Density-Based Empirical Likelihood Procedures for Testing Symmetry of Data Distributions and K-Sample Comparisons

Albert Vexler,
Department of Biostatistics, New York State University at Buffalo, Buffalo, NY 14124, USA
avexler@buffalo.edu

Hovig Tanajian,
Department of Biostatistics, New York State University at Buffalo, Buffalo, NY 14124, USA
hovigtan@buffalo.edu

Alan D. Hutson,
Department of Biostatistics, New York State University at Buffalo, Buffalo, NY 14124, USA
ahutson@buffalo.edu

Abstract. In practice, parametric likelihood based techniques are commonly most powerful statistical tools. This paper proposes and examines novel and simple distribution-free test statistics that efficiently approximate parametric likelihood ratios to analyze and compare distributions of K-groups of observations. Using the density-based empirical likelihood (DBEL) methodology, we develop a STATA package that can be applied to test for symmetry of data distributions as well as compare K-sample distributions. Recognizing the fact that recent statistical software packages do not sufficiently address K-sample nonparametric comparisons of data distributions, we propose a new STATA command VX_DBEL to execute exact DBEL ratio tests based on K-samples. To calculate p-values of the proposed exact tests, we use the following methods: (1) a classical technique based on Monte Carlo (MC) p-value evaluations; (2) an interpolation technique based on tabulated critical values; (3) a new hybrid technique that combines (1) and (2). The third cutting edge method is shown to be very efficient in the context of exact-tests' p-value computations. This Bayesian type method considers tabulated critical values as a prior information and MC generations of test statistic values as data used to depict the likelihood function. In this case, a nonparametric Bayesian method is proposed to compute critical values of exact tests.

Keywords: empirical likelihood, likelihood ratio, nonparametric tests, exact tests, K-sample comparisons, symmetry, p-value computation

Introduction

The classical statistical literature proposes the parametric likelihood ratio (LR) based decision making procedures as techniques for testing simple and composite hypotheses. In a variety of scenarios the parametric LR tests oftentimes are the most powerful statistical tools (e.g., Lehman and Romano, 2005; Vexler *et al*., 2010).

The likelihood principle is arguably the most important concept for inference under parametric model assumptions. As an example, consider the goodness-of-fit testing problem, where given a sample of *n*

independent identically distributed (i.i.d.) observations $X_1,\ldots, X_n$, we are interested in testing the hypothesis

$$H_0 : X_1,\ldots, X_n \sim F_0 \qquad \text{versus} \qquad H_1 : X_1,\ldots, X_n \sim F_1,$$

where F_0 and F_1 are some hypothesized distributions with the density functions $f_0(x)$ and $f_1(x)$, respectively. By virtue of the Neyman-Pearson Lemma, the most powerful test statistic for this hypothesis is the LR $\prod_{i=1}^{n} f_1(X_i) / f_0(X_i)$, where the density functions f_0 and f_1 are assumed to be completely known.

One of the very attractive properties of the parametric LR methodology is that the likelihood principle provides researchers with clear algorithms for constructing efficient statistical tests across various complicated problems as they relate to practical applications. For example, we can easily extend the solution of the goodness-of-fit problem considered above to the two sample context with i.i.d. observations $X_1,..., X_{n_1}$ and i.i.d. observations $Y_1,..., Y_{n_2}$ relative to testing

$$H_0 : F_X = F_Y \text{ vs. } H_1 : F_X = F_1, F_Y = F_2, F_1 \neq F_2,$$

where F_X, F_Y and F_1, F_2 are distribution functions of observed data under the null and alternative hypothesis, respectively. The resulting most powerful test statistic has the form

$$\prod_{i=1}^{n_1} \frac{f_1(X_i)}{f_X(X_i)} \prod_{j=1}^{n_2} \frac{f_2(Y_j)}{f_X(Y_j)},$$

where f_X, f_1, and f_2 are the corresponding density functions that are assumed to be known.

Our research is focused on modifying the traditional parametric LR testing to the nonparametric setting to gain a degree of robustness while not sacrificing substantial power. In general, the concept of the parametric LR testing method may not be applicable in the nonparametric setting or when the number of unknown parameters involved in hypothesis testing is relatively large (e.g., Fan, *et al.*, 2001). It is also known that when key assumptions are not met, parametric approaches may be non-robust to the underlying assumptions, suboptimal or extremely biased. A very important goal in our research is to preserve the efficiency of statistical testing while maintaining robustness via the use of robust distribution-free likelihood type methods. As part of our approach towards developing new methods we extend and develop new methods based on empirical likelihood (EL) methods (e.g., Owen, 2001; Yu, *et al.*, 2010).

The EL methodology provides an efficient nonparametric analog to the most powerful parametric likelihood methods. An outline of the EL approach is as follows: For i.i.d. observations

$X_1,...,X_n$, the EL function has the form $L_p = \prod_{i=1}^{n} p_i$, where the components $p_i \in (0,1)$, $i=1,\ldots,n$ maximize the likelihood L_p. The maximization process is conditional on a set of empirical constraints defined under the null hypothesis. For example, define the null hypothesis of interest as $H_0 : EX_1 = 0$. It follows that the constraints of interest are $\sum_{i=1}^{n} p_i = 1$ and $\sum_{i=1}^{n} p_i X_i = 0$ given by the empirical counterpart of $H_0 : EX_1 = 0$. Computation of $p_i's$ is based on a straightforward implementation of the method of Lagrange multipliers. This approach is a result of consideration of the 'distribution function'–based likelihood $\prod_{i=1}^{n}(F(X_i) - F(X_i-))$ over all distribution functions F (e.g., Owen 2001).

Whereas the Neyman-Pearson principle uses a density-based structure of the likelihood ratio, the classical EL approach employs a distribution-based likelihood. Towards this end, we develop exact tests by modifying the classical EL approach to be based directly on density functions (e.g., Vexler and Yu, 2010). This method provides nonparametric approximations to Neyman–Pearson type tests.

Vexler and Gurevich (2010) proposed to modify the main idea of the EL technique to develop density-based empirical approximations to the likelihood having the form

$$L_f = \prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} f(X_{(i)}) = \prod_{i=1}^{n} f_i ,$$

where $f_i = f(X_{(i)})$ and $X_{(i)} \le ... \le X_{(n)}$ are the order statistics derived from the sample $X_1,...,X_n$. In this case, following the maximum EL technique, we can obtain estimated values f_i, $i=1,\ldots,n$ that maximize L_f satisfying an empirical constraint that corresponds to $\int f(u)du = 1$. Gurevich and Vexler (2010), Vexler and Yu (2010), as well as Vexler *et al*., (2013a) employed this approach to create very powerful DBEL ratio tests for two-sample comparisons and symmetry. Miecznikowski *et al*., (2013) developed an R package (R Development Core Team, 2007) for DBEL ratio goodness-of-fit tests. Tsai *et al*. (2013) showed the DBEL tests are robust and significantly outperform classical procedure, which include the Kolmogorov-Smirnov test, Wilcoxon rank-sum test and t-test.

In this paper we extend these results to K-sample comparisons, including considerations for ordered alternatives (Gurevich, 2012). Note that, despite the fact that in various clinical trials, investigators need to compare K samples in nonparametric settings, to our knowledge software procedures based on the Kruskal-Wallis test are only techniques to execute nonparametric test for a three-sample comparison.

The proposed tests are exact, i.e. their null distributions are defined independently of data distributions and their critical values can be evaluated without using asymptotic approximations.

In terms of the inferential procedure, we note that the following two methods are commonly used for calculating the p-values in statistical software products. The Monte Carlo (MC) method is a well-known technique for accurately approximating the critical values and/or p-values of exact tests. For some testing situations the use of the MC technique can be computationally intensive. For example, a relatively large number of MC repetitions, say *t*, are needed to evaluate critical values that correspond to the 1% significance level, since in this case the common 95% confidence interval of such evaluation can be calculated as $\left(0.01 \pm 1.96\sqrt{(0.01)(1-0.01)/t}\right)$. The use of tables with corresponding critical values is also a standard method applied in various statistical software routines. Commonly, this method is referred to as "interpolation" in the literature. Providing the tables for use within the testing algorithm improves the execution speed of the test. However, the interpolation method becomes less reliable when real data characteristics (e.g. sample sizes) differ from those used to tabulate the critical values (e.g., Pearson and Hartley, 1966).

As advancement to the methods described above we propose a new hybrid method that combines both the interpolation and MC methods. The result of the implementation of this hybrid method is an innovation related to applications of exact tests in statistical software, which can be applied in the broader setting.

