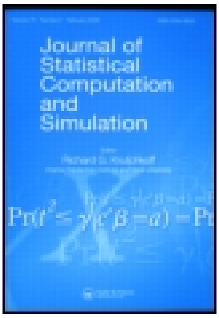
This article was downloaded by: [Regina Bispo]
On: 20 October 2011, At: 04:44
Publisher: Taylor & Francis
Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Statistical power of goodness-of-fit tests based on the empirical distribution function for type-I right-censored data

Regina Bispo [a] [b], Tiago A. Marques [b] [c] & Dinis Pestana [b] [d]

[a] Departamento de Estatística, ISPA-Instituto Universitário, Rua Jardim do Tabaco, 34, 1149-041, Lisboa, Portugal

[b] CEAUL-Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal

[c] Centre for Research into Ecological and Environmental Modeling, The Observatory, Buchanan Gardens, Saint Andrews, Scotland, UK

[d] Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa, Portugal

Available online: 20 Oct 2011

To cite this article: Regina Bispo, Tiago A. Marques & Dinis Pestana (2011): Statistical power of goodness-of-fit tests based on the empirical distribution function for type-I right-censored data, Journal of Statistical Computation and Simulation, DOI:10.1080/00949655.2011.624519

To link to this article: http://dx.doi.org/10.1080/00949655.2011.624519

Journal of Statistical Computation and Simulation
iFirst, 2011, 1–9

Statistical power of goodness-of-fit tests based on the empirical distribution function for type-I right-censored data

Regina Bispo[a,b]*, Tiago A. Marques[b,c] and Dinis Pestana[b,d]

[a] *Departamento de Estatística, ISPA-Instituto Universitário, Rua Jardim do Tabaco, 34, 1149-041 Lisboa, Portugal;* [b] *CEAUL-Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal;* [c] *Centre for Research into Ecological and Environmental Modeling, The Observatory, Buchanan Gardens, Saint Andrews, Scotland, UK;* [d] *Departamento de Estatística e Investigação Operacional, Faculdade de Ciências, Universidade de Lisboa, Campo Grande, Lisboa, Portugal*

(*Received 15 March 2011; final version received 14 September 2011*)

In this study, the power of common goodness-of-fit (GoF) statistics based on the empirical distribution function (EDF) was simulated for single type-I right-censored data. The statistical power of the Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling statistics was investigated by varying the null and the alternative distributions, the sample size, the degree of censoring and the significance level. The exponential, Weibull, log-logistic and log-normal lifetime distributions were considered as they are among the most frequently distributions used when modelling censored data. We conclude by giving some general recommendations for testing the distributional assumption of parametric survival models in homogeneous populations when using EDF-based GoF statistics.

Keywords: censored data; goodness-of-fit; lifetime distributions; power; type-I censoring

AMS Subject Classifications: 62G10; 62N99

1. Introduction

Lifetime or failure time data refer to the time until the occurrence of some event of interest. This type of data emerges frequently in many areas, such as engineering, bioscience and biomedical sciences. A difficulty frequently associated with this type of data is the presence of some subjects for which the exact time of failure is unknown, with only a lower (or an upper) lifetime bound being available. In particular, when data are obtained over a fixed time period, some subjects' time is only noted as being less than some predetermined value. This type of data is said to be type-I censored. More formally, a type-I censored sample arises when n subjects are observed for the limited periods of time $L_1, L_2, \ldots, L_n$ so that an individual's lifetime $T_i \leq L_i$ $(i = 1, \ldots, n)$.

*Corresponding author. Email: rbispo@ispa.pt

ISSN 0094-9655 print/ISSN 1563-5163 online

http://dx.doi.org/10.1080/00949655.2011.624519
http://www.tandfonline.com

Such data can be represented by the n pairs of variables (t_i, δ_i), where

$$t_i = \min(T_i, L_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if } T_i \leq L_i, \\ 0 & \text{if } T_i > L_i. \end{cases}$$

When $L_1 = L_2 = \cdots = L_n$, data are said to be single type-I censored.

