On Some Goodness-of-Fit Tests and Their Connection to Graphical Methods with Uncensored and Censored Data

Claudia Castro-Kuriss[1], Mauricio Huerta[3], Víctor Leiva[3(✉)], and Alejandra Tapia[2]

[1] Instituto Tecnológico de Buenos Aires, Buenos Aires, Argentina
[2] Faculty of Basic Sciences, Universidad Católica del Maule, Talca, Chile
[3] School of Industrial Engineering, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile
http://www.victorleiva.cl

Abstract. In this work, we present goodness-of-fit tests related to the Kolmogorov-Smirnov and Michael statistics and connect them to graphical methods with uncensored and censored data. The Anderson-Darling test is often empirically more powerful than the Kolmogorov-Smirnov test. However, the former one cannot be related to graphical tools by means of probability plots, as the Kolmogorov-Smirnov test does. The Michael test is, in some cases, more powerful than the Anderson-Darling and Kolmogorov-Smirnov tests and can also be related to probability plots. We consider the Kolmogorov-Smirnov and Michael tests for detecting whether any distribution is suitable or not to model censored or uncensored data. We conduct numerical studies to show the performance of these tests and the corresponding graphical tools. Some comments related to big data and lifetime analysis, under the context of this study, are provided in the conclusions of this work.

Keywords: Anderson-Darling Kolmogorov-Smirnov and Michael tests · Big data · Censored data · Test power

1 Introduction

Several efforts have been conducted to develop goodness-of-fit (GOF) techniques that allow us to handle the problem of fitting distributions to different types of data. In general, GOF tests permit us to assess whether the distribution under a null hypothesis (H_0) is adequate to model a data set or not. For this hypothesis, there are two options: (i) the distribution can be completely specified (known parameters) or (ii) some (or all) of its parameters are unknown. In the second case, the parameters need to be properly estimated, for example, with the maximum likelihood (ML) method. Depending on the distribution under H_0, ML estimates of its parameters cannot be easily calculated and iterative numerical

procedures must be used. However, problems of convergence can arise, which are not yet completely studied; see [6] for more details about this.

According to [12, pp. xi–xiii], GOF methods can be of graphical or inferential (tests) type, where the corresponding tests are based on distances which use the empirical cumulative distribution function (ECDF), among other types of GOF methods. Most of the test statistics used for assessing GOF, such as Anderson-Darling (AD) and Kolmogorov-Smirnov (KS), compare the ECDF and the hypothesized theoretical cumulative distribution function (CDF) assumed for the data. The AD test is often more powerful than the KS test. Note that the AD test is more sensitive to detect discrepancies at the tails of the distribution, whereas the KS test does it at the center of the distribution. For more details about the AD and KS statistics, see [12, Chap. 4].

A graph that allows us to relate the ECDF with a specified theoretical CDF is the probability versus probability (PP) plot. Analogously, ordered observations corresponding to empirical quantiles can be plotted versus the theoretical quantiles of a specified distribution in a graph known as the quantile versus quantile (QQ) plot [23]. The KS test is related to the PP and QQ plots [10]. However, a disadvantage of the PP plot associated with the KS test is that some points in this graph can have more variability than others. Michael [24] proposed a modification of the KS test based on the arcsin transformation to stabilize the variance of the points in the PP plot. The graph related to this variance stabilizing transformation is known as the stabilized probability (SP) plot and the statistic associated with the Michael test is denoted as MI. [24] studied the MI test and showed that it results to be more powerful than the KS test for certain alternative hypotheses.

In reliability and survival analysis, and also in other areas, it is frequent to find situations where not all the individuals or instruments on examination complete the event under study, which can be called without loss of generality as a "failure". Samples involving such situations are named as "censored". Several books and articles on the GOF topic related to censored and uncensored can be cited; see, for example, [15–18].

When a parametric statistical analysis with censored data needs to validate its distributional assumption, the classical GOF test statistics need to be adapted to consider censorship following two options. The first of them consists of using GOF tests for uncensored data adapting the type-II right censored data to become an uncensored (complete) data sample, while the second option adapts the test statistics to type-II right censored data [5,12,21].

[8,9] proposed GOF tests for the lognormal and normal distributions with type-II right censored data. [7] studied GOF tests for location-scale distributions with type-II right censored data and unknown parameters. Other works on the topic based on different types of censoring are attributed to [4,21,25,26].

The main objectives of this work are: (i) to present GOF tests related to the KS and MI statistics and (ii) to connect them to PP and QQ plots with censored and uncensored data. These tools can be used for any distribution, as long as its parameters are known or properly estimated. With the provided tools, it is

possible to decide what distribution fits best the data set with censoring or not based not only on existing GOF tests, but also on graphical methods.

The remainder of this work is organized as follows. Section 2 introduces a methodology to perform GOF tests, in addition to presenting some useful transformations to carry out graphical GOF tools. Furthermore, in this section, we establish the hypotheses of interest and the corresponding test statistics to assess goodness of fit for any distribution to censored and uncensored data. In Sect. 3, applications with real-world data are provided. In Sect. 4, we discuss conclusions and future works, including a connection between big data and lifetime analysis, under the context of this study.

2 Methodology

2.1 Hypotheses

Consider the hypotheses: H_0: "the data were generated from a model with CDF F" versus H_1: "the data were not generated from that model". The hypothesized model (distribution) with CDF F is indexed by a parameter vector $\boldsymbol{\theta}$ that can contain location (μ), scale (β), shape (α) parameters, or any other parameter not necessarily of location, scale, or shape. This means that the random variable (RV) of interest T can follow any distribution. If the CDF is completely specified in H_0, that is, $\boldsymbol{\theta}$ is assumed to be known, the data must be transformed for testing uniformity. Otherwise, the parameters must be consistently estimated and the data transformed to test normality from the distribution under H_0.

2.2 GOF Test Statistics with Uncensored Data

In order to test H_0 established in Subsect. 2.1, when F is completely specified, and then to assess goodness of fit for a distribution to a censored or uncensored data set, we consider test statistics based on the ECDF defined as

$$F_n(t) = \frac{1}{n}\#\{j\colon t_j \leq t\},$$

where n is the size of simple and $\#A$ denotes de cardinality of the set A. The most common statistics constructed with the ECDF use vertical distances, that is, between F_n and F by means of the supremum and quadratic classes. Statistics that consider the supremum class are KS and MI given by

$$\mathrm{KS} = \sup_t \left|F_n(t) - F(t)\right| = \max\left\{\sup_t \left\{F_n(t) - F(t)\right\}, \sup_t \left\{F(t) - F_n(t)\right\}\right\}, \tag{1}$$

$$\mathrm{MI} = \max\left\{\sup_t \left\{\frac{2}{\pi}\arcsin(F_n(t)) - \frac{2}{\pi}\arcsin(F(t))\right\}, \sup_t \left\{\frac{2}{\pi}\arcsin(F(t)) - \frac{2}{\pi}\arcsin(F_n(t))\right\}\right\}. \tag{2}$$

Now, consider

$$U = F(T) \tag{3}$$

follows a uniform distribution on [0,1], denoted by U(0,1), for any continuous F, which is known as probability integral transformation. Then, KS and MI statistics defined in (1) and (2) can be implemented in practice by the formulas

$$\mathrm{KS} = \max\left\{\max_{1\le j\le n}\left\{w_{j:n} + \frac{1}{2n} - U_{j:n}\right\},\ \max_{1\le j\le n}\left\{U_{j:n} - w_{j:n} + \frac{1}{2n}\right\}\right\}, \quad (4)$$

$$\mathrm{MI} = \max\left\{\max_{1\le j\le n}\left\{\frac{2}{\pi}\arcsin\left(w_{j:n} + \frac{1}{2n}\right) - \frac{2}{\pi}\arcsin(U_{j:n})\right\},\right.$$
$$\left.\max_{1\le j\le n}\left\{\frac{2}{\pi}\arcsin(U_{j:n}) - \frac{2}{\pi}\arcsin\left(w_{j:n} - \frac{1}{2n}\right)\right\}\right\}, \quad (5)$$

where

$$w_{j:n} = \frac{j - 0.5}{n} \quad (6)$$

and $U_{j:n} = F(T_{j:n})$ is the jth order statistic (OS) of a sample of size n extracted from an RV $U \sim \mathrm{U}(0,1)$, with $u_{j:n} = F(t_{j:n})$ being its observed value, for $j = 1,\ldots,n$. More details about expressions provided in (1) and (5) can be found in [12, Chap. 4] and [24]. Quantiles of the distribution of the KS statistic must be obtained under H_0. However, if the distribution under this hypothesis is not completely specified, its parameters must be properly estimated and the KS and MI statistics must be modified for the distribution under H_0. These modified statistics are denoted by $\mathrm{KS}^\star$ and $\mathrm{MI}^\star$, whereas their calculated values by $\mathrm{ks}^\star$ and $\mathrm{mi}^\star$, respectively. In this case, new quantiles of the distributions of $\mathrm{KS}^\star$ and $\mathrm{MI}^\star$ must be computed under H_0.

