Testing Methods of Mean Difference for Longitudinal Data Based on Stationary Bootstrap

Hirohito Sakurai[1,3] and Masaaki Taguri[2]
[1] National Center for University Entrance Examinations, Tokyo, JAPAN
[2] Chuo University, Tokyo, JAPAN
[3] Corresponding author: Hirohito Sakurai, e-mail: sakurai@rd.dnc.ac.jp

Abstracts

We propose testing methods for detecting the difference of two mean curves in longitudinal data using the stationary bootstrap when the data of two groups are not paired. For the detection of mean difference of two groups, we here consider the following four types of test statistics: (i) sum of absolute values of difference between two mean sequences, (ii) sum of squares of difference between two mean sequences, (iii) estimator of area-difference between two mean curves, and (iv) difference of kernel estimators based on two mean sequences. The stationary bootstrap is used to approximate the null distribution of each test statistic. Our approaches of block resampling generate a resample with replacement from blocks of observations. Monte Carlo simulations are conducted in order to investigate finite sample behavior of the sizes and powers of the proposed tests. We also show an example of how to use the above methods for analyzing a real data.

Keywords: block resampling, comparison of mean curves, sizes and powers of tests

1. Introduction

Comparing two means or regression curves of two samples is an important problem in statistical inference. Suppose now that there are two samples given by $\{Y_i(t)\}_{i=1}^{q_1}$ and $\{X_j(t)\}_{j=1}^{q_2}$ for $t = 1, \ldots, n$, and assume that they are mutually independent, where q_1 and q_2 are numbers of subjects, and n is the number of observed points. We also assume that, for fixed t, $Y_i(t)$ and $X_j(t)$ are independent over q_1 and q_2 subjects, respectively. Then we consider the model

$$\begin{cases} Y_i(t) = p_1(t) + \varepsilon_i(t), & i = 1, \ldots, q_1, \\ X_j(t) = p_2(t) + \eta_j(t), & j = 1, \ldots, q_2, \end{cases} \tag{1}$$

where $p_1(t)$ and $p_2(t)$ are unknown regression functions, and $\varepsilon_i(t)$ and $\eta_j(t)$ are the error terms having means 0 and finite variances, respectively. Then, we are interested in a testing problem

$$H_0 : p_1(t) = p_2(t) \text{ for all } t \quad \text{vs.} \quad H_1 : p_1(t) \neq p_2(t) \text{ for some } t, \tag{2}$$

where H_0 and H_1 denote the null and alternative hypotheses.

In this paper, we propose testing methods to detect the significant difference between $p_1(t)$ and $p_2(t)$ using the stationary bootstrap (Politis and Romano, 1994). In Section 2 we propose testing methods using four types of test statistics and stationary bootstrap. In order to investigate the properties of sizes and powers of the proposed testing methods, Monte Carlo simulations are carried out in Section 3, and some concluding remarks and results of a real data analysis are summarized in Section 4.

2. Testing Methods

There are some approaches to detecting the difference between two mean curves, $p_1(t)$ and $p_2(t)$, in (1). The following statistic is proposed by Hall and Hart (1990):

$$S_n = S_n(D_1,\dots,D_n) = \left[\sum_{j=0}^{n-1}\left(\sum_{t=j+1}^{j+g} D_t\right)^2\right]\left[n\sum_{t=1}^{n-1}\frac{(D_{t+1}-D_t)^2}{2}\right]^{-1}, \tag{3}$$

where $D_t = Y_t - X_t$ for $t = 1,\dots,n$ or $D_t = Y_{t-n} - X_{t-n}$ for $t = n+1,\dots,n+g$, $Y_t = \sum_{i=1}^{q_1} Y_i(t)/q_1$, $X_t = \sum_{j=1}^{q_2} X_j(t)/q_2$, $g = [np]$ is the integer part of np, and p is a tuning constant satisfying $0 < p < 1$ which is determined by the fully data-driven approach in Hall and Hart (1990, pp.1043–1044). The statistic (3) is essentially based on kernel estimators of $p_1(t)$ and $p_2(t)$. As another type of test statistics, we here consider

