Large number of rare events: Diversity analysis in multiple choice questionnaires and related topics

by

Giorgi Kvizhinadze

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Statistics.

Victoria University of Wellington
2010

Abstract

The statistical analysis of a large number of rare events, (LNRE), which can also be called statistical theory of diversity, is the subject of acute interest both in statistical theory and in numerous applications. A careful eye will quickly see the presence of a large number of very rare objects almost everywhere: large numbers of rare species in ecosystems, large numbers of rare opinions in any opinion pool, large numbers of small admixtures in any solution and large numbers of rare words in any text are only few examples.

In studying such objects, the interest for mathematical statisticians lies in the fact that most of the frequencies are small and, therefore, difficult to deal with. It is not immediately clear how one should be able to derive consistent and reliable inference from a large number of such frequencies.

In this thesis we study the diversity of questionnaires with multiple answers. It has been demonstrated that this is a particular model of LNRE theory. In our analysis, the theories of large deviation, contiguity and Edgeworth expansion were employed, and limit theorems have been established.

Acknowledgments

I would like to thank my supervisor Professor Estate Khmaladze for giving me an opportunity to do a PhD with his financial support, for warm and friendly guidance throughout my research and for encouraging me to work on very challenging and interesting topics.

I would also like to thank Victoria university's faculty of science staff and School of Mathematics, Statistics and operations Research staff for valuable assistance they were providing.

I am very grateful for Victoria university's accommodation service, namely for University Hall office for providing excellent conditions to live and study.

And finally want to thank my family back home in Georgia for the continuous encouragement and support I was feeling from them.

Contents

Chapter 1

Introduction

The background of the LNRE theory can be traced back in the 1980's. The pioneering work of formal statistical analysis of the concept of large number of rare events was [14], where Khmaladze introduced the notions, studied its different forms and found various necessary and sufficient conditions for results. Also, some significant connections between this area and several other areas of statistical theory were found.

1.1 Examples of statistical data with a large number of rare events.

From some points of view the presence of a large number of rare events is a rather fundamental feature of nature. In particular, in any statistical analysis devoted to the study of the variety of words in the large text or variety of species, one has to deal with what might be called "a large number of rare events." The common feature of examples we will present below is that along with several frequent events there is also a very large number of very rare events, say, with frequency 0,1,2. The total amount of these rare events compared to the number of observations typically is not large but the number of these events among all different observed events is always

very significant. These rare events are usually very important. For instance the number of words used in the book only once, can be considered not of vital importance for this book, but it is very clear that these words are absolutely important because they constitute half of the author's vocabulary. "...Most of us agree that mankind must preserve a rich variety in biology, i.e. must protect a large number of rare animals and rare plants. "[16].

1.1.1 LNRE in Linguistics

Let n be the total number of words in a text, μ_n be number of different words used in the text, or size of vocabulary, $\mu_n(k)$ - the number of words used in the text k times.
Table 1 contains data taken from [1]. It illustrates frequencies of different words in separate novels. *Tables 2* and 3 contain data from the BNC (British National Corpus) showing word frequency distributions in casual and formal English.

Table 1.1: LNRE in Linguistics

Works	n	μ_n	$\mu_n(1)$
L. Carroll			
Alice in Wonderland	26505	2651	1176
H.G. Wells			
War of the worlds	59938	7112	3613
A. Conan-Doyle			
Hound of the Baskervilles	59241	5741	ca. 2836

Table 1.2: Word frequencies in casual English language usage

n	μ_n	$\mu_n(1)$	$\mu_n(2)$	$\mu_n(3)$
4188576	26618	6718	3616	2259

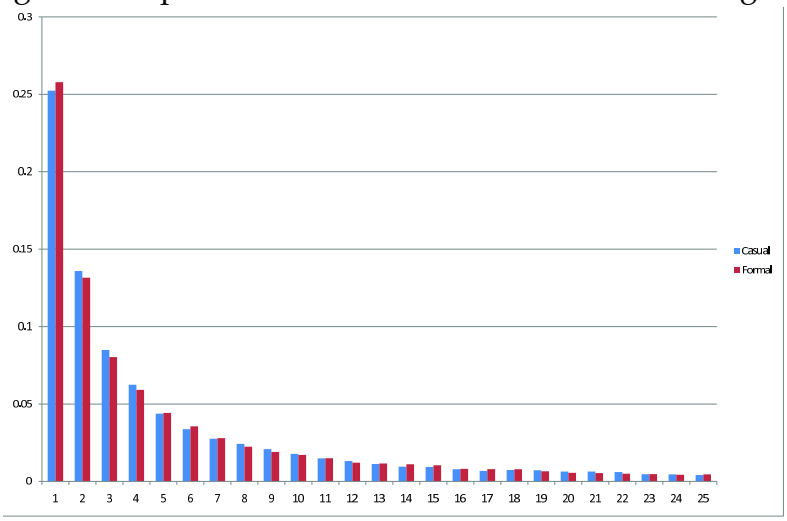
The following chart shows the ratio of number of words used k-times $k = 1, \cdots, 25$, over number of different words (vocabulary) for two corpora of English, for casual and formal English sources. We see similarities

Table 1.3: Word frequencies in formal English language usage

n	μ_n	$\mu_n(1)$	$\mu_n(2)$	$\mu_n(3)$
4187647	34455	8883	4534	2766

between them, although the actual words are different. This demonstrates a common feature of linguistic data.

Figure 1.1: Spectral statistics for casual and formal English



1.1.2 LNRE in Nature

Williams in [37] analyzed the distribution of the number of head lice found on 461 prisoners.

Table 1.4: Number of head lice on prisoners

Lice per head	Number of heads	Lice per head	Number of heads
1	106	7	12
2	50	8	18
3	29	9	11
4	33	10	11
5	20	11-12	13
6	14	13-14	14

1.1.3 LNRE in Chemistry

Data on chemical analysis of a substance, taken from [35], shows that there always is a large number of rare admixtures. For instance "...in ocean water one can find ions of all elements of the periodic system of Mendeleev, though the main part of all inorganic substances dissolved in the ocean water contains only nine ions... The total amounts of these nine ions exceeds 99.9% of the whole amount of all dissolved salts"[35].

Table 1.5: Data on chemical analysis of ocean water

% of total amount of inorganic admixture								
Cl^-	Na^+	SO_2	Mg^{++}	Ca^{++}	K^+	$HCO_3 + CO_2$	Br^-	H_3BO_3
55.04	30.61	7.68	3.69	1.16	1.10	0.41	0.19	0.07

According to Table 5 the remainder of more than a hundred dissolved elements take up only 0.05% of the total amount of inorganic admixture. A large variety of data of the same character is available in demography: for instance, data of the population of different nationalities in a given community, say, in a city or in a whole state.

1.1.4 LNRE in Demography

Table 6 is created from the data taken from Statistics New Zealand website and illustrates ethnical diversity of New Zealand population in 2001.

Table 1.6: Ethnic diversity in New Zealand

No. of different nationalities in the population of New Zealand	Nationalities	$\%$
over 65	NZ European	68.03
	Maori	13.31
	Samoan	2.89
	Chinese	2.53
	Indian	1.51
	Cook Islander	1.29
	Tongan	1.02
	English	0.86
and only 8.5% for more than 57 others		

1.1.5 LNRE in Bibliography

Table 7 contains data extracted from the Author index of *A Journal of Modern Society and Culture* 2003-2007. $(http://www.logosjournal.com/)$

Table 1.7: Author index data

No. of publications, m	1	2	3	4	5	6	7
No. of authors publishing m papers	167	31	12	3	3	2	1

1.1.6 Artificial source of LNRE

An artificial source of a large number of rare events can be created in the following way: let $X_1, \cdots, X_n$ be i.i.d. continuous random variables, distributed over a finite interval $[a, b]$. Divide this interval into N equal subintervals $[a+i\Delta, a+(i+1)\Delta]$, $\Delta = \frac{1}{N}$, $i = 0, \cdots, N-1$. For small Δ the events $X_j \in [a+i\Delta, a+(i+1)\Delta]$ have small probabilities, but the number N of such events is large. Let ν_{in} be the frequency of X_i with values in the i-th subinterval. The main question is behavior of the vector $\nu_n = (\nu_{1n}, \cdots, \nu_{Nn})$ of frequencies when both n and N are large. In the second part of this thesis we will see that properties of statistical methods based on grouped data (such as the χ^2 test) are changed very essentially if N is not much smaller than n [14].

1.2 Definition of a large number of rare events.

In this section we formulate key definitions and results obtained by E. Khmaladze in [14], where LNRE theory was shaped as an independent and significant area of statistics.

Consider a random vector of frequencies $\nu_n = (\nu_{1n}, \cdots, \nu_{Nn})$ which has a multinomial distribution with vector of probabilities $p_n = (p_{1n}, \cdots, p_{Nn})$ and sample size n,

$$\mathbb{P}\{\nu_{in} = k_i, i = 1, \cdots, N\} = \frac{n!}{\prod_{i=1}^{N} k_i!} \prod_{i=1}^{N} p_{in}^{k_i},$$

$$\sum_{i=1}^{N} k_i = n,\ k_i \geq 0,\ \sum_{i=1}^{N} p_{in} = 1.$$

The number N of different events might be finite or infinite. The random variable ν_{in} is called the frequency of the i-th event.

Consider the statistics

$$\mu_n(m) = \sum_{i=1}^{N} I\{\nu_{in} = m\}$$

and

$$\mu_n = \sum_{i=1}^{N} I\{\nu_{in} > 0\}$$

where

$$I\{\nu_{in} = m\} = \begin{cases} 1, & \nu_{in} = m \\ 0, & \nu_{in} \neq m, \end{cases}$$

so that $\mu_n(m)$ is the number of events observed in n trials exactly m times, and μ_n is the number of different observed events in n trials. The vector $\{\mu_n(1), \mu_n(2), \cdots, \mu_n(n)\}$ is sometimes called the set of spectral statistics.

The marginal distribution of each frequency is binomial:

$$\mathbb{P}\{\nu_{in} = k\} = \frac{n!}{k!(n-k)!} p_{in}^k (1 - p_{in})^{n-k}.$$

Hence

$$\mathbb{E}\mu_n(m) = \frac{n!}{m!(n-m)!} \sum_{i=1}^{N} p_{in}^m (1 - p_{in})^{n-m}$$

and

$$\mathbb{E}\mu_n = \sum_{i=1}^{N} [1 - (1 - p_{in})^n].$$

1.2.1 d1 and d2 zones of LNRE

Definition 1. *(d1) A sequence of random vectors $\{\nu_n\}$ is called a sequence with a large number of rare events (LNRE sequence) if*

$$\lim_{n \to \infty} \frac{\mathbb{E}\mu_n(1)}{n} > 0$$

Definition 2. *(d2) A sequence of random vectors $\{\nu_n\}$ is called an LNRE sequence if*

$$\lim_{n\to\infty} \frac{\mathbb{E}\mu_n(1)}{\mathbb{E}\mu_n} > 0 \ \ and \ \ \lim_{n\to\infty} \mathbb{E}\mu_n = \infty$$

For convenience, we will say that we are in $(d1)$ or $(d2)$ zone of LNRE if $(d1)$ or $(d2)$ is satisfied respectively. These two definitions are not equivalent: namely $(d1) \Rightarrow (d2)$, but not vice versa. Later on we will see examples when $(d2)$ is satisfied but not $(d1)$. It is easy to observe that for a fixed finite N and fixed vector of probabilities p each frequency $\nu_{in} \to \infty$ a.s. as $n \to \infty$, and therefore $\mu_n(1) \to 0$ and $\mu_n \to N$ a.s. Consequently $(d1)$ and $(d2)$ cannot be satisfied.

If $X_1, \cdots, X_n$ are i.i.d. random variables with some absolutely continuous distribution supported on interval $[0, 1]$ and if ν_{in} are frequencies accumulated in the subinterval $[\frac{i-1}{N}, \frac{i}{N})$, then for $N = cn$, $n \to \infty$, the sequence of the vectors of frequencies $\{\nu_n\}_{n\geq 1}$ satisfies $(d1)$ and therefore $(d2)$ as well [14].

Following two functions, introduced in [16], play crucial role in the study of LNRE theory:

$$G_n(z) = \sum_{i=1}^{N} I\{np_{in} > z\},$$

$$Q_n(z) = \sum_{i=1}^{N} p_{in} I\{np_{in} \leq z\}.$$

The function $G_n(z)$, called the structural distribution function, is one of the main objects of interest in LNRE theory. Estimation of this function was widely considered in [36], [26] and [18].

1.2.2 Some conditions for d1 and d2 zones of LNRE

Condition 1. *(c1) For some $z < \infty$*

$$\lim_{n\to\infty} Q_n(z) > 0$$

Condition 2. *(c2) For some* $z < \infty$

$$\lim_{n\to\infty} \frac{G_n(z)}{nQ_n(z)} < \infty$$

and

$$\lim_{n\to\infty} nQ_n(z) = \infty.$$

Lemma 1. $(c1) \Leftrightarrow (d1)$.

$$\lim_{n\to\infty} \frac{\mathbb{E}\mu_n(1)}{n} > 0 \quad \Rightarrow (c1). \ [14]$$

Lemma 2. $(c2) \Leftrightarrow (d2)$. *[14]*

As Khmaladze indicates, some other definitions also may correspond to the intuitive understanding of the expression "large number of rare events". For instance $\{\nu_n\}$ could be called an LNRE sequence if

$$\lim_{n\to\infty} \mathbb{E}\mu_n(1) = \infty$$

or if

$$\lim_{n\to\infty} \mathbb{E}\mu_n(1) > 0$$

or if

$$\lim_{n\to\infty} \mathbb{E}\mu_n = \infty.$$

1.2.3 Some artificial sources for d1 and d2 zones of LNRE

Case 1. *Let* $X_1, \cdots, X_n$ *be independent random variables, identically distributed on* $[0, 1]$ *and* f *be the density of the distribution of* X_i*. Consider the uniform partition of* $[0, 1]$ *by* N *points and denote*

$$p_{in} = \Delta F(\frac{i}{N}), \ \ f_n(t) = Np_{in}, \ \ \frac{i-1}{N} \le t < \frac{i}{N}$$

If $N \to \infty$ *then the frequencies* $\nu_{1n}, \cdots, \nu_{Nn}$*, where*

$$\nu_{in} = \sum_{j=1}^{N} I\{\frac{i-1}{N} \le X_j < \frac{i}{N}\}$$

satisfy $(d1)$. *[14].*

Case 2. *Let p be a non-increasing density on $(0, \infty)$ and let*

$$p_{in} = p_i = \int_{i-1}^{i} p(t)dt$$

Let $X_1, \cdots, X_n$ be i.i.d. random variables with density p and, finally, let

$$\nu_{in} = \sum_{j=1}^{N} I\{i - 1 \leq X_j < i\}.$$

Lemma 3. *For any fixed p the sequence $\{\nu_n\}$ of vectors of frequencies (6) does not satisfy $(d1)$.*

Condition 3. $(c.3)$ *For some $\rho \in (0, 1]$*

$$p(t) = t^{-\rho} L(t)$$

where L is a slowly varying function, that is $\dfrac{L(tc)}{L(t)} \to 1$ as $t \to \infty$ for any $c > 0$

Lemma 4. *Let p be a fixed density and p_i and ν_{in} be defined as in Case2. Then $(c.3) \Leftrightarrow (d2)$.[14]*

Chapter 2

Models and Laws of LNRE

2.1 Distribution functions of probabilities

Assume that $\{\nu_{in}\}_{i=1}^{N}$ are drawn from a multinomial distribution with probabilities $\{p_{in}\}_{i=1}^{N}$ which form an array with respect to n. Asymptotic behaviour of statistics $\{\nu_{(k,n)}\}_{k=1}^{\mu_n}$, $\{\mu_n(k)\}_{k=1}^{\mu_n}$ is governed by the distribution function of probabilities $\{p_{in}\}_{i=1}^{N}$ [14]. Let us recall the function $G_n(x)$ introduced in previous section,

$$G_n(x) = \sum_{i=1}^{N} I\{np_{in} \geq x\}.$$

If

$$\frac{1}{n} G_n \xrightarrow{w} G$$

then

$$\frac{\mu_n(k)}{n} \xrightarrow{P} \int_0^\infty \frac{\lambda^k}{k!} e^{-\lambda} G(d\lambda).$$

Let us define new measure $R_n(x)$ as follows

$$R_n(x) = \frac{G_n(x)}{\int_0^\infty (1 - e^{-z}) G_n(dz)}.$$

If

$$R_n \xrightarrow{w} R$$

and

$$\lim_{\epsilon \to 0} \sup_n \int_0^\epsilon z R_n(dz) = 0$$

then

$$\frac{\mu_n(k)}{\mu_n} \xrightarrow{P} - \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} R(d\lambda).$$

This expression shows that we have as many possible limiting expressions for the ratio of spectral statistics $\frac{\mu_n(k)}{\mu_n}$ as there are measures which satisfy

$$\int_0^\infty (1 - e^{-\lambda}) R(d\lambda) = 1$$

[16].

2.2 Zipf's Law

The French stenographer J.B. Estoup observed that word frequencies in a long text, or in a corpus, fall off inversely with the word's rank according to a simple power law. Later on this phenomenon was systematically studied by the American linguist and philologist G.K. Zipf.

2.2.1 Definition of Zipf's Law

Let $\{\nu_{in}\}_{i=1}^N$ be frequencies of N, $N \leq \infty$, different disjoint events in a sample of size n, for example, occurrences of different words in a text of n running words. So called "empirical vocabulary", or number of different words in a text, can be defined as follows:

$$\mu_n = \sum_{i=1}^N I\{\nu_{in} \geq 1\}.$$

Number of words that occurred in the text exactly k times can be written as follows:

$$\mu_n(k) = \sum_{i=1}^N I\{\nu_{in} = k\}.$$

The following assertion is called *Zipf's law*

$$\frac{\mu_n(k)}{\mu_n} \approx \frac{1}{k(k+1)}, \quad k = 1, 2, \cdots$$

For example, when $k = 1$, then, $\frac{\mu_n(1)}{\mu_n} \approx \frac{1}{2}$. In other words it means that half of empirical vocabulary is built up from the words which are used in the text only once. The words that are used only twice constitute $\frac{1}{6}$ of vocabulary and so on.

Notice that in this case definition $(d2)$ is satisfied, but not $(d1)$. Indeed

$$\frac{n}{\mathbb{E}\mu_n} = \sum_{k=1}^{\infty} k \frac{\mathbb{E}\mu_n(k)}{\mathbb{E}\mu_n} \to \infty, \; n \to \infty$$

and therefore

$$\frac{\mu_n}{n} \to 0.$$

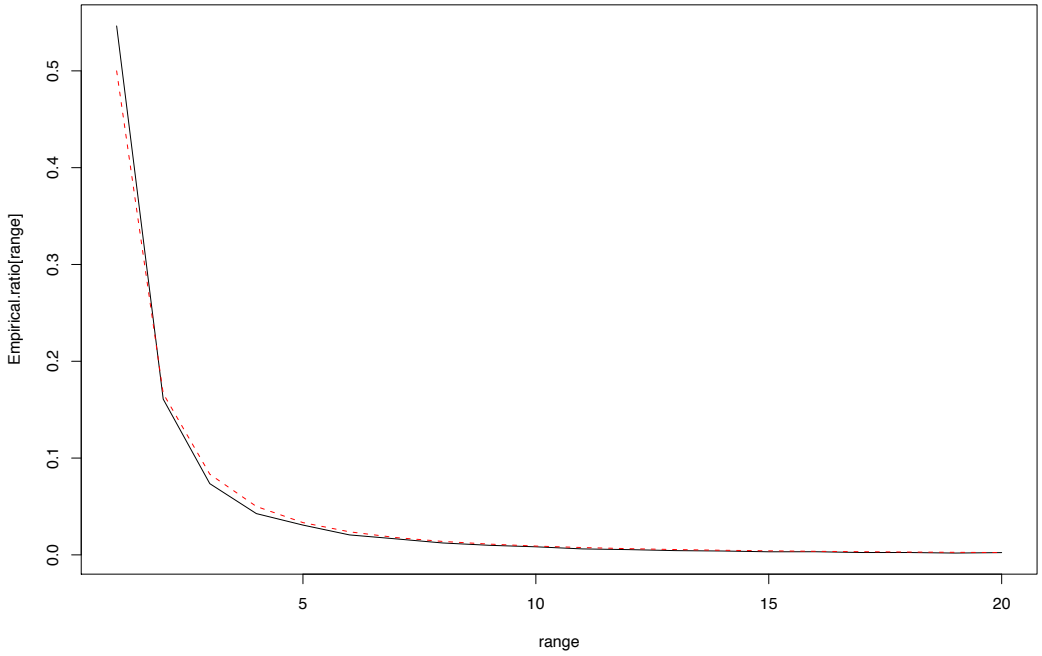
2.2.2 Data satisfying Zipf's Law

Below is given data from James Joyce's famous novel *Ulysses*. As we can see empirical data agrees with the theoretical law quite well.

Table 2.1: Spectral statistics for James Joyce's *Ulysses*

Total number of words $n = 264217$									
Vocabulary $\mu_n = 30030$									
$\mu_n(1)$	$\mu_n(2)$	$\mu_n(3)$	$\mu_n(4)$	$\mu_n(5)$	$\mu_n(6)$	$\mu_n(7)$	$\mu_n(8)$	$\mu_n(9)$	$\mu_n(10)$
16409	4832	2209	1280	924	619	496	369	297	250

Figure 2.1: Zipf's law in James Joyce's *Ulysses*



2.3 Zipf-Mandelbrot Law

Mandelbrot in [24] used considerations based on information theory. He took the "effort" or "cost" of words as the delay resulting from their transmission as a sequence of letter patterns or phonemes, separated by spaces or pauses. Assuming that the aim of language is to allow transmission of the largest variety of signals as possible with the least delay, he used the technique for matching codes to message usage [3]. The statement

$$\frac{\mu_n(k)}{\mu_n} \approx \frac{1}{(a+bk)^2}$$

is frequently called *Zipf-Mandelbrot law*. Here a and b are constants for the texts being analyzed.

2.4 Pareto distribution

The Pareto distribution in its original form, or "Pareto distribution of the first kind", is defined as follows:

$$F(x) = 1 - \left(\frac{k}{x}\right)^a, a > 0; x \geq k \geq 0$$

where a and k are parameters. Corresponding density function is:

$$f(x) = \frac{ak^a}{x^{a+1}}, a > 0; x \geq k \geq 0$$

The economist Vilfredo Pareto formulated following Law:

$$N = Ax^{-a}$$

where N is number of those in the community with income equal, or exceeding x, and A, a are parameters. Notice that it is very general distribution and applies to a very wide range of phenomena outside economics.

