Transactions on Ecology and the Environment vol 6, © 1995 WIT Press, www.witpress.com, ISSN 1743-3541

The problem of assessing and achieving normality: an application to environmental data

J. Mateu
Departamento de Matemáticas, Escuela Superior de Tecnología y Ciencias Experimentales, Universidad Jaume I, Castellón, Spain

Abstract

It is well known that many statistical works in environmental analysis are based on methods that need normal variables. Data transformation methods capable of achieving normality of distributions have a crucial role in statistical analysis, especially towards an efficient application of techniques such as analysis of variance and multiple regression analysis.

Different techniques to test the normality of distributions have been studied in this paper. Since normal distribution has value zero in skewness and kurtosis, we develope confidence intervals over them to contrast their nullity. Moreover different nonparametric tests are applied such as Chi-square and K-S.

On the other hand we suggest Box-Cox transformations as a method to achieve normal variables.

1 Introduction

Box and Cox [1] suggested a technique for choosing a transformation of available data which could simultaneously achieve: (a) normality of distributions, (b) constancy of error variance or, equivalently, independence between cell mean and cell variance,i.e., between the sample mean and the sample variance of the observations in each experimental trial and (c) simplicity (linearity) of the model structure.

However, it is unreasonable to expect that, for any type of data, there will always be a transformation so that assumptions (a)-(c) are simultaneously satisfied. So, in many instances, the final choice of transformation will depend on which of the assumptions (a),(b) or (c) is considered most important (Perry [2]). We shall focus on condition (a) as it is an important condition towards an efficient application of techniques such as analysis of variance or multiple regression analysis. Normality is a basic assumption in many of the statistical methods used in the Environmental Sciences.

The computation work has been done using the statistical packages S-Plus for WorkStation and SPSS\PC for Windows.

2 Testing Normality

It is examined three complementary methods. None of them is conclusive. Usually it is used some (or all) of them to test normality.

2.1 Graphical methods

Since the normal distribution has a characteristic probability density function, it is plotted for each variable its histogram and density function just to compare how close are they to the normal one.

2.2. Confidence Intervals

It is known that the values of skewness and kurtosis for a normal distribution are zero. A non-zero value of skewness states that the distribution is concentrated on the right or on the left of the mean value (negative or positive skewness,respectively). A non-zero value of kurtosis is related to the flattenning of the distribution. So, we propose to make different contrasts about these values to see if they can be values of a normal distribution.

a) We can write the confidence interval for skewness ($I_{skewness}$) and the one for kurtosis ($I_{kurtosis}$) and if 0 belongs to two of them we can accept normality. Depending on what we are concerned we can accept normality if 0 belongs just to one of them.

$$I_{skewness} = \left]skewness \pm t_{1-\alpha/2}(s.e.\ skew.)\right[\qquad I_{kurtosis} = \left]kurtosis \pm t_{1-\alpha/2}(s.e.\ kurt.)\right[$$

where s.e.=standard error , 1-α=confidence level , t=value for T-Student distribution.

b) We compute the following values for skewness and kurtosis

$$Coef(skew.) = \left[(n-2)skewness\right] / \sqrt{n(n-1)}$$

$$Coef.(kurt.) = \left[kurtosis - \frac{n-1}{(n-2)(n-3)} \right] / \frac{(n+1)(n-1)}{(n-2)(n-3)}$$

where n=sampling size.

Looking at the statistical tables ¨for testing skewness and kurtosis¨ we accept normality if Coef.(skew.)$\in$[-0.534,0.534] and Coef.(kurt.)$\in$[-0.85,0.99] under the value $\alpha = 0.05$.

2.3 Non-parametric methods

Usually, it is used two different methods (Dixon [3], Whitney [4]).

a) *The χ^2 goodness-of-fit test.*(Cochran [5], Scheffé [6]).

A single random sample of size n is drawn from apopulation with unknown cumulative distribution function FX .We wish to test the null hypothesis

$$H_0 : F_X(x) = F_0(x) \ \ for\ all\ x$$

where F0(x) is the normal distribution function, against the general alternative

$$H_0 : F_X(x) \neq F_0(x) \ \ for\ some\ x$$

In order to apply the chi-square test, the sample data must first be grouped according to some scheme in order to form a frequency distribution. When the sample observations are quantitative, the categories would be numerical classes chosen by the experimenter. Even though the hypothesized distribution is most likely continuous, the data must be categorized for analysis. Assuming that the population distribution is completely specified by the null hypothesis, one can calculate the probability that a random observation will be classified into each of the chosen categories. These probabilities multiplied by n give the frequencies for each category which would be expected if the null hypothesis were true.

