Published in final edited form as:

J Stat Plan Inference. 2011 January 1; 141(1): 335–344. doi:10.1016/j.jspi.2010.06.011.

A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data

Liansheng Larry Tang^{a,*} and N. Balakrishnan^{b,1}

^aDepartment of Statistics, George Mason University, Fairfax, VA 22030, USA

^bDepartment of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada L8S 4K1

Abstract

The Wilcoxon–Mann–Whitney statistic is commonly used for a distribution-free comparison of two groups. One requirement for its use is that the sample sizes of the two groups are fixed. This is violated in some of the applications such as medical imaging studies and diagnostic marker studies; in the former, the violation occurs since the number of correctly localized abnormal images is random, while in the latter the violation is due to some subjects not having observable measurements. For this reason, we propose here a random-sum Wilcoxon statistic for comparing two groups in the presence of ties, and derive its variance as well as its asymptotic distribution for large sample sizes. The proposed statistic includes the regular Wilcoxon rank-sum statistic. Finally, we apply the proposed statistic for summarizing location response operating characteristic data from a liver computed tomography study, and also for summarizing diagnostic accuracy of biomarker data.

Keywords

Diagnostic accuracy; Distribution-free comparison; Random-sum statistic; Tomography study

1. Introduction

The Wilcoxon–Mann–Whitney statistic is a commonly used distribution-free method for the comparison of two groups. This statistic is the sum of ranks of all observations in group 1 within the pooled sample. In the case of ties, the statistic is modified by allowing midranks for the tied observations. The use of the Wilcoxon statistic requires the sample sizes for the two groups to be fixed (Lehmann, 1998). However, this requirement is violated in some of the applications. For example, medical imaging studies use either a human or a computer algorithm as an observer to identify a lesion location on every image and give a rating to the image according to the belief of lesion existence. The data consist of ratings of a random number of abnormal images correctly identified by the rater and ratings of a fixed number of normal images. Examples of such data can be found in recent studies involving mammogram scans in breast cancer screening and PET scans in human liver or lung tumor screening. A study by Seltzer et al. (1991) considered a radiologist's performance in reviewing liver computed tomographic (CT) images. In this study, 46 normal CT images without lesions and 46 abnormal CT images with simulated tumors were presented to a radiologist observer. The observer was told to specify the most suspicious location and give

a six-category rating to the location for each image. Since each image has only one rated location, the rating of the location is used as the rating of the associated image. Table 1 presents that rating data from the observer. As mentioned earlier, since the number of correctly localized abnormal images is random, the usual Wilcoxon statistic cannot be used in this case to summarize the observer performance. Here, we propose a random-sum Wilcoxon-type statistic that is suitable for this type of imaging data.

There are limitations in the use of the usual Wilcoxon statistic for other types of diagnostic data as well. For example, it is used to evaluate the accuracy of markers in marker studies (Zhou et al., 2002) when a marker is measured on both diseased and non-diseased patients. The magnitude of the Wilcoxon statistic indicates how accurately the marker can distinguish the diseased subjects from the non-diseased ones. However, in some of these studies, marker measurements have zero values due to the limit of detection (Vexler et al., 2008), and non-zero values which are the actual observed values. This situation results in random numbers of actual measurements, which means the usual Wilcoxon statistic cannot be used. Still, if one were to use the usual Wilcoxon statistic for the comparison of the two groups, it would result in an incorrect Type-I error rate. The construction of a Wilcoxon-type statistic would facilitate a distribution-free comparison between the two groups in these cases.

The rest of the paper is organized as follows. In Section 2, we propose a random-sum Wilcoxon statistic for the comparison of two groups with random sample sizes, which would be useful in the aforementioned situations. We examine the asymptotic properties of the proposed statistic in this section and also discuss the relationship between the proposed statistic and the usual Wilcoxon statistic under both null and alternative hypotheses. In Section 3, we describe how the proposed statistic can be applied to data from LROC studies and biomarker studies. In Section 4, we evaluate the finite-sample performance of the asymptotic variance of the proposed statistic under different distributions. In Section 5, we apply the proposed random-sum Wilcoxon statistic for evaluating the localizing accuracy of the observer in the example involving liver CT, and for summarizing the diagnostic accuracy of cancer biomarkers with the limit of detection. Finally, some concluding remarks are made in Section 6. The proofs of the key results are relegated to the Appendix.

2. Notation and methods

We start this section with a brief introduction to the usual Wilcoxon rank-sum statistic.

2.1. Wilcoxon statistic

Suppose there are n_1 subjects in group 1 and n_2 subjects in group 2, with observations from different subjects being independent, and that the sample sizes n_1 and n_2 are fixed. Let X_i be an observation on the i -th subject in group 1 (for $i = 1, \dots, n_1$) and Y_j be an observation on the j -th subject in group 2 (for $j = 1, \dots, n_2$). Let F and G denote the cumulative distribution functions of X_i and Y_j , respectively. Then, the usual Wilcoxon rank-sum statistic is given by

$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij}$, where

$$U_{ij} = \begin{cases} 1 & \text{if } X_i > Y_j, \\ \frac{1}{2} & \text{if } X_i = Y_j, \\ 0 & \text{if } X_i \leq Y_j. \end{cases}$$

In terms of probability, W gives an estimator for $n_1 n_2 \{P(X > Y) + \frac{1}{2}P(X = Y)\}$. An expression for $\text{var}(W)$ is given by Lehmann (1998) as

$$\text{var}(W) = n_1 n_2 \xi_1 (1 - \xi_1) + n_1 n_2 (n_2 - 1) (\xi_2 - \xi_1^2) + n_1 (n_1 - 1) n_2 (\xi_3 - \xi_1^2) \quad (1)$$

when n_1 and n_2 are large, where

$$\xi_1 = P(X > Y) + \frac{1}{2} P(X = Y), \quad \xi_2 = E(U_{ij} U_{ij'}) \quad \text{and} \quad \xi_3 = E(U_{ij} U_{i'j'}).$$

