Diversity of responses in questionnaires and similar objects

Estáte V. Khmaladze
Victoria University of Wellington

July 28, 2009

Abstract. Consider a questionnaire with q $\{0,1\}$ questions, which is filled by N individuals, thus providing N "opinions". Probabilities of the answer 1 to each question can be more or less arbitrary. Out of 2^q, how many different opinions, μ_q, would one expect to see in the sample? How many of these opinions, $\mu_q(k)$, will occur exactly k times?

The paper gives asymptotic expression for $\mu_q/2^q$ and the limit for the ratios $\mu_q(k)/\mu_q$, when the number of questions q increases along with the sample size N so that $N = \lambda 2^q, \lambda = const.$

In the context of questionnaires, the q is often not too big, and we show how the asymptotic expressions work numerically for q of order 10-30.

Running title. Diversity of questionnaires.
AMS 2000 subject classifications. 62D05, 62E20, 60E05, 60F10.
Key words and phrases. Number of unique outcomes, sparse tables, systems with large number of components, Karlin-Rouault law, Zipf's law, Good-Turing indices, large deviations, contiguity.

1 Formulation of the problem

Suppose a person is asked to fill in a form with q binary (yes/no) questions. Obviously, there are 2^q possible ways to fill such form. Suppose $N = \lambda 2^q$

persons were asked to fill it in, so that we have a sample of size N of all possible responses. We will be interested in the diversity of responses in this sample when $q \to \infty$ and $\lambda = const$.

Suppose this fixed λ, a "rate per cell", equals, say, 10, so that we have asked 10 times more persons than there are possible ways to fill the form. How many different responses will we see in the sample? And how many of them will we see only once, or twice, or any fixed k number of times? The first impression well may be that we will see basically all possible responses, some of them, say, 3 times and some of them, say, 16 times. More careful intuition may suggest that some outcomes may occur 30 times and some will not occur at all, but at least the number of different outcomes in the sample will be of the same order of magnitude as 2^q and that the fraction of unique responses (with frequency 1) will not be that big.

Although the latter situation may indeed occur, in a majority of cases the number of different outcomes in the sample will be much less than 2^q, and the number of unique outcomes will constitute not small fraction of it, but often close to a half. In general, in the total number of different observed outcomes the fraction of outcomes observed exactly k times will tend to certain limits, which we show below. Strangely enough, these limits do not depend on λ.

Let us formalize the problem. Suppose $\overrightarrow{\xi} = (\xi_1, \xi_2, \ldots, \xi_q)$ is a vector of q Bernoulli random variables. The set Ξ_q of its possible values is the set of all sequences $\overrightarrow{x} = (0, 1, 0, \ldots, 1)$ of length q consisting of 0-s and 1-s. Suppose we have now a sample of such $\overrightarrow{\xi}$, i.e. a sequence of N i.i.d. vectors $\overrightarrow{\xi}_1, \ldots, \overrightarrow{\xi}_N$. Each of 2^q possible values will have then its frequency

$$f_q(\overrightarrow{x}) = \sum_{j=1}^{N} \mathbb{I}_{\{\overrightarrow{\xi}_j = \overrightarrow{x}\}}$$

in this sample. Let

$$\mu_q = \sum_{\overrightarrow{x} \in \Xi_q} \mathbb{I}_{\{f_q(\overrightarrow{x}) \geq 1\}}, \quad \mu_q(k) = \sum_{\overrightarrow{x} \in \Xi_q} \mathbb{I}_{\{f_q(\overrightarrow{x}) = k\}}, \quad k = 1, 2, \ldots.$$

Then we are interested in the asymptotic behaviour of the quotients

$$\frac{\mu_q}{2^q} \quad \text{and} \quad \frac{\mu_q(k)}{\mu_q}, \tag{1}$$

when $q \to \infty$ and $N \to \infty$. In the situations, where sample size N is much smaller than the number of possible values 2^q, it is obvious that most frequencies will be 0. On the other hand, the case with N much larger than 2^q and q increasing can have a matching situation in practice only very rarely. So, we assume here that N is of the order 2^q, that is, $N \sim \lambda 2^q, \lambda = const, q \to \infty$.

Usually, rare and infrequent objects or outcomes are placed in a group called "others". From some point of view, these objects may seem unimportant. But as soon as we ask questions about diversity, that is, "how many different objects do we have", or "how many different species do we observe", these rare objects become most important. At the same time, they may look awkward to study from statistical point of view, as their frequencies, even in large samples, remain small and, therefore, unstable.

The problems where we meet sequences of 0 and 1 of increasing length occur in many situations. The case of large opinion pools with relatively long sequence of $\{0, 1\}$ questions we mentioned already. In this context we would mean to prove that the saying "as many men as many minds" is mostly incorrect: typically, the number of "minds" will be much smaller than the number of "men". It is no less common a situation in classification problems - for example, in the medical diagnostic problems, where they check presence of absence of q symptoms in each patient. Then we ask, how many different cases should they expect to see in the large data-bases, if q is also large? Similar situation with the so called "sparse tables" is very common in, say, economic statistics: if companies are classified using q-dimensional parameter, then many of possible "sells" will remain empty, and our question is how many of the cells will have any object, or a given number of objects, in them? In the cases like this, cells with only one object are often of special interest, as the corresponding objects will be unique and, therefore, easily indentifiable. How many of them should occur "naturally", then?

Different class of problems, where potentially q can be very large, is mentioned in the beginning of Sec. 4. For still another class of problems, studied in mathematical methods of taxonomy, we refer the reader to Gyllenberg and Koski (1996) and Gyllenberg and Koski (2001) and references therein.

The questions we ask here were studied in Khmaladze and Tsigroshvili (1993) in the context, close to questionnaires and to the so-called "McArthur's stick"

(see McArthur(1957)). Namely, suppose we fix some positive $a < 1/2$ and break the interval $[0,1]$ in proportion $a : 1-a$. At the second step each of the resulting two intervals we break into two in the same proportion and repeat this steps q times. Consider the lengths of the 2^q resulting intervals as probabilities and generate a sample of size N from this distribution. This is equivalent to generation of N i.i.d. $\overrightarrow{\xi}$, when the coordinates of $\overrightarrow{\xi}$ are independent Bernoulli random variables with the same probability $\mathbb{P}(\xi_i = 1) = a$. Then Khmaladze and Tsigroshvili (1993) considered the questions: will this sample behave in the way similar to the situation where we would have broken $[0,1]$ into 2^q subintervals using spacings formed by $2^q - 1$ uniformly distributed points, as in McArthur(1957)? The authors showed that for any $a \neq 1/2$ the former situation is very different form the latter and obtained the asymptotics for the quotients (1). In particular, unlike McArthur's stick, for any $a \neq 1/2$ the ratio $\mu_q/2^q$ converges to 0.

In this paper we obtain more general result with different tools to prove it. Although we keep the assumption of independence of the coordinates $\xi_1, \ldots, \xi_q$ of each $\overrightarrow{\xi}$, otherwise their distribution can be arbitrary: $\mathbb{P}(\xi_i = 1) = a_i$, possibly different for different i. We show that, basically, there are two possibilities. Namely, if the probabilities $a_1, \ldots, a_q$ form, in q, a triangular array converging to $1/2$ so that the distributions, given on Ξ_q by

$$p_q(\overrightarrow{x}) = \Pi_{i=1}^q a_i^{x_i}(1-a_i)^{1-x_i}$$

form a sequence, contiguous to the uniform distributions, given by $p_0(\overrightarrow{x}) = (1/2)^q$, then $E\mu_q/2^q$ has the positive limit – see Theorem 1, and see the form of the limit of $E\mu_q(k)/E\mu_q$ in (5). If $a_1, \ldots, a_q$ are fixed, that is, they form just a sequence in q, then $E\mu_q/2^q \to 0$, and we show the rate. The quotients $E\mu_q(k)/E\mu_q$ still converge to limits, which we describe in Theorem 2.

There is a long tradition of probabilistic research, of which the asymptotic behaviour of $E\mu_q(k)/E\mu_q$ is one of the main objects. Perhaps the most known key word for this research is "Zipf's Law". This law was observed in very large number of different situations. In particular, it is often considered in the statistical analysis of texts – the state of art of this research is represented by the monograph Baayen(2000). The quotients $E\mu_q(k)/E\mu_q$ are said to follow Zipf's law if they would tend to $1/k(k+1)$. However, as Theorem 2 shows, in the situation we consider they converge to a different

"law". The expression they converge to is known in the literature under the name of Karlin-Rouault Law (Rouault(1978) obtained this expression in the situation very different from ours – for frequencies of different events in a Markov chain).

