Biometrika (2012), **99**, 1, *pp.* 1–5

Printed in Great Britain

Advance Access publication on 31 July 2012

General type-token distribution

BY S. HIDAKA

School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

shhidaka@jaist.ac.jp

SUMMARY

We consider the problem of estimating the number of types in a corpus using the number of types observed in a sample of tokens from that corpus. We derive exact and asymptotic distributions for the number of observed types, conditioned upon the number of tokens and the latent type distribution. We use the asymptotic distributions to derive an estimator of the latent number of types and we validate this estimator numerically.

Some key words: Poisson-binomial distribution; Species sampling; Type-token ratio

1. INTRODUCTION

Estimation of the number of unique types or distinct species in a group is required in many fields. A linguist may study the vocabulary size of an author (Jarvis, 2002; Malvern & Richards, 2002, 2012; McCarthy & Jarvis, 2007, 2010; Tweedie & Baayen, 1998; Zipf, 1949). An ecologist may estimate species abundance in a region (Chao, 1984, 1992; Good, 1953; Huillet & Paroissin, 2009). In such situations, the potential types are unknown a priori. We derive the asymptotic distribution of the number of observed types in a sample, which may be used to estimate this number of latent types.

Consider a sequence of independent and identically distributed random variables $X_1, \ldots, X_M$, where each of these is an integer $X_i \in \bar{N} \equiv \{1, \ldots, N\}$ drawn with probability $p_k \equiv \mathrm{pr}(X_i = k)$. We associate several quantities with this sequence: the number $f_{k,M}$ of integers which appear exactly k times in the sequence, the number of tokens $M = \sum_{k=1}^{M} k f_{k,M}$, the number of distinct types $K = \sum_{k=1}^{M} f_{k,M}$ observed in the sample of M tokens, the latent number of types $N = \sum_{k=0}^{M} f_{k,M}$, and the word distribution

$$\mathcal{N} \equiv (p_1, \ldots, p_N), \quad p_i > 0, \quad \sum_{i=1}^{N} p_i = 1.$$

Past studies have taken two distinct approaches (Bunge & Fitzpatrick, 1993). The first approach utilizes the observation that, if prior samples reflect the probability that a subsequent one is of a given type, then this implies that the frequencies $f_{k,M}$ satisfy certain relations (Good, 1953; Goodman, 1949; Ewens, 1972; Pitman, 1995). Typically, the number of tokens M is fixed. The second approach is to fit a curve to pairs (K, M) of the number of types K observed in M tokens (Brainerd, 1982; Chao, 1992; Herdan, 1960; Malvern & Richards, 2002; McCarthy & Jarvis, 2010; Tweedie & Baayen, 1998). The pairs (K, M) used in this approach are derived from an empirical data set, and the number of latent types is a parameter in the fitting model.

Our work builds upon the second approach by deriving the probability distribution of the pairs (K, M). This distribution is implicit in Brainerd (1982), who derived its first- and second-order moments.

2. TYPE-TOKEN DISTRIBUTION

2·1. *Exact probability distribution*

Suppose that M tokens are drawn from a corpus with word distribution $\mathcal{N} = (p_1, \ldots, p_N)$. For a subset $s \subseteq \bar{N}$, the probability that a sampled word has a type in s is $\mathrm{pr}(s) = \sum_{i \in s} p_i$, with $\mathrm{pr}(\emptyset) = 0$. By the inclusion-exclusion principle (Allenby & Slomson, 2011), the probability that the types observed in M tokens are precisely those in s is

$$\mathrm{pr}(s \mid M, \mathcal{N}) = \sum_{k=0}^{|s|-1} (-1)^k \sum_{\{t \subseteq s : |t| = k\}} \mathrm{pr}\left(s \setminus t\right)^M, \tag{1}$$

where $s \setminus t$ denotes these elements of s not in t. Equation (1) also follows from the Chapman–Kolmogorov equations (Brainerd, 1972)

$$\mathrm{pr}(s \mid M, \mathcal{N}) = \mathrm{pr}(s \mid M-1, \mathcal{N})\mathrm{pr}(s) + \sum_{i \in s} \mathrm{pr}(s \setminus \{i\} \mid M-1, \mathcal{N}) p_i.$$

