Extra Chapter: Survival Analysis

- Lifetime data (or "time-to-event" data) are data in which the random variable is the time until some event of interest.
- This event could be:
  * death
  * relapse of a disease
  * recovery from a disease
  * failure of a mechanical part
  * an insurance claim
  * eradication of an outbreak, etc.
- In many studies, the random variable measures time until patient death, so the analysis of lifetime data is often called survival analysis.
- In the context of industrial or machine failure times, it is often called reliability analysis.
- In most survival studies, the researchers cannot wait until all patients die before concluding the study.

- When data collection ends before all failure times are observed, we have <u>censored data</u>: The survival times that are not yet observed are called <u>censored</u>.

<u>Note</u>: Since patients may enter the study at different times (staggered entry), the amount of time under observation may differ for each subject.

<u>Types of Censoring</u>

- When the experiment ends after a fixed amount of time, this may produce <u>Type I censoring</u>.

- When the experiment ends after a fixed number of failure times (somewhat less than the total sample size) are observed, this may produce <u>Type II censoring</u>.

- When an individual's survival time T is censored and we only know it must be <u>greater than</u> some known length of time [e.g., $T \in (c, \infty)$], we call it <u>right-censored</u>.
- If T is only known to be <u>less than</u> some length of time [e.g., $T \in [0, c)$], we call it <u>left-censored</u>.
- If T is known only to fall between two known times [e.g., $T \in (c, d)$], we call it <u>interval-censored</u>.
- In our discussions, we will mainly focus on Type I right-censored data.

Models for Survival Data

- We denote our lifetime r.v. by T and will assume T is <u>continuous</u> with <u>positive</u> support:

- We could describe the distribution of T by:

  1. The cdf of T:

  2. The survival function (or survivor function) of T:

  3. The pdf of T:

- Note that:

Example 1: A simple model for the distribution of T is the exponential distribution, with pdf

- The exponential distribution has cdf

- So the exponential survival function is

Plot:

- The survival function at time t measures the probability that a random subject will be alive at time t (or the proportion of subjects expected to survive until at least time t).

- For a population having $T \sim \text{expon}(1)$, say, a mean survival time of 1 year:

  $S(1.5) =$

- So a random subject has probability of surviving at least

- This population's median survival time is found by:

Defn: A r.v. T_1 is **stochastically larger** than T_2 [Shorthand: $T_1 \geq_{st} T_2$] if:

Examples:

Defn: The mortality rate at time t is the conditional probability that an individual alive at time t will die between times t and t+1:

- The time t here is measured in some prespecified units, e.g., days, or months, or years, etc.

- The hazard rate is the limiting case of the mortality rate as the time interval approaches 0.

Defn: The hazard function of T is

- The hazard rate is the instantaneous rate of failure and is not a true probability.

- The hazard function characterizes how individuals' death risk varies with time:
  * An item with a <u>constant</u> hazard has the same instantaneous chance of failure at any time t.
  * An item with an <u>increasing</u> hazard tends to "wear out" and to have a higher failure rate as it grows older.
  * An item with a <u>decreasing</u> hazard "improves with age".
  * Human beings may have a bathtub-shaped hazard function:

Connection Between the Hazard and Survival Functions

- Note that

- If we integrate both sides, we see

where $\Lambda_T(t)$ is called the cumulative hazard function.

<u>Note:</u> For a continuous lifetime r.v. T, given any one of:

$f(t)$, $F(t)$, $S(t)$, $\lambda(t)$, or $\Lambda(t)$,

we can find the other four functions and completely specify the distribution of T.

<u>Example:</u> If $T \sim \text{expon}(\beta)$, then T has a <u>constant</u> hazard function.

<u>Proof:</u>

Other Parametric Survival Distributions

- The exponential model is simple to work with, but not especially flexible.
- Generalizations of the exponential may be better models for real survival data.

- The ________ distribution is one such generalization: Recall that

- Another common distribution for lifetime data is the <u>Weibull</u> distribution.

<u>Defn</u>: A r.v. T follows a Weibull(α, β) distribution if its pdf is:

- The Weibull cdf and survival function are:

<u>Note</u>: The Weibull distribution when $\alpha = 1$ is simply the ________ distribution.

- The Weibull is a flexible lifetime model: Depending on α, the Weibull hazard could be increasing, constant, or decreasing:

$\lambda(t) =$

We see:

Pictures: