Journal of Crop and Weed, 13(1): 07-16 (2017)

Different methods for judging the normality assumption for univariate and bivariate data and its remedial measure

S. S.DAS, [1]A. K.DAS, [1]A.MAZUMDER AND M. K. DEBNATH

Department of Agricultural Statistics,Uttar Banga Krishi Viswavidyalaya, Pundibari, Coochbehar, West Bengal - 736165

[1]Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, West Bengal-741252

Received: 14-03-2017; Revised :20-04-2017; Accepted : 24-04-2017

ABSTRACT

Methods of finding density functions for statistics often lead to intractable mathematical expressions but the calculations are relatively simple and a good deal of work has been done using the basic assumption of normality. The most popular tests among practical Statisticians are tests which depend on normality in the parent population. When there is grave doubt about the assumption, non-parametric tests should be used, even at some sacrifice of power. Parametric tests like τ-test, t-test and F-test are applied under the assumption that the data are normally distributed. One should test the data whether they follow normal or not before conducting the parametric tests. If there is reason to suspect non-normality, it is advisable to try a transformation. In this work different methods for testing the normality are discussed and eight data sets are taken and they are tested for normality. The data sets 2 and 5 are found normal. The other data sets i.e.3,4 and 6 are non-normal. The bivariate data set 7 is normal where as bivariate data set-8 is non-normal. Box-Cox- power transformation is used for all non-normal data and it is found that all the transformed data follow normality. But it is not necessary that the Box-Cox-power transformation will always makes the data normal.

Keywords : Bivariate data, Box-Cox transformation, Box-Whisker plots, Shapiro and Wilk test, goodness of fit Jarque – Bera (JB) test

One of the earliest applications in the history of the normal distribution was provided by Laplace and Guass (1968) to describe the measurement error in the observation of the motions of planets. Since then, the normal distribution is usually applied in the measurement error assumption. Thus, the normal distribution is called "Law of Errors" by some scientists. Quetelet (1796-1874) used it to describe physiological and behavioral phenomena and Galton in the early 1900's, used it to describe anthropometric measurements. Until now, the normal distribution is the most widely used probability distribution based on three main reasons. The first reason is the mathematical properties of the normal distribution. The second reason is that many scientists have noted that random variables often have normal or approximately normal distributions. The third reason is the central limit theorem. The logarithm of the variate or the square root or the inverse sine may be more nearly normal (Bartlett, 1947).

Statistical inference procedures have been systematically developed under the assumption of normality; in addition, in a large number of applications, the normal mode1 is fitted, at least approximately, to underlying situations based on the consequence of the central limit theorem. It is therefore important to devise tests for normality or to have methods for verifying the reasonableness of the normality assumption. Thus normal distribution plays a very important role in statistical theory. Normal distribution has got the tremendous importance in the theory and application of statistics (Sahu, 2010). Most the distribution can be brought under the normal distribution such as normal distribution is a limiting case of Poisson distribution with the parameter $\lambda \to \infty$ (Sahu, 2010). Similarly, normal distribution can also be obtained as a limiting case of Binomial distribution when $n \to \infty$ and neither p nor q is very small (Sahu, 2010).

It is very essential to judge the data before any analysis whether the data follow normal distribution or not. If the data do not follow normal distribution, then there are some remedial measures for converting the data to be normalized. Even if a variable is not normally distributed, it can sometimes be brought to normal form by simple transformation of variable. For example, if the distribution of X is skewed, the distribution of $\sqrt{x}$ might come out to be normal. The entire theory of small sample tests, *viz.*, *t*, F, x^2 tests, *etc*., is based on the fundamental assumption that the parent populations from which the samples have been drawn follow normal distribution.The normal distribution is easy to work with mathematically. In many practical cases, the methods developed using normal theory work quite well even when the distribution is not normal. Box-Cox (1964) power transformations are used for non-normal data.

Email: soumitradas30@gmail.com

MATERIALS AND METHODS

The following data sets are considered for testing the normality. The feature and sources of the collected data sets are given below.

1. The information about number of pod per plant for 12 plants (n=12):

   Source: Jaguli farm, BCKV

2. The height of 12 students (in inches) from New PG hostel (n=12):

   Source: students are selected from New PG hostel in BCKV.

3. The information about the number of leaves in a shoot of wood apple plant for 27 plants (n=27):

   Source: Jaguli farm, BCKV.

4. The information about forest cover 2009 as revised (forest cover reported in ISFR 2009 + Interpretational changes) in km^2 (n=35):

   Source: Report of Forest Survey of India 2009.

5. The information about the number of aphids per three leaves(n=49):

   Source: Jaguli farm, BCKV.

6. The information about the rainfall distribution from 1901 to 2002 in Kolkata district (n=102):

   Source: www.india water portal.org.

7. The information about organic carbon ($g\ kg^{-1}$) (x_1) and CEC[c mole $(p)^+\ kg^{-1}$] (x_2) for soil samples (n=30) :

   Source: Journal of the Indian Society of Soil Science, Vol. 59, No. 2, pp 125-133 (2011).

