Homework 1

P8108 - Survival Analysis

William Donovan

Question 1

```
library(tidyverse)
library(survival)
library(survminer)

# Load in Q1 data
q1_df = read_csv("data/Q1data_extracted.csv")
```

a. The MLE $\hat{\lambda}$ for an Exponential distribution is given by:

$$\hat{\lambda} = \frac{d}{\sum_i t_i} = \frac{\text{The number of events}}{\text{Person-time: total number of time units observed on all individuals}}$$

Using R to calculate:

```
# Calculate number of events and person-time for relapse
d_relapse = sum(pull(q1_df, Relapse))
sum_time_relapse = sum(pull(q1_df, Relapse_Time))

# Calculate relapse MLE
mle_relapse = d_relapse/sum_time_relapse

# Calculate number of events and person-time for death
d_death = sum(pull(q1_df, Death))
sum_time_death = sum(pull(q1_df, Death_Time))

# Calculate death MLE
mle_death = d_death/sum_time_death
```

`mle_relapse` $\hat{\lambda}_{relapse} = 0.032$

`mle_death` $\hat{\lambda}_{death} = 0.013$

The maximum likelihood estimator $\hat{\lambda}$ is an estimator for the hazard rate parameter, λ, which is constant in an exponential distribution. The estimated hazard rate of relapse $\hat{\lambda}_{relapse}$ is 0.032 events per month of person-time. The estimated hazard rate of death $\hat{\lambda}_{death}$ is 0.013 events per month of person-time.

b. We can use the MLE to calculate the quantities below.

i. **Mean**
The expectation, or mean, of the exponential distribution is $\frac{1}{\lambda}$.

$$\mu_{relapse} = 1/0.032 = 31.25$$
$$\mu_{death} = 1/0.013 = 74.333$$

ii. **Median**
The median of an exponential distribution is given by $\tau = \frac{-log(0.5)}{\lambda}$.

$$\tau_{relapse} = \frac{-log(0.5)}{0.032} = 21.488$$

$$\tau_{death} = \frac{-log(0.5)}{0.013} = 51.524$$

iii. **1 & 2 Year Relapse-Free & Survival Probabilities**
These are calculated using the survival functions $S_R(t)$ and $S_D(t)$. Under the exponential distribution $S(t) = e^{-\lambda t}$.

$$S_R(12) = e^{-0.032(12)} = 0.679$$
$$S_R(24) = e^{-0.032(24)} = 0.461$$
$$S_D(12) = e^{-0.013(12)} = 0.851$$
$$S_D(24) = e^{-0.013(24)} = 0.724$$

iv. **1 & 2 Year Relapse and Death Probabilities**
This is easily calculated from the survival function since $F(t) = 1 - S(t)$.

$$F_R(12) = 1 - S_R(12) = 0.321$$
$$F_R(24) = 1 - S_R(24) = 0.539$$
$$F_D(12) = 1 - S_D(12) = 0.149$$
$$F_D(24) = 1 - S_D(24) = 0.276$$

v. **Probability of Staying Relapse-Free 2 Years Given 1 Year Relapse-Free**
This is a conditional probability denoted as $S_R(24|12)$ and is easily calculated using $S_R(24|12) = S_R(24)/S_R(12)$ since it is certain $S_R(12|24) = 1$. This simplification is shown below.

$$S_R(24|12) = \frac{S_R(24 \cap 12)}{S_R(12)} = \frac{S_R(12|24)S_R(24)}{S_R(12)} = \frac{S_R(24)}{S_R(12)} = \frac{0.461}{0.679} = 0.679$$

As expected, $S_R(24|12) = S_R(12)$ since the hazard rate λ of an exponential distribution is constant.

vi. **Median (Using Non-Parametric Methods)**
If an exponential distribution is not assumed, the median time-to-event can be calculated using a Kaplan-Meier estimate. However, in this case, only the median time-to-relapse can be calculated. The median time-to-event is given by the smallest t where $\hat{S}(t) \leq 0.5$. For deaths, the KM survival estimator $\hat{S}(t)$ never reaches 0.5, since 7 of 10 observations are censored, and can therefore not be estimated. For relapse, $\hat{S}(t)$ drops to 0.5 at 27 months, so the median time-to-relapse is calculated to be 27 months. This can be confirmed with R:

```
km_q1 = survfit(
  Surv(Relapse_Time, Relapse) ~ 1,
  data = q1_df)

surv_median(km_q1)
```

```
##   strata median lower upper
## 1    All     27    12    NA
```

