

# **GSERM - Oslo 2019**

## Introduction to Survival Data

January 9, 2019 (afternoon session)

# Survival Analysis

- Models for *time-to-event data*.
- Roots in biostats/epidemiology, plus engineering, sociology, economics.
- Examples...
  - Political careers, confirmation durations, position-taking, bill cosponsorship, campaign contributions, policy innovation/adoption, etc.
  - Cabinet/government durations, length of civil wars, coalition durability, etc.
  - War duration, peace duration, alliance longevity, length of trade agreements, etc.
  - Strike durations, work careers (including promotions, firings, etc.), criminal careers, marriage and child-bearing behavior, etc.

# Characteristics of Time-To-Event Data

- Discrete events (i.e., not continuous),
- Take place over time,
- May not (or *never*) experience the event (i.e., possibility of censoring).

# Survival Data Basics: Terminology

$Y_i$  = the duration until the event occurs,

$Z_i$  = the duration until the observation is “censored”

$T_i$  =  $\min\{Y_i, Z_i\}$ ,

$C_i$  = 0 if observation  $i$  is censored, 1 if it is not.

# Survival Data Basics: The Density

$$f(t) = \Pr(T_i = t)$$

Issues:

- $T_i = t$  iff  $T_i > t - 1, t - 2$ , etc.
- $C_i = 0$  (censoring)

# Survival Data Basics: Survivor Function

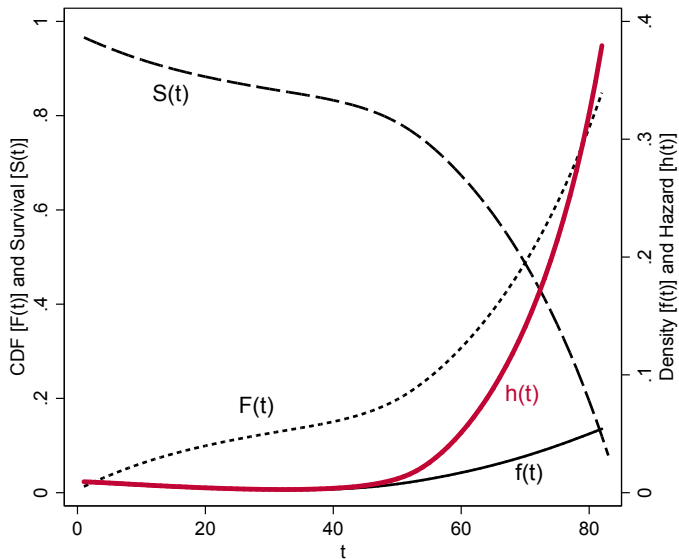
$$\Pr(T_i \leq t) \equiv F(t) = \int_0^t f(t) dt$$

$$\begin{aligned}\Pr(T_i \geq t) \equiv S(t) &= 1 - F(t) \\ &= 1 - \int_0^t f(t) dt\end{aligned}$$

## Survival Data Basics: The Hazard

$$\begin{aligned}\Pr(T_i = t | T_i \geq t) \equiv h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{f(t)}{1 - \int_0^t f(t) dt}\end{aligned}$$

## Example: Human Mortality





# Some Useful Equivalencies

$$f(t) = \frac{-\partial S(t)}{\partial t}$$

Implies

$$\begin{aligned} h(t) &= \frac{\frac{-\partial S(t)}{\partial t}}{S(t)} \\ &= \frac{-\partial \ln S(t)}{\partial t} \end{aligned}$$