8. The information about available Phosphorus (mg kg^{-1}) (x_1) and Nitrogen (kg ha^{-1}) (x_2) content in soil samples (n=66) :

Source: Journal of the Indian Society of Soil Science, Vol. 59, No. 2, pp 125-133 (2011)

The data set 1 to 6 are univariate but the last two sets (*i.e.* 7 & 8) are bivariate and n represents the number of observations for a particular data sets.

For applying parametric test one need the assumption that the data are normally distributed. Assumption of normality can be tested by various graphical methods like- histogram, box-whisker plot, normal probability plots and statistical test like – chi-square, Anderson-Darling, Kolmogorov-Smirnov tests etc. None of the methods, however, is absolutely definite.In this article some tests for normality, like-Box-plot, Q-Q plot, W-test, Jarque- Bera (JB) test, χ^2 test for goodness of fit for univariate data are discussed.

Chi-square plot for bivariate data

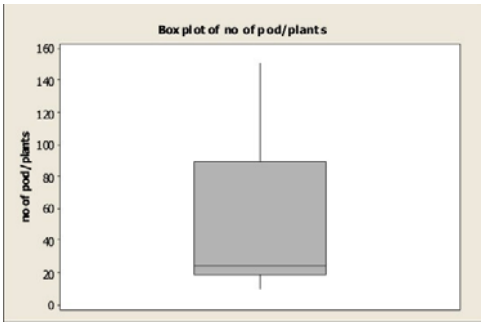
If the observations were generated from a multivariate normal distribution, each bivariate distribution would be normal and the contours of constant density would be ellipses. We know that for the bivariate normal distribution $\left(\underset{\sim}{x}-\underset{\sim}{\mu}\right)'\sum^{-1}\left(\underset{\sim}{x}-\underset{\sim}{\mu}\right)$ follows χ^2 distribution with 2 d.f. Moreover, the set of bivariate outcomes $\underset{\sim}{x}$ such that $\left(\underset{\sim}{x}-\underset{\sim}{\mu}\right)'\sum^{-1}\left(\underset{\sim}{x}-\underset{\sim}{\mu}\right)\le\chi^2_{0.5,2}$ has probability 0.5. Thus we should expect roughly 50 per cent of sample observations to lie in the ellipse $\left(\underset{\sim}{x}-\underset{\sim}{\bar{x}}\right)'S^{-1}\left(\underset{\sim}{x}-\underset{\sim}{\bar{x}}\right)\le\chi^2_{0.5,2}$ where we have replaced $\underset{\sim}{\mu}$ by it's estimate $\underset{\sim}{\bar{x}}$ and $\sum^{-1}$ by its estimate S^{-1}. If it does not hold the normality assumption is suspect. A formal method for judging the joint normality of a data set is based on the square generalized distances $d_j^2=\left(\underset{\sim}{x}_j-\underset{\sim}{\bar{x}}\right)'S^{-1}\left(\underset{\sim}{x}_j-\underset{\sim}{\bar{x}}\right)$, j = 1, 2,……., n. where $\underset{\sim}{x}_1, \underset{\sim}{x}_2, \ldots\ldots\ldots\ldots, \underset{\sim}{x}_n$ are the sample observations. To construct the chi-square plot we order the squared distances from smallest to largest as $d_2^1\le d_2^2\le\ldots\ldots\le d_n^2$ and graph the pairs $d_{(j)}^2, \chi^2_{p(j-0.05)/n}$, where $\chi^2_p\left(\left(j-\frac{1}{2}\right)/n\right)$ is the $100\left(j-\frac{1}{2}\right)/n$ percentile of the chi-square distribution with p d.f. The plot should resemble a straight line. A systematic curved pattern suggests lack of normality.

Transformation of data for normality

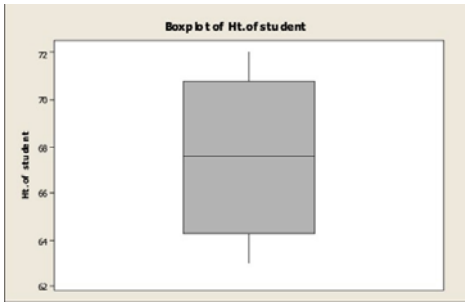
When a variable does not conform to the assumption of normality, but the data analysis method requires the data to come from a normal distribution, then it is advisable to try a transformation. The logarithm of the variate or the square root or the inverse sine or the reciprocal of the variate, may be more nearly normal.If normality is not a viable assumption, one alternative is to ignore the findings of a normality check and proceed as if the data were normally distributed. This practice is not recommended since, in many instance, it could lead to incorrect conclusions. A second alternative is to make non-normal data more "normal looking" by considering transformations of the data. Normal theory analyses can then be carried out with the suitably transformed data.In many instances the choice of

