Survival Analysis Workshop

Elizabeth M. Sweeney, PhD
Assistant Professor
Weill Cornell

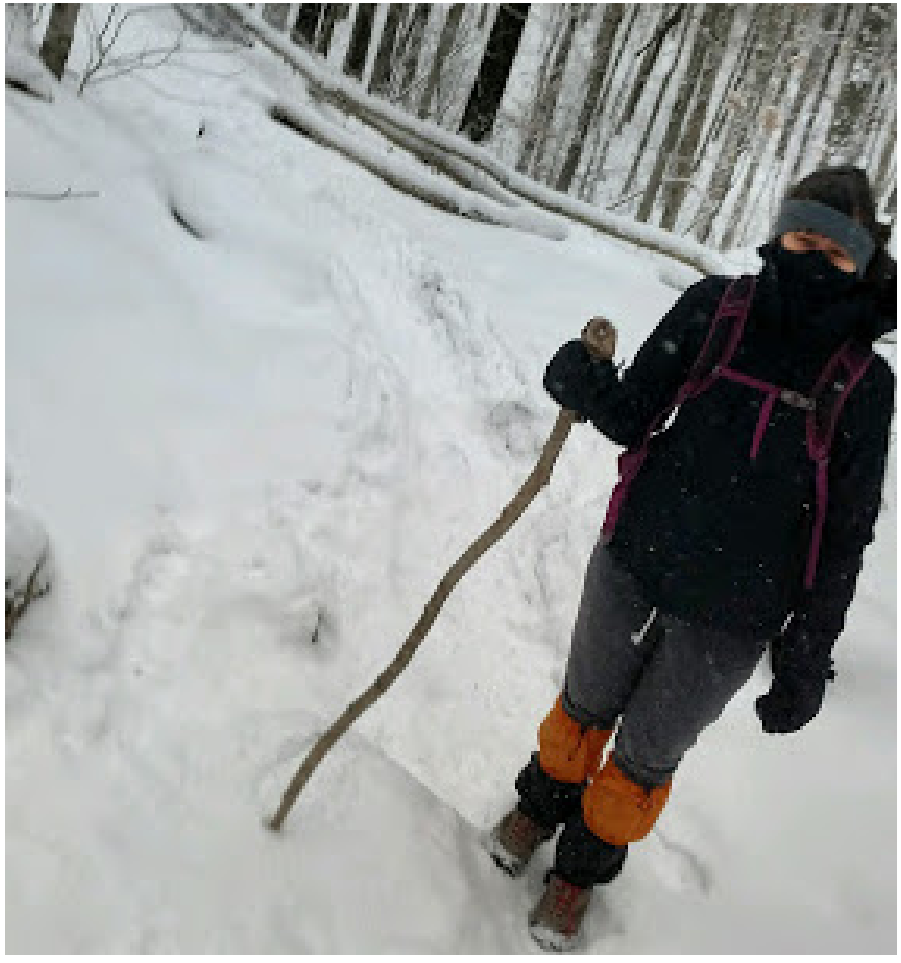
elizabethmargaretsweeney@gmail.com
@emsweene57

May 9, 2019

Meet the Instructor

- PhD in Biostatisitcs from Johns Hopkins Bloomberg School of Public Health
- Industry Data Scientist at Flatiron Health and Covera Health
- Lecturer in Biostatistics at Columbia University
- Assistant Professor at Weill Cornell (for 9 days!)

Meet the Instructor



- I bike commute everywhere
- I once biked over 100 miles in a single day from Brooklyn to the Hamptons
- I am 8 mountains away from finishing the Catskills 3500 club

Agenda

1. Lecture
   - Slides
   - R Code
   - In-Class Exercises
2. In-Workshop Assignment

- Introduction to survival time data
- The survival function
- Cox proportional hazards regression
- Competing risks methods

- **Introduction to survival time data**
  - Types of censoring
  - Components of survival data
  - Dealing with dates in R
  - Introduction to data examples
- The survival function
- Cox proportional hazards regression
- Competing risks methods

Survival analysis is a branch of statistics for analyzing the expected duration of time until one or more events happen, such as death in biological organisms and failure in mechanical systems.

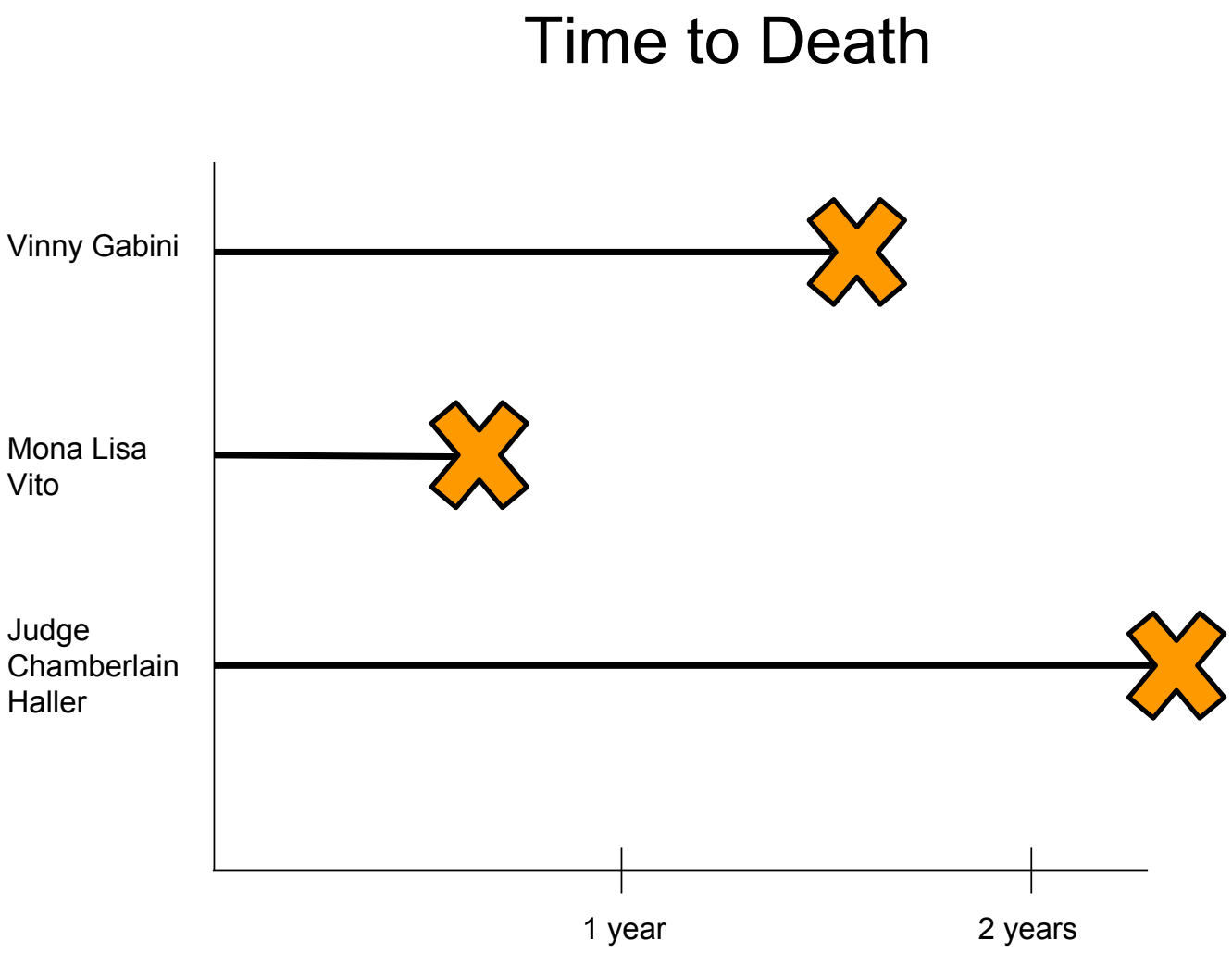
These data can have censoring. Censoring occurs when the value of a measurement is only partial known. Data with censoring requires special statistical techniques.

Example: Death Times of Kidney Patients

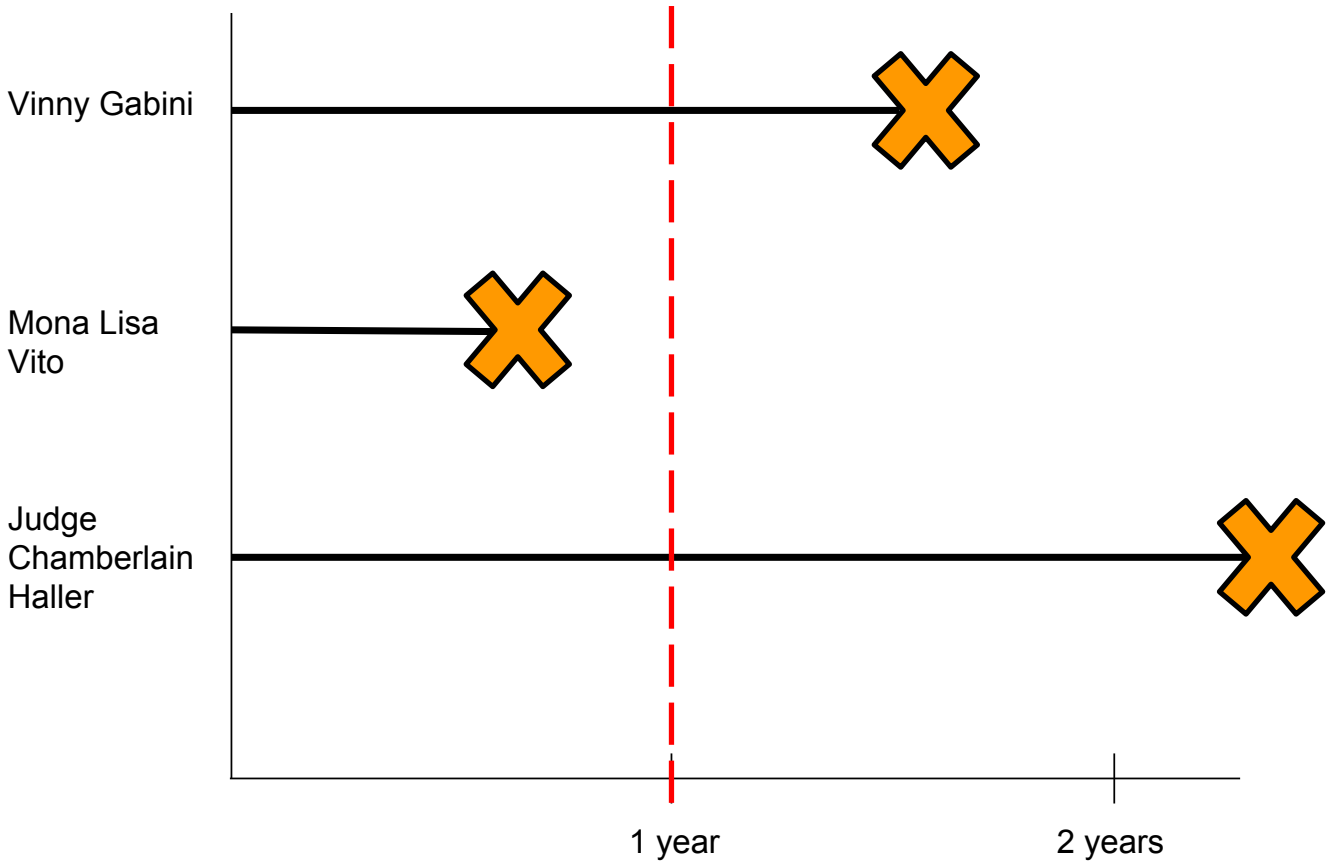
Data on the time of death of patients who had a kidney transplant at The Ohio State University Transplant Center during the period of 1982 to 1992. Patients were censored if they moved from Columbus (lost-to follow-up) or if they were alive on June 30, 1992.

