AUSTRIAN JOURNAL OF STATISTICS
Volume 34 (2005), Number 2, 225–232

Notes on Statistical Tests for Nonidentically Distributed Observations

Andrew M. Zubkov[1]
Steklov Mathematical Institute of RAS, Moscow, Russia

Abstract: We discuss a problem of testing simple statistical hypotheses on the nonidentically distributed observations. A possibility of using the method Monte-Carlo for choosing the critical value in Neyman – Pearson test is pointed out. We propose also an approach permitting to reduce the problem considered to a problem of testing hypotheses on uniformity vs nonuniformity of distributions on the interval (0,1).

Keywords: Hypotheses Testing, Nonhomogeneous Observations, Likelihood Ratio Ordering, Variational Distance.

1 Introduction

Let us consider a classical problem: we observe a sequence $\xi_1, \dots, \xi_n$ of random variables (the structure of the set of their values is unessential) and wish to test two statistical hypotheses (H_0 and H_1) on the distributions of these random variables. According to the Neyman – Pearson fundamental lemma (see Lehmann, 1986) the most powerful test in this problem should be based on the likelihood ratio

$$L_n = L_n(\xi_1, \dots, \xi_n) = \frac{p_0(\xi_1, \dots, \xi_n)}{p_1(\xi_1, \dots, \xi_n)},$$

where p_0 and p_1 are densities (probabilities in discrete case) of the outcome $(\xi_1, \dots, \xi_n)$. Given the probability α of choosing H_1 when H_0 is valid we have to find a number $c(\alpha)$ such that $\mathbf{P}\{L_n \le c(\alpha) \mid H_0\} = \alpha$ (with possible randomization if the distribution of L_n has an atom at $c(\alpha)$). If such an $c(\alpha)$ is found (and randomization is unnecessary) we may choose the hypothesis H_0 if $L_n > c(\alpha)$ and choose H_1 in the opposite case. This criterion has minimal probability of choosing H_0 when H_1 is valid in the set of all criteria such that the probability of choosing H_1 when H_0 is valid equals to α.

From a practical viewpoint a serious drawback of Neyman – Pearson criterion is the necessity to compute $c(\alpha)$. This problem has an exact solution if the equation for $c(\alpha)$ is solvable. In the case of independent identically distributed observations (with known mean and variance of $\log L_1$) and large n it is possible to use normal approximation for the distribution of $\log L_n$. But just in the simple case when $\xi_1, \dots, \xi_n$ are independent but nonidentically distributed (for example, if $\xi_1, \dots, \xi_n$ correspond to different characteristics of random objects) the computation of $c(\alpha)$ may become a hard problem.

This note consists of two parts. The content of the first part is an almost trivial idea to use the Monte-Carlo method to estimate $c(\alpha)$. In the second part we suggest a hint permitting to reduce the statistical problem with nonidentically distributed observations to a problem of testing hypothesis on uniformity vs nonuniformity of distributions on the interval (0,1).

[1]This research was supported by the Leading Scientific Schools Support Fund of the President of Russia Federation, grant 1758.2003.1, and by the Program of RAS "Modern Problems of Theoretical Mathematics".

2 Monte-Carlo Estimates

First of all note that it is possible to use Monte-Carlo approximation to $c(\alpha)$. Indeed, let us simulate N independent samples $(\eta_1^{(j)}, \dots, \eta_n^{(j)})$, $j = 1, \dots, N$, according to the hypothesis H_0. For each $j = 1, \dots, N$ we compute $\lambda_j = L_n(\eta_1^{(j)}, \dots, \eta_n^{(j)})$. These random variables are independent and identically distributed:

$$\mathbf{P}\{\lambda_j \le x\} = F_0(x) := \mathbf{P}\{L_n \le x \,|\, H_0\}.$$

So, $\mathbf{P}\{F_0(\lambda_j) < y\} \le y$ for all $y \in [0,1]$, and $\mathbf{P}\{F_0(\lambda_j) < y\} = y$ if $F_0^{-1}(y) := \sup\{x\colon F_0(x) < y\}$ is a continuity point of F_0. For large N the empirical distribution function of $\lambda_1, \dots, \lambda_N$ will approximate the distribution function $F_0(x)$. For example, let $\lambda_{1:N} \le \lambda_{2:N} \le \dots \le \lambda_{N:N}$ be the order statistics of $\lambda_1, \dots, \lambda_N$. If we choose $\lambda_{[\alpha N]:N}$ ($[x]$ denotes an integer part of x) as an approximation to $c(\alpha)$ then

$$\mathbf{P}\{F_0(\lambda_{[\alpha N]:N}) < \alpha - \varepsilon\} = \mathbf{P}\left\{\sum_{j=1}^{N} I(F_0(\lambda_j) < \alpha - \varepsilon) \ge [\alpha N]\right\} \le \mathbf{P}\{\beta_{N,\alpha-\varepsilon} \ge \alpha N\},$$

$$\mathbf{P}\{F_0(\lambda_{[\alpha N]:N}) > \alpha + \varepsilon\} = \mathbf{P}\left\{\sum_{j=1}^{N} I(F_0(\lambda_j) \le \alpha + \varepsilon) < [\alpha N]\right\} \le \mathbf{P}\{\beta_{N,\alpha+\varepsilon} < \alpha N\},$$

where $\beta_{N,p}$ denotes random variable having the binomial distribution with parameters (N, p). These equations may be used to estimate the accuracy of approximation of $c(\alpha)$ as a function of α and N.

The same idea of using the Monte-Carlo method may be applied not *before*, but *after* observing the sequence $\xi_1, \dots, \xi_n$. Let us simulate N independent samples

$$(\eta_1^{(j)}, \dots, \eta_n^{(j)}), \quad j = 1, \dots, N,$$

according to the hypothesis H_0 and compute the number ν_N of $j = 1, \dots, N$ such that $\lambda_j \le L_n(\xi_1, \dots, \xi_n)$. Then the value $\frac{\nu_N}{N}$ is a consistent (as $N \to \infty$) statistical estimate of $F_0(L_n(\xi_1, \dots, \xi_n))$. So the inequality $\frac{\nu_N}{N} \ge \alpha$ is asymptotically (as $N \to \infty$) equivalent to the Neyman – Pearson criterion. It may be shown that

$$-\max_{0 \le y \le 1} \mathbf{P}\{F_0(L_n) = y\} \le \mathbf{P}\{\nu_N \le m\} - \frac{m+1}{N+1} \le 0, \qquad m = 1, \dots, N.$$

Computation of estimates $\frac{\nu_N}{N}$ for each observation of $\xi_1, \dots, \xi_n$ takes more time but it rules out inevitable systematic bias appearing when a single Monte-Carlo estimate of $c(\alpha)$ is used.

3 Reduction to a Simpler Problem

Consider another possibility to test hypotheses H_0 and H_1 on independent nonidentically distributed random variables $\xi_1, \dots, \xi_n$.

We begin with the case when for each $j = 1, \dots, N$ the distributions of ξ_j under both H_0 and H_1 are absolutely continuous with common support. Let $p_{i,j}(x)$ be a density of ξ_j

under hypothesis $H_i,\ i=0,1,\ j=1,\dots,N$. Denote $r_j(x)=\frac{p_{0,j}(x)}{p_{1,j}(x)},\ j=1,\dots,N$, and introduce distribution functions

$$G_j(x)=\mathbf{P}\{r_j(\xi_j)\le x|H_0\}=\int\limits_{u:\,r_j(u)\le x} p_{0,j}(u)du,\ j=1,\dots,N. \tag{1}$$

