Event History Data Analysis

X. Joan Hu

Department of Statistics and Actuarial Science
Simon Fraser University

UBCO-MDS Short Course
April 28, 2025

Event History Data Analysis

1.1 Introduction: What is Event History Data Analysis?

It often focuses on **analysis of event times**.

- If the event is a failure,
  $\Longrightarrow$ **failure time analysis**.
  e.g. *The Statistical Analysis of Failure Time Data* by Kalbeisch and Prentice
- If the event is death,
  $\Longrightarrow$ **survival analysis**.
  e.g. *Survival Analysis* by Klein and Moeschberger
- To be positive,
  $\Longrightarrow$ **lifetime data analysis**.
  e.g. *Statistical Models and Methods for Lifetime Data* by Lawless
  e.g. the journal of *Lifetime Data Analysis*; ASA-LiDS section

1.1 Introduction: Why to Study Event History Data Analysis?

There are so many events to deal with. For example,

- death/failure
  e.g. people's death, products' failure, ...
- during the COVID-19 pandemic
  e.g. infection, hospitalization, vaccination, restoration to health, ...
- terrorist attacks, soccer corner kicks, car accidents, emergency department visits, ...

$\Longrightarrow$ **demands** of statistical learning from event times

1.1 Introduction: What to Study in Event History Data Analysis?

The focus of this short course:

- to study how to analyze the data (observations) on a continuous r.v. $T \geq 0$ (*time to an event*)
- to study how to analyze the data (observations) on $T \geq 0$ conditional on covariates Z

The special features of event time data

- various data structures – rarely there are iid observations from the population in practice; it's particularly so with event times.
- medical settings require more robust approaches – it's always desirable *to play safe* there.

1.2 Basic Concepts: Hazard Function and Survivor Function

Consider a continuous r.v. $T \geq 0$, time to an event: for $t \geq 0$,

- probability density function (pdf): $f(t)$
- cumulative distribution function (cdf): $F(t) = P(T \leq t)$
- **survivor (survival) function**: $S(t) = P(T \geq t) = 1 - F(t)$
- **hazard function**

$$h(t) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\big(T \in [t, t + \Delta t) \big| T \geq t\big)$$

  - *The instantaneous occurrence rate of an event at a fixed time given that the event has not already occurred.*
  - $h(t) = f(t)/S(t)$; $S(t) = \exp\{-\int_0^t h(u)du\}$

1.2 Basic Concepts: *Conditional* Hazard Function and Survivor Function

Consider a continuous r.v. $T \geq 0$ conditional on $Z = z$: for $t \geq 0$,

- conditional probability density function (pdf): $f(t|z)$
- conditional cumulative distribution function (cdf): $F(t|z) = P(T \leq t|Z = z)$
- **conditional survivor (survival) function**: $S(t|z) = P(T \geq t|Z = z) = 1 - F(t|z)$
- **hazard function**

$$h(t|z) = \lim_{\Delta t \to 0+} \frac{1}{\Delta t} P\big(T \in [t, t + \Delta t) \big| Z = z, T \geq t\big)$$

  - *The conditional instantaneous occurrence rate of an event at a fixed time given that the event has not already occurred.*
  - $h(t|z) = f(t|z)/S(t|z)$; $S(t|z) = \exp\{-\int_0^t h(u|z)du\}$

1.2 Basic Concepts: Censoring

A Reliability Example: To conduct an experiment to assess the quality of a certain make of (LED) light bulb ... ... (the distn of T, the lifetime of such light bulb?)

randomly select n such light bulbs, plug in them at the same time

- wait till all of them burned out: record the lifetimes $T_1, \ldots, T_n$; take them as iid observations on T.
  (if so, one may need to wait for longer than 50,000 hours)

- alternatively, choosing a time c before the experiment, stop the experiment after time c elapses: only available are T_i if $T_i \leq c$, $i = 1, \ldots, n \Leftarrow$ **type I censoring**

- or, choosing an interger $r < n$ before the experiment, stop the experiment after r number of light bulbs burn out: only available are $T_{(1)} < T_{(2)} < \ldots < T_{(r)}$. $\Leftarrow$ **type II censoring**

1.2 Basic Concepts: Censoring

What if it is in a clinical trial ...

... staggered entries of the study subjects, with a predetermined study duration?

$\Longrightarrow$ *one of the often confronted incomplete data structures:*

- **right-censoring** Let C_i be the censoring time associated with study unit i. The observed is $U_i = \min(T_i, C_i)$ (or denoted by $T_i \wedge C_i$).
  - **type I censoring.** $C_i \equiv c$ for all unit i
  - **type II censoring.** $C_i = T_{(r)}$ for all unit i

In general, the right-censored data are presented as

$$\{(U_i, \delta_i) : i = 1, \ldots, n\}\text{: } U_i = \min(T_i, C_i);\ \delta_i = \begin{cases} 1, & T_i \le C_i \\ 0, & \textit{otherwise} \end{cases}$$

e.g. $n = 3$ and $\{(4,1),(9,0),(10,1)\}$

Special features?

2.1 Commonly Used Parametric Distributions: Exponential distribution

Exponential distribution $T \sim NE(\lambda)$ with the rate $\lambda > 0$

$$f(t;\lambda) = \lambda \exp(-\lambda t), \quad t \geq 0$$

(or $f(t;\theta) = \frac{1}{\theta}\exp(-t/\theta),\ t \geq 0$ with the *scale* $\theta > 0$)

- $E(T) = 1/\lambda = \theta$ and $V(T) = 1/\lambda^2$; $S(t) = \exp(-\lambda t)$; $h(t) = \lambda$
  - the only distribution with a constant hazard function.
  - the central role in LiDA
- memoryless property: $P(T > a + b | T > a) = P(T > b)$
- In R,

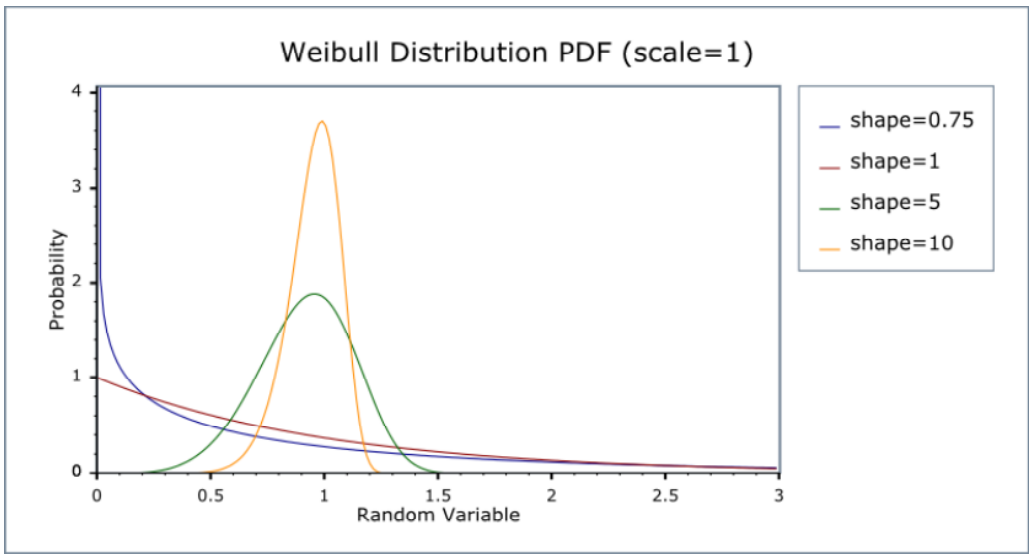
```
dexp(x, rate = 1, log = FALSE)
pexp(q, rate = 1, lower.tail = TRUE, log.p = FALSE)
qexp(p, rate = 1, lower.tail = TRUE, log.p = FALSE)
rexp(n, rate = 1)
```

2.1 Commonly Used Parametric Distributions: Weibull distribution

Weibull distribution $T \sim Weibull(k, \theta)$ with the scale $\theta > 0$ and shape $k > 0$.

