A Kolmogorov-Smirnov type test for shortfall dominance against parametric alternatives

Michel Denuit
Institut des Sciences Actuarielles & Institut de Statistique
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
denuit@stat.ucl.ac.be

Anne-Cécile Goderniaux
Institut Supérieur Industriel Pierrard
Haute Ecole Blaise Pascal
Virton, Belgium
anne.cecile.goderniaux@isipierrard.be

Olivier Scaillet
HEC Genève and Swiss Finance Institute
Université de Genève
Genève 4, Suisse
olivier.scaillet@hec.unige.ch

This version: June 2006 (first version: April 2005)

Abstract

This paper introduces a Kolmogorov-type test for the shortfall order (also known in the literature as the right-spread or excess-wealth order) against parametric alternatives. In the case of the null hypothesis corresponding to the Negative Exponential distribution, this provides a test for the new better than used in expectation (NBUE) and for the new worse than used in expectation (NWUE) properties. Such a test is particularly useful in reliability applications as well as duration and income distribution analysis. The theoretical properties of the testing procedure are first established for uncensored data, and then for censored and truncated data. Simulation studies reveal that the test based on a bootstrap procedure performs well, even with moderate sample sizes. Applications to real data, namely chief executive officer (CEO) compensation data, flight delay data and throttle failure data, illustrate its empirical relevance.

Key words and phrases: Right-spread order, Excess-wealth order, New better than used in expectation, New worse than used in expectation, Bootstrap, Reliability, Random censorship, Truncation.

AMS 2000: 60E15, 62F40, 62N05, 62P30, 90B25.

1 Introduction and motivation

An important aspect of reliability analysis is to find a lifetime distribution that can adequately describe the ageing behaviour of the item under study (a motor, an electronic component, a light bulb, etc.). Often, engineers are interested in the reliability of a non-repairable item (that can be anything from a small component to a large system). For such an item, the variable of interest is the time to failure or lifetime. This is the time elapsing from when the item is put into operation until it fails for the first time. For an introduction to reliability theory, we refer the reader, e.g., to Barlow & Proschan (1981) and Rausand & Hoyland (2004).

Ageing notions are used to explain how functioning items get used. Different ageing criteria have been used to classify positive and negative ageing properties. For example increasing failure rate (IFR), new better than used (NBU), decreasing mean remaining life (DMRL) and new better than used in expectation (NBUE), and their duals are the main existing ageing criteria; see Barlow & Proschan (1981). Ageing classes of life distributions are often based on comparison between survival functions of new and used items.

The following ageing notions will be encountered throughout the text. Let X be a non-negative random variable with distribution function F and survival function $\overline{F} \equiv 1 - F$. Then,

(i) X is said to be increasing failure rate (IFR, in short) if $\overline{F}$ is logconcave. It is said to be decreasing failure rate (DFR, in short) if $\overline{F}$ is logconvex. These notions correspond to the increasingness and decreasingness of the failure rate, respectively, when it exists.

(ii) X is said to be new better than used in expectation (NBUE, in short) if

$$\frac{\int_t^\infty \overline{F}(s)\,ds}{\overline{F}(t)} = \mathbb{E}[X-t|X>t] \leq \mathbb{E}[X] = \int_0^\infty \overline{F}(s)\,ds$$

for all $t \geq 0$. It is said to be new worse than used in expectation (NWUE, in short) if the reverse inequality holds for all $t \geq 0$. The quantity $\mathbb{E}[X-t|X>t]$ is called the mean residual life of the item with lifetime X: considering an item with time to failure X that is still functioning at time t, it gives the expected extra time during which the item will be working. Thus, X is NBUE if for all $t \geq 0$ the mean residual life at time t is not greater than the mean life of a new item.

A machine with lifetime X that is IFR or NBUE will age with the passage of time, in the sense that its expected remaining lifetime will diminish as it gets older. On the contrary, if the lifetime X is DFR or NWUE, this means that the reliability of the machine increases as it gets older. Note that IFR implies NBUE, and that DFR implies NWUE. The boundary members of each of these classes are the Negative Exponential distributions which, of course, are appropriate for models where lifetimes neither improve nor deteriorate with age.

Several ageing notions can also be characterized by means of a stochastic comparison with respect to the Negative Exponential distribution. We will see that the NBUE class can be obtained in this way using the shortfall order. Through its links with the NBUE/NWUE properties, the shortfall order also finds interesting and natural applications in the analysis of duration and income distributions (see below).

The present contribution proposes a Kolmogorov-Smirnov type test for the shortfall order. References for this stochastic ordering include SHAKED & SHANTHIKUMAR (1998), FAGIUOLI, PELLEREY & SHAKED (1999), and FERNANDEZ-PONCE, KOCHAR & MUÑOZ-PEREZ (1998). As it will be pointed out in the conclusion, our approach remains nevertheless applicable for many other stochastic orderings, defined by the pointwise comparison of a transform associated with the underlying distribution functions.

The IFR/DFR classes have received the most study from an inferential point of view. There are fewer papers devoted to the development of tests for NBUE/NWUE. The test proposed in this paper can be applied to check for the validity of the NBUE/NWUE assumptions. In contrast to most procedures found in the literature, Exponentiality will not be taken as our null hypothesis to be tested against the NBUE/NWUE alternatives. In this paper, the null hypothesis H_0 supports NBUE/NWUE whereas the alternative hypothesis H_1 is the violation of one of these ageing properties. Let us briefly comment on this issue. HOLLANDER & PROSCHAN (1975,1980) seem to have been the first authors to derive a test of Exponentiality versus NBUE alternatives. Specifically, H_0 corresponds to the Negative Exponential distribution whereas H_1 is NBUE but not exponential. The test obtained is the total time on test procedure, and is consistent against NBUE alternatives. Similarly, FERNANDEZ-PONCE, INFANTE-MACIAS & MUNOZ-PEREZ (1996), AHMAD, ALWASEL & MUGDADI (2001), AHMAD, HENDI & AL-NACHAWAK (1999), and BELZUNCE, PINAR & RUIZ (2001) take H_0 as the Negative Exponential distribution, as well as KOUL & SUSARLA (1980) who consider incomplete data.

The approach developed in this paper differs from previous works in two respects. First, we reverse the roles of the hypotheses: NBUE is now in H_0 contrary to all previous works, whereas the alternative hypothesis H_1 is the converse of the null (and thus corresponds to the violation of NBUE). Therefore, $H_0 \cup H_1$ contains all possible distributions, which is intuitively more appealing. Note that the null hypothesis is composite in our setting (in the sense that it is true for many different distributions, namely all the NBUE ones). Second, we do not test a necessary condition for NBUE/NWUE (as HOLLANDER & PROSCHAN (1975,1980), KOUL & SUSARLA (1980), AHMAD, ALWASEL & MUGDADI (2001), AHMAD, HENDI & AL-NACHAWAK (1999), and BELZUNCE, PINAR & RUIZ (2001) do), but we test a null hypothesis that corresponds directly to the NBUE property.

Let us also mention that all the papers cited above (except KOUL & SUSARLA (1980)) only deal with complete data whereas an extension to censored and truncated data is worked out here as well. Since the testing procedure remains valid for incomplete data (which are common in practice) this is a clear advantage of our approach.

The paper is organized as follows. Section 2 gathers some fundamental results about the shortfall order. Section 3 presents the testing procedure. Section 4 is devoted to an extension of the results to censored and truncated data. In Section 5, a simulation study is provided to evaluate the finite sample performance of the test in terms of size. The power of the test is also computed against different parametric alternatives, namely Weibull, Gamma, and Pareto alternatives. Section 6 contains the empirical illustrations. First, we consider compensation data for chief executive officers (CEO) of some of the largest US companies (randomly sampled from the Executive Paywatch database). It will be seen that these data support the NWUE property, and this feature will lead to interesting interpretations about the allocation of compensations among CEO. Second, we study data about European flights gathered by Eurocontrol. Delays before departure are measured in time (minutes), and delays after departure are measured as the increase in the route length (kilometers). Both types of delays will be analyzed. In the latter case, NWUE will be detected, but not in the former where NBUE seems to hold. Finally, we examine a censored data set about throttle failures. In this case, the NBUE property cannot be rejected. Section 7 contains some concluding remarks. Proofs are gathered in an appendix at the end of the paper.