The Wilcoxon statistic is commonly used to test the hypothesis $H_0: F = G$ against the alternative $H_a: F \neq G$, especially when the normal assumption for the population distributions is suspect. The test statistic $(W - E(W)) / \sqrt{\text{var}(W)}$ can be used to evaluate the power of the test and also to determine the sample size based on the alternative. Under H_0 , it is known that $E(W) = \frac{1}{2} n_1 n_2$ and $\text{var}(W)$ is given by

$$\text{var}_0(W) = \frac{1}{12} n_1 n_2 (n_1 + n_2 + 1) - \frac{n_1 n_2}{12(n_1 + n_2)(n_1 + n_2 + 1)} \sum_{i=1}^R (d_i^3 - d_i) \quad (2)$$

in the presence of ties, where R is the number of distinct values and d_i is the number of observations equalling the i -th smallest value in the pooled sample. When we consider X_i and Y_j as measurements of a diagnostic marker on a diseased group and a non-diseased group, another commonly used version of the Wilcoxon statistic is $W/n_1 n_2$. It gives an empirical estimate of the area under the receiver operating characteristic (ROC) curve, which is a common measure of accuracy of the marker to distinguish the diseased from the non-diseased subjects (Hanley and McNeil, 1982). Clearly, a continuous marker with no distinguishing power results in an AUC of $\frac{1}{2}$, which is also $E[W/n_1 n_2]$ under the null hypothesis $H_0: F = G$.

2.2. Proposed statistic

Suppose K_1 and K_2 are independent non-negative random variables taking on integer values. Further, suppose a_ℓ and b_ℓ^2 denote $E(K_\ell)$ and $\text{var}(K_\ell)$, for $\ell = 1, 2$, respectively. Let X_i ($i = 1, \dots, K_1$) be observations from group 1 and Y_j ($j = 1, \dots, K_2$) be observations from group 2, where K_1 and K_2 are random sample sizes. Allowing ties, we define the random-sum Wilcoxon statistic as

$$W_r = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} U_{ij},$$

where U_{ij} is as defined earlier. For continuous data, the probability of having a tie is 0.

The expectation of W_r is given by

$$E(W_r) = E[E(W_r | K_1, K_2)] = a_1 a_2 \xi_1,$$

and the variance of W_r is given by

$$\begin{aligned} \text{var}(W_r) &= E \left[\text{var} \left(\sum_{i=1}^{K_1} \sum_{j=1}^{K_2} U_{ij} \right) \middle| K_1, K_2 \right] + \text{var}(K_1 K_2 \xi_1) \\ &= a_1 a_2 \xi_1 (1 - \xi_1) + a_1 (b_2^2 + a_2^2 - a_2) (\xi_2 - \xi_1^2) + a_2 (b_1^2 + a_1^2 - a_1) (\xi_3 - \xi_1^2) + (b_1^2 b_2^2 + b_1^2 a_2^2 + a_1^2 b_2^2) \xi_1^2, \end{aligned} \quad (3)$$

where ξ_1 , ξ_2 and ξ_3 are as defined in the preceeding section.

Suppose the distribution of K_i depends on a parameter τ_i and has an asymptotic expression as $\tau_i \rightarrow \infty$, for $i=1,2$. As an example, we could consider $K_i \sim \text{Binomial}(n_i, p_i)$, wherein the parameter τ_i is n_i and we have in this case K_i to be asymptotically distributed as

$N(a_i = n_i p_i, b_i^2 = n_i p_i (1 - p_i))$. Another example is when we consider $K_i \sim \text{Poisson}(\lambda_i)$, wherein the parameter τ_i is the same as λ_i and that we have in this case K_i to be distributed asymptotically as $N(a_i = \lambda_i, b_i^2 = \lambda_i)$. Let

$$U(x, y) = \begin{cases} 1 & \text{if } x > y \\ \frac{1}{2} & \text{if } x = y, \\ 0 & \text{if } x \leq y. \end{cases}$$

Now, let us define

$$T = W_r - K_1 K_2 \xi_1 \quad \text{and} \quad S = a_2 \sum_{i=1}^{K_1} \psi_1(X_i) + a_1 \sum_{j=1}^{K_2} \psi_2(Y_j),$$

where $\psi_1(x) = E[U(x, Y) - \xi_1]$ and $\psi_2(y) = E[U(X, y) - \xi_1]$. Further, let $S^* = \tau_1^{-3/2\rho} S$, and $T^* = \tau_1^{-3/2\rho} T$. To establish Lemma 1 and Theorem 1, we need the following assumptions:

- Condition 1.** K_i depending on a parameter τ_i has its asymptotic distribution as $\text{Normal}(a_i, b_i^2)$, as $\tau_i \rightarrow \infty$;
- Condition 2.** $a_2/a_1 \rightarrow \gamma$, for some finite non-zero constant γ , as $\tau_i \rightarrow \infty$;
- Condition 3.** $b_i^2/a_i \rightarrow \delta_i$, for some finite constant δ_i , as $\tau_i \rightarrow \infty$;
- Condition 4.** $a_i^\rho/\tau_i \rightarrow \lambda_i$, for some finite non-zero constant λ_i and $\rho > 1$, as $\tau_i \rightarrow \infty$.

Then, given the above regularity conditions, the following lemma states that S^* is a Hajek-like projection of T^* , and the proof of the lemma is given in the Appendix.

Lemma 1—Under Conditions 2–4, we have

$$E[(T^* - S^*)^2] \rightarrow 0$$

as $\tau_1, \tau_2 \rightarrow \infty$.

