

# An Exact Test for Hazard-Similarity

Kais Hamza<sup>1</sup>

Monash University

<sup>1</sup>Address: School of Mathematical Sciences, Monash University, VIC  
3800, Australia. Phone: (03) 99054453. Fax: (03) 9905 9520. Email:  
Kais.Hamza@sci.monash.edu.au

**Summary.** We develop an exact test for hazard-similarity and in particular for exponentiality. This test is distinct from more common goodness of fit tests such as the Kolmogorov-Smirnov goodness of fit test, as it does not require full specification of the null distribution. We obtain this test through a characterization of hazard-similar distributions and a generalization of Fisher's test for association.

**KEYWORDS:** Exponential distribution, Fisher's exact test, hazard-similar distributions, test for exponentiality. ■

# 1 Introduction

The prominence of the exponential distribution is probably only second to the normal distribution. Being able to test for exponentiality becomes an essential ingredient in many statistical studies, in particular in the areas of survival analysis and reliability theory. We stress here that a test for exponentiality is one that does not prescribe the value of the parameter  $\lambda$  of the exponential distribution under the null hypothesis.

While many studies have access to large data sets, many others are limited to small samples, and being able to carry out a test on a small sample is often one of the challenges statisticians have to face.

The survey papers of Ascher (1990) and, Henze and Meintanis (2005) examine a wide range of statistical tests for exponentiality. They range from the classical Kolmogorov-Smirnov test, to Gnedenko's F-test, to more modern approaches such as the test based on the empirical Laplace transform introduced by Henze and Meintanis (2002).

With the exception of Gnedenko's F-test (and its variants), all other tests are based on a limiting result, and are only of practical use in the case of large samples.

In this paper, we develop an alternate test for exponentiality. As for other so-called scale invariant tests, this test allows us to check whether the data at hand come from an exponential distribution without having to specify its parameter or estimating it. Unlike many other tests, such as the chi-

square goodness of fit test, it is an exact test, based on a precise theoretical distributional statement. Its use is therefore perfectly warranted in situations where most other tests fail, the case of small samples.

This test for exponentiality is derived in Section 4. It is obtained by developing a generalization of Fisher's exact test for association (Section 3) and uncovering, as far as we know, a previously undocumented characterization of exponential distributions (Section 2). In fact this characterization is not restricted to the exponential family, it holds true for families of hazard-similar distributions. Naturally, the test for exponentiality extends to become a test for hazard-similarity.

## **2 Hazard-similar distributions**

Hazard functions play an important role in survival analysis, reliability theory, and many other areas of applied probability and statistics. As a model for age, the hazard function describes the distribution conditional on survival up to a given time.

It is, for example, well known that exponential random variables have constant hazard functions, and that as such exponential random variables are well-suited to model survival-independent ageing.

While the notion of hazard function can be extended to general distributions – see for example Kotz and Shanbhag (1980), we shall focus our attention on absolutely continuous distributions.

The hazard function of an absolutely continuous distribution is the ratio of its density to the upper tail.

**Definition 1** *The hazard function of a distribution  $\mathbb{D}$  with density  $f$ , distribution function  $F$  and upper bound  $\delta \doteq \inf\{x : F(x) = 1\}$  ( $\inf \emptyset = +\infty$ ) is*

$$h(x) = \frac{f(x)}{1 - F(x)} \quad (1)$$

for  $x < \delta$ , and say, 0 everywhere else.

It is also well known that hazard functions determine their distributions; in fact we have

$$F(x) = 1 - \exp\left(-\int_{-\infty}^x h(u)du\right). \quad (2)$$

It immediately follows that exponential random variables are the only positive random variables with constant hazard functions.

Throughout this section, we use  $h$ ,  $f$ ,  $F$ ,  $\gamma$  and  $\delta$  indexed in various ways to denote respectively the hazard, density, distribution function, lower bound ( $\gamma \doteq \sup\{x : F(x) = 0\}$ ,  $\sup \emptyset = -\infty$ ) and upper bound of a distribution  $\mathbb{D}$ . Next, we introduce the key concept in this paper.

**Definition 2** *Two distributions,  $\mathbb{D}_1$  and  $\mathbb{D}_2$ , are said to be hazard-similar if their hazard functions differ by a multiplicative constant; that is one can find  $k > 0$  such that*

$$\forall u, h_2(u) = kh_1(u). \quad (3)$$

*Two random variables are said to be hazard-similar if their distributions are hazard-similar.*

Obviously any two identically distributed random variables are hazard-similar. Also, any two exponential random variables are necessarily hazard-similar. Rayleigh random variables, for which

$$\forall x > 0, f(x) = 2\alpha x e^{-\alpha x^2} \text{ and } h(x) = 2\alpha x,$$

constitute another example of a family of hazard-similar random variables. In general, starting from any “baseline” distribution with hazard function  $h(u)$ , the family of distributions whose hazard functions are  $kh(u)$ ,  $k > 0$ , are all hazard-similar. For example, starting from a standard normal distribution, we construct the family of so-called power normal distributions. Their hazard functions are:

$$h(x) = \frac{k\phi(x)}{1 - \Phi(x)},$$

where  $\phi$  and  $\Phi$  are the density and distribution function of the standard normal distribution.

Proportional hazards are in fact the starting point of Cox’s survival regression – see for example Cox and Oakes (1984), where the constant of proportionality  $k$  is modelled as a (log-linear) function of the explanatory variables under investigation.

The term “hazard-similar” is, of course, technically equivalent to Cox’s description of distributions satisfying (3) as proportional hazards. However, as we are not interested in estimating  $k$  or investigating its dependence on other variables, and rather use  $k$  to represent an “arbitrary” constant within a class of distributions we do not wish to discriminate, we feel that characterizing distributions satisfying (3) as hazard-similar is more appropriate.

Using (2), we see that hazard-similar families can also be constructed by “power-twisting” their distribution functions.

**Proposition 3**  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are hazard-similar if and only if there exists  $k > 0$  such that

$$\forall u, F_2(u) = 1 - (1 - F_1(u))^k. \quad (4)$$

**Corollary 4**  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are hazard-similar if and only if (3) or (4) are satisfied for  $\max(\gamma_1, \gamma_2) \leq u \leq \min(\delta_1, \delta_2)$ . Furthermore, if  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are hazard-similar then  $\gamma_1 = \gamma_2$  and  $\delta_1 = \delta_2$ .

Now, it is well known that if  $Y$  is distributed as the minimum of a random sample of integer size  $k$  on  $X$ , then

$$\forall u, F_Y(u) = 1 - (1 - F_X(u))^k.$$

In other words  $X$  and its sample minimums are always hazard-similar. The concept of hazard-similarity is therefore closely related to that of the distribution of sample minimums. The following definition generalizes this notion.

**Definition 5** Let  $\mathbb{D}$  be a distribution and  $k > 0$ . We call fractional sample minimum (FSM) a distribution whose distribution function is given by  $1 - (1 - F(y))^k$ . We denote such a distribution  $\min_k(\mathbb{D})$ .

