18.465, further revised November 27, 2012

Survival analysis and the Kaplan–Meier estimator

1. Definitions

Ordinarily, an unknown distribution function F is estimated by an empirical distribution function F_n based on observations $X_1, \ldots, X_n$ i.i.d. (F). The survival function $S(t) \equiv 1 - F(t)$ is a nonincreasing function, approaching 0 as $t \to +\infty$. It would then be estimated by $S_n = 1 - F_n$.

In survival analysis, we assume that $F(0) = 0$, so that $S(0) = 1$ and $X_j > 0$ for all j. Here X_j is the time until an "endpoint" occurs for the jth individual. Some much-studied endpoints are deaths of humans or other organisms, or failures of manufactured devices. Or, one can consider the time (starting at time 0) to complete some given task.

In survival analysis, not all X_j may be observed. It's assumed that there are i.i.d. pairs (X_j, Y_j) of positive random variables where also each X_j is independent of Y_j and Y_j have another distribution G. (Venables and Ripley, p. 354, mention possibly less stringent assumptions with regard to the joint distributions of (X_j, Y_j).) If $Y_j < X_j$, that fact is observed but we have no information beyond that about the value of X_j. The Y_j are called "censoring times." They may be, for example, the times individuals are "lost to follow-up" such as by stopping their participation in a study, or stop working on a task without completing it. Or, because a study ended at a fixed time, or because each subject (patient) is only followed up for a fixed amount of time (such as 5 years) after treatment, endpoints after that time are not included in the analysis. Let $V_j := \min(X_j, Y_j)$ and $I_j = 1_{\{V_j = X_j\}}$. Thus $I_j = 0$ if the jth individual is censored $(Y_j < X_j)$ and 1 if an endpoint is observed for the individual $(X_j \le Y_j)$. What are observed are (V_j, I_j) for $j = 1, ..., n$. Individual j will be said to be "uncensored" at time t if $t < Y_j$ and to have "survived" beyond time t if $t < X_j$. Both hold if $V_j > t$.

There are two commonly used models for a sequence of times $t_1 < t_2 < \cdots$ with $t_1 > 0$ at which observations occur. In what I'll call *Model I*, t_j are non-random times, usually equally spaced, with $t_j - t_{j-1} = \Delta$ for $j \ge 1$, where $t_0 := 0$, so that $t_j = j\Delta$ for $j = 0$ up to some finite integer. For example, $\Delta = 1$ year for life insurance actuarial tables, $\Delta = 1$ month for some medical studies where patients in a study are seen once a month, $\Delta = 1$ day for some mortality studies of insects, and Δ may be a short time for some studies of bacteria.

In Model II, the t_i for $i \geq 1$ are the random times when endpoints are observed to occur, namely, some $V_j = X_j = t_i$, so that $I_j = 1$. The t_i are arranged in increasing order (more than one endpoint may be observed at a given t_i). Here again $t_0 = 0$ by definition.

To define some notation, the cardinality (number of elements) of a finite set S will be denoted by $|S|$. In either model, let $M_i := \{j : V_j > t_{i-1}\}$ and $m_i := |M_i|$, i.e. the number of individuals who survived, uncensored, beyond time t_{i-1}. By the assumptions, $m_1 = n$. Let $S_i = \{j \in M_i : \ V_j \leq t_i\}$ and $s_i = |S_i|$. Then $m_{i+1} = m_i - s_i$. Let $D_i := \{j \in S_i : \ I_j = 1\}$, $d_i = |D_i|$, $C_i := \{j \in S_i : \ V_j < t_i, \ I_j = 0\}$, $c_i = |C_i|$, and $C_i' := \{j \in S_i : \ V_j = t_i, \ I_j = 0\}$, $c_i' = |C_i'|$. For $j \in C_i'$ we know that $X_j > t_i$.

For $i \geq 1$ we have $s_i \equiv c_i + c_i' + d_i$. Let $n_i := m_i - c_i$. Thus: n_i is the number of individuals who have survived beyond t_{i-1} and have not been censored before time t_i. Then n_i is called the number of individuals *at risk* at time t_i.

