



Working Paper nº 17

A qualidade dos dados em estatística
Outliers e Observações Discordantes
numa amostra

Fernando Rosado
Working Paper nº 17

ISSN: 0872 - 895X

Depósito Legal nº: 84842/94

Fernando Rosado
DEIO, Universidade de Lisboa

Novembro 1994

Trabalho apresentado no âmbito do Seminário de Estatística - coordenado pelo
Professor Doutor Bento Murteira

A qualidade dos dados em estatística Outliers e Observações Discordantes numa Amostra

Fernando Rosado

Faculdade de Ciências da Universidade de Lisboa

Resumo

Os princípios e a metodologia estatística estão hoje em dia bastante consolidados. As questões práticas sobre a análise dos dados são, em vista das disponibilidades computacionais, cada vez mais importantes. É fundamental o estudo de valores discordantes, por exemplo em problemas de modelação estatística. Aqui tem particular relevo o uso de software "identificando" observações. Em sintonia com um básico estudo de teoria dos outliers todas estes problemas serão abordados.

Palavras chave - outliers; modelação estatística; observações influentes; software estatístico.

Abstract

Statistical principles and methods has for many years been an important part of training programmes for statisticians, and there is now broad agreement on how it should be presented.

Good data analysis starts from carefully phrased questions. A very common mistake is to start instead from a predecided method of analysis.

It is very easy to fall into the trap of doing an statistical inference because the data have a form amenable. In practice, there are no standard problems, only standard solutions.

From the statistical point of view it is very important to study discordant values, for instance, in model identification and model fitting.

Most statistical software deal with those observations, "identifying" them. Using subjective criteria the influence on statistical inference may be decisive. We need to pay careful attention to the data - and its quality - and to the underlying theory or structure of the problem.

Clearly computers have an important role to play in teaching statistics. We study how important is to generate genuine data in which we have confidence to provide satisfactory statistical inference.

Keywords: statistical software; outliers; model identification and model fitting.

1. Sobre a noção de observação discordante

Começamos com algumas citações retiradas da vasta literatura estatística sobre observações discordantes numa amostra (as referências em itálico são nossas).

" The problem of how to deal with data which contain 'outliers', i.e. observations which *look suspicious* in some way, has long been a source of concern to experimenters and data analysts." (Guttman e Smith, 1969)

" It is well recognised by those who collect or analyse data that values occur in a sample... which are so far removed from the remaining values that the analyst is not willing to believe that these values have come from the same population. Many times values occur which are dubious *in the eyes of the analyst* and he feels that he should make a decision as to *whether to accept or reject these values as part of his sample.*" (Dixon, 1950)

"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 surprisingly far away from the main group...*" (Ferguson, 1961)

"The problem ... is to introduce some degree of objectivity into the rejection of *the outlying observations.*" (Ferguson, 1961)

"The outliers are values which *seem* either too large or too small as compared to the rest of the observations." (Gumbel, 1960)

"One or more of the observations may *have the appearance* of being outliers and we are interested here in determining ... whether *such observations* should be retained in the sample for interpreting results or whether they should be regarded as being inconsistent with remaining observations." (Grubbs, 1950)

"An outlying observation, or outlier, is one that *appears* to deviate markedly from other members of the sample in which it occurs." (Grubbs, 1969)

Estas citações são suficientemente elucidativas da natureza subjectiva dos métodos de rejeição e identificação de outliers numa amostra. Em síntese, todos consideram que, um outlier é sempre uma observação inconsistente com os dados.

Pesquisas na literatura dos séculos XVIII e XIX têm provado ser bastante antiga a sensibilidade dos estatísticos relativamente ao problema "outlier".

Uma completa discussão dos primeiros escritos sobre rejeição de outliers é apresentada por Harter (1974-76).

Um dos primeiros trabalhos salientando a necessidade de tratar de modo especial as observações surpreendentes foi elaborado por Bernoulli (1777).

O estudo de Beckman e Cook (1983) faz uma excelente síntese sobre o tratamento estatístico de outliers, quer do ponto de vista histórico quer das aplicações aos modelos standard na estatística e ironicamente, concluindo que: "Although much has been written, the notion of an outlier seems as vague today as it was 200 years ago". Por exemplo, Edgeworth em 1887 escreveu: "Discordant observations may be defined as those which present the appearance of differing in respect of their law of frequency from other observations with which they are combined"; e Grubbs em 1969, oitenta e dois anos mais tarde, afirma que "An outlying observation, or outlier, is one that appears to deviate markedly from the other members of the sample in which it occurs".

