ERCIM2013, London

Statistical inference based on the NPMLE under double-truncation

Takeshi Emura

Graduate Institute of Statistics,

National Central University, Taiwan

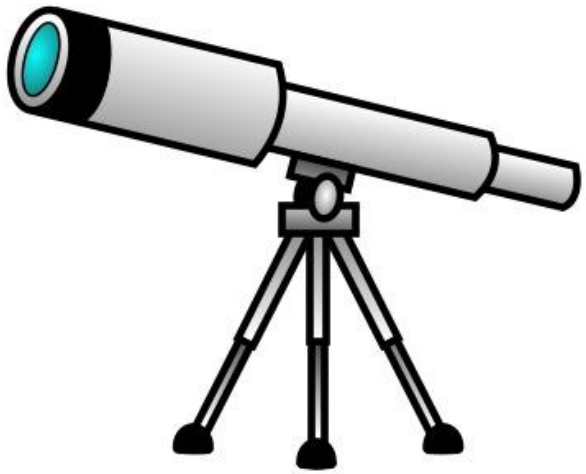
Outlines

- Doubly Truncated data: Review
- The NPMLE: Review
- Derivation of the proposed methods

  (Large sample theory)
- Simulation & Data analysis

- Example (Efron & Petrosian, 1992):

T^* : Quasar's luminosity (brightness)



