# MODULE 11: INTRODUCTION TO SURVIVAL ANALYSIS

Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Susanne May, Ph.D.
Professor
Department of Biostatistics
University of Washington

## OVERVIEW

- Session 1
  - Introductory examples
  - The survival function
  - Survival Distributions
  - Mean and Median survival time
- Session 2
  - Censored data
  - Risk sets
  - Censoring Assumptions
  - Kaplan-Meier Estimator and CI
  - Median and CI
- Session 3
  - Two-group comparisons: logrank test
  - Trend and heterogeneity tests for more than two groups
- Session 4
  - Introduction to Cox regression

# OVERVIEW – MODULE 13

Module 13: Survival analysis in Clinical Trials

- Quick review of basics
- Estimating survival after Cox model fit
- More two-sample tests
  - Weighted logrank
  - Additional tests based on functionals and metrics
- Adjustment, precision and post-randomization variables
- Power
- Choice of outcome
- Information accrual in sequential monitoring



# OVERVIEW – MODULE 17

Module 17: Survival analysis for Observational Data

- More complicated Cox models
  - Adjustment
  - Interaction
- Competing Risks
- Choice of time variable
- Left Entry/Truncation
- Immortal time bias
- Index event bias
- Time-dependent covariates

# MODULE 11
# INTRODUCTION TO SURVIVAL ANALYSIS

## SESSION 1:
## SURVIVAL DATA: EXAMPLES

## PRELIMINARIES

- No prior knowledge of survival analysis techniques assumed
- Familiarity with standard one- and two-sample statistical methods (estimation and testing) is assumed
- Emphasis on application rather than mathematical details
- Examples

## SESSIONS/BREAKS

- 8:30 – 10:00
  - Break until 10:30
- 10:30 – 12:00
  - Break until 1:30
- 1:30 – 3:00
  - Break until 3:30
- 3:30 – 5:00



## WHAT IS SURVIVAL ANALYSIS ABOUT?

- Studies the occurrence of an event over time
  - Time from randomization to death (cancer RCT)
  - Time from acceptance into a heart transplant program to death
  - Time from randomization to diagnosis of Alzheimer's Disease in a prevention trial
  - Time from randomization to ovarian cancer death in a randomized screening trial
  - Time from birth to removal of supplementary oxygen therapy
  - Time from first VTE diagnosis to recurrent VTE

## YOUR EXAMPLES



## WHAT IS SURVIVAL ANALYSIS ABOUT?

- Explores factors that are thought to influence the chance that the event occurs
  - Treatment
  - Age
  - Gender
  - Body Mass Index
  - Diet
  - Etc.

# EXAMPLE 1

- Levamisole and Fluorouracil for adjuvant therapy of resected colon carcinoma
  Moertel et a. l, 1990, 1995
- 1296 patients, enrolled 1 – 5 weeks after surgery
- Stage $B_2$ or C
- 3 unblinded treatment groups in stage C (2:1:1 ratio)
  - Observation only
  - Levamisole (oral, 1yr)
  - Levamisole (oral, 1yr) + fluorouracil (intravenous 1yr)

Moertel CG, Fleming TR, Macdonald JS, et al. (1990) NEJM: 322(6):352–358.
Moertel CG. et al (1995). Annals of Internal Medicine: 122(5):321.



# EXAMPLE 1

- Randomization
  - Dynamic method based on accrued:
  - For $B_2$, extent of invasion, time since surgery
  - For C, extent of invasion, time since surgery, number of lymph nodes involved

# EXAMPLE 1

- Statistical analysis
  - Survival primary outcome (recurrence secondary)
  - Kaplan-Meier survival curves
  - Log-rank statistic
  - Cox proportional-hazards model for all multivariable analysis
  - Backward regression, maximal partial-likelihood estimate statistic
  - O'Brien-Fleming boundary for sequential monitoring; stopped early for stage C



# EXAMPLE 1

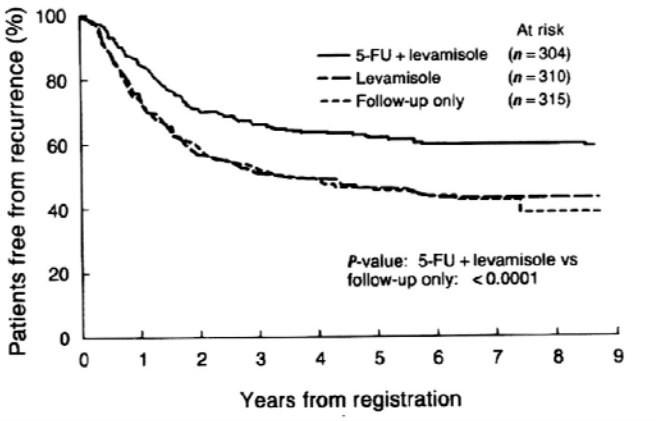



**Figure 1: Recurrence-free interval according to treatment arm. Patients who died without recurrence have been censored. 5-FU = fluorouracil.**

# EXAMPLE 1

- Results (stage C) after 2nd interim analysis
- Fluorouracil + Levamisole reduced the
  - Recurrence rate by 41% (95% CI 23% - 54%) (p<0.0001)
  - Death rate by 33% (95% CI 10% - 50%) (p<0.006)
- Levamisole reduced the
  - Recurrence rate by 2%
  - Death rate by 6%
- Toxicity was mild (with few exceptions)
- Patient compliance excellent



# EXAMPLE 1

- R survival package data "colon"
  - 929 eligible stage C patients (971 randomized – 42 ineligible)
  - Treatment groups (rx)
  - Sex, age
  - Obstruction of colon by tumor (obstruct)
  - Perforation of colon (perfor)
  - Adherence to nearby organs (adhere)
  - Number of lymph nodes with detectable cancer (nodes)
  - Days until event or censoring (time)
  - Censoring status (status)

# EXAMPLE 1

- Multivariable analysis:
  - Proportional hazards model
  - "we kept the variable of treatment in the model and used backward regression for other covariates"
  - Other covariates ($P < 0.01$)
    - Depth of primary tumor invasion,
    - Invasion of adjacent structures
    - Regional implants
    - Number of metastatic lymph nodes
    - Histological differentiation
    - Preoperative carcinoembryonic antigen level



# EXAMPLE 1

- Multivariable results: "After correction for the influence of prognostic factors through the use of a proportional hazards model, patients receiving fluorouracil plus levamisole were again found to have a significant survival advantage when compared with patients assigned to observation only; they had a 33% reduction in mortality rate (95% CI, 16% to 47%; *P* = 0.0007). Therapy with levamisole alone showed essentially no effect (6% reduction in death rate; *P* = 0.57."

  Moertel et al (1995)

## EXAMPLE 2 – ALZHEIMER'S

- Petersen et al. 2005, NEJM
- Subjects with amnestic subtype of mild cognitive impairment
- Adaptive randomization based on MMSE score, age, Apo ε4 genotype
- Three arms: Vitamin E, Donepezil, and Placebo
- Primary outcome: Time from randomization to possible or probable AD diagnosis
- Length of double-blind treatment: 3 years

Petersen RC, Thomas RG, Grundman M. et al. (2005) NEJM. 352(23):2379–2388.



## EXAMPLE 2 – ALZHEIMER'S

- Primary analysis: Cox regression adjusted for randomization influencing variables MMSE score, age and Apo E genotype
- 769 enrolled: 253 donepezil, 257 vitamin E, 259 placebo
- 230 dropped out: 92 donepezil, 74 vitamin E, 66 placebo
  - Treatment related toxicity: GI complaints, muscle aches, insomnia
- Dropout was observed to be related to MMSE score

## EXAMPLE 2 – ALZHEIMER'S

- 212 developed possible or probable AD
- "There were no significant differences … during the three years of treatment"
- Vitamin E vs Placebo
  - Hazard Ratio 1.02 (95% CI, 0.74, 1.41), p-value 0.91
- Donepezil vs Placebo
  - Hazard Ratio 0.80 (95% CI, 0.57, 1.13), p-value 0.42

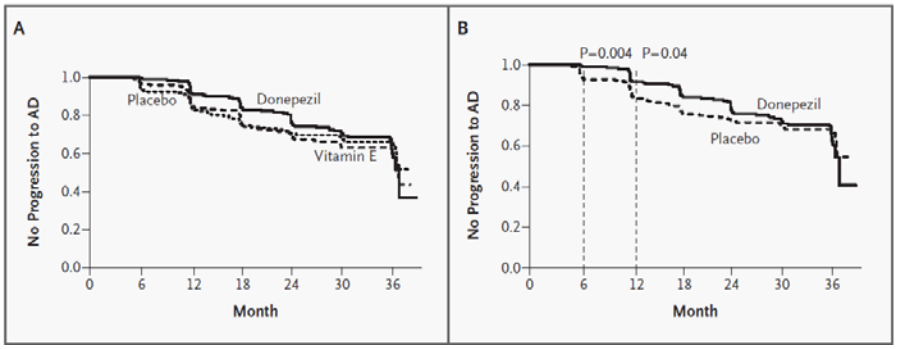


## EXAMPLE 2 – ALZHEIMER'S

- Prespecified analyses
- At 6 months intervals
  - Donepezil vs Placebo significantly reduced likelihood of progression to AD during the first 12 months (p-value 0.04)
  - Finding supported by secondary outcome measures
  - Subgroup ≥ 1 apolipoprotein E ϵ4 alleles significantly reduced likelihood of progression to AD over 3 years
  - Vitamin E vs Placebo: no significant differences
  - Vitamin E vs Placebo: also no significance for above subgroup
- Simulations assuming informative treatment-related dropout did not change primary conclusions

# EXAMPLE 2 – RESULTS

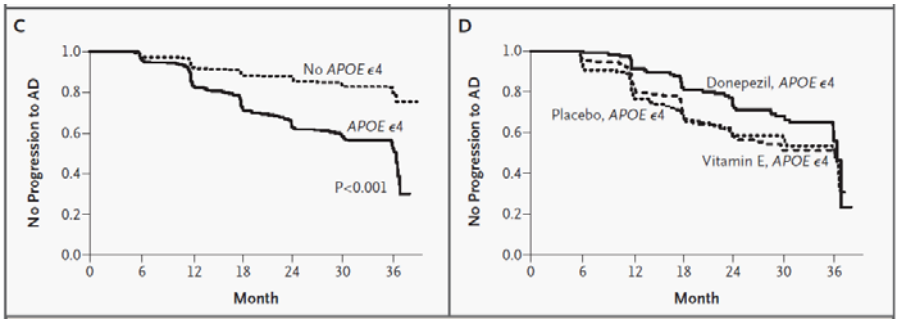
- Overall and at 6 and 12 months






# EXAMPLE 2 – RESULTS

- APOE ϵ4 results



## EDITORIAL

- "long-awaited results"
- Donepezil standard therapy for AD
- "Implications …. Enormous"
  - Clear-cut negative findings for Vitamin E
  - Especially noteworthy
  - Despite dearth of evidence of its efficacy
  - Findings for donepezil "much less clear"
  - "not quite as disappointing"

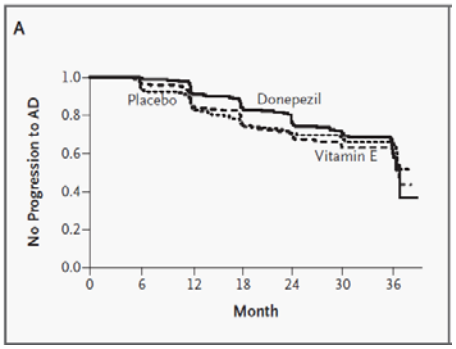


## EDITORIAL COMMENTS

- "rate of progression … somewhat lower in the treatment group during the first year of the study"
- "by two years, even this small effect had worn off"
- Possible explanation: "Reduced statistical power later in the study as the number of subjects at risk declined owing to death, withdrawal and development of AD
- Secondary analyses suggest… benefits wore off

# EXAMPLE 2 – RESULTS

- Interesting steps…..






# SCREENING TRIAL

- 202,546 women 50-72 years of age, England, Wales, Northern Ireland
- Randomized to one of three arms in 1:1:2 ratio between June 1, 2001 and Oct 21, 2005.
  - Annual multimodal screening (serun CA 125 + algorithm)
  - Annual transvaginal ultrasound
  - No screening
- Screening ended Dec 31, 2011.
- Not blinded
- Primary outcome: death from ovarian cancer (by end of 2014)

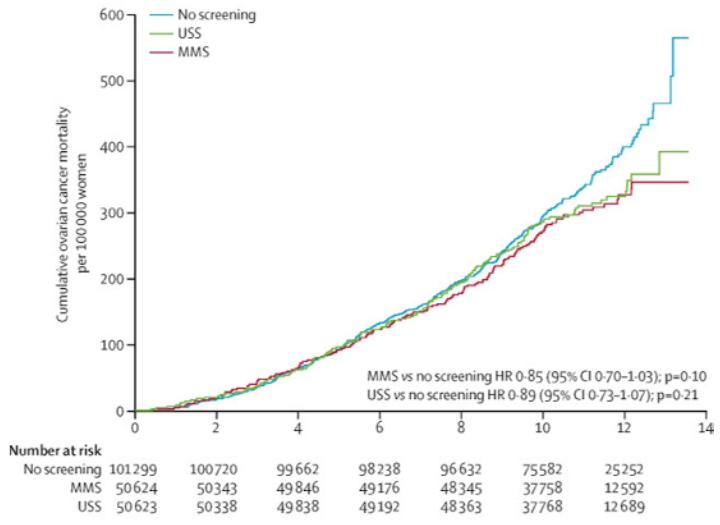
Jacobs IJ, Menon U, Ryan A, et al. (2016) The Lancet. 387(10022):945–956.

# OVARIAN CANCER SCREENING TRIAL

- Primary analysis: Cox regression (proportional hazards)
  - MMS vs. no screening: Mortality reduction = (1 – HR)100 = 15% (95% CI: -1% – 33%) P = .10
  - USS vs. no screening: Mortality reduction = (1 – HR) 100 = 11% (95% CI: -7% - 27%) P = .21



# OVARIAN CANCER SCREENING TRIAL



## OVARIAN CANCER SCREENING TRIAL

- Why the delayed difference?



