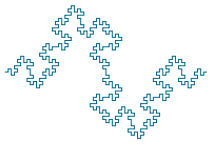
High-Throughput Sequencing Course

Statistics with Censored Data

Biostatistics and Bioinformatics



Summer 2017

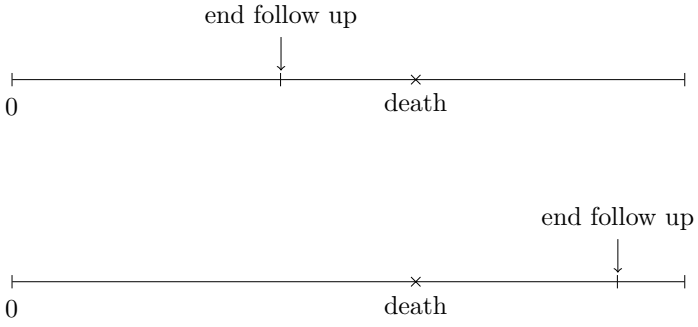
Duke University
School of Medicine

Censoring

- In many experiments, the event of interest may not have been realized at the time of the analysis
- Example: Time of death
- At the time of the analysis, the time of death for mice who are still alive is unkown
- Death will occur in the future
- All we can say is that the time of death will be greater than the current observed lifetime
- In Statistics, we use the term censoring to describe this type of data
- There are multiple types of censoring mechanisms
- We will look at three standard mechanisms

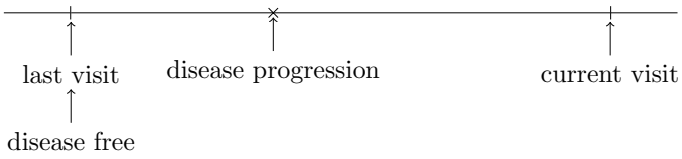
Right Censoring

Large values are censored (e.g., time of death)



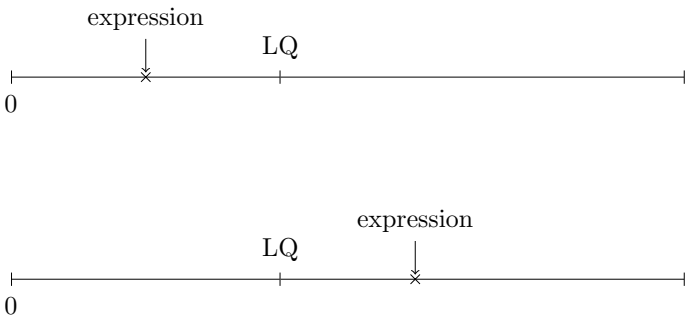
Interval Censoring

Disease progression occurs after the *last* visit (where patient was assessed to be in remission) and before the *current* visit (where patient was assessed to have relapsed)



Left Censoring

Small values are censored (e.g., Below Quantifiable Limit; low sequencing depth)



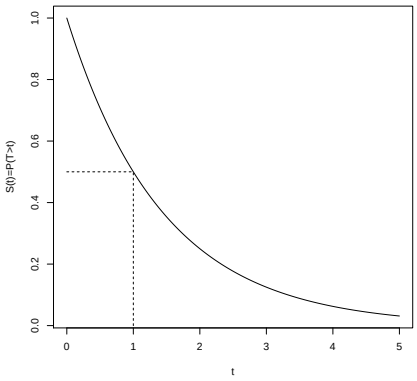
Survival Distribution

- Let T denote time of death
- Then $T > t$ denotes the event of surviving longer than time t
- $P(T > t)$ denote the probability of the event of surviving longer than time t
- How does one estimate $S(t) = P(T > t)$
- Let's consider this question assuming that there is no censoring.
- In other words, at the time of the analysis the death time for each mouse has been observed

Survival Distribution: Properties

- Let T denote the time of death measured in weeks
- $P(T > 0) = 1$
- Why?
- $P(T > 1)$: Probability of surviving longer than one week
- $P(T > 2)$: Probability of surviving longer than two weeks
- $P(T > 1) \geq P(T > 2)$
- Why?
- More generally, if $t_1 < t_2$ then $P(T > t_1) \geq P(T > t_2)$
- In other words, the survival function is a decreasing (actually non-increasing) fuction of time
- It decreases from 1 to 0

Survival Distribution: Example



Survival Distribution: Example

Simulate death times from an exponential distribution with median 1

```
set.seed(12316)
deathtimes <- rexp(10, rate = log(2))
sort(deathtimes)
```

```
##  [1] 0.1596587 0.5393022 0.6823364 0.7314445 0.7541258 1.6637266 1.8274204
##  [8] 2.3545131 3.7831739 4.0069001
```