In either model, m_i is already known from observations up to and at time t_{i-1}. In Model I, the observer learns at time t_i just the values (c_i, c_i', d_i).

In Model II, each (V_j, I_j) is observed at time V_j. For $t_{i-1} < V_j < t_i$, we have $I_j = 0$ and $j \in C_i$, and for $j \in D_i$, $V_j = X_j = t_i$, both by definition of Model II. (The observed values of $V_j = Y_j$ for $j \in C_i$ are not useful in estimating $S(t)$.)

Now, in either model, we want to estimate the conditional probability

$$p_i := P(X_1 > t_i | X_1 > t_{i-1}) = S(t_i)/S(t_{i-1}).$$

Observations on individual j are useful in the estimation if we know that $X_j > t_{i-1}$ and whether or not $X_j > t_i$.

For the m_{i+1} values of j with $V_j > t_i$ and the c_i' with $j \in C_i'$ we know that $X_j > t_i > t_{i-1}$.

For $j \in D_i$ we know that $t_{i-1} < X_j \leq t_i$.

For $j \in C_i$, or for j with $V_j = t_{i-1}$ and $I_j = 0$, we know that $X_j > t_{i-1}$ but don't know whether $X_j > t_i$.

For all other values of j, either we know that $X_j \leq t_{i-1}$ or don't know whether this is true or not.

Thus we have $m_{i+1} + c_i' + d_i = n_i$ useful values of j, and the fraction of these that survived beyond time t_i is $(n_i - d_i)/n_i$, if $n_i > 0$, so that gives the estimate

$$\hat{p}_i = \frac{n_i - d_i}{n_i}, \quad n_i > 0.$$

Thus for $r = 1, 2, \ldots$ such that $S(t_{r-1}) > 0$,

$$S(t_r) \;=\; \prod_{i=1}^{r} \frac{S(t_i)}{S(t_{i-1})} \;=\; \prod_{i=1}^{r} p_i \tag{1}$$

is estimated by the *Kaplan–Meier (KM) estimator*, defined if $n_r > 0$,

$$\widehat{S}(t_r) = \prod_{i=1}^{r} \frac{n_i - d_i}{n_i}. \tag{2}$$

Also, $\widehat{S}(t)$ is defined as $\widehat{S}(t_r)$ for $t_r \le t < t_{r+1}$.

Sometimes, it's assumed that the Y_j (and/or X_j) have a continuous distribution, in which case each c_i' would equal 0 with probability 1 in Model I. But it seems that $j \in C_i'$ in Model I occurs in real-life situations. Suppose a research study is done with follow-up of human subjects by monthly or yearly telephone calls. A subject in the study who had not yet reached an endpoint might, just at the time of getting a follow-up call, decide not to cooperate in answering questions and so censor themselves out of further participation.

1.1. **Examples.** Wikipedia's article "Kaplan–Meier estimator" shows on one graph two Kaplan–Meier estimates of survival for "Gene A signature" and "Gene B signature." The graphs are shown only up to less than three and a half years. The steps downward in either graph indicate one or more deaths. The vertical bars through the graph where there is no step downward indicate censoring times. In this study it seems that about 10 of the individuals were lost to follow-up before the study was over. Mortality appears to have been more severe in the Gene B group than in the Gene A group. As numbers are not given, p-values cannot be computed from the graphs.

The Garber et al. paper (the source of "lung") in Fig. 4 on p. 13788 shows in one graphic three Kaplan–Meier estimators for Adeno groups 1, 2, and 3 as found by hierarchical clustering in the study, The follow-up time was evidently 5 years (60 months) as is often true in medical studies. So, all censoring times $Y_j = 60$ months. There are no vertical bars indicating earlier censoring. All 6 patients in Adeno Group 2 survived the full 5 years. None of the 9 patients in Group 3 survived more than 14 months after their biopsy. For the 16 patients in Group 1, only about half survived for 3 years, but then they also survived through 5 years. One can see that there was mortality soon after the biopsy for a few patients in Groups 1 and 3.

