Nonparametric analysis of recurrent events with incomplete observation gaps

Jinheum Kim
University of Suwon
Korea

2011-12-17

Joint with Yang-Jin Kim, Eun Hee Choi, Chung Mo Nam

Recurrent event data

- Recurrent events arise frequently in clinical trials in which patients are followed longitudinally and response of interest is transient

- Examples
  - Transient ischemic attacks in patients with cerebrovascular disease
  - Seizures in epileptic patients
  - Tumor recurrences in cancer patients

Motivating example

- Young Traffic Offenders Program(YTOP) data

- What is YTOP?
  - Conducted at the University of Missouri-Columbia Health Sciences Center and other medical centers since late 1987
  - One-day educational intervention for people under 24 convicted of speeding by 20+ mph over the posted speed limit

YTOP: objective

To evaluate the effect of the program on reducing the rates of speeding violation of young drivers

YTOP: characteristics

- Data were collected on 192 young drivers about their speeding violation information since they obtained their driving license
- 29 subjects(13: YTOP, 16: non-YTOP) received suspensions
- These subjects dropped out of the study when the suspension began, and then came back to the study after completing suspension
- These intermittent dropouts are called **observation gaps**
- The observation gap is determined by the starting and terminating time of the suspension
- BUT, the observation gaps of YTOP data are **incomplete** due to missing terminating times

Previous works & their limitations

- Farmer et al.(2000, Brain Injury): two-sample rank test
  - Ignore the detailed information about the conviction process
  - Can assess only the long-term impact

- Sun et al.(2001, JASA): nonparametric & semi-parametric methods
  - Treat YTOP data as recurrent event data
  - Evaluate both short- and long-term effects applying the conviction process information
  - Ignore the observation gaps

Previous works & their limitations

- Zhao & Sun(2006, CSDA)
  - Similar to Sun et al. (2001, JASA), but consider the observation gaps
  - Suspension periods: NOT deterministic, BUT varying and possibly unknown

- Kim & Jhun(2008, StatMed)
  - Treat incomplete observation gaps as terminating times being interval-censored
  - Only use the first suspension and ignore all the other suspensions following the first one.

Objective of this talk

- Use the multivariate-interval censored data to estimate the distribution of the terminating time