Assume that the n observations have been grouped into k mutually exclusive categories, and denote by fi and ei the observed and expected frequencies, respectively, for the ith group, i=1,...,k. The following statistic is then calculated

$$Q = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

The distribution of Q is chi-square with k-1 degrees of freedom.We reject the hypothesis of normal distribution if $Q > \chi^2{}_{k-1;1-\alpha/2}$ or the p-value of Q< α.

b) *The Kolmogorov-Smirnov test.*(Birnbaum [7], Massey [8]).

A random sample $X_1, X_2, \ldots, X_n$ is drawn from a population with unknown cumulative distribution function FX(x). For any value of x, the empirical distribution function of the sample, Sn(x), provides a consistent point estimate for FX(x).(Gibbons [9]). The Glivenko-Cantelli theorem (See page 75 Gibbons [9]) states that the step function Sn(x), with jumps occurring at the values of the order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ for the sample, approaches the true distribution function for all x. Therefore, for large n, the deviations between the true function and its statistical image, $|S_n(x) - F_X(x)|$, should be small for all values of x. This result suggests that the statistic

$$\mathrm{D_n} = \sup_x |S_n(x) - F_X(x)| \qquad \text{(K-S statistic)}$$

is, for any n, a reasonable measure of the accuracy of our estimate.

If the p-value of Dn $<\alpha$ then we can reject the normality of distribution. For values of Dn see page 431 of [7].

3 Box-Cox Transformations

Box and Cox [1], in an important paper, considered two classes of transformations: a single-parameter family indexed by λ and defined by

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & (\lambda \neq 0) \\ \log\ y & (\lambda = 0) \end{cases} \tag{1}$$

which hold for y > 0, and a two-parameter family indexed by $\lambda = (\lambda_1, \lambda_2)$

$$y^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & (\lambda_1 \neq 0) \\ \log(y+\lambda_2) & (\lambda_1 = 0) \end{cases} \tag{2}$$

which hold for y > $-\lambda_2$. Estimation of the parameter λ was discussed from a sampling theory and Bayesian point of view. The fundamental assumption made was that for some λ the transformed observations defined by (1) (or by equation (2) in the shifted location case) can be treated as independently normally distributed with constant variance $\sigma^2{}_\lambda$ and with expectations defined by a model of simple structure (linear). The cases for different λ's were made comparable by working with the normalized transformation

$$z^{(\lambda)} = y^{(\lambda)} / J^{1/n}$$

where J=J(λ ;y) is the Jacobian of the transformation defined by

$$J(\lambda; y) = \prod_{i=1}^{n} \left| \frac{dy_i^{(\lambda)}}{dy_i} \right|$$

The resulting normalized values were then expressed as (corresponding to equation (1) and (2) respectively)

$$z^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda(y^*)^{\lambda-1}} & (\lambda \neq 0) \\ y^*\ \log y & (\lambda = 0) \end{cases} \tag{1a}$$

where y^* is the geometric mean of the observations,or in the shifted location case

$$z^{(\lambda)} = \begin{cases} \dfrac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1 gm(y+\lambda_2)^{\lambda_1 -1}} & (\lambda_1 \neq 0) \\ gm(y+\lambda_2)\log(y+\lambda_2) & (\lambda_1 = 0) \end{cases} \tag{2a}$$

where $gm(y+\lambda_2)$ is the sample geometric mean of the $(y+\lambda_2)s$.

Invoking standard least squares theory, the maximized log-likelihood function with respect to λ can be found to be proportional to $S(\lambda;z)^{-n}$, where $S(\lambda;z)$ is the residual sum of squares of $z^{(\lambda)}$; in particular, except for a constant,

$$L_{max}(\lambda) = -\frac{1}{2}n\log\left\{\frac{S(\lambda;z)}{n}\right\}$$

and so the maximum likelihood estimate for λ is obtained by minimizing $S(\lambda;z)$ with respect to λ. In practice, in the majority of the cases, values for λ of 1/2 (square-root transformation), 0 (logarithmic transformation), -1 (inverse), 2 (square) or 1 (no transformation) are the common values to be used although any real value for λ is possible.

If α is a level of significance, a $100(1-\alpha)\%$ confidence interval can be found for λ by calculating a critical sum of squares SSc from

$$SS_c = S_{min}(\lambda;z)\left\{1+\frac{t^2{}_\nu(\alpha/2)}{\nu}\right\}$$

where $S_{min}(\lambda;z)$ is the minimum residual sum of squares with respect to λ, with the associated ν degrees of freedom and t_ν is the corresponding value from the t tables.

