Section 1 Assignment

Instructions (read carefully):

- Each student must submit one assignment as. Members of the same group can submit the same assignment.
- Each member of the same group will receive the same grade.
- Please put the name of each group member on the first page.
- Assignments must be done using RMarkdown.
- Submissions must include the .pdf file and the reproducible .rmd file used to do the homework. R code for all applied questions must be provided and be executable in the .rmd file.
- This assignment is due electronically through CANVAS on *?* at the beginning of class.

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.

The key to correctly answering this question is to recognize that prospective studies are less likely to suffer from left truncation. The reason is that in a prospective study, the start time is determined, and eligible individuals are identified as a result of this. On the other hand, with a retrospective study, the converse is usually the case: individuals are identified, and then the start time is determined. The problem with this is that individuals would have to make it at least to the point where they can be identified for inclusion in the retrospective study.

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

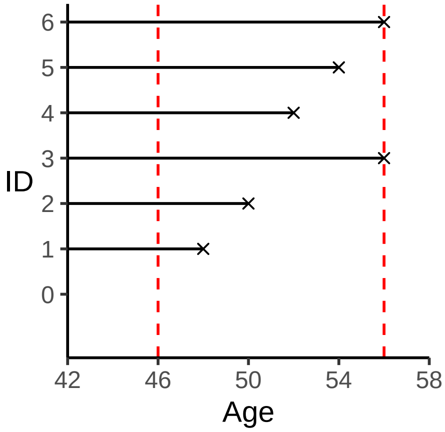


Figure 1: Line diagram with six hypothetical individuals who enter into a study at age 46 and who exit at age 56.

Question 3) 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.

This should be an easy question to answer. I am looking for basic descriptives for the exposure, the confounder, and the outcome, in tabular and/or graphical form.

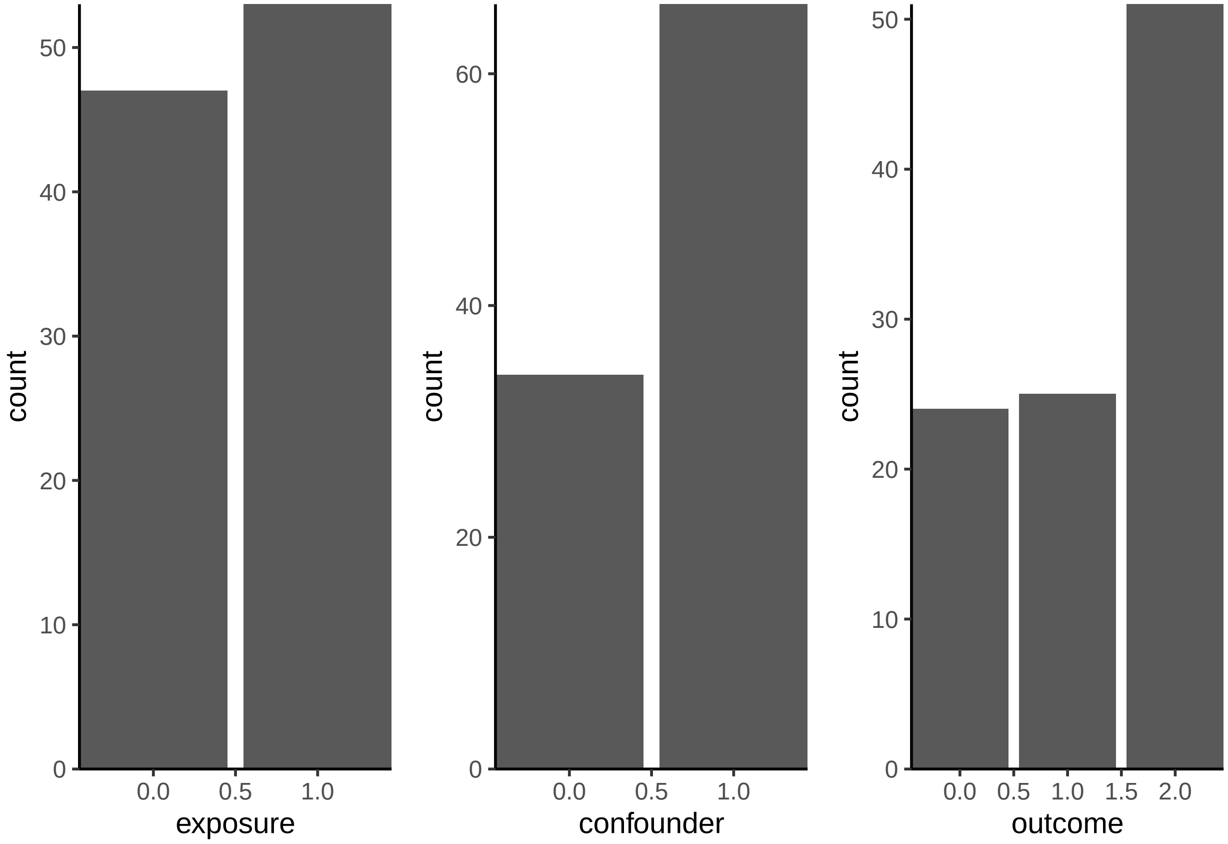
```
library(tidyverse)
library(gridExtra)

cohort <- read_csv("../../data/section1_cohort.csv")

thm <- theme_classic() +
  theme(
    legend.position = "top",
    legend.background = element_rect(fill = "transparent", colour = NA),
    legend.key = element_rect(fill = "transparent", colour = NA)
  )
theme_set(thm)

p1 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  geom_bar(aes(exposure))
p2 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  geom_bar(aes(confounder))
p3 <- ggplot(cohort) +
  scale_y_continuous(expand=c(0,0)) +
  scale_x_continuous(expand=c(0,0)) +
  geom_bar(aes(outcome))


grid.arrange(p1,p2,p3,nrow=1)
```



```
cohort %>%
  group_by(outcome) %>%
  summarize_at(vars(exposure,confounder,stop),mean)
```

```
## # A tibble: 3 x 4
##   outcome exposure confounder  stop
##     <dbl>    <dbl>      <dbl> <dbl>
## 1       0    0.25       0.5    5
## 2       1    0.72       0.68   1.16
## 3       2    0.569      0.725  2.28
```

