SURVIVAL ANALYSIS WITH DISINFORMATION

Esther Portnoy
Associate Professor, Mathematics
State Farm Companies Foundation Scholar
University of Illinois at Urbana-Champaign
Email: portnoy@math.uiuc.edu
and
Rose M. Reynolds, co-author

By "disinformation" I mean the inclusion of data that should not be considered in a study. For example, in a study of patients with a particular disease, records of some other patients (perhaps with a similar disease) are wrongly included in the data set; or a study of infant mortality is confounded by the inclusion of some records that should have been classified as fetal deaths. It might or might not be feasible to identify the extraneous records. As with censoring, no single method can be relied upon to detect, let alone correct, all possible types of disinformation. The starting point in any analysis must be a careful consideration of the data, and of possible errors. In this paper we discuss a case in which the presence of disinformation was clear, but it was not possible to determine with certainty which records were extraneous.

Data collection for this project was funded by a grant obtained by Kimberly Hughes from the National Institute on Aging (RO3 AG18122-02)

In the summer of 2002 I was contacted by Rose Reynolds, a Ph.D. student in life sciences who was studying the survival patterns of fruit flies. In particular, she was concerned about a plateau, where mortality rates that have been rising with age level off or begin to drop. She wanted to know how various factors affect the existence or location of such a plateau, and had accordingly conducted a number of experiments on homogeneous groups of newly emerged adult flies, recording their times of death. I said that I would take a look at the data and see what advice I could offer.

Table 1 shows the essential information from one of her experiments. Note that although she began with 190 flies (all having emerged from the pupal state the same day), there are 206 recorded times of death, and one censored observation. Clearly there are some extraneous data entries; but before attempting any analysis it was essential to understand how they might have arisen.

Ms. Reynolds told me that one factor that clearly increases mortality is crowding in the cages. Thus, if you start with 200 flies in a cage and observe decreasing mortality rates after, say, 50 days, it is not clear whether the surviving flies are intrinsically more robust, or whether they are simply lucky to have survived into a less stressful environment. So in about half of the experiments she had maintained a constant density (usually of about 200 flies) in the cage, by replacing dead flies with others. Of course the replacements were not immortal, and some of them died. But they were marked, and they were not supposed to be included in the death counts. Clearly some mistakes were made; but when? That is, how many of the deaths recorded at the various times should be removed from the data set? Perhaps a preliminary question should be, "Does it matter?" That depends on the use to be made of the estimated survival function. The median age at death will be almost the same after removing the first 17 deaths, or the last 17. Higher percentiles are affected more dramatically. As the example here illustrates, the misrecorded deaths tend to mask a plateau.

Table 1: Recorded deaths from a cohort of 190 flies; and adjusted death counts by two different methods. Note there was one censored observation on day 31. In order to simplify the analysis I assumed that was one of the original flies, so the adjusted deaths in columns A and B add up to 189 (except for roundoff error).

Day	Deaths Recorded	Adjusted Deaths (A)	Adjusted Deaths (B)
3	0	0	0
5	2	2	2
7	0	-0.02	0
9	1	0.98	0.97
11	1	0.97	0.96
13	4	3.96	3.94
15	12	11.91	11.89
17	10	9.78	9.72
19	14	13.68	13.58
21	53	52.52	52.39
23	34	32.95	32.65
25	19	17.59	17.19
27	12	10.39	9.94
29	20	18.28	17.80
31	11*	9.08	8.55
33	1	-1.03	0.00
35	0	-2.02	0
37	2	0.00	0.00
39	10	8.00	7.41
41	0	0	0

The Kaplan-Meier estimator, used where there is censoring, is based on the intuitive idea of "redistributing to the right" – each censored observation is redistributed among various times when that individual might have died, based on the other information that you have about the sample. In that spirit, I would like here to "undistribute" a total of 17 deaths, not necessarily in integer pieces, that I think might be the mistaken observations, again based on what I have in the data.

With an understanding about how the mistakes arose, I can say definitely that neither of the deaths recorded on the fifth day were errors, because at that time there were no marked flies in the cage. The likely number of errors is certainly larger when there are more of the marked replacement flies in the cage. Unfortunately, next to nothing was known about the replacement flies. In particular, they were not necessarily the same age as the original cohort, so if a death was erroneously recorded on day 20 it was not necessarily a death at age (ie duration since emergence from the pupa) 20 days. Accordingly, I felt I had no option but to use a very simple model for the combined risk of death-and-misrecord for the replacement flies, and thus assumed that there was a constant (but unknown) hazard rate α. Table 1 shows the results of two different efforts to recalculate the correct number of deaths.

Column A is based on the assumption that the (expected) number of misreported deaths on any given day is α times the (expected) number of replacement flies in the cage at the beginning of the day. The value of α is then determined so that the total expected number of misrecorded deaths equals 17. Since there is nothing in this method to force the number of misrecorded deaths on a particular day to be less than the total number of deaths reported that day, this method occasionally resulted in an adjusted number of deaths less than 0. Obviously, this method is inadequate.

