GOODNESS-OF-FIT AND SYMMETRY TESTS BASED ON CHARACTERIZATIONS.

Yakov Nikitin

St.Petersburg State University, Russia

2008

1 Introduction

Coinsider the sample $X_1, \ldots, X_n$ with continuous df F. Two classical problems of Statistics are:

A. **Goodness-of-fit problem.**

We should test the hypothesis $H_0 : F = F_0$ against some alternative $A_0 : F \neq F_0$. More general is the testing problem $F \in \mathcal{F}$, where $\mathcal{F}$ is some parametric family of distributions, say, normal or exponential family. The alternative is $A : F \notin \mathcal{F}$ or more narrow hypothesis.

B. **Symmetry hypothesis.**

We test the hypothesis $H_1 : F \in \mathfrak{F}$, where $\mathfrak{F}$ is the set of all symmetric continuous df's,

$$\mathfrak{F} = \{F : 1 - F(x) - F(-x) = 0, \forall\, x.\}$$

We can take as alternative the hypothesis A_1 that F is non-symmetric (shift, skew, contamination, etc.)

There are numerous tests for these problems: the sign test (1702), the Pearson's χ^2-test (1900), the Kolmogorov-Smirnov test (1933), the Cramér - von Mises - Smirnov test (1937), the Wilcoxon test (1945), the Anderson-Darling test (1952), the Watson test (1961), the Bickel-Rosenblatt test (1973), etc., if we mention only most famous tests.

Observe that all these tests were proposed from certain *empirical point of view,* all of them are based on some "heuristic" idea, and only later came their deeper analysis and study.

However, no one of these tests dominate the others in sense of power, and the ordering of tests can be different when changing the alternative. For large samples the notion of asymptotic efficiency is often used.

It is generally recognized that any test should be analyzed from the point of view of its power and efficiency

in order to give the recommendations for practitioners.

We shall present in this talk some tests mainly based on the idea of characterization of distributions by the property of "equidistribution"of statistics. The idea ascends to the paper of famous Russian mathematician Yu.V.Linnik "Linear forms and statistical tests", published in 1953.

It seems that mathematical technique, especially connected with $U-$statistics and their large deviations, was developed not enough at this time. That's why the realization of Linnik ideas became possible only recently.

Characterization of distributions began in 1923 by celebrated Polya's theorem.

Polya's Theorem. *Let X and Y be two independent and identically distributed rv's with zero mean. Then two rv's (statistics) $(X+Y)/\sqrt{2}$ and X have the same distribution if and only if the distribution of X and Y is normal.*

We cite another simple result of this kind which belongs to Desu (1973).

Desu's Theorem. *Let X and Y be independent non-negative rv's with common df F. Then two rv's (statistics) $2\min(X_1, X_2)$ and X_1 are equidistributed iff $F(x) = 1 - \exp(-\lambda x), x \geq 0$ for some $\lambda > 0$.*

Later many other results were obtained, mainly in 1960-70-s, some of them will be used later.

How to use these results for testing?

Consider the sample $X_1, \ldots, X_n$ with unknown df $F(t) = \mathbb{P}(X_1 < t)$. The classical and well-known estimate of F is the empirical distribution function (Cramér, 1928)

$$F_n(t) = n^{-1} \sum_{j=1}^{n} \mathbf{1}\{X_j < t\},\ t \in R^1.$$

Now suppose we have a function of observations $h(X_1, \ldots, X_k)$ where k is not large and we want to estimate its df

$$H(t) = \mathbb{P}(h(X_1, \ldots, X_k) < t).$$

The natural estimator of $G(t)$ is the $U-$statistical df

$$H_n(t) = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \ldots < i_k \le n} \mathbf{1}\{h(X_{i_1}, ..., X_{i_k}) < t\}.$$

Sometimes we can use the similar $V-$statistical edf:

$$G_n(t) = n^{-k} \sum_{i_1, \ldots, i_k = 1}^{n} \mathbf{1}\{h(X_{i_1}, ..., X_{i_k}) < t\}.$$

Due to the versions of Glivenko-Cantelli theorem obtained around 1980 both empirical df's converge uniformly to $H(t) = \mathbb{P}(h(X_1, \ldots, X_k) < t)$.

