E11-2003-116

M. Stehlík[1], G. A. Ososkov[2]

EFFICIENT TESTING OF THE HOMOGENEITY, SCALE PARAMETERS AND NUMBER OF COMPONENTS IN THE RAYLEIGH MIXTURE

[1]Department of Probability and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia; E-mail: Stehlik@fmph.uniba.sk
[2]E-mail: Ososkov@jinr.ru

1 Introduction

This paper is devoted to the statistical problem of expanding an experimental distribution of transverse momenta $P_\perp$ into a series of Rayleigh distributions and can be considered as a continuation of [Efimova et al 89]. The physical background of this problem arises in the emulsion experiment studying the dynamics of inelastic collision of fast heavy particles as nuclei ^{22}Ne with the photoemulsion nuclei by momenta 4.1 A geV/c. The spectrum of transverse momenta for inclusive experiment can bear the quite important information about the generation process of secondary particles, whether this process is direct or is going through some intermediate stages. As it is known (see, for example, [Efimova et al 89]), transverse momenta are distributed according to the Rayleigh law. However depending on the collision model (one of more than one channels of the particle generation) the $P_\perp$ distribution can be described by just one Rayleigh distribution or by a series

$$f(y; P_\perp) = \sum_{i=1}^{k} a_i \frac{y}{\sigma_i^2} \exp(-\frac{y^2}{2\sigma_i^2}), y > 0, \sum a_i = 1$$

with some unknown k, σ_i and a_i. The formulation of mathematical problem is complicated due to experimental restrictions caused by different conditions of registering secondary particles depending on the emanating angle θ of those particles in respect to the collision axis. That was taken into account in [Efimova et al 89] by inventing corresponding statistical weights of measured $P_\perp$ depending on θ and allowed to elaborate a method of expanding the experimental $P_\perp$-distribution into one or the mixture of two Rayleigh distributions.

However the generalization of the Rao-Smirnov ω^2 test proposed in the previous paper to choose the hypothesis about the expansion type (one or the mixture of two Rayleigh distributions) was not proven to be optimal. Therefore at the present paper we focus ourselves on two problems:

- Constructing of the high efficient testing procedure of the homogeneity hypothesis with general and mixture alternatives. In the case of general alternative is constructed test AOBS (asymptotically optimal in the Bahadur sense (see [Bahadur 65])).

- if homogeneity holds, constructing the exact likelihood ratio (LR) test of the scale parameter of the Rayleigh distribution optimal in the Bahadur sense.

The paper is organized as follows. In section *Subpopulation Model* we approximate the mixture model by the subpopulation one when sample size is between 40 and 50. In section *Homogeneity testing* we express the distribution function of the exact LR test of the homogeneity in the terms of spherical coordinates and provide the procedure for critical values computation by simulations. We also provide the table of obtained critical values. In section *Asymptotical optimality in the Bahadur sense* we discuss the asymptotical optimality of the derived test. In section *The maximum likelihood estimation and efficient exact testing of the common scale parameter* we provide the ML estimator of the common scale parameter σ^2 in the homogeneous Rayleigh sample. We also derive the exact distribution function and the density of the LR test of the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_1 : \sigma^2 \neq \sigma_0^2$$

of the common parameter. We also derive the exact power function of the test and compare the exact distribution of the LR test with the asymptotic one. In section *Efficient testing of the number of components in the Rayleigh mixture* we construct the procedure for the LR testing of the hypotheses of the number of components m in the Rayleigh mixture for $m = 2$ and 3. In Appendix we provide some properties of the Lambert W function.

2 Subpopulation model

We have physical reasons for considering that $40 \leq N \leq 50$. In smaller samples we recommend the exact LR testing of the considered hypothesis about the number of components m in the mixture. Such a procedure, however, leads to a rather laborious computation. Practical difficulties arise specially due to the likelihood frequently having multiple local extremes. In our approach we approximate the exact mixture model given by the mixture density

$$f(y|\sigma^2) = \sum_{i=1}^{m} \pi_i f(y|\sigma_i^2), \ \pi_1 + ... + \pi_m = 1 \tag{1}$$

with the subpopulation model that is frequently used as motivation for the mixture density (see [Susko 03]). The subpopulation model supposes that there are m subpopulations, that π_j is the probability of selecting an individual from subpopulation j and $f(y|\sigma_j^2)$, the component density, is the conditional density for Y given that the observation is from the jth subpopulation. Since the true classification of observations into subpopulations is unobserved, the marginal density (1) is typically used for the observations.

3 Homogeneity testing

In this section we derive the exact distribution of the LR test of the homogeneity for the Rayleigh distribution. We consider a statistical model with N independent observations $y_1, ..., y_N$ which are distributed according to Rayleigh densities

$$f(y_i|\sigma_i^2) = \begin{cases} \frac{y_i}{\sigma_i^2}\exp\{-\frac{y_i^2}{2\sigma_i^2}\}, & \text{for } y_i > 0, \\ 0, & \text{for } y_i \le 0. \end{cases} \tag{2}$$

