J. Appl. Prob. **48,** 1095–1113 (2011)
Printed in England

CONVERGENCE PROPERTIES IN CERTAIN OCCUPANCY PROBLEMS INCLUDING THE KARLIN–ROUAULT LAW

ESTÁTE V. KHMALADZE,* *Victoria University of Wellington*

Abstract

Let $\boldsymbol{x}$ denote a vector of length q consisting of 0s and 1s. It can be interpreted as an 'opinion' comprised of a particular set of responses to a questionnaire consisting of q questions, each having {0, 1}-valued answers. Suppose that the questionnaire is answered by n individuals, thus providing n 'opinions'. Probabilities of the answer 1 to each question can be, basically, arbitrary and different for different questions. Out of the 2^q different opinions, what number, μ_n, would one expect to see in the sample? How many of these opinions, $\mu_n(k)$, will occur exactly k times? In this paper we give an asymptotic expression for $\mu_n/2^q$ and the limit for the ratios $\mu_n(k)/\mu_n$, when the number of questions q increases along with the sample size n so that $n = \lambda 2^q$, where λ is a constant. Let $p(\boldsymbol{x})$ denote the probability of opinion $\boldsymbol{x}$. The key step in proving the asymptotic results as indicated is the asymptotic analysis of the joint behaviour of the intensities $np(\boldsymbol{x})$. For example, one of our results states that, under certain natural conditions, for any $z > 0$, $\sum \mathbf{1}_{\{np(\boldsymbol{x})>z\}} = d_n z^{-u}$, $d_n = o(2^q)$.

Keywords: Number of unique outcomes; sparse tables; Karlin–Rouault law; Zipf's law; Good–Turing index; large deviations; contiguity

2010 Mathematics Subject Classification: Primary 62D05; 62E20; 60E05; 60F10

1. Introduction

Consider the multinomial vector $\nu_n = (\nu_{1n}, \nu_{2n}, \ldots, \nu_{Nn})$ of frequencies of N disjoint events, with sample size n and vector of probabilities $p = (p_1, p_2, \ldots, p_N)$. Consider the statistics, sometimes called 'spectral statistics', based on these frequencies:

$$\mu_n(k) = \sum_{x=1}^{N} \mathbf{1}_{\{\nu_{xn}=k\}}, \qquad k = 1, 2, \ldots,$$

and

$$\mu_n = \sum_{x=1}^{N} \mathbf{1}_{\{\nu_{xn}\geq 1\}}.$$

Here and elsewhere, $\mathbf{1}_A$ is the indicator function of an event A and, therefore, $\mu_n(k)$ is the number of events occurring exactly k times and μ_n is the number of different events observed in the sample. These statistics have been central to classical occupancy problems. The bibliography is large; for a review and recent developments, we refer the reader to [3], [8], and [20]. Earlier references can be found in, e.g. [9].

Received 6 January 2011; revision received 10 August 2011.
* Postal address: School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, PO Box 600, Wellington, 2052, New Zealand. Email address: estate.khmaladze@msor.vuw.ac.nz

There are two main groups of problems associated with the spectral statistics, namely, various forms of the central limit theorem (CLT) and law of large numbers (LLNs). Papers [3] and [8] cited above established relatively difficult (local) forms for the CLT by extending the method called Poissonization. For this, the multinomial frequencies ν_{xn} are replaced by independent Poisson random variables with the same expected values np_x, $x = 1, 2, \dots, N$. For example, Lemma 2.1 of [3] states general conditions under which this replacement is valid. Integral CLT, but for the generalized occupancy problem, with several types of allocated particle, was recently presented in [20].

In this paper, however, we are interested in statements of the LLNs type, namely, in limits of the quotients

$$\frac{\mu_n(k)}{\mu_n}, \qquad k = 1, 2, \dots,$$

and the asymptotic behaviour of the ratio μ_n/n for the case that the vector p becomes 'essentially' dependent on n and changes as n increases. Specifically, we consider triangular arrays of multinomial distributions as n, N, and p change simultaneously.

For fixed p, the range of possibilities for the LLNs is narrow: for example, Lemma 3.2 of [12] shows that a positive limit for $\mu_n(k)/\mu_n$ exists if and only if the probabilities p_x, arranged in decreasing order, form a regularly varying function of x. In this case the limits of $\mu_n(k)/\mu_n$ form a discrete version of the Karlin–Rouault law, derived in [14].

Prior to analysis of any particular triangular array, it would be convenient to have some classification of such arrays. In [12] two main classes have been defined and investigated: the class of multinomial distributions, called (d1), such that

$$\liminf_{n\to\infty} \frac{\mu_n(1)}{n} > 0,$$

and another class, called (d2), such that

$$\lim_{n\to\infty} \mu_n = \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{\mu_n(1)}{\mu_n} > 0.$$

These definitions are not equivalent, with (d1) $\subset$ (d2); very interesting cases are found in (d2) $\setminus$ (d1). Indeed, no other arrays can lead to such well-known laws as Zipf's law (see, e.g. [1] and [13]), or the Karlin–Rouault law (see, e.g. [24] or Theorem 2 below). In both of these cases $\mu_n/n \to 0$ and, hence, they cannot be obtained within the simpler class (d1). Note that the earlier classifications, which we know about, do not reflect the class (d2) $\setminus$ (d1). For example, in the frequently cited classification given in [17] there are three 'zones': a 'central zone', where, for n and $N \to \infty$ and some constants C, C_1, and C_2,

$$n \max_x p_x < C, \qquad C_1 < \frac{n}{N} < C_2;$$

a 'left m-zone' for $m \geq 2$, where, for a constant λ,

$$n \max_x p_x \to 0, \qquad \mathrm{E}\,\mu_n(m) \to C;$$

and a 'right m-zone' for $m \geq 0$, where

$$n \min p_x \to \infty, \qquad \mathrm{E}\,\mu_n(m) \to C.$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

The latter 'right' zones are irrelevant here as then $\mathrm{E}\,\mu_n(k)/\mathrm{E}\,\mu_n \to 0$ for all k and, therefore, even (d2) is not possible, while in the case of Zipf's law or Karlin–Rouault's law only $\mathrm{E}\,\mu_n(1)/\mathrm{E}\,\mu_n$ is half or close to half (see Section 5). In the first two zones above only (d1) is possible and in the left m-zone only the degenerate limit $\lim_{n\to\infty} \mathrm{E}\,\mu_n(1)/n = 1$ is possible (see the functional limit theorems for this case in [21]). In the situation we study in this paper $n \max_x p_x \to \infty$ while $n \min_x p_x \to 0$. Moreover, for overwhelmingly many x, $np_x \to 0$, while although $np_x \to \infty$ for increasingly many x, the number of such x is $o(N)$; see Corollary 4.

Below we focus on one particular form of probabilities given not on positive integers but on the set Ξ_q of all vectors $\boldsymbol{x}$ of length q consisting of 0s and 1s, as, for example, $(0, 1, 0, \ldots, 1)^\top$. Let $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_q)^\top$ denote a q-dimensional random vector with Bernoulli random variables as coordinates. Heuristically, $\boldsymbol{\xi}$ can be interpreted in many ways: as a random response to a questionnaire with q questions with binary-valued answers, as a randomly changing state of a system with q 'on/off' components, or as a result of, say, taxonomic evaluation with q 'present/absent' classifiers, and so on. In the latter case observation often leads to the so-called q-dimensional 'sparse tables'. The sparse tables are also commonly found in classifications of industrial companies or financial institutions. Usual practice consists of reducing such high-dimensional tables to lower-dimensional tables, say, 2×2 or 3×3 tables. In this paper, however, the approach is to take high-dimensional sparse tables as they are, in order not to lose diversity in the underlying sample. Below we use the first example of questionnaires when needed.

Let $p_q(\boldsymbol{x}) = \mathrm{P}\{\boldsymbol{\xi} = \boldsymbol{x}\}$ denote the distribution of $\boldsymbol{\xi}$ on Ξ_q, and let $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n$ denote an independent and identically distributed (i.i.d.) sequence of n vectors with this distribution. In this setting our $N = 2^q$ and, for each of the 2^q possible vectors $\boldsymbol{x}$, we denote the frequency in this sample by

$$\nu_n(\boldsymbol{x}) = \sum_{j=1}^{n} \mathbf{1}_{\{\boldsymbol{\xi}_j = \boldsymbol{x}\}}\,.$$

From now on we consider spectral statistics based on these frequencies $\nu_n(\boldsymbol{x})$ and study their asymptotic behaviour as n and $N \to \infty$ (hence, $q \to \infty$) in such a way that $n = \lambda N$, where λ is a constant.