Now return to the Polya characterization. Under condition of normality we have

$$\mathbb{P}(X_1 < t) = \mathbb{P}((X_1 + X_2 < t\sqrt{2}), \quad t \in R^1.$$

Hence also the empirical df $F_n(t)$ and $U-$statistical edf

$$H_n(t) = \binom{n}{2}^{-1} \sum_{1 \le i < j \le n} \mathbf{1}\{X_i + X_j < t\}$$

should be close for all t. The significance test can be based on some functional of the difference $F_n(t) - H_n(t)$, for instance on

$$T_{1n} = \int_{R^1} [H_n(t) - F_n(t)] dF_n(t),$$

$$T_{2n} = \int_{R^1} [H_n(t) - F_n(t)]^2 dF_n(t),$$

and

$$T_{\infty n} = \sup_{R^1} |H_n(t) - F_n(t)|.$$

Large values of these statistics are critical.

Similar statistics can be introduced for testing of exponentiality using Desu's characterization, etc. There are many other characterizations which lead to corresponding tests. Consider some examples.

1) *Baringhaus-Henze characterization of symmetry.*

Baringhaus and Henze (1991) proposed the following characterization of symmetry with respect to zero for testing.

Let X and Y be i.i.d. rv's with common continuous df. Then $|X|$ and $|\max(X, Y)|$ are equidistributed iff X and Y are symmetric.

We can use this characterization in the following way.

Let introduce the empirical df based on $|X_1|, ..., |X_n|$,

$$Q_n(t) = n^{-1} \sum_{i=1}^{n} \mathbf{1}\{|X_i| < t\},$$

and the U-statistical empirical df

$$H_n(t) = \binom{n}{2}^{-1} \sum_{1 \leqslant j < k \leqslant n} \mathbf{1}\{|\max(X_j, X_k)| < t\}.$$

Now we propose the statistics:

$$S_{n,1} = \int_0^\infty [H_n(t) - Q_n(t)] dQ_n(t),$$

$$S_{n,2} = \int_0^\infty [H_n(t) - Q_n(t)]^2 dQ_n(t)$$

and

$$S_{n,\infty} = \sup_{R^+} |H_n(t) - Q_n(t)|.$$

2) *Rao-Ramachandran characterization of Cauchy Law.*

In 1970 Rao and Ramachandran proved the following theorem.

Let $X_1, ..., X_m$ – *be nondegenerate i.i.d. rv's, while constants* $a_1, ..., a_m$ *satisfy the conditions* $0 < |a_i| < 1$, $\sum_{i=1}^{m} |a_i| = 1$, *and at least two numbers of* $-\ln |a_1|, ..., -\ln |a_m|$ *are incommensurable. Then two statistics* X_1 *and* $\sum_{i=1}^{m} a_i X_i$ *are equidistributed iff* X_i *have symmetric Cauchy distribution with arbitrary scale factor.*

We can use this characterization to construct goodness-of-fit tests in the same way as above. The practical choice is $m = 2, a_1 = \frac{2}{3},\ a_2 = -\frac{1}{3}$.

3).*Lack of memory (memory-loss) property of the exponential law.*

Let F be a df of a non-negative random variable X. Put $\bar{F}(x) = 1 - F(x), x \geq 0$. Consider the well-known *memory-loss property*:

$$\bar{F}(x + y) = \bar{F}(x)\bar{F}(y), \quad \forall\, x, y \geq 0.$$

This equation (Cauchy functional equation) characterizes the exponential distribution.

Replacing F by empirical distribution function, we arrive to some empirical field

$$\xi_n(x,y) = (1 - F_n(x+y)) - (1 - F_n(x)(1 - F_n(y)),$$

and suitable functionals of this field may serve as significance tests for the exponentiality hypothesis.

