久留米大学バイオ統計セミナー講演

Testing Quasi-Independence for Truncation Data

国立交通大学統計学研究所、江村剛志

Joint work with Weijing Wang

発表内容

- 自己紹介
- Literature review, quasi-independence
- 既存の検定方法

  Tsai法, Marting & Betensky法
- 提案する検定方法 (JMVA, Emura & Wang, 2010)
- 3つの方法の比較
- Doubly-truncated data への拡張 (時間があれば)
- 台湾での写真×２

自己紹介

- 千葉大院生時代;
  今野先生の下でFleming &Harrington（1991）, *Counting Process and Survivial Analysis* を読む

- 博士課程の途中で国立交通大学へ留学
  博士論文;Statistical Inference for Dependent Truncation Data（2007, advisor; Weijing Wang）

- 北里大学、臨床統計部門へ1年勤務

- 現在、国立交通大学でポスドク2年目
  （Supervisor; Weijing Wang）

- 今後・・・
  Academia Sinicaの Assistant Research Fellow及び
  ポスドクの申請中

Literature review

Censored Data (for comparison)

- In survival analysis, censoring is assumed "independent"

- Observe $\{X_j \wedge Y_j, I(X_j \le Y_j);\ j = 1, \ldots, n\}$

  where $X_j \wedge Y_j \equiv \min\{X_j, Y_j\}$ and $X_j \perp Y_j$

- Nelson-Aalen type estimator

$$\frac{d\bar{N}(u)}{\bar{Y}(u)} = \frac{\sum_{j=1}^{n} I(X_j = u, Y_j \ge u)}{\sum_{j=1}^{n} I(X_j \ge u, Y_j \ge u)} \xrightarrow{n\to\infty} \frac{\Pr(X_j = u, Y_j \ge u)}{\Pr(X_j \ge u, Y_j \ge u)} = \frac{\Pr(X_j = u)}{\Pr(X_j \ge u)} = d\Lambda(u)$$

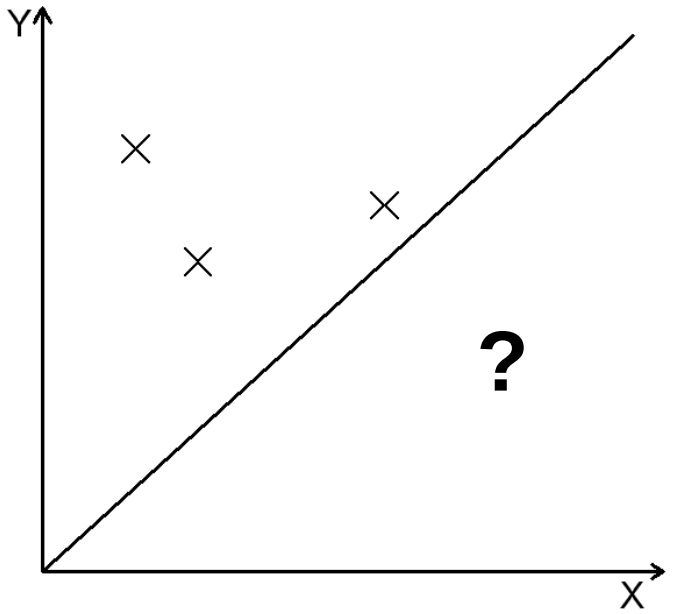
Truncation vs. Censoring

- Truncation data:

$\{(X_j, Y_j)\,(j = 1, \ldots, n)\}$

subject to $X_j \le Y_j$

Assumption; $X \perp Y$

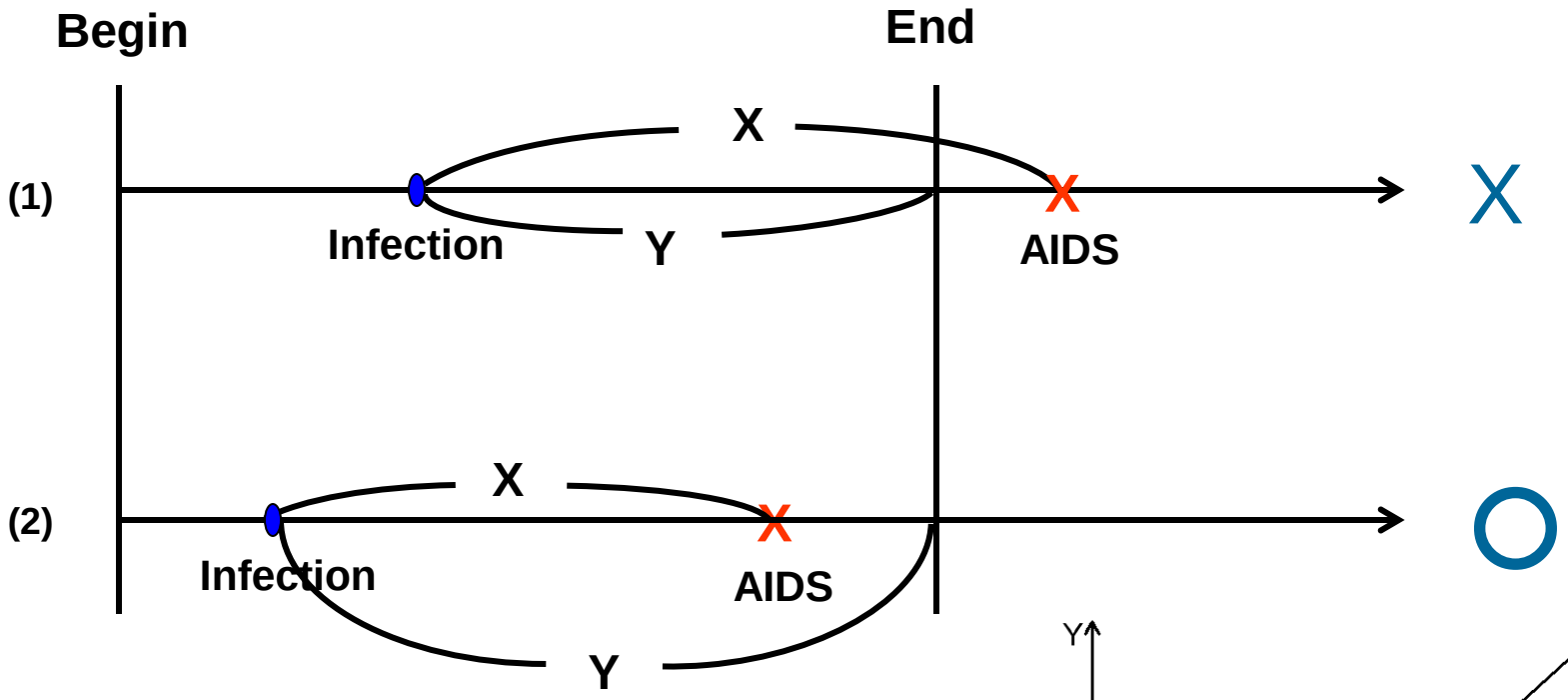


- Censored data

$\{\min(X_j, Y_j), I(X_j \le Y_j);\ j = 1, \ldots, n\}$

Example: transfusion related AIDS

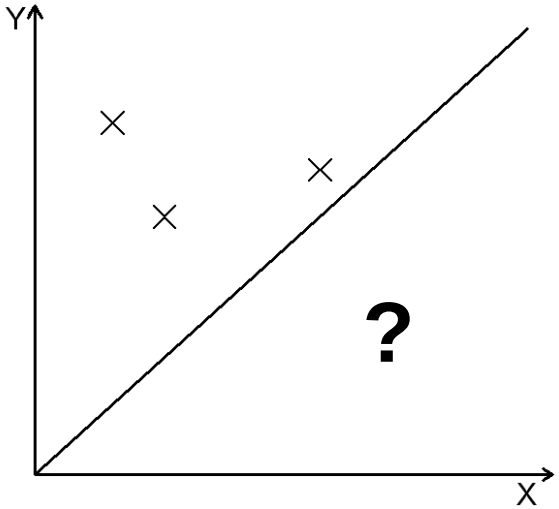
Lagakos et al. (1988)



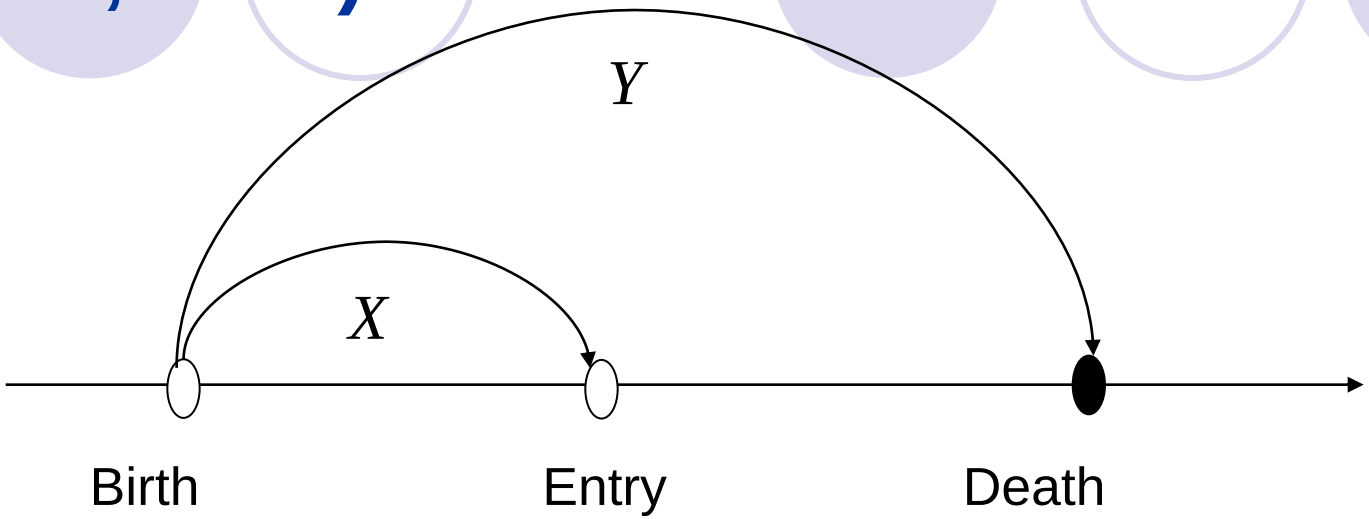
X : Infection to AIDS (interest)

Y : Infection to end of study

* Are X and Y independent?



Example 2: Channing House data (Hyde, 1977)



- Objectıve: wnether living in the retirement center (Channing house) prolonged the lifetime
  ➔Estimate S(t)=Pr(Y>t)
- Only X < Y can be included in the study
- Are X and Y independent?

Literature Review on Truncation

- Under independence between X and Y
  - estimate marginal distribution (Lynden-Bell, 1971)
  - estimate $\Pr(X \le Y)$ (He and Yang, 1998)
- **Test quasi-independence**
  - Tsai (1990); Martin & Betensky (2005)
  - Emura and Wang (2010, JMVA)
- Investigate association between X and Y
  - Copula-based approach
    (Chaieb et al., 2006),
    (Emura et al. Accepted to *Statistica Sinica*)
  - Bivariate normal approach (Emura & Konno, In revision)

★発表の目的;
3つの方法を説明

Under complete data

- Data: $\{(X_i, Y_i); i = 1, \ldots, n\} \sim F(x, y)$

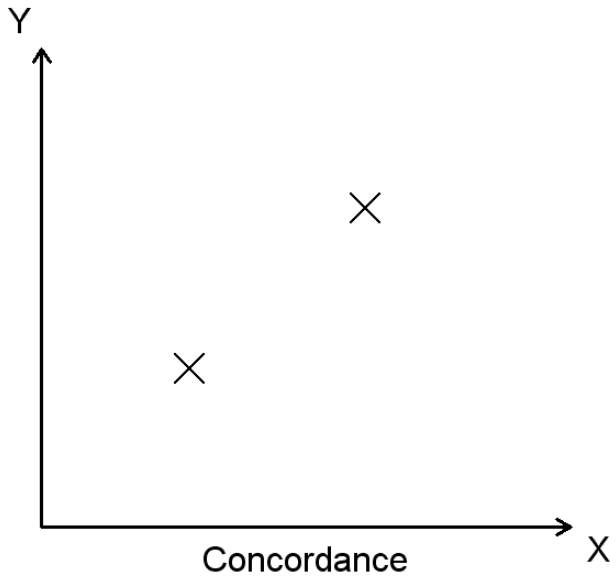
- Testing independence:

$$H_0 : F(x, y) = F_X(x) F_Y(y)$$

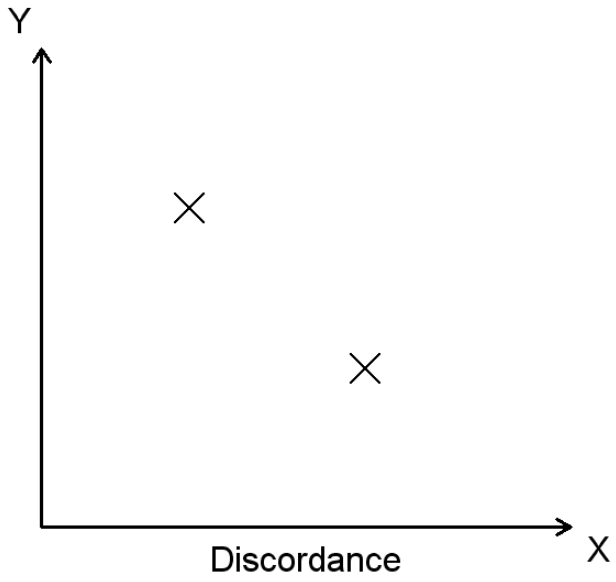
Pairwise relationship: without truncation

- Concordance indicator

$$\Delta_{ij} = I\{(X_i - X_j)(Y_i - Y_j) > 0\}$$



$\Delta_{ij} = 1$



$\Delta_{ij} = 0$

Kendall's tau for typical bivariate data

- Definition

$$\tau = \Pr(\Delta_{ij} = 1) - \Pr(\Delta_{ij} = 0) = 2E[\Delta_{ij}] - 1$$

- Properties

$$\tau = 0 \quad \text{if } X \perp Y$$

$$-1 \le \tau \le 1$$

- Estimator

$$\hat{\tau} = \binom{n}{2}^{-1} \sum_{i<j} (2\Delta_{ij} - 1)$$

$\hat{\tau}$ の値を見て独立性
が検定できる (robust test)

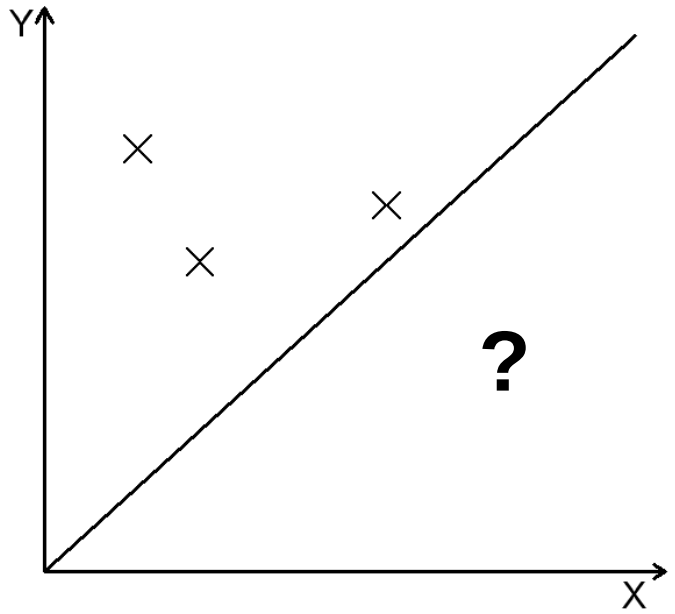
Association measure under Truncation

- Truncation data:

$$\{(X_j, Y_j)\,(j = 1, \ldots, n)\}$$

subject to $X_j \le Y_j$

$\hat{\tau}$ は計算出来るが
意味を成さない



- 実際 $\tau = \Pr(\Delta_{ij} = 1) - \Pr(\Delta_{ij} = 0) = 2E[\Delta_{ij}] - 1$
を計算すると、

$$\tau > 0 \quad \text{even if } X \perp Y$$

Association measure under Truncation

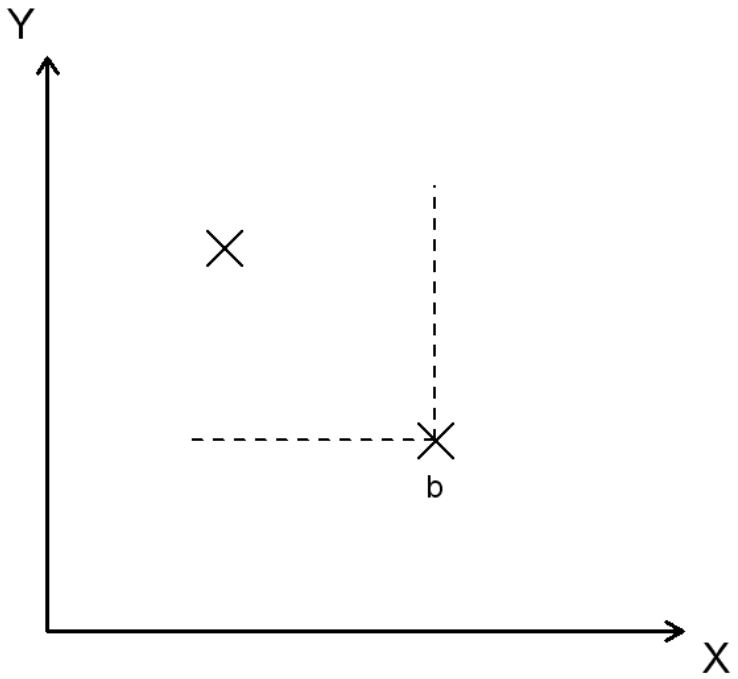
- Kendall's tau $\tau = \Pr(\Delta_{ij} = 1) - \Pr(\Delta_{ij} = 0) = 2E[\Delta_{ij}] - 1$ is not suitable for truncation data

