DEEP LEARNING FOR PATIENT-SPECIFIC KIDNEY GRAFT SURVIVAL ANALYSIS

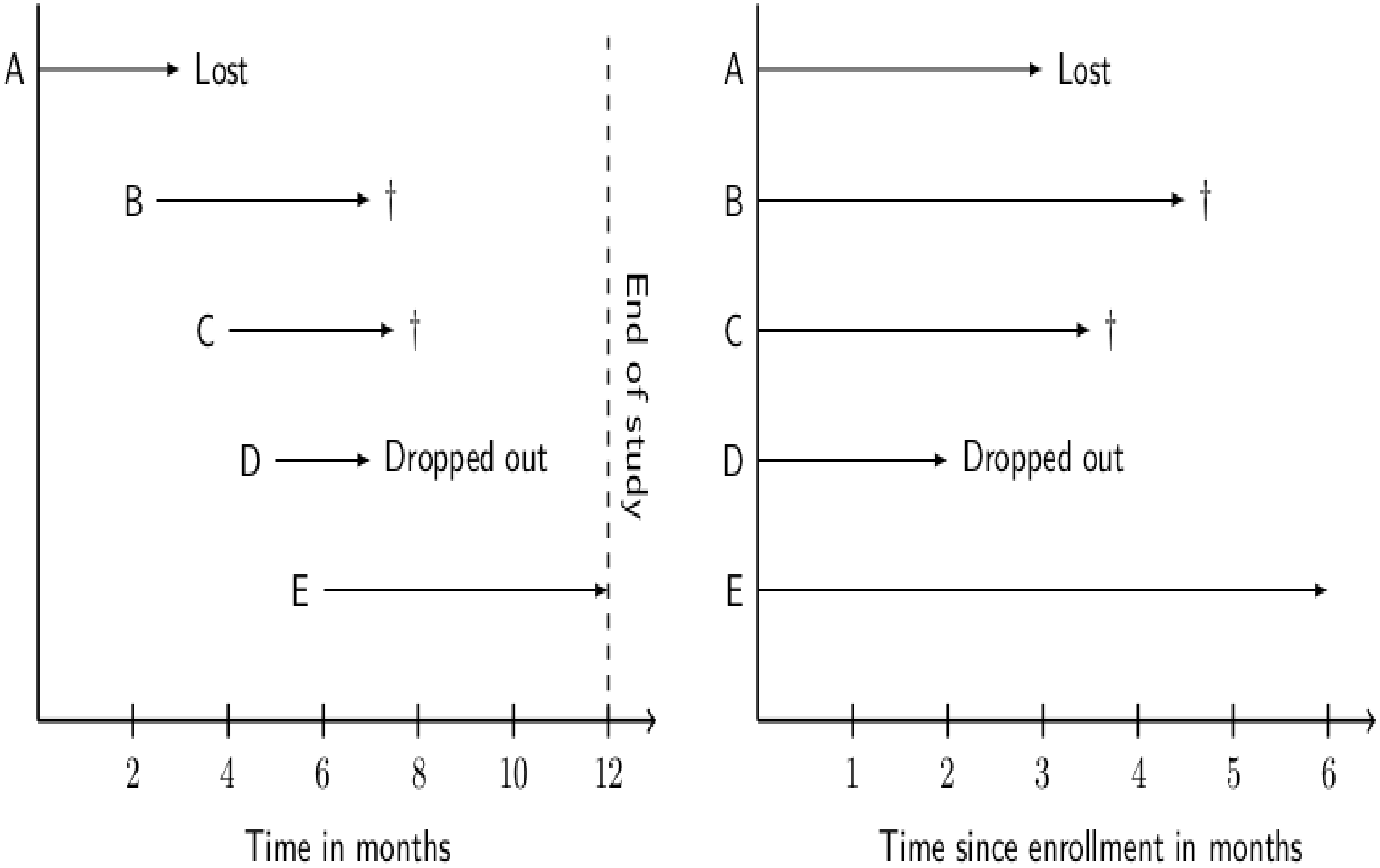
Group Members:

Nishat Zaman	(011 151 264)
Nawshiba Tasnim Ahmed	(011 161 145)
Ferdous Zaman	(011 162 125)
Ebnul Mahmood Shovan	(011 162 058)

Survival Analysis

Survival analysis is a type of regression problem (one wants to predict a continuous value), but with a twist. It differs from traditional regression by the fact that parts of the training data can only be partially observed – they are *censored*.

