NRES_798_14_201501

Survival Analysis

		Lecture	Due
Friday	March 20	Generalized additive models	
Monday	March 23		Return paper 1
Wednesday	March 25	Survival analysis	
Friday	March 27	Model selection	
Monday	March 30		
Wednesday	April 1	Time series analysis	
Friday	April 3	UNBC closed good Friday	
Monday	April 6	UNBC closed Easter Monday	
Wednesday	April 8	Spatial statistics	
Friday	April 10	Regression trees (data mining)	
Monday	April 13		Lab/lecture Final
Wednesday	April 15	Multivariate statistics	
Friday	April 17	Mixed-effects models	Final paper

First report

- Hypothesis/model to be tested (15%)
  - Scientific rational for analysis
  - Framing the scientific question in statistically appropriate way
- Description of data (5%)
  - Experimental design, dependent & independent variables
  - Descriptive statistics, distributions, outliers
- Limitations of data (10%)
  - Problems
  - Experimental design limitation, sampling restraints, measurement error
- Sources of uncertainty/variability (10%)
  - What types of uncertainty can be examined, and what is unknown
- Historical approaches used for analysis (20%)
- Alternative statistical approaches (40%)
  - Comparative: strengths, weaknesses and differences of alternative approaches
  - Limitation (inappropriate because …)

Final report

- Scientific paper with **heavy** emphasis placed on statistical analysis
  - Statistics methods paper
    - Intro (15)
      - Scientific question, emphasis on statistical framing of hypothesis being tested
      - Description of statistical "problem"
      - Description of why stats matter
      - Description of statistical approaches
    - Methods (15)
      - Statistically oriented, clear description of stats applied
    - Results (30)
      - Presentation interpretation
    - Discussion (30
      - Detailed interpretation of statistical results
      - Evaluation of shortcomings of analysis
      - Discussion of results in the context of
    - Literature cited (5)
    - Appendix: R code for analysis (5)

Survival analysis

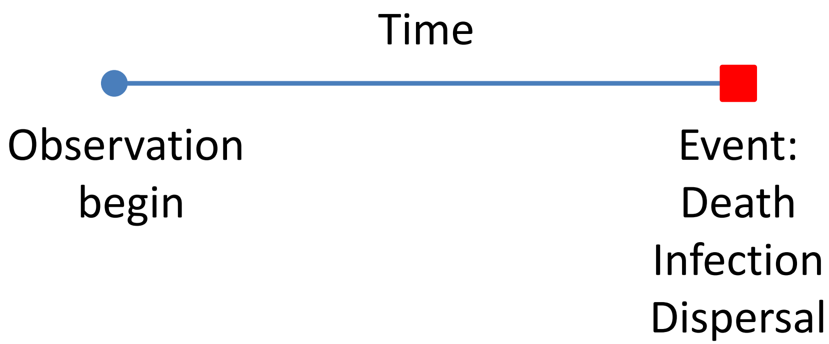
- Examines and models the time it take for events to occur
  - The event can be death, therefor “Survival analysis”
- Other names
  - “event-history analysis” : sociology
  - “failure-time analysis” : engineering

Survival analysis

- Classically, the analysis focuses on time to death
  - But can be used anywhere you want to know what factors affect the time **for** an event to occur:
    - Germination timing
    - Arrival of a migrant or parasite
    - Dispersal of seeds or offspring
    - Failure time in mechanical systems
    - Response to stimulus

Survival data

- Start of observation period (not real time)
- Time from start that event occurs



Challenges with this type of data?