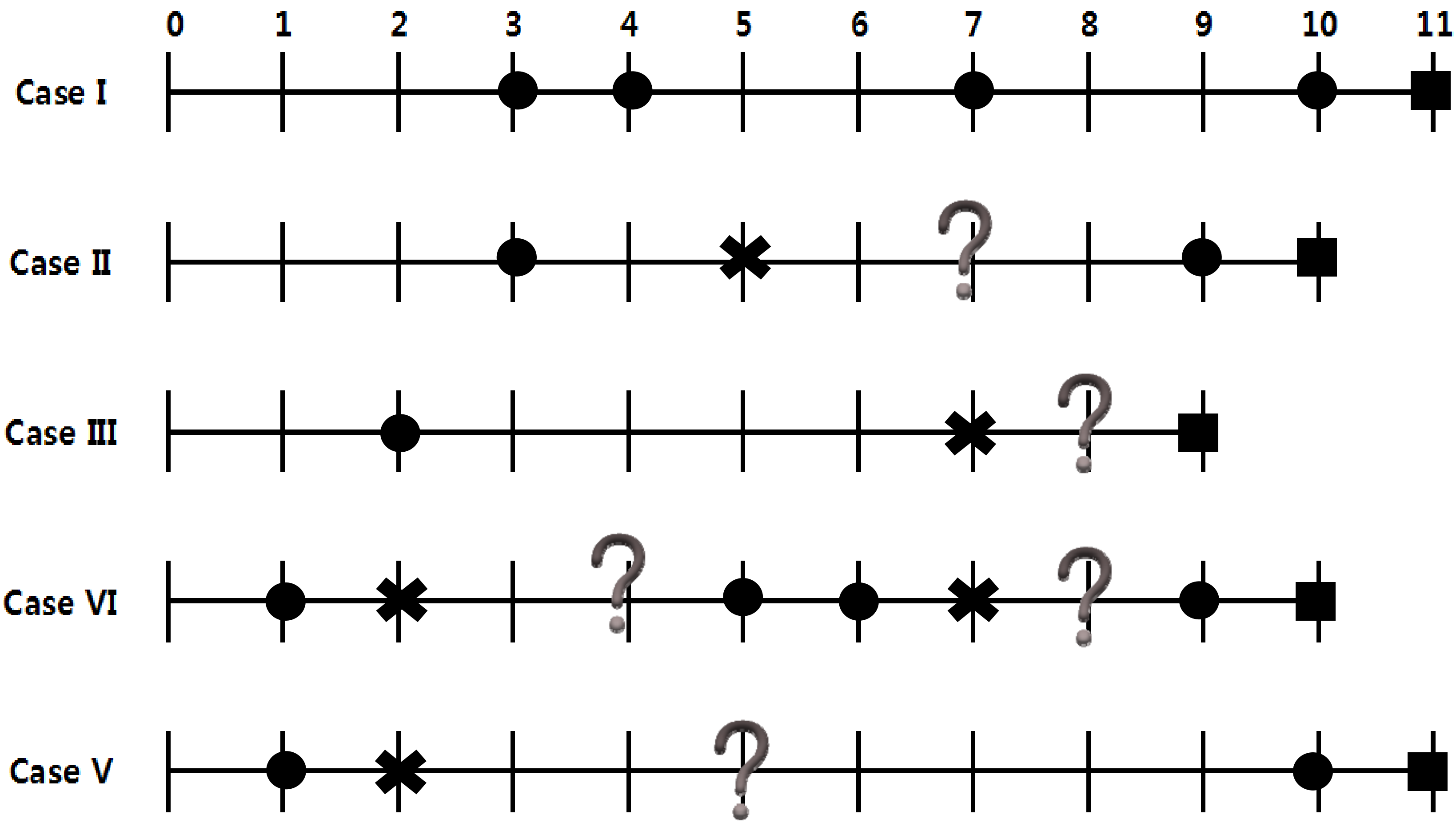
- Propose a non-parametric test to compare the conviction rates of two groups

- Evaluate the proposed test via simulation & reanalyze the YTOP data

Illustrative data: notation

- ●: conviction time
- X: starting time of the suspension
- ?: terminating times of the suspension **(unknown)**
- ■ : end of study time

Diagram



IC data: objective

Want to estimate the distribution of the terminating time using the interval-censored data to incorporate the observation gaps

Two types of IC data

- Univariate IC data
  - cases II, III, & V having one observation gap
  - eg: (5,9], (7, ∞), (2,10]

- Multivariate IC data
  - case IV having two observation gaps
  - eg: (2,5]. BUT, …

Multivariate IC data

How can we define the origin of the second terminating time?

Constructing IC data

- Set the origin to be the first conviction time after the previous suspension
  - eg: For the second suspension of case IV,

$$(7-5=2,9-5=4]$$

- Hence, estimate the distribution of the terminating time using

$$(5,9],\ (7,\infty),(2,10],(2,5]\quad+\quad(2,4]$$

EM-ICM

- Hybrid algorithm(Wellner & Zhan, 1997, JASA)
- Use R package **interval**

Notation & assumption

- $N_i(t)$:the number of occurrences of the event up to time $t, \ i = 1, \ldots, n$
- $Y_i(t)$: function indicating if subject i is under observation at time t
- $\Lambda_i(t) = E\{N_i(t)\}$: cumulative mean (frequency) function (CMF)
- N_i & Y_i : independent

Data

- $\mathrm{D} = \{(t_{ik}, \tilde{Y}_i, h_{ij}), i = 1, \ldots, n; k = 1, \ldots, n_i; j = 1, \ldots, m_i\}$
- t_{ik} : the kth recurrent event time of subject i
- $\tilde{Y}_i$: 0-1 indicator whether or not subject i has experienced observation gap
- h_{ij} : starting time of the jth observation gap of subject i

Unobservable terminating times

- BUT, the terminating time of each observation gaps is NOT available, that is, interval-censored

