Testing the Approximation of Hypergeometric Distribution by the Binomial Distribution

Kayode R. Fowobaje, Anthony I. Wegbom, Igboye S. Aboko And Emmanuel Teju Jolayemi

1Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan
2Department of Statistics, Rivers State College of Art and Science, Port Harcourt
3Department of Statistics, University of Ilorin, Ilorin

Abstracts: This paper examines the approximation of Hypergeometric population by binomial distribution. The influence of the sample size k without replacement under various values of the parameters m and n such that m+n<∞ of the Hypergeometric distribution was considered. Simulation studies using RGui software was used to generate Hypergeometric random variate for various values of the parameters m, n, k starting from m=2, n=2, k=2 in 100 replicates. The goodness-of-fit of the binomial distribution as an approximation to the Hypergeometric distribution simulated data was examined using Pearson chi-square goodness-of-fit test and Likelihood Ratio Test (LRT) were used to test the approximation of the discrete (count) data. Results from this study revealed that the Hypergeometric random variate could asymptotically approach the binomial distribution at m+n≥10 and the absolute difference between m and n is not large i.e. |m-n|= δ for a bounded δ≥0 which changes as m and n changes such that \( \frac{n}{n+m} \approx \frac{m}{n+m} \). In addition as the sample size k approaches n the Hypergeometric population ceases to become binomial. As a measure of goodness-of-fit, the results from chi-square test are more consistent than that of the Likelihood Ratio Test (LRT). We therefore recommend the use of Binomial distribution with parameter \( k, \frac{n}{n+m} \) for a discrete (count) data thought to be from Hypergeometric distribution with parameter (m, n, k) having m+n≥10 and |m-n|= δ which changes as m and n changes such that \( \frac{n}{n+m} \approx \frac{m}{n+m} \).

Keywords: Hypergeometric distribution, Binomial distribution, Simulation, Likelihood ratio test

I. Introduction

Mathematical statistics is the study of statistics from the mathematical point of view, using branches of mathematics such as linear algebra and analysis as well as probability theory Roussas, G.G. (1997). We formulate hypothesis and search for rules to govern our behaviour as regards to them, which will ensure that in the long period of time we shall not be too often wrong.

The Hypergeometric distribution is a finite discrete distribution which arises in situations like; for example, a box contains \( n+m \) balls out of which \( n \) are of one type and the other \( m \) of another type. A sample of size \( k \) is to be randomly chosen (without replacement) in the sense that the set of sampled ball are equally likely to be any of the \( \binom{n+m}{k} \) subset of size \( k \). Roussas, G.G. (1997).

If X represents the number of red balls in the sample, then the probability mass function of X is given as

\[
    f(X = x) = \binom{\binom{n+m}{k}}{\binom{n}{x}} = \max(0, k - n), \ldots, \min(n, k),
\]

Where

\( n \) is the number of variable of interest (good, acceptable, non-defective, etc).
\( m \) is the number of any other variable aside the variable of interest (bad, defective, etc).
\( k \) is the sample size taken.
\( x \) is the number of variable of interest in the sample.

Then the distribution of X is called the Hypergeometric Distribution.

The Binomial distribution is a finite discrete distribution. The Binomial distribution arises in situations where one is observing a sequence of what are known as Bernoulli trials. A Bernoulli trial is an experiment which as only two possible outcomes: success and failure. Furthermore the probability of success \( p \) is fixed and is constant as a result of the sampling scheme (with replacement). A binomial distribution variable counts the number of successes in a sequence of \( k \) independent Bernoulli trials. For \( k \) trials one can obtain between 0 and \( k \) successes.
If X is the random variable denoting the number of successes in k Binomial experiments, then the probability mass function of X is given as

\[ f(X = x) = \binom{k}{x} p^x (1 - p)^{k-x} \]  if \( 0 \leq x \leq k \).

Where
- \( k \) is the sample size.
- \( x \) is the number of success in the sample.
- \( p \) is the probability of success.

\( \binom{k}{x} \) is the binomial coefficient which can be express as \( \frac{k!}{(k-x)!x!} \).

The Distribution of X is called the Binomial Distribution.

The Hypergeometric distribution with parameter \((m, n, k)\) is a very common discrete distribution and is known to have many combinatorial terms whose evaluation becomes cumbersome for large values of term(s). Hence, an approximation is sought for this distribution if possible.

