Semi-Sequential Rank Based Test for Location for Bivariate Population

Amitava Mukherjee

Department of Mathematics and Mathematical Statistics,
Umeå University,
MIT-Huset, Plan 3,
Umeå, SE-901 87 , Sweden
amitava.mukherjee@math.umu.se,
amitmukh2@yahoo.co.in.

Abstract. Partially sequential or semi-sequential tests have been widely studied over past three decades for location problems for two or more independent populations. In the present work, we address semi sequential location problem for Bi-variate population. Let (X, Y) be the jointly distributed random variables. X denotes the records of the first treatment and Y that of the second treatment on same individual. Suppose we already observed a set of random samples $Xi, (i = 1, , m)$ of fixed size m from X population. In order to save time and cost of the experiment, it is decided to collect samples from Y population sequentially with the restriction that at most $q(q << m)$ samples may be observed from Y population. Therefore, we use inverse sampling scheme to collect samples from Y population. We design a stopping rule and propose a simple but powerful technique to test for the difference in location using all the available sample observations from one variable and partially observed records from other variables. The proposed test is based on ranks. We study the asymptotic performance of the proposed test. Some numerical findings obtained through Monte-Carlo studies are presented. The proposed test may be used when one has m observations regarding pre-monsoon contaminated arsenic level in ground water collected from different tube-wells or bore-holes. While geologists need early and efficient decision regarding post monsoon change in mean arsenic contamination level based on the data from those same tub wells or bore holes. Another example is, monitoring acid-rain impact based on pH factor of the water bodies.

Keywords. Acidic precipitation, Arsenic contamination, Bivariate population, Rank, Semi sequential test, Test for location.

1 Introduction

Wolfe (1977) introduced the partially sequential sampling scheme for two sample location problem. Costello and Wolfe (1980) extended partially sequential test for more than two populations. Various aspects of such a sampling scheme, tests based on them under different assumptions and alternatives are widely studied by a host of researchers. Orban and Wolfe (1978, 1980, 1982), Chatterjee and Bandyopadhyay (1984), Chu et al. (1996), Chattopadhyay (2002), Bandyopadhyay and Mukherjee (2007), Bandyopadhyay et al. (2007, 2008a, 2008b), Mukherjee (2009), among others, considered different versions of partially sequential tests. However, in all previous communications partially sequential or semi sequential tests are designed for two or more independent populations. The present study deals with the semi sequential test for the difference in locations in a bi-variate population. The problem is motivated from some geo-statistical issues. However it can be used in several other contexts as well. Suppose we are interested to study the monsoon effect on Arsenic contamination in ground water or wish to monitor the level acidic precipitation on water bodies.

It is believed by a section of geologists that during monsoon in the whole aquifer, the capillaries of soils are filled-up with water through revitalization of the aquifer by rain and river water. As a consequence, the water level raises up 2-3 m and arsenic concentration may somehow get a bit diluted at the concentrated hot-spots. This may increase the concentration in the nearby area by migration. As as a reason sometimes, level of concentration sometimes changes from pre-monsoon to post monsoon. Suppose one has collected a huge number of water samples from different wells or bore holes prior to beginning of monsoon. Now we want to test weather arsenic contamination has increases after monsoon. We obviously need a quick decision in this aspect as an increase in arsenic contamination in groundwater can cause severe disaster to the human race particularly if such a water is used for drinking purpose. To save the time and cost of the experiment we would like to employ a sequential sampling in post monsoon phase. we start collecting samples from the same wells or bore holes used in previous pre monsoon season using some stopping rule. In this stage, we only observe as many samples as required to reach a decision at a specified level. This will obviously lead to a inverse sampling scheme under partially

sequential framework. Only difference is that here a pre monsoon water sample has an obvious dependence with corresponding post monsoon sample when they are collected from same wells or bore holes. So we realize a bi-variate location problem under semi sequential sampling design.

Similarly if we may think of an experiment related to monitoring the acid snow in winter time. Acid rain is the common name for many phenomena including acid fog, acid sleet, and acid snow. When atmospheric pollutants such as sulphur dioxide and nitrogen oxides mix with water vapour in the air, they are transformed to sulphuric and nitric acids. These acids make the rain acidic, hence the term acid rain. Rain returns the sulphur and nitrogen acids to Earth. If large quantities of acid rain are deposited they may have detrimental consequences for wildlife, forests, soils, freshwater and buildings. Acid rain acidifies the soils and waters where it falls, killing off plants and animals. Surface water acidification can lead to a decline in, and loss of, fish populations and other aquatic species including frogs, snails and crayfish. Acid rain affects trees, usually by weakening them through damage to their leaves. Certain types of building stone can be dissolved in acid rain. Acid precipitation in winter in some areas may increase the acidity of the water which can be measured by pH level [the cologarithm of the activity of dissolved hydrogen ions (H+)]. Increase in acidity will reduce the pH factor of the water samples collected from a specific water bodies. If we have a large number of observation on pH level of water bodies, say before snow in pre winter, we can quickly make decision in spring whether there is a decline in pH factor. We just need to employ a sequential stopping rule to collect samples in spring from some of the water bodies included in the experiment before.