- $h(t) = \frac{k}{\theta}\left(\frac{t}{\theta}\right)^{k-1}$
- $T^k \sim NE(1/\theta)$
- In R,

```
dweibull(x, shape, scale = 1, log = FALSE)
pweibull(q, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)
qweibull(p, shape, scale = 1, lower.tail = TRUE, log.p = FALSE)
rweibull(n, shape, scale = 1)
```



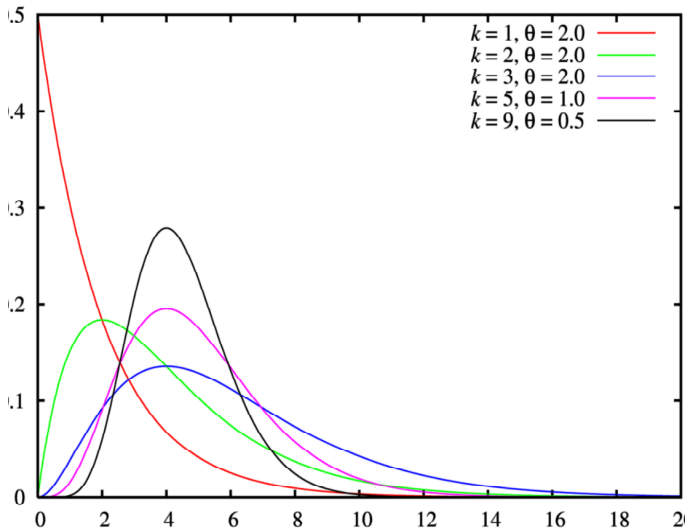
– *valelab4.ucsf.edu*

2.1 Commonly Used Parametric Distributions: Gamma distribution

Gamma distribution $T \sim \Gamma(k, \lambda)$ with the rate $\lambda > 0$ and shape $k > 0$.

- $E(T) = k/\lambda$, $Var(T) = k/\lambda^2$
- T_1, T_2 indpt and $T_j \sim \Gamma(\alpha_j, \lambda)$ for $j = 1, 2$: $T_1 + T_2 \sim \Gamma(\alpha_1 + \alpha_2, \lambda)$
- $\frac{2}{\theta} T \sim \chi^2(2k)$
- In R,

```
dgamma(x, shape, rate = 1, scale = 1/rate, log = FALSE)
pgamma(q, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
qgamma(p, shape, rate = 1, scale = 1/rate, lower.tail = TRUE, log.p = FALSE)
rgamma(n, shape, rate = 1, scale = 1/rate)
```



– *Wikimedia Commons*

2.1 Commonly Used Parametric Distributions: Other distributions

- **Log-normal distribution** $T \sim logN(\mu, \sigma)$, ie $logT \sim N(\mu, \sigma)$ with $\sigma > 0$.
  - $E(T) = \exp(\mu + \sigma^2/2)$; $S(t) =?$; $h(t) =?$
- **Extreme value distribution**
- **Gumbel distribution**
- **... ...**

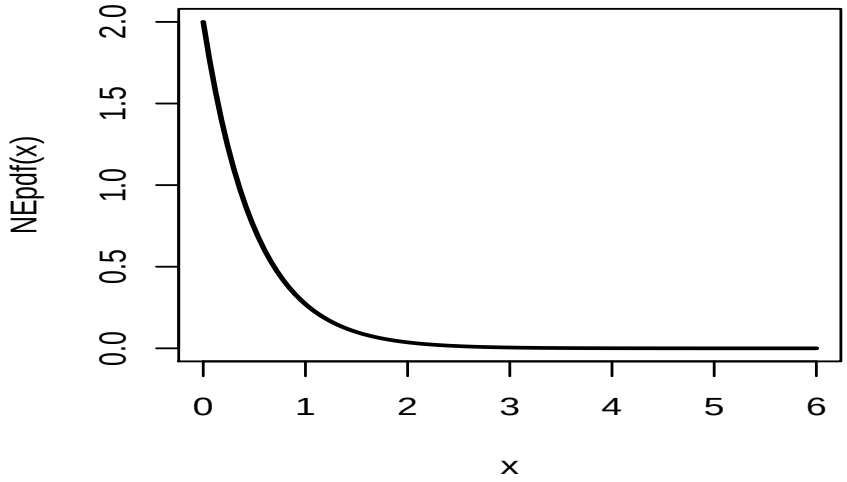
See books on reliability, such as Lawless (2003), for more examples of parametric models for event time

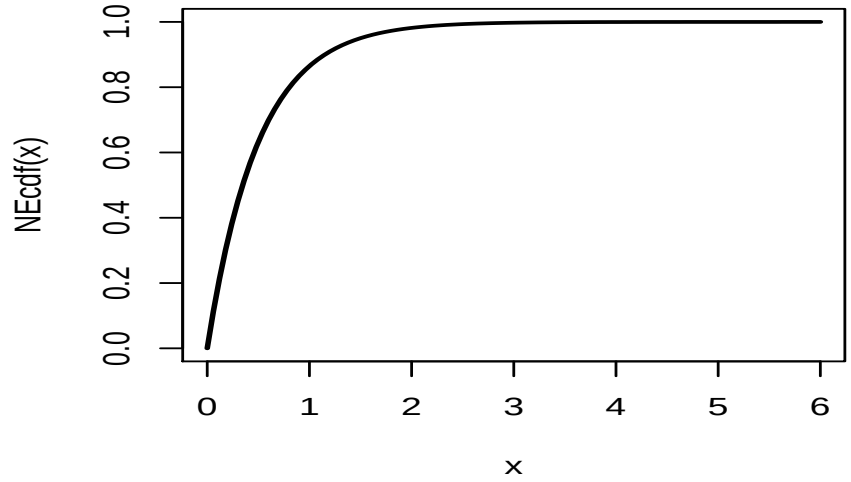
2.1 Commonly Used Parametric Distributions: Exercise 1

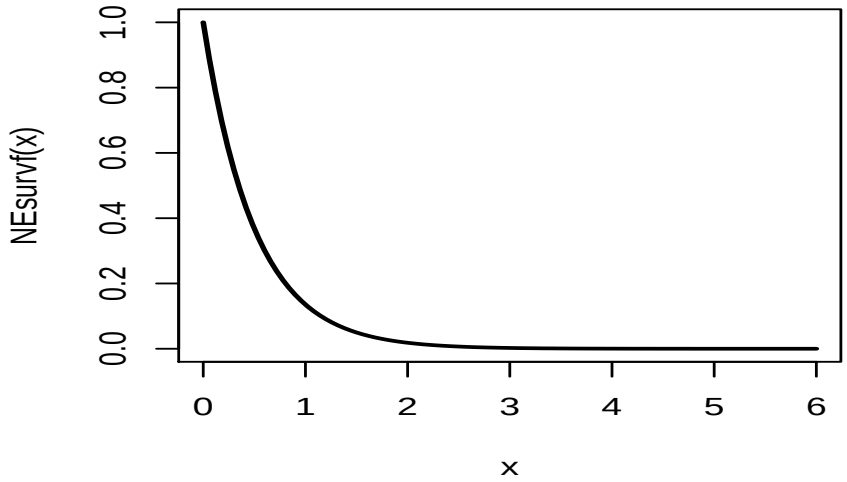
Using R to do the following if T follows (i) $NE(\lambda)$ with $\lambda = 0.5$ or (ii) $T \sim Weibull(k, \theta)$ with $\theta = 0.5, k = 3$ [**Homework 1**]