Initially, asymptotic analysis of the integrals $E\mu_q(k)$ and $E\mu_q$ looked not so simple. It was a certain challenge to find relatively transparent and purely probabilistic proofs.

In sections 2 and 3 we prove our main limit theorems. In section 4 we consider the "reverse" question: if Karlin-Rouault law for $\mu_q(k)/\mu_q$ is observed, what can be said about underlying probabilities? - how many how small probabilities there were? Exact answer to this question is given in Theorem 4. We will see, in passing, that the usual estimates $\hat{p}(\overrightarrow{x}) = f_q(\overrightarrow{x})/N$ for "most" $\overrightarrow{x}$ are not good. We also show implications of Theorems 1 and 2 for Good-Turing indices.

2 The approach and the case of contiguity.

The key step in asymptotic analysis of the ratio $\mu_q/2^q$ and of $\mu_q(k)/\mu_q$ consists of the analysis of $E\mu_q/2^q$ and $E\mu_q(k)/E\mu_q$, because the ratios $\mu_q/E\mu_q$ and $\mu_q(k)/E\mu_q(k)$ converge to 1 a.s. Therefore we concern ourselves with the expressions for the expected values.

Denote $b(k, N, p)$ binomial probability of k with parameters N and p. Since each frequency $f_q(\overrightarrow{x})$ has the distribution $b(\cdot, N, p(\overrightarrow{x}))$ we have

$$E\mu_q = \sum_{\overrightarrow{x} \in \Xi_q} (1 - b(0, N, p(\overrightarrow{x})))$$

and

$$E\mu_q(k) = \sum_{\overrightarrow{x} \in \Xi_q} b(k, N, p(\overrightarrow{x})), k = 1, 2, \ldots.$$

It is clear that, as $q \to \infty$, all probabilities $p(\overrightarrow{x}) \to 0$. Therefore, it is natural to expect that binomial probabilities in above can be replaced by the Poisson probabilities. As this step is only of a technical importance for us, we use

the easiest way to justify this replacement – we assume that the sample size N is Poisson random variable with expected value $\lambda 2^q$. As we know, then all $f_q(\overrightarrow{x})$ become independent Poisson random variables (or Poisson processes in q) with parameters $\lambda 2^q p(\overrightarrow{x})$). Consequently, we can use the expressions

$$E\mu_q = \sum_{\overrightarrow{x} \in \Xi_q} (1 - \pi(0, \lambda 2^q p(\overrightarrow{x})))$$

and

$$E\mu_q(k) = \sum_{\overrightarrow{x} \in \Xi_q} \pi(k, \lambda 2^q p(\overrightarrow{x})), \quad k = 1, 2, \ldots,$$

where $\pi(k, z)$ denotes Poisson probability, of k, with parameter z.

One can study the sums above as they are. For example, if all a_i were equal $1/2$, i.e. if all $p_q(\overrightarrow{x}) = 1/2^q$ we immediately obtain that

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

This, in particular, implies that one should expect the number of different opinions in a sample to be of the same order as the number of possible opinions.

In general situation, the asymptotic analysis of the sums $E\mu_q$ and $E\mu_q(k)$ looked for this author a challenging problem. However, some, technically insignificant, change in the point of view on the quantity $2^q p(\overrightarrow{x})$ transforms the situation and makes it possible to use powerful probabilistic tools, which otherwise would seem irrelevant and distant to the problem.

Namely, denote $\mathbb{P}_q$ and $\mathbb{P}_{0q}$ the distributions on Ξ_q defined by $p_q(\overrightarrow{x})$ and $p_0(\overrightarrow{x}) = 1/2^q$ respectively. Then

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

is the likelihood ratio of $\mathbb{P}_q$ and $\mathbb{P}_{0q}$ and we can write

$$\begin{aligned} E\mu_q &= 2^q E_0(1 - \pi(0, \lambda M_q(\overrightarrow{\xi}))), \\ E\mu_q(k) &= 2^q E_0 \pi(k, \lambda M_q(\overrightarrow{\xi}))), k = 1, 2, \ldots, \end{aligned} \tag{3}$$

where E_0 denotes expected value calculated with respect to the uniform distribution $\mathbb{P}_{0q}$ of $\overrightarrow{\xi}$.

We rewrite the previous display as

$$\begin{aligned} \frac{E\mu_q}{2^q} &= E_0(1-\pi(0,\lambda M_q(\overrightarrow{\xi}))), \\ \frac{E\mu_q(k)}{E\mu_q} &= \frac{E_0\pi(k,\lambda M_q(\overrightarrow{\xi}))}{E_0(1-\pi(0,\lambda M_q(\overrightarrow{\xi})))},\ k=1,2,\dots, \end{aligned} \tag{4}$$

and in what follows will consider asymptotic behaviour of the expectations on the right hand side.

As an immediate consequence of this point of view we obtain the following statement: if the sequence of distributions $\mathbb{P}_q$ is contiguous with respect to the sequence of uniform distributions $\mathbb{P}_{0q}$, then M_q typically converges in distribution, under $\mathbb{P}_{0q}$, to a random variable e^L, where $L \sim \mathcal{N}(-c^2/2, c^2)$ and c^2 is specified below. But since $\pi(k,\lambda e^z)$, for each k, are bounded and continuous functions in z, the expected values in (4) converge to the corresponding limits. Theorem below recalls the formal conditions and specifies c^2. In this theorem and everywhere below, $\Phi_{\mu,\sigma^2}(z)$ and $\phi_{\mu,\sigma^2}(z)$ denote normal distribution function and normal density, respectively, with mean μ and variance σ^2.

Theorem 1. *Suppose probabilities $a_{1q},\dots,a_{qq}$ form in q a triangular array, such that $\max_i |a_{iq}-1/2| \to 0$ and*

$$a_{iq} = \frac{1}{2} + \frac{c_{iq}}{\sqrt{q}}, \quad \textit{with} \quad \limsup_{q\to\infty} \sum_{i=1}^{q} c_{iq}^2/q < \infty.$$

Then

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

If the finite limit

$$\lim_{q\to\infty} \sum_{i=1}^{q} c_{iq}^2/q = c^2$$

exist, then

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

and

$$\frac{E\mu_q(k)}{E\mu_q} = \frac{\int \pi(k,\lambda e^z)\Phi_{-c^2/2,c^2}(dz)}{\int (1-\pi(0,\lambda e^z))\Phi_{-c^2/2,c^2}(dz)}\,. \tag{5}$$

Proof. The condition on the upper limit of $\sum_{i=1}^q [a_{iq}-1/2]^2 = \sum_{i=1}^q c_{iq}^2/q$ guaranties contiguity of the sequence of distributions, given by $p_q(\overrightarrow{x})$, to the sequence of uniform distributions given by $p_0(\overrightarrow{x})$. In its turn, the contiguity implies that the sequence of distributions of M_q is weakly compact. Hence the result on $E_0\mu_q$ follows. Existence of the limit, together with the condition on $\max_i |a_{iq}-1/2|$ guaranties asymptotic normality of the log-likelihood ratio $\ln M_q$ (see, e.g., Oosterhof and van Zwet (1979)) with parameters, under the null distribution, equal $-c^2/2$ and c^2. This asymptotic normality implies convergence of the expected values as the integrands are continuous and bounded functions of $\ln M_q$. □

Note that the result extends to very general class of distributions. Namely, whether $\xi_1,\ldots,\xi_q$ are independent and $p_q(\overrightarrow{x})$ is a product of Bernoulli distributions or not does not matter much. 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 stands for 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 Greenwood, Shiryayev (1985)). Therefore, the statement of the theorem remains still true.

3 The case of arbitrary a_i-s. Large deviations.

Suppose now that probabilities $a_1, a_2, \ldots, a_q$ are arbitrary, that is, they form some sequence in q. In this case the behaviour of the likelihood ratio M_q becomes somewhat erratic: under $\mathbb{P}_{0q}$ we have $M_q \to 0$ in probability, but $E_0 M_q = 1$, and therefore increasingly large values of M_q are unavoidable. As

the consequence of this, in asymptotic analysis of the expected value $E_0(1-\pi(0,\lambda M_q))$ one can not rely, for example,on Taylor series approximation like, say,

$$1-\pi(0,\lambda M_q)\sim\lambda M_q,$$

because in mean it is not correct: we will see below that

$$E_0(1-\pi(0,\lambda M_q))\to 0,\quad \text{while}\quad E_0\lambda M_q=\lambda.$$

How small is the quantity $E_0(1-\pi(0,\lambda M_q))$ is not immediately obvious: for large q, although the random variable M_q is small with high probability, the integrand $1-\pi(0,\lambda M_q)$ also becomes small, while although M_q is large with only small probability, the integrand is close to 1, that is to say, not small.

In a little bit more detail, for small ϵ

$$E_0(1-\pi(0,\lambda M_q))\mathbb{I}_{\{\lambda M_q<\epsilon\}}\approx\lambda E_0 M_q\mathbb{I}_{\{\lambda M_q<\epsilon\}}=\mathbb{P}_q\{\lambda M_q<\epsilon\}$$

becomes, in the terminology of testing statistical hypothesis, almost the probability of the type two error, while, for large n

$$E_0(1-\pi(0,\lambda M_q))\mathbb{I}_{\{\lambda M_q>n\}}\approx E_0\mathbb{I}_{\{\lambda M_q>n\}}=\mathbb{P}_0\{\lambda M_q>n\}$$

becomes, in the same terminology, almost the probability of the type one error. Therefore, with interpretation of M_q as a likelihood ratio, both tails are of comparable size and both could be comparable in size with the "central" part of the expectation $E_0(1-\pi(0,\lambda M_q))$. We did not find it fruitful to try and locate a part where the main contribution to the integral $E_0(1-\pi(0,\lambda M_q))$ is made directly. Instead we suggest to give it a form of a certain probability, which, as we will see, is naturally connected with the theory of large deviations.

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

$$E_0(1-\pi(0,\lambda M_q))=\mathbb{P}_0\{L_q>\eta_1-\ln\lambda\}.$$

Moreover, if T_k is a Gamma-distributed random variable with shape parameter k, that is, if T_k is a sum of k independent copies of T_1, independent from

M_q, and if $\eta_k = \ln T_k$, then in the similar way

$$E_0 \sum_{j=k}^{\infty} \pi(j, \lambda M_q) = \mathbb{P}_0\{L_q > \eta_k - \ln \lambda\}$$

Here and above $\mathbb{P}_0$ denotes, obviously, the joint distribution of L_q, under uniform distribution on Ξ_q, and η_k.

It is clear that

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

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

$$\begin{aligned} \psi_i(u) &= \ln E_0 \exp u[\xi_i \ln 2a_i + (1-\xi_i) \ln 2(1-a_i)] \\ &= \ln[(2a_i)^u + (2(1-a_i))^u] - \ln 2. \end{aligned}$$

As we know, $\psi_i(u)$ is convex, infinitely differentiable function of u and $\psi_i(0) = \psi_i(1) = 0$. Then so is the sum $\sum_{i=1}^{q} \psi_i(u)$, which is logarithm of the moment generating function of L_q.

Consider the sequence $a_1, a_2, ..., a_q$ and denote

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

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

$$\begin{gathered} F_q(a) \to F(a) \ \text{ for all } \ a \in [0,1]\,, \\ \int_0^1 \ln^2 \frac{a}{1-a} dF_q(a) \to \int_0^1 \ln^2 \frac{a}{1-a} dF(a). \end{gathered} \tag{6}$$

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

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

Lemma 1. *Suppose condition* (6) *is satisfied. Let* $u = u_q$ *be such that* $\sum_{i=1}^q \psi_i'(u) = 0$. *Denote*

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

Then,

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

Proof. It is easy to see that condition (6) implies convergence

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

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

$$\sum_{i=1}^q \psi_i'(0)/q \to \int_0^1 (\ln 4a(1-a))dF(a)/2 > -\infty$$

and

$$\sum_{i=1}^q \psi_i'(1)/q \to \int_0^1 (a\ln a + (1-a)\ln(1-a))dF(a) + \ln 2 < \infty\,.$$

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

Essential step in the theorem below is given by the following lemmas.

Lemma 2. *Suppose condition* (6) *is satisfied. Then, with* u *as in Lemma 1,*

$$\mathbb{P}_0\{L_q > z\} \sim e^{\sum_{i=1}^q \psi_i(u) - uz} \frac{1}{u\sqrt{q}} \phi_{0,\sigma_q^2}(z/\sqrt{q})[1 + r_q(z)] \tag{7}$$

with

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

for any fixed $\beta > 0$.

Since

$$E_0 L_q/q \to \int_0^1 (\ln 4a(1-a))dF(a)/2 < 0,$$

and therefore $L_q \to -\infty$, the probability, $\mathbb{P}_0\{L_q > c\}$ is, for any given c, probability of large deviations for L_q. Although the proof basically follows the known pattern (cf., e.g., Kallenberg (2003), ??) the lemma shows the asymptotic expression for this probability and not its logarithm, as is commonly stated in the literature. For the i.i.d. case, the idea can already be seen in Bahadur and Ranga Rao (1960), sec. 5, and for general case it was carried through, with the aid of some assumptions, in Chaganty and Sethuraman (1993). The lemma states, in addition, that the asymptotic expression is correct uniformly in c in increasing intervals of the length $\sqrt{q}$. This is sufficient for the application of (7) in Theorem 2 below, although it could be easily extended to the length $o(q^{3/4})$.

Proof. Consider the adjoint to $\mathbb{P}_0$ distribution $\mathbb{Q}$ defined by the relationship

$$\mathbb{P}_0\{L_q > z\} = e^{\sum_{i=1}^q \psi_i(u)} \int_z^\infty e^{-ux} d\mathbb{Q}(x) \tag{8}$$

One can see that the moment generating function of $\mathbb{Q}$ is

$$\int e^{rt} d\mathbb{Q}(t) = e^{\sum_{i=1}^q \psi_i(u+r)-\psi_i(u)},$$

and therefore, with the choice of u as in the lemma, the expected value of $\mathbb{Q}$ is 0 and the variance is $q\sigma_q^2$. Denote $\mathbb{Q}_{L_q/\sqrt{q}}$ the distribution of $L_q/\sqrt{q}$ under the distribution $\mathbb{Q}$. Then (8) can be re-written as

$$\begin{aligned} \mathbb{P}_0\{L_q > z\} &= e^{\sum_{i=1}^q \psi_i(u)} \int_{z/\sqrt{q}}^\infty e^{-u\sqrt{q}y} d\mathbb{Q}_{L_q/\sqrt{q}}(y) \\ &= e^{\sum_{i=1}^q \psi_i(u)-uz} \int_0^\infty e^{-u\sqrt{q}x} d\mathbb{Q}_{L_q/\sqrt{q}}(x + z/\sqrt{q}), \end{aligned} \tag{9}$$

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

$$\begin{aligned} u\sqrt{q}\int_{z/\sqrt{q}}^{\infty} e^{-u\sqrt{q}y}\phi_{0,\sigma_q^2}(y)dy &= e^{-uz}u\sqrt{q}\int_0^{\infty} e^{-u\sqrt{q}x}\phi_{0,\sigma_q^2}(x+z/\sqrt{q})dx \\ &= e^{-uz}\phi_{0,\sigma_q^2}(z/\sqrt{q})[1+r_q(z)], \end{aligned} \tag{10}$$

where

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

Note, that to obtain non-zero asymptotics we had to normalise the integral above by $\sqrt{q}$. Therefore we have to consider normalised difference

$$u\sqrt{q}\int_{z/\sqrt{q}}^{\infty} e^{-u\sqrt{q}y}(\mathbb{Q}_{L_q/\sqrt{q}}(dy) - \Phi_{0,\sigma_q^2}(dy)).$$

The difference $\sqrt{q}(\mathbb{Q}_{L_q/\sqrt{q}}(y) - \Phi_{0,\sigma_q^2}(y))$ does not have to be a small function. Therefore we need better approximation for $\mathbb{Q}_{L_q/\sqrt{q}}(y)$, which one can obtain in the form of Edgeworth expansion (see the next lemma). According to this expansion

$$\sup_y |\mathbb{Q}_{L_q/\sqrt{q}}(y) - C_q(y)| = o(1/\sqrt{q}), \tag{11}$$

where

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

and where $P(y) = y^3 - 3y$ is third Hermite polynomial. The term $P(y)\phi_{0,\sigma_q^2}(y)/\sqrt{q}$ will not change the asymptotics in (19), while using integration by parts we get

$$\sqrt{q}\int_0^{\infty} e^{-u\sqrt{q}x} d(\mathbb{Q}_{L_q/\sqrt{q}}(x+\frac{z}{\sqrt{q}}) - C_q(x+\frac{z}{\sqrt{q}})) = -\sqrt{q}(\mathbb{Q}_{L_q/\sqrt{q}}(\frac{z}{\sqrt{q}}) - C_q(\frac{z}{\sqrt{q}}))$$

$$+uq\int_0^{\infty} e^{-u\sqrt{q}x}(\mathbb{Q}_{L_q/\sqrt{q}}(x+\frac{z}{\sqrt{q}}) - C_q(x+\frac{z}{\sqrt{q}}))dx \to 0$$

uniformly in z. □

The next lemma shows that the Edgeworth expansion (11) for distribution $\mathbb{Q}_{L_q/\sqrt{q}}(z)$ indeed exists.

