How many different HLA genotypes exist in a population ?

F. Fève and J.P. Florens

Toulouse School of Economics [1]

8th June 2010

[1]Address of the authors: TSE, Université Toulouse Capitole, Manufacture des Tabacs - Aile Jean–Jacques Laffont, 31000 Toulouse, France, Tel: +33(0)5.61.12.85.90, Fax: +33(0)5.61.12.86.37, email: feve@cict.fr, florens@cict.fr

Summary

This paper proposes a new method to estimate the number of distinct values of genotypes in a population using a unique sample or several increasing samples. The method rests on a Polya urn scheme which is assumed to be the generator of the genotypes. This model depends on a parameter estimated by the maximum likelihood method or by moments methods. The main conclusion of our analysis is to estimated the number of genotypes in the French population (three locus, low resolution typing) to be around 500 000.

Keywords: Bayesian non parametric models, forecasts of the number of genotypes, Polya urn scheme, random generation of genotypes.

1 Introduction

This statistical study is part of a larger research project (see MADO, POSEIDON, Fève , 2006) [1] that aims at analysing the efficiency of the registries of voluntary donors of hematopoietic stem cells. Voluntary donors are registered with their HLA genotype recorded at a given precision which may vary between registries. The efficiency of a registry may be measured by the probability for potential receivers to find a compatible donor (see Oudshoorn et al., 1997 and 2007 where many references on bone marrow transplantation may be found). Different probability models have been elaborated (see Fève, Florens, 2009; [2] Fève et al., 2007) which highlight the key role played by the

[1] MADO, European contract 2001-2005 QLG7-CT-2001-00065; POSEIDON, European contract, Optimisation, Safety, Experience sharing and quality Implementation for Donation Organisation and Networking in Unrelated Haematopoietic Stem Cell Transplantation in Europe, 2007-2010.

[2] Fève F., Florens J.P. (2009). "Matching Models and Optimal Registry for Voluntary Organ Donation Registries", http://www.idei.fr/doc/by/ffeve/matchingmodels.pdf, University of Toulouse

number of HLA genotypes[3] in the relevant population (see also Single et alii, 2002; Speiser et alii, 1994). This paper presents a method to estimate this number and applies this approach to the French population.

Before the presentation of our approach, let us first give some results coming from previous methods. Our observation consists into a sample of $n = 107\ 925$ individuals where the number of observed different types J_n is equal to 66 164[4]. The size of the French population is approximated by $N = 63\ 000\ 000$. We have applied several methods as they are recalled in Haas et alii (1995)(see this paper for detailed formulae and references). The results are:

$J_{63M}(Chao) = 225\ 213$. (see Chao, 1984)

$J_{63M}(Jacknife) = 117\ 513$. (see Burnham and Overton, 1978 and 1979)

$J_{63M}(ChaoLee) = 1\ 866\ 500$ (see Chao and Lee, 1992)

$J_{63M}(Schlosser) = 378\ 610$. (see Schlosser, 1981)

$J_{63M}(Moment)$ solution of $(66\ 164 = J(1 - exp^{-\frac{107\ 925}{J}})) = 100\ 510$.

Another method is coming from the paper of Charikar et al., 2000 and gives the result: $J_{63M}(Chalikar) = 1\ 255\ 400$. On an other way, Gourraud, 2006 has estimated an upper bound of the number of types in the French population by constructing all possible combinations of previously observed haplotypes. This upper bound is evaluated at 3.4 millions. Up to our knowledge no statistic study of the evolution of the number of HLA genotypes has been previously done. The message coming from this wide range of results is clearly very ambiguous and is a motivation to propose a new approach for this problem.

More precisely, we consider a population of N individuals where each individual has an unknown "type" (namely the HLA genotype (A, B, DR low precision)). We only observe a sample of n individuals, among which J_n different types are observed. The question is how to deduce from this in-

[3]The genotype are here characterized by the observation of three locus (A, B, DR) at a low resolution ("two digits") where only the ordered pairs for the two chromosomes are observed.

