Journal of Statistical Software

August 2013, Volume 54, Issue 3. *http://www.jstatsoft.org/*

dbEmpLikeGOF: An R Package for Nonparametric Likelihood Ratio Tests for Goodness-of-Fit and Two-Sample Comparisons Based on Sample Entropy

Jeffrey C. Miecznikowski
University at Buffalo

Albert Vexler
University at Buffalo

Lori Shepherd
Roswell Park Cancer Institute

Abstract

We introduce and examine **dbEmpLikeGOF**, an R package for performing goodness-of-fit tests based on sample entropy. This package also performs the two sample distribution comparison test. For a given vector of data observations, the provided function `dbEmpLikeGOF` tests the data for the proposed null distributions, or tests for distribution equality between two vectors of observations. The proposed methods represent a distribution-free density-based empirical likelihood technique applied to nonparametric testing. The proposed procedure performs exact and very efficient p values for each test statistic obtained from a Monte Carlo (MC) resampling scheme. Note by using an MC scheme, we are assured exact level α tests that approximate nonparametrically most powerful Neyman-Pearson decision rules. Although these entropy based tests are known in the theoretical literature to be very efficient, they have not been well addressed in statistical software. This article briefly presents the proposed tests and introduces the package, with applications to real data. We apply the methods to produce a novel analysis of a recently published dataset related to coronary heart disease.

Keywords: empirical likelihood, likelihood ratio, goodness-of-fit, sample entropy, nonparametric tests, normality, two-sample comparisons, uniformity.

1. Introduction

1.1. Empirical likelihood

Empirical likelihood (EL) allows researchers the benefit of employing powerful likelihood methods (maximizing likelihoods) without having to choose a parametric family for the data.

A thorough overview of empirical likelihood methods can be found in Owen (2001). The research in this area continues to grow while empirical likelihood methods are being extended to many statistical problems as in, for example, Vexler, Yu, Tian, and Liu (2010) or Yu, Vexler, and Tian (2010).

In short, an outline of the EL approach can be presented as follows. Given independently identically distributed observations $X_1, \ldots, X_n$, the EL function has the form of $L_p = \Pi_{i=1}^n p_i$ where the components $p_i, i = 1, \ldots, n$ maximize the likelihood L_p (maximum likelihood estimation) provided that empirical constraints, based on $X_1, \ldots, X_n$ are in effect ($\sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i X_i = 0$, under the hypothesis $EX_1 = 0$). Computation of the EL's components $p_i, i = 1, \ldots, n$ used to be an exercise in Lagrange multipliers. This nonparametric approach is a product of the consideration of the 'distribution-functions'-based likelihood $\Pi_{i=1}^n F(X_i) - F(X_i-)$ over all distribution functions F where $F(X_i-)$ denotes the left hand limit of F at X_i.

The following extensions from these methods involve a density-based likelihood methodology for goodness-of-fit testing. The proposed extensions have been motivated by developing test statistics that approximate nonparametrically most powerful Neyman-Pearson test statistics based on likelihood ratios. A density-based EL methodology can be introduced utilizing the EL concept as in Vexler and Gurevich (2010a), Vexler and Gurevich (2010b), Gurevich and Vexler (2011). Following the EL methodology, the likelihood function $L_f = \Pi_{i=1}^n f(X_i)$ where $f(\cdot)$ is a density function of X_i can be approximated by $\Pi_{i=1}^n f_i$, where values of f_i should maximize $\Pi_{i=1}^n f_i$ provided that an empirical constraint which corresponds to $\int f(u)du = 1$ under an underlying hypothesis is in effect. Outputs of the density based EL approach have a structure that utilize sample entropy (for example, Vexler and Gurevich 2010a). To date, density based EL tests have not been presented in R packages (R Core Team 2013) but are known to be very efficient in practice. Moreover, despite the fact that many theoretical articles have considered very powerful entropy-based tests, to our knowledge there does not exist software procedures to execute procedures based on sample entropy in practice.

1.2. Goodness-of-fit tests

Goodness-of-fit tests commonly arise when researchers are interested in checking whether the data come from an assumed parametric model. In certain situations, this question manifests to test whether two datasets come from the same parametric model. Commonly used goodness-of-fit tests include the Shapiro-Wilks (SW) test, Kolmogorov-Smirnov (KS) test, the Lilliefors (L) test, Wilcoxon rank sum test (WRS), the Cramér-von-Mises test, and the Anderson-Darling test (Darling 1957; Lilliefors 1967; Hollander, Wolfe, and Wolfe 1973; Royston 1991).

Recently several new goodness-of-fit tests have been developed using density based empirical likelihood methods. These powerful new tests offer exact level α tests with critical values that can be easily obtained via Monte Carlo approaches.

1.3. EL ratio test for normality

The derivation of the EL ratio test for normality can be found in Vexler and Gurevich (2010b). To outline this method, we suppose that the data consist of n independent and identically distributed observations $X_1, \ldots, X_n$. Consider the problem of testing the composite hypothesis that a sample $X_1, \ldots, X_n$ is from a normal population. Notationally, the null hypothesis is

$$H_0 : X_1, \ldots, X_n \sim N(\mu, \sigma^2), \tag{1}$$

where $N(\mu, \sigma^2)$ denotes the normal distribution with unknown mean μ and unknown standard deviation σ. Generally speaking when the density functions f_{H_1} and f_{H_0} corresponding to the null and alternative hypotheses, are completely known, the most powerful test statistic is the likelihood ratio:

$$\frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{\prod_{i=1}^{n} f_{H_0}(X_i)} = \frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{(2\pi\sigma^2)^{-n/2}\exp\left(-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2\right)}, \tag{2}$$

where, under the null hypothesis, $X_1, \ldots, X_n$ are normal with mean μ and variance σ^2. In the case of the unknown μ and σ^2, the maximum likelihood estimation applied to (2) changes the ratio to,

$$\frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{(2\pi e s^2)^{-n/2}}, \tag{3}$$

where s represents the sample standard deviation.

Applying the maximum EL method to (3) forms the likelihood ratio test statistic

$$T_{mn} = (2\pi e s^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)}, \tag{4}$$

where m is assumed to be less than $n/2$. Using empirical likelihood modifications, the maximum EL method applied to (3), and following Vexler and Gurevich (2010b), to test the null hypothesis at (1) we can use the test statistic,

$$V_n = \min_{1 \le m < n^{1-\delta}} (2\pi e s^2)^{n/2} \prod_{i=1}^{n} \frac{2m}{n\left(X_{(i+m)} - X_{(i-m)}\right)} \tag{5}$$

where $0 < \delta < 1$, s denotes the sample standard deviation, and $X_{(1)}, \ldots, X_{(n)}$ represent the order statistics corresponding to the sample $X_1, \ldots, X_n$. Note, here, $X_{(j)} = X_{(1)}$ if $j \le 1$ and $X_{(j)} = X_{(n)}$ if $j \ge n$.

We employ the following decision rule, we reject the null hypothesis if and only if

$$\log(V_n) > C, \tag{6}$$

where C is a test threshold and V_n is the test statistic defined in (5).

