SAMPLE DISTRIBUTION OF AGRICULTURAL STATISTICS: A CASE STUDY IN SOUTHEASTERN BRAZIL

Francisco Alberto PINO[1]

- ABSTRACT: The probability distributions of some agricultural variables, namely planted or cultivated area, number of plants, plant stand (number of plants per hectare), production and productivity or yield (production per hectare), were studied and characterized, using census data. The hypothesis of normality was rejected for all variables and all cultivations. The main transformation to normality showed to be the logarithmic, followed by root transformations. An archetypal distribution was suggested for planted area: defined for non-negative values, heavy tail at right, mean > median > mode. It is shown to be relevant for planted area itself (91% of the cases), number of plants (87%) and production (85%), a little less for plant stand (60%) and yield (48%).

- KEYWORDS: Test for normality; Box-Cox transformation; skewness and kurtosis; agricultural data.

1 Introduction

Assuming a positivist approach, the scientific method comprises, among other points, the formulation of scientific hypotheses, testable predictions, data collection (observations not necessarily quantitative) and hypotheses testing. Although a scientific hypothesis must not be a statistical hypothesis, nor the data and the tests must be statistical, it is true that Statistics, as a branch of Mathematics, has been developed essentially to deal with scientific method (it can be used also in practical applications). Therefore, currently, natural science (biological and physical) use statistical methods to obtain data (experimentally or surveying an existing situation) and to analyze data (mainly through parameter estimation, modelling and hypothesis testing).

Any statistical method is based on a set of statements or conditions that are imagined to be true (assumptions, postulates, axioms). If one of these assumptions is not true, such method should not be used, because it will lead to erroneous results, invalid conclusions and spurious interpretations. This has to do with the rudiments of statistical methods use and, by extension, the use of the scientific method.

However, not always a scientist has enough knowledge to deal with such question of invalid assumptions. Moreover, sometimes it is not a scientist, but a technician working in a company who need to get results, who cannot afford to invest a lot of time to resolve the question of an unsatisfied premise. Thus, there are some papers in the literature showing

[1] Retired scientific researcher - Institute of Agricultural Economics, São Paulo, SP, Brazil, E-mail: *drfapino@gmail.com*

more or less general solutions for what to do when a given assumption is not satisfied. This greatly facilitates the work of scientists and applied technicians.

To know the shape of a random variable probability distribution is essential in the maximum likelihood estimation. The assumption that this distribution is normal (or Gaussian) is required in the estimation of parameters by least squares method, and it is fundamental in some hypothesis tests (such as the Student t test and *F* test), in the analysis of variance, and when comparing groups for statistical differences. Non-normality is also associated to the asymmetry of distribution when the location measurements (mean, median and mode) fail to coincide, as well as in the cases of heteroscedasticity or lack of homogeneity of variances.

Attempts to explain the subject to nonspecialists appear in Indrayan (2014) and Lee (2007). Non-normality is showed to be a relevant concern in biological and medical studies data[2], as well as in other fields, like engineering, cosmology, meteorology and geology[3].

Agricultural variables constitute an important field where normal distribution may not be true, due to restrictions on the values that observations can assume: a) planted or cultivated area (also known as acreage) is positive real; b) agricultural production (also known as output) and productivity or yield (production per hectare) are non-negative real; c) number of plants, as well as plant stand (number of plants per hectare), are positive integers. Furthermore, two of these variables are defined as the ratio of two of the others, namely:

a) The productivity or field crop yield is calculated as the production (output, yield) divided by the planted area;
b) The plant stand is calculated as the number of plants divided by the planted area and it depends on the plant spacing.

If the production and the cultivated area are normal, then the productivity will follow a Cauchy distribution, with infinite mean and variance, which is unacceptable for productivity. Similarly, if the number of plants and the cultivated area are normal, then the plant stand will follow a Cauchy distribution, which is also unacceptable. In summa, if two of the variables in a ratio are normal, the third one will not be normal.

The non-normality of agricultural productivity or field crop yield (production per hectare) and related statistics has been tested and eventually demonstrated in different situations by several authors[4], using data on a small number of agricultural products, like cotton, corn, oats, soybeans and potatoes yields, as well as milk production. Three classical studies established the basis for discussion. Day (1965) showed that in general: a) the distribution of field crop yields are non-normal nor lognormal; b) "the degree of skewness and kurtosis depends upon the specific crop and on the amount of available nutrients"; c) "mode or median estimates of yields may be preferred to mean estimates".

[2] As seen in BAKER *et al*. 1987; BERNIER *et al*., 2011; BRITTON, 1989; BULMER, 1974; CRAWFORD *et al*., 2006; DeBOER *et al*., 2009; DURAZO-ARVIZU *et al*., 1998; DUTY *et al*., 2005; GILES and KIPLING, 2003; GUAN *et al*., 2012; LACOURCIÈRE *et al*., 2000; NUSSER *et al*., 1996; OMARIBA, 2011; POLLOCK *et al*., 1990; ROBERTS *et al*., 2009; STROWIG *et al*., 2002; SWARTZ *et al*., 2008.

[3] See, for example, DINEEN and COLES (2008); GABRIEL and FEDER, 1969; HONG, 1998; KATZ and PARLANGE, 1998; McGRATH *et al*., 2004.

[4] ANTLE and GOODGE (1984); BLYTHE and MERHAUT (2007); BRAH *et al*. (1982); HARRAR and GUPTA (2007); MOSS (2015); ZHU *et al*. (2011).

According to Taylor (1984): a) “there appear to be few cases in agriculture where one can appeal to the Central Limit Theorem in order to theoretically justify a normal distribution”; b) generally, we cannot use theoretical arguments to establish that variables like weather, crop yield, gross returns, or equipment failure follow a common distribution like lognormal, Gamma, Beta, Poisson or normal; c) “with many agricultural relationships an independent variable may influence the parameters of the probability distribution function of a dependent variable”, but not its functional form, e.g., “the fertilization rate may influence the moments of a crop yield” distribution, but the form of the distribution may not be conditional on the fertilization. Finally, Nelson and Preckel (1989) proposed the conditional beta distribution “as a parametric model of the probability distribution of agricultural output”[5]. According to the authors, “this distribution is consistent with agronomic models of field crop production, and it is supported by previous research on the distribution of field crop yields”. The beta distribution allows a flexibility to model the fact that field crop yield varies from zero to a given maximum value, depending on the plant genetic potential, as well as to allow the yield distribution to be right or left skewed.

Several procedures have been used or suggested to deal with non-normality and related problems in agricultural studies, such as arc sine transformation, inverse hyperbolic sine transformation, hyperbolic tangent transformation, logarithmic transformation, non-linear Bayesian models, and non-parametric procedures[6]. These examples represent a small sample of non-normality cases in different scientific fields, particularly in agronomical issues.

