

A robust approach to one-sample tests for recurrent events

Paola Rebora and Stefania Galimberti

Abstract In clinical and epidemiological studies one might face the problem of contrasting a study sample to a specified reference population. One-sample non-parametric tests are proposed here for inference on recurring events. The focus is on the marginal mean function of events and the basis for inference is the standardized distance between the observed and the expected number of events under a specified reference rate. In order to achieve greater robustness for inference, an alternative variance estimate was also considered. Both versions of the test were generalized to include weights and strata. The performance of these tests was investigated through simulation studies under various underlying event generation processes, such as homogeneous and non-homogeneous Poisson processes, autoregressive and renewal processes, with and without frailty effects. The robust versions of the test have been shown to be suitable in a wide variety of event generating processes. The motivating context has been a study on gene therapy in a very rare immunodeficiency in children, where a major end-point is the recurrence of severe infections.

Key words: recurrent events, one-sample statistical hypothesis test, Nonparametric statistical methods, survival analysis

1 Introduction

In some circumstances, one might face the problem of contrasting a study sample to a specified reference population. In 1975 Breslow developed the one-sample log-rank test [2] to compare the survival of a defined group of patients to a reference, but, to our knowledge, there is no corresponding test for a recurrent event end-

Paola Rebora e-mail: paola.rebora@unimib.it · Stefania Galimberti e-mail: stefania.galimberti@unimib.it

Center of Biostatistics for Clinical Epidemiology Department of Clinical Medicine and Prevention, University of Milan - Bicocca, via Cadore, 48 20052 Monza- Italy

point. Nonetheless, there is an increasing interest in methods for the analysis of situations where subjects may experience the outcome of interest more than once in time. Examples cover many areas of medicine, from cardiovascular to neurological disorders or cancer. In the two-sample context, both parametric and non-parametric statistical methods are available for inference, when the outcome of interest is a recurrent event [10, 7, 4]. The main aim of this work is to develop non-parametric tests to compare a recurrent event outcome in one study sample to a reference value.

2 Methods

Consider a context in which n subjects are observed for the occurrence of recurrent events up to τ , that is the largest of the observed follow-up times $(\tau_1, \tau_2, \dots, \tau_n)$ across the sample. Let $t_{i1} < \dots < t_{ij} < \dots < t_{iJ_i}$ be the subsequent event times of subject i ($i=1\dots n$), with t_{ij} being the time of occurrence of event j among J_i . The right continuous counting process $N_i(s)$ records the number of events experienced by subject i up to time s : $N_i(s) = \sum_{j=1}^{J_i} I(t_{ij} \leq s)$, where I represents the indicator function. Over the time interval $(s, u]$, the number of events for subject i is denoted by $N_i(s, u)$ with $u > s > 0$, and for infinitesimal time intervals, such as $(s - h, s]$, it is contracted as $dN_i(s) = \lim_{h \rightarrow 0} [N_i(s) - N_i(s - h)]$.

$Y_i(s)$ indicates whether subject i is at risk and under observation at time s . For example, if subject i is censored at time τ_i , then $Y_i(s) = 1$ for $s \leq \tau_i$ and $Y_i(s) = 0$ for $s > \tau_i$. Considering the whole sample, $N.(s) = \sum_{i=1}^n N_i(s)$ represents the total number of events in the sample up to time s and $Y.(s) = \sum_{i=1}^n Y_i(s)$ the total number of subjects at risk at time s . The standard assumption is made here, that for every subject i the observation pattern is independent of the event process.

Denote with $\mu_0(s)$ the completely specified reference mean function and with $\mu(s)$ the mean function of events in the study population: $\mu(s) = E[N_i(s)] = \int_0^s d\mu(u)$, where $d\mu(u)$ is the instantaneous rate function. The mean function of events can be consistently estimated by the non-parametric Aalen-Nelson estimator $\hat{\mu}(s) = \sum_{i=1}^n \int_0^s \frac{Y_i(u) dN_i(u)}{Y.(u)} = \int_0^s \frac{dN.(u)}{Y.(u)}$, where $dN.(s) = \sum_{i=1}^n Y_i(s) dN_i(s)$ [7, 9].