## OVARIAN CANCER SCREENING TRIAL

- Secondary analyses, excluding prevalent cases:
- Post-hoc Weighted* logrank test:
  - MMS mortality reduction = 22% (3-38%) P = .023
  - USS mortality reduction = 20% (0 – 35%) P = .049

* by pooled cumulative mortality

## "COUNTER" EXAMPLE

- Resuscitation Outcomes Consortium
  - Out-of-hospital cardiac arrest
  - Traumatic injury
- Prehospital interventions
- Exception from informed consent
- 10 Regional Centers
  - 7 US
  - 3 Canada



## "COUNTER" EXAMPLE

- Times
  - Event (cardiac arrest, traumatic injury)
  - 911 call
  - Arrival of EMS
  - Treatment start
  - Potential outcomes
    - Return of spontaneous circulation (Cardiac arrest)
    - ED admission
    - Survival to hospital discharge
    - Neurologically intact survival
    - 28-day survival
    - 6-month neurological outcomes

## "COUNTER" EXAMPLE

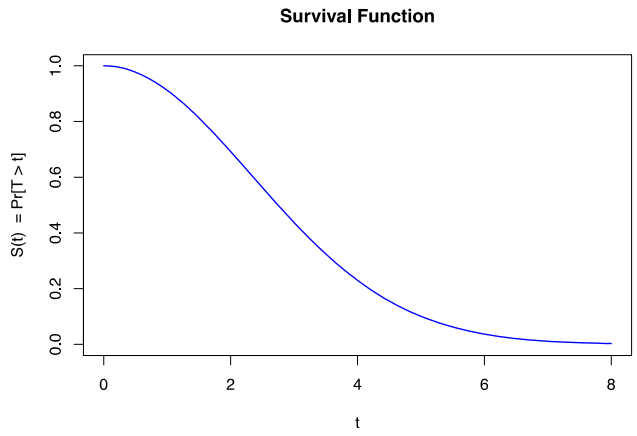
- Time of injury/cardiac arrest (ordinarily unknown)
- 911 call
- Cardiac arrest: Many deaths before admission to hospital
- Trauma: Many deaths within the first 24 – 48 hours



## SURVIVAL DATA AND FUNCTION

- Original applications in biometry were to survival times in cancer clinical trials
- Many other applications in biometry: eg. disease onset ages
- Interest centers not only on average or median survival time but also on probability of surviving beyond 2 years, 5 years, 10 years, etc.
- Best described with the entire survival function S(t).
  - For T = a subject's survival time, $S(t) = P[T > t]$.
  - Characterizes the entire distribution of survival times T.
  - Gives useful information for each t.

# SURVIVAL FUNCTION

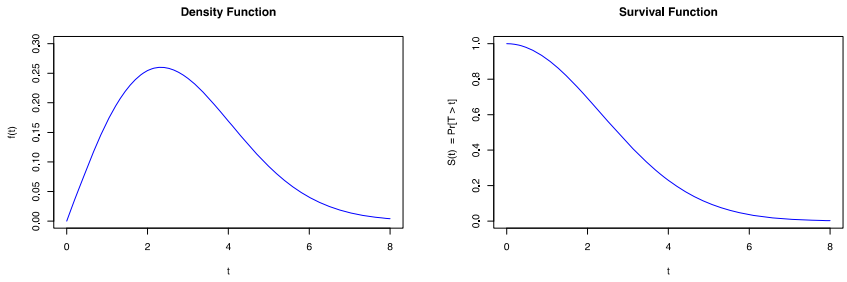





# SURVIVAL DISTRIBUTION

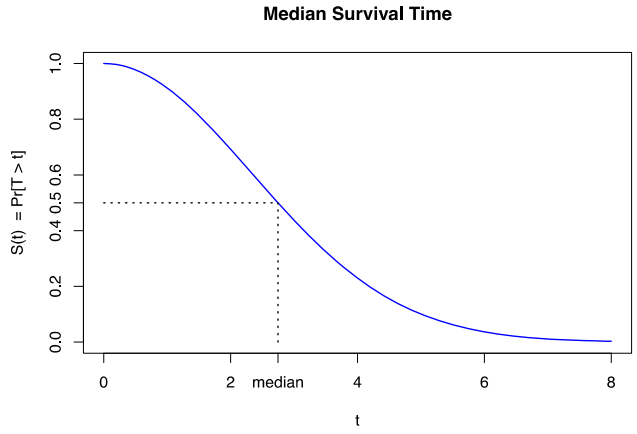
- Continuous probability distribution of times $T$
- Only non-negative $T$'s are possible: $\Pr(T<0)=0$
- Density function $f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr(t \le T < t + \Delta t)$
- Area under the $f(t)$ curve between two points is the probability $T$ is between the two points.

# DENSITY AND SURVIVAL FUNCTIONS

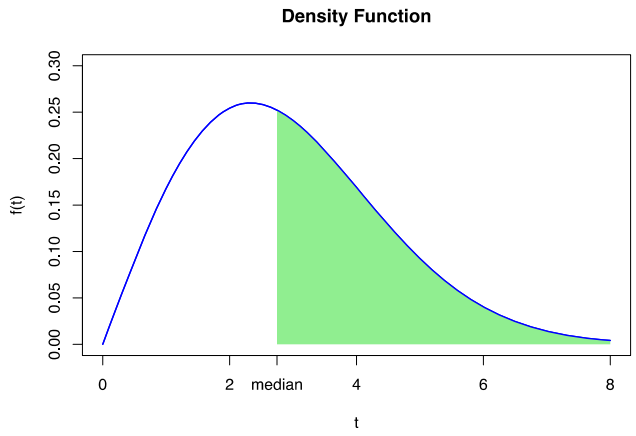





# MEDIAN SURVIVAL TIME

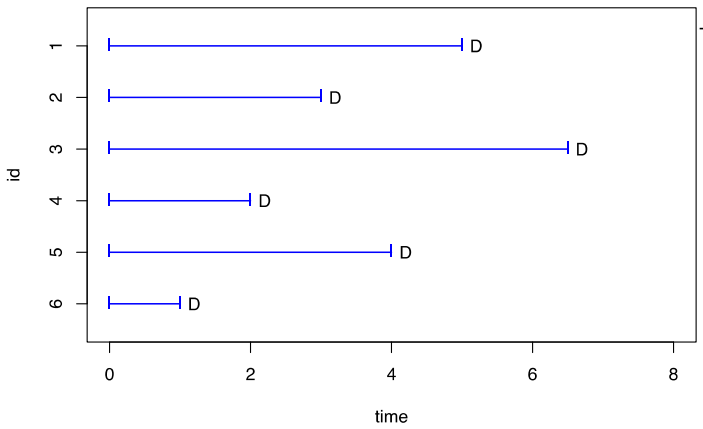


# MEDIAN SURVIVAL TIME





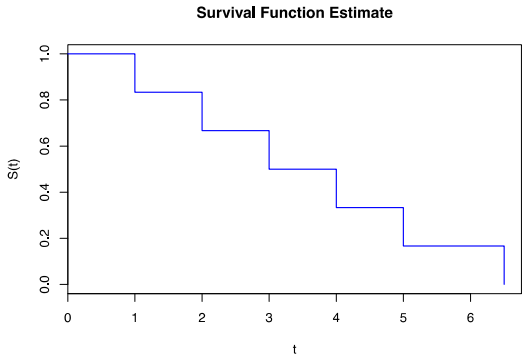

# ILLUSTRATIVE DATA




| id | Y |
|---|---|
| 1 | 5 |
| 2 | 3 |
| 3 | 6.5 |
| 4 | 2 |
| 5 | 4 |
| 6 | 1 |

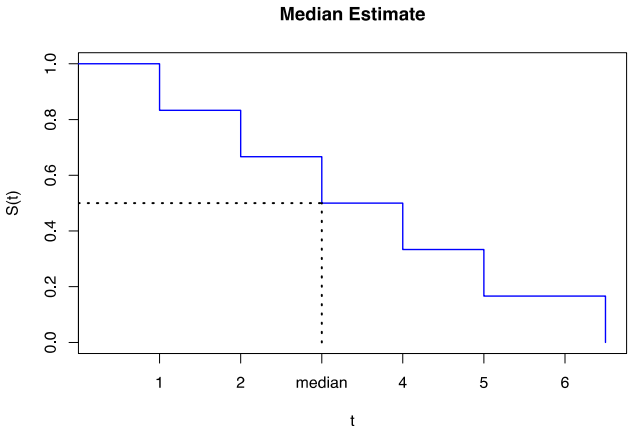
# SURVIVAL FUNCTION ESTIMATE

- Nonparametric Estimate: reduce estimate by 1/n every time there is an event (death): Empirical survival function estimate






# MEDIAN ESTIMATE




By convention: median is earliest time where survival estimate ≤ .5

## OTHER WAYS TO DESCRIBE A SURVIVAL DISTRIBUTION

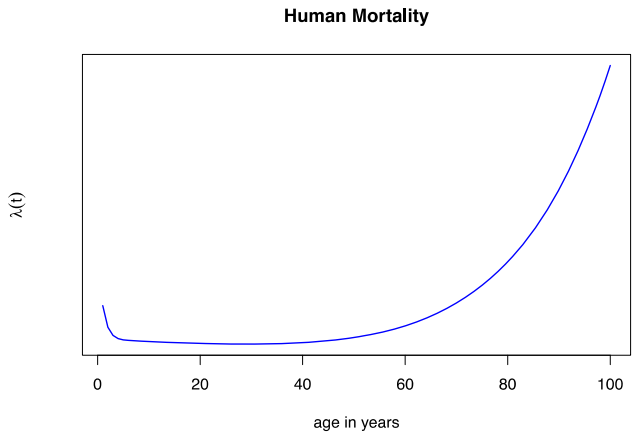
- So far we have looked at the density function and survival function S(t).
- Also of interest: "hazard" function λ(t)

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr[t \le T < t + \Delta t | T \ge t]$$

- Instantaneous rate at which death occurs at t in those who are alive at t
- Examples:
  - Age-specific death rate
  - Age-specific disease incidence rate



## HAZARD FUNCTION FOR HUMANS

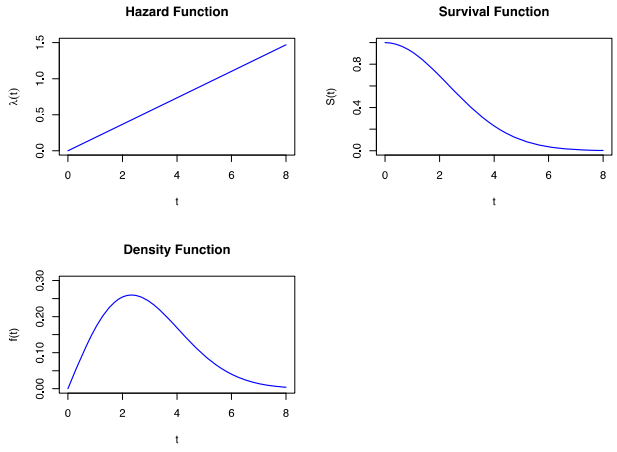


# EQUIVALENT CHARACTERIZATIONS

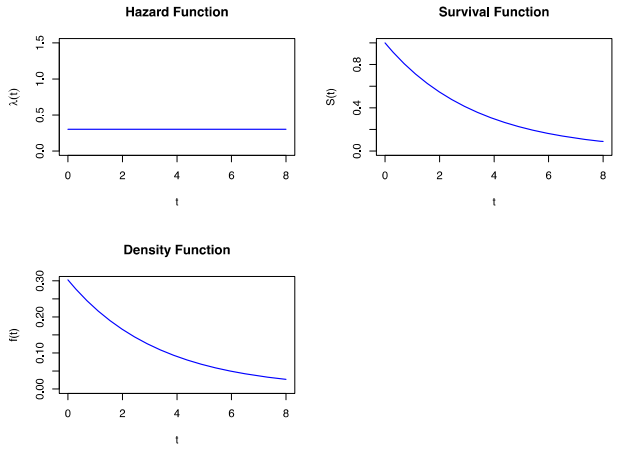
- Any <u>one</u> of the density function( f(t)), the survival function(S(t)) or the hazard function(λ(t)) is enough to determine the survival distribution.
- They are each functions of each other:
  - $S(t) = \int_t^\infty f(s)ds = e^{-\int_0^t \lambda(s)ds}$
  - $f(t) = -\frac{d}{dt}S(t) = \lambda(t)e^{-\int_0^t \lambda(s)ds}$
  - $\lambda(t) = \frac{f(t)}{S(t)}$



# EQUIVALENT CHARACTERIZATIONS



# EQUIVALENT CHARACTERIZATIONS






# QUESTIONS?

# SESSION 2:
# ONE-SAMPLE METHODS

Module 11: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

## OUTLINE

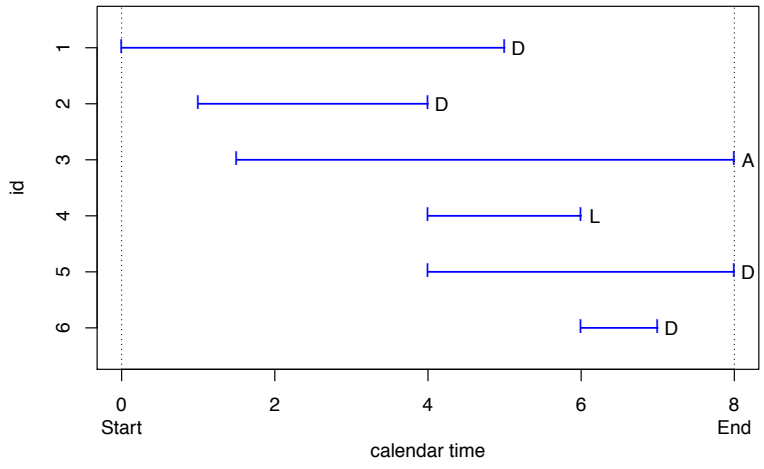
- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and Cis
  - Example

