1-27-2016

A Test of Symmetry Based on the Kernel Kullback-Leibler Information with Application to Base Deficit Data

Hani M. Samawi
Georgia Southern University, hsamawi@georgiasouthern.edu

Robert L. Vogel
Georgia Southern University, rvogel@georgiasouthern.edu

MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

A Test of Symmetry Based on the Kernel Kullback-Leibler Information with Application to Base Deficit Data

Abstract

The assumption of the symmetry of the underlying distribution is important to many statistical inference and modeling procedures. This paper provides a test of symmetry using kernel density estimation and the Kullback-Leibler information. Based on simulation studies, the new test procedure outperforms other tests of symmetry found in the literature, including the Runs Test of Symmetry. We illustrate our new procedure using real data.

Keywords: Test of symmetry; Power of the test; Overlap coefficients; Kernel Density estimation; Kullback-Leibler information

Research Article

Volume 3 Issue 2 - 2016

Hani M Samawi* and Robert Vogel

Department of Biostatistics, Georgia Southern University, USA

***Corresponding author:** Hani M Samawi, Department of Biostatistics, Jiann Ping Hsu College of Public Health, PO Box 8015, Georgia Southern University Statesboro, GA 30460, USA, Email: hsamawi@georgiasouthern.edu; rvogel@georgiasouthern.edu

Received: January 07, 2016 | **Published:** January 27, 2016

Introduction

Many statistical applications and inferences rely on the validity of the underlying distributional assumption. Symmetry of the underlying distribution is essential in many statistical inference and modeling procedures. There are several tests of symmetry in the literature; however most of these tests suffer from low statistical power. Tests have been suggested by Butler [1], Rothman & Woodroofe [2], Hill & Roa [3], Baklizi [4], and McWilliams [5]. McWilliams [5] showed, using simulation, that his runs test of symmetry is more powerful than those provided by Butler [1], Rothman & Woodroofe [2], and Hill & Roa [3] for various asymmetric alternatives. However, Tajuddin [6] introduced a distribution-free test for symmetry based on Wilcoxon two-sample test which is more powerful than the runs test.

Moreover, Modarres & Gastwirth [7] modified McWilliams [5] runs test by using Wilcoxon scores to weight the runs. The new test improved the power for testing for symmetry about a known center but did not perform well when the asymmetry is focused in regions close to the median for a given distribution. Mira [8], introduced a distribution free test for symmetry based on Boferroni's Measure. She showed that her test outperform tests introduced by Modarres & Gastwirth [7] and Tajauddin [6]. Recently, Samawi et al. [9] provided a test of symmetry based on a nonparametric overlap measure. They demonstrated that the test of symmetry based on an overlap measure outperformed other tests of symmetry in the literature, including the runs test. Samawi & Helu [10] introduced a runs test of conditional symmetry which is reasonably powerful to detect even small asymmetry in the shape of the conditional distribution. In addition, the Samawi & Helu [10] test does not need any approximation nor extra computations such as kernel estimation of the density function as in the other tests that are found in the literature.

This paper uses the Kullback-Leibler information to test for the symmetry of the underlying distribution. Let $f_1(x)$ and $f_2(x)$ be two probability density functions. Assume samples of observations are drawn from continuous distributions. The Kullback-Leibler discrimination information function is given by

$$D(f_1,f_2)=\int_{-\infty}^{\infty} f_1(x)\ln\left(\frac{f_1(x)}{f_2(x)}\right)dx,=\int_{-\infty}^{\infty} f_1(x)\ln(f_1(x))dx-\int_{-\infty}^{\infty} f_1(x)\ln(f_2(x))dx, \quad (1)$$

as defined by Kullback & Leibler [11]. For simplicity we will write (1) as

$$D(f_1,f_2)=D_{11}(f_1,f_1)-D_{12}(f_1,f_2), \text{ where}$$

$$D_{11}(f_1,f_1)=\int_{-\infty}^{\infty} f_1(x)\ln(f_1(x))dx \text{ and } D_{12}(f_1,f_2)=\int_{-\infty}^{\infty} f_1(x)\ln(f_2(x))dx.$$

This measure can be directly applied to discrete distributions by replacing the integrals with summations. It is well known that $D(f_1,f_2)\geq 0$, and the equality holds if and only if $f_1(x)=f_2(x)$ almost everywhere. The discrimination function $D(f_1,f_2)$ measures the disparity between f_1 and f_2.

Many authors used the discrimination function $D(.,.)$ for testing goodness of fit of some distributions. For example see Alizadeh & Arghami [12,13].

In this paper we consider testing the null hypothesis of symmetry for an underlying absolutely continuous distribution $F(.)$ with known location parameter and density denoted by $f(.)$ $H_0: f(x)=f(-x)$ versus $H_a: f(x)\neq f(-x)$; for some x. Under the null hypothesis of symmetry, if we let

$f_1(x)=f(x)$ and $f_2(x)=f(-x)$ then $D(f_1,f_2)=0$.

Since kernel density estimation procedures are readily available in various statistical software packages such as SAS, STATA, S-Plus and R, we were interested in exploring the development of a new test of symmetry using kernel density estimation of $D(f_1,f_2)$. This paper will introduce a powerful test of symmetry based on Kullback-Leibler discrimination information function. The Kullback-Leibler information test of symmetry and its asymptotic properties are introduced in Section 2. A simulation study is provided in Section 3. Illustrations of the test using base deficit score data and final comments are given in Section 4.

