Censoring mechanisms

Patrick Breheny

August 31

Fixed vs. random censoring

- In the previous lecture, we derived the contribution to the likelihood from fully observed, censored, and truncated observations under a variety of situations (left/right/interval)
- We didn't explicitly state it as an assumption, but our derivations treated the censoring time c_i for ith individual as a fixed quantity, known in advance
- In most real settings, however, censoring times are random variables, not fixed constants

Fixed vs. random censoring (cont'd)

- Even in settings where the date of censoring is fixed in advance, the time on study until censoring is random because it depends on when the subject entered the study
- To realistically account for the effect of censoring on inference, then, we must consider censoring as a random variable, meaning that we must consider the distributions of two random variables (time to event and time to censoring) and how they relate to one another

Random censorship model

- Let $\widetilde{T}_i$ denote the true survival time and C_i denote the censoring time, with

$$\widetilde{T}_i | x_i \overset{\perp\!\!\!\perp}{\sim} S(\theta, x_i)$$
$$C_i | x_i \overset{\perp\!\!\!\perp}{\sim} G(\eta, x_i)$$
$$\widetilde{T}_i \perp\!\!\!\perp C_i | x_i,$$

  where θ is the parameter(s) of interest and x_i is a potentially vector-valued covariate
- We observe $\{t_i, d_i, x_i\}_{i=1}^n$, where

$$t_i = \widetilde{t}_i \wedge c_i = \min(\widetilde{t}_i, c_i)$$
$$d_i = 1\{\widetilde{t}_i \leq c_i\}$$

Random censoring: Likelihood

- Under the assumptions on the previous slide, which collectively are known as *random censoring*, each observation is independent and therefore each observation makes an independent contribution $L_i(\theta)$ to the likelihood:

$$\text{for } T_i = t, D_i = 1: \; L_i(\theta) = f(t|x_i, \theta)G(t|x_i, \eta)$$
$$\text{for } T_i = t, D_i = 0: \; L_i(\theta) = g(t|x_i, \eta)S(t|x_i, \theta)$$

- Thus,

$$\begin{aligned} L(\theta) &= \prod_i L_i(\theta) \\ &\propto \prod_i f(t_i|x_i, \theta)^{d_i} S(t_i|x_i, \theta)^{1-d_i} \end{aligned}$$

Remarks on random censoring

- Thus, we arrive at the same likelihood we derived previously in the case of fixed censoring: under the random censorship model, the likelihood contributions from censoring amount to a constant with respect to θ and do not affect inference
- This is very convenient since, in practice, one generally does not care about the censoring mechanism or wish to spend any effort modeling it

Random entry

- It is worth noting that the random censorship model covers the important special case of random entry into the study, with a fixed censoring date at the end of the study:

$$\text{Entry date: } A_i \overset{\perp\!\!\!\perp}{\sim} G'(\eta)$$
$$\text{Censoring date: } C_i' = c$$
$$\text{True failure date: } \widetilde{T}_i' = A_i + \widetilde{T}_i,$$

  where $\widetilde{T}_i \sim S(\theta)$ is the true failure time, as discussed earlier
- Thus, in terms of time on study, under the assumption $A_i \perp\!\!\!\perp \widetilde{T}_i$, we have

$$\widetilde{T}_i \sim S(\theta)$$
$$C_i \sim G(\eta)$$
$$\widetilde{T}_i \perp\!\!\!\perp C_i$$

