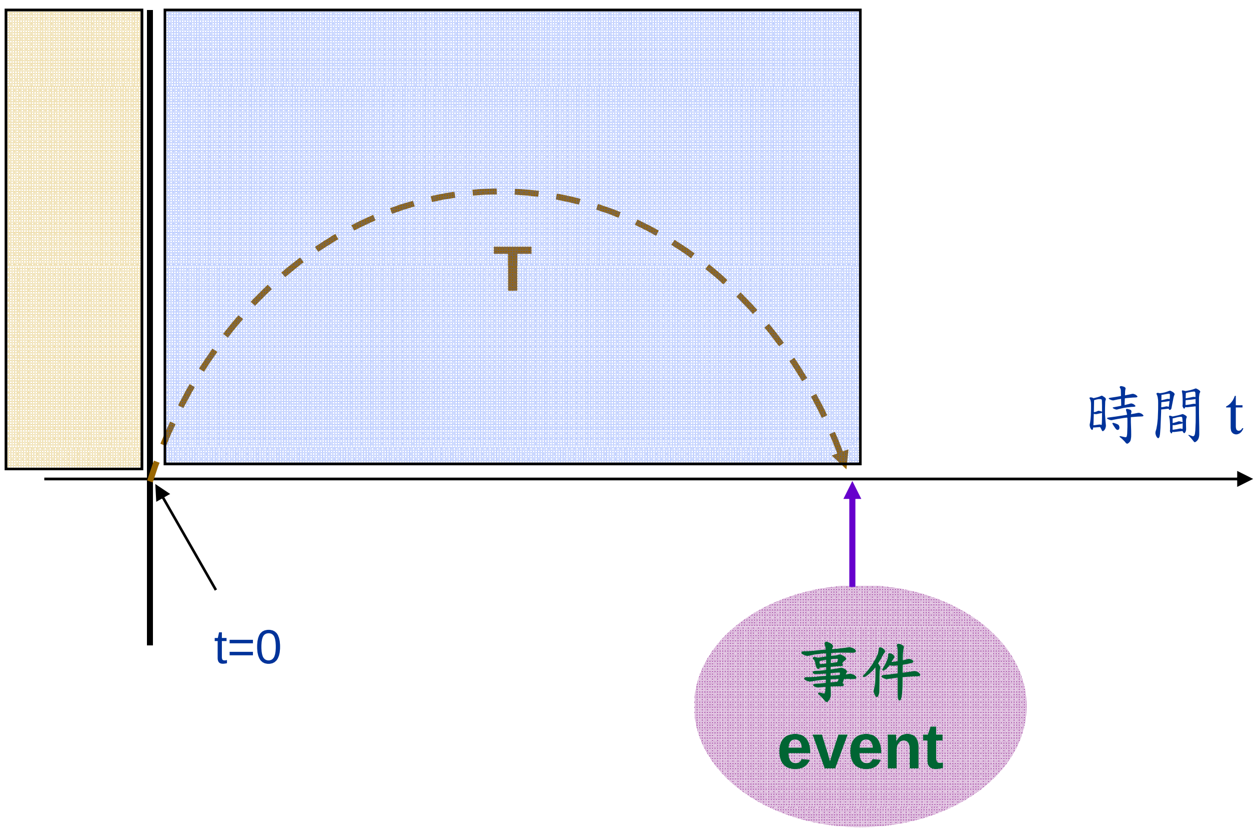
A Review on Event-History Data Analysis

Event history（事件史）



Examples of ‘event’

- 車禍 (保險費率) accident
- 離婚 divorce
- 破產 bankrupt
- 燈泡、門鎖、主機板…壞掉 (元件壽命) failure
- 病人(癌症、器官移植)死亡 death
- 疾病復發 relapse
- (住院)出院 discharge
- ……

several core functions

- hazard rate (mortality) [λ(t)]

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{Pr(t < T < t + \Delta t \mid t < T)}{\Delta t}$$

