Lecture 15: More Non-Parametric Estimation Methods and Testing in Survival Analysis

Lin Chen

Department of Public Health Sciences
The University of Chicago

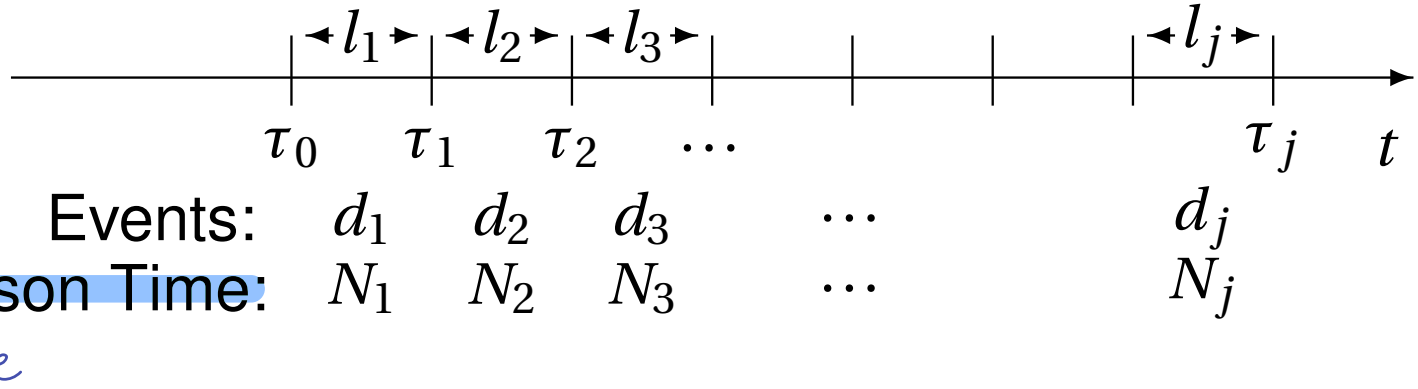
Non-parametric procedures

- In Lecture 14, we described non-parametric (distribution-free) methods to estimate survivor function and hazard function: Kaplan-Meier method and life-table method.
- We briefly described hazard estimator derived after KM or life-table survival curve estimator
- In this lecture, we will first describe a few additional approaches and related quantities and then introduce commonly used testing procedures after nonparametric estimation

The Hazard Function - Estimating the hazard rate $\lambda(t)$

An alternate way of formulating the question (similar to actuarial, but also like KM)

- Imagine dividing time into small intervals:



where d_j is the number of events in interval j and N_j is the total person-time observed in interval j.

The Hazard Function - Estimating the hazard rate $\lambda(t)$

- **Assume**:
  - λ_j = event (hazard) rate for subjects in interval j is constant within the interval (conditioning on surviving previous intervals)
  - d_j = Poisson(μ_j) where $\mu_j = N_j \lambda_j$
- Then the **maximum likelihood** estimate for λ_j is given by:

$$\hat{\lambda}_j = d_j / N_j.$$

(# of events / # of person-years)

The estimated rate in interval j equals the observed number of cases divided by the person-time at risk (compare with the life-table estimate of the hazard)
This method accounts for both the number of events and the total person-time within the interval.

Example IUD data: *Estimating the Hazard rate $\lambda(t)$*

- Estimate interval hazard rates using average hazard idea:

```
* Need to stset with id() option - make a unique pt identifier
. use "discontinuation_IUD.dta", clear
. gen PTid=_n
. stset time, failure (status) id(PTid)
.....
. stptime, at (10 20:110)
        failure _d:  status
  analysis time _t:  time
                id:  PTid
```

person-time

Cohort	Person-time	Failures	Rate	[95% Conf.	Interval]
(0 - 10]	180	1	.00555556	.0007826	.0394393
(10 - 20]	160	1	.00625	.0008804	.0443692
(20 - 30]	133	1	.0075188	.0010591	.0533765
(30 - 40]	114	1	.00877193	.0012356	.0622726
(40 - 50]	100	0	0	.	.
(50 - 60]	89	1	.01123596	.0015827	.0797648
(60 - 70]	70	0	0	.	.
(70 - 80]	65	1	.01538462	.0021671	.1092165
(80 - 90]	60	0	0	.	.
(90 - 1.0e+02]	50	2	.04	.0100039	.1599375
> 1.0e+02	25	1	.04	.0056345	.2839629
Total	1046	9	.00860421	.0044769	.0165365

The hazards estimates from actuarial

- Based on actuarial methods (Eqn. 12 Slide 38 Lecture 14):

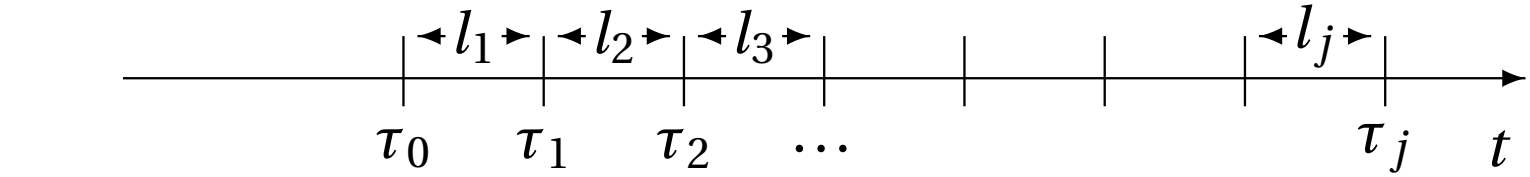
```
. ltable time status, interval(10) hazard
```

Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
10	20	18	0.1176	0.0781	0.0125	0.0088	0.0000	0.0298
20	30	14	0.1176	0.0781	0.0000	.	.	.
30	40	13	0.2588	0.1126	0.0174	0.0123	0.0000	0.0414
50	60	10	0.3412	0.1267	0.0118	0.0117	0.0000	0.0348
70	80	7	0.4353	0.1392	0.0154	0.0153	0.0000	0.0454
90	100	6	0.6235	0.1429	0.0400	0.0277	0.0000	0.0943
100	110	4	0.7741	0.1448	0.0500	0.0484	0.0000	0.1449

- The interval hazard rate estimates for $\lambda(t)$ are a bit different from the hazard estimates for $h(t)$ from actuarial methods (due to how censoring being treated). But the two methods are consistent even in a small data set.