The binomial distribution is less cumbersome to use. The Binomial distribution with sample size \((n + m)\) and probability of success \((p)\) is assumed when taking samples with replacement. The Binomial distribution could be a good approximation to Hypergeometric distribution. Our concern is to determine what values of \( m, n \) and \( k \) are required for this approximation to be valid.

II. Computational Procedure

Recall that the Hypergeometric Distribution is

\[ f(X = x) = \binom{n}{x} \binom{m}{k-x} \binom{n+m}{k}^{-1} \]  \( x = \max(0, k-n), ..., \min(n, k) \),

which we may write as

\[ f(x) = \frac{\Pi_{i=0}^{x-1}(n-i)\Pi_{i=0}^{k-x-1}(m-i)}{\Pi_{i=0}^{x-1}(n + m - i)} \]

Since there are \( k \) terms in the numerator and the denominator we may divide both by \((n + m)^k\) to obtain

\[ f(x) = \frac{\Pi_{i=0}^{x-1}\left(\frac{n}{n+m} - \frac{i}{n+m}\right)\Pi_{i=0}^{k-x-1}\left(1 - \frac{n}{n+m} - \frac{i}{n+m}\right)}{\Pi_{i=0}^{x-1}\left(1 - \frac{i}{n+m}\right)} \]

Since

\[ \left(1 - \frac{n}{n+m} - \frac{k-x-1}{n+m}\right) \leq \left(\frac{n}{n+m} - \frac{i}{n+m}\right) \leq \left(1 - \frac{n}{n+m}\right) \]  \( for \ i = 0, 1, ..., x - 1 \)

\[ \left(1 - \frac{n}{n+m} - \frac{k-x-1}{n+m}\right) \leq \left(\frac{n}{n+m} - \frac{i}{n+m}\right) \leq \left(1 - \frac{n}{n+m}\right) \]  \( for \ i = 0, 1, ..., k - x - 1 \)

We have that

\[ \left(\frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^x \leq \prod_{i=0}^{x-1}\left(\frac{n}{n+m} - \frac{i}{n+m}\right) \leq \left(\frac{n}{n+m}\right)^x \]

\[ \left(1 - \frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^{k-x} \leq \prod_{i=0}^{k-x-1}\left(1 - \frac{n}{n+m} - \frac{i}{n+m}\right) \leq \left(1 - \frac{n}{n+m}\right)^{k-x} \]

it follows that

\[ \left(\frac{k}{x}\right)\left(\frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^x \left(1 - \frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^{k-x} \leq f(x) \leq \left(\frac{k}{x}\right)\left(\frac{n}{n+m}\right)^x \left(1 - \frac{n}{n+m}\right)^{k-x} \]

If we assume that \( \lim_{(n + m) \to \infty} \left(\frac{n}{n+m}\right) = p \) and that \( k \) and \( x \) are fixed then

\[ \lim_{(n + m) \to \infty} \left(\frac{k}{x}\right)\left(\frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^x \left(1 - \frac{n}{n+m} - \frac{k-x-1}{n+m}\right)^{k-x} = \left(\frac{k}{x}\right)p^x (1 - p)^{k-x} \]

and
Testing The Approximation Of Hypergeometric Distribution By The Binomial Distribution

\[
\lim_{(n+m)\to\infty} \frac{k \binom{n}{(n+m)} x \left(1 - \frac{n}{(n+m)}\right)^{k-x}}{\left(1 - \frac{k}{(n+m)}\right)^k} = \binom{k}{x} p^x (1-p)^{k-x}
\]

It follows that

\[
\lim_{(n+m)\to\infty} \frac{k \binom{n}{x} m_{k-x}}{(n+m)_k} = \binom{k}{x} p^x (1-p)^{k-x}
\]

So that the Hypergeometric Distribution can be approximated by Binomial Distribution with \( p = \frac{n}{n+m} \)

