Quaderni di Statistica
Vol. 3, 2001

Some Approximations for the Asymptotic Variance of the Maximum Likelihood Estimator of the Parameter in the Inverse Hypergeometric Random Variable

Domenico Piccolo
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
Centro per la Formazione, Portici
E-mail: dopiccol@unina.it

Summary: We derive some approximations for the asymptotic variance of the Maximum Likelihood estimator for the parameter of the Inverse Hypergeometric random variable. For most statistical models, the asymptotic variance is usually derived after some algebraic manipulations. In this paper, we show that this lengthy calculations can be overcome by simple and accurate linear approximations. The interest for this result arises from a statistical model for preferences that has been recently proposed for evaluation studies, preferences analyses and marketing researches.

Keywords: Inverse Hypergeometric Random Variable, Maximum Likelihood Estimation, Asymptotic Variance

1. Introduction

The Inverse Hypergeometric (IHG) random variable is a discrete probabilistic model firstly discussed in Wilks (1963, pp. 141-143) and fully developed by Guenther (1975). As for most of the discrete models, the IHG random variable is generated by consecutive drawing of balls from an urn that contains a known composition of coloured balls.

In fact, while the Binomial and Hypergeometric random variables are generated by drawing a *fixed* number of balls (with and without replacement, respectively), the Inverse Binomial and Inverse Hypergeometric random variables are defined as the number of drawings (with and without replacement, respectively) necessary in order to achieve *a prefixed number of balls*. In this sense, Binomial and Hypergeometric probabilistic models refer to a direct drawing while Inverse Binomial and Inverse Hypergeometric refer to an inverse drawing. As pointed out by Wilks, the inverse scheme of drawing can be interpreted as a *discrete waiting time modeling.*

The interest in modeling procedures for the preferences has recently arisen in the literature as it is shown by the works by Marden (1995) and Taplin (1997). In this area, D'Elia (1999; 2000) has proposed to use the IHG random variable to model the preferences that a sample of subjects expresses towards an ordered collection of objects.

In fact, the IHG random variable can be defined as the number of drawings necessary in order to achieve a "first success": this number can be interpreted as the rank associated with an item in a set of ordered objects, services, brands, opinions, etc. Also, the performance of the model has been successfully exploited in many fields for the evaluation of people, methods and structures (D'Elia, 2001a).

The paper is organized as follows. In the next section, we discuss the probabilistic implications of the IHG random variable; then, in section 3, we derive the asymptotic variance of the maximum likelihood (ML) estimator of the parameter of interest providing some recursive formula. Then, in section 4, we develop an approximation for this variance and discuss its accuracy. Finally, some concluding remarks end the paper.

2. The probabilistic model and its implications

In a series of consecutive drawings without replacement of balls from an urn containing B white balls and $m-1$ not-white balls, we define R as the number of drawings necessary to firstly obtain a white ball. Of course, the support of R is $\{1,2,\ldots,m\}$ depending on the circumstance if the white ball is drawn at first, ..., or at most after m drawings.

For a fixed m, the sample space consists of a partition of $\binom{B+m-1}{m-1}$ equiprobable elementary events and $\binom{B+m-1-r}{m-r}$ of them generate exactly the event $(R=r)$, $r=1,2,\ldots,m$. Thus, the probability mass function is given by:

$$Pr(R=r)=\frac{\binom{B+m-1-r}{m-r}}{\binom{B+m-1}{m-1}},\ r=1,2,\ldots,m.$$

This expression can be interpreted in a straight way since:

$$Pr(R=1)=\frac{B}{B+m-1};Pr(R=2)=\frac{m-1}{B+m-1}\frac{B}{B+m-2};$$

$$Pr(R=3)=\frac{m-1}{B+m-1}\frac{m-2}{B+m-2}\frac{B}{B+m-3};\ldots$$

These results confirm the structure of the random experiment and, moreover, they show that any probability is a function of the ratio $B/(B+m-1)$, that is the probability of drawing a white ball firstly. In a preference model, this expression is the probability of the best selection $(R=1)$ of an item among m similar items. Thus, we define this quantity as the *preference parameter* of the model by letting:

$$Pr(R=1) = \frac{B}{B+m-1} = \theta.$$

Then, in the following, for a fixed m, we re-parameterize the IHG random variable by means of the θ parameter instead of B.