Lifetime data can be modelled by using either semiparametric or parametric approaches. Parametric methods, assuming a specific form for the underlying data distribution, may achieve more precise inferences [1]. In particular, Efron [2] and Oakes [3] showed that under certain conditions, parametric survival models can lead to more efficient parameter estimates than a semiparametric modelling approach. The drawback is that in this framework the used model is assumed to be correct and estimates, and hence the corresponding variances, depend on the validity of the distributional assumption and robustness concerns do arise. The assumption of an *a priori* specific distribution model may affect the accuracy of the estimation and the inference procedures [4]. Hence, one of the most important aspects when using parametric survival methods is the selection of the lifetime distribution that one expects to be governing the generation of the data.

Goodness-of-fit (GoF) tests are a formal procedure to test the significance of the discrepancy between an empirical distribution function (EDF) $(F_n(x))$ and a particular distribution function $(F_0(x))$. Let F be a continuous cumulative distribution function. The hypothesis under test are $H_0 : F(x) = F_0(x)$ against the alternative $H_1 : F(x) \neq F_0(x)$. There are several types of statistical GoF tests. The reader can find a broad discussion on these tests in [5]. GoF statistics based on the EDF such as the Kolmogorov–Smirnov (D), the Cramér–von Mises (W^2) and the Anderson–Darling (A^2) are among the most commonly used statistics.

Although these types of procedures are widely used in research, investigations about the power of these test statistics for censored data are scarce. The power of a GoF statistic is the conditional probability of correctly rejecting a null distribution given a true alternative distribution. Information about the power of GoF statistics is important as it governs the choice of the test to be used when checking the GoF of the models. In this study, the power of the traditional GoF statistics D, W^2 and A^2 is investigated by varying the null and the alternative distributions (completely specified), the sample size, the significance level and the degree of censoring.

1.1. *GoF statistics based on the EDF*

1.1.1. *Kolmogorov–Smirnov statistic*

For single type-I censoring at point L, the Kolmogorov–Smirnov statistic [5–7] is defined by

$$D_{n,p} = \sup_{-\infty < x \leq L} |\tilde{F}_n(x) - F_0(x)|, \tag{1}$$

where $p = F_0(L)$. Given the order statistics $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, this statistic has the useful alternative form for computational purposes

$$D_{n,p} = \max_{1 < x \leq r} \left[\max \left\{ \frac{i}{n} - F_0(x_{(i)}), F_0(x_{(i)}) - \frac{i-1}{n} \right\} \right], \tag{2}$$

with r representing the number of observations less than or equal to L (i.e. the number of uncensored observations).

1.1.2. *Anderson–Darling statistic*

The Anderson–Darling statistic [5,7–9] is, in fact, a modification of the Kolmogorov–Smirnov test statistic introduced to give a different weight to the distance $|\tilde{F}_n(x) - F_0(x)|$ placing a higher weight in the tails of the underlying distribution [10]. This statistic is defined as

$$A^2_{n,p} = n \int_{-\infty}^{L} \frac{[\tilde{F}_n(x) - F_0(x)]^2}{F_0(x)[1 - F_0(x)]} \,\mathrm{d}F_0(x), \tag{3}$$

with an alternative form

$$\begin{aligned} A^2_{n,p} = &-\sum_{i=1}^{r} \left(\frac{2i-1}{n}\right) [\log(1 - F_0(x_{(i)})) - \log(F_0(x_{(i)}))] - 2\sum_{i=1}^{r} \log(1 - F_0(x_{(i)})) \\ &+ n\left[\frac{2r}{n} - \left(\frac{r}{n}\right)^2 - 1\right] \log(1 - F_0(L)) + \frac{r^2}{n}\log p - nF_0(L). \end{aligned} \tag{4}$$

1.1.3. *Cramér–von Mises statistic*

The Cramér–von Mises statistic [5,7] is defined by

$$W^2_{n,p} = n \int_{-\infty}^{L} [\tilde{F}_n(x) - F_0(x)]^2 \,\mathrm{d}F_0(x), \tag{5}$$

with an alternative form for computational simplification given by

$$W^2_{n,p} = \sum_{i=1}^{r} \left(F_0(x_{(i)}) - \frac{i-0.5}{n}\right)^2 + \frac{r}{12n^2} - \frac{n}{3}\left(\frac{r}{n} - F_0(L)\right)^3. \tag{6}$$