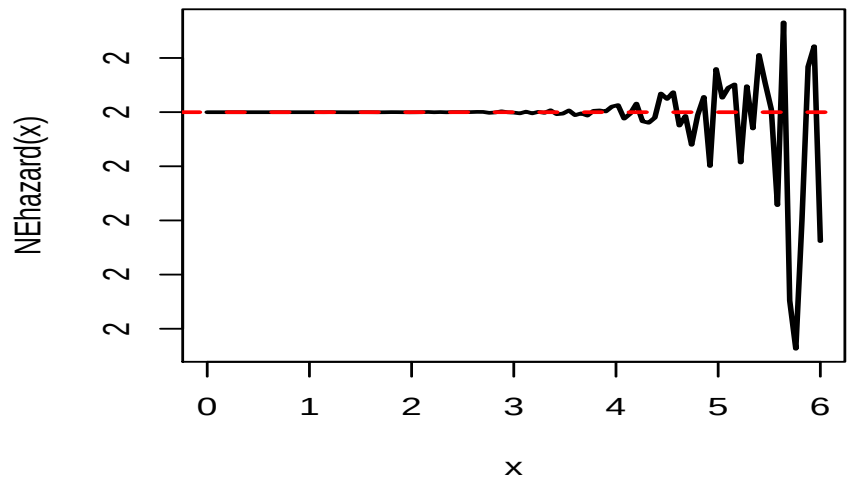
- Generate a random sample with size=1000, and obtain the sample mean, the sample variance, and the sample standard derivation.
- Plot the density, the cdf, the survivor, and the hazard functions.

```
###Exercise 1.(i)
NEobs<-rexp(n=1000, rate=2)
mean(NEobs);var(NEobs); sd(NEobs)

par(mfrow=c(2,2))
NEpdf<-function(x){dexp(x,rate=2)}
curve(NEpdf, xlim=c(0,6))
NEcdf<-function(x){pexp(x,rate=2)}
curve(NEcdf, xlim=c(0,6))
NEsurvf<-function(x){1-pexp(x,rate=2)}
curve(NEsurvf, xlim=c(0,6))
NEhazard<-function(x){dexp(x,rate=2)/(1-pexp(x,rate=2))}
curve(NEhazard,xlim=c(0,6))
abline(h=2,lty=2,col="red")
```


NEpdf(x)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
6
x


NEcdf(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
x


NEsurvf(x)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
x


NEhazard(x)
2
2
2
2
2
2
0
1
2
3
4
5
6
x

2.2 Analysis with Right-Censored Data

Consider event time r.v. $T \sim f(\cdot;\theta)$: to make inference on θ with a set of right-censored data $\{(U_i, \delta_i) : i = 1, \ldots, n\}$, arising from n indpt individuals.

- Assume **Independent Censoring**, the situations with indpt T_i and C_i for $i = 1, \ldots, n$.
- **What is the likelihood function** $L(\theta|data)$**?**

Recall that if there are iid observations $T_1, \ldots, T_n$ on T, ...

2.2 Analysis with Right-Censored Data

If knowing the likelihood function $L(\theta|data)$ with the right-censored data,

$\Longrightarrow$ **applications of MLE/likelihood-based testing proceudres**

Consider event time r.v. $T \sim f(\cdot;\theta)$: to make inference on θ with a set of right-censored data $\{(U_i, \delta_i) : i = 1, \ldots, n\}$, arising from n indpt individuals. Plus T_i and C_i are indpt.

$L(\theta|data) = \prod_{i=1}^{n} L_i(\theta)$ with $L_i(\theta)$ is the contribution from unit i: $L_i(\theta) = [U_i = u_i, \delta_i]$ is $[U_i = u_i, \delta_i = 1]$ if $\delta_i = 1$, and $[U_i = u_i, \delta_i = 0]$ if $\delta_i = 0$.

Provided $C_i \sim g(\cdot)$ with cdf $G(\cdot)$,

- $[U_i = u_i, \delta_i = 1] = [T_i = u_i, u_i \leq C_i]$
  $= [T_i = u_i \big| C_i \geq u_i][C_i \geq u_i] \propto f(u_i;\theta)\bar{G}(u_i)$;
- $[U_i = u_i, \delta_i = 0] = [C_i = u_i, u_i \leq T_i]$
  $= [C_i = u_i \big| T_i \geq u_i][T_i \geq u_i] \propto g(u_i)\bar{F}(u_i;\theta)$

Thus

$$L(\theta|data) \propto \prod_{i=1}^{n} f(u_i;\theta)^{\delta_i} S(u_i;\theta)^{1-\delta_i} = \prod_{i=1}^{n} h(u_i;\theta)^{\delta_i} S(u_i;\theta)$$

2.2 Analysis with Right-Censored Data: Example

r.v. $T \sim NE(1/\theta)$ with observations from a random sample $\{t_1, \ldots, t_n\}$ subject to right-censoring: the right-censored data $\{(u_i, \delta_i) : i = 1, \ldots, n\}$, assuming indpt censoring.

- Can we use $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T_i$, the sample mean to estimate the population mean of T, $E(T) = \theta$? ($\bar{T} \sim AN(\theta, \theta^2/n)$)

- Can the observed sample mean be a 'good estimator' $\tilde{\tilde{\theta}} = \frac{\sum_i \delta_i u_i}{\sum_i \delta_i}$?

- How about the mean of the observed event times $\tilde{\theta} = \frac{1}{n}\sum_i u_i$?

- What is the MLE of θ with the censored data?

Example. cont'd

$$L(\theta|data) = \prod_{i=1}^{n} \left(\frac{1}{\theta} e^{-u_i/\theta}\right)^{\delta_i} \left(e^{-u_i/\theta}\right)^{1-\delta_i} = \frac{1}{\theta}^{\sum_i \delta_i} \exp(-\sum_i u_i/\theta)$$

$\log L(\theta) = -\sum_i \delta_i log(\theta) - \sum_i u_i/\theta$: concave?

- the MLE $\hat{\theta} = argmax \log L(\theta)$ (e.g. R: `optimx`, `nlmin`, `nlminb`)

- alternatively,
  $\frac{\partial logL(\theta)}{\partial \theta} = -\frac{\sum_i \delta_i}{\theta} + \frac{\sum_i u_i}{\theta^2}$: decreasing?
  $\frac{\partial^2 logL(\theta)}{\partial \theta^2} = \frac{\sum_i \delta_i}{\theta^2} - 2\frac{\sum_i u_i}{\theta^3}$: negative?

Solving $\frac{\partial logL(\theta)}{\partial \theta} = 0 \Longrightarrow$ the MLE $\hat{\theta} = \frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} \delta_i}$:

$\hat{\theta} \sim AN(\theta, 1/nFI(\theta)), \quad n >> 1$

How to compare the efficiency of MLE $\hat{\theta}$ with $\bar{T} \sim AN(\theta, \theta^2/n)$?

