Introduction to Survival Analysis

Shannon Pileggi

STAT 417

OUTLINE

Characteristics of time to event data

Censoring and truncation

Parametric models for time to event data

Group Exercise

Survival analysis is statistical methodology only applied to the study of whether or not someone lives or dies.

1. True
2. False

Chocolate chip activity

- Outcome of interest:
- Response variable of interest:
- Noticeable feature(s) about the recorded values of the response variable:

Survival data

- The collection of lengths of times it takes for the chocolate chips to melt is an example of **survival data** (also called **time-to-event data** or **failure-time data**).
- Time-to-event data are the times until the experimental (or observational) units experience a particular event of interest.
- Examples of events of interest:
  1.
  2.
  3.
  4.
  5.
- Examples of units experiencing the event:

  animate:

  inanimate:

Incomplete data

Special features of survival data include **incomplete** data:

1. **Censoring**: Some event times may only partially known (the exact time until the event occurs is unknown)
2. **Truncation**: Certain subjects are excluded (screened) from the investigation due to some selection condition

Note: We'll refer to an exact known event time as **complete**.

Chocolate chip example:

- Some *complete* melting times:
- Some *censored* melting times:

Survival time random variable

The response variable of interest is the survival time random variable (T), also called the

- survival time
- failure time
- time-to-event random variable

$T =$

The observed values of T include *both* the complete and censored survival times.

Group Exercise

Consider the chocolate chip activity. Which of the following is the response variable for survival analysis?

1. whether or not the chip is melted by 80 seconds
2. the time until the chocolate chip melts (seconds)
3. the amount of the chip that is left over after 80 seconds
4. the time at which we stop monitoring the chip melting

Examples of time to event random variables

1. Time until death after diagnosis of lung cancer (Medicine)
2. Age at which first alcoholic drink is taken (Public Health)
3. Time until students graduate from college (Education)
4. Other examples from your questionnaire:

    ▶

    ▶

    ▶

Identify the time to event random variable

1. Diekmann et al. (1996) investigated the association between driver characteristics and social status of cars to aggressive driver responses by measuring the time that elapsed between the being blocked and honking the horn.
   $T =$

2. Hepburn (2005) found that the time to rearrest for drug-using criminal offenders depended on various socio-economic characteristics of the individual, as well as whether the drug-user was exposed to treatment.
   $T =$

3. Dickersin et al. (2002) investigated whether manuscripts submitted for publication in the *Journal of the American Medical Association* that reported positive results from clinical trial studies were published more quickly than manuscripts reported negative results.
   $T =$

Research questions addressed

- What proportion of subjects will take longer than a particular time to experience an event of interest, i.e. survive beyond a certain time?

- What is the typical time at which individuals experience the event of interest, e.g. how long does it take until half the chocolate chips are melted?

- Of those subjects who survive to a particular time, at what rate do they experience the event at that instant?

- Does survival experience depend on particular characteristics of the individual. For example, do milk chocolate and white chocolate chips melt at the same rates over time?

Features of a survival study

1. A well-defined **event of interest** whose occurrence is being explored with individuals (or objects) *at risk* of experiencing the event (e.g. death due to lung cancer, graduation from college).
2. A clearly defined **beginning of time**: a point in time when no one under study has yet to experience the target event (e.g. date student enters a post-secondary institution).
3. A meaningful **metric** for time (e.g. seconds, minutes, days, months, years, etc.)

Chocolate chip example:

1. Event of interest
2. Beginning of time
3. Time metric

Example survival analysis study: Lung Cancer

- Lung cancer is the number one cause of death from cancer each year in both men and women
- The data file `lung` contains measurements on survival time (in days) of 228 patients diagnosed with advanced lung cancer, *censoring* status, and 7 explanatory variables (data in `R survival` package).
- The original study was conducted by the North Central Cancer Treatment Group of the Mayo Clinic, and data from the study were subsequently analyzed in Loprinzi, et al. (1994).

