

On the distribution function of extreme flood discharge

MIKHAIL V. BOLGOV

Water Problems Institute, Russian Academy of Science, Gubkin Street 3, 117971 Moscow, Russia

e-mail: bolgov@iwapr.msk.su

VLADILEN F. PISARENKO

International Institute of Earthquake Theory and Prediction, Russian Academy of Science, Moscow, Russia

MARINA I. FORTUS

Institute of Atmospheric Physics, Russian Academy of Science, Moscow, Russia

Abstract The examination of sample histograms of river runoff data for the Pacific coast of Russia revealed that it is reasonable to consider the hypothesis that the distribution density for annual maximum of runoff values must include a component with the so-called "fat tail". To describe the tails of maximum runoff histograms, we recommend using the distribution in the form of a mixture of two densities: the normal distribution for the main region and the Pareto distribution providing the slow decrease of the density in the tail region. To estimate the parameters of the mixture, the numerical method of solution of the maximum likelihood equations is proposed.

INTRODUCTION

There is long-standing experience on the application of probabilistic methods for maximum runoff estimation. A number of standard guides regulate the use of specific types of probability distributions. Nevertheless, the models for extreme runoff are not finally justified.

A great deal of papers beginning with the early publications by Kritsky & Menkel (1948) and including Kalinin (1967) and Rozhdestvensky *et al.* (1990) consider the agreement between the maximum water discharge data and the fitted statistical model. As a rule, this agreement is judged either by statistical tests, or more simply, by comparing some numerical characteristics of the empirical and theoretical distribution functions. In recent years, characteristics related to the so-called order statistics and their linear combinations have been used (Hosking, 1990). In a manner similar to the method of moments, systems of equations can be obtained for a wide class of distribution functions. These equations include certain sample order statistics and the parameters to be estimated. More complex statistics derived from the order statistics can also be helpful in making inferences about the underlying theoretical distribution. However, if the observed series are short, neither this technique nor the tests of goodness of fit can provide reliable inferences about the type of distribution function. This is the main difficulty in the case of trying to fit the distribution function to the whole range of discharge values.

Other approaches to this problem are known in the literature. Some authors suggest classifying the pre-smoothed histograms depending on their behaviour in a particular interval of discharge values. For example, Adamovsky & Pilon (1995) have attempted to study qualitatively the behaviour of the tails of the histograms of maximum runoff values. But, their findings have shown that the variety of admissible types of distributions can be very wide, adversely affecting the reliability of the statistical model. We consider as constructive the idea of analysing empirical histograms within a certain range of values and apply this idea here. To make the set of acceptable models as narrow as possible, we consider only geographical regions where hydrological conditions are similar. Moreover, we implement the preliminary data processing so that it is possible to consider that many series belong to the same parent population and can be pooled to form one homogeneous sample.

EMPIRICAL DISTRIBUTION OF MAXIMUM RUNOFF

The processes of runoff formation are highly complex and heterogeneous by their genetic nature as related to both different seasons and the years of different streamflow. Due to this fact, the observational series to be analysed make up statistically heterogeneous samples. This complicates the probabilistic analysis, particularly taking into account that the samples are limited in size.

As for the specific features of maximum runoff, (here we consider rivers of the Pacific coast of Russia as an example, Fig. 1), it should be noted that the heterogeneity of runoff series has been mentioned by many investigators. A visual inspection of empirical distribution functions and histograms points clearly to the heterogeneity of the samples leading to significant underestimation of rare probabilities if the standard estimation technique is employed (for example, the approximation by the three-parametric gamma distribution, Fig. 2). Of course, we can reduce the extent of the discrepancy between an "empirical" and a "theoretical" curve by assigning a large enough value to the skewness of the distribution function. However, this value would vary considerably from river to river. Evidently, it is unlikely that a great spatial variability of skewness is hydrologically meaningful. So, such an approach would be subjective in its character.

We proceed from the assumption that it is impossible to make an inference about one or another type of the preferred distribution, taking as a basis only data for one river. In order to reveal the specific features of the underlying probabilistic models, a great body of data must be analysed jointly.

To increase the sample size, we suggest to jointly analyse the series either originated from genetically homogeneous regions or reduced so as to be independent of morphometrical or other hydrological characteristics. As for the maximum runoff data, these can be, for example, specific runoff series reduced to a unit basin area and adjusted to nonregulated runoff conditions. But the reducing procedure depends on the values of runoff themselves, and in order to diminish this dependence, we consider standardizing each data series by its empirical mean and standard deviation.

