Tests for trend in more than one repairable system.[1]

Jan Terje Kvaløy
Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim

ABSTRACT: If failure time data from several systems of the same kind are available it is often desirable to analyze these data simultaneously. Two concepts for doing simultaneous trend testing in data from more than one system are presented. These two concepts are compared in a simulation study, and a general strategy for trend testing in data from several systems that draw benefits from both concepts is proposed.

1 INTRODUCTION

For repairable systems it is important to detect possible changes in the pattern of failures, which for instance can be caused by various aging effects or reliability growth. We say that there is a trend in the pattern of failures if the inter-failure times tend to alter in some systematic way, which means that the inter-failure times are not identically distributed. By using statistical trend tests it is possible to decide whether such an alteration is statistically significant or not.

If we have failure time data from several repairable systems, we can either test for trend in each separate system, or do a simultaneous analysis. Doing a simultaneous analysis of data from more than one system requires that all the systems are judged to be sufficiently equal for a simultaneous analysis to make sense. This decision must primarily be based on knowledge of the systems and their operating conditions. Graphical techniques for exploring the pattern of failures in each system can also be useful in this stage of the analysis.

If doing a simultaneous analysis can be justified, this will in general be much more powerful than analyzing each system separately. Changes in the pattern of failures which is impossible to reveal when each system is analyzed separately may become possible to detect as statistical significant changes when a simultaneous analysis is performed.

Two strategies for doing simultaneous trend analysis in data from more than one repairable system are reviewed and compared. Attention is restricted to nonhomogeneous Poisson process (NHPP) models, i.e. minimal repair is assumed. A number of statistical trend tests based on Poisson process models of a single system exists (Ascher & Feingold 1984). If we have more than one system and we want to do a simultaneous analysis, testing for trend can be done for example by combining the individual test statistics, or by using a one-system trend tests on the total time on test (Barlow & Davis 1977) transformed observations. The difference is that the last approach rely on the stronger assumption of identical intensity functions for each system, while the first approach allows for heterogeneities from system to system. Relying on a stronger assumption, the total time on test approach leads to more powerful tests if the assumption holds, but can be quite misleading in the presence of heterogeneities. On the other hand, with no or minimal heterogeneities these tests can be favorable to use due to their good power properties.

For both strategies the tests can be extended to more general models than the NHPP by using resampling techniques (Elvebakk 1998).

[1]To appear in proceedings of ESREL 1998.
E–mail: `jtk@stat.ntnu.no`

The total time on test (TTT) approach for testing for trend in data from more than one repairable systems was presented by Kvaløy & Lindqvist (1998). The main issue in that paper was to show how tests for trend in repairable systems can be based on the TTT transformed observations, and to derive the Anderson Darling test for trend which is a new test arising from that approach. In the simulation study in Kvaløy & Lindqvist (1998) the main focus was on comparing the Anderson-Darling test for trend with other trend tests in the single process case. Testing for trend in more than one system was more briefly covered, though some simulations which explored the behavior of the various tests in a few important situations was given. The purpose of the present paper is to complement the study of trend testing in data from more than one system initiated in Kvaløy & Lindqvist (1998), and to suggest a general strategy for trend testing in data from more than one repairable system.

The two well known and popular trend tests called the Laplace test and the Military Handbook test are studied. These tests have both desirable theoretical properties (Cohen & Sackrowitz 1993) and have shown good performance compared to other tests in simulation studies (Bain et al. 1985). In the simulation study in Kvaløy & Lindqvist (1998) the Anderson-Darling test for trend mentioned above outperformed these tests in cases with nonmonotonic trend, but since only monotonic trend is considered in the present paper the Anderson-Darling test for trend is not included. In cases with monotonic trend the Anderson-Darling test for trend behave very similar to the Laplace test.

An important issue in trend testing in data from several systems is the possible presence of heterogeneities between the systems, either observed or unobserved heterogeneities. Since the difference between the two approaches for trend testing in data from several repairable systems is that the TTT-based approach does not allow scale heterogeneities, it is important to be able to detect such heterogeneities. Attention is restricted to unobserved heterogeneities, and included in the simulations is a standard test for scale heterogeneities.

The general strategy suggested for trend testing in data from more than one system is to first use the heterogeneity test to decide if there are heterogeneities in the data, and then either use a combined test if the heterogeneity test detects statistically significant heterogeneities or otherwise use a TTT-based test. This is apparently a straightforward approach, but some adjustments are needed to get a test with desirable properties under the null hypothesis.

2 MODEL AND TESTS

2.1 *Notation and terminology*

Throughout it is assumed that we have data from $m \geq 1$ independent systems. The ith system is observed in the time interval $(a_i, b_i]$ with n_i independent failures occurring at times T_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. The total number of failures observed is $N = \sum_{i=1}^{m} n_i$. The systems are assumed to either all be time truncated (all intervals $(a_i, b_i]$ fixed in advance and n_i random) or all be failure truncated (all a_i and n_i fixed in advance and b_i random).