* Telescope cannot detect quaser if

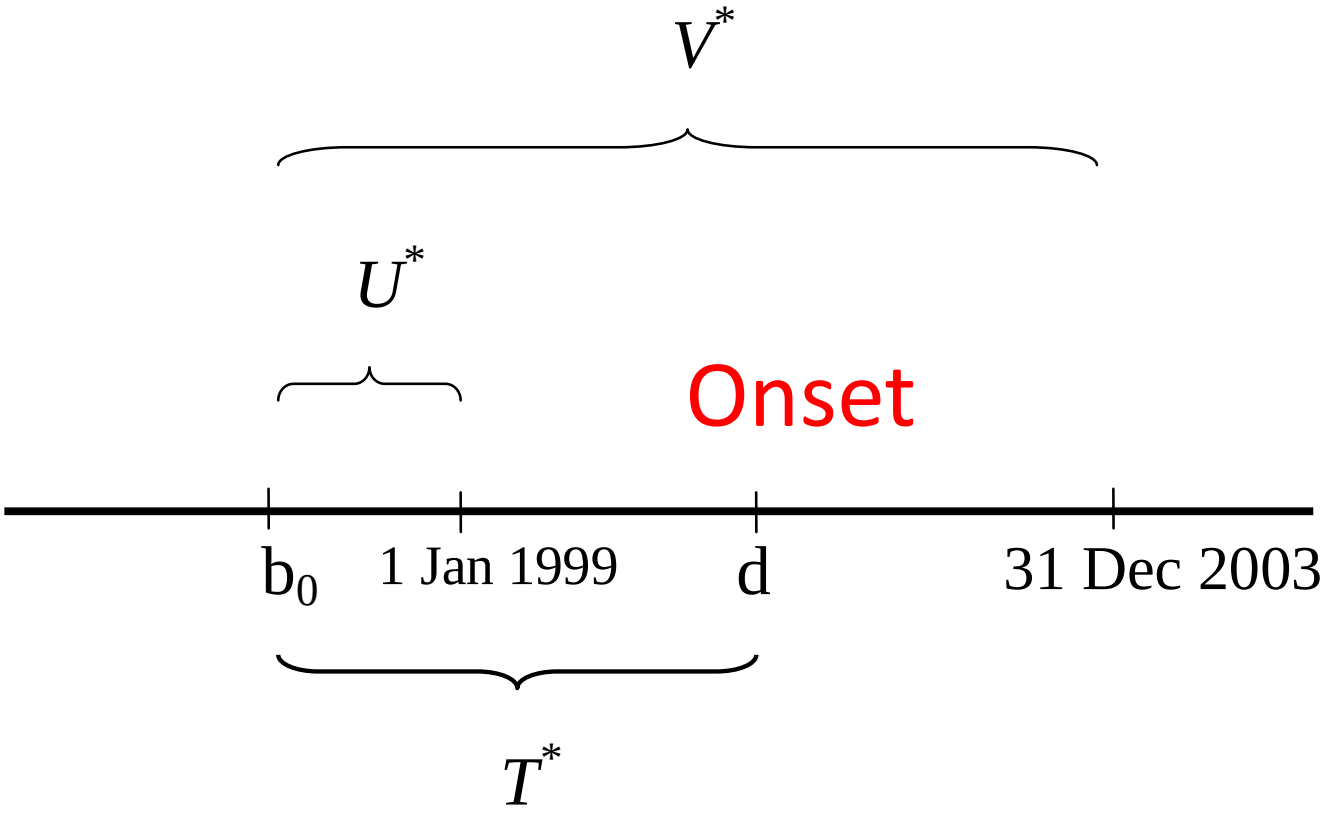
Too dim : $T^* < U^*$

Too bright : $T^* > V^*$

If $U^* \le T^* \le V^* \Rightarrow$ observed;

otherwise, nothing is available (truncated)

Example 2:
Childhood cancer (Moreira & Uña-Álvarez 2010)



- If $U^* \leq T^* \leq V^* \Rightarrow$ observed;

  otherwise, nothing is available (truncated)

- **Population:** Tri-variate random variables

$$(U^*, T^*, V^*) \quad \text{with} \quad (U^*, V^*) \perp T^*$$

$$* \text{If} \quad U^* \le T^* \le V^* \quad \Rightarrow \quad \text{observed;}$$

$$\text{otherwise,} \quad \text{nothing is available !}$$

- **Observation:**

$$\{ (U_j, T_j, V_j) : j = 1, \ldots, n \} \quad \text{subject to} \quad U_j \le T_j \le V_j$$

- **Parameter of interest:** $F(t) = \Pr(T^* \le t)$

as in Efron & Petrosian (1999); Shen (2010)

Moreira & Uña-Álvarez (2010)

Moreira & Keilegom (2013); Stovring & Wang (2007)

Nonparametric MLE (NPMLE)

by Efron & Petrosian (1999)

- T^* has probability masses at $(T_1, \ldots, T_n)$:

$$\mathbf{f} = (f_1, \ldots, f_n)^{\mathrm{T}}, \qquad \sum_{j=1}^{n} f_j = 1$$

- Nonparametric likelihood

$$L_n(\mathbf{f}) = \prod_{j=1}^{n} \Pr(T^* = T_j \mid T^* \in [U_j, V_j]) = \prod_{j=1}^{n} \frac{f_j}{F_j}$$

where $F_j = \sum_{m=1}^{n} f_m J_{jm}$, and $J_{jm} = \mathbf{I}\{U_j \le T_m \le V_j\}$

Nonparametric MLE (NPMLE) by Efron & Petrosian (1999)

- Likelihood equation is solved numerically:

$$\frac{\partial \log L_n(\mathbf{f})}{\partial \mathbf{f}} = \frac{1}{\mathbf{f}} - J^T \frac{1}{\mathbf{F}} = 0 \quad , \quad \mathbf{F} = (F_1, \ldots, F_n)^{\mathrm{T}}$$

(Self-consistency algorithm)

- Target parameter: $F(t) = \Pr(T^* \le t)$

- Estimator: $\hat{F}(t) = \sum_{j=1}^{n} \mathbf{I}(T_j \le t)\, \hat{f}_j \ , \quad \hat{\mathbf{f}} = (\hat{f}_1, \ldots, \hat{f}_n)^{\mathrm{T}}$

- Shen (2010 *AISM*) : For fixed t,

$$\sqrt{n}\{\hat{F}(t)-F(t)\}\rightarrow_d N(0,V(t))$$

where the form of $V(t)$ is not given explicitly.

- Moreira and Uña-Álvarez(2010 *J. of Nonpar.*) suggest Bootstrap methods to construct the confidence interval (C.I.) of $F(t)$
- Shen (2012 *J. of App. Stat.*) invert the empirical likelihood ratio test to derive C.I. of $F(t)$
- Our contribution:

  1) Alternative formula of $V(t)$

  2) Explicit covariance estimator $Cov\{\hat{F}(s),\hat{F}(t)\}$

  3) Various inferences using $Cov\{\hat{F}(s),\hat{F}(t)\}$

Our argument

1. Derive the weak convergence

$$\sqrt{n}(\hat{F} - F) \rightarrow_d G_F$$

2. Estimate the covariance structure by the plug-in method

$$Cov\{\hat{F}(s), \hat{F}(t)\} \approx \hat{E}[G_F(s)G_F(t)]$$

Our asymptotic derivation

Following Murphy (1995 AS),

$$\sqrt{n}(\,\hat{F}(t) - F(t)\,) \xrightarrow{d} G_F(t)$$

where G_F is zero-mean Gaussian with

$$E[\,G_F(s)G_F(t)\,] = \int w_s(x)\sigma_F^{-1}(w_t)(x)dF(x)$$

$$w_s(x) \equiv \mathbf{I}(x \le s)$$

$$\sigma_F(x)[h] =$$

$$E\left[I(U \le x \le V)\left\{ \frac{h(x)}{\int I(U \le s \le V)dF(s)} - \frac{\int I(U \le s \le V)h(s)dF(s)}{\{\int I(U \le s \le V)dF(s)\}^2} \right\} \right]$$

Our asymptotic derivation

- Plug-in estimator

$$\hat{\sigma}_F(x)[h] = \frac{1}{n}\sum_{i=1}^{n} I(U_i \le x \le V_i)\left\{\frac{1}{\hat{F}_i}h(x) - \frac{1}{\hat{F}_i^{\,2}}\sum_{k=1}^{n} J_{ik}h_k\,\hat{f}_k\right\}$$

- Covariance estimator

$$\hat{E}[\,G_F(s)G_F(t)\,] = \int w_s(x)\hat{\sigma}_F^{-1}(w_t)(x)\,d\hat{F}(x)$$

Technical but exact

$$= \sum_{j=1}^{n} w_s(T_j)\hat{\sigma}_F^{-1}(w_t)(T_j)\,\hat{f}_j$$

$$= \mathbf{W}_s^{\mathrm{T}}\left[D\left\{\mathrm{diag}\left(\frac{1}{\hat{\mathbf{f}}^{\,2}}\right) - J^{\mathrm{T}}\mathrm{diag}\left(\frac{1}{\hat{\mathbf{F}}^{\,2}}\right)J\right\}D^{\mathrm{T}}\right]^{-1}\mathbf{W}_t$$

where $\mathbf{W}_t = (\,\mathbf{I}(T_{(1)} \le t) - \mathbf{I}(T_{(n)} \le t), \ldots, \mathbf{I}(T_{(n-1)} \le t) - \mathbf{I}(T_{(n)} \le t)\,)^{\mathrm{T}}$

$$D = [I_{n-1} \vdots -\mathbf{1}_{n-1}]$$

Our asymptotic derivation

- Covariance estimator

$$\hat{C}ov\{\,\hat{F}(s),\hat{F}(t)\,\}$$

$$= \mathbf{W}_s^{\mathrm{T}}\left[D\left\{\mathrm{diag}\left(\frac{1}{\hat{\mathbf{f}}^2}\right) - J^{\mathrm{T}}\mathrm{diag}\left(\frac{1}{\hat{\mathbf{F}}^2}\right)J\right\}D^{\mathrm{T}}\right]^{-1}\mathbf{W}_t$$