### III. Results And Discussion

The scheme used for the simulation studies are described as follows: We simulated a set of \( r \) random variate from Hypergeometric family with parameter \((m, n, k)\). The data were generated at various values of \( m, n, k \) and at different sample sizes. The RGui statistical package was used for simulation; in addition a vcd package which is a goodness-of-fit test used for count data was downloaded and installed on the RGui package. Bates, D.M. (2001). Some complementary tasks were equally performed using some other statistical software like SPSS 17.0 and Microsoft Excel. Example of the RGui code used for simulation studies is given by:

```r
library(vcd)
f = rhyper(r,m,n,k)
gf <- goodfit(f,type= "binomial",par=list(prob=\frac{n}{n+m},size=k))
summary(gf)
plot(gf,main="Count data vs Binomial distribution")
```

where

- `library(vcd)` is use to load the goodness of fit package.
- `f` is the name under which the data set generated are stored.
- `rhyper(r,m,n,k)` generates \( r \) random variates from hypergeometric family with parameter \((m,n,k)\).
- `type` a character string indicating which distribution should be fit (for good–of–fit) or indicating the type of prediction (fitted response or probabilities in predict) respectively.
- `par` a named list giving the distribution parameters (named as in the corresponding density function)
- `summary` is a generic function used to produce result summaries of the results of various model fitting functions.
- `plot` is a generic function for plotting of R objects.
- `main` a character string indicating the title of the plotted R object.

Simulation was done at increasing magnitude of parameters \( m,n,k \) interchangeably until the data set generated is approximate by Binomial.

```r
library(vcd)
f = rhyper(100,2,2,2)
gf <- goodfit(f,type= "binomial",par=list(prob=1/2,size=2))
summary(gf)
```

**Goodness-of-fit test for binomial distribution**

<table>
<thead>
<tr>
<th></th>
<th>X²</th>
<th>df</th>
<th>Pr(X² &gt; X²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>13.68000</td>
<td>20.001070134</td>
<td></td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>14.38723</td>
<td>20.0007513677</td>
<td></td>
</tr>
</tbody>
</table>

**PEARSON CHI-SQUARE TEST**

\( H_0: f \sim B(2, \frac{1}{2}) \) vs \( H_1: \text{Not } H_0 \)

**Decision Rule:** Reject \( H_0 \) if \( P(X^2) \leq \alpha=0.05 \), otherwise do not reject \( H_0 \).

**P-value:** \( P(X^2) = 0.0010701034 \)

**Decision:** Since \( P(X^2) = 0.0010701034 \) is less than \( \alpha=0.05 \), we reject \( H_0 \).

**Conclusion:** We therefore conclude that the data does not follow \( B(2, \frac{1}{2}) \).

**LIKELIHOOD RATIO TEST**

\( H_0: f \sim B(2, \frac{1}{2}) \) vs \( H_1: \text{Not } H_0 \)

**Decision Rule:** Reject \( H_0 \) if \( P(X^2) \leq \alpha=0.05 \), otherwise do not reject \( H_0 \).

**P-value:** \( P(X^2) = 0.0007513677 \)

**Decision:** Since \( P(X^2) = 0.0007513677 \) is less than \( \alpha=0.05 \), we reject \( H_0 \).

**Conclusion:** We therefore conclude that the data does not follow \( B(2, \frac{1}{2}) \).

The goodness-of-fit test implies that using either Pearson \( X^2 \) or Likelihood ratio statistic there is poor fit and it cannot be said that the generated data come from binomial distribution.
> plot(gf, main="Count data vs Binomial distribution")

Fig 3.2.1: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.

> f=rhyper(100,3,2,2)
> gf<-goodfit(f,type= "binomial",par=list(prob=2/5,size=2))
> summary(gf)

Goodness-of-fit test for binomial distribution

<table>
<thead>
<tr>
<th>X²</th>
<th>df</th>
<th>Pr(x² &gt; X²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>27.08333</td>
<td>2</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>31.12879</td>
<td>2</td>
</tr>
</tbody>
</table>

