UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2009, Vol. 2, No. 1

A Note About Maximum Likelihood Estimator in Hypergeometric Distribution

Una nota sobre los estimadores de máxima verosimilitud en la distribución hipergeométrica

Hanwen Zhang[a]
hanwenzhang@usta.edu.co

Abstract

The method of maximum likelihood estimation is one of the most important statistical techniques, and it is widely used by statistical scientists. However for the hypergeometric distribution $Hg(n, R, N)$, the maximum likelihood estimators of N and R are not clear in most of the statistical texts. In this paper, rigorous procedures in order to find the maximum likelihood estimator of N and R in a hypergeometric distribution are presented.

***Key words*:** Hypergeometric distribution, maximum likelihood estimation..

Resumen

El método de la estimación de máxima verosimilitud es una de las técnicas más importantes de la estadística y es utilizado ampliamente por los profesionales estadísticos. Sin embargo, para la distribución hipergeométrica $Hg(n, R, N)$, los estimadores de máxima verosimilitud de N y R no son claros en la mayoría de los textos estadísticos. En este artículo, se presentan los procedimientos rigurosos para encontrar estos estimadores de máxima verosimilitud.

***Palabras clave*:** Distribución hipergeométrica, estimación de máxima verosimilitud..

1 Introduction

The method of maximum likelihood estimation is one of the most important statistical technique, and is one of the concepts that every student of statistics should

[a] Docente. CIEES, Universidad Santo Tomás

be familiarized with. The maximum likelihood estimation of parameters in most of the probabilistic distribution is easy to handel, but the problem of estimating the population size in the capture-recapture problem (see Ardilly & Tillé (2006)), that is, the problem of estimating N in a hypergeometric distribution $Hg(n, R, N)$, is not very clear in statistical inference texts, since this problem is not considered in most of them, for example, Mood & Boes (1974), Bickel & Doksum (2001) and Casella & Berger (2002). Although the estimator of maximum likelihood of N is given by Shao (2003), the procedure is not quite clear. Also we may need to estimate the parameter R, that is the size of a subpopulation, and the same situation occurs with this problem of estimation. The goal of this paper is to provide the rigorous procedures in order to find the maximum likelihood of the parameters N and R.

2 Estimating N

Suppose that $X \sim Hg(n, R, N)$, and we need to estimate the poblation size N, the likelihood function is:

$$L(N) = L(x, N) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}, \tag{1}$$

if $\max(n - N + R, 0) \leq x \leq \min(R, n)$. The function $L(N)$ is a discreet function, so the way to find the maximum is not taking derivatives with respect to N, but finding out where $L(N)$ is a creasing function of N, and where is decreasing. In this way, we compute the ratio $D(N) = L(N)/L(N-1)$, and we find out where $D(N) > 1$ and where $D(N) < 1$. With a simple algebraic reasoning, we have:

$$D(N) = \frac{\binom{N-R}{n-x}\binom{N-1}{n}}{\binom{N}{n}\binom{N-R-1}{n-x}} > 1,$$

which is equivalent to

$$D(N) = \frac{(N-R)!(N-1)!(N-n)!(N-R-1-n+x)!}{(N-R-1)!N!(N-n-1)!(N-R-n+x)!} > 1.$$

Canceling all posible factorials, it can be shown that $D(N) > 1$ if and only if $(N - R)(N - n) > N(N - R - n + x)$, so the values of N where $D(N) > 1$ are defined by $N < Rn/x$. Similarly, $D(N) < 1$ for $N > Rn/x$.

Based on the former argument, some teachers mistakenly establish that the maximum likelihood estimator of N is Rn/X. The previous statement is not correct, since according to the definition of the maximum likelihood estimator, the realization of the estimator must fall in the parameter space. And clearly the realization of the statistic Rn/X may be fractional, and it is immediately disqualified to be the maximum likelihood estimator of N.

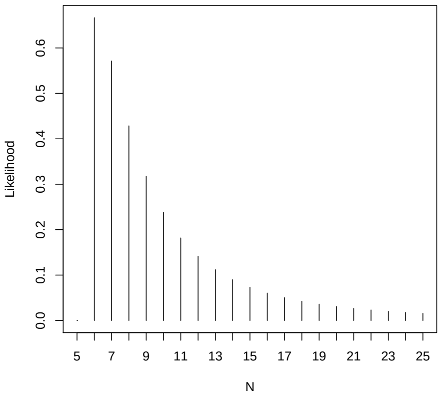

Figure 1: Likelihood function with $R = 5$, $n = 4$ and $x = 3$.

Shao (2003, pg.277) gives the correct maximum likelihood estimator of N, that is, $\hat{N}_{MV} = [Rn/X]$ where $[a]$ denote the integer part of a. Note that the creasing and decreasing property of $L(N)$ leads two candidates to be the maximum likelihood estimator of N: $[Rn/X]$ and $[Rn/X] + 1$. However Shao draw the conclusion without a further comment.

To verify that the maximum likelihood estimator N is, indeed, $[Rn/X]$, we recall that $D(N) < 1$ if $N > Rn/x$. It is clear that $[Rn/X]+1 > Rn/x$, so $D([Rn/X]+1) < 1$ that is $L([Rn/X] + 1)/L([Rn/X]) < 1$, so we conclude that

$$L([Rn/X] + 1) < L([Rn/X]). \tag{2}$$

Further more we can conclude that $[Rn/X]$ is the unique maximum likelihood estimator since in (2) we can never have the equality.

Suppose that $R = 5$, $n = 4$ and $x = 3$, in this case, $[Rn/x] = [6.67] = 6$ is the maximum likelihood estimate of N. Figure 1 illustrates the likelihood function $L(N)$ where clearly the maximum situates at 6.

3 Estimating R

Suppose that the parameter of interest is R, the number of individuals with some specific characteristic in a population of size N, and in a sample of size n, x individuals with this characteristic are selected. So we consider the likelihood function as a function of R, that is

$$L(R) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{N}{n}}, \tag{3}$$

and according to the discussed in the previous section, we compute the ratio $D(R) = L(R)/L(R+1)$, and we find out for what values of R, $D(R) > 1$ and for what values, $D(R) < 1$. We have

$$D(R) = \frac{\binom{R}{x}\binom{N-R}{n-x}}{\binom{R+1}{x}\binom{N-R-1}{n-x}} > 1 \tag{4}$$

is equivalent to

$$D(R) = \frac{(R+1-x)(N-R)}{(R+1)(N-R-n+x)} > 1, \tag{5}$$

which is equivalent to $(R+1-x)(N-R) > (R+1)(N-R-n+x)$. And this leads to $D(R) > 1$ if and only if $R > \dfrac{x(N+1)-n}{n}$, that is, $L(R)$ is decreasing for integers larger than $\dfrac{x(N+1)-n}{n}$ and increasing for integers smaller than $\dfrac{x(N+1)-n}{n}$. And the intuition leads that $\hat{R}_{MV} = \dfrac{x(N+1)-n}{n}$, but this is true when $\dfrac{x(N+1)-n}{n}$ is an integer, and in this case $D(R) = 1$, so

$$L\left(\frac{x(N+1)-n}{n}\right) = L\left(\frac{x(N+1)-n}{n} + 1\right),$$

and the maximum likelihood estimator may be either of $\dfrac{x(N+1)-n}{n}$ and $\dfrac{x(N+1)-n}{n} + 1$, and is not unique.

If $\dfrac{x(N+1)-n}{n}$ is not integer, the value of R that maximizes $L(R)$ is $[\dfrac{x(N+1)-n}{n}]$ or $[\dfrac{x(N+1)-n}{n}]+1$ considering that the estimator of R must take values in the set of the integers. However, recalling the previous argument we can state that

$$L\left([\frac{x(N+1)-n}{n}]\right) < L\left([\frac{x(N+1)-n}{n}] + 1\right)$$

since

$$[\frac{x(N+1)-n}{n}] < \frac{x(N+1)-n}{n},$$

and in this case, $\hat{R}_{MV} = [\dfrac{x(N+1)-n}{n}] + 1 = [\dfrac{x(N+1)}{n}]$.

Summarizing, we have that

$$\hat{R}_{MV} = \begin{cases} \dfrac{x(N+1)}{n} - 1 \text{ or } \dfrac{x(N+1)}{n} & \text{if } \dfrac{x(N+1)}{n} \text{ is an integer} \\ [\dfrac{x(N+1)}{n}] & \text{if } \dfrac{x(N+1)}{n} \text{ is not an integer} \end{cases} \tag{6}$$

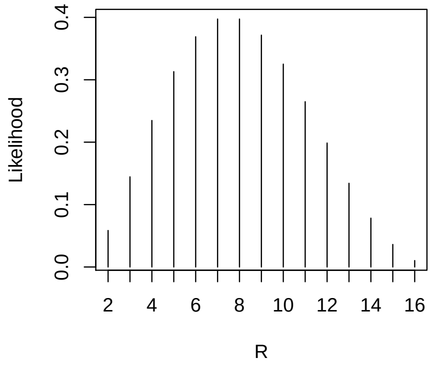

Figure 2: Likelihood function with $N = 19$, $n = 5$ and $x = 2$.

Suppose that $N = 19$, $n = 5$ and $x = 2$, in this case, $\dfrac{x(N+1)}{n} = 8$ is an integer, so the maximum likelihood estimate of R may be 7 or 8. Figure 2 illustrates the likelihood function $L(R)$ where clearly, the maximum of the function situates at 7 and 8.

Now suppose that the $N = 20$, $n = 5$ and $x = 2$, then $\dfrac{x(N+1)}{n} = 8.4$, so the maximum likelihood estimate of R is $[8.4] = 8$. Figure 3 illustrates the likelihood function $L(R)$ where the maximum of the function situates at 8.

4 Conclusion

In this paper, the procedures of maximum likelihood estimation in hypergeometric distribution are presented, and we point out that because the parameters N and R are integers, the maximization of the likelihood function needs to be handled carefully with more detailed analysis.

References

Ardilly, P. & Tillé, Y. (2006), *Sampling Methods: Exercises and Solutions.*, Springer.

Bickel, P. & Doksum, K. (2001), *Mathematical Statistics. Basic Ideas and Selected Tipics.*, Vol. I, second edn, Prentice-Hall.

Casella, G. & Berger, R. (2002), *Statistical Inference.*, second edn, Duxbury Pres.

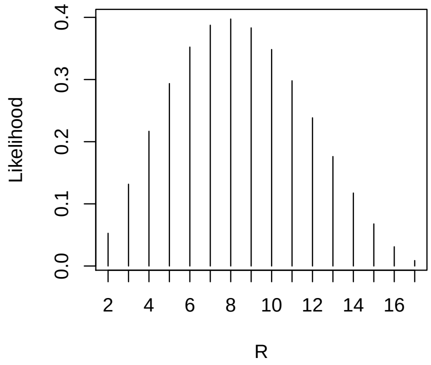

Figure 3: Likelihood function with $N = 20$, $n = 5$ and $x = 2$.

Mood, A.M, G. F. & Boes, D. (1974), *Introduction to the theory of statistics.*, International edition. McGraw Hill.

Shao, J. (2003), *Mathematical statistics.*, second edn, Springer.