2.3 GOF Test Statistics with Censored Data

To test H_0 when F is completely specified and then to assess goodness of fit in practice with r uncensored data and $n - r$ type-II right censored data, we use the results presented in [12, Chap. 4] and adapt the statistics given in (4) and (5) as

$$\mathrm{KS}_{r,n} = \max\left\{\max_{1\le j\le r}\left\{w_{j:n} + \frac{1}{2n} - U_{j:n}\right\}, \max_{1\le j\le r}\left\{U_{j:n} - w_{j:n} + \frac{1}{2n}\right\}\right\}, \quad (7)$$

$$\mathrm{MI}_{r,n} = \max\left\{\max_{1\le j\le r}\left\{\frac{2}{\pi}\arcsin\left(w_{j:n} + \frac{1}{2n}\right) - \frac{2}{\pi}\arcsin(U_{j:n})\right\},\right.$$
$$\left.\max_{1\le j\le r}\left\{\frac{2}{\pi}\arcsin(U_{j:n}) - \frac{2}{\pi}\arcsin\left(w_{j:n} - \frac{1}{2n}\right)\right\}\right\}. \quad (8)$$

The quantiles of the distribution of the $\mathrm{KS}_{r,n}$ and $\mathrm{MI}_{r,n}$ statistics given in (7) and (8) must be obtained under H_0. However, if the distribution under H_0 is not completely specified, its parameters must be properly estimated, taking into account the censorship, and the statistics must be modified for each case under H_0. We denote these statistics by $\mathrm{KS}^\star_{r,n}$ and $\mathrm{MI}^\star_{r,n}$, and their calculated values by $\mathrm{ks}^\star_{r,n}$ and $\mathrm{mi}^\star_{r,n}$, respectively. Also, new quantiles of the distribution of $\mathrm{KS}^\star_{r,n}$

and $\mathrm{MI}^{\star}_{r,n}$ must be computed under H_0. For more details about how to obtain the quantiles of the distributions of the corresponding test statistics under H_0, which have been studied for different distributions of the location-scale family with uncensored and censored, see [7–9,12]. In the next subsections, we mention that, for any distribution, analogous results for assessing GOF with both uncensored and censored data can be considered.

2.4 GOF Tests for any Distribution with Uncensored Data

If the hypotheses of interest H_0 is $F(t) = \Phi((t-\mu)/\beta)$ with unknown parameters, we can consider the procedure detailed in Algorithm 1.

Algorithm 1 GOF test for normality with uncensored data

1: Collect data $t_1, \ldots, t_n$ and order them as $t_{1:n}, \ldots, t_{n:n}$.
2: Estimate μ and β of $\Phi((t-\mu)/\beta)$ by $\widehat{\mu}$ and $\widehat{\beta}$, respectively, with $t_1, \ldots, t_n$.
3: Obtain $\widehat{u}_{j:n} = \Phi(\widehat{z}_j)$, with $\widehat{z}_j = (t_{j:n} - \widehat{\mu})/\widehat{\beta}$, for $j = 1, \ldots, n$.
4: Evaluate $\mathrm{KS}^{\star}$ and $\mathrm{MI}^{\star}$ statistics at $\widehat{u}_{j:n}$.
5: Compute the p-values of the $\mathrm{KS}^{\star}$ and $\mathrm{MI}^{\star}$ statistics.
6: Reject H_0: $F(t) = \Phi((t-\mu)/\beta)$ for a specified significance level based on the obtained p-values.

We consider a procedure that can be applied to any distribution based on the work proposed by Chen and Balakrishnan [11], which provides an approximate GOF method. This method first transforms the data to normality and then applies Algorithm 1, generalizing it. Testing normality in H_0 allows us to compute the critical values of the corresponding test statistics, independently of the parameter estimators, if they are consistent and the sample size is large enough. To test the hypotheses of interest defined in Subsect. 2.1, for $\alpha > 0$ and $\beta > 0$ unknown, we consider a generalization of Algorithm 1 detailed in Algorithm 2. Following [11], we recommend in general to use a sample size $n > 20$, so that the approximations work well. This is also valid for the algorithms presented in the next sections.

Algorithm 2 GOF test for any distribution with uncensored data

1: Collect data $t_1, \ldots, t_n$ and order them as $t_{1:n}, \ldots, t_{n:n}$.
2: Estimate α and β of $F(t;\alpha,\beta)$ by $\widehat{\alpha}$ and $\widehat{\beta}$, respectively, with $t_1, \ldots, t_n$.
3: Compute $\widehat{v}_{j:n} = F(t_{j:n};\widehat{\alpha},\widehat{\beta})$, for $j = 1, \ldots, n$.
4: Calculate $\widehat{y}_j = \Phi^{-1}(\widehat{v}_{j:n})$, where Φ^{-1} is the N(0, 1) inverse CDF.
5: Obtain $\widehat{u}_{j:n} = \Phi(\widehat{z}_j)$, with $\widehat{z}_j = (\widehat{y}_j - \bar{y})/s_y$, $\bar{y} = \sum_{j=1}^{n} \widehat{y}_j/n$ and $s_y = (\sum_{j=1}^{n}(\widehat{y}_j - \bar{y})^2/(n-1))^{1/2}$.
6: Repeat Steps 4-6 of Algorithm 1 with $F(t) = F(t;\alpha,\beta)$.

2.5 GOF Tests for any Distribution with Censored Data

As mentioned, GOF tests for any distribution with uncensored data can be considered for censored data adapting them or the GOF statistics.

To test the hypotheses of interest defined in Subsect. 2.1, for $\alpha > 0$ and $\beta > 0$ both of them unknown, with type-II right censored data, we first transform censored data into uncensored data by using

$$V_{j:n} = \frac{U_{j:n}(B_{r,n-r+1}(U_{r:n}))^{1/r}}{U_{r:n}}, \quad j = 1, \ldots, r, \quad r = 1, \ldots, n, \tag{9}$$

where $B_{r,n-r+1}(x) = I_x(r, n-r+1)$ is the Beta$(r, n-r+1)$ CDF, with I_x being the incomplete beta ratio function. Hence, the OSs $V_{1:n}, \ldots, V_{r:n}$ obtained from the transformation given in (9) are distributed as the OSs from an uncensored sample of size r from $V \sim \mathrm{U}(0, 1)$. Algorithm 3 details the corresponding GOF procedure.

Algorithm 3 GOF test 1 for any distribution with censored data

1: Repeat Steps 1-3 of Algorithm 2.
2: Determine $\widehat{v}'_{j:n} = \widehat{v}_{j:n}(B_{r,n-r+1}(\widehat{v}_{r:n}))^{1/r}/\widehat{v}_{r:n}$, for $j = 1, \ldots, r$ and $r = 1, \ldots, n$.
3: Repeat Steps 4-6 of Algorithm 2 replacing $\widehat{v}_{j:n}$ by $\widehat{v}'_{j:n}$ in Step 4.

Second, as mentioned, another way to perform a GOF test for any distribution with censored data can be obtained adapting the GOF statistics, which is detailed in Algorithm 4.

Algorithm 4 GOF test 2 for any distribution with censored data

1: Repeat Steps 1-5 of Algorithm 2.
2: Evaluate $\mathrm{KS}^{\star}_{r,n}$ and $\mathrm{MI}^{\star}_{r,n}$ statistics at $\widehat{u}_{j:n}$.
3: Determine the p-values of $\mathrm{KS}^{\star}_{r,n}$ and $\mathrm{MI}^{\star}_{r,n}$ statistics.
4: Reject the corresponding H_0 for a specified significance level depending on the obtained p-values.

Next, based on Algorithm 4, we provide acceptance regions for the KS and MI statistics which allow graphical tools to be obtained for assessing goodness of fit in any distribution.

2.6 PP and SP Plots

PP and QQ plots are well known, but this is not the case of the SP plot. We recall that, if the distribution under H_0 is U(0,1), then the corresponding QQ plot is essentially the same as the PP plot [8]. [24] used the arcsin transformation to stabilize the variance of the points on probability graphs associated with the

KS test to propose the SP plot. This is due to that, if $U \sim \mathrm{U}(0,1)$, then the RV given by the SP transformation

$$S = \frac{2}{\pi} \arcsin(\sqrt{U}) \tag{10}$$

follows a distribution with probability density function (PDF) given by

$$f_S(s) = \frac{\pi}{2} \sin(\pi s), \quad 0 < s < 1.$$

The OSs $S_{1:n} \leq \cdots \leq S_{n:n}$, associated with a sample of size n from the distribution of the transformed RV S given in (10), have a constant asymptotic variance, because as n goes to infinity and j/n to q, $\mathrm{Var}(n\, S_{j:n})$ goes to $1/\pi^2$, which is independent of q, for $j = 1, \ldots, n$ [24]. Formulas to construct PP and SP plots are provided in Table 1. In this table, $u_{j:n}$ is given as in (3), $w_{j:n}$ in (6) and $s_{j:n}$ as in (10).

Table 1. Formulas for the indicated probability plot.