Since

$$\sup_{\mu,\sigma} P_{H_0}\{\log(V_n) > C\} = P_{X_1,\ldots,X_n \sim N(0,1)}\{\log(V_n) > C\}, \tag{7}$$

the type I error of the test at (6) can be calculated exactly using a Monte Carlo approach. Type I error for the test in (6) refers to the probability of rejecting the null hypothesis in (1) when, in fact, the null hypothesis is true. Figure 1 displays the Monte Carlo roots C_α of the

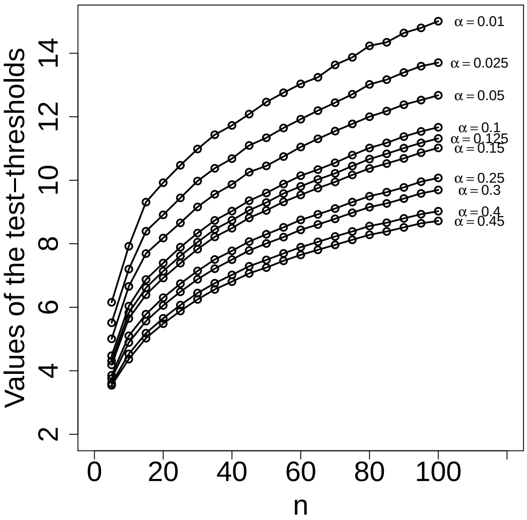

Figure 1: The curves display the value of the thresholds C_α for the test statistic $\log(V_n)$ with $\delta = 0.5$ corresponding to the significance (α) levels of $\alpha = 0.01, 0.025, 0.05, 0.125, 0.15, 0.25, 0.3, 0.4, 0.45$ that are plotted against the sample sizes $n = 5, 10, 15, \ldots, 100$.

equation $P_{X_1,\ldots,X_n \sim N(0,1)} \{\log(V_n) > C_\alpha\} = \alpha$ for different values of α and n. (For each value of α and n, the solutions were derived from 75,000 samples of size n.) The setting of $\delta = 0.5$ is motivated by the work presented in Vexler and Gurevich (2010b). In general, the choice of δ is not critical for these goodness-of-fit tests.

1.4. EL ratio test for uniformity

One can show that tests for uniformity correspond to general goodness-of-fit testing problems when the null hypothesis is based on completely known distribution functions. The full derivation of the EL ratio test for uniformity can be found in Vexler and Gurevich (2010b). We consider the test for the uniform distribution on the interval $[0, 1]$ $(Uni(0, 1))$, specifying the null distribution

$$H_0 : Y_1, \ldots, Y_n \sim Uni(0, 1) \tag{8}$$

versus the alternative that $Y_1, \ldots, Y_n$ are from a nonuniform distribution $F(y)$.

Before considering the hypothesis in (8), consider the problem of testing

$$H_0 : f = f_{H_0} \text{ vs } H_1 : f = f_{H_1}, \tag{9}$$

where, under the alternative hypothesis, f_{H_1} is completely unknown and under the null hypothesis $f_{H_0}(x) = f_{H_0}(x; \boldsymbol{\theta})$ is known up to the vector of parameters $\vec{\theta} = (\theta_1, \ldots, \theta_d)$, where $d \geq 1$ defines a dimension of the vector $\boldsymbol{\theta}$. In accordance with maximizing EL, for the test in

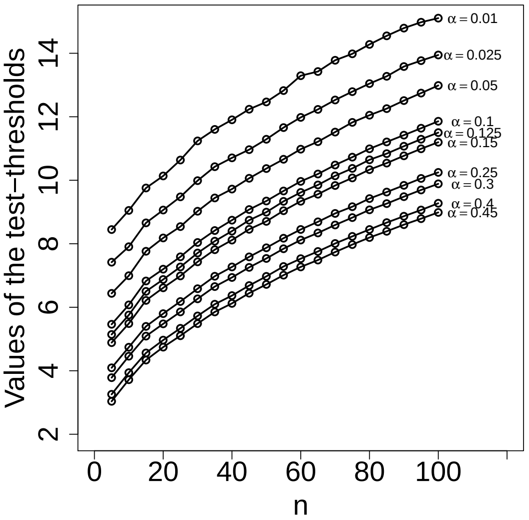

Figure 2: The curves display the value of the thresholds C_α for the test statistic $\log(U_n)$ with $\delta = 0.5$ corresponding to the significance (α) levels of $\alpha = 0.01, 0.025, 0.05, 0.125, 0.15, 0.25, 0.3, 0.4, 0.45$ that are plotted against the sample sizes $n = 5, 10, 15, \ldots, 100$.

(9) we obtain the statistic,

$$G_n = \min_{1 \le m < n^{1-\delta}} \frac{\displaystyle\prod_{i=1}^{n} \frac{2m}{n(X_{(i+m)} - X_{(i-m)})}}{\displaystyle\prod_{i=1}^{n} f_{H_0}(X_i; \hat{\theta})}. \tag{10}$$

Applying the result in (10) to the specific hypothesis in (8) and using the outputs from Vexler and Gurevich (2010b), we suggest the following EL ratio test statistic

$$U_n = \min_{1 \le m < n^{1-\delta}} \prod_{i=1}^{n} \frac{2m}{n\left(Y_{(i+m)} - Y_{(i-m)}\right)}, \tag{11}$$

where $0 < \delta < 1$ and $Y_{(1)}, \ldots, Y_{(n)}$ correspond to the order statistics from the sample $Y_1, \ldots, Y_n$. Note, $Y_{(j)} = Y_{(1)}$ if $j \le 1$ and $Y_{(j)} = Y_{(n)}$ if $j \ge n$. The event

$$\log(U_n) > C \tag{12}$$

implies that H_0 is rejected, where C is a test threshold. The significance level of this test can be calculated according to the following equation,

$$P_{H_0}\left\{\log(U_n) > C_\alpha\right\} = P_{X_1,\ldots,X_n \sim Uni(0,1)}\left\{\log(U_n) > C_\alpha\right\} = \alpha. \tag{13}$$

Figure 2 shows the roots C_α of the equation in (13) for different values of α and n. (For each value of α and n, the solution is derived from 75,000 samples of size n).

Note, the test for uniformity in (12) will cover a generalized version of the goodness-of-fit problem when the distribution in H_0 is completely known. In other words, if we consider the random sample $X_1, \ldots, X_n$ from a population with a density function f and a finite variance we can test the hypotheses:

$$H_0 : F = F_{H_0} \quad \text{vs } H_1 : F = F_{H_1}, \tag{14}$$

where, under the alternative hypothesis, F_{H_1} is completely unknown, whereas under the null hypothesis, $F_{H_0}(x) = F_{H_0}(x; \boldsymbol{\theta})$ is known up to the vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$. Note, $d \geq 1$ defines the dimension for $\boldsymbol{\theta}$. Although a strong assumption, by assuming that densities exist under the alternative, we are able to demonstrate asymptotic consistency of the proposed test statistic. By employing the probability integral transformation (Dodge 2006), if $X_1, \ldots, X_n \sim f_{H_0}$, with f_{H_0} completely known, then $Y_i = F_{H_0}^{-1}(X_i) \sim Uni(0, 1)$. Hence, the uniformity test in (12) can be employed on data $Y_1, \ldots, Y_n$ to test whether $X_1, \ldots, X_n$ conforms with density f_{H_0}.

