Statistics, Vol. 43, No. 1, February 2009, 1–12

Taylor & Francis
Taylor & Francis Group

A location- and scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations

Aurea Grané[a]* and Josep Fortiana[b]

[a] *Statistics Department, Universidad Carlos III de Madrid, Getafe Spain;* [b] *Departament de Probabilitats Lògica i Estadística, Universitat de Barcelona, Spain*

(*Received 2 July 2007; final version received 3 March 2008*)

We propose a goodness-of-fit statistic, Q_n, based on Hoeffding's maximum correlation (Fortiana and Grané 2003) to test the composite hypothesis that the data come from the two-parameter exponential family. We study its small and large sample properties, and we obtain tables of the critical values of Q_n and some power curves. We compare our statistic with the Shapiro-Wilk statistic for exponentiality and the Gini statistic.

Keywords: goodness-of-fit; exponentiality; L-statistics; maximum correlation

AMS Classifications 2000: 62G10; 62G30; 62E15

1. Introduction

Most goodness-of-fit statistics can be interpreted as measures of proximity between two distributions: empirical and hypothesized. The statistic we propose in this paper is based on Hoeffding's maximum correlation between two probability distribution functions F_1 and F_2, with second-order moments, which is equal to

$$\rho^+(F_1, F_2) = \frac{\int_0^1 F_1^-(p)\, F_2^-(p)\, dp - \mu_1\, \mu_2}{\sigma_1\, \sigma_2}, \tag{1}$$

where F_i^- is the left-continuous pseudoinverse of F_i, $\mu_i = E(F_i)$ and $\sigma_i^2 = \mathrm{var}(F_i)$, $i = 1, 2$ [1]. Here the notation $E(F)$ represents the expected value of any random variable whose probability distribution function is F, and analogously for $\mathrm{var}(F)$. Since $\rho^+(F_1, F_2)$ equals 1 if and only if $F_1 = F_2$ (almost everywhere) up to a scale and location change, the quantity $\rho^+(F_n, F)$ has been used in previous works [2–4] as a qualitative measure of goodness-of-fit of an iid sample, with empirical distribution function F_n, to a given distribution F. In this article, we use Hoeffding's maximum correlation to construct a statistic for testing the composite hypothesis of exponentiality when the location and scale parameters are both unknown. More precisely, given n iid $\sim F$ random variables, we will test

$$H_0 : F = \mathrm{Exp}(\alpha, \beta), \quad (\alpha, \beta) \in \mathbb{R} \times \mathbb{R}_+, \tag{2}$$

*Corresponding author. Email: aurea.grane@uc3m.es

ISSN 0233-1888 print/ISSN 1029-4910 online

DOI: 10.1080/02331880802187667
http://www.informaworld.com

where $\mathrm{Exp}(\alpha, \beta)$ is the exponential distribution with unknown location and scale parameters, $F(x; \alpha, \beta) = 1 - \exp -(x - \alpha)/\beta, x \geq \alpha, -\infty < \alpha < +\infty$ and $\beta > 0$. Note that the standard exponential corresponds to $\alpha = 0$, $\beta = 1$. The two-parameter exponential distribution is closely related to several well-known distributions with statistical applications, as the generalized Pareto, the gamma and the Weibull distributions and it is frequently used in reliability theory, life testing and the theory of stochastic processes. In the context of reliability theory, the parameter α is also referred as a threshold parameter. There are many types of problems in which either the response will not occur below some threshold value or the phenomenon can not be measured below that value. Sometimes, the location parameter α is interpreted as the minimum (or the guarantee) time before which no failure occurs; a negative location parameter is indicative of quiescent failures (failures that occur before a product is used for the first time) or of problems with the manufacturing, packaging or shipping processes (see [5–7] for some examples and applications).

Let $y_1, \dots, y_n$ be n iid $\sim F$ random variables with empirical distribution function F_n and let $y_{(1)}, \dots, y_{(n)}$ be the order statistic. We propose the statistic

$$Q_n = \frac{s_n \rho^+(F_n, F)}{1/n \sum_{i=1}^n b_i y_{(i)}}, \tag{3}$$

where s_n^2 is the sample variance and coefficients b_i, in the denominator, are such that $\sum_{i=1}^n b_i = 0$, to test the null composite hypothesis of exponentiality.

In Section 2, we provide a formula for computing the numerator of Equation (3) under the null composite hypothesis (see Lemma 2.1) which does not require the knowledge or estimation of the location and scale parameters. This is the reason why we state that statistic (3) can be used to test the composite hypothesis that the data come from the two-parameter exponential family.

