Fully Nonparametric Method for Clustered Data and Multivariate Data

Yue Cui

Department of Statistics
University of Kentucky

yue.cui@uky.edu

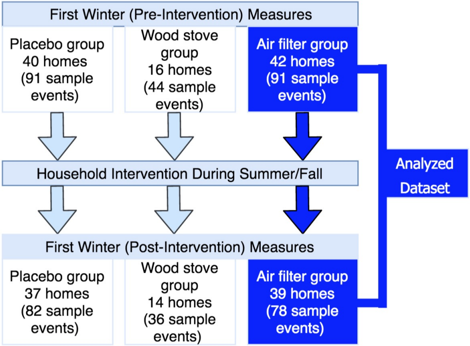
October 05, 2019

Overview

Motivating Example

Asthma Randomized Trial of Indoor Wood Smoke (ARTIS)

Figure 1: Trial Profile of ARTIS



- Pediatric Asthma Quality of Life Questionnaire (PAQLQ)
- **Question**: Does air filter intervention improve quality-of-life measures for kids with asthma?

More on ARTIS Data

- Answers to questionnaire questions: 1,2,3,4,5,6,7.
- Among all 42 homes in air-filter group, 4 of them have at least 2 kids who participate in the trial.
- At most 4 visits were made per kid both before and after air-filter intervention.
- A considerable portion of kids may miss some even all visits either before or after intervention, i.e. not all kids have data paired pre/post-intervention.
- Multiple visits on each kid before and after intervention are correlated.

Analytics Strategies

1. Item-by-item analysis with clustering.
2. Joint analysis.

Item-by-Item Analysis with Clustering

Table 1: Schematic representation.

Intervention	Pre-	Post-
Distribution	F_1	F_2
1	$x \dots x$	$x \dots x$
$\vdots$	$\vdots$	$\vdots$
n_c	$x \dots x$	$x \dots x$
$n_c + 1$	$x \dots x$	
$\vdots$	$\vdots$	
$n_c + n_1$	$x \dots x$	
$n_c + n_1 + 1$		$x \dots x$
$\vdots$		$\vdots$
$n_c + n_1 + n_2$		$x \dots x$
Count	N_1	N_2

- Keep one kid per household (randomly selected).
- Each kid serves as a cluster.
- Assume all visits follow the same distribution in each intervention period.

Statistical Model

- Data from g^{th} treatment and j^{th} cluster

$$\mathbf{X}_{gj}^{(c)} = (X_{gj}^{(c)(1)}, \cdots, X_{gj}^{(c)(m_{gj}^{(c)})})^T$$

$$\mathbf{X}_{gj}^{(i)} = (X_{gj}^{(i)(1)}, \cdots, X_{gj}^{(i)(m_{gj}^{(i)})})^T$$

  for the complete and incomplete cases where

$$X_{gj}^{(c)(d)} \sim F_g \quad \text{and} \quad X_{gj}^{(i)(d)} \sim F_g$$

  under assumption of MCAR.
- Cluster sizes are $m_{gj}^{(c)}$ and $m_{gj}^{(i)}$ for $g = 1, 2$.
- No assumption on intra-cluster dependence within or across treatment groups.

Statistical Model Cont'd

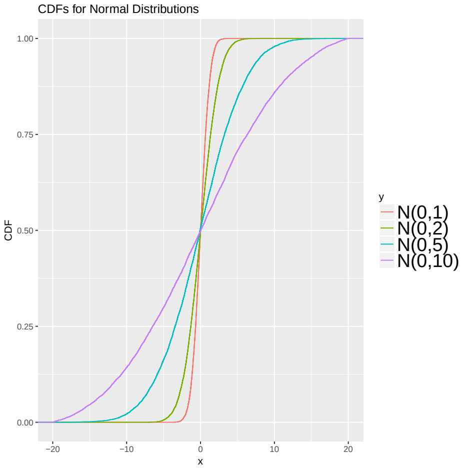
- The nonparametric effect size is

$$p = \int F_1 dF_2 = P(X_{11}^{(1)} < X_{22}^{(1)}) + \frac{1}{2} P(X_{11}^{(1)} = X_{22}^{(1)}).$$