For $K = 1, \ldots, N$, the probability that exactly K types occur in a sample of M tokens is

$$\mathrm{pr}(K \mid M, \mathcal{N}) = \sum_{\{u \subseteq \bar{N} : |u| = K\}} \mathrm{pr}(u \mid M, \mathcal{N}). \tag{2}$$

For each set $s \subseteq \{1, \ldots, N\}$ with $|s| = k \leq K$, upon making the substitutions specified by (1), the expression $\mathrm{pr}(s)^M$ occurs $(N-k)!/\{(N-K)!(K-k)!\}$ times in (2). Therefore,

$$\mathrm{pr}(K \mid M, \mathcal{N}) = \sum_{k=1}^{K} (-1)^{K-k} \binom{N-k}{N-K} \sum_{\{s \subseteq \bar{N} : |s| = k\}} \mathrm{pr}(s)^M. \tag{3}$$

We call $\mathrm{pr}(K \mid M, \mathcal{N})$ the type-token distribution.

2·2. *Moment-generating function*

LEMMA 1. *The moment-generating function of the type-token distribution (3) is*

$$\mathcal{M}_{P,M}(t) = \sum_{k=1}^{N} \sum_{\{s \subseteq \bar{N} : |s| = k\}} \mathrm{pr}(s)^M e^{kt} (1 - e^t)^{N-k}.$$

Proof. By (3),

$$\begin{aligned}
\mathcal{M}_{P,M}(t) &\equiv \sum_{K=1}^{N} e^{Kt} \mathrm{pr}(K \mid M, \mathcal{N}) \\
&= \sum_{k=1}^{N} \sum_{\{s \subseteq \bar{N} : |s| = k\}} \mathrm{pr}(s)^M \sum_{K=k}^{N} (-1)^{K-k} \binom{N-k}{N-K} e^{K't} \\
&= \sum_{k=1}^{N} \sum_{\{s \subseteq \bar{N} : |s| = k\}} \mathrm{pr}(s)^M e^{kt} \sum_{K'=0}^{N'} (-1)^{K'} \binom{N'}{N'-K'} e^{K't}.
\end{aligned}$$

This yields Lemma 1 as, by the binomial theorem,

$$\sum_{K'=0}^{N'} (-e^t)^{K'} \binom{N'}{N'-K'} = (1-e^t)^{N'}.$$

2·3. *Asymptotic distribution*

Exact calculation of the type-token distribution (3) is intractable when sampling from corpora with large numbers of types. It is useful to have a reasonable approximation to this distribution which can be computed efficiently. We show that Poisson-binomial distributions (Chen & Liu, 1997; Shah, 1994; Wang, 1993) provide such approximations.

Poisson-binomial distributions can be computed efficiently (Fernandez & Williams, 2010; Shah, 1994). Le Cam's (1960) theorem, which provides a Poisson approximation to Poisson-binomial distributions, can make computation even more efficient at the cost of accuracy.

THEOREM 1. *For each positive integer M and $i = 1, \ldots, N$, write $s_i = \{1, \ldots, N\} \setminus \{i\}$ and $q_{M,i} = 1 - \mathrm{pr}\,(s_i)^M$. Consider the family of Poisson-binomial distributions*

$$Q(K \mid M, \mathcal{N}) = \sum_{\{s \subseteq \bar{N} : |s| = K\}} \prod_{i \in s} q_{M,i} \prod_{j \in \bar{N} \setminus s} (1 - q_{M,j}). \tag{4}$$

For a fixed probability distribution $\mathcal{N}$,

$$\lim_{M \to \infty} \max_{K=1,\ldots,N} |\mathrm{pr}(K \mid M, \mathcal{N}) - Q(K \mid M, \mathcal{N})| = 0.$$

Proof. The moment-generating function of $Q(K \mid M, \mathcal{N})$ is (Wang, 1993)

$$\mathcal{M}_{Q,M}(t) = \prod_{i=1}^{N} \left\{ e^t + (1-e^t)\mathrm{pr}(s_i)^M \right\}.$$