Since clearly  $\min_k(\min_l(\mathbb{D})) = \min_{k \times l}(\mathbb{D})$ , we have  $\min_k(\min_{1/k}(\mathbb{D})) = \mathbb{D}$ , and we obtain an interpretation of  $\min_{1/k}(\mathbb{D})$  as the distribution the  $k$ -sample minimum of which is  $\mathbb{D}$ .

Observe that  $\min_k(\mathbb{D})$  stochastically dominates  $\min_l(\mathbb{D})$ , for  $k < l$ . Also, as  $k \uparrow \infty$ ,  $\min_k(\mathbb{D})$  approaches  $\gamma$  in distribution, and therefore in probability. Likewise, as  $k \downarrow 0$ ,  $\min_k(\mathbb{D})$  approaches  $\delta$ .

Hazard-similar distributions are those distributions that can be obtained by taking FSMs of each other. Of course, if  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are not hazard-similar, then there is no FSM that turns  $\mathbb{D}_1$  into  $\mathbb{D}_2$  or vice versa. We are therefore lead to look for the “best approximating”  $k$  instead; that is a  $k$  such that  $\min_k(\mathbb{D}_1)$  is as close as possible to  $\mathbb{D}_2$ , in some sense yet to be defined. Before we make these concepts precise, we look at pairs of distributions essentially unaffected by any amount of power-twisting. In fact, for these distributions the optimum  $k$  is either nil or infinite.

**Definition 6** Two distributions,  $\mathbb{D}_1$  and  $\mathbb{D}_2$ , are said to be completely hazard-dissimilar if for any two independent random variables, say  $X_1$  and  $X_2$  with distributions  $\mathbb{D}_1$  and  $\mathbb{D}_2$ , respectively,  $\Pr(X_1 < X_2)$  equals 0 or 1.

Two random variables are said to be completely hazard-dissimilar if their distributions are completely hazard-dissimilar.

Notice that since  $X_1$  and  $X_2$  are independent,

$$\Pr(X_1 < X_2) = \int_{-\infty}^{+\infty} [1 - F_2(x)] f_1(x) dx,$$

and this quantity vanishes if and only if  $(1 - F_2)f_1 \equiv 0$ . This in turn implies that  $\delta_2$  must be finite and that  $F_1(\delta_2) = 0$ . In other words, we must have  $\delta_2 \leq \gamma_1$ . This immediately leads to the following characterization of completely hazard-dissimilar distributions.

**Proposition 7**  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are completely hazard-dissimilar if and only if either  $\gamma_1 \leq \delta_1 \leq \gamma_2 \leq \delta_2$  or  $\gamma_2 \leq \delta_2 \leq \gamma_1 \leq \delta_1$ .

The terminology “ $\mathbb{D}_1$  and  $\mathbb{D}_2$  are completely hazard-dissimilar” is therefore understood to signify that if  $\mathbb{D}_1$  dominates  $\mathbb{D}_2$ , then for any  $k, l > 0$ ,  $\min_k(\mathbb{D}_1)$  dominates  $\min_l(\mathbb{D}_2)$ . In other words,  $\mathbb{D}_1$  and  $\mathbb{D}_2$  bear no resemblance to each other.

We are now ready to introduce the hazard-similarity scale alluded to above. For any distributions  $\mathbb{D}_1$  and  $\mathbb{D}_2$ , we define

$$\mathcal{H}(\mathbb{D}_1, \mathbb{D}_2) = \min_{k,l} \int_{-\infty}^{+\infty} |(1 - F_2(u))^l - (1 - F_1(u))^k| du. \quad (5)$$

**Proposition 8** If  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are hazard-similar then  $\mathcal{H}(\mathbb{D}_1, \mathbb{D}_2) = 0$ . If  $\mathbb{D}_1$  and  $\mathbb{D}_2$  are completely hazard-dissimilar, then  $\mathcal{H}(\mathbb{D}_1, \mathbb{D}_2) = \max(\gamma_1 - \delta_2, \gamma_2 - \delta_1)$ .

**Proof.** The first statement immediately follows from Corollary 4. To prove

the second one, assume that  $X_1$  and  $X_2$  are independent random variables with distributions  $\mathbb{D}_1$  and  $\mathbb{D}_2$ , respectively, such that  $X_1$  stochastically dominates  $X_2$ ; the alternative is shown in an identical manner. Proposition 7 yields  $\delta_2 \leq \gamma_1$ . It follows that

$$\begin{aligned}
& \int_{-\infty}^{+\infty} |(1 - F_2(u))^l - (1 - F_1(u))^k| du \\
&= \int_{-\infty}^{\delta_2} [1 - (1 - F_2(u))^l] du + \gamma_1 - \delta_2 + \int_{\gamma_1}^{+\infty} (1 - F_1(u))^k du \\
&= \int_{-\infty}^{\delta_2} F_{\min_l(\mathbb{D}_2)}(u) du + \gamma_1 - \delta_2 + \int_{\gamma_1}^{+\infty} [1 - F_{\min_k(\mathbb{D}_1)}(u)] du \\
&= -\mathbb{E}(\min_l(\mathbb{D}_2)) + \mathbb{E}(\min_k(\mathbb{D}_1))
\end{aligned}$$

where we have used the facts that, if  $\gamma > -\infty$  then  $\int_{\gamma}^{+\infty} (1 - F(u)) du = \mathbb{E}(\mathbb{D}) - \gamma$ , and if  $\delta < +\infty$ , then  $\int_{-\infty}^{\delta} F(u) du = \delta - \mathbb{E}(\mathbb{D})$ , and where we have loosely used  $\mathbb{E}(\mathbb{D})$  to denote the mean of the distribution  $\mathbb{D}$ .

Since  $\min_l(\mathbb{D}_2)$  is a stochastically monotone sequence that converges, as  $l \downarrow 0$ , to  $\delta_2$ , and  $\min_k(\mathbb{D}_1)$  is also stochastically monotone and converges, as  $k \uparrow \infty$ , to  $\gamma_1$ , then  $\mathcal{H}(\mathbb{D}_1, \mathbb{D}_2) = \min_{k,l} (-\mathbb{E}(\min_l(\mathbb{D}_2)) + \mathbb{E}(\min_k(\mathbb{D}_1))) = \gamma_1 - \delta_2$ . ■

**Remark 9** *Although we focus on absolutely continuous distributions, (5) extends to general random variables. In particular, we can compute the hazard-similarity scale of constants. It is simply their Euclidean distance: for any pair of constants  $(x, y)$ ,  $\mathcal{H}(x, y) = |x - y|$ .*

Apart from the cases detailed in Proposition 8, hazard-similarity scales are

quite tedious to compute. In most cases, numerical solutions are all one can hope for.

The next result identifies a property of hazard-similar distributions which will be used in Section 4 to construct a test for hazard-similarity.