Since we want to develop a test based on Q_n, our first aim is to determine its critical values, that is,

$$a \in \mathbb{R} \quad \text{such that} \quad P(Q_n > a) = \varepsilon,$$

where $\varepsilon \in (0, 1)$ is a fixed significance level. Defining the auxiliary function

$$L_n^a = s_n \rho^+(F_n, F) - \frac{a}{n} \sum_{i=1}^n b_i y_{(i)}, \tag{4}$$

the problem of finding the critical values of Q_n reduces to find

$$a \in \mathbb{R} \quad \text{such that} P(L_n^a > 0) = \varepsilon.$$

In Section 2, we compute the exact expectation and variance of L_n^a and we prove that this auxiliary function depends linearly on the scale parameter and does not depend on the location parameter (see Propositions 2.1 and 2.2). In fact, the restriction imposed on coefficients b_i's in the denominator of Equation (3) is necessary to avoid the dependence of L_n^a on the location parameter. As a consequence, since both the numerator and denominator of Q_n depend linearly only on the scale parameter (and they do not depend on the location parameter), the Q_n statistic is location- and scale-free.

In Sections 3 and 4, we select a set of coefficients b_i's and we determine the asymptotic distribution of the resulting auxiliary function L_n^a and the critical values of the test based on the corresponding Q_n statistic. For a sample size of $n = 20$, we study the power of the test based on this Q_n and we compare it with those of the tests based on the Shapiro–Wilk statistic and the Gini statistic. In Section 5, we analyse two data sets, and in Table 1 we reproduce some of the approximate asymptotic critical values of the test based on this Q_n.

Table 1. Approximate asymptotic critical values of the bilateral and unilateral test based on Q_n for a 5% significance level.

	Two-tail		One-tail	
n	Lower	Upper	Lower	Upper
20	3.146304	4.320201	3.251234	4.192144
21	3.160088	4.314727	3.264053	4.192066
22	3.173227	4.310192	3.276208	4.192224
23	3.185770	4.306386	3.287757	4.192541
24	3.197759	4.303155	3.298751	4.192963
25	3.209234	4.300380	3.309236	4.193450
26	3.220228	4.297969	3.319251	4.193972
27	3.230775	4.295852	3.328831	4.194508
28	3.240903	4.293973	3.338008	4.195045
29	3.250639	4.292289	3.346810	4.195570
30	3.260006	4.290764	3.355261	4.196076
40	3.337582	4.280157	3.424729	4.199531
50	3.394754	4.272683	3.475477	4.200289
60	3.439147	4.266099	3.514688	4.199350
70	3.474911	4.259991	3.546184	4.197461
80	3.504560	4.254222	3.572218	4.195059
90	3.529652	4.248786	3.594213	4.192399
100	3.551117	4.243842	3.613121	4.189627
150	3.626893	4.222365	3.679269	4.176155
200	3.671637	4.208274	3.720112	4.164680

2. Definition of the test statistic and small sample properties of L_n^a

Let $y_1, \dots, y_n$ be n random variables iid $\sim F = \mathrm{Exp}(\alpha, \beta)$ with empirical distribution function F_n, and let $y_{(1)}, \dots, y_{(n)}$ be the order statistic.

LEMMA 2.1 *If s_n^2 is the sample variance, then $s_n\, \rho^+(F_n, F)$ is an L-statistic, $1/n \sum_{j=1}^n l_j\, y_{(j)}$, with coefficients*

$$l_j = (n - j)\log(n - j) - (n - j + 1)\log(n - j + 1) + \log(n), \quad 1 \le j \le n,$$

where, conventionally, $0\ \log 0 = 0$.

Proof Using formula (1) for F_n and $F = \mathrm{Exp}(\alpha, \beta)$ and that the expectation and variance of $\mathrm{Exp}(\alpha, \beta)$ are $\alpha + \beta$ and β^2, respectively, we have that

$$s_n\, \rho^+(F_n, F) = \frac{1}{\beta}\left(\int_0^1 F_n^-(p)\, F^-(p)\, dp - (\alpha + \beta)\, \bar{y}_n\right). \tag{5}$$

The part that corresponds to the integral in (5) is

$$\int_0^1 F_n^-(p)\, F^-(p)\, \mathrm{d}p = \sum_{i=0}^{n-1} \int_{\frac{i}{n}}^{(i+1)/n} (\alpha - \beta\ \log(1 - p))\ y_{(i+1)}\, \mathrm{d}p$$

$$= \frac{\alpha}{n}\sum_{j=1}^{n} y_{(j)} - \frac{\beta}{n}\sum_{i=0}^{n-1} y_{(i+1)}\left[(n - i)\log(n - i) - (n - i - 1)\log(n - i - 1) - \log(n) - 1\right].$$

Letting $j = i + 1$,

$$\int_0^1 F_n^-(p)\,F^-(p)\,\mathrm{d}p = \frac{\alpha+\beta}{n}\sum_{j=1}^{n} y_{(j)}$$
$$+\frac{\beta}{n}\left(\sum_{j=1}^{n}\left[(n-j)\log(n-j)-(n-j+1)\log(n-j+1)+\log(n)\right]y_{(j)}\right).$$

