Available online at www.sciencedirect.com
ScienceDirect
Journal of Informetrics 2 (2008) 101–105
Journal of INFORMETRICS
An International Journal
www.elsevier.com/locate/joi

Short communication

Some comments on "The estimation of lost multi-copy documents: A new type of informetrics theory" by Egghe and Proot

Quentin L. Burrell

Isle of Man International Business School, The Nunnery, Old Castletown Road, Douglas, Isle of Man IM2 1QB

Received 25 May 2007; received in revised form 11 July 2007; accepted 12 July 2007

Abstract

Egghe and Proot [Egghe, L., & Proot, G. (2007). The estimation of the number of lost multi-copy documents: A new type of informetrics theory. *Journal of Informetrics*] introduce a simple probabilistic model to estimate the number of lost multi-copy documents based on the numbers of retrieved ones. We show that their model in practice can essentially be described by the well-known Poisson approximation to the binomial. This enables us to adopt a traditional maximum likelihood estimation (MLE) approach which allows the construction of (approximate) confidence intervals for the parameters of interest, thereby resolving an open problem left by the authors. We further show that the general estimation problem is a variant of a well-known unseen species problem. This work should be viewed as supplementing that of Egghe and Proot [Egghe, L., & Proot, G. (2007). The estimation of the number of lost multi-copy documents: A new type of informetrics theory. *Journal of Informetrics*]. It turns out that their results are broadly in line with those produced by this rather more robust statistical analysis.

Keywords: Multi-copy documents; Truncated Poisson distribution; Maximum likelihood; Unseen species problem

1. The Egghe–Proot model

The model talks of "documents" (=particular editions of books, journals, music scores, theatre programmes, etc.)—we refer the reader to the original work of Egghe and Proot (2007), hereafter referred to as E&P, for a fuller description. The idea is that originally n copies of the document are "published" but that after a (long) period of time, each individual copy has only a (small) probability of surviving. If we denote this survival probability by θ then X, the number of surviving copies of this document, has (assuming the same survival probability for each copy and independence between copies) a binomial distribution.

Remark. E&P say that their model assumes that "... each copy has an equal chance to be lost (in other words that copies are lost independently)". Technically, this is incorrect. For instance in rolling a die, a one-spot and a six-spot have the same probability but they are not independent. This does not affect anything in the subsequent analysis in E&P, or in what follows.

In fairly standard notation we would write $X \sim \text{Bin}(n, \theta)$. The E&P notation is slightly different since they focus on p, the probability that a copy is lost. Thus $\theta = 1 - p$. They also write a instead of n. Although acknowledging that

E-mail address: q.burrell@ibs.ac.im.

1751-1577/$ – see front matter
doi:10.1016/j.joi.2007.07.002

n varies greatly from document to document and given that there is no reason to suppose that p also is constant E&P implicitly assume that both are constant. In fact, this is not too crucial. (E&P give some calculations showing that, at least in some respects, the role of n is not too important.) In what follows we assume that, for each document, n is large and θ is small. More precisely, we make use of the following well-known result.

Proposition (Poisson approximation to binomial). *If* $X_n \sim \mathrm{Bin}(n, \theta_n)$ *then if* $n \to \infty$, $\theta_n \to 0$ *in such a way that* $n\theta_n \to \lambda$, *a constant*, *then*

$$P(X_n = r) \to \mathrm{e}^{-\lambda}\frac{\lambda^r}{r!} \quad \text{for } r = 0, 1, 2, \ldots$$

This last distribution is, of course, simply a Poisson distribution with mean λ.

More informally, and sufficient for what follows, this result can be stated as: A binomial distribution $\mathrm{Bin}(n, \theta)$ with n large and θ small can be approximated by a Poisson distribution with mean $\lambda = n\theta$. A form of this result can be found in almost any introductory text on probability, we mention just Chung's classic (Chung & AitSahlia, 2003, p. 205ff) and Feller (1968, p. 153ff).

Remark. The proof of the Proposition in E&P is essentially a special case of that behind the above general Poisson approximation.

