ON A TWO-SAMPLE NON-PARAMETRIC TEST IN THE CASE THAT TIES ARE PRESENT

By Hirosi Hudimoto

(Received Nov. 25, 1959)

1. In this paper, we shall be concerned with a two-sample non-parametric test in the case that ties are present and we shall consider the statistic $\frac{\hat{h}(F, G)}{(1-\hat{p})^2}=\frac{\sum_i[\hat{F}_i\hat{g}_i-\hat{G}_i\hat{f}_i]}{(1-\hat{p})^2}$ for any two compound distributions $F(x)=f_0+(1-f_0)F^*(x)$ and $G(x)=g_0-(1-g_0)G^*(x)$, where $\hat{F}_i$, $\hat{G}_i$, $\hat{f}_i$, $\hat{g}_i$ and $\hat{p}$ are the empirical distributions and relative frequencies corresponding to the theoretical ones defined in sections 2 and 3. Then a probablistic inequality between the empirical result and the theoretical will be given in section 3. In the appendix, the necessary relation for this evaluation is derived.

Such testing problems as we shall try to investigate in this paper can frequently occur in statistical investigations related to sociology.

Example. We have a result of the past five weeks obtained with respect to whether the individual involved in a random sample had bought weekly magazines with the specified kinds of names or not. Then we shall be interested in their sales and the aspects of continuing purchase of each individual at the same time.

This example will be offer a typical model about our pourpose of this paper.

2. In the part of [2] on the Wilcox test, J. Putter described as a known result that when $F(x)$ and $G(x)$ are any two continuous distributions, we have, from Mann-Whitney [1], if $F\equiv G$,

$$ES_{nm}=\frac{n(n+m+1)}{2}=\mu_{nm}$$

and in general

$$ES_{nm}=\mu_{nm}+nm\theta\ ,$$

$$\theta=\theta(F, G)=P(X>Y)-\frac{1}{2}=\int_{-\infty}^{\infty}G(t)dF(t)-\frac{1}{2}\ ,$$

where S_{nm} shows the sum of the ranks assigned to the X's and, n and m are the sample sizes of independent observations from F and G,

respectively. He also showed using Lemma 5.1 of Lehmann [3] in the case where F and G are discontinuous distributions that we have

$$ES_{nm}=\mu_{nm}+\theta^*{}_{nm}$$

$$\theta^*=P(X>Y)+\frac{1}{2}P(X=Y)-\frac{1}{2}\ .$$

Even if the assumption that all the samples involved are drawn from continuous distributions is satisfied, this assumption frequently is not realistic because of limitations on the precision of measurement or because of grouping data for the sake of conveniences of numerical treatment. Therefore we shall propose to replace the above-described θ^* by an empirical result.

We shall assume the observations of sizes n and m on the continuous distributions F and G are grouped into the classes of a finite number. Denote the midpoint of the i-th class by ξ_i and set

$$\hat{f}_i=\text{emp. } P\{X=\xi_i\} \text{ and } \hat{g}_i=\text{emp. } P\{Y=\xi_i\}\ ,$$

where X and Y denote the random variables with F and G, respectively, and the sign emp. means the empirical corresponding to the theoretical ones. We look upon as the frequencies concentrate to the midpoint in the each class for the sake of concise description.

In the following, we shall mainly be concerned with the statistic

$$\text{emp. } P\{X>Y\}+\frac{1}{2}\text{ emp. } P\{X=Y\}-\frac{1}{2}$$

and denote this statistic by $\hat{h}(F, G)$. However, this quantity is seen by a slight modification to be written as follows:

$$\hat{h}(F, G)=\sum_i [\hat{G}_i\hat{f}_i-\hat{F}_i\hat{g}_i]/2$$

where $\hat{F}_i$ and $\hat{G}_i$ denote the empirical cumulative distributions. The following relations are evident,

$$\hat{h}(F, G)=0 \qquad \text{for } \hat{F}\equiv\hat{G}$$

and

$$\hat{h}(F, G)=-\hat{h}(G, F)\ .$$

Thus the statistic $\hat{h}(F, G)$ will be suitable for testing the hypotheses $F=G$ against the alternative $F<G$.

