STAT347: Generalized Linear Models

Lecture 14

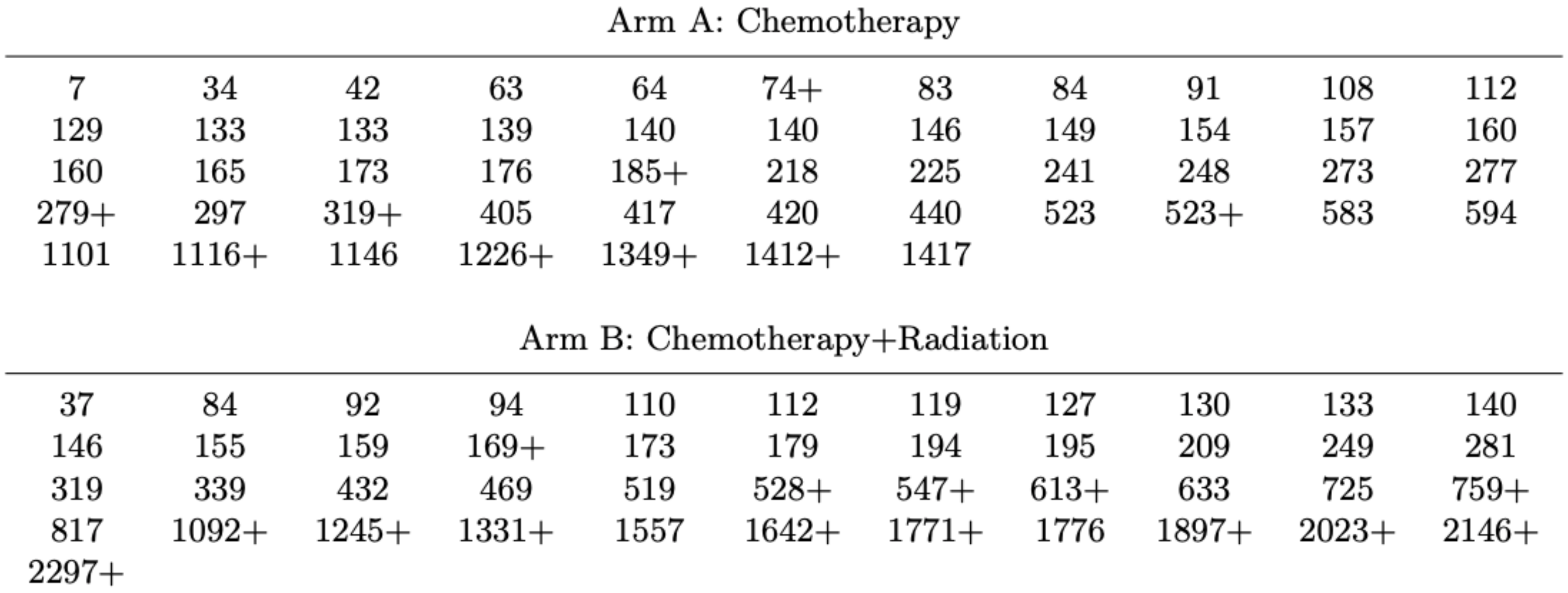
Winter, 2024
Jingshu Wang

Today’s topics:

- Survival analysis
  - Examples of survival analysis datasets
  - Basic concepts in survival analysis: survival function, hazard rate, censoring
  - Kaplan-Meier estimator of the survival function
  - Log-rank test

Example 1:
Northern California Oncology Group (NCOG) study

- Two treatments for head and neck cancer:
  Arm A: Chemotherapy; Arm B: Chemotherapy + Radiation
- Data: censored survival time in days
  ‘+’ indicate patients still alive on their final day of observation



Arm A: Chemotherapy

7	34	42	63	64	74+	83	84	91	108	112
129	133	133	139	140	140	146	149	154	157	160
160	165	173	176	185+	218	225	241	248	273	277
279+	297	319+	405	417	420	440	523	523+	583	594
1101	1116+	1146	1226+	1349+	1412+	1417				

Arm B: Chemotherapy+Radiation

37	84	92	94	110	112	119	127	130	133	140
146	155	159	169+	173	179	194	195	209	249	281
319	339	432	469	519	528+	547+	613+	633	725	759+
817	1092+	1245+	1331+	1557	1642+	1771+	1776	1897+	2023+	2146+
2297+										

Example 1:
Northern California Oncology Group (NCOG) study

- Two treatments for head and neck cancer:
  Arm A: Chemotherapy; Arm B: Chemotherapy + Radiation
- Data: censored survival time in days
  + indicate patients still alive on their final day of observation

Main questions:

- Is the Arm B more effective treatment than Arm A?
- Instead of just compare the mean survival time, we would like to know more information about the survival time distribution (the survival curve)
- How to deal with "lost to follow-up" (censoring)?

Example 2: duration of nursing home stay

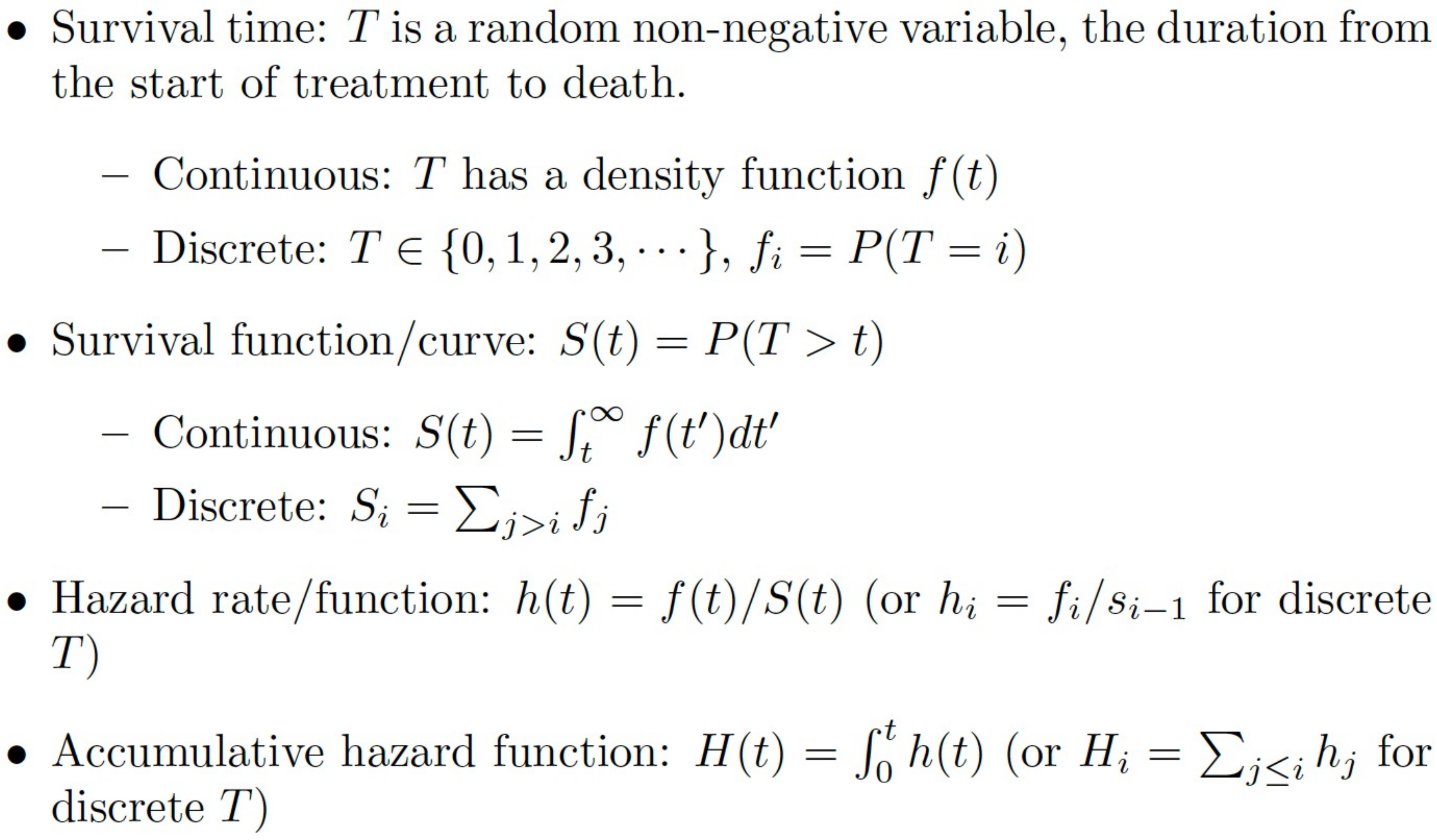
- Goal: assess the effects of different financial incentives on length of
- stay.
- Treated nursing homes received higher per diems for Medicaid patients, and bonuses for improving a patient's health and sending them home.
- Study included 1601 patients admitted between May 1, 1981 and April 30, 1982.

Measured variables:

- LOS - Length of stay of a resident (in days)
- AGE - Age of a resident
- RX - Nursing home assignment (1:bonuses, 0:no bonuses)
- gender, age, married or not, heath status
- CENSOR - Censoring indicator (1:censored, 0:discharged)

Goal: treatment effect on stay length after adjusting for other covariates and censoring?

Basic concepts



- Survival time: T is a random non-negative variable, the duration from the start of treatment to death.
  - Continuous: T has a density function $f(t)$
  - Discrete: $T \in \{0, 1, 2, 3, \cdots\}$, $f_i = P(T = i)$
- Survival function/curve: $S(t) = P(T > t)$
  - Continuous: $S(t) = \int_t^\infty f(t')dt'$
  - Discrete: $S_i = \sum_{j>i} f_j$
- Hazard rate/function: $h(t) = f(t)/S(t)$ (or $h_i = f_i/s_{i-1}$ for discrete T)
- Accumulative hazard function: $H(t) = \int_0^t h(t)$ (or $H_i = \sum_{j \leq i} h_j$ for discrete T)