Question 2

```
# Load in Q2 data
q2_df = read_csv("data/Q2data_extracted.csv")
```

a. Kaplan-Meier Survival Estimate

t_j	d_j	c_j	r_j	$\lambda_j = (d_j/r_j)$	$\hat{S}(t_j) = \prod_j (1-\lambda_j)$
2	1	0	17	$\frac{1}{17}$	$1.000(1-\frac{1}{17}) = 0.941$
3	1	0	16	$\frac{1}{16}$	$0.941(1-\frac{1}{16}) = 0.882$
4	1	0	15	$\frac{1}{15}$	$0.882(1-\frac{1}{15}) = 0.824$
12	1	0	14	$\frac{1}{14}$	$0.824(1-\frac{1}{14}) = 0.765$
22	1	0	13	$\frac{1}{13}$	$0.765(1-\frac{1}{13}) = 0.706$
48	1	0	12	$\frac{1}{12}$	$0.706(1-\frac{1}{12}) = 0.647$
51	0	1	11	$\frac{0}{11}$	$0.647(1-\frac{0}{11}) = 0.647$
56	0	1	10	$\frac{0}{10}$	$0.647(1-\frac{0}{10}) = 0.647$
80	2	0	9	$\frac{2}{9}$	$0.647(1-\frac{2}{9}) = 0.503$
90	1	0	7	$\frac{1}{7}$	$0.503(1-\frac{1}{7}) = 0.431$
94	0	1	6	$\frac{0}{6}$	$0.431(1-\frac{0}{6}) = 0.431$
160	1	0	5	$\frac{1}{5}$	$0.431(1-\frac{1}{5}) = 0.345$
161	1	0	4	$\frac{1}{4}$	$0.345(1-\frac{1}{4}) = 0.259$
180	1	1	3	$\frac{1}{3}$	$0.259(1-\frac{1}{3}) = 0.173$
238	1	0	1	1	$0.173(1-1) = 0$

b.
```
# Log-log CI
km_loglog = survfit(
  Surv(Value, Binary) ~ 1,
  data = q2_df,
  conf.type = "log-log")

summary(km_loglog)
```

```
## Call: survfit(formula = Surv(Value, Binary) ~ 1, data = q2_df, conf.type = "log-log")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2     17       1    0.941  0.0571       0.6502        0.991
##     3     16       1    0.882  0.0781       0.6060        0.969
##     4     15       1    0.824  0.0925       0.5471        0.939
##    12     14       1    0.765  0.1029       0.4883        0.904
##    22     13       1    0.706  0.1105       0.4315        0.866
##    48     12       1    0.647  0.1159       0.3771        0.823
##    80      9       2    0.503  0.1272       0.2436        0.716
##    90      7       1    0.431  0.1277       0.1870        0.656
```

```
##   160      5       1    0.345  0.1280       0.1216        0.584
##   161      4       1    0.259  0.1217       0.0691        0.505
##   180      3       1    0.173  0.1074       0.0296        0.416
##   238      1       1    0.000     NaN           NA           NA
```

```r
# Linear CI
km_linear = survfit(
  Surv(Value, Binary) ~ 1,
  data = q2_df,
  conf.type = "plain")

summary(km_linear)
```

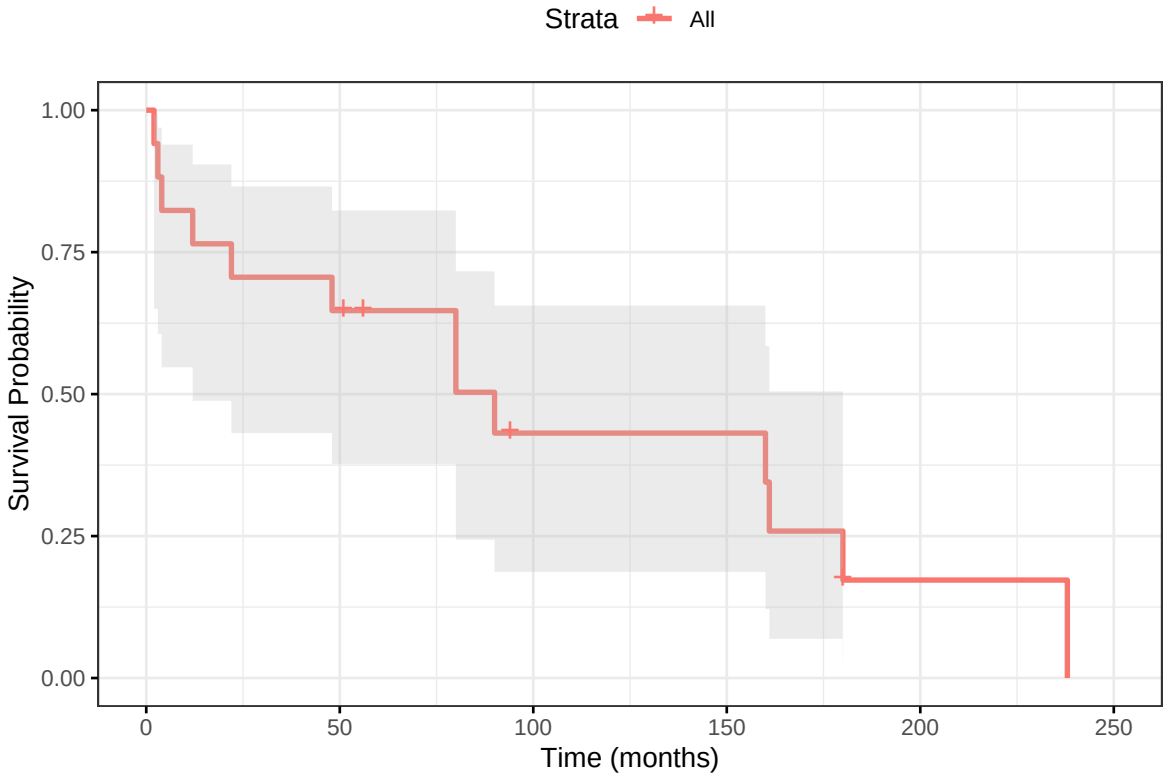
```
## Call: survfit(formula = Surv(Value, Binary) ~ 1, data = q2_df, conf.type = "plain")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2     17       1    0.941  0.0571       0.8293        1.000
##     3     16       1    0.882  0.0781       0.7292        1.000
##     4     15       1    0.824  0.0925       0.6423        1.000
##    12     14       1    0.765  0.1029       0.5631        0.966
##    22     13       1    0.706  0.1105       0.4893        0.922
##    48     12       1    0.647  0.1159       0.4199        0.874
##    80      9       2    0.503  0.1272       0.2541        0.752
##    90      7       1    0.431  0.1277       0.1811        0.682
##   160      5       1    0.345  0.1280       0.0942        0.596
##   161      4       1    0.259  0.1217       0.0204        0.497
##   180      3       1    0.173  0.1074       0.0000        0.383
##   238      1       1    0.000     NaN          NaN           NaN
```

The "log-log" approach to calculating the 95% confidence intervals is done in order to keep the interval within the [0, 1] bounds of probability. The "linear" approach however is a simple $\hat{S}(t) \pm z_{1-\alpha/2}(SE)$ which can often lead to confidence intervals out of the [0, 1] interval. This does indeed happen with the above Linear CI calculation but the shown interval is truncated at 0.000 and 1.000 by the `survfit()` function. Using the linear CI calculation, the upper 95% CI at $t_j = 2$ is:

$$\hat{S}(t) \pm z_{1-\alpha/2}(SE) = 0.941 + 1.96(0.0571) = 1.053$$

c. KM Plot

```r
ggsurvplot(
  km_loglog,
  data = q2_df,
  conf.int = TRUE,
  ggtheme = theme_bw(),
  xlab = "Time (months)",
  ylab = "Survival Probability")
```


Strata
All
1.00
0.75
0.50
0.25
0.00
Survival Probability
0
50
100
150
200
250
Time (months)

d.