Describing Continuous Time Event Occurrence Data

(Chapter 13, *ALDA*)

Judy Singer & John Willett

Harvard University Graduate School of Education
May, 2003

What we will cover

What are the salient features of continuous-time event data and how do they differ from discrete-time event data?

(ALDA, Section 13.1, p. 471)

When we assess event occurrence in continuous time:

- We know the *precise instant* when events occur—e.g., Jane took her first drink at 6:19 after release from an alcohol treatment program
- There exists an infinite number of these instants. Any division of continuous time—weeks, days, hours, etc—can always be made finer (in contrast to the finite—and usually small—number of values for TIME in discrete-time)
- The probability of observing any particular event time is infinitesimally small (it approaches 0 as time's divisions get finer).
  - This has serious implications for the definition of hazard, the lynchpin of discrete-time survival analysis.
  - Continuous-time hazard must be defined differently, and it is difficult to estimate and display it in data analysis
- The probability of ties—two or more individuals sharing an event time—is therefore infinitesimally small
  - Continuous-time survival methods assume *no ties*. When they exist—and they inevitably do—they can cause difficulties.
  - Why are ties inevitable in continuous time? Continuous-time data are really not continuous. Because they are collected to the nearest unit (year, month, week, etc), they are really "rounded." The theory of continuous-time survival analysis, however, is developed assuming that the probability of ties approaches 0.

Data example: Time to horn honk

(ALDA, Table 13.1, p. 471)

- *Research Question*: How long does it take for drivers to honk at a VW Jetta purposefully blocking their car at a busy intersection?
- *Citation*: Diekmann, Jungbauer-Gans, Krassnig & Lorenz (1996).
- *Design*:
  - 57 motorists: 43 honked; 14 are censored (*)
  - Time measured to the nearest *hundredth* of a second
  - Only 1 tie—at 1.41 seconds—the earliest recorded event time.
  - About as close to truly continuous event data as we ever get.

Table 13.1: Known and censored (*) event times for 57 motorists blocked by another automobile (reaction times are recorded to the nearest hundredth of a second)

1.41	2.12	2.54	2.83	3.14	3.56	4.18	4.71*	6.03	12.29
1.41*	2.19	2.56	2.88	3.17	3.57	4.30*	4.96	6.21*	13.18
1.51	2.36*	2.62	2.89	3.21	3.58	4.44	5.12*	6.30	17.15*
1.67	2.48	2.68	2.92	3.22	3.78	4.51	5.39	6.60*	
1.68	2.50	2.76*	2.98	3.24	4.01*	4.52	5.73	7.20	
1.86	2.53	2.78*	3.05*	3.46*	4.10	4.63*	5.88	9.59	

Notation for continuous-time event data

- **T** is a *continuous random variable* representing event time
- $\mathbf{T}_i$ indicates the event time for individual i
- $\mathbf{CENSOR}_i$ indicates whether $\mathbf{T}_i$ is censored
- t_j clocks the infinite number of instants when the event *could* occur

Defining continuous-time survivor and hazard functions

(ALDA, Sections 13.2 & 13.3, p. 472)

The survivor function

The survival probability for individual i at time t_j is the probability that his or her event time, T_i will exceed t_j

$$S(t_{ij}) = \Pr\left[T_i > t_j\right]$$

Note that this definition is essentially identical to that in discrete-time

The hazard function

Hazard assesses the *risk*—at a particular moment—that an individual who has not yet done so will experience the event

Can't be defined as a (conditional) probability because that probability $\rightarrow 0$

Instead divide time into an infinite number of vanishingly small intervals:

includes t_j ⟶ $[t_j\ ,\ t_j + \Delta t)$ ⟵ excludes $t_j + \Delta t$

Idea: Compute the probability that T_i falls in this interval as $\Delta t \rightarrow 0$ and divide by the interval width to derive an estimate of hazard per unit time:

$$h(t_{ij}) = limit\ as\ \Delta t \rightarrow 0 \left\{ \frac{Pr\left[T_i\ is\ in\ the\ interval\ [t_j, t_j + \Delta t) \mid T_i \geq t_j\right]}{\Delta t} \right\}$$

