# The format of Sp(t)

*2019-06-13*

I think the format of $S_p(t)$ should be:

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho(X_i)}{n_i})\}$$

The process is showing below.

## Deriving $\hat{S}_p(t)$

The old equation in Slud paper

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_i - 1}\} \text{ equation (0)}$$

To solve it, let's begin with:

$$\begin{aligned}
&P(T > X_{(j)}, C < X_{(j-1)}) + P(T > X_{(j)}, C > X_{(j-1)}) \\
&= P(T > X_{(j)}) \\
&= P(T > X_{(j)}, C > X_{(j)}) + P(T > X_{(j)}, C < X_{(j)})
\end{aligned}$$

And

1. $$\begin{aligned}
P(T > X_{(j)}, C < X_{(j-1)}) =& P(T > X_{(j)}, T > X_{(j-1)}, C < X_{(j-1)}) \\
=& P(T > X_{(j)} | T > X_{(j-1)}, C < X_{(j-1)}) \times P(T > X_{(j-1)}, C < X_{(j-1)}) \\
=& (1 - P(T < X_{(j)} | T > X_{(j-1)}, C < X_{(j-1)})) \times P(T > X_{(j-1)}, C < X_{(j-1)})
\end{aligned}$$

2. If $X_{(j)} - X_{(j-1)} \to 0$

$$\begin{aligned}
P(T > X_{(j)}, C > X_{(j-1)}) =& P(T > X_{(j)}, T > X_{(j-1)}, C > X_{(j-1)}) \\
=& P(T > X_{(j)} | T > X_{(j-1)} C > X_{(j-1)}) P(T > X_{(j-1)} C > X_{(j-1)}) \\
=& P(T > X_{(j)} | T > X_{(j-1)}, C > X_{(j-1)}) \times P(W > X_{(j-1)}) \\
\approx& \frac{n_{j-1} - 1}{n_{j-1}} \times \frac{n_{j-1}}{N} = \frac{n_{j-1} - 1}{N}
\end{aligned}$$

Therefore,

$$(1 - P(T < X_{(j)} | T > X_{(j-1)}, C < X_{(j-1)})) \times P(T > X_{(j-1)}, C < X_{(j-1)}) + \frac{n_{j-1} - 1}{N}$$

$$= P(T > X_{(j)}, C > X_{(j)}) + P(T > X_{(j)}, C < X_{(j)})$$

$$\begin{aligned}
P(T > X_{(j)}, C < X_{(j)}) =& (1 - \frac{\rho(X_{j-1})}{n_{j-1}}) \times P(T > X_{(j-1)}, C < X_{(j-1)}) + \frac{n_{j-1} - 1}{N} - \frac{n_j}{N} \\
=& (1 - \frac{\rho(X_{j-1})}{n_{j-1}}) \times P(T > X_{(j-1)}, C < X_{(j-1)}) + \frac{c_{j-1}}{N}
\end{aligned}$$

Let $Y_j = P(T > X_{(j)}, C < X_{(j)})$, $A_j = 1 - \frac{\rho(X_j)}{n_j}$, $B_j = \frac{c_j}{N}$ to make it is easier to see.

Since $Y_0 = P(T > X_{(0)}, C < X_{(0)})$, we can treat it as something that will never happen and probability = 0. Therefore,

- $Y_1 = A_0 Y_0 + B_0 = B_0$, since $Y_0 = 0$. Begin with 0 since $k = 0$ in the equation 0