3. In this section, we shall consider the testing problem that

applies $\hat{h}(F, G)$ discussed in the preceding section to two compound distribution functions

$$F(t)=\begin{cases} f_0 & \text{for } X\leqq 0 \\ f_0+(1-f_0)F^*(t) & \text{for } X>0 \end{cases}$$

$$G(t)=\begin{cases} g_0 & \text{for } Y\leqq 0 \\ g_0+(1-g_0)G^*(t) & \text{for } Y>0 . \end{cases}$$

Such cases frequently appear in statistical investigations related to sociology; for instance, the quantities referring to the number of persons who regularly subscribe to the newspapers and quantities referring to the manner of their reading.

Consider non-negative random variables X and Y with any compound distribution functions $F(x)$ and $G(y)$ as described above. Let us assume that the random samples of equal size N are taken from the populations characterized by these distributions and the real observations are grouped into the common finite classes denoted by the midpoints ξ_i's, and arranged in ascending order of magnitude on ξ_i's; i.e.

$$\hat{f}_i=\text{emp. } P\{X=\xi_i\} , \qquad \hat{g}_i=\text{emp. } P\{Y=\xi_i\} ,$$

and

$$0=\xi_0<\xi_1<\xi_2\cdots<\xi_k ,$$

$$\hat{F}_i=\sum_{\alpha=0}^{i}\hat{f}_\alpha , \qquad \hat{G}_i=\sum_{\beta=0}^{i}\hat{g}_\beta , \qquad i=0, 1, 2, \cdots, k ,$$

where the sign $\wedge$ means the empirical result corresponding to the population characteristic. Therefore, the test statistic for our two sample problem is expressed as follows:

$$\hat{h}(F, G)=\sum_{i=0}^{k}[\hat{G}_i\hat{f}_i-\hat{F}_i\hat{g}_i]/2 .$$

In the following we shall confine ourselves to the case where $F(x)$ and $G(x)$ are continuous distributions, because the resulting relation (*) also is valid for the discrete one. That is, we assume that

$$P\{X=\xi_0\}=f_0 , \qquad P\{Y=\xi_0\}=g_0$$

and

$$P\{X\subset(t, t+dt)\}=(1-f_0)f^*(t)dt , \qquad \text{for } X \text{ and } Y \text{ such as}$$
$$P\{Y\subset(t, t+dt)\}=(1-g_0)g^*(t)dt \qquad X\neq\xi_0 \text{ and } Y\neq\xi_0 ,$$

respectively. Then we have

$$\frac{\hat{h}(F, G)-(\hat{g}_0-\hat{f}_0)}{(1-\hat{g}_0)(1-\hat{f}_0)}=\hat{h}(F^*, G^*)\ ,$$

where $\hat{h}(F^*, G^*)=\sum_{i=1}^{k}[\hat{G}_i^*\hat{f}_i^*-\hat{F}_i^*\hat{g}_i^*]/2$.

We shall be interested in our two sample testing problem where the alternative consists of at least any one of the following situations:

$$(f_0<g_0 \text{ and } F^*<G^*) \text{ and } (f_0<g_0 \text{ or } F^*<G^*)\ .$$

Such a situation frequently appears in practical applications and is not uncommon. For instance, let $\hat{f}_0$ and $\hat{g}_0$ be the proportions of persons in sample who do not subscribe to the newspaper A and B and $\hat{F}^*$ and $\hat{G}^*$ the proportions of newspaper readers with same amount of reading. Then desirable situation for each of the newspaper publishers is that his newspapers is more subscribed to and more read than the other. Thus we can consider the above alternative.