Thus, T^* has the same asymptotic distribution as S^* , which is the random summation of iid random variables. Robbins (1948a, 1948b) was the first to discuss the normality of random summation of iid random variables. Since his seminal work, theoretical properties of random summations have been studied by many authors. While Guiasu (1971) and Szasz (1972)

presented general conditions for the convergence, Gnedenko and Korolev (1996) provided a book-length account of the literature on random summations. Most of these existing methods consider the summation of independent random variables, while the problem considered here involves a random summation of dependent random variables. Thus, the existing results cannot be directly utilized here. By combining the lemma with the results of Robbins (1948a, b), the asymptotic normality of the proposed statistic can then be established, as stated in the following theorem.

Theorem 1—Under Conditions 1–4, $\tau_1^{-3/2\rho}(W_r - a_1 a_2 \xi_1)$ is asymptotically normal with mean 0 and variance

$$\lambda_1^3 \gamma^2 (\xi_2 - \xi_1^2) + \lambda_1^3 \gamma (\xi_3 - \xi_1^2) + (\delta_1 \lambda_1^3 \gamma^2 + \delta_2 \lambda_1^3 \gamma_1) \xi_1^2$$

as $\tau_1, \tau_2 \rightarrow \infty$.

The proof of this theorem is given in the Appendix. Since Lemma 1 states that T^* converges to S^* in probability, what is left is only to prove the normality of S^* . Because of random sample sizes, S^* consists of a summation of random number of independent random variables. Hence, by applying results on the sum of random number of variables of Robbins (1948a, b), the projection T^* can be shown to be asymptotically normal.

As mentioned in the preceding subsection, the Wilcoxon statistic is commonly used for testing $H_0: F = G$ against the alternative $H_a: F \neq G$. When the null hypothesis H_0 is true, it is evident that

$$P(X > Y) + \frac{1}{2}P(X = Y) = 0.5$$

indicating that there is no difference between the two groups. Using the asymptotic variance of W_r presented in Theorem 1, the power of the random-sum Wilcoxon test can be easily studied. Under the null hypothesis, the asymptotic variance in Eq. (3) can be simplified to an explicit form. In the absence of ties, we have $\xi_1 = \frac{1}{2}$ and $\xi_2 = \xi_3 = \frac{1}{3}$, and consequently we get $\text{var}_0(W_r)$ under the null hypothesis to be

$$\frac{1}{12}a_1 a_2^2 + \frac{1}{12}a_1^2 a_2 + \frac{1}{4}(b_1^2 b_2^2 + b_1^2 a_2^2 + a_1^2 b_2^2). \quad (4)$$

We now present some typical examples for K_1 and K_2 , and discuss the relationship between the proposed random-sum Wilcoxon statistic and the usual Wilcoxon rank-sum statistic.

2.2.1. Example 1—Suppose n_1 subjects are in group 1 and n_2 subjects are in group 2, where n_1 and n_2 are some fixed positive integers. Further, suppose a subject in group 1 (group 2) has the probability p_1 (p_2) of being observed. Since the observations on subjects are independent, the number of observable subjects K_1 in group 1 follows the Binomial (n_1, p_1) distribution. Similarly, the number of observable subjects K_2 in group 2 follows the Binomial (n_2, p_2) distribution. In this case, we see that n_1 and n_2 are the aforementioned τ_1 and τ_2 , and $\rho = 1$. If n_2/n_1 converges to some constant η , as $n_1, n_2 \rightarrow \infty$, Assumption 1 is satisfied. Since $a_2/a_1 = p_2/p_1$, and $b_i^2/a_i = 1 - p_i$, Conditions 2 and 3 are satisfied as well. If we set $\rho = 1$, it also follows that $a_i/n_i = p_i$, which means Assumption 4 is satisfied. Hence, Theorem 1 gives the asymptotic normality of W_r and the following specific result.

Corollary 1—If $n_2/n_1 \rightarrow \eta$ as $n_1, n_2 \rightarrow \infty$, $n_1^{-3/2}(W_r - n_1 n_2 p_1 p_2 / 2)$ is asymptotically normal with mean 0 and variance

$$p_1 p_2^2 \eta^2 (\xi_2 - \xi_1^2) + p_1^2 p_2 \eta (\xi_3 - \xi_1^2) + [p_1(1 - p_1) p_2^2 \eta^2 + \eta p_1^2 p_2 (1 - p_2)] \xi_1^2.$$

When $p_2 = 1$, which means that every subject in group 2 is observable, we have $K_2 = n_2$, and the asymptotic variance expression of W_r in this case becomes

$$n_1 n_2^2 p_1 (\xi_2 - \xi_1^2) + n_1^2 n_2 p_1^2 (\xi_3 - \xi_1^2) + n_1 n_2^2 p_1 (1 - p_1) \xi_1^2. \quad (5)$$

When we have both $p_1 = 1$ and $p_2 = 1$, W_r becomes the usual Wilcoxon rank-sum statistic in the presence of ties. The asymptotic normality of W_r and the asymptotic variance in this case become the same as those corresponding to the usual Wilcoxon statistic.

Under the null hypothesis $H_0: F = G$, $\text{var}_0(W_r)$ converges to

$$\frac{1}{12} n_1 n_2^2 p_1 (4 - 3p_1) p_2^2 + \frac{1}{12} n_1^2 n_2 p_1^2 p_2 (4 - 3p_2) + \frac{1}{4} m n p_1 (1 - p_1) p_2 (1 - p_2), \quad (6)$$

which further simplifies as

$$\frac{1}{12} \{n_1 n_2^2 p_1 (4 - 3p_1) + n_1^2 n_2 p_1^2\} \quad (7)$$

when $p_2 = 1$.