If for a given system all the interfailure times $T_{ij} - T_{i,j-1}$, are not equally distributed we say there is a trend in the pattern of failures for this system. If the expected length of the interfailure times is monotonically increasing or decreasing with time, corresponding to an improving or a deteriorating system, there is a monotone trend (an increasing or a decreasing trend), otherwise the trend is nonmonotone.

2.2 *The nonhomogeneous Poisson process*

Let $N(t)$ be the number of failures occurring in one process in the time interval $[0, t]$. The counting process $\{N(t), t \geq 0\}$ is called a nonhomogeneous Poisson process (NHPP) with intensity function $\lambda(t)$ if (1) $N(0) = 0$, (2) the number of failures in disjoint time intervals are stochastically independent, (3) $P(N(t + \Delta t) - N(t) = 1) = \lambda(t)\Delta t + o(\Delta t)$ as $\Delta t \to 0$, and (4) $P(N(t+\Delta t)-N(t) \geq 2) = o(\Delta t)$ as $\Delta t \to 0$.

The NHPP model implies assuming minimal repair. It is also well known that the intensity function $\lambda(t)$ coincides with the ROCOF (Rate of Occurrence of Failures) associated with the repairable system (Ascher & Feingold 1984).

We consider two different parameterizations of $\lambda(t)$, the power law intensity:

$$\lambda(t) = \alpha\beta t^{\beta-1}, \quad \alpha, \beta > 0, \quad t \geq 0 \qquad (1)$$

and the log-linear intensity:

$$\lambda(t) = \alpha e^{\beta t}, \quad \alpha > 0, \; -\infty < \beta < \infty, \quad t \geq 0 \qquad (2)$$

In the above intensity functions α is called a scale parameter and β is called a shape parameter. An NHPP with constant intensity, i.e. no trend, is called a homogeneous Poisson process (HPP).

2.3 *TTT transformation for repairable systems*

The total time on test (TTT) transformation for repairable systems data was introduced by Barlow & Davis (1977). Let S_k denote the kth arrival time in the process obtained by superposing the observations from each system. That is, S_k is a failure time in one of the systems and $0 < S_1 \leq S_2 \leq \ldots \leq S_N \leq S$, where $S = \max_{i \in \{1,\ldots,m\}} b_i$. Let $p(u)$ denote the number of systems under observation at time u. Then $\mathcal{T}(t) = \int_0^t p(u)du$ is the total time on test from time 0 to time t, and the scaled total time on test statistic is:

$$\frac{\mathcal{T}(S_k)}{\mathcal{T}(S)} = \frac{\int_0^{S_k} p(u)du}{\int_0^S p(u)du} \tag{3}$$

2.4 *Trend tests*

The combined and the TTT-based versions of the trend tests mentioned in the introduction are presented below. The test statistic for the combined tests are combinations of single system test statistics, while the test statistics for the TTT-based tests is using the total time on test statistic (3). Derivation of distributional results for the TTT-based tests and other details are given in Kvaløy & Lindqvist (1998). The combined tests are tests of the null hypothesis H_0: HPPs with possibly different intensities, while the TTT-based tests are tests of H_0: HPPs with identical intensities. The alternative hypothesis is in general "not H_0", but the Laplace test is optimal for the alternative of an NHPP with log-linear intensity and the Military Handbook test is optimal for the alternative of an NHPP with power law intensity.

2.4.1 *The Laplace test*

Let $\hat{n}_i = n_i$ if the ith process is time truncated, and $\hat{n}_i = n_i - 1$ if the ith process is failure truncated. Then under the null hypothesis of HPPs with possibly different intensities:

$$L_C = \frac{\sum_{i=1}^m \sum_{j=1}^{\hat{n}_i} T_{ij} - \sum_{i=1}^m \frac{1}{2}\hat{n}_i(b_i + a_i)}{\sqrt{\frac{1}{12}\sum_{i=1}^m \hat{n}_i(b_i - a_i)^2}} \tag{4}$$

is approximately standard normally distributed. We call the test based on (4) the combined Laplace test.

Let $\hat{N} = N$ if the processes are time truncated, and $\hat{N} = N - 1$ if the processes are failure truncated. Then if we assume that all HPPs have identical intensity the test statistic:

$$L_T = \frac{\sum_{k=1}^{\hat{N}} \frac{\mathcal{T}(S_k)}{\mathcal{T}(S)} - \frac{1}{2}\hat{N}}{\sqrt{\frac{1}{12}\hat{N}}} \tag{5}$$

is approximately standard normally distributed. We call the test based on this statistic the TTT-based Laplace test.

