Kaplan-Meier Estimator, Alternative Variance Formula and RMST based tests

Mai Zhou
University of Kentucky

RMST = Restricted Mean Survival Time

First, a quick review of the Kaplan-Meier estimator.

The (1958) JASA paper of Kaplan and Meier is THE MOST CITED statistics paper.

The second most cited paper is Cox (1972) (for Cox regression). The baseline survival function in a Cox regression model is also related to the Kaplan-Meier estimator.

(2005) (Journal of Applied Statistics) “The Most-Cited Statistical Papers” by T. P. Ryan & W. H. Woodall

Up to 2004, the Kaplan-Meier paper was cited 25,869 times.

 Paul Meier in 2006.

My small contribution to the 25,869 counts :-)

Zhou (1988) Two-sided bias bound of the Kaplan-Meier estimator. *Probability Theory and Related Fields*

Zhou (1991) Some Properties of the Kaplan-Meier Estimator for Independent Nonidentically Distributed Random Variables. *The Annals of Statistics*

Given n iid right censored survival data (T_i, δ_i), (like $3, 5^+, 7, \cdots$) the formula for the Kaplan-Meier estimator:

$$1 - \hat{F}_{km}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{R_i}\right) = \prod_{t_i \le t} \left(1 - \frac{\Delta N(t_i)}{R(t_i)}\right) \ .$$

It is also called the 'Product-Limit Estimator'.

The Kaplan-Meier estimator is the maximizer among all distributions (discrete or continuous) of the (nonparametric) likelihood function:

$$Lik(F) = \prod_i [\Delta F(t_i)]^{d_i} [1 - F(t_i)]^{R_i - d_i} = \prod_{i:\, \delta_i = 1} \Delta F(t_i) \prod_{i:\, \delta_i = 0} [1 - F(t_i)] \, .$$

Questions:

Expectation $\mathbf{E}\widehat{F}_{km}(t) = ?$ — We know the Kaplan-Meier is almost unbiased (bias is exponentially small).

Var $\widehat{F}_{km}(t) = ?$ — Greenwood formula for estimating (APPROXIMATE) variance of $\widehat{F}_{km}(t)$ at a fixed time t.

What is the (approximate) covariance between $\widehat{F}_{km}(t)$ and $\widehat{F}_{km}(s)$?

What is the variance of

$$\int_0^\infty g(t) d\widehat{F}_{km}(t)\,; \qquad \text{and/or}$$

$$\int_0^\tau [1 - \widehat{F}_{km}(t)] dt \text{ ?}$$

The reason for asking this question is the recent surge of interest in Restricted Mean Survival Time (RMST) as a parameter of interest, in place of hazard ratio.

Restricted Mean Survival Time

In two sample comparisons, the standard test with right censored survival data is the log-rank test. But it is optimal only when the "Proportional Hazards" assumption hold. Works OK for 'hazards not crossing each other' type situation. But can be very bad for 'Crossing Hazards' situation.

Example of CH: treatment delays the onset of symptoms; or delayed benefits (immuno-therapy)

What makes things worse is that it is not easy to tell them apart (from data): PH? CH?

Many recent references recommend using RMST test over log-rank test, when PH assumption is in doubt.

- Royston P, Parmar MKB. (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine*. 30:2409–2421.

- Zhao L, Tian L, Uno H, Solomon SD, Pfeffer MA, Schindler JS, Wei LJ. (2012). Utilizing the integrated difference of two survival functions to quantify the treatment contrast for designing, monitoring, and analyzing a comparative clinical study. *Clinical Trials*. 9:570–577.

- Uno, H., Claggett, B., Tian, L., Inoue, E., Gallo, P., Miyata, T., Schrag, D., Takeuchi, M., Uyama, Y., Zhao, L., Skali, H., Solomon, S., Jacobus, S., Hughes, M., Packer, M. and Wei, L.-J. (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*, 32, pp. 2380–2385.

