BIOMETRICS 55, 1059–1063
December 1999

Analysis of Experimental Data with Repeated Measurements

Jenö Reiczigel

Department of Biomathematics and Informatics, University of Veterinary Science,
István u. 2., H-1078 Budapest, Hungary
e-mail: jreiczig@ns.univet.hu

SUMMARY. Experimental data often consist of serial measurements on subjects after a treatment. Typical questions concerning such data are: (A) Do subjects really react to treatment or are the fluctuations just random? (B) What are the numerical characteristics of the response? (C) Is the response identical in all groups? Differences between the individuals in the dynamics of the reaction make it difficult to apply standard statistical procedures. This paper proposes to answer questions (A) and (B) at the individual level, then to give an answer to (C) on the basis of this information. This kind of analysis may be useful since it can separate subjects giving response from those that do not and can identify individual response patterns and compare treatments with respect to each numerical characteristic separately. To answer question (A), a permutation test is proposed and its power is evaluated by simulation.

KEY WORDS: Reaction curve; Repeated measures; Response curve; Serial measurements.

1. Introduction

Experimental data in medical and veterinary science are often like those from an experiment by Varga (unpublished data), partially shown in Figure 1, which presents the body temperature profile of three calves during 10 days after a bovine rhinotracheitis infection. Data of a similar kind occur in pharmacokinetics studies or when various measurable reactions to a certain surgery or other intervention are studied. Their common feature is that several measurements are made on each individual at consecutive time points within a certain observation period. Such data are usually presented in the form of curves, which are sometimes called reaction curves. A couple of questions naturally arise concerning such reaction curves.

(A) Do the animals really have a response to treatment or are the fluctuations simply random? In such cases, when a definite type of reaction is expected, the question changes to "Do the animals react in the way they are expected to?" In the example of Figure 1, a temporary increase in body temperature is expected, but in other cases, curves of other shapes can be considered. For simplicity, the present paper focuses on the temporary increase type of reaction.

(B) If the answer to question (A) is positive, i.e., if the experimental animals do show the expected reaction to treatment, the next question that arises is about the numerical characteristics of their reaction. In the example of Figure 1, if the body temperature really increases first and then returns to the normal value, on which day does it reach its maximum value, how much is this maximum, and how much is the rise in body temperature compared to its baseline value?

(C) If the aim of the study is a comparison of two groups (e.g., a vaccinated group and a control group in the example of Figure 1) or two or more treatments, then the further question arises of whether or not the reaction is the same in the different groups. In other words, can the expected reaction be observed in each group, and if it can be, are its numerical characteristics the same or not? The individual differences between the experimental animals in the dynamics of the process (different length of latency, different intensity of reaction, etc.) make it difficult, however, to use the standard statistical procedures.

Clinicians are often interested in evaluating the reaction of a single individual to a certain treatment, so in some practical situations, questions (A) and (B) refer to one single individual rather than to a sample of individuals. In such cases, evaluation involves the comparison of observed data series to a reference profile that is derived theoretically or is based on former experience.

Section 2 summarizes the conventional analysis methods for this kind of data. Section 3 describes the proposed method in detail. Section 4 presents an application to real data. Section 5 gives an evaluation of the proposed analysis.

2. Usual Ways of Analysis

If there is an acceptable parametric model for the particular process, it is quite natural to make all kinds of inferences on the basis of that model. For the temporary increase type of reaction, usually polynomials, beta or gamma functions, or the difference between two exponential functions are considered. Also, the numerical characteristics of the response are calculated from the parameters of the curve. Though this is the most reasonable thing one can do, in some unlucky cases, in particular with polynomials, these estimates may be worse than nonparametric ones even when the model holds.