Notice that the $(\alpha+\beta)/n\sum_{j=1}^{n} y_{(j)} = (\alpha+\beta)\bar{y}_n$; hence the part between parenthesis in Equation (5) is equal to

$$\frac{\beta}{n}\left\{\sum_{j=1}^{n}\left[(n-j)\log(n-j)-(n-j+1)\log(n-j+1)+\log(n)\right]y_{(j)}\right\}.$$

Finally, dividing by β, we get an expression for $s_n\,\rho^+(F_n, F)$, which does not require the knowledge or the estimation of the parameters. ■

As a consequence of Lemma 2.1, under the null composite hypothesis H_0: $F = \mathrm{Exp}(\alpha, \beta)$, where $(\alpha, \beta) \in \mathbb{R}\times\mathbb{R}_+$, the statistic Q_n defined in Equation (3) can be written as:

$$Q_n = \frac{\sum_{i=1}^{n} l_i\, y_{(i)}}{\sum_{i=1}^{n} b_i\, y_{(i)}},$$

where $l_i = (n-i)\log(n-i)-(n-i+1)\log(n-i+1)+\log(n)$, for $i = 1, \ldots, n$ and b_i's coefficients are such that $\sum_{i=1}^{n} b_i = 0$. Hence, in order to test the composite hypothesis that the data come from the two-parameter exponential family, and for a given set of coefficients b_i's, it is possible to compute the value of Q_n without knowing or estimating the values of the location and scale parameters.

2.1. *Small sample properties of L_n^a under exponentiality*

The transformation $x_i = (y_i - \alpha)/\beta$, for $i = 1, \ldots, n$, gives n iid $\sim$ Exp(0, 1) random variables. If $x_{(1)}, \ldots, x_{(n)}$ is the order statistic, we denote the expected value of $x_{(i)}$ by

$$m_i = E(x_{(i)}),\quad i = 1, 2, \ldots, n,$$

and the covariance of $x_{(i)}$ and $x_{(j)}$ by

$$v_{ij} = E\left((x_{(i)} - m_i)(x_{(j)} - m_j)\right),\ i, j = 1, 2, \ldots, n,$$

and let

$$\mathbf{m} = (m_1, \ldots, m_n)',\quad \mathbf{V} = \left(v_{ij}\right)_{1\le i,j\le n},$$

be the mean vector and the covariance matrix of the order statistic from the standard exponential, respectively. Explicit formulae for m_i and v_{ij} for the Exp(0, 1) distribution [8] are

$$m_i = \sum_{k=1}^{i}(n-k+1)^{-1},\ i = 1, 2, \ldots, n,$$
$$v_{ij} = \sum_{k=1}^{\min(i,j)}(n-k+1)^{-2}.$$

Proposition 2.1 *The function L_n^a defined in Equation (4) does not depend on the location parameter and can be written as a linear combination of the order statistic from the standard exponential,*

$$\frac{\beta}{n}\sum_{j=1}^{n} c_{nj}^a\, x_{(j)},$$

with coefficients

$$c_{nj}^a = (n-j)\log(n-j) - (n-j+1)\log(n-j+1) + \log(n) - a\,b_j,\ 1 \le j \le n,$$

where, conventionally, $0 \log 0 = 0$.

Proof From Lemma 2.1, definition (4) is written as

$$L_n^a = \frac{1}{n}\sum_{j=1}^{n} c_{nj}^a y_{(j)},$$

where

$$c_{nj}^a = (n-j)\log(n-j) - (n-j+1)\log(n-j+1) + \log(n) - a\,b_j,\ 1 \le j \le n.$$

Using the transformation above,

$$L_n^a = \frac{1}{n}\sum_{j=1}^{n} c_{nj}^a(\alpha + \beta\, x_{(i)}) = \frac{\alpha}{n}\sum_{j=1}^{n} c_{nj}^a + \frac{\beta}{n}\sum_{j=1}^{n} c_{nj}^a\, x_{(i)}.$$

To prove the proposition, we need to see that the first summatory is null:

$$\begin{aligned}\sum_{j=1}^{n} c_{nj}^a &= \sum_{j=1}^{n}\left[(n-j)\log(n-j) - (n-j+1)\log(n-j+1) + \log(n) - a\,b_j\right]\\ &= -n\,log(n) + n\,log(n) - a\sum_{j=1}^{n} b_j = 0,\end{aligned}$$

where we have used that $\sum_{j=1}^{n} b_j = 0$. ■

So far we have proved that the auxiliary function L_n^a depends linearly only on the scale parameter, and consequently, the Q_n statistic is location- and scale-free and can be written as a quotient of L-statistics from the standard exponential distribution.

Proposition 2.2 *The exact expectation and variance of the function L_n^a defined in Equation (4) are*

$$E(L_n^a) = \frac{\beta}{n}\,\mathbf{c}_n'\,\mathbf{m}, \quad \mathrm{Var}(L_n^a) = \frac{\beta^2}{n^2}\,\mathbf{c}_n'\,\mathbf{V}\,\mathbf{c}_n,$$

where $\mathbf{c}_n = (c_{n1}^a, \ldots, c_{nn}^a)'$.