1.5. EL ratio test for distribution equality

In this section we present the EL ratio test for examining if two datasets are from the same distribution. The complete derivation for this case can be found in Gurevich and Vexler (2011). In short, let $X_1 = (X_{11}, X_{12}, \ldots, X_{1n_1})$ denote independent observations in the first dataset and $X_2 = (X_{21}, X_{22}, \ldots, X_{2n_2})$ denote independent observations in another dataset. Under H_0 (equal distributions), we assume that both groups are identically distributed. That is, our null hypothesis is

$$H_0 : F_{X_1} = F_{X_2} \tag{15}$$

where F_{X_1} and F_{X_2} denote the cumulative density function (CDF) for the observations in X_1 and X_2, respectively.

To derive the test for (15), we consider that the likelihood ratio can be expressed as,

$$R = \frac{\displaystyle\prod_{i=1}^{n} \prod_{j=1}^{n_i} f_{X_i}(x_{i(j)})}{\displaystyle\prod_{i=1}^{n} \prod_{j=1}^{n_i} f_X(x_{i(j)})}, \tag{16}$$

where the $x_{i(j)}$ indicates the j-th order statistic for the group i. Following the EL concept, we approximate the likelihoods and integrals and obtain the non parametric approximation to (16) as,

$$\tilde{R}_{m,v,n_1,n_2} = \prod_{i=1}^{n_1} \frac{2m}{n_1 \delta_{m1j}} \prod_{j=1}^{n_2} \frac{2v}{n_2 \delta_{v2j}}. \tag{17}$$

The proper selection of m and v in the current literature of entropy-based decision making recommends selecting values utilizing information regarding alternative distributions when sample sizes are finite. Ultimately using the work in Yu *et al.* (2010), we look at selecting m and v by minimizing $\tilde{R}$ over appropriate ranges. This suggests the following test statistic for

n_1	α					n_2				
		10	15	20	25	30	35	40	45	50
10	0.01	5.2249	6.4875	7.0649	7.0628	7.6580	8.2567	8.8410	8.8204	9.4151
	0.03	5.2230	6.4845	7.0614	7.0608	7.6534	8.2508	8.8375	8.8152	9.4125
	0.05	5.2198	6.4803	7.0584	7.0572	7.6496	8.2469	8.8329	8.8116	9.4082
	0.10	5.2174	6.4684	7.0490	7.0461	7.6363	8.2369	8.8207	8.7972	9.3969
	0.30	5.1986	6.4048	7.0166	7.0008	7.5933	8.1853	8.7725	8.7450	9.3487
15	0.01	6.4819	7.7007	8.3128	8.2926	8.8968	9.4823	10.0879	10.0353	10.6478
	0.03	6.4788	7.6977	8.3097	8.2876	8.8930	9.4770	10.0844	10.0316	10.6442
	0.05	6.4738	7.6960	8.3061	8.2844	8.8895	9.4713	10.0813	10.0272	10.6403
	0.10	6.4632	7.6907	8.2988	8.2757	8.8791	9.4588	10.0694	10.0170	10.6299
	0.30	6.4025	7.6443	8.2673	8.2359	8.8400	9.4163	10.0130	9.9653	10.5839
20	0.01	7.0562	8.3157	8.9355	8.9828	9.5897	10.1909	10.7870	10.7672	11.3622
	0.03	7.0526	8.3126	8.9331	8.9784	9.5860	10.1878	10.7825	10.7625	11.3543
	0.05	7.0495	8.3095	8.9308	8.9756	9.5836	10.1844	10.7777	10.7584	11.3503
	0.10	7.0411	8.3003	8.9246	8.9667	9.5749	10.1745	10.7673	10.7456	11.3425
	0.30	7.0082	8.2699	8.8990	8.9270	9.5337	10.1290	10.7282	10.6870	11.2925
25	0.01	7.0695	8.3077	8.9712	9.0562	9.6745	10.2986	10.8985	10.9047	11.5256
	0.03	7.0663	8.3045	8.9673	9.0535	9.6710	10.2929	10.8947	10.9012	11.5200
	0.05	7.0624	8.3010	8.9630	9.0486	9.6678	10.2889	10.8918	10.8976	11.5167
	0.10	7.0505	8.2918	8.9537	9.0396	9.6587	10.2795	10.8830	10.8851	11.5060
	0.30	6.9997	8.2478	8.9175	9.0002	9.6139	10.2385	10.8425	10.8401	11.4572
30	0.01	7.6563	8.8949	9.5886	9.6636	10.2997	10.9110	11.5353	11.5588	12.1544
	0.03	7.6530	8.8907	9.5851	9.6606	10.2951	10.9073	11.5326	11.5553	12.1514
	0.05	7.6490	8.8878	9.5825	9.6574	10.2922	10.9026	11.5282	11.5511	12.1485
	0.10	7.6378	8.8785	9.5727	9.6474	10.2821	10.8922	11.5200	11.5402	12.1387
	0.30	7.5943	8.8380	9.5305	9.6089	10.2397	10.8484	11.4847	11.5012	12.0907
35	0.01	8.2490	9.4841	10.1938	10.2783	10.9105	11.5302	12.1554	12.1973	12.7912
	0.03	8.2455	9.4804	10.1893	10.2739	10.9080	11.5270	12.1509	12.1922	12.7877
	0.05	8.2403	9.4754	10.1853	10.2705	10.9037	11.5232	12.1460	12.1861	12.7847
	0.10	8.2309	9.4664	10.1743	10.2608	10.8926	11.5143	12.1342	12.1759	12.7741
	0.30	8.1824	9.4163	10.1294	10.2220	10.8488	11.4732	12.0924	12.1343	12.7228
40	0.01	8.8545	10.0826	10.7819	10.8933	11.5321	12.1583	12.7709	12.8080	13.4148
	0.03	8.8515	10.0789	10.7777	10.8902	11.5284	12.1546	12.7677	12.8044	13.4108
	0.05	8.8476	10.0752	10.7730	10.8878	11.5241	12.1511	12.7637	12.7998	13.4063
	0.10	8.8348	10.0629	10.7614	10.8759	11.5147	12.1413	12.7521	12.7897	13.3969
	0.30	8.7839	10.0101	10.7171	10.8357	11.4727	12.0982	12.7155	12.7384	13.3528
45	0.01	8.8205	10.0466	10.7657	10.9100	11.5541	12.1744	12.8068	12.8520	13.4708
	0.03	8.8143	10.0425	10.7615	10.9062	11.5501	12.1724	12.8022	12.8457	13.4644
	0.05	8.8094	10.0392	10.7581	10.9025	11.5465	12.1689	12.7976	12.8409	13.4605
	0.10	8.7982	10.0290	10.7463	10.8940	11.5378	12.1581	12.7868	12.8303	13.4505
	0.30	8.7423	9.9804	10.6967	10.8490	11.4923	12.1150	12.7384	12.7843	13.4009
50	0.01	9.4276	10.6427	11.3561	11.4996	12.1675	12.7951	13.4223	13.4760	14.0931
	0.03	9.4231	10.6376	11.3523	11.4937	12.1628	12.7914	13.4176	13.4709	14.0873
	0.05	9.4187	10.6328	11.3484	11.4895	12.1587	12.7859	13.4130	13.4663	14.0842
	0.10	9.4037	10.6219	11.3374	11.4794	12.1460	12.7737	13.4054	13.4546	14.0708
	0.30	9.3520	10.5700	11.2834	11.4316	12.0993	12.7285	13.3588	13.4088	14.0186

Table 1: The critical values for $\log(R_{n_1,n_2})$ with $\delta = 0.10$ for the two sample comparison with various sample sizes n_1 and n_2 at significance level α.

