Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

E.V. Khmaladze

The statistical analysis of a large number of rare events

Department of Mathematical Statistics Report MS-R8804 April

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O.).

The Statistical Analysis of a Large Number of Rare Events

E.V. Khmaladze
Steklov Mathematical Institute, Moscow
and
Razmadze Mathematical Institute, Tbilisi

V.A. Steklov Mathematical Institute
Vavilova 42, 117966 Moscow GSP-1, USSR

This paper presents a theory of the statistical analysis of a large number of rare events, linking together problems of the variety of species and vocabulary, chi-square tests with a large number of cells, and models with a growing number of parameters.

1980 Mathematics Subject Classification: 62E20
Key Words and Phrases: chi-square test, histogram, spectral statistics.

PREFACE

This paper contains results of a not very intensive but rather long consideration of specific statistical problems which are sometimes called "statistical problems in relatively small samples", but we prefer to call them "the statistical analysis of a large number of rare events".

It was some five years ago that prof. R. Chitashvili, who was already involved jointly with prof. Yu. Orlov in the statistical analysis of classical texts, asked me the question "what is the difference between water and watermelon from the statistical point of view?" And when I realized that a good answer is not trivial I found myself involved in this peculiar area, where frequencies do not estimate probabilities.

The present paper is not focussed on the solution of important specific problems like estimation of "vocabulary" or "number of different species". Neither is it a review - in fact, our list of references is very narrow. The aim of this paper is to demarcate the situation of a large number of rare events by appropriate definitions and to clarify its pecularity, and also to show some significant connections between this area and several other areas of statistical theory.

The present text was prepared with the helpful attention of my colleagues at 'Centrum voor Wiskunde en Informatica' (CWI) in Amsterdam for a lecture there. In particular, it was my pleasant discovery that prof. R.D. Gill also contributed to the area of the analysis of a large number of rare events by dealing with a beautiful problem of coinage in medieval Holland (GILL 1983, 1984, STAM 1987). Critical comments of dr. K. Dzaparidze and prof. R.D. Gill lead to improvements to the manuscript. Also Mr. Roos (CWI, Library) attracted my attention to a whole field of application of Zipf-Mandelbrot's law, Lotka's law, etc. - the study of the structure of documentation (see the sequence of corresponding publications in the *Journal of Documentation* for many years).

This text, as my part of a joint paper with prof. R. Chitashvili, will be published also in Tbilisi in the memorial volume for prof. G. Mania.

Report MS-R8804
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

Contents

§ 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 nearly any statistical analysis devoted to the study of the variety of species one has to deal with what might be called “a large number of rare events”.

Consider a few examples.

(a) Data concerning frequencies of different words in separate novels, other pieces of text and even in a whole language are illustrated by Table 1.

Works	Total word usage	No. of different words (vocabulary)	No. of words used only once
Dante *Divina Comedia*	101.554	13.004	ca. 6.500
Pushkin *Complete Works*	544.777	21.197	6.388
Byron *Don Juan*	130.745	14.411	ca. 7.200

“Woord frequenties in geschreven en gesproken Nederlands” * ed. P.C. Uit den Boogaart Oosthoek, Scheltema & Holkema (Utrecht).	Total no. of words in sample	No. of words used only once	No. of words used only twice	No. of words used three times
	720.000	51.372	10.306	4.544

Frequency of word	
De, de	32.843
Hoogleraar **	20
Voetballer ***	2

* Word frequencies in written and spoken Dutch.

** Professor.

*** Soccerplayer.

Table 1.

(b) Data concerning variety of species for various kinds of animals and plants

(c) Data on chemical analysis of a substance 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 amount of these nine ions exceeds 99,9% of the whole amount of all dissolved salts" (see V. Skirstymonskaia, M. Sofer, Water and ice of oceans, Science and Life (*Nauka i Zhizn)* 1980, Vol. 8, pp. 42-49, the same source for Table 2 below).

% of total amount of inorganic admixtures								
Cl^-	Na^+	SO_2	Mg^{++}	Ca^{++}	K^+	$HCO_3 + CO_2$	Br^-	$H_3 BO_3$
55.04	30.61	7.68	3.69	1.16	1.10	0.41	0.19	0.07
Total 99.95								

TABLE 2.

According to Table 2 the remaining of more then a hundred dissolved elements take up only 0.05% of the total amount of inorganic admixtures.

(d) 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.

No. of different nationalities in the population of Tbilisi	Nationalities	%
80	Georgians	62.2
	Armenians	14.2
	Russians	12.2
	Ossetians	2.7
	Greeks	1.5
	Jews	1.4
	Azerbajanians	1.2
	Ukrainians	1.2
and only 0.9% for more then 70 others		

TABLE 3.

The common feature of these and similar examples is that except for several frequent items (events) there is also a very large number of very rare items (events). The total amount of these rare events compared to the number of observations is sometimes not large but the number of these events among all different observed events is always very significant. In spite of being rare these events are usually very important. For instance the number of words used in *Divina Comedia* only once, according to Table 1, is approximately 6.500. A text of 6.500 words, which is approximately 6-7 chapters, can be considered not of vital importance for *Divina Comedia* which contains 100 chapters. But it is very clear that these rare 6.500 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. And if we dissolve in distilled water the nine ions of Table 2 what we get will be quite far from what the ocean water means for any of us. In particular, Tabel 2 does not contain iodine, to whose presence man is so sensitive.

Sometimes it is useful for statisticians to study themselves. Table 4 contains data extracted from the Author Index of *Theoria Verojatn. i primen*, 1955-1970, by I. Urinov. According to this data the number of authors who published only one paper in the journal during 25 years is nearly one-half of all 741 authors of papers in the journal.

No. of publications, m	1	2	3	4	5	6	$\cdots$	32	39	43
No. of authors publishing m papers	366	135	67	41	31	19	$\cdots$	1	1	1

TABLE 4.

The present paper is mainly concerned with another, this time artificial, source of a large number of rare events. Namely, let $X_1, \ldots, X_n$ be i.i.d. continuous random variables, distributed over a finite interval $[a, b]$. Divide this interval into N equal subintervals $[a+i\Delta, a+(i+1)\Delta]$, $\Delta = 1/N$, $i=0, \ldots, N-1$. For small Δ the event $X_j \in [a+i\Delta, a+(i+1)\Delta]$ has small probability, but the number N of such events is large. Let ν_{in} be the frequency of X's with values in the i-th subinterval. The main part of this paper is connected with the behaviour of the vector $\nu_n = (\nu_{1n}, \ldots, \nu_{Nn})$ of frequencies when both n and N are large. It will be observed, in particular, 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 then n.

§ 2. DEFINITIONS OF A LARGE NUMBER OF RARE EVENTS. NECESSARY AND SUFFICIENT CONDITIONS THROUGH G- AND Q-FUNCTIONS

Consider a random vector $\nu_n = (\nu_{1n}, \ldots, \nu_{nn})$ which has a multinomial distribution $\mathfrak{M}(\mathbf{p}_n, n)$ with vector of probabilites $\mathbf{p}_n = (p_{1n}, \ldots, p_{Nn})$ and sample size n, i.e.

$$\mathrm{P}\{\nu_{in} = k_i,\ i=1, \ldots, N\} = \frac{n!}{\prod_{i=1}^{N} k_i!} \prod_{i=1}^{N} p_{in}^{k_i},\ \sum_{i=1}^{N} k_i = n,\ k_i \geqslant 0,\ \sum_{i=1}^{N} p_{in} = 1\,.$$

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

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

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. The vector $\{\mu_n(1), \mu_n(2), \ldots, \mu_n(n)\}$ is sometimes called the set of spectral statistics. The statistic μ_n is called the observed vocabulary, or, sometimes, simply, the vocabulary.