- So, need to estimate the distribution of the terminating time of the observation gap

IC data

- $I = \{(L_{ij}, R_{ij}], i = 1, \ldots, n; j = 1, \ldots, m_i\}$
  - $L_{ij} = h_{ij} - t_{ij0}, R_{ij} = t_{ij1} - t_{ij0}$

- t_{ij0} : minimum of t_{ik} greater than $h_{i,j-1}$
- t_{ij1} : minimum of t_{ik} greater than h_{ij}
- R_{ij} can be right-censored

Terminating time distribution

- S_{ij} : the terminating time for the jth observation gap of subject i
- $a_l \ (l = 1, \ldots, d)$: the mid-point of the lth equivalence class
- $f_l = P(S_{ij} = a_l)$: mass of S_{ij} at a_l

Redefine a risk set

Replace $Y_i(t)$ by the estimated expected risk set defined as

$$Y_i^*(t) = 1 - \tilde{Y}_i \sum_{j=1}^{m_i} \sum_{l=1}^{q} \frac{J_{ij}(t) I(t - t_{ij0} \le a_l \le R_{ij}) \hat{f}_l}{\sum_{j=1}^{m_i} \sum_{r=1}^{q} J_{ij}(t) I(L_{ij} < a_r \le R_{ij}) \hat{f}_r}$$

Equivalence classes

Case II: (L R]

Case III: (L

Case IV: (L R]

(L R]

Case V: (L R]

Combine L R L L R R

R

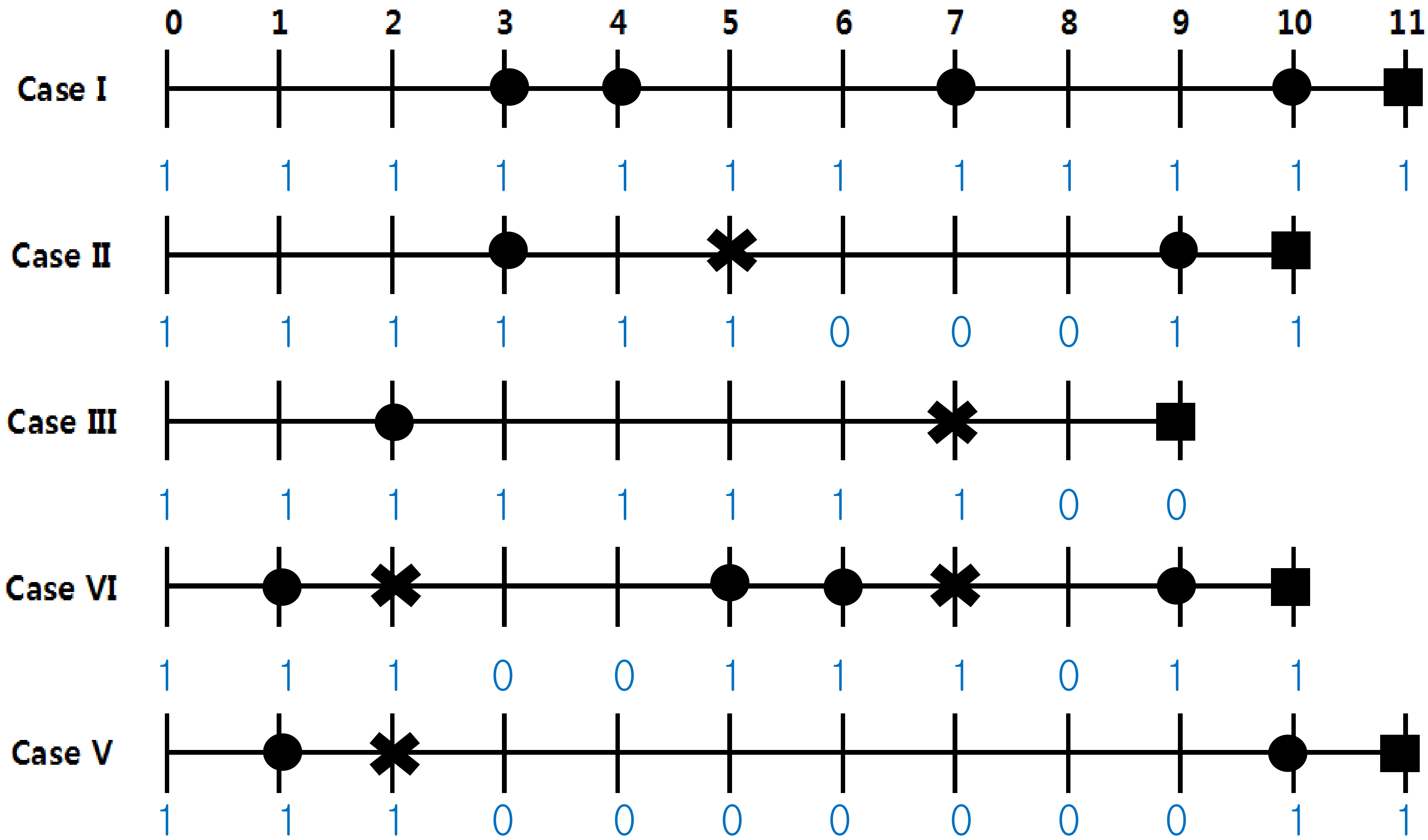
Equivalence classes: (2,4], (5,5], (7,9]