Censoring occurs when the value of a measurement is only partial known. There are three types of censoring:

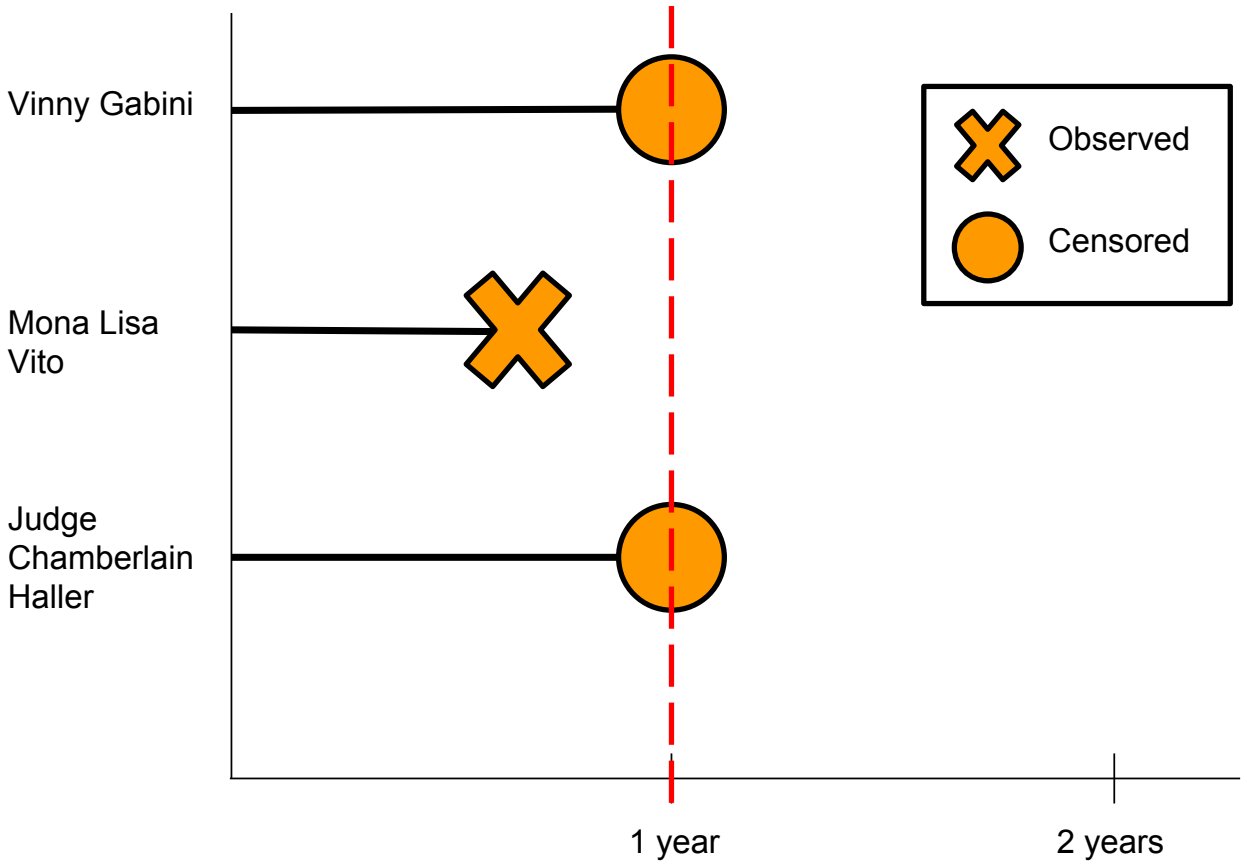
- Right censoring
- Left censoring
- Interval censoring


Time to Death
Vinny Gabini
Mona Lisa
Vito
Judge
Chamberlain
Haller
1 year
2 years