With a lack of a parametric model, question (A) is usually answered by applying a repeated measures ANOVA over

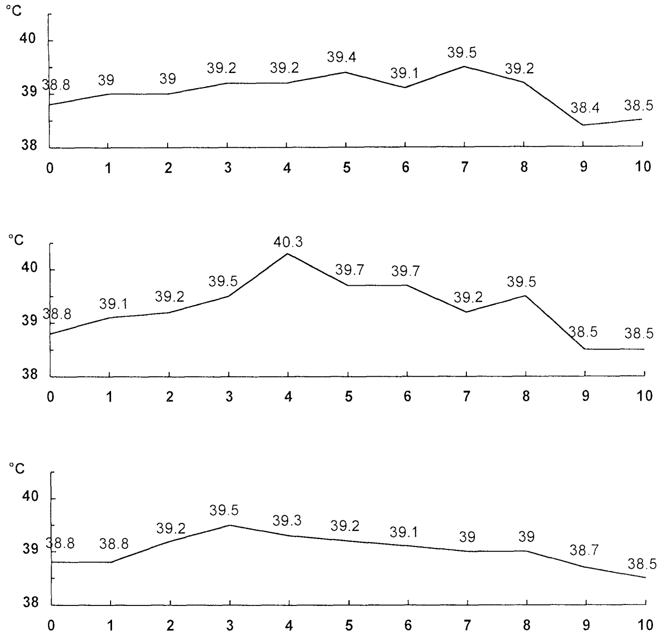

Figure 1. Daily records of body temperature of three calves during 10 days after a rhinotracheitis infection.

time or by selecting an appropriate time point or interval and comparing its values to the initial (pretreatment) values using a paired t-test or one of its nonparametric equivalents. If this latter approach is planned, it is useful to make more than one measurement on each individual before treatment. The individual differences in the dynamics of the reaction, however, cannot be controlled in such a way; they may bring considerable variance into measurements at any single time point, which can lead to an increase of type II error (i.e., loss of power). On the other hand, if the time period is selected optimally, i.e., a time period is chosen in which actual data show the largest deviation from the initial values, the type I error will increase.

The usual way of answering question (B) without a parametric model is to consider individual observed values (in the case of a single subject) or their sample means, e.g., for a temporary increase, the time of the maximum is usually estimated by that time point at which the measured value (for a single subject) or the sample mean of the measurements is maximal. There are several objections to considering sample means at each time point (Matthews et al., 1990). First, the main interest lies in how subjects respond over time and the curve consisting of the mean values does not usually represent a typical individual response. Second, due to differences between subjects, the standard error of such an estimate may increase considerably. Additionally, calculating with mean values makes the construction of confidence intervals quite problematic. Therefore, it seems to be better to calculate using the individual numerical characteristics (the individual maxima in the above example).

The same methods are applied (and similar problems arise) in answering question (C) in the two- or multiple-sample setting. When there are more than two groups, multiple range tests can also be used.

ANOVA has often been criticized for its restrictive assumptions (e.g., normality of errors) and for being unable to exploit the information in the ordering of the time points (Raz, 1989), and alternative models and analysis methods have been proposed.

Fischer and Csáki (1962) proposed an analysis, based on the minimum and the maximum values of the individual curves, in which a couple of hypotheses could be tested, but they did not deal with the estimation of the numerical characteristics of the curves.

Matthews et al. (1990) proposed a general framework for analyzing this kind of data by calculating first some summary measures from the individual curves, which quantify important aspects of the individual responses, and then analyzing these summary measures as if they were raw data. The present approach is fully in accordance with that framework. Dawson and Lagakos (1993) studied the performance of some distribution-free tests based on summary measures in case of missing data. Their paper contains further references to other approaches to the analysis of repeated measurements.

3. The Proposed Method of Analysis

The main principle of the proposed method is to analyze the individual curves separately to find answers to questions (A) and (B) at the individual level and then to answer (C) on the basis of this information. In the following, a procedure is given for the analysis of the individual curves. Though this analysis is described here only for the temporary increase type of reaction, its adaptation to other kinds of curves is straightforward. The basic idea comes from smoothing using a moving average. Working with averages of m consecutive measurements instead of the single measurements themselves may reduce the standard error of the estimates. For a temporary increase, usual parameters to be estimated are the baseline value, the maximum value, and the time of maximum. The procedure for their estimation is as follows.