Mid-points: 3, 5, 8

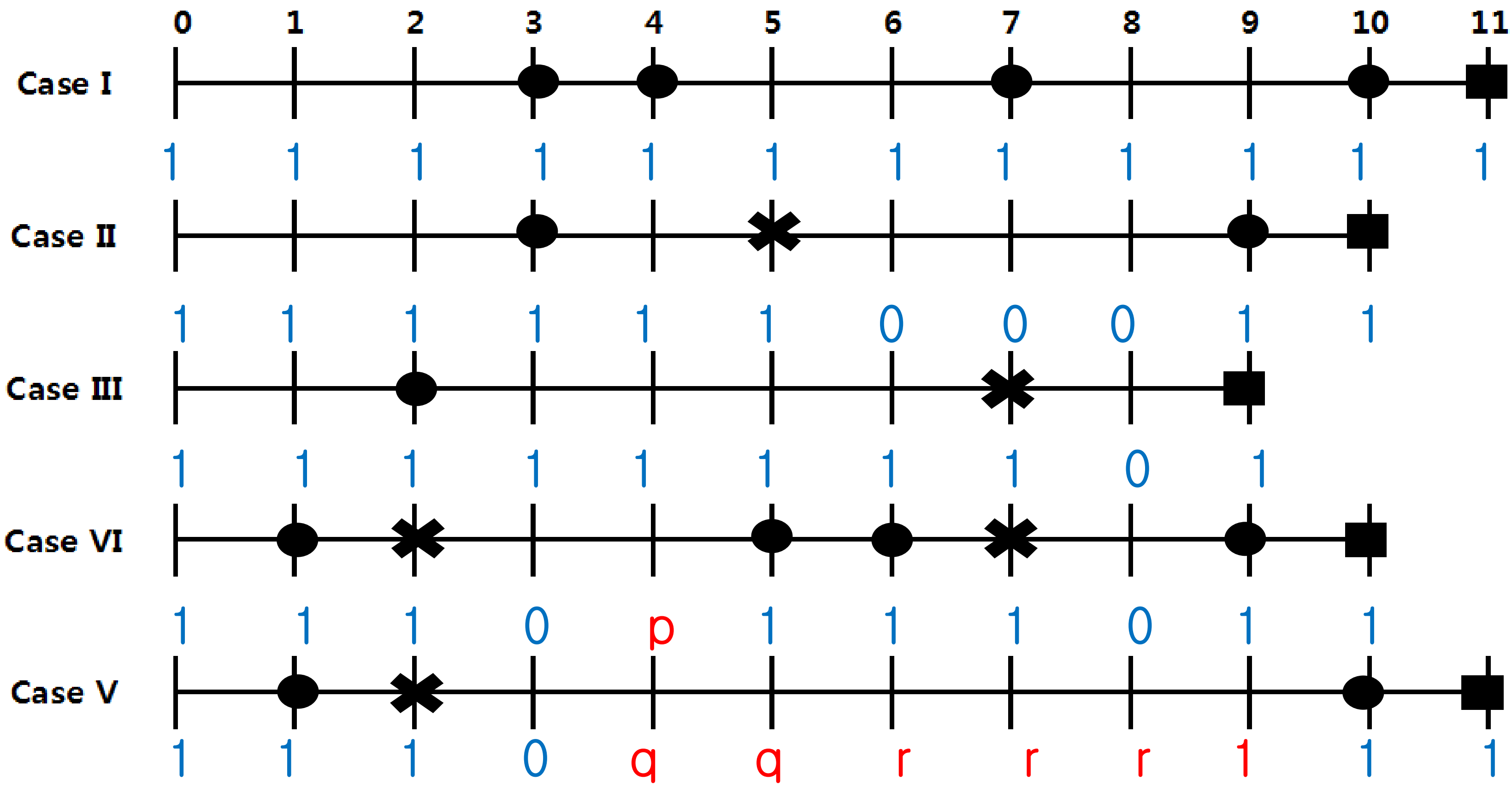
Assume $P(S=3)=f_1, P(S=5)=f_2, P(S=8)=f_3,$ with $f_1+f_2+f_3=1$