Instead of trying to arrange the probabilities $p(\boldsymbol{x})$ in decreasing order, as is commonly done for scalar $\boldsymbol{x}$ but would not seem natural in this case, we study the tail of their distribution function

$$H_n(z) = \frac{1}{N} \sum_{\boldsymbol{x} \in \Xi_q} \mathbf{1}_{\{Np_q(\boldsymbol{x}) \geq z\}}\,. \tag{1}$$

It is easy to see that H_n provides a complete symmetric characteristic of the probabilities $p_q(\boldsymbol{x})$: any symmetric function of these probabilities is a functional from H_n. We will also see that it is fruitful to shift our attention from $np(\boldsymbol{x})$, usually perceived as an expected value, to $Np_q(\boldsymbol{x})$, which differs from it under our asymptotics only by a constant. The latter quantity, however, can be viewed as a likelihood ratio of p_q and the uniform distribution on Ξ_q, given by the probabilities $p_0(\boldsymbol{x}) = 1/2^q$. Then H_n is the distribution of this likelihood ratio under this uniform measure. Hence, in its asymptotic analysis we can use asymptotic methods developed for likelihood ratios, as we indicate in Sections 2 and 3.

To be precise, we show in Section 3 that if the coordinates of $\boldsymbol{\xi}$ are independent, that is, if the $p(\boldsymbol{x})$ are given by (4) below with, basically, arbitrary probabilities $a_1, \ldots, a_q$ for the answer

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

'yes' to each question, then

$$H_n(z) \sim c_n(\lambda z)^{-u}, \quad \text{with} \quad c_n = \frac{\mathrm{E}\,\mu_n/N}{\Gamma(1-u)} \to 0. \tag{2}$$

In the limit, the parameter u is the only trace of the probabilities $p_q(\boldsymbol{x})$, $\boldsymbol{x} \in \Xi_q$; numerically, it is surprisingly stable: its typical values lie between 0.4 and 0.5. According to (2), it is not true that the majority of frequencies $\nu_n(\boldsymbol{x})$, $\boldsymbol{x} \in \Xi_q$, have asymptotically any nondegenerate Poisson distribution. Rather, the contrary holds in that the intensities $np(\boldsymbol{x})$ for the overwhelming majority converge to 0.

Although one does not have to insist on any sort of randomness of the probabilities $p(\boldsymbol{x})$, it may help to gain some insight if we use familiar terminology and say that the intensities $np(\boldsymbol{x})$ overall behave as 'asymptotically small random variables' and that 'their distribution belongs to the domain of attraction' of the u-stable law, of which $R(z) = z^{-u}$ is the Lévy–Khinchine measure.

For the triangular arrays, we consider it is possible that H_n may behave differently from (2) and actually converge to a nondegenerate limit distribution. Conditions for this, in terms of the contiguity of the distributions $p_q(\boldsymbol{x})$ and $p_0(\boldsymbol{x})$, are given in Section 2.

In principle, one might expect that if in a sample of size n an event $\boldsymbol{x}$ occurs $\nu_n(\boldsymbol{x})$ times then its probability would be estimated best by the maximum likelihood estimator (MLE) $\hat{p}_q(\boldsymbol{x}) = \nu_n(\boldsymbol{x})/n$. As a corollary of Theorem 2, we conclude that these MLEs would be unsatisfactory estimators: the asymptotic behaviour of the function

$$\widehat{H}_n(z) = \frac{1}{N} \sum_{\boldsymbol{x} \in \Xi_q} \mathbf{1}_{\{\nu_n(\hat{\boldsymbol{x}}) > z\}} \tag{3}$$

as $n \to \infty$ is different from that of H_n; the two are similar only for large values of z. There are other corollaries which we give in Section 4. One of them shows the asymptotic expressions for the so-called Good–Turing indices (see [6]) and, in particular, for the overall probability of the outcomes not seen in the sample. The other corollary shows how μ_n increases with n, which is an important question in studying the diversity of biological species for example.

As a final aside, we recall that the case where all the a_i are equal but different from $\frac{1}{2}$, was studied in [14]. The proofs in that paper are quite different from what we give below.

2. The approach and the case of contiguity

To begin with, let us assume that the coordinates $\xi_1, \ldots, \xi_q$ of each $\boldsymbol{\xi}$ are independent with $\mathrm{P}\{\xi_i = 1\} = a_i$. Then the probability of an outcome $\boldsymbol{x}$ equals

$$p_q(\boldsymbol{x}) = \prod_{i=1}^{q} a_i^{x_i} (1 - a_i)^{1 - x_i}. \tag{4}$$

As $q \to \infty$, these probabilities tend to 0; we shall see that the expectations $np_q(\boldsymbol{x})$, mostly, remain bounded or may even be small. Therefore, an assumption of asymptotically Poisson behaviour of the frequencies $\nu_n(\boldsymbol{x})$ looks natural; in the expressions below we assume that the $\nu_n(\boldsymbol{x})$, $\boldsymbol{x} \in \Xi_q$, are independent Poisson random variables with means $np_q(\boldsymbol{x})$. Although, for triangular arrays, this Poissonization can essentially change the CLT statement for spectral statistics, more so for the functional CLT (see, e.g. [11]); for the LLNs, it makes no difference.

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

The key step in our asymptotic analysis of the ratio μ_n/N and $\mu_n(k)/\mu_n$ consists of the analysis of $\mathrm{E}\,\mu_n/N$ and $\mathrm{E}\,\mu_n(k)/\mathrm{E}\,\mu_n$, because it is possible to prove that the ratios $\mu_n/\mathrm{E}\,\mu_n$ and $\mu_n(k)/\mathrm{E}\,\mu_n(k)$ converge to 1 almost surely. We therefore concern ourselves with the following expressions for the expected values:

$$\mathrm{E}\,\mu_n(k) = \sum_{\boldsymbol{x}\in\Xi_q} \pi(k, np_q(\boldsymbol{x})) \quad \text{and} \quad \mathrm{E}\,\mu_n = \sum_{\boldsymbol{x}\in\Xi_q} [1 - \pi(0, np_q(\boldsymbol{x}))].$$

Here $\pi(k, z) = z^k \mathrm{e}^{-z}/k!$ denotes the Poisson probability of k with mean z.

In some cases we can study these expected values as they are. For example, if all the a_i equal $\frac{1}{2}$, i.e. if all the $p_q(\boldsymbol{x})$ equal $1/2^q$, we immediately deduce that

$$\frac{\mathrm{E}\,\mu_n}{2^q} = 1 - \pi(0, \lambda) \quad \text{and} \quad \frac{\mathrm{E}\,\mu_n(k)}{\mathrm{E}\,\mu_n} = \frac{\pi(k, \lambda)}{1 - \pi(0, \lambda)}. \tag{5}$$

In particular, this implies that we should expect the number of different opinions in a sample to be of the same order as the number of all possible opinions.

In the general situation, where all the a_i can be different, or even simply not equal to $\frac{1}{2}$, the asymptotic analysis of the sums $\mathrm{E}\,\mu_n$ and $\mathrm{E}\,\mu_n(k)$ may not appear to be that simple. However, we can turn it into a purely probabilistic problem, using general and simple tools to study it, which otherwise would seem irrelevant and distant to the problem. Specifically, let P_q and P_0 denote the distributions on Ξ_q defined by $p_q(\boldsymbol{x})$ and $p_0(\boldsymbol{x}) = 1/2^q$, respectively. Then

$$M_q(\boldsymbol{x}) := 2^q\, p_q(\boldsymbol{x}) = \frac{p_q(\boldsymbol{x})}{p_0(\boldsymbol{x})}$$

is the likelihood ratio of P_q and P_0, and we can write

$$\begin{aligned} \frac{\mathrm{E}\,\mu_n(k)}{2^q} &= \int_0^\infty \pi(k, \lambda z)\,\mathrm{d}H_n(z) = \mathrm{E}_0\,\pi(k, \lambda M_q(\boldsymbol{\xi})), \qquad k = 1, 2, \ldots, \\ \frac{\mathrm{E}\,\mu_n}{2^q} &= \int_0^\infty [1 - \pi(0, \lambda z)]\,\mathrm{d}H_n(z) = \mathrm{E}_0[1 - \pi(0, \lambda M_q(\boldsymbol{\xi}))], \end{aligned} \tag{6}$$

where E_0 denotes the expectation calculated with respect to the uniform distribution P_0 of $\boldsymbol{\xi}$.

