Lecture 4

Monday, September 25, 2023 10:01

BIOS522_Slides4

EMORY ROLLINS SCHOOL OF PUBLIC HEALTH | Department of Biostatistics and Bioinformatics

BIOS 522: Survival Analysis Methods

Lecture 4:

The hazard and cumulative hazard functions

Previously

- *Introduced the survival function*
- *Defined the Kaplan-Meier estimator*
- *Calculated the log-rank test for comparing survival curves*
- *Used R to implement these procedures*

2

Survival random variable

- Non-negative random variable T
- For a given population, we may want to summarize:
  - The mean of T
  - The median of T
  - The variance of T
  - The density function (pdf) for T
  - The cumulative distribution function (CDF) for T

3

Survival random variable

- For time-to-event random variables, we are also interested in summarizing other key quantities
  - The survival function $S(t)$
    - Probability of surviving beyond time t
  - The **hazard function** $h(t)$
    - Instantaneous rate of failure among those still at risk at time t
  - The **cumulative hazard function** $H(t)$
    - The accumulated hazard from time 0 to time t

4

Hazard function

- Among those still at risk at time t, what is the instantaneous rate of failure at time t?

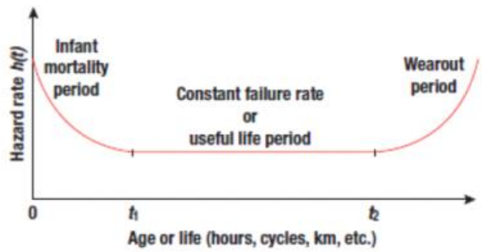
$$h(t) = \lim_{\Delta \downarrow 0} \frac{1}{\Delta} \Pr(t \le T < t + \Delta | \underbrace{T \ge t}_{\text{conditional on not having failed before time } t})$$

$$= \frac{f(t)}{S(t)} = \frac{\text{density function}}{\text{survival function}}$$

5

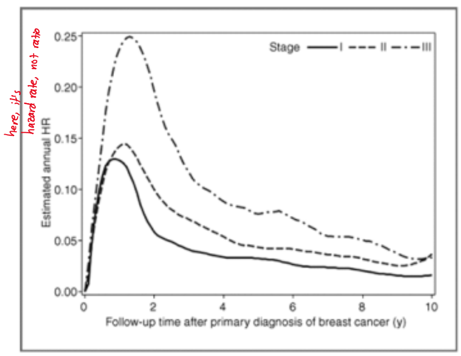
Hazard is a speedometer for risk

- Can use the hazard function to identify periods of highest risk



6

Example: Breast cancer recurrence



- Smoothed hazard functions by tumor stage for first recurrence among women after primary breast cancer treatment
- Early period of elevated risk of recurrence
- At all times, highest risk is for Stage III cancers

7

Source: Cheng et al. (2012) Cancer, Epi, Biomarkers & Prevention 10.1158/1055-9965.EPI-11-1089

Example: Seasonal mortality in wildlife


DOI: 10.1111/2041-210X.13305

APPLICATION

Methods in Ecology and Evolution

For everything there is a season: Analysing periodic mortality patterns with the `cyclomort` R package

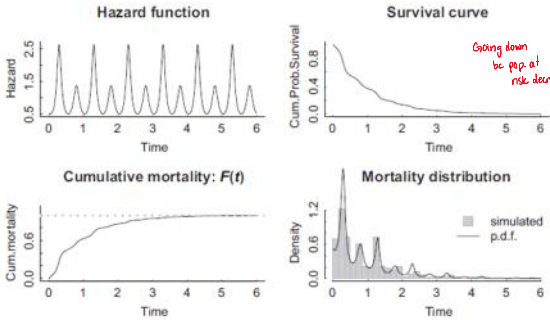
Eliezer Gurarie[1] | Peter R. Thompson[1,2] | Allicia P. Kelly[3] | Nicholas C. Larter[4] | William F. Fagan[1] | Kyle Joly[5]

- For many species, mortality risk follows a seasonal pattern
- For example, during certain times of the year, resources may be scarce, or susceptibility to predators or disease is high
- The authors create an R package to model seasonal mortality patterns

Source: Gurarie et al. (2020) Methods Ecol Evol, doi: 10.1111/2041-210X.13305

8

Example: Seasonal mortality in wildlife



Going down
bc pop. at
risk decreasing

FIGURE 2 Example of a simulated multi-seasonal periodic mortality process, outputted by the `simulate_cycloSurv()` function

Check your understanding: Why do the peaks of the pdf decline over time?