S: terminating time

Unadjusted(naive) risk set



Modified(proposed) risk set



$$p=\frac{f_1}{f_1+f_2},\qquad q=\frac{f_1}{f_1+f_2+f_3}\qquad r=\frac{f_1+f_2}{f_1+f_2+f_3}$$

Estimated CMF

- Assuming that all subjects have the common CMF,

$$\Lambda_1(t) = \Lambda_2(t) = \cdots = \Lambda_n(t) = \Lambda(t),$$

$$\hat{\Lambda}(t) = \int_0^t \frac{dN_\bullet^*(s)}{Y_\bullet^*(s)}$$

- $dN_\bullet^*(t) = \sum_{i=1}^{n} Y_i^*(t)\,dN_i(t)$: total number of events observed at time t
- $Y_\bullet^*(t) = \sum_{i=1}^{n} Y_i^*(t)$: total number of subjects at risk at time t

Representation of estimated CMF

- $t_1 < t_2 < \cdots < t_d$: distinct event times across all individuals

$$\hat{\Lambda}(t) == \sum_{j:t_j \le t} \frac{dN_{\bullet}^{*}(t_j)}{Y_{\bullet}^{*}(t_j)}$$

- Similar to the Nelson-Aalen estimator from survival analysis
- Cook & Lawless(2007)

Asymptotic properties

- Following Lin et al.(2001, JRSSB),
  - Consistent
  - Asymptotically normal with consistently estimated variance,

$$\hat{v}ar\{\hat{\Lambda}(t)\} = \sum_{i-1}^{n}\{\sum_{j:t_j \le t} \frac{Y_i^*(t_j)}{Y_\bullet^*(t_j)}[dN_i(t_j) - \frac{dN_\bullet^*(t_j)}{Y_\bullet^*(t_j)}]\}^2$$

- Approximate $1-\alpha$ pointwise CI:

$$\exp(\log \hat{\Lambda}(t) \pm z_{\alpha/2}\sqrt{\hat{v}ar\{\log \hat{\Lambda}(t)\}})$$

Comparison of two groups

- $\Lambda_1(t), \Lambda_2(t)$: the CMFs for group 1 and 2
- $n = n_1 + n_2$: the number of subjects in the gth group with n_1, n_2
- $\hat{\Lambda}_1(t), \hat{\Lambda}_2(t)$: the estimated CMF for the gth group

Two-sample test statistic

- Hypothesis: $H_0 : \Lambda_1(t) = \Lambda_2(t), \forall t > 0$

- Test statistic:

$$T_w = \int_0^{\tau} w(s)\{d\hat{\Lambda}_1(s) - d\hat{\Lambda}_2(s)\}$$

  - $\tau(>0)$: maximum follow-up time across two groups
  - $w(t)$: non-negative predictable function