The value of L_C or L_T indicates the direction of the trend. If L_C or L_T is negative this indicates a decreasing trend, while a positive L_C or L_T indicates an increasing trend.

2.4.2 *The Military Handbook test*

The test statistic for the combined Military Handbook test is:

$$M_C = 2\sum_{i=1}^{m}\sum_{j=1}^{\hat{n}_i} \ln\left(\frac{b_i - a_i}{T_{ij} - a_i}\right) \tag{6}$$

which is (exactly) χ^2_{2q}-distributed, where $q = \sum_{i=1}^m \hat{n}_i$, under the null hypothesis of HPPs with possibly different intensities.

Under the assumption of HPPs with identical intensities, the test statistic:

$$M_T = 2\sum_{k=1}^{\hat{N}} \ln\left(\frac{\mathcal{T}(S)}{\mathcal{T}(S_k)}\right) \tag{7}$$

is $\chi^2_{2\hat{N}}$-distributed, and we call the test based on this statistic the TTT-based Military Handbook test.

If M_C or M_T is large this indicates a decreasing trend, while a small value of M_C or M_T indicates an increasing trend.

Notice that both (4) and (5), as well as (6) and (7), are equal in the case of only one process, and also in the case of several processes if the processes are time truncated and all observation intervals $(a_i, b_i]$ are equal.

2.5 *Test for heterogeneities*

Even if the repairable systems we are considering apparently are identical, they will often be operating under different conditions. This may alter the pattern of failures from system to system. Moreover it is not obvious that presumably identical systems really are identical, differences in manufacturing etc. may affect the failure intensities. Such differences in the failure intensity are called heterogeneities, and can be either observed or not.

Typically heterogeneities are assumed to lead to differences in the scale parameter from system to system. Being able to detect such heterogeneities is important if we want to use the TTT-based tests for trend since these tests do not allow for such heterogeneities under the null hypothesis. This implies that if the assumption of equal scale parameters is violated, the TTT-based tests may reject the null hypothesis even if there is no trend.

A nice reference on estimation of the intensity function and testing for heterogeneities in the case of data from several NHPPs with either observed, unobserved or both observed and unobserved heterogeneities is Lawless (1987). In the present paper only unobserved heterogeneities are considered.

The standard way to model unobserved scale heterogeneities within the NHPP framework is to let $\lambda_0(t)$ be a baseline intensity function which is common for all systems, and let the intensity for the ith process be $\lambda_i(t) = z_i\lambda_0(t)$. Here z_i is an unobserved random effect which typically is modeled to be gamma-distributed with expectation 1 and variance η. Thus, η is a measure of the degree of heterogeneity, and likelihood ratio tests for heterogeneity based on this quantity can be derived (Lawless 1987).

A standard model is to use a power law parameterization for $\lambda_0(t)$. In the likelihood function for the problem the unobserved z_i-values must be integrated out, leading to the likelihood function (Lawless 1987):

$$L(\alpha, \beta, \eta) =$$

$$\prod_{i=1}^{m} \frac{1}{\Gamma(\frac{1}{\eta})\eta^{1/\eta}} \frac{\Gamma(\frac{1}{\eta} + n_i)}{(\frac{1}{\eta} + \alpha(b_i^{\beta} - a_i^{\beta}))^{1/\eta + n_i}} \prod_{j=1}^{n_i} \alpha\beta T_{ij}^{\beta-1}$$

The likelihood ratio test of H_0: $\eta = 0$ (no heterogeneities) can be based on the likelihood ratio statistic $R = 2(\ln L(\hat{\alpha}, \hat{\beta}, \hat{\eta}) - \ln L(\hat{\alpha_0}, \hat{\beta_0}, 0))$ where the parameter estimates are found respectively by maximizing the full likelihood function and the likelihood function under the null hypothesis. But since $\eta = 0$ is not in the interior of the parameter space ($\eta < 0$ not allowed) R does not have the usual asymptotic χ_1^2-distribution. In fact asymptotically R is χ_1^2-distributed with probability 0.5 and has a probability mass of 0.5 at $R = 0$. See Lawless (1987) for details. This means that on a 5% significance level, H_0 is rejected if $R \geq 2.706$ which is the 10% quantile in the χ_1^2-distribution.

2.6 *Two step test*

In cases with no trend but heterogeneities between the systems the TTT-based tests may reject the null hypothesis too often since these tests, as opposed to the combined tests, do not allow such heterogeneities under the null hypothesis. On the other hand, in cases with no heterogeneities but trend, the TTT-based tests are much more powerful than the combined tests since these tests make the stronger assumption of equal intensities under the null hypothesis.

The following algorithm, named the two step test, tries to combine the best properties of the TTT-based and the combined tests: 1. Apply the heterogeneity test. 2. If the heterogeneity test rejects then use a combined test, otherwise use a TTT-based test.

