Biometrical Journal **41** (1999) 5, 517–531

Confidence Limits for the Onset and Duration of Treatment Effect

ROGER L. BERGER and DENNIS D. BOOS

Department of Statistics
North Carolina State University
Raleigh
USA

Summary

Studies of biological variables such as those based on blood chemistry often have measurements taken over time at closely spaced intervals for groups of individuals. Natural scientific questions may then relate to the first time that the underlying population curve crosses a threshold (onset) and to how long it stays above the threshold (duration). In this paper we give general confidence regions for these population quantities. The regions are based on the intersection-union principle and may be applied to totally nonparametric, semiparametric, or fully parametric models where level-α tests exist pointwise at each time point. A key advantage of the approach is that no modeling of the correlation over time is required.

Key words: Concentration curve; Crossover; Intersection-union; Population function; Regression function; Survival curve; Threshold.

1. Introduction

Early studies with a new active compound (drug) often evaluate blood concentration levels every hour or half hour or even continuously for up to a day. There may be an absolute level which characterizes adequate concentration or comparisons may be made with a placebo. For example a crossover design might be used so that n patients receive both placebo and active compound and are measured at k time points on separate days with an adequate washout period between the days.

Figure 1 displays results from a crossover trial with $n = 36$ subjects and $k = 15$ time points where a concentration of 100 is assumed to be an adequate level. There were actually two active compounds, but here we are displaying only the sample means for the concentration of one of the compounds. Notice that the sample mean is above 100 for values $t = 0.7$ through $t = 8.0$. In this example it also turns out that a t-test is highly significant at each time point in the interval [0.7,8.0]. What kind of statistical statement can be made about this time interval?

A naive approach suggests that we just take the time points where the t-test is significant and declare that we are confident that on the resulting region the con-

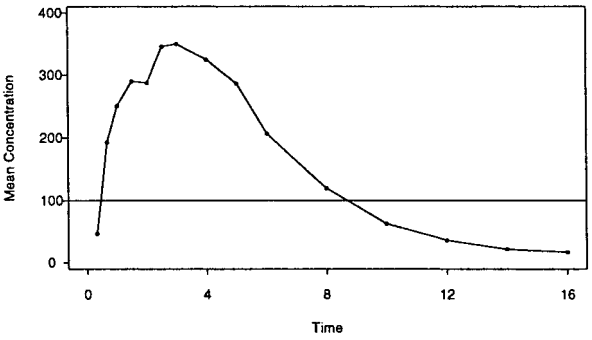


Fig. 1. Mean concentration versus time

centration is above 100. But what is the confidence level? Do we need to worry about multiple testing or the fact that each individual's results might be highly correlated?

Onset is defined in this example as the first time point when the population mean is greater than 100. *Duration* is the amount of time from onset until the population mean is no longer greater than 100. (In Section 2 we will make these definitions more rigorous.) This paper concerns confidence statements about the onset and duration of a treatment effect.

Our second example is from a crossover study where an active compound for stomach acid suppression was compared to a placebo. In Figure 2 we have plotted the difference in means for the $n = 75$ subjects at $k = 34$ time points. Time *zero* is when dosing occurred so that the values at $t = -1.5, -1.0, -0.5$, and 0.0 are baseline values. The active compound is effective for acid suppression when the population mean difference is greater than zero. *Onset* is defined here as the first time point when the population mean difference is greater than zero. *Duration* is the amount of time from onset until the population difference is no longer greater than zero. A paired t-test was highly significant on the interval $[1.0, 13.0]$ suggesting an onset of at least $t = 1.0$ and a duration of at least $13.0 - 1.0 = 12.0$ hours.

In these first two examples the treatment effect was defined as *mean* $-$ 100 and the *mean difference*, respectively. In other situations one might want to use medians or hazard rates or proportions or differences of these. The results of this paper

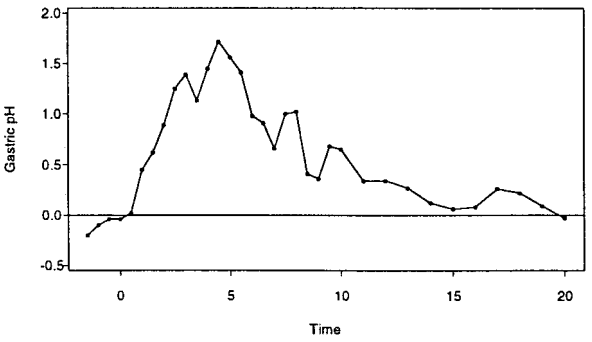
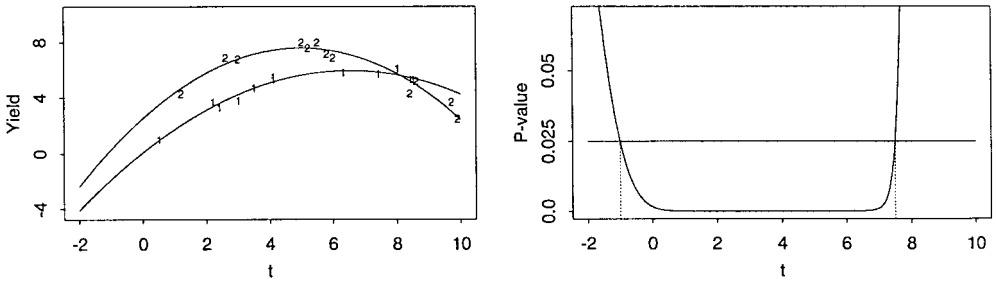


Fig. 2. Mean difference (active-placebo) in Gastric pH



(a) Data and Regression Lines (b) P-value as a Function of t

Fig. 3. (a) Yield versus covariate t for quadratic fits (1 = control, 2 = additive). (b) The interval on which the p-value is less than 0.025

apply to any setting where a treatment effect can be tested at multiple time points. Along these lines we have studied differences in mean residual life (BERGER, BOOS, and GUESS, 1988) and survival curves (DINSE, PIEGORSCH, and BOOS, 1993).