Here $\sigma^2 := (\sigma_1^2, ..., \sigma_N^2)$ is vector of unknown scale parameters. Let us introduce the notation $X \sim R(\sigma^2)$ when X is distributed according to density (2) with the scale parameter σ^2 and $X \sim Exp(\lambda)$ when X is distributed according to the exponential density

$$f(x|\lambda) = \begin{cases} \lambda\exp\{-\lambda x\}, & \text{for } x > 0, \\ 0, & \text{for } x \le 0 \end{cases}$$

with the scale parameter λ. Now let us construct the efficient test of the homogeneity in the model (2). The null hypothesis has the form

$$H_0 : \ \sigma_1^2 = ... = \sigma_N^2. \tag{3}$$

The LR of the homogeneity test has the form

$$\lambda_N(y) = \frac{\max_{\sigma_1^2=...=\sigma_N^2} f(y, \sigma^2)}{\max_{\sigma^2} f(y, \sigma^2)},$$

where $f(y, \sigma^2) = \prod_{i=1}^N f(y_i|\sigma_i^2)$. After the optimization we obtain that

$$\lambda_N(y) = \frac{N^N(y_1...y_N)^2}{(y_1^2 + ... + y_N^2)^N}. \tag{4}$$

If $X \sim R(\sigma^2)$ holds, then we have

$$\frac{X^2}{2\sigma^2} \sim Exp(1). \tag{5}$$

Under the homogeneity hypothesis, the distribution of the likelihood ratio (4) does not depend on the unknown parameter σ^2. Furthermore, due to (5) we have that $\lambda_N(y)$ has under H_0 the same distribution as the homogeneity LR statistics

$$\frac{N^N x_1...x_N}{(x_1 + ... + x_N)^N}$$

of the homogeneous exponential sample $x_1, ..., x_N$ (see [Stehlík 03]). Due to the monotonous transformation $g(x) = \sqrt[N]{x}$ of the likelihood ratio (4) we obtain the interesting statistics of the homogeneity,

$$\frac{\sqrt[N]{y_1^2 \cdots y_N^2}}{\frac{y_1^2+\ldots+y_N^2}{N}}.$$

that is the ratio of the geometric and aritmetic mean of the squares of observations. What is more, the LR test of the homogeneity hypothesis (3) is asymptotically optimal in the Bahadur sense, as it is shown in the following section. The distribution of the LR test statistics $-\ln \lambda_N$ of the homogeneity under the null hypothesis is derived in the following theorem.

Theorem 1 *Let $y_1, ..., y_N$ are iid (independent, identically distributed) according to the Rayleigh distribution with the unknown scale parameter σ^2. Then the LR test statistics $-\ln \lambda_N$ has the form*

$$-\ln \lambda_N(y) = N \ln(\sum_{i=1}^{N} y_i^2) - N \ln N - \sum_{i=1}^{N} \ln(y_i^2) \tag{6}$$

and it has the same distribution as the function

$$\Phi_N(\varphi) = -N \ln N - 2 \ln(\prod_{j=1}^{N-1} \sin^{N-j} \varphi_j \prod_{k=1}^{N-1} \cos \varphi_k)$$

where the vector $\varphi = (\varphi_1, ..., \varphi_{N-1})$ is distributed on set $[0, \frac{\pi}{2}]^{N-1}$ according to the density

$$f_N(\varphi) = 2^{N-1}(N-1)! \prod_{j=1}^{N-1} \sin^{N-j} \varphi_j \prod_{k=1}^{N-1} \cos \varphi_k \prod_{l=1}^{N-2} \sin^{N-l-1} \varphi_l.$$

Remark

The main advantage of the provided distribution of the random variable Φ_N is the possibility of simulation of the density of the LR statistics (6) based on the random vector φ distributed on the compact set $[0, \frac{\pi}{2}]^{N-1}$ although the sample space of vector y is the unbounded positive cone $R^{+N} := \{y \in \mathbf{R}^N : y_1 > 0, \ldots, y_N > 0\}$ of the N-dimensional Euclidean space.

Proof.

Into the characteristic function ψ of the random variable $\lambda_N(y)$ we introduce the spherical coordinates $(r, \varphi_1, ..., \varphi_{N-1})$ of the R^{+N}. We have

$$y_1 = r\cos\varphi_1$$

$$y_2 = r\sin\varphi_1\cos\varphi_2$$

$$\ddots$$

$$y_{N-1} = r\sin\varphi_1\sin\varphi_2...\sin\varphi_{N-2}\cos\varphi_{N-1}$$

$$y_N = r\sin\varphi_1\sin\varphi_2...\sin\varphi_{N-2}\sin\varphi_{N-1}$$

where $r > 0$ and $\varphi_i \in [0, \frac{\pi}{2}]$ for $i = 1, ..., N-1$. After expressing the terms we obtain

$$\psi(t) = \int_{[0,\frac{\pi}{2}]^{N-1}} \exp\{it\Phi_N(\varphi)\} f_N(\varphi) d\varphi,$$

which is the characteristic function of the random variable defined in Theorem 1. This completes the proof. □

Our computation of the critical values of the test uses the fact that $\lambda_N(y)$ has under H_0 the same distribution as the homogeneity LR statistics

$$\frac{N^N x_1...x_N}{(x_1 + ... + x_N)^N}$$

where x_i are iid $Exp(1)$. For small dimensions we can compute the critical values from the exact c.d.f.s F_N of the test statistics $-\ln\lambda_N$. In [Stehlík 03] we can find, that in dimension 2 and 3 the c.d.f. has form

$$F_2(x) = \begin{cases} \sqrt{1-\exp(-x)}, & \text{for } x > 0, \\ 0, & \text{for } x \le 0 \end{cases}$$

and

$$F_3(x) = \begin{cases} \int_{a(x)}^{b(x)} \frac{1}{s}\sqrt{s^2(1-s)^2 - \frac{4}{27}s\exp(-x)}ds, & \text{for } x > 0, \\ 0, & \text{for } x \le 0 \end{cases}$$

where $0 < a(x) < b(x) < 1$ are solutions of the algebraic equation

$$t(1-t)^2 = \frac{4}{27}\exp(-x).$$

In high dimensions the c.d.f.s and densities are much more complicated and estimates of the critical values can be obtained by the simulation. We simulate the critical values using the S-plus 4 software, number of simulations is $n = 500\ 000$. The following Table 1 contains the critical values $c_{\alpha,N}$ of the homogeneity LR test

statistics (6) in samples $40 \leq N \leq 44$ for various values of the level of significance α obtained from the simulations. Our simulation essentially use the fact, that the LR statistics (6) under the H_0 does not depend on the unknown value of the parameter σ^2.