Tips for interpreting continuous-time hazard

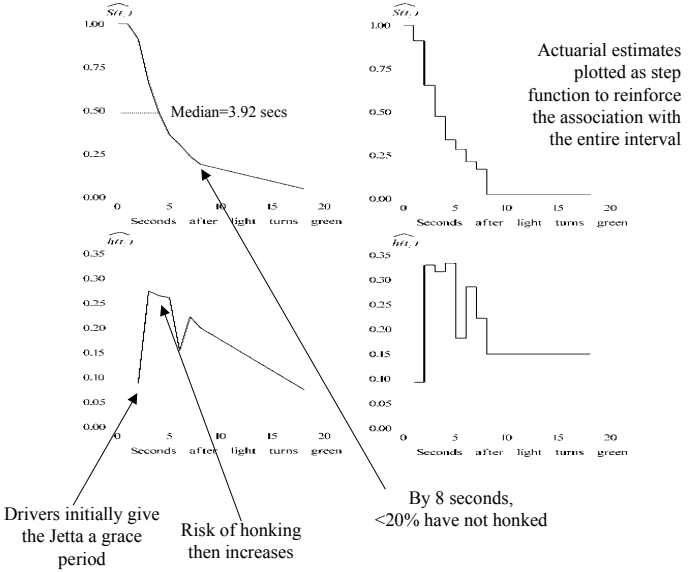
- Rate per unit of time—**it is not a probability**
- Rates must be attached to a unit of measurement—60 mph, 60K/yr
- Intuition using repeatable events—estimate the number of events in a finite period (e.g., if monthly hazard =.10, annual hazard = 1.2)
- Unlike probabilities, rates can exceed 1 (has implications for modeling—instead of modeling *logit* hazard we model *log* hazard)

Grouped estimates of the survivor and hazard functions

(ALDA, Section 13.2, p. 475, Fig 13.1, p. 479)

Discrete-time method
Group the event times into intervals and use discrete-time methods

Actuarial/Life-table method
Adapt these estimates by assuming that times are distributed randomly throughout the interval



Why categorize *continuous* data? ⟹ Kaplan Meier method

Kaplan Meier estimates of the Survivor Function

(ALDA, Section 13.3, p. 483, Table 13.3, p. 484)

Key idea: **Use observed event times to construct intervals so that *each interval contains just one event time* and then use standard discrete-time methods.**

- Since the first 3 observed event times are 1.41, 1.51 and 1.67, construct two intervals: [1.41, 1.51), [1.51, 1.67)
- By convention, construct an initial interval [0, 1.41)
- Continue through all the observed event times

Interval	[Start	End)	n at risk	n events	n censored	$\hat{p}(t)$	$\hat{S}(t)$
0	0.00	1.41	57	0	0	—	1.0000
1	1.41	1.51	57	1	1	0.0175	0.9825
2	1.51	1.67	55	1	0	0.0182	0.9646
3	1.67	1.68	54	1	0	0.0185	0.9467
4	1.68	1.86	53	1	0	0.0189	0.9289
5	1.86	2.12	52	1	0	0.0192	0.9110
6	2.12	2.19	51	1	0	0.0196	0.8931
7	2.19	2.48	50	1	1	0.0200	0.8753
8	2.48	2.50	48	1	0	0.0208	0.8570
9	2.50	2.53	47	1	0	0.0213	0.8388
10	2.53	2.54	46	1	0	0.0217	0.8206
11	2.54	2.56	45	1	0	0.0222	0.8023
12	2.56	2.62	44	1	0	0.0227	0.7841
13	2.62	2.68	43	1	0	0.0233	0.7659
14	2.68	2.83	42	1	2	0.0238	0.7476
15	2.83	2.88	39	1	0	0.0256	0.7285
16	2.88	2.89	38	1	0	0.0263	0.7093
17	2.89	2.92	37	1	0	0.0270	0.6901
18	2.92	2.98	36	1	0	0.0278	0.6710
19	2.98	3.14	35	1	1	0.0286	0.6518
...							
25	3.57	3.58	26	1	0	0.0384	0.5121
26	3.58	3.78	25	1	0	0.0400	0.4916
...							
35	4.96	5.39	14	1	1	0.0714	0.3349
36	5.39	5.73	12	1	0	0.0833	0.3070
37	5.73	6.03	11	1	1	0.0909	0.2791
38	6.03	6.30	9	1	1	0.1111	0.2481
39	6.30	7.20	7	1	1	0.1429	0.2126
40	7.20	9.59	5	1	0	0.2000	0.1701
41	9.59	12.29	4	1	0	0.2500	0.1276
42	12.29	13.18	3	1	0	0.3333	0.0851
43	13.18	∞	2	1	1	0.5000	0.0425