As indicated above, the shapes of individual histograms exhibit appreciable random scatter, therefore an additional averaging should be made over the collection of

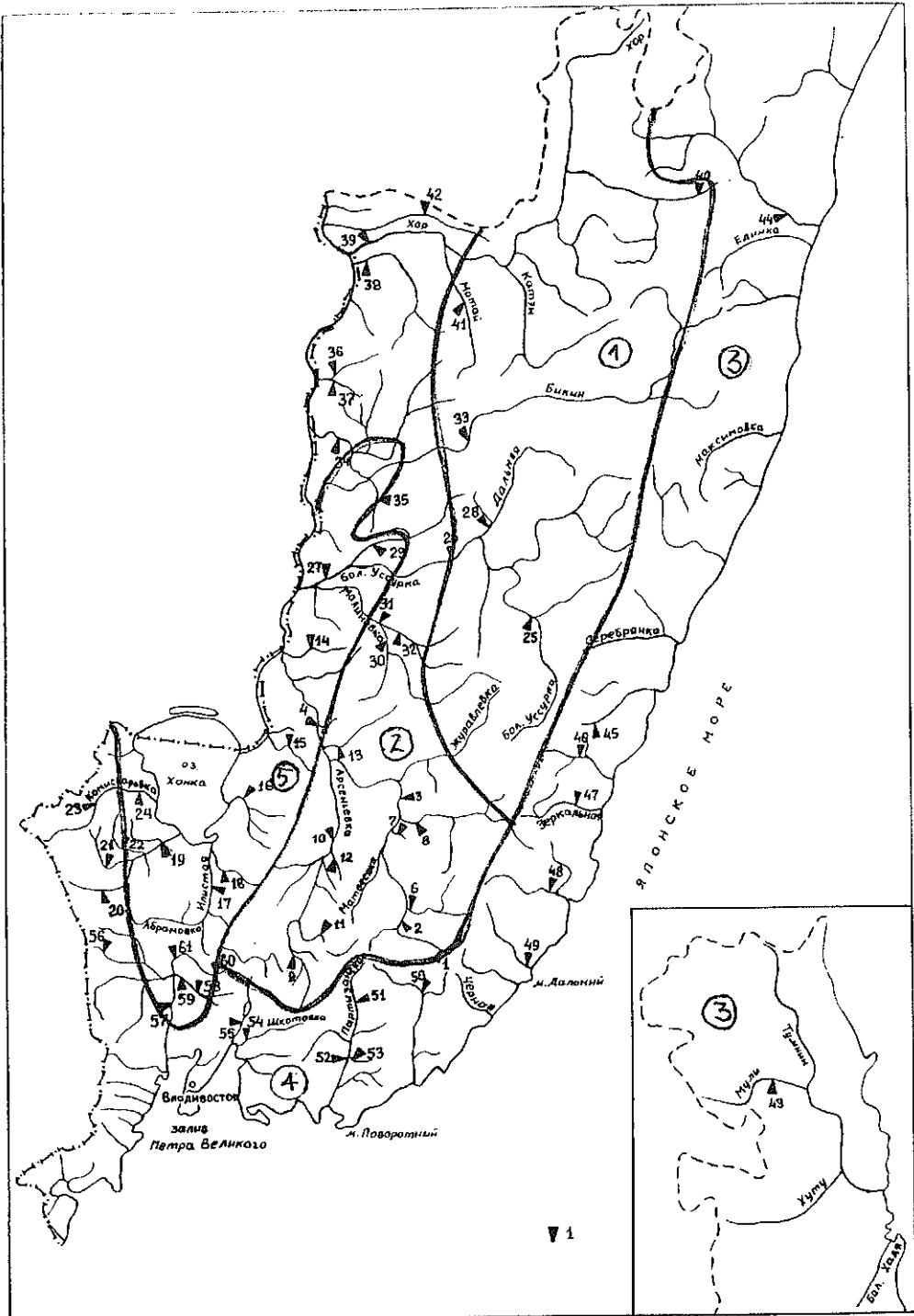


Fig. 1 Primor'ye territory subdivided into hydrologically homogeneous regions and allocation of points of runoff observation.