Definition of $G_j(x)$ seems complex, but it may be simplified in concrete cases. For example, if $r_j(x)$ is monotonically decreasing and $r_j^{-1}(y)=\sup\{x\colon r_j(x)\ge y\}$ then

$$G_j(x)=\int\limits_{r_j^{-1}(x)}^{\infty} p_{0,j}(u)du,\ j=1,\dots,N;$$

if $r_j(x)$ is unimodal (i.e. $r_j(x)$ is increasing from $r_j(-\infty)$ to $r_j(a_j)$ on $[-\infty,a_j]$ and decreasing from $r_j(a_j)$ to $r_j(\infty)$ on $[a_j,\infty)$) and

$$r_j^+(y)=\{\sup\{x\colon r_j(x)\ge y\},\quad r_j(\infty)<y\le r_j(a_j),\infty,\quad r_j(0)\le y\le r_j(\infty),$$

$$r_j^-(y)=\{\inf\{x\colon r_j(x)\ge y\},\quad r_j(0)\le y\le r_j(a_j),\ -\infty,\quad r_j(\infty)<y\le r_j(a_j),$$

then

$$G_j(x)=1-\int\limits_{r_j^-(x)}^{r_j^+(x)} p_{0,j}(u)du,\quad \min\{r_j(0),r_j(\infty)\}\le x\le r_j(a_j),\quad j=1,\dots,N.$$

Let $\zeta_j=G_j(r_j(\xi_j)),\ j=1,\dots,N$. If distributions of $\xi_1,\dots,\xi_n$ satisfy the hypothesis H_0 then random variables $\zeta_1,\dots,\zeta_N$ are independent and uniformly distributed on [0,1]:

$$Z_0(x):=\mathbf{P}\{\zeta_j\le x|H_0\}=\mathbf{P}\{G_j(r_j(\xi_j))\le x|H_0\}= \tag{2}$$
$$=\mathbf{P}\{r_j(\xi_j)\le G_j^{-1}(x)|H_0\}=G_j(G_j^{-1}(x))=x,\quad x\in[0,1],\ j=1,\dots,N,$$

where G^{-1} denotes inverse function for G.

If hypothesis H_1 is valid then random variables $\zeta_1,\dots,\zeta_N$ are independent and have nonuniform distributions. The distribution function of ζ_j under the hypothesis H_1 takes the form

$$Z_{1,j}(x):=\mathbf{P}\{\zeta_j\le x|H_1\}=\mathbf{P}\{G_j(r_j(\xi_j))\le x|H_1\}=\mathbf{P}\{r_j(\xi_j)\le G_j^{-1}(x)|H_1\}. \tag{3}$$

Theorem 1. *For any $j=1,\dots,N$ the function $Z_{1,j}(x)$ is concave on $[0,1]$ and $\rho_j=\max\limits_{0\le x\le 1}(Z_{1,j}(x)-Z_0(x))$ is equal to the variational distance between distributions of ξ_j under hypotheses H_0 and H_1. Further,*

$$\mathbf{E}\{\zeta_j|H_0\}=\frac{1}{2},\quad \mathbf{E}\{\zeta_j|H_1\}\le\frac{1-\rho_j}{2},$$

$$\mathbf{D}\{\zeta_j|H_0\}=\frac{1}{12},\quad \mathbf{D}\{\zeta_j|H_1\}\le\frac{1}{4}.$$

Corollary. *We have* $Z_{1,j}(x) > Z_0(x) = x$ *for all* $x \in (0,1)$ *and all* $j = 1, \ldots, N$.