[4]France Greffe de Moelle (2004). The authors are grateful to D^r Colette Raffoux and Marie–Lorraine Appert for helpful comments.

formation an estimation of J_N. This paper develops a statistical model of type's generation which leads to a prediction of J_N. If several observations are available the model introduces overidentification constraints which may be tested.

Two approaches may be adopted to analyse this problem. The first one is a finite population approach. It assumes a large population where individuals have types and consider the observed population as a subsample obtained through a random survey. In this case the only stochastic element of the data generating process is the survey mechanism (see references in the survey by Haas et al., 1995, more recently in Chalikar et al., 2000 and the applications of most of the methods are given above). The statistical problem is then to infer a characteristic of the total population (namely the number of different types) from the observed sample. The second one is adopted in this note and consists of the specification of a model for the generation of types which allows ties. The sample is then used to estimate (and to test the model) and this model is used for forecasting the number of types in the whole population. Our model has many interests. First, it is based on a process generating a type (new or previously realized) for each new individual. Second, it is a parsimonious (in terms of number of parameters) specification and the estimation is undertaken by maximum likelihood (which is the most commonly accepted method in statistics). Third the model can be evaluated by different ways described in section 4. We can simulate subsamples and analyse the fit of the model for the total number of types and fit the prediction of increments.

The paper is organized as follows. Section 2 presents the probabilistic foundations of the Polya urn model and section 3 details the statistical procedures for estimation and forecasting. Several results concerning the French population are given in section 4. More possible extensions are joint in the last section.

2 Methodology: A statistical model for type's generation

Our approach is based on the following methodology: we assume a simple one parameter model explaining types'generation and we estimate this parameter using our dataset. The model is then used to forecast the number of types for the whole population. Estimation and forecast are given with confidence intervals. The model describing types' generation is a Polya urn scheme:

1- The first individual has a type generated randomly

2- Assume that i individuals ($i \geq 1$) have been generated and denoted by P_i the empirical distribution of the first i types (P_i (type l) is equal to the frequency of type l in the first i individuals).

Then the $(i+1)$th individual receive a new type with a probability $\frac{n_0}{n_0+i}$ or drawn a type in the distribution P_i with a probability $\frac{i}{n_0+i}$

The unknown parameter is n_0.

This Polya urn scheme has been studied in the literature in many contexts. In particular in the Bayesian non parametric model using a Dirichlet prior distribution (see Ferguson , 1973, 1974; or Antoniak, 1974), the sample is marginally distributed according to this Polya urn model . Despite the sequential description of this scheme the distribution of the types among the individuals is exchangeable: the distribution is invariant by the permutation of the individuals (individuals are anonymous). It has been proved (see Hill et al., 1987) that our model is the unique exchangeable Polya process.

In order to estimate the parameter n_0 we need to recall some properties of the probability to observe J_n different types in a sample of size n generated by our model.

We may first remark that:

$$P(J_n = k) = \frac{n_0}{n_0 + n - 1} P(J_{n-1} = k - 1) + \frac{n - 1}{n_0 + n - 1} P(J_{n-1} = k)$$

$$k = 1, ..., n$$

from which some characteristics of the distribution may be derived.

Moreover following Antoniak (1974), it may be reminded that the probability to observe a sample of size n containing J_n different types is proportional to:

$$n_0^{J_n} \frac{\Gamma(n_0)}{\Gamma(n_0 + n)}$$

The multiplicative factor does not depend on n_0 and is function of the configuration of ties.

Following Rolin, 1993 we may reconstruct J_n by the following way. Let u_i be a random element in $\{0, 1\}$ which takes the value 1 if the individual i has a new type and 0 else.