In the current context, n is known in practice to be large. We need to make the assumption, implicit in E&P, that $\lambda = n\theta$ is similar for each document.

Applying this model in the E&P context, we actually need to consider the zero-truncated form of the Poisson distribution since documents for which $X = 0$ are not observed, since by definition every copy has been lost. Hence the random variable that is observed has probability mass function given by

$$P(X = r|\lambda) = (\mathrm{e}^{\lambda} - 1)^{-1}\frac{\lambda^r}{r!} \quad \text{for } r = 1, 2, \ldots. \tag{1}$$

Given observed values $x_1, x_2, \ldots x_m$ the log-likelihood function is easily found to be

$$L(\lambda|\underline{x}) = -m\ \ln(\mathrm{e}^{\lambda} - 1) + \ln\lambda\sum_{j=1}^{m} x_j + \text{constant} \tag{2}$$

Here m is the number of extant documents, and $x_1, x_2, \ldots x_m$ are the numbers of copies of each.

Differentiating (2) with respect to λ and setting the derivative equal to zero shows that the maximum likelihood estimate (MLE) of λ, i.e. the solution of $\partial L/\partial\lambda = 0$, satisfies $g(\hat{\lambda}) = 0$ where

$$g(\lambda) = \frac{\lambda}{1 - \mathrm{e}^{-\lambda}} - \bar{x} \tag{3}$$

and $\bar{x} = \left(\sum_{j=1}^{m} x_j\right)/m$ is the mean of the observed x-values.

(For the theory of maximum likelihood estimation we refer the reader to, for instance, Bain & Engelhardt, 1992, Chapter 9; DeGroot, 1986, Chapter 6, but almost any intermediate text on mathematical statistics will do.)

Note that from the E&P data, the number of documents is $m = 804$, the total number of copies is $\sum_{j=1}^{m} x_j = 907$, and hence $\bar{x} = \left(\sum_{j=1}^{m} x_j\right)/m = 907/804 = 1.1281$.

The important point is that Eq. (3) for $g(\lambda)$ cannot be solved by analytic means so we have to resort to numerical methods. Traditionally, this would have been done by using a technique such as the Newton-Raphson approach, see, e.g. Press, Flannery, Tuekolsky, and Vetterling (1986, p. 254). Nowadays most computer packages, and even pocket calculators, have an "equation solver" facility which makes the problem trivial. Whatever method is adopted, we find a maximum likelihood estimate $\hat{\lambda} = 0.2461$ (to 4 d.p.).

Using this MLE to generate expected values of the observed frequencies leads to the values in Table 1.

Clearly the agreement is very close. Since the parameter has been estimated by MLE it is legitimate to use the chi-squared approach to test goodness of fit. Unfortunately, using the usual rule that expected frequencies should be at

Table 1
Observed and expected frequencies

No. of surviving copies	Observed no. of documents	Expected no. of documents
1	714	709.11
2	82	87.27
3	4	7.16
4	3	0.44
5	1	0.02
≥6	0	0.00

least 5, requires further groupings so that there are in fact only three categories and just one degree of freedom (d.f.) available for the goodness of fit test. The calculated chi-squared value is 0.370 which, with 1 d.f., has a p-value of 0.89. Although this suggests a very good fit, the small number of categories and minimal number of d.f. should not be taken as a demonstration of the correctness of the model, merely that the data are consistent with the assumed model.

The invariance theorem for MLEs then gives straight away that the MLE of the proportion of documents which have been lost is

$$\hat{P}_0 = \mathrm{e}^{-\hat{\lambda}} = \mathrm{e}^{-0.2461} = 0.7818$$

Note therefore that this estimate is slightly less than, but comparable to, the value of 0.7946 reported by E&P using their ad hoc estimation approach. In fact we can go further.

Theorem 1. *If $\hat{\lambda}$ is the MLE of λ, then for n large, $\sqrt{n}(\hat{\lambda} - \lambda) \to N(0, i(\lambda)^{-1})$ where the convergence is in distribution and $i(\lambda)$ denotes the Fisher information in a single observation on X.*

Proof. This is a standard result from the asymptotic theory for MLEs, see e.g. Bain and Engelhardt (1992, p. 316) and DeGroot (1986, p. 428).