$$T_{1n} = T_{1n}(D_1,\dots,D_n) = \sum_{t=1}^{n} |D_t|, \tag{4}$$

$$T_{2n} = T_{2n}(D_1,\dots,D_n) = \sum_{t=1}^{n} D_t^2. \tag{5}$$

In addition to (3), (4) and (5), we also consider the following test statistic:

$$T_{3n} = T_{3n}(D_1,\dots,D_n) = \frac{1}{2}\sum_{t=1}^{n-1}(|D_t|+|D_{t+1}|)I_{3,1} + \frac{1}{2}\sum_{t=1}^{n-1}\frac{|D_t|^2+|D_{t+1}|^2}{|D_t|+|D_{t+1}|}I_{3,2}, \tag{6}$$

where $I_{3,1} = I\{D_tD_{t+1} \geq 0\}$, $I_{3,2} = I\{D_tD_{t+1} < 0\}$ and $I\{\cdot\}$ is the indicator function, respectively. The statistic (6) is an estimator of $A = \int |p_1(t) - p_2(t)|\,dt$ constructed by the trapezoidal rule with linear interpolations of adjacent observations. The quantity A is 0 under H_0 and positive under H_1. Thus, the hypothesis of our interest reduces to testing

$$H_0 : A = 0 \quad \text{vs.} \quad H_1 : A > 0. \tag{7}$$

Note that the values of S_n and T_{rn} $(r = 1,2,3)$ will be small when H_0 is true, and large when H_0 is false. Using these four statistics, we can measure the discrepancy between $p_1(t)$ and $p_2(t)$.

In this section, we propose testing methods for the problem (2) or (7) using (3), (4), (5) and (6). We call them "Mixed Stationary Bootstrap (MSB) Test." The main ideas of MSB test are to make blocks of observations in each sample similar to the stationary bootstrap, and to generate resamples corresponding to two samples by drawing blocks with replacement from the mixed (pooled) stationary bootstrap blocks. The latter is motivated from the technique that can reflect the null hypothesis by resampling from a combined sample. For i.i.d. data, the idea of combining observations of two samples and drawing resamples with replacement from the combined sample is previously considered by Boos et al. (1989) and Wang and Taguri (1996). The former is the test of homogeneity of scale, and the latter is that of equality of two means.

For simplicity, let T be a generic notation for statistics (3), (4), (5) and (6). For a given significance level α, the unified testing algorithm for T together with Monte Carlo method is described as follows.

1. Calculate $t_{obs} = T(Y,X) = T(D_1,\dots,D_n)$.
2. Put $C_{y,t} = Y_t - \overline{Y}$ and $C_{x,t} = X_t - \overline{X}$ for $t = 1,\dots,n$, where $\overline{Y} = \sum_{t=1}^{n} Y_t/n$ and $\overline{X} = \sum_{t=1}^{n} X_t/n$.

3. Divide $\{C_{y,1},\ldots,C_{y,n}\}$ and $\{C_{x,1},\ldots,C_{x,n}\}$ into n collections of blocks as follows:
$$\xi_y = \{\xi_y(1,L_1),\ldots,\xi_y(n,L_n)\}, \quad \xi_x = \{\xi_x(1,M_1),\ldots,\xi_x(n,M_n)\},$$
where $\xi_y(t,\ell)$ and $\xi_x(t,\ell)$ are the blocks starting from $C_{y,t}$ and $C_{x,t}$ with length $\ell(\geq 1)$, that is,
$$\xi_y(t,\ell) = \begin{cases} \{C_{y,t},\ldots,C_{y,t+\ell-1}\}, & t = 1,\ldots,n-\ell+1, \\ \{C_{y,t},\ldots,C_{y,n},C_{y,1},\ldots,C_{y,t+\ell-n-1}\}, & t = n-\ell+2,\ldots,n, \end{cases}$$
$$\xi_x(t,\ell) = \begin{cases} \{C_{x,t},\ldots,C_{x,t+\ell-1}\}, & t = 1,\ldots,n-\ell+1, \\ \{C_{x,t},\ldots,C_{x,n},C_{x,1},\ldots,C_{x,t+\ell-n-1}\}, & t = n-\ell+2,\ldots,n. \end{cases}$$
$L_1,\ldots,L_n,\ M_1,\ldots,M_n$ are independent and identically distributed to a geometric distribution with parameter $p = 1/\ell$.