In developing the STATA package, we used the mata matrix programming language. Utilizing the proposed package we provide MC comparisons between classical procedures and our newly proposed command. The results are shown in Section 4. Section 5 presents applications of the proposed three-sample test to real data examples. Section 6 provides some concluding remarks.

2 Method

In this section, we develop the DBEL ratio tests for symmetry of data distributions as well as K-sample distributions comparisons.

2.1 Tests for Symmetry

Consider the problem of testing symmetry of one-sample distribution about zero. We suppose that the data consists of a sample of i.i.d. observations $X_1,...,X_n$. Our hypothesis of interest in this setting is $H_0 : F_x(u) = 1 - F_x(-u)$, for all $-\infty \le u \le \infty$ vs. $H_1 : F_x(u) \ne 1 - F_x(-u)$, for some $-\infty \le u \le \infty$, where the distribution F_x of the observations is assumed unknown. In this case, the LR has the form

$$\text{LR}=\frac{\textit{likelihood under } H_1}{\textit{likelihood under } H_0}=\frac{\prod_{i=1}^{n} f_{H_1}(X_i)}{\prod_{i=1}^{n} f_{H_0}(X_i)}=\frac{\prod_{j=1}^{n} f_{H_1}(X_{(j)})}{\prod_{j=1}^{n} f_{H_0}(X_{(j)})}=\frac{\prod_{j=1}^{n} f_{H_1,j}}{\prod_{j=1}^{n} f_{H_0,j}}, \quad f_{H_k}(X_{(j)})= f_{H_k,j}, k=0,1,$$

where $X_{(j)}$ are the order statistics based on the observations $X_1,...,X_n$.

This LR is the most powerful test statistic in the parametric setting. However, in the nonparametric setting the LR statistic is not computable since the density functions are unknown. To generate the nonparametric test for the hypothesis above, we consider the likelihood $L_f = \prod_{j=1}^{n} f_{H_1,j}$. We will estimate the values of $f_{H_1,j}$ that maximize L_f given an empirical version of the constraint $\int f_{H_1} = 1$. Following Vexler and Gurevich (2010) this constraint has the form of $\frac{1}{2m}\sum_{j=1}^{n}\int_{X_{(j-m)}}^{X_{(j+m)}} \frac{f_{H_1}(u)}{f_{H_0}(u)} f_{H_1}(u)du \le 1$. The empirical constraint on $f_{H_1,j}$'s may then be reformulated as

$$\frac{1}{2m}\sum_{j=1}^{n}\frac{f_{H_1,j}}{f_{H_o,j}}\Delta_{jm} = 1-\frac{m+1}{2n}, \quad \Delta_{jm} := \frac{1}{2n}\sum_{i=1}^{n}\Big(I(X_{(j-m)} \le X_i \le X_{(j+m)})) + I(X_{(j-m)} \le -X_i \le X_{(j+m)})\Big),$$

where *I*() is an indicator function, $X_{(i)} = X_{(1)}$, if $i \le 1$, and $X_{(i)} = X_{(n)}$, if $i \ge n$. The method of Lagrange multipliers is then used to find values of $f_{H_1,j}$, leading to

$$f_{H_1,j} = f_{H_o,j}\frac{2m(1-(m+1)(2n)^{-1})}{n\Delta_{jm}}, j=1,...n.$$

Hence, the EL approximation to L_f and LR can be presented as

$$\prod_{j=1}^{n} f_{H_0,j}\frac{2m(1-(m+1)(2n)^{-1})}{n\Delta_{jm}} \text{ and } V_{nm} = \prod_{j=1}^{n}\frac{2m(1-(m+1)(2n)^{-1})}{n\Delta_{jm}},$$

respectively. Thus, it follows that the maximum EL method forms the LR test statistic

$$V_n = \min_{a_n \le m \le b_n} V_{nm}, \quad a_n = n^{0.5+\delta}, \; b_n = \min(n^{1-\delta}, n/2), \; \delta \in (0,0.25)$$

(see for details Vexler *et al*., 2013b). For practical purposes we suggest a value of $\delta = 0.1$ for our applications. It was shown in Tsai *et al.* (2013) and Vexler *et al*. (2013b) that the power of the test statistic does not differ substantially for values of $\delta \in (0,0.25)$. Similar to the rational provided in Gurevich and Vexler (2011) we will set $\Delta_{jm} = 1/n$, if $\Delta_{jm} = 0$ in terms of practical applications.

The proposed test is now designed to reject the null hypothesis iff $\log(V_n) > C$, where C is the critical value of the test. As shown by Vexler *et al*. (2013b), a test based on the statistic $\log(V_n)$ has an asymptotic power one, i.e. it is a consistent test. It turns out that the null distribution of the test statistic

V_n is independent of the distribution of observations $X_1,...,X_n$. Our test statistic is based on indicator functions involved in the definition of Δ_{jm}. Since $I(X > Y) = I(F_X(X) > F_X(Y))$ it follows that

$$P_{H_0}\{\log(V_n) > C\} = P_{X_1,...,X_n \sim norm(0,1)}\{\log(V_n) > C\}.$$

By virtue of this result, the proposed test is exact, the corresponding critical values can be tabulated for fixed sample sizes. In the interpolation method presented in Section 3 the MC approach is used to obtain the related critical values which are stored in STATA beforehand to increase the execution speed of test.

Next, we consider the one-sided version of the two-sided problem (see Vexler *et al*., 2013b). We are interested in verifying that the sample is generated from a distribution that is stochastically greater than zero. That is, we want to test for

$H_0 : F_x(u) = 1 - F_x(-u)$, for all $-\infty \le u \le \infty$ vs. $H_1 : F_x(u) \le 1 - F_x(-u)$, for some $-\infty \le u \le \infty$,

where the distribution F_x is assumed to be unknown.

Applying the DBEL concept, we have the EL ratio test statistic

$$V_n^* = \min_{a_n \le m \le b_n} \prod_{j=1}^{n} \frac{2m(1-(m+1)(2n)^{-1})}{n\Delta_{jm}}, \text{ with } \Delta_{jm} = \max \frac{1}{n}\left(\sum_{i=1}^{n} I(X_i \le X_{(j+m)}), \sum_{i=1}^{n} I(-X_i \le X_{(j+m)})\right)$$

$$\text{-}\max \frac{1}{n}\left(\sum_{i=1}^{n} I(X_i \le X_{(j-m)}), \sum_{i=1}^{n} I(-X_i \le X_{(j-m)})\right).$$

The proposed test rejects the null hypothesis iff $\log(V_n^*) > C$, where C is the test threshold. Similarly to the two-sided test' consideration mentioned above, the critical values of the one-side test can be calculated according to the following equation $P_{H0}\{\log(V_n^*) > C\} = P_{X_1,...,X_n \sim norm(0,1)}\{\log(V_n^*) > C\}$.

2.2 Two-sample Comparison

The two-sample DBEL ratio test has been dealt with extensively in the recent literature (e.g., Gurevich and Vexler (2011); Vexler and Yu, 2011; Vexler *et al*., 2012a; Miecznikowski *et al*., 2013). To outline this method, we suppose that data consists of two-samples of i.i.d. observations $X_1,...,X_n$ and $Y_1,...,Y_k$. We are interested in verifying that both samples are from the same distribution. That is, we want to test for $H_0 : F_x = F_y = F_z$ vs. $H_1 : F_x \ne F_y$, where the distributions F_x, F_y, and F_z of the observations are unknown. In this case, the most powerful LR statistic has the form of $\text{LR} = \prod_{i=1}^{n} \frac{f_{X,i}}{f_{ZX,i}} \prod_{j=1}^{k} \frac{f_{Y,j}}{f_{ZY,j}}$,

where $f_X(X_{(i)}) = f_{X,i}$, $f_Z(X_{(i)}) = f_{ZX,i}$, $f_Y(Y_{(j)}) = f_{Y,j}$, $X_{(i)}$ and $Y_{(j)}$ are the order statistics based on the observations $X_1,...,X_n$ and $Y_1,...,Y_k$, respectively.