Est. ML
3.5759

Conditional probability of event occurrence

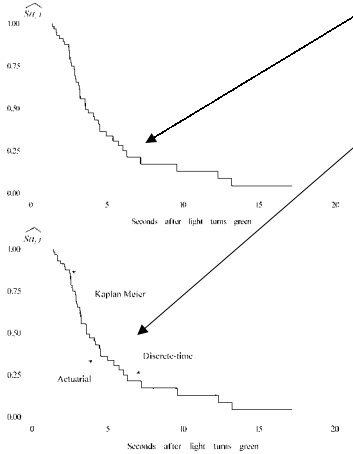
$$\hat{p}(t_j) = \frac{n\ events_j}{n\ at\ risk_j}$$

Kaplan-Meier estimate of the survivor function

$$\hat{S}(t_j) = (1 - \hat{p}(t_1))(1 - \hat{p}(t_2)) \cdots (1 - \hat{p}(t_j))$$

Kaplan Meier estimates—displays, pros and cons

(ALDA, Fig 13.2, p. 485)



Pros of KM approach

- Uses all of the observed information on event times without categorization
- If event occurrence is recorded using a truly continuous metric, the estimated survivor function appears almost 'continuous.'
- Estimates are as refined as the fine-ness of data collection—certainly finer than DT and Actuarial/Life Table approaches

Drawbacks of KM approach

- When examining plots for subgroups, the "drops" will occur in different places making visual comparison trickier
- No corresponding estimate of hazard. You can compute:

$$\hat{h}_{KM}(t_j) = \frac{\hat{p}_{KM}(t_j)}{width_j}$$

but the estimates are generally too erratic to be much direct use.

?? Question ??
Knowing the value of the hazard function during data analysis, is their any way to discern its shape over time?

Understanding the cumulative hazard function

(ALDA, Section 13.4, p. 488, Fig 13.3, p. 489)

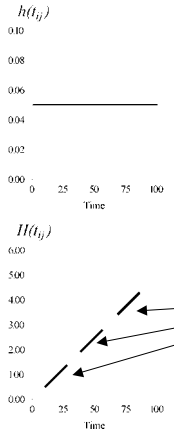
Cumulative hazard function

- Assesses the *total amount of accumulated risk* individual *i* has faced from the beginning of time (t_0) to the present (t_j)

$$H(t_{ij}) = \underset{\text{between } t_0 \text{ and } t_j}{cumulation} \left[h(t_{ij})\right],$$

- By definition, begins at 0 and rises over time (never decreasing).
- Has no directly interpretable metric, and is not a probability
- Cumulation prevents it from directly assessing unique risk (hazard)
- But, examining its changing shape allows us to deduce this information

A: Constant hazard



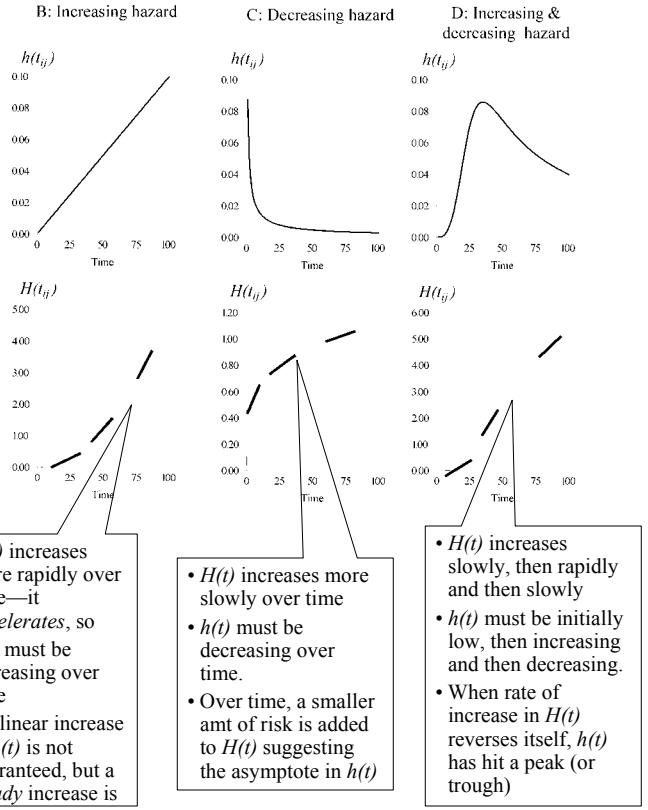
From *H(t)* to *h(t)*

To deduce the shape of *h(t)*, study how the rate of increase in *H(t) changes over time.* Any change in its rate of increase reflects a corresponding change in *h(t)*