Our third example will motivate a somewhat different application where the time points are no longer discrete (or even time points!). The data are from Table 3.5.1 of MILLIKEN (1992) who reports the results from a completely randomized experiment with three treatments (a control and two additives) and a covariate t. Here we just use the control and first additive treatment group. Figure 3a shows separate quadratic fits to the control and to the first additive treatment group.

In this example we might be interested in knowing over what range of the covariate the mean yield with the first additive is greater than the mean yield of the control. This problem is quite different from Examples 1 and 2 because here we use a parametric quadratic regression model, and we can actually carry out a test for each point in the continuous set $(-\infty, +\infty)$ (assuming of course that the model holds over that range). Because the first two examples do not use parametric models for the curves, tests there can only be carried out at time points where there are actually data. Of course in this third example we might not use *onset* and *duration* terminology to describe the population quantities of interest.

The general inferential situation is as follows. Let $g(t)$ denote an unknown population function of interest. The function $g(t)$ can be a regression function, survival curve, concentration curve, or any other population function. In many situations we wish to compare two populations, and $g(t)$ is the difference between two regression functions, survival curves, concentration curves, etc. Let δ be a numeric value fixed by the experimenter. We wish to make a confidence statement of the form "$g(t) > \delta$ for all t in the interval $L \leq t \leq U$." Here, L and U are statistics calculated from the data, and the interval $[L, U]$ upon which we declare $g(t) > \delta$ is a random interval. Associated with this confidence statement, we wish to have a guaranteed confidence level. That is, we wish to ensure that, for a specified confidence level, $1 - \alpha$,

$$P(g(t) > \delta \text{ for all } t \text{ in the interval } L \leq t \leq U) \geq 1 - \alpha$$

regardless of the true value of the population parameters. Equivalently, we wish the error probability to be bounded above, that is,

$$P(g(t) \leq \delta \text{ for any } t \text{ in the interval } L \leq t \leq U) \leq \alpha.$$

We will describe a simple procedure that provides this kind of confidence statement.

Usually, $\delta = 0$ if $g(t)$ is the difference in two population functions. That is, if $g(t) = g_1(t) - g_2(t)$ where g_1 and g_2 are the population functions from populations 1 and 2, respectively, then the confidence statement is "$g_1(t) > g_2(t)$ for all t in the interval $L \leq t \leq U$". On the other hand, if $g(t)$ is a single population function, then δ represents a threshold value of interest. For example, in Figure 1, concentrations greater than $\delta = 100$ are of interest. The inferences in this method are one-sided, "$g_1(t) > g_2(t)$". In some cases, an inference of the form "$g_1(t) \neq g_2(t)$ for all t in the interval $L \leq t \leq U$" might be more appropriate. This paper does not address this type of inference.

We will give two variations of this confidence statement corresponding to different modeling scenarios. In Section 2 we will explain the method for situations without parametric models for the population function $g(t)$ (as in Examples 1 and 2). In Section 3 we discuss the method for situations with parametric models for continuous $g(t)$ as in Example 3. Section 4 makes comparisons with other approaches and discusses details of the method. Section 5 reports on a small Monte Carlo experiment. The Appendix contains proofs of the results in Sections 2 and 3.

2. Multiple Observations at Discrete Values

2.1 *Confidence procedure*

Suppose that for each of k discrete values of the independent variable, $t_1 < \cdots < t_k$, there exists a level-$\alpha/2$ test of $H_{0i} : g(t_i) \leq \delta$ versus $H_{ai} : g(t_i) > \delta$. Call this test ϕ_i. This situation frequently arises when, at each value t_i, multiple observations are obtained and a test about $g(t_i)$ can be conducted based only on the observations at t_i, not using the observations at other values of t. The data in Figure 1 are like this. At each of the $k = 15$ time points, there are between 28 and 36 observations. A test about the concentration at a particular time point can be conducted based only on the measurements at that time point.

For this scenario, we do not assume a specific parametric form for $g(t)$ but rather assume only the following.

Assumption 1. If, for any two consecutive t_i's, $g(t_i) > \delta$ and $g(t_{i+1}) > \delta$, then $g(t) > \delta$ for all t in $t_i \leq t \leq t_{i+1}$.

Practically speaking, Assumption 1 says that the t_i's are closely enough spaced so that, if $g(t)$ ever dips below δ, then there will be a t_i at that place with $g(t_i) \leq \delta$. So long as this is true, $g(t)$ can oscillate above and below δ.