Table 1. Critical values $c_{\alpha,N}$.

$\alpha \backslash N$	40	41	42	43	44
0.001	18.6919	18.9155	19.1999	19.4025	19.5655
0.005	15.6891	15.8459	15.9999	16.1951	16.4221
0.01	14.3399	14.3659	14.6599	14.9012	14.9599
0.05	10.3899	10.4045	10.5099	10.6412	10.7378

The following Figure 1 displays the dependence of critical values on the level α (x-axis) in dimensions $N = 40,\ 42$ and 43.

4 Asymptotical optimality in the Bahadur sense

In this section we briefly discuss the asymptotical optimality in the Bahadur sense. The test of the homogeneity derived in the previous section will be shown to be an AOBS.

Consider a testing problem $H_0 : \vartheta \in \Theta_0$ *vs* $H_1 : \vartheta \in \Theta_1 \setminus \Theta_0$, where $\Theta_0 \subset \Theta_1 \subset \Theta$. Further consider sequence $T = \{T_N\}$ of test statistics based on measurements $y_1, ..., y_N$ which are iid acording to an unknown member of an family $\{P_\vartheta : \vartheta \in \Theta\}$. We assume that large values of test statistics give evidence against H_0, which is the case of our test statistics $-\ln \lambda_N$ of homogeneity. For ϑ and t denote $F_N(t, \vartheta) := P_\vartheta\{s : T_N(s) < t\}$; $G_N(t) := \inf\{F_N(t, \vartheta) : \vartheta \in \Theta_0\}$. The quantity $L_n(s) = 1 - G_n(T_n(s))$ is called the attained level or the p-value. Suppose that for every $\vartheta \in \Theta_1$ the equality

$$\lim \frac{-2 \ln L_n}{n} = c_T(\vartheta)$$

holds a.e. P_ϑ. Then the nonrandom function c_T defined on Θ_1 is called the Bahadur exact slope of the sequence $T = \{T_n\}$. According to the theorem of Raghavachari and Bahadur (see [Raghavachari 70]) the inequality

$$c_T(\vartheta) \leq 2K(\vartheta, \Theta_0) \tag{7}$$

holds for each $\vartheta \in \Theta_1$. Here $K(\vartheta, \Theta_0) := \inf\{K(\vartheta, \vartheta_0) : \vartheta_0 \in \Theta_0\}$ and $K(\vartheta, \vartheta_0)$ denotes the Kullback-Leibler information number defined by the formula

$$K(\vartheta, \vartheta_0) := \begin{cases} \int \ln \frac{dP_\vartheta}{dP_{\vartheta_0}} dP_\vartheta, & \text{if } P_\vartheta << P_{\vartheta_0}, \\ +\infty, & \text{otherwise.} \end{cases}$$

If (7) holds with the equality sign for all $\vartheta \in \Theta_1$, then the sequence T is said to be asymptotically optimal in the Bahadur sense. The maximization of $c_T(\vartheta)$ is a nice statistical property, because the greater the exact slope is, the more one can be convinced that the rejected null hypothesis is indeed false. The class of such statistics is apparently narrow, though it contains under certain conditions the LR statistics (see [Bahadur 65], [Bahadur 67], [Rublík 89, 1] and [Rublík 89, 2]). Rublík proved AO of the LR statistic under regularity condition which is shown to be fullfiled by regular normal, exponential and Laplace distribution under additional assumption that Θ_0 is a closed set and Θ_1 is either closed or open in metric space Θ. In [Stehlík 03] is proved, that the homogeneity test is AOBS in the case of observations distributed exponentially. Due to the connection (5) between the Rayleigh and exponential distribution is the homogeneity test of the Rayleigh distribution also AOBS. For more extensive discussion on asymptotical optimality see also monograph [Nikitin 95].

5 The maximum likelihood estimation and efficient exact testing of the common scale parameter

In this section we derive the exact LR test of the common scale parameter σ^2 in the homogeneous iid Reley sample. Typically, after the acceptation of the homogeneity hypothesis, one wants to know the probable value of the unknown common scale parameter σ^2. We will consider the maximum likelihood (ML) estimation of the common scale parameter. The ML estimator of the parameter σ^2 in the homogeneous iid Rayleigh sample $y_1, ..., y_N$ has the form

$$\hat{\sigma}^2 = \frac{y_1^2 + ... + y_N^2}{2N}.$$

Let us introduce the notation $y \sim \Gamma(N, 1)$ to express that y is distributed according to the density

$$f(y) = \frac{y^{N-1}}{(N-1)!} e^{-y}, \; y > 0, \; N = 1, 2, 3, ...$$

In the following theorem we derive the exact distribution of the LR test of the hypothesis

$$H_0 : \sigma^2 = \sigma_0^2 \text{ versus } H_1 : \sigma^2 \neq \sigma_0^2 \tag{8}$$

in the homogeneous iid Rayleigh sample.