2 Shortfall order

2.1 Expected shortfall and related notions

Let $Q(u) = \inf\{x \in \mathbb{R} | F(x) \geq u\}$ be the inverse of the distribution function F (also called the quantile function). Having a probability level u and a random variable X with distribution function F, the positive part $(X - Q(1-u))_+$ of $X - Q(1-u)$ is called the shortfall, denoted by $S(u)$. The quantile $Q(1-u)$ is the level that is exceeded with probability u (at most). Hence, the shortfall $S(u)$ represents the possible exceedance of X over the threshold $Q(1-u)$ exceeded by only $100u\%$ of similar devices. The expected shortfall ES is then defined as the average shortfall, that is

$$ES(u) = \mathbb{E}[S(u)] = \mathbb{E}[(X - Q(1-u))_+] = \int_{Q(1-u)}^{+\infty} \overline{F}(x)dx.$$

The expected shortfall is widely used in reliability, finance and insurance. It possesses several nice properties, that make it appealing for practical applications. It is also closely related to the so-called excess-wealth transform defined by

$$W(u) = \frac{1}{\mathbb{E}[X]} \int_{Q(u)}^{+\infty} \overline{F}(x)dx = \frac{ES(1-u)}{\mathbb{E}[X]}, \quad u \in [0,1]. \tag{2.1}$$

In the context of economics, if X is thought of as an income, then $W(u)$ can be viewed as the proportion of the additional wealth (on top of the u-th percentile) of the richest $100(1-u)\%$ individuals in the population.

The expected shortfall is also closely related to the so-called total time on test transform. Recall that the transform W has been introduced by Kochar, Li & Shaked (2002) to compare probability distributions. Alternatively, $u \mapsto ES(1-u)$ is also called the right-spread function. This function is used to measure the average exceedance over a quantile $Q(u)$. It has been considered in Belzunce, Pinar & Ruiz (2001).

The total time on test (TTT) transform of the non-negative random variable X is defined for $u \in (0,1)$ as

$$T(u) = \mathbb{E}[X] - ES(1-u) = \int_0^{Q(u)} \overline{F}(x)dx.$$

The TTT transform is the theoretical counterpart of the empirical TTT transform that is often used in statistical reliability theory. Broadly speaking, $T(u)$ gives the average time that an item spends on test if the test is terminated when a fraction u of all the items on the test fail. For more details about the TTT transform, we refer the interested reader e.g. to Pham & Turkkan (1994).

2.2 Shortfall order

Having two random variables X and Y, X is said to be smaller than Y in the shortfall order if the expected shortfall for X is always smaller than the corresponding expected shortfall for Y, whatever the probability level u. Specifically, denoting by F and G the distribution functions of X and Y, respectively, X precedes Y in the shortfall order, which is denoted as $X \preceq_{SH} Y$, if $ES_F(u) \leq ES_G(u)$ for all u. When X and Y are two times to failure, $X \preceq_{SH} Y$ means that on average, the extra time elapsed after $100u\%$ of the similar devices fail is larger for Y than for X, whatever u.

There is also a scaled version of the shortfall order. Specifically, X is said to be smaller than Y in the scaled shortfall order, denoted as $X \preceq_{SH,=} Y$ if the inequality $ES_F(u)/\mathbb{E}[X] \leq ES_G(u)/\mathbb{E}[Y]$ holds for all probability levels u. If $\mathbb{E}[X] = \mathbb{E}[Y]$ then we obviously have that $X \preceq_{SH} Y \Leftrightarrow X \preceq_{SH,=} Y$.

The scaled shortfall order is called the excess wealth order in Shaked & Shanthikumar (1998) where it is defined by means of the excess-wealth transform. The shortfall order is also termed as the right-spread order e.g., in Belzunce, Pinar & Ruiz (2001). This name originates in the fact that $\preceq_{SH}$ is based on the pointwise comparison of the right-spread functions associated with the distribution functions to be ordered.

2.3 Ageing notions and shortfall order

The classification of a lifetime according to the type of ageing structure it represents is useful, e.g., to decide about the suitability of a parametric model for the data to be analyzed. By ageing, we mean the phenomenon whereby an older system has a shorter lifetime, in a statistical sense, than a younger one (after Bryson & Siddiqui (1969)). Some orderings of distributions have been used to give characterizations of ageing classes. The idea is to compare the actual distribution to the Negative Exponential distribution.

The Negative Exponential distribution is often taken as a benchmark in reliability theory. An assumption of Exponentially distributed lifetimes implies that a used item is stochastically as good as new, so there is no reason to replace a functioning item. The Negative Exponential distribution is the most commonly used life distribution in applied reliability analysis (mainly because of its mathematical simplicity). It is therefore of interest to detect possible departures from exponentiality in the data, such as NBUE/NWUE, for instance.

Belzunce, Pinar & Ruiz (2001) proved that the class of the NBUE/NWUE distributions can be characterized with the help of the shortfall order with respect to the Negative Exponential distribution with the same mean. Specifically, given a non-negative random variable X with finite mean,

$$X \text{ NBUE} \Leftrightarrow X \preceq_{SH} \mathcal{E}xp(\mathbb{E}[X]),$$

where $\mathcal{E}xp(\mathbb{E}[X])$ represents a random variable with survival function $t \mapsto \exp(-t/\mathbb{E}[X])$. Similarly,

$$X \text{ NWUE} \Leftrightarrow \mathcal{E}xp(\mathbb{E}[X]) \preceq_{SH} X.$$

It is easy to check that if X is Exponentially distributed with mean μ, then $ES(u) = u\mu$. The NBUE property thus means that the expected shortfalls for X are smaller than the straight line with slope $\mathbb{E}[X]$, whereas the NWUE property places the expected shortfalls for X above this straight line.

In the remainder of this paper, we only describe the testing procedure in the NBUE case. The modification to accomodate the NWUE case is straightforward through changing the sense of the inequalities.

3 Testing procedure

3.1 Nonparametric estimation of the expected shortfall

In this paper we consider an i.i.d. sampling scheme. Hence, we work on the basis of the following hypothesis.

Assumption 3.1. $\{x_i\}_{i=1}^N$ *is a random sample from a continuous distribution with distribution function* F *and mean* m.

The probability density function is denoted as usual with a lower case, namely f.

Let us introduce the empirical distribution $\hat{F}(z) := \frac{1}{N}\sum_{i=1}^N \mathbb{I}\{x_i \leq z\}$. From the theory of empirical processes we know that $\sqrt{N}(\hat{F} - F)$ converges weakly to a Brownian bridge process $\mathcal{B}_F \circ F$ (see van Der Vaart & Wellner (1986), henceforth referred to as VW).

To build the test statistic we need empirical counterparts of the moments involved in the definition of the shortfall order. Let us first deal with the nonparametric part $ES(u)/m$. The mean m will be estimated with the empirical mean:

$$\hat{m} \;:=\; \frac{1}{N}\sum_{i=1}^N x_i. \tag{3.1}$$

Note that

$$ES(u) = \int_{Q(1-u)}^{+\infty} (x - Q(1-u))dF(u) = \int_{Q(1-u)}^{+\infty} xdF(u) - uQ(1-u).$$

The expected shortfall $ES(u)$ can therefore be estimated as

$$\hat{E}S(u) \;:=\; \frac{1}{N}\sum_{i=1}^N x_i\mathbb{I}\{x_i > \hat{Q}(1-u)\} - u\hat{Q}(1-u), \tag{3.2}$$

where $\hat{Q}(1-u) := \hat{F}^{-1}(1-u)$ is the empirical quantile.

