Developments in Applied Statistics
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 19, Ljubljana: FDV, 2003

Exploratory Data Analysis as an Efficient Tool for Statistical Analysis: A Case Study From Analysis of Experiments

Katarina Košmelj[1], Andrej Blejec[2], and Drago Kompan[1]

Abstract

Goat breeders investigated the effects of three different additives on the number of somatic cells in goats' milk, additive DHA which is of fish origin, additive ALFA of plant origin, and additive EPA of fish origin. A control treatment contained no additive. The objective of this experiment was to answer two questions: Do any of these additives significantly reduce the number of SC in goats' milk? For treatments resulting in a reduction, how long does the effect persist?

Standard statistical methods used in the first phase did not give satisfactory results, therefore we analyzed the experiment in the context of exploratory data analysis. We used several graphical displays to establish which transformations have to be used in the preliminary phase and what kind of statistical methods should be applied to answer the questions of the experimenters. The results showed that ALFA treatment is the best, its effect was significant up to 54^{th} day of the experiment. Exploratory approach generated a new hypothesis: the procedure used for drug administration may increase the number of somatic cell due to the stress caused to the animal.

In this paper we tried to show that in the analysis of experiments the common approach mostly based on modeling and hypothesis testing can be expanded by an intense exploration of data.

1 Introduction

Goat-breeders carried out an experiment with three new treatments and a control. Two questions were put to the statistician: 'Which treatment is the best?' 'How long does the effect of the best treatment persist?' These are typical questions in

[1] Biotechnical Faculty, University of Ljubljana, Slovenia.
[2] National Institute of Biology, Slovenia.

the analysis of experiments where statistical methods such as ANOVA are commonly used.

In this case study, other statisticians undertook standard statistical analysis based on modelling, however the results were not satisfactory to the experimenters. When the experimenters addressed us for an alternative view, we decided to undertake an approach based on exploratory data analysis (Tukey, 1977).

In the next section we describe the experiment in detail, and in the third we present the pathway followed to obtain the answers. We started our analysis with graphical displays of the data. These plots were very informative and led us step-by-step to the next phase of the statistical analysis. In the last section some conclusions are drawn.

2 Description of the experiment

2.1 Objectives of the experiment

The quality of milk is determined on the basis of the content of microorganisms and somatic cells. The somatic cells considered here are leukocytes and epithelial cells from the udder of the animal. Increased numbers of leukocytes may be due to inflammation, to bacterial infection, or to other factors such as the status of lactation (during the lactation period the number of somatic cells increases), age of the animal, period of the year, udder lesion, quantity of milk and diet. Factors such as stress and anxiety increase the number of somatic cells as well. The number of somatic cells per ml (SC) in goats' milk is considerably higher than in cows' milk, and variations in content are also higher. Standards for the number of SC in goats' milk have not yet been set (Haenlein, 2002).

Researchers at the Zootechnical Department of the Biotechnical Faculty in Ljubljana, Slovenia suggested that some kind of food additive might decrease the number of SC. This led to the following experiment: the breeders investigated the effects of three different additives on the number of SC in goats' milk, additive DHA which is of fish origin, additive ALFA of plant origin, and additive EPA of fish origin. A control treatment contained no additive. The objective of this experiment was to answer two questions:

Question 1: Do any of these additives significantly reduce the number of SC in goats' milk?

Question 2: For treatments resulting in a reduction, how long does the effect persist?

Sixty goats were involved in the experiment. Animals were of the same age, same breed and at the beginning of lactation.

2.2 Time schedule of the experiment

The experiment lasted for 64 days, from Sept 17th, 2000 to November 19th, 2000 and was conducted in three phases.

a) *The preliminary phase* lasted for 10 days and was designed as the adaptation process. According to the experience of the experimenters it takes about a week for the animals to adapt to a new environment and to new personnel. All the animals were put in the same shed. On the 10th day the goats were divided into 4 groups: DHA, ALFA, EPA, CONTROL.
b) *The treatment phase,* during which additives were administered, lasted from the 11th to the 14th day. After the morning milking the animals were chased from one part of the shed to a specific restricted area where a technician put the additive directly in the muzzle of each animal. The control goats were also let through this area.
c) *The observation phase,* during which the animals were observed, lasted up to the 64th day.