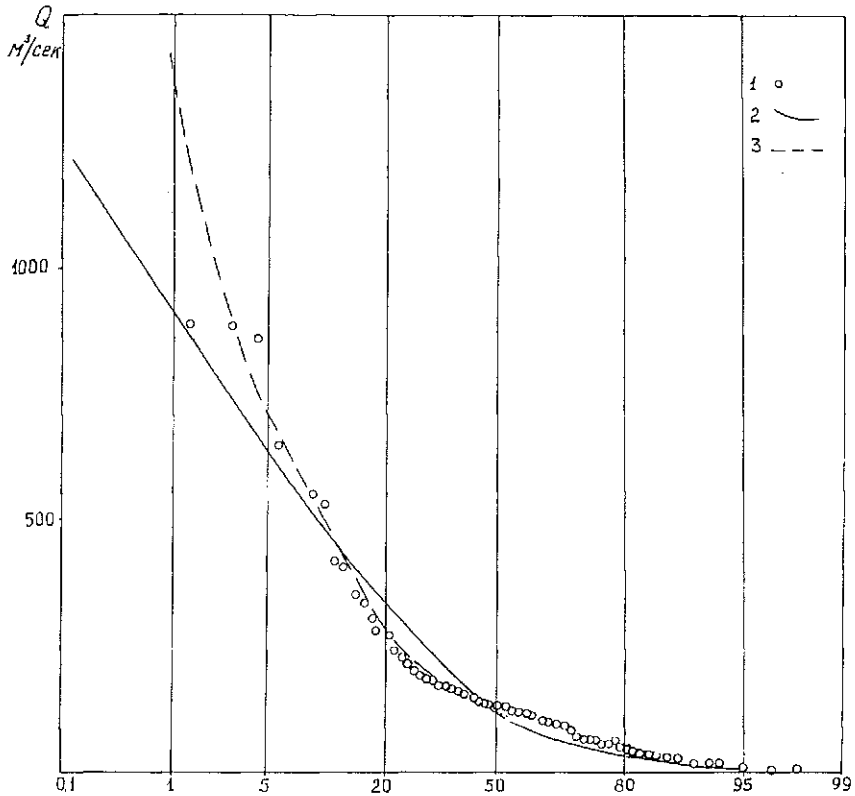


Fig. 2 Cumulative distribution function for maximum runoff for the River Komarovka near the point Sakharny zavod. 1: empirical points; 2: the approximation by Kritsky-Menkel distribution (sample estimation); 3: the approximation by the mixed distribution including the "fat tail" component (equation (4)).

histograms. The empirical probability density functions generalized in this way reveal the following specific feature: when moving from moderate streamflow values to extremely high values, the rate of the density decrease slows down markedly.

The histograms of this type are hardly likely to be approximated by a simple analytical expression describing empirical curves adequately in the whole range of values. Moreover, the gamma-distribution densities (including the three-parameter modification) most frequently used in hydrology, are monotonically decreasing functions in the entire range of values if the coefficient of variation exceeds 1.0 (this is just the case for the majority of the rivers in question). This monotone decrease has not been observed in maximum runoff histograms displaying one significant local maximum value.

PROBABILITY MODEL

The examination of the histograms suggests that it is reasonable to consider the hypothesis that the probability density for annual maximum runoff values must include a component having a so-called "fat tail". We define as a tail the part of the histogram

to the right of its median point a . Then, to describe the tail for $x > a$, we recommend using the distribution in the form of mixture of two densities: the normal distribution

$$f(x, a, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a)^2}{2\sigma^2}} \frac{1}{[1 - \Phi(a, \sigma)]} \quad (1)$$

$$\Phi(a, \sigma) = \int_{-\infty}^a \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\xi-a)^2}{2\sigma^2}} d\xi \quad (2)$$

(where we see that the mean value of the normal component is additionally supposed to coincide with the median a), and the Pareto distribution

$$\varphi(x, a, \beta) = \frac{a^\beta}{x^{1+\beta}}; \quad x \geq a \quad (3)$$

The Pareto distribution serves in our case as the fat component of the mixture providing the slow decrease in the tail region. The sought-for mixed density can be written in the form:

$$\psi(x, a, \sigma, \lambda) = \lambda f(x, a, \sigma) + (1 - \lambda)\varphi(x, a, \beta) \quad (4)$$

Having chosen the truncation point a beforehand, we reduce to three the number of unknown parameters in equation (4) to be estimated from the truncated sample (i.e. the sample consisting of runoff values greater than a). So, first, the value a is determined from the total sample. Then the parameter σ of the normal distribution, the weight coefficient λ , and the Pareto parameter β are determined from the corresponding truncated samples.

