PH 716 Applied Survival Analysis

Part 1: Basics of Survial Analysis

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/02/03 20:35:57

Recall the workflow of statistical inference (making decision based on data)

1. Prior to statistical inference
   - Research gap
   - Research question
   - Research hypotheses
2. Collect data (realizations of random variables of your interest)
3. Assume a statistical model for dada
4. Translate research hypotheses into statistical terms
5. Hypothesis testing
6. Make decision

Why do we need survival analysis?

Many scientific studies ask questions not just about whether an event occurs, but when it occurs.

Examples:

- How long do patients remain progression-free after treatment?
- How long does it take for smokers to relapse?
- How long does a device function before failure?

In all of these settings:

- the outcome is time to an event, and
- the event may not be observed for everyone during the study period.

These features require specialized statistical tools.

Motivating examples for survial analysis

([DM] Example 1.2) Cancer clinical trial (time to progression)

This is a Phase II clinical trial of Xeloda and oxaliplatin (XELOX) chemotherapy given before surgery to 48 advanced gastric cancer patients with paraaortic lymph node metastasis (Wang et al., 2014). The outcome of interest is the progression-free survival, which is the time from entry into the clinical trial until progression or death, whichever comes first.

```
head(asaur::gastricXelox)
```

```
##   timeWeeks delta
## 1         4     1
```

```
## 2         8     1
## 3         8     1
## 4         8     1
## 5         9     1
## 6        11     1
```

Scientific questions that may be of interest include:

- How long do patients typically remain progression-free?
- What fraction of patients remain progression-free beyond a given time?
- How should we handle patients whose disease has not progressed by the end of follow-up?

([DM] Example 1.5) Smoking cessation trial (time to relapse)

The purpose of Steinberg et al. (2009) was to evaluate extended duration of a triple-medication combination versus therapy with the nicotine patch alone in smokers with medical illnesses. Here, the event of interest is relapse. Of course, not all participants relapse during follow-up.

```
head(asaur::pharmacoSmoking[, 2:8])
```

```
##   ttr relapse         grp age gender     race employment
## 1 182       0   patchOnly  36   Male    white         ft
## 2  14       1   patchOnly  41   Male    white      other
## 3   5       1 combination  25 Female    white      other
## 4  16       1 combination  54   Male    white         ft
## 5   0       1 combination  45   Male    white      other
## 6 182       0 combination  43   Male hispanic         ft
```

Scientific questions that may be of interest include:

- How does relapse risk evolve over time?
- Do different treatments change the timing of relapse?
- How do we incorporate incomplete follow-up?

What counts as an "event"?

Depending on the application, the event of interest may be:

- death, cancer incidence, disease remission
- HIV infection, AIDS onset
- organ rejection, transplant failure
- bankruptcy, economic recovery

What unifies these problems is time-to-event structure, not the nature of the event itself.

What makes time-to-event data different?

Two key features distinguish survival data from standard outcomes:

- Time is non-negative
  - Measured from a well-defined time origin (e.g., diagnosis, treatment initiation)
  - May be measured in various units (e.g., days, months, years)
- Incomplete observation (censoring): for some subjects, the event has not occurred by the end of follow-up.
  - We only know that their event time exceeds a certain value (the censoring time)
  - Ignoring censoring can lead to biased results

These features motivate the core quantities of survival analysis.

Random variable (RV) of interest

- T (≥ 0): survival time or, more specifically, the time elapsed from time 0 to the event of interest (abbr. time to event); the RV of interest in survival analysis
  - The time origin, 0, often represents
    - the time point when a disease was first detected,
    - the initiation time of therapy/procedure, or
    - certain meaningful time point
  - E.g.,
    - In a clinical trial: time 0 might be the time of randomization;
    - Among heart transplant patients: time 0 is the end of transplantation
- As an RV, T has a distribution that can be characterized by various functions
  - Probability density function (pdf) or probability mass function (pmf)
  - Cumulative distribution function (cdf)
  - Survival function
  - Hazard function
  - Cumulative hazard function

pdf, pmf and cdf of T

- The cumulative distribution function (cdf) of T, say F_T, is defined as

$$F_T(t) = \Pr(T \leq t), \quad t \in \mathbb{R}.$$

  - Interpretation: How likely for the event occur before or at time t
- If T is a discrete RV, then the pmf of T is defined as

$$p_T(t) = \Pr(T = t)$$

  - Interpretation: How likely for the event occur at time t
- If T is a continuous RV, then its pdf is defined as

$$f_T(t) = \mathrm{d}F_T(t)/\mathrm{d}t$$

  - Interpretation: The instantanous likelihood of the event at time t

Survival function: "how likely for a subject remains event-free up to time t"

The survival function is defined as

$$S_T(t) = \Pr(T > t) = 1 - F_T(t)$$

- Interpretation: How likely for the event to NOT occur by time t

Hazard function (aka. hazard rate): "how risky the subject is at time t?"

