# A BOOTSTRAP TEST FOR INFORMATIVE INTRA - CLUSTER GROUP SIZES IN CLUSTERED DATA

Hasika K.Wickrama Senevirathne[1], Sandipan Dutta[2]

Email :[1] hwick003@odu.edu, [2] s1dutta@odu.edu

Department of Mathematics and Statistics, Old Dominion University

## INTRODUCTION

- Clustered data are observed in various domains.
- Units within a cluster are correlated while units between clusters are independent.

Example: in dental studies, individuals are clusters and teeth in an individual are units within a cluster.

- Informative intra-cluster group size (IICCGS) [1, 2] : Outcomes from a group in a cluster can be associated with the no. of units belonging to that group in that cluster.
- There does not exist a statistical method to test the existence of IICGS in a clustered data.
- We propose a bootstrap based hypothesis testing of IICGS in clustered data - assuming exchangeability within groups in a cluster [3].
- Through simulation studies, we show that our method can accurately detect IICGS in clustered data.

## METHODOLOGY

$M =$ No. of clusters

$B =$ No. of bootstrap samples

$Y_{ik} = k^{th}$ observation in the $i^{th}$ cluster

$N_i =$ No. of observations in the $i^{th}$ cluster.

$N_{i0} =$ No. of observations of group 0 in the $i^{th}$ cluster

$N_{i1} =$ No. of observations of group 1 in the $i^{th}$ cluster.

$G_{ik} =$ Group membership indicator.

$\mathbb{V}_i = \{N_i, Y_{ik}, G_{ik}\}, k = 1, \dots, N_i, i = 1, \dots, M$

- $H_0$: $\hat{F}(y) = \tilde{F}(y)$ vs. $H_1$: $\hat{F}(y) \neq \tilde{F}(y)$

where $\hat{F}(y) = \frac{\sum_{i=1}^{M}\sum_{k=1}^{N_i} I(Y_{ik} \le y, G_{ik}=d)}{\sum_{i=1}^{M}\sum_{k=1}^{N_i} I(G_{ik}=d)}$ and

$\tilde{F}(y) = \frac{\sum_{i=1}^{M}\frac{1}{2N_{id}}\sum_{k=1}^{N_i} I(Y_{ik} \le y, G_{ik}=d)}{\sum_{i=1}^{M}\frac{1}{2N_{id}}\sum_{k=1}^{N_i} I(G_{ik}=d)}$ ; where $d = 0,1$ [1]

- Test statistics:

*1)* $T_F = \sup_y |\hat{F}(y) - \tilde{F}(y)|$

*2)* $T_{CM} = \sum_{k \in \mathcal{J}}\left[k M_k \int (\hat{F}_k(y) - \hat{F}(y))^2 \, dy\right]$

where $\mathcal{J}$: set of unique cluster size,

$M_k$: no. of clusters of size $k$,

$\hat{F}_k(y) = \frac{1}{KM_k}\sum_{i=1}^{M}\sum_{j=1}^{N_i} \mathbb{I}(N_i = k, Y_{ij} \le y)$: estimator of the distribution of cluster size $k$ [3].

**Algorithm**

➢ ***Step 1***: Test statistic $T = T(\mathbb{V})$, where $\mathbb{V} = (V_1, \dots, V_M)$.

➢ ***Step 2:*** Consider $j^{th}$ bootstrap sample, $j = 1, \dots, B$ .

- Permute the units in each group within a cluster.
- Resample $M$ clusters from the permuted data set by repeating for every $i = 1, \dots, M$.
  - Draw a random cluster $V$ with index $i^*$ from $\{1, \dots, M\}$.
  - If $(N_{i^*1} \ge N_{i1}) \cap (N_{i^*0} \ge N_{i0})$ then the bootstrap cluster is $V^*_{ji} = \begin{Bmatrix} N_{i1}; Y^{(1)}_{i^*1}, \dots, Y^{(1)}_{i^*N_{i1}} \\ N_{i0}; Y^{(0)}_{i^*0}, \dots, Y^{(0)}_{i^*N_{i0}} \end{Bmatrix}$.