2.5 Yule-Willis taxonomy model

Yule in [38], analyzing data of J.C. Willis, assumed that new species within given genus arise from a specific mutation. If p is probability that in some small assigned interval of time Δt the species will mutate, then starting with N species of different genera, after Δt time we will have $N(1-p)$ genera without new species and Np genera with new species. Proceeding to the limit, taking the time-interval Δt as indefinitely small but the number of such intervals n as large, so that the time $n\Delta t = t$ is finite, Yule obtained: $p = s\Delta t$, $pn = st$ and

$$q^n = (1-p)^n = (1 - \frac{st}{n})^n \sim e^{-st}$$

Now, according to our notations introduced above, let $\mu_n(k)$ be number of genera containing k species and μ_n be number of different genera respectively. Then according our scheme μ_n is N. After certain transformations, Yule obtained following result:

$$\frac{\mu_n(k)}{\mu_n} \xrightarrow{P} \frac{\gamma\Gamma(\gamma+1)\Gamma(k)}{\Gamma(k+\gamma+1)},$$

where γ is some parameter.

2.6 Lotka's distribution of literary productivity

Lotka's law describes the frequency of papers published by authors in any given field. In a study of literary output, Lotka in [21] found that the number of authors who had published k papers in a given field was roughly $\frac{1}{k^2}$ of the number of authors who had published one paper only [3].

2.6.1 Data satisfying Lotka's Law

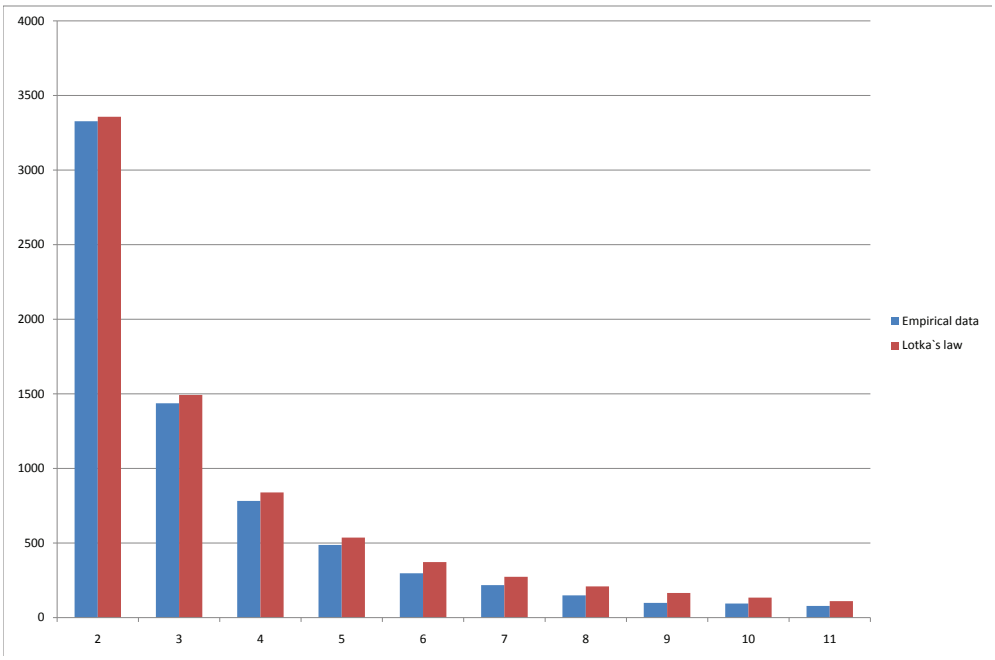
In the following Table we illustrate the data taken from [31]. It comprises the numbers of times that individual authors had published with Emerald Group Publishing Limited.

Table 2.2: Bibliographic data from Emerald Group Publishing Limited

Total number of authors, $n = 20624$									
$\mu_n(1) = 13428$									
$\mu_n(2)$	$\mu_n(3)$	$\mu_n(4)$	$\mu_n(5)$	$\mu_n(6)$	$\mu_n(7)$	$\mu_n(8)$	$\mu_n(9)$	$\mu_n(10)$	$\mu_n(11)$
3327	1437	782	487	297	218	149	99	95	79

The chart below shows the matching of empirical $\mu_n(k)$ to $\mu'_n(k)$ calculated from Lotka's law. $\mu'_n(k) = \frac{\mu_n(1)}{k^2}$, $k = 2, \cdots, 11$.

Figure 2.2: Lotka's law for bibliographic data from Emerald Group Publishing Limited



2.6.2 Lotka's law versus Zipf's law

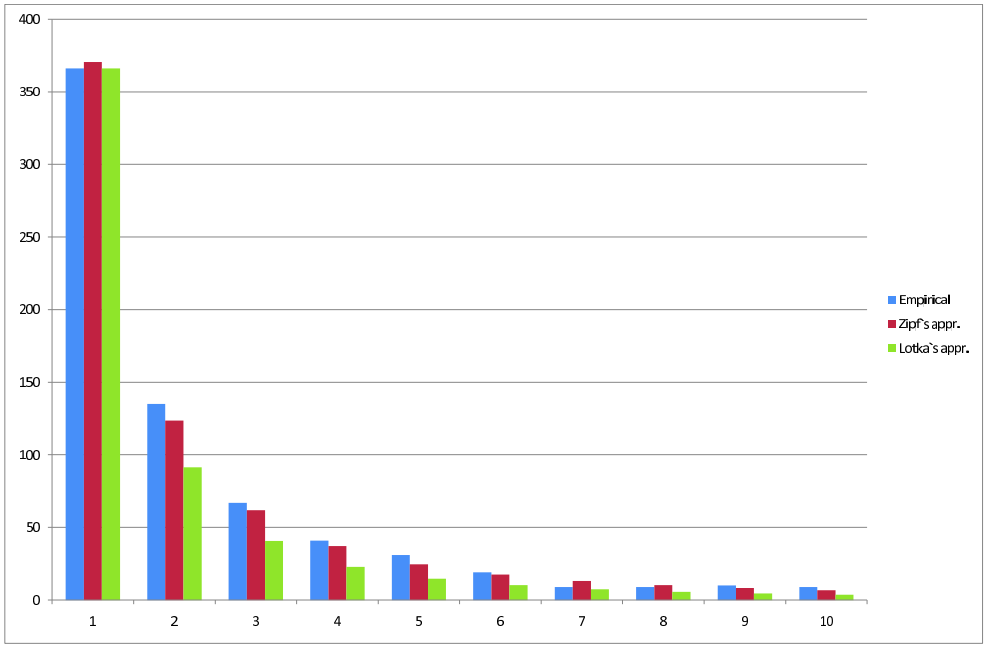
Lotka's law is not the only one which describes scientific productivity of authors in a given field. For example, Khmaladze and Tsigroshvili in [17] analyzed data collected by Dr I. Urinov. It contains the authors index of the *Theory of Probability and Applications* from 1955-1980. Khmaladze and Tsigroshvili reproduce this data jointly with the Zipf's approximation for $\mu_n(k)$ which is $\frac{\mu_n}{k(k+1)}$ with $\mu_n = 741$ being the total number of the authors during 25 years.

Table 2.3: Bibliographic data from *Theory of Probability and Applications*

$\mu_n = 741$									
$\mu_n(1)$	$\mu_n(2)$	$\mu_n(3)$	$\mu_n(4)$	$\mu_n(5)$	$\mu_n(6)$	$\mu_n(7)$	$\mu_n(8)$	$\mu_n(9)$	$\mu_n(10)$
366	135	67	41	31	19	9	9	10	9

In the chart below we demonstrate comparison of Zipf's and Lotka's approximations. Blue column correspond to empirical data, red column - Zipf's approximation and green column - Lotka's approximation. We can see clearly that Zipf's law gives much better matching.

Figure 2.3: Lotka's law vs Zipf's law



2.7 MacArthurs Stick Model

A simple way to study the structure of animal communities is to plot the rank of species from commonest to rarest along abscissa and their abundances along ordinate. The idea is to to predict curves on the basis of simple biological hypotheses such as there being equilibrium or near-equilibrium in population.

MacArthur in [22] considered following scheme:
"The environment is compared with a stick of unit length on which $n-1$ points are thrown at random. The stick is broken at these points and lengths of the n resulting segments are proportional to the abundance of the n species".

For uniform partition of $[0,1]$ interval, when ordered from smallest to largest, the expected length of the rth shortest interval is given by

$$\frac{1}{n}\sum_{i=1}^{r}\frac{1}{n-i+1}.$$

This is a natural ordering to use when listing the species in "order", from rarest to most common. The expected abundance of the rth rarest species among n species and m individuals is

$$\frac{m}{n}\sum_{i=1}^{r}\frac{1}{n-i+1}.$$

Quite frequently common species are too abundant and rare species are too rare, so the curve is very steep. These steep curves can be duplicated by considering the community as composed of two sticks of very different length (totaling unit length), each broken uniformly into $n/2$ pieces. That could be generalized to communities which are composed of several smaller ones, each obeying the original hypotheses.

2.8 Hill's model

B. M. Hill in [7] suggested the following model. Suppose n individuals are distributed to μ_n non-empty genera with Bose-Einstein distribution - all allocations are equiprobable with probability

$$(C_{n-1}^{\mu_n-1})^{-1}.$$

If $\mathbb{P}\{\frac{\mu_n}{n} \leq x\} \xrightarrow{w} F(x)$, $0 \leq x \leq 1$, $F(0) = 0$, then the number of genera with exactly k species $\mu_n(k)$ satisfy following expression

$$\frac{\mu_n(k)}{\mu_n} \xrightarrow{d} U(1-U)^{k-1}$$

where the random variable U has distribution F. Notice that if U has uniform distribution, then

$$\mathbb{E}\frac{\mu_n(k)}{\mu_n} \to \frac{1}{k(k+1)}$$

which is exactly Zipf's law.

2.9 Karlin-Rouault's Law

S. Karlin in [12] considered following scheme: n balls are thrown independently at a fixed infinite array of cells with probability p_i of hitting the i-th cell. Probabilities, without loss of generality, were ordered in such a way that $p_i \geq p_{i+1}$ for $i = 1, 2, \ldots$ and $p_i > 0$ for all i. Let ν_{in} be number of balls in the i-th cell after n tosses. Consider a Poisson process $\{N(t); t \in [0, \infty)\}$ with parameter 1 and let $\nu_{N(t)i}$ be number of balls in the i-th cell at time t. The stochastic process $\{\nu_{N(t)i;t\geq 0}\}$ for $i = 1, 2, \ldots$ is composed of mutually independent homogeneous Poisson processes with parameters p_i, $i = 1, 2, \ldots$ respectively

$$\mathbb{P}\{\nu_{N(t)i} = k\} = e^{-tp_i}\frac{tp_i^k}{k!}.$$

Notice that, in sharp contrast, the frequencies $\{\nu_{in}\}$ with n fixed and i varying are not mutually independent.

Karlin proved central limit theorems and determined the asymptotic behavior of the moments for several special functionals of the processes $\{\nu_{N(t)i;t\geq 0}\}_{i=1}^{\infty}$ and $\{\nu_{ni}\}_{i=1}^{\infty}$. What is more interesting for us, he considered $\mu_n(k)$ - the number of cells containing exactly k balls after n tosses and $\mu_n = \sum_{k=1}^{\infty}\mu_n(k)$ - number of occupied cells. The limit theorem he obtained is as follows:

$$\frac{\mu_n(k)}{\mu_n} \xrightarrow{P} \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}$$

where u is some parameter typically close to $\frac{1}{2}$.

A. Rouault in [30] considered certain Markov process as a model for formation of a text word by word and obtained a similar result to that of Karlin. This expression is called *Karlin-Rouault's Law*

2.10 Good-Turing Estimators

I.J. Good in [5] tried to answer the following question:
Suppose a random sample is drawn from an infinite set of outcomes Ξ. Let sample size be n, frequency of outcome ξ in the sample be ν_ξ and $\mu_n(k)$ be number of outcomes represented in the sample exactly k times, so that

$$\sum_{k=1}^{\infty} k\mu_n(k) = n.$$

Suppose number of different outcomes we have seen in the sample is μ_n, then what can one say about underlying probabilities $p(\omega)$, $\omega \in \Omega$?
I.J Good introduced following quantities

$$G_n(k) = \sum_{\xi\in\Xi} p(\xi) I\{\nu_\xi = k\}$$

and

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

These two quantities can be interpreted as follows. $G_n(k)$ is the total probability of outcomes represented in the sample k times and $p_n(k)$ is an "average" probability of each such outcome.

As an estimate for $p_n(k)$ one would take $\frac{k}{n}$, but this naive estimator leads to one awkward result: namely, for unseen outcomes in the sample it would imply the estimate $\widehat{G_n}(0) = 0$.

Assume that frequencies ν_ξ follow binomial distribution with parameters $p(\xi)$ and n, then

$$\mathbb{E}G_n(k) = \mathbb{E}\sum_{\xi\in\Xi} p(\xi)I\{\nu_\xi = k\}$$

$$= \sum_{\xi\in\Xi} p(\xi)\frac{n!}{k!(n-k)!}p^k(\xi)(1-p(\xi))^{n-k}$$

$$= \sum_{\xi\in\Xi} \frac{n!}{(k+1)!(n-k-1)!}p^{k+1}(\xi)(1-p(\xi))^{n-k-1}\frac{k+1}{n-k}(1-p(\xi))$$

but

$$\mathbb{E}\mu_n(k+1) = \sum_{\xi\in\Xi} \frac{n!}{(k+1)!(n-k-1)!}p^{k+1}(\xi)(1-p(\xi))^{n-k-1}$$

and

$$\frac{k+1}{n-k}(1-p(\xi)) \to \frac{k+1}{n}$$

as $n \to \infty$. So we obtained following expression:

$$\mathbb{E}G_n(k) \sim \frac{k+1}{n}\mathbb{E}\mu_n(k+1).$$

More precisely, as $p(\xi) \to 0$ as $n \to \infty$, we can assume that frequencies ν_ξ follow Poisson distribution with parameters $np(\xi)$.

$$\mathbb{E}G_n(k) = \mathbb{E}\sum_{\xi\in\Xi} p(\xi)I\{\nu_\xi = k\}$$

$$= \sum_{\xi\in\Xi} p(\xi)e^{np(\xi)}\frac{(np(\xi))^k}{k!}$$

$$= \sum_{\xi \in \Xi} p(\xi) e^{np(\xi)} \frac{(np(\xi))^{k+1}}{(k+1)!} \frac{k+1}{np} = \frac{k+1}{n} \mathbb{E}\mu_n(k+1)$$

so here we obtained exact equality:

$$\mathbb{E}G_n(k) = \frac{k+1}{n} \mathbb{E}\mu_n(k+1).$$

Based on this Good proposed to estimate $G_n(k)$ and $p_n(k)$ as

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

and

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

respectively.

Chapter 3

LNRE in questionnaires

3.1 On multinomial distributions with LNRE

Khmaladze and Tsigroshvili, in [17], obtained *the Karlin-Rouault law* in a context very different from that of Karlin and Rouault.
" Let unit interval be divided into two in the ratio $a : 1 - a$. Let each of these two be again sub-divided in the same ratio, and so on. On q-th step one obtains probabilities p_i, $i = 1, 2, \cdots, 2^q$, of the form $a^k(1-a)^{q-k}$ for some $k = 0, 1, \cdots q$. One can think of filling a questionnaire with "yes-no"-questions at random with probability a for one of the answers in each question". The p_i's defined above are the probabilities of each particular answer to the questions. It was proved that if $a \neq \frac{1}{2}$, then

$$\mu_n \to \infty,$$

but

$$\frac{\mu_n}{n} \to 0,$$

and that

$$\frac{\mu_n(k)}{\mu_n} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}.$$

The authors give to this statement very interesting heuristic interpretation. If each particular answer is regarded as an "opinion", then if $a \neq \frac{1}{2}$, the

proverb "as many men as many minds" is incorrect though number of "minds" is infinitely large it still is asymptotically smaller than number of "men". Notice that in this case we have definition $(d2)$ satisfied but not $(d1)$.

In 2009 in his paper "Diversity of responses in questionnaires and similar objects",[15], E.Khmaladze considered similar scheme with some generalization, namely he assumed that probabilities for "yes" or "no" for each question may be different. Investigation of limit for $\frac{\mu_n(k)}{\mu_n}$ becomes significantly more difficult, but using relatively transparent probabilistic tools, such as, large deviation theory and contiguity theory, the author showed that

$$\frac{\mu_n(k)}{\mu_n} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}$$

in other words he again obtained *Karlin-Rouault's Law*.

3.2 Formulation of the problem

We intend to replace the ratio of statistics, $\frac{\mu_n(k)}{\mu_n}$, with the ratio of their expected values, $\frac{\mathbb{E}\mu_n(k)}{\mathbb{E}\mu_n}$. To make it legitimate, in Section 5.1 we will prove that $\frac{\mu_n}{\mathbb{E}\mu_n}$ and $\frac{\mu_n(k)}{\mathbb{E}\mu_n(k)}$ converge to 1 a.s.

The scheme considered in [15] was following: A person is asked to fill in a form with q binary (yes/no) questions and probability that an answer to i-th question is "yes" is equal to a_i. So we have a vector $\overrightarrow{\xi} = (\xi_1, \xi_2 \cdots, \xi_q)$ of q Bernoulli random variables. The set Ξ_q of its possible values is the set of all sequences $\overrightarrow{x} = (0, 1, 0, \cdots, 0)$ of length q. It is obvious that cardinality of the set Ξ_q is 2^q, which means we have 2^q possible values of vector $\overrightarrow{\xi}$ with corresponding frequencies

$$\nu_q(\overrightarrow{x}) = \sum_{j=1}^{2^q} I\{\overrightarrow{\xi_j} = \overrightarrow{x}\}$$

Probability of a particular vector $\overrightarrow{x}$ ("opinion") could be written as follows:

$$p(\overrightarrow{x}) = \prod_{i=1}^{q} a_i^{x_i}(1-a_i)^{1-x_i}.$$

It is obvious that each frequency $\nu_q(\overrightarrow{x})$ has a binomial distribution with parameters n and $p(\overrightarrow{x})$, where $n = \lambda 2^q$. Therefore

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

and

$$\mathbb{E}\mu_q(k) = \sum_{\overrightarrow{x}\in\Xi_q} b(k, n, p(\overrightarrow{x})), \quad k = 1, 2, \cdots$$

Because $p(\overrightarrow{x}) \to 0$ as $q \to \infty$, the author assumes that frequencies $\nu_q(\overrightarrow{x})$ behave like Poisson random variables with parameter $\lambda 2^q p(\overrightarrow{x})$. Consequently, following expressions were used

$$\mathbb{E}\mu_q = \sum_{\overrightarrow{x}\in\Xi_q} (1 - \pi(0, np(\overrightarrow{x})))$$

and

$$\mathbb{E}\mu_q(k) = \sum_{\overrightarrow{x}\in\Xi_q} \pi(k, np(\overrightarrow{x})), \quad k = 1, 2, \cdots$$

The asymptotic behavior of these sums is extremely awkward; however, after a certain interpretation the author gave to the expression $2^q p(\overrightarrow{x})$, suddenly everything became relatively clear and transparent. Namely, he introduced a new, artificial measure $\mathbb{P}_{0q}$ of $\overrightarrow{\xi}$ on Ξ_q as follows

$$p_0(\overrightarrow{x}) = \frac{1}{2^q}$$

in other words he assumed that under measure $\mathbb{P}_{0q}$ all $a_i = \frac{1}{2}$.

The gain from this transformation is that it allowed one to consider an expression $2^q p(\overrightarrow{x})$ as a likelihood ratio of $\mathbb{P}_q$ and $\mathbb{P}_{0q}$

$$M_q(\overrightarrow{x}) \equiv 2^q p(\overrightarrow{x}) = \frac{p_q(\overrightarrow{x})}{p_{0q}(\overrightarrow{x})}.$$

Now $\mathbb{E}\mu_q$ and $\mathbb{E}\mu_q(k)$ could be written as following

$$\mathbb{E}\mu_q = 2^q \mathbb{E}_0(1 - \pi(0, \lambda M_q(\overrightarrow{\xi})))$$

and

$$\mathbb{E}\mu_q(k) = 2^q \mathbb{E}_0 \pi(k, \lambda M_q(\overrightarrow{\xi})), \quad k = 1, 2, \cdots$$

where $\mathbb{E}_0$ denotes expected value with respect to the new, uniform distribution $\mathbb{P}_{0q}$ of $\overrightarrow{\xi}$.

3.3 Contiguity approach

Before we start to investigate asymptotic behaviour of M_q, we would like to overview some basic facts and results from contiguity theory.

Definition 3. *The sequence $\mathbb{P}_q$ is contiguous with respect to the sequence $\mathbb{P}_{0q}$ if $\lim_{q\to\infty} \mathbb{P}_{0q}(A_q) = 0$ implies $\lim_{q\to\infty} \mathbb{P}_q(A_q) = 0$ for any sequence of measurable sets A_q. This one-sided contiguity is denoted by $\mathbb{P}_q \triangleleft \mathbb{P}_{0q}$.*

The sequences are said to be contiguous with respect to each other if both $\mathbb{P}_q \triangleleft \mathbb{P}_{0q}$ and $\mathbb{P}_{0q} \triangleleft \mathbb{P}_q$. This two-sided contiguity concept is denoted by $\mathbb{P}_q \triangleleft\triangleright \mathbb{P}_{0q}$ [28]

The Hellinger distance $H(\mathbb{P}, \mathbb{P}_0)$ between two probability measures $\mathbb{P}$ and $\mathbb{P}_0$ is defined as follows:

$$H(\mathbb{P}, \mathbb{P}_0) = (\int (\sqrt{p} - \sqrt{p_0})^2 d\mu)^{\frac{1}{2}} = (2 - 2\int \sqrt{p}\sqrt{p_0} d\mu)^{\frac{1}{2}}$$

where $p = \dfrac{d\mathbb{P}}{d\mu}$ and $p_0 = \dfrac{d\mathbb{P}_0}{d\mu}$ are corresponding Radon-Nikodym derivatives with respect to the σ-finite measure μ dominating $\mathbb{P} + \mathbb{P}_0$.

Suppose $\mathbb{P}_q = \prod_{i=1}^q P_{qi}$ and $\mathbb{P}_{0q} = \prod_{i=1}^q P_{0qi}$, so $\mathbb{P}_q$ and $\mathbb{P}_{0q}$ are product measures. Then, the Hellinger distance between product measures and that of their marginals are connected by the relationship

$$H^2(\mathbb{P}_q, \mathbb{P}_{0q}) = 2 - 2\prod_{i=1}^{q}(1 - \frac{1}{2}H^2(P_{qi}, P_{0qi})).$$

We formulate here two theorems, without proofs, from [28].

Theorem 1. *[28]* $\mathbb{P}_q \triangleleft \mathbb{P}_{0q}$ *iff*

$$\limsup_{q\to\infty} \sum_{i=1}^{q} H^2(P_{qi}, P_{0qi}) < \infty$$

and

$$\limsup_{q\to\infty} \sum_{i=1}^{q} P_{qi}(\frac{p_{qi}}{p_{0qi}}(X_{qi}) \geq c_q) = 0$$

whenever $c_q \to \infty$.