2. Life tables

. In this application of survival analysis, the sets of individuals observed may be disjoint for different values of i. Let $t_i = i$ in years (so this is a Model I situation). Suppose there are m_i individuals in a certain population who at some time (say, early January of some year) have reached age $i-1$ but have not yet reached age i, for each $i = 1, \ldots, 99$ say. The sets S_i with m_i members are disjoint, as people in different sets have different ages. Suppose that one year later, d_i of the m_i are known to have died and for c_i, it is not known whether they survived (e.g., they may have emigrated from the area where study data are being gathered). Let $n_i = m_i - c_i$. Let B_i be the event that an individual lives until the January after their $(i-1)$st birthday and $S(i) = P(B_i)$. Here let's assume that $S(0) = 1$, i.e. consider only the population alive in some January, omitting infants who did not live until the January after they were born. Then for the given population, the conditional probability $P(B_{i+1}|B_i)$, namely $S(i)/S(i-1)$, can again be estimated by $(n_i - d_i)/n_i$ provided that $n_i > 0$. Suppose that $n_i > 0$ for all $i = 1, \ldots, r$. Then the probability $P(B_{r+1})$ for an infant alive in January of age less than 1 to survive until January after their rth birthday can be estimated by the Kaplan–Meier estimator (2), based on data collected during only a little more than one year on the population of all ages up to r. (Whether being censored, i.e. being one of the c_i, is actually independent of survival, could be questioned.)

For a large value of i such that n_i is small, it could happen that $d_i = n_i$, so that $\widehat{S}(t) = 0$ for $t \geq i$, even though possibly $n_{i'} > 0$ for some $i' > i$, so we know from the data that $S(t) > 0$ for all $t < i'$. For example, it could happen that the (few) people in the set S_{107} of people of age 107 at the beginning of the year all passed away, so that $d_{107} = n_{107}$, there could well have been people of age 108 or more (some sets S_i for $i \geq 108$ were not empty). Among Medicare enrollees born in the years 1872 through 1875, 100 reached age 109 or more (Bayo and Faber, 1983).

3. Greenwood's variance estimator

Greenwood in 1926 (as mentioned by Wikipedia) gave an estimator of the variance of $\widehat{S}(t)$. Here two formulas will be given, one for $\log(\widehat{S}(t))$ and the other for $\widehat{S}(t)$ itself. Under suitable hypotheses, to be seen in the derivation, in Model I, for $t > 0$, $\log(\widehat{S}(t))$ is approximately normal

with distribution

(3)

$$\log(\widehat{S}(t)) \ \sim \ N(\log(S(t)), \hat{\sigma}_l^2(t)), \quad \text{where} \ \ \hat{\sigma}_l^2(t) = \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}.$$

Under the same or possibly stronger hypotheses, $\widehat{S}(t)$ is approximately normal with mean $S(t)$ and variance

$$\widetilde{\sigma}^2(t) = S(t)^2 \hat{\sigma}_l^2(t), \tag{4}$$

which then can be estimated by

$$\hat{\sigma}^2(t) = \widehat{S}^2 \hat{\sigma}_l^2(t) = \widehat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}. \tag{5}$$

If $\widehat{S}(t) = 0$, formula (5) is not useful. A non-rigorous derivation of the approximations and formulas will be given.

Once one has the approximation (3), one gets (4) by exponentiating both sides and applying the delta-method, then (5) by plugging in the estimator $\widehat{S}(t)$ of $S(t)$. From (5), one can get a confidence interval for $S(t)$ given $\widehat{S}(t)$, symmetric around $\widehat{S}(t)$, with endpoints

$$\widehat{S}(t) \pm z_{\alpha/2} \hat{\sigma}(t) \tag{6}$$

where for $\alpha = 0.05$ we have as usual $z_{0.025} \doteq 1.96$.

Another way to proceed, giving the "logarithmic" confidence intervals which R "survival" uses as the default, is to apply (3) directly, to give a $1 - \alpha$ confidence interval for $\log(S(t))$ with endpoints

$$\log(\widehat{S}(t)) \pm z_{\alpha/2} \hat{\sigma}_l. \tag{7}$$

Then one takes the exponentials of these endpoints to get endpoints $[a_l, b_l]$ of a confidence interval for $S(t)$,

$$a_l = \widehat{S}(t) \exp(-z_{\alpha/2} \hat{\sigma}_l), \quad b_l = \widehat{S}(t) \exp(z_{\alpha/2} \hat{\sigma}_l). \tag{8}$$

Although the interval with endpoints (7) for $\log(S(t))$ is symmetric, the one (8) for $S(t)$ will not be in general. The logarithmic confidence interval may be preferred to the classical Greenwood one with endpoints (6) as it avoids the further approximations involved in getting first (4), then (5).

An analogous situation arises in finding confidence intervals for odds ratios. It is treated in the 18.443 handout deltameth-oddsratios.pdf from spring 2012 (although not in the abridged version on the 18.465 website). Namely, one shows that the log of an odds ratio is approximately normal if all four entries n_{ij} in a contingency table are large

enough, by the delta-method. The normal distribution has mean the log of the true log odds ratio and variance the sum of $1/n_{ij}$, which is observed. Thus one gets a confidence interval for the log of the true odds ratio. One takes the exponential of the endpoints to get a confidence interval for the true odds ratio. Although odds ratios themselves may be approximately normal under some conditions, that is not the usual way to treat them even for large n_{ij}.

A short outline of a derivation of Greenwood's formulas is given in Cox and Oakes (1984, pp. 50–51). The idea of taking a logarithm of the product to get a sum, finding variances of the summands, adding them, and then exponentiating again, is indicated there, under an assumption that one is an an asymptotic situation where all d_i are large. Greenwood was considering life tables, a special case of Model I. Here a somewhat different rationale will be given, but still only applying to Model I.

It's possible that a random variable X is approximately $N(\mu, \sigma^2)$ even if X has infinite variance, for example if the distribution of X is "contaminated normal," namely $\lambda N(\mu, \sigma^2) + (1-\lambda)Q$ where $0 < \lambda < 1$, λ is close to 1, and $\int_{-\infty}^{\infty} x^2 dQ(x) = +\infty$.

A derivation of the formulas is as follows. To see if it is approximately correct in a given case one would need to check that each of the approximations used, including applications of the delta-method, is reasonably valid.

From (2) we have for $t_r \le t < t_{r+1}$ that

$$\log \widehat{S}(t) = \log \widehat{S}(t_r) \;=\; \sum_{i=1}^{r} \log\left(\frac{n_i - d_i}{n_i}\right). \tag{9}$$

Here $n_i - d_i$ has a binomial (n_i, p_i) distribution given n_i where $p_i = S(t_i)/S(t_{i-1})$. This distribution has variance $n_i p_i q_i$ where $q_i = 1 - p_i$. The argument assumes that the binomial distribution is approximately normal, which is valid if $n_i p_i q_i$ is reasonably large, say, at least 5. Plugging in estimates of the unknown probabilities, one would like that $d_i(n_i - d_i)/n_i \ge 5$. It's necessary, but not sufficient, for this that $n_i \ge 20$, as $x(n_i - x)$ is maximized for $x = n_i/2$ where it equals $n_i^2/4$.

In Model II, $d_i \equiv 1$, so $d_i(n_i - d_i)/n_i < 1$ and cannot be ≥ 5.

If the approximate normality of $n_i - d_i$ holds, then $(n_i - d_i)/n_i$ is, given n_i, approximately $N(p_i, p_i q_i/n_i)$. In more detail, we have for $n_i > 0$

$$\frac{n_i - d_i}{n_i} = \frac{n_i p_i + (n_i q_i - d_i)}{n_i} = p_i - r_i$$

where $r_i := (d_i/n_i) - q_i$ has $E(r_i|n_i) = 0$ and $\mathrm{Var}(r_i|n_i) = p_i q_i/n_i$. We also have

$$L_i := \log\left(\frac{n_i - d_i}{n_i}\right) = \log(p_i - r_i) = \log p_i + \log\left(1 - \frac{r_i}{p_i}\right) \tag{10}$$

where

$$\log\left(1 - \frac{r_i}{p_i}\right) = -\frac{r_i}{p_i} + O\left((r_i/p_i)^2\right) = -\frac{r_i}{p_i} + O_p(1/(n_i p_i)). \tag{11}$$