Censoring: dealing with uncertain data

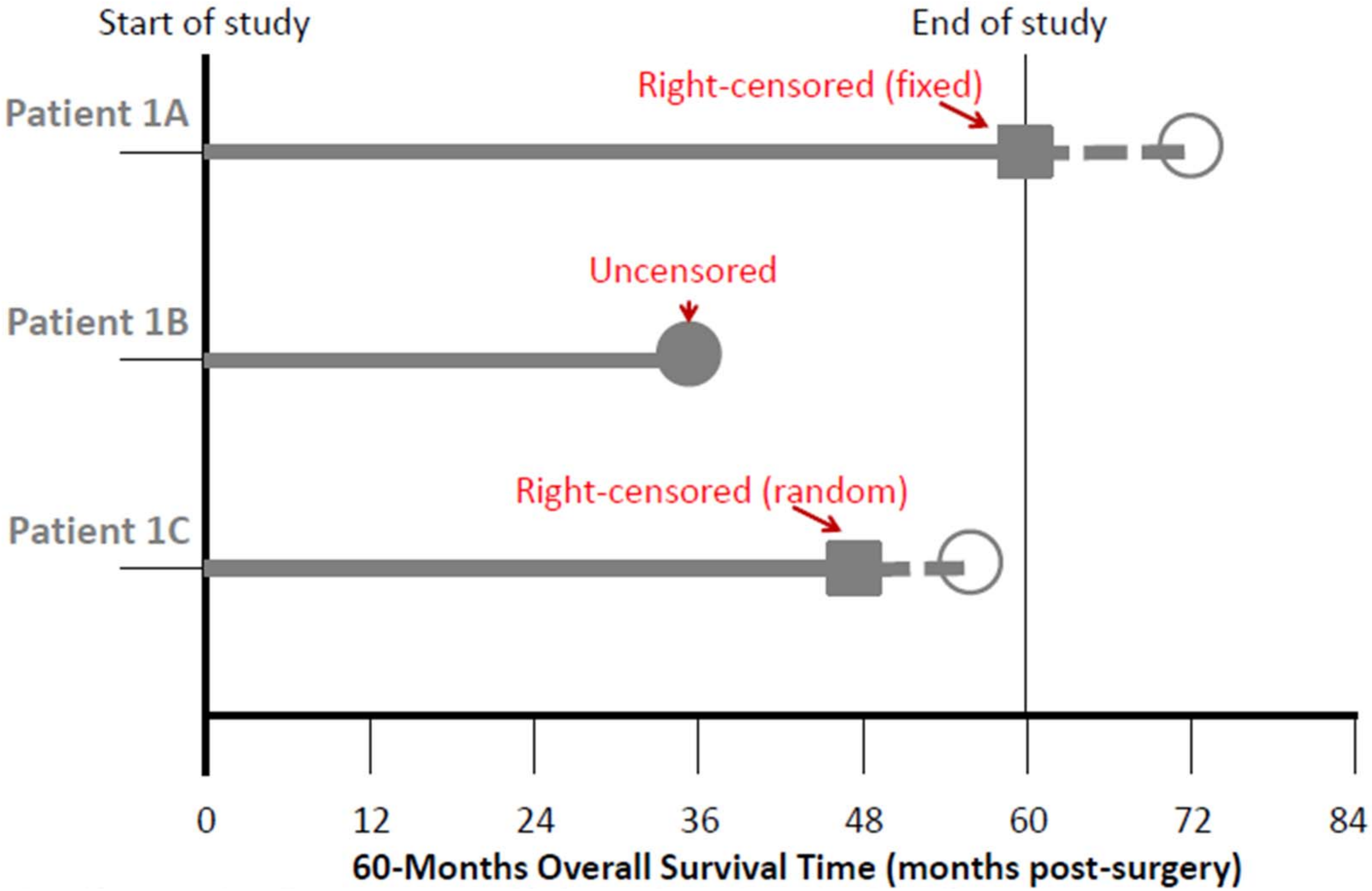
- Censored survival times:
  - problem when event has not occurred (within the observation time) or the exact time of event is not known.

- Right censoring:
  - Where the date of death is unknown but is after some known date
  - true survival time > observed survival time

e.g.

- Organism alive at end of the observation period (study)
- Subject is removed from the study
  - animal escapes, animal gets lost, plant gets eaten, etc.

Censoring



Start of study
End of study
Patient 1A
Right-censored (fixed)
Patient 1B
Uncensored
Patient 1C
Right-censored (random)
0
12
24
36
48
60
72
84
60-Months Overall Survival Time (months post-surgery)

