Homework 2

Denis Ostroushko

Problem 1

(i)

We are not able to calculate variance with the information given in this problem. We need the number of events and number of observations at risk for each time point leading up to the time point j.

(ii)

For an asymptotically normal random variable, either the delta method or an endpoint transformation produce desired outcome.

We will use the endpoint transformation method to find a confidence interval for the log-odds of surviving past a time point t_j.

First, obtain a confidence interval for the probability of surviving, which is given by

$$\widehat{S(t_j)} \pm 1.96 * \sqrt{\widehat{Vat(S(t_j))}} = (lb, up)$$

The, for the log-odds of survival, a confidence interval is given by

$$(log\frac{lb}{1-lb}, log\frac{ub}{1-ub})$$

Delta Method

if we have to determine the variance, as stated in the Note section, we would use the delta method approach to get variance of $T(S(t_j))$.

First, define $g(S(t_j)) = log\frac{S(t_j)}{1-S(t_j)}$

Then, $g'(S(t_j))^2 = [\frac{d}{dS}g(S(t_j))]^2 = (\frac{1}{S(t_j)} + \frac{1}{1-S(t_j)})^2$

Therefore, $Var[T(S(t_j))] = (\frac{1}{S(t_j)} + \frac{1}{1-S(t_j)})^2 * Var(S(t_j))$

We will need to compute this variance at t_j using the value of $S(t_j)$ and its variance.

Then we compute a 95% confidence interval: $T(S(t_j)) \pm 1.96 * \sqrt{Var[T(S(t_j))]}$

(iii)

Alternative hypothesis: $H_a : S_1(t_j) \neq S_0(t_j)$

Pooled standard to test the difference, taking into account the known sample sizes for group 0 and 1 at time t_j: pooled $SE = \sqrt{\frac{(n_1-1)*Var(S_1(t_j))}{n_0+n_1-2} + \frac{(n_0-1)*Var(S_0(t_j))}{n_0+n_1-2}}$

Test statistic: $\frac{(S_1(t_j)-S_0(t_j))-0}{SE}$

Dealing with finite sample sizes, we will need to use t-distribution to test, so the cutoff for significance is $T^* = t_{.975,df=n_0+n_1-2}$. We check the absolute value of our test statistic against T^*

(iv)

Confidence interval for the difference for the two sample t-test is: $S_1(t_j) - S_0(t_j) \pm t_{.975,df=n_0+n_1-2} * SE$, where SE was defined in the previous section.

However, comparing just the two point estimates means that we do not need to account for sample size, and can treat $S_1(t_j)$ and $S_0(t_j)$ as two independent normal random variables.

Therefore, by properties of independent normal random variables we obtain their pooled standard error as: $SE = \sqrt{Var(S_1(t_j)) + Var(S_0(t_j))}$

Then, the confidence interval is given by: $S_1(t_j) - S_0(t_j) \pm 1.96 * SE$

Problem 2

(i)

