IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 2 Ver. VI (Mar. - Apr. 2016), PP 10-16
www.iosrjournals.org

Testing the Approximation of Hypergeometric Distribution by the Binomial Distribution

[1]Kayode R. Fowobaje, [2]Anthony I. Wegbom, [2]Igboye S. Aboko And [3]Emmanuel Teju Jolayemi

[1]*Department of Epidemiology and Medical Statistics, College of Medicine, University of Ibadan*
[2]*Department of Statistics, Rivers State College of Art and Science, Port Harcourt*
[3]*Department 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 $m+n< \infty$ 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 $m+n\geq10$ and the absolute difference between m and n is not large i.e. $|m-n|= \delta$ for a bounded $\delta\geq0$ which changes as m and n changes such that $\frac{n}{n+m} \simeq \frac{m}{n+m}$. 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 $(k,\frac{n}{m+n})$ for a discrete (count) data thought to be from Hypergeometric distribution with parameter (m, n, k) having $m+n\geq10$ and $|m-n|= \delta$ which changes as m and n changes such that $\frac{n}{n+m} \simeq \frac{m}{n+m}$.

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 theoryRoussas, 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

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

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

$f(X = x) = \binom{k}{x} p^x (1-p)^{k-x} \;\; if\; 0 \le x \le k,$

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 (**k**) 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

$$f(X = x) = \begin{cases} \frac{\binom{k}{x}(n)_x (m)_{k-x}}{(n+m)_k} = \frac{\binom{n}{x}\binom{m}{k-x}}{\binom{n+m}{k}} & x = \max(0, k-n), \dots \min(n,k), \\ 0 \end{cases}$$

which we may write as

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

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

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

Since

$$\left(\frac{n}{(n+m)} - \frac{x-1}{(n+m)}\right) \le \left(\frac{n}{(n+m)} - \frac{i}{(n+m)}\right) \le \left(\frac{n}{(n+m)}\right) \quad for\; i = 0,1,\dots,x-1$$

$$\left(1 - \frac{n}{(n+m)} - \frac{k-x-1}{(n+m)}\right) \le \left(1 - \frac{n}{(n+m)} - \frac{i}{(n+m)}\right) \le \left(1 - \frac{n}{(n+m)}\right) \; for\; i = 0,1,\dots,k-x-1$$

$$\left(1 - \frac{k-1}{(n+m)}\right) \le \left(1 - \frac{i}{(n+m)}\right) \le 1 \; for\; i = 0,1,\dots,k-1$$

We have that

$$\left(\frac{n}{(n+m)} - \frac{x-1}{(n+m)}\right)^x \le \prod_{i=0}^{x-1}\left(\frac{n}{(n+m)} - \frac{i}{(n+m)}\right) \le \left(\frac{n}{(n+m)}\right)^x$$

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

$$\left(1 - \frac{k-1}{(n+m)}\right)^k \le \prod_{i=0}^{k-1}\left(1 - \frac{i}{(n+m)}\right) \le 1$$

it follows that

$$\binom{k}{x}\left(\frac{n}{(n+m)} - \frac{x-1}{(n+m)}\right)^x \left(1 - \frac{n}{(n+m)} - \frac{k-x-1}{(n+m)}\right)^{k-x} \le f(x) \le \frac{\binom{k}{x}\left(\frac{n}{(n+m)}\right)^x \left(1 - \frac{n}{(n+m)}\right)^{k-x}}{\left(1 - \frac{k-1}{(n+m)}\right)^k}$$

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

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

and

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

It follows that

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

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:

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).
typea 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)
summaryis a generic function used to produce result summaries of the results of various model fitting functions.
plotis a generic function for plotting of R objects.
maina 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.