Test of Symmetry Based on the Kullback-Leibler Discrimination Information Function

Assume that a random sample, $X_1,X_2....,X_n$, is drawn from absolutely continuous distribution $F(.)$ having known median, assumed to be 0. In the case of an unknown median, or if the center of the distribution is not known, then the data can be centered by a consistent estimate of the median. However, the implications of centering the data around a consistent estimator of the median on the asymptotic properties are not straightforward. Therefore, further investigations are needed to study the robustness of the proposed test of symmetry and compare it with other available tests of symmetry when the median is unknown. In this paper we will discuss only the case where the median of the underlying distribution is assumed known.

Consider testing for symmetry $H_0:f(x)=f(-x)$ versus $H_a:f(x)\neq f(-x);$for some $x.$ Let $f_1(x)=f(x)$ and $f_2(x)=f(-x)$. Under the null hypothesis, $D(f_1,f_2)=0$. An equivalent hypothesis for testing the symmetry is $H_0:D(f_1,f_2)=0$ versus $H_a:D(f_1,f_2)>0$ let $\hat{D}$ be a consistent nonparametric estimator of $D(f_1,f_2)$. Under the null hypothesis of symmetry and some regularity assumptions, which will be discussed later in this paper, we propose the following test of symmetry:

$$z_0=\frac{\hat{D}-0}{\hat{\sigma}_{\hat{D}}}\xrightarrow{L}N(0,1), \qquad (2)$$

For large *n*, where $\hat{\sigma}_{\hat{D}}$ is a consistent estimator of the standard error of $\hat{D}$. An asymptotic significant test procedure at level α is to reject H_0 if $z_0>z_\alpha$, where z_α is the upper α percentile of the standard normal distribution.

Kernel estimation of $D(f_1,f_2)$

For the i.i.d. sample $X_1,X_2....,X_n$, let $\hat{D}_{11}(f_1,f_1)$ be an estimate of $D_{11}(f_1,f_1)$. To address which estimator of $D_{11}(f_1,f_1)$ will be appropriate to our inference procedure we need to state some necessary conditions:

C1: f is continuous. (Smoothness conditions)

C2: f is k times differentiable. (Smoothness conditions)

C3: $D_{11}([X],[X])<1$, where [X] is the integer part of *X*. (Tail condition)

C4: $Inf_{f(x)>0}f(x)>0$ (Tail condition)

C5: $\int f(\ln f)^2<\infty$ (Peak condition) (Note that, this is also a mild tail condition.)

C6: f is bounded. (Peak condition)

Some suggested estimators for $-D_{11}(f_1,f_1)=\int_{-\infty}^{\infty}f_1(x)\ln(f_1(x))dx$ may be found in the literature. These include the plug-in estimates of entropy which are based on a consistent density estimate f_n of f. For example, the integral estimate of entropy introduced by Dmitriev & Tarasenko [14]. Joe [15] considers estimating $-D_{11}(f_1,f_1)$ when f_1 is a multivariate pdf, but he points out that the calculation when $\hat{f}_1$ is a kernel estimator gets more difficult when the dimension of the integral is more than two. He therefore excludes the integral estimate from further study. The integral estimator can however be easily calculated if, for example, $\hat{f}_1$ is a histogram.

The re-substitution estimate is proposed by Ahmad & Lin [16] as follows:

$$-\hat{D}_{11}(\hat{f}_1,\hat{f}_1)=-\frac{1}{n}\sum_{i=1}^{n}\ln\hat{f}_1(X_i), \qquad (3)$$

Where $\hat{f}_1$ is a kernel density estimator? They showed the mean square consistency of (3), such that $\lim_{n\to\infty}E\{(\hat{D}_{11}(\hat{f}_1,\hat{f}_1)-D_{11}(f_1,f_1))^2\}=0$ Joe [15] considers the estimation of $-D_{11}(f_1,f_1)$ for multivariate pdfs by an entropy estimate of the re-substitution type (3), also based on a kernel density estimate. He obtained asymptotic bias and variance terms, and showed that non-unimodal kernels satisfying certain conditions can reduce the mean square error. His analysis and simulations suggest that the sample size needed for good estimates increases rapidly when the dimension of the multivariate density increases. His results rely heavily on conditions C4 and C6. Hall & Morton [17] investigated the properties of an estimator of the type (3) both when f_n is a histogram density estimator and when it is a kernel estimator. For the histogram estimation they showed that $\lim_{n\to\infty}n^{1/2}(\hat{D}_{11}(\hat{f}_1,f_1)-D_{11}(f_1,f_1))\sim N(0,\sigma^2)$ under certain tail and smoothness conditions with.

$$\sigma^2=Var(\ln(f(X))) \qquad (4)$$

Other estimators using sampling-spacing are investigated by Tarasenko [18], Beirlant & van Zuijlen [19], Hall [20], Cressie [21], Dudewicz & van der Meulen [22], and Beirlant [23]. Finally, other nonparametric estimator has been discussed by many authors including Vasicek [24], Dudewicz & Van der Meulen [22], Bowman [25] and Alizadeh [26]. Among these various entropy estimators, Vasicek's sample entropy has been most widely used in developing entropy based statistical procedures. However, deriving the asymptotic distribution for there $\hat{D}$ is hard to establish. Therefore, in this paper we will adopt the kernel re-substitution estimate which is proposed by Ahmad & Lin [16].