2.3 Data acquisition

From each animal a sample of milk was taken at morning and evening milking. It was analyzed with respect to several milk parameters, including the number of SC in a standard milk sample. After the 19th day, laboratory analyses were done every fifth day. Some animals had to be eliminated from the statistical analysis due to the lack of data (as explained further). The data for 54 animals were taken into account for statistical analysis.

3 Statistical analysis

3.1 Data presentation and transformation

The data were first presented graphically. For each animal we plotted the number of SC as a function of time. Figure 1 displays the time series for four selected goats, each from a different treatment group.

Figure 1, and other plots not presented, reveal several interesting facts:

– there is great difference in the number of SC between animals (note that the scales on the y-axis are different);
– great oscillations of SC for each animal; the peaks are quite prominent;
– variability during the same day is visible: in the morning SC is generally higher than in the evening.

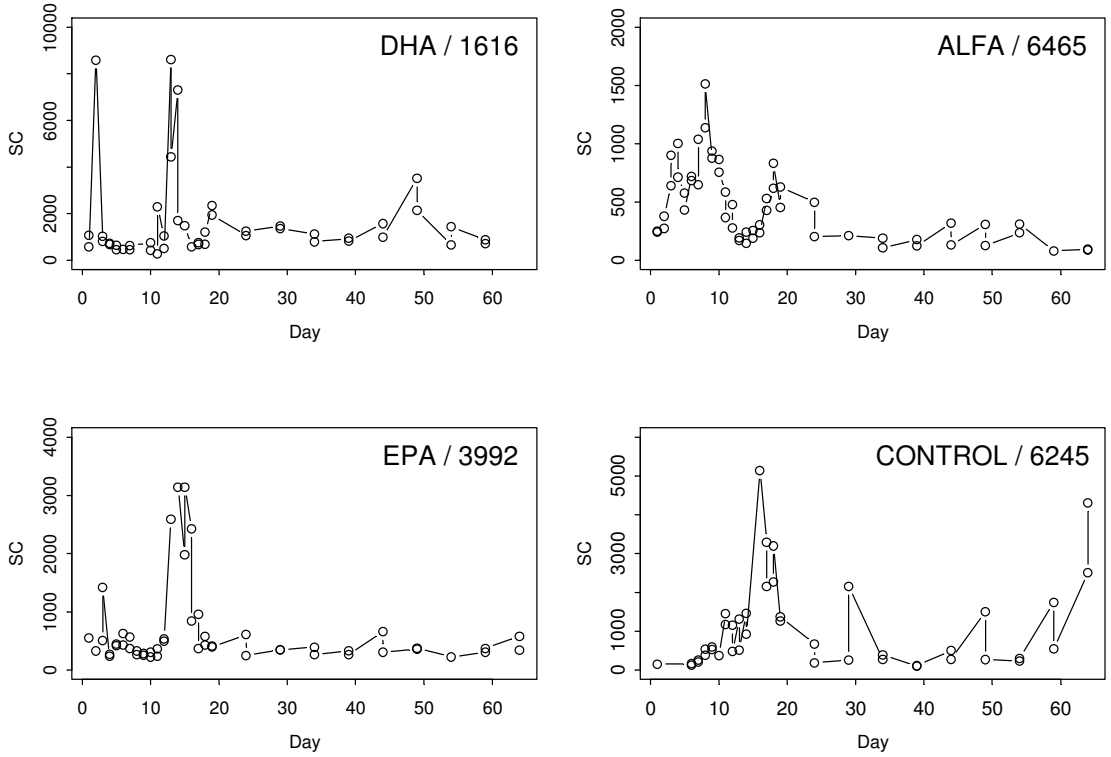

Figure 1: Number of somatic cells per ml (SC) for four goats, each from a different treatment group. The labels refer to the individual animals.

The first step in the analysis was to transform the data. The number of SC measures the concentration which is defined as a ratio of two quantities. Therefore the obvious transformation to start with was the logarithmic transformation. For ease of interpretation we used logarithm to base 10.

Figure 2 shows that the animals were not comparable in the preliminary phase, therefore the effect of different treatments can not be assessed directly. Some kind of standardization is necessary to obtain the comparability of the animals' data before the treatment phase. The adaptation phase was seen to end before the day 7, as the breeders expected. Visual inspection shows modest variability with no noise (outliers) in this period. Therefore the period from days 7 to 10 was considered as the representative period, i.e. when the animals were in their standard (inherent) state. The vertical lines on Figure 2 delineate this period.