Lemma 3. *If* (6) *is satisfied, then there exists Edgeworth expansion for the distribution function* $\mathbb{Q}_{L_q/\sqrt{q}}(z)$.

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

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

is characteristic function of i-th summand of L_q in measure $\mathbb{Q}$. For the proof we need to show (12), because the rest basically follows the lines of the proof for the i.i.d. case as given in Feller (19??), Ch.XVI.2-4. We give only a sketch of it. If $G_q(z)$ is as in the previous lemma and $\gamma_q(t)$ is its Fourier transform, then for arbitrarily small ϵ there is large enough constant b, such that

$$|\mathbb{Q}_{L_q/\sqrt{q}}(z) - G(z)| \leq \int_{-b\sqrt{q}}^{b\sqrt{q}} \frac{|\Pi_{i=1}^q \xi_i(\frac{t}{\sqrt{q}}) - \gamma_q(t)|}{t} dt + \frac{\epsilon}{\sqrt{q}},$$

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

$$\sup_{\delta<|t|/\sqrt{q}<a} |\Pi_{i=1}^q \xi_i(\frac{t}{\sqrt{q}})| < c^q, \quad \text{for some } \ 0 < c < 1. \tag{12}$$

However, for the norm of this characteristic function we have

$$\frac{1}{q} \ln \Pi_{i=1}^q |\xi_i(s)| = \frac{1}{q} \sum_{i=1}^q \ln(1 + 2q_i(1 - q_i)(\cos s\omega_i - 1))$$

$$= \int_0^1 \ln(1 + 2q(a)(1 - q(a))(\cos s\omega(a) - 1))dF_q(a))$$

$$\leq 2 \int_0^1 q(a)(1 - q(a))(\cos s\omega(a) - 1)dF_q(a)$$

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

$$\int_0^1 (1 - \cos s\omega(a))(dF_q(a) - dF(a)) = \int_{-\infty}^{\infty} (1 - \cos s\omega)(dH_q(\omega) - dH(\omega))$$

and integration by parts leads to

$$s|\int_{-\infty}^{\infty}\sin s\omega(H_q(\omega)-H(\omega))d\omega| \le s\int_{-\infty}^{\infty}|H_q(\omega)-H(\omega)|d\omega$$

Conditions (6) imply that the last integral converges to 0, because they guarantee that $H_q(\omega)\to H(\omega)$ uniformly in ω and the second, and, hence, the first absolute moments converge. Obviously, this is true uniformly in $s\in[\delta,b]$. On the other hand, for any continuous distribution

$$\int_{-\infty}^{\infty}\cos s\omega\; dH(\omega) < 1-2\epsilon$$

for $s>\delta$ and therefore (6) is true with $c=1-\epsilon$. □

Note that the form of condition (12) vary in the literature. Often it may seem simpler to require this inequality uniformly for $s>\delta$ (see, e.g., Kolassa (1994), p.34). However, this requirement would be restrictive for us: under (6) it will not be generally true. To see this one can consider F_q which attaches equal weight $1/q$ to regularly spaced points $i/q, j=1,\dots,q$. However, if $a_1,\dots,a_q$ were assumed to be independent random variables, then (12) will be true for $s>\delta$.

Now we are ready to formulate the following theorem.

Theorem 2. *If condition* (6) *is satisfied, then*

$$\begin{aligned}\frac{E\mu_q}{2^q} &\sim e^{\sum_{i=1}^q\psi_i(u_q)}\frac{\lambda^{2u}}{u\sqrt{q}}\phi_{0,\sigma_q^2}(0)\Gamma(1-u),\\ \frac{E\mu_q(k)}{E\mu_q} &\to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)},\end{aligned} \tag{13}$$

where $u=\lim u_q$.

Proof. We start with asymptotic expression for $E_0\sum_{j=k}^{\infty}\mu_q(j)=\mathbb{P}_0\{L_q>\eta_k-\ln\lambda\}$. The result will then follow from (4). Denote F_k gamma distribution function with the shape parameter k (and scale parameter 1). Then $F_k(e^x)$ is the distribution function of η_k. We have

$$\mathbb{P}_0\{L_q>\eta_k-\ln\lambda\}=\int_{-\infty}^{\infty}\mathbb{P}_0\{L_q>z-\ln\lambda\}dF_k(e^z). \tag{14}$$

Let us split the domain of the integration into three parts. Using (9), for the integral over $(-\infty, -\beta\sqrt{q}]$ we have

$$\int_{-\infty}^{-\beta\sqrt{q}} \mathbb{P}_0\{L_q > z - \ln\lambda\} dF_k(e^z) = F_k(e^{-\beta\sqrt{q}})\mathbb{P}_0\{L_q > -\beta\sqrt{q} - \ln\lambda\} + \\ + e^{\sum_{i=1}^q \psi_i(u)} \int_{-\infty}^{-\beta\sqrt{q}} F_k(\lambda e^{\sqrt{q}z}) e^{-u\sqrt{q}z} d\mathbb{Q}_{L_q/\sqrt{q}}(z).$$

Since $F_k(\epsilon) < \epsilon^k/2 < \epsilon/2$ for all sufficiently small ϵ, we obtain

$$\int_{-\infty}^{-\beta\sqrt{q}} \mathbb{P}_0\{L_q > z - \ln\lambda\} dF_k(e^z) < e^{-\beta\sqrt{q}}\, \mathbb{P}_0\{L_q > -\beta\sqrt{q} - \ln\lambda\} + \\ + e^{\sum_{i=1}^q \psi_i(u)} \lambda \int_{-\infty}^{-\beta\sqrt{q}} e^{\sqrt{q}(1-u)z} d\mathbb{Q}_{L_q/\sqrt{q}}(z)$$

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

$$\int_{\beta\sqrt{q}}^{\infty} \mathbb{P}_0\{L_q > z - \ln\lambda\} dF_k(e^z) < \mathbb{P}_0\{L_q > \beta\sqrt{q} - \ln\lambda\} e^{-e^{\beta\sqrt{q}}}.$$

For the middle part we can use Lemma 2:

$$\begin{aligned} \int_{-\beta\sqrt{q}}^{\beta\sqrt{q}} \mathbb{P}_0\{L_q > z - \ln\lambda\} dF_k(e^z) \sim \\ \sim e^{\sum_{i=1}^q \psi_i(u)} \frac{\lambda^u}{u\sqrt{q}} \int_{-\beta\sqrt{q}}^{\beta\sqrt{q}} e^{-uz} \phi_{0,\sigma_q^2}(\frac{z - \ln\lambda}{\sqrt{q}}) dF_k(e^z) \\ \sim e^{\sum_{i=1}^q \psi_i(u)} \frac{\lambda^u}{u\sqrt{q}} \int_{-\infty}^{\infty} e^{-uz} \phi_{0,\sigma_q^2}(\frac{z - \ln\lambda}{\sqrt{q}}) dF_k(e^z) \\ = e^{\sum_{i=1}^q \psi_i(u)} \frac{\lambda^u}{u\sqrt{q}} \int_0^{\infty} s^{-u} \phi_{0,\sigma_q^2}(\frac{\ln s - \ln\lambda}{\sqrt{q}}) dF_k(s) \end{aligned} \tag{15}$$

Therefore, altogether

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

and

$$\frac{\mathbb{P}_0\{L_q > \eta_k\}}{\mathbb{P}_0\{L_q > \eta_1\}} \sim \frac{\Gamma(k-u)}{\Gamma(1-u)\Gamma(k)} \tag{16}$$

Taking the difference in k will end the proof. □

It is interesting to note how the statements of Theorems 1 and 2 are related to each other. It is, of course, not true that in studying asymptotic behaviour of the tail of the distribution of L_q when z and q increase simultaneously we can use sequential limit, first in $q \to \infty$ and then in $z \to \infty$. However, as the corollary below shows, if we consider the limit of the ratio in (5), as the distributions $\mathbb{P}$ become "less and less" contiguous to $\mathbb{P}_0$, it agrees with (13) in a very natural way.

