The Economic Journal of Nepal
Vol. V No. 1 Jan-Mar. 1982 EIC-TU

An Easy Approach to Some Exact Tests for Normality

A. K. L. Das and K. B. Basnyat★

The concept of normality, which was first developed by English mathematician De-Moivre in 1773, has wide applications in social and behavioral sciences. As such, the use of normality in statistics, econometrics and research works would appear unquestionable. Not to talk of other fields, the concept of normality is assumed even when we should have to find out the correlation between two variables. Besides, it is obvious that in most of the cases we assume that a particular sample has come from a normal population despite the fact that no statistical test is carried out to verify this argument. According to central limit theorem, most of the large samples tend to normality; but in case of small samples, we cannot say definitely whether a sample is normal or not unless we carry out a particular statistical test. As there is a growing importance of normality in research works, an attempt has been made in this paper regarding the use of some statistical tests with a view to testing normality both in case of large and small samples. The following are some exact statistical tests to investigate whether a particular sample has come from a normal population or not.

1. K. Pearson's Probability Product Test:

This test was developed in 1970 and is used both in case of large and small samples.

Let us take n real independent random variables x_1 , x_2 , ... x_n which satisfy

★ Mr. Das is a Deputy Research Officer and Mr. Basnyat is a Statistician at the Centre for Economic Development and Administration, Tribhuvan University, Kathmandu.

the condition $P\{(x_k = o)\} = 0$ and have symmetrical distribution about zero. Let it be noted that $n \geqslant 3$. In order that x_k to be independently identically distributed normal with mean zero and common standard deviation σ, the necessary and sufficient condition is that $y_1, y_2, \ldots\ldots y_{n-1}$ are independently distributed according to student's law with $1, 2, \ldots\ldots n-1$ degree of freedoms respectively

$$\left.\begin{aligned} \text{Where } y_1 &= \frac{x_2}{1\ x_1\ 1} \\ y_2 &= \frac{\sqrt{2}\ x\ 3}{\sqrt{(x_1^2 + x_2^2)}} \\ y_{n-1} &= \frac{\sqrt{n-1.}\ x_n}{\sqrt{x_1^2 + x_2^2 + \ldots\ldots + x_{n-1}^2}} \end{aligned}\right\} \ldots (I)$$

Suppose that $x = (x_1 \ldots\ldots x_n)$ is a set consisting of n independent random variables and let P be the set of all normal distributions with mean u and variance σ^2. It is assumed that the random variables $x_1, x_2, \ldots\ldots x_n$ are identically distributed in P. Now we make the transformation $y = f(x)$ so that the random variables $y_1, y_2, \ldots\ldots y_{n-2}$ are independently distributed following student's law degree of freedoms $1, 2, \ldots\ldots$ and $n-1$ respectively. Thereafter the original random variables $x_1, x_2 \ldots\ldots x_n$ will have a distribution belonging to the set P.

Lemma:

Consider n real independent random variables with mean $M (-\infty < M < \infty)$ and variance σ^2. Let us take a set of other variables $z_1, z_2, \ldots\ldots z_{n-1}$ which are given by:

$$z_1 = \frac{x_1 - x_2}{\sqrt{2}}$$

$$z_2 = \frac{x_1 + x_2 - 2\times 3}{\sqrt{2}}$$

$$z_i = \frac{x_1 + x_2 + \ldots\ldots\ldots i.x_{i+1}}{\sqrt{i(i+1)}}$$

$$z_{n-1} = \frac{x_1 + x_2 + \ldots\ldots (n-1)\, x_n}{\sqrt{n(n-1)}}$$

... (II)

It is to be noted that z_k ($k = 1, 2 \ldots\ldots\ldots n-1$) are i.i.d. normal with mean zero and variance σ^2 if and only if x_k ($k = 1, 2, \ldots\ldots\ldots n$) are normal with mean u and variance σ^2 . In order to have y_1 , y_2 , $\ldots\ldots\ldots y_{n-2}$ as mentioned above, we apply kotwaski's result to the above-mentioned random variables z_1 , z_2 , $\ldots\ldots\ldots z_{n-1}$. Then only y_1 , y_2 $\ldots\ldots\ldots y_{n-2}$ are independent random variables each distributed according to student's Law with 1. 2, $\ldots\ldots\ldots$ n–2 degree of freedoms respectively. But it should be taken into account that the orginal variables x_1 , x_2 , $\ldots\ldots\ldots x_n$ should be normal with mean u and variance σ^2 .