For this scenario, the confidence statement is defined as follows. Let t_{ap} be a fixed *a priori* starting value satisfying $t_1 < t_{ap} < t_k$, where t_{ap} is specified by the experimenter prior to the data analysis and cannot be chosen based on the data. Let

$$i* = \max \{i : t_i \leq t_{ap} \text{ and } \phi_i \text{ accepts } H_{0i}\}$$

and

$$i^* = \min \{i : t_i \geq t_{ap} \text{ and } \phi_i \text{ accepts } H_{0i}\}.$$

Testing sequentially downward from t_{ap}, H_{0i*} is the first hypothesis that is accepted. If H_{0i} is rejected for all i with $t_i \leq t_{ap}$, $i*$ is defined to be 0. Similarly, testing sequentially upward from t_{ap}, H_{0i^*} is the first hypothesis that is accepted. If H_{0i} is rejected for all i with $t_i \geq t_{ap}$, i^* is defined to be $k+1$. Then, let $L = t_{i*+1}$ and $U = t_{i^*-1}$. If $L > U$, no confidence statement can be made. But, if $L \leq U$, the confidence statement "$g(t) > \delta$ for all t in the interval $L \leq t \leq U$" is made. This confidence procedure controls the error rate at α in that, for any population satisfying Assumption 1,

$$P(L \leq U \text{ and } g(t) \leq \delta \text{ for any } t \text{ in the interval } L \leq t \leq U) \leq \alpha. \qquad (1)$$

This error rate is verified in the Appendix.

There are some properties to note about this procedure.

1. The tests, $\phi_1, \ldots, \phi_k$ are one-sided, level-$\alpha/2$ tests. Level-$\alpha/2$ tests are used because the total error probability, α, is divided between $\alpha/2$ probability that L is too small and $\alpha/2$ probability that U is too large.
2. The tests, $\phi_1, \ldots, \phi_k$ may be correlated. The error rate (1) is still valid. In Figure 1, observations on the same individual are taken at the different time points. This would, typically, induce a correlation in the tests at the different time points. This correlation does not need to be specifically modeled to control the error rate at α.
3. The starting point, t_{ap}, is usually in the interval $[L, U]$. If t_{ap} equals one of the values $t_1, \ldots, t_k$, then it will always be the case that $L \leq t_{ap} \leq U$, when $L \leq U$. If t_{ap} is not one of the values $t_1, \ldots, t_k$, say $t_i < t_{ap} < t_{i+1}$, then it can be the case that $L \leq U = t_i < t_{ap} < t_{i+1}$ or $t_i < t_{ap} < t_{i+1} = L \leq U$. That is, t_{ap} might be just outside the interval.
4. It is better to choose t_{ap} between two t_i values rather than equal to a t_i value. It will be seen from the proof of (1) that, if t_{ap} is equal to one of the t_i values and if $g(t_{ap}) \leq \delta$, then the error rate is bounded above by $\alpha/2$, and the procedure is conservative. On the other hand, if t_{ap} is not equal to any t_i, then the error probability might be as large as α. So, choosing t_{ap} between two t_i values avoids unnecessary conservativeness. If $t_i < t_{ap} < t_{i+1}$, it does not matter what the exact value of t_{ap} is. The same confidence statement will be made in all cases. In fact, a stronger statement can be made. The confidence statement obtained by choosing $t_i < t_{ap} < t_{i+1}$ is never any worse than

the confidence statement made by choosing $t_i = t_{ap}$. If ϕ_i rejects, then the same confidence statement is made in both cases. But, if ϕ_i accepts, then no confidence statement is made if $t_i = t_{ap}$, while a confidence statement will be made if $t_i < t_{ap} < t_{i+1}$ and ϕ_{i+1} rejects.

It will be seen that the sequential use of tests to define these intervals is similar to a method used by HSU and BERGER (1999) to define stepwise confidence procedures. The fact that, although multiple tests are performed, they can all be done at the same level, $\alpha/2$, is related to the theory of intersection-union tests (IUTs) in BERGER (1982). BERGER, BOOS, and GUESS (1988) and DINSE, PIEGORSCH, and BOOS (1993) used sequential IUTs. MAURER, HOTHORN, and LEHMACHER, (1995) and KOCH and GANSKY (1996) discuss sequential IUTs in general settings.

2.2 *Example 1*

Table 1 displays summary data for the Example 1 experiment on concentrations of a blood analyte. There were 36 subjects with a few missing observations at times 0.3 and 16.0. Preliminary analysis suggested that the square root of concentration is approximately normally distributed.

Letting X_{ij} denote the square root of concentration on the jth subject at the ith timepoint, we might use the model

$$X_{ij} = g(t_i) + \epsilon_{ij}\,,$$

where we assume that the vectors of errors for each subject, $(\epsilon_{1j}, \ldots, \epsilon_{15,j})$ are a random sample from a multivariate normal population with mean vector $\mathbf{0}$ and unknown covariance matrix. We might expect that the variance of X_{ij} is smaller for $i = 1$ or 2 or 14 or 15 than for timepoints in the middle of the observed range. And we might expect that the covariance of two observations on the same subject, X_{ij} and $X_{i'j}$, is higher if i and i' are close together rather than farther apart. But, to use our confidence limits, we do not need to model any of these relationships.

All that is important is that, at a given timepoint, $X_{i1}, \ldots, X_{i,36}$ are a random sample from a normal population with mean $g(t_i)$. Thus we may use the pointwise t-statistics based on the square root of concentration compared to the square root of the threshold, $\delta = \sqrt{100} = 10$. We also need to specify t_{ap} and α before collecting the data. Typically, the scientists running the experiment would choose t_{ap} in a region where they expect the concentration to be high and choose α in the standard range of 0.01 to 0.10. Suppose that $t_{ap} = 4$ and $\alpha = 0.05$ were chosen.

