Order Statistics, Quantile Processes and Extreme Value Theory

Oberwolfach Meeting from 25th to 31st March 1984

by R. Helmers

This meeting, organized by R.D. Reiss (Siegen) and W.R. van Zwet (Leiden), brought together 43 researchers working in three different, though closely related, fields in probability and statistics: order statistics, quantile processes and extreme value theory. In this report I shall survey a number of recent developments that were discussed at the conference. Specifically I shall deal with the following subjects: spacings theory, empirical and quantile processes, estimation of the tail of a distribution, extreme value theory, and a miscellaneous category.

Spacings theory

The talks in this section dealt with functions of uniform k-spacings. Statistics of this type play an important role in a number of different contexts, such as tests for uniformity, Poisson processes and non parametric density estimation.

Consider a sequence $U_1, U_2, \ldots$ of independent random variables (observations), each of them distributed according to the uniform distribution on [0,1], and let, for each $n \geq 1$, $U_{1:n} \leq \cdots \leq U_{n:n}$ denote the first n U_i's ordered in ascending order of magnitude. Set $U_{0:n} = 0$ and $U_{n+1:n} = 1$. *Uniform spacings* are defined by $D_{in} = U_{i+1:n} - U_{i:n}$ $(0 \leq i \leq n)$, the gaps induced by the 'random points' $U_1, \ldots, U_n$ in the interval [0,1], and, more generally, *uniform k-spacings* by $D_{ink} = U_{i+k:n} - U_{i:n}$ $(0 \leq i \leq n+1-k)$. Let $M_n = \max_{0 \leq i \leq n} D_{in}$, the maximal spacing or maximal gap, and $M_{nk} = \max_{0 \leq i \leq n+1-k} D_{ink}$ the maximal k-spacing. Of course $M_{n1} = M_n$.

As early as in 1939 P. Levy found the limit distribution of M_n: $\lim_{n \to \infty} P(nM_n \leq \log n + x) = \exp(-e^{-x})$ for all $x \in \mathbb{R}$. It is well-known that the distribution function $\exp(-e^{-x})$ has mean $\gamma = 0.5772 \cdots$, Euler's constant, and variance $\frac{\pi^2}{6}$. In Devroye (1981) it was noted that indeed the expected value $E(nM_n - \log n) \to \gamma$ and the variance $\sigma^2(nM_n) \to \frac{\pi^2}{6}$, as $n \to \infty$, as one may expect from Levy's result. However, the question remains: how small or large can the maximal gap M_n be as n gets large? In Devroye (1982) it was proved that, with probability 1,

$$\liminf_{n\to\infty} (nM_n - \log n + \log\log\log n) = -\log 2$$

and (1)

$$\limsup_{n\to\infty} \frac{nM_n - \log n}{2\log\log n} = 1$$

At the conference Deheuvels (joint work with L. Devroye) proved similar strong limit laws ($n\to\infty$) for M_{nk} and related statistics, both when k is fixed and when it is allowed to increase with n at a rate not exceeding $\log n$. If k is *fixed,* then, with probability 1,

$$\liminf_{n\to\infty} \frac{nM_{nk} - \log n - (k-1)\log\log n}{\log\log\log n} = -1$$

and (2)

$$\limsup_{n\to\infty} \frac{nM_{nk} - \log n - (k-1)\log\log n}{2\log\log n} = 1$$

Note that (2) includes the case $k = 1$, for which (2) is implied by the more refined result (1). Clearly k affects only the second order terms in (2). This means that if the maximal spacing M_n is either 'small' or 'large', the maximal k-spacing M_{nk} is likely to be of the same order of magnitude. If, on other hand, $k = k(n) \to \infty$, *but* $k = o(\log n)$, then, with probability 1,

