Introduction to Survival Analysis

David M. Rocke

September 25, 2025

STA/BST 222

Course Information

Class Meetings:	TR 2:10pm–3:30pm, 115 Wellman Hall
Lab:	TR 3:40pm–4:00pm, 115 Wellman Hall
Office Hours:	By appointment, in person or by Zoom 140B Med Sci 1C
Contact Information	cell: (530) 304-1019 email: dmrocke@ucdavis.edu web site: dmrocke.ucdavis.edu/Class/BST222.2025.Fall/BST222-Fall-2025.html Canvas: BST 222
TA:	Chen Le (czle@ucdavis.edu)
Prerequisites:	Statistical theory courses such as STA 231AB, STA 200ABC, or STA 131ABC

Texts and Software

Required:	**Survival Analysis : Techniques for Censored and Truncated Data.**
	Klein, John P. and Moeschberger, Melvin L., Springer, 2005.
Optional:	**Statistical Analysis of Failure Time Data, Second Edition.**
	John Kalbfleisch and Ross Prentice, Wiley, 2002.
	Most useful for parametric survival analysis.
	Applied Survival Analysis Using R.
	Dirk Moore, Springer, 2016.
Software:	Example analyses will be in R.
	You may use any software capable of the analyses,
	but R is suggested.
Data:	Data sets from KM and Moore are
	in the R packages `KMsurv` and `asaur`.

Grades

- Grades will be based on homework, exams, class attendence, and possibly projects.
- Class attendence is required.
- Each student will be expected to produce their own homework and/or projects, though discussion among students is allowed, even encouraged.
- For doctoral students, material from this course will also be tested during the preliminary qualifying exam.

- Lecture slides will be posted on my web site, but lectures will not be recorded. Class attendance is required.
- Homework assignments and other useful documents will also be posted on my web site:

  `dmrocke.ucdavis.edu/Class/BST222.2025.Fall/BST222-Fall-2025.html`
- A class e-mail list `bst-sta222-f25@ucdavis.edu` will be used for some communications. This will use your official UCD email address.
- Homework and other grades will be posted on the Canvas site `BST 222 001 FQ 2025`.

Time to Event Data

- *Survival Analysis* is a term for analyzing time-to-event data.
- This is used in clinical and epidemiological studies, where the event is often death or incidence or recurrence of disease.
- It is used in engineering reliability analysis, where the event is often failure of a device or system.
- It is used in insurance, particularly life insurance, where the event is death, disability, or damage from an accident.

Time to Event Data

- The distribution of 'failure' times is usually asymmetric and can be long-tailed.
- The base distribution is not normal, but exponential. This is the simplest distribution to model failure time data.
- There are often *censored* or *truncated* observations, which are ones in which the failure time is not observed.
- Typically, this is because the failure has not yet happened, though there are other patterns.

Time to Event Data

- Often, these are *right-censored*, meaning that we know that the event occurred after some known time t, but we don't know the actual event time, as when a patient is still alive at the end of the study.
- Observations can also be *left-censored*, meaning we know the event has already happened at time t, or *interval-censored*, meaning that we only know that the event happened between times t_1 and t_2.
- Analysis is difficult if censoring is associated with treatment or other predictors of the event in question.

Right Censoring

- Patients are in a clinical trial for cancer, some on a new treatment and some on standard of care.
- Some patients in each group have died by the end of the study. We know the survival time (measured for example from time of diagnosis—each person on their own clock).
- Patients still alive at the end of the study are right censored.
- Patients who are lost to follow-up or withdraw from the study may be right-censored.

Left and Interval Censoring

- An individual tests positive for HIV.
- If the event is infection with HIV, then we only know that it has occurred before the testing time t, so this is left censored.
- If an individual has a negative HIV test at time t_1 and a positive HIV test at time t_2, then the infection event is interval censored.

Censoring vs. Truncation

- Suppose that all individuals in the population or sample of interest are observed, and for some subjects we know the exact time t of the event, and for some subjects we have a time t, but the exact time is not known, only that it occurred before time t. Then some of the the data are left censored at t.
- If individuals in the population are not observed or even known to exist if the event occurs before time t, then the data are said to be left truncated at t.
- This means that the potentially observed population is not the same as the actual population.