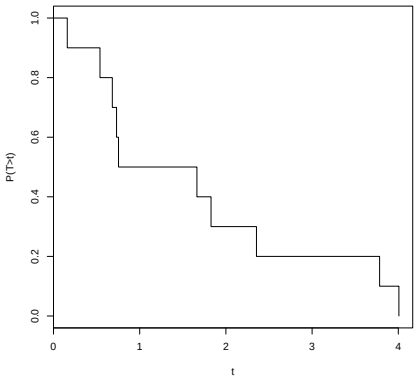
Note: If $P(T > m) = 0.5$, we say that m is the median tim of death

Survival Distribution: Example

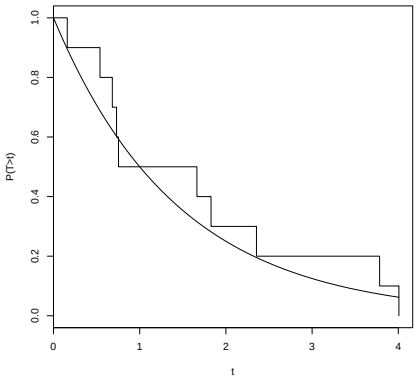
```
round(sort(deathtimes), 3)

##  [1] 0.160 0.539 0.682 0.731 0.754 1.664 1.827 2.355 3.783 4.007
```

- How many death times are greater than 0.16?
- 10/10
- How many death times are greater than 0.539?
- 9/10
- ...
- How many death times are greater than 3.783?
- 1/10
- How many death times are greater than 4.007?
- 0

Empirical Survival Function



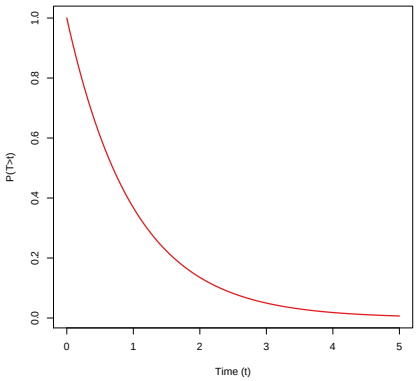
Empirical Survival Function



Survival Distribution

- We will focus on right censoring
- T= Time of Death
- C= Censoring Time
- $Y = \min\{T, C\}$= observed time
- What we want to study is the survival distribution
- $P(T > t)$ the proportion of mice in the population whose lifetime exceeds t time units
- Note that we only observe T (the time of interest) if $T \leq C$
- We define the event indicator as $D = 1$ (e.g. dead) if $T \leq C$ or $D = 0$ (e.g. alive or censored) otherwise
- We observe the pair (Y, D) *not* T

Survival Distribution



Survival Distribution: Censoring

- Note that we only observe T (the time of interest) if $T < C$
- So we have to estimate $P(T > t)$ not on the basis of T, but rather (Y, D)
- The Kaplan-Meier estimator is a standard method for estimating $P(T > t)$ on the basis of (Y, D)

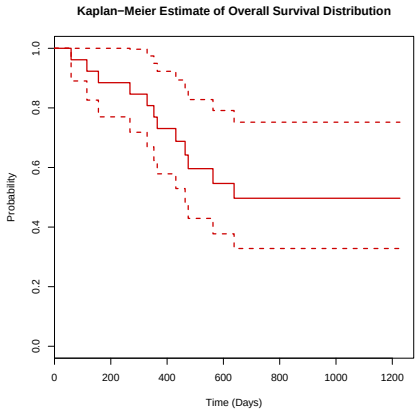
Example: Treatments for Ovarian Cancer (Edmunson *et al*, 1979)

```
head(ovarian)
```

```
##   futime fustat     age resid.ds rx ecog.ps
## 1     59      1 72.3315        2  1       1
## 2    115      1 74.4932        2  1       1
## 3    156      1 66.4658        2  1       2
## 4    421      0 53.3644        2  2       1
## 5    431      1 50.3397        2  1       1
## 6    448      0 56.4301        1  1       2
```

