*Statistical Modeling and Analysis of Repairable Systems**

Bo Henry Lindqvist
Department of Mathematical Sciences
Norwegian University of Science and Technology, Trondheim

Abstract: A repairable system can briefly be characterized as a system which is repaired rather than replaced after a failure. The most commonly used models for the failure process of a repairable system are nonhomogeneous Poisson processes (NHPP), corresponding to minimal repairs, and renewal processes (RP), corresponding to perfect repairs. The paper reviews models for more general repair actions, often called "better-than-minimal repair" models. In particular we study the class of so called trend-renewal processes (TRP), which has both the NHPP and the RP as special cases. Parametric inference in TRP models is considered, including cases with several systems involving unobserved heterogeneity. Trend testing is discussed when the null hypothesis is that the failure process is an RP. It is shown how Monte Carlo trend tests for this case can be made from the commonly used trend tests for the null hypothesis of a homogeneous Poisson process (e.g. the Laplace test and the Military Handbook test). Simulations show that the Monte Carlo tests have favorable properties when the sample sizes are not too small.

Keywords and phrases: Repairable system, nonhomogeneous Poisson process, renewal process, better-than-minimal repair, heterogeneity, trend, Monte Carlo test.

---

1 Introduction

A repairable system is a system which, after failing to perform one or more of its functions satisfactorily, can be restored to fully satisfactory performance by any method, other than replacement of the entire system [Ascher and Feingold (1984)].

Much of the literature on repairable systems seems to be motivated from applications to mechanical systems. However, the theory is of course not limited to such situations. An important application area is software reliability. Software systems will usually experience an increasing reliability due to correction of errors, and a number of different models have been suggested for the failure processes observed under testing as well as ordinary use. Finally, the theory of repairable systems will have applications in such areas as medicine and finance, and many other disciplines involving recurrent events.

A particular example of an application area, which has motivated much of the present work, is the oil and gas installations at the Norwegian continental shelf. The oil companies

*Invited lecture given at 1st International Conference on Mathematical Methods in Reliability, Bucharest, Romania, September 16-19, 1997, (MMR'97). To appear in conference proceedings.

possess large data bases containing survival- and maintenance data for various kinds of equipment. An example of one such database is the OREDA (=Offshore REliability DAtabase), which is a cooperation between the largest oil companies involved in the North Sea. Analyses of the data are used for, e.g., descriptions and comparisons of equipment of different types and manufacturers; predictions for new installations; identification of problems and trends with the aim of improving maintenance procedures; control and improvement of safety and cost-effectiveness; exchange of technology.

The purpose of the present paper is to review some basic models and methods used for the analysis of failure data from repairable systems, with emphasis put on our own research interests. The paper is therefore not particularly intended to give a broad review of the field.

In particular we shall consider "better-than-minimal repair models", i.e. non-Poisson models. Another topic considered is trend testing, when the null hypothesis is that the failure process is a renewal process, and the alternative is a monotone trend in the failure intensity. A brief survey of some important papers in the field is given in the following section.

---

2 "Major events" in the history of repairable systems reliability

The choice of "events" is of course subjective, and the list is by no means meant to be comprehensive.

A natural point of departure for a brief historic survey is the talk given by Harold Ascher and Harry Feingold at the Office of Naval Research Reliability Workshop in 1983, with the title *Repairable systems: Reliability's stepchild.* The year after appeared their famous and important book, *"Repairable Systems Reliability – Modeling, inference, misconceptions and their causes"*

This is a remarkable book, a statistical book written with passion and anger. Their starting point is the observation that reliability researchers, and practitioners using their methods, have not recognized the crucial difference between treatment of *repairable* systems and *nonrepairable* components. They illustrate this by their "happy" and "sad" systems, with successive times between failures given by, respectively,

$$\begin{array}{rl} & 15,\ 27,\ 32,\ 43,\ 51,\ 65,\ 177 \\ \text{and} & 177,\ 65,\ 51,\ 43,\ 32,\ 27,\ 15 \end{array}$$

Although the two *sets* of numbers are the same, considering them sequentially in time the first system is apparently getting better and better while the second is becoming increasingly worse. Nevertheless, it is believed that even today the method most frequently used by reliability engineers is to treat the numbers as i.i.d. from some distribution and then fitting a suitable parametric model.

The lesson to learn from Ascher and Feingold's book is that repairable systems need to be analysed by nonstationary stochastic point processes rather than the more elementary methods for random samples. However, the book was of course not the first source of such methods. Statistical models and methods appropriate for analysis of repairable systems were known at least from the fifties (see below). But no reliability textbook before 1978 provided *any* test for distinguishing between cases like the "sad" and "happy" systems.

Still, as pointed out by Ascher and Feingold, the world's oldest known formal statistical test, Laplace's test from 1773, would be appropriate to distinguish between the cases. This test is believed to still be the one most used for trend testing (see section 7). Moreover, the book by Ascher and Feingold is still the basic reference on repairable systems reliability.

The next "event" that we shall mention is the talk by David R. Cox, read before the Royal Statistical Society in London in March 1955 and published in Cox (1955). This is one of the first comprehensive treatments of appropriate statistical methods, where reliability applications are explicitly mentioned.

Cox touches a large number of problems, most of them motivated from the clothing industry. Examples of topics of particular importance for reliability applications are: Trend tests, testing whether a failure process is Poisson (which Cox calls *randomness*), autocorrelated intervals, doubly stochastic Poisson processes, heterogeneity (called *components of variance*), correlations between different types of events, mean repair times, availability.

In addition to demonstrating a large number of useful statistical techniques, the paper is special – as expressed by one of the discussants – "because it recognizes that the application of statistics is an *art*; that we cannot merely follow standardized procedures for any situation and always find an answer". Many results from the paper are contained in the subsequent book by Cox and Lewis (1966), which still is one of the most cited sources on the subject.

Another early contribution to the theory, and indeed a very much cited paper, is the 1963 paper by Frank Proschan, "Theoretical explanation of observed decreasing failure rate". Reliability engineers analysing failure data frequently experience an apparent *decreasing* failure rate, often counterintuitive in view of wear and ageing effects. Proschan's paper is important because it is presumably the first to point out that such observed decreasing rates could possibly be caused by *heterogeneity* between systems. Proschan presents failure data from 13 airconditioner systems on Boeing 720 airplanes. He argues that failures occur according to a homogeneous Poisson process for each separate plane, but with different rates for each plane. He demonstrates this statistically by using a result from Barlow, Marshall and Proschan (1963) which implies that a mixture of exponential distributions has a decreasing failure rate.