Censoring

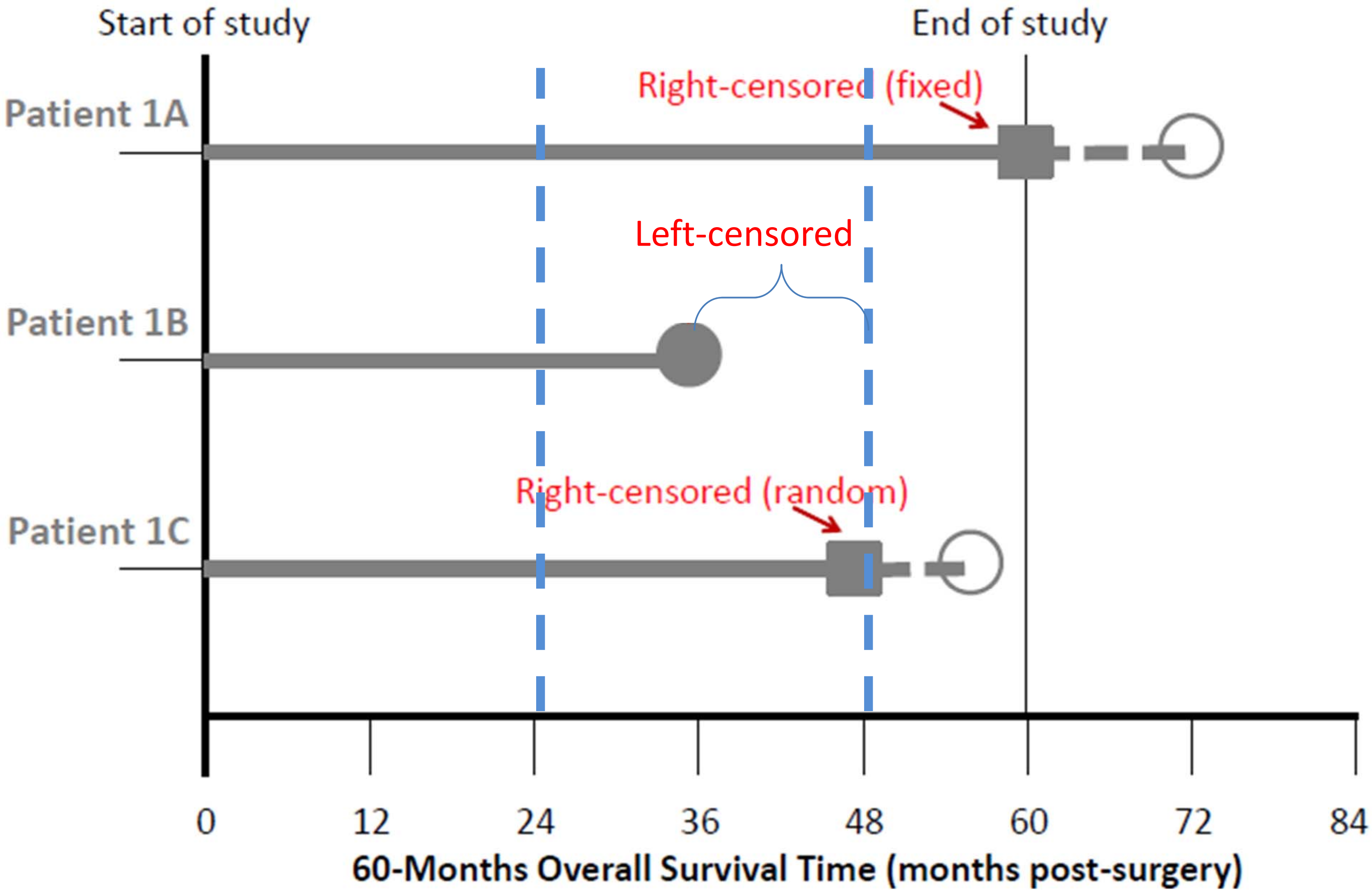
- Left censoring:
  - Occurs when a subject's survival time is incomplete on the left side of the follow-up period.
  - True survival time < Observed survival time
  - Exact timing of event is uncertain: e.g..

e.g.

