

THE STATISTICAL ANALYSIS OF BEHAVIOURAL LATENCY MEASURES

Sergey V. Budaev

Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences, Moscow

Author's current address: Centre for Neuroscience, School of Life Sciences, University of Sussex, Brighton. BN1 9QG, UK
E-mail: s.budaev@sussex.ac.uk

This paper has been published in: **Budaev, S. V. (1997). The statistical analysis of behavioural latency measures. ISCP Newsletter, 14, No. 1, 1-4.**

Abstract: This article concerns two important problems with the statistical analysis of behavioural latency measures: they typically have severely skewed distributions, and are often censored (truncated). These problems, however, were not generally recognised by animal behaviour researchers: most people either allot an arbitrary score to all censored values or simply ignore them. Yet, such treatments could easily lead to dubious conclusions because of reduction of power and spuriously significant p-values. Thus, one should always use specially devised survival analysis methods whenever the study involves the measurement of censored latencies. The present article provides a short catalogue of some appropriate references, concentrating on the methods which are not “standard” for the common biomedical applications of survival analysis, but may be crucial in many behavioural studies. The statistical analysis of uncensored latencies is also discussed, with a particular attention to the analysis of variance.

INTRODUCTION

Latency measures are widely used in studies of animal behaviour. Typically, latencies are routinely analysed as all other behavioural measures, by applying standard parametric or non-parametric tests implemented in various statistical packages. There exist, however, several major problems with this approach, both statistical and methodological.

Many behavioural measures, such as the time devoted to a particular behavioural pattern, represent, probably, a gross outcome of numerous behavioural decisions and therefore the argument of the central limit theorem underpins the normality assumption. Unlike this, the latency reflects a single decision to evoke particular behaviour, even though the underlying mechanisms may be very complex. Therefore, a random decision-making process similar to radioactive decay (when the event may occur at any time with some constant probability) would result in an exponential distribution of the corresponding latency measures. The similar logic is used in modelling temporal and sequential dynamics of animal behaviour on the basis of continuous-time Markov chains (see Metz, 1981; Haccou & Meelis, 1992; Langton et al., 1995). Thus, extremely asymmetric and skewed (e.g. exponential, gamma or Weibull) rather than normal distributions are most typical for the latency data.

Furthermore, the observational period is often limited, so that in some individuals the desired event is likely not to occur. In such a case the exact latency is unknown (censored), although it is known that its actual value is greater than the total period of observation. Worse still, sometimes it may prove impractical or even impossible to avoid censoring at all, since an exponential or similarly skewed distribution may have a very long “tail” – one would simply have to wait for hours for the behaviour to occur!

ANALYSIS OF CENSORED LATENCIES

Thus, specialised statistical techniques are necessary for an analysis of censored behavioural latencies to be valid. Survival analysis has been especially devised for this sort of data (see Eland-Johnson & Johnson, 1980; Kalbfleisch & Prentice, 1980; Lawless, 1982; Allison, 1984; Cox & Oakes, 1984; Blossfeld et al., 1989; Lee, 1992, and also Haccou & Meelis, 1992 for general overviews), and some widespread methods were previously discussed in both animal behaviour (Fagen & Young, 1978; Bressers et al., 1991; Haccou & Meelis, 1992) and behaviour ecology (Muenchow, 1986; Pyke & Thompson, 1986) literature.

This is an extremely important issue, as it is known that the power is greatly reduced (by up to 60% and even more in some circumstances, see Bressers et al., 1991 for instance) if one applies ordinary statistical methods without the necessary adjustments for censors (e.g. treating them as if they were uncensored or merely omitting altogether). In some cases adjustment for censoring does not increase power, however. For example, there is no difference between unadjusted and censor-adjusted tests based on ranks (e.g. on the Wilcoxon statistic), provided all censored times are exactly the same (i.e. if the latencies are truncated), since in both cases the actual values are replaced by their ranks (see Bressers et al., 1991). Yet, in this case a large reduction of power may take place because the tied points cannot be ranked. Unfortunately, it is generally impossible to determine the degree to which censoring affects the results of tests and estimates; this depends on the sort of problem being analysed, type of the censoring mechanism and other factors. But in most cases simply omitting all censored values would lead to the greatest loss of the data analysis efficiency. Thus, applying standard statistical methods to censored data one must expect biased estimates and a very high risk of not detecting any effect while, in fact, it is significant. And even worse, spuriously significant effects might

appear in many circumstances, particularly when the censoring mechanism is not consistent across treatment groups. Finally, it is worth noting that complex parametric statistical procedures like ANOVA and ANOVA with repeated measures are likely to lead to particularly misleading results due to inconsistent estimation of variance components in the presence of censors (see Kimber & Crowder, 1990, for example). Because of inherent assumptions of linearity and zero expectation of residuals, Pearson product-moment correlation is also highly inappropriate in these cases (Amemiya, 1984; Muthén, 1989).