- Kim DH, Uno H, Wei LJ. (2017). Restricted mean survival time as a measure to interpret clinical trial results. *JAMA Cardiol*; 2(11): 1179–1180. doi: 10.1001/jamacardio.2017.2922

- Abulizi, X; Ribaudo, H J.; Flandre, P. (2019). The Use of the Restricted Mean Survival Time as a Treatment Measure in HIV/AIDS Clinical Trial: Reanalysis of the ACTG A5257 Trial. *Journal of Acquired Immune Deficiency Syndromes* Vol. 81 Issue 1 pp. 44–51. doi: 10.1097/QAI.0000000000001978

• Royston, P., Parmar, M.K. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Med Res Methodol* 13, 152 (2013). https://doi.org/10.1186/1471-2288-13-152

• On the Restricted Mean Survival Time Curve in Survival Analysis Lihui Zhao, Brian Claggett, Lu Tian, Hajime Uno, Marc A. Pfeffer, Scott D. Solomon, Lorenzo Trippa, and L. J. Wei *Biometrics*. (2016) Mar; 72(1): 215–221. doi: 10.1111/biom.12384

• Analyzing Restricted Mean Survival Time Using SAS/STAT
SAS Paper 3013-2019

— and many more.

The computation of the test is available in the R package `survRM2` — (but I have some improvements), and

In SAS/STAT 15.1 or later, you can use the new RMST option in the LIFETEST procedure to estimate and compare the RMST, or use SAS Proc RMSTreg.

My (1986) dissertation include the Chapter 7: Difference of Means Test, and Chapter 8: Comparison of Tests.

Alternative Variance Formula

If $\widehat{F}_{km}(t)$ is the Kaplan-Meier estimator, based on n i.i.d. right censored data, what is the variance of

$$\text{Var} \int g(t) d\widehat{F}_{km}(t) \text{ ?}$$

Answer:

$$\text{AsyVar } \sqrt{n} \int g(t) d\widehat{F}_{km}(t) = \int [g(t) - \bar{g}(t)]^2 \frac{dF(t)}{1 - G(t-)}$$

where G is the CDF of the censoring variable.

This is proved by Akritas (2000) (Bernoulli). This is called the Asymptotic variance of the Kaplan-Meier mean or Kaplan-Meier integral.

For the RMST, (a specific $g(t) = \min(t, \tau)$), other people used a different formula:

`survRM2` package (By people from Harvard):

$$\widehat{\mathrm{Var}}_H = \sum_{t_i \le \tau} \left[\int_{t_i}^{\tau} 1 - \widehat{F}_{km}(s) ds \right]^2 \frac{d_i}{R_i(R_i - d_i)}$$

SAS: with an extra $\frac{m}{m-1}$ factor ... where m is the number of events before restriction time τ.

$$\widehat{\mathrm{Var}}_{sas} = \frac{m}{m-1} \sum_{t_i \le \tau} \left[\int_{t_i}^{\tau} 1 - \widehat{F}_{km}(s) ds \right]^2 \frac{d_i}{R_i(R_i - d_i)} .$$

Asymptotically, $n \cdot \mathrm{Var}_H$ has limit

$$\int_0^{\tau} \frac{\left[\int_t^{\tau} 1 - F(s) ds \right]^2}{[1 - F(t-)]^2} \frac{dF(t)}{1 - G(t-)} .$$

Some Comments (against the SAS factor $m/(m-1)$):

The RMST is $\int \min(t, \tau) d\widehat{F}_{km}(t)$.

The Kaplan-Meier at τ is $\int I[t \leq \tau] d\widehat{F}_{km}(t) = \widehat{F}_{km}(\tau)$.

They both are of the type $\int g(t) d\widehat{F}_{km}(t)$. (for different choice of $g(t)$)

While it is (almost) universally agreed that the Greenwood formula is THE estimate of choice for the variance of $\widehat{F}_{km}(\tau)$.

Why we need to add a factor $m/(m-1)$ in estimate the variance when the $g(t)$ function is the RMST? (but not for the $\widehat{F}_{km}(\tau)$)?

Should we change the factor as the restriction time τ change? Censoring percentage change?

What about other $g(t)$'s?

Advanced time (no censoring in the next 3 pages)

Suppose $X \sim F$ is a random variable. Assume the CDF F is continuous. Let $g(\cdot)$ be a function.

How do we compute the variance of $g(X)$? mean of $g(X)$?