- We want to know time to infection, but only assess infection when tested

Censoring must be independent of the event being looked at

Censoring

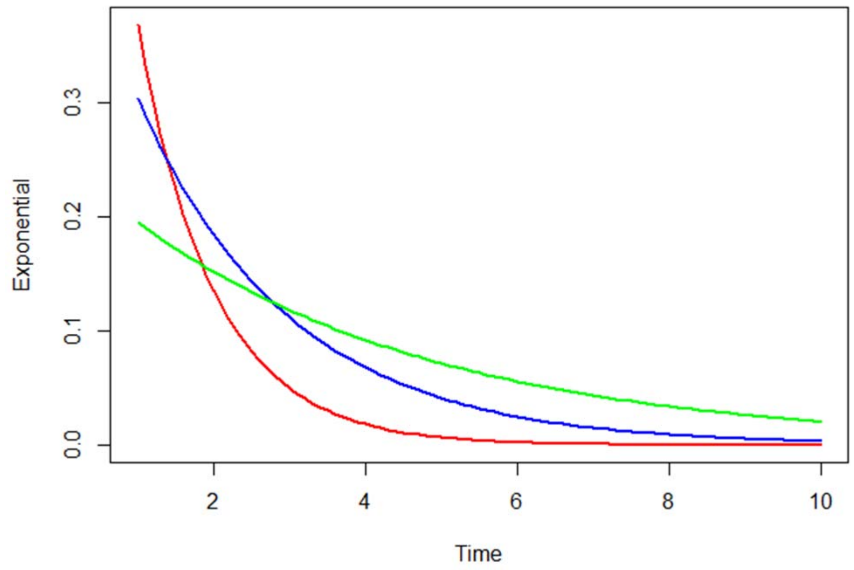


Start of study
End of study
Patient 1A
Right-censored (fixed)
Left-censored
Patient 1B
Right-censored (random)
Patient 1C
0
12
24
36
48
60
72
84
60-Months Overall Survival Time (months post-surgery)

Survival

- Survival time T may be though of as a random variable
- T can be represented as a probability density function
- The simplest parametric model is the exponential distribution, with density function:

$$p(t) = \lambda e^{-\lambda t}$$



- In this distribution there is a single rate parameter (λ)
  - In this distribution the rate is assumed to be constant over time
- Other distributions (that are based on more biologically/ecologically sound principles) can also be used: Gompertz, Weibull, Gamma

Survival function: S(t), survival curve

- The survival function gives probability of surviving to time t.
  - i.e. the proportion of the population still without the event by time t.

- The survival function is the complement of the cumulative distribution function.

$$S(t) = \Pr(T > t) = 1 - P(t)$$

- **Hazard rate** is the continuous analog of an age-specific mortality rate.
  - i.e. the probability of dying at time t (death between time t1 and t2)

- **Hazard function [h(t)]** is the hazard rate as a function of survival time.
  - Give the instantaneous potential per unit time for the event to occur, given the individual has survived to time t

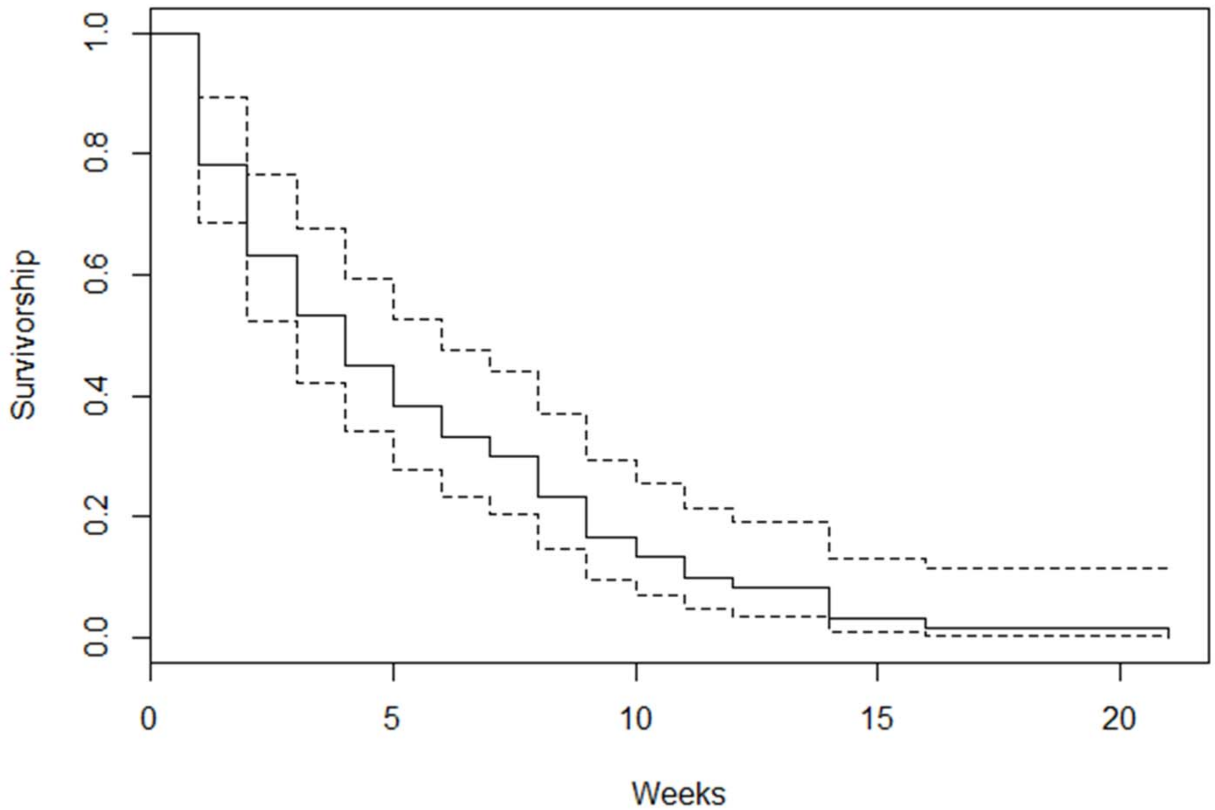
  - e.g. the hazard of death in human populations is relatively high in infancy, declines during childhood, stays relatively steady during early adult hood, and rises through middle and old age.