Table 1:Descriptive statistics for 6 univariate data sets

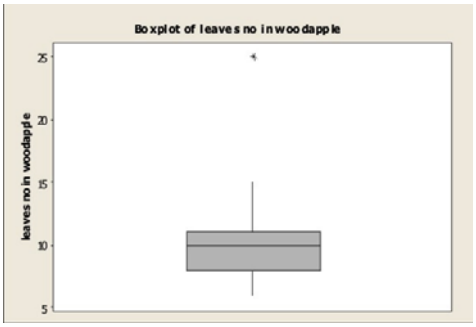
Data set no.	Mean $(\bar{x})$	Min $(x_{(1)})$	Max (x_n)	Range (R)	***St.dev* (S)**	Q_1	Q_2	Q_3	β_1	β_2	$\sqrt{\beta_1}=\gamma_1$	$\beta_2-3=\gamma_2$	***IQR***
1	48.8	10	151	141	52.2	18	24	89.5	1.96	3.19	1.40	0.19	71.5
2	67.667	63	72	9	3.284	64.25	67.5	70.750	0	1.38	0.00	-1.62	6.5
3	10.444	6	25	19	3.662	8	10	11	6.76	12.22	2.60	9.22	3
4	19211	6	77700	77694	20534	2212	14620	24459	1.7956	4.18	1.34	1.18	22247
5	42.49	33.080	54.090	21.010	5.391	38.890	41.995	46.295	0.0169	2.58	0.13	-0.42	7.405
6	435.8	229.8	892.1	662.4	122.7	345.5	229.8	492.9	1.4884	5.12	1.22	2.12	147.4



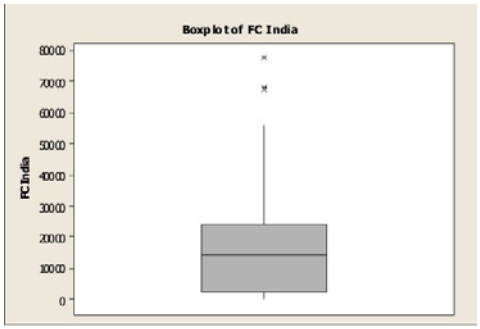
Data set-1



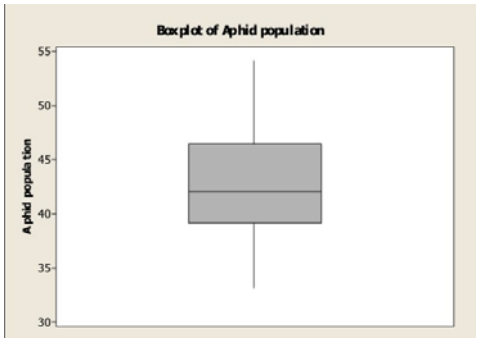
Data set-2



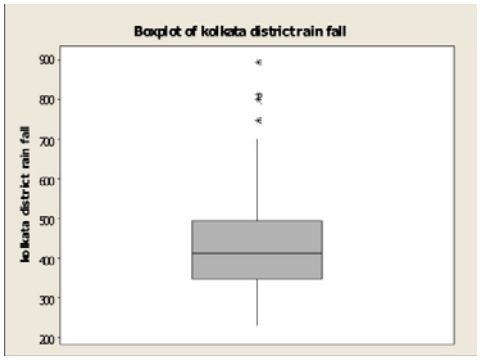
Data set-3



Data set-4

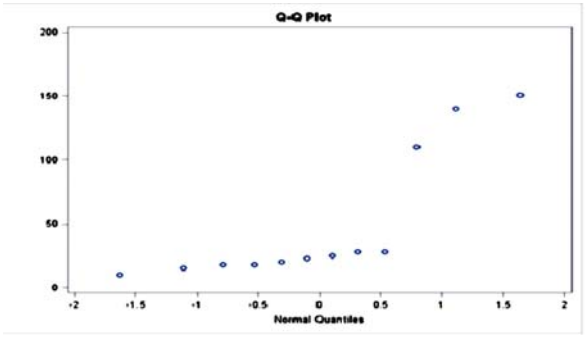


Data set-5

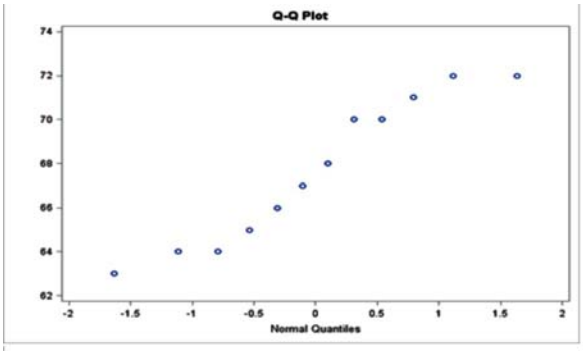


Data set-6

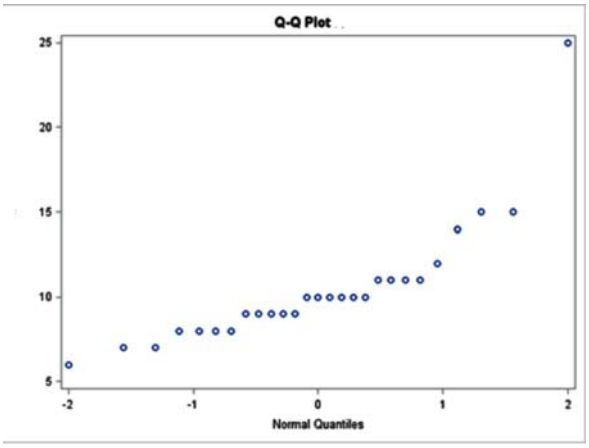
Fig.1. Box-Whisker plots for six univariate data