Mean:

$$\mu = \mathbf{E}g(X) = \int g(s) dF(s) \, .$$

Variance: Usually

$$\text{Var}[g(X)] = \int [g(x) - \mu]^2 dF(x) \, .$$

Alternatively (less popular)

$$\text{Var}[g(X)] = \int [g(x) - \bar{g}(x)]^2 dF(x)$$

where

$$\bar{g}(s) = \frac{\int_{(s,\infty)} g(x) dF(x)}{1 - F(s)} = E_F[g(X) | X > s] \ .$$

This $\bar{g}(s)$ is called 'advanced-time transformation' by Efron & Johnstone (1990). Their paper also contains a proof of this formula.

Efron, B. & Johnstone, I. (1990). Fisher's information in terms of the hazard rate. Annals of Statistics

For covariance, between $g(X)$ and $h(X)$, there is a similar formula:

$$\text{cov}(g(X), h(X)) = \int [g(x) - \bar{g}(x)][h(x) - \bar{h}(x)]dF(x).$$

This can be proved similarly to Efron & Johnstone (1990).

If $X_1, \cdots, X_n$ i.i.d. $\sim F$. The variance of

$$\text{Var}\left\{\frac{1}{n}\sum_{i=1}^{n} g(X_i)\right\} = \frac{1}{n}\text{Var}[g(X_1)].$$

Define the empirical distribution function based on $X_1, \cdots, X_n$,

$$\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \leq t]$$

then

$$\int g(t) d\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

Thus

$$\text{Var}\left\{\int g(t) d\widehat{F}(t)\right\} = \frac{1}{n}\int [g(x) - \bar{g}(x)]^2 dF(x).$$

It turns out, that for right censored data, with the Kaplan-Meier estimator taking the place of the empirical distribution,

we have

$$\text{Var}\left\{\int g(t)d\widehat{F}_{km}(t)\right\} \approx \frac{1}{n}\int [g(x) - \bar{g}(x)]^2 \frac{dF(x)}{1 - G(x-)}.$$

The exact equality holds in the limit as $n \to \infty$.

A similar covariance formula holds. (can be used to figure out the cov between $\widehat{F}_{km}(t)$) and $\widehat{F}_{km}(s)$.)

Remark The censored and un-censored version of variance are SO SIMILAR.

This simpler variance formula allows us to use the Cauchy-Schwarz inequality.
(and show the optimality of the Kaplan-Meier mean, etc.)

Variance of an MLE can be approximated by the inverse of the Fisher information number.

Recall, the Kaplan-Meier estimator is an MLE and the likelihood function is available. (Even though it is a Non-parametric MLE)

When there is a likelihood function and MLE, it is natural to try some known methods based on likelihood functions. i.e. Fisher Information and likelihood ratio test.

• Find the Fisher Information? (Not easy, since the derivative is not easily defined. Parameter is $F(\cdot)$ itself — infinite dimensional.)

• Try likelihood ratio test? YES. Empirical Likelihood Ratio test (Owen 1988, 2001), Zhou (2016).

Great features of likelihood ratio tests/confidence interval:

1. No need to estimate the variance.

2. No need to apply a transformation to the parameter.

Transformation:

To compute a confidence interval for the Kaplan-Meier estimator at t_0, R package `survival` recommend/default to a log transformation. (with several others available).

SAS recommend/default to a log-log transformation. (with even more different transformations available).

For computing the confidence interval for RMST1 − RMST2, people from Harvard commented in a paper that using some transformation may be a better approach but not sure what is the appropriate transformation.

For the ratio RMST1/RMST2 they used a log transform.

Likelihood ratio confidence intervals automatically apply the best transformation without you worrying about it.

Now let us focus on the **information number**, since this is also related to the variance of the MLE.

We shall not compute the information for the parameter $F(\cdot)$ but for $\mu = \int g(t)dF(t)$.

The latter is 1-dim, and should be easier to deal with.

Inference for a finite dim parameter inside an infinite dim model is usually called semiparametric.

Reference Books:

(1993, 1998) Efficient and Adaptive Estimation for Semiparametric Models, by BKRW

(2006) Semiparametric Theory and Missing Data, by Tsiatis

(2008) Introduction to Empirical Processes and Semiparametric Inference, by Kosorok

Yet the semiparametric information is still a difficult topic, involving Hilbert spaces, Hadamard derivatives, etc.