Corollary 3. *If $c \to \infty$, then*

$$\frac{\int \pi(k, \lambda e^z)\Phi_{-c^2/2,c^2}(dz)}{\int (1-\pi(0,\lambda e^z))\Phi_{-c^2/2,c^2}(dz)} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}\Big|_{u=1/2} = 0.282\frac{\Gamma(k-1/2)}{\Gamma(k+1)}.$$

Proof. In equation (14) we can now replace $\mathbb{P}_0\{L_q > z - \ln\lambda\}$ directly by the tail of normal distribution function and use its asymptotics for $c \to \infty$:

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

Taking the integral will then produce the result. □

4 On Good-Turing indices and behavior of underlying probabilities.

The question we consider now is reverse to the question we studied so far and is more of statistical than of probabilistic nature: given the statistics $\mu_q(k), k = 1, 2, \ldots,$ and μ_q agree with Karlin-Rouautl law, what can one say about overall behavior of the underlying probabilities $p(\overrightarrow{x}), \overrightarrow{x} \in \Xi_q$? To answer this question we start with the Good-Turing indices.

In his famous paper Good (1953), with reference to A.Turing, I.J.Good introduced quantities

$$G_q(k) = \sum_{\overrightarrow{x} \in \Xi_q} p(\overrightarrow{x}) \mathbb{I}_{\{f_q(\overrightarrow{x})=k\}} \quad \text{and} \quad p_q(k) = \frac{G_q(k)}{\mu_q(k)}.$$

Intuitive meaning of these quantities is very appealing and transparent: $G_q(k)$ is the total probability of outcomes (in our case - opinions) that happened to appear k times in the sample while $p_q(k)$ is an "average" or "typical" probability of each of such outcomes. The definitions extend to $k = 0$, and $G_q(0)$ is the total probability of outcomes that did not appear in the sample, while $p_q(0)$ is an "average" probability of any such outcome.

Prior to any asymptotic analysis of these quantities it may look quite reasonable to say that since the frequency $f_q(\overrightarrow{x})$ of $\overrightarrow{x}$ is equal to k, its expected value is best estimated by k and, therefore, the sample suggests that there were $\mu_q(k)$ probabilities, each equal

$$\bar{p}_q(k) = \frac{k}{N}.$$

This is, basically, to say that as an estimation of $p(\overrightarrow{x})$ we take $f_q(\overrightarrow{x})/N$. Since $\sum_{k=1}^{\infty} \bar{p}_q(k)\mu_q(k) = 1$, for events that did not appear in a sample this would imply the estimate $\bar{G}_q(0) = 0$.

Despite of being MLE, such a pessimistic estimate does not look satisfactory, and one would hope to produce reasonable positive estimator for $G_q(0)$ and $p_q(0)$. The paper Orlitsky *et al.* (2003) recalls that P. Laplace, in his "Philosophical Esseys on Probabilities" of 1825 (see translation Laplace(1995)) suggested to use the quantities

$$\tilde{p}_q(k) = \frac{k+1}{N+\mu_q+1}, \quad k = 1, 2 = \ldots.$$

Since $\sum_{k=1}^{\infty} \tilde{p}_q(k)\mu_q(k) = (N+\mu_q)?(N+\mu_q+1)$, this leaves the value

$$\tilde{G}_q(0) = 1/(N+\mu_q+1)$$

to the total probability of unseen outcomes.

Based on easily obtainable equality

$$EG_q(k) = \frac{k+1}{N} E\mu_q(k+1),$$

Good(1953) proposed to estimate $G_q(k)$ and $p_q(k)$ as

$$\hat{G}_q(k) = \frac{k+1}{N}\mu_q(k+1) \quad \text{and} \quad \hat{p}_q(k) = \frac{k+1}{N}\frac{\mu_q(k+1)}{\mu_q(k)}$$

respectively. Since then several authors investigated the statistical quality of these estimators. For example, their rate of convergence was recently studied in McAllester and Schapire (2000).

Notwithstanding importance of this work, we note, however, that Theorems 1 and 2 imply that for a sample, which agrees with Karlin-Rouault law, there is no need to use any estimators. In particular, under conditions of Theorem 2 it follows that

$$EG_q(k) \sim \frac{1}{N}\frac{u\Gamma(k+1-u)}{\Gamma(1-u)\Gamma(k+1)} E\mu_q \quad \text{and} \quad p_q(k) \sim \frac{k-u}{N}, \qquad (17)$$

and thus

$$EG_q(0) \sim \frac{u}{N} E\mu_q \quad \text{and} \quad Ep_q(0) \sim \frac{u}{N}\frac{E\mu_q}{2^q - E\mu_q},$$

as $q \to \infty$. This, in its turn, leads to conclusion that if Karlin-Rouault law is satisfied, then in the underlying probabilities there should have been approximately $\mu_q(k)$ probabilities, equal $(k-u)/N$, while the total probability of unseen outcomes was $u\mu_q/N$, more optimistic (small but infinitely larger) than value $\tilde{G}_q(0)$ suggested by Laplace.

Note that the right hand side of (17) provides "smoothing" of $G_q(k)$ in k: even if $\mu_q(k+1) = 0$ (while $\mu_q(k) \neq 0$), unlike $G_q(k)$, which then also is equal 0, the right hand side of (17) is not and behaves in k "smoothly". The need for "smoothing" was discussed in Good(1956) and more recently in Gale and Sampson(1995) and, among other things, in Orlitski *et al* (2003).

The possible spread of values of the underlying probabilities, which can be extracted from Good-Turing indices, is, however, inherently an approximation only. Indeed, every probability $p(\overrightarrow{x})$ can be estimated as a $p_q(k)$ or,

say, $p_q(k+1)$, depending on whether $f_q(\overrightarrow{x}) = k$ or, say, $f_q(\overrightarrow{x}) = k+1$. However, for most $\overrightarrow{x}$ its frequency $f_q(\overrightarrow{x})$ have Poissonian behavior and can take these different values with no small probability. Thus, even for large q, classification for $p(\overrightarrow{x})$ remains random and misclassification very possible. The statement below shows, however, that more accurate and complete evaluation of overall behavior of probabilities is possible.

Let

$$H_q(z) = \frac{1}{2^q} \sum_{\overrightarrow{x} \in \Xi_q} \mathbb{I}_{\{Np(\overrightarrow{x}) > z\}} \quad \text{and} \quad R_q(z) = \frac{1}{\int_0^\infty (1 - e^{-y}) dH_q(y)} H_q(z)$$

be (tail of) empirical distribution function of $Np(\overrightarrow{x})$ under $\mathbb{P}_{0q}$ and its normalized form respectively. Theorem 4 shows the limit of $R_q(z)$. Although this limit is lurking behind Lemma 2 or Theorem 2, in Theorem 4 we do not use any assumption about the structure of probabilities $p(\overrightarrow{x})$.

Theorem 4. *If, as $q \to \infty$ and sample size $N = \lambda 2^q$ with $\lambda = const$,*

$$\frac{\mu_q(k)}{\mu_q} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}, \quad k = 1, 2, \ldots, \ 0 < u < 1,$$

then for all $z > 0$

$$R_q(z) \to R(z) = \frac{1}{\Gamma(1-u)} z^{-u}.$$

One might think that the overall spread of probabilities $p(\overrightarrow{x})$, $\overrightarrow{x} \in \Xi$, in the questionnaire model above could be more or less arbitrary and, for example, in the case of independent questions, should essentially depend on the distribution function F of individual probabilities $a_1, \ldots, a_q$. Contrary to this, Theorem 4 shows that the spread can be described through very narrow class of functions and the dependence on F is very weak - only through the value of the parameter u. In the next section we will see that numerically u changes very little.