By Lemma 1, $\mathcal{M}_{P,M}(t) = \sum_{k=0}^{N} e^{t(N-k)} (1-e^t)^k \sum_{\{s \subseteq \bar{N} : |s| = k\}} \mathrm{pr}(\bar{N} \setminus s)^M$. Writing $\Delta_{s,M} \equiv \mathrm{pr}(\bar{N} \setminus s)^M - \prod_{i \in s} \mathrm{pr}(s_i)^M$,

$$\mathcal{M}_{P,M}(t) - \mathcal{M}_{Q,M}(t) = \sum_{k=2}^{N} e^{t(N-k)} \left(1 - e^t\right)^k \sum_{\{s \subseteq \bar{N} : |s| = k\}} \Delta_s^M,$$

Since $-\prod_{i \in s} \mathrm{pr}(s_i)^M \leq \Delta_{s,M} \leq 0$, and since the number of subsets s is independent of M. $\lim_{M \to \infty} \mathcal{M}_{P,M}(t) - \mathcal{M}_{Q,M}(t) = 0$. As the probability distributions $\mathrm{pr}(K \mid M, \mathcal{N})$ and $Q(K \mid M, \mathcal{N})$ have the same support, this proves the theorem. □

3. ESTIMATION

Given n independent pairs of numbers of types and tokens (K_i, M_i) $(i = 1, \ldots, n)$, the likelihood of the parameter $\mathcal{N} = (p_1, \ldots, p_N)$ is

$$L(\mathcal{N}) = \prod_{i=1}^{n} Q(K_i \mid M_i, \mathcal{N}), \tag{5}$$

where Q is the Poisson-binomial distribution of (4). We obtain an estimator $\bar{N}$ for the number of latent types by maximizing the likelihood $L(\mathcal{N})$.

Suppose that infinitely many tokens are sampled from the distribution $\nu = (\pi_1, \ldots, \pi_{N_\nu})$ and that, for each positive integer M, there are $K(M)$ types observed amongst the first M tokens. For $\mathcal{N} = (p_1, \ldots, p_N)$ and $i = 1, \ldots, N$, by the law of large numbers, the proportion of the tokens of i amongst the first M tokens tends to p_i as $M \to \infty$. Therefore, as $M \to \infty$, $L_M(\mathcal{N})$ tends to 1 if $N = N_\nu$ and to 0 otherwise. This proves that the maximum likelihood estimator consistently estimates the number of types.

As a consequence, the optimization of the likelihood function (5) may be restricted to any family of distributions in which, for any positive integer N, there is at least one distribution with N types. When analyzing data from a natural corpus, one may restrict the maximization to the family of Zipf distributions. This is justified by the prevalence of these distributions in such data (Kornai, 2002; Zipf, 1949).

In our analysis, we compared this estimator to the Good–Turing estimator (Good, 1953; Gale & Sampson, 1995) and the Horvitz–Thompson (1952) estimator. We observed that the Poisson-binomial estimator was less biased than the other estimators. See the Supplementary Material.

ACKNOWLEDGEMENTS

The author is grateful to Takuma Torii, Akira Masumi and Dr. Neeraj Kashyap for their helpful discussions and comments on early versions of the manuscript. This work was supported by the Artificial Intelligence Research Promotion Foundation, JSPS KAKENHI Grant-in-Aid for Scientific Research B and the Grant-in-Aid for Challenging Exploratory Research.

4. SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online describes practical use of the Poisson-binomial estimator and compares it to two other commonly used type estimators.

REFERENCES

ALLENBY, R. B. & SLOMSON, A. (2011). *How to Count: An Introduction to Combinatorics*. Florida, USA: CRC Press.

BRAINERD, B. (1972). On the relation between types and tokens in literary text. *Journal of Applied Probability* **9**, pp. 507–518.

BRAINERD, B. (1982). On the relation between the type-token and species-area problems. *Journal of Applied Probability* **19**, pp. 785–793.

BUNGE, J. & FITZPATRICK, M. (1993). Estimating the number of species: A review. *Journal of the American Statistical Association* **88**, 364–373.

CHAO, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11**, pp. 265–270.

CHAO, M.-T. (1992). From Animal Trapping to Type-Token. *Statistica Sinica* **2**, 189–201.