As an example, consider a clinical study, which investigates coronary heart disease and has been carried out over a 1 year period as in the figure below.


A
Lost
B
†
C
†
D
Dropped out
E
End of study
2
4
6
8
10
12
Time in months
A
Lost
B
†
C
†
D
Dropped out
E
1
2
3
4
5
6
Time since enrollment in months

Survival Analysis

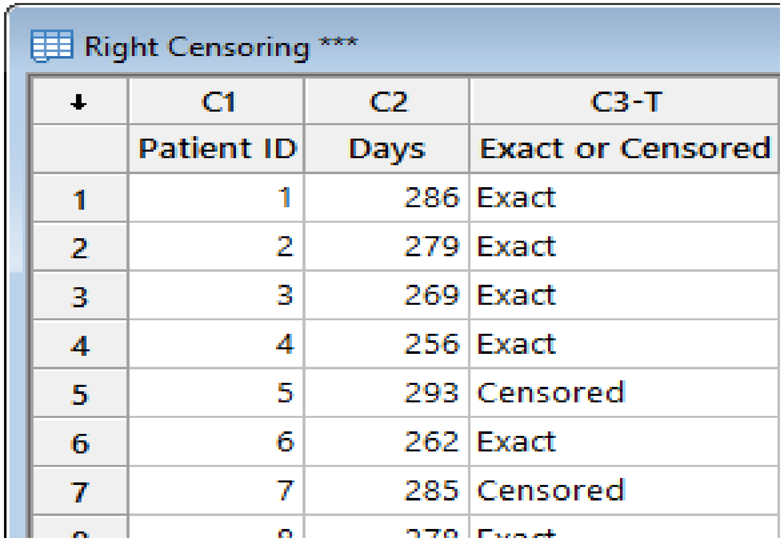
Patient A was lost to follow-up after three months with no recorded cardiovascular event, patient B experienced an event four and a half months after enrollment, patient D withdrew from the study two months after enrollment, and patient E did not experience any event before the study ended. Consequently, the exact time of a cardiovascular event could only be recorded for patients B and C; their records are *uncensored*. For the remaining patients it is unknown whether they did or did not experience an event after termination of the study. The only valid information that is available for patients A, D, and E is that they were event-free up to their last follow-up. Therefore, their records are *censored*.

Censoring

Censoring is a form of missing data problem in which time to event is not observed for reasons such as termination of study before all recruited subjects have shown the event of interest or the subject has left the study prior to experiencing an event. Censoring is common in survival analysis.

Right Censored

Suppose you're conducting a study on pregnancy duration. You're ready to complete the study and run your analysis, but some women in the study are still pregnant, so you don't know exactly how long their pregnancies will last.



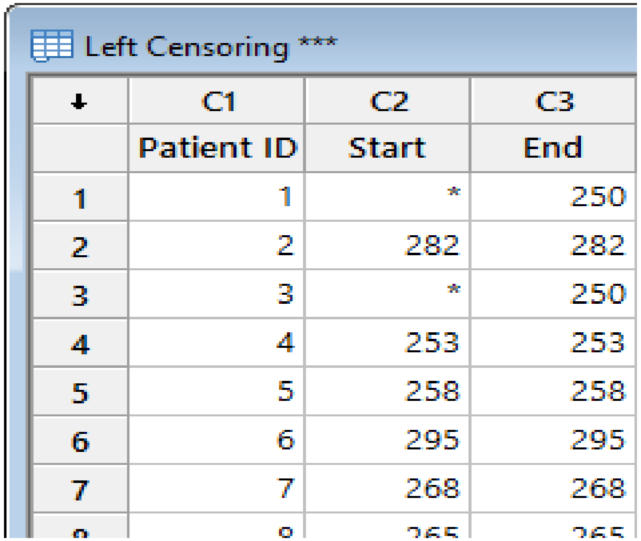
Right Censoring ***

↓	C1	C2	C3-T
	Patient ID	Days	Exact or Censored
1	1	286	Exact
2	2	279	Exact
3	3	269	Exact
4	4	256	Exact
5	5	293	Censored
6	6	262	Exact
7	7	285	Censored
8	8	278	Exact