Consider log-likelihood $\mathbb{P}_q$ with respect to $\mathbb{P}_{0q}$

$$L_q = \sum_{i=1}^{q} \ln \frac{p_{qi}}{p_{0qi}}(X_{qi}).$$

Theorem 2. *[28] For a given* $\sigma \geq 0$

$$L_q \xrightarrow{w} \mathbb{N}(-\frac{1}{2}\sigma^2; \sigma^2)$$

under measure $\mathbb{P}_{0q}$ *and*

$$\lim_{q\to\infty} \max_{1\leq i\leq q} P_{0qi}(|\ln \frac{p_{qi}}{p_{0qi}}(X_{qi})| \geq \epsilon) = 0$$

for every $\epsilon > 0$, *iff for every* $\epsilon > 0$

$$\lim_{q\to\infty} \sum_{i=1}^{q} H^2(P_{qi}, P_{0qi}) = \frac{1}{4}\sigma^2$$

and

$$\lim_{q\to\infty} \sum_{i=1}^{q} \int_{|p_{0qi}-p_{qi}|\geq\epsilon p_{qi}} (\sqrt{p_{0qi}} - \sqrt{p_{qi}})^2 d\mu_{qi} = 0.$$

As an immediate consequence of those two theorems, Khmaladze stated following result:

Theorem 3. *[15] Suppose probabilities $a_{1q}, \cdots, a_{qq}$ form in q triangular array, such that* $\max_{1\le i\le q} |a_{iq} - \frac{1}{2}| \to 0$ *and*

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

Then

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

If the finite limit

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

exists, then

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

and

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

"Note that the result extends to very general class of distributions. Namely, whether $\xi_1, \cdots, \xi_q$ are independent and $\mathbb{P}_q$ is a product of Bernoulli distributions or not does not matter much."
Indeed, a_{iq} could stand for conditional probability of $\xi_i = 1$, given $\xi_1, \cdots, \xi_{i-1}$ and, in this case, conditions for asymptotic normality for $\ln M_q$ are well known (see, e.g,[6]).

3.4 Arbitrary underlying distribution. Large deviations approach

For arbitrary a_i-s, $i = 1, \cdots, q$, the behavior of the likelihood ratio M_q becomes unstable as, under $\mathbb{P}_{0q}$, $M_q \to 0$ in probability but $\mathbb{E}_0 M_q = 1$ which indicates that, although with very small probability, M_q takes extremely

large values. The asymptotic analysis of $\mathbb{E}_0(1-\pi(0,\lambda M_q))$ and $\mathbb{E}_0\pi(k,\lambda M_q)$ becomes again nontrivial.

" How small is the quantity $\mathbb{E}_0(1-\pi(0,\lambda M_q))$ 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." [15]

To avoid these difficulties, the author suggested following transformation: " 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 $1-\pi(0,e^x) = 1-e^{-e^x}$"[15]

These interpretations transforms $\mathbb{E}_0(1-\pi(0,\lambda M_q))$ into certain probability. Namely

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

where $L_q = \ln M_q$ and $\mathbb{P}_0$ denotes the joint distribution of L_q and η_1 under uniform distribution on Ξ_q.

Similarly if T_k is a Gamma-distributed random variable with shape parameter k and $\eta_k = \ln T_k$, we can write

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

These probabilities are naturally connected with the theory of large deviations and the author uses corresponding technique to establish their asymptotic. We will not go here in details but will just formulate the main theorem.

$$L_q = \ln \frac{p_q(\overrightarrow{\xi})}{p_{0q}(\overrightarrow{\xi})} = \sum_{i=1}^{q} [\xi_i \ln 2a_i + (1-\xi_i)\ln 2(1-a_i)]$$

where $\xi_1, \cdots, \xi_q$ are symmetric Bernoulli random variables under $\mathbb{P}_0$.

Let $\psi_i(u)$ denote the logarithm of the moment generating function of each summand

$$\psi_i(u) = \ln \mathbb{E}_0 e^{u[\xi_i \ln 2a_i + (1-\xi_i)\ln 2(1-a_i)]}$$

$$= \ln[(2a_i)^u + (2(1-a_i))^u] - \ln 2.$$

Let

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

and suppose as $q \to \infty$, for all $a \in [0,1]$

$$F_q(a) \to F(a)$$

and

$$\int_0^1 \ln^2 \frac{a}{1-a} dF_q(a) \to \int_0^1 \ln^2 \frac{a}{1-a} dF(a).$$

Notice that, the last condition excludes the possibility of having too many a_i-s too close to 0 or 1. Now we can formulate the theorem given in [15].

Theorem 4. *[15]*

If

$$\int_0^1 \ln^2 \frac{a}{1-a} dF_q(a) \to \int_0^1 \ln^2 \frac{a}{1-a} dF(a)$$

Then

$$\frac{\mathbb{E}\mu_q(k)}{\mathbb{E}\mu_q} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)}$$

where $u = \lim u_q$ *and* u_q *is such that* $\sum_{i=1}^{q} \psi'(u_q) = 0$.

3.5 Remark on Good-Turing Estimation

As an immediate consequence of previous two theorems, Khmaladze obtained estimators for quantities which I.J.Good introduced in early 1950's. Namely, recall that $G_q(k)$ is the total probability of outcomes that happened to appear k times and $p_q(k)$ is an "average" probability of each

of such outcomes. Then if a sample agrees with the Karlin-Rauault law, Khmaladze showed, that

$$\mathbb{E}G_q(k) \sim \frac{1}{n}\frac{u\Gamma(k+1-u)}{\Gamma(1-u)\Gamma(k+1)}\mathbb{E}\mu_q$$

and

$$p_q(k) \sim \frac{k-u}{n},$$

and thus

$$\mathbb{E}G_q(0) \sim \frac{u}{n}\mathbb{E}\mu_q$$

and

$$p_q(0) \sim \frac{u}{n}\frac{\mathbb{E}\mu_q}{2^q - \mathbb{E}\mu_q}$$

as $q \to \infty$.

"This, in 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 $\frac{k-u}{n}$, while the total probability of unseen outcomes was $\frac{u\mu_q}{n}$." [15]

In [15] it was shown that more accurate and complete evaluation of overall behaviour of probabilities is possible.

Let

$$H_q(x) = \frac{1}{2^q}G_q(x)$$

where

$$G_q(x) = \sum_{\overrightarrow{x}\in\Xi_q} I\{np(\overrightarrow{x}) > x\}.$$

And as in Section 2.1 of Chapter 2, let

$$R_q(x) = \frac{1}{\int_0^\infty (1-e^y)dH_q(y)}H_q(x),$$

or equivalently,

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

be the tail of empirical distribution function of $np(\overrightarrow{x})$ under $\mathbb{P}_{0q}$ and its normalized form respectively.

Khmaladze in [15] obtained in a sense a very strange result. The dependence of the spread of probabilities $p(\overrightarrow{x})$ on distribution function F of individual probabilities $a_1, \cdots, a_q$ is very weak. It essentially depends on very narrow class of functions through the parameter u, which in itself is quite stable as it varies around 0.5 for whatever Beta distribution of $a_1, \cdots, a_q$.

Theorem 5. *[15]*

If, as $q \to \infty$ and sample size $n = \lambda 2^q$ with $\lambda = const$,

$$\frac{\mu_q(k)}{\mu_q} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)},$$

then for all $x > 0$

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

Chapter 4

Measures of Diversity

4.1 Diversity as a property of a population

As we mentioned in Section 1.1, in any statistical analysis devoted to the study of variety of species, one will observe the presence of a large number of rare events. Diversity in populations is of particular interest to ecologists.

As an illustration, we quote Williams from [37]: "If one goes into a natural forest in a cold temperate climate such as Northern Europe and selects at random two trees, the chances are high that both will belong to the same species, because in such an environment the vegetation is undiversified. If one makes the same experiment in a tropical forest, it may be necessary to select quite a number of pairs before getting two of the same kind: here the vegetation is highly diverse. If one collected a few thousand mosquitoes in the far north, it is likely that only a few species would be represented, but in the tropics forty or fifty species might easily be found in a sample of the same size."

Magguran in [23] partitions biological diversity into two components: species richness and evenness. The term "species richness" was coined by McIntosh in [25] and represents one of the oldest and most intuitive measures of biological diversity. Species richness is simply the number of

species in the particular study. A community in which all species have approximately equal number of individuals would be considered as extremely even. A large disparity in the relative abundances of species would result in the description "uneven"

A "diversity index" or "diversity measure" or heterogeneity (see,e.g., [5]) is a measure that incorporates information on richness and evenness and therefore takes species abundances into account.

4.2 The Shannon diversity index and evenness measure

In the theory of probability an entropy is a quantity which measures an indeterminacy of distribution. Let $(p_1, p_2, \cdots, p_N)$ be a finite probability distribution. The entropy, or Shannon's index, of this distribution is defined as follows:

$$H = -\sum_{i=1}^{N} p_i \ln p_i.$$

It is clear that $H \geq 0$, and $H = 0$ if and only if every p_i, with one exception, is equal to zero and "exception" is equal to 1. The function $-p \ln p$, on the interval $[0, 1]$ is convex, therefore

$$H = -\sum_{i=1}^{N} p_i \ln p_i \leq -N \frac{\sum_{i=1}^{N} p_i}{N} \ln \frac{\sum_{i=1}^{N} p_i}{N} = \ln N.$$

In other words, the entropy attains its maximum for $p_1 = p_2 = \cdots = p_N = \frac{1}{N}$.

The Shannon's index is widely used to measure biological diversity. Let n be a sample size and N be number of different outcomes. The Shannon index is estimated as follows:

$$\hat{H}_n = -\sum_{i=1}^{N} \frac{\nu_i}{n} \ln \frac{\nu_i}{n}$$

where the quantity ν_i is the number of occurrences of i-th outcome. Some commentators who discuss this measure underline disadvantages of it. For example, Sager and Hasler in [32] complain that Shannon's index is inadequate because it is insensitive to rare species which may play important role in the ecosystem. This argument is incorrect as diversity index can not in itself incorporate species importance.

As a heterogeneity measure the Shannon index takes into account the degree of evenness in species abundances. Pielou in [29] suggested calculating a separate evenness measure based on following arguments. The maximum diversity $(\hat{H}_n)_{max}$ that could possibly occur would be found in a situation where all species had equal abundances, in other words if $\hat{H}_n = (\hat{H}_n)_{max} = \ln N$. The ratio of observed diversity to maximum diversity can therefore interpreted as a measure for evenness (see, e.g., [23])

$$\hat{J}_n = \frac{\hat{H}_n}{(\hat{H}_n)_{max}} = \frac{\hat{H}_n}{\ln N}.$$

Hurlbert in [8] uses the value $(\hat{H}_n)_{min}$, instead of $(\hat{H}_n)_{max}$, as a measure of Shannon evenness. For a given N and n, a simple method to calculate $(\hat{H}_n)_{min}$ is given in [2].

$$\begin{aligned}
\hat{H}_n &= -\sum_{i=1}^{N} \frac{\nu_i}{n} \ln \frac{\nu_i}{n} \\
&= -\frac{1}{n} \sum_{i=1}^{N} \nu_i (\ln \nu_i - \ln n) \\
&= -\frac{1}{n} \sum_{i=1}^{N} \nu_i \ln \nu_i + \frac{\ln n}{n} \sum_{i=1}^{N} \nu_i \\
&= \ln n - -\frac{1}{n} \sum_{i=1}^{N} \nu_i \ln \nu_i.
\end{aligned}$$

The lower limit of $\hat{H}_n$ corresponds to the case for which $\nu_i = n - (N-1)$ for one i-th outcome and $\nu_j = 1$ for all the rest $j \neq i$ outcomes. The $N-1$

outcomes j with $\nu_j = 1$ do not contribute to the calculation. Thus minimal diversity value depends only on the outcome i with $n-(N-1)$ individuals. Substituting ν_i with $n - (N - 1)$ we will obtain

$$(\hat{H}_n)_{min} = \ln n - \frac{(n - N + 1)\ln(n - N + 1)}{N}.$$

4.3 McIntosh's measure of diversity

McIntosh in [25] treated a population as point in a N-dimensional space and used Euclidean distance from the origin to define a measure of diversity. It is calculated as follows:

$$U = \sqrt{\sum n_i^2}$$

where, again, n_i is number of i-th species. And a measure of diversity is

$$D = \frac{n - U}{n - \sqrt{n}}$$

(see,e.g, [23]). And a further evenness measure can be obtained from the formula:

$$E = \frac{n - U}{n - n/\sqrt{N}}$$

[29].

4.4 Simpson's diversity index and measure of evenness

One of the simplest diversity measures was suggested by Simpson in [34]: it is the probability of any two individuals drawn at random from an infinitely large community belonging to the same species, or equivalently "It is calculation of the number of pairs that would have to be selected

at random from a particular population in order to give an even chance of getting one pair with both individuals belonging to the same species“ [37].

$$D = \sum p_i^2$$

where p_i is the probability of i-th species.

Another form

$$D = \sum \frac{\nu_i(\nu_i - 1)}{n(n-1)}$$

where n is a total number of individuals and ν_i is a frequency of i-th species [23].

Simpson‘s index is usually expressed as $1 - D$ or $1/D$. The last one could be interpreted as an expected value of geometric distribution with parameter D, that is, an average number of trials before we get a pair with both individuals belonging to the same species. This index is heavily weighted towards the most abundant species in the sample, while being less sensitive to species richness.[23]

Simpson‘s measure of evenness can be calculated by dividing the reciprocal form of Simpson‘s index by the number of species in the sample N

$$E_{1/D} = \frac{(1/D)}{N}.$$

Chapter 5

Results

5.1 Preliminary analysis and discussion

Consider again N disjoint events with probabilities $p_1, p_2, ..., p_N$, $\sum_{i=1}^{N} p_i = 1$ and let $\nu_n = (\nu_{1n}, ..., \nu_{Nn})$ be the vector of frequencies of these events in n independent trials. The so-called "spectral statistics" or "empirical vocabulary" (see, e.g. [1], [14] and [16]) are defined by,

$$\mu_n(m) = \sum_{i=1}^{N} I\{\nu_{in} = m\}, \;\; m = 1, ..., n \;,$$

$$\mu_n = \sum_{i=1}^{N} I\{\nu_{in} \geq 1\} = \sum_{m=1}^{n} \mu_n(m) \;.$$

Before we start analysis of the spectral statistics, we will prove that the ratios $\frac{\mu_n}{\mathbb{E}\mu_n}$ and $\frac{\mu_n(k)}{\mathbb{E}\mu_n(k)}$ converge to 1 a.s. as $n \to \infty$.

Theorem 6. *If for every $k = 1, 2, \ldots$, $\mathbb{E}\mu_n(k)$ and $\mathbb{E}\mu_n \geq a(\ln n)^2$, then*

$$\frac{\mu_n}{\mathbb{E}\mu_n} \to 1 \; a.s.$$

and

$$\frac{\mu_n(k)}{\mathbb{E}\mu_n(k)} \to 1 \; a.s.$$

Remark: Later in Lemma 6 of Section 5.6 we will show that actually $\mathbb{E}\mu_n(k)$ and $\mathbb{E}\mu_n$ are of order of N^u, $0 < u < 1$, which is much stronger than what is required in the theorem.

Proof:

$$\mu_n(k) = \sum_{i=1}^{N} I\{\nu_{in} = k\} \equiv \sum_{i=1}^{N} X_i$$

$$\mathbb{E}\mu_n(k) = \mathbb{E}\sum_{i=1}^{N} I\{\nu_{in} = k\} = \mathbb{E}\sum_{i=1}^{N} X_i$$

$$A_n^{\varepsilon} \equiv \{\omega :| \frac{\mu_n(k)}{\mathbb{E}\mu_n(k)} - 1 |\geq \varepsilon\}$$

$$= \{\omega :| \sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i |\geq \varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\}.$$

Using exponential inequality we can obtain:

$$\mathbb{P}\{(\sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i) \geq \varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\}$$

$$= \mathbb{P}\{e^{t\sum_{i=1}^{N} X_i} \geq e^{(\varepsilon+1)t\mathbb{E}\sum_{i=1}^{N} X_i}\} \leq \frac{\mathbb{E}e^{t\sum_{i=1}^{N} X_i}}{e^{(\varepsilon+1)t\mathbb{E}\sum_{i=1}^{N} X_i}}$$

with

$$\mathbb{E}e^{t\sum_{i=1}^{N} X_i} = \prod_{i=1}^{N}(1 - q_i + q_i e^t)$$

where

$$q_i = \mathbb{P}\{\nu_{in} = k\}.$$

Using

$$\mathbb{E}e^{t\sum_{i=1}^{N} X_i} = e^{\sum_{i=1}^{N} \ln(1-q_i+q_i e^t)}$$

we obtain

$$\mathbb{P}\{(\sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i) \geq \varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\}$$
$$\leq e^{\sum_{i=1}^{N} \ln(1-q_i+q_i e^t)-(\varepsilon+1)t\mathbb{E}\sum_{i=1}^{N} X_i}$$

$$\leq e^{\sum_{i=1}^{N}(q_i e^t - q_i) - (\varepsilon+1)t\mathbb{E}\sum_{i=1}^{N} X_i}$$

$$= e^{(e^t - 1 - (\varepsilon+1)t)\mathbb{E}\sum_{i=1}^{N} X_i}.$$

So far, parameter t was arbitrary. After minimizing in t, at $t = \ln(\varepsilon + 1)$, we will obtain:

$$\mathbb{P}\{(\sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i) \geq \varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\}$$

$$\leq e^{(\varepsilon+1-1-(\varepsilon+1)\ln(\varepsilon+1))\mathbb{E}\sum_{i=1}^{N} X_i}$$

$$= e^{(\varepsilon-(\varepsilon+1)\ln(\varepsilon+1))\mathbb{E}\sum_{i=1}^{N} X_i}$$

$$= e^{(\varepsilon-(\varepsilon+1)\ln(\varepsilon+1))\mathbb{E}\mu_n(k)}.$$

Denote $(\varepsilon-(\varepsilon+1)\ln(\varepsilon+1))$ with $-\alpha$, then using condition from theorem we obtain,

$$\mathbb{P}\{(\sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i) \geq \varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\}$$

$$\leq e^{-\alpha a(\ln n)^2} = (e^{\ln n})^{-\alpha a \ln n} = n^{-\alpha a \ln n}.$$

Similarly, we can prove that

$$\mathbb{P}\{(\sum_{i=1}^{N} X_i - \mathbb{E}\sum_{i=1}^{N} X_i) \leq -\varepsilon\mathbb{E}\sum_{i=1}^{N} X_i\} \leq n^{-\alpha a \ln n}$$

therefore

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n^{\varepsilon}) \leq \infty$$

which implies that $\frac{\mu_n(k)}{\mathbb{E}\mu_n(k)}$ converge to 1 a.s.. Similarly we can show that $\frac{\mu_n}{\mathbb{E}\mu_n}$ converge to 1 a.s.

□

From now on we concentrate on asymptotic behaviour of $\frac{\mathbb{E}\mu_n(k)}{\mathbb{E}\mu_n}$ instead of $\frac{\mu_n(k)}{\mu_n}$.

We are interested in a specific framework in which the disjoint events can be indexed by q-dimensional vectors $\vec{x}_q = (x_1, \ldots, x_i, \ldots, x_q)$ with coordinates x_i changing from 1 to k_i respectively. Then $p_i (i = 1, ..., N)$ in previous setting becomes $p(\vec{x}_q)$, with $\vec{x}_q \in \Xi_q = \times_{i=1}^{q} \{1, \ldots, k_i\}$ and $N = \prod_{i=1}^{q} k_i$ is the cardinality of Ξ_q. Hence $\mu_n(m)$ becomes,

$$\mu_n(m) = \sum_{\vec{x}_q \in \Xi_q} I\{\nu(\vec{x}_q) = m\}.$$

Therefore,

$$\mathbb{E}\mu_n(m) = \sum_{\vec{x}_q \in \Xi_q} \mathbb{P}\{\nu(\vec{x}_q) = m\}$$

$$= \sum_{\vec{x}_q \in \Xi_q} \binom{n}{m} p(\vec{x}_q)^m \left(1 - p(\vec{x}_q)\right)^{n-m}$$

and

$$\mathbb{E}\mu_n = \sum_{\vec{x}_q \in \Xi_q} \left(1 - (1 - p(\vec{x}_q))^n\right).$$

This framework can be found in many applications. For example, $\vec{x}_q$ can be interpreted as an opinion in a questionnaire with q multi-option or multi-choice questions (the i-th question has k_i options). And the ratios,

$$\frac{\mathbb{E}\mu_n(m)}{n} \quad \text{and} \quad \frac{\mathbb{E}\mu_n(m)}{\mathbb{E}\mu_n} \tag{5.1}$$

can be interpreted as, the proportion of the number of opinions with m supporters in all n responses, and in the total number of opinions with at least 1 supporter, respectively.

The main setting of the framework was given in [15]. In that paper all x_i were binary. However, in this chapter, we want to take advantage of the fact that the proofs given by Khmaladze are of more general nature. We demonstrate this by extending settings for questionnaires with a general structure.

5.2 Formulation of problem

There are three variables n, q and N, which control the asymptotic behavior of the ratios. Among them, q and N are directly associated with each other by the definition of N. Therefore, it is sufficient that we discuss the relation between n and N.

Since N is the number of disjoint events (opinions) and n is sample size (the number of responses), when $n = o(N)$, most frequencies tend to zero and those nonzero frequencies will mostly be 1. On the other hand, in the situation of $N = o(n)$, most frequencies are nonzero and tend to infinity eventually. However, it is more interesting to investigate the situation where N and n are comparable, particularly, $n = \lambda N$ for some constant $0 < \lambda < \infty$. In this chapter we only focus on the last case.