  - If $p < \frac{1}{2}$, observations generated from distribution F_2 tend to be smaller than observations generated from distribution F_1.
  - Similar results when $p > \frac{1}{2}$.
  - If $p = \frac{1}{2}$, observations generated from F_1 and F_2 are tendentiously equal.
- No treatment effect $\rightleftharpoons H_0 : p = \frac{1}{2}$.
- Nonparametric Behrens-Fisher problem is addressed since

$$H_0 : p = \frac{1}{2} \nRightarrow F_1 = F_2.$$

Nonparametric Behrens-Fisher Problem Illustration



Estimating Nonparametric Effect p

- Plug-in weighted estimator of p is

$$\widehat{p}^{(\phi)} = \int \widehat{F}_1^{(\phi)} d\widehat{F}_2^{(\phi)}.$$

- An example of weight is

$$\phi_{gj}^{(c)} = \frac{m_{gj}^{(c)}}{N_g} \quad and \quad \phi_{gj}^{(i)} = \frac{m_{gj}^{(i)}}{N_g}.$$

- The estimator takes the form

$$\widehat{p}^{(\phi)} = \frac{1}{N_1 N_2} \cdot \text{weighted sum of} \left\{ \sum_{d=1}^{m_{1j}^{(A)}} \sum_{d'=1}^{m_{2j'}^{(B)}} c(X_{2j'}^{(B)(d')} - X_{1j}^{(A)(d)}) \right\}$$

where $A, B \in \{c, i\}$ and $c(x) = 0, \frac{1}{2}, 1$ if $x < 0, x = 0, x > 0$ is the normalized comparison function.

Assumptions

Assumption 1

$0 \leq m_{gj}^{(A)} \leq M < \infty$ for $g = 1, 2$, $A \in \{c, i\}$, where M is some constant.

Assumption 2

$n \to \infty$ such that $\frac{n}{n_c} < \infty$ or $\frac{n}{n_g} < \infty$ for $g = 1, 2$.

Note: Assumption 2 covers the practical-oriented patterns of sample sizes below:

(i) $n_c \leq K < \infty$ and $n_1, n_2 \to \infty$
(ii) $n_c \to \infty$ but $n_1, n_2 \leq K < \infty$
(iii) $n_c, n_1 \to \infty$ but $n_2 \leq K < \infty$
(iv) $n_c, n_2 \to \infty$ but $n_1 \leq K < \infty$

Asymptotic Theory

Theorems (Cui and Harrar, 2019b)

Under Assumption 1 and 2,

- $E(\widehat{p}^{(\phi)}) = p + O(\frac{n_c}{N_1 N_2})$
- $\widehat{p}^{(\phi)} \xrightarrow{a.s.} p$
- $\sqrt{N}(\widehat{p}^{(\phi)} - p) \xrightarrow{D} N(0, \sigma^{2(\phi)})$ where

$$\sigma^{2(\phi)} = \lambda_1^{(\phi)}\sigma_1^{2(\phi)} + \lambda_2^{(\phi)}\sigma_2^{2(\phi)} + \lambda_c^{(\phi)}\sigma_c^{2(\phi)}$$

and $\sigma_s^{2(\phi)}$ depends on F_1 and F_2 for $s \in \{1, 2, c\}$.

Test Procedures

Theorem (Cui and Harrar, 2019b)

Under Assumption 2,

$$\widehat{\sigma}^{2(\phi)} \xrightarrow{L_2} \sigma^{2(\phi)}.$$

- For large sample sizes, under $H_0 : p = \frac{1}{2}$,

$$T = \sqrt{N}\frac{\widehat{p}^{(\phi)} - 1/2}{\widehat{\sigma}^{(\phi)}} \to N(0,1).$$

- For small sample sizes, under $H_0 : p = \frac{1}{2}$,

$$T_{app} = \sqrt{N}\frac{\widehat{p}^{(\phi)} - 1/2}{\widehat{\sigma}^{(\phi)}} \approx t_v$$

where $v = \dfrac{\left(\frac{\widehat{\sigma}_c^{2(\phi)}}{n_c^2} + \frac{\widehat{\sigma}_1^{2(\phi)}}{N_1^2} + \frac{\widehat{\sigma}_2^{2(\phi)}}{N_2^2}\right)^2}{\left(\frac{\widehat{\sigma}_c^{2(\phi)}}{n_c^2}\right)^2/(n_c-1) + \left(\frac{\widehat{\sigma}_1^{2(\phi)}}{N_1^2}\right)^2/(n_1-1) + \left(\frac{\widehat{\sigma}_2^{2(\phi)}}{N_2^2}\right)^2/(n_2-1)}.$

Simulation Results

Table 2: Type-I Error Rate($\times$100), Discretized Multivariate Normal, $n_c = 10$, $n_1 = 20$ and $n_2 = 10$ ($\alpha = 0.05$)