This step deserves some comments. We introduced the parameter B as the (*discrete*) number of white balls in the urn, so that the parameter space of B is $\Omega(B)=\{B: B=0,1,2,\ldots\}$. On the other end, we are now introducing θ, which is a probability, as a (*continuous*) parameter expressing the likelihood of the event, and then the parameter space of θ is $\Omega(\theta)=\{\theta: 0<\theta\leq 1\}$.

To overcome the problem, for any fixed m, we introduce a theoretical urn such that, for any $\theta\in(0,1]$, there exists a B^* such that $|\frac{B^*}{B^*+m-1}-\theta|<\epsilon$, for any small $\epsilon>0$. If B^* is an integer number then the urn has a physical meaning; otherwise, the model refers to an unfeasible urn with a real number of B^* white balls. In this paper, the IHG random variable will be examined in term of the preference parameter θ and a theoretical urn system will be associated to it.

An immediate advantage of this approach is that the preference parameter is invariant with respect of the number of items; thus, we could compare the degree of preference for different objects, brands, professions, colours, and so on.

After some algebra, the new parameterization is the following:

$$Pr(R=r) = \begin{cases} \theta, & r=1; \\ c_r\,\theta\,(1-\theta)^{r-1}\prod_{s=1}^{r-1}(m-s-1+s\theta)^{-1}, & r=2,\ldots,m; \end{cases}$$

where $c_r = \prod_{s=1}^{r-1}(m-s) = (m-1)!/(m-r)!$, $r=2,\ldots,m$.

Alternative parameterizations can be derived. For instance, one is given by:

$$Pr(R=r)=\theta\left[\mathcal{J}(r=1)+\mathcal{J}(r\neq 1)\prod_{s=1}^{r-1}\frac{(m-s)(1-\theta)}{m-1-s(1-\theta)}\right],r=1,2,\ldots,m;$$

where $\mathcal{J}(.)$ is the indicator function:

$$\mathcal{J}(A)=\begin{cases}1\text{, if the condition } A \text{ is } true;\\ 0\text{, if the condition } A \text{ is } false.\end{cases}$$

Notice that, from a computational point of view, it is more efficient to introduce the recursive formula :

$$Pr(R=1)=\theta;\ \ Pr(R=r+1)=Pr(R=r)(1-\theta)\frac{m-r}{m-r-1+r\theta},r=1,\ldots,m-1.$$

From the previous expression, we deduce that the IHG random variable has a mode either in $R=1$, if $\theta>1/m$, or in $R=m$, if $\theta<1/m$. If $\theta=1/m$, then the IHG random variable coincides with a discrete Uniform random variable defined over the support $\{1,2,\ldots,m\}$. Thus, the IHG random variable distribution is:

i) a monotone decreasing function when $\theta>1/m$ (positive asymmetry);

ii) a monotone increasing function when $\theta<1/m$ (negative asymmetry);

iii) a symmetric constant distribution over the first m integers when $\theta=1/m$.

The mean value and the variance are, respectively:

$$\mathbb{E}(R)=\frac{m-\theta}{1+\theta(m-2)};\ Var(R)=\frac{(m-1)^2(m-\theta)\theta(1-\theta)}{[1+\theta(m-2)]^2[2+\theta(m-3)]}.$$

At this point, it is useful to derive explicitely the probability distributions of the IHG random variables for the first values of m. They are shown in the Table 1 for $m=1,2,\ldots,5$.

First of all, we note that the case $m=1$ refers to a degenerate random variable that assumes the value $(R=1)$ with probability 1, since it refers to an urn consisting of B white balls, so that $\theta=1$.

Table 1. Probability distributions, mean value and variance for the Hypergeometric random variables, m=1,2,...,5.