# More Useful Things: Integrated Hazard

Define

$$H(t) = \int_0^t h(t) dt.$$

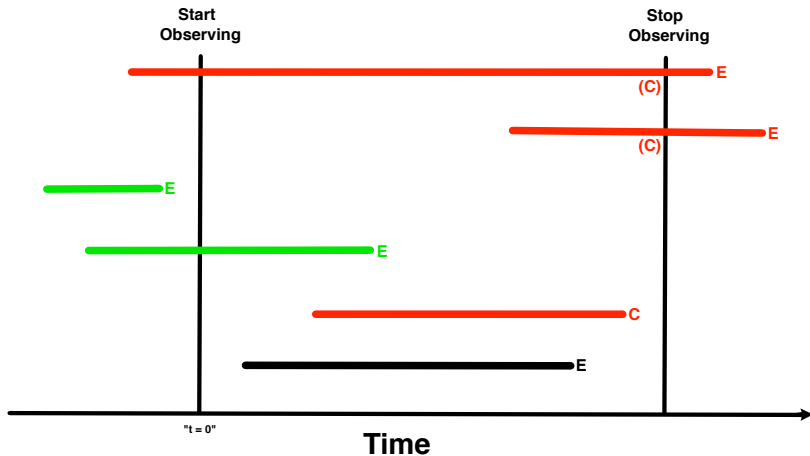
Implies

$$\begin{aligned} H(t) &= \int_0^t \frac{-\partial \ln S(t)}{\partial t} dt \\ &= -\ln[S(t)] \end{aligned}$$

and

$$S(t) = \exp[-H(t)]$$

# Censoring and Truncation



- Defined by the researcher
- Conditionally independent of both  $T_i$  and  $\mathbf{X}_i$
- Doesn't mean that the observation provides no information

# Estimating $S(t)$

Assume  $N$  observations, *absorbing* events, and no ties. Then define

- $n_t$  = number of observations “at risk” for the event at  $t$ , and
- $d_t$  = number of observations which experience the event at time  $t$ .

Then

$$\widehat{S(t_k)} = \prod_{t \leq t_k} \frac{n_t - d_t}{n_t}$$

## Variance of $\widehat{S}(t)$

$$\text{Var}[\widehat{S}(t_k)] = \left[\widehat{S}(t_k)\right]^2 \sum_{t \leq t_k} \frac{d_t}{n_t(n_t - d_t)}$$

Note:

- $\text{Var}[\widehat{S}(t_k)]$  is increasing in  $S(t)$ ,
- is also increasing in  $d_t$ , but
- is decreasing in  $n_t$ .

“Nelson-Aalen”:

$$\widehat{H(t_k)} = \sum_{t \leq t_k} \frac{d_t}{n_t}$$

...which gives an alternative estimator for the survival function equal to:

$$\begin{aligned}\widehat{S(t_k)} &= \exp[-\widehat{H(t_k)}] \\ &= \exp \left[ - \sum_{t \leq t_k} \frac{d_t}{n_t} \right]\end{aligned}$$

# Bivariate Hypothesis Testing

	Treatment	Placebo	Total
Event	$d_{1t}$	$d_{0t}$	$d_t$
No Event	$n_{1t} - d_{1t}$	$n_{0t} - d_{0t}$	$n_t - d_t$
Total	$n_{1t}$	$n_{0t}$	$n_t$

Log-Rank Test:

$$Q = \frac{\left[ \sum (d_{1t} - \frac{n_{1t}d_t}{n_t}) \right]^2}{\left[ \frac{n_{1t}n_{0t}d_t(n_t - d_t)}{n_t^2(n_t - 1)} \right]}$$
$$\sim \chi_1^2$$



# A Diversion: Survival Models and Counting Processes

Assume

- Event is *absorbing*,
- $Y_i$  is duration to the event
- $Z_i$  is duration to censoring
- Observe  $T_i = \min(Y_i, Z_i)$ , and
- $C_i$ :
  - $C_i = 0$  if  $T_i = Z_i$ ,
  - $C_i = 1$  if  $T_i = Y_i$ .
- $T_i \neq T_j \forall i \neq j$  (no “ties”)