We, instead, shall compute the OBSERVED semiparametric Fisher information for μ in the likelihood function Lik given by Kaplan-Meier (1958). Much Easier. (expected info needs to compute expectation...for Kaplan-Meier this is difficult)

Expected Fisher Information $$I_e(\theta_0) = \mathbf{E}\left[\frac{\partial \log L(\theta)}{\partial \theta}\right]^2_{\theta=\theta_0}$$

Observed Fisher Information $$I(\widehat{\theta}_{mle}) = -\frac{\partial^2 \log L(\theta)}{(\partial \theta)^2}\Big|_{\theta=\widehat{\theta}_{mle}}$$

In above 3 books, they all seem to work exclusively with expected Fisher information.

Charles Stein (1956) told us how to proceed for semiparametric information:

• Obviously, the information in the nonparametric model is less than those of ANY parametric model contained in the nonparametric model. (called a sub-model)

• We compute the information for μ in a parametric sub-model, and minimize the information over ALL possible such parametric sub-models.

The minimum is then the μ-information of the nonparametric model.

We compute the observed Fisher information for estimating μ in $Lik(F)$:

$$- \frac{\partial^2 \log Lik(F)}{(\partial \mu)^2} |_{F=\hat{F}_{km}} \tag{1}$$

We use one-parameter sub-models: (λ is the parameter, $\lambda \in (-\epsilon, \epsilon)$)

$$\Delta F_\lambda(t_i) = \Delta \hat{F}_{km}(t_i)[1 - \lambda f(t_i)] \ . \tag{2}$$

By chain rule for differentiate compound functions, we compute the derivative of $\log Lik$ with respect to λ and then derivative of λ with respect to μ.

After long and tedious calculation/simplifications, using 'self-consistency identity', recursive formula for Kaplan-Meier jumps, and alternative variance/covariance formula, etc. the negative second derivative is

$$\frac{\sum [f(t_i) - \bar{f}(t_i)]^2 [1 - \hat{G}_{km}(t_i-)] \Delta \hat{F}_{km}(t_i)}{[\sum g(t_i) f(t_i) \Delta \hat{F}_{km}(t_i)]^2}$$

Minimize the above or (1) over all possible 1-parameter sub-models (here minimizing over all possible f using Cauchy-Schwarz inequality), we get the semiparametric observed information

$$I(\mu, \hat{F}_{km}) = \left\{ \frac{1}{n} \int [g(t) - \bar{g}(t)]^2 \frac{d\hat{F}_{km}(t)}{1 - \hat{G}_{km}(t-)} \right\}^{-1}$$

This is exact, no approximation, not limit. For a given sample size n.

No fancy derivatives.

We see that the Kaplan-Meier mean/integral (asymptotically) reach the information lower bound.

The ‘least favorable’ parametric sub-model for estimating μ is given by (2) with:

$$f(t) - \bar{f}(t) \propto \frac{g(t) - \bar{g}(t)}{1 - \widehat{G}_{km}(t-)}$$

The expected information seems hard, since we cannot even compute the exact expectation for Kaplan-Meier, worse for variance. Need to resort to Big O, Small o, Weak convergence, sup-tea, exchange of expectation and integration etc.

With censoring, I can compute both the expected and observed semiparametric information for estimating $\theta = \int g(t) d\Lambda(t)$ — 'mean hazard', or Nelson-Aalen integrals.

Without censoring, I can compute the expected semiparametric information for estimating μ.

As for the likelihood ratio test:

We already have the Lik given by Kaplan-Meier (1958) above and know it is maximized by the Kaplan-Meier estimator.

To find the likelihood ratio, what is missing is the same Lik but maximized among all CDFs that satisfy a null hypothesis H_0. (for example, true RMST = 50)

We shall leave out the details. Or see my paper 'Empirical likelihood ratio tests for RMSTs'. (with some example R-codes)

Two improvements over existing R/SAS computation for two-sample RMST difference tests:

1. The (empirical) likelihood ratio based RMST-test is better.

2. If you insist on using the Wald test to compare two samples, i.e. a confidence interval given by

$$est \pm Z_{\alpha/2} \cdot SE$$

then I suggest to use a Welch-Satterthwaite calibration.