**Theorem 10** *Let  $X$  and  $Y$  be two independent random variables.  $Z = \min(X, Y)$  is independent of  $A = \{X < Y\}$  if and only if  $X$  and  $Y$  are either hazard-similar or completely hazard-dissimilar.*

**Proof.** First observe that

$$\begin{aligned} f_Z(z) &= f_X(z) [1 - F_Y(z)] + f_Y(z) [1 - F_X(z)] \\ &= h_X(z) [1 - F_X(z)] [1 - F_Y(z)] + h_Y(z) [1 - F_Y(z)] [1 - F_X(z)] \\ &= (h_X(z) + h_Y(z)) [1 - F_X(z)] [1 - F_Y(z)] \end{aligned}$$

and for any  $g$  for which the following expectation exists,

$$\mathbb{E}(g(Z)) = \int_{-\infty}^{+\infty} g(z) [1 - F_Y(z)] [1 - F_X(z)] [h_X(z) + h_Y(z)] dz.$$

Using the independence of  $X$  and  $Y$ , we write

$$\begin{aligned}
\mathbb{E}(g(Z)1_A) &= \mathbb{E}(\mathbb{E}(g(Z)1_A | X)) = \mathbb{E}(g(X)\Pr(X < Y | X)) \\
&= \int_{-\infty}^{+\infty} g(x)\Pr(X < Y | X = x) f_X(x)dx \\
&= \int_{-\infty}^{+\infty} g(x)\Pr(x < Y | X = x) f_X(x)dx \\
&= \int_{-\infty}^{+\infty} g(x) [1 - F_Y(x)] f_X(x)dx \\
&= \int_{-\infty}^{+\infty} g(x) [1 - F_Y(x)] [1 - F_X(x)] h_X(x)dx
\end{aligned}$$

For  $Z = \min(X, Y)$  to be independent of  $A = \{X < Y\}$ , we must have, for any bounded  $g$ ,  $\mathbb{E}(g(Z)1_A) = \mathbb{E}(g(Z))\Pr(X < Y)$ , that is

$$\begin{aligned}
&\int_{-\infty}^{+\infty} g(x) [1 - F_Y(x)] [1 - F_X(x)] h_X(x)dx \\
&= \Pr(X < Y) \int_{-\infty}^{+\infty} g(z) [1 - F_Y(z)] [1 - F_X(z)] [h_X(z) + h_Y(z)] dz
\end{aligned}$$

It follows that a necessary and sufficient condition for  $Z$  and  $A$  to be independent is

$$\begin{aligned}
&\Pr(X > Y) [1 - F_Y(x)] [1 - F_X(x)] h_X(x) \\
&= \Pr(X < Y) [1 - F_Y(x)] [1 - F_X(x)] h_Y(x).
\end{aligned}$$

The result follows immediately from Corollary 4. ■

**Corollary 11** *If  $X$  and  $Y$  are hazard-similar, then the constant of propor-*

tionality  $k \doteq \frac{h_Y(u)}{h_X(u)}$  is the odds number

$$k = \frac{\Pr(Y > X)}{\Pr(Y < X)}.$$

In view of Theorem 10, an alternative measure of hazard-similarity is obtained by applying the following measure of dependence to  $A = \{X < Y\}$  and  $Z = X \wedge Y$ .

For an event  $A$ , with  $0 < \Pr(A) < 1$ , and a random variable  $Z$ ,  $\mathcal{D}(A, Z) = \frac{\text{var}(\Pr(A|Z))}{\Pr(A)(1 - \Pr(A))}$  measures the dependence of  $A$  and  $Z$  in the sense that

1.  $\mathcal{D}(A, Z) \in [0, 1]$ ;
2.  $\mathcal{D}(A, Z) = 0$  if and only if  $A$  and  $Z$  are independent;
3.  $\mathcal{D}(A, Z) = 1$  if and only if  $A$  is a (measurable) function of  $Z$ .

If  $Z$  is an event  $B$ ,  $\mathcal{D}(A, B) = \frac{(\Pr(A \cap B)\Pr(B^c) - \Pr(A \cap B^c)\Pr(B))^2}{\Pr(A)\Pr(A^c)\Pr(B)\Pr(B^c)}$  and  $\mathcal{D}(A, B) = 1$  if and only if  $A = B$  or  $A = B^c$ . Note that  $\mathcal{D}(A, B)$  differs from the square of the correlation of  $1_A$  and  $1_B$ .

Now let  $\mathbb{D}$  and  $\mathbb{D}'$  be two non-completely hazard-dissimilar distributions. Then, with  $X$  and  $X'$  independent and distributed like  $\mathbb{D}$  and  $\mathbb{D}'$  respectively,

$$\mathcal{K}(\mathbb{D}, \mathbb{D}') = \mathcal{D}(X < X', X \wedge X') = \frac{\text{var}(\Pr(X < X'|X \wedge X'))}{\Pr(X < X')\Pr(X \geq X')}$$

defines a measure of hazard-similarity in the sense that  $\mathcal{K}(\mathbb{D}, \mathbb{D}') = 0$  if and only if  $\mathbb{D}$  and  $\mathbb{D}'$  are hazard-similar. Note that, for two non-completely

hazard-dissimilar distributions  $\mathcal{K}(\mathbb{D}, \mathbb{D}') < 1$ . Indeed,  $\text{var}(\Pr(X < X'|X \wedge X')) = \Pr(X < X')\Pr(X \geq X')$  if and only if  $\{X < X'\}$  is a (measurable) function of  $X \wedge X'$ . In other words, there exists a (measurable) set  $C$  such that  $X < X' \iff X \wedge X' \in C$  (a.s.). It immediately follows that  $\mathbb{D}$  and  $\mathbb{D}'$  must be completely hazard-dissimilar.

To compute explicitly  $\mathcal{K}(\mathbb{D}, \mathbb{D}')$ , we observe that, for any positive  $\phi$ ,

$$\int \phi(z)[1 - F_{X'}(z)]f_X(z)dz = \int \phi(z)\Pr(X < X'|X \wedge X' = z)f_{X \wedge X'}(z)dz.$$

It immediately follows that, whenever  $f_{X \wedge X'}(z) \neq 0$ ,

$$\Pr(X < X'|X \wedge X' = z) = \frac{(1 - F_{X'}(z))f_X(z)}{f_{X \wedge X'}(z)}.$$

**Example 12** Let  $\mathbb{D}_\alpha$ ,  $\alpha \in [0, 1]$ , be the continuous uniform distribution over the interval  $[0, \alpha]$ . Then, for  $0 < \alpha < 1$ ,

$$\mathcal{K}(\mathbb{D}_1, \mathbb{D}_\alpha) = \frac{(1 - \alpha)^2[\ln(1 + \alpha) - \ln(1 - \alpha)] - 2\alpha(1 - 2\alpha)}{2(2 - \alpha)\alpha^2}.$$

Note that the degenerate case  $\alpha = 0$  corresponds to a pair of completely hazard-dissimilar random variables ( $\lim_{\alpha \downarrow 0} \mathcal{K}(\mathbb{D}_1, \mathbb{D}_\alpha) = 0$ ), while the limiting case  $\alpha = 1$  corresponds to a pair of identically distributed (and thus hazard-similar) random variables ( $\lim_{\alpha \uparrow 1} \mathcal{K}(\mathbb{D}_1, \mathbb{D}_\alpha) = 1$ ).