The significance level of this two step test under the null hypothesis of no trend (but possibly heterogeneities) can be calculated as follows. Let H, C and T respectively be the event that the heterogeneity test, the combined test and the TTT-based Military Handbook test rejects the null hypothesis, and let $\bar{H}$, $\bar{C}$ and $\bar{T}$ be the events that these tests do not reject the null hypothesis. Then under the hypothesis of no trend the two step test procedure rejects this null hypothesis with probability: $P(H \cap C \cup \bar{H} \cap T) = P(H \cap C) + P(\bar{H} \cap T) = P(C|H)P(H) + P(T|\bar{H})P(\bar{H}) = P(C) \cdot P(H) + P(T|\bar{H})P(\bar{H})$. Intuitively $P(T|\bar{H}) \leq P(T)$ in the no trend case, but $P(T)$ will be larger than the significance level of the test in the presence of heterogeneities. To compensate for this, the TTT-based test could be evaluated on a lower significance level than the overall significance level we want the two step test to have. The two step test can also be made more conservative by evaluating the heterogeneity test on a higher significance level than the overall significance level wanted. By reducing $P(\bar{H})$ the contribution from a possibly too high $P(T|\bar{H})$ is weighted down.

The appropriate significance levels of the heterogeneity test and the TTT-based test can be found by simulations. These levels may vary according to the number of systems and the number of observations in each system. In the simulation study the heterogeneity test is evaluated on a 15% significance level and the TTT-based test on a 2.5% significance level. These values was found by preliminary simulations, and leads to a test with desirable significance level properties in a wide range of situations. The combined Laplace test is chosen as combined test and the TTT-based Military Handbook test is chosen as TTT-based test.

3 SIMULATION STUDY

By simulating the rejection probability of the different tests in various situations some insight is gained into the properties of the various tests for trend in data from more than one system.

The rejection probabilities are estimated by simulating 10000 data sets with the same setup and recording the relative number of rejections of each test. In the curves presented the rejection probability is simulated in a number of points, and straight lines are drawn between the points. Let $\hat{p}$ denote the estimated rejection probability. Then the standard deviation of $\hat{p}$ is $\sqrt{\hat{p}(1-\hat{p})/10000} \leq 0.005$. All simulations are done in C. The significance level has been set to 5%, and only simulations of failure truncated processes are reported.

3.1 *Level properties*

If for all systems the data come from HPPs with identical intensities, all tests, except the two step test, will, exactly or asymptotically, have the correct 5% significance level irrespectively of observation intervals or number of processes. The Military Handbook test is exact and the Laplace test achieves the correct significance level even for very small sample sizes. For small and moderate sample sizes the heterogeneity test has an actual significance level of less than 5%, but achieves the correct significance level for larger samples. In cases with no heterogeneities the two step test has a significance level slightly above the 2.5% significance level chosen for the TTT-based test in step 2 of the test. In other words, the two step test is more conservative than necessary in such cases.

Two illustrative examples of what can happen when data for different systems come from HPPs with different intensities are given in Figure 1. In the left plot 10 processes all observed from time 0 and until 10 failures have occurred are simulated. For each of the 10 processes the intensity is $\lambda_i = z_i$, $i = 1, \ldots, 10$, where z_i is gamma-distributed with expectation 1 and variance η. The rejection probability is plotted as a function of $1/\eta$. On the right plot in Figure 1 the number of processes is varying from 2 to 20 and all processes are observed from time 0 and until 10 failures have occurred. In this case the variance in the heterogeneity distribution equals $\frac{1}{5}$ in all cases.

As expected the combined tests remain exactly on the 5% significance level and the two step test has a significance level of approximately 5%. The TTT-based tests on the other hand have increasing rejection probability as the heterogeneity or number of processes increases. This is not surprising as the TTT-based tests are constructed for the null hypothesis of identical intensities, which implies that in the presence of heterogeneities these tests will as test for trend have an actual significance level of more than 5%. The TTT-based Military Handbook test is least affected by the heterogeneities. This motivates why the TTT-based Military Handbook test is used in the two step test. In practically all cases the heterogeneity test has a rejection probability higher than the TTT-based Laplace test, and in all cases a much higher rejection probability than the TTT-based Military Handbook test.