4. Combine ξ_y and ξ_x, and put
$$\xi_{\text{pooled}} = \{\xi_y(1,L_1),\ldots,\xi_y(n,L_n),\ \xi_x(1,M_1),\ldots,\xi_x(n,M_n)\}.$$

5. Draw $K_{y,b}$ and $K_{x,b}$ blocks with replacement from ξ_{pooled}, and put
$$\xi_y^{*b} = \{\xi(I_1^{*b},L_1^{*b}),\ldots,\xi(I_{K_{y,b}}^{*b},L_{K_{y,b}}^{*b})\}, \quad \xi_x^{*b} = \{\xi(J_1^{*b},M_1^{*b}),\ldots,\xi(J_{K_{x,b}}^{*b},M_{K_{x,b}}^{*b})\},$$
where $b = 1,\ldots,B$,
$$\xi(t,\ell) = \begin{cases} \xi_y(t,\ell), & 1 \leq t \leq n, \\ \xi_x(t,\ell), & n+1 \leq t \leq 2n. \end{cases}$$
$I_1^{*b},\ldots,I_{K_{y,b}}^{*b},\ J_1^{*b},\ldots,J_{K_{x,b}}^{*b}$ are independent and identically distributed to a discrete uniform distribution on $\{1,\ldots,n,\ n+1,\ldots,2n\}$. A pair of random variables, (I_i^{*b},L_i^{*b}) or (J_i^{*b},M_i^{*b}), is one of $\{(1,L_1),\ldots,(n,L_n),\ (1,M_1),\ldots,(n,M_n)\}$, and $K_{y,b} = \min\{k : \sum_{i=1}^{k} L_i^{*b} \geq n\}$, $K_{x,b} = \min\{k : \sum_{j=1}^{k} M_j^{*b} \geq n\}$, respectively.

6. Construct resamples,
$$Y^{*b} = \{Y_1^{*b},\ldots,Y_n^{*b}\}, \quad X^{*b} = \{X_1^{*b},\ldots,X_n^{*b}\},$$
by putting the first n elements of ξ_y^{*b} and ξ_x^{*b}.

7. Calculate $t^{*b} = T(Y^{*b},X^{*b}) = T(D_1^{*b},\ldots,D_n^{*b})$ based on step 6, where $D_t^{*b} = Y_t^{*b} - X_t^{*b},\ t = 1,\ldots,n$.

8. Repeating steps 5–7 an appropriate number of times B, calculate $t^{*1},\ldots,t^{*B}$.

9. From steps 1 and 8,
$$\begin{cases} \text{reject } H_0 \text{ if } \widehat{\text{ASL}} \leq \alpha, \\ \text{accept } H_0 \text{ if } \widehat{\text{ASL}} > \alpha, \end{cases}$$
where
$$\widehat{\text{ASL}} = \frac{1}{B}\sum_{b=1}^{B} I\{t^{*b} \geq t_{obs}\}.$$

3. Numerical Study

In this section, we carry out Monte Carlo simulations in order to investigate finite sample behavior of the sizes and powers of the proposed tests in Section 2. Our study also includes the comparison with Bowman and Young's (1996) test for unpaired data (hereafter termed "BY" for short).