### 3 An exact test for independence

In this section, we develop an exact non-parametric test for the independence of a random variable  $Z$  and an event  $A$ . The approach is very similar to (and generalizes) Fisher's exact test for association.

Let  $\xi = 1_A$  and  $(Z_1, \xi_1), \dots, (Z_n, \xi_n)$  be a random sample on  $(Z, \xi)$ . We introduce the following notation. Let  $F(z) = \Pr(Z \leq z)$ ,  $p = \Pr(A)$  and,  $\nu(z)$ ,  $N(z)$  and  $M$  be the counts as described by the contingency Table 1.

	$A$	$A^c$	
$\{Z \leq z\}$	$\nu(z)$	$N(z) - \nu(z)$	$N(z)$
$\{Z > z\}$	$M - \nu(z)$	$n - N(z) - M + \nu(z)$	$n - N(z)$
	$M$	$n - M$	$n$

Table 1: Cross-tabulation of counts

Under the assumption of independence of  $Z$  and  $A$ ,  $\{Z \leq z\}$  and  $A$  are independent for each  $z$ . It follows that  $(\nu(z), N(z) - \nu(z), M - \nu(z))$  has a multinomial distribution with parameters  $n$  and  $(pF(z), (1 - p)F(z), p(1 - F(z)))$ , that the conditional distribution of  $\nu(z)$  given  $N(z)$  and  $M$ , is hypergeometric with parameters  $N(z)$ ,  $M$  and  $n$ , and that  $E(\nu(z)|N(z), M) = N(z)\frac{M}{n}$ .

As for Fisher's exact test, the principle of the statistical test we are about to develop, is that any significant departure of  $\nu(z)$  from its conditional expectation would lead to the assumption of independence to be deemed unsustainable. We shall adopt a uniform approach to our definition of departure

and introduce the following statistic

$$\begin{aligned} T &= \sup_z \left| \nu(z) - N(z) \frac{M}{n} \right| \\ &= \sup_{1 \leq k \leq n} \left| \nu(Z_{(k)}) - k \frac{M}{n} \right| \end{aligned}$$

where, as usual,  $Z_{(1)}, \dots, Z_{(n)}$  denote the order statistics of the sequence  $Z_1, \dots, Z_n$ , and where we have used the fact that  $N(Z_{(k)}) = k$ .

**Proposition 13** *Under the null hypothesis that  $A$  and  $Z$  are independent,*

$$\Pr(T > c | M = m, N(z), \forall z) = 1 - \frac{1}{\binom{n}{m}} \times W_{n,c}(m),$$

where  $W_{n,c}(m)$  is the number of sequences  $(\varepsilon_1, \dots, \varepsilon_n) \in \{0, 1\}^n$  such that  $\varepsilon_1 + \dots + \varepsilon_n = m$  and  $\forall k \leq n, km/n - c \leq \sum_{i=1}^k \varepsilon_i \leq km/n + c$ .

**Proof.** First, we order  $\xi_1, \dots, \xi_n$  according to the ranking of  $Z_1, \dots, Z_n$ , and denote the resulting sequence  $\xi_1^*, \dots, \xi_n^*$ . Then  $\nu(Z_{(k)}) = \sum_{i=1}^k \xi_i^*$ ,  $\xi_1^* = \nu(Z_{(1)})$  and  $\xi_k^* = \nu(Z_{(k)}) - \nu(Z_{(k-1)})$ ,  $k = 2, \dots, n$ .

Furthermore, for  $\varepsilon_1, \dots, \varepsilon_n \in \{0, 1\}$  such that  $\varepsilon_1 + \dots + \varepsilon_n = m$ ,

$$\begin{aligned} &\Pr(\xi_1^* = \varepsilon_1 \dots \xi_n^* = \varepsilon_n | M = m, N(z), \forall z) \\ &= \Pr(\xi_1^* = \varepsilon_1 \dots \xi_n^* = \varepsilon_n | M = m, Z_{(1)}, \dots, Z_{(n)}) \end{aligned}$$

which, under the assumption of independence of  $\{Z_1, \dots, Z_n\}$  and  $\{\xi_1, \dots, \xi_n\}$

yields

$$\begin{aligned}
& \Pr(\xi_1^* = \varepsilon_1, \dots, \xi_n^* = \varepsilon_n | M = m, N(z), \forall z) \\
&= \frac{\Pr(\xi_1^* = \varepsilon_1, \dots, \xi_n^* = \varepsilon_n, M = m | Z_{(1)}, \dots, Z_{(n)})}{\Pr(M = m)} \\
&= \frac{\Pr(\xi_1^* = \varepsilon_1, \dots, \xi_n^* = \varepsilon_n | Z_{(1)}, \dots, Z_{(n)})}{\Pr(M = m)} \\
&= \frac{1}{\binom{n}{m} p^m (1-p)^{n-m}} \Pr(\xi_1^* = \varepsilon_1, \dots, \xi_n^* = \varepsilon_n | Z_{(1)}, \dots, Z_{(n)})
\end{aligned}$$

Let  $\sigma$  be any permutation of  $\{1, \dots, n\}$ . Then, under the assumption of independence of  $\{Z_1, \dots, Z_n\}$  and  $\{\xi_1, \dots, \xi_n\}$ ,

$$(\xi_{\sigma(1)}, \dots, \xi_{\sigma(n)}, Z_{(1)}, \dots, Z_{(n)}) \stackrel{d}{=} (\xi_1, \dots, \xi_n, Z_{(1)}, \dots, Z_{(n)}).$$

Therefore for any random permutation  $\sigma$  independent of  $(\xi_1, \dots, \xi_n)$  and  $(Z_{(1)}, \dots, Z_{(n)})$ ,

$$(\xi_{\sigma(1)}, \dots, \xi_{\sigma(n)}, Z_{(1)}, \dots, Z_{(n)}) \stackrel{d}{=} (\xi_1, \dots, \xi_n, Z_{(1)}, \dots, Z_{(n)}).$$

Now  $(\xi_1^*, \dots, \xi_n^*)$  is the result of applying to  $(\xi_1, \dots, \xi_n)$  the (random) permutation obtained from the ordering of  $(Z_1, \dots, Z_n)$ . This permutation is known to be independent of the order statistics  $(Z_{(1)}, \dots, Z_{(n)})$ . It follows that

$$(\xi_1^*, \dots, \xi_n^*, Z_{(1)}, \dots, Z_{(n)}) \stackrel{d}{=} (\xi_1, \dots, \xi_n, Z_{(1)}, \dots, Z_{(n)})$$

and

$$\begin{aligned}
& \Pr(\xi_1^* = \varepsilon_1 \dots \xi_n^* = \varepsilon_n | M = m, N(z), \forall z) \\
&= \frac{1}{\binom{n}{m} p^m (1-p)^{n-m}} \Pr(\xi_1^* = \varepsilon_1, \dots, \xi_n^* = \varepsilon_n | Z_{(1)}, \dots, Z_{(n)}) \\
&= \frac{1}{\binom{n}{m} p^m (1-p)^{n-m}} \Pr(\xi_1 = \varepsilon_1, \dots, \xi_n = \varepsilon_n | Z_{(1)}, \dots, Z_{(n)}) \\
&= \frac{1}{\binom{n}{m} p^m (1-p)^{n-m}} \Pr(\xi_1 = \varepsilon_1, \dots, \xi_n = \varepsilon_n) \\
&= \frac{1}{\binom{n}{m}}.
\end{aligned}$$

