SESSION 3:
TWO AND K-SAMPLE METHODS

Module 12: Introduction to Survival Analysis
Summer Institute in Statistics for Clinical Research
University of Washington
June, 2017

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

OVERVIEW

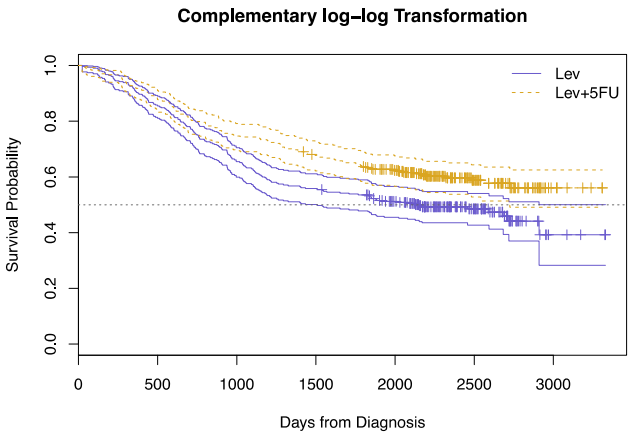
- Session 1
  - Introductory examples
  - The survival function
  - Survival Distributions
  - Mean and Median survival time
- Session 2
  - Censored data
  - Risk sets
  - Censoring Assumptions
  - Kaplan-Meier Estimator and CI
  - Median and CI
- Session 3
  - Two-group comparisons: logrank test
  - 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

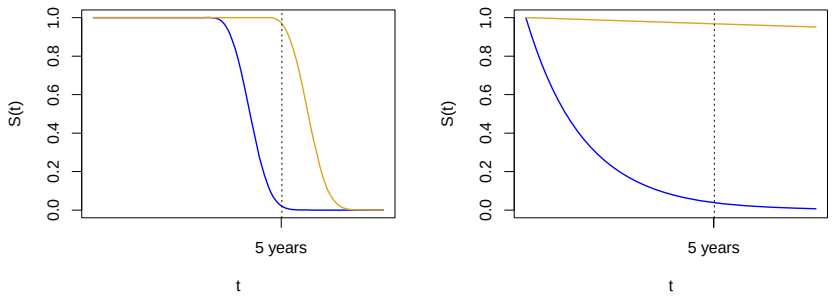
- Statistical significance?

COMPARING SURVIVAL DISTRIBUTIONS

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

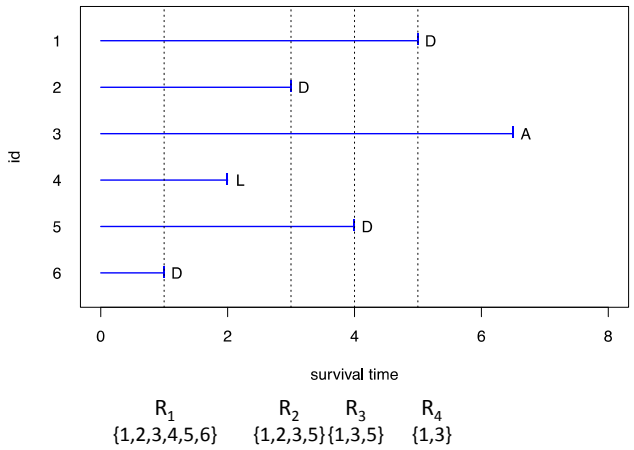
COMPARISON AT 5 YEARS



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.

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

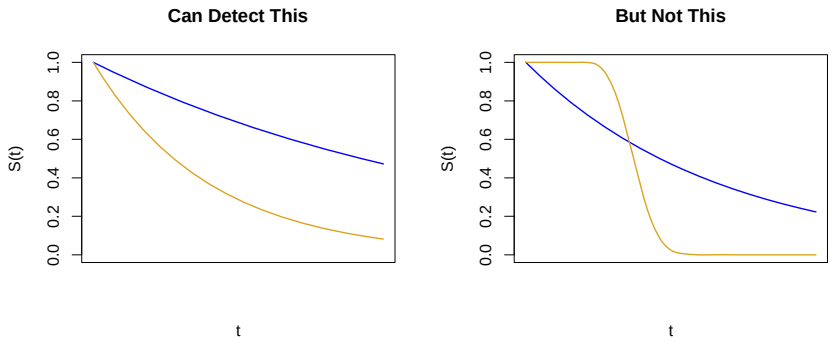
COLON CANCER EXAMPLE

- 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 20)

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 ~ χ^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)

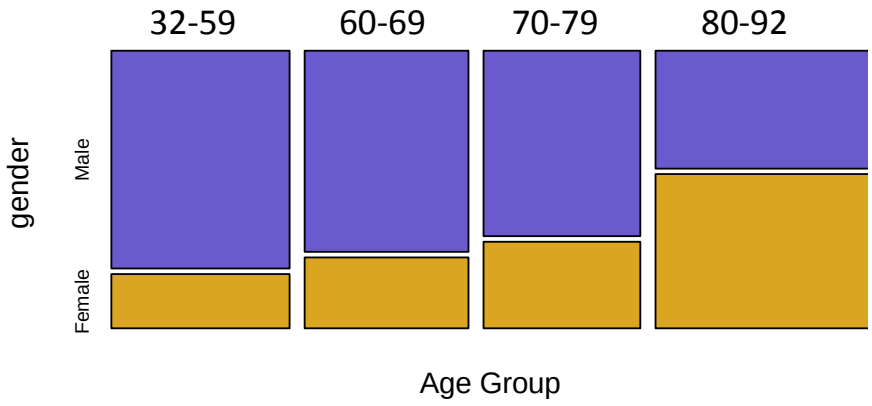
	N	Obs	Exp	$(O-E)^2/E$	$(O-E)^2/V$
Male	65	28	34.7	1.29	4.06
Female	35	23	16.3	2.74	4.06

Chisq = 4.1, 1 df, p=0.044

WHY?

```
> with(whas100,table(age_trend, gender))
         gender
age_trend Male Female
       46   20      5
       65   17      6
       75   15      7
       86   13     17
```

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: $\not\equiv$ 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
- $\lambda_1(t) \leq \lambda_2(t) \leq \ldots \leq \lambda_k(t)$ with < somewhere, or
- $\lambda_1(t) \geq \lambda_2(t) \geq \ldots \geq \lambda_k(t)$ with > somewhere
- Placebo and two or more doses of a therapeutic agent
- Pre-hypothesized

TREND

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

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

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

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

- 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