The hazard function is denoted by $\lambda_T(t)$, defined differently for continuous and discrete T:

- If T is continuous with pdf f_T
  - $\lambda_T(t) = f_T(t)/S_T(t)$
  - Conditional density (not a probability), given that the event has not yet occurred prior to time t
- If T is discrete with pmf p_T
  - $\lambda_T(t_j) = \Pr(T = t_j \mid T \geq t_j) = p_T(t_j)/S_T(t_{j-1})$ and zero elsewhere, $j = 1, 2, \ldots$.
    - Conditional failure probability, given that the event has not yet occurred prior to time t

Interpretation: the instantaneous risk of experiencing the event of interest at time t, assuming the subject has survived up to t.

Property: $\lambda_T(t) \geq 0$ for all t

- Ex. 1.1 ([KM] Ex. 2.3): Suppose that T has pmf $p_T(t_j) = 1/3$, $j = 1, 2, 3$. Find the survival and hazard functions.
- Ans. to Ex. 1.1: The survival function is

$$S_T(t) = \Pr(T > t) = \sum_{j:t_j > t} p_T(t_j) = \begin{cases} 1 & \text{if } 0 \leq t < t_1 \\ 2/3 & \text{if } t_1 \leq t < t_2 \\ 1/3 & \text{if } t_2 \leq t < t_3 \\ 0 & \text{if } t \geq t_3. \end{cases}$$

  The hazard function is then

$$\lambda(t_j) = \frac{p_T(t_j)}{S_T(t_{j-1})} = \begin{cases} 1/3 & j = 1 \\ 1/2 & j = 2 \\ 1 & j = 3 \\ 0 & \text{elsewhere.} \end{cases}$$

Cumulative hazard: "How much total risk has accumulated up to time t?"

The cumulative hazard is denoted by $\Lambda_T(t)$, defined differently for continuous and discrete T:

- $\Lambda_T(t) = \int_0^t \lambda_T(s)\mathrm{d}s$ if T is continuous
- $\Lambda_T(t) = \sum_{j:t_j \leq t} \lambda_T(t_j)$ if T is discrete

Interpretation: the integrated risk of the event occurring from time 0 to t

---

- Ex. 1.2 (exponential distribution): Suppose that T has pdf $f_T(t) = \lambda \exp(-\lambda t)$, $\lambda > 0$. Find the survival, hazard and cumulative hazard functions.
  - Memorylessness: constant hazard rate $\Rightarrow \Pr(T > t + t_0 \mid T > t_0) = \Pr(T > t)$ i.e., the survival probability would not be influenced by the history
  - This is the only continuous distribution with constant hazard rate
  - Not often used to model human subjects unless followed for relatively short duration; more often used in engineering or for illustrative purposes.
- Ans. to Ex. 1.2: $F_T(t) = 1 - \exp(-\lambda t), t > 0 \Rightarrow S_T(t) = \exp(-\lambda t)$ if $t > 0$ and 1 otherwise. So, $\lambda_T(t) = \lambda$ and $\Lambda_T(t) = \lambda t$.

- Ex. 1.3 (Weibull distribution): Find survival, hazard and cumulative hazard functions for the following three equivalent parameterizations of Weibull pdf and cdf: for $t \geq 0$,
  - $f_T(t) = (k/\lambda)(t/\lambda)^{k-1}e^{-(t/\lambda)^k}$ and $F_T(t) = 1 - \exp\{-(t/\lambda)^k\}$ (used in `R` with k as "shape" and λ as "scale")
  - $f_T(t) = bkt^{k-1}e^{-bt^k}$ and $F_T(t) = 1 - \exp(-bt^k)$ (i.e., $b = \lambda^{-k}$)
  - $f_T(t) = \beta k(\beta t)^{k-1}e^{-(\beta t)^k}$ and $F_T(t) = 1 - \exp\{-(\beta t)^k\}$ (i.e., $\beta = \lambda^{-1}$)
- For the Weibull distribution with any of the three parameterizations
  - $k > 1$: monotone increasing hazard rate
  - $k < 1$: monotone decreasing hazard rate
  - $k = 1$: reducing to the exponential distribution