In order to test the following hypothesis: $H_0 : \mu(s) = \mu_0(s), \forall s > 0$, we propose a non-parametric test that derives from the two-sample test introduced by Lawless and Nadeau [7], and borrows the one-sample approach suggested by Breslow in the field of survival analysis [2].

The test statistic we developed is based on the weighted distance between the mean cumulative number of events observed in time in the sample and that expected under the null hypothesis of no difference with the reference rate:

$$U(\tau) = \int_0^\tau w(s) \{d\hat{\mu}(s) - d\mu_0(s)\} = \sum_{i=1}^n \sum_{j=1}^{J_i} w(t_{ij}) \frac{Y_i(t_{ij}) \cdot dN_i(t_{ij})}{Y.(t_{ij})} - \int_0^\tau w(s) d\mu_0(s). \quad (1)$$

Here, $w(s)$ is a predictable non-negative weight function with the property that $w(s) = 0$ whenever $Y(s) = 0$.

In the situation in which events are generated according to a memory-less process ($\text{Cov}(N_i(s, u), N_i(v, z)) = 0$ with $s < u < v < z$), it can be shown that, under the null, $U(\tau)$ is a mean zero martingale and its variance can be estimated by:

$$\text{Var}[U(\tau)] = \int_0^\tau \frac{w(s)^2}{Y(s)} d\mu_0(s). \quad (2)$$

The proposed test has the standardized form: $T = \frac{[U(\tau)]^2}{\text{Var}[U(\tau)]}$ and behaves asymptotically as a chi-square with 1 degree of freedom under the null hypothesis of no difference between the mean and the reference functions.

More generally, following Lin et al.[8], a robust version of the test can be obtained using an alternative variance estimate, without assuming independent increments in the number of events and predictability of $w(s)$:

$$\text{Var}_R[U(\tau)] = \sum_{i=1}^n \left[\int_0^\tau \frac{w(s)Y_i(s)}{Y(s)} (dN_i(s) - \frac{dN(s)}{Y(s)}) \right]^2. \quad (3)$$

This variance is a consistent estimator of the variance of $U(\tau)$, provided that $\text{Cov}(dN_i(s), dN_i(u))$ exists for all s, u in $[0, \tau]$ and $Y(t) \rightarrow \infty$ as $n \rightarrow \infty$. As n increases, the test approaches a chi-squared random variable with 1 degree of freedom under H_0 , for a wide class of underlying point processes.

In the spirit of the class of rank tests for censored survival data proposed by Harrington and Fleming [6], four different weights are considered in order to gain power under various types of alternative hypotheses on the mean function of the recurrent event process. A natural choice for the weight function is $w_{(1)}(s) = Y(s)$, that yields: $U_{(1)}(\tau) = \int_0^\tau Y(s) \left\{ \frac{dN(s)}{Y(s)} - d\mu_0(s) \right\} = N(\tau) - \int_0^\tau Y(s) d\mu_0(s)$. The corresponding test statistic represents the log-rank version for the one-sample recurrent events problem and is expected to be powerful against rate functions proportional to the reference. Alternative weight functions are also explored: $w_{(2)}(s) = Y(s)^2$, $w_{(3)}(s) = Y(s)(\tau - s)$ and $w_{(4)}(s) = Y(s)\hat{\mu}(s)$. We will refer to $[U_k(\tau)]$ as the expression (1) with these four weights ($k = 1, 2, 3, 4$) and consider the corresponding test statistic $T_{(k)}$ and $T_{R(k)}$ with variance (2) and (3), respectively.

3 Simulations

3.1 Simulations protocol

The theoretical properties of these tests were investigated through simulation studies under different scenarios, characterized by different event generating processes: homogeneous and non-homogeneous Poisson, renewal and autoregressive. A varying

rate across subjects was also considered in order to account for unmeasured covariates, as subject's susceptibility to the event, through a subject specific frailty effect z_i , simulated from a Gamma distribution with mean 1 and variance $\theta = 0.25, 1$ (homogeneous and non-homogeneous Poisson) or $\theta = 0.1, 0.2$ (renewal, autoregressive process).