PROOF. Let

$$G_{1,j}(x) = \mathbf{P}\{r_j(\xi_j) \le x | H_1\} = \int\limits_{u:\, r_j(u) \le x} p_{1,j}(u)du,\ j = 1, \ldots, N. \tag{4}$$

Then $Z_{1,j}(x) = G_{1,j}(G_j^{-1}(x))$. So,

$$\frac{d}{dx} Z_{1,j}(x) = \frac{d}{dx} G_{1,j}(G_j^{-1}(x)) = \frac{d}{dv} G_{1,j}(v)\Big|_{v=G_j^{-1}(x)} \frac{d}{dx} G_j^{-1}(x) = \tag{5}$$

$$= \frac{\frac{d}{dv} G_{1,j}(v)\Big|_{v=G_j^{-1}(x)}}{\frac{d}{dv} G_j(v)\Big|_{v=G_j^{-1}(x)}} = \frac{\lim\limits_{\Delta \downarrow 0} \frac{1}{\Delta} \int\limits_{u:\, G_j^{-1}(x)-\Delta \le r_j(u) \le G_j^{-1}(x)} p_{1,j}(u)du}{\lim\limits_{\Delta \downarrow 0} \frac{1}{\Delta} \int\limits_{u:\, G_j^{-1}(x)-\Delta \le r_j(u) \le G_j^{-1}(x)} p_{0,j}(u)du} = \frac{1}{G_j^{-1}(x)}$$

is nonincreasing, i.e. $Z_{1,j}(x)$ is concave.

Note that distribution functions $Z_0(x) = x$, $Z_{1,j}(x)$, $x \in [0,1]$, are absolutely continuous with densities $z_0(x) = 1$, $z_{1,j}(x) = \frac{1}{G_j^{-1}(x)}$, $x \in [0,1]$, and satisfy conditions $Z_0(0) = Z_{1,j}(0) = 0$, $Z_0(1) = Z_{1,j}(1) = 1$. Let $c_j = \sup\{x \in [0,1]\colon z_{1,j}(x) \ge 1\}$. Due to monotonicity of $z_{1,j}(x)$ we have

$$\rho_j = \max_{0 \le x \le 1} (Z_{1,j}(x) - Z_0(x)) = Z_{1,j}(c_j) - Z_0(c_j) =$$

$$= \frac{1}{2} \left(\int\limits_0^{c_j} (z_{1,j}(x) - 1)dx + \int\limits_{c_j}^1 (1 - z_{1,j}(x))dx \right) = \frac{1}{2} \int\limits_0^1 |z_{1,j}(x) - 1|dx.$$

Further, taking into account (5) and applying change of variable $x = G_j(y)$ we find

$$\int\limits_0^1 |z_{1,j}(x) - 1|dx = \int\limits_0^1 \left| \frac{1}{G_j^{-1}(x)} - 1 \right| dx = \int\limits_0^1 \left| \frac{\frac{d}{dv} G_{1,j}(v)\Big|_{v=G_j^{-1}(x)}}{\frac{d}{dv} G_j(v)\Big|_{v=G_j^{-1}(x)}} - 1 \right| dx =$$

$$= \int\limits_{-\infty}^{\infty} \left| \frac{\frac{d}{dy} G_{1,j}(y)}{\frac{d}{dy} G_j(y)} - 1 \right| dG_j(y) = \int\limits_{-\infty}^{\infty} \left| \frac{d}{dy} G_{1,j}(y) - \frac{d}{dy} G_j(y) \right| dy =$$

$$= \int\limits_{-\infty}^{\infty} \left| \frac{d}{dy} \int\limits_{u:\, r_j(u) \le y} p_{1,j}(u)du - \frac{d}{dy} \int\limits_{u:\, r_j(u) \le y} p_{0,j}(u)du \right| dy = \int |p_{1,j}(u) - p_{0,j}(u)|du,$$

i.e., ρ_j equals to the variational distance between distributions of ξ_j under hypotheses H_0 and H_1.