Estas afirmações são relevantes no sentido de que um outlier é de facto um conceito subjectivo e "post-data".

2. Modelo de discordância com alternativa natural

Consideremos uma amostra aleatória simples $x_1, \dots, x_n$ de dimensão n numa população com função densidade de probabilidade $f(x_i; \delta_i)$. Seja $L(x_1, \dots, x_n; \delta_1, \dots, \delta_n)$ a função de verosimilhança.

Definamos o seguinte modelo geral:

- Pela hipótese H_0 , de homogeneidade, admitimos que todas as observações $x_1, \dots, x_n$ têm a mesma densidade $f(x_i; \delta)$, ($i=1, \dots, n$). Observada uma amostra, seja L_0 função de verosimilhança correspondente na hipótese H_0 .

- Pela hipótese H , a alternativa natural, iremos admitir a presença de um valor discordante na amostra e que, em princípio, pode ser qualquer uma das n observações.

Seja então x_j com densidade $f(x_j; \delta')$ para algum $j \in (1, \dots, n)$ e sejam $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n$ observações com a mesma função densidade de probabilidade $f(x_i; \delta)$, ($i \neq j$). Representemos por L_j a função de verosimilhança da amostra quando válida a hipótese alternativa natural.

a) Teste de homogeneidade

Por aplicação do teste da razão de verosimilhanças nas hipóteses H_0 e H , obtemos o quociente

$$I_n = \frac{\hat{L}_0}{\max (\hat{L}_0, \max \hat{L}_j)}$$

onde $\hat{L}_0$ e $\hat{L}_j$ são os máximos das funções de verosimilhança naquelas hipóteses respectivamente e a partir do qual uma regra de teste

$$I_n < c_0$$

pode então ser construída e correspondente à rejeição da hipótese H_0 de homogeneidade da amostra. Uma estatística de teste é, por exemplo,

$$T(x_1, \dots, x_n) = \frac{\max \hat{L}_j}{\hat{L}_0}$$

A região de rejeição da hipótese H_0 é então definida por

$$T(x_1; \dots; x_n) > c.$$

Se para uma amostra $x_1, \dots, x_n$ a correspondente estatística de teste $T(x_1, \dots, x_n)$ não verifica a anterior condição então o método generativo com alternativa natural conduz à aceitação da homogeneidade, negando-se portanto a existência de alguma observação discordante e terminando nesta segunda fase a pesquisa de outliers nessa amostra.

b) Seleção do Outlier

Supondo rejeitada a hipótese H_0 de homogeneidade, aceitando-se então a existência de uma observação discordante como conduzidos pelo próprio método, ao valor x_j que deve ser considerado como outlier.

Na metodologia proposta, a observação que deve ser considerada discordante, se tal for o caso, é definida a posteriori. Este resultado torna-se particularmente importante na pesquisa de "outliers inesperados" como acontece em alguns modelos para populações normais.

3. Aplicação em POPULAÇÕES NORMAIS

a) Modelo de discordância com outlier por μ

Este modelo de discordância considera que na hipótese H_0 de homogeneidade, todas as observações $x_1, \dots, x_n$ têm a mesma densidade normal $n(x; \mu, \sigma)$ e admite σ conhecido.

Pela alternativa natural, existe uma observação x_j com função densidade de probabilidade $n(x_j; \mu', \sigma)$ e todas as restantes têm densidade $n(x; \mu, \sigma)$.

O MODELO GERAL acima definido permite construir a seguinte tabela onde se indica uma estatística de teste S equivalente a $T(x_1, \dots, x_n)$, para decidir sobre a discordância de um candidato a outlier.

