TEKA Kom. Mot. Energ. Roln. – OL PAN, 2011, 11, 181–188

APPLICATION OF THE AKAIKE INFORMATION CRITERION FOR AN ASSESSMENT OF SUGAR BEET CROP DISTRIBUTION.

Andrzej Kornacki[1], Katarzyna Ostroga [2]

University of Life Sciences in Lublin
[1] Department of Applied Mathematics and Informatics
[2] Department of Agricultural Machinery

Summary. Normal distribution is commonly applied in natural sciences, as well as in technical, humanistic or social. It is also very often used for the description of phenomena and processes in investigations in the domain of agricultural engineering. Wide application of this distribution is justified by the principle of great numbers. Great importance of normal distribution results also from the following facts: 1) Normal distribution is a model for random error of measurements. 2) Many physical phenomena, even if they do not have normal distribution, can be described in terms of this distribution after suitable transformation. 3) Normal distribution is a good approximation for different distributions, 4) Every linear combination of independent random variables with normal distribution has a normal distribution.

These desirable properties cause, that research workers want to check whether the results of their experiments have normal distribution. In literature many tests checking normality of distribution of the studied feature are described both in one [Anderson at al 1954, Bowman at al 1975, Geary 1947, Green at al 1976,Tench at al 1980, Oja 1983, Pearson 1930, Shapiro at al 1965, Uthoff 1973, Vasicek, 1976] and in multidimensional case [Andrews at al 1973, Bera ay 1986, Cox at al 1978, Hawkins 1981, Henze at al 1990, Buck 1986, Royston 1983]. All of them require the application of suitable tables of critical values.

In the present work we suggest using the selection between normal and logarythm - normal distribution: Akaike information criterion (AIC). This criterion originates from the theory of information. The suggested method was applied to dataset obtained in an experiment with crops of sugar beets. Conclusions resulting from criterion Akaike were confirmed by the classic Shapiro-Wilka test.

Key words: Crop of sugar beet, Akaike information criterion, normal distribution, lognormal distribution, selection of the models.

AKAIKE INFORMATION CRITERION.

The concept of Akaike information criterion (AIC) has its source in the theory of information. This notion is based on the concept of Kullback-Leibler (K-L) information [Kulback.and Leibler 1951]. Let us consider a continuous probability distribution. Let $g(x)$ be the true density function and $f(x)$ be a model density function. Then the Kullback-Leibler information is given by [17]

$$I(g,f) = \int_{-\infty}^{+\infty} \log \frac{g(x)}{f(x)} g(x)dx, \tag{1}$$

where: log denotes the natural logarithm. $I(g,f)$ describes the distance between two probability distributions and has the following properties [Akaike 1974]:

1. $I(g,f) \geq 0$,
2. $I(g,f) = 0 \Leftrightarrow f(x) = g(x)$ almost everywhere.

Obviously, the lower the value of $I(g,f)$, the better the fitting of the model. Notice that the value of $I(g,f)$ depends on the distribution of $g(x)$, which is usually unknown. So we need an estimator of $I(g,f)$. From the definition of K-L information we have:

$$I(g,f) = \int_{-\infty}^{+\infty} g(x)\log g(x)dx - \int_{-\infty}^{+\infty} g(x)\log f(x)dx.$$

The first term in the last equality does not depend on the model distribution $f(x)$, so finding the minimum of K-L information is equivalent to finding the maximum of the value:

$$\int_{-\infty}^{+\infty} g(x)\log f(x)dx, \tag{2}$$

which is called an expected log likelihood. If we have n independent realizations $\{X_1,....,X_n\}$, then the value :

$$\sum_{i=1}^{n} \log f(x_i), \tag{3}$$

is called the log likelihood of the model. It is easy to prove that an expected log likelihood can be approximated by (3) times 1/n [24]. So, the higher the value of (3), the better the fitting of the model.

Assume that $f(x_1,...,x_n \mid \theta)$ is a joint distribution function of the vector $(X_1,...,X_n)$ θ, where is a parameter of the distribution. If we have n observations $\{x_1,....,x_n\}$ then the function $L(\theta) = f(x_1,...x_n \mid \theta)$ is called the likelihood function. When we consider the independent random variables $\{X_1,....,X_n\}$, then $L(\theta) = f(x_1 \mid \theta) \cdot ... \cdot f(x_n \mid \theta)$, where $f(x_i \mid \theta)$ is a density function of X_i, we can define the log likelihood function as

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i \mid \theta).$$

Let us identify $g(x) = f(x \mid \theta^*)$, where f is a model with k parameters and θ^* is a vector of true parameters. Then the expected log likelihood of the distribution $f(\cdot \mid \theta)$ is given by [24]:

$$l^*(\theta) = E_Z[\log f(Z \mid \theta)],$$

where Z is a random variable with the same distribution as X_i and independent of X. As we said, this value is a criterion of fitting of the distribution. The higher the value of $l^*(\theta)$, the better the fitting of the model.