# OUTLINE

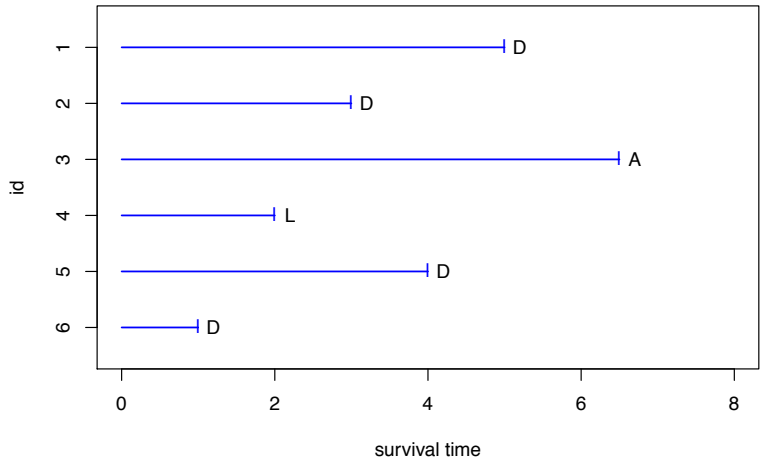
- Session 2:
  - **Censored data**
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and Cis
  - Example



# CLINICAL TRIAL

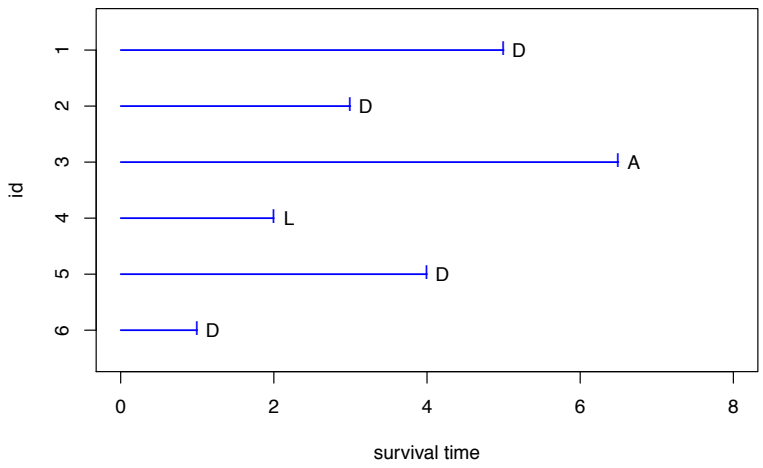


# CENSORED DATA





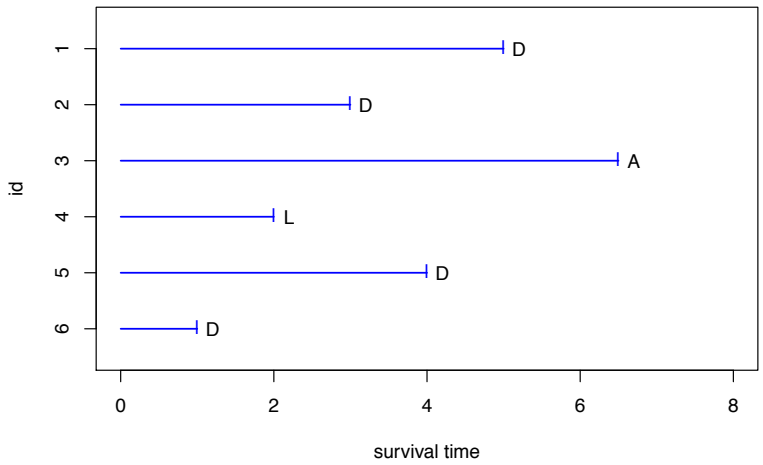

# CENSORED DATA




| id | Y | $\delta$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 3 | 1 |
| 3 | 6.5 | 0 |
| 4 | 2 | 0 |
| 5 | 4 | 1 |
| 6 | 1 | 1 |

"Censored" observations give some information about their survival time.

# CENSORED DATA




| id | Y | $\delta$ |
|---|---|---|
| 1 | 5 | 1 |
| 2 | 3 | 1 |
| 3 | 6.5 | 0 |
| 4 | 2 | 0 |
| 5 | 4 | 1 |
| 6 | 1 | 1 |

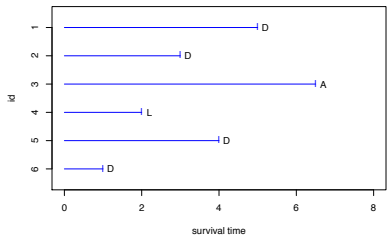
"Censored" observations give some information about their survival time.



# ESTIMATION

- Can we use the partial information in the censored observations?
- Two off-the-top-of-the-head answers:
  - Full sample: Yes. Count them as observations that did not experience the event ever and estimate S(t) as if there were not censored observations.
  - Reduced sample: No. Omit them from the sample and estimate S(t) from the reduced data as if they were the full data.

# CENSORED DATA




Problem: How to estimate:

| | $\Pr[T > 3.5]$ | $\Pr[T > 6]$ |
|---|---|---|
| Full Sample: | $\frac{4}{6} = .67$ | $\frac{2}{6} = .33$ |
| Reduced Sample: | $\frac{2}{4} = .5$ | $\frac{0}{4} = 0$ |



# CENSORED DATA

Based on the data and estimates on the previous page,

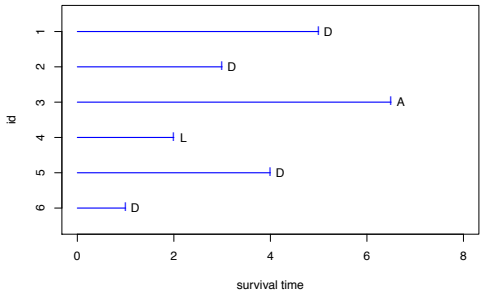
**Q:** Are the Full Sample estimates biased? Why or why not?

**A:**

**Q:** Are the Reduced Sample estimates biased? Why or why not?

**A:**

# CENSORED DATA




Problem: How to estimate:

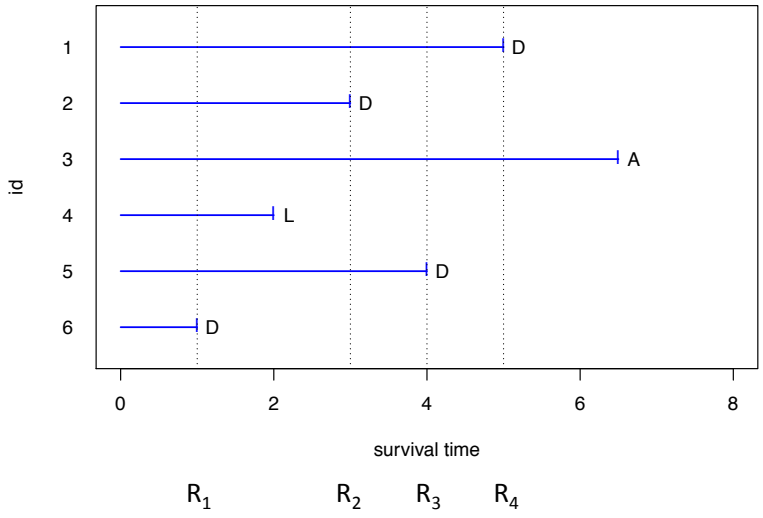
| | $\Pr[T > 3.5]$ | $\Pr[T > 6]$ | |
|---|---|---|---|
| Full Sample: | $\frac{4}{6} = .67$ | $\frac{2}{6} = .33$ | ← too high |
| Reduced Sample: | $\frac{2}{4} = .5$ | $\frac{0}{4} = 0$ | ← too low |

Need a good way to use the partial information in the censored observations.

IMPORTANT ASSUMPTION: Subjects who are censored at time t are representative of all subjects at risk of dying at time t.



# RISK SETS

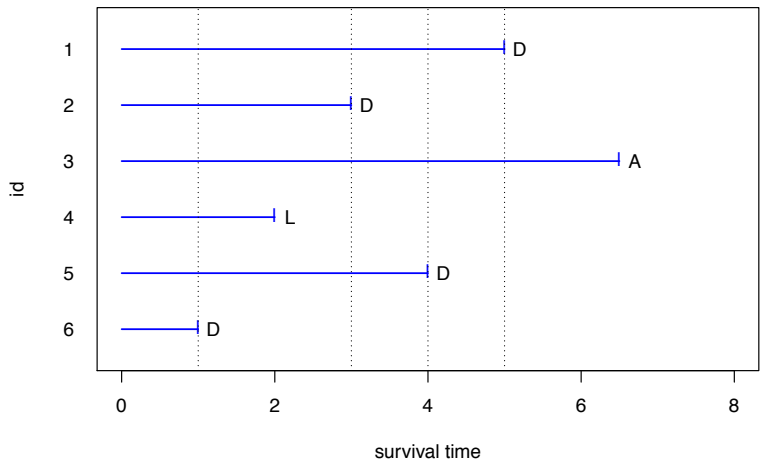


# OUTLINE

- Session 2:
  - Censored data
  - **Risk sets**
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and Cis
  - Example



# RISK SETS




$R_1$ {1,2,3,4,5,6}  $R_2$ {1,2,3,5}  $R_3$ {1,3,5}  $R_4$ {1,3}

# CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.
  - When would you expect this to be true (or false) for subjects lost to follow-up?

  - When would you expect this to be true (or false) still alive at the time of the analysis?



# OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - **Censoring assumptions**
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and Cis
  - Example

# CENSORED DATA ASSUMPTION

- Important assumption: subjects who are censored at time t are at the same risk of dying at t as those at risk but not censored at time t.
- This means the risk set at time t is an unbiased sample of the population still alive at time t.
- Can use information from the unbiased risk sets to estimate S(t) using the method of Kaplan and Meier (Product-Limit Estimator).

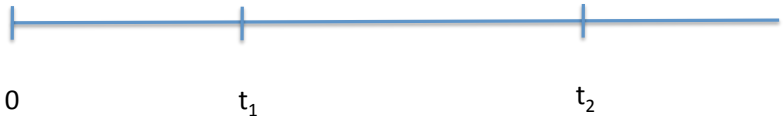


# USING RISK SETS INFO TO ESTIMATE S(t)

- Repeatedly use the fact that for $t_2 > t_1$,

$$Pr[T > t_2] = Pr[T > t_2 \text{ and } T > t_1] = Pr[T > t_2 \,|\, T > t_1]Pr[T > t_1]$$

- An observation censored between $t_1$ and $t_2$ can contribute to the estimation of $Pr[T > t_2]$ by its unbiased contribution to estimation of $Pr[T > t_1]$.



# OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - **Kaplan-Meier Estimator**
  - Median estimator
  - Standard errors and Cis
  - Example



# PRODUCT-LIMIT (KAPLAN-MEIER) ESTIMATE

Notation: Let $t_{(1)}, t_{(2)} \ldots, t_{(J)}$ be the ordered failure times in the sample in ascending order.

$t_{(1)}$ = smallest $Y_i$ for which $\delta_i = 1$ ($t_{(1)} = 1$ )
$t_{(2)}$ = $2^{nd}$ smallest $Y_i$ for which $\delta_i = 1$ ($t_{(2)} = 3$ )
⋮
$t_{(J)}$ = largest $Y_i$ for which $\delta_i = 1$ ($t_{(4)} = 5$ )

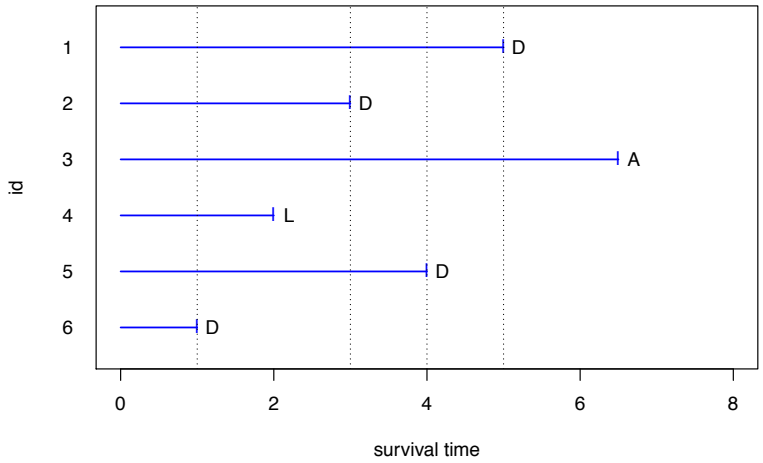
Q: Does J = the number of observed deaths in the sample?

A:

Q:When does J = n?