# Three Key Variables

1. *Counting Process* Indicator:

$$N_i(t) = I(T_i \leq t, C_i = 1)$$

2. *Risk* Indicator:

$$R_i(t) = I(T_i > t)$$

3. *Intensity Process*:

$$\lambda_i(t) dt = R_i(t)h(t)$$

With

$$\Lambda_i(t) = \int_0^t \lambda_i(t) dt$$

we can think of

$$N_i(t) = \Lambda_i(t) + M_i(t)$$

or

$$M_i(t) = N_i(t) - \Lambda_i(t).$$

$$E(X_{t+s} | X_0, X_1, \dots, X_i, \dots, X_t) = X_t \quad \forall s > 0$$

# Data Structure and Organization: Non-Time-Varying

id	durat	censor	timein	timeout	X
1	4	0	30	34	0.12
2	2	1	12	14	0.19
3	5	1	5	10	0.09
...	...	...	...	...	...
N	10	1	21	31	0.22

# Time-Varying Data

id	durat	censor	timein	timeout	X	Z
1	1	0	30	31	0.12	331
1	2	0	31	32	0.12	412
1	3	0	32	33	0.12	405
1	4	0	33	34	0.12	416
2	1	0	12	13	0.19	226
2	2	1	13	14	0.19	296
3	1	0	5	6	0.09	253
3	2	0	6	7	0.09	311
3	3	0	7	8	0.09	327
3	4	0	8	9	0.09	344
3	5	1	9	10	0.09	301
...	...	...	...	...	...	...

# Analyzing Survival Data in R

survival object (non-time-varying):

```
library(survival)
NonTV<-read.csv(NonTVdata.csv)
NonTV.S<-Surv(NonTV$duration, NonTV$censor)
```

survival object (time-varying):

```
TV<-read.csv(TVdata.csv)
TV.S<-Surv(TV$starttime, TV$endtime, TV$censor)
```

# An Example

OECD Cabinet survival [Strom (1985); King et al. (1990)],

$N = 314$  cabinets in 15 countries

Outcome: Duration of cabinet, in months

Covariates (all non-time varying):

- *Fractionalization*
- *Polarization*
- *Formation Attempts*
- **Investiture**
- *Numerical Status*
- *Post-Election*
- *Caretaker*

Also: Indicator for whether the cabinet ended within 12 months of the end of the “constitutional inter-election period” (→ censored)



```
> head(KABL)
  id country durat ciep12 fract polar format invest numst2 eltime2 caret2
1  1      1  0.5     1   656    11     3      1      0      1      0
2  2      1  3.0     1   656    11     2      1      1      0      0
3  3      1  7.0     1   656    11     5      1      1      0      0
4  4      1 20.0     1   656    11     2      1      1      0      0
5  5      1  6.0     1   656    11     3      1      1      0      0
6  6      1  7.0     1   634     6     4      1      1      1      0
```

```
> KABL.S<-Surv(KABL$durat,KABL$ciep12)
```

```
> KABL.S[1:50,]
```

```
[1] 0.5  3.0  7.0 20.0  6.0  7.0  2.0 17.0 27.0 49.0+
[11] 4.0 29.0 49.0+ 6.0 23.0 41.0+ 10.0 12.0  2.0 33.0
[21] 1.0 16.0  2.0  9.0  3.0  5.0  5.0  6.0 45.0+ 23.0
[31] 41.0  7.0 49.0+ 46.0  9.0 51.0+ 10.0 32.0 28.0  3.0
[41] 53.0+ 17.0 59.0+  9.0 52.0+  3.0 23.0 33.0  1.0 30.0
```