REGIONAL ESTIMATORS

On the whole, we can state that the distribution of mixture type (4) is acceptable for maximum runoff values for the rivers on the Primor'ye territory. However, because of the fact that the conditions of the runoff formation vary distinctly even over this relatively small territory, we can expect that the distribution parameters would vary in their magnitude. We discuss this problem in more detail below.

According to the Russian Federal Programme (1994), there are five homogeneous regions in the Primor'ye territory (Fig. 1, Table 1) differing from one another by

Table 1 Hydrological characteristics of the regions.

No.	Name of region	Coefficient of variability of annual runoff	Annual runoff (mm)	Percentage of storm runoff (%)
1	Central part of Sikhote-Alin ridge	10-25	300-650	28-30
2	West slope of Sikhote-Alin ridge	30-50	150-300	25-27
3	East slope of Sikhote-Alin ridge	10-30	300-600	22-27
4	Southwest Primor'ye	80-200	150-450	18-21
5	West Primor'ye plain	80-100	60-150	15-17

specific features of their hydrological conditions. Among these, there is the region of central Sikhote-Alin (region 1), characterized by the greatest runoff magnitude. The estimated values of the distribution parameters for the maximum rainflow runoff also differ for the regions in question. Here, the "rainflow runoff" refers to a peak discharge originating from rainfall rather than other mechanisms. For example, for region 1 characterized by the greatest and least variable annual runoff, the variability of annual maximum rainflow runoff is a minimum. For the regions with more variable annual runoff, this variability is greater (Table 2). The revealed specific hydrological behaviour allows pooling of the data to provide regional samples.

After fitting the "fat tail"-model to the data, we conclude that the rate of decrease of the fat component is less for the regions with the larger coefficient of variation. We see from Table 2 that, in order to increase the reliability of the estimated parameters λ and β , it is reasonable to calculate these estimates from the unified sample consisting of regions 2–5. The result is that two different regionally-averaged parameters of mixture can be recommended for the typical form of the distribution tail (see Table 2, region 1 and regions 2–5).

Since the experience of application of the distribution recommended here is not wide, we consider, in addition, the behaviour of the tails of the probability distributions of maximum runoff for several neighbouring geographical regions where the rainflow flood is the significant component of the hydrological regime. The regionally-averaged parameters of the distributions of maximum rainflow runoff are given in Table 3, where the numerator is the fraction of the normal component in the mixture, λ , and the

Table 2 Regional estimates of parameters of mixture distribution (equation (4)) for the Primor'ye regions.

Region number from Fig. 1	Root mean square deviation for the normal distribution, σ	Pareto parameter, β	The weight coefficient for the normal component in the mixture λ	Coefficient of variation for maximum water discharge, C_V
1	1.07	1.54	0.63	0.61
2	-	1.16	0.0	0.92
3	-	1.21	0.0	0.95
4	1.06	1.28	0.45	1.09
5	1.54	1.38	0.42	1.07
Averaged over all the regions	1.23	1.46	0.49	
Averaged over the regions 2–5	1.29	1.47	0.47	

Table 3 Regional estimates of parameters of mixture distribution (equation (4)) for some Siberian regions.

Region name	λ	β
Northeast	0.78	1.39
Lena	0.68	1.56
Baikal	0.46	1.68
Upper Amur	0.0	1.42
Sakhalin	0.0	1.39

denominator is the Pareto parameter, β . These parameters do not vary noticeably from one region to another and we may conclude that the parameters obtained here for the Primor'ye territory do not contradict those for other regions.

CONCLUSION

In conclusion, it should be emphasized that the use of a standard estimation technique often leads to underestimation for discharges of low probability. An acceptable approximation can be obtained by taking truncated samples and utilizing a mixture of different distributions making allowance for the genetic differences in the flood-formation conditions in the years characterized by different stream flow. Our results are based on the observed values standardized with respect to their mean values and standard deviations. Since the moment of the second order for the recommended probability density function (equation (4)) (i.e. the variance) does not exist, the theoretical considerations underlying the outlined procedure require an additional justification presented in the Appendix.

Acknowledgements This work was supported by the Russian Foundation for Basic Research, project no. 96-05-64674.