The one-sided p-values in Table 1 are highly significant on the range $[0.7, 8.0]$. Thus, the 95% confidence region would be $[0.7, 8.0]$. Of course the same confidence region would have been obtained for any t_{ap} chosen in that range and for any α greater than 0.006.

Table 1
Data and pointwise t-tests for Example 1

Time	Concentration	n	t	p-value
0.3	46.5	32	−8.8	1.000
0.7	192.3	36	7.6	0.000
1.0	250.7	36	8.5	0.000
1.5	290.0	36	9.1	0.000
2.0	287.2	36	10.0	0.000
2.5	345.8	36	11.3	0.000
3.0	349.8	36	12.0	0.000
4.0	325.0	36	12.9	0.000
5.0	286.2	36	13.7	0.000
6.0	206.2	36	9.9	0.000
8.0	119.0	36	2.9	0.003
10.0	62.5	36	−10.4	1.000
12.0	36.4	36	−24.5	1.000
14.0	21.8	36	−39.6	1.000
16.0	17.0	28	−46.4	1.000

Note: t-statistics are based on the square root of concentration $-\sqrt{100}$. P-values are one-sided.

3. Continuous Modeled Functions

3.1 *Confidence procedure*

In this second scenario, we assume that $g(t)$ has a specific functional form. Based on all the data, we can estimate $g(t)$ for every value of t. For each value of t, let ϕ_t denote a level-$\alpha/2$ test of $H_{0t} : g(t) \leq \delta$ versus $H_{at} : g(t) > \delta$. The data in Figure 3 are like this. That is, we can estimate the difference between the two regression functions, for every t, and, for each t, we can test if the difference is nonpositive versus the alternative that it is positive.

We make only the following assumption about $g(t)$.

Assumption 2. $g(t)$ is a continuous function of t.

For this scenario, the confidence statement is defined as follows. Again, let t_{ap} be a fixed value. If $\phi_{t_{\text{ap}}}$ accepts $H_{0t_{\text{ap}}}$, no confidence statement is made. If $\phi_{t_{\text{ap}}}$ rejects $H_{0t_{\text{ap}}}$, let I denote the largest interval containing t_{ap} for which ϕ_t rejects H_{0t} for all $t \in I$. Then, the confidence statement "$g(t) > \delta$ for all t in I" is made. This confidence procedure controls the error rate at α in that, for any population satisfying Assumption 2,

$$P(\phi_{t_{\text{ap}}} \text{ rejects and } g(t) \leq \delta \text{ for any } t \in I) \leq \alpha. \tag{2}$$

This error rate is also verified in the Appendix.

Typically, the test ϕ_t will be defined in terms of a p-value, $p(t)$; ϕ_t rejects H_{0t} if and only if $p(t) \leq \alpha/2$. Then, graphically and/or numerically, it will be easy to determine the interval I where $p(t) \leq \alpha/2$. In most problems, $p(t)$ is a continuous function of t. In this case, I will be a closed interval, and, if L and U denote the endpoints of I, the confidence statement can be expressed as "$g(t) > \delta$ for all t in the interval $L \leq t \leq U$", just as in Section 2. In some unusual problems, the interval I might be open or half-open.

Note these facts about this procedure.

1. The tests, ϕ_t, are one-sided, level-$\alpha/2$ tests, just as for the procedure in Section 2.
2. The tests, ϕ_t, may be correlated, and the correlation may depend upon which pair of tests, ϕ_{t_1} and ϕ_{t_2}, we consider. The error rate (2) is still valid. In Figure 3, the tests for the regression function at the different values of t are correlated. This correlation does not need to be specifically calculated to control the error rate at α.

3.2 *Example*

Consider the data from Treatment 1 and Treatment 2 in Table 3.5.1 of Milliken (1992). Treatment 1 is the "no additive" data, and Treatment 2 is the "additive" data. It is believed that the additive might increase the yield. The data consist of observations on yield, y, and a covariate t (Milliken calls it X). Following Milliken, we model the regression relationship between t and y using quadratic functions,

$$y_{ij} = \alpha_i + \beta_{1i} t_{ij} + \beta_{2i} t_{ij}^2 + \epsilon_{ij}; \qquad i = 1, 2; \qquad j = 1, \ldots, 12.$$

Denoting the regression parameter vector as $\boldsymbol{\beta} = (\alpha_1, \beta_{11}, \beta_{21}, \alpha_2, \beta_{12}, \beta_{22})'$ and using ordinary least squares regression, we obtain the estimate $\hat{\boldsymbol{\beta}} = (0.073, 1.809, -0.139, 2.528, 2.039, -0.205)'$. This yields the two estimated quadratic regression functions shown in Figure 3a. The function of interest is

$$g(t) = (\alpha_2 + \beta_{12} t + \beta_{22} t^2) - (\alpha_1 + \beta_{11} t + \beta_{21} t^2),$$

and we wish to determine an interval of t values where $g(t) > 0$, i.e., those values of the covariate where the average yield with the additive exceeds the average yield without the additive. The test statistic used to test $H_{0t} : g(t) \leq 0$ versus $H_{at} : g(t) > 0$ is the usual t-statistic,

$$T_t = \frac{\boldsymbol{\lambda}_t' \hat{\boldsymbol{\beta}}}{s \sqrt{\boldsymbol{\lambda}_t' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{\lambda}_t}},$$

where $\boldsymbol{\lambda}_t = (-1, -t, -t^2, 1, t, t^2)'$, $s = 0.423$ is the root mean square error, and