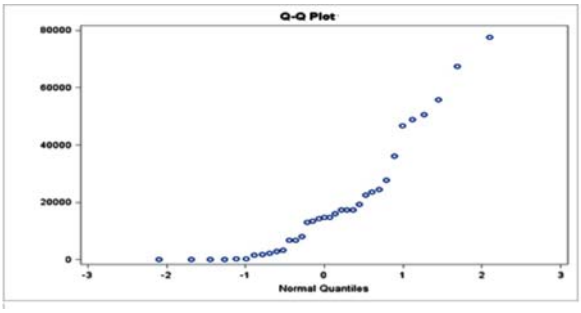
Data set-3 (r_Q = 0.82623)



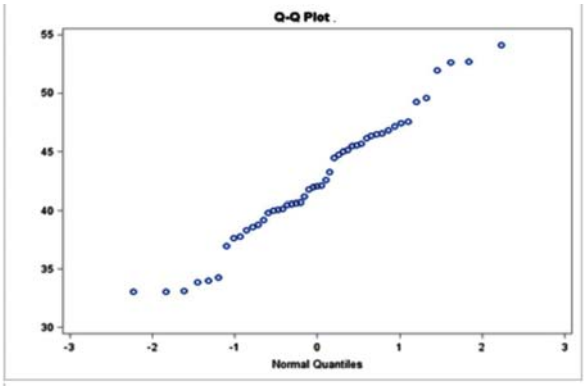
Data set-2 (r_Q = 0.96628)



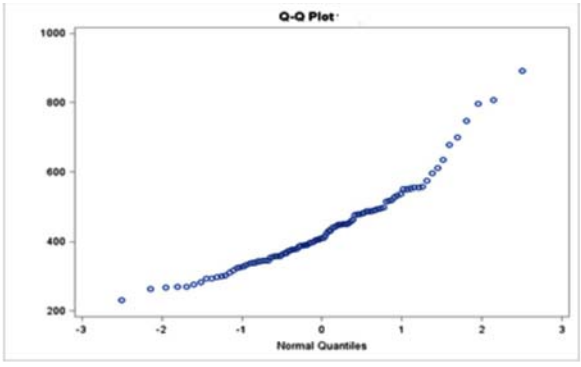
Data set-3 (r_Q = 0.8567)



Data set-4 (r_Q = 0.9158)



Data set-5 (r_Q = 0.989507)



Data set-6 (r_Q = 0.95978)

Fig. 2 : Q-Q Plots and the correlation coefficients for the Q-Q plots for six univariate data.

transformation to improve the approximation to normality is not obvious. For such cases it is convenient to let the data suggest a transformation. A useful family of transformation for this purpose is the family of power transformation.

Power transformation

For detailed theory on power transformation any standard book may be followed.

RESULTS AND DICUSSIONS

The descriptive statistics of First 6 univariate (data sets) are calculated and presented in the table 1. A common rule – of- thumb for test of normality is to run descriptive statistics to get skewness and kurtosis, and then divide these by the standard errors. Skewnsee should be within the range +2 to -2 (some authors used +1 to -1 as a more stringent criterion when normality is critical) and kurtosis within the range +2 to -2. A few authors use the more lenient +3 to -3 where other authors use +1 to -1 as a more stringent criterion (Garson, 2012). Using the thumb rule, it appears from table 1 that two of the data sets follow normal distribution having skewness in the range 0.00 to 0.13. The other data sets *i.e.* data sets 1, 4 & 6 were skewed to the right with skewness value ranging from 1.2 to 1.40 except for data set 3 where both skewness and kurtosis were higher. So it can be conclude that the data set 2 and 5 are normal.

Box-Whisker plots

From the Box-Whisker graphs (Fig. 1) it has been observed that only in case of data set-2 and data set-5, the distance between median to first-quartile and median to third-quartile are nearly same and also the distance between median to lower value and median to upper value are also same. So we can conclude that, these two data sets are normal. Other data sets don't follow the above criteria. Hence we can conclude that, other data sets (*i.e.* data sets 1, 3, 4 & 6) are not normal.

Any observation falling beyond $(Q_1 - 1.51QR,\ Q_3 + 1.5QR)$ is identified as a suspect outlier and any value falling outside $(Q_1 - 3QR,\ Q_3 + 3QR)$ is known as a extreme outlier. It is also observed that outliers are observed in data set 1, 3, 4 and 6. If outliers are identified, they should be examined very carefully. Depending upon the nature of outlier and the objectives of the investigation, outliers may be deleted or appropriately weighted in a subsequent analysis.

Q-Q Plots

Q-Q Plots and the correlation coefficients for the Q-Q plots are presented in figure 2 for the above mentioned six univariate data sets.