In the Poisson Process scenarios we independently simulated inter-event times for each subject by the inverse method according to a constant (homogeneous) or time-varying rate (non-homogeneous). Here, we considered an increasing Weibull rate with shape parameter equal to 2 and varying scale $d\mu_0(s) = 2l_0^2s$, with $l_0 = 1, 5$. We also explored situations in which the hypothesis of independence between non-overlapping intervals no longer holds: a renewal process, where event intensity depends on the time of previous events, and a self-exciting process, where the occurrence of an event increases the risk of the subsequent ones. For this last scenario, an autoregressive process was considered, whose event intensity depends on the number of previous events $N(s^-)$. In this setting, inter-event times were simulated as realizations of exponential distribution with rates $d\mu_0(s) = \phi \cdot \exp[(0.1) \cdot N_i(s^-)]$ and $\phi = 1, 3$, by means of the inversion method as before.

Samples of different size ($n=25, 50, 100$) were considered and the maximum time of follow-up was set at 1 unit of time (i.e. year, $\tau = 1$). Individual censoring times (τ_i) were generated from an exponential distribution with rate δ , that was chosen in order to be 0 (no censoring) or one tenth of the initial event rate (mild amount of censoring). We considered the type I error rate, at a nominal level of 0.05, and the power for a two-tailed test. Power was evaluated with alternative hypotheses resulting in a reduction of around 20% and 50% in the mean number of events at 1 year in the experimental group with respect to the reference.

In each configuration, we applied in 2000 independent simulations the proposed non-parametric test, with the different weight processes, in the non-robust and robust version.

3.2 Results from simulations

In Table 1 the results of a non-homogeneous Poisson process with a Weibull shape are reported. In the setting without frailty ($\theta = 0$), the empirical type I error rates are consistent with the nominal level of 0.05 for the non-robust tests (T) and in larger samples (100 subjects) for the robust tests. In the mixed scenarios ($\theta > 0$), T presents inflated type I error rates, while the robust versions T_R generally maintain good control, even for sample sizes above 50 subjects. With increasing overdispersion and larger event rates, type I error inflation increases. For what concern power, in the non-mixed Poisson settings (with $\theta = 0$) the robust versions of the test are fairly more powerful than the non-robust versions. In the mixed process scenarios, interpretation of type II error rates of the non-robust versions are flawed because of the inflation in type I error rates. As expected, the power of the robust tests increases as event rates increase and decreases with increasing overdispersion. Compatibly

with the type of weights, in the proportional means setting here considered (1), the robust one-sample log-rank test $T_{R(1)}$ is the most powerful.

Under an autoregressive process, whose results are reported in Table 2, both the non-robust and robust versions of the test present inflated type I error rates for small sample sizes, but robust tests perform better for bigger sample sizes. With regard to power, robust tests behave similarly as in the Poisson settings.

The results shown here are only part of the extensive simulation study described in section 3.1, yet they well represent the general behavior.

Table 1 Empirical type I error rates (a) and power (b) for non-mixed and mixed non-homogeneous Poisson processes with Weibull shape for varying reference rate ($d\mu_0(s) = 2l_0^2s$), variance of the frailty effect (θ), censoring rate (δ) and sample size (n). The power refers to the alternative hypothesis $d\mu(s) = 0.8d\mu_0(s)$.