CHEN, S. X. & LIU, J. S. (1997). Statistical applications of the Poisson-binomial and conditional Bernoulli distributions. *Statistica Sinica* **7**, 875–892.

EWENS, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3**, 87–112.

FERNANDEZ, M. & WILLIAMS, S. (2010). Closed-form expression for the Poisson-binomial probability density function. *IEEE Transactions on Aerospace Electronic Systems* **46**, 803–817.

GALE, W. A. & SAMPSON, G. (1995). Good–Turing frequency estimation without tears. *Journal of Quantitative Linguistics* **2**, 217–237.

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

GOODMAN, L. A. (1949). On the estimation of the number of classes in a population. *Annals of Mathematical Statistics* **20**, 572–579.

HERDAN, G. (1960). *Type-Token Mathematics: A Textbook of Mathematical Linguistics*. Hague, Netherlands: Mouton & Co.

HORVITZ, D. G. & THOMPSON, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

HUILLET, T. & PAROISSIN, C. (2009). Sampling from Dirichlet partitions: estimating the number of species. *Environmetrics* **20**, 853–876.

JARVIS, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* **19**, 57–84.

KORNAI, A. (2002). How many words are there? *Glottometrics* **4**, 2002.

LE CAM, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics* **10**, 1181–1197.

MALVERN, D. & RICHARDS, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* **19**, 85–104.

MALVERN, D. & RICHARDS, B. (2012). *Measures of Lexical Richness*. Oxford, UK: Blackwell Publishing Ltd.

MCCARTHY, P. & JARVIS, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* **42**, 381–392. 10.3758/BRM.42.2.381.

MCCARTHY, P. M. & JARVIS, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing* **24**, 459–488.

PITMAN, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102**, 145–158.

SHAH, B. K. (1994). On the distribution of the sum of independent integer valued random variables. *American Statistician* **27**, 123–124.

TWEEDIE, F. & BAAYEN, R. (1998). How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities* **32**, 323–352. 10.1023/A:1001749303137.

WANG, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica* **3**, 295–312.

ZIPF, G. K. (1949). *Human Behavior and the Principle of Least-Effort.* Oxford, UK: Addison-Wesley Press.

[*Received April* 2012. *Revised September* 2012]

Biometrika (2012), **99**, 1, *pp.* 1–3

Printed in Great Britain

Advance Access publication on 31 July 2012

Supplementary material to General type-token distribution

BY S. HIDAKA

School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, Japan

shhidaka@jaist.ac.jp

1. PRACTICAL TYPE ESTIMATION

Consider the problem of estimating the size of Lewis Carroll's vocabulary when he wrote "Alice's Adventures in Wonderland". The number of tokens we have from this corpus is limited to the 24,168 words which appear in the novel, and there is little hope of adding to this sample. In practice, one often has to deal with such limitations on sampling. The conventional method of dealing with this problem is to generate multiple samples from the same data set for use in estimation. For example, in the case of "Alice's Adventures in Wonderland", one would sample data sets $D_1, \ldots, D_n$ from the text, with each data set D_i consisting of M_i tokens. These data sets would not be independent as required by most estimators. It has been observed empirically, however, that the use of such data sets increases the accuracy of estimators when additional sampling is difficult.

There are many schemes one could use to generate the data sets $D_1, \ldots, D_n$. Our objective is to compare type estimators. We therefore adopt the strategy of sampling successive tokens: if the original sample consists of M tokens, we decide upon a target number $n \leq M$ of data sets and take for D_i the first $[M/n] \times i$ tokens, where $[x]$ denotes the greatest integer less than or equal to x. We take care to choose n so that the overlap between data sets does not impede estimation.