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

	X^2df	Pr($x_2^2 > X^2$)
Pearson	13.68000	20.001070134
Likelihood Ratio	14.38723	20.0007513677

PEARSON CHI-SQUARE TEST

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

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)= 0.0010701034
Decision: Since P(>X^2)= 0.0010701034 is less than α=0.05, we reject H_o.
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: $Not\ H_o$

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)= 0.0007513677
Decision: Since P(>X^2)= 0.0007513677 is less than α=0.05, we reject H_o.
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")

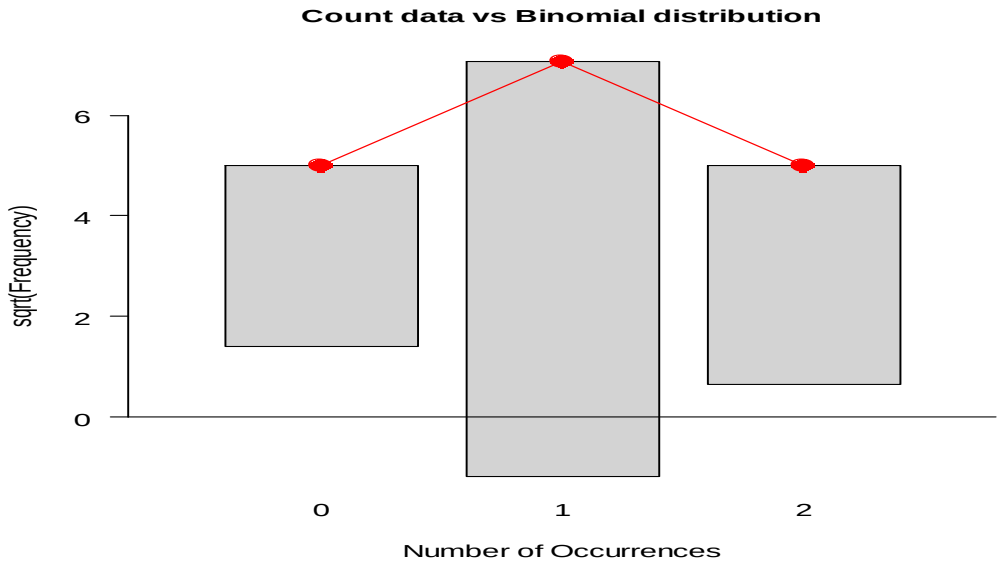

Fig 3.2.1: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.

> 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

	X^2	df	$\Pr(x_2^2 > X^2)$
Pearson	27.08333	2	1.315010e-06
Likelihood Ratio	31.12879	2	1.739677e-07

PEARSON CHI-SQUARE TEST

H_0: $f \sim B(2, \frac{2}{5})$ vs H_1: $Not\ H_o$

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=1.315010e-06
Decision: Since P(>X^2)=1.315010e-06 is less than α=0.05, we reject H_o.
Conclusion: We therefore conclude that the data does not follow$B(2, \frac{2}{5})$.

LIKELIHOOD RATIO TEST

H_0: $f \sim B(2, \frac{2}{5})$ vs H_1: $Not\ H_o$

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=1.739677e-07
Decision: Since P(>X^2)=1.739677e-07 is less than α=0.05, we reject H_o.
Conclusion: We therefore conclude that the data does not follow$B(2, \frac{2}{5})$.