Suppose that x_1 , x_2 , $\ldots\ldots\ldots x_n$ are n real independent random variables with mean M and variance σ^2 ($-\infty < M < \infty, \sigma > 0$), Let z_i , y_1 and y_k be defined by

$$z_i = \frac{(x_1 + x_2 + \ldots\ldots x_i - i\, x_{i+1})}{\sqrt{i\,(i+1)}}, \quad i=1, 2, \ldots\ldots n-1$$

$$y_1 = \frac{z_1}{1\, z_2\, 1}$$

$$y_k = \frac{\sqrt{k}\; z_{k+1}}{\sqrt{z_1^2 + z_2^2 + \ldots\ldots + z_k^2}}, \quad k=2, \ldots\ldots n-2 \qquad \text{(III)}$$

Then $x_1, x_2, \ldots\ldots x_n$ are normally distributed with mean u and variance σ^2 if and only if $y_1, y_2, \ldots\ldots y_{n-2}$ are identically distributed, obeying student's law with degree of freedoms 1, 2, ... & n-2 respectively. Now we are in position to replace the composite null hypothesis H_o: $x = \{(x_1, x_2, \ldots\ldots x_n)\ n \geqslant 4$ is a random sample from $N < M, \sigma^2$ by simple equivalent null hypothesis H_o:$y = (y_1, y_2, \ldots\ldots y_{n-2})$ are independently distributed according to students' how with 1, 2, ... n-2 d.f's respectively.

In order to test H_o', the random vairables $y_1, y_2, \ldots\ldots y_{n-2}$ and are transformed to their probability integral $y_1 = F_1(y_1), y_2 = F_2(y_2), \ldots\ldots y_{n-2} = F_{n-2}(y_{n-2})$ where $y_1, y_2, \ldots\ldots y_{n-2}$ are independent U (0,1) random variables.To test H_o', we can use K. Pearson's Probability Product test using the fact that

$$2\, P_{n-2} = -2 \log \prod_{j=1}^{n-2} y_j \sim x^2_{2(n-2)} \qquad \ldots\ldots\ldots\ldots \text{(IV)}$$

Decision rule:

If the calculated value of x^2 at a particular degree of freedom is greater than the tabular value of x^2 at 5% level of significance, the hypothesis is rejected. But the reverse case emerges if the calculated value of x^2 at a particular degree of freedom is less than the tabular value of x^2 at 5% level of significance.

Illustrative Example:

x: 6, 1, – 4, 2, 5, 0

In order to test the normality of the above sample, K. Pearson's Probability product test can be applied.

Table No. 1

H_o: The sample has come from the normal population

Vs H_1 : The sample has not come from the normal population

Value of z_i	Value of y_i	Value of f_i $(y_i$	Calculated Value of x^2
$z_1 = 3.54$	$y_1 = 0.5784$	0.6669	
$z_2 = 6.12$	$y_2 = -1.2132$	0.1744	
$z_3 = -6.06$	$y_3 = 0.7905$	0.7565	
$z_4 = 4.25$	$y_4 = 0.5705$	0.2994	8.10
$z_5 = -2.92$	$y_5 = 0.4528$	0.6655	
$z_6 = 2.16$			

Since the calculated value x^2 is less than the tabular value of x^2 at 5% level of significance, the null hypothesis is accepted, that is the given sample has come from the normal population.

2. Csorgo and Seshadri's Test:

This test was developed by csorgo and seshadri in 1970 and is used both for large and small samples.

Let $x_1, x_2, \ldots\ldots x_n$, $n = 2k + 3, k \geqslant 2$ be n real independent random variables with mean u and variance σ^2 $(-\infty < M < \infty, \sigma > 0)$.