Theorem 2 *The statistics* $-\ln\lambda$ *of the LR test of the hypothesis (8) has the form*

$$-\ln\lambda_N(y) = G_N(\frac{1}{2\sigma_0^2}\sum_{i=1}^{N} y_i^2) - G_N(N),$$

where for $N = 1, 2, \ldots$ *we introduce the function*

$$G_N(x) = \begin{cases} x - N\ln(x), & \text{for } x > 0, \\ 0, & \text{for } x \leq 0. \end{cases}$$

Under the null hypothesis the c.d.f. of $-\ln\lambda_N$ *has the form*

$$F_N(\rho) = \begin{cases} \mathcal{F}_N(-N\ W_{-1}(-\exp(-1-\frac{\rho}{N}))) - \mathcal{F}_N(-N\ W_0(-\exp(-1-\frac{\rho}{N}))), & \rho > 0, \\ 0, & \rho \leq 0 \end{cases}$$

and the density of $-\ln\lambda_N$ *has the form*

$$f_N(\rho) = \begin{cases} h_N(1,\rho) - h_N(0,\rho), & \text{for } \rho > 0, \\ 0, & \text{for } \rho \leq 0. \end{cases}$$

Here for $k \in \{-1, 0\}$ *and* $w \in \mathbf{R}$ *is* $W_k(w)$ *the value of the k-th branch of the Lambert W function at the point w (see Appendix),* $\mathcal{F}_N$ *is the c.d.f. of the* $\Gamma(N,1)$*-distribution and for* $r > 0$ *we define*

$$h_N(k,r) = \frac{(-N)^{N-1}}{\Gamma(N)} \frac{\{W_{-k}(-\exp(-1-\frac{r}{N}))\}^N}{1 + W_{-k}(-\exp(-1-\frac{r}{N}))} \exp(NW_{-k}(\exp(-1-\frac{r}{N}))).$$

The Wilks statistics $-2\ln\lambda$ *has under the* H_0 *c.d.f. of the form*

$$F_N(\tau) = \begin{cases} \mathcal{F}_N(-NW_{-1}(-e^{-1-\frac{\tau}{2N}})) - \mathcal{F}_N(-NW_0(-e^{-1-\frac{\tau}{2N}})), & \tau > 0, \\ 0, & \tau \leq 0, \end{cases} \tag{9}$$

and the density of the form

$$f_N(\tau) = \begin{cases} \frac{1}{2}\{h_N(1,\frac{\tau}{2}) - h_N(0,\frac{\tau}{2})\}, & \text{for } \tau > 0\ , \\ 0, & \text{for } \tau \leq 0. \end{cases}$$

Proof. Under the H_0 is the sample $\frac{x_1^2}{2\sigma_0^2}, \ldots, \frac{x_N^2}{2\sigma_0^2}$ iid from the $Exp(1)$ distribution (see(5)). Theorem 6 in [Stehlík 03] completes the proof. □

The test given by the Theorem 2 has the UUMP property (see [Lehmann 64]). The useful property of this test is its asymptotical optimality in the sense of the Bahadur exact slopes (see previous section), which is proved due to the connection (5) between the Rayleigh and exponential distributions in [Stehlík 03]. In the following theorem we derive the exact power of the test of the hypothesis (8) .

Theorem 3 *The exact power $p(\sigma^2, \alpha)$ of the LR test based on the Wilks statistics of the hypothesis (8) in the Rayleigh homogeneous iid sample on the level α at the point σ^2 of the alternative has the form*

$$p(\sigma^2, \alpha) = 1 - \mathcal{F}_N(-N\frac{\sigma_0^2}{\sigma^2}W_{-1}(-e^{-1-\frac{c_{\alpha,N}}{2N}})) + \mathcal{F}_N(-N\frac{\sigma_0^2}{\sigma^2}W_0(-e^{-1-\frac{c_{\alpha,N}}{2N}})),$$

where $c_{\alpha,N}$ denotes the critical value of the exact test of the hypothesis (8) on the level α.

Proof

The critical region based on the Wilks statistics of the LR test of the hypothesis (8) on the level of significance α has the form $W_c = \{y \in Y : -2\ln\lambda_N(y) > c\}$ such that $P\{W_c|\sigma^2 = \sigma_0^2\} = \alpha$, where Y denotes the sample space. The power $p(\sigma_1^2, \alpha)$ of the test of the hypothesis (8) at the point σ_1^2 of the alternative is equal to $P\{W_c|\sigma^2 = \sigma_1^2\}$. Applying Theorem 2 in [Stehlík 03] we obtain the equality

$$1 - P\{W_c|\sigma^2 = \sigma_1^2\} = \mathcal{F}_N(-N\frac{\sigma_0^2}{\sigma_1^2}W_{-1}(-e^{-1-\frac{c_{\alpha,N}}{2N}})) - \mathcal{F}_N(-N\frac{\sigma_0^2}{\sigma_1^2}W_0(-e^{-1-\frac{c_{\alpha,N}}{2N}})).$$

This completes the proof. □

The ML estimator $\hat{\sigma}_N^2$ of the parameter σ^2 is consistent and $-2\ln\lambda_N$ has asymptotically χ_1^2-distribution (see [Wilks 67]). Let us briefly investigate how the exact distribution of the LR test differs from the asymptotic one. The test based on the asymptotics is oversized. The following Table 2 gives the oversizing of the asymptotical test for small samples. Here α is the size of the test given from the Wilks asymptotics while $\alpha_{e,N}$ is the exact size of the same test. We calculate from the formula $\alpha_{e,N} = 1 - F_N(\chi_{\alpha,1}^2)$. Here $\chi_{\alpha,1}^2$ denotes $(1-\alpha)$-quantile of the asymptotical χ_1^2-distribution and F_N is the exact cdf of the Wilks statistics $-2\ln\lambda$ of the LR test of the hypothesis (8) under the H_0 given by the formula (9) .