As we have just seen in (5), if the questionnaire is 'symmetric', or 'balanced', that is, if each a_i is equal to $\frac{1}{2}$, the ratio $\mu_n/2^q$ has a positive limit. The same should be true then for 'nearly symmetric' or 'nearly balanced' questionnaires, when the a_is are close enough to $\frac{1}{2}$. Using M_q, we immediately obtain the tool to describe this situation in a relatively complete form.

Specifically, if the sequence of distributions P_q is contiguous with respect to the sequence of uniform distributions P_0, then, under P_0, M_q typically converges in distribution to a random variable e^L, where L is normal $N(-c^2/2, c^2)$, and the expected values in (6) converge to the corresponding limits. In Theorem 1 below we formally state the conditions and specify the constant c^2. In this theorem and everywhere below, $\Phi_{\mu,\sigma^2}(z)$ and $\phi_{\mu,\sigma^2}(z)$ denote the normal distribution function and normal density with mean μ and variance σ^2, respectively.

Theorem 1. *Suppose that the probabilities $a_{1q}, \ldots, a_{qq}$ form a triangular array in q such that*

$$\max_i \left| a_{iq} - \frac{1}{2} \right| \to 0 \quad \textit{and} \quad \limsup_{q\to\infty} \sum_{i=1}^{q} [1 - \sqrt{2a_{iq}}] < \infty.$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

Then

$$\liminf_{q\to\infty} \frac{\mathrm{E}\,\mu_n}{2^q} > 0.$$

If the finite limit

$$\lim_{q\to\infty} \sum_{i=1}^{q} [1 - \sqrt{2a_{iq}}] = \frac{c^2}{2}$$

exists then

$$\frac{\mathrm{E}\,\mu_n}{2^q} \sim \int (1 - \pi(0, \lambda \mathrm{e}^z)) \Phi_{-c^2/2, c^2}(\mathrm{d}z)$$

and

$$\frac{\mathrm{E}\,\mu_n(k)}{\mathrm{E}\,\mu_n} \sim \frac{\int \pi(k, \lambda \mathrm{e}^z) \Phi_{-c^2/2, c^2}(\mathrm{d}z)}{\int (1 - \pi(0, \lambda \mathrm{e}^z)) \Phi_{-c^2/2, c^2}(\mathrm{d}z)}. \tag{7}$$

Proof. The condition on the limit supremum of $\sum_{i=1}^{q} [1 - \sqrt{2a_{iq}}]$ guarantees contiguity of the sequence of distributions P_q to the sequence of uniform distributions P_0. In its turn, the contiguity implies that the sequence of distributions of the log-likelihood ratio $\ln M_q$ is weakly compact. Hence, the result on $\mathrm{E}_0\,\mu_n$ follows.

Existence of the limit, together with the condition on $\max_i |a_{iq} - \frac{1}{2}|$, guarantees asymptotic normality of $\ln M_q$ (see, e.g. [22]) with parameters, under the null distribution, equal to $-c^2/2$ and c^2. This asymptotic normality implies convergence of the expected values as the integrands are continuous and bounded functions of $\ln M_q$.

Remark. Note that the result extends to a very general class of distributions. Namely, whether the coordinates $\xi_1, \ldots, \xi_q$ are independent and $p_q(\boldsymbol{x})$ is a product of Bernoulli distributions or matters little. For any distribution on Ξ_q, the quantity M_q still remains a likelihood ratio and, hence, a martingale in q. The conditions of asymptotic normality of $\ln M_q$, if M_q is a positive martingale, are well known: if now a_{iq} is random and denotes the conditional probability of $\xi_i = 1$ given $\xi_1, \ldots, \xi_{i-1}$ then, notationally, the same conditions, with convergence replaced by convergence in probability, imply asymptotic normality for L_q (see [7]). Therefore, the statement of Theorem 1 remains true.

Note also that, under the conditions of Theorem 1, the array of distributions P_q with $n = \lambda 2^q$ belongs to the class (d1). Under the conditions of Theorem 2 below, it belongs to the class (d2) \ (d1).

3. The case of arbitrary a_is

Suppose now that the probabilities $a_1, \ldots, a_q$ are arbitrary, i.e. they form some sequence in q. In this case the behaviour of the likelihood ratio M_q becomes somewhat erratic: under P_0, we have $M_q \to 0$ in probability, but $\mathrm{E}_0\, M_q = 1$, and, therefore, increasingly large values of M_q are unavoidable. Consequently, in the asymptotic analysis of the expectation $\mathrm{E}_0[1 - \pi(0, \lambda M_q)]$ we can no longer rely, say, on Taylor series approximations like

$$1 - \mathrm{e}^{-\lambda M_q} \sim \lambda M_q.$$

Indeed, in the mean the asymptotic behaviour as just indicated is not correct: we show shortly that

$$\mathrm{E}_0[1 - \pi(0, \lambda M_q)] \to 0, \quad \text{while} \quad \mathrm{E}_0\, \lambda M_q = \lambda.$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

The size of the quantity $\mathrm{E}_0[1-\pi(0, \lambda M_q)]$ is not immediately obvious: for large q, while the random variable M_q is small with large probability, the integrand $1-\pi(0, \lambda M_q)$ also becomes small, and although M_q is large with only a small probability, the integrand is close to 1, i.e. it is not small. We did not find it fruitful to try and locate a part where the main contribution to the integral $\mathrm{E}_0[1-\pi(0, \lambda M_q)]$ is made directly. Instead, we express it as a certain probability which, as we shall see, is connected in a natural way with the theory of large deviations.

Let T_1 be an exponential random variable with scale parameter 1, independent of M_q, and let $\eta_1 = \ln T_1$. The distribution function of η_1 is $1-\pi(0, \mathrm{e}^x) = 1-\mathrm{e}^{-\mathrm{e}^x}$. As above, let $L_q = \ln M_q$ denote the log-likelihood. Then we can write

$$\mathrm{E}_0[1-\pi(0, \lambda M_q)] = \mathrm{P}_0\{L_q > \eta_1 - \ln\lambda\}.$$

Similarly, if T_k is a gamma-distributed random variable with shape parameter k, i.e. if T_k is a sum of k independent copies of T_1, independent of M_q, and if $\eta_k = \ln T_k$, then

$$\mathrm{E}_0 \sum_{j=k}^{\infty} \pi(j, \lambda M_q) = \mathrm{P}_0\{L_q > \eta_k - \ln\lambda\}.$$

With some abuse of notation, in the last two displayed formulae (and within some proofs later on) we use P_0 for the joint distribution of L_q, under the uniform distribution on Ξ_q, and η_k with the appropriate k.

It is clear that

$$L_q = \ln \frac{p(\boldsymbol{\xi})}{p_0(\boldsymbol{\xi})} = \sum_{i=1}^{q} [\xi_i \ln 2a_i + (1-\xi_i)\ln 2(1-a_i)],$$

where $\xi_1, \ldots, \xi_q$, under P_0, are independent symmetric Bernoulli random variables: $\mathrm{P}_0\{\xi_i = 1\} = \frac{1}{2}$. Let $\psi_i(u)$ denote the logarithm of the moment generating function of each summand

$$\begin{aligned}\psi_i(u) &= \ln(\mathrm{E}_0 \exp u[\xi_i \ln 2a_i + (1-\xi_i)\ln 2(1-a_i)]) \\ &= \ln[2^u(a_i^u + (1-a_i)^u)] - \ln 2.\end{aligned}$$

Now $\psi_i(u)$ is a convex, infinitely differentiable function of u and $\psi_i(0) = \psi_i(1) = 0$ (see, e.g. [10]). Then so too is the sum $\sum_{i=1}^{q} \psi_i(u)$, which is the logarithm of the moment generating function of L_q.