Basic concepts

- The survive function and hazard rate provide more information than $E(T)$.
- An important fact is that knowing one of the three functions of $H(t), h(t)$ and $S(t)$ will enable inferring the other two functions.
- For discrete T:

$$S_i = \prod_{j=0}^{i} P[T \geq j+1 \mid T \geq j] = \prod_{j=0}^{i} (1 - h_j)$$

- For continuous T:

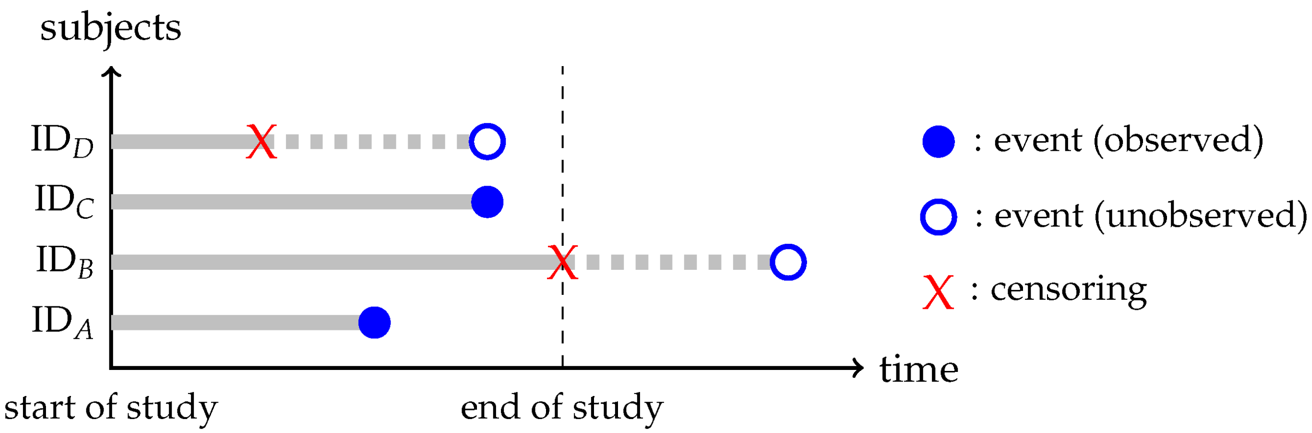
$$S(t) = e^{-H(t)}$$

Concept of censoring

Censoring

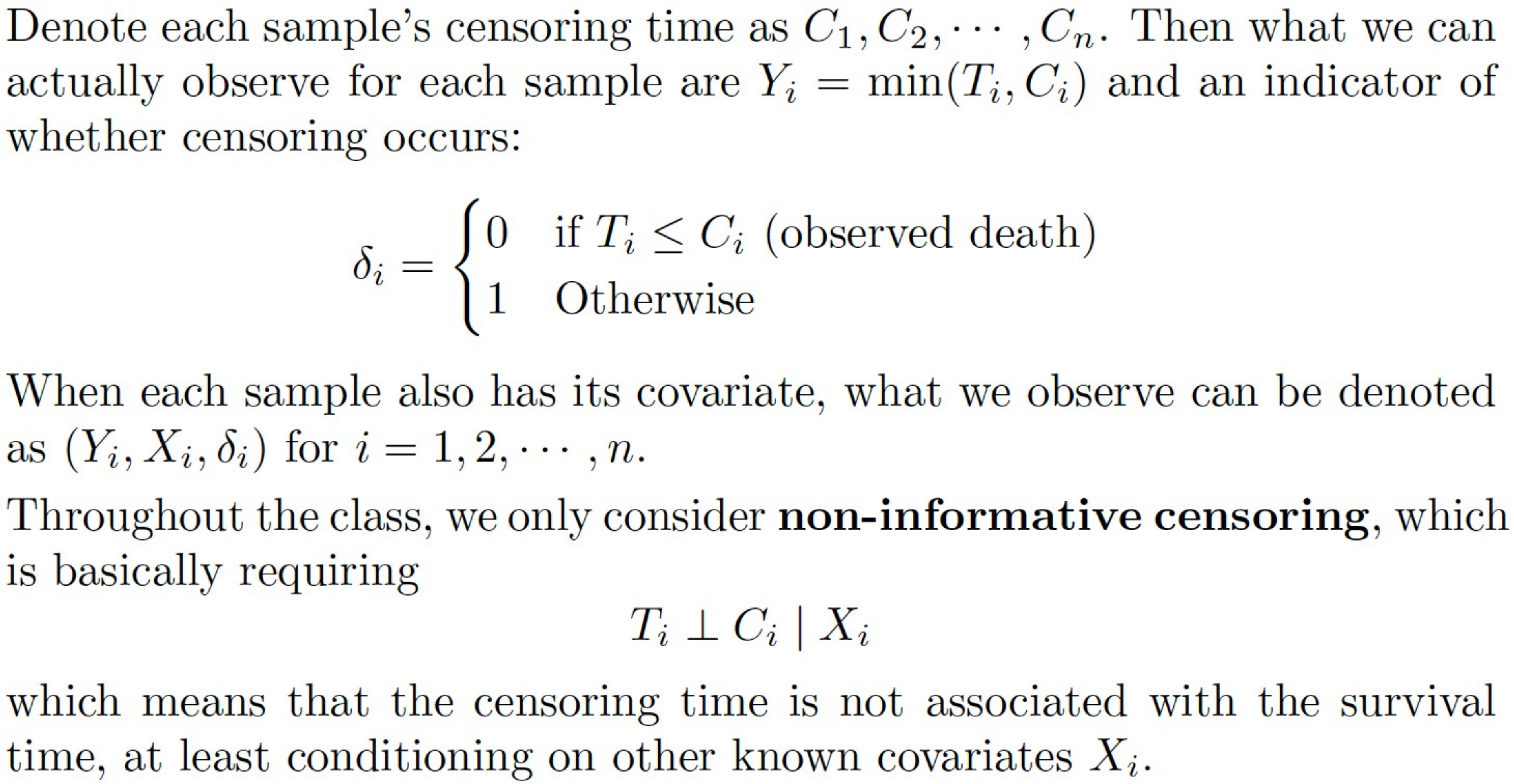
- We may not be able to observe every T_i where i is an individual.

