STAT 641
Homework 5

Keegan Smith

July 7, 2025

Problem 1

First of all, in the data, none of the entries are marked with status == 0, so I will be treating this data as uncensored.

1. a) the trimmed mean is 886.7853403141361 and the untrimmed mean is 955.3713080168776. This would suggest that the data has some extreme values to the right, and is right skewed.

   b) the survival function is:

   $$S(t) = 1 - F(t)$$

   We can derive $F(t)$ from the pdf, (and since t is a time any probability for $t < 0$ is 0, so we are only considering $t \geq 0$):

   $$\begin{aligned} F(t) &= \int_0^t \lambda e^{-\lambda x} \\ &= (-e^{-\lambda x})_0^t \\ &= (-e^{-\lambda t} - (-1)) \\ &= 1 - e^{-\lambda t} \end{aligned}$$

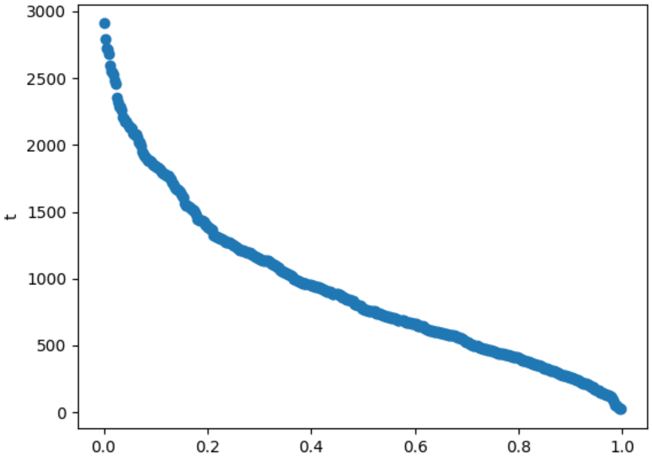
   So we have:

   $$\begin{aligned} S(t) &= e^{-\lambda t} \\ \ln(S(t)) &= -\lambda t \end{aligned}$$

   So if $S(t)$ is a good estimate, then the plot t vs $\ln(S(t))$ should be a linear plot with slope:

$$\frac{t}{-\lambda t} = \frac{1}{-\lambda}$$

The actual plot is:



c) We have the likelihood function:

$$\begin{aligned} L(\lambda; y) &= \Pi_{k=0}^{n-1} \lambda e^{-\lambda y_k} \\ &= \lambda^n e^{\sum_{k=0}^{n-1} -\lambda y_k} \\ &= \lambda^n e^{-\lambda \cdot \sum_{k=0}^{n-1} y_k} \end{aligned}$$

The log likelihood function is then:

$$\begin{aligned} \ln(L(\lambda; y)) &= \ln(\lambda^n e^{-\lambda \cdot \sum_{k=0}^{n-1} y_k}) \\ &= \ln(\lambda^n) + (-\lambda \cdot \sum_{k=0}^{n-1} y_k) \\ &= n \cdot \ln(\lambda) - \lambda \cdot \sum_{k=0}^{n-1} y_k \end{aligned}$$

The derivative of the log likelihood function w.r.t λ is:

$$\begin{aligned}\frac{d}{d\lambda}(n \cdot \ln(\lambda) - \lambda \cdot \sum_{k=0}^{n-1} y_k) &= n \cdot \frac{1}{\lambda} - \sum_{k=0}^{n-1} y_k \\ &= \frac{n}{\lambda} - \sum_{k=0}^{n-1} y_k\end{aligned}$$

Solving for the maximum:

$$\begin{aligned}\frac{n}{\lambda} - \sum_{k=0}^{n-1} y_k = 0 \\ \lambda &= \frac{n}{\sum_{k=0}^{n-1} y_k} \\ &= \frac{1}{\mu}\end{aligned}$$

Thus the MLE estimate for λ is $\frac{1}{955.3713} \approx 0.001047$

d) The exponential distribution estimated probability that a male will have a survival time > 1200 is:

$$\begin{aligned}P(t > 1200) &= 1 - P(t \leq 1200) \\ &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-0.001047 \cdot 1200} \\ &= 0.2846770212809416\end{aligned}$$

The distribution free estimate is:

$$\begin{aligned}P(t > 1200) &= 1 - P(t \leq 1200) \\ &= 1 - F(1200) \\ &= 1 - 0.7257383966244726 \\ &= 0.2742616033755274\end{aligned}$$

e) First I'm just going to derive the Quantile function to make this slightly less painful:

$$\begin{aligned}z &= 1 - e^{-\lambda t} \\ -z + 1 &= e^{-\lambda t} \\ \ln(-z + 1) &= -\lambda t \\ t &= \frac{\ln(-z + 1)}{-\lambda}\end{aligned}$$

So:

$$Q(z) = \frac{\ln(-z+1)}{-\lambda}$$

Thus we have the MLE median:

$$\begin{aligned} Q(.5) &= \frac{\ln(-.5+1)}{-0.001047} \\ &= 662.0316910792219 \end{aligned}$$

MLE IQR:

$$\begin{aligned} Q(.75) - Q(.25) &= \frac{\ln(-.75+1)}{-0.001047} - \frac{\ln(-.25+1)}{-0.001047} \\ &= 1049.2954046495795 \end{aligned}$$

Distribution free median is 833, the IQR is 793. These vary greatly from the MLE estimates for the exponential distribution, implying that the exponential distribution may not be a great fit for the distribution of survival times.