Cumulative hazard function

- How much hazard has accumulated between time 0 and time t?

$$H(t) = \int_{u=0}^{u=t} h(u)du$$

- *Has a convenient relationship to* $S(t)$

$$S(t) = e^{-H(t)}$$

Any one function fully describes the distribution…

Survival function
$S(t) = \Pr(T > t)$
$S(t) = 1 - F(t)$
$S(t) = e^{-H(t)}$

Cumulative distribution function
$F(t) = \Pr(T \leq t)$

Probability density function
$f(t) = \frac{d}{dt}F(t)$

$f(t) = h(t)S(t)$

Hazard function
$h(t) = \frac{f(t)}{S(t)}$

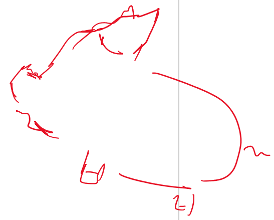
$h(t) = -\frac{d}{dt}\ln[1 - F(t)]$

$h(t) = -\frac{d}{dt}\ln S(t)$

Cumulative hazard function
$H(t) = \int_{u=0}^{u=t} h(u)du$
$H(t) = -\log S(t)$

KM is non-parametric, but :

Common parametric survival distributions



- Failure time random variable T $(T \geq 0)$

- $T \sim Exponential(\lambda)$ or $T \sim Exp(\lambda)$
- $T \sim Weibull(\lambda, \gamma)$
- $T \sim LogLogistic(\lambda, \gamma)$

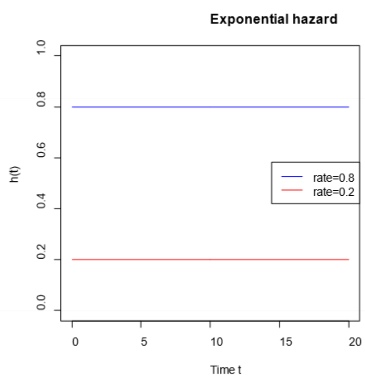
- There are other parametric survival distributions out there (e.g. gamma, Gompertz-Makeham, log-normal, generalized F, Pareto), but we won't discuss these in this course

Exponential distribution

- Constant hazard function

$$h(t) = \lambda$$

- λ is called the **rate parameter**
- For the exponential distribution, λ is also the hazard rate



higher

lower

Exponential distribution

- "Memoryless" bc risk of failure doesn't depend on the past

- It can be hard to justify the constant hazard assumption in practice

- **Examples of exponential distributions in the real world:**
  - Time until an earthquake occurs
  - Length (in minutes) of long-distance business telephone calls
  - The amount of time (in months) a car battery lasts
  - The amount of time (in minutes) a postal clerk spends with a customer

Weibull distribution

- By adding a second parameter, we allow for greater flexibility in our hazard function

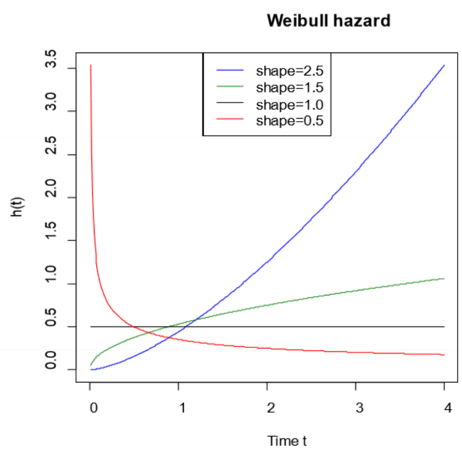
$$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$$

- λ is still called the **rate parameter**
- γ is called the **shape parameter** → second parameter alters the shape
- Note that the hazard rate for the Weibull distribution is not λ, but rather the hazard rate is calculated from λ, γ, t Hazard rate is not λ!