The proposed technique may be very useful in clinical trial if one wish study the level of improvement with the second dose of a specific drug. Suppose one already have some data on the effect of the drug after the treatment with first dose. For ethical reason we might not be interested to administer second dose to all of them but only a few of them chosen sequentially as per requirement to reach a conclusion. Obviously a inverse sampling scheme under partially sequential framework revisits in this case with a bi-variate location problem.

2 Statistical framework

Let us consider an experiment with two states. As for example, pre-medication state and post medication state in a clinical trial experiment or pre-monsoon Arsenic contamination and post monsoon arsenic contamination level in groundwater. Let X be random variable denting the records corresponding to initial state, say, State A. Further suppose Y be the random variable denoting the post treatment records or the records or final state, say, State B. Then we may assume that (X,Y) are jointly distributed as $F_{X,Y}(x,y)$, where $F_{X,Y}(x,y)$ is a bivariate probability distribution function. Let μ_X and μ_Y be the median corresponding to the random variable X and Y respectively. Our problem is to test

$$H_0 : \mu_X = \mu_Y$$

against one-sided alternative

$$H_1 : \mu_X < \mu_Y.$$

For the present treatise, our working assumptions are X and Y are continuous and are homoscedastic.

Let $\mathbf{X_m} = (X_1, ..., X_m)$ be a random sample of size m from X. That is, $\mathbf{X_m}$ corresponds to a random sample from the initial state or State A. Observations from State B are to be observed one by one sequentially as per requirement following a simple rhythm. First observation Y_1 will be the record related to final state corresponding to X_1, Second observation Y_2 similarly corresponds to X_2 and so on. However, for some reason, out of those m individual samples, we may at most observe q $(q < m)$ random samples from the State B or from Y population. That is, we may say that in the best possible scenario, we have q proper paired samples as $(X_1, Y_1), ..., (X_q, Y_q)$. Besides this we $(m - q)$ random samples from the X population only, say $X_{m-q+1}, ..., X_m$. For this $(m-q)$ samples Y observations are either missing or not recorded because of induced censoring. Necessarily, a sequential inverse sampling scheme comes into operation. Thus, if we think of usual nonparametric tests for paired sample location problems like sign test, Wilcoxon signed rank test etc, we can only make use of q samples for which

records related to both the variables are available. As a result, atleast $(m-q)$ additional information related to State A will be under utilized. In the present context, we aim at designing a test procedure where we can optimally utilize all the available m samples related to State A. Even we like to make use of all the q available resources corresponding to State B, particularly when null hypothesis is true. If there are sufficient evidence against the null hypothesis at a given level we do not mind if some information related to State B are overlooked. For this, we consider a fixed set of m observation from State A a priori and consider samples from State B one by one sequentially, with $j-th$ samples from State B, i.e. Y_j represents the final state records of X_j, the $j-th$ sample from State A for $j=1,...,q$.

Now, for $j=1,...,q$, let us define,

$$\Psi_j = \text{ Rank of } \quad X_j \quad \text{ among } \quad X_1, X_2, ..., X_{j-1}, X_j, X_{j+1}, ..., X_m$$

and

$$\Phi_j = \text{ Rank of } \quad Y_j \quad \text{ among } \quad X_1, X_2, ..., X_{j-1}, Y_j, X_{j+1}, ..., X_m.$$

Note that both $\Psi_j's$ and $\Phi_j's$ assume positive integer values between 1 and m. Further Suppose, for $j=1,...,q$,

$$\Delta_j^* = \Phi_j - \Psi_j.$$

Obviously, Δ_j^*'s may be negative and assume integral values between (1-m) to (m-1). It is easy to note that that under null hypothesis, most of the Δ_j^*'s are expected to be smaller and take values around 0. While under alternative hypothesis of right shift of the median of Y population, majority of the Δ_j^*'s are expected to be larger and positive. We Consider

$$\Delta_j = \frac{m + \Delta_j^*}{2m-1},$$

and for any positive integer n, define the partial sums

$$S_n = \sum_{j=1}^{n} \Delta_j.$$

Now, we introduce a stopping variable N depending on the sequence of partial sums S_n. We define,

$$N = \min_{n \geq n_0} \{n : S_n > \frac{r}{2}\},$$

where r is a prefixed positive number depending on q and n_0.