Group Exercise

For the lung cancer study, what do you think is the *beginning of time*?

1. birth date of the individual
2. date entered into the Mayo Clinic research study
3. date lung cancer treatment began
4. date diagnosed with lung cancer

Lung cancer: features of a survival study

1. Event of interest

2. Beginning of time

3. Time metric

4. Survival time random variable

5. Observed data
   - complete event times:

   - incomplete event times:

Example survival analysis study: First alcoholic drink

- When do individuals consume their first drink of alcohol? The legal age for consuming alcohol is 21 years, but some individuals claim to have had their first alcoholic drink when they were as young as 1 year old!
- Participants were asked to recall the age at which they had their first drink of alcohol.
- Data source: National Comorbidity Survey (1990-1992)

First alcoholic drink: features of a survival study

1. Event of interest

2. Beginning of time

3. Time metric

4. Survival time random variable

5. Observed data
   - complete event times:
   - incomplete event times:

Example survival analysis study: Motorist reaction time

- Researchers intentionally blocked 57 motorists at a green light by a Volkswagen Jetta
- Recorded the time it took for motorists to show signs of aggression
- Signs of aggression included honking their horn or beaming the headlights at the Jetta
- Study performed by sociologists in Germany (Diekmann et al., 1996)

Motorist reaction time: features of a survival study

1. Event of interest
2. Beginning of time
3. Time metric
4. Survival time random variable
5. Observed data
   - complete event times:
   - incomplete event times:

Characteristics of time to event data

Censoring and truncation

Parametric models for time to event data

Censoring vs truncation

Survival data may be *incomplete* due to:

1. **Censoring**:
   - The exact event time (time until the event of interest occurs) is unknown.
   - Pertains to when subjects *leave* a study
   - Censoring can be **right**, **left**, or **interval**
2. **Truncation**:
   - Systematic exclusion of subjects from the study because their event times are either smaller than a threshold value or larger than some threshold value. Subjects whose event times fall outside the threshold values are unknown to the investigator.
   - Refers to when subjects *enter* a study
   - Truncation can be **left** or **right**

Right censoring

- **Right censoring** occurs when the “observed” event time is less than the “actual” event time.
- Explain how right censoring occurred in the following studies:
  - chocolate chip
  - lung cancer
  - motorists

Displaying right-censored event times

- Right censored are often displayed with a “+”

  (e.g., `motorist reaction times:` $2.88, 4.63+, 2.36+, 2.68$)
- For statistical analysis, we use two variables:
  1. $T =$ time to event random variable
  2. $C =$ censoring indicator variable, such that

$$C = \begin{cases} 1, & \text{complete time} \\ 0, & \text{right censored time} \end{cases}$$

- We then have the pair of random variables, (T, C).
- Fill in the censoring status for the motorist reaction times:

  `(2.88,_), (4.63,_), (2.36,_), (2.68,_)`

Left censoring

- **Left censoring** occurs when the “observed” event time is greater than the “actual” event time.
- Can be common in interviews or surveys
- Example study: In one study conducted at the Stanford-Palo Alto Peer Counseling Program, 191 CA high school boys were asked “When did you first use marijuana?” The time-to-event variable is defined to be the age at first use of marijuana. One possible response was “I have used it but cannot recall just when the first time was.”
- Explain how this is left censoring:

Interval censoring

- **Interval censoring** occurs when the event of interest is only known to have occurred between two time points.
- Interval censoring is a generalization of left and right censoring.
- Any combination of left, right, or interval censoring may occur in a study.
- Example study: Consider investigating the lifetimes (in hours) of light bulbs, i.e. the time until the light bulb burns out. The light bulbs are illuminated and we inspect them every 50 hours for 2000 hours. Explain how interval censoring is present.

Group Exercise

Consider a study in which a researcher is trying to ascertain the age of first alcohol consumption. Consider three high school seniors being interviewed at the age of 18. Identify the following individuals as left, right, or interval censored.