The Nelson-Aalen Estimator of $\Lambda(t)$

- Recall that the cumulative hazard function $\Lambda(t) = -\log\{S(t)\}$ Therefore, a natural estimator of the cumulative hazard function is $\hat{\Lambda}(t) = -\log(S_{KM}(t))$
- Another estimator: Recall also that $\Lambda(t) = \int_0^t \lambda(s)\,ds$ which provides an additional estimator



Events:	d_1	d_2	d_3	$\cdots$	d_j
Person Time:	N_1	N_2	N_3	$\cdots$	N_j

$$\hat{\lambda}_j = d_j / N_j$$

$$\hat{\Lambda}(t) = \int_0^t \hat{\lambda}(s)\,ds = \sum_{t_j \le t} l_j \frac{d_j}{N_j}$$

The Nelson-Aalen Estimator of $\Lambda(t)$

- Let intervals shrink: $J \to \infty$ so that $l_j = \tau_j - \tau_{j-1} \to 0$
- Each interval contains a single subject with $d_{ij} = 1$ (or d_j subjects if ties)
- No one enters or leaves (approximately)
- For sufficiently small intervals

$$N_j \approx l_j n_j$$

The Nelson-Aalen Estimator of $\Lambda(t)$

Definition: The **Nelson-Aalen** estimator of interval hazard rate $\Lambda(t)$ is:

$$\hat{\Lambda}(t) = \sum_{t_j \le t} \frac{d_j}{n_j}$$

where the sum is over the *distinct* observed event times.

- As the total sample size increases, the total number of jumps grows and the jump sizes get smaller, resulting in a better approximation to a continuous function.
- Variance from Greenwood's formula:

$$\text{Var}\hat{\Lambda}(t) = \sum_{t_j \le t} \frac{d_j}{n_j^2}$$

- In small samples, the Nelson-Aalen estimator has better properties than Kaplan-Meier estimator (Eqn. 9 on Slide 30 Lecture 14) and life-table estimator (Eqn. 12 on Slide 38 Lecture 14).

Estimating the Nelson-Aalen (N-A) Hazard

- Calculate the N-A estimator for cumulative hazard at specified time points and plot it

```
. sts list, na at (0 10 :110)

        failure _d:  status
  analysis time _t:  time
```

Time	Beg. Total	Fail	Nelson–Aalen Cum. Haz.	Std. Error	[95% Conf.	Int.]
0	0	0	0.0000	0.0000	.	.
10	18	1	0.0556	0.0556	0.0078	0.3944
20	15	1	0.1222	0.0868	0.0304	0.4915
30	13	1	0.1991	0.1160	0.0636	0.6235
40	11	1	0.2825	0.1428	0.1049	0.7608
50	11	0	0.2825	0.1428	0.1049	0.7608
60	8	1	0.4075	0.1898	0.1636	1.0152
70	8	0	0.4075	0.1898	0.1636	1.0152
80	7	1	0.5503	0.2375	0.2362	1.2824
90	7	0	0.5503	0.2375	0.2362	1.2824
100	5	2	0.9170	0.3524	0.4318	1.9476
110	3	1	.	.	.	.

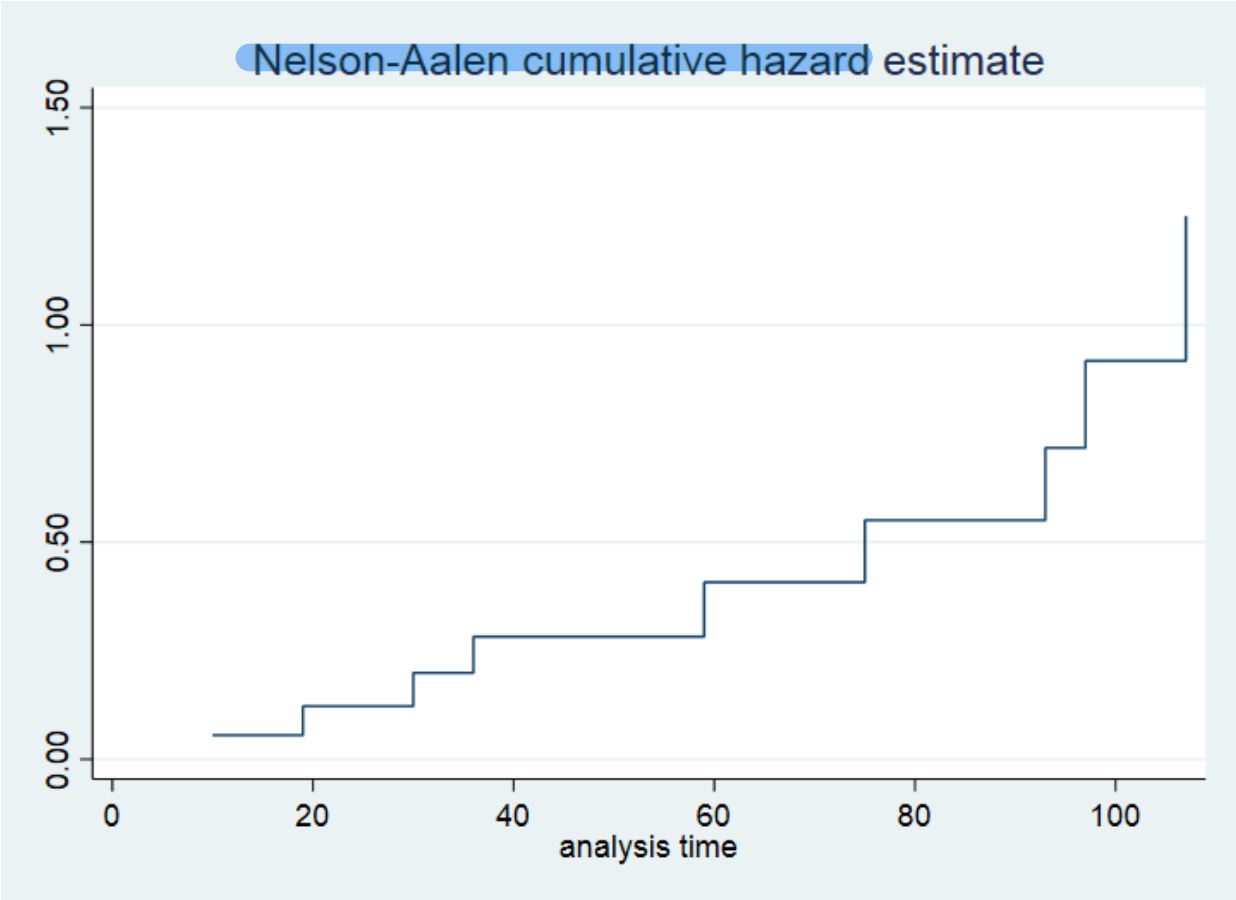
Note: Nelson–Aalen function is calculated over full data and evaluated at indicated times; it is not calculated from aggregates shown at left.

- The cumulative hazard is important for model form diagnostic purposes.

Estimating the N-A Cumulative Hazard

- Plot the cumulative hazard estimate

```
. sts graph , na
```

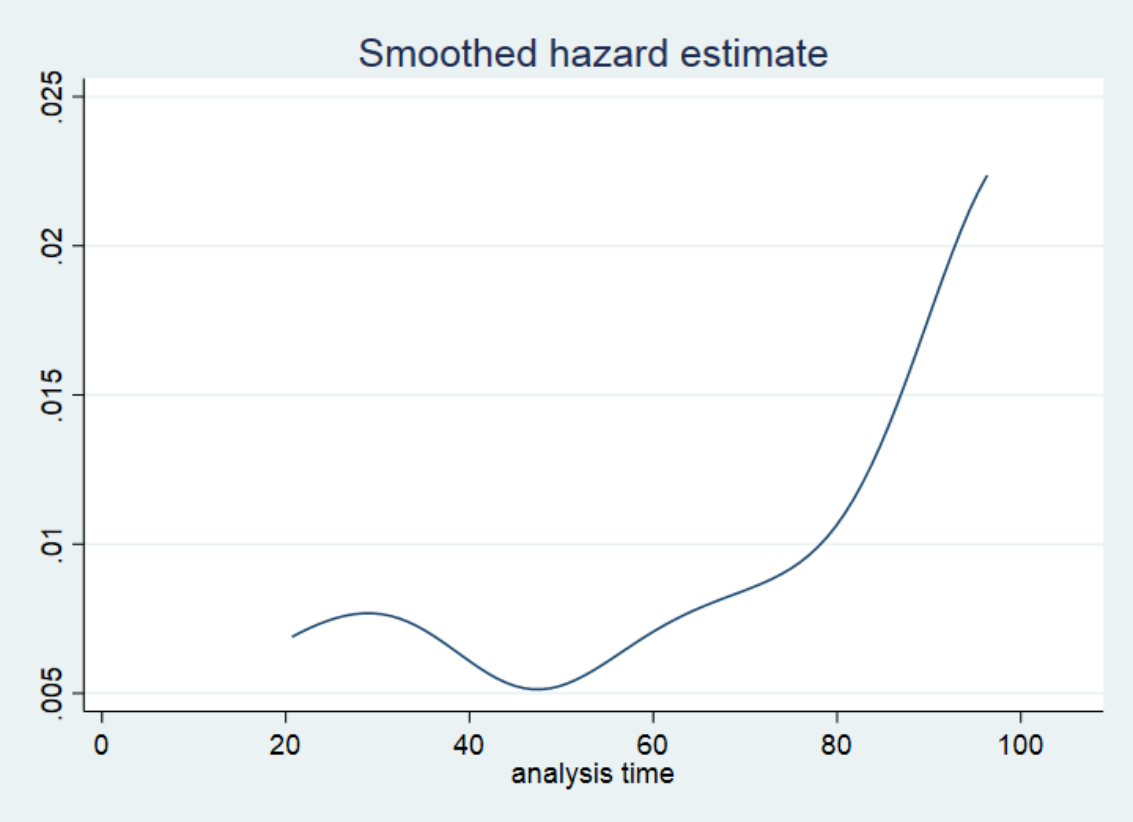


Another hazard estimator – Kernel Estimate of $\lambda(t)$

- Basic Idea: At each time t, let the estimate of $\lambda(t)$ be a weighted average of the jumps in $\hat{\Lambda}(t)$ at nearby time points per unit of time
- Choose a *bandwidth* $\pm b$ within which jumps are averaged.
- Choose a kernel or weight function $K(\cdot)$. Here we used a Gaussian kernel.
- As b becomes larger, the estimate of $\hat{\lambda}$ becomes smoother. We trade off roughness and bias by manipulating b

Kernel Estimator for hazard

```
. sts graph, hazard kernel(gaussian) width(10)
```



The Average Hazard Rate as a Data Summary

From time to event data, we can estimate the incidence rate (λ) for an interval or for overall data. To compute the overall λ, we divide the number of responses by the sum of observation time over all individuals:

$$\hat{\lambda} = \frac{D}{\sum_{i=1}^{N} t_i},$$

where

- t_i is the follow-up time for patient i
- Add up each patient's time observed until failure or censoring gives the total patient-time (total person-week, total person-year)
- D is the number of failure events (relapses, deaths, etc). When events are recorded using a $0/1$ indicator d_i, then $D = \sum_{i=1}^{N} d_i$
- Here λ is called the *average hazard rate* or incidence rate
- The estimator $\hat{\lambda}$ is the maximum likelihood estimator for the exponential survival model

Estimation of the Average Hazard Rate

We can obtain the average hazard rate using the *stsum* or *strate* command in Stata. For the IUD data:

```
. stsum

        failure _d:  status
  analysis time _t:  time

        |                incidence          no. of    |------ Survival time ---|
        | time at risk      rate           subjects        25%       50%       75%
--------+-------------------------------------------------------------------------
  total |         1046   .0086042                 18         36        93       107

. * same as if one adds up needed quantities directly, 9 death, 1046 total time
. egen events = total(status)
. egen pmonths = total(time)
. display events/pmonths
.00860421
```

Comparing the survival of two groups

Next, we will use an example to illustrate how to compare the survival of two groups, formally, informally, and visually.

Example – 6-MP Leukemia data: Remission times, in weeks, of leukemia patients randomized to receive either 6-mercaptopurine (6-MP) or not (control).

6-MP (N=21)	Control (N=21)
6,6,6,6*,7,	1,1,2,2,3,
9*,10,10*,11*,	4,4,5,5,8,
13,16,17*,19*,	8,8,8,11,
20*,22,23,25*,	11,12,12,15,
32*,32*,34*,35*	17,22,23

* denotes a censored observation

Average Hazard Rate Comparison

We can compute within relevant subject strata. Then to contrast two groups, we can form a difference or ratio of average hazard rates. In the tamoxifen trial, the ratio of mortality by treatment group is:

```
. stset time, failure (relapse)
...
. sort txgrp

. stsum, by(txgrp)

        failure _d:  relapse
  analysis time _t:  time
```

txgrp	Time at risk	Incidence rate	Number of subjects	Survival time 25%	Survival time 50%	Survival time 75%
0	182	.1153846	21	4	8	12
1	359	.0250696	21	13	23	.
Total	541	.0554529	42	6	12	23

- The rate ratio (6-mp/control) is $\frac{.0250696}{.1153846} = 0.217$ or a 78.3% relative remission reduction. The absolute difference is 0.090315 per person week. These are summaries upon which we may want to make inference about treatment.

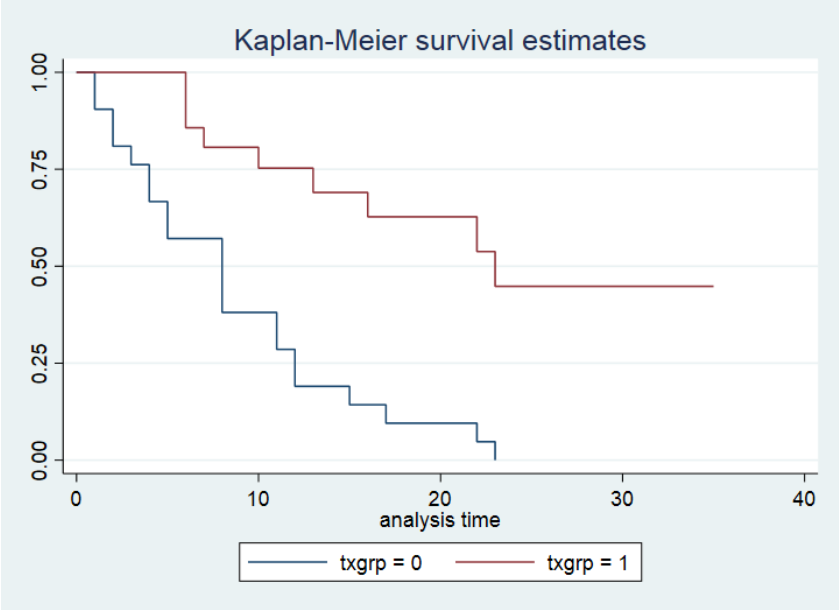
Comparing Two Survival Distributions

- A question of general interest, how to compare two survival distributions for two groups and make inference?

  E.g., is there a difference between these survival curves, or, is 6-MP treatment associated with a longer time in cancer remission?

```
. sts graph, by(txgrp)

        failure _d:  relapse
  analysis time _t:  time
```



Comparing Two Survival Distributions

- One may compare the survival probabilities at a particular time point t_0:
  - Use the Kaplan-Meier estimate to obtain: $\hat{S}_0(t_0) - \hat{S}_1(t_0)$
  - Use Greenwood's variance estimator for the difference in survival probabilities.
  - Compute the Z statistic and the statistic is approximately normally distributed under the null
- Issues:
  - Which time we should compare? Subjective.
  - One time point is insufficient to represent the whole distribution
- Goal: compare the entire survival curves

The logrank test for Hazard Ratio (HR)

The logrank test is a widely-used test for comparing the survival distributions of two (or more) groups. It tests for the hazard rate ratio between groups.

$$\begin{aligned} H_0: &\quad HR = 1.0 \quad \text{versus} \\ H_A: &\quad HR \neq 1.0 \end{aligned}$$

- The **logrank test** is a non-parametric test assuming non-informative censoring. It works for time-independent hazards or time-varying hazards if the hazard ratio is constant across time.
- Under the null $H_0: HR = 1$. Under the alternative, the two hazard rate ratio is assumed to be constant, we have $H_A: \lambda_1(t) = \phi \times \lambda_0(t)$.

The Logrank Test

To facilitate, we specify the following:

For $t_1 < t_2 < ... < t_D$ denote the set of distinct survival times occurring in the pooled sample (both groups combined)

We want to test $H_0 : S_1(t) = S_0(t)$ or $H_0 : \lambda_1(t) = \lambda_0(t)$ for all t

vs. $H_A : S_1(t) \neq S_0(t)$ or $H_A : \lambda_1(t) \neq \lambda_0(t)$

Specifically, the alternative we will specify is of the form

vs. $H_A : S_1(t) = [S_0(t)]^{\phi}$ or $H_A : \lambda_1(t) = \phi\lambda_0(t)$

The Logrank Test (cont.)

- At each failure time, consider a 2×2 table of the form:

	Failure		
	Yes$^+$	No$^-$	Risk Set
Group 0$^-$	d_{0k}	$y_{0k} - d_{0k}$	y_{0k}
Group 1$^+$	d_{1k}	$y_{1k} - d_{1k}$	y_{1k}
Total	d_k	$y_k - d_k$	y_k

- For failure time t_k compute the observed number of deaths in group 1 and the expected number of deaths under H_0.

observed
$$\text{Observed:} \quad o_{1k} = d_{1k}$$

expected
$$\text{Expected:} \quad e_{1k} = \frac{y_{1k} d_k}{y_k}$$

- At each failure time, compute:

$$u_k = (o_{1k} - e_{1k})$$

The Logrank Test (cont.)

- Under H_0:

$$\mathrm{E}(u_k) = 0$$

$$\mathrm{Var}(u_k) = \frac{y_{1k} y_{0k} (y_k - d_k) d_k}{y_k^2 (y_k - 1)} = v_k$$

- The **logrank statistic** is given by:

$$T_L = \frac{\left[\sum_{k=1}^{D} (o_k - e_k)\right]^2}{\sum_{k=1}^{D} v_k} = \frac{\left[\sum_{k=1}^{D} u_k\right]^2}{\sum_{k=1}^{D} v_k}$$

- Assuming independent failure times and for large n,

$$T_L \dot{\sim} \chi_1^2$$

or

$$Z = \sqrt{T_L} \dot{\sim} \mathcal{N}(0, 1)$$

Logrank Statistic - 6-MP leukemia example

To conduct a logrank test in Stata, use `sts test`:

```
. stset time, failure (relapse)
....
. sts test txgrp

        failure _d:  relapse
  analysis time _t:  time


Log-rank test for equality of survivor functions

      |    Events         Events
txgrp |  observed       expected
------+------------------------
0     |        21          10.75
1     |         9          19.25
------+------------------------
Total |        30          30.00

           chi2(1) =       16.79
           Pr>chi2 =      0.0000
```

compare survival of 2 groups

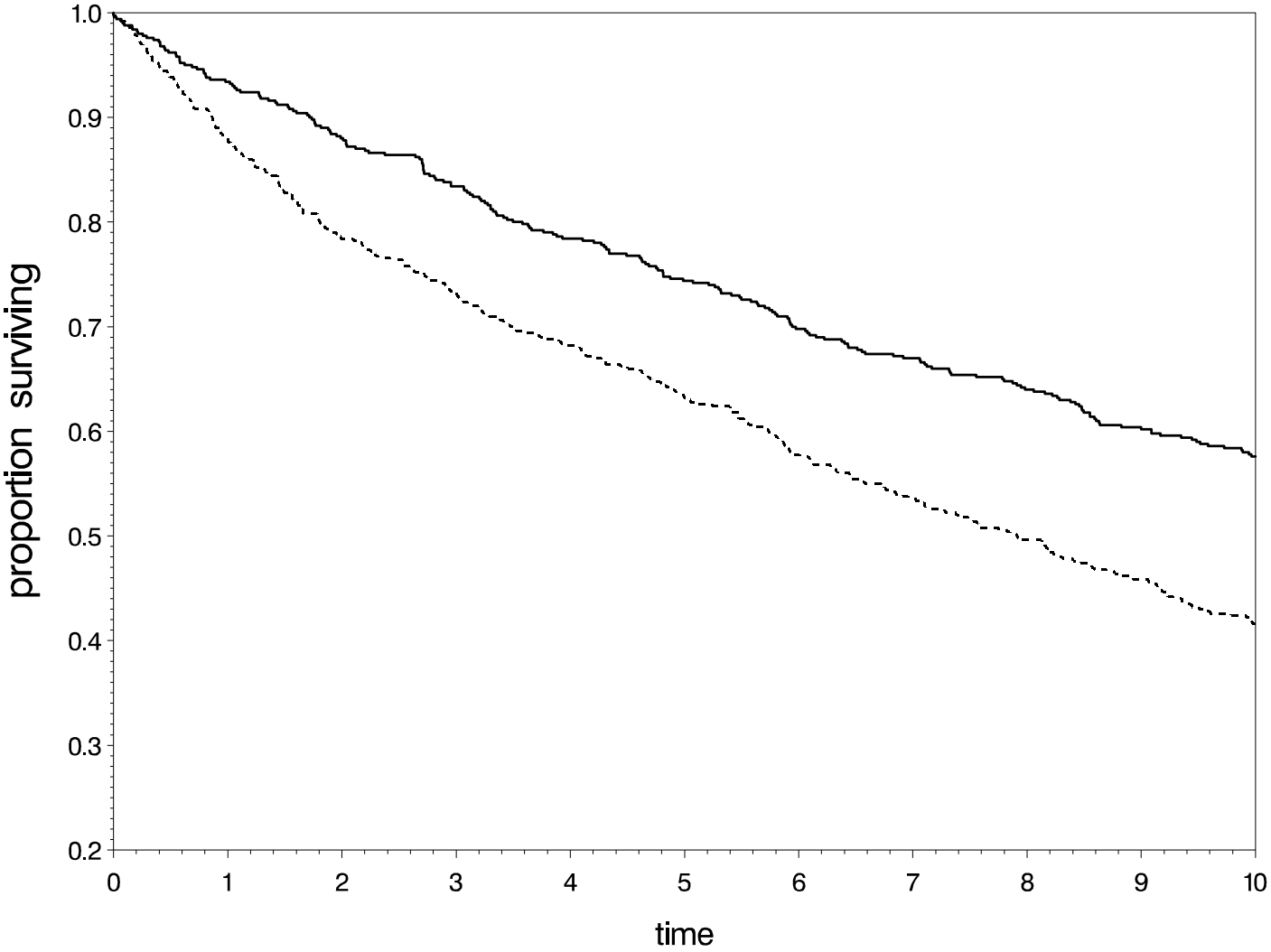
There is a significant difference in survival between the 6-MP versus the control group.

Logrank test and the Proportional Hazards (PH) assumption

- When using the logrank test, the alternative hypothesis we specified earlier was of the form $H_A : \lambda_1(t) = \phi\lambda_0(t)$
  It assumes that the hazard ratio ϕ does not depend on time and is a constant. This is the **proportional hazards (PH)** assumption.
  But the logrank test is robust to non-severe violations to the PH assumption.
- The logrank test has the most power when the PH assumption hold. The test will still detect differences deviating (to a reasonable extent) from this form.

