DOI: 10.1515/bile-2015-0007

DE GRUYTER
OPEN

Biometrical Letters
Vol. 52 (2015), No. 2, 75 - 84

The use of outlier detection methods in the log-normal distribution for the identification of atypical varietal experiments

Andrzej Kornacki, Andrzej Bochniak

Department of Applied Mathematics and Computer Science, University of Life Sciences in Lublin, Akademicka 15, 20-950 Lublin, Poland, e-mail: andrzej.kornacki@up.lublin.pl

SUMMARY

In this study the Akaike information criterion for detecting outliers in a log-normal distribution is used. Theoretical results were applied to the identification of atypical varietal trials. This is an alternative to the tolerance interval method. Detection of outliers with the help of the Akaike information criterion represents an alternative to the method of testing hypotheses. This approach does not depend on the level of significance adopted by the investigator. It also does not lead to the masking effect of outliers.

Key words: outliers, log-normal distribution, atypical variety trials, hypothesis testing, masking of outliers, wheat, entropy.

1. Introduction

In analyzing the results of series of varietal experiments, a question often arises concerning the disqualification of some experiments. This is especially true when the precision of the experiments is not uniform. One may then reject an experiment with a large experimental error, which is a vague procedure, based on the suspicion that errors were made in the conduct of the experiment. For the detection of atypical experiments, Ohanowicz and Pilarczyk (1985) proposed tolerance intervals for the smallest significant difference expressed as a percentage of the average yield (known NRIP). It turned out that the distribution of the NRIP is clearly asymmetrical. It gives a good approximation to the log-normal distribution. In this paper we propose to treat atypical experiments (observations) as outliers, and it is suggested that the Akaike information criterion may be used to detect such outliers.

For the detection of outliers, hypothesis testing methods are most frequently used (Barnett and Lewis, 1994; Breuning et al., 2000; Ferguson, 1961; Grubbs, 1960, 1969; Ramaswamy et al., 2000; Rousseeuw and Leroy, 2000; Srivastava and Von Rosen, 1998). However, in the hypothesis testing method the conclusions may differ depending on the assumed significance level. Moreover, the “masking” effect for outliers may be encountered. Grubbs (1969), in relation to data on the strength of plastic materials, describes a situation where the tests do not detect one least observation, whereas two least observations are identified as outliers (an apparent contradiction).

The use of the Akaike information criterion, as suggested here, allows one to choose, out of the models describing the experimental data, that model which maximizes entropy (Akaike, 1973, 1977). According to Sakamoto et al. (1986), the value for this criterion equals:

$$\text{AIC} = -2\, ln(\text{max likelihood}) + 2\text{K} \tag{1}$$

where max likelihood denotes the likelihood calculated for parameter estimators obtained using the maximum likelihood method, and K denotes the number of these parameters. We select the model for which the AIC value is the smallest. This way of proceeding does not depend on the significance level, the number of outliers, or whether the “suspicious” observations are the lowest or the highest.

The aim of the study was to use an outlier detection method based on the Akaike information criterion to find atypical experiments on wheat varieties, and to compare this method with the method based on tolerance intervals.

2. Log-normal distribution

Definition: A random variable X has a two-parameter log-normal distribution with parameters (μ,σ^2), denoted LN(μ,σ^2), when its logarithm has a normal distribution, i.e. Y=lnX ~ N(μ,σ^2). Thus we have X ~ LN(μ,σ^2)$\Leftrightarrow$ Y=lnX ~ N(μ,σ^2).

Limpert, Stahel and Abbt (2001) tabulate some of the areas where the log-normal distribution is applied: in geology (concentration of the elements Co, Cu, Cr, ^{226}Ra, Au and U), in medicine (latent period of infectious diseases, survival periods after being diagnosed with cancer), in environmental science (rainfall, air pollution, particle decomposition, environmental chemistry and organisms), in food technology, in ecology (species resources), in linguistics (length of words uttered in telephone conversations), in social science (marriage age, income), in operations research (time distribution in queuing), and others.