Let $z_1, z_2, \ldots\ldots z_{n-1}$ be the set of variables as defined above and let $y_1, y_2, \ldots - y_{k+1}$ be an another set of random variables defined by

$$y_1 = z_1^2 + z_2^2$$

$$y_2 = z_3^2 + z_4^2$$

....

and $y_{k+1} = z_{n-2}^2 + z_{n-1}^2$ where $k+1 = \frac{n-1}{2}$ --(I)

Let $s_{k+1} = \sum_{i=1}^{k+2} y_i$ and let us define $n_y^* = \prod_{r=1}^{Y} \frac{y_i}{s_{k+1}}$, r=1.2.....k

Where $n_1^*, n_2^*, \ldots\ldots n_k^*$ act like K order statistics of K independent U (0,1) random variables if and only if $x_1, x_2, \ldots\ldots x_n$ are normally distributed with mean u and variance σ^2. Thus,

$$2.P_k^* = -2 \sum_{r=1}^{k} \log n_r^* \quad \ldots \ldots \ldots \ldots \ldots \text{(II)}$$

is a x_{2k}^2 random variable if and only if $x_1, x_2, \ldots\ldots x_n$ are normally distributed with mean u and variance σ^2.

Decision rule:

If the calculated value of x^2 is greater than the tabular value at 5% level of significance, the hypothesis is rejected. But if the calculated value of x^2 is less than the tabular value at 5% level of significance, the reverse case emerges.

Illustrative Example:

x: 32, 34, 35, 37, 42, 43, 47, 58, 49, 62, 69.

To test the normality of this sample, the above- mentioned test, that is, Csorgo and seshadri's can be applied.

Table No. 2

H_o: The sample has come from normal population
Vs H_1 : The sample has not come from the normal population

Value of z_i 's	Value of y_i 's	Value of n_r	Calculated Value of x^2
$z_1 = -1.4142$	$y_1 = 4.67$	$n^*_1 = 0.0028$	
$z_2 = -1.6330$	$y_2 = 53.33$	$n^*_2 = 0.0351$	
$z_3 = -2.8868$	$y_3 = 123.71$	$n^*_3 = 0.1099$	24.32
$z_4 = -6.7082$	$y_4 = 618.29$	$n^*_4 = 0.4840$	
$z_5 = -6.3901$	$y_5 = 852.41$		
$z_6 = -9.1039$	S_5 = total = 1652.91		
$z_7 = -18.1738$			
$z_8 = -16.0706$			
$z_9 = -18.0250$			
$z_{10} = -22.9784$			

Since the calculated value of x^2 is greater than the tabular value of x^2_{n-2} at 5% level of significance, the hypothesis is rejected, that is the given sample has not come from the normal population

3. Kolmogrov - Smirnov One-Sample test:

The Kolmogrov - smirnow one-sample test shows the degree of agreement between the distribution of a set of observed sample values and some specified theoritical distribution. In fact, this test specifies the cumulative frequency distribution and makes comparison with the observed cumulative frequency distribution.

Let F_o (x) be a completely specified cumulative frequency distribution function and let it be the theoritical cumulative distribution under H_o. Let S_N (x) be the observed cumulative frequency distribution of a random sample of N observations. In this case, x is any possible score and

$S_N(x) = \frac{K}{N}$ where K is the number of observations equal to or less than x.

It is expected that the difference between S_N (x) and F_o (x) are small and largest value of F_o (x) – S_N (x) is called the maximum deviation (D) which is given by

$$D = \text{maximum} \left| F_o(x) - S_N(x) \right| \text{ ————(i)}$$

Decision rule:

To find the probability associated with the occurrence under H_o of values as large as the observed value of D, the concerned table is referred. If that p is equal to or less than 2, the null hypothesis is rejected.

Illustrative Example:

X: 16, 12, 19, 20, 24, 15, 13, 10, 14.

In order to test the normality of this sample, the Kolmorgrov–Smirnov one-sample test is used.

Table No. 3

H_o: The sample has come from the normal population

Vs H_1 : The sample has not come from the normal population

Value of x	F_o (x)	S_N (x)	F_o (x)–S_N (x)	Value of D
– ∞	0.0000	0	0	
10	0.0901	1/9	– 0.02	
12	0.1894	2/9	– 0.03	
13	0.2546	3/9	– 0.08	0.16
14	0.3336	4/9	– 0.11	
15	0.4207	5/9	– 0.13	
16	0.5080	6/9	– 0.16	
19	0.7611	7/9	– 0.02	
20	0.8238	8/9	– 0.07	
24	0.9671	9/9	– 0.03	
+ ∞	1.0000	1	0	

Since the probability associated with the observed value of D is greater than $\alpha = .10$, the null hypothesis is accepted. Hence it can be concluded that the sample has come from the normal population.