Proof Defining $\mathbf{x} = (x_{(1)}, \ldots, x_{(n)})'$, then $L_n^a = \frac{\beta}{n}\,\mathbf{c}_n'\,\mathbf{x}$. ■

The expressions for the expectation and variance of the auxiliary function L_n^a can be made more explicit if we consider the vectors $\mathbf{b} = (b_1, \dots, b_n)'$ and $\mathbf{l} = (l_1, \dots, l_n)'$, where

$$l_j = (n-j)\log(n-j) - (n-j+1)\log(n-j+1) + \log(n),\ 1 \le j \le n.$$

Since $\mathbf{c}_n = (\mathbf{l} - a\,\mathbf{b})$, then

$$E(L_n^a) = \frac{\beta}{n}\,(\mathbf{l} - a\,\mathbf{b})'\,\mathbf{m},$$

$$\mathrm{Var}(L_n^a) = \frac{\beta^2}{n^2}\,(\mathbf{l}'\mathbf{V}\mathbf{l} - 2a\,\mathbf{l}'\mathbf{V}\mathbf{b} + a^2\,\mathbf{b}'\mathbf{V}\mathbf{b}).$$

If we could compute the exact distribution of Q_n, we could find the exact critical values of the test based on Q_n. In [2], we were able to compute the distribution of a similar statistic for testing exponentiality when the location parameter α was specified. Unfortunately, in the general case, it is not easy to compute this distribution, even for a fixed set of coefficients b_i's. This is the reason why in the following sections, we select a specific set of coefficients b_i's and we determine the asymptotic critical values of the resulting Q_n using the auxiliary function L_n^a and applying the general theory of L-statistics.

3. Asymptotic distribution of L_n^a under exponentiality

In order to compute the asymptotic distribution of L_n^a, applying the general asymptotic theory for L-statistics, coefficients c_{ni}^a should be of the form $c_{ni}^a = J_n^a(i/n)$, where J_n^a is bounded and continuous a.e. (F^-) (see [9] or [10] for a general explanation of this theory). Since $c_{ni}^a = l_i - a\,b_i$, and coefficients l_i already satisfy these conditions, these restrictions should only be imposed on coefficients b_i. As far as here, the only restriction imposed on coefficients b_i is that $\sum_{i=1}^n b_i = 0$. Of course, there are many sets of b_i's that could be selected, but in order to illustrate the methodology we have chosen:

$$b_i = \frac{i}{n} - \frac{n+1}{2n}, \quad \text{for } i = 1, 2, \dots, n.$$

With these coefficients, we have the following result:

Proposition 3.1 *Under the composite null hypothesis $H_0 : F = Exp(\alpha, \beta)$, where $(\alpha, \beta) \in \mathbb{R} \times \mathbb{R}_+$,*

(i) *Let $\mu_n^a = E(L_n^a)$, $\sigma_{n,a}^2 = n\,\mathrm{var}(L_n^a)$, then*

$$\mu_n^a = \beta\left[\log\left(\frac{n^n}{n!}\right)^{1/n} - \frac{a}{n}\left(\frac{n}{2} - \frac{n+1}{4}\right)\right],$$

$$\sigma_{n,a}^2 = \beta^2[A_0(n) + A_1(n)\,a + A_2(n)\,a^2],$$

with $A_0(n)$, $A_1(n)$, and $A_2(n)$ functions depending only on n, and the sequence $\{\sigma_{n,a}^2\}$ converges to $\sigma_a^2 = \beta^2(2 - 37/36\,a + a^2/6)$, as $n \to \infty$.

(ii) *We have the following convergences in law*

$$\sqrt{n}\left[L_n^a - \mu_n^a\right] \xrightarrow[n\longrightarrow\infty]{\mathcal{L}} N\left(0, \sigma_a^2\right), \tag{6}$$

$$\sqrt{n}\frac{\left[L_n^a - \mu_n^a\right]}{\sigma_{n,a}^2} \xrightarrow[n\longrightarrow\infty]{\mathcal{L}} N(0,1). \tag{7}$$

Proof (i) From the asymptotic theory of L-statistics (see [10, Ch. 19] for the notation and the constructions used), L_n^a can be written as

$$L_n^a = \int_0^1 J_n^a(t)\, F_n^-(t)\, \mathrm{d}t,$$

where F_n^- is the pseudo–inverse of the empirical distribution and

$$J_n^a(t) = n(1-t)\log\left(\frac{1-t}{1-t+1/n}\right) + \log\left(\frac{1}{1-t+1/n}\right) - a\left(t - \frac{n+1}{2n}\right).$$

