

OVERVIEW

Description of the goals and material

The typical description of survival analysis is mathematically complex because the topic is mathematically complex. The primary goal of these notes, however, is a sophisticated introduction to survival analysis concepts using only elementary mathematics and relying heavily on examples and intuitive explanations. No previous background is assumed. The mathematical level is completely accessible with knowledge of high school algebra, a bit of calculus (one semester) and a one-year course in basic statistics (for example, t-tests, chi-square analysis, correlation and some experience with linear regression models). This material is part of a Masters degree course in epidemiology at the University of California, Berkeley. With a minimal background in mathematics and statistics, the reader should be able to appreciate the analytic methods and with the help of modern computer systems use these techniques to understand much of biologic and medical survival data.

A secondary goal is the introduction (perhaps, the review) of important statistical methods that are key elements of survival analysis but are also important to statistical analysis in general. Such techniques as statistical tests, transformations, confidence

intervals, analytic modeling techniques and likelihood methods are presented in the context of survival data but, in fact, are statistical tools that apply to the analysis of many kinds of data. Similarly, discussions of statistical concepts such as bias, confounding and interactions are presented in terms of survival analysis but also clearly apply to the understanding of a broad range of analytic techniques.

To achieve these two goals these notes are divided into nine topics:

- LECTURE 1: **Rates and their properties**
- LECTURE 2: **Product-limit estimation**
- LECTURE 3: **Life tables**
- LECTURE 4: **Two especially useful statistical tools**
- LECTURE 5: **Exponential Survival Probability Distribution**
- LECTURE 6: **Weibull Survival Probability Distribution**
- LECTURE 7: **Analysis of Two-sample Data**
- LECTURE 8: **General Survival Model: parametric**
- LECTURE 9: **General Survival Model: nonparametric**

The first topic is a description of rates. Rates are fundamental to epidemiologic and medical research as well as the basis for much of survival analysis. Following this description, two ways to estimate rates are discussed (product-limit and life table methods).

Then, slightly digressing, the principles of likelihood estimation are discussed. This valuable estimation technique is then applied in the remaining material; starting with an application to the exponential survival distribution. The exponential and the Weibull (next topic) survival distributions are central to parametric survival analysis. After their introduction, these two survival distributions play major roles as components of survival analysis models. The parametric comparison of two groups ("two-sample" models) and the analysis of multivariable data make-up the next two topics. The notes end with a

discussion of semiparametric analysis of survival data (sometimes called the Cox proportional hazards model).

All techniques are extensively illustrated with both analytic and graphic examples from the San Francisco Men's Health Study. This unique study was established in 1983 to conduct a population based prospective investigation of the epidemiology and natural history of the newly emerging disease, Acquired Immunodeficiency Syndrome (AIDS). The collected data is a source of valuable and extensive information on the AIDS epidemic in its earliest years (available from their website <http://socrates.berkeley.edu:7502/rdata/>). These data are used to illustrate realistically the discussed techniques. A "workbook" of non-computer problems is included to further explore the practical side of survival analysis. Finally, a bit of computer code is presented to give a sense of survival analysis software. The statistical analysis system called "R" is chosen because it is extensive, fully documented and both the software and documentation can be obtained without cost (<http://www.r-project.org/>).

Clearly many kinds of phenomena fail. The failure of equipment, machine components, numerous kinds of products and the structural integrity of various materials are frequently analyzed with survival analysis techniques (sometimes called time-to-failure data and methods). For the following description of survival analysis, however, the terminology is by and large in terms of human mortality (alive/dead). For example, rates are described in the context of mortality risk (risk of death). The language of human mortality was chosen strictly for simplicity. The theory and applications of the methods discussed are not affected. Using general terminology complicates the issues under study

and is avoided to more clearly focus on the statistical issues important in the analysis of epidemiologic and medical survival data. Also for simplicity, all confidence intervals are set at the 95% level of confidence. The general notation for a confidence interval is not used. That is, all confidence levels = $z_{1-\alpha/2} = z_{0.975} = 1.960$ for $\alpha = 0.05$ because the 95% level is used almost exclusively in published data.