This allows us to write

$$\begin{aligned}
& \Pr(T > c | M = m, N(z), \forall z) \\
&= \Pr\left(\sup_{k \leq n} |\nu(Z_{(k)}) - km/n| > c \mid M = m, N(z), \forall z\right) \\
&= 1 - \Pr(\forall k \leq n, |\nu(Z_{(k)}) - km/n| \leq c \mid M = m, N(z), \forall z) \\
&= 1 - \Pr(\forall k \leq n, km/n - c \leq \nu(Z_{(k)}) \leq km/n + c \mid M = m, N(z), \forall z) \\
&= 1 - \frac{1}{\binom{n}{m}} \times W_{n,c}(m)
\end{aligned}$$

where  $W_{n,c}(m)$  is as described in the statement. ■

Next, we display some basic properties of the conditional distribution of  $T$ .

**Proposition 14** *Under the null hypothesis that  $A$  and  $Z$  are independent,*

1. conditional on  $M$ ,  $T$  and  $(Z_1, \dots, Z_n)$  are independent and

$$\Pr(T > c|M = m) = 1 - \frac{1}{\binom{n}{m}} \times W_{n,c}(m);$$

2. the conditional distribution of  $T$  given  $M = m$  equals that of  $T$  given  $M = n - m$ ;

3.  $\Pr(T = 0|M = 0) = \Pr(T = 0|M = n) = 1$ ;

4. for  $m = 1, \dots, n - 1$ ,  $\Pr(T > 0|M = m) = 1$ ;

5. for  $m = 1, \dots, n - 1$ , given  $M = m$ , the largest possible value of  $T$  is  $m(n - m)/n$ .

**Proof.** 1–4 are obvious. To prove 5, we note that the largest possible value of  $T$  is obtained by the two most “extreme” sequences  $\varepsilon_1 = \dots = \varepsilon_m = 1$ ,  $\varepsilon_{m+1} = \dots = \varepsilon_n = 0$ , and  $\varepsilon_1 = \dots = \varepsilon_{n-m} = 0$ ,  $\varepsilon_{n-m+1} = \dots = \varepsilon_n = 1$ . For these,  $T$  equals  $m(n - m)/n$ . That is  $\Pr(T \leq m(n - m)/n|M = m) = 1$  and  $\Pr(T = m(n - m)/n|M = m) = \frac{2}{\binom{n}{m}}$ . ■

When testing for the dependence between an event,  $A$ , and a scale variable,  $Z$ , it is then natural to reject the null hypothesis of  $A$  and  $Z$  being independent if the observed  $T$  exceeds the critical value of the conditional distribution of  $T$  given the observed marginals,  $M$  and  $\{N(z), \forall z\}$ .

We shall refer to this statistical test as the Extended Fisher Test (EFT).

**Example 15 (Eye Colour and Flicker Frequency)** *We apply the EFT to a data set (given in Table 2) of critical flicker frequency and iris colour of the eye for  $n = 19$  subjects (Source: OzDASL). An EFT for the dependence of*

Flicker	Colour	Flicker	Colour	Flicker	Colour
26.8	brown	24.5	brown	27.2	blue
27.9	brown	26.4	green	29.9	blue
23.7	brown	24.2	green	28.5	blue
25.0	brown	28.0	green	29.4	blue
26.3	brown	26.9	green	28.3	blue
24.8	brown	29.1	green		
25.7	brown	25.7	blue		

Table 2: Eye Colour and Flicker Frequency data

*the event “brown eyes” on the critical flicker frequency produces an observed  $M$  of 8 and an observed  $T$  of 53/19. It follows that the EFT has a  $P$ -value of 0.0468. On the other hand the  $P$ -value for testing the independence of the event “green eyes” on the critical flicker frequency is 0.8108 ( $m = 5$  and  $t = 21/19$ ).*

We end this section with a simulation-based comparative study of the power of the EFT. Undoubtedly, the most popular test for the dependence between a binary variable,  $\xi$ , and a scale variable,  $Z$ , is that based on the logistic regression model:

$$\Pr(\xi = 1|Z) = \frac{e^{\beta_0 + \beta_1 Z}}{1 + e^{\beta_0 + \beta_1 Z}}. \quad (6)$$

A test for independence becomes a test for  $\beta_1 = 0$ . We refer to this test as the Logistic Test (LT).

Figures 1 and 2 are two scatter plots for the P-values of the logistic test against the extended Fisher test for 100 simulated samples of size  $n = 13$  with  $\beta = (\beta_0, \beta_1) = (2, 0)$  for the first one and  $\beta = (2, 2)$  for the second one. Samples are produced using the logistic model (6) with  $Z$  simulated as a logistic random variable with location parameter equal to zero, and scale parameter equal to 1.

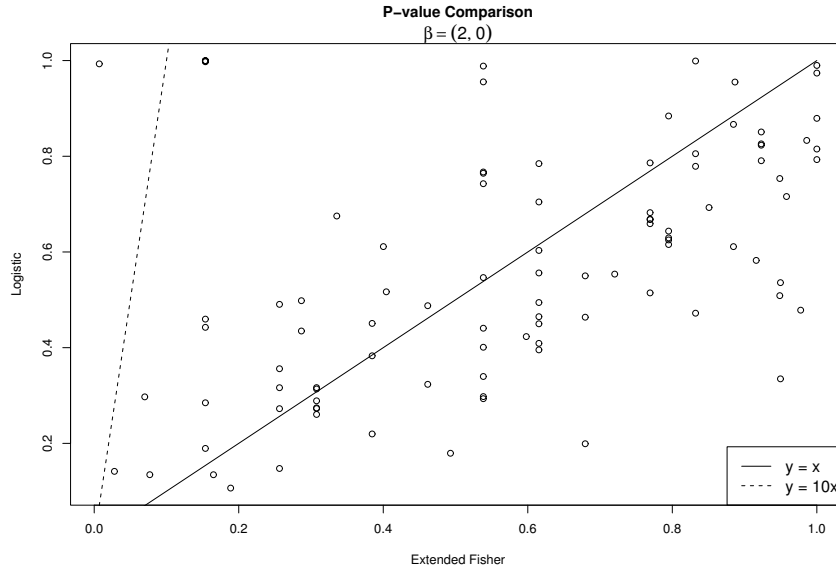


Figure 1:  $P_L \vee P_{EF}$  for  $\beta = (2, 0)$

We see that the extended Fisher test does at least as well as the Logistic test when  $\xi$  and  $Z$  are independent ( $\beta_1 = 0$ ). However, when  $\xi$  and  $Z$  are dependent ( $\beta_1 \neq 0$ ), the EFT clearly outperforms the LT. In all but 6 samples (94%), the EFT P-value is smaller than the LT P-value and, in 49 out of the 100 samples, the EFT P-value is more than 10 times smaller than the LT