All our results are based on independent 2000 pairs of two samples, $\{Y_i(t)\}$ and $\{X_j(t)\}$, where $B = 2000$ replications of resampling in our tests are applied to every two samples, and the nominal level is $\alpha = 0.05$. We generate initial two samples according to (1) whose means are specified by $p_1(t) = 0$ and $p_2(t) = c$, where $c = 0, 0.2, 0.4, 0.6, 0.8, 1.0$; $c = 0$ or $c \neq 0$ corresponds to the null hypothesis or the alternative hypothesis being true. As for the error terms, $\varepsilon_i(t)$ and $\eta_j(t)$, we consider the following Gaussian AR(1) errors: $\varepsilon_i(t) = \phi\varepsilon_i(t-1) + z_i(t)$ and $\eta_j(t) = \phi\eta_j(t-1) + z_j(t)$, where $z_i(t) \overset{i.i.d.}{\sim} N(0, \tau_1^2)$, $z_j(t) \overset{i.i.d.}{\sim} N(0, \tau_2^2)$, $\phi = 0, \pm 0.1, \pm 0.2$, $\tau_1^2 = \tau_2^2 = (1-\phi^2)\mathrm{V}(\varepsilon_i(t))$, and $\mathrm{V}(\varepsilon_i(t)) = 1, 3, 5$. Due to limitations of space, we restrict ourselves to discussing the case of $\mathrm{V}(\varepsilon_i(t)) = 3$. For $n = 10$ points, the cases of $(q_1, q_2) = (10, 10)$, $(10, 20)$, $(10, 30)$, $(20, 20)$, $(20, 30)$, $(30, 30)$ are examined.

Since it is preferable that the empirical level is nearly equal to the nominal level α, our choice of ℓ in MSB test is done so that the empirical level is close to α. If there are some candidates which have the same level errors, we make the conservative choice, viz., we choose ℓ such that the empirical level is less than the nominal level. Further, if there are some candidates whose empirical levels are equal, we select ℓ to maximize the empirical power among them. The resulting choices of ℓ are summarized in Table 1.

We first summarize the results of the level studies in Table 2. The table shows that the empirical levels of MSB test with T_{rn} $(r = 1, 2, 3)$ and S_n tend to keep the nominal level α, however it is not true for most cases of $\phi = 0.2$. When $\phi > 0$, the level error of T_{3n} and S_n seems to be slightly larger. On the other hand, BY test shows a tendency to underestimate the nominal level except for the case of $(q_1, q_2) = (10, 10)$. S_n does not need longer ℓ than T_{rn} $(r = 1, 2, 3)$ to keep the nominal level as is shown in Table 1.

Table 1: Optimum ℓ in MSB test for $\mathrm{V}(\varepsilon_i(t)) = 3$

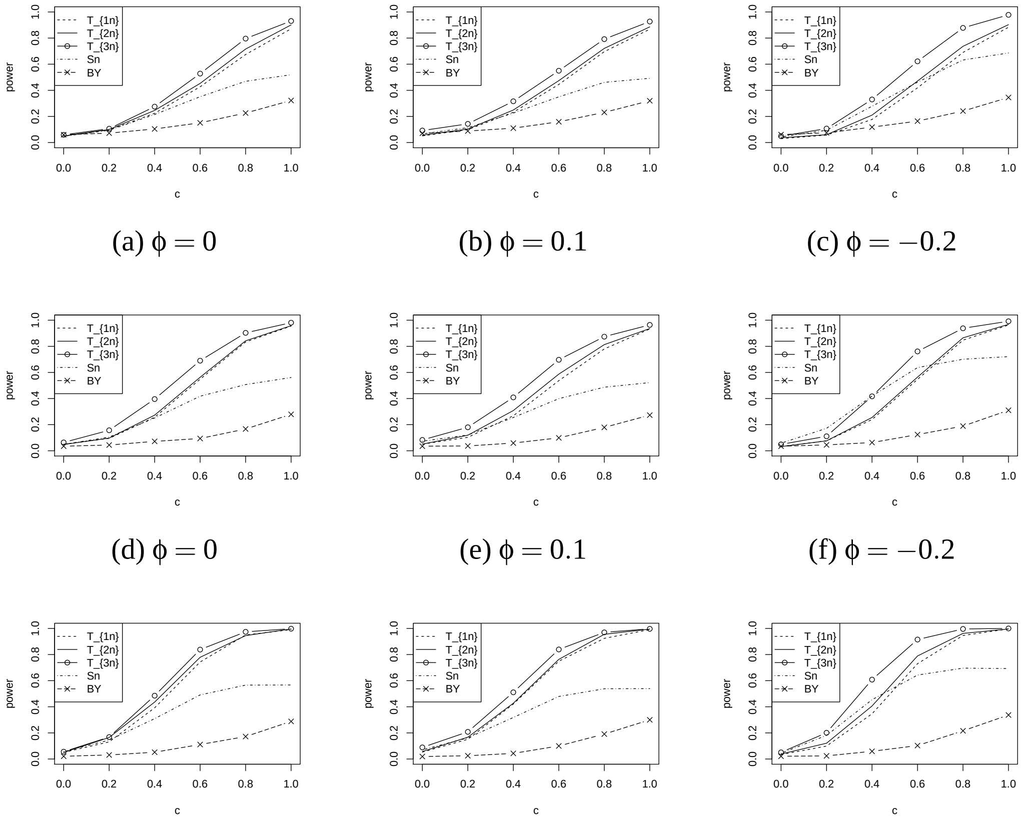
	$(q_1,q_2)=(10,10)$				$(q_1,q_2)=(10,20)$				$(q_1,q_2)=(10,30)$			
ϕ	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n
-0.2	6	6	7	2	1	1	9	3	2	1	6	2
-0.1	8	8	4	2	1	9	6	2	3	4	6	1
0	8	3	1	1	1	1	2	1	3	4	2	1
0.1	1	1	1	1	4	6	6	1	3	6	9	1
0.2	1	1	4	1	6	6	6	1	9	9	9	2