1. Beatrice doesn't remember exactly what age she first consumed alcohol, but she knows it was between 11 and 13.
2. Billy has consumed alcohol, but he doesn't remember the first age of occurrence.
3. Sally has not yet consumed alcohol.

Sketch of left versus right censoring

Noninformative vs. informative censoring

Regardless of censoring type (e.g., left/right/interval), it can be:

- **noninformative censoring**, where the censoring operates independently of event occurrence. More formally,
- **informative censoring**, where the censoring does not operate independently of event occurrence. More formally,

For this course, we will be assuming that we have **noninformative** censoring. This can occur when censoring is caused by planned termination or by a subject leaving a study for reasons unrelated to the risk of event occurrence.

Example: noninformative vs. informative censoring

Researchers were interested in the time to relapse for recently treated alcoholics. Patients were observed from the day they left the hospital treatment program for two years. The event of interest was "heavy drinking," defined as consuming three or more ounces of alcohol. Survival time was measured as the number of days between the release date from the program and the first day of heavy drinking.

- **noninformative censoring** occurs when

- **informative censoring** occurs when

Left truncation

- **Left truncation**, also known as **delayed entry**, occurs when only subjects whose event times are *greater* than a "threshold" time enter the study.
- Those subjects who do not experience the condition are not included the study.
- *Example:* Consider a survival study of residents in a retirement center located somewhere in California. Ages at death are recorded, as well the ages at which individuals entered the retirement community. How could this result in left truncation?

Right truncation

- **Right truncation** occurs when only subjects whose event times are *less* than a "threshold" time enter the study.
- Any subject who has yet to experience the target event is not included in the study, and the investigator is unaware of this subject.
- *Example:* A study looked at data on the infection times for adults who were infected with the HIV virus and developed AIDS by June 30, 1986. The data consist of the time in years measured from April 1, 1978 when adults were infected by the virus from a contaminated blood transfusion, and the waiting time to development of AIDS, measured from the date of infection. How could this result in right truncation?

Group Exercise

Consider going to the next SLO bike night to interview participants and study the time since birth (in years) at which individuals learned how to ride a bike. Which of the following could apply to this study, and how?

1. right censoring
2. left censoring
3. interval censoring
4. left truncation
5. right truncation

Group Exercise

A randomised controlled trial evaluated the effectiveness of an integrated care program compared with usual care in facilitating the return to work of patients with chronic low back pain. The event of interest is fully sustained return to work. Trial participants were followed for 12 months.

Which of the following describes an individual whose survival time would be censored?

1. Someone who moved away during the study period and was lost to follow-up
2. Someone killed in an accident unrelated to their condition before a fully sustained return to work
3. Someone who did not return to work within 12 months after entering the trial
4. Someone with a fully sustained return to work within the 12 month study period

Characteristics of time to event data

Censoring and truncation

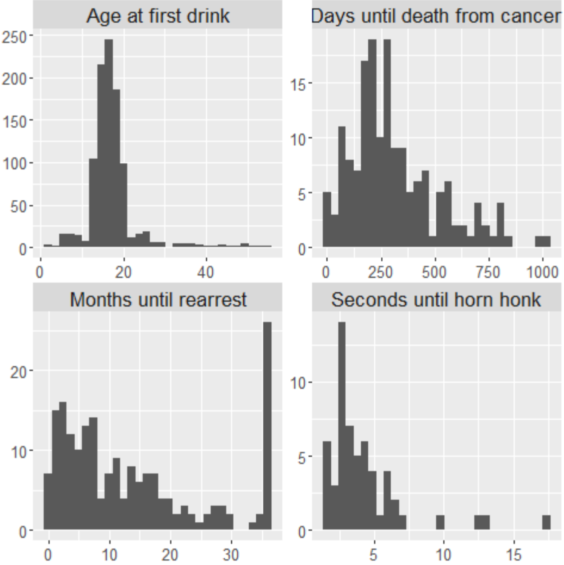
Parametric models for time to event data

Describing survival data

- numerical measures:
- graphical displays:

What could be problematic about these methods?