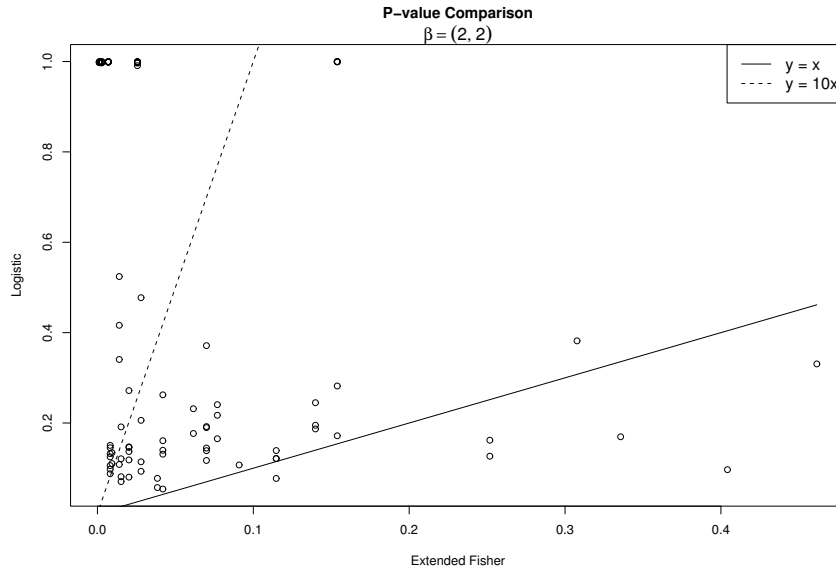


Figure 2:  $P_L \vee P_{EF}$  for  $\beta = (2, 2)$

P-value. Furthermore, when we look at those samples for which the EFT P-value is smaller than 0.1 (Figure 3), in the case  $\beta = (2, 2)$ , we see that all 80 EFT P-values are smaller than the LT P-values. In fact a cross-tabulation of the EFT and LT P-values for all 100 samples (Table 3)

	$.05 < P_L \leq .1$	$P_L > .1$	
$P_{EF} \leq .01$	2	33	35
$.01 < P_{EF} \leq .05$	7	26	43
$.05 < P_{EF} \leq .1$	0	12	12
$P_{EF} > .1$	2	18	20
	11	89	100

Table 3: Cross-tabulation of EFT and LT P-values

clearly demonstrates that, in the case  $\beta = (2, 2)$ , the EFT P-values are overall

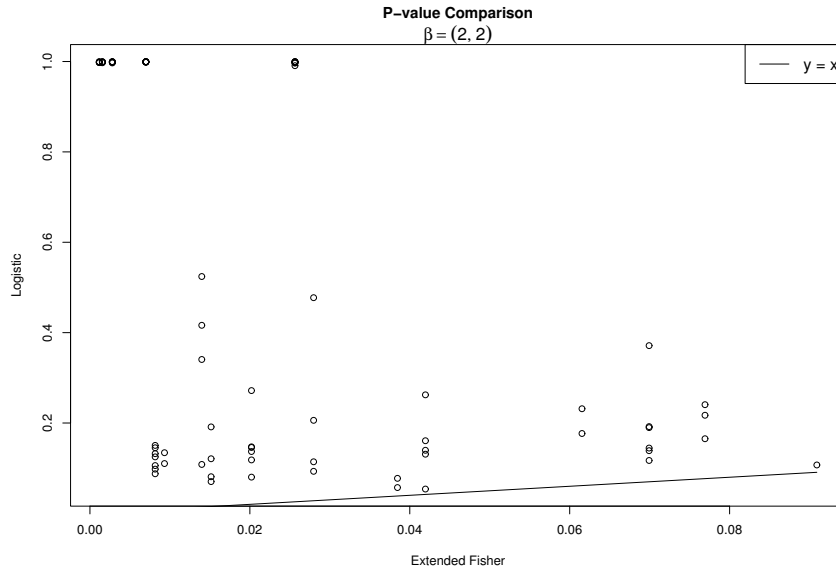


Figure 3:  $P_L \vee P_{EF}$  for  $P_{EF} < 0.1$  and  $\beta = (2, 2)$

significantly smaller than the LT P-values.

Finally, Figure 4 shows the result of a simulation of the EFT P-values, as a function of  $\beta_1$  ( $0 \leq \beta_1 \leq 4$ ). The simulation is made up of 1000 samples of size  $n = 13$  for each value of  $\beta_1$ . Those samples with an observed  $M$  of 0 were not counted in the tally. Their numbers ranged from 20%, for  $\beta_1 = 0$ , to 0%, for  $\beta_1 = 4$ , with an average of about 4% and a median of about 1%.

## 4 An exact test for hazard-similarity

We are now ready to tackle the main objective of this paper; to develop an exact test for hazard-similarity. The idea is simple and combines the

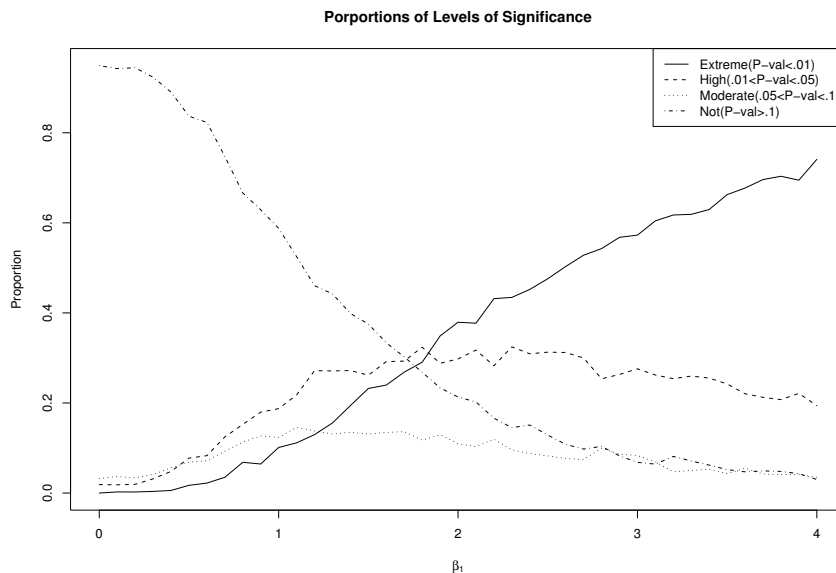


Figure 4:  $P_{EF} \vee \beta_1$  for  $\beta_1 \in [0, 4]$

extended Fisher test introduced in the previous section with Theorem 10.

Let  $X_1, \dots, X_n$  be a random sample from a given but unknown distribution  $\mathbb{D}$ . We wish to test the null hypothesis that  $\mathbb{D}$  is hazard-similar to some distribution  $\mathbb{D}_0$  (for example, the exponential distribution). According to Theorem 10,  $\mathbb{D}$  and  $\mathbb{D}_0$  are hazard-similar if and only if for any pair of independent random variables,  $X$  and  $Y$ , with distributions  $\mathbb{D}$  and  $\mathbb{D}_0$ , respectively,  $Z = \min(X, Y)$  and  $\xi = 1_{X < Y}$  are independent. The hazard-similarity test (HS) then consists of generating a random sample of size  $n$  from  $\mathbb{D}_0$ ,  $Y_1, \dots, Y_n$ , and applying the extended Fisher test to  $((Z_1, \xi_1), \dots, (Z_n, \xi_n))$ , where  $Z_i = \min(X_i, Y_i)$  and  $\xi_i = 1_{X_i < Y_i}$ .

