**Censoring, Survival, and Hazards**

* In survival analysis, we are interested in the time until an event occurs
  + Also called the failure time
  + We care about “how long it took to happen”
* Censoring 🡪 when you have information on that person, but the event is yet to happen
  + Censored data is incomplete, but not missing
  + We will be using right censoring
  + Is noninformative 🡪 this means that patients who are censored should have the same future risk for the occurrence of the event of interest, conditional on exposure, as those who continue to be followed (assumption of independent censoring)
* Time 🡪 tenure for an observation
  + Continuous and always positive
  + Not actually interested in time, but interested in tenure

Survival Function 🡪 probability of surviving past time t

* + Always starts at 1
  + Never increases
  + Bounded below by 0 (or 0%)
* Kaplan-Meier Method
  + Want to estimate the proportion of individuals “still alive” at any given time t
  + Is the maximum likelihood estimate for the non-parametric estimation of the survival curve
  + S(t) = ∏ \* (1-)
    - dk = # of events occurring at time t
    - rk = # of observations available right before time t (risk set)
    - censoring does effect risk set
  + s(1) = s(0) \* (1-)
  + s(2) = s(1) \*(1-)
* summary statistics
  + due to censoring, the mean is difficult to estimate, but the median is still valid as long as long as the event occurs for at least half of the sample
  + the median (also called half-life) is the time t that s(t) drops below 0.5 (or 50%)
    - interpretation: 50% of observations survive beyond time t

**Stratified Analysis**

* can also create separate/stratified curves by group
* different curves result in different estimates for each group
* both the log-rank test(developed by Mantel-Haenszel) and Wilcoxon test have same hypothesis
  + Ho: all survival curves are equal
  + Ha: at least one curve is different
* Log rank test 🡪 for each group, calculate expected events and compare to observe events. This is a chi-square statistic with k-1 df(k is the number of groups being compared)
  + Rho = 0
* Wilcoxon Test 🡪 places a larger emphasis on earlier event times
  + Rho = 1

Hazard Function

* We use the hazard function to summarize the data
* Hazard probabilities
  + Are common in a business setting
  + Example: A customer has survived for a certain length of time, so the customers tenure is t. What is the probability that the customer leaves at time t given they lasted until time t
  + h(t) = P(T = t | T >= t) =
  + to calculate at time t, divide the number of events occurring at time t by the risk set at time t
  + inverse hazard probability 🡪
* Hazard Rates
  + Slightly different interpretation than the hazard probabilities because they are based on a continuous distribution
  + h(t) =
  + is the instantaneous event rate for the risk set at time t
  + given survival up to time t, it is the instantaneous rate of the event happening at time t
  + is the instantaneous event rate for the risk set at time t
  + bounded below by 0, but not bounded above by 1
  + “Assuming the event has not occurred yet, this is the probability the event occurs at this time point
    - Say the hazard probability for a given point in time for contracting a sinus infection is 0.2 with a time measured in months
    - Assuming the individual has not contacted a sinus infection, the probability the contract a sinus infection this month is 0.2
  + The interpretation of the inverse of the hazard function is the length of time until the expected next occurrence
    - Hazard for some point in time for contracting a sinus infection is 0.2 with a time measured in months
    - “Assuming an individual has not yet contracted a sinus infection, we expect this individual to go (1/0.2) = 5 months before contracting a sinus infection (assuming the hazard stays constant)

**Accelerated Failure Time Model**

* Are parametric. We assume failure time has a particular structure and distribution, and will be modeling time until failure
  + Makes it easier to estimate medians, survival, and hazard functions
* Length of time is always positive
* Assumptions:
  + Correct distribution is specified
  + Distribution belongs to the location, scale family
  + Independence across observations
* Natural log of both sides to transform AFT
* Parameter interpretation:
  + Positive 🡪 increase in that variable increases the expected survival time
  + Negative 🡪 increase in that variable decreases expected survival time
  + Zero 🡪 variable has no impact on survival time
  + 100 \*(e^Beta -1) is the % increase in expected survival time for each one unit increase in that variable
* Distributions
  + Exponential
  + Weibull
  + Log normal hazard
  + Log logistic hazard
  + Gamma
* AFT can also predict survival quantiles
  + Length of time until failure occurs

**Cox Regression Model**

* Creates a linear model for the hazard function
* Models the log of the hazard directly
  + Predictions are the hazard rate rather than the failure time like in the AFT model
  + No longer depends on time (likelihood ratio is same across time)
  + Constant proportions on hazards
* Hazard Ratio coefficient interpretations
  + Positive parameter 🡪 increases in that variable increase the expected hazard
    - Increases the rate/risk of failure
  + Negative parameter 🡪 increase in that variable decreases the expected hazard
    - Decrease in the rate/risk of failure
  + E^beta is the hazard ratio 🡪 the ratio of hazards for each one unit increase in the variable
  + 100\*(E^beta – 1) is the % increase in the expected hazard for each one unit increase in the variable
* In AFT and PH models, estimation depends on some distributional assumption around either the failure time or the baseline hazard
  + However, in PH models, Cox noticed the likelihood can be split into two pieces
    - 1st piece depends on h(t)
      * Is non-parametric (no assumptions about form or distribution)
    - 2nd piece only depends on the parameters
      * Is a parametric
    - Hence why it is semi-parametric
* Assumptions
  + Proportional hazards
    - No interactions with time
    - Use Schoenfeld residuals
  + Linearity
    - Use Martingale residuals
* Martingale residuals
  + Difference between the observed number of events and the expected number of events at a specific point in time
    - Not symmetrical around zero
  + For linearity
* Schonfeld Residuals
  + Calculated for each variable for each individual
  + Difference between the actual value of the variable and the expected value for someone who had the event occur for the first time
  + For proportional hazards
  + Coefficients should not depend on time, and we want to see a strait line
    - Want to fail to reject H0, which implies there is no relationship with time