Of more importance in practical applications is the following:

Corollary 1. For large n, an approximate 95% confidence interval for λ is given by

$$\hat{\lambda} - \frac{1.96}{\sqrt{ni(\hat{\lambda})}} \leq \lambda \leq \hat{\lambda} + \frac{1.96}{\sqrt{ni(\hat{\lambda})}}. \tag{4}$$

Proof. This again is a fairly standard result based on Theorem 1 and the 95% limits of the standard Normal distribution. (Note that in the variance term we are obliged to replace λ by its MLE, hence introducing a further approximation.)

Theorem 1 and Corollary 1 are general results (subject only to certain mild regularity conditions). In order to implement them in the current context we need an expression for the Fisher information in the case of the truncated Poisson distribution. As this does not seem to appear in the standard literature, we include it here.

Proposition. If X has a truncated Poisson distribution as in (1), the Fisher information in a single observation is given by

$$i(\lambda) = \frac{1 - (1+\lambda)\mathrm{e}^{-\lambda}}{\lambda(1 - \mathrm{e}^{-\lambda})^2} \tag{5}$$

Proof. By definition, $i(\lambda) = E[(\partial f(X|\lambda)/\partial\lambda)^2]$ where $f(r|\lambda) = P(X = r|\lambda)$ as in (1).

Thus $\ln f(X|\lambda) = -\ln(e^{\lambda} - 1) + X \ln\lambda - \ln(\mathrm{X!})$

$$\text{Then} \quad \frac{\partial \ln f(X|\lambda)}{\partial\lambda} = \frac{X}{\lambda} - \frac{\mathrm{e}^{\lambda}}{\mathrm{e}^{\lambda} - 1}$$

Squaring both sides and taking expectations leads to

$$E\left[\left(\frac{\partial \ln f(X|\lambda)}{\partial\lambda}\right)^2\right] = \frac{1}{\lambda^2}E[X^2] - \frac{2\mathrm{e}^{\lambda}}{\lambda(\mathrm{e}^{\lambda} - 1)}E[X] + \left(\frac{\mathrm{e}^{\lambda}}{\mathrm{e}^{\lambda} - 1}\right)^2 \tag{6}$$

Now for a standard Poisson distributed random variable Y with parameter λ, we have $E[Y]=\lambda$, $E[Y^2]=\lambda+\lambda^2$ so for the zero-truncated version the same two moments are given by these expressions each divided by $(1-e^{-\lambda})$. Upon substitution into (6) we find, after a little algebra

$$E\left[\left(\frac{\partial \ln f(X|\lambda)}{\partial\lambda}\right)^2\right]=\frac{1+\lambda}{\lambda(1-e^{-\lambda})}-\frac{1}{(1-e^{-\lambda})^2}=\frac{1-(1+\lambda)e^{-\lambda}}{\lambda(1-e^{-\lambda})^2}$$ as required.

2. Application to the Egghe and Proot data

(i) For the published data, we have $n=804=N_f$ (in the E&P notation) and, as reported earlier, $\hat{\lambda}=0.2461$. Substituting into (5) gives $i(\hat{\lambda})=2.1978$ so that $\sqrt{ni(\hat{\lambda})}=42.0357$. Finally, substitution into (4) gives the 95% approximate confidence interval for λ as $0.1995<\lambda<0.2928$.

(ii) A further application of the invariance theorem for MLEs allows us to use these lower and upper limits for λ to construct a 95% confidence interval for the proportion of lost documents as $0.746<P_0<0.819$. This includes the value 0.7946 calculated by E&P.

(iii) If initially N documents were published, then the (expected) number of still existing documents is $N(1-P_0)$, whereas we know the observed number of survivors to be $m=804$. Thus our best estimate for N is given by

$$\hat{N}=\frac{m}{1-\hat{P}_0}=\frac{804}{1-0.7818}=\frac{804}{0.2182}=3684.7\approx 3685$$

Making use of the confidence limits for P_0 reported above similarly leads to an approximate 95% confidence interval for N as $3165<N<4442$. Similarly, our point estimate of the number of lost documents is $3685-804=2881$, with an approximate confidence interval (2361, 3638). The calculated values given by E&P are 3903 and 3099, respectively, in both cases lying within our calculated confidence intervals.

3. The "unseen species" viewpoint

Although the E&P problem is to estimate those items which have been lost, we claim that mathematically it is equivalent to the so-called unseen species problem where we are seeking to estimate the number of species which have not yet been found. To see this, imagine the Egghe and Proot context but with time reversed. In other words we have an extant collection of documents then as we look back over time further examples are recovered or, rather, cease to be lost. In both cases we are looking at a situation in which there is a non-productive/invisible/lost class of, in all cases (currently) zero-producers which may be uncovered as the search is extended. In ecology the unseen species problem dates back at least to Fisher, Corbet, and Williams (1943), see also Engen (1978), in which context the search extension may be in the sense of widening the geographical area being surveyed. In bibliometrics Kendall (1960), in his discussion of Bradford's work on journal productivity, posed the problem "there is also a non-observed class of journals which have not carried a relevant article in the period examined but may do so at any moment in the future. One would like to be able to estimate the size of this potentially contributory class". Thus in this context the search extension corresponds to extending the time period. Even within informetrics there are others, see Burrell (2003) for a discussion. In our current analysis we are adopting a model based parametric approach, but one could also use a non-parametric approach. For instance, Kendall's problem was addressed by Brookes (1975) who essentially, but independently, demonstrated a special case of the so-called Good & Toulmin formula (see Good, 1953; Good & Toulmin, 1956; Burrell, 1988). Efron and Thisted (1976) extended this empirical approach which was further developed and applied within bibliometrics by Burrell (1989, 1990).

4. Concluding remarks

Although it is true that Egghe and Proot (2007) have proposed a novel informetric application, we have shown that classical statistical analysis yields a more complete solution to the problem than they achieved. We trust that future

studies in this area will adopt this more robust approach. Furthermore, we have shown that the problem can be framed in terms which have already been extensively researched in informetrics.

References

Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics* (2nd ed.). Belmont: Duxbury Press.

Brookes, B. C. (1975). A sampling theorem for finite discrete distributions. *Journal of Documentation*, *31*, 26–35.

Burrell, Q. L. (1988). A simple empirical method for predicting library circulations. *Journal of Documentation*, *44*, 302–314.

Burrell, Q. L. (1989). On the growth of bibliographies with time: an exercise in bibliometric prediction. *Journal of Documentation*, *45*, 302–317.

Burrell, Q. L. (1990). Empirical prediction of library circulations based on negative binomial processes. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90: Selection of papers submitted for the Second International Conference on Bibliometrics, Scientometrics and Informetrics* (pp. 57–64). Amsterdam: Elsevier.

Burrell, Q. L. (2003). The sample size dependency of statistical measures in informetrics? Some comments. *Journal of the American Society for Information Science and Technology*, *54*(11), 1076–1077.

Chung, K. L., & AitSahlia, F. (2003). *Elementary probability* (4th ed.). New York: Springer Verlag.

DeGroot, M. H. (1986). *Probability and statistics* (2nd ed.). Reading, MA: Addison-Wesley.

Efron, B., & Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, *63*, 435–477.

Egghe, L., & Proot, G. (2007). The estimation of the number of lost multi-copy documents: A new type of informetrics theory. *Journal of Informetrics*, *1*, 257–268.

Engen, S. (1978). *Stochastic abundance models*. London: Chapman and Hall.

Feller, W. (1968). (3rd ed.). *An introduction to probability theory and its applications* New York: Wiley.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample from an animal population. *Journal of Animal Ecology*, *12*, 42–58.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264.

Good, I. J., & Toulmin, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, *43*, 45–63.

Kendall, M. G. (1960). The bibliography of operational research. *Operational Research Quarterly*, *11*, 31–36.

Press, W. H., Flannery, B. P., Tuekolsky, S. A., & Vetterling, W. T. (1986). *Numerical recipes: The art of scientific computing*. Cambridge: Cambridge University Press.