- Censoring can occur when
  - the study ends, some individual have not had the event yet (still alive)
  - Some individuals dropout or get lost in the middle of the study.



- Typically, individuals do not enter the study at the same time
  - Not a concern as T_i is the length of duration
  - can adjust for starting time by add it as a covariate

Concept of censoring



Denote each sample's censoring time as $C_1, C_2, \cdots, C_n$. Then what we can actually observe for each sample are $Y_i = \min(T_i, C_i)$ and an indicator of whether censoring occurs:

$$\delta_i = \begin{cases} 0 & \text{if } T_i \leq C_i \text{ (observed death)} \\ 1 & \text{Otherwise} \end{cases}$$

When each sample also has its covariate, what we observe can be denoted as (Y_i, X_i, δ_i) for $i = 1, 2, \cdots, n$.

Throughout the class, we only consider **non-informative censoring**, which is basically requiring

$$T_i \perp C_i \mid X_i$$

which means that the censoring time is not associated with the survival time, at least conditioning on other known covariates X_i.

Estimating the survival function

- We consider the scenario with no observed covariates X_i and the survival time T_i are i.i.d.

- A non-parametric way with no censoring

$$\widehat{S}_n(t) = \frac{1}{n}\sum_i 1_{T_i > t}$$

  - This does not work if there are censored data
  - Example:
    survival times: 1, 1, 2, 2+, 3+, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
    We don't know how to estimate $S(3)$ from the empirical cdf approach

Kaplan-Meier estimator

- Assume we have discrete time points
- Make use of the equation:

$$S_i = \prod_{j=0}^{i} P[T \geq j+1 \mid T \geq j] = \prod_{j=0}^{i} (1 - h_j)$$

- How to estimate a hazard rate h_i? For time bin i, assume
    - r_i samples that are still alive at the beginning of this time bin
    - d_i death during this time bin
    - c_i drop-outs at the end of this time bin
    - No drop-outs during thins time bin

$$d_i \sim \mathrm{Bernoulli}(r_i, h_i) \qquad \widehat{h}_i = \frac{d_i}{r_i}$$

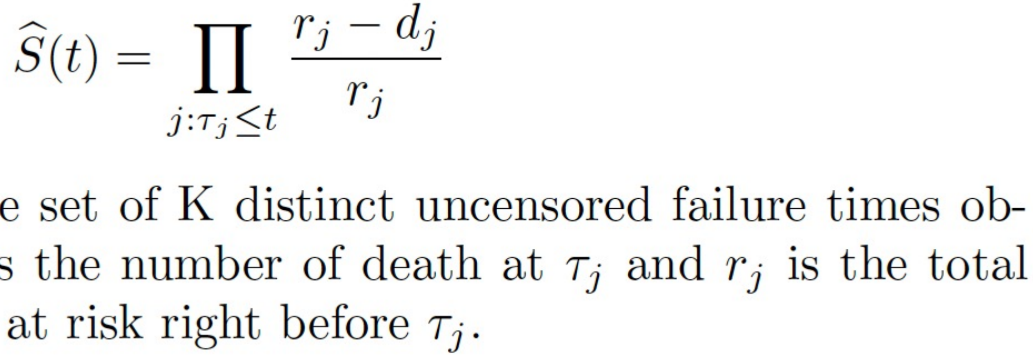
- Kaplan-Meier estimator

$$\widehat{S}_i = \prod_{j \leq i} (1 - \widehat{h}_j)$$

Kaplan-Meier estimator

- For continuous T, we can discretize time into bins and make the bin size smaller and smaller

