Computer Lab 10: Survival Analysis and Bayesian Inference

Complete all of the following questions, adding your inputs as code chunks (enclose within triple accent marks) within Rmarkdown.

The exercises are not marked and will not be factored into your course grade, but it is important to complete them to make sure you have the skills to answer assessment questions. You may consult any resource, including other students and the instructor. Please Knit this document to a PDF and upload your work via Canvas at the end of the session. Solutions will be posted for you to check your own answers.

Survival Analysis

The data file aids.csv contains data on survival times for patients diagnosed with autoimmune deficiency syndrome (AIDS) in the 1980s in Australia. The variables are: state - Australian state sex - sex (F or M) diganosis - day of diagnosis (Julian day) death - day of death (or last day of monitoring if patient did not die) status - whether alive (A) or deceased (D) at the end of the study period T.categ - Transmission category (how the patient became infected) age - Age at diagnosis

1. Load this file in from disk and produce a summary table.

```
df=read.csv('aids.csv')
head(df)
```

```
##   state sex  diag death status T.categ age
## 1   NSW   M 10905 11081      D      hs  35
## 2   NSW   M 11029 11096      D      hs  53
## 3   NSW   M  9551  9983      D      hs  42
## 4   NSW   M  9577  9654      D    haem  44
## 5   NSW   M 10015 10290      D      hs  39
## 6   NSW   M  9971 10344      D      hs  36
```

```
summary(df)
```

```
##     state               sex                 diag           death      
##  Length:2814        Length:2814        Min.   : 8302   Min.   : 8469  
##  Class :character   Class :character   1st Qu.:10167   1st Qu.:10677  
##  Mode  :character   Mode  :character   Median :10666   Median :11237  
##                                        Mean   :10587   Mean   :10997  
##                                        3rd Qu.:11104   3rd Qu.:11504  
##                                        Max.   :11503   Max.   :11504  
##     status            T.categ              age       
##  Length:2814        Length:2814        Min.   : 0.00  
##  Class :character   Class :character   1st Qu.:30.00  
##   Mode  :character   Mode  :character   Median :37.00
##                                          Mean   :37.38
##                                          3rd Qu.:43.00
##                                          Max.   :82.00
```

2. The "status" column distinguishes patients who passed away from patients who were still alive at the end of the study. What proportion of patients died during the study?

```
#str(df)

result=subset(df, status=='D')
prop_die=nrow(result)/length(df$status)
prop_die
```