Asymptotic distribution

- Following Cook et al.(1996, BCS) & Lin et al.(2001, JRSSB),
  - Asymptotically normal with consistently estimated variance,

$$\hat{var}(T_w) = \sum_{g=1}^{2} \sum_{k=1}^{n_g} [\int_0^{\tau} w(s) \frac{Y_{gk}^{*}(s)}{Y_{g\bullet}^{*}(s)} \{dN_{gk}(s) - d\hat{\Lambda}_g(s)\}]^2$$

- Hence, $T_w^2 / \hat{var}(T_w) \sim \chi^2(1)$ asymptotically

Simulation: data generation

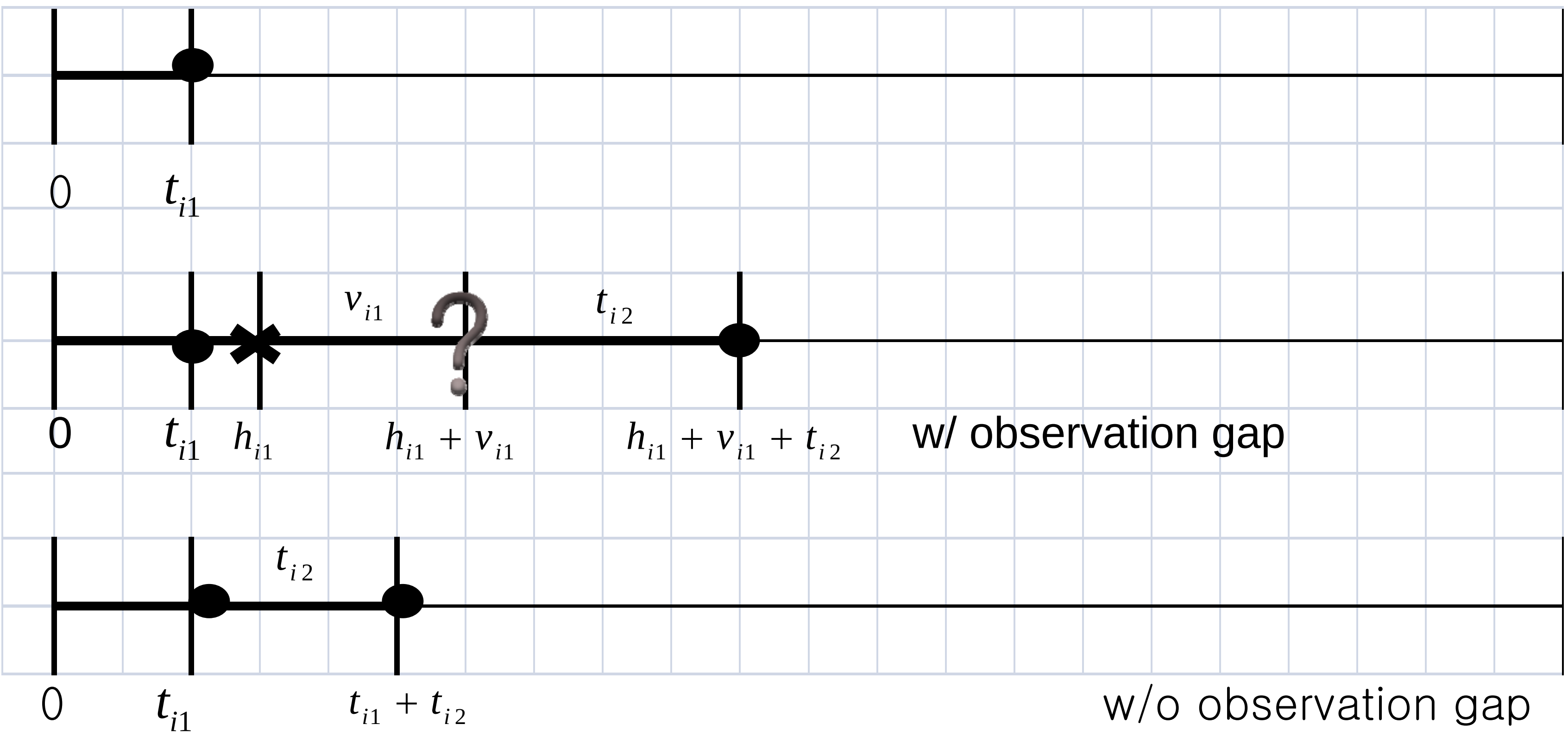
- Generate the first event time from t_{i1} an exponential distribution with a hazard rate of λ_i