- Variance estimator ($s = t$)

$$\hat{V}_{\mathrm{Info}}\{\,\hat{F}(t)\,\}$$

$$= \mathbf{W}_t^{\mathrm{T}}\left[D\left\{\mathrm{diag}\left(\frac{1}{\hat{\mathbf{f}}^2}\right) - J^{\mathrm{T}}\mathrm{diag}\left(\frac{1}{\hat{\mathbf{F}}^2}\right)J\right\}D^{\mathrm{T}}\right]^{-1}\mathbf{W}_t$$

- Related to information matrix via

$$\hat{C}ov\{\,\hat{F}(s),\,\hat{F}(t)\,\} = \mathbf{W}_s^{\mathrm{T}}\left[i_n(\mathbf{f})\right]^{-1}\mathbf{W}_t$$

where

$$i_n(\mathbf{f}) = -\frac{\partial^2 \log L_n(\mathbf{f})}{\partial \mathbf{f}_{(-n)}\partial \mathbf{f}_{(-n)}^{\mathrm{T}}}$$

$$= D\left\{\mathrm{diag}\left(\frac{1}{\mathbf{f}^2}\right) - J^{\mathrm{T}}\mathrm{diag}\left(\frac{1}{\mathbf{F}^2}\right)J\right\}\Bigg|_{f_n = 1 - \mathbf{1}_{n-1}^{\mathrm{T}}\tilde{\mathbf{f}}} D^{\mathrm{T}}$$

$$:\ (n-1)\times(n-1)\ \ \text{matrix}$$

This is not usual Fisher info. matrix since

the dimension grows with *n*

Statistical Inference based on

$$\hat{C}ov\{\hat{F}(s), \hat{F}(t)\} = \mathbf{W}_s^{\mathrm{T}} \left[i_n(\mathbf{f}) \right]^{-1} \mathbf{W}_t$$

1. Confidence interval of $F(t)$
2. Goodness-of-fit test $H_0 : F = F_0$

} Focus

3. Confidence band of $F(\cdot)$

4. Others

Confidence Interval

- Delta method

$$\log \hat{F}(t) - \log F(t) \sim N(\,0, \hat{V}_{\text{Info}}\{\,\hat{F}(t)\,\}/\hat{F}(t)^2\,)$$

- Invert the Wald test:

➔log-transformed interval of F(t):

$$\left(\hat{F}(t)\exp\left[-z_{\alpha/2}\frac{\hat{V}_{\text{Info}}^{1/2}\{\hat{F}(t)\}}{\hat{F}(t)}\right],\ \hat{F}(t)\exp\left[z_{\alpha/2}\frac{\hat{V}_{\text{Info}}^{1/2}\{\hat{F}(t)\}}{\hat{F}(t)}\right] \right)$$

*Log-transformation popular in K-M survival curves

"log-transform" is default in R survfit routine.

(pp.104-108, Klein & Moeschberger, 2003)

Goodness-of-fit test

- Testing $H_0: F = F_0$ vs. $H_1: F \neq F_0$
- Continuous Mapping Theorem

$$\sqrt{n}\sup_t |\hat{F}(t) - F(t)| \xrightarrow{d} \sup_t |G_F(t)|$$

- Kolmogorov-Smirnov statistics

$$K = \sup_t |\hat{F}(t) - F_0(t)| \overset{d}{\approx} \sup_t |\hat{G}_{F_0}(t)/\sqrt{n}|$$

where $E[\hat{G}_{F_0}(s)/\sqrt{n} \times \hat{G}_{F_0}(t)/\sqrt{n}] =$

$$\mathbf{W}_s^{\mathrm{T}}\left[D\left\{\mathrm{diag}\left(\frac{1}{\hat{\mathbf{f}}^2}\right) - J^{\mathrm{T}}\mathrm{diag}\left(\frac{1}{\hat{\mathbf{F}}^2}\right)J\right\}D^{\mathrm{T}}\right]^{-1}\mathbf{W}_t$$