l_0	θ	δ	n	(a) TYPE I ERROR RATE					(b) POWER				
				$T_{(1)}$	$T_{R(1)}$	$T_{R(2)}$	$T_{R(3)}$	$T_{R(4)}$	$T_{(1)}$	$T_{R(1)}$	$T_{R(2)}$	$T_{R(3)}$	$T_{R(4)}$
1	0	0	25	0.048	0.075	0.075	0.077	0.094	0.166	0.282	0.282	0.249	0.309
1	0	0	50	0.066	0.069	0.069	0.065	0.069	0.310	0.400	0.400	0.334	0.381
1	0	0	100	0.050	0.061	0.061	0.057	0.060	0.527	0.598	0.598	0.477	0.536
1	0	0.1	25	0.054	0.074	0.074	0.085	0.093	0.140	0.268	0.270	0.249	0.289
1	0	0.1	50	0.048	0.058	0.058	0.062	0.083	0.264	0.384	0.381	0.310	0.362
1	0	0.1	100	0.052	0.059	0.060	0.060	0.062	0.499	0.595	0.592	0.442	0.512
1	1	0	25	0.171	0.097	0.097	0.109	0.111	0.258	0.244	0.244	0.228	0.270
1	1	0	50	0.163	0.068	0.068	0.077	0.082	0.354	0.305	0.305	0.262	0.312
1	1	0	100	0.157	0.052	0.052	0.055	0.055	0.536	0.426	0.426	0.379	0.407
1	1	0.1	25	0.156	0.100	0.100	0.102	0.122	0.213	0.242	0.241	0.227	0.258
1	1	0.1	50	0.155	0.070	0.070	0.082	0.085	0.320	0.294	0.293	0.260	0.294
1	1	0.1	100	0.166	0.068	0.067	0.073	0.074	0.512	0.428	0.430	0.374	0.381
5	0	0	25	0.054	0.066	0.066	0.069	0.071	1.000	1.000	1.000	0.991	0.997
5	0	0	50	0.044	0.053	0.053	0.050	0.058	1.000	1.000	1.000	1.000	1.000
5	0	0	100	0.046	0.056	0.056	0.055	0.047	1.000	1.000	1.000	1.000	1.000
5	0	0.5	25	0.058	0.080	0.075	0.082	0.080	0.990	0.994	0.992	0.966	0.968
5	0	0.5	50	0.045	0.056	0.058	0.061	0.062	1.000	1.000	1.000	0.999	1.000
5	0	0.5	100	0.045	0.051	0.051	0.050	0.054	1.000	1.000	1.000	1.000	1.000
5	1	0	25	0.707	0.096	0.096	0.100	0.100	0.827	0.360	0.360	0.359	0.353
5	1	0	50	0.667	0.075	0.075	0.072	0.075	0.902	0.458	0.458	0.452	0.458
5	1	0	100	0.701	0.063	0.063	0.063	0.059	0.978	0.661	0.661	0.654	0.656
5	1	0.5	25	0.691	0.098	0.098	0.095	0.102	0.800	0.343	0.342	0.337	0.328
5	1	0.5	50	0.683	0.077	0.075	0.074	0.079	0.876	0.419	0.422	0.427	0.399
5	1	0.5	100	0.697	0.074	0.074	0.074	0.073	0.955	0.585	0.590	0.597	0.554

Table 2 Empirical type I error rates (a) and power (b) for autoregressive processes with varying initial rate (ϕ), variance of the frailty effect (θ), censoring rate (δ) and sample size (n). The power refers to the alternative hypothesis with inter-event rate equals to $0.8\phi e^{0.1 \cdot N(s^-)}$.