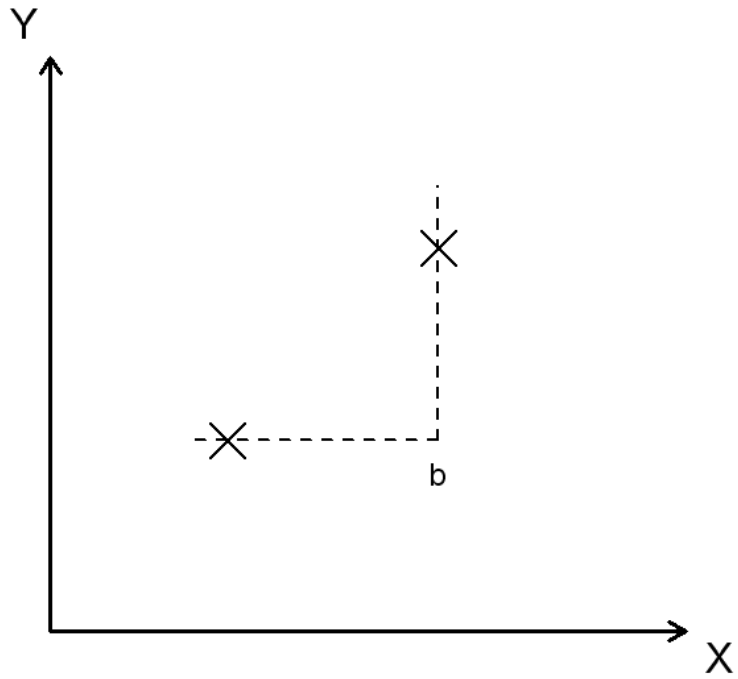
- Tsai (1990, Biometrika) introduced the so called Conditional Kendall's tau, suitable for truncation data (以降のスライドで説明)

- Recently, conditional Kendall's tau is a basic tool for analyzing truncation data
  (Martin & Betensky, 2005; Chaieb et al., 2006; Efron & Petrosian, 1999)

Test quasi-independence (Tsai, 1990)

- $\text{Corner "}b\text{"} = (X_i \vee X_j, Y_i \wedge Y_j)$





Test quasi-independence (Tsai, 1990)

- Comparable condition

$$B_{ij} = \{X_i \vee X_j \le Y_i \wedge Y_j\}$$

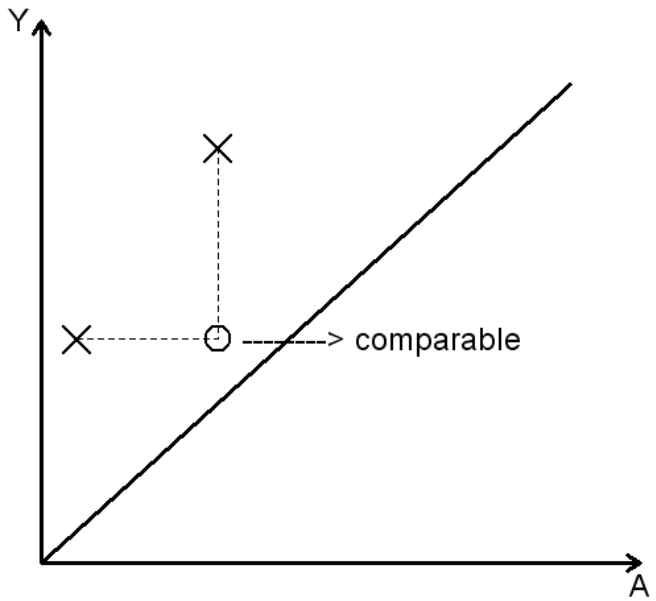
$$\text{"b"} \in \{(x, y) : 0 < x \le y < \infty\} \Leftrightarrow B_{ii} \text{ occurs}$$

- Conditional Kendall's tau

$$\tau_b = 2E(\Delta_{ij} \mid B_{ij}) - 1$$

- Note:

$$\tau_b = 0 \quad \text{if} \quad X \perp Y$$



Test quasi-independence (Tsai, 1990)

- **Definition (Quasi-independence)**

$$H_0 : \Pr(X \le x, Y > y \mid X \le Y) = F_X(x) S_Y(y) / c_0 \qquad (x \le y)$$

F_X, S_Y; distribution/survival functions

$$c_0 = -\iint_{x \le y} dF_X(x) dS_Y(y)$$

- **Remarks:**
  - If H_0 holds, then $\tau_b = 0$
  - If $X \perp Y$, then H_0 holds

As a result,

Reject $\tau_b = 0$ ➔ Reject $X \perp Y$

Test based on Conditional Kendall's tau

- Non-parametric estimator

$$\hat{\tau}_b = \frac{\sum_{i<j} I\{B_{ij}\}(2\Delta_{ij} - 1)}{\sum_{i<j} I\{B_{ij}\}} \rightarrow \tau_b = 2E[\Delta_{ij} \mid B_{ij}] - 1$$

- Under quasi-independence,

$$\hat{\tau}_b / SD(\hat{\tau}_b) \sim N(0,1) \qquad \text{(by CLT for U-statistics)}$$

- Two different way to estimate SD
  - Based on properties of ranks (Tsai)
  - Based on properties of U statistics (Martin and Betensky)

Tsai と Marting & Betensky の違いは分散推定量のみ！

Our method: two-by-two tables

- Consider odds ratio

$$\theta(x,y) = \frac{\Pr(X = x, Y = y \mid X \le Y)\Pr(X \le x, Y > y \mid X \le Y)}{\Pr(X \le x, Y = y \mid X \le Y)\Pr(X = x, Y > y \mid X \le Y)}$$

$$= \frac{dF_X(x)dS_Y(y)\cdot F_X(x)S_Y(y)}{F_X(x)dS_Y(y)\cdot dF_X(x)S_Y(y)} \quad (\text{Under } \mathrm{H}_0)$$

$$= 1$$

- **Comparison:**
  - Tsai's method: $\tau_b = 0$ when H_0
  - Our method: $\theta(x,y) = 1$ when H_0

Our method: two-by-two tables

- Emura and Wang (2010, J. of Mult. Anal.)