Consider the sequence $a_1, a_2, \ldots, a_q$, and let

$$F_q(a) = \frac{1}{q} \sum_{i=1}^{q} \mathbf{1}_{\{a_i < a\}}$$

denote the empirical distribution function of this sequence. Again, by using the term 'empirical distribution function' we do not imply that $a_1, a_2, \ldots, a_q$ are to be considered as independent random variables. We assume only a certain ergodic property, namely, that there is a continuous distribution function F on the interval $[0, 1]$ such that, as $q \to \infty$,

$$\begin{gathered} F_q(a) \to F(a) \quad \text{for all } a \in [0, 1], \\ \int_0^1 \left(\ln \frac{a}{1-a}\right)^2 \mathrm{d}F_q(a) \to \int_0^1 \left(\ln \frac{a}{1-a}\right)^2 \mathrm{d}F(a) < \infty. \end{gathered} \tag{8}$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

In the second condition we assume that, asymptotically, we do not have too many a_i too close to 0 or 1. For example, F can be any beta distribution.

Let $\psi_i'(u)$ and $\psi_i''(u)$ denote the first and second derivatives of $\psi_i(u)$.

Lemma 1. *Suppose that the conditions in (8) are satisfied. Let $u = u_q$ be such that*

$$\sum_{i=1}^{q} \psi_i'(u) = 0.$$

Define

$$\sigma_q^2 = \sum_{i=1}^{q} \frac{\psi_i''(u_q)}{q}.$$

Then $\lim_{q\to\infty} u_q$ and $\lim_{q\to\infty} \sigma_q^2$ exist, with

$$0 < \lim_{q\to\infty} u_q < 1 \quad \text{and} \quad 0 < \lim_{q\to\infty} \sigma_q^2 < \infty.$$

Proof. It is easy to see that the conditions in (8) imply the convergence

$$\sum_{i=1}^{q} \frac{\psi_i(u)}{q} \to \int_0^1 \ln[2^u(a^u + (1-a)^u)]\,\mathrm{d}F(a) - \ln 2$$

for all $u \in [0, 1]$ together with the convergence for the first two derivatives. In particular,

$$\sum_{i=1}^{q} \frac{\psi_i'(0)}{q} \to \frac{1}{2}\int_0^1 (\ln 4a(1-a))\,\mathrm{d}F(a) > -\infty$$

and

$$\sum_{i=1}^{q} \frac{\psi_i'(1)}{q} \to \int_0^1 [a\ln a + (1-a)\ln(1-a)]\,\mathrm{d}F(a) + \ln 2 < \infty.$$

Therefore, both limits in the lemma exist, and since the limit of $\sum_{i=1}^{q} \psi_i(u)/q$ is also a convex function, equal to 0 at $u = 0$ and 1, the limit of u_q cannot equal 0 or 1.

An essential step in Theorem 2 below is given by the following lemma.

Lemma 2. *Suppose that the conditions in (8) are satisfied. Then, with u as in Lemma 1,*

$$\mathrm{P}_0\{L_q > z\} \sim \exp\left[\sum_{i=1}^{q} \psi_i(u) - uz\right] \frac{1}{u\sqrt{q}}\,\phi_{0,\sigma_q^2}\left(\frac{z}{\sqrt{q}}\right)[1 + r_q(z)], \tag{9}$$

where, for any fixed $\beta > 0$,

$$\sup_{-\beta\sqrt{q}<z<\beta\sqrt{q}} |r_q(z)| = o(1) \quad \text{as } q \to \infty.$$

Remark. Since

$$\frac{\mathrm{E}_0\, L_q}{q} \to \frac{1}{2}\int_0^1 [\ln 4a(1-a)]\,\mathrm{d}F(a) < 0,$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

and, therefore, $L_q \to -\infty$, the probability $\mathrm{P}_0\{L_q > z\}$ for any given z is a large deviation probability for L_q. Lemma 2 exhibits the asymptotic expression for this probability and not its logarithm, as is more often stated in the literature. For the i.i.d. case, the idea can already be seen in [2], and, for the general case, it was carried through, with the aid of some assumptions, in [4]. Lemma 2 also states that the asymptotic expression is correct uniformly in z in increasing intervals of length $\sqrt{q}$. We could have extended its length to $o(q^{3/4})$, but do not need this—the rate $\sqrt{q}$ is sufficient for the application of (9) in Theorem 2 below.

Proof of Lemma 2. Consider the distribution Q that is adjoint to P_0. It is defined by

$$\mathrm{P}_0\{L_q > z\} = \exp\left[\sum_{i=1}^{q} \psi_i(u)\right] \int_z^\infty \mathrm{e}^{-ux}\, \mathrm{dQ}(x). \tag{10}$$

Then the moment generating function of Q is given by

$$\int \mathrm{e}^{rt}\, \mathrm{dQ}(t) = \exp\left[\sum_{i=1}^{q} [\psi_i(u+r) - \psi_i(u)]\right],$$

and, therefore, with the choice of u as in the lemma, the expected value of Q is 0 and its variance is $q\sigma_q^2$. Denote the distribution of $L_q/\sqrt{q}$ under the distribution Q by $\mathrm{Q}_{L_q/\sqrt{q}}$. Then (10) can be rewritten as

$$\begin{aligned}\mathrm{P}_0\{L_q > z\} &= \exp\left[\sum_{i=1}^{q} \psi_i(u)\right] \int_{z/\sqrt{q}}^\infty \mathrm{e}^{-u\sqrt{q}y}\, \mathrm{d}\, \mathrm{Q}_{L_q/\sqrt{q}}(y) \\ &= \exp\left[\sum_{i=1}^{q} \psi_i(u) - uz\right] \int_0^\infty \mathrm{e}^{-u\sqrt{q}x}\, \mathrm{d}\, \mathrm{Q}_{L_q/\sqrt{q}}\left(x + \frac{z}{\sqrt{q}}\right). \end{aligned} \tag{11}$$

Since $\mathrm{Q}_{L_q/\sqrt{q}}$ is the distribution of a normalized sum of independent and bounded random variables with mean 0 and variance σ_q^2, it can be approximated by a normal distribution with the same moments. First we replace $\mathrm{Q}_{L_q/\sqrt{q}}(x)$ by $\Phi_{0,\sigma_q^2}(x)$ and then justify this replacement. We obtain

$$\begin{aligned} u\sqrt{q} \int_{z/\sqrt{q}}^\infty \mathrm{e}^{-u\sqrt{q}y} \phi_{0,\sigma_q^2}(y)\, \mathrm{d}y &= \mathrm{e}^{-uz} u\sqrt{q} \int_0^\infty \mathrm{e}^{-u\sqrt{q}x} \phi_{0,\sigma_q^2}\left(x + \frac{z}{\sqrt{q}}\right) \mathrm{d}x \\ &= \mathrm{e}^{-uz} \phi_{0,\sigma_q^2}\left(\frac{z}{\sqrt{q}}\right)[1 + r_q(z)], \end{aligned} \tag{12}$$

where

$$\sup_{|z|<\beta\sqrt{q}} |r_q(z)| \to 0 \quad \text{as } q \to \infty.$$

Note that, to obtain nonzero asymptotics, we have normalized the integral above by $\sqrt{q}$. Therefore, we must consider the normalized difference

$$u\sqrt{q} \int_{z/\sqrt{q}}^\infty \mathrm{e}^{-u\sqrt{q}y} [\mathrm{Q}_{L_q/\sqrt{q}}(\mathrm{d}y) - \Phi_{0,\sigma_q^2}(\mathrm{d}y)],$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

in which the difference $\sqrt{q}[\mathrm{Q}_{L_q/\sqrt{q}}(y) - \Phi_{0,\sigma_q^2}(y)]$ need not be small; we need a better approximation for $\mathrm{Q}_{L_q/\sqrt{q}}(y)$, which we can obtain in the form of an Edgeworth expansion (see the next lemma). According to this expansion,

$$\sup_y |\mathrm{Q}_{L_q/\sqrt{q}}(y) - C_q(y)| = o\left(\frac{1}{\sqrt{q}}\right), \tag{13}$$

where

$$C_q(y) = \Phi_{0,\sigma_q^2}(y) + \frac{P(y)\phi_{0,\sigma_q^2}(y)}{\sqrt{q}}$$

with $P(y) = y^3 - 3y$, the third Hermite polynomial. The asymptotics in (12) are not affected by the term $P(y)\phi_{0,\sigma_q^2}(y)/\sqrt{q}$, while using integration by parts leads to

$$\begin{aligned}
&\sqrt{q}\int_0^\infty \mathrm{e}^{-u\sqrt{q}x}\,\mathrm{d}\left[\mathrm{Q}_{L_q/\sqrt{q}}\left(x+\frac{z}{\sqrt{q}}\right) - C_q\left(x+\frac{z}{\sqrt{q}}\right)\right] \\
&\quad = -\sqrt{q}\left[\mathrm{Q}_{L_q/\sqrt{q}}\left(\frac{z}{\sqrt{q}}\right) - C_q\left(\frac{z}{\sqrt{q}}\right)\right] \\
&\qquad + uq\int_0^\infty \mathrm{e}^{-u\sqrt{q}x}\left[\mathrm{Q}_{L_q/\sqrt{q}}\left(x+\frac{z}{\sqrt{q}}\right) - C_q\left(x+\frac{z}{\sqrt{q}}\right)\right]\mathrm{d}x \\
&\quad \to 0,
\end{aligned}$$

uniformly in z.