Proschan's test is constructed by pooling all the interfailure times, from all the planes, and then using a test statistic which is sensitive to a decreasing failure rate. In this way he gets a P-value of 0.007 for the null hypothesis of *identical* homogeneous Poisson processes. We also tried the test on "variance components" suggested by Cox (1955), and got approximately the same P-value. We return to the airconditioner data later.

We close the introductory historical remarks by mentioning a definite breakthrough in the analysis of lifetime data. This is the martingale approach to statistical inference in point processes, as introduced by Aalen in his Ph.D. thesis in 1975, and then further refined and applied in the subsequent years. A comprehensive source is the book Andersen et al. (1992). One of the main features is the ability of taking into consideration rather complicated history processes and censoring mechanisms, thus enabling a realistic and intuitive way of modeling.

---

3 Notation and basic definitions

We shall in this paper restrict attention to systems which following a failure are immediately restored to a functioning state. For simplicity of notation we shall consider systems observed from time $t = 0$, and let $T_1, T_2, \ldots$ denote the successive *failure times*. The assumption of 0 repair time is reasonable if repair times are negligible compared to times between failures. The assumption can also be justified provided the time scale is taken to be the *operation time* of the system. In the latter case, a marked point process may keep track of the repair times, and possibly additional information such as production conditions, maintenance and external conditions. However, marked processes will not be considered in this paper.

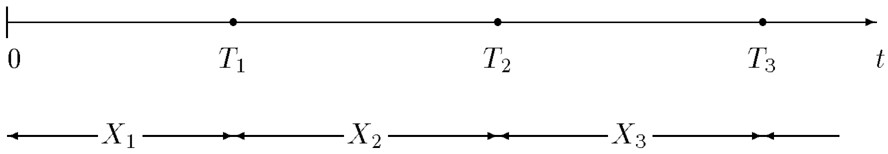

Figure 1: Failure times T_i and interfailure times X_i of a repairable system.

We shall let the *interfailure times*, i.e. times between failures, be denoted $X_1, X_2, \ldots$. Thus we have

$$X_i = T_i - T_{i-1}, \quad i = 1, 2, \ldots$$

where for convenience we define $T_0 \equiv 0$. (See figure 1). The sequence $T_1, T_2, \ldots$ of failure times and the sequence $X_1, X_2, \ldots$ of interfailure times thus contain exactly the same information about a particular realization of the process. This also holds for the counting process representation

$$N(t) = \text{ number of failures in } (0, t]$$

A plot of $N(t)$ as a function of t (or equivalently, plotting i vs. T_i) is the usual way of depicting data from single repairable systems. An example is given in figure 2 which shows the observed failure process of the airconditioner system of one of the airplanes in the Proschan (1963) data.

The most popular stochastic models for repairable systems are the *renewal process*, RP, and the *nonhomogeneous Poisson process*, NHPP. The defining property of an RP is that the interfailure times X_i are i.i.d. with some distribution F. We shall denote this process by RP(F). The NHPP is defined in terms of an intensity function $\lambda(t)$ such that the probability of a failure in a (small) time interval $[t, t+\Delta t)$ is $\lambda(t)\Delta t + o(\Delta t)$, while the number of events in disjoint time intervals are stochastically independent. We denote this process by NHPP($\lambda(\cdot)$). A nice reference on point process models in reliability is Thompson (1988).

As we shall see in the next section, the RP and NHPP can be viewed as corresponding to the two extreme repair actions, *perfect repair* and *minimal repair*, respectively. In order to describe models for more general repair strategies, we need some more notation from the theory of point processes.

Let $\mathcal{F}_{t-}$ dentote the *history* of the failure process up to, but not including, t. Usually we shall think of $\mathcal{F}_{t-}$ as $\{N(s) : 0 \leq s < t\}$ which keeps record of all the failures before time t. It could, however, contain more information, e.g. in the case where a marked point process is used as mentioned earlier. The conditional intensity $\lambda(t|\mathcal{F}_{t-})$ is defined by [Andersen et al. (1992)]

$$\lambda(t|\mathcal{F}_{t-}) = \lim_{\Delta t \downarrow 0} \frac{Pr(\text{failure in } [t, t+\Delta t)|\mathcal{F}_{t-})}{\Delta t}$$

In other words, $\lambda(t|\mathcal{F}_{t-})\Delta t$ is approximately the probability of failure in the time interval $[t, t+\Delta t)$, conditional on the failure history before time t.

For statistical inference in models defined by the conditional intensity we need to write down the *likelihood function.* Suppose that a single repairable system with failure process $N(t)$ and conditional intensity function $\gamma(t) \equiv \lambda(t|\mathcal{F}_{t-})$ is observed from time 0 to time τ, resulting in failure times $T_1, T_2, \ldots, T_{N(\tau)}$. The *likelihood* function is then given by [Andersen et al. (1992)]

$$L_\tau = \{ \prod_{i=1}^{N(\tau)} \gamma(T_i)\} \exp(-\int_0^\tau \gamma(u)\mathrm{du}) \tag{1}$$

This likelihood function is valid under the assumption that τ is a so-called *stopping time*, which holds for most censoring schemes occuring in practice. In particular it holds for the two most common observation schemes for repairable systems, *time truncation* in which τ is a prespecified point in time, and *failure truncation* where the process is observed until a given number n of failures have occurred.

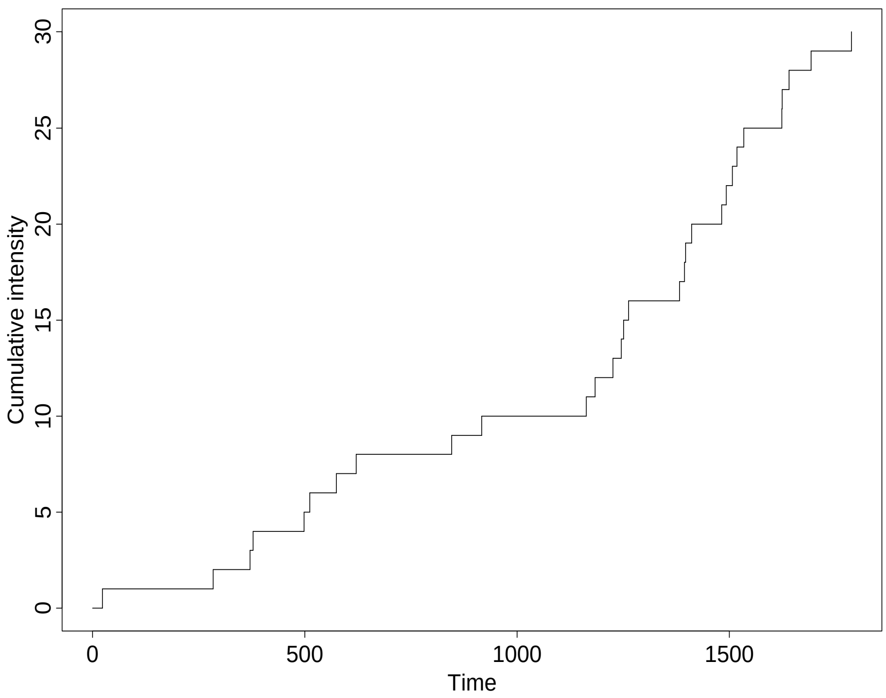

Figure 2: Plot of observed cumulative intensity, $N(t)$, for airconditioner failures of plane 7912 in the Proschan (1963) data.