There are many other characterizations, for instance, by independence of two statistics. Most known is possibly the Geary-Lukacs characterization: *independence of* $\bar{x}$ *and* s^2 *implies normality.*

Another famous theorem of this kind is the Darmois-Skitovich theorem:*independence of non-degenerate linear forms* $\sum_i a_i X_i$ *and* $\sum_j b_j X_j$ *also implies normality.* Corresponding goodness-of-fit tests are almost unexplored.

So we have numerous new test statistics. Natural questions appear:

I How can we find their limiting distributions?

II Are these tests consistent against common alternatives?

III How to compare these tests and which is better?

(Usually the answer to III can be given using the notion of *asymptotic efficiency*.)

These questions are difficult and were studied partially only in last years.

Let return to integral and Kolmogorov-Smirnov-type statistics we introduced above:

$$T_{1n} = \sqrt{n}\int_{R^1}[H_n(t) - F_n(t)]dF_n(t),$$

$$T_{2n} = n\int_{R^1}[H_n(t) - F_n(t)]^2 dF_n(t),$$

and

$$T_{\infty n} = \sqrt{n}\sup_{R^1}|H_n(t) - F_n(t)|.$$

I **Limiting distributions**

First of them is in fact the *non-degenerate* $U-$ or $V-$statistic. Hence, by Hoeffding theorem (1948), it is asymptotically normal.

For instance, in the case of Polya test for normality we obtain

$$L_n = n^{-3}\sum_{i,j,k=1}^{n}[\mathbf{1}\{X_i + X_j < X_k\sqrt{2}\} - \frac{1}{2}].$$

There are some difficulties when calculating the asymptotic variance Δ^2. For the Polya test we get after long calculations

$$\Delta^2 = \frac{117}{108} - \frac{4}{\pi}(\arctan\sqrt{\frac{3}{5}} + \frac{9}{2}\arctan\frac{1}{\sqrt{7}}) \approx 1.414\cdot 10^{-2} > 0.$$

(Muliere and Nikitin, Metron, 2002).

The second statistic is equivalent to the *degenerate* $U-$ or $V-$statistic. Hence the limiting distribution is the weighted chi-square distribution $\sum_{k\geq 1}\lambda_k N_k^2$ with the standard Gaussian N_k and coefficients λ_k which are the eigenvalues of the Fredholm integral equation with the complicated kernel, depending to the kernel of the $U-$statistic.

Such limiting distributions are too complicated to be computed.

The third statistic converges to the supremum of special zero-mean Gaussian process U, defined by Silverman (1976). It has the covariance function

$$Cov(U(x),U(y)) =$$
$$k^2 Cov[\mathbb{P}(h(X_1,...,X_k)<x|X_1),\mathbb{P}(h(X_1,\ldots,X_k)<y|X_1)].$$

Its distribution is hardly computable analytically, but one can use simulations to get critical values.

Same conclusions are valid for other statistics of these types.

II **Consistency**

Consistency means that the power of the test tends to zero under the alternative when n tends to infinity. Hence the alternative can be distinguished from the null-hypothesis with high probability in large samples.

The Kolmogorov-Smirnov statistic and the L_2-type statistic are always consistent for standard alternatives. Under standard alternatives for the hypothesis of exponentiality we understand, for instance, the following families of densities:

the Gamma family with the density

$$g(x;\theta) = (\Gamma(\theta+1))^{-1} x^{\theta} \exp(-x),\ x \geqslant 0;$$

the Weibull family with the density

$$g(x;\theta) = (\theta+1) x^{\theta} \exp(-x^{1+\theta}),\ x \geqslant 0;$$

the Makeham family with the density

$$g(x;\theta) = (1+\theta(1-\exp(-x)))\exp(-x-\theta(\exp(-x)-1+x)),\ x \geqslant 0;$$

the "linear failure rate"density

$$g(x;\theta) = (1+\theta x)\exp[-x-\frac{1}{2}\theta x^2],\ x \geqslant 0.$$

It is not always the case for the statistic

$$T_{1n} = \sqrt{n}\int_{R^1}[H_n(t)-F_n(t)]dF_n(t),$$

but this drawback is compensated by the simplicity of this statistic and its standard limiting distribution.