Note that R_q can be written as

$$R_q(z) = \frac{1}{E\mu_q} \sum_{\overrightarrow{x} \in \Xi_q} \mathbb{I}_{\{Np(\overrightarrow{x}) > z\}}.$$

It is interesting to compare the statement of Theorem 4 with what one could get, in the same conditions, if one uses naive estimators $\bar{p}_q(k)$ or the estimators $p_q(k)$ that follow from Good-Turing indices. We formulate it as a corollary (of Theorem 2.

Define functions $R_{q,MLE}(z)$ and $R_{q,GT}(z)$ as

$$R_{q,MLE}(z) = \frac{1}{E\mu_q}\sum_{k=0}^{\infty}\mathbb{I}_{\{N\bar{p}_q(k)\geq z\}} \quad \text{and} \quad R_{q,GT}(z) = \frac{1}{E\mu_q}\sum_{k=0}^{\infty}\mathbb{I}_{\{Np_q(k)\geq z\}}$$

Corollary 5. *Suppose the condition of Theorem 4 is satisfied. Then*

$$R_{q,MLE}(z) \sim \begin{cases}\Gamma(k-u)/\Gamma(1-u)\Gamma(k), & k-1<z\leq k,\ k=1,2,\dots,\\ 2^q/E\mu_q\to\infty, z=0,\end{cases}$$

while

$$R_{q,GT}(z) \sim \begin{cases}R_{q,MLE}(z+u), & u\frac{E\mu_q}{2^q}<z,\\ 2^q/E\mu_q\to\infty, & 0\leq z\leq u\frac{E\mu_q}{2^q}.\end{cases}$$

Proof of Theorem 4. It is easy to check that

$$-\int_0^{\infty}\pi(k,z)dR(z) = \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}, \quad k=1,2,\dots \tag{18}$$

and, integrating by parts, that

$$-\int_0^{\infty}(1-\pi(0,z)dR(z) = \frac{1}{\Gamma(1-u)}\int_0^{\infty}z^{-u}e_z dz = 1. \tag{19}$$

Equations (18) determine the mixing measure R uniquely, given it has no weight at the point $z=0$, or, equivalently, up to summand $a\delta_{\{0\}}(z)$. On the other hand, the ratio $E\mu_q(k)/E\mu_q$ can be viewed, see (4), as a mixture of Poisson probabilities

$$\frac{\mu_q(k)}{\mu_q} = -\int_0^{\infty}\pi(k,z)dR_q(z).$$

The sequence $-\int_0^z(1-e^{-y})dR_q(y), q=1,2,\dots$, forms a sequence of probability distribution functions in z. For any its sub-sequence, weakly converging

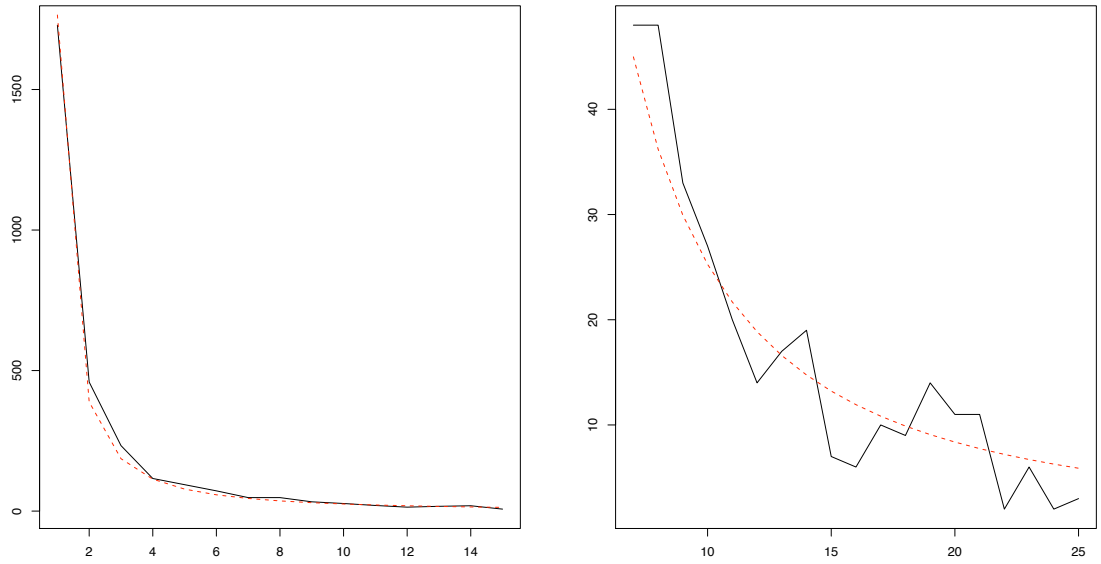

Figure 1: Statisitcs $\mu_q(k)$ from NLTCS disability data (solid line) and its approximation by Karlin-Rouault law. On the left graph $k = 1, \ldots, 15$ while $k = 7, \ldots, 25$ are shown, on larger scale, separately on the right graph

to a (possibly deficient) distribution function $-\int_0^z (1-e^{-y}) dR'(y)$, one should have

$$\int_0^\infty \pi(k,z) dR_{q'}(z) \to \int_0^\infty \pi(k,z) dR'(z), \quad \text{as} \quad q' \to \infty,$$

which therefore must equal R. □

Consider a practical example on disability data from NLTCS, which recently received careful consideration in Erosheva *et al.* (2007). The data represented (0-1) responses on presence or absence of $q = 16$ parameters in $N = 21574$ disability patients. Among other things, H. Erosheva and co-authors demonstrated that the parameters were not independent random variables and probabilities $p(\overrightarrow{x})$ are of more complex structure. At the same time, one can see that the ratios $\mu_q(k)/\mu_q$ in the data follow, as it is shown in Fig 1, Karlin-Rouault law with $u = 0.55$ quite closely. Therefore, Theorem 4 can be

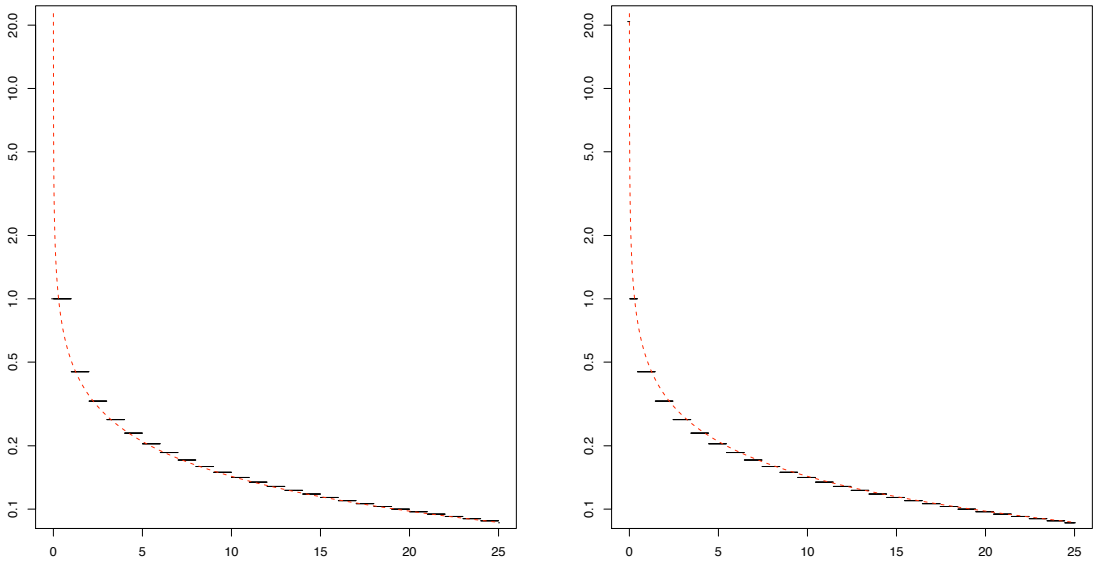

Figure 2: The graph on the left compares $R(z)$ of Theorem 4, corresponding to NLTCS, dotted line, and $R_{q,MLE}$ as in the corollary. The graph on the right compares $R(z)$ with $R_{q,GT}$.

applied to show how many how small have been the underlying probabilities. Fig 2 shows this and illustrates the mutual behavior of the function R and asymptotic form of $R_{q,MLE}$ and $R_{q,GT}$. We replaced $E\mu_q$ by $\mu_q = 3152$ from the NLTCS data.