For parametric models for $\gamma(t)$ there is a well developed theory for maximum likelihood estimation based on the martingale approach to point processes, see Andersen et al. (1992). Other relevant references for statistical inference in reliability models are Crowder et al. (1991) and Høyland and Rausand (1994).

4 Classification of repair actions

A common recipe for analysis of a set of data from a repairable system seems to be as follows. First, apply a test for *trend* in the interfailure times X_i (see section 7). If no significant trend is found, then model the X_i by an RP using the well established machinery for analysis of independent and identically distributed (i.i.d.) observations. Otherwise, use an NHPP model which is able to handle various kinds of trends through specification of the intensity function $\lambda(t)$. For example, a *deteriorating* system corresponds to an increasing function $\lambda(t)$, while an *improving* system corresponds to a decreasing $\lambda(t)$. Of course a *homogeneous* Poisson process, HPP(λ), corresponds to a constant intensity $\lambda(t) \equiv \lambda$.

As mentioned in the previous section, the RP and NHPP models correspond to, respectively, perfect and minimal repair. We shall now consider a unifying approach which makes these two models the extremes of a more general model. The unifying approach involves the specification of a *hazard rate* function $z(t)$, which we think of as being the hazard function of the time to first failure of the system. This function will then be used in describing the various models by means of their conditional intensity function $\lambda(t|\mathcal{F}_{t-})$. Recall that if T is a positive, continuous random variable, then the hazard rate is $z(t) = \lim_{\Delta t \downarrow 0} Pr(T \le t + \Delta t | T > t)/\Delta t$.

For a renewal process, RP(F), it is well known that

$$\lambda(t|\mathcal{F}_{t-}) = z(t - T_{N(t-)})$$

where z is the hazard rate corresponding to F, and $t - T_{N(t-)}$ is the time since the last failure strictly before time t (often called *local time*). Thus, after each failure, the rate of failures is returned back to the state at time 0, i.e. the system is as good as new immediately after each failure. This is the reason for calling the RP a *perfect repair* model.

On the other hand, an NHPP has the conditional intensity

$$\lambda(t|\mathcal{F}_{t-}) = z(t)$$

which means that after a failure, the failure rate is exactly as if no failure had ever occurred. Thus we have a reason for calling this a "minimal repair", since the system is only restored to a state where it is exactly as good (or bad) as it was immediately before the failure. It should also be remarked that for an NHPP the intensity function $\lambda(t)$ equals, for all t, the hazard rate $z(t)$ of the time to first failure.

For many applications it is more reasonable to model the repair action by something in between the two given extremes, so called "better-than-minimal repair" models. One such model is suggested by Brown and Proschan (1983), where at the time of each failure a perfect repair occurs with probability p and a minimal repair occurs with probability $1 - p$. More general suggestions were given by Kijima (1989).

In a recent paper, Dorado, Hollander and Sethuraman (1997) suggest a general repair model which contains all the above models as special cases. The definition in their paper is given in terms of conditional distributions of the X_i. An equivalent definition in terms of conditional intensities is

$$\lambda(t|\mathcal{F}_{t-}) = \theta(t) z\left(a(t) + \theta(t)(t - T_{N(t-)})\right)$$

which in addition to the basic hazard rate $z(t)$ is defined in terms of two random processes

$$\begin{aligned} a(t) &= \text{effective age at last failure before } t \text{, i.e. at time } T_{N(t-)} \\ \theta(t) &= \text{repair characteristic of last repair before } t \\ &\quad\ \text{(called "life supplement")}, 0 < \theta(t) \le 1. \end{aligned}$$

It is easy to see that an RP corresponds to $\theta(t) \equiv 1$, $a(t) \equiv 0$, while an NHPP has $\theta(t) \equiv 1$, $a(t) = T_{N(t-)}$. Kijima's (1989) Model I has $\theta(t) \equiv 1$, while the effective age $a(t)$ increases by the amount $D_j X_j$ at the jth failure time T_j, where $D_1, D_2, \ldots$ is a sequence of random variables independently distributed on $[0, 1]$ and independent of other processes. In Kijima's Model II, we have $\theta(t) \equiv 1$, while the effective age $a(t)$ immediately after the jth failure is set to $\sum_{k=1}^{j}(\prod_{i=k}^{j} D_i)X_k$. Moreover, when D_j is 1 with probability p and 0 with probability $1 - p$, we obtain the Brown-Proschan model.

Dorado et al. (1997) consider estimation in the general repair model when the effective ages $a(t)$ and life supplements $\theta(t)$ are observed at each failure, in addition to the failure times T_i. Using the general counting process framework they derive estimates and confidence bands for the distribution of time to first failure, under the condition that this has increasing failure rate. Statistical inference in the Brown-Proschan model was first studied by Whitaker and Samaniego (1989).

Berman and Turner (1992) consider estimation in the fairly general parametric model

$$\lambda(t|\mathcal{F}_{t-}) = g\left\{\sum_{k=1}^{p} \theta_k Q_k(t|\mathcal{F}_{t-})\right\} \tag{2}$$

where g is a known monotonic continuous function and the $Q_k(t|\mathcal{F}_{t-})$ are known functions of t and the history $\mathcal{F}_{t-}$. The data needed for estimation are just the failure times T_i.

The special case when $g(y) = e^y$ is applied to repairable sytems by Lawless and Thiagarajah (1996). In particular they consider the model

$$\lambda(t|\mathcal{F}_{t-}) = e^{\alpha+\beta g_1(t)+\gamma g_2(t-T_{N(t-)})} \tag{3}$$

where α, β, γ are parameters and g_1 and g_2 are known functions. This is, in fact, an example of what Cox (1972) called a *modulated renewal process.*

---

5 The trend-renewal process

An interesting feature of the model (3) is that it contains both an NHPP and an RP as special cases. This makes it possible to choose between a minimal and perfect repair model by use of statistical (likelihood ratio) tests for the parameters β and γ. We shall present and study an alternative model, which at the outset may seem very similar to (3), but which is not even covered by the general formula (2).