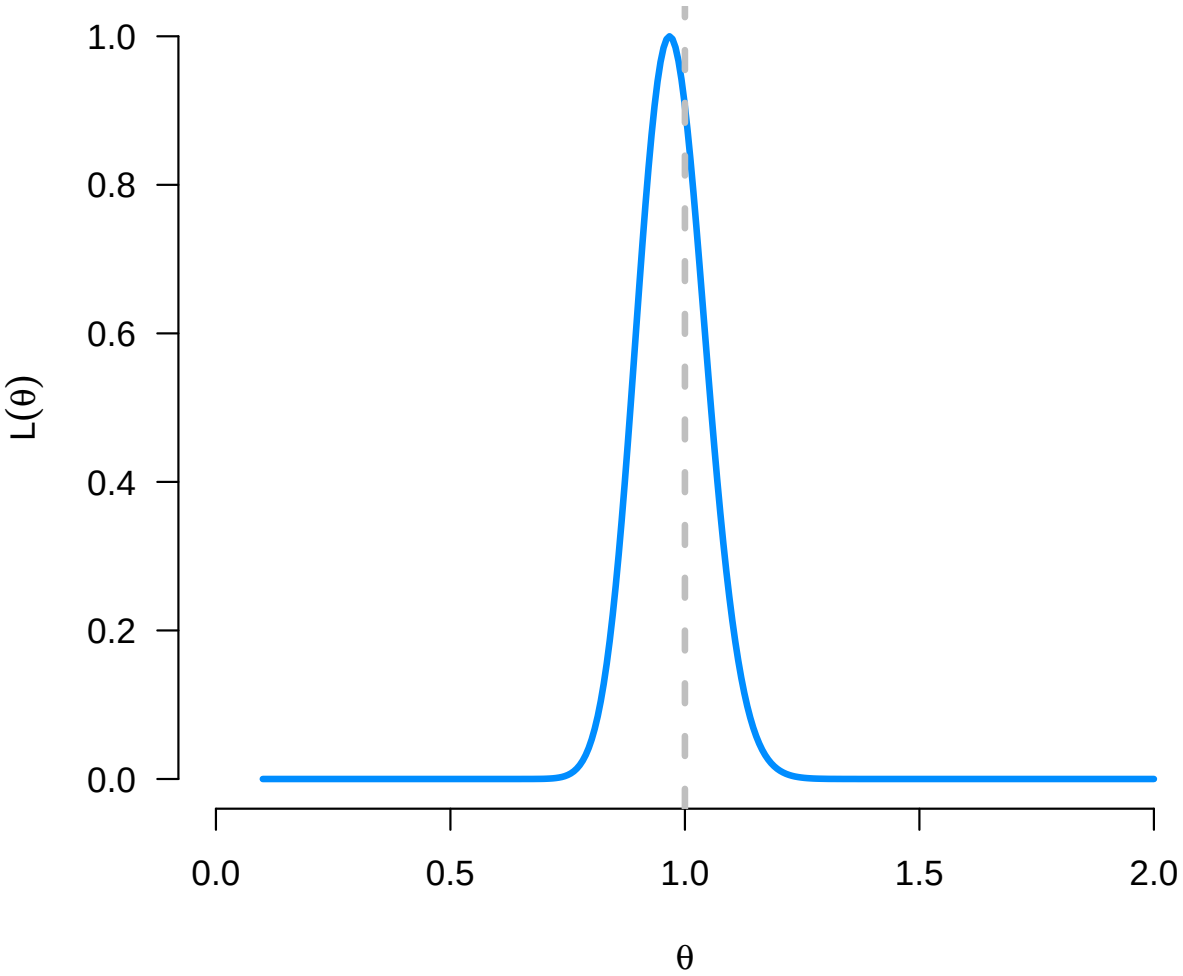
Random censoring: Example

- To illustrate the assumptions of the random censorship model and the bias that can result when those assumptions are violated, we'll now go through a few examples
- The examples will generally follow the random entry scenario just described, in which subjects enter the model randomly over the interval $(0, 1)$ and then the analysis is carried out at time $t' = 1$
- To begin with, let's satisfy ourselves that what we just derived does, in fact, work:

$$\widetilde{T}_i \overset{\perp\!\!\!\perp}{\sim} \text{Exp}(1)$$
$$A_i \overset{\perp\!\!\!\perp}{\sim} \text{Unif}(0, 1),$$

so that $\widetilde{T}_i \perp\!\!\!\perp C_i$

Random censoring: Simulated results ($n = 500$)