The Q-Q plot for the data set-1, which is a plot of the ordered data $x_{(j)}$ against the normal quantiles $q_{(j)}$, is shown in fig. 2. The pairs $(q_{(j)}, x_{(j)})$ do not lie in a straight line. It appears from the fig. 2 that the data as a whole are not normally distributed. The straightness of the Q-Q plot can also be measured by calculating the correlation coefficient of the points in the plot. We have $r_Q = 0.82623$. For n=12 and $\alpha = 0.05$ the table value of $r_Q = 0.92744$. Since $r_Q = 0.82623 < 0.92744$, we can reject the hypothesis of normality. The Q-Q plot for the data set-2, lies nearly straight line and it appears in the fig. 2 for data set 2. So, we can conclude that it is normal. The correlation coefficient value is also greater then table value. So, we can also conclude that the data follow normal distribution. Similarly from the fig. 1 for data set 5, the pairs of points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line and we would not reject the notion that the data set-5 is normally distributed. Also $r_Q = 0.989507 > 0.97642$ which is the critical point for the Q-Q plot correlation coefficient test for normality for n = 49 and $\alpha = 0.05$. It is also observed from the fig. 1 for the data sets 3, 4, 6 that the pairs of points $(q_{(j)}, x_{(j)})$ do not lie along a straight line and we reject the normality assumption for the data sets 3,4 and 6. Here also the calculated values of r_Q are less than table values for $\alpha = 0.05$ and corresponding values of n. Thus the hypothesis of normality is rejected.

Shapiro and Wilk (W-test) test for normality

In case of dataset 1, the tabulated value of W-test for normality for n =12 and alfa = 0.05 is 0.859 which is greater than the calculated value of W *i.e.* 0.6794. So it can conclude that the hypothesis of normality is rejected.Similarly the values of W-test statistics for data sets 2 to 5 are given in table 2.

Comparing the calculated values of W with table values of W at 5 per cent level it is shown that data sets 2 and 5 are normal and data sets 3 and 4 are non-normal. Since the sample sizes for the data sets 6 is 102, W-test

Table 2: W-test statistics for data sets 2 to 5.

	Data set-2	Data set-3	Data set-4	Data set-5
W-statistic	0.9196	0.754	0.836	0.974
Critical value	0.859	0.923	0.934	0.947

Table 3: χ^2 test for goodness of fit for data set 6.

C.I(m)	f	x(upper class boundary)	$u = \frac{x - \bar{x}}{s}$	$\Phi(u)$	Probability $= \Delta\ \Phi(u)$ (p)	e =np Expected frrequency	$\frac{f^2}{e}$
200-275	5	275	-1.274	0.102	0.102	10.408	2.402
275-350	21	350	-0.667	0.255	0.153	15.564	28.33
350-425	28	425	-0.060	0.480	0.225	22.994	34.09
425-500	26	500	0.548	0.705	0.225	22.985	29.41
500-575	13	575	1.155	0.875	0.170	17.292	9.773
575-650	3	650	1.762	0.961	0.086	8.759	1.028
650-725	2	725	2.370	0.991	0.030	3.067	1.304
725-800	2	800	2.977	0.999	0.008	0.780	5.127
800-875	1	875	3.585	1.000	0.001	0.135	7.406
875-900	1	900	3.787	1.000	0.0002	0.017	59.31
					1.000	102	110.36

Table 4: Values of d_j^2 for the data set 7.

Obs.	x_{1j}	x_{2j}	d_j^2	Obs.	x_{1j}	x_{2j}	d_j^2
1	17.2	4.4	4.932	16	9.1	2.8	1.163
2	13.5	4.2	1.182	17	7.9	2.6	2.273
3	12.3	3.4	0.187	18	7.3	4.8	3.845
4	10.8	4	0.299	19	13.7	5.3	4.871
5	8.8	3.7	0.619	20	12.3	3.3	0.243
6	6.2	4.5	3.972	21	9.72	2	3.433
7	15.7	3	2.754	22	9.42	2.5	1.765
8	13.7	2.5	2.285	23	8.82	4.7	2.462
9	10.8	3.7	0.045	24	7.98	4.7	2.976
10	10.5	3.5	0.042	25	17.5	3.1	4.775
11	9.2	2.2	2.817	26	13.5	3.9	0.78
12	7.1	2.2	4.105	27	11.7	4.2	0.622
13	15.7	3.6	2.293	28	10.8	3.3	0.09
14	13.7	3.4	0.768	29	9.72	3.5	0.21
15	10.8	3.1	0.277	30	7.68	4.2	1.915

is not applied. We know that W-test can be effective when the sample size is small.

Jarque – Bera (JB) test of normality

The JB test of normality is an asymptotic or large sample test. We can apply this test on only data sets 6.

The test statistic under, $H_0 : \beta_1 = 0, \beta_2 = 3$, is given by

$$JB = n\left[\frac{\left(\sqrt{\beta_1}\right)^2}{6} + \frac{(\beta_2 - 3)^2}{24}\right]$$

Using table 1 the calculated test statistic for Jarque – Bera (JB) is 44.24. The table value of χ^2 at 5 per cent level of significance with 2 d.f. is 5.991. Since the JB values are greater than 5.991, we reject H_0. So it can be conclude that the observations in the data set 6 are not normally distributed.

χ^2 test for goodness of fit

This test is applicable only when sample size is large. At first we can construct the grouped frequency distribution tables for the data sets 6. Remembering that in fitting the normal distribution, two parameters has to be estimated from the sample. It is found from the table

Table 5: Values of d_j^2 for the data set 8.