Example distributions of time to event data



Approaches to analyzing survival data

Survival methods account for censored event times with:

1. Nonparametric methods
   - treat observed event times as a random sample from an *unknown* probability distribution
   - does not require an underlying probability distribution for T
2. Parametric methods
   - treat observed event times as a random sample from a *known* probability distribution
   - requires an underlying probability distribution for T

Example questions parametric models can answer

- **Chocolate chips example:** What is the probability that a randomly selected chocolate chip from a bag will take longer than 90 seconds to melt?
- **Motorist reactions example:** What is the population average time it takes for motorists to react aggressively?
- **Lung cancer example:** Assuming that patients have survived for 400 days after being diagnosed with advanced stage lung cancer, at what rate are they dying at that *instant*?

Distributions for T

- Suppose the observed event times are treated as a random sample from a known probability distribution for T.
- Restrictions on T:
  1.
  2.
- What are some examples of continuous random variables that you have seen before that possess the above characteristics?

Functions of continuous random variables

- Functions for continuous random variables:
  1.
  2.
- New functions for time-to-event random variables:
  3.
  4.
  5.
- If you know the expression for one function, then all others can be derived.
- The survival, hazard, and cumulative hazard functions have *nonparametric* analogs.

Probability density function

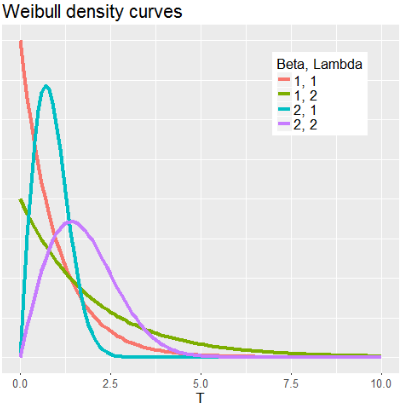
- The *probability density function (pdf)* of T, denoted by $f(t)$, is a function such that for any two constants a and b, with $a \leq b$,

$$\begin{aligned} P(a \leq T \leq b) &= \text{Area under the curve between } a \text{ and } b \\ &= \int_a^b f(t)dt \end{aligned}$$

- For $f(t)$ to be a proper pdf, it must satisfy:
  1.
  2.

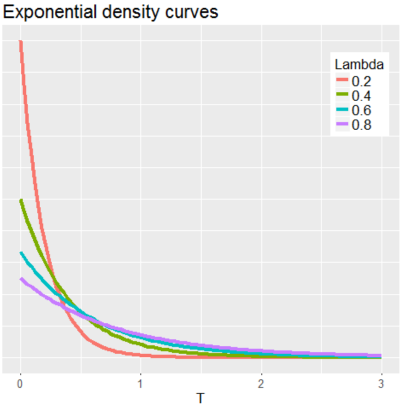
PDF of Weibull random variable, T

$$f(t) = \frac{\beta t^{\beta-1}}{\lambda^{\beta}} e^{-(t/\lambda)^{\beta}}, \; t \geq 0, \; \lambda > 0, \; \beta > 0$$