The marginal distribution of each frequency is binomial:

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

Hence

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

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

DEFINITION 1. *A sequence* $\{\boldsymbol{\nu}_n\}$ *(of random vectors* $\boldsymbol{\nu}_n=(\nu_{1n}, \ldots, \nu_{Nn})$*) is called a sequence with a large number of rare events (LNRE) if*

$$\liminf_{m\to\infty}\frac{\mathrm{E}\mu_n(1)}{n}>0 . \tag{d.1}$$

DEFINITION 2. *A sequence* $\{\boldsymbol{\nu}_n\}$ *is called an LNRE sequence if*

$$\liminf_{n\to\infty}\frac{\mathrm{E}\mu_n(1)}{\mathrm{E}\mu_n}>0 \textit{ and } \lim_{n\to\infty}\mathrm{E}\mu_n=\infty . \tag{d.2}$$

These two definitions are not equivalent, namely (d.1) ⇒ (d.2), but not vice versa.

It is easy to observe that for a fixed finite N and fixed vector of probabilities $\mathbf{p}$ each frequency $\nu_{in}\to\infty$ a.s. as $n\to\infty$ and therefore $\mu_n(1)\to0$ and $\mu_n\to N$ a.s. Consequently (d.1) and (d.2) cannot be satisfied. So the question is in what way should $\mathbf{p}_n$ (and N) vary as $n\to\infty$ to satisfy (d.1) and (d.2).

To answer this question it is convenient to introduce the following two functions:

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

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

CONDITION 1. *For some* $z<\infty$

$$\liminf_{n\to\infty} Q_n\left(\frac{z}{n}\right)>0 \tag{c.1}$$

CONDITION 2. *For some* $z<\infty$

$$\limsup_{n\to\infty}\frac{G_n\left(\frac{z}{n}\right)}{nQ_n\left(\frac{z}{n}\right)}<\infty \textit{ and } \lim_{n\to\infty}nQ_n\left(\frac{z}{n}\right)=\infty . \tag{c.2}$$

The following lemmas state that these are necessary and sufficient conditions for (d.1) and (d.2) respectively.

LEMMA 1. (*c.* 1)⇔(*d.* 1). *Besides,*

$$\liminf_{n\to\infty}\frac{\mathrm{E}\mu_n(m)}{n}>0\Rightarrow(c.\ 1) .$$

PROOF. From the equality

$$\mathrm{E}\mu_n(1)=n\sum_{i=1}^{N}p_{in}(1-p_{in})^{n-1}\left[I\left\{p_{in}\leqslant\frac{z}{n}\right\}+I\left\{p_{in}>\frac{z}{n}\right\}\right]$$

one can easily derive that

$$\mathrm{E}\mu_n(1)\geqslant n[e^{-z}+o(1)]Q_n(\frac{z}{n}) \tag{2}$$

which means that (c.1) ⇒ (d.1). On the other hand for $z\geqslant 1$

$$\mathrm{E}\mu_n(1)\leqslant nQ_n(\frac{z}{n})+[ze^{-z}+o(1)]G_n(\frac{z}{n})\,. \tag{3}$$

Since $G_n(\frac{z}{n})\leqslant n$ and ze^{-z} can be chosen arbitrarily small for sufficiently large z, from (3) it follows that (d.1) ⇒ (c.1). The proof of the last statement of Lemma 1 we leave to the reader. □

LEMMA 2. (c.2) ⇔ (d.2).

PROOF. Using the inequality

$$1-(1-p)^n\leqslant np$$

one can easily derive that

$$\mathrm{E}\mu_n\leqslant nQ_n(\frac{z}{n})+G_n(\frac{z}{n})\,. \tag{4}$$

From (2) and (4) it follows that

$$\frac{\mathrm{E}\mu_n(1)}{\mathrm{E}\mu_n}\geqslant\frac{[e^{-z}+o(1)]Q_n(\frac{z}{n})}{Q_n(\frac{z}{n})+\frac{1}{n}G_n(\frac{z}{n})}\,. \tag{5}$$

It is also true that

$$\mathrm{E}\mu_n\geqslant n\sum_{i=1}^{N}\frac{1-(1-p_{in})^n}{np_{in}}p_{in}I\{p_{in}\leqslant\frac{z}{n}\}\geqslant n\left[\frac{1-e^{-z}}{z}+o(1)\right]Q_n(\frac{z}{n})\,. \tag{6}$$

From (5) and (6) follows (c.2) ⇒ (d.2). Now consider the inequality

$$\mathrm{E}\mu_n\geqslant\sum_{i=1}^{N}[1-(1-p_{in})^n]I\{p_{in}>\frac{z}{n}\}\geqslant n[1-e^{-z}+o(1)]G_n(\frac{z}{n}) \tag{7}$$

which jointly with (3) gives

$$\frac{\mathrm{E}\mu_n(1)}{\mathrm{E}\mu_n}\leqslant\frac{nQ_n(\frac{z}{n})}{[1-e^{-z}+o(1)]G_n(\frac{z}{n})}+\frac{ze^{-z}}{1-e^{-z}}+o(1)\,.$$

Choose z in such a way that

$$\liminf_{n\to\infty}\frac{\mathrm{E}\mu_n(1)}{\mathrm{E}\mu_n}>\frac{ze^{-z}}{1-e^{-z}}\,.$$

Then

$$\liminf_{n\to\infty}\frac{\mathrm{E}\mu_n(1)}{\mathrm{E}\mu_n}>0\Rightarrow\liminf_{n\to\infty}\frac{G_n(\frac{z}{n})}{nQ_n(\frac{z}{n})}<\infty\,.$$

From this inequality and inequality (4) it follows that (d.2) ⇒ (c.2). □

Let us remark in conclusion of this section that 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

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

or if

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

or if

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

But it seems at present that Definitions 1 and 2 are the most interesting ones among these many possibilities.

§ 3. Properties of G- and Q-functions

We have seen that the functions G_n and Q_n defined in (1) of §2 proved to be of some use because they led to necessary and sufficient conditions for LNRE. These functions will also be useful below. In the present §3 we will rewrite these functions in a more natural way and study some of their properties.

Since a discrete variable might be sometimes less convenient then a continuous one let us associate with the vector of probabilities $\mathbf{p}_n$ two densities, p_n and f_n:

$$p_n(t)=\sum_{i=1}^{N} p_{in} I\{i-1\leqslant t<i\}\ ,$$

$$f_n(t)=\sum_{i=1}^{N} np_{in} I\{\frac{i-1}{n}\leqslant t<\frac{i}{n}\}=np_n(nt)\ .$$

Now consider a density f of a distribution F absolutely continuous with respect to, say, Lebesgue measure, and define

$$G_f(z)=\int I\{f(t)>z\}\mathrm{d}t$$

$$Q_f(z)=\int I\{f(t)\leqslant z\}F(\mathrm{d}t)\ . \qquad (1)$$

Let us call these functions G- and Q-functions of a density f.