3.2 *Power properties, no heterogeneities*

The TTT-based tests have the serious drawback of being unable to distinguish between heterogeneities and trend, but the simulations displayed in Figure 2 show that the TTT-based tests have superior power properties in many situations. No heterogeneities are present, and the left plot show rejection probabilities as a function of β in the intensity function $\beta t^{\beta-1}$ when 10 partly overlapping processes with 10 observed failures are simulated. The first process is observed from time 0, the second process from the time when in expectation 4 failures have occurred in the first process, the third from the time when in expectation 4 failures have occurred in the second process and so on. The right plot in Figure 2 shows rejection probabilities as a function of total number of failures when data are simulated using the intensity function $e^{0.5t}$ and four processes are observed from time 0 and until 2 failures have occurred, and one process is observed from time 0 and until n_1 failures have occurred where $n_1 \in \{2, \ldots, 50\}$

The plots in Figure 2 demonstrate the power of the TTT-based approach for trend testing in more than one system if the assumption of no heterogeneities is valid. Moreover the plots also show that the two step test is able to catch most of the power of the TTT-based tests. In such cases with no heterogeneities the two step test is

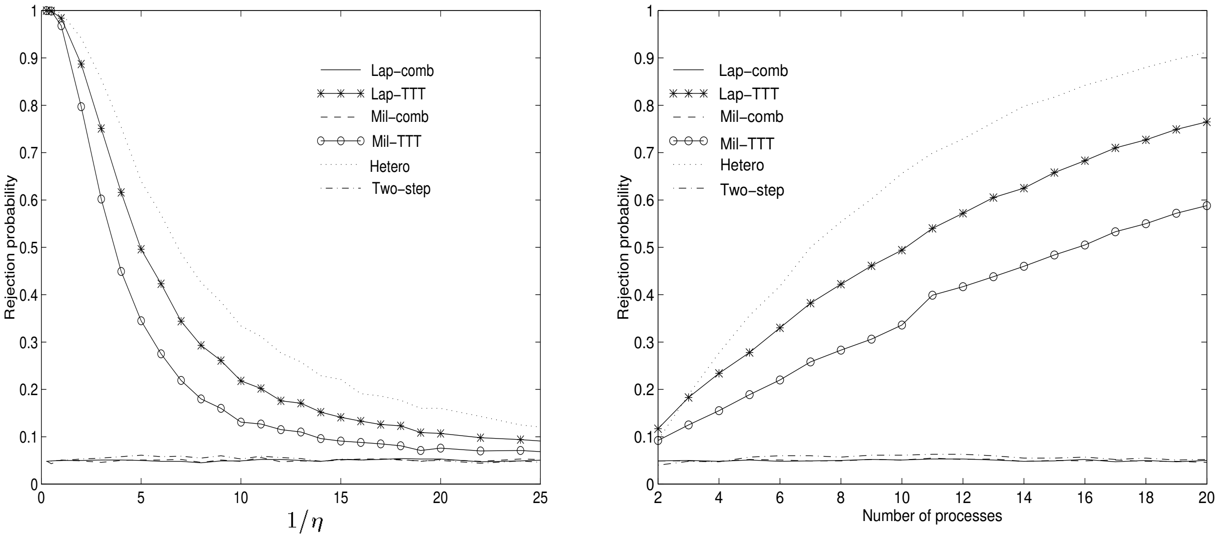

Figure 1: Simulations of HPPs with respectively varying degree of heterogeneity and varying number of processes. All processes are observed from time 0 and until 10 failures have occurred. In the left plot 10 processes are simulated and the rejection probability is plotted as a function of $1/\eta$. In the right plot $\eta = \frac{1}{5}$ and the rejection probability is plotted as a function of number of processes.

essentially a TTT-based Military Handbook test evaluated on a 2.5% significance level. The differences in power between the TTT-based tests and the combined tests are in general large when the observation intervals are unequal. The bigger the difference in observation intervals gets, the larger is the difference in power. If the processes are time truncated and the observation intervals are equal, then the TTT-based and the combined versions of the Laplace and the Military Handbook tests are equal. Otherwise the TTT-based tests have higher rejection probability since they under the null hypothesis make the stronger assumption of identical intensities.

3.3 *Power properties, heterogeneity*

To check the ability of the heterogeneity test to detect heterogeneities in cases with trend and to study the behavior of the other tests in cases with both trend and scale heterogeneities, the two simulation series displayed in Figure 3 are done. In both cases 10 processes are observed from time 0 and until 10 failures have occurred, and the variance in the heterogeneity distribution is $\frac{1}{10}$. In the left plot the baseline intensity function $\beta t^{\beta-1}$ is used, while the baseline intensity function $e^{\beta t}$ is used in the right plot. The rejection probability is plotted as a function of β, and for the log-linear intensity function $e^{\beta t}$ only positive values of β (increasing trend) are used since for negative values there is a positive probability of getting infinite interfailure times.

