
Department of Biostatistics

# Association of Bivariate Survival Data.

Svetlana Eden

# Examples of early work in survival analysis.

Ignaz Semmelweis, 1846

Florence Nightingale, 1854

# Survival data:

- Yes or No
- Time to event
- Time to event in the presence of censoring

$Y$ – time to prostate cancer diagnosis

$X$ – paternal history of prostate cancer

$$Y \sim \beta \cdot X$$

$X$ – time to event possibly censored

$Y$ – time to event possibly censored

**How to measure association of $X$ and $Y$ ?**

# I. *Cross ratio*

# II. *Bivariate survival surface*

# III. *Linear rank tests*

*Probability Scale Residuals* approach by *Shepherd, Li, Liu, 2016*

# Modeling *Cross Ratio*

# Definitions for univariate case:

- $T$ is time to event
- $F(t) = Pr(T \leq t)$ is a *cumulative distribution function* (*CDF*)
- $S(t) = Pr(T > t) = 1 - F(t)$ is a *survival function*
- $f(t) = \frac{dF(t)}{dt}$ *probability density function* (*PDF*)
- $\lambda(t) = \lim_{h \to 0} \frac{1}{h} Pr(T \in [t, t+h] | T \geq t) = \frac{f(t)}{S(t)}$ is a *hazard rate*

# Definitions for bivariate case

- $T_1, T_2$ are times to event
- $F(t_1, t_2) = Pr(T_1 \leq t_1, T_2 \leq t_2)$ is a bivariate *CDF*
- $F_1(t_1)$, $F_2(t_2)$ marginal *CDFs*
- $S(t_1, t_2) = Pr(T_1 > t_1, T_1 > t_2)$ is a bivariate *survival function*
- $f(t_1, t_2) = \frac{\partial^2 F(t_1, t_2)}{\partial t_1, \partial t_2}$ is a bivariate *PDF*
- $\lambda(t_1, t_2) = \frac{f(t_1, t_2)}{S(t_1, t_2)}$ is a bivariate *hazard rate*

# Cox Proportional Hazard Model, *1972*

$$\lambda(t|x) = \lambda_0(t) \cdot e^{\beta \cdot x}$$

where $\lambda_0(t)$ is a *baseline hazard* and $x$ is a covariate of interest.

*Hazard ratio*:

$$\theta = \frac{\lambda(t|x=1)}{\lambda(t|x=0)} = \frac{\lambda_0(t) \cdot e^{\beta}}{\lambda_0(t)} = e^{\beta}$$

# Modeling *Cross Ratio*

*Cox, 1972*

$$Hazard\ \ ratio: \quad \theta = \frac{\lambda(t|x=1)}{\lambda(t|x=0)} = e^{\beta}$$

*Clayton, 1978*

$$Cross\ ratio: \quad \theta = \frac{\lambda_s(s_0|t=t_0)}{\lambda_s(s_0|t>t_0)} = \frac{\lambda_t(t_0|s=s_0)}{\lambda_t(t_0|s>s_0)}$$

# Modeling Cross Ratio

*Clayton, 1978:*

$$S(t_1, t_2) = \left[ \left\{ \frac{1}{S_1(t_1)} \right\}^{\theta-1} + \left\{ \frac{1}{S_2(t_2)} \right\}^{\theta-1} - 1 \right]^{-\frac{1}{\theta-1}}$$

$$\theta = 1, \theta < 1, \theta > 1$$

# Modeling Cross Ratio

***Oakes, 1982:***

The author demonstrated that:

$$\frac{\theta - 1}{\theta + 1} = \tau, \text{ where } \tau \text{ is an association measure introduced by Kendal [1938]}$$

# Modeling Cross Ratio

***Kendal, 1938:***

If two independent pairs of variables $(U_1, Z_1)$ and $(U_2, Z_2)$ are from the same bivariate distribution, the pairs are *concordant* if $(U_1 - U_2)(Z_1 - Z_2) > 0$ and *discordant* if $(U_1 - U_2)(Z_1 - Z_2) < 0$.