We will adopt the notation of Samawi et al. [9]. Our proposed test of symmetry is as follow: Let $X_1,X_2....,X_n$ be a random sample from absolutely continuous distribution $F(.)$ which is continuously differentiable with uniformly bounded derivatives and having known median.

Let K be a kernel function satisfying the condition

$$\int_{-\infty}^{\infty} K(x)dx=1. \quad (5)$$

For simplicity, the kernel *K* will be assumed to be a symmetric density function with mean 0 and finite variance; an example is the standard normal density. The kernel estimators for $f(w_i)$ and $f(-w_i),i=1,2,...,C$, are:

$$\hat{f}_K(-w_i)=\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{-w_i-x_j}{h}\right) \quad (6)$$

and

$$\hat{f}_K(w_i)=\frac{1}{nh}\sum_{j=1}^{n}K\left(\frac{w_i-x_j}{h}\right), \quad (7)$$

Respectively, where C is the number of bins and depends on the sample size. As in Samawi et al. [9], we suggest to take the integer of $C=\sqrt{n}$. In addition, h is the bandwidths of the kernel estimators satisfying the conditions that $h>0,h\to 0$ and $(nh\to\infty)$ as $n\to\infty$. There are many choices of the bandwidths (h). In our procedure we use the method suggested by Silverman [27] Using the normal distribution as the parametric family, the bandwidths of the kernel estimators are

$$h=0.9A(n)^{-1/5}, \quad (8)$$

Where A =min{standard deviation of ($x_1,x_2....,x_n$), interquantile range of ($x_1,x_2....,x_n$)/1.349}. This form of (8) is found to be adequate choices of the bandwidth for many purposes which minimizes the integrated mean squared error (IMSE),

$$IMSE=\int E[\hat{f}_K(x)-f(x)]^2dx. \quad (9)$$

We will use the Samawi et al. [9] suggestion to calculate the bins as follows: Let $R=range(x_1,x_2,...,x_n)$, then bins will be selected as $w_i=w_{i-1}+\delta_x$, where $i=2,...,C$, $w_1=\min(x_1,x_2,...,x_n)$ and $\delta_x=\frac{R}{C}$.

Using the above kernel estimator, the nonparametric kernel estimator of $D(f_1,f_2)$ under the null hypothesis is given by

$$\hat{D}=\int \hat{f}_K(x)\ln\left(\frac{\hat{f}_K(x)}{\hat{f}_K(-x)}\right)dx,=\hat{D}_{11}(\hat{f}_K(x),\hat{f}_K(x))-\hat{D}_{12}(\hat{f}_K(x),\hat{f}_K(-x)), \quad (10)$$

Which can be approximated by?

$$\hat{D}=\frac{1}{C}\sum_{i=1}^{C}\ln\hat{f}_K(w_i)-\frac{1}{C}\sum_{i=1}^{C}\ln\hat{f}_K(-w_i) \quad (11)$$

The approximate variance of $\hat{D}$ is given by

$$Var(\hat{D})=\frac{Var(\sum_{i=1}^{C}\ln\hat{f}_K(w_i))}{C^2}+\frac{Var(\sum_{i=1}^{C}\ln\hat{f}_K(-w_i))}{C^2}.$$

Asymptotic properties of $\hat{D}$

The nonparametric kernel estimator of $D(f_1,f_2)$ ($\hat{D}$) is based on the univariate kernel for density estimation, $K:\mathbb{R}\to\mathbb{R}$. The necessary regularity conditions imposed on the univariate kernel for density estimation are:

I. $\int_R K(z)dz=1.$

II. $\int_R z^\beta K(z)dz=0$ for any $\beta=1,...,r-1$, and $\int_R|z|^r K(z)dz<\infty.$

III. $R=\int_R K^2(z)dz<\infty.$

IV. $h>0,h\to 0$, $(nh\to\infty)$ and $(\frac{nh}{\log n}\to\infty)$

These conditions may be found in Silverman [27] (Chapter 3) or Wand & Jones [28] (Chapter 2).

To show consistency of $\hat{D}$, apply the kernel density asymptotic properties found in Silverman [27], (Chapter 3) or Wand & Jones [28], (Chapter 2). Under assumptions 1-4 and assuming that the density $f:\mathbb{R}\to\mathbb{R}$ is continuous at each w_i , i=1, 2,... C,

$$Bias(\hat{f}_K(-w_i))=o(1)_- \text{ and } Bias(\hat{f}_K(w_i))=o(1)_+ \quad (12)$$

$$Var(\hat{f}_K(-w_i))=\frac{f(-w_i)}{nh}\int_{\mathbb{R}}K^2(z)dz+o(\frac{1}{nh}) \text{ and } Var(\hat{f}_K(w_i))=\frac{f(w_i)}{nh}\int_{\mathbb{R}}K^2(z)dz+o(\frac{1}{nh}), \quad (13)$$