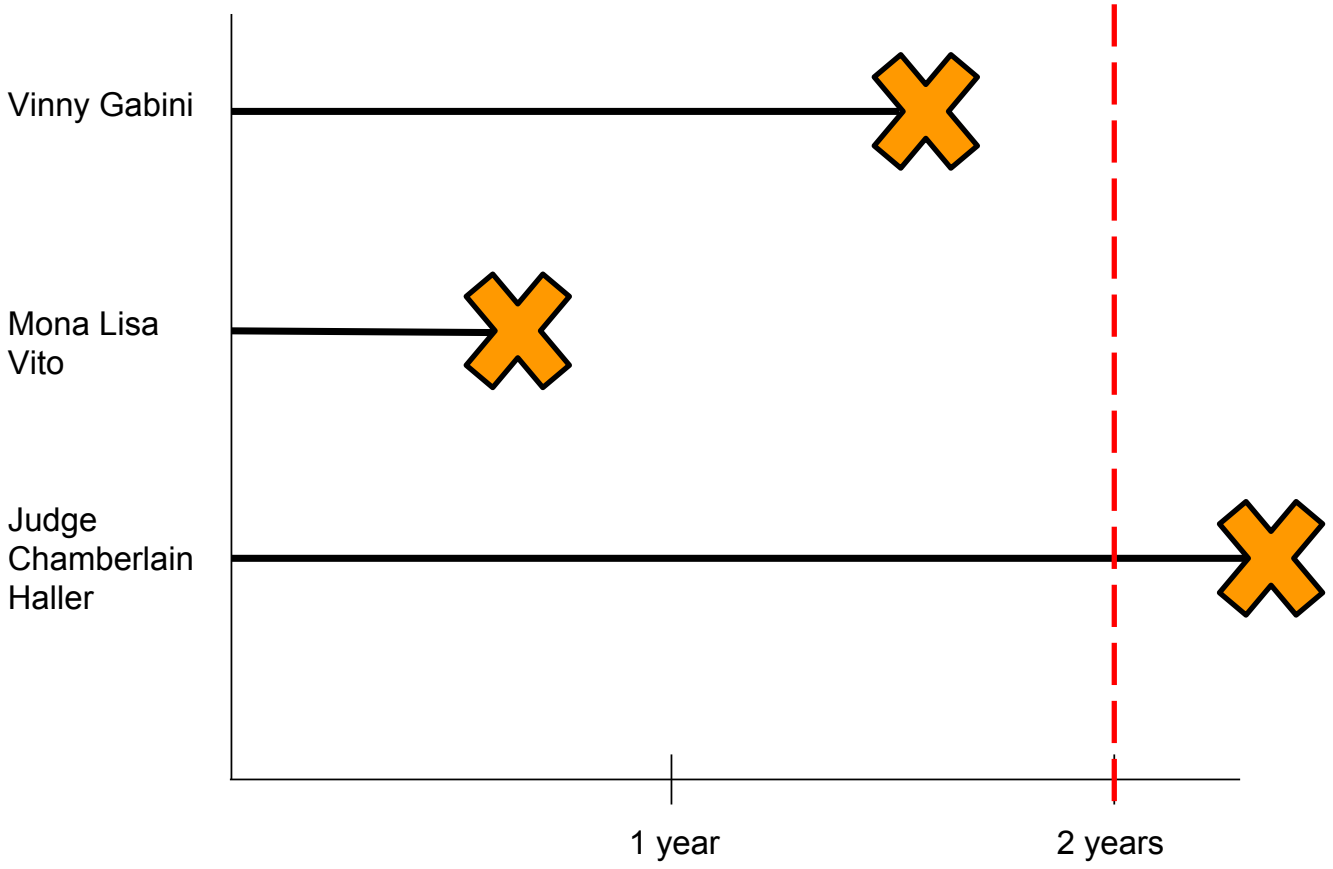
Right Censoring: 1 Year of Follow Up



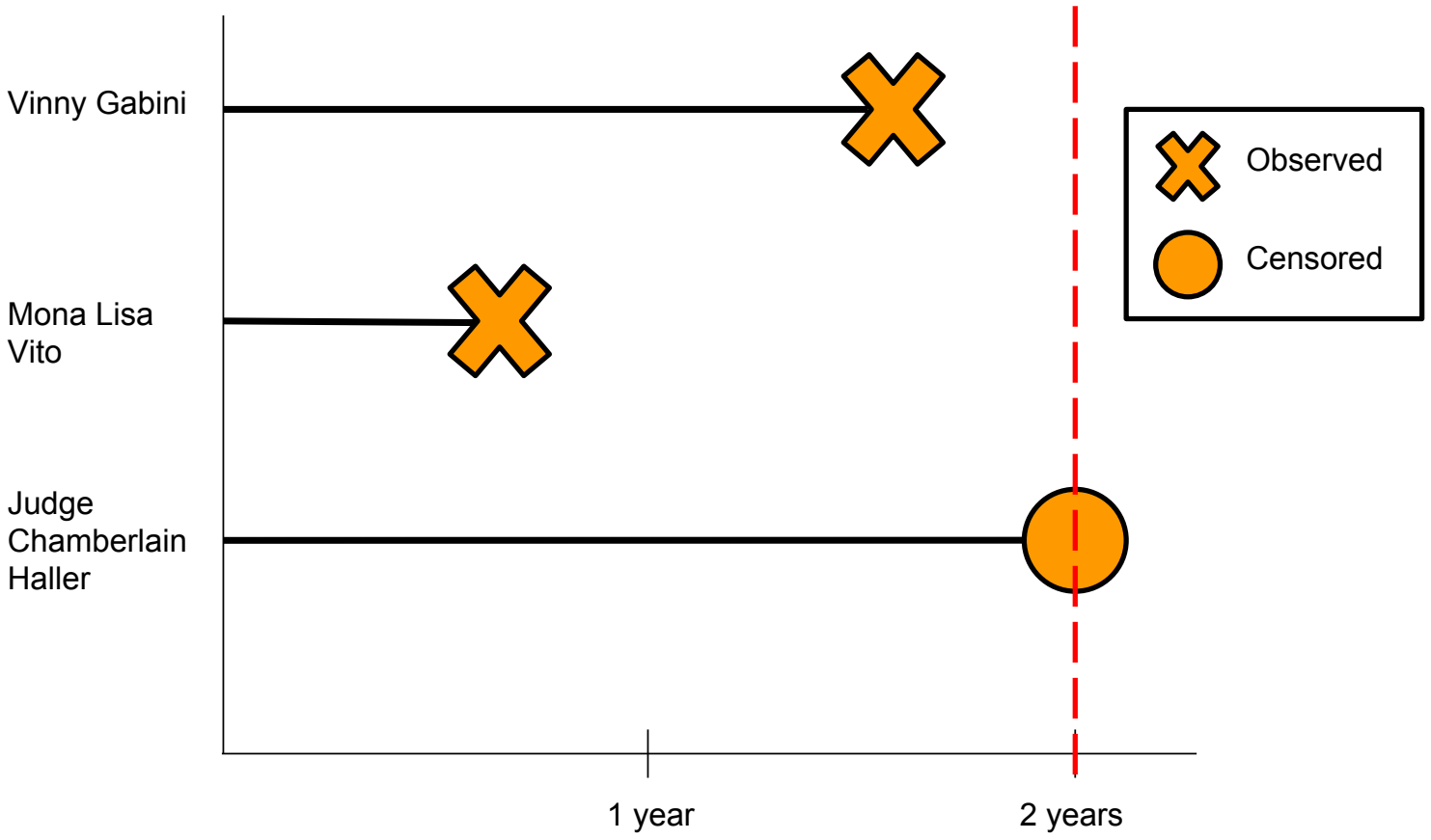
Right Censoring: 1 Year of Follow Up



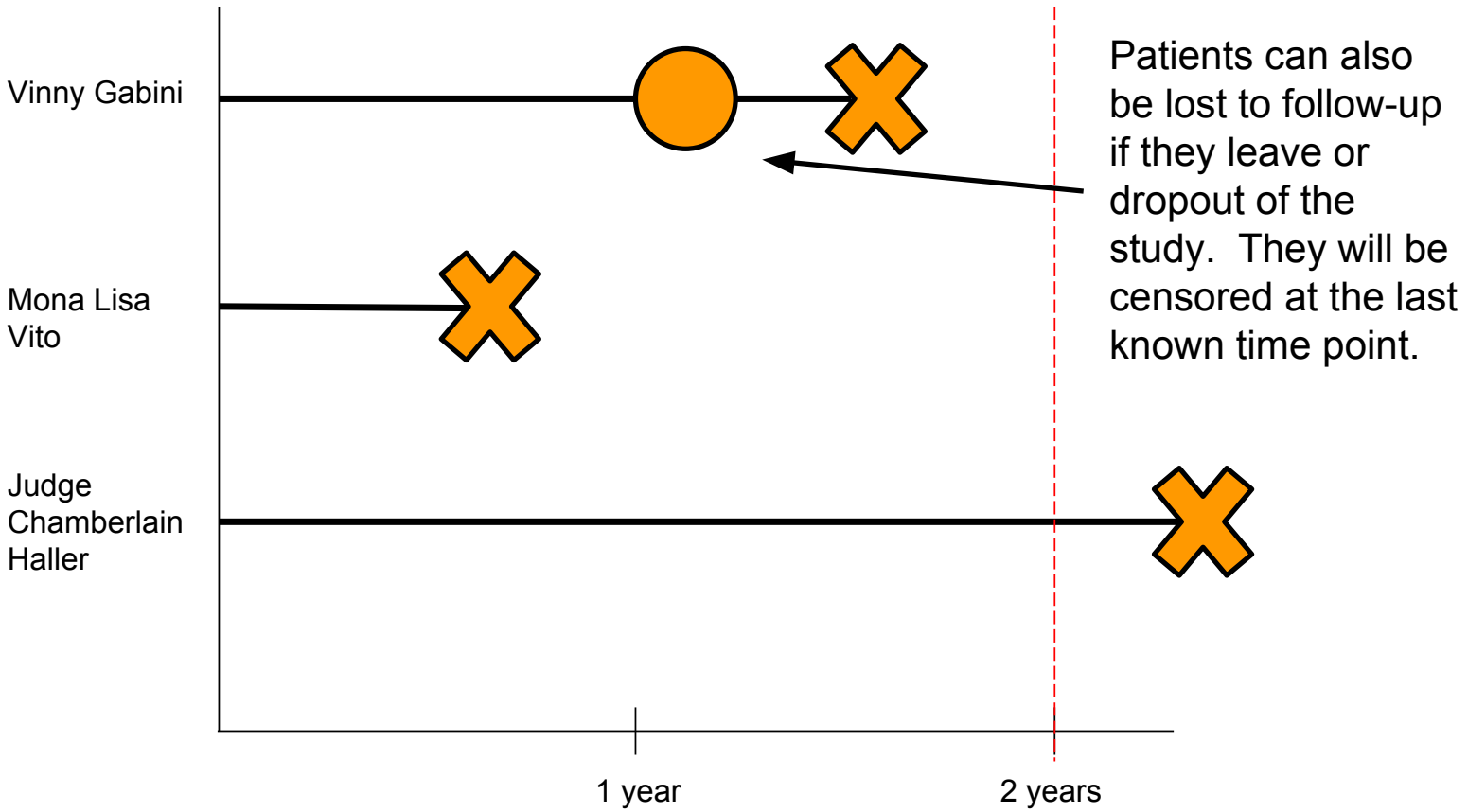
Right Censoring: 2 Years of Follow Up



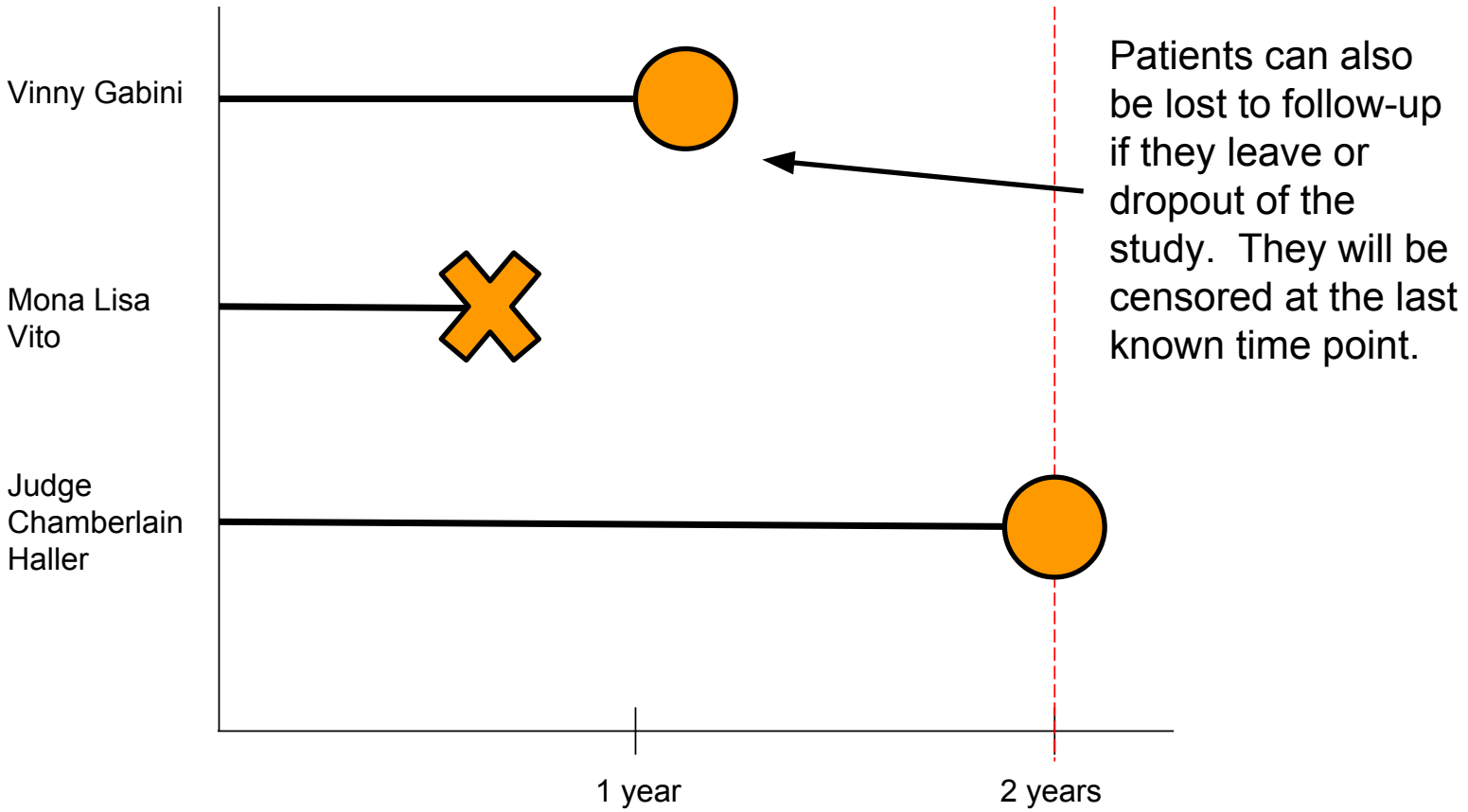
Right Censoring: 2 Years of Follow Up



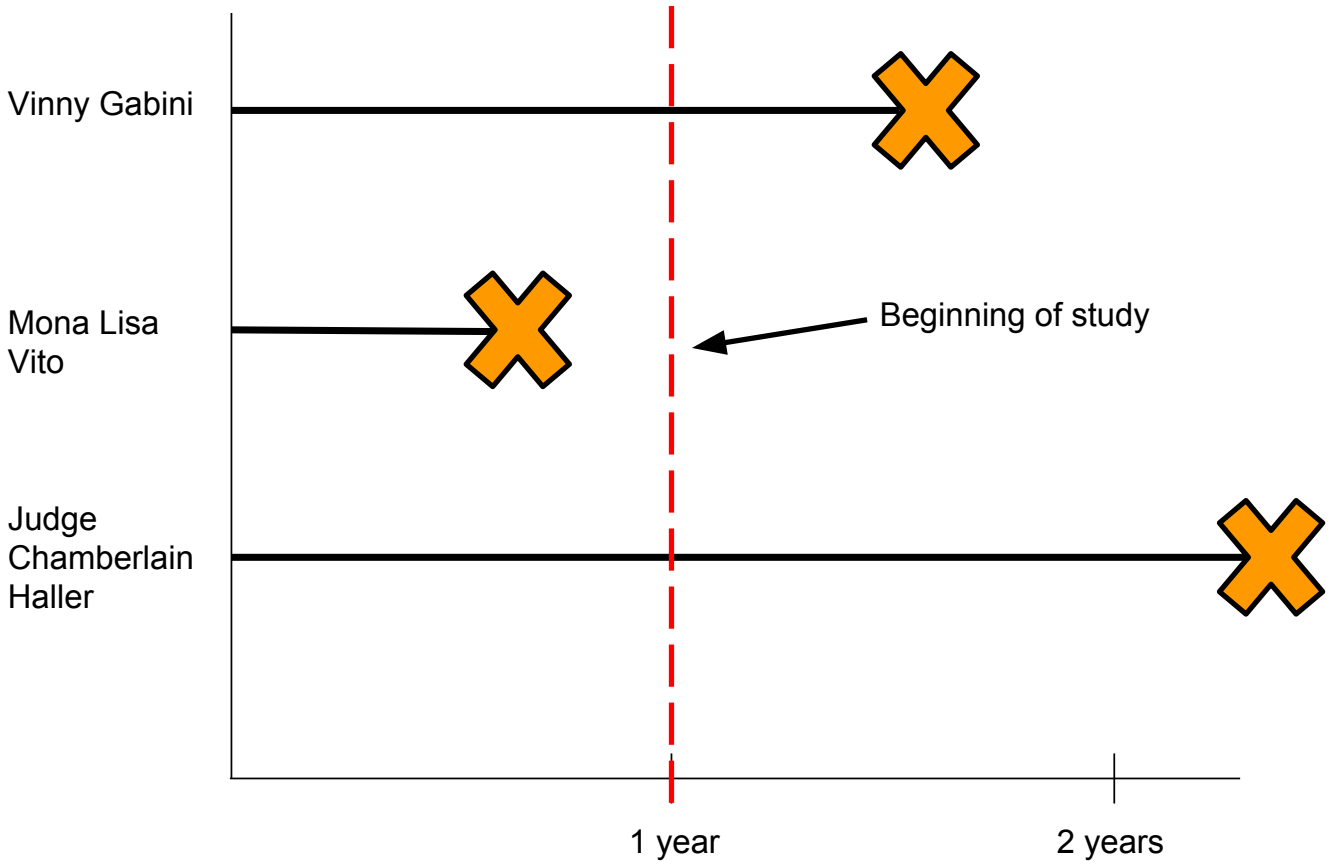
Right Censoring: 2 Years of Follow Up



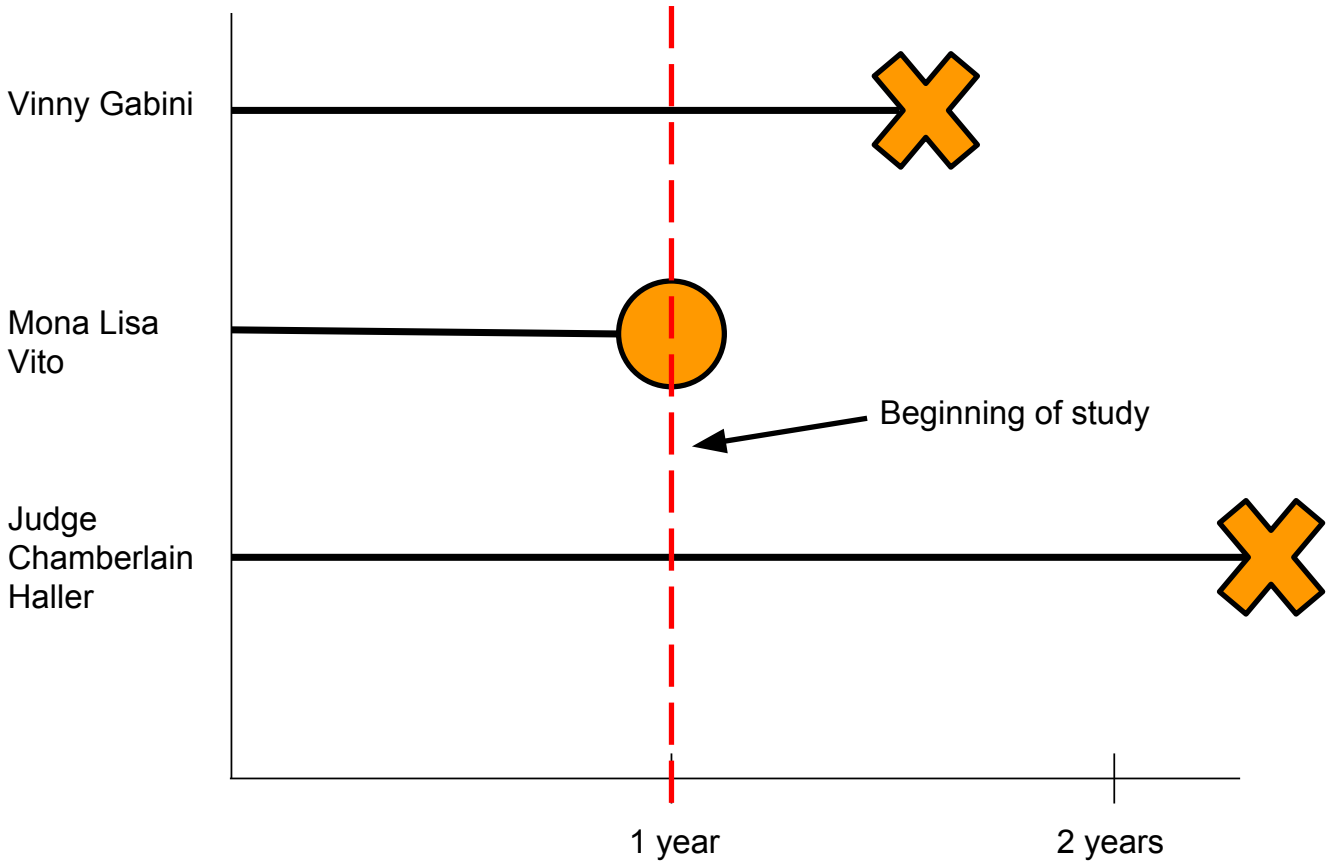
Right Censoring: 2 Years of Follow Up

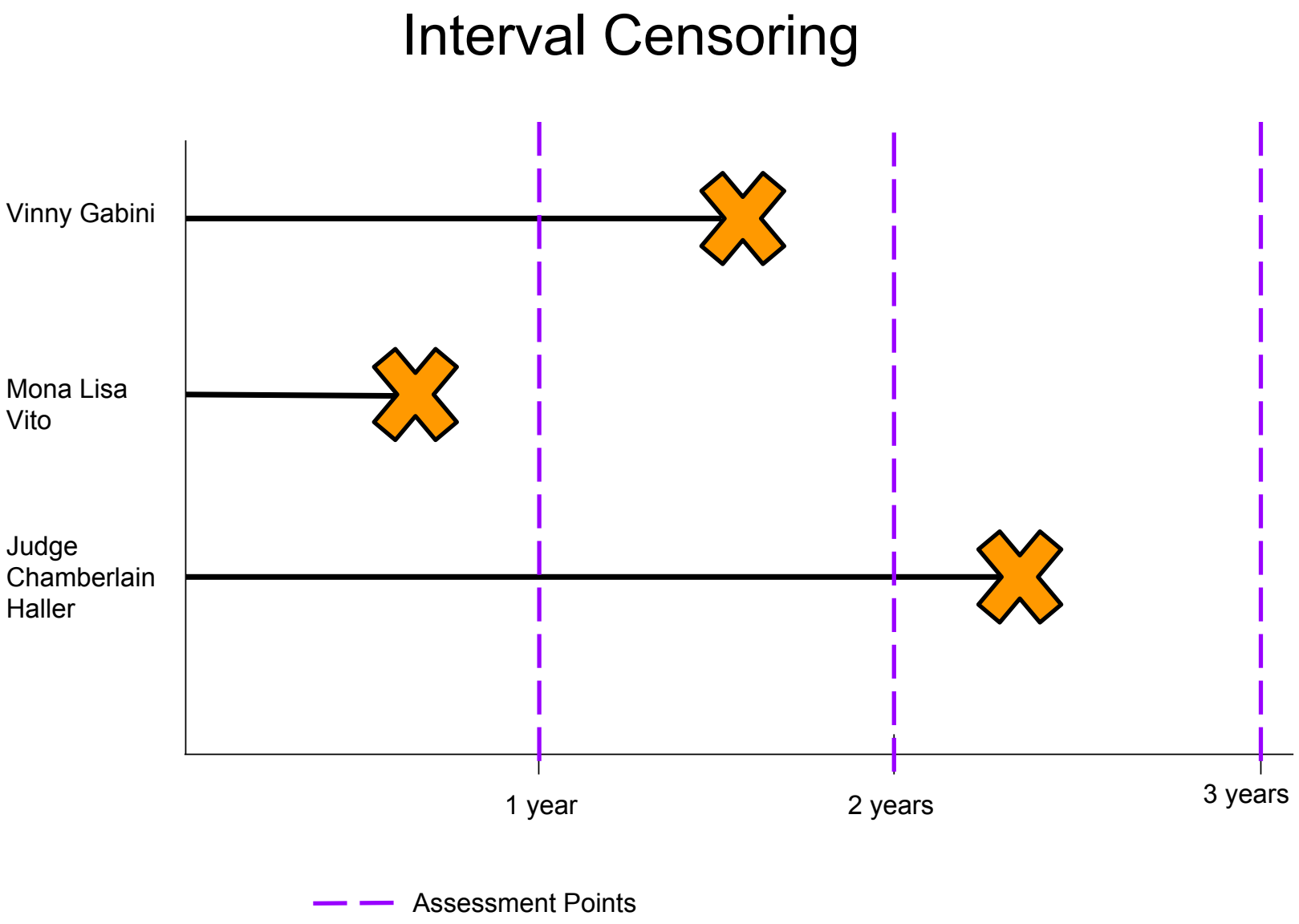


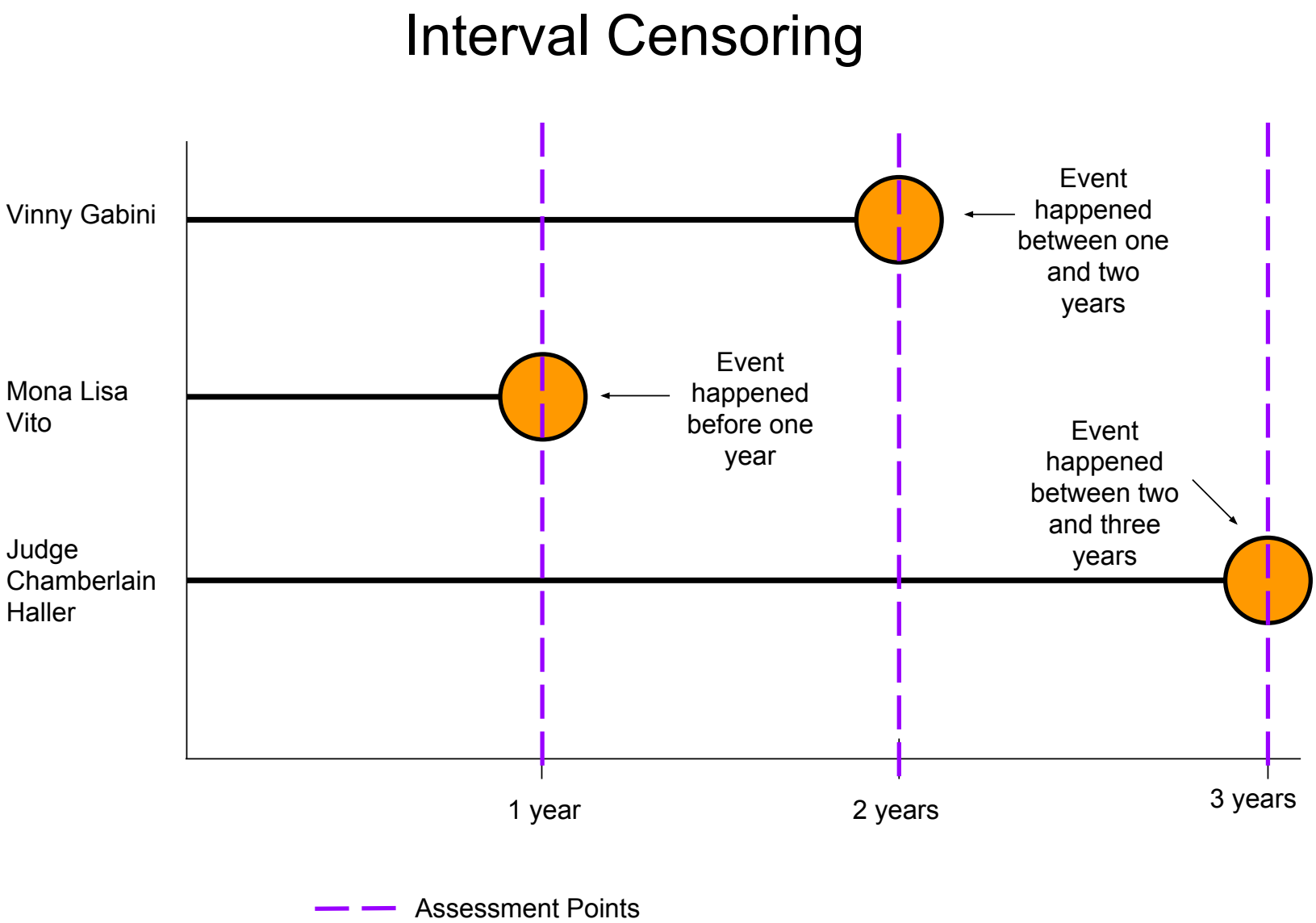
Left Censoring: Study Starts at 1 Year



Left Censoring: Study Starts at 1 Year




Interval Censoring