Table 2: The exact sizes $\alpha_{\epsilon,N}$

$\alpha \backslash N$	1	2	3	4	5
0.00001	0.229211e-4	0.178154e-4	0.155230e-4	0.142351e-4	0.134204e-4
0.00002	0.445707e-4	0.347445e-4	0.303720e-4	0.279358e-4	0.264021e-4
0.00005	0.1070863e-3	0.838629e-4	0.736765e-4	0.680620e-4	0.645496e-4
0.0001	0.2073771e-3	0.1630849e-3	0.1439040e-3	0.1334225e-3	0.1268967e-3
0.0002	0.4007327e-3	0.3167122e-3	0.2808525e-3	0.2614285e-3	0.2493896e-3
0.0005	0.9536900e-3	0.7598856e-3	0.6789730e-3	0.6356518e-3	0.6089504e-3
0.001	0.18315866e-2	0.14706397e-2	0.13227325e-2	0.12442198e-2	0.11960162e-2
0.002	0.35061415e-2	0.28417427e-2	0.25749120e-2	0.24344361e-2	0.23484946e-2
0.005	0.82247349e-2	0.67718253e-2	0.62044830e-2	0.59087828e-2	0.57286067e-2
0.01	0.015599286	0.013037809	0.012058871	0.011552053	0.011244013
0.02	0.029448482	0.025065314	0.023424550	0.022579936	0.022067611
0.05	0.067701923	0.059361294	0.056314364	0.054754992	0.053810812
$\alpha \backslash N$	6	7	8	9	10
0.00001	0.128631e-4	0.124589e-4	0.121533e-4	0.119145e-4	0.117233e-4
0.00002	0.253549e-4	0.245975e-4	0.240255e-4	0.235786e-4	0.232200e-4
0.00005	0.621610e-4	0.604363e-4	0.591357e-4	0.581205e-4	0.573069e-4
0.0001	0.1224689e-3	0.1224689e-3	0.1168726e-3	0.1149965e-3	0.1134937e-3
0.0002	0.2412404e-3	0.2353750e-3	0.2309574e-3	0.2275143e-3	0.2247565e-3
0.0005	0.5909296e-3	0.5779784e-3	0.5682339e-3	0.5606426e-3	0.5545638e-3
0.001	0.11635466e-2	0.1140235e-2	0.11227052e-2	0.11090527e-2	0.10981231e-2
0.002	0.22907016e-2	0.22492459e-2	0.22180864e-2	0.21938248e-2	0.21744048e-2
0.005	0.56076661e-2	0.55209911e-2	0.54558724e-2	0.54051793e-2	0.53646073e-2
0.01	0.0110374652	0.0108895082	0.0107783725	0.0106918635	0.010622626
0.02	0.0217243535	0.0214785418	0.0212939228	0.021150213	0.0210351898
0.05	0.0531785833	0.0527258996	0.0523858952	0.0521212038	0.0519093212
$\alpha \backslash N$	11	20	30	40	50
0.00001	0.115663e-4	0.108587e-4	0.105711e-4	0.104276e-4	0.103418e-4
0.00002	0.229270e-4	0.216047e-4	0.210675e-4	0.207998e-4	0.206392e-4
0.00005	0.566409e-4	0.536422e-4	0.524229e-4	0.518152e-4	0.514510e-4
0.0001	0.1122631e-3	0.1067253e-3	0.1044757e-3	0.1033529e-3	0.1026806e-3
0.0002	0.2224988e-3	0.2123424e-3	0.2082143e-3	0.2061548e-3	0.2049209e-3
0.0005	0.5495886e-3	0.5272108e-3	0.5181143e-3	0.5135745e-3	0.5108540e-3
0.001	0.10891785e-2	0.10489474e-2	0.10325913e-2	0.10244260e-2	0.10195322e-2
0.002	0.21585122e-2	0.20870295e-2	0.20579594e-2	0.20434441e-2	0.20347421e-2
0.005	0.53314054e-2	0.51820403e-2	0.51212668e-2	0.50909105e-2	0.50727079e-2
0.01	0.0105659642	0.010311006	0.010207222	0.010155366	0.010124268
0.02	0.0209410536	0.00517344	0.020344780	0.020258532	0.020206799
0.05	0.0517358893	0.050954881	0.050636560	0.050477398	0.050381907

6 Efficient testing of the number of components in the Rayleigh mixture

In this section we construct the efficient testing procedure of the number of components m in the Rayleigh mixture for $m = 2$ and 3.

