CompBayes [MTH2133] Survival Analysis Notes

Sources:
http://conference.scipy.org/proceedings/scipy2015/pdfs/allen_downey.pdf
https://www.wikiwand.com/en/Survival_analysis

Survival Analysis:

"Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems." -wikipedia

"Survival analysis is used to study and predict the time until an event: in medicine, the event might be the death of a patient, hence "survival"; but more generally we might be interested in the time until failure of a mechanical part, the lifetimes of civilizations, species, or stars; or in this study the time from birth until first marriage." -Allen Downey

A **survival function** will be produced after a survival analysis.

$$S(t) \equiv \Pr(T > t)$$

Complement of the **cumulative distribution function (CDF)**

CDF: the probability that X will take a value less than or equal to x.

$$S(t) = 1 - CDFT\ (t)$$

If you have a survival function, you can compute a **hazard function.**
Hazard function: the instantaneous death rate at time t; the fraction of people who survive until time t and then die at time t

$$\lambda(t) = -S\ 0\ (t)/S(t)$$

Missing data where an event has not occurred yet is called **censored** data.
When we do know when the event occurred or until how long the person "survived," then we are dealing with the simple case.

Kaplan-Meier estimation

> "The fundamental idea is that at each time, t, we know the number of events that occurred and the number of respondents who were "at risk"; that is, known to to be unmarried. The ratio of these factors estimates the hazard function." -Allen Downey

*With this estimated HazardFunction, we can compute the SurvivalFunction.

The probability of surviving until t is the cumulative product of the complementary hazard function.

$$S(t) = \prod_{ti<t} [1-\lambda(ti)]$$

*Kaplan-Meier estimation and other survival analysis algorithms are also available in a Python package called **Lifelines**
https://github.com/CamDavidsonPilon/lifelines

Lifelines documentation and tutorials
http://lifelines.readthedocs.io/en/latest/index.html

Conepts:
Resampling
Jittering

For our analysis of Airbnb data, we can define the **birth** of surrounding homes/sublets as the date the first home/sublet in a neighborhood is listed on Airbnb. We can define **death** as the date when a neighboring unlisted home/sublet is officially listed on Airbnb.

> The individuals in a population who have not been subject to the death event are labeled as ***right-censored***, i.e. we did not (or can not) view the rest of their life history due to some external circumstances. All the information we have on these individuals are their current lifetime durations (which is naturally *less* than their actual lifetimes) There is also left-censorship, where an individual's birth event is not seen. -Lifelines Intro to Survival Analysis