In the left plot the heterogeneity test has the same rejection probability for any value of β, while the heterogeneity test in the right plot shows the unpleasant property that the rejection probability is rapidly decreasing for increasing β-values. This causes the two step test to be the least powerful test in this region. Since the heterogeneity test has very low rejection power the two step test is essentially a TTT-based Military Handbook test evaluated on a 2.5% significance level in this region. The reason why the heterogeneity test has this weakness in this case with log-linear intensity is probably because the heterogeneity test implemented is using a power law parameterization for the baseline intensity function. A heterogeneity test with a log-linear model for the baseline intensity would presumably not have had this weakness, but possibly similar weaknesses in cases with other intensity functions than the log-linear.

It is also interesting to note in Figure 3 that the TTT-based tests seem to be a bit "skewed". They have very low power against weakly increasing trend. The TTT-based Military Handbook test is least affected both by the skewness and by the rise in actual significance level. The two step test is also slightly affected by the skewness, but has correct significance level. The differences in power between the combined tests and the TTT-based tests are not so big since the observation intervals tend to be fairly equal with the setup

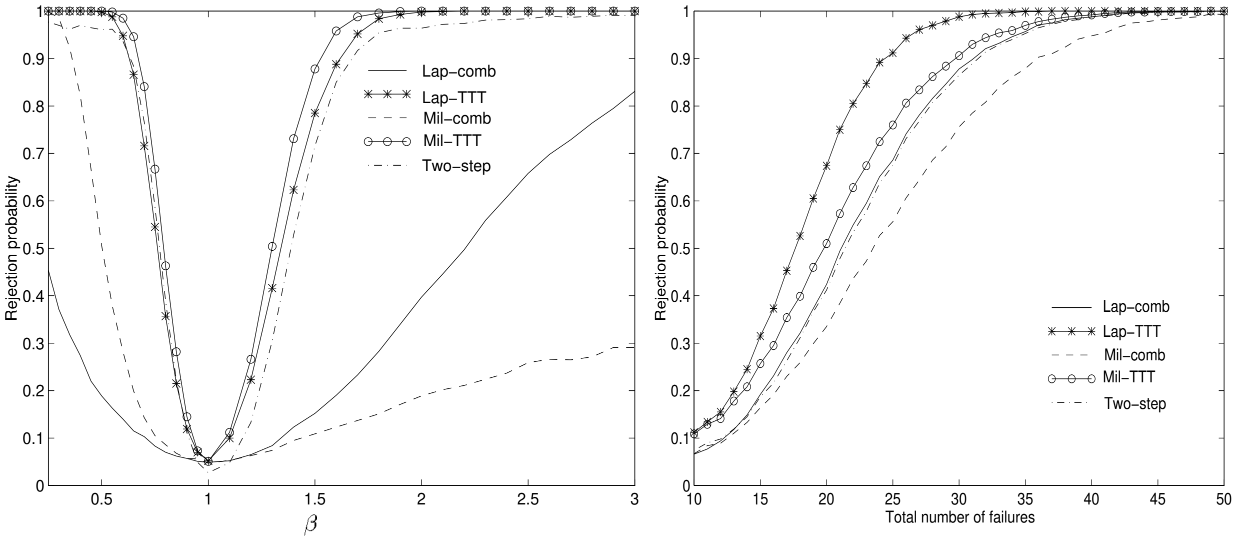

Figure 2: Simulations of NHPPs with no heterogeneities. In the left plot 10 partly overlapping processes with 10 observed failures and intensity function $\beta t^{\beta-1}$ are simulated, and the rejection probability is plotted as a function of β. In the right plot 5 processes with intensity $e^{0.5t}$ are simulated. Four of the processes are observed from time 0 and until 2 failures have occurred while a varying number of failures are observed in the last process. The rejection probability is plotted as a function of total number of failures.

used here.

3.4 *Shape heterogeneity*

Until now we have only considered heterogeneities that affect the scale parameter. But it is not unreasonable to think that, for the same reasons as mentioned in Section 2.5, there may be some differences in shape parameter from system to system as well. One way to simulate such differences is to add random noise to the shape parameter. In Figure 4 the simulations in Figure 3 have been repeated, but instead of having a multiplicative scale heterogeneity, a random noise ϵ_i where $\epsilon_i \sim \mathrm{N}(0, \frac{1}{20})$, is added to the shape parameter in each process $i = 1, \ldots, 10$.

By adding noise to the shape parameter there is no "no trend" situation in the plots displayed in Figure 4. The probably most interesting thing to notice is the high rejection probability of the heterogeneity test in most parts of the left plot and in the middle of the right plot. It is not surprising that something like this could happen since the heterogeneity test is constructed under the assumption of equal baseline intensity in each process, and this assumption is violated here. The heterogeneity test in fact seems to be more sensitive to this kind of heterogeneity than the scale kind of heterogeneity that it is constructed to detect, but this is not critical. It only implies that the two step test is using the combined test in cases where in fact the TTT-based test could have been used.