- To develop an intuition, first move from *h(t)* to *H(t).* Because *h(t)* is constant, *H(t)* increases linearly as the same fixed amount of risk—the constant value of hazard—is added to the prior cumulative level at each successive instant (making *H(t)* linear).
- Next, move from *H(t)* to *h(t)* *because this is what you need to do in practice.*
  - Guesstimate the rate of increase in *H(t) at different points in* time
  - Because the slopes are identical, the rate of change in *H(t)* is constant over time, indicating that *h(t)* is constant over time

From cumulative hazard to hazard: Developing intuition

(ALDA, Fig 13.3, p. 489)



- *H(t)* increases more rapidly over time—it *accelerates*, so
- *h(t)* must be increasing over time
- the linear increase in *h(t)* is not guaranteed, but a *steady* increase is

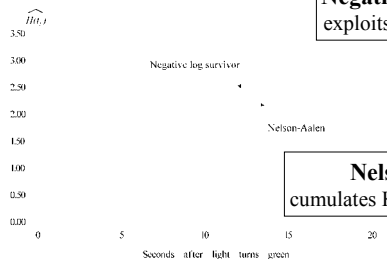
- *H(t)* increases more slowly over time
- *h(t)* must be decreasing over time.
- Over time, a smaller amt of risk is added to *H(t)* suggesting the asymptote in *h(t)*

- *H(t)* increases slowly, then rapidly and then slowly
- *h(t)* must be initially low, then increasing and then decreasing.
- When rate of increase in *H(t)* reverses itself, *h(t)* has hit a peak (or trough)

Estimating sample cumulative hazard functions

(ALDA, Section 13.4.2, p. 491, Fig 13.4, p. 493)

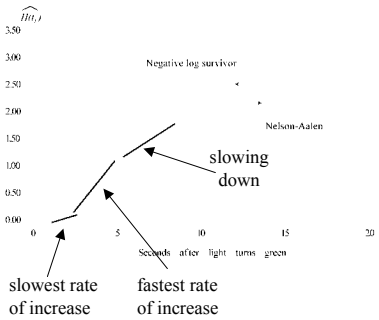
Negative log survivor method
exploits the identity $H(t_j)=-ln\ S(t_{ij})$



Nelson-Aalen method
cumulates KM-type estimates of hazard

- **Careful**: Focus on early estimates because later estimates are usually based upon relatively small risk sets

Deducing the shape of hazard from sample cumulative hazard
Compare several "guesstimated" slopes drawn on sample plots



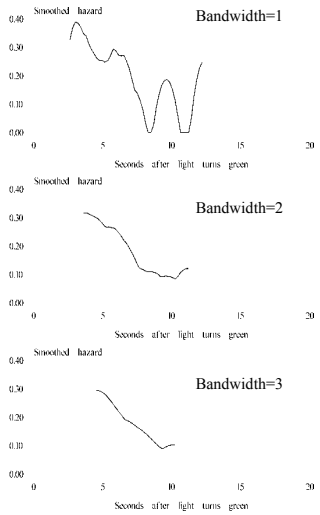
Conclusion
Hazard is initially low, increases until the 5th second, and then decreases again

Kernel-smoothed estimates of the hazard function

(ALDA, Section 13.5, p 494, Fig 13.5, p. 496

Key idea: **Estimate hazard's *average* value at many points in time by aggregating together estimates in the temporal vicinity**

- Successive differences in cumulative hazard yield *pseudo-slope* estimates of hazard
- Aggregate these estimates within a temporal window—the *bandwidth*
- Yields approximate values of hazard based upon estimates nearby



Finally
a clear window on hazard
(especially with wider bandwidths)

But, as the bandwidth widens,

- The link between the smoothed function and hazard diminishes because it is estimating hazard's average within a broader timeframe
- The estimates can only be computed at intermediate time frames (a big problem if hazard is highest initially)

Developing your data analytic intuition--overview

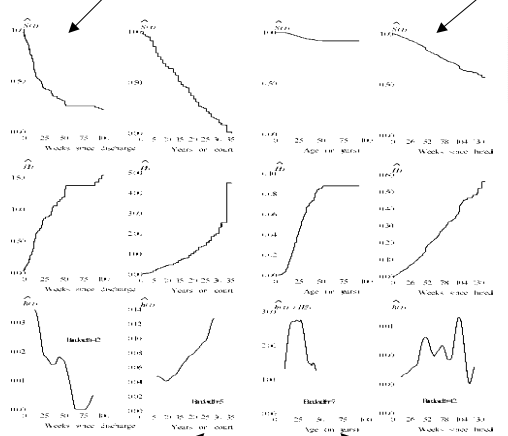
(ALDA, Section 13.6, p 497, Fig 13.6, p. 499)