Let $\mathbf{P}_q$ denote the probability measure on Ξ_q which is defined by probabilities $p(\vec{x}_q)$:

$$\mathbf{P}_q\left(\vec{X}_q = \vec{x}_q\right) = p(\vec{x}_q)$$

and let $\mathbf{P}_{0,q}$ denote the uniform measure on Ξ_q:

$$\mathbf{P}_{0,q}\left(\vec{X}_q = \vec{x}_q\right) = p_0(\vec{x}_q) = \frac{1}{N}.$$

Then,

$$\begin{aligned}\mathbb{E}\mu_n(m) &= \sum_{\vec{x}_q \in \Xi_q} \binom{n}{m} p(\vec{x}_q)^m \left(1 - p(\vec{x}_q)\right)^{n-m} \\ &= N\mathbb{E}_{\mathbf{P}_{0,q}}\left[\binom{n}{m} p(\vec{X}_q)^m \left(1 - p(\vec{X}_q)\right)^{n-m}\right]\end{aligned}$$

and

$$\mathbb{E}\mu_n = N\mathbb{E}_{\mathbf{P}_{0,q}}\left[1 - \left(1 - p(\vec{X}_q)\right)^n\right].$$

Define M_q as likelihood ratio of the alternative measure $\mathbf{P}_q$ to null measure $\mathbf{P}_{0,q}$, i.e.

$$M_q(\vec{x}_q) = \frac{d\mathbf{P}_q}{d\mathbf{P}_{0,q}}(\vec{x}_q) = \frac{p(\vec{x}_q)}{p_0(\vec{x}_q)} = Np(\vec{x}_q). \tag{5.2}$$

Then we have

$$\mathbb{E}\mu_n(m) = N\binom{n}{m}\frac{1}{n^m}\mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m\left(1-\frac{\lambda M_q(\vec{X}_q)}{n}\right)^{n-m}\right] \tag{5.3}$$

and

$$\mathbb{E}\mu_n = N\left(1-\mathbb{E}_{\mathbf{P}_{0,q}}\left[\left(1-\frac{\lambda M_q(\vec{X}_q)}{n}\right)^n\right]\right). \tag{5.4}$$

At first sight, it looks artificial that we introduce such a likelihood ratio M_q. However, the benefit of this introduction is significant. Although "physical" measure of $\vec{X}_q$ is $\mathbf{P}_q$, using M_q we can exploit its asymptotic properties as if $\vec{X}_q$ has uniform distribution $\mathbf{P}_{0,q}$. As a likelihood ratio and a martingale in q, $M_q(\vec{X}_q)$ possesses some good and well-known asymptotic properties, which is very convenient.

Further, according to the Lemma5 below, expressions in the right hand side of (5.3) and (5.4) can be replaced by Poissonian limits. This suggests that we can lay aside the role of sample size n in the asymptotic behavior of the ratios and focus on the limit behavior of distribution of $M_q(\vec{X}_q)$, or equivalently, $Np(\vec{X}_q)$ under the measure induced by $\mathbf{P}_{0,q}$.

Lemma 5. *For $M_q(\vec{X}_q)$ defined by (5.2),*

$$\mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m\left(1-\frac{\lambda M_q(\vec{X}_q)}{n}\right)^{n-m}\right] = \mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m e^{-\lambda M_q(\vec{X}_q)}\right] + O\left(\frac{1}{n}\right).$$

Proof: Let us denote the distribution function of $\lambda M_q(\vec{X}_q)$ with $F_{M_q}(x)$. Since $0 \leqslant \lambda M_q(\vec{X}_q) \leqslant n$,

$$\sup_{0\leqslant x\leqslant n}\left|x^m\left(1-\frac{x}{n}\right)^{n-m} - x^m e^{-x}\right| = O\left(\frac{1}{n}\right),$$

$$\left|\mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m\left(1-\frac{\lambda M_q(\vec{X}_q)}{n}\right)^{n-m}\right] - \mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m e^{-\lambda M_q(\vec{X}_q)}\right]\right|$$
$$\leqslant \int_0^\infty \left|x^m\left(1-\frac{x}{n}\right)^{n-m} - x^m e^{-x}\right| dF_{M_q}(x) \leqslant O\left(\frac{1}{n}\right).$$

□

5.3 On the probabilities of large deviations

In this subsection we give a brief overview of the problem known as a probabilities of large deviations. We will discuss the motivation of the problem and explain, in a very simple way, the technique used to solve the problem. Consider i.i.d. Bernoulli random variables ξ_i, $i = 1, ..., n$ with parameter p. Let $S_n = \sum_{i=1}^{n} \xi_i$ and F_n be the distribution function of $\frac{S_n - np}{\sqrt{np(1-p)}}$. It is well known that $\dfrac{S_n - np}{\sqrt{np(1-p)}}$ converges, in distribution, to standard normal random variable,

$$\sup_{-\infty \leq x \leq \infty} \mid F_n(x) - \Phi(x) \mid \to 0$$

but at the same time we know that, this information is valuable for moderate values of x, as for large x both $F_n(x)$ and $\Phi(x)$ are close to unity and the statement of the limit theorem loses its power. (see,e.g, [4])
In many cases we would like to consider following ratio:

$$\frac{1 - F_n(x)}{1 - \Phi(x)}.$$

When x is fixed and $n \to \infty$ this ratio converges to 1, but when x increases along with n the limit of the ratio is not 1 any more.
For example, suppose we have a sum of 100 Bernoulli random variables with parameter $p = 0.1$ and we are interested in the following probability $\mathbb{P}\{S_n \geq 20\}$. Direct calculation of this binomial probability gives us value 0.0008075739, while calculating corresponding normal probability $\mathbb{P}\{Y \geq 3.3\}$ gives us value 0.0004290603. As we see, both of these probabilities are small, as they should be, but at the same time one is twice as big as the other. Applying so called continuity correction gives us even worse approximation, 0.000232691. Now the question is: can we approximate binomial probability any better?

Let's write down the probability we want to calculate:

$$\mathbb{P}\{S_n \geq k\} = \sum_{l=k}^{n} b(l; p, n) = \sum_{l=k}^{n} C_l^n p^l (1-p)^{n-l}.$$

Using simple modifications we can rewrite it as follows:

$$\mathbb{P}\{S_n \geq k\} = \sum_{l=k}^{n} C_l^n (\frac{p}{q})^l (\frac{1-p}{1-q})^{n-l} q^l (1-q)^{n-l}$$

$$= (\frac{1-p}{1-q})^n \sum_{l=k}^{n} (\frac{p(1-q)}{q(1-p)})^l C_l^n q^l (1-q)^{n-l}$$

$$= (\frac{1-p}{1-q})^n \sum_{l=k}^{n} (\frac{p(1-q)}{q(1-p)})^l b(l; q, n),$$

$$\mathbb{P}\{S_n \geq k\} = (\frac{1-p}{1-q})^n \sum_{l=k}^{n} (\frac{p(1-q)}{q(1-p)})^l b(l; q, n).$$

So we ended up with another binomial distribution with parameters q and n. Notice, that here we are free in choice of q, we could make it whatever we want, but we will make it equal to $\frac{k}{n}$, which will make expected value of new binomial random variable equal to k where the limit theorem gives the best approximation. Using limit theorem for the new binomial distribution, denoting $\alpha = \frac{p(1-q)}{q(1-p)}$ and $c = \sqrt{nq(1-q)} \ln \alpha$ we will obtain:

$$\mathbb{P}\{S_n \geq k\} = (\frac{1-p}{1-q})^n \sum_{l=k}^{n} (\frac{p(1-q)}{q(1-p)})^l b(l; q, n)$$

$$= (\frac{1-p}{1-q})^n \int_{\frac{k-nq}{\sqrt{nq(1-q)}}}^{\infty} \alpha^{(x\sqrt{nq(1-q)}+nq)} e^{-\frac{x^2}{2}} dx$$

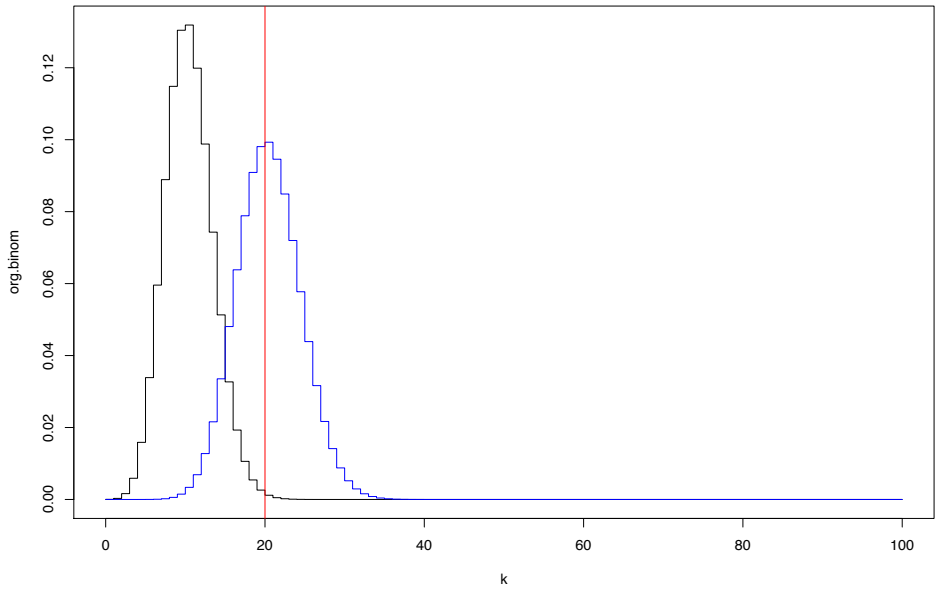
$$= (\frac{1-p}{1-q})^n \int_{\frac{k-nq}{\sqrt{nq(1-q)}}}^{\infty} e^{(x\sqrt{nq(1-q)}+nq)\ln\alpha - \frac{x^2}{2}} dx$$

$$= (\frac{1-p}{1-q})^n \int_{\frac{k-nq}{\sqrt{nq(1-q)}}}^{\infty} e^{-\frac{x^2 - 2x\sqrt{nq(1-q)}\ln\alpha}{2} + nq\ln\alpha} dx$$

$$= (\frac{1-p}{1-q})^n e^{nq\ln\alpha} e^{\frac{nq(1-q)\ln^2\alpha}{2}} \int_{\frac{k-nq}{\sqrt{nq(1-q)}}}^{\infty} e^{-\frac{x^2 - 2x\sqrt{nq(1-q)}\ln\alpha + nq(1-q)\ln^2\alpha}{2}}$$

$$= (\frac{1-p}{1-q})^n \alpha^{nq} e^{\frac{c^2}{2}} \int_{\frac{k-nq}{\sqrt{nq(1-q)}}}^{\infty} e^{-\frac{(x-c)^2}{2}}$$

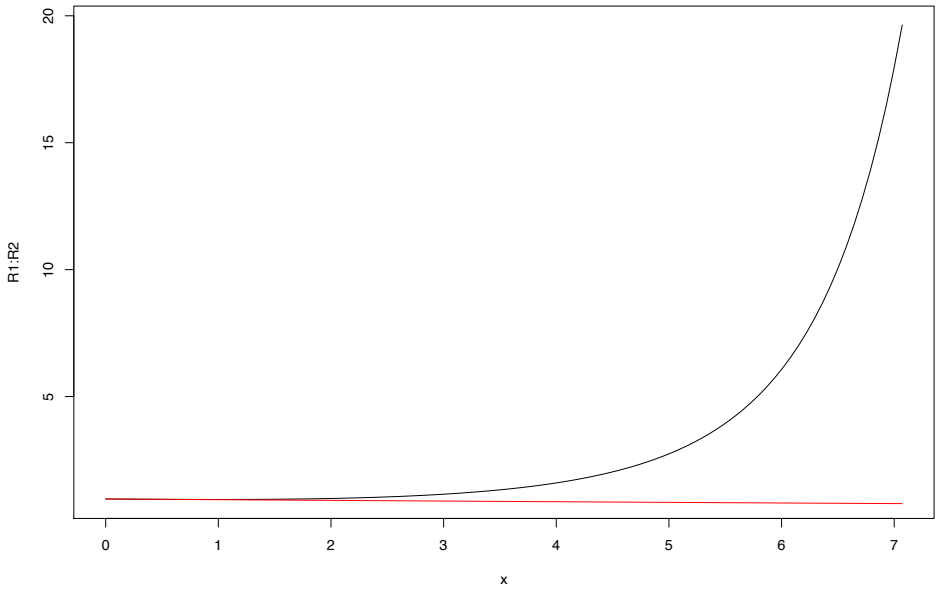
Figure 5.1: Shifting the distribution using Esscher transform



$$= (\frac{1-p}{1-q})^n \alpha^{nq} e^{\frac{c^2}{2}} [1 - \Phi(\frac{k-nq}{\sqrt{nq(1-q)}} - c)] \equiv \mathbb{P}_{LD}\{S_n \geq k\}.$$

Let R_1 be the ratio of $\mathbb{P}\{S_n \geq k\}$ and $1 - \Phi(x)$, where $x = \dfrac{k-np}{\sqrt{np(1-p)}}$ and R_2 be the ratio of $\mathbb{P}\{S_n \geq k\}$ and $\mathbb{P}_{LD}\{S_n \geq k\}$. The following graph illustrates what happens to these ratios when k, and respectively x, increases. For this experiment we assumed that $n = 2000$, $p = 0.1$ and k changes from 200 to 300, or x changes from 0 to approximately 7. We can clearly see that for large x, in this case when $x > 4$, $\mathbb{P}_{LD}\{S_n \geq k\}$ approximates $\mathbb{P}\{S_n \geq k\}$ much better then $1 - \Phi(x)$. One would say that we do not need to consider probabilities which are that small; but in finance, for example, it is very common to deal with risks which have probabilities as small as 10^{-7}. This figure shows the graph of $R_1 = \dfrac{\mathbb{P}\{S_n \geq k\}}{1 - \Phi(x)}$, where

Figure 5.2: Comparison of CLT and large deviations approach



$$x = \frac{k - np}{\sqrt{np(1-p)}} \text{ and } R_2 = \frac{\mathbb{P}\{S_n \geq k\}}{\mathbb{P}_{LD}\{S_n \geq k\}}.$$

Now let us consider arbitrary i.i.d. random variables ξ_i, $i = 1, ..., n$ such that $\mathbb{E}\xi_i = 0$ and $\mathbb{E}(\xi_i)^2 = \sigma^2$. Let $S_n = \sum_{i=1}^{n} \xi_i$ with distribution F_n let $\varphi(s) = \ln \mathbb{E}e^{s\xi_i}$ be cumulant-generating function of random variable ξ_i, then Esscher's transform is defined as follows:

$$\frac{dG_n}{dF_n}(x) = e^{sx - n\varphi(s)}.$$

To make sure that G_n is a probability distribution, we need to show that

$$\int_{-\infty}^{\infty} dG_n(x) = 1.$$

Indeed,

$$\int_{-\infty}^{\infty} dG_n(x) = \int_{-\infty}^{\infty} e^{sx-n\varphi(s)} dF_n(x)$$

$$= e^{-n\varphi(s)} \int_{-\infty}^{\infty} e^{sx} dF_n(x) = e^{-n\varphi(s)} e^{n\varphi(s)} = 1.$$

So, we introduced new distribution which we now could approximate by the corresponding normal distribution. The gain here is that, unlike for the previous distribution F_n, relative error committed in this approximation is minimal.

$$P\{S_n \geq k\} = e^{n\varphi(s)} \sum_{l=k}^{n} e^{-sx} Q_s\{S_n = l\}.$$

Suppose now S_n is the sum of arbitrary i.i.d random variables, the cumulant-generating function is $n\varphi(s)$, where $\varphi(s)$ is a cumulant-generating function of each random variable under initial $\mathbb{P}$ measure, or $\varphi(s) = \ln(1 - p + pe^s)$. Cumulant-generating function of each random variable under shifted measure Q will be,

$$\psi(r) = \ln \sum_{l=0}^{n} e^{rl} Q_s\{S_n = l\} = \ln \sum_{l=0}^{n} e^{rl} e^{sl-\varphi(s)} P\{S_n = l\}$$

$$= \ln \sum_{l=0}^{n} e^{l(r+s)-\varphi(s)} P\{S_n = l\} = \varphi(r+s) - \varphi(s).$$

5.4 The structure of $p(\vec{x}_q)$

By definition, $p(\vec{x}_q)$ is the probability of $\{\vec{X}_q = \vec{x}_q\}$, and we can define

$$a_i(j) = \mathbf{P}_q\left(X_i = j\right)$$

to be the probability of answering "j" to i-th question. In the cases that $X_1, \ldots, X_q$ are independent,

$$p(\vec{x}_q) = \prod_{i=1}^{q} a_i(x_i).$$

Correspondingly,

$$M_q(\vec{x}_q) = Np(\vec{x}_q) = \prod_{i=1}^{q} k_i a_i(x_i).$$

If we consider,

$$\xi_i = \ln(k_i a_i(X_i)),$$

then we can define,

$$L_q(\vec{X}_q) = \ln M_q(\vec{X}_q) = \sum_{i=1}^{q} \ln(k_i a_i(X_i)) = \sum_{i=1}^{q} \xi_i.$$

In principle, discussions based on M_q and L_q are equivalent. Since L_q can be expressed as a sum of q random variables, it is more convenient to discuss the limit distribution of L_q. Let us call a questionnaire "neutral" if the distribution of each X_i is uniform on its possible values. In this case $a_i(x_i) = \dfrac{1}{k_i}$ and there is no need to study M_q , as it is simply 1. In this case the limits of the ratios are:

$$\frac{\mathbb{E}\mu_n(1)}{n} \to e^{-\lambda}$$

and

$$\frac{\mathbb{E}\mu_n(m)}{\mathbb{E}\mu_n} = \frac{N\binom{n}{m}\frac{1}{n^m}\lambda^m(1-\frac{\lambda}{n})^{n-m}}{N(1-(1-\frac{\lambda}{n})^n)} \to \frac{\lambda^m e^{-\lambda}}{m!(1-e^{-\lambda})}.$$

Note that $\mathbb{E}\mu_n(1) \sim n$ in this case, and hence the frequencies defined here form a sequence of large number of rare events in sense of both d1 and d2, of section 1.2.

In practice, the questionnaires are often neither absolutely neutral nor too "far" from the neutral case. In other words, they are "nearly neutral". In this case, we assume the sequence of measure $\mathbf{P}_q$, which is defined by probabilities $p(\vec{x}_q)$, is contiguous to the sequence of measure $\mathbf{P}_{0,q}$.

In more general situations, where $\{a_i(j)\}$ were assumed to be an arbitrary sequence, the asymptotic behavior of ratios in (5.1) is more complicated. We will show that, under certain condition, the limit theorems can still be established.

5.5 Limit theorem for contiguous neighborhood of neutral questionnaires

As mentioned, one reason for introducing likelihood ratio M_q is its possession of good asymptotic properties. The asymptotic normality of log-likelihood ratio (see e.g., [28] and [6]) shows that if $\{\mathbf{P}_q\}$ is contiguous to $\{\mathbf{P}_{0,q}\}$ (denoted by $\{\mathbf{P}_q\} \triangleleft \{\mathbf{P}_{0,q}\}$), and satisfies some additional conditions, the distribution of L_q converges to the normal distribution $\mathcal{N}(-\frac{1}{2}\sigma^2, \sigma^2)$, i.e. the distribution of M_q converge to a log-normal distribution. The limit theorem under this condition can therefore be formulated.

Define the Hellinger distance between $\mathbf{P}_{qi}$ and $\mathbf{P}_{0,qi}$ as follows:

$$
\begin{aligned}
H(\mathbf{P}_{qi}, \mathbf{P}_{0,qi}) &= \Big(\int (\sqrt{p} - \sqrt{p_0})^2 d\mu\Big)^{\frac{1}{2}} \\
&= \left(2 - 2\int \left(\frac{d\mathbf{P}_{qi}}{d\mathbf{P}_{0,qi}}\right)^{\frac{1}{2}} d\mathbf{P}_{0,qi}\right)^{\frac{1}{2}} \\
&= \left(2 - 2\int \sqrt{k_i a_i(x_i)}\, d\mathbf{P}_{0,qi}\right)^{\frac{1}{2}}.
\end{aligned}
$$

Theorem 7. *If*

$$
\lim_{q\to\infty} \sum_{i=1}^{q} H(\mathbf{P}_{qi}, \mathbf{P}_{0,qi})^2 = \frac{1}{4}\sigma^2 < \infty \tag{5.5}
$$

and for every $\epsilon > 0$,

$$
\lim_{q\to\infty} \sum_{i=1}^{q} \int_{|k_i a_i(x_i) - 1| \geqslant \epsilon} \left(\left(\sqrt{k_i a_i(x_i)}) - 1\right)\right)^2 d\mathbf{P}_{0,qi} = 0 \tag{5.6}
$$

then

$$
\frac{\mathbb{E}\mu_n(m)}{n} \to \frac{1}{\lambda m!}\mathbb{E}\left[\lambda^m e^{mL} e^{-\lambda e^L}\right] \tag{5.7}
$$

and

$$
\frac{\mathbb{E}\mu_n(m)}{\mathbb{E}\mu_n} \to \frac{\mathbb{E}\left[\lambda^m e^{mL} e^{-\lambda e^L}\right]}{m!\left(1 - \mathbb{E}\left[e^{-\lambda e^L}\right]\right)} \tag{5.8}
$$

with $L \sim \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma^2\right)$.

Proof: Condition (5.5) and (5.6) implies $\{\mathbf{P}_q\} \triangleleft \{\mathbf{P}_{0,q}\}$, and they guarantee the asymptotic normality of L_q ([28], Theorem 2),

$$L_q = \ln M_q \xrightarrow{d(\mathbf{P}_{0,q})} \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma^2\right)$$

and combine with Lemma 5, we can get (5.7), (5.8) thereafter. □

In this case, both ratios are strictly greater than 0, and $\mathbb{E}\mu_n \to \infty$. Hence the conditions of both definitions of large number of rare events are satisfied.

Example: Suppose we have

$$\mathbf{P}_{qi}(j) = a_i(j) = \frac{1}{k_j}(1 + \frac{e_{ij}}{\sqrt{q}})$$

where $\{e_{ij}\}$ satisfies $-1 \leqslant e_{ij} \leqslant 1$ and

$$\lim_{q\to\infty} \frac{1}{q} \sum_{i=1}^{q} \frac{1}{k_i} \sum_{j=1}^{k_i} e_{ij}^2 \to \sigma^2 < \infty$$

and constraint $\sum_{j=1}^{k_i} e_{ij} = 0$. Then square of Hellinger distance between $\mathbf{P}_{qi}$ and $\mathbf{P}_{0,qi}$ becomes

$$H(\mathbf{P}_{qi}, \mathbf{P}_{0,qi})^2 = 2 - 2\int \sqrt{k_i a_i(x_i)} d\mathbf{P}_{0,qi} = 2 - 2\frac{1}{k_i} \sum_{j=1}^{k_i} \sqrt{1 + \frac{e_{ij}}{\sqrt{q}}}$$

using Taylor's expansion we get

$$\sqrt{1 + \frac{e_{ij}}{\sqrt{q}}} = 1 + \frac{1}{2}\frac{e_{ij}}{\sqrt{q}} - \frac{1}{8}\frac{e_{ij}^2}{q} + \frac{1}{16}\frac{e_{ij}^3}{q\sqrt{q}} \cdots$$

So

$$\sum_{i=1}^{q} H(\mathbf{P}_{qi}, \mathbf{P}_{0,qi})^2 = \frac{1}{4}\sigma^2 + O(\frac{1}{\sqrt{q}}) \to \frac{1}{4}\sigma^2$$

and since when $q > \frac{1}{\epsilon^2}$, $|k_i a_i(x_i) - 1| < \epsilon$ for all i, it is easy to see that (5.6) is satisfied. These imply the asymptotic normality of L_q.

Remark 1. In our treatment in this section, we assumed that the components of $\vec{X}$ are independent. However, this is not a necessary condition. In the case that components of $\vec{X}$ are dependent, we can simply replace $k_i a(x_i)$ by conditional probabilities $k_i a(x_i|\vec{x}_{i-1})$, to achieve the same result (see [6]).