	$m=1$	$m=2$	$m=3$	$m=4$	$m=5$
$Pr(R=1)$	1	θ	θ	θ	θ
$Pr(R=2)$		$1-\theta$	$\frac{2\theta(1-\theta)}{(1+\theta)}$	$\frac{3\theta(1-\theta)}{(2+\theta)}$	$\frac{4\theta(1-\theta)}{(3+\theta)}$
$Pr(R=3)$			$\frac{(1-\theta)^2}{(1+\theta)}$	$\frac{6\theta(1-\theta)^2}{(2+\theta)(1+2\theta)}$	$\frac{6\theta(1-\theta)^2}{(1+\theta)(3+\theta)}$
$Pr(R=4)$				$\frac{2(1-\theta)^3}{(2+\theta)(1+2\theta)}$	$\frac{12\theta(1-\theta)^3}{(1+\theta)(1+3\theta)(3+\theta)}$
$Pr(R=5)$					$\frac{3(1-\theta)^4}{(1+\theta)(1+3\theta)(3+\theta)}$
$Pr(R=6)$					
$Pr(R=7)$					
$\mathbb{E}(R)$	1	$2-\theta$	$\frac{3-\theta}{1+\theta}$	$\frac{4-\theta}{1+2\theta}$	$\frac{5-\theta}{1+3\theta}$
$Var(R)$	0	$\theta(1-\theta)$	$\frac{2\theta(1-\theta)(3-\theta)}{(1+\theta)^2}$	$\frac{9\theta(1-\theta)(4-\theta)}{(2+\theta)(1+2\theta)^2}$	$\frac{8\theta(1-\theta)(5-\theta)}{(1+\theta)(1+3\theta)^2}$

Then, the case $m=2$ refers to a Bernoulli trial ("success-failure" system, or an urn with 2 balls one of which is white). The IHG random variable is defined as the number of drawings necessary for a first "success" (that is the drawing of a white ball). Thus, its support is $\{1,2\}$ with a probability distribution defined by:

$$Pr(R=1)=\theta; \quad Pr(R=2)=1-\theta.$$

The random variable R is a shifted Bernoulli random variable in the sense that if $X \sim Ber(\theta)$ than $R=2-X$. The mean value and the variance are:

$$\mathbb{E}(R) = 2 - \theta; \, Var(R) = \theta(1-\theta).$$

The case $m = 3$ has been fully discussed by Piccolo (2000) as a convenient model for ranking preferences in many political and sociological fields. For this distribution it is possible to derive an explicit expression for the ML estimator of the preference parameter. Moreover, the moment generating function is:

$$G(t) = \frac{e^t}{1+\theta}\Big\{\theta + \big[\theta + e^t(1-\theta)\big]^2\Big\},$$

and the mean value and the variance are:

$$\mathbb{E}(\mathcal{R}) = \frac{3-\theta}{1+\theta}; \quad Var(\mathcal{R}) = \frac{2\,\theta\,(1-\,\theta)(3-\theta)}{(1+\theta)^2}.$$

The case $m = 4$ is quite common in evaluation studies, mostly in Education and Marketing, where respondents are asked to choose among two ordered disagreement and two ordered agreement answers. Some experiences have shown that when the number of alternatives is even and small, an amount of undesirable variability can mask the real agreement of the subjects since an unknown proportion of indifference answers may be uncorrectly reported as likeness or disagreement.

Finally, we find interesting to relate the IHG random variable to an *index of positive evaluation (IVP)*, defined as the relative frequency of respondents which agree with the service to be rated. This measure has been recently adopted by official institutions (as the Italian Ministry of Education, University and Research, for instance).

Then, when the scale is rated on m points:

$$IVP = Pr\Big(R \leq \Big[\frac{m}{2}\Big]\Big).$$

Some current studies set $m = 4$ which implies:

$$IVP = Pr\,(R \leq 2\,).$$

Thus, if we adopt the IHG random variable as a coherent probabilistic tool for modeling the ranks, it is immediate to derive the following formulas:

$$IVP = \frac{\theta(5 - 2\,\theta)}{2 + \theta};$$

$$\theta = \frac{1}{4}\left[(5 - IVP) - \sqrt{(5 - IVP)^2 - 16\,IVP}\;\right].$$

This shows the one-to-one relationship between the IVP measure and the preference parameter θ of the IHG random variable.

3. Maximum likelihood estimation of the preference parameter

In this section, we discuss the ML estimation of the preference parameter of the IHG random variable and derive the formulas for the asymptotic variance of its estimator.