  - This is why the exponential distribution (which assumes a constant hazard rate) is not appropriate to use in a survival analysis of human (biological) populations

Estimated/Empirical survival curves

- Survival curve is estimated by Kaplan-Meier (KM) estimator, also know as "product estimator"
- The Kaplan-Meier estimate is a nonparametric maximum likelihood estimate of the survival function, S(t)
- The estimate is a step function with jumps at observe event times

Kaplan-Meier estimate



Events (deaths)

$$\hat{S}_{KM} = \prod_{t_i < t} \frac{r(t_i) - d(t_i)}{r(t_i)}$$

At risk

Using explanatory variable to inform survival time estimates

- Parametric models
  - GLM framework using:
    - Exponential, gamma, lognormal or Weibull distribution
    - Use function survfit()

- Non-parametric models
  - Cox proportional hazards model
    - Use function coxph()

Cox Proportional Hazard Model

- Popular model for survival analysis because its simple and makes no assumption about the survival distribution

$$h_i(t) = h_o(t)\exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

$$h_i(t|age) = h_o(t)\exp(\beta_1 X_{i1} + age * \beta_1)$$

Age at beginning of observation

- Is a semi-parametric model
  - The baseline hazard function is unspecified
  - The effects of the covariates are multiplicative
  - The model doesn't make any arbitrary assumptions about the shape/form of the baseline hazard function

Cox proportional hazards model

Assumptions

- Covariates multiply the hazard by some constant
  - e.g. drug may halve a subjects risk of death at any time
- The effect of the covariate is the same at any point in time.

Goals of survival analysis

1. Estimate and interpret survival and hazard functions from survival data
(descriptive statistics)

2. Compare survival and/or hazard functions
(two-sample mean test)

3. Assess the relationship of explanatory variables to survival time
(regression analysis)

Survival analysis in R

- “survival”package
- Survival analysis components (functions)

  - Surv(): Defines a survival object
  - survdiff(): determines if two survival curves differ using a log-rank test
  - survfit(): fits a survival curve to a model or function, using Kaplan-Meier estimates. Parametric
  - coxph(): Runs a cox PH regression (Cox proportional hazards model). Non-parametric

Survival in R

- The response variable defined by Surv() includes:
  - Start time (after study start)
  - Stop time (after study start)
  - Whether or not an event occurred

