Lecture Three: Standard Error of KM Estimate and Estimating Hazard Function

1. Standard Error and Confidence interval for $\hat{S}(t)$

We also need to know about how good it's the (KM) estimate. A common way is to estimate the sample variation or standard error of the estimate $\hat{S}(t)$.

Use the derivation at page 26-27: Steps:

- Take log transformation of KM estimate
- # of survivals, $n_j - d_j$, through the interval beginning at $t_{(j)}$ has Binomial(n_j, $\hat{p}_j$), where $\hat{p}_j = 1 - d_j/n_j$
- Obtain the variance of log $\hat{p}_j$ by the delta-method:

$$\text{var}\{g(X)\} \approx \{\frac{dg(X)}{dX}\}^2 \text{ var}(X),$$

which is known as the *Taylor series approximation* to the variance of a function of a random variable.

- Standard error (S.E.): square-root of variance estimate.

With the estimated standard error, a **(1 - α)100% confidence interval for** $\hat{S}(t)$ **at each time point t** can be easily constructed, based on a typical normal approximation (meaning?). When we link the upper and lower confidence limits together along the time axis, we form a so-called **confidence band**. This can be done on different scales as implemented in Splus and SAS (PROC LIFETEST: conftype, confband options in SURVIVAL statement).

- **Original scale: S (t).**
  - Confidence interval for $\hat{S}(t_j)$ at t_j

    CI = $\hat{S}(t_j) \pm z_{\alpha/2}$*S.E. ($\hat{S}(t_j)$)
  - Although S (t) should be in [0, 1], the lower and upper limit can be out of the range. A practical solution to this problem is to replace any limit that greater than 1 by 1, and any limit that is less than zero by 0.0.

- **Log-scale: log S (t).**
  - Confidence interval for $\log \hat{S}(t_j)$

    $CI_{log} = \log \hat{S}(t_j) \pm z_{\alpha/2}$*S.E. ($\log \hat{S}(t_j)$)
  - Converting CI_{log} back to the original scale to have CI for $\hat{S}(t_j)$

    CI = exp (CI_{log}) =?

- Where the lower bound is always nonnegative, the upper bound may exceed 1

- **Log-log scale: log (-log $\hat{S}(t_j)$).**

  - Obtain the standard error for log(-log $\hat{S}(t_j)$) by the delta-method
  - Confidence interval $CI_{\text{log-log}}$ for log(-log $\hat{S}(t_j)$) by the normal approximation
  - Convert $CI_{\text{log-log}}$ to have CI for $\hat{S}(t_j)$

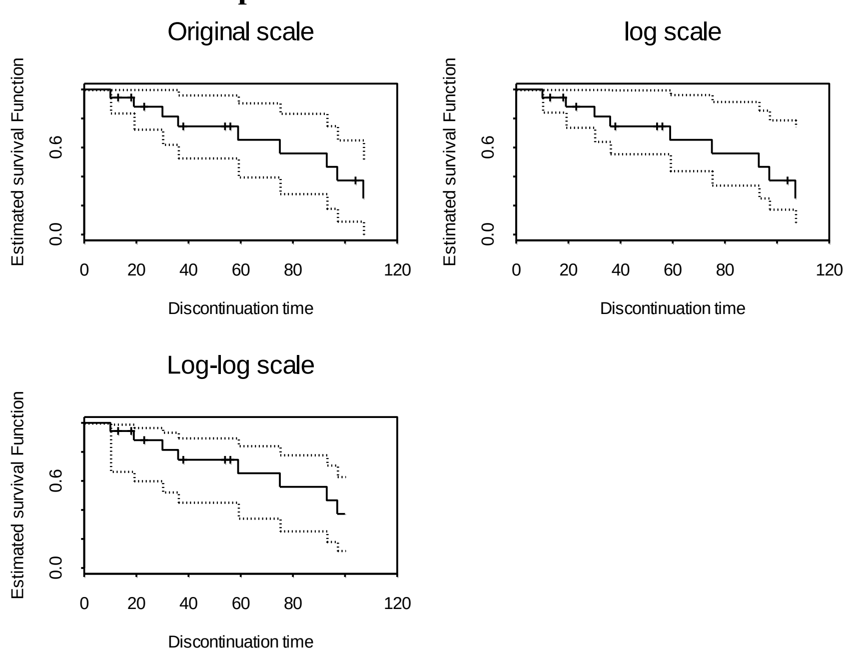
$$CI = \exp\left(-\exp\left(CI_{\text{log-log}}\right)\right)$$