To construct the corresponding nonparametric test we consider the likelihood $L_f = \prod_{i=1}^{n} f_{X,i}$. Towards this end we find values of $f_{X,i}$ that maximize L_f given an empirical version of the constraint $\int f_X(u)du = 1$. This is quite similar to the one-sample setting described in detail earlier. Define the H_0-empirical distribution function $F_{Z(n+k)}(u) = \left(\sum_{i=1}^{n} I(X_i \le u) + \sum_{j=1}^{k} I(Y_j \le u)\right)/(n+k)$. Following the work of Gurevich and Vexler (2011) one can show that by virtue of the hypotheses setting, $H_0 : F_x = F_y = F_z$ vs. $H_1 : F_x \neq F_y$, the empirical constraint $\tilde{\Delta}_m \le 1$, with $\tilde{\Delta}_m = (2m)^{-1}\sum_{i=1}^{n}\left(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})\right) f_{X,i} / f_{ZX,i}$, approximates the condition $\Delta_m = (1/2m)\sum_{i=1}^{n}\int_{X_{(i-m)}}^{X_{(i+m)}} f_Z(u) f_X(u) / f_Z(u)du \le 1$ that reflects the property $\int f_Z(u) f_X(u) / f_Z(u)du = \int f_X(u)du$ =1. Then the Lagrange multiplier method may be used to directly find values

$$f_{X,i} = \frac{2m \ f_{ZX,i}}{n\left(F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})\right)},\ i=1,\ldots,n,$$

that maximize L_f satisfying the empirical constraint $\tilde{\Delta}_m \le 1$. This implies that the EL estimator of $\prod_{i=1}^{n} f_{X,i} / f_{ZX,i}$ is given as

$$ELR_{X,m,n} = \prod_{i=1}^{n} \frac{2m}{n[F_{Z(n+k)}(X_{(i+m)}) - F_{Z(n+k)}(X_{(i-m)})]}.$$

Similarly, the maximum EL method forms the EL estimator of $\prod_{j=1}^{k} f_{Y,j} / f_{ZY,j}$ as

$$ELR_{Y,r,k} = \prod_{i=1}^{k} \frac{2r}{k[F_{Z(n+k)}(Y_{(i+r)}) - F_{Z(n+k)}(Y_{(i-r)})]},$$

where $Y_{(i)} = Y_{(1)}$, if $i \le 1$, and $Y_{(i)} = Y_{(k)}$, if $i \ge k$.

Thus, the proposed test statistic is

$$V_{nk} = ELR_{X,n} ELR_{Y,k}, \text{ where}$$

$ELR_{X,n} = \min_{a_n \le m \le b_n} ELR_{X,m,n}$, $ELR_{Y,k} = \min_{a_k \le r \le b_k} ELR_{Y,r,k}$, $a_l = l^{0.5+\delta}$, $b_l = \min(l^{1-\delta}, l/2)$, $\delta \in (0,1/4)$, and $l = n,k$.

Here, the operation "min" is utilized to provide the EL estimators of the parameters *m* and *r* in the expresssions $ELR_{X,m,n}$ and $ELR_{Y,r,k}$, respectively. The bounds a_l and b_l are defined to preserve the asymptotic consistency of the test, e.g. see Vexler and Yu (2010) for details. As before, in this paper we

set a value of $\delta = 0.1$. It is shown in Tsai *et al*. (2013), Vexler and Yu (2010), Vexler and Gurevich (2010), and Vexler *et al*. (2013b) that the power of the test statistic does not differ substantially for values of $\delta \in (0,1/4)$. Similar to what was previously derived in Gurevich and Vexler (2011) we define $F_{Z(n+k)}(x) - F_{Z(n+k)}(y) = 1/(n+k)$, if $F_{Z(n+k)}(x) = F_{Z(n+k)}(y)$.

The proposed test rejects the null hypothesis iff $\log(V_{nk})$>C, where C is the test threshold. In a similar manner to Section 2.1, it turns out that the null distribution of the test statistic V_{nk} is independent of the sample distributions. Therefore we have $P_{H_0}\{\log(V_{nk}) > C\} = P_{X_1,...,X_n,Y,...,Y_k \sim unif\,(0,1)}\{\log(V_{nk})>C\}$. Thus, this expression indicates that the probability of a Type I error of the test can be calculated exactly.

Now consider the one-sided version of the two sample problem (Gurevich, 2012). We are interested in testing the hypotheses $H_0 : F_x = F_y = F_z$ vs. $H_1 : F_x(u) \le F_y(u)$ for all $-\infty \le u \le \infty$, $F_x(u) < F_y(u)$ for some $-\infty \le u \le \infty$.

Applying the maximum BDEL concept one can define the test statistic

$$V_{nk}^{*} = ELR_{X,n}^{*} ELR_{Y,k}^{*}, \text{ where}$$

$$ELR_{X,n}^{*} = \min_{a_n \le m \le b_n} \prod_{i=1}^{n} \frac{2m}{n[F_Z^{*}(X_{(i+m)}) - F_Z^{*}(X_{(i-m)})]},\ ELR_{Y,k}^{*} = \min_{a_k \le r \le b_k} \prod_{i=1}^{k} \frac{2r}{k[F_Z^{**}(Y_{(i+r)}) - F_Z^{**}(Y_{(i-r)})]},$$

$F_Z^{*}(u) = \max\left(F_{X,n}(u), F_{Y,k}(u)\right)$ and $F_Z^{**}(u) = \min\left(F_{X,n}(u), F_{Y,k}(u)\right)$ with $F_{X,n}(u) = n^{-1}\sum_{i=1}^{n} I(X_i \le u)$ and $F_{Y,k}(u) = k^{-1}\sum_{j=1}^{k} I(Y_j \le u)$. Further, if $F_Z^{*}(X_{(i+m)}) - F_Z^{*}(X_{(i-m)})$ is less than zero then it is set to $(n+k)^{-1}$. Similarly, if $F_Z^{**}(Y_{(i+r)}) - F_Z^{**}(Y_{(i-r)}) \le 0$ then it is set to $(n+k)^{-1}$. Here, in comparing with the test V_{nk}, instead of $F_{Z(n+k)}$, the empirical distribution function (edf) of identically H_0-distributed *X* and *Y*, we use F_Z^{*} and F_Z^{**} to depict the fit between data distributions and our alternative hypothesis.

Similar to the two-sided test, the critical values can be calculated exactly according to the following equation: $P_{H_0}\{\log(V_{nk}^{*} > C\} = P_{X_1,...,X_n,Y,...,Y_k \sim norm\,(0,1)}\{\log(V_{nk}^{*})>C\}$.

2.3 K-sample Comparison

To outline the K-sample procedure we suppose that the data consists of K independent samples. We are interested in testing whether or not all K-samples are distributed identically in a nonparametric fashion. Let $n_1,\ldots,n_k$ denote the respective sample sizes corresponding to the K samples being compared. Assume that the K-samples are represented by the vectors of observations given as $\{X_{11},..., X_{1n_1}\},\ldots,$

$\{X_{k1},..., X_{kn_k}\}$ from the corresponding distribution functions $F_{X_1},\ldots, F_{X_k}$. We now want to test the hypothesis $H_0 : F_{X_1} = ... = F_{X_k} = F_Z$ vs. H_1 : not all $F_{Xi} = F_{Xj}$, $i \neq j$. If the corresponding density functions are known the LR statistic has the form

$$\text{LR} = \prod_{j=1}^{k}\prod_{i=1}^{n_j} f_{X_j}(X_{ji}) / f_Z(X_{ji}) = \prod_{j=1}^{k}\prod_{i=1}^{n_j} f_{X_j,i} / f_{ZX_j,i},$$

where f_{X_j} denotes the density function of the *j*-th sample under H_1, f_Z is the theoretical density function of observations under H_0 and $X_{j(i)}$, *j=1,…,k,* are the order statistics per sample based on the observations $X_{j1},..., X_{jn_j}$. As before we denote $f_{X_j}(X_{j(i)}) = f_{X_j,i}$ and $f_Z(X_{j(i)}) = f_{ZX_j,i}$, *j=1,…,k.*

We apply the maximum EL method to obtain the proposed DBEL ratio test statistic

$V_{n_1 n_2 \ldots n_k} = \prod_{j=1}^{k} ELR_{X_j,n_j}$, where the EL estimator of $\prod_{i=1}^{n_j} f_{X_j,i} / f_{ZX_j,i}$, for *j=1,…,k* is

$$ELR_{X_j,n_j} = \min_{a_{n_j} \le m_j \le b_{n_j}} \prod_{i=1}^{n_j} \frac{2m_j}{n_j[F_{Z(N)}(X_{j(i+m_j)}) - F_{Z(N)}(X_{j(i-m_j)})]}, a_l = l^{0.5+\delta}, b_l = \min(l^{1-\delta}, l/2), \delta \in (0,1/4)$$

and the corresponding edf under H_0 is given as

$$F_{Z(N)}(u) = \frac{1}{N}\sum_{j=1}^{k}\sum_{i=1}^{n_j} I\left(X_{ji} \le u\right), N = \sum_{j=1}^{k} n_j,\ X_{j(i)} = X_{j(1)}, \text{ if } i \le 1, \text{ and } X_{j(i)} = X_{j(n_j)}, \text{ if } i \ge n_j.$$

As in the two-group setting we define $F_{Z(N)}(x) - F_{Z(N)}(y) = 1/N$, if $F_{Z(N)}(x) = F_{Z(N)}(y)$.