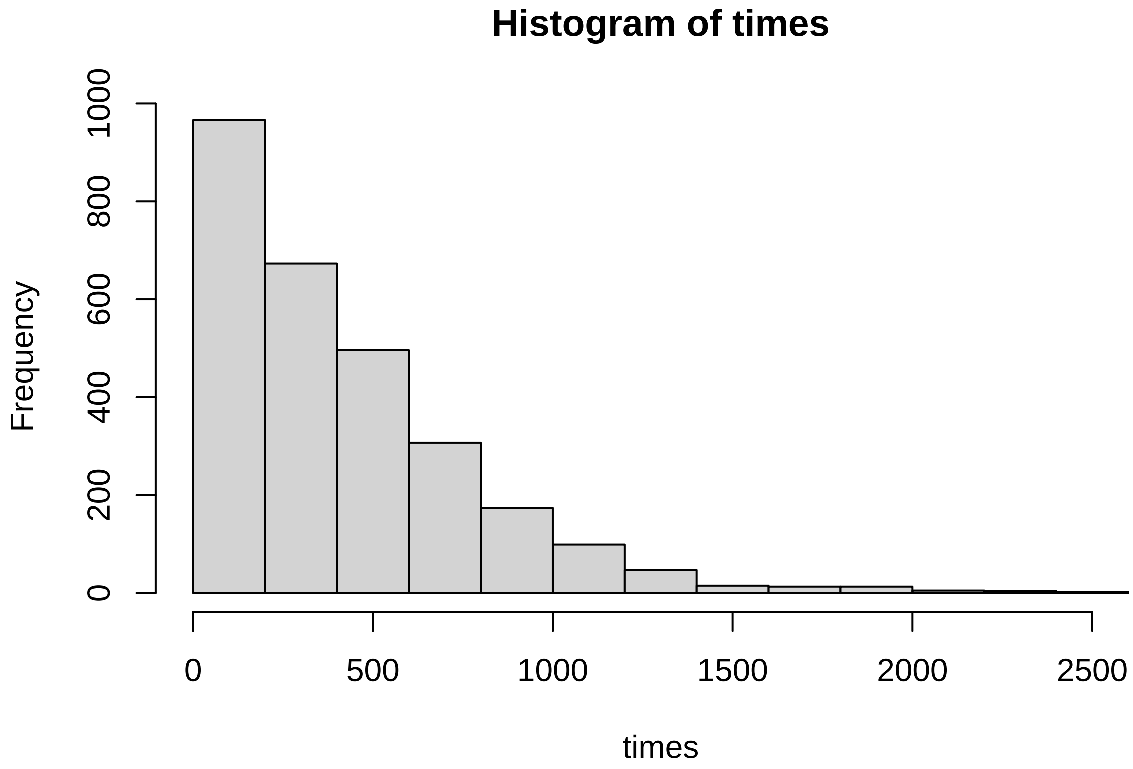
```
## [1] 0.6158493
```

3. Calculate elapsed times in days by taking the difference of the day of "death" minus the day of diagnosis. (For patients who actually died this is the survival time, for other patients it is the time until the end of the study).

```
times = df$death-df$diag
```

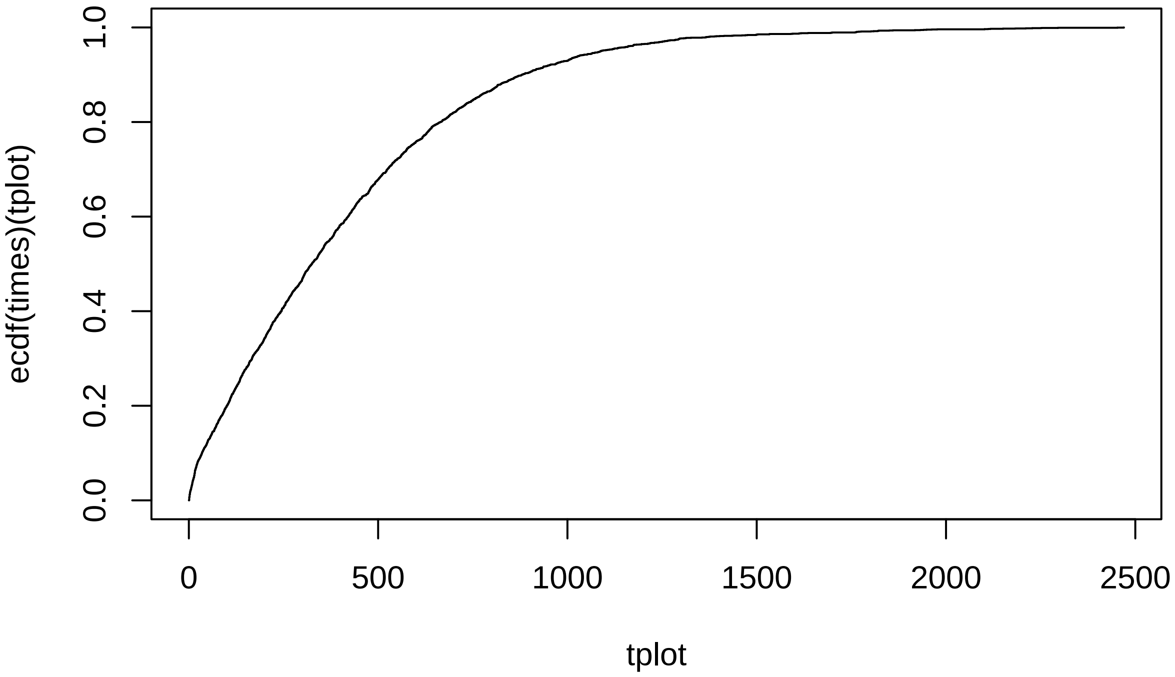
4. Plot a histogram of all elapsed times.

```
hist(times)
```