1.1 Objectives

The general purpose of this paper is the study and characterization of probability distributions of the following agronomical variables: cultivated area, plant stand, number of plants, production and crop yield. The specific aim is to apply the non-normality tests and to find the best lambda for Box-Cox transformation over farm data on a set of many different crops.

2 Material and Methods

Data from an agricultural census were used rather than from a sample survey, to avoid sampling errors and the complexity of sampling schemes[7]. The first agricultural census in the 21st century in the state of Sao Paulo, Brazil, described by Torres et al. (2009), provided the data about 129 crops or crop groups, each one with at least 50 observations[8].

[5] Other possibilities are the log-normal distribution (DAY, 1965), the Weibull distribution (CHEN and MIRANDA, 2004; OZAKI *et al*., 2010) and the logistic distribution (ZANINI et al., 2001).

[6] JUNQUEIRA *et al*. (1982); KLEIN *et al*. (2003); MOSS and SHONKWILER (1993); PINO *et al*. (1979); TAYLOR (1984); VAN RADEN (2006); VERCHOT *et al*. (2000).

[7] An earlier unpublished report showed the non-normality of the same kind of variables, using goodness-of-fit chi-square tests applied over data from a stratified sample survey (PINO, 1979).

[8] For sample size 30 or less, the powers of the tests at 5% significance level are less than 40% (RAZALI and WAH, 2010).

The null hypothesis of normality was tested by Shapiro-Wilk, Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling tests[9]. The best value for Box-Cox transformation lambda parameter (BOX and COX, 1964) may be chosen by a maximum likelihood criterion (DRAPER and SMITH, 1981, p. 225-226; SAS, 2008). Eventually, the most convenient value may be used, provided that this value is within the confidence interval. Conventionally, we adopted the following decreasing order of convenience (SAS, 2008): $\lambda = 1.0$; $\lambda = 0.0$; $\lambda = 0.5$; $\lambda = -1.0$; $\lambda = -0.5$; $\lambda = 2.0$; $\lambda = -2.0$; $\lambda = 3.0$; $\lambda = -3.0$. Calculations were done by procedures UNIVARIATE and TRANSREG from SAS® – Statistical Analysis Software (SAS, 2008, 2010; LaLONDE, 2012).

3 Results

Cultivated area data were available for every crop, but production and yield data were not available for horticulture crops, crop groups and plant nurseries; moreover, number of plants and plant stand were available only for perennial crops.

A warning on the mode estimated values: when observed data distribution present more than one maximum peak, the computer software takes the first one as being the mode, resulting low values. Therefore, in this paper mode estimates were considered valid only when they have approximately the same order as the mean and the median.

The hypothesis of normality was rejected at 1% significance level in all the tests, for all the variables and all the crops with 50 or more observations[10]. The best (or more convenient) value for Box-Cox transformation λ parameter resulted to be different for each variable or culture (Tables 1, 2, 3, 4): the value was $\lambda = 0$ for cultivated area (53% of crops) and number of plants (47%), but $\lambda > 0$ for productivity or yield (87%), production (77%), and plant stand (40%).

The values of λ were concentrated in a narrow interval around zero for cultivated area, between –0.1 and 0.1, and a little more spread for plant stand, between –0.3 and 0.3 (Figure 1). On the other hand, the values of λ were concentrated in a positive interval for number of plants, between 0.0 and 0.2, for production, between 0.1 and 0.3, and for productivity or yield, between 0.25 and 0.75.

[9] A theoretical review of non-normality is presented in Pino (2014).

[10] The p-value for all variables and crops in this study were the following: <0.0100 (Kolmogorov-Smirnov), <0.0050 (Anderson-Darling), <0.0050 (Cramér-von Mises) and <0.0001 (Shapiro-Wilk). Different values were obtained only for crops with less than 50 observations, but they were discarded in this paper.

Table 1 - Best/more convenient value of lambda in Box-Cox transformation to normality, for each variable and crop, state of Sao Paulo, Brazil, 2007/08

Species	Crop	Cultivated area	Plant stand	Number of plants	Production	Yield
Agaricus spp.	Mushroom	0.0	...	...	...	...
Allium cepa L.	Onion	0.1	...	...	0.2	1.00
Allium fistulosum L.	Scallion	–0.3	...	...	...	...
Allium sativum L.	Garlic	–0.2	...	...	...	...
Anacardium occidentale L.	Cashew	0.4/0.5	0.0	0.2	0.2	0.25
Ananas comosus (L.) Merr.	Pineapple	0.1	...	...	0.2	0.25
Annona spp.	Custard apple	0.1	0.3	0.1/0.0	0.3	0.50
Arachis hypogaea L.	Peanut	0.1	...	...	0.1	0.50
Araucaria angustifolia (Bertol.) Kuntze	Parana pine	–0.2	-0.1	-0.2	0.2	0.25
Arracacia xanthorrhiza Bancr.	White carrot	0.0	...	...	0.2	0.50
Avena sativa L.	Oats	0.0	...	...	0.1	0.25
Averrhoa carambola L.	Star-fruit	0.3	0.0	0.6/0.5	0.3	0.50
Bactris gasipaes Kunth	Peach-palm	0.0	...	...	0.1	0.25
Bambuseae Kunth ex Dumort (tribo)	Bamboo	–0.1/0.0	0.0	0.0	...	...
Beta vulgaris L.	Swiss chard	0.0	...	...	...	...
Beta vulgaris L. subsp. *vulgaris*	Beet	0.0	...	...	...	...
Bixa orellana L.	Annatto	0.4	0.1/0.0	0.3	0.2	0.25
Brachiaria spp.	Brachiaria (signalgrass)	0.0	...	...	...	...
Brassica oleracea L. *var. acephala*	Kale	–0.1	...	...	...	...
Brassica oleracea L. *var. botrytis cauliflora*	Cauliflower	0.0	...	...	...	...
Brassica oleracea L., *var. italica*	Broccoli	0.0	...	...	...	...
Brassica oleracea L.*var. capitata*	Cabbage	0.0	...	...	0.2	0.50
Capsicum annuum L.	Sweet pepper	–0.2	...	...	–0.2	0.50
Capsicum spp.	Pepper	–0.2	...	...	...	...
Carica papaya L.	Papaya	0.1/0.0	...	...	0.1/0.0	0.00
Cichorium endivia L.	Endive	–0.2	...	...	...	...
Citrullus lanatus (Thunb.) Matsum. & Nakai	Watermelon	0.2	...	...	0.2	0.25
Citrus aurantifolia (Christm. et Panz.) Swingle	Sweet lime	0.0	-0.1/0.0	-0.1/0.0	0.2	0.50
Citrus aurantium L.	Bitter orange (sour orange)	0.0	-0.2	0.1/0.0	0.1	0.25
Citrus hybridum	Tangor	0.0	0.0	-0.1/0.0	0.2	0.75
Citrus reticulata Blanco	Tangerine	0.0	-0.3	0.0	0.2	0.50
Citrus sinensis Pers	Orange	0.0	-0.3	0.0	0.2	0.75
Citrus spp.	Lemon	0.1	0.1	0.1	0.3	0.50
Cocos nucifera L.	Coconut	0.0	0.1/0.0	0.0	...	...
Coffea spp.	Coffee	0.0	0.1	0.0	0.2	0.50
Colocasia antiquorum Schott	Yam	0.1	...	...	0.2	1.00
Crotalaria juncea L.	Brown hemp (Indian hemp, Madras hemp, or sunn hemp)	0.0	...	...	...	...
Cucumis anguria L.	Bur cucumber (West Indian gourd)	–0.1/0.0	...	...	...	...
Cucumis sativus L.	Cucumber	–0.2	...	...	–0.2	–0.25
Cucurbita spp.	Pumpkin	–0.1	...	...	0.0	0.25