5.6 Limit theorem for general cases

In general, if $\{a_i(j)\}$ is an arbitrary sequence of distributions, then unlike the contiguity case in previous section where $L_q(\vec{X}_q)$ converge in distribution to normal random variable, the expectation of $L_q(\vec{X}_q)$ usually tend to $-\infty$ while the variance goes to ∞. In this situation we can use similar technique which typically is used in the theory of large deviations (see, e.g., [4] and [11]). After applying Esscher's transform, the random variable

$$Y_q = \frac{L_q(\vec{X}_q)}{\sqrt{q}}$$

will converge in distribution under the adjoint measure, and can be approximated by Edgeworth series (see, e.g., [4] and [19]).

Under necessary conditions, we shall see that, in this case, the limit theorem can be established and result agrees with Karlin-Rouault's law.

For any fixed sequence $\{a_i(j)\}$, the cumulant generating function of ξ_i under $\mathbf{P}_{0,qi}$ is defined by

$$\psi_i(u) = \ln \mathbb{E}_{\mathbf{P}_{0,qi}} e^{u\xi_i} = \ln \left(\sum_{j=1}^{k_i} [k_i a_i(j)]^u \right) - \ln(k_i)$$

and the cumulant generating function of $L_q(\vec{X}_q)$ is therefore,

$$\ln \mathbb{E}_{\mathbf{P}_{0,q}} e^{uL_q(\vec{X}_q)} = \sum_{i=1}^{q} \psi_i(u).$$

By Esscher's transform, the adjoint to $\mathbf{P}_{0,q}$ distribution, $\mathbf{Q}_{u,q}$, is defined as follows,

$$\frac{d\mathbf{Q}_{u,q,L_q(\vec{X})}}{d\mathbf{P}_{0,q,L_q(\vec{X})}}(z) = e^{uz-\sum_{i=1}^{q}\psi_i(u)}.$$

Consequently, the logarithm of moment generating function of $Y_q = \frac{L_q(\vec{X}_q)}{\sqrt{q}}$ under $\mathbf{Q}_{u,q,L_q(\vec{X})}$ is,

$$\ln \mathbb{E}_{\mathbf{Q}_{u,q}} e^{rY_q} = \sum_{i=1}^{q}\psi_i(u+\frac{r}{\sqrt{q}}) - \sum_{i=1}^{q}\psi_i(u).$$

Therefore, expected value of Y_q under $\mathbf{Q}_{u,q,L_q(\vec{X})}$ is equal to

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

Each function $\psi_i(u)$ is convex and $\psi_i(0) = \psi_i(1) = 0$. Therefore, we can choose $u = u_q$ such that

$$\mathbb{E}_{\mathbf{Q}_{u_q,q}} L_q(\vec{X}_q) = \sum_{i=1}^{q}\psi_i'(u_q) = 0.$$

The variance of Y_q under $\mathbf{Q}_{u_q,q}$ is

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

and therefore $Y_q = \frac{L_q(\vec{X}_q)}{\sqrt{q}}$ becomes a random variable with mean 0 and variance σ_q^2 under $\mathbf{Q}_{u_q,q}$.

Theorem 8. *Assume u_q is the solution of $\sum_{i=1}^{q}\psi_i'(u) = 0$. If $\{a_i(j)\}$ is such that*

$$c < \frac{1}{q}\sum_{i=1}^{q}\psi_i''(u_q) < C \tag{5.9}$$

and if there exists $\delta > 0$ such that

$$\left|e^{\sum_{i=1}^{q}[\psi_i(u_q+r)-\psi_i(u_q)]}\right| = o\left(\frac{1}{\sqrt{q}}\right) \quad \textit{uniformly in} \quad r > \delta > 0 \tag{5.10}$$

are satisfied, then

$$\frac{\mathbb{E}\mu_n}{n} \to 0 \tag{5.11}$$

and

$$\frac{\mathbb{E}\mu_n(m)}{\mathbb{E}\mu_n} \to \frac{u^*\Gamma(m-u^*)}{\Gamma(m+1)\Gamma(1-u^*)}, m = 1, 2, \ldots \tag{5.12}$$

where $u^* = \lim_{q\to\infty} u_q$.

Proof: Applying Esscher's transform,

$$\mathbb{E}_{\mathbf{P}_{0,q}}\left[(\lambda M_q(\vec{X}_q))^m e^{-\lambda M_q(\vec{X}_q))}\right]$$

$$= e^{\sum_{i=1}^q \psi_i(u_q)} \int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)x} e^{-\lambda e^x} d\mathbf{Q}_{u_q,q,L_q(\vec{X})}(x) \tag{5.13}$$

then replace $L_q(\vec{X})$ by Y_q,

$$\int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)x} e^{-\lambda e^x} d\mathbf{Q}_{u_q,q,L_q(\vec{X})}(x)$$

$$= \int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\mathbf{Q}_{u_q,q,Y_q}(y) \tag{5.14}$$

In Lemma 6 , we will prove that under condition (5.9) and (5.10),

$$\int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\mathbf{Q}_{u_q,q,Y_q}(y)$$

$$= \int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y) + o\left(\frac{1}{\sqrt{q}}\right) \tag{5.15}$$

where Φ_{0,σ_q^2} is normal distribution function with mean 0 and variance σ_q^2. Then by Lemma 12,

$$\int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y)$$

$$\sim \frac{\lambda^{u_q}}{\sqrt{q}} \phi_{0,\sigma_q^2}(0)\Gamma(m-u_q) = O\left(\frac{1}{\sqrt{q}}\right). \tag{5.16}$$

Combine (5.3), Lemma5, (5.13), (5.14), (5.15), (5.16), and note that $\frac{1}{n} = o\left(\frac{1}{\sqrt{q}}\right)$, we conclude that for any $m \geqslant 1$,

$$\mathbb{E}\mu_n(m) \sim Ne^{\sum_{i=1}^{q}\psi_i(u_q)}\frac{\lambda^{u_q}}{\sqrt{q}}\phi_{0,\sigma_q^2}(0)\frac{\Gamma(m-u_q)}{m!}.$$

It is easy to show that

$$\sum_{m=1}^{\infty}\frac{\Gamma(m-u_q)}{m!} = \frac{\Gamma(1-u_q)}{u_q}.$$

Indeed

$$\sum_{m=1}^{\infty}\frac{\Gamma(m-u_q)}{m!} = \sum_{m=1}^{\infty}\frac{\int_0^{\infty}y^{m-u_q-1}e^{-y}dy}{m!}$$

$$= \int_0^{\infty}\sum_{m=1}^{\infty}\frac{y^m}{m!}y^{-u_q-1}e^{-y}dy$$

$$= \int_0^{\infty}(e^y-1)e^{-y}y^{-u_q-1}dy = \int_0^{\infty}(1-e^{-y})y^{-u_q-1}dy$$

using integration by parts we obtain

$$\int_0^{\infty}(1-e^{-y})y^{-u_q-1}dy = -\frac{1}{u_q}\int_0^{\infty}(1-e^{-y})d(y^{-u_q})$$

$$= -\frac{1}{u_q}(1-e^{-y})y^{-u_q}\,|_0^{\infty} - \int_0^{\infty}y^{-u_q}e^{-y}dy = \frac{\Gamma(1-u_q)}{u_q}.$$

Finally we obtain,

$$\mathbb{E}\mu_n \sim Ne^{\sum_{i=1}^{q}\psi_i(u_q)}\frac{\lambda^{u_q}}{\sqrt{q}}\phi_{0,\sigma_q^2}(0)\frac{\Gamma(1-u_q)}{u_q}$$

hence (5.11) and (5.12) holds. □

Lemma 6. *Assume u_q is the solution of $\sum_{i=1}^{q}\psi_i'(u) = 0$. If $\{a_i(j)\}$ is such that the conditions*

$$c < \frac{1}{q}\sum_{i=1}^{q}\psi_i''(u_q) < C \tag{5.17}$$

and

$$\left|e^{\sum_{i=1}^{q}[\psi_i(u_q+r)-\psi_i(u_q)]}\right| = o\left(\frac{1}{\sqrt{q}}\right) \quad \textit{uniformly in} \quad r > \delta > 0 \tag{5.18}$$

are satisfied, then

$$\mathbb{E}\mu_n \gtrsim N^{u_q}$$

and therefore

$$\mathbb{E}\mu_n(k) \gtrsim N^{u_q}$$

Proof: Let us prove the lemma for $\mathbb{E}\mu_q$.

$$\mathbb{E}\mu_n \sim Ne^{\sum_{i=1}^{q}\psi_i(u_q)}\frac{\lambda^{u_q}}{\sqrt{q}}\phi_{0,\sigma_q^2}(0)\frac{\Gamma(1-u_q)}{u_q}$$

or

$$\mathbb{E}\mu_n \sim Ne^{\sum_{i=1}^{q}\psi_i(u_q)} = N\prod_{i=1}^{q}e^{\psi_i(u_q)}.$$

$$e^{\psi_i(u_q)} = \sum_{j=1}^{k_i}\frac{[k_i a_i(j)]^{u_q}}{k_i}$$

$$= \frac{(k_i)^{u_q}}{k_i}\sum_{j=1}^{k_i}[a_i(j)]^{u_q} \geq \frac{1}{k_i^{1-u_q}}$$

therefore

$$\mathbb{E}\mu_n \geq N\prod_{i=1}^{q}\frac{1}{k_i^{1-u_q}} = N^{u_q}.$$

□

Lemma 7. *If conditions (5.9) and (5.10) are satisfied, then (5.15) holds.*

Proof: Denote $g(y,q) = \lambda^m e^{(m-u_q)\sqrt{q}y}e^{-\lambda e^{\sqrt{q}y}}$ and $g'(y,q) = \frac{\partial g(y,q)}{\partial y}$, then since $\lim_{y\to\infty} g(y,q) = 0$ and $\lim_{y\to-\infty} g(y,q) = 0$,

$$\int_{-\infty}^{\infty} g(y,q)d\mathbf{Q}_{u_q,q,Y_q}(y) - \int_{-\infty}^{\infty} g(y,q)d\Phi_{0,\sigma_q^2}(y)$$
$$= \int_{-\infty}^{\infty}\left(\mathbf{Q}_{u_q,q,Y_q}(y) - \Phi_{0,\sigma_q^2}(y)\right)g'(y,q)dy.$$

Under the condition (5.9) and (5.10), the Edgeworth expansion (see [4]) shows,

$$\mathbf{Q}_{u_q,q,Y_q}(y) = \Phi_{0,\sigma_q^2}(y) - \frac{\sum_{i=1}^{q} \psi_i^{(3)}(u_q)}{6\sigma_q^3 q^{\frac{3}{2}}} H_2(\sigma_q y)\phi(\sigma_q y) + o\left(\frac{1}{\sqrt{q}}\right)$$

where $H_2(y) = y^2 - 1$ is the second Hermite polynomial. Therefore, using differentiation by parts

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left(\mathbf{Q}_{u_q,q,Y_q}(y) - \Phi_{0,\sigma_q^2}(y)\right) g'(y,q)dy \\
=& -\int_{-\infty}^{\infty} \frac{\sum_{i=1}^{q} \psi_i^{(3)}(u_q)}{6\sigma_q^3 q^{\frac{3}{2}}} H_2(\sigma_q y)\phi(\sigma_q y) g'(y,q)dy + o\left(\frac{1}{\sqrt{q}}\right) \\
=& -\frac{\frac{1}{q}\sum_{i=1}^{q} \psi_i^{(3)}(u_q)}{6\sigma_q^2\sqrt{q}} \int_{-\infty}^{\infty} H_3(\sigma_q y)\phi(\sigma_q y) g(y,q)dy + o\left(\frac{1}{\sqrt{q}}\right).
\end{aligned} \tag{5.19}$$

Since

$$\begin{aligned}
&\int_{-\infty}^{\infty} H_3(\sigma_q y)\phi(\sigma_q y) g(y,q)dy \\
=& \int_{-\infty}^{\infty} ((\sigma_q y)^3 - 3\sigma_q y)\phi(\sigma_q y)\lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} dy \to 0
\end{aligned}$$

then if $\lim_{q\to\infty} \frac{1}{q}\sum_{i=1}^{q} \psi_i^{(3)}(u_q) < \infty$, the right side of (5.19) is $o(\frac{1}{\sqrt{q}})$ and hence (5.15) holds. □

Lemma 8. *Suppose u_q is solution of $\sum_{i=1}^{q} \psi_i'(u) = 0$, then*

$$\int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y) \sim \frac{\lambda^{u_q}}{\sqrt{q}} \phi_{0,\sigma_q^2}(0)\Gamma(m - u_q).$$

Proof: Since for any $\beta > 0$ and $m \geqslant 1 > u_q$, if q large enough

$$\begin{aligned}
&\int_{-\infty}^{-\beta q^{-\frac{1}{4}}} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y) \\
\leqslant& \lambda^m e^{-(m-u_q)\beta q^{\frac{1}{4}}} e^{-\lambda e^{-\beta q^{\frac{1}{4}}}} \int_{-\infty}^{-\beta q^{-\frac{1}{4}}} d\Phi_{0,\sigma_q^2}(y) < o(\frac{1}{\sqrt{q}})
\end{aligned}$$

and

$$\int_{\beta q^{-\frac{1}{4}}}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y)$$

$$\leqslant \lambda^m e^{(m-u_q)\beta q^{\frac{1}{4}}} e^{-\lambda e^{\beta q^{\frac{1}{4}}}} \int_{\beta q^{-\frac{1}{4}}}^{\infty} d\Phi_{0,\sigma_q^2}(y) < o(\frac{1}{\sqrt{q}})$$

while,

$$\int_{-\beta q^{-\frac{1}{4}}}^{\beta q^{-\frac{1}{4}}} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y)$$

$$= \lambda^{u_q} \int_{-\beta q^{\frac{1}{4}}}^{\beta q^{\frac{1}{4}}} (\lambda e^z)^{m-u_q} e^{-\lambda e^z} d\Phi_{0,q\sigma_q^2}(z)$$

$$= \lambda^{u_q} \int_{-\beta q^{\frac{1}{4}}}^{\beta q^{\frac{1}{4}}} (\lambda e^z)^{m-u_q} e^{-\lambda e^z} \frac{1}{\sigma_q \sqrt{2\pi q}} e^{-\frac{z^2}{2q\sigma_q^2}} dz$$

$$\sim \frac{\lambda^{u_q}}{\sqrt{q}} \phi_{0,\sigma_q^2}(0) \int_{-\infty}^{\infty} (\lambda e^z)^{m-u_q} e^{-\lambda e^z} dz$$

$$= \frac{\lambda^{u_q}}{\sqrt{q}} \phi_{0,\sigma_q^2}(0) \Gamma(m-u_q)$$

we conclude that,

$$\int_{-\infty}^{\infty} \lambda^m e^{(m-u_q)\sqrt{q}y} e^{-\lambda e^{\sqrt{q}y}} d\Phi_{0,\sigma_q^2}(y) \sim \frac{\lambda^{u_q}}{\sqrt{q}} \phi_{0,\sigma_q^2}(0) \Gamma(m-u_q).$$

□

5.7 Non-classical asymptotics for Shannon diversity index

In Section 4.2 we considered entropy, proposed by Shannon, as a measure of biological diversity. Namely, we stated that biologists use estimator of entropy

$$H = -\sum_{i=1}^{N} p_i \ln p_i$$

with

$$\hat{H}_n = -\sum_{i=1}^{N} \frac{\nu_i}{n} \ln \frac{\nu_i}{n}.$$

In this case we were in the so called classical situation when number of different events (species) N was fixed. It would be interesting to see what happens if we use entropy to measure diversity in non-classical assumption that not only n is large, but at the same time N is large and all or majority of p_i-s are very small. For this purpose let us consider data in which elements are binary vectors, for example, responses for a questionnaire with q "Yes/No" questions, the state of any mechanical device with q "On/Off" components and so on. Let us consider asymptotic behaviour of $\hat{H}_n$ and H under non classical situation. Rewrite estimator of entropy in terms of spectral statistics. Suppose $\nu_{\overrightarrow{x}}$ is frequency of opinion $\overrightarrow{x}$ and Ξ_q the set of all possible opinions.

$$\hat{H}_n = -\sum_{\overrightarrow{x} \in \Xi_q} \frac{\nu_{\overrightarrow{x}}}{n} \ln \frac{\nu_{\overrightarrow{x}}}{n}$$

$$= -\frac{1}{n} \sum_{k=1}^{n} k\mu_n(k) \ln \frac{k}{n}$$

$$= -\frac{\mu_n}{n} \sum_{k=1}^{n} k \frac{\mu_n(k)}{\mu_n} \ln \frac{k}{n}$$

where $\mu_q(k)$ is number of opinions seen in a sample k times and μ_n is number of different opinions.

Lemma 9. *If, as $q \to \infty$ and sample size $n = \lambda 2^q$ with $\lambda = const$,*

$$\frac{\mu_q(k)}{\mu_q} \to \frac{u\Gamma(k-u)}{\Gamma(k+1)\Gamma(1-u)},$$

then

$$\hat{H}_n \sim \frac{u\mu_n}{n^u \Gamma(1-u)(u-1)^2}.$$

Proof. Using results obtained in chapter 5 we can write:

$$\hat{H}_n \sim -\frac{\mu_n}{n}\frac{u}{\Gamma(1-u)}\sum_{k=1}^{n}\frac{\Gamma(k-u)}{\Gamma(k)}\ln\frac{k}{n}.$$

Let us consider asymptotic behaviour of $\dfrac{\Gamma(k-u)}{\Gamma(k)}$ when k is large.

$$\frac{\Gamma(k-u)}{\Gamma(k)} = \frac{\int_0^\infty e^{-t}t^{k-u-1}dt}{\Gamma(k)} = \mathbb{E}T^{-u}$$

where T is Gamma random variable with scale parameter 1 and shape parameter k. From the law of large numbers $\dfrac{T}{k} \to 1$, and therefore

$$\frac{\Gamma(k-u)}{\Gamma(k)} = \mathbb{E}T^{-u} \sim \frac{1}{k^u}.$$

Let us now go back to $\hat{H}_n$,

$$\frac{1}{n}\sum_{k=1}^{n}\frac{\Gamma(k-u)}{\Gamma(k)}\ln\frac{k}{n} = \frac{1}{n}\sum_{x=\frac{1}{n}}^{1}\frac{\Gamma(nx-u)}{\Gamma(nx)}\ln x$$

$$\sim \int_{\frac{1}{n}}^{1}\frac{\Gamma(nx-u)}{\Gamma(nx)}\ln x\,dx$$

$$= \int_{\frac{1}{n}}^{\frac{1}{n}+\varepsilon}\frac{\Gamma(nx-u)}{\Gamma(nx)}\ln x dx + \int_{\frac{1}{n}+\varepsilon}^{1}\frac{\Gamma(nx-u)}{\Gamma(nx)}\ln x dx$$

$$\sim \int_{\frac{1}{n}+\varepsilon}^{1}\frac{\Gamma(nx-u)}{\Gamma(nx)}\ln x dx$$

$$\sim \int_{\frac{1}{n}+\varepsilon}^{1}(nx)^{-u}\ln x dx$$

$$\sim \frac{1}{n^u}\int_{\frac{1}{n}+\varepsilon}^{1}(x)^{-u}\ln x dx$$

$$\sim -\frac{1}{n^u}\frac{1}{(u-1)^2}.$$

Finally we obtain that

$$\hat{H}_n \sim \frac{u\mu_n}{n^u\Gamma(1-u)(u-1)^2}.$$

□

Notice that this was the case when underlying probabilities are arbitrary. In case when these probabilities are from contiguous neighbourhood of $\frac{1}{2}$ or $\frac{1}{m_i}$, depending on whether we have binary or multiple choice questionnaire, we have different limit for $\frac{\mu_n(k)}{\mu_n}$ and respectively we have different asymptotics for $\hat{H}_n$, namely,

$$\hat{H}_n \sim -\frac{\mu_n}{n}\sum_{k=1}^{n} k\frac{\int \pi(k,\lambda e^z)\Phi_{-\frac{c^2}{2},c^2}(dz)}{\int(1-\pi(0,\lambda e^z))\Phi_{-\frac{c^2}{2},c^2}(dz)}\ln\frac{k}{n}.$$

One possible, although not very rigorous, way to derive asymptotics of entropy H, is by assuming that $R_q(z) \sim z^{-u}$.

$$H = -\sum_{\overrightarrow{x}\in\Xi} p_{\overrightarrow{x}}\ln p_{\overrightarrow{x}} = -\frac{1}{n}\sum_{\overrightarrow{x}\in\Xi} np_{\overrightarrow{x}}\ln\frac{np_{\overrightarrow{x}}}{n}.$$

$$H = \frac{\mathbb{E}\mu_n}{n\Gamma(1-u)}\int_0^n z\ln\frac{z}{n}dz^{-u} = -\frac{u\mathbb{E}\mu_n}{n\Gamma(1-u)}\int_0^n z^{-u}\ln\frac{z}{n}dz$$

denoting $\frac{z}{n}$ by x we will obtain,

$$H = -\frac{u\mathbb{E}\mu_n}{n^u\Gamma(1-u)}\int_0^1 x^{-u}\ln x dx \sim \frac{u\mathbb{E}\mu_n}{n^u\Gamma(1-u)(u-1)^2}.$$

Chapter 6

Some numerical observations.

6.1 d1 and d2 zones of LNRE

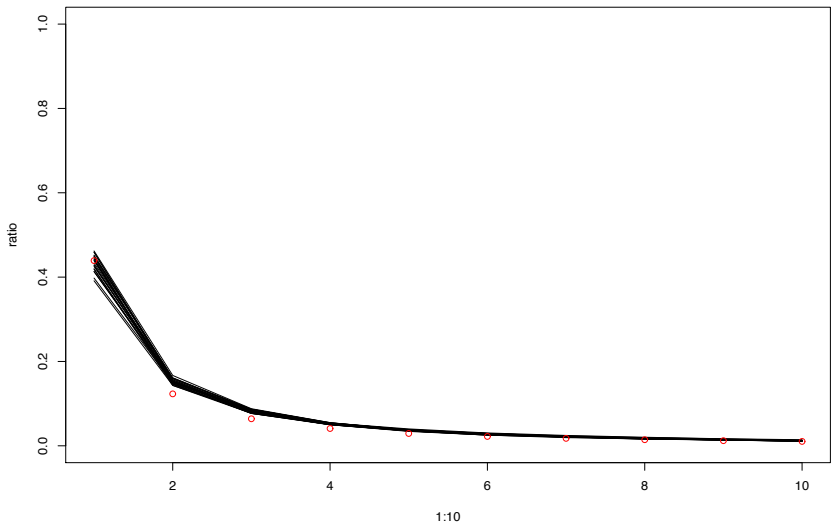
In this chapter, for simplicity we concentrate on binary questionnaires. Suppose we have a survey which contains, say, 20 binary "yes/no" questions. This implies that there are $N = 2^{20}$ possible "opinions". Suppose sample size is equal to n, $n = \lambda N$.