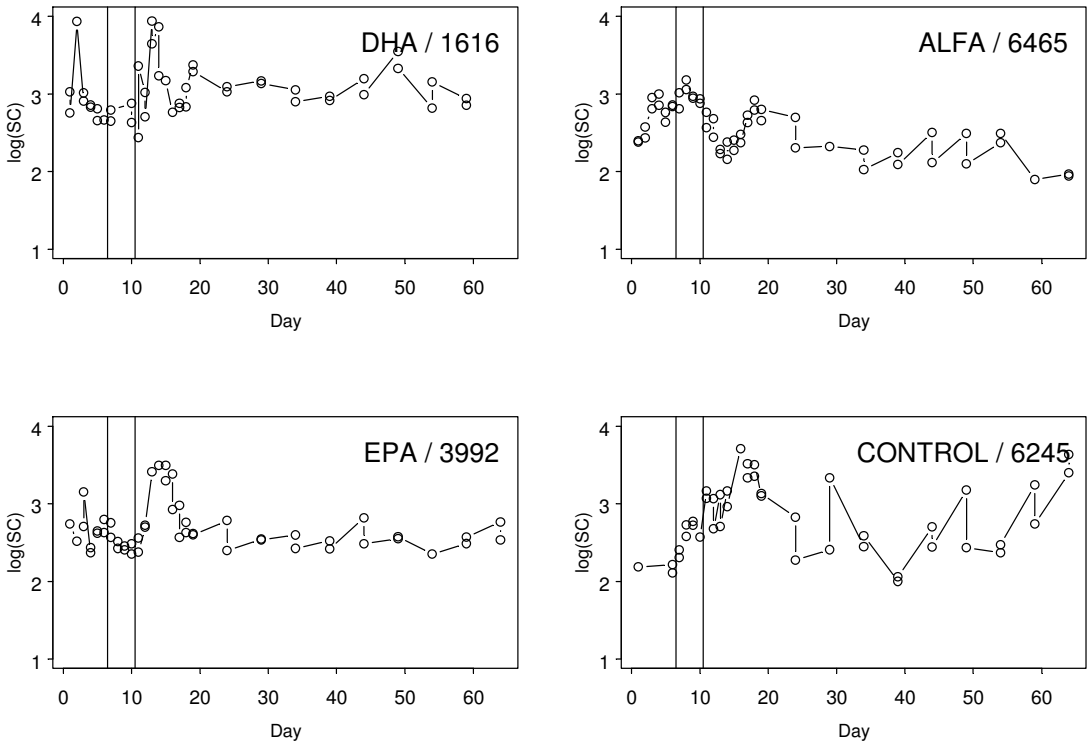

Figure 2: Logarithm of the number of somatic cells per ml (SC) for four goats, each from a different treatment group. The vertical lines delineate the representative period (see text) of the animals (day 7 to day 10).

The data is seen to be highly skewed, and there are several extreme outliers. Thus, the commonly used standardization procedure based on the mean and standard deviation was not a good choice for our data. We replaced the mean by the median (*Me*) and the standard deviation by the average absolute deviate from the median (*AAD*). The transformed time series $S(t)$ for each animal is calculated as follows:

$$S(t) = \frac{y(t) - Me_0}{AAD_0},$$

where Me_0 and AAD_0 are, respectively, the median and average absolute deviate for the representative period, and $y(t)$ is the logarithm of the original time series. For the representative period (i.e. for the period from day 7 to day 10) the transformed data is standardized: for each animal the median for the representative period (Me_0) is zero and the corresponding average absolute deviate (AAD_0) is one. Outside the representative period the transformed time series $S(t)$ is adjusted to the representative period.

It should be also noted that some animals had to be eliminated from the subsequent analysis due to the lack of data in the representative period. The final database for the statistical analysis was as follows: 13 animals for DHA, 12 animals for EPA and 12 for ALFA, and 17 animals for the control.