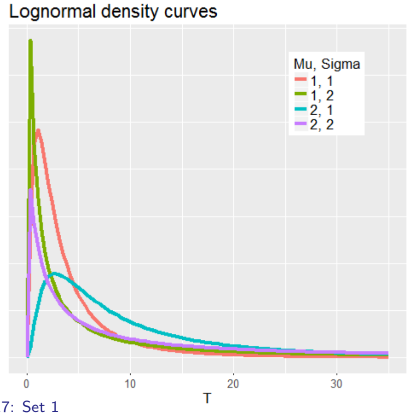
PDF of Exponential random variable, T

$$f(t) = (1/\lambda)e^{-t/\lambda},\ t \geq 0,\ \lambda > 0$$

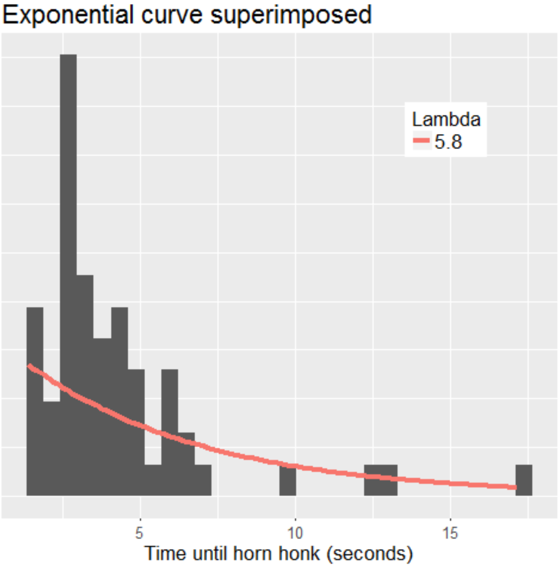


PDF of Lognormal random variable, T

$$f(t) = \frac{\exp\left[-\frac{1}{2}\left(\frac{\ln(t)-\mu}{\sigma}\right)^2\right]}{t(2\pi)^{1/2}\sigma},\ t \geq 0,\ \sigma > 0,\ -\infty < \mu < \infty$$



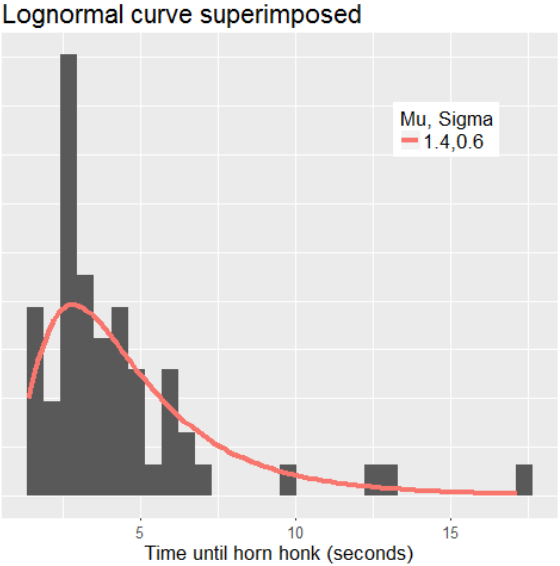
Group Exercise



Does the exponential density with $\lambda = 5.8$ appear to provide a reasonable fit to the data?

1. yes
2. no

Group Exercise



Does the lognormal density with $\mu = 1.4$ and $\sigma = 0.6$ appear to provide a reasonable fit to the data?

1. yes
2. no

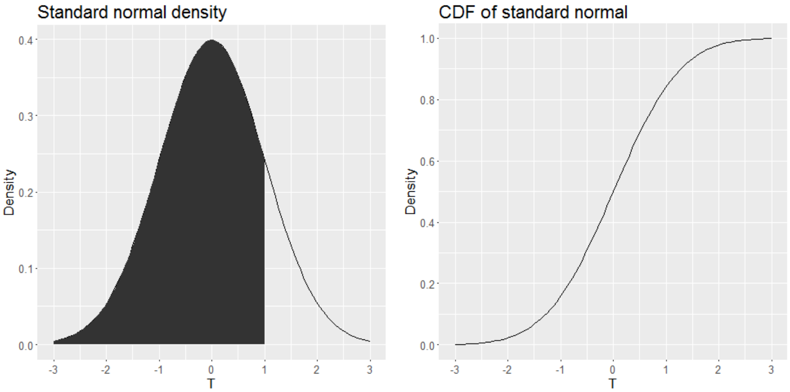
Cumulative distribution function

The *cumulative distribution function* (cdf) of a continuous time-to-event random variable T, $F(t)$, is the unconditional probability that an individual experiences the event of interest before time t.