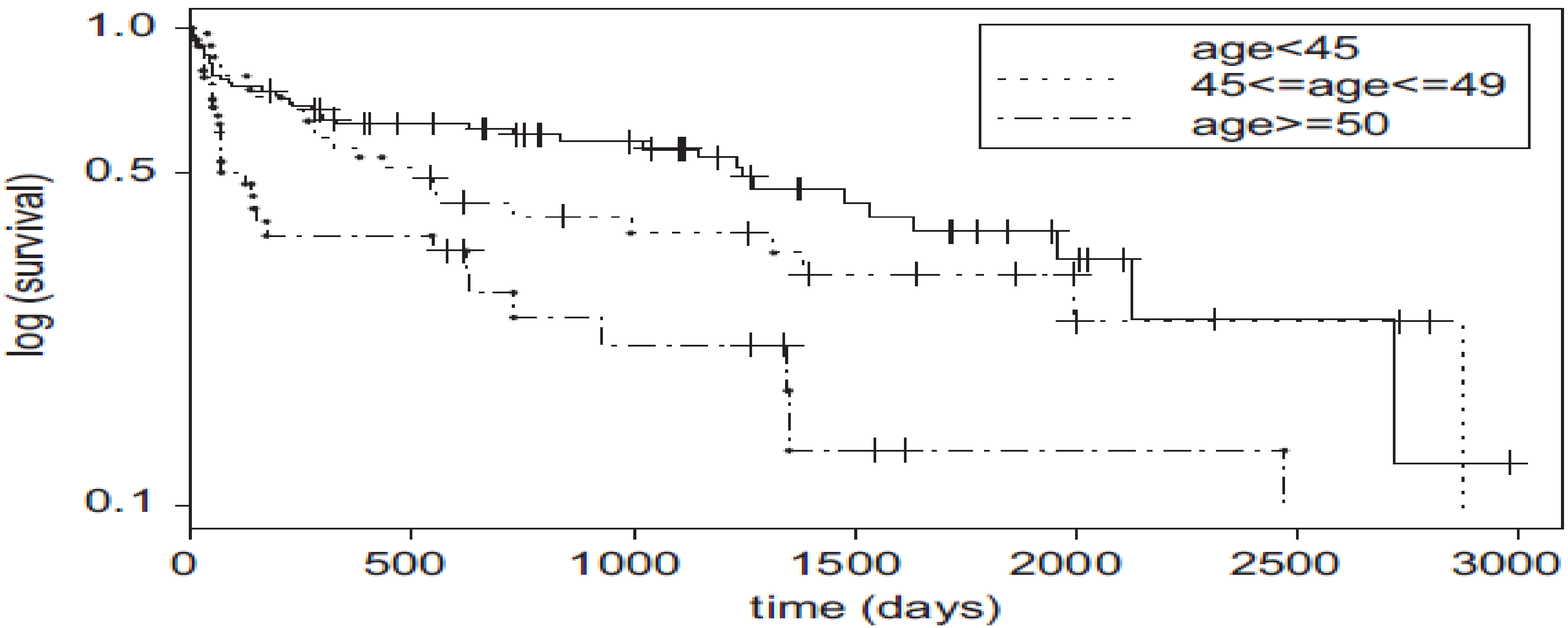
$$\lambda(t) = f(t)/S(t)$$

- cumulative hazard $\Lambda(t) = \int_0^t \lambda(u)du$
- survival function **P(T>t)=1-F(T≦t)**

$$f(t) = \lambda(t)\exp\{-\Lambda(t)\}$$

Kaplan-Meier survival estimate: nonparametric

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t}\left[1 - \frac{d_i}{Y_i}\right], & \text{if } t_1 \leq t \end{cases}.$$



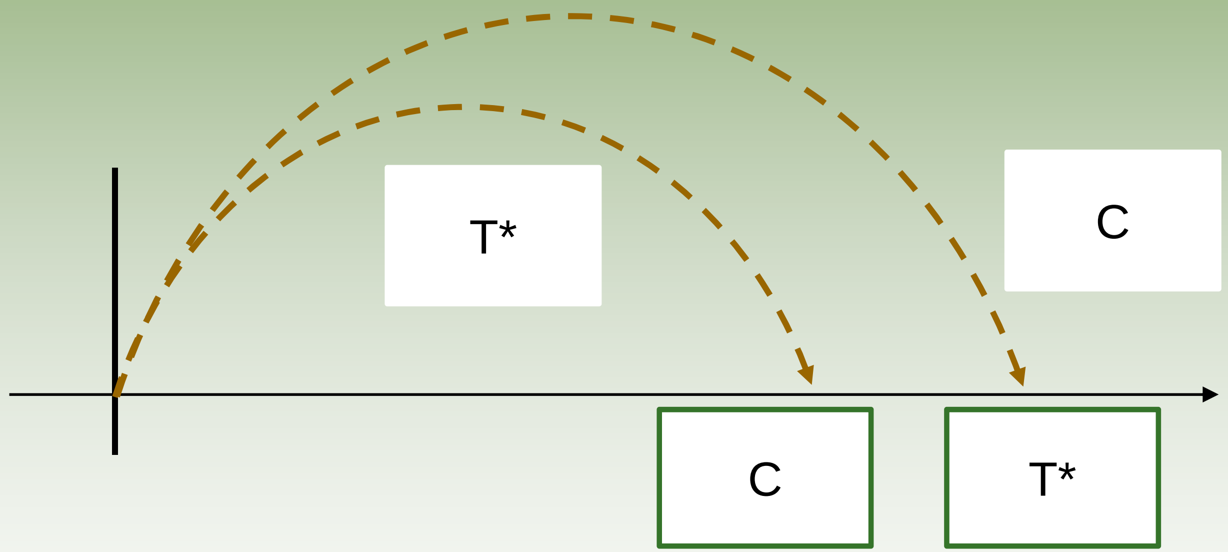
Parametric approach

- Weibull
- log-normal
- Pareto
- log-logistic
- Gompertz
- ….and many others

MLE

Types of censoring (設限型態)

