Scientific Papers Series Management, Economic Engineering in Agriculture and Rural Development
Vol. 14, Issue 3, 2014
PRINT ISSN 2284-7995, E-ISSN 2285-3952

RESEARCH ON THE NORMAL DISTRIBUTION OF VARIABLES AS A CONDITION FOR THE CORRECT APPLICATION OF THE STATISTICAL MODELS FOR PROCESSING DATA

Agatha POPESCU

University of Agricultural Sciences and Veterinary Medicine Bucharest, 59 Marasti, District 1, Zip code 011464, Bucharest, Romania, Phone: +40 213182564/232, Fax:+40213182888, Email:agatha_popescu@yahoo.com

Corresponding author: agatha_popescu@yahoo.com

Abstract

The paper aimed to test the normality of the distribution of the variables using Pearson's central moment coefficients. The data were referring to the milk and meat production characters collected from 5,817 individuals, offspring from 106 Frisian bulls under breeding value estimation. The variables were used both in terms of metric characters and logarithmic characters. In case of milk production and reproduction traits, the use of logarithmic characters was useless, and the values of the Beta coefficients showed a distribution relatively closer to the normal one. In case of meat production characters, both the use of metric and logarithmic characters reflected that the distribution of variables was not a normal one, on the contrary it was sharply vaulted. As a conclusion, it is important to check the normality of the variables distribution in order to avoid the errors of calculation by eliminating the extreme values in case of the skewed distribution.

Key words**:** *normal distribution, statistical processing methods, study case, testing*

INTRODUCTION

In many research works it is required a large range of data collected from many individuals, which could be an impediment to carry out the experiments or studies. For this reason, it is simplier to use samples of individuals and the conclusions to be extended at the whole population or colectivity.

Thesample selection with "n" individuals and X_1, X_2,... X_n variables imposes to study their type of distribution in order as later to choose the most adequate statistical methods for processing data.

A normal distribution of variables looks like Gauss' curve where variables are bell-shaped, and the mean, $\mu = 0$ and standard deviation is S=1 and its variance is S^2.

This is the so called "standard normal distribution"[7,13,15].

Sometimes, the variables are not normally distributed, that is we can discuss about "skewness" as a measure of the asymmetry of the probability distribution of real values around its mean, μ. In case of skewness, the distribution of variables could be:(a)*skewed to the left*, where the mean is less than the median and (b) *skewed to the right*, where the mean is higher than the median [14].

In a symmetric distribution, the mean is equal to the median and the distribution has zero skewness [9,11].

In case of a sample, the mean, "m_3" is the sample 3rd central moment and the sample variance is "m_2", the 2nd central moment.

The sample skewness, g_1 is given by the formula: $g_1 = m_3 / m_2^{3/2}$, where $m_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$ and

$m_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The common estimator of the population skewness is:

$G_1 = \frac{K_3}{K_2^{3/2}} = (\sqrt{n(n-1)}/(n-2))\, g_1$

In this case, teh variance of the skewness of a sample of size n from a normal distribution is: $var(G_1) = (6n(n-1))/((n-2)(n+1)(n+3))$ [6,12].

Karl Pearson (1895and later 1920) established two skewness coefficients to simplify

calculations:
-1st skewness coefficient =(mean-mode)/standard deviation;
-2nd skewness coefficient = 3(mean-median)/standard deviation [16,17].
At present, Excel and Minitab, SAS, SPSS, Stata, Visual Statistics and other statistical packages facilitate the calculation of the "*adjusted Fisher-Pearson standardized moment coefficient*" based on the formula:

$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} ((x_i - \bar{x})/s)^3$, where n= the sample size and s= standard deviation [16,17].
In order to establish if thedistributionof variables is a normal one, one can think that (a) the null hypothesis, H_0, is that the variables are normally distributed with an unspecific mean, μ, and variance S^2, and (b) the alternative hypothesis, H_a, is that the variables are notnormally distributed.
In this purpose, more than 40 normality tests are at our disposal, among which the most commonly used are: Q-Q plot, P-P plot, Shapiro-Wilk test, normal probability test, D'Agostino's k-squared test, Jarque-Bera test, Pearson's beta coefficients test, Kolmogorov-Smirnov test, Liliefors test, Anderson-Darling test, Sarkady test, Massey test, Chi Squares test [6].
Since 1895, when Karl Pearson started studying the various statistics of skewness, statisticians approached this topic and pointed out its strengths and weaknesses [1,5,6,8,20,22, 23,24].
In this context, the paper aimed to check the distribution of the variables in the field of animal breeding, in a cattle population, using a large sample of individuals, used for breeding value estimation both for milk and meat production traits. Pearson's Beta coefficients were used in order to test the normality of the distribution of variables.
The purpose is demonstrate that from a practical point of view, before starting to process data by statistical mathematical methods, we need to know what kind of distribution the variables have in order to apply the correct mathematical models and avoid wrong results.
If the variables are not normally distributed, they have to be normalized by eliminating the extreme variables, in general the source of error.
Many times, metric variables are transformed in logarithmic variables or probit, more suitable tobe later processed [10].