These observations would be *right-censored*. The "failure," or birth in this case, will occur after the recorded time.

Left Censored

Now suppose you survey some women in your study at the 250-day mark, but they already had their babies. You know they had their babies before 250 days, but don't know *exactly* when.



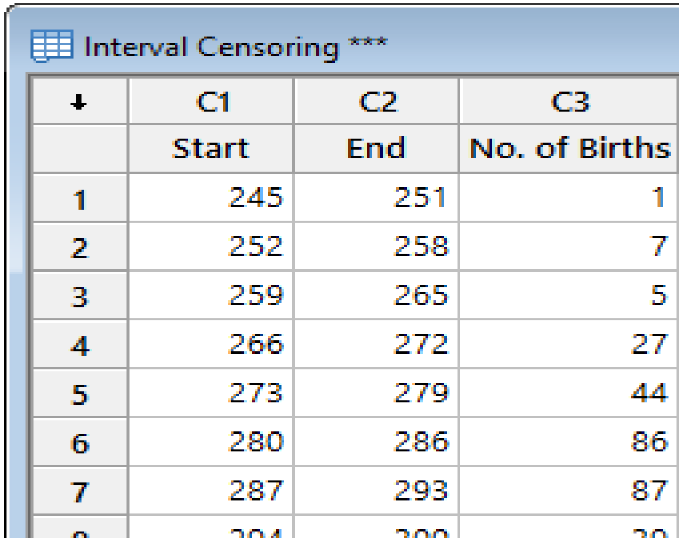
Left Censoring ***

↓	C1	C2	C3
	Patient ID	Start	End
1	1	*	250
2	2	282	282
3	3	*	250
4	4	253	253
5	5	258	258
6	6	295	295
7	7	268	268
8	8	265	265

These are therefore *left-censored* observations, where the "failure" occurred before a particular time.

Interval Censored

If we don’t know exactly when some babies were born but we know it was within some interval of time, these observations would be *interval-censored*. We know the “failure” occurred within some given time period

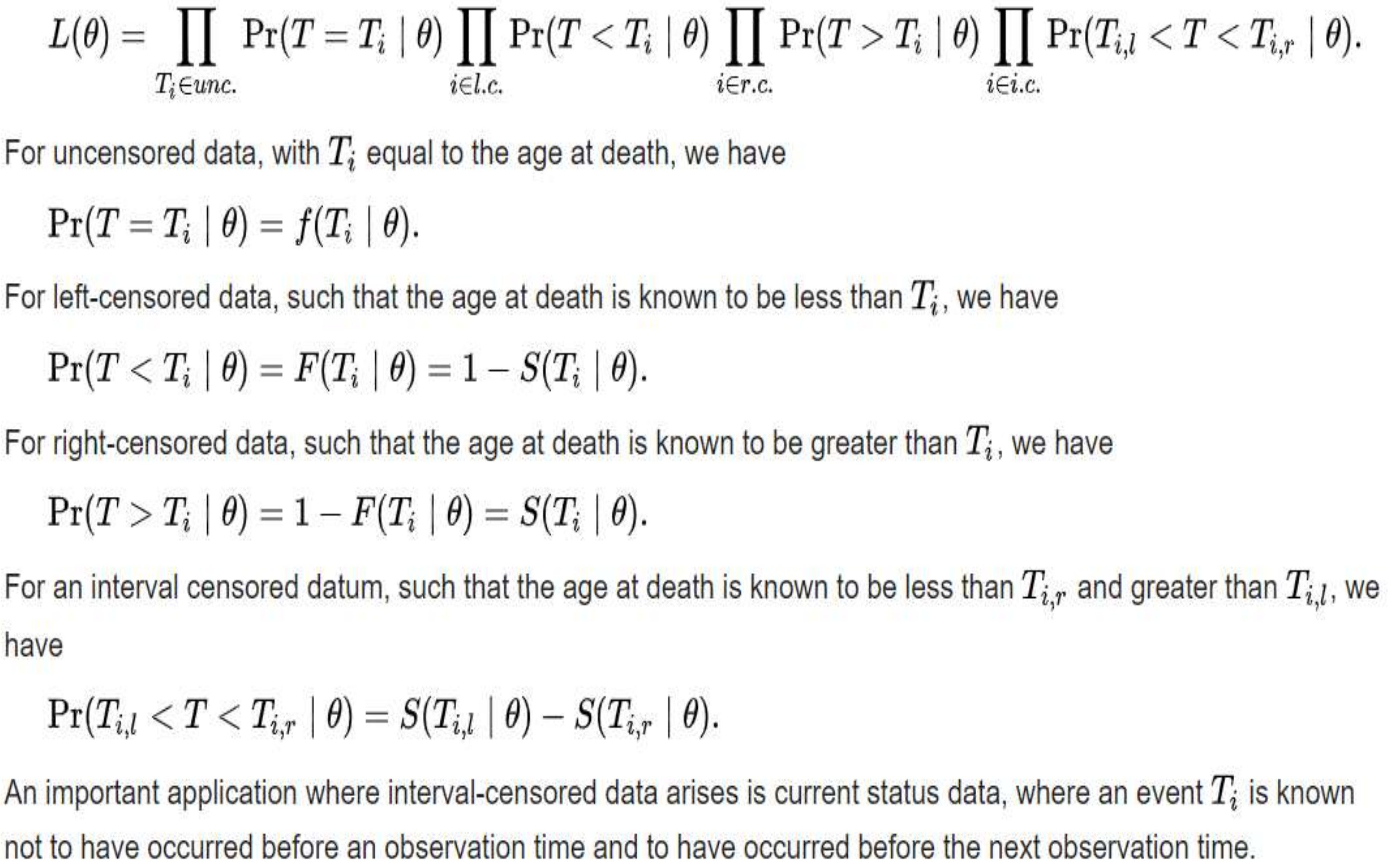


↓	C1	C2	C3
	Start	End	No. of Births
1	245	251	1
2	252	258	7
3	259	265	5
4	266	272	27
5	273	279	44
6	280	286	86
7	287	293	87

. For example, we might survey expectant mothers every 7 days and then count the number who had a baby within that given week.

Fitting parameters to data

Survival models can be usefully viewed as ordinary regression models in which the response variable is time. However, computing the likelihood function (needed for fitting parameters or making other kinds of inferences) is complicated by the censoring. The likelihood function for a survival model, in the presence of censored data, is formulated as follows. By definition the likelihood function is the conditional probability of the data given the parameters of the model. It is customary to assume that the data are independent given the parameters. Then the likelihood function is the product of the likelihood of each datum. It is convenient to partition the data into four categories: uncensored, left censored, right censored, and interval censored. These are denoted "unc.", "l.c.", "r.c.", and "i.c." in the equation below.



$$L(\theta) = \prod_{T_i \in unc.} \Pr(T = T_i \mid \theta) \prod_{i \in l.c.} \Pr(T < T_i \mid \theta) \prod_{i \in r.c.} \Pr(T > T_i \mid \theta) \prod_{i \in i.c.} \Pr(T_{i,l} < T < T_{i,r} \mid \theta).$$