$$\tau = P[(U_1 - U_2)(Z_1 - Z_2) > 0] - P[(U_1 - U_2)(Z_1 - Z_2) < 0] =$$

$$= E[sign((U_1 - U_2)(Z_1 - Z_2))]$$

# Modeling Cross Ratio

*Oakes, 1982:*

$$\frac{\theta - 1}{\theta + 1} = \tau$$

# Modeling Cross Ratio

*Oakes, 1989:*

$$Pr(T_1 > t_1 | W = w) = \{B_1(t_1)\}^w$$
$$Pr(T_2 > t_2 | W = w) = \{B_2(t_2)\}^w$$

$$S(t_1, t_2) = \int \{B_1(t_1) B_2(t_2)\}^w dF(w)$$

# Modeling Cross Ratio

*Oakes, 1989:*

$$S(t) = p\left[q\left\{S_1(t_1)\right\} + q\left\{S_2(t_2)\right\}\right] \qquad , p'(u) \geq 0,$$

$p''(u) \geq 0$, $p(0) = 1$, and $q(u)$ is an inverse function of $p(v)$

$$\theta(t_1, t_2) = \frac{\lambda_{t_1}(t_1 | T_2 = t_2)}{\lambda_{t_1}(t_1 | T > t_2)} = \frac{\frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} S(t_1, t_2)}{\frac{\partial S(t_1,t_2)}{\partial t_2} \frac{\partial S(t_1,t_2)}{\partial t_1}}$$

$$\theta(v) = -v \cdot q''(v)/q'(v), \quad \text{where } v = S(t_1, t_2)$$

# Modeling Cross Ratio

**Copulas and Bivariate Survival**

***Nelsen, 2007:***

*Copula* is a function from $I \times I$ to $I$, such that for every $u$ and $v$ $C(u, v)$:

$$C(u, 0) = 0 = C(0, v)$$

and

$$C(u, 1) = u \text{ and } C(1, v) = v$$

Also, for every $u_1 \leq \$u_2$ and $v_1 \leq v_2$:

$$C(u_2, v_2) - C(u_2, v_1) - C(u_2, v_1) + C(u_1, v_1) \geq 0$$

# Modeling Cross Ratio

**Copulas and Bivariate Survival**

***Nelsen, 2007:***

*Sklar's theorem*:
X, Y are two random variables with marginal distributions $F(x)$, and $G(y)$, and joint distribution $H(x, y)$. Then
$H(x, y) = C(F(x), G(y))$, where $C(\cdot)$ is a copula.

# Modeling Cross Ratio

## Copulas and Bivariate Survival

*Nelsen, 2007:*

**Examples:**

$$H(x,y) = xy$$

$$H(x,y) = min(x,y)$$

$$C(u,v) = log_\alpha \left[ 1 + \frac{(\alpha^u - 1)(\alpha^v - 1)}{\alpha - 1} \right]$$

# Modeling Cross Ratio

## Copulas and Bivariate Survival

*Nelsen, 2007*

$$\tau = 4 \int \int_{I \times I} C(u, v) dC(u, v) - 1$$

$$\rho_S = 12 \int \int_{I \times I} \left[ C(u, v) - uv \right] du dv$$

# Modeling Cross Ratio

***Fan, Hsu, Prentice, 1998*:**

$$C(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \frac{c(s_1, s_2) S(ds_1, ds_2)}{\int_0^{t_1} \int_0^{t_2} S(du_1, du_2)}, \text{ where } c(s_1, s_2) = \frac{1}{\theta(s_1, s_2)}$$

The following estimator is consistent and asymptotically normally distributed:

$$\hat{C}(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \frac{\hat{S}(s_1^-, s_2^-) \hat{\Lambda}_{10}(ds_1, s_2^-) \hat{\Lambda}_{01}(s_1^-, ds_2)}{1 - \hat{S}(t_1, 0) - \hat{S}(0, t_2) + \hat{S}(t_1, t_2)}$$

# Modeling Cross Ratio

***Fan, Hsu, Prentice, 1998*:**

$$\tau(t_1, t_2) = E\{sign(T_{11} - T_{12})(T_{21} - T_{22}) \mid T_{11} \wedge T_{12} = t_1,\ T_{21} \wedge T_{22} = t_2\}$$

was weighted by $2 \cdot S(t_1^-, t_2^-)S(dt_1, dt_2) + 2 \cdot S(t_1^-, dt_2)S(dt_1, t_2^-)$
gives:

$$\mathcal{T}(t_1, t_2) = E\{sign(T_{11} - T_{12})(T_{21} - T_{22}) \mid T_{11} \wedge T_{12} \leq t_1,\ T_{21} \wedge T_{22} \leq t_2\}$$