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

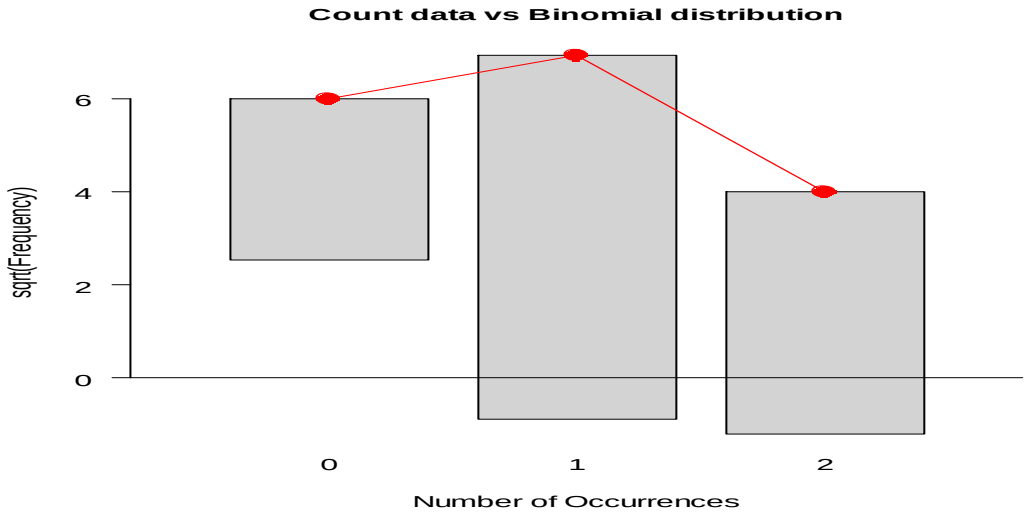

Fig 3.2.2: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.> f=rhyper(100,5,5,2)
>gf<-goodfit(f,type= "binomial",par=list(prob=1/2,size=2))
>summary(gf)
Goodness-of-fit test for binomial distribution

	X^2	df	$\Pr(x_2^2 > X^2)$
Pearson	2.460000	2	0.2922926
Likelihood Ratio	2.549166	2	0.2795475

PEARSON CHI-SQUARE TEST

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

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=0.2922926
Decision: Since P(>X^2)=0.2922926 is greater than α=0.05, we do not reject H_o.
Conclusion: We therefore conclude that the data follow $B(2, \frac{1}{2})$.

LIKELIHOOD RATIO TEST

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

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=0.2795475
Decision: Since P(>X^2)=0.2795475 is greater than α=0.05, we do not reject H_o.
Conclusion: We therefore conclude that the data follow $B(2, \frac{1}{2})$.
>plot(gf,main="Count data vs Binomial distribution")

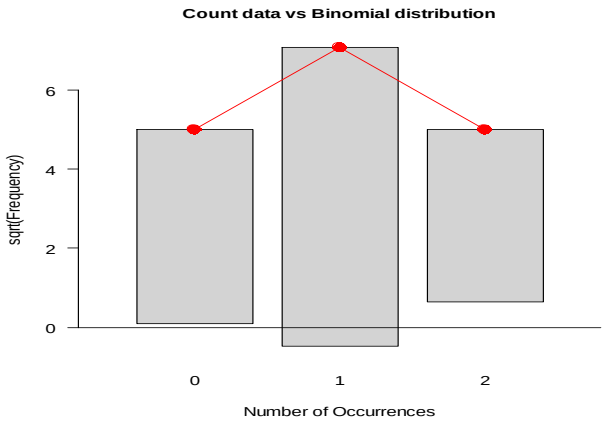

Fig 3.2.3: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial.

> 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

	X^2	df	$Pr(x_3^2 > X^2)$
Pearson	2.1333333		0.5451987
Likelihood Ratio	2.123368	3	0.5471997

PEARSON CHI-SQUARE TEST

H_0: $f \sim B(3, \frac{1}{2})$ vs H_1: $Not\ H_o$

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=0.5451987
Decision: Since P(>X^2)=0.5451987 is greater than α=0.05, we do not reject H_o.
Conclusion: We therefore conclude that the data follow$B(3, \frac{1}{2})$.

LIKELIHOOD RATIO TEST

H_0: $f \sim B(3, \frac{1}{2})$ vs H_1: $Not\ H_o$

Decision Rule: Reject H_o if P(>X^2) ≤ α=0.05, otherwise do not reject H_o
P-value: P(>X^2)=0.5471997
Decision: Since P(>X^2)=0.5471997 is greater than α=0.05, we do not reject H_o.
Conclusion: We therefore conclude that the data follow$B(3, \frac{1}{2})$.