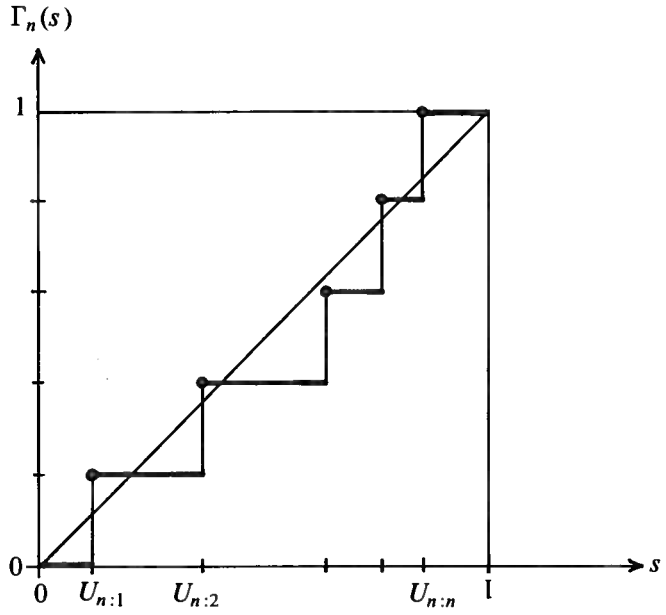
$$\frac{nM_{nk} - \log n}{(k-1)\log(\frac{e\log n}{k})} \to 1 \qquad (3)$$

Again, as in (2), nM_{nk} is of the order of $\log n$. A correction term $(k-1)\log(\frac{e\log n}{k})$ is also established. In reference [3] the deviation of $nM_{nk} - \log n$ from this correction term is studied. Also the case $k = c\log n$, for some constant $c > 0$, was considered in the talk as well as results for the non-uniform case.

Berry-Esseen bounds and Edgeworth expansions for statistics of the form $\Sigma_{i=0}^{n} g((n+1)D_{in})$ for a fixed function g, were derived by Does (joint work with R. Helmers and C.A.J. Klaassen). Klaassen proved a general limit theorem for conditional statistics, with an application to uniform k-spacings.

Empirical and quantile processes

In a three part talk M. Csörgö, S. Csörgö and D.M. Mason (joint work with L. Horvath), introduce a new Brownian Bridge approximation to the uniform empirical and quantile processes and discuss applications in probability and statistics. To explain their result we need a bit of notation. Let Γ_n denote



the empirical distribution function based on $U_1,...,U_n$; i.e. $\Gamma_n(s)$ is the proportion of the U_i's $(1 \leq i \leq n)$ which are $\leq s$ $(0 \leq s \leq 1)$. The *uniform empirical process* α_n is given by $\alpha_n(s) = n^{1/2}(\Gamma_n(s) - s)$, $0 \leq s \leq 1$. Also let Q_n be the uniform empirical quantile function; $Q_n(s) = U_{i:n}$ for $\frac{i-1}{n} < s \leq \frac{i}{n}$ $(1 \leq i \leq n)$ and $Q_n(0) = 0$. So Q_n is the left-continuous inverse of Γ_n. The *uniform quantile process* β_n is given by $\beta_n(s) = n^{1/2}(s - Q_n(s))$, $0 \leq s \leq 1$. Finally let $B(s)$, $0 \leq s \leq 1$, denote a *Brownian Bridge,* i.e. a real-valued, zero-mean Gaussian process with continuous sample paths and covariance function $EB(s)B(t) = \min(s,t) - st$, $0 \leq s,t \leq 1$.