The next lemma shows that the Edgeworth expansion (13) for the distribution $\mathrm{Q}_{L_q/\sqrt{q}}(z)$ does indeed exist.

Lemma 3. *If the conditions in (8) are satisfied then there exists an Edgeworth expansion for the distribution function* $\mathrm{Q}_{L_q/\sqrt{q}}(z)$.

Proof. Use the notation $q(a_i) = q_i = \mathrm{Q}(\xi_i = 1)$ and $\omega(a_i) = \omega_i = \ln[a_i/(1-a_i)]$. Note that $q(a) = a^u/[a^u + (1-a)^u]$. Then

$$\xi_i(t) = \mathrm{e}^{-\mathrm{i}tq_i\omega_i}[q_i(\mathrm{e}^{\mathrm{i}t\omega_i} - 1) + 1]$$

is the characteristic function of the ith summand of L_q in the measure Q. For the proof, we need to show (14) below, while the rest basically follows the lines of the proof for the i.i.d. case given in [5, Chapter XVI.2–4]. We give only a sketch. If $G_q(z)$ is as in Lemma 2 and $\gamma_q(t)$ is its Fourier transform, then, for arbitrarily small ε, there exists a large enough constant b such that

$$|\mathrm{Q}_{L_q/\sqrt{q}}(z) - G(z)| \le \int_{-b\sqrt{q}}^{b\sqrt{q}} \frac{|\prod_{i=1}^q \xi_i(t/\sqrt{q}) - \gamma_q(t)|}{t}\,\mathrm{d}t + \frac{\varepsilon}{\sqrt{q}},$$

and we can split the domain of integration into $|t| < \delta\sqrt{q}$ and $\delta\sqrt{q} < |t| < b\sqrt{q}$. For $|t| < \delta\sqrt{q}$, the expansion of the characteristic function $\prod_{i=1}^q \xi_i(t/\sqrt{q})$ of $L_q/\sqrt{q}$, just as in the case of i.i.d. random variables, shows that the corresponding integral is $o(1/\sqrt{q})$. For intervals $\delta\sqrt{q} < |t| < b\sqrt{q}$, it is sufficient to show that

$$\sup_{\delta < |t|/\sqrt{q} < a} \left|\prod_{i=1}^q \xi_i\left(\frac{t}{\sqrt{q}}\right)\right| < c^q \quad \text{for some } 0 < c < 1. \tag{14}$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

However, for the norm of this characteristic function, we have

$$\begin{aligned}\frac{1}{q}\ln\prod_{i=1}^{q}|\xi_i(s)| &= \frac{1}{q}\sum_{i=1}^{q}\ln[1+2q_i(1-q_i)(\cos s\omega_i-1)]\\ &= \int_0^1 \ln[1+2q(a)(1-q(a))(\cos s\omega(a)-1)]\,\mathrm{d}F_q(a)\\ &\le 2\int_0^1 q(a)(1-q(a))(\cos s\omega(a)-1)\,\mathrm{d}F_q(a).\end{aligned}$$

Now we need to show that this integral becomes less than some negative number $-\varepsilon$, uniformly for $s\in[\delta,b]$. If H_q and H are respectively the empirical and limit distribution functions of the ω_is, then

$$\int_0^1[1-\cos s\omega(a)][\mathrm{d}F_q(a)-\mathrm{d}F(a)]=\int_{-\infty}^{\infty}(1-\cos s\omega)[\mathrm{d}H_q(\omega)-\mathrm{d}H(\omega)],$$

and integration by parts leads to

$$s\left|\int_{-\infty}^{\infty}\sin s\omega[H_q(\omega)-H(\omega)]\,\mathrm{d}\omega\right|\le s\int_{-\infty}^{\infty}|H_q(\omega)-H(\omega)|\,\mathrm{d}\omega.$$

The conditions in (8) imply that the last integral converges to 0, because they guarantee both that $H_q(\omega)\to H(\omega)$ uniformly in ω and that the second moment (hence, also the first absolute moment) converges. Obviously, this is true uniformly in $s\in[\delta,b]$. On the other hand, for any continuous distribution,

$$\int_{-\infty}^{\infty}\cos s\omega\,\mathrm{d}H(\omega)<1-2\varepsilon$$

for $s>\delta$ and, therefore, (14) is true with $c=1-\varepsilon$.

Note that the form of condition (14) varies in the literature. Often, it may seem simpler to require this inequality to hold uniformly for $t>\delta$ (see, e.g. [16, p. 34]). However, this requirement would be restrictive for us: under (8), it will not be true generally. To see this, consider F_q, which attaches equal weight $1/q$ to regularly spaced points j/q, $j=1,\dots,q$. However, if $a_1,\dots,a_q$ were assumed to be independent random variables then (14) would be true for $t>\delta$.

Now we are ready to formulate the following theorem. The expression $R_u(k)$ is known in the literature as the Karlin–Rouault law.

Theorem 2. *If, with $N=2^q$, $n=\lambda N$, and $q\to\infty$, the conditions in (8) are satisfied and $u=\lim u_q$, then*

$$\begin{aligned}\frac{\mathrm{E}\,\mu_n}{2^q}&\sim\exp\left[\sum_{i=1}^{q}\psi_i(u_q)\right]\frac{\lambda^u}{u\sqrt{q}}\phi_{0,\sigma_q^2}(0)\Gamma(1-u),\\ \frac{\mathrm{E}\,\mu_n(k)}{\mathrm{E}\,\mu_n}&\to R_u(k)=\frac{u\Gamma(k-u)}{\Gamma(1-u)\Gamma(k+1)},\end{aligned}\tag{15}$$

for every fixed $k=1,2,\dots$.

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

Proof. We start with the asymptotic expression for

$$\mathrm{E}_0 \sum_{j=k}^{\infty} \mu_n(j) = \mathrm{P}_0\{L_q > \eta_k - \ln\lambda\}, \qquad k \geq 1.$$

Let F_k denote the gamma distribution function with shape parameter k and scale parameter 1. Then $F_k(\mathrm{e}^x)$ is the distribution function of η_k. We have

$$\mathrm{P}_0\{L_q > \eta_k - \ln\lambda\} = \int_{-\infty}^{\infty} \mathrm{P}_0\{L_q > z - \ln\lambda\}\,\mathrm{d}F_k(\mathrm{e}^z), \tag{16}$$

where $\int_{-\infty}^{\infty} = \int_{-\infty}^{-\beta\sqrt{q}} + \int_{-\beta\sqrt{q}}^{\beta\sqrt{q}} + \int_{\beta\sqrt{q}}^{\infty}$. Using (11), for the integral over $(-\infty, -\beta\sqrt{q}]$, we have

$$\begin{aligned}
&\int_{-\infty}^{-\beta\sqrt{q}} \mathrm{P}_0\{L_q > z - \ln\lambda\}\,\mathrm{d}F_k(\mathrm{e}^z) \\
&\quad = F_k(\mathrm{e}^{-\beta\sqrt{q}})\,\mathrm{P}_0\{L_q > -\beta\sqrt{q} - \ln\lambda\} \\
&\qquad + \exp\Bigl[\sum_{i=1}^{q} \psi_i(u)\Bigr] \int_{-\infty}^{-\beta\sqrt{q}} F_k(\lambda \mathrm{e}^{\sqrt{q}z}) \mathrm{e}^{-u\sqrt{q}z}\,\mathrm{d}\,\mathrm{Q}_{L_q/\sqrt{q}}(z).
\end{aligned}$$

