What is the probability of survival?

Shannon Pileggi
Statistics Department, California Polytechnic State University San Luis Obispo

Jeff Sklar
Statistics Department, California Polytechnic State University San Luis Obispo

Keywords: survival analysis, interpretation

Abstract

This is our abstract.

1. Introduction

Much in literature about correct and incorrect interpretations of p-values. Anything else that gets a lot of attention on interpretation? Not much on interpreting survival probabilities, and most texts books address it in one line and move on. However, it is easy to incorrectly interpret survival probabilities and this merits careful thought and attention. Other topics: interpreting probabilities, complements. Scope: teaching survival analysis to undergraduate statistics majors, can be applied to graduate students as well.

2. About survival analysis

Survival analysis is a set of statistical methodologies used to model survival data, also known as time to event data or failure time data. The time to event data are the times until the experimental, or observational, units experience a particular event of interest. Applications of survival analysis can include time to death after diagnosis with lung cancer (Loprinzi et al., 1994), time to re-arrest after release from prison (Hepburn and Albonetti, 1994), time to answer a question after posted on an online forum, time until students graduate from college, and time until aggressive reaction by motorists (Diekmann et al., 1996). (reference papers that use parametric modeling?)

A particular characteristic of survival data is that the observed event times may be either complete or incomplete. A complete event time is recorded when the actual time to event is known. However, an incomplete time occurs when the actual time to event is unknown, but we record some sort of observed event time. Incomplete event times are said to be censored, and censoring can be left, right, or interval. For this purposes of this paper, we will only address right-censoring as it pertains to interpretations of survival probabilities. If an observation is right-censored, then the actual time to event is greater than the observed time.

Other features of a survival study include a well defined event of interest, a clearly defined beginning of time, and a meaningful time metric. Provided all of this, the random variable T modeled in survival analysis is defined as

$$T = \text{time to event}$$

For the purposes of this paper, restrictions on T include:

1. T is a non-negative random variable (i.e. $T \geq 0$)
2. T is continuous (we will not address discrete time survival analysis methods)

The random variable T can be modeled either with parametric or non-parametric methods. Non-parametric methods treat observed(???) event times as a random sample from an unknown probability distribution and do not require an underlying probability distribution for T. A common non-parametric method for analysis of survival data is Kaplan-Meier curves. However, this paper will focus on interpretations from parametric methods, in which we treat observed(???) event times as a random sample from a known probability distribution and requires an underlying probability distribution for T. Examples of probability distributions used to model survival data include weibull, exponential, lognormal, and gompertz (capitalization on any of these?).

Common questions that arise regarding the random variable T include both *failure* and *survival* probabilities, where failure indicates that the event of interest occurs before some time t and survival indicates that the event of interest does not occur after some time t. The three basic functions that can be utilized to address these questions include the *probability density function*, the *cumulative density function*, and the *survival function*.

The probability density function (pdf) of T, denoted by $f(t)$, is a function such that for any two constants a and b, with $a \leq b$,

$$Pr(a \leq T \leq b) = \int_a^b f(t)dt \tag{1}$$

such that $f(t) \geq 0$ for all t and $\int_0^t f(t)dt = 1$. This represents the unconditional probability that the time to event is in the interval (a, b).

The cumulative density function (cdf) of T, denoted by $F(t)$, is the unconditional probability that an **(individual experiences the event of interest before time** t**)** that the time to event is less than t. As T is non-negative, $F(t)$ is given by:

$$Pr(0 \leq T \leq t) = Pr(T \leq t) = F(t) = Pr(T \leq t) = \int_0^t f(y)dy \tag{2}$$

The height of $F(t)$ at point t gives $Pr(T \leq t)$ and also corresponds to the area under the density curve $f(t)$ to the left of t. This corresponds to the probability that an individual *does not* survive beyond time t.

Lastly, the survival function, denoted by $S(t)$, is the unconditional probability that the time to event is greater than T. This is given by one minus the cdf:

$$Pr(t \leq T \leq \infty) = Pr(T \geq t) = S(t) = 1 - F(t) \tag{3}$$

This corresponds to the probability that an individual *does* survive beyond time t.

Clean up equations (1), (2), and (3) for simpler presentation.

3. Example application

Should we find a paper that actually uses parametric methods? Or just apply parametric methods to data from any general paper?

- Event of interest:
- Beginning of time:
- Time metric:
- Complete event times:
- Incomplete (right-censored) event times:

Show distribution of event times, fit parametric model, provide sample calculations and interpretations for $S(t)$ and $F(t)$.

4. Correct/incorrect interpretations

While this may all seem very straightforward, notice how diligent we have been to describe the various probabilities in terms of the random variable $T =$ time to event. However, many individuals unconsciously re-phrase the probability interpretations in terms of whether or not the event is experienced, which is a different random variable. This is likely because survival analysis is considered to be a more advanced topic instructed to students who are already very familiar with bernoulli or binomial type calculations.

So if we contemplate the example probability calculations presented in Section 3, there are actually two components that this can be written in terms of:

1. Whether or not the event of interest occurs.
2. Whether or not the time period under examination is before or after t.

Table 1: Alternative interpretations of survival probabilities

Outcome	Occurrence	Time Period	Interpretation	Represents
A.	Event **does not** occur	**after** time t	survival	$S(t)$
B.	Event **does** occur	**before** time t	failure	$F(t)$
C.	Event **does not** occur	**before** time t	?	?
D.	Event **does** occur	**after** time t	?	?

This leads to four potential interpretations of survival and failure probabilities.

Would interpretations slightly differ between parametric and nonparametric methods? Ie, conditional versus unconditional probabilities?

5. Textbook definitions of survival probabilities

Various ways others have defined survival probabilities - do they align with our discussion?

6. What about the word until?

7. Results from the classroom

Give multiple choice question on midterm exam, report results. Apply intervention, re-examine students?

8. Acknowledgements

9. References

References

Diekmann, a., Jungbauer-Gans, M., Krassnig, H., and Lorenz, S. (1996), "Social status and aggression: a field study analyzed by survival analysis.,".

Hepburn, J. R., and Albonetti, C. A. (1994), “Recidivism among drug offenders: A survival analysis of the effects of offender characteristics, type of offense, and two types of intervention,” *Journal of Quantitative Criminology*, 10(2), 159–179.

Loprinzi, C. L., Goldberg, R. M., Su, J. Q., Mailliard, J. a., Kuross, S. a., Maksymiuk, a. W., Kugler, J. W., Jett, J. R., Ghosh, C., and Pfeifle, D. M. (1994), “Placebo-controlled trial of hydrazine sulfate in patients with newly diagnosed non-small-cell lung cancer.,” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 12(6), 1126–9.
URL: *http://www.ncbi.nlm.nih.gov/pubmed/8201374*

10. Author contact information