5. Plot the ECDF of all elapsed times.

```
tplot = c(0,sort(times))
plot(tplot,ecdf(times)(tplot),typ='s')
```



6. Create a numeric vector that is 0 for patients who are alive and 1 for patients who died.

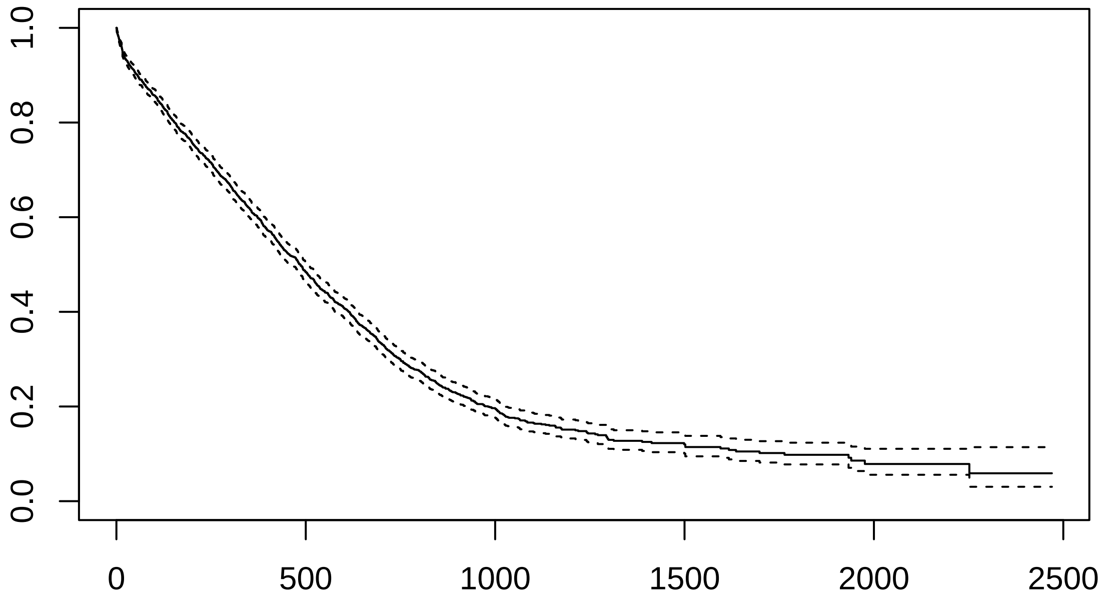
```
dead=as.numeric(df$status== 'D')
```

7. Create a survival object using `Surv()` combining the elapsed times and the numeric survival status. (You will need to load the survival library).

```
library(survival)
survival = Surv(times, dead)
```

8. Calculate and plot the Kaplan-Maier estimator of the survivorship curve using `survfit()` (treat all participants together as a single group).

```
survivalkm = survfit(survival~1)
plot(survivalkm)
```



9. Taking into account censorship, what is the median survival time after diagnosis (in days) for these patients?

```
survivalkm
```

```
## Call: survfit(formula = survival ~ 1)
##
##          n events median 0.95LCL 0.95UCL
## [1,] 2814   1733    485     458     506
```

10. Based on this plot, what is the approximate survival fraction for AIDS patients 6 years (~2200 days) after diagnosis? What is the expected fatality rate?

```
survivalkm$surv[survivalkm$time==2183]
```

```
## [1] 0.07853335
```

```
survivalkm$surv[survivalkm$time==2183]
```

```
## [1] 0.07853335
```

11. Carry out a survival regression using `survreg()` with no independent variables and a Weibull survival-time distribution model. Examine the output summary. Is there evidence for a non-zero `Log(scale)` parameter? What does this mean? (Remember: this parameter actually describes the shape; the intercept describes the scale. Yes, this is confusing.)

```
sr = survreg(survival~1)
summary(sr)
```

```
##
## Call:
## survreg(formula = survival ~ 1)
##              Value Std. Error      z      p
## (Intercept) 6.5088     0.0256 254.37 <2e-16
## Log(scale)  0.0560     0.0198   2.83 0.0046
##
## Scale= 1.06
##
## Weibull distribution
## Loglik(model)= -12995.4   Loglik(intercept only)= -12995.4
## Number of Newton-Raphson Iterations: 6
## n= 2814
```

12. Confirm your conclusion about the `Log(scale)` parameter from #11 by fitting an exponential-distribution model and comparing via `anova()` or `AIC()`.

```
srexp = survreg(survival~1,dist='exponential')
anova(sr,srexp,test='Chi')
```

```
##   Terms Resid. Df    -2*LL Test Df  Deviance    Pr(>Chi)
## 1     1      2812 25990.86      NA        NA          NA
## 2     1      2813 25999.09    = -1 -8.228356 0.004124088
```

```
AIC(sr,srexp)
```

```
##       df      AIC
## sr     2 25994.86
## srexp  1 26001.09
```