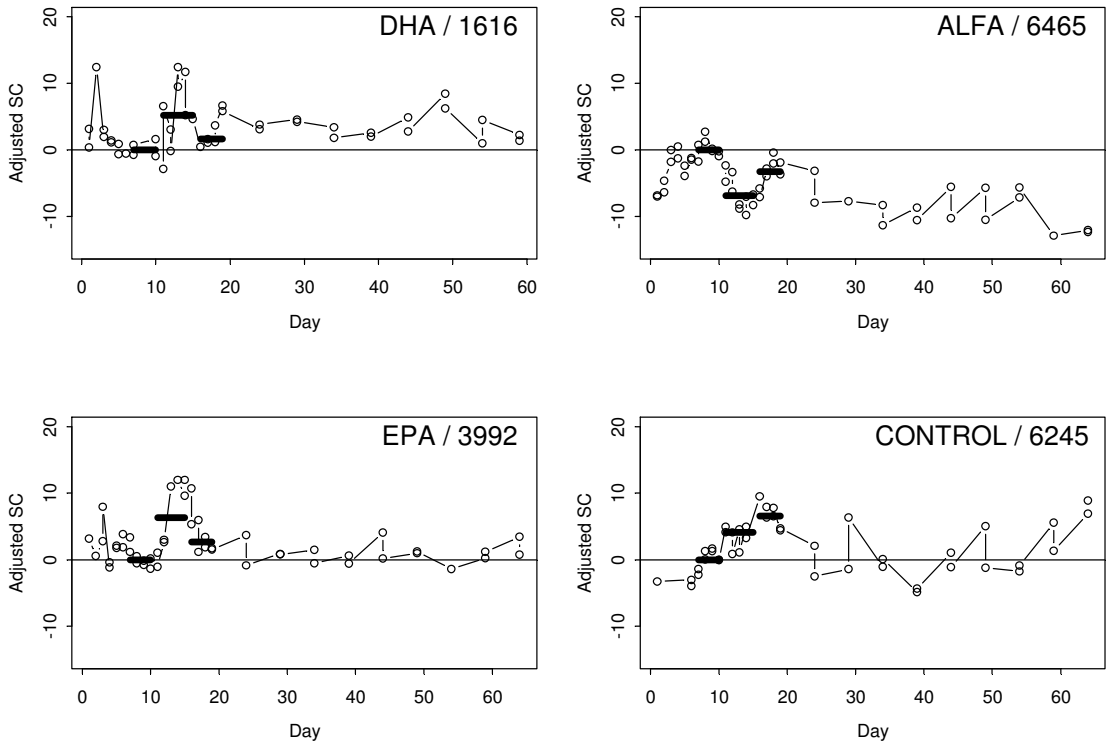

Figure 3: Adjusted time series $S(t)$ for four goats, each from a different treatment group. The thick horizontal lines indicate the medians for three periods: the representative period (days 7 to 10 – this value is always zero), the period when the additives were administered (days 11 to 14), and that from days 15 to 19.

We calculated two additional medians from the adjusted time series: the median for the time at which the additives were administered (11^{th} to 14^{th} day) and the median for the period from the 15^{th} to the 19^{th} day. These are denoted by Me_1 and Me_2, respectively and shown in Figure 3 by thick horizontal lines.

For the goat from the ALFA treatment group negative medians are evident (Figure 3). Plots for the animals from EPA and DHA are very similar, the two medians for the control goat are positive as well.

3.2 Question 1: Which treatment is best?

The distributions of the medians Me_1 and Me_2 obtained for all the animals are shown in Figure 4 as a series of box-plots.