For uncensored data, with T_i equal to the age at death, we have

$$\Pr(T = T_i \mid \theta) = f(T_i \mid \theta).$$

For left-censored data, such that the age at death is known to be less than T_i, we have

$$\Pr(T < T_i \mid \theta) = F(T_i \mid \theta) = 1 - S(T_i \mid \theta).$$

For right-censored data, such that the age at death is known to be greater than T_i, we have

$$\Pr(T > T_i \mid \theta) = 1 - F(T_i \mid \theta) = S(T_i \mid \theta).$$

For an interval censored datum, such that the age at death is known to be less than $T_{i,r}$ and greater than $T_{i,l}$, we have

$$\Pr(T_{i,l} < T < T_{i,r} \mid \theta) = S(T_{i,l} \mid \theta) - S(T_{i,r} \mid \theta).$$

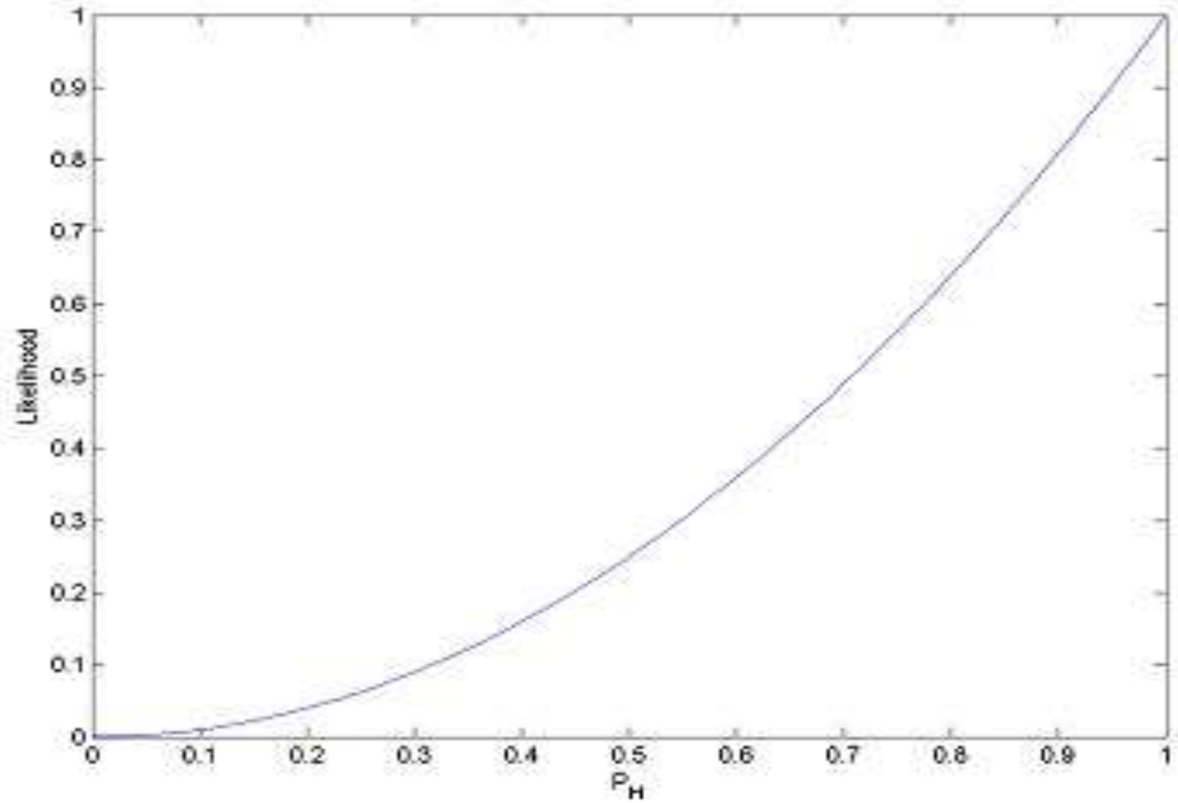
An important application where interval-censored data arises is current status data, where an event T_i is known not to have occurred before an observation time and to have occurred before the next observation time.

Likelihood Function

Let’s say you’re interested in creating a probability density function that represents binomial probabilities for getting a heads (or tails) in a single coin toss. You’re going to estimate the likelihood of getting heads from your data, so you run an experiment.