The Figure 6.1 shows a bundle of trajectories of $\frac{\mu_q(k)}{\mu_q}$ for $k = 1, \cdots, 10$ and a_{iq}-s are uniformly distributed on the interval $[0, 1]$

The dots correspond to the Karlin-Rouault law with parameter $u = 0.442$, which is the mean value of u_q for uniformly distributed a_{iq}-s. Even though $q = 20$ is not very large, we see that convergence is quite satisfactory.

It would be interesting to answer following questions: If we are given values of $\frac{\mu_q(k)}{\mu_q}$ and $\frac{\mu_q}{n}$, then what can one say about underlying probabilities of questions? or if we know these probabilities a priori, what values should we expect $\frac{\mu_q(k)}{\mu_q}$ and $\frac{\mu_q}{n}$ to take? In other words, we want to investigate empirically if the ratios of spectral statistics depend on underlying distribution of a_{iq}. It is quite possible to give a legitimate answer to these questions by combining analytical results we obtained in previous chap-

Figure 6.1: Simulation of Karlin-Rouault law.



ter and some numerical results we will obtain here. In fact the ratios $\frac{\mu_q(k)}{\mu_q}$ and $\frac{\mu_q}{n}$ can take any value between 0 and 1, as for as underlying probabilities can be any number between 0 and 1. In the previous chapter we obtained two different asymptotic for the ratios $\mathbb{E}\frac{\mu_q(k)}{\mu_q}$ and $\mathbb{E}\frac{\mu_q}{n}$. Namely, for underlying probabilities a_{iq} from contiguous neighbourhood of $\frac{1}{2}$ we obtained:

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

and

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

According to Definitions 1 and 2 of Chapter 1 these results imply that we are in d1 and therefore in d2 zone of LNRE.

For arbitrary underlying probabilities a_{iq} we have following asymp-

totic expressions:

$$\frac{\mathbb{E}\mu_q(k)}{\mathbb{E}\mu_q} \to \frac{u_q\Gamma(k-u_q)}{\Gamma(k+1)\Gamma(1-u_q)}$$

and

$$\frac{\mathbb{E}\mu_q}{n} \to 0$$

which implies that we are in d2 zone of LNRE, but not in d1.

These results suggest that so called Karlin-Rouault's law is not the only possible limiting law we can obtain. Now we can rephrase the questions asked above as follows. If the data shows that we are in d1 and therefore in d2 or only in d2 zone, what can be said about underlying probabilities a_{iq} and the other way round, if a_{iq}-s are given, should we expect to find ourselves in d1 or in d2? To answer these questions we start with investigating behavior of function $\psi_i(u)$ considered in previous chapters

$$\psi_i(u) = \ln[(2a_{iq})^u + (2(1-a_{iq}))^u] - \ln 2$$

as we know the parameter u_q is the solution of following equality;

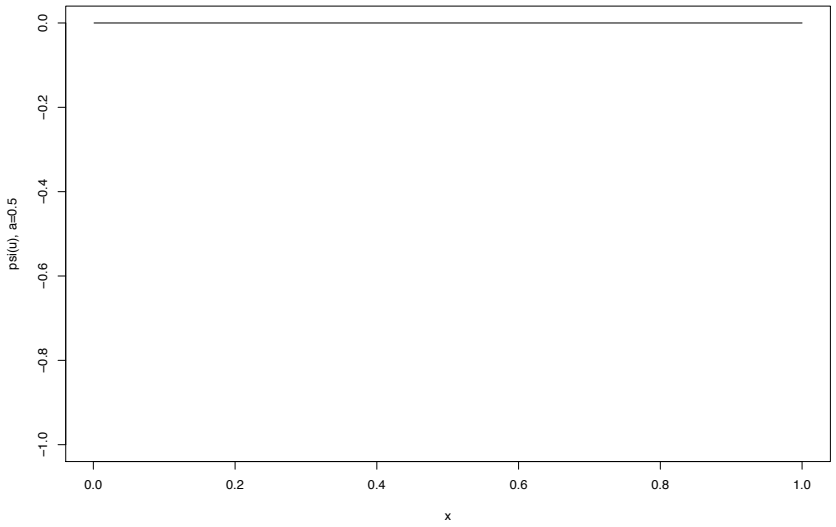
$$\sum_{i=1}^{q} \psi_i'(u) = 0$$

in other words it is the arg min of $\sum_{i=1}^{q} \psi_i(u)$. With respect a_{iq}-s, the parameter u_q is very stable not only for the $\sum_{i=1}^{q} \psi_i(u)$, but even for the summand $\psi_i(u)$. Notice here that the function $\psi_i(u)$ and therefore $\sum_{i=1}^{q} \psi_i(u)$ is symmetric with respect to the a_{iq}-s.

On Figure 6.2 is a graph of $\frac{1}{q}\sum_{i=1}^{q} \psi_i(u)$ for 17 a_{iq}-s equal to $\frac{1}{2}$ and 3 a_{iq}-s taking extreme values, greater then 0.9. The graph is lying on $y = 0$ line, therefore looking at the arg min does not make sense as its exact value will not matter much.

On the Figure 6.3 a_{iq}-s change uniformly in $[0.45, 0.55]$ or we can say they are in the contiguous neighbourhood of $\frac{1}{2}$, but still $\frac{1}{q}\sum_{i=1}^{q} \psi_i(u)$ is so flat that it is almost impossible to distinguish from $y = 0$ line, therefore the arg min is very volatile.

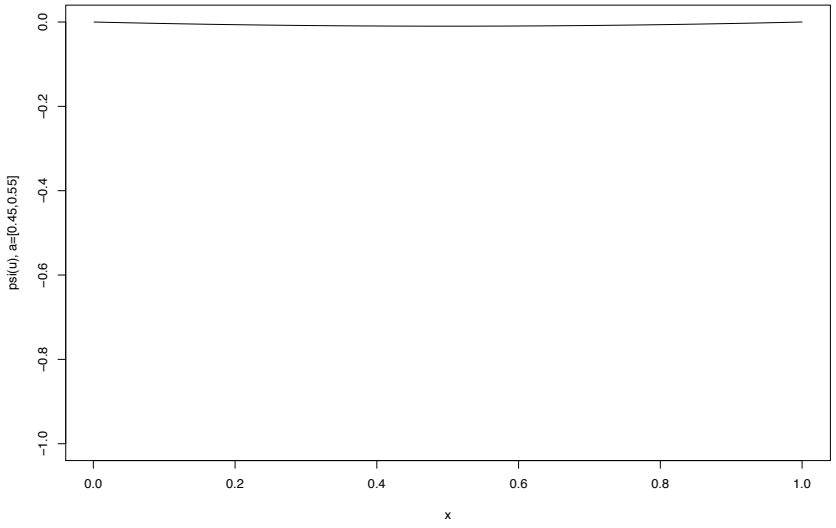
Figure 6.2: $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ for 17 a_{iq}-s equal to $\frac{1}{2}$ and 3 a_{iq}-s take extreme values, greater then 0.9.



Only in the Figure 6.4, when a_{iq}-s change uniformly in $[0.4, 0.6]$, one can clearly see the concave shape of $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ and consequently it is reasonable to talk about arg min. In this case it is approximately 0.5.

These values of a_{iq}-s, could be regarded as a some kind of "boundary" between contiguity and large deviations situations or the "boundary" between d1 and d2 zones of LNRE. If a_{iq}-s are not deviated from 0.5 more then 0.1, then we will stay in contiguity case and therefore in d1 and d2 zones.

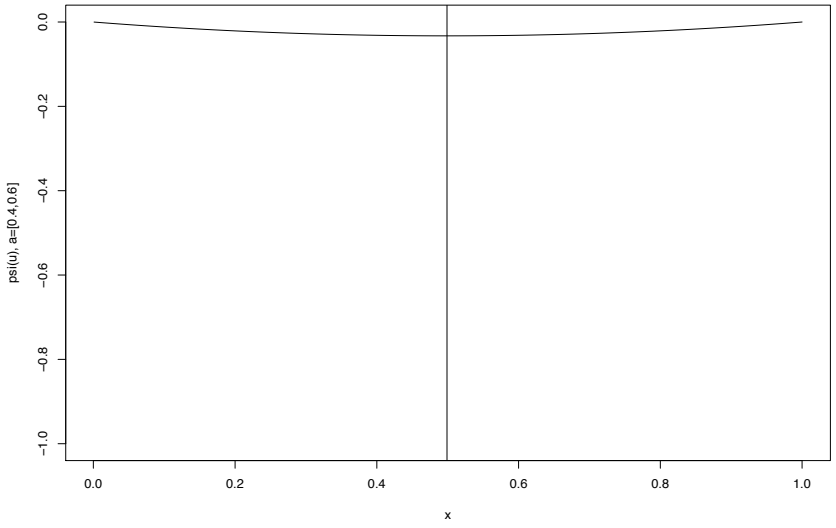
Figure 6.3: $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ for a_{iq}-s change uniformly in $[0.45, 0.55]$.



6.2 Moving from contiguity to large deviations

In this context we can answer another interesting question. In the Theorem 3 of chapter 3 we have following condition on a_{iq}-s: $a_{iq} = \frac{1}{2} + \frac{c_{iq}}{\sqrt{q}}$ and $\lim_{q\to\infty}\sum_{i=1}^{q}\frac{c_{iq}^2}{q} = c^2$. This condition leads us to the contiguity situation and therefore to d1 and d2 zone. Now one could ask the following question: how big can c^2 be that we will stay in d1? As our simulation shows, for $q = 20$, we can take values from interval $[0.4, 0.6]$, which means that $|\frac{c_{iq}}{\sqrt{q}}|$ can be as big as 0.1, which on the other hand means that $c^2 = 20 * 0.01 = 0.2$. So we can conclude that , for $q = 20$, if $c^2 \leq 0.2$ then we definitely will be in d1 zone. In other words c^2 can be another kind of "boundary" between d1 and d2 zones of LNRE. I have to say that this "boundary" is a bit vague, as its hard to say for what value of c^2 we will

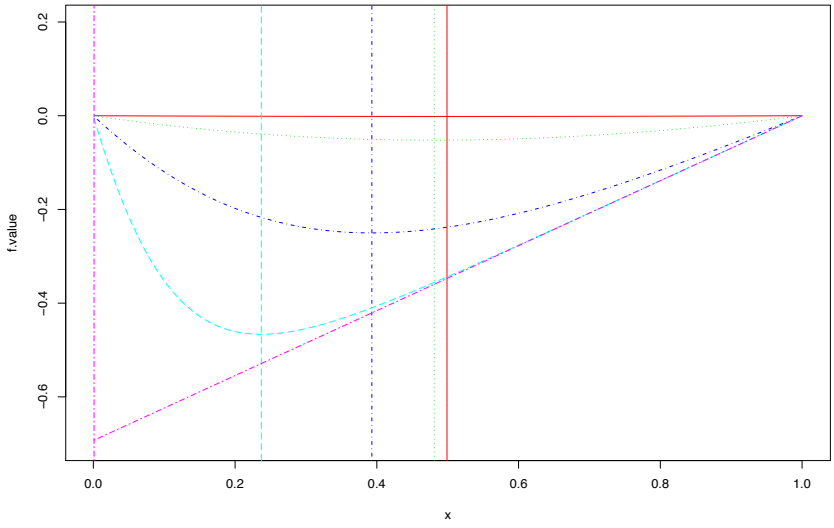
Figure 6.4: $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ for a_{iq}-s change uniformly in $[0.4, 0.6]$.



leave d1 zone and move to d2. The reason for this inaccuracy is that q is not large enough. Notice that to stay in d1, as $q \to \infty$ the size of deviation from 0.5 should be decreasing to maintain c^2 finite.

Figure 6.5 contains graphs of $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ for different collections of a_{iq}-s. The most "flat" curve corresponds to a_{iq}-s changing in the interval $[0.4, 0.6]$ (we are in d1); then we slowly leave d1 zone, but stay in d2. The most concave shape corresponds to a_{iq}-s changing in the interval $[0, 0.00002]$ with corresponding $u_q \simeq 0.25$ and finally the straight line corresponds to the case when all a_{iq}-s are equal to 0.

Figure 6.5: $\frac{1}{q}\sum_{i=1}^{q}\psi_i(u)$ for different collections of a_{iq}-s.

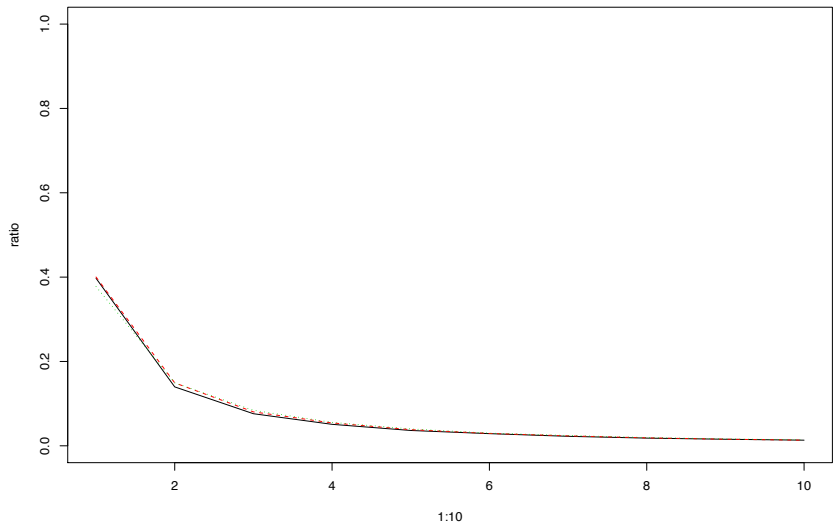


6.3 The role of λ-"rate per cell"

It is also very interesting to investigate the role of the λ in the asymptotic behaviour of $\frac{\mu_q(k)}{\mu_q}$. Looking at the analytical expressions we can see that λ participates only in the contiguity case. Following simulations give us empirical proof that influence of λ for arbitrary a_{iq}-s is very insignificant, while for contiguity situation λ plays crucial role. On the Figure 6.6 $q = 20$, a_{iq}-s change uniformly in $[0, 1]$ and λ takes values 1,2,3. Black line corresponds to $\lambda = 1$, red line to $\lambda = 2$ and green line to $\lambda = 3$. On the Figure 6.7, $q = 20$, a_{iq}-s change uniformly in $[0.4, 0.6]$ and λ takes values 1,2,3. Black line corresponds to $\lambda = 1$, red line to $\lambda = 2$ and green line to $\lambda = 3$.

As we mentioned before for questionnaire with q "Yes/No" questions we have 2^q possible "opinions". Anyone who conducts a survey would try to get at least that many 'opinions" as many are possible. That would

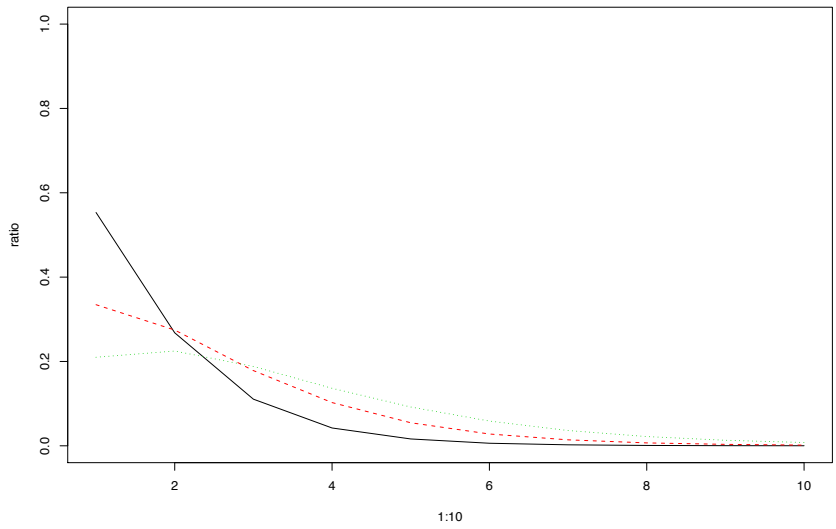
Figure 6.6: Influence of "rate per cell"-λ for uniform a_{iq}-s.



be fair in a sense as it would give a chance to each possible "opinion" to appear. So in this case $\lambda = 1$ and $n = 2^q$. Ideally he or she would question twice or three times more people. However, when q is very large, say greater then 50 or greater then 100, it is not always possible to obtain such a big sample size. In this context we have to consider the case when $\lambda < 1$.

For arbitrary a_{iq}-s, as we mentioned before, λ does not play big role and we demonstrated this in Figure 6.1. Figure 6.7 shows the result of similar simulation, but for $\lambda < 1$. When $\lambda < 1$ its role becomes more significant as it allows the ratio $\frac{\mu_q(1)}{\mu_q}$, which in this case is same as u, to take the value greater then 0.5, which on the other hand is impossible for any $\lambda \geq 1$. Therefore when we observe $\frac{\mu_q(1)}{\mu_q} \geq \frac{1}{2}$ we don't know a priori whether we are in the contiguity case or in Karlin-Rouault case, unless we know the value of λ. For a_{iq}-s changing from $[0.4, 0.6]$, or in the contiguity case, the influence of $\lambda < 1$ remains significant as it allows $\frac{\mu_q(1)}{\mu_q}$ to be as close to

Figure 6.7: Influence of "rate per cell"-λ in contiguity case.



1 as possible, depending how small λ is. To find the "boundary" between d1 and d2 zones of LNRE, one can consider following approach. We know the asymptotic behaviour of ratio $\frac{\mu_q(k)}{\mu_q}$ in contiguity case.

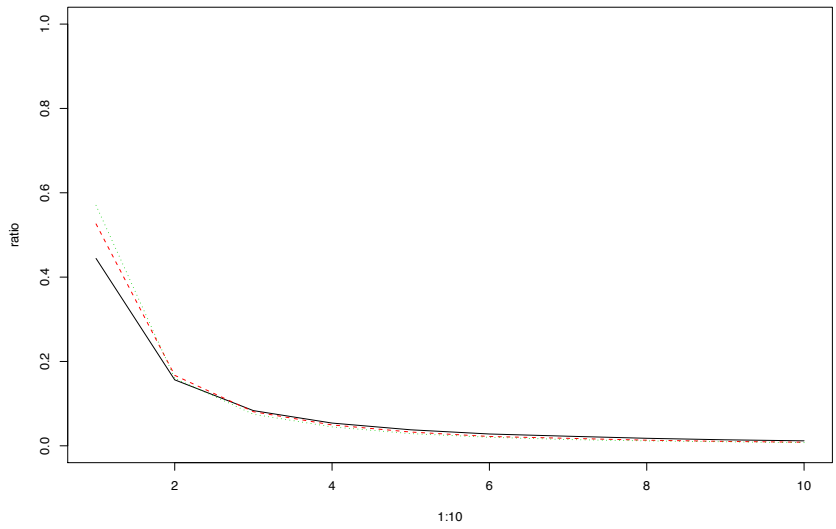
$$\frac{\mu_q(k)}{\mu_q} \sim \frac{\int \pi(k, \lambda e^z)\Phi_{-\frac{c^2}{2},c^2}(dz)}{\int (1-\pi(0,\lambda e^z))\Phi_{-\frac{c^2}{2},c^2}(dz)}.$$

On the other hand we know that for $\lambda = 1$ the value $\frac{\mu_q(1)}{\mu_q} = \frac{1}{2}$ sets the boundary between Karlin-Rouault law and contiguity case and therefore between d1 and d2 zones of LNRE. Taking advantage of this fact, one can look at the ratio

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

as a function of c^2 and find out for what value of c^2 it becomes smaller

Figure 6.8: Influence of "rate per cell"-λ in general case, $\lambda < 1$.



or equal to $\frac{1}{2}$. Figure 6.10 shows that for the value $c = 1.1$ or equivalently $c^2 = 1.21$ the ratio $\frac{\mu_q(1)}{\mu_q}$ becomes $\frac{1}{2}$. Consequently one can consider $c^2 = 1.21$ as a boundary between d1 and d2 zones of LNRE. Notice that the requirement that $\lambda = 1$ plays crucial role here, as for $\lambda < 1$ the ratio $\frac{\mu_q(1)}{\mu_q}$ can be smaller than $\frac{1}{2}$ even in case of contiguity. As a conclusion we can formulate that, we have only two cases. One is Contiguity case, when underlying probabilities are in the contiguous neighbourhood of "uniform" probabilities ($\frac{1}{2}$, or $\frac{1}{k_i}$, depending on questionnaire); and second situation, when a_{iq}-s are arbitrary. In the latter case the variation of the parameter u, and therefore the values of the ratios of spectral statistics, is extremely stable as was demonstrated by our simulations.

Figure 6.9: Influence of "rate per cell"-λ in contiguity case, $\lambda < 1$.

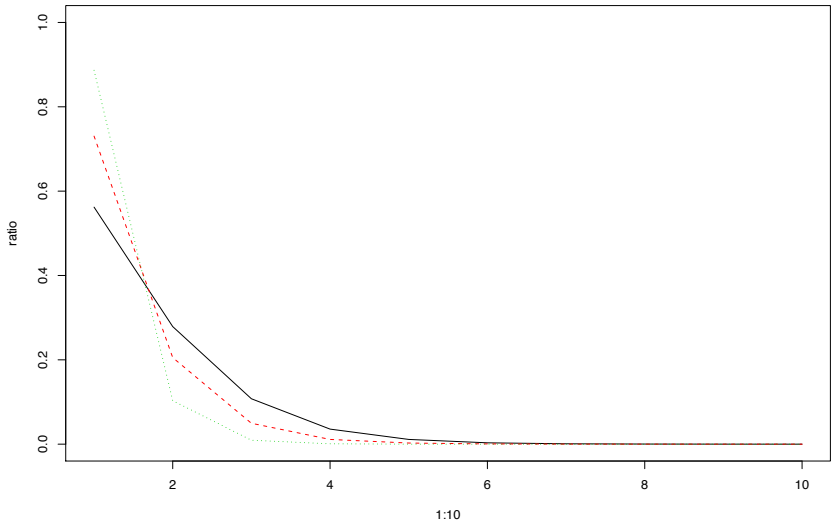
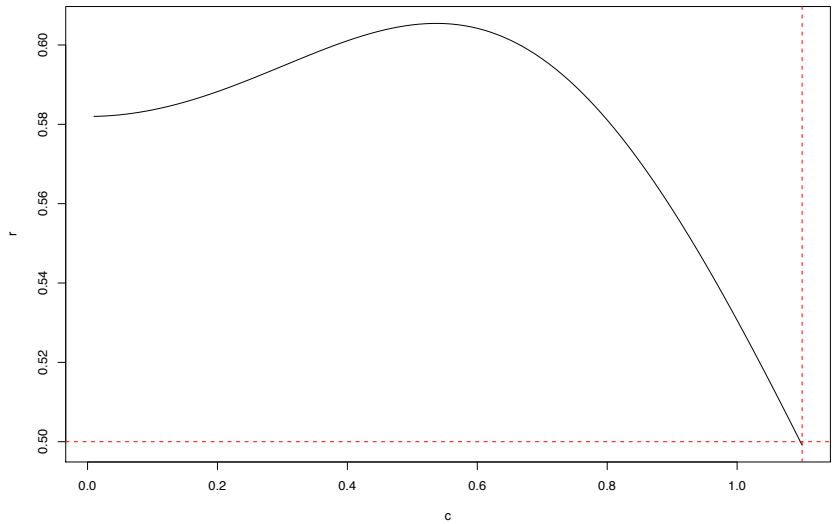


Figure 6.10: "Boundary" between d1 and d2 zones of LNRE in terms of c.