- The Kaplan-Meier estimator in the limiting case becomes



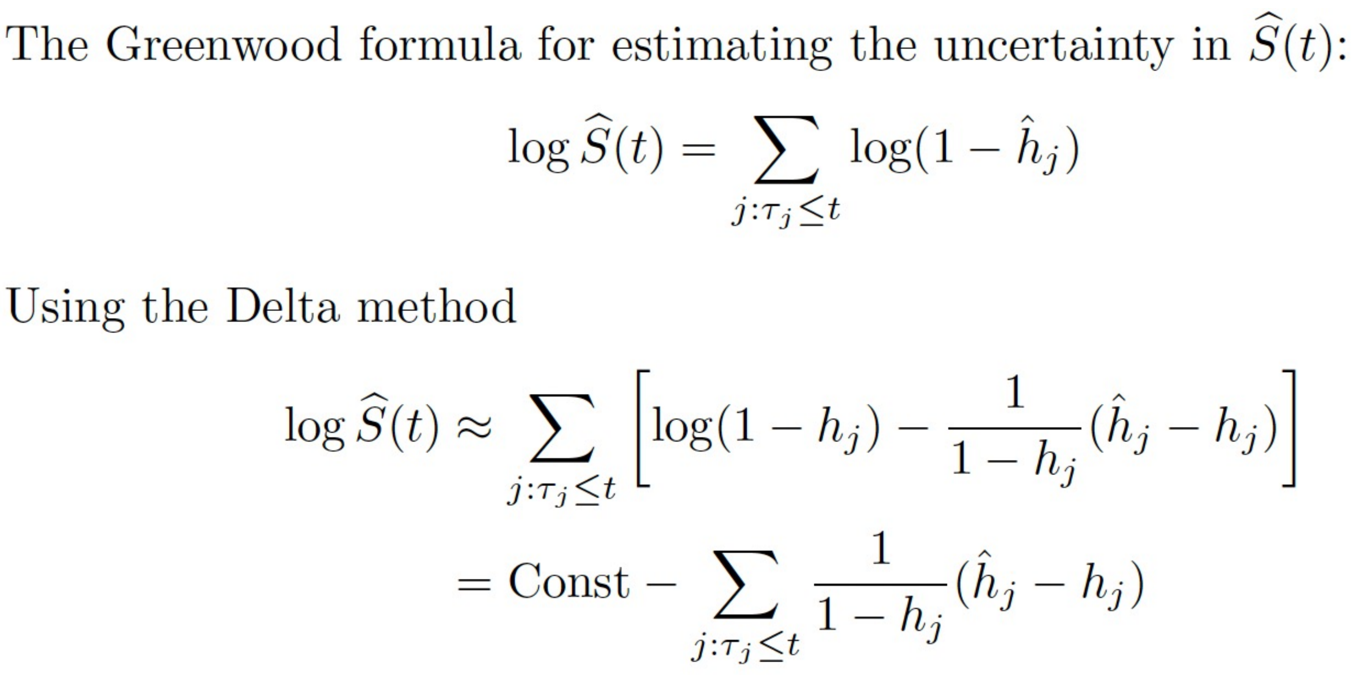
$$\widehat{S}(t) = \prod_{j:\tau_j \leq t} \frac{r_j - d_j}{r_j}$$

where $\{\tau_1, \tau_2, \cdots \tau_K\}$ is the set of K distinct uncensored failure times observed in the sample, d_j is the number of death at τ_j and r_j is the total number of people who are at risk right before τ_j.

- The above formula also works for discrete time points

Variance of $\hat{S}(t)$

- The estimates $\hat{h}_1, \cdots, \hat{h}_K$ are not independent: $r_{j+1} = r_j - d_j - c_j$, $\quad \widehat{h}_i = \dfrac{d_i}{r_i}$



The Greenwood formula for estimating the uncertainty in $\widehat{S}(t)$:

$$\log \widehat{S}(t) = \sum_{j:\tau_j \leq t} \log(1 - \hat{h}_j)$$

Using the Delta method

$$\log \widehat{S}(t) \approx \sum_{j:\tau_j \leq t} \left[\log(1 - h_j) - \frac{1}{1 - h_j}(\hat{h}_j - h_j) \right]$$

$$= \text{Const} - \sum_{j:\tau_j \leq t} \frac{1}{1 - h_j}(\hat{h}_j - h_j)$$

- Though the estimates $\hat{h}_1, \cdots, \hat{h}_K$ are not independent, we always have

$$E\left[\hat{h}_i - h_i \middle| \hat{h}_1, \cdots, \hat{h}_{i-1}\right] = 0$$

  - The partial sums form a martingale
  - $\hat{h}_1, \cdots, \hat{h}_K$ are pairwise uncorrelated

Variance of $\hat{S}(t)$

- When calculating the variance, we can treat $\hat{h}_1, \cdots, \hat{h}_K$ as “independent” and K as fixed.

$$\widehat{\text{Var}}\left(\log \widehat{S}(t)\right) \approx \sum_{j:\tau_j \le t} \left(\frac{1}{1-\widehat{h}_j}\right)^2 \widehat{\text{Var}}(\widehat{h}_j)$$

$$= \sum_{j:\tau_j \le t} \frac{\widehat{h}_j}{(1-\widehat{h}_j)r_j} = \sum_{j:\tau_j \le t} \frac{d_j}{(r_j-d_j)r_j}$$

Using Delta method on $\widehat{S}(t) = e^{\log \widehat{S}(t)}$, we get

$$\widehat{\text{Var}}\left(\widehat{S}(t)\right) = [\widehat{S}(t)]^2 \widehat{\text{Var}}\left(\log(\widehat{S}(t))\right)$$

$$= [\widehat{S}(t)]^2 \sum_{j:\tau_j \le t} \frac{d_j}{(r_j-d_j)r_j}$$

- Under some conditions, we can also have CLT of $\log \hat{S}(t)$

Comparison between two survival survival curves

- In the NCOG data, we may want to know if the whole survival curve of Arm B is significantly larger than the whole curve of Arm A.

