A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 131, 1993

Czesław Domański, Wiesław Wagner***

UNIVARIATE NORMALITY TESTS BASED ON STOCHASTIC PROCESSES

1. INTRODUCTION

The paper discusses univariate normality tests based on stochastic processes. The theory of these tests has been developed extensively since 1973, i.e. the year of D u r b i n´s study publication (1973), devoted to the weak convergence of distribution of sample elements function for unknown parameters.

Here we present three normality tests based upon stochastic processes. The selection of these tests was done on the ground of their application, comparing with the classical normality tests.

Let us start with a short presentation of same notions referring to the theory of stochastic processes.

DEFINITION 1. A set of random variables $X(t)$, $t \in T$ depending on parameter $t \in T \subset R$, where $R = (-\infty, \infty)$ is called a univariate stochastic process.

DEFINITION 2. Stochastic process is called real (complex) when random variables $X(t)$ are variables having real (complex) values.

* Professor at the Institute of Econometrics and Statistics, University of Łódź.

** Professor at the Department of Mathematical and Statistical Methods Applications, Academy of Agriculture, Poznań.

DEFINITION 3. Stochastic process is called continuous (discrete) if scalar t takes only continuous (discrete) values. Distribution of stochastic process is defined when probability distributions of random variables $X(t_1)$, $X(t_2)$, ..., when t_1, t_2, ... $\in T$ are known. The knowledge of the expected value $E[X(t)]$ and autocovariance cov $[X(t_1), X(t_2)]$ $t_1, t_2 \in T$ is required for that purpose.

DEFINITION 4. Stochastic process $\{X(t)\}$, $t \in T$ is called the univariate Gaussian process (normal) when $X(t)$, $t \in T$ are random variables of univariate normal distribution for each $t \in T$. Therefore, the Gaussian process is defined simultaneously by its normality and $E[X(t)]$ and $E[X(t_1)\ X(t_2)]$ since cov $[X(t_1), X(t_2)] = E[X(t_1)\ X(t_2)] - E[X(t_1)]\ E[X(t_2)]$.

The interest in stochastic processes for the construction of univariate normality tests results from the fact that each distribution function of this process is invariant with respect to parameter t. This property allows to use invariance of affine transformation.

Normality tests based on the stochastic process take into account:

a) empirical distribution function - real stochastic process,

b) empirical characteristic function - complex stochastic process.

Generally speaking, the univariate stochastic process which is being considered here is expressed by the following functional

$$Z_n(t) = \sqrt{n}\{F_n(t) - F(t)\}, \quad t \in R,$$

where $F_n(t)$ is the empirical function of the sample elements and $F(t) = E\{F_n(t)\}$ is its expected value. This process expresses the difference between the empirical distribution and the theoretical (expected) distribution defined in the set $T = R$. Under the null hypothesis of normality for $n \to \infty$ the process $Z_n(t) \to Z(t)$ for n where $Z(t)$ is an univariate random variable normally distributed with a zero expected value and known covariance. The aim of the paper is to show how the stochastic process can be used for the construction of univariate normality tests. Making use of different properties of the characteristic function, empirical distribution function and empirical characteristic function we discuss normality tests.

2. PREREQUISITES

Let X be a random variable with an unknown distribution function $G_X(x, \theta) = G_X(x)$, where $x \in R$ and θ is a set of unknown parameters belonging to a certain parameter space Θ. Let, subsequently, a sequence of independent realizations of random variable X be $X_1, \ldots, X_n$ and the values in non-decreasing order $X_{(1)} \leq \cdots \leq X_{(n)}$ be sample order statistics. The distribution function of the normal distribution is denoted by $F_X(x) = F_X(x; \mu, \sigma^2) = F_X(x, \theta)$, where $\theta = (\mu, \sigma^2)$, $\mu \in R$ and $\sigma^2 \in R_+$. If the parameters μ and σ^2 are known, we write $\theta_o = (\mu_o, \sigma_o^2)$. If they are unknown, we find unbiased estimators $\bar{X}$ and S^2, as arithmetic mean and sample variance, respectively, and denote $\hat{\theta} = (\bar{X}, S^2)$.

