**The Many Different Types of Survival Analysis and the Challenges in Learning Them**

Survival analysis isn't just a single model.

It's a whole set of tests, graphs, and models that are all used in slightly different data and study design situations.

Choosing the most appropriate model can be challenging.

In this article I will describe the most common types of tests and models in survival analysis, how they differ, and some challenges to learning them.

What they share is all of these allow you to test how predictor variables predict an outcome variable that measures the time until an event. They are all based on a few central concepts that are important in any time-to-event analysis, including censoring, survival functions, the hazard function, and cumulative hazards.

**The Kaplan-Meier curve**

A Kaplan-Meier curve is an estimate of survival probability at each point in time. It has very few assumptions and is a purely descriptive method. This is often your first graph in any survival analysis.

You can get confidence intervals for your Kaplan-Meier curve and these intervals are valid under a very few easily met assumptions. You can easily extract quartiles and medians (and their confidence limits) from the Kaplan-Meier curve.

**The log rank test**

The log rank test is a direct comparison of the Kaplan-Meier curves for two or more groups. You can think of it as a one-way ANOVA for survival analysis. It is easy to calculate, has very few assumptions, and for many settings, it may be the only test you need.

The only limitation of the log rank test is that it does not extend easily to more complex settings like continuous covariates and interactions.

**Cox proportional hazards regression**

This is the model that most of us think of when we think Survival Analysis. It's a pretty revolutionary model in statistics and something most data analysts should understand.

Cox proportional hazards models are unique in that they're semi-parametric. That's right--not entirely parametric and not entirely non-parametric. This creates a lot of flexibility, but it also creates an assumption that is so important it's right in the name--proportional hazards.

This is a concept that is a little bit strange and takes a bit of explaining. But once you can wrap your head around it (and you know how to check it and what to do instead if it doesn't fit), you'll see how incredibly useful Cox models can be.

**Parametric models**

Parametric models (also known as accelerated time models) make an even stronger assumption than the Cox proportional hazards model. They force you to choose an appropriate survival distribution for your data. The most commonly used survival distributions are exponential and Weibull and these provide fundamental insights into the mechanistic structure of your data.

For a parametric model, this choice of a survival distribution represents the methods greatest strength and biggest potential weakness. When you select a parametric distribution, you can make strong conclusions about survival patterns over time and you can even (very carefully) extrapolate beyond the range of the observed data. But the distribution you choose will affect all of your results, especially your extrapolations, so you need to make sure you pick one that works.

**Frailty models**

Frailty/cluster models account for correlation within a group by introducing a random effect called a frailty term. These are useful when you have

* Multiple events for the same person
* Multiple sites on the same person
* Multiple animals in the same litter.

If you're at all familiar with running mixed or random effects linear models, you know how hard these can be. So with frailty models, you need to understand not just the survival part of the model, but how to fit and interpret the random effects.

**Competing risk models**

Just like there’s more than one way to skin a cat, there is more than one way that you can die. Cigarettes can kill you in many ways: lung cancer, emphysema, heart attacks, or strokes, to name a few. This type of data can occur for events other than mortality. A high school dropout can eventually return to school to get a diploma, complete the requirements of a GED, or stay a dropout (a censored observation).

A competing risk model allows you to partition the events that occur in your model into discrete competing risks and examine how different factors influence not just the risk of the event, but the mix of competing causes.

**Discrete Time Model using logistic regression**

All the models we've talked about so far assume that time is measured continuously, but sometimes it just isn't. If you are measuring time until a graduate student finishes their PhD, they can't actually graduate any day of the year--only at the end of a semester. So we measure time until finishing as the number of semesters.

Cox models don't work when times is discrete because there too many ties--too many students who finished in the exact same number of semesters.

So here we use an entirely different type of model: logistic regression. If you're familiar with logistic regression, this isn't hard to apply. Each semester, the student graduates or doesn't.

The trick is to set up the data correctly so that you incorporate the censoring. Unfortunately this is not the data set up that is used for Cox models, so you may have to spend some time doing data management.

(This is the one type of model that isn't included in our survival analysis workshop, though we do include a 90-minute webinar recording on it as a bonus).

**Conclusion**

The biggest challenge in data analysis is knowing what statistic to use for what type of data. Survival analysis, with its unique focus on censoring, offers a variety of tools for data that often cannot be analyzed any other way. Within the realm of survival analysis, there are simple methods with very few assumptions which may be all that you need. If your data is more complex, however, you have a wide range of methods that you can choose from, though each of these comes with a different set of assumptions. In our survival analysis workshop, you’ll learn how each of these models work, when to use which ones, and how you can use your favorite statistical software (R, SAS, SPSS, or Stata) to fit these models.