Now, we denote the numbers of observations which take on the value ξ_0 in these samples of size N by n and m: i.e. $\hat{f}_0=\frac{n}{N}$, $\hat{g}_0=\frac{m}{N}$. At the first step in our testing procedure, we assume $f_0=g_0=p$ and adopt the test statistic

$$\frac{\hat{h}(F, G)}{(1-\hat{p})^2}\ ,$$

where $\hat{p}=\frac{n+m}{2N}$.

Now, denote $\frac{\hat{h}(F, G)-\varepsilon}{(1-\hat{p})^2}$, $|\hat{f}_0-p|+|\hat{g}_0-p|$, $(A\leqq\eta)\cup(B\leqq\varepsilon)$ and $(A\leqq\eta)\cup(B>\varepsilon)$ by A, B, S and $\bar{S}$ respectively, where $\eta>0, \varepsilon>0$. Then we have

$$P\{A\leqq\eta\}=P\{S\}+P\{\bar{S}\}$$

and

$$P\{\hat{h}(F, G)\leqq\varepsilon+\eta(1-p-\varepsilon)^2, B\leqq\varepsilon\}\leqq P\{S\}\ .$$

Therefore, the following relation holds

$$P\{A\leqq\eta\}\geqq P\{S\}\geqq P\{\hat{h}(F, G)\leqq\varepsilon+\eta(1-p-\varepsilon)^2, B\leqq\varepsilon\}$$

On the other hand, we have

$$\hat{h}(F, G) \leqq \sup(\hat{F}-\hat{G})$$

At the second step, if we assume $F^*\equiv G^*$, where $F^*(x)$ and $G^*(x)$ are the two continuous distribution functions defined before, we have

$$(*)\quad P\left\{\frac{\hat{h}(F, G)-\varepsilon}{(1-\hat{p})^2} \leqq \eta\right\} > \sum_{\left|\frac{m}{N}-p\right| \leqq \varepsilon/2} \frac{2(N!)}{m!(N-m)!} p^m \Bigg((1-p)^{N-m} - (\xi+\eta') \sum_{j=0}^{[N(1-\eta')]-m} \binom{N-m}{j} \left(\xi+\eta'+\frac{j}{N}\right)^{j-1} \times \left\{1-p-\left(\xi+\eta'+\frac{j}{N}\right)\right\}^{N-m-j}\Bigg) - 1$$

for any $\eta'>0, \varepsilon>0$ where $\eta'=\frac{1}{2}\{\varepsilon+\eta(1-p-\varepsilon)^2\}$, $\xi=\frac{m}{N}-p$ and the sign [ ] denotes the greatest integer such as is less than or equal to the number enclosed by the square brackets (see appendix on this derivation).

Then determine $\varepsilon>0$ *and* $\eta>0$ *such that*

$$\sum_{\left|\frac{m}{N}-p\right| \leqq \varepsilon/2} \frac{2(N!)}{m!(N-m)!} p^m \Bigg((1-p)^{N-m} - (\xi+\eta') \sum_{j=0}^{[N(1-\eta')]-m} \binom{N-m}{j} \left(\xi+\eta'+\frac{j}{N}\right)^{j-1} \times \left\{1-p-\left(\xi+\eta'+\frac{j}{N}\right)\right\}^{N-m-j}\Bigg) - 1 \geqq 1-\alpha$$

If $\dfrac{\hat{h}(F, G)-\varepsilon}{(1-\hat{p})^2} > \eta$, *we reject the hypothesis* ($f_0 \geqq g_0$ *and/or* $F^* \geqq G^*$) *or* ($f_0 \geqq g_0$ *and* $F^* \geqq G^*$) *with significant level smaller than* α *and accept the alternative that* ($f_0 < g_0$ *and* $F^* < G^*$) *or* ($f_0 < g_0$ *or* $F^* < G^*$), *because* $\hat{h}(F, G) = -\hat{h}(G, F)$ *for the case that* $f_0 > g_0$ *and* $F^* > G^*$, *and because there is no contribution of* $(f_0-g_0)>0$ *for the case that* $f_0 > g_0$ *or* $F^* > G^*$.