Proportional Hazards

Proportional hazards looks like this on the survival curve scale:



Weighted Logrank Statistics

- What if PH does not hold, or we are interested in other alternatives? Consider weighting $(Obs - Exp)$ differently over time
- This will enable us to inflate the influence of early or late differences (note: weight =1 under standard logrank test)

  → Potential for increased power under non-proportional hazards

$$T_W = \frac{\left[\sum_{k=1}^{D} w_k(o_k - e_k)\right]^2}{\sum_{k=1}^{D} w_k^2 v_k} = \frac{\left[\sum_{k=1}^{D} w_k u_k\right]^2}{\sum_{k=1}^{D} w_k^2 v_k}$$

Weighted Logrank Statistics

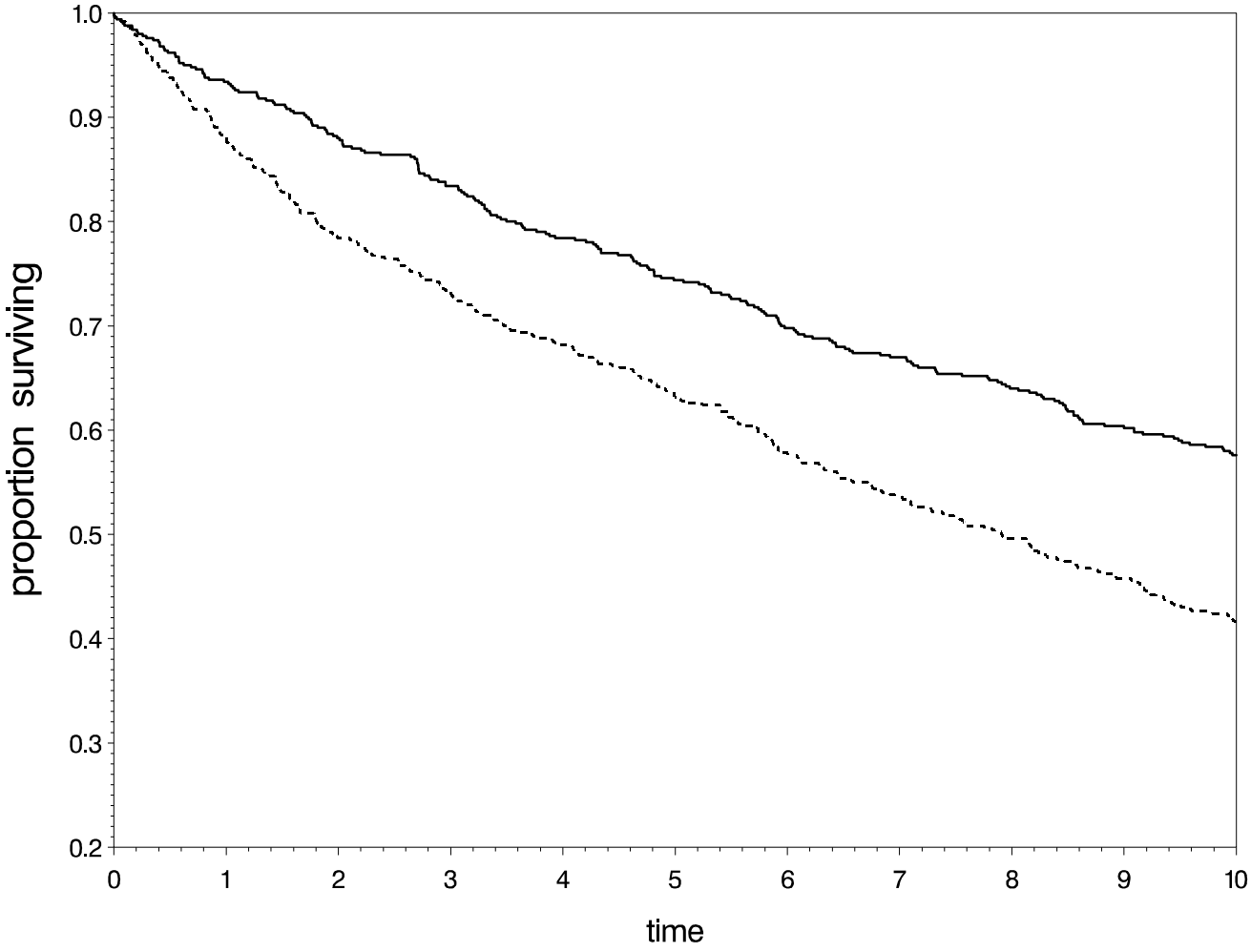
Choices for w_k:

1. $w_k = n_k$ gives the Gehan-Breslow test (weights equal to the total number of subjects at risk at each failure time). Applies greater weight to early failure times.
2. $w_k = \hat{S}_{KM}(t_k-)$ gives the generalized Wilcoxon test (weights equal to the pooled estimate of survival (across groups) just prior to time t_k). Applies greater weight to early failure times.
   - Equivalent to the Wilcoxon rank sum statistic when there is no censoring.