Vinny Gabini
Mona Lisa
Vito
Judge
Chamberlain
Haller
1 year
2 years
3 years
Assessment Points


Interval Censoring
Vinny Gabini
Mona Lisa Vito
Judge Chamberlain Haller
Event happened between one and two years
Event happened before one year
Event happened between two and three years
1 year
2 years
3 years
Assessment Points

For the rest of the workshop we will be dealing with the most common type of censoring, right censoring. Special methods exist for left and interval censored data.

Let X be the time to event.

Let C be the censoring time.

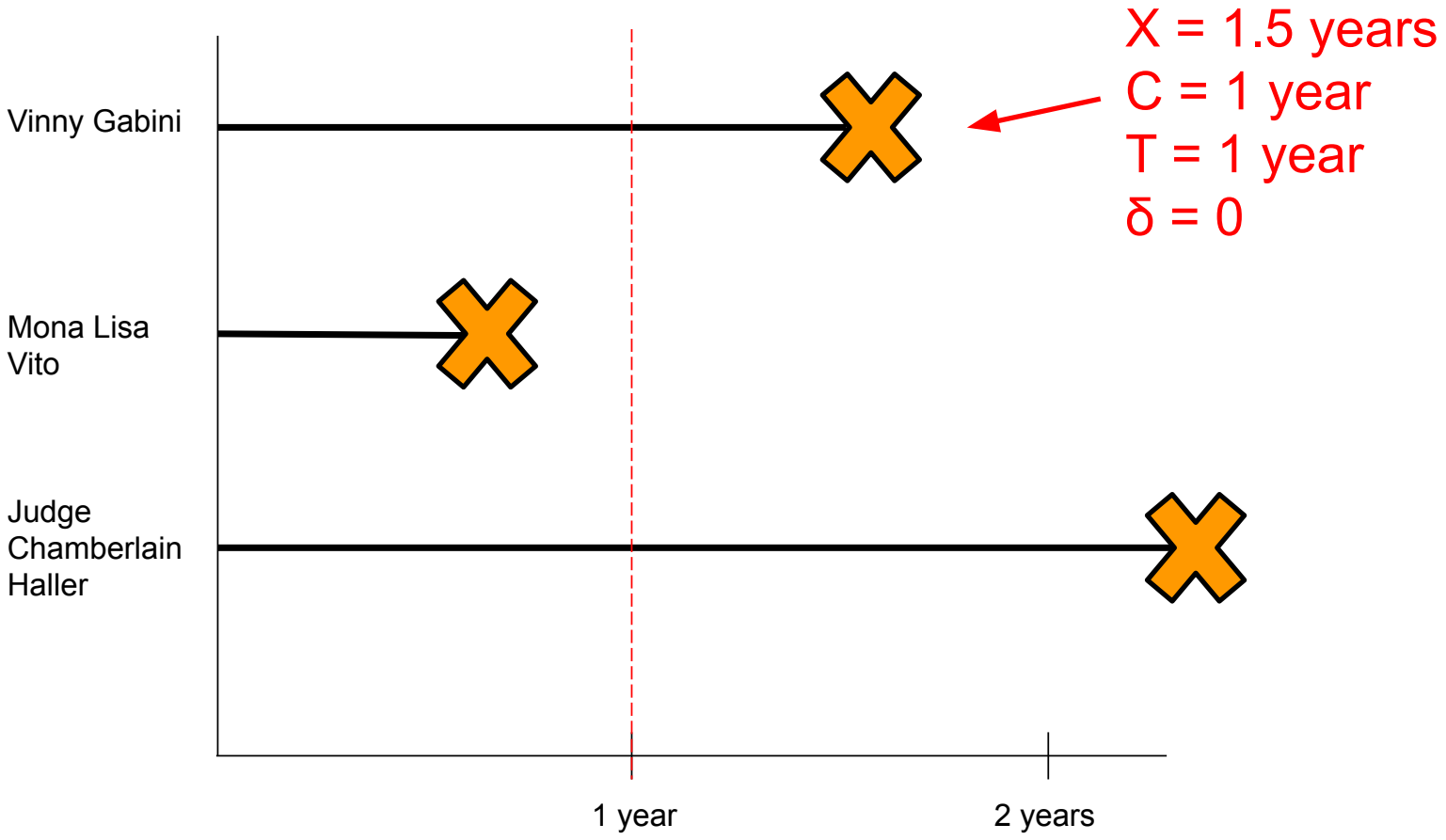
Let $T = min(X, C)$

Let δ be an indicator of if the observed data corresponds to an event:

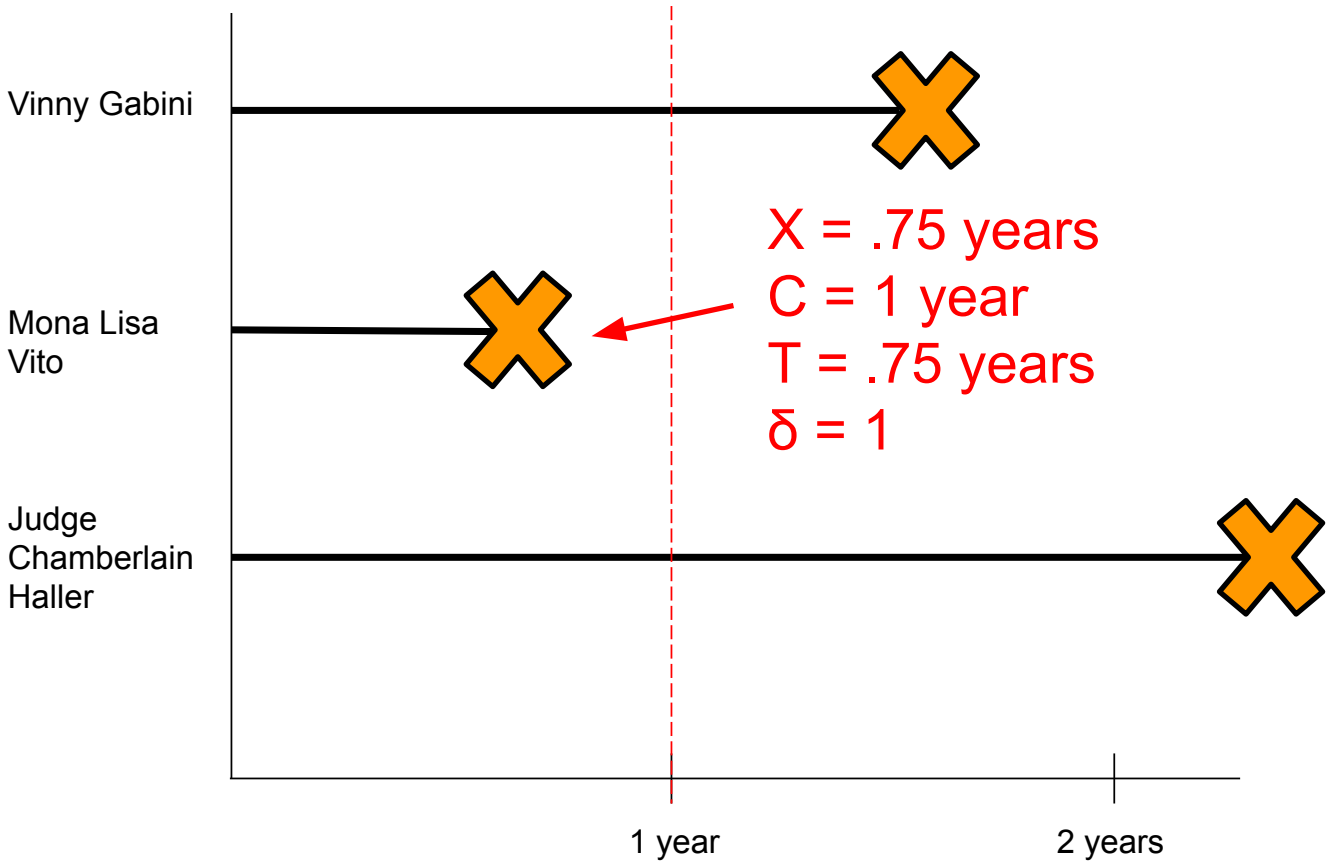
$$\delta = \begin{cases} 1 & \text{if } T = X \\ 0 & \text{if } T = C \end{cases}$$

The observed data is then (T, δ)

Right Censoring: 1 Year of Follow Up



Right Censoring: 1 Year of Follow Up



Refer to R code chunk 1

Components of Survival Data: The Survival Function

The basic quantity used to describe time-to-event data is the survival function. The survival function is the probability of surviving beyond time x (i.e. experiencing the event beyond time x).

$$S(x) = Pr(X > x)$$

The survival function takes value 1 at the origin and 0 at infinity.

When X is a continuous random variable, the survival function is the compliment of the cumulative distribution function (cdf):

$$S(x) = 1 - F(x)$$

Also, the survival function is the integral of the probability density function(pdf) , f(x), that is:

$$S(X) = P(X > x) = \int_x^{\infty} f(t)dt$$

Say we have data that follows an exponential distribution. We have a cdf of

$$F(x) = 1 - \exp^{-\lambda x}$$

and a pdf of

$$f(x) = \lambda \exp^{-\lambda x}$$

It follows that the expected lifetime (or mean of the exponential distribution) is $\frac{1}{\lambda}$.

Let's calculate the survival function using the cdf:

$$F(x) = 1 - \exp^{-\lambda x}$$

The survival function is:

$$\begin{aligned} S(x) &= Pr(X > x) \\ &= 1 - F(x) \\ &= 1 - (1 - \exp^{-\lambda x}) \\ &= \exp^{-\lambda x} \end{aligned}$$

Let's calculate the survival function using the pdf:

$$f(x) = \lambda \exp^{-\lambda x}$$

The survival function is:

$$\begin{aligned} S(x) &= Pr(X > x) \\ &= \int_x^\infty f(t)dt \\ &= \int_x^\infty \lambda \exp^{-\lambda t}\, dt \\ &= -\exp^{-\lambda \infty} - (-\exp^{-\lambda x}) \\ &= \exp^{-\lambda x} \end{aligned}$$

Example

Now let's see what this looks like in R!

Refer to R code chunk 2

Derive the survival function for a Weibull distribution. Then plot the cdf, pdf, and survival function in R.

The pdf for a Weibull distribution is:

$$f(x) = \alpha\lambda x^{\alpha-1} exp(-\lambda x^{\alpha})$$

The cdf for a Weibull distribution is:

$$F(x) = 1 - exp(-\lambda x^{\alpha})$$

One of the fundamental quantities in survival analysis is the hazard function. The hazard function is defined as:

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

If X is continuous the hazard function is equal to

$$h(x) = \frac{f(x)}{S(x)}$$

The hazard function can be viewed as the approximate probability of an individual of age x experiencing the event at the next instant. The hazard may take many shapes and the only restriction on the hazard is that $h(x) \geq 0$.

The hazard function for the exponential distribution is:

$$\begin{aligned} h(x) &= \frac{f(x)}{S(x)} \\ &= \frac{\lambda \exp^{-\lambda x}}{\exp^{-\lambda x}} \\ &= \lambda \end{aligned}$$

The exponential distribution has a constant hazard function.

Let's plot the hazard function.

Refer to R code chunk 3

Calculate and plot the hazard for the Weibull distribution.

We will deal with dates in R using the R package lubridate!

Date-time data can be frustrating to work with in R. R commands for date-times are generally unintuitive and change depending on the type of date-time object being used. Moreover, the methods we use with date-times must be robust to time zones, leap days, daylight savings times, and other time related quirks, and R lacks these capabilities in some situations. Lubridate makes it easier to do the things R does with date-times and possible to do the things R does not.

Refer to R code chunk 4

Refer to R code chunk 5

Explore and visualize the veteran dataset from the survival package.

- Introduction to survival time data
- **The survival function**
  - Kaplan-Meier estimate
  - Estimating median survival times
  - Estimating survival times at specific timepoints
  - Testing for between group differences
  - Kaplan-Meier plots in R
- Cox proportional hazards regression
- Competing risks methods

The basic quantity used to describe time-to-event data is the survival function. The survival function is the probability of surviving beyond time x (i.e. experiencing the event beyond time x).

$$S(x) = Pr(X > x)$$

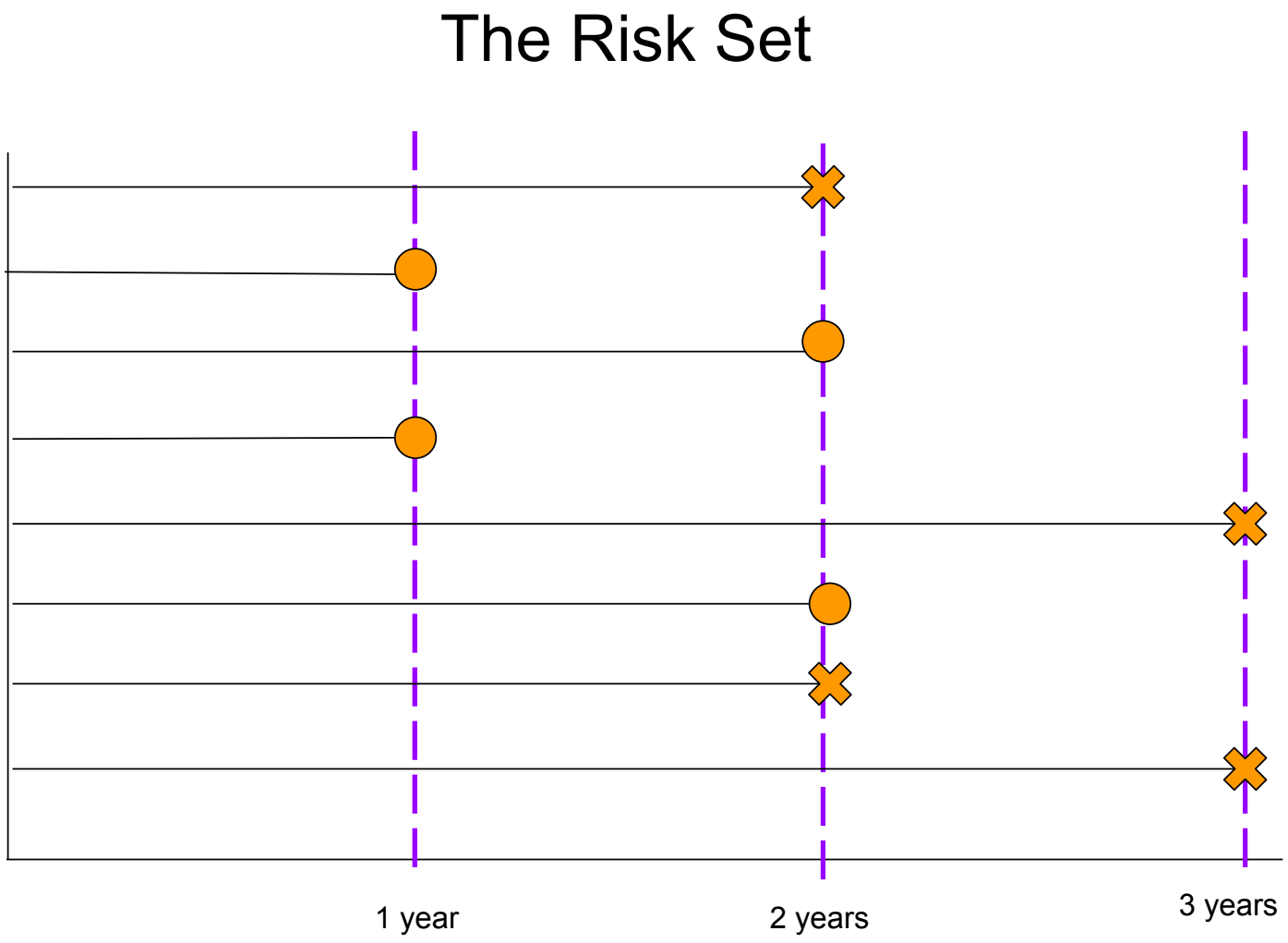
The survival function takes value 1 at the origin and 0 at infinity.