The Good–Turing (Good, 1953; Gale & Sampson, 1995) and Horvitz–Thompson estimators (Horvitz & Thompson, 1952) are most commonly used in practice. These estimators make use of the frequency $f_{k,M}$ defined in the introduction to our article. We denote by $\hat{N}_{\mathrm{GT}}$ the Good–Turing estimate of the latent number of types, and by $\hat{N}_{\mathrm{HT}}$ the Horvitz–Thompson estimate. These are

$$\hat{N}_{\mathrm{GT}} \equiv f_{1,M} + K, \quad \hat{N}_{\mathrm{HT}} \equiv \sum_{k=1}^{\infty} \frac{f_{k,M}}{1 - \left(1 - \frac{k}{M}\right)^M} .$$

We compared these estimators to the maximum likelihood estimator for the likelihood function $L(\mathcal{N})$ of (5). In maximizing this likelihood, we assumed that $\mathcal{N}$ was a Zipf distribution on the set $\bar{N}$ for some positive integer N. This constraint makes the optimization tractable and, as noted in the main article, it does not affect the consistency of the estimator.

The Zipf distributions form a two-parameter family. Each distribution $\mathrm{pr}(k) \propto k^{-a}, k = 1, \ldots, N$, is specified by its exponent a and the size N of its support. For such a distribution $\mathcal{N}$, we write $L(\mathcal{N}) = L(a, N)$. We obtained maximum likelihood estimates $\hat{a}$ and $\hat{N}$ of these parameters, using $\hat{N}_{\mathrm{PB}} \equiv \hat{N}$ as the Poisson-binomial estimate of the latent number of types.

2. NUMERICAL EXPERIMENTS

We assessed these estimators using two classes of numerically generated data sets D. The data sets in the first class consisted of $M = 1000, 1500, 2000$ tokens sampled from a corpus of

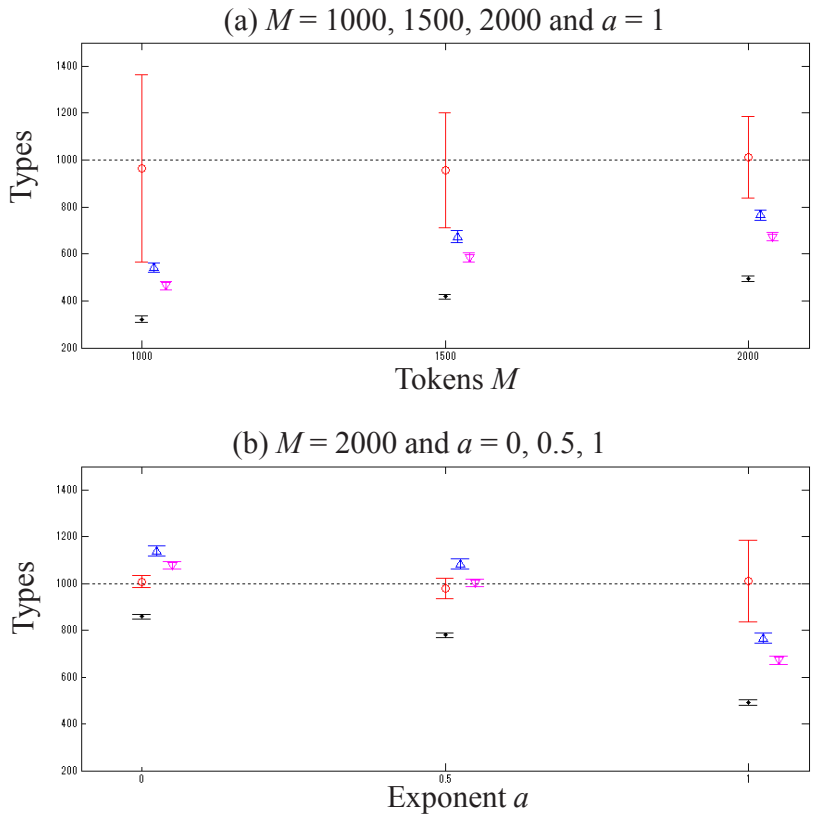

Fig. 1. Results of simulation study. Panels (a) and (b) show the average estimates produced by the three estimators for each set of parameters. Red circles represent average Poisson-binomial estimates $\hat{N}_{\text{PB}}$; blue, upward-pointing triangles represent average Good–Turing estimates $\hat{N}_{\text{GT}}$; and purple, downward-pointing triangles represent average Horvitz–Thompson estimates $\hat{N}_{\text{HT}}$. Black dots represent average numbers of observed types for each set of parameters. The dashed lines reflect the true number of types, $N = 1000$. The vertical lines around each marker indicate the standard deviation of the estimates on the hundred data sets corresponding to that set of parameters.