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \begin{pmatrix} 1.222 & -0.537 & 0.048 & 0 & 0 & 0 \\ -0.537 & 0.281 & -0.027 & 0 & 0 & 0 \\ 0.048 & -0.027 & 0.003 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1.670 & -0.584 & 0.044 \\ 0 & 0 & 0 & -0.584 & 0.236 & -0.019 \\ 0 & 0 & 0 & 0.044 & -0.019 & 0.002 \end{pmatrix}.$$

The p-value is $p(t) = P(T \geq T_t^{\text{obs}})$, where T has a Student's t distribution with 18 degrees of freedom. The test of size $\alpha/2 = 0.025$ rejects H_{0t} if $p(t) \leq 0.025$.

The p-value $p(t)$, as a function of t, is graphed in Figure 3b. Suppose the starting value $t_{\text{ap}} = 6.5$ was chosen. Then it can be seen that $p(t) \leq 0.025$ for all t between $t = -0.997$ and $t = 7.489$. Thus, we make the confidence statement that the additive mean exceeds the no-additive mean for all t in $[-0.997, 7.489]$. (Of course, if the covariate is logically constrained to be nonnegative, then the interval $[0.000, 7.489]$ would be stated.)

4. Remarks

The two confidence statements we have proposed do not look like typical confidence sets. But, in fact, they are closely related to a more conventional confidence set. Define two parameters

$$t_* = \sup\{t : t \leq t_{\text{ap}} \text{ and } g(t) \leq \delta\} \tag{3}$$

and

$$t^* = \inf\{t : t \geq t_{\text{ap}} \text{ and } g(t) \leq \delta\}. \tag{4}$$

Then, it can be shown by arguments like those in the Appendix that $t_* < L$ is a $100(1-\alpha/2)\%$ upper confidence bound for t_*, and $t^* > U$ is a $100(1-\alpha/2)\%$ lower confidence bound for t^*. So, by Bonferroni's Inequality, $\{t_* < L, t^* > U\}$ is a $100(1-\alpha)\%$ confidence set for (t_*, t^*). The confidence statement we have proposed is a logical consequence of this confidence set.

The confidence statements we have described are similar to three confidence procedures described in BERGER, BOOS, and GUESS (1988) for mean residual life functions. The intervals there had one of three forms, $(-\infty, t_{\text{ap}}]$, $[t_{\text{ap}}, \infty)$, or $(t_{\text{ap}} - l, t_{\text{ap}} + l)$. Unlike the third of these intervals, the intervals we have described herein need not be symmetric about t_{ap}. This can be a real advantage. For the quadratic regression example in Section 3.2, a symmetric interval constructed similar to that described in BERGER, BOOS, and GUESS (1988) would state that $g(t) > 0$ for t in $t_{\text{ap}} \pm 1.099$. Because $t_{\text{ap}} = 6.5$ is close to the point where the two regression functions cross, the interval cannot extend far to the right. And, because it must be symmetric, it cannot extend far to the left, either. The interval $[-0.997, 7.489]$, calculated in Section 3.2, is much longer. The choice of t_{ap} has much less effect for the new intervals we have described.

Also, the new procedures are less dependent on the choice of t_{ap} in another way. If one of these procedures obtains the interval $L \leq t \leq U$, then the same interval would be obtained for any starting point t_{ap}^* satisfying $L \leq t_{ap}^* \leq U$. This is not the case for the symmetric intervals in BERGER, BOOS, and GUESS (1988). For those, each different starting value, t_{ap}, defines a different interval.

The starting value t_{ap} should always be chosen before the experiment in a region likely to have significant test results. This requires some knowledge of the physical process. A statistical approach which requires less knowledge is to repeat the confidence region method at k different starting points $t_{ap1}, \ldots, t_{apk}$, but using level α/k for each one. (That is, level-$\alpha/(2k)$ tests are used.) The union of these confidence regions will have level α, and the experimenter is protected against an unlucky choice of t_{ap}.

Multiple starting values might also be used if the experimenter suspects that $g(t)$ is multimodal. Starting values near the suspected modes could be chosen. The resulting confidence set could then consist of intervals containing each mode.

TSUTAKAWA and HEWETT (1978), SPURRIER, HEWETT, and LABABIDI (1982), and BERGER (1984) discuss hypothesis tests for comparing two regression functions. Rejection of the null hypothesis corresponds to the statement that one regression function is greater than the other on a specified set. An advantage of our new procedures is that the set that is stated is determined by the data, not prespecified in the hypotheses. Also, our new procedures are more general and can be used for other kinds of functions besides regression functions.

5. Monte Carlo Results

We report here on a small Monte Carlo experiment used to illustrate how the proposed confidence procedures perform in a simple setting. We shall consider a repeated measurement study with data taken at 24 equally-spaced time points on $n = 30$ individuals. All individuals are independent, and the ith individual's data will be generated as

$$X_{ij} = \mu_j + \epsilon_{ij}, \qquad j = 1, \ldots, 24,$$

where the ϵ_{ij} form an AR(1) normal times series process, that is, the ϵ_{ij} are standard normal random variables with $\text{cov}(\epsilon_{ij}, \epsilon_{i,j+h}) = \varrho^h$. Suppose we are interested in finding an interval of values where the means are greater than 0.

