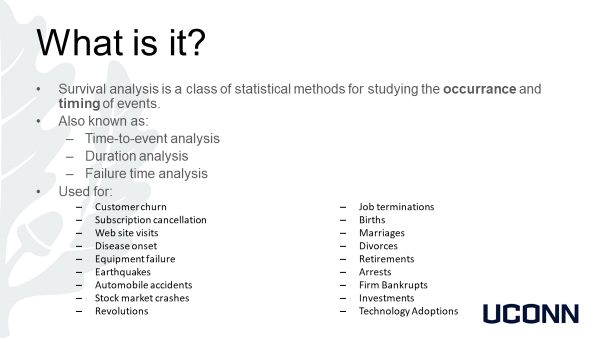
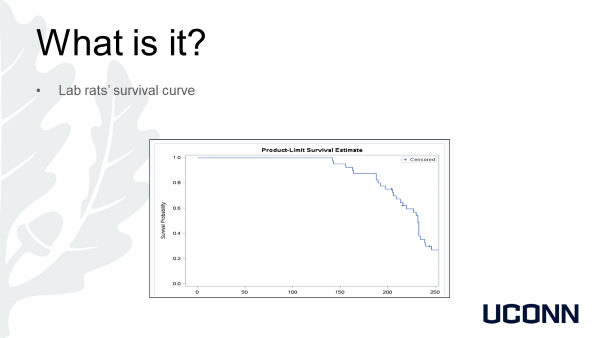
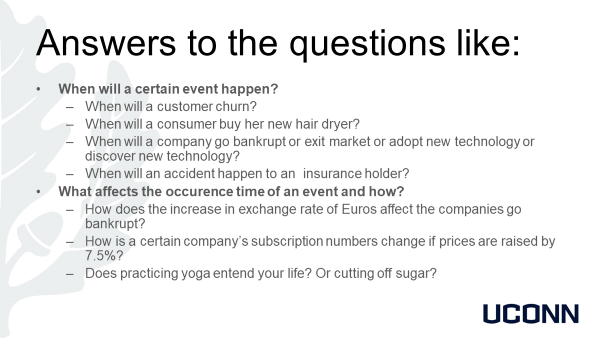
# 

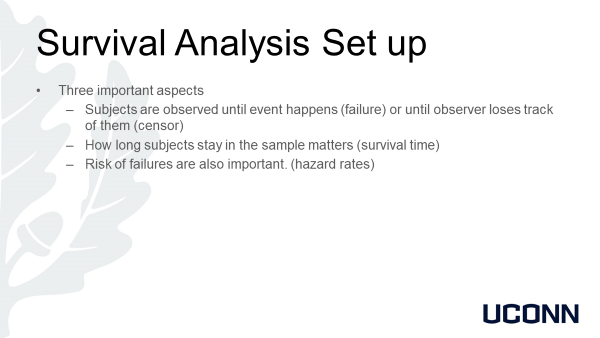
# Problems

* If we try to model with Logistic regression
  + Dichotomous dependent variable: arrested, not arrested
  + This method ignores the timing of arrests. A person who is arrested one week after the release have , on average, higher propensity to be arrested than the person who are not arrested until week 52.
* If we try to model with Linear regression
  + Dependent variable is length of time between release and first arrest.
  + What about people who are not arrested during the 52 week follow-up but entered 52 weeks in their arrested column?
    - We can take out the observations that are not arrested, and model the data (**loss of important information**)
    - We can assume that not arrested are arrested on week 52 and model the data. (**a big underestimation and may cause bias**)
* Employment column
  + In both methods above, it is not clear how to deal with time-dependent variables like employment.
    - We can create dummy variables from 1 to 52 for employment but it can be inefficient and awkward.
    - More importantly, Employment variable may be the consequence of the arrest rather than cause. In particular, someone who is jailed cannot work full-time.
* Next chapters will explain how survival analysis methods deal with both censoring and time-dependent variables.