Figure 1 shows the two KM survival curves for the two groups. It appears that group 2, with a percutaneously placed catheter, has more evidence that this method of placement delays time of infection more effectively.
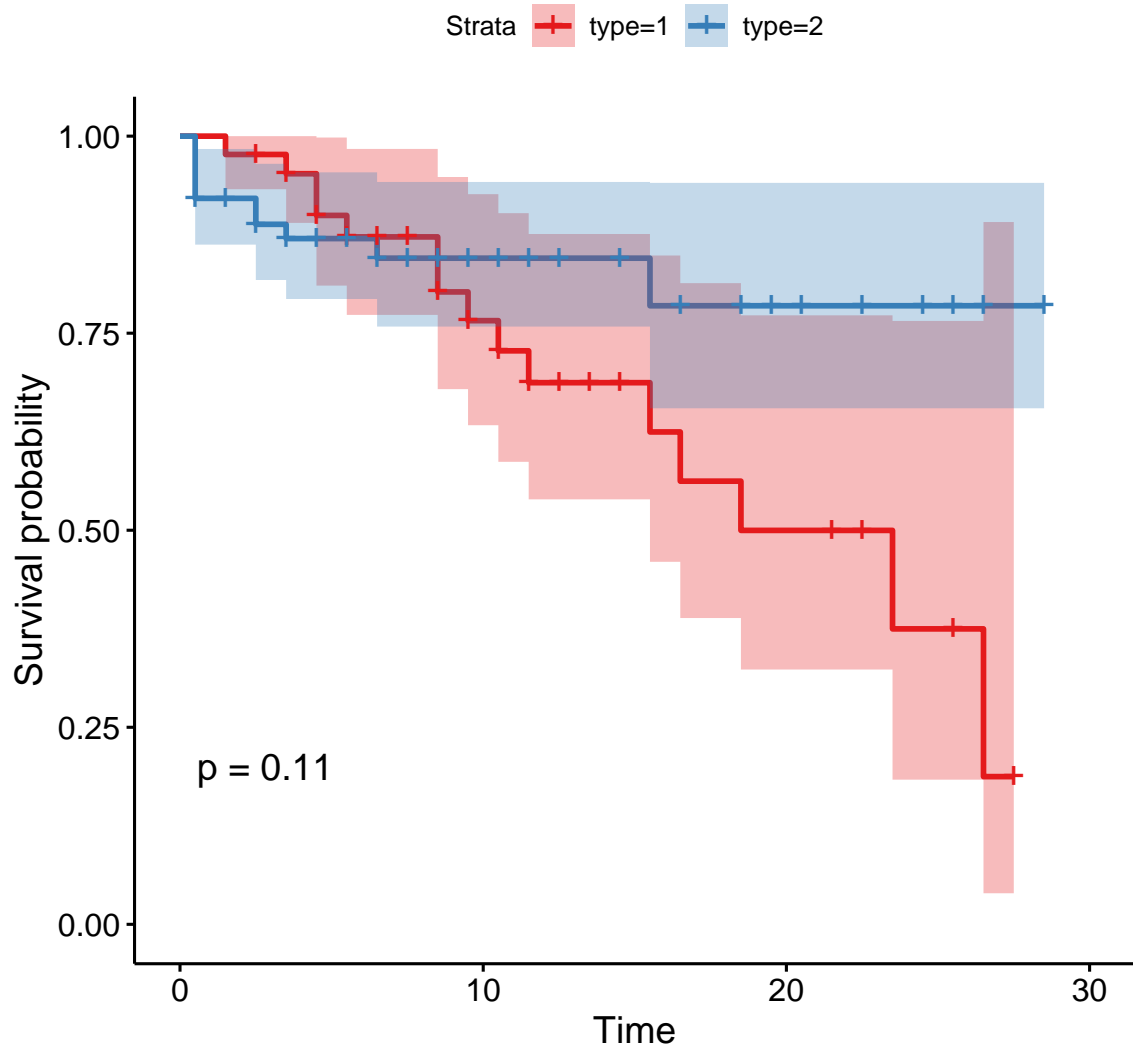
Figure 1: Kaplan Meier Estiamtes for time to infection in weeks the two groups

(ii)

Figure 2 provides cumulative hazard estimates for the two groups. We can see that type 2 group has a much flatter cumulative hazard curve since there are less events in that group.