- Here, we consider testing for the simple null hypothesis

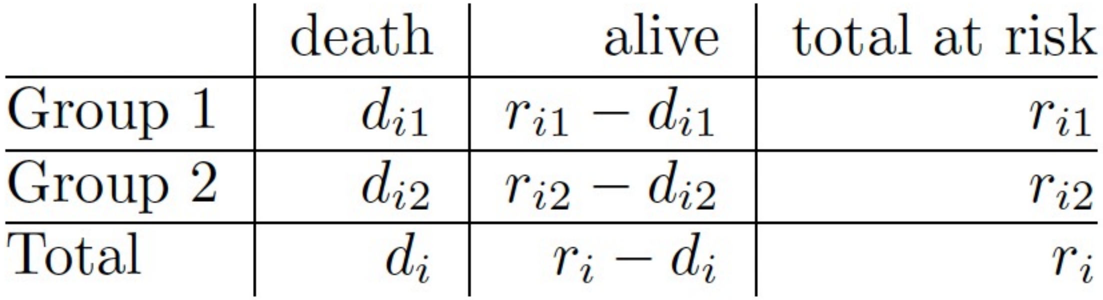
$$H_0 : S_1(t) \equiv S_2(t)$$

This tests if the two curves are exactly the same

The Cochran-Mantel-Haenszel log-rank test

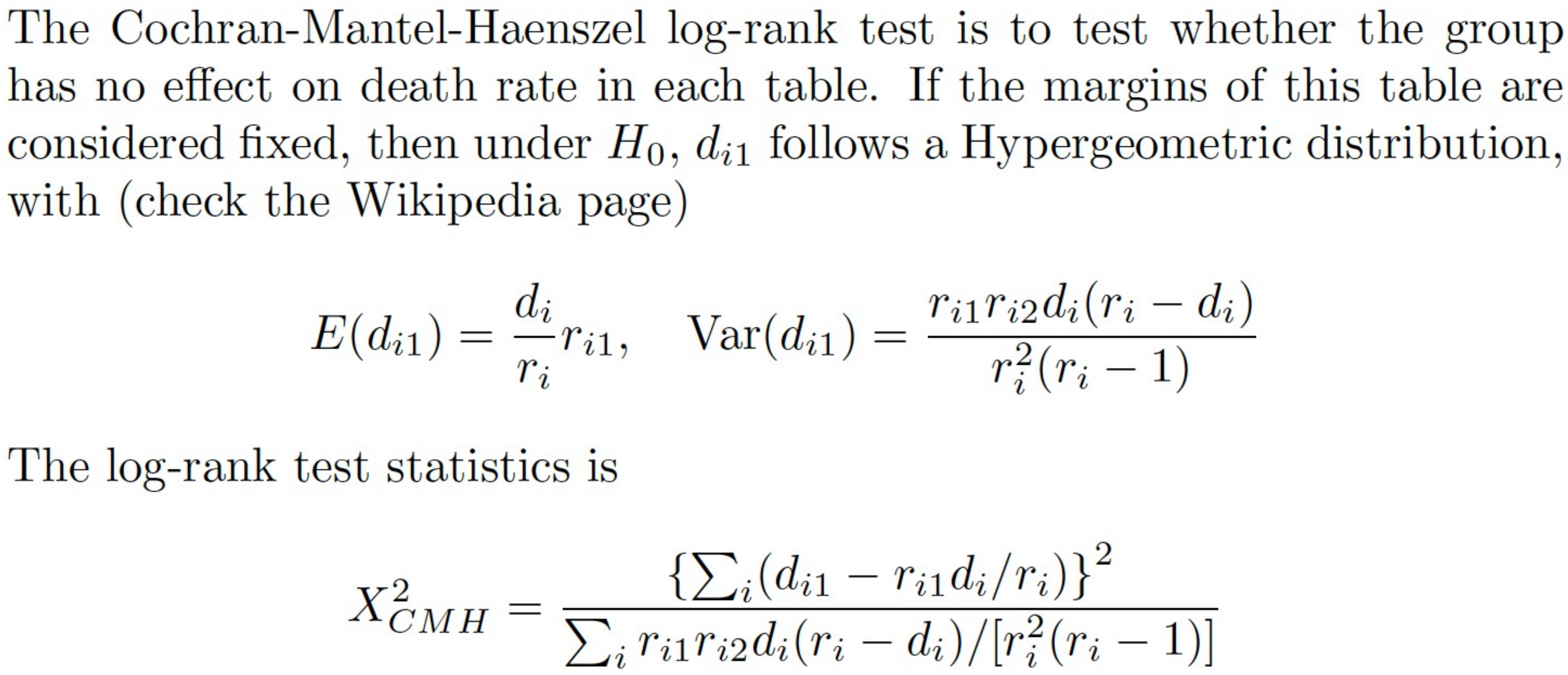
- Assume we have discrete time points

- For each discrete survival time i,
  - We observe r_{i1} and r_{i2} samples that are still alive at the beginning of this time bin for each group respectively
  - Observe d_{i1} and d_{i2} death during this time bin for two groups respectively.
  - Assume that drop-outs happen at the end of each time bin. (so we don’t need to consider it)



	death	alive	total at risk
Group 1	d_{i1}	$r_{i1} - d_{i1}$	r_{i1}
Group 2	d_{i2}	$r_{i2} - d_{i2}$	r_{i2}
Total	d_i	$r_i - d_i$	r_i