A:

# $t_{(j)}$






# MORE NOTATION

For each $t_{(j)}$:

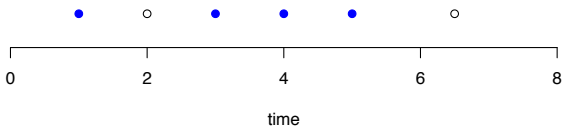
$D_{(j)}$ = number that die at time $t_{(j)}$

$S_{(j)}$ = number known to have survived beyond $t_{(j)}$
(by convention: includes those known to have been censored at $t_{(j)}$)

$N_{(j)}$ = number "at risk" of being observed to die at time $t_{(j)}$
(ie: number still alive and under observation just before $t_{(j)}$)

$S_{(j)} = N_{(j)} - D_{(j)}$

# FOR EXAMPLE DATA




| $t_{(j)}$ | $N_{(j)}$ | $D_{(j)}$ | $S_{(j)}$ |
|---|---|---|---|
| 1 | 6 | 1 | 5 |
| 3 | 4 | 1 | 3 |
| 4 | 3 | 1 | 2 |
| 5 | 2 | 1 | 1 |

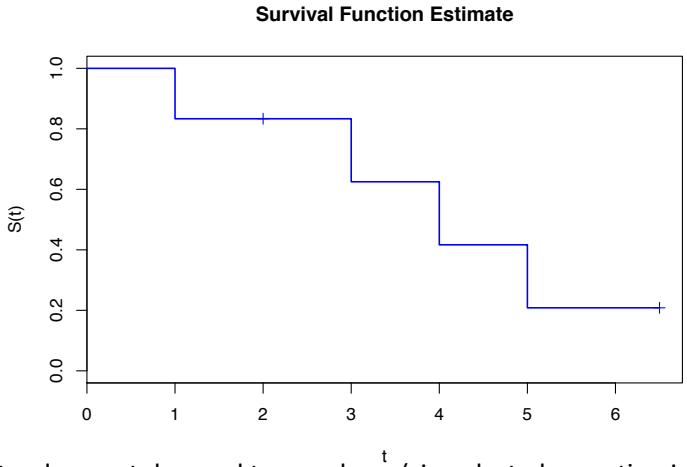
Product-limit (Kaplan-Meier) Estimator:

$$\hat{S}(t) = \Pi_{j:t_{(j)} \le t}(1 - \frac{D_{(j)}}{N_{(j)}}) = \Pi_{j:t_{(j)} \le t}(\frac{S_{(j)}}{N_{(j)}})$$

| for t in | $\hat{S}(t)$ | |
|---|---|---|
| $[0, 1)$ | 1 | (empty product) |
| $[1, 3)$ | $1 \times \frac{5}{6} = .833$ | |
| $[3, 4)$ | $1 \times \frac{5}{6} \times \frac{3}{4} = .625$ | |
| $[4, 5)$ | $1 \times \frac{5}{6} \times \frac{3}{4} \times \frac{2}{3} = .417$ | |
| $[5, \infty)$ | $1 \times \frac{5}{6} \times \frac{3}{4} \times \frac{2}{3} \times \frac{1}{2} = .208$ | |



# K-M ESTIMATOR




Note: does not descend to zero here (since last observation is censored).

Q: Since the estimate jumps only at observed death times, how does information from the censored observations contribute to it?

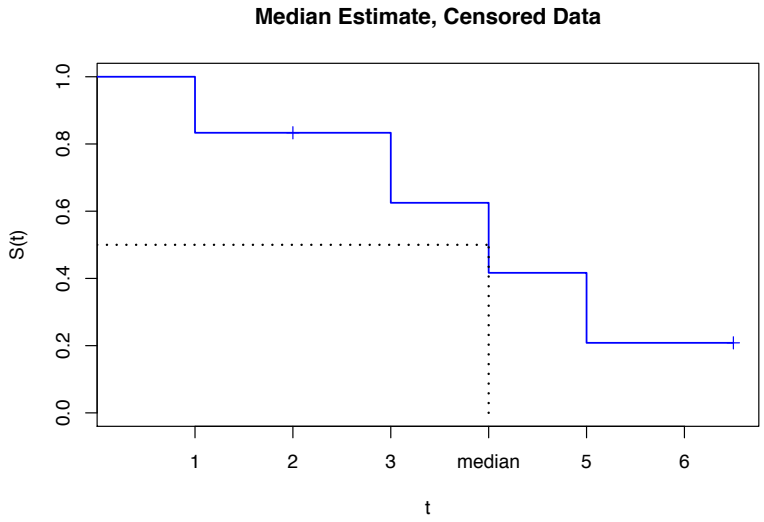
A:

# OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - **Median estimator**
  - Standard errors and Cis
  - Example



# MEDIAN SURVIVAL CENSORED DATA



# OUTLINE

- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - **Standard errors and CIs**
  - Example



# KM STANDARD ERRORS

Greenwood's Formula:

- $\widehat{Var}(\hat{S}(t)) = \hat{S}^2(t) \sum_{j:t_{(j)} \leq t} \frac{D_{(j)}}{N_{(j)} S_{(j)}}$
- $\text{se}(\hat{S}(t)) = \sqrt{\widehat{Var}(\hat{S}(t))}$
- Pointwise CI: $(\hat{S}(t) - z_{\frac{\alpha}{2}} \text{se}(\hat{S}(t)), \quad \hat{S}(t) + z_{\frac{\alpha}{2}} \text{se}(\hat{S}(t)))$
  - Can include values < 0 or > 1.

# LOG –LOG KM STANDARD ERRORS

Use complementary log log transformation to keep CI within (0,1):

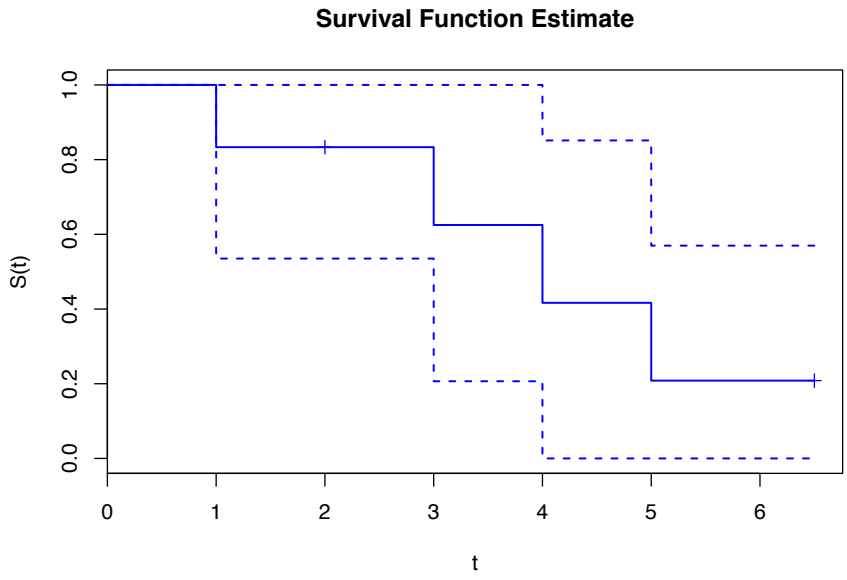
- $\widehat{Var}(\log(-\log(\hat{S}(t)))) = \frac{\sum_{j:t_{(j)} \le t} \frac{D_{(j)}}{N_{(j)} S_{(j)}}}{[\log(\hat{S}(t))]^2}$
- se = $\sqrt{\widehat{Var}(\log(-\log(\hat{S}(t))))}$
- CI for $\log(-\log(S(t)))$ :

  $$(\log(-\log(\hat{S}(t))) - z_{\frac{\alpha}{2}}\text{se}, \quad \log(-\log(\hat{S}(t))) + z_{\frac{\alpha}{2}}\text{se})$$

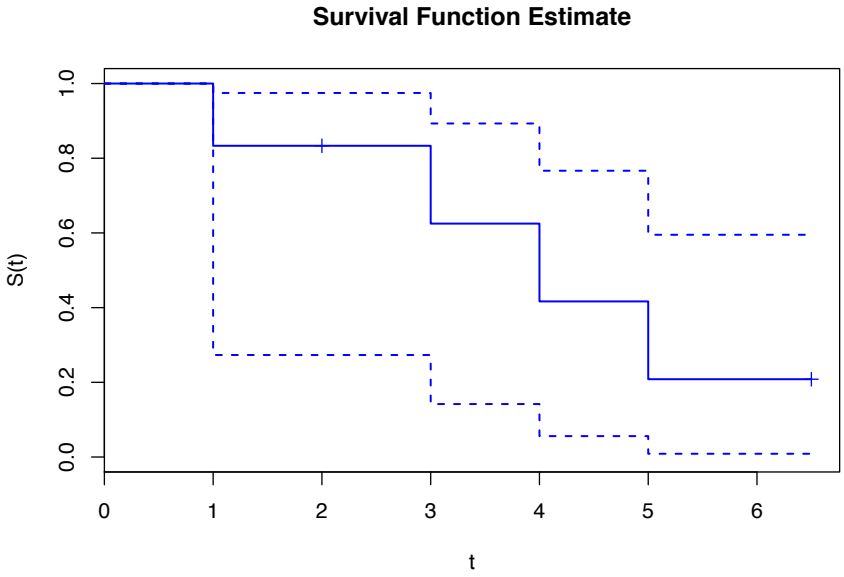
- CI for $\hat{S}(t)$ : $([\hat{S}(t)]^{e^{z\alpha/2}\text{se}}, [\hat{S}(t)]^{e^{-z\alpha/2}\text{se}})$
  - CI remains within (0,1).



# GREENWOOD'S FORMULA



# COMPLEMENTARY LOG-LOG






# MEDIAN CONFIDENCE INTERVAL

Confidence interval for the median is obtained by inverting the sign test of $H_0$ : median = $M$ (Brookmeyer and Crowley, 1982).

- With complete data $T_1, T_2, \ldots, T_n$, the sign test of $H_0$ : median $= M$ is performed by seeing if the observed proportion, $\hat{P}[Y > M]$ is too big or too small (Binomial Distribution or Normal Approximation).
- With censored data $(Y_1, \delta_1), (Y_2, \delta_2), \ldots, (Y_n, \delta_n)$ giving incomplete data about $T_1, T_2, \ldots, T_n$, we cannot always tell whether $T_i > M$:

| | | |
|---|---|---|
| When $Y_i \leq M, \delta_i = 1$ | observed death before $M$ | we know $T_i \leq M$ |
| When $Y_i > M$ | observed death after $M$ | we know $T_i > M$ |
| When $Y_i \leq M, \delta_i = 0$ | censored before $M$ | we don't know if $T_i \leq M$ or $T_i > M$ |

# MEDIAN CONFIDENCE INTERVAL

Solution: Following Efron (self-consistency of KM), we estimate $\Pr[T > M]$ when $Y_i \leq M, \delta_i = 0$ using $\frac{\hat{S}(M)}{\hat{S}(Y_i)}$.

- For complete data, we let $U_i = \begin{cases} 1 & T_i > M \\ 0 & T_i \leq M \end{cases}$

  and our test is based on $\sum_{i=1}^{n} U_i$.

- For censored data, we let $U_i = \begin{cases} 1 & Y_i > M \\ \frac{\hat{S}(M)}{\hat{S}(Y_i)} & Y_i \leq M; \delta_i = 0 \\ 0 & Y_i \leq M; \delta_i = 1 \end{cases}$

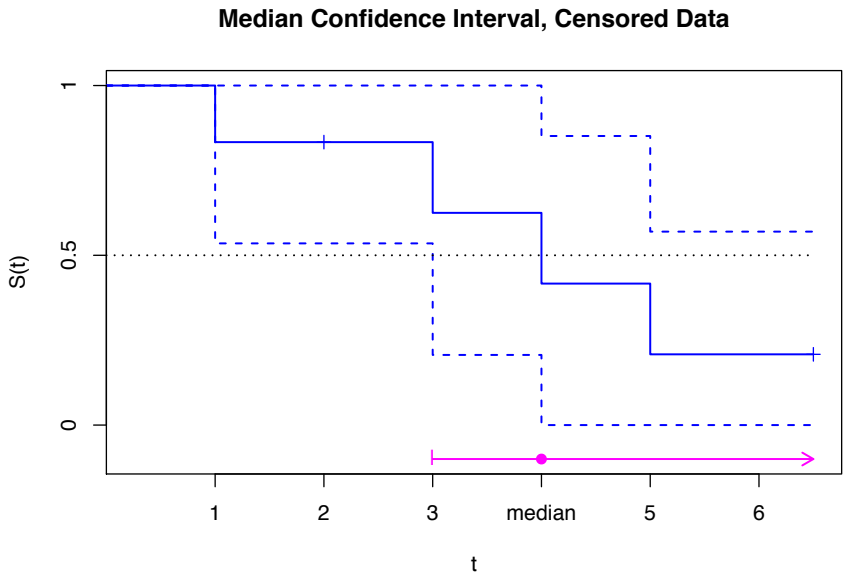
  and our test is based on $\sum_{i=1}^{n} U_i$.



# MEDIAN CONFIDENCE INTERVAL

- It turns out, this is the same as basing our test of $H_0$ : median $= M$ on a test of $H_0 : S(M) = \frac{1}{2}$.
- So a 95% CI for the median contains all potential $M$ for which the test of $H_0 : S(M) = \frac{1}{2}$ cannot reject at $\alpha = .05$ (2 sided).
- Since $\hat{S}(M)$ only changes value at observed event times, the test need only be checked at $M = t_{(1)}, t_{(2)}, \ldots, t_{(J)}$.
- Originally proposed for Greenwood's formula CIs for $\hat{S}(M)$, but any good CIs are OK.
- Implemented in many software packages.