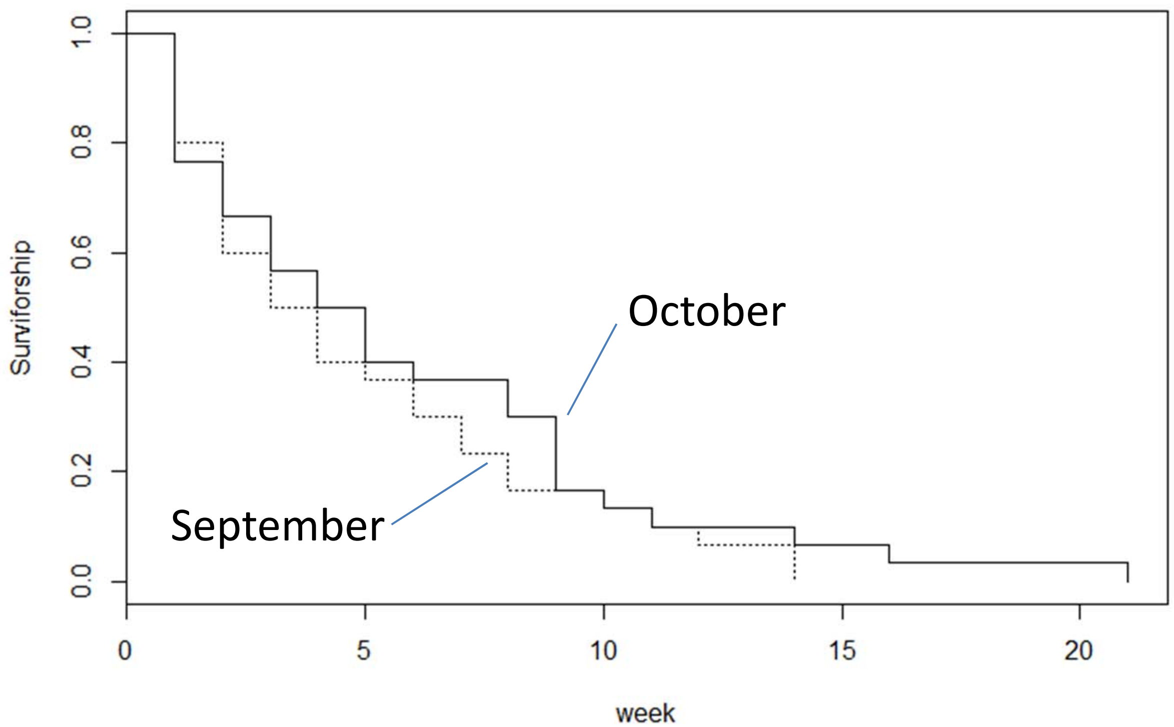
- Allows for censoring issues to be accounted for in data structure

Survival Analysis

- Example 1
  - Survival of tree seedlings
  - Does size of canopy gap influence survival

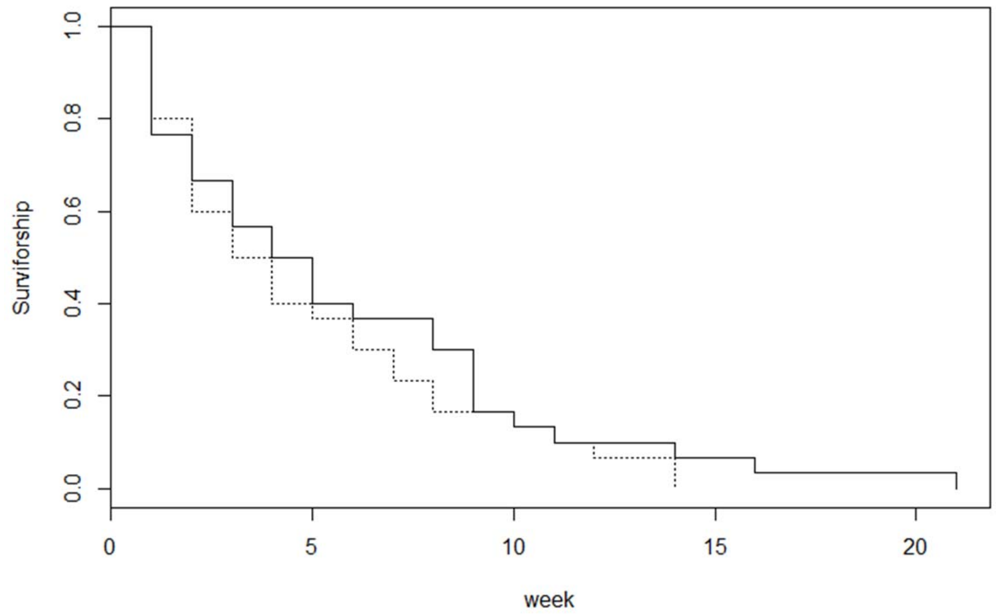
```
> head(seedlings)
   cohort     death  gapsize
1  September    7    0.5889
2  September    3    0.6869
3  September   12    0.9800
4  September    1    0.1921
5  September    4    0.2798
6  September    2    0.2607
```

Survival analysis



Survival differences between cohorts?

model <- survfit(Surv(death,status)~cohort,data=seedlings)




```
model <- survfit(Surv(death,status)~cohort,data=seedlings)
```

```
Call: survfit(formula = Surv(death, status) ~ cohort, data = seedlings)

                  records n.max n.start events median 0.95LCL 0.95UCL
cohort=October        30    30      30     30    4.5       3       9
cohort=September      30    30      30     30    3.5       2       7
```

Differences between cohorts?