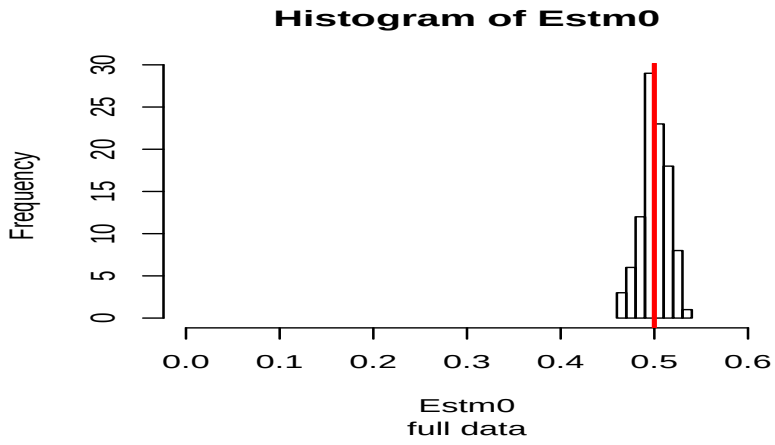
2.2 Analysis with Right-Censored Data: Exercise

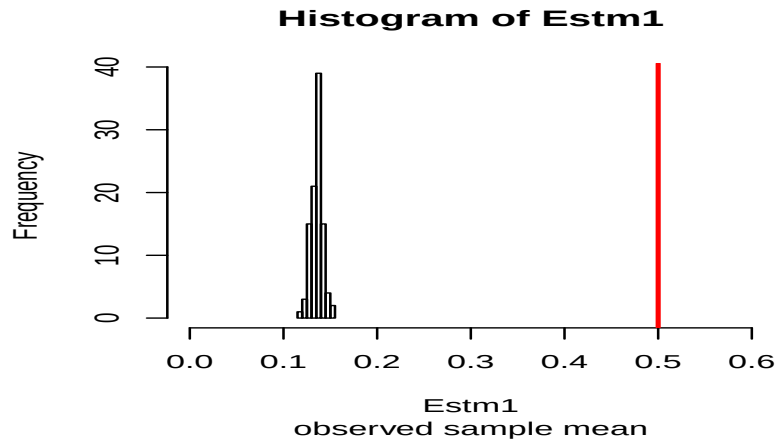
Exercise 2. Consider r.v. T following the exponential distn with scale $\theta = 0.5$. Use the generated right-censored observations on T to estimate the population mean θ:

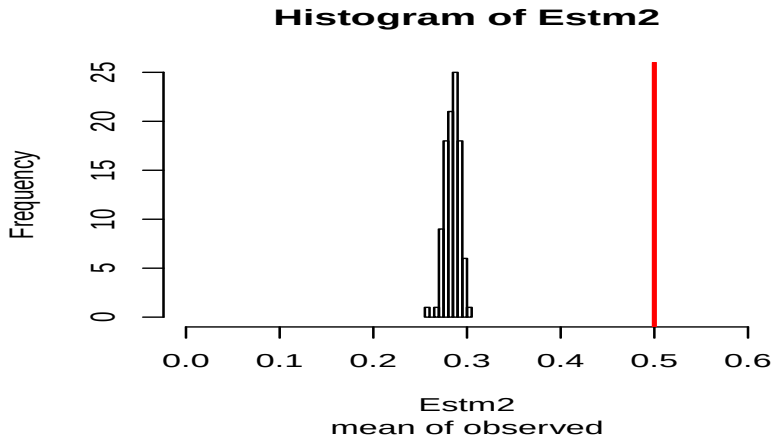
- generate a set of right-censored data with indpt censoring from $n = 1000$ indpt individuals: $\{(u_i, \delta_i) : i = 1, \ldots, n\}$:
  - sample iid $T_1, \ldots, T_n \sim f(.)$, iid $C_1, \ldots, C_n \sim Unif(0, 1)$; obtain $U_i = min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$.
- calculate (A) $\bar{T} = \sum_i T_i/n$, (B) $\tilde{\tilde{\theta}} = \sum_i \delta_i T_i / \sum_i \delta_i$, (C) $\tilde{\theta} = \sum_i U_i/n$, and (D) the MLE $\hat{\theta}$.
- repeat the two steps above $m = 100$ times and plot the histograms of the obtained estimates.

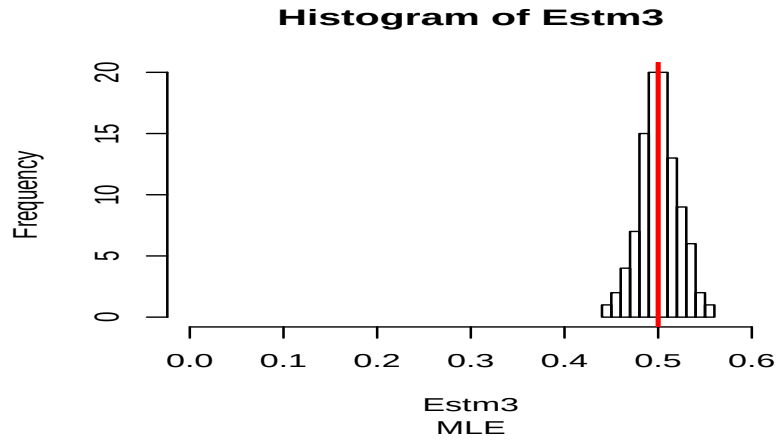
```
###Ex2.(i)A Generate n=1000 iid observations from NE(rate=2);
from Unif(0,1). Form a collection of right-censored observatio

NEobs<-rexp(n=1000, rate=2)
Censoring<-runif(n=1000, min=0,max=1)
Observed<-apply(cbind(NEobs,Censoring),1,min)
Delta<-ifelse(NEobs>Censoring,0,1)

####Ex2.(i)B Calculate the 4 estimates
estm0<-mean(NEobs)
estm1<-mean(NEobs*Delta)
estm2<-mean(Observed)
estm3<-sum(Observed)/sum(Delta)
estimates<-cbind(estm0,estm1,estm2,estm3)

> estimates
          estm0     estm1     estm2     estm3
[1,] 0.4993561 0.1467705 0.2881204 0.4993422
```

```
###Ex2.(i)C Plot the histograms of the 4 sets of estimates,
each with m=100 repetitions

Estm0<-rep(0,100);
Estm1<-rep(0,100);
Estm2<-rep(0,100);
Estm3<-rep(0,100);
for(j in 1:100){
tmpNEobs<-rexp(n=1000, rate=2)
tmpCensoring<-runif(n=1000, min=0,max=1)
tmpObserved<-apply(cbind(tmpNEobs,tmpCensoring),1,min)
tmpDelta<-ifelse(tmpNEobs>tmpCensoring,0,1)

Estm0[j]<-mean(tmpNEobs)
Estm1[j]<-mean(tmpNEobs*tmpDelta)
Estm2[j]<-mean(tmpObserved)
Estm3[j]<-sum(tmpObserved)/sum(tmpDelta)
}
#Estimates<-cbind(Estm0,Estm1,Estm2,Estm3)

par(mfrow=c(2,2))
hist(Estm0,xlim=c(0,0.6),sub="full data")
abline(v=0.5,lty=1,col="red")
hist(Estm1,xlim=c(0,0.6),sub="observed sample mean")
abline(v=0.5,lty=1,col="red")
hist(Estm2,xlim=c(0,0.6),sub="mean of observed")
abline(v=0.5,lty=1,col="red")
hist(Estm3,xlim=c(0,0.6),sub="MLE")
abline(v=0.5,lty=1,col="red")
```









What do you see? Why?