2. Simulation study

To investigate the power of the mentioned GoF statistics, we conducted a Monte Carlo simulation study by varying the null and the alternative distributions. The tested lifetime models included the exponential, Weibull, log-logistic and log-normal distributions as these are among the most frequently used distributions when modelling censored data [1]. The probability density function for the exponential distribution with parameter ρ ($\rho > 0$) and mean $1/\rho$ is defined by

$$f(t) = \rho \exp\{-\rho t\} \quad (t > 0). \tag{7}$$

For the Weibull distribution with scale parameter ρ ($\rho > 0$) and shape parameter γ ($\gamma > 0$), the density function is given by

$$f(t) = \gamma\rho t^{(\gamma-1)} \exp\{-\rho t^{\gamma}\} \quad (t > 0). \tag{8}$$

The log-logistic distribution, with scale ρ ($\rho > 0$) and shape κ ($\kappa > 0$), has density

$$f(t) = \frac{\kappa\rho^{\kappa} t^{(\kappa-1)}}{[1 + (\rho t)^{\kappa}]^2} \quad (t > 0). \tag{9}$$

The density function of a log-normal random variable T with location and scale parameters μ and σ ($\sigma > 0$) can be written as

$$f(t) = \frac{1}{\sqrt{2\pi}\sigma t} \exp\left\{-\frac{1}{2}\left(\frac{\log t - \mu}{\sigma}\right)^2\right\} \quad (t > 0). \tag{10}$$

As alternative lifetime distributions, we used the exponential distribution with a parameter of 0.3, the Weibull distribution with a shape parameter of 0.5 and a scale parameter of 2, the log-logistic distribution with a shape parameter of 2 and a scale parameter of 1.5 and the log-normal distribution with location 2 and a scale parameter of 0.8. These specific distributions cover monotonic decreasing and asymmetric unimodal densities. The models exhibit a range of possible hazard behaviours (constant, monotonic decreasing and asymmetric with positive mode) covering frequent situations in many study areas. The use of these models enables a systematic comparison to show the strong and weak points arising from the use of each of the models when an alternative model would be a better fit.

The power of the GoF tests was estimated by the proportion of the correct rejections of H_0. The power of each statistic was simulated from 10,000 replications. The sample size included the values from 10 to 100 (with a step of 10) and from 100 to 200 (with a step of 50) to cover a wide range of possible sizes for data sets. The significance level was fixed at 0.05 and 0.10 and the proportion of uncensored observations (r/n) was taken between 0.3 and 0.9 (with a step of 0.2).

To find the critical values for the GoF statistics, we simulated 10,000 random samples that were type-I right censored, for each of the censoring scenarios, from a uniform $U(0, 1)$ population. If F_0 represents the cumulative distribution function of the $U(0, 1)$, then for the test statistic Q, critical values for the upper tail can be found, such as $\alpha = P(Q > q_{Q,n,1-\alpha} | F \equiv F_0)$ [11]. To check the accuracy of the critical values, we tested the data under the null hypothesis.

In this study, we describe the results obtained by simulating censored samples drawn from the exponential (Case I), the Weibull (Case II), the log-logistic (Case III) and the log-normal (Case IV) populations. Although we have performed the simulations using both the significance levels of 0.05 and 0.10, only the results concerning the 0.05 level are presented as a similar behaviour was observed under both conditions.

3. Results

Table 1 shows the proportions of correct H_0 rejections for the significance level of 0.05. As in this case, the null hypothesis is true, it is expected that the statistics maintain the type-I error rate. Overall, we found very small differences between the nominal level 0.05 and the actual levels, which shows a reliable performance of the studied GoF statistics. The highest deviations were consistently found for smallest sample sizes and/or higher censoring rates.

Table 1. Significance levels under the null hypothesis at a nominal level of 0.05 (ranges reflect variation according to sample size).