and for $h>0,h\to 0$ and $(nh\to\infty)$ as $n\to\infty$ $\hat{f}_K(-w_i)\to^P f(-w_i)$ and $\hat{f}_K(w_i)\to^P f(w_i)$ If f(.) uniformly continuous, then the kernel density estimate is strongly consistent. Moreover, as in Ahmad & Lin [16], $_C\underline{\lim}_\infty E\{(\hat{D}_{11}(\hat{f}_K(x),\hat{f}_K(x))-D_{11}(f_K(x),f_K(x)))^2\}=0,$ and hence $\hat{D}_{11}(\hat{f}_K(x),\hat{f}_K(x))\xrightarrow{P}D_{11}(f_K(x),f_K(x))$, as $C\to\infty$ and $\hat{D}_{12}(\hat{f}_K(x),\hat{f}_K(-x))\xrightarrow{P}D_{12}(f_K(x),f_K(-x))$, as $C\to\infty$. However, since $\hat{D}=\hat{D}_{11}(\hat{f}_K(x),\hat{f}_K(x))-\hat{D}_{12}(\hat{f}_K(x),\hat{f}_K(-x))$ therefore $\hat{D}\xrightarrow{p}D(f(w),f(-w))$, as $C\to\infty.$

To drive the asymptotic distribution of $\hat{D}$, we will define $D(f_1,f_2)$ as a functional

$$D(f_1,f_2)=\int_{-\infty}^{\infty}f_1(w)\ln(f_1(w))dw-\int_{-\infty}^{\infty}f_1(w)\ln f_2(w)dw=\int_{-\infty}^{\infty}\ln(f_1(w))dF_1-\int_{-\infty}^{\infty}\ln f_2(w)dF_1.$$

Using the previously stated regularity conditions, some regularity conditions given by Serfing [29] and assuming that the Gâteaux derivatives of the functional $D(f_1,f_2)$ exist, we can show that the partial influence function of the functional $D(f_1,f_2)$ [30] are as follows:

$$L_1(w;F_1,F_1)=\ln f_1(w)-\int_{-\infty}^{\infty} f_1(w)\ln f_1(w)dw,$$

and

$$L_2(w;F_1,F_2)=\ln f_2(w)-\int_{-\infty}^{\infty} f_1(w)\ln f_2(w)dw.$$

Note that

$\int L_1(w;F_1(w),F_1(w))dF_1(w)=0$ and $\int L_2(w;F_1(w),F_2(w))dF_1(w)=0.$

Now using this functional representation of $D(f_1,f_2)$, then as in Samawi et al. [30] and Serfing [29],

$$\sqrt{C}(\hat{D}-D(f_1,f_2))\xrightarrow{L}N(0,\sigma_{\hat{D}}^2), \qquad (14)$$

Where $\sigma_{\hat{D}}^2=\int L_1^2(w;F_1,F_1)dF_1+\int L_2^2(w;F_1,F_2)dF_1$

A consistent estimate for $\sigma_{\hat{D}}^2$ is given by

$$\hat{\sigma}_{\hat{D}}^2=\frac{1}{C}\sum_{i=1}^{C}L_1^2(w;\hat{F}_1,\hat{F}_1)+\frac{1}{C}\sum_{i=1}^{C}L_2^2(w;\hat{F}_1,\hat{F}_2),$$ Where,

$$L_1(w_i;\hat{F}_1,\hat{F}_1)=\ln\hat{f}_1(w_i)-\hat{D}_{11}(\hat{f}_1(w_i),\hat{f}_1(w_i)) \text{ and } L_2(w_i;\hat{F}_1,\hat{F}_2)=\ln\hat{f}_2(w_i)-\hat{D}_{12}(\hat{f}_1(w_i),\hat{f}_2(w_i)),i=1,2,...,C,$$

Where in our case $f_1(w_i)=f(w_i)$ and $f_2(w_i)=f(-w_i)$.

For discussions about different methods addressing the issue of the performance of kernel density estimation at the boundary, see Hall & Park [31].

Simulation Study

As in Samawi et al. [9], to gain some insight of our procedure, a simulation study was conducted to investigate the performance of our new test of symmetry based on $\hat{D}$. We compared our proposed test of symmetry with the test proposed by McWilliams [5], Modarres & Gastwirth [32], Mira [8] Bonferroni's test, and Samawi et al. [9] tests of symmetry.

As in McWilliams [5], the runs test is described as follows: For any random sample of size n, let $Y_{(1)},Y_{(2)},...,Y_{(n)}$ denote the sample values ordered from the smallest to largest according to their absolute value (signs are retained), and $S_1,S_2,...,S_n$ denote indicator variables designating the sign of the $Y_{(j)}$ values [S_j=1 if $Y_{(j)}$ is nonnegative, 0 otherwise]. Thus, the test statistic used for testing symmetry is = the number of runs in $S_1,S_2,...,S_n$ sequence= $1+\sum_{j=2}^{n} I_j$, where

$$I_j=\begin{cases}0 & \text{if } S_j=S_{j-1}\\ 1 & \text{if } S_j\neq S_{j-1}\end{cases}.$$

We reject the null hypothesis if R^* is smaller than a critical value (c_α) at the pre-specified value of α. Moreover, Mira [8] Bonferroni's test is $\gamma_1(F_n)=2(\bar{X}_n-X_{s:n})$, where $X_{s:n}=Median(X_1,X_2,...,X_n)$. The process is to reject the null hypothesis if $|\gamma_1(F_n)|\geq\frac{a_n}{\sqrt{n}}S_c(\gamma_1,F_n)$, where

$a_n\to z_{1-\frac{\alpha}{2}}$ as $n\to\infty$, $S_c^2(\gamma_1,F_n)=4\hat{\sigma}^2+(D_{n,c})^2-4D_{n,c}S_{\bar{\mu}_F}$, $\hat{\sigma}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X}_n)^2$, $S_{\bar{\mu}_F}=\bar{X}_n-\frac{2}{n}\sum_{i=1}^{n}X_iI(X_i\leq X_{s:n})$, $D_{n,c}=\frac{n^{1/5}}{2c}(X_{[(n/2)+cn^{4/5}]:n}-X_{[(n/2)+cn^{4/5}+1]:n})$, and $c=0.5$.