One can easily see that

$$G_{f_n}(z)=\frac{1}{n}G_{p_n}(z)\ ,$$

$$G_{p_n}(z)=G_n(\frac{z}{n}) \qquad (2)$$

$$Q_{f_n}(z)=Q_n(\frac{z}{n})\ .$$

Consider some of the properties of G_f and Q_f.

(G.1). $G_f\downarrow$, $G_f(z)\leqslant 1/z$, $\inf\{z:G_f(z)=0\}=$ ess sup f and $G_f(0+)$ is equal to the length of the support of f.

(G.2). Let λ be a Lebesgue measure preserving transformation, i.e.

$$\int_B \mathrm{d}t=\int_{\lambda^{-1}(B)} \mathrm{d}t$$

and let $f\circ\lambda^{-1}$ be the density of $F\circ\lambda^{-1}(B)=F[\lambda^{-1}(B)]$. Then

$$G_f = G_{f \circ \lambda^{-1}}$$

EXAMPLE.

$$f_a(x) = f(x-a) \Rightarrow G_{f_a} = G_f\ ;$$

$$f_-(x) = f(-x) \Rightarrow G_{f_-} = G_f\ .$$

(G.3). Formula of change of variable:

$$\int \phi[f(t)]\mathrm{d}t = -\int \phi(z) G_f(\mathrm{d}z)\ .$$

EXAMPLE.

$$\int f(t)\mathrm{d}t = 1 \Rightarrow -\int z G_f(\mathrm{d}z) = 1\ .$$

(G.4). A function $G(z), z \geqslant 0$, is the G-function of a density f iff $G\downarrow$ and $-\int zG(\mathrm{d}z) = 1$. Namely, under these conditions the function

$$G^{-1}(t) = \inf\{z : G(z) \leqslant t\}$$

is a density and $G = G_f$ with $f = G^{-1}$.

Consider also continuity properties of G_f with respect to f.

(G.5). Let I_a be the distribution concentrated at a point a and let $\{F_n\}$ be a sequence of absolutely continuous distributions, then

$$F_n \xrightarrow{\mathscr{W}} I_a \Rightarrow G_{f_n} \to 0 \text{ on } (0, \infty]\ .$$

Denote

$$\|f_1 - f_2\| = \int |f_1(t) - f_2(t)|\,\mathrm{d}t\ .$$

(G.6).

$$\|f_n - f\| \to 0 \Rightarrow G_{f_n} \xrightarrow{\mathscr{W}} G_f\ .$$

EXAMPLE.

(a) Let

$$f_n(t) = (1 + \tfrac{1}{n}) I\{0 \leqslant t \leqslant \tfrac{1}{2}\} + (1 - \tfrac{1}{n} I\{\tfrac{1}{2} < t \leqslant 1\}\ .$$

Then $G_{f_n}(z) \not\to G_f(z)$ for $z = 1$. Therefore $\|f_n - f\| \to 0 \not\Rightarrow G_{f_n} \to G_f$ for all z.

(b) Let $f_n = 1 + \chi_n$, where χ_n is the n-th Haar function on [0,1], and let $\mathscr{U}$ be the uniform distribution on [0,1]. Then $F_n \xrightarrow{\mathscr{W}} \mathscr{U}$ but

$$G_{f_n}(z) = \tfrac{1}{2} I\{0 < z \leqslant 2\} \not\to G_f(z) = I\{0 < z \leqslant 1\}\ .$$

Therefore

$$F_n \xrightarrow{\mathscr{W}} F \not\Rightarrow G_{f_n} \xrightarrow{\mathscr{W}} G_f\ .$$

These examples show that (G.6) cannot easily be strengthened.

The following property is in some sense an inverse of (G.6).

(G.7). Let $\mathfrak{F}_G=\{f:G_f=G\}$ be the class of all densities with a given G-function. Then

$$G_n\xrightarrow{\mathfrak{W}}G\Rightarrow\exists\{f_n\}\text{ and }f:f_n\in\mathfrak{F}_{G_n},\ f\in\mathfrak{F}_G,\ \text{ and } t-\text{a.e. } f_n(t)\rightarrow f(t)\,.$$

In particular,

$$G_n^{-1}(t)\rightarrow G^{-1}(t)\quad t-\text{a.e.}\,.$$

Turning to Q-functions one can observe that

$$Q_f(z)=-\int_0^z xG_f(\mathrm{d}x) \tag{3}$$

which determines many of Q_f's properties. We formulate only one.

(Q.1) Any distribution function Q on $[0,\ \infty]$ continuous at $z=0$ is Q-function of some density. In particular,

$$Q=Q_f \text{ with density } f=G^{-1}$$

which is the inverse function of

$$G(z)=-\int_z^\infty \frac{1}{x}Q(\mathrm{d}x)\,.$$

The functions G_n and Q_n defined in (1) of § 2 are usually considered as something very specific for the statistical analysis of LNRE. Definitions (1) and (2) of this § 3 show that these functions occur much more widely. As a matter of fact if f is the density of a distribution F with respect to H, $f=\mathrm{d}F/\mathrm{d}H$, then the distribution functions of the likelihood $f(X)$ under H and under F are precisely the G- and Q-functions of the density f respectively.

Consider now two specific cases when condition (c.1) and (c.2) are satisfied.

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

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

If $N\rightarrow\infty$ then $f_n(t)\rightarrow f(t)$ for a.e. $t\in[0,1]$ and, according to (G.6) and (3),

$$G_{f_N}\xrightarrow{\mathfrak{W}}G_f,\quad Q_{f_N}\xrightarrow{\mathfrak{W}}Q_f\,. \tag{4}$$

But since $Q_n(\frac{z}{n})=Q_{f_n}(\alpha z)$ for $N=\alpha n$, α being a positive constant, (4) means that condition (c.1) is satisfied. Therefore the frequencies $\nu_{1n}, \ldots, \nu_{Nn}$, where

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

satisfy Definition 1.