A probability space is constructed with a sequence of independent uniform (0,1) random variables $U_1,U_2,...$ and a sequence of Brownian Bridges $B_1,B_2,...$ defined on it, such that for all $0 \leq \nu < 1/4$

$$\sup_{0 \leq s \leq 1} |\alpha_n(s) - \bar{B}_n(s)| / (s(1-s))^{1/2-\nu} = O_p(n^{-\nu}) \tag{4}$$

where $\bar{B}_n(s) = B_n(s)$ for $n^{-1} \leq s \leq 1 - n^{-1}$ and zero elsewhere; in addition one also has for all $0 \leq \nu < 1/2$

$$\sup_{1/n+1 \leq s \leq n/n+1} |\beta_n(s) - B_n(s)| / (s(1-s))^{1/2-\nu} = O_p(n^{-\nu}) \tag{5}$$

Here $O_p(n^{-\nu})$ has the standard meaning that n^{ν} times the l.h.s. of (4) and (5) remain bounded in probability as n gets large.

In a way these Brownian Bridge approximations improve upon the well-known 'KMT-embedding' for α_n, due to Komlos, Major and Tusnady and also known as the Hungarian embedding, and the parallel result for β_n; the

improvement is in the 'tails', i.e. in neighbourhoods of zero and one. The constructions (4) and (5) were successfully applied to certain asymptotic problems involving the tails of α_n and β_n. Examples of such problems, which were mentioned in the talks, include: a refined Chibisov-O-Reilly theorem for the weak convergence of *weighted* empirical and quantile processes, a new proof of the Jaeschke-Eicker limit theorems, and central limit theorems for sums of extreme order statistics.

In the talk of Révész a very delicate strong invariance principle was presented for the *local time* γ_n of α_n; here γ_n is the number of zero crossings of α_n. Révész proved that one has with probability 1,

$$|n^{-\frac{1}{2}}\gamma_n - \eta(n)| = 0(n^{-\frac{1}{4}+\epsilon}) \tag{6}$$

for any $\epsilon>0$, where the process $\eta(t)$, $t \geqslant 0$ denotes the (carefully defined) local time of a Kiefer process. As an application of (6) we have, with probability 1,

$$\limsup_{n\to\infty} \frac{\gamma_n}{\sqrt{n \log\log n}} = \frac{1}{\sqrt{2}} \tag{7}$$

Ruymgaart (joint work with J. Einmahl and J.A. Wellner) established a Chibisov O-Reilly result for the weak convergence of weighted empirical processes for the multidimensional case, both when the process is indexed by points (the classical case) and when it is indexed by rectangles. Steinebach proved an improved Erdös - Rényi strong law for moving quantiles.

The talks in this category discussed so far emphasise a probabilistic point of view. In contrast, the two part talk of Basset and Koenker contributes significantly to problems of statistical applications, e.g. in econometrics. Their idea is to estimate the error structure in linear models with the aid of an empirical regression quantile function. The behaviour of associated quantile processes is studied and applications discussed. Other talks in this category were by Boos (estimation of large quantiles) and by Falk (kernel type estimators of a population quantile).

Estimation of the tail of a distribution

Consider a distribution function F with regularly varying upper tail, i.e.

$$1 - F(x) = x^{-\lambda}L(x), \quad x > 0 \tag{8}$$

where $\lambda > 0$ denotes the tail index of F and L is slowly varying at infinity. The problem is to estimate λ on the basis of n independent observations $X_1,...,X_n$ from F. Hill [4] proposed the estimator

$$\lambda_n = (k^{-1} \Sigma_{i=1}^{k} \log X_{n-i+1:n} - \log X_{n-k:n})^{-1} \tag{9}$$

where $X_{1:n} \leqslant \cdots \leqslant X_{n:n}$ are the ordered X_i's. For simplicity we suppose $F(0) = 0$. The integer k $(1 \leqslant k \leqslant n)$ depends on n in such a way that

$$k = k(n) \to \infty, \quad k(n) = o(n) \tag{10}$$

Three talks were devoted to the problem of the asymptotic normality of λ_n. Haüsler showed that $\sqrt{(n)}\ (\lambda_n - \lambda)$ is asymptotically normally distributed, provided $k(n) \to \infty$ sufficiently slow. If some knowledge about L is available then the sequence $k = k(n)$ can be determined explicitly. Smith dealt with the same problem from a different point of view: $k = k(n)$ is now the random number of exceedances of a given threshold level x_n. Again to determine x_n information about L is required. S. Csörgö (joint work with D.M. Mason) showed that the Brownian Bridge approximation (5) can be applied to establish the asymptotic normality of λ_n.