First consider results for the case where the means μ_j are all 0, and within an individual the data are 1) independent ($\varrho = 0$) and 2) correlated ($\varrho = 0.8$). Since the data were assumed to be normally distributed, we used a one-sample t-statistic at each time point.

Recall that we form our confidence sets by moving left and right from t_{ap} including all time points where the t-statistic is significant at the 0.025 level. These sets can be empty, or consist of one or more consecutive time points. In

this null case where all the means are identically 0, a mistake is committed if a nonempty set is obtained. Thus for our method of set construction, a mistake will occur when we reject the null hypothesis at either of the t_i's adjacent to t_{ap}. The probability of a mistake is bounded by $2(0.025) = 0.05$.

Now consider the naive procedure that uses the first set of consecutive points (starting from $t_{ap} = 1$) where the t-statistic is significant at the 0.025 level. We can calculate the probability of a mistake for the $\varrho = 0$ case directly as P (nonempty set) $= P$ (maximum of 24 independent t-statistics is significant at $\alpha = 0.025$ level) $= 1 - (0.975)^{24} = 0.455$. The naive method is thus wildly liberal in that the confidence set is nonempty 45.5% of the time when it would be advertised as having nonempty results only 5% of the time.

Moving to Case 2 with $\varrho = 0.8$, using the same naive procedure from the previous paragraph, a simulation of 1000 replications yielded a nonempty set 243 times or an estimated error rate of 24.3%. Other schemes can be constructed such as taking sets for which we get at least two consecutive significant t-tests. The latter also fails in the $\varrho = 0.8$ case yielding two or more consecutive test rejections in 8.2% of the replications.

Now we turn to several alternatives where the mean function μ_j is not exactly zero. The first alternative pattern we call the "box" pattern because the means have a box shape: $\mu_j = \mu_0$ for $j = 5, \ldots, 20$ and $\mu_j = 0$ otherwise. Table 2 has results for the box pattern using $\mu_0 = 0.5$ and $\mu_0 = 1.0$. The second alternative pattern we call "sine" because $\mu_j = \mu_0 \sin(\pi(j - 4.5)/16)$ for $j = 5, \ldots, 20$ and $\mu_j = 0$ otherwise. The same μ_0 values were used for the sine pattern as were used for the box pattern. Independent ($\varrho = 0.0$) and correlated ($\varrho = 0.8$) data were used as in the null case discussed above.

The "Proportion Non-empty" column of Table 2 reports the proportion of cases where the t-statistic rejects H_{0i} for at least one t_i adjacent to t_{ap}. Of course for the "box" pattern there is no difference between starting at $t_{ap} = 12.5$ and at $t_{ap} = 7.5$ since the true means are the same at both places. There is a large difference for the "sine" pattern (0.94 compared to 0.52 and 0.84 compared to 0.43) since the means at $t = 12$ and $t = 13$ are higher than those at $t = 7$ and $t = 8$. This illustrates the advantage of having a good guess for t_{ap}.

Recall that the left endpoint of our interval is also a 97.5% upper confidence bound for the population onset value defined by t_* in Section 4. The fifth column in Table 2 is the proportion of left endpoints which were $t = 1$, 2, 3, or 4. These are "misses" in the sense that the left endpoint should never be lower than $t = 5$ because $t = 5$ is the first nonzero population value where we obtain data. Similarly, the 6th column is the proportion of cases where the right endpoint is $t = 21$, 22, 23, or 24. The values in columns five and six of Table 2 are bounded by 0.025 except for random variation.

The "Aver. Onset" column of Table 2 reports the average values of the left endpoint of our 95% confidence region for all replications where a nonempty result occurred. Similarly the "Aver. Endpoint" is the average value of the right endpoint

Table 2
Results for 95% intervals when population means are not zero

Mean Pattern	μ_0	Start (t_{ap})	Prop. Non-empty	Prop. Miss Left	Prop. Miss Right	Aver. Onset*	Aver. Endpoint*	Aver. Length**
			$\varrho = 0.0$					
Box	0.5	12.5	0.94	0.001	0.007	10.0	14.8	4.5
Box	0.5	7.5	0.94	0.017	0.002	6.1	10.1	3.8
Box	1.0	12.5	1.00	0.020	0.025	5.0	20.0	15.0
Box	1.0	7.5	1.00	0.020	0.025	5.0	20.0	15.0
Sine	0.5	12.5	0.94	0.000	0.000	10.9	14.0	2.9
Sine	0.5	7.5	0.52	0.000	0.000	7.5	8.9	0.7
Sine	1.0	12.5	1.00	0.000	0.000	7.3	17.7	10.4
Sine	1.0	7.5	0.98	0.000	0.000	7.0	16.9	9.7
			$\varrho = 0.8$					
Box	0.5	12.5	0.85	0.022	0.023	7.7	17.3	8.2
Box	0.5	7.5	0.83	0.026	0.015	5.5	14.3	7.3
Box	1.0	12.5	1.00	0.024	0.028	5.0	20.0	15.1
Box	1.0	7.5	1.00	0.024	0.028	5.0	20.0	15.1
Sine	0.5	12.5	0.84	0.005	0.006	9.6	15.4	4.9
Sine	0.5	7.5	0.43	0.008	0.004	7.2	13.3	2.6
Sine	1.0	12.5	1.00	0.016	0.013	7.0	18.0	11.0
Sine	1.0	7.5	0.93	0.016	0.012	6.8	17.9	10.4

