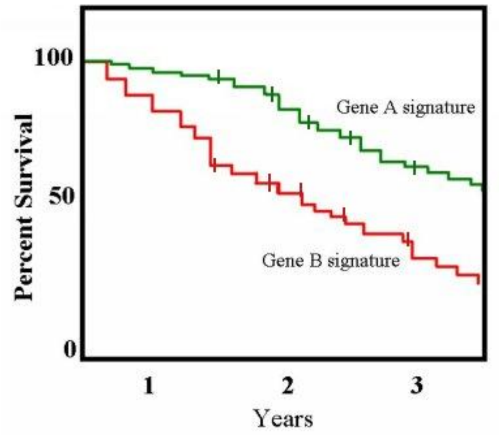
WIKIPEDIA

Kaplan–Meier estimator

The **Kaplan–Meier estimator**,[1][2] also known as the **product limit estimator**, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. In other fields, Kaplan–Meier estimators may be used to measure the length of time people remain unemployed after a job loss,[3] the time-to-failure of machine parts, or how long fleshy fruits remain on plants before they are removed by frugivores. The estimator is named after Edward L. Kaplan and Paul Meier, who each submitted similar manuscripts to the *Journal of the American Statistical Association*. The journal editor, John Tukey, convinced them to combine their work into one paper, which has been cited about 55,000 times since its publication.[4][5]



An example of a Kaplan–Meier plot for two conditions associated with patient survival.

The estimator of the survival function $S(t)$ (the probability that life is longer than t) is given by:

$$\widehat{S}(t) = \prod_{i:\ t_i \le t} \left(1 - \frac{d_i}{n_i}\right),$$

with t_i a time when at least one event happened, d_i the *number of events* (e.g., deaths) that happened at time t_i and n_i the *individuals known to have survived* (have not yet had an event or been censored) up to time t_i.

Contents

Basic concepts

A plot of the Kaplan–Meier estimator is a series of declining horizontal steps which, with a large enough sample size, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant.

An important advantage of the Kaplan–Meier curve is that the method can take into account some types of censored data, particularly *right-censoring*, which occurs if a patient withdraws from a study, is lost to follow-up, or is alive without event occurrence at last follow-up. On the plot, small vertical tick-marks indicate individual patients whose survival times have been right-censored. When no truncation or censoring occurs, the Kaplan–Meier curve is the complement of the empirical distribution function.

In medical statistics, a typical application might involve grouping patients into categories, for instance, those with Gene A profile and those with Gene B profile. In the graph, patients with Gene B die much more quickly than those with Gene A. After two years, about 80% of the Gene A patients survive, but less than half of patients with Gene B.

In order to generate a Kaplan–Meier estimator, at least two pieces of data are required for each patient (or each subject): the status at last observation (event occurrence or right-censored) and the time to event (or time to censoring). If the survival functions between two or more groups are to be compared, then a third piece of data is required: the group assignment of each subject.[6]

Problem definition

Let $\tau \geq 0$ be a random variable, which we think of as the time until an event of interest takes place. As indicated above, the goal is to estimate the survival function S underlying τ. Recall that this function is defined as

$$S(t) = \mathrm{Prob}(\tau > t), \text{ where } t = 0, 1, \ldots \text{ is the time.}$$

Let $\tau_1, \ldots, \tau_n \geq 0$ be independent, identically distributed random variables, whose common distribution is that of τ: τ_j is the random time when some event j happened. The data available for estimating S is not $(\tau_j)_{j=1,\ldots,n}$, but the list of pairs $\left((\tilde{\tau}_j, c_j)\right)_{j=1,\ldots,n}$ where for $j \in [n] := \{1, 2, \ldots, n\}$, $c_j \geq 0$ is a fixed, deterministic integer, the **censoring time** of event j and $\tilde{\tau}_j = \min(\tau_j, c_j)$. In particular, the information available about the timing of event j is whether the event happened before the fixed time c_j and if so, then the actual time of the event is also available. The challenge is to estimate $S(t)$ given this data.