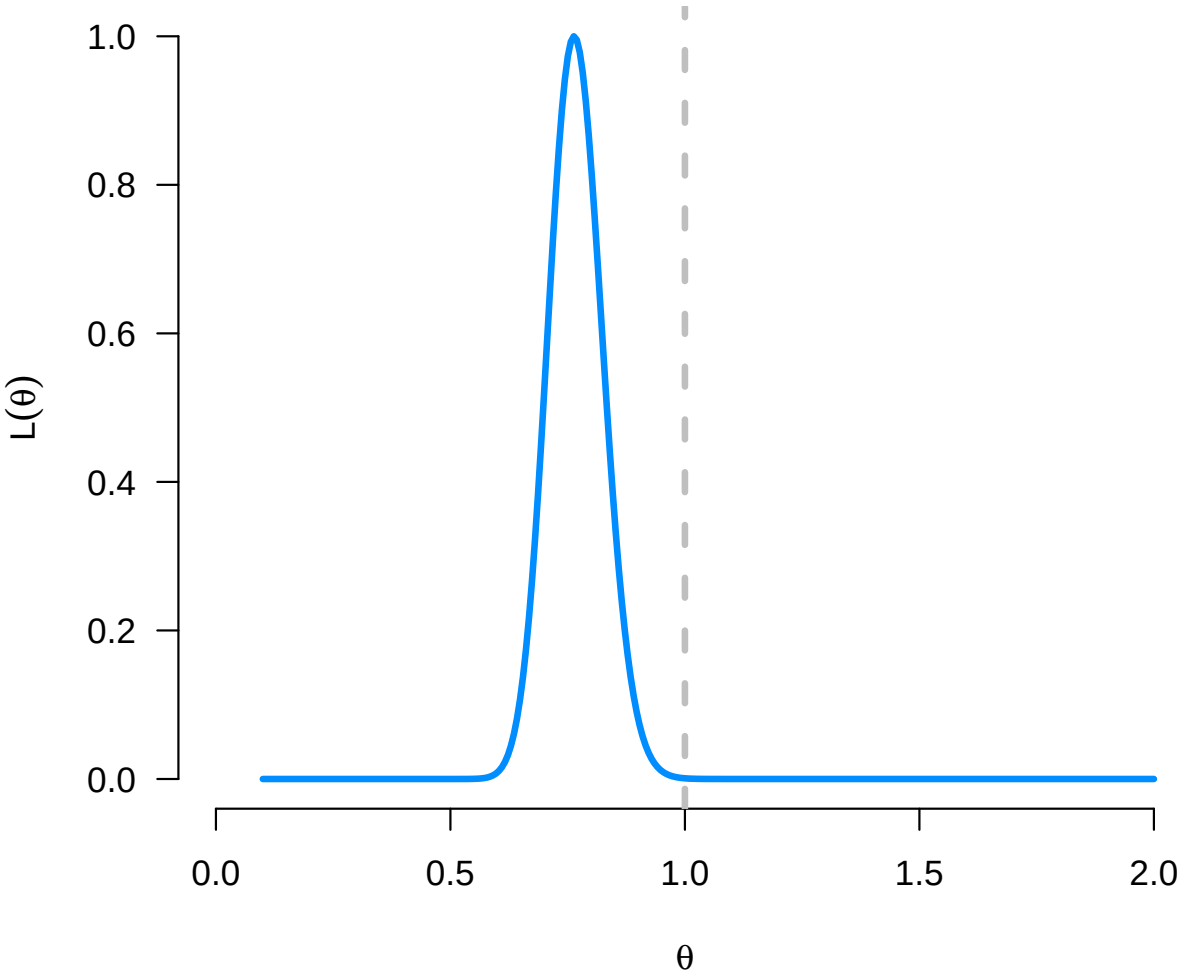
Example: Entry and failure dependent

- Now, let's introduce some dependence between the entry time and the failure time, so that the random censorship assumption of $\widetilde{T}_i \perp\!\!\!\perp C_i$ is violated
- In particular, let's let $\widetilde{T}_i$ follow an $\text{Exp}(1)$ distribution as before, but

$$A_i \sim \text{Unif}(0, 1) \qquad \text{if } \widetilde{T} < 1$$
$$A_i \sim \text{Unif}(0, \tfrac{1}{2}) \qquad \text{if } \widetilde{T} \geq 1$$