Suppose the individual has a data series of length k. Find those m consecutive data points with d points deleted next to them (d_1 from the left and d_2 from the right, where $d_1 + d_2 = d$) for which the difference between the averages of those m and of the remaining $k - m - d$ measurements is maximal. Those m points are hereafter referred to as the top period, those d_1 and d_2 points as the transient periods, and those $k - m - d$ points as the baseline period. The time of the maximum is estimated by the median time calculated from the m points of the top period. The maximal and baseline values are estimated by the average of the m values in the top period and that of the $k - m - d$ values of the baseline period, respectively.

It is important to note that m and d are fixed parameters of the procedure that must be established before the observation of data (d_1 and d_2, however, are optimally chosen with respect to the maximization, provided that $d_1 + d_2 = d$). The best choice of m and d depends on the shape of the curve and on the frequency of sampling. Some experiences suggest that choosing $m = 2$, 3, or 4 with an even d (to allow for symmetric transient periods) performs well, provided that the baseline period has a sufficient length ($\geq m$).

Both the maximum and the baseline values are estimated with a bias (the maximum is underestimated, the baseline is overestimated) due to the nature of the procedure. The degree

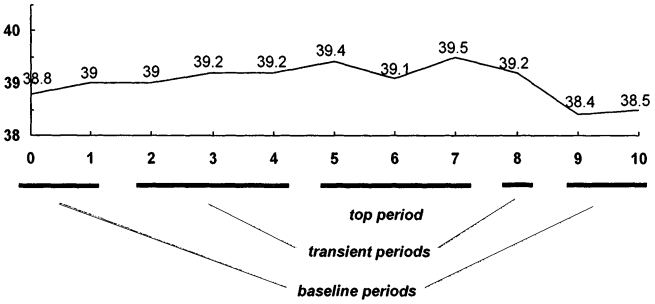

Figure 2. The top, baseline, and transient periods for the first data series of Figure 1 (for $m = 3$ and $d = 4$).

of the bias depends on the frequency of sampling compared to the shape of the curve and cannot be expressed easily.

Since the analysis operates with blocks of consecutive measurements, irrespective of the actual times, it works without any modification even with irregularly timed measurements and/or missing values (but irregular timing or missing values may reduce the accuracy).

Figure 2 shows these periods determined for the first data series of Figure 1, calculated with $m = 3$ and $d = 4$. The estimated time point of the maximum is day 6. The calculated maximal value of body temperature is 39.3°C, the baseline value is 38.7°C, and the increase is 0.66°C.

To answer question (C), the individual numerical characteristics (estimated by the above procedure) in the different groups can be compared using standard statistical methods (e.g., t-test or Mann–Whitney test for the two-sample situation or ANOVA or Kruskal–Wallis test for more groups). The individual numerical characteristics can also be combined into common values or used for construction of confidence intervals (but for the construction of a confidence interval, a larger sample is needed).

It is often the case that the null profile representing no response is assumed, on the basis of former experience, to be constant with i.i.d. random errors added. In such cases, question (A) about an individual can be answered by applying a permutation test on the individual data series to test $\mathrm{H}_0^{(\mathrm{indiv})}$: there is no change over time (values are constant with i.i.d. random error) versus $\mathrm{H}_1^{(\mathrm{indiv})}$: values show an increase followed by a return to the initial value.

As a test statistic, the difference between the average values of the top and baseline periods can be used. The idea behind the test is that, under $\mathrm{H}_0^{(\mathrm{indiv})}$, permuting the points does not affect the distribution of the test statistic, so its null distribution can be determined by considering all possible permutations (or can be approximated by Monte Carlo simulation, taking a large random sample from the set of all permutations). Significance can be assessed by comparing the actual value of the test statistic to that null distribution. (Actually, permutation tests require a little less than i.i.d. errors [see Good, 1994, Chapters 2 and 14].)