Obviously Δ_j's are positive random variables lying in the half open interval $(0,1]$. For, moderately large m Δ_j's are expected to take more and more values around 0.5 and under alternative of right shift in median, it is expected to be larger. Therefore, for any given n, the partial sums are are expected to be stochastically larger under the alternative than under the null hypothesis. As a consequence, a left tailed test based on N will be appropriate in the present scenario.

Let N_α, depending on r, be the lower 100α percent point of the distribution of N. Then, we reject H_0 in favour of the alternative at the level α, if observed

$$N < N_\alpha.$$

Note that if

$$S_{N_\alpha} < \frac{r}{2},$$

we can accept the null hypothesis at the level α even without considering any further samples from Y-population. This is an advantage of inverse sampling scheme considered here. Capitalizing this feature, we set $N_\alpha = q$ and choose a suitable r accordingly. This will ensure that we can reach at a conclusion at a given level with the at most available q records from Y-population.

If alternative is of the type $H_2 : \mu_X > \mu_Y$, then just use $\Delta_j^{**} = \Psi_j - \Phi_j$, instead of Δ_j^* and proceed as before. In the subsequent sections we discuss asymptotic normality of the stopping variable and obtain numerical results through simulation study. We provide a small data study to illustrate the proposed procedure before presenting the concluding remarks.

3 Some asymptotic results

Result 3.1. There exists a positive integer ν such that, for large m, the asymptotic distribution of $\frac{r-N}{2\sqrt{r}}$ is equivalent to that of $\frac{S_\nu - \frac{\nu}{2}}{\sqrt{\nu}}$.

Proof. Note that, for any x,

$$\begin{aligned} Prob\ (N > x) &= Prob\ (\max_{n=n_0,\ldots,[x]} S_n < \frac{r}{2}) \\ &= Prob\ (S_{[x]} < \frac{r}{2}), \end{aligned} \tag{1}$$

since, Δ_j's being positive, partial sums $\{S_n\}$ forms a strictly monotonic increasing sequence. Here $[\cdot]$ stands for largest integer contained in it. This implies

$$Prob\ (\frac{N-r}{2\sqrt{r}} > \frac{x-r}{2\sqrt{r}}) = Prob\ (\frac{S_{[x]} - \frac{x}{2}}{\sqrt{r}} < -\frac{x-r}{2\sqrt{r}}). \tag{2}$$

Let $\xi = \frac{x-r}{2\sqrt{r}}$, so that $x = r + 2\xi\sqrt{r}$. Define ν to be the largest integer contained in x. Therefore for every r and finite ξ, Assuming as m is sufficiently large, r is also very large and thus, we get a ν such that as r tends to ∞, ν also tends to ∞, but $\frac{\nu}{r}$ tends to 1. Hence (3.2) immediately gives

$$Prob\ (\frac{N-r}{2\sqrt{r}} > \xi) = Prob\ (\frac{r-N}{2\sqrt{r}} < -\xi) = Prob\ (\frac{S_\nu - \frac{\nu}{2}}{\sqrt{\nu}} < -\xi).$$

Hence the result follows.

Result 3.2. Under null hypothesis, for any $j = 1, 2, ...$, $E[\Delta_j^*]$ equals 0. Also, for any $j = 1, 2, ...$, and $j\prime(\neq j) = 1, 2, ...$, $E[\Delta_j^* \Delta_{j\prime}^*]$ equals 0 when null hypothesis is true.

Proof. Proof is trivial and hence is omitted.

Result 3.3. For large m, and consequently a large ν, the asymptotic null distribution of $\frac{S_\nu - \frac{\nu}{2}}{\sqrt{\nu}}$ can be approximated by normal with mean 0 and variance $\frac{1-3\tau}{6}$, where $\tau = Prob\ (X_j > X_i, Y_j > X_{i'})$ with $i \neq i' = 1, ...m$, and $j \neq i \neq i' = 1, ...$

Proof. Using Result 3.2 and some algebric computation, the result follows straightway from the classical central limit theorem.