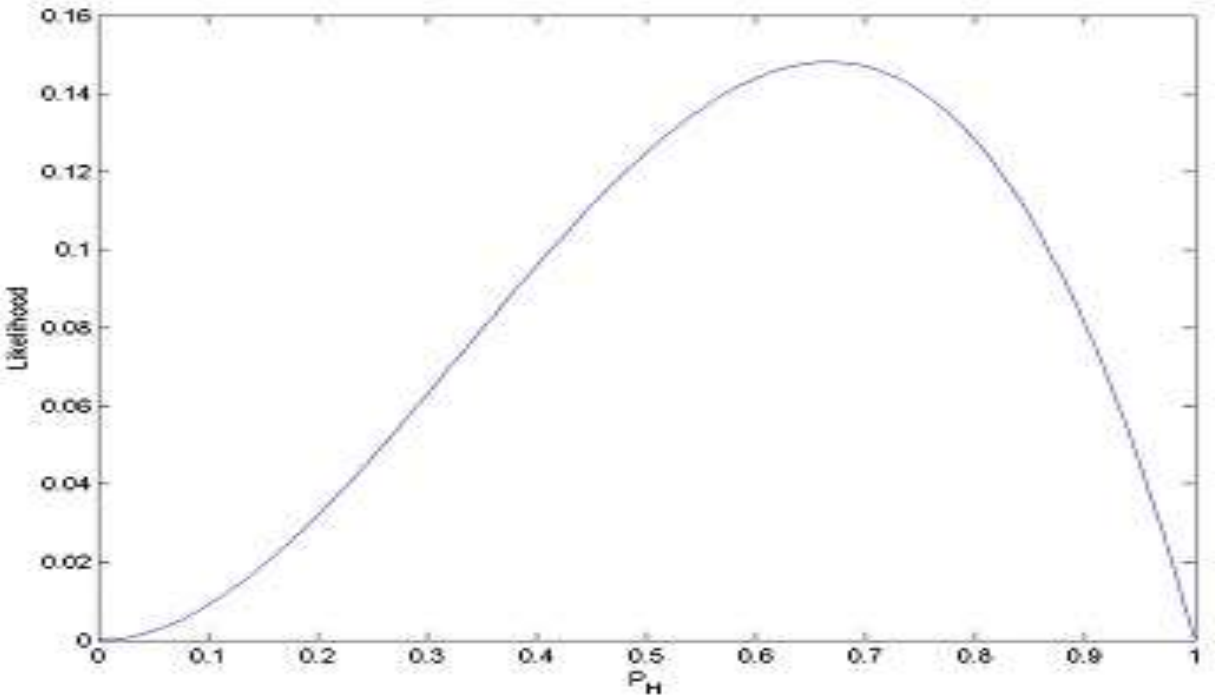
If you get two heads in a row, your likelihood function for the probability of a coin landing heads-up will look like this:

Likelihood Function



Likelihood Function

If you toss once more and get tails (making HHT), your function changes to look like this:



Likelihood Function

Although a likelihood function might look just like a probability density function, it’s fundamentally different. A probability density function is a function of x, your data point, and it will tell you how likely it is that certain data points appear. A likelihood function, on the other hand, takes the data set as a given, and represents the likeliness of different parameters for your distribution.

Ordinal Regression

Ordinal logistic regression (often just called 'ordinal regression') is used to predict an ordinal dependent variable given one or more independent variables. ordinal regression can also use interactions between independent variables to predict the dependent variable.

For example, you could use ordinal regression to predict the belief that "tax is too high" (your ordinal dependent variable, measured on a 4-point Likert item from "Strongly Disagree" to "Strongly Agree"), based on two independent variables: "age" and "income". Alternately, you could use ordinal regression to determine whether a number of independent variables, such as "age", "gender", "level of physical activity" (amongst others), predict the ordinal dependent variable, "obesity", where obesity is measured using three ordered categories: "normal", "overweight" and "obese".