Let $p_r(\theta) = Pr(R = r \mid \theta)$, and consider the random sample of observed ranks $(r_1, r_2, \ldots, r_n)$. The latter conveys an amount of information about the parameter θ equivalent to that of the sample collection $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)'$ of the absolute frequencies of the ranks $(R = 1)$, $(R = 2)$, ..., $(R = m)$, respectively.

Thus, the log-likelihood function can be written as:

$$l(\theta; \boldsymbol{n}) = \sum_{r=1}^{m} n_r \, log(p_r(\theta)).$$

D'Elia (2001b) showed that the ML estimator T_n of θ is always well defined since it is the unique solution of an $(m-1)$-degree polynomial equation in θ, being $\theta \in \Omega(\theta) = \{\theta : 0 \leq \theta \leq 1\}$. This solution can be found analitically for $m \leq 5$, although it has a simple form only for $m = 3$ (Piccolo, 2000).

Exploiting standard results on the ML estimation (Serfling, 1980), the asymptotic variance of T_n can be derived by the Cramér theorem :

$$Var(T_n) = \frac{1}{n}\left(\sum_{r=1}^{m}\frac{\{p_r'(\theta)\}^2}{p_r(\theta)}\right)^{-1}.$$

Because of the consistency of the ML estimator and the continuity of $p_r(\theta)$, we can use $Var(T_n)$ for asymptotic inference about θ.

The previous formula can be evalued efficiently by recursion. Given m, it is possible to compute jointly the probabilities $p_r(\theta)$, their derivatives $p_r'(\theta)$ with respect to θ, and the addends $v_r(\theta) = \frac{\{p_r'(\theta)\}^2}{p_r(\theta)}$. Some simple but lengthy algebra yields the following recursion:

$$\boldsymbol{g}_{r+1} = \boldsymbol{A}_r\, \boldsymbol{g}_r\,,\ r = 1, 2, \ldots, m-1;$$

where for $r = 1, 2, \ldots, m-1$:

$$\boldsymbol{g}_r = \begin{pmatrix} p_r(\theta) \\ p_r'(\theta) \\ v_r(\theta) \end{pmatrix};$$

$$\boldsymbol{A}_r = \begin{pmatrix} a_r(\theta) & 0 & 0 \\ a_r'(\theta) & a_r(\theta) & 0 \\ \frac{a_r'(\theta)}{a_r(\theta)} & 2\, a_r'(\theta) & a_r(\theta) \end{pmatrix};$$

$$a_r(\theta) = (1-\theta)\frac{m-r}{m-r-1+r\theta};$$

$$a_r'(\theta) = \frac{-(m-r)(1+r-r\theta)}{(m-r-1+r\theta)^2}.$$

We observe that the asymptotic variance $Var(T_n)$, for $m \geq 2$, includes the factor $\theta(1-\theta)/n$. Thus, it is convenient to remove it before studying an approximation.

The Table 2 shows the quantities $g_m(\theta) = \frac{n}{\theta(1-\theta)} Var(T_n)$ for $m=2,3,\ldots,5$. These expressions have been obtained by a procedure written in the Maple V language (reported in Appendix).

Table 2. *Factors in the asymptotic variance of the ML estimator*

m	$g_m(\theta) = \frac{n}{\theta(1-\theta)} Var(T_n)$
2	1
3	$\frac{\theta^2+2\theta+1}{\theta^2+3}$
4	$\frac{4\theta^4+20\theta^3+33\theta^2+20\theta+4}{4\theta^4+2\theta^3+33\theta^2+20\theta+22}$
5	$\frac{9\theta^6+78\theta^5+247\theta^4+356\theta^3+247\theta^2+78\theta+9}{9\theta^6+12\theta^5+145\theta^4+224\theta^3+379\theta^2+180\theta+75}$

4. A simple approximation for the asymptotic variance

Although recursive expressions can be derived for any m, they are rather cumbersome to be used in the inferential procedures (common situations range up to $m=14$, for instance). Thus, our next objective is to derive a simple and effective expression for the asymptotic variance of the ML estimator for $m \geq 4$. In fact, we exclude from the analysis the cases

$m = 2$ and $m = 3$, since the expressions for the asymptotic variance is very simple and, of course, it is best than any approximation.

