Survival heterogeneity review

Gordon A. Fox and Bruce E. Kendall

January 9, 2026

```
library(survival)
library(ggplot2)
library(ggsurvfit)
library(ggpubr)
```

This is a working draft. Hell, it's not even that yet . . . TEST

1 Initial questions

- Why models?
- Why covariates?
- But there are many possible covariates, most of which can't be measured in practice.
- What if you ignore some, or just don't have enough data? This is one reason for using frailties

2 Introduction

Individuals vary in their survival probabilities. Not only because of variation in age, size, or stage, but because of their histories of nutrition, parasites, or disease, their varied access to resources or exposure to toxicants, and their varied positions in social hierarchies. There may also be genetic variation that contributes to varied survival propensities.

This said, distinguishing between individual variation and stochasticity can be a knotty problem. Some individuals die quickly: did they have poor survival probabilities, or bad luck? In individual cases, we generally can't say, but appropriate sampling and modeling can address this problem. Similarly, survival is usually affected by many things, some of which we can readily identify and measure, like family, site, or presence of a disease. Accounting for these is necessary, both to make appropriate estimates about the survival process, and to understand some of the causes underlying variation in survival.

There are several types of regression models used for survival data. These are somewhat specialized for survival data because survival times are not normally distributed, and the data are typically censored. By censored, we mean that we know only inequalities about the data (an individual survival time x is greater than some value ($x \geq \tau$), between two values ($\tau_1 \leq x \leq \tau_2$), or less than some value($x \leq \tau_2$)) It is also possible for the time of death to be known, but not the time of "birth." Censorship and non-normality generally make it impossible to use well-known approaches like GLM.

Mark-recapture data

But there are powerful methods for survival analysis, mainly developed in biostatistics, industrial reliability testing, and sociology (where it is often called event history analysis).

2.1 Predictors in survival models

There's a bestiary of names – covariates, factors, random factors, independent variables, and so on. For the present, just call them all "predictors."

- Because sometimes that's where a lot of biological interest lies. E.g., we may want to know the effect of predator density on prey survival. Sometimes these have been manipulated experimentally.
- Because these factors induce variation in survival within the population.
- Because otherwise we may delude ourselves by e.g., just taking the mean survival.

How is it that we may be deluding ourselves by ignoring underlying variation? Consider a simple example: there are two types in a population. Both have constant survival probabilities, but one (perhaps in poor microsites) survives each interval at $p \approx 0.86$, while the better survivors have $p \approx 0.96$. The simulated survival process is shown in Figure 1.

```
source("SimpleSurvHetSimulation.R")
```

This example (Figure 1) may seem extreme, but it provides several important conclusions. Most important, the cumulative survival probability for the pooled population does not reflect that for any individual. Moreover, reliance simply on the survival estimate for the pooled population would miss the biological processes underlying the difference. Ignoring the predictors (in this case, type) can lead to misestimation. For example, if the initial proportions of the two types were different (in this example they are equal), the right-hand figure would be unchanged, but the left-hand figure (for the pooled population) would be different.

An important conceptual distinction is between measurable and nonmeasurable predictors. We can measure, at least in principle, the height of an organism. On the other hand, while we can posit that there are predictors that make an individual more or less prone to some disease (these may be, e.g., genetic, site-related, or socially induced), in practice it may be often be impossible to measure these. Similarly, while we may find that some families live longer than others, characterizing the genetic and environmental factors that cause this is typically beyond our grasp in practice.

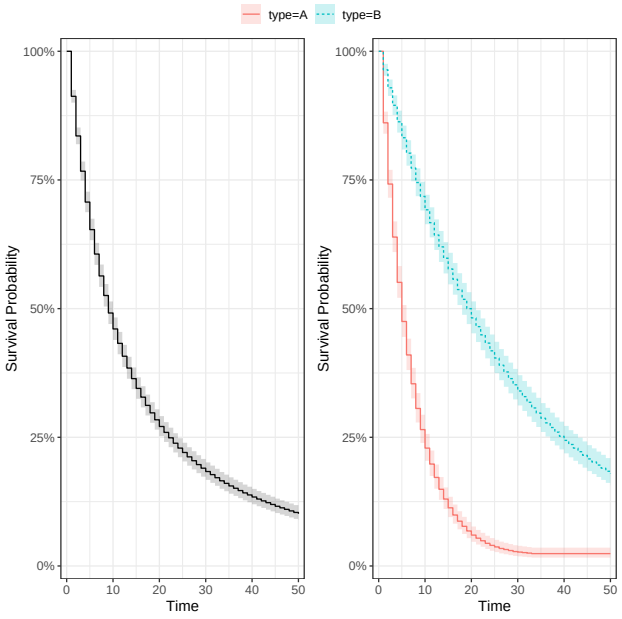

Figure 1: Survival in a population composed of two types, one of which has superior survival. Survival estimated for the entire population, ignoring the heterogeneity (left) does not represent that of any individual.

3 Frailty

4 Notes made earlier, to be incorporated or tossed

By its nature, estimating survival probabilities, and heterogeneity in them, presents some conceptual and technical challenges. Individuals have single sample - there's no replication. For individuals, we can observe date of death, but not the survival probability. And the process is inherently stochastic. So for individuals without any covariates, we can't distinguish between a high probability of survival and good luck, or vice versa.

But there are ways of estimating meaningful quantities. They all involve some combination of model-based estimation, measurement of meaningful covariates, and aggregation of individuals.

There are 3 main types of models:

- AFT
- Hazard-based regression
- Logistic regression
- Mark/recapture approaches

They each have strengths and weaknesses; for an introduction, see XXX.

One difficulty is that there are relatively few ecologists or evolutionists who are deeply familiar with the statistical methods used for analyzing survival data. Deep assumption in much of ecol/evol that GLM and its relatives covers most of the statistics needed. Most of the survival literature is in biostatistics and human demography; these are approachable, but the problems and the jargon are somewhat different.

Frailty models do not provide an estimate of the amount of heterogeneity, unless one can defend certain assumptions. This is because the variance of the random effect depends on the model specified and on the questions asked. CHECK ON THIS the random terms can be inflated/deflated by other terms in model. Obviously, its magnitude is meaningful in a qualitative sense.

There are multiple frailty models, and multiple senses of the word. Individual frailty, shared frailty, correlated frailty are the most common. It seems likely there will be more.

Frailty distibutions

Value of studying covariates which you *may* treat as random.

Study of heterogeneity in survival presents additional challenges.

********** Individuals vary in their survival probabilities. Not only because of variation in age, size, or stage, but because of their histories of nutrition, parasites, or disease, their varied access to resources or exposure to toxicants, and their varied positions in social hierarchies. There may also be genetic variation that contributes to varied survival propensities.

This said, distinguishing between individual variation and stochasticity can be a knotty problem. Some individuals die quickly: did they have poor survival probabilities, or bad luck? In individual cases, we generally can't say, but appropriate sampling and modeling can address this problem.

There are several types of regression models used for survival data. These are somewhat specialized for survival data because survival times are not normally distributed, and the data are typically censored. By censored, we mean that we know only inequalities about the data (an individual survived at least as long as $\tau < x$, or between two values ($x_1 \leq \tau \leq x_2$), or less than some value ($\tau < x_2$). Censorship and non-normality usually make it impossible to use well-known approaches like GLM. Mark-recapture data

But there are powerful methods for survival analysis, mainly developed in biostatistics and industrial reliability testing.

Empirical studies: scrub-jays.