Table 2 - Best/more convenient value of lambda in Box-Cox transformation to normality, for each variable and crop, state of Sao Paulo, Brazil, 2007/08

Species	Crop	Cultivated area	Plant stand	Number of plants	Production	Yield
Cynara cardunculus L. *var. scolymus*	Artichoke	0.1/0.0	...	...	...	...
Daucus carota L.	Carrot	0.0	...	...	0.2	0.50
Diospyros kaki Thunb.	Kaki	0.0	0.0	0.1/0.0	0.2	0.75
Eriobotrya japonica (Thunb.) Lindl.	Loquat (Japanese plum)	0.3	-0.1/0.0	0.1/0.0	0.2	0.50
Eucalyptus spp.	Eucalyptus	–0.1	0.4	-0.1	0.0	0.25
Euterpe edulis Mart.	Palm heart	–0.1	...	...	0.1	0.00
Ficus carica L.	Fig	0.3	0.6/0.5	0.3	0.4	0.75/0.50
Fragaria vesca L.	Strawberry	–0.1	...	...	...	...
Glycine hispida Maxim.	Soybean	0.0	...	...	0.0	0.50
Gossypium sp.	Cotton	–0.1	...	...	0.1	0.75
Helianthus annuus L.	Sunflower	0.0	...	...	0.1	0.25
Hevea brasiliensis (Willd. ex A.Juss.) Müll.Arg.	Rubber tree	0.0	-0.3	0.0	0.2	0.25
Hibiscus esculentus L.	Okra	–0.1	...	...	0.0	0.25
Hyparrhenia rufa (Nees) Stapf	Thatching grass	0.0	...	...	...	...
Ipomoea batatas (L.) Lam.	Sweet potato	0.0	...	...	0.1	0.50
Lactuca sativa L.	Lettuce	0.0	...	...	...	...
Litchi chinensis Sonn.	Litchi	0.0	0.0	0.0	0.2	0.25
Luffa cylindrica M.Roem.	Luffa (loofah)	0.0	...	...	...	...
Macadamia spp.	Macadamia nut	0.1	-0.3	0.1	0.0	0.00
Malpighia glabra L. L.	Barbados cherry	0.0	0.2	0.0	0.2	0.50
Mangifera indica L.	Mango	0.1	-0.3	0.1	0.2	0.50
Manihot utilissima Pohl	Manioc (cassava, manihot)	–0.1	...	...	–0.1	0.25
Medicago sativa L.	Alfalfa (lucerne)	0.0	...	...	...	...
Melinis minutiflora P.Beauv.	Molasser grass	0.0	...	...	...	...
Morus alba L.	White mulberry	0.0	0.2	0.2	...	...
Musa spp.	Banana	0.0	0.7	0.1	0.1	0.50
Myrciaria cauliflora (Mart.) O.Berg	Jaboticaba (Brazilian grapetree)	0.1	-0.3	0.1	0.2	0.50
Opuntia ficus-indica (L.) Mill.	Prickly pear (Indian fig)	0.4/0.5	0.1/0.0	0.2	0.4/0.5	0.75/1.00
Oryza sativa L.	Rice	–0.2	...	...	-0.2	0.00
Panicum maximum Jacq.	Guinea grass	–0.1	...	...	...	...
Passiflora spp.	Passion fruit	0.0	...	...	0.2	0.25
Pennisetum glaucum (L.) R.Br.	Millet	0.0	...	...	0.0	0.00
Pennisetum purpureum Schumach.	Elephant grass	–0.1	...	...	...	...
Persea americana Mill.	Avocado	0.1	-0.6	0.0	0.2	0.50
Phaseolus vulgaris L.	Bean	–0.1	...	...	-0.1	0.00
Phaseolus vulgaris L.	Pole bean	–0.1	...	...	...	...
Pinus spp.	Pine tree	0.0	0.3	0.0	0.0	0.25
Pisum sativum L.	Pea	–0.1/0.0	...	...	...	...
Prunus persica (L.) Batsch	Peach	0.1/0.0	-0.2	0.0	0.2	0.75
Prunus persica (L.) Batsch var. *nucipersica*	Nectarine	0.3/0.5	0.0	0.3	0.3	0.75/1.00

Prunus spp.	Prune	0.1	0.4/0.5	0.1	0.3	0.75

Table 3 - Best/more convenient value of lambda in Box-Cox transformation to normality, for each variable and crop, state of Sao Paulo, Brazil, 2007/08