13. Calculate the actual shape (a) and scale (s) parameters for the Weibull model. [Reminder: The shape a is equal to the exponential of what `survreg` calls `Log(scale)`, or the reciprocal of what `survreg` calls `Scale`. The scale s is equal to the exponential of the intercept term. You can also access the parameter variables directly within the model object via $scale and $coefficients.]

```
a = 1/sr$scale # shape
s = exp(sr$coeff[1]) # scale
a
```

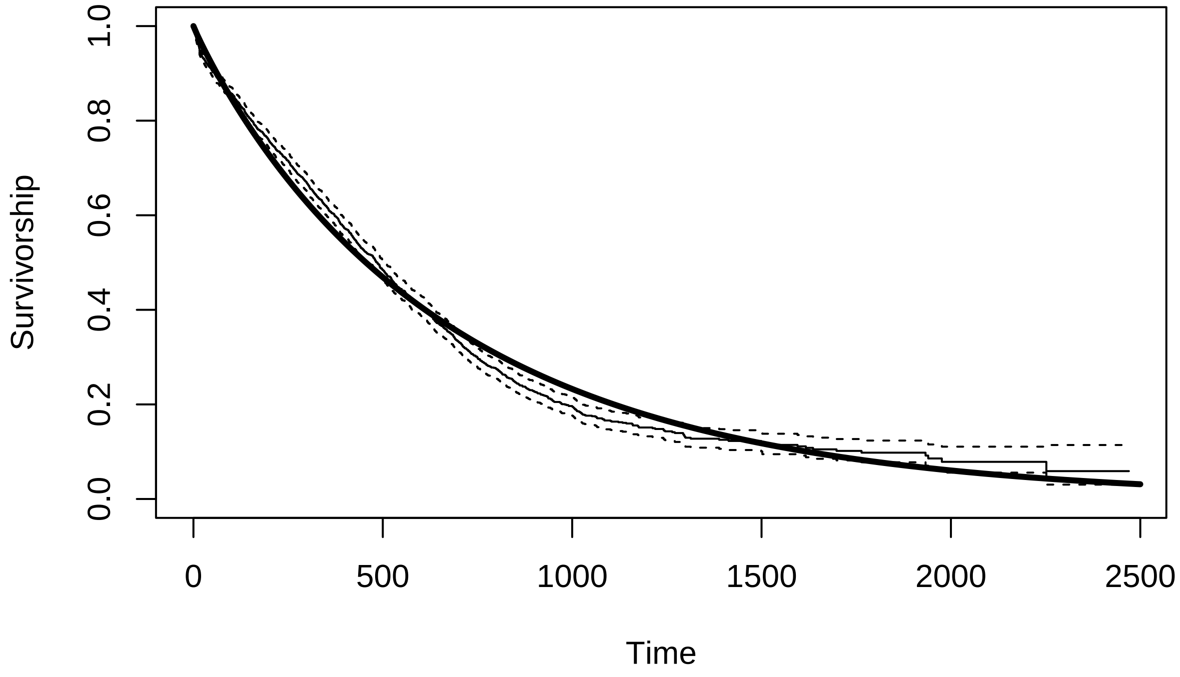
```
## [1] 0.9455319
```

```
s
```

```
## (Intercept)
##    671.0229
```