- Thus, there is a systematic bias in which patients who live longer tend to enter the study earlier

Dependent entry and failure ($n = 500$)



Example: Conditional independence

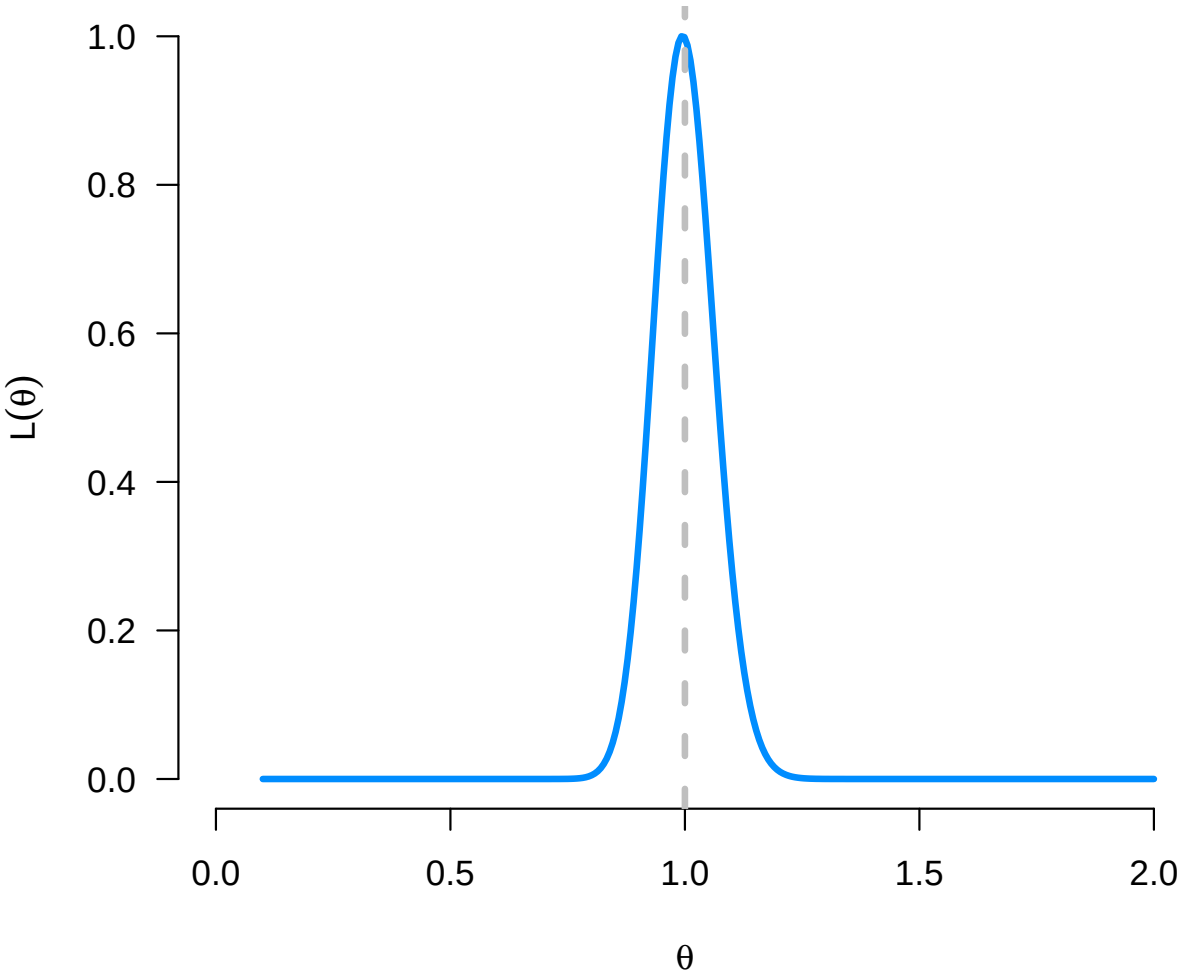
- Note, however, that C_i and $\widetilde{T}_i$ do not have to be strictly independent; they can be *conditionally independent* given x_i
- To see how this distinction matters, let's introduce a covariate such that C_i and $\widetilde{T}_i$ are marginally dependent, but conditionally independent given x_i:

$$\begin{aligned} X_i &\sim \text{Bern}(\tfrac{1}{2}) \\ \widetilde{T}_i | x_i = 1 &\sim \text{Exp}(\tfrac{2}{5}\theta) \\ \widetilde{T}_i | x_i = 0 &\sim \text{Exp}(2\theta) \\ A_i | x_i = 1 &\sim \text{Unif}(0, \tfrac{1}{2}) \\ A_i | x_i = 0 &\sim \text{Unif}(0, 1) \end{aligned}$$

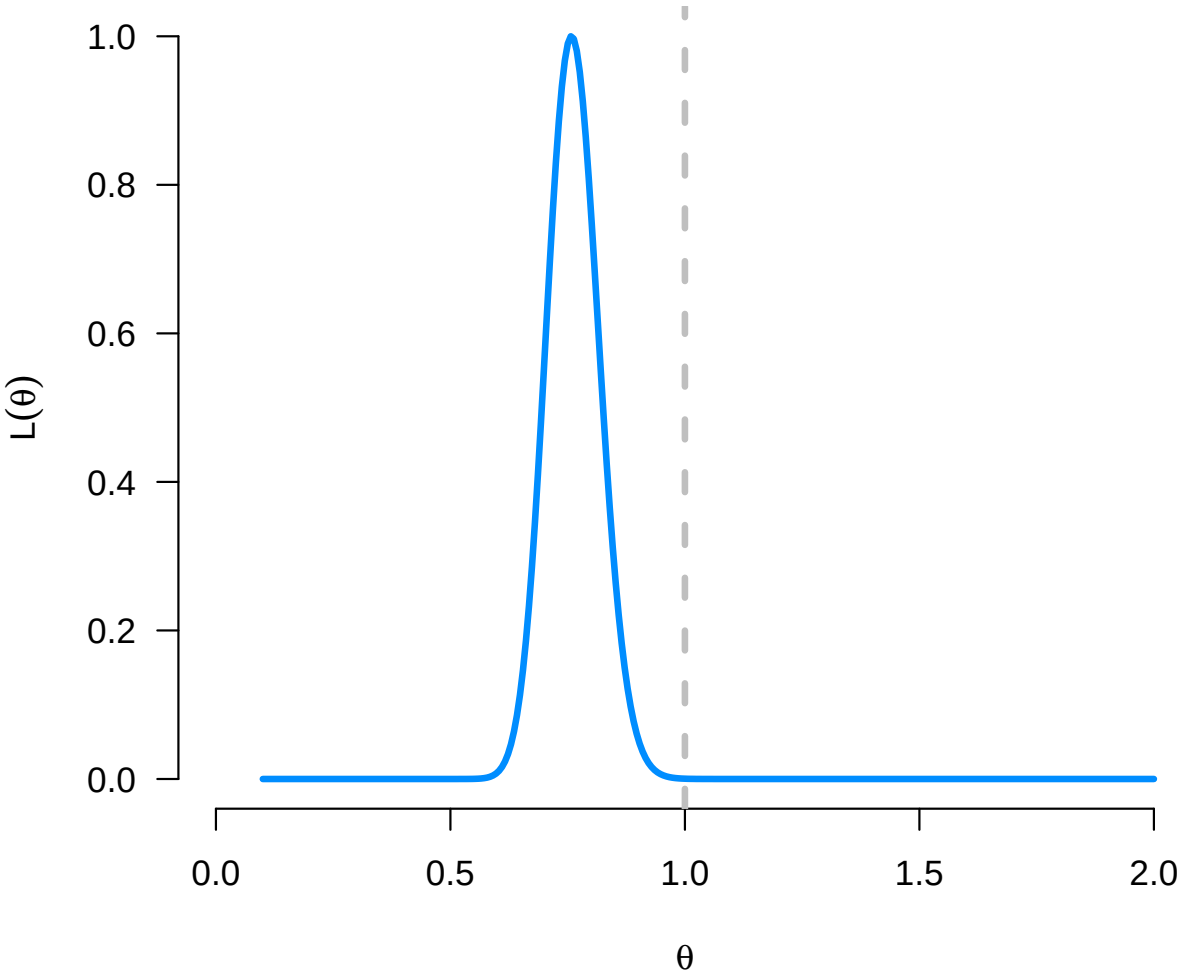
Conditional independence: Remarks

- Note that, as in the first example, there is a systematic bias in which patients who live longer tend to enter the study earlier
- In this example, however, there is a covariate, x_i, that can explain and account for this phenomenon
- Let's see what happens to the likelihood in this second example when we use the information from the covariate, and also what happens when we ignore x_i (although admittedly, this isn't a perfect comparison, as we'll discuss in class)

Using the covariate ($n = 500$)



Ignoring the covariate ($n = 500$)



Generality of the RC likelihood

- We have seen that the likelihood

$$L(\theta) = \prod_i f(t_i|x_i,\theta)^{d_i} S(t_i|x_i,\theta)^{1-d_i}$$

  is correct under the random censorship model
- However, it is also the correct likelihood in many other situations that do not fall under random censorship
- In other words, random censorship is a sufficient condition for the above likelihood, but not a necessary one

Type II censoring

- For example, suppose we enrolled n subjects in a study and then continued the study until a prespecified number d of events occurred (this is known as "Type II censoring")
- This is clearly outside the earlier framework; in particular, the censoring times depend on the failure times for other subjects