The idea behind the process, which was termed the *trend-renewal process* by Lindqvist (1993) and Lindqvist et al. (1994), is to generalize the following well known property of the NHPP. First, define the *cumulative* intensity function corresponding to an intensity $\lambda(\cdot)$ by $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$.

The result is that if $T_1, T_2, \ldots$ is a NHPP$(\lambda(\cdot))$, then the time-transformed stochastic process $\Lambda(T_1), \Lambda(T_2), \ldots$ is HPP(1).

We define the *trend-renewal process* (TRP) simply by allowing the above HPP(1) to be *any* renewal process RP(F). Thus, in addition to the intensity function $\lambda(t)$, for a TRP we need to specify a distribution function F of the interarrival times of this renewal process. Formally we can define the process TRP$(F, \lambda(\cdot))$ as follows:

Let $\lambda(t)$ be a nonnegative function defined for $t \geq 0$, and let $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$. The process $T_1, T_2, \ldots$ is called TRP$(F, \lambda(\cdot))$ if the time-transformed process $\Lambda(T_1), \Lambda(T_2), \ldots$ is RP(F), i.e. if the $\Lambda(T_i) - \Lambda(T_{i-1})$; $i = 1, 2, \ldots$ are i.i.d. with distribution function F.

Figure 3 illustrates the definition. Note that for an NHPP$(\lambda(\cdot))$, the RP(F) is replaced by HPP(1). Thus TRP$(1 - e^{-x}, \lambda(\cdot))$ = NHPP$(\lambda(\cdot))$. Also, TRP$(F, 1)$ = RP(F), which shows that the TRP generalizes both the RP and the NHPP.

Remark, however, that the representation TRP$(F, \lambda(\cdot))$ is unique only up to a scale factor in the following manner,

$$\mathrm{TRP}(F(x), \lambda(t)) = \mathrm{TRP}(F(cx), c^{-1}\lambda(t)) \tag{4}$$

for any $c > 0$. This means in particular that if the expected value of F exists, it can without loss of generality be assumed to be 1 (which it is, indeed, for the HPP(1)).

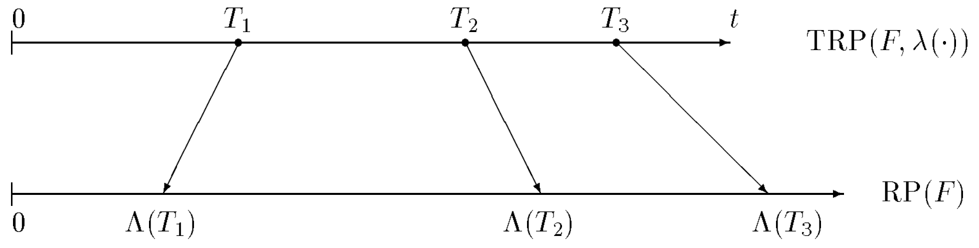

Figure 3: The defining property of the trend-renewal process.

The *inhomogeneous gamma process* of Berman (1981) is a special case of the TRP, given by TRP$(G, \lambda(\cdot))$ with G corresponding to a gamma-distribution with scale parameter 1. Berman studied in particular what he calls a *modulated gamma process*, with intensity function given by

$$\lambda(t) = \rho \exp\{\sum_{i=1}^{p} \beta_i z_i(t)\} \tag{5}$$

where ρ and the β_i are unknown parameters and the $z_i(\cdot)$ are given ("explanatory") functions. Of particular interest is the model for "cyclic trend of fixed frequency", for which

$$\lambda(t) = \rho \exp\{\beta_1 \cos(\omega t) + \beta_2 \sin(\omega t)\}$$

with ω fixed. Testing the null hypothesis $\beta_1 = \beta_2 = 0$ here leads to a test for the null hypothesis of an RP with gamma-distributed interfailure times vs. a *non*-monotonic (cyclic) trend.

We obtain the conditional intensity function for the TRP$(F, \lambda(\cdot))$ from the following informal derivation, where $z(t)$ here denotes the hazard rate corresponding to F. We have

$$\begin{aligned}
\lambda(t|\mathcal{F}_{t-})\Delta t &\doteq P(\text{failure in } [t, t+\Delta t)|\mathcal{F}_{t-}) \\
&= P(\text{failure in } RP(F) \text{ in } [\Lambda(t), \Lambda(t+\Delta t))|\mathcal{F}_{t-}) \\
&\doteq P(\text{failure in } RP(F) \text{ in } [\Lambda(t), \Lambda(t) + \lambda(t)\Delta t))|\mathcal{F}_{t-}) \\
&= z(\Lambda(t) - \Lambda(T_{N(t-)}))\lambda(t)\Delta(t)
\end{aligned}$$

so that

$$\lambda(t|\mathcal{F}_{t-}) = z(\Lambda(t) - \Lambda(T_{N(t-)}))\lambda(t) \tag{6}$$

It is seen that the intensity function of the TRP is a product of one factor, $\lambda(t)$, which depends on the age t of the system and one factor which depends on a transformed time from the last previous failure. In fact, time since last failure is measured on a scale depending on the current intensity of failures, in contrast to the model (3) of Lawless and Thiagarajah (1996).

We conclude the present section by showing how the trend-renewal process naturally arises in connection with bootstrapping of NHPP models. The idea of *bootstrapping* in statistics can briefly be explained as follows [Efron and Tibshirani (1993)]:

Suppose that the observed phenomenon obeys the unknown probability law $\mathcal{P}$, and let $\hat{\mathcal{P}}$ be an estimate of $\mathcal{P}$ obtained from the observations. Bootstrapping means to estimate a desired characteristic of $\mathcal{P}$ by deriving the corresponding characteristic for $\hat{\mathcal{P}}$. In practice, these characteristics are usually found by simulating a large number of realizations from the estimated law $\hat{\mathcal{P}}$.

Now consider an NHPP($\lambda(\cdot)$), which we for simplicity assume is observed until the nth failure, for a fixed number n (failure truncation). Let $T_1, T_2, \ldots, T_n$ be the observed failure times, and let $\mathcal{P}$ be the unknown probability law of these observations.

Suppose we have found an estimate $\hat{\lambda}(\cdot)$ of $\lambda(\cdot)$ based on the observations $T_1, \ldots, T_n$. Then a natural estimate $\hat{\mathcal{P}}$ of $\mathcal{P}$ is NHPP($\hat{\lambda}(\cdot)$), from which it would be easy to simulate new realizations, so called bootstrap samples.

Another way of constructing an estimated model, $\tilde{\mathcal{P}}$, is the following. Again, let $\hat{\lambda}(\cdot)$ be an estimate of $\lambda(\cdot)$. Then we can compute *estimated* residuals by

$$\hat{R}_i = \hat{\Lambda}(T_i) - \hat{\Lambda}(T_{i-1}) \; ; \; i = 1, \ldots, n$$

Let $\hat{F}$ denote the empirical distribution function corresponding to the $\hat{R}_i$, i.e. the distribution putting mass $1/n$ to each of $\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_n$. Then we shall define $\tilde{\mathcal{P}}$ as TRP($\hat{F}, \hat{\lambda}(\cdot)$), and may instead simulate bootstrap samples from this process. The motivation for this method is that the theoretical residuals of the NHPP are i.i.d. exponential with parameter 1.

To compare the two ways of bootstrapping, the $\tilde{\mathcal{P}}$ has the advantage of being more robust against departures from the NHPP assumption in the direction of TRP. This is quite similar to bootstrapping in ordinary regression models, where bootstrapping of residuals is more model-robust than a parametric bootstrapping with normally distributed residuals.

Bootstrapping in NHPP and TRP models is studied in more detail in Elvebakk (1998).

---

6 Statistical inference in trend-renewal processes

We shall consider statistical inference in cases where failure times $T_1, T_2, \ldots$ are observed for one or more systems of the same kind, and the failure process of each system is modeled by a TRP as defined in the previous section.

Recall that a TRP model is specified in terms of a distribution function F and an intensity function $\lambda(\cdot)$. A parametric model can thus conveniently be described in terms of parametric representations for each of F and $\lambda(\cdot)$. Semiparametric models may involve either a parametric F and nonparametric λ, or vice versa. Completely nonparametric models may also be considered.

An S-Plus library, called REIDAR (=REpaIrable systems Data Analysis Routines), containing routines for analysis of repairable systems, has been developed at the Norwegian University of Science and Technology as part of Elvebakk (1998), (for information, contact this author). In particular, the program has routines for graphical analyses, parametric estimation in TRP models, estimation of heterogeneity, trend testing and model checking.

Parametric estimation in TRP models using REIDAR is possible with F either a Weibull-, gamma- or exponential distribution, and $\lambda(\cdot)$ being either a constant, a *power law* intensity $\lambda(t) = abt^{b-1}$, or a *log linear* intensity $\lambda(t) = \exp\{a + bt\}$.

As an example we shall describe what we call the *Weibull-trend-renewal process*, WTRP(a, b, β), which in the notation of the previous section is given by TRP$(F_W(\cdot|1, \beta), \lambda_W(\cdot|a, b))$, where $F_W(t|\alpha, \beta) = 1 - \exp(-(\alpha t)^\beta)$ is the Weibull distribution with parameters α, β, and λ_W is the power law intensity given above. Note that the scale parameter α in the Weibull distribution is set to 1, since by (4) the scale is taken care of by the parameter a in λ_W.

Three special cases are easily identified: WTRP$(a, 1, \beta)$ = RP$(F_W(\cdot|a, \beta))$, WTRP$(a, b, 1)$= NHPP$(\lambda_W(\cdot|a, b))$, WTRP$(a, 1, 1)$= HPP(a).

Using (6) we find the following expression for the intensity function of the WTRP,

$$\gamma(t) = \beta a^\beta b t^{b-1}[t^b - T^b_{N(t-)}]^{\beta-1}$$

which when substituted into (1) leads to the *log* likelihood

$$\begin{aligned} l(a, b, \beta) &= N(\tau)\{\log\beta + \log b + \beta\log a\} \qquad (7) \\ &+ \sum_{i=1}^{N(\tau)} \left\{(b-1)\log T_i + (\beta - 1)\log[T_i^b - T_{i-1}^b] - a^\beta[T_i^b - T_{i-1}^b]^\beta\right\} \\ &- a^\beta[\tau^b - T^b_{N(\tau)}]^\beta \end{aligned}$$

The unknown parameters are estimated by maximum likelihood.

If *several* systems of the same kind are under observation, and if the corresponding failure processes are taken as independent and identically distributed, then the full log likelihood consists of a sum of terms (7) for each system.

However, in the case of several systems, even being of same type, they may well exhibit different failure rates due to various causes. For example, systems may be used under different environmental or operational conditions, like compressors or pumps installed on different oil platforms situated in different fields of the North Sea.

If the differing conditions can be quantified, then they may be included in the model as *observable* covariates, and treated with models like (2) and (5). For estimation in NHPP models with covariates, see e.g. Lawless (1987).

Unobserved heterogeneity is more difficult to handle, but can be incorporated into the WTRP process by allowing, e.g., the scale parameter a to vary from system to system. The common approach [Andersen et al. (1992)] is to model a by a gamma-distributed random variable with shape parameter γ and scale parameter δ, say. Thus what we shall call the GWTRP has four parameters, b, β, γ, δ, where b, β have the same meaning as for the WTRP. The likelihood function for each single system is now obtained by taking the expected value of the likelihood function (exponential of (7)), with respect to the gamma distributed a. It is not possible to do the integration explicitly, and the program REIDAR uses numerical integration

Model	Parameters in model	Log likelihood
GWTRP	b, β, γ, δ	-1172.1
GNHPP	$b, 1, \gamma, \delta$	-1172.3
GHPP	$1, 1, \gamma, \delta$	-1174.9
GRP	$1, \beta, \gamma, \delta$	-1174.8
HPP	$b = 1, \beta = 1, a =$ constant	-1178.9

Table 1: Maximum log likelihood values for various models for the Proschan (1963) airconditioner data.

here and computes the maximum likelihood estimates by a simplex method. Heggland (1994) studied heterogeneity in the inhomogeneous gamma process of Berman (1981).

As an example of applying the GWTRP model, we consider the airconditioner data studied by Proschan (1963) (see section 2). In order to determine the "best" model for the data, we have used REIDAR to compute maximum values of the log likelihood for the GWTRP and various submodels. For notation, note that a G in front of the earlier notation for NHPP etc., means a model where the scale parameter a is replaced by a gamma-distributed random variable with parameters γ and δ. The results are given in table 1.

For testing hypotheses of the parameters we can use the general result that *two times differences in log likelihood* are approximately chi-square distributed, with degrees of freedom (df) equal to the difference in number of parameters in the largest model and the null hypothesis model, when the null hypothesis holds.

Recall from section 2 that Proschan (1963) tested the null hypothesis of a common HPP model for all aircrafts, against the alternative of HPPs with intensities differing from plane to plane. In the WTRP model this translates to testing HPP vs. GHPP. Using table 1 we get that twice the difference in log likelihood is 7.84, which with df=1 gives a P-value of 0.005. This is not much different from what Proschan got (0.007). However, from table 1 it seems that the best model is an NHPP with heterogeneity, since testing GHPP vs. GNHPP gives a P-value of 0.022. Follmann and Goldberg (1988) studied Weibull renewal processes with gamma heterogeneity (i.e. GRP) and concluded for the Proschan data that a GHPP would be the best model, since the log likelihood increases by 0.1 only when going from GHPP to GRP.

7 Trend testing

For the proper operation of repairable systems it is important to detect possible changes in the pattern of failures occurring over time. For example, reliability growth corresponds to times between failures becoming longer as time goes (improving system), whereas ageing effects often lead to decreasing interfailure times (deteriorating system). Various types of nonmonotonic trend can also be present, e.g. *cyclic* trend or *bathtub shaped* trend.

A simple but informative way of checking for a possible trend in the pattern of failures is to study plots like figure 2. Departures from a straight line may indicate some kind of trend. In particular, a convex shape will be indicative of a deteriorating system, while a concave plot may indicate an improving system. In figure 2 there seems to be an increase in the failure intensity from about 1200 hrs.

By a *trend test* we shall mean a statistical test for the null hypothesis that the failure process is stationary, in some sense to be made precise, versus alternatives depending on the kind of trend one would like to detect. Here we shall consider the null hypotheses either "the process is an HPP" or "the process is an RP", whereas the alternative will for brevity of exposition be restricted to "monotone trend". Note that trend tests of "RP" versus "cyclic trend" already are mentioned in connection with Berman's inhomogeneous gamma process of section 5. Trend tests versus bathtub shaped trend are considered for the NHPP case in Kvaløy and Lindqvist (1998).

In parametric modeling of failure processes one may often represent trend by means of one or more of the parameters. For example in the WTRP(a, b, β) model, the parameter b represents trend and equals 1 for the case of no trend. Thus a likelihood ratio test of the null hypothesis $b = 1$ is in fact a trend test, with null hypothesis corresponding to a Weibull renewal process. An example of such a test is the likelihood ratio test for H_0 : GHPP vs. H_1 : GNHPP that we performed for the Proschan data in table 1.

Two classical trend tests will be given much attention in the following, For simplicity we shall always assume that a single system is under observation, with failure truncation after n failures.

The Laplace test. This is a test for the null hypothesis of "HPP" vs. the alternative of a monotonic trend. The test statistic is

$$U = \frac{\frac{1}{n-1}\sum_{i=1}^{n-1} T_i - \frac{1}{2}T_n}{T_n\sqrt{\frac{1}{12(n-1)}}}$$

which under the null hypothesis is approximately standard normally distributed. The null hypothesis is rejected for too small or too large values of U. Moreover, $U > 0$ (resp. $U < 0$) indicate deteriorating (respectively improving) system.

The Military Handbook test. The null hypothesis is "HPP" and the test statistic is $Z = 2\sum_{i=1}^{n-1}\log(T_n/T_i)$, which under the null hypothesis is approximately χ^2-distributed with $2(n-1)$ degrees of freedom. The null hypotheses is rejected for small or large values of Z, low values indicating deteriorating system and large values indicating improving system.

Many users seem to forget the fact that the Laplace- and Military Handbook tests really are tests for the null hypothesis that the data come from an HPP. Thus *rejection* of the null hypothesis in strict terms means just that the process is *not an HPP* – but it could still in principle be an RP and thus still have no trend. That such false rejections may occur with considerable probability will be seen from simulations reported below.

There exist, however, tests which are constructed to test the null hypothesis of an RP vs. a monotone trend, e.g. the *Mann test* and the *Lewis-Robinson test* [Ascher and Feingold (1984)]. The latter test is defined and studied below.

The Lewis-Robinson test. This is a modification of the Laplace test, with test statistic $U_{LR} = U/\widehat{CV}(X)$. Here U is the Laplace test statistic, while $\widehat{CV}(X)$ is an estimate of the *coefficient of variation* of the X_i. Note that the coefficient of variation of the exponential distribution is 1. The reason for introducing the denominator in U_{LR} is to account for non-exponential distributions of the interfailure times. In fact, the denominator assures that U_{LR} is approximately standard normally distributed whenever the data come from an RP.

β	Laplace	MilHbk	Lew-Rob
0.2	.591	.738	.018
0.5	.289	.333	.043
0.8	.120	.110	.051
1.0	.046	.049	.046
1.2	.022	.023	.053
1.5	.003	.003	.055
2.0	.000	.000	.053

Table 2: Simulated rejection probabilities for $n = 20$ failures of an RP with Weibull$(1, \beta)$ interfailure times.

Table 2 shows the results of simulations reported in Lindqvist et al. (1994), performed in order to show the deficiency of the Laplace- and Military Handbook tests as tests for the null hypothesis of RP. An RPs with Weibull$(1, \beta)$-distribution and varying shape parameter β was simulated. The number n of failures per case was set to 20, and the nominal significance level was $\alpha = 0.05$. Each estimated value is based on 2000 simulations. Table 2 shows the effective significance level (i.e. relative number of rejections of the null hypothesis made in the 2000 simulations).

The effective significance level of the Lewis Robinson test correspond very well to the nominal level, which should be expected. For the two classical tests, however, the effective level is 0.05 only when $\beta = 1$, which corresponds to an HPP. The tests reject much too often for $\beta < 1$ (improving system), but are surprisingly conservative, i.e. have too low effective significance level, when $\beta > 1$ (deteriorating system).

Lindqvist et al. (1994) also compare the power of the tests against alternatives with monotonic trend. As is well known [Ascher and Feingold (1984)] the Laplace test and the Military Handbook test are optimal tests for alternatives of a NHPP with log linear intensity and power law intensity functions, respectively. The general conclusion regarding testing vs. NHPP trend is that the two classical tests outperform the Lewis-Robinson test for NHPP alternatives with log linear and power law intensities.

In a study of the properties of the three tests with respect to rejection power against non-NHPP cases with trend, it is concluded, however, that the Laplace and Military Handbook tests should be used only with extreme care. Tests like the Lewis Robinson test and the Mann test are to be preferred here, due to the fact that they behave correctly under the null hypothesis of an RP. Alternative tests based on Monte Carlo simulation are suggested in the next section.

---

8 Monte Carlo trend tests

As mentioned above, the Lewis-Robinson test is essentially a modified Laplace test, with the aim of attaining the correct significance level under the null hypothesis of an RP. We shall in this section see that corresponding modifications can be done with any trend test by the use of Monte Carlo simulations. The work described in this section will appear in more detail in Elvebakk (1998). A general reference on Monte Carlo tests is Ripley (1987).