  - Lower limit >= 0 and upper limit <= 1
  - Appropriate with moderate to large sample size because of repeated use of the delta-method.

The Greenwood variance estimate is appropriate only when the expected risk set size n_j is fairly large at each time point $t_{(j)}$ because the use of the delta-method requires large sample size. As n_j gets smaller with increasing time, the Greenwood estimate becomes unstable at the tail. (Cut the tail out requested by investigators, reasonable?)

- In Splus, use option "conf.type" in "survfit()" to choose different methods
- In SAS, use *conftype* option in the PROC LIFETEST statement.

- **Example: IUD**



Splus code:

```
 iud.s<-function (){
tmpdf <- importData("../sdata/iud.sas7bdat")
motif()
par(mfrow=c(2,2))
iud.km1 <- survfit(Surv(survt, censor), conf.type="plain",
            type="kaplan-meier", data=tmpdf)
plot(iud.km1,xlab="Discontinuation time",
          ylab="Estimated survival Function", xlim=c(0, 120),
               ylim=c(0,1),mark.time=T, conf.int=T,
                 main="Original scale")
iud.km2 <- survfit(Surv(survt, censor), conf.type="log",
                 type="kaplan-meier", data=tmpdf)
plot(iud.km2,xlab="Discontinuation time",
                ylab="Estimated survival Function", xlim=c(0, 120) ,
             ylim=c(0,1),mark.time=T, conf.int=T, main="log scale")
iud.km3 <- survfit(Surv(survt, censor), conf.type="log-log",
          type="kaplan-meier", data=tmpdf)
plot(iud.km3,xlab="Discontinuation time",
        ylab="Estimated survival Function", xlim=c(0, 120),
        ylim=c(0,1),mark.time=T, xmax= 100, conf.int=T,
        main="log-log scale")
}
```

2. Estimating the hazard function

- **Life-table estimate of the hazard function**

    - Dividing the period of observation into a series of time intervals: $t_j^{'}$ to $t_{j+1}^{'}$, j = 1, 2, …,m, with length τ_j
    - d_j deaths, c_j censored in $(t_j^{'},\ t_{j+1}^{'}]$ and n_j at risk at the start of the j'th interval
    - Assume censored times occur uniformly (i.e. U(0, c_j)) through the j'th interval, then average number of individual at risk is $n_j^{'} = n_j - c_j / 2$
    - Assuming the death rate is constant during the j'th interval
    - The average hazard of death per unit time can be estimated by

$$h^*(t) = \frac{d_j}{(n_j^{'} - d_j / 2)\tau_j},$$

for $t_j^{'} \le t < t_{j+1}^{'},\ j = 1,2,...,m$, where $(n_j^{'} - d_j / 2)\tau_j$ is the average time survived in $(t_j^{'},\ t_{j+1}^{'}]$.

- Kaplan-Meier Type Estimate

  Let the observed survival times: $t_1, t_2, \ldots, t_n$ and r ordered death times: $t_{(1)} < t_{(2)} < \ldots < t_{(r)}$; n_j at risk just before $t_{(j)}$, d_j deaths at the j'th death time
  - Assuming constant hazard between successive death times
  - The hazard can be estimated by

  $$\hat{h}(t) = \frac{d_j}{n_j \tau_j},$$

  for $t_{(j)} \le t < t_{(j+1)}$, where $\tau_j = t_{(j+1)} - t_{(j)}$

  - No estimate for $t > t_{(r)}$
  - Proof: The conditional death probability for $t_j \le T < t_{(j+1)}$ is $\hat{h}(t)\tau_j$, which is d_j/n_j
- **Kernel-smoothed estimate**
  - Above estimates are rather irregular
  - Using smoothing techniques (ref: Smoothing Methods in Statistics, 1996, Simonoff JS).
  - A weighted average of values of the estimated hazard $\hat{h}(t)$ at death times in the neighborhood of t.