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

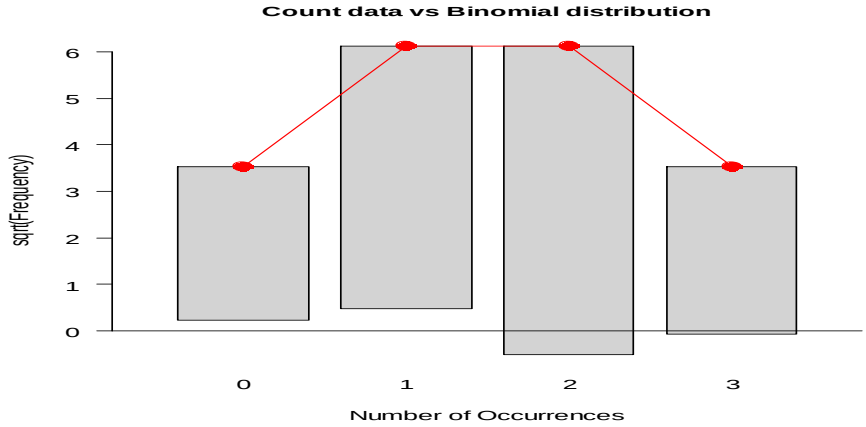

Fig 3.2.4: The bar chat of counts simulated from Hypergeometric population with the density plot of Binomial. The rest can be obtained in the table, Table 3.2.1

Table 3.2.1

1. SIMULATION

Hyper(r, m, n, k)	**H_o**	**df**	**Pearson Chi-Square**	**P-value**	**Likelihood Ratio**	**P-value**
100,2,2,2	$B(2, \frac{1}{2})$	2	13.68	0.001	14.39	0.0008
100,3,2,2	$B(2, \frac{2}{5})$	2	27.08	1.3e-06	31.13	1.7e-07
100,4,5,2	$B(2, \frac{5}{9})$	2	14.11	0.001	15.28	0.0004
100,4,5,5	$B(5, \frac{5}{9})$	5	51.59	6.55e-10	63.06	6.60e-13
100,5,5,2	$B(2, \frac{1}{2})$	2	2.46	0.292	2.55	0.280
100,5,5,3	$B(3, \frac{1}{2})$	3	2.13	0.545	2.12	0.547
100,5,5,4	$B(4, \frac{1}{2})$	4	9.92	0.042	16.12	0.001

100,5,5,5	$B(5,\frac{1}{2})$	5	18.50	2.38e-03	24.65	1.83e-05
100,10,5,2	$B(2,\frac{1}{3})$	2	118.34	2.00e-26	84.01	5.71e-19
100,10,5,5	$B(5,\frac{1}{3})$	5	451.43	2.42e-95	217.05	8.10e-46
100,10,8,2	$B(2,\frac{4}{9})$	2	7.41	0.025	7.55	0.023
100,10,8,4	$B(4,\frac{4}{9})$	4	15.43	0.004	16.35	0.003
100,10,8,8	$B(8,\frac{4}{9})$	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$|\mathrm{m} - \mathrm{n}| = \delta$ which changes as m and n changes,the Hypergeometric data set will approximate by Binomial.
3. As $k \to n$, 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|= δ for a bounded δ≥0 which changes as m and n changes such that $\frac{n}{n+m} \simeq \frac{m}{n+m} = p \approx \frac{1}{2}$.
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 $k \to n$

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

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≥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

$\frac{n}{n+m} \simeq \frac{m}{n+m}$.

References

[1]. D.M. BATES, (2001). "Using Open Source Software to Teach Mathematical Statistics", http://www.stat.wisc.edu/~bates/JSM2001.pdf

[2]. R CORE DEVELOPMENT TEAM, (2004). An introduction to R, Release 2.0.1, http://cran.r-project.org/doc/manuals/R-intro.pdf

[3]. Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". The Annals of Mathematical Statistics9: 60–62. doi:10.1214/aoms/1177732360.edit

[4]. Chernoff, H. and Lehmann E.L. (1954). "The use of maximum likelihood in χ^2 tests for goodness-of-fit". The Annals of Mathematical Statistics 25: 579-586.

[5]. Roussas, G. G. (1997)."A course in Mathematical Statistics", Intercollege Division of Statistics, California

[6]. Cassella, G. and Berger, R.L. (2002). Statistical Inference. Duxbury:Pacific Grove, CA.

[7]. Feller, W. (1968). "The Hypergeometric Series." An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd edition. New York: Wiley, pp. 41-45.

[8]. Spiegel, M. R. (1992). Theory and Problems of Probability and Statistics. New York: McGraw-Hill, pp. 113-114, 1992.

[9]. Shao, Jun (2003).Mathematical Statistics, second edition, Springer.