2 Asymptotic efficiency

First let give the synopsis of the notion of asymptotic efficiency of tests. Its development began in 1948 after seminal work by Pitman and was continued by Hodges and Lehmann, Chernoff and Bahadur in 1952 - 1970.

Let $\{T_n\}$ and $\{V_n\}$ be two sequences of statistics based on the sample $X_1, \ldots, X_n$ with distribution P_θ, where $\theta \in \Theta \subset R^1$, and we are testing the null-hypothesis $H_0 : \theta \in \Theta_0 \subset \Theta$ against the alternative
$A : \theta \in \Theta_1 = \Theta \backslash \Theta_0$.

Let $N_T(\alpha, \beta, \theta)$ be the minimal sample size of $X_1, \ldots, X_n$, for which the sequence $\{T_n\}$ attains the power $\beta < 1$ under given significance level $\alpha > 0$ for the alternative value of parameter $\theta \in \Theta_1$. In the same way $N_V(\alpha, \beta, \theta)$ is introduced.

The relative efficiency of the test based on the statistic T_n, with respect to the test based on V_n is the quantity

$$e_{T,\,V}(\alpha,\beta,\theta)=\frac{N_V(\alpha,\beta,\theta)}{N_T(\alpha,\beta,\theta)}.$$

This quantity is too complicated and cannot be calculated. Therefore it is generally agreed to consider the limits:

$$\lim_{\alpha\to 0} e_{T,\,V}(\alpha,\beta,\theta),\quad \lim_{\beta\to 1} e_{T,\,V}(\alpha,\beta,\theta),\ \lim_{\theta\to\partial\Theta_0} e_{T,\,V}(\alpha,\beta,\theta).$$

In the first case we obtain the Bahadur efficiency, the second limit corresponds to Hodges-Lehmann efficiency while the third one leads to Pitman efficiency. Note that just small levels, large powers and close alternatives are interested for practice.

In our case many statistics have non-normal limiting distribution, hence the Bahadur efficiency seems to be most adequate. The key point for the calculation of Bahadur efficiency is the large deviation asymptotics of tests statistics under the null-hypothesis.

Despite great successes of Large Deviation theory, large deviations of $U-$ and $V-$statistics were studied only recently.

Theorem 1. (Nikitin and Ponikarov, 1999). Suppose the kernel h of the $U-$statistic or of von Mises functional V_n is bounded,

$$|h(s_1, \dots, s_k)| \leqslant M,$$

suppose that $\mathbb{E}h = 0$ and that h has rank 1, that is $\Delta^2 = \mathbb{E}\psi^2(X_1) > 0, \quad$ where $\quad \psi(s) = \mathbb{E}(h|X_1 = s)$.

Then it is true that

$$\lim_{n\to\infty} n^{-1} \ln \mathbb{P}\{V_n \geqslant a\} = \sum_{j=2}^{\infty} b_j a^j,$$

where the series with numerical coefficients b_j converges for sufficiently small $a > 0$, and $b_2 = -\frac{1}{2m^2\Delta^2}$.

The proof uses the large deviation principle for $U-$statistics with subsequent minimization of Kullback-Leibler information on a special set of measures.

In case the kernel is degenerate, $\Delta^2 = 0$, the answer is more complicated and depends on the spectrum of the Fredholm integral equation

$$\int_{R^1} h^*(s,t) f(s) ds = \lambda f(t)$$

(Here we set

$$h^*(s,t) = \int_{R^{m-2}} h(s_1, \ldots, s_k) dF(s_3) \ldots dF(s_k), \quad \text{if} \quad k > 2,$$

$$h^*(s,t) = h(s,t), \qquad \text{if} \quad k = 2.)$$

In case λ_0 is the first eigenvalue, then

$$\lim_{n \to \infty} n^{-1} \ln \mathbb{P}\{V_n \geqslant a\} = \sum_{j=2}^{\infty} b_j a^{j/2}, \; b_2 = -\lambda_0 / k(k-1).$$

These results enable to calculate and analyze the Bahadur efficiency of integral statistics based on $H_n(t) - F_n(t)$.