Case 2. Let p be a nonincreasing density on $(0,\ \infty)$, i.e. $p=G_p^{-1}$, and let

$$p_{in}=p_i=\int_{i-1}^i p(t)\mathrm{d}t\,. \tag{5}$$

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

$$\nu_{in} = \sum_{j=1}^{n} I\{i-1 \leqslant X_j < i\}\ , \tag{6}$$

Lemma 1. *For any fixed p the sequence $\{\nu_n\}$ of vectors of frequencies (6) does not satisfy (d.1).*

Proof. This follows from Lemma 1 of § 2 since $Q(\frac{z}{n}) \to 0$ for all $z < \infty$ and $n \to \infty$. □

Now consider

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

$$p(t) = t^{-\rho} L(t) \tag{c.3}$$

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

Lemma 2. *Let p be a fixed density and p_i and ν_{in} be defined by (5) and (6). Then (c.3) $\Leftrightarrow$ (d.2).*

Proof. According to Theorem 1.a) in ch. VIII, § 9 of (Feller, 1971) condition (c.3) is equivalent to

$$\frac{tp(t)}{1-P(t)} \to 1-\rho, \quad P(t) = \int_0^t p(s)\mathrm{d}s\ . \tag{7}$$

Denote $t = G_p(z)$ so that $z = G^{-1}(t) = p(t)$ and $Q_p(t) = 1 - P(t)$. Then

$$\frac{tp(t)}{1-p(t)} = \frac{zG_p(z)}{Q_p(z)}$$

so that (7) is just (c.2). Therefore Lemma 2 follows from Lemma 2 of § 2. □

To conclude this section let us introduce the empirical analogues of the G- and Q-functions. Consider density estimators

$$\hat{p}_n(t) = \sum_{i=1}^{N} \frac{\nu_{in}}{n} I\{i-1 \leqslant t < i\} \tag{8}$$

$$\hat{f}_n(t) = \sum_{i=1}^{N} \nu_{in} I\{\frac{i-1}{n} \leqslant t < \frac{i}{n}\}\ .$$

Clearly, $\hat{p}_n$ and $\hat{f}_n$ are simply histograms. Consider now

$$G_{\hat{f}_n}(z) = \frac{1}{n} \sum_{i=1}^{N} I\{\nu_{in} > z\} \tag{9}$$

$$Q_{\hat{f}_n}(z) = \frac{1}{n} \sum_{i=1}^{N} \nu_{in} I\{\nu_{in} \leqslant z\}\ .$$

These two functions could be considered as the natural empirical analogues of G_f and Q_f and, in fact, they have been considered in the literature for a long time. In particular, $G_{\hat{f}_n}^{-1}$ is the so called "empirical rank distribution", while the jumps $n\Delta G_{\hat{f}_n}(m)$ at integer points m are just the spectral statistics $\mu_n(m)$, and $nG_{\hat{f}_n}(0+)$ is the vocabulary μ_n.

The study of the asymptotic behaviour of $G_{\hat{f}_n}$ and $Q_{\hat{f}_n}$ is, no doubt, one of the main problems in a statistical analysis of LNRE. That is why it is a little surprising how many papers are connected with the wrong conception of these functions. It is frequently supposed that $G_{\hat{f}_n}$ consistently estimates G_{f_n}

as $n \to \infty$, that is the difference $G_{\hat{f}_n} - G_{f_n}$ is small if n is large. But this is a mistake. In § 5 we consider this question in more detail.

In the next section, § 4, we turn to another description of LNRE.

§ 4. Inconsistent density estimates

Consider whether or not $\hat{p}_n$ and $\hat{f}_n$ are consistent estimators for p_n and f_n respectively. Namely, consider the L_1-distance

$$\|\hat{p}_n - p_n\| = \|\hat{f}_n - f_n\| = \sum_{i=1}^{N} \left| \frac{\nu_{in}}{n} - p_{in} \right| .$$

Condition 4.

$$\liminf_{n \to \infty} \mathrm{E}\, \|\hat{p}_n - p_n\| > 0 . \tag{c.4}$$

Lemma 1. (*c.* 1)⇔(*c.* 4). *Consequently (c.4) ⇔ (d.1).*

This lemma means that to consider LNRE in the sense of Definition 1 is the same as to deal with inconsistent histograms.

Proof.

$$\mathrm{E}\sum_{i=1}^{N} \left| \frac{\nu_{in}}{n} - p_{in} \right| \leqslant \sum_{i=1}^{N} [\mathrm{E}(\frac{\nu_{in}}{n} - p_{in})^2]^{1/2} I\{p_{in} > \frac{z}{n}\} + 2\sum_{i=1}^{N} p_{in} I\{p_{in} \leqslant \frac{z}{n}\}$$
$$\leqslant [\sum_{i=1}^{N} \mathrm{E}(\frac{\nu_{in}}{n} - p_{in})^2]^{1/2} G_n^{1/2}(\frac{z}{n}) + 2Q_n(\frac{z}{n}) .$$

From this and from (G.1) we have

$$\mathrm{E}\|\hat{p}_n - p_n\| \leqslant G_{p_n}^{1/2}(z) + 2Q_{p_n}(z) \leqslant z^{-1/2} + 2Q_{p_n}(z)$$

which proves the implication (c.4) ⇒ (c.1). On the other hand

$$\mathrm{E}\|\hat{p}_n - p_n\| \geqslant \sum_{i=1}^{N} \mathrm{E} \left| \frac{\nu_{in}}{n} - p_{in} \right| I\{p_{in} \leqslant \frac{z}{n}\} \geqslant$$
$$\geqslant \sum_{i=1}^{N} p_{in}(1 - p_{in})^n I\{p_{in} \leqslant \frac{z}{n}\} \geqslant (1 - \frac{z}{n})^n Q_{p_n}(z)$$

which proves the implication (c.1) ⇒ (c.4). The equivalence of (c.4) and (d.1) now follows from Lemma 1 of § 2. □

The equivalence (c.1) ⇔ (c.4) is proved by Abou-Jaoude (1976).

According to this lemma it is rather unreasonable to use the vector of relative frequencies $\frac{1}{n}\boldsymbol{\nu}_n$ as an estimator of a vector of probabilities $\mathbf{p}_n$ in the case of LNRE in the sense of (d.1). But it makes it even more worthwile to remark that the statistics $\|\hat{p}_n - p_n\|$ could be succesfully used for testing a hypothesis about p_n even against contiguous alternatives. We consider this fact in more detail in § 6.

Consider now $\hat{f}_n$ - a histogram derived according to Case 1 of § 3. It should be intuitively clear that $\hat{f}_n$ does not consistently estimate f_n because the subdivision of [0,1] by points $\{i/n\}$ is too fine.*) If we compare the histogram $\hat{f}_n$ with a kernel estimator

*) An extremely large real data set, essentially connected, in my view, with inconsistent histogramms is reported in (Udalcova, Colombet & Shnoll (1987)).

$$\tilde{f}_n(t)=\int \frac{1}{h}K(\frac{t-s}{h})F_h(\mathrm{d}s),\quad F_n(s)=\frac{1}{n}\sum_I\{X_i\leqslant s\}\ ,$$

where K is, say, a finite density, it becomes clear that a "too fine partition" corresponds to "too small h", i.e. the third of the conditions

$$n\rightarrow\infty,\quad h\rightarrow 0,\quad nh\rightarrow\infty$$

is not satisfied and hence $\tilde{f}_n$ does not converge to f. "Too small h" means that the kernel $(1/h)K((t-s)/h)$ is "too close" to the δ-function. The precise investigation of this phenomenon leads to conditions which are necessary and sufficient for the convergence

$$\mathrm{E}\|\bar{f}_n-f\|\rightarrow 0,\ n\rightarrow\infty$$

for density estimators of quite general form:

$$\bar{f}_n(t)=\int K_n(t,s)F_n(\mathrm{d}s)$$

with a general δ-sequence of kernels $\{K_n(\cdot,\cdot)\}$. These conditions are given, e.g., in Mnacakanov & Khmaladze (1981). Mnacakanov (1984) considered the convergence in distribution of Parzen-Rosenblatt-type inconsistent estimators $\tilde{f}_n$. One of his results is this: all one-dimensional limiting distributions of $\tilde{f}_n$ belong to the same continuous convolution semigroup of distributions $\{G_\lambda,\ \lambda\geqslant 0\}$ (cf. ch. IX in Feller (1971)), this semigroup is determined by the function K only and does not depend on the estimated density f. The value $f(t)$ determines which G_λ is the limiting distribution of $\tilde{f}_n(t)$, namely the value of λ must be equal to $cf(t)$ (where the positive constant c is unimportant and could be replaced by 1 under a suitable normalization). For instance if $K(t)=I\{0\leqslant t\leqslant 1\}$ then $\{G_\lambda,\ \lambda\geqslant 0\}$ is the semigroup of Poisson distributions, no matter what the estimated density is.