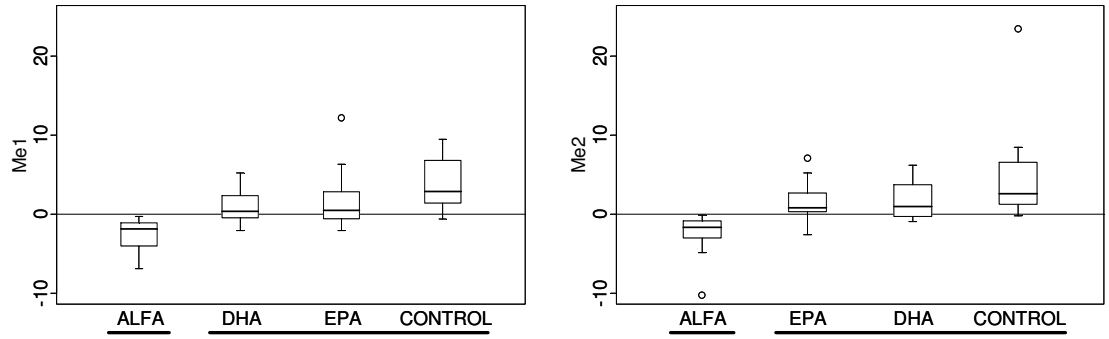

Figure 4: Box-plots for Me_1 (left) and for Me_2 (right) for each of four treatment groups. The lines under the treatment labels mark the groups obtained by multiple comparison test.

For the ALFA treatment, the values for Me_1 and for Me_2 are all negative, while for DHA and EPA treatments the majority of values are positive. In contrast, the majority of values for the control group are positive, some of them unexpectedly high (we would expect them to be around zero). When the distributions for Me_1 and Me_2 are compared, the same situation is identified, though the values for Me_2 are higher than for Me_1.

We used distribution-free tests (Siegel and Castellan, 1988; Hutchinson, 2000) to assess the effect of the treatments. Kruskal-Wallis one-way analysis of variance showed highly significant differences between the distributions of the medians for the different treatments. A multiple comparison test revealed two disjoint groups: the first group consists of the ALFA treatment, the second group treatments with EPA and DHA, and the Control. The result is identical for Me_1 and Me_2 (see Figure 4).

The answer to Question 1 is straightforward: ALFA treatment is shown to be the best – it significantly reduces the number of somatic cells in the goats' milk. Its effect is evident both at the time it is administered and afterwards. No significant reduction is effected by DHA or EPA treatment (both of fish origin).

3.2 Question 2: How long does the effect of ALFA last?

Again we started with a graphical presentation. For the ALFA group we plotted the values for the adjusted time series $S(t)$ for all animals (see Fig. 5 left). For each time point we calculated the median of these values and displayed the time series of the grand medians (solid line). The grand median is negative over the whole period, and significantly so up to the 54th day (Wilcoxon's test).

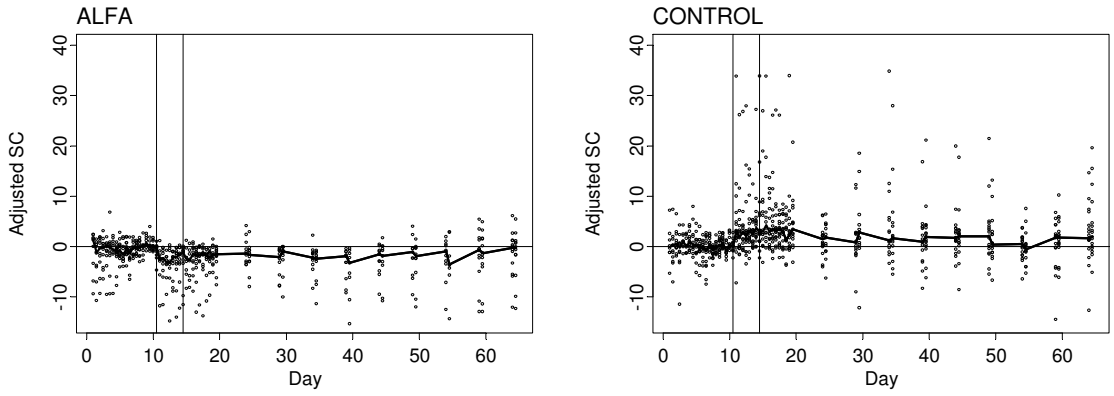

Figure 5: Values for the adjusted time series $S(t)$ for all animals at each time point. Left ALFA group, right CONTROL group. The solid line connects the grand medians at each time point. The vertical lines define the period during which the additives were administered.

3.3 Ancillary findings of the analysis

The results show that the control group is an essential part of the analysis. In the control group, the medians Me_1 and Me_2 are all positive rather than around zero, as expected. This is confirmed by the plot in Figure 5, where it is evident that, after the period of administration started, the grand median is positive up to the 50th day.

The breeders suspected that the procedure of chasing the animals from one side of the shed to the technician's area may be very stressful for the animals and may result in the increase of the number of somatic cells. If this is true, then the effect of the other additives might be underestimated.