As a consequence of the above results, we may say that asymptotic null distribution of N can be approximated by normal with mean r and variance $\frac{2r}{3}(1-3\tau)$. Therefore, it is legitimate to think that we reject H_0 against H_1 if observed N is less than $r - [\frac{2r}{3}(1-3\tau)]^{\frac{1}{2}}\tau_\alpha$, where τ_α is the upper 100α percent point of the standard normal distribution. Note that, τ is, in general, not known. So we need to estimate it from the observed data. Suppose, $c(A) = 1$ or 0 according as A or A^c occurs. Then $3\sum_{i=1}^{m}\sum_{i'(\neq i)=1}^{m}\sum_{j(\neq i\ i')=1}^{n_0} c(X_j > X_i, Y_j > X_{i'})/[(m(m-1)(m-2) - (m-n_0)(m-n_0-1)(m-n_0-2)]$ may be considered to be the basic estimator of τ.

One may further think of improving this based on incoming observations sequentially, but for most the practical situations this basic estimator serves the purposes quite well. One should note that when X and Y are perfectly positively associated with $X_j = Y_j$ for all j, we have $\tau = 1/3$, while in case of perfect negative association with $X_j = -Y_j$ for all j, we have $\tau = 1/6$. In case X and Y are independent, it is see to see that $\tau = 1/4$. Nevertheless, while estimating τ from data, one may find that the estimate is more than 1/3 and consequently encounter a negative estimate of variance. To avoid this hazard and also degeneracy, we suggest to take the minimum of 0.33 and basic estimate as the working estimate of τ.

4 Numerical results

Extensive simulation studies have been carried out to verify the asymptotic results. Findings based on 10000 replicates of Monte Carlo experiments are extremely encouraging. Data are generates from bi-variate normal as well as bi-variate Cauchy distributions using R-packages (MASS and fMultivar respectively). We draw a sample of size m from X population with m equals 25, 50 and 100. We study two situations when $r = 0.6m, n_0 = 0.4m$ and when $r = 0.4m, n_0 = 0.2m$. For sake of brevity, we omit the details and just note few points. We see that the type I error or level actually attained based on asymptotic cut off points at 5% level are pretty satisfactory in most of the situations. Actual type I error is bit low when there is a high order of positive association between two variables. At the same time, remaining other things fixed, if we consider power under right shift in second population, it slowly increase as correlation increases from -1 to +1. For example, if we consider bivariate normal with $(0, 0, 1, 1, \rho)$ as null distribution and bivariate normal with $(0, 1, 1, 1, \rho)$ as the alternative, we see that with $m = 25, r = 15$, and $n_0 = 10$, power increases from 0.6 tot 0.9 when correlation coefficient (ρ) is increases from -0.8 to 0.8. Further as desired, power increases with r for fixed m as well with m for fixed r. Power is close to unity with even small shift of order 0.5 with unit variances, if m is about 50, $r = 0.6m, n_0 = 0.4m$ and there is a good positive association. But for large m and r there are chances that level actually attained may be 1-1.5% more than desired when association between the two variables is less. It might be a because of over estimation of τ in presence of spurious correlation. However the test is indeed a great choice when there is high order of positive association and we are looking for a right shift in second population. Same may be said when there is high order of negative association and we are looking for a left shift in second population.

5 Data study

The objective of present data study is to show the gain in sample size through proposed sequential test while reaching a valid decision. The gain is sample size from the second population invariably minimizes cost of the experiment and also save considerable time. The data is taken from Massachusetts Water Resources Research Center at UMASS/Amherst Acid Rain Monitoring Project carried out with funding from Massachusetts Division of Fisheries, Wildlife, Massachusetts Department of Environmental Protection, Trout Unlimited and the U.S. Geological Surveys Water Resources Institute Program. 156 observations corresponding to X populations from various water bodies in Cape cod watershed (A drainage basin is an extent of land where water from rain or snow melt drains downhill into a body of water, such as a river, lake, reservoir, estuary, wetland, sea or ocean) are collected prior to snowfall in the mid of October in 1984. Second samples corresponding to Y population are collected after snow melt in mid April of 1985. See website (http://umatei.resuo.ads.umass.edu/armproject1/bsearch.cfm) for the details of data. data shows that correlation between two variable is as high as 0.779 while the variability in both the samples are amazingly similar up to 4 places of decimal and is 0.5443. Thus homogeneity in variance assumption is not all problem. However assumption of normality is not so justified as a Kolmogorov-Smirnov test based on 156 first sample observations returns a low p-value 0.04367. Non-normality and dependence structure rules out various possibilities of two sample location test. A Wilcoxon signed rank test returns a very high p-value (0.9977) indicating there is no right shift of location in second sample.