Code chunk that produces cumulative hazard plot also outputs instant crude hazard for each group at time 5.

```
[1] "Crude hazard for group type 1 at time = 5"


        haz
0.03030303


[1] "Crude hazard for group type 2 at time = 5"


        haz
0.00952381
```
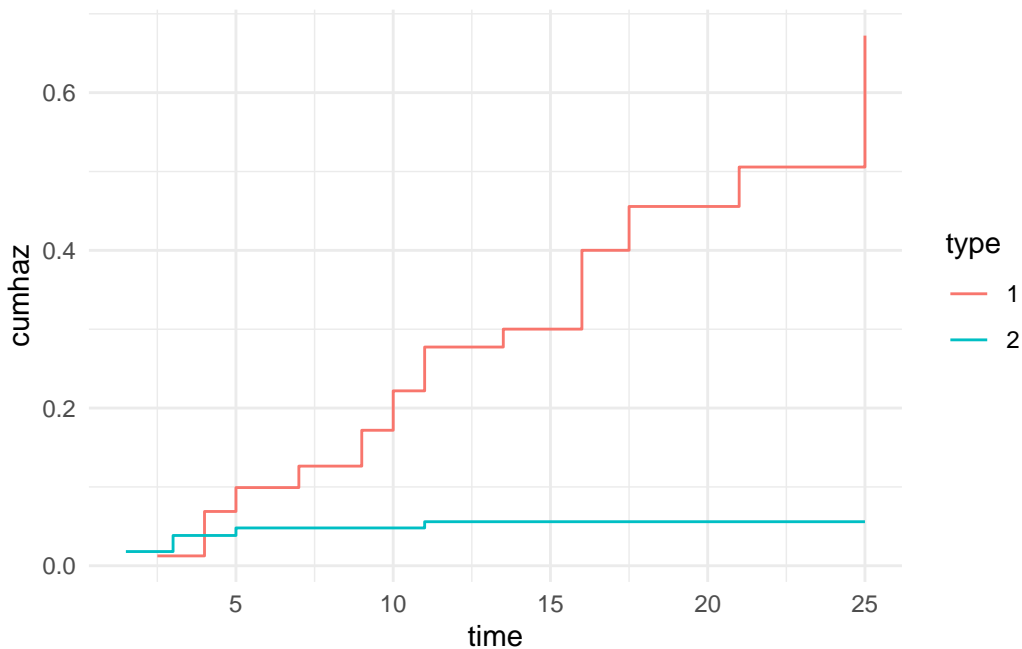

Figure 2: Hazard functions for two groups

(iii)

We can use `R`'s `survfit` and summary functions to obtain a 36-month restricted mean survival time

```
fit_1 <- survfit(Surv(time, delta) ~1 ,data = kidney %>% filter(type == 1))

s.km.fit = summary(fit_1, rmean=36)

fit_2 <- survfit(Surv(time, delta) ~1 ,data = kidney %>% filter(type == 2))

s.km.fit2 = summary(fit_2, rmean=36)
```

Restricted mean survival time for group 1 is: 20.12, and the standard error is 2.46. Therefore, a confidence interval is given by (15.31, 24.93)

Restricted mean survival time for group 2 is: 29.54, and the standard error is 1.9. Therefore, a confidence interval is given by (25.8, 33.27)

(iv)

- Null hypothesis: restricted to 36 month mean survival time for group 1 is equal to mean for group 2

- Alternative hypothesis: restricted means are not equal

- $H_0 : \hat{\mu}_2^{36} = \hat{\mu}_2^{36}$

- $H_a : \hat{\mu}_2^{36} \neq \hat{\mu}_2^{36}$

- We will conduct a two-sample t-test, so we need a pooled standard error weighted by the sample sizes.

- Test statistic is $\frac{\mu_2^{36} - \hat{\mu}_2^{36} - 0}{pooled\ SE}$

- Test statistic is 4.4406072

- Significance threshold is given by the 97.5th quantile fo the t distribution on the 117 degrees of freedom. So, T^* is 1.9800999

- So, the two restricted survival mean times are different for the two groups. Time to infection is higher for group 2.