We checked for the outliers in the control group. It turned out that the majority of extreme values (over 20 on Figure 5 right) belong to one particular animal CONTROL/6464. Figure 6 reveals its characteristics on the original scale and on the adjusted scale. For this goat, SC in the preliminary period was extremely low and stable, when the treatment phase started it increased to about 2000. The

extreme values in the adjusted time series are caused by the low variability in the representative period.

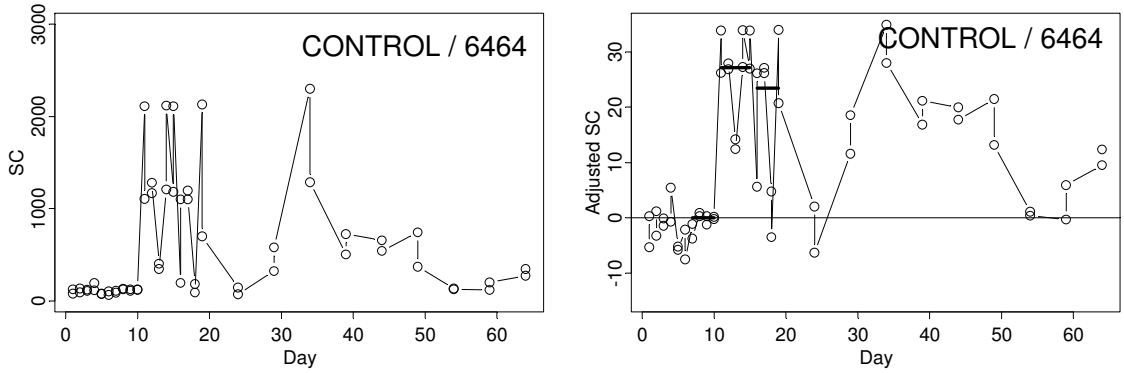

Figure 6: Time series for all animal CONTROL/6464: original time series left, adjusted time series right.

4 Discussion

The exploratory approach we followed is based on a step-by-step exploration of data. Graphics is its essential part. We started with graphical display of the time-series for each goat. The plots we made spoke for themselves – they revealed several interesting facts about the data under study and showed the way ahead: use of appropriate transformations and distribution-free tests.

The choice of transformations was based on the subject matter knowledge (e.g. logarithmic transformation is appropriate for the variables measuring concentration) and on the careful inspection of the plots. The latter revealed the need for standardization to obtain the comparability of the animals' data before the treatment phase. Subject matter knowledge was also necessary for this procedure: the experimenters defined the representative period for the animals. We have used two alternative measure of variation in the standardization procedure: the average absolute deviate AAD (as described in the text) and the median of the absolute deviates, *MAD* (not presented in the text), as proposed by (Huber, 1981). Both transformation produced the same rank order of values, therefore the results of the distribution-free tests were the same.

We considered the medians as the representative values for the specific time intervals. We are aware of the fact the variability due to subject is not taken into account in the Kruskal-Wallis test. We tried to compensate this deficiency by a careful inspection of the within group variability (see Figure 5). The inspection of the control goats' data generated a new research hypothesis: the procedure used for

drug administration may increase the number of somatic cells due to the stress caused to the animal.

To sum up: in this paper we tried to show that in the analysis of experiments the common approach mostly based on modeling and hypothesis testing can be expanded by an intense exploration of data. The case study from the field of animal breeding illustrates this. However intense collaboration among experimenters and statisticians is the prerequisite for that.

Acknowledgments

We express thanks to an anonymous referee whose comments helped us to improve the paper considerably.

References

[1] Haenlein, G.F.W. (2002): Relationship of somatic cells counts in goats milk to mastitis and productivity. *Small Ruminant Research*, **45**, 163-178.

[2] Huber, P.J. (1981): *Robust Statistics.* New York: John Wiley and Sons.

[3] Hutchinson, T.P. (2000): Letter to the editor: ANOVA with skewed data. *Environmetrics 2000,* **11**, 121-124.

[4] Siegel, S. and Castellan, N.J. (1988): *Nonparametric Statistics for the Behavioral Sciences.* New York: Mc-Graw-Hill.

[5] Tukey, J.N. (1977): *Exploratory Data Analysis*. Reading: Addison-Wesley.