Sample size $n = 30$, $k = 24$ time points, 1000 replications.
* Based only on non-empty cases.
** Empty regions counted as length = 0.

of our 95% region for all replications where a nonempty result occurred. These average onset and endpoint values should always be viewed in light of the proportion of nonempty replications. Thus the average onset value of 7.5 in Row 6 of Table 2 is the average over 520 nonempty results. It is a little hard to compare the average value 10.9 in the fifth row with the value 7.5 in the sixth row because the result for the fifth row is based on $940 - 520 = 420$ more intervals than for the sixth row. We can compare average lengths, however, by using 0 for the length of the empty sets: $[(940)(14.0 - 10.9) + (60)(0)]/1000 = 2.9$ for row five versus $[(520)(8.9 - 7.7) + (480)(0)]/1000 = 0.7$ for row six. These average lengths appear in the last column of Table 2.

Comparing Rows 1 and 2 of Table 2 is much easier because the μ_j are both 0.5, and thus they have the same nonempty rate (0.94). The comparison of these two rows shows that starting our procedure at a t_{ap} in the middle of the region where μ_j is positive ($t_{ap} = 12.5$) as compared to more to the left side ($t_{ap} = 7.5$) gives a

longer average length, 4.5 versus 3.8, a better average right endpoint, 14.8 versus 10.1, but not as good an average left endpoint, 10.0 versus 6.1. Note that the starting point made little difference when $\mu_0 = 1.0$.

The $\varrho = 0.8$ portion of the table indicates that the procedure works quite well for positively correlated data. This is important because repeated observations on the same individual are often positively correlated. The Proportion Non-empty columns are about the same for the $\varrho = 0$ and $\varrho = 0.8$ portions of Table 2. This is as it should be because whether or not the interval is nonempty is determined by only the tests on each side of t_{ap}. It has nothing to do with the correlation structure. The positive correlation actually helps the procedure to estimate the onset and endpoint better. In all eight cases, the average estimated onset for the $\varrho = 0.8$ data is closer to the true value of 5 than for the $\varrho = 0$ data. Similarly, in all eight cases, the average estimated endpoint for the $\varrho = 0.8$ data is closer to the true value of 20 than for the $\varrho = 0$ data. For some cases, e.g., all the sine cases, the procedure is very conservative for the $\varrho = 0$ data. That is, the proportions of misses are very small. For these cases, the procedure is much less conservative for the $\varrho = 0.8$ data, with error rates on each end closer to the target value of 0.025. Thus, the confidence procedure appears to perform quite well for positively correlated data. This is achieved without any complicated modeling of the correlation structure. In fact there is no estimation of the correlations in our procedure.

6. Conclusions

In this paper we have presented very general methods for making inference about the onset and duration defined by a population function such as a concentration curve or survival curve or the difference between two such functions. The key features of the approach are

1. only pointwise tests are needed (because of intersection-union test theory)
2. models may be parametric or nonparametric
3. modeling of correlation is not required.

These methods are widely applicable because of these nonrestrictive features.

Acknowledgement

We would like to thank El Giefer, Bob Small, and Yin Yin for motivating discussions and examples.

Appendix: Error Rates of the Two Procedures

The proofs of the two error rates, (1) and (2), follow. The two probabilities on the left-hand sides of (1) and (2) will be denoted by $P(\text{error})$.

Proof of (1). There are several cases to consider.

Case 1, $g(t_i) > \delta$ for all $i = 1, \ldots, k$: Then, by Assumption 1, $g(t) > \delta$ for all t in $t_1 \le t \le t_k$. It is always the case that $L \ge t_1$ and $U \le t_k$. So any confidence statement that is made will be correct, and $P(\text{error}) = 0$.

Case 2, $t_{ap} = t_i$ for some $i \in \{1, \ldots, k\}$ and $g(t_i) \le \delta$: If ϕ_i accepts H_{0i}, then $i* = i^*$. Hence, $U = t_{i^*-1} < t_{i_*+1} = L$, and no confidence statement is made. So, the only way that an error can be made is if ϕ_i rejects H_{0i}. But, then $P(\text{error}) \le P(\phi_i \text{ rejects}) \le \alpha/2$, because ϕ_i is a level-$\alpha/2$ test of H_{0i} and H_{0i} is true.

Case 3, for some $j*$ and j^* satisfying $1 \le j* < j^* \le k$, $t_{j_*} < t_{ap} < t_{j^*}$, $g(t_{j_*}) \le \delta$, $g(t_{j^*}) \le \delta$, and $g(t_i) > \delta$ for all i in $j* < i < j^*$: (Note, this case includes the case that $j* = j^* - 1$ and there are no i's in $j* < i < j^*$.) If $j* < j^* - 1$, then by Assumption 1, $g(t) > \delta$ for all t in $t_{j_*+1} \le t \le t_{j^*-1}$. If ϕ_{j_*} and ϕ_{j^*} both accept, then $L \ge t_{j_*+1}$ and $U \le t_{j^*-1}$. So, either $L > U$ and no confidence statement is made, or $t_{j_*+1} \le L \le U \le t_{j^*-1}$ and the confidence statement is true.