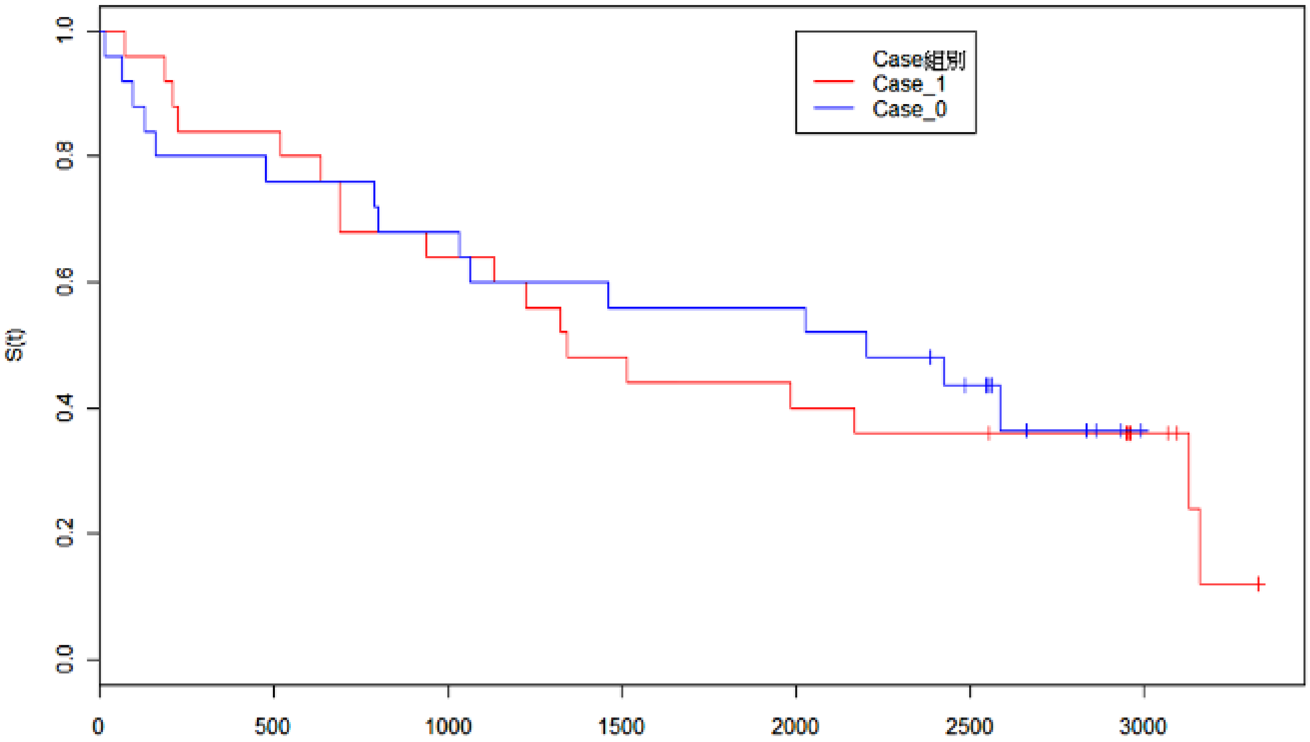
- right censoring, random censoring

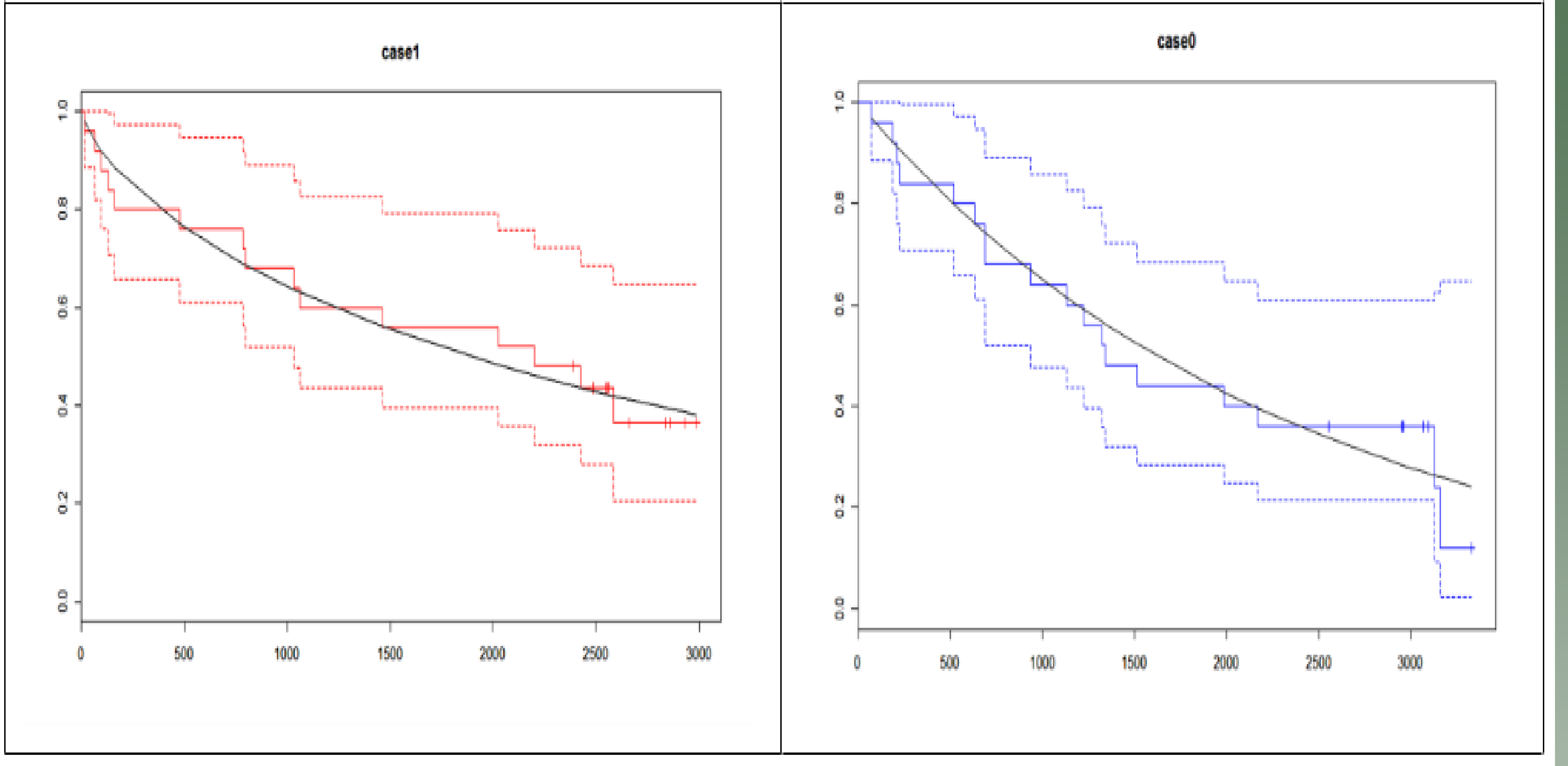


Data: T=min(T*,C), 1(T*<=C)= Δ

- left censoring
- interval censoring
- type I censoring ; type II censoring
- left truncation
- HBeAg HCC patients: data

Examples of parametric estimate: Weibull





$f(t) = abt^{b-1}$		
Case1	a= 0.003230118	b= 0.7119216
Case0	a= 0.0004526262	b= 0.992764

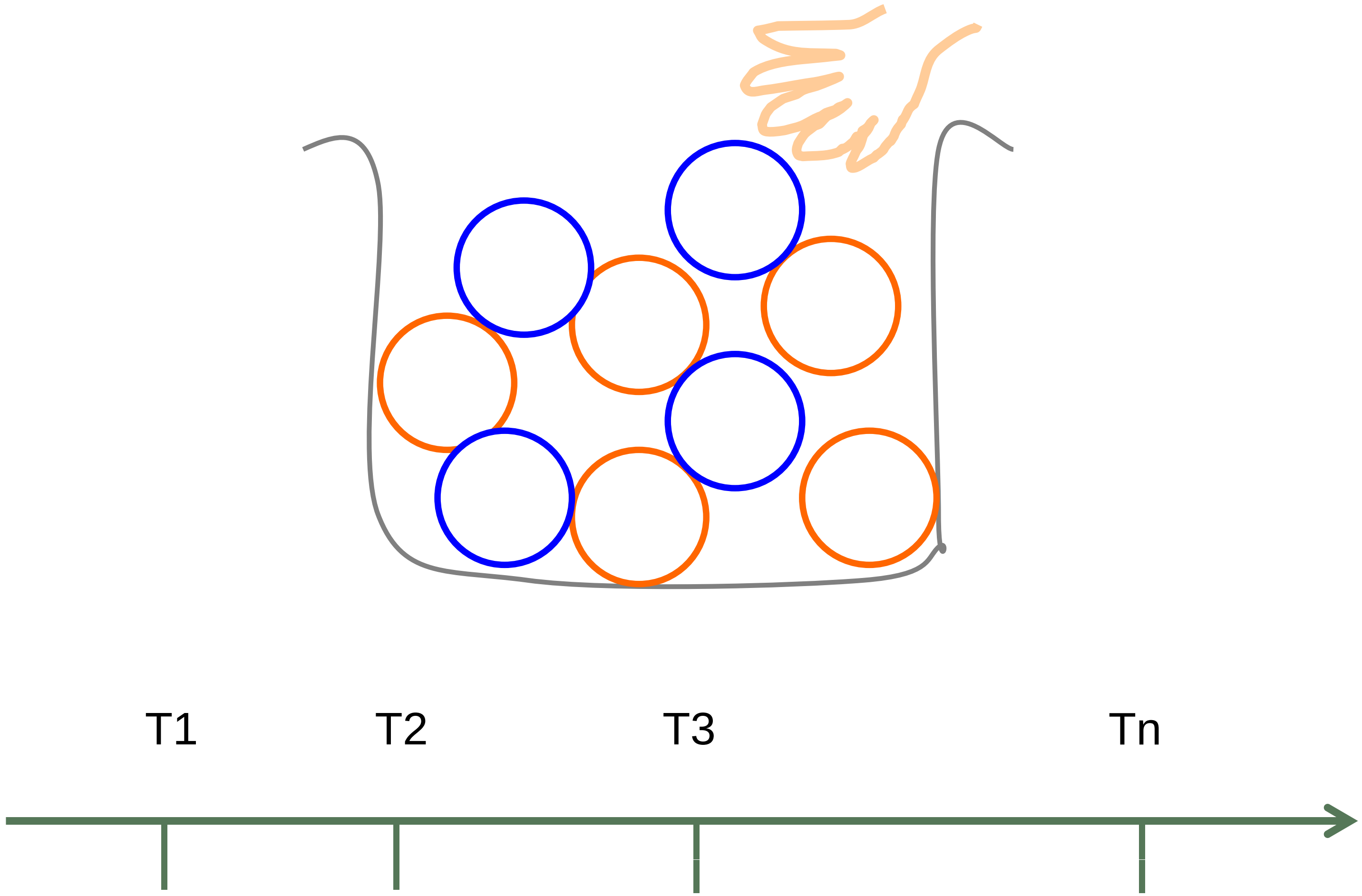
Likelihood subject to right censoring

$$Lik = \Pi \ \{f(T_i)S_c(T_i)\}^{\Delta i} \ \{S(T_i)g_c(T_i)\}^{1-\Delta i}$$

MLE

Contain information of the parameters of the survival time distribution

Two-sample comparison (non-parametric): **urn model**



Forming a sequence of 2 by 2 tables

T_1 -	failed	surv	
Trt (z=1)	a	n_1-a	n_1
Ctrl (z=0)	1-a	n_0-1+a	n_0
	1	N-1	N

- If the event occurred is ‘blue’
  the coding $\mathbf{D_i}$=1, otherwise $\mathbf{D_i}$=0;
- there are $\mathbf{N_i}$ blue and $\mathbf{M_i}$ orange at time $\mathbf{t_i}$ **-** ;
- there are one event (blue or orange) at $\mathbf{t_i}$;
- $T=\Sigma_i(\mathbf{D_i}-E[\mathbf{D_i}])$; $E(\mathbf{D_i})= \mathbf{N_i}/(\mathbf{N_i}+\mathbf{M_i})$
  $=\Sigma_i$ (observed-expected)
- **Log-rank statistic=T/√[VarT]**

Contributions from the first and subsequent observations

- $E(a_1)=1*n_1/N$;
  $Var(a_1)=1*(N-1)n_1n_0/N^2(N-1)$
- LR=

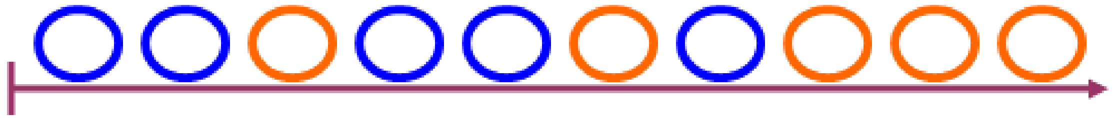
$$\left\{\frac{\sum[a_i - E(a_i)]}{[\sum Var(a_i)]^{1/2}}\right\}^2$$

Under H_0, the LR statistic ~ χ_1^2

Log-rank test statistic

which alignment is possible?

(**The most impossible** (under H_0) ; $p<0.0001$, say)



(**Very impossible** (under H_0) ; $p<0.01$, say)

(**Impossible** (under H_0) to some extent; $p=0.096$, say)

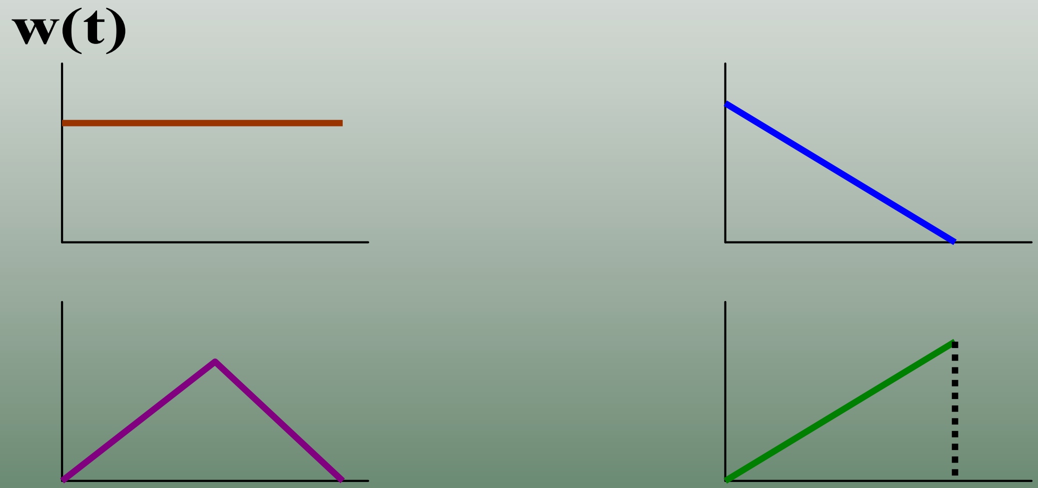
(**Possible** or **do not reject H_0** ; $p=0.82$, say)

- data history
- filtration: increasing sigma algebra
- martingale structure and CLT

Weighted log rank statistic

$$\left\{\frac{\sum w(t_i)[a_i - E(a_i)]}{[\sum w^2(t_i)Var(a_i)]^{1/2}}\right\}^2$$

$$W(t_i)=\{\hat{S}(t_i-)\}^{\rho}\{1-\hat{S}(t_i-)\}^{\gamma}$$



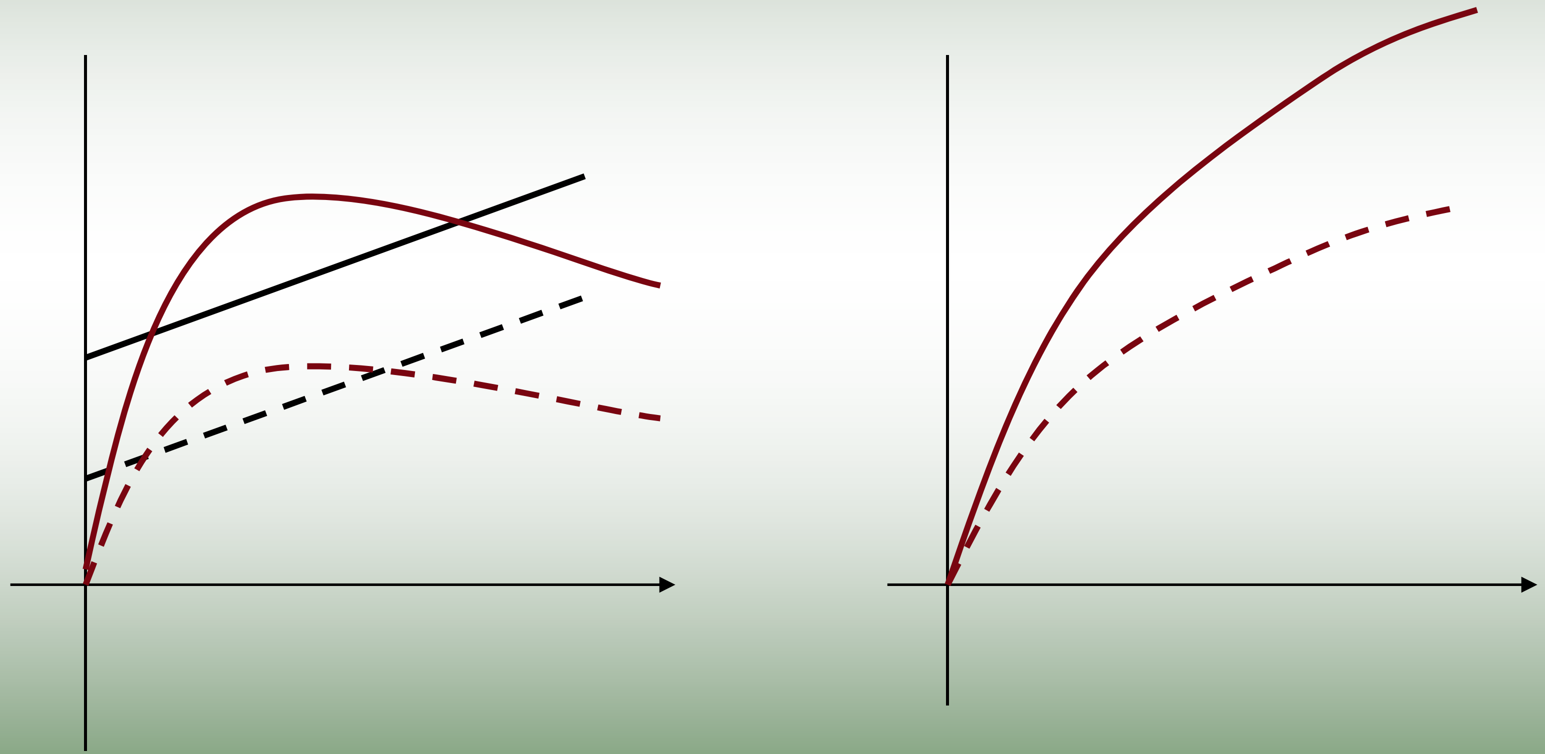
Fleming and Harrington (1997, Chapter7)

Proportional hazards (PH) model: Cox 1972, JRSS-B

- $\lambda_Z(t) = \lambda_0(t)\exp[\beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p]$
- $Z = (Z_1, Z_2, \ldots, Z_p)$
- Relative risk (RR) between an individual *i* (with $Z_i = (Z_{1i}, Z_{2i}, \ldots, Z_{pi})$) and the baseline (referent group):

$$RR(t) = \lambda_{Zi}(t)/\lambda_0(t)$$
$$= \exp[\beta_1 Z_{1\,i} + \beta_2 Z_{2\,i} + \ldots + \beta_p Z_{p\,i}]$$

Illustrating PH: $\lambda(t)$ and $\Lambda(t)$



RR between individuals *i* and *j*, with covariates Z_i and Z_j respectively:

(1) $RR(t) = \lambda_{Zi}(t)/\lambda_{Zj}(t)$

$$= \exp[\beta_1(Z_{1\,i} - Z_{1\,j}) + \ldots + \beta_p(Z_{p\,i} - Z_{p\,j})]$$

(2) If $Z_i = (Z_{1i}, Z_2, \ldots, Z_p)$ and $Z_j = (Z_{1j}, Z_2, \ldots, Z_p)$:

$RR(t) = \lambda_{Zi}(t)/\lambda_{Zj}(t) = \exp[\beta_1(Z_{1\,i} - Z_{1\,j})]$

The effect of Z_1 on the hazard (or incidence) of an event, adjusted for confounding variables.

- Under the PH model, **log-rank test** can be viewed as a **score test** for H_0: β_1=0 ; if your two-sample problem is represented by coding Z_1=1,0.
- **Log-rank test is most powerful** if the alternative is the PH-class of models
- The covariates can be **time-dependent**

Partial likelihood inference

- **partial likelihood**: (Cox, 1975, Biometrika)

  plik=$\Pi_{ti}\{\lambda_i/\Sigma_j\ Y_j(t_i)\lambda_j\}$, where

  $\lambda_i = \lambda_{Zi}(T_i) = \lambda_0(T_i)\exp[\boldsymbol{\beta}^T\mathbf{Z}]$
- log(plik) = $\mathbf{pl}(\boldsymbol{\beta})$
- partial score equation: (for $\boldsymbol{\beta}$)

  $\partial\, \mathbf{pl}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}=0$, $\boldsymbol{\beta}=(\beta_1,\beta_2,\ldots,\beta_p)^T$

Breslow estimator for the baseline cumulative hazard

- Breslow, 1972, JRSS-B

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{I(\widetilde{T}_i \le t)\Delta_i}{\sum_{j \in \mathcal{R}_i} e^{\widehat{\beta}^{\mathrm{T}} Z_j(\widetilde{T}_i)}}.$$

$$\widehat{\Lambda}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} Y_j(u) e^{\widehat{\beta}^{\mathrm{T}} Z_j(u)}}$$

$$\Delta N_i(t) = N_i(t) - N_i(t-)$$

validation of the PH model

- diagnostic plots
- local tests (for specific parameters)
- **omnibus** goodness-of-fit (GOF) tests

specification tests (1): two-sample PH test

- Gill and Schumacher (1987, Biometrika)

$$\hat{\theta}_K = \int K(t)\, d\hat{\Lambda}_2(t) \Big/ \int K(t)\, d\hat{\Lambda}_1(t)$$