Since $F_k(\varepsilon) < \frac{1}{2}\varepsilon^k < \frac{1}{2}\varepsilon$ for all sufficiently small ε, we obtain

$$\begin{aligned}
\int_{-\infty}^{-\beta\sqrt{q}} \mathrm{P}_0\{L_q > z - \ln\lambda\}\,\mathrm{d}F_k(\mathrm{e}^z) &< \mathrm{e}^{-\beta\sqrt{q}}\,\mathrm{P}_0\{L_q > -\beta\sqrt{q} - \ln\lambda\} \\
&\quad + \exp\Bigl[\sum_{i=1}^{q} \psi_i(u)\Bigr] \lambda \int_{-\infty}^{-\beta\sqrt{q}} \mathrm{e}^{\sqrt{q}(1-u)z}\,\mathrm{d}\,\mathrm{Q}_{L_q/\sqrt{q}}(z).
\end{aligned}$$

The last integral on the right-hand side is $O(\mathrm{e}^{-q(1-u)\beta})$, where u stays strictly inside $[0, 1]$ for all large enough q. Similarly, for the interval $[\beta\sqrt{q}, \infty)$, we have

$$\int_{\beta\sqrt{q}}^{\infty} \mathrm{P}_0\{L_q > z - \ln\lambda\}\,\mathrm{d}F_k(\mathrm{e}^z) < \mathrm{P}_0\{L_q > \beta\sqrt{q} - \ln\lambda\}\mathrm{e}^{-\mathrm{e}^{\beta\sqrt{q}}}.$$

For the integral on $|z| \leq \beta\sqrt{q}$, we use Lemma 2:

$$\begin{aligned}
&\int_{-\beta\sqrt{q}}^{\beta\sqrt{q}} \mathrm{P}_0\{L_q > z - \ln\lambda\}\,\mathrm{d}F_k(\mathrm{e}^z) \\
&\quad \sim \exp\Bigl[\sum_{i=1}^{q} \psi_i(u)\Bigr] \frac{\lambda^u}{u\sqrt{q}} \int_{-\beta\sqrt{q}}^{\beta\sqrt{q}} \mathrm{e}^{-uz} \phi_{0,\sigma_q^2}\Bigl(\frac{z - \ln\lambda}{\sqrt{q}}\Bigr)\,\mathrm{d}F_k(\mathrm{e}^z) \\
&\quad \sim \exp\Bigl[\sum_{i=1}^{q} \psi_i(u)\Bigr] \frac{\lambda^u}{u\sqrt{q}} \int_{-\infty}^{\infty} \mathrm{e}^{-uz} \phi_{0,\sigma_q^2}\Bigl(\frac{z - \ln\lambda}{\sqrt{q}}\Bigr)\,\mathrm{d}F_k(\mathrm{e}^z) \\
&\quad = \exp\Bigl[\sum_{i=1}^{q} \psi_i(u)\Bigr] \frac{\lambda^u}{u\sqrt{q}} \int_{0}^{\infty} s^{-u} \phi_{0,\sigma_q^2}\Bigl(\frac{\ln s - \ln\lambda}{\sqrt{q}}\Bigr)\,\mathrm{d}F_k(s).
\end{aligned}$$

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

Thus, we can rewrite (16) as

$$\mathrm{P}_0\{L_q > \eta_k - \ln\lambda\} \sim \exp\left[\sum_{i=1}^{q}\psi_i(u)\right]\frac{\lambda^u}{u\sqrt{q}}\phi_{0,\sigma_q^2}(0)\frac{\Gamma(k-u)}{\Gamma(k)},$$

and, hence,

$$\frac{\mathrm{P}_0\{L_q > \eta_k\}}{\mathrm{P}_0\{L_q > \eta_1\}} \sim \frac{\Gamma(k-u)}{\Gamma(1-u)\Gamma(k)}. \tag{17}$$

Taking the difference in k completes the proof.

Since the average $\sum_{i=1}^{q}\psi_i(u)/q$ converges to a negative number, the first statement of Theorem 2 implies that the number of different outcomes in a sample is asymptotically $o(N)$ as $N \to \infty$: only a negligible portion of possible different opinions will be seen in a sample. The second statement implies that, no matter how large is the 'rate per cell' λ, as soon as it is fixed, the limit of the ratios $\mu_n(k)/\mu_n$ does not depend on it.

In the next section we present further corollaries and discussion of Theorems 1 and 2.

4. Some corollaries, an inverse problem, and Good–Turing indices

We start by noting how the statements of Theorems 1 and 2 are inter-related. It is, of course, the case that in studying the asymptotic behaviour of the tail of the distribution of L_q when z and q increase simultaneously we cannot use the sequential limit, first for $q \to \infty$ and then $z \to \infty$. However, as the corollary below shows, if we consider the limit of the ratio in (7) then as the distributions P become 'less and less' contiguous to P_0, the sequential limit does agree with (15) in a very natural way.

Corollary 1. *If $c \to \infty$ then*

$$\frac{\int \pi(k, \lambda e^z)\Phi_{-c^2/2,c^2}(\mathrm{d}z)}{\int[1-\pi(0,\lambda e^z)]\Phi_{-c^2/2,c^2}(\mathrm{d}z)} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}\bigg|_{u=1/2}.$$

Proof. In (16) replace $\mathrm{P}_0\{L_q > z - \ln\lambda\}$ directly by the tail of the normal distribution function and use its asymptotics for $c \to \infty$:

$$1 - \Phi_{-c^2/2,c^2}(z - \ln\lambda) = 1 - \Phi_{0,1}\left(\frac{z-\ln\lambda}{c} + \frac{1}{2}c\right) \sim \lambda e^{-z}\left[1 - \Phi_{0,1}\left(\frac{1}{2}c\right)\right].$$

Taking the integral produces the result.

Next we consider the following question. If in a sample of size n_0 there are μ_{n_0} different opinions in a sample, how many more will we expect to see if the sample size is increased to n? The following corollary supplies the answer.

Corollary 2. *For sample sizes $n_0 = \lambda_0 N$ and $n = \lambda N$, as $N \to \infty$,*

$$\frac{\mu_n}{\mu_{n_0}} \sim \left(\frac{n}{n_0}\right)^u.$$

Proof. The assertion follows from the first display in (15).

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

To carry this question further, consider a testing problem for a system of loosely independent 'on/off' components. In real systems of this type there will be a large number of states of overall small probability when the system will fail, while in other states it will continue to function. If the composition of the system is unknown or complex, it is important, during trials of the system, that we see as many different states as possible, or, at least, a 'significant proportion' of all possible states. Under the conditions of Theorem 2, this will not happen. Therefore, it is necessary to either design a testing procedure with probabilities a_i close to $\frac{1}{2}$, and then the number of trials of order 2^q will be sufficient (cf. Theorem 1), or use some appropriately large rate λ for large q. The statement below specifies the rate of such λ when the probabilities of the 'on'-position of the components are arbitrary, apart from satisfying the conditions in (8).

Corollary 3. *Suppose that the conditions in (8) are satisfied. If*

$$\lambda = \lambda_q \gtrsim \mathrm{e}^{m_0 q + b\sqrt{q}},$$

where

$$m_0 = -\frac{1}{2}\int_0^1 \ln[4a(1-a)]\,\mathrm{d}F(a)$$

and b is a constant, then

$$\frac{\mathrm{E}\,\mu_n}{2^q} \sim 1 - \Phi_{0,\sigma_0^2}(-b) \quad \textit{and, for every fixed } k \geq 1, \quad \frac{\mathrm{E}\,\mu_n(k)}{\mathrm{E}\,\mu_n} \to 0. \tag{18}$$

Proof. Again, we can rewrite $\mathrm{P}_0\{L_q > z - \ln\lambda_q\}$ in (16) as

$$\mathrm{P}_0\left\{\frac{L_q + qm_0}{\sqrt{q}} > \frac{z}{\sqrt{q}} - b\right\} \to 1 - \Phi_{0,\sigma_0^2}(-b).$$

Therefore, $\mathrm{P}_0\{L_q > \eta_1 - \ln\lambda\} \to 1 - \Phi_{0,\sigma_0^2}(-b)$ and

$$\frac{\mathrm{P}_0\{L_q > \eta_k - \lambda\}}{\mathrm{P}_0\{L_q > \eta_1 - \lambda\}} \to 1,$$

which proves both statements in (18).