Censoring vs. Truncation

- Truncation is a form of selection bias and like other sources of selection bias one can account for it in the analysis to the extent possible, or redefine the population of interest. This requires careful thought beforehand.
- If UC Davis runs a clinical trial of patients diagnosed with stage 3 colon cancer at a regularly scheduled colonoscopy, then strictly speaking the population from which these patients come is *not* the set of people with stage 3 colon cancer.

Censoring vs. Truncation

- The population here is that of patients who get regularly scheduled colonoscopies and who have stage 3 colon cancer at the time of the examination.
- This group may differ from the general population of the same age in income, occupation, race/ethnicity, and other factors which might influence the course of the disease.

Censoring vs. Truncation

- The *internal validity* of comparisons within the clinical trial between a new treatment and standard of care is not compromised.
- But the *external validity* of the study, the population that it can be generalized to, is compromised since all the patients are those who routinely receive colonoscopies.
- When patients in a study are selected by the time of the event, where patients with events occurring at an early age are implicitly excluded, both problems can occur.

Censoring vs. Truncation

- Suppose that patients are selected for a clinical trial of a new cancer therapy intended to prevent recurrence from those who have been on standard of care for one year after diagnosis and have not had a recurrence.
- Then the population of interest perforce excludes those who had died or recurred before one year, and thus may exclude the sickest patients.

Mechanisms of Censoring

- An important assumption of survival analysis methods is that censoring is not related to survival time conditional on the treatment and covariates; that is, with the risk of the event. This assumption is crucial but hard to check.
- Both the survival process and the censoring process require a clock. Time 0 may be birth, the beginning of the study, or the entry of an individual subject into the study.
- The progression of time in the statistical model may not be in physical time units, but in some related measure.

Mechanisms of Censoring

- **Type I Censoring** is when the censoring time of all subjects is fixed and known in advance, such as the known end of an animal study when the study start time is the same for all animals.
- We might study mice with explanted human tumor tissue. Six mice are controls and six receive a form of chemotherapy. At 12 weeks, mice that have not died are sacrificed so the time at which they would have died is unknown.
- All censored observations are censored at 12 weeks.

Mechanisms of Censoring

- If the censoring time is fixed in advance for each animal, but may differ between animals, then this is called **Progressive Type I Censoring**.
- We might have the same design as the previous slide, but at 6 weeks a pre-specified half of each group of mice is sacrificed to obtain pathology tissues at that time.
- Some of the censored observations are censored at 6 weeks and some at 12 weeks, and the possible censoring time is fixed for each animal at the start of the study.

Mechanisms of Censoring

- **Generalized Type I Censoring** is when the censoring time is the end of the study and is known in advance, but subjects enter the study at different times.
- Usually, we define time 0 as the time of entry for each subject, so that t is the time on study which may be different at the same physical time for different subjects.
- This might happen in human clinical trials with a known end date.

Mechanisms of Censoring

- **Type II Censoring** is when the end of the study is determined by the accumulation of a fixed number of events. For example, this is the case if a study using 100 mice ends at the death of the 25th mouse.
- This is often used in engineering reliability analysis in which one may not want to wait until the last unit on test has failed, but it is important for analysis that at least some failures have occurred.

Mechanisms of Censoring

- Censoring can also be the result of a random process which generally should be independent of the event process. For example, patients may drop out of a clinical trial or be otherwise lost to followup.
- The chance of censoring from this random process should be statistically independent from the process leading to the main event (e.g., death), conditional on the treatment or other variables used to predict the event.
- Complications can ensue if this is not true. We call this *noninformative censoring*

Mechanisms of Censoring

A more formal way to represent random censoring is that, for each subject i, there are two associated random variables, T_i, the event time for subject i and C_i, the random censoring time. We observe only the smaller of the two times since each subject either has an event or is censored. If Z_i is a collection of potential predictors of the event time, then the censoring is non-informative iff C_i is independent of T_i conditional on Z_i.

Engineering Reliability Example

- Engineering reliability studies often use parametric survival models, the simplest of which is the exponential distribution.
- The following example is based on information provide by Seagate about one of their disk drive models in terms of likelihood of failure.
- A common statistics given is MTBF = mean time between failures, which is equal to the mean lifetime under the exponential distribution.