§ 5. Convergence of $G_{\hat{f}_n}$

In spite of the fact that the histograms $\hat{f}_n$ are not consistent estimates of anything, the G-and Q-functions of these histograms usually converge to limits.

The following lemma exploits a slightly more strong condition then (c.1).

Lemma 1. *If $Q_{f_n}\xrightarrow{\mathfrak{W}} Q_f$, then as $n\rightarrow\infty$*

$$\sup_{z>0}\ |G_{\hat{f}_n}(z)-C(z)|\xrightarrow{\mathrm{P}}0$$

where the limiting function C is

$$C(z)=\int_0^\infty \Pi^+(z,x)\frac{1}{x}Q_f(\mathrm{d}x)=-\int_0^\infty \Pi^+(z,x)G_f(\mathrm{d}x) \tag{1}$$

and

$$\Pi^+(z,x)=\sum_{k>z}\frac{x^k}{k!}e^{-x}\ .$$

Proof. Both $G_{\hat{f}_n}$ and C are nonincreasing step functions with jumps only at integer values of z, and both functions can be bounded by $1/z$ for large z (see (G.1) and verify that C satisfies (G.4)). Therefore it is sufficient to prove that for each $z>0$

$$G_{\hat{f}_n}(z)-C(z)\xrightarrow{\mathrm{P}}0\ . \tag{2}$$

But the convergence

$$G_{\hat{f}_n}(z) - \mathrm{E}G_{\hat{f}_n}(z) \xrightarrow{\mathrm{P}} 0 \tag{3}$$

is an immediate consequence of the limit result

$$\sqrt{n}\,[G_{\hat{f}_n}(z) - \mathrm{E}G_{\hat{f}_n}(z)] \xrightarrow{\mathfrak{D}} \mathfrak{N}(0,\sigma^2),\ \sigma^2 < \infty$$

which is a very particular case of a limit theorem for separable statistics (cf. § 6). Consider if

$$\mathrm{E}G_{\hat{f}_n}(z) - C(z) \to 0 \tag{4}$$

is true. Denote

$$B^+(z,\ p,\ n) = \sum_{k>z} \binom{n}{k} p^k (1-p)^{n-k}\ .$$

$$\mathrm{E}G_{\hat{f}_n}(z) = \int_0^\infty B^+\!\left(z,\ \frac{x}{n},\ n\right) \frac{1}{x} Q_{f_n}(\mathrm{d}x)\ . \tag{5}$$

But the integrand on the right-hand side is a bounded continuous function in x and it converges to the continuous bounded function $\Pi^+(z,x)/x$. This implies that (4) is true, hence (2) is true. □

Let us remark that (3) is true without any assumptions on the behaviour of f_n. We need the condition of the lemma only to prove (4).

From Lemma 1 it clearly follows that if according to the observed data $G_{\hat{f}_n}$ shows some regularity, e.g. if it might be supposed that

$$\frac{\mathrm{E}G_{\hat{f}_n}(z)}{\mathrm{E}G_{\hat{f}_n}(0+)} = \frac{1}{z}\ ,\quad \text{i.e.} \frac{\mathrm{E}\mu_n(z)}{\mathrm{E}\mu_n} = \frac{1}{z(z+1)} \tag{6}$$

(Zipf's law — cf. Tabel 4 of § 1), it does not mean at all that p_n or f_n follow the same regularity, that is, that

$$\frac{G_{f_n}(z)}{G_{f_n}(0+)} \approx \frac{1}{z}\ .$$

Probably the first paper which discusses this phenomenon was the paper of ORLOV & CHITASHVILI (1983).

For the purpose of estimation of G_{f_n} it is natural to consider the integral equation

$$G_{\hat{f}_n}(z) = \int_0^\infty \Pi^+(z,x)G(\mathrm{d}x)\ . \tag{7}$$

The problem of finding the solution of this equation could be the theme of a separate discussion. But we avoid it in the present paper.

§ 6. SEPARABLE STATISTICS. EXAMPLES — χ^2STATISTIC, MULTINOMIAL MAXIMUM LIKELIHOOD STATISTIC ETC. LIMIT THEOREMS

A very interesting class of statistics connected with the statistical analysis of LNRE is made up of so called additively separable or, simply, separable statistics*) of the form

*) The term "separable statistics" seems to me more adequate then the term "divisible statistics" used by the translator of the paper (KHMALADZE 1983) in English edition of *Theorija Verojatn. i Primen.*

$$\sum_{i=1}^{N} g(\nu_{in}, np_n) \tag{1}$$

(the term was introduced by MEDVEDEV (1970)). The spectral statistics $\mu_n(m)$ are one example with $g(\nu)=I\{\nu=m\}$, and $G_{\hat{f}_n}(z)$ and $Q_{\hat{f}_n}(z)$ are other examples with $g(\nu)=I\{\nu>z\}$ and $g(\nu)=\nu I\{\nu\leqslant z\}$ respectively. Further examples are the χ^2statistics

$$X_{n,N}^2=\sum_{i=1}^{N}\frac{(\nu_{in}-np_{in})^2}{np_{in}}$$

and the maximum likelihood statistics for multinomial distributions

$$L_{n,N}=\sum_{i=1}^{N}\nu_{in}\ln\frac{\nu_{in}}{np_{in}} .$$

The L_1- distance between $\hat{f}_n$ and f_n is also a separable statistic

$$\|\hat{f}_n-f_n\|_1=\sum_{i=1}^{N}\left|\frac{\nu_{in}}{n}-p_{in}\right|$$

and, obviously, linear functionals of $G_{\hat{f}_n}$ are separable statistics

$$\int_0^\infty g(z)G_{\hat{f}_n}(\mathrm{d}z)=\frac{1}{n}\sum_{i=1}^{N}g(\nu_{in}) .$$

The bibliography on the asymptotic theory of separable statistics is very large and a good review paper by IVANOV, IVCHENKO & MEDVEDEV (1984) is available, see also IVCHENKO & MEDVEDEV (1980). Our aim in this § 6 is only
(a) to illustrate the specific behaviour of statistics (1) in the case of LNRE;
(b) to discuss the not quite traditional centering for the statistics (1).