Goodness-of-fit test

- Testing

$$H_0 : F = F_0 \quad \text{vs.} \quad H_1 : F \neq F_0$$

- Cramér-von Mises test based on

$$C = n\int_0^{\infty} \{ \hat{F}(t) - F_0(t) \}^2 dF_n(t)$$

$$= \sum_{j=1}^{n} \{ \hat{F}(T_j) - F_0(T_j) \}^2$$

The null distribution similarly approximated as the K-S statistics

- Simulations

1) Generate (from a design of Moreira & Uña-Álvarez 2010)

$$T^* \sim Uniform(0,1)$$
$$U^* \sim Uniform(0, 0.25), \quad V^* \sim Uniform(0.75, 1)$$

2) Get

$$\{(U_j, T_j, V_j) : j = 1, \ldots, n\} \text{ subject to } U_j \le T_j \le V_j$$

3) Calculate 3 variance estimators

$$\hat{V}_{\text{Info}}\{\hat{F}(t)\}, \quad \hat{V}_{\text{Boot}}\{\hat{F}(t)\} \text{ and } \hat{V}_{\text{Jack}}\{\hat{F}(t)\}$$

4) Repeat 500 times and compare 3 methods

Simple bootstrap algorithm (Moreira and Uña-Álvarez, 2010):

Step 1: For each $b = 1, \ldots, B$, draw bootstrap resamples $\{(U_{jb}^*, T_{jb}^*, V_{jb}^*) : j = 1, \ldots, n\}$ from $\{(U_j, T_j, V_j) : j = 1, \ldots, n\}$, and then compute the NPMLE $\hat{F}_b^*(t)$ from them.

Step 2: Compute the bootstrap variance estimator

$$\hat{V}_{\text{Boot}}\{\hat{F}(t)\} = \frac{1}{B-1}\sum_{b=1}^{B}\{\hat{F}_b^*(t) - \overline{F}^*(t)\}^2 ,$$

where $\overline{F}^*(t) = \frac{1}{B}\sum_{b=1}^{B}\hat{F}_b^*(t)$

Jackknife algorithm

Step 1: For each $i = 1, \ldots, n$, delete the i th sample from $\{(U_j, T_j, V_j) : j = 1, \ldots, n\}$, and then compute the NPMLE $\hat{F}_{(-i)}(t)$ from the remaining n-1 samples.

Step 2: Compute the jackknife variance estimator

$$\hat{V}_{\text{Jack}}\{\hat{F}(t)\} = \frac{n-1}{n}\sum_{i=1}^{n}\{\hat{F}_{(-i)}(t) - \overline{F}_{(\cdot)}(t)\}^2 ,$$

where $\overline{F}_{(\cdot)}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{F}_{(-i)}(t)$

Simulations

Target : $\mathrm{SD}\{\hat{F}(t)\} = \sqrt{\frac{1}{500}\sum_{r=1}^{500}(\hat{F}(t) - \bar{\hat{F}}(t))^2}$

- ESD = $\frac{1}{500}\sum_{r=1}^{500}\sqrt{\hat{V}\{\hat{F}(t)\}_{(r)}}$

(estimated SD)

- MSE = $\frac{1}{500}\sum_{r=1}^{500}(\sqrt{\hat{V}\{\hat{F}(t)\}_{(r)}} - \mathrm{SD}\{\hat{F}(t)\})^2$

Table 2 Simulation results under $U^* \sim \text{Unif}(0, a)$, $T^* \sim \text{Unif}(0,1)$, and $V^* \sim \text{Unif}(b,1)$ with $a = 0.25$ and $b = 0.75$ based on 500 replications.