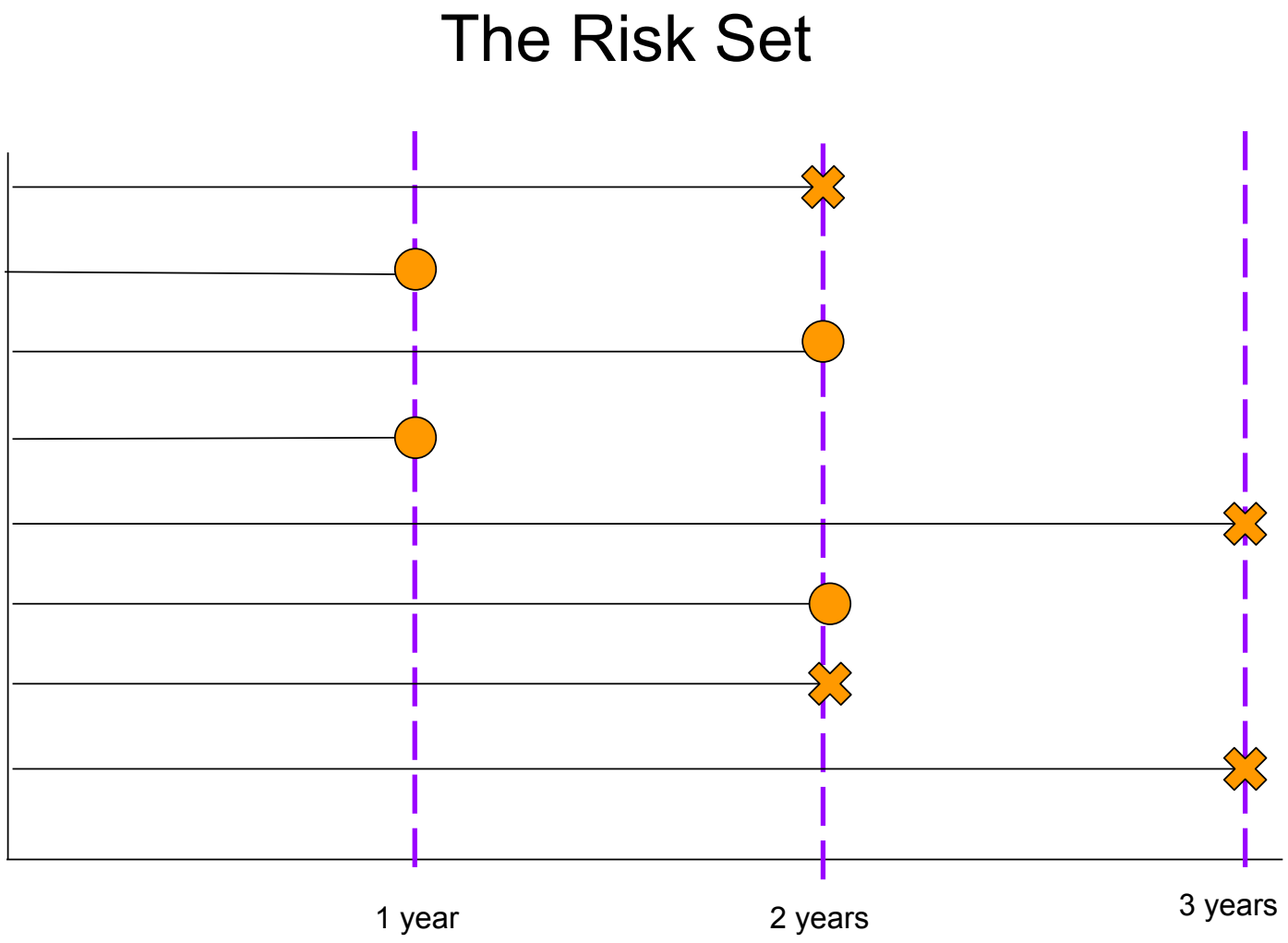
Question: In the last section we discussed survival functions when you know the distribution that your data follows. What if you do not know this distribution? What do you do?!?!?!

Answer: Kaplan-Meier Estimator

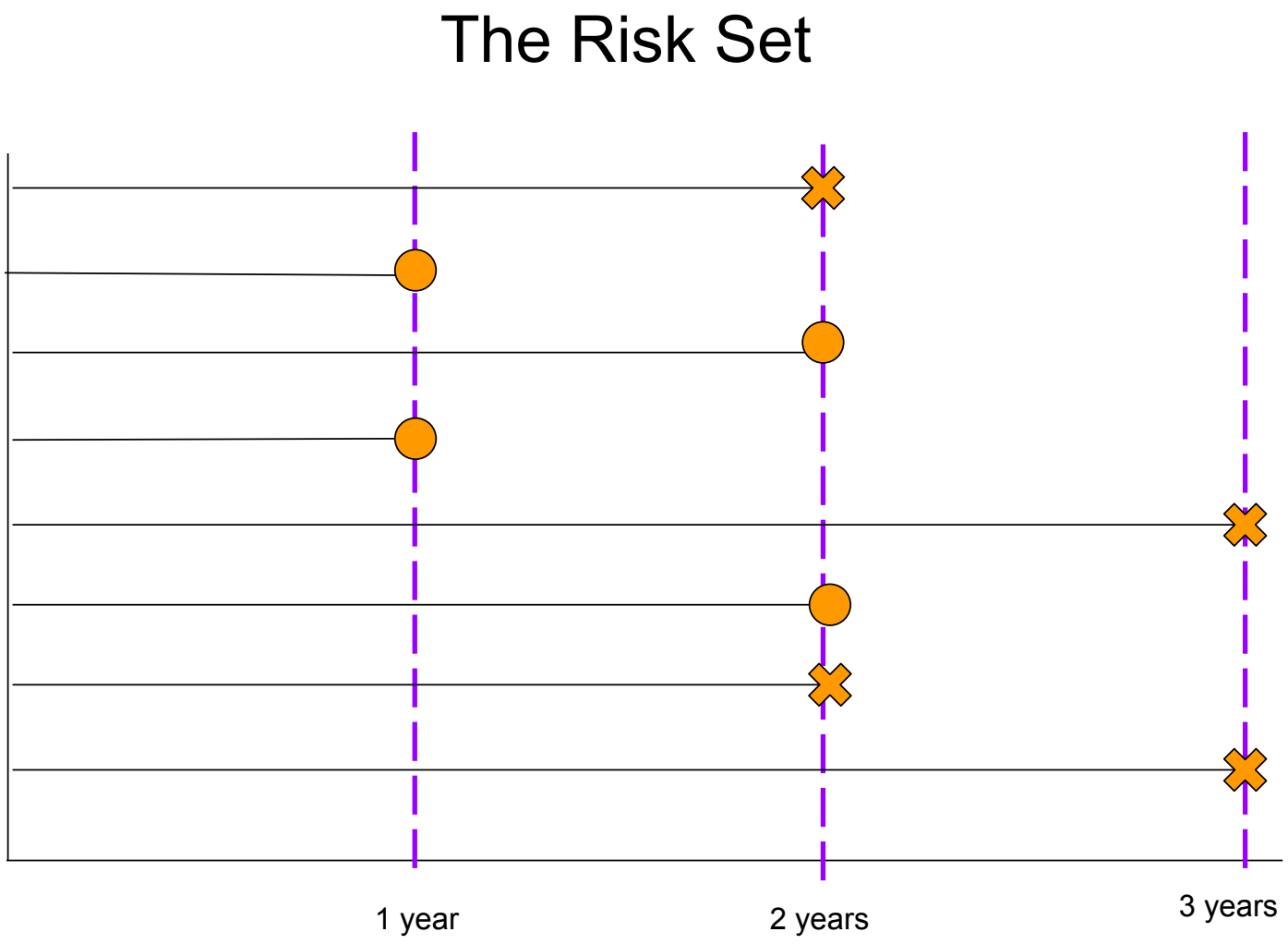
The first thing we need to understand to calculate the Kaplan-Meier estimator is the notion of a 'risk set'.

The risk set is all of the individuals who are at risk to have an event at time t. This includes individuals who are known to be alive at time t and those who have the actual event at time t.


The Risk Set
1 year
2 years
3 years


The Risk Set
1 year
2 years
3 years

Who is in the risk set at 3 years?



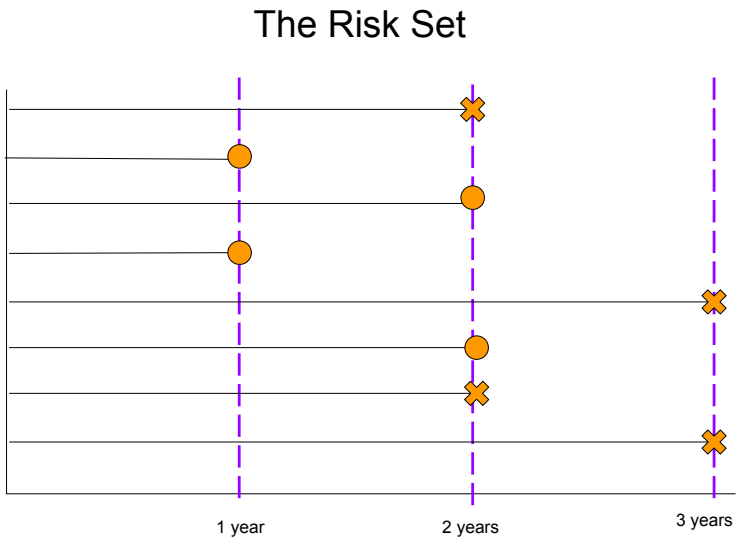
Say we have distinct follow-up time points $t_1 < t_2 < ... < t_D$. Let there be d_i events at time t_i. Let Y_i be the number of individuals who are at risk at time t_i. This includes individuals who are known to be alive at time t_i and those who have the actual event at time t_i.

The Kaplan-Meier estimator is:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if } t_1 \leq t \end{cases}$$

Assumptions:

- that censoring is unrelated to prognosis
- the survival probabilities are the same for subjects recruited early and late in the study
- the events happened at the times specified



$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if } t_1 \leq t \end{cases}$$

$$\hat{S}(0) = 1$$
$$\hat{S}(2) = (1 - 1/2) = 1/2$$
$$\hat{S}(3) = (1 - 1/2)(1 - 2/2) = 0$$

Death times of psychiatric patients: Survival data for 26 psychiatric inpatients admitted to the University of Iowa hospital during the years 1935 - 1948 (Wolson, 1981).

1, 1, 2, 22, 30+, 28, 32, 11, 14, 36+, 31+, 33+, 33+, 37+, 35+, 25, 31+, 22, 26, 24, 35+, 34+, 30+, 35, 40, 39

We will make a Kaplan Meier estimator for the survival function for this data.

Step 1: Order the data

1, 1, 2, 11, 14, 22, 22, 24, 25, 26, 28, 30+, 30+, 31+, 31+, 32, 33+, 33+, 34+, 35+, 35+, 35, 36+, 37+, 39, 40

Step 2: Determine the time points of events

1, 1, 2, 11, 14, 22, 22, 24, 25, 26, 28, 30+, 30+, 31+, 31+, 32, 33+, 33+, 34+, 35+, 35+, 35, 36+, 37+, 39, 40

t_i	d_i	Y_i	$\hat{S}(t) = \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}]$
1			
2			
11			
14			
22			
24			
25			
26			
28			
32			
35			
39			
40			

Step 3: Fill in the chart!

1, 1, 2, 11, 14, 22, 22, 24, 25, 26, 28, 30+, 30+, 31+, 31+, 32, 33+, 33+, 34+, 35+, 35+, 35, 36+, 37+, 39, 40

t_i	d_i	Y_i	$\hat{S}(t) = \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}]$
1			
2			
11			
14			
22			
24			
25			
26			
28			
32			
35			
39			
40			