We now consider a converse to the question considered in the previous section: given that statistics $\mu_n(k)$, $k = 1, 2, \ldots$, and μ_n agree with the Karlin–Rouault law, what can be said about the overall behaviour of the underlying probabilities $p(\boldsymbol{x})$, $\boldsymbol{x} \in \Xi_q$?

In a sense, a complete answer can be formulated as follows.

Theorem 3. *If, for any fixed $k = 1, 2, \ldots$, $\mu_n(k)/\mu_n \to R_u(k)$ then, for every $z > 0$,*

$$\frac{N}{\mathrm{E}\,\mu_n} H_n(z) \to (\lambda z)^{-u}.$$

We omit proof of this statement, noting however that assumption (4) is not needed here. Klaassen and Mnatsakanov [15] studied the problem of convergence of the normalized H_n as part of a general inverse problem.

Given the tradition of how this inverse question has been studied in the literature, we consider it in more detail from a somewhat different angle. Recall first (cf. the introduction) that it may seem reasonable to think that the best we can do is to rely on the vector of relative frequencies

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

$\nu_n(\boldsymbol{x})/n$, $\boldsymbol{x} \in \Xi_q$, as an estimator of the vector of probabilities $p_q(\boldsymbol{x})$, $\boldsymbol{x} \in \Xi_q$. However, this would not be satisfactory: first, this would imply the estimate 0 for the overall probability of a very large number of outcomes that did not occur in the sample; and second, as highlighted in the next corollary, the two vectors $\nu_n(\boldsymbol{x})/n$ and $p_q(\boldsymbol{x})$ differ in their overall asymptotic behaviours.

Corollary 4. *For the functions $\widehat{H}_n(z)$ and $H_n(z)$ defined in (1) and (3), with $k = [z] + 1$,*

$$\mathrm{E}\,\widehat{H}_n(z) \sim \frac{\mathrm{E}\,\mu_n}{N} \frac{\Gamma(k-u)/\Gamma(k)}{\Gamma(1-u)}$$

and

$$H_n(z) \sim \frac{\mathrm{E}\,\mu_n}{N} \frac{(\lambda z)^{-u}}{\Gamma(1-u)}.$$

Proof. Note that

$$\sum_{\boldsymbol{x} \in \Xi_q} \mathbf{1}_{\{\nu_n(\boldsymbol{x}) \geq z\}} = \sum_{j=[z]+1}^{\infty} \mu_n(j).$$

Then the proof of the first relation is included around (17) in the proof of Theorem 2. For the second, since $H_n(z) = \mathrm{P}_0\{L_q \geq \mathrm{e}^z - \lambda\}$, its proof follows from Lemma 2.

A well-known way of making inferences about probabilities p_q is to consider the so-called Good–Turing indices. Good [6], referring to A. Turing, introduced the quantities

$$G_n(k) = \sum_{\boldsymbol{x} \in \Xi_q} p(\boldsymbol{x})\, \mathbf{1}_{\{\nu_n(\boldsymbol{x}) = k\}}$$

and

$$p_n(k) = \frac{G_n(k)}{\mu_n(k)}.$$

The intuitive meaning of these quantities is both appealing and transparent: $G_n(k)$ is the total probability of outcomes (in our case, 'opinions') that happen to appear k times in a sample, while $p_n(k)$ is an 'average' or 'typical' probability of each of these outcomes. The definitions extend to $k = 0$, in which case $G_n(0)$ is the total probability of outcomes that do not appear in the sample, while $p_n(0)$ is an 'average' probability of any such outcome.

Based on the simple equality

$$\mathrm{E}\,G_n(k) = \frac{k+1}{n}\,\mathrm{E}\,\mu_n(k+1),$$

Good [6] proposed the estimation of $G_n(k)$ and $p_n(k)$ by

$$\widehat{G}_n(k) = \frac{k+1}{n}\mu_n(k+1)$$

and

$$\widehat{p}_n(k) = \frac{k+1}{n}\frac{\mu_n(k+1)}{\mu_n(k)},$$

respectively. Since then, several authors have investigated the statistical properties of these estimators (e.g. their rate of convergence was studied recently in [19]).

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

Notwithstanding the importance of this work, note that Theorems 1 and 2 imply that, for a sample which agrees either with (7) or with the Karlin–Rouault law, there is no need to use any estimator. In particular, for the latter case, we have the following statement.

Corollary 5. *If* $\mu_n(k)/\mu_n \to R_u(k)$ *for* $k = 1, 2, \dots$ *then*

$$\mathrm{E}\, G_n(k) \sim \frac{u\Gamma(k+1-u)}{\Gamma(1-u)\Gamma(k+1)} \frac{\mathrm{E}\,\mu_n}{n} \quad \textit{and} \quad p_q(k) \sim \frac{k-u}{n}, \tag{19}$$

and, for $k = 0$,

$$\mathrm{E}\, G_n(0) \sim \frac{u}{n} \mathrm{E}\,\mu_n \quad \textit{and} \quad \mathrm{E}\, p_q(0) \sim \frac{u}{n} \frac{\mathrm{E}\,\mu_n}{2^q - \mathrm{E}\,\mu_n}. \tag{20}$$

Orlitzky *et al.* [23] recalled that Laplace [18] suggested the use of the quantities

$$\widetilde{p}_n(k) = \frac{k+1}{n+\mu_n+1}, \qquad k = 1, 2, \dots,$$

which leads to the value

$$\widetilde{G}_n(0) = \frac{1}{n+\mu_n+1}$$

for the total probability of unseen outcomes. Corollary 5 shows that if the sample agrees with the Karlin–Rouault law then the estimation of the total probability of unseen outcomes is more optimistic (small but infinitely larger) than the value $\widetilde{G}_n(0)$ suggested by Laplace.

In conclusion, we remark that the approach used in this paper depends not so much on the form of the probabilities $p_q(\boldsymbol{x})$ or where they are defined, but rather on the asymptotic properties of likelihood ratios. It may, therefore, be applicable to occupancy problems in other situations.

5. Numerical behaviour of the asymptotic formulae for moderate q

As mentioned in the introduction, for the systems of q 'on/off' components, it is possible that q will be of the order of several hundreds. However, in the context of questionnaires or classifications, the number of questions or the number of classifying parameters q will rarely be larger than several tens. For this reason, we would prefer to stay within the case of not very large q and consider how good the asymptotic expressions above work for q between only 10 and 20.

Stability of u_q. The arg min, defined in Lemma 1, is surprisingly stable numerically—not only for the sum $\sum_{i=1}^{q} \psi_i(u)$, but even for one single summand $\psi_i(u)$. For a_i changing in the interval $[0.55, 0.90]$, and by symmetry, in the interval $[0.1, 0.45]$, the value of u, where $\psi_i(u)$ attains its minimum, changes only in the interval $[0.46, 0.50]$. If we choose a_i uniformly distributed on $[0, 1]$ and $q = 10$, when values considerably larger than 0.9 (or smaller than 0.1) can easily occur, the mean value of u_q turned out to be 0.442 with the standard deviation of only 0.024. For values of a_i closer to 0.5, $\psi_i(u)$, as a function in u, becomes quite 'flat' and, therefore, its arg min will be more volatile. However, in this case its exact value will not matter much.

Convergence of $\mu_q(k)/\mu_q$. The plots in Figures 1 and 2 show that this convergence, although not too quick, is reasonable. The bundle of plots of the ratio $\mu_q(k)/\mu_q$, $k = 1, \dots, 10$, for $q = 10$ uniformly distributed probabilities a_i along with the limiting expression is given in Figure 1. For $q = 20$, Figure 2 shows closer approximations and much smaller spread in the bundle of trajectories of $\mu_q(k)/\mu_q$.