This is writing out some details of the delta-method in this case, which gives that $\sqrt{n_i}(L_i - \log p_i)$ given n_i has distribution $N(0, q_i/p_i)$, approximately, with an error of order $O_p(1/\sqrt{n_i p_i})$. (Errors come not only from $O_p(1/(n_i p_i))$ at the end of (11) but the fact that a binomial random variable is only approximately normal.) This approximation also is only useful for n_i large enough. Plugging in the estimates of p_i and q_i (another approximation) gives that L_i is approximately

$$N\left(\log p_i, \frac{p_i q_i}{n_i} \cdot \frac{n_i^2}{(n_i - d_i)^2}\right) \doteq N\left(\log p_i, \frac{d_i}{n_i(n_i - d_i)}\right).$$

Next we need to see why the variance of a sum is approximately the sum of variances of the terms. Cox and Oakes (1984, p. 50) argue that the terms are "asymptotically independent," but they are not exactly independent, as the value of n_{i+1} depends on n_i and d_i as well as other variables. What needs to be shown is that the covariances of the terms are approximately negligible relative to the variances (which are themselves small). The covariance of L_i and L_{i+1} given n_{i+1} large enough is approximately that of r_i/p_i and r_{i+1}/p_{i+1} by (10) and (11). Let $\mathcal{F}_i$ be the set of random variables (V_j, I_j) for $V_j \le t_{i-1}$ and also n_i. Then r_i/p_i is a function of the random variables in $\mathcal{F}_{i+1}$, and r_{i+1}/p_{i+1} has conditional expectation 0 given that $n_{i+1} > 0$, also given $\mathcal{F}_{i+1}$, so by a fact about conditional expectations, that of the product $(r_i/p_i) \cdot (r_{i+1}/p_{i+1})$ is also 0, as desired.

At this point, we have given a derivation of formula (3).

Using the delta-method again, now for the exponential function, one gets that the approximation (4) holds. Then plugging in the estimate $\widehat{S}(t)$ in place of $S(t)$ gives Greenwood's (approximate) formula for the variance (5). Also, as $\prod_{i=1}^{r} p_i = S(t)$ by (1), the argument, to the extent it is valid, shows that $\widehat{S}(t)$ is approximately an unbiased estimator of $S(t)$ since the approximating normal distribution has mean $S(t)$.

The approximate normality in (5) can only hold if $\widehat{S}(t)$ is distant from 0 and 1 by at least $2\hat{\sigma}(t)$, say.

Even if n_i is large, there is a strictly positive, though possibly small, probability that $d_i = n_i$ and then $\log((n_i - d_i)/n_i)$ would usually be defined as $-\infty$. So this random variable always has infinite variance. Moreover, if for example $d_r = n_r$, then $\widehat{S}(t_r) = 0$, but the conditions mentioned for normal approximation of the binomial distribution are definitely violated, and use of $\hat{\sigma}(t_r) = 0$ as a standard error would lead to a confidence interval of $[0, 0]$, in principle with arbitrarily high confidence such as 0.999. But in fact we are not at all sure that $S(t) = 0$. This is more complex but essentially the same as using the plug-in confidence interval for a binomial success probability p when we observe no successes, which also is $[0, 0]$ but is unjustified.

To estimate an unknown binomial success probability p given that one has observed X successes in n independent trials, one can try to respect the conditions on n and X for different approximations of binomial probabilities. For example, one can use a normal approximation if $X(n-X)/n$ is large enough, say, at least 5. If n is large but $X(n-X)/n$ is not, one can use a Poisson approximation to either the distribution of X if X is not large, or to that of $n - X$ if it is not large. If n is not large, say $n \leq 19$, one can just give a table for $0 \leq X \leq n/2$, using symmetry, of adjusted "Clopper–Pearson" confidence intervals. Such a procedure is given with details in the handout binomial.pdf on the 18.443 Spring 2012 website. The properties of the method are not yet fully clear, and an extension of the method to survival analysis has not yet been worked out,

4. Bias of the Kaplan–Meier estimator

. It had been claimed in older literature that the estimator is unbiased, but that is not correct, although it is nearly unbiased under some conditions.