- Notation; Counts

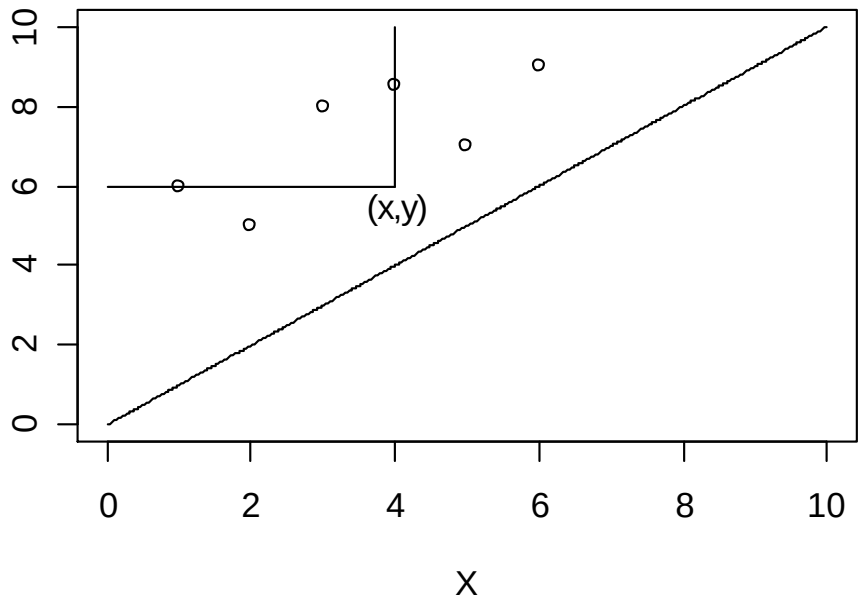
$$\Delta(x,y)=\sum_{j=1}^{n} I(X_j=x,Y_j=y), \qquad N_{1\bullet}(x,dy)=\sum_{j=1}^{n} I(X_j=x,Y_j\geq y)$$

$$N_{\bullet 1}(x,dy)=\sum_{j=1}^{n} I(X_j\leq x,Y_j=y) \qquad R(x,y)=\sum_{j=1}^{n} I(X_j\leq x,Y_j\geq y)$$

Truncation Data

- Two-by-two tables for x < y

	$Y=y$	$Y>y$	
$X=x$	$\Delta(x,y)$		$N_{1\bullet}(dx,y)$
$X<x$			
	$N_{\bullet 1}(x,dy)$		$R(x,y)$



Odds ratio of above table is

$$\theta(x,y)=1$$

Our method: two-by-two tables

- Under quasi-independence

$$E\{\Delta(x,y)\,|\,\text{margins}\} = \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}$$

- Weighted log-rank statistics

$$L_w = \iint_{x\le y} w(x,y)\left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y)N_{\bullet 1}(x,dy)}{R(x,y)}\right]$$

How to choose $w(x,y)$ is important!

Our method: two-by-two tables

- Relationship between two statistics

$$\iint_{x \le y} w(x, y)\left\{\Delta(dx, dy) - \frac{N_{1\bullet}(dx, y)N_{\bullet 1}(x, dy)}{R(x, y)}\right\}$$

$$= -\sum_{i<j} I\{B_{ij}\}\frac{w(\breve{X}_{ij}, \tilde{Y}_{ij})}{R(\breve{X}_{ij}, \tilde{Y}_{ij})}(2\Delta_{ij} - 1)$$

where $B_{ij} = \{X_i \vee X_j \le Y_i \wedge Y_j\} = \{\breve{X}_{ij} \le \tilde{Y}_{ij}\}$

- Choice $w(\breve{X}_{ij}, \tilde{Y}_{ij}) = R(\breve{X}_{ij}, \tilde{Y}_{ij})$ leads to Tsai's (or Martin & Betensky) statistics

Choices of the weight

- weight : $\hat{\pi}(x, y-)^{\rho} = \left\{ \frac{R(x,y)}{n} \right\}^{\rho}$

$$L_{\rho} = \iint_{x \le y} \hat{\pi}(x, y-)^{\rho} \left\{ \Delta(dx, dy) - \frac{N_{1\bullet}(dx, y) N_{\bullet 1}(x, dy)}{R(x, y)} \right\}$$

- Examples

  - Log-rank test: $\rho = 0$

  - Tsai's (or Marting & Betensky) test: $\rho = 1$

- How to choose a good ρ?

  重みは対立仮説(H_1)に依存

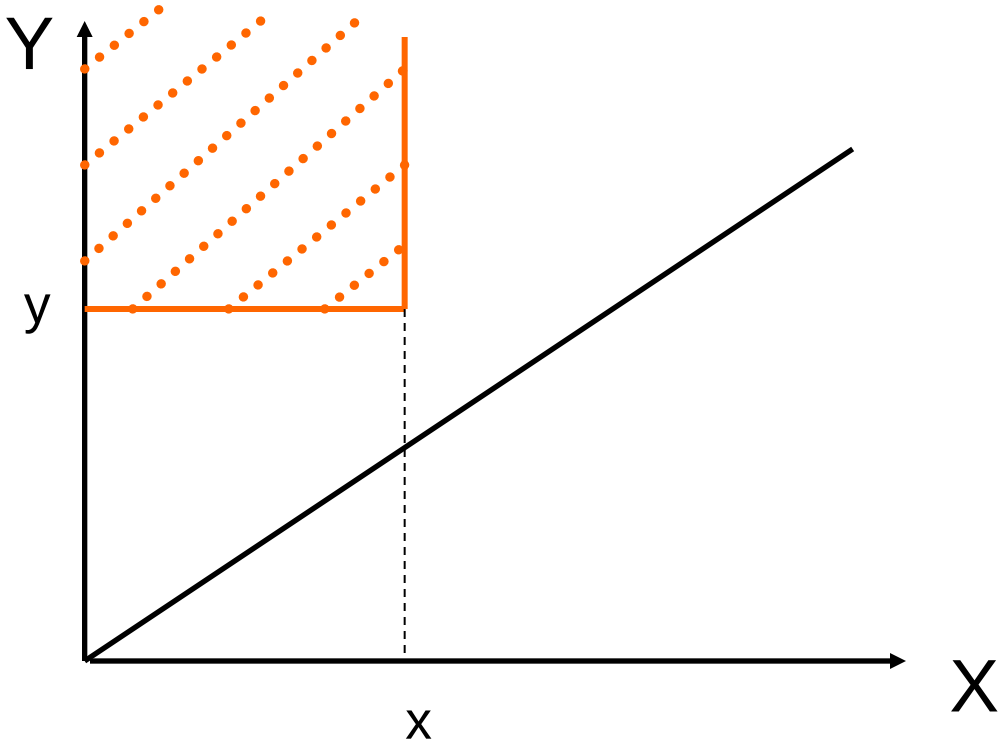
  →対立仮説をCouplaモデルで定式化

Model Assumption for Score Test

- "Semi-survival" copula model (Chaieb et al. 2006)

$$H_1 : \pi(x,y) = \Pr(X \le x, Y > y \mid X \le Y) = \phi_\alpha^{-1}[\phi_\alpha\{F_X(x)\} + \phi_\alpha\{S_Y(y)\}]/c$$