Survival analysis
Cox's Proportional Hazard

model1 <- coxph(Surv(death,status)~strata(cohort)*gapsize)

```
Call:
coxph(formula = Surv(death, status) ~ strata(cohort) * gapsize)

  n= 60, number of events= 60

                                        coef  exp(coef) se(coef)      z  Pr(>|z|)
gapsize                              -1.1863     0.3054   0.6210 -1.910    0.0561 .
strata(cohort)cohort=September:gapsize  0.5795   1.7852   0.8264  0.701    0.4831
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                                       exp(coef) exp(-coef) lower .95 upper .95
gapsize                                   0.3054     3.2749   0.09042     1.031
strata(cohort)cohort=September:gapsize    1.7852     0.5602   0.35341     9.018

Concordance= 0.659  (se = 0.077 )
Rsquare= 0.076   (max possible= 0.993 )
Likelihood ratio test= 4.73  on 2 df,   p=0.09372
Wald test            = 4.89  on 2 df,   p=0.08682
Score (logrank) test = 5.04  on 2 df,   p=0.08046
```

Survival Analysis

- Example 2
  - Survival of Cockroaches to three insecticide applications (A,B,C)
  - Does weight of the animal influence their survivorship?

	death	status	weight	group
1	20	1	5.385	A
2	34	1	7.413	A
3	1	1	9.266	A
4	2	1	6.228	A
5	3	1	5.229	A
6	3	1	9.699	A

```
> summary(insects)
    death            status            weight            group
 Min.   : 1.00    Min.   :0.0000    Min.   : 0.055     A:50
 1st Qu.: 1.00    1st Qu.:1.0000    1st Qu.: 2.459     B:50
 Median : 7.00    Median :1.0000    Median : 6.316     C:50
 Mean   :15.17    Mean   :0.8667    Mean   : 9.390
 3rd Qu.:21.00    3rd Qu.:1.0000    3rd Qu.:11.955
 Max.   :50.00    Max.   :1.0000    Max.   :42.090
```

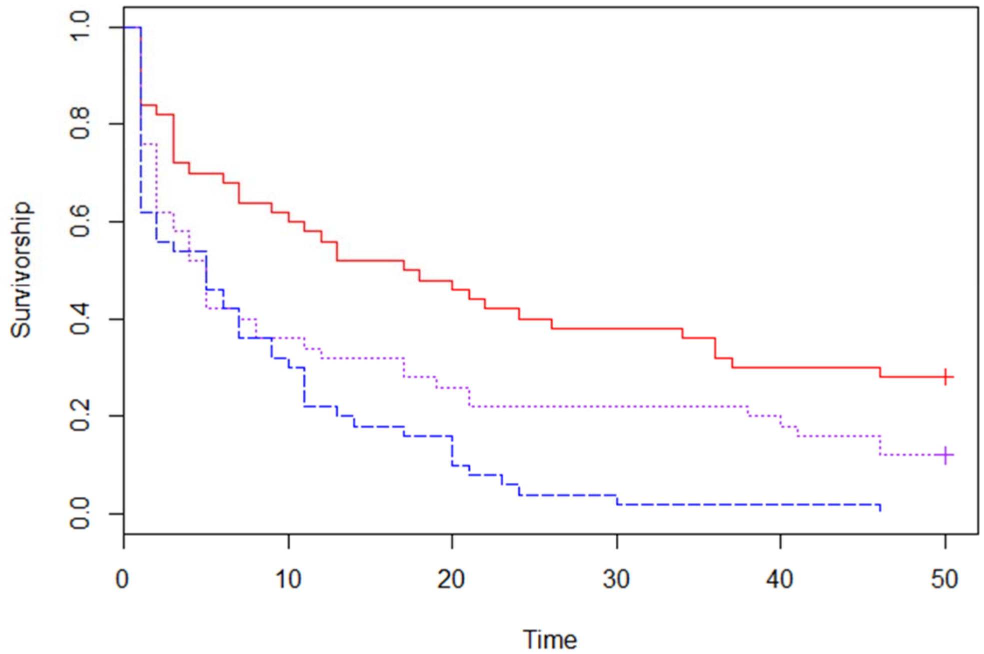
Survival analysis

	death	status
1	20	1
2	34	1
3	3	1

- Create a survival analysis data object
  - sdat <- Surv(insects$death,insects$status)