The Type I error of the K-sample test can be monitored exactly by way of the following probability statement: $P_{H_0}\{\log(V_{n_1 n_2 \ldots n_k}) > C\} = P_{X_{11},\ldots X_{1n_1},\ldots,X_{k1},\ldots,X_{kn_k} \sim unif\,(0,1)}\{\log(V_{n_1 n_2 \ldots n_k}) > C\}$.

As above, we set a value of $\delta = 0.1$. The proposed test statistic approximates nonparametrically the most powerful parametric LR test statistic, consequently one can assume heuristically that the proposed test also has good relative efficiency.

3 Evaluations of critical values of the exact tests

In this section, we outline three methods to obtain critical values for exact tests proposed in this note. The MC, Interpolation, and Hybrid methods will be described in Sections 3.1, 3.2, and 3.3, respectively. These techniques are applied within a newly developed STATA command.

3.1 Monte Carlo Method

The MC method is a well-known approach for obtaining accurate approximations to the critical values (CVs) of exact tests (see, e.g., Metropolis and Ulam (1949) and Rubinstein (2008) for details).

The proposed STATA command provides the MC method option using 10,000 MC repetitions to calculate a critical value per each application of the command. In the developed procedures, the CVs are calculated by utilizing data generated from a standard normal distribution for one-sample tests and a Unif(0,1) distribution for the two-sample and three-sample tests. The generated values of *V*, the test statistic, are used to determine the critical value *C*, the test threshold, for the desired significance level $\alpha = P_{H_0}\{\log V > C_\alpha\}$ via calculating quantiles of the MC H_0-distribution of *logV*.

In addition, using the MC procedure we tabulated tables of CVs. Each critical value was calculated for each proposed test, each set of sample sizes, and each significance level. The MC procedure was based on 50,000 replicate samples of the corresponding test statistics *logV*. The resulting tables of CVs are to be used by the command VX_DBEL, a product of this package.

3.2 An approach based on interpolation

Interpolation differs from MC method in that tables of CVs were calculated beforehand for each proposed test, for various sample sizes, and various significance levels, α . Following the previous section, the resulting tables were obtained and saved in the STATA data format. Pearson and Hartley (1966) demonstrated the method of interpolation for calculating values within tables. For example, suppose we have data consisting of one-sample with size *n*, say *n=37*, to be tested for symmetry. The needed critical value can be interpolated based on tabulated CVs of *n=30, 35, 40 and 50*.

For our procedure, the tables with CVs are provided for sample sizes: the one-sample tests with sample size *n* from the set *{A}*, where *A={1,2,3,…29, 30, 35, 40, 50, 60, 80, 100, 150, 200}*; the two-sample tests with sample sizes *(m, n)* from the set *{AxA}*; the three-sample test with sample sizes *(m, n, p)* from the set *{AxAxA}*. The applications of the interpolation/extrapolation procedures decrease the accuracy of the estimates of the CVs, when actual samples have the sample sizes that differ significantly from those used to tabulate the tables. Providing the tables for use by the test increases the speed to execute the test compared with the MC method, but the interpolation method becomes less reliable.

In the context of the proposed procedures, in the cases of sample sizes that are not tabulated within the tables, an appropriate subset of the table data is selected. The data to be selected has been defined to be related to sample sizes available in the tables within a radius of two values around the values of the sample sizes that are needed. For example, if we are interested in the one-sample test based on a sample with size *n=78*, we utilize CVs related to *n=50, 60, 80, and 100* to estimate the required critical value. Based on extensive MC experiments, we conclude that to interpolate for needed CVs the radius of two values around the needed CVs was found to be best to minimize a bias of the

interpolation/extrapolation procedures. To outline this method in detail for the case of one-sample, we define $C_{n\alpha}$ to be the critical value corresponding to the simple size *n* and the significance level α. Using the selected table data, we fit $C_{n\alpha}$ via the regression model

$$C_{n\alpha} \cong \beta_0 + \beta_1 n + \beta_2 1/n + \beta_3 n^2 + \beta_4 \alpha + \beta_5 \alpha^{1/2},$$

employing the local maximum likelihood methodology (e.g., Fan *et al*., 1998). In this manner the coefficient $\beta' s$ are estimated yielding estimated values of $C_{n\alpha}$ as a function of *n* and α.

Similarly for the two- and three-sample tests, we define $C_{nm\alpha}$ and $C_{nmk\alpha}$ to be the CVs corresponding to the sample sizes *n*, *m*, *k*, and the significance level α, respectively. Using the selected table data, we fit $C_{nm\alpha}$ and $C_{nmk\alpha}$ via the regression models

$C_{nm\alpha} \cong \beta_0 + \beta_1 n + \beta_2 m + \beta_3 1/n + \beta_4 1/m + \beta_5 n^2 + \beta_6 m^2 + \beta_7 \alpha + \beta_8 \alpha^{1/2}$ and

$C_{nmk\alpha} \cong \beta_0 + \beta_1 n + \beta_2 m + \beta_3 k + \beta_4 1/n + \beta_5 1/m + \beta_6 1/k + \beta_7 n^2 + \beta_8 m^2 + \beta_9 k^2 + \beta_{10} \alpha + \beta_{11} \alpha^{1/2}$.

Here we assume $C_{n\alpha} = G_1(n, \alpha)$, $C_{nm\alpha} = G_2(n, m, \alpha)$ and $C_{nmk\alpha} = G_3(n, m, k, \alpha)$, where the functions G_1, G_2 and G_3 are unknown, but approximated via the equations shown above.

The regression equation with the parameter estimates is solved backwards for an estimate of the Type I error, α, using the equation solvers mm_root and optimize in STATA. This is accomplished by plugging in the value of the test statistic, *logV*, based on the observed data for the critical value and the sample sizes of the observed data into the regression equation, and solving for α, i.e. *logV*= $\hat{\beta}_0 + \hat{\beta}_1$ n + $\hat{\beta}_2$ 1/n+ $\hat{\beta}_3 n^2 + \hat{\beta}_4 \alpha + \hat{\beta}_5 \alpha^{1/2}$ should be solved with respect to α for the one-sample test. Thus, the value obtained for α is an estimate of the p-value for the test.

3.3 A novel hybrid technique that combines MC and Interpolation

This method combines the MC and interpolation methods depicted in Sections 3.1 and 3.2, respectively. Towards this end, we propose a nonparametric Bayesian type approach for constructing the posterior expectations of the needed CVs. Thereby, we incorporate the efficiency of the interpolation method and the accuracy of the MC method.

Lazar (2003) and Vexler *et al*. (2013) showed that the ELs can be utilized in Bayesian statistical inferences instead of the corresponding parametric likelihoods. This provides nonparametric Bayes procedures. This concept is applied in the proposed command to calculate the CVs. Since distributions of test statistic values are unknown, the likelihoods are presented in the EL form.

The proposed algorithm is conducted in two stages, and repeated until a stopping condition is met. The following notations are used in the description of the procedures: t_k denotes the number of MC simulations related to stage *k=1,2*; α : the level of significance (we use α =0.05 as default in the command); $T_{(1)}^k < T_{(2)}^k < ... < T_{(t_k)}^k$: the order statistics based on the test statistic values $T_1^k, T_2^k, ..., T_{t_k}^k$, generated on stage *k* (*T=logV*); J_k : the interval $\left[\max\left\{1,(1-\alpha)t_k - t_k^{1/2}\log(t_k)\right\}, \min\left\{(1-\alpha)t_k + t_k^{1/2}\log(t_k), t_k\right\}\right]$; $L_k(q) = \exp\left[t_k F_{k,t_k}(q)\left\{\log(1-\alpha) - \log(F_{k,t_k}(q))\right\} + t_k\left\{1-F_{k,t_k}(q)\right\}\left\{\log\alpha - \log(1-F_{k,t_k}(q))\right\}\right]$: the EL function = $\max\left\{\prod_{i=1}^{t_k} p_i : \sum_{i=1}^{t_k} p_i = 1, \sum_{i=1}^{t_k} p_i I(T_i < q) = \alpha\right\}$, where $F_{k,t_k}(q) = \sum_{i=1}^{t_k} I(T_{(i)}^k < q)/t_k$; the quantile q_α $P_{H_0}\{\log V > q_\alpha\} = \alpha$ defines the needed critical value of the test statistic $\log V$.