The Modarres & Gastwirth [32] test is the hybrid test of sign test in the first stage and a percentile-modified two-sample Wilcoxon see Gastwirth [33] test in the second stage. Finally, Samawi et al. [9] test of symmetry is based on kernel estimate of the overlap measure.

In the following simulation, SAS version 9.3 {proc kde; method=srot} is used. As in McWilliams [5], the generalized lambda distribution see, Ramberg & Schmeiser [34] is used in our simulation with following set of parameters:

1- $\lambda_1=0,\lambda_2=0.197454,\lambda_3=0.134915,\lambda_4=0.134915$, (Symmetric)

2- $\lambda_1=0,\lambda_2=1,\lambda_3=1.4,\lambda_4=0.25$,

3- $\lambda_1=0,\lambda_2=1,\lambda_3=0.00007,\lambda_4=0.1$,

4- $\lambda_1=3.586508,\lambda_2=0.04306,\lambda_3=0.025213,\lambda_4=0.094029$,

5- $\lambda_1=0,\lambda_2=-1,\lambda_3=-0.0075,\lambda_4=-0.03$,

6- $\lambda_1=-0.116734,\lambda_2=-0.351663,\lambda_3=-0.13,\lambda_4=-0.16$,

7- $\lambda_1=0,\lambda_2=-1,\lambda_3=-0.1,\lambda_4=-0.18$,

8- $\lambda_1=0,\lambda_2=-1,\lambda_3=-0.001,\lambda_4=-0.13$,

9- $\lambda_1=0,\lambda_2=-1,\lambda_3=-0.0001,\lambda_4=-0.17$.

To generate the observations we used $x_i=\lambda_1+\frac{1}{\lambda_2}(u_i^{\lambda_3}-(1-u_i)^{\lambda_4}$, $i=1,...,m$, where u_i a uniform random number. The significance level used in the simulation is $\alpha=0.05$, with sample sizes n=30, 50, and 100. To investigate the Type I error, the symmetric distributions used in the simulation are the first case of the generalized lambda and the normal. Our simulation is based on 5000 simulated samples. The 95% confidence intervals of the true probability of type I error under the null hypothesis with $\alpha=0.05$ are (0.04396, 0.05504).

Table 1.1 shows the estimated probability of type I error. Our test is an asymptotic test with a slight bias in D(., .) and in the variance estimation for small sample size. For sample sizes more than 30, the test seems to have an estimated probability of type I error close to the nominal value 0.05. However, Bonferroni's test seems to be conservative test procedure, while Modarres, Gastwirth test is slightly conservative for small sample size. Table 1.2 and Table 1.3 show that using D(., .) based test is more powerful than McWilliams [5], Bonferroni's, Modarres & Gastwirth [32] and Samawi et al. [9] tests in all of the presented cases. The efficiency increases as the sample size increases.

Note: The values of skewness (α_3) and kurtosis (α_4) are from McWilliams [5].

Note: The values of skewness (α_3) and kurtosis (α_4) are from McWilliams [5].

Table 1.1: Probability of Type I Error under the Null Hypothesis. (α=0.05).

Distribution	n	Run Tests	Test Based on the Overlap	Bonferroni's $\gamma_1(F_n)$	Modarres and Gastwirth (1998) Test $W_{0.80}$	Test Based on Kullback-Leibler Information
Case #1 generalized lambda $\lambda_1 = 0, \lambda_2 = 0.197454, \lambda_3 = 0.134915,$ $\lambda_4 = 0.134915,\ \alpha_3 = 0, \alpha_4 = 3.0$	30	0.046	0.056	0.03	0.027	0.051
	50	0.052	0.051	0.032	0.044	0.047
	100	0.058	0.052	0.027	0.046	0.051
Normal (0, 1)	30	0.052	0.057	0.03	0.03	0.052
	50	0.048	0.055	0.03	0.043	0.051
	100	0.051	0.052	0.032	0.048	0.052

Table 1.2: Power of Kullback-Leibler Information based test, with comparison with other tests Under Alternative Hypotheses (α=0.05).