Consider again for simplicity a single failure truncated system, observed until n failures have occurred. Thus the observations are the failure times $T_1, T_2, \ldots, T_n$, or equivalently the interfailure times $X_1, X_2, \ldots, X_n$. The null hypothesis H_0 is that the observations come from an RP.

Suppose now that $W \equiv W(X_1, \ldots, X_n)$ is a test statistic, for which H_0 is rejected for large values. We may, for example, think of W as $|U|$, where U is the Laplace test statistic. The problem in practice is to find the critical value for a given significance level. We shall describe two types of Monte Carlo tests for this problem.

Permutation test. Let F be the cumulative distribution function of the interfailure times under the null hypothesis of RP. It is well known that the order statistic $S \equiv (X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ is statistically sufficient for F. The permutation test with given significance level α is defined as a conditional test given S, which rejects H_0 if

$$Pr(W \geq w_{obs} | S = s) \leq \alpha \tag{8}$$

where w_{obs} is the observed value of W and s is the observed value of S. By sufficiency, the distribution of W given $S = s$ is independent of the unknown distribution F, and hence the left hand side of (8) can in principle be found. In practice the permutation test can be performed as follows:

1. Record interfailure times $x_1, x_2, \ldots, x_n$.
2. Compute $w_{obs} = W(x_1, x_2, \ldots, x_n)$
3. Compute $W(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)})$ for a large number K of randomly chosen permutations π of $x_1, x_2, \ldots, x_n$. Order the computed values of W as $w_1 \leq w_2 \leq \ldots \leq w_K$.
4. Reject H_0 if $w_{obs} \geq w_{[(1-\alpha)(K+1)]}$ (where $[a]$ denotes the integer part of a).

Bootstrap test. As above, let F be the distribution of the RP under H_0. If F were known, then we would reject H_0 with significance level α if $P_F(W \geq w_{obs}) \leq \alpha$. In practice, however, the distribution F is unknown. The idea of the bootstrap test is to replace F by the empirical distribution $\hat{F}$ giving mass $1/n$ to each of the observations $x_1, \ldots, x_n$, and then reject H_0 if $P_{\hat{F}}(W \geq w_{obs}) \leq \alpha$. It can be shown that, when certain conditions hold (which will not be discussed here) we have an asymptotic α level test. The practical recipe of the bootstrap test turns out to be as for the permutation test, but with step 3 replaced by 3*: Compute $W(x_1^*, x_2^*, \ldots, x_n^*)$ for a large number K of random, ordered samples $(x_1^*, x_2^*, \ldots, x_n^*)$ drawn *with replacement* from $x_1, x_2, \ldots, x_n$. Order the computed values of W as $w_1 \leq w_2 \leq \ldots \leq w_K$.

It is interesting to note that the only difference between the permutation test and the bootstrap test is that the bootstrap test draws *with* replacement, while the permutation test draws *without* replacement from the observed set of interfailure times.

Note also that the permutation tests are exact conditional tests, which in principle achieve the correct significance level, while the bootstrap tests only have approximately correct level, due to the approximation of F by $\hat{F}$. In practice, however, the level of the permutation test also becomes approximate, due to the limited number K of permutations considered.

Comprehensive simulation studies of the Monte Carlo tests described above are found in Kvaløy (1994) and Elvebakk (1998). We merely extract a few numbers from their studies in order to illustrate the main features of the Monte Carlo tests.

n	b	MilHbk	MilH Boot	MilH Permut	Lew-Rob
	0.50	0.976	0.967	0.967	0.867
	0.75	0.409	0.393	0.398	0.281
35	1.00	0.045	0.050	0.049	0.049
	1.50	0.637	0.602	0.576	0.468
	2.00	0.983	0.973	0.962	0.880
15	0.50	0.750	0.677	0.693	0.459
	1.50	0.310	0.268	0.236	0.228
5	2.50	0.396	0.121	0.138	0.083

Table 3: Simulated rejection probabilities for NHPP alternatives with $\lambda(t) = bt^{b-1}$.

β	b	MilHbk	MilH Boot	MilH Permut	Lew-Rob
	0.5	0.780	0.908	0.912	0.766
1.5	1.0	0.006	0.051	0.049	0.047
	1.5	0.299	0.496	0.450	0.375
	2.0	0.842	0.899	0.851	0.721
	0.5	0.749	0.505	0.528	0.266
	1.0	0.136	0.053	0.052	0.047
0.75	1.5	0.299	0.172	0.145	0.125
	2.0	0.621	0.415	0.355	0.291
	2.5	0.844	0.634	0.551	0.443

Table 4: Simulated rejection probabilities for $n = 15$ failures of a WTRP$(1, b, \beta)$.

The first question one may ask is whether much is lost in power for NHPP alternatives when going from the classical Laplace- and Military Handbook tests to the corresponding Monte Carlo tests. Table 3 gives estimated rejection probabilities when data are simulated from an NHPP with $\lambda(t) = bt^{b-1}$, with varying n and b. The tests compared are the ordinary Military Handbook test, its two Monte Carlo versions, and the Lewis Robinson test.

We conclude that rather little power seems to be lost in this case by performing Monte Carlo tests, but that more is lost for small sample sizes. Moreover, there is only a small difference between the power of the bootstrap- and the permutation tests.

It is furthermore seen that the Monte Carlo tests are always better than the Lewis-Robinson test. But this can be explained by the fact that the Military Handbook test is optimal for the NHPP alternative considered here. In fact, simulations of Elvebakk (1998) show small differences between the power of the Lewis-Robinson test and the Monte Carlo tests based on the Laplace test, but with the Monte Carlo tests always slightly better.

Recall from the previous section that the ordinary Laplace and Military Handbook tests may behave badly in non-Poisson cases. Table 4 shows estimated rejection probabilities for alternatives given by WTRP$(1, b, \beta)$, with $\beta = 1.5$ and 0.75, and varying b. As before we can conclude that the Military Handbook test does not achieve the correct significance level (see $b = 1.0$). On the other hand, the results for the Monte Carlo tests based on the Military Handbook are satisfactory.

The general conclusion of Elvebakk (1998) is that Monte Carlo trend tests are favorable to the Lewis-Robinson test and the Mann test in any case considered. As regards the Laplace test and the Military Handbook test, the corresponding Monte Carlo tests have lower power than the original tests for NHPP alternatives, but the difference is small for large sample sizes. This fact, together with the fact that the Laplace and Military Handbook tests are invalid in non-Poisson cases, leads to a recommendation of the Monte Carlo tests as an overall conclusion.