4 An Application and Results

In order to show the usefulness of these methods, a practical case with two variables was analysed. The variables are: *direc* (wind direction) and *SO2* (sulphur dioxide concentration measured in $\mu g/m^3$). Each of them has a sampling size of 365 data. We have chosen these variables because they are part of a regression analysis in a paper still unpublished. For each variable methods in sections 2 and 3 have been applied.

To compute a Box-Cox transformation it is used formula (1a) with values of λ ranging between -2 and 2 with a step of 0.25 (the step can be changed in practice). It is chosen that λ that makes formula (1a) have the least variance. Then the final transformation with formula (1) is computed and represented as b-c( ). And to compute the χ^2 goodness-of-fit test it is sufficient to group each variable into 4 categories with equal expected frequencies.

In tables it is specified in parenthesis if the method fails to accept normality (No) or (Yes) if it accepts it.

Results with direc

Results in table1 show that this variable is not normal at all for all the methods fail to support normality. The histogram in Fig.1 explains graphically what is happening. A Box-Cox transformation has been developed with results in table 2. The value of λ is 2. Now, normality can be accepted and if we see Fig.1 we can see how close is the histogram to the normal one.

Table 1. Results of normality for direc under different methods.

	$I_{skew.}$	$I_{kurt.}$	$Coef(skew.)$	$Coef(kurt.)$	$K-S(2-tailed\,prob.)$	$\chi^2\,(2-tailed\,prob.)$
direc	]−1.374,−0.872[	]1.318,2.318[	−1.1184	1.7904	4.32(0.0)	48.578(0.0)
	(*No*)	(*No*)	(*No*)	(*No*)	(*No*)	(*No*)

Table 2. Results of normality for b-c(direc) under different methods.

	$I_{skew.}$	$I_{kurt.}$	$Coef(skew.)$	$Coef(kurt.)$	$K-S(2-tailed\,prob.)$	$\chi^2\,(2-tailed\,prob.)$
$b-c$	]−0.224,0.285[	]0.20,1.199[	0.0298	0.6877	1.337(0.06)	5.59(0.133)
(*direc*)	(*Yes*)	(*No*)	(*Yes*)	(*Yes*)	(*Yes*)	(*Yes*)

where $b-c(direc) = \dfrac{(direc)^2 - 1}{2}$.

Results with SO2

This variable has normal values of kurtosis. Nevertheless, it can not be considered to have a normal distribution. See table 3. A Box-Cox transformation is applied with a value of λ of 0.5 giving the transformed variable $b-c(SO2) = 2\sqrt{SO2} - 2$. Normality can be seen graphically in the histogram of b-c(SO2) in Fig.2 and tested in table 4.

5 Conclusions

We claim the simplicity in practice of the Box-Cox transformation despite its mathematical background.

The acceptance or rejection of the hypothesis of normality of distribution is up to the experimenter depending on the results of the different methods applied. That is why it is recommended to use all (or most) of them and base our conclussions on them.

It is important to state that we can accept normality even though some method says (No). Finally say that in this paper we have worked with a statistical significance level of 0.05 and, of course, this value can be changed.

Table 3. Results of normality for SO2 under different methods.

	$I_{skew.}$	$I_{kurt.}$	$Coef\,(skew.)$	$Coef\,(kurt.)$	$K-S(2-tailed\,prob.)$	$\chi^2\,(2-tailed\,prob.)$
$SO2$	$]0.525, 1.035[$ (No)	$]-0.119, 0.881[$ (Yes)	0.776 (No)	0.373 (Yes)	$1.64(0.009)$ (No)	$19.94(0.0)$ (No)

Table 4. Results of normality for b-c(SO2) under different methods.

	$I_{skew.}$	$I_{kurt.}$	$Coef\,(skew.)$	$Coef\,(kurt.)$	$K-S(2-tailed\,prob.)$	$\chi^2\,(2-tailed\,prob.)$
$b-c(SO2)$	$]-0.365, 0.145[$ (Yes)	$]-1.01, 0.02[$ (Yes)	-0.1095 (Yes)	-0.5156 (Yes)	$0.812(0.525)$ (Yes)	$2.529(0.470)$ (Yes)

References

1. Box, G.E.P. & Cox, D.R. An analysis of transformations (with discussion), *J.R.Statist.Soc. B*, 1964, **26**, 211-246.

2. Perry, J.N. Iterative improvement of a power transformation to stabilise variance, *Appl. Statist.*, 1987, **36,** 15-21.

3. Dixon, W.J. Power under normality of several non-parametric tests, *Ann. Math. Statist.*, 1954, **25,** 610-614.

4. Whitney, D.R. A comparison of the power of non-parametric tests and tests based on the normal distribution under non-normal alternatives, Unpublished doctor's dissertation, Ohio State Univ. 1948.