$m_{gj}^{(A)}$	ρ_1	ρ_2	ρ_{12}	σ_1^2	σ_2^2	T	T_{app}
Binom(2,0.3)+1	0.9	0.9	0.1	1	1	9.2	7.6
					5	7.8	6.5
	0.1	0.9	0.9	1	1	6.7	4.6
					5	8.2	5.8
	0.1	0.1	0.9	1	1	7.1	5.3
					5	8.9	6.8
Binom(9,0.3)+1	0.9	0.9	0.1	1	1	6.7	5.7
					5	8.5	6.7
	0.1	0.9	0.9	1	1	7.7	5.3
					5	8.9	6.5
	0.1	0.1	0.9	1	1	7.2	5.3
					5	6.7	4.9

Simulation Results Cont'd

Table 3: Type-I Error Rate($\times$100), Multivariate Cauchy Distribution, $n_c = 10$, $n_1 = 20$ and $n_2 = 10$ ($\alpha = 0.05$)

$m_{gj}^{(A)}$	ρ_1	ρ_2	ρ_{12}	σ_1^2	σ_2^2	T	T_{app}
Binom(2,0.3)+1	0.9	0.9	0.1	1	1	6.3	5.2
					5	6.7	5.9
	0.1	0.9	0.9	1	1	6.2	5.6
					5	8.8	7.3
	0.1	0.1	0.9	1	1	6.3	4.7
					5	7.2	5.3
Binom(9,0.3)+1	0.9	0.9	0.1	1	1	6.7	5.9
					5	7.7	6.5
	0.1	0.1	0.9	1	1	6.8	5.4
					5	8.7	7.1
	0.1	0.9	0.9	1	1	7.7	5.9
					5	6.4	4.8

Simulation Results Cont'd

Table 4: Power($\times$100), $n_c = 20$, $n_1 = n_2 = 10$, $\rho_1 = 0.1$, $\rho_2 = \rho_{12} = 0.9$, F_1 and F_2 are linear combinations of CDFs of $N(0, 1)$ and $N(-15, 5)$

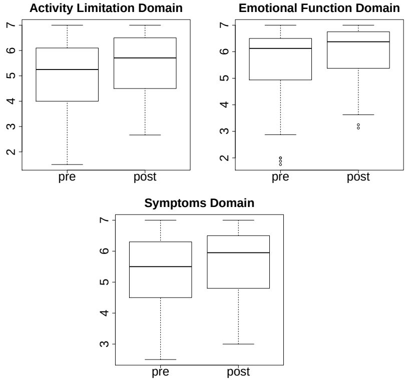
$m_{gj}^{(A)}$	p	T	T_{app}
Binom(2,0.3)+1	0.5	5.3	4.4
	0.55	12.8	11.0
	0.6	32.0	30.2
	0.65	56.8	54.6
	0.7	82.4	80.8
	0.75	94.7	93.7
	0.8	99.5	99.5
	0.849	100.0	100.0
Binom(9,0.3)+1	0.5	7.0	6.1
	0.55	10.8	9.6
	0.6	31.3	29.1
	0.65	59.1	56.8
	0.7	82.2	80.1
	0.75	96.4	96.2
	0.8	99.4	99.2
	0.849	100.0	100.0

Trial Profile of ARTIS

Table 5: Summary Table for Measures on Selected Kids

	Family Counts	Events Count
Pre	Complete 35	79
	Incomplete 7	12
Post	Complete 35	78
	Incomplete 0	0

Figure 2: Box plot of the Quality of Life scores for each domain.



Tests on Domain Variables

Table 6: Summary Test Results for Domain Variables in ARTIS data.

Domain	Test Statistic	p-value	$\widehat{p}^{(\phi)}$	95% CI
Activity Limitation	3.96	0.0003	0.604	(0.553, 0.656)
Emotional Function	2.89	0.0057	0.579	(0.526, 0.633)
Symptoms	2.75	0.0060	0.580	(0.523, 0.637)

Joint Multivariate Analysis

Simple Missing Pattern (Brunner et al., 2002)

- Consider a clinical trial where d endpoints are assessed at two treatments:

Treatment	TX=1			TX=2		
Component	1	$\cdots$	d	1	$\cdots$	d
Distribution	$F_1^{(1)}$	$\cdots$	$F_1^{(d)}$	$F_2^{(1)}$	$\cdots$	$F_2^{(d)}$
1	x	$\cdots$	x	x	$\cdots$	x
$\vdots$		$\vdots$			$\vdots$	
n_c	x	$\cdots$	x	x	$\cdots$	x
$n_1 + 1$	x	$\cdots$	x			
$\vdots$		$\vdots$				
$n_c + n_1$	x	$\cdots$	x			
$n_c + n_1 + 1$				x	$\cdots$	x
$\vdots$					$\vdots$	
$n_c + n_1 + n_2$				x	$\cdots$	x

- Are the two treatments different on all component jointly?