the hypothesis in (15),

$$\tilde{R}_{n_1,n_2} = \min_{l_{n_1} \le m \le u_{n_1}} \prod_{j=1}^{n_1} \frac{2m}{n_1 \Delta_{m1j}} \min_{l_{n_2} \le v \le u_{n_2}} \prod_{j=1}^{n_2} \frac{2v}{n_2 \Delta_{v2j}}, \tag{18}$$
$$l_n = n^{0.5+\delta}, u_n = \min(n^{1-\delta}, n/2), \delta \in (0, 0.25).$$

The Δ_{mij} function is defined as:

$$\Delta_{mij} = \frac{1}{n_1 + n_2} \sum_{k-1}^{2} \sum_{i=1}^{n_i} \left(I(x_{kl} \le x_{i(j+m)}) - I(x_{kl} \le x_{i(j-m)}) \right), \tag{19}$$

where $I()$ denotes an indicator function that takes the value 1 if the condition in the parenthesis is satisfied and takes the value 0, otherwise. Note, here, $x_{i(j+m)} = x_{i(n_i)}$, if $j + m \ge n_i$ and $x_{i(j-m)} = x_{i(1)}$ if $j - m \le 1$.

The test rejects the null hypothesis for large values of $\log \tilde{R}_{n_1,n_2}$. Note that we define $\Delta_{lij} = 1/(n_1 + n_2)$ if $\Delta_{lij} = 0$.

Significance of level α can be determined since $I(X > Y) = I(F(X) > F(Y))$ for any distribution function F. Hence, the null distribution of $\tilde{R}_{n_1,n_2}$ is independent with respect to the form of the underlying distributions given H_0. Hence, we can tabulate universal critical values regardless of the null distribution of the X_{ij}'s.

Table 1 shows the critical values for the logarithm of $\tilde{R}_{n_1,n_2}$ for common sample sizes and significance levels. These critical values were obtained from deriving Monte Carlo roots of

$$P_{H_0}(\log(\tilde{R}_{n_1,n_2}) > C_\alpha) = \alpha$$

based on 75,000 repetitions of sampling $X_{1j} \sim N(0,1)$ and $X_{2j} \sim N(0,1)$.

In the following we present the structure and functioning of the package, with applications to real datasets.

2. What is package dbEmpLikeGOF?

In summary, the **dbEmpLikeGOF** package provides a function `dbEmpLikeGOF` to be used for empirical likelihood based goodness-of-fit tests based on sample entropy. The function can also perform the two sample EL ratio test for the hypothesis in (15). The output of `dbEmpLikeGOF` analysis is an object containing the test statistic and the p value. Standard bootstrap options can be used in conjunction with the object (statistic) in order to make confidence sets a straightforward and automated task.

The proposed function provides the test statistic and p value, where the user can specify an option for the p value to be obtained from a Monte Carlo simulation or via interpolation from stored tables. A complementary function is also included in this package to compute the cut-off value for the appropriate tests of normality and uniformity.

To perform the goodness of fit function, we call the `dbEmpLikeGOF` function:

```
dbEmpLikeGOF(x = data, y = na, testcall = "uniform", pvl.Table = FALSE,
  num.mc = 1000)
```

where data represents a vector of data and the testcall option allows the user to perform the goodness-of-fit test for uniformity (uniform) or normality (normal). The pvl.Table option when set to TRUE employs a stored table of p values to approximate the p value for the given situation, when set to FALSE, a Monte Carlo simulation scheme is employed to estimate the p value. The number of simulations in the Monte Carlo scheme can be controlled using the num.mc option.

In the event that the user specifies both x and y in dbEmpLikeGOF the two sample distribution equality hypothesis in (15) is performed using the logarithm of the statistic in (18).

Further input options for dbEmpLikeGOF include specifying δ (delta) in (11) and δ (delta.equality) in (18). We recommend using the default settings and note that these procedures are fairly robust to the specification of δ.

In certain situations the user may simply be interested in obtaining the cut-off value for a given test and sample size. The function returnCutoff is designed to return the cut-off value for the specified goodness-of-fit test at a given α significance level. For example, the following code:

```
returnCutoff(samplesize, testcall = "uniform", targetalpha = 0.05,
  pvl.Table = FALSE, num.mc = 200)
```

will return the Monte Carlo based test statistic cutoff for determining significance at level 0.05 for the null hypothesis in (8) with decision rule in (12).

The required input for returnCutoff requires the user to specify the sample size (samplesize) and targetalpha represents the significance level of the test. If the user specifies samplesize as a two element vector, then it is assumed that the user is specifying the two sample sizes for the distribution equality test. Note, num.mc represents the number of Monte Carlo simulations performed to estimate the cut-off value. Similar to the dbEmpLikeGOF, there is an option to use stored tables to obtain the cutoff rather than Monte Carlo simulations. The logical variable pvl.Table when true will determine the cut-off from interpolation based on stored tables. Importantly, note that the cutoff values for the test statistics in (5), (11), and (18) are returned on the logarithm scale with base e.

Using the methodology developed by North, Curtis, and Sham (2003), for each test statistic, T_{obs}, the Monte Carlo p value is computed according to the equation below:

$$p \text{ value} = \frac{1 + \sum_{j=1}^{M} I\left(T(x_1, \ldots, x_n) > T_{obs}\right)}{M+1} \tag{20}$$

where M represents the number of simulations and $T(x_1, \ldots, x_n)$ is the statistic from the simulated data $(x_1, \ldots, x_n)$ and T_{obs} is the observed statistic.

2.1. Availability

The **dbEmpLikeGOF** package is available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=dbEmpLikeGOF and also available for download from the author's department webpage (http://sphhp.buffalo.edu/biostat/research/software/dbEmpLikeGOF/index.php).

3. Examples

This section provides examples of using **dbEmpLikeGOF** with the corresponding R code. Using several publicly available datasets and a novel dataset we compare our results with results from other goodness-of-fit tests including Shapiro-Wilks (SW), Kolmogorov-Smirnov (KS), Wilcoxon rank sum (WRS), and Lilliefors (L) tests. The SW test for normality was implemented using the R function `shapiro.test`. The one and two sample KS tests were implemented using the R function `ks.test`. The Lilliefors test introduced in Lilliefors (1967) is an adaptation of the Kolmogorov-Smirnov test for normality. We have included the Lilliefors test for normality as implemented in the R package **nortest** in our simulations (Gross 2012). The two sample WRS test was implemented using the R function `wilcox.test` (R Core Team 2013).

Note that Monte Carlo studies presented in Vexler and Gurevich (2010a) and Gurevich and Vexler (2011) showed various situations when the density based EL test clearly outperformed the classical procedures.