ϕ §	θ	δ	n	(a) TYPE I ERROR RATE					(b) POWER				
				$T_{(1)}$	$T_{R(1)}$	$T_{R(2)}$	$T_{R(3)}$	$T_{R(4)}$	$T_{(1)}$	$T_{R(1)}$	$T_{R(2)}$	$T_{R(3)}$	$T_{R(4)}$
1	0	0	25	0.070	0.085	0.085	0.088	0.060	0.173	0.270	0.270	0.243	0.187
1	0	0	50	0.064	0.054	0.054	0.051	0.051	0.323	0.402	0.402	0.333	0.297
1	0	0	100	0.073	0.054	0.054	0.056	0.049	0.588	0.616	0.616	0.503	0.498
1	0	0.1	25	0.066	0.072	0.075	0.074	0.058	0.165	0.287	0.294	0.254	0.189
1	0	0.1	50	0.075	0.069	0.072	0.068	0.066	0.308	0.392	0.395	0.328	0.277
1	0	0.1	100	0.070	0.056	0.055	0.058	0.062	0.577	0.622	0.620	0.506	0.492
1	0.2	0	25	0.093	0.092	0.092	0.093	0.069	0.192	0.281	0.281	0.261	0.198
1	0.2	0	50	0.089	0.068	0.068	0.059	0.067	0.355	0.385	0.385	0.336	0.293
1	0.2	0	100	0.096	0.057	0.057	0.057	0.049	0.641	0.611	0.611	0.499	0.492
1	0.2	0.1	25	0.095	0.094	0.098	0.095	0.074	0.200	0.289	0.295	0.243	0.217
1	0.2	0.1	50	0.101	0.074	0.076	0.068	0.063	0.341	0.381	0.383	0.325	0.283
1	0.2	0.1	100	0.097	0.072	0.072	0.063	0.058	0.594	0.581	0.579	0.481	0.470
3	0	0	25	0.109	0.064	0.064	0.063	0.068	0.614	0.598	0.598	0.466	0.533
3	0	0	50	0.128	0.066	0.066	0.063	0.060	0.874	0.830	0.830	0.691	0.781
3	0	0	100	0.125	0.056	0.056	0.060	0.059	0.991	0.988	0.988	0.911	0.973
3	0	0.3	25	0.123	0.088	0.085	0.082	0.084	0.546	0.564	0.565	0.446	0.468
3	0	0.3	50	0.112	0.062	0.059	0.054	0.067	0.808	0.778	0.772	0.620	0.709
3	0	0.3	100	0.100	0.050	0.052	0.054	0.053	0.983	0.971	0.967	0.868	0.947
3	0.2	0	25	0.298	0.110	0.110	0.087	0.107	0.662	0.537	0.537	0.439	0.496
3	0.2	0	50	0.292	0.083	0.083	0.082	0.076	0.881	0.731	0.731	0.623	0.709
3	0.2	0	100	0.293	0.065	0.065	0.060	0.066	0.987	0.931	0.931	0.861	0.911
3	0.2	0.3	25	0.270	0.109	0.111	0.104	0.102	0.631	0.514	0.518	0.431	0.462
3	0.2	0.3	50	0.275	0.085	0.085	0.070	0.082	0.850	0.681	0.683	0.609	0.648
3	0.2	0.3	100	0.279	0.076	0.073	0.064	0.073	0.978	0.895	0.894	0.809	0.870

§ For $\phi = 3$, the maximum number of observable events was fixed at 20 to constrain the event rate of subjects with high frailty

4 Application to a clinical study

The motivating clinical context is represented by a study on gene therapy in a very rare immunodeficiency in children. The disease, Adenosine Deaminase (ADA) deficiency, is an autosomal recessive disease belonging to the group of severe combined immunodeficiencies (SCID). ADA-SCID is a fatal disorder, characterized by immunological defects and systemic toxicity. In the exceptional context of an experimental gene therapy for a very rare disease (around 1 every 1000000 live births) on a fragile population of pediatric patients, a non-randomized trial was designed to assess efficacy and safety, with an expected recruitment of 2 cases/year.

Recurrent infections are the hallmark of ADA-SCID and the inevitable result of the combined immune deficiency, thus the recurrence of severe infections, defined as requiring hospitalization, is one of the major study end-points. The presented tests were applied to infection data of the first 9 ADA-SCID patients treated with gene-

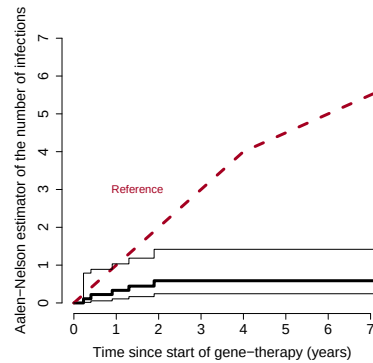
therapy and the reference was a decreasing piecewise constant rate (1 severe event per year in the first 4 years and 0.5 events for each following year), based on expert opinion and sparse historical data. The 9 patients experienced 5 events in a total of 38.97 person-years of observation and were followed for a maximum observation time of 7.6 years after gene-therapy [1]. The Aalen-Nelson estimate of the mean number of infections per person observed in the experimental group is plotted in Figure 1 and is far below the reference. The results obtained applying the log-rank version of the test are $T_{(1)} = 23.19$ ($P < 0.0001$). We obtained consistent results with the robust version ($T_{R(1)} = 115.54$, $P < 0.0001$), that accounts for subject-specific characteristics. Despite the small sample size, the magnitude of the effect provides evidence of the fact that gene therapy may markedly change the prognosis of these patients.