A rank version of the test can also be used if the distribution of the measured values makes it necessary. In this version, the top and baseline periods can be defined so that the rank sum of the top period after deletion of the transient periods is maximal. As the test statistic, this rank sum can be used. Given there are no ties, exact significance levels can be calculated relatively simply using combinatorics. With ties, however, calculations become more complicated. Of course, both numerical evaluation of all permutations and Monte Carlo simulation work in the case of the rank version as well.

For the first data series of Figure 1, simulation was used for assessing significance. Three hundred random permutations of the data were generated by a computer program. For each permutation, the top, transient, and baseline periods were determined according to the above procedure and the test statistic was calculated. (The program for IBM PC compatible computers is available from the author.) Among these 300 simulated values of the test statistic, which represent the null distribution, there were only six values greater than the actual value of the test statistic, so H_0 could be rejected at $p = 0.02$, i.e., this particular animal had a statistically significant increase in body temperature. In the same way, significance can be assessed for each individual data series.

The p values obtained from these individual tests can serve as a basis of an overall test with $\mathrm{H}_0^{(\mathrm{overall})}$: no individual in the population shows a reaction (for each individual in the population, $\mathrm{H}_0^{(\mathrm{indiv})}$ holds, i.e., values are constant with i.i.d. random error) versus $\mathrm{H}_1^{(\mathrm{overall})}$: some (maybe all) individuals in the population show a temporary increase type of reaction.

The test is enabled by the fact that, under $\mathrm{H}_0^{(\mathrm{overall})}$ (i.e., when $\mathrm{H}_0^{(\mathrm{indiv})}$ holds for all individuals in the population), p values obtained from the individual tests follow a uniform distribution in $[0, 1]$, while under $\mathrm{H}_1^{(\mathrm{overall})}$, they are more likely to be near zero. In the meta-analysis context, several methods are known to combine p values from independent experiments (see Hedges and Olkin, 1985), which can be applied to the present situation as well. In Edgington's (1972) method, the decision is based on the sum of the p values ($\Sigma\, p_i$). According to Edgington, it has good power properties, is proven to be exact, and gives the overall p value immediately, without the need to use probability tables, applying the formula

$$p^{(\mathrm{overall})} = \frac{\left(\sum p_i\right)^n}{n!} - \binom{n}{1}\frac{\left(\sum p_i - 1\right)^n}{n!} + \binom{n}{2}\frac{\left(\sum p_i - 2\right)^n}{n!} - \binom{n}{3}\frac{\left(\sum p_i - 3\right)^n}{n!} + \cdots,$$

where n is the number of individual p values and summation goes as long as the number subtracted from $\Sigma\, p_i$ is less than $\Sigma\, p_i$. The formula can be derived from the distribution of the sum of n independent observations from the uniform distribution in $[0, 1]$.

If a greater number of individuals is involved in the experiment, even the population proportion of those showing the reaction can be estimated from the sample on the basis of the individual tests.

4. Example

In Sterczer, Vörös, and Karsai (1996), the effect of cholagogues on changes in gallbladder volume (GBV) in dogs was studied by two-dimensional ultrasonography. Three different kinds of cholagogues and tap water (administered orally) as control were used in the experiment. Six healthy dogs were

Table 1
GBV data (cm^3) *by Sterczer et al.* (1996)