Question 4) 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.

The interpretation for the CDF in the given context is the probability of the death from any cause occurring at time 64 years of age or before.

The interpretation of the survival function in the given context is the probability of death from any cause occurring after 64 years of age. (note: not at or after 64 years, but after 64 years).

To get full points for this question, the students must: 1) interpret F(t) and S(t) in the specified context (i.e., all cause mortality, 64 years of age); and 2) they must correctly (and precisely) interpret the equations.

Question 5) 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.

$$\hat{S}(t) = \prod_{k \in t_k \leq t}(1 - d_k/n_k) = \underset{\text{for } t=0.03}{\left(1 - \frac{1}{10}\right)} \times \underset{\text{for } t=0.49}{\left(1 - \frac{1}{9}\right)} \times \underset{\text{for } t=0.65}{\left(1 - \frac{1}{8}\right)} \times \underset{\text{for } t=2.61}{\left(1 - \frac{1}{6}\right)}$$

The key here is to recognize that the risk-set for the last term excludes the censored observation occuring in ID = 4.

Question 6) 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.

```r
library(tidyverse)
library(survival)

cohort <- read_csv("../../data/section1_cohort.csv") %>% mutate(outcome = as.numeric(outcome>0))

surv_obj <- survfit(Surv(time=start,time2=stop,event=outcome)~1,data=cohort)

names(surv_obj)
```

```
##  [1] "n"         "time"      "n.risk"    "n.event"   "n.censor"  "surv"
##  [7] "std.err"   "cumhaz"    "std.chaz"  "n.enter"   "type"      "logse"
## [13] "conf.int"  "conf.type" "lower"     "upper"     "call"
```

```r
str(surv_obj)
```

```
## List of 17
##  $ n        : int 100
##  $ time     : num [1:77] 5.02e-05 3.03e-03 5.52e-03 2.82e-02 4.97e-02 ...
##  $ n.risk   : num [1:77] 100 99 98 97 96 95 94 93 92 91 ...
##  $ n.event  : num [1:77] 1 1 1 1 1 1 1 1 1 1 ...
##  $ n.censor : num [1:77] 0 0 0 0 0 0 0 0 0 0 ...
##  $ surv     : num [1:77] 0.99 0.98 0.97 0.96 0.95 0.94 0.93 0.92 0.91 0.9 ...
##  $ std.err  : num [1:77] 0.0101 0.0143 0.0176 0.0204 0.0229 ...
##  $ cumhaz   : num [1:77] 0.01 0.0201 0.0303 0.0406 0.051 ...
##  $ std.chaz : num [1:77] 0.01 0.0142 0.0175 0.0203 0.0228 ...
##  $ n.enter  : num [1:77] 1.00e+02 8.28e-313 1.25e-312 1.89e-312 1.87e-312 ...
##  $ type     : chr "counting"
##  $ logse    : logi TRUE
##  $ conf.int : num 0.95
##  $ conf.type: chr "log"
##  $ lower    : num [1:77] 0.971 0.953 0.937 0.922 0.908 ...
##  $ upper    : num [1:77] 1 1 1 0.999 0.994 ...
##  $ call     : language survfit(formula = Surv(time = start, time2 = stop, event = outcome) ~ 1,
##  - attr(*, "class")= chr "survfit"
```

```
med_time <- tibble(time=surv_obj$time,risk=surv_obj$surv) %>% filter(round(risk,2)==0.50)

med_time
```

```
## # A tibble: 1 x 2
##    time  risk
##   <dbl> <dbl>
## 1  2.74 0.500
```

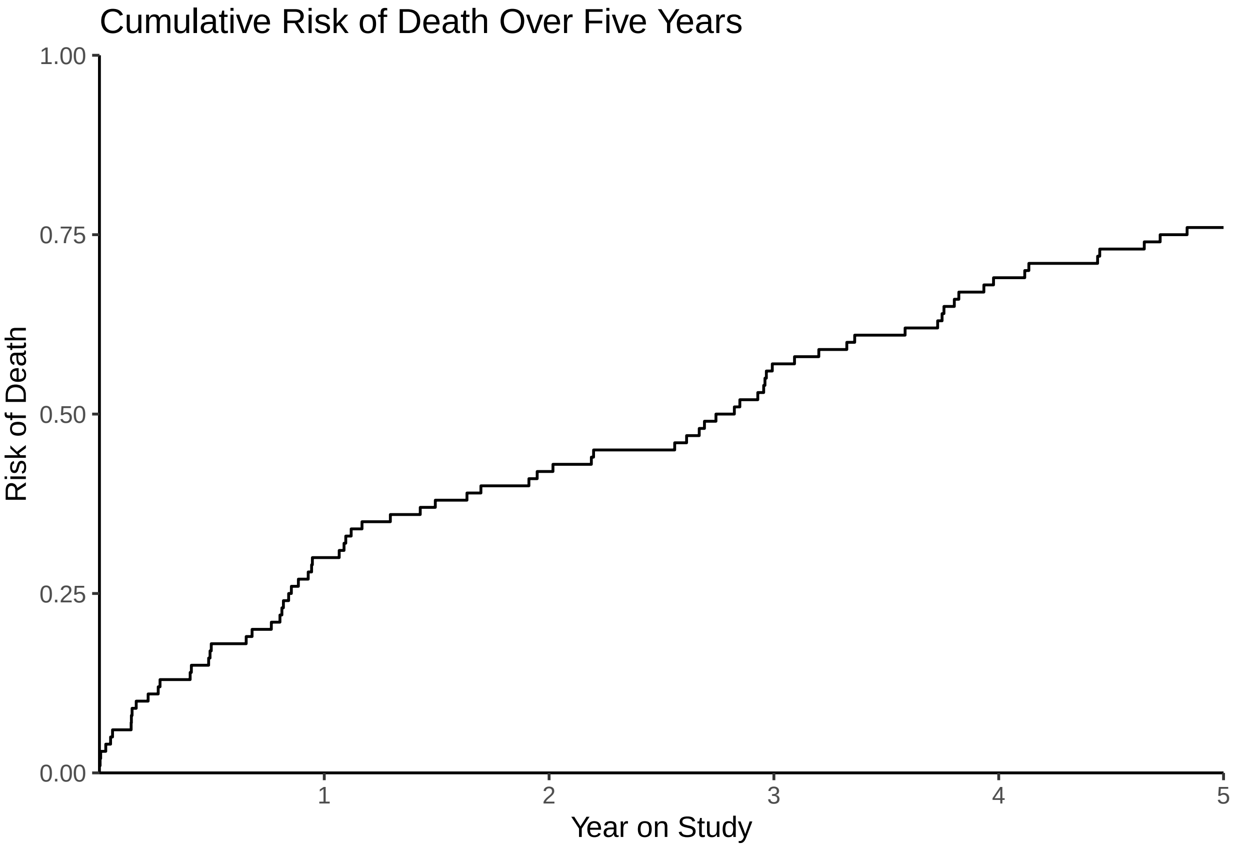
There are 17 elements in the object created by `survfit`. The first six elements are: n, time, n.risk, n.event, n.censor, surv. These are the original sample size, the unique event times, the number at risk at each event time, the number of events at each event time, the number of censored observations at each event time, and the estimated survival probability at each event time. There is, in fact, enough information to evaluate median survival time in this cohort, which is the time at which 50 percent of the sample survives. The median survival time is 2.74.

Some important elements to answer this question: the median survival time should be reported to 1 or 2 decimal places. Any more is unwarranted, and should probably result in loss of points.

Question 7) Using the fit from Question 6, 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.

```
plot_dat <- tibble(time = surv_obj$time,
                   risk = 1 - surv_obj$surv,
                   LCL = 1 - surv_obj$lower,
                   UCL = 1 - surv_obj$upper)

ggplot(plot_dat) +
  scale_y_continuous(expand=c(0,0), limits=c(0,1)) +
  scale_x_continuous(expand=c(0,0)) +
  ylab("Risk of Death") +
  xlab("Year on Study") +
  ggtitle("Cumulative Risk of Death Over Five Years") +
  geom_step(aes(x=time,y=risk))
```



This figure shows a steady increase in the risk of death from the start to the end of study. By year 3 on study, just over 50% of the sample died, and by the 5th year on study, 76% of the sample died.