15

Weibull distribution

- More flexible than the exponential
- The Weibull distribution accommodates three distinct possibilities*:
  1. If something is going to fail it will most likely fail at the start
  2. The rate of failure is fairly constant
  3. Failure becomes more likely as time goes on.



*https://doi.org/10.1111/j.1740-9713.2018.01123.x

16

Log-logistic distribution

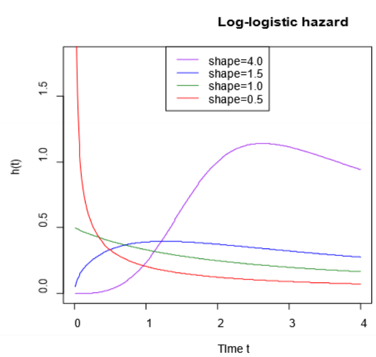
Even more complex

- Another (different) two-parameter distribution

Not a generalization of exponential, no flat line!

$$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1 + (\lambda t)^{\gamma}}$$

$$\begin{cases} \text{decreasing from } \infty, & \text{if } \gamma < 1 \\ \text{decreasing from } \lambda, & \text{if } \gamma = 1 \\ \text{increasing then decreasing}, & \text{if } \gamma > 1 \end{cases}$$



17

Distribution	Hazard Function	Cumulative Hazard Function	Survival Function
Any	$h(t)$	$H(t) = \int_0^t h(u)du$	$S(t) = e^{-H(t)}$

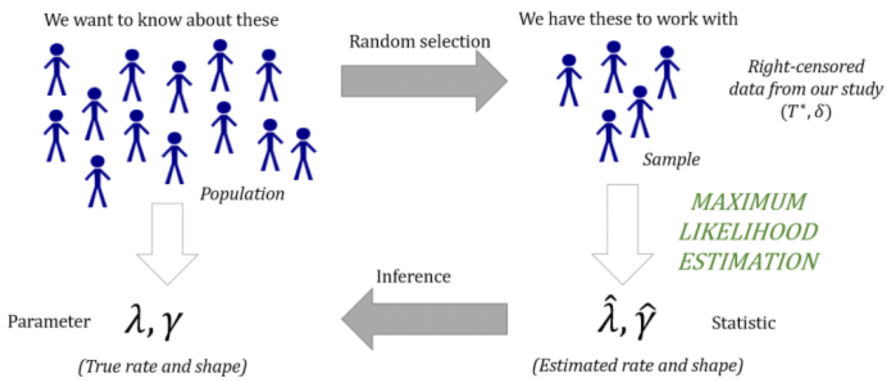
18

Distribution	Hazard Function	Cumulative Hazard Function	Survival Function
Exponential	$h(t) = \lambda$	$H(t) = \lambda t$	$S(t) = e^{-\lambda t}$
Weibull	$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$	$H(t) = (\lambda t)^{\gamma}$	$S(t) = e^{-(\lambda t)^{\gamma}}$
Log-logistic	$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$	$H(t) = \log(1 + (\lambda t)^{\gamma})$	$S(t) = \frac{1}{1+(\lambda t)^{\gamma}}$

integrate from 0 → t

19

Statistical inference

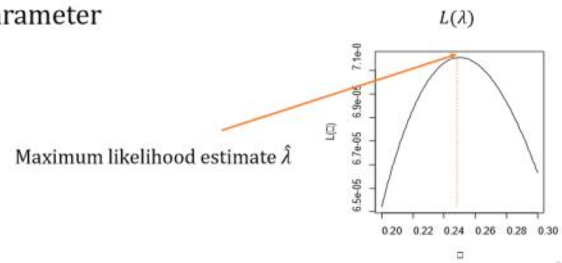


20

https://www.cliffsnotes.com/study-guides/statistics/sampling/populations-samples-parameters-and-statistics

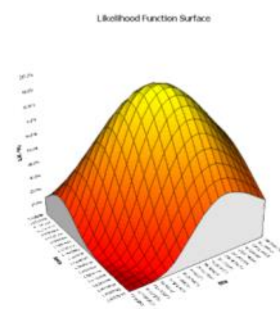
Maximum likelihood estimation

- Likelihood function $L(\lambda)$ is the probability of observing data if the true parameter is λ
- Function of the *data* and the *parameter(s)*
- Example with <u>one</u> parameter



Maximum likelihood estimation

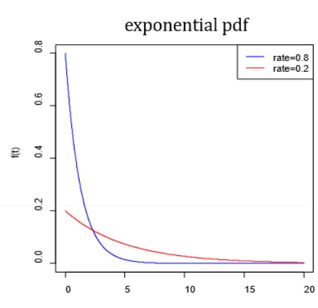
- If there is more than one parameter, then the maximum likelihood estimate is where all parameters are simultaneously maximized
- Where does $L(\lambda, \gamma)$ maximize?



Building the likelihood function

- General likelihood function for time-to-event data
- If there were NO censoring in the data:

$$L(\lambda) = \prod_{i=1}^{n} f(T_i|\lambda)$$



Likelihood for right-censored data

Observed failures
$T_i^*, \delta_i = 1$

Likelihood of failure at time T_i^* given by the pdf

$$L_i(\lambda) = f(T_i^*|\lambda) = h(T_i^*|\lambda)S(T_i^*|\lambda)$$

hazard fcn * survival fcn

Censored observations
$T_i^*, \delta_i = 0$

Likelihood of being at risk at time T_i^* given by $S(t)$

Likelihood
Survival up until time T_i^* (continued to be at risk until T_i^*)

$$L_i(\lambda) = S(T_i^*|\lambda)$$

General likelihood for individual i

$$L_i(\lambda) = [h(T_i^*|\lambda)]^{\delta_i} S(T_i^*|\lambda)$$

Likelihood for all n individuals

$$L(\lambda) = \prod_{i=1}^{n} [h(T_i^*|\lambda)]^{\delta_i} S(T_i^*|\lambda)$$

24

General Likelihood: $L(\lambda) = \prod_{i=1}^{n}[h(T_i^*|\lambda)]^{\delta_i} S(T_i^*|\lambda)$

Distribution	Hazard Function	Survival Function	Likelihood Function
Exponential	$h(t) = \lambda$	$S(t) = e^{-\lambda t}$	$L(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda T_i^*}$ (hazard fcn raised to deltas; → survival fcn)
Weibull	$h(t) = \lambda\gamma(\lambda t)^{\gamma-1}$	$S(t) = e^{-(\lambda t)^{\gamma}}$	$L(\lambda,\gamma) = \prod_{i=1}^{n} [\lambda\gamma(\lambda T_i^*)^{\gamma-1}]^{\delta_i} e^{-(\lambda T_i^*)^{\gamma}}$
Log-logistic	$h(t) = \frac{\lambda\gamma(\lambda t)^{\gamma-1}}{1+(\lambda t)^{\gamma}}$	$S(t) = \frac{1}{1+(\lambda t)^{\gamma}}$	$L(\lambda,\gamma) = ?$ (write out for in-class activity)

25

In practice…

- Most of the time we will rely on statistical software to compute the maximum likelihood estimate
- For the exponential, we can derive the maximum likelihood estimate from the likelihood function
  - The maximum likelihood estimate is the incidence rate: → will prove this in HW

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} T_i^*}$$

total # failures / total amount of person-time

26

Another nonparametric estimator of $S(t)$

e.g.

like reimann sum

break time into intervals

- With a nonparametric estimator of $H(t)$, we can estimate $S(t)$

$$H(t) = \int_{u=0}^{u=t} h(u)du$$

- What if we broke time into small intervals, and assumed that the hazard is constant within each interval?