- Nevertheless, it can be shown that one still arrives at the same likelihood as under the random censorship model

Independent censoring

- Censoring mechanisms for which the aforementioned likelihood remains correct are called *independent censoring* mechanisms; random censoring is a special case of independent censoring
- A detailed description of the conditions under which independent censoring holds is beyond the scope of this course, but the general idea to consider the likelihood as being built up by a collection of stochastic processes unfolding in time, leading to

$$L(\theta) = \left\{ \prod_i \lambda(t_i|\theta, x_i)^{d_i} \right\} \exp\left\{ -\int_0^\infty \sum_{j \in R(u)} \lambda(u|\theta, x_j)\, du \right\},$$

where $R(u)$ is the *risk set* at time u, consisting of all the individuals still alive and uncensored at time u

Noninformative censoring

- Finally, let us briefly revisit the assumption of the random censorship model that $C_i \sim G(\eta, x_i)$
- Specifically, we are assuming here that the distribution of the censoring time does not depend on θ, the parameter of interest
- This assumption is referred to as *noninformative censoring*; in other words, that the censoring mechanism does not contain any information about the parameter we are studying

Violations of noninformative censoring

- We have already examined the consequences when the assumption that $C_i \perp\!\!\!\perp \widetilde{T}_i$ is violated; what if the assumption of noninformative censoring is violated?
- Generally speaking, if the other assumptions of the random censorship model hold, informative censoring does not necessarily introduce any bias
- Instead, its main consequence is a loss of efficiency with regard to the estimation of θ