Obs.	x_{1j}	x_{2j}		Obs.	x_{1j}	x_{2j}	
1	37.6	98.5	3.301	34	5.11	178	1.302
2	26	190.9	1.348	35	4.41	129	0.294
3	16.4	104.7	0.328	36	3.25	70.8	1.17
4	12.3	67.7	1.3	37	7.68	231	3.956
5	4.8	46.2	2.179	38	4.46	178	1.367
6	3.2	43.1	2.357	39	2.68	157	0.875
7	47.4	36.9	9.969	40	2.14	144	0.651
8	23.3	73.9	1.912	41	1.61	117	0.477
9	10	126	0.035	42	0.89	104	0.599
10	8.9	30.8	3.138	43	8.39	237	4.319
11	6.9	24.6	3.483	44	2.86	157	0.857
12	5.1	21.5	3.656	45	2.5	126	0.422
13	59.4	240	8.985	46	2.14	113	0.445
14	55.6	227	7.425	47	2.14	104	0.508
15	53.9	163	6.044	48	2.68	77	1
16	52.4	157	5.649	49	2.78	184	1.803
17	39.8	144	2.645	50	2.32	175	1.489
18	38.9	123	2.81	51	2.09	150	0.771
19	15	181	0.818	52	1.86	129	0.494
20	6.25	166	0.796	53	2.78	117	0.387
21	2.68	150	0.712	54	18.57	101	0.527
22	2.5	126	0.422	55	9.05	181	1.086
23	2.5	104	0.483	56	5.8	175	1.126
24	2.32	95.4	0.608	57	4.87	166	0.914
25	49.4	209	5.305	58	4.87	135	0.305
26	31.2	181	1.519	59	3.5	117	0.336
27	11.9	117	0.055	60	3	55.4	1.771
28	17.4	64.6	1.746	61	4.87	209	2.841
29	4.64	55.4	1.738	62	4.64	175	1.236
30	1.79	43.1	2.388	63	4.41	150	0.556
31	56.1	227	7.561	64	4.18	113	0.309
32	10.91	187	1.2	65	4.18	86.2	0.696
33	7.66	181	1.189	66	3.95	55.4	1.75

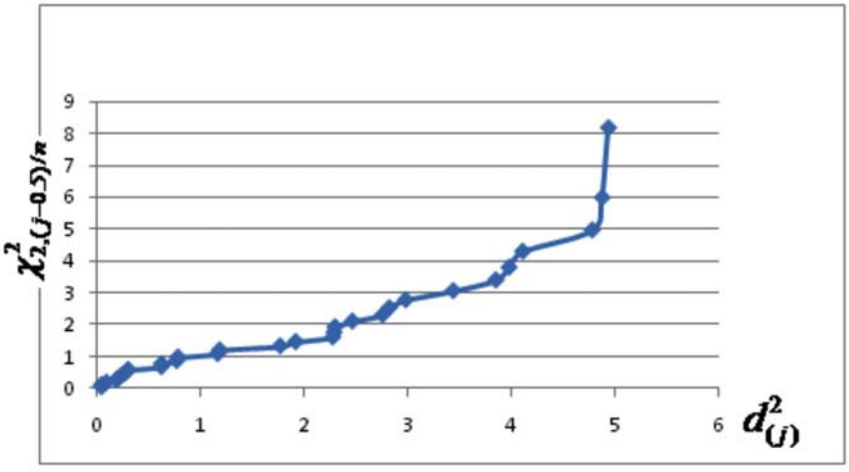


Fig. 3. The chi-square plot of the pairs $(d^2_{(j)}, \chi^2_{2,(j-0.5)/n})$ for the data set

3 that the last 5 class intervals of the variable have been amalgamated to form only one interval. Thus, the calculated χ^2 test statistic is 8.3663 with 3 d.f. which is greater than tabulated $\chi^2_{0.05,3}$. Hence the null hypothesis is rejected. So we can conclude that the data has not come from a normal population.

Normality test for bivariate data

The data sets 7 and 8 are bivariate. The tabulated $\chi 2$ value at 5 per cent level of significance and 2 d.f. is 1.39. Any observation satisfying the condition $d_j^2 \leq 1.39$ is on or inside the estimated 50 per cent contour. Otherwise the observation is outside the contour. It is observed from the table 4 that 14 generalized distances (d_j^2) are less than 1.39 *i.e.* a proportion 46.67 per cent of the data falls within the 50 per cent contour. If these observation are normally distributed we would expected 50 per cent of the observations to be within this contour. So the deviation is not too much. Hence we can conclude that the bivariate data has come from a bivariate normal population.

To draw a chi-square plot at first the squared distances are ordered from smallest to largest and graph the pairs $(d^2_{(j)}, \chi^2_{2,(j-0.5)/n})$ where $\chi^2_{2,(j-0.5)/n}$ is the 100(j-0.5)/n percentile of the chi-square distribution with 2 d.f.

From the figure 3 it can be concluded that the plot is not very straight. However it is difficult to reach a definite conclusion by chi-square plot.