Distribution	r/n	$D_{n,p}$	$A^2_{n,p}$	$W^2_{n,p}$
Exponential	0.90	0.042–0.055	0.048–0.060	0.039–0.055
	0.70	0.041–0.052	0.045–0.055	0.038–0.054
	0.50	0.041–0.053	0.041–0.054	0.041–0.051
	0.30	0.043–0.050	0.042–0.054	0.041–0.051
Weibull	0.90	0.047–0.056	0.045–0.064	0.043–0.055
	0.70	0.040–0.055	0.045–0.053	0.037–0.052
	0.50	0.044–0.052	0.042–0.050	0.036–0.049
	0.30	0.034–0.048	0.042–0.058	0.034–0.046
Log-logistic	0.90	0.043–0.055	0.043-0.058	0.039–0.055
	0.70	0.043–0.055	0.043–0.053	0.040–0.051
	0.50	0.041–0.050	0.044–0.053	0.042–0.049
	0.30	0.029–0.051	0.042–0.056	0.029–0.046
Log-normal	0.90	0.044–0.056	0.043–0.066	0.040–0.055
	0.70	0.040–0.053	0.044–0.053	0.040–0.053
	0.50	0.041–0.057	0.042–0.053	0.042–0.049
	0.30	0.032–0.050	0.040–0.061	0.033–0.047

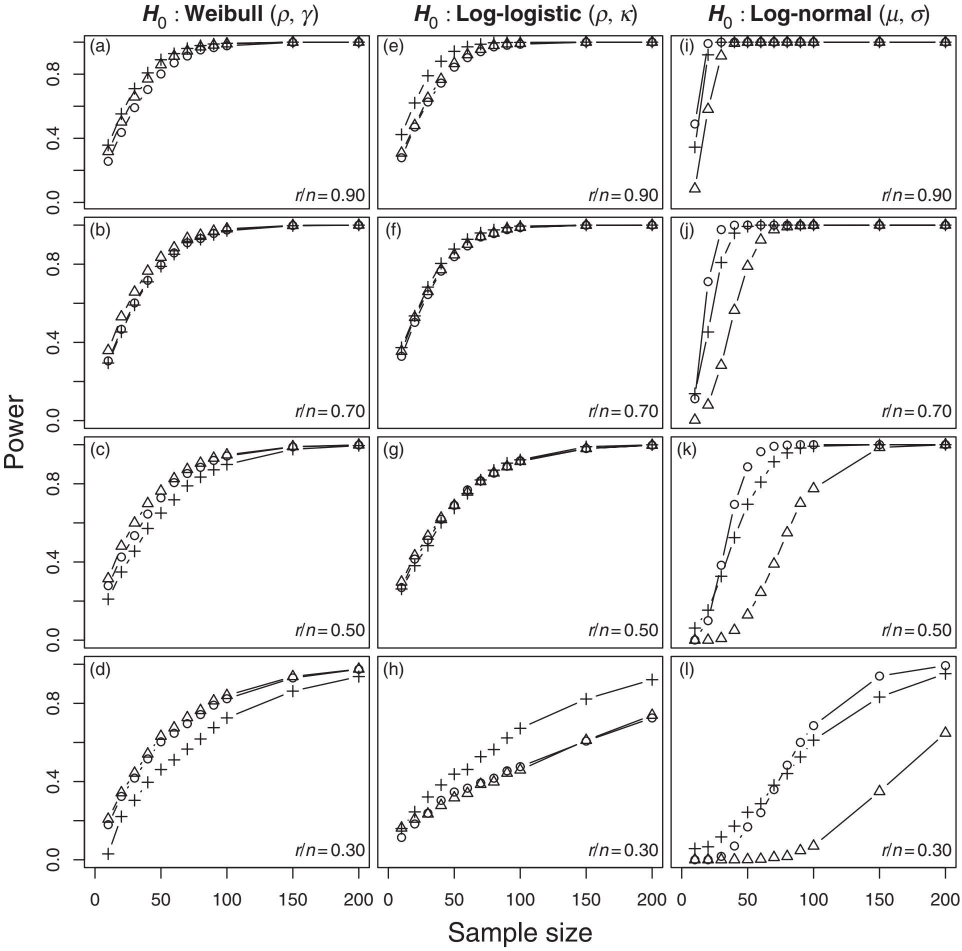


Figure 1. Statistical power when testing H_0: Weibull (ρ, γ), H_0: log-logistic (ρ, κ) and H_0: log-normal (μ, σ) versus H_1: exponential (0.3) as a function of the sample size and the proportion of uncensored observations (r/n) at a 0.05 significance level ($\circ$, $D_{n,p}$; $\triangle$, $W^2_{n,p}$; $+$, $A^2_{n,p}$).

We now describe the results found for each of the previously described scenarios.