Let $\hat{\theta}$ be the maximum likelihood estimator of the parameters of the model maximizing the log likelihood function $l(\theta)$. So the goodness of fitting of the model can be expressed in terms of $l^*(\hat{\theta})$. Observe that this value depends on the realization of the random variable X. Therefore, in order to lose this dependence, define the mean expected log likelihood as:

$$l_n^*(k) = E_X[l^*(\hat{\theta}_k)] = \int l^*(\hat{\theta}) \prod_{i=1}^{n} g(x_i)dx.$$

As before, the higher the value of $l_n^*(k)$, the better the fitting of the model. At first sight, it would seem that the maximum log likelihood is a good estimator of the mean expected log likelihood. However, as Akaike has shown, it is a biased estimator of this value and its bias is equal to the number of the parameters. Akaike [25] has also shown that the asymptotically unbiased estimator of the mean expected log likelihood is:

AIC(k) = the maximum log likelihood – number of estimated parameters.

By historical reasons [Rao 1980] we take:

$$AIC(k) = -2l(\hat{\theta}_k) + 2k. \tag{4}$$

Summarizing, the model which has the minimal value of AIC(k) is considered to be the most appropriate model.

ESTIMATION OF PARAMETERS OF NORMAL DISTRIBUTION BY THE MAXIMUM LIKELIHOOD METHOD

Let $X_1, X_2, ..., X_n$ be the independent n observations of the normal distribution $N(\mu, \sigma^2)$, where $\mu \in R, \sigma^2 > 0$. Let us consider the normal density function of the random variable X_i:

$$f(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}.$$

The likelihood function is given by

$$L(\mu, \sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \tag{5}$$

Thus

$$l(\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma^2.$$

We also have

$$\frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \text{ and } \frac{\partial l}{\partial \sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^2}. \tag{6}$$

We will find the maximum value of the function $l(\cdot,\cdot)$ and the corresponding maximum argument. The necessary condition for the existence of the maximum of this function is $\frac{\partial l}{\partial \mu} = 0$ and $\frac{\partial l}{\partial \sigma^2} = 0$. Solving the system

$$\begin{cases} \sum_{i=1}^{n}(x_i-\mu) = 0 \\ \sum_{i=1}^{n}(x_i-\mu)^2 - n\sigma^2 = 0, \end{cases} \tag{7}$$

we get $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$.

We can write the solution of the above system as:

$$\hat{\mu}=\frac{1}{n}\sum_{i=1}^{n}x_i,\ \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2. \tag{8}$$

Now, we should check if the function $l(\cdot,\cdot)$ has the maximum in $(\hat{\mu},\hat{\sigma}^2)$. We will check if $l(\hat{\mu},\hat{\sigma}^2)-l(\mu,\sigma^2)\geq 0$. Let us consider the difference:

$$l(\hat{\mu},\hat{\sigma}^2)-l(\mu,\sigma^2)==-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})^2-\frac{n}{2}\log\hat{\sigma}^2+\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{2}\log\sigma^2=$$

$$=-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(x_i-\hat{\mu})^2-\frac{n}{2}\log\frac{\hat{\sigma}^2}{\sigma^2}+\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\hat{\mu})^2+n(\hat{\mu}-\mu)^2\right]=$$

$$=-\frac{n}{2}\log\frac{\hat{\sigma}^2}{\sigma^2}+\left(\frac{1}{2\sigma^2}-\frac{1}{2\hat{\sigma}^2}\right)\sum_{i=1}^{n}(x_i-\hat{\mu})^2+\frac{n(\hat{\mu}-\mu)^2}{2\sigma^2}=\frac{n}{2}\left(\frac{\hat{\sigma}^2}{\sigma^2}-1-\log\frac{\hat{\sigma}^2}{\sigma^2}\right)+\frac{n(\hat{\mu}-\mu)^2}{2\sigma^2}.$$

The last term in the equation is nonnegative, so we can observe the function $f(x)=\log x-x$. The function f has a maximum in $x=1$ and $f(1)=-1$. Thus $\log x-x\leq -1$ for $x>0$.