	$(q_1,q_2)=(20,20)$				$(q_1,q_2)=(20,30)$				$(q_1,q_2)=(30,30)$			
ϕ	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n	T_{1n}	T_{2n}	T_{3n}	S_n
-0.2	4	7	7	2	2	9	9	2	2	9	5	3
-0.1	4	7	3	2	2	9	3	2	4	9	3	2
0	5	9	1	1	1	4	1	1	9	7	1	1
0.1	9	1	1	1	7	1	1	1	2	1	3	1
0.2	6	1	8	2	8	2	9	1	2	1	6	1

Next, we discuss the power studies based on Figure 1. The vertical and horizontal axes of Figure 1 are the empirical power of tests and c $(0 \leq c \leq 1)$ defined above. Since we found similar tendencies among the six cases of (q_1, q_2), we show the results for $(q_1, q_2) = (10, 10)$, $(10, 20)$, $(20, 20)$ with $\phi = 0, 0.1, -0.2$.

Table 2: Empirical level for $\mathrm{V}(\varepsilon_i(t)) = 3$

ϕ	T_{1n}	T_{2n}	T_{3n}	S_n	BY	T_{1n}	T_{2n}	T_{3n}	S_n	BY
			$(q_1,q_2) = (10,10)$					$(q_1,q_2) = (20,20)$		
-0.2	0.030	0.036	0.049	0.043	0.061	0.034	0.038	0.051	0.044	0.021
-0.1	0.041	0.046	0.049	0.057	0.067	0.043	0.046	0.048	0.058	0.025
0	0.049	0.050	0.059	0.048	0.060	0.051	0.050	0.056	0.056	0.021
0.1	0.051	0.056	0.092	0.066	0.071	0.053	0.057	0.089	0.070	0.019
0.2	0.070	0.074	0.122	0.091	0.070	0.072	0.080	0.121	0.103	0.023
			$(q_1,q_2) = (10,20)$					$(q_1,q_2) = (20,30)$		
-0.2	0.033	0.032	0.049	0.058	0.035	0.027	0.035	0.046	0.044	0.028
-0.1	0.036	0.040	0.052	0.049	0.035	0.035	0.042	0.052	0.052	0.026
0	0.050	0.048	0.064	0.049	0.035	0.045	0.050	0.061	0.047	0.022
0.1	0.051	0.051	0.083	0.073	0.035	0.050	0.054	0.092	0.071	0.029
0.2	0.064	0.068	0.107	0.106	0.034	0.069	0.075	0.119	0.093	0.026
			$(q_1,q_2) = (10,30)$					$(q_1,q_2) = (30,30)$		
-0.2	0.022	0.021	0.041	0.044	0.036	0.035	0.040	0.049	0.058	0.026
-0.1	0.028	0.032	0.050	0.043	0.036	0.042	0.046	0.045	0.049	0.021
0	0.039	0.037	0.049	0.059	0.033	0.049	0.049	0.058	0.048	0.021
0.1	0.050	0.050	0.074	0.070	0.037	0.055	0.056	0.086	0.072	0.022
0.2	0.052	0.056	0.096	0.102	0.036	0.070	0.082	0.118	0.100	0.021



