Lecture 4:
Kaplan-Meier estimator of the survival function

Eben Kenah

STA 6177: Applied Survival Analysis
UF College of Public Health and Health Professions
September 16, 2014

1 Nonparametric survival function estimation

The cumulative distribution function (CDF) of a random variable X is the function

$$F(x) = \Pr(X \leq x). \tag{1}$$

If $X_1, \ldots, X_n$ are observations of the random variable X, the *empirical CDF* is the function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \leq x}, \tag{2}$$

where $\mathbf{1}_P = 1$ if P is true and $\mathbf{1}_P = 0$ if P is false. For a fixed value of x, $\hat{F}_n(x)$ is just the proportion of the observations that are $\leq x$. At each x, its expected value of $\hat{F}_n(x)$ is $F(x)$ and its variance is

$$\operatorname{Var} \hat{F}_n(x) = \frac{1}{n} F(x)\big(1 - F(x)\big). \tag{3}$$

As $n \to \infty$, we have

$$\frac{\sqrt{n}\big(\hat{F}_n(x) - F(x)\big)}{\sqrt{\hat{F}_n(x)\big(1 - \hat{F}_n(x)\big)}} \rightsquigarrow N(0, 1), \tag{4}$$

where $\rightsquigarrow$ denotes convergence in distribution and $N(0, 1)$ is the standard normal distribution. For large n, this gives us the pointwise 95% confidence interval

$$\hat{F}_n(x) \pm 1.96 \sqrt{\frac{1}{n} \hat{F}_n(x)\big(1 - \hat{F}_n(x)\big)} \tag{5}$$

for the true value of $F(x)$. To force the confidence interval to stay inside $(0, 1)$, we could use a logit or log-log transformation with the delta method. This is an estimate of a function $F(x)$, so it gives us a different point and interval estimate for each value of x.

1.1 Kaplan-Meier estimator

In right-censored data, we cannot calculate the empirical CDF directly. If we have a censored observation with survival time T_i, we do not know whether $F_i - O_i > t$ for $t > T_i$. However, it turns out that we can use conditional probabilities to estimate the survival function

$$S(x) = 1 - F(x) = \Pr(X > x). \tag{6}$$

For right-censored data, we have the survival time $T_i = \min(F_i, C_i) - O_i$ where O_i is the time origin, F_i is the failure time, and C_i is the censoring time. Let $T_i^F = F_i - O_i$ and $T_i^C = C_i - O_i$. Our event indicator $\delta_i = 1$ if $T_i = T_i^F$ and $\delta_i = 0$ if $T_i = T_i^C$. Let

$$Y_i(t) = \begin{cases} 1 & \text{if } i \text{ is at risk of failure and under observation at time } O_i + t, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

The set $\mathcal{R}(t) = \{i : Y_i(t) = 1\}$ is called the *risk set* at time t. The time scale used to define the risk set is *analysis time*, which starts at time O_i for each individual i. The risk set $\mathcal{R}(t)$ includes everyone who is censored at time t and everyone with a survival time greater than t.

Suppose we have right-censored survival data $(T_1, \delta_1), (T_2, \delta_2), \ldots, (T_n, \delta_n)$. Let $t_1 < t_2 < \ldots < t_m$ denote the distinct survival times in our data set (so $m \leq n$). Then

$$\Pr(T^F > t_2) = \Pr(T^F > t_2 | T^F > t_1) \Pr(T^F > t_1), \tag{8}$$

$$\Pr(T^F > t_3) = \Pr(T^F > t_3 | T^F > t_2) \Pr(T^F > t_2), \tag{9}$$

$$\vdots$$

$$\Pr(T^F > t_m) = \Pr(T^F > t_m | T^F > t_{m-1}) \Pr(T^F > t_{m-1}). \tag{10}$$

Under independent censoring, all of these conditional probabilities can be calculated using right-censored data.

Let $n_j = \sum_{i=1}^n Y_i(t_j)$ be the number of people in the risk set $\mathcal{R}(t_j)$ and let $d_j \geq 0$ be the number of failures that occur at time t_j. Then

$$\Pr(T^F > t_j) | T^F > t_{j-1}) = \frac{n_j - d_j}{n_j} = 1 - \frac{d_j}{n_j}, \tag{11}$$

For a general t, the Kaplan-Meier estimator of $S(t)$ is

$$\hat{S}(t) = \prod_{j : t_j \leq t} \Big(1 - \frac{d_j}{n_j}\Big). \tag{12}$$

It is named after American statisticians Edward L. Kaplan (1920–2006) and Paul Meier (1924–2011). Their 1958 paper that described this estimator[1] is the most-cited paper in statistics.

[1] Edward L. Kaplan and Paul Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53: 457–481. It has 44,254 citations on Google Scholar as of September 15.