3.1. Real data examples

Snowfall dataset

We consider the 63 observations of the annual snowfall amounts in Buffalo, New York, as observed from 1910/11 to 1972/73 (data in Figure 3 and Table 2); see, for example, Parzen (1979). Importantly, we note that the snowfall data is not normally distributed, namely due to the impossibility of negative snowfall amounts. However, we have examined numerous publications that have studied this data and all of them recognize that the data is sufficiently close to normally distributed. In Carmichael (1976) the Buffalo snowfall dataset was "chosen to illustrate the response of the different methods to approximately normal data." More notably, Parzen (1979) states "all our $\tilde{D}$ and $|\hat{\varphi}|^2$-based diagnostic tests of the hypothesis H_0

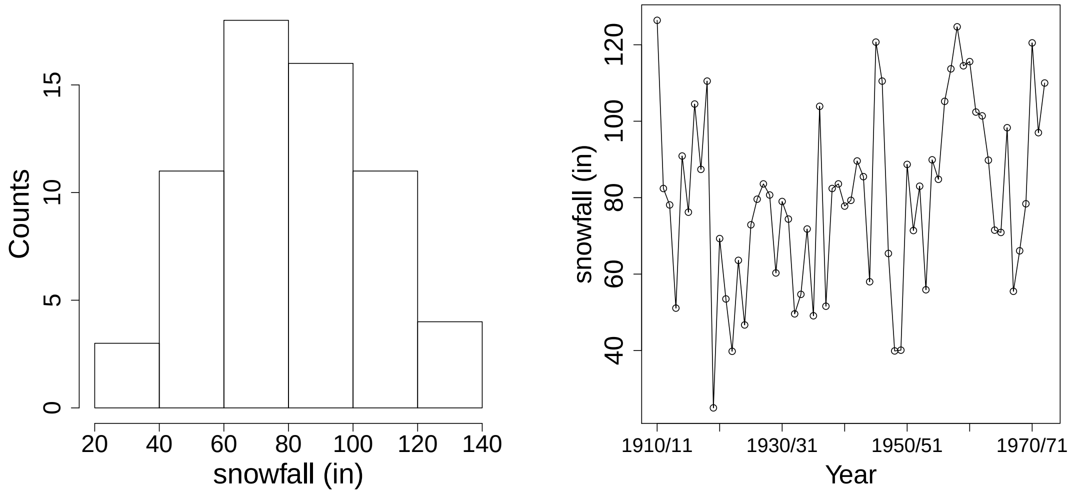

Figure 3: Left: Histogram of snowfall data in Buffalo, NY from 1910/11 to 1972/73. Right: Snowfall data displayed as a time series. Using the EL test for normality, we conclude that the distribution for the data is consistent with a normal distribution (p value $= 0.321$).

Buffalo snowfall dataset ($n = 63$)								
126.4	82.4	78.1	51.1	90.9	76.2	104.5	87.4	110.5
25.0	69.3	53.5	39.8	63.6	46.7	72.9	79.6	83.6
80.7	60.3	79.0	74.4	49.6	54.7	71.8	49.1	103.9
51.6	82.4	83.6	77.8	79.3	89.6	85.5	58.0	120.7
110.5	65.4	39.9	40.1	88.7	71.4	83.0	55.9	89.9
84.8	105.2	113.7	124.7	114.5	115.6	102.4	101.4	89.8
71.5	70.9	98.3	55.5	66.1	78.4	120.5	97.0	110.0

Table 2: The amount of snowfall in Buffalo, New York, for each of 63 winters from 1910/11 to 1972/73. See Parzen (1979) for more details.

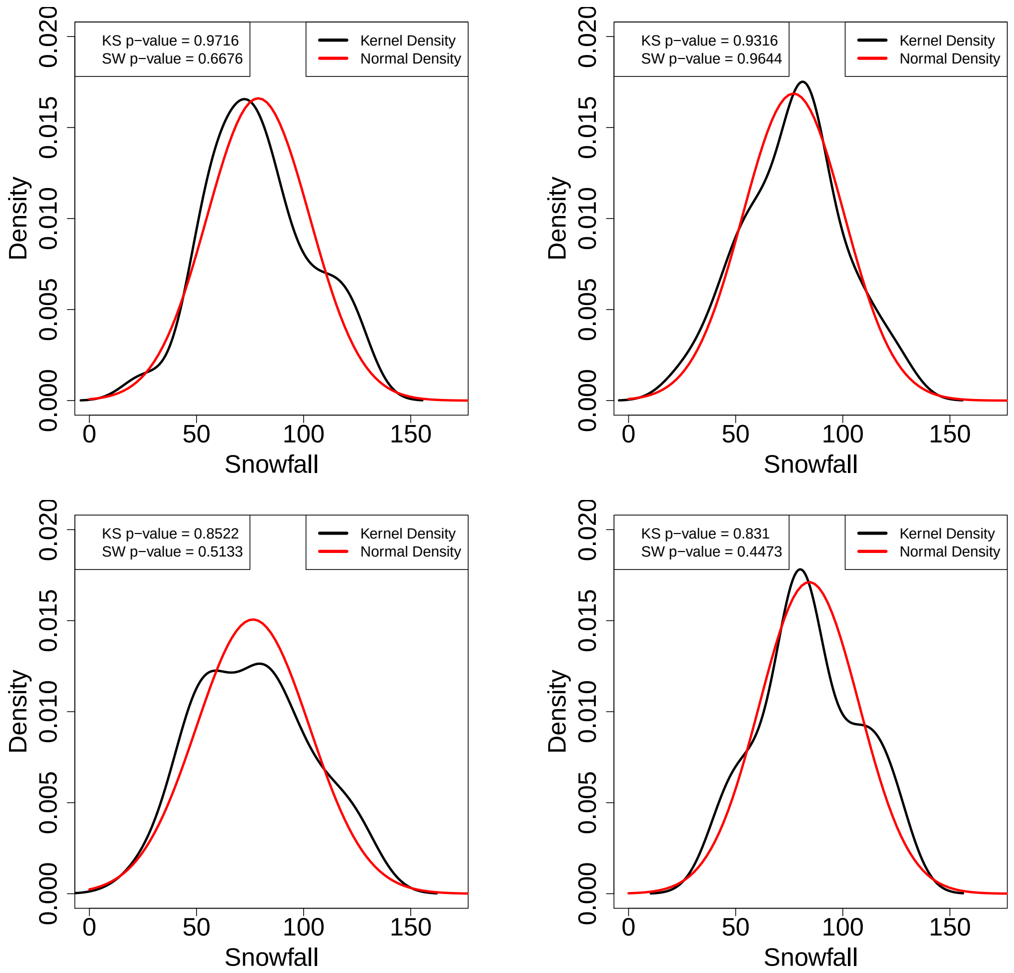

Figure 4: Snow fall dataset examples where the density-based EL statistic is significant (p value < 0.05), but the Kolmogorov-Smirnov (KS) test and Shapiro Wilks (SW) test are not significant (p values > 0.05). The normal density (red curve) is determined using the sample mean and sample standard deviation to estimate the mean and standard deviation. The black curve represents the kernel density for a randomly chosen subset from the snowfall dataset.

that Buffalo snowfall is normal confirm that it is. The quantile-box plot of Buffalo snowfall, given in Figure E, also indicates that it is normal." For these reasons, we believe the snowfall dataset is sufficiently close to Gaussian.

We perform the proposed test for (1) with the statistic in (5). We obtain the value of the test statistic to be 8.49 with an MC based p value of 0.321 using the following command,

```
R> data("Snow")
R> dbEmpLikeGOF(x = snow, testcall = "normal", pvl.Table = FALSE,
+    num.mc = 5000)
```

where `snow` represents the vector of annual snowfall amounts. Note, when using a KS test to examine the same hypothesis for the snowfall dataset we obtain a p value of 0.9851 and a SW p value of 0.5591. Thus, we conclude that there is not significant evidence to conclude that the snowfall data is inconsistent with a normal distribution.