(v)

Figure 3 provides a plow with survival curves and HW confidence bands for each group. We can see that unlike equal probability bands, HW bands remain constant over time.
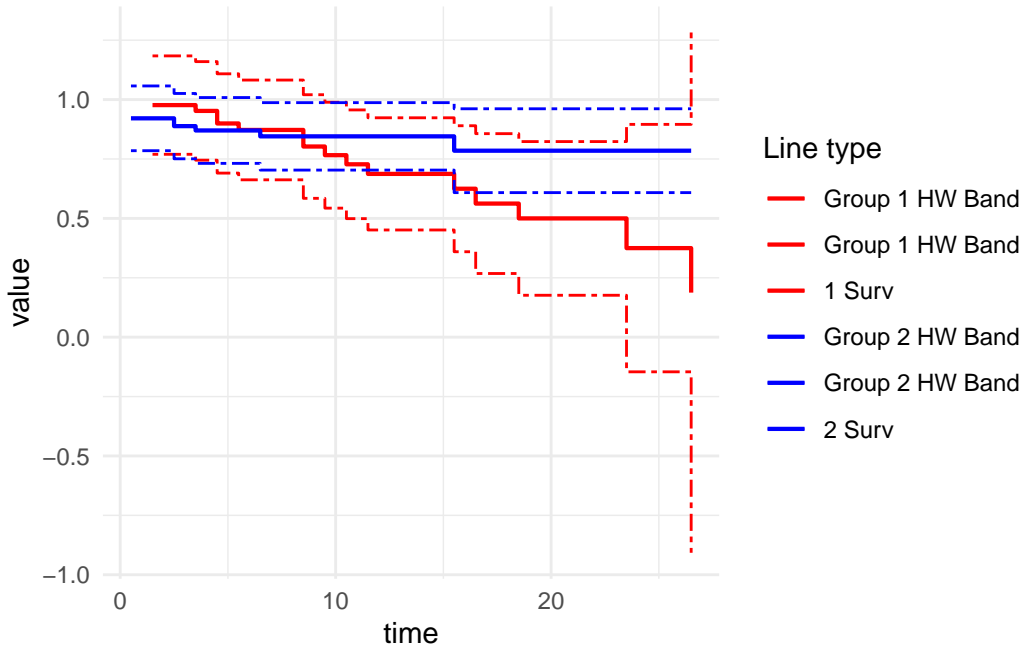

Figure 3: HW confidence bands for two group

Problem 3

(i)

I presume we need to identify probability distribution that fits our data. We will obtain a plot of time against survival probabilities on the logarithmic scale. A linear fit will indicate that a simple exponential model with one parameter is an appropriate fit for the data.

Figure 4 shows that an exponential model is an appropriate fit for the data. We will use the number of observed events and the total observed exposure time to estimate $\hat{\lambda}$