Species	Crop	Cultivated area	Plant stand	Number of plants	Production	Yield
Psidium guajava L.	Guava	0.1	-0.1	0.1	0.3	0.50
Pyrus communis L.	Pear	0.4/0.5	-0.1/0.0	0.2	0.4	0.75
Ricinus communis L.	Castor bean	0.0	...	...	0.2	0.25
Rubus spp.	Blackberry	0.4/0.5	0.1	0.2	0.2	0.25
Saccharum officinarum L.	Sugar cane	0.0	...	...	0.1	0.75
Sechium edule (Jacq.) Sw.	Chayotte	0.0	...	...	0.2	0.50
Setaria italica (L.) P.Beauv.	Italian millet (foxtail millet)	0.1/0.0	...	...	0.1	0.00
Solanum gilo Req. ex Dunal	Scarlet eggplant	–0.1	...	...	...	...
Solanum lycopersicum	Tomato (vine or indeterminate)	–0.1	...	...	–0.1	0.25
Solanum lycopersicum	Tomato for industrial processing (bush or determinate)	0.0	...	...	0.1	0.25/0.50
Solanum melongena L.	Egg-plant (aubergine)	–0.1	...	...	...	...
Solanum tuberosum L.	Potato	0.0	...	...	0.1	0.75
Sorghum spp.	Forage sorghum	0.0	...	...	...	...
Sorghum vulgare Pers.	Sorghum	0.0	...	...	0.2	0.50
Sorghum vulgare Pers. var. *technicum*	Sorghum for broom industry	0.0	...	...	...	...
Spinacia oleracea L.	Spinach	–0.1	...	...	...	...
Spondias purpurea L.	Red mombin	0.0	0.0	0.0	0.4	0.50
Tectona grandis L.f.	Teak	0.1/0.0	0.5	0.0	0.0	0.00
Thea sinensis L.	Tea	0.0	2.0	0.0	0.2	0.75/0.50
Triticum vulgare Vill.	Wheat	0.1/0.0	...	...	0.1	0.25
Triticum x Secale	Triticale	0.1/0.0	...	...	0.1	0.25
Vigna sinensis Endl. ex Hassk.	Black-eyed bean (black-eyed pea)	–0.1/0.0	...	...	...	...
Vitis spp.	Fine table-grape	0.1	-0.2	0.0	0.3	0.75
Vitis spp.	Rustic grape	0.3	0.5	0.4	0.3	0.50
Zea mays L.	Corn silage	–0.1	...	...	0.2	0.25
Zea mays L.	Maize (corn)	–0.1	...	...	0.0	0.75
Zea mays L.	Maize (corn) 2nd harvest	0.0	...	...	0.1	0.50
Zea mays L.	Sweet corn	0.0	...	...	0.0	0.25
Zea mays var. *everta*	Popcorn	0.2	...	...	0.2	0.25
Zingiber officinale Roscoe	Ginger	0.1	...	...	0.2	0.25
...	Citrus nursery	–0.2	0.3	0.2	...	...
...	Cut flower species	0.1	...	...	...	...
...	Flowers and ornamental plants nursery	0.1/0.0	...	...	...	...
...	Forest species nursery	0.0	0.1	0.3	...	...
...	Grass	0.0	...	...	...	...

...	Home garden	–1.2	...	...	...	...
...	Home orchard	–0.6	0.0	-0.2	...	...

Table 4 - Best/more convenient value of lambda in Box-Cox transformation to normality, for each variable and crop, state of Sao Paulo, Brazil, 2007/08

Species	Crop	Cultivated area	Plant stand	Number of plants	Production	Yield
...	Medicinal and aromatic plants	–0.2	...	...	...	...
...	Other annual species	0.0	...	...	...	...
...	Other forest species	0.0	0.0	0.0	0.0	0.00
...	Other fruit trees	–0.6	0.2	-0.3	...	...
...	Other fruits nurseries	0.4	0.2	0.4/0.5	...	...
...	Other grasses for grazing	0.0	...	...	...	...
...	Other leguminous species for grazing	0.0	...	...	...	...
...	Other nurseries	0.0	...	...	...	...
...	Other vegetable crops	0.0	...	...	...	...
...	Pot flower species	0.1	...	...	...	...
...	Rubber tree nursery	0.3	0.3	0.3	...	...

In general lines, the *logarithmic transformation* showed to be the best for cultivated area (53% of crops), number of plants (47%), and plant stand (34%). The square root transformation ($\lambda = 0.5$) in the case of number of plants, usual for count data, was indicated for only two crops, probably due to the fact that a very large number of plants are counted, not only success and failure trials, like in a binomial distribution (Tables 1, 2, 3, 4). The most common value for productivity or yield was $\lambda = 0.5$, the *square root transformation* (although it is not a case of count data and there is no evidence that yield follows a Poisson distribution). On the other hand, the most common value for production was $\lambda = 0.2$, that is, the transformation given by $y^{(\lambda)} = 5\sqrt[5]{y}$. It is remarkable that in Cobb-Douglas production function modelling (COBB and DOUGLAS, 1928) for agriculture the logarithm of the production is taken in order to linearize (linear anamorphosis) the model before estimating the parameters. This procedure can at the same time make the production a normal and homoscedastic variable, in the cases the logarithmic transformation is convenient.

Also (Table 5), the mean is showed to be larger than median for cultivated area and production (100% of crops), number of plants (98%), plant stand (94%), and productivity or yield (90%), whereas the median was larger than the mode for cultivated area (91% of crops), number of plants (87%), production (85%), plant stand (66%), and productivity or yield (63%).