The LR tests can also have some calibration ...... but at least it got the shape of confidence region right.

A typical situation:

Wald confidence interval:

Wilks confidence interval: Same $\widehat{\text{Est}}$

Some Simulations

First one is from my 1986 dissertation. n1=n2=50.

It shows that mean based tests for right censored data is very competitive to log-rank test, even in Proportional Hazards case. (as long as the true DFs have light tail).

Here the null distribution is Weibull $1 - F(t) = 1 - \exp(-t)^2$.

POWER SIMULATIONS FOR PROPORTIONAL HAZARD ALTERNATIVES				
	logrank	RANK	mean	censoring%
H_0: λ=1.0	5.043% (5.058%)	5.041% (5.041%)	5.039%	20.0%
H_A: λ=1.1	13.03% (13.05%)	13.39% (13.65%)	12.98%	23.2%
H_A: λ=1.2	34.28% (34.31%)	34.78% (35.02%)	34.42%	26.4%
H_A: λ=1.3	59.88% (59.98%)	60.04% (60.21%)	60.28%	29.7%
H_A: λ=1.4	80.27% (80.31%)	79.88% (80.01%)	80.74%	32.9%
H_A: λ=1.5	91.68% (91.76%)	91.20% (91.39%)	92.10%	36.0%

Recent Simulations

For both samples, the lifetimes X_i are generated from a Weibull distribution with shape parameter 0.9 and scale parameter 12.7. The censoring times C_i are generated from a uniform (0, 12.5) distribution.

Simulated type I errors are based on 5000 runs. We have used the t-distribution calibration for empirical likelihood as suggested by Owen (2001) equation (3.3): i.e. the $-2\log LR$ is deemed 5% significant if it is larger than $[qt(0.975, df = n1 + n2 - 2)]^2$, etc. where $qt(\,)$ denotes the quantile function of a student t-distribution.

sample sizes	restriction time	nominal 5%	nominal 10%	Wald test (5%)
25, 25	$\tau = 8$	5.74%	11.08%	7.1%
25, 25	$\tau = 10$	6.12%	11.52%	7.02%
30, 30	$\tau = 8$	5.30%	10.44%	6.52%
30, 30	$\tau = 10$	5.73%	11.34%	6.44%
40, 40	$\tau = 8$	5.30%	10.06%	6.02%
40, 40	$\tau = 10$	5.36%	10.72%	6.08%

Table 1. Nominal versus actual type I error.

Wald test results obtained by using `survRM2` package.

The type I errors for empirical likelihood tests are slightly inflated compare to the nominal, but they are less profound than the results of asymptotic Wald test.

Even without the t-distribution calibration, empirical likelihood tests type I error are more accurate: in sample sizes 30+30, $\tau = 8$ case is 5.92% vs. Wald test 6.52%, and for 30+30, $\tau = 10$, 5.73% vs. Wald test 6.44%, etc.

For the same sample size, we see that when the restriction time are higher ($\tau = 10$), the type I errors of EL are less accurate than a lower restriction time ($\tau = 8$).

There are some recent researches on the choice of the restriction time τ.

Next, we look at the Welch–Satterthwaite calibration (from two sample t-test) and try the alternative hypothesis (when two true RMSTs are not equal), and un-equal sample sizes.

Sample 1: Weibull shape=1.9, scale =12.7

Sample 2: Weibull shape =0.9, scale =12.7

Percentage of Simulated confidence intervals NOT covering the true RMST difference: nominal 95% confidence interval and 99% intervals.

sample size	restriction	No calib.	W-S calib.	No calib.	W-S calib.
25, 20	$\tau = 10$	7.47%	6.56%	2.19%	1.58%
30, 25	$\tau = 10$	6.86%	6.23%	1.71%	1.36%
30, 20	$\tau = 10$	7.29%	6.5%	2.31%	1.76%
35, 25	$\tau = 10$	6.83%	6.19%	1.95%	1.56%

Table 2. Welch–Satterthwaite correction for Wald test from `survRM2`.