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

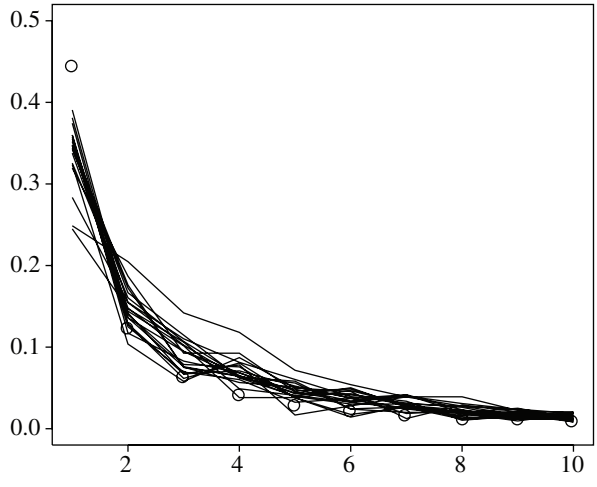

FIGURE 1: The bundle of trajectories of $\mu_q(k)/\mu_q$ for $k = 1, \ldots, 10$. The number of questions $q = 10$ and probabilities $a_1, \ldots, a_{10}$ are uniformly distributed on [0,1]. Open circles indicate the limits of $\mu_q(k)/\mu_q$.

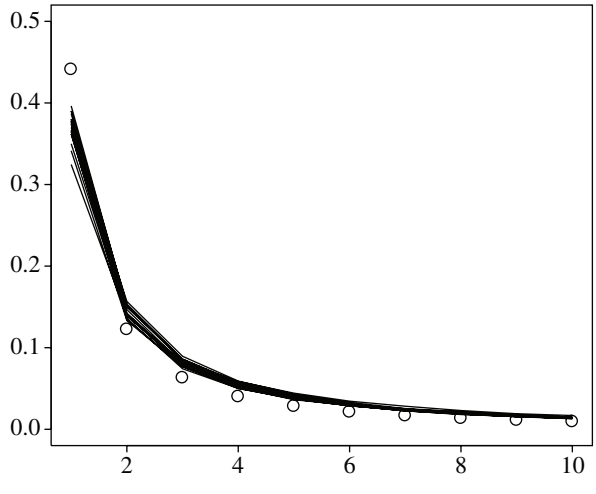

FIGURE 2: The bundle of trajectories of $\mu_q(k)/\mu_q$ for $k = 1, \ldots, 10$. The number of questions $q = 20$ and probabilities $a_1, \ldots, a_{20}$ are uniformly distributed on [0,1].

Transition from the contiguity case to the Karlin–Rouault law. It is interesting to see which limiting values c of the Hellinger distance correspond to the contiguity case, and which ones would already correspond to the large deviations. The plots in Figure 3 show the ratio of integrals (7) for three values $c = 1, 3, 6$. For $c = 1$, the distance (uniform and in the total variation) between $\Phi_{-c^2/2,c^2}$ and $\Phi_{c^2/2,c^2}$ is equal to only 0.3829, while, for $c = 6$, it is equal to 0.9973, so the latter case can be thought of as the case of 'large deviations'. The corresponding, uppermost at $k = 1$, graph in Figure 3 is quite close to the limit, while the graph for $c = 1$ is very far from it.

https://doi.org/10.1239/jap/1324046021 Published online by Cambridge University Press

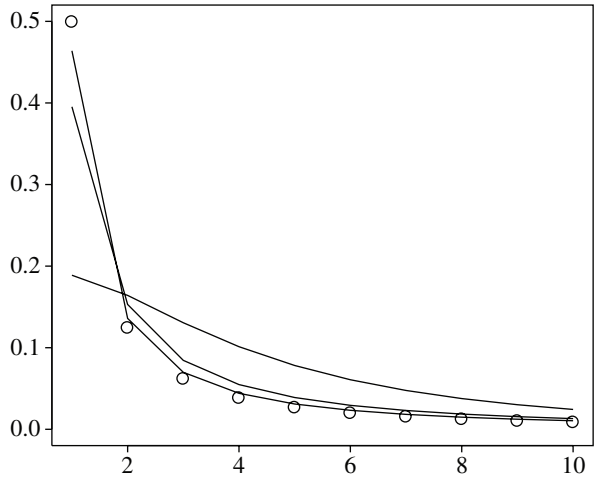

FIGURE 3: Plots of the ratio (7) in k for $c = 1, 3, 6$. Here $\lambda = 5$.

Acknowledgements

I am grateful to Professor D. Daley for interesting and fruitful discussions and his many comments that led to an improvement of the text. I am also grateful to the anonymous referee for his/her positive approach.

References

[1] BAAYEN, R. H. (2002). *Word Frequency Distribution*. Kluwer, Dordrecht.
[2] BAHADUR, R. R. AND RANGA RAO, R. (1960). On deviations of the sample mean. *Ann. Math. Statist.* **31,** 1015–1027.
[3] BARBOUR, A. D. AND GNEDIN, A. V. (2009). Small counts in the infinite occupancy scheme. *Electron. J. Prob.* **14,** 365–384.
[4] CHAGANTY, N. R. AND SETHURAMAN, J. (1993). Strong large deviation and local limit theorems. *Ann. Prob.* **21,** 1671–1690.
[5] FELLER, W. (1986). *Introduction to Probability Theory*, Vol. 2. John Wiley, New York.
[6] GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrica* **40,** 237–264.
[7] GREENWOOD, P. E. AND SHIRYAEV, A. N. (1985). *Contiguity and the Statistical Invariance Principle*. Gordon and Breach, New York.
[8] HWANG, H.-K. AND JANSON, S. (2008). Local limit theorems for finite and infinite urn models. *Ann. Prob.* **36,** 992–1022.
[9] IVANOV, V. A., IVCHENKO, G. I. AND MEDVEDEV, Y. I. (1985). Discrete problems of probability theory (a survey). *J. Soviet Math.* **31,** 2759–2795.
[10] KALLENBERG, O. (1997). *Foundations of Modern Probability*. Springer, New York.
[11] KHMALADZE, È. V. (1983). Martingale limit theorems for divisible statistics. *Theory Prob. Appl.* **28,** 530–549.
[12] KHMALADZE, È. V. (1988). The statistical analysis of a large number of rare events. Tech. Rep. MS-R8804, CWI, Amsterdam.
[13] KHMALADZE, È. V. (2002). Zipf's law. In *Encyclopaedia of Mathematics, Supplement III*, Kluwer, Dordrecht.
[14] KHMALADZE, È. V. AND TSIGROSHVILI, Z. P. (1993). On polynomial distributions with a large number of rare events. *Math. Meth. Statist.* **2,** 240–247.
[15] KLAASSEN, C. A. J. AND MNATSAKANOV, R. M. (2000). Consistent estimation of the structural distribution function. *Scand. J. Statist.* **27,** 733–746.
[16] KOLASSA, J. E. (1994). *Series Approximation Methods in Statistics* (Lecture Notes Statist. **88**), Springer, New York.
[17] KOLCHIN, V. F., SEVASTYANOV, B. A. AND CHISTYAKOV, V. P. (1978). *Random Allocations*. Halsted Press, New York.

[18] Laplace, P.-S. (1995). *Philosophical Essays on Probability* (translation of 5th (1825) French edn.). Springer, New York.
[19] McAllester, D. and Schapire, R. E. (2000). On the convergence rate of Good–Turing estimators. In *Proc. COLT 2000*, pp. 1–6.
[20] Mirakhmedov, S. M. (2007). Asymptotic normality associated with generalized occupancy problem. *Statist. Prob. Lett.* **77,** 1549–1558 .
[21] Mnatsakanov, R. M. (1986). Functional limit theorem for additively separable statistics in the case of very rare events. *Theory Prob. Appl.* **30,** 622–631.
[22] Oosterhoff, J. and van Zwet, W. R. (1979). A note on contiguity and Hellinger distance. In *Contributions to Statistics*, ed. J. Jurechkova, Reidel, Dordrecht, pp. 157–166.
[23] Orlitsky, A., Santhanam, N. P. and Zhang, J. (2003). Always good Turing: asymptotically optimal probability estimation. *Science* **302,** 427–431.
[24] Rouault, A. (1978). Loi de Zipf et sources markoviennes. *Ann. Inst. H. Poincaré* **14,** 169–188.