2.2.2. Example 2—Suppose in a clinical trial, subjects enter the study serially and are randomly assigned to two treatment arms. Suppose the numbers of observations in two arms, K_1 and K_2 , follow Poisson distributions with parameters τ_1 and τ_2 , respectively. In this case, these τ_1 and τ_2 are the aforementioned τ_1 and τ_2 , and the assumptions in Theorem 1 are satisfied when τ_1/τ_2 converges to some finite constant as $\tau_1, \tau_2 \rightarrow \infty$. So, we have $W_r - \tau_1 \tau_2 \xi_1$ to be asymptotically normal with mean 0 and variance

$$\tau_1 \tau_2^2 (\xi_2 - \xi_1^2) + \tau_1^2 \tau_2 (\xi_3 - \xi_1^2) + (\tau_1 \tau_2^2 + \tau_1^2 \tau_2) \xi_1^2.$$

To test $H_0: F = G$ against the alternative $H_a: F \neq G$, we use the test statistic

$$z = \frac{W_r - \frac{1}{3} \tau_1 \tau_2}{\sqrt{\text{var}_0(W_r)}},$$

where $\text{var}_0(W_r)$ is given by (obtained upon substituting $a_i = b_i = \tau_i$ in (4) for $i=1,2$)

$$\text{var}_0(W_r) = \frac{1}{3} \tau_1 \tau_2^2 + \frac{1}{3} \tau_1^2 \tau_2 + \frac{1}{4} \tau_1 \tau_2.$$

3. Application to different types of diagnostic data

In this section, the statistic proposed in the preceding section is applied to a medical imaging study and in the analysis of a biomarker data.

3.1. LROC studies

A problem of great interest in medical imaging studies is to learn whether an observer has good performance on localizing lesions (Kallergi et al., 2006). A good observer, either a computer algorithm or a radiologist, should be accurate in diagnosing medical images and localizing lesions. In a lesion localization study, an observer is asked to identify and mark suspicious lesion locations on images in radiological studies. In addition, the observer provides a rating for every lesion location according to the confidence level in marking the location. In some studies, a radiologist is asked to mark the most suspicious malignant location in each image, and to give rating to the image according to the radiologist's belief of the lesion existence. In such studies, each image has one rating and one marked location. In some other studies, a radiologist freely marks multiple lesion locations and gives ratings for all marked locations, and consequently each image then has several marked locations and associated ratings. The highest malignancy rating of all marked locations in each image is then used as the rating of the image, and the location associated with the rating is used as the observer's localization.

Suppose in an imaging study, an observer is asked to find the most suspicious lesion location in an image which is either normal or abnormal and give rating to that lesion. That rating is then used for both the lesion and the image. Let X_i denote the rating of a correctly localized lesion in the i -th image (for $i = 1, \dots, K_1$), where K_1 is random denoting the number of abnormal images with correct localization out of the total n_1 abnormal images. Let Y_j denote the rating of a suspicious location in the j -th normal image for $j = 1, \dots, n_2$, where n_2 is the number of normal images. Further, suppose $X_i \sim F$, $Y_j \sim G$ and $K_1 \sim \text{Binomial}(n_1, p_1)$, where F and G are cumulative distribution functions and p_1 is the probability of correctly localizing a lesion in an abnormal image. Starr et al. (1975) introduced a location response receiver operating characteristic (LROC) curve, which combines the fraction of abnormal images detected with lesion localized correctly (the ratio of the number of correctly localized and marked abnormal images to the number of abnormal images) and FPRs. The area under the LROC curve is a useful quantity for evaluating the performance of a radiologist on his/her ability to correctly localize a lesion on an abnormal image. Based on Swensson's (1996) parametric modeling, the LROC curve is given by

$$LROC(u) = p_1 [1 - F\{G^{-1}(1 - u)\}],$$

where $0 \leq u \leq 1$. So, the empirical LROC curve is given by

$$\widehat{LROC}(u) = \widehat{p}_1 [1 - \widehat{F}\{\widehat{G}^{-1}(1 - u)\}], \quad (8)$$

where $\widehat{F}$ and $\widehat{G}$ are empirical functions of F and G . Based on the data, (8) can be expressed as

$$\widehat{LROC}(u) = \frac{K_1}{n_1} \sum_{i=1}^{K_1} I(X_i \geq Y_{[n_2(1-u)]}),$$

where $\lfloor z \rfloor$ indicates the greatest integer less than or equal to z . Therefore, a non-parametric estimator for the area under the LROC curve is given by

$$A_L = \int_0^1 \widehat{LROC}(u) du \frac{1}{n_1 n_2} = \sum_{i=1}^{K_1} \sum_{j=1}^{n_2} U_{ij},$$

which is indeed a random-sum Wilcoxon statistic (multiplied by a constant). The variance of A_L is given by (5). An A_L of $p_1/2$ means that the observer is performing the same as random guessing. For testing whether an observer can distinguish an abnormal image from a normal image, we can use the test statistic, $(A_L - p_1/2) / \sqrt{\text{var}_0(A_L)}$, where an explicit expression for $\text{var}_0(A_L)$ is as given in (7).