- the sample mean $\sum_i T_i/n \to \theta$ (by SLLN)
- the observed sample mean $\sum_i \delta_i U_i / \sum_i \delta_i \to E(T|T \le C)$
- the sample mean of the observed $\sum_i U_i/n \to E(T \wedge C)$
- the MLE is consistent: $\hat{\theta} \to \theta$

2.2 Analysis with Right-Censored Data: Two additional issues

In general, suppose $T \sim f(t;\theta)$. Provided that if the data collection subject to indpt right-censoring:
$\{(U_i, \delta_i) : i = 1, \ldots, n\}$

$L(\theta|\mathbf{U}, \boldsymbol{\delta}) = \prod_{i=1}^{n} f(U_i;\theta)^{\delta_i} S(U_i;\theta)^{1-\delta_i}$

$\Longrightarrow$ **applications of MLE/likelihood-based testing proceudres**

... ...

- $\Longrightarrow$ Issue 1. how to obtain MLE $\hat{\theta}$?
- $\Longrightarrow$ Issue 2. how to estimate $V(\hat{\theta})$?

Issue 1. EM (Expectation-Maximization) Algorithm (cf: Dumpster, Laird and Rubin, 1977; Self-Consistency Algorithm, cf: Turnbull, 1976) an iterative procedure for computing MLE

e.g. in the setting with right-censored data ... ...
Define $Q(\theta, \theta^*) = E\{ \log L_0(\theta|\mathbf{T})|\mathbf{U}, \boldsymbol{\delta}; \theta^* \}$

Given $\theta^{(j-1)}$, $j \geq 1$,

- E-step. $Q(\theta, \theta^{(j-1)}) = E\{ \log L_0(\theta|\mathbf{T})|\mathbf{U}, \boldsymbol{\delta}; \theta^{(j-1)} \}$
- M-step. Obtain $\theta^{(j)}$ such that $Q(\theta^{(j)}, \theta^{(j-1)}) = \max_{all\ \theta} Q(\theta, \theta^{(j-1)})$

iterating ... $\Longrightarrow \{\theta^{(j)} : j = 1, 2, \dots\}$

The sequence converges to $\hat{\theta}$, the maximum point of $\log L(\theta|\mathbf{U}, \boldsymbol{\delta})$, provided convergence.

Remarks:

- Why does it work? $\log L(\theta^{(j)}|\mathbf{U}, \boldsymbol{\delta}) \nearrow$ as $j \nearrow$
- When $\log L(\theta|\mathbf{T})$ is a linear function of $T_1, \ldots, T_n$, "E-step" is to get $E(T_i|U_i, \delta_i)$.
- "M-step" is replaced with an "S-step" when to max $Q(\theta, \theta^*)$ with fixed θ^* can be achieved by solving the equation $\partial Q(\theta, \theta^*)/\partial\theta = 0$.
- Why is it so popular?
  intuitive; not very efficient, though
- MCEM algorithm

Issue 2. Variance Estimation for MLE $\hat{\theta}$

- Recall $\hat{\theta} \sim AN\big(\theta, AV(\hat{\theta})\big)$ when $n >> 1$
  - if iid case, $AV(\hat{\theta}) = \frac{1}{n} FI(\theta)^{-1}$;
    in general,

$$AV(\hat{\theta}) = E\Big(- \frac{\partial^2 \log L(\theta)}{\partial \theta^2}\Big)^{-1} = V\Big(\frac{\partial \log L(\theta)}{\partial \theta}\Big)^{-1}$$

- Estimating $AV(\hat{\theta})$ by $- \frac{\partial^2 \log L(\theta)}{\partial \theta^2}^{-1}\Big|_{\hat{\theta}}$

- Robust Variance Estimator: the Huber sandwich estimator is based on

$$E\Big(- \frac{\partial^2 \log L(\theta)}{\partial \theta^2}\Big)^{-1} V\Big(\frac{\partial \log L(\theta)}{\partial \theta}\Big) E\Big(- \frac{\partial^2 \log L(\theta)}{\partial \theta^2}\Big)^{-1}$$

- Alternative variance estimator?

  Bootstrap, Jackknife resampling variance estimation

  - e.g. Bootstrap variance estm (cf. Efron and Tibshirani, 1993)

Viewing $\theta = \theta(F)$ and thus $\hat{\theta} = \theta(\hat{F})$ ...

data $\mathbf{X} \Rightarrow \hat{\theta}$: $V(\hat{\theta}) =$?

bootstrap samples $\mathbf{X}_1^*, \ldots, \mathbf{X}_B^* \Rightarrow \hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$:

$$\bar{\hat{\theta}}^* = \sum_{b=1}^{B} \hat{\theta}_b^* / B; \quad \hat{V}(\hat{\theta}) = \sum_{b=1}^{B} (\hat{\theta}_b^* - \bar{\hat{\theta}}^*)^2 / (B-1)$$

2.2 Analysis with Right-Censored Data: Exercise 3/Homework 3

Homework 3. Consider r.v. T following the lognormal distribution with $\log(T) \sim N(\mu, 1)$, $\mu = 0$. Use the generated right-censored observations on T to estimate μ:

- Generate a set of right-censored data with indpt censoring from $n = 1000$ indpt individuals:
  - $\{(u_i, \delta_i) : i = 1, \ldots, n\}$: sample iid
    $T_1, \ldots, T_n \sim lognormal(\mu = 0, \sigma = 1)$, iid
    $C_1, \ldots, C_n \sim Unif(0, 1.5)$; obtain $U_i = min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$;
- Calculate the MLE of μ using the generated data by the EM/MCEM algorithm.
- Obtain a bootstrap estimate of the MLE's variance using $B = 1000$.

2.2 Analysis with Right-Censored Data: Final Remarks

- What if the goal is to estimate the conditional distn of $T|Z = z \sim f(\cdot|z;\theta)$ with right-censored data $\{(U_i, \delta_i, z_i) : i = 1, \ldots, n\}$?

- Recall parametric inference in LIDA ...
  - What if the parametric model is not plausible?
  - What if it's desirable not to take much risk of model-misspecification?

$\Longrightarrow$ the demand of approaches with loose assumptions on the model structure: nonparametric/semi-parametric inference procedures
"*Modern Survival Analysis*"

Part 3. Nonparametric/Semiparametric Analysis

Overview

- Kaplan and Meier (1958, JASA)

  product-limit (Kaplan-Meier) estimator for $S(t)$ with right-censored event times – nonparametric estimator

- Mantel (1966, Cancer Chem); Gehan (1965, Biometrika)

  logrank test (extended Wilcoxon test) with right-censored event times – nonparametric test

- Cox (1972, JRSSB; 1975, Biometrika)

  Cox's proportional hazards model and partial likelihood approach – semiparametric inference

3.1 Kaplan-Meier Estimator

Motivation

$T_1, \ldots, T_n \sim F(\cdot)$ iid
the empirical distribution $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t)$, the nonparametric MLE (Kiefer's version)

the empirical distribution $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t)$

- $\forall t \in [0, \infty)$,
  - $E\{\hat{F}_n(t)\} = F(t)$
  - $Var\{\hat{F}_n(t)\} = F(t)[1 - F(t)]/n$
  - $\sqrt{n}\{\hat{F}_n(t) - F(t)\} \to N(0, F(t)[1 - F(t)])$ in distn, as $n \to \infty$
- $\sup_{t \geq 0} |\hat{F}_n(t) - F(t)| \to 0$ a.s.
- $\sqrt{n}\{\hat{F}_n(t) - F(t)\} \to$ Gaussian Process with mean zero and variance function $F(t)[1 - F(t)]$ in distribution (weak convergence)

What if $\{(U_i, \delta_i) : i = 1, \ldots, n\}$?