3.2 Sample distribution of the empirical expected shortfalls

Note that the estimators (3.1) and (3.2) can be viewed as particular functionals of the empirical distribution $\hat{F}$:

$$\begin{aligned} \hat{m} &= \int x d\hat{F}(x), \\ \hat{E}S(u) &= \int x\mathbb{I}\{x > \hat{F}^{-1}(1-u)\}d\hat{F}(x) - u\hat{F}^{-1}(1-u). \end{aligned}$$

To make explicit the dependence on $\hat{F}$, we will use the notation $\mathcal{J}(u; \hat{F}) := \hat{ES}(u)/\hat{m}$.

The characterisation of the test statistic in terms of a given map of the empirical distributions $\hat{F}$ is instrumental in the proof of the following lemma. The lemma describes the limiting behaviour of $\mathcal{J}(u; \hat{F})$, and is useful to deduce the properties of the testing procedure (see proof of Proposition 3.7).

Lemma 3.2. *Under Assumption 3.1, $\sqrt{N}(\mathcal{J}(u; \hat{F}) - \mathcal{J}(u; F))$ converges weakly in $C((0, 1))$ (the space of continuous functions on $(0, 1)$) to a mean zero Gaussian process $\mathcal{J}'_F(u; \mathcal{B}_F \circ F)$ with covariance kernel given by:*

$$\begin{aligned}
&\Omega_F(u_1, u_2) = E[\mathcal{J}'_F(u_1; \mathcal{B}_F \circ F)\mathcal{J}'_F(u_2; \mathcal{B}_F \circ F)] \\
&= \mathbb{C}ov \Bigg[\frac{1}{m}(X - F^{-1}(1 - u_1) + \frac{u_1}{f(F^{-1}(1 - u_1))})\mathbb{I}\{X > F^{-1}(1 - u_1)\} - \frac{ES(u_1)}{m^2}X, \\
&\qquad \frac{1}{m}(X - F^{-1}(1 - u_2) + \frac{u_2}{f(F^{-1}(1 - u_2))})\mathbb{I}\{X > F^{-1}(1 - u_2)\} - \frac{ES(u_2)}{m^2}X \Bigg].
\end{aligned}$$

Remark 3.3. If we do not want to check the scaled version of the order, we can simply rely on $\bar{\mathcal{J}}(u; \hat{F}) := \hat{ES}(u)$ to build the testing procedure. A weak convergence result similar to Lemma 3.2 holds, but with a limiting process $\bar{\mathcal{J}}'_F(u; \mathcal{B}_F \circ F)$, whose covariance kernel is:

$$\begin{aligned}
\bar{\Omega}_F(u_1, u_2) &= E[\bar{\mathcal{J}}'_F(u_1; \mathcal{B}_F \circ F)\bar{\mathcal{J}}'_F(u_2; \mathcal{B}_F \circ F)] \\
&= \mathbb{C}ov \Bigg[(X - F^{-1}(1 - u_1) + \frac{u_1}{f(F^{-1}(1 - u_1))})\mathbb{I}\{X > F^{-1}(1 - u_1)\}, \\
&\qquad (X - F^{-1}(1 - u_2) + \frac{u_2}{f(F^{-1}(1 - u_2))})\mathbb{I}\{X > F^{-1}(1 - u_2)\} \Bigg].
\end{aligned}$$

3.3 Sample distribution of the estimated parametric expected shortfalls

Let us now examine the parametric case $ES_\theta(u)/m_\theta$. An estimate $ES_{\hat{\theta}}(u)/m_{\hat{\theta}}$ obtained by plug-in may be viewed as a functional of $F_{\hat{\theta}}$, i.e. $\mathcal{J}(u; F_{\hat{\theta}})$, or as a functional of $\hat{\theta}$ itself, i.e. $J_{\hat{\theta}}(u) := \mathcal{J}(u; F_{\hat{\theta}})$. Hence we can easily characterize the limiting behavior of $\sqrt{N}(\mathcal{J}(u; F_{\hat{\theta}}) - \mathcal{J}(u; F_{\theta_0}))$ if we know the limiting behaviour of $\sqrt{N}(\hat{\theta} - \theta_0)$.

Assumption 3.4. *$\sqrt{N}(\hat{\theta} - \theta_0)$ converges to a mean zero Gaussian random variable $-\boldsymbol{V}_{\theta_0}^{-1}\boldsymbol{G}_{\theta_0}$.*

Note that this assumption is satisfied by classical M-estimators (VW Section 3.2) and Z-estimators (VW Section 3.3).

The next lemma is a direct consequence of the delta-method (VW Section 3.9), and is analogous to Lemma 3.2. Hence we omit its proof.

Lemma 3.5. *Under Assumptions 3.1 and 3.4, $\sqrt{N}(\mathcal{J}(u; F_{\hat{\theta}}) - \mathcal{J}(u; F_{\theta_0}))$ converges weakly in $C((0, 1))$ (the space of continuous functions on $(0, 1)$) to a mean zero Gaussian process $J'_{\theta_0}(u; -\boldsymbol{V}_{\theta_0}^{-1}\boldsymbol{G}_{\theta_0})$ with covariance kernel given by:*

$$\Omega_{\theta_0}(u_1,u_2) = E[\dot{\boldsymbol{J}}_{\theta_0}(u_1)\boldsymbol{V}_{\theta_0}^{-1}\boldsymbol{G}_{\theta_0}\boldsymbol{G}_{\theta_0}^{\top}\boldsymbol{V}_{\theta_0}^{-1}\dot{\boldsymbol{J}}_{\theta_0}(u_2)^{\top}]$$

where

$$\dot{\boldsymbol{J}}_{\theta_0}(u) = \frac{\partial}{\partial\theta}\boldsymbol{J}_{\theta}(u)\Big|_{\theta=\theta_0}.$$

3.4 Testing procedure

Since we wish to test for a dominance in the $\preceq_{SH}$-sense with respect to a parametric model, namely

$$\begin{aligned} H_0: &\quad \mathcal{J}(u;F) \le \mathcal{J}(u;F_\theta) \quad \text{for all}\, u \in (0,1), \\ H_1: &\quad \mathcal{J}(u;F) > \mathcal{J}(u;F_\theta) \quad \text{for a}\, u \in (0,1), \end{aligned}$$

we consider the test statistic

$$\hat{S} := \sqrt{N}\sup_u(\mathcal{J}(u;\hat{F}) - \mathcal{J}(u;F_{\hat{\theta}})),$$

and a test based on the decision rule:

$$\text{“ reject} \quad H_0 \quad \text{if} \quad \hat{S} > c\text{”},$$

where c is a critical value that will be approximated by an empirical quantile of bootstrap values of the test statistic.

The following result characterizes the properties of the test, where

$$\bar{S} := \sup_u(\mathcal{J}'_F(u;\mathcal{B}_F \circ F) - J'_{\theta_0}(u; -\boldsymbol{V}_{\theta_0}^{-1}\boldsymbol{G}_{\theta_0})).$$

Proposition 3.6. *Let c be a positive finite constant, then under Assumptions 3.1 and 3.4:*

i) if H_0 is true,

$$\lim_{N\to\infty} P[\textit{reject}\, H_0] \le P[\bar{S} > c] := \alpha(c),$$

with equality when $\mathcal{J}(u;F) = \mathcal{J}(u;F_{\theta_0})$ for all $u \in (0,1)$;

ii) if H_0 is false,

$$\lim_{N\to\infty} P[\textit{reject}\, H_0] = 1.$$

The first part of the result provides a random variable that dominates the limiting random variable corresponding to the test statistic under the null hypothesis. The inequality tells us that the test will never reject more often than $\alpha(c)$ when the null hypothesis is satisfied. Furthermore the probability of rejection will asymptotically be exactly $\alpha(c)$ when we have strict equality. The first part also implies that if one could find a c to set the $\alpha(c)$ to a desired probability level (say the conventional 0.05 or 0.01) then this would be the significance level for composite null hypotheses in the sense described by Lehmann (1986). The second part of the result indicates that the test is capable of detecting any violation of the full set of restrictions of the null hypothesis.