Then:

$$J_n = \sum_{i=1}^{n} u_i. \tag{2.1}$$

The u_i are independent and

$$P(u_i = 1) = \frac{n_0}{n_0 + i - 1}. \tag{2.2}$$

Then

$$E(J_n) = \sum_{i=1}^{n} \frac{n_0}{n_0 + i - 1} = n_0 \left\{ \psi(n_0 + n) - \psi(n_0) \right\}, \tag{2.3}$$

where $\psi(t)$ is the digamma function $\psi(t) = \frac{d}{dt} ln \Gamma(t)$ which satisfies $\psi(n) = \sum_{i=1}^{n-1} \frac{1}{i} - \gamma$ (where γ is the Euler-Mascheroni constant).

Thank to the approximation $\frac{\psi(t)}{ln\ t)} \to 1$ if t is large, we have :

$$E(J_n) \simeq n_0 [ln\ (n_0 + n) - ln\ n_0] \tag{2.4}$$

Note that this approximation is extremely accurate.

The variance of J_n is obtained by :

$$V(J_n) = \sum_{i=1}^{n} V(u_i) = \sum_{i=1}^{n} \frac{n_0\ (i-1)}{(n_0+i-1)^2}.$$

$$= n_0(\psi(n_0+n) - \psi(n_0)) + n_0^2(\psi'(n_0+n) - \psi'(n_0)) \tag{2.5}$$

where ψ' is the derivative of ψ (also called the trigamma function) which satisfies

$$\psi'(n) = constant - \sum_{j=1}^{n-1} \frac{1}{j^2}. \tag{2.6}$$

The description of the exact distribution of the random variable J_n is complex but asymptotic distributions of J_n and of $J_N - J_n$ given J_n (the incremental number of different types when the sample increases from n to N) are easily obtained from (3.5).

Actually, an application of Lindeberg Feller theorem shows that:

Proposition 1: (Rolin 1993)

i) Asymptotically

$$J_n \sim \mathcal{N}(E(J_n), V(J_n)) \tag{2.7}$$

ii) $J_N - J_n$ is independent in distribution to J_n and, asymptotically

$$J_N|J_n \sim \mathcal{N}(J_n + n_0(\psi(n_0+N) - \psi(n_0+n)), n_0(\psi(n_0+N) - \psi(n_0+n))$$

$$+n_0^2(\psi'(n_0+N) - \psi'(n_0+n))). \tag{2.8}$$

3 Methodology: Estimation and prediction

Let us consider first the case where a single sample of size n is available. We observe in the sample J_n distinct types. The log likelihood derived from (2.1) is then equal to :

$$ln\ell = J_n\ ln\ n_0 + ln\Gamma(n_0) - ln\Gamma(n_0+n) +\ \text{constant} \tag{3.1}$$

and the first order conditions of the likelihood maximization reduces to :

$$n_0 = \frac{J_n}{\psi(n_0+n) - \psi(n_0)} \ , \tag{3.2}$$

which has no solution in a closer form. Let us denote $\hat{n_0}$ the numerical solution of this first order condition.

This estimator $\hat{n_0}$ may also be viewed as the solution of the moment condition

$$E(J_n) = n_0[\psi\ (n_0+n) - \psi\ (n_0)] \ , \tag{3.3}$$

where the expectation is replaced by the observed value. The function $n_0\ [\psi(n_0+n) - \psi(n_0)]$ is strictly increasing. This implies that the solution of the moment condition is unique.

Proposition 2: The estimator $\hat{n}_0$ is almost surely convergent and its asymptotic distribution is:

$$\hat{n_0} \approx N(n_0, \frac{\hat{n}_0}{(\psi(\hat{n}_0+n) - \psi(\hat{n}_0)) + \hat{n}_0(\psi'(\hat{n}_0+n) - \psi'(\hat{n}_0))}).$$

The proof is given in Appendix I.

The prediction of J_N is constructed by the following way:

$$\hat{J_N} = J_n + \hat{n}_0\ [\psi(\hat{n}_0+N) - \psi(\hat{n}_0+n)] \tag{3.4}$$

This prediction is equal to the conditional mean of J_N given J_n where the unknown n_0 is replaced by its estimator. Then this prediction has two sources of errors: firstly J_N is not equal to its mean and secondly n_0 is only estimated and the estimation error has an impact on the forecast error. However we may compute the estimated variance of the forecast error:

$$Var(\hat{J_N} - J_N \mid J_n) = [\hat{n_0}(\psi(\hat{n_0}+N) - \psi(\hat{n_0}+n)) + \hat{n_0}^2(\psi'(\hat{n_0}+N) - \psi'(\hat{n_0}+n))]$$

$$[1 + \frac{\psi(\hat{n_0}+N) - \psi(\hat{n_0}+n) + \hat{n_0}(\psi'(\hat{n_0}+N) - \psi'(\hat{n_0}+n))}{\psi(\hat{n_0}+n) - \psi(\hat{n_0}) + \hat{n_0}(\psi'(\hat{n_0}+n) - \psi'(\hat{n_0}))}] \tag{3.5}$$

The proof is given in Appendix II.

We now assume that we observe increasing samples of individuals at different dates $t_1, ..., t_p$. At t_i the file contain n_i individual and J_{n_i} different types. Using previous properties we have the asymptotic approximation:

$$J_{n_j} - J_{n_{j-1}} \sim N\ [n_0(\psi(n_0+n_j) - \psi(n_0+n_{j-1})),$$

$$n_0(\psi(n_0+n_j) - \psi(n_0+n_{j-1})) + n_0^2\ (\psi'(n_0+n_j) - \psi'(n_0+n_{j-1}))]$$

and all these distributions are independent. We then propose a weighted mean square estimation of n_0 based on the increments of J_n [5],

$$\widehat{n_0} = \arg\min$$

$$\sum_{j=2}^{p} \frac{\left[(J_{n_j} - J_{n_{j-1}}) - n_0(\psi\ (n_0+n_j) - \psi\ (n_0+n_{j-1}))\right]^2}{n_0(\psi\ (n_0+n_j) - \psi\ (n_0+n_{j-1})) + n_0^2(\psi'\ (n_0+n_j) - \psi'\ (n_0+n_{j-1}))} \tag{3.6}$$

An elementary extension of the proof given in Appendix I shows that this estimator verifies asymptotically:

$$\hat{n_0} \approx N[0, \frac{\hat{n_0}}{\sum_{j=2}^{p} \psi(\hat{n_0}+n_j) - \psi(\hat{n_0}+n_{j-1}) + \hat{n_0}(\psi'(\hat{n_0}+n_j) - \psi'(\hat{n_0}+n_{j-1}))}] \tag{3.7}$$

4 Results

We use to apply our approach the France Greffe de Moelle file of volunteer donors observed in 2003 where 107 925 individuals are registered and generates 66 164 different genotypes as there was defined in the introduction. Solving equation (2.10) with $n = 107\ 925$ and $J_n = 66\ 164$, the estimator $\widehat{n_0}$ is then equal to $\widehat{n_0} = 72\ 702$ with a standard deviation equal to 482. A confidence interval at 95 % is then $[71\ 757, 73\ 647]$.

[5] An estimation based on the increments is more robust than an estimation which incorporates the initial conditions. This argument is standard in stochastic processes. Moreover the model will be used to predict the increment $J_N - J_n$ only.

The forecast of the number of types for a population of 63 000 000 individuals is :

$$\widehat{J_n} = 66\ 164 + 72\ 702[\psi\ [63\ 000\ 000 + 72\ 702] - \psi\ [107\ 925 + 72\ 702]] = 491\ 880$$

with a standard deviation equal to 2 705.

Finally the probability that a new donor added to the actual registry does not have a previously observed type is equal to 0.40

Let us consider the FGM file of 107 925 individuals. We have not the history of the file but in order to test our model we propose the following strategy: from the original file we may draw (uniformly without replacement) increasing subfiles. Note that this approach is not feasible if n and J_n only are observable but in our case we observe the complete list of the types of the whole sample. For example we have done the simulations reported in Table 4.1.