# MEDIAN CONFIDENCE INTERVAL






# OUTLINE

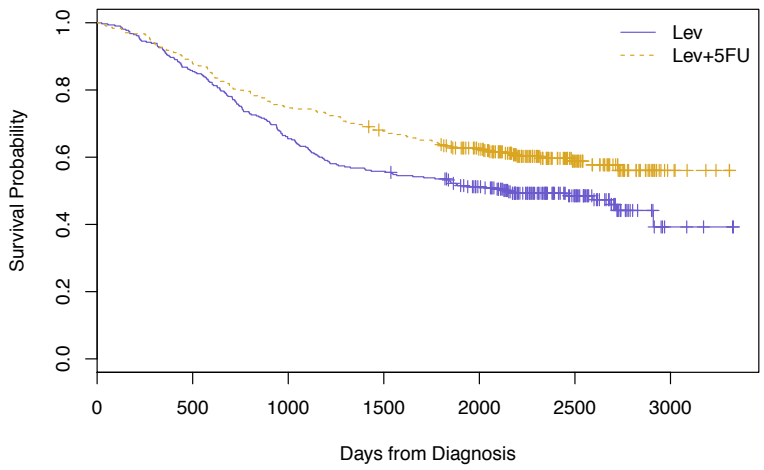
- Session 2:
  - Censored data
  - Risk sets
  - Censoring assumptions
  - Kaplan-Meier Estimator
  - Median estimator
  - Standard errors and Cis
  - **Example**

# COLON CANCER EXAMPLE

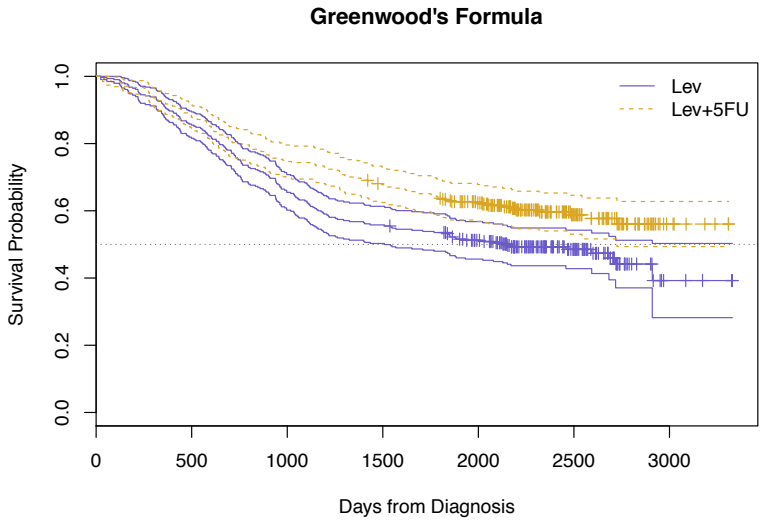
- Clinical trial at Mayo Clinic (Moertel et al. (1990) NEJM)
- Stage $B_2$ and C colon cancer patients; adjuvant therapy
- Three arms
  - Observation only
  - Levamisole
  - 5-FU + Levamisole
- Stage C patients only
- Two treatment arms only



# COLON CANCER EXAMPLE

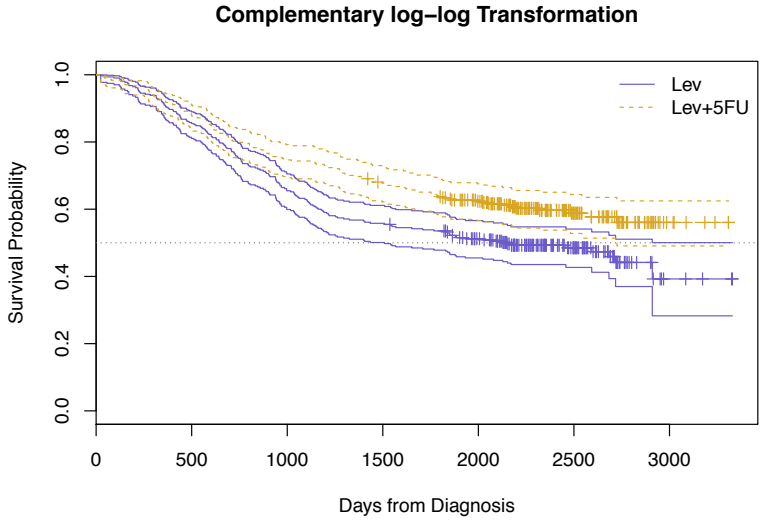


# COLON CANCER EXAMPLE






# COLON CANCER EXAMPLE

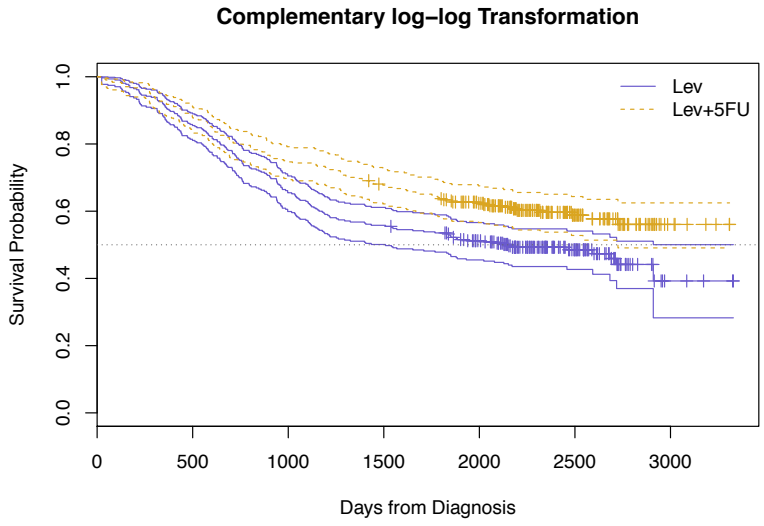


# PRESENTATION

| | N | Events | Median (days) | 95% CI |
|---|---|---|---|---|
| Levamisole Only | 310 | 161 | 2152 | (1509, ∞) |
| 5FU + Levamisole | 304 | 123 | -- | (2725, ∞) |



# COLON CANCER EXAMPLE



# ESTIMATION

- Estimate S(t) using KM curve (nonparametric).
  - Pointwise standard errors and Cis
  - Almost always presented
  - Not appropriate when the event of interest happens only to some (more on this Friday)
- Median: based on KM curve: often presented (too often?)



# TO WATCH OUT FOR

- Mean survival time hard to estimate without parametric assumptions
  - Censoring means incomplete information about largest times
  - Mean over restricted time interval may be useful in some settings (some on this tomorrow)
- Median estimate more complicated than median of times
- Even with CIs, evaluating differences between curves visually is subjective
- Interpretation of survival function estimates depends on validity of censoring assumptions

# SESSION 3:
# TWO AND K-SAMPLE METHODS

Module 11: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Susanne May, Ph.D.
Professor
Department of Biostatistics
University of Washington

## OVERVIEW

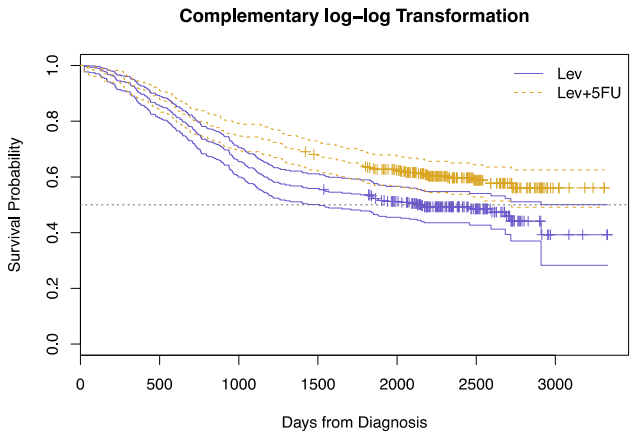
- Session 1
  - Introductory examples
  - The survival function
  - Survival Distributions
  - Mean and Median survival time
- Session 2
  - Censored data
  - Risk sets
  - Censoring Assumptions
  - Kaplan-Meier Estimator and CI
  - Median and CI
- Session 3
  - Two-group comparisons: logrank test
  - Trend and heterogeneity tests for more than two groups
- Session 4
  - Introduction to Cox regression

# TESTING

- Group comparisons
  - Two groups
  - k- group heterogeneity
  - k- group trend
- Assume, $H_0$ : no differences between groups



# COLON CANCER EXAMPLE



## THE P-VALUE QUESTION

- Statistical significance?

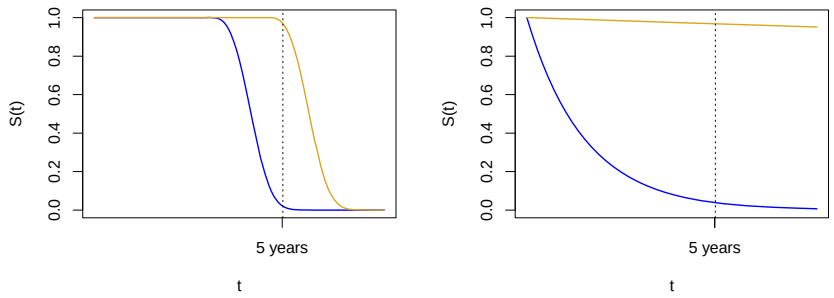


## COMPARING SURVIVAL DISTRIBUTIONS

- Two-sample data: comparing $S_1(t)$ and $S_2(t)$
  - $(Y_{1i}, \delta_{1i})$, $i=1,\ldots,n_1$, $T \sim S_1(t)$
  - $(Y_{2i}, \delta_{2i})$, $i=1,\ldots,n_2$, $T \sim S_2(t)$

- Could look at $S_2(t) - S_1(t)$ at a single time *t*, but this might be misleading unless all you care about is survival at that time.

# COMPARISON AT 5 YEARS

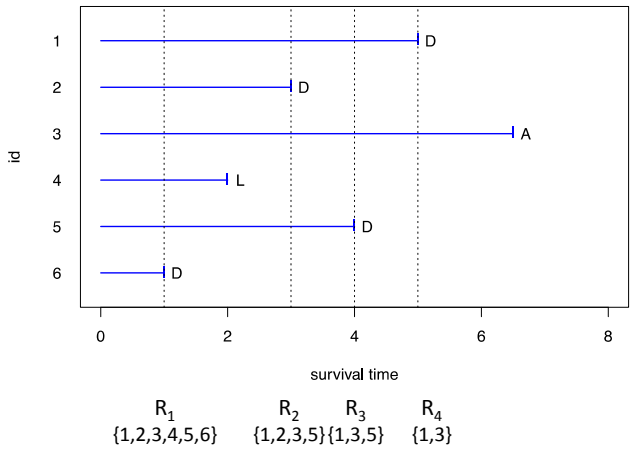





# COMPARING SURVIVAL DISTRIBUTIONS

- There are many ways to measure $S_2(t) - S_1(t)$, the distance between two functions of time
- Here: focus on most commonly used test: the logrank test, which compares consistent ratios of hazard functions
- Module 16 will consider other tests

# RISK SETS






# LOGRANK TEST

- The test is based on a 2x2 table of group by current status at each observed failure time (i.e. for each risk set)
- $T_{(j)}$, $j=1,\ldots m$, as shown in the Table below.

| Event/Group | 1 | 2 | Total |
|---|---|---|---|
| Die | $d_{1(j)}$ | $d_{2(j)}$ | $D_{(j)}$ |
| Survive | $n_{1(j)}-d_{1(j)}= s_{1(j)}$ | $n_{2(j)}-d_{2(j)} = s_{2(j)}$ | $N_{(j)}-D_{(j)} = S_{(j)}$ |
| At Risk | $n_{1(j)}$ | $n_{2(j)}$ | $N_{(j)}$ |

# TWO-GROUP COMPARISONS

- The contribution to the test statistic at each event time is obtained by calculating the expected number of deaths in one group, assuming that the risk of death at that time is the same in each of the two groups.
- This yields the usual *"row total times column total divided by grand total"* estimator. For example, for group 1, the expected number is

$$\hat{E}_{1(j)} = \frac{n_{1(j)} D_{(j)}}{N_{(j)}}$$

- Most software packages base their estimator of the variance on the hypergeometric distribution, defined as follows:

$$\hat{V}_{(j)} = \frac{n_{1(j)} n_{2(j)} D_{(j)} \left( N_{(j)} - D_{(j)} \right)}{N_{(j)}^2 \left( N_{(j)} - 1 \right)}$$



# LOGRANK TWO-GROUP COMPARISONS

- Each test may be expressed in the form of a ratio of sums over the observed survival times as follows

$$Q = \frac{\left[ \sum_{j=1}^{J} \left( d_{1(j)} - \hat{E}_{1(j)} \right) \right]^2}{\sum_{j=1}^{J} \hat{V}_{(j)}} = \frac{\left[ \sum_{j=1}^{J} \left( \frac{n_{1(j)} n_{2(j)}}{n_{1(j)} + n_{2(j)}} \right) \left( \frac{d_{1(j)}}{n_{1(j)}} - \frac{d_{2(j)}}{n_{2(j)}} \right) \right]^2}{\sum_{j=1}^{J} \hat{V}_{(j)}}$$

- Where $t_j$, $j = 1,\ldots, J$, are the unique ordered event times
- Under the null hypothesis of no difference in survival distribution, the *p*-value for *Q* may be obtained using the chi-square distribution with one degree-of-freedom, when the expected number of events is large.

$$p = \Pr\left( \chi^2_1 \geq Q \right)$$

# COLON CANCER EXAMPLE

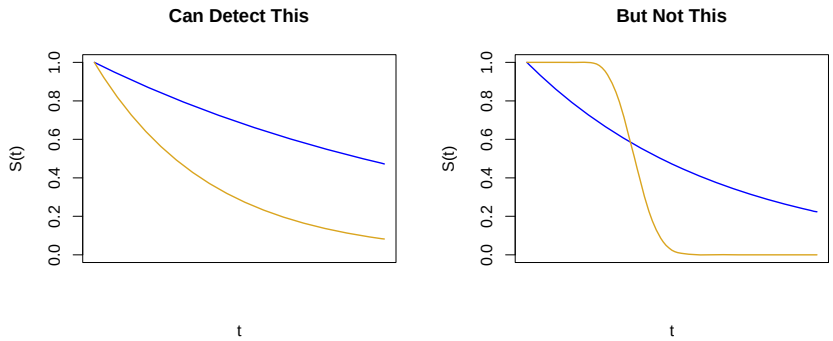
- Comparing Lev and Lev+5FU:

| Group | N | Obs | Exp |
|---|---|---|---|
| Lev | 310 | 161 | 136.9 |
| Lev+5FU | 304 | 123 | 147.1 |
| Total | 614 | 284 | 284.0 |

- Log-rank test: $\chi^2_1 = 8.2$, p-value = 0.0042



# LOGRANK TEST




Other tests (generalized Wilcoxon and others) can give more weight to early or late differences.