Case #	n	Run Test	Test Based on the Overlap	Bonferroni's $W_{0.80}$	Modarres and Gastwirth (1998) Test $W_{0.80}$	Test based on Kullback-Leibler Information
-2 λ_1=0,λ_2=1,λ_3=1.4,λ_4=0.25 α_3=0.5,α_4=2.2	30	0.282	0.501	0.253	0.495	0.948
	50	0.456	0.839	0.352	0.941	0.992
	100	0.781	0.999	0.5	1	1
-3 $\lambda_1 = 0, \lambda_2 = 1, \lambda_3 = 0.00007, \lambda_4 = 0.1, \alpha_3 = 1.5, \alpha_4 = 5.8$	30	0.444	0.846	0.508	0.61	0.98
	50	0.678	0.953	0.756	0.99	0.999
	100	0.913	1	0.966	1	1
-4 $\lambda_1 = 3.586508, \lambda_2 = 0.04306, \lambda_3 = 0.025213, \lambda_4 = 0.094029$ α_3=0.9,α_4=4.2	30	0.12	0.38	0.154	0.179	0.684
	50	0.134	0.541	0.26	0.474	0.854
	100	0.245	0.761	0.488	0.845	0.946
-5 $\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.0075, \lambda_4 = -0.03, \alpha_3 = 1.5, \alpha_4 = 7.5$	30	0.141	0.451	0.231	0.247	0.81
	50	0.201	0.601	0.41	0.652	0.92
	100	0.336	0.839	0.741	0.954	0.98

Table 1.3: Power of Overlap based test and Run Tests under Alternative Hypotheses. (α=0.05).

Case #	n	Runs Test	Test Based on the Overlap	Bonferroni's $\gamma_1(F_n)$	Modarres and Gastwirth (1998) Test $W_{0.80}$	Test Based on Kullback-Leibler Information
$\lambda_1 = -0.116734, \lambda_2 = -0.351663, \lambda_3 = -0.13, \lambda_4 = -0.16,$ α_3=0.8,α_4=11.4	30	0.051	0.161	0.034	0.033	0.191
	50	0.055	0.174	0.04	0.055	0.225
	100	0.053	0.21	0.059	0.12	0.331
$\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.1, \lambda_4 = -0.18, \alpha_3 = 2.0, \alpha_4 = 21.2$	30	0.101	0.189	0.091	0.092	0.452
	50	0.111	0.241	0.155	0.21	0.611
	100	0.122	0.361	0.336	0.478	0.737
$\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.001, \lambda_4 = -0.13, \alpha_3 = 3.16, \alpha_4 = 23.8$	30	0.544	0.98	0.643	0.655	0.993
	50	0.752	0.999	0.888	0.992	1
	100	0.961	1	0.996	1	1
$\lambda_1 = 0, \lambda_2 = -1, \lambda_3 = -0.0001, \lambda_4 = -0.17 \alpha_3 = 3.88, \alpha_4 = 40.7$	30	0.571	1	0.685	0.676	0.993
	50	0.81	1	0.916	0.995	0.999
	100	0.963	1	0.999	1	1

Illustration Using Base Deficit Data

We applied our new test procedure of symmetry to the base deficit (bd) data as in Samawi et al. [9]. The base deficit score refers to a deficit of "base" present in the blood. Base deficit scores were first established by Davis et al. [35]. The base deficit score has been found correlated to many variables in the trauma population, such as, mechanism of injury, the presence of intra-abdominal injury, transfusion requirements, mortality, the risk of complications, and the number of days spent in the intensive care unit as indicated by Tremblay et al. [36] and Davis et al. [37].

The samples used in this illustration are part from the data collected based on a retrospective study of the trauma registry at a level 1 trauma center between January, 1998 and May, 2000. The primary concern was to determine at what point we can differentiate between life and death based on a base deficit score. A first step in this analysis is to determine if there is a difference in location for the base deficit score of those who survive and those who fail to survive. As is frequently the case in such studies, the underlying distribution is assumed "normal" or at least symmetric and a t-test or a nonparametric test would be performed without checking the assumptions. In either case a test of symmetry is almost never considered as a means of determining how one may proceed in the analysis. Based on the conclusions of a test of symmetry, the analyst can chose the most powerful test for location. The goal is to test the hypothesis that, on average, the base deficit score is the same for those who survive and those who fail to survive their injuries. The injuries of interest in this group of patients are either penetrating injury or blunt injury. However, before deciding on the test procedure, we need to check the assumptions of underlying distribution of the base deficit for both penetrating injury and blunt injury groups of patients. In particular, the assumption of symmetry of the underlying distribution needs to be verified. The data will be centered about the estimated measure of location to perform the tests of symmetry.

Figure 1.1 and Figure 1.2 show the box plot for penetrating injury and blunt injury groups for dead and alive patients respectively. Clearly there is some asymmetry on all four distributions. Also, Table 2.1 and Table 2.2 show summery statistics for penetrating injury and blunt injury groups for dead and alive patients respectively. Table 2.3 shows the overlap based test, the runs test and the proposed test of symmetry based on the Kullback-Leibler information of symmetry for the underlying distribution for patients discharged alive and dead patients of blunt trauma and penetrating trauma. We reject the assumption of symmetry for underlying distribution of these groups.

The proposed test of symmetry based on the Kullback-Leibler information, appears to outperform the other tests of symmetry in the literature in terms of power. Our test is more sensitive to detect a slight asymmetry in the underlying distribution than other tests proposed in the literature. Moreover, the kernel density estimation literature is very rich and many of the proposed methods and the improved methods are available on statistical software, such as SAS™, S-plus, Stata and R. Since based on the Kullback-Leibler information can be used in multivariate cases as well as in univariate cases, our proposed test of symmetry can be extended to multivariate cases for diagonal symmetry, conditional symmetry and other types of symmetry.

