Seminar in Statistics: Survival Analysis

Chapter 2

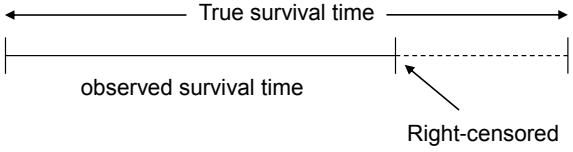
Kaplan-Meier Survival Curves and the Log-Rank Test

Linda Staub & Alexandros Gekenidis

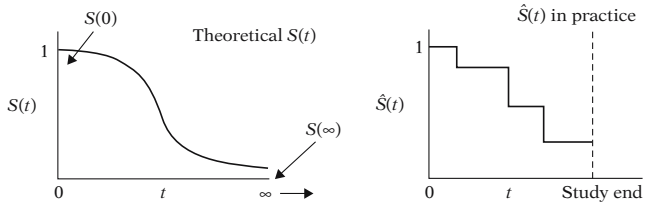
March 7th, 2011

1 Review

- Outcome variable of interest: time until an event occurs
- Time = survival time
  Event = failure
- Censoring: Don't know survival time exactly



- **Model**
  T = failure time with distribution F, density f
  C = censoring time with distribution G, density g
  Assume that the censoring time C and the failure time T are independent
  $X = \min(T, C)$, $\Delta = 1_{\{T \leq C\}}$. We observe n i.i.d. copies of (X, Δ).
- Survivor function: $S(t) = P[T > t]$



- $t_{(j)}$ = ordered failure times
  m_j = # of failures at $t_{(j)}$
  $q_{(j)}$ = # censored in $[t_{(j)}, t_{(j+1)})$
  $R(t_{(j)})$ is the set of individuals for whom $T \geq t_{(j)}$
  $n_j = |R(t_{(j)})|$

2 KM-Curves

General KM-formula

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \hat{Pr}(T > t_{(i)} \mid T \geq t_{(i)}) = \hat{S}(t_{(j-1)}) \cdot \hat{Pr}(T > t_{(j)} \mid T \geq t_{(j)})$$

In the case of no censoring this reduces to

$$\hat{S}(t_{(j)}) = \frac{\#\text{ surviving past } t_{(j)}}{\text{total number of people at the beginning}} = \frac{|R(t_{(j+1)})|}{|R(t_{(0)})|}$$

***R*-Command:** survfit(Surv(time, status) $\sim$ 1)

Remark: KM-estimator is the nonparametric MLE.

3 Log-Rank Test

- Want to find out whether the true survival curves differ from group to group
- We look at 2 groups ($\rightarrow$ extension to several groups possible)
- Nullhypothesis
  H_0 : no difference between survival curves
- Goal: To find an expression (depending on the data) from which we know the distribution (or at least approximately) under the nullhypothesis
  $\rightsquigarrow$ Test statistic: $\frac{(O_2-E_2)^2}{Var(O_2-E_2)} \sim \chi^2_1$
  where O_2 = total # failures in group 2, E_2 = expected # failures in group 2

***R*-Command:** survdiff(Surv(time, status) $\sim$ treatment)

Remark: Extension to $G \geq 2$ groups possible
$\rightarrow$ needs (co)variances of $O_i - E_i \rightarrow$ log-rank statistic $\sim \chi^2$ with $G-1$ df

3.1 Stratified Log-Rank Test

- Variation of Log-Rank Test
- Allows controlling for additional ("stratified") variable
- Split data into stratas, depending on value of stratified variable
- Calculate $O - E$ scores within strata
- Sum $O - E$ across strata
- **Limitation:** Sample size may be small within strata

***R*-Command:** survdiff(Surv(time, status) $\sim$ treatment + strata(strat))

References

[1] Kleinbaum, D.G. and Klein, M. (2005). *Survival Analysis. A self-learning text.* Springer.

[2] Maathuis, M. (2007). *Survival analysis for interval censored data. Part I.*