MATERIALS AND METHODS

This case study was applied on a sample of 5,817 individuals of which 4,112 half daughters and 1,705 half broters, descendants from 106 Friesian bulls under the estimation of the breeding value both for milk and meet production traits [18,19].
The observed variables regarded the following characters:(a) *for milk production:*milk yield, fat % and fat yield, (b) *for reproduction:*age at the 1st calving, and (c) *for meat production:* body live weight at the age of 180 days, body live weight at the age of 365 days, and weight daily gain during the period of strees fattening.
Pearson's beta coefficients were used in order to test the normality of the variables distribution.
As we mentioned before, Pearson's beta coefficients are based on the central slectionmoments of the order "k" (k=2,3,4) which allow the synthetic presentation ofthe distribution, idetifying some specific features linked to its shape. The distribution of the variables couldbe normal, skewed to the left, skewed to the right, vaulted or flattened. When the variables is normally distributed, the assymetry is equal to zero (B_1=0), the curve of the frequences is perfecly symmetrical having the shape of Gauss-Laplace's curve. In case where the variables are not normally distributed around the mean, itsi about positive assymetry (B_1 over 0) or negative assymetry (B_1 lower than 0). The degree of assymetry on the horizontal of the frequences polygon versus the normal $N(\bar{X}, S_X^2)$ is reflected therefore by Pearson's assymetry coefficient, B_1, calculated as: $B_1 = M_3^2 / M_2^3$, where M_2 and M_3 are the central 2nd and 3rd moments, calculated according

tothe formula:

$$M_K = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^K$$

The vertical analysis of the frequences polygon versus the normal allows to establish the value of Pearson's B_2 coefficient, which could reflect either a vaulted distribution or a flattened distribution. In case of a normal distribution, the values distribution is a normal one and B_2=3, while in case of a flattened distribution, B_2 is lower than 3, and in case of a vaulted distribution, B_2 >3.

The B_2 coefficient was calculated with the formula $B_2=M_4/M_3^2$, where M_3 and M_4 are the central selectionmoments of the 3rd and 4th order.

The normality test was applied both on metric variables and logarythmic variables [21].

In order to validate the conclusions, the null hypothesis was tested and the intervals over the critical values of the skewness and vaultness distribution were established, where significant and very significant differences have appeared. Inthispurpose, the figures presented in the Statistical Tables for P=0.05 were used [2,3,4].

RESULTS AND DISCUSSIONS

Normality test for milk production and reproduction traits

(a)The distribution of variables in terms of metric characters

The calculated values of Pearson's Beta coefficients, based on metric characters, are presented in Table 1.

Milk yield for 305 days of lactation regitered almost a normal value of the B_1 coefficient on the horizontal. Makingthe vertical analysis, we noticed that the B_2 value is a little smaller than 3, indicating a slight flattened curve. However, the difference is not so far away from the normal situation.

The fat percentage registered almost a normal assymetry coefficient (B_1=0.12) on the horizontal, while in case ofthe vertical analysis B_2 = 12.68, reflecting a sharpness of the curve of values for this trait.

The fat yield recorded almost normal values for Pearson's Beta coefficients, B_1 = 0.11 and B_2 =2.80, but with a slight flattened shape.

The age at the 1st calving had a distribution which did not fit to Gauss curve. The value B_1 = 3.94 reflect a right skewed distribution of the metric variables, therefore a strong positive skewness, and the value B_2 = 12.68 reflected a substantial sharpness on the vertical of the distribution curve.

The affirmations presented above were confirmed by the results obtained by testing the null hypothesis (Table 1).

Table 1. Skewness and vaulting coefficients for the metric characters of milk production and reproduction (N=4,112)

Character	B_1	H_0	B_2	H_0
Milk yield	0.14	No/Yes	2.83	Yes
Fat %	0.12	No/Yes	12.68	No
Fat yield	0.13	No/Yes	2.80	Yes
Age at the 1st calving	3.94	No	2.81	No

Source:Own calculations

(b)The distribution of variables in terms of logarythmic characters

Because it was obtained both a normal and an skewed distribution of the variables interms of metric characters, it was compulsory to recalculate Pearson's Beta coefficients based on data expressed in logarythmic characters, thinking that in this way the distribution could be normalized. The values of the B_1 and B_2 coefficients values based on logarithmic characters are given in Table 2.