It is known that the maximum likelihood estimators of the parameters in the log-normal distribution are equal (Krzyśko, 2004):

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

Let us consider a sample of n observations which, when arranged by increasing value, form the set $x_{(1)}\le x_{(2)}\le\ldots\le x_{(n)}$. Therefore, $x_{(k)}$ denotes the value of the k-th order statistics $X_{k:n.}$

In the remainder of the paper the following notation will be used: $\Phi(x,\mu,\sigma^2)$ denotes the cumulative density function for the distribution LN (μ,σ^2), $\Psi(x,\mu,\sigma^2)$ denotes the probability density function for the distribution LN (μ,σ^2), $f_{r,n}(x,\mu,\sigma^2)$ denotes the probability density function for the i-th order statistic $X_{i,n.}$

So we have:

$$\Psi(x,\mu,\sigma^2)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right],\ x>0 \tag{2}$$

$$\Phi(x,\mu,\sigma^2)=\frac{1}{2}+\frac{1}{2}erf\left[\frac{\ln x-\mu}{\sigma\sqrt{2}}\right], \text{ x>0} \qquad (3)$$

$$f_{r,n}(x,\mu,\sigma^2)=B(r,n-r+1)^{-1}\Phi(x,\mu,\sigma^2)^{r-1}(1-\Phi(x,\mu,\sigma^2)^{n-r}\Psi(x,\mu,\sigma^2)\ (4)$$

(David and Nagaraja, 2003), where *B(p,q)* denotes the beta function:

$$B(p,q)=\int_0^1 t^{p-1}(1-t)^{q-1}dt, \text{ p>0 q>0}$$

It is known that:

$$B(p,q)=\frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}=\frac{(p-1)!(q-1)!}{(p+q-1)!}$$

for natural numbers *p* and *q*, where $\Gamma(x)$ denotes Euler's function.

We write $EX=\mu^*=e^{\mu+\frac{\sigma^2}{2}}$, $VAR(X)=(\sigma^2)^*=(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$.

3. Outliers model

Let us consider the following situation: we have an ordered sample:

$$\underbrace{x_{(1)}\le x_{(2)}\le\ldots\le x_{(n_1)}}_{n_1\ lowest\ observations}\le\overset{main\ part\ of\ sample}{\overbrace{\underbrace{x_{n_1+1}\le\ldots\le x_{n-n_2}}_{n-n_1-n_2\ observations}}}\le\underbrace{x_{n-n_2+1}\le\ldots\le x_{(n)}}_{n_2\ highest\ observations}$$

from a log-normal distribution, where $x=(x_1,x_2,\ldots,x_n)$. Outliers may be the lowest or the highest observations coming from populations characterized by various mean values μ_1^* and μ_2^*. The 'main part' of the sample comes from the population with the mean μ^*. Detecting outliers usually involves using the hypothesis testing procedure at some significance level. In such a case the hypotheses take on the following form: H_0 – there are no outliers, that is to say $\mu_1^*=\mu^*=\mu_2^*$; H_{1a} – there are low outliers, that is to say $\mu_1^*<\mu^*$; H_{1b} – there are high outliers, that is to say $\mu^*<\mu_1^*$; H_{1c} – there are low and high outliers, that is to say $\mu_1^*<\mu^*<\mu_2^*$.

By setting the parameter values $\mu_1 < \mu < \mu_2$ respectively and σ the same for all groups, we obtain: $\mu_1^* < \mu^* < \mu_2^*$ and $\sigma_1^* < \sigma^* < \sigma_2^*$.