**PEARSON CHI-SQUARE TEST**

H₀: \( f \sim B(2, \frac{2}{5}) \) vs H₁: Not H₀

**Decision Rule:** Reject H₀ if \( P(X^2 > X^2) \leq \alpha = 0.05 \), otherwise do not reject H₀

**Decision:** Since \( P(X^2) = 1.315010e-06 \) is less than \( \alpha = 0.05 \), we reject H₀.

**Conclusion:** We therefore conclude that the data does not follow \( B(2, \frac{2}{5}) \).

**LIKELIHOOD RATIO TEST**

H₀: \( f \sim B(2, \frac{2}{5}) \) vs H₁: Not H₀

**Decision Rule:** Reject H₀ if \( P(X^2 > X^2) \leq \alpha = 0.05 \), otherwise do not reject H₀

**Decision:** Since \( P(X^2) = 1.739677e-07 \) is less than \( \alpha = 0.05 \), we reject H₀.

**Conclusion:** We therefore conclude that the data does not follow \( B(2, \frac{2}{5}) \).

> plot(gf, main="Count data vs Binomial distribution")
Fig 3.2.2: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.

```r
> f=rhyper(100,5,5,2)
> gf<-goodfit(f,type= "binomial",par=list(prob=1/2,size=2))
> summary(gf)

Goodness-of-fit test for binomial distribution

<table>
<thead>
<tr>
<th></th>
<th>X^2</th>
<th>df</th>
<th>Pr(x^2 &gt; X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>2.460000</td>
<td>2</td>
<td>0.2922926</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>2.549166</td>
<td>2</td>
<td>0.2795475</td>
</tr>
</tbody>
</table>

PEARSON CHI-SQUARE TEST

H₀: f~B(2, \frac{1}{2}) vs H₁: Not H₀

**Decision Rule:** Reject H₀ if P(>X^2) ≤ α=0.05, otherwise do not reject H₀.

**P-value:** P(>X^2)=0.2922926

**Decision:** Since P(>X^2)=0.2922926 is greater than α=0.05, we do not reject H₀.

**Conclusion:** We therefore conclude that the data follow B(2, \frac{1}{2}).

LIKELIHOOD RATIO TEST

H₀: f~B(2, \frac{1}{2}) vs H₁: Not H₀

**Decision Rule:** Reject H₀ if P(>X^2) ≤ α=0.05, otherwise do not reject H₀.

**P-value:** P(>X^2)=0.2795475

**Decision:** Since P(>X^2)=0.2795475 is greater than α=0.05, we do not reject H₀.

**Conclusion:** We therefore conclude that the data follow B(2, \frac{1}{2}).

Fig 3.2.3: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.
> f=rhyper(100,5,5,3)
> gf<-goodfit(f,type= "binomial",par=list(prob=1/2,size=3))
> summary(gf)

Goodness-of-fit test for binomial distribution

<table>
<thead>
<tr>
<th></th>
<th>X^2</th>
<th>df</th>
<th>Pr(x^2 &gt; X^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>2.1333333</td>
<td>2</td>
<td>0.5451987</td>
</tr>
<tr>
<td>Likelihood Ratio</td>
<td>2.123368</td>
<td>3</td>
<td>0.5471997</td>
</tr>
</tbody>
</table>

PEARSON CHI-SQUARE TEST

H_0: f ~ B(3, 1/2) vs H_1: Not H_0

Decision Rule: Reject H_0 if P(>X^2) ≤ α=0.05, otherwise do not reject H_0.

Decision: Since P(>X^2)=0.5451987 is greater than α=0.05, we do not reject H_0.

Conclusion: We therefore conclude that the data follow B(3, 1/2).

LIKELIHOOD RATIO TEST

H_0: f ~ B(3, 1/2) vs H_1: Not H_0

Decision Rule: Reject H_0 if P(>X^2) ≤ α=0.05, otherwise do not reject H_0.

Decision: Since P(>X^2)=0.5471997 is greater than α=0.05, we do not reject H_0.

Conclusion: We therefore conclude that the data follow B(3, 1/2).

> plot(gf,main="Count data vs Binomial distribution")

Fig 3.2.4: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial. The rest can be obtained in the table, Table 3.2.1

Table 3.2.1

<table>
<thead>
<tr>
<th>Hyper(r, m, n, k)</th>
<th>H_o</th>
<th>df</th>
<th>Pearson Chi-Square</th>
<th>P-value</th>
<th>Likelihood Ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>100, 5, 5, 2</td>
<td>B(2, 5/3)</td>
<td>2</td>
<td>13.68</td>
<td>0.001</td>
<td>14.39</td>
<td>0.0008</td>
</tr>
<tr>
<td>100, 3, 2, 2</td>
<td>B(2, 1/2)</td>
<td>2</td>
<td>27.08</td>
<td>1.3e-06</td>