Weighted Logrank Statistics

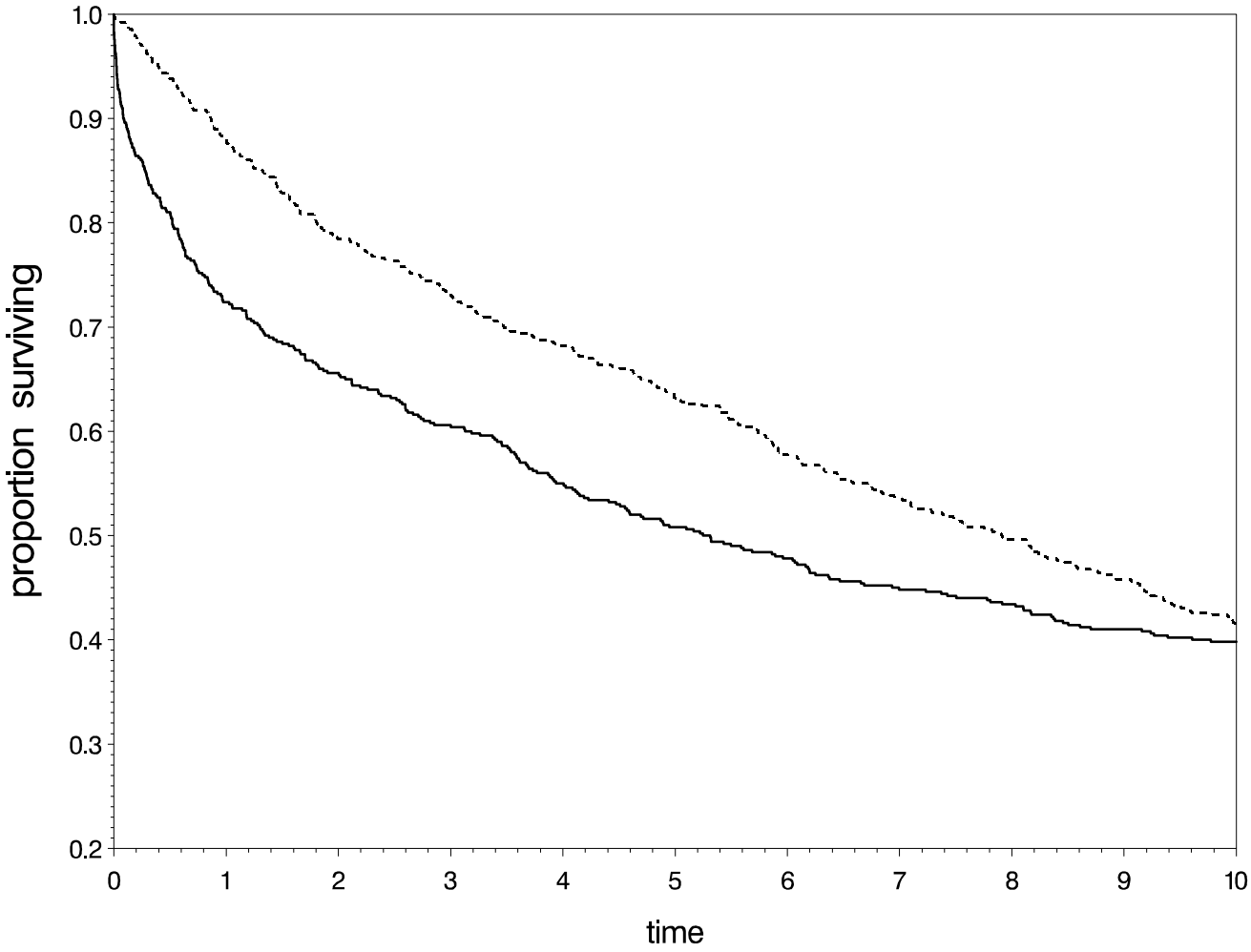
- A more general approach to weighted tests is the $G^{\rho,\gamma}$ family by Fleming and Harrington (1991)
  - $w_k = \left[\hat{S}_{KM}(t_k-)\right]^{\rho}\left[1-\hat{S}_{KM}(t_k-)\right]^{\gamma}$
  - $\rho = \gamma = 0$ gives the usual logrank statistic
  - $\rho = 1$ and $\gamma = 0$ gives the generalized Wilcoxon test (greater weight to earlier time)
  - $\rho = 0$ and $\gamma > 0$ more weight to late differences

Weighted Logrank Statistics - under proportional hazards



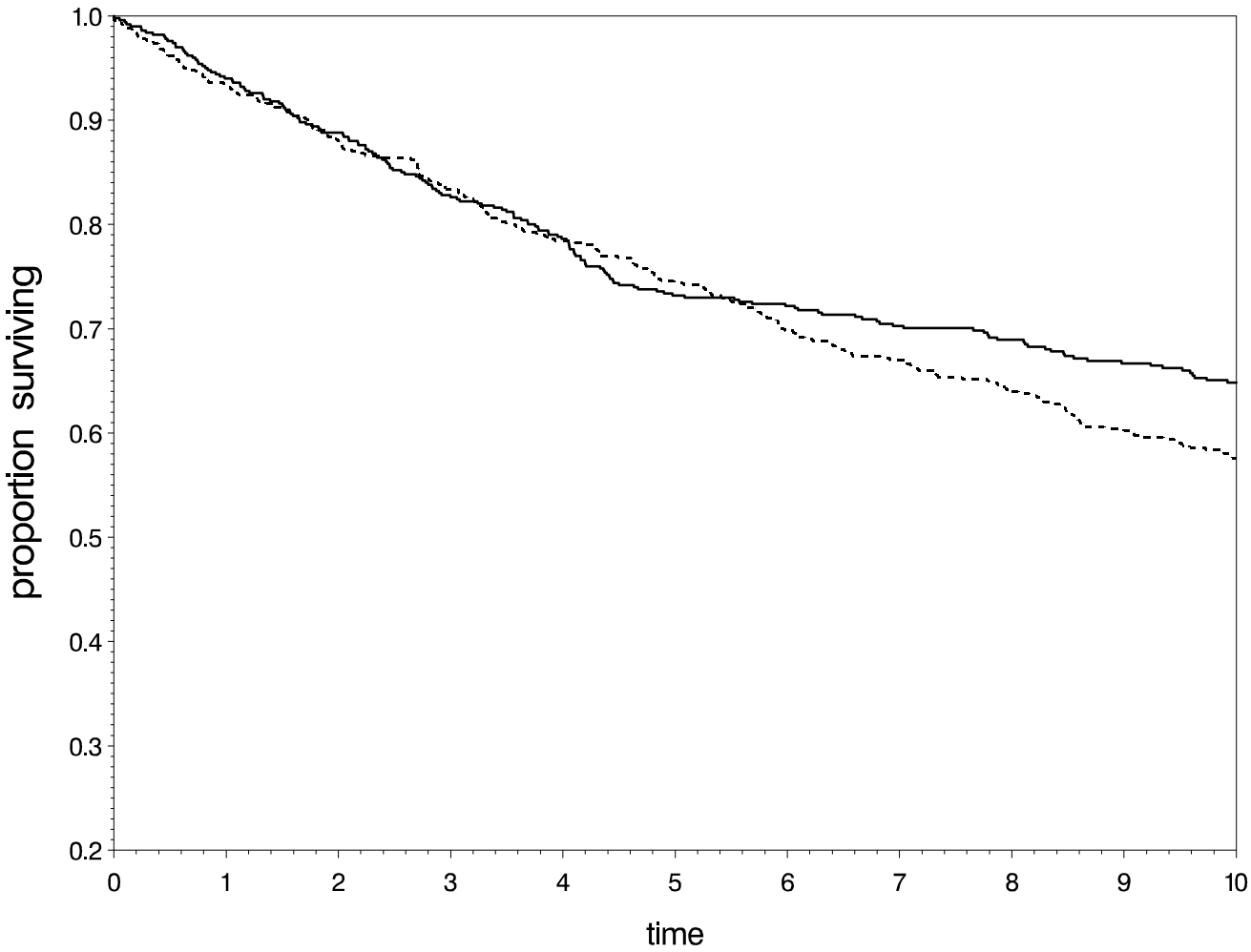
logrank test: $p < 0.0001$ Wilcoxon test: $p < 0.0001$