		$n = 100$	$n = 150$	$n = 200$	$n = 250$	$n = 300$
$F(t) = 0.5$						
SD		0.083	0.064	0.053	0.050	0.046
ESD	Proposed	0.070	0.057	0.050	0.045	0.042
	Bootstrap	0.070	0.059	0.051	0.046	0.043
	Jackknife	0.075	0.061	0.053	0.047	0.044
MSE	Proposed	0.00219	0.00086	0.00033	0.00028	0.00026
	Bootstrap	0.00104	0.00070	0.00048	0.00038	0.00035
	Jackknife	0.00296	0.00185	0.00094	0.00075	0.00073

Boot is best for n<=150

Proposed is best for n>=200

Table 2 Simulation results under $U^* \sim \text{Unif}(0, a)$, $T^* \sim \text{Unif}(0,1)$, and $V^* \sim \text{Unif}(b,1)$ with $a = 0.25$ and $b = 0.75$ based on 500 replications.

		$n = 100$	$n = 150$	$n = 200$	$n = 250$	$n = 300$
$F(t)$ =0.2						
SD		0.090	0.065	0.057	0.052	0.0462
ESD	Proposed	0.069	0.056	0.049	0.045	0.042
	Bootstrap	0.069	0.058	0.051	0.046	0.043
	Jackknife	0.074	0.061	0.053	0.047	0.044
MSE	Proposed	0.00394	0.00091	0.00073	0.00052	0.00026
	Bootstrap	0.00213	0.00113	0.00094	0.00067	0.00035
	Jackknife	0.00522	0.00248	0.00189	0.00115	0.00073

Boot is best for n<=100

Proposed is best for n>=150

Table 2 Simulation results under $U^* \sim \text{Unif}(0, a)$, $T^* \sim \text{Unif}(0, 1)$, and $V^* \sim \text{Unif}(b, 1)$ with $a = 0.25$ and $b = 0.75$ based on 500 replications.

		$n = 100$	$n = 150$	$n = 200$	$n = 250$	$n = 300$
$F(t) = 0.5$						
95%Cov	Proposed	0.930	0.942	0.950	0.946	0.938
	Bootstrap	0.920	0.938	0.950	0.942	0.948
	Jackknife	0.930	0.950	0.948	0.946	0.940
$F(t) = 0.2$						
95%Cov	Proposed	0.938	0.942	0.946	0.932	0.938
	Bootstrap	0.898	0.910	0.928	0.908	0.948
	Jackknife	0.940	0.948	0.952	0.938	0.940

Bootstrap percentile perform poorly for <=250

Consistent with the result of Moreira and Uña-Álvarez(2010)