In Figure 4 we see that the two step test is almost identical to the combined Laplace test in regions with a very high rejection probability for the heterogeneity test, whereas in regions with a very low rejection probability of the heterogeneity test the two step test is clearly essentially a TTT-based Military Handbook test evaluated on a lower significance level than the 5% TTT-based Military Handbook test.

It is also worth noting in Figure 4 the big difference in rejection probability between the combined tests and the TTT-based tests near $\beta = 0$ in the right plot. This represents a situation where some of the systems have an increasing trend and some a decreasing trend. The combined tests have very low rejection probability since the terms in the test statistic from processes with increasing trend tends to "cancel" the terms from processes with decreasing trend. The TTT-based tests on the other hand have a very high rejection probability since the observation intervals typically will be of different length for processes with increasing trend versus processes with decreasing trend (since the processes are failure truncated).

If we want to test whether there are heterogeneity in the shape parameter likelihood ratio

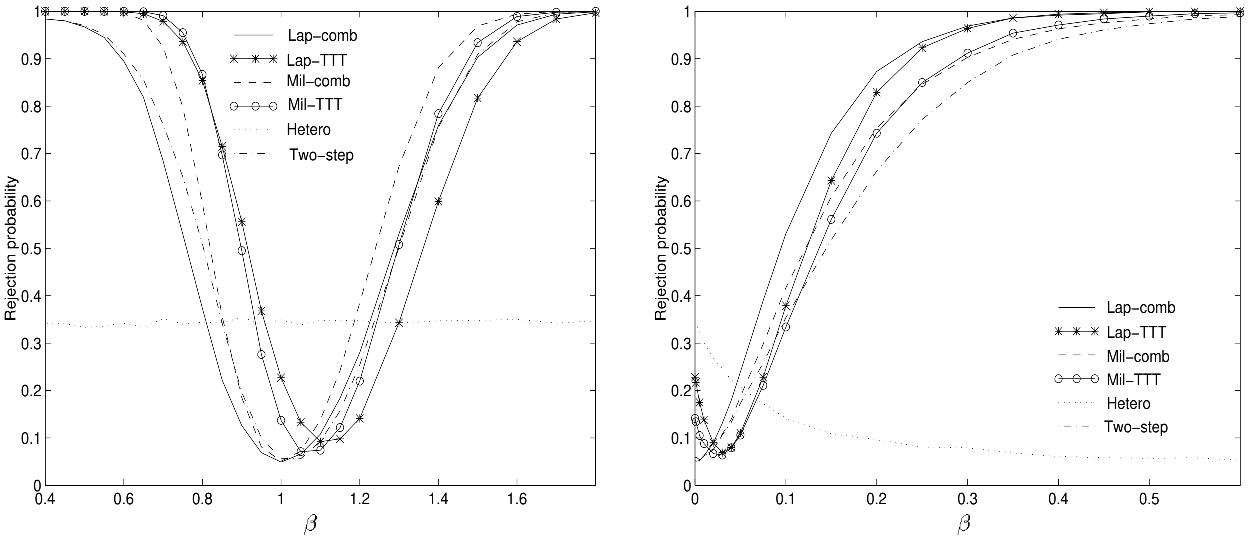

Figure 3: Simulations of NHPPs with scale heterogeneities with variance $\frac{1}{10}$. In both plots 10 process are observed from time 0 and until 10 failures have occurred. The baseline intensity functions used are respectively $\beta t^{\beta-1}$ and $e^{\beta t}$ and the rejection probability is plotted as a function of β.

tests for heterogeneity in the shape parameter can easily be derived (Tveit 1993).

4 CONCLUSION

The TTT-based tests are the best tests in cases with no heterogeneities, while the combined tests are the best tests in cases with heterogeneities. The two step test is a successful combination of the best properties of the TTT-based tests and the combined tests.

The TTT-based tests are far more powerful than the combined tests if the observation intervals are a bit different. The problem with the TTT-based tests is that these tests by their construction are unable to distinguish between data with trend and data with no trend but scale heterogeneities. Based on the simulation study the TTT-based Military Handbook tests is recommended over the TTT-based Laplace test since the TTT-based Military Handbook test in all cases studied is least sensitive to heterogeneities in the data.

Within the NHPP framework the combined tests can always be used since they allow scale heterogeneities between the systems. It is difficult to recommend one of the combined tests over the other. Which one is best varies from situation to situation.

The heterogeneity test shows satisfying properties as a test for heterogeneities in cases with no trend, but has weaknesses in cases with trend of other kind than the power law baseline intensity function model used in the parameterization of the heterogeneity test.

The two step test tries to draw benefit from both the good power properties of the TTT-based tests and the significance level properties of the combined tests, and is the test which in general has the best properties. It achieves a significance level very close to the correct significance level in all cases with heterogeneities studied, and does in general have better power properties than the combined tests. In the two step test the TTT-based Military Handbook test should be used as TTT-based test, while any of the combined tests could be used as combined test.