So we have $l(\hat{\mu},\hat{\sigma}^2)-l(\mu,\sigma^2)\geq 0$, which implies that function $l(\mu,\sigma^2)$ has the maximum at $\mu=\hat{\mu}$ and $\sigma^2=\hat{\sigma}^2$. To summarize, $\hat{\mu}$ and $\hat{\sigma}^2$ are the maximum likelihood estimators of the parameters μ and σ^2. Also the maximum log likelihood is given by:

$$l(\hat{\mu},\hat{\sigma}^2)=-\frac{n}{2}\log 2\pi\hat{\sigma}^2-\frac{n}{2}. \tag{9}$$

ESTIMATION OF PARAMETERS OF LOGNORMAL DISTRIBUTION BY THE MAXIMUM LIKELIHOOD METHOD.

Def. Let random variable Y have a normal distribution: $Y\sim N(\mu,\sigma^2)$. Then a random variable $X=e^Y$ has a lognormal distribution with parameters μ,σ^2. The probability density function of the lognormal random variable is given by the formula;

$$f(x)=\begin{cases}0, & x\leq 0\\ \dfrac{1}{x\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, & x>0\end{cases} \tag{10}$$

Let $X_1,X_2,...,X_n$ be the sample from lognormal distribution with parameters μ,σ^2. $\mu\in R,\sigma^2>0$. Let us consider the lognormal density function of the random variable X_i:

$$f(x_i\mid\mu,\sigma^2)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(\ln x_i-\mu)^2}{2\sigma^2}\right\}\text{ dla x>0.}$$

The likelihood function is given by:

$$\begin{cases}L(\mu,\sigma^2)=f(x_1,x_2,\ldots,x_n I\sigma^2,\mu)=\prod_{i=1}^{n}f(x_i I\sigma^2,\mu)\\ =(2\pi\sigma^2)^{-\frac{n}{2}}\prod_{i=1}^{n}(x_i)^{-1}\cdot e^{-\frac{\sum_{i=1}^{n}(\ln x_i-\mu)^2}{2\sigma^2}}.\end{cases} \tag{11}$$

Thus log likelihood is equal to:

$$l(x|\sigma^2,\mu) = -\frac{n}{2}(\ln 2\pi + \ln\sigma^2) - \sum_{i=1}^{n}\ln x_i - \frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{2\sigma^2} \tag{12}$$

In order to find the maximum likelihood estimators $\hat{\mu}\ i\ \hat{\sigma}^2$ of parameters of distribution we differentiate the function with respect to μ i σ^2 i and then equate it to zero. We get the following:

$$\begin{cases} \dfrac{\partial l}{\partial \mu} = -\dfrac{1}{2\sigma^2}\sum_{i=1} 2(\ln x_i - \mu)(-1) = \dfrac{1}{\sigma^2}\sum_{i=1}^{n}(\ln x_i - \mu) = 0 \\ \dfrac{\partial l}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{2}\cdot\dfrac{1}{\sigma^4} = 0. \end{cases} \tag{13}$$

After the transformation of the system (13) we obtain equalities:

$$\begin{cases} \sum_{i=1}^{n}(\ln x_i - \mu) = 0 \\ -n\sigma^2 + \sum_{i=1}^{n}(\ln x_i - \mu)^2 = 0, \end{cases} \tag{14}$$

and from it the following estimators:

$$\begin{cases} \hat{\mu} = \dfrac{\sum_{i=1}^{n}\ln x_i}{n} = \overline{\ln x} \\ \hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2}{n} = S^2_{\ln x}. \end{cases} \tag{15}$$

Moreover, the maximum log likelihood is equal to (from(12) and (15)):

$$\{l(x|\hat{\sigma}^2,\hat{\mu}) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\hat{\sigma}^2 - \sum_{i=1}^{n}\ln x_i - \frac{n}{2}. \tag{16}$$

SELECTION OF NORMAL DISTRIBUTION BY THE AKAIKE METHOD

We want to verify if the investigated feature in population has a normal or lognormal distribution. We state hypotheses:

$$\begin{cases} H_0 : F(x) \in F_N \\ H_{1_1} : F(x) \in F_{LN}, \end{cases} \tag{17}$$

where: F_N denotes a family of normal cumulative distribution functions, F_{LN} a family of lognormal cumulative distribution functions and F(x) a cumulative distribution function of feature in population.

In Akaike method we match to each of hypotheses a model connected with specified probability distribution, evaluate Akaike information criterion AIC for each model and then we choose this hypothesis, which corresponds to the lower value of AIC.