2. The distribution free estimates for the mean and standard deviation are simply the sample mean and standard deviation respectively, which are: 955.371308 and 634.253830 for males, and 896.506977 and 569.495299 for females.

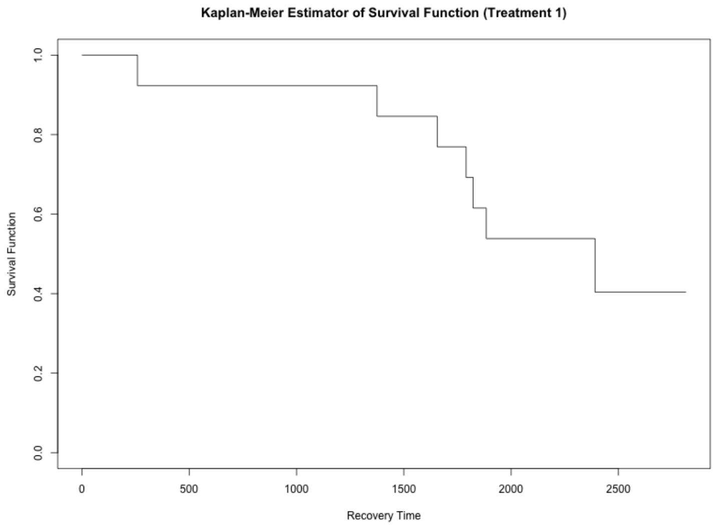
3. For males the estimated median is 833 and the MAD is 567.8280207561156. For females the estimated median is 759 and the MAD is 558.9325426241661.

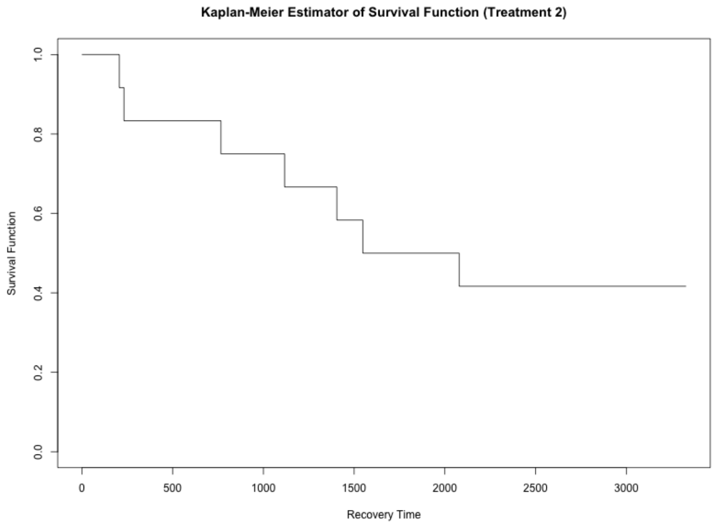
```
def mad(data):
  data = np.asarray(data)
  c = np.median(data)
  deviations = np.abs(data - c)
  raw_mad = np.median(deviations)
  return raw_mad / 0.6745
```

4. The above results would imply that the female average survival time is less than the male average survival time and that the variability in survival times is less for females than for males.

Problem 2

1. This experiment is using random censoring since there isn't a specified time when the experiment was stopped, patients could leave at random times. Knowing this, we will use Kaplan-Meier to approximate the survival function. Using the R code from handout 7, we get the following survival functions for treatments 1 and 2:





2. The estimated mean from the survival function for treatment 1 is 2135 and the estimated median is 2391. The estimated mean from the survival function for treatment 2 is 2000 and the estimated median is 1814.

3. From the estimated mean recovery times, it would appear that treatment 2 is more effective than treatment 1.

4. The estimated mean and median for treatment 1 are 1596.857142857143 and 1375 respectively, for treatment 2 they are 1050.4285714285713 and 1117 respectively.

The code used for this problem is below:

```
library(survival)

df <- data.frame(
  time   = c(2391, 2815, 1884, 1656, 2184, 2118, 1905, 1375, 259, 1790, 2413, 2761, 1823),
  status = c(1, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1)
)

km.fit <- survfit(Surv(time, status) ~ 1, data = df)

print(km.fit)
png("treatment1_survival.png", width=800, height=600)
```

```
plot(km.fit,conf.int=FALSE,log=FALSE,
main="Kaplan-Meier Estimator of Survival Function (Treatment 1)",xlab="Recovery Time",
ylab="Survival Function")
summary(km.fit)
print(km.fit, print.rmean=TRUE, rmean="individual")

df <- data.frame(
  time   = c(2312, 2501, 2691, 1548, 3329, 2154, 766, 1405, 1117, 232, 206, 2079),
  status = c(0, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1)
)

km.fit <- survfit(Surv(time, status) ~ 1, data = df)

print(km.fit)
png("treatment2_survival.png", width=800, height=600)

plot(km.fit,conf.int=FALSE,log=FALSE,
main="Kaplan-Meier Estimator of Survival Function (Treatment 2)",xlab="Recovery Time",
ylab="Survival Function")
summary(km.fit)
print(km.fit, print.rmean=TRUE, rmean="individual")
```

Problem 3

1. A (type 1 censoring)

2. B (type 2 censoring)

3. C (right censoring)

4. C (random censoring)

5. A (left censoring since you know the patient got the disease at a time less than or equal to the start of the study)

6. C (random censoring)

7. B (type 2 censoring)

8. A (random censoring)

9. D (random censoring)

10. C (random censoring)