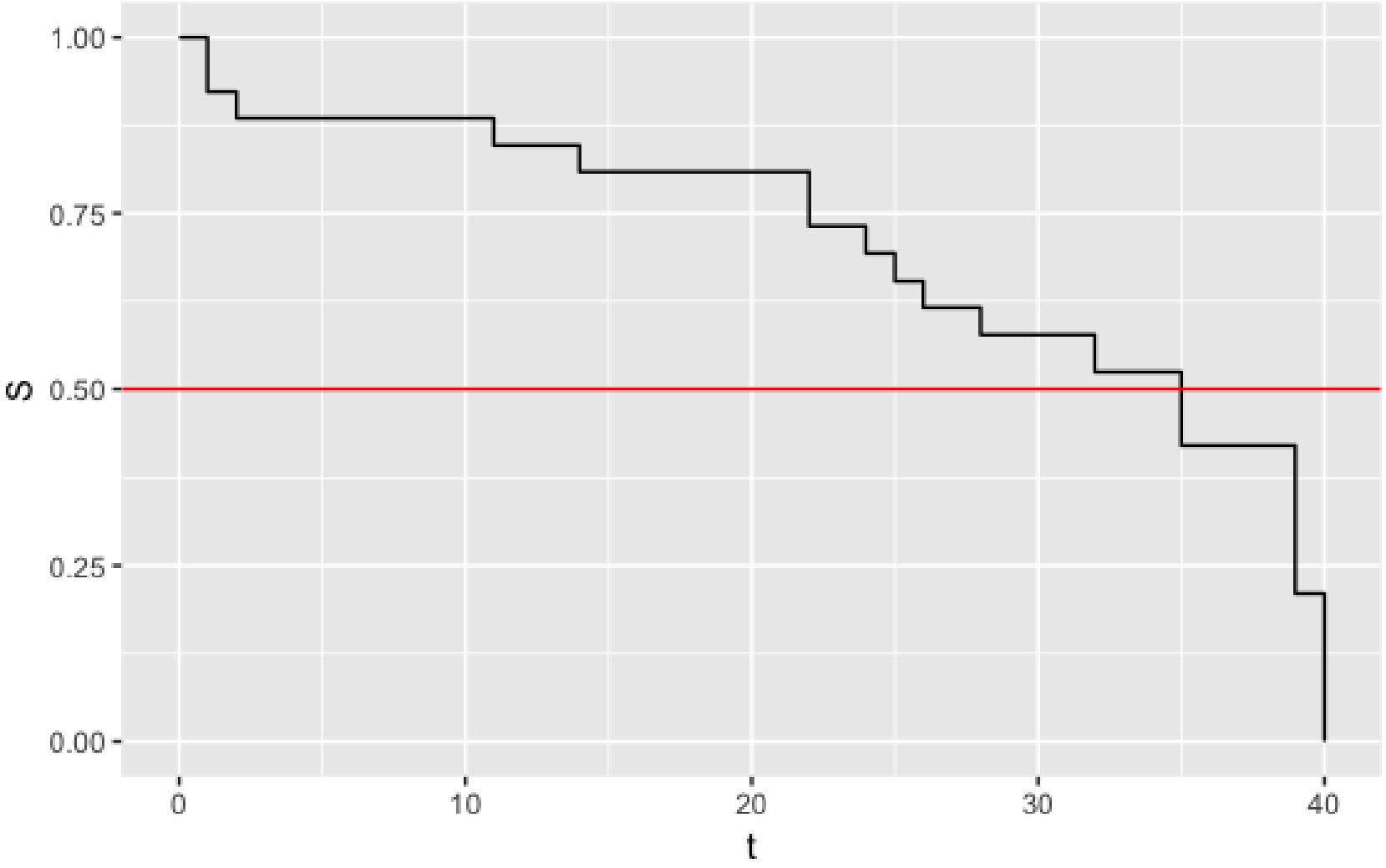
Plot the Kaplan-Meier Estimator

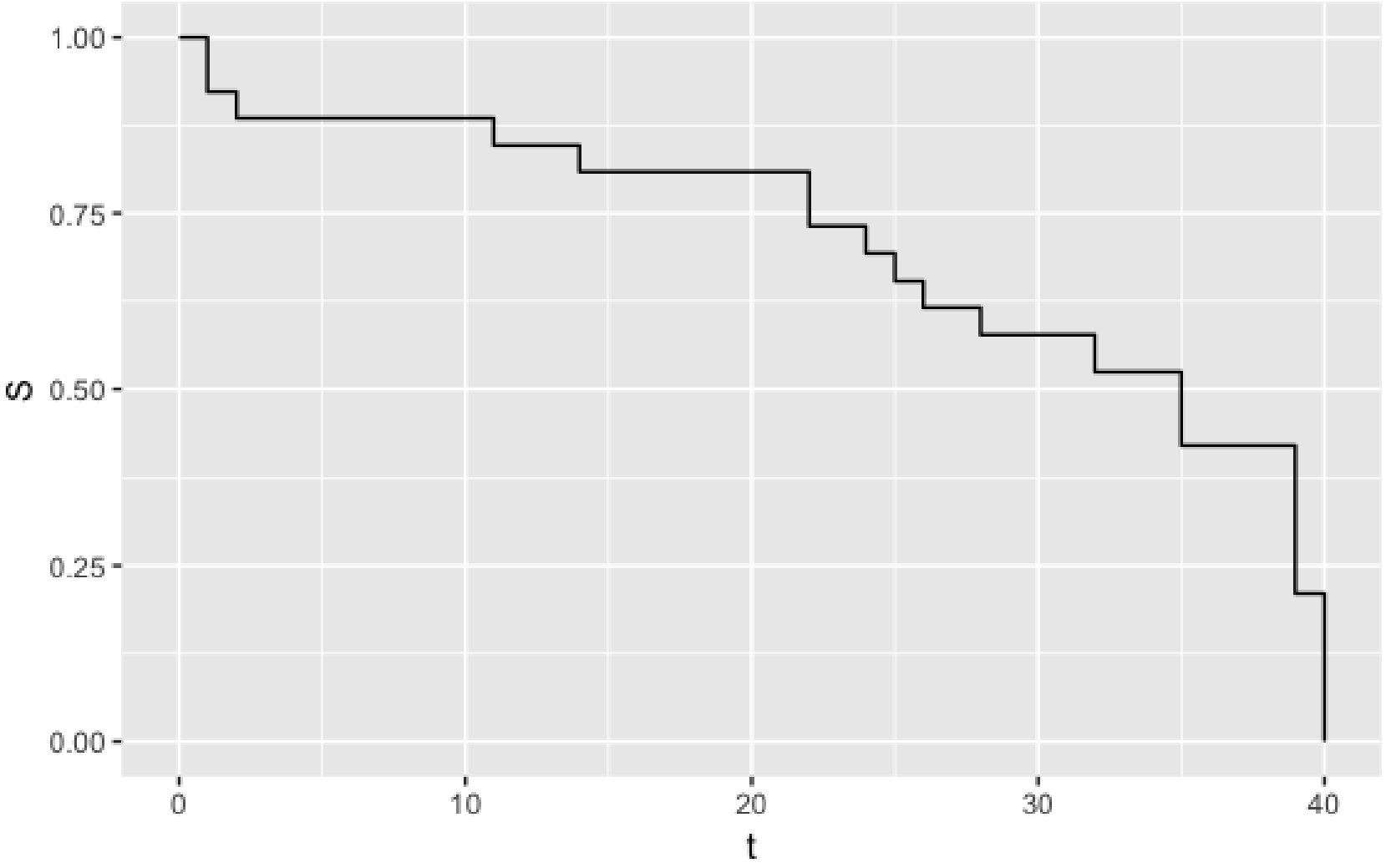
Refer to R code chunk 6

The median survival time is defined to be

$$\text{median} = \inf\{t : S(t) \leq t\}$$

Estimating median survival times




1.00
0.75
0.50
0.25
0.00
S
0
10
20
30
40
t

We have the following data (in years):

1, 2, 3+, 3, 5, 5+, 6, 7, 8+, 10, 11+

1. Calculate the Kaplan-Meier estimator for the data.
2. Plot the Kaplan-Meier curve.
3. Estimate the median survival time.
4. What is the survival probability at time 8? At time 6?

We have the following data (in years):

1, 2, 3+, 3, 5, 5+, 6, 7, 8+, 10, 11+

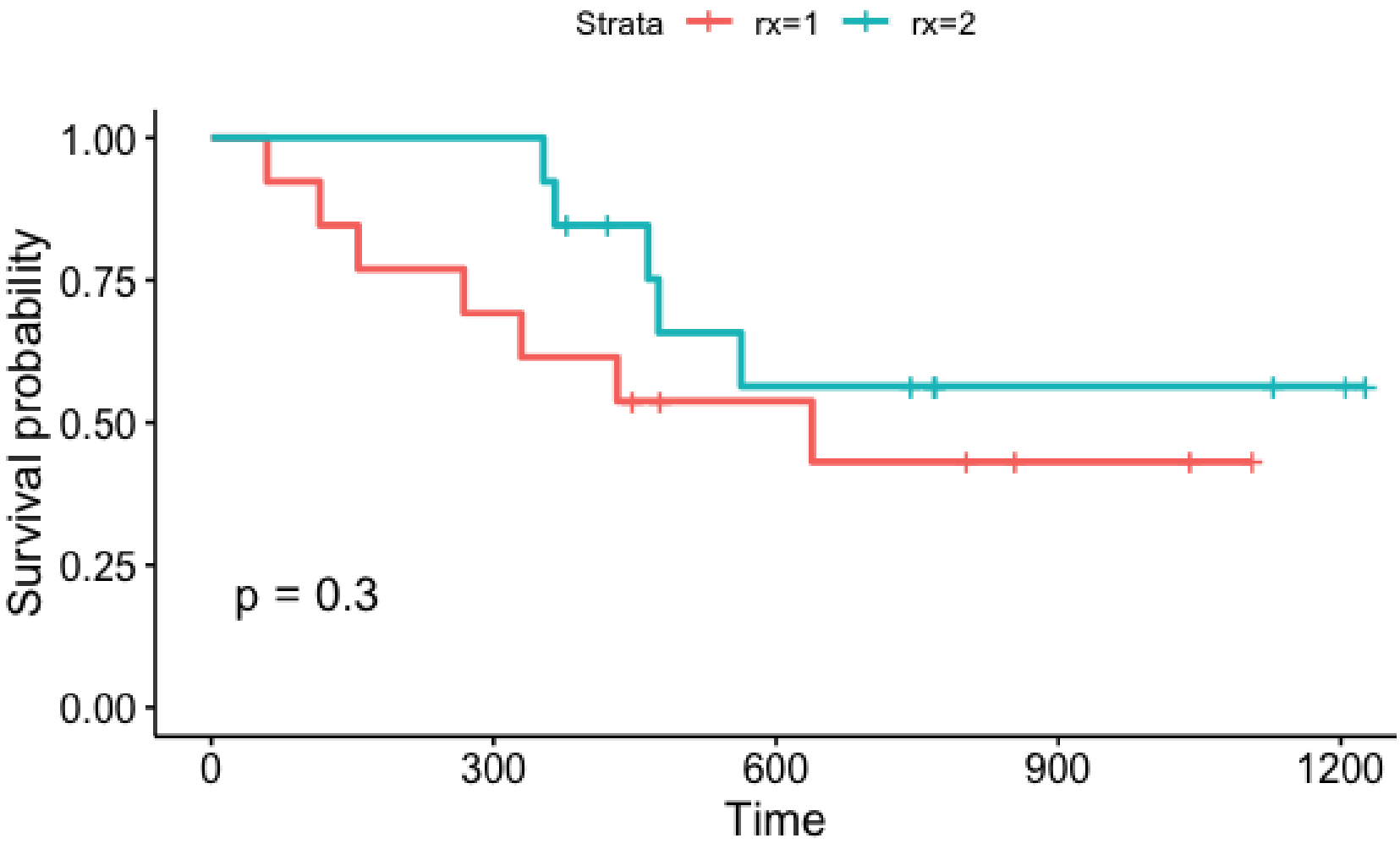
t_i	d_i	Y_i	$\hat{S}(t) = \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}]$
1			
2			
3			
5			
6			
7			
10			

We have the following data (in years):

1, 2, 3+, 3, 5, 5+, 6, 7, 8+, 10, 11+

t_i	d_i	Y_i	$\hat{S}(t) = \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}]$
1			
2			
3			
5			
6			
7			
10			

Ovarian data example for two different treatment groups:



To test for differences between two groups we will use the Logrank test. The Logrank test tests if the hazard function of two groups are the same.

$H_0 : h_1(t) = h_2(t)$ for all t

$H_a : h_1(t) \neq h_2(t)$ for some t

The Logrank test:

Let $j = 1, ..., J$ index the event times in either group.

Let Y_{1j} and Y_{2j} be the be the number of subjects at risk in either group at each time point.

Let d_{1j} and d_{2j} be the number of events at each time point in each group.

Let $Y_j = Y_{1j} + Y_{2j}$ and $d_j = d_{1j} + d_{2j}$

Under the null hypothesis d_{1j} follows a hypergeometric distribution with

$$E_{1j} = \frac{d_j}{Y_j} Y_{1j}$$

$$V_j = \frac{d_j (Y_{1j}/Y_j)(1 - Y_{1j}/Y_j)(Y_j - d_j)}{Y_j - 1}$$

It follows that

$$Z = \frac{\sum_{j=1}^{J}(d_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}} \rightarrow N(0,1)$$

Let's do this test in R for the ovarian dataset

Refer to R Code chunk 7

Perform the logrank test for the veterans and jasa datsets (based on treatment in the veterans dataset and transplant in the jasa dataset).

We will now make Kaplan-Meier plots in R using the 'ggsurvplot' command from the survminer R package!

Refer to R Code chunk 7

Make Kaplan-Meier curves for the jasa dataset.

1. The entire dataset
2. Stratified by transplant status
3. Stratified by age $>=$ 50

- Introduction to survival time data
- The survival function
- **Cox proportional hazards regression**
  - The Cox proportional hazards model
  - The proportional hazards assumption
  - Incorporating time dependent covariates
- Competing risks methods

One of the fundamental quantities in survival analysis is the hazard function. The hazard function is defined as:

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

If X is continuous the hazard function is equal to

$$h(x) = \frac{f(x)}{S(x)}$$

The hazard function can be viewed as the approximate probability of an individual of age x experiencing the event at the next instant. The hazard may take many shapes and the only restriction on the hazard is that $h(x) \geq 0$.

For a group of n individuals, we have the data (T_j, δ_j, Z_j) where $Z_j = (Z_{j1}, Z_{j2}, ..., Z_{jp})^t$ is a set of p covariates for the j^{th} individual. The Cox proportional hazards model models the hazard at time t for an individual with covariates vector Z as:

$$h(t|Z) = h_0(t)exp\left(\sum_{k=1}^{p} \beta_k Z_k\right)$$

Where $h_0(t) = h(t|z = 0)$ is the baseline hazard rate.

$$h(t|Z) = h_0(t)exp\left(\sum_{k=1}^{p}\beta_k Z_k\right)$$

This model is considered to be 'semi-parametric'. The baseline hazard can take any form as a function of t (so long as $h_0(t) \geq 0$) and is the non-parametric part of the model. The $\sum_{k=1}^{p}\beta_k Z_k$ is the parametric part.