At the second step in our testing procedure, if we do not assume $F \equiv G$ is the partially modified result is obtained, i.e.

$$P\left\{\frac{\hat{h}(F, G)-\varepsilon}{(1-\hat{p})^2} - \frac{h(F, G)}{(1-p)^2} \leqq \eta \mid B \leqq \varepsilon\right\}$$

$$\geqq P\left\{\hat{h}(F, G)-h(F, G) \leqq \varepsilon+\eta(1-p-\varepsilon)^2 - h(F, G)\left\{2\left(\frac{\varepsilon}{1-p}\right)-\left(\frac{\varepsilon}{1-p}\right)^2\right\} \mid B \leqq \varepsilon\right\}$$

$$\geqq P\left\{\hat{h}(F, G)-h(F, G) \leqq \varepsilon+\eta(1-p-\varepsilon)^2 - \frac{1}{2}\left\{2\left(\frac{\varepsilon}{1-p}\right)-\left(\frac{\varepsilon}{1-p}\right)^2\right\} \mid B \leqq \varepsilon\right\}$$

where $h(F, G)$ is θ described in the section 1. On the other hand, we have

$$\hat{h}(F, G)-h(F, G)\leqq[\sup(\hat{F}-F)+\sup(\hat{G}-G)] .$$

Therefore, to evaluate (*) we have only to replace η' by

$$\eta''=\frac{1}{2}\left(\varepsilon+\eta(1-p-\varepsilon)^2-\frac{1}{2}\left\{2\left(\frac{\varepsilon}{1-p}\right)-\left(\frac{\varepsilon}{1-p}\right)^2\right\}\right)$$

in (*).

Appendix

The right-hand expression of (*) in the preceding section can be calculated by an analogous procedure to that which is carried out by Z. W. Birnbaum and F. H. Tingey [4].

Let X be a non-negative random variable with a continuous distribution function except the origin $X=0$, i.e. let the distribution be

$$F(x)=\begin{cases} p, & \text{for } X\leqq 0 \\ p+(1-p)F^*(x), & \text{for } X>0 . \end{cases}$$

An ordered sample $X_1\leqq X_2\leqq X_3\leqq\cdots\leqq X_N$ of X determines the empirical distribution function

$$\hat{F}(x)=\begin{cases} 0 & \text{for } x<X_1, \\ \dfrac{k}{n} & \text{for } X_k\leqq x<X_{k+1} \\ 1 & \text{for } X_N\leqq x . \end{cases}$$

Now, we denote the minimum value of either value of $\hat{F}(x)+\eta'$ or 1 by $\hat{F}^+(x)$, $\eta'=\frac{1}{2}\{\varepsilon+\eta(1-p-\varepsilon)^2\}>\frac{\varepsilon}{2}>0$. Then what we want is:

$$P(\eta', m, p)=\text{prob.}\left\{F(x)\leqq\hat{F}^+(x) \quad \text{for all } x>0,\ \hat{p}=\frac{m}{N}\right\} ,$$

$$\text{under } \left|\frac{m}{N}-p\right|\leqq\frac{\varepsilon}{2} ,$$

where $\hat{p}$ is the empirical value of p. Since $P(\eta', m, p)$ does not depend on $F(x)$ at $x>0$, we assume that X has the distribution function

$$F(x)=\begin{cases} p & \text{for } x\leqq 0 \\ p+(1-p)x & \text{for } 0<x<1, \\ 1 & \text{for } 1\leqq x \end{cases}$$