Chapter 7

Divisible Statistics

7.1 Introduction

Consider a random vector $\nu_{n1}, \ldots, \nu_{nN}$ which follows multinomial distribution with sample size n and probabilities $p_{n1}, \ldots, p_{nN}$. In general, the divisible statistics can be defined by the sum of some functions g_{ni} of the frequencies and probabilities,

$$\sum_{i=1}^{N} g_{ni}(\nu_{ni}).$$

Let's consider the normalized frequencies

$$Y_{ni} = \frac{\nu_{ni} - np_{ni}}{\sqrt{np_{ni}}}, \tag{7.1}$$

where the function g_{ni} has been expressed as a function of argument Y_{ni}, i.e.,

$$h_{ni}(Y_{ni}) = g_{ni}(\nu_{ni}).$$

Examples of the divisible statistics include the maximum likelihood statistics with

$$g_{ni}(\nu_{ni}) = \nu_{ni} \log\left(\frac{\nu_{ni}}{np_{ni}}\right),$$

the chi-square statistics with

$$g_{ni}(\nu_{ni}) = \frac{(\nu_{ni} - np_{ni})^2}{np_{ni}},$$

the so-called spectrum with

$$g_{ni}(\nu_{ni}) = \mathbf{I}\{\nu_{ni} \in A\},$$

and so on.

In classical statistical analysis, the limit theorems for the divisible statistics when $n \to \infty$ but N is fixed have been well-studied. However, it is also of great interest, from both the practical and theoretical points of view, to investigate the limit behaviour of these divisible statistics when both n and N tend to infinity. Research in this direction includes [10], [9], [27] etc.

In 1980, Khmaladze firstly developed an innovative approach to study this kind of problems. Instead of focusing on the asymptotic behaviour of the total sum $\sum_{i=1}^{N} g_{ni}(\nu_{ni})$, he considered the process $X_{n,N}(t)$ formed by the normalized partial sum

$$X_{n,N}(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[g_{ni}(\nu_{ni}) - \mathbb{E} g_{ni}(\nu_{ni}) \right].$$

It was shown that this process can be conveniently viewed as a semi-martingale with respect to the natural filtration $\{\mathcal{F}_i^n\}_{0 \leqslant i \leqslant N}$ with $\mathcal{F}_i^n = \sigma\{\nu_{nk} : k \leqslant i\}$ and $\mathcal{F}_0^n = \{\emptyset, \Omega\}$, and can be easily decomposed into a martingale part,

$$W_{n,N}(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[g_{ni}(\nu_{ni}) - \mathbb{E}\left(g_{ni}(\nu_{ni}) | \mathcal{F}_{i-1}^n \right) \right],$$

and a compensator part,

$$K_{n,N}(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[\mathbb{E}\left(g_{ni}(\nu_{ni}) | \mathcal{F}_{i-1}^n \right) - \mathbb{E} g_{ni}(\nu_{ni}) \right].$$

In Khmaladze's paper, the condition $n \sim N$ and $\sup |Np_{ni}| < \infty$ implies that the asymptotic behaviour of the frequencies ν_{ni} are Poissonian. Under these conditions, if g_{ni} are functions such that $|g_{ni}(\nu_{ni})| < ce^{a\nu_{ni}}$, both $W_{n,N}$ and $K_{n,N}$ converge in distribution to some Gaussian processes (see[13] for detail).

7.2 Limit theorems for spectral statistics

Before we start to analyze limit behavior for general $\sum_{i=1}^{N} g_{ni}(\nu_{ni})$ case, it is interesting to consider one example of divisible statistics, namely spectral statistics $\mu_n(k) = \sum_{i=1}^{N} I\{\nu_{ni} = k\}$ we were dealing with in previous chapters. Taking advantage of powerful theory developed by Khmaladze we can obtain limit theorem for this statistics. Lets consider normalized version of spectral statistics:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} [I\{\nu_{ni} = k\} - \mathbb{E}I\{\nu_{ni} = k\}]$$

$$= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} [I\{\nu_{ni} = k\} - P\{\nu_{ni} = k\}].$$

If we assume that ν_{ni}-s are independent Poisson random variables, then limit theorem for normalized spectral statistics becomes trivial, namely it converges to normal random variable with expected value equal to 0 and variance σ^2

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} [I\{\nu_{ni} = k\} - P\{\nu_{ni} = k\}] \sim \mathbb{N}(0, \sigma^2)$$

where

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} [\pi(k; np_i) - \pi^2(k; np_i)]$$

and $\pi(k; np_i)$ denotes Poisson probability with intensity np_i.

Suppose now that we are interested in spectral statistics for "opinions" we considered in previous chapters, namely let $\mu_n(k)$ be number of "opinions" (questionnaire with q questions) in a sample with k supporters, then σ^2 can be written as follows:

$$\sigma^2 = \mathbb{E}_0\pi(k; np_i) - \mathbb{E}_0\pi^2(k; np_i)$$

where $\mathbb{E}_0$ denotes expected value taken with respect artificial uniform measure, similar to one we have in Section 5.1. Then again asymptotic behaviour of σ^2 depends on distribution of $M_{ni} = np_i$. Its asymptotic theory we have carefully investigated in Chapter 5.

7.2.1 Limit theorem for spectral statistics for independent frequencies in contiguity case

Let us consider spectral statistics for "opinions" we have defined in section 3.2 and suppose underlying distribution of probabilities is contiguous to uniform distribution as it was defined in Chapter 3 and 5. We can formulate following theorem:

Theorem 9. *Suppose probabilities $a_{1q}, \cdots, a_{qq}$ form a q triangular array, such that $\max_{1\le i\le q}|a_{iq} - \frac{1}{2}| \to 0$ and*

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

If the finite limit

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

exists, then

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} [I\{\nu_{ni} = k\} - P\{\nu_{ni} = k\}] \sim \mathbb{N}(0, \sigma^2)$$

with

$$\sigma^2 \sim \int \pi(k, \lambda e^z)\Phi_{-\frac{c^2}{2},c^2}(dz) - \int \pi^2(k, \lambda e^z)\Phi_{-\frac{c^2}{2},c^2}(dz)$$

7.2.2 Limit theorem for spectral statistics for independent frequencies in arbitrary distribution case

Now suppose underlying probabilities a_{iq}-s have arbitrary distribution. Notice that in this case the variance, $\sigma^2 \to 0$ as

$$\frac{1}{N}\sum_{i=1}^{N} \pi(k; np_i) \to 0$$

and

$$\frac{1}{N}\sum_{i=1}^{N} \pi^2(k; np_i) \to 0,$$

therefore we have to normalize the statistics $\mu_n(k)$ differently. Namely we will consider

$$\frac{1}{\sqrt{\mathbb{E}\mu_n}}\sum_{i=1}^{N}[I\{\nu_{ni} = k\} - P\{\nu_{ni} = k\}]$$

this statistics again converges to Gaussian random variable with expected value equal to 0 and variance σ^2,

Theorem 10. *Assume u_q is the solution of $\sum_{i=1}^{q} \psi_i'(u) = 0$. If $\{a_i(j)\}$ is such that the conditions*

$$c < \frac{1}{q}\sum_{i=1}^{q} \psi_i''(u_q) < C, \tag{7.2}$$

and

$$\left|e^{\sum_{i=1}^{q}[\psi_i(u_q+r)-\psi_i(u_q)]}\right| = o\left(\frac{1}{\sqrt{q}}\right) \quad \textit{uniformly in} \quad r > \delta > 0 \tag{7.3}$$

are satisfied, then

$$\frac{1}{\sqrt{\mathbb{E}\mu_n}}\sum_{i=1}^{N}[I\{\nu_{ni} = k\} - P\{\nu_{ni} = k\}] \sim \mathbb{N}(0, \sigma^2)$$

where

$$\sigma^2 \sim \frac{u\Gamma(k-u)}{\Gamma(1-u)\Gamma(k+1)} - \frac{u2^{2k-u-2}}{(k!)^2}\Gamma(2k-u)$$

and $u = \lim_{q\to\infty} u_q$.

Proof.

$$\frac{1}{\mathbb{E}\mu_n}\sum_{i=1}^{N}\pi(k;np_i) = \frac{\mathbb{E}\mu_n(k)}{\mathbb{E}\mu_n} \sim \frac{u\Gamma(k-u)}{\Gamma(1-u)\Gamma(k+1)}.$$

To analyze asymptotic behaviour of

$$\frac{1}{\mathbb{E}\mu_n}\sum_{i=1}^{N}\pi^2(k;np_i),$$

again, as in Section 1.2, let $G_n(x)$ be defined as

$$G_n(x) = \sum_{i=1}^{N} I\{np_i \geq x\}.$$

Then

$$\frac{1}{\mathbb{E}\mu_n}\sum_{i=1}^{N}\pi^2(k;np_i) \sim -\frac{1}{\mathbb{E}\mu_n}\int_0^\infty \pi^2(k;x)dG_n(x).$$

Using Theorem 5 from Section 3.5, we can write

$$\begin{aligned}
&-\frac{1}{\mathbb{E}\mu_n}\int_0^\infty \pi^2(k;x)dG_n(x)\\
&\sim -\int_0^\infty \pi^2(k;x)dR(x)\\
&\sim -\frac{1}{\Gamma(1-u)}\int_0^\infty \frac{e^{-2x}x^{2k}}{(k!)^2}dx^{-u}\\
&= \frac{u}{\Gamma^2(k+1)\Gamma(1-u)}\int_0^\infty e^{-2x}x^{2k-u-1}dx\\
&= \frac{u2^{2k-u-2}}{\Gamma^2(k+1)\Gamma(1-u)}\Gamma(2k-u).
\end{aligned}$$

Finally we can write

$$\sigma^2 \sim \frac{u\Gamma(k-u)}{\Gamma(1-u)\Gamma(k+1)} - \frac{u2^{2k-u-2}}{\Gamma(1-u)\Gamma^2(k+1)}\Gamma(2k-u).$$

□

7.2.3 Limit theorem for spectral statistics for dependent frequencies

Now let us drop the assumption about independence of ν_{ni}-s. In this case the methodology proposed in [13] becomes particularly important.

$$X_n(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} [I\{\nu_{ni} = k\} - \mathbb{E}I\{\nu_{ni} = k\}].$$

We will split $X_n(t)$ into two parts, martingale part $W_n(t)$ and compensator part $K_n(t)$.

$$W_n(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} [I\{\nu_{ni} = k\} - \mathbb{E}[I\{\nu_{ni} = k\} \mid \mathcal{F}_{i-1}^n]],$$

$$K_n(t) = \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} [\mathbb{E}[I\{\nu_{ni} = k\} \mid \mathcal{F}_{i-1}^n] - \mathbb{E}I\{\nu_{ni} = k\}].$$

Let us define $\tilde{n}_i$ and $\tilde{p}_i$ as follows

$$\tilde{p}_i = \frac{p_i}{1 - \sum_{j=1}^{i-1} p_j}$$

and

$$\tilde{n}_i = n - \sum_{j=1}^{i-1} \nu_{nj}.$$

In this example, for simplicity, instead of partial sums we will consider total sum. Then W_n is value of martingale in the last point with expected value 0 and variance

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}[(I\{\nu_{ni} = k\} - \pi(k; \tilde{n}_i \tilde{p}_i))^2 \mid \mathcal{F}_{i-1}^n]$$

or

$$\begin{aligned} VarW_n &= \frac{1}{N} \sum_{i=1}^{N} (\pi(k; \tilde{n}_i \tilde{p}_i) - \pi^2(k; \tilde{n}_i \tilde{p}_i)) \\ &= \mathbb{E}_0 \pi(k; \tilde{n}_i \tilde{p}_i) - \mathbb{E}_0 \pi^2(k; \tilde{n}_i \tilde{p}_i) \end{aligned}$$

$$\tilde{n}_i\tilde{p}_i \to np_i \ a.s.$$

and $\pi(k;\tilde{n}_i\tilde{p}_i)$ is uniformly integrable, therefore

$$\mathbb{E}_0\pi(k;\tilde{n}_i\tilde{p}_i) \to \mathbb{E}_0\pi(k;np_i)$$

$$\mathbb{E}_0\pi^2(k;\tilde{n}_i\tilde{p}_i) \to \mathbb{E}_0\pi^2(k;np_i).$$

Consequently

$$VarW_n \to \mathbb{E}_0\pi(k;np_i) - \mathbb{E}_0\pi^2(k;np_i).$$

Let us now consider compensator random variable K_n

$$K_n = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\pi(k;\tilde{n}_i\tilde{p}_i) - \pi(k;np_i)].$$

Let us denote np_i and $\tilde{n}_i\tilde{p}_i$ with λ_i and $\tilde{\lambda}_i$ respectively. Then

$$K_n = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}[\pi(k;\tilde{\lambda}_i) - \pi(k;\lambda_i)]$$

$$= -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\lambda_i^k e^{-\lambda_i} - \tilde{\lambda}_i^k e^{-\tilde{\lambda}_i}}{k!}.$$

Consider function $g(\tilde{\lambda}_i) = \tilde{\lambda}_i^k e^{-\tilde{\lambda}_i}$. Using linear approximation $g(\tilde{\lambda}_i)$ around λ_i, we can obtain

$$g(\tilde{\lambda}_i) - g(\lambda_i) \approx (\tilde{\lambda}_i - \lambda_i)g'(\lambda_i),$$

consequently

$$K_n \approx -\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\lambda_i - \tilde{\lambda}_i}{k!}[k\lambda_i^{k-1}e^{-\lambda_i} - \lambda_i^k e^{-\lambda_i}]$$

Consider the summands

$$\frac{\lambda_i - \tilde{\lambda}_i}{k!}[k\lambda_i^{k-1}e^{-\lambda_i} - \lambda_i^k e^{-\lambda_i}]$$

$$= (\lambda_i - \tilde{\lambda}_i)\pi(k;\lambda_i)(\frac{k}{\lambda_i} - 1)$$
$$= -(\frac{\tilde{\lambda}_i}{\lambda_i} - 1)\pi(k;\lambda_i)(k - \lambda_i)$$

$$\sqrt{N}(\frac{\tilde{\lambda}_i}{\lambda_i} - 1) = \sqrt{n}(\frac{\tilde{n}_i\tilde{p}_i}{np_i} - 1) =$$
$$= \sqrt{N}(\frac{1 - \hat{F}_n(\frac{i-1}{N})}{1 - F_n(\frac{i-1}{N})} - 1) = -\frac{v_n(\frac{i-1}{N})}{1 - F_n(\frac{i-1}{N})}$$

where $v_n(t) = \sqrt{N}(\hat{F}_n(t) - F_n(t))$ is empirical process. Finally if $\lambda_{Ns} \to f(s)$ we will obtain:

$$K_n \sim -\int_0^1 \frac{v_n(s)}{1 - F_n(s)}\pi(k; f(s))(k - f(s))ds.$$

7.3 Limit theorem for martingale part

In this section, we intend to remove the constraints of $n \sim N$ and $\sup |Np_{ni}| < \infty$. Under this framework, both Poissonian and Gaussian frequencies (in the sense that Y_{ni} can be approximated by a normal random variable) can be observed, which is more general and can be found in some real applications. To achieve this, we consider that h_{ni} satisfy

$$\lim_{n,N\to\infty} h_{n[Nt]}(y) = h(y,t)$$

($[s]$ being the integer part of s) and

$$|h_{ni}(y)| < be^{a|y|} \quad \text{for some} \quad a, b > 0 \tag{C1}$$

We will investigate some properties of the frequency ν_{ni} and the normalized frequency Y_{ni}.

Marginally, the distribution of ν_{ni} follows binomial distribution with sample size n and probability p_{ni}. It is well-known that as $\lim_{n\to\infty} np_{n[Nt]} = \lambda(t) < \infty$

$$Y_{n[Nt]} \xrightarrow{d} Y(t) = \frac{Z_{\lambda(t)} - \lambda(t)}{\sqrt{\lambda(t)}}$$

with

$$Z_{\lambda(t)} \sim \text{Poi}(\lambda(t))$$

and as $np_{n[Nt]} \to \infty$,

$$Y_{n[Nt]} \xrightarrow{d} Y(t) \sim \mathcal{N}(0,1),$$

under condition,

$$\sup_i p_{ni} \to 0. \tag{C2}$$

An important but not so obvious fact is,

Lemma 10. *If*

$$\inf_{n,i}(np_{ni}) \geqslant \delta^2 > 0 \tag{C3}$$

then for any $a, b > 0$, $\{be^{a|Y_{ni}|}\}$ is a sequence of uniformly integrable random variables over n and i.

Proof. Apply exponential inequality to Y_{ni}, it can be shown that for $y > 0$,

$$1 - F_{Y_{ni}}(y) = \mathbb{P}(Y_{ni} > y) \leqslant \exp\left(-\frac{y^2}{2}\psi\left(\frac{y}{\sqrt{np_{ni}}}\right)\right)$$

with

$$\psi(\lambda) = (2/\lambda^2)[(1+\lambda)\ln(1+\lambda) - \lambda] = (2/\lambda^2)\int_0^\lambda \ln(1+x)dx.$$

Since ψ is $\downarrow$ and

$$\psi(\lambda) \sim 2\log(\lambda)/\lambda$$

as $\lambda \to \infty$, for $\sqrt{np_{ni}} > \delta$,

$$\psi\left(\frac{y}{\sqrt{np_{ni}}}\right) \geqslant \psi\left(\frac{y}{\delta}\right)$$

and hence as $y \to \infty$,

$$\mathbb{P}(Y_{ni} > y) \leqslant e^{-\frac{y^2}{2}\psi\left(\frac{y}{\delta}\right)} \sim e^{-\delta y \log\left(\frac{y}{\delta}\right)} \ll e^{-(a+1)y},$$

Since $e^{-y^2\psi(y/\delta)/2}$ is monotonic on $y > 0$, there exists $c_1 > 0$ such that for all $y > \ln(c_1/b)/a$,

$$e^{-y^2\psi(y/\delta)/2} \leqslant e^{-(a+1)y}.$$

Therefore,

$$\int_{be^{ay}>c>c_1} be^{ay} dF_{Y_{ni}}(y) \leqslant b\left(\frac{c}{b}\right)^{-\frac{1}{a}} + \int_{be^{ay}>c_1} be^{-y} dy = 2b\left(\frac{c}{b}\right)^{-\frac{1}{a}} \tag{7.4}$$

On the other hand, for $y < 0$

$$F_{Y_{ni}}(y) = \mathbb{P}(Y_{ni} < y) \leqslant \exp\left(-\frac{y^2}{2}\psi\left(\frac{y}{\sqrt{np_{ni}}}\right)\right)$$

Since $(1+x)\ln(1+x) - x = \frac{x^2}{2} - \frac{x^3}{6} + \frac{x^4}{12} - \frac{x^5}{20} + \ldots \geqslant \frac{x^2}{2}$, $\psi(x) \geqslant 1$ for $x < 0$. Therefore,

$$-\frac{y^2}{2}\psi\left(\frac{y}{\sqrt{np_{ni}}}\right) \leqslant -\frac{y^2}{2}.$$

and as $y \to -\infty$,

$$\mathbb{P}(Y_{ni} < y) \leqslant e^{-\frac{y^2}{2}} \ll e^{(a+1)y}.$$

Since $e^{-y^2/2}$ is monotonic on $y < 0$, there exists $c_2 > 0$ such that for all $y < -\ln(c_2/b)/a$, $e^{-y^2/2} \leqslant e^{(a+1)y}$. Therefore,

$$\int_{be^{-ay}>c>c_2} be^{-ay} dF_{Y_{ni}}(y) \leqslant b\left(\frac{c}{b}\right)^{-\frac{1}{a}} + \int_{be^{-ay}>c} be^{y} dy = 2b\left(\frac{c}{b}\right)^{-\frac{1}{a}} \tag{7.5}$$

Since the right side of both (7.4) and (7.5) are uniform over n and i, and tend to 0 as $c \to \infty$, this implies,

$$\lim_{c\to\infty} \sup_{n,i} \mathbb{E}\left[be^{a|Y_{ni}|}\mathbf{I}\left\{be^{a|Y_{ni}|} > c\right\}\right] \to 0$$

hence the lemma has been proved.

□

If we consider the conditional distribution of ν_{ni} under $\mathcal{F}_{i-1}^n$, it is again a binomial random variable, only with sample size $\tilde{n}_{ni} = n - \sum_{j=1}^{i-1} \nu_{nj}$ and probability $\tilde{p}_{ni} = p_{ni}/(1 - \sum_{j=1}^{i-1} p_{nj})$. If we consider $F_n(t) = \sum_{i=1}^{Nt} p_{ni}$ as the distribution function defined by $\{p_{ni}\}$, $\hat{F}_n(t) = \frac{1}{N}\sum_{i=1}^{Nt} \nu_{ni}$ can be regarded as the empirical distribution of n F^n-distributed random variables and $\sup_t \left|\hat{F}_n(t) - F_n(t)\right| \to 0$ almost surely by Glivenko-Cantelli theorem. Therefore,

$$r_{ni} = \frac{\tilde{n}_{ni}\tilde{p}_{ni}}{np_{ni}} = \frac{1 - \hat{F}_n\left(\frac{i-1}{N}\right)}{1 - F_n\left(\frac{i-1}{N}\right)} \xrightarrow{a.s.} 1$$

and

$$\inf_i(\tilde{n}_{ni}\tilde{p}_{ni}) \geqslant \delta^2 \tag{7.6}$$

almost surely.

It is noteworthy that, for some $T < 1$ such that $\inf_n(1 - F_n(T)) > 0$,

$$\sup_{i \leqslant NT} |r_{ni} - 1| \xrightarrow{a.s.} 0 \tag{7.7}$$

If define $v_n(t) = \sqrt{n}(\hat{F}_n(t) - F_n(t))$ as the empirical process, and assume that

$$\sup_t |F_n(t) - F(t)| \to 0, \tag{C4}$$

then $v_n(t)$ converges to Brownian bridge $v(t)$ with respect to time $F(t)$.

By Dvoretzky-Kiefer-Wolfowitz inequality (see, e.g., [33]),

$$\mathbb{P}\left(\sup_t |v_n(t)| > \lambda\right) \leqslant 58e^{-2\lambda^2} \tag{7.8}$$

the following lemma can be established.