Recognizing that the number of misrecorded deaths will also be larger on days when the total number of deaths is larger, I used the following formula to estimate d_x^*, the number of misrecorded deaths on day x:

$$\begin{aligned} d_x^* &= d_x \frac{\alpha P_x^*}{\alpha P_x^* + q_x P_x} \\ P_0^* &= 0 \\ P_{x+1}^* &= P_x^* + d_x - d_x^* \\ P_{x+1} &= P_x - d_x + d_x^* = n - P_{x+1}^* \\ q_x &= 1 - P_{x+1}/P_x \\ \sum d_x^* &= k \end{aligned}$$

Here, P_x is the number of survivors to day x from the original cohort of n flies, P_x^* is the number of replacement flies in the cage on day x, d_x is the recorded number of deaths, and k is the number of extraneous deaths (sum of recorded deaths and censored observations, minus n).

Clearly I need to use an iterative method to find the α and q_x that satisfy these equations, but I've written an S-plus program that does it quickly. The resulting adjusted deaths are shown in Column B of Table 1.

We now have an initial estimate of the survival function, essentially analogous to what the Kaplan-Meier method would give us. We can do all sorts of things with it. But we also need an indication of how good this estimate is. Figure 1 shows the estimate and a (pointwise) 95%

confidence interval, based on bootstrapping the original data, for the survival function; Figure 2 shows the corresponding estimated mortality rates and estimated confidence intervals. The theory justifying this approach is yet to be worked out.

Does this data set exhibit a plateau? The rates do seem to drop off (temporarily) after 21 days, but not with much significance. For the questions of interest, one probably needs samples of a few thousand flies, not a few hundred. However, Ms. Reynolds has data on about 20 similar experiments, with various factors. Combining them may lead to some interesting results, but it is also likely to be difficult (much as censoring complicates regression).

There is another piece of information in the data sets, that I have not mentioned: the initials of the person who collected the data each time. It would not be surprising to find a "collector effect" on the likely errors; but the data may be too skimpy for any significance. Analysis is yet to be carried out.

Figure 1: Estimated survival function, with approximate 95% confidence intervals determined by 200 bootstrap iterations. Asymmetric intervals are 2.906 times the interquartile differences.

