},$$

where $\hat{p}_1 = X/m$, $\hat{p}_2 = Y/n$, and $\hat{p} = (X+Y)/(m+n)$.

Then, the χ^2 test rejects H_0 if $|Z| \geq z_{\alpha/2}$.

What is "asymptotic size-α" for χ^2 test?

For any fixed α, $0 < \alpha < 1$, and any fixed p, $0 < p < 1$, $p_1 = p_2 = p$

$$\lim_{\substack{m \to \infty \\ n \to \infty}} P_p(\chi^2 \geq \chi^2_{\alpha,1}) = \alpha.$$

"P_p" refers to the exact binomial probability.

What is the size of the χ^2 test?

$$\text{size}(m,n) = \sup_{0<p<1} P_p(\chi^2 \geq \chi^2_{\alpha,1})$$

Is it true that

$$\lim_{\substack{m \to \infty \\ n \to \infty}} \text{size}(m,n) = \alpha?$$

This talk is about size(m,n).

Numeric Computation of size(m, n)

We computed size(m, n) for

- $\alpha = .05$, $\chi^2_{.05,1} = 3.84$
- every m and n with $1 \leq m, n \leq 400$, 160,000 points

Note:

- $P_p(\chi^2 \geq \chi^2_{\alpha,1})$ is a polynomial of degree $N = m + n$ in p. It has many local maxima.
- size$(m, n) = \sup_{0<p<1} P_p(\chi^2 \geq \chi^2_{\alpha,1})$.
- How we found the global maximum is the topic of another talk.

Sample Sizes for which size$(m, n) \leq \alpha = .05$

- There are about 660 cases (.4%) for which size$(m, n) \leq .05$.
- There are no cases with $\min\{m, n\} \geq 150$ for which size$(m, n) \leq .05$.
- Having both m and n large does not guarantee size$(m, n) \leq .05$.

(m, n) Sample Size Pairs in Different Ranges

size(m, n)	number of (m, n) pairs	percentage
.00 – .05	667	0.4%
.05 – .06	126,478	79.1%
.06 – .07	8,227	5.1%
.07 – .08	8,331	5.2%
.08 – .09	8,996	5.6%
.09 – .10	7,301	4.6%
total	160,000	100.0%

Recall: $\text{size}(m, n) = \sup_{0<p<1} P_p(\chi^2 \geq \chi^2_{\alpha,1})$

When m and n are not tiny, if $\text{size}(m, n)$ is much bigger than $\alpha = .05$, then the "sup" takes place at a p near 0 or 1.

Explanation

$$P_p(\chi^2 \geq \chi^2_{\alpha,1})$$
$$= \sum_{r=0}^{m+n} \underbrace{P(\chi^2 \geq \chi^2_{\alpha,1} | X+Y=r)}_{W(r)} \underbrace{P_p(X+Y=r)}_{\text{binomial}(N=m+n,p)}.$$

- $W(r)$ does not depend on p; it is a hypergeometric probability.
- When r is small,, e.g., $r = 1, 2, 3, 4$, there are only a few points in the hypergeometric sample space and $W(r)$ can be much bigger than $\alpha = .05$.
- If p is small ($p \approx 1/N$, $2/N$, etc.) most of the binomial probability is on the values $r = 1$, 2, 3, and 4.
- In this way the sum can be much larger than $\alpha = .05$.

Explanation, continued

- But for r moderate, near $N/2$, there are more points in the hypergeometric sample space and $W(r)$ tends to be closer to $\alpha = .05$.
- For p near $1/2$ most of the binomial probability is on these moderate r's. Also, for p near $1/2$ the binomial probability is more spread out and one or two large values of $W(r)$ will not have much effect on the sum.

Example: $m = 400$, $n = 100$, $p = .003$

r	$W(r)$	$P_p(X+Y=r)$
0	.000	.223
1	.200	.335
2	.040	.251
3	.103	.126
4	.027	.047
5	.057	.014
6	.016	.003
7	.032	.001
8	.055	.000
9	.019	.000
10	.031	.000

$$\sum_{r=0}^{10} P(\chi^2 \geq \chi^2_{\alpha,1} | X+Y=r) P_p(X+Y=r) = .092.$$

- Does this pattern persist for larger m and n?
- Or for larger (m, n) does the size converge to α?

This pattern persists as m and $n \to \infty$.

Why?

Clue: The regions seem to be defined by straight lines.

$$\text{straight line:} \quad \frac{n}{m} = f \text{ (a constant)}$$

Example: Boundary at $n/m = 1/4$ or $= 4$.

Actually, $n/m = \chi^2_{.05,1} = 3.84$.