To describe the distribution of continuous-time event data, plot:

- The survivor function (Kaplan-Meier estimates)—usually begin here because we can interpret its level and it can be estimated at all times
- The cumulative hazard function (either –LS or Nelson-Aalen estimates)
- If possible, kernel smoothed estimates of the hazard function

A: Time to first heavy drinking day
89 recently treated alcoholics
(Cooney et al, 1991)

D: Employment duration
2,074 health care workers
(Singer et al, 1998)



Survivor functions differ markedly reflecting differences in event frequency

Cumulative hazard functions differ markedly reflecting differences in shape of underlying hazard

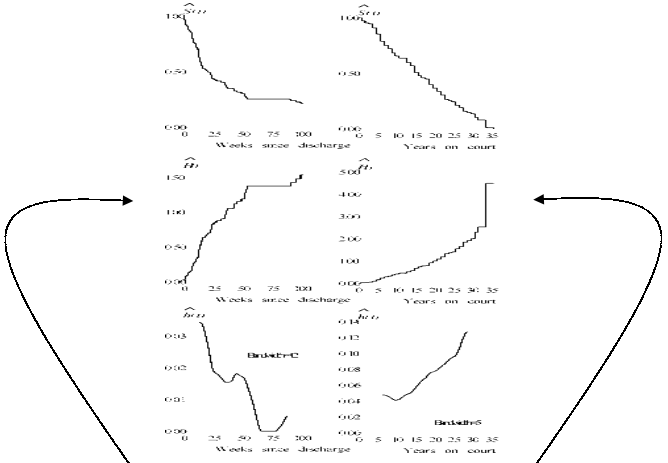
Kernel smoothed hazard functions differ markedly suggesting major differences in the shape of underlying hazard

B: Time on bench
107 US Supreme Ct justices
(Zorn & Van Winkle, 2000)

C: Age at first depressive episode
1974 adults ages 18 - 94
(Sorenson, Rutter, Aneshensel, 1991)

Developing your data analytic intuition—two examples

(ALDA, Section 13.6, p 497, Fig 13.6, p. 499)



A: Weeks to first heavy drinking day

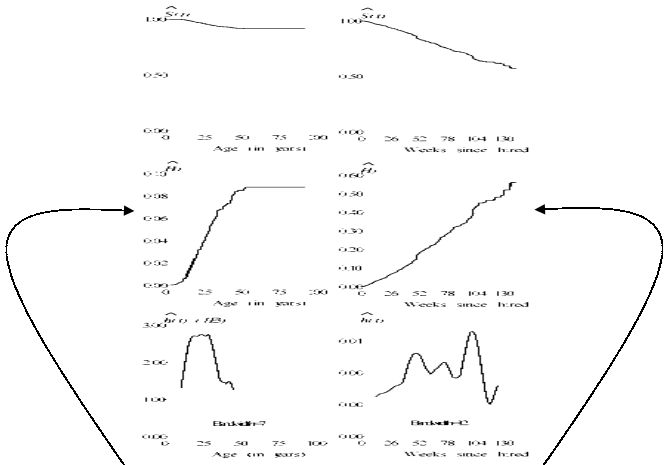
- Relapse very common: ML=22 weeks, final S(t)=.2149 (at 100 wks)
- Cumulative hazard rises sharply initially and then decelerates. Suggests great initial risk of relapse that declines over time
- Kernel-smoothed hazard shows steady decline over time (although can't comment on first 12 weeks because of bandwidth)

B: Years on the Supreme Court

- Event occurrence very common: All eventually retire or die, ML=16 years
- Cumulative hazard rises slowly initially and then accelerates ~10yrs. Suggests low immediate risk followed by steady increase in risk
- Kernel-smoothed hazard shows increasing risk over time (although can't comment on first 5 years because of bandwidth)

Developing your data analytic intuition—two more examples

(ALDA, Section 13.6, p 497, Fig 13.6, p. 499)



C: Age at first depressive episode

- Onset very rare: no ML, S(t)=.92 at age 54
- Cumulative hazard rises slowly, then sharply, then slowly. Suggests that hazard is first low, then rises to a peak, and then declines
- Kernel-smoothed hazard shows this inverted-U shape with a peak between 15 and 30

D: Weeks of employment at CMHCs

- Over 3 years, many people stay: no ML, S(t)=.57 at 139 weeks
- Cumulative hazard seems almost linear, with a few 'bumps'. Suggests relatively steady risk
- Kernel-smoothed hazard reveals pronounced peaks which correspond to anniversaries—12mos, 18mos, and 24 mos after hire.