Now let us set $r = 100, n_0 = 60$ and carry out the proposed test. We see basic estimate of τ is 0.3392. As it marginally exceeds 0.33 we take estimated $\tau = 0.33$ and at 5% level we have cut off points as 98.65686. We may reach at the same conclusion of no right shift in second population saving 57 sample observations and using only 99 second sample observations as the sequential drawing does not terminate early on or before 98 observations. If we set $r = 65, n_0 = 35$ the basic estimate of τ becomes $0.3025 (< 0.33)$. So we take $\tau = 0.3025$ and at 5% level we find cut off points as 61.70833. In this case also, we reach at the same conclusion saving 94 sample observations and using only 62 second sample observations as the sequential drawing does not terminate before the cut off. Choice of r may be made depending on the available resource, time and cost with of course amount of precision required. This is left to the statistician's choice. Nevertheless, we advise to take $r < (m - \sqrt{3m})$.

6 Concluding remarks

One should note that even though the propose test is based on ranks, it is not truly nonparametric in nature. This is because τ is model dependent. One may refer this test as near nonparametric. However, if there is a high order of positive association between the two variables and that is known a priori one should set a prefixed value of τ to achieve a purely nonparametric test. Interesting fact is that the asymptotic results are applicable to a class of bi-variate populations starting from thin tailed Normal to heavy tailed Cauchy for testing difference in location. Future researches on partially sequential framework assuming various dependent set-up are highly warranted. An immediate consequence of the present work should be to extend the test in case of heterogeneous variance of the two populations. Improved estimation of τ is also a worth considering problem.

Acknowledgements

Author is grateful to Prof. Nitis Mukhopadhyay of University of Connecticut, USA, and Dr. Sara Stöstedt de-Luna of Umeå University, Sweden, for their encouragements as well as a referee for his/her suggestions. Author is thankful to Wallenberg foundation for the financial support towards his conference participation.

References

Bandyopadhyay, U. and Mukherjee, A. (2007). Nonparametric Partial Sequential Test for Location Shift at an Unknown Time Point. *Sequential Analysis*, **26**, 99-113.

Bandyopadhyay, U., Mukherjee, A., and Purkait, B. (2007). Nonparametric Partial Sequential Tests for Patterned Alternatives in Multi-Sample Problems. *Sequential Analysis*, **26**, 443-466.

Bandyopadhyay, U., Mukherjee, A., and Biswas, A. (2008a). Controlling Type-I Error Rate in Monitoring Structural Changes Using Partially Sequential Procedures. *Communications in Statistics: Simulation and Computation*, **37**, 466-485.

Bandyopadhyay, U., Mukherjee, A., and Purkait, B. (2008b). Simultaneous Tests for Patterned Recognition using Nonparametric Partially Sequential Procedure. *Statistical Methodology*, **5**, 535-551.

Chatterjee, S. K. and Bandyopadhyay, U. (1984). Inverse Sampling based on General Scores for Nonparametric Two-sample Problems. *Calcutta Statistical Association Bulletin*, **33**, 35–58.

Chattopadhyay, G. (2002). Partially Sequential Nonparametric Two-Sample Test for Ordered Categorical Data. *Journal of Nonparametric Statistics*, **14**, 539-553.

Chu, C.J., Stinchcombe M. and White, H. (1996). Monitoring Structural Change. *Econometrica*, **64**, 1045–1065.

Costello, P. and Wolfe, D. A. (1980). Partially Sequential Treatment versus Control Multiple Comparison. *Biometrika*, **67**, 403-412.

Mukherjee, A. (2009). Some Rank-Based Two-Phase Procedures in Sequential Monitoring of Exchange Rate. *Sequential Analysis*, **28**, To Appear.

Orban, J. and Wolfe, D. A. (1978). Optimality Criteria for the Selection of Partially Sequential Indicator set. *Biometrika*, **65**, 357–362.

Orban, J. and Wolfe, D. A.(1980). Distribution Free Partially Sequential Placement Procedure. *Communications in Statistics-Theory and Methods*, **9**, 883–902.

Orban, J. and Wolfe, D. A. (1982). A Class of Distribution Free Two-Sample Tests Based on Placements. *Journal of American Statistical Association*, **77**, 666–672.

Randles, H. R. and Wolfe, D. A. (1979). *Introduction to the Theory of Nonparametric Statistics.* Jonn Wiley & Sons, Inc., New York.

Wolfe, D. A. (1977). On a Class of Partially Sequential Two Sample Test Procedure. *Journal of American Statistical Association*, **72**, 202–205.