3.2. Biomarker data with limit of detection

The result derived in Corollary 1 can be used to estimate the area under ROC curves (AUC) in the presence of limit of detection (LOD). LOD is due to equipment limitations (Schisterman et al., 2006). LOD in diagnostic marker studies has attracted considerable attention recently; see Perkins et al. (2007) and Vexler et al. (2008). Their methods are parametric in nature, and may not perform well if the models are misspecified. Suppose $X_1, \dots, X_{K_1}$ denote non-zero marker measurements for the diseased group, $X_{K_1+1}, \dots, X_{n_1}$ denote zero measurements for diseased subjects who are not observable, $Y_1, \dots, Y_{K_2}$ denote non-zero measurements for the non-diseased, and $Y_{K_2+1}, \dots, Y_{n_2}$ denote zero measurements for non-diseased subjects. If the LOD is d , then we have $p_1 = P(X > d)$ to be the probability of observing an actual non-zero observation for X , and $p_2 = P(Y > d)$ to be the probability of observing an actual non-zero observation for Y . Then, Schisterman et al. (2006) have given the AUC as

$$\theta = (1 - p_2) \{1 - \frac{1}{2}(1 - p_1)\} + p_1 p_2 P(X > Y)$$

and then proceeded to estimate $P(X > Y)$ from the random samples of non-zero measurements based on the assumption of normality for the distributions. However, by using the proposed statistic, we can present a non-parametric estimator for θ as

$$\widehat{\theta} = \left(1 - \frac{K_2}{n_2}\right) \left(\frac{1}{2} + \frac{K_1}{2n_1}\right) + \frac{1}{n_1 n_2} W_r, \quad (9)$$

where W_r is the random-sum Wilcoxon statistic

$$W_r = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} U_{ij}$$

as introduced earlier in Section 2.2. Since K_1 and K_2 are approximately independent of W_r as n_1 and n_2 get large, a modification of Corollary 1 yields $\sqrt{n_1}(\widehat{\theta} - \theta)$ to be asymptotically normal with mean zero and variance

$$\frac{1}{4}[p_2(1-p_2)(1+p_1)^2 + \eta p_1(1-p_1)(1-p_2)^2] + p_2 p_1^2 (\xi_2 - \xi_1^2) + \eta p_2^2 p_1 (\xi_3 - \xi_1^2) + [p_2(1-p_2)p_1^2 + \eta p_2^2 p_1 (1+p_1)] \xi_1^2$$

when $n_2/n_1 \rightarrow \eta$ as $n_1, n_2 \rightarrow \infty$

4. Simulations and results

In the first simulation study, we evaluate the accuracy and robustness of the proposed statistic as well as the finite-sample performance of the asymptotic variance. In the second simulation study, we apply the proposed random-sum Wilcoxon statistic as well as the usual Wilcoxon statistic to simulated data in the presence of ties, and compare the two methods based on their simulated Type-I error rates.

4.1. Finite-sample performance of W_r and its variance

We simulated continuous data from the following models:

- bi-normal model: $X \sim N(0,1)$, $Y \sim N(1.5,2)$;
- bi-lognormal model: $X \sim \text{Lognormal}(0,1)$, $Y \sim \text{Lognormal}(1,2)$;
- bi-exponential model: $X \sim \text{Exponential}(1)$, $Y \sim \text{Exponential}(4)$.

For each model, we simulated 1000 data sets with sample sizes $n_1 = n_2 = (50, 100)$. The probability of observing a subject from either group was set to be 0.4 or 0.8. For each data set, we first simulated the number of observable subjects, I_ℓ from Binomial (n_ℓ, p_ℓ) for group ℓ and then I_ℓ observations from group ℓ were simulated. To generate ties, the simulated continuous data were divided into rating data with R categories by using cutoff points $-\infty = c_1 < c_2 < \dots < c_R < c_{R+1} = \infty$. For the bi-normal data, we chose $c_i = \Phi^{-1}\{0.01 + (i-2)/(R-2)\} + 0.75$, $i = 2, \dots, R$. For the bi-lognormal data, we chose $c_i = \exp(\Phi^{-1}\{0.01 + (i-2)/(R-2)\} + 0.75)$, and similarly for the bi-exponential data, we chose $c_i = \Phi^{-1}\{0.01 + (i-2)/(R-2)\} + 2$. A subject is assigned a rating r if the observation falls between c_r and c_{r+1} for $r = 1, \dots, R$. We used $R=5$ and 8 in our simulation study. Based on 1000 data sets under each combined setting, we calculated the bias and mean squared error (MSE) of W_r . We also determined the coverage percentages of 95% confidence intervals based on the asymptotic variance in (5). These simulation results are all presented in Table 2, and from these results it is clear that the proposed random-sum Wilcoxon statistic has small values for bias and MSE under these settings. Table 2 shows that the bias and MSE for $p_1 = p_2 = 0.8$ are larger than the corresponding results for $p_1 = p_2 = 0.4$. A possible explanation is that when p_1 and p_2 are small, the resulting true measure $p_1 p_2 \xi_1$ gets smaller. This may lead to smaller bias. The coverages based on the asymptotic variance are close to the nominal level even for sample sizes as small as 50, suggesting a good convergence of the asymptotic normality of the proposed statistic.

4.2. Simulated Type-I error rates of W_r and W

We also carried out simulation studies for comparing the Type-I error rates of the proposed random-sum Wilcoxon statistic and the usual Wilcoxon statistic, especially when the assumption of fixed sample sizes of the latter is violated. For this purpose, we used the model settings bi-normal model: $X, Y \sim N(0,1)$, bi-lognormal model: $X, Y \sim \text{Lognormal}(0,1)$, and bi-exponential model: $X, Y \sim \text{Exponential}(4)$. We applied the proposed statistic with its variance expression in (6) to these simulated data, and also the usual Wilcoxon statistic with its variance expression in (2) by assuming that the sample sizes are fixed. With the nominal Type-I error rate fixed as $\alpha = 0.05$, the rejection boundaries of the test statistics were determined, and then the simulated Type-I error rates were computed as the number of rejections divided by the total number of simulations (taken to be 1000). We used (p_1, p_2) to

be (0.9, 0.4) and (0.7, 0.9). Once again, the rating categories we used were $R=5$ and 8, and sample sizes were taken as $n_1 = n_2 = (50, 100)$. Table 3 presents the simulated Type-I error rates for the proposed W_r statistic as well as the usual Wilcoxon statistic W . The statistic W_r possesses stable Type-I error rates (close to the nominal level of 5%) for all the settings considered, while the Type-I error rates of the usual Wilcoxon statistic turn out to be lower than the nominal level in many cases and especially so in the case of exponential model. A possible explanation for such lower error rates is that the variance of the usual Wilcoxon statistic does not capture the variation in the random sample size, and that the resulting sampling distribution of the statistic under the null tends to have a higher peak and flatter tails than the true one.