REFERENCES

- Adamovskii, K. & Pilon, P. (1995) Nonparametric analysis of hydrologic extremes. In *Statistical and Bayesian Methods in Hydrological Sciences: an International Conference in Honour of Jacques Bernier* (11–13 September 1995, Paris), 9. UNESCO, Paris, France.
- Hosking, J. R. M. (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Statist. Soc.* **52**(1), 105–124.
- Kalinin, G. P. (1967) *Problems of Global Hydrology*. Gidrometeoizdat, Leningrad.
- Kritsky, S. N. & Menkel, M. F. (1948) On an agreement between probability distribution curves and river runoff data. In *Problems of Control of River Runoff*, vol. 3, 5–69.
- Rozhdestvensky, A. V., Ezhov, A. V. & Sakharyuk, A. V. (1990) *Estimation of Accuracy of Hydrological Computations*. Gidrometeoizdat, Leningrad.
- The Russian Federal Programme (1994) Protection of Primor'ye against floods. Report of the Russian Research Institute of Water Resources Management, Far East Branch, Vladivostok, Russia.

APPENDIX

On a probability distribution of random variables normalized by the sample standard deviations in the case that their theoretical variance is infinite

Our purpose here is to find out how the probability distribution transforms as a result of the transition to variables normalized by the sample standard deviations. We concentrate our particular attention to the distributions having fat tails. Our consideration is not mathematically rigorous, it is rather of heuristic character, but we believe that it is possible to make the arguments outlined here into the rigorous mathematical proof.

1. First, we consider a ratio of two independent random variables ξ/η . Let the probability density functions (PDFs) of the variables ξ , η , and ξ/η be $f(x)$, $g(x)$, and $w(x)$, respectively. Then

$$w(x) = \int_0^{\infty} f(xy)yg(y)dy \tag{A1}$$

In particular, if the distributions of ξ and η are of Pareto type

$$f(x) = \begin{cases} \frac{\alpha}{x^{1+\alpha}} & x \geq 1 \\ 0 & x < 1 \end{cases} \quad \alpha > 0$$

$$g(x) = \begin{cases} \frac{\beta}{x^{1+\beta}} & x \geq 1 \\ 0 & x < 1 \end{cases} \quad \beta > 0$$
(A2)

then we can obtain from (A1):

$$w(x) = \frac{\beta\alpha}{\beta + \alpha} \frac{1}{x^{1+\alpha}} \frac{1}{[\kappa(x)]^{\alpha+\beta}} \tag{A3}$$

where $\kappa(x) = \max(1; \frac{1}{x})$. Whence it follows that

$$w(x) \cong \frac{\beta\alpha}{\beta + \alpha} \frac{1}{x^{1+\alpha}} \quad x \rightarrow \infty \tag{A4}$$

It should be mentioned that the PDF of $w(x)$ tends to infinity as $x \rightarrow 0$ if $\beta < 1$ (but it is obviously always integrable). It is essential for us that its asymptotic behaviour as $x \rightarrow \infty$ is: $w(x) \sim \frac{1}{x^{1+\alpha}}$, i.e. it does not depend on the distribution of the denominator η , in other words, on the parameter β . This is the consequence of the fact that the variable η in the denominator is separated from zero ($\eta \geq 1$).

2. For the variables normalized by standard deviations the denominator has the following form:

$$\sqrt{\eta_n} = \sqrt{\frac{\zeta_1^2 + \dots + \zeta_n^2}{n}} \tag{A5}$$

Suppose that ζ_1, \dots, ζ_n are mutually independent, identically distributed variables, and, as in Section 1, ζ_k are independent of the nominator ξ (we shall discuss this assumption later). Denote the PDF of ξ_k^2 by $h(x)$ and the PDF of η_n by $g_n(x)$. The PDF of the sum

$$\zeta_1^2 + \dots + \zeta_n^2 \tag{A6}$$

is equal to the n -fold convolution $h^{[n]}(x)$ of the function $h(x)$

$$\begin{aligned}
 h^{[2]}(x) &= \int h(y)h(x-y)dy \\
 h^{[3]}(x) &= \int h(y)h^{[2]}(x-y)dy \\
 &\dots \\
 h^{[n]}(x) &= \int h(y)h^{[n-1]}(x-y)dy
 \end{aligned}
 \tag{A7}$$

The PDF $g^n(x)$ of the normalized sum $\frac{\zeta_1^2 + \dots + \zeta_n^2}{n}$ takes the form:

$$g_n(x) = h^{[n]}(nx)n \tag{A8}$$

According to equation (A1), the PDF of the ratio $\xi/\sqrt{\eta_n}$ is

$$w_n(x) = \int_0^\infty f(\sqrt{yx})\sqrt{y}g_n(y)dy \tag{A9}$$

We assume that $f(x)$ (the PDF of the nominator) is of Pareto type (A2). Then

$$w_n(x) = \frac{\alpha}{x^{1+\alpha}} \int_{1/x^2}^\infty \frac{g_n(y)}{y^{\alpha/2}} dy \tag{A10}$$