The best way to judge whether the functions $R_{q,MLE}$ and $R_{q,GT}$ are sufficiently accurate to describe the underlying probabilities is to generate samples from all three and compare. This was done in Kvizhinadze and Wu (2009) and the samples produced from $R_{q,MLE}$ and $R_{q,GT}$ were quite different from each other and from the NLTCS disability data (Fig 1); the former produced far too small, while the latter produced far too large $\mu_q(1)$. The differences faded away for $\mu_q(k)$ after $k = 10$.

5 Numerical behaviour of the asymptotic formulas for moderate q.

It would be quite possible to consider our problem within a context, where q is very large. For example, in the context of complex systems, with q different "on/off" components, the value of q can easily be of several hundreds or thousands. For such systems, each $\overrightarrow{x}$ will describe the state of the system (the list of states of its components), while μ_q and $\mu_q(k)$ will be the number of the different observed states and the number of states observed k times in a long sequence of trials, respectively.

In a testing problem for such a system, the states $\overrightarrow{x}$, in which the system will fail, for practically interesting cases will be of quite small probability, but there most likely will be many of such states. Therefore, the question of what proportion of them will we see in testing trials is of considerable interest, when high reliability is needed. There is a number of interesting questions, apart from those we consider in this paper, arising in the context of such systems. We intend to study some of them in a separate work.

In the context of questionnaires or classifications, although, e.g., Loughin and Scherer(1998) and Agresti and Liu(1999) demonstrated advantages of viewing a questionnaire with, say, q' questions with multiple responses as a longer one with binary responses, the number of questions or 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 the exact value will not matter much.

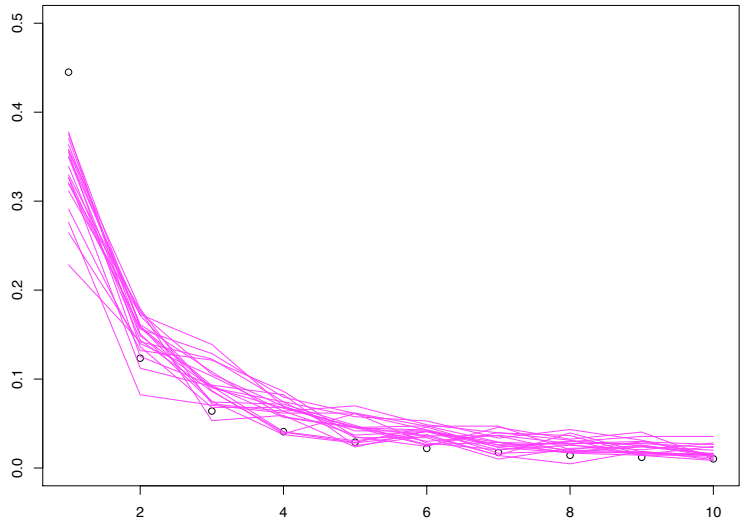

Figure 3: The bundle of trajectories of $\mu_q(k)/\mu_q$ for $k = 1, \ldots, 10$. The number of questions $q = 10$ and probabilities $a_1, \ldots, a_{10}$ are uniformly distributed on [0,1]. Dots show the limits of $\mu_q(k)/\mu_k$.

Convergence of $\mu_q(k)/\mu_q$: the plots in Fig 3 and 4 show that this convergence, although not too quick, is reasonable. The bundle of graphs of the ratio $\mu_q(k)/\mu_q$ for $q = 10$ uniformly distributed probabilities a_i along with the limiting expression is given on Fig 3. For $q = 20$, the next Fig 4 shows closer approximation and smaller spread in the bundle of trajectories of $\mu_q(k)/\mu_q$.

Transition from the contiguity case to Karlin - Rouault law: it is interesting to see which limiting values c of Hellinger distance correspond to the contiguity case, and which ones would already correspond to large deviations. It is also interesting to see, what is the influence of "rate per cell" λ on the transition from one case to another as c increases. The graphs below show the ratio of integrals (5) 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 0.9973, so the latter case can be thought of as the case of "large deviations". The corresponding, uppermost at $k = 1$, graph is quite close to the limit, while the graph for $c = 1$ is very far from it. In Fig 5 the value of $\lambda = 5$ is not very high. As Fig 6 shows,

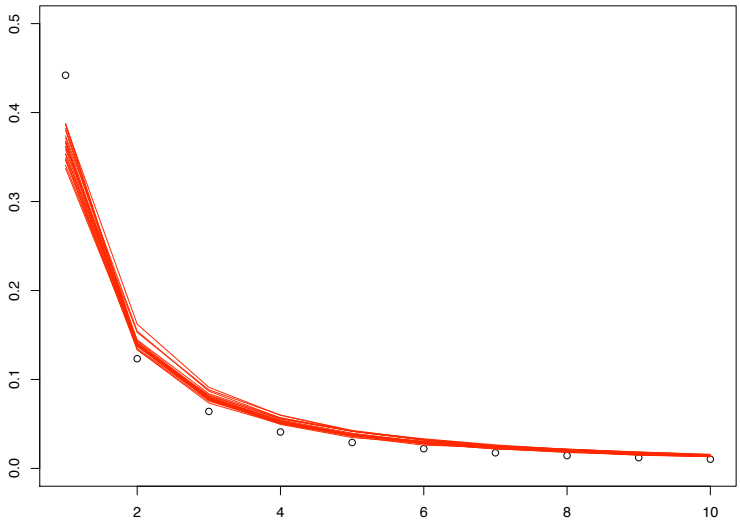

Figure 4: The bundle of trajectories of $\mu_q(k)/\mu_q$ for $k = 1, \ldots, 10$. The number of questions $q = 20$ and probabilities $a_1, \ldots, a_{20}$ are uniformly distributed on [0,1].

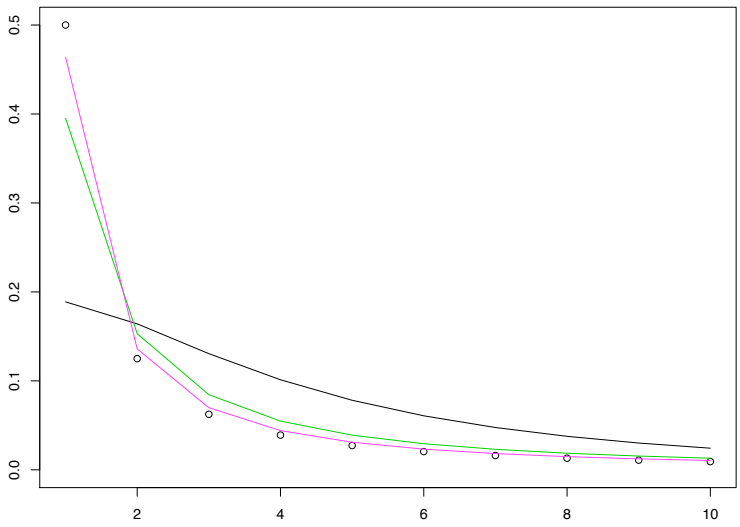

Figure 5: Three graphs of the ratio (5) in k for $c = 1, 3, 6$. Here $\lambda = 5$.

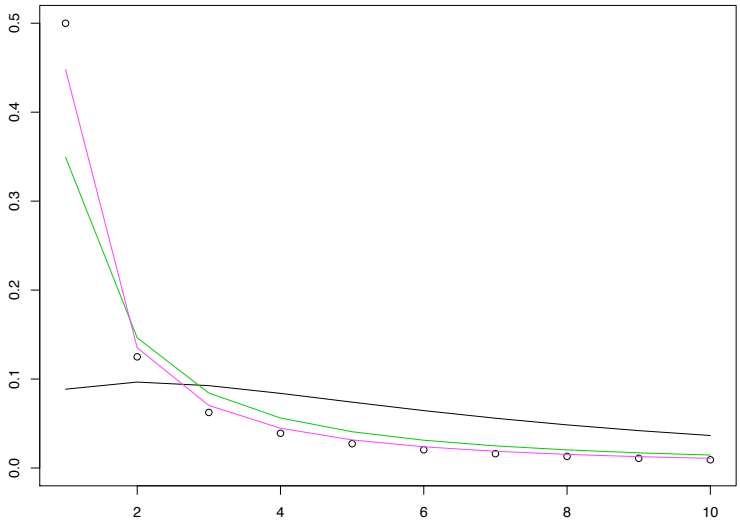

Figure 6: The same situation as in the previous graph, but with $\lambda = 10$.

for higher λ and not large c the ratio (5) can behave very differently from its limit - in this graph $\lambda = 10$. However, with c increasing, the influence of large λ fades away noticeably.