# Modeling Survival Surface

# Modeling Survival Surface

*Dabrowska, 1988*

$$S(t_1, t_2) = S(t_1, 0)S(0, t_2) \cdot \frac{S(0,0)S(t_1, t_2)}{S(t_1, 0)S(0, t_2)}$$

$$\theta(t_1, t_2) = \frac{\lambda_{t_1}(t_1 | T_2 = t_2)}{\lambda_{t_1}(t_1 | T > t_2)} = \frac{\frac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2} S(t_1, t_2)}{\frac{\partial S(t_1,t_2)}{\partial t_2} \frac{\partial S(t_1,t_2)}{\partial t_1}}$$

# Modeling Survival Surface

*Dabrowska, 1988*

$$S(t_1, t_2) = S(t_1, 0) S(0, t_2) \cdot \frac{S(0,0) S(t_1, t_2)}{S(t_1, 0) S(0, t_2)}$$

$$S(t_1, t_2) = S(t_1, 0) S(0, t_2) \cdot \prod_{0 < u \leq t_1} \prod_{0 < v \leq t_2} \{1 - L(du, dv)\}$$

$$L(du, dv) = \frac{\Lambda_{10}(du, v^-) \Lambda_{01}(u^-, dv) - \Lambda_{11}(du, dv)}{(1 - \Lambda_{10}(du, v^-))\,(1 - \Lambda_{01}(u^-, \Delta v))}$$

# Modeling Survival Surface

*Dabrowska, 1988*

$$L(du, dv) = \frac{\Lambda_{10}(du, v^-)\Lambda_{01}(u^-, dv) - \Lambda_{11}(du, dv)}{(1 - \Lambda_{10}(du, v^-))\,(1 - \Lambda_{01}(u^-, \Delta v))}$$

$$\Lambda_{11}(du, dv) = \frac{P(u \in du, v \in dv)}{P(u \geq u, v \geq v)} = \frac{S(du, dv)}{S(u^-, v^-)}$$

$$\Lambda_{10}(du, v) = \frac{P(u \in du, v > v)}{P(u \geq du, v > v)} = \frac{-S(du, v)}{S(u^-, v)}$$

$$\Lambda_{01}(u, dv) = \frac{P(u > u, v \in dv)}{P(u > u, v \geq v)} = \frac{-S(u, dv)}{S(u, v^-)}$$