Derivation of the Kaplan–Meier estimator

Here, we show two derivations of the Kaplan–Meier estimator. Both are based on rewriting the survival function in terms of what is sometimes called **hazard**, or **mortality rates**. However, before doing this it is worthwhile to consider a naive estimator.

A naive estimator

To understand the power of the Kaplan–Meier estimator, it is worthwhile to first describe a naive estimator of the survival function.

Fix $k \in [n] := \{1, \ldots, n\}$ and let $t > 0$. A basic argument shows that the following proposition holds:

> **Proposition 1:** If the censoring time c_k of event k exceeds t ($c_k \geq t$), then $\tilde{\tau}_k = t$ if and only if $\tau_k = t$.

Let k be such that $c_k \geq t$. It follows from the above proposition that

$$\operatorname{Prob}(\tau_k \geq t) = \operatorname{Prob}(\tilde{\tau}_k \geq t).$$

Let $X_k = \mathbb{I}(\tilde{\tau}_k \geq t)$ and consider only those $k \in C(t) := \{1 \leq k \leq n \,:\, c_k \geq t\}$, i.e. the events for which the outcome was not censored before time t. Let $m(t) = |C(t)|$ be the number of elements in $C(t)$. Note that the set $C(t)$ is not random and so neither is $m(t)$. Furthermore, $(X_k)_{k \in C(t)}$ is a sequence of independent, identically distributed Bernoulli random variables with common parameter $S(t-1) = \operatorname{Prob}(\tau \geq t)$. Assuming that $m(t) > 0$, this suggests to estimate $S(t-1)$ using

$$\hat{S}_{\text{naive}}(t-1) = \frac{1}{m(t)} \sum_{k : c_k \geq t} X_k = \frac{|\{1 \leq k \leq n \,:\, \tilde{\tau}_k \geq t\}|}{m(t)},$$

where the last equality follows because $\tilde{\tau}_k \geq t$ implies $c_k \geq t$.

The quality of this estimate is governed by the size of $m(t)$. This can be problematic when $m(t)$ is small, which happens, by definition, when a lot of the events are censored. A particularly unpleasant property of this estimator, that suggests that perhaps it is not the "best" estimator, is that it ignores all the observations whose censoring time precedes t. Intuitively, these observations still contain information about $S(t)$: For example, when for many events with $c_k < t$, $\tilde{\tau}_k < c_k$ also holds, we can infer that events often happen early, which implies that $\operatorname{Prob}(\tau \leq t)$ is large, which, through $S(t) = 1 - \operatorname{Prob}(\tau \leq t)$ means that $S(t)$ must be small. However, this information is ignored by this naive estimator. The question is then whether there exists an estimator that makes a better use of all the data. This is what the Kaplan–Meier estimator accomplishes. Note that the naive estimator cannot be improved when censoring does not take place; so whether an improvement is possible critically hinges upon whether censoring is in place.

The plug-in approach

By elementary calculations,

$$\begin{aligned} S(t) &= \operatorname{Prob}(\tau > t \mid \tau > t-1) \operatorname{Prob}(\tau > t-1) \\ &= (1 - \operatorname{Prob}(\tau \leq t \mid \tau > t-1)) \operatorname{Prob}(\tau > t-1) \\ &= (1 - \operatorname{Prob}(\tau = t \mid \tau \geq t)) \operatorname{Prob}(\tau > t-1) \\ &= q(t) S(t-1)\,, \end{aligned}$$

where the one but last equality used that τ is integer valued and for the last line we introduced

$$q(t) = 1 - \operatorname{Prob}(\tau = t \mid \tau \geq t).$$

By a recursive expansion of the equality $S(t) = q(t)S(t-1)$, we get

$$S(t) = q(t)q(t-1)\cdots q(0).$$

Note that here $q(0) = 1 - \operatorname{Prob}(\tau = 0 \mid \tau > -1) = 1 - \operatorname{Prob}(\tau = 0)$.