The expectation of L_n^a under the null hypothesis, can be expressed as

$$\mu_n^a = \int_0^1 J_n^a(t)\, F^-(t)\, \mathrm{d}t,$$

where $F^-(t) = \alpha - \beta\, \log(1-t),\ 0 \le t \le 1$. Alternatively, it can be computed directly,

$$\mu_n^a = \frac{\beta}{n}\left(\sum_{j=1}^n l_j m_j - a \sum_{j=1}^n b_j m_j\right), \tag{8}$$

where $m_j = E(x_{(j)})$, for $j = 1, \ldots, n$. The first part of Equation (8) is the expectation of $s_n \rho^+$,

$$\sum_{j=1}^m l_j\, m_j = \sum_{j=1}^{n-1}\left(\big[(n-j)\log(n-j) - (n-j+1)\log(n-j+1)\big]\sum_{k=1}^j \frac{1}{n-k+1}\right) + \log(n)\sum_{j=1}^n\left(\sum_{k=1}^j \frac{1}{n-k+1}\right). \tag{9}$$

Note that the first summatory of Equation (9) can be written as

$$\begin{aligned}
&\sum_{j=1}^{n-1}\left(\big[(n-j)\log(n-j) - (n-j+1)\log(n-j+1)\big]\sum_{k=1}^j \frac{1}{n-k+1}\right)\\
&\quad = -\log n + (n-1)\log(n-1)\left[\sum_{n}^{n}\frac{1}{k} - \sum_{n-1}^{n}\frac{1}{k}\right] + (n-2)\log(n-2)\left[\sum_{n-1}^{n}\frac{1}{k} - \sum_{n-2}^{n}\frac{1}{k}\right]\\
&\qquad + \cdots + 2\,\log 2\left[\sum_{3}^{n}\frac{1}{k} - \sum_{2}^{n}\frac{1}{k}\right] = -\log n - \sum_{2}^{n-1}\log j = -\log(n!).
\end{aligned}$$

And the second summatory of Equation (9)

$$\log(n)\sum_{j=1}^n\left(\sum_{k=1}^j \frac{1}{n-k+1}\right) = \log(n)\, n.$$

The second part of Equation (8) is the expectation of $\sum_{j=1}^n b_j x_{(j)}$,

$$\sum_{j=1}^n b_j\, m_j = \sum_{j=1}^n\left[\left(\frac{j}{n} - \frac{n+1}{2n}\right)\sum_{k=1}^j \frac{1}{n-k+1}\right]. \tag{10}$$

The first summand of Equation (10) is

$$\sum_{j=1}^{n}\left(\frac{j}{n}\sum_{k=1}^{j}\frac{1}{n-k+1}\right)=\left(\frac{n}{2}+\frac{n+1}{4}\right),$$

and the second summand of Equation (10) is equal to

$$-\frac{n+1}{2n}\sum_{j=1}^{n}\sum_{k=1}^{j}\frac{1}{n-k+1}=\frac{1}{n}\,\frac{n+1}{2}.$$

Finally, we have

$$\mu_n^a=\frac{\beta}{n}\left[-\log(n!)+n\,\log(n)-a\,\left(\frac{n}{2}+\frac{n+1}{4}-\frac{n+1}{2}\right)\right].$$

The variance of L_n^a is $\operatorname{var}(L_n^a)=\sigma_{n,a}^2/n$, where

$$\sigma_{n,a}^2=\int_0^1\int_0^1 J_n^a(s)\,J_n^a(t)[\min(s,t)-s\,t]\,\mathrm{d}F^-(s)\,\mathrm{d}F^-(t). \tag{11}$$

Because of the symmetry of the function in the integrand, the region to integrate can be reduced to

$$\sigma_{n,a}^2=2\,\beta^2\int_0^1\left(J_n^a(s)\int_0^s J_n^a(t)\frac{t}{1-t}\,\mathrm{d}t\right)\,\mathrm{d}s.$$

We divide the region $\{(s,t)\in\mathbb{R}^2:0\le s\le 1, 0\le t\le s\}$ in three parts, namely A, B and C,

$$\mathrm{A}=\left\{(s,t)\in\mathbb{R}^2:0\le s\le 1-\frac{1}{n},\ 0\le t\le s\right\},$$
$$\mathrm{B}=\left\{(s,t)\in\mathbb{R}^2:1-\frac{1}{n}\le s\le 1,\ 0\le t\le 1-\frac{1}{n}\right\},$$
$$\mathrm{C}=\left\{(s,t)\in\mathbb{R}^2:1-\frac{1}{n}\le s\le 1,\ 1-\frac{1}{n}\le t\le s\right\},$$

and we consider the following approximations for $\log(1-s+1/n)$:

$$\log\left(1-s+\frac{1}{n}\right)\approx\log(1-s)+\sum_{k=1}^{m}\frac{(-1)^{k+1}}{k}(n(1-s))^{-k}, \tag{12}$$

when $s<1-1/n$, and

$$\log\left(1-s+\frac{1}{n}\right)\approx\log\left(\frac{1}{n}\right)+\sum_{k=1}^{m}\frac{(-1)^{k+1}}{k}(n(1-s))^{k}, \tag{13}$$

when $s\ge 1-1/n$.