To examine the robustness of our tests, we employed a resampling technique where we randomly removed 10, 20 and 50 percent of the data and examined the significance of the test statistics derived from the remaining dataset. For each test, we repeated this technique 2000 times where the results are summarized in Table 3. When randomly removing 10, 20, and 50 percent of the data, we obtained a significant density based EL test statistic in 3, 5.8 and 6.6 percent of the simulations, respectively. From the work in Parzen (1979), it is suggested that the Snowfall dataset follows a normal distribution. Table 4 displays the average p value for each of the tests when randomly removing the data. Ultimately, the results from randomly removing a percentage of observations demonstrates the robustness of the proposed test in controlling the type I error.

To study the power of the EL statistic, we examine four snowfall datasets where each dataset is obtained by randomly removing 50 percent of the snowfall data. These datasets represent examples where the EL based test is significant (p value < 0.05), while the KS and SW tests are not significant (p values > 0.05). These examples are summarized by displaying the kernel density estimates and the hypothesized distributions as shown in Figure 4. From the examples in Figure 4, there is the potential for the EL tests to be more powerful than KS and SW tests.

Birth dataset

As another example of `dbEmpLikeGOF`, we examine a baby boom dataset summarizing the time of birth, sex, and birth weight for 44 babies born in one 24-hour period at a hospital in Brisbane, Australia. These data appeared in an article entitled "Babies by the Dozen for Christmas: 24-Hour Baby Boom" in the newspaper The Sunday Mail on 1997-12-21. According to the article, a record 44 babies were born in one 24-hour period at the Mater Mothers' Hospital, Brisbane, Australia, on 1997-12-18. The article listed the time of birth, the sex, and the weight in grams for each of the 44 babies where the full dataset can be found at Dunn (1999). We examine whether an exponential distribution can be used to model the times between births. From the work in Dunn (1999), it is suggested that this data is exponentially distributed. The data summarizing the time between births is shown in Table 5. Using the KS test for an exponential distribution, we obtain a p value of 0.3904. We transform the data in Table 5 using the inverse exponential distribution and thus the transformed data can be examined using the EL ratio test for uniformity. The following command returns the test statistic (11) and p value,

Dataset	Test	10% removed	20% removed	50% removed
Snowfall	EL	0.0315	0.0705	0.0575
	KS	0.0000	0.0000	0.0000
	SW	0.0000	0.0000	0.0010
	L	0.0000	0.0000	0.0065
Birth	EL	0.0020	0.0035	0.0165
	KS	0.0020	0.0070	0.0125

Table 3: Resampling results for Snowfall dataset and Birth dataset. With 2000 simulations, we randomly remove 10, 20, and 50 percent of the observations in the original dataset (Snowfall or Birth). In each remaining dataset, we compute the test statistic and the percentage of significant test statistics (at level 0.05) are summarized in each cell. EL refers to the density-based empirical likelihood test; KS, SW, and L denote the Kolmogorov-Smirnov, Shapiro-Wilks, and Lilliefors tests, respectively.

Dataset	Test	10% removed	20% removed	50% removed
Snowfall	EL	0.2981	0.3001	0.3396
	KS	0.9589	0.9309	0.8792
	SW	0.5576	0.5480	0.5454
	L	0.7704	0.6960	0.6062
Birth	EL	0.6108	0.5686	0.5080
	KS	0.4625	0.5151	0.5547

Table 4: Resampling results for Snowfall dataset and Birth dataset. With 2000 simulations, we randomly remove 10, 20, and 50 percent of the observations in the original dataset (Snowfall or Birth). In each remaining dataset, we compute the test statistic and the mean p values are summarized in each cell. EL refers to the density-based empirical likelihood test; KS, SW, and L denote the Kolmogorov-Smirnov, Shapiro-Wilks, and Lilliefors tests, respectively.

```
R> data("Baby")
R> dbEmpLikeGOF(x = baby, testcall = "uniform", pvl.Table = FALSE,
+    num.mc = 5000)
```

where `baby` represents the vector of transformed data. When this test is employed, we observe a MC based p value of 0.6626. Ultimately, for this data the time between births can be adequately modeled using an exponential distribution.

Similar to the snowfall dataset, we examine the robustness of our results by employing a bootstrap scheme where the bootstrap resamplings are taken when removing 10, 20, or 50 percent of the original dataset. The results are summarized in Tables 3 and 4. With 2000 simulations where we randomly remove 10, 20, and 50 percent of the observations from the original dataset, we find significant statistics in 0, 0.2, and 2 percent of the simulated datasets, respectively.

To examine the power of the EL statistic, we examine four birth datasets where the EL test is significant (p value < 0.05), while the KS test is not significant (see Figure 5). Figure 5 displays the data driven kernel density estimate against the hypothesized distribution. From these examples, there may be situations where the EL test for uniformity may be more powerful than the traditional KS test.

Inter-time births ($n = 43$)		
Time between births (minutes)	Tally	Empirical probability
0–19	18	0.419
20–39	12	0.279
40–59	6	0.140
60–79	5	0.116
80+	2	0.047
Total	43	1.001

Table 5: The time between births for 44 babies born in one 24 hour period at the Mater Mothers' Hospital, Brisbane, Australia, on 1997-12-18. See Dunn (1999) for more details.