# Modeling Survival Surface

*Dabrowska, 1988*

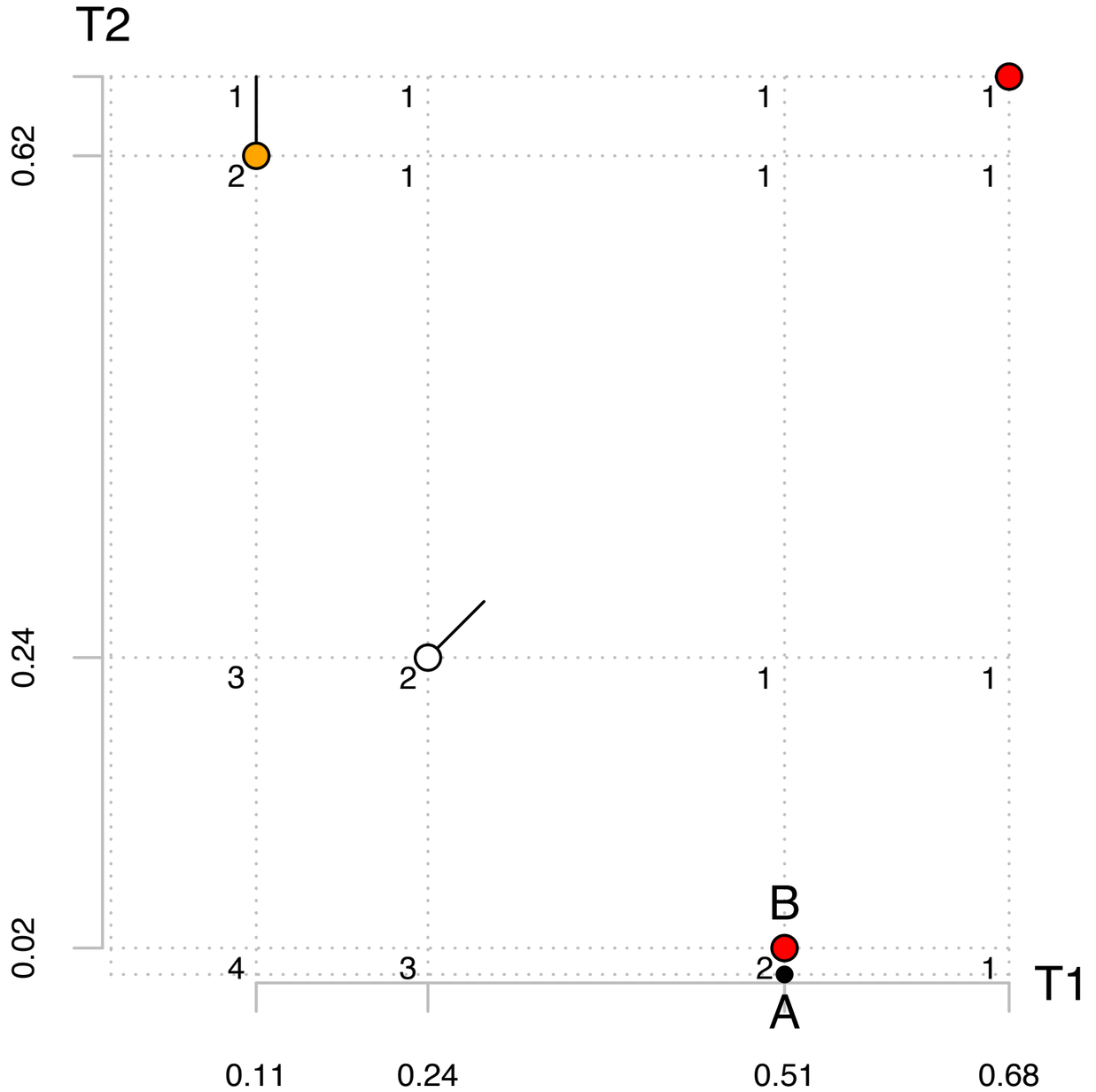
$$\hat{S}(t_1, t_2) = \hat{S}(t_1, 0)\hat{S}(0, t_2) \cdot \prod_{0<u\leq t_1} \prod_{0<v\leq t_2} \{1 - \hat{L}(\Delta u, \Delta v)\}$$

$$\hat{L}(\Delta u, \Delta v) = \frac{\hat{\Lambda}_{10}(\Delta u, v^-)\hat{\Lambda}_{01}(u^-, \Delta v) - \hat{\Lambda}_{11}(\Delta u, \Delta v)}{\Big(1 - \hat{\Lambda}_{10}(\Delta u, v^-)\Big)\Big(1 - \hat{\Lambda}_{01}(u^-, \Delta v)\Big)}$$

$$\hat{\Lambda}_{11}(\Delta u, \Delta v) = \#(T_1 = u, T_2 = v, \delta_1 = \delta_2 = 1)/\#(T_1 \geq u, T_2 \geq v)$$
$$\hat{\Lambda}_{10}(\Delta u, v^-) = \#(T_1 = u, \delta_1 = 1, T_2 \geq v)/\#(T_1 \geq u, T_2 \geq v)$$
$$\hat{\Lambda}_{01}(u^-, \Delta v) = \#(T_2 = v, \delta_2 = 1, T_1 \geq u)/\#(T_1 \geq u, T_2 \geq v)$$