We have used `Mathematica` to compute these integrals, using the approximations up to $m=2$ for (12) and $m=5$ for (13) and we have obtained that the variance of L_n^a is a polynomial of second degree in a.

Since J_n^a is continuous and bounded a.e. (F^-), to compute $\sigma_a^2 = \lim_{n\to\infty} \sigma_{n,a}^2$, we can commute the limit with the integral, hence we substitute

$$J^a(t) = \lim_{n\to\infty} J_n^a(t) = -\left[1 + \log(1-t) + a\left(t - \frac{1}{2}\right)\right]$$

for J_n^a in integral (11). Analogously, since the function in the integrand is symmetrical,

$$\begin{aligned}\sigma_a^2 &= 2\,\beta^2 \int_0^1 \left(J^a(s) \int_0^s J^a(t) \frac{t}{1-t}\,\mathrm{d}t\right) \mathrm{d}s \\ &= 2\beta^2 \int_0^1 \left[1 + \log(1-s) + a\left(s - \frac{1}{2}\right)\right] \int_0^s \left[1 + \log(1-t) + a\left(t - \frac{1}{2}\right)\right] \frac{t}{1-t}\,\mathrm{d}t\,\mathrm{d}s \\ &= \beta^2 \left(2 - \frac{37}{36}\,a + \frac{a^2}{6}\right).\end{aligned}$$

(ii) The convergence of Equation (6) is obtained from Theorem 1 of [10, pp. 664–665]. Convergence (7) is immediate from Equation (6) and from the fact that $\sigma_a^2 = \lim_{n\to\infty} \sigma_{n,a}^2$. ■

To compute the critical values, we have used the normal approximation based on Equation (7) which, having a higher convergence rate than Equation (6), gives a more powerful test.

The critical value a is obtained from solving the equation

$$-\sqrt{n}\,\frac{\mu_n^a}{\sigma_{n,a}} = c_\varepsilon,$$

where c_ε is the $(1-\varepsilon)$–quantile of the standard normal distribution. We have developed a `Mathematica` program that computes these critical values, given n. Some of them are reproduced in Table 1.

4. Power of Q_n

In this section, we consider three families of probability distributions, frequently used as alternatives to the two-parameter exponential distribution, in the context of reliability theory and life testing. These alternatives are of the form $F(x; \theta_1, \theta_2)$. In each case, we fix one of the parameters, and let the other vary in its specific range.

- A1. Generalized Pareto, whose distribution function is

$$F(x; a, k) = 1 - \left(1 - \frac{k}{a}\,x\right)^{1/k}, \quad a > 0,$$

with $0 \le x < \infty$, if $k \le 0$, and $0 \le x \le a/k$, if $k > 0$. See [11] for notation and relations.

- A2. Gamma with shape parameter α and scale parameter β, whose density function is

$$f(x; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha}\,x^{\alpha-1}\,\mathrm{e}^{-x/\beta}, \quad \alpha, \beta > 0, x \ge 0.$$

- A3. Weibull with shape parameter α and scale parameter β, whose density function is

$$f(x; \alpha, \beta) = \frac{\alpha}{\beta^\alpha}\,x^{\alpha-1}\,\exp\left\{-\left(\frac{x}{\beta}\right)^\alpha\right\}, \alpha, \beta > 0, \quad x \ge 0.$$

We have compared the power of the test based on Q_n with the tests based on other location- and scale-free statistics for testing exponentilaity, such as the Shapiro–Wilk statistic [12]:

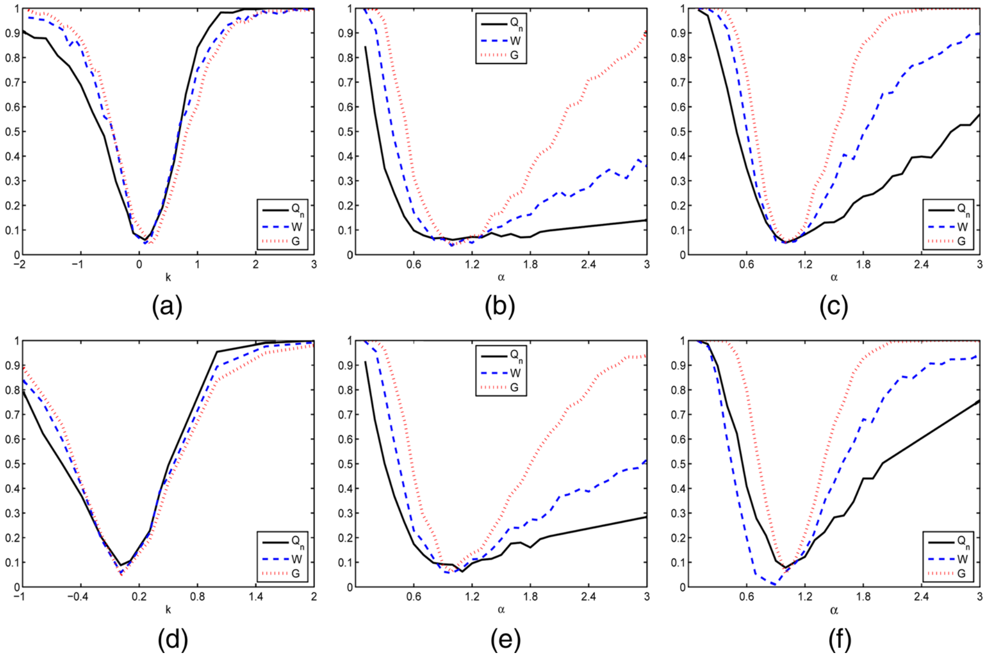
$$W = \frac{n\,(\bar{x} - x_{(1)})^2}{(n-1)\,\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{14}$$