6.1 Case of the alternative $H_1 : m = 2$

In this section we consider the alternative of the form $H_1 : m = 2$. The hypothesis

$$H_0 : m = 1 \text{ versus } H_1 : m = 2 \tag{10}$$

in the mixture model (1) can be approximate due to the subpopulation model by the hypothesis

$$H_0 : \sigma_1^2 = ... = \sigma_n^2 \text{ versus } approx H_1 : \exists M_1, M_2, M_1 \cup M_2 = \{1, ..., n\}, \tag{11}$$

$$M_1 \cap M_2 = \emptyset, M_1, M_2 \neq \emptyset, \forall j \in M_1 : \sigma_j^2 = \sigma_1^2, \forall j \in M_2 : \sigma_j^2 = \sigma_2^2, \sigma_1^2 \neq \sigma_2^2$$

e.g. by the null hypothesis of the homogeneity with the modified alternative, which is actually a subset of the alternative of the hypothesis of the homogeneity. We construct the LR test of the hypothesis (11) which approximates the hypothesis (10) . Let $y_1, ..., y_N$ are distributed according to Rayleigh densities. The LR of the test of the hypothesis (11) has the form

$$\lambda_N(y) = \frac{\max_{\sigma_1^2 = ... = \sigma_N^2} f(y, \sigma^2)}{\max_{approx H_1} f(y, \sigma^2)}.$$

To compute the denominator $\max_{approx H_1} f(y, \sigma^2)$ we proceed as follows. Suppose that $\{y_{i_1}, ..., y_{i_K}\}$, $0 < K < N$ are the observations from the Rayleigh distribution with the scale parameter σ_1^2 and the other observations are distributed according to the Rayleigh distribution with the scale parameter σ_2^2. Without lost of generality we can suppose that $i_j = j, j = 1, ..., K$. Then the ML estimators of parameters σ_1^2 and σ_2^2 have the form

$$\hat{\sigma}_1^2 = \frac{y_1^2 + ... + y_K^2}{2K}$$

and

$$\hat{\sigma}_2^2 = \frac{y_{K+1}^2 + ... + y_N^2}{2(N - K)}.$$

In such case we have

$$\lambda_N(y) = \frac{N^N}{K^K (N - K)^{N-K}} \frac{(y_1^2 + ... + y_K^2)^K (y_{K+1}^2 + ... + y_N^2)^{N-K}}{(y_1^2 + ... + y_N^2)^N}.$$

In practice is the classification of measured data into subpopulations unobserved and we must consider the all possibilities by the finding the maximum. For $0 < K < N$ let $P(K)$ denotes the all K-subsets $\{i_1, ..., i_K\}$ of the set $\{1, 2, ..., N\}$. For $p = \{i_1, ..., i_K\} \in P(K)$ we denote $approxL(p) :=$

$$= 2^N K^K (N-K)^{N-K} \exp(-N) y_1 ... y_N (y_{i_1}^2 + ... + y_{i_K}^2)^{-K} (y_{i_{K+1}}^2 + ... + y_{i_N}^2)^{-N+K}.$$

The LR of the test of the hypothesis (11) has form

$$\lambda_N(y) = \frac{\max_{\sigma_1^2=...=\sigma_N^2} f(y, \sigma^2)}{\max_{0<K<N,\ p\in P(K)} approxL(p)} = \min_{0<K<N,\ p\in P(K)} \frac{\max_{\sigma_1^2=...=\sigma_N^2} f(y, \sigma^2)}{approxL(p)}.$$

Finally we obtain the formula

$$\lambda_N(y) = \min_{0<K<N, p\in P(K)} \frac{N^N}{K^K (N-K)^{N-K}} \frac{(y_{i_1}^2 + ... + y_{i_K}^2)^K (y_{i_{K+1}}^2 + ... + y_{i_N}^2)^{N-K}}{(y_1^2 + ... + y_N^2)^N}. \quad (12)$$

The main advantages of the test statistic (12) is that under the H_0 it does not depend on the unknown value of the parameter σ^2. The distribution of the LR test statistics $-\ln \lambda_N$ where λ_N is given by the formula (12) under the null hypothesis is derived in the following theorem.

Theorem 4 *Let $y_1, ..., y_N$ are iid according to the Rayleigh distribution with the unknown scale parameter σ^2. Then the LR test statistics $-\ln \lambda_N$ where λ_N is given by the formula (12) has the form*

$$-\ln \lambda_N(y) = -\min_{0<K<N,\ p\in P(K)} \{N \ln N - K \ln K - (N-K) \ln(N-K) +$$

$$+K \ln(\sum_{n=1}^{K} y_{i_n}^2) + (N-K) \ln(\sum_{n=1}^{N-K} y_{i_n}^2) - N \ln(\sum_{n=1}^{N} y_n^2)\}$$

and it has the same distribution as the random variable

$$U_N = -\min_{0<K<N,\ p\in P(K)} \{N \ln N - K \ln K - (N-K) \ln(N-K) +$$

$$+K \ln(\sum_{n=1}^{K} u_{i_n}) + (N-K) \ln(\sum_{n=1}^{N-K} u_{i_n}) - N \ln(\sum_{n=1}^{N} u_n)\}$$

where $u_1, ..., u_N$ are iid according to $Exp(1)$.

Remark

The main advantage of the provided distribution of the random variable U_N is the possibility of simulation of the density of the LR statistics $-\ln\lambda_N$ based on the $Exp(1)$ simulations.