H_0: $\lambda_2(t)/\lambda_1(t) = \theta$ for some positive θ,

H_1: $\lambda_2(t)/\lambda_1(t) \neq \theta$ for any positive θ.

$$\hat{\theta}_{K_i} = \hat{K}_{i2}/\hat{K}_{i1},$$

$$\hat{K}_{ij} = \int K_i(t)\, d\hat{\Lambda}_j(t)$$

Instead of the difference $\hat{K}_{22}/\hat{K}_{21} - \hat{K}_{12}/\hat{K}_{11}$

$$Q_{K_1K_2} = \hat{K}_{11}\hat{K}_{22} - \hat{K}_{21}\hat{K}_{12},$$

$$T_{K_1K_2} = \{\text{est var}\,(Q_{K_1K_2})\}^{-\frac{1}{2}} Q_{K_1K_2}$$

specification tests (2): Cox regression

- Lin (JASA, 1991)

$$U(\beta) = \sum^{n} \Delta_i \{Z_i(X_i) - E(\beta, X_i)\},$$

$$U_w(\beta) = \sum_{i=1}^{n} \Delta_i W(X_i)\{Z_i(X_i) - E(\beta, X_i)\}$$

$$Q_w = n(\hat{\beta}_w - \hat{\beta})' D_w(\hat{\beta})^{-1} (\hat{\beta}_w - \hat{\beta})$$

$$D_w(\hat{\beta}) = C_w(\hat{\beta}) - C(\hat{\beta})$$

time-dependent covariates

- $\lambda_{\mathbf{Z}}(t)= \lambda_0(t)\exp[\boldsymbol{\beta}^T\mathbf{Z}(t)]$
- $\mathbf{Z}(t)=(Z_1(t),Z_2(t),\ldots,Z_p(t))$
- **partial likelihood**:

  $plik=\Pi_{ti}\{\lambda_i/\Sigma_j\ Y_j(t_i)\lambda_j\}$, where

  $\lambda_i =\lambda_{Zi}(T_i) =\lambda_0(T_i)\exp[\boldsymbol{\beta}^T Z(t)]$

time-varying coefficients

- $\lambda_{\mathbf{Z}}(t) = \lambda_0(t)\exp[\boldsymbol{\beta}(\mathbf{t})^T\mathbf{Z}(t)]$

$$\mathcal{L}(\boldsymbol{\beta}, t) = (nh_n)^{-1}\sum_{i=1}^{n}\int_0^{\tau} K\left(\frac{s-t}{h_n}\right)\left[\boldsymbol{\beta}'\mathbf{Z}_i(s) - \log\left(\sum_{j=1}^{n} Y_j(s)e^{\boldsymbol{\beta}' Z_j(s)}\right)\right] dN_i(s),$$

$$\mathbf{U}(\boldsymbol{\beta}, t) = (nh_n)^{-1/2}\sum_{i=1}^{n}\int_0^{\tau}(\mathbf{Z}_i(s) - \mathbf{E}(\boldsymbol{\beta}, s))K\left(\frac{s-t}{h_n}\right) dN_i(s),$$

$$\mathbf{E}(\boldsymbol{\beta}, t) = S^{(1)}(\boldsymbol{\beta}, t)/S^{(0)}(\boldsymbol{\beta}, t),$$

$$S^{(r)}(\boldsymbol{\beta}, t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\mathbf{Z}_i(t)^{\otimes r} e^{\boldsymbol{\beta}'\mathbf{Z}_i(t)}, r = 0, 1, 2.$$

alternatives to the PH model

- AFT: accelerated failure time model
- Prentice and Kalbfleisch (1979, Biometrika)

$$\log T_i = \beta' x_i + e_i$$

$$U_\phi(\beta) = \Sigma_{i=1}^{n} \Delta_i \phi(\hat{F}_\beta(e_i(\beta))\Big(x_i - \frac{\sum_{k=1}^{n} x_k I(e_k(\beta) \geq e_i(\beta))}{\sum_{k=1}^{n} I(e_k(\beta) \geq e_i(\beta))}\Big)$$

- additive risk model
- Lin and Ying, 1994, Biometrika

$$\lambda(t|Z) = \lambda_0(t) + \beta_0' Z,$$

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} Z_i \{dN_i(t) - Y_i(t) d\hat{\Lambda}(\beta, t) - Y_i(t)\beta' Z_i dt\} = 0,$$

$$\left[\sum_{i=1}^{n} \int_0^{\infty} Y_i(t)\{Z_i - \bar{Z}(t)\}^{\otimes 2} dt\right] \hat{\beta} = \left[\sum_{i=1}^{n} \int_0^{\infty} \{Z_i - \bar{Z}(t)\} dN_i(t)\right],$$

$$\bar{Z}(t) \ = \ \textstyle\sum_{i=1}^{n} Y_i(t) Z_i / \sum_{i=1}^{n} Y_i(t)$$

$$\hat{\Lambda}(\hat{\beta}, t) = \int_0^{t} \frac{\{dN_i(u) - Y_i(u)\hat{\beta}' Z_i du\}}{\sum_{i=1}^{n} Y_i(u)}.$$