Empirical distribution function from sample $X_1, \ldots, X_n$ is defined by

$$F_n(X) = \begin{cases} 0, & X < X_{(1)} \\ i/n, & X_{(i-1)} \leq X < X_{(i)}, \quad i = 2, \ldots, n \\ 1, & X \geq X_{(n)} \end{cases}$$

while empirical characteristic function is expressed by

$$C_n(t) = \frac{1}{n} \sum_{j=1}^{n} \exp(it X_j)$$

where $i = \sqrt{-1}$.

The compound null hypothesis of normality is expressed as H_o: $G_X(x, \theta) = F_X(x, \theta)$ against H_1: $G_X(x, \theta) \neq F_X(x, \theta)$. In the case when $\theta = \theta_o$ the sample null hypothesis of normality is denoted as H_o^o: $G_X(x, \theta) = F_X(x, \theta)$. Characteristic function $c(t)$, $t \in R$ of random variable X with a distribution defined by $G_X(x)$ is written as

$$C(t) = \int_{-\infty}^{\infty} \exp(itx)\, dG_X(x)$$

Its basic characteristics are as follows:

a) $C(0) = 1$,

b) $|C(t)| \leq 1$,

c) $\overline{C(t)} = C(-t)$,

d) $C(t)\,C(-t) \approx |C(t)|^2$,

e) $C(t) = \exp\{P(t)\}$ if $P(t)$ is a polynominal of degree ≤ 2 (Marcinkiewicz theorem),

f) for $C(t)$ there exists such $\delta > 0$ that $C(t) \neq 0$ for $|t| < \delta$.

The properties given above are used for constructing normality tests based on empirical characteristic function, which is given in 4.1. It is worth to note that for $X \sim N(\mu, \sigma^2)$ there is $C(t) = \exp(it\mu - \sigma^2 t^2/2)$ and that a distribution with the characteristic function $C(t)$ is normal if and only if $\{-\ln|C(t)|^2\}^{1/2}$ is linear with respect to $t \geq 0$.

3. CRAMER - VON MISES TEST FOR NORMALITY

Let $\psi(X, \theta) = (X - \mu)^2/\sigma^2$. Under the assumption that hypothesis H_o^o is true there is $\psi(X, \theta) \sim \chi_1^2$. Transformation of variables X_i onto $Y_i = \psi(X_i, \theta_o)$ leads to random variables, each of which has the distribution χ_1^2. Let $F_n(y)$ and $F_{\chi_1}(x) = V(x)$ denote distribution functions: empirical for $Y_1, \ldots, Y_n$ sample and distribution χ_1^2, respectively.

The measure of divergence of empirical distribution of a normally distributed random sample $Y_1, \ldots, Y_n$ is given as the stochastic process:

$$Z_n(t) = \sqrt{n}\{\hat{F}_n(t) - V(t)\}, \qquad t \in \langle 0, \infty \rangle$$

This process is used to check H_o^o hypothesis. The well-known goodness - of - fit tests e.g. Kolmogorov - Smirnov and Cramer - von Mises tests are taken into account.