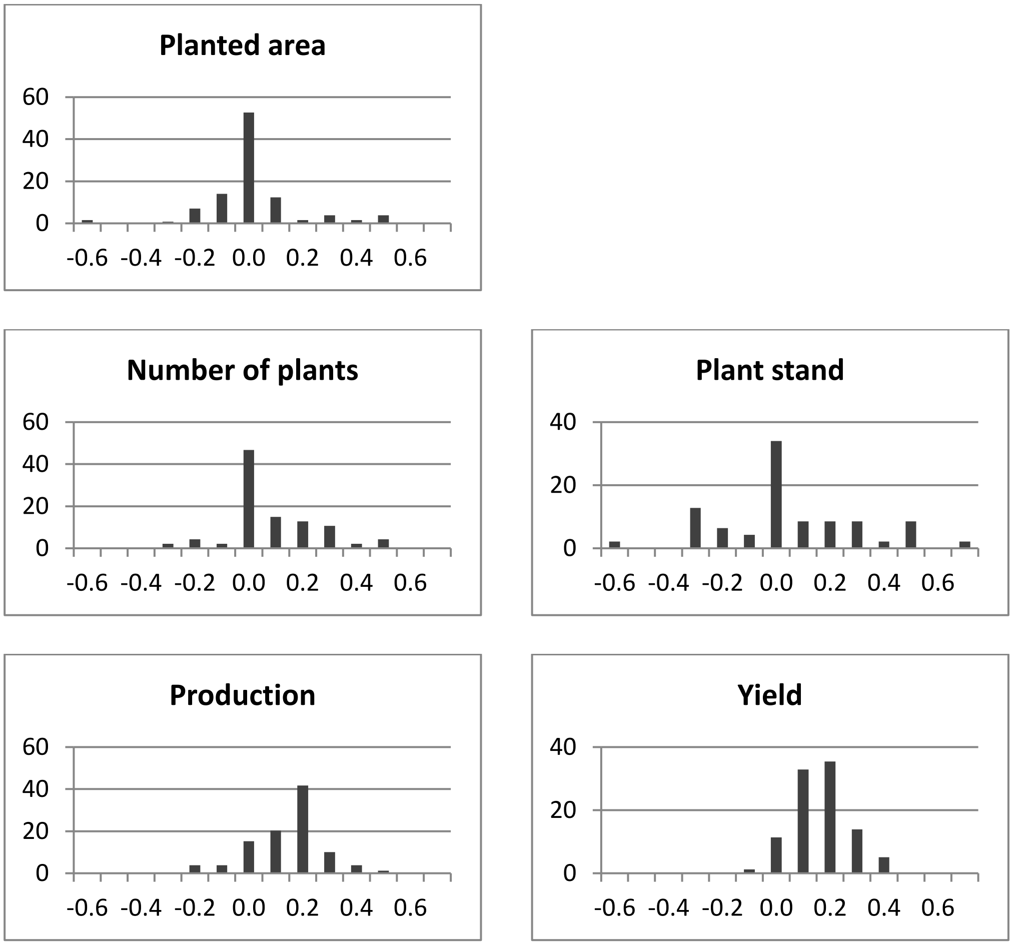

Figura 1 - Percentage of crops (vertical axis) by lambda value for Box-Cox transformation (horizontal axis) and by agronomic variable, state of Sao Paulo, Brazil, 2007/08.

All the crops (100%) exhibited positive skewness (heavy tail at right), and positive kurtosis for cultivated area, number of plants and production, and the same occurred in most crops for plant stand (98%) and for productivity or yield (skewness in 90% of crops, and kurtosis in 95% of the cases). Negative or left skewness was observed in plant stand (tea), and in productivity or yield (sugar cane, cotton, Parana pine, yam, nectarine, pear, Indian fig and red mombin, but there is no evidence of any pattern or relationship among these cases). A few cases of negative kurtosis data indicate that the sample distribution of yield has lighter tails and a flatter peak than the normal distribution in these crops (Parana pine, Indian fig, red mombin and fine table grape).

Table 5 - Comparison among location measures, and linear correlation coefficient among moments, for each variable, state of Sao Paulo, Brazil, 2007/08.

Statistics	Cultivated area	Plant stand	Number of plants	Production	Yield
Mean > Median (%)	100.00	93.62	97.87	100.00	89.87
Median > mode (%)	91.47	65.96	87.23	84.81	63.29
Skewness > 0 (%)	100.00	97.87	100.00	100.00	89.87
Kurtosis > 0 (%)	100.00	97.87	100.00	100.00	94.94
Correlation SD x Mean	0.85	0.92	0.17	0.96	0.57
Correlation Skewness x Coef. variation	0.68	0.20	0.50	0.68	0.80
Correlation Skewness x Mean	0.05	0.17	-0.13	-0.02	-0.14
Correlation Skewness x Kurtosis	0.86	0.20	0.06	0.92	0.93
Correlation Kurtosis x Coef. variation	0.35	-0.08	-0.05	0.46	0.84
Correlation Kurtosis x Mean	0.03	0.94	0.74	-0.03	-0.13

When the standard deviation varies proportionally to the mean, heterocedasticity may be assessed by calculating the linear correlation between standard deviation and mean among the different farms with each crop. This was not done in this paper, but the linear correlation was calculated between crops as a proxy (Table 5): high correlation was found for production (0.96), plant stand (0.92) and cultivated area (0.85), but lower for yield (0.57) and number of plants (0.17). Hence, a schematic classification of the variables is obtained by putting in the same column the variables with similar transformations and in the same line the variables with similar correlation between standard deviation and mean (Figure 2).

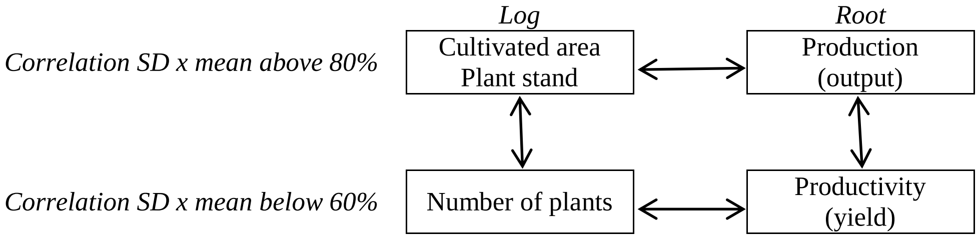

Figura 2 - Schematic classification of agricultural variables by transformation to normality and by correlation between standard deviation and mean, state of Sao Paulo, Brazil, 2007/08.

It is expected the high correlation between standard deviation and mean problem to be solved as a consequence of transformation to normality in all the variables, except for the number of plants, in which this question is irrelevant.

Skewness showed to be linear and positively correlated to kurtosis (0.86), to a lesser extent to the coefficient of variation (0.68), but not to the mean, in the cases of cultivated area, production and yield. Skewness was averagely correlated to the coefficient of variation (0.50), but not to kurtosis and mean, for the number of plants. For plant stand, skewness showed to be little correlated to kurtosis, coefficient of variation and the mean.

Kurtosis is little correlated to coefficient of variation and the mean (cultivated area and production), or highly correlated to the coefficient of variation, but not to the mean (yield), or highly correlated to the mean, but not to the coefficient of variation (plant stand and number of plants).

These results for cultivated area suggest a distribution with heavy tail at right and extreme high values, from which one may infer a sample *archetypal distribution* (Figure 3). A distribution with heavy tail at right and extreme high values also follows for number of plants, production and productivity our yield, but for plant stand the values may be extremely high or low, according to the crop (as in banana, rustic grape and tea). The sample archetypal distribution fits well for number of plants (87% of the crops), production (85%), plant stand (60%), and productivity or yield (52%).

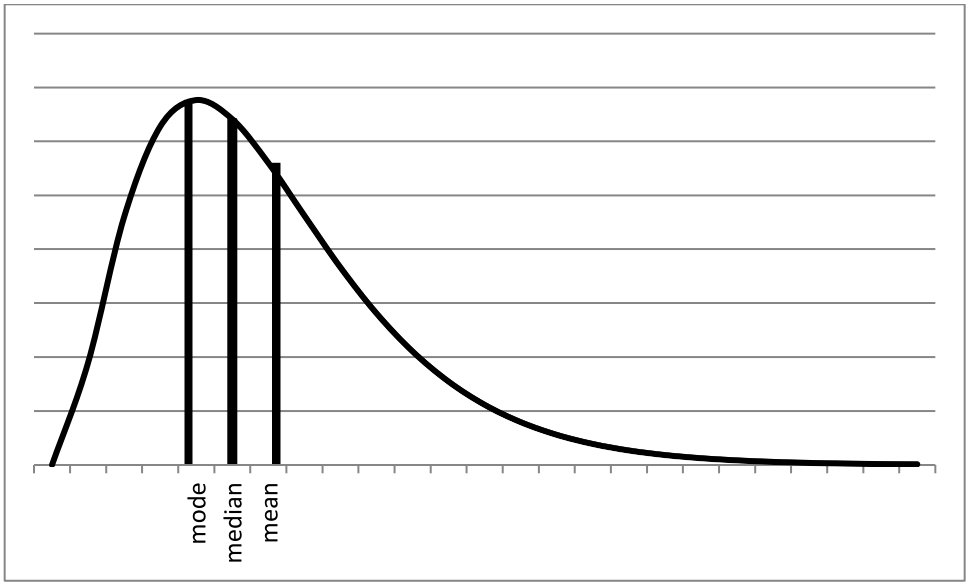

Figura 3 - Archetypal distribution proposed for cultivated area.