5. Cochran, W.G. The χ^2 test of goodness of fit, *Ann. Math. Statist.*, 1952, **23**, 315-345.

6. Scheffé, H. Statistical inference in the non-parametric case, *Ann. Math. Statist.*, 1943, **14**, 305-332.

7. Birnbaum, Z.W. Numerical Tabulation of the distribution of Kolmogorov's statistic for finite sample values, *J. Amer. Statist. Ass.*, 1952, **47**, 425-441.

8. Massey, Jr., F.J. The Kolmogorov-Smirnov test for goodness of fit, *J. Amer. Statist. Ass.*, 1951a, **46**, 68-78.

9. Gibbons, J.D. *Nonparametric Statistical Inference,* Second Edition. Vol.65. Marcel Dekker, 1985.

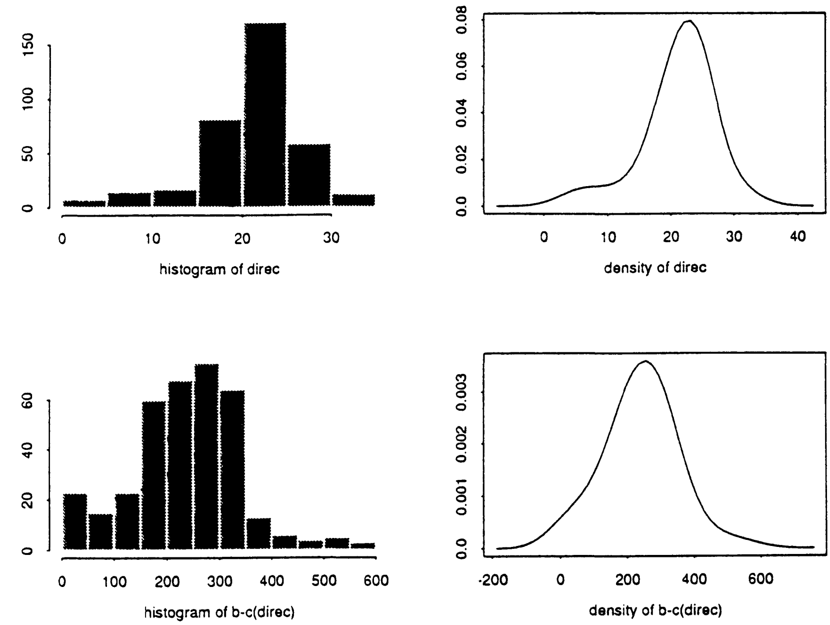

Figure 1: *Upper plots*: histogram and density function of variable "wind direction". Lower plots: histogram and density function of the transformed "wind direction" by Box-Cox (b-c(direc)).

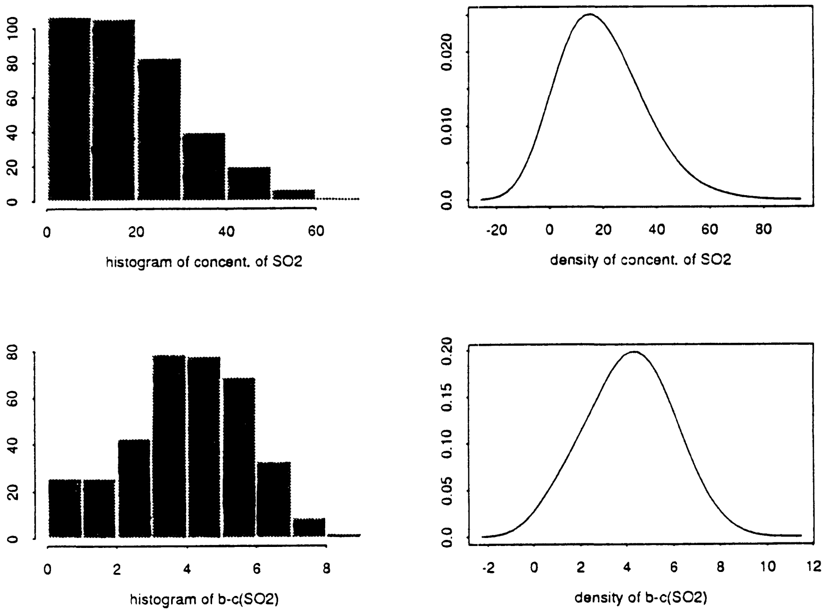

Figure 2: *Upper plots*: histogram and density function of variable "sulphur dioxide concentration". Lower plots: histogram and density function of the transformed "SO2" by Box-Cox (b-c(SO2)).