# Survival Data

* Event: a *qualitative* change that can be **situated in time**. A qualitative change is a transition from one state to another.
  + Marriage
  + Promotion
  + Arrest
  + Growing to adulthood is not an event because it can’t be situated in time.
* To apply survival analysis, we need to know **the duration** subject stayed in one status before moving to the new one.
* The unit of time is important. And we must use the same interval for each observation. Find another way if the onset of event is vague.
  + If accident in a race, every second is significant.
  + If revolution, the year is significant.

# Process of Survival Observation

* Observe a set of individuals at some well defined time
* Follow them for some period of time
* Record the times at which the events of interests occur
* Calculate the time between as survival time
* **Note: It is okay that some of the individuals don’t experience the event. Then we mark them as they didn’t experience it but still it carries very important information. (i.e. that individual survived at least the observation time)** 
  + **people who can’t get a job (event is getting a job)**
  + **active subscribers (event is canceling subscription ).**

# Aims of Survival Analysis

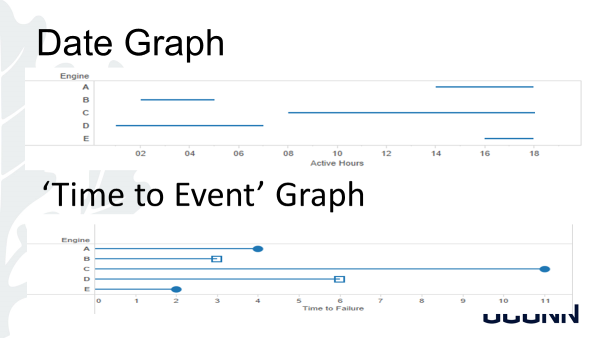
* Probability of an event happening at some time
* Estimate causal(prescriptive) or predictive models in which the risk of an event depends on covariates.
  + Constant covariates
    - Race
    - Gender
  + Time-varying covariates
    - Income
    - Marital status
    - Blood pressure

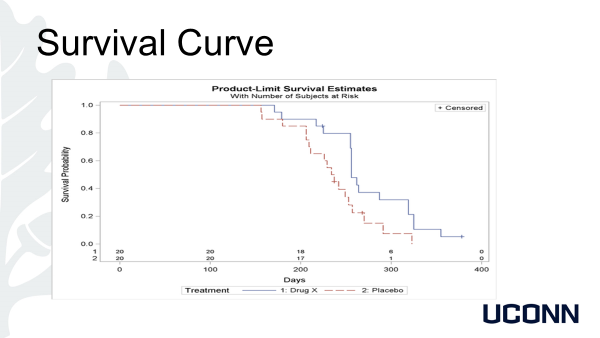
# Two Common Features

* Censoring (marking that an individual hasn’t experience the event)
* Time-dependent covariates
* **These two features are very difficult to handle with conventional statistics!!!**

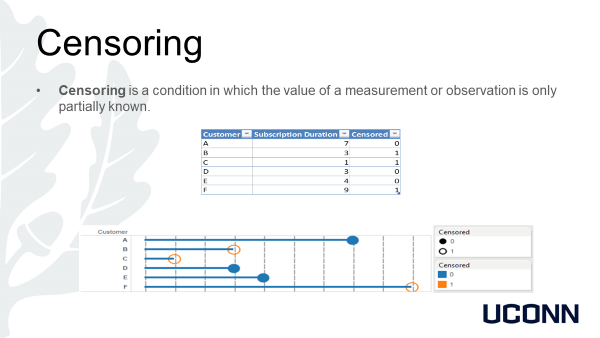
# Approaches to Survival Analysis

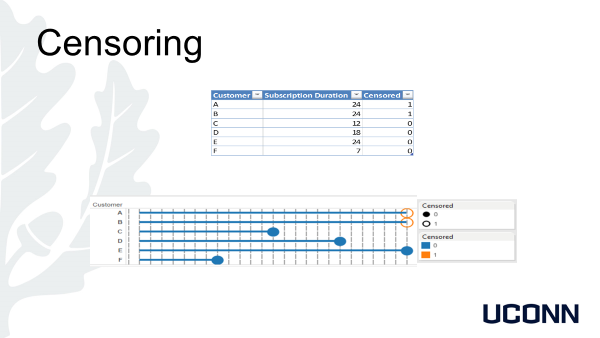
* Many different models including:
  + Life table method
  + Kaplan-Meier estimation
  + Exponential regression
  + Log-normal regression
  + Proportional hazards regression
  + Competing risk models
  + Discrete time models
* Some of these methods are complementary, some are alternatives.
* Next chapters will teach you how to use these models for your specific data needs.