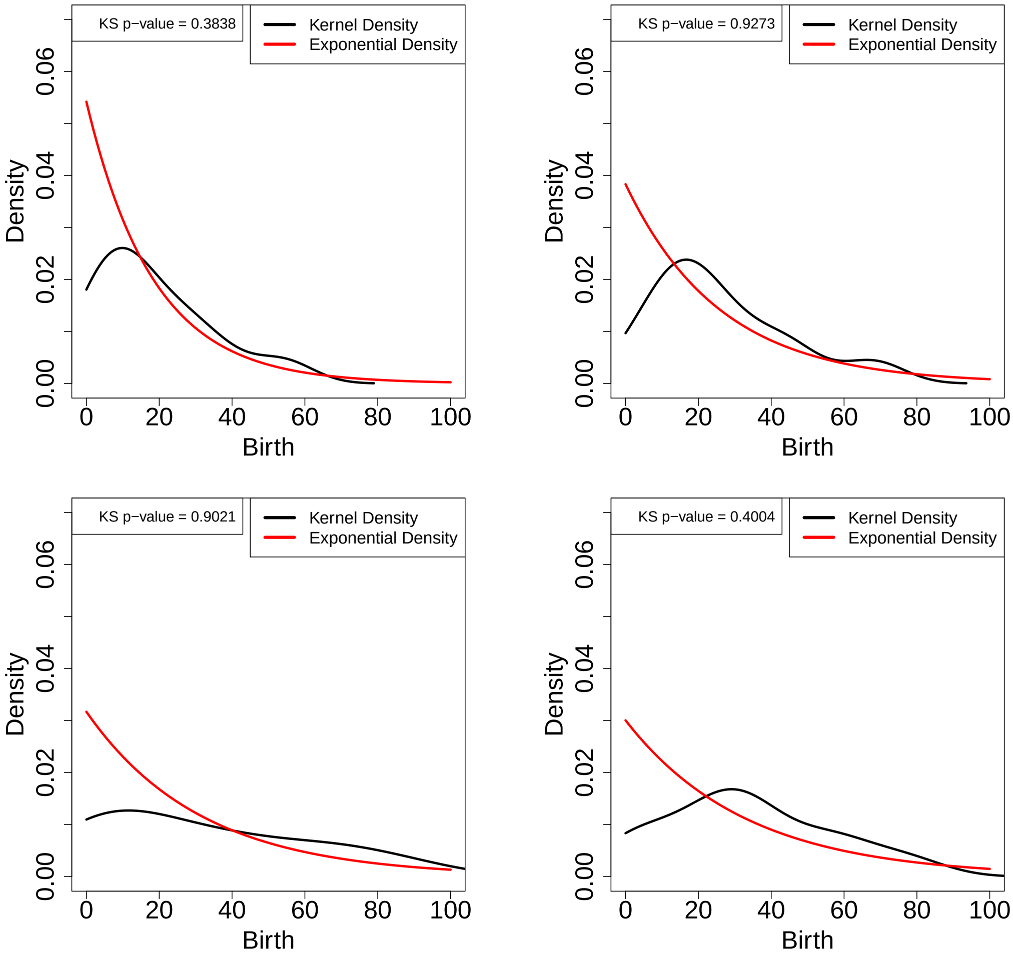

Figure 5: Birth dataset examples where the density-based EL statistic is significant (p value < 0.05), but the Kolmogorov-Smirnov (KS) test is not significant (p values > 0.05). The exponential density (red curve) has a rate parameter of the inverse of the sample mean. The black curve represents the kernel density for a randomly chosen subset from the birth dataset.

A TBARS data example

For a novel analysis using the density-based EL software, we consider data from a study evaluating biomarkers related to atherosclerotic (CHD) coronary heart disease (see acknowledgments). A population-based sample of randomly selected residents of Erie and Niagara counties of the state of New York, United States of America, was the focus of this investigation. The New York State Department of Motor Vehicles drivers' license rolls were utilized as the sampling frame for adults between the ages of 35 and 65; where the elderly sample (age 65–79) was randomly selected from the Health Care Financing Administration database. Participants provided a 12-hour fasting blood specimen for biochemical analysis at baseline, and a number of parameters were examined from fresh blood samples. A complete description of this dataset is available at Schisterman, Faraggi, Browne, Freudenheim, Dorn, Muti, Armstrong, Reiser, and Trevisan (2001).

A cohort of 5620 men and women were selected for the analyses yielding 1209 cases (individuals that had a heart attack) and 4411 controls (no heart attack history). In a subset of this dataset, we examine the significance of the thiobarbituric acid reactive substances (TBARS) variable which is known to play a role in atherosclerotic coronary heart disease process. TBARS was measured in patient serum samples using reverse-phase high performance liquid chromatography and spectrophotometric approaches.

For the analysis of the TBARS dataset, we would like to test the claim that the TBARS distribution is different between the cohort of patients that have suffered a heart attack and the cohort of patients that have not suffered a heart attack. If the null hypothesis is true, we expect the empirical distribution of the TBARS variable to be very similar in the two cohorts, if the null hypothesis is not true, we expect the empirical distributions to be very different (e.g., TBARS is stochastically greater in the heart attack population). A quantile-quantile (QQ) plot of this data is shown in Figure 6 (left).

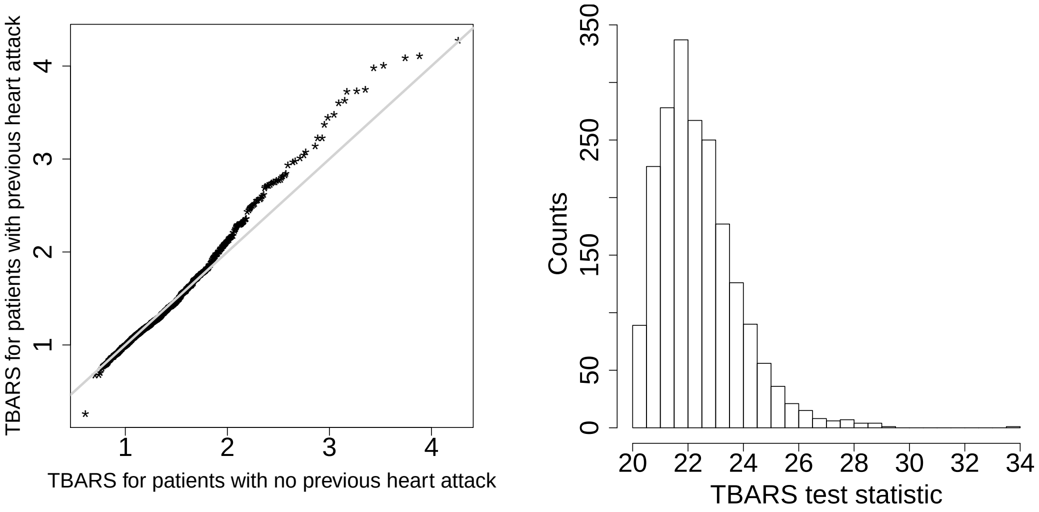

Figure 6: Left: A quantile-quantile (QQ) plot comparing the distribution of TBARS for patients with a previous heart attack against the distribution of TBARS for patients without a previous heart attack. Right: Histogram of the test statistic for TBARS distribution equality based on 2000 bootstrap resamplings.