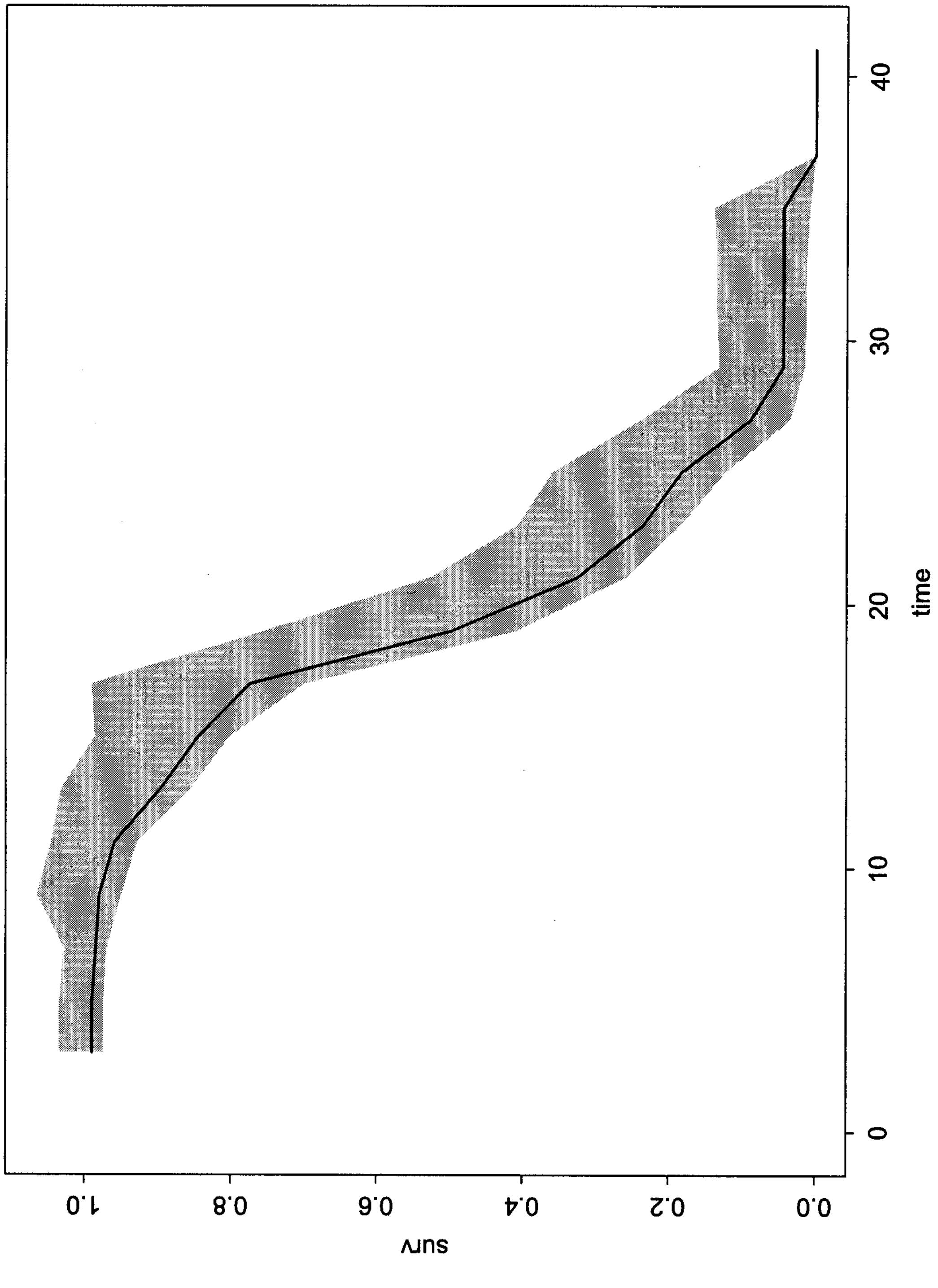
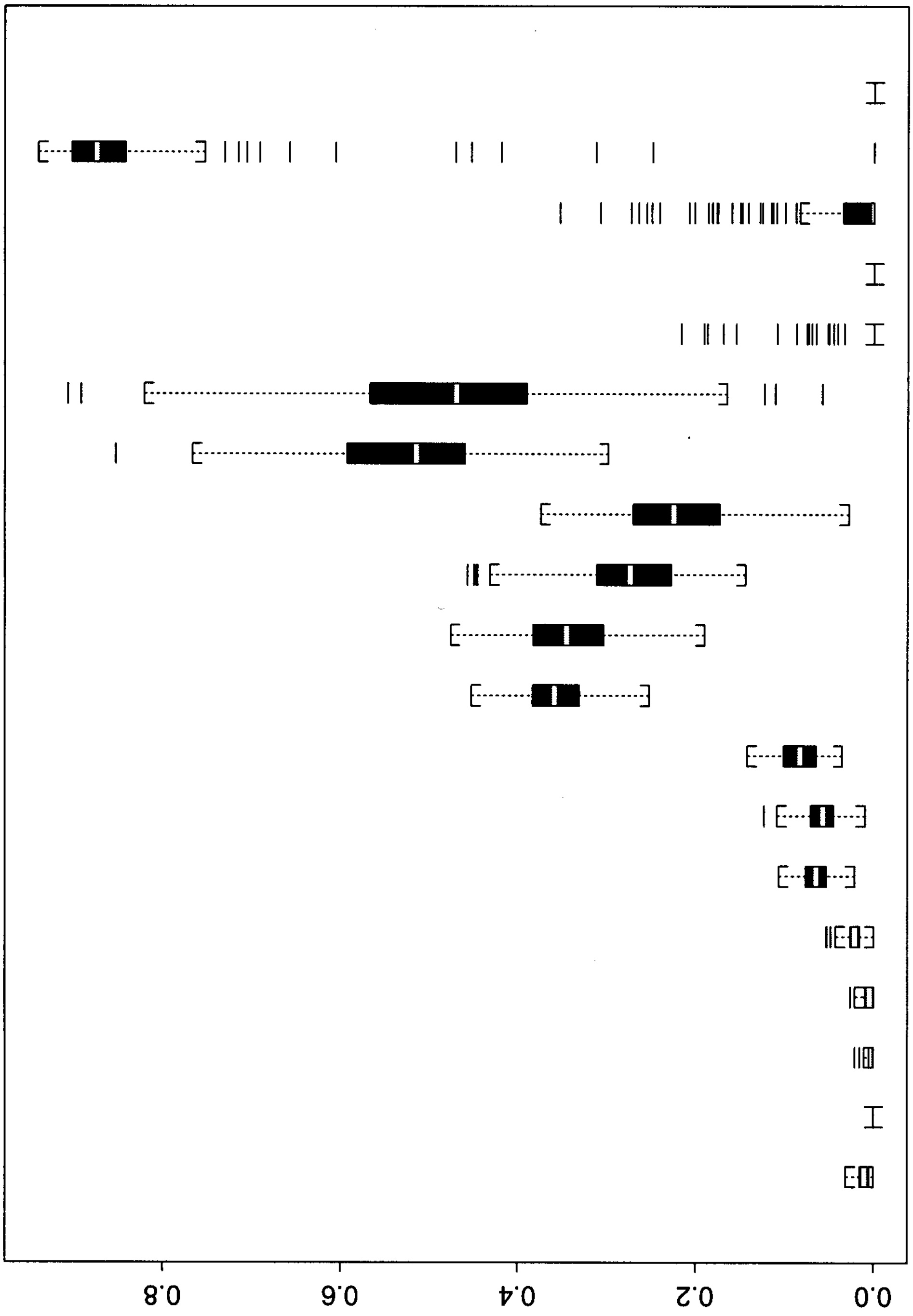


Figure 2: "Box and whisker" plot of estimated mortality rates. The central line is the median value, the "box" covers the 25th to 75th percentiles. Brackets enclose median ± 2.906 times the interquartile range, giving a 95% confidence interval. Horizontal lines indicate outliers.


0.8
0.6
0.4
0.2
0.0