In case of dataset 8, It is observed (Table 5) that 39 generalized distances (d_j^2) are less than 1.39 *i.e.* a proportion 59.09 per cent of the data falls within the 50 per cent contour. If these observation are normally distributed we would expect 50 per cent of the observations to be within this contour. Here the deviation prevails *i.e.*the bivariate data has not come from a bivariate normal population.

To draw a chi-square plot at first the squared distances are ordered from smallest to largest and graph the pairs $(d^2_{(j)}, \chi^2_{2,(j-0.5)/n})$ where $\chi^2_{2,(j-0.5)/n}$ is the 100(j-0.5)/n percentile of the chi-square distribution with 2 d.f. The ordered distances and the corresponding chi-square percentiles for 2 d.f. and n=66 are given in figure 4.

Since from figure 4 it is found that the Chi-square plot is not very straight, the data do not appear to be bivariate normal.

Transformation of data and Test for Normality

From the above test it has been found that some data sets *i.e.* data set 1,3,4,6 & 8 are non-normal. In this section first we use Box-Cox transformation to the observation for data sets-1, 3, 4, 6 and 8. Then the transformed data are cheeked for normality. Since all the observation are positive, let us performed a power transformation of the data which will produce result that are more nearly normal. We restrict our attention to the family of power transformation. We find that value of λ which maximizes the function $l(\lambda)$ .Using Statistica software we obtain the value of λ which maximizes for data sets 1,3,4 and 6, which are given in table 6.

Since for data set 1, r_Q value in Q-Q PLOT is 0.961914 which is greater than table value at 5 per cent level of significance indicates the acceptance of the hypothesis of normality. For Data set -3 the calculated Shapiro-Wilk (W) value is 0.950675 which is greater

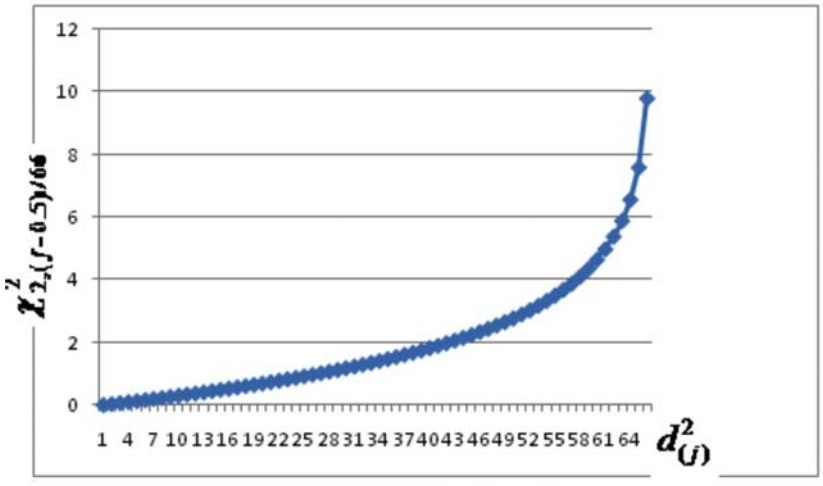


Fig.4. The chi-square plotsof the pairs $(d^2_{(j)}, \chi^2_{2,(j-0.5)/n})$ for the data set 8.

Table 6: Value of λ which maximizes *l* (λ) for data sets 1,3, 4, 6 and 8.

Parameter	Data set 1	Data set 3	Data set 4	Data set 6	data sets 8	
					X_1	X_2
λ	-0.60	-0.43	0.37	-0.31	0.89	

Table 7: Values of d_j^2 for transformed data set 8.

Obs.	y_1	y_2	d_j^2	Obs.	y_1	y_2	d_j^2
1	2.181	66.8	2.925	34	1.281	114.208	0.824
2	2.054	121.659	2.075	35	1.19	85.341	0.084
3	1.873	70.613	1.269	36	0.988	49.426	1.228
4	1.746	47.441	2.19	37	1.512	144.498	2.98
5	1.243	33.379	2.31	38	1.197	114.208	0.951
6	0.977	31.301	2.525	39	0.85	101.955	1.286
7	2.254	27.095	7.076	40	0.678	94.284	1.693
8	2.013	51.402	3.04	41	0.443	78.108	2.518
9	1.647	83.54	0.296	42	-0.119	70.183	6.534
10	1.589	22.883	4.086	43	1.559	147.878	3.312
11	1.454	18.51	4.232	44	0.897	101.955	1.119
12	1.28	16.282	4.167	45	0.798	83.54	0.96
13	2.32	149.564	5.033	46	0.678	75.68	1.364
14	2.301	142.24	4.339	47	0.678	70.183	1.385
15	2.292	105.472	2.683	48	0.85	53.369	1.291
16	2.284	101.955	2.61	49	0.877	117.68	2.086
17	2.199	94.284	2.166	50	0.741	112.467	2.312
18	2.192	81.734	2.353	51	0.659	97.833	1.916
19	1.835	115.946	1.191	52	0.565	85.341	1.988
20	1.399	107.226	0.412	53	0.877	78.108	0.678
21	0.85	97.833	1.121	54	1.924	68.34	1.569
22	0.798	83.54	0.96	55	1.598	115.946	0.835
23	0.798	70.183	0.962	56	1.356	112.467	0.662
24	0.741	64.884	1.233	57	1.252	107.226	0.525
25	2.266	132.025	3.485	58	1.252	88.93	0.054
26	2.118	115.946	2.09	59	1.039	78.108	0.3
27	1.731	78.108	0.579	60	0.932	39.466	1.934
28	1.897	45.447	2.97	61	1.252	132.025	2.125
29	1.222	39.466	1.782	62	1.222	112.467	0.815
30	0.533	31.301	3.574	63	1.19	97.833	0.281
31	2.304	142.24	4.351	64	1.156	75.68	0.155
32	1.69	119.412	1.101	65	1.156	59.159	0.599
33	1.511	115.946	0.795	66	1.119	39.466	1.774

than table value of W at 5 per cent level of significance. So, the hypothesis of normality can be accepted.