We begin by fitting the prior information with a functional form. The tabulated values are assumed to provide prior information regarding the target CVs. We obtain the prior distribution function $\pi(q)$, with the parameters (μ_0, σ_0), using the local maximum likelihood method (Fan *et al*., 1998) based on tabulated CVs. Here,

$$\pi(q) = (2\pi\sigma_0)^{-0.5}\exp\left\{-(q-\mu_0)^2\left(2\sigma_0{}^2\right)^{-1}\right\}.$$

The normal function form of $\pi(q)$ was utilized, since quantile estimators are commonly normally distributed when sample sizes are relatively large. For example, in the context of the two sample test, following Section 3.2, we can present $\mu_0 = \hat{\beta}_0 n_0 + \hat{\beta}_1 m_0 + \hat{\beta}_2 1/n_0 + \hat{\beta}_3 1/m_0 + \hat{\beta}_4 n_0{}^2 + \hat{\beta}_5 m_0^2 + \hat{\beta}_6 \alpha$, where it is assumed that the observed data consists of two samples of sizes n_0 and m_0 . In this case, the σ_0 can be estimated using the standard regression analysis.

In the first MC simulation step, 200 generations $(t_1 = 200)$ of test statistic values are proposed to be conducted as a first stage, and then test statistics $T^1 = (T_1^1, ..., T_{t_1}^1)$ based on the generated data are calculate. Next, the posterior expectation, $\hat{q}_{1,\alpha}$, of the quantile q_α is calculated. We compute the posterior expectation of quantiles as follows:

$$\hat{q}_{k,\alpha} = \frac{\int_{T_{(1)}^1}^{T_{(t_1)}^k} qL_k(q)\pi(q)dq}{\int_{T_{(1)}^k}^{T_{(t_1)}^k} L_k(q)\pi(q)dq},\ k=1$$

and then one can show that

$$\hat{q}_{1,\alpha} = \frac{\sum_{j\in J_1} \exp\{-t_1(1-\alpha-\frac{j}{t_1})^2(2\alpha(1-\alpha))^{-1}\}\int_{T^c_{(j-1)}}^{T^1_{(j)}} q\pi(q)dq}{\sum_{j\in J_1} \exp\{-t_1(1-\alpha-\frac{j}{t_1})^2(2\alpha(1-\alpha))^{-1}\}\int_{T^1_{(j-1)}}^{T^1_{(j)}} \pi(q)dq},$$

where, in general, $\int_a^b q\pi(q)dq = \sqrt{\sigma_0{}^2/2\pi}\left[\exp\left\{-(a-\mu_0)^2/(2\sigma_0{}^2)\right\} - \exp\left\{-(b-\mu_0)^2/(2\sigma_0{}^2)\right\}\right] +$ $\mu_0\int_a^b \pi(q)dq$ and $\int_a^b \pi(q)dq$ = normal($(b-\mu_0)/\sigma_0$) - normal($(a-\mu_0)/\sigma_0$). In this equation, the STATA Normal function, *normal*, returns the cumulative standard normal distribution.

During the second stage, 200 additional generations $(t_2 = 200)$ of the test statistic values are proposed to be conducted and then test statistics $T^2 = (T_1^2,...,T_{t_2}^2)$ based on simulated data are calculated. We then estimate $f_{2,t_2}(\hat{q}_{1,\alpha})$, an estimate of density function of the test statistic, using the following kernel estimator (see, e.g., Gibbons and Chakraborti (2005) for details)

$$f_{2,t_2}(\hat{q}_{1,\alpha}) = (t_2)^{-1}\sum_{j=1}^{t_2}\left(2\pi h^2\right)^{-1/2}\exp\left\{-\left(\hat{q}_{1,\alpha}-T_j^2\right)^2\left(2h^2\right)^{-1}\right\},$$

where $h = 1.06\hat{\sigma}_{t_2}t^{-1/5}$, and $\hat{\sigma}_{t_2}$ is the standard deviation of the test statistics T_j^2. Then, we calculate the estimated variance of $\hat{q}_{1,\alpha}$, as $V_k \equiv F_{k+1,t_{k+1}}(\hat{q}_{k,\alpha})\{1-F_{k+1,t_{k+1}}(\hat{q}_{k,\alpha})\}f_{k+1,t_{k+1}}(\hat{q}_{k,\alpha})^{-2}/t_{k+1},\ k=1$. (For details of this approximation to the variance see Serfling, 1980.)

To define the stopping rule of the procedure we evaluate if $V_1 \le \sigma_0^2$, then we stop the procedure and calculate, based on the combined values of the test statistics $T_c = (T_1^1,...,T_{t_1}^1,T_1^2,...,T_{t_2}^2)$ and $t_c = t_1 + t_2$, the posterior expectation of the quantiles

$$\hat{q}_{c,\alpha} = \frac{\int_{T^c_{(1)}}^{T^c_{(t_c)}} qL_c(q)\pi(q)dq}{\int_{T^c_{(1)}}^{T^c_{(t_c)}} L_c(q)\pi(q)dq} \cong \frac{\sum_{j\in J_c}\exp\{-t_c(1-\alpha-\frac{j}{t_c})^2/(2\alpha(1-\alpha))\}\int_{T^c_{(j-1)}}^{T^c_{(j)}} q\pi(q)dq}{\sum_{j\in J_c}\exp\{-t_c(1-\alpha-\frac{j}{t_c})^2/(2\alpha(1-\alpha))\}\int_{T^c_{(j-1)}}^{T^c_{(j)}} \pi(q)dq}.$$

In this case, we reach the estimated value of the variance of the CVs comparable with the variance of the CVs found in the table. If the value of the test statistic based on the data, T_0, is greater than $\hat{q}_{c,\alpha}$ then we reject the null.

If $V_1 > \sigma_0^2$, combine T^1 and T^2 into a new T^1, so that the new t_1 is equal to $t_1 + t_2$. Repeat stages one and two of the procedure until the stop condition ($V_1 \le \sigma_0^2$) is reached or the number of the new combined values of test statistics, T_1, is greater than 35,000.

In summary, the algorithm is performed iteratively, and the decision to perform another iteration of the scheme is based on a comparison of the variance estimator V_k, and the parameter σ_0^2.

3 VX_DBEL Command

3.1 Description

In this paper we present a STATA implementation of the DBEL ratio test in the command vx**_dbel**. The new commands vx**_dbel** and **help vx_dbel** are freely available for download at the Department of Biostatistics website for the University at Buffalo. The vx**_dbel** command conducts a one-sample test of symmetry, two-sample comparison, or three-sample comparison as described above. The command output presents applicable test statistic and p-value.

3.2 Syntax

The syntax of the command to execute the test of symmetry and K-sample Distribution Free DBEL Ratio Test is

vx_dbel [varlist] [, options],

where *varlist* specifies up to three variable names.

3.3 Options

Sided() is required. It specifies the test as one-sided or two sided. Sided("greater") and sided("less") performs a one sided test. Sided("two.sided") conduct a two-sided test.

Method() is required. It specifies the test method as Monte Carlo, interpolation, or hybrid. Method("mc") defines the method as Monte Carlo, as described in Section 3.1 Method("interpolation") defines the method as interpolation, as described in Section 3.2. Method("hybrid") defines the method as hybrid, as described as in Section 3.3.

repsnum(#) specifies the number of Monte Carlo iterations to obtain the critical value of the test in the Monte Carlo method. The default is repsnum(10000).

alpha(#) specifies significance level for the hybrid method. The default is alpha(0.05), meaning that a significance level of 0.05 is used for alpha in the hybrid method.

3.4 Remarks

Three STATA tables are provided with the command. For the *mc* method, the critical values are obtained by Monte Carlo simulations for each test conducted as described in Section 3.1. For the *interpolation* method, the provided tables of CVs are used to obtain the p-value as described in Section 3.2. For the hybrid method, which is based on Bayesian approach, the provided tables of CVs are used to obtain initial parameters estimates (μ_0, σ_0) as described in Section 3.3.

Further, assuming the method option is set to interpolation or hybrid. It is noted that the provided tables of CVs are complete for sample sizes of less than or equal to thirty, but are incomplete for sample sizes greater than thirty. If all the sample sizes are less than thirty then the CVs are directly available in the tables. In this case, the CVs are directly obtained from the appropriate table. If all sample sizes are not less than thirty then the CVs are not directly available in the tables. In this case, the hybrid and interpolation methods are utilized to obtain the desired CVs.