Table 2.1: Summery statistics for base deficit for dead patients.

Descriptives					
	Type of Wound			Statistic	Std. Error
BD	Penetrating	Mean		-10.81	0.846
		95% Confidence Interval for Mean	Lower Bound	-12.49	
			Upper Bound	-9.12	
		5% Trimmed Mean		-10.68	
		Median		-10	
		Variance		52.904	
		Std. Deviation		7.274	
		Minimum		-29	
		Maximum		9	
		Range		38	
		Interquartile Range		10	
		Skewness		-0.21	0.279
		Kurtosis		0.102	0.552
	Blunt	Mean		-7.59	0.444
		95% Confidence Interval for Mean	Lower Bound	-8.46	
			Upper Bound	-6.71	
		5% Trimmed Mean		-7.3	
		Median		-6	
		Variance		60.65	
		Std. Deviation		7.788	
		Minimum		-37	
		Maximum		23	
		Range		60	
		Interquartile Range		10	
		Skewness		-0.518	0.139
		Kurtosis		1.368	0.277

Table 2.2: Summery statistics for base deficit for alive patients.

Descriptives					
Base Deficit	**Type of Wound**			**Statistic**	**Std. Error**
	penetrating	Mean		-3.52	0.202
		95% Confidence Interval for Mean	Lower Bound	-3.91	
			Upper Bound	-3.12	
		5% Trimmed Mean		-3.06	
		Median		-2.7	
		Variance		24.683	
		Std. Deviation		4.968	
		Minimum		-28	
		Maximum		12	
		Range		40	
		Interquartile Range		5	
		Skewness		-1.75	0.099
		Kurtosis		5.079	0.199
	Blunt	Mean		-1.8	0.059
		95% Confidence Interval for Mean	Lower Bound	-1.92	
			Upper Bound	-1.69	
		5% Trimmed Mean		-1.61	
		Median		-1.3	
		Variance		11.601	
		Std. Deviation		3.406	
		Minimum		-27	
		Maximum		13	
		Range		40	
		Interquartile Range		3	
		Skewness		-1.22	0.043
		Kurtosis		4.39	0.085

Table 2.3: Test of symmetry with summary statistics.

	Injury Type	**N**	**Test**	**Significance**
Kullback-Leibler Information	Penetrating - Dead	74	3.989	<0.0001
	Penetrating - alive	603	13.057	<0.0000
Overlap test*	Penetrating - Dead	74	-2.09	0.0183
	Penetrating - alive	603	-16.928	<0.0001
Run test*	Penetrating - Dead	74	-2.065	0.0195
	Penetrating - alive	603	-16.41	<0.0001
Kullback-Leibler Information	Blunt - Dead	306	13.92	<0.0001
	Blunt - alive	3275	8.053	<0.0001
Overlap test*	Blunt - Dead	306	-13.264	<0.0001
	Blunt - alive	3275	-79.074	<0.0001
Run test*	Blunt - Dead	306	-10.29	<0.0001
	Blunt - alive	3275	-52.405	<0.0001

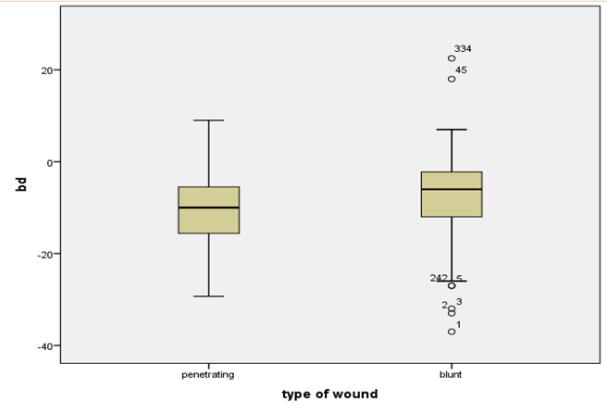

Figure 1.1: Box plot to base deficit for dead patients.

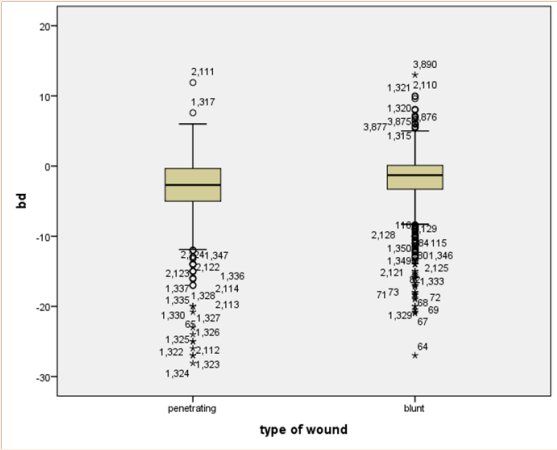

Figure 1.2: Box plot to base deficit for alive patients.

Acknowledgement

None.

Conflict of Interest

None.