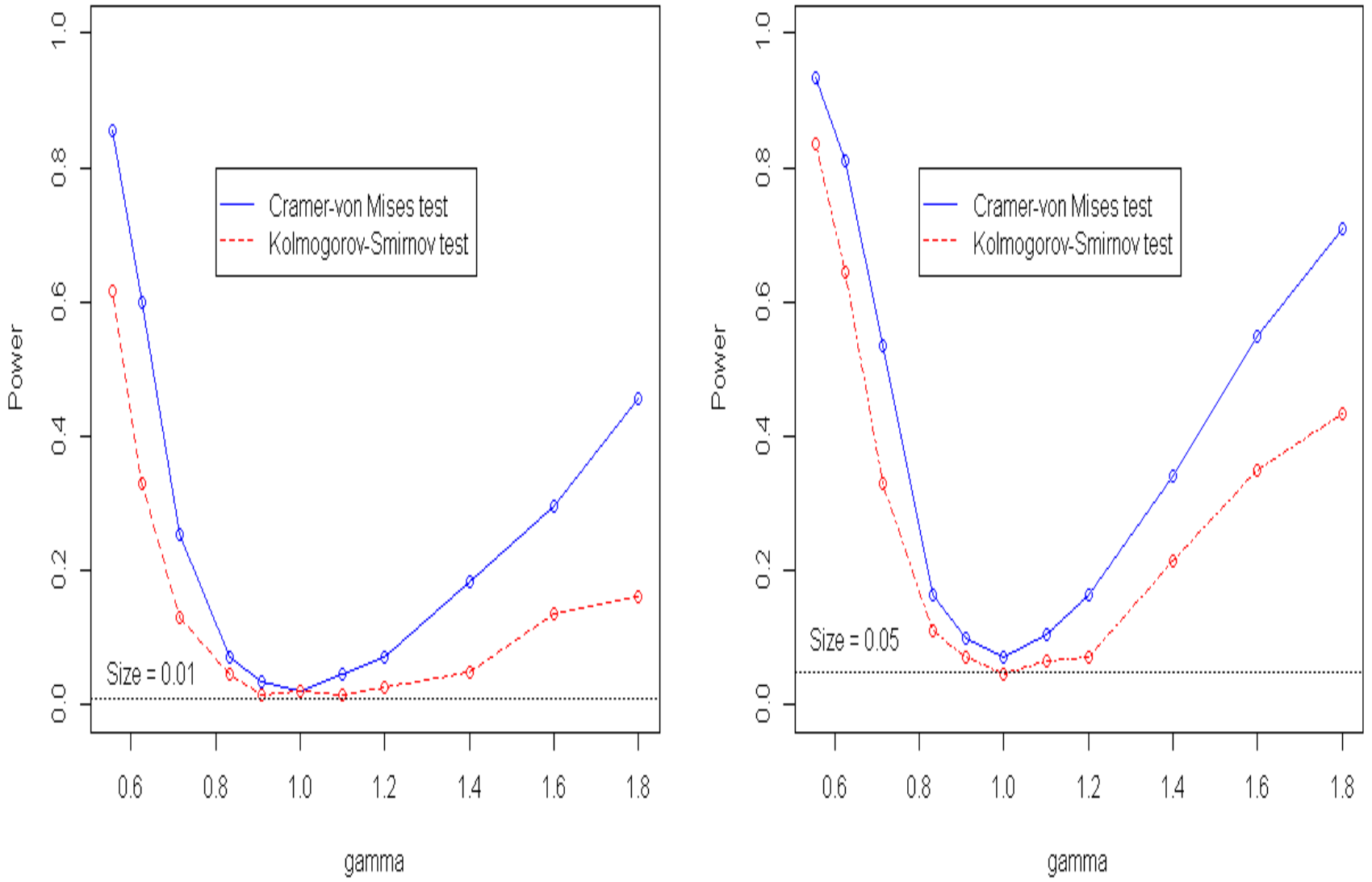


Fig. 2 The power curves for the proposed goodness-of-fit tests with sizes $\alpha = 0.01$ (left panel) and $\alpha = 0.05$ (right panel) based on $n = 150$.

- Data Analysis

  n = 409 childhood cancer cases:

  $$\{ (U_j, T_j, V_j) : j = 1, \ldots, 409 \}$$

  from Moreira and Uña-Álvarez (2010)

- Objective:

  Inference on the distribution of cancer onset

  $$F(t) = \Pr(T^* \leq t)$$

  $$T^* : \text{Time to onset of cancer (from birth)}$$

- Goodness-of-fit test

- $H_{01}: \; F(t) = \dfrac{t}{5475}\mathbf{I}(0 < t < 5475) + \mathbf{I}(t \geq 5475)$

: Cancer occurs *uniformly* below age 15 years

- $H_{02}: \; F(t) = \left(\dfrac{t}{5475}\right)^{3/4}\mathbf{I}(0 < t < 5475) + \mathbf{I}(t \geq 5475)$