Cramer-von Mises test is presented for $\theta = \hat{\theta}$, i.e. when unknown parameters μ and σ^2 are estimated from the sample $X_1, \ldots, X_n$. The stochastic process $Z_n(t)$ is replaced by

$$\tilde{Z}_n(t) = \sqrt{n}\{\tilde{F}_n(t) - V(t)\},$$

where $\tilde{F}_n(t)$ is an empirical distribution function from the sample $\tilde{Y}_1, \ldots, \tilde{Y}_n$ when $\tilde{Y}_i = \psi(X_i, \hat{\theta})$. The stochastic process $Z(t)$, $t \in \langle 0, \infty)$ is, in terms of methods given by D u r b i n (1973) and N e u h a u s (1974), convergent to the Gaussian process $Z(t)$ with covariance for H_o hypothesis (K o z i o l 1982).

$$\mathrm{cov}(t, t') = V(\min(t, t')) - V(t)V(t') - 2tt'V(t)V(t'),$$

where $V(t) = V(t')$ is a derivative of a distribution function $V(t)$ i.e. it is a density of distribution χ^2_1. For the hypothesis H_o we have

$$\int_{-\infty}^{\infty} \bar{Z}_n^2(t)\, dV(t) \to \int_{-\infty}^{\infty} \bar{Z}_n^2(t)\, dV(t)$$

the statistic of Cramer-von Mises - type. The last functional was studied by D u r b i n (1973) and S t e p h e n s (1976). Their studies were aimed at presenting integral in the form of certain sum which depends on eigen-values and eigen-vectors of covariance matrix

$$\mathrm{cov}(t, t'), \qquad t, t' \in \langle 0, \infty).$$

The summation form of Cramer-von Mises test statistic for verifying the hypothesis H_o is given as

$$M^2 = \frac{1}{12n} \sum \{ V(\hat{Y}_{(i)}) - \frac{2i-1}{2n} \}^2,$$

where $\hat{Y}_i = \psi(x_i, \hat{\theta})$ and $\hat{Y}_1 \leq \ldots \leq \hat{Y}_{(n)}$. The critical values for M^2 were given by A n d e r s o n and D a r l i n g (1952).

4. TESTS FOR NORMALITY BASED ON THE EMPIRICAL CHARACTERISTIC FUNCTION

4.1. DEFINITIONS OF THE EMPIRICAL CHARACTERISTIC FUNCTION

If $F_n(x)$ is an empirical distribution function based on the sample $X_1, \ldots, X_n$ then the function having complex values

$$C_n(t) = \int_{-\infty}^{\infty} e^{itX}\, dFn(x) = \frac{1}{n} \sum_{j=1}^{n} \exp(it\, X_j) = \frac{1}{n} \sum_{j=1}^{n} \cos tX_j + \frac{i}{n} \sum_{j=1}^{n} \sin tX_j$$

is called the empirical characteristic function (ECF). Its basic properties are as follows:

a) $C_n(0) = 1$;

b) $|C_n(t)| \leq 1$;

c) $C_n(t) = \overline{C_n(-t)}$;

d) $A_n(t) = |C_n(t)|^2 = \frac{1}{n} + \frac{2}{2} \sum_{j<j} \cos [t(X_j - X_{j,})]$;

e) if $Y_j = aX_j + b$, a, b - constant, $a \neq 0$ and $C_{X,n}(t)$ $C_{Y,n}(t)$ denote the ECF of random variables X and Y, then

$$C_{Y,n}(t) = e^{i+b} C_{X,n} \quad (at);$$

f) $E [C_n(t)] = C(t)$;

g) $\tilde{A}_n(t) = |\tilde{C}_n(t)|^2$ is invariant with respect to the shift and change of scale of the parameters, where $\tilde{C}_n(t) = C_n(t/s)$ while $S = \sqrt{S^2}$ is a standard deviation from the sample $X_1, \ldots, X_n$.

h) for the fixed $\tilde{T} < \infty$ (F e n r v e r g e r and M u r e i k a 1977)

$$P\{\lim_{n\to\infty} \sup_{t<T} |C_n(t) - C(t)|\} = 0 = 1;$$

i) $n_o \int^{T_n} |C_n(t) - C(t)|^P \, dt \to 0, \; 0 < P \leq 2.$

Some of the above properties are analogous to those presented in section 2.