## LOGRANK TEST

- Detects consistent differences between survival curves over time.
- Best power when:
  - $H_0$: $S_1(t) = S_2(t)$ for all t vs $H_A$: $S_1(t) = [S_2(t)]^c$ , or
  - $H_0$: $\lambda_1(t) = \lambda_2(t)$ for all t vs $H_A$: $\lambda_1(t) = c\,\lambda_2(t)$
- Good power whenever survival curve difference is in consistent direction



## STRATIFIED LOGRANK TEST

- In a large-enough clinical trial, confounding bias due to imbalance between treatment arms is unlikely.
- However, better power can be obtained by adjusting for strongly prognostic variables.
- One way to adjust: stratified logrank test
- Can also use Cox regression (Module 17)

## STRATIFIED LOGRANK TEST

- Assume $R$ strata ($r = 1,\ldots,R$)
- Recall (non-stratified) log-rank test statistic

$$Q = \frac{\left[\sum_{j=1}^{J}\left(d_{1(j)} - \hat{E}_{1(j)}\right)\right]^2}{\sum_{j=1}^{J}\hat{V}_{(j)}}$$

- Stratified log-rank test

$$Q = \frac{\left[\sum_{j_1=1}^{J_1}\left(d_{1,1(j)} - \hat{E}_{1,1(j)}\right) + \ldots + \sum_{j_r=1}^{J_r}\left(d_{1r(j)} - \hat{E}_{1r(j)}\right) + \ldots + \sum_{j_R=1}^{J_R}\left(d_{1R(j)} - \hat{E}_{1R(j)}\right)\right]^2}{\sum_{j_1=1}^{J_1}\hat{V}_{1(j)} + \ldots + \sum_{j_r=1}^{J_r}\hat{V}_{r(j)} + \ldots + \sum_{j_R=1}^{J_R}\hat{V}_{R(j)}}$$



## STRATIFIED LOG-RANK TEST

- $H_0$: $\lambda_{1r}(t) = \lambda_{2r}(t)$ for all $t$ and r = 1,…,$R$
- $H_A$: $\lambda_{1r}(t) = c\lambda_{2r}(t)$, $c \neq 1$, for all $t$ and $r = 1,\ldots,R$

- Under $H_0$ test statistic ~ $\chi^2_1$ when the number of events is large
- The $d_{1r(j)}, \hat{E}_{1r(j)}$ and $\hat{V}_{r(j)}$ are based solely on subjects from the $r^{th}$ stratum
- Will be powerful when direction of group difference is consistent across strata and over time.

## EXAMPLE - WHAS

- Example: The Worcester Heart Attack Study (WHAS)
- Goal: study factors and time trends
  associated with long term survival
  following acute myocardial infarction (MI)
  among residents of the Worcester, Massachusetts
  Standard Metropolitan Statistical Area (SMSA)
- Study began in 1975
- Data collection approximately every other year
- Most recent cohort: subjects who experienced an MI in 2001
- The main study: over 11,000 subjects
- Here: a small sample from the main study with n = 100



## EXAMPLE - WHAS

- $t_0$: time of hospital admission following an acute myocardial infarction (MI)
- Event: Death from any cause following hospitalization for an MI
- Time: Time from hospital admission to
  - Death
  - End of study
  - Last contact
- Interest in effect of gender adjusted for age

## GENDER BY AGE GROUPS

| Age | Male | Female | Total |
|---|---|---|---|
| 32-59 | 20 | 5 | 25 |
| 60-69 | 17 | 6 | 23 |
| 70-79 | 15 | 7 | 22 |
| 80-92 | 13 | 17 | 30 |
| Total | 65 | 35 | 100 |



## TESTING GENDER BY AGE

- Log rank test for age group 32-59

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 20 | 5 | 6.53 | 0.357 | 1.95 |
| Female | 5 | 3 | 1.47 | 1.584 | 1.95 |

Chisq = 1.9, 1 df, p=0.163

## TESTING GENDER BY AGE

- Log rank test for age group 60-69

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 17 | 4 | 5.6 | 0.458 | 2.41 |
| Female | 6 | 3 | 1.4 | 1.833 | 2.41 |

Chisq = 2.4, 1 df, p=0.121



## TESTING GENDER BY AGE

- Log rank test for age group 70-79

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 15 | 10 | 9.07 | 0.0947 | 0.273 |
| Female | 7 | 4 | 4.93 | 0.1743 | 0.273 |

Chisq = 0.3, 1 df, p=0.602

## TESTING GENDER BY AGE

- Log rank test for age group 80-92

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 13 | 9 | 8.83 | 0.0032 | 0.0057 |
| Female | 17 | 13 | 13.17 | 0.0021 | 0.0057 |

Chisq = 0, 1 df, p=0.94



## STRATIFIED TEST

- Log rank test stratified by age

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 65 | 28 | 30 | 0.138 | 0.402 |
| Female | 35 | 23 | 21 | 0.197 | 0.402 |

Chisq = 0.4, 1 df, p=0.53

# UN-STRATIFIED TEST

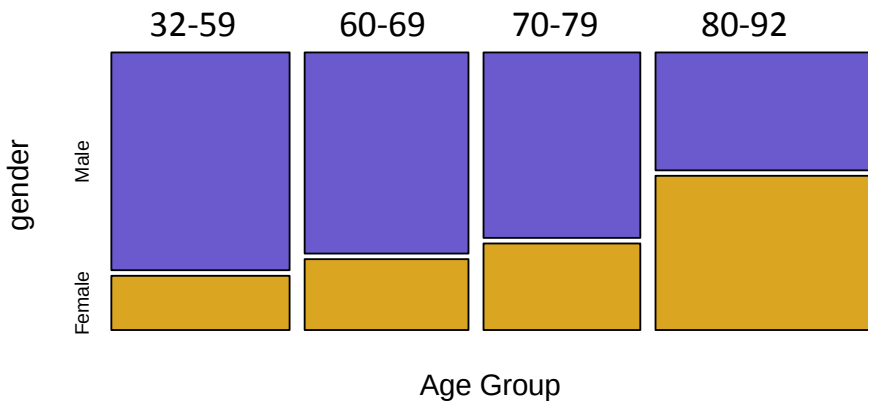
- Log rank test (not stratified by age)

| | N | Obs | Exp | $(O-E)^2/E$ | $(O-E)^2/V$ |
|---|---|---|---|---|---|
| Male | 65 | 28 | 34.7 | 1.29 | 4.06 |
| Female | 35 | 23 | 16.3 | 2.74 | 4.06 |

Chisq = 4.1, 1 df, p=0.044



# WHY?



## HETEROGENEITY

- When there are more than two groups, can test for difference somewhere between groups:
- Null hypothesis: $\lambda_1(t) \equiv \lambda_2(t) \equiv ... \equiv \lambda_k(t)$
- Alternative hypothesis: $\neq$ somewhere



## COLON DATA: THREE TREATMENT GROUPS

| | Observed Events | Expected Events |
|---|---|---|
| Obs | 161 | 146.1 |
| Lev | 123 | 157.5 |
| Lev+5FU | 168 | 148.4 |
| | 452 | 452 |

- $\chi^2_2 = 11.7$ (df = one fewer than number of groups)
- P-value: 0.003

## TREND

- When there are more than two "ordered" groups, it is sometimes of interest to test the null hypothesis of no difference against a "trend" alternative
- $\lambda_1(t) \le \lambda_2(t) \le ... \le \lambda_k(t)$ with < somewhere, or
- $\lambda_1(t) \ge \lambda_2(t) \ge ... \ge \lambda_k(t)$ with > somewhere
- Placebo and two or more doses of a therapeutic agent
- Pre-hypothesized



## TREND

- The test statistic for trend uses "scores": $s_1, s_2, \ldots, s_k$

$$\frac{\left(\sum_{i=1}^{k} s_i \sum_{j=1}^{J_k} (d_{ij} - E_{ij})\right)^2}{s'Vs}$$

- Null hypothesis: $\lambda_1(t) \equiv \lambda_2(t) \equiv ... \equiv \lambda_k(t)$
- Specific alternative hypothesis:

$$c^{s_1}\lambda_1(t) \equiv c^{s_2}\lambda_2(t) \equiv ... \equiv c^{s_k}\lambda_k(t), c \neq 1$$

- Good power when average difference between observed and expected events grows or diminishes with increasing $s_i$

# TREND

Tumor differentiation and all-cause mortality:

| Group | N | Observed | Expected |
|---|---|---|---|
| Well Differentiated | 93 | 42 | 47.5 |
| Moderately Differentiated | 663 | 311 | 334.9 |
| Poorly Differentiated | 150 | 88 | 58.6 |

Tarone trend test: $\chi_1^2 = 11.57$, $P = 6.6 \times 10^{-4}$



# SUMMARY

- Can use logrank test to detect consistent differences (over time) in the hazard of dying (the event occurring) using censored survival data
  - Can stratify on prognostic variables
- Can test for differences between more than two groups
- When alternative is ordered by prior hypothesis, can test for trend rather than heterogeneity

## TO WATCH OUT FOR:

- Only ranks are used for “standard” tests
- Observations with time = 0
- Crossing hazard functions
- P-value not valid if you decide between trend and heterogeneity test after looking at the data
  - Data told you what your hypothesis was

# SESSION 4:
# INTRODUCTION TO COX REGRESSION

Module 11: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
July, 2018

Barbara McKnight, Ph.D.
Professor
Department of Biostatistics
University of Washington

# OVERVIEW

- Session 1
  - Introductory examples
  - The survival function
  - Survival Distributions
  - Mean and Median survival time
- Session 2
  - Censored data
  - Risk sets
  - Censoring Assumptions
  - Kaplan-Meier Estimator and CI
  - Median and CI
- Session 3
  - Two-group comparisons: logrank test
  - Trend and heterogeneity tests for more than two groups
- Session 4
  - Introduction to Cox regression

# OUTLINE

- Motivation:
  - Confounding in observational studies
  - Stratified randomization designs
- Cox Regression model
  - Coefficient interpretation
  - Estimation and testing
  - Relationship to 2- and K-sample tests
  - Examining non-proportionality
- Examples throughout



# OUTLINE

- **Motivation:**
  - **Confounding in observational studies**
  - **Stratified randomization designs**
- Cox Regression model
  - Coefficient interpretation
  - Estimation and testing
  - Relationship to 2- and K-sample tests
  - Examining non-proportionality
- Examples throughout

# CONFOUNDING

- Observational data: sometimes observed associations between an explanatory variable and outcome can be due to their joint association with another variable.
  - Age related to both sex and risk of death.
  - Other examples?



# PRECISION IN RCTS

- Because of randomization, confounding/imbalance usually not an issue except in small trials.
- As in linear regression, regression models for censored survival data allow group comparisons among subjects with similar values of adjustment or "precision" variables (more later).
- Fairer and possibly more powerful comparison as long as adjustment variables are not the result of treatment.

# STRATIFIED RANDOMIZATION

- For strong predictors: concern about possible randomization imbalance
  - Clinic or center
  - Stage of disease
  - Sex
  - Age
- Adjust for stratification variables in analysis
  - More powerful if predictors are strong
  - Same conditioning as the sampling



# OUTLINE

- Motivation:
  - Confounding in observational studies
  - Stratified randomization designs
- **Cox Regression model**
  - **Coefficient interpretation**
  - Estimation and testing
  - Relationship to 2- and K-sample tests
  - Examining non-proportionality
- Examples throughout

# COX REGRESSION MODEL

- Usually written in terms of the hazard function
- As a function of independent variables $x_1, x_2, \ldots x_k$,

$$\lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

↑ relative risk / hazard ratio

$$\log \lambda(t) = \log \lambda_0(t) + \beta_1 x_1 + \cdots + \beta_k x_k$$

↑ intercept



# RELATIVE RISK / HAZARD RATIO

$$\lambda(t|x_1, \ldots, x_k) = \lambda_0(t)e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

$$\frac{\lambda(t|x_1,\ldots,x_k)}{\lambda(t|0,\ldots,0)} = e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

# REGRESSION MODELS

LS Linear Regression: $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$

| | | |
|---|---|---|
| Linear: | $Y \sim N(\mu, \sigma^2)$ | $\mu = EY = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ |
| Cox: | $T \sim S(t)$ | $\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \cdots + \beta_k x_k}$ |
| | ↑ | ↑ |
| | Distribution of outcome variable | Dependence of distribution on $x_1, \ldots x_k$ |



# PROPORTIONAL HAZARDS MODEL

$$\lambda(t) = \lambda_0(t) e^{\beta_1 x_1 + \cdots + \beta_k x_k}$$

Interpretation of $e^{\beta_1}$ in general:

"Relative risk (or hazard ratio) associated with a one unit higher value of $x_1$, holding $x_2, \ldots, x_k$ constant".

| | |
|---|---|
| $\lambda(t)$ for $x_1 + 1$: | $\lambda_0(t) e^{\beta_1 (x_1 + 1) + \cdots + \beta_k x_k}$ |
| $\lambda(t)$ for $x_1$: | $\lambda_0(t) e^{\beta_1 x_1 + \cdots + \beta_k x_k}$ |
| ratio: | $e^{\beta_1 (x_1 + 1 - x_1)} = e^{\beta_1}$ |

# EXAMPLE

Single binary $x$:

$$x = \begin{cases} 1 & \text{Test treatment} \\ 0 & \text{Standard treatment} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta x}$$

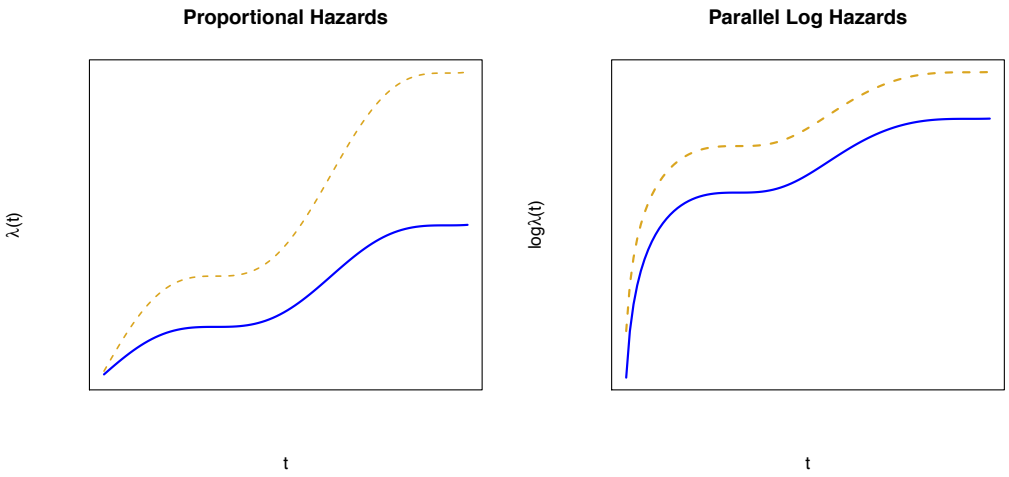
Interpretation of $e^{\beta}$:

"Relative risk (or hazard ratio) comparing test treatment to standard".