4 Discussion

The main reason for non-normality in variables like the cultivated area, production, number of plants, plant stand and yield is the *restriction* given by the fact that they assume only strictly positive values.

Variability. The *decision* about how much to cultivate, in terms of area or number of plants, in order to produce a given amount, is a human responsibility, specifically a *farmer's* decision, according to his interests and the resources availability. Thus, these variables can take on extreme values, especially too high, resulting in large coefficients of variation for them, as well as heavy tails at right. Nevertheless, the cultivated area was less variable than production (lower coefficient of variation in 91% of the crops[11]) because the land (and the corresponding productive soil properties) remains in the same place, does not change, except for very long periods of time, although the ownership of the land may vary, as well as the cultivated species. Therefore, the cultivated area depends only on a producer's decision, while production depends not only on that decision (through the expected productivity), but also on other environmental factors such as weather, pests, etc., that can reduce the potential productivity.

On the other hand, the variables productivity and plant stand, both defined as a ratio of two of the above three variables, are limited by the plant biology, the soil and weather conditions and the technological level. Hence, they cannot be easily established by the farmer as it is done for cultivated area. Usually, their coefficients of variation are smaller and their tails are less heavy, although asymmetric. It follows that variables such as yield and plant stand are "better behaved" than the other three, in the sense they exhibit minor discrepancies.

Skewness. The use of data from a fairly complete census, in this paper, which included commercial plantations, but also smaller crops for self-consumption, or for demonstration, or collection, explains the existence of lower values for the mode of these variables in relation to the median and the mean. However, when the cultivation is mechanized, large areas of monoculture can be planted, suggesting that in these cases the asymmetry can happen to the left. It would be interesting to verify, in some future work, if it happens now with soybeans planted in the Brazilian Midwest, but it is not the case in the state of São Paulo, where the average area of an agricultural production unit is equal to 63.3 ha, and where more than half of the units have up to 20 ha (TORRES et al., 2009). The majority of cultures showed productivity with positive skewness, which means the yield below the average is more likely than above the average. Consider the case of corn with skewness slightly above zero (0.18) while in a study in the United States it showed negative skewness (NELSON and PRECKEL, 1989). One possible explanation is the different technological level, which leads to very different average productivities: in 2007, the reference year for the data used in this paper, the average corn yield in Brazil was 3.79 t / ha, while in the US it was 9.46 t / ha (these values were calculated over data from FAOSTAT, 2007). The great variability of technological level, even among corn producers in the state of Sao Paulo, also may be because corn is a widespread culture: actually, most farmers have at least a small area, even if the corn is not important for his income.

The variables were quite regular on the fact they are not normally distributed, but unlike that, the skewness and especially the kurtosis, vary greatly in value. The high *correlation between standard deviation and mean* may indicate excessive asymmetry: the higher the average, the larger the maximum value, but lower values tend to be small, close to zero. Notwithstanding some crops may be typical of large producers and other may be typical of small ones, a census survey usually detect even very small planted areas of almost all crops. This seems to explain why the cultivated area has a high correlation between standard deviation and mean (0.85). Differently, the plant stand depends on each species and cultivation technique, not on the size of the area planted with that culture; this seems to explain why the number of plants does not show a high correlation between

[11] Exceptions: cabbage, Parana pine, sweet pepper, okra, custard apple, fig and red mombin.

standard deviation and mean (0.17). The estimated kurtosis is not linearly related to the estimated mean of the variables, except for number of plants and plant stand, wherein the larger the value of the variable, the heavier the tail of the distribution tends to be. On the other side, the kurtosis seems to grow with the variability (measured by the coefficient of variation) only for the yield.

Median. It follows that in all these variables, the median is shown to be a useful measure of central tendency or location, because it is less affected than the average by extreme values resulting from heavy tails of the variable distribution. Confidence intervals around the mean should be asymmetric on the original data, since they have to be calculated over data transformed to normality. In the case of direct modeling of these variables, where the tails are heavy, the *estimation* of minimum absolute deviation will work better than the usual least squares estimation, but the latter can be used if the data are transformed to normality before the estimates are calculated.

Distribution. The distribution suggested as *archetypal*, with positive skewness and mean ≥ median ≥ mode, is applicable for cultivated area (91% of the cases), number of plants (87%) and production (85%), but to a lesser extent for plant stand (60%) and yield (48%). Thus, the empirical results support the idea of the proposed theoretical archetypal distribution in many cases. Moreover, the shape of this distribution is consistent with a Beta distribution, as proposed by Nelson and Preckel (1989).

Criticizing papers on agriculture. At this point, a pertinent question (in the sense of apposite, relevant, applicable) and simultaneously an impertinent question (in the sense of rude) arises: do these results invalidate all the thousands of works and studies published in the last half century on agricultural issues, but did not considered the lack of normality of these variables? Obviously, this is not true for papers that have taken into account the issue of non-normality. Nor can they be invalidated the papers done with rigor and care about the other statistical aspects and whose conclusions were based on tests in which the null hypothesis was rejected with high significance. However, estimates of model parameters and variance analysis based on results with low significance may eventually have led to erroneous conclusions.

The chances of valid results are high in biological and physical sciences, applied to agricultural issues, usually based on data arising from carefully designed experiments measurements. It may not be the case in human sciences, applied to agricultural issues, commonly based in surveyed data from available real situations. Even if the problem is mitigated in econometric models, which generally use the logarithmic transformation, this paper strongly suggests special attention to the question of non-normality in future studies, mainly in agricultural economics.

5 Conclusions

The main conclusions are:

a) Agronomical variables, such as cultivated area, plant stand, number of plants, production and crop yield do not follow a normal distribution.

b) Non-normality, great variability, asymmetry and heavy tails of these agronomical variables are a consequence of several factors: restrictions to positive values; some variables are defined as the ratio of two others; farmer decisions; environmental conditions (weather, pests, etc.); and technological level.

c) More adequate probability distributions should be used, like the Beta distribution or an archetypal distribution defined for non-negative values, with heavy tail at right, and mean > median > mode.

d) The Box-Cox transformation should be used on these agronomical variables in search for normality. Despite of the fact that each individual crop requires a particular transformation, some general results emerge from this paper: the logarithmic transformation is the best/more convenient in most cases for cultivated area, plant stand and number of plants, but root transformations perform better for crop production and yield.

e) The median is a useful measure of central tendency or location for these agronomical variables.

f) Least absolute deviations estimation should be preferred rather than minimum squares estimation when dealing with models for these agronomical variables without data transformation to normality.

g) Scientists and technicians are strongly recommended to deal with the question of non-normality of the data before the analysis and the conclusions of their reports.