Weighted Logrank Statistics - early difference in hazards that later converges



logrank test: p = 0.082 Wilcoxon test: p = 0.0003

Weighted Logrank Statistics - late diverging hazards



logrank test: p = 0.033 Wilcoxon test: p = 0.084

Weighted Logrank Statistics

Back to the leukemia example:

- We know that the (unweighted) logrank statistic will be most powerful under proportional hazards. How can we (informally) check the proportional hazards assumption?
  - If we have proportional hazards, then

$$\lambda_1(t) = \phi\lambda_0(t)$$

so that

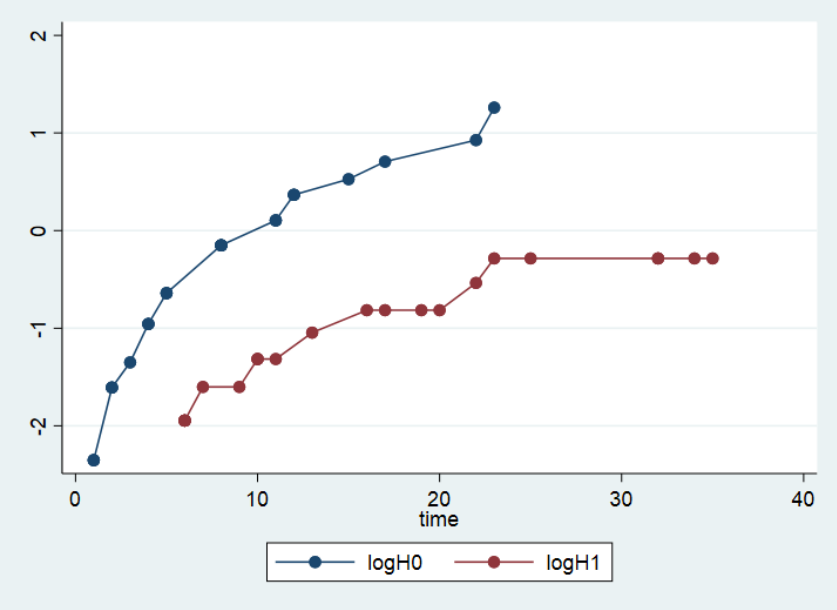
$$\log\Lambda_1(t) = \log(\phi) + \log\Lambda_0(t)$$

- So, if the log cumulative hazards are roughly parallel, the logrank test will be most powerful

Weighted Logrank Statistics

Look at the log cumulative hazard lines in Stata

```
** Compute the Nelson-Aalen estimate for each group
. sts gen cumhaz=na, by(txgrp)
. gen logH = log( cumhaz )
. gen logH0 = logH if (txgrp==0)
. gen logH1 = logH if (txgrp==1)
. scatter logH0 time, connect(l) || scatter logH1 time, connect(l)
```



- Plot showed two roughly parallel log cumulative hazard lines. Do we expect the generalized Wilcoxon test to be more or less powerful than the logrank test for this situation?

Weighted Logrank Statistics

- First, the unweighted logrank test

```
. sts test txgrp
Log-rank test for equality of survivor functions
       |    Events          Events
txgrp  |  observed        expected
-------+-------------------------
0      |        21           10.75
1      |         9           19.25
-------+-------------------------
Total  |        30           30.00
                chi2(1) =     16.79
                Pr>chi2 =    0.0000
```

- Fleming-Harrington test with higher weight to earlier time

```
. sts test txgrp, fh(1,0)
         failure _d:  relapse
   analysis time _t:  time
Fleming-Harrington test for equality of survivor functions

       |    Events          Events         Sum of
txgrp  |  observed        expected          ranks
-------+-----------------------------------------
0      |        21           10.75       6.877045
1      |         9           19.25      -6.877045
-------+-----------------------------------------
Total  |        30           30.00              0

                chi2(1) =     14.46
                Pr>chi2 =    0.0001
```

Weighted Logrank Statistics

- Fleming-Harrington test with higher weight to later time

```
. sts test txgrp, fh(0,2)

        failure _d:  relapse
  analysis time _t:  time

Fleming-Harrington test for equality of survivor functions

       |    Events         Events          Sum of
txgrp  |   observed       expected          ranks
-------+-------------------------------------------
0      |         21          10.75      1.7088049
1      |          9          19.25     -1.7088049
-------+-------------------------------------------
Total  |         30          30.00              0

           chi2(1) =      11.14
           Pr>chi2 =     0.0008
```

When the PH assumption is satisfied, upweighting or downweighting early large difference diminishes significance (a tiny bit) relative to the unweighted log-rank test

Choosing the Right Logrank Statistic

How should weights be chosen?

- For scientific inference, it is **not** reasonable to look at the survival curves first, then choose weights.
- Standard (unweighted) logrank test by far the most commonly used version
- First, ask whether there is a reason to believe we will have non-proportional hazards
  - If not, use the logrank test - most powerful under PH and reasonable under minor to moderate deviations
  - If so, consider what survival differences are most meaningful (early vs late)
    → Childhood cancer – late differences may be more important
    → Advanced stage cancer – early differences may be more important

Summary

Nonparametric Estimation and Tests

- Logrank test (like chi-squared test for 2x2 table) only provides a p-value and it does not provide the magnitude of survival difference. It is most powerful under the PH assumption. And the (log) cumulative hazard estimates are helpful for checking the PH assumption.
- Nonparametric estimation and testing are the most common form of analysis in one major research construct: the *randomized trial.* This is because, aside from stratification factors that can be accommodated easily, covariates do not typically influence the main comparison of interest

completely random

- Next: modeling survival data - needed to study or account for covariates related to outcome. The hazard function will largely be the basis on which we model factors related to survival time