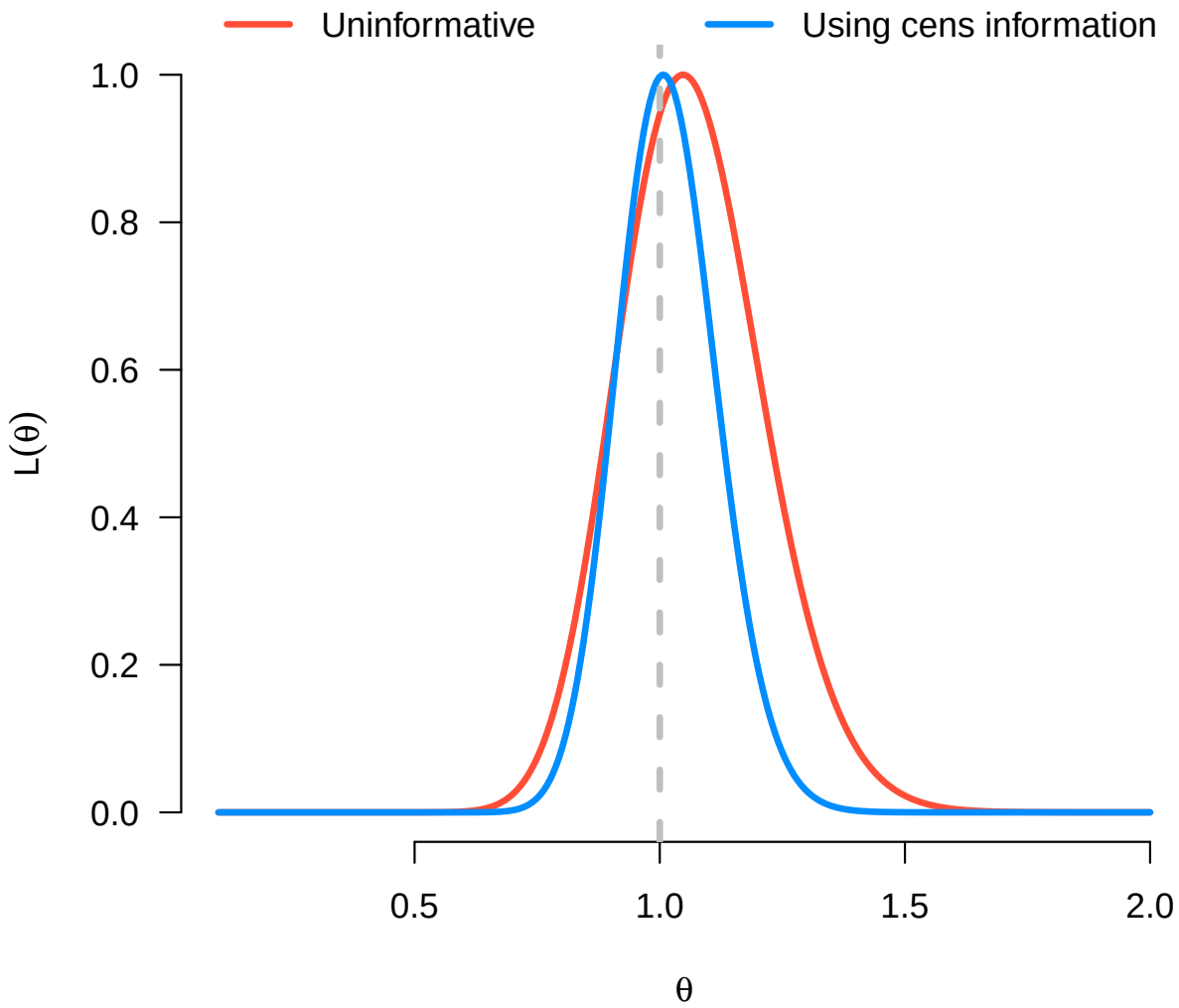
Informative censoring: Example

- As an example of informative censoring, suppose we have:

$$\widetilde{T}_i \sim \text{Exp}(\theta)$$
$$C_i \sim \text{Exp}(\theta)$$
$$\widetilde{T}_i \perp\!\!\!\perp C_i$$

- In this case, as you showed in homework, $T_i \sim \text{Exp}(2\theta)$ and all observations provide the same amount of information about θ
- Let's compare this likelihood with the likelihood we get without making any assumptions concerning G (i.e., assuming random censorship)

Informative censoring ($n = 100$)



Final remarks

- This week, we addressed the question of constructing likelihoods in the presence of censoring (and truncation), and examined various censoring mechanisms with respect to how they influence the likelihood and potentially introduce bias
- Our illustrations have come from the exponential distribution so far out of convenience, but the main points have broad applicability
- Next week, we will discuss nonparametric approaches to estimating survival distributions