In the case where  $\mathbb{D}_0$  is an exponential distribution, this test reduces to a

test of exponentiality. This is distinct from a goodness of fit test such as the chi-square ( $\chi^2$ ) or Kolmogorov-Smirnov (KS) goodness of fit tests, as it does not require specifying the parameter of the exponential distribution under the null hypothesis. Unlike Gnedenko's F test (GF), HS does not require the splitting of the data set (see below), thus avoiding a reduction in power. Another benefit of HS, and for that matter of GF, is that it is an exact test (valid for any sample size, in particular, small samples) that does not require the prior estimation of any parameter, whereas  $\chi^2$  and KS either use approximations that involve the estimation of a number of parameters and only apply to large samples, or require the null distribution be fully specified. Finally, in contrast to Lilliefors' (1969) adaptation of the Kolmogorov-Smirnov test for the exponential distribution in the case of an unknown  $\lambda$ , HS is based on an analytically tractable distributional statement the critical values of which are obtained through a precise counting procedure rather than a Monte Carlo simulation. However, HS does assume available, as is often the case, exogenous information on the value of the parameter  $\lambda$ . More specifically, we assume that we exogenously know that the true value of the parameter  $\lambda$  belongs to some reasonably sized neighbourhood of a reference value,  $\lambda_0$ . As this information cannot be used in the GF setup, we introduce a variant that does take into account such a knowledge. We call it the no-split Gnedenko F test (NSGF).

The remainder of the paper is devoted to an assessment of HS's potential as a competitive test, which we conduct by comparing it to the three exact

tests, KS, GF and NSGF. To this end, and for completeness, we briefly recall the two procedures, KS and GF – see for example, Ascher (1990).

### **The (classical) Kolmogorov-Smirnov test**

KS tests the null hypothesis that the data come from an exponential distribution with a given  $\lambda_0$ . It uses the empirical distribution function and measures its distance to the null distribution. The critical values are well documented and readily available in most statistical software.

### **The Gnedenko F test**

GF is a scale invariant test that examines whether the data come from some exponential distribution. It uses the so-called normalized spacings

$$D_i = (n - i + 1)(X_{(i)} - X_{(i-1)}),$$

where, as usual,  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  are the order statistics. Under the null hypothesis, the  $2\lambda D_i$  are independent exponential 1/2 (i.e.  $\chi_2^2$ ) random variables and the test statistic

$$Q(r) = \frac{\sum_{i=1}^r D_i / r}{\sum_{i=r+1}^n D_i / (n - r)}$$

has an  $F_{2r, 2(n-r)}$  distribution, where  $r$  is usually taken to equal  $n/2$  (or its integer part, if  $n$  is odd).

### **The no-split Gnedenko F test (NSGF)**

Instead of splitting the data as above, we generate a sample of size  $n$  from a reference exponential distribution with parameter  $\lambda_0$ , thus increasing the sample size from  $n$  to  $2n$ . The GF test is then applied.

### **A comparative analysis of the hazard-similarity test**

We simulate 1000 samples of  $n = 9$  observations from different exponential distributions, and compare the abilities of all four tests to recognize these null distributions as being exponential. The reference value we use in HS, KS and NSGF, is  $\lambda_0 = 1$ .

Figure 5 shows the dependence on  $\lambda$  of the distribution of P-values for all four tests.

We see, through this simulation, that in the cases where  $\lambda_0$  is reasonably close to the true  $\lambda$ , all four tests return similar results. However, when the difference is important (the exogenous information on the value of  $\lambda$  is erroneous), HS and GF markedly outperform KS and NSGF; thus highlighting the robustness of HS and GF (in comparison to KS and NSGF) to variations in the values of  $\lambda$ .

Similar results are obtained for  $n = 13$  (see Figure 6) and  $n = 20$  (see Figure 7).

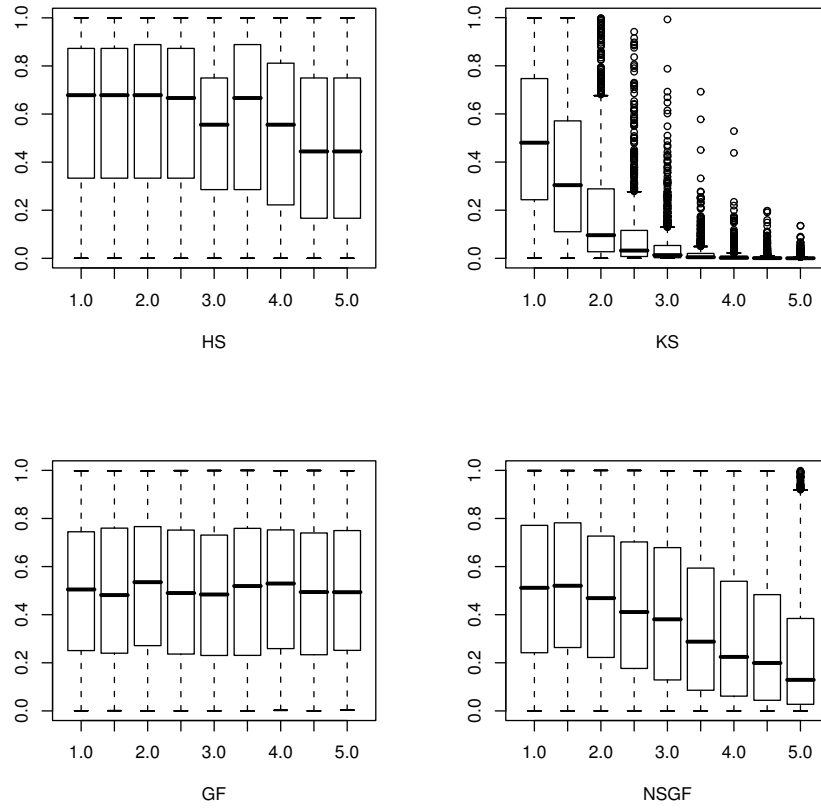


Figure 5: P-values by  $\lambda$  when the true distribution is exponential ( $n=9$ )

A closer investigation of the relative performances of HS and GF shows that the former is “significantly” larger than the latter for reasonable guesses of the true value of  $\lambda$ . Indeed, we simulated 1000 samples of  $n = 9$  observations from exponential distributions with  $\lambda$  varying between .5 and 1.5. For each sample, using a reference value of  $\lambda_0 = 1$ , the difference  $\text{P-value}_{HS} - \text{P-value}_{GF}$  was computed and the hypothesis that the difference in P-values is greater than