References

1. Butler CC (1969) A test for symmetry using sample distribution function. The Annals of Mathematical Statistics 40: 2211-2214.
2. Rothman ED, Woodroofe M (1972) A Cramer-Von Mises type statistic for testing symmetry. The Annals of Mathematical Statistics 43: 2035-2038.
3. Hill DL, Rao PV (1977) Test of Symmetry based on Cramer-Von Mises statistics. Biometrika 64(3): 489-494.
4. Baklizi A (2003) A conditional distribution free runs test for symmetry. Journal of Nonparametric Statistics 15(6): 713-718.
5. McWilliams TP (1990) A distribution free test of symmetry based on a runs statistic. Journal of the American Statistical Association 85(412): 1130-1133.
6. Tajuddin IH (1994) Distribution-Free test for symmetry based on Wilcox on two-sample test. J Applied Statistics 21(5): 409-415.
7. Modarres R, Gastwirth JL (1996) A modified runs test of symmetry. Statistics & Probability Letters 31(2): 107-112.
8. Mira A (1999) Distribution-free test for symmetry based on Bonferroni's measure. Journal of Applied Statistics 26(8): 959-971.
9. Samawi HM, Helu A, Vogel R (2011) A nonparametric test of symmetry based on the overlapping coefficient. Journal of Applied Statistics 38(5): 885-898.
10. Samawi HM, Helu A (2011) Distribution-Free Runs Test for Conditional Symmetry. Communications in Statistics Theory and Methods 40(15): 2709-2718.
11. Kullback S, Leibler RA (1951) On information and sufficient. Annals of Mathematical Statistics 22(1): 79-86.
12. Alizadeh Noughabi H, Arghami NR (2011a) Testing exponentially using transformed data. Journal of Statistical Computation and Simulation 81(4): 511-516.
13. Alizadeh NH, Arghami NR (2011b) Monte Carlo comparison of five exponentially tests using different entropy estimates. Journal of Statistical Computation and Simulation 80(11): 1579-1592.
14. Dmitriev, Yu G, Tarasenko FP (1973) On the estimation functions of the probability density and its derivatives. Theory Probab Appl 18: 628-633.
15. Joe H (1989) On the estimation of entropy and other functional of a multivariate density. Ann Inst Statist Math 41(4): 683-697.
16. Ahmad IA, Lin PE (1976) A nonparametric estimation of the entropy for absolutely continuous distributions. Information Theory, IEEE Transactions 22(3): 372-375.
17. Hall P, Morton SC (1993) On the estimation of the entropy. Ann Inst Statist Math 45(1): 69-88.
18. Tarasenko FP (1968) On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit. Proceedings of the IEEE 56(11): 2052-2053.
19. Beirlant J (1985) Limit theory for spacing statistics from general univariate distributions. Pub Inst Stat Univ Paris XXXI fasc 1: 27-57.
20. Hall P (1984) Limit theorems for sums of general functions of m-spacing. Math Proc Camb Phil Soc 96(3): 517-532.
21. Cressie N (1977) The minimum of higher order gaps. Australian Journal of Statistics 19(2): 132-143.
22. Dudewicz E, Van der Meulen E (1981) Entropy based tests of uniformity. Journal of American Statistical Association 76(376): 967-974.
23. Beirlant J, Zuijlen MCA (1985) The empirical distribution function and strong laws for functions of order statistics of uniform spacings. Journal of Multivariate Analysis 16(3): 300-317.
24. Vasicek O (1976) A test for normality based on sample entropy. Journal of the Royal Statistical Society 38(1): 54-59.
25. Bowman AW (1992) Density based tests for goodness-of-fit. Journal of Statistical Computation and Simulation 40: 1-13.
26. Alizadeh Noughabi H (2010) A new estimator of entropy and its application in testing normality. Journal of Statistical Computation and Simulation 80(10): 1151-1162.

27. Silverman BW (1986) Density estimation for statistics and data analysis. London Chapman and Hall.

28. Wand MP, Jones MC (1995) Kernel Smoothing London. Chapman and Hall.

29. Serfing RJ (1980) Approximation theorems of mathematical statistics. John Wiley & Sons, Inc, USA.

30. Samawi HM, Woodworth GG, Lemke J (1998) Power estimation for two-sample tests using importance and antithetic resampling. Biometrical Journal 40(3): 341-354.

31. Hall P, Park BU (2002) New methods for bias correction at endpoints and boundaries. The Annals of Statistics 30(5): 1460-1479.

32. Modarres R, Gastwirth JL (1998) Hybrid test for the hypothesis of symmetry. Journal of Applied Statistics 25(6): 777-783.

33. Gastwirth JL (1965) Percentile modification of two sample ranked test. Journal of the American Statistical Association 60(312): 1127-1141.

34. Ramberg JS, Schmeiser BW (1974) An approximate method for generating Asymmetric random variables. Communications of the ACM 17: 78-82.

35. Davis JW, Shackford SR, Mackersie RC, Hoyt DB (1988) Base deficit as a guide to volume Resuscitation. J Trauma 28(10): 1464-1467.

36. Tremblay LN, Feliciano DV, Rozycki GS (2002) Assessment of initial base deficit as a predictor of outcome: mechanism does make a difference. Am Surg 68(8): 689-694.

37. Davis JW, Mackersie RC, Holbrook TL, Hoyt DB (1991) Base deficit as an indicator of significant abdominal injury. Ann Emerg Med 20(8): 842-844.

Citation: Samawi, Vogel R (2016) A Test of Symmetry Based on the Kernel Kullback-Leibler Information with Application to Base Deficit Data. Biom Biostat Int J 3(2): 00060. DOI: 10.15406/bbij.2016.03.00060