Case I We simulated samples from an exponential population and tested the GoF for the Weibull, log-logistic and log-normal distributions. The obtained results are presented graphically in Figure 1.

The statistical power of the studied GoF statistics increased, as expected, with the increase in the sample size. For small sample sizes, namely under 50, the power of all statistics was almost always under 0.8. Overall, the power decreased with the censoring degree. This effect was particularly clear under the null log-logistic (Figure 1(e)–(h)) or the null log-normal (Figure 1(i)–(l)) distributions. The performance of the Cramér–von Mises statistic was greatly affected by the increase in the proportion of censored observations, which occurs particularly when testing a null log-normal distribution.

Case II We simulated samples from a Weibull population and tested the GoF for the exponential, log-logistic and log-normal distributions. Figure 2 shows the results obtained at a 0.05 significance level.

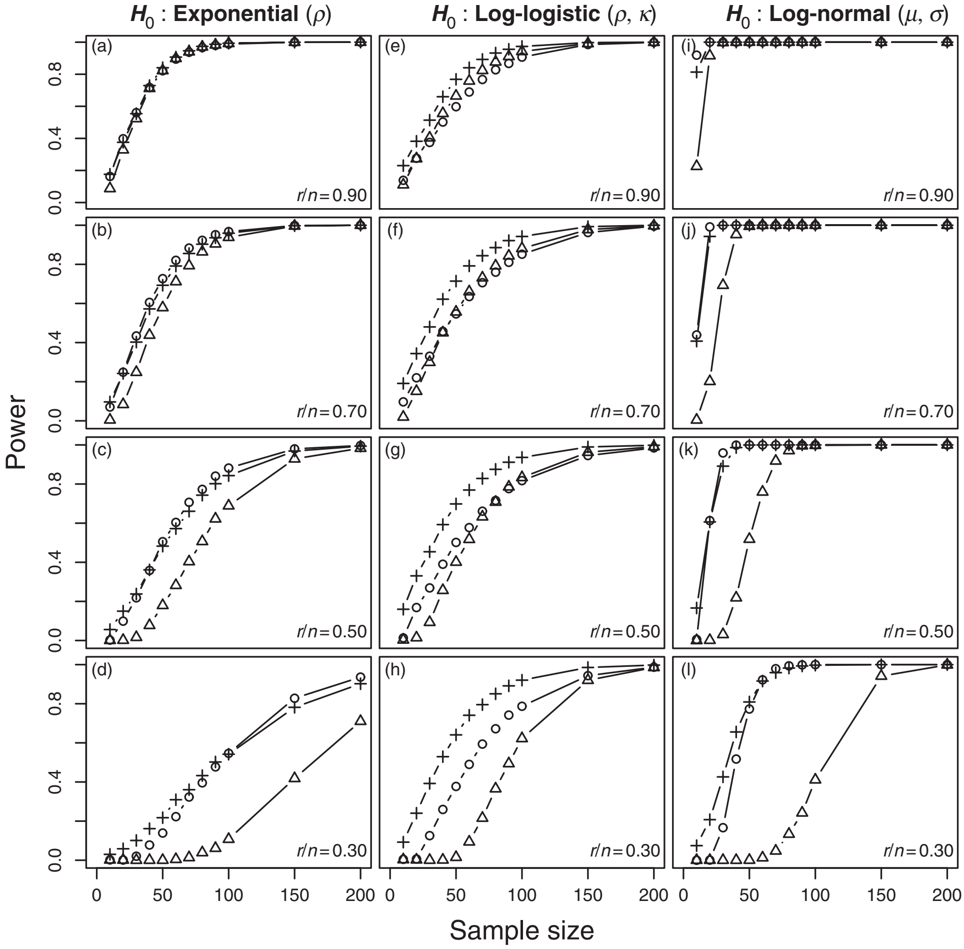


Figure 2. Statistical power when testing H_0: exponential (ρ), H_0: log-logistic (ρ, κ) and H_0: log-normal (μ, σ) versus H_1: Weibull $(2, 0.5)$ as a function of the sample size and the proportion of uncensored observations (r/n) at a 0.05 significance level (∘, $D_{n,p}$; △, $W^2_{n,p}$; +, $A^2_{n,p}$).