- Generate $\delta_{i1} \sim B(1, p)$
  - p : probability of having an observation gap

Simulation: data generation

- If $\delta_{i1} = 1$, generate the starting time h_{i1} of the first observation gap from $U(t_{i1}, t_{i1} + 005)$
- If $\delta_{i1} = 1$, generate the duration tine v_{i1} from an exponential distribution with a hazard rate of 2 resulting in the terminating time $h_{i1} + v_{i1}$
- So, we get t_{i1} & h_{i1} (if $\delta_{i1} = 1$)

Simulation: the first cycle of data generation



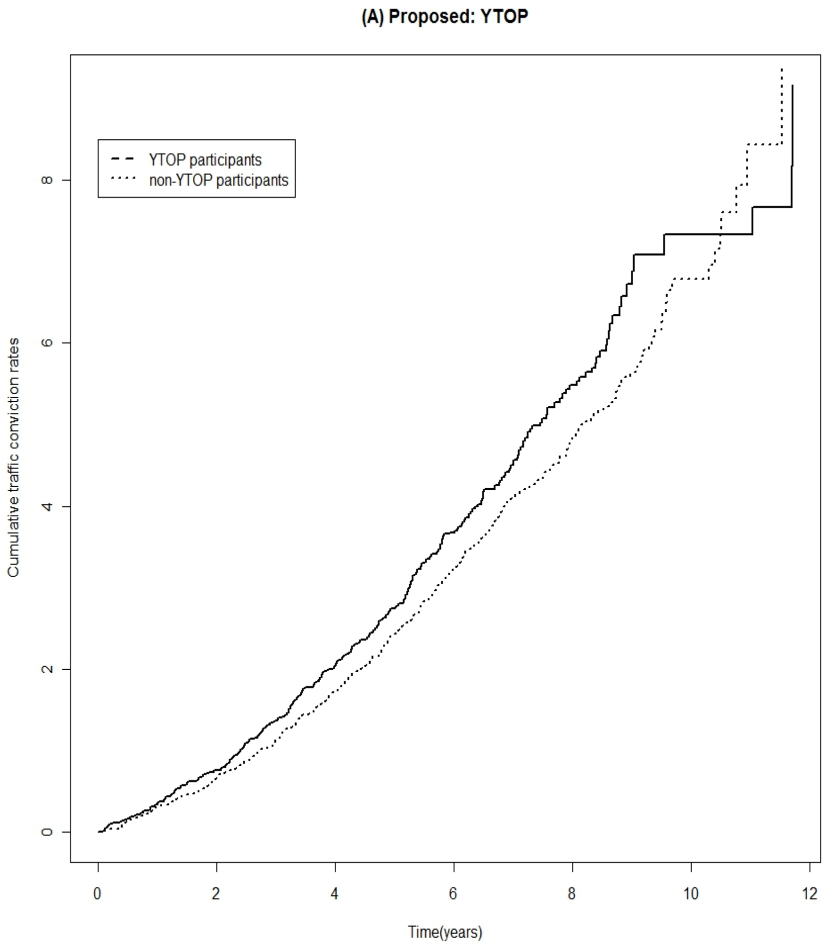
Two-sample tests: design params

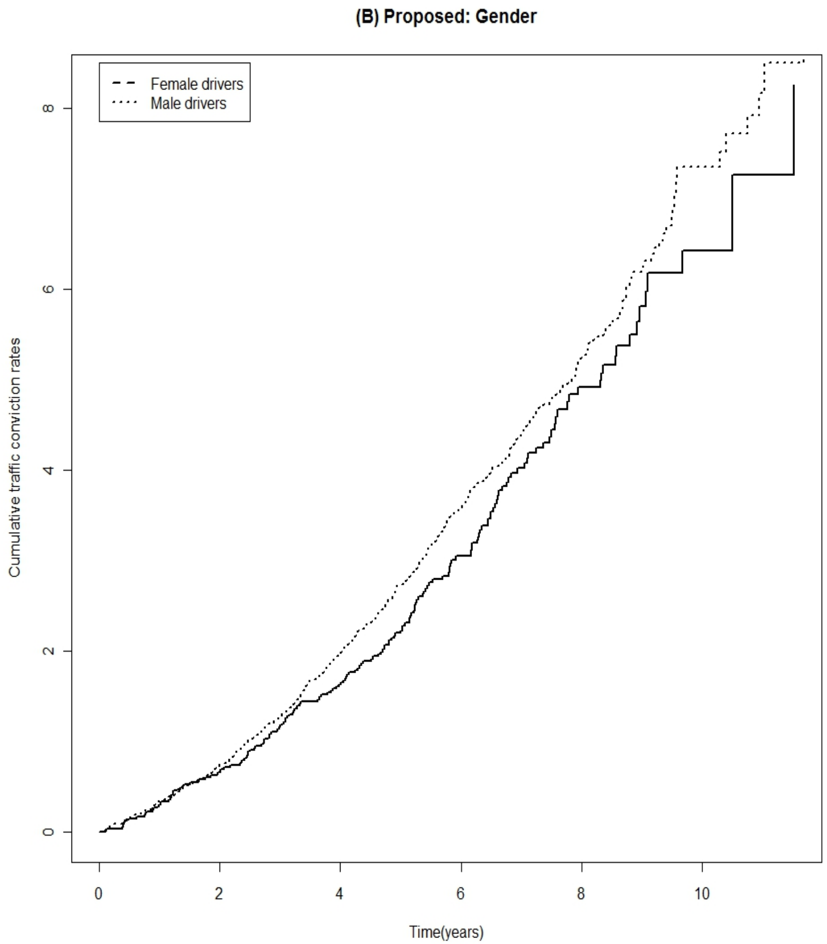
- Event rate of each group
  - $\lambda_1 = 1.5$
  - $\lambda_2 = 1.5, 1.4, 1.3, 1.2, 1.1$
- Weight: $w(t) = 1$
- p =0.05, 0.1, 0.2, 0.4
- Fixed censoring at 5
- Sample size: $n = 50, 100$
- Replication: 1,000

Two-sample tests: results(n=100)

p	λ_2	Tests			
		Proposed	Naive	Sun et al.	Farmer et al.
0.05	1.5	0.047	0.055	0.047	0.047
	1.4	0.271	0.273	0.263	0.190
	1.3	0.755	0.740	0.740	0.593
	1.2	0.978	0.972	0.979	0.918