Plot	Abscissa	Ordinate
PP	$w_{j:n}$	$u_{j:n}$
SP	$x_{j:n} = \frac{2}{\pi} \arcsin\left(\sqrt{w_{j:n}}\right)$	$s_{j:n} = \frac{2}{\pi} \arcsin\left(\sqrt{u_{j:n}}\right)$

2.7 Acceptance Regions for Probability Plots

Acceptance regions for PP and SP plots can be constructed by means of KS and MI statistics. Thus, we can display acceptance bands to assess whether the data can come from the distribution under H_0 with these two statistics [8,9]. Formulas to construct $100 \times \varrho\%$ acceptance regions on PP and SP plots with right type-II censored data, based on $\mathrm{KS}^{\star}_{r,n}$ and $\mathrm{MI}^{\star}_{r,n}$ and where ϱ is the significance level, are displayed in Table 2. In this table, w and x are continuous versions of $w_{j:n}$ and $x_{j:n}$, respectively, given in Table 1 to construct the acceptance bands. If all of the r data points lie inside the constructed acceptance bands, then H_0 cannot be rejected at the ϱ level. Also, if a noticeable curvature is detected, we can question such a hypothesis. Table 2 may be adapted to the uncensored case with $r = n$ and the quantiles must be replaced by the quantiles of the distribution of the corresponding statistics without censorship.

To test H_0 defined in Subsect. 2.1, for some $\alpha > 0$ and $\beta > 0$, with type-II right censored data, we consider a graphical tool whose procedure is detailed in Algorithm 5 based on Algorithm 4 and Tables 1 and 2, which is valid for censored or uncensored data. We consider the general case for unknown parameters of the distribution under H_0, but it can also be used when the parameters are known.

Table 2. $100\times\varrho\%$ acceptance regions for the indicated plot and statistic with $100\times\varrho$th quantiles $\mathrm{ks}^{\star}_{r,n,\varrho}$ and $\mathrm{mi}^{\star}_{r,n,\varrho}$.

Plot	Stat	Bands defining acceptance regions
PP	KS*	$[\max\{w-\mathrm{ks}^{\star}_{r,n,\varrho}+\frac{1}{2n},0\},\min\{w+\mathrm{ks}^{\star}_{r,n,\varrho}-\frac{1}{2n},1\}]$
PP	MI*	$[\max\{\sin^2(\arcsin(w^{\frac{1}{2}})-\frac{\pi}{2}\mathrm{mi}^{\star}_{r,n,\varrho}),0\},\min\{\sin^2(\arcsin(w^{\frac{1}{2}})+\frac{\pi}{2}\mathrm{mi}^{\star}_{r,n,\varrho}),1\}]$
SP	KS*	$[\max\{\frac{2}{\pi}\arcsin(\{\sin^2(\frac{\pi}{2}x)-\mathrm{ks}^{\star}_{r,n,\varrho}+\frac{1}{2n}\}^{\frac{1}{2}}),0\},\min\{\frac{2}{\pi}\arcsin(\{\sin^2(\frac{\pi}{2}x)+\mathrm{ks}^{\star}_{r,n,\varrho}-\frac{1}{2n}\}^{\frac{1}{2}}),1\}]$
SP	MI*	$[\max\{x-\mathrm{mi}^{\star}_{r,n,\varrho},0\},\min\{x+\mathrm{mi}^{\star}_{r,n,\varrho},1\}]$

Algorithm 5 Acceptance regions to test goodness of fit for any distribution with censored data

1: Repeat Step 1 of Algorithm 4.
2: Draw the PP plot with points $w_{j:n}$ versus $\widehat{u}_{j:n}$, for $j=1,\ldots,r$ and $r=1,\ldots,n$.
3: Display the SP plot with points $x_{j:n}=(2/\pi)\arcsin(\sqrt{w_{j:n}})$ versus $s_{j:n}=(2/\pi)\arcsin(\sqrt{\widehat{u}_{j:n}})$.
4: Construct acceptance bands according to Table 2 specifying a ϱ significance level.
5: Decide if H_0 must be rejected for the specified significance level.
6: Corroborate decision in Step 5 with the p-values after evaluating $\mathrm{KS}^{\star}_{r,n}$ and $\mathrm{MI}^{\star}_{r,n}$ statistics at $\widehat{u}_{j:n}$.

3 Applications

In this section, we consider several real-world data sets and the Birnbaum-Saunders (standard, truncated and generalized [14]), gamma, truncated normal (TN) and Weibull distributions under H_0 to decide whether these data can reasonably come from the hypothesized distribution. The results are also displayed by means of the probability plots with the acceptance bands proposed in Sect. 2.

3.1 Example 1: Uncensored Sea Data

These data correspond to the sea surface temperature (in °K), which are generated by a radiometer of high resolution. We call these data as "sea". The sample size is $n=88$ and the truncation point is $\kappa=278.187$°K; see details in [13]. An exploratory data analysis (EDA) for sea data is provided in Table 3, including the coefficients of variation (CV), skewness (CS) and kurtosis (CK), as well as the standard deviation (SD), minimum (Min) and maximum (Max) values. Our EDA is also based on Fig. 1, which displays their histogram and boxplot. This EDA indicates that the truncated Birnbaum-Saunders (TBS) distribution can

Table 3. Descriptive statistics for sea data.

Median	Mean	SD	CV	CS	CK	Range	Min	Max	n
279.5	279.6	0.787	0.003	0.008	3	3.8	278.2	282	88

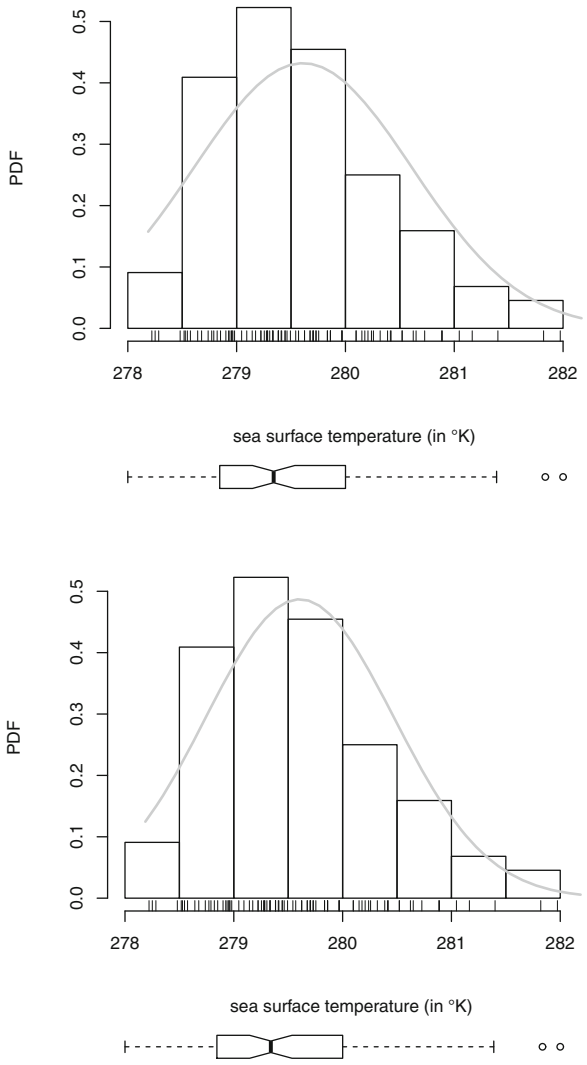


Fig. 1. Histogram, boxplot and estimated PDF from TN (top) and TBS (bottom) distributions for sea data.

be suitable for describing these data, as a competitor of the TN distribution. Some atypical data are detected in the boxplot, but their study is not considered here because it is beyond the objective of this work. We consider the TBS and TN models under H_0 with parameters estimated. The `R` packages named `tbs` and `truncnorm` are used to estimate the corresponding parameters with the ML method. The associated estimates and p-values from the GOF tests are displayed in Table 4. According to these p-values, both distributions perform a good fitting to the data. We show the PP and SP plots in Fig. 2 for the TBS

Table 4. Estimated parameters and values of the statistics for the indicated distribution under H_0 with sea data.

Model	Parameter	Estimate	Estimated statistic	p-value
TBS	α	0.0057	$ks^\star = 0.0653$	$[0.4, 0.5]$
	β	278.8488	$mi^\star = 0.0414$	$[0.7, 0.8]$
TN	μ	279.6093	$ks^\star = 0.0600$	$[0.5, 0.6]$
	σ	1.0011	$mi^\star = 0.0369$	$[0.8, 0.9]$

distribution. From Fig. 2, note that the points are well aligned, as expected, due to the high p-values obtained for the TBS distribution, and all the points fall inside the 95% acceptance bands, which confirms the good fitting of the TBS distribution to sea data.

3.2 Example 2: Uncensored Forestry Data

These data correspond to the diameter at breast height (DBH, in cm) of trees of loblolly pine from a plantation in the Western Gulf Coast. We call these data as "forestry". The sample size is $n = 75$ and the left-truncation point is $\kappa = 6$ cm; see details in [19]. Table 5 provides an EDA of forestry data, which indicates once again that the TBS distribution can be a good model for these data.