Acknowledgements

The author thanks Denise Viani Caser, Dr. José Roberto Vicente, and Dr. Antonio Ambrosio Amaro for the suggestions on an early version, and Dr. Ingrid Silva Pino for data visualization consultancy.

PINO, F. A. Distribuições amostrais de variáveis agrícolas: um estudo de caso no sudeste do Brasil. *Rev. Bras. Biom*., Lavras, v.35, n.4, p.714-731, 2017.

- *RESUMO: As distribuições de probabilidade de algumas variáveis agronômicas, a saber, área plantada ou cultivada, número de plantas, densidade de plantio (número de plantas por hectare), produção e produtividade (produção por hectare) foram estudadas e caracterizadas, usando dados censitários. A hipótese de normalidade foi rejeitada para todas as variáveis e todas as culturas. A principal transformação para normalidade foi a logarítmica, seguida de transformações por raiz. Uma distribuição arquetípica foi sugerida para área plantada: definida para valores não negativos, com cauda pesada à direita, média > mediana > moda. Ela mostra-se relevante para a própria área plantada (91% dos casos), número de plantas (87%) e produção (85%), um pouco menos para densidade de plantio (60%) e produtividade (48%).*

- *PALAVRAS-CHAVE: Teste para normalidade; transformação de Box-Cox; assimetria e curtose; dados agrícolas.*

References

ANTLE, J. M.; GOODGER, W. J. Measuring stochastic technology: the case of Tulare milk production. *Am. J. Agr. Econ.*, v.66, p.342-350, 1984.

BAKER, D. J.; CROSS, N. L.; SEDGWICK, E. M. Normality of single fibre electromyographic jitter: a new approach. *J. Neurol. Neurosur.*, v.50, p.471-475, 1987.

BERNIER, J.; FENG, Y.; ASAKAWA, K. Strategies for handling normality assumptions in multi-level modeling: a case study estimating trajectories of Health Utilities Index Mark 3 scores. *Health Report*, Ottawa, v.22, p.45-51, 2011.

BLYTHE, E. K.; MERHAUT, D. J. Testing the assumption of normality for pH and electrical conductivity of substrate extract obtained using the pour-through method. *HortScience*, v.42, p.661-669, 2007.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *J. Roy. Stat. Soc. B*, v.26, n.2, p.211-252, 1964.

BRAH, G. S.; LANZA, G. M.; POTTS, P. L.; WASHBURN, K. W. Effects of deviations from normality on selection intensities for shell deformation and egg weigh in chickens. *Poultry Sci.*, v.61, p.424-428, 1982.

BRITTON, J. R. Effects of social class, sex, and region of residence on age at death from cystic fibrosis. *Brit. Med. J.*, v.298, p.483-487, 1989.

BULMER, M.G. On fitting the Poisson lognormal distributions to species abundance data. *Biometrics*, v.30, p.101-110, 1974.

CHEN, S. -L.; MIRANDA, M. J. Modeling multivariate crop yield densities with frequent extreme events. In: AMERICAN AGRICULTURAL ECONOMICS ASSOCIATION ANNUAL MEETING, 2004, Denver, Colorado. Paper 19970. Available: http://ageconsearch.umn.edu/bitstream/19970/1/sp04ch12.pdf.

COBB, C. W.; DOUGLAS, P. H. A theory of production. *Am. Econ. Rev.*, v.18, Supplement, p.139-165, 1928.

CRAWFORD, J. R.; GARTHWAITE, P. H.; AZZALINI, A.; HOWELL, D. C.; LAWS, K. R. Testing for a deficit in single-case studies: effects of departures from normality. *Neuropsychologia*, v.44, p.666-677, 2006.

DAY, R. H. Probability distributions of field crop yields. *J. Farm Econ.*, v.47, p.713-741, 1965.

DeBOER, W.; VOET, H.; BOON, P.; BOKKERS, B.; BAKKER, M. A comparison of two models for estimating usual intake addressing zero consumptions and non-normality. *Food Addit. Contam.*, v.26, p.1433-1449, 2009.

DINEEN, P.; COLES, P. Multivariate non-normality in WMAP 1st year data. *Mon. Not. R. Astron. Soc.*, 2008. Available: http://www.researchgate.net/publication/1825787_Multivariate_Non-Normality_in_the_WMAP_1st_Year_Data.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. 2d. New York: Wiley. 1981.

DURAZO-ARVIZU, R. A.; McGEE, D. L.; COOPER, R. S.; LIAO, Y.; LUKE, A. Mortality and optimal body mass index in a sample of the US population. *Am. J. Epidemiol.*, v.147, p.739-749, 1998.

DUTY, S. M.; CALAFAT, A. M.; SILVA, M. J.; RYAN, L.; HAUSER, R. Phthalate exposure and reproductive hormones in adult men. *Hum. Reprod.*, v.20, p.604-610, 2005.

FAOSTAT. Food and Agriculture Organization of the United Nations. Statistics Division. Available: http://faostat3.fao.org/browse/Q/QC/E, 2007.

GABRIEL, K. R.; FEDER, P. On the distribution of statistics suitable for evaluating rainfall stimulation experiments. *Technometrics*, v.11, n.1, p.149-160, 1969.

GILES, P. J.; KIPLING, D. Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, v.19, p.2254-2262, 2003.

GUAN, N. C.; YUSOFF, M. S. B.; ZAINAL, N. Z.; YUN, L. W. Analysis of two independent samples with non-normality using non parametric method, data transformation and bootstrapping method. *Int. Med. J.*, v.19, p.227-229, 2012.

HARRAR, S. W.; GUPTA, A. K. Asymptotic expansion for the null distribution of the *F*-statistic in one-way ANOVA under non-normality. *Ann. I. Stat. Math.*, Tokyo, v.59, p.531-556, 2007.

HONG, H. P. Application of polynomial transformation to normality in structural reliability analysis. *Can. J. Civil Eng.*, v.25, p.241-249, 1998.

INDRAYAN, A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr.*, v.51, p.37-43, 2014.

JUNQUEIRA, A. A. B.; CRISCUOLO, P. D.; PINO, F. A. Use of energy by agriculture in the state of Sao Paulo. *Agricultura em São Paulo*, SP, v.29, p.55-100, 1982. (in Portuguese).

KATZ, R. W.; PARLANGE, M. B. Overdispersion phenomenon in stochastic modeling of precipitation. *J. Climate*, v.11, p.591-601, 1998.

KLEIN, A. -M.; STEFFAN-DEWENTER, I.; TSCHARNTKE, T. Fruit set of highland coffee increases with the diversity of pollinating bees. *Proc. Roy. Soc. Lond. B Bio.*, v.270, p.955-961, 2003.

LACOURCIÈRE, Y.; BÉLANGER, A.; GODIN, C.; HALLÉ, J. -P.; ROSS, S.; WRIGHT, N.; MARION, J. Long-term comparison of losartan and enalapril on kidney function in hypertensive type 2 diabetics with early nephopathy. *Kidney Int.*, v.58, p.762-769, 2000.

LaLONDE, S. M. Transforming variables for normality and linearity – when, how, why and why not's. In SAS GLOBAL FORUM, April 22-25, 2012, Orlando. *Proceedings…* 2012. Paper 430-2012. 8 p.

LEE, S. -G. Statistical concepts and techniques for testing departures from normality in the Mathematics teacher preparation. *Honam Math. J.*, Kwandju, v.29, p.83-100, 2007.

McGRATH, D.; ZHANG, C.; CARTON, O. T. Geostatistical analyses and hazard assessment on soil lead in Shilvermines area, Ireland. *Environ. Pollut.*, v.127, p.239-248, 2004.

MOSS, C. B. Defining the general transformation to normality: a proposal to correlate general nonnormal distributions. In: SCC-76 ANNUAL MEETING, Pensacola Beach, March 26-28, 2015,.

MOSS, C. B.; SHONKWILER, J. S. Estimating yield distributions using a stochastic trend model and nonnormal errors. *Am. J. Agr. Econ.*, v.75, p.1056-1062, 1993.

NELSON, C. H.; PRECKEL, P. V. The conditional beta distribution as a stochastic production function. *Am. J. Agr. Econ.*, v.71, p.370-378, 1989.

NUSSER, S. M.; CARRIQUIRY, A. L.; DODD, K. W.; FULLER, W. A. A semiparametric transformation approach to estimating usual daily intake distributions. *J. Am. Stat. Assoc.*, v.91, p.1440-1449, 1996.

OMARIBA, W. R. Gender differences in functional limitations among Canadians with arthritis: the role of disease duration and comorbidity. *Health Report*, Ottawa, v.22, p.7-14, 2011.

OZAKI, V. A.; FARIA, P.; OLINDA, R.; CAMPOS, R. C. Análise estatística da probabilidade de perda agrícola: uma aplicação da teoria dos valores extremos. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE ECONOMIA, ADMINISTRAÇÃO E SOCIOLOGIA RURAL, 48, 2010, Campo Grande, MS. *Anais*... SOBER, 2010. (in Portuguese).

PINO, F.A. *Distribuição de probabilidade de algumas variáveis de interesse agronômico*: área plantada, produção e produtividade. São Paulo, Instituto de Economia Agrícola, 1979. 64p. Relatório não publicado de projeto de bolsa de pesquisa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). (in Portuguese).

PINO, F. A. The question of non-normality: a review. *Revista de Economia Agrícola*, São Paulo, v.61, n.2, p.17-33, jul.-dez. 2014. (in Portuguese).

PINO, F. A.; CAMARGO, M. L. B.; VIANI, D. N. *Changing of farm owners and farm sizes in the state of São Paulo during the period of 1972-77*. São Paulo: Instituto de Economia Agrícola, 1979. (Relatório de Pesquisa, 16). (in Portuguese).

POLLOCK, V. E.; SCHNEIDER, L. S.; LYNESS, S. A. EEG amplitudes in healthy, late-middle-aged and elderly adults: normality of the distributions and correlations with age. *Electroen. Clin. Neuro.*, v.75, p.276-288, 1990.

RAZALI, N. M.; WAH. Y. B. Power comparisons of some selected normality tests. In: REGIONAL CONFERENCE ON STATISTICAL SCIENCES, 2010. *Proceedings*... 2010, p. 126-138.

ROBERTS, S. G. B.; DUNBAR, R. I. M.; POLLET, T. V.; KUPPENS, T. Exploring variation in active network size: constraints and ego characteristics. *Soc. Networks*, v.31, p.138-146, 2009.

SAS INSTITUTE INC. *SAS/STAT® 9.2 user's guide*. The Transreg procedure. 2ed. Cary, NC: SAS Institute Inc. 2008.

SAS INSTITUTE INC. *BASE SAS® 9.2 procedures guide*: *Statistical procedures*. 3ed. Cary, NC: SAS Institute Inc. 2010.

STROWIG, S. M.; AVILÉS-SANTA, M. L.; RASKIN, P. Comparison of insulin monotherapy and combination therapy with insulin and metformin or insulin and troglitazone in type 2 diabetes. *Diabetes Care*, v.25, p.1691-1698, 2002.

SWARTZ. R. H.; STUSS, D. T.; GAO, F.; BLACK, S. Independent cognitive effects of atrophy and diffuse subcortical and thalamico-cortical cerebrovascular disease in dementia. *Stroke*, v.39, p.822-830, 2008.

TAYLOR, C. R. A flexible method for empirically estimating probability functions. *Western J. Agr. Econ.*, v.9, p.66-76, 1984.

TORRES, A.J. et al. (orgs.). *Projeto LUPA 2007/08*: censo agropecuário do Estado de São Paulo. São Paulo: IEA, CATI, SAA, 2009. 381p. (in Portuguese).

VanRADEN, P. M. Normality and skewness of genetic evaluations. *Interbull Bulletin*, v.35, p.164-167, 2006.

VERCHOT, L. V.; DAVIDSON, E. A.; CATTÂNIO, J. H.; ACKERMAN, I. L. Land-use change and biogeochemical controls of methane fluxes in soils of Eastern Amazonia. *Ecosystems*, v.3, p.41-56, 2000.

ZANINI, F. C.; SHERRICK, B. J.; SCHNITKEY, G. D.; IRWIN, S. H. Crop insurance valuation under alternative yield distributions. In: NCR-134 CONFERENCE ON APPLIED COMMODITY PRICE ANALYSIS, FORECASTING AND MARKET RISK MANAGEMENT, St. Louis, MO. *Proceedings***…** 2001.

ZHU, Y.; GOODWIN, B. K.; GHOSH, S. K. Modeling yield risk under technological change: dynamic yield distributions and the U.S. Crop Insurance Program. *J. Agr. Resour. Econ.*, v.36, p.192-210, 2011.

Received on 25.08.2016

Approved after revised on 23.01.2017