The efficiency of GoF statistics increased with the sample size. The rate of this increase was, nonetheless, smaller when testing the adjustment to the exponential (Figure 2(a)–(d)) or the log-logistic (Figure 2(e)–(h)) distribution than when testing the adjustment to the log-normal distribution (Figure 2(i)–(l)). The power of the studied statistics was clearly smaller for higher degrees of right censoring, regardless of the null distribution. The Cramér–von Mises statistic consistently presented the lowest power results. For low proportions of uncensored observations (Figure 2(d), (h) and (l)) and small sample sizes ($n < 50$), this statistic did not discriminate the null from the alternative distribution. High censoring rates (Figure 2(d), (h) and (l)) required large samples to correctly reject the null distribution in favour of the alternative distribution, particularly in the case of testing a null exponential distribution (Figure 2(d)).

Case III We simulated samples from a log-logistic population and tested the GoF for the exponential, Weibull and log-normal distributions. Figure 3 shows the results obtained at a 0.05 significance level.

The statistical power of the GoF statistics increased with the sample size. The three statistics presented very similar patterns for low censoring rates (Figure 3(a), (e) and (i)). In this case

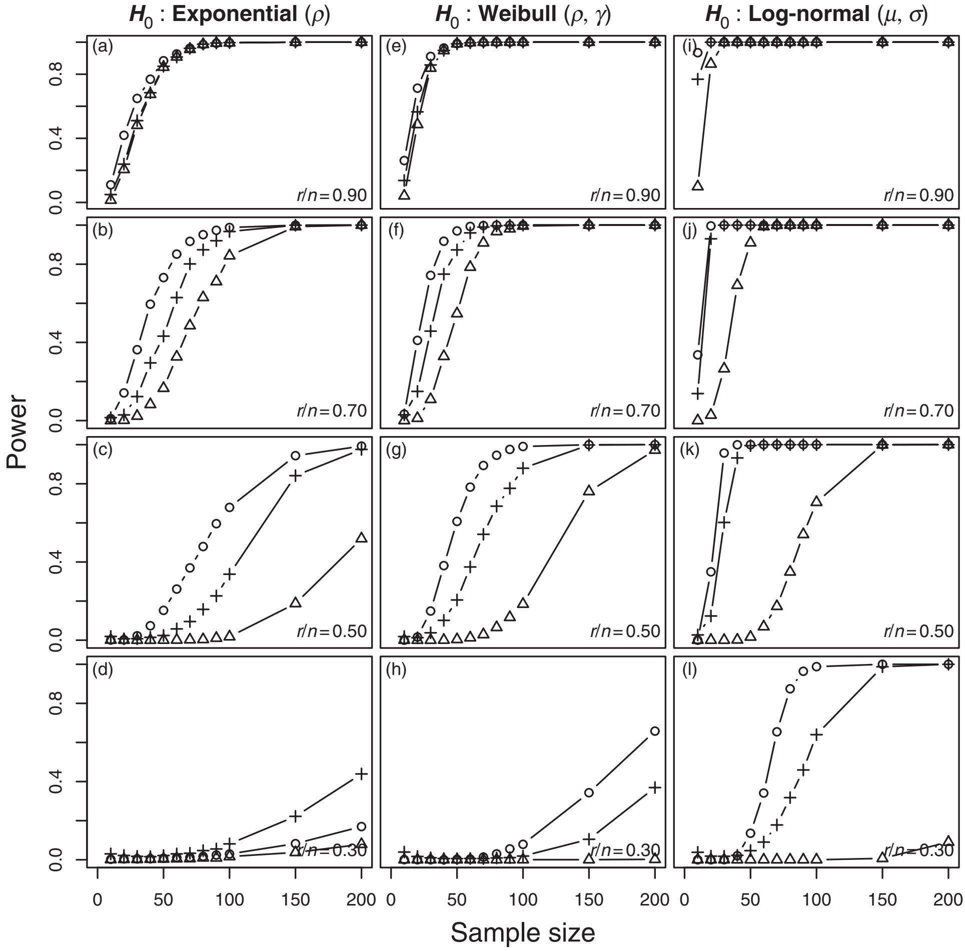


Figure 3. Statistical power when testing H_0: exponential (ρ), H_0: Weibull (ρ, γ) and H_0: log-normal (μ, σ) versus H_1: log-logistic (1.5, 2) as a function of the sample size and the proportion of uncensored observations (r/n) at a 0.05 significance level (○, $D_{n,p}$; △, $W^2_{n,p}$; +, $A^2_{n,p}$).