Random variable ζ_j under hypothesis H_0 is uniformly distributed on $[0,1]$, so

$$\mathbf{E}\{\zeta_j | H_0\} = \frac{1}{2}, \quad \mathbf{D}\{\zeta_j | H_0\} = \frac{1}{12}.$$

To estimate $\mathbf{E}\{\zeta_j|H_1\}$ we introduce piecewise linear function $L_j(x)$: the graph of $L_j(x)$ connects points $(0,0)$, $(c_j, Z_{1,j}(c_j))$, $(1,1)$. Due to the concavity of $Z_{1,j}(x)$ we have $Z_{1,j}(x) \geq L_j(x)$ for all $x \in [0,1]$. Now

$$\mathbf{E}\{\zeta_j|H_1\} = \int_0^1 (1 - Z_{1,j}(x))dx \leq \int_0^1 (1 - L_j(x))dx =$$

$$= \int_0^1 (1-x)dx - \int_0^1 (L_j(x) - x)dx = \frac{1}{2} - \frac{\rho_j}{2}$$

because the area of triangle formed by graphs of $L_j(x)$ and $Z_0(x) = x$ equals to

$$\frac{1}{2}(Z_{1,j}(c_j) - c_j) = \frac{\rho_j}{2}.$$

Inequality $\mathbf{D}\{\zeta_j|H_1\} \leq \frac{1}{4}$ is valid for any random variable with values in $[0,1]$. The theorem is proved.

Now consider the case when distributions of ξ_j under both H_0 and H_1 are discrete with common support $T_j,\ j = 1, \ldots, N$. Let

$$P_{i,j}(t) = \mathbf{P}\{\xi_j = t \,|\, H_i\}, \quad t \in T_j = \{t_{j,1}, t_{j,2}, \ldots\}, \quad j = 1, \ldots, N,$$

be the distribution of ξ_j under hypothesis $H_i,\ i = 0, 1$. The sets $T_1, \ldots, T_N$ are at most countable. Let $R_j(t) = \frac{P_{0,j}(t)}{P_{1,j}(t)},\ t \in T_j$; the sets R_j of values $R_j(t),\ t \in T_j$, are at most countable also.

It is convenient to introduce right-continuous and left-continuous distribution functions for each $j = 1, \ldots, N$:

$$S_{0,j}(x) = \mathbf{P}\{R_j(\xi_j) \leq x \,|\, H_0\} = \sum_{t \in T_j : R_j(t) \leq x} P_{0,j}(t), \tag{6}$$

$$S^-_{0,j}(x) = \mathbf{P}\{R_j(\xi_j) < x \,|\, H_0\} = \sum_{t \in T_j : R_j(t) < x} P_{0,j}(t), \tag{7}$$

$$S_{1,j}(x) = \mathbf{P}\{R_j(\xi_j) \leq x \,|\, H_1\} = \sum_{t \in T_j : R_j(t) \leq x} P_{1,j}(t), \tag{8}$$

$$S^-_{1,j}(x) = \mathbf{P}\{R_j(\xi_j) < x \,|\, H_1\} = \sum_{t \in T_j : R_j(t) < x} P_{1,j}(t), \tag{9}$$

here $x \in [0, \infty]$. Let S_j be a set of points with coordinates $(S_{0,j}(r), S_{1,j}(r)), r \in R_j$.

Lemma. *For each $j = 1, \ldots, N$ the set S_j is lying on a concave curve.*

PROOF. It is sufficient to prove that for any $r_1 < r_2 < r_3\, (r_1, r_2, r_3 \in R_j)$ the slope of a chord $[(S_{0,j}(r_1), S_{1,j}(r_1)), (S_{0,j}(r_2), S_{1,j}(r_2))]$ is greater than the slope of a chord $[(S_{0,j}(r_2), S_{1,j}(r_2)), (S_{0,j}(r_3), S_{1,j}(r_3))]$, i.e. that

$$\frac{S_{1,j}(r_2) - S_{1,j}(r_1)}{S_{0,j}(r_2) - S_{0,j}(r_1)} > \frac{S_{1,j}(r_3) - S_{1,j}(r_2)}{S_{0,j}(r_3) - S_{0,j}(r_2)}.$$