(a). The classical result about convergence in distribution of the χ^2statistic $\chi_{n,N}^2$ is this: for any fixed **p**

$$X_{n,N}^2\xrightarrow{\mathfrak{D}}\chi_{N-1}^2 \quad \text{as } n\to\infty ,$$

where χ_{N-1}^2 denotes a χ^2-distributed random variable with $N-1$ degrees of freedom. If then $N\to\infty$, we get

$$\frac{\chi_{N-1}^2-(N-1)}{\sqrt{2(N-1)}}\xrightarrow{\mathfrak{D}}\mathcal{N}(0,1) .$$

Therefore

$$\frac{X_{n,N}^2-(N-1)}{\sqrt{2(N-1)}}\xrightarrow{\mathfrak{D}}\mathcal{N}(0,1) \text{ if } n\to\infty \text{ and then } N\to\infty .$$

But if we consider the simultaneous limit as we should in the case of LNRE the result will be entirely different:

$$\frac{X_{n,N}^2-(N-1)}{\sqrt{2(N-1)}}\xrightarrow{\mathfrak{D}}\mathcal{N}(0,\sigma^2),\text{ if } n\to\infty,\ \frac{N}{n}\to 1$$

where the variance $\sigma^2>1$. More precisely,

$$\mathrm{E}\left[\frac{\chi_{N-1}^2-(N-1)}{\sqrt{2(N-1)}}\right]^2=2\left[\frac{N-1}{N}+\frac{1-N-N^2/2}{nN}+\frac{1}{2nN}\sum_{i=1}^{N}\frac{1}{p_{in}}\right]$$

and therefore, if $n \to \infty$ and $N/n \to 1$ in the Case 1 of § 3 we get

$$\sigma^2 = \begin{cases} 1+\int\limits_0^1 \frac{1}{f(t)}\mathrm{d}t, & \text{if } \int\limits_0^1 \frac{1}{f(t)}\mathrm{d}t < \infty \\ \infty, & \text{if } \int\limits_0^1 \frac{1}{f(t)}\mathrm{d}t = \infty \end{cases} \tag{2}$$

Therefore in the case of LNRE the statistic $X^2_{n,N}$ loses its main property — it is no longer asymptotically distribution free. The same is true for the statistics $L_{n,N}$.

Remark. One faces a similar and more unpleasant situation in the two-sample case: if ν_n and ν_n' are two vectors of frequencies based on two independent samples, say, of the same size n, and if $\mathcal{Y}^2_{n,N}$ is two sample χ^2statistic

$$\mathcal{Y}^2_{n,N} = 2\sum_{i=1}^{N} \frac{(\nu_{in}-\nu_{in}')^2}{\nu_{in}+\nu_{in}'} ,$$

then the limit distribution of $\mathcal{Y}^2_{n,N}$ as $n \to \infty$, $N/n \to \alpha > 0$ under the null hypothesis depends on the unknown underlying distribution. Recently Urinov (1988) suggests some asymptotically distribution free tests for this case. His approach shares the idea of conditioning used long ago by Bickel (1969) and the martingale approach proposed by Khmaladze (1983).

This example illustrates that limit theorems for statistics (1) cannot be derived by means of classical asymptotics when $n \to \infty$ (and then $N \to \infty$).

In connection with this discussion it might be appropriate to consider just another situation when we have to deal with LNRE.

Case 3. Let $\mathrm{X}_1, \ldots, \mathrm{X}_n$ be i.i.d. random vectors taking values in m-dimensional space $\mathbb{R}^m$. Consider the problem of testing a hypothesis about the distribution of these random vectors based on grouped data. If we divide the range of each single coordinate of the vector X_i into only 2 or 3 cells we get 2^m or 3^m different subsets in $\mathbb{R}^m$ and hence $N=2^m$ or $N=3^m$ different frequencies.

But for $m=10$ we get $N=2^{10}=1024$ or $N=3^{10}=59.049$. It is clear that for a large number of real statistical problems it is hard to expect the sample size n to be much greater then N, and one should consider asymptotics as $n/N \to \alpha < \infty$ as $n \to \infty$, or even $n/N \to 0$ as $n \to \infty$. (This problem was pointed out in Autumn 1985 by prof. Yu. Prohorov in the seminar of Mathematical statistics at the V.A. Steklov Mathematical Institute. Functional limit theorems for the asymptotics $n/N \to 0$ were studied in the one-dimensional case by Mnacakanov (1985, 1987), for earlier references see these papers).

(b). In Khmaladze (1983) instead of fixed sums (1) partial sums are considered:

$$X_{n,N}(t) = \frac{1}{\sqrt{n}} \sum_{i \le Nt} [g(\nu_{in}, \frac{i}{N}) - \mathrm{E}g(\nu_{in}, \frac{i}{N})] ,$$

and it was suggested to adjoin to the process $X_{n,N}$ the filtration $\{\mathcal{F}_i^n\}$, $i=1,\ldots,N$, where the σ-algebra $\mathcal{F}_i^n$ is generated by the first i frequencies:

$$\mathcal{F}_i^n = \sigma\{\nu_{1n}, \ldots, \nu_{in}\} .$$

The evolution of the semimartingale $\{X_{n,N}(\frac{i}{N}), \mathcal{F}_i^n\}$ as a process in $t=\frac{i}{n}$ is determined by the conditional distribution of the frequency ν_{in} given $\mathcal{F}_{i-1}^n$. But this distribution is very simple — a binomial one:

$$\mathrm{P}\{\nu_{in}=k\,|\,\mathcal{F}^n_{i-1}\}=\mathcal{B}(k,\ \frac{\Delta F(\frac{i}{N})}{1-F(\frac{i}{N})},\ n-\sum_{j=1}^{i-1}\nu_{jn})\ , \qquad (3)$$

where $\mathcal{B}(k,\ p,\ m)$ stands for the binomial probability of k with parameter p and number of trials m. Therefore it is convenient to study the symptotic behaviour of $X_{n,N}$ by means of martingale limit theorems - this was the starting point of the paper KHMALADZE (1983).

But let us compare here the process $\{X_{n,N}(\frac{i}{N}),\ \mathcal{F}^n_i\}$ with its martingale component only $\{W_{n,N}(\frac{i}{N}),\ \mathcal{F}^n_i\}$:

$$W_{n,N}(T)=\frac{1}{\sqrt{n}}\sum_{i\leqslant Nt}(g(\nu_{in},\ \frac{i}{N})-\mathrm{E}[g(\nu_{in},\ \frac{i}{N})\ |\mathcal{F}^n_{i-1}])\ , \qquad (4)$$

so that

$$X_{n,N}=W_{n,N}+K_{n,N}$$

with $K_{n,N}$ being a compensator of the $X_{n,N}$. Recall that the marginal distribution of ν_{in} is also binomial:

$$P\{\nu_{in}=k\}=\mathcal{B}(k,\ \Delta F(\frac{i}{N}),\ n)\ . \qquad (5)$$

Now:

— the comparison of (3) and (5) shows that the actual calculation of the trajectories of $W_{n,N}$ is not much more difficult then that of $X_{n,N}$;

— the limit theorem for $W_{n,n}$ is an easy consequence of the CLT for martingales, while a limit theorem for the compensator $K_{n,N}$, and hence for $X_{n,N}$ itself required rather more effort;

— the limit process of $W_{n,N}$ is simply a Wiener process, and it is relatively easy to calculate the limit distribution of statistics based on $W_{n,N}$. For example, the limit distribution of the statistics $\sup W_{n,N}(t)$ is simply $2\Phi(\cdot/\sigma)-1$, where σ^2 is the limit of $\mathrm{E}W^2_{n,N}(1)$ as $n,\ N\to\infty$ and $\Phi(\cdot)$ denotes a normal distribution function with expectation 0 and variance 1, while for the limit distribution of statistics $\sup\limits_{0\leqslant t\leqslant 1} X_{n,N}(t)$ which is a distribution of an upper bound $\sup\limits_{0\leqslant t\leqslant 1} X(t)$ of a gaussian process X one cannot in general get something better then a difficult to calculate approximation for high percentage points;

— one more remark, but in favour of the process $X_{n,N}$ this time, is this — the expectation

$$\frac{1}{n}\sum_{i\leqslant Nt}\mathrm{E}g(\nu_{in},\ \frac{i}{N})$$

or its limit as $n,\ N\to\infty$ might be itself a function of interest, hence, $X_{n,N}$ is a difference between estimator and estimated function and so consideration of $X_{n,N}$ is natural and even necessary.