Some algebraic considerations support the following statements:

i) the quantities $g_m(\theta)$ are the ratio of two homogenous polynomials of degrees $2(m-2)$;

ii) the leading coefficients of these polynomials are the same;

iii) the polynomial in the numerator is always symmetric since it is the product of the square of simple expressions;

iv) $g_m(1) = 1, \forall m = 2, 3, \ldots;$

v) $g_m(0) \to c > 0$, for $m \to \infty;$

vi) although the numerical value of the coefficients increases with the degree of the polynomials, the numerator and the denominator of the quantities $g_m(\theta)$ are bounded since they are linear combinations of the variable θ which belongs to $(0,1]$;

vii) the functions $g_m(\theta)$ are well approximated by linear relationships over the whole range of $\Omega(\theta)$, with coefficients depending by m, as the Figure 1 shows for $m = 4, 5, \ldots, 15.$

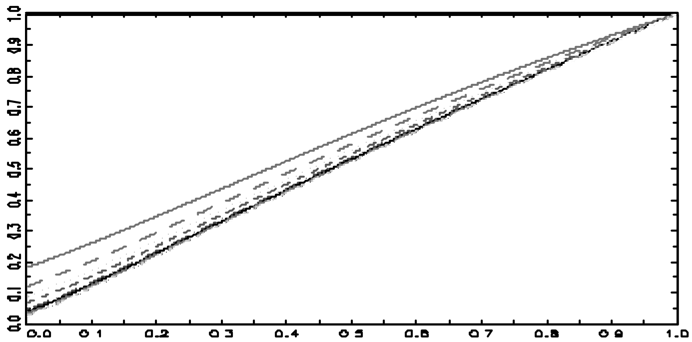

Figure 1. Plots of the function $g_m(\theta)$, for m=4, 5, ...,15.

Then, the last consideration suggests to regress, for $m{=}4{,}5{,}\ldots{,}15$ (say), the exact $g_m(\theta)$ on the approximation $\widehat{g}_m(\theta) = A_m + B_m\theta,$ over some discrete range (θ $=0.001$,...,$0{,}999$). Of course, since we need $g_m(1) = 1$, the coefficients are constrained by $A_m + B_m = 1$.

The results are quite encouraging and briefly summarized by the following points:

i) the simplest and best result we obtained is the following:

$$\widehat{g}_m(\theta) \simeq 1 - B_m(1-\theta);$$

$$B_m = 0.992958 - 0.087813m^{-1} - 2.673147m^{-2}.$$

ii) for any $m{=}4{,}5{,}\ldots$ the goodness of fit of this approximation increases quickly starting from R^2 $=0.996$ when $m = 4$;

iii) in the worst case $(m = 4)$, we obtained:

$$|g_m(\theta) - \widehat{g}_m(\theta)| \leq 0.00489$$

which confirms that our approximation is uniformly accurate;

iv) a graphical evidence of the fitting quality of our proposal is displayed in the previous Figure 2, where the asymptotic variances and their approximations $\widehat{g}_m(\theta)\theta(1-\theta)$, for $m{=}4{,}5{,}\ldots{,}15$, are plotted.

Finally, we observe that we have excluded the size n from the previous consideration. Thus, since the sample size is generally moderate, the absolute value of the error implied by our approximation is substantially smaller in real case studies.