Lemma 11. *If conditions (C2) are satisfied, then as $n \to \infty$, for $i \leqslant NT$,*

$$\sup_i |\sqrt{np_{ni}}(r_{ni} - 1)| \leqslant \frac{\sup_i \sqrt{p_{ni}}}{\inf_n(1 - F_n(T))} \sup_i \left|v_n\left(\frac{i-1}{N}\right)\right| \xrightarrow{P} 0$$

Proof. Since $\sup_i p_{ni} \to 0$, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\sup_i \sqrt{p_{ni}}}{\inf_n(1 - F_n(T))} \sup_i \left|v_n\left(\frac{i-1}{N}\right)\right| > \epsilon\right) \leqslant 58e^{-2\left(\frac{\epsilon \inf_n(1-F_n(T))}{\sup_i p_{ni}}\right)^2} \to 0 \tag{7.9}$$

□

Based on these properties, if let $\xi_{ni} = h_{ni}(Y_{ni})$, we shall be able to establish the uniform integrability of ξ_{ni}^2 under conditional measure.

Lemma 12. *If conditions (C1-C3) are satisfied, then as* $n \to \infty$,

$$\lim_{c\to\infty} \sup_{i\leqslant NT} \mathbb{E}\left[\xi_{ni}^2 \mathbf{I}\left\{\xi_{ni}^2 > c\right\} |\mathcal{F}_{i-1}^n\right] \xrightarrow{P} 0 \tag{7.10}$$

Proof. Let $\tilde{Y}_{ni} = (\nu_{ni} - \tilde{n}\tilde{p}_{ni})/\sqrt{\tilde{n}\tilde{p}_{ni}}$ then

$$Y_{ni} = \sqrt{r_{ni}}\tilde{Y}_{ni} + \sqrt{np_{ni}}(r_{ni} - 1)$$

By lemma 11 and (7.7), we have,

$$\sup_{i\leqslant NT} \left|Y_{ni} - \tilde{Y}_{ni}\right| \xrightarrow{P} 0.$$

Since

$$\xi_{ni}^2 \leqslant b^2 e^{2aY_{ni}} \leqslant b^2 e^{2a\sup_{i\leqslant NT}\left|Y_{ni}-\tilde{Y}_{ni}\right|} e^{2a\tilde{Y}_{ni}},$$

by lemma 10 and (7.6), we shall get (7.10).

□

Then, we can establish the limit theorem for martingale part.

Theorem 11. *If the conditions (C1-C3) are satisfied, as* $n, N \to \infty$, *for* $t \leqslant T$,

$$W_{n,N}(t) \xrightarrow{d} W(t) = w\left(\tau(t)\right)$$

with w *being a standard Brownian motion and*

$$\tau(t) = \int_0^t \sigma^2(s)ds$$

with $\sigma^2(t)$ *being variance of* $h(Y(t), t)$.

Proof. Let $\eta_{ni} = \mathbb{E}\left(\xi_{ni}|\mathcal{F}_{i-1}\right)$. Then $W_{n,N}$ is a martingale with martingale differences

$$\frac{1}{\sqrt{N}}\left(\xi_{ni} - \eta_{ni}\right)$$

According to corollary 6 in [20], to prove the theorem, it is necessary and sufficient that

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}((\xi_{ni}-\eta_{ni})^2\,\mathbf{I}((\xi_{ni}-\eta_{ni})^2>\epsilon N)|\mathcal{F}_{i-1}^n)\xrightarrow{P}0 \qquad (\alpha)$$

and

$$\frac{1}{N}\sum_{i=1}^{Nt}\mathbb{E}((\xi_{ni}-\eta_{ni})^2\,|\mathcal{F}_{i-1}^n)\xrightarrow{P}\tau(t) \qquad (\beta)$$

to verify (α), it is sufficient to show that

$$\sup_i\mathbb{E}((\xi_{ni}-\eta_{ni})^2\,\mathbf{I}((\xi_{ni}-\eta_{ni})^2>\epsilon N)|\mathcal{F}_{i-1}^n)\xrightarrow{P}0 \qquad (7.11)$$

Since

$$(\xi_{ni}-\eta_{ni})^2\leqslant 2(\xi_{ni})^2+2(\eta_{ni})^2$$

and

$$\mathbf{I}((\xi_{ni}-\eta_{ni})^2>\epsilon N)\leqslant\mathbf{I}\left((\xi_{ni})^2>\frac{\epsilon N}{2}\right)+\mathbf{I}\left((\eta_{ni})^2>\frac{\epsilon N}{2}\right)$$

It is sufficient that all the following conditions are satisfied.

$$\sup_i\mathbb{E}\left[2(\xi_{ni})^2\mathbf{I}\left((\xi_{ni})^2>\frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right]\xrightarrow{P}0 \qquad (a)$$

$$\begin{aligned}&\sup_i\mathbb{E}\left[2(\eta_{ni})^2\mathbf{I}\left((\xi_{ni})^2>\frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right]\\ \leqslant&2\sup_i(\eta_{ni})^2\sup_i\mathbb{E}\left(\mathbf{I}\left((\xi_{ni})^2>\frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right)\xrightarrow{P}0\end{aligned} \qquad (b)$$

$$\begin{aligned}&\sup_i\mathbb{E}\left[2(\eta_{ni})^2\mathbf{I}\left((\eta_{ni})^2>\frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right]\leqslant\sup_i\mathbb{E}\left[2(\xi_{ni})^2\mathbf{I}\left((\eta_{ni})^2>\frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right]\\ &\leqslant 2\sup_i\mathbb{E}\left((\xi_{ni})^2|\mathcal{F}_{i-1}^n\right)\sup_i\mathbf{I}\left((\eta_{ni})^2>\frac{\epsilon N}{2}\right)\xrightarrow{P}0\end{aligned} \qquad (c)$$

(a) follows from (7.10) immediately. For (b) and (c), (7.10) imply that $\sup_i\mathbb{E}\left((\xi_{ni})^2|\mathcal{F}_{i-1}^n\right)$ and $\sup_i(\eta_{ni})^2$ are bounded and hence $\sup_i\mathbf{I}\left((\eta_{ni})^2>\frac{\epsilon N}{2}\right)$

and $\sup_i \mathbb{E}\left(\mathbf{I}\left((\xi_{ni})^2 > \frac{\epsilon N}{2}\right)|\mathcal{F}_{i-1}^n\right)$ vanishes in the limit. Therefore, (α) is satisfied.

For (β), consider the step functions

$$\varphi_n(t) = \mathbb{E}\left[\left(\xi_{n[Nt]} - \eta_{n[Nt]}\right)^2 |\mathcal{F}_{[Nt]-1}^n\right]$$

Then (7.11) implies that $\varphi_n(t) \xrightarrow{P} \sigma^2(t)$. By lemma 2, for all sufficiently large n, and $t \leqslant T$,

$$\varphi_n(t) \leqslant \mathbb{E}\left[\xi_{n[Nt]}^2 |\mathcal{F}_{[Nt]-1}^n\right] \leqslant \sup_{t \leqslant T} \mathbb{E}\left[\xi_{n[Nt]}^2 |\mathcal{F}_{[Nt]-1}^n\right] < \infty$$

Obviously, $\sup_{t\leqslant T} \mathbb{E}\left[\xi_{n[Nt]}^2 |\mathcal{F}_{[Nt]-1}^n\right]$ is integrable with respect to $t \in [0, T]$. By dominated convergence theorem, for $t \leqslant T$,

$$\int_0^t \varphi_n(s)ds \xrightarrow{P} \tau(t)$$

hence (β) and the theorem is proved. □

7.4 Limit theorems for compensator process

As in section 2, we develop some preliminary lemmas before we establish the theorem.

Lemma 13. *For $N \leqslant n$ and $c = o\left(\frac{N^{1/4}}{\sqrt{p}}\right)$, a binomial density $B(k, n, p)$ can be approximated by a Poisson density by*

$$B(k, n, p) = \frac{\pi(k, np)}{\sqrt{1-p}}\left(1 + o(1/\sqrt{N})\right) \tag{7.12}$$

in the range of $\frac{|k-np|}{\sqrt{np}} < c$.

Proof. Apply the Stirling's approximation,

$$\frac{B(k, n, p)}{\pi(k, np)} = e^{np-k}(1-p)^{n-k}\left(\frac{n}{n-k}\right)^{n-k+1/2} \frac{\left(1 + O\left(\frac{1}{n}\right)\right)}{\left(1 + O\left(\frac{1}{n-k}\right)\right)}$$

Since $\left|\frac{k}{n}-p\right| = O\left(c\sqrt{\frac{p}{n}}\right) \to 0$, Taylor's expansion shows that,

$$\ln\left(e^{np-k}(1-p)^{n-k}\left(\frac{n}{n-k}\right)^{n-k+1/2}\right)$$

$$= -\frac{1}{2}\ln(1-p) + \frac{1}{2(1-p)}\frac{k-np}{n} + \left(\frac{1}{2(1-p)^2} - \frac{n}{1-p}\right)\left(\frac{k-np}{n}\right)^2 + O\left(\left(\frac{k-np}{n}\right)^2\right)$$

$$= -\frac{1}{2}\ln(1-p) + \frac{1}{2(1-p)}O\left(c\sqrt{\frac{p}{n}}\right) - \left(\frac{1}{1-p}\right)O(c^2p).$$

Therefore,

$$B(k,n,p) = \frac{\pi(k,np)}{\sqrt{1-p}}\left(1 + O\left(\sqrt{\frac{c^2p}{n}}\right) + O(c^2p) + O\left(\frac{1}{n}\right) + O\left(\frac{1}{n-k}\right)\right)$$

For $N \leqslant n$ and $c = o\left(\frac{N^{1/4}}{\sqrt{p}}\right)$,

$$c^2p = o\left(\frac{1}{\sqrt{N}}\right)$$

and hence (7.12) holds. □

Let $\mathbb{E}_{np}$ denoting the expectation when ν follows binomial distribution with parameter n and p, while $\mathbb{E}_\lambda$ being the expectation when ν follows Poisson distribution with parameter $\lambda = np$. Then based on lemma10 and lemma13, we can show that

Lemma 14. *If conditions (C1-C3) hold,*

$$\sqrt{N}\sup_{i\leqslant NT} |\,\mathbb{E}_{np_{ni}}\xi_{ni} - \mathbb{E}_{\lambda_{ni}}\xi_{ni}| \to 0$$

Proof. let $c_{ni} = \frac{N^{1/8}}{\sqrt{p_{ni}}}$, then $\inf_i c_{ni} = \frac{N^{1/8}}{\sqrt{\sup_i p_{ni}}} \to \infty$. By lemma13, and note that $1 - 1/\sqrt{1-p} = O(p)$,

$$\sqrt{N}\sup_{i\leqslant NT} |\,\mathbb{E}_{np_{ni}}\left[\xi_{ni}\mathbf{I}\left\{|Y_{ni}| \leqslant c_{ni}\right\}\right] - \mathbb{E}_{\lambda_{ni}}\left[\xi_{ni}\mathbf{I}\left(|Y_{ni}| \leqslant c_{ni}\right)\right]|$$

$$= \left(o(1) + O(\sup_i p_{ni})\right)\sup_{i\leqslant NT}|E_{np_{ni}}\left[\xi_{ni}\mathbf{I}\left(|Y_{ni}| \leqslant c_{ni}\right)\right]| \to 0$$

On the other hand, by (7.4) and (7.5)

$$\begin{aligned}&\sqrt{N}\sup_{i\leqslant NT}\mathbb{E}_{np_{ni}}\left[|\xi_{ni}|\,\mathbf{I}\{|Y_{ni}|>c_{ni}\}\right]\\ \leqslant&\sqrt{N}\sup_{i\leqslant NT}\mathbb{E}_{np_{ni}}\left[be^{a|Y_{ni}|}\mathbf{I}\{be^{a|Y_{ni}|}>be^{ac_{ni}}\}\right]\to 0\end{aligned}$$

Since (7.4) and (7.5) also apply when ν_{ni} follows Poisson distribution with parameter np_{ni},

$$\sqrt{N}\sup_{i\leqslant NT}\mathbb{E}_{\lambda_{ni}}\left[|\xi_{ni}|\,\mathbf{I}\{|Y_{ni}|>c_{ni}\}\right]\to 0$$

and therefore the lemma holds. □

Lemma 14 easily applies to the case when n, p_{ni}, λ_{ni} are replaced by $\tilde{n}, \tilde{p}_{ni}, \tilde{\lambda}_{ni}$, if we realize that

$$\tilde{n}\geqslant\tilde{N}=N(1-F_n^n(\frac{i-1}{N}))\to\infty$$

and

$$\inf_i\tilde{c}_{ni}=\frac{N^{1/8}}{\sqrt{\sup_i\tilde{p}_{ni}}}\to\infty$$

almost surely.

Theorem 12. *If conditions (C1-C4) hold, and for $t<T$*

$$K_{n,N}(t)\xrightarrow{d}K(t)=\int_0^t\frac{\mathbb{E}\left[h(Y)Y\right]}{\sqrt{f(s)}}\frac{v(s)}{1-F(s)}dF(s)$$

Proof. Let

$$K_{n,N}^{\lambda}(t)=\frac{1}{\sqrt{N}}\sum_{i=1}^{Nt}\left[\mathbb{E}_{\tilde{\lambda}_{ni}}\xi_{ni}-\mathbb{E}_{\lambda_{ni}}\xi_{ni}\right]$$

by lemma14, for $t\leqslant T$,

$$\left|K_{n,N}(t)-K_{n,N}^{\lambda}(t)\right|\to 0.$$

If let $y=(k-r\lambda)/\sqrt{r\lambda}$ and denote $\pi(k,\lambda)$ by poisson probability with intensity λ, then Taylor's expansion of $\pi(k,r\lambda)$ w.r.t r around 1 shows,

$$\pi(k,r\lambda)-\pi(k,\lambda)=y\pi(k,\lambda)\sqrt{\lambda}(r-1)+O\left(\lambda(r-1)^2\right)$$

(7.8) implies that $\sup_i n(r_{ni}-1)^2/\sqrt{N} \xrightarrow{p} 0$, and hence $\lambda_{ni}(r_{ni}-1)^2/\sqrt{N} = o_p(p_{ni})$. Therefore,

$$\begin{aligned}
&\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[\mathbb{E}_{\tilde{\lambda}_{ni}} \xi_{ni} - \mathbb{E}_{\lambda_{ni}} \xi_{ni} \right] - \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[\mathbb{E}_{\lambda_{ni}} \left[\xi_{ni} Y_{ni} \right] \sqrt{np_{ni}} (r_{ni}-1) \right] \right| \\
= &\left| \frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} O\left(\lambda (r-1)^2 \right) \right| \xrightarrow{p} 0
\end{aligned}$$

Since for $t \leqslant T$, $NP_{n[Nt]} \to f(t)$, $\mathbb{E}_{\lambda_{n[Nt]}} \left[\xi_{n[Nt]} Y_{n[Nt]} \right] \to \mathbb{E}\left[h(Y(t)) Y(t) \right]$, $\sqrt{n}(r_{n[Nt]} - 1) \xrightarrow{d} \frac{v(t)}{1-F(t)}$, and

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{Nt} \left[\mathbb{E}_{\lambda_{ni}} \left[\xi_{ni} Y_{ni} \right] \sqrt{np_{ni}} (r_{ni}-1) \right] = \sum_{i=1}^{Nt} \frac{\mathbb{E}_{\lambda_{ni}} \left[\xi_{ni} Y_{ni} \right]}{\sqrt{Np_{ni}}} \sqrt{n} (r_{ni}-1) p_{ni} \xrightarrow{d} K(t)$$

the theorem has been proved.

□

Chapter 8

Conclusion

This thesis studies statistical analysis of the diversity of multiple choice questionnaires in the context of a large number of rare events (LNRE). In Chapter 1 (Introduction) the background of LNRE was discussed, corresponding definitions, conditions and results were given; and several data sources of LNRE were discussed. In Chapter 2 (Models and Laws of LNRE) we reviewed some of the existing models, laws and distributions related to LNRE. The topic of Chapter 3 is questionnaire in LNRE, in which we discussed results obtained in [15], namely, analysis and review of LNRE in the case of binary (Yes/No) questionnaires. Chapter 4 (Measures of Diversity) contains survey of diversity measures in the context of LNRE.

The main results of the thesis are presented in Chapter 5. It contains the statistical analysis of questionnaires with multiple answers, which is the generalization of the problem considered in Chapter 3. In the process of discussion, advanced mathematical and probabilistic tools such as Contiguity theory, probability of large deviations, Esscher's transform and Edgeworth Expansion were employed. It was shown that approach and techniques used in [15] are universal and can be generalized to more complex situations. Chapter 6 (Some numerical observations) contains empirical justification of the theoretical results obtained in Chapter 5. Data from

several simulations were applied to the model. Chapters 5 and 6 together present one of the most striking results of the thesis. It is demonstrated particularly that dependence of the diversity of responses on the underlying distribution of the questionnaire is quite insignificant. In Chapter 7 we discuss functional martingale limit theorems for divisible statistics. Results for divisible statistics in the LNRE context were obtained. Also, limit theorems for general divisible statistics considered by Khmaladze in [13] were generalized.

List of Tables

List of Figures

Bibliography

[1] BAAYEN, R. H. *Word Frequency Distributions*. Springer, 2001.

[2] BEISEL, J., AND MORETEAU, J. A simple formula for calculating the lower limit of shannon's diversity index. *Ecological Modelling 99*, 2-3 (1997), 289–292.

[3] FAIRTHORNE, R. A. Empirical hyperbolic distributions (bradford-zipf-mandelbrot) for bibliometric description and prediction. *Journal of Documentation 61*, 2 (2005), 171–193.

[4] FELLER, W. *An introduction to probability theory and its applications, Vol. 2*, 2nd ed., vol. 2 of *Wiley series in probability and mathematical statistics*. New York, Wiley, 1971.

[5] GOOD, I. The population frequencies of species and the estimation of population parameters. *Biometrica 40*, 3/4 (1953), 237–264.

[6] GREENWOOD, P. E., AND SHIRYAYEV, A. N. *Contiguity and the Statistical Invariance Principle*, vol. 1 of *Stochastics Monographs - Theory and Applications of Stochastic Processes*. Gordon and Breach Science Publishers, 1985.

[7] HILL, B. M. Zipf's law and prior distrbution for the compostion of a population. *Journal of Amercan Statstcal Association. 65* (1970), 1220–1232.

[8] HURLBERT, S. H. The nonconcept of species diversity: A critique and alternative parameters. *Ecology 52*, 4 (1971), 577–586.

[9] IVCHENKO, G. I., AND MEDVEDEV, Y. I. Separable statistics and hypothesis testing. the case of small samples. *Theory of probability and its applications 23*, 4 (1978), 764–775.

[10] IVCHENKO, G. I., AND MEDVEDEV, Y. I. Decomposable statistics and hypothesis testing for grouped data. *Theory of probability and its applications 25*, 3 (1980), 540–551.

[11] KALLENBERG, O. *Foundations of Modern Probability*, 2nd ed. Probability and Its Applications. Springer, 2001.

[12] KARLIN, S. Central limit theorems for certain infinite urn schemes. *Indiana University Mathematics Journal 17* (1968), 373–401.

[13] KHMALADZE, E. V. Martingale limit theorems for divisible statistics. *Theory of Probability and its Applications 28*, 3 (1984), 530–548.

[14] KHMALADZE, E. V. The statistical analysis of a large number of rare events. *CWI Report MS-R8804* (1988).

[15] KHMALADZE, E. V. Diversity of responses in questionnaires and similar objects. *MSOR Report 09-3, Victoia University of Wellington* (2009).

[16] KHMALADZE, E. V., AND CHITASHVILI, R. J. The statistical analysis of a large number of rare events and the related problem. *Proc. Tbilisi Mathematical Institute 92* (1989), 196–245.

[17] KHMALADZE, E. V., AND TSIGROSHVILI, Z. P. On polinomial distributions with large number of rare events. *Math. Methods Stat. 2*, 3 (1993), 240–247.

[18] KLAASSEN, C. A., AND MNATSAKANOV, R. M. Consistent estimation of the structural distribution function. *Scandinavian Journal of Statistics 27*, 4 (2000), 733–746.

[19] KOLASSA, J. E. *Series Approximation Methods in Statistics*, vol. 88 of *Lecture Notes in Statistics*. Springer-Verlag New York, 2006.

[20] LIPSTER, R., AND SHIRYAEV, A. A functional central limit theorem for semimartingales. *Theory of probability and its applications 25*, 4 (1980), 667–688.

[21] LOTKA, A. J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences 16*, 12 (1926), 317–324.

[22] MACARTHUR, R. H. On the relative abundace of bird species. *Proceedings of the National Academy of Sciences of the United States of America 43*, 3 (1957), 293–295.

[23] MAGURRAN, A. E. *Measuring Biological Diversity*. Blackwell Science Ltd., 2004.

[24] MANDELBROT, B. An information theory of the statistical structure of language. *Proceedings of the Symposium on Applications of communication Theory* (1952), 486–500.

[25] MCINTOSH, R. P. An index of diversity and the relation of certain concepts to diversity. *Ecology 48*, 3 (1967), 392–404.

[26] MNATSAKANOV, R. M., AND KLAASSEN, C. A. Estimation of the mixing distribution in poisson mixture models: uncensored and censored samples. *Proceedings of Hawaii international conference on Statistics and Related fields*.

[27] MORRIS, C. Central limit theorems for multinomial sums. *Annals of statistics 3*, 1 (1975), 165–188.

[28] OOSTERHOFF, J., AND ZWET, W. R. V. A note on contiguity and hellinger distance. *Contributions to Statistics, Reidel, Dordrecht* (1979), 157–166.

[29] PIELOU, E. *Ecological Diversity*. New York, Wiley., 1975.

[30] ROUAULT, A. Loi de zipf et sources markoviennes. *Annales de l'Institute H. Poincare 14* (1978), 169–188.

[31] ROWLANDS, I. Emerald authorship data, lotka's law and research productivity. *Aslib Proceedings 57*, 1 (2005), 5–10.

[32] SAGER, P. E., AND HASLER, A. D. Species diversity in lacustrine phytoplankton. i. the components of the index of diversity from shannon's formula. *The American Naturalist 103*, 929 (1969), 51–59.

[33] SHORACK, G. R., AND WELLNER, J. A. *Empirical Processes With Applications to Statistics*. Wiley series in probability and Mathematical Statistics, 1986.

[34] SIMPSON, E. H. Measurement of diversity. *Nature 163* (1949), 688.

[35] SKIRSTYMONSKAIA, V., AND SOFER, M. Water and ice of oceans. *Science and Life (Nauka i Zhizn) 8* (1980), 42–49.

[36] VAN ES, B., KLAASSEN, C. A., AND MNATSAKANOV, R. Estimating the structural distribution function of cell probabilities. *Austrian Journal of Statistics 32* (2003), 85–98.

[37] WILLIAMS, C. *Patterns in the Balance of Nature*. Academic Press, 1964.

[38] YULE, G. U. A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philosophical Transactions of the Royal Society of London. 212* (1924), 21–87.