Of course, in order to make the result operational, we need to find an appropriate critical value c. Since the distribution of the test statistic depends on the underlying unknown distributions, this is not an easy task, and we decide hereafter to rely on the bootstrap method to simulate p-values.

3.5 Simulating p-values

We rely on the traditional bootstrap (see Barrett & Donald (2003) and Abadie (2002) for use in stochastic dominance tests). Alternatively we could use a subsampling method (as presented in Politis, Romano & Wolf (1999)) instead of a bootstrap method to get simulated p-values. The approach outlined in this section can be easily adapted to that framework.

A bootstrap sample $\{x_i^*\}_{i=1}^N$ is built from drawing N pairs with replacement from $\{x_i\}_{i=1}^N$. Let $\hat{F}^*$ denote the empirical distribution function associated to this bootstrap sample and $F_{\hat{\theta}^*}$ the distribution function associated with parametric estimation on the bootstrap sample. Let us further take

$$\hat{S}^* := \sqrt{N} \sup_u ((\mathcal{J}(u; \hat{F}^*) - \mathcal{J}(u; \hat{F})) - (\mathcal{J}(u; F_{\hat{\theta}^*}) - \mathcal{J}(u; F_{\hat{\theta}})),$$

and define

$$p^* := P[\hat{S}^* > \hat{S}].$$

Then the bootstrap method is justified by the next statement.

Proposition 3.7. *Assuming that $\alpha < 1/2$, a test for dominance in the $\preceq_{SH}$-sense based on the rule:*

$$\text{“ } \textit{reject}\, H_0 \quad \textit{if} \quad p^* < \alpha \text{ ”},$$

satisfies the following

$$\begin{aligned} \lim P[\textit{reject}\, H_0] &\leq \alpha \quad \textit{if} \quad H_0 \textit{ is true}, \\ \lim P[\textit{reject}\, H_0] &= 1 \quad \textit{if} \quad H_0 \textit{ is false}. \end{aligned}$$

In practice we need to use Monte-Carlo methods to approximate the probability and a grid to approximate the supremum. The p-value is simply approximated by

$$p^* \approx \frac{1}{R} \sum_{r=1}^{R} \mathbb{I}\{\hat{S}_r^* > \hat{S}\},$$

where the averaging is made on R bootstrap replications and $\hat{S}_r^*$ is computed from a fine grid on $(0, 1)$. Note that the replication number and the grid mesh can be chosen to make the approximations as accurate as one desires given time and computer constraints.

4 Extension to censored and truncated data

In this section we briefly discuss the extension of previous results when data are incomplete. We first examine right censored data in the standard random censorship model before examining truncated data.

Assumption 4.1. *$\{z_i, \delta_i\}_{i=1}^N$ is a random sample where $z_i = \min(x_i, y_i)$ is either an observed value, x_i, or an observed censoring value, y_i, and $\delta_i = \mathbb{I}\{x_i = z_i\}$. The observed values and observed censoring values are independent with continuous distribution functions F and G, respectively.*

The most commonly used estimator of F is the Kaplan-Meier estimator (see, e.g., ANDERSEN, BORGAN, GILL & KEIDING (1993), hereafter ABGK) for a review of its properties and applications) defined by :

$$\tilde{F}(x) := 1 - \prod_{z_{(i)} \leq x} \left(\frac{N-i}{N-i+1} \right)^{\delta_{(i)}},$$

where $z_{(1)} \leq z_{(2)} \leq \ldots \leq z_{(N)}$ denote the ordered values of $z_1, z_2, \ldots, z_N$, and $\delta_{(i)}$ is the indicator function associated with $z_{(i)}$.

In such a setting, GILL (1983) suggests to estimate the mean m with

$$\tilde{m} := \int_0^{z_{(N)}} (1 - \tilde{F}(x))dx = z_{(1)} + \sum_{i=1}^{n-1} (z_{(i+1)} - z_{(i)}) \prod_{j=1}^{i} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}.$$

Similarly we suggest to estimate the expected shortfall $ES(u)$ with

$$\tilde{E}S(u) := \int_{\tilde{Q}(1-u)}^{z_{(N)}} (1 - \tilde{F}(x))dx,$$

where $\tilde{Q}(1-u) := \tilde{F}^{-1}(1-u)$ is the empirical quantile induced by the Kaplan-Meier estimator. As before we can put $\mathcal{J}(u; \tilde{F}) := \tilde{E}S(u)/\tilde{m}$.

In order to describe the limiting behaviour of $\mathcal{J}(u; \tilde{F})$ we need to introduce the distribution $H := 1 - (1-F)(1-G)$ from which the random sample $\{z_i\}_{i=1}^N$ is drawn, and the (possibly infinite) bound $\tau_H := \sup\{x : H(x) < 1\}$. Let us also introduce the continuous, nonnegative, nondecreasing functions:

$$\begin{aligned} \Lambda(x) &:= \int_0^x \frac{dF(s)}{1-F(s)}, \\ C(x) &:= \int_0^x \frac{dF(s)}{(1-F(s))^2(1-G(s))} = \int_0^x \frac{d\Lambda(s)}{1-H(s)}. \end{aligned}$$

The function Λ is the so-called cumulative hazard function. Then we know that $\sqrt{N}(\tilde{F} - F)$ converges weakly to a zero-mean Gaussian martingale $\mathcal{Z} = \mathcal{B}_C \circ C$, whose covariance function is $\mathbb{C}\mathrm{ov}[\mathcal{Z}(z_1), \mathcal{Z}(z_2)] = C(\min(z_1, z_2))$, while $\sqrt{N}(\tilde{F}^{-1} - F^{-1})$ converges weakly to $-(1-\iota)(\mathcal{Z} \circ F^{-1})/(f \circ F^{-1})$, where ι denotes the identity mapping (see, e.g., DOSS & GILL (1992), and ABGK Example IV.3.7).

Let us define the functions:

$$m(x) := \int_0^x (1 - F(s))ds, \qquad \bar{m}(x) := m - m(x).$$

Then GILL (1983) (ABGK Example IV.3.8) shows that $\sqrt{N}(\tilde{m} - m)$ converges to $\int_0^{\tau_H} \bar{m} d\mathcal{Z}$ if $\int_0^{\tau_H} \bar{m}^2 dC < \infty$ and $\sqrt{N}\bar{m}(z_{(N)})$ goes to zero when $N \longrightarrow \infty$.

In an analogous way we define the functions:

$$ES(u, x) := \int_{Q(1-u)}^x (1 - F(s))ds, \qquad \bar{E}S(u, x) := ES(u) - ES(u, x),$$

in order to deduce a result similar to Lemma 3.2, namely that, if $\sqrt{N}\bar{E}S(u; z_{(N)})$ goes to zero when $N \longrightarrow \infty$, $\sqrt{N}(\mathcal{J}(u; \tilde{F}) - \mathcal{J}(u; F))$ converges weakly to a zero-mean Gaussian martingale given by

$$\frac{1}{m}\left[\int_{F^{-1}(1-u)}^{\tau_H} \bar{E}S(u,x)d\mathcal{Z}(x) + \frac{u^2}{f(F^{-1}(1-u))}\mathcal{Z}(F^{-1}(1-u))\right] - \frac{ES(u)}{m^2}\int_0^{\tau_H} \bar{m}(x)d\mathcal{Z}(x).$$

Note that the presence of censoring does not affect the validity of Lemma 3.5 since Assumption 3.4 remains true for classical parametric estimators under random censorship (Borgan (1984)). Therefore we can easily design a test based on

$$\tilde{S} := \sqrt{N}\sup_u(\mathcal{J}(u; \tilde{F}) - \mathcal{J}(u; F_{\hat{\theta}})),$$

whose properties are as in Proposition 3.6.