Thus the model with outliers can be described in the following way

$$h_i(x) = \begin{cases} \Psi(x,\mu_1,\sigma^2) & i = 1,\ldots,n_1 \\ f_{i-n_1,n-n_1-n_2}(x,\mu,\sigma^2) & i = n_1+1,\ldots,n-n_2 \\ \Psi(x,\mu_2,\sigma^2) & i = n-n_2+1,\ldots,n \end{cases} \tag{5}$$

4. Akaike information criterion

In order to calculate the Akaike criterion (1) we find the model likelihood function (5)

$$L(x;n_1,n_2,\mu,\mu_1,\mu_2,\sigma^2) = \prod_{i}^{n_1} \Psi(x_{(i)},\mu_1,\sigma^2) \cdot \prod_{i=n_1+1}^{n-n_2} f_{i-n_1,n-n_1-n_2}(x_{(i)},\mu,\sigma^2) \cdot$$

$$\cdot \prod_{n-n_2+1}^{n} \Psi(x_{(i)},\mu_2,\sigma^2)$$

Thus the likelihood logarithm with respect to (2–4) is equal to:

$$\begin{cases} l(x;n_1,n_2,\mu,\mu_1,\mu_2,\sigma^2) = \ln \prod_{i=1}^{n_1} \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x_i - \mu_1)^2}{2\sigma^2}\right] \cdot \\ \ln \prod_{i=n_1+1}^{n-n_2} B(i-n_1, n-n_1-n_2-i+n_1+1)\Phi(x_i,\mu,\sigma^2)^{i-n_1-1}(1-\Phi(x_i,\mu,\sigma^2)^{n-n_1-n_2-i+n_1} \cdot \\ \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right] \ln \prod_{i=n-n_2+1}^{n} \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln x_i - \mu_2)^2}{2\sigma^2}\right] \end{cases}$$

that is to say:

$$\begin{cases} l(x;n_1,n_2,\mu,\mu_1,\mu_2,\sigma^2) = -\frac{1}{2}\left\{ n\ln(2\pi) + n\ln(\sigma^2) + \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln x_i - \mu^i)^2 \right\} \\ -\sum_{i=1}^{n}\ln x_i - \sum_{i=n_1+1}^{n-n_2}[\ln B(j,k-j+1) - (j-1)\ln\{\Phi(x_i)\} - (k-j)\ln\{1-\Phi(x_i)\}] \end{cases}$$

where $j = i - n_1,\ k = n - n_1 - n_2$ and:

$$\mu^i = \begin{cases} \mu_1 \ dla \quad 1 \le i \le n_1 \\ \mu \ dla \quad n_1 < i \le n - n_2 \\ \mu_2 \ dla \quad n - n_2 < i \le n \end{cases}$$

Finally, the value of the Akaike information criterion is equal to

$$\begin{cases} -2l(x; i, j, \hat{\mu}, \hat{\sigma}^2) + 2 \times 2 \quad (i = 0, j = 0) \\ -2l(x; i, j, \hat{\mu}, \hat{\mu}_1, \hat{\sigma}^2) + 2 \times 3 \quad (i \ne 0, j = 0) \\ -2l(x; i, j, \hat{\mu}, \hat{\mu}_2, \hat{\sigma}^2) + 2 \times 3 \quad (i = 0, j \ne 0) \\ -2l(x; i, j, \hat{\mu}, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}^2) + 2 \times 4 \quad (i \ne 0, j \ne 0) \end{cases}$$

where $\hat{\mu}$, $\hat{\mu}_1$, $\hat{\mu}_2$, $\hat{\sigma}^2$ are the maximum likelihood estimators of the parameters.

5. Materials and methods

To compare the results obtained by the method used here and by the tolerance interval method, the original data presented by Pilarczyk (1988) were used in calculations. The basic experiment and exploratory evaluation of winter wheat varieties were carried out in 1983. It was a series of WGO experiments (study of the economic value of varieties). Each of them was conducted with 4 replications in a 1-resoluble design. In each incomplete block 15 to 33 different varieties were compared. The yield (average of all the tested cultivars) ranged from about 40 q/ha to 70 q/ha depending on the experimental station. Grain yield calculated for a common humidity of 15% was the studied feature. Next, values of NRIP were calculated for all experimental stations. Yields of individual varieties do not participate in the NRIP calculation. Table 1 shows the values of ln (NRIP) in the experiment with wheat.