Suppose that the survival and censoring distributions are both concentrated in a finite set $t_1 < t_2 < ... < t_k$. If n_j is the number of individuals alive and being observed just before t_j, and d_j the number of deaths observed at t_j, then the conditional probability p_j of dying at t_j given that one survived past t_{j-1} has an unbiased estimator d_j/n_j, IF n_j is not 0. By iterated conditional expectation, one gets a proof that the KM estimator of the survival function is unbiased at each t, which is actually conditional on the event A_t that there is at least one individual alive and uncensored at all times less than t. On the complementary event A_t^c that there are no such individuals, it seems that there is no way to give an unbiased estimator. The event A_t^c has probability > 0, though possibly small, for $t > t_1$ if $F(t_1) > 0$ or $G(t_1) > 0$.

$\widehat{S}(t)$ will have a constant value for all $t \geq T$ where T is the largest t_j in Model II, or the largest t_j such that $d_j > 0$ in Model I. This constant value will be 0 if the largest V_j has $I_j = 1$ and will be larger than 0 otherwise.

Under conditions implying that A_t^c has small probability, such as $F(t)$ and $G(t)$ well away from 1 and n large, it has been shown that the bias of the KM estimator, although non-zero, is small: Zhou (1988), see also Stute (1994).

5. The Kaplan–Meier estimator and confidence intervals in R

The package "survival" in R can be loaded by
> library(survival)
Given survival data (V_j, I_j), $j = 1, ..., n$, create the two vectors $v = (V_1, ..., V_n)$ and ind $= (I_1, ..., I_n)$. Then
> Surv(v,ind)
will give a vector of the numbers V_j, followed by a + sign if $I_j = 0$. This means we know for such j about X_j only that $X_j > V_j = Y_j$. If there is no + sign it means we observe the actual $X_j = V_j$, so $I_j = 1$ (see the example in Venables and Ripley, p. 537, surv(time, cens)). The V_j are not sorted and don't need to be (in further programs, R will sort them).

Sometimes one tries to give confidence bands $[g, h]$ for an unknown function, such that one has confidence $1 - \alpha$ that the graph of the function f on an interval $[a, b]$ satisfies $g(t) \leq f(t) \leq h(t)$ for all $t \in [a, b]$. The confidence intervals for the survival function $S(t)$ are only for one t at a time.

On the last page of this handout is a plot of a Kaplan–Meier (KM) estimator (dashed line) with upper and lower confidence bounds (dotted lines). The confidence intervals are logarithmic, as given by (8).

In the example $n = 30$. To compute the value of the KM estimator at 1.5 (which one sees in the plot is roughly 0.1) I first looked at fit$times to see that $V_{(28)} < 1.5 < V_{(29)}$. So the value is KM(1.5) = fit$surv[28] = 0.1101. The upper endpoint of the confidence interval is fit$upper[28] = 0.3194 and the lower endpoint is fit$lower[28] = 0.037965.

By the way $V_{(28)}$ is a censoring time shown by a vertical mark through the graph, horizontal at that point. At endpoint times the graph takes steps downward. Only three censoring times are shown in the graph.

Similarly, for 2 I found that $V_{(29)} < 2 < V_{(30)}$ and so KM(2) = fit$surv[29] = 0.05506, with lower endpoint of the confidence interval fit$lower[29] = 0.009589, which is small but not 0, and upper endpoint

fit$upper[29] = 0.31614, only slightly different from the previous upper endpoint.

At $V_{(30)} = 2.26$, the last surviving uncensored individual has an endpoint, the KM estimator becomes 0 and no further confidence bounds are given.

REFERENCES

Bayo, F. R., and Faber, J. F., "Mortality experience around age 100," *Trans. Soc. Actuaries* **35**, 37–59, with Discussion pp. 60–64.

Cox, D. R., and Oakes, D., *Analysis of Survival Data*, Chapman and Hall, London, 1984.

Stute, W., "The bias of Kaplan-Meier integrals," *Scandinavian Journal of Statistics* **21** (1994), 475–484.

Zhou, M., "Two-sided bias bound of the Kaplan-Meier estimator," *Probability Theory and Related Fields* **79** (1988), 165–173.