Convergence of $\mu_q/2^q$ to 0: the rate of this convergence, numerically, is not too high. For example, for $q = 20$ and $\lambda = 3$ one would still see in a sample about 20% of all possible "opinions". Also, one would expect that this rate will strongly depend on the overall spread of $a_1, \ldots, a_q$, that is, on F. As for a_i-s all equal $1/2$ the expectation $E\mu_q$ reaches its maximum value $1 - e^{-\lambda}$, one would expect that for a_i-s tending to $1/2$ the ratio $\mu_q/2^q$ should be essentially larger that for the case when they tend to the end-points of $[0, 1]$. This effect, although visible on Fig 7, is, however, not too strong.

Related question is whether the asymptotic expression for $\mu_q/2^q$ in (13) is accurate for our choice of not very large q. The graph in Fig 8 illustrates the convergence of $\mu_q/2^q$ and its asymptotic form to each other. However, we can increase this convergence sharply if we note there is a loss in numerical accuracy in the step from last integral in (15) to the final form, because, as a function in s, $\phi_{0,\sigma_q^2}((\ln s - \ln \lambda)/\sqrt{q})$ for not too large q decreases noticeably as s/λ deviates from 1. To replace it by its maximum value $\phi_{0,\sigma_q^2}(0)$

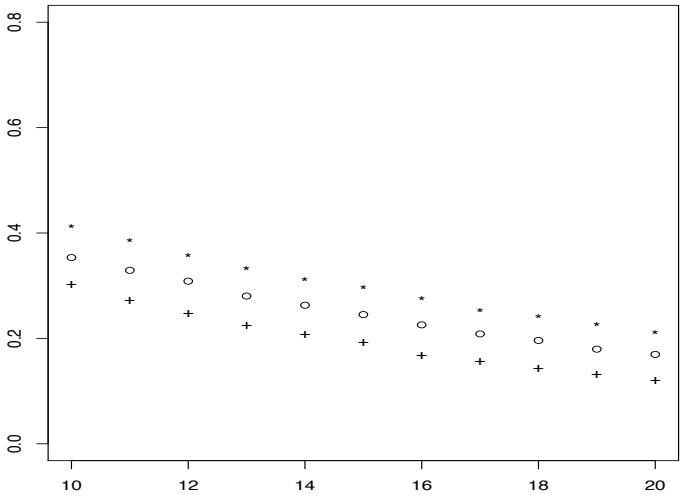

Figure 7: The graph of $\mu_q/2^q$ in q for three different distributions F of a_i-s: the middle graph is for uniform F, the upper one - for B-distribution with the density $a^{\beta-1}(1-a)^{\beta-1}/B(\beta,\beta)$ and $\beta = 1.2$ while the lower graph is for U-shaped B-distribution with $\beta = 0.8$.

with sufficient accuracy requires $q \geq 50$. Since σ_q^2 is numerically stable, one can calculate the integral without much difficulty and use the asymptotic expression

$$\frac{\mu_q}{2^q} \sim e^{\sum_{i=1}^q \psi_i(u)} \frac{\lambda^u}{u\sqrt{q}} \int_0^\infty s^{-u} \phi_{0,\sigma_q^2}(\frac{\ln s - \ln \lambda}{\sqrt{q}}) e^{-s} ds. \tag{20}$$

The third graph in Fig 8 shows that this latter expression is quite accurate.

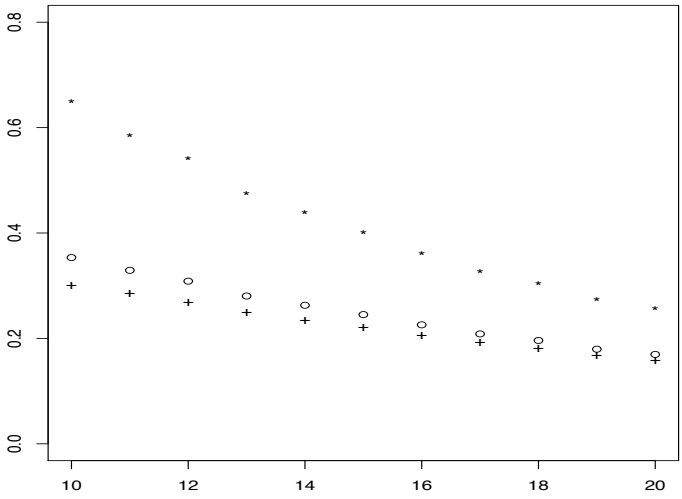

Figure 8: The graph of $\mu_q/2^q$ in q (in the middle) along with two asymptotic expressions: given in Theorem 2 (the upper graph) and in (15) and (20) above. The graphs show that the approximation based on large deviations works reasonably well.

References

[1] A. Agresti and I.M. Liu (1999) Modelling a categorical variable allowing arbitrarly many category choices. *Biometrics*, **55**, 936-943.

[2] R.H. Baayen (2002) *Word frequency distribution*, Kluiver Acad. Publishers.

[3] R.R. Bahadur and R. Ranga Rao (1960) On deviation of sample mean, *The Annals Mathematical Statistics*, 31, 1015-1027.

[4] N.R. Chaganty and J. Sethuraman (1993) Large deviation and local limit theorems, *The Annals of Probability*, **1**, 3, 1671-1690.

[5] Elena A. Erosheva, Stephen E. Fienberg and Cyrille Joutard (2007) Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, **1**, 2, 502-537.

[6] W. Feller (?)*Introduction to probability theory,* vol.2, John Wiley

[7] W.A. Gale and G. Sampson (1995) Good-Turing estimation without tears *J. Quantitative Linguistics* , **2**, 217-237.

[8] I.J. Good (1953) The population frequencies of species and the estimation of population parameters *Biometrica*, **40**, 3/4, 237-264.

[9] P. Greenwood, A.N. Shiryaev (1985) *Contiguity and the statistical invariance principle,* Gordon and Breach, London.

[10] M. Gyllenberg, T. Koski (1996) Numerical taxonomy and the principle of maximum entropy, *J. Classification*, **13** , 2, 213–229.

[11] M. Gyllenberg and T. Koski, (2001) Probabilistic Models for Bacterial Taxonomy, *International Statistical Review*, **69**, 249-276.

[12] O. Kallenberg (1997) *Foundations of Modern Probability*, Springer-Verlag.

[13] E.V. Khmaladze (2002) Zipf's Law, *Encyclopaedia of Mathematics, Supplement III,* Kluwer Academic Publishers, Dordrecht.

[14] E.V. Khmaladze and Z.P. Tsigroshvili (1993) On Polynomial Distributions with Large Number of Rare Events, *Mathematical Methods of Statistics,* **2**, 3, 240-247.

[15] John E. Kolassa (1994) *Series approximation methods in statistics*, 1994, Lecture Notes in Statistics, Springer-Verlag.

[16] Giorgi Kvizhinadze and Haizhen Wu (2009) *??* SMSOR Report ??, Victoria University of Wellington.

[17] P. Laplace (1825, 1995) *Philosophical Essays on Probability,* (transl. by A.I. Dale), Springer-Verlag.

[18] T.M. Loughin and P. Scherer (1998) Testing for association in contingency tables with multiple column responses. *Biometrics*, **54**, 630-637.

[19] R.H. MacArthur (1957) On relative abundancy of bird species, *Proceedings of National Academy of Sciences of USA*, **43**, 293-295.

[20] David McAllester, Robert E. Schapire (2000) On the convergence rate of Good-Turing estimators. *COLT 2000*, 1-6.

[21] J. Oosterhoff and W. van Zwet (1979) A note on contiguity and Hellinger distance, *In: Contribution to Statistics. Jaroslav Hajek Memorial Volume (J.Jurechkova, ed.). Reidel, Dordrecht.*

[22] A. Orlitsky, Narayana P. Santhanam, Juan Zhang (2003) Always Good Turing: Asymptotically Optimal Probability Estimation. *Science*, **302**, 5644, 427-431.

[23] A. Rouault (1978) Loi de Zipf et sources markoviennes, *Annales de l'Institute H. Poincare*, **14**, 169-188.

Address:
E.Khmaladze
School of Mathematics, Statistics and Operations Research
Victoria University of Wellington
PO Box 600
Wellington, New Zealand
E-mail: Estate.Khmaladze@msor.vuw.ac.nz