Statistical Model

- For the complete as well as incomplete data for k^{th} subject of ℓ^{th} component in g^{th} group

$$X_{gk}^{(c)(\ell)}, X_{gk'}^{(i)(\ell)} \overset{iid}{\sim} F_g^{(\ell)}$$

for $g = 1, 2,\ k = 1, \ldots, n_c,\ k' = 1, \ldots, n_g$ and $\ell = 1, \ldots, d$.

- The total number of subjects is n.
- Assumption 1 and 2 still apply here.
- No assumptions on dependence structures among components.

- Derive nonparametric relative effect for each component, i.e. for ℓ^{th} component, $\ell = 1, \cdots, d$,

$$p^{(\ell)} := \int F_1^{(\ell)} dF_2^{(\ell)} = p(X_{11}^{(c)(\ell)} < X_{21}^{(c)(\ell)}) + \frac{1}{2} p(X_{11}^{(c)(\ell)} = X_{21}^{(c)(\ell)}).$$

- The plug-in estimator of $p^{(\ell)}$

$$\widehat{p}^{(\ell)} = \int \widehat{F}^{(\ell)}_{1,\theta_1^{(\ell)}} d\widehat{F}^{(\ell)}_{2,\theta_2^{(\ell)}}.$$

- Nonparametric relative effect size vector: $\boldsymbol{p} = (p^{(1)}, \ldots, p^{(d)})'$.
- $H_0 : \boldsymbol{p} = \frac{1}{2}\mathbf{1}_d$ v.s. $H_a : \boldsymbol{p} \neq \frac{1}{2}\mathbf{1}_d$.

Theoretical Results

Assumption 3

Let $\lambda_1, \ldots, \lambda_d$ denote eigenvalues of $\boldsymbol{V}_n = \mathrm{Cov}(\sqrt{n}(\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d))$ and let $\lambda_{min} = min\{\lambda_1, \ldots, \lambda_d\}$ denote the smallest eigenvalue, then

$$\lambda_{min} \geq \lambda_0 > 0$$

where λ_0 is some constant.

Theorems (Cui and Harrar, 2019a)

Under Assumption 1,2 and 3,

- $E(\widehat{\boldsymbol{p}}) = \boldsymbol{p} + O(\frac{n_c}{m_1 m_2})$ where $m_g = n_c + n_g$ for $g = 1, 2$
- $||\widehat{\boldsymbol{p}} - \boldsymbol{p}||_2 = O(\frac{1}{n})$
- $\sqrt{n}(\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{V}_n)$

Test Procedures

Under Assumption 3 and Theorems (Cui and Harrar, 2019a),

- For large sample sizes, under the null hypothesis $H_0 : \boldsymbol{p} = \frac{1}{2}\mathbf{1}_d$,

$$Q_n = n \cdot (\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d)' \widehat{\boldsymbol{V}}_n^{-1} (\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d) \sim \chi^2_d.$$

- For small sample sizes, under the null hypothesis $H_0 : \boldsymbol{p} = \frac{1}{2}\mathbf{1}_d$,

$$F_n = \frac{n}{tr(\widehat{\boldsymbol{V}}_n)} (\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d)' (\widehat{\boldsymbol{p}} - \frac{1}{2}\mathbf{1}_d) \sim F(\widehat{F}, \infty)$$

where $\widehat{F} = \frac{[tr(\widehat{\boldsymbol{V}}_n)]^2}{tr(\widehat{\boldsymbol{V}}_n^2)}$.

Table 7: Type-I error rates after multiple imputation with 5 chains, sample sizes are $n_c = n_2 = 30$ and $n_1 = 10$. $\alpha = 0.05$.

			Rounded Multivariate Normal			Multivariate Log-Normal			Multivariate Cauchy		
$(\rho_1, \rho_2, \rho_{12})$	(σ_1^2, σ_2^2)	d	Q_n	F_n	Multi-Impute	Q_n	F_n	Multi-Impute	Q_n	F_n	Multi-Impute
(-0.4,-0.4,-0.4)	(1,1)	2	7.3	6.1	5.6	6.6	5.6	4.9	7.2	6.3	2.8
		3	6.2	5.4	6.2	5.8	4	6.1	8.9	6.2	3.2
		5	7.4	4.6	6.4	8.7	4.5	4.7	10.5	5.5	1.3
	(1,5)	2	7.1	6.7	6.5	5.6	4.7	**69.4**	6.7	5	2.3
		3	8.5	6.8	7.7	6.9	5.3	**76.4**	6.3	4.3	2.7
		5	8.2	4.8	7.8	7.5	4.8	**93.2**	8.6	4.5	1.1
(0.4,0.4,0.4)	(1,1)	2	6.4	5.4	6.7	4.7	4.5	6.9	7.2	6.7	2.8
		3	7.1	5.5	5.4	6.6	5.6	5.9	6.4	4.9	2.2
		5	8.1	5.1	5.7	9.9	6.4	6.1	8.7	5.5	1.2
	(1,5)	2	4.6	4.6	6.8	6.1	5.1	**69.2**	6.7	5.1	2.3
		3	6.2	5.9	5.6	6.4	6.1	**78.9**	7.2	6.2	1.2
		5	9.1	6.0	5.8	8.9	7.5	**84.1**	10.1	5.9	1.2