In order to implement the test in practice we can exploit the bootstrap procedure of Efron (1981) relying on N drawings with replacement from $\{z_i, \delta_i\}_{i=1}^N$ to build the bootstrap sample $\{z_i^*, \delta_i^*\}_{i=1}^N$. Efron (1981) shows that such a bootstrap for data subject to right censoring inherits the characteristics of the standard bootstrap. Therefore the use of

$$\tilde{S}^* := \sqrt{N}\sup_u((\mathcal{J}(u; \tilde{F}^*) - \mathcal{J}(u; \tilde{F})) - (\mathcal{J}(u; F_{\hat{\theta}^*}) - \mathcal{J}(u; F_{\hat{\theta}})),$$

yields a bootstrap testing procedure matching the properties of Proposition 3.7.

Another case of practical interest concerns left truncated data, namely a random sample $\{x_i\}_{i=1}^N$ only observed if x_i is above a constant threshold T. Since none of the data are observed below the threshold, we have no information on the part of the distribution below T, and we cannot identify that part nonparametrically. Nevertheless we might still be interested in testing the NBUE property for the observable part above T, and in obtaining a consistent test for that part. To this aim we suggest to modify the procedure described in Section 3 as follows. For the parametric part we have a) to estimate the hypothesized parametric distribution on $[0, \infty)$ taking into account the presence of a truncation, b) to rescale the estimated distribution on $[T, \infty)$ so that it is a properly defined distribution (integrate to one), c) to compute the parametric expected shortfall under that truncated distribution. For the nonparametric part the procedure remains unchanged since the empirical cdf of the observed data converges to the true truncated distribution. The bootstrap method remains unaffected as well. Here the only difference with respect to Section 3 lies in working with truncated parametric distributions instead of distributions defined on $[0, \infty)$.

5 Monte Carlo results

In this section we examine the performance of the test in small and moderate samples. The replication number R to approximate the p-value is set equal to 1,000. A total of 250 Monte Carlo simulations are performed, and the rejection rates are computed for the bootstrap method with respect to the standard significance levels of $\alpha = 0.01$, 0.05 and 0.1.

5.1 Size

We first evaluate the Type I error (i.e. the probability to reject H_0 when it is true). Table 5.1 presents the results when the parametric distribution is the Negative Exponential with unit mean. In this case, the true distribution and the parametric distribution coincide. Then Proposition 3.6 suggests that the test should reject the null hypothesis H_0 with a frequency close to the chosen nominal significance level. This experiment should give us an idea about the validity of the asymptotic theory and the bootstrap method used to simulate the p-values in small samples in terms of size. The values displayed in Table 5.1 indicate that the test tends to reject the true H_0 less often than prescribed by the significance level in small samples, but the distortion is smaller for samples of sizes $N = 500$ and 1,000.

Table 5.1: Proportion of Negative Exponential samples where H_0 was rejected for different sample sizes with 250 simulations and 1,000 bootstrap replications.

Significance	Sample size				
level	N=25	N=50	N=100	$N = 500$	$N = 1,000$
α=0.01	0.000	0.004	0.000	0.004	0.008
	(0.000)	(0.004)	(0.000)	(0.004)	(0.006)
α=0.05	0.020	0.012	0.020	0.024	0.028
	(0.009)	(0.007)	(0.009)	(0.010)	(0.010)
α=0.1	0.032	0.056	0.048	0.052	0.056
	(0.011)	(0.015)	(0.014)	(0.014)	(0.015)

5.2 Power

Weibull alternative Table 5.2 gathers the results concerning power properties with a Weibull alternative. The Weibull distribution is one of the most widely used life distribution in reliability analysis. It has been initially developed for modelling the strength of materials.

Here, without loss of generality, we assume a value of one for the scale parameter. Thus we use a one-parameter Weibull distribution with shape parameter less than 1, to mimic a DFR behaviour. The associated probability density function is

$$f(x) = \theta x^{\theta-1} \exp\left(-x^{\theta}\right), \ \ x > 0,$$

for a positive parameter θ. Note that when $\theta = 1$ the Weibull distribution reduces to the Negative Exponential one. When $\theta = 2$, the resulting distribution is known as the Rayleigh distribution. This distribution is DFR when the shape parameter θ is less than 1, and IFR when θ is greater than 1. Therefore, the Weibull distributions are not NBUE for $\theta < 1$, and we expect a rejection of H_0 in these cases.

We see from Table 5.2 that for small values of θ (0.25 and 0.5), the testing procedure indeed rejects H_0 for the vast majority of the samples, even for small sample sizes. The power of the test is quite high (even for small sample sizes) and increases with the sample size. For larger values of θ (i.e., when the true parent distribution is "closer" to the Exponential one), the power is lower in small samples, but becomes reasonable in moderate samples (for a sample of size 100 and $\theta = 0.75$, the power for the usual levels of 0.05 and 0.1 are respectively equal to 74.0% and 83.2%).

Table 5.2: Proportion of samples where H_0 was rejected for different Weibull distributions and for different sample sizes with 250 simulations and 1,000 bootstrap replications.

Weibull parameter θ	Significance level	Sample size N=25	N=50	N=100
0.25	α=0.01	0.772	0.968	1
		(0.027)	(0.011)	(0)
	α=0.05	0.936	0.996	1
		(0.015)	(0.004)	(0)
	α=0.1	0.984	0.996	1
		(0.008)	(0.004)	(0)
0.5	α=0.01	0.512	0.880	1
		(0.032)	(0.021)	(0)
	α=0.05	0.792	0.984	1
		(0.026)	(0.008)	(0)
	α=0.1	0.876	0.996	1
		(0.021)	(0.004)	(0)
0.75	α=0.01	0.088	0.128	0.392
		(0.018)	(0.021)	(0.031)
	α=0.05	0.252	0.368	0.740
		(0.027)	(0.031)	(0.028)
	α=0.1	0.360	0.576	0.832
		(0.030)	(0.031)	(0.024)

Gamma alternative The Gamma density is also commonly used in lifetime analysis. The density function is

$$f(x) = \frac{\lambda}{\Gamma(\theta)}(\lambda x)^{\theta-1}\exp(-\lambda x), \;\; x > 0,$$

where $\theta > 0$ and $\lambda > 0$. When $\theta \geq 1$, the Gamma distributions are IFR. On the contrary, if $\theta \leq 1$ the the Gamma distributions are DFR. For $\theta = 1$, the Gamma distribution is reduced to the Negative Exponential distribution. Here, we simulate Gamma samples with $\theta < 1$ and $\lambda = 1$. These samples are not NBUE, and we expect the test procedure rejects the null hypothesis. Table 5.3 is the analogue of Table 5.2 for Gamma alternatives.

The same conclusions apply. Note that the power is lower for Gamma distributions than for Weibull distributions. However even if they share the same shape parameter value, they are not directly comparable.

Table 5.3: Proportion of samples where H_0 was rejected for different Gamma distributions and for different sample sizes with 250 simulations and 1,000 bootstrap replications.

Gamma parameter θ	Significance level	Sample size N=25	N=50	N=100
0.25	α=0.01	0.476	0.920	1
		(0.032)	(0.017)	(0)
	α=0.05	0.840	0.996	1
		(0.023)	(0.004)	(0)
	α=0.1	0.928	1	1
		(0.016)	(0)	(0)
0.5	α=0.01	0.112	0.284	0.748
		(0.020)	(0.029)	(0.027)
	α=0.05	0.332	0.644	0.916
		(0.030)	(0.030)	(0.018)
	α=0.1	0.504	0.832	0.972
		(0.032)	(0.024)	(0.010)
0.75	α=0.01	0.012	0.028	0.048
		(0.007)	(0.010)	(0.014)
	α=0.05	0.064	0.132	0.228
		(0.015)	(0.021)	(0.027)
	α=0.1	0.152	0.228	0.376
		(0.023)	(0.027)	(0.031)

Pareto alternative The Pareto distribution can be obtained as a mixture of Negative Exponential distributions, or as the exponential transform of a Negative Exponentially distributed random variable (suitably shifted to have the positive half real line as support). The survival function corresponding to the one-parameter Pareto distribution is given by $\overline{F}(x) = (1+x)^{-\theta}$, $x \geq 0$, for a parameter $\theta > 0$.