		Minutes after treatment												
Treatment	Dog	0	10	20	30	40	50	60	70	80	90	100	110	120
Cholechystokynin	1	17.70	10.35	10.78	11.44	11.20	12.38	12.68	12.30	14.00	14.64	14.96	14.18	16.78
	2	17.22	11.30	11.30	13.28	14.08	13.98	14.74	15.63	17.60	17.34	17.38	17.36	17.64
	3	14.24	9.20	9.40	9.62	10.10	10.08	9.60	9.70	11.23	11.20	11.96	12.20	13.98
	4	39.58	26.88	26.20	29.80	31.50	32.75	34.45	35.64	36.62	38.65	38.56	39.20	39.36
	5	13.33	7.15	7.82	7.94	8.40	8.94	9.28	9.95	10.40	10.95	11.70	12.10	12.35
	6	16.16	8.36	9.53	9.80	9.64	9.84	10.70	11.26	12.12	12.60	13.98	14.52	14.78
Clanobutin	1	16.35	13.65	13.10	13.58	14.03	15.45	15.58	15.56	15.62	16.10	16.28	16.74	16.25
	2	15.65	13.08	12.35	12.76	13.78	13.76	13.54	14.18	14.40	15.16	15.20	15.18	13.40
	3	12.68	9.68	10.70	10.98	11.12	11.78	12.02	11.95	12.16	12.25	12.40	12.55	12.54
	4	21.88	15.92	15.18	17.04	18.60	18.98	19.26	20.38	21.32	21.03	21.80	21.08	22.65
	5	12.78	9.03	9.28	9.54	9.38	9.88	9.94	10.14	10.34	11.50	11.83	11.78	12.08
	6	15.58	11.50	11.88	12.06	12.58	12.98	13.00	13.00	13.04	13.18	13.88	13.43	13.85
Control	1	20.75	19.83	19.98	18.84	19.10	19.50	19.75	19.64	20.00	19.13	20.15	19.45	19.43
	2	13.88	13.60	13.73	13.16	13.44	13.62	13.86	13.58	14.28	14.10	13.12	13.53	13.42
	3	11.92	11.74	11.84	10.90	11.75	11.45	11.98	12.38	11.70	11.48	11.80	11.20	12.03
	4	26.38	26.90	27.73	27.73	27.56	28.43	27.54	26.50	27.94	27.58	27.56	27.64	28.83
	5	13.30	13.18	13.52	13.43	13.40	13.25	13.28	13.24	13.44	12.98	12.60	13.48	13.08
	6	13.80	13.86	13.06	13.76	13.82	13.80	13.86	13.84	13.76	13.82	13.50	13.72	13.70

treated with each substance. GBV was determined immediately before the administration of the test substance and at 10-minute intervals for 120 minutes thereafter.

In the following, the proposed analysis is applied to groups treated with cholechystokynin and clanobutin and to the control group (Table 1).

In this case, though the type of expected reaction is not a temporary increase but a temporary decrease, data can be analyzed using the same procedure after a sign change. Results of an analysis with $m = 2$ and $d = 4$ are as follows. In the control group, no individuals showed a significant reaction. The individual p values (each one determined from 300 Monte Carlo replications) were 0.21, 0.69, 0.77, 0.34, 0.08, 0.88, which result in an overall value $p = 0.49$ according to the calculations above.

In the group treated with cholechystokynin, all individual p values were less than 0.01, which result in an overall value $p < 10^{-14}$. The estimated numerical characteristics were the following: the period of the greatest decrease was from 10 to 20 minutes after treatment for all dogs and the decreases in GBV were 4.34 (29%), 5.86 (34%), 2.77 (23%), 11.70 (31%), 4.06 (35%), and 4.69 (34%). As the magnitude of the decrease is correlated with the baseline value, the percentage is more informative.

The effect of clanobutin was also significant (again $p < 0.01$ for all dogs), and the time period of the greatest decrease was quite similar (from 10 to 20 minutes after treatment for four dogs and from 20 to 30 minutes after treatment for two dogs). The decreases in GBV, namely 2.79 (17%), 2.16 (15%), 2.19 (18%), 5.91 (28%), 2.34 (20%), and 2.01 (15%), were, however, significantly less than those caused by cholechystokynin ($p < 0.05$ by paired t-test as well as by Wilcoxon's rank test).

In Sterczer et al. (1996), decreases were expressed as percentages of the pretreatment (time point 0) values, then the average decreases of the groups were compared in each time point. Though the main conclusions of the two kinds of analysis were the same, the present method could provide more information due to having taken into account individual behavior, namely, it pointed out that all individuals in the treated groups but no individuals in the control group showed a significant reaction, it demonstrated the very concentrated distribution of the time of the minimum in the treated groups, and it could compare the groups with respect to the time of the minimum and to the magnitude of the decrease separately.