Kaplan-Meier

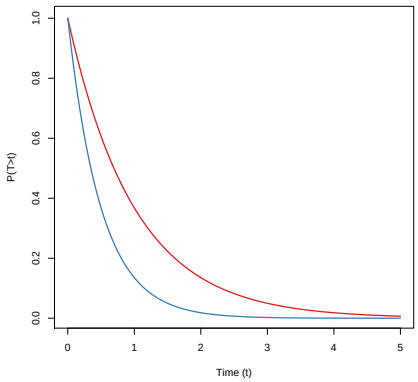
```
plot(survfit(Surv(futime, fustat) ~ 1, data = ovarian), lwd = 2, col = "red3",
    xlab = "Time (Days)", ylab = "Probability", main = "Kaplan-Meier Estimate of Overall Survival Distribu
```



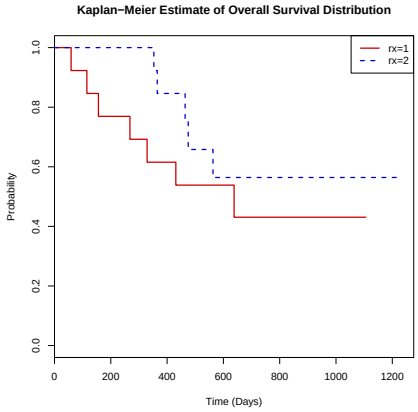
Two-sample hypothesis for Survival Data

- Let $P(T > t|Z = 0)$ denote the survival probability, at time t, if mouse is not treated
- Let $P(T > t|Z = 1)$ denote the survival probability, at time t, if mouse is treated
- Null: $H_0 : P(T > t|Z = 0) = P(T > t|Z = 1)$ for *all* t
- Alternative: $H_1 : P(T > t|Z = 0) \neq P(T > t|Z = 1)$ for *some* t

Survival Distribution: Two sample



Example: Treatments for Ovarian Cancer (Edmunson *et al*, 1979)



Logrank Test: Example

The log-rank test can be used to test if the survival probability depends on a factor

```
survdiff(Surv(futime, fustat) ~ rx, data = ovarian)
```

```
## Call:
## survdiff(formula = Surv(futime, fustat) ~ rx, data = ovarian)
##
##       N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=1 13        7     5.23     0.596      1.06
## rx=2 13        5     6.77     0.461      1.06
##
##  Chisq= 1.1  on 1 degrees of freedom, p= 0.303
```

Cox Score Test

The log-rank statistic is also called the Cox score statistic

```
coxmod <- coxph(Surv(futime, fustat) ~ rx, data = ovarian)
summary(coxmod)


## Call:
## coxph(formula = Surv(futime, fustat) ~ rx, data = ovarian)
##
##   n= 26, number of events= 12
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## rx -0.5964    0.5508   0.5870 -1.016     0.31
##
##    exp(coef) exp(-coef) lower .95 upper .95
## rx    0.5508      1.816    0.1743      1.74
##
## Concordance= 0.608  (se = 0.078 )
## Rsquare= 0.04   (max possible= 0.932 )
## Likelihood ratio test= 1.05  on 1 df,   p=0.3052
## Wald test            = 1.03  on 1 df,   p=0.3096
## Score (logrank) test = 1.06  on 1 df,   p=0.3026
```

Cox Proportional Hazards Model

The Cox Proportional Hazards Model can be used to model the hazard of the event as a function of baseline covariates

```
coxmod <- coxph(Surv(futime, fustat) ~ rx + log10(age) + resid.ds + ecog.ps,
    data = ovarian)
summary(coxmod)


## Call:
## coxph(formula = Surv(futime, fustat) ~ rx + log10(age) + resid.ds +
##     ecog.ps, data = ovarian)
##
##   n= 26, number of events= 12
##
##                  coef  exp(coef)   se(coef)      z Pr(>|z|)
## rx         -1.028e+00  3.576e-01  6.476e-01 -1.588   0.1123
## log10(age)  1.545e+01  5.111e+06  6.114e+00  2.526   0.0115 *
## resid.ds    9.203e-01  2.510e+00  7.869e-01  1.170   0.2422
## ecog.ps     3.720e-01  1.451e+00  6.460e-01  0.576   0.5646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## rx         3.576e-01  2.796e+00    0.1005 1.273e+00
## log10(age) 5.111e+06  1.956e-07   31.9232 8.184e+11
## resid.ds   2.510e+00  3.984e-01    0.5368 1.174e+01
## ecog.ps    1.451e+00  6.893e-01    0.4090 5.145e+00
##
## Concordance= 0.803  (se = 0.091 )
## Rsquare= 0.472   (max possible= 0.932 )
## Likelihood ratio test= 16.61  on 4 df,   p=0.002302
## Wald test            = 13.7  on 4 df,   p=0.008322
## Score (logrank) test = 19.27  on 4 df,   p=0.0006955
```