    where $\{Y^{(1)}_{i1}, \dots, Y^{(1)}_{iN_{i1}}\}$ and $\{Y^{(0)}_{i0}, \dots, Y^{(0)}_{iN_{i0}}\}$ represent the observations of the group 1 and group 0, respectively.
  - If $(N_{i^*1} \ge N_{i1}) \cap (N_{i^*0} < N_{i0})$ then the bootstrap cluster is merged from two matching clusters;

    $$V^*_{ji} = \begin{Bmatrix} N_{i1}; Y^{(1)}_{i^*1}, \dots, Y^{(1)}_{i^*N_{i1}} \\ N_{i0}; Y^{(0)}_{i^*0}, \dots, Y^{(0)}_{i^*N_{i0}}, Y^{(0)}_{k0(N_{i^*0}+1)}, \dots Y^{(0)}_{k0N_{i0}} \end{Bmatrix}$$

    where $k0 = argmin_{k0}(D_0(V_{i^*0}, V_{k0}): N_{k0} \ge N_{i0})$
  - If $(N_{i^*0} \ge N_{i0}) \cap (N_{i^*1} < N_{i1})$ then the bootstrap cluster is merged from two matching clusters;

    $$V^*_{ji} = \begin{Bmatrix} N_{i0}; Y^{(0)}_{i^*0}, \dots, Y^{(0)}_{i^*N_{i0}} \\ N_{i1}; Y^{(1)}_{i^*1}, \dots, Y^{(1)}_{i^*N_{i1}}, Y^{(1)}_{k1(N_{i^*1}+1)}, \dots Y^{(1)}_{k1N_{i1}} \end{Bmatrix}.$$

    where $k1 = argmin_{k1}(D_1(V_{i^*1}, V_{k1}): N_{k1} \ge N_{i1})$
  - If $(N_{i^*1} < N_{i1}) \cap (N_{i^*0} < N_{i0})$ then the bootstrap cluster is merged from two matching clusters;

    $$V^*_{ji} = \begin{Bmatrix} N_{i1}; Y^{(1)}_{i^*1}, \dots, Y^{(1)}_{i^*N_{i1}}, Y^{(1)}_{k1(N_{i^*1}+1)}, \dots Y^{(1)}_{k1N_{i1}} \\ N_{i0}; Y^{(0)}_{i^*0}, \dots, Y^{(0)}_{i^*N_{i0}}, Y^{(0)}_{k0(N_{i^*0}+1)}, \dots Y^{(0)}_{k0N_{i0}} \end{Bmatrix}$$

- $j^{th}$ bootstrap sample: $\mathbb{V}^*_j = (V^*_{j1}, \dots, V^*_{jM})$ and test statistic: $T^*_j = T(\mathbb{V}^*_j)$.

➢ ***Step 3***: Compute the $p$ −value as $\frac{1}{B}\sum_{j=1}^{B}\mathbb{I}(T^*_j \ge T)$.

Note: The distance between two clusters is defined as:

$$D_1(V_{i1}, V_{j1}) = (\min\{N_{i1}, N_{j1}\})^{-1} \sum_{k1=1}^{\min\{N_{i1}, N_{j1}\}} (Y^{(1)}_{i1k1} - Y^{(1)}_{j1k1})^2$$

$$D_0(V_{i0}, V_{j0}) = (\min\{N_{i0}, N_{j0}\})^{-1} \sum_{k0=1}^{\min\{N_{i0}, N_{j0}\}} (Y^{(0)}_{i0k0} - Y^{(0)}_{j0k0})^2$$

For tied distances, choose one of the clusters at random.

## RESULTS

**Simulation studies**

- $M = 50$ and 100 clusters, $B = 500$ and 1000 bootstrap samples and 500 Monte Carlo iterations.
- Test statistics for group 0 are $T_{F0}, T_{CM0}$, and for group 1 are $T_{F1}, T_{CM1}$.
- Let $Y_{i1} = 0.5 + a_i + e_1$ and $Y_{i0} = 0.5 + a_i + e_0$ where $a_i \sim N(0,1)$; where $a_i$ = random cluster effect $e_1 \sim N(0,0.3)$ and $e_0 \sim N(0.01,0.3)$ , $i = 1,2, \dots, M$

➢ ***Empirical size calculation***

$(N_{i1}-1) \sim Poi(15)$, $(N_{i1}-1) \sim Poi(12)$

- Nominal size = 0.05

| $M$ | $B$ | $T_{F0}$ | $T_{CM0}$ | $T_{F1}$ | $T_{CM1}$ |
|---|---|---|---|---|---|
| 50 | 500 | 0.050 | 0.090 | 0.060 | 0.076 |
| | 1000 | 0.050 | 0.088 | 0.072 | 0.076 |
| 100 | 500 | 0.062 | 0.114 | 0.058 | 0.090 |
| | 1000 | 0.064 | 0.112 | 0.058 | 0.092 |

Table 1: Empirical sizes

➢ ***Power calculation***

$(N_{i1}-1) \sim Poi(15(\exp(\gamma a_i))$, $(N_{i0}-1) \sim Poi(12(exp(\gamma a_i))$

$\gamma = 0.1, 0.2, 0.3, 0.4, 0.5$

| $M$ | $B$ | $T_{F0}$ | $T_{CM0}$ | $T_{F1}$ | $T_{CM1}$ |
|---|---|---|---|---|---|
| 50 | 500 | 0.522 | 0.262 | 0.654 | 0.342 |
| | 1000 | 0.524 | 0.282 | 0.650 | 0.342 |
| 100 | 500 | 0.834 | 0.496 | 0.902 | 0.556 |
| | 1000 | 0.834 | 0.504 | 0.906 | 0.566 |