The general recommendation is to use the two step test. The exception is if the assumption of negligible heterogeneities is unquestionable in which case a TTT-based test should be used, or if the observation intervals are very equal in which case a combined test should be used.

Finally it should be pointed out that if the null hypothesis is rejected in a simultaneous trend analysis, this only implies that some of the systems have a trend, not necessarily all of them. Some of the systems may even have trend of an opposite kind of the overall trend detected by the simultaneous trend analysis. If the null hypothesis is rejected we can only conclude that the systems in general tend to have a trend of a certain kind.

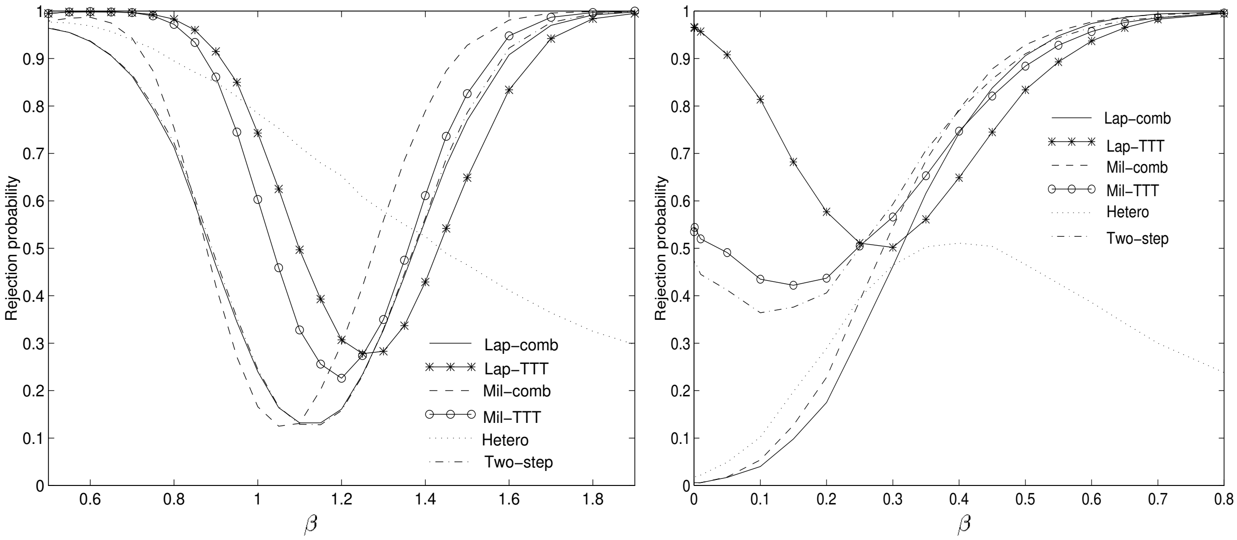

Figure 4: Simulations of NHPPs with shape heterogeneities with variance $\frac{1}{20}$. In both plots 10 process are observed from time 0 and until 10 failures have occurred. The baseline intensity functions used are respectively $\beta t^{\beta-1}$ and $e^{\beta t}$ and the rejection probability is plotted as a function of β.

ACKNOWLEDGEMENT

I would like to thank Bo Lindqvist for valuable suggestions and discussions. The author is supported by a grant from the Norwegian Council of Research.

REFERENCES

Ascher, H. & H.Feingold 1984. *Repairable Systems Reliability. Modeling, Inference, Misconceptions and Their Causes.* New York: Marcel Dekker.

Bain, L.J., M.Engelhardt & F.T.Wright 1985. Tests for an Increasing Trend in the Intensity of a Poisson Process: A Power Study. *Journal of the American Statistical Association.* 80:419-422.

Barlow, R.E. & B.Davis 1977. Analysis of Time Between Failures for Repairable Components. In J.B.Fussel & G.R. Burdick (ed.), *Nuclear Systems Reliability Engineering and Risk assessment*: 543-561. Philadelphia: SIAM.

Cohen, A. & H.B.Sackrowitz 1993. Evaluating Tests for Increasing Intensity of a Poisson Process. *Technometrics.* 35:446-448.

Elvebakk, G. 1998. Robustification of trend tests by resampling techniques. *Proceedings of ESREL 1998.* Rotterdam: Balkema

Kvaløy J.T. & B.H.Lindqvist 1998. TTT-based Tests for Trend in Repairable Systems Data. *Reliability Engineering and System Safety.* To appear.

Lawless, J.F. 1987. Regression Methods for Poisson Process Data. *Journal of the American Statistical Association.* 82:808-815.

Tveit, I.M. 1993. *Heterogenity in reliability analysis.* Project work, Norwegian Institute of Technology. (In Norwegian).