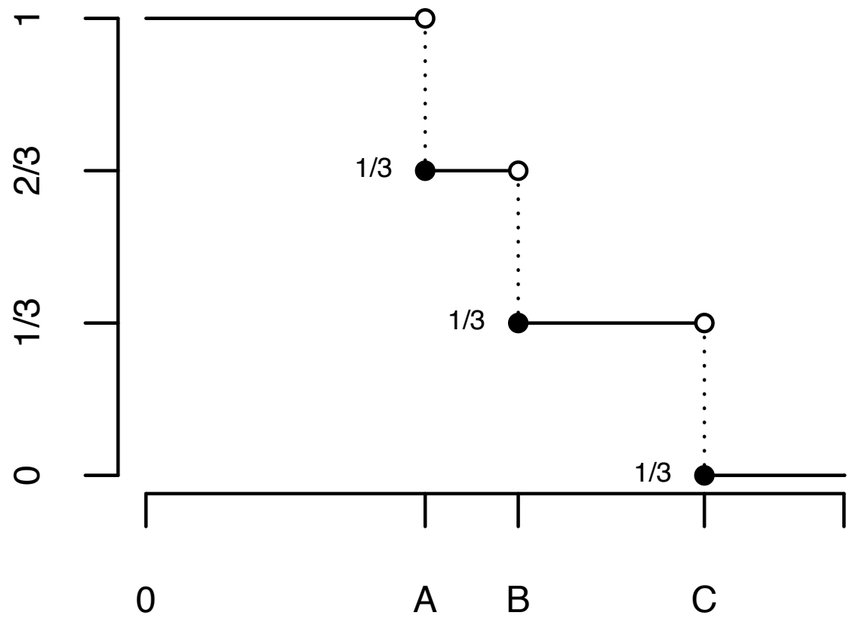
# Modeling Survival Surface

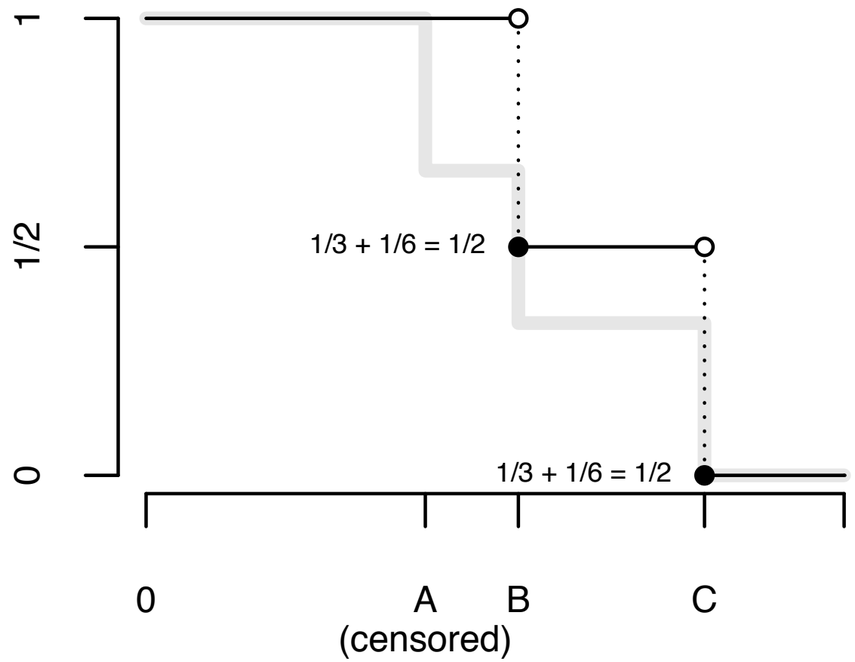
*Example of data with negative mass assignment*



# Modeling Survival Surface

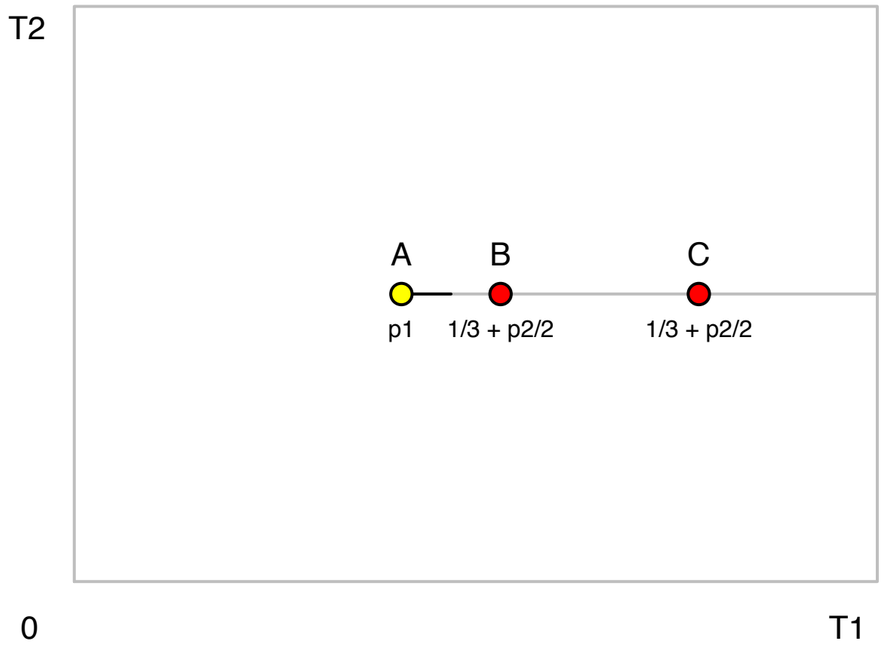
***Mass redistribution, univariate case*:**