: Cancer occurs more frequently on early ages

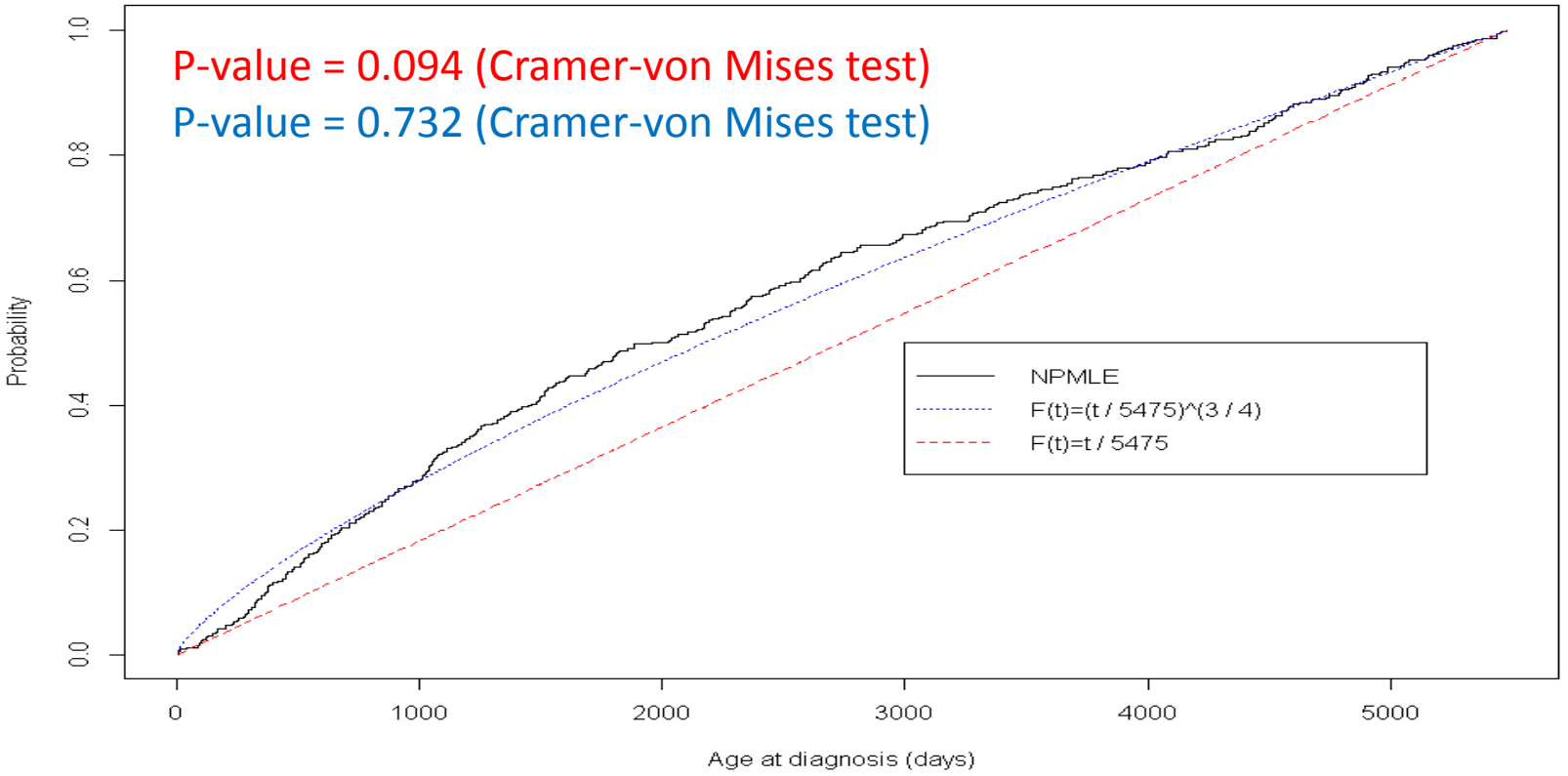


Fig.2: The NPMLE of the distribution of the age at diagnosis for the childhood cancer and the two hypothesized curves for $H_{01}: F(t) = (t/5475)\,\mathbf{I}(0 < t < 5475) + \mathbf{I}(t \geq 5475)$ and $H_{02}: F(t) = (t/5475)^{3/4}\mathbf{I}(0 < t < 5475) + \mathbf{I}(t \geq 5475)$.

Summary

- We derive a closed-form variance-covariance estimator of the NPMLE
- Proposed estimator has competitive performance with the bootstrap and jackknife.
- Proposed estimator is applied to develop a goodness-of-fit test
- Further statistical inference will be possible using the proposed estimator