5 Conclusions

Non-parametric one-sample tests are proposed here for inference on recurring events. As the focus of the analysis is on the marginal mean function of events, we used, as basis for inference, the standardized distance between the observed and the expected number of events under a specified reference rate. The non-robust versions of the test were shown to be suitable for non-mixed Poisson processes, while robust versions were adequate for a wide variety of event generating processes (mixed, renewal and autoregressive), but less efficient for small sample sizes. While the log-rank type weight was shown to give more power under proportional mean rates, the remaining three weights were more suitable in situations departing from proportionality (data not shown). It is important to note that, as in the two-sample context, they cannot properly handle cases with the rate function crossing the reference.

The relative performance of the tests considered was evaluated based on a wide range of models for event generation and they have been shown to be suitable under

Fig. 1 Aalen-Nelson estimate of the mean number of infections per children treated with gene-therapy with point-wise 95% confidence intervals (continuous lines) and reference mean considered (dashed line).



very general processes, with a varying intensity in time and different models of auto-correlation. These tests can easily be extended to deal with a stratified design, which would be particularly useful in order to adjust for possible confounders. Moreover, these tests can also easily handle intermittent 'exposures' if, after the occurrence of an event, the subject temporarily ceases to be at risk.

Future work will address the derivation of formulas to calculate sample size for planning one-arm trials, based on this class of test statistics, while this can now be done only through simulations, as applied in the ADA-SCID study. The proposed procedure will also be extended to deal with a possible terminating event, by considering a mean function that accounts for the fact that the process ceases among subjects who experience that specific event [3, 5]. This, for example, would be important if ADA-SCID patients had died for the disease.

References

1. Aiuti, A., Cattaneo, F., Galimberti, S., Benninghoff, U., Cassani, B., Callegaro, L., Scaramuzza, S., Andolfi, G., Mirolo, M., Brigida, I., Tabucchi, A., Carlucci, F., Eibl, M., Aker, M., Slavin, S., Al-Mousa, H., Al, G.A., Ferster, A., Duppenhaler, A., Notarangelo, L., Wintergerst, U., Buckley, R.H., Bregni, M., Marktel, S., Valsecchi, M.G., Rossi, P., Ciceri, F., Miniero, R., Bordignon, C., Roncarolo, M.G.: Gene Therapy for Immunodeficiency Due to Adenosine Deaminase Deficiency. *N.Engl.J.Med.* **360**(5), 447 – 458 (2009). DOI 10.1056/NEJMoa0805817
2. Breslow, N.E.: Analysis of survival data under the proportional hazards model. *Int.Stat.Rev.* **43**(1), 45 – 58 (1975)
3. Cook, R.J., Lawless, J.F.: Marginal analysis of recurrent events and a terminating event. *Stat.Med.* **16**, 911 – 924 (1997)
4. Cook, R.J., Lawless, J.F., Nadeau, C.: Robust tests for treatment comparisons based on recurrent event responses. *Biometrics* **52**, 557 – 571 (1996)
5. Ghosh, D., Lin, D.Y.: Marginal regression models for recurrent and terminal events. *Stat.Sin.* **12**, 663 – 688 (2002)
6. Harrington, D.P., Fleming, T.R.: A class of rank test procedures for censored survival data. *Biometrika* **69**(3), 553 – 566 (1982). DOI 10.1093/biomet/69.3.553
7. Lawless, J.F., Nadeau, C.: Some simple robust methods for the analysis of recurrent events. *Technometrics* **37**(2), 158–168 (1995)
8. Lin, D.Y., Wei, L.J., Yang, I., Ying, Z.: Semiparametric Regression for the Mean and Rate Functions of Recurrent Events. *J.R.Stat.Soc.Ser.B-Stat.Methodol.* **62**(4), 711 – 730 (2000)
9. Nelson, W.: Graphical analysis of system repair data. *J.Qual.Technol.* **20**(1), 24 – 35 (1988)
10. Pepe, M., Cai, J.: Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *J.Am.Stat.Assoc.* **88**(423), 811 – 820 (1993)