4. Chi-square test:-

This test is used for large samples and it tests whether the observed frequencies are sufficiently close to the expected ones to be likely to have occurred under H_o.

The null hypothesis may be tested by

$$x^2 = \sum_{i=1}^{k} \frac{(0_i - E_i)^2}{E_i} \quad \ldots\ldots\ldots\ldots \text{(i)}$$

Where 0_i = observed number of cases categorized in the ith category.

E_i = expected number of cases in the ith category under H_o.

$\sum_{i=1}^{k}$ directs one to sum over all K categories.

Decision rule:

If the calculated value of x^2 is greater than the tabular value of x^2 at 5% level of significance, the null hypothesis is rejected, whereas the reverse case is experienced when the calculated value of x^2 is less than the tabular value at the same level of significance. In case of sample size greater than 30 the Z – test is used.

Illustrative Example:

X: 15, 20, 22, 23, 20, 24, 19, 18, 19, 18, 16, 21, 15, 15, 16, 12, 19, 20, 24, 15, 13, 10, 14, 9, 13; 17, 18, 18, 9, 16, 18, 13, 9, 22, 17, 17, 15, 16, 16, 16, 19, 11, 14, 19, 18, 20, 10, 9, 16, 12, 15, 17, 18, 13, 20, 11, 11, 18, 14, 17, 23, 23, 14, 15, 15, 19, 15, 20, 20, 18, 23, 21, 17, 20, 20, 20, 21, 21, 25, 17, 18, 20, 21, 16, 15, 14, 18, 18, 22, 17, 17, 24, 18, 12, 16, 13, 15, 17, 17.

Table No. 4

H_o: The sample has come from the normal population
Vs H_1 : The sample has not come from the normal population

Class Interval	Middle point	Observed frequency	z_1	z_2	Probability	Expected frequency	Calculated value of x_2
$-\infty - 8.5$	∞	0	$-\infty$	-2.31	0.0104		
8.5 – 10.5	9.5	6	-2.31	-1.77	0.0279	3.79	
10.5 – 12.5	11.5	6	-1.77	-1.23	0.0710	7.03	
12.5 – 14.5	13.5	10	-1.23	-0.68	0.1389	13.75	
14.5 – 16.5	15.5	20	-0.68	-0.14	0.1961	19.41	6.98
16.5 – 18.5	17.5	24	-0.14	0.40	0.2111	20.90	
18.5 – 20.5	19.4	17	0.40	0.95	0.1735	17.18	
20.5 – 22.5	21.5	8	0.95	1.49	0.1029	10.19	
22.5 – 24.5	23.5	7	1.49	2.03	0.0469	4.64	
24.5 – 26.5	25.5	1	2.03	2.58	0.0162		
$26.5 - +\infty$	∞	0	2.58	$+\infty$	0.0049	2.09	

Footnote: The first three and the last three expected frequencies have been pooled since the expected frequencies in these caces are less than five.

Since the tabular value of x^2 for the degree of freedom more than 70 is not available in the statistical table, the formula $\sqrt{2x^2} - \sqrt{2n-1}$ is used as a normal deviate. Using this formula, the value of z is found to be 1.09 and the corresponding probability p becomes 0.1379. As this probability is greater than the pre-assigned probability .05, the null hypothesis is accepted; that is, the given sample comes from a normal population.

Recommendation:

In fact, the superiority of different tests for normality can be determined only by finding out the power of these tests but because of the non-availability of the tabuler value of non-central chi-square, the power of the abovementioned four tests could not be calculated.

However, K. Pearson's Probability Product Tests seems to be reliable compared to other statistical tests as it specifies the parameters and distributions as well. Besides, Csorgo and Seshadri's test is simple and reliable. But it can be used only in case of odd sample size. According to sidney and siegal, the Kolmogrov-smirnow one-sample test is the most powerful test for small sample and the Chi-square test for the large samples. However, the parameters are not specified in these cases.

Selected References

1. Sydney and Siegal, Non-parametric Statistics.
2. Sankhiya, Vol. 25, 26, 28, 29. 30, 31, 32.
3. The Annals of Mathematical Statistics, Vol. 28 and 38.
4. Journal of American Statistical Association, Vol. 67 to 69.
5. Journal of Royal Statistical Society, Vol. 35, No. 1.