- Fit a survival curve to the raw data, seperating by group (treatment)
  sdat_fit <- survfit(sdat~insects$group)

- Plot the fitted curves
  plot(sdat_fit,lty=c(1,3,5),col=c("red","purple","blue"),ylab="Survivorship",xlab="T
  ime")



Survival analysis

- Parametric and non-parametric models

```
# Create the response variable
sdat <- Surv(insects$death,insects$status)

# Parametric model
pmod <- survreg(sdat~insects$weight*insects$group,dist="weibull")

# Cox proportional hazards regression model
non_pmod <- coxph(sdat~insects$weight*insects$group)
```

Parametric survival analysis

```
> summary(pmod)

Call:
survreg(formula = sdat ~ insects$weight * insects$group, dist = "weibull")
                                Value    Std. Error        z        p
(Intercept)                     3.9506   0.5308           7.443    9.84e-14
insects$weight                  -0.0973   0.0909          -1.071    2.84e-01
insects$groupB                  -1.1337   0.6207          -1.826    6.78e-02
insects$groupC                  -1.9841   0.6040          -3.285    1.02e-03
insects$weight:insects$groupB  0.0826    0.0929           0.889    3.74e-01
insects$weight:insects$groupC  0.0931    0.0930           1.002    3.16e-01
Log(scale)                      0.3083   0.0705           4.371    1.24e-05

Scale= 1.36

Weibull distribution
Loglik(model)= -469.6   Loglik(intercept only)= -483.3
        Chisq= 27.42 on 5 degrees of freedom, p= 4.7e-05
Number of Newton-Raphson Iterations: 5
n= 150
```

Cox ph survival analysis

```
> summary(non_pmod)
Call:
coxph(formula = sdat ~ insects$weight * insects$group)

  n= 150, number of events= 130

                                 coef      exp(coef) se(coef)      z Pr(>|z|)
insects$weight                   0.06330   1.06535  0.06738      0.940    0.34747
insects$groupB                   0.79098   2.20555  0.45641      1.733    0.08309 .
insects$groupC                   1.28634   3.61953  0.45243      2.843    0.00447 **
insects$weight:insects$groupB   -0.05568   0.94585  0.06878     -0.809    0.41824
insects$weight:insects$groupC   -0.05869   0.94300  0.06897     -0.851    0.39481
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                     exp(coef) exp(-coef) lower .95 upper .95
insects$weight                 1.0654     0.9387    0.9336     1.216
insects$groupB                 2.2056     0.4534    0.9016     5.395
insects$groupC                 3.6195     0.2763    1.4912     8.785
insects$weight:insects$groupB  0.9458     1.0573    0.8266     1.082
insects$weight:insects$groupC  0.9430     1.0604    0.8238     1.079

Concordance= 0.608  (se = 0.034 )
Rsquare= 0.135   (max possible= 0.999 )
Likelihood ratio test= 21.83  on 5 df,   p=0.0005645
Wald test            = 20.75  on 5 df,   p=0.000903
Score (logrank) test = 22.05  on 5 df,   p=0.0005132
```

```
pmod1 <- survreg(sdat~insects$group,dist="weibull")
non_pmod1 <- coxph(sdat~insects$group)
```

```
> summary(pmod1)
```

	Value	Std. Error	z	p
(Intercept)	3.459	0.2283	15.15	7.20e-52
insects$groupB	-0.822	0.3097	-2.65	7.94e-03
insects$groupC	-1.540	0.3016	-5.11	3.28e-07
Log(scale)	0.314	0.0705	4.46	8.15e-06

```
> summary(non_pmod1)
```

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
insects$groupB	0.5607	1.7520	0.2257	2.485	0.013 *
insects$groupC	1.0084	2.7412	0.2263	4.456	8.33e-06 ***