and the Gini statistic [13]:

$$G = \frac{\sum_{i=1}^{n}(2\,i - n - 1)\,x_{(i)}}{n\,(n-1)\,\bar{x}}. \tag{15}$$

Power against an alternative distribution $F(x;\theta_1,\theta_2)$ has been estimated by the relative frequency of values of the statistic in the critical region for $N = 500$ simulated samples of size n of $F(x;\theta_1,\theta_2)$. For $F(x;\theta_1,\theta_2)$, we have taken distributions in each of the families A1–A3 with one fixed parameter and let the other vary within its range. For each family, we have taken 30 different values of the free parameter.

As it is explained in [14], a useful measure in describing distributions is the coefficient of variation, C_{V}, which is defined as the quotient between the mean and the standard deviation of the distribution being considered. This coefficient is closely related to the failure rate (or hazard rate) since $C_{\mathrm{V}} > 1$ for decreasing failure rate (DFR) distributions, $C_{\mathrm{V}} < 1$ for increasing failure rate (IFR) distributions and $C_{\mathrm{V}} = 1$ for distributions with constant failure rate. The gamma and the Weibull alternatives are IFR for $\alpha > 1$ and DFR for $0 < \alpha < 1$. For $\alpha = 1$, they both reduce to the Exp$(0, \beta)$ with constant failure rate. The generalized Pareto is IFR for $k > 0$ and DFR $k < 0$. When $k = 0$, the generalized Pareto reduces to the Exp$(0, a)$. We have observed that the right tail of Q_n is significant for DFR alternatives, whereas the left tail is significant for IFR alternatives. Consequently, if C_{V} is known for a given distribution, we can develop a one-sided test more powerful than the two-sided default test.



(a) (b) (c)

(d) (e) (f)

Figure 1. (a)–(c) Two-tail power curves for the tests based on Q_n, W and G for (a) the A1 family with $a = 1$, (b) the A2 family with $\beta = 1$ and (c) the A3 family with $\beta = 1$. (d)–(f): One-tail power curves for the tests based on Q_n, W and G for (d) the A1 family with $a = 1$, (e) the A2 family with $\beta = 1$, and (f) the A3 family with $\beta = 1$.

Figure 1 shows the power for both bilateral and unilateral tests at 5% significance level, for $n = 20$. Looking at the graphs, we can conclude that the Q_n statistic is good in detecting family A1, whereas for families A2 and A3, it is only reasonably good when $C_V < 1$. In the other cases, the Gini statistic is the best one. It also can be observed that the test based on the Shapiro–Wilk statistic is biased for the Weibull alternative.

5. Data analysis

In this section, we apply the test of exponentiality based on the statistic

$$Q_n = \frac{\sum_{i=1}^{n} l_i y_{(i)}}{\sum_{i=1}^{n} (i/n - (n+1)/(2n)) y_{(i)}},$$

where $l_i = (n-i)\log(n-i) - (n-i+1)\log(n-i+1) + \log(n)$, for $i = 1, \dots, n$, with the convention $0 \log 0 = 0$, to two real data sets, in order to determine whether the data come from the two-parameter exponential distribution. In Figure 2, we depict the histograms of these data sets.

5.1. *Data set 1*

Grubbs [15] gives the following mileages for the failure times of 19 personnel carriers: 162, 200, 271, 302, 393, 508, 539, 629, 706, 777, 884, 1008, 1101, 1182, 1463, 1603, 1984, 2355, 2880. He considers these data to follow a two-parameter exponential distribution.

We compute the value of the test statistic obtaining $Q_n = 3.5571$. For a 5% significance level, the approximate asymptotic critical values are $\text{cv}_{\text{left}} = 3.13182$ and $\text{cv}_{\text{right}} = 4.32689$; hence, we conclude that the data follow a two-parameter exponential distribution.