§ 7. EMPIRICAL BAYES APPROACH AND STATISTICAL PROBLEMS WITH LARGE NUMBER OF PARAMETERS: CONNECTION WITH LNRE

In the author's opinion the title of the present section promises more then its content is able to deliver: here only some remarks are gathered that help to recognise the same mathematical contents in entirely different heuristic contexts. But no advanced study is presented, e.g. no mathematical statement is formulated.

The description of a typical problem of the *empirical Bayes approach* has little in common with the analysis of frequencies of words in a literary texts. This typical problem can be formulated as follows: suppose in each i'th batch of some manufactured items ν_i items are found to be defective and that these ν_i's have Poisson distributions with intensities θ_i:

$$P\{\nu_i > z \mid \theta_i\} = \Pi^+(z, \theta_i) \tag{1}$$

(cf. Lemma 1 of § 5). The intensity θ_i characterizes the quality of the i'th batch and may vary from batch to batch. Let us describe it as a random variable with prior distribution Λ. This distribution characterises the given technology and the given manufacturer and is the subject of interest. One has to estimate it from the data, that is, from the sequence $\{\nu_i\}$ $i=1, \ldots, n$, of numbers of defective items in a sequence of n independent batches. Since the conditional distribution of ν_i given θ_i is defined by (1), the unconditional distribution of ν_i is

$$F^+(z) = P\{\nu_i > z\} = \int \Pi^+(z, \theta)\Lambda(d\theta). \tag{2}$$

Now, if n is large F^+ can be estimated by the empirical distribution function of $\{\nu_i\}$:

$$F_n^+(z) = \frac{1}{n}\sum_{i=1}^{n} I\{\nu_i > z\},$$

and, hence, the solution Λ_n of the equation

$$F_n^+(z) = \int \Pi^+(z, \theta)\Lambda_n(d\theta) \tag{3}$$

can be considered as an estimator for the prior distribution Λ. This is the central proposal in the empirical Bayes approach.

But equations (2) and (3) differ from equations (1) and (7) of § 5 only in notations. It is quite clear that the present Λ and F_n^+ are essentially the same as G_f and $G_{\hat{f}_n}$ (see (1) and (9) of § 3), and therefore the mathematical setting of the empirical Bayes approach and of the statistical analysis of LNRE is, in fact, very similar.

Sometimes one supposes that the conditional distribution of ν_i given the batch is not Poisson but hypergeometric or binomial. This will change the Kernel Π^+ in (2) and (3), but this is not essential for the similarity just mentioned.

Consider now some typical statistical *problems with a large number of parameters*. Suppose, e.g., $X_1, \ldots, X_n$ are *i.i.d.* N-dimensional random vectors and let R be the covariance matrix of each X_i. Denote $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_N)$ the set of eigenvalues of the matrix R and

$$\mu_N(x) = \frac{1}{N}\sum_{i=1}^{N} I\{\theta_i \leqslant x\}.$$

The spectral function μ_N is frequently considered as the function of the main interest in inference problems concerning R. For example, functions like det R and trace R, which describe the dispersion of X_i, can be easily expressed through μ_N:

$$(\det R)^{1/N} = \left[\prod_{i=1}^{N} \theta_i\right]^{1/N} = \exp\{\int \log x \;\; \mu_n(dx)\},$$

$$\frac{1}{N} \operatorname{trace} R = \frac{1}{N}\sum_{i=1}^{N} \theta_i = \int x \;\; \mu_n(dx).$$

Denote by $\hat{R}_n$ the sample covariance matrix

$$\hat{R}_n = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^T, \quad \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and consider

$$\hat{\mu}_{n,N}(x) = \frac{1}{N}\sum_{i=1}^{N} I\{\hat{\theta}_{in} \leqslant x\}$$

where $\{\hat{\theta}_{in}\}$ denotes the eigenvalues of $\hat{R}_n$. The main starting point of many researches on random matrices (the most recent papers the author knows about are (GIRKO, (1987) and

(BAI, YIN & KRISHNAIAH, 1987)) is the fact that if N increases with the same order as n then under proper conditions the difference $\hat{\mu}_{n,N}-\mathrm{E}\hat{\mu}_{n,N}$ converges to 0 (in probability), but the difference $\mathrm{E}\hat{\mu}_{n,N}-\mu_N$ does not. That is, $\hat{\mu}_{n,N}$ consistently estimates its expectation but not μ_N.

Examples of statistical problems with random matrices of increasing order can be found, e.g., in (GIRKO, 1980). Problems with increasing number of parameters can frequently be found in the context of medical diagnosis. One particular situation can be described as follows (see, e.g., (KHMALADZE & NEFEDOV, 1977): suppose for each of n patients N different symptoms are measured and X_{ij} is a result of the measurement of the j'th symptom on the i'th patient. Suppose for simplicity that $X_{i1},\ldots,X_{iN}$ are independent Bernoulli random variables and let $p_j=\mathrm{P}\{X_{ij}=1\}$. Very often there is no reason to assume p_j equal for different j and, therefore, the outcome $\mathbf{X}_i=(X_{i1},\ldots,X_{iN})$ of treatment of the i'th patient has distribution

$$\mathrm{P}\{\mathbf{X}_i=x\}=\prod_{j=1}^{N}p_j^{x_j}(1-p_j)^{1-x_j},\ \mathrm{x}=(x_1,\ldots,x_n),\ x_j=0 \text{ or } 1\,.$$

Hence if we have many intensively studied patients we get a problem with a large sample size n and a large number of parameters N.

The problems with a large number of parameters in the context of regression analysis was considered recently in (TORONDJADZE & MDZINARISHVILI, 1988).

But if we look back at the LNRE-sequences it becomes obvious, that the quantities $np_{1n},\ldots,np_{Nn}$ are just these unknown parameters in a large amount, and the functions G_{f_n} and $G_{\hat{f}_n}$ defined by (2) and (9) of § 3 are quite similar to the functions $\mu_{n,N}$ and $\hat{\mu}_{n,N}$ of the present section. Not exaggerating the similarity too much one can still say that the principal way used in problems with an increasing number of parameters (as studied, e.g., in (GIRKO, 1987)) has a natural analogy with the theory of separable statistics.