Under null hypothesis, type I error. Different sample sizes.

Nominal 5% and 1% error.

sample size	restriction	No calib.	W-S calib.	No calib.	W-S calib.
30, 20	$\tau = 10$	6.87%	6.13%	2.06%	1.62%
35, 25	$\tau = 10$	6.51%	5.96%	1.77%	1.47%

Table 3. Welch–Satterthwaite calibration for Wald test from `survRM2`.

Inflation of error is worse under alternative hypothesis (compare to null) ...... where skewness is worse.

In general, the Welch–Satterthwaite calibration is more visible for 1. smaller and unequal sample sizes; 2. unequal and large SE in two samples.

But Welch-Satterthwaite still do not solve the skewness problem for Wald test.

Three Levels of P-value Accuracy (null distribution)

1. Tests using one null distribution (Normal(0, 1) or Chi-square df=1).

(One-size-fit-all)

2. Tests using null distributions depend on the sample sizes n_1 and n_2 (t-distributions, square of t-distributions, F-distributions)

(Different sizes of shoes)

3. Tests using null distributions depend on the sample sizes AND some sample estimates (Welch-Satterthwaite, Bartlett, Bootstrap etc.)

(Within the same size shoe, you still have Wide, Narrow, custom fit etc. to capture specific shape)

Empirical likelihood ratio RMSTs tests, under H_A. (based on 5000 runs)

Nominal	sample size	restriction	Chi-square	$t^2(n1+n2-2)$	W-S calib.
5%	35, 25	$\tau = 10$	6.7%	6.12%	5.9%
1%	35, 25	$\tau = 10$	1.76%	1.4%	1.32%

Table 4. Simulated type II error of EL test using Chi-square, $t^2(n1+n2-2)$ and Welch–Satterthwaite

Under H_0. Different sample sizes.

Nominal	sample size	restriction	Chi-square	$t^2(n1+n2-2)$	W-S calib.
5%	35, 25	$\tau = 10$	6.54%	6.04%	5.94%
1%	35, 25	$\tau = 10$	1.64%	1.34%	1.28%

Table 5. Simulated type I error of EL test using Chi-square, $t^2(n1+n2-2)$ and Welch–Satterthwaite

Compute Empirical Likelihood ratio Tests

1. How do one compute such a Lik? (EM algorithm or recursive + Newton-Raphson)

The recent update of the (much faster, recursive based) package `kmc` version 0.2-4 is broken. :-(

For now, have to use `emplik` package (EM-algorithm based), or use `kmc` version 0.2-3 on older R.

2. Does the Wilks Theorem still hold? ($-2 \log LR \to \chi^2$) .

Computational time for EL test (computing p-values) for RMST1 − RMST2:

On a Dell OptiPlex 9020 desktop computer with i7-4790 and R version 4.0.2

Sample size 200/200; about 1.1 second per p-value.

Sample size 50/50; about 0.9 second per p-value.

The Wald test is almost instant.

Wilks confidence intervals have many advantages over the Wald type confidence intervals. See for example

> Meeker & Escobar, (1995). Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation *The American Statistician*, Vol. 49, No. 1, pp. 48–53.

Empirical likelihood-based confidence intervals share the same nice properties with those based on parametric likelihoods.

The disadvantage of the Wilks type confidence intervals often cited is the difficulty of computation.

Faster computers, innovative algorithms (EM-algorithm, recursive formula, Newton-Raphson iteration) certainly made this less painful.

Additional Recent References:

Zhou, M. (2020). Restricted Mean Survival Time and Confidence Intervals by Empirical Likelihood Ratio. Forthcoming.

Zhou, M. (2020). Semiparametric Observed Information Bound for Kaplan-Meier Integrals and Nelson-Aalen Integrals. Submitted

Horiguchi, M. & Uno, H. (2020). On permutation tests for comparing restricted mean survival time with small sample from randomized trials. *Statistics in Medicine*

Nemes, S., Bülow, E. and Gustavsson, A. (2020). A Brief Overview of Restricted Mean Survival Time Estimators and Associated Variances. *MDPI Stats* 3(2), 107-119; https://doi.org/10.3390/stats3020010