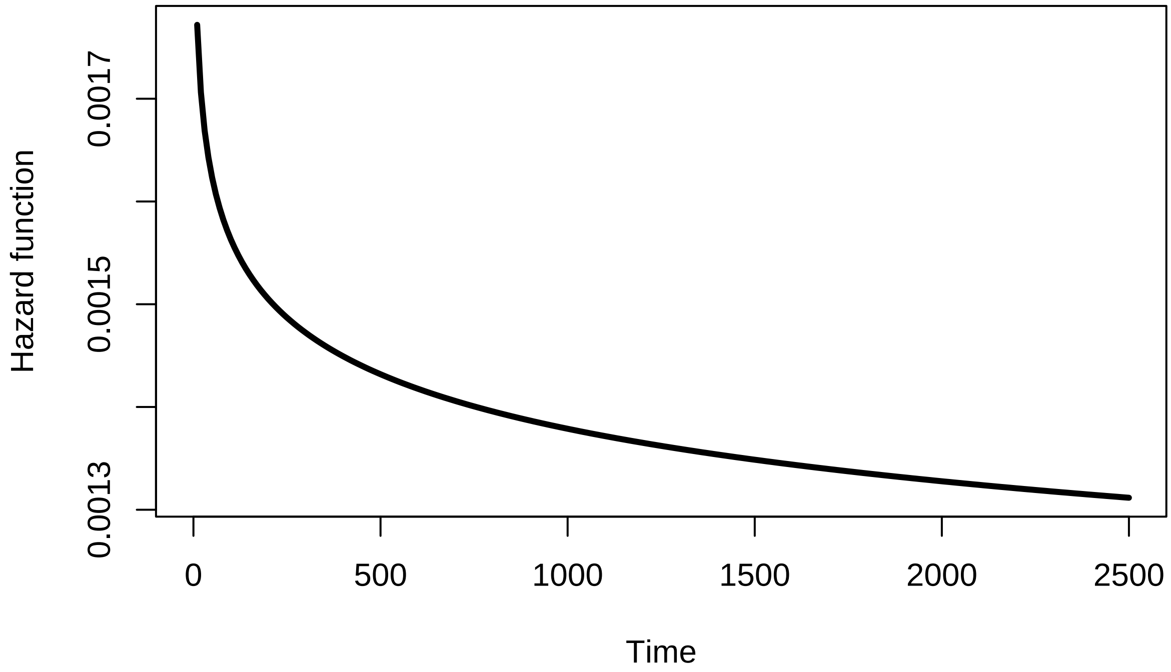
14. Calculate the model survivorship curve using the formula from lecture (or using `pweibull()`), given the values of s and a. Re-plot the K-M estimator (from #8), and then overplot this model survivorship curve on top of it.

```
plot(survivalkm,xlab='Time',ylab='Survivorship')
tplot = seq(0,2500,10)
lines(tplot, 1-pweibull(tplot,shape=a,scale=s), lwd=3)
```



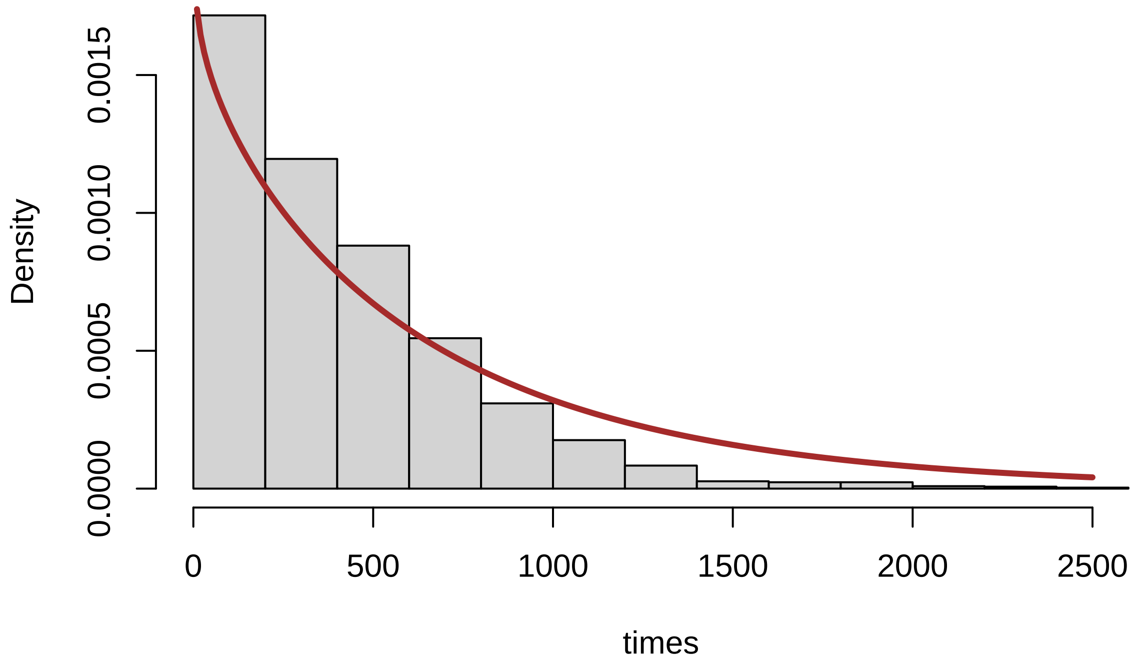
15. Plot the hazard function h(t) of the best-fit Weibull model using the formula from lecture (slide #28).

```
plot(tplot, (a/s)*(tplot/s)^(a-1),typ='l', xlab='Time', ylab='Hazard function', lwd=3)
```