---

9 Concluding remarks and topics for further study

Goodness of fit tests for NHPP models. The literature contains several suggestions of how to test the adequacy of parametric NHPP models. Particularly much studied are NHPP with power law intensity. Recent references are Baker (1996) and Gaudoin (1997). Less studied is the case of NHPP with log linear intensity, apparently due to mathematical intractability. Lindqvist and Elvebakk (1998) study a general setup for using conditional tests given the minimal sufficient statistic under the null hypothesis.

Nonparametric and semiparametric inference. Boswell (1966) considered nonparametric estimation in NHPP($\lambda(\cdot)$) under the assumption of a monotone $\lambda(\cdot)$. Heggland (1994) considered the semiparametric case of estimating a monotone $\lambda(\cdot)$ nonparametrically in a TRP($F, \lambda(\cdot)$), with F being a parametric Weibull($1, \beta$) distribution. As mentioned in section 6, there are several possibilities for non- and semiparametric inference in the TRP model. Kernel estimators and other smoothing techniques applied to the $\lambda(\cdot)$ or the F are of particular interest.

Complex modeling for repairable systems. The TRP model and in particular the NHPP model can in principle easily be extended by assuming that the intensity function $\lambda(\cdot)$ itself is stochastic. For the NHPP, this is the well known *doubly stochastic* Poisson process.

For repairable systems one could also imagine using multiple time scales, e.g. *time* and *number of cycles*. In practice there may also be several competing failure modes. It would be nice to see how this can be incorporated in the common models for repairable systems.

For many applications it is natural to model the interfailure times in more detail. For example this could be done by Markov models, which would lead to phase type distributions for times between failures. Hokstad and Frøvig (1996) suggest simple models of this kind.

References

1. Aalen, O. O. (1975) Statistical inference for a family of counting processes. PhD thesis, University of California, Berkeley.

2. Andersen, P., Borgan, O., Gill, R., Keiding, N. (1992). *Statistical Models Based on Counting Processes*. Springer.

3. Ascher, H. and Feingold, H. (1984). *Repairable Systems – Modeling, inference, misconceptions and their causes.* Marcel Dekker, New York.

4. Baker, R. D. (1996). Some new tests for the power law process. *Technometrics*, **38**, 256-265.

5. Barlow, R. E., Marshall, A. W. and Proschan, F. (1963). Properties of probability distributions with monotone hazard rate. *Annals of Mathematical Statistics*, **34**, 375-389.

6. Berman, M. (1981). Inhomogeneous and modulated gamma processes. *Biometrika*, **68**, 143-152.

7. Berman, M. and Turner, T. R. (1992). Approximate point process likelihoods with GLIM. *Applied Statistics*, **41**, 31-38.

8. Boswell, M. T. (1966). Estimating and testing trend in a stochastic process of Poisson type. *Annals of Mathematical Statistics*, **37**, 1564-1573.

9. Brown, M. and Proschan, F. (1983). Imperfect repair, *Journal of Applied Probability*, **20**, 851–859.

10. Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *Journal of the Royal Statistical Society, Series B*, **17**, 129-164.

11. Cox, D. R. (1972). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes,* Ed. P.A. Lewis, pp. 55-66. Wiley, New York.

12. Cox, D. R. and Lewis, P. W. (1966). *The Statistical Analysis of Series of Events.* Methuen, London.

13. Crowder, M. J., Kimber, A. C., Smith, R. L. and Sweeting, T. J. (1991). *Statistical Analysis of Reliability Data.* Chapman & Hall, Great Britain.

14. Dorado, C., Hollander, M. and Sethuraman, J. (1997). Nonparametric estimation for a general repair model. *Annals of Statistics*, **25**, 1140-1160.

15. Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

16. Elvebakk, G. (1998). Resampling methods for analysis of repairable systems. Thesis work. Norwegian University of Science and Technology. (In preparation).

17. Follmann, D. A. and Goldberg, M. S. (1988). Distinguishing heterogeneity from decreasing hazard rates. *Technometrics*, **30**, 389-396.

18. Gaudoin, O. (1997). CPIT goodness-of-fit tests for the power-law process. *Communications in Statistics. Theory and Methods.* To appear.

19. Heggland, K. (1994). Analysis of repairable systems: Modeling of heterogeneity, especially in trend-renewal processes. (In Norwegian). Project work. Norwegian Institute of Technology.

20. Heggland, K. (1994). Non-parametric estimation in trend-renewal processes. Diploma thesis. Norwegian Institute of Technology.

21. Hokstad, P. and Frøvig, A. T. (1996). The modelling of degraded and critical failures for components with dormant failures. *Reliability Engineering and System Safety*, **51**, 189-199.

22. Høyland, A. and Rausand, M. (1994). *System Reliability Theory. Models and Statistical Methods.* Wiley, New York.

23. Kijima, M. (1989). Some results for repairable systems with general repair. *Journal of Applied Probability*, **26**, 89–102.

24. Kvaløy, J. T. (1994). Analysis of repairable systems: Trend testing. Diploma thesis. Norwegian Institute of Technology.

25. Kvaløy, J. T. and Lindqvist, B. H. (1998). TTT-based tests for trend in repairable systems data. *Reliability Engineering and System Safety.* To appear.

26. Lawless, J. F. (1987). Regression methods for Poisson process data. *Journal of American Statistical Association*, **82**, 808-815.

27. Lawless, J. F. and Thiagarajah, K. (1996). A point-process model incorporating renewals and time trends, with application to repairable systems. *Technometrics*, **38**, 131-138.

28. Lindqvist, B. H. (1993). The trend-renewal process, a useful model for repairable systems. Society of Reliability Engineers, Scandinavian Chapter, Annual Conference, Malmö, Sweden.

29. Lindqvist, B. H., Kjønstad, G. A. and Meland, N. (1994). Testing for trend in repairable systems data. *Proceedings of ESREL'94.* La Baule, France.

30. Lindqvist, B. H. and Elvebakk, G. (1998). Exact tests for deviation from nonhomogeneous Poisson processes. In preparation.

31. Proschan, F. (1963). Theoretical explanation of observed decreasing failure rates. *Technometrics*, **5**, 375-383.

32. Ripley, B. D. (1997). *Stochastic Simulation.* Wiley, New York.

33. Thompson, W. A. Jr. (1988). *Point Process Models with Applications to Safety and Reliability.* Chapman and Hall, London.

34. Whitaker, L. R. and Samaniego, F. J. (1989). Estimating the reliability of systems subject to imperfect repair. *Journal of American Statistical Association*, **84**, 301-309.