From the histograms displayed in Fig. 3, note that the fit of the TBS distribution seems to be better than for the TN distribution. The corresponding estimates and p-values from the GOF tests are displayed in Table 6. According to these p-values, both distributions perform a reasonable fit to the data, but clearly the TBS distribution has a better performance. Figure 4 shows the PP and SP plots for the TBS distribution. From this figure, note that once again the points are well aligned, as expected, due to the p-values obtained for the TBS distribution, and all the points fall inside the 95% acceptance bands, confirming the good fitting of the TBS distribution to forestry data.

3.3 Example 3: Uncensored Survival Data

These data correspond to the survival times (in days) of pigs injected with a dose of tubercle bacilli, under a regimen corresponding to 4.0×10^6 bacillary units per 0.5 ml ($\log(4.0 \times 10^6) = 6.6$). We call these data as "survival". The sample size is $n = 72$ guinea pigs infected with tubercle bacilli in regimen 6.6; see details in [1,2]. Table 7 and Fig. 5 provide an EDA of survival data. From this EDA, we detect a distribution skewed to the right with the presence of some outliers. We propose the BS and BS-t (BS based on the Student-t) distributions for analyzing survival data. ML estimates of the BS and BS-t parameters, which are obtained using an `R` package named `gbs`, and the p-values from the GOF tests, are displayed in Table 8. Clearly the BS distribution does not fit properly these data, whereas the BS-t distribution performs a better fit. Figures 6 and 7

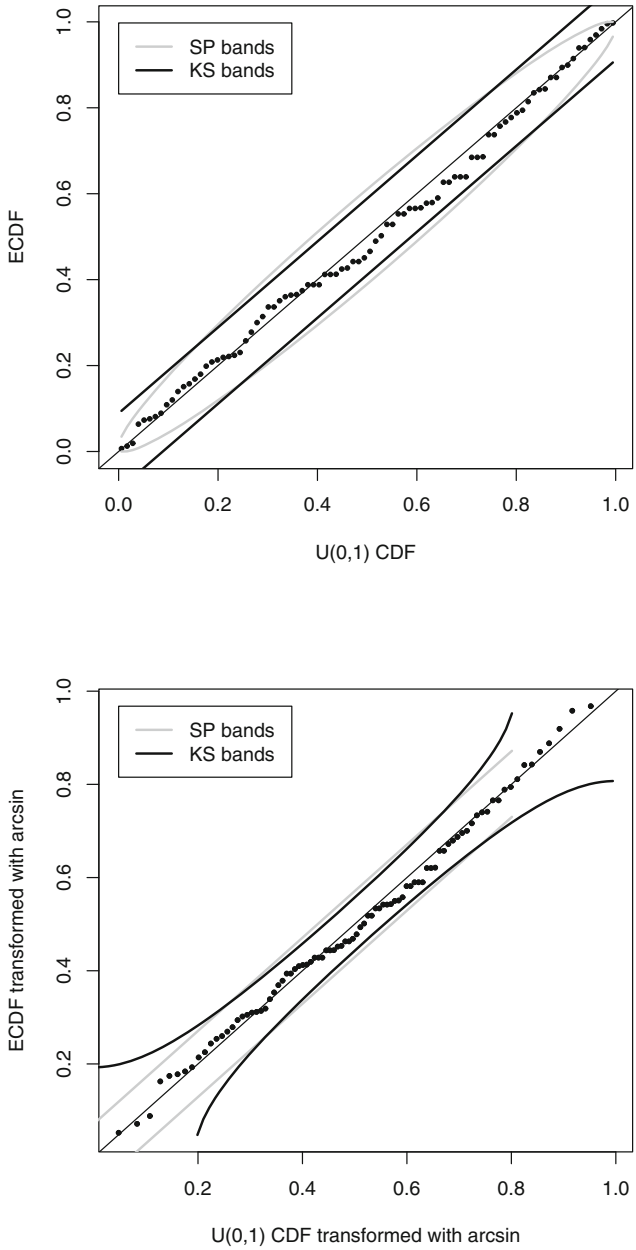


Fig. 2. PP (top) and SP (bottom) plots with 95% acceptance bands for the TBS distribution with sea data.

show the PP and SP plots for the BS and the BS-t distributions, respectively. From Fig. 6, note that the points are not well aligned, specially in the center, and that one observation (case #15) is outside the KS band. In Fig. 7, the points are better aligned, as expected, due to the p-values obtained for the BS-t distribution, and all the points fall inside the 95% acceptance bands, indicating the good fitting of the BS-t distribution to survival data.

Table 5. Descriptive statistics for forestry data.

Median	Mean	SD	CV	CS	CK	Range	Min	Max	n
8.20	8.19	1.013	0.124	0.053	2.253	4.1	6.2	10.3	75

3.4 Example 4: Uncensored Survival Data with Outliers

Next, we conduct a simple empirical robustness study. First, we add a large value (outlier) to the data and call them as "survival1", so that we have now a sample of size $n = 73$. This new observation is greater than all the observed values ($t_{73:73} = 580$). Second, we add another large value ($t_{74:74} = 750$) to the data and call them as "survival2", so that we have now a sample of size $n = 74$. Then, we input one more outlier, corresponding to the value $t_{75:75} = 1000$, and call these data as "survival3". Figure 8 displays usual and adjusted boxplots, as well as stripcharts, for survival, survival1, survival2 and survival3 data sets. The adjusted boxplot is often used for skewed data because it includes a robust measure of skewness [22]. The stripchart is a scatterplot in one dimension, where all the observations are plotted. From these graphs we can visualize the effect of the outliers added to the data. In the adjusted boxplot for survival1 (see Fig. 8-center), it is possible to note that the first added value is an outlier, but when a second atypical value is added for survival2, only this second value is detected as outlier, but the first one is no longer an outlier for this data set. With the third value being part of the sample, which is much greater than the others, the presence of two outliers is detected. The ML estimates and p-values of the GOF tests for survival1, survival2 and survival3 are provided in Table 9. Note that little changes in the estimated parameters and in the bounds for the p-values of the tests are detected. For survival3 data, the differences are more noticeable. We conclude that the GOF tests are relatively robust to outliers when the BS-t distribution is considered under H_0, specially the MI test, but a more extensive study about this issue should be carried out.

Figure 9 shows the PP and SP plots for the BS-t distribution using the survival data with three outliers added. From this figure, note that the points are not still well aligned, specially in the center, but now the case # 15 is not near the bands and there is one observation (case # 42) outside the KS band. If we compare Figs. 7 and 9, there are no apparently visual differences with minor distinct alignments in the points, specially because the rejection is due to points in the center and not in the tails of the sample where the outliers are located.

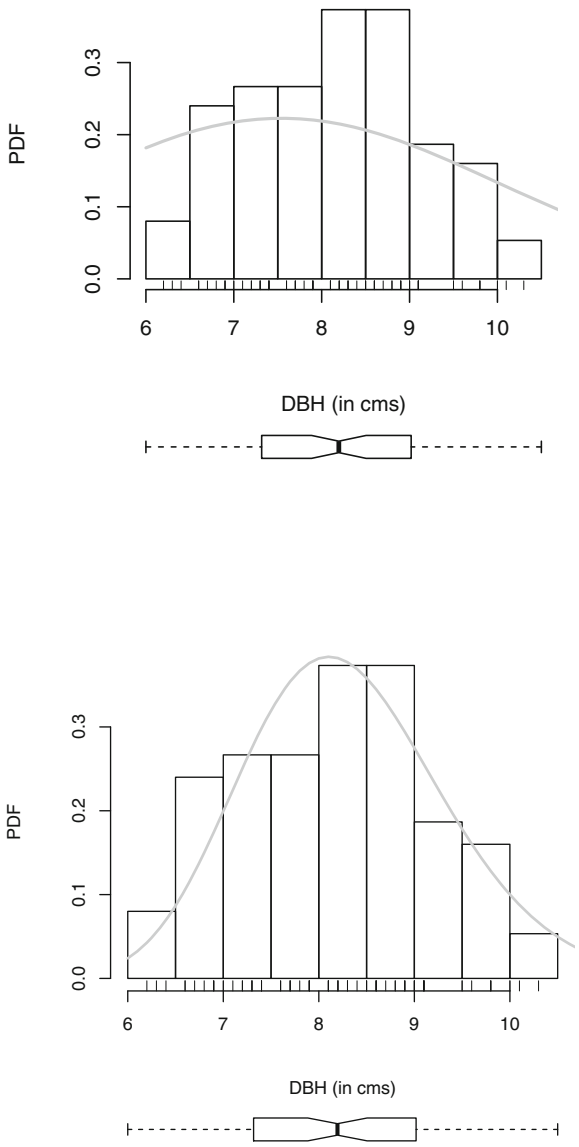


Fig. 3. Histogram, boxplot and estimated PDF from TN (top) and TBS (bottom) models for forestry data.