Candidato a outlier		Estatística S
μ conh.	μ' conh.	
	$x_{(1)}$ se $\mu > \mu'$	$(\mu - \mu') \min ((x_j - \mu)/\sigma)$
	$x_{(n)}$ se $\mu < \mu'$	$(\mu' - \mu) \max ((x_j - \mu)/\sigma)$
μ conh.	μ' desc.	
	$x_{(1)}$ ou $x_{(n)}$	$\max ((x_j - \mu)/\sigma)^2$
μ desc.	μ' desc.	
	$x_{(1)}$ ou $x_{(n)}$	$\max ((x_j - \bar{x})/\sigma)^2$

b) Modelo de discordância com outlier por σ

Este modelo de discordância considera que na hipótese H_0 de homogeneidade, todas as observações $x_1, \dots, x_n$ têm a mesma densidade normal $n(x; \mu, \sigma)$ e admite μ conhecido. Pela alternativa natural, existe uma observação x_j com função

densidade de probabilidade $n(x; \mu, \sigma')$ e todas as restantes observações têm densidade $n(x; \mu, \sigma)$.

O MODELO GERAL acima definido permite construir a seguinte tabela onde se indica uma estatística de teste S equivalente a $T(x_1, \dots, x_n)$, para decidir sobre a discordância de um candidato a outlier.

Candidato a outlier	Estatística S
σ conh. σ' conh.	
$x_{(1)}$ ou $x_{(n)}$ se $\sigma < \sigma'$	$(\sigma^2 - \sigma'^2) \max ((x_j - \mu)/\sigma)^2$
$x_{(\mu)}$ se $\sigma > \sigma'$	$(\sigma^2 - \sigma'^2) \min ((x_j - \mu)/\sigma)^2$
σ conh. σ' desc.	
$x_{(\mu)}, x_{(1)}$ ou $x_{(n)}$	$\max \frac{s}{ x_j - \mu } \exp \frac{(x_j - \mu)^2}{2\sigma^2}$
σ desc. σ' desc.	
$x_{(\mu)}, x_{(1)}$ ou $x_{(n)}$	$\min \left(\frac{(x_j - \mu)^2}{\sum (x_i - \mu)^2} \right)^{1/2} \left(1 - \frac{(x_j - \mu)^2}{\sum (x_i - \mu)^2} \right)^{(n-1)/2}$

e onde representamos por $x_{(\mu)}$ a observação mais próxima do valor médio conhecido μ .

Referências

- Barnett, V. e Lewis, T. (1994) - Outliers in Statistical Data. (3rd edition). John Wiley.
- Beckman, R.J. e Cook, R.D. (1983) - Outlier.....s. *Technometrics* 25, 119-63 (com discussão)
- Dixon, W.J. (1950) - Analysis of Extreme Values. *Ann. Math. Stat.* 21, 488-506.
- Ferguson, T.S. (1961) - On the Rejection of Outliers. *Proceedings of the 4th Berkeley Symposium*, 1, 253-87.
- Grubbs, F.E. (1950) - Sample Criteria for Testing Outlying Observations. *Ann. Math. Stat.* 21, 27-58.
- Harter, H.L. (1974-76) - The Method of Least Squares and some alternatives. *Rev. Inst. Int. de Stat.* 42,147-74 (part I); 42,235-64 (part II); 43,1-44 (part III); 43,125-90 (part IV); 43,269-78 (part V); 44,113-59 (part VI).
- Kimber, A.C. (1990) - Exploratory Data Analysis for Possibly Censored Data from Skewed Distributions. *Applied Statistics*, 39, 21-30.
- Muñoz-Garcia, J., Moreno-Rebollo, J.L. e Pascual-Acosta, A. (1990) - Outliers: A Formal Approach. *Int. Stat. Rev.* 58, 215-26.
- Rosado, F.F. (1984) - Existência e Detecção de Outliers - Uma Abordagem Metodológica. Tese de Doutoramento. Universidade de Lisboa.
- Rosado, F.F. e Mendes, Z. (1993) - Sobre a Obtenção de Pontos Críticos para Testes de Discordância em Populações Normais. Nota 6/93 do Centro de Estatística e Aplicações da Universidade de Lisboa.
- Rosado, F.F. e Mendes, Z. (1994) - Sobre a Detecção de Observações Discordantes em Populações Normais. *Actas do I Congresso Anual da Sociedade Portuguesa de Estatística - A Estatística e o Futuro e o Futuro da Estatística* (p 271-286). Edições Salamandra.
- Rosado, F.F. e Alpiarça, I. (1994) - Outliers Múltiplos em Modelos de Discordância para Populações Exponenciais. Nota 7/94 do CEAUL.