Linking survival, hazard, and cumulative hazard functions

- When T is continuous
  - $\lambda_T(t) = -\mathrm{d}\{\ln S_T(t)\}/\mathrm{d}t$ (Why?)
    - Proof: $\lambda_T(t) = f_T(t)/S_T(t) = F_T'(t)/S_T(t) = \frac{\mathrm{d}\{1-S_T(t)\}}{\mathrm{d}t}\frac{1}{S_T(t)} = -\mathrm{d}\{\ln S_T(t)\}/\mathrm{d}t$
      - The last equality holds due to the chain rule
  - $\Lambda_T(t) = -\ln S_T(t)$
  - $S_T(t) = \exp\{-\Lambda_T(t)\}$
  - $f_T(t) = S_T(t)\lambda_T(t)$
- When T is discrete with potential values $0(= t_0) < t_1 < t_2 < \cdots$
  - $\lambda_T(t_j) = 1 - S_T(t_j)/S_T(t_{j-1})$, $j = 1, 2, \ldots$ (Why?)
  - If $t_J \le t < t_{J+1}$, then $S_T(t) = \prod_{j:t_j \le t}\{1 - \lambda_T(t_j)\} = \prod_{j=1}^{J}\{1 - \lambda_T(t_j)\}$ (Why?)
    - Noting that $1 - \lambda_T(t_j) = S(t_j)/S_T(t_{j-1})$
- Any one of f_T, S_T, λ_T and Λ_T fully determines the distribution of T.

Other meaningful summaries of the distribution of T

Mean residual lifetime (only for continuous T)

$$r(t) = \mathrm{E}(T - t \mid T \ge t) = \int_t^\infty (u - t) f_T(u)\mathrm{d}u / S_T(t) = \int_t^\infty S_T(u)\mathrm{d}u / S_T(t)$$

- Interpretation: the expected remaining lifetime for an individual alive at time t
- Proof: $\int_t^\infty (u-t) f_T(u)\mathrm{d}u/S_T(t) = \int_t^\infty (\int_t^u \mathrm{d}v)\mathrm{d}F_T(u)/S_T(t) \overset{\text{Tonelli's theorem}}{=} \int_t^\infty (\int_v^\infty \mathrm{d}F_T(u))\mathrm{d}v/S_T(t) = \int_t^\infty \Pr(T > v)\mathrm{d}v/S_T(t) = \int_t^\infty S_T(v)\mathrm{d}v/S_T(t)$

Mean lifetime (or life expectancy, i.e., $\mathrm{E}(T)$) is defined as

- (For continuous T) $\mathrm{E}(T) = \int_0^\infty t f_T(t)\mathrm{d}t = \int_0^\infty S_T(t)\mathrm{d}t = r(0)$
- (For discrete T with potential event times $0 < t_1 < t_2 < \cdots$) $\mathrm{E}(T) = \sum_{j=1}^\infty t_j p_T(t_j)$

Median lifetime: the smallest t such that $S_T(t) \le .5$

- Interpretation: the earliest time at which the survival probability falling below .5

Censoring

- T is potentially unobserved in full for all the subjects
  - Subjects whose survival times are unobserved during follow-up are said to be censored.
- Cause of censoring
  - Administrative censoring
    - Study ends before the event of interest has occurred
    - Usually independent of survival time
  - Withdrawal from study
    - E.g.,
      - a patient drops out of clinical trial because he/she is too sick to participate
      - a subject discontinues participation in trial because her symptoms have subsided
    - Depending on the censoring
- Types of censoring
  - Right-censoring (most common and by default here):
    - For a censored subject, T is not known exactly, but $\ge$ the censoring time;
    - i.e., there is a lower bound for the event time.
  - Left-censoring
    - For a censored subject there is a upper bound for the event time
    - E.g.,

            - ([DM] Example 3.4) "In early childhood learning centers, interest often focuses upon testing children to determine when a child learns to accomplish certain specified tasks. The age at which a child learns the task would be considered the time-to-event. Often, some children can already perform the task when they start in the study. Such event times are considered left censored."
    - Interval-censoring
        * For a censored subject there are both lower and upper bounds for the event time
        * E.g.,
            - ([DM] Example 3.5) "In the Framingham Heart Study, the ages of first occurrence of the subcategory angina pectoris may be known only to be between two clinical examinations, approximately two years apart (Odell et al., 1992). Such observations would be interval-censored."
- Independent censoring (by default here)
    - The censoring time is independent of survival time