Now, the later versions of all comprehensive general-purpose statistical packages (such as BMDP, SAS, Solo, SPSS, Statistica and Systat) incorporate procedures to perform the common types of survival analysis, sometimes with its advanced extensions (e.g. competing risks analysis in BMDP 7). The user's manuals and on-line help systems of all these packages contain informal introductions to the respective methods and the basic examples of data analysis. In addition, McCullagh & Nelder (1983) showed how censored data could be put into the framework of generalized linear models, so that the software like GLIM can easily be adapted for some kinds of survival analysis.

However the survival analysis is borrowed from a very different field of study (primarily, human mortality and equipment failures) and does not meet some specific requirements of comparative psychology and ethology. For example, while a lot of techniques was developed for computing various descriptive statistics, distribution fitting, group comparisons and regression (in which the dependent variable is the survival time and predictors represent some risk factors) (see ref. above), relatively less was done for analysing repeated latency measures (also, these are never discussed in the context of ethological analysis of behavioural sequences). None the less, they do exist and may be readily used in the studies of animal behaviour. Schemper (1984 a,b) and Krauth (1988), for instance, developed generalised nonparametric correlation coefficients (based on

Kendall τ and Spearman ρ statistics, respectively), and Schemper (1984c) – a generalised Friedman test applicable to censored data. Furthermore, an ANOVA-like repeated measurements regression model (Crowder, 1985; Kimber & Crowder, 1990) with a flexible error structure, and a new approach to factor analysis of non-normal variables that are skewed and censored (Muthén, 1989) were recently developed. Finally, several years ago two extremely simple techniques were described (Theobald & Goupillot, 1990), which allow to collapse several repeated latencies to a single composite score, as well as to extend the Page test for ordered alternatives to censored data.

A minor problem might be that the methods of survival analysis are often based on the assumption of random censoring, but in most experiments the observational period is fixed, which would lead to fixed censoring times. Despite this, most techniques are relatively robust in cases of moderate censoring, and one could easily design an experiment of randomized length, assuring, of course, some fixed minimum duration to avoid unusually short observations (see Budaev, 1997 for an example).

In fact, survival analysis provides a powerful approach for analysis of the latency data, which can answer many important questions completely not recognised otherwise (also see Fagen & Young, 1978). For instance, in the context of “free” exploration of a novel adjacent arena in the guppy (*Poecilia reticulata*) I found (Budaev, 1997) that the distribution of the latency to enter a novel environment verged upon exponential distribution with repeated exposures to the same test situation. Exactly identical trend was also observed in case of the latency to perform predator inspection behaviour (Budaev, unpublished data). This means that after some experience the fish were entering (and inspecting) in a way resembling radioactive decay (that is, with a constant hazard rate), which may be meaningfully interpreted in terms of a reduction of curiosity.

Furthermore, survival analysis may be applied to a wide range of research problems far beyond the mere analysis of the latency data. For example, in studies of

learning, some portion of individuals often fail to reach the necessary criterion, inevitably leading to censored data. Within a very different context, Kimber & Crowder (1990) and Muthén (1989) showed (see also Amemiya, 1984) how censor-adjusted models can be employed in cases when substantial “ceiling effect” heavily undermines most parametric assumptions – all values reaching either of the scale bounds may be legitimately viewed left- or right-censored. Sometimes even missing values may be handled in this way (e.g. simply setting zero censored values if all these normally exceed zero, see Kimber & Crowder, 1990 for more discussion). This provides an important possibility to design repeated-measurements experiments, while each subject has one or more missing components in its data vector (e.g. because of ethical concerns, to diminish the carry-over effect of traumatic procedures).

Thus, one should always use the appropriate survival analysis methods whenever the study involves the measurement of latencies which are censored. To assist a broader use of the appropriate statistical approaches, I provide here a short list of the most straightforward alternatives to the ordinary statistical methods for censored data (Table 1).

ANALYSIS OF UNCENSORED LATENCIES

What if all latencies turned out uncensored, however, and how should one cope with the severe non-normality, typical in this case? Of course, nonparametric methods (e.g. Krauth, 1988) and, particularly, randomization tests (Manly, 1991) will work satisfactory in most such circumstances. Due to their advantages with small samples, the distribution-free statistical methods should be preferably applied to the latency data.