Table 2: Statistical power when $\gamma = 0.1$

## CONCLUSIONS

- Our bootstrap based nonparametric hypothesis testing for IICGS detection is robust in terms of being free from any distributional assumptions.
- Our method, based on TF statistic, maintains the type-1 error rate (size) at the target level of 0.05.
- Our test has high power under a variety of simulation settings. The power increases with the increase in the number of clusters.
- In future, we plan to extend our method to account for covariate(s) in addition to the grouping factor.

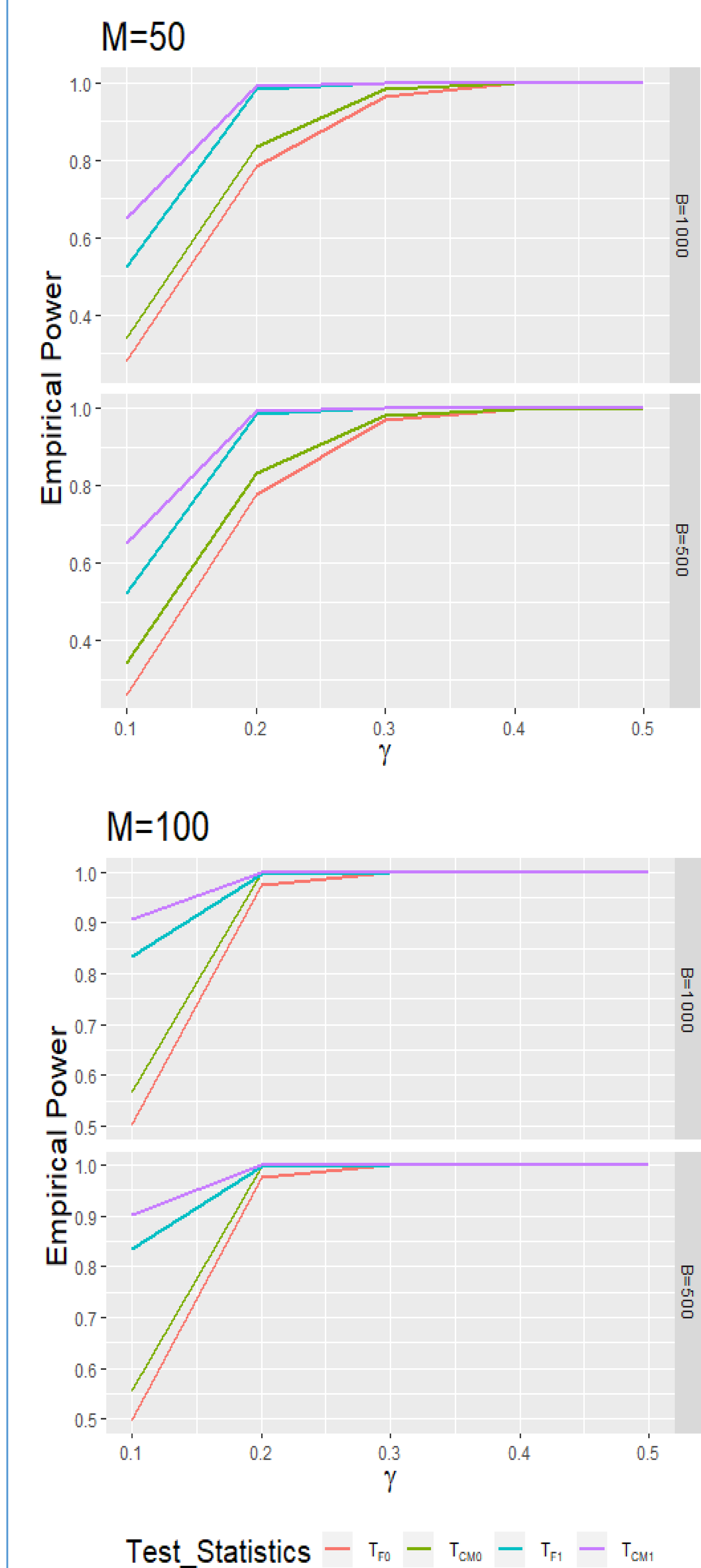


Figure 1: Power curves for different choices of M

## REFERENCES

[1] Dutta S, Datta S. *Biometrics.* 2016;72:432–40.

[2] Dutta S, Datta S. *Stat. Med.* 2018;72:4807–22.

[3] Nevalainen J, Oja H, Datta S. *Stat. Med.* 2017; 36:2630-40.