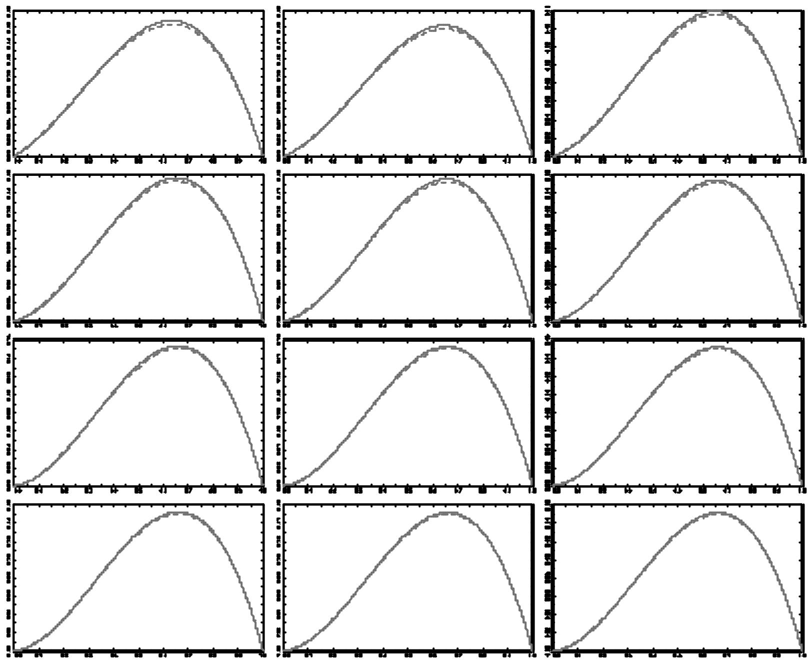

Figure 2. A comparison among asymptotic variances and $their$ linear approximations, for m=4,5,...,15.

5. Concluding remarks

In this paper, we proposed an approximate formulation for the asymptotic variance of the ML estimator of the parameter of the IHG random variable. Specifically, we suggested to operate following a mixed strategy:

i) if $m = 2, 3$, it is convenient to apply the explicit expressions for the asymptotic variance of T_n, as they are shown in the Table 1;

ii) if $m \geq 4$, it is convenient to approximate the asymptotic variance by a linear function with varying coefficients.

The fitting we obtained is very good. Of course, it could be improved by studying the difference $g_m(\theta) - \widehat{g}_m(\theta)$, which is clearly autocorrelated as it often happens when one fits mathematical functions. However, we believe that the increase in the goodness of the approximations will not be balanced by the cost of more complex expressions.

Acknowledgments: This work was possible thanks to the financial support received from MURST and University research funds at the Dipartimento di Scienze Statistiche, University of Naples Federico II. We thank the referees for many useful suggestions that improved a preliminary version of the paper.

Appendix

Maple V procedure for computing the asymptotic variance of the maximum likelihood estimator of the preference parameter in the Hypergeometric random variable, for m fixed.

```
> varexact:=proc(m)
> local prob, adderiv, r, varianza, vvv, uuu, www;
> global th;
> prob[1]:=th;
> adderiv:=1/th;
> for r from 1 to m-1 do
>    prob[r+1]:=prob[r]*(1-th)*(m-r)/(m-r-1+r*th);
>    adderiv:=adderiv+normal(((diff(prob[r+1],th))^2)/(prob[r+1]));
> od;
> varianza:=1/adderiv;
> vvv:=simplify(varianza)/((-th)*(-1+th));
> uuu:=sort(normal(vvv,'expanded'));
> www:=th*(1-th)*uuu;      print(www);
> end:
```

References

D'Elia A. (1999) A Proposal for Ranks Statistical Modelling, *Proceedings of the 14th International Workshop on Statistical Modelling,* (Friedl H., Berghold A., Kauermann G. editors), Graz, Austria, 468-471.

D'Elia A. (2000) Un modello lineare generalizzato per i ranghi: aspetti statistici, problemi computazionali e verifiche empiriche, *Italian Journal of Applied Statistics*, 12, 205-227.

D'Elia A. (2001a) Efficacia didattica e strutture universitarie: la valutazione mediante un approccio modellistico, *Atti Convegno SIS su "Processi e Metodi Statistici di Valutazione"*, Università di Tor Vergata, Roma, 21-24.

D'Elia A. (2001b) A Statistical Model for Studying Preferences, *submitted for publication*.

Guenther W. C. (1975) The Inverse Hypergeometric - A Useful Model, *Statistica Neerlandica*, 29, 129-144.

Marden J. I. (1995) *Analyzing and Modeling Rank Data,* Chapman & Hall, London.

Piccolo D. (2000) Analisi statistica di un modello per le preferenze nel caso di tre alternative, *Quaderni di Statistica,* 2, 241-267

Serfling R. J. (1980) *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York.

Taplin R. H. (1997) The Statistical Analysis of Preferences Data, *Applied Statistics*, 46, 49-512.

Wilks S. (1963) *Mathematical Statistics,* John Wiley & Sons, New York.