Table 6. Estimated parameters and statistic values for the indicated distribution under H_0 with forestry data.

Model	Parameter	Estimate	Estimated statistic	p-value
TBS	α	0.12804	$ks^\star = 0.07031$	$[0.4, 0.5]$
	β	8.23963	$mi^\star = 0.05923$	$[0.25, 0.4]$
TN	μ	7.54905	$ks^\star = 0.08201$	$[0.2, 0.25]$
	σ	2.42369	$mi^\star = 0.06697$	$[0.1, 0.2]$

Table 7. Descriptive statistics for survival data.

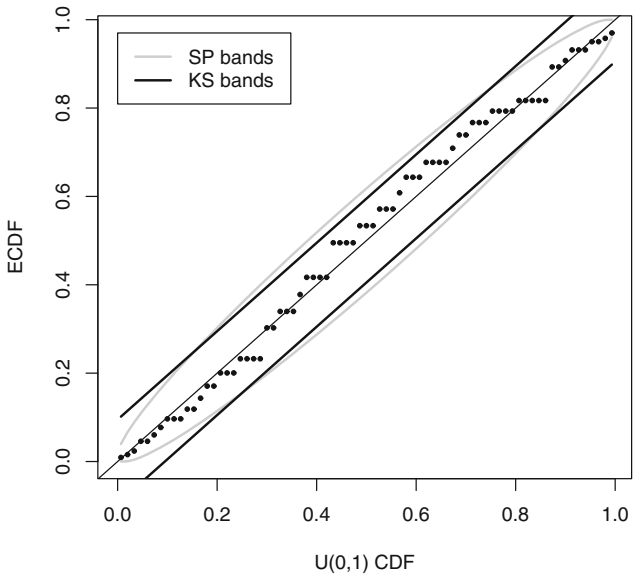
Median	Mean	SD	CV	CS	CK	Range	Min	Max	n
70.00	99.82	81.12	0.81	1.76	5.46	364	12	376	72

Table 8. Estimated parameters and statistic values for the indicated distribution under H_0 with survival data.

Model	Parameter	Estimate	Estimated statistic	p-value
BS	α	0.7600	$ks^\star = 0.08848$	$[0.01, 0.05]$
	β	77.5348	$mi^\star = 0.07318$	$[0.05, 0.1]$
BS-t	α	0.6085	$ks^\star = 0.08201$	$[0.1, 0.2]$
	β	75.5880	$mi^\star = 0.05908$	$[0.25, 0.4]$
	ν	5.0000	–	–

3.5 Example 5: Comparison of Two Treatments

Certain clinical trials are aimed at shortening the time-to-discharge. In a double-blind placebo controlled drug study, times (in hours) of 23 patients on drug and of 25 patients on placebo were reported. No censoring occurred on this trial. The hypothesis was that a 4-day ambulatory femoral nerve block decreases the length of stay after a total knee arthroplasty compared to the usual treatment. In Fig. 10, the placebo data show a distribution skewed to the right. We consider 12 possible distributions including the Birnbaum-Saunders, gamma and Weibull models. Notice that the drug data seem to be generated by two different populations. Only two models fit well the data: the mixture normal and mixture gamma distributions, being better the last one. We reject the Weibull model with a p-value < 0.001. The ML estimates and corresponding observed statistics are omitted. By means of the selected distributions, we estimate that 43.4% of the patients that received the conventional treatment and 4.4% of the patients that received the new drug stay in the hospital more than 3 days (the usual estimated time): the drug works very well reducing the length of stay. Figure 10 shows the histogram and estimated PDF of the indicated distributions for placebo and drug data in different scales. We omit here the PP and SP plots for each group.



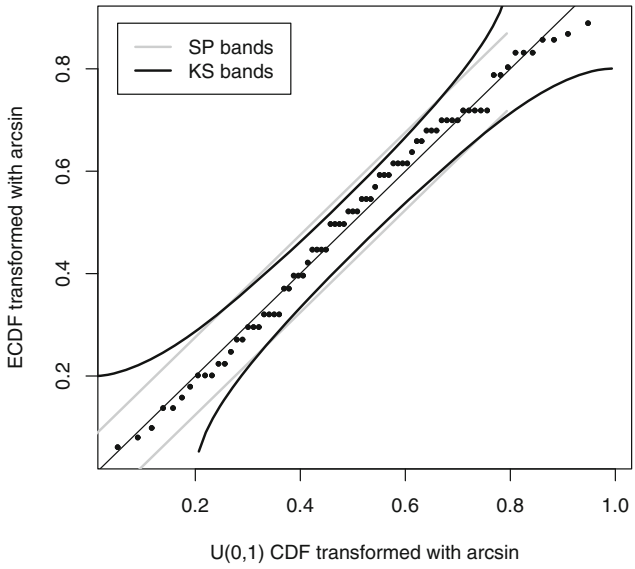


Fig. 4. PP (top) and SP (bottom) plots with 95% acceptance bands for the TBS distribution with forestry data.

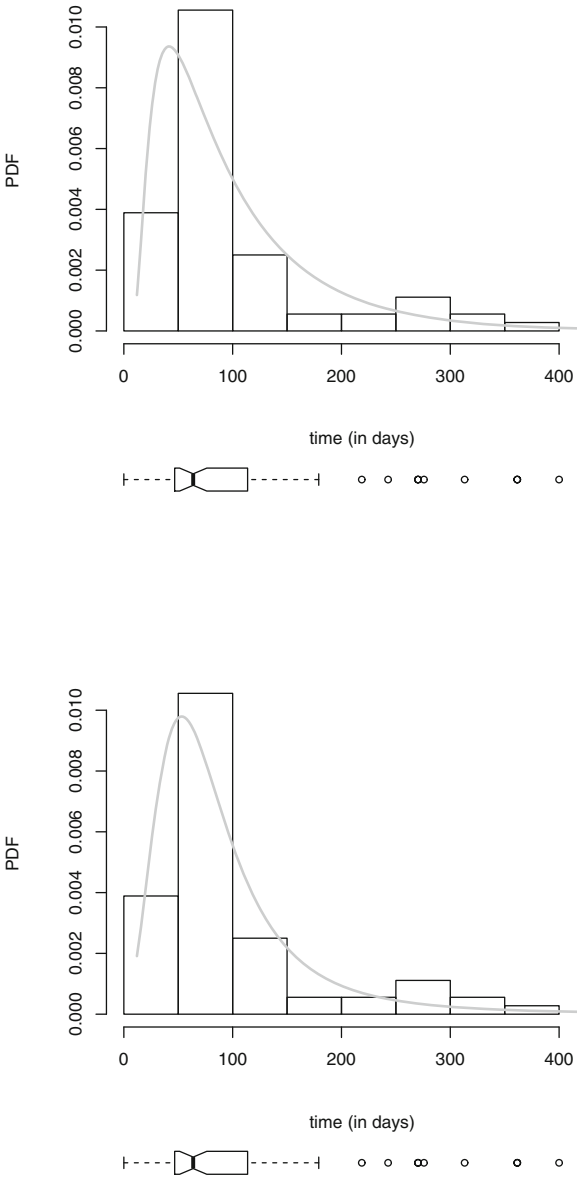


Fig. 5. Histogram, boxplot and estimated PDF from BS (top) and BS-t (bottom) model for survival data.

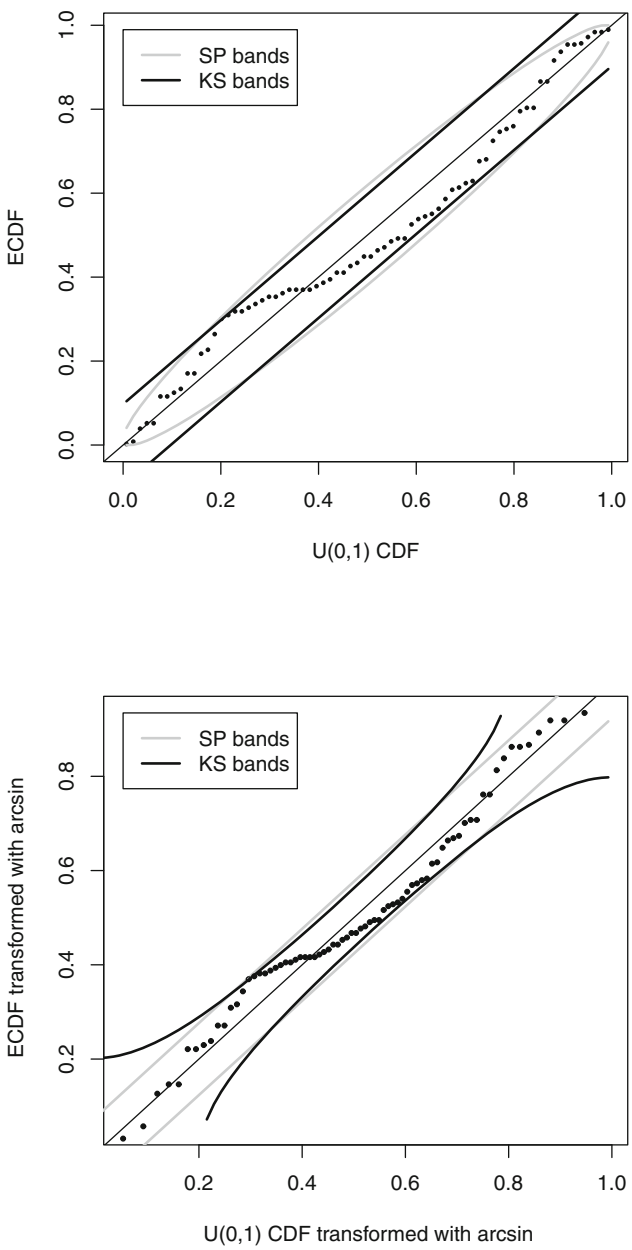


Fig. 6. PP (top) and SP (bottom) plots with 95% acceptance bands for the BS distribution using survival data.