- Now our purpose is to show that the asymptotic behaviour of $w_n(x)$ as $x \rightarrow \infty$ is $(1/x(1 + \alpha))$ (i.e. the same as of the PDF $f(x)$). To do this, we must show that the integral

$$\int_{1/x^2}^\infty \frac{g_n(y)}{y^{\alpha/2}} dy \tag{A11}$$

is bounded by some constant.

If $\alpha > 2$, the PDF $f(x)$ is not of fat tail type: the variance is finite, and by virtue of the law of large numbers, we have: $\frac{\xi_1^2 + \dots + \xi_n^2}{n} \rightarrow E\xi_1^2$, so that the function $g_n(y)$ is delta-function (which is equal to zero in the vicinity of $y = 0$) resulting in the boundedness of the integral (A11).

Consider now the fat tail case when $\alpha < 2$. It is sufficient to show that the function $g_n(y)$ (or $h^{[n]}(nx)$, which is the same—see equation (A8)) does not tend to infinity as $x \rightarrow 0$. Here we must consider two cases:

Case one: the PDF $h(x)$ of any variable ξ_k^2 in the sum (A6) is such that $h(0) < \infty$. The convolution results in smoothing of the functions $h^{[2]}(x)$, $h^{[3]}(x)$ in such a way that $h^{[k]}(0) = 0$, $k = 2, 3, \dots$

Case two: the PDF $h(x)$ of any variable ξ_k^2 in the sum (A6) is such that $h(x) \rightarrow \infty$ as $x \rightarrow 0$ (for example, $h(x) \sim \frac{1}{x^{1-\gamma}}$, $0 < \gamma < 1$). But it is practically improbable to have such a singularity for the variables ζ_k , which were obtained from the observed variables $\zeta_k \sim$ as the deviations from their means: $\zeta_k = \tilde{\zeta}_k - b$. If it were so (i.e. if $h(x) \sim \frac{1}{x^{1-\gamma}}$, $\gamma > 0$), then, as a result of the summation procedure, we would have $h^{[k]}(x) \sim x^{k\alpha-1}$ and $h^{[k]}(0) = 0$.

Thus, we have proved that the asymptotic behaviour of the PDF $w_n(x)$ of the normalized ratio $\xi / \sqrt{\eta_n}$ as $x \rightarrow \infty$ is the same as that for the PDF of the nominator ξ . Now we return to our supposition on the independence between the nominator ξ and variables ζ_k . It is clear that this is not the case for the normalized variables in question. Nevertheless, it is clear that as n tends to infinity, the nominator ξ will become more and more independent of the sum (A6). So, our supposition about such an independence will hold asymptotically as $n \rightarrow \infty$.

In conclusion we shall prove one more statement: If $\beta < 2$, then the PDF of the ratio $\xi / \sqrt{\frac{\zeta_1^2 + \dots + \zeta_n^2}{n}}$ does depend on the number n . Suppose that the PDFs of the variables ζ are from the class of *stable* PDF with the parameter $\beta/2$ (the PDF of the sum $\zeta_1^2 + \dots + \zeta_n^2$ will tend to the stable law if, for example, ζ_k are distributed according to the Pareto distribution with the parameter β). Then in consequence of the main property of stable laws, we have:

$$\frac{\zeta_1^2 + \dots + \zeta_n^2}{n} \stackrel{d}{=} \left(\sum_{j=1}^n \left(\frac{1}{n} \right)^{\beta/2} \right)^{2/\beta} \zeta_1 = n^{\frac{2}{\beta}-1} \zeta_1^2 \tag{A12}$$

Here the symbol $\stackrel{d}{=}$ means equivalence in distribution. Therefore the ratio

$$\xi / \sqrt{\frac{\zeta_1^2 + \dots + \zeta_n^2}{n}}$$

is distributed as the variable

$$\frac{\xi}{n^{1/\beta-1/2} |\zeta_1|} \tag{A13}$$

It is seen from equation (A13) that the distribution in question does depend on n , moreover, because of $\left(\frac{1}{\beta} - \frac{1}{2} \right) > 0$, this distribution is concentrated in zero point as $n \rightarrow \infty$.