5. Illustrative examples

In this section, we first apply the proposed statistic to a LROC imaging data and then to a cancer diagnostic biomarker data with LOD in order to illustrate the distribution-free method developed here based on the random-sum Wilcoxon statistic.

5.1. LROC imaging data

The numbers of abnormal and normal images, $n_1 = n_2 = 46$, are known in the chest film example mentioned earlier. A natural assumption for the random number of correctly localized images, K_1 , is that $K_1 \sim \text{Binomial}(n_1, p_1)$. The realization of K_1 in this particular example is $K_1 = 23$. The non-parametric area under the LROC curve is then $A_L = 0.3922$ with its variance $\text{var}(A_L) = 0.0047$. Suppose under the null hypothesis of no localization accuracy, the hypothesized value is $A_L = p_1/2$. An investigator then needs to specify the null value of p_1 for the hypothesis testing problem. If we suppose $p_1 = \frac{1}{2}$ in this case, we find the test statistic to be

$$z = \frac{A_L - \frac{1}{4}}{\sqrt{\text{var}_0(A_L)}} = 2.73,$$

with the corresponding p -value of 0.0031. From this, we can conclude that there is sufficient evidence to say that the observer is more accurate than random guessing in localizing lesions and in discriminating abnormal images from normal images.

5.2. Cancer biomarker data with LOD

Cancer diagnostic data were used by Goddard and Hinberg (1990) and Lloyd (1998) to illustrate the drawbacks of using parametric bi-normal models to fit ROC curves. The data were collected by taking seven diagnostic tests on $n_1 = 135$ cancer patients and $n_2 = 218$ non-cancer patients. Neither the test results nor their log-transformed values seem to be distributed as normal, as noted by Lloyd (1998) and Goddard and Hinberg (1990). Although these biomarker data are fully observed, we assigned zero values to those observations which were less than certain threshold values so that the data look similar to typical biomarker data with LOD. Thresholds were chosen so that 10% of observations for each biomarker were trimmed. Since the assumption of normal distributions is not appropriate in this case, the parametric methods of Schisterman and Rotnitzky (2001) and Perkins et al. (2007) cannot be used here. We applied the modified statistic in (9) and its variance in (10) to each biomarker in the data, and the results are presented in Table 4. The table reveals that all the biomarkers have better accuracy than random guessing and that Marker B has the best accuracy among all markers.

6. Concluding remarks

In this paper, we have proposed a random-sum Wilcoxon statistic and discussed its asymptotic properties. The statistic is shown to have important applications in diagnostic studies wherein one is interested in the localization accuracy and in biomarker studies in which observations are truncated due to limit of detection. The proposed statistic can also be applied to free-response receiver operating characteristic (FROC) data, which is similar to the second LROC data setting described at the beginning of Section 3. Unlike LROC studies that require raters to mark only one abnormal location, raters freely mark suspicious locations in FROC studies. Parametric approaches for this situation have been provided by Chakraborty (1989) and Chakraborty and Winter (1990). For applying the random-sum Wilcoxon statistic, we consider the maximum rating of all the ratings in an abnormal image as the rating of that image, and the location associated with this rating as the solely marked location in the image. For normal images, we simply pick the maximum rating out of all ratings in an image. By doing this for all the images, we modify FROC data into LROC-like data. The proposed statistic then becomes suitable for analyzing such a modified FROC data and can provide an accurate estimate of the rater performance. In terms of testing of hypotheses, it has also been shown that the statistic can be used to determine if an observer is accurate in localizing lesions or not.

The method presented here could be used to estimate the expected power when the sample sizes can only be approximated a priori. Also, the Wilcoxon-type statistic proposed here for the two-sample comparison could be generalized to the k -sample comparison. In addition, the proposed statistic can be used to test whether an observer or a biomarker is better than a random guesser. It can be extended to compare two diagnostic tests. Work is currently under progress on these issues and we hope to report these findings in a future paper.

Acknowledgments

The authors thank Xiao-Hua Andrew Zhou and Bill Rosenberger for making some valuable suggestions. The authors also thank two anonymous referees for their constructive comments and useful suggestions. The project described here was supported by Award Number R15CA150698 from the National Cancer Institute under the American Recovery and Reinvestment Act of 2009. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

- Chakraborty D, Winter L. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*. 1990; 174(3):873–881. URL: <<http://radiology.rsna.org/cgi/content/abstract/174/3/873>>. [PubMed: 2305073]
- Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (fROC) data. *Medical Physics*. 1989; 16(4):561–568. URL: <<http://link.aip.org/link/?MPH/16/561/1>>. [PubMed: 2770630]
- Gnedenko, BV.; Korolev, VY. *Random Summation: Limit Theorems and Applications*. CRC Press; 1996.
- Goddard MJ, Hinberg I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. *Statistics in Medicine*. 1990; 9:325–337. [PubMed: 2343225]
- Guiasu S. On the asymptotic distribution of the sequences of random variables with random indices. *The Annals of Mathematical Statistics*. 1971; 42(6):2018–2028. URL: <<http://www.jstor.org/stable/2240129>>.
- Hanley J, McNeil B. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
- Kallergi M, Lucier BJ, Berman CG, Hersh MR, Kim JJ, Szabunio MS, Clark RA. High-performance wavelet compression for mammography: localization response operating characteristic evaluation.