ϕ_α; generator function



Example

1. $\phi_\alpha(t) = (t^{-(\alpha-1)} - 1)/(\alpha - 1)$

Clayton - Copula model

2. $\phi_\alpha(t) = \log\{(1-\alpha)/(1-\alpha^t)\}$

Frank - Copula model

Useful Properties of AC Models

- Odds ratio

$$\theta(x,y)=\frac{\Pr(X=x,Y=y\mid X\le Y)\Pr(X\le x,Y>y\mid X\le Y)}{\Pr(X\le x,Y=y\mid X\le Y)\Pr(X=x,Y>y\mid X\le Y)}=1 \quad (\text{Under } \mathrm{H}_0)$$

- Under semi-survival copula model indexed by $\phi_\alpha(.)$

$$\theta(x,y)=\theta_\alpha(c\pi(x,y)) \quad \text{under } \mathrm{H}_1$$

$$\theta_\alpha(\eta)=-\eta\frac{\phi''_\alpha(\eta)}{\phi'_\alpha(\eta)}$$

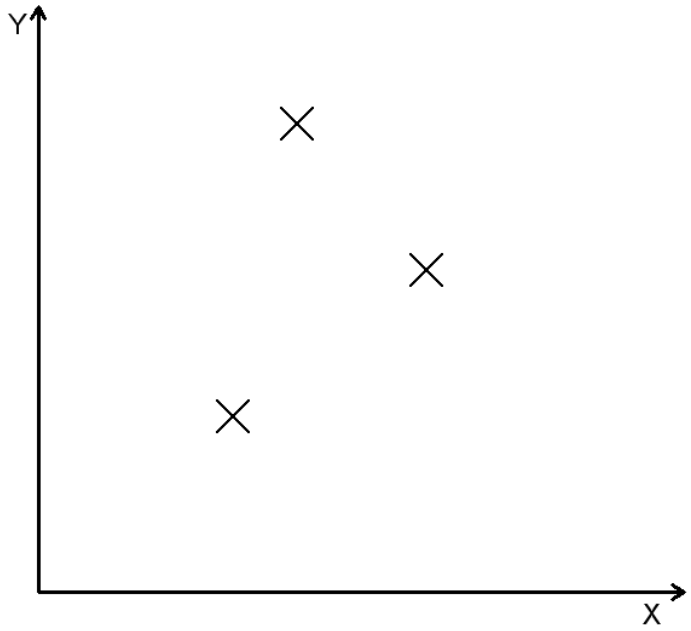
- Quasi-independence $\theta_\alpha(\eta)=1 \quad \text{if} \quad \alpha\to 1$

$$H_0: \quad \alpha=1$$

Likelihood Construction

- Define grid points:

$$\varphi = \left\{ (x,y) \,|\, x \le y, \sum_{j=1}^{n} I(X_j \le x, Y_j = y) = 1, \sum_{j=1}^{n} I(X_j = x, Y_j \ge y) = 1 \right\}$$



Random Variable on a Grid Point

- Failure indicator: $\Delta(x,y)=\sum_{i=1}^{n} I(X_j = x, Y_j = y)$
- Number in the “risk set”:

$$R(x,y)=\sum_{i=1}^{n} I(X_j \le x, Y_j \ge y)$$

- Probability property

$$\Pr\{\Delta(x,y)=1 \mid R(x,y)=r,(x,y)\in\varphi\}=\frac{\theta_\alpha\{c\pi(x,y)\}}{r-1+\theta_\alpha\{c\pi(x,y)\}}$$

Conditional likelihood under AC model

- Likelihood

  (under independence working assumption)

$$L(\alpha, \pi(x,y), c)$$

$$= \prod_{(x,y)\in\varphi} \left[\frac{\theta_\alpha\{c\pi(x,y)\}}{R(x,y)-1+\theta_\alpha\{c\pi(x,y)\}} \right]^{\Delta(x,y)} \left[\frac{r-1}{R(x,y)-1+\theta_\alpha\{c\pi(x,y)\}} \right]^{1-\Delta(x,y)}$$

- Score function: plug in estimate of nuisance parameter

$$\frac{\partial \log L(\alpha, \hat{\pi}(x,y), c)}{\partial \alpha}$$

$$= \iint \frac{\dot{\theta}_\alpha\{c\hat{\pi}(x,y)\}}{\theta_\alpha\{c\hat{\pi}(x,y)\}} \left[\Delta(x,y) - \frac{N_{1\bullet}(dx,y) N_{\bullet 1}(x,dy) \theta_\alpha\{c\hat{\pi}(x,y)\}}{R(x,y)-1+\theta_\alpha\{c\hat{\pi}(x,y)\}} \right]$$

Score Tests based on conditional likelihood

- Assumption: semi-survival AC models
- Score test for $H_0 : \alpha = 1$

  → Log-rank test with a special weight

$$w^*(x, y) = \lim_{\alpha \to 1} \dot{\theta}_\alpha \{\hat{c}\hat{\pi}(x, y-)\}$$

- Suggested weight based on the score test:
  - Clayton model: $w^*(x, y) = 1 \to \rho = 0$
  - Frank model: $w^*(x, y) = \hat{\pi}(x, y-) \to \rho = 1$