6. Results and discussion

The distribution of NRIP values is distinctly asymmetrical (see Fig. 1). The skewness is equal to 1.35408.

Table 1. Values of ln(NRIP) in experiments with wheat

Values									
1.221	1.247	1.435	1.497	1.504	1.530	1.533	1.554	1.556	1.558
1.633	1.668	1.670	1.688	1.708	1.728	1.728	1.730	1.761	1.763
1.770	1.772	1.782	1.828	1.842	1.845	1.847	1.881	1.896	1.910
1.918	1.923	1.926	1.930	1.939	1.949	1.954	1.966	1.970	1.971
1.971	1.973	1.975	1.980	2.006	2.006	2.012	2.019	2.020	2.036
2.036	2.055	2.58	2.069	2.069	2.076	2.083	2.091	2.097	2.107
2.108	2.117	2.131	2.133	2.134	2.147	2.149	2.158	2.174	2.191
2.201	2.204	2.213	2.217	2.260	2.266	2.268	2.273	2.274	2.309
2.312	2.350	2.372	2.376	2.417	2.437	2.442	2.450	2.470	2.520
2.526	2.530	2.612	2.711	2.750	2.756	2.988			

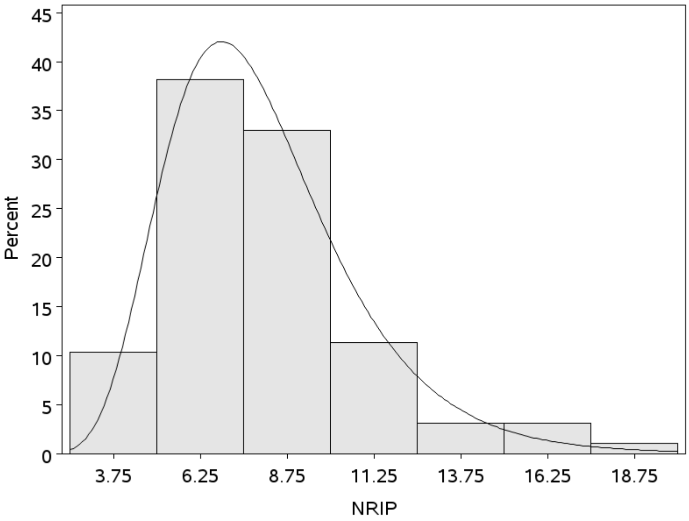

Figure 1. Distribution of NRIP values

Hence we use a transformation based on evaluation of the natural logarithms of the NRIP values. After the transformation, the results are well described by a normal distribution. This means that the NRIP values have a log-normal distribution. A verification of this fact is shown in Figure 1 and Table 2.

Table 2. Compatibility of NRIP with the log-normal distribution according to test statistics

Test of log-normal distribution		
Name of test	Statistics	p-value
Kolmogorov–Smirnov	D=0.0546	> 0.150
Cramer–von Mises	W^2=0.0531	0.474
Anderson–Darling	A^2=0.3018	> 0.500

It can be seen that all of the tests indicate no basis to reject the hypothesis of the log-normal distribution of the tested feature.

Using the theory presented in section 4, we can find outlier observations for the NRIP values of winter wheat. The results are presented in Table 3.