- $Y_2 = A_1Y_1 + B_1 = A_1B_0 + B_1$
- $Y_3 = A_2Y_2 + B_2 = A_2A_1B_0 + A_2B_1 + B_2$
- ...
- Therefore the equation is:

$$\begin{aligned} Y_n =& B_0 \prod_{i=1}^{n-1} A_i + B_1 \prod_{i=2}^{n-1} A_i + B_2 \prod_{i=3}^{n-1} A_i + ...B_{n-2} \prod_{i=n-1}^{n-1} A_i + B_{n-1} \\ =& [\sum_{k=0}^{n-2} B_k \prod_{i=k+1}^{n-1} A_i] + B_{n-1} \qquad \text{(equation (1))} \\ =& [\sum_{k=0}^{n-2} \frac{c_k}{N} \prod_{i=k+1}^{n-1} (1 - \frac{\rho(X_i)}{n_i})] + \frac{c_{n-1}}{N} \end{aligned}$$

However, it does not equal to: $\sum_{k=0}^{d(t)-1} \frac{c_k}{N} \prod_{i=k+1}^{d(t)} \left(1 - \frac{\rho(X_i)}{n_i}\right)$ (equation (2))

But equation (2) is $(1 - \frac{\rho(X_{d(t)})}{n_{d(t)}})$ times equation (1)

We can show from the simulated examples that the newly derived formula has the best performance. Therefore, I think the equation should be

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho(X_i)}{n_i})\}$$

How does it work?

## Simulation

### Example 1

Suppose we simulate the data from the following joint distribution:

$$f(s,t) = \frac{1}{1000}(s+t)$$

where $s \in (0, 10)$ and $t \in (0, 10)$.

1. The mean difference between KM estimater and the true $S(t)$

```
fit1 = survfit(Surv(time, status) ~ 1, data=data)
fit2 = km.ci(fit1)
mean(abs(S(fit1$time) - fit1$surv))
```

```
## [1] 0.03647195
```

2. The mean difference between Slud equation and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_i - 1}\}$$

```
mean(abs(s0 - s1))
```

```
## [1] 0.03464802
```

3. The mean difference between Slud equation (after correction of the $\rho(t)$) and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} (1 - \frac{\rho_i}{n_i})\}$$

```r
mean(abs(s0 - s2))
```

```
## [1] 0.03445972
```

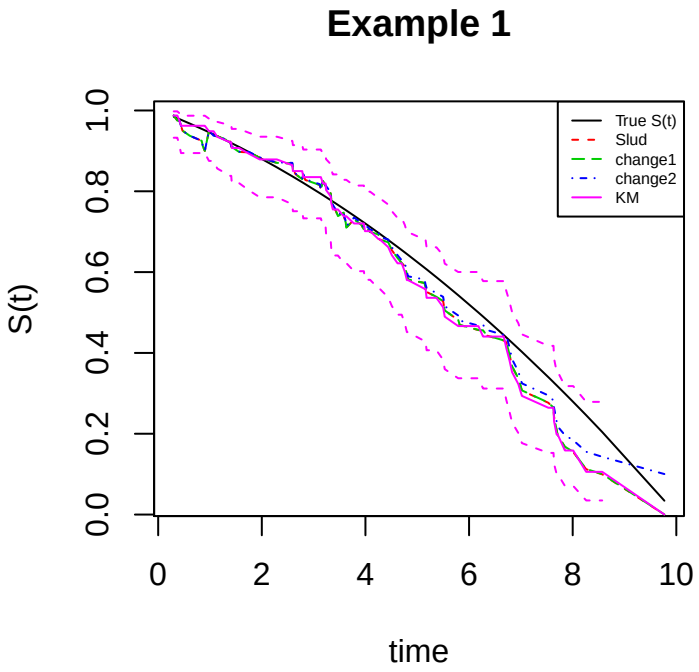
4. The mean difference between new equation and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1}(1 - \frac{\rho(X_i)}{n_i})\}$$

```r
mean(abs(s0 - s3))
```

```
## [1] 0.02828857
```

The plots:




In the plot,

- The purple lines are the KM estimates and their confidence intervals.
- The three estimates of true $S(t)$ look similar, but the new formula has the smallest difference with true $S(t)$
- The true $S(t)$ as well as the three estiamtes are within the confidence intervals of KM. Therefore, we may say that in this scenario, the KM can work well, even the independent assumption is not satisfied.