0.1	1.5	0.060	0.055	0.056	0.039
	1.4	0.246	0.231	0.248	0.182
	1.3.	0.704	0.668	0.677	0.538
	1.2	0.981	0.966	0.980	0.907
0.4	1.5	0.062	0.049	0.068	0.054
	1.4	0.190	0.151	0.179	0.138
	1.3.	0.633	0.440	0.576	0.423
	1.2	0.941	0.793	0.914	0.772

YTOP data: estimated CMFs





YTOP data: p-values

Tests	Covariate	
	YTOP	Gender
Proposed	0.836	0.135
Naive	0.735	0.161
Sun et al.	0.456	0.008
Farmer et al.	0.113	0.533

Summary

- Propose the expected risk set to incorporate the observation gaps

- Our two-sample test shows to satisfy the nominal level & has the higher power than the other tests

- No significant difference in traffic conviction rates between YTOP and non-YTOP participants & between male and female

Extend to …

- A regression model to include a time-dependent covariate $Z(t) = I(T \leq t)$, where T denotes the time at which a subject undergoes the traffic education program

$$\Lambda_i(t) = E\{N_i(t) \mid z_i(t)\} = \lambda_0(t)\exp\{\beta' z_i(t)\}$$

- A case to allow a termination event such as death

Thank you!