($r/n = 0.90$), samples with a number of cases higher than 30 gave good power levels, correctly rejecting the null distribution in favour of the alternative distribution. With the increase of the censoring rate, the efficiency of statistics clearly decreased. For a proportion of uncensored observations of only 30%, the studied statistics did not discriminate the exponential (Figure 3(d)) or the Weibull (Figure 3(h)) distribution from the log-logistic distribution, regardless of the sample size. In particular, the Cramér-von Mises statistic was again the statistic that was most affected by the censoring degree with power levels close to zero for $r/n = 0.30$ (Figure 3(d), (h) and (l)).

Case IV We simulated samples from a log-normal population and tested the GoF for the exponential, Weibull and log-logistic distributions. Figure 4 shows the results obtained at a 0.05 significance level.

Again as expected, power increased with the sample size and decreased when the proportion of censored observations increased. For relatively low censoring rates (r/n above 0.7), the decreasing effect of the increasing proportion of censored observations on the power test was difficult to be observed under a null Weibull distribution (Figure 4(e) versus (f)) or a null log-logistic distribution (Figure 4(i) versus (j)). Higher censoring rates imposed a more evident power decrease,

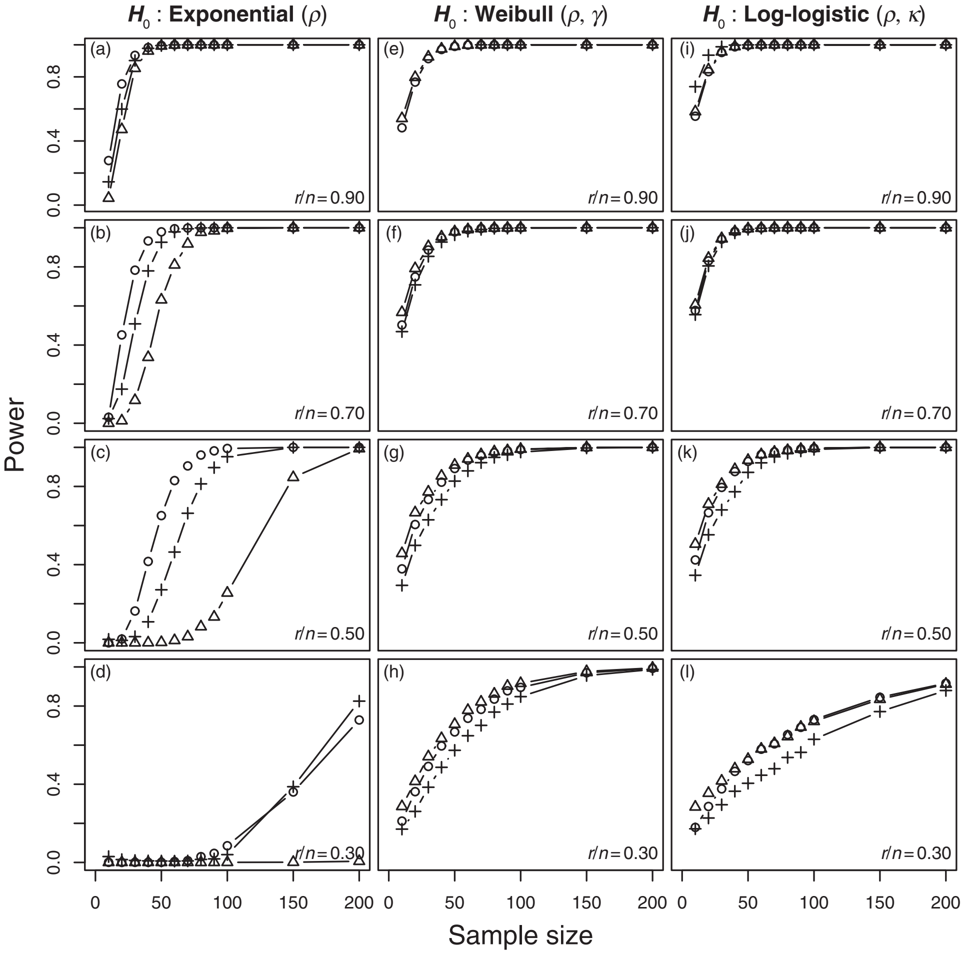


Figure 4. Statistical power when testing H_0: exponential (ρ), H_0: Weibull (ρ, γ) and H_0: log-logistic (ρ, κ) versus H_1: log-normal $(2, 0.8)$ as a function of the sample size and the proportion of uncensored observations (r/n) at a 0.05 significance level ($\circ$, $D_{n,p}$; $\triangle$, $W^2_{n,p}$; $+$, $A^2_{n,p}$).

depending, however, on the null distribution. For a null exponential (Figure 4(a)–(d)), the fact that the power decreased with the increase in the censoring degree is quite evident. For a low censoring degree ($r/n = 0.90$), the exponential distribution was correctly rejected for sample sizes above 30 (Figure 4(a)). For higher censoring rates, acceptable power levels required much larger samples.

Differences between the statistics performance were minor when testing the GoF for the Weibull (Figure 4(e)–(h)) and the log-logistic (Figure 4(i)–(l)) distributions. When testing a null exponential, the Cramér–von Mises statistic had the poorest performance, particularly low, for higher censoring rates (Figure 4(d)).

4. Concluding remarks