- heteroscedastic hazards regression model
- Hsieh (2001, JRSS-B)

$$\Lambda(Z,X;t) = \{\Lambda_0(t)\}^{\sigma}\exp\{\beta' Z\}, \sigma = \exp(\gamma' X),$$

$$\lambda(Z,X;t) = \lambda_0(t)\exp\{\beta' Z\}\sigma\{\Lambda_0(t)\}^{\sigma-1}.$$

$$\sum \int_0^t \left\{ Z_i - \frac{S_Z(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)} \right\} dN_i(u) = 0,$$

$$\sum \int_0^t \left\{ V_i - \frac{S_V(u;\Lambda_0,\theta)}{S_1(u;\Lambda_0,\theta)} \right\} dN_i(u) = 0,$$

where $Y_i(t)$ is the at-risk indicator for individual i at time t; $\sigma_i = \exp\{\gamma' X_i\}$; $V_i(t) = X_i(t)[1 + \exp\{\gamma' X_i\}\log\{\Lambda_0(t)\}]$; and $S_K(t) = (1/n)\Sigma Y_i(t) K_i(t)\exp\{\beta' Z_i\}\sigma_i\{\Lambda_0(t)\}^{\sigma_i - 1}$, for

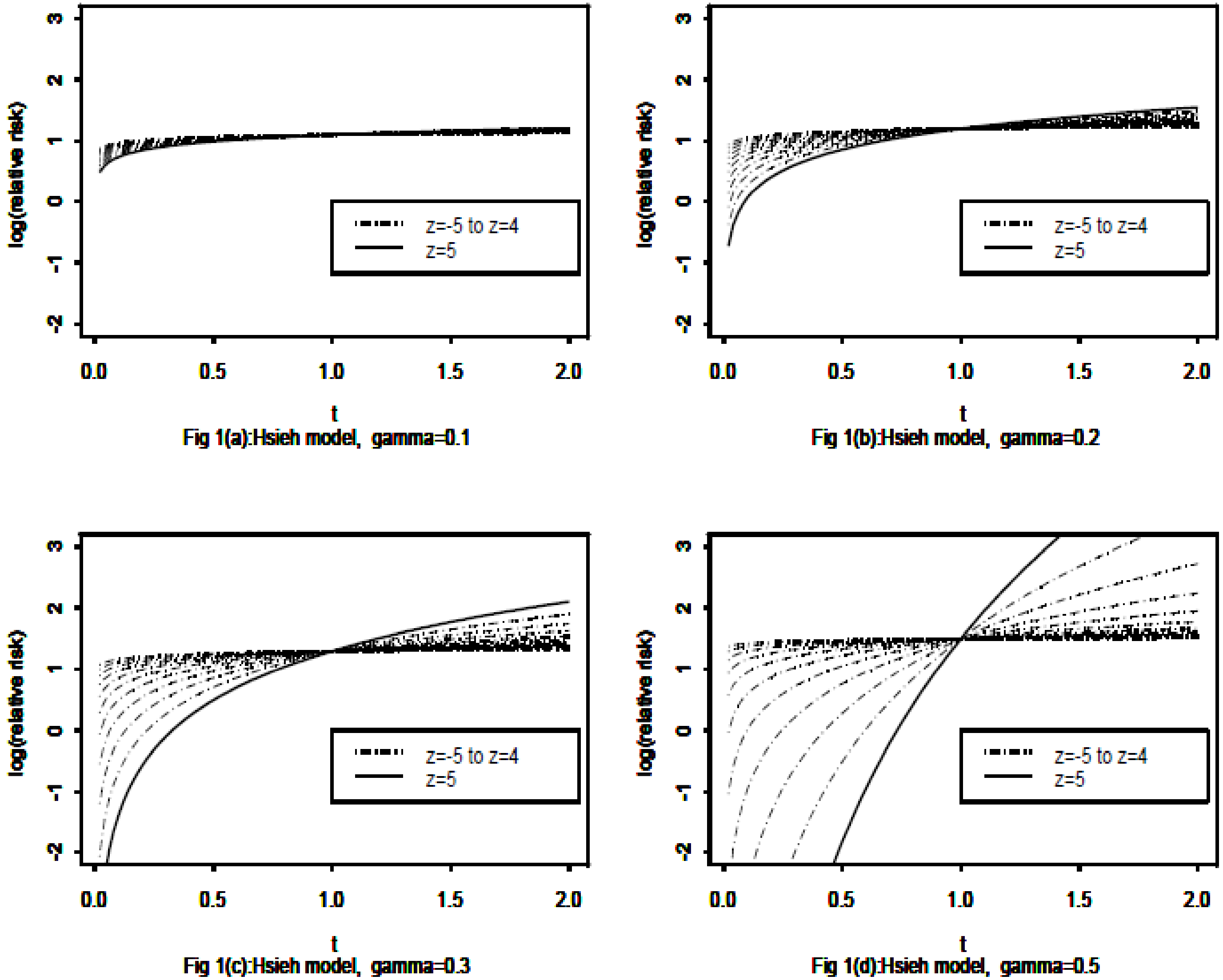


Fig. 1. Heterogeneous log-relative risks in the Hsieh model.

- proportional odds model
- Murphy and Rossini (1997, JASA)

$$-\text{logit}(S_{\mathbf{Z}}(t)) = G(t) + \mathbf{Z}^T\boldsymbol{\beta},$$

$$\text{logit}(x) = \log(x/(1-x)).$$

- frailty model
- Vaupel (1979), Hougaard (1986, Biometrika)

$$\lambda_0(t)\exp\left(\beta' z + \rho' w\right)$$

$Y = \exp(\rho' w)$, the so-called frailty.

discussions

- many other models useful for specific problems
- clinical trial applications: e.g., group sequential tests for treatment effect
- competing risk analysis
- mMultivariate survival analysis
- demographic data analysis