Test based on normal approximation

- Large sample property

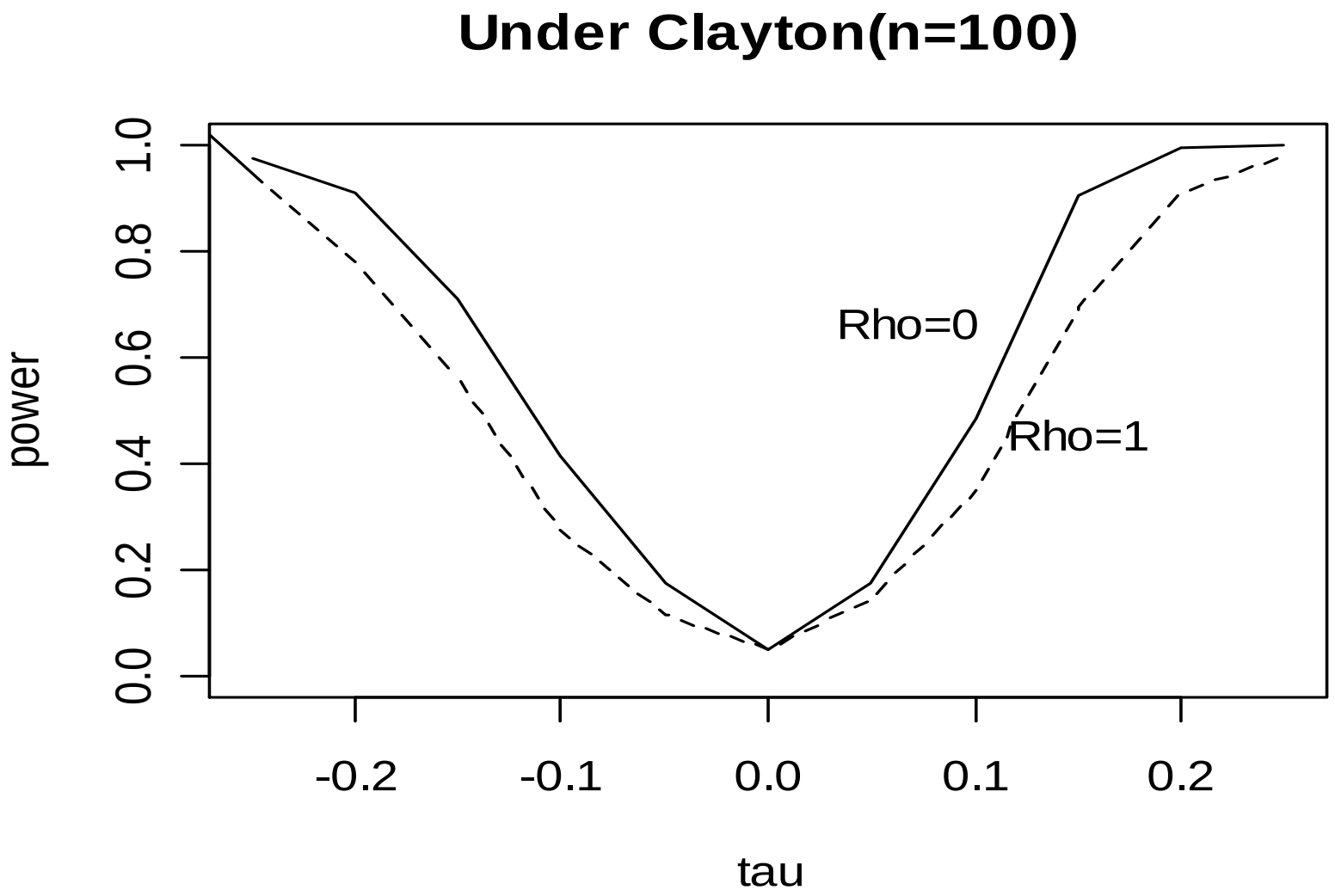
$$L_\rho \xrightarrow{H_0} N(0, \sigma_\rho^2)$$

- Reject H_0 if $|L_\rho / \hat{\sigma}_\rho| > 1.96$

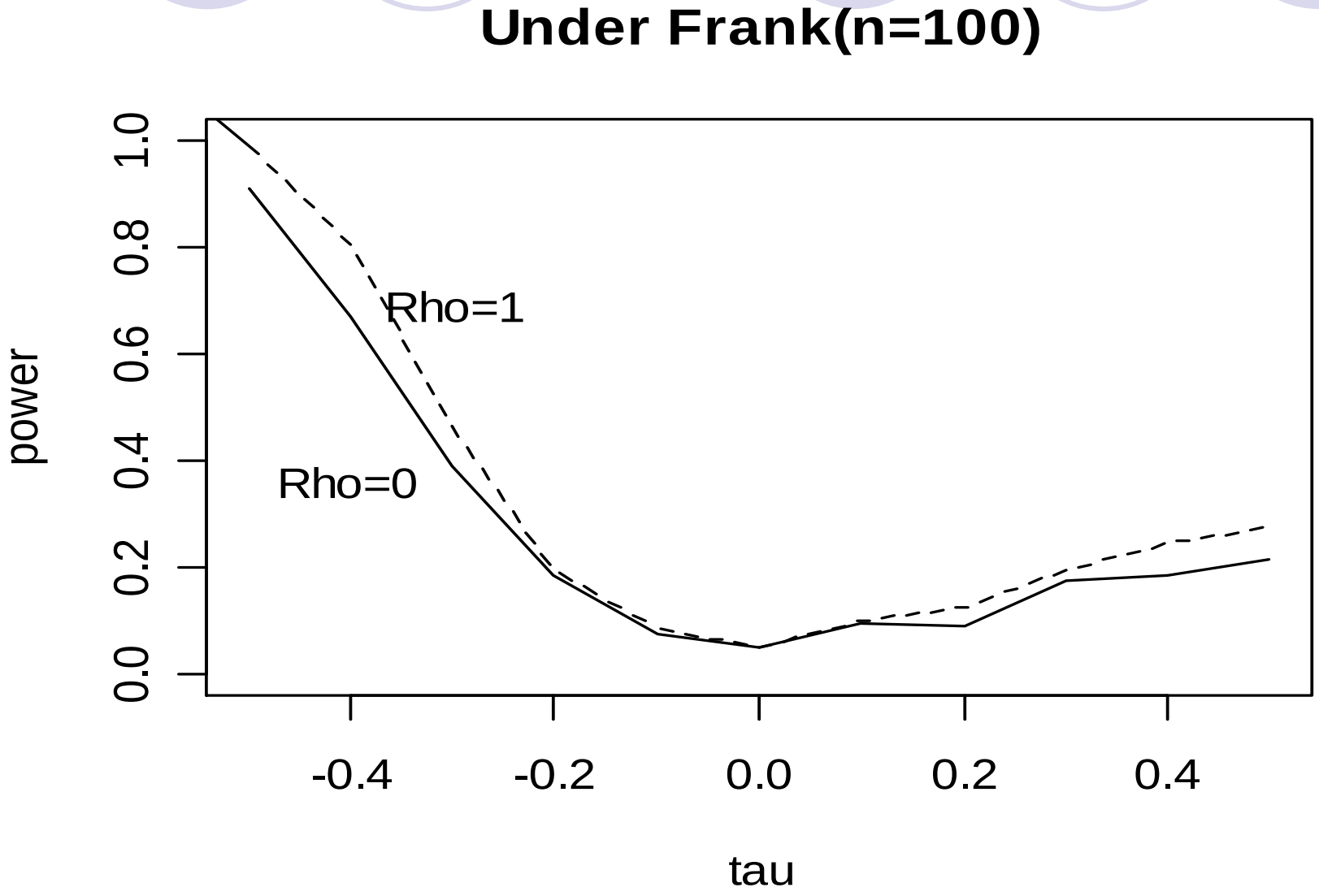
$$\hat{\sigma}_\rho^2 = n/(n-1)\sum_j (L_\rho^{(-j)} - L_\rho^{(\cdot)})^2$$