4 Simulations

In this section, we examine the power of the proposed tests. In particular, we compare the power of our tests with that of the classical tests, and evaluate the new hybrid method against the Monte Carlo and interpolation methods. Simulations were conducted for 10,000 repetitions for all tests with varying sample sizes, at a significance level of 0.05. In this paper, in the interest of economy of space we present just a part of the obtained Monte Carlo results to show the general picture of the evaluations. The power of the proposed tests was simulated under various alternatives. To describe the MC simulations we will use the following abbreviations: N(μ, σ) is the normal distribution with mean μ and standard deviation σ ; u(a,b) is the uniform distribution between a and b; e(β) is the exponential distribution with parameter β ; $\log N(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ .

To complete Table 1, we compare the power of the proposed one-sample tests (Prop) with the Wilcoxon test (W). In accordance with the statistical literature, in the case of normally distributed data and the change of location the W is known to be close to an optimal test. We can see that our tests provide the power that is not significantly different than that of W. However, in the cases that are different than the above case, the proposed tests are superior to the W test. For example, when the two-sided test is based on u(-1,0.8) distributed data with *n*=40, the proposed test provides power of 0.397 vs. the W test power of 0.199. An increase of 50% in observed power.

In Tables 2a, 2b and 3, the power of the proposed two-sample tests are compared to that of the W and Kolmogorov-Smirnov (KS) tests. In accordance with the statistical literature, in the case of normally distributed data and change of location the W is very powerful. In such cases, we observed that our tests provided the powers that are not significantly different than those of W. However, in the cases that are different than the above case, the proposed tests are superior to the W and KS test. For example, when the two-sided test is based on N(1,1) and N(0,2) distributed data with sample sizes of 10 and 25, respectively. The power of the proposed test was significantly higher than that of W and KS. The proposed, W, and KS tests powers were 0.66, 0.32, and 0.37, respectively. An increase of 52% and 53%

in observed power over W and KS, respectively. Similar results were seen for the two-sided test with sample distributions u(1,2) and logN(0,1). The powers of the proposed test at sample sizes of 10 and 20 were 0.99, 0.23, and 0.54 for our proposed, W, and KS tests, respectively. An increase of 76% and 45% in observed powers over W and KS, respectively. The same situations were observed, but not shown, when the test is based on two exponentially distributed data or two Cauchy distributed data.

--- Tables 1, 2a, 2b and 3 ---

To complete Table 4, we compared the power of the proposed three-sample test with the ANOVA F-test and Kruskal-Wallis (KW) test. In accordance with the statistical literature, in the case of normally distributed data with the change of location the F-test is known to be close to an optimal test. We observed that our tests provide the power that is not significantly different than that of the F-test. However, in the cases that are different than the above case, the proposed tests are superior to the F-test. For example, when the test is based on N(0,1), N(0,1) and u(-1,1) distributed data with sample sizes of 37, 45 and 50, respectively, the proposed test provides power of 0.98 vs. the F-test power of 0.061 and KW test power of 0.057. An increase of 94% in observed power. The proposed three-sample test outperformed the F and KW tests under all the tested alternatives.

--- Table 4 ---

To analyze the methods for calculating critical values of exact tests mentioned in Sections 3.1, 3.2 and 3.3, we conducted the following MC. In Table 5, the probability of a Type I errors for the proposed two-sample test were compared for the Monte Carlo, interpolation, and hybrid method. The two-sample test was based on N(0,1) and N(0,1) distributed data samples. The Type I errors for the test was appropriately controlled at 5% for all three methods. For example, when the test is based on sample sizes of 35 and 37, the actual Type I errors were 0.048, 0.0526 and 0.0526, for the Monte Carlo, interpolation, and hybrid method, respectively. The Type I errors for the proposed three-sample test were compared for the Monte Carlo, interpolation, and hybrid method in Table 6. The three-sample test was based on sample distributions of N(0,1), N(0,1) and N(0,1). The Type I errors for the test were appropriately controlled at 5% for all three methods. The hybrid and MC methods are comparable, but the hybrid method on average utilizes five times less Monte Carlo repetitions.

--- Tables 5 and 6---

The interpolation and hybrid methods make use of CVs tables as mentioned in Sections 3.1-3.3.

5 Application

In this section, the proposed three-sample comparisons test is illustrated via a dataset of blood test results for patients with anemia (Wians *et al*., 2001). A total of 134 patients with anemia underwent

a series of blood tests. To eliminate the bias which might be caused by gender, the analysis was limited to the 55 female study patients. Ferritin concentration provides a useful screening test for iron deficiency anemia (IDA). Non-pregnant women with anemia and a ferritin concentration less than 20 $\mu g/l$ were assigned to the IDA group, while those with anemia and a ferritin concentration greater than 240 $\mu g/l$ were assigned to be the anemia of chronic disease (ACD) group. The intermediate group consists of the women with ferritin concentration between 20 and 240 $\mu g/l$. There were 12, 14, 29 female study subjects in ACD, intermediate and IDA groups, respectively. We are interested in comparison of sample distributions of the two rapid blood tests, i.e. total iron binding capacity (TIBC) and percent transferrin saturation (%TS), for discriminating between the ACD, intermediate and IDA groups (Tian *et al*., 2011).

Tian *et al*. (2011) focuses on the confidence interval estimation of the differences in paired volumes under surfaces (VUS) and paired partial volumes under surfaces. The 95% confidence intervals for Tian *et al*. approach were (0.1103, 0.5139) for ΔVUS and (0.038, 0.1515) for $\Delta PVUS$. Both confidence intervals showed that TIBC had better diagnostic ability than %TS. The approaches assumption of multivariate normality for each group was tested and not rejected.

We use **vx_dbel** command to conduct three-sample comparison tests, comparing the three groups within the two rapid blood tests. In addition, we use **vx_dbel** command to conduct two-sample comparison tests, comparing each group between the two rapid blood tests

First, we conduct the three-sample comparison of the three groups for the TIBC. The test results indicate that the distributions of the three groups are not equal.

vx_dbel IDA Intermediate ACD, sided("three.sample") method("mc")

	T0	P-value
DBEL (3-sample)	25.153864	0.003

Similarly, we conduct the three-sample comparison of the three groups for the %TS. The test results indicate that the distributions of the three groups are not equal.

vx_dbel IDA Intermediate ACD, sided("three.sample") method("mc")

	T0	P-value
DBEL (3-sample)	44.255267	<0.001

Thus, the proposed three sample test is able to discriminate between the three groups for both the TIBC and %TS blood tests.

In addition, we conducted two-sample comparison of TIBC vs. %TS for each group. We note that the %TS and TIBC result values have completely different scales. The %TS results range within single digits and the TIBC ranges within hundreds, so that a standardization transformation is applied to put the result of the blood tests in comparable ranges. The **vx_dbel** outputs are not presented for brevity. The p-values for the three two-sample tests were 0.4556, 0.9647 and 0.0939 for the IDA, intermediate and ACD groups, respectively. The test results suggest that the distributions between TIBC and %TS for each group are equal.

In conducting clinical experiments, it is often desired to compare populations receiving different treatments to establish equivalence, or test identical distributions assumptions of various statistical tests. Thus, the proposed STATA command may be used in such experiments without assumptions on data distributions (Tian *et al*., 2011 assumed normal distributions of observations). In general, there is excellent applicability for discriminating between samples in any context.

6 Concluding Remarks

In this article, we proposed and examined the one-sample symmetry and K-sample comparisons DBEL ratio tests. The proposed tests are shown to be exact and robust nonparametric tests that approximate the optimal LR test statistic. The powers of the proposed tests were comparable and in many cases outperformed the classical tests.

To date, simple DBEL tests have not been presented in the STATA package, but are known to be very efficient in practice. We developed and presented a STATA package to perform the discussed approaches. The data example was used to demonstrate that it is simple to use. The STATA command performs both one-sided and two-sided alternatives for one-sample symmetry and two-sample comparison tests, and a three-sample comparison test. Three methods can be performed by the package, Monte Carlo for best accuracy, and Interpolation and Hybrid methods for improved speed. The STATA package is already freely available for download at the Department of Biostatistics website for the University at Buffalo. The material can be found under the Research and Facilities tab. The web address is sphhp.buffalo.edu/biostatistics/research-and-facilities/software/stata.html.

The Monte Carlo study performed in this article confirmed powerful properties of the proposed tests. We demonstrated that our one and two sample tests outperform the classic nonparametric tests of Wilcoxon and KS, and the three-sample test outperforms the F-test and KW test under various alternatives. Further, this is accomplished while appropriately controlling the Type I error for all the tests. The data example shows that the proposed test can be easily and efficiently used by practitioners.