- $F(t)$ is defined as:
- Since the time-to-event random variable T is non-negative, i.e. $t \geq 0$, $F(t)$ is given by:

Group Exercise

What is $F(1) = Pr(T \leq 1)$?

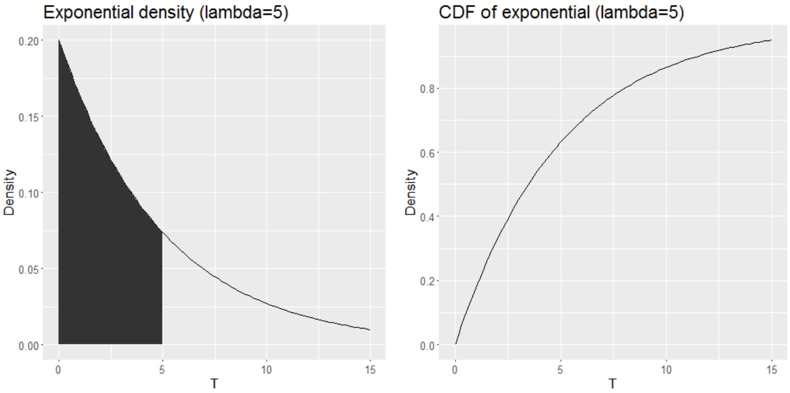


Standard normal density

CDF of standard normal

Group Exercise

What is $F(5) = Pr(T \leq 5)$?



Group Exercise

Suppose the time until death (in days) for males with a particular type of inoperable lung cancer follows an exponential distribution with $\lambda = 125$.

1. Derive the cdf for the time-to-event random variable.

Group Exercise, continued

2. Use the cdf to find the probability that a randomly selected male with inoperable lung cancer will die in less than 75 days.

3. Use the cdf to find the probability that a randomly selected male with inoperable lung cancer will die between the 100th and 150th days.

Gompertz random variable

If T follows a Gompertz distribution with parameters $\theta > 0$ and $\alpha > 0$ then its pdf is given by:

$$f(t) = \theta e^{\alpha t} \exp\left[\frac{\theta}{\alpha}\left(1 - e^{\alpha t}\right)\right], \; t \geq 0$$

Suppose that the time to death in months for mice exposed to a high dose of radiation follows a Gompertz distribution with $\theta = .01$ and $\alpha = .25$

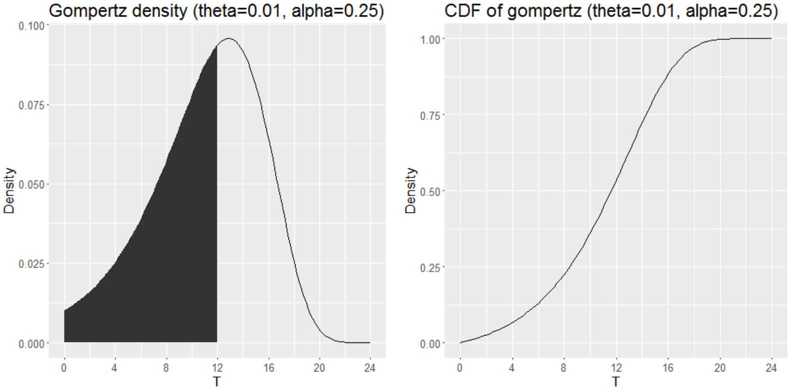
1. Derive the cdf for T.
2. Use the cdf to determine the probability that a randomly chosen mouse will die within 1 year of exposure.

Gompertz random variable, continued

Gompertz random variable, continued

Gompertz random variable, continued

What is $F(12) = Pr(T \leq 12)$?


Gompertz density (theta=0.01, alpha=0.25)
Density
T
CDF of gompertz (theta=0.01, alpha=0.25)
Density
T