The Pareto distribution always exhibits a long-tailed behavior, and it is DFR for all the values of the parameter θ. Therefore, the testing procedure should reject the null hypothesis. Considering the results of the simulation study displayed in Table 5.4, we see that the power of the test increases as the sample gets bigger and as the value of θ decreases. The performance of the test is remarkable in moderate samples.

5.3 Power in the censored case

Let us now briefly investigate the power of the test in the incomplete case. The censoring time Y is taken to be Exponentially distributed with mean μ. The proportion of censored observations in the sample is then given by

$$P[Y < X] = 1 - \int_0^\infty \exp(-x/\mu) dF(x) = L(1/\mu),$$

Table 5.4: Proportion of samples where H_0 was rejected for different Pareto distributions and for different sample sizes with 250 simulations and 1,000 bootstrap replications.

Pareto parameter θ	Significance level	Sample size N=25	N=50	N=100
1.50	α=0.01	0.420	0.740	0.924
		(0.031)	(0.028)	(0.017)
	α=0.05	0.616	0.880	0.992
		(0.031)	(0.021)	(0.006)
	α=0.1	0.724	0.928	0.996
		(0.028)	(0.016)	(0.004)
2.00	α=0.01	0.324	0.520	0.768
		(0.030)	(0.032)	(0.027)
	α=0.05	0.528	0.696	0.924
		(0.032)	(0.029)	(0.017)
	α=0.1	0.600	0.796	0.956
		(0.031)	(0.025)	(0.013)
2.50	α=0.01	0.220	0.296	0.580
		(0.026)	(0.029)	(0.031)
	α=0.05	0.400	0.548	0.848
		(0.031)	(0.032)	(0.023)
	α=0.1	0.492	0.656	0.912
		(0.032)	(0.030)	(0.018)

where $L(\cdot)$ is the Laplace transform of the observation X. In the Monte Carlo experiments, we consider Gamma samples with $\theta < 1$ and $\lambda = 1$. We take 30% of censored observations. The value of μ to reach this proportion is given by

$$\left(1 + \frac{1}{\mu}\right)^{-\theta} = 0.7 \Rightarrow \mu = \frac{1}{(0.7)^{-1/\theta} - 1}.$$

Considering the results of the simulation study displayed in Table 5.5, we see that the power is higher in the censored case compared with complete Gamma samples (see Table 5.2). An intuitive explanation is as follows: the censoring mechanism tends to induce a higher value for the test statistic, which in turn facilitates the rejection of H_0.

6 Empirical illustrations

Before analyzing the data sets used in this paper to demonstrate the usefulness of the test procedure in practice, let us mention that in many cases the NBUE behavior is so clear (in the sense that the empirical expected shortfalls are dominated by the exponential ones for all probability levels) that there is no need for a formal testing procedure (the p-values are

Table 5.5: Proportion of samples where H_0 was rejected in the censored case (about 30% of censored observations, the censoring variable being exponentially distributed) for different Gamma distributions and for different sample sizes with 250 simulations and 1,000 bootstrap replications.

Gamma parameter θ	Significance level	Sample size N=25	N=50	N=100
0.25	α=0.01	0.964	0.972	1
		(0.012)	(0.010)	(0)
	α=0.05	0.996	0.992	1
		(0.004)	(0.006)	(0)
	α=0.1	0.996	0.992	1
		(0.004)	(0.006)	(0)
0.5	α=0.01	0.344	0.596	0.896
		(0.030)	(0.031)	(0.019)
	α=0.05	0.608	0.848	0.980
		(0.031)	(0.023)	(0.009)
	α=0.1	0.692	0.900	0.992
		(0.029)	(0.019)	(0.006)
0.75	α=0.01	0.160	0.240	0.420
		(0.023)	(0.027)	(0.031)
	α=0.05	0.312	0.484	0.664
		(0.029)	(0.032)	(0.030)
	α=0.1	0.416	0.576	0.724
		(0.031)	(0.031)	(0.028)

close or equal to unity). Therefore, we focus on cases where at least one crossing between the empirical and exponential expected shortfalls is observed.

6.1 CEO compensation data

Compensation data of CEOs of some of the largest companies in the United States are included in the Executive PayWatch database. This database is available online from http://www.aflcio.org/corporateamerica/paywatch/ceou/database.cfm. CEO compensations are listed for companies whose common stock comprises the Standard & Poor's Super 1500. We have randomly extracted 49 observations from the website database on March 8, 2005.

We have considered the total compensation, that is determined by adding the salary, bonus, other compensation, the value of restricted stock awards, long-term incentive payouts and the value of stock option awards in the fiscal year.

Figure 6.1 plots the (scaled) empirical shortfalls for the CEO compensation data (dotted line) together with the linear shortfall corresponding to the unit Negative Exponential distribution. Clearly, the empirical shortfalls dominate their Negative Exponential counterparts, except for probability levels near 1. A crossing between the two shortfall curves is visible on

Figure 6.1 for probability levels higher than 0.9.

We have performed the testing procedure to decide whether this crossing invalidate the shortfall dominance of the CEO compensation parent distribution over Negative Exponential (that is, the NWUE property). Note that we use here the NWUE version of the testing procedure. A p-value of 98.2% is obtained (on the basis of 1,000 bootstrap samples drawn from the original data set), so that the null hypothesis is not rejected (at any usual level).

Hence the CEO compensation data appear as NWUE. This indicates the "heavy-tailed behavior" of the CEO compensations. This conclusion is quite appealing. Indeed, the NWUE property ensures that $\mathbb{E}[X - t|X > t] \geq \mathbb{E}[X]$. On average, the excess of the CEO compensation above t is higher than the salary of the Standard & Poor's Super 1500, whatever t. This suggests that a large part of the total compensation is concentrated on payrolls of the best paid CEOs.

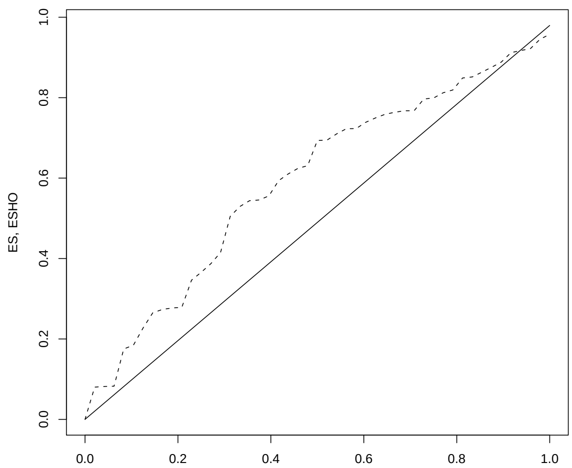

Figure 6.1: *Empirical shortfalls for the CEO compensation data set (broken line) and the corresponding linear Negative Exponential shortfalls (continuous line).*

6.2 Delays in European flights

The data set comes from Eurocontrol. It contains all the characteristics of each flight over Europe, day by day. In this paper, we consider September 5, 2004 (but the results are similar for other days). We measure the delay first as the difference (in minutes) between the real departure time and the scheduled one, and second as the difference (in kilometers) between the real route length and the initially planned one.

Figure 6.2 displays the graphs of the (scaled) empirical expected shortfalls for the two types of delay, together with the unit Negative Exponential benchmark. For the time delay, we have randomly sampled 800 delayed flights among the 14,475 flights delayed on September 5, 2004. For the kilometer delay, we have worked with the 1,201 flights with a longer route than initially planned. Considering the empirical shortfalls displayed in Figure 6.2, we expect that the delays before departure measured in time exhibit a NBUE behavior whereas the flight delays measured in kilometers exhibit a NWUE behavior. Therefore, we test the null hypothesis of NBUE for the time delay and the null hypothesis of NWUE for the kilometer delay, against the violation of these ageing properties. In each case, we have performed 1,000 bootstrap replications of the data set. We obtain a p-value of 59% for the time delay, and of 69.1% for the length delay supporting H_0 in both cases.