Proof. Under the H_0 is the sample $\frac{x_1^2}{2\sigma_0^2},...,\frac{x_N^2}{2\sigma_0^2}$ iid from the $Exp(1)$ distribution (see(5)). The independence of the LR statistics (12) on the real value of the scale parameter σ^2 under the null hypothesis completes the proof. □

6.2 Case of the alternative $H_1 : m = 3$

In this section we consider the alternative of the form $H_1 : m = 3$. The hypothesis

$$H_0 : m = 1 \text{ versus } H_1 : m = 3 \tag{13}$$

in the mixture model (1) can be approximate due to the subpopulation model by the hypothesis

$$H_0 : \ \sigma_1^2 = ... = \sigma_n^2 \ \text{ versus } approxH_1 : \exists \text{ nonempty disjoint subsets } M_1, M_2, M_3 \tag{14}$$

of the set $1,...,n$ such that $\forall j \in M_1 : \sigma_j^2 = \sigma_1^2, \forall j \in M_2 : \sigma_j^2 = \sigma_2^2, \forall j \in M_3 : \sigma_j^2 = \sigma_3^2$, where σ_1^2, σ_2^2 and σ_3^2 are different scale parameters.

We construct the LR test of the hypothesis (14) which approximates the hypothesis (13) . Let $y_1,...,y_N$ are distributed according to Rayleigh densities. The LR of the test of the hypothesis (14) has the form

$$\lambda_N(y) = \frac{\max_{\sigma_1^2=...=\sigma_N^2} f(y,\sigma^2)}{\max_{approxH_1} f(y,\sigma^2)}.$$

To compute the denominator $\max_{approxH_1} f(y,\sigma^2)$ we proceed like in previous subsection. Suppose that $\{y_{i_1},...,y_{i_K}\}$, $0 < K < N-1$, are the observations from the Rayleigh distribution with the scale parameter σ_1^2, $\{y_{j_1},...,y_{j_L}\}$, $0 < L < N-K$, are the observations from the Rayleigh distribution with the scale parameter σ_2^2 and the other observations are distributed according to the Rayleigh distribution with the scale parameter σ_3^2. For $0 < K < N-1$, $0 < L < N-K$ let $P(K,L)$ denotes the all disjoint pairs of K-subsets $\{i_1,...,i_K\}$ and L-subsets $\{j_1,...,j_L\}$ of the set $\{1,2,...,N\}$. Then the LR of the test of the hypotheses (14) has the form

$$\lambda_N(y) = \min_{0<K<N-1,\ 0<L<N-K,\ p\in P(K)} \{\frac{N^N}{K^K L^L (N-K-L)^{N-K-L}} \times \tag{15}$$

$$\times \frac{(y_{i_1}^2+...+y_{i_K}^2)^K (y_{j_1}^2+...+y_{j_L}^2)^L (y_{l_1}^2+...+y_{l_{N-K-L}}^2)^{N-K-L}}{(y_1^2+...+y_N^2)^N}\}.$$

The main advantages of the test statistic (15) is that under the H_0 it does not depend on the unknown value of the parameter σ^2. The distribution of the LR test statistics $-\ln\lambda_N$ where λ_N is given by the formula (15) under the null hypothesis is derived in the following theorem.

Theorem 5 *Let $y_1, ..., y_N$ are iid according to the Rayleigh distribution with the unknown scale parameter σ^2. Then the LR test statistics $-\ln\lambda_N$ where λ_N is given by the formula (15) has the form*

$$-\ln\lambda_N(y) = -\min_{0<K<N-1,\ 0<L<N-K,\ p\in P(K)}\{N\ln N - K\ln K - L\ln L+$$

$$-(N-K-L)\ln(N-K-L) + K\ln(\sum_{n=1}^{K} y_{i_n}^2) + L\ln(\sum_{n=1}^{L} y_{j_n}^2)+$$

$$+(N-K-L)\ln(\sum_{n=1}^{N-K-L} y_{l_n}^2) - N\ln(\sum_{n=1}^{N} y_n^2)\}$$

and it has the same distribution as the random variable

$$V_N = -\min_{0<K<N-1,\ 0<L<N-K,\ p\in P(K)}\{N\ln N - K\ln K - L\ln L+$$

$$-(N-K-L)\ln(N-K-L) + K\ln(\sum_{n=1}^{K} u_{i_n}) + L\ln(\sum_{n=1}^{L} u_{j_n})+$$

$$+(N-K-L)\ln(\sum_{n=1}^{N-K-L} u_{l_n}) - N\ln(\sum_{n=1}^{N} u_n)\}$$

where $u_1, ..., u_N$ are iid according to $Exp(1)$.

Remark

The main advantage of the provided distribution of the random variable V_N is the possibility of simulation of the density of the LR statistics $-\ln\lambda_N$ based on the $Exp(1)$ simulations.

Proof. Under the H_0 is the sample $\frac{x_1^2}{2\sigma_0^2}, ..., \frac{x_N^2}{2\sigma_0^2}$ iid from the $Exp(1)$ distribution (see(5)). The independence of the LR statistics (15) on the real value of the scale parameter σ^2 under the null hypothesis completes the proof. □

7 Appendix

The Lambert W function is defined to be the multivalued inverse of the complex function $f(y) = ye^y$. As the equation $ye^y = z$ has an infinite number of solutions for each (non-zero) value of $z \in \mathbf{C}$, the Lambert W has an infinite number of branches. Exactly one of these branches is analytic at 0. Usually this branch is referred to as the principal branch of the Lambert W and is denoted by W or W_0. The other branches all have a branch point at 0. These branches are denoted by W_k where $k \in \mathbf{Z} \setminus \{0\}$. The principal branch and the pair of branches W_{-1} and W_1 share an order 2 branch point at $z = -e^{-1}$. A detailed discussion of the branches of the Lambert W can be found in [Corless 96].