Furthermore, when the sample size is not too small, ANOVA, MANOVA and related statistical methods are fairly robust in cases of moderate deviations from

normality, for example, pronounced kurtosis and skewness. There is a common belief that, provided all samples are of equal size, mild variance inhomogeneity (with $\sigma_{\max} / \sigma_{\min} \leq \sqrt{3}$, see Wilcox, 1987) may also be inconsequential – it is the correlation between means and variances, that is most important (Lindman, 1974; Rencher, 1995 and many other textbooks on ANOVA). Yet, blindly assuming variance homogeneity when the deviations are, in fact, excessive will almost certainly have detrimental effects on both power and the probability of Type I error (e.g. Wilcox, 1987 cited several examples when violations of this assumption reduced power or, when sample sizes were different, inflated the p-values). Unfortunately, the correlation between means and variances is very likely to occur in cases of exponential and similar distributions, typical for the latency data.

Thus, the use of data transformations is generally unavoidable when the latency measures are analysed. Most often the common logarithmic and square-root transformations work quite well, although the resulting scores might sometimes be difficult to interpret meaningfully. In addition, Box & Cox (1964) and Lindman (1974) pointed out that the reciprocal transformation has a natural appeal for the analysis of survival times and latencies, which become easily interpretable in terms of “rate of dying” or risk (see also McCullagh & Nelder, 1983). Furthermore, in cases where the analysis of individual means and comparisons between them (by constructing suitable contrasts or employing multiple comparison procedures) rather than the overall significance of a treatment effects are of primary interest, several innovative ANOVA techniques specifically adjusted for various kinds of inhomogeneity and not requiring data transformations may be particularly appropriate (see McCullagh & Nelder, 1983; Wilcox, 1987 and Bechhofer et al., 1995).

DISCUSSION

To see how the students of animal behaviour treat behavioural latencies in their empirical research I surveyed several journals publishing research papers on animal behaviour (the 1995 volumes). The analysis showed (Budaev, 1996) that among the papers in which various latency measures were recorded and analysed (ranging from the latency to death to various display latencies) only about 10% used the appropriate statistical techniques (and even in these instances they were limited to the methods which are routinely used in medical sciences). Most often the authors allotted the total observational duration (or the maximum test length) to all censored cases, merely excluded all censored cases from the data analysis, or provided no information about the treatment of censored values (even though sometimes the actual data clearly implied that the censoring was rather heavy). Also, in all these investigations standard statistical methods were utilised (e.g. Kruskal-Wallis, Mann-Whitney, t- tests and ANOVA), although with parametric statistics the values were typically log-, square-root- or rank-transformed.

Thus, the statistical treatment of behavioural latencies is typically far from correct. Whilst many volumes specifically devoted to the survival analysis are available (see ref. above), the general textbooks most often used by animal behaviour researchers (e.g. Martin & Bateson, 1993) frequently do not even note them. This was, probably, the cause why censoring has not been generally recognised in the study of animal behaviour. However, special considerations are needed whenever the study involves the measurement of latency measures, both censored and uncensored. And inappropriate statistical analysis would at best result in a reduction of power and ineffective analysis, and at worst might lead to completely misleading inferences.

ACKNOWLEDGEMENTS

I thank Steve Langton for his helpful comment on an earlier draft of the manuscript.

REFERENCES

- Allison, P. (1984). Event history analysis: Regression for longitudinal event data. Beverly Hills, CA: Sage Publications.
- Amemiya, T. (1984). Tobit models: a survey. Journal of Econometrics, *24*, 3-61.
- Bechhofer R.E., Santner T.J., & Goldsman D.M. (1995). Design and analysis of experiments for statistical selection, screening, and multiple comparisons. New York: John Wiley.
- Blossfeld, H.-P., Hammerle, A., & Mayer, R. (1989). Event history analysis: Statistical theory and application in the social sciences. Hillsdale, NJ: Lawrence Erlbaum.
- Box, G.E.P., & Cox, D.R. (1964) An analysis of transformations. Journal of the Royal Statistical Society, Series B, *26*, 211-243.
- Bressers, M., Meelis, E., Haccou, P., & Kruk, M. (1991). When did it really start and stop: the impact of censored observations on the analysis of duration. Behavioural Processes, *23*, 1-20.
- Budaev, S.V. (1996, April). The statistical analysis of censored behavioural latency measures. Paper presented at the ASAB Easter Conference, Bolton, Lancashire, UK.
- Budaev, S.V. (1997). "Personality" in the guppy (Poecilia reticulata): A correlational study of exploratory behaviour and social tendency. Journal of Comparative Psychology. In Press.
- Cox, D., & Oakes, D. (1984). Analysis of survival data. London: Chapman & Hall.