- Radiology. 2006; 238 (1):62–73. URL: <<http://radiology.rsnajnl.org/cgi/content/abstract/238/1/62>>. [PubMed: 16373759]
- Lehmann, EL. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day; San Francisco: 1998.
- Lloyd CJ. Using smoothed receiver operating characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical Association. 1998; 93:1356–1364.
- Perkins NJ, Schisterman EF, Vexler A. Receiver operating characteristic curve inference from a sample with a limit of detection. American Journal of Epidemiology. 2007; 165(3):325–333. [PubMed: 17110640]
- Robbins H. The asymptotic distribution of the sum of a random number of random variables. Bulletin of the American Mathematical Society. 1948a; 54(12):1151–1161.
- Robbins H. On the asymptotic distribution of the sum of a random number of random variables. Proceedings of the National Academy of Sciences of the United States of America. 1948b; 34(4): 162–163. URL: <<http://www.pnas.org/content/34/4/162.short>>. [PubMed: 16588799]
- Schisterman E, Reiser B, Faraggi D. ROC analysis for markers with mass at zero. Statistics in Medicine. 2006; 25(4):623–638. [PubMed: 16345033]
- Schisterman E, Rotnitzky A. Estimation of the mean of a k-sample u-statistic with missing outcomes and auxiliaries. Biometrika. 2001; 88(3):713–725. URL: <<http://www.jstor.org/stable/2673441>>.
- Seltzer S, Swensson R, Nawfel R, Lentini J, Kazda I, Judy P. Visualization and detection-localization on computed tomographic images. Investigative Radiology. 1991; 26(4):285–294. [PubMed: 2032815]
- Starr SJ, Metz CE, Lusted LB, Goodenough DJ. Visual detection and localization of radiographic images. Radiology. 1975; 116:533–538. [PubMed: 1153755]
- Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Medical Physics. 1996; 23:1709–1725. [PubMed: 8946368]
- Szasz D. Limit theorems for the distributions of the sums of a random number of random variables. The Annals of Mathematical Statistics. 1972; 43(6):1902–1913. URL: <<http://www.jstor.org/stable/2240201>>.
- Vexler A, Liu A, Eliseeva E, Schisterman EF. Maximum likelihood ratio tests for comparing the discriminatory ability of biomarkers subject to limit of detection. Biometrics. 2008; 64(3):895–903. [PubMed: 18047527]
- Zhou, XH.; McClish, DK.; Obuchowski, N. Statistical Methods in Diagnostic Medicine. John Wiley & Sons; New York: 2002.

Appendix

Proof of Lemma 1

Since T^* and S^* have the same expectation 0, the second moment of $T^* - S^*$ need to go to 0 in order for T^* to have the same asymptotic distribution as S^* . To establish this, we need to show that $E[(T^* - S^*)^2] = \text{var}[(T^*)^2] - \text{var}[(S^*)^2]$ first. For this purpose, let us write

$$E[(T^*)^2] = E[(T^* - S^*)^2] + 2E[(T^* - S^*)S^*] + E[(S^*)^2].$$

We have

$$E[(T^* - S^*)S^*] = a_2 E \left[\sum_{i=1}^{K_1} \psi_1(X_i)(T^* - S^*) \right] + a_1 E \left[\sum_{j=1}^{K_2} \psi_2(Y_j)(T^* - S^*) \right]. \quad (\text{A.1})$$

For (A.1) to be zero, it is sufficient to show that

$$E[\psi_1(X_i)(T^* - S^*)] = 0 \quad (\text{A.2})$$

and

$$E[\psi_2(Y_j)(T^* - S^*)] = 0. \quad (\text{A.3})$$

The left-hand side of (A.2) can be written as

$$\tau_1^{-3/2\rho} E \left\{ E \left(E \left[\psi_1(X_i) \left(T - a_2 \psi_1(X_i) - a_2 \sum_{i' \neq i}^{K_1} \psi_1(X_{i'}) - a_1 \sum_{j=1}^{K_2} \psi_2(Y_j) \right) \middle| K_1, K_2 \right] \middle| X_i \right) \right\}.$$

Based on the definition of $\psi_1(\cdot)$, $\tau_1^{-3/2\rho} E[(T - a_2 \psi_1(X_i)|X_i)]$ is 0. Since X_i is independent from $X_{i'}$ and Y_j , we have $\tau_1^{-3/2\rho} E[a_2 \psi_1(X_{i'})|X_i]$ and $\tau_1^{-3/2\rho} E[a_1 \psi_1(Y_j)|X_i]$ to be both 0. It is then clear that (A.2) is true. Similarly, we see (A.3) to be true. Combining the results in (A.2) and (A.3), we get

$$E[(T^* - S^*)^2] = \text{var}[(T^*)^2] - \text{var}[(S^*)^2] \quad (\text{A.4})$$

as $\tau_1, \tau_2 \rightarrow \infty$. The variance of T is given by

$$\begin{aligned} \text{var}(W_r - K_1 K_2 \xi_1) &= E[\text{var}(W_r) - 2\text{cov}(W_r, K_1 K_2 \xi_1) + \text{var}(K_1 K_2 \xi_1)|K_1, K_2] \\ &= a_1 a_2 \xi_1 (1 - \xi_1) + a_1 (b_2^2 + a_2^2 - a_2)(\xi_2 - \xi_1^2) + a_2 (b_1^2 + a_1^2 - a_1)(\xi_3 - \xi_1^2) - 2E[\text{cov}(W_r, K_1 K_2 \xi_1)|K_1, K_2]. \end{aligned}$$

