The *survival function* is one of the basic tools used in assessing time-to-event data and comparing survival curves across groups. The following sections define the survival function, present the Kaplan-Meier estimator and present and justify Greenwood's formula for the approximate variance of the KM estimator.

# 1. The Survival Function

Survival (time-to-event) data is relevant for research where the outcome variable is the time $T$ until an event occurs. The survival function is defined by

$$S(t) = P(T > t). \tag{1}$$

Intuitively, $S(t)$ describes the fraction of subjects for whom the survival time is greater than or equal to $t$. In general, $S(0) = 1$ and $S(t)$ then decreases along the time axis.

# 2. Survival Data

For the $i$th subject we record the time $t_i$ that the subject is observed, an indicator $\delta_i$ of whether the observation time ended with an event, and some potential explanatory variables. If the $i$th subject has an event, then $\delta_i = 1$, otherwise $\delta_i = 0$. Note that if the subject has an event, then $t_i$ is equal to the survival time and $\delta_i = 1$; but if the subject does not have an event, then all we know is that $t_i \leq$ the survival time.

# 3. The Kaplan-Meier Estimator

Let $S(t)$ denote the survival function for some population. Kaplan and Meier (1959) proposed a simple non-parametric estimator $\hat{S}(t)$ of the survival function $S(t)$. Their idea was to look at the distinct times when events were recorded in the data. At each such time, they reduce the survivor function, multiplying it by the fraction of subjects who were still under observation (also known as *at risk*) at the time but did not have an event.

We need some notation to define the estimator. Suppose we denote the event times by $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. We write the index in parentheses

as a reminder that $t_{(j)}$ is the $j$th smallest event time; it is *not* the follow-up time for the $j$th subject, which we denote by $t_j$, without the parentheses. Look at all the subjects for whom $t_i \geq t_{(j)}$. These subjects could have had an event at time $t_{(j)}$ and we call them the *Risk Set* at that time, which we label by $R_j$. It is important to know how many people were at risk and we denote that number by $r_j$. Finally, let $d_j$ denote the number of subjects who had events at time $t_{(j)}$.

The idea of the Kaplan-Meier estimator is to keep $\hat{S}(t)$ constant between the event times and then to reduce $\hat{S}(t)$ at the event times by the probability of surviving at time $t_{(j)}$. At that time, there are $r_j$ subjects at risk for an event and $d_j$ actually have an event. So the natural estimator of the immediate event probability is $\hat{q}_j = d_j/r_j$ and the estimator of the survival probability is $1 - \hat{q}_j$.

The Kaplan-Meier estimator starts off at 1, i.e. $\hat{S}(0) = 1$ and it remains equal to 1 until the first event time. Then at time $t_{(1)}$, $\hat{S}(t)$ drops to $1 - \hat{q}_1$. It remains at that level until time $t_{(2)}$, when there is a further reduction by a factor of $1 - \hat{q}_2$, so that $\hat{S}(t_{(2)}) = (1 - \hat{q}_1)(1 - \hat{q}_2)$. The estimator remains at this level until the next event time, $t_{(3)}$, when there is a further reduction by a factor of $1 - \hat{q}_3$. Continue this process throughout the set of event times.

The general formula for $\hat{S}(t)$ involves products of the terms $1 - \hat{q}_j$. The only question is which terms to include in the product. The answer is that, for any fixed time $t$, the relevant reduction factors are the ones for which $t_{(j)} \leq t$. This gives the general formula

$$\hat{S}(t) = \prod_{j:t_{(j)} \leq t} (1 - \hat{q}_j) \tag{2}$$

# 4. The Standard Error of the KM Estimator

As with all statistical estimators, it is important to assess the precision of $\hat{S}(t)$ – what is its standard error? The first step in calculating a standard error is to determine the variance of the estimator; then take a square root to get the standard error. Statisticians are very good at computing the variance of sums; but the KM estimator is a product, not a sum, and we have no direct methods for working with products. The obvious solution: take logarithms and convert the product to a sum.

Solving one problem creates another: we will get $\mathrm{Var}[\ln \hat{S}(t)]$ and not

$\text{Var}[\hat{S}(t)]$. So we need to know what is the relationship between these two variances. Here an idea known as the *delta method* is helpful. This idea tells us, approximately, what is $\text{Var}[g(\hat{\theta})]$, when $\hat{\theta}$ is an estimator of a parameter $\theta$ and $g$ is a function (in our case the natural logarithm). There is one special case where we can give an exact answer and that is when $g(\hat{\theta})$ is a linear function $g(\hat{\theta}) = a + b\hat{\theta}$. Then we know that $\text{Var}[g(\hat{\theta})] = b^2\text{Var}(\hat{\theta})$. What if $g$ is not linear? Then we *approximate* it by a linear function: $g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$, where $g'$ is the derivative of the function $g$. Those who remember some calculus should recognize this as the *Taylor series expansion* of the function $g(\hat{\theta})$ about the value $\theta$. This leads to an approximate formula for $\text{Var}[g(\hat{\theta})]$,

$$\text{Var}[g(\hat{\theta})] \approx [g'(\theta)]^2\text{Var}(\hat{\theta}). \tag{3}$$

The formula requires us to evaluate the derivative at $\theta$, the true but unknown value of the parameter. So we make a further approximation, replacing $\theta$ in the formula by $\hat{\theta}$, to get

$$\text{Var}[g(\hat{\theta})] \approx [g'(\hat{\theta})]^2\text{Var}(\hat{\theta}). \tag{4}$$

Let's look at a specific and relevant example. Suppose the parameter $\theta$ that interests us is a probability, which we denote by $p$, that we have an estimator $\hat{p}$ and that we want to know the variance of $\ln(\hat{p})$. The derivative of $\ln(\hat{p})$ is $1/\hat{p}$. So the approximation formula (4) tells us that

$$\text{Var}[\ln(\hat{p})] \approx [1/\hat{p}]^2\text{Var}(\hat{p}). \tag{5}$$

From here we conclude that

$$\text{Var}(\hat{p}) \approx \hat{p}^2\text{Var}[\ln(\hat{p})]. \tag{6}$$

Now use result (6) when the probability we want to estimate is $S(t)$ and we use KM to estimate it by $\hat{S}(t)$. We find

$$\text{Var}[\hat{S}(t)] \approx [\hat{S}(t)]^2\text{Var}[\ln(\hat{S}(t))]. \tag{7}$$

The next step is to find $\text{Var}[\ln(\hat{S}(t))]$. We use the fact that

$$\ln(\hat{S}(t)) = \sum_{j:t_{(j)} \le t} \ln(1 - \hat{q}_j). \tag{8}$$

Each term in the sum is independent of the other terms, as it relates to the results at a new event time. So we can compute the variance of the sum as the sum of the variances,

$$\text{Var}[\ln(\hat{S}(t))] = \sum_{j:t_{(j)} \leq t} \text{Var}[\ln(1 - \hat{q}_j)]. \tag{9}$$

To find $\text{Var}[\ln(1-\hat{q}_j)]$ we again use the delta method. Note that $\hat{q}_j$ is a simple proportion (the proportion of those in the current risk set with an event at time $t_{(j)}$), so we can easily estimate its variance. We have $\text{Var}(1 - \hat{q}_j) = \text{Var}(\hat{q}_j) \approx \hat{q}_j(1 - \hat{q}_j)/r_j$. Using this along with the delta method gives

$$\text{Var}[\ln(1 - \hat{q}_j)] \approx [1/(1 - \hat{q}_j)]^2 \hat{q}_j(1 - \hat{q}_j)/r_j = \hat{q}_j/[r_j(1 - \hat{q}_j)]. \tag{10}$$

Insert this in equation (9) to get

$$\text{Var}[\ln(\hat{S}(t))] \approx \sum_{j:t_{(j)} \leq t} \hat{q}_j/[r_j(1 - \hat{q}_j)]. \tag{11}$$

Finally, use relation (7) along with (11) to conclude that

$$\text{Var}[\hat{S}(t)] \approx [\hat{S}(t)]^2 \sum_{j:t_{(j)} \leq t} \hat{q}_j/[r_j(1 - \hat{q}_j)]. \tag{12}$$

This is known as *Greenwood's formula* for the variance of the Kaplan-Meier estimator and it is the formula used by SPSS to compute the standard errors in the output.

# 5. References

Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 475-481.