4.2. A TEST FOR NORMALITY BASED ON THE SQUARED ABSOLUTE VALUE OF THE ECF

Let $A(t) = |C(t)|^2$ and $A_n(t) = |C_n(t)|^2$ denote absolute values of the characteristic function for the distribution function of the random variable and ECF based on the sample $X_1, \ldots, X_n$. Function $A_n(t)$ is invariant with respect to location parameter, therefore it can be used to test the hypothesis H_o when μ is unknown and σ is known. Let us first define the complex stochastic process

$$Z_n(t) = \sqrt{n} \{C_n(t) - C(t)\}$$

which is weakly convergent to the Gaussian process characterized by the following properties:

a) $Z(t) = \overline{Z(-t)}$,

b) $E [Z(t) = 0$,

c) $E\,[Z(t)\,Z(t')] = C(t + t') - C(t)\,(t')$.

Next we shall define the real stochastic process,

$$Z_n^1(t) = \sqrt{n}\,\{A_n(t) - A(t)\}$$

which is also weakly convergent to the Gaussian process characterized by the following properties:

a) $Z^1(t) = Z^1(-t)$;

b) $E(Z^1(t)) = 0$,

c) $E\,[Z^1(t)Z^1(t')] = 2\mathrm{Re}\,\{C(-t)\,C(-t')\,C(t+t') + C(-t)\,C(t')\,C(t-T') - 4A(t)\,A(t')\}$, where Re (:) denotes the real part of the complex number which is the argument of the Re operator.

For the hypothesis H_o': $G_x(x, 0) = F_x(x, \mu, 1)$ the process $Z^1(t')$ is transformed into (M u r o t a and T a k e u c h i 1981)

$$Z_n^1 = \sqrt{n}\,(A_n(t) - \exp(-t^2))$$

since then

$$C(t) = C(-t') = |C(t)|^2 = A(t) = \exp(-t^2)$$

and

$$E[Z^1(t)\,Z^1(t')] = 4\exp(-t^2-t'^2)\cosh(tt') - 1).$$

A simple test for normality is obtained when $A_n(t)$, for a fixed t, is treated as the test statistic instead of certain functionals which make use of $Z_n^1(t)$.

It is possible to determine moments of the statistic $A_n(t)$ under the hypothesis together and the skewness and kurtosis measures which e.g. for t = 0.5 assume values - 1,83 $\sqrt{n}$ and 2.76/n. The H_o' hypothesis is verified in such a way that it is rejected when $A_n(t) > A_n$ for t close to zero, where α is a pre-assigned significance level. The critical values for the test $A_n(t)$ can be established for different t close to zero. M u r o t a (1981) accepted 1,0 as an appropriate parameter t and he fixed for it the critical values by means of computer simulation. These values are contained in Table 1.

Table 1

Critical values $A_n(\alpha, 1,0)$ of $A_n(1,0)$ test

n \ α	0.05	0.10	0.50	0.90	0.95
10	0.1653	0.2155	0.4239	0.6557	0.7174
20	0.2075	0.2459	0.3959	0.5575	0.6025
50	0.2571	0.2830	0.3191	0.4798	0.5086

4.3. TEST FOR NORMALITY BASED ON THE SQUARED ABSOLUTE VALUES OF THE STUDENTIZED ECF

Now we give a test for normality to verify the hypothesis

$$H_o: G_X(x; \theta) = F_X(x: \mu, \sigma^2),$$

where μ and σ^2 are unknown. The studentized form of the ECF is defined as $C_n t = C(t/s)$, where S is a standard deviation from the sample $X_1, \ldots, X_n$. Then the square module of the studentized ECF is denoted as $\tilde{A}(t) = |C_n(t)|^2$. A change of $A_n(t)$ is invariant with respect to the change of location and variability parameters which results from

$$\tilde{A}_n(t) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp\{it(X_j - X_k)/S\}.$$

Therefore, the squared absolute value of $A_n(t)$ will be an appropriate test for both H_n and H_o' hypotheses.