$$\begin{aligned} \lambda(t) \text{ for } x = 1: \quad & \lambda_0(t)e^{\beta \cdot 1} = \lambda_0(t)e^{\beta} \\ \lambda(t) \text{ for } x = 0: \quad & \lambda_0(t)e^{\beta \cdot 0} = \lambda_0(t) \\ \text{ratio:} \quad & e^{\beta(1-0)} = e^{\beta} \end{aligned}$$



# EXAMPLE



# RELATIONSHIP TO SURVIVAL FUNCTION

Single binary $x$:

$$x = \begin{cases} 1 & \text{Test treatment} \\ 0 & \text{Standard treatment} \end{cases}$$

$$\lambda(t) = \lambda_0(t)e^{\beta x} \implies S(t) = [S_0(t)]^{e^{\beta x}}$$

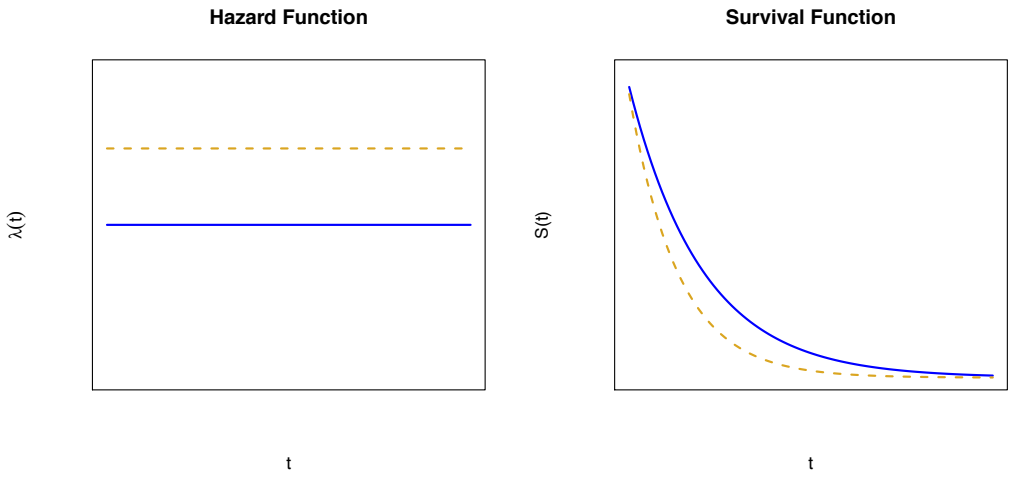
In terms of $S_0(t)$:

$$S(t) \text{ for } x = 1: \quad [S_0(t)]^{e^{\beta \cdot 1}} = [S_0(t)]^{e^{\beta}}$$

$$S(t) \text{ for } x = 0: \quad [S_0(t)]^{e^{\beta \cdot 0}} = [S_0(t)]^{1} = S_0(t)$$



# PICTURE



# OUTLINE

- Motivation:
  - Confounding in observational studies
  - Stratified randomization designs
- **Cox Regression model**
  - Coefficient interpretation
  - **Estimation and testing**
  - Relationship to 2- and K-sample tests
  - Examining non-proportionality
- Examples throughout



# ESTIMATES AND CONFIDENCE INTERVALS

- We estimate $\beta$ by maximizing the "partial likelihood function"
- Requires iteration on computer
- $\hat{\beta}$ is a MPLE (Maximum Partial Likelihood Estimator)
- We do not need to estimate $\lambda_0(t)$ to do this

- Most packages will estimate se($\hat{\beta}$) using the information matrix from this PL.
- 95% CI for $\beta$: $(\hat{\beta} - 1.96\text{se}(\hat{\beta}),\ \hat{\beta} + 1.96\text{se}(\hat{\beta}))$
- 95% CI for RR = $e^{\beta}$ : $(e^{\hat{\beta}-1.96\text{se}(\hat{\beta})}, e^{\hat{\beta}+1.96\text{se}(\hat{\beta})})$

# PARTIAL LIKELIHOOD

Data for the $i^{th}$ subject: $(t_i, \delta_i, x_{1i}, \ldots x_{ki})$

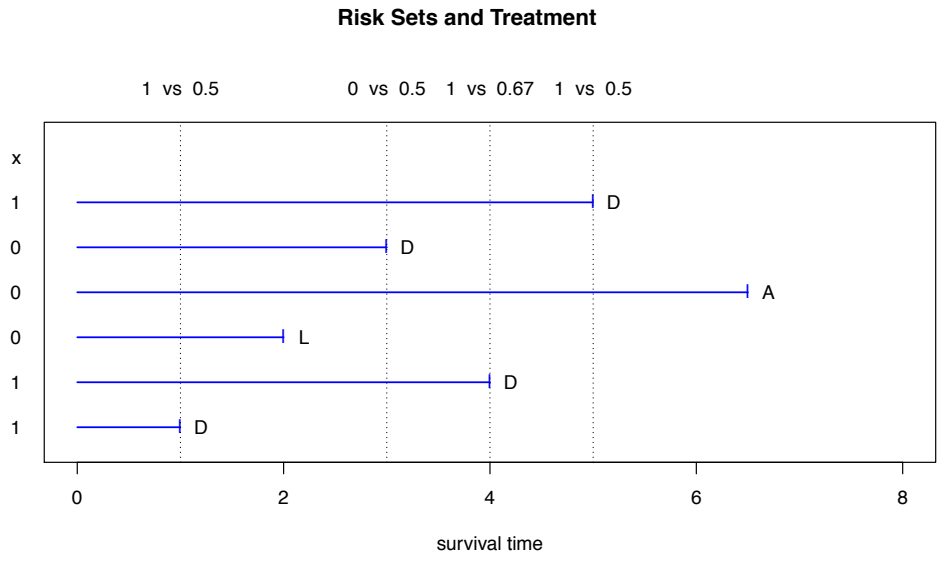
For subject with the $j^{th}$ ordered failure time : $(t_{(j)}, 1, x_{1(j)}, \ldots, x_{k(j)})$

$$\text{PL}(\beta_1, \ldots, \beta_k) = \prod_{j=1}^{J} \frac{e^{\beta_1 x_{1(j)} + \cdots + \beta_k x_{k(j)}}}{\sum_{i:t_i \geq t_{(j)}} e^{\beta_1 x_{1i} + \cdots + \beta_k x_{ki}}}$$

- $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$ are the values of $(\beta_1, \ldots, \beta_k)$ that maximize $\text{PL}(\beta_1, \ldots, \beta_k)$. (MPLEs)
- Compares $x$ values for the subject who failed at time $t_{(j)}$ to those of all subjects at risk at time $t_{(j)}$.
- Does not depend on the values of the $t_i$, only on their order.
- Does not depend on $\lambda_0(t)$.



# RISK SET PICTURE



# FULL LIKELIHOOD

$$
\begin{aligned}
L(\beta, \lambda_0(t)) &= \prod_{\text{Failures}} \Pr[T = t_i] \prod_{\text{Censorings}} \Pr[T > t_i] \\
&= \prod_{\text{Failures}} \lambda(t_i|x_i) S(t_i|x_i) \prod_{\text{Censorings}} S(t_i|x_i) \\
&= \prod_{i=1}^{n} [\lambda(t_i|x_i)]^{\delta_i} S(t_i|x_i) \\
&= \prod_{i=1}^{n} [\lambda_0(t_i) e^{\beta x_i}]^{\delta_i} e^{-\int_0^{t_i} \lambda_0(s) e^{\beta x} ds}
\end{aligned}
$$



# PARTIAL LIKELIHOOD

Let $H_t$ represent the entire history of failure, censoring and $x$ in the sample before time $t$.

Then the likelihood can be rewritten as follows:

$$
\begin{aligned}
L(\beta, \lambda_0(t)) &= \prod_{j=1}^{J} \Pr[i^{th} \text{ subject fails at } t_{(j)} | H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \cdot \\
&\qquad \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \\
&= \prod_{j=1}^{J} \frac{\lambda(t_{(j)}|x_{(j)})}{\sum_{i:t_i \geq t_{(j)}} \lambda(t_{(j)}|x_i)} \cdot \prod_{j=1}^{J} \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \\
&= \prod_{j=1}^{J} \frac{\lambda_0(t_{(j)}) e^{\beta x_{(j)}}}{\sum_{i:t_i \geq t_{(j)}} \lambda_0(t_{(j)}) e^{\beta x_i}} \cdot \prod_{j=1}^{J} \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \\
&= \prod_{j=1}^{J} \frac{e^{\beta x_{(j)}}}{\sum_{i:t_i \geq t_{(j)}} e^{\beta x_i}} \cdot \prod_{j=1}^{J} \Pr[H_{t_{(j)}}, \text{ some subject fails at } t_{(j)}] \\
&= \underbrace{\qquad\qquad}_{\substack{\text{Partial Likelihood} \\ \text{Depends only on } \beta}} \quad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{Depends on } \lambda_0(\cdot) \text{ and } \beta}
\end{aligned}
$$

# OUTLINE

- Motivation:
  - Confounding in observational studies
  - Stratified randomization designs
- **Cox Regression model**
  - Coefficient interpretation
  - **Estimation and testing**
  - **Relationship to 2- and K-sample tests**
  - Examining non-proportionality
- Examples throughout



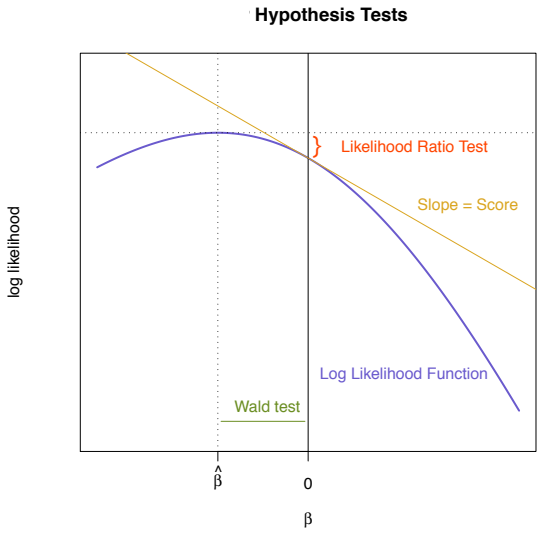
# HYPOTHESIS TESTS

Three tests of $H_0 : \beta = 0$ are possible:

1. Wald test: $\frac{\hat{\beta}}{\text{se}(\hat{\beta})}$
2. (Partial) Likelihood ratio test
3. Score test: ($\approx$ logrank test)

Likelihood ratio test is best, but requires fitting full ($\beta = \hat{\beta}$) and reduced ($\beta = 0$) models.

# LIKELIHOODS AND TESTS

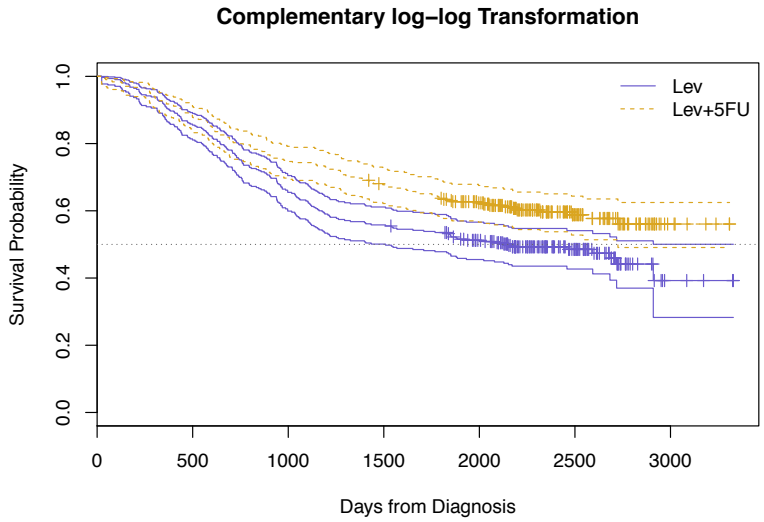





# COLON CANCER EXAMPLE

- Clinical trial at Mayo Clinic
- Stage $B_2$ and C colon cancer patients; adjuvant therapy
- Three arms
  - Observation only
  - Levamisole (stage C only)
  - 5-FU + Levamisole at Mayo Clinic
- Stage C patients only
- Two treatment arms only

# COLON CANCER EXAMPLE






# COLON CANCER EXAMPLE

| Variable | n | Deaths | Hazard ratio | CI | P-value |
|---|---|---|---|---|---|
| Levamisole Only | 310 | 161 | 1.0 (reference) | -- | -- |
| Levamisole + 5FU | 304 | 123 | 0.71 | (0.56, 0.90) | .004 |

Q: Which group has better survival?