- Crowder, M.J. (1985). A distributional model for repeated failure time measurements. Journal of the Royal Statistical Society, Series B, 47, 447-452.
- Eland-Johnson, R., & Johnson, N. (1980). Survival models and data analysis. New York: John Wiley.
- Fagen, R.M., & Young, D.Y. (1978). Temporal patterns of behavior: durations, intervals, latencies and sequences. In P.W. Colgan (Ed.), Quantitative ethology (pp. 79-114). New York: John Wiley.
- Haccou, P., & Meelis, E. (1992). Statistical analysis of behavioural data. Oxford: Oxford University Press.
- Kalbfleisch, J.D., & Prentice, R.L. (1980). The statistical analysis of failure time data. New York: John Wiley.
- Kimber, A.C., & Crowder, M.J. (1990). A repeated measurements model with applications in psychology. British Journal of Mathematical and Statistical Psychology, 43, 283-292.
- Krauth, J. (1988). Distribution-free statistics: An application-oriented approach. Amsterdam: Elsevier.
- Langton S.D., Collett D., & Sibly R.M. (1995). Splitting behaviour into bouts: a maximum likelihood approach. Behaviour, 132, 781-800.
- Lawless, J. (1982). Statistical models and methods for lifetime data. New York: John Wiley.
- Lee, E.T. (1992). Statistical methods for survival data analysis. New York: John Wiley.
- Lindman, H.R. (1974). Analysis of variance in complex experimental designs. San Francisco: W.H. Freeman.
- Manly, B.F.J. (1991). Randomization and Monte Carlo methods in biology. London: Chapman & Hall.

- Martin, P., & Bateson, P. (1993). Measuring behaviour, 2nd ed. Cambridge: Cambridge University Press.
- McCullagh, P., & Nelder, J.A. (1983). Generalized linear models. London: Chapman & Hall.
- Metz, H. (1981). Mathematical representations of the dynamics of animal behaviour. Ph.D. Thesis, Mathematisch Centrum, Amsterdam.
- Muenchow, G. (1986). Ecological use of failure time analysis. Ecology, *67*, 246-250
- Muthén, B. (1989). Tobit factor analysis. British Journal of Mathematical and Statistical Psychology, *42*, 241-250.
- Pyke, D.A., & Thompson, J.N. (1986). Statistical analysis of survival and removal rate experiments. Ecology, *67*, 240-245.
- Rencher, A.C. (1995). Methods of multivariate analysis. New York: John Wiley.
- Schemper, M. (1984a). Analyses of associations with censored data by generalized Mantel and Breslow tests and generalized Kendall correlation coefficients. Biometrical Journal, *26*, 309-318.
- Schemper, M. (1984b). Exact test procedures for generalized Kendall correlation coefficients. Biometrical Journal, *26*, 305-308.
- Schemper, M. (1984c). A generalized Friedman test for data defined by intervals. Biometrical Journal, *26*, 305-308.
- Theobald, C.M., & Goupillot, R.P. (1990). The analysis of repeated latency measures in behavioural studies. Animal Behaviour, *40*, 484-490.
- Wilcox, R.R. (1987). New designs in analysis of variance. Annual Review of Psychology, *38*, 29-60.

Table 1. A list of some alternatives to standard statistical methods applied in cases where the data values are censored, * indicates “standard” survival analysis methods, that are implemented in many statistical packages

Problem and the standard approach	Appropriate survival analysis methods	References
Analysis of distribution patterns, estimating and fitting parameters of distributions	Log-survivor plot, Kaplan-Meier estimate of the survival function, generalized least-squares estimates of distribution parameters (unweighted and weighted)*	Bressers et al. (1991); many tests are discussed by Haccou & Meelis (1992) and Lee (1992)
Comparing groups: t-test, Mann-Whitney or Kruskal-Wallis test	Cox F-test, Gehan’s Wilcoxon test, log-rank test, Prentice’s Wilcoxon test, Peto and Peto’s Wilcoxon test*	see Lee (1992) for an overview of many tests; see also Bressers et al. (1991); Pyke & Thompson (1986) provided an informal discussion in the ecological context
Aggregating several censored variables into a single composite	A simple scoring method	Theobald & Goupillot (1990)
Friedman test for repeated measures	Generalized Friedman test	Schemper (1984c)
Testing a monotonous trend in repeated measures	Page test	Theobald & Goupillot (1990)
Calculation of correlation between two censored variables (or one censored and another uncensored)	W-test (a generalized Spearman correlation test) Generalized Kendall correlation coefficient	see Krauth (1988) for a simple description Schemper (1984 a,b)
Multiple regression analysis in which the dependent variable is censored and predictors are uncensored, ANOVA	Cox proportional hazard regression model*	Allison (1984); Blossfeld et al. (1989); Lee (1992); an informal discussion in the ecological context is given by Muenchow (1986)
Multi-way repeated measures ANOVA	The multivariate Burr model	see Kimber & Crowder (1990) for an example of its application in psychology