For this random variable X, $P(\eta', m, p)$ is the product of the probability that an ordered sample

$$0 \leqq X_1 \leqq X_2 \leqq \cdots \leqq X_N \leqq 1$$

falls into the region

$$0 \leqq X_{m+1} \leqq \frac{1}{1-p}\left(\frac{m}{N}-p+\eta'\right)$$

$$X_{m+k-1} \leqq X_{m+k} \leqq \frac{1}{(1-p)}\left(\frac{m+k-1}{N}-p+\eta'\right) \quad \text{for } k=2, \cdots, K+1 ,$$

$$X_{m+k-1} \leqq X_{m+k} \leqq 1 \qquad \text{for } k=K+2, \cdots, N, \text{ by } p^m ,$$

where $m+K=[N(1-\eta')]$. Then we have

$$P(\eta', m, p)=\frac{N!}{m!}p^m\phi(\eta', m, p) ,$$

and

$$\phi(\eta', m, p)=(1-p)^{N-m}\int_0^{\lambda(\xi+\eta')}\int_{X_{m+1}}^{\lambda(\xi+\eta'+(1/N))}\cdots\int_{X_{m+K}}^{\lambda(\xi+\eta'+(K/N))}\int_{X_{m+K+1}}^{1}$$
$$\cdots\int_{X_{N-1}}^{1} dX_N dX_{N-1}\cdots dX_{m+K}\cdots dX_{m+2}dX_{m+1} ,$$

where $\lambda=\dfrac{1}{1-p}$ and $\xi=\dfrac{m}{N}-p$. Moreover,

$$\phi(\eta', m, p)=(1-p)^{N-m}\int_0^{\lambda(\xi+\eta')}\int_{X_{m+1}}^{\lambda(\xi+\eta'+(1/N))}$$
$$\cdots\int_{X_{m+K}}^{\lambda(\xi+\eta'+(K/N))}\frac{(1-X_{m+K+1})^{N-m-K-1}}{(N-m-K-1)!}dX_{m+K+1}\cdots dX_{m+1}$$
$$=(1-p)^{N-m}\int_0^{\lambda(\xi+\eta')}\cdots\int_{X_{m+K-1}}^{\lambda(\xi+\eta'+(K-1/N))}\frac{(1-X_{m+K})^{N-m-K}}{(N-m-K)!}dX_{m+K}\cdots dX_{m+1}$$
$$-\frac{(1-p)^{N-m-K}(1-\lambda(\xi+\eta'+(K/N))^{N-m-K}}{(N-m-K)!}\cdot\frac{(\xi+\eta')}{K!}\left\{\left(\xi+\eta'+\frac{K}{N}\right)^{K-1}\right\}.$$

Therefore, we have

$$P(\eta', m, p)$$
$$=\frac{N!}{m!\,(N-m)!}p^m\Big((1-p)^{N-m}-(\xi+\eta')^{[N(1-\eta')]-m}\sum_{j=0}^{[N(1-\eta')]-m}\binom{N-m}{j}\left(\xi+\eta'+\frac{j}{N}\right)^{j-1}$$
$$\times\left\{1-p-\left(\xi+\eta'+\frac{j}{N}\right)\right\}^{N-m-j}\Big) .$$

On the other hand, setting $\hat{F}^-(x)=\max[\hat{F}(x)-\eta', 0]$,

$$\text{prob.}\left\{F(x)\geqq\hat{F}^{-}(x) \text{ for all } x>0,\ \hat{p}=\frac{m}{N}\right\}$$

is equal to $P(\eta', m, p)$ as far as $\left|\frac{m}{N}-p\right|\leqq\frac{\varepsilon}{2}$ holds.

REFERENCES

[1] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 50–60.

[2] J. Putter, "The treatment of ties in some nonparametric tests" *Ann. Math. Stat.*, Vol. 26 (1955), pp. 369–386.

[3] E. L. Lehmann, "Consistency and Unbiasedness of certain non-parametric tests", *Ann. Math. Stat.*, Vol. 22 (1951) pp. 165–179.

[4] Z. W. Birnbaum and F. H. Tingey, "One-sided confidence contours for probability distribution functions," *Ann. Math. Stat.*, Vol. 22 (1951) pp. 592–596.