But according to definitions (6) – (9) we have

$$\frac{S_{1,j}(r_2)-S_{1,j}(r_1)}{S_{0,j}(r_2)-S_{0,j}(r_1)}=\frac{\sum\limits_{t\in T_j: r_1<R_j(t)\le r_2} P_{1,j}(t)}{\sum\limits_{t\in T_j: r_1<R_j(t)\le r_2} P_{0,j}(t)}\ge\frac{1}{r_2}>$$

$$>\frac{\sum\limits_{t\in T_j: r_2<R_j(t)\le r_3} P_{1,j}(t)}{\sum\limits_{t\in T_j: r_2<R_j(t)\le r_3} P_{0,j}(t)}=\frac{S_{1,j}(r_3)-S_{1,j}(r_2)}{S_{0,j}(r_3)-S_{0,j}(r_2)},$$

and Lemma is proved.

The proof of Lemma is applicable to the absolutely continuous case also, but in that case we have used an explicit formula for the density of $Z_{1,j}(x)$. To get a full analogy with the absolutely continuous case we define functions $Z^*_{1,j}(x), x\in[0,1], j=1,\dots,N$, such that their graphs are convex hulls of corresponding sets S_j. In other words, these functions are continuous, satisfy conditions

$$Z^*_{1,j}(S_{0,j}(r))=S_{1,j}(r),\quad r\in R_j,\quad Z^*_{1,j}(0)=0,\quad Z^*_{1,j}(1)=1,$$

and are piecewise linear on intervals $(S^-_{0,j}(r),S_{0,j}(r)),\ r\in R_j$, between points of the set $\bar S_j=\{0\}\cup\{S_{0,j}(r),\ r\in R_j\}\cup\{1\}$. Evidently, $Z^*_{1,j}$ for each $j=1,\dots,N$ is a distribution function of the mixture of uniform distributions on intervals $(S^-_{0,j}(r),S_{0,j}(r)),\ r\in\bar S_j$, with weights $S_{1,j}(r)-S^-_{1,j}(r)$.

It is obvious that mixture of uniform distributions on intervals $(S^-_{0,j}(r),S_{0,j}(r)),\ r\in\bar S_j$, with weights $S_{0,j}(r)-S^-_{0,j}(r)$ is a uniform distribution on $[0,1]$ and has distribution function $Z^*_{0,j}(x)=x,\ x\in[0,1]$.

Now to define randomized statistics $\zeta_j=\zeta_j(\xi_j)$ we introduce auxiliary random variables $\alpha_1,\dots,\alpha_N$ uniformly distributed on an interval $[0,1]$ and put

$$\zeta_j=\alpha_jS^-_{0,j}(r)+(1-\alpha_j)S_{0,j}(r)\quad\text{if}\quad R_j(\xi_j)=r.$$

Being defined in such a way random variable ζ_j is uniformly distributed on $[0,1]$ under hypothesis H_0 and has distribution function $Z^*_{1,j}(x)$ under hypothesis H_1.

Theorem 2. *For any $j=1,\dots,N$ the function $Z^*_{1,j}(x)$ is concave on $[0,1]$ and $\rho_j=\max\limits_{0\le x\le1}(Z^*_{1,j}(x)-Z^*_{0,j}(x))$ is equal to the variational distance between distributions of ξ_j under hypotheses H_0 and H_1. Further,*

$$\mathbf{E}\{\zeta_j|H_0\}=\frac{1}{2},\quad \mathbf{E}\{\zeta_j|H_1\}\le\frac{1-\rho_j}{2},$$

$$\mathbf{D}\{\zeta_j|H_0\}=\frac{1}{12},\quad \mathbf{D}\{\zeta_j|H_1\}\le\frac{1}{4}.$$