Left-truncated data The Kaplan-Meier estimator extends easily to data that is left-truncated as well as right-censored. In right-censored data, the risk indicator $Y_i(t) = 1$ for all $t \in (0, T_i]$. In other words, person i is at risk and under observation from his or her time origin until his or her survival time. If person i is not under observation until time $A_i > O_i$, we simply let $Y_i(t) = 0$ for $t \in (0, A_i - O_i]$. Thus, person i is not included in risk sets $\mathcal{R}(t)$ for $t \leq A_i - O_i$.

1.2 Greenwood's formula and confidence intervals

Calculating the variance of a product is difficult, but calculating the variance of a sum is easy. Taking logarithms in equation (12), we get

$$\ln \hat{S}(t) = \sum_{j:t_j \leq t} \ln \hat{p}_j, \tag{13}$$

where $\hat{p}_j = 1 - \frac{d_j}{n_j}$ is the estimated probability of surviving from time t_{j-1} to time t_j. For each j, the estimated variance of $\hat{p}_j$ is

$$\mathrm{Var}(\hat{p}_j) = \frac{1}{n_j} \hat{p}_j \big(1 - \hat{p}_j\big). \tag{14}$$

By the delta method, we have $\mathrm{Var} f(X) \approx f'(\mathbb{E}X)^2\, \mathrm{Var}(X)$. Since $\ln x$ has the derivative $\frac{1}{x}$,

$$\mathrm{Var}(\ln \hat{p}_j) \approx \frac{1}{\hat{p}_j^2}\, \mathrm{Var}\big(\hat{p}_j\big) = \frac{d_j}{n_j(n_j - d_j)}. \tag{15}$$

Under independent censoring, the survival probabilities in each time interval are independent, so we have

$$\mathrm{Var}\big(\ln \hat{S}(t)\big) = \sum_{t_j \leq t} \mathrm{Var}(\ln \hat{p}_j) = \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{16}$$

Since $\hat{S}(t) = \exp\big(\ln \hat{S}(t)\big)$, we can use the delta method again to get an estimated variance for $\hat{S}(t)$. The function $\exp(x) = e^x$ is its own derivative, so we get

$$\mathrm{Var}\, \hat{S}(t) = \hat{S}(t)^2 \sum_{t_j \leq t} \mathrm{Var} \ln \hat{p}_j = \hat{S}(t)^2 \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}. \tag{17}$$

This is called the *Greenwood formula* for the variance of $\hat{S}(t)$. It is named after the English epidemiologist and statistician Major Greenwood (1880–1949).

For each t, a Wald 95% confidence interval for $S(t)$ is simply

$$\hat{S}(t) \pm 1.96 \sqrt{\mathrm{Var}\, \hat{S}(t)}. \tag{18}$$

This confidence interval can include values below zero or above one. Better confidence intervals are based on the log-log transformation and the delta method using $\mathrm{Var}\, \hat{S}(t)$ from equation (17).

1.3 Software

R In R, survival data is set up using the function Surv. For right-censored data, there are two possibilities. For purely right-censored data, use Surv(time, event) where time is the survival time and event is the event indicator. For data that is both right-censored and left-truncated, use Surv(time, time2, event) where time is the beginning of observation and time2 is the failure or censoring time. In both cases, the time origin is implicitly assumed to be time $= 0$. For left-truncated observations, time > 0.

Kaplan-Meier curves are calculated using the function survfit. For a single survival curve, the formula argument should be Surv(time, event) ~ 1 or Surv(time, time2, event) ~ 1, depending on whether the data are left-truncated. To get a Kaplan-Meier curve, use the option type = "kaplan-meier". To calculate a Wald confidence interval, use the option conf.type = "plain". To get log-log-transformed confidence intervals, use conf.type = "log-log". The default confidence interval is log-transformed (conf.type = "log"), which prevents a negative lower bound but can allow an upper bound > 1.

To plot the survival function, use plot(survfit.obj) where survfit.obj is the object produced by survfit. The option conf.int determines whether confidence intervals are plotted (it plots them by default when there is only one survival curve), and the option mark.time determines whether censored times are marked on the survival curve. Many other options are available, including all regular plot options (see help for plot.survfit).

SAS For right-censored data, PROC LIFETEST will calculate and plot a Kaplan-Meier curve. It cannot handle left-truncated data. The TIME statement on the line after the PROC LIFETEST statement is used to specify the survival time and the event indicator. The statement TIME T*delta(0) identifies TIME as the survival time and delta as the event indicator, with delta $= 0$ indicating a right-censored observation.

To get a Wald confidence interval, use the option CONFTYPE=LINEAR. To get a log-log confidence interval, use CONFTYPE=LOGLOG (this is the default). Logit-transformed and log-transformed confidence intervals are also available.

To plot the survival function, use the option PLOTS=SURVIVAL. To get pointwise confidence limits, use the CL option for the survival plot (PLOTS=SURVIVAL(CL). The NOCENSOR option for the survival plot suppresses plotting of censored failure times. Don't forget to turn ODS GRAPHICS on and off.