Different values of n represent the size of the sample and J_n the observed number of types.

Table 4.1:

n	J_n	$\hat{J}_n(72\ 702)$	$\hat{J}_n(78\ 225)$
107 925	66 164	66 164	66 112
86 427	55 617	56 951	56 512
64 789	44 139	46 325	45 491
43 170	31 529	33 888	32 670
21 551	17 328	18 874	17 328

Using numbers in table 1, an application of procedure (3.6) gives a new estimation equal to 78 225 (with a standard deviation equal to 549). In order to compare the predictions in the sample of two values of n_0 we have computed the predicted values of J_n obtained by:

$$\hat{J}_{n_j} = \hat{J}_{n_{j-1}} + \hat{n_0}(\psi(\hat{n_0} + n_j) - \psi(\hat{n_0} + n_{j-1})) \tag{4.1}$$

with $\hat{J}_{21551}$ equal to the value 17 328. These values are given in the last two columns of Table 4.1. We see that the new estimator of n_0 improves the in the sample predictions.

Outside the sample the new value of $\hat{n_0}$ predicts that a new donor has 0.42 probability to have a new type (if the file has 107 925 individuals) and that the number of types in the French population is about 523 520. If the size of the registry increases to 130 000 donors the model predicts around 76 500 different HLA genotypes (the model predicts 112 180 different types for a registry of 250 000 voluntary donors).

5 Discussion

This paper develops a statistical model for HLA genotypes generation in a population. This model is used for the prediction of the number of types for large registries or for the whole population. A possible direction to extend our model is to consider heterogenous populations. Even if this is eventually a topic for future researches, let us consider the following example. Let us assume that the population of size N is separated into two subpopulations of sizes N_1 and N_2 and the preceding Polya urn scheme applied to each population with two different parameters n_{01} and n_{02}. We assume moreover that the two populations are independently generated. Finally the two populations has no common types.

Then the expected number of types will be:

$$E(J_N) = E(J_{N_1}) + E(J_{N_2}) \tag{5.1}$$

where

$$E(J_{N_1}) = n_{01} \left(\psi(N_1 + n_{01}) - \psi\left(n_{01}\right)\right) \tag{5.2}$$

and

$$E(J_{N_2}) = n_{02} \left(\psi(N_2 + n_{02}) - \psi\left(n_{02}\right)\right) \tag{5.3}$$

using repeated observations (or resampling in a file) the three parameters of the model (n_{01}, n_{02} and α, the proportion of the two subpopulations) may be

estimated using at least three observations.

An other possible extension of our approach will be to relax the specific parametrized form of the probability that individual i has a new type $\frac{n_0}{n_0+i-1}$ and to consider more general specifications. The estimation of this sequence of probabilities will also be made using more information on the structure of ties in the sample. This will also be the topic of a future paper.

Acknowledgments

The authors thank Anne–Cambon Thomsen and Pierre-Antoine Gourraud (INSERM U558, Toulouse) for initially raising the question. They are grateful to Jean–Marie Rolin, Sébastien Van Belleghem, Fabrice Collard and to anonymous referee for helpful discussions and comments

References

Antoniak, Ch.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non parametric problems. *Annals of Statistics*, **2**, 1152-1174.

Burnham, K.P., Overton, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals, *Biometrika*, **65**, 625-633.

Burnham, K.P., Overton, W.S. (1979). Robust estimation of population size when capture probabilities vary among animals, *Ecoloy*, **60**, 927-936.

Chao, A. (1984). Non parametric estimation of the number of classes in a population, Scandinavian *J. Statist., Theory and Applications*, **11**, 265-270.

Chao, A., Lee S. (1992). Estimating the number of classes via sample coverage, *J. Amer. Statis. Assoc.*, **87**, 210-217.

Charikar, M., Chaudhuri, S., Motwani, S., Narasayya, V. (2000). Towards estimation error guarantees for distinct values. *PODS*, 278-279.

Ferguson, T.S. (1973). A Bayesian analysis of some non parametric problems *Annals of Statistics*, **1**, 209-230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615-629.

Fève, F. (2006)." Economie et Statistique du Don de Cellules Souches Hématopoiétiques : Contributions à la Gestion de Registres de Donneurs Volontaires", Thèse de Doctorat de Sciences Economiques, Université Toulouse III.

Fève, F.,Cambon-Thomsen A., Eliaou J.F, Raffoux C., Florens, J.P.(2007).

Economic evaluation of the organization of a registry of haematopoietic stem cell donors. *Revue d'Epidémiologie et de Santé Publique*, **55**, 275-284.

France Greffe De Moelle, Fichier National de donneurs de Cellules Souches Hématopoïétiques, Rapport d'activité (2004). Hôpital Saint-Louis, Paris, France.

Gourraud, P.A. (2006). Note sur la diversité des types HLA à travers le registre de donneurs volontaires de CSH, *private communication,INSERM U558*.

Haas, P.J., Naughton, J.F., Seshadri, S., Stokes, L.,(1995). Sampling-Based Estimation of the Number of Distinct Values of an Attribute, in VLDB'95, Proceedings of 21^{th} International Conference on very Large data Bases, sept 11-15, Zurich, Zwitzerland, 311-322, Morgan Kaufman.

Oudshoorn, M., Cornelissen, J.J., Fibbe, W.E., de Graeff-Meeder, E.R., Lie, J.L., Schreuder, G.M., Sintnicolaas, K., Willemze R., Vossen, J.M., Van Rood, J.J., (1997). Problems and possible solutions in finding an unrelated bone marrow donor. Results of consecutive searches for 240 Dutch patients, *Bone Marrow Transplant*, **20(12)**, 1011-7.

Oudshoorn, M., Van Rood, J.J.(2007). Eleven million donors in bone marrow donors worldwide! Time for reassessment ? *Bone Marrow Transplant*, **1-9**

Hill Bruce, M., Lane, D., Sudderth W. (1987). Exchangeable urn processes, The Annals of Probability, **(15)4**, 1586-1592.

Rolin, J.M. (1993). On the distribution of jumps of the Dirichlet process, Institut de Statistique DP 9302, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Serfling, R. (1980). Approximation Theorems of Mathematical Statistics, Wiley, New York.

Schlosser, A. (1981). On estimation of the size of the dictionary of a long text on the basis of a sample, *Engrg*, Cybernetics, **19**, 97-102.

Single, R.M., Meyer, D., Hollenback, J., Nelson, M.P., Noble J., Erlich, H., Thomson, G., (2002), Haplotype Frequency Estimation in Patient Populations: The effect of Departures from Hardy-Weinberg Proportions and Collasping over a Locus in the HLA Region, *Genetic Epidemiology* **22**, 186-195.

Speiser, D.E, Tiercy, J.M., Rufer, N., Chapuis, B.,Morell, A., Kern M., Gmur, J, Gratwohl, A., Roosnek, E., Jeannet M., (1994). Relation between the resolution of HLA-typing and the chance of finding an unrelated bone marrow donor, *Bone Marrow Transplant*, **13(6)**, 805-9.