The Kolmogorov-Smirnov statistics are more complicated. Note, that for any t the difference $H_n(t) - F_n(t)$ is a $U-$statistic, too, with a different kernel, say $\mathcal{H}(\cdot; t)$.

Here we meet the *families* of $U-$statistics $\{U_n(t)\}$ depending on t, and we must find the asymptotics of $\mathbb{P}(\sup_t |U_n(t)| > a)$ when $n \to \infty$. One needs rather refined methods using variational calculus, nonlinear analysis and exponential inequalities for the tails of $U-$statistics.

Suppose that the kernel $\mathcal{H}(\cdot; t)$ is non-degenerate for any t and define again the variance function

$$\delta^2(t) = E(E(\mathcal{H}(X_1, \ldots, X_k; t)|X_1)^2) > 0.$$

Denote $\mu_0^2 := \sup_t \delta^2(t)$.

Theorem (Nikitin, 2008). Suppose that the kernels $\{\mathcal{H}(\cdot;t)\}$ *are uniformly bounded, centered and non-degenerate for any* t*. Then there exists such continuous function* v *that*

$$\lim_{n\to\infty} n^{-1}\ln\mathbb{P}\{V_n \geqslant a\} = v(a) = -(2k^2\mu_0^2)^{-1}a^2 + O(a^3), a \to 0.$$

Now take as an example the Baringhaus-Henze statistics for testing symmetry. Using Theorem 1, we get

$$\lim_{n\to\infty} n^{-1}\ln\mathbb{P}\{S_{n,1} \geqslant a\} = -10a^2 + O(a^3),\ a \to 0.$$

In case of Kolmogorov-Smirnov-type statistic we use Theorem 2 and obtain

$$\lim_{n\to\infty} n^{-1}\ln\mathbb{P}\{S_{n,\infty} \geqslant a\} = -\frac{27}{8}a^2 + O(a^3),\ a \to 0.$$

Another example is the Desu statistic. Here we have

$$\lim_{n\to\infty} n^{-1} \ln \mathbb{P}\{D_{n,1} \geqslant a\} = -\frac{210}{11}a^2 + O(a^3),\ a \to 0,$$

$$\lim_{n\to\infty} n^{-1} \ln \mathbb{P}\{D_{n,\infty} \geqslant a\} = -2a^2 + O(a^3),\ a \to 0.$$

Similar results were obtained for other statistics introduced above.

This enables calculating the Bahadur efficiency for various local alternatives. We present the typical table of efficiencies.

Alternative	Distribution		
	Normal	Logistic	$\frac{8}{3\pi(1+x^2)^3}$
location	0.977	0.938	0.905
skew	0.977	0.962	0.925
Lehmann	0.962	0.962	0.962
contamination, $r = 1$	0.988	0.988	0.988
contamination, $r = 5$	0.997	0.997	0.997

Table 1: Local Bahadur efficiency for the statistic S_n^1.

Similar tables can be obtained for other statistics.

Let summarize briefly the results.

• The symmetry tests based on Baringhaus-Henze are very efficient.

• Usually supremum-type tests demonstrate weaker results than integral tests besides special alternatives.

• The tests of exponentiality based on Desu's characterization have medium values of efficiency from 0.5 till 0.8 depending on the alternative.

• Tests of normality based on Polya's characterization are very efficient for location, skew and Lehmann alternatives, but very poor for contamination alternatives.

• Tests for Cauchy distribution have moderately high efficiency from 0.6 till 0.9 depending on the alternative.

This shows that the tests based on characterizations deserve both attention and use in practice.

□