The Cochran-Mantel-Haenszel log-rank test



The Cochran-Mantel-Haenszel log-rank test is to test whether the group has no effect on death rate in each table. If the margins of this table are considered fixed, then under H_0, d_{i1} follows a Hypergeometric distribution, with (check the Wikipedia page)

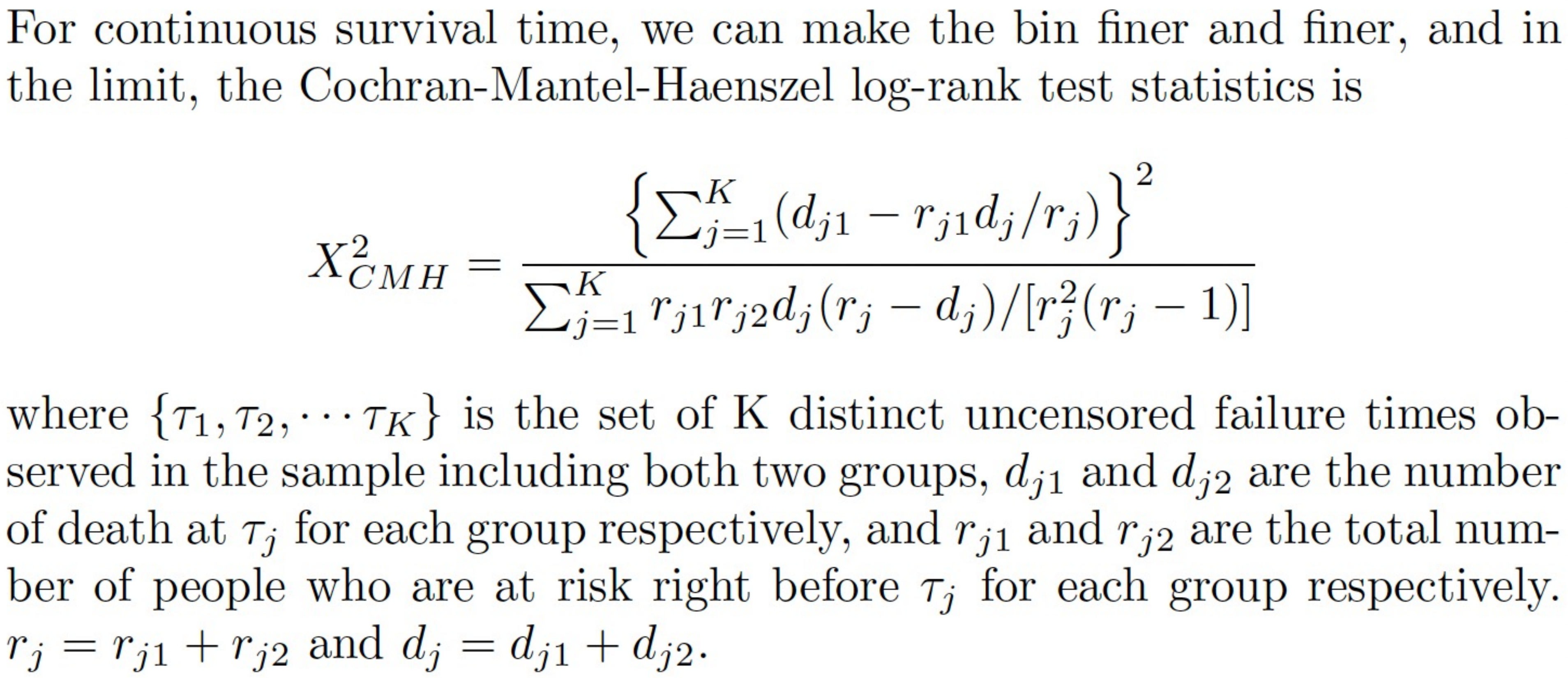
$$E(d_{i1}) = \frac{d_i}{r_i} r_{i1}, \quad \mathrm{Var}(d_{i1}) = \frac{r_{i1} r_{i2} d_i (r_i - d_i)}{r_i^2 (r_i - 1)}$$

The log-rank test statistics is

$$X^2_{CMH} = \frac{\left\{\sum_i (d_{i1} - r_{i1} d_i / r_i)\right\}^2}{\sum_i r_{i1} r_{i2} d_i (r_i - d_i) / [r_i^2 (r_i - 1)]}$$