Table 2. Skewness and vaulting coefficients for the logarithmic characters of milk production and reproduction (N=4,112)

Character	B_1	H_0	B_2	H_0
Milk yield	0.19	No/Yes	3.05	Yes
Fat %	14.96	No	72.49	No
Fat yield	0.37	No/Yes	3.98	No
Age at the 1st calving	0.47	No	4.96	No

Source:Own calculations

Milk yield for 305 days of lactation registered higher values of the Beta coefficients than in case when the variables were expressed in terms of metric characters, but these values are closer to a normal distribution.

The fat percentage recorded an expected values for the both coefficients, placed far away outside of the critical values ofthe skewness and valutness for P = 0.05.

The fat yield also achieved higher values for the Beta coefficients tha in case when we used metric characters. But, these values do not exceed too much the critical thresholds, reflecting a slight positive skewness and a slight vaulting of the normal distribution as well.

The age at the 1st calving recorded normal values for the both coefficients, existing just a slight right assymetry and a slight trend of sharpness of the curve.

Therefore, the transformation and use of the primary data as logarithmic characters did not lead to the distribution normalization, on the contrary, it increased the deviation of the Beta coefficients values from the critical thresholds and normal values. We could affirm that the use of logarithmic characters contributed to the omogenization of the variances and imposes to look for other solutions and methods to normalize the distribution of the variables.

Normality test for meat production traits

(a) The distribution of variables in terms of metric characters

The body weight atthe age of 180 days registered values which were not normally. The distribution was skewed to the right and a little vaulted, a reason for which the null hypothesis was not accepted for P = 0.05.

The body weight at the age of 365 days achieved values closer to a normal distribution, but, also, like in the previous case, the distribution was a little positive (B_1 = 0.045), and also slightly flattened (B_2 = 2.748).

The weight daily gain recorded values which were normally distributed around the average on the horizontal line, but on the vertical line iot was noticed a strong vaulting trend (B_2 =4.848).

Testing the H_o hypothesis, we noticed various situations form a trait to another and from a coefficient to another.

Because the differences regarding the critical values ofthe assymetry and vaulting for various probability degrees were sometimes un significant, significant and substantial significant, but not too much, it was considered that, in genral, this sample of individuals is enough representative, allowing to determine the statistical parameters like average, standard deviation, variance and variability based on the usual statistical models without facing any risk of error (Table 3).

Table 3. Skewness and vaulting coefficients for the metric characters of meat production (N=1.705)

Character	B_1	H_0	B_2	H_0
Weight at the age of 180 days	0.170	No/Yes	3.511	No
Weight at the age of 365 days	0.045	Yes	2.748	No/Yes
Weight daily gain	0.006	Yes	4.848	No

Source:Own calculations

(b)The distribution of variables in terms of logarythmic characters

The body weight atthe age of 180 days recorded Beta coefficients with values closer to thenormal distribution than in case when there were used metric characters.

The body weight atthe age of 180 days registered higher values of the Beta coefficients, reflecting a positive skewness and vaulting.

The body daily gain carried out an distribution far away froma normal one, becuase the values of the Beta coefficients reflect a sharp vaulting curve.

Table 4. Skewness and vaulting coefficients for the logarithmic characters of meat production (N=1.705)

Character	B_1	H_0	B_2	H_0
Weight at the age of 180 days	0.95	No/Yes	2.98	Yes
Weight at the age of 365 days	2.73	No	3.91	No/Yes
Weight daily gain	0.25	No/Yes	10.02	No

Source:Own Calculations

Therefore, in case of meat production, the use of the orimary data in terms of logarithmic

characters hgas contributed to the normalization of the distribution only for the trait body weight atthe age of 180 days.
For the other two traits of fattening, this method did not assure the normalization ofthe variables distribution in the studied population, on the contrary, it increased the deviation ofthe Beta coefficients from the normal values.