APPENDIX I: Asymptotic properties of $\hat{n_0}$

Even if n_0 is derived from maximum likelihood estimation, the general consistency result does not apply because the variables u_i are not identically generated.

i) The a.s. convergence of $\frac{J_n}{ln\, n}$ to n_0 has been prooved by Rolin (1993) among others. Then using $\frac{\psi(x)}{ln\, x} \to 1$ we first have $\frac{J_n}{\psi(n)} \to n_0$ a.s.

ii) We consider now the true value n_0 and an $\varepsilon \geq 0$. Let us define the function:

$$A_n(x) = \frac{x[\psi(x+n) - \psi(x)] - J_n}{\psi(n)}.$$

Using the previous convergence theorem, we get:

$$\exists\, n_1 \; such\; that\; n > n_1 \; A_n(n_0 - \varepsilon) < 0 \; a.s.$$

$$\exists\, n_2 \; such\; that\; n > n_2 \; A_n(n_0 + \varepsilon) > 0 \; a.s.$$

and then

$$\forall\, \varepsilon\; n > \sup(n_1, n_2)\; \exists\, \hat{n_0} \in]n_0 - \varepsilon, n_0 + \varepsilon[\; A_n(\hat{n_0}) = 0$$

and the result is prooved.

The asymptotic normality is deduced from Serfling (1980, chapter 3). The asymptotic variance may be obtained using a Taylor expansion of

$$J_n = \widehat{n_0}[\psi\,(\widehat{n_0} + n) - \psi\,(\widehat{n_0})].$$

Then:

$$V(\hat{n_0}) \simeq [\frac{\partial}{\partial n_0} E(J_n)_{n_0 = \hat{n_0}}]^{-2}\, V(J_n).$$

$$= [\psi\,(\hat{n}_0 + n) - \psi\,\hat{(}n_0) + \hat{n_0}[\psi'\,(\hat{n}_0 + n) - \psi'\,\hat{(}n_0))]^{-2}\, V(J_n)$$

$$= \frac{\hat{n}_0^2}{V(J_n)}.$$

Note that this result is coherent with the usual maximum likelihood variance. Indeed, the score of the model is:

$$\frac{\partial \ln l}{\partial n_0} = \frac{J_n}{n_0} + \psi(n_0) - \psi(n_0 + n),$$

and the Fisher information is equal to

$$V(\frac{\partial \ln l}{\partial n_0}) = E(\frac{\partial \ln l}{\partial n_0})^2 = E(-\frac{\partial^2 \ln l}{\partial^2 n_0}) = \frac{V(J_n)}{n_0^2}.$$

The estimated variance of the estimator $\hat{n_o}$ is then the estimated inverse of the variance of the score.

APPENDIX II: Variance of the forecast error

$$Var(\hat{J_N} - J_N \mid J_n) = [Var(J_N - J_n) - \hat{n_0}(\psi(\hat{n_0} + N) - \psi(\hat{n_0} + n) \mid J_n]$$

$$= Var[(J_N - J_n) \mid J_n] + [Var[\hat{n_0}(\psi(\hat{n_0} + N) - \psi(\hat{n_0} + n) \mid J_n]$$

Indeed the two terms are independent. The second term depends on $\hat{n_0}$ which is function of the sample and the first term depends only on out of sample types. We have first:

$$Var[J_N - J_n \mid J_n] = n_0[\psi(n_0 + N) - \psi(n_0 + n)] + \hat{n_0}^2[\psi'(n_0 + N) - \psi'(n_0)].$$

using a Taylor expansion, we have:

$$\hat{n_0}(\psi(\hat{n_0} + N) - \psi(\hat{n_0} + n) \simeq n_0(\psi((n_0 + N) - \psi(n_0 + n)))$$

$$+[\psi(n_0 + N) - \psi(n_0 + n) + n_0(\psi'(n_0 + N) - \psi'(n_0 + n))](\hat{n_0} - n_0)$$

Then

$$Var\ \hat{n_0}(\psi(\hat{n_0} + N) - \psi(\hat{n_0} + n))$$

is approximated by:

$$[\psi(n_0 + N) - \psi(n_0 + n) + n_0(\psi'(n_0 + N) - \psi'(n_0 + n))]^2$$

$$\times \frac{n_0^2}{n_0[\psi(n_0 + N) - \psi(n_0 + n) + n_0^2(\psi'(n_0 + N) - \psi'(n_0 + n))}$$

The result obtained by replacing n_0 by its estimated value.