Using Maximum Likelihood in the Detection of Outliers

Fernando Rosado
Faculdade de Ciencias de Lisboa
Departamento de Estatistica
Campo Grande Bloco C2
1749-016 Lisboa
Portugal
fernando.rosado@fc.ul.pt

Keywords: Outliers; Discordancy tests; Maximum Likelihood; Generative Model; Natural Alternative Hypothesis.

The problem of how to deal with data which contain outliers has long been a source of concern to experimenters and data analysts.

The general problem (of rejection of outliers) is a very old and common one. In its simplest form, it may be stated as follows: in a sample of moderate size taken from a certain population, it appears that one or two values are surprising far away from the main group. After selecting "a priori" one observation, the analyst performs some test to evaluate the discordancy of this observation.

The main problem is to introduce some degree of objectivity into the rejection of the outlying observation.

Outliers are usually defined as those observations that seem to be inconsistent with the rest of the data. They can be caused by an error on the records. But it is important to know if a discordant observation is authentic or if it indicates the presence of another distribution on the data.

Graphic methods are often used to identify those aberrant values. The several approaches and the other types of graphics turn strongly subjective that preliminary analysis of outliers which may leed to the wrong conclusion about the presence or absence of discrepant values in a sample. On that first analysis of the data there is right from the start, an influence of the method for the search of possible discordant observations. Rosado (1984) approaches and analyses that question.

The outlier notion is obviously influenced by the considered discordancy model. There has been several the models that allow the justification of the presence of discordant observations in a sample, of which depending the outlier's theory - Barnett and Lewis (1994) and Rosado (1984).

From a theoretical point of view, the most important general method in the statistical estimation so far known is the method of maximum likelihood. This is a construtive method of obtaining estimators which, under certain conditions, have desirable properties. A method of test construction closely allied to it is the likelihood ratio method. It has played an important role in the theory of tests.

Apart from intuitively based procedures, the maximum likelihood ratio principle is a widely applicable method for setting up discordancy tests for outliers. Using this principle we construct one general discordancy test assuming one general discordancy model. An objective criterion for the detection and selection of outliers it is suggested by Rosado (2001) - Generative Model with Natural Alternative (GAN Model) and the maximum likelihood ratio principle gives

$$(1) \qquad T(x_1,\ldots,x_n) = \frac{\max L_i}{L_0}$$

as a statistic to test the homogeneity of one sample and

$$T(x_1,\ldots,x_n) > c$$

is the region of rejection.

REFERENCE

Barnett, V. e Lewis, T. (1994) - Outliers in Statistical Data. (3rd edition). Wiley.
Rosado, F. (1984) - Existência e Detecção de Outliers. Uma Abordagem Metodológica. Tese de

Doutoramento. Universidade de Lisboa.
Rosado, F.(2001) - Outliers em Dados Estatí sticos - O Passado e o Presente. E o futuro? Proceedings of the VII Congresso Anual da Sociedade Portuguesa de Estatí stica. Um Olhar sobre a Estatí stica, p. 90-110.
Rosado, F.(2001) - A General Approach to Detect Outliers. (to be submitted).

RESUME

Using the statistic (1) we clarify the definition of outlier. Some important results on the influence of the dimension of the sample we can get with that statistic. We apply this results to normal and exponential populations.