2 Parametric likelihoods

The Kaplan-Meier estimator is a nonparametric estimator of the survival function. In Lecture 3, we talked about fitting parametric likelihoods. This can be done in both SAS and R. When the parametric assumptions are valid, parametric models have several advantages over nonparametric models:

- They can easily handle left, right, and interval censoring as well as left truncation. The Kaplan-Meier estimator can only handle right censoring and left truncation.

- A parametric model fit gives you the entire failure time distribution, allowing the calculation of moments (means, variances, etc.) and quantiles (median, quartiles, etc.). A Kaplan-Meier curve gives you an estimate of the survival function only in the range of the available data.

- Parametric maximum-likelihood estimates are slightly more efficient than nonparametric estimates when the parametric assumptions are valid.

Despite these advantages, nonparametric estimators like the Kaplan-Meier survival curve used much more often than parametric models. It is difficult to know that your parametric assumptions are correct, and the efficiency advantage of parametric estimators is often very small.

R To fit the parametric distributions from Lecture 3 to right-censored data, use the `survreg` function. The `formula` argument should be `Surv(time, event) ~ 1` or `Surv(time, time2, event) ~ 1`. The distribution can be set using the `dist` argument, which can be set to exponential (`exponential`), Weibull (`weibull`), and log-logistic (`loglogistic`). Let $\hat{\beta}_0$ denote the estimated intercept term and $\hat{\sigma}$ denote the estimated scale parameter. The parameters of our distributions are estimated as follows:

- For the exponential distribution, the scale parameter is assumed to be one. The rate parameter $\hat{\lambda} = \exp(-\hat{\beta}_0)$.

- For the Weibull and log-logistic distributions, the estimated shape parameter is $\hat{\gamma} = \hat{\sigma}^{-1}$. The estimated rate parameter $\hat{\lambda} = \exp(-\hat{\beta}_0)$.

SAS To fit the parametric failure time distributions from Lecture 3 to right-censored data, use `PROC LIFEREG`. The model is specified in the `MODEL` statement. We are fitting models with no covariates, so the `MODEL` statement needs no covariates on the right-hand side of the equals sign. In the `MODEL` statement, you can specify a distribution using the `DISTRIBUTION` option. Exponential (`EXPONENTIAL`), Weibull (`WEIBULL`), and log-logistic (`LLOGISTIC`) distributions are all available. Let $\hat{\beta}_0$ denote the estimated intercept term and $\hat{\sigma}$ denote the estimated "scale parameter." The parameters of our distributions are estimated as follows from the regression output:

- For the exponential distribution, the scale parameter is assumed to be one. The rate parameter $\hat{\lambda} = \exp(-\hat{\beta}_0)$.

- For the Weibull and log-logistic distributions, the estimated shape parameter is $\hat{\gamma} = \hat{\sigma}^{-1}$. The estimated rate parameter $\hat{\lambda} = \exp(-\hat{\beta}_0)$.

Exercises

1. Using the data set `exponential.csv` on the class website:

   a. Fit an exponential distribution. Get a point estimate and 95% confidence interval for the rate parameter λ.

   b. Calculate and plot the survival function predicted by the exponential model.

   c. Calculate and plot the Kaplan-Meier survival curve with log-log 95% confidence limits. If possible, superimpose the plots from (b) and (c).

2. Using the data set `Weibull.csv` on the class website:

   a. Fit a Weibull distribution. Get point estimates and 95% confidence intervals for the shape parameter γ and the rate parameter λ.

   b. Calculate and plot the survival function predicted by the parametric model.

   c. Calculate and plot the Kaplan-Meier survival curve with log-log 95% confidence limits. If possible, superimpose the plots from (b) and (c).

3. Repeat Exercise 2 using the data set `loglogistic.csv` on the class website. Use a log-logistic model instead of a Weibull model.

4. Fit a Weibull distribution to the data set `exponential.csv`. Do we reject the null hypothesis that $\gamma = 1$ (i.e., the distribution is exponential)?

5. Fit an exponential distribution to the data set `loglogistic.csv`. Calculate and plot the survival function and compare it to the Kaplan-Meier curve. Is the exponential distribution a good fit for this data?

6. When there is no censoring, the Kaplan-Meier estimator $\hat{S}(t) = 1 - \hat{F}(t)$ where $\hat{F}(t)$ is the empirical CDF. Confidence intervals are also identical.

   a. Show that

   $$\hat{S}(t) = 1 - \hat{F}(t) \tag{19}$$

   under no censoring. (Hint: $d_{j+1} = n_j - d_j$ for all j.)

   b. Show that

   $$\mathrm{Var}\big(\hat{S}(t)\big) = \frac{\hat{S}(t)\big(1 - \hat{S}(t)\big)}{n} \tag{20}$$

   under no censoring. (Hint: $\frac{d_j}{n_j(n_j-d_j)} = \frac{n_j - n_{j+1}}{n_j n_{j+1}} = \frac{1}{n_{j+1}} - \frac{1}{n_j}$ for all j.)