CONCLUSIONS

Normality test is compulsory in case of the use of samples for statistical populations, even though the volume of the sample is enough high, because we do not know how the variables are distributed.
If the estimated values for the Beta coefficients, based on the metric characters, are close to the normal thresholds of variability, B_1=0 and B_2 =3, and the critical values presented in Tables do not attest significant differences, then we could consider that the distribution of the variables is a normal one, which allows us to use usual statistical methid sto process the primary data.
If the values of the Beta coefficients substantially exceed the thresholds for a normal variability, we are obliged to eliminate the potential errors proceeding to the normalization of the distribution, either using the logarithmic characters in the calculations or probit and then applying corresponding mathematical methods.
In this study, the use of logarithmic characters for calculating the Beta coefficients did not led to the normalization of the distribution, on the contrary, the skewness and vaulting have become more evident.
The normalization of the variables distribution offers a certain guarantee that the results which are expected to be obtained in the research work could supply valuable information which could be extended to the whole population of individuals.
In other case, our research work is in danger to lead to random results and the conclusions and recommendations to be under the risk of error.

ACKNOWLEDGMENTS

The author thanks the farmers who provided the primary data regarding the bulls' descendants tested for milk and meat production traits, and which were used for the application of the normality test.

REFERENCES

[1]Arnold, B.C., Groeneveld, R. A.,1995, "Measuring Skewness with Respect to the Mode," The American Statistician, 49:34-38
[2]Craiu, V., 1982, Statistical hypothesis testing. Didactical and Pedagogical Press House, Bucharest, 1982
[3]Craiu, V., Ciucu, G., Stefanescu, A., 1974, Mathematical statistics and operation research. Didactical and Pedagogical Press House, Bucharest
[4]Craiu, V., Mihoc, G., 1976, Teatise of mathematical statistics. Romanian Academy Press House, Bucharest
[5]D'Agostino, R.B. and Stephens, M.A. (1986), Goodness of Fit Techniques, Marcel Dekker, Inc., 11-12 and 24-33.
[6]Doane David P., Seward, Lori E., 2011, Measuring Skewness: A Forgotten Statistic? Journal of Statistics Education, 19(2)
[7]Gauss, 1809, Theoria motus corporum coelestium in sectionibus conicis solem ambientium (http://en.wikipedia.org/wiki/Normal_distribution)
[8]Groeneveld, R.A. and Meeden, G.(1984), "Measuring Skewness and Kurtosis," Journal of the Royal Statistical Society. Series D (TheStatistician), 33: 391-399
[9]Hazewinkel Michiel, ed., 2001, Normal Distribution, Encyclopedia of Mathematics, Springer
[10]Iosifescu, M., Moineagu, C., Trebici, V., Ursianu Emiliana, 1985, Small encyclopedia of statistics. Scientific and Encyclopedic Press House, Bucharest
[11]Jagdish Patel K., Campbell Read B., 1996, Handbook of the Normal Distribution (2nd ed.). CRC Press
[12]Krishnamoorthy Kalimuthu, 2006, Handbook of Statistical Distributions with Applications. Chapman & Hall/CRC
[13]Kruskal William H., Stigler Stephen M., 1997, Spencer, Bruce D., ed. Normative Terminology: 'Normal' in Statistics and Elsewhere. Statistics and Public Policy. Oxford University Press.
[14]Mac Gillivray, H.L., 1986, Skewness and Asymmetry: Measures and Orderings, The Annals of Statistics, 14: 994-1011
[15]Marsaglia George, 2004, Evaluating the Normal Distribution, Journal of Statistical Software 11 (4)
[16]Pearson, K., 1895, Contributions to the Mathematical Theory of Evolution, II: Skew Variation in Homogeneous Material, Transactions of the Royal Philosophical Society, Series A, 186, 343-414

[17]Pearson Karl, 1920, Notes on the History of Correlation, Biometrika 13 (1): 25–45
[18]Popescu Agatha, 1988, Study regarding the normality testing using Pearson's Beta coefficients in a cattle population. The Symposium „Actualities in Animal Science, Vol. XIII:145-152
[19]Popescu Agatha, 2014, Increase of the precision in the breeding value estimation by B.L.U.P.-Best Linear Unbiased Prediction, Eikon Cluj-Napoca, Ed.RawexComs, Bucharest Press Houses, p.74-75, 91-99
[20]Rayner, J.C.W., Best, D.J., and Matthews, K. L. (1995), "Interpreting the Skewness Coefficient," Communications in Statistics –Theory and Methods, 24, 593-600
[21]Rumsiski, L. Z.,1974, Mathematical processing of the experimantal data. Technical Press House, Bucharest
[22]Stigler, Stephen M., 1986, The History of Statistics: The Measurement of Uncertainty before 1900. Harvard University Press.
[23]Tabor, J., 2010, Investigating the Investigative Task: Testing for Skewness -An Investigation of Different Test Statistics and their Power to Detect Skewness, Journal of Statistics Education, 18, 1-13
[24]Vessereau, B., 1960, Recherche et expérimentation en agricuLture. Methode statistique en biologie et en agronomie, Paris, 1960