Model (0)-the feature in population has normal distribution.

Model (1)-the feature in population has lognormal distribution.

From (4) we have $AIC = -2l(\hat{\theta}) + 2k$. As the number of free parameters in the model (0) is 2 so AIC for it is equal to (from (9)):

$$\left\{AIC(0) = AIC(\mu,\sigma^2) = -2l(\hat{\mu},\hat{\sigma}^2) + 2x2 = n[\log 2\pi + \log\hat{\sigma}^2 + 1] + 4\right\} \tag{18}$$

where $\hat{\sigma}^2$: is given by (8).

Similarly, for model (1) we find (from (16)):

$$AIC(1) = -2l(\hat{\mu},\hat{\sigma}^2) + 2x2 = n[\log 2\pi + \log\hat{\sigma}^2 + 1] + 2\sum_{i=1}^{n}\ln x_i + 4 \tag{19}$$

where in its time $\hat{\sigma}^2$ is given by (15).

APPLICATION

The Akaike theory given in previous sections we apply to an analysis of experimental data from the crop of sugar beets. Proper investigations were carried out in the years 2006-2008. [Ostroga 2010, Bzowska-Bakalarz et al 2008]. The tool for assembling the data in farms were questionnaires filled through growers. Area of questionnaires' investigations contained region of working of four sugar factories - Krasnystaw, Lublin, Werbkowice and Strzyżów. Respondents were farmers who had a chance of survival of production of sugar beets in next years. They became chosen at co-operation with workers of sections of raw material sugar factory on the basis of positive productive results. Investigations were carried out by the method of standardized interview. Analyzing dataset related to crops of sugar beet received through 100 manufacturers in 2006 year we obtain $\hat{\sigma}^2$ =75,035, n=100 while from example (18) we get:

$$AIC(0) = 100\cdot\left[1,8379 + \ln(75,035) + 1\right] + 4 = 719,5855. \tag{20}$$

For dataset $\hat{\sigma}^2$ =0,0364, $\sum_{i=1}^{n}\ln x_i$=381,89, n=100 from example (19) we receive:

$$AIC(1) = 100\cdot[1,8379 + \ln(0,0364) + 1] + 2\cdot 381,89 + 4 = 720,2513 \tag{21}$$

In the face of inequality AIC(1)>AIC(0) we state, that normal distribution better than log normal distribution describes crops of beets. Now we will try to apply the traditional method of testing of hypotheses for analysis of experimental dataset. We will use classic test Shapiro-Wilka to this end. Test function of this test has the form [Domański 1990]:

$$W = \frac{\{\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{(i)}(x_{(n-i+1)} - x_{(i)})\}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{22}$$

where: $x_{(i)}$ denotes i-th; within growing crop of beet., $a_{(i)}$ fixed value in Shapiro-Wilka test read from statistical tables, meanwhile symbol [z] denotes the largest integer number less or equal to z. In our example we have: n=100, $\sum_{i=1}^{\left[\frac{n}{2}\right]} a_{(i)}(x_{(n-i+1)} - x_{(i)}) = 85,39$, $nS^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 == 7503,5202$.

Therefore in the face of (22) we get:

$$W = \frac{(85,39)^2}{7503,5202} = 0,9717. \tag{23}$$

Assuming the level of significance α =0,05 we read the critical value $W_{0,05}$=0,9641. In the face of W> $W_{0,05}$ we state, that it does not have bases for rejection of hypothesis about normality of distribution of crop of sugar beets.

REFERENCES

Akaike H.1974.: A new look at the statistical model identification. IEEE.Trans.Autom.Contr, AC-19, 716-723.

Anderson T.W., Darling D.A. 1954.: A test of goodness of fit JASA 49, 765-769.

Andrews D.F., Gnanadesikan R., Warner J.L. 1973.:Methods for assessing multivariate normality. In P.R Krishnaiah, ed, Multivariate Analysis, 3rd ed., Academic Press, New York.

Bera A., John. 1986.:Test for multivariate normality with Pearson alternatives. Commun in Statist. 12, 103-117.

Box G.E.P., 1953.:A note on regions for tests of kurtosis. Biometrika 40, 465-468.

Bowman K.O., Shenton L.R. 1975.:Omnibus test contours for depature from normality based on $\sqrt{b_1}$ and b_2.

Bzowska-Bakalarz M., Gil K. 2008.:Analiza funkcjonowania bazy danych do monitorowania systemu produkcji buraków cukrowych. Inzynieria Rolnicza 2,15-21.

Cox D.R., Small N.J.H. 1978.:Testing Multivariate normality. Biometrika 65, 263-262.