- **Estimating the cumulative hazard function**
  - Use relation $H(t) = -\log S(t)$ and KM estimate of survivor function $\hat{S}(t) = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j}$, for
  - Use Taylor series expansion of log(1 – x), and ignore higher-order terms when x is small
  - $\hat{H}(t) \approx \sum_{j=1}^{k} \frac{d_j}{n_j}$, which is called Nelson-Aalen estimate.

3. **Estimating the median, mean and percentiles of survival times**

- **Median survival time**: defined as smallest observed survival time for which the value of the estimated survival function is less than 0.5
- In math term

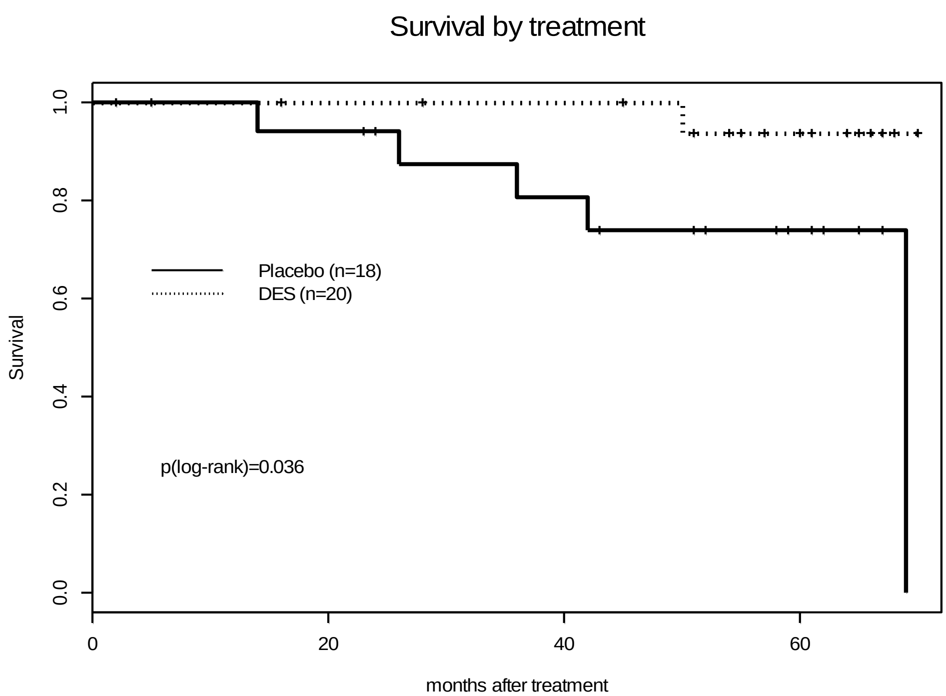
$$\hat{t}(50) = \min\{t_i \mid \hat{S}(t_i) \le 0.5\}$$

where t_i is the observed survival time for the *i*'th individual, I = 1, …, n

- What if $\hat{S}(t) > 0.5$ for any t > 0?
- Mean: $E(T) = \int_0^\infty tf(t)dt = -\int_0^\infty tdS(t) = -tS(t)|_0^\infty + \int_0^\infty S(t)dt = \int_0^\infty S(t)dt$
- **p'th percentile:** Defined to be the value t(p), such that F{t(p)} = p/100. In terms of survival, t(p) is such that S{t(p)} = 1 – (p/100)
- The p'th percentile of the estimated survival:

$$\hat{t}(p) = \min\{ t_i \mid \hat{S}(t_i) \le 1 - (p/100)\}$$

- Example: Medians of two treatment groups of prostatic cancer patients (Table 1.4, p10). Use the plot from lecture one



- **Confidence intervals for the median and percentiles by the delta-method.**