Table 1a Life table for the first year of life, Canada, 2000 to 2002: males

Age x	l_x	d_x	p_x	q_x	cv(q_x)	L_x	T_x	e_x	cv(e_x)
0 to 1 day	100000	252	0.99748	0.00252	4.8	273	7691798	76.92	0.04
1 to 2 days	99748	23	0.99977	0.00023	16.0	273	7691525	77.11	0.04
2 to 3 days	99725	20	0.99998	0.00020	17.1	273	7691252	77.12	0.04
3 to 4 days	99705	14	0.99986	0.00014	20.9	273	7690979	77.14	0.04
4 to 5 days	99691	11	0.99989	0.00011	23.1	273	7690706	77.15	0.04
5 to 6 days	99680	8	0.99991	0.00009	26.1	273	7690433	77.15	0.04
6 to 7 days	99672	7	0.99994	0.00006	30.6	273	7690160	77.15	0.04
0 to 7 days	100000	335	0.99665	0.00335	4.2	1911	7691798	76.92	0.04
7 to 14 days	99665	40	0.99959	0.00041	12.1	1909	7689887	77.16	0.04
14 to 21 days	99625	23	0.99977	0.00023	16.1	1908	7687978	77.17	0.04
21 to 28 days	99602	13	0.99987	0.00013	21.2	1909	7686070	77.17	0.04
0 to 28 days	100000	411	0.99589	0.00411	3.8	7637	7691798	76.92	0.04
28 days to 2 months	99589	48	0.99952	0.00048	11.1	8963	7684161	77.16	0.04
2 to 3 months	99541	33	0.99967	0.00033	13.3	8294	7675198	77.11	0.04
3 to 4 months	99508	21	0.99979	0.00021	16.7	8291	7666904	77.05	0.04
4 to -5 months	99487	16	0.99984	0.00016	19.5	8290	7658613	76.98	0.04
5 to 6 months	99471	12	0.99988	0.00012	22.0	8289	7650323	76.91	0.04
6 to 7 months	99459	12	0.99987	0.00013	21.6	8288	7642034	76.84	0.04
7 to 8 months	99447	6	0.99994	0.00006	32.2	8287	7633746	76.76	0.04
8 to 9 months	99441	5	0.99995	0.00005	34.0	8286	7625459	76.68	0.04
9 to 10 months	99436	7	0.99994	0.00006	30.6	8286	7617173	76.60	0.04
10 to 11 months	99429	5	0.99995	0.00005	34.0	8286	7608887	76.53	0.04
11 to 12 months	99424	4	0.99996	0.00004	37.8	8285	7600601	76.45	0.04

Note: Estimates with a coefficient of variation (cv) greater than 33.3% are to be used with caution
F too unreliable to be published (indicates a cv of at least 100.0%).

Recall "Actuarial Life Table"

time interval	number of death	number of withdrawal	number at risk	$\hat{q}_j$	$\hat{p}_j$	$\hat{P}_j$
I_1			... ...			
			... ...			
I_j	D_j	W_j	N_j	$\frac{D_j}{N_j - \frac{1}{2}W_j}$	$1 - \hat{q}_j$	
			... ...			
I_K			... ...			

p_j=P(an individual survives beyond I_j|beyond I_{j-1})
$q_j = 1 - p_j$=P(an individual dies in I_j|beyond I_{j-1})
P_j=P(an individual survives beyond I_j)

Rationale?

3.1 Kaplan-Meier Estimator

In general, $F \in \mathcal{F} = \{all\ cdfs\}$
With the right-censored data, the likelihood function

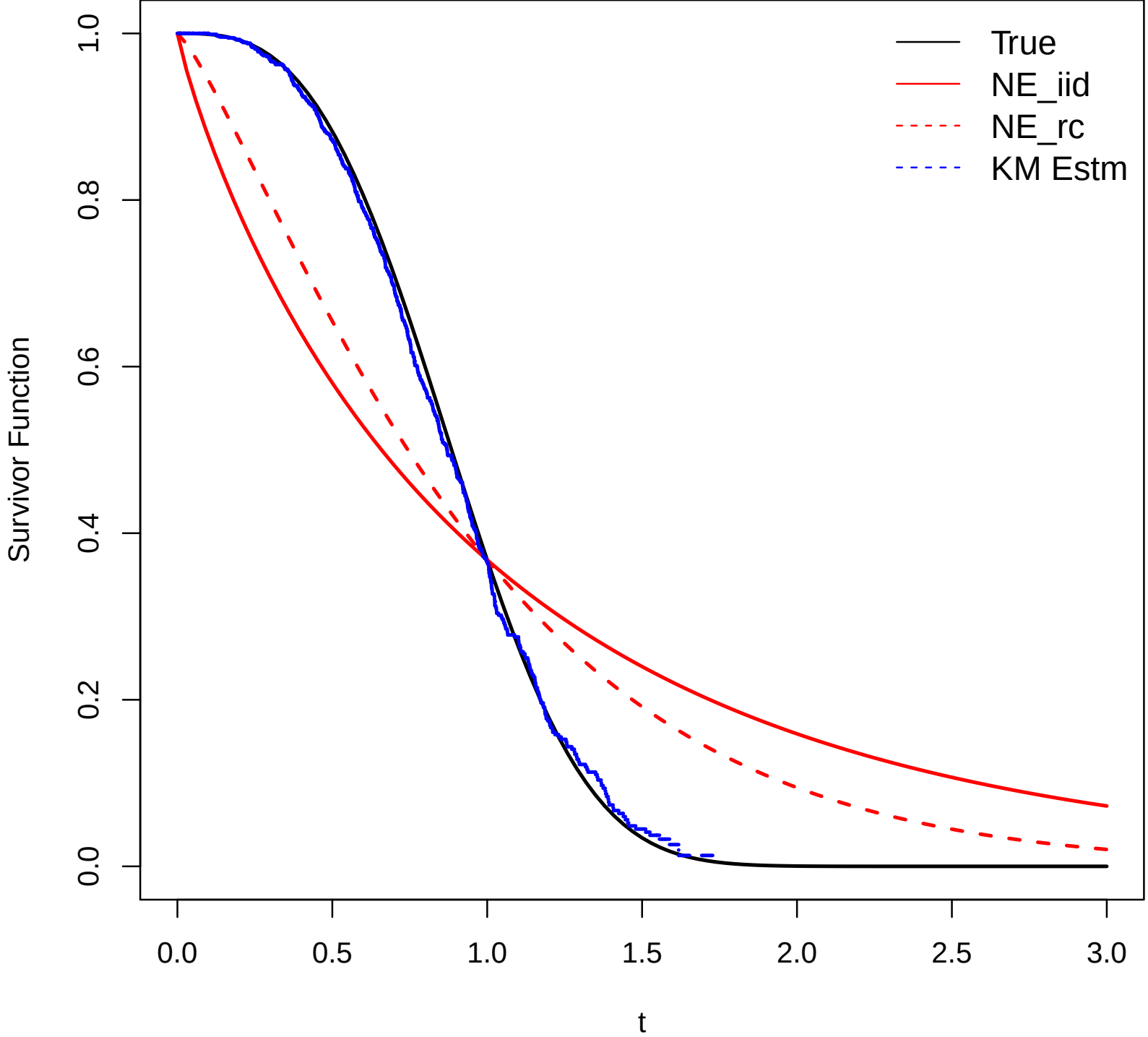
$$L(F) = \prod_{i=1}^{n} dF(u_i)^{\delta_i} [1 - F(u_i)]^{1-\delta_i}$$