Acknowledgement

This research is supported by the NIH grant 1R03DE020851-01A1 (the National Institute of Dental and Craniofacial research). The authors are grateful to the Editor and the referee for suggestions that led to a substantial improvement in this paper and the proposed STATA package.

References

Fan J, Farmen M, Gijbels I (1998). Local maximum likelihood estimation and inference. J. R. Stat. Soc. Ser. B 60, pp. 591–608.

Gibbons JD, Chakraborti S. Nonparametric Statistical Inference 4th ed. Marcel Dekker, Inc.: New York 2005.

Gurevich G (2012). Two-sample density-based empirical likelihood tests for stochastically ordered alternatives. The 6th International Days of Statistics and Economics, Prague, September 13-15, 2012.

Gurevich G, Vexler A (2011). A two-sample empirical likelihood test based on samples entropy. Statistics and Computing, Volume 21, Number 4, pp 657-670.

Lazar NA (2003). Bayesian Empirical Likelihood. Biometrika 90, 319-326.

Lehmann EL, Romano JP (2005). Testing statistical hypotheses. Springer, New York.

Metropolis N, Ulam S (1949). The Monte Carlo Method. Journal of the American Statistical Association, Volume 44, Number 247, pp 335-341.

Miecznikowski JC, Vexler A, Shepherd L (2013). dbEmplikeGOF: An R package for nonparametric likelihood ratio tests for goodness-of-fit and two-sample comparisons based on sample entropy. Journal of Statistical Software in press.

Obuchowski, N., (2006). An ROC-type measure of diagnostic accuracy when the gold standard is continuous-scale. Statistics in Medicine, Volume 25, 481–493.

Owen AB. Empirical Likelihood. Chapman and Hall/CRC: New York, 2001

Pearson K (1900). On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag Ser 5 50: 157–172.

Pearson E, Hartley H (1966). Biometrika Tables for Statisticians Volume I. Cambridge: New York 1966

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0 (htpp://www.R-project.org).

Rubenstein R, Kroese D. Simulation and the Monte Carlo Method 2nd ed. John Wiley & Sons, Inc.: New Jersey 2008

Serfling RJ. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, Inc.: New Jersey 1980

Tian L, Xiong C, Lai C, Vexler A (2011). Exact confidence interval estimation for the difference in diagnostic accuracy with three ordinal diagnostic groups. Journal of Statistical Planning and Inference 2011, Volume 141, pp 549–558.

Tsai W, Vexler A, Gurevich G (2013). An extensive power evaluation of a novel two-sample density-based empirical likelihood ratio test for paired data with an application to a treatment study of attention-deficit/hyperactivity disorder and severe mood dysregulation. Journal of Applied Statistics 2013, Volume 40, Issue 6, pp 1189-1208

Vexler A, Deng W, Wilding GE (2013a). Nonparametric Bayes factors based on empirical likelihood ratios. Journal of Statistical Planning and Inference, Volume 143, pp 611-620.

Vexler A, Gurevich G (2010). Empirical likelihood ratios applied to goodness-of-fit tests based on sample entropy. Computational Statistics and Data Analysis, Volume 54, pp 531-545.

Vexler A, Gurevich G (2011). A note on optimality of hypothesis testing. Journal Mesa, Volume 2, Number 3, pp 243-250.

Vexler A, Gurevich G, Hutson AD (2013b). An exact density-based empirical likelihood ratio test for paired data. Journal of Statistical Planning and Inference, Volume 143, pp 334-345.

Vexler A, Tsai W, Gurevich G, Yu J (2012a). Two-sample density-based empirical likelihood ratio tests based on paired data, with application to a treatment study of attention-deficit/hyperactivity disorder and severe mood dysregulation. Statistics in Medicine, Volume 31, 1821-1837.

Vexler A, Tsai W, Malinovsky Y (2012b). Estimation and testing based on data subject to measurement errors: from parametric to non-parametric likelihood methods. Statistics in Medicine, Volume 31, 2498-2512.

Vexler A, Wu C, Yu KF (2010). Optimal hypothesis testing: from semi to fully Bayes factors. Metrika , Volume 71, Issue 2, pp 125-138.

Vexler A, Yu J (2010). Two-sample density-based empirical likelihood tests for incomplete data in application to a pneumonia study. Biometrical Journal, Journal 52, Number 3, pp 348-361.

Vexler A, Yu J, Kim S, Hutson A (2011). Two-sample empirical likelihood ratio tests for medians in application to biomarker evaluation. The Canadian Journal of Statistics, Volume 39, Number 4, pp 671-689.

Wians FH, Urban JE, Keffer JH, Kroft SH (2001). Discriminating between iron deficiency anemia and anemia of chronic disease using traditional indices of iron status vs transferrin receptor concentration. American Journal of Clinical Hematopathology, Volume 115, 112–118.

Table 1: Monte Carlo power comparisons between the proposed and Wilcoxon (W) one-sample tests

	two-sided				one-sided (greater)							
	N(1,1)		u(-1,0.8)		N(0.5,1)		u(-0.3,0.7)		u(-0.8,1)		u(-1,1)	
n	Prop	W	Prop	W	Prop	W	Prop	W	Prop	W	Prop	W
5	0.4179	<0.001	0.0734	<0.0001	0.2782	0.1592	0.3272	0.1638	0.1003	0.0557	0.0661	0.0302
10	0.7871	0.7833	0.0805	0.0716	0.4190	0.3846	0.5740	0.5301	0.1272	0.1076	0.0481	0.0418
15	0.9373	0.9356	0.1176	0.0953	0.5614	0.5500	0.8134	0.7413	0.1730	0.1535	0.0551	0.0454
25	0.9959	0.9969	0.2056	0.1339	0.7433	0.7595	0.9800	0.9170	0.2739	0.2129	0.0463	0.0486
35	>0.999	>0.999	0.3219	0.1837	0.8587	0.8802	0.9993	0.9763	0.4178	0.2667	0.0501	0.0457
40	>0.999	>0.999	0.3965	0.1997	0.8943	0.9109	>0.999	0.9888	0.4888	0.3030	0.0492	0.0476
45	>0.999	>0.999	0.4621	0.2231	0.9198	0.9425	>0.999	0.9945	0.5499	0.3274	0.0508	0.0496

Table 2a: The Monte Carlo power comparisons between the proposed, Wilcoxon (W), and Kolmogorov-Smirnov (KS) two-sample two-sided tests

Design		N(1,1) & N(0,0.5)			N(1,1) & N(0,2)			u(-1,1)+e(1) & u(-1,1)		
n	m	Proposed	W	KS	Proposed	W	KS	Proposed	W	KS
10	5	0.5872	0.4967	0.3542	0.198	0.1878	0.1384	0.3939	0.3852	0.2084
	10	0.7906	0.7045	0.5221	0.3792	0.2477	0.1332	0.6565	0.595	0.2954
	15	0.8223	0.7935	0.805	0.5445	0.2931	0.3716	0.7224	0.7057	0.5983
	20	0.8663	0.8255	0.7809	0.5864	0.3174	0.3095	0.7933	0.7554	0.5547
	25	0.8502	0.8577	0.8632	0.6603	0.3257	0.3727	0.7866	0.7844	0.7028
	30	0.8613	0.8623	0.8605	0.7065	0.3485	0.3943	0.8083	0.7997	0.6898
	35	0.8657	0.8771	0.8929	0.7275	0.3589	0.4476	0.8136	0.8155	0.729
15	5	0.641	0.5672	0.4629	0.2039	0.2205	0.1874	0.4062	0.4534	0.2628
	15	0.9417	0.8772	0.8527	0.6125	0.3604	0.3507	0.8659	0.8101	0.6292
	20	0.968	0.9083	0.9334	0.6852	0.4113	0.5209	0.9241	0.8542	0.79
	25	0.9671	0.9345	0.9532	0.7707	0.4505	0.5606	0.9246	0.8889	0.8161
	30	0.9699	0.9426	0.9529	0.8181	0.4657	0.5659	0.9425	0.9064	0.8088
	35	0.9721	0.9489	0.9616	0.8492	0.4892	0.6014	0.9413	0.9144	0.8298
20	5	0.6805	0.6162	0.6556	0.2152	0.2542	0.2477	0.4333	0.4832	0.41
	20	0.9815	0.9596	0.9611	0.7655	0.4746	0.5153	0.974	0.9148	0.8253
	25	0.9862	0.9675	0.977	0.8629	0.506	0.6271	0.9783	0.9324	0.897
	30	0.9908	0.9729	0.9848	0.9051	0.5498	0.6837	0.9849	0.9425	0.9135
	35	0.99	0.9809	0.9887	0.9361	0.5734	0.7234	0.9876	0.9567	0.9198
30	5	0.7258	0.7084	0.6853	0.2084	0.2907	0.2537	0.4472	0.5644	0.4123
	25	0.9987	0.9928	0.9972	0.893	0.5939	0.7415	0.9975	0.9771	0.9587
	30	0.9991	0.9949	0.9963	0.9344	0.6505	0.7492	0.9989	0.9824	0.9598
	35	0.9996	0.9965	0.9983	0.9628	0.6846	0.8295	0.999	0.9875	0.9817