- Compare X^2_{CMH} with a χ^2_1 distribution to get p-value

The Cochran-Mantel-Haenszel log-rank test

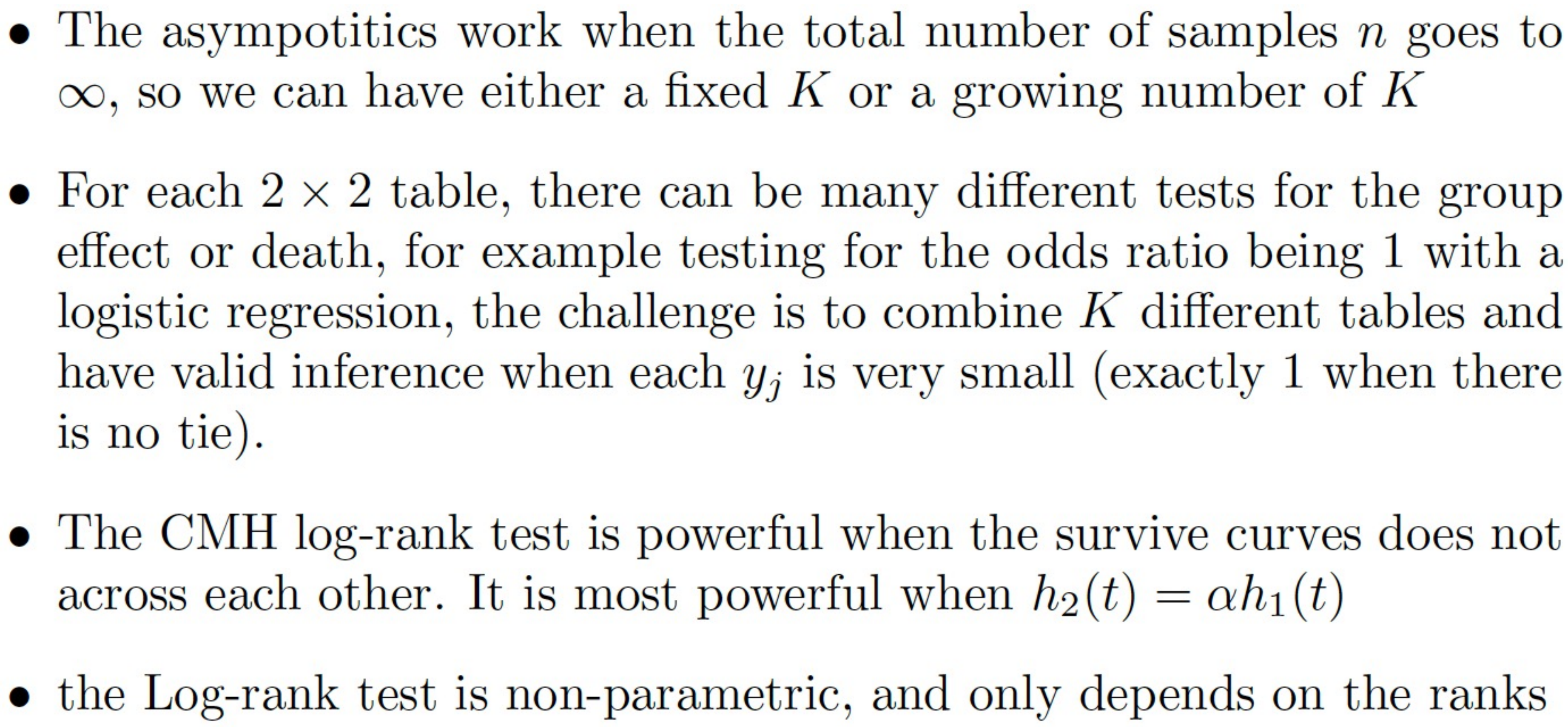


For continuous survival time, we can make the bin finer and finer, and in the limit, the Cochran-Mantel-Haenszel log-rank test statistics is

$$X^2_{CMH} = \frac{\left\{\sum_{j=1}^{K}(d_{j1} - r_{j1}d_j/r_j)\right\}^2}{\sum_{j=1}^{K} r_{j1}r_{j2}d_j(r_j - d_j)/[r_j^2(r_j - 1)]}$$

where $\{\tau_1, \tau_2, \cdots \tau_K\}$ is the set of K distinct uncensored failure times observed in the sample including both two groups, d_{j1} and d_{j2} are the number of death at τ_j for each group respectively, and r_{j1} and r_{j2} are the total number of people who are at risk right before τ_j for each group respectively. $r_j = r_{j1} + r_{j2}$ and $d_j = d_{j1} + d_{j2}$.

Some remarks



- The asympotitics work when the total number of samples n goes to ∞, so we can have either a fixed K or a growing number of K

- For each 2×2 table, there can be many different tests for the group effect or death, for example testing for the odds ratio being 1 with a logistic regression, the challenge is to combine K different tables and have valid inference when each y_j is very small (exactly 1 when there is no tie).

- The CMH log-rank test is powerful when the survive curves does not across each other. It is most powerful when $h_2(t) = \alpha h_1(t)$

- the Log-rank test is non-parametric, and only depends on the ranks

Data example

- Check Example10 R notebook