M u r o t a and T a k e u c h i (1981) prove the following theorem.

THEOREM 1. Assume that the distribution of variable X has the finite fourth moment $\mu_4 = E(X^4)$ with $E(X) = 0$ and $D^2(X) = 1$. Then, due to the properties of ECF, the process $Z_n(t)$ is weakly convergent to the Gaussian process $\tilde{Z}(t)$, i.e.

$$Z_n(t) = \sqrt{n}\{\tilde{C}(t) - C(t)\} \to \tilde{Z}(t)$$

and the process $Z_n(t)$ has the following properties:

a) $E[Z(t)] = 0$;

b) $\tilde{Z}(t) = \tilde{Z}(-t)$;

c) $E\,[\tilde{Z}(t)\tilde{Z}(t')] = C(t+t') - C(t)C(t') + \frac{1}{2}\,t\,C(t)\,[C(t') +$

$+ C''(t') + \frac{1}{2}t'C'(t)[C(t) + C''(t')] + \frac{1}{4}\,(\mu_4 - 1)tt'C'(t)C'(t'),$

where $C'(\ )$, $C''(\ )$ denote the first and second derivative of $C(\ )$ respectively.

A similar theorem can be formulated for the stochastic process of the real values. It makes use of the square module of the studentized ECF

$$Z_n^2(t) = \sqrt{n}\,\{\tilde{A}_n(t) - A(t)\} \to Z^2(t)$$

and is weakly convergent to the Gaussian process $Z(t)$. The Gaussian process has the following properties:

a) $Z^2(t) = Z^2(-t)$,

b) $E\,[Z^2(t)] = 0$,

c) $E\,[Z^2(t)Z^2(t')] = 2\mathrm{Re}\{C(-t)\,C(-t')\,E\,[\tilde{Z}(t)\tilde{Z}(t')] +$

$+ C(-t)\,C(t')\,E\,[\tilde{Z}(t)\tilde{Z}(t')]\}.$

The last stochastic process for the H_o hypothesis is transformed into

$$Z_n^2(t) = \sqrt{n}\,\tilde{A}(t) - \exp(-t^2) \to Z(t)$$

and it is the process which is weakly convergent to the Gaussian process with covariance:

$$E\,[Z^2(t)Z((t')] = 4\,\exp(-t^2 - t'^2)\,[\cosh(tt') - 1 - t^2t'^2/2].$$

Hence, it is possible to construct a test for normality based on the statistic $\tilde{A}_n(t)$. The moments for $A_n(t)$ were determined by M u r o t a (1981), while the critical values for $t = 1.0$ were given by M u r o t a and T a k e u c h i (1981) (of Table 2).

T a b l e 2

Critical values $\tilde{A}_n(\alpha, 1, 0)$ of statistics $A_n(1, 0)$

n \ α	0.05	0.10	0.50	0.90	0.95
10	0.3604	0.3650	0.3883	0.4325	0.4527
15	0.3512	0.3557	0.3792	0.4241	0.4440
20	0.3475	0.3523	0.3753	0.4172	0.4365
35	0.3462	0.3505	0.3717	0.4059	0.4192
50	0.3466	0.3509	0.3701	0.3991	0.4094

The comparative studies of the power of tests for normality based on $\tilde{A}_n(t)$ and $A_n(t)$ show that the test $\tilde{A}_n(t)$ has an advantage over the $A_n(t)$ and the power is the greatest for $t = 1$.

5. FINAL REMARKS

The tests presented above do not discuss comprehensively the problem of application of stochastic processes to the construction of goodness-of-fit tests. The studies on this problem originated as early as in 1955 by D a r l i n g (1955) and then developed by D u r b i n, K n o t t and T a y l o r (1975). They aimed at different possible ways of defining Cramer-von Mises test. The basic results include the expression of the functional, being Cramer-von Mises statistic in the form of a non-finite series of normal variables with N(0, 1) distribution with coefficients which are the eigen-values of Fredholm integral equation.