Let's fit the model in R.

Refer to R code chunk 8.

$$h(t) = h_0(t)exp(\beta_1 rx + \beta_2 resid.ds + \beta_3 age\_group + \beta_4 ecog.ps)$$

	coef	exp(coef)	se(coef)	z	p
rx.B	-1.38	0.25	0.64	-2.14	0.03
resid.ds.yes	1.45	4.25	0.73	1.98	0.05
age_group.young	-2.20	0.11	1.11	-1.99	0.05
ecog.ps.bad	0.59	1.80	0.63	0.93	0.35

- coef – the estimate of β_i
- exp(coef) – the estimate of e^{β_i}
- se(coef) – standard error of the estimate of β_i
- z = coef / se(coef) – the Wald statistic for the testing the null hypothesis that $\beta_i = 0$; under the null z follows a standard normal distribution
- p – two sided p-value

Let's interpret the coefficients!

$$h(t) = h_0(t)exp(\beta_1 rx + \beta_2 resid.ds + \beta_3 age\_group + \beta_4 ecog.ps)$$

Assume that we have two individuals with the same covariates for resid.ds, age_group, and ecog.ps. Say

- resid.ds = yes,
- age_group = young
- ecog.ps = bad

The only difference between the two individuals is that one is on treatment A while the other is on treatment B.

$$h(t) = h_0(t)exp(\beta_1 rx + \beta_2 resid.ds + \beta_3 age\_group + \beta_4 ecog.ps)$$

Let's look at the hazards for these two individuals:

$$h_1(t) = h_0(t)exp(\beta_1 + \beta_2 + \beta_3 + \beta_4)$$

$$h_2(t) = h_0(t)exp(\beta_2 + \beta_3 + \beta_4)$$

The ratio of these hazards is

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t)exp(\beta_1 + \beta_2 + \beta_3 + \beta_4)}{h_0(t)exp(\beta_2 + \beta_3 + \beta_4)} = exp(\beta_1)$$

Taking the log of both sides gives us

$$log\left(\frac{h_1(t)}{h_2(t)}\right) = \beta_1$$

The interpretation of β_1 is the log of the hazard ratio of death for people on treatment B compared to treatment A, keeping all else constant.

$$log\left(\frac{h_1(t)}{h_2(t)}\right) = \beta_1$$

Note that

- $\beta_1 > 0$ higher hazard (indicates poorer survival)
- $\beta_1 < 0$ lower hazard (indicates better survival)
- $\beta_1 = 0$ no association

The sign of β is typically interpretable, but interpreting the magnitude is difficult. We therefore often prefer the interpretation of the exponentiated coefficients.

The interpretation of e^{β_1} is the hazard ratio of death for people on treatment B compared to treatment A, keeping all else constant.

$$\frac{h_1(t)}{h_2(t)} = e^{\beta_1}$$

Note that

- $e^{\beta_1} > 1$ higher hazard (indicates poorer survival)
- $e^{\beta_1} < 1$ lower hazard (indicates better survival)
- $e^{\beta_1} = 1$ no association

$$h(t) = h_0(t)exp(\beta_1 rx + \beta_2 resid.ds + \beta_3 age\_group + \beta_4 ecog.ps)$$

	coef	exp(coef)	se(coef)	z	p
rx.B	-1.38	0.25	0.64	-2.14	0.03
resid.ds.yes	1.45	4.25	0.73	1.98	0.05
age_group.young	-2.20	0.11	1.11	-1.99	0.05
ecog.ps.bad	0.59	1.80	0.63	0.93	0.35

- coef – the estimate of β_i
- exp(coef) – the estimate of e^{β_i}
- se(coef) – standard error of the estimate of β_i
- z = coef / se(coef) – the Wald statistic for the testing the null hypothesis that $\beta_i = 0$; under the null z follows a standard normal distribution
- p – two sided p-value

Let's use the 'ggforest' function from the survminer package to a plot as an alternative to our table!

Refer to R Code Chunk 9

We will fit the cox proportional hazards model for the veterans dataset.

Refer to R Code Chunk 10

Fit a Cox proportional hazard model to the jasa dataset. Fit a first model that compares people who have a transplant and do not, adjusting for available confounders. Then fit a second model to understand survival in the group that received a transplant.

Implicit in the Cox model is the proportional hazards assumption. For simplicity, say we have a model with one covariate of treatment, T, with T=1 being a subject is on a treatment and T=0 being that a subject is not on treatment:

$$h(t|T) = h_0(t)exp(\beta T)$$

Now say we have two individuals, one who is on the treatment and one who is not. If we look at the ratio of the hazards between these two individuals we have

$$\frac{h(t|T=1)}{h(t|T=0)} = \frac{h_0(t)exp(\beta(1))}{h_0(t)exp(\beta(0))} = exp(\beta)$$

Note that the ratio of the hazards is constant, which indicates that the hazard rates are proportional.

In general, this holds. Lets look at two individuals with covariates Z and Z^*. The ratio of their hazards is

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k Z_k\right]}{h_0(t)\exp\left[\sum_{k=1}^{p}\beta_k Z_k^*\right]} = exp\left[\sum_{k=1}^{p}\beta_k(Z_k - Z_k^*)\right]$$

Note that the ratio of the hazards is constant, which indicates that the hazard rates are proportional.

We will need to test the proportional hazards assumption when we are fitting a cox proportional hazards model.

The proportional hazards assumption can be tested in a few ways. We will explore two of these ways

- Visual inspection of the Kaplan Meier curves
- Scaled Schoenfeld Residuals

Calculating the Schoenfeld residuals is outside of the scope of this workshop. All we need to know is that these residuals should be independent of time.

Refer to R Code chunk 11

What should you do when the proportional hazards assumption is violated?!?!?!

Fit a model with time-varying coefficients.

Let's think about a model with only one coefficient. What if we allowed β to be a time-varying coefficient? We now have the model:

$$h(t|Z) = h_0(t)exp\left(\beta(t)Z\right)$$

In general, this holds. Let's look at two individuals with covariates Z and Z^*. The ratio of their hazards is

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp\left[\beta(t)Z\right]}{h_0(t)\exp\left[\beta(t)Z^*\right]} = exp\left[\beta(t)(Z - Z^*)\right]$$

Note that the ratio of the hazards is no longer constant and can vary with time!

The simplest time varying coefficient is to turn β into a step function. We will do this with the veteran's dataset.

Refer to R Code Chunk 12

What if we have a covariate that we would like to put into the model that changes over time?

What if we have a covariate that we would like to put into the model that changes over time?

We now have the model

$$h[t|Z(t)] = h_0(t)exp\left[\sum_{k=1}^{p}\beta_j Z(t)\right]$$

In general, this holds. Let's look at two individuals with covariates $Z(t)$ and $Z^*(t)$. The ratio of their hazards is

$$\frac{h(t|Z)}{h(t|Z^*)} = \frac{h_0(t)\exp\left[\beta Z(t)\right]}{h_0(t)\exp\left[\beta Z^*(t)\right]} = exp\left[\beta(Z(t) - Z^*(t))\right]$$

Note that the ratio of the hazards is not constant and can vary with time!

Let's motivate this with a data example in R!

Refer to R Code Chunk 13

- Introduction to survival time data
- The survival function
- Cox proportional hazards regression
- **Competing risks methods**
  - Cause-specific hazards
  - Cumulative incidence
  - Cumulative incidence plots in R
  - Competing risks regression

In standard survival analysis, subjects are supposed to experience only one type of event over follow-up. An example of this is death due to cancer.

In real life, subjects can potentially experience more than one type of a certain event. Returning to our example, senior patients at an oncology department could possibly die from a heart attack or even traffic accident before they die from cancer.

How do we account for this in our analysis?

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

- Competing Events – When there are multiple events but only one of the events can occur we call this 'competing events'. This is because the events compete with each other and the occurrence of one type of event will prevent the occurrence of the others. An example of this is death from cancer, which competes with death from other causes.
- Competing Risks – We refer to the probabilities of competing events as 'competing risks', in a sense that the probability of each competing event is somehow regulated by the other competing events, which has an interpretation suitable to describe the survival process determined by multiple types of events.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

Some examples...

- A patient can die from breast cancer or from stroke, but he cannot die from both
- A breast cancer patient may die after surgery before they can develop hospital infection
- A soldier may die during a combat or in a traffic accident

In the examples above, there are more than one pathway that a subject can fail, but the failure, either death or infection, can only occur once for each subject (without considering recurring event). Therefore, the failures caused by different pathways are mutually exclusive and hence called competing events.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

One of the fundamental quantities in survival analysis is the hazard function. The hazard function is defined as:

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

If X is continuous the hazard function is equal to

$$h(x) = \frac{f(x)}{S(x)}$$

The hazard function can be viewed as the approximate probability of an individual of age x experiencing the event at the next instant. The hazard may take many shapes and the only restriction on the hazard is that $h(x) \geq 0$.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

In competing event data, the typical approach involves the use of Kaplan-Meier estimator to separately estimate probability for each type of event, while treating the other competing events as censored in addition to those who are censored from loss to follow-up or withdrawal. This leads to the cause-specific hazard function:

$$h_c(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X_c < x + \Delta x | X_c \geq x]}{\Delta x}$$

The random variable X_c denotes the time to failure from event type c, therefore the cause-specific hazard function $h_c(t)$ gives the instantaneous failure rate at time t from event type c, given not failing from event c by time t.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

By the law of total probability, we have:

$$h(x) = \sum_{c=1}^{C} h_c(x)$$

[1]Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

Let's calculate the cause specific Kaplan-Meier estimator in R $S_c(t)$.
This is calculated just like the regular Kaplan-Meier estimator by treating competing risks as censored observations.

Refer to R Code Chunk 14

Using these methods, one can separately estimate failure rate for each one of competing events. For instance, in our breast cancer mortality example, when death from breast cancer is the event of interest, the death from heart attack and all other causes should be treated as censored in addition to conventional censored observations. This would allow us to estimate the cause-specific hazard for cancer mortality rate, and go on to fit a cause-specific hazard model on cancer mortality. The same procedure can apply to death from heart attack when it becomes event of interest.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

We even have the cause-specific hazards model based upon the Cox regression model:

$$h_c(t|x) = h_{0c}(t) exp\left[\sum_{k=1}^{K} \beta_k Z_k\right]$$

[1]Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

PROBLEM! The cause specific hazards approach assumes independent censoring! (An assumption that we have been making in all of our work up to this point).

What is independent censoring?

Independent censoring essentially means that within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience. In other words, censoring is independent provided that it is random within any subgroup of interest.

Independent censoring means that, conditional on covariates at each duration, the censored items are ?representative? of those under observation at the same time.

[1]Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

Suppose this assumption is true, when focusing on cause-specific death rate from breast cancer, then any censored subject at time t would have the same death rate from breast cancer, regardless of whether the reason for censoring is either CVD or other cause of death, or loss to follow-up. This assumption is equivalent to saying competing events are independent, which is the foundation for the KM type of analysis to be valid.

There is no way to test this assumption in your data and it is probably note true :(

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

The solution to this problem is to look at the Cumulative Incidence Function, which estimates the marginal probability for each competing event. Marginal probability is defined as the probability of subjects who actually developed the event of interest, regardless of whether they were censored or failed from other competing events.

[1]Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

We denote the Cumulative Incidence Function as CIF_c. It is defined to be:

$$CIF_c(x) = Pr(\text{Failure Time} X \leq x, \text{cause } = c)$$

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

The cumulative incidence function is easy to calculate! Say we have distinct follow-up time points $t_1 < t_2 < ...t_D$. Let $d_i c$ be the number of events for risk c at time t_i and Y_{ic} be number of subjects at risk for event c at time t_i.

$$CIF_c(t) = \sum_{t_i \leq t} \hat{S}(t_i - 1)\frac{d_{ic}}{Y_{ic}}$$

[1]Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

- The CIF is equivalent to 1-Kaplan - Meier estimator when there is no competing event.
- When there is competing event, the CIF differs from 1-Kaplan -Meier estimator in that it uses overall survival function $S(t)$ that counts failures from competing events in addition to the event of interest, whereas the 1-Kaplan-Meier estimator uses the event-type specific survival function $S_c(t)$, which treats failures from competing events as censored.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

- By using the overall survival function, CIF bypasses the need to make unverifiable assumptions of independence of censoring on competing events.
- Since the $S(t)$ is always less than $S_c(t)$, in competing event data, the CIF is always smaller than 1-KM estimates, which means the 1-KM tends to overestimate the probability of failure from the event type of interest.

[1] Source: https://www.mailman.columbia.edu/research/population-health-methods/competing-risk-analysis

Let's plot the Cumulative Incidence Function in R!

Refer to R Code Chunk 15

We can also perform competing risk regression in R.

Refer to R Code Chunk 16