Thus, the only way that an error can be made is if either ϕ_{j_*} or ϕ_{j^*} rejects. Therefore,

$$P(\text{error}) \le P(\phi_{j_*} \text{ or } \phi_{j^*} \text{ rejects}) \le \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha\,,$$

because both H_{0j_*} and H_{0j^*} are true, and both tests are level-$\alpha/2$ tests.

Case 4, for some $j^* \in \{2, \ldots, k\}$ satisfying $t_{ap} < t_{j^*}$, $g(t_{j^*}) \le \delta$, and $g(t_i) > \delta$ for all $i = 1, \ldots, j^* - 1$

Case 5, for some $j* \in \{1, \ldots, k-1\}$ satisfying $t_{ap} > t_{j_*}$, $g(t_{j_*}) \le \delta$, and $g(t_i) > \delta$ for all $i = j* + 1, \ldots, k$:

These two cases are similar. We will only prove the first. By Assumption 1, $g(t) > \delta$ for all t in $t_1 \le t \le t_{j^*-1}$. If ϕ_{j^*} accepts, then $U \le t_{j^*-1}$ and either no confidence statement is made or the confidence statement is correct. Thus, the only way that an error can be made is if ϕ_{j^*} rejects. Therefore, $P(\text{error}) \le P(\phi_{j^*} \text{ rejects}) \le \alpha/2$, because H_{0j^*} is true and ϕ_{j^*} is a level-$\alpha/2$ test.

These five cases exhaust all the possibilities. Therefore, in all cases, the error bound (1) is true.

Proof of (2).

Case 1, $g(t_{ap}) \le \delta$: If $\phi_{t_{ap}}$ accepts $H_{0t_{ap}}$, then no confidence statement is made. So, the only way that an error can be made is if $\phi_{t_{ap}}$ rejects $H_{0t_{ap}}$. But, then $P(\text{error}) = P(\phi_{t_{ap}} \text{ rejects}) \le \alpha/2$, because $\phi_{t_{ap}}$ is a level-$\alpha/2$ test of $H_{0t_{ap}}$ and $H_{0t_{ap}}$ is true.

Case 2, $g(t_{ap}) > \delta$: By the continuity Assumption 2, there is an open interval $(t*, t^*)$ containing t_{ap} on which $g(t) > \delta$. The endpoints $t*$ and t^* are defined in

(3) and (4). If $t_* = -\infty$ or $t^* = \infty$ or both, then (2) is bounded above by $\alpha/2$ or 0, as in Cases 1, 4, or 5 in the proof of (1). We will consider only the case $-\infty < t_* < t^* < \infty$. By Assumption 2, $g(t_*) = \delta = g(t^*)$. If ϕ_{t_*} and ϕ_{t^*} both accept, then either no confidence statement is made (if $\phi_{t_{ap}}$ accepts) or the interval I is a subset of the open interval (t_*, t^*). If the confidence statement is made, it is correct because $g(t) > \delta$ for all t in $t_* < t < t^*$. Thus, the only way that an error can be made is if either ϕ_{t_*} or ϕ_{t^*} rejects. Therefore,

$$P(\text{error}) \leq P(\phi_{t_*} \text{ or } \phi_{t^*} \text{ rejects}) \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha,$$

because both H_{0t_*} and H_{0t^*} are true, and both tests are level-$\alpha/2$ tests.

References

BERGER, R. L., 1982: Multiparameter Hypothesis Testing and Acceptance Sampling. *Technometrics* **24**, 295–300.

BERGER, R. L., 1984: Testing Whether One Regression Function is Larger than Another. *Communications in Statistics – Theory and Methods* **13**, 1793-1810.

BERGER, R. L., BOOS, D. D., and GUESS, F. M., 1988: Tests and Confidence Sets for Comparing Two Mean Residual Life Functions. *Biometrics* **44**, 103–115.

DINSE, G. E., PIEGORSCH, W. W., and BOOS, D. D., 1993: Confidence Statements about the Time Range Over which Survival Curves Differ. *Applied Statistics* **42**, 21–30.

HSU, J. C. and BERGER, R. L., 1999: Stepwise Confidence Intervals without Multiplicity Adjustment for Dose-Response and Toxicity Studies. *Journal of the American Statistical Association* **94**, 468–482.

KOCH, G. G. and GANSKY, S. A., 1996: Statistical Considerations for Multiplicity in Confirmatory Protocols. *Drug Information Journal* **30**, 523-534.

MAURER, W., HOTHORN, L. A., and LEHMACHER, W., 1995: Multiple Comparisons in Drug Clinical Trials and Preclinical Assays: A-Priori Ordered Hypotheses. In: J. Vollman (ed.): *Biometrie in der chemisch-pharmazeutischen Industrie* 6. Gustav Fischer Verlag, Stuttgart, 3–18.

MILLIKEN, G. A., 1992: *Analysis of Messy Data, Vol.* 3. Chapman and Hall, New York.

SPURRIER, J. D., HEWETT, J. E., and LABABIDI, Z., 1982: Comparison of Two Regression Lines Over a Finite Interval. *Biometrics* **38**, 827–836.

TSUTAKAWA, R. K. and HEWETT, J. E., 1978: Comparison of Two Regression Lines Over a Finite Interval. *Biometrics* **34**, 391–398.

DENNIS D. BOOS
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203
USA
E-mail: boos@stat.ncsu.edu

Received, November 1998
Revised, March 1999
Accepted, April 1999