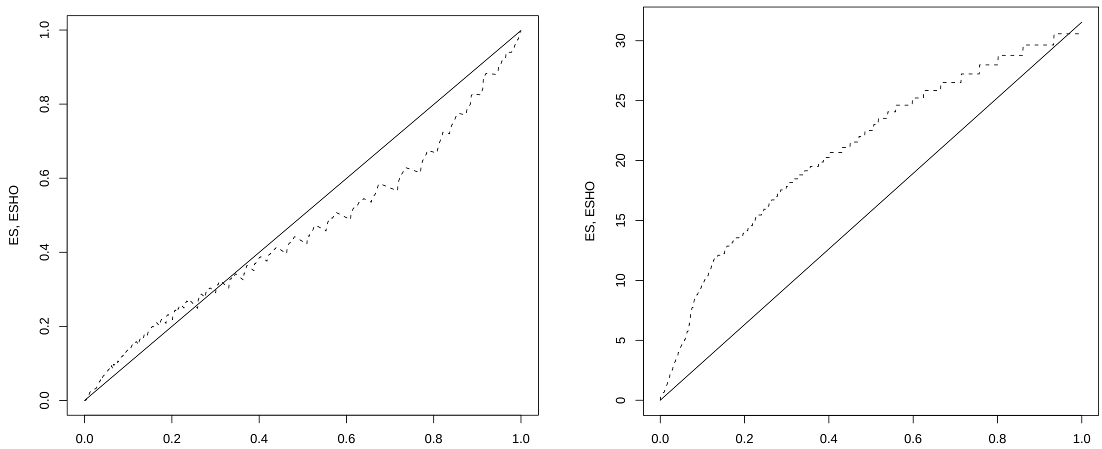

Figure 6.2: *Empirical shortfalls for the Eurocontrol data set, September 5, 2004 (broken line) for the delay measured in minutes (left panel) and in kilometers (right panel) and the corresponding linear Negative Exponential shortfalls (continuous line).*

We have then reached the somewhat surprising conclusion that when the delay before departure is measured in minutes, it exhibits a NBUE behavior, whereas when the delay during the flight is measured in kilometers, the data support the NWUE hypothesis. The NBUE behavior might be explained by the effort of catching up with the initial schedule by reducing lost time as much as possible. The NWUE might be explained by the initial planned route being, in general, the shortest one and the alternative available routes being relatively invariant in distance (fixed flight zones).

6.3 Throttle failures

We consider data on throttle failures in prototype models of general purpose load carrying vehicles taken from Blischke & Murthy (2000, Case 2.19). The data consist of failure times (measured in kilometers driven prior to failure) for 25 units and service times for 25 units that had not failed at the time of observation (and are thus censored).

Figure 6.3 displays the graph of the empirical expected shortfall for the throttle failures, together with the Negative Exponential ones. Even if the empirical expected shortfalls are below the Negative Exponential ones for most probability levels, we observe some crossings for moderate levels. The question is therefore whether these crossings invalidate NBUE. The

p-value obtained with 1,000 bootstrap replications is 24.1%, so that the null hypothesis of NBUE is not rejected.

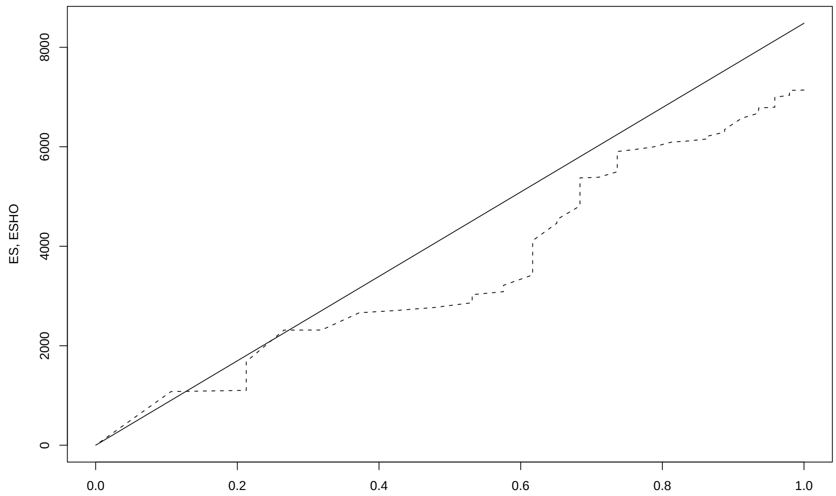

Figure 6.3: *Empirical shortfalls for the throttle failures data set (dotted line) and the corresponding linear Negative Exponential shortfalls (continuous line).*

7 Concluding remarks

In this paper, we introduce a method to test for the shortfall dominance against parametric alternatives, with primary interest in testing for NBUE/NWUE ageing classes. The test is of a Kolmogorov-Smirnov type, and distributional aspects of the test statistic for determining the rejection region are determined by a bootstrap technique. Asymptotic properties of the approach have been established, and finite sample performance is assessed through Monte Carlo experiments. Empirical applications illustrate the practical relevance of the approach followed in this paper.

It is worth mentioning that the procedure developed in this paper is applicable to any stochastic ordering defined by means of the pointwise comparison of transforms associated with the probability distributions. In particular, this applies to the moment generating function order among two non-negative random variables X and Y defined as

$$X \preceq_{MGF} Y \Leftrightarrow \mathbb{E}[\exp(tX)] \leq \mathbb{E}[\exp(tY)] \text{ for all } t > 0.$$

The order $\preceq_{MGF}$ can be tested using a Kolmogorov-Smirnov type test statistic based on empirical moment generating function functions. The theoretical properties of the empirical moment generating function process are derived in Csorgo (1982). It has been successfully applied in various testing procedures for exponentiality, e.g., by Baringhaus & Henze (1991,1992) and Henze (1993).

Note that such a Kolmogorov-Smirnov type test should outperform the procedure proposed by Klar (2005) to test for the so-called $\mathcal{M}$-class of life distributions. For other applications to testing for stochastic dominances, we refer the reader to Barrett & Donald (2003), Horvat, Kokoszka & Zitikis (2005) and the references therein. Testing procedures for stochastic dominances are also of interest in reliability applications since several ageing notions are defined with the help of such comparisons with respect to the Negative Exponential distribution. This is the case for instance with the ageing class known as HNBUE (for Harmonic New Better than Used in Expectation) consisting of the distributions larger than the exponential one in the second order stochastic dominance (or, equivalently, smaller than the exponential distribution in the convex order). For relationships between ageing classes and convex order, see, e.g., Ahmad, Hendi & Al-Nachawati (1999).

APPENDIX

All limits are taken as N goes to infinity.

A Proof of Lemma 3.2

The result is a direct consequence of the weak convergence of the empirical process $\sqrt{N}(\hat{F} - F)$, the Hadamard differentiability of the map $\mathcal{J}$, and the delta-method (see VW Section 3.9).

B Proof of Proposition 3.6

1. Proof of Part i):

From the definitions of $\hat{S}$ and the fact that under H_0, $\mathcal{J}(u; F) - \mathcal{J}(u; F_{\theta_0}) \leq 0$ for all $u \in (0, 1)$, we get that

$$\begin{aligned} \hat{S} &\leq \sup_u \sqrt{N}((\mathcal{J}(u; \hat{F}) - \mathcal{J}(u; F_{\hat{\theta}})) - (\mathcal{J}(u; F) - \mathcal{J}(u; F_{\theta_0}))) \\ &\qquad + \sup_u \sqrt{N}(\mathcal{J}(u; F) - \mathcal{J}(u; F_{\theta_0})) \\ &\leq \sup_u \sqrt{N}((\mathcal{J}(u; \hat{F}) - \mathcal{J}(u; F_{\hat{\theta}})) - (\mathcal{J}(u; F) - \mathcal{J}(u; F_{\theta_0}))). \end{aligned}$$

Hence the results follows from the weak convergence of $\sqrt{N}((\mathcal{J}(u; \hat{F}) - \mathcal{J}(u; F_{\hat{\theta}})) - (\mathcal{J}(u; F) - \mathcal{J}(u; F_{\theta_0})))$ induced by Lemmas 3.2 and 3.5, and the definition of $\bar{S}$.