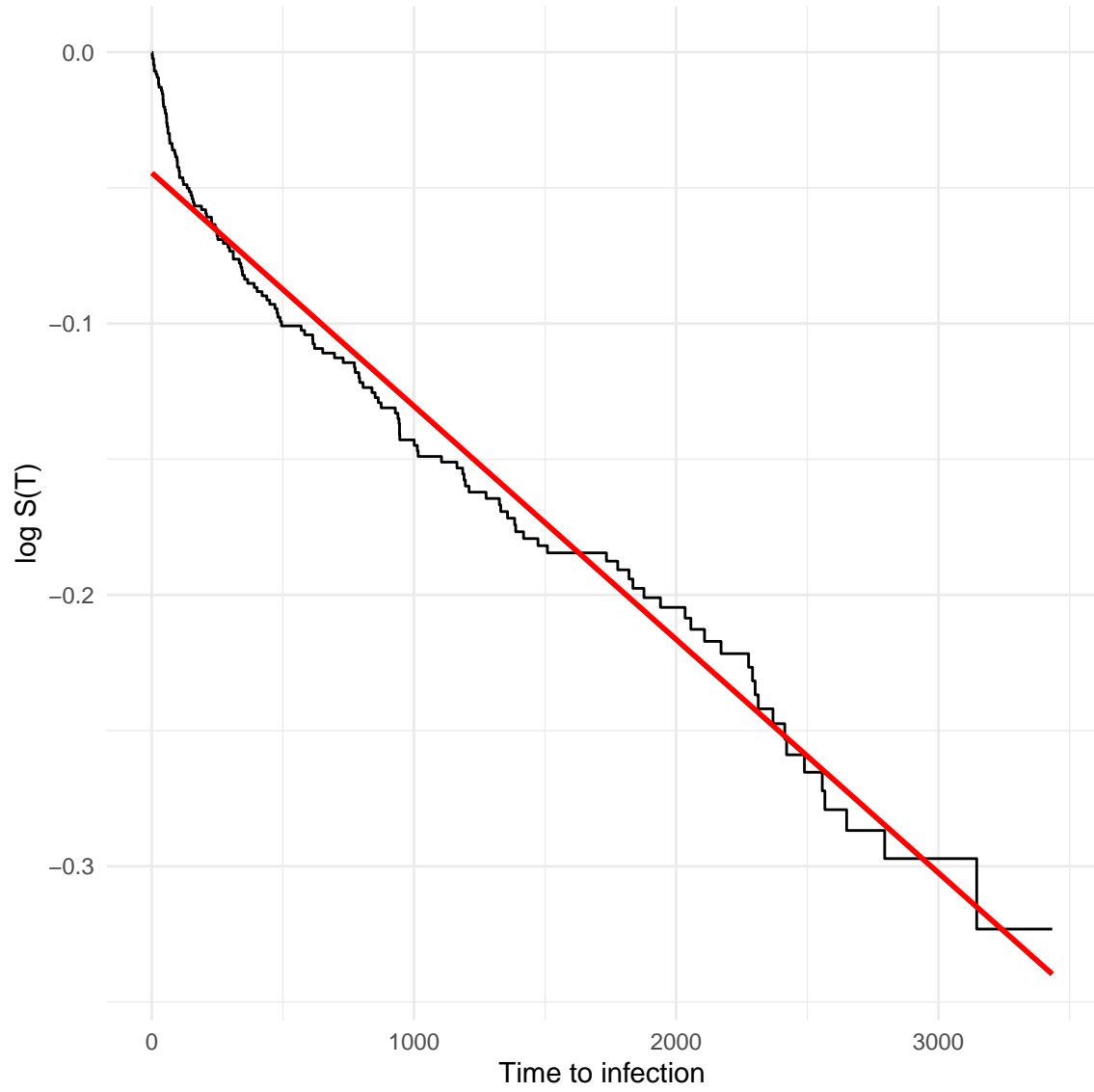```
[1] 0.00009399878
```

$\hat{\lambda} = 0.000094$

Figure 4: Validation of exponential model fit to the data

**(ii)**

Using exponential survival model $S(t) = exp(-0.000094 * t)$, we use $t = 365$ and obtain a 12-month survival probability, which is 0.96627.

Now, using variance of parameter $\hat{\lambda}$ we can compute a 95% linear confidence interval, which is given in the `R` output below:

```
     2.5%      97.5%
0.9547297 0.9764791
```

(iii)

Using log transformation of survival probability at time 365 days, and exponential transformation of confidence interval endpoints, we use variance of lambda again to compute the 95% confidence interval, given below:

```
     2.5%      97.5%
0.9553151 0.9761448
```

(iv)

Due to large sample size, and small variance of $\hat{\lambda}$ the two 95% confidence intervals are approximately equal. We know that in small sample sizes log-transformed confidence intervals perform better, but we are not able to see this distinction.

(v)

We use a KM estimated survival curve for to display EP confidence band. Figure 5 shows the results. We can see that a time increases, the band gets wider. That is because EP bands are proportional to the point wise confidence interval estimates, which get larger as time goes on.

(vi)

Use `R` code below to get restricted survival means