While attending the meeting S. Csörgö, P. Deheuvels and D.M. Mason wrote paper [5]. In this paper these authors introduced a new class of estimators of λ, which can be viewed as the inverse of the convolution of a kernel function with the logarithm of the empirical quantile function of the X_i's. Asymptotic normality is established with the aid of (5) and optimal kernels are selected that give more weights to the extreme observations then Hill's λ_n.

Extreme value theory

Extreme value theory in concerned with the asymptotic behaviour of sample extremes. The central result in this area is of course Gnedenko's theorem stating that the limit distribution of the normalized maximum $X_{n:n}$ of independent and identically distributed random variables $X_1,...,X_n$, if it exists, must be one of three types: of the well-known extreme value distributions. Balkema (joint work with L. de Haan and S. Resnick) gave rate of convergence results for the case that the X_i's are in the domain of attraction of the extreme value distribution $\exp(-e^{-x})$.

It is known that under appropriate 'long range' and 'local' dependence conditions, classical extreme value theory remains true for dependent (stationary) sequences. Leadbetter discusses what happens if the 'local' dependence condition is relaxed. The answer turns out to be that the asymptotic distribution of the maximum is essentially unchanged, whereas the distributions of the other extreme order statistics are altered in a specific way. This phenomenon is caused by the fact that the point processes of high exceedances, which are classically Poisson, now involve clustering.

Other talks in this category were by de Haan (records from an improving population), Hüsler (extreme value theory for non-stationary sequences), Meizler (extreme value theory for independent but not identically distributed random variables), Lindgren (optimal prediction of upcrossing of a critical level by a stationary Gaussian process), Rootzén (behaviour of extremes of moving averages) and Tiago de Oliveira (extreme value theory for bivariate random variables).

Miscellaneous

Daniels discussed the joint distribution of the maximum of a random walk

and the time at which it is attained. Asymptotically the problem is related to one involving Brownian motion in presence of a parabolic boundary. This bears some resemblance with recent work at CWI by P. Groeneboom and N.M. Temme on the statistical problem of estimating a monotone density.

The other talks dealt with the saddle point method and M-estimators (Dinges), the asymptotic behaviour of the L_1-error of density estimators (Györfi), generalized L-statistics (Helmers; joint work with P. Janssen and R.J. Serfling), order statistics in the non i.i.d. case and the influence of outliers (David, Gather, Mathar), order statistics in insurance mathematics (Teugels), order statistics and symmetric functions (Rüschendorf), test for a change point model (Mittal), and a multivariate two sample test based on nearest neighbours (Henze).

References

1 DEVROYE, L. (1981). Laws of the iterated logarithm for order statistics of uniform spacings. *Ann. of Probability 9,* 860-876.

2 DEVROYE, L. (1982). A log log law for maximal uniform spacings. *Ann. of Probability 10,* 863-868.

3 DEHEUVELS, P. & L. DEVROYE (1984). Strong laws for the maximal k-spacing when $k \leqslant c \log n$. *Z. Wahrscheinlichkeitstheorie verw. Gebiete 66,* 315-334.

4 HILL, B.M. (1975). A simple approach to inference about the tail of a distribution. *Ann. of Statistics 3,* 1163-1174.

5 CSÖRGÖ, S., P. DEHEUVELS, D.M. MASON (1984). *Kernel estimates of the tail index of a distribution,* Techn. report Laboratoire de Statistique Theorique et Applique, Université Pierre et Marie Curie, Paris.

6 REISS, R.D. (1984). *Tagungsbericht 14,* Mathematisches Forschungs institut, Oberwolfach.