2. Proof of Part ii):

If the alternative is true, then there is a u, say $\bar{u} \in (0, 1)$, for which $\mathcal{J}(\bar{u}; F) - \mathcal{J}(\bar{u}; F_{\theta_0}) = \delta > 0$. Then the result follows using the inequality $\hat{S} \geq \sqrt{N}(\mathcal{J}(\bar{u}; \hat{F}) - \mathcal{J}(\bar{u}; F_{\hat{\theta}}))$ and almost sure uniform convergence.

C Proof of Proposition 3.7

We know that $\sqrt{N}(\hat{F}^* - \hat{F})$ converges weakly to an independent copy of $\mathcal{B}_F \circ F$ (VW Theorem 3.6.3). Then Hadamard differentiability, via the delta-method for bootstrap (VW Theorem 3.9.11), and the CMT, yield that $\hat{S}^*$ converges in probabilility to a random variable, which is an independent copy of $\bar{S}$. Note that the distribution $P^0(t)$ of this random variable is absolutely continuous (Tsirel'son (1975)), while its median is strictly positive and finite. Moreover $c(\alpha)$ defined by $P(\bar{S} > c(\alpha)) = \alpha$ is finite and positive for any $\alpha < 1/2$ (VW Proposition A.2.7).

Note that the event $\{p^* < \alpha\}$ is equivalent to the event $\{\hat{S} > \hat{c}^*(\alpha)\}$ where

$$\inf\{t : \hat{P}^*(t) > 1 - \alpha\} = \hat{c}^*(\alpha) \xrightarrow{p} c(\alpha), \tag{C.1}$$

by the convergence of $\hat{S}^*$ and the aforementioned properties of P^0. Then:

$$\begin{aligned} \lim P[\text{reject } H_0|H_0] &= \lim P(\hat{S} > \hat{c}^*(\alpha)) \\ &= \lim P(\hat{S} > c(\alpha)) + \lim(P(\hat{S} > \hat{c}^*(\alpha)) - P(\hat{S} > c(\alpha))) \\ &\leq P(\bar{S} > c(\alpha)) := \alpha, \end{aligned}$$

where the last statement comes from (C.1), part *i*) of Proposition 3.6 and $c(\alpha)$ being a continuity point of the distribution of $\bar{S}$. On the other hand part *ii*) of Proposition 3.6 and $c(\alpha) < \infty$ ensure that $\lim P[\text{reject } H_0|H_1] = 1$.

References

[1] Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association* **97**, 284-292.

[2] Ahmad, I.A., Alwasel, I.A., & Mugdadi, A.R. (2001). A goodness-of-fit approach to major lifetesting problems. *International Journal of Reliability and Applications* **2**, 81-97.

[3] Ahmad, I.A., Hendi, M., & Al-Nachawati, H. (1999). Testing new better than used classes of life distribution derived from a convex ordering using kernel methods. *Journal of Nonparametric Statistics* **11**, 393-411.

[4] Andersen, P.K., Borgan, O., Gill, R., & Keiding, N. (1993). *Statistical Models based on Counting Processes.* Springer Verlag, New York.

[5] Baringhaus, L., & Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transforms. *Annals of the Institute of Statistical Mathematics* **43**, 551-564.

[6] Baringhaus, L., & Henze, N. (1992). An adaptive omnibus test for exponentiality. *Communications in Statistics: Theory and Methods* **21**, 969-978.

[7] Barlow, R.E., & Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing.* Holt, Rinehart and Winston, New York.

[8] Barrett, G., & Donald, S. (2003). Consistent tests for stochastic dominance. *Econometrica* **71**, 71-104.

[9] Belzunce, F., Pinar, J.F., & Ruiz, J.M. (2001). A family of tests for right spread order. *Statistics and Probability Letters* **54**, 79-92.

[10] Blischke, W.R., & Murthy, D.N.P. (2000). *Reliability: Modeling, Prediction and Optimization.* Wiley Series in Probability and Statistics, Wiley, New York.

[11] Borgan, O. (1984). Maximum likelihood estimation in parametric counting process models, with applications to censored failure time data. *Scandinavian Journal of Statistics* **11**, 1-16.

[12] Bryson, M.C., & Siddiqui, M.M. (1969). Some criteria for ageing. *Journal of the American Statistical Association* **64**, 1472-1482.

[13] Csorgo, S. (1982). The empirical moment generating function. In *Nonparametric Statistical Inference*, edited by B.V. Gnedenko, M.L. Puri & I. Vincze, Vol. 32, pp. 139-150, Coll. Math. Soc. J. Bolyai, Elsevier, Amsterdam.

[14] Doss, H., & Gill, R. (1992). An elementary approach to weak convergence for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association* **87**, 869-877.

[15] Fagiuoli, E., Pellerey, F., & Shaked, M. (1999). A characterization of the dilation order and its applications. *Statistical Papers* **40**, 393–406.

[16] Fernandez-Ponce, J. M., Infante-Macias, R., & Munoz-Perez, J. (1996). Characterization of lifetime distributions based on a quantile dispersion measure. *Computational Statistics and Data Analysis* **21**, 547–561.

[17] Fernandez-Ponce, J. M., Kochar, S. C., & Munoz-Perez, J. (1998). Partial orderings of distributions based on right-spread functions. *Journal of Applied Probability* **35**, 221–228.

[18] Gill, R. (1983). Large sample behaviour of the product-limit estimator on the whole line. *Annals of Statistics* **11**, 49-58.

[19] Henze, N. (1993). A new flexible class of omnibus tests for exponentiality. *Communications in Statistics: Theory and Methods* **22**, 115-133.

[20] Hollander, M., & Proschan, F. (1975). Test for the mean residual life. *Biometrika* **62**, 585-592.

[21] Hollander, M., & Proschan, F. (1980). Test for the mean residual life. Amendments and corrections. *Biometrika* **67**, 259.

[22] Horvath, L., Kokoszka, P., & Zitikis, R. (2005). Testing for stochastic dominance using the weighted McFadden-type statistic. *Journal of Econometrics*, in press.

[23] Klar, B. (2005). Tests for exponentiality against the $\mathcal{M}$ and $\mathcal{LM}$-classes of life distributions. *Test* **14**, 1-23.

[24] Klefsjö, B. (1994). Letter to the Editor on “An exact test for NBUE class of survival functions”. *Communications in Statistics: Theory and Methods* **23**, 2417-2421.

[25] Koul, H.L., & Susarla, V. (1980). Testing for new better than used in expectation with incomplete data. *Journal of the American Statistical Association* **75**, 952-956.

[26] Kochar, S.C., Li, X., & Shaked, M. (2002). The total time on test transform and the excess wealth stochastic orders of distributions. *Advances in Applied Probability* **34**, 826-845.

[27] Lehmann, E. (1986). *Testing Statistical Hypotheses.* Wiley & Sons, New York.

[28] Pham, T.G., & Turkkan, N. (1994). The Lorenz and the scaled total-time-on-test transform curves: a unified approach. *IEEE Transactions on Reliability* **43**, 76-84.

[29] Politis, D., Romano, J., & Wolf, M. (1999). *Subsampling.* Springer Verlag, New York.

[30] Rausand, M, & Hoyland, A. (2004). *System Reliability Theory.* Wiley, New York.

[31] Shaked, M., & Shantikumar, J.G. (1994). *Stochastic Orders and Their Applications.* Academic Press, New York.

[32] Shaked, M., & Shanthikumar, J.G. (1998). Two variability orders. *Probability in the Engineering and Informational Sciences* **12**, 1–23.

[33] Tsirel'son, V. (1975). The density of the distribution of the maximum of a gaussian process. *Theory of Probability and their Applications* **16**, 847-856.

[34] van der Vaart, A., & Wellner, J. (1996). *Weak Convergence and Empirical Processes.* Springer Verlag, New York.