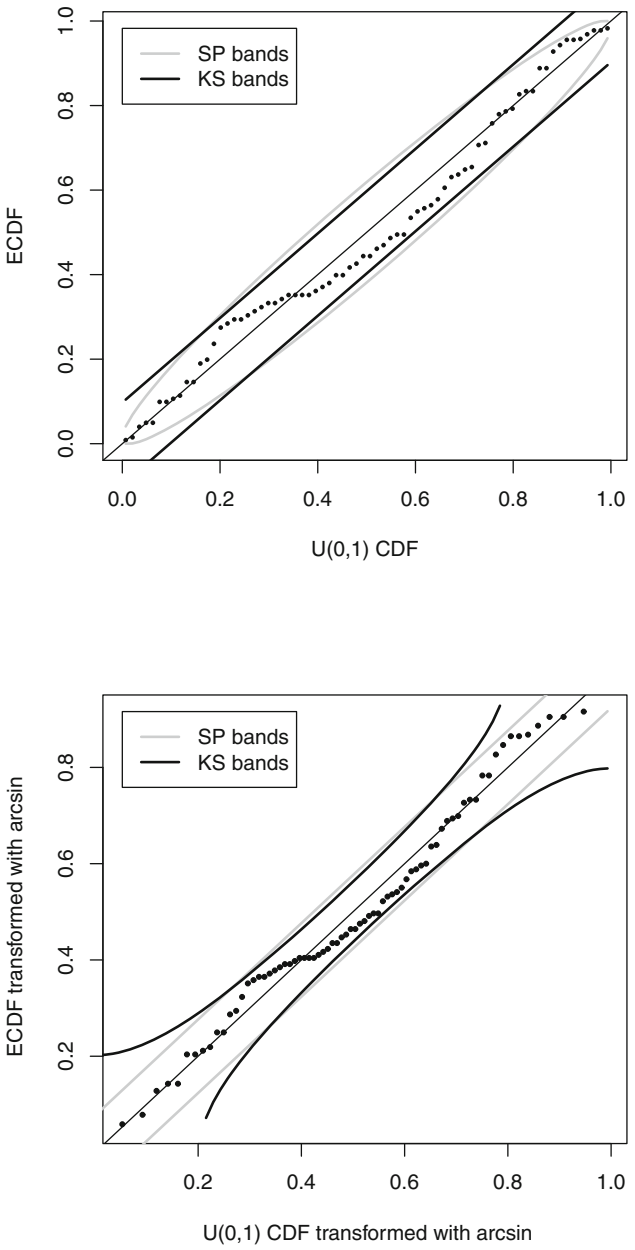


Fig. 7. PP (top) and SP (bottom) plots with 95% acceptance bands for the BS-t distribution using survival data.

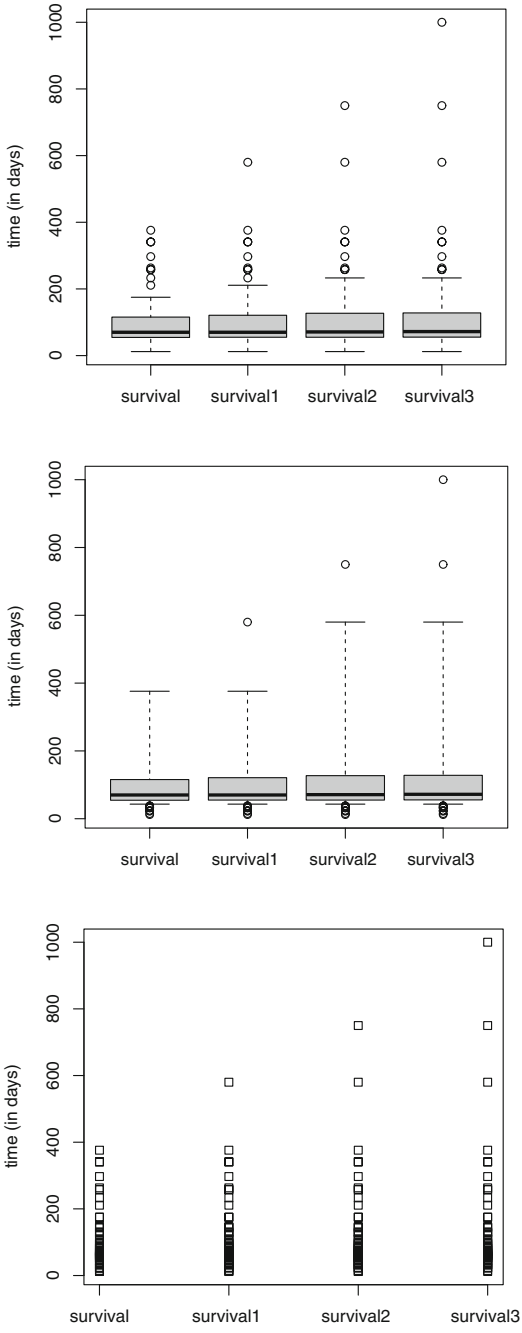


Fig. 8. Boxplots (top), adjusted boxplots (center) and stripchart (bottom) for survival data and three outliers added.

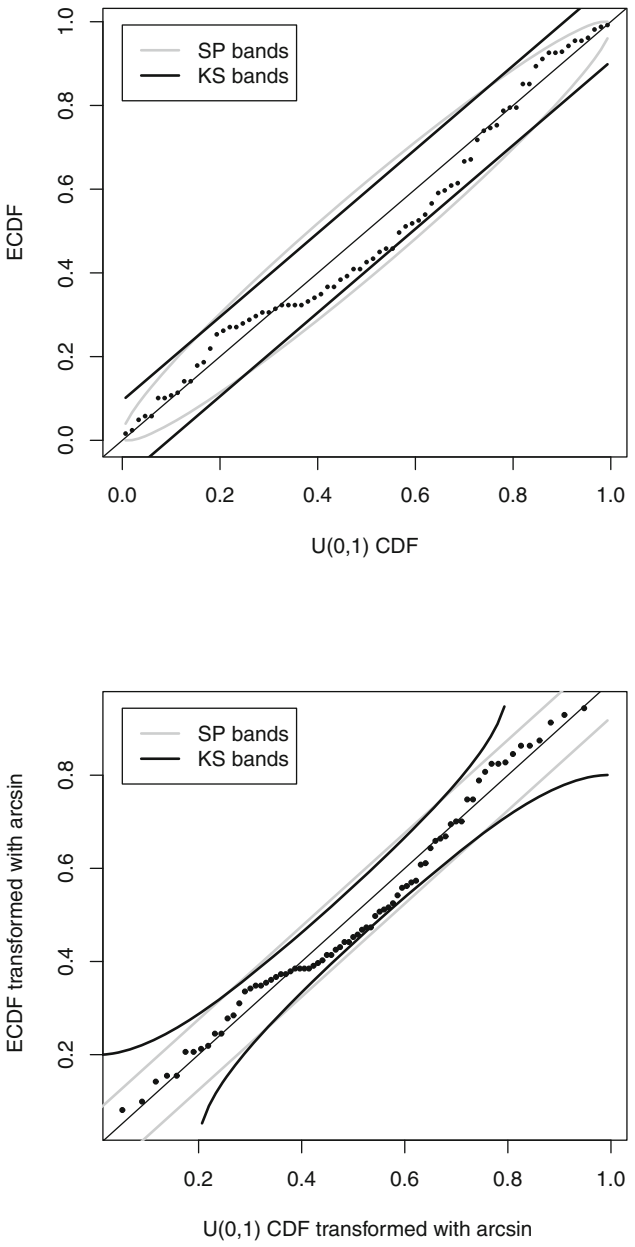


Fig. 9. PP (top) and SP (bottom) plots with bands for the BS-t model using survival data and three outliers added.