PROOF. Concavity of the function $Z^*_{1,j}(x)$ follows from Lemma. The function $Z^*_{1,j}(x)$ is linear on intervals $(S^-_{0,j}(r),S_{0,j}(r)), r\in R_j$, so $\max\limits_{0\le x\le1}(Z^*_{1,j}(x)-Z^*_{0,j}(x))$ is attained on the closure of the set $\{0\}\cup\{S_{0,j}(r), r\in R_j\}$. Further, for any $r\in R_j$

$$Z^*_{1,j}(S_{0,j}(r))-Z^*_{0,j}(S_{0,j}(r))=S_{1,j}(r)-S_{0,j}(r)=$$

$$= \sum_{t\in T_j: R_j(t)\le r} P_{1,j}(t) - \sum_{t\in T_j: R_j(t)\le r} P_{0,j}(t) = \sum_{t\in T_j: P_{0,j}(t)\le r P_{1,j}(t)} (P_{1,j}(t) - P_{0,j}(t)).$$

The summands in the last sum are positive for $r < 1$ and negative for $r > 1$; it follows that $\max\limits_{0\le x\le 1}(Z^*_{1,j}(x) - Z^*_{0,j}(x))$ is attained at the point $x = \sup\{S_{0,j}(r)\colon r \le 1\}$ and is equal to

$$\frac{1}{2}\sum_{t\in T_j} |P_{1,j}(t) - P_{0,j}(t)|,$$

i.e. to the variational distance between distributions of ξ_j under hypotheses H_0 and H_1.

Moments of ζ_j are estimated as in Theorem 1. Theorem 2 is proved.

Theorems 1 and 2 reduce the problem of testing two simple statistical hypotheses on the distribution of independent nonidentically distributed random variables $\xi_1, \ldots, \xi_N$ (taking values of any nature) to the problem of testing hypothesis H'_0: independent random variables $\zeta_1(\xi_1), \ldots, \zeta_N(\xi_n)$ are uniformly distributed on $[0, 1]$ against hypothesis H'_1: independent random variables $\zeta_1(\xi_1), \ldots, \zeta_N(\xi_n)$ taking values in $[0, 1]$ have concave nonuniform distribution functions $Z_{1,j}(x), j = 1, \ldots, N$. So under H'_0 random variables $\zeta_1(\xi_1), \ldots, \zeta_N(\xi_n)$ are identically distributed and under H'_1 random variables $\zeta_1(\xi_1), \ldots, \zeta_N(\xi_n)$ have biases of the same sign compared to their distributions under H'_0.

The simplest way to test H'_0 against H'_1 by means of these biases is to use statistics $V_N = \sum\limits_{j=1}^{N} \zeta_j$. If H_0 is valid then V_N is the sum of N independent random variables uniformly distributed on the interval $[0, 1]$, in particular,

$$\mathbf{E}\{V_N|H_0\} = \frac{N}{2}\,, \quad \mathbf{D}\{V_N|H_0\} = \frac{N}{12}\,.$$

If H_1 is valid then V_N is the sum of N independent random variables which are stochastically smaller than random variables uniformly distributed on the interval $[0, 1]$, and

$$\mathbf{E}\{V_N|H_1\} \le \sum_{j=1}^{N} \frac{1-\rho_j}{2}\,, \quad \mathbf{D}\{V_N|H_1\} = \sum_{j=1}^{N} \mathbf{D}\{\zeta_j|H_1\} \le \frac{N}{4}\,.$$

For large N normal approximations may be used to estimate critical levels and error probabilities (especially if explicit formulas for $\mathbf{E}\{V_N|H_1\}$ and $\mathbf{D}\{V_N|H_1\}$ are known). But this approach, of course, is not the best possible.

Another way is to compute statistics of goodness-of-fit criteria (for example, Kolmogorov – Smirnov statistics) for H'_0 and H'_1 separately and compare their values.

However it seems that the problem of testing H'_0 against H'_1 is far from being completely solved.

References

E.L. Lehmann. *Testing Statistical Hypotheses*. Springer-Verlag, New York e.a., 2nd edition, 1986.

Author's address:

Dr. Andrew M. Zubkov
Department of Discrete Mathematics
Steklov Mathematical Institute
Gubkina Str. 8
119991 Moscow
Russia

Tel. +7 095 1351519
Fax +7 095 1350555
E-mail: zubkov@mi.ras.ru