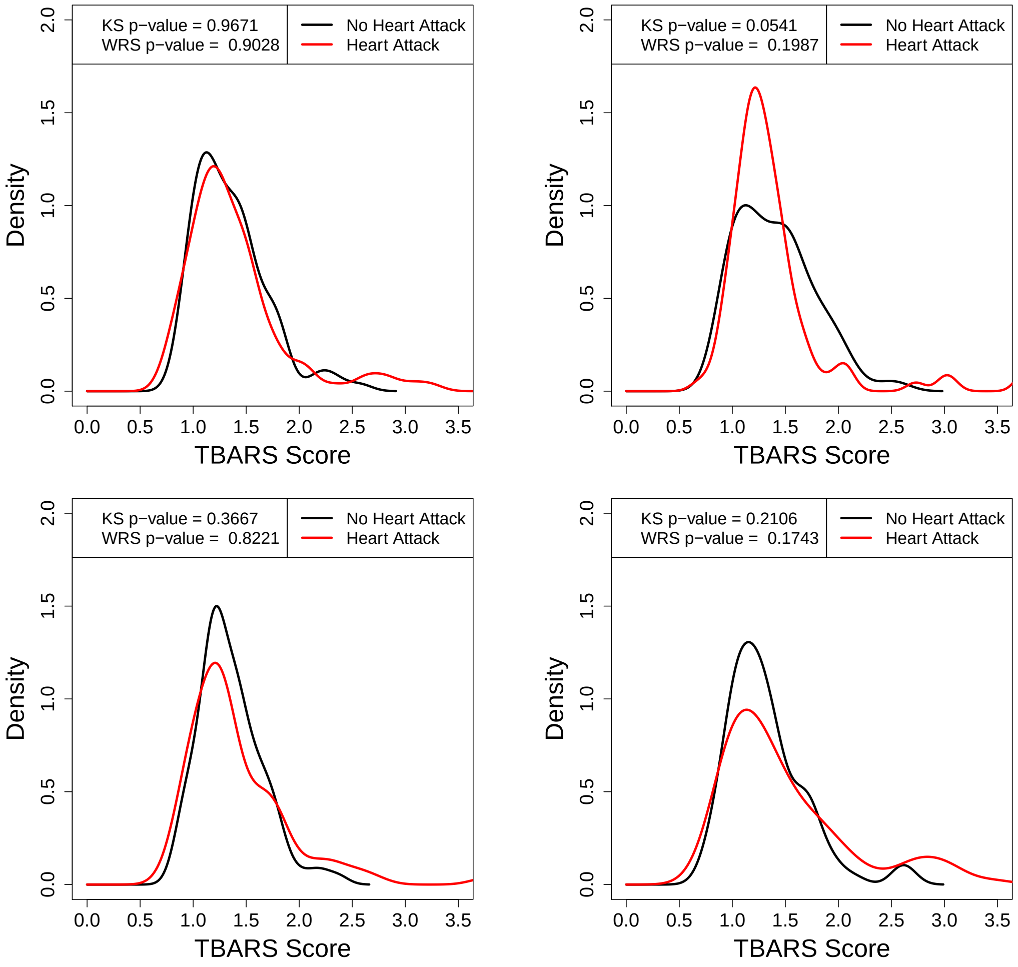

Figure 7: TBARS examples where the density-based EL distribution equality statistic is significant (p value < 0.05), but the two-sample Kolmogorov-Smirnov (KS) test and two-sample Wilcoxon rank sum (WRS) test are not significant (p values > 0.05).

We employed a bootstrap strategy to study the TBARS variable using the statistic in (18). The strategy was based on randomly choosing 200 patients, where 100 patients had previously suffered a heart attack and 100 patients did not have a heart attack. The distribution of TBARS was examined for equality between the heart attack patient cohort and the no heart attack patient cohort. We repeated this procedure 2000 times calculating the frequency of the event of a significant statistic. Rather than obtain a p value associated with each statistic, we employed the `returnCutoff` command to obtain the cutoff for significance,

```
R> tbar.cut <- returnCutoff(100, testcall = "distribution.equality",
+    targetalpha = 0.05, num.mc = 5000)
```

Figure 6 (right) displays the histogram of the logarithm of the two sample test statistic calculated in (18). 5.9 percent of the bootstrap resamplings yielded a significant test statistic. These results were also compared against the two sample KS test and two sample WRS test to compare distribution equality. Using the KS test, 3.4 percent of the resamplings were significant (p value < 0.05). Using the two sample WRS test, 4.5 percent of the resamplings were significant (p value < 0.05).

Thus all of the statistical tests suggest that TBARS is not significantly different in heart attack patients. This is further confirmed when examining the mean p values in the resampled datasets. The mean p values are 0.5262, 0.5002, and 0.4489 for the KS, WRS, and empirical likelihood tests, respectively.

To examine the power of the two sample EL test, we focus on four examples comparing 100 patients that had previously suffered a heart attack and 100 patients that did not have a heart attack, where the EL statistic is significant, however, the KS and WRS tests are not significant. Figure 7 displays the density for each cohort when a kernel density smoother is employed with the bandwidth chosen according to Equation 3.31 in Silverman (1986). These examples highlight situations where there may be a power advantage obtained using the density-based EL statistics over traditional goodness-of-fit tests such as the two sample KS test and two sample WRS test.

4. Conclusions

The package **dbEmpLikeGOF** provides R users with a new and powerful way to perform goodness-of-fit tests using empirical likelihood ratios. We focus on two sample tests and tests for normality and uniformity which are common distributions to test in applied studies. Monte Carlo methods and interpolation are used to estimate the cutoff values and *exact* p values for the proposed tests. The proposed procedure can execute entropy based structured tests that have not been addressed in statistical software. We believe that the **dbEmpLikeGOF** package will help investigators to use density based empirical likelihood approaches for goodness-of-fit tests in practice.

Acknowledgments

The authors are grateful to Dr. Jo L. Freudenheim, chairperson of the Department of Social and Preventive Medicine at the SUNY University at Buffalo, for generously allowing us to use the TBARS dataset.

Further, the authors would also like to acknowledge and thank the Editors and referees for their helpful comments and suggestions that improved this article.

This research is supported by the NIH grant 1R03DE020851 – 01A1 (the National Institute of Dental and Craniofacial Research).

References

Carmichael JP (1976). “The Autoregressive Method: A Method of Approximating and Estimating Positive Functions.” *Technical Report 45*, Statistical Science Division, State University of New York at Buffalo.

Darling D (1957). “The Kolmogorov-Smirnov, Cramér-von Mises Tests.” *The Annals of Mathematical Statistics*, **28**(4), 823–838.

Dodge Y (2006). *The Oxford Dictionary of Statistical Terms.* Oxford University Press, USA.

Dunn PK (1999). "A Simple Data Set for Demonstrating Common Distributions." *Journal of Statistics Education*, **7**(3).

Gross J (2012). ***nortest**: Tests for Normality.* R package version 1.0-2, URL http://CRAN.R-project.org/package=nortest.

Gurevich G, Vexler A (2011). "A Two-Sample Empirical Likelihood Ratio Test Based on Samples Entropy." *Statistics and Computing*, **21**(4), 657–670.

Hollander M, Wolfe DA, Wolfe DA (1973). *Nonparametric Statistical Methods.* John Wiley & Sons.

Lilliefors HW (1967). "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown." *Journal of the American Statistical Association*, **62**(318), 399–402.

North B, Curtis D, Sham P (2003). "A Note on the Calculation of Empirical p values from Monte Carlo Procedures." *American Journal of Human Genetics*, **71**(2), 439–441.

Owen AB (2001). *Empirical Likelihood.* CRC Press, New York.

Parzen E (1979). "Nonparametric Statistical Data Modeling." *Journal of the American Statistical Association*, **74**(365), 105–121.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Royston P (1991). "Estimating Departure from Normality." *Statistics in Medicine*, **10**(8), 1283–1293.

Schisterman EF, Faraggi D, Browne R, Freudenheim J, Dorn J, Muti P, Armstrong D, Reiser B, Trevisan M (2001). "TBARS and Cardiovascular Disease in a Population-Based Sample." *Journal of Cardiovascular Risk*, **8**(4), 219–225.

Silverman BW (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall/CRC.

Vexler A, Gurevich G (2010a). "Density-Based Empirical Likelihood Ratio Change Point Detection Policies." *Communications in Statistics – Simulation and Computation*, **39**(9), 1709–1725.

Vexler A, Gurevich G (2010b). "Empirical Likelihood Ratios Applied to Goodness-of-Fit Tests Based on Sample Entropy." *Computational Statistics & Data Analysis*, **54**(2), 531–545.

Vexler A, Yu J, Tian L, Liu S (2010). "Two-Sample Nonparametric Likelihood Inference Based on Incomplete Data with an Application to a Pneumonia Study." *Biometrical Journal*, **52**(3), 348–361.

Yu J, Vexler A, Tian L (2010). "Analyzing Incomplete Data Subject to a Threshold Using Empirical Likelihood Methods: An Application to a Pneumonia Risk Study in an ICU Setting." *Biometrics*, **66**(1), 123–130.

Affiliation:

Jeffrey Miecznikowski
Department of Biostatistics
University at Buffalo
Kimball Tower Rm 723
3435 Main Street
Buffalo NY 14214, United States of America
E-mail: jcm38@buffalo.edu
URL: http://sphhp.buffalo.edu/biostatistics/faculty-and-staff/faculty-directory/jcm38.html

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 54, Issue 3 *Submitted:* 2011-03-02
August 2013 *Accepted:* 2013-07-26