The Exact Size of the $\chi^2$ Test for Comparing Two Binomial Proportions

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

This talk represents the views only of the author and in no way represents the views of the National Science Foundation or any other Federal government agency.
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”

\[ H_0: p_1 = p_2 \]
\[ H_a: p_1 \neq p_2 \]
Data often presented in a $2 \times 2$ table

<table>
<thead>
<tr>
<th>Cured</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug 1</td>
<td>$X = O_{11}$</td>
<td>$m - X = O_{12}$</td>
</tr>
<tr>
<td>Drug 2</td>
<td>$Y = O_{21}$</td>
<td>$n - Y = O_{22}$</td>
</tr>
<tr>
<td></td>
<td>$R = O_1$</td>
<td>$N - R = O_2$</td>
</tr>
</tbody>
</table>
The $\chi^2$ 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

$$E_{ij} = \frac{O_iO_j}{N}.$$

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

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

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

Then, the $\chi^2$ test rejects $H_0$ if $|Z| \geq z_{\alpha/2}$. 
What is “asymptotic size-α” for $\chi^2$ test?

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

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

“$P_p$” refers to the exact binomial probability.

What is the size of the $\chi^2$ test?

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

Is it true that

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

This talk is about $\text{size}(m, n)$. 

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.
- \(\text{size}(m, n) = \sup_{0 < p < 1} P_p(\chi^2 \geq \chi^2_{\alpha,1})\).
- How we found the global maximum is the topic of another talk.
Sample Sizes for which $\text{size}(m, n) \leq \alpha = .05$

‡ There are about 660 cases (.4%) for which $\text{size}(m, n) \leq .05$.

‡ There are no cases with $\min\{m, n\} \geq 150$ for which $\text{size}(m, n) \leq .05$.

‡ Having both $m$ and $n$ large does not guarantee $\text{size}(m, n) \leq .05$. 
### Sample Size Pairs in Different Ranges

<table>
<thead>
<tr>
<th>size(m, n)</th>
<th>(m, n) pairs</th>
<th>percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>.00 – .05</td>
<td>667</td>
<td>0.4%</td>
</tr>
<tr>
<td>.05 – .06</td>
<td>126,478</td>
<td>79.1%</td>
</tr>
<tr>
<td>.06 – .07</td>
<td>8,227</td>
<td>5.1%</td>
</tr>
<tr>
<td>.07 – .08</td>
<td>8,331</td>
<td>5.2%</td>
</tr>
<tr>
<td>.08 – .09</td>
<td>8,996</td>
<td>5.6%</td>
</tr>
<tr>
<td>.09 – .10</td>
<td>7,301</td>
<td>4.6%</td>
</tr>
<tr>
<td><strong>total</strong></td>
<td><strong>160,000</strong></td>
<td><strong>100.0%</strong></td>
</tr>
</tbody>
</table>
Recall: size\((m, n) = \sup_{0 < p < 1} P_p(\chi^2 \geq \chi^2_{\alpha, 1})\)

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

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

- \( W(r) \) 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: $m = 400, n = 100, p = .003$

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

\[ \sum_{r=0}^{10} P(\chi^2 \geq \chi^2_{\alpha,1}|X + Y = r)P_p(X + Y = r) = .092. \]
• Does this pattern persist for larger $m$ and $n$?
• Or for larger $(m, n)$ does the size converge to $\alpha$?

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

Why?

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

straight line: $\frac{n}{m} = f$ (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 \( f \).

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

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 
\[ X + Y \sim \text{binomial}(N, p), \quad N = m + n = m + fm, \]
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

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

can be approximated by something that depends only of \( f = n/m \).
Example: $m = 400, n = 100, p = .003, f = .25$

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

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

† The exact size of this test is very close to $\alpha$ 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 \( \alpha \), e.g., \( .047 \leq \text{size}(m, n) \leq .050 \) for almost all \( m \) and \( n \).