5.2. *Data set 2*

Example 6.2 of [5]: Number of cycles (in thousands) of fatigue life for 67 Alloy T7987 specimens that failed before 300,000 cycles: 94, 96, 99, 99, 104, 108, 112, 114, 117, 117, 118, 121, 121, 123, 129, 131, 133, 135, 136, 139, 139, 140, 141, 141, 143, 144, 149, 149, 152, 153, 159, 159, 159, 159, 162, 168, 168, 169, 170, 170, 171, 172, 173, 176, 177, 180, 180, 184, 187, 188, 189,

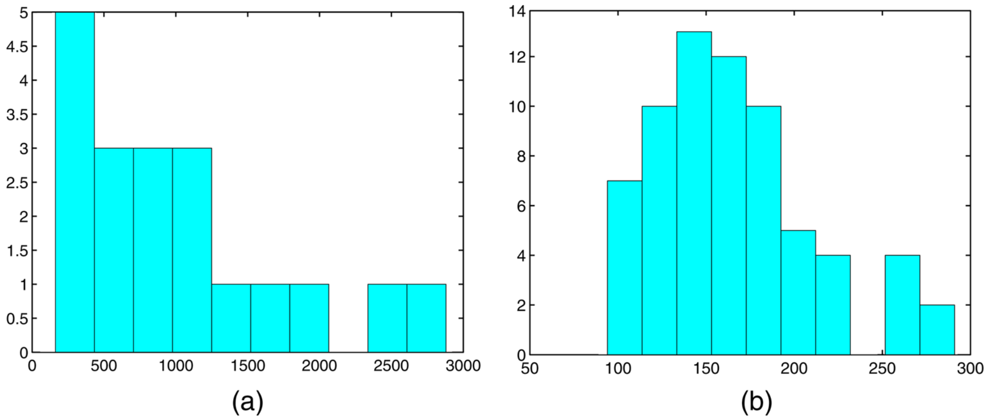


Figure 2. Histograms for the (a) data set 1 and (b) data set 2.

190, 196, 197, 203, 205, 211, 213, 224, 226, 227, 256, 257, 269, 271, 274, 291. They conclude that the lognormal distribution provides a reasonable fit for these data, but that the inclusion of a threshold parameter would improve the fit.

We compute the value of the test statistic obtaining $Q_n = 3.4337$. For a 5% significance level, the approximate asymptotic critical values are $\text{cv}_{\text{left}} = 3.46492$ and $\text{cv}_{\text{right}} = 4.26178$; hence, we conclude that the data do not follow a two-parameter exponential distribution.

Acknowledgement

This research project was partially supported by the Spanish grants SEJ2007-64500 and MTM2006-09920 (Ministry of Education and Science-FEDER).

References

[1] S. Cambanis, G. Simons, and W. Stout, *Inequalities for $\mathcal{E}k(x, y)$ when the marginals are fixed*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 36 (1976), pp. 285–294.

[2] J. Fortiana, and A. Grané, *A scale-free goodness-of-fit statistic for the exponential distribution based on maximum correlations*, J. Statist. Plann. Inference 108 (2002), pp. 85–97.

[3] ———, *Goodness-of-fit tests based on maximum correlations and their orthogonal decompositions*, J. Roy. Statist. Soc. B 65 (2003), pp. 1–12.

[4] A. Grané, and J. Fortiana, *An adaptive goodness-of-fit test*, Commun. Statist. A, Theory and Methods 35 (6) (2006), pp. 1141–1155.

[5] W.Q. Meeker, and L.A. Escobar, *Statistical Methods for Reliability Data*, John Wiley & Sons, New York, (1998).

[6] L.J. Bain, and M. Engelhardt, *Statistical Analysis of Reliability and Life-Testing Models*, Marcel Deker Inc., New York, (1991).

[7] K.C. Kapur, and L.R. Lamberson, *Reliability in Engineering Design*, John Wiley & Sons, New York, (1977).

[8] M. Kendall, and A. Stuart, *The Advanced Theory of Statistics, 2*, C. Griffin and Co, London, (1961).

[9] R.J. Serfling, *Approximation Theorems of Mathematical Statistics*, John Wiley & Sons, New York, (1980).

[10] G.R. Shorack, and J.A. Wellner, *Empirical Processes with Applications to Statistics*, John Wiley & Sons, New York, (1986).

[11] V. Choulakian, and M. Stephens, *Goodness-of-fit tests for the generalized Pareto distribution*, Technometrics 43 (2001), pp. 478–484.

[12] S. Shapiro, and M. Wilk, *An analysis of variance test for the exponential distribution (complete samples)*, Technometrics 14 (1972), pp. 355–370.

[13] M. Gail, and J. Gastwirth, *A scale-free goodness-of-fit test for the exponential distribution based on the Gini statistic*, J. Roy. Statist. Soc. B 40 (1978), pp. 350–357.

[14] M.A. Stephens, *Tests for the exponential distribution*, in Goodness-of-fit Techniques M.A. S. R. B. D'Agostino, ed., Marcel Dekker, Inc., New York, (1986), pp. 421–459.

[15] F.E. Grubbs, *Fiducial bounds on reliability for the two parameter negative exponential distribution*, Technometrics 13 (1971), pp. 873–876.