at 5% level of significance reveals the acceptance of the hypothesis of normality. JB test statistics value (0.02958) for data set 6 also conclude that the data has come from the normal population. Since the data set 8 is not bivariate normal, we select appropriate transformation for the marginal distribution than for the joint distribution. The values of λ which maximizes *l* (λ) for x_1 and x_2 are given in table 7.

The tabulated χ^2 value at 5 per cent level of significance and 2 degrees of freedom is found 1.39. For any observation if generalized distances (d_j^2) are less than 1.39 is consider to be inside or on the estimated 50 per cent contour, otherwise the observation is outside

the contour.It is observed from the Table 7 that 31 generalized distances (d_j^2) are less than 1.39 *i.e.* a proportion 46.97 per cent of the data falls within the 50 per cent contour. If these observations are normally distributed one should expect 50 per cent of the observations to be within this contour. So in this case the deviation is not too much. Hence it can conclude that the bivariate data has come from a bivariate normal population. In case of Data set 4 the calculated w value from W-test is 0.9513 which is greater than the tabulated value (0.934)

The following observations are made on the basis of empirical studies:

It is observed from the Box-Whisker plot for data set 1, the distances from median to first quartile and median to third quartile are not same. So, the data set-1 is not normal. But it also reveals from fig. 2 that the Q-Q plot do not lie in a straight line and the value of r_Q *i.e.* 0.82623 which is less than the table value of r_Q at 5 per cent level. So the data set 1 is not normal. Shapiro and Wilk W-test also reveals that the data set 1 is not normal. Box-Cox transformation is used for this data and r_Q value is calculated. The calculated r_Q value (0.961) is found much greater than the table value of at 5 per cent level indicates the normality of data. So it can be conclude that the Box-Cox transformation can be used to make the data normal.The calculated value of r_Q for data set 2 and 5 are much greater than table value of r_Q at 5 per cent level indicates the normality of data. It is also justified by Shapiro and Wilk W-test. The value r_Q and W-test statistic for dataset 3 and 4 shows that data are non-normal. Since, data sets 6 are large samples the JB test and χ^2 test for goodness of fit were used to test the normality and found that the data sets are non-normal. For data set 7, almost 47 per cent of the observations is on or inside the estimated 50 per cent contour but for data set-8, nearly 59 per cent of the data falls within the 50 per cent contour. So, it can be concluded that data set 7 is nearly normal, but the data set 8 is not normal. This result also reveals from χ^2 plots. Finally it can be summarized that the data sets 2 and 5 are normal. The other data sets 3, 4 and 6 are non-normal. The bivariate data set 7 is normal whereas bivariate data set-8 is non-normal. Box-Cox- power transformation is used for all non-normal data and it is found that all the transformed data follow normality.But it is not necessary that theBox-Cox- power transformation will always makes the data normal.

REFERENCES

Bartlett, M.S.1947: "The use of transformations". *Biometrics*, **3**, pp.39-52.

Box, G. E. P. and Cox, D. R. 1964. "An analysis of transformations". *J .Roy. Statist. Soc. B*,**26:** 211-52.

Fisher, R.A. 1958. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh, Thirteenth Edition, Section- 21, 4th Edition (1932).

Garson, G. D. 2012. "Testing statistical assumptions." Asheboro, NC: Statistical Associates Publishing.

Gun, A.M., Gupta, M. K. and Dasgupta, B. 1968. Fundamentals of Statistics, Vol.-**1**, World Press Private Limited, Kolkata.

Jarque, C. M. and Bera, A. K. 1987. "A test for normality of observations and regression residual". *Int. Stat.Rev*, **55**: 163-72.

Johnson, Richard A & Wichern, Dean W. 1992.Applied multivariate statistical analysis **.**Prentice-Hall, Inc.

Kendal, S.N. and Stuart, Alan 1977. The advanced theory of Statistics, Vol.1.

Quetelet A. 1842. "A Treatise on Man and the Development of his Faculties".Reprinted in1968 by Burt Franklin, New York.

Rao, C.R., Mitra, S.K., Matthai, A. and Ramamurthy, K.G. 1975. Formulae and Tables for Statistical Work, Statistical Publishing Society, Kolkata.

Sahu, P.K. 2010. Agriculture and Applied Statistics. Kalyani Publishers, Ludiana.

Shapiro, S. S. and Wilk, M. B. 1965. "An analysis of variance test for normality (complete samples)". *Biometrika*,**52**:591-611.