Simulation Results Cont'd

Table 8: Obtained power after multiple imputation with 5 chains, sample sizes are $n_c = n_2 = 30$ and $n_1 = 10$.

$(\rho_1, \rho_2, \rho_{12})$	(σ_1^2, σ_2^2)	δ_1	δ_2	Q_n	F_n	Multiple Imputation
(-0.4,-0.4,-0.4)	(1,1)	0	0.3	20.7	17.9	20.2
		0.3	0.3	40.5	34.7	37.8
		0.6	0.6	93.8	92.6	94
		0.9	0.9	99.9	100	99.9
		0.3	0.6	80.9	78.1	77.5
		0.3	0.9	97.6	96.6	96.9
	(1,5)	0	0.3	11.8	9.9	10.9
		0.3	0.3	19.5	16.5	20.6
		0.6	0.6	56	49.5	59.6
		0.9	0.9	90.6	87.5	93.8
		0.3	0.6	38	33.4	43.7
		0.3	0.9	64.8	59.1	66.7
(0.4,0.4,0.4)	(1,1)	0	0.3	28.9	25.7	28.7
		0.3	0.3	45.5	50	46.8
		0.6	0.6	98	98.7	96.8
		0.9	0.9	100	100	100
		0.3	0.6	88.2	89.7	86.3
		0.3	0.9	99.8	99.6	99.6
	(1,5)	0	0.3	12.5	11.2	13.2
		0.3	0.3	19.8	19.7	20.7
		0.6	0.6	60	64	65.2
		0.9	0.9	91.5	92.9	92.9
		0.3	0.6	43.2	45	46.6
		0.3	0.9	74.5	75.5	79.1

Tests on Domain Variables

Table 9: Summary Test Results for Domain Variables in ARTIS

	All			Complete			Multiple Imputation
Domain	$\widehat{p}$	Q_n	F_n	$\widehat{p}$	Q_n	F_n	p-value
Activity Limitation	0.581	<0.001	<0.001	0.564	<0.001	<0.001	0.322
Emotional Function	0.568			0.581			
Symptoms	0.545			0.567			

Flexible Missing Patterns

- Example of $d = 2$, ? represents missing value.

Treatment	TX=1		TX=2	
Component	1	2	1	2
Distribution	$F_1^{(1)}$	$F_1^{(2)}$	$F_2^{(1)}$	$F_2^{(2)}$
1	x	x	x	x
2	?	x	x	x
3	x	?	x	x
4	x	x	?	x
5	x	x	x	?
6	?	?	x	x
7	?	x	?	x
8	?	x	x	?
9	x	?	?	x
10	x	?	x	?
11	x	x	?	?
12	x	?	?	?
13	?	x	?	?
14	?	?	x	?
15	?	?	?	x

- For $d > 2$, $2^{2d} - 1$ different missing patterns in total.

Conclusions and Summary

Conclusions

- Accommodates binary, ordinal, discrete and continuous data seamlessly; allows meaningful probabilistic comparison of treatments with flexible and precise alternatives; provides numeric measurement of effect size.
- No assumptions needed on between-cluster or intra-cluster dependence structures.
- Favorable performance for data generated from different distributions and also for multiple cluster sizes settings.
- Introduce nonparametric test to comparisons between more than two groups and consider more missing structures.
- Propose a nonparametric test procedure which can be used directly to analyze ARTIS data and also consider about gender effect.

References I

Brunner, E., U. Munzel, and M. L. Puri (2002). The multivariate nonparametric behrens-fisher problem. *Journal of Statistical Planning and Inference 108*, 37–533.

Cui, Y. and S. W. Harrar (2019a). The multivariate nonparametric behrens-fisher problem.

Cui, Y. and S. W. Harrar (2019b). Nonparametric behrens-fisher for partially complete dependent replicates.

Thank you!