Since the Lambert W function has many applications in pure and applied mathematics, the branches of the LW function are implemented to many mathematical computational softwares, e.g. the Maple, Matlab, Mathematica and Mathcad. For more information about the implementation and some computational aspects see [Corless 93].

8 Conclusion

In the present paper we construct the efficient testing procedure of the hypotheses of homogeneity, scale parameter under the homogeneity and the number of components in the Rayleigh mixture. We also discuss the properties of such tests and give the procedure for the computation of the critical values. The test of the homogeneity and scale of the Rayleigh distribution is shown to be asymptotically optimal in the Bahadur sense. The obtained results can be applied to expanding the experimental distribution of transverse momenta into Rayleigh distribution. As we see, our results for $m = 1, 2$ and 3 can be generalized to the case of an auxiliary alternative $m = j$, $j \leq N$. Our next goal is to evaluate estimations for the mixture parameters.

9 Acknowledgements

Research is supported by the VEGA grant (Slovak Grant Agency) No 1/0264/03.

References

[Bahadur 65] Bahadur RR, *An optimal property of the likelihood ratio statistic*, Proc. 5th Berkeley Sympos. on Probab. Theory and Mathem. Statist., vol. 1, eds. L. Le Cam and J. Neyman, Berkeley and Los Angels: Univ. of California Press, 1965, 13-26

[Bahadur 67] Bahadur RR, *Rates of convergence of estimates and test statistics*, Ann. Mathem. Statist. **38**, 1967, 303-324enddocument

[Bahadur 71] Bahadur RR, *Some Limit Theorems in Statistics*, Philadelphia, SIAM.

[Corless 96] Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ and Knuth DE, *On the Lambert W function*, Advances in Computational mathematics **5** (1996), 329-359

[Corless 93] Corless RM, Gonnet GH, Hare DEG, Jeffrey DJ and Knuth DE, *Lambert's W Function in Maple*, Maple Technical Newsletter **9**, Spring 1993, 12-22

[Efimova et al. 89] Efimova TG, Leskin VA, Ososkov GA, Tolstov KD and Chernov NI, *Expansion of Transverse Momenta in Inelastic Collisions of Particles into Rayleigh Distributions*, JINR Rapid Communications No.**3**[36] 1989.

[Lehmann 64] Lehmann EL, *Testing Statistical Hypotheses*, John Wiley & Sons, New York, 1964.

[Nikitin 95] Nikitin Y, *Asymptotic Efficiency of Nonparametric Tests*, Cambridge University Press, New York, 1995, 1-36

[Raghavachari 70] Raghavachari M, *On a theorem of Bahadur on the rate of convergence of test statistics*, Ann. Mathem. Statist. **41**, 1970, 1695-1699

[Rublík 89, 1] Rublík F, *On optimality of the LR tests in the sense of exact slopes*, Part 1, General case. Kybernetika **25**, 1989, 13-25

[Rublík 89, 2] Rublík F, *On optimality of the LR tests in the sense of exact slopes*, Part 2, Application to Individual Distributions. Kybernetika **25**, 1989, 117-135

[Stehlík 03] Stehlík M, *Distributions of exact tests in the exponential family*, Metrika **57**(2003), Springer-Verlag, Heidelberg, 145-164

[Susko 03] Susko E, *Weighted tests of homogeneity for testing the number of components in a mixture*, Computational Statistics and Data Analysis **41**(2003), 367-378.

[Wilks 67] Wilks SS, *Mathematical Statistics*, Nauka, Moscow, 1967, (Russian), 410-419

Received on June 20, 2003.

Стеглик М., Ососков Г. А. E11-2003-116
Эффективные критерии проверки однородности, параметров масштаба и числа компонентов смеси распределений Рэлея

Рассматривается статистическая проблема представления экспериментального распределения поперечных моментов в виде смеси рэлеевских распределений. Разработан высокоэффективный критерий для проверки гипотезы однородности выборки, оптимальный в смысле Бахадура. Для случаев выполнения гипотезы предложен точный критерий отношения правдоподобия для параметра масштаба. Для иных случаев предложен эффективный тест для числа компонентов смеси.

Работа выполнена в Лаборатории информационных технологий ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна, 2003

Stehlík M., Ososkov G. A. E11-2003-116
Efficient Testing of the Homogeneity, Scale Parameters and Number of Components in the Rayleigh Mixture

The statistical problem to expand the experimental distribution of transverse momenta into Rayleigh distribution is considered. A high-efficient testing procedure for testing the hypothesis of the homogeneity of the observed measurements which is optimal in the sense of Bahadur is constructed. The exact likelihood ratio (LR) test of the scale parameter of the Rayleigh distribution is proposed for cases when the hypothesis of homogeneity holds. Otherwise the efficient procedure for testing the number of components in the mixture is also proposed.

The investigation has been performed at the Laboratory of Information Technologies, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna, 2003

Макет *Т. Е. Попеко*

Подписано в печать 07.07.2003.
Формат 60 × 90/16. Бумага офсетная. Печать офсетная.
Усл. печ. л. 1,18. Уч.-изд. л. 1,79. Тираж 310 экз. Заказ № 54002.

Издательский отдел Объединенного института ядерных исследований
141980, г. Дубна, Московская обл., ул. Жолио-Кюри, 6.
E-mail: publish@pds.jinr.ru
www.jinr.ru/publish/