# Censoring





# Types of Censoring

Right, Left and Interval Censoring

* Right censoring
  + Person’s age at death is larger than 50. T>50
* Left censoring
  + Girls’ menarche time is less than 12. T<12
* Interval Censoring
  + HIV infection time is censored between 2 and 3. 2<T<3

Type I, Type II and Random Censoring

* Type I censoring (observations are censored if they survived until a designated time)
* Type II censoring (observations are censored once a number of events happened)
* Random censoring(tracks of subjects are lost or unable to get information)
  + Informative Censoring – Ph. D. students getting a degree
* For the sake of simplicity, **Type II censoring** is not covered in this course. For further information on **Type II censoring**; please refer to P. Allison’s book.

# Survival Distribution

* All standard approaches to survival analysis are stochastic. In other words, the collection of times at when events happen are assumed to be realizations of a random process. Which means ***T* is a random variable having a probability distribution**.
* *There are many models in survival data. These models are distinguished by the probability distribution of T.*
* Three different ways to describe probability distributions
  + Cumulative distribution functions
  + Probability density function
  + Hazard function

# Cumulative Distribution Function

* Denoted by
* It is a function that tells us the probability of the variable (in our case the probability of event time) will be less than or equal to any value t we choose.
* Thus,
* F is non-decreasing, i.e. . For proof, see <http://www.essie.ufl.edu/~alex/courses/ocp6168/section1_HW3.pdf> (Optional)
* Relation with **Survivor Function**
  + Probability of survival **beyond** t
  + Never increases (since is non-decreasing.)
  + We’ll talk about survivor functions later in more detail.

# Probability Density Function (p.d.f.)

* It is the derivative (or the slope) of c.d.f.
* (2.1)
* (2.2)
* More related to histograms and distribution shapes than probabilities
  + Normal distribution has a bell like shape.

# Hazard Function

* What is hazard?
  + **Number of events per interval of time.** Also called ***rate.***
* It is more popular than the p.d.f. as a way of describing distributions of event time.
* (2.3)
* Can be denoted by or in other sources.
* It is used to quantify the instantaneous risk that an event will occur at time *t*.
* t is a point, so probability is zero at t. We need a very small interval around t. We use Δt for defining an interval.
* It is sometimes described as *conditional density* because of the condition on its numerator.

# Interpretations of the Hazard Function

* Although it may be helpful to think hazard as instantenous probability of an event at time t**; in reality, it is not probability**, it can be greater than 1 but it is **non-negative.** The reason for that is it has also a time dimension, i.e. the interval related to the hazard.
* **Hazard unit** is probability, i.e. it is defined in terms of probability.
* It is a characteristic of individuals, not populations.
* h(t) is rate.
* 1/h(t) is the expected length of time until the event occurs again.
* Accident hazard (h(t))
  + 0.1 per year (assuming the hazard stays constant)
* Accident interval (1/h(t))
  + 10 years until the next accident.

# An Example

* Suppose;
  + Sample of 10000 people observed for one month.
  + 300 cases of influenza
  + Total exposure time is 10000 months. (10000 people \* 1 month)
* Assuming that the hazards is constant over the month and across individuals:
  + An estimate of hazard is 300 / 10000 = 0.03
* If a person cathches flu, and everyting stays the same(hazard stays constant), the next time that person cathes flu again is 1/0.03 = 33 months later.

# Some Insights on Hazard

* Every **individual** has different hazards for different **kind of events** in different **environments.**
* **Example**
  + Hazard of death for me exercising and my father exercising. (individual)
  + Hazard of accident for me sitting at home and for me scuba diving. (environment)
  + Hazard of marriage for me today, hazard of promoted for me today. (different events)
* Hazard funtion for a specific individual and a specific event varies greatly with the ambient condition.

