Interval and Double Censoring

David M. Rocke

November 19, 2024

Interval Censoring

Interval censored data are when each individual's time to event is described only as within a time interval $(L, R]$ rather than at a specific time. The time R can be `Inf` or `NA`, meaning that the event is right censored at L. Similarly, if L is `-Inf` or `NA`, this means that the event is left censored at R. This type of data naturally arises from medical studies with periodic follow-up. If the event is known to have occurred by the time of an exam at time R, but had not occurred at the previous exam at time L, then we know it occurred in the interval $(L, R]$, including R and excluding L.

Breast Cosmesis Data

Finkelstein, D.M., and Wolfe, R.A. (1985). A semiparametric model for regression analysis of interval-censored failure time data. Biometrics 41: 731-740.

Interval-censored data arise naturally when the response times come from a medical study in which there is a periodic follow-up. An individual who is monitored weekly or monthly for response may miss visits and return in a changed state.

In the example considered here, a retrospective study was carried out to compare early breast cancer patients who had been treated with primary radiation therapy and adjuvant chemotherapy to those treated with radiotherapy alone with respect to the cosmetic effects of their treatment (Beadle et al., 1984a, 1984b). Excisional biopsy followed by irradiation is becoming an increasingly practiced alternative to mastectomy. Since the primary reason for avoiding mastectomy is the enhanced cosmetic outcome, it is of considerable importance to document the cosmetic results of various treatments.

Adjuvant chemotherapy improves the relapse-free and overall survival in at least some subgroups of patients treated initially by mastectomy. However, there is experimental and clinical evidence which indicates that chemotherapy enhances the acute response of normal tissue to radiation treatment. Acute skin reactions are worse when adjuvant chemotherapy is administered in conjunction with either postoperative radiation or primary radiation treatment for breast cancer. The long-term impact of adjuvant chemotherapy on the radiation response of the breast is uncertain.

The objective of this analysis is to compare the patients who received adjuvant chemotherapy following the initial radiation treatment ($X = 1$) to those who received only the (comparable dose of) radiation treatment ($X = 0$), to determine whether chemotherapy affects the rate of deterioration of the cosmetic state. In this study, patients were seen at clinic visits every 4 to 6 months. With increasing time after completion of primary irradiation treatment, and for those patients who were geographically remote, follow-up intervals were often longer.

Physicians recorded, on a scale of 0 to 3 (none, minimal, moderate, severe) the cosmetic appearance of the patient with respect to breast edema, telangiectasia (spider veins), breast retraction, and the overall cosmetic result. Breast retraction was highly correlated with a negative overall cosmetic appearance, and was one of the least subjective of the endpoints that were followed. Therefore, we chose to compare the effect of the two-treatment regimen on the time until cosmetic deterioration, as determined by the appearance of breast retraction.

The authors of this paper received the data from the authors of the two Beadle (1984) et al. papers (Finkelstein being also at Harvard/Dana Farber), and the subset of the data used in the Finkelstein paper is the data set `bcos` in the package `interval`. A similar data set in 1.18 in the text is `bcdeter`, which differs in some mostly minor respects from the first mentioned data set, and is described as originating with the two Beadle papers (which do not contain the data nor a link to it because it was 1984).The Finkelstein paper contains the data as given in `bcos`. Both data sets consist of a left, or lower time point, a right or higher time point and a treatment (radiation or chemorads). We will use `bcdeter`. Note that an interval for which the right side is infinity or missing means that the patient was censored at the left side of the interval.

Survival Function Estimation

Similar procedures are used for data with left censoring, interval censoring, and double (left and right) censoring. These procedures due to Turnbull, are nonparametric maximum likelihood, in the way the the Kaplan-Meier and Nelson-Aalen estimaters are for right censored data, but there is no closed form solution. We can apply this algorithm to the radiation and chemorads data separately.

The algorithm as described in KM begins (for the 46 radiation patients) by making a sorted list of all unique times (in months since radiation treatment) that occur on either side of an interval. For the start of the algorithm, we take $1/46$ from each patient and divide it equally among all times that are within the the $(lower, upper]$ interval. There is also another method that first reduces the intervals that need to be considered, but we will first look at the equivalent calculation as described in KM.

```
> install.packages("interval")
> if (!requireNamespace("BiocManager", quietly = TRUE))
+     install.packages("BiocManager")
> BiocManager::install("Icens")
> library(interval)
> library(KMsurv)
> data(bcdeter)
> summary(bcdeter)
     lower             upper             treat
 Min.   : 0.00   Min.   : 5.00   Min.   :1.000
 1st Qu.:11.00   1st Qu.:15.25   1st Qu.:1.000
 Median :18.00   Median :24.00   Median :2.000
 Mean   :22.34   Mean   :24.93   Mean   :1.516
 3rd Qu.:34.00   3rd Qu.:34.00   3rd Qu.:2.000
 Max.   :48.00   Max.   :60.00   Max.   :2.000
                 NA's   :37
> times <- with(bcdeter,sort(unique(c(lower[treat==1],
   upper[treat==1]))))
> times
 [1]  0  4  5  6  7  8 10 11 12 14 15 16 17 18 19 22 24 25
     26 27 32 33 34 35 36 37 38 40 44 45 46 48
```

```
> times
 [1]  0  4  5  6  7  8 10 11 12 14 15 16 17 18 19 22 24 25
     26 27 32 33 34 35 36 37 38 40 44 45 46 48
> head(bcdeter)
  lower upper treat
1     0     5     1
2     0     7     1
3     0     8     1
4     4    11     1
5     5    11     1
6     5    12     1
```

For example, patient 4 generates 1/6 of 1/46 to the six times 5, 6, 7, 8, 10, and 11 and patient 3 contributes 1/5 of 1/46 to the five times 4, 5, 6, 7 and 8. We can use that to derive a first approximation to the probability of an event at each time. That generates an estimate of the number of deaths at each time, then the number at risk, and an updated set of probabilities, etc.

Toy Example

Suppose we have four subjects that are interval censored as follows:

left	right
0	5
0	7
0	8
6	10

so that the possible times are 0, 5, 6, 7, 8, 10. These are the subjects in `bcdeter` in the radiation group with right time less than or equal to 10. In this case, there are no censored intervals (with right = infinity or NA) so the survival function would be just the 1 - CDF if the exact times of events were known.

left	right	0	5	6	7	8	10	weight
0	5		1/4					1/4
0	7		1/12	1/12	1/12			1/4
0	8		1/16	1/16	1/16	1/16		1/4
6	10				1/12	1/12	1/12	1/4
			19/48	7/48	11/48	7/48	4/48	1

Each patient has an event (in this case no censoring) and there are four patients, so each has a probability fraction of 1/4 to distribute among possible event times. The bottom row are the initial estimates of the probability that an individual would die at each of the times heading the column. We can iterate to find improved estimates because these are based on the assumption that an interval censored event is equally likely to occur at each time point within the interval, but the overall estimated event chances of each time point differ.

left	right	0	5	6	7	8	10	weight
0	5		1/4					1/4
0	7		1/12	1/12	1/12			1/4
0	8		1/16	1/16	1/16	1/16		1/4
6	10				1/12	1/12	1/12	1/4
			19/48	7/48	11/48	7/48	4/48	1

The weight per patient stays the same, but the probability of the patient/time point combination will now not be distributed equally, but in proportion to the overall probabilities of the time points. The number of events at time 5 from the first patient is 1. The number of events at time 5 from the second patient is $19/(19+7+11) = 19/37$. And the third patient adds $19/(19+7+11+7)$. The total expected events at time point 5 is then the sum of these three number, which is 1.945, and the estimated probability of an event at time 5 is $1.945/4 = 0.486$. Compare this to $19/48 = 0.396$

The expected event contributions at time 5 are 19/19, 19/37, and 19/44. These need to be divided by the number of patients to get the estimated probabilities of an event in each of the three cases, which is 0.250, 0.1284, and 0.1080. The sum of these, 0.4864, is the new estimate of the probabiity of an event at time 5.

