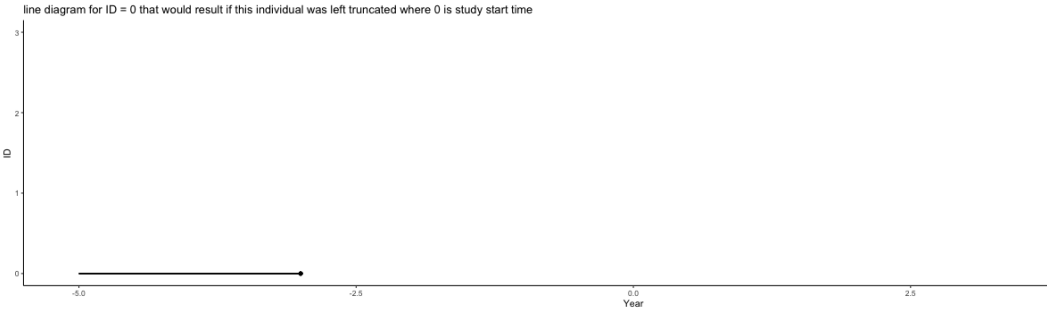
Section 1 Assignment

Question 1) Using the language of "censoring" and/or "truncation" (left, right, and/or interval), explain why a prospective cohort study is often seen as higher quality than a retrospective cohort study.

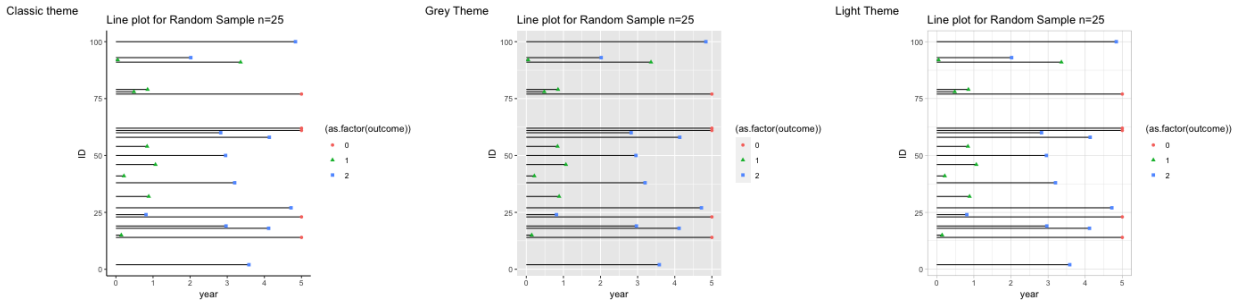
Prospective cohort studies are generally designed with specific collection methods in mind so are inherently less susceptible to censoring as the data collection methods are specific to the study vs retrospective cohort studies which are limited by the original study design. Prospective cohort studies are less susceptible to truncation as participants in retrospective studies must have the outcome at the start of the study, hypothetically people who died from the outcome of interest would be left trunkated data. Ergo, the risk estimate would not be accurate. Prospective studies are more accurate in their understanding of risk due to following diseasee free people over time. ***

Question 2) Using Figure 1 from the Section 1 notes, draw the line diagram for ID = 0 that would result if this individual was left truncated.



***

Question 3) For a randomly selected subset of 25 observations out of the N = 100 observations in the "section1_cohort.csv" data, fit a line plot for the time-to-event outcome using ggplot. With this line plot, explore different themes. Pick your two favorite themes and compare them to the classic theme (i.e., `theme_classic()`). Title each plot with the name of the theme used. Plot each of the three themes in a grid with one row and three columns. Save the plot as a pdf or png file to directory. Include this plot in your homework output and provide an informative caption with the plot.



***

Question 4) Please do a basic exploratory analysis of the "section1_cohort.csv" dataset. No more than 1/2 page of results. Provide results for the exposure, the confounder, and the outcome.

	no outcome		outcome 1		outcome 2		Ove
	no confounder	confounder	no confounder	confounder	no confounder	confounder	no confounder
	(N=12)	(N=12)	(N=8)	(N=17)	(N=14)	(N=37)	(N=34)
section1_expl$exposure1							
no exposure	8 (66.7%)	10 (83.3%)	5 (62.5%)	2 (11.8%)	6 (42.9%)	16 (43.2%)	19 (55.9%)
exposure	4 (33.3%)	2 (16.7%)	3 (37.5%)	15 (88.2%)	8 (57.1%)	21 (56.8%)	15 (44.1%)

Of the sample (n=100), there were 24 subjects with no outcome, 25 subjects with outcome 1, and 51 subjects with outcome 2. The risk of outcome 1 was .33 among those who were exposed. The risk of outcome 2 among those exposed was .54 The risk of getting either outcome 1 or 2 among those exposed was .88

Question 5) Describe, in words, the interpretation of the CDF:

$$F(t) = P(T \leq t)$$

AND the survival function:

$$S(t) = P(T > t)$$

if T represents age at death from all causes, and t represents 64 years of age.

F(t) represents the probability of death *before* 64 years of age. S(t) represents the probability of death *after* age 64. ***

Question 6) Using the first five observations from the synthetic data in Table 1 of the course notes, write out (but do not solve for) the terms for the Kaplan-Meier estimator $\hat{S}(t) = \prod_{k \in t_k \leq t}(1 - d_k/n_k)$. Assume that the total population at risk includes all 10 observations in Table 1.

$$\hat{S}(t) = (1 - 1/10)(1 - 1/9)(1 - 1/8)(1 - 1/6)$$

we excluded the one withdraw which is why we go from 8 -> 6 in denominator

---

Question 7) Please explain the difference between the `Surv()` function, and the `survfit()` function in the survival package.

surv creates a survival object

the *survfit* function uses the object created in the surv function to create the curve while handling left truncation and right censoring automatically

---

Question 8) Refer to Figure 3 in the Section 1 course notes. Note that the dashed blue line in Figure 3 is from the Kaplan-Meier estimator, while the solid black line is from the simple calculation shown in the equations above the Figure (on page 10). Why don't these figures align exactly? - calculates average vs km actual start and stop time KM takes into account the actual start and stop time but the solid black line only calculates the average time spent which will generally be different than the actual total start and stop times especially when there are more significant outliers.