# Example survfit Object

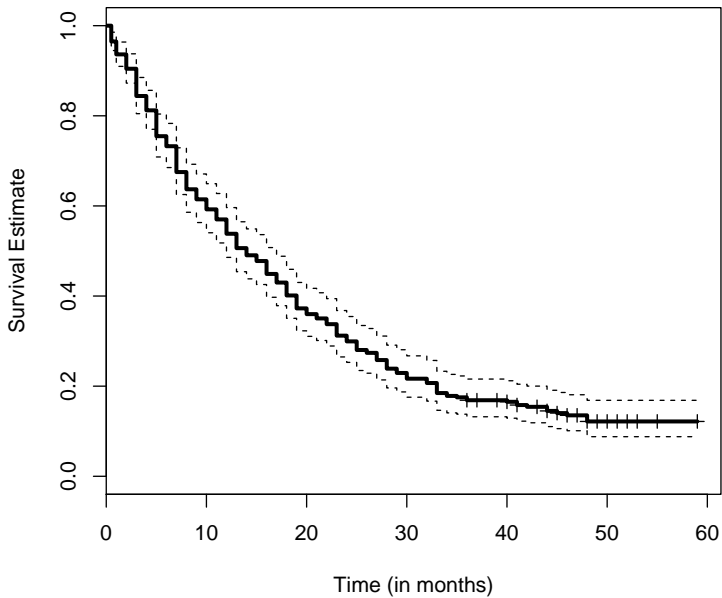
```
> KABL.fit<-survfit(KABL.S~1)
```

```
> str(KABL.fit)
```

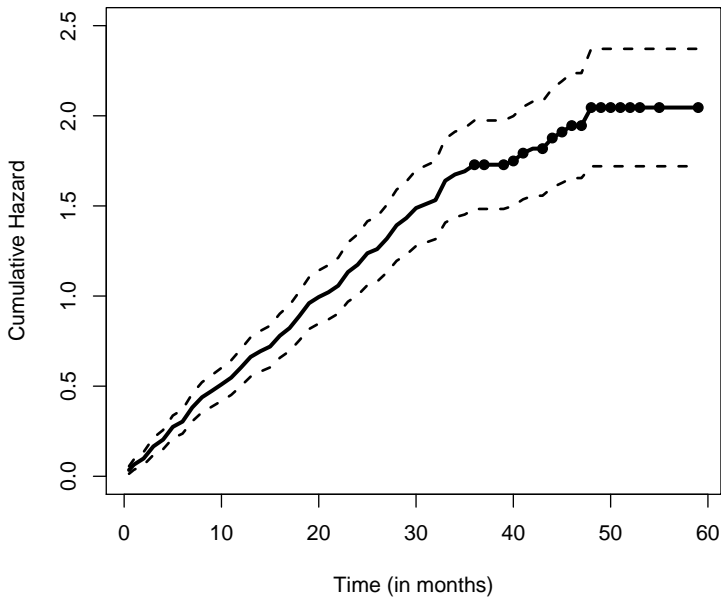
List of 13

```
$ n      : int 314
$ time   : num [1:54] 0.5 1 2 3 4 5 6 7 8 9 ...
$ n.risk  : num [1:54] 314 303 294 284 265 255 237 230 212 200 ...
$ n.event : num [1:54] 11 9 10 19 10 18 7 18 12 7 ...
$ n.censor : num [1:54] 0 0 0 0 0 0 0 0 0 0 ...
$ surv    : num [1:54] 0.965 0.936 0.904 0.844 0.812 ...
$ type    : chr "right"
$ std.err  : num [1:54] 0.0108 0.0147 0.0183 0.0243 0.0271 ...
$ upper    : num [1:54] 0.986 0.964 0.938 0.885 0.856 ...
$ lower    : num [1:54] 0.945 0.91 0.873 0.805 0.77 ...
$ conf.type: chr "log"
$ conf.int : num 0.95
$ call     : language survfit(formula = KABL.S ~ 1)
- attr(*, "class")= chr "survfit"
```

Plotting  $\widehat{S}(t)$



Plotting  $\widehat{H}(t)$



Log-rank test:

```
> survdiff(KABL.S~invest,data=KABL,rho=0)
```

Call:

```
survdiffformula = KABL.S ~ invest, data = KABL, rho = 0)
```

	N	Observed	Expected	(O-E) <sup>2</sup> /E	(O-E) <sup>2</sup> /V
invest=0	172	137	178.7	9.72	30.5
invest=1	142	134	92.3	18.81	30.5

Chisq= 30.5 on 1 degrees of freedom, p= 3.26e-08

# Comparing $\widehat{S}(t)$ s