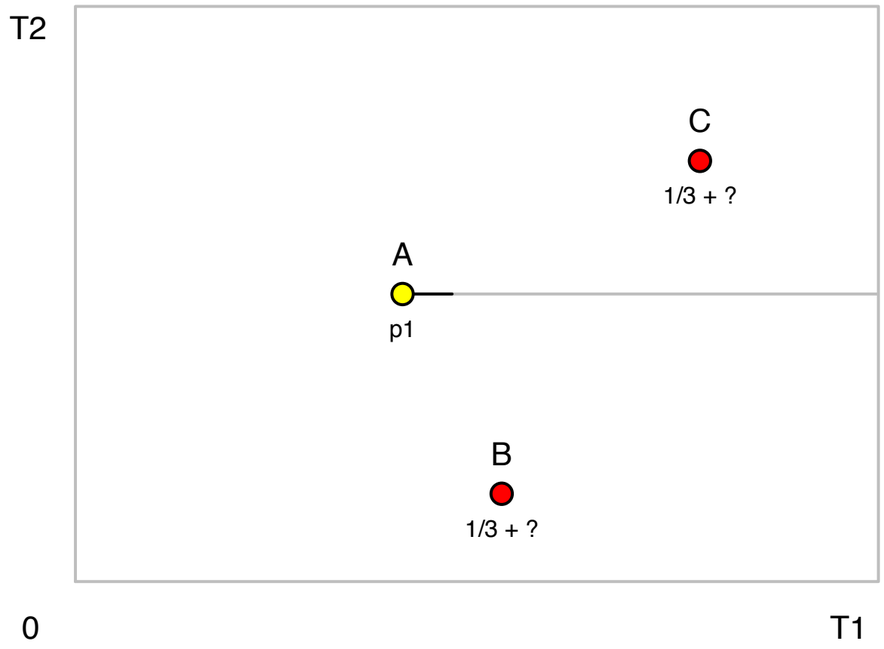


**On the left: Uncensored observations are contained in the half-line**

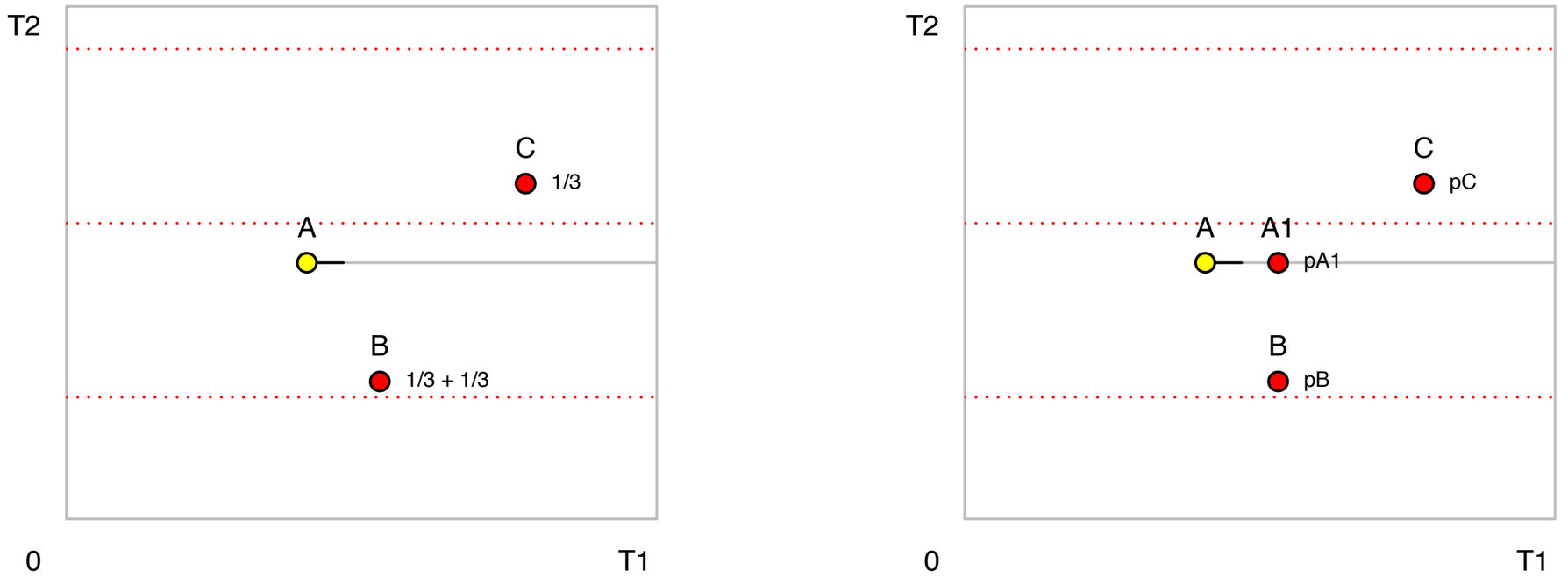
**On the right: Uncensored observations are not contained in the half-line**