Initial Iterate

left	right	0	5	6	7	8	10	weight
0	5		0.2500					1/4
0	7		0.0833	0.0833	0.0833			1/4
0	8		0.0625	0.0625	0.0625	0.0625		1/4
6	10				0.0833	0.0833	0.0833	1/4
			0.3958	0.1458	0.2292	0.1458	0.0833	1

Second Iterate

left	right	0	5	6	7	8	10	weight
0	5		0.2500					1/4
0	7		0.1284	0.0473	0.0743			1/4
0	8		0.1080	0.0398	0.0625	0.0398		1/4
6	10				0.1250	0.0795	0.0455	1/4
			0.4864	0.0871	0.2618	0.1193	0.0455	1

Second Iterate

left	right	0	5	6	7	8	10	weight
0	5		0.2500					1/4
0	7		0.1284	0.0473	0.0743			1/4
0	8		0.1080	0.0398	0.0625	0.0398		1/4
6	10				0.1250	0.0795	0.0455	1/4
			0.4864	0.0871	0.2618	0.1193	0.0455	1

Final Iterate

left	right	0	5	6	7	8	10	weight
0	5		0.250					1/4
0	7		0.125	0	0.125			1/4
0	8		0.125	0	0.125	0		1/4
6	10				0.250	0	0	1/4
			0.5	0	0.5	0	0	1

```
> summary(icfit(c(0,0,0,6),c(5,7,8,10)))
  Interval Probability
1    (0,5]         0.5
2    (6,7]         0.5
```

left	right	(0, 5]	(6, 7]	weight
0	5	0.250		1/4
0	7	0.125	0.125	1/4
0	8	0.125	0.125	1/4
6	10		0.250	1/4
		0.5	0.5	1

This is one syntax for estimating a survival function from interval censored data. Note that, instead of a list of six times, 0, 5, 6, 7, 8, 10, there are only two intervals. With two intervals, the calculation is done in one iteration.

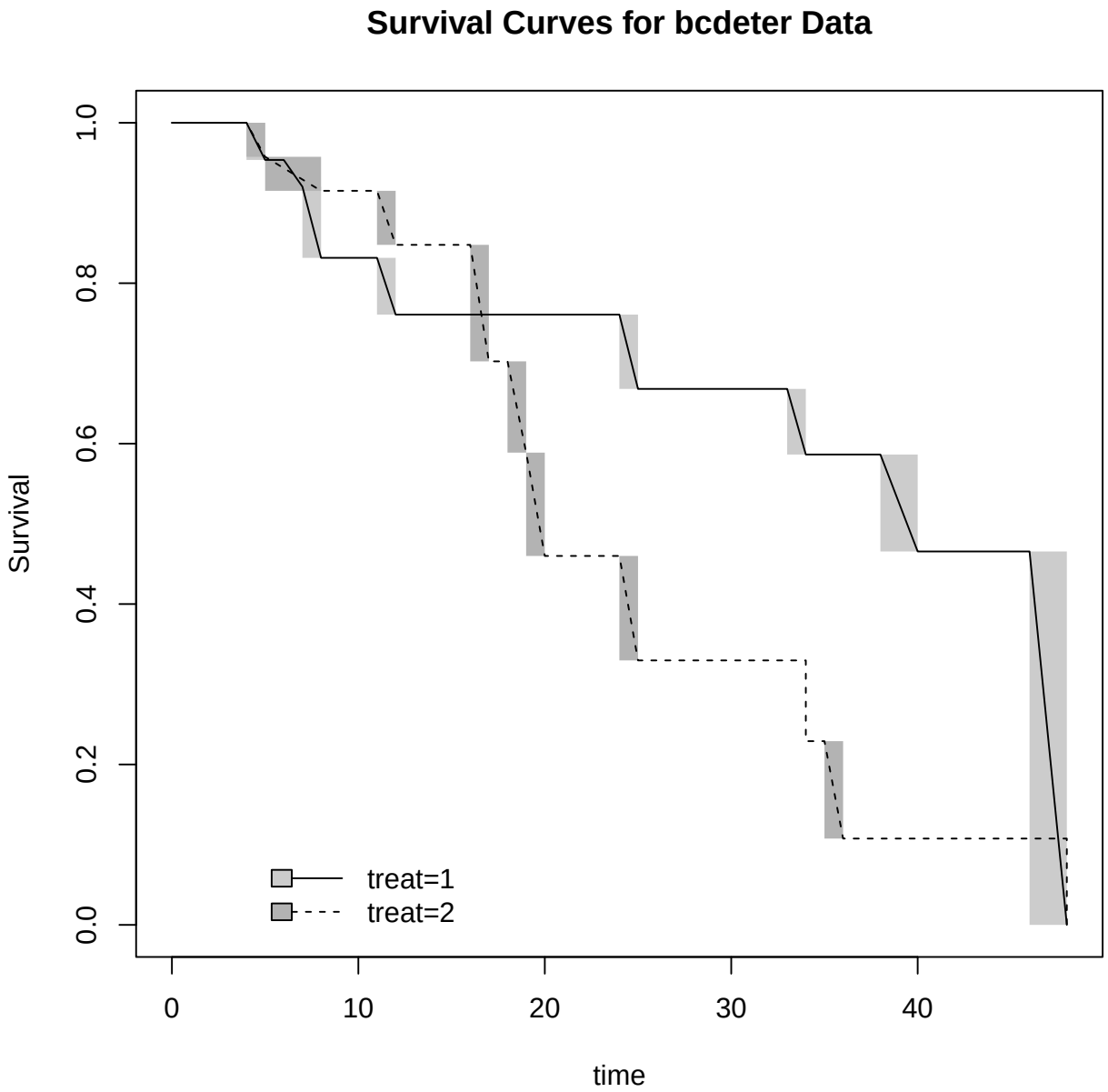
```
> bc.out <- icfit(Surv(lower,upper,type="interval2")~treat,data=bcdeter)
> print(summary(bc.out))
treat=1:
  Interval Probability
1    (4,5]      0.0463     #There are 32 times in the combined lower/upper vector
2    (6,7]      0.0334     #for the radiation only patients
3    (7,8]      0.0887     #but there are only 8 intervals in the output!
4  (11,12]      0.0708
5  (24,25]      0.0926
6  (33,34]      0.0818
7  (38,40]      0.1209
8  (46,48]      0.4656
treat=2:
   Interval Probability
1     (4,5]      0.0424
2     (5,8]      0.0424
3   (11,12]      0.0673
4   (16,17]      0.1453
5   (18,19]      0.1138
6   (19,20]      0.1288
7   (24,25]      0.1302
8   [34,34]      0.1007
9   (35,36]      0.1215
10  [48,48]      0.1076
```

```
plot1 <- function(){
  pdf("bc.pdf")
  plot(bc.out)
  dev.out()
}
```

```
> print(ictest(Surv(left,right,type="interval2")~treat,data=bcdeter))
        Asymptotic Logrank two-sample test (permutation form), Sun's scores

data:  Surv(lower, upper, type = "interval2") by treat
Z = -2.8993, p-value = 0.00374
alternative hypothesis: survival distributions not equal

   n Score Statistic*
1 46        -10.06046
2 49         10.06046
* like Obs-Exp, positive implies earlier failures than expected
```


Survival Curves for bcdeter Data
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
time
treat=1
treat=2

Turnbull (1976)

Suppose that the data set consists of n independent observations X_i generated from a distribution F, with the proviso that all that is known of X_i is that it lies in a set A_i which we will assume to be an interval $(L_i, R_i]$. (The assumption in Turnbull is a little more general.) Let $\mathcal{L}$ be the set of all values L_i and $\mathcal{R}$ the set of values R_i and $\mathcal{B} = \mathcal{L} \cup \mathcal{R}$. We then construct all intervals $(q_j, p_j]$ with $q_j \in \mathcal{L}$, $p_j \in \mathcal{R}$ for which $\nexists b \in \mathcal{B}$ such that $q_j < b < p_j$. We order the intervals such that $q_1 \leq p_1 \leq q_2 \leq \cdots \leq q_m \leq p_m$ and let $C = \cup_{j=1}^{m}(q_j, p_j]$.

Turnbull (1976)

Then

1. any CDF which increases anywhere outside the set C cannot be the NPMLE of F and
2. The likelihood is independent of the behavior of F inside the intervals $(q_j, p_j]$.

This means that all we can determine is the probability content of each interval in C and the CDF (and survival function) is flat in between these intervals.

Klein and Moeschberger			`icfit()` in `interval`	
Interval	$S(t)$	Δ	Interval	Probability
0–4	1.000			
		0.046	$(4, 5]$	0.0463
5–6	0.954			
		0.034	$(6, 7]$	0.0334
7	0.920			
		0.088	$(7, 8]$	0.8887
8–11	0.832			
		0.071	$(11, 12]$	0.0708
12–24	0.761			
		0.093	$(24, 25]$	0.0926
25–33	0.668			
		0.082	$(33, 34]$	0.0818
34–38	0.586			
		0.119	$(38, 40]$	0.1209
40–48	0.467			
		0.467	$(46, 48]$	0.4656
≥ 48	0.000			

KM method had 32 times by 46 patients and took 305 iterations. Interval method had 8 intervals by 46 patients and took 137 iterations. Error estimates for interval were around 10^{-6}, and the apparent error in the KM values was more like 10^{-3}.

The full set of times for the radiation only data is

left = 0 4 5 6 7 11 15 17 18 19 22 24 25 26 27 32 33 34 36 37 38 40 45 46
right = 5 7 8 10 11 12 14 15 16 18 25 26 34 35 37 40 44 48

No interval can begin with 0 because the shortest one is (0, 5] and that contains 4.
Only interval beginning with 4 is (4, 5].
No interval with 5 since (5, 7] contains 6.
Next is (6, 7], (7, 8], and (11, 12].
The interval (17, 18] exists for the initial iterates, but converges to 0.
No interval with 18 because (18,25] contains 19.
No interval with 19, 22.
Next is (24, 25]
...
last interval is (46, 48]

The initial iteration has 13 intervals, of which 5 drop out with estimated probability 0. After 10 iterations, there are 10 intervals, after 20, there are 9 intervals, and after 32 iterations the final 8.

Toy Example 2

Suppose we have six subjects that are interval censored as follows:

lower = left	upper = right
0	5
0	7
0	8
4	11
5	11
6	10

These are the subjects in `bcdeter` in the radiation group with right time less than or equal to 11. Possible inner intervals from Turnbull are $(4, 5]$ and $(6, 7]$. The list of times for the KM method is 0, 4, 5, 6, 7, 8, 10, 11, comprising seven possible times.

interval	$(4,5]$	$(6,7]$	weight
$(0,5]$	$1/6$		$1/6$
$(0,7]$	$1/12$	$1/12$	$1/6$
$(0,8]$	$1/12$	$1/12$	$1/6$
$(4,11]$	$1/12$	$1/12$	$1/6$
$(5,11]$		$1/6$	$1/6$
$(6,10]$		$1/6$	$1/6$
	$5/12$	$7/12$	1

interval	$(4,5]$	$(6,7]$	weight
$(0,5]$	$1/6$		$1/6$
$(0,7]$	$5/72$	$7/72$	$1/6$
$(0,8]$	$5/72$	$7/72$	$1/6$
$(4,11]$	$5/72$	$7/72$	$1/6$
$(5,11]$		$1/6$	$1/6$
$(6,10]$		$1/6$	$1/6$
	$3/8$	$5/8$	1

interval	$(4,5]$	$(6,7]$	weight
$(0,5]$	$1/6$		$1/6$
$(0,7]$	$1/18$	$1/9$	$1/6$
$(0,8]$	$1/18$	$1/9$	$1/6$
$(4,11]$	$1/18$	$1/9$	$1/6$
$(5,11]$		$1/6$	$1/6$
$(6,10]$		$1/6$	$1/6$
	$1/3$	$2/3$	1

This is the converged array, which took 20 iterations with error 10^{-6} although the table above is exact, meaning that iterating leads to the exact same array.

Regression with Interval Censored Data

The R package `icenReg` can fit to interval censored data the NPMLE as with `interval`, semi-parametric models with proportional hazards or proportional odds, and fully parametric models in those two cases as well as accelerated failure time. The distributional choices are

- Exponential
- Gamma
- Weibull
- log normal
- log logistic
- generalized gamma

Double Censored Data

Double censored data have some observations left censored, some right censored, and some exact. In a way, this is a subset of interval censored data with the left side of the interval for left-censored observations and the right side of the interval for right-censored observations written as NA or $\pm$Inf. Using the data on first-time use of marijuana in KM 1.17, we can analyze it with the interval package.

KM 1.17

Turnbull and Weiss (1978) report part of a study conducted at the Stanford-Palo Alto Peer Counseling Program (see Hamburg et al. [1975] for details of the study). In this study, 191 California high school boys were asked, "When did you first use marijuana?" The answers were the exact ages (uncensored observations); "I never used it," which are right-censored observations at the boys' current ages; or "I have used it but can not recall just when the first time was," which is a left-censored observation (see section 3.3). Notice that a left-censored observation tells us only that the event has occurred prior to the boy's current age. The data is in Table 1.8. This data is used in section 5.2 to illustrate the calculation of the survival function for both left- and right-censored data, commonly referred to as doubly censored data.

Table 1.8

Marijuana use in high school boys			
Age	Exact	Not Yet	Started earlier
10	4	0	0
11	12	0	0
12	19	2	0
13	24	15	1
14	20	24	2
15	13	18	3
16	3	14	2
17	1	6	3
18	0	0	1
>18	4	0	0

```
> print(mj)
   left right weight
1    10    10      4
2    11    11     12
3    12    12     19
4    13    13     24
5    14    14     20
6    15    15     13
7    16    16      3
8    17    17      1
9    19    19      4
10   12    NA      2
11   13    NA     15
12   14    NA     24
13   15    NA     18
14   16    NA     14
15   17    NA      6
16   NA    13      1
17   NA    14      2
18   NA    15      3
19   NA    16      2
20   NA    17      3
21   NA    18      1
```

These are the data from 1.17 where weight is the number of observations of the kind. Exact observations are represented by equal left and right ages. Left-censored observations have NA on the left and right-censored observations have NA on the right. The code below replicates each left-right pair the number of times listed as weight.

```
mrows <- function(df){
  mvec <- c(0,0) #temporary first row
  nrows <- dim(df)[1]
  for (i in 1:nrows){
    wt <- df[i,3] #the number of times to replicate
    newrow <- df[i,1:2]
    for (j in 1:wt){
      mvec <- rbind(mvec,newrow) #add row
    }
  }
  mvec <- mvec[-1,] #remove temporary first row
  return(mvec)
}

> mj2 <- mrows(mj)
```

```
> print(mj)
   left right weight
1    10    10      4
2    11    11     12
3    12    12     19
4    13    13     24
5    14    14     20
6    15    15     13
7    16    16      3
8    17    17      1
9    19    19      4
10   12    NA      2
11   13    NA     15
12   14    NA     24
13   15    NA     18
14   16    NA     14
15   17    NA      6
16   NA    13      1
17   NA    14      2
18   NA    15      3
19   NA    16      2
20   NA    17      3
21   NA    18      1
```

```
> head(mj2,20)
    left right
2     10    10    #four of these
3     10    10
4     10    10
5     10    10
21    11    11    #12 of these
22    11    11
23    11    11
24    11    11
25    11    11
26    11    11
27    11    11
28    11    11
29    11    11
210   11    11
211   11    11
212   11    11
31    12    12    #19 of these
32    12    12
33    12    12
34    12    12
```

```
> mj2.fit <- icfit(Surv(left,right,type="interval2")~1,data=mj2)
> summary(mj2.fit)
  Interval Probability
1  [10,10]      0.0235
2  [11,11]      0.0705
3  [12,12]      0.1116
4  [13,13]      0.1431
5  [14,14]      0.1355
6  [15,15]      0.1236
7  [16,16]      0.0467
8  [17,17]      0.0375
9  [19,19]      0.3079
```

Note that age 18 is omitted (there were no events at age 18 and one left censored observation.

Here is the comparison with the results from `icfit()` with the results reported in KM Table 5.3. The error in the `icfit()` results is reported to be 10^{-6}

KM Table 5.3			`icfit()`
Age	$S(Age)$	Δ	Probability
0	1.000		0.0000
10	0.977	0.023	0.0235
11	0.906	0.071	0.0705
12	0.794	0.112	0.1116
13	0.651	0.143	0.1431
14	0.516	0.135	0.1355
15	0.392	0.124	0.1236
16	0.345	0.047	0.0467
17	0.308	0.037	0.0375
18	0.308	0.000	0.0000
$>$ 18	0.000	0.308	0.3079

One issue with the interpretation of this study is the > 18 group. Exact times were entered for the 4 students that were in this group ($4/191 = 2.1\%$) but most of the 30.8% probability content represents students who reported at age 17 or younger that they had not yet used marijuana. Some of those (about 29%) will never use marijuana and some will try it for the first time at ages greater than 18. The right-censored group (79 students or 41.4% of the sample) includes about 10% probability mass imputed to ages under 18 as well.

Although this analysis is the best we can do with the sample, we can only be certain (subject to correct reporting by the students) that about 59% of them have already tried it. In addition, there might be concerns about what population these students represent and how they were selected.