```
# mean survival for females restrited to 60 months
wf_model <- survfit(Surv(time, delta) ~1 ,data = kidtran %>% filter(gender == 2, race == 1
wf_model.sum = summary(wf_model, rmean=1825)
```
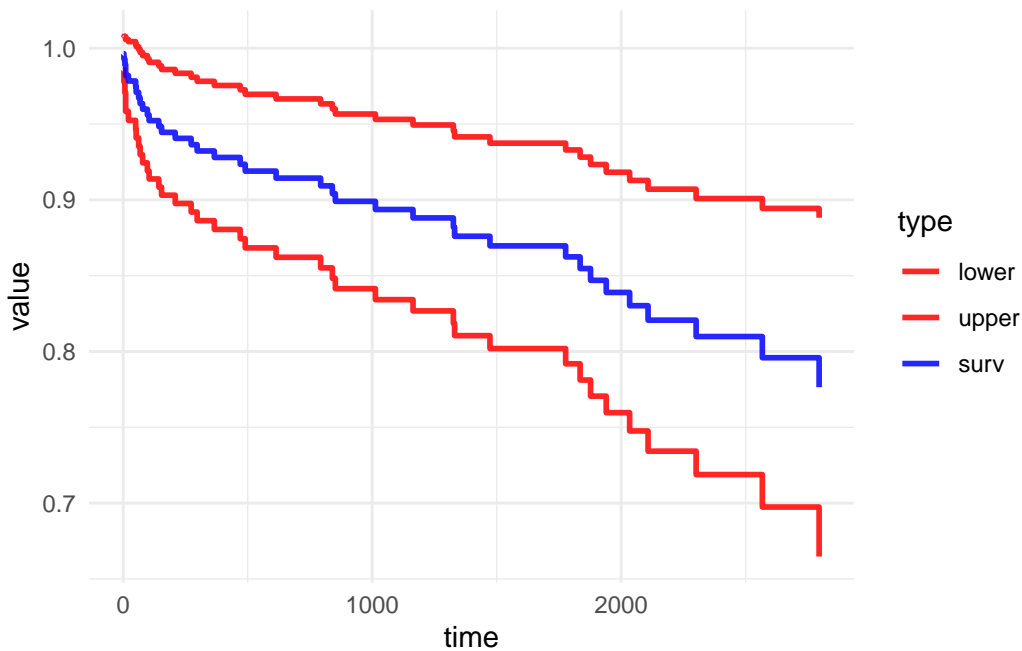
8

Figure 5: Equal probability confidence band

```
w_mean <- wf_model.sum$table["rmean"]
w_variance <- wf_model.sum$table["se(rmean)"]^2
```

- Mean survival time for white females is is 1651.93, the standard error is 30.43. Confidence interval for the mean is (1592.29, 1711.57)

```
# mean survival for males restrited to 60 months
wm_model <- survfit(Surv(time, delta) ~1 ,data = kidtran %>% filter(gender == 1, race == 1
wm_model.sum = summary(wm_model, rmean=1825)

m_mean <- wm_model.sum$table["rmean"]
m_variance <- wm_model.sum$table["se(rmean)"]^2
```

- Mean survival time for white males is is 1586.31, the standard error is 27.62. Confidence interval for the mean is (1532.18, 1640.43)

(vii)

We will use the t-test approach to find the confidence interval for the mean survival time between white males and white females. We will do so to account for varying sample sizes for these two groups.

```
    rmean
-2.282251
```

- The difference is Male mean - Female mean = -65.62

- Pooled standard error for this difference is 28.75

- We use 97.5 quantile of the t-distribution with 710 degrees of freedom in the calculation of the confidence interval

- Confidence interval is given by -65.62 $\pm$ 1.9633 * 28.75 = (-122.08, -9.17)

- Based on this confidence interval we can conclude that the survival times for males are consistently lower on averages then the survival times of females.