Namely, let $\hat{\boldsymbol{\theta}}_n$ denote an estimator of the vector of parameters $\boldsymbol{\theta}$ and let $f(\boldsymbol{\theta})$ be some function of $\boldsymbol{\theta}$. Then

(1) under some conditions

$$f(\hat{\boldsymbol{\theta}}_n)-\mathrm{E}f(\hat{\boldsymbol{\theta}}_n)\overset{\mathrm{P}}{\rightarrow}0,\quad n,\quad N\rightarrow\infty$$

but usually

$$\mathrm{E}f(\hat{\boldsymbol{\theta}}_n)-f(\boldsymbol{\theta})\not\rightarrow 0,\quad n,\quad N\rightarrow\infty$$

(GIRKO, 1987). Similarly for separable statistics it is well known that

$$\frac{1}{N}\sum_{i=1}^{N}[g(\nu_{in})-\mathrm{E}g(\nu_{in})]\overset{\mathrm{P}}{\rightarrow}0,\quad n,\quad N\rightarrow\infty$$

(see § 6), but usually $\sum\limits_{i=1}^{N}\mathrm{E}g(\nu_{in})/N$ and $\sum\limits_{i=1}^{N}g(np_{in})/N$ have different limits (cf. Lemma 1 of § 5):

$$\frac{1}{N}\sum_{i=1}^{N}\mathrm{E}g(\nu_{in})\rightarrow-\int g(z)C(\mathrm{d}z)\,,$$

while

$$\frac{1}{N}\sum_{i=1}^{N}g(np_{in})=-\int g(z)G_{f_n}(\mathrm{d}z)\rightarrow-\int g(z)G_f(\mathrm{d}z)\,,$$

as n, $N\rightarrow\infty$.

(2) If the distribution of $\boldsymbol{\xi}=\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}$ is $\mathfrak{N}(\mathbf{0},\ R)$ then the function

$$u(\boldsymbol{\theta},t)=\mathrm{E}f(\boldsymbol{\theta}+tR^{\frac{1}{2}}\boldsymbol{\xi})$$

is a solution of the equation

$$\frac{\partial u(\boldsymbol{\theta},t)}{\partial t}=\frac{1}{2}\sum_{i,j=1}^{N}R_{i,j}\frac{\partial^2}{\partial\theta_i\partial\theta_j}u(\boldsymbol{\theta},t)\,,\quad R=(R_{ij})\,,$$

and the initial condition $u(\boldsymbol{\theta},0)=f(\boldsymbol{\theta})$ is the quantity of interest. Now, if one knew $u(\boldsymbol{\theta},1)$ one could calculate $u(\boldsymbol{\theta},0)$ by some inverse procedure. But $u(\boldsymbol{\theta},1)$ can be estimated by $f(\hat{\boldsymbol{\theta}}_n)$, which leads to the estimation of $u(\boldsymbol{\theta},0)$ (Girko, 1987). But similarly, we already have mentioned in § 5 that $G_{\hat{f}_n}$ can be used as an estimation of C and then G_f can be estimated as an inverse of (7) of § 5.

Some readers will find this § 7 too general and undeveloped. The author agrees with them. The author would refrain from adding it to the written text of the lecture if he knew a single cross-reference between these three circles of investigations — empirical Bayes approach, problems with increasing number of parameters and analysis of LNRE.

References

Abou-Jaoude, S., (1976). Conditions necessaries et suffisantes de convergence L_1 en probabilite de l'histogramme pour une densité. *Ann. Inst. H. Poincare, B,* **12,** 213-216.

Bai, Z.D., Yin, Y.Q., Krishnaiah, P.R., (1987). On limiting empirical distribution function of the eigenvalues of a multivariate F matrix. *Teorya Verojatn. i Primen,* **32,** 538-548.

Bickel, P., (1969). A distribution free version of the Smirnov two sample tests in p-variate case. *Ann. Math. Statist.,* **40,** 1-23.

Feller, W., (1971). *An introduction to probability theory and its applications, Vol. II,* Wiley, New-York.

Gill, R.D., (1983). *First solution of the coin problem,* (in Dutch). Unpublished report, CWI, Amsterdam.

Gill, R.D., (1984). *Comparison of methods for determining the number of dies,* (in Dutch). Unpublished report, CWI, Amsterdam.

Girko, V.L., (1980). *Theory of random determinants.* Kiev State Univ., Kiev.

Girko, V.L., (1987). Introduction to general statistical analysis. *Teorya Verojatn. i Primen,* **32,** 252-265.

Ivanov, V.A., Ivchenko, G.I., Medvedev, Yu.I., (1984). Discrete problems of probability theory (in Russian). *Itogi Nauki i Techn.,* Series *Theory Probab., Mathemat. Statist., Teor. Cybernetics,* **22,** VINITI, Moskow.

Ivchenko, G.I., Medvedev, Yu.I., (1980). Separable statistics and hypothesis testing for grouped data. *Theory Probab. Appl.* **25,** 540-551.

Khmaladze, E.V., (1983). Martingale limit theorems for divisible statistics. *Theory Probab. Appl.* **28,** 530-548.

Khmaladze, E.V., Nefedov, F., (1977). On errors related to the use of expert estimates. *Internat. J. Bio-Medical Computing,* **8,** 11-20.

Medvedev, Yu.I., (1970). Some theorems on asymptotic distribution of statistic χ^2. *Soviet Math. Dokl.* (Doklady Acad. Nauk), **192,** 87-989.

Mnacakanov, R.M., (1984). *Limit theorems for statistics connected with large number of rare events.* Ph.D. Thesis., Tbilisi.

Mnačakanov, R.M., (1985). Functional limit theorems for additively-separable statistics in the case of very rare events. *Theory Probab. Appl.,* **30,** 584-588.

Mnacakanov, R.M., (1987). On convergence of separable statistics to Wiener process. *Teorya Verojatn. i Primen,* **32,** 392-396.

Mnacakanov, R.M. & Khmaladze E.V., (1981). *On L_1-convergence of statistical kernel estimators of distribution densities. Soviet Math. Dokl. (Doklady Acad. Nauk)* **23,** 633-636.

ORLOV, YU., CHITASHVILI, J., (1983). On the statistical interpretation of Zipf's law. *Bulletin Acad. Sc. Georgian SSR,* **109,** 505-508.

STAM, A.J., (1987). Statistical problems in ancent numismatics, *Statistica Neerlandica,* **41,** 151-171.

TORONDJADZE, A.F., MDZINARISHVILI, P.G. (1982). Constrained maximal likelihood method and some of its application in astronomy. *Soviet-Japan Symposium on Probab. Theory and Math. Statistics.* Abstract of comm. 2. Meznierela, Tbilisi.

UDALCOVA, N.V., COLOMBET, V.A. SHNOLL, S.E., (1987). *On possible cosmophysical conditionality of macroscopie fluctuations of processes of various natures,* (in Russian). Nauchn. Center Biolog. Issledovanii (Center of Biolog. Studies, Acad. Sc. USSR, Pushchino).

URINOV, I.K., (1988). Martingale limit theorems for tests of homogeneity of two multinomial populations. *Teorya Verojatn. i Primen,* **33,** (to appear).