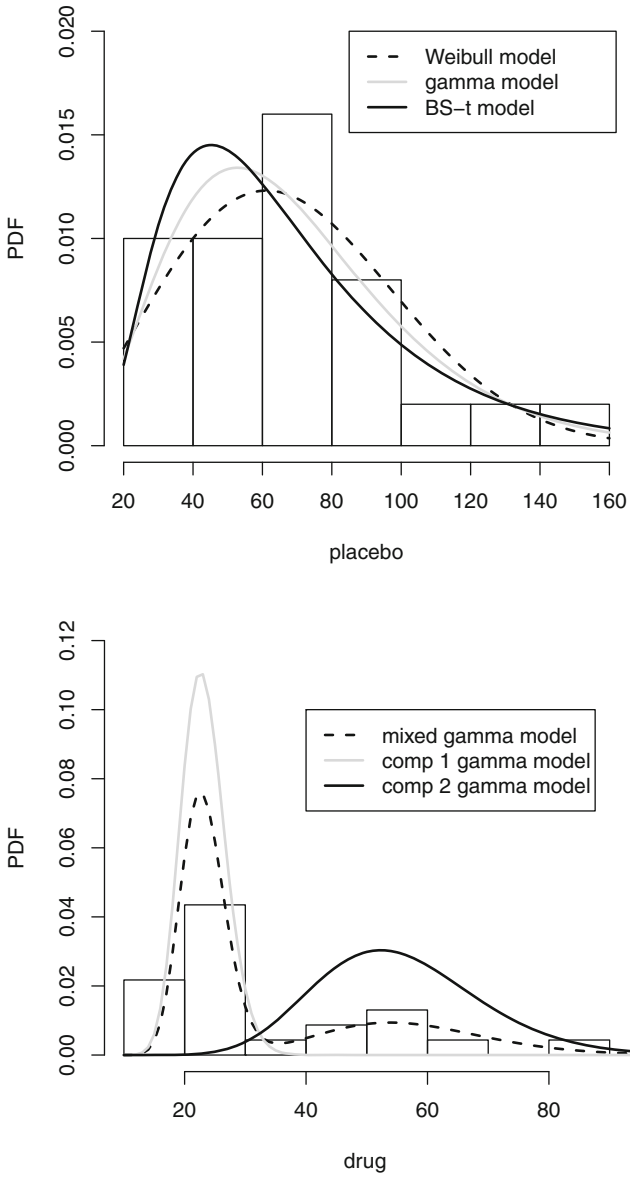


Fig. 10. Histogram and estimated PDF of the indicated distribution for data in the placebo (top) and drug (bottom) groups.

Table 9. Estimated parameters and statistic values for the indicated distribution under H_0 with survival data.

Model	Parameter	Estimate	Estimated statistic	p-value	n	Data set
BS-t	α	0.6036	ks* = 0.09313	[0.10, 0.2]	73	survival1
	β	76.2200	mi* = 0.05916	[0.25, 0.4]		
	ν	4.0000	–	–		
BS-t	α	0.5806	ks* = 0.095664	[0.05, 0.1]	74	survival2
	β	76.0500	mi* = 0.058160	[0.25, 0.4]		
	ν	3.0000	–	–		
BS-t	α	0.6049	ks* = 0.102022	[0.01, 0.05]	75	survival3
	β	77.2200	mi* = 0.060802	[0.20, 0.25]		
	ν	3.0000	–	–		

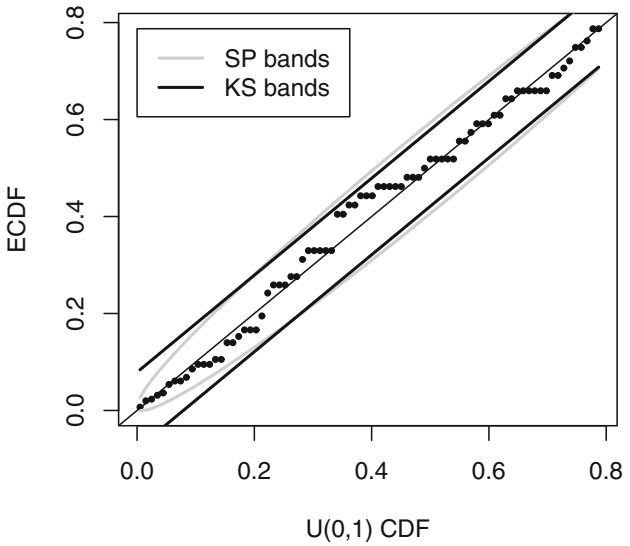
3.6 Example 6: Censored Fatigue Data

These data correspond to fatigue life (in cycles $\times 10^{-3}$) of coupons of aluminium. We call these data as "fatigue"; see details in [1,14,20]. We consider a censored fatigue data sample, such as in [5], so that we have $r = 80$ failures and $n - r = 21$ data censored. From the histogram displayed in Fig. 12, note that the distribution of fatigue data is clearly skewed to the right, with the BS distribution showing a good fitting to these data, which must be corroborated.

The parameters are estimated at $\widehat{\alpha} = 0.1751$ and $\widehat{\beta} = 132.2525$, by the ML method, considering the presence of type-II right censoring. The KS and MI statistics are computed such as in Examples 1 and 2, and their corresponding p-values based on the BS distribution are in $[0.2, 0.25]$ and $[0.4, 0.5]$. According to these p-values, we cannot reject the null hypothesis that the censored sample comes from a BS distribution, which can be confirmed from the PP and SP plots shown in Fig. 11. This result is consistent with those obtained by other authors.

3.7 Example 7: Times to Failure in an Accelerated Life Test

We analyze the times to failure in a temperature-accelerated life test for a device; see details in [15]. The sample is singly censored to the right with 33 failures and four censored observations at 5000 h. We evaluate the adequacy of five life distributions, estimate their parameters and then use Algorithm 2. The BS distribution provides the best fit to these data. The ML estimates of the corresponding parameters and the obtained observed statistics are omitted here, but for the three of them we obtain $0.9 <$ p-value < 0.95, indicating an excellent agreement between the model and the data. Figure 13 shows the PP and SP plots of the times to failure for device according to the selected model with 95% acceptance bands. As expected, all the observations fall inside the bands with a very good alignment.



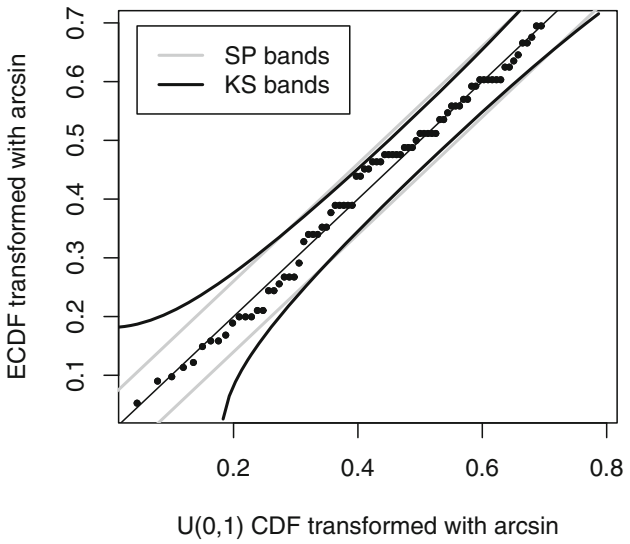


Fig. 11. PP (top) and SP (bottom) plots with 95% acceptance bands for the BS distribution using fatigue data.

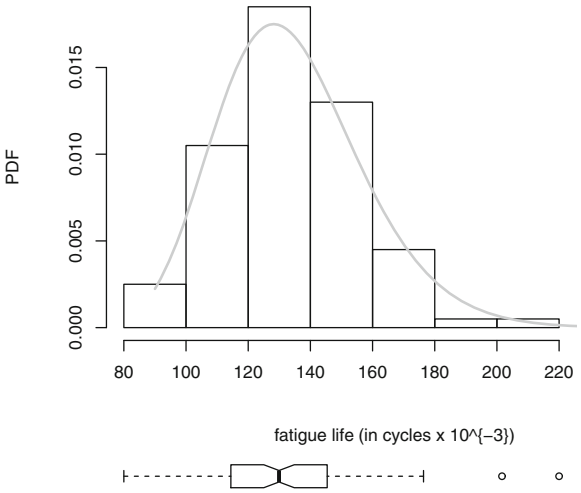


Fig. 12. Histogram, boxplot and estimated BS PDF for fatigue data.

4 Conclusions and Future Research

We have presented goodness-of-fit tests related to the Kolmogorov-Smirnov and Michael statistics and connected them to graphical methods with uncensored and censored data. Although the Anderson-Darling test is often more powerful than the Kolmogorov-Smirnov test, it cannot be related to graphical tools by means of probability plots, as the Kolmogorov-Smirnov test does. The stabilized probability plot is related to the Michael test, which is, in some cases, more powerful than the Anderson-Darling and Kolmogorov-Smirnov tests. We have considered the Kolmogorov-Smirnov and Michael tests for detecting whether any distribution is suitable or not to model censored or uncensored data using graphical tools. We have conducted numerical studies for showing potential applications of these tests and their corresponding graphical versions.