Computer Disk Drives

Here is an example excerpt from a Product Manual, in this case for the Seagate Barracuda ES.2 Near-Line Serial ATA drive:

The product shall achieve an Annualized Failure Rate (AFR) of 0.73% (Mean Time Between Failures (MTBF) of 1.2 Million hrs) when operated in an environment that ensures the HDA case temperatures do not exceed 40° C. Operation at case temperatures outside the specifications in Section 2.9 may increase the product Annualized Failure Rate (decrease MTBF).

AFR and MTBF are population statistics that are not relevant to individual units.
AFR and MTBF specifications are based on the following assumptions for business critical storage system environments:

- *8,760 power-on-hours per year.*
- *250 average motor start/stop cycles per year.*
- *Operations at nominal voltages.*

Systems will provide adequate cooling to ensure the case temperatures do not exceed 40° C. Temperatures outside the specifications in Section 2.9 will increase the product AFR and decrease MTBF.

Computer Disk Drives

- 1.2 million hours at 8,760/hours per year (365×24) is 137 years! The exponential parameter in years is $1/137 = 0.0073$.
- How can this be tested!
- Assuming exponential failures, the average time in years until the first failure out of n units is $137/n$ and an estimate of the exponential parameter is n times the first failure time.

Computer Disk Drives

- With 1000 disk drives, the mean waiting time would be 0.137 years or less than 2 months.
- To find the mean time until failure k, we need to use the gamma distribution. The chance of 4 or more failures in 6 months is about 0.5.
- Accelerated failure time methods vs. temperature is more feasible (later).
- These figures are not credible to anyone who has ever had a disk drive!

Poisson Process

- A *Poisson Process* places random points in a continuous space, usually $\mathbb{R}, \mathbb{R}^2, \mathbb{R}^3$ or a subset thereof.
- The *complete independence* property of a Poisson process is that for any pre-specified collection of disjoint, bounded subregions of the space, the random variables indicating the number of points in each subregion are statistically independent.
- For a *homogeneous Poisson process*, the probability that there are n points in a region depends only on the measure (length, area, volume) of the region.

Poisson Process

- For survival analysis, we are interested in a counting process on the positive real line. This can be defined as a random process that generates a random function $N(t), t \geq 0$, defined as the number of points in the interval $(0, t]$.
- For a homogeneous Poisson process, there is a parameter λ such that the mathematical expectation of the number of points in an interval of length t is λt.
- The probability mass function for the number of points n in an interval of length t is

$$f(n; \lambda t) = \frac{(\lambda t)^n e^{-\lambda t}}{n!}$$

which is called the *Poisson distribution*.

Exponential Distribution

If the points on the line are generated by a homogeneous Poisson process with parameter λ and $t_0 \geq 0$ is any pre-chosen point on the line, then the distance between t_0 and the point of the process that has the smallest distance forward from t_0 has a distance x defined by the exponential density

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

This is the *waiting time* until the next event.

Gamma Distribution

- In cases where more than one event can happen, and where the distribution of the time to the next event does not change when an event occurs, we can measure the time to the k^{th} event forward from a particular time. That waiting time has a gamma distribution defined by

$$f(x;\lambda,k) = \frac{x^{k-1}e^{-\lambda x}\lambda^k}{\Gamma(k)}$$

- For R, using the gamma distribution, scale $= \lambda$ and shape $= k$.

Disk Drive Failure

- If the failures of a particular type of disk drive form a homogeneous Poisson process on the real line with parameter λ and if we have m disk drives on test with independent failures, then the pooled failure times form a Poisson process with parameter $\lambda^* = m\lambda$.
- The probability that an interval of length T contains n points is Poisson with parameter $m\lambda$.
- The time until the k^{th} failure has a gamma distribution with scale $m\lambda$ and shape k.

Counting Process

- In general, the time to an event can be viewed as the result of a counting process, but one without necessarily the same value of λ thoughout time.
- We will learn how to make the hazard λ depend on characteristics of the individual and to vary over time.
- But the exponential model is still interesting as the simplest example of time to event data.