**Example 2**

$$f(t,s) = \begin{cases} exp(-t-s) & (t \le s) \\ 10exp(8s-10t) & (t > s) \end{cases}$$

And $\rho(t) = 10$

1. The mean difference between KM estimater and the true $S(t)$

```r
fit1 = survfit(Surv(time, status) ~ 1, data=data)
fit2 = km.ci(fit1)
mean(abs(S(fit1$time) - fit1$surv))
```

```
## [1] 0.3172037
```

2. The mean difference between Slud equation and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} \frac{n_i - 1}{n_i + \rho_i - 1}\}$$

```
mean(abs(s0 - s1))
```

```
## [1] 0.0520484
```

3. The mean difference between Slud equation (after correction of the $\rho(t)$) and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + \sum_{k=0}^{d(t)-1} c_k \prod_{i=k+1}^{d(t)} (1 - \frac{\rho_i}{n_i})\}$$

```
mean(abs(s0 - s2))
```

```
## [1] 0.06191508
```

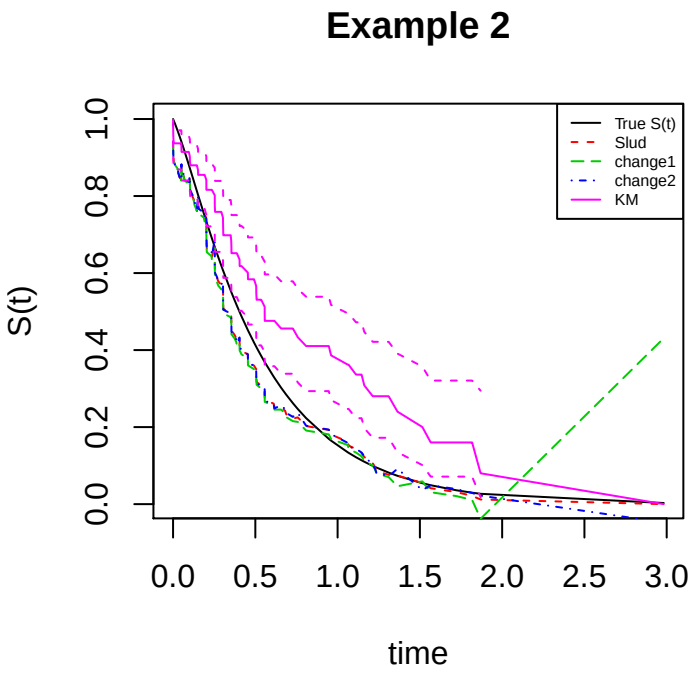
4. The mean difference between new equation and the true $S(t)$

$$\hat{S}_p(t) = \frac{1}{N}\{n(t) + c_{d(t)-1} + \sum_{k=0}^{d(t)-2} c_k \prod_{i=k+1}^{d(t)-1} (1 - \frac{\rho(X_i)}{n_i})\}$$

```
mean(abs(s0 - s3))
```

```
## [1] 0.05125131
```

The plots:




In this plot:

- The difference becomes larger for the scenario 3, which we correct $\frac{n_i-1}{n_i+\rho_i-1}$ into $1 - \rho_i/n_i$.
- It is true that $\frac{n_i-1}{n_i+\rho_i-1}$ can control the value's range. When $\rho$ is large, $1 - \rho_i/n_i$ can smaller than 0 and make the values to be negative. That is why the green line, which is the the method that just change $\frac{n_i-1}{n_i+\rho_i-1}$ to $\frac{n_i-\rho_i}{n_i}$, increased at the end.

```r
tail(s2)
```

```
## [1]  0.05908528  0.03091472  0.02091472  0.01091472 -0.03786940  0.43082457
```

- The true $S(t)$ is out of the confidence intervals of KM. In this scenario, the KM doesn't work well.
- The three estimaters also do not have big differences here. And the new one seems to have the best result.