Data science is being an important topic where statistics plays a relevant role. A challenging issue in data science is the handling of large amount of data, known as massive data or big data [3]. Goodness-of-fit methods have been often used for data sets with a low frequency sampling (small amount of observations). Today there are instruments that generate big data. Due to the rapid advancement of computers and information technologies, automatic data acquisition is becoming increasingly common, moving data collection away from historically low-dimensional approaches. With current technologies, such as digital equipment, analytical sensors and live health monitoring, data generated from these technologies may be used to detect whether a distribution is suitable to describe these data. The term big data is often used to describe large, diverse and complex data sets that are generated from different types of instruments, sensors or computer-based transactions. Big data are information assets, characterized by

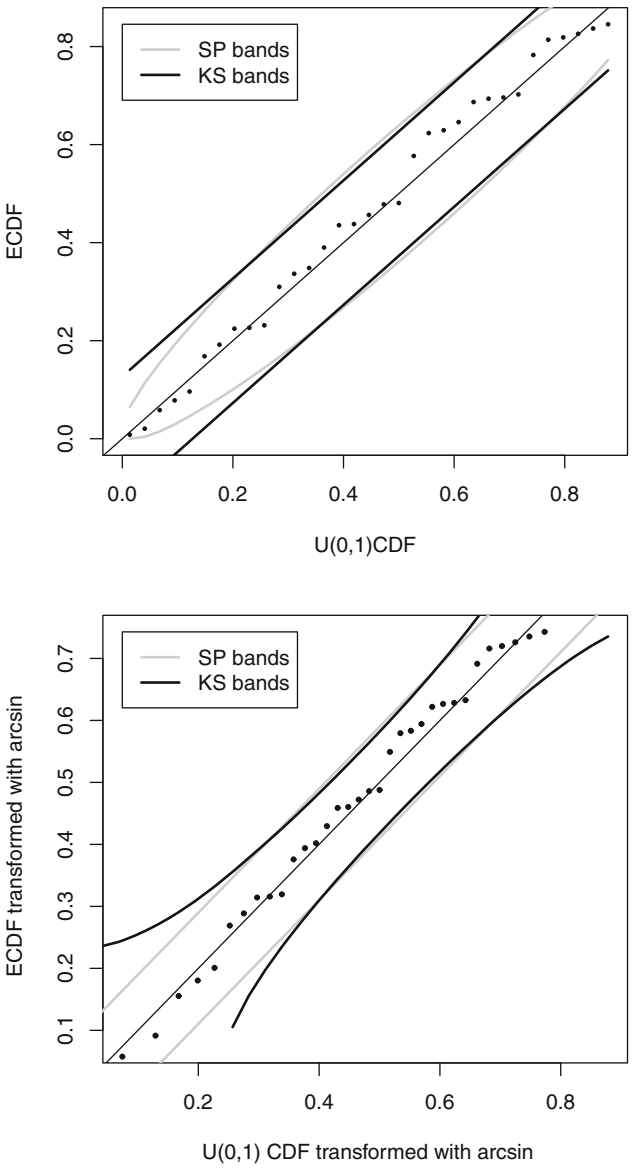


Fig. 13. PP (top) and SP (bottom) plots with 95% acceptance bands for times to failure data of a device using the BS model.

its large volume, velocity and variety (3Vs), requiring innovative and efficient solutions to improve the knowledge process when making decisions in organizations. The objective of big data is to provide high technology (hardware and software) to store, process and analyze large amounts of data (mega, giga, tera, peta, exa, zetta and yottabyte) and to create value in an organization. Facing on big data, many statistical concepts need to be updated, particularly, goodness-of-fit methods. For example, it is known that the power of a goodness-of-fit test depends on the sample size, because a large number of observations provides a larger test power. Thus, considering the large number of observations that the big data brings to the current era, it leads to a great advantage when deciding whether a distribution fits a large data set, especially in the procedures used in this work. This advantage affects the process to make decisions, that is, to choose which is the distribution that best fits the data. Note that a large amount of data increases the probability of detecting unusual anomalies that cannot be modeled by some distribution, making it possible to reduce the probability of obtaining false positives in the hypothesis test. This allows us to filter a large number of candidate distributions, until finding the adequate model. For this reason, considering big data in the methodologies of the current times results in great opportunities for discovering knowledge, particularly when performing goodness-of-fit tests.

Acknowledgements. The authors thank the editors and reviewers for their comments on this manuscript. This research work was partially supported by FONDECYT 1160868 grant from the Chilean government.

References

1. Athayde, E., Azevedo, C., Barros, M., Leiva, V.: Failure rate of Birnbaum-Saunders distributions: shape, change-point, estimation and robustness. Braz. J. Probab. Stat. **33**, 301–328 (2019)
2. Azevedo, C., Leiva, V., Athayde, E., Balakrishnan, N.: Shape and change point analyses of the Birnbaum-Saunders-t hazard rate and associated estimation. Comput. Stat. Data Anal. **56**, 3887–3897 (2012)
3. Aykroyd, R.G., Leiva, V., Ruggeri, F.: Recent developments of control charts, identification of big data sources and future trends of current research. Technological Forecasting and Social Change (pages in press) (2019)
4. Balakrishnan, N., Ng, H., Kannan, N.: Goodness-of-fit tests based on spacings for progressively type-II censored data from a general location-scale distribution. IEEE Trans. Reliab. **53**, 349–356 (2004)
5. Barros, M., Leiva, V., Ospina, R., Tsuyuguchi, A.: Goodness-of-fit tests for the Birnbaum-Saunders distribution with censored reliability data. IEEE Trans. Reliab. **63**, 543–554 (2014)
6. Castillo, J., Puig, P.: Testing departures from gamma, Rayleigh, and truncated normal distributions. Ann. Inst. Stat. Math. **49**, 255–269 (1997)
7. Castro-Kuriss, C.: On a goodness-of-fit test for censored data from a location-scale distribution with application. Chil. J. Stat. **37**, 115–136 (2011)

8. Castro-Kuriss, C., Kelmansky, D., Leiva, V., Martinez, E.: A new goodness-of-fit test for censored data with an application in monitoring processes. Commun. Stat. Simul. Comput. **38**, 1161–1177 (2009)
9. Castro-Kuriss, C., Kelmansky, D., Leiva, V., Martinez, E.: On a goodness-of-fit test for normality with unknown parameters and type-II censored data. J. Appl. Stat. **37**, 1193–1211 (2010)
10. Castro-Kuriss, C., Leiva, V., Athayde, E.: Graphical tools to assess goodness-of-fit in non-location-scale distributions. Colomb. J. Stat. **37**, 341–365 (2014)
11. Chen, G., Balakrishnan, N.: A general purpose approximate goodness-of-fit test. J. Qual. Technol. **27**, 154–161 (1995)
12. D'Agostino, C., Stephens, M.: Goodness of Fit Techniques. Marcel Dekker and Routledge, New York (2017)
13. DePriest, D.: Using the singly truncated normal distribution to analyse satellite data. Commun. Stat. Theory Methods **12**, 263–272 (1983)
14. Leiva, V.: The Birnbaum-Saunders distribution. Academic Press, New York (2016)
15. Meeker, W.Q., Escobar, L.A.: Statistical Methods for Reliability Data. Wiley, New York (1998)
16. Feuerverger, A.: On goodness of fit for operational risk. Int. Stat. Rev. **84**, 434–455 (2016)
17. Goldmann, C., Klar, B., Meintanis, S.G.: Data transformations and goodness-of-fit tests for type-II right censored samples. Metrika **78**, 59–83 (2015)
18. Lawless, J.: Statistical Models and Methods for Lifetime Data. Wiley, New York (2003)
19. Leiva, V., Ponce, M., Marchant, C., Bustos, O.: Fatigue statistical distributions useful for modeling diameter and mortality of trees. Colomb. J. Stat. **35**, 349–367 (2012)
20. Leiva, V., Rojas, E., Galea, M., Sanhueza, A.: Diagnostics in Birnbaum-Saunders accelerated life models with an application to fatigue data. Appl. Stoch. Model. Bus. Ind. **30**, 115–131 (2014)
21. Lin, C., Huang, Y., Balakrishnan, N.: A new method for goodness-of-fit testing based on type-II right censored samples. IEEE Trans. Reliab. **57**, 633–642 (2008)
22. Meatless, M., Rousseeuw, P.J., Croux, C., et al.: robustbase: Basic Robust Statistics. R package version 0.93-3 (2018). http://robustbase.r-forge.r-project.org
23. Marden, J.: Positions and QQ-plots. Stat. Sci. **19**, 606–614 (2004)
24. Michael, J.: The stabilized probability plot. Biometrika **70**, 11–17 (1983)
25. Pakyari, R., Balakrishnan, N.: A general purpose approximate goodness-of-fit test for progressively type-II censored data. IEEE Trans. Reliab. **61**, 238–244 (2012)
26. Rad, A., Yousefzadeh, F., Balakrishnan, N.: Goodness-of-fit test based on Kullback-Leibler information for progressively type-II censored data. IEEE Trans. Reliab. **60**, 570–579 (2011)