**Left panel: Repaired NPMLE of Van der Laan, 1996**

**Right panel: Repaired NPMLE with adjustment Moodie and Prentice, 2005.**



# Linear Rank Tests

## Probability Scale Residuals

# Probability Scale Residuals (PSR)

*Li* and *Shepherd*, 2012

$$r(t, F) = E\{sign(t, T)\} =$$

$$= pr(T < t) - pr(T > t) =$$

$$= F(t^-) - (1 - F(t)) =$$

$$= F(t^-) - 1 + F(t)$$

# Measure of association using PSR

***Liu, Shepherd, Wanga, Li, 2015***

Pointed out that *Spearman* $\rho_S$ is equivalent to correlation of PSR:

$$\rho_S = \rho_{PSR}$$

where

$$\rho_S = cor(F_1(T_1),\ \ F_2(T_2))$$

$$\rho_{PSR} = cor(r(x, F_1, \delta),\ \ r(y, F_2, \epsilon))$$

# PSR for censored data

***Shepherd, Li, Liu, 2016***

$$r(y, F) = F(y) + F(y^-) - 1, \qquad \delta = 1$$

$$r(y, F) = E\{r(T, F) | T > y\} = F(y), \quad \delta = 0$$

or

$$r(y, F, \delta) = F(y) - \delta(1 - F(y^-))$$

where $\delta \in \{0, 1\}$

Does the equality $\rho_S = \rho_{PSR}$ still hold for censored data?

# Censored version of *Spearman* correlation

***Cuzick, 1982***

Suggested to define $\rho_S$ for continuous censored data as:

$$\rho_S = cor(\delta \cdot F_1(x) + (1-\delta)E_{F_1}(F_1(x)|T_1 > x),$$

$$\delta \cdot F_2(y) + (1-\delta)E_{F_2}(F_2(y)|T_2 > y))$$

# Continuous case: $\rho_S$ vs $\rho_{PSR}$

$$F_1(x) = F_1(x^-)$$
$$F_2(y) = F_2(y^-)$$

$$E_{F_1}[F_1(x)|T_1 > x] = \frac{1 + F_1(x)}{2}$$
$$E_{F_2}[F_2(y)|T_2 > y] = \frac{1 + F_2(y)}{2}$$

# Continuous case: $\rho_S$ vs $\rho_{PSR}$

$$\begin{aligned}\rho_S &= cor\left(\delta F_1(x) + (1-\delta)\frac{1+F_1(x)}{2},\ \epsilon F_2(y) + (1-\epsilon)\frac{1+F_2(y)}{2}\right) = \\ &= cor\left((1+\delta)F_1(x) + 1 - \delta,\ (1+\epsilon)F_2(y) + 1 - \epsilon\right)\end{aligned}$$

$$\rho_{PSR} = cor\left((1+\delta)F_1(x) - \delta,\ \ (1+\epsilon)F_2(y) - \epsilon\right)$$