Consistency of $\hat{\sigma}_\rho^2$ proven by Emura and Wang (2010)

Power comparison for two weights



Power comparison for two weights



Comparison

- Our method（2×2table method）

1. Take advantage of odds ratio
2. Optimal weight, relation to the score test
3. Weight selection via likelihood

- Tsai; Martin & Betensky (pairwise method)

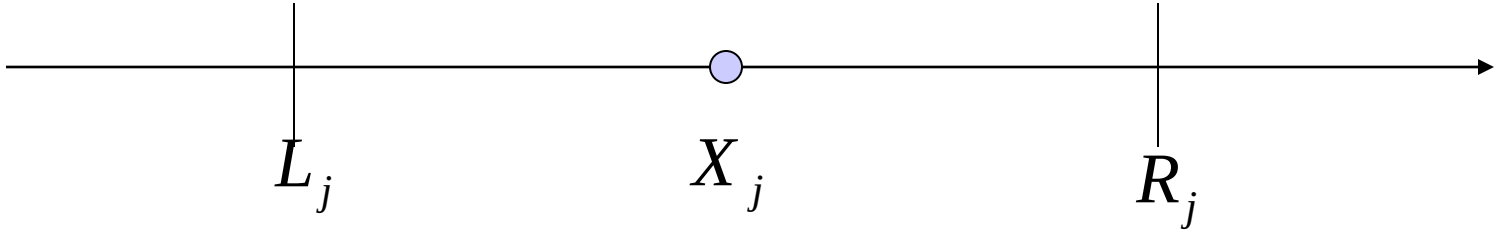
1. Take advantage of conditional Kendall's tau
2. The value of conditional tau is informative
3. U-statistics expression, beautiful asymptotics

Data Analysis: AIDS Example

- Adults (258 subjects): $\hat{\tau}_b = 0.111$
  - $L_{\rho=0}$, $L_{\rho=1}$, Tsai, M&B tests all reject at 5%
  - Clayton model has the best fit among three copula models (maximum likelihood)
    ➔ $L_{\rho=0}$ is recommended
- Children (37 subjects): $\hat{\tau}_b = 0.117$
  - $L_{\rho=0}$ and M&B test reject at 10% level of significance, other three tests does not rejects
  - Clayton model has the best fit among three coupla models (maximum likelihood)
    ➔ $L_{\rho=0}$ is recommended
- Earlier infection time → longer incubation time

Double truncation

- Data: $\{(X_j, L_j, R_j) : j = 1, \ldots, n\}$
  subject to $L_j \le X_j \le R_j$



- Right-truncation: $L_j \equiv 0$

- Quasi-independence

  $H_0 : X \perp_Q (L, R)$

  $: \Pr(L \le l, X \le x, R > r \mid L \le X \le R) = F_X(x)\pi(l, r)/c_0$

Double-truncation

- Concordance

$$\Delta_{XR}(i,j) = \mathbf{I}\{(X_j - X_i)(R_j - R_i) > 0\}$$

$$\Delta_{XL}(i,j) = \mathbf{I}\{(X_j - X_i)(L_j - L_i) > 0\}$$

- Martin & Betensky (2005) define

$$\tau_{XR} = E[2\Delta_{XR}(i,j) - 1 | \Omega_{ij}]$$

$$\tau_{LR} = E[2\Delta_{XL}(i,j) - 1 | \Omega_{ij}]$$

$$\text{where} \quad \Omega_{ij} = \{L_i \vee L_j \le X_i \wedge X_j, X_i \vee X_j \le R_i \wedge R_j\}$$

- And they show

$$\tau_{XR} = \tau_{XL} = 0 \quad \text{under} \quad \mathrm{H}_0 : X \perp_Q (L,R)$$

Double-truncation

- Empirical estimator

$$\hat{\tau}_{XR} = \frac{\sum_{i<j} I\{\Omega_{ij}\}\{2\Delta_{XR}(i,j)-1\}}{\sum_{i<j} I\{\Omega_{ij}\}} \qquad \hat{\tau}_{XL} = \frac{\sum_{i<j} I\{\Omega_{ij}\}\{2\Delta_{XL}(i,j)-1\}}{\sum_{i<j} I\{\Omega_{ij}\}}$$

- CLT for the U-statistics

$$(\hat{\tau}_{XL}, \hat{\tau}_{XL})\hat{\Sigma}^{-1}(\hat{\tau}_{XL}, \hat{\tau}_{XL})' \xrightarrow{H_0} \chi^2_{df=2}$$

注；Powerに関する理論研究がなされていない

My current interest

- Define odds ratio

$$\vartheta(x,l,r)=\frac{\Pr(X=x,L=l,R=r)\Pr(l\le X\le x,L\le l,R\ge r)}{\Pr(X=x,L\le l,R\ge r)\Pr(l\le X\le x,L=l,R=r)}$$

- Measure interaction between X and (L,R)
- Properties

$$\vartheta(x,l,r)=1 \quad \text{if} \quad H_0 \qquad\qquad \vartheta(x,l,r)=\frac{h(x\mid L=l,R=r)}{h(x\mid L\le l,R\ge r)}$$

- How to utilize $\vartheta(x,l,r)$ to test H_0?

Summary (testing quasi-independence)

- Truncation data解析におけるQuasi-indepenenceの必要性を説明した
- Pairwise法 (Tsai; Martin & Betensky) と two-by-two table法 の違いを説明した
- 2つの方法のDouble-truncation dataへの拡張について議論した

ご清聴ありがとうございました

2009年度、数学会年会にて、
一番後ろ;Chen Yi-Hau (Sinica)
中央左;Wang Weijing with son
右;Hsieh, Jin-Jieng（中正大学）
左; Huang, Su-Ying (Sinica)

清華大学、GLM講義後

正面右；K-Y, Liang (Johns Hopkins大)、正面左；Anne Chao （清華大学）

その他；清華大学統計学研究所、博士課程の学生