<td>31.13</td>
<td>1.7e-07</td>
</tr>
<tr>
<td>100, 4, 5, 2</td>
<td>B(2, 5/3)</td>
<td>2</td>
<td>14.11</td>
<td>0.001</td>
<td>15.28</td>
<td>0.0004</td>
</tr>
<tr>
<td>100, 4, 5, 5</td>
<td>B(5, 5/3)</td>
<td>5</td>
<td>51.59</td>
<td>6.55e-10</td>
<td>63.06</td>
<td>6.60e-13</td>
</tr>
<tr>
<td>100, 5, 5, 2</td>
<td>B(2, 1/2)</td>
<td>2</td>
<td>2.46</td>
<td>0.292</td>
<td>2.55</td>
<td>0.280</td>
</tr>
<tr>
<td>100, 5, 5, 3</td>
<td>B(3, 1/2)</td>
<td>3</td>
<td>2.13</td>
<td>0.545</td>
<td>2.12</td>
<td>0.547</td>
</tr>
<tr>
<td>100, 5, 5, 4</td>
<td>B(4, 1/2)</td>
<td>4</td>
<td>9.92</td>
<td>0.042</td>
<td>16.12</td>
<td>0.001</td>
</tr>
</tbody>
</table>

DOI: 10.9790/5728-1202061016 www.iosrjournals.org 15 | Page
Testing The Approximation Of Hypergeometric Distribution By The Binomial Distribution

| 100,5,5,5 | $B(5,\frac{1}{2})$ | 5 | 18.50 | 2.38e-03 | 24.65 | 1.83e-05 |
| 100,10,5,2 | $B(2,\frac{1}{2})$ | 2 | 118.34 | 2.00e-26 | 84.01 | 5.71e-19 |
| 100,10,5,5 | $B(5,\frac{1}{2})$ | 5 | 451.43 | 2.42e-95 | 217.05 | 8.10e-46 |
| 100,10,8,2 | $B(2,\frac{4}{5})$ | 2 | 7.41 | 0.025 | 7.55 | 0.023 |
| 100,10,8,4 | $B(4,\frac{2}{5})$ | 4 | 15.43 | 0.004 | 16.35 | 0.003 |
| 100,10,8,8 | $B(8,\frac{4}{5})$ | 8 | 48.18 | 9.11e-08 | 48.01 | 3.12e-11 |
| 100,15,16,2 | $B(2,\frac{16}{31})$ | 2 | 1.11 | 0.574 | 1.10 | 0.577 |
| 100,15,16,5 | $B(5,\frac{16}{31})$ | 5 | 8.46 | 0.133 | 9.92 | 0.077 |
| 100,15,16,10 | $B(10,\frac{16}{31})$ | 10 | 16.97 | 0.750 | 18.17 | 0.011 |
| 100,15,16,16 | $B(16,\frac{16}{31})$ | 16 | 25.11 | 6.79e-03 | 32.18 | 8.63e-05 |

IV. Conclusion

After series of simulations studies, the following can be inferred from the results obtained in approximating the generated Hypergeometric data by the Binomial distribution.

1. A good approximation of Hypergeometric distribution by Binomial is found if $m + n \geq 10$, $m$ and $n$ being the parameter of the Hypergeometric distribution.

2. If $|m - n| = \delta$ which changes as $m$ and $n$ changes, the Hypergeometric data set will approximate by Binomial.

3. As $k \rightarrow n$, the Hypergeometric data cannot be approximated by Binomial distribution, $k$ being the number of the selected samples.

4. Likelihood ratio test is more critical than the Pearson Chi-Square in the approximation.

We can conclude that:

1. Hypergeometric data are approximated by Binomial if the absolute difference between the Hypergeometric parameters $m$ and $n$ is not large i.e. $|m-n|=\delta$ for a bounded $\delta \geq 0$ which changes as $m$ and $n$ changes such that $\frac{n}{n+m} \approx \frac{m}{n+m} = p \approx \frac{1}{2}$.

2. The Hypergeometric data set will approximate to Binomial if the sum of $m$ and $n$ is at least ten i.e. $m + n \geq 10$

3. As the sample size tends towards the size of the variable of interest the Hypergeometric data set cease to become Binomial i.e. As $k \rightarrow n$

$$f = rhyper(m, n, k) \neq rbinom\left(k, \frac{n}{m+n}\right)$$

4. In the approximation of Hypergeometric data set to Binomial, the Pearson Chi-square goodness-of-fit test is more robust than the Likelihood ratio test.

V. Recommendation

We therefore recommend the use of Binomial distribution with parameter $(k, \frac{n}{m+n})$ for a discrete (count) data generated from Hypergeometric distribution with parameter $(m, n, k)$ with a population of at least ten i.e. $m+n \geq 10$ and the absolute difference between $m$ and $n$ is not large i.e. $|m-n|=\delta$ for a bounded $\delta \geq 0$ which changes as $m$ and $n$ changes such that $\frac{n}{n+m} \approx \frac{m}{n+m}$.

References