(a) $\phi = 0$ (b) $\phi = 0.1$ (c) $\phi = -0.2$

(d) $\phi = 0$ (e) $\phi = 0.1$ (f) $\phi = -0.2$

(g) $\phi = 0$ (h) $\phi = 0.1$ (i) $\phi = -0.2$

Figure 1: Empirical power for $\mathrm{V}(\varepsilon_i(t)) = 3$ and $(q_1,q_2) = (10,10),(10,20),(20,20)$ The panels (a)–(c), (d)–(f) and (g)–(i) are the cases of $(q_1,q_2) = (10,10)$, $(10,20)$ and $(20,20)$, respectively.

We can observe that the empirical power of T_{3n} is most powerful among those corresponding to four statistics, and that the overall relationship among powers corresponding to MSB and BY tests is given by $T_{3n} \geq T_{2n} \geq T_{1n} \geq S_n \geq \mathrm{BY}$. This indicates the numerical superiority of MSB test using the four statistics in power. Especially, the superiority of T_{3n} in power is confirmed from Figure 1. As the number of subjects increases, the empirical power is improved. For $0 \leq c \leq 1$, the empirical power of T_{1n} is nearly equal to that of T_{2n}, though T_{2n} is slightly higher than T_{1n} for most cases.

4. Concluding Remarks

In this paper we have proposed testing methods for detecting the difference of two means in longitudinal data based on the stationary bootstrap. Our numerical studies indicate the applicability of MSB test for weakly dependent data even when the observed points are quite few. In some cases the effectiveness of application of block resampling could be confirmed.

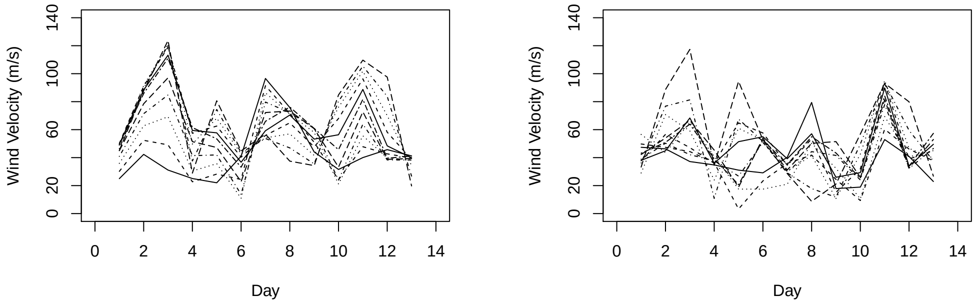


Figure 2: Wind velocity data (left: satellite, right: radar)

Figure 2 is a real data of wind velocity measured by an artificial satellite and a radar on the earth, where $q_1 = q_2 = 11$ and $n = 13$. Applying MSB with every possible ℓ and BY tests to the data given in Figure 2, we obtain the results that BY test rejects the null hypothesis, however MSB tests do not. Therefore there is a possibility of the significant difference between the satellite and radar in measuring wind velocity. However, the problem on the selection of ℓ in the block resampling is very important, and the development of a fully data-driven approach to selecting block length in MSB test will be needed for practical data analyses.

References

Boos, D., Janssen, P. and Veraverbeke, N. (1989) “Resampling from centered data in the two sample problem,” *Journal of Statistical Planning and Inference*, 21, 327–345.

Bowman, A. and Young, S. (1996) “Graphical comparison of nonparametric curves,” *Applied Statistics*, 45, 83–98.

Hall, P. and Hart, J. D. (1990) “Bootstrap test for difference between means in nonparametric regression,” *Journal of the American Statistical Association*, 85, 1039–1049.

Politis, D. N. and Romano, J. P. (1994) “The stationary bootstrap,” *Journal of the American Statistical Association*, 89, 1303–1313.

Wang, J. and Taguri, M. (1996) “Bootstrap method — An introduction from a two sample problem (in Japanese),” *Proceedings of the Institute of Statistical Mathematics*, 44, 3–18.