27

Interval

- Use the same time intervals as the Kaplan-Meier estimator
- Within each interval, we assume the hazard rate is constant and estimate the incidence rate

$$\frac{\text{\# of events in the interval}}{\text{person-time follow-up in the interval}} = \frac{d_j}{n_j(t_j - t_{j-1})}$$

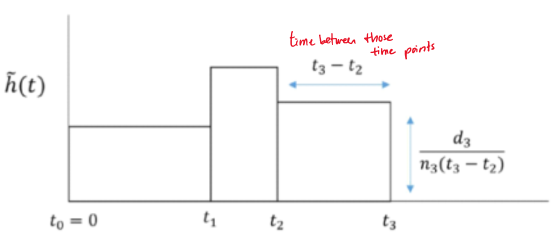
→ failures

ppl at risk

→ one interval

28

Nelson-Aalen estimator



The estimated cumulative hazard is:

$$\tilde{H}(t) = \sum_{j:t_j \le t} \frac{d_j}{n_j}$$

Each rectangle has area:

$$\frac{d_j}{\cancel{n_j(t_j - t_{j-1})}} \times \cancel{(t_j - t_{j-1})} = \frac{d_j}{n_j}$$

29

Example: AIDS hemophiliac cohort

Ordered follow-up times: 2, 3+, 6, 6, 8, 10+, 15, 15, 16, 27, 30, 32 months

Unique failure/ censoring time	Number at risk n_j during $(t_{j-1}, t_j]$	Number of deaths d_j at t_j	Number censored c_j at t_j	Cumulative hazard contribution $\frac{d_j}{n_j}$	Nelson-Aalen estimate
$t_0 = 0$					$t = [0,2)$ $\tilde{H}(t) = 0$
$t_1 = 2$	$n_1 = 12$	$d_1 = 1$	$c_1 = 0$	$\frac{d_1}{n_1} = \frac{1}{12}$	$t = [2,3)$ $\tilde{H}(t) = 0.083$
$t_2 = 3$	$n_2 = 11$	$d_2 = 0$	$c_2 = 1$	$\frac{d_2}{n_2} = \frac{0}{11}$	$t = [3,6)$ $\tilde{H}(t) = 0.083$
$t_3 = 6$	$n_3 = 10$	$d_3 = 2$	$c_3 = 0$	$\frac{d_3}{n_3} = \frac{2}{10}$	$t = [6,8)$ $\tilde{H}(t) = 0.283$

b.c. this is before anyone fails
Cannot look to the future

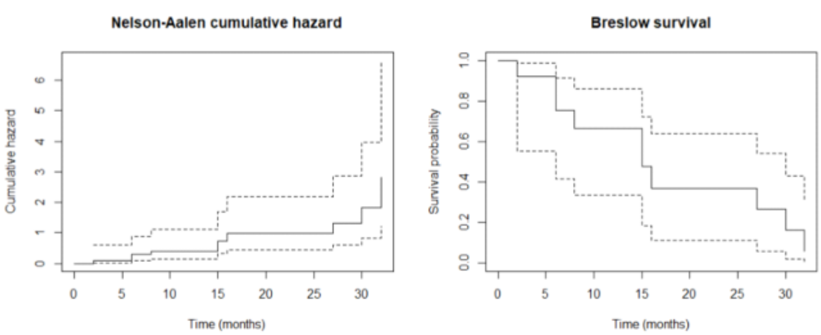
Breslow estimator

- Use the Nelson-Aalen estimator to estimate survival

$$\tilde{S}(t) = \exp\left(-\tilde{H}(t)\right)$$

- Produces an estimate of survival that is similar to, though not identical to, the Kaplan-Meier estimator