D'Agostino R.B. 1971.:An omnibus test for normality for moderate and large size samples. Biometrika 58, 341-348.

Domański C. 1990.:Testy statystyczne. PWE, Warszawa.

Franck W.E. 1981.:The most powerful invariant test of normal versus Cauchy with applications to stable alternatives. JASA 76, 1002-1005.

Geary, R.C. 1947.: Testing for normality. Biometrika.34, 209-242.

Green J.R., Hegazy Y.A.S. 1976. Powerful modified EDF goodness of fit test. JASA 71, 204-209.

Hawkins D.M. 1981.:A new test for multivariate normality and homoscedasticity. Technometrics 23, 105-110.

Henze, N., Zirkler, B. 1990.:A class of invariant consistent tests for multivariate normality. Commun in Statist. 19, 3595-3617.

Kozioł J.A. 1986.:Assesing multivariate normality: a compendium Commun in Statist 15, 2763-2783.

Kulback Leibler. 1951.: On information and sufficiency. Ann.Math.Statist, 22, 79-86.

Lin, C.C. Mudholkar G.S. 1980.:A simple test for normality against non-symmetric alternatives. Buiometrika 67, 455-461.

Locke C., Spurrier J.D. 1976.:The use of U-Statistics for testing normality against non-symmetric alternatives. Biometrika 63, 143-147.

Oja H. 1983**.:**New tests for normality. Biometrika 70, 297-299.
Ostroga K 2010.:Techniczne i technologiczne uwarunkowania produkcji buraków cukrowych na tle aktualnej organizacji rynku rolnego**.** Rozprawa doktorska. Wyd UP Lublin.
Pearson E.S. 1930:A further development of tests for normality. Biometrika 22, 239-249.
RaoC.R. 1980.:Modele liniowe statystyki matematycznej PWN Warszawa.
Royston J.P. 1983.:Some techniques for assessing multivariate normality based on the
Shapiro-Wilk W. Appl Statist 32, 121-133.
Sakamoto .Y., Ishiguro M., Kitagawa G. 1986.:Akaike Information Criterion Statistics. Mathematics and Its Application.

ZASTOSOWANIE KRYTERIUM AKAIKE DO OCENY ROZKŁADU PLONÓW BURAKÓW CUKROWYCH

Streszczenie. Rozkład normalny jest powszechnie stosowany w naukach przyrodniczych, technicznych jak również humanistycznych i społecznych. Bardzo często jest on również używany do opisu zjawisk i procesów w badaniach z zakresu inżynierii rolniczej. Szerokie wykorzystanie tego rozkładu tłumaczą prawa wielkich liczb. Znaczenie rozkładu normalnego wynika też z następujących faktów: 1) Rozkład normalny jest modelem dla losowych błędów pomiarów. 2) Wiele zjawisk fizycznych, choć nie podlega rozkładowi normalnemu może być opisanych za pomocą tego rozkładu po odpowiedniej transformacji. 3) Rozkład normalny jest dobrym przybliżeniem dla innych rozkładów, 4) Każda kombinacja liniowa niezależnych zmiennych losowych o rozkładzie normalnym ma rozkład normalny.

Te pożądane własności powodują, iż badacze dysponując wynikami eksperymentu chcą sprawdzić czy mają one rozkład normalny. W literaturze opisanych jest wiele testów sprawdzających normalność rozkładu badanej cechy zarówno w przypadku jedno [Anderson i in 1954, Bowman i in 1975, Geary 1947,Green i in 1976,Lin i in 1980,Oja 1983,Pearson 1930, Shapiro i in 1965, Uthoff 1973, Vasicek, 1976] jak i wielowymiarowym [Andrews i in 1973, Bera i in 1986, Cox i in 1978, Hawkins 1981, Henze i in 1990, Kozioł 1986, Royston 1983]. Wszystkie one wymagają stosowania odpowiednich tablic wartości krytycznych.

W niniejszej pracy proponuje się zastosowanie do selekcji między rozkładem normalnym i logarytmo-normalnym kryterium informacyjnego Akaike. Kryterium to wywodzi się z teorii informacji. Sugerowaną metodę zastosowano do danych uzyskanych w doświadczeniu z plonami buraków cukrowych .Wnioski wynikające z kryterium Akaike potwierdzono klasycznym testem Shapiro-Wilka.

Słowa kluczowe: Plony buraków cukrowych, kryterium informacyjne Akaike, rozkład normalny, rozkład logarytmo-normalny, selekcja modeli.