Table 2b: The Monte Carlo power comparisons between the proposed, Wilcoxon (W), and Kolmogorov-Smirnov (KS) two-sample two-sided tests

Design		e(1)+u(0,0.5) & e(1)			u(1,2) & logN(0,1)		
n	m	Proposed	W	KS	Proposed	W	KS

10	5	0.0944	0.0936	0.0466	0.2835	0.1614	0.2617
	10	0.1360	0.1056	0.0366	0.8104	0.2149	0.2450
	15	0.1810	0.1314	0.1251	0.9571	0.2175	0.6398
	20	0.1859	0.1424	0.0869	0.9934	0.2351	0.5446
	25	0.1961	0.1437	0.1042	0.9971	0.2380	0.6575
	30	0.2028	0.1459	0.1043	0.9996	0.2392	0.7225
	35	0.2207	0.1522	0.1184	0.9998	0.2249	0.7686
15	5	0.0898	0.1116	0.0662	0.2930	0.1783	0.2806
	15	0.1885	0.1618	0.0931	0.9859	0.2877	0.6142
	20	0.2189	0.1728	0.1546	0.9994	0.3096	0.8510
	25	0.2450	0.1894	0.1671	0.9997	0.3340	0.8792
	30	0.2793	0.1935	0.1537	>0.9999	0.3484	0.8976
	35	0.2826	0.2041	0.1693	0.9999	0.3490	0.9218
20	5	0.0889	0.1208	0.1063	0.3168	0.2857	0.3235
	20	0.2671	0.2054	0.1414	0.9998	0.3801	0.8458
	25	0.3284	0.2225	0.1811	>0.9999	0.3980	0.9511
	30	0.3649	0.2283	0.2111	>0.9999	0.4134	0.9639
	35	0.3842	0.2452	0.2147	>0.9999	0.4377	0.9854
30	5	0.0841	0.1454	0.1044	0.3059	0.3244	0.3256
	25	0.3479	0.2715	0.2471	>0.9999	0.4773	0.9851
	30	0.3762	0.2866	0.2255	>0.9999	0.5066	0.9859
	35	0.4133	0.318	0.2906	>0.9999	0.5456	0.9950

Table 3: The Monte Carlo power comparisons between the proposed, Wilcoxon (W), and Kolmogorov-Smirnov (KS) two-sample one-sided tests (greater)

Design		N(1,1) & N(0,0.5)			N(1,1) & N(0,2)			u(-1,1)+e(1) & u(-1,1)			e(1)+u(0,0.5) & e(1)		
n	m	Proposed	W	KS	Proposed	W	KS	Proposed	W	KS	Proposed	W	KS
10	5	0.7073	0.6792	0.5926	0.3756	0.3085	0.233	0.5236	0.5866	0.4055	0.1695	0.1749	0.0991
	10	0.9014	0.8169	0.7506	0.4183	0.3572	0.3111	0.7563	0.7295	0.5594	0.1843	0.188	0.1147
	15	0.9477	0.8618	0.8787	0.4529	0.3984	0.4295	0.865	0.8053	0.7326	0.1892	0.208	0.1705
	20	0.9665	0.9015	0.878	0.4724	0.4603	0.4471	0.8993	0.8431	0.724	0.1939	0.2327	0.1475
	30	0.9798	0.9201	0.9096	0.5106	0.485	0.5062	0.9554	0.8727	0.7641	0.1956	0.2267	0.1581
	35	0.9825	0.9195	0.9216	0.5252	0.5114	0.547	0.953	0.8827	0.7809	0.2185	0.2468	0.1746
15	5	0.793	0.7453	0.6346	0.4179	0.3436	0.2377	0.6235	0.6438	0.4103	0.1896	0.1907	0.1023
	10	0.944	0.8921	0.8869	0.5336	0.4175	0.4685	0.8371	0.8197	0.7054	0.241	0.2138	0.1779
	15	0.9809	0.9396	0.9394	0.5812	0.5043	0.5384	0.9277	0.8855	0.8048	0.2569	0.2534	0.1901
	20	0.9913	0.9551	0.9461	0.6344	0.5368	0.5609	0.9635	0.9227	0.8117	0.262	0.2795	0.1846
	30	0.995	0.9705	0.9762	0.6822	0.61	0.6831	0.9859	0.9465	0.8915	0.2758	0.3009	0.2399
20	5	0.8151	0.7891	0.651	0.4648	0.3599	0.2418	0.6393	0.6741	0.4181	0.2132	0.1935	0.1016
	10	0.9643	0.9396	0.9084	0.611	0.4687	0.4584	0.8848	0.8715	0.7349	0.2695	0.2486	0.1733
	15	0.9907	0.9703	0.9596	0.6739	0.5438	0.5417	0.9576	0.9289	0.8374	0.3174	0.2942	0.1877
	20	0.9964	0.9793	0.9827	0.7428	0.5967	0.6869	0.9809	0.9582	0.9206	0.3125	0.3027	0.2574
	30	0.999	0.9889	0.9899	0.7941	0.6715	0.7514	0.9945	0.9749	0.9317	0.345	0.3424	0.2614
	35	0.9994	0.9925	0.9942	0.8263	0.7062	0.7766	0.997	0.978	0.9554	0.3646	0.3655	0.2766
30	5	0.8424	0.8464	0.7714	0.5489	0.3691	0.3356	0.6866	0.7097	0.5197	0.257	0.2054	0.141

15	0.9975	0.9907	0.9901	0.7902	0.601	0.6571	0.9837	0.9627	0.9289	0.3809	0.3297	0.2516
20	0.9992	0.9948	0.996	0.8511	0.6576	0.7265	0.9958	0.9818	0.9611	0.4251	0.353	0.2712
30	0.9999	0.9987	0.9988	0.9104	0.7589	0.8426	0.9994	0.9923	0.9851	0.4734	0.4074	0.3358
35	0.9999	0.9988	0.9992	0.9248	0.7869	0.883	0.9995	0.995	0.9877	0.5068	0.4242	0.3717

Table 4: The Monte Carlo power comparisons between the proposed, F and Kruskal-Wallis (KW) three-sample tests

distributions	n	m	k	Proposed	F	KW
N(0,1)x2: u(-1,1)	25	15	20	0.3745	0.0573	0.0526
	25	25	25	0.5578	0.0529	0.0521
	37	45	50	0.9809	0.0610	0.0577
e(1)x2: logN(0,1)	25	15	20	0.2313	0.2306	0.2386
	37	45	50	0.6023	0.5481	0.5241
N(0,1)x2: N(0,2.25)	25	15	20	0.5908	0.0685	0.0599
	25	25	25	0.7719	0.0624	0.0576
	37	45	50	0.9941	0.0497	0.0544
N(0,1)x2: N(0,0.25)	25	15	20	0.9950	0.0610	0.0664
	25	25	25	0.9998	0.0584	0.0651
e(1):logN(0,1): u(0,1)	25	15	20	0.9645	0.8004	0.6745
	25	25	25	0.9980	0.9121	0.8311

Table 5: The Monte Carlo Type I error comparisons between proposed methods for two-sample test

Design	n	m	Monte Carlo	Interpolation	Hybrid
N(0,1)x2	32	35	0.0538	0.0520	0.0472
		40	0.0533	0.0583	0.0513
		50	0.0519	0.0446	0.0498
	35	37	0.0483	0.0526	0.0526
		50	0.0476	0.0327	0.0488
	37	50	0.0513	0.0467	0.0526
	40	50	0.0469	0.0351	0.0498

Table 6: The Monte Carlo Type I error comparisons between proposed methods for three-sample test

Design	n	m	k	Monte Carlo	Interpolation	Hybrid
N(0,1)x3	32	40	50	0.0493	0.0454	0.0519
	35	35	35	0.0546	0.0415	0.0501
	50	37	40	0.0477	0.0441	0.0466