If we set $n/m = f$ $(n = fm)$ and let $m \to \infty$, all the important quantities depend on m and n only through f.

$$P_p(\chi^2 \geq \chi^2_{\alpha,1})$$
$$= \sum_{r=0}^{m+n} \underbrace{P(\chi^2 \geq \chi^2_{\alpha,1} | X+Y=r)}_{W(r)} \underbrace{P_p(X+Y=r)}_{\text{binomial}(N=m+n,p)}.$$

Look at each term and see why it depends only on f as $m \to \infty$.

Test statistic: For each fixed sample point (x, y),

$$Z(x,y) = \frac{(x/m) - (y/fm)}{\sqrt{\frac{x+y}{m+fm}\left(1 - \frac{x+y}{m+fm}\right)\left(\frac{1}{m} + \frac{1}{fm}\right)}}$$

$$\lim_{m\to\infty} Z(x,y) = \frac{xf - y}{\sqrt{(x+y)f}}$$

So, for fixed $r = x + y$ and $f = n/m$, the set of (x, y) values that satisfy $Z^2(x, y) = \chi^2 \geq \chi^2_{\alpha,1}$ converges to a fixed set as $m \to \infty$.

$$W(r) = P(\chi^2 \geq \chi^2_{\alpha,1} | X + Y = r)$$

is a hypergeometric probability:

m white balls,

fm black balls,

sample size r

For large m (large number of balls), this probability can be approximated by a binomial(r, p) probability where

$$p = \frac{m}{m + fm} = \frac{1}{1 + f}.$$

Again, depends only on $f = n/m$.

The other term is $P_p(X+Y=r)$ where $X+Y \sim \text{binomial}(N,p)$, $N = m+n = m+fm$, and $p = p_1 = p_2$.

Earlier we explained that if the size is much bigger than $\alpha = .05$, the "sup" will occur at a p near 0 or 1, in the range of $1/N$, $2/N$, $3/N$, etc.

To approximate this probability at $p = a/N$, use the usual Poisson(a) approximation.

Does not depend on p, f, m, or anything.

So, as $m \to \infty$, all the terms in

$$P_p(\chi^2 \geq \chi^2_{\alpha,1})$$
$$= \sum_{r=0}^{m+n} \underbrace{P(\chi^2 \geq \chi^2_{\alpha,1} | X+Y=r)}_{W(r)} \underbrace{P_p(X+Y=r)}_{\text{binomial}(N=m+n,p)}.$$

can be approximated by something that depends only of $f = n/m$.

Example: $m = 400$, $n = 100$, $p = .003$, $f = .25$

r	$W(r)$	$P_p(X+Y=r)$	approx $W(r)$	$\frac{1.5^r e^{-1.5}}{r!}$
0	.000	.223	.000	.223
1	.200	.335	.200	.335
2	.040	.251	.040	.251
3	.103	.126	.104	.126
4	.027	.047	.027	.047
5	.057	.014	.058	.014
6	.016	.003	.017	.004
7	.032	.001	.033	.001
8	.055	.000	.056	.000
9	.019	.000	.020	.000
10	.031	.000	.033	.000

$$\sum_{r=0}^{10} W(r) P_p(X+Y=r) = .092 \approx .092.$$

Exact Unconditional Tests

There are exact, unconditional tests that

- ‡ maintain size$(m, n) \leq \alpha$ for all m and n
- ‡ have size(m, n) very close to α for most m and n

My favorite (Berger & Boos (1994, *JASA*), Berger (1996, *Amer. Statist.*)):
Let $C = C(X, Y)$ be a $100(1 - \beta)\%$ confidence interval for $p = p_1 = p_2$. Let

$$p(x, y) = \sup_{p \in C(x,y)} P_p(\chi^2 \geq \chi^2_{\text{obs}}) + \beta.$$

Then

- † $p(x, y)$ is a valid p-value.
- † The test that rejects H_0 if and only if $p(X, Y) \leq \alpha$ is a level-α test.
- † The exact size of this test is very close to α for most m and n.
- † The computation of this p-value requires a fraction of a second for $m + n = 1000$.

Conclusions

1. For only very few sample sizes (m, n) is the exact size$(m, n) \leq \alpha = .05$.
2. We found no cases with m and n both large for which size$(m, n) \leq \alpha = .05$.
3. If $n/m \leq 1/4$ or $n/m \geq 4$, size(m, n) is usually much greater than $\alpha = .05$.
4. For large m and n, size(m, n) is a function only of $f = n/m$.
5. These extremely high P(Type I Error) values occur only for p very close to 0 or 1; they may not be of much practical concern.
6. There are exact unconditional tests for which size$(m, n) \leq \alpha$ for all m and n and for which size(m, n) is very close to α, e.g., $.047 \leq$ size$(m, n) \leq .050$ for almost all m and n.