In general, under severe censoring conditions, the Kolmogorov–Smirnov and Anderson–Darling statistics show higher power levels than the Cramér–von Mises statistic. Hence, it seems advisable to avoid the use of this statistic when dealing with type-I right-censored data. In most cases, sample sizes above 50 are required to reach power levels above 0.80. The degree of right censoring influence on the performance of the statistics varies with the sample size. Larger sample sizes may cushion the negative impact effect that censoring has on the statistics performance. Whenever possible, a 0.10 level of significance seems to be advisable, especially when dealing with small sample sizes and higher censoring rates, to increase power levels. In this study, the power of some classical GoF statistics was established for the situation of $F_0(x)$ completely known. In this case, the sampling distribution of the GoF statistics was independent of the population probability distribution from which observations were drawn. These statistics are, therefore, said to be distribution free. However, if $F_0(x)$ has unknown parameters, this is no longer true, which limits the application of the present study results. Hence, further research is needed to extend the study to the case of complex hypotheses.

Acknowledgements

This research was partially sponsored by national funds through the Fundação Nacional para a Ciência e Tecnologia, Portugal – FCT under the project PEst-OE/MAT/UI0006/2011. We are thankful to Professor M. Ivette Gomes for useful advice. Comments during the presentation at LinStat 2010 have been helpful in improving the results presented in this paper.

References

[1] D. Collett, *Modelling Survival Data in Medical Research*, Chapman and Hall, Boca Raton, FL, 2003.

[2] B. Efron, *The efficiency of Cox's likelihood function for censored data*, J. Amer. Statist. Assoc. 72 (1977), pp. 557–565.

[3] D. Oakes, *The asymptotic information in censored survival data*, Biometrika 64 (1977), pp. 441–448.

[4] X. Romao, R. Delgado, and A. Costa, *An empirical power comparison of the univariate goodness-of-fit tests for normality*, J. Stat. Comput. Simul. 80 (2010), pp. 545–591.

[5] R.B. D'Agostino and M.A. Stephens, *Goodness-of-fit Techniques*, Dekker, New York, 1986.

[6] W.J. Conover, *Pratical Nonparametric Statistics*, Wiley, New York, 1999.

[7] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, Wiley, New York, 2003.

[8] T.W. Anderson and D.A. Darling, *Asymptotic theory of certain 'goodness of fit' criteria based on stochastic processes*, Ann. Math. Statist. 23 (1952), pp. 193–212.

[9] A. Pettitt, *Goodness-of-fit tests for discrete and censored data based on the empirical distribution function*, PhD Thesis, University of Nottingham, 1973.

[10] T. Thadewald and H. Buning, *Jarque–Bera test and its competitors for testing normality – a power comparison*, J. Appl. Stat. 34 (2007), pp. 87–105.

[11] Y. Marhuenda, D. Morales, and C. Pardo, *A comparison of uniformity tests*, Statistics 39 (2005), pp. 315–328.