5. Discussion

The proposed analysis seems to be more informative than usual procedures as (1) it can handle different individual response dynamics, (2) it can separate those individuals that give a response from those that do not, (3) it can identify individual response patterns, (4) it can give account of the distribution of the numerical characteristics in the different groups, and (5) it can compare treatments with respect to each numerical characteristic separately.

The power of the permutation test was studied by simulation with respect to a special H_1 in which the response profiles were the following gamma type functions:

$$f_1(t) = 0.15t^3e^{-t} + 1,$$

$$f_2(t) = 0.043t^4e^{-t} + 1,$$

and

$$f_3(t) = 0.01t^5e^{-t} + 1,$$

representing three different types of response dynamics. Data series were generated using 10 values (for $t = 0, 1, 2, \ldots, 9$) of one of the functions with i.i.d. random error added (normal with $\mu = 0$ and $\sigma = 0.1$ or 0.15). The effect of missing data on the power was also analyzed using the same model. Missingness was simulated as independent of the response (one or two values per individual were randomly deleted). Simulation results show that the proposed permutation test is powerful even for small samples (n = 3 and 5) and is not very sensitive to missing data (in case of noninformative missingness). Compared to an analysis by repeated measures ANOVA, the permutation test was found to be superior even when there were no individual differences in the dynamics of the process (i.e., when the same f_k was used for all individuals). In the case of such differences, its superiority was even more pronounced: It had the same power with a single individual as ANOVA with a sample of three and the same power with a sample of three as ANOVA with a sample of five. So, in such cases when former experience has proven that the null profile (that without any reaction) is constant, the proposed test is a powerful way to test either $\mathrm{H}_0^{(\mathrm{indiv})}$ or $\mathrm{H}_0^{(\mathrm{overall})}$ given above. Of course, when the null profile is not known, testing of a treatment effect must involve a comparison between a treated and a control group.

Acknowledgements

The author is grateful to János Varga and Ágnes Sterczer for making the data sets available and to Rolf Holle for his constructive criticism of an earlier version of the manuscript.

Résumé

Les données expérimentales consistent souvent en une série de mesures sur des sujets après l'application d'un traitement. Les habituelles questions sur de telles données sont: A) Observons-nous vraiment les réactions des sujets aux traitements ou observons-nous uniquement des fluctuations aléatoires. B) Quelles sont les caractéristiques numériques de la réponse? C) La réponse est-elle identique dans tous les groupes? La différence entre les individus dans la dynamique de la réaction rend difficile l'application de procédures statistiques standards. Cet article propose de répondre aux questions (A) et (B) au niveau de l'individu, puis de répondre à la question (C) sur la base de cette information. Ce type d'analyse peut être très utile pour distinguer les sujets qui répondent de ceux qui ne répondent pas, puis pour identifier les modèles de réponses individuelles et comparer les traitements en accord avec chaque caractéristique numérique. Pour répondre à la question (A), un test de permutation et sa puissance est évaluée par simulation.

References

Dawson, J. D. and Lagakos, S. W. (1993). Size and power of two-sample tests of repeated measures data. *Biometrics* **49,** 1022–1032.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *Journal of Psychology* **80,** 351–363.

Fischer, J. and Csáki, P. (1962). Analysis of reaction curves by extreme values. *Acta Medica Hungarica* **15,** 363–370.

Good, P. (1994). *Permutation Tests.* New York: Springer-Verlag.

Hedges, L. V. and Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic.

Matthews, J. N. S., Altman, D. G., Campbell, M. J., and Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal* **300,** 230–235.

Raz, J. (1989). Analysis of repeated measurements using nonparametric smoothers and randomization tests. *Biometrics* **45,** 851–871.

Sterczer, A., Vörös, K., and Karsai, F. (1996). Effect of cholagogues on the volume of the gallbladder of dogs. *Research in Veterinary Science* **60,** 44–47.

Received November 1997. *Revised December* 1998.
Accepted January 1999.