Table 3. Values of AIC for models with different numbers of outliers

		Number of high outliers					
		0	1	2	3	4	5
Number of low outliers	0	23.275	16.268	15.239	16.423	20.478	26.045
	1	25.147	15.995	13.278	12.828	15.426	19.782
	2	28.705	17.752	13.476	11.576*	12.920	16.180
	3	33.826	21.347	15.977	12.955	13.261	15.696
	4	38.320	24.423	18.089	14.040	13.369	15.049
	5	43.610	28.296	20.972	15.870	14.201	15.095

* smallest value of AIC

The outlier observations appear to be the three largest NRIP values (15.643; 15.174; 19.846) and the two least (3.391; 3.480). Therefore these indicate atypical experiments. The results obtained here are largely consistent with those of Pilarczyk (Pilarczyk 1988). That author found the following 95% interval of tolerance for ln (NRIP):

$$1.347 \le \ln(NRIP) \le 2.768 \tag{6}$$

or equivalently

$$3.846 \leq \text{NRIP} \leq 15.927. \tag{7}$$

There are three experiments that are located beyond the tolerance interval, those having the two smallest and the highest NRIP value. The same points are identified in this paper by the outlier detection method. Differences in the results concern only the experiments with the second and third highest NRIP values, which lie within the tolerance intervals (6) and (7), but were identified as outlier observations. We can notice that for these the values of ln(NRIP) are almost equal, at 2.750 and 2.756, and differ from the right end-point of the tolerance interval by the very small amounts of 0.018 and 0.012.

7. Conclusions

- A method of detecting outliers based on the Akaike information criterion is used here to find atypical specific experiments. This is an alternative to the tolerance interval method.
- The proposed method is an objective procedure independent of the adopted significance level, the number of outliers and whether the “suspicious” observations are the lowest or the highest.
- The conclusions obtained by the proposed method are largely consistent with the results of Pilarczyk (1988) based on the tolerance interval method**.**

REFERENCES

Akaike H. (1973): Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory. eds. B.N. Petrv and F. Csaki. Budapest; Akademia Kiado: 267-281.

Akaike H. (1977): On entropy maximization principle. Proc Symposium on Applications of Statistics. ed. P.R. Krishnaiah. Amsterdam: North Holland: 27-47.

Barnett V., Lewis T. (1994): Outliers in Statistical Data. John Wiley & Sons.

Breuning M., Kriegel H.P, Sander J. (2000): LOF: Identifying Density-Based Local. Proceedings of the ACM SIGMOND Conference: 93-104.

David H.A., Nagaraja H.N. (2003): Order Statistics. Wiley Series in Probability and Statistics.

Ferguson T.S. (1961): On the rejection of outliers. Proc. Fourth Berkeley Symposium Math. Statist. Prob.1: 253-287.

Grubbs F.E. (1960): Sample criteria for testing outlying observations. Ann. Math. Statist. 21: 27-58.

Grubbs F.E. (1969): Procedures for detecting outlying observations in samples. Technometrics 11: 1-21.

Krzyśko. M. (2004): Mathematical Statistics. Poznań, Wydawnictwo Naukowe UAM (in Polish).

Limpert. E., Stahel W., Abbt M. (2001): Log-normal distribution across the sciences: Kees and Clues. Bioscience. 51(5): 341-352.

Ohanowicz T., Pilarczyk W. (1985): Precision experiments with potato and detection of unusual experiments. XV Colloquium Metodologiczne z Agrobiometrii: 106-115 (in Polish).

Pilarczyk W. (1988): The effectiveness of varietal trials with cereals and detection of untypical experiments. Biuletyn Oceny Odmian. 13: 115-123 (in Polish).

Ramaswamy S., Rastogi R., Shim K. (2000): Efficient algorithms for mining outliers from large data sets. Proceedings of the ACM SIGMOND Conference: 427-438.

Rousseeuw P., Leroy A. (2003): Robust Regression and Outlier Detection. John Wiley & Sons.

Sakamoto Y., Ishiguro M., Kitagawa G. (1986): Akaike Information Criterion Statistics. Tokyo Reidel Publishing Company.

Srivastava M.S., Von Rosen D. (1998): Outliers in Multivariate Regression Models. J. Mult. Anal. 65: 195-208.

Stefansky W. (1972): Rejecting outliers in factorial designs. Technometrics 14: 469-479.