# Simple Hazard function

* Every hazard function has a corresponding probability distribution. But, as we have seen earlier, hazard functions can be very complicated with the effect of changing ambient conditions. This fact is reflected by different hazard functions.
* Some very basic simple hazard models are:
  + Case when hazard is constant.
  + Case when log of hazard is a linear function. ()

# Constant Hazard Function

* Simplest hazard function is the one with the constant hazard across time.
* is constant, or
* Eq. 2.6.,
* Eq. 2.1.,

* This is the **well-known exponential distribution** with parameter .
* So, a constant hazard implies an **exponential distribution of time-to-event variable T.**

# Simple Hazard Models (cont.)

Let’s assume natural logaritm be a linear function of time.

**(1)**

**(2)**

This hazard is a **Gompertz distribution**.

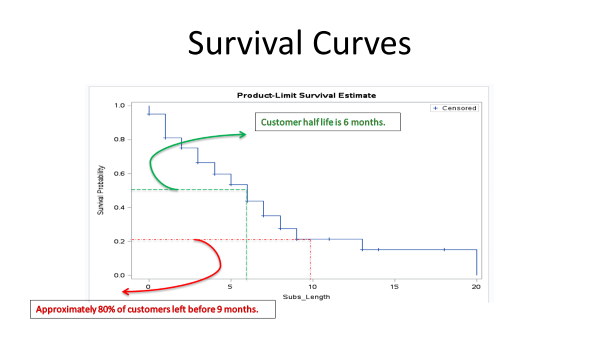
**(3)**

**(4)**

This hazard is a **Weibull distribution**.

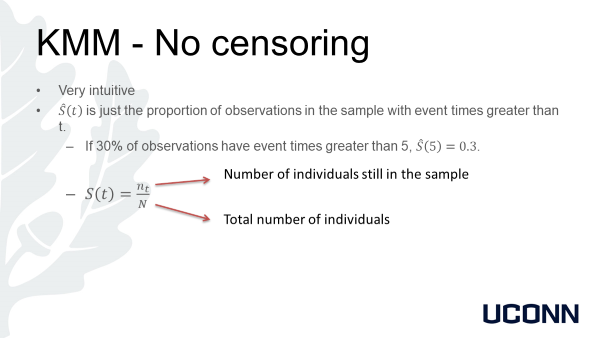
* We will be using PROC LIFEREG to estimate Hazard models in the later chapters.

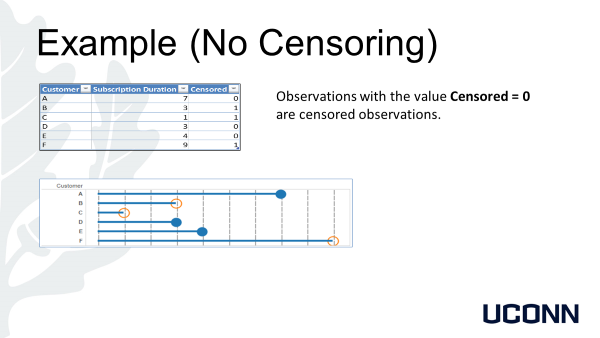
# Survival Curves

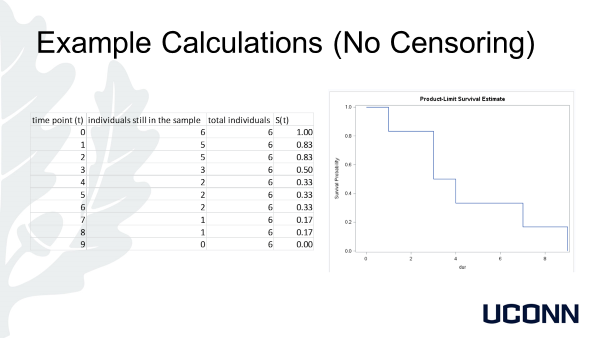


# Kaplan-MEIER Method

* It was known long before Kaplan and Meier.
* They showed that this method was, in fact, the nonparametric maximum likelihood estimator.
* Three cases:
  + No censoring
  + Right censoring
  + Other kinds of censoring(not covered in this course)







# KMM – Right Censoring

* Also simple
* is the number of observations ‘died’ at time t.
* is the number of observations remaining at time t-1.
* Note that