Maximize $L(F)$ as $F(\cdot)$ having only masses at the distinct observed event times: $0 = V_0 \leq V_1 < \ldots < V_J \leq V_{J+1}$
$\Longrightarrow$ the Kaplan-Meier estimator (left-continuous)

$$\hat{S}(t) = \prod_{j:V_j<t} \left(1 - \frac{n_j}{N_j}\right) = \begin{cases} 1 & t \leq V_1 \\ \prod_{l=1}^{j}(1 - \hat{h}_j) & V_j < t \leq V_{j+1} \\ ? & t > V_{J+1} \end{cases}$$

3.1 Kaplan-Meier Estimator: Exercise 4

```
####Ex4.A Generate n=1000 iid observations from Weibull(shape=
#### from Unif(0,2). Then form a collection of right-censored
WBobs2<-rweibull(n=1000, shape=3, scale=1)
Censoring2<-runif(n=1000, min=0,max=2)
Observed2<-apply(cbind(WBobs2,Censoring2),1,min)
Delta2<-ifelse(WBobs2>Censoring2,0,1)
sum(Delta2);
[1] 527

###KM estm with censored data
### R package 'survival':
library("survival")
KMestm<-survfit(Surv(Observed2,Delta2)~1)
> objects(KMestm)
 [1] "call"      "conf.int"  "conf.type" "lower"
 [5] "n"         "n.censor"  "n.event"   "n.risk"
 [9] "std.err"   "surv"      "time"      "type"
[13] "upper"
```



$T \sim Weibull(shape = 3, scale = 1);\ C \sim U(0, 2);\ n = 1000, \sum \delta_i = 527$

3.1 Kaplan-Meier Estimator: Remarks

- Recall the alternative pointwise CI: for $t > 0$,
  $(\hat{S}_{KM}(t)e^{-1.96\sqrt{\hat{Var}(\hat{S}_{KM}(t))}}, \hat{S}_{KM}(t)e^{1.96\sqrt{\hat{Var}(\hat{S}_{KM}(t))}})$

- For comparing two populations' distn with censored data
  e.g. $\sup_{t>0} |\hat{S}_{1,KM}(t) - \hat{S}_{2,KM}(t)|$? an extension of the Kolmogorov-Smirnov test statistic $\sup_{t>0} |F_{1,n}(t) - F_{2,m}(t)|$
  no need to specify the population distributions into parametric models

- for assessing parametric goodness-of-fit with censored data
  - e.g. is $T \sim NE(\lambda)$ $(H(t) = \lambda t)$?
    $\Longrightarrow$ to check if $\log S(t) = -\lambda t$?
  - e.g. is $T \sim Weibull(\lambda, \rho)$ $(H(t) = \lambda t^{\rho})$?
    $\Longrightarrow$ to check if $\log\big(-\log S(t)\big) = \log\lambda + \rho \log t$?

3.2 Logrank Test

Introduction

Consider to compare two groups wrt the event time distns ... ...

For example,

- in the placebo group, iid $T_{0i} \sim F_0(\cdot)$: $i = 1, \ldots, n$
- in the treatment group, iid $T_{1j} \sim F_1(\cdot)$: $j = 1, \ldots, m$

$\Longrightarrow H_0 : F_0(\cdot) = F_1(\cdot)$

... Many different ways to differ: any UMP?

- *directional tests*: designated/oriented to a specific type of difference between the two population distns
  e.g. $S_1(t) = S_0(t)^c$

- *omnibus tests*: there is power to detect all or most types of differences but not with great power for a specific difference

Early work with censored data ... ...

- Gehan (1965, Biometrika): modifying rank tests to allow censoring
  $\Longrightarrow$ Wilcoxon-Gehan testing procedure ...

- Mantel (1966, Cancer Chem): adapting data to use methods for several 2×2 tables
  $\Longrightarrow$ Logrank testing procedure ...

- Application of the Cox partial likelihood approach (Cox, 1975)*

3.2 Logrank Test

with all observed distinct event times: $0 < V_1 < \ldots, V_K$

First, consider what happens at time $t = V_l$ ...

	at $t = V_l$		
Group	failure	not	at risk
placebo	n_{0l}	–	N_{0l}
treatment	n_{1l}	–	N_{1l}
total	$n_{.l}$	–	$N_{.l}$

- the expected number of failures from treatment group $E_l = E(O_l) = n_l \frac{N_{1l}}{N_l}$ under H_0
- $V(O_l) = \frac{N_l - N_{1l}}{N_l - 1} N_{1l} \left(\frac{n_l}{N_l}\right)\left(1 - \frac{n_l}{N_l}\right)$ under H_0

Now, pull together the information at all the observed failure times ...

$$Z = \frac{\sum_{l=1}^{K} (O_l - E_l)}{\sqrt{\sum_{l=1}^{K} V(O_l)}} \sim N(0, 1)$$

approximately under H_0

$\Longrightarrow$ the Mantel (logrank) testing procedure ...

3.2 Logrank Test

Example. Group 0: $3.1,\ 6.8^{+},\ 9,\ 9,\ 11.3^{+},\ 16.2$

Group 1: $8.7,\ 9,\ 10.1^{+},\ 12.1^{+},\ 18.7,\ 23.1^{+}$

3.2 Logrank Test: Variants of Logrank Test

- What if the subjects are stratified according to a factor, say, gender?

  Stratified Logrank Test with the factor of K levels

$$Z = \frac{\sum_{k=1}^{K}(O^{(k)} - E^{(k)})}{\left(\sum_k V^{(k)}\right)^{1/2}} \sim N(0,1)$$

  approximately under H_0.

- What if there is a need to weight the information at different times differently?

  Weighted Logrank Test

$$Z_W = \frac{\sum_{l=1}^{L} w_l(O_l - E_l)}{\left(\sum_l w_l^2 V_l\right)^{1/2}} \sim N(0,1)$$

  approximately under H_0.

  How to choose the weights in general?

  - If $w_l = N_{\cdot l}$, the test is similar to Gehan test.

3.2 Logrank Test: Variants of Logrank Test

- What if to compare p treatment groups with the placebo group?

  $H_0 : S_0(\cdot) = S_1(\cdot) = \ldots = S_p(\cdot)$

  Given all the distinct failure times are $0 < V_1 < \ldots < V_L < \infty$,

	at $t = V_l$		
Group	failure	not	at risk
placebo	n_{0l}	...	N_{0l}
treatment 1	n_{1l}	...	N_{1l}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
treatment p	n_{pl}	...	N_{pl}
total	$n_{.l}$	...	$N_{.l}$

$$\mathbf{O}_l = \begin{pmatrix} n_{1l} \\ \vdots \\ n_{pl} \end{pmatrix};\ \mathbf{E}_l = E\{\mathbf{O}_l\} = \begin{pmatrix} N_{1l} \\ \vdots \\ N_{pl} \end{pmatrix} \frac{n.l}{N_{.l}};\ \mathbf{V}_l = Var\{\mathbf{O}\}$$

$$\tilde{\mathbf{O}} = \sum_{l=1}^{L} \mathbf{O}_l,\ \tilde{\mathbf{E}} = \sum_{l=1}^{L} \mathbf{E}_l,\ \tilde{\mathbf{V}} = \sum_{l=1}^{L} \mathbf{V}_l$$

$$\left(\tilde{\mathbf{O}} - \tilde{\mathbf{E}}\right)' \tilde{\mathbf{V}}^{-1} \left(\tilde{\mathbf{O}} - \tilde{\mathbf{E}}\right) \sim \chi^2(p)$$

approximately under H_0, provided the sample size is large.

- The test is *omnibus*.
- If a trend test is intended?

3.3 Cox Proportinal Hazards Model

- Recall the two-sample problem $\rightarrow$ testing on $H_0 : h_1(\cdot) = h_0(\cdot)$
  - $Z = \begin{cases} 1 & treatment \\ 0 & placebo \end{cases}$,
    to study event time $T|Z = z$?
  - with general covariates Z, to explore event time $T|Z = z$?
    $\Longrightarrow$ regression modeling?

- Feigl and Zelen (1965)
  $T|Z = z \sim NE(\lambda_z)$: $h(t|z) = \lambda_z = \lambda_0 e^{\beta z}$
  $\beta = 0 \rightarrow$ no effect of Z

$\Longrightarrow$ **Cox Proportional Hazards Model** (Cox, JRSSB 1972)

Cox Proportional Hazards Model: (Cox, JRSSB 1972)
The hazard function of event time $T|Z = z$ is

$$h(t|z) = h_0(t)e^{\beta z}, \quad t > 0$$

The conditional survivor function is

$$S(t|z) = \exp(-\int_0^t h_0(u)e^{\beta z}du) = \exp(-H_0(t)e^{\beta z}), \quad t > 0$$

Remark:

- the hazard ratio $h(t|Z = z_1)/h(t|Z = z_0) = e^{\beta(z_1 - z_0)}$ for all $t > 0$ *proportional*!

3.3 Cox Proportional Hazards Model: Estimation of β

Often is interested to estm β in the Cox PH model, for comparison/evaluate/assess effect ... ...

With right-censored event times along with the covariates

$$\{(U_i, \delta_i, Z_i) : i = 1, \ldots, n\}$$

from n indpt subjects and indpt censoring

$$L(\beta, h_0(\cdot)|data) = \prod_{i=1}^{n} \left(h_0(u_i)e^{\beta z_i}\right)^{\delta_i} \exp(-H_0(u_i)e^{\beta z_i})$$

$$L(\beta, h_0(\cdot)|data) = L_1(\beta|data)L_2(\beta, h_0(\cdot)|data)$$

$\Longrightarrow$ **the Cox partial likelihood function** (Cox, Biometrika 1975)

the Cox partial likelihood function (Cox, Biometrika 1975)

$$L_1(\beta|data) = \prod_{i=1}^{n} \left(\frac{e^{\beta z_i}}{\sum_{l \in \mathcal{R}_i} e^{\beta z_l}} \right)^{\delta_i}$$

the risk set at time u_i: $\mathcal{R}_i = \{j : u_j \geq u_i\}$

$\Longrightarrow$ the MPLE (maximum partial likelihood estimator) of β:

$$\hat{\beta} = argmax_{all\ \beta} L_1(\beta|data)$$

With some conditions, as $n \to \infty$

- $\hat{\beta} \to \beta$ a.s.
- $\sqrt{n}(\hat{\beta} - \beta) \to N(0, ?)$ in distn

Example. $n = 5$ indpt subjects and $Z = \begin{cases} 1 & \textit{treatment} \\ 0 & \textit{placebo} \end{cases}$

(u_i, δ_i, z_i) : $(16,1,1),\ (13,0,0),\ (21,1,1),\ (11,1,0),\ (12,1,1)$

$$L_1(\beta) \propto \frac{e^{\beta}}{(3e^{\beta}+2)(3e^{\beta}+1)},\ \ \partial \log L_1(\beta)/\partial\beta = 1 - \frac{9e^{\beta}(2e^{\beta}+1)}{(3e^{\beta}+2)(3e^{\beta}+1)}$$

$$\Longrightarrow \hat{\beta} = \tfrac{1}{2}\log 2 - \log 3$$

3.3 Cox Proportional Hazards Model: Testing on β

Consider $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
the partial score test

$$U(\beta) = \partial \log L_1(\beta)/\partial\beta = \sum_{i=1}^{n} \delta_i \Big[z_i - \frac{\sum_{l \in \mathcal{R}_i} z_l e^{\beta z_l}}{\sum_{l \in \mathcal{R}_i} e^{\beta z_l}} \Big]$$

Based on $U(\beta)/\sqrt{n} \sim AN(0, ??)$ as $n \to \infty$ with some conditions,
$\Longrightarrow$ the partial score testing procedure ...

Remark.

- e.g. when $Z = \begin{cases} 1 & treatment \\ 0 & placebo \end{cases}$
  $U(\beta)\big|_{\beta=0} = \sum_{l=1}^{L} \left(O_l - n_{.l} \frac{N_{1l}}{N_{.l}} \right) = O - E$, the numerator of the logrank test statistic

3.3 Cox Proportional Hazards Model: Exercise 5

```
####Ex5.A Generate n=1000 observations from each
####NE(rate=1) and NE(rate=exp(0.5)); from each Unif(0,1),Unif
####for censoring times. Then form a collection of right-censo
sum(Delta3a); sum(Delta3b);
[1] 365
[1] 632

###Cox PH model fits with censored data
### R package 'survival':
library("survival")
Coxphoverall<-coxph(Surv(c(Observed3a,Observed3b),
c(Delta3a,Delta3b))~Zindicator)
> summary(Coxphoverall)
  n= 2000, number of events= 997
               coef exp(coef) se(coef)     z Pr(>|z|)
Zindicator 0.47304   1.60486  0.06639 7.125 1.04e-12 ***

           exp(coef) exp(-coef) lower .95 upper .95
Zindicator     1.605     0.6231     1.409     1.828

Concordance= 0.553  (se = 0.009 )
Rsquare= 0.026   (max possible= 0.999 )
Likelihood ratio test= 52.31  on 1 df,   p=5e-13
Wald test            = 50.76  on 1 df,   p=1e-12
Score (logrank) test = 51.7  on 1 df,   p=6e-13
```

Part 4. Further Topics

- More unconventional data structures
  - interval censoring
  - current status data
  - truncated data
  - competing risks
  - ... ...
- Beyond survival analysis
  - what if the event is recurrent?
  - what if there are multiple types of events?
  - what if events take place spatio-temporally?
  - ... ...

Thank-you for your participation in this course!

What have we studied?

- *Part 1. Preliminaries*
  - Introduction
  - Basic concepts
- *Part 2. Parametric Interence in LIDA*
  - Commonly used parametric models
  - Inference with right-censored data
- *Part 3. Nonparametric/Semi-parametric Approaches*
  - Kaplan-Meier estimator
  - Logrank test
  - Cox proportional hazards model
- *Part 4. Further Topics*