Since $E(W_r K_1 K_2 \xi_1|K_1, K_2) = K_1^2 K_2^2 \xi_1^2$, it is then easy to show that $E[\text{cov}(W_r, K_1 K_2 \xi_1)|K_1, K_2] = 0$. Thus, as $\tau_1, \tau_2 \rightarrow \infty$, the asymptotic variance of T^* is given by

$$\text{var}(T^*) = \lambda_1^3 \gamma^2 (\xi_2 - \xi_1^2) + \lambda_1^3 \gamma (\xi_3 - \xi_1^2), \quad (\text{A.6})$$

which is the same as the variance of S^* given by

$$\tau_1^{-3/2\rho} \left[\text{var} \left(a_2 \sum_{i=1}^{K_1} \psi_1(X_i) + a_1 \sum_{j=1}^{K_2} \psi_2(Y_j) \right) \right] = \lambda_1^3 \gamma^2 (\xi_2 - \xi_1^2) + \lambda_1^3 \gamma (\xi_3 - \xi_1^2). \quad (\text{A.7})$$

From these variances, we see that (A.4) converges to zero as $\tau_1, \tau_2 \rightarrow \infty$.

Proof of Theorem 1

In order to show the asymptotic normality of W_n , we start with the asymptotic property of the Hajek-like projection S , which consists of two random sums

$$\sum_{i=1}^{K_1} \psi_1(X_i) \quad (\text{A.8})$$

and

$$\sum_{j=1}^{K_2} \psi_2(Y_j). \quad (\text{A.9})$$

Robbins (1948b, a) has studied asymptotic properties of the sum of a random number of random variables. Let $\sigma_1^2 = a_1 \text{var}(\psi_1(X_i) + b_1^2 E(\psi_1(X_i)))$; then, $\sigma_1^2 = a_1(\xi_2 - \xi_1^2)$. Similarly, we obtain $\sigma_2^2 = a_2 \text{var}(\psi_2(Y_j) + b_2^2 E(\psi_2(Y_j)))$ as $\sigma_2^2 = a_2(\xi_3 - \xi_1^2)$. It follows that Assumption 3 is a simplified condition of those in Corollary 2 of Robbins (1948a). It can also be seen that Assumption 3 satisfies Condition (C0) of Guiasu (1971). Hence, the results in Robbins (1948a) and Guiasu (1971) yield (A.8) to be asymptotically normal with mean 0 and variance σ_1^2 , as $\tau_1 \rightarrow \infty$. Similarly, (A.9) is asymptotically normal with mean 0 and variance σ_2^2 . The limit distribution of S is, therefore, normal with mean 0 and variance

$$a_2^2 a_1 (\xi_2 - \xi_1^2) + a_1^2 a_2 (\xi_3 - \xi_1^2). \quad (\text{A.10})$$

Thus, S^* converges to a mean zero normal distribution with variance as in (A.7). Lemma 1 yields that T^* has the same limiting distribution. Since under Assumption 1 both (A.8) and K_1 are asymptotically normal, when $\tau_1 \rightarrow \infty$, these uncorrelated normal random variables are approximately independent. Similar arguments can be used to claim that (A.9) and K_2 are approximately independent normal variables, when $\tau_2 \rightarrow \infty$. From these facts, the result in Theorem 1 follows.

Table 1

Ratings and localizations on liver CT images.

Rating	<u>Abnormal images</u>		Normal images
	CL	IL	
1	3	9	19
2	3	2	8
3	0	1	8
4	6	4	8
5	5	5	3
6	6	2	0

CL denotes correct localization; IL denotes incorrect localization.

Table 2

Bias (in %), RMSE, and coverage probability.

	n_1 (n_2)	$p_1 = p_2 = 0.4$			$p_1 = p_2 = 0.8$		
		Bias	MSE	Coverage (%)	Bias	MSE	Coverage (%)
Normal	50	-0.31	0.03	93.60	-0.95	0.06	95.20
	100	-0.23	0.02	93.70	-1.00	0.04	94.70
LN	50	-0.29	0.03	95.60	-1.03	0.06	95.80
	100	-0.22	0.02	93.90	-0.90	0.04	94.10
Exp	50	-0.60	0.03	92.90	-1.30	0.05	92.80
	100	-0.65	0.02	93.30	-1.23	0.04	93.20

LN denotes lognormal, Exp denotes exponential.

Table 3

Simulated Type-I error rates (in %).

n_1 (n_2)	$p_1=0.9, p_2=0.4$				$p_1=0.7, p_2=0.9$				
	W_r		W		W_r		W		
	$R=5$	8	$R=5$	8	$R=5$	8	$R=5$	8	
50	Normal	5.10	5.40	2.90	3.80	5.30	5.40	2.60	4.10
	LN	4.30	5.40	4.50	4.40	5.30	3.50	3.60	4.50
	Exp	5.10	4.10	1.80	2.80	4.30	5.40	2.00	2.70
100	Normal	5.00	4.80	3.10	3.40	3.80	4.70	3.10	3.00
	LN	4.30	5.00	4.20	3.80	4.70	3.70	3.80	4.20
	Exp	5.00	5.60	1.90	3.00	4.60	5.20	2.50	2.70

The rejection rate based on 1000 realizations. The 95% prediction interval is $(5.0\% \pm 1.4\%)$.

Table 4

Estimated AUCs for biomarker data with LOD.

	$\hat{\theta}$	$var(\hat{\theta})$	95% CI for $\hat{\theta}$
Marker A	0.7159	0.0011	(0.6521, 0.7797)
Marker B	0.7662	0.0008	(0.7109, 0.8214)
Marker C	0.7431	0.0009	(0.6855, 0.8006)
Marker D	0.7645	0.0008	(0.7107, 0.8183)
Marker E	0.7528	0.0008	(0.6961, 0.8094)
Marker F	0.7479	0.0008	(0.6922, 0.8036)
Marker G	0.7565	0.0009	(0.6987, 0.8143)