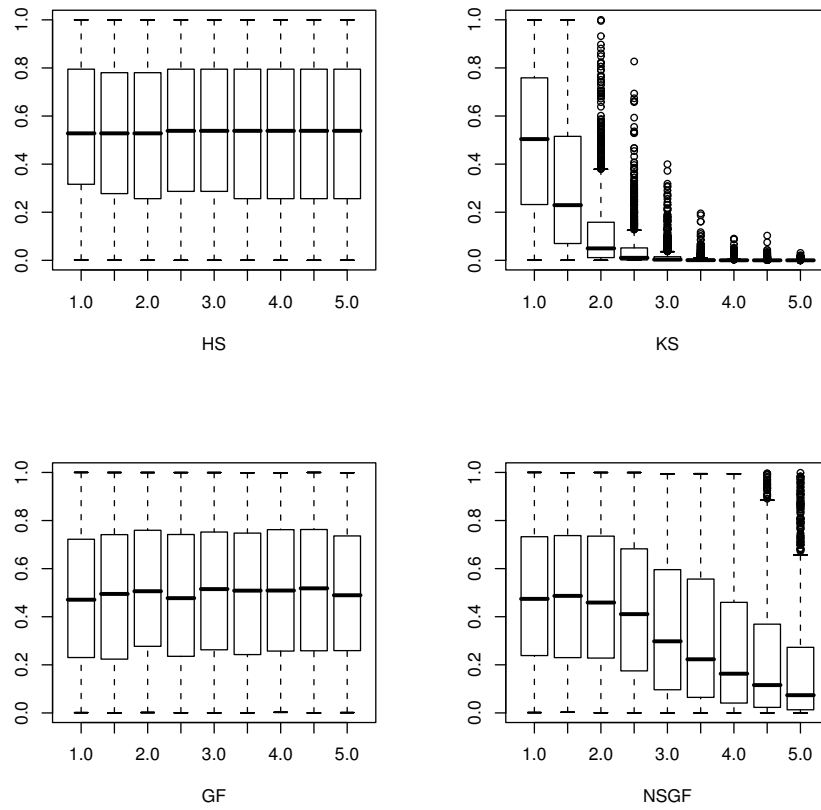


Figure 6: P-values by  $\lambda$  when the true distribution is exponential ( $n=13$ )

0.05 (one-sided) was tested for each value of  $\lambda$ . The table below summarises the results:

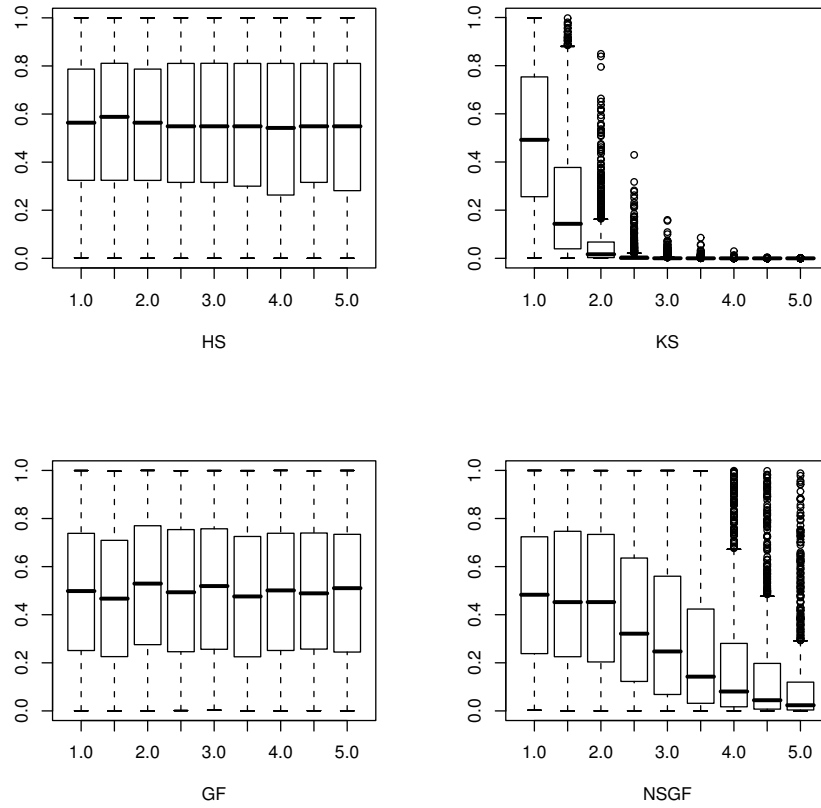


Figure 7: P-values by  $\lambda$  when the true distribution is exponential ( $n=20$ )

$\lambda$	Mean	P-values	
		t Test	Wilcoxon Test
0.5	0.06298	0.16243	0.13319
0.6	0.08567	0.00301	0.00220
0.7	0.09487	0.00016	0.00004
0.8	0.09776	0.00007	0.00004
0.9	0.08439	0.00357	0.00479
1.0	0.08376	0.00300	0.00423
1.1	0.08489	0.00347	0.00428
1.2	0.09584	0.00024	0.00022
1.3	0.08944	0.00086	0.00037
1.4	0.07363	0.03007	0.01884
1.5	0.08811	0.00413	0.00163

This simulation shows that, although both HS and GF seem to have similar behaviours overall, for reasonable guesses of the true value of  $\lambda$ , HS P-values tend to be at least 0.05 larger than GF P-values.

This investigation of the hazard-similarity test is of course not complete and should be followed by a power analysis. We shall not attempt this in this paper but only point out that either scales,  $\mathcal{H}$  or  $\mathcal{K}$ , introduced in Section 2, could be used in such an analysis.

We end this discussion by observing that, while HS performs very well for reasonable guesses of the true value of  $\lambda$  (true value), it is vulnerable to large differences between  $\lambda$  and  $\lambda_0$  (reference value used in simulation). Indeed, if  $\lambda$  is overly underestimated (resp. overestimated) then one runs the risk of having each and every data point,  $x_i$ , bigger (resp. smaller) than its simulated counterpart,  $y_i$ , thus returning a value of  $M$  equal to  $n$  (resp. 0). Now, since according to Proposition 14,  $\Pr(T = 0|M = 0) = \Pr(T = 0|M = n) = 1$ , whenever  $M = 0$  or  $M = n$ , the hazard-similarity P-value will be nil and the null hypothesis will be automatically rejected.

## References

- [1] Ascher S. (1990), A survey of tests for exponentiality, *Commun. Statist. – Th. Meth.*, **19**, pp 1811–1825.
- [2] Cox D. R., Oakes D. (1984), *Analysis of Survival Data*, Chapman-Hall.

- [3] Henze N., Meintanis S. (2002), Tests of fit for exponentiality based on the empirical Laplace transform, *Statistics*, **36**, pp 147–161.
- [4] Henze N., Meintanis S. (2005), Recent and classical tests for exponentiality: a partial review with comparisons, *Metrika*, **61**, pp 21–45.
- [5] Kotz S., Shanbhag D. N. (1980), Some new approaches to probability distributions, *Adv. Appl. Prob.* **12**, pp 903–921.
- [6] Lilliefors H. W. (1969), On the Kolmogorov-Smirnov test for the exponential distribution with unknown mean, *JASA*, **64**, pp 387–389.