The introduction of the ECF made it possible for the research on tests for normality based on stochastic processes to take a new direction. Earlier the empirical distribution function characterized the properties of distribution and now this role was taken over by the ECF.

Along with the tests discussed in this paper there are many other tests for univariate normality. They were given by e.g. K o n t r o c c v e l i s (1980), K o n t r o c c v e l i s and K e l l e r m e i e r (1981) E p p s and P u l l e y (1983), and H a l l and W e l s h (1983). The theory of ECF was also studied in a multivariate case (e.g. C s ö r g o 1984), which makes it possible to construct test of for multivariate normality. The examination of properties of these tests is also the subject of interest of the authors.

REFERENCES

A n d e r s o n T. W., D a r l i n g D. A. (1952), *Asymptotive Theory of Certain Goodness Criteria Based on Stochastic Processes*, "Annals of Mathematical Statistics", No. 23, p. 193-212.

C s ö r g ö S. (1986), *Testing for Normality in Arbitrary Dimension*, "Annals of Statistics", No. 14, p. 708-723.

D a r l i n g D. A. (1955), *The Cramer-Smirnov Test in the Parametric Case*, "Annals of Mathematical Statistics", No. 26, p. 1-20.

D u r b i n J. (1973), *Weak Convergence of the Sample Distribution Function when Parameters Are Estimated*, "Annals of Statistics", No. 1, p. 279-290.

D u r b i n J., K o n t t M., T a y l o r C. C. (1975), *Components of Cramer-von Misses Statistics*, "Journal of Royal Statistical Society", Ser. B, No. 37, p. 216-237.

E p p s T. W., P u l l e y L. B. (1983), *A Test for Normality Based on the Empirical Characteristics Function"*, "Biometrika", No. 70, p. 723-726.

F e n e r v e r g e r A., M u r e i k a R. A. (1977), *The Empirical Characteristic Function and its Applications*, "Annals of Statistics" p. 88-97.

H a l l P., W e l s h A. H. (1983), *A Test for Normality Based on the Empirical Characteristic Function*, "Biometrika", No. 70, p. 485-489.

K o n t r o c c e v e l i s I. A., K e l l e r m e i e r J. (1981), *A Goodness-of-Fit Test Based on the Empirical Characteristic Function when Parameters Must Be Estimated*, "Journal of Royal Statistical Society", Ser. B, p. 173-176.

K o z i o l J. A. (1982), *A Class of Invariant Procedures for Assessing Multivariate Normality*, "Biometrika", No. 69, p. 423-427.

M u r o t a K. (1981), *Test for Normality Based on the Empirical Characteristic Function*, Rep. Stat. Appl. res. JUSE, No. 28, p. 1-14.

M u r o t a K., T a k e u c h i K. (1981), *The Studentized Empirical Characteristic Function and Its Application to Test for the Shape of Distribution*, "Biometrika", No. 68, p. 55-65.

N e u h a u s G. (1974), *Asymptotic Properties of the Cramer-von Mises Statistic when Parameters Are Estimated*, Proceedings Prague Symposium Asymptotic Statistics, ed. J. Hajek, p. 257-297.

S t e p h e n s M. A. (1976), *Asymptotic Results for Goodness-of-Fit Statistics with Unknown Parameters*, "Annals of Statistics", No. 4, p. 369-357.

Czesław Domański, Wiesław Wagner

TESTY NORMALNOŚCI OPARTE NA PROCESACH STOCHASTYCZNYCH

Artykuł przedstawia testy normalności oparte na procesach stochastycznych. W szczególności zaprezentowany został test Cramera-van Misesa i dwa testy normalności oparte na empirycznej funkcji charakterystycznej rozkładu Studenta. Podane wartości krytyczne umożliwiają ich praktyczne zastosowanie i analizę ich własności.