The Kaplan–Meier estimator can be seen as a "plug-in estimator" where each $q(s)$ is estimated based on the data and the estimator of $S(t)$ is obtained as a product of these estimates.

It remains to specify how $q(s) = 1 - \operatorname{Prob}(\tau = s \mid \tau \geq s)$ is to be estimated. By Proposition 1, for any $k \in [n]$ such that $c_k \geq s$, $\operatorname{Prob}(\tau = s) = \operatorname{Prob}(\tilde{\tau}_k = s)$ and $\operatorname{Prob}(\tau \geq s) = \operatorname{Prob}(\tilde{\tau}_k \geq s)$ both hold. Hence, for any $k \in [n]$ such that $c_k \geq s$,

$$\operatorname{Prob}(\tau = s | \tau \geq s) = \operatorname{Prob}(\tilde{\tau}_k = s) / \operatorname{Prob}(\tilde{\tau}_k \geq s).$$

By a similar reasoning that lead to the construction of the naive estimator above, we arrive at the estimator

$$\hat{q}(s) = 1 - \frac{|\{1 \leq k \leq n \,:\, c_k \geq s, \tilde{\tau}_k = s\}|}{|\{1 \leq k \leq n \,:\, c_k \geq s, \tilde{\tau}_k \geq s\}|} = 1 - \frac{|\{1 \leq k \leq n \,:\, \tilde{\tau}_k = s\}|}{|\{1 \leq k \leq n \,:\, \tilde{\tau}_k \geq s\}|}$$

(think of estimating the numerator and denominator separately in the definition of the "hazard rate" $\operatorname{Prob}(\tau = s | \tau \geq s)$). The Kaplan–Meier estimator is then given by

$$\hat{S}(t) = \prod_{s=0}^{t} \hat{q}(s).$$

The form of the estimator stated at the beginning of the article can be obtained by some further algebra. For this, write $\hat{q}(s) = 1 - d(s)/n(s)$ where, using the actuarial science terminology, $d(s) = |\{1 \leq k \leq n \,:\, \tilde{\tau}_k = s\}|$ is the number of known deaths at time s, while $n(s) = |\{1 \leq k \leq n \,:\, \tilde{\tau}_k \geq s\}|$ is the number of those persons who are alive at time $s - 1$.

Note that if $d(s) = 0$, $\hat{q}(s) = 1$. This implies that we can leave out from the product defining $\hat{S}(t)$ all those terms where $d(s) = 0$. Then, letting $0 \leq t_1 < t_2 < \cdots < t_m$ be the times s when $d(s) > 0$, $d_i = d(t_i)$ and $n_i = n(t_i)$, we arrive at the form of the Kaplan–Meier estimator given at the beginning of the article:

$$\hat{S}(t) = \prod_{i: t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

As opposed to the naive estimator, this estimator can be seen to use the available information more effectively: In the special case mentioned beforehand, when there are many early events recorded, the estimator will multiply many terms with a value below one and will thus take into account that the survival probability cannot be large.

Derivation as a maximum likelihood estimator

Kaplan–Meier estimator can be derived from maximum likelihood estimation of hazard function.[7] More specifically given d_i as the number of events and n_i the total individuals at risk at time t_i, discrete hazard rate h_i can be defined as the probability of an individual with an event at time t_i. Then survival rate can be defined as:

$$S(t)=\prod_{i:\ t_i\le t}(1-h_i)$$

and the likelihood function for the hazard function up to time t_i is:

$$\mathcal{L}(h_{j:j\le i}\mid d_{j:j\le i},n_{j:j\le i})=\prod_{j=1}^{i}h_j^{d_j}(1-h_j)^{n_j-d_j}$$

therefore the log likelihood will be:

$$\log(\mathcal{L})=\sum_{j=1}^{i}\left(d_j\log(h_j)+(n_j-d_j)\log(1-h_j)\right)$$

finding the maximum of log likelihood with respect to h_i yields:

$$\frac{\partial\log(\mathcal{L})}{\partial h_i}=\frac{d_i}{\widehat{h}_i}-\frac{n_i-d_i}{1-\widehat{h}_i}=0\Rightarrow\widehat{h}_i=\frac{d_i}{n_i}$$

where hat is used to denote maximum likelihood estimation. Given this result, we can write:

$$\widehat{S}(t)=\prod_{i:\ t_i\le t}\left(1-\widehat{h}_i\right)=\prod_{i:\ t_i\le t}\left(1-\frac{d_i}{n_i}\right)$$

Benefits and limitations

The Kaplan–Meier estimator is one of the most frequently used methods of survival analysis. The estimate may be useful to examine recovery rates, the probability of death, and the effectiveness of treatment. It is limited in its ability to estimate survival adjusted for covariates; parametric survival models and the Cox proportional hazards model may be useful to estimate covariate-adjusted survival.

Statistical considerations

The Kaplan–Meier estimator is a statistic, and several estimators are used to approximate its variance. One of the most common estimators is Greenwood's formula:[8]

$$\widehat{\operatorname{Var}}\left(\widehat{S}(t)\right)=\widehat{S}(t)^2\sum_{i:\ t_i\le t}\frac{d_i}{n_i(n_i-d_i)},$$

where d_i is the number of cases and n_i is the total number of observations, for $t_i<t$.

For a mathematical derivation of the equation above, click on "show" to reveal

Greenwood formula is derived[9] by noting that probability of getting d_i failures out of n_i cases follows a binomial distribution with failure probability h_i. As a result for maximum likelihood hazard rate $\widehat{h}_i=d_i/n_i$ we have $E\left(\widehat{h}_i\right)=h_i$ and $\operatorname{Var}\left(\widehat{h}_i\right)=h_i(1-h_i)/n_i$. To avoid dealing with multiplicative probabilities we compute variance of logarithm of $\widehat{S}(t)$ and will use the delta method to convert it back to the original variance:

$$\begin{aligned}\operatorname{Var}\left(\log\widehat{S}(t)\right) &\sim \frac{1}{\widehat{S}(t)^2}\operatorname{Var}\left(\widehat{S}(t)\right) \Rightarrow \\ \operatorname{Var}\left(\widehat{S}(t)\right) &\sim \widehat{S}(t)^2\operatorname{Var}\left(\log\widehat{S}(t)\right)\end{aligned}$$

using martingale central limit theorem, it can be shown that the variance of the sum in the following equation is equal to the sum of variances:[9]

$$\log\widehat{S}(t) = \sum_{i:\ t_i\le t}\log\left(1-\widehat{h}_i\right)$$

as a result we can write:

$$\begin{aligned}\operatorname{Var}(\widehat{S}(t)) &\sim \widehat{S}(t)^2\operatorname{Var}\left(\sum_{i:\ t_i\le t}\log\left(1-\widehat{h}_i\right)\right)\\ &\sim \widehat{S}(t)^2\sum_{i:\ t_i\le t}\operatorname{Var}\left(\log\left(1-\widehat{h}_i\right)\right)\end{aligned}$$

using the delta method once more:

$$\begin{aligned}\operatorname{Var}(\widehat{S}(t)) &\sim \widehat{S}(t)^2\sum_{i:\ t_i\le t}\left(\frac{\partial\log\left(1-\widehat{h}_i\right)}{\partial\widehat{h}_i}\right)^2\operatorname{Var}\left(\widehat{h}_i\right)\\ &= \widehat{S}(t)^2\sum_{i:\ t_i\le t}\left(\frac{1}{1-\widehat{h}_i}\right)^2\frac{\widehat{h}_i\left(1-\widehat{h}_i\right)}{n_i}\\ &= \widehat{S}(t)^2\sum_{i:\ t_i\le t}\frac{\widehat{h}_i}{n_i\left(1-\widehat{h}_i\right)}\\ &= \widehat{S}(t)^2\sum_{i:\ t_i\le t}\frac{d_i}{n_i(n_i-d_i)}\end{aligned}$$

as desired.

In some cases, one may wish to compare different Kaplan–Meier curves. This can be done by the log rank test, and the Cox proportional hazards test.

Other statistics that may be of use with this estimator are the Hall-Wellner band[10] and the equal-precision band.[11]

Software

- Mathematica: the built-in function `SurvivalModelFit` creates survival models.[12]
- SAS: The Kaplan–Meier estimator is implemented in the `proc lifetest` procedure.[13]
- R: the Kaplan–Meier estimator is available as part of the `survival` package.[14][15][16]
- Stata: the command `sts` returns the Kaplan–Meier estimator.[17][18]
- Python: the `lifelines` package includes the Kaplan–Meier estimator.[19]

- MATLAB: the `ecdf` function with the `'function','survivor'` arguments can calculate or plot the Kaplan–Meier estimator.[20]
- StatsDirect: The Kaplan–Meier estimator is implemented in the `Survival Analysis` menu.[21]

See also

- Frequency of exceedance
- Median lethal dose
- Nelson–Aalen estimator

References

1. Kaplan, E. L.; Meier, P. (1958). "Nonparametric estimation from incomplete observations". *J. Amer. Statist. Assoc.* **53** (282): 457–481. doi:10.2307/2281868 (https://doi.org/10.2307%2F2281868). JSTOR 2281868 (https://www.jstor.org/stable/2281868).
2. Kaplan, E.L. in a retrospective on the seminal paper in "This week's citation classic". *Current Contents* **24**, 14 (1983). Available from UPenn as PDF. (http://www.garfield.library.upenn.edu/classics1983/A1983QS51100001.pdf)
3. Meyer, Bruce D. (1990). "Unemployment Insurance and Unemployment Spells" (http://www.nber.org/papers/w2546.pdf) (PDF). *Econometrica*. **58** (4): 757–782. doi:10.2307/2938349 (https://doi.org/10.2307%2F2938349). JSTOR 2938349 (https://www.jstor.org/stable/2938349).
4. "- Google Scholar" (https://scholar.google.com/scholar?cites=14181649205747775124&as_sdt=5,28&sciodt=0,28&hl=en). *scholar.google.com*. Retrieved 2017-03-04.
5. "Paul Meier, 1924–2011" (http://articles.chicagotribune.com/2011-08-18/news/ct-met-meier-obit-20110818_1_clinical-trials-research-experimental-treatment). *Chicago Tribune*. August 18, 2011.
6. Rich JT, Neely JG, Paniello RC, Voelker CC, Nussenbaum B, Wang EW (2010). "A practical guide to understanding Kaplan–Meier curves" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959). *Otolaryngol Head Neck Surg*. **143** (3): 331–6. doi:10.1016/j.otohns.2010.05.007 (https://doi.org/10.1016%2Fj.otohns.2010.05.007). PMC 3932959 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959). PMID 20723767 (https://www.ncbi.nlm.nih.gov/pubmed/20723767).
7. (PDF) https://web.stanford.edu/~lutian/coursepdf/STAT331unit3.pdf (https://web.stanford.edu/~lutian/coursepdf/STAT331unit3.pdf). Missing or empty `|title=` (help)
8. Greenwood, M. (1926). "The natural duration of cancer". *Reports on Public Health and Medical Subjects*. London: Her Majesty's Stationery Office. **33**: 1–26.
9. (PDF) https://www.math.wustl.edu/%7Esawyer/handouts/greenwood.pdf (https://www.math.wustl.edu/%7Esawyer/handouts/greenwood.pdf). Missing or empty `|title=` (help)
10. Hall WJ and Wellner JA (1980) Confidence bands for a survival curve for censored data. Biometrika 69
11. Nair VN (1984) Confidence bands for survival functions with censored data: A comparative study. Technometrics 26: 265–275
12. "Survival Analysis – Mathematica SurvivalModelFit" (http://reference.wolfram.com/language/ref/SurvivalModelFit.html). *wolfram.com*. Retrieved 2017-08-14.
13. The LIFETEST Procedure (https://support.sas.com/documentation/cdl/en/statug/68162/HTML/default/viewer.htm#statug_lifetest_overview.htm)
14. "survival: Survival Analysis" (https://cran.r-project.org/web/packages/survival/index.html). *R Project*. April 2019.
15. Willekens, Frans (2014). "The *Survival* Package" (https://books.google.com/books?id=Cd2CBAAAQBAJ&pg=PA135). *Multistate Analysis of Life Histories with R*. Springer. pp. 135–153. doi:10.1007/978-3-319-08383-4_6 (https://doi.org/10.1007%2F978-3-319-08383-4_6). ISBN 978-3-319-08383-4.

16. Chen, Ding-Geng; Peace, Karl E. (2014). *Clinical Trial Data Analysis Using R* (https://books.google.com/books?id=fGnRBQAAQBAJ&pg=PA99). CRC Press. pp. 99–108. ISBN 9781439840214.
17. "sts — Generate, graph, list, and test the survivor and cumulative hazard functions" (https://www.stata.com/manuals15/ststs.pdf) (PDF). *Stata Manual*.
18. Cleves, Mario (2008). *An Introduction to Survival Analysis Using Stata* (https://books.google.com/books?id=xttbn0a-QR8C&pg=PA93) (Second ed.). College Station: Stata Press. pp. 93–107. ISBN 978-1-59718-041-2.
19. lifelines docs (https://lifelines.readthedocs.io/en/latest/)
20. "Empirical cumulative distribution function – MATLAB ecdf" (http://mathworks.com/help/stats/ecdf.html). *mathworks.com*. Retrieved 2016-06-16.
21. [1] (https://www.statsdirect.co.uk/help/Default.htm#survival_analysis/kaplan_meier.htm)

Further reading

- Aalen, Odd; Borgan, Ornulf; Gjessing, Hakon (2008). *Survival and Event History Analysis: A Process Point of View*. Springer. pp. 90–104. ISBN 978-0-387-68560-1.
- Greene, William H. (2012). "Nonparametric and Semiparametric Approaches" (https://books.google.com/books?id=-WFPYgEACAAJ&pg=PA909). *Econometric Analysis* (Seventh ed.). Prentice-Hall. pp. 909–912. ISBN 978-0-273-75356-8.
- Jones, Andrew M.; Rice, Nigel; D'Uva, Teresa Bago; Balia, Silvia (2013). "Duration Data" (https://books.google.com/books?id=7tdcCol9mNEC&pg=PA141). *Applied Health Economics*. London: Routledge. pp. 139–181. ISBN 978-0-415-67682-3.
- Singer, Judith B.; Willett, John B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (https://books.google.com/books?id=PpnA1M8VwR8C&pg=PA483). New York: Oxford University Press. pp. 483–487. ISBN 0-19-515296-4.

External links

- Dunn, Steve (2002). "Survival Curves: Accrual and The Kaplan–Meier Estimate" (http://www.cancerguide.org/scurve_km.html). *Cancer Guide*. Statistics.
- Staub, Linda; Gekenidis, Alexandros (Mar 7, 2011). "Kaplan–Meier Survival Curves and the Log-Rank Test" (http://stat.ethz.ch/education/semesters/ss2011/seminar/contents/handout_2.pdf) (PDF). *Survival Analysis* (http://stat.ethz.ch/education/semesters/ss2011/seminar/contents/presentation_2.pdf) (PDF). *Handout and presentation*. Seminar for Statistics (SfS). Eidgenössische Technische Hochschule Zürich (ETH) [Swiss Federal Institute of Technology Zurich].
- Three evolving Kaplan–Meier curves (https://www.youtube.com/watch?v=5C_zzD1pOAg) on YouTube

Retrieved from "https://en.wikipedia.org/w/index.php?title=Kaplan–Meier_estimator&oldid=922696746"

This page was last edited on 23 October 2019, at 18:43 (UTC).