Question 9) Fit the `survfit()` function to the "section1_cohort.csv" data. Before you fit, be sure to re-code the outcome so that any non-zero event counts as an event (i.e., re-code outcome=2 to outcome=1). Examine the R object that you get from this fit. How many elements are in this object? What are the first six elements (describe them briefly, don't just provide their element names). Is there enough information in this object for you to determine the median survival time for the outcome? If so, what is the median survival time?

create dataset for plotting

```
## # A tibble: 3 x 2
##   outcome     n
##     <dbl> <int>
## 1       0    24
## 2       1    25
## 3       2    51
```

```
## # A tibble: 2 x 2
##   cs_outcome     n
##        <dbl> <int>
## 1          0    24
## 2          1    76
```

```
## [1] 1.634748
```

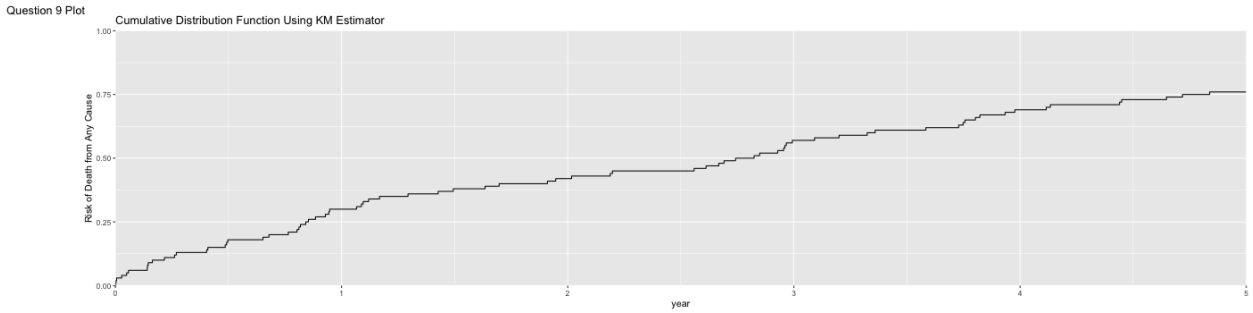
examine dataset

```
## Call: survfit(formula = Surv(time = start, time2 = stop, event = cs_outcome) ~
##     1, data = cohort_s1q9)
##
##        n events median 0.95LCL 0.95UCL
## [1,] 100     76   2.78    1.95    3.36
```

There are 17 elements in this object. The first six are as followed $n is the count of the population $time is the stop time $n.risk number of people at risk $n.event number of people who got the event $n.censor

number of people censored $surv percentage of people alive from the original population The median survial time is 1.63.

Question 10) Using the fit from Question 9, plot the cumulative distribution function (not the survival function) using the KM estimator. Interpret the curve assuming that the outcome is death from any cause and the time-scale is year on study.



Question 11) Referring to Figure 6 of the section 1 course notes, why is the cumulative risk represented by the dashed line higher than the cumulative risks represented by the solid black line, even though they are the same events?

The black line represents outcome = 1, dashed line represents the *cumulative cause specific risk* obtained by km estimator if we prevent outcome=2 from occurring. The KM estimator has a "redistribution to the right algorithm" as well as a censored individual in that graph specifically. Essentially, the algorithm redistributes the risk from censored participants to all of the remaining participants. Therefore, the KM estimator usually has a higher risk by the end of the study in the presence of censoring than the end of study risk of the empirical risk function.

Question 12) What is the main problem with using the cause-specific risk to understand the causal effects of exposures on outcomes of interest?

The main problem with using the cause-specific risk to understand the causal effects of exposures on the outcome of interest is that there is no clear way that the competing event is actually prevented (particularly

when the competing event is death due to any other cause). In most settings, it is impossible to prevent 100% the other event from occuring, so it's not generalizable to the 'real world.'

Question 13) Provide a single plot of the cause-specific and sub-distribution risk for "outcome = 1'' in the "section1_cohort.csv" using the Kaplan-Meier, Aalen-Johansen, and Gray's CIF estimators.

```
## List of 2
##  $ 1 1:List of 3
##   ..$ time: num [1:52] 0.00 5.02e-05 5.02e-05 3.03e-03 3.03e-03 ...
##   ..$ est : num [1:52] 0 0 0.01 0.01 0.02 0.02 0.03 0.03 0.04 0.04 ...
##   ..$ var : num [1:52] 0 0 0.0001 0.0001 0.000198 ...
##  $ 1 2:List of 3
##   ..$ time: num [1:104] 0 0.0282 0.0282 0.1636 0.1636 ...
##   ..$ est : num [1:104] 0 0 0.01 0.01 0.02 0.02 0.03 0.03 0.04 0.04 ...
##   ..$ var : num [1:104] 0 0 0.0001 0.0001 0.000198 ...
##  - attr(*, "class")= chr "cuminc"
```

```
## Rows: 100 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): ID, exposure, confounder, start, stop, outcome
##
## i Use 'spec()' to retrieve the full column specification for this data.
## i Specify the column types or set 'show_col_types = FALSE' to quiet this message.
```

light pink = Kaplan-Meier curve; dark pink = Gray's CIF Estimator; dashed line = Aalen-Johansen