Therefore $\rho_S = \rho_{PSR}$

# Discrete case: $\rho_S$ vs $\rho_{PSR}$

$$\rho_S = cor\left(\frac{F_1(t_1)+F_1(t_1^-)}{2}, \frac{F_2(t_2)+F_T(t_2^-)}{2}\right)$$

$$\rho_S = cor\left(\delta\frac{F_1(t_1)+F_1(t_1^-)}{2} + (1-\delta)E\left[\frac{F_1(t_1)+F_1(t_1^-)}{2}|T_1 > t_1\right],\right.$$
$$\left.\epsilon\frac{F_2(t_2)+F_2(t_2^-)}{2} + (1-\epsilon)E\left[\frac{F_2(t_2)+F_2(t_2^-)}{2}|T_2 > t_2\right]\right)$$

# Discrete case: $\rho_S$ vs $\rho_{PSR}$

$$\frac{E_{F_1}\left[F_1(x)|T_1 > x\right] + E_1\left[F_1(x^-)|T_1 > x\right]}{2} = \frac{1 + F_1(x)}{2}$$

$$\frac{E_{F_2}\left[F_2(y)|T_2 > y\right] + E_2\left[F_2(y^-)|T_2 > y\right]}{2} = \frac{1 + F_2(y)}{2}$$

# Discrete case: $\rho_S$ vs $\rho_{PSR}$

$$\rho_S = cor\left(F_1(x) + \delta F_1(x^-) + 1 - \delta,\ F_2(y) + \epsilon F_2(y^-) + 1 - \epsilon\right)$$

$$\rho_{PSR} = cor\left(F_1(x) + \delta F_1(x^-) - \delta,\ F_2(y) + \epsilon F_2(y^-) - \epsilon\right)$$

Therefore $\rho_S = \rho_{PSR}$

# Linear Rank Tests

***Prentice, 1978, Cuzick, 1982***

$$Y_1 = aZ + e_1 \qquad Y_2 = bZ + e_2$$

$$b = a\lambda, \qquad 0 < |\lambda| < \infty$$

Where $Z$, $e_1$, $e_2$ are independent and $S_k(x) = Pr(e_k > x)$,
$f_k(x) = \dfrac{S_k(x)}{dx}, \quad k = \{1, 2\}.$

He tested: $H_0 : a = 0$ vs $H_1 : a \neq 0$

# Linear Rank Tests

***Prentice, 1978, Dabrowska, 1986***

It is also assumed that $F = F_\theta(x_{1n}, x_{2n})$ and $H_0 : \ F = F_1 F_2$ is equivalent to $H_0 : \ \theta = 0$

$$S_n = \sum_{n=1}^{N} (\delta_{1n} - (1 + \delta_{1n}) F_1) \cdot (\delta_{2n} - (1 + \delta_{2n}) F_2)$$

Which is equivalent to:

$$S_n = \sum_{n=1}^{N} (F_1 - \delta_{1n}(1 - F_1)) \cdot (F_2 - \delta_{2n}(1 - F_2))$$

## *Cuzick, 1982* and *Dabrowska, 1986*

Semiparametric model setting

Assumption about ranking of censored observations

Results were not generalized to discrete data

# Advantages of using PSR correlation as a measure of association

No assumptions regarding shapes of $F(x)$ and $F(y)$

No assumptions about ranking of censored observations

Can be generalized to discrete data

Can be generalized to partial and conditional association measures

# Future research

Estimating partial and conditional censored versions of *Spearman* correlation and evaluating their performance for different proportions of censored data and different distributions

Comparing performance of PSR correlation with martingale-type residuals by **Shih** and **Louis**, 1996.

Deriving if possible a local censored version of *Spearman* correlation and evaluating its performance.

# Acknowledgements

I would like to thank the committee members, **Dr. Harrell**, **Prof. Lasko**, **Prof. Chen**, and **Prof. Shepherd**, for taking your time and participating in this exam.

The class in Survival Analysis by **Prof. Chen** was of great value and I really appreciate her comments and encouragement regarding this project.

I am also very grateful to **Prof. Shepherd** for his class in Probability Theory, for giving me an opportunity to work on this project, and for pushing me to explore material that seemed incomprehensible.

# Things to know:

Survival function and stuff: $S(t) = exp(-\Lambda(t))$

Product limit estimator (KM) and Greenwood var.:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i}\right) \quad \hat{V}[\hat{S}(t)] = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$$

Nelsen-Aalen: $\hat{\Lambda}(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i} \quad \sigma^2_\Lambda(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}$