A:

# TEST COMPARISON

| Test | Statistic | P-value |
|---|---|---|
| Wald's | 8.13 | .004 |
| Score | 8.21 | .004 |
| Likelihood Ratio | 8.21 | .004 |

Two-sided tests



# ANOTHER EXAMPLE

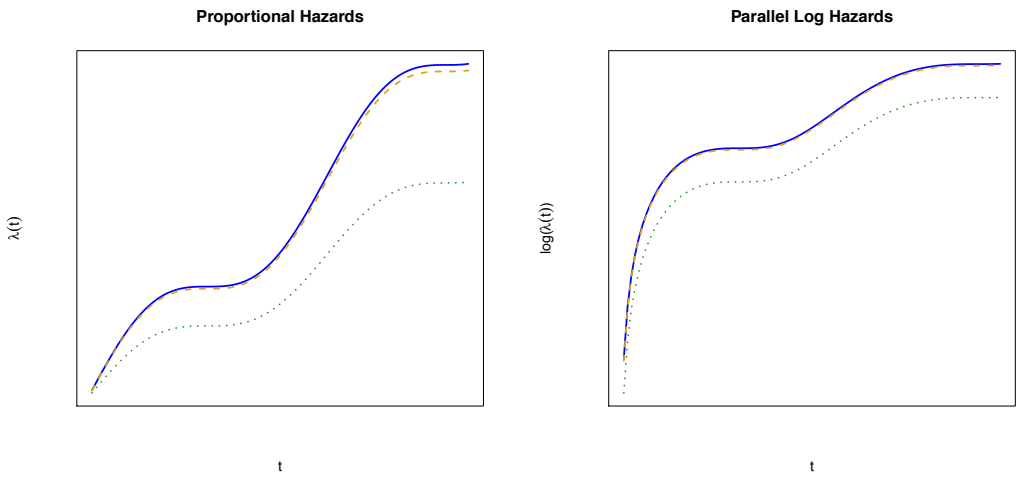
Three groups: use indicators for two

$$x_1 = \begin{cases} 1 & \text{Levamisole Only} \\ 0 & \text{otherwise} \end{cases} \qquad x_2 = \begin{cases} 1 & \text{Levamisole + 5FU} \\ 0 & \text{otherwise} \end{cases}$$

Model: $\lambda(t) = \lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$

| | | | | |
|---|---|---|---|---|
| RRs: | Levamisole Only | vs. | Observation | $e^{\beta_1}$ |
| | Levamisole + 5FU | vs. | Observation | $e^{\beta_2}$ |
| | Levamisole + 5FU | vs. | Levamisole Only | $e^{\beta_2 - \beta_1}$ |

# HEURISTIC HAZARDS






# COLON CANCER

| Variable | n | Deaths | Hazard Ratio | 95% CI | P-value |
|---|---|---|---|---|---|
| Observation Only | 315 | 168 | 1.0 (reference) | -- | -- |
| Levamisole Only | 310 | 161 | 0.97 | (0.78, 1.21) | 0.81 |
| Levamisole + 5FU | 204 | 123 | 0.69 | (0.55, 0.87) | 0.002 |

Q: Which group has best survival?

A:

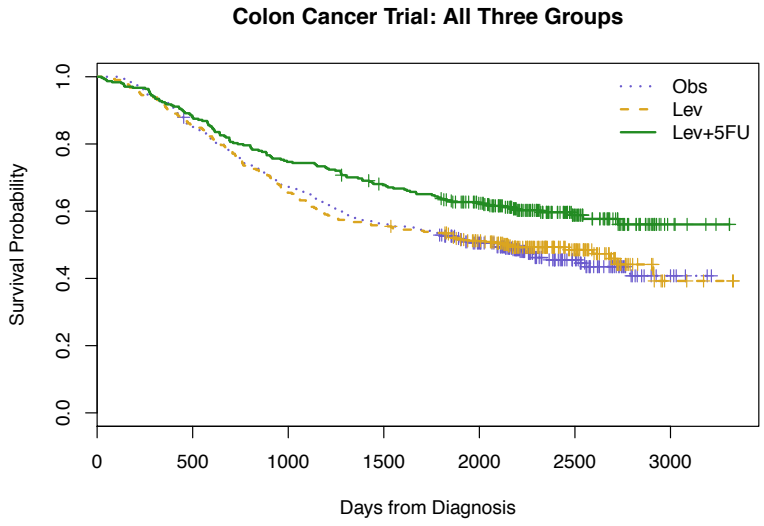
# TEST COMPARISON

| Test | Statistic | P-value |
|---|---|---|
| Wald's | 11.56 | .003 |
| Score | 11.68 | .003 |
| Likelihood Ratio | 12.15 | .002 |

Same hypothesis as 3-group heterogeneity test. Score test is same in large samples.



# COLON CANCER TRIAL DATA



# TREND

- When there are several groups, it is sometimes of interest to test whether risk increases from one group to the next:
  - Several dose groups
  - Other ordered variable
  - Example: tumor differentiation
- For $x = \begin{cases} 1 & \text{well differentiated} \\ 2 & \text{moderately differentiated} \\ 3 & \text{poorly differentiated} \end{cases}$

$$\text{Model: } \lambda(t) = \lambda_0(t)e^{\beta x}$$

- Score test is the same as the trend test
- Could use other values for $x$ (actual dose levels)



# TREND

For $x = \begin{cases} 1 & \text{well differentiated} \\ 2 & \text{moderately differentiated} \\ 3 & \text{poorly differentiated} \end{cases}$

$$\text{Model: } \lambda(t) = \lambda_0(t)e^{\beta x}$$

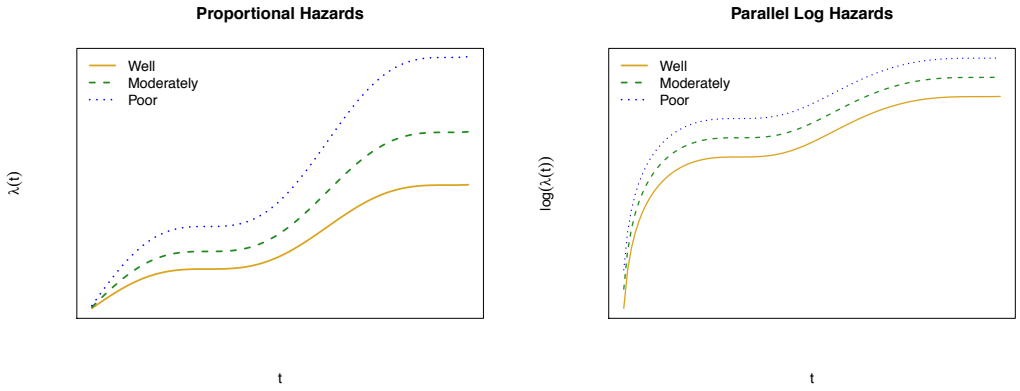
Interpretation of $e^{\beta}$: HR associated with the comparison of one worse differentiation group to one better:

- poorly differentiated to moderately differentiated, or
- moderately differentiated to well differentiated

Q: What is HR comparing poorly differentiated to well differentiated?

A:

# TREND






# TREND WITH DIFFERENTIATION

One presentation based entirely on trend ("grouped linear") model:

| | Hazard Ratio | 95% CI |
|---|---|---|
| One category worse differentiation (well, moderately, poor) | 1.4 | (1.1, 1.8) |
| P = .003 (trend) | | |

I prefer presenting hazard ratios and CI's based on dummy variable model, and providing P-value for trend.

# TREND WITH DIFFERENTIATION

My preferred presentation based on dummy variable mode with trend P-value:

| | n | Deaths | Hazard Ratio | 95% CI |
|---|---|---|---|---|
| Well differentiated | 66 | 26 | 1.0 (reference) | -- |
| Moderately differentiated | 434 | 196 | 1.2 | (0.80, 1.8) |
| Poorly differentiated | 98 | 54 | 1.8 | (1.2, 3.0) |
| P = .003 (trend) | | | | |

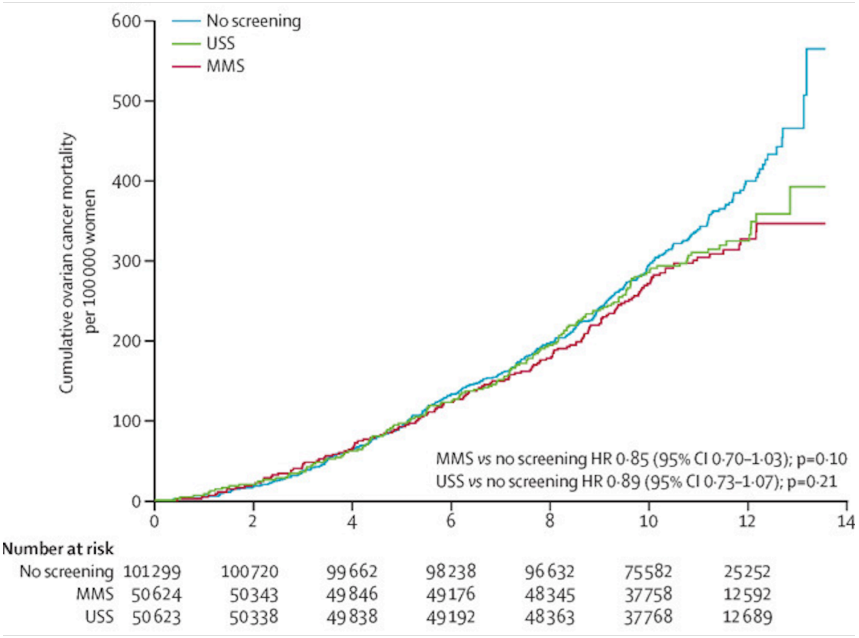
I usually would not present this for an *a priori* trend hypothesis, but for comparison here, the heterogeneity P-value (2 df ) is 0.009.



# OUTLINE

- Motivation:
  - Confounding in observational studies
  - Stratified randomization designs
- Cox Regression model
  - Coefficient interpretation
  - Estimation and testing
  - Relationship to 2- and K-sample tests
  - **Examining non-proportionality**
- Examples throughout

# OVARIAN CANCER SCREENING TRIAL






# PROPORTIONAL HAZARDS

- One way to examine evidence against proportional hazards is to look at plots of scaled Schoenfeld residuals and perform tests based on them.
- For each failing subject there is a Schoenfeld residual for each x variable in the model.
- At the subject's failure time, the residual measures how the value of x for the subject who fails differs from a weighted average of x values for those still at risk. (Weights depend on estimated HR for each subject at risk).
- If consistently high or low over an interval of time, this is evidence that the hazard at that time is even higher (lower) for the subject with that x than the model indicates.

# SCHOENFELD RESIDUALS

## Formula for Schoenfeld residuals

Let $r_i(t) = e^{\hat{\beta} x_i(t)}$ be the estimated hazard ratio for the $i^{th}$ subject at $t$ compared to $x(t) = 0$.

Then for $\overline{x}(\hat{\beta}, t) = \frac{\sum_{\text{at risk at } t} r_i(t) x_i(t)}{\sum_{\text{at risk at } t} r_i(t)}$,

The Schoenfeld residual for the $k^{th}$ subject failing at time $t$ is given by $x_k(t) - \overline{x}(\hat{\beta}, t)$.

The scaled Schoenfeld residual is the Schoenfeld residual divided by a variance estimate.



# SCHOENFELD RESIDUALS

- Grambsch and Therneau (1994) showed that the scaled Schoenfeld residual measures the deviation of a time-dependent log hazard ratio $\beta(t)$ from time-constant $\hat{\beta}$.
- Can use linear regression comparing scaled Schoenfeld residuals to functions of time to examine evidence for lack of constant hazard ratio over time.
- Grambsch PM, Therneau TM. Biometrika. 1994 Sep 1;81(3):515–526.

# COLON CANCER TRIAL DATA

Observation Arm Omitted

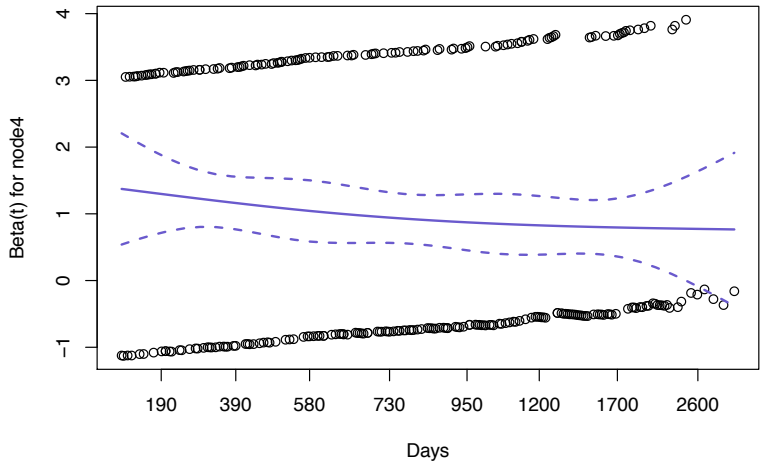
| | $\hat{\beta}$ | exp($\hat{\beta}$) | se($\hat{\beta}$) | z | Pr(>\|z\|) |
|---|---|---|---|---|---|
| 5FU + Lev | -0.34 | 0.71 | 0.12 | -2.83 | 0.0064 |
| 4+ Nodes Pos | 0.98 | 2.67 | 0.12 | 8.08 | <0.0001 |

$e^{\beta_{Rx}}$ CI: (0.5629, 0.9008)

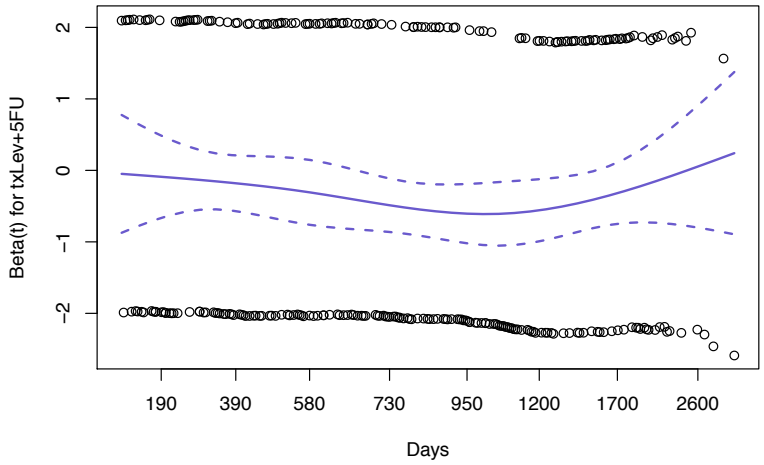
LRT: 8.098 on 1 df, P = 0.0044



# FOR NODE 4 POSITIVITY



# FOR TREATMENT






# TEST FOR NON-PROPORTIONALITY

| Variable | P-value |
|---|---|
| node4 | 0.158 |
| txLev+5FU | 0.560 |

No strong evidence for non-proportionality based on scaled Schoenfeld residuals correlation with "time" S(t).

# TO WATCH OUT FOR:

- Coefficients in Cox regression are positively associated with risk, not survival.
  - Positive β means large values of x are associated with shorter survival.
- Without certain types of time-dependent covariates, Cox regression does not depend on the actual times, just their order.
  - Can add a constant to all times to remove zeros (some packages remove observations with time = 0) without changing inference
- For LRT, nested models must be compared based on same subjects.
  - If some values of variables in larger model are missing, these subjects must be removed from fit of smaller model.
- Hazards may not always be proportional