Nelson-Aalen curve example



Point estimation

$g^{-1}(y) = e^{-y}$

Nelson-Aalen estimator $\widetilde{H}(t)$

Breslow estimator $\tilde{S}(t)$

$g(x) = -\log(x)$

33

Confidence intervals

$g^{-1}(y) = e^{-y}$

Wald interval for $H(t)$

$$\widetilde{H}(t) \pm 1.96\sqrt{\widehat{Var}\left(\widetilde{H}(t)\right)}$$

Log-transformed interval for $S(t)$

$$\tilde{S}(t)\exp\left(\mp 1.96\sqrt{\widehat{Var}\left(\widetilde{H}(t)\right)}\right)$$

$g(x) = -\log(x)$

34

Confidence intervals, continued

$g^{-1}(y) = e^{y}$ $\qquad$ $g^{-1}(y) = e^{-y}$

Wald interval for $\log(H(t))$

$$\log\widetilde{H}(t) \pm 1.96\sqrt{\widehat{Var}\left(\log\widetilde{H}(t)\right)}$$

Log-transformed interval for $H(t)$

$$\widetilde{H}(t)\exp\left(\frac{\pm 1.96\sqrt{\widehat{Var}\left(\widetilde{H}(t)\right)}}{\widetilde{H}(t)}\right)$$

Log-log transformed interval for $S(t)$

$$\tilde{S}(t)^{\exp\left(\mp 1.96\sqrt{\widehat{Var}\left(\widetilde{H}(t)\right)}/\widetilde{H}(t)\right)}$$

→ how can we connect this to 36

$g(x) = \log(x)$ $\qquad$ $g(x) = -\log(x)$

<u>Complementary log-log</u>: $\log(-\log(x))$

35

Example: Prostate cancer

The NEW ENGLAND JOURNAL of MEDICINE

Prostate-Cancer Mortality at 11 Years of Follow-up

- Goal: Investigators conducted a large **randomized** study to examine the value of monitoring **prostate-specific antigen (PSA)** as part of routine screening for reducing prostate-cancer mortality.

- Population: The study included **162,388 men** between the ages of 55 and 69 years at entry randomized to receive either **PSA-based screening or standard screening** (control group). The trial was conducted in eight European countries.

Source: Schroder et al. (2012) NEJM 10.1056/nejmoa1113135

36

Example: Prostate cancer

- Outcome variable: The primary outcome was **date of prostate-cancer mortality**, assessed using national registries to identify the official cause of death in participants with prostate cancer diagnosis. The time origin was time of randomization.
- Predictor variables: For the primary analysis, the only predictor variable considered was screening arm (PSA-based screening or control).
- Statistical analysis: Researchers used the **Nelson-Aalen method** to calculate the cumulative hazard of death from prostate cancer.

37

Example: Prostate cancer

- Results: Figure 2 summarizes the Nelson-Aalen cumulative hazard curve. These two curves begin to gradually separate starting approximately 7 years after randomization. Authors note that there is evidence that PSA-based screening significantly reduced mortality from prostate cancer but did not affect all-cause mortality (*not shown in figure*).

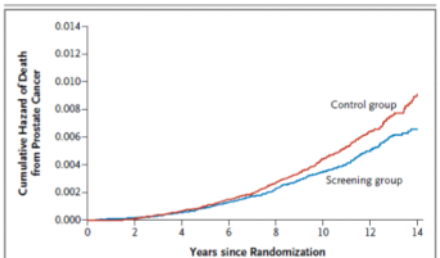


Figure 2. Cumulative Hazard of Death from Prostate Cancer among Men 55 to 69 Years of Age.
Values are not included for centers in France because of the short follow-up period (median, 4.6 years). The Nelson–Aalen method was used to calculate the cumulative hazard of death from prostate cancer.

38

Looking ahead

- The hazard function is the basis of the **Cox proportional hazards model** – the most popular regression model for time-to-event data
- Regression models allow us to model the effects of multiple covariates simultaneously, including continuous covariates
- Next week will be the first of several weeks on the Cox proportional hazards regression model

39

Today's activity

- Small groups
- Wordle word problem!
- More transformed confidence intervals
- Sketching hazard functions

40