$N = 1000$ types according to the Zipf distribution $p_k \propto k^{-a}, a = 1$. The data sets in the second class consisted of $M = 2000$ tokens sampled from a corpus of $N = 1000$ types according to the Zipf distributions $p_k \propto k^{-a}, a = 0, 0.5, 1$.

For each sample D, we generated data sets $D_1, \ldots, D_{M/50}$ by successively sampling tokens as described above. We obtained the estimate $\hat{N}_{\text{PB}}$ for the corpus corresponding to D by maximizing the product of the likelihood functions corresponding to each of the data sets D_i.

In the family of Zipf distributions, the exponent a is a smooth parameter. Consequently, it is easy to maximize the conditional likelihood $L(a \mid N)$. However, as N is a discrete parameter, and this does not translate to easy maximization of $L(a, N)$. In these experiments, we assumed that $N \leq 2000$ and performed the optimization on N by brute force.

For each choice of parameters M and a, we independently generated one hundred data sets D which we used to estimate the size of the underlying corpus. The result of this analysis are shown in Fig. 1. These results indicate that the Good–Turing and Horvitz–Thompson estimators are more biased for such data than the Poisson-binomial estimator. Moreover, their biases increase with the exponent of the Zipf distributions whereas the mean Poisson-binomial estimates consistently reflect the true number of types.

3. Alice in Wonderland

We used the Good–Turing, Horvitz–Thompson, and Poisson-binomial estimators to estimate the size of Lewis Carroll's vocabulary when he wrote "Alice's Adventures in Wonderland". The text consists of 24,168 words with 4,920 distinct types. The Good–Turing and Horvitz–Thompson estimates for the size of the underlying corpus were, respectively,

$$\hat{N}_{\text{GT}} = 8346, \quad \hat{N}_{\text{HT}} = 6988.8.$$

To put this in context, the vocabulary of an average adult native English speaker has been estimated to contain over 20,000 words (Zechmeister et al., 1995). Taken together, "Alice's Adventures in Wonderland" and "Through the Looking Glass" (Carroll, 1865, 1871) contain 8869 distinct words, already exceeding these estimates.

We derived the Poisson-binomial estimate by successively sampling $n = 48$ data sets $D_1, \ldots, D_{48}$ from the text and maximizing the product of corresponding likelihoods. In this case, we did not find it appropriate to set a hard bound on the number of types and used the brute force approach of the previous section. As the difficulty of optimization stems from the discrete nature of the parameter N for the family of Zipf distributions, we introduced a smooth parameter λ which determines N. We did this by assuming that N is a Poisson random variable with parameter λ, so that

$$\text{pr}\,(N = k \mid \lambda) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0, \quad k = 0, 1, \ldots.$$

Under this assumption, we write $L(\mathcal{N}) = L(a, \lambda)$. We use the expectation-maximization algorithm (Dempster et al., 1977) to maximize $L(a, \lambda)$. Given the maximum likelihood estimate $\hat{\lambda}$, the estsimate for the latent number of types was the expected values of the corresponding Poisson random variable, $\hat{N}_{\text{PB}} \equiv \hat{\lambda}$.

The Poisson-binomial estimate of the size of Lewis Carroll's vocabulary when he wrote "Alice's Adventures in Wonderland" is $\hat{N}_{\text{PB}} = 41,647.128$ (its standard error 191.748).

References

Carroll, L. (1865). Alice's Adventures in Wonderland. `http://www.gutenberg.org/ebooks/11`. Accessed: March 9th, 2010.

Carroll, L. (1871). Through the Looking-Glass. `http://www.gutenberg.org/ebooks/12`. Accessed: March 9th, 2010.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society Series B* **39**, 1–38.

Gale, W. A. & Sampson, G. (1995). Good–Turing frequency estimation without tears. *Journal of Quantitative Linguistics* **2**, 217–237.

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

Horvitz, D. G. & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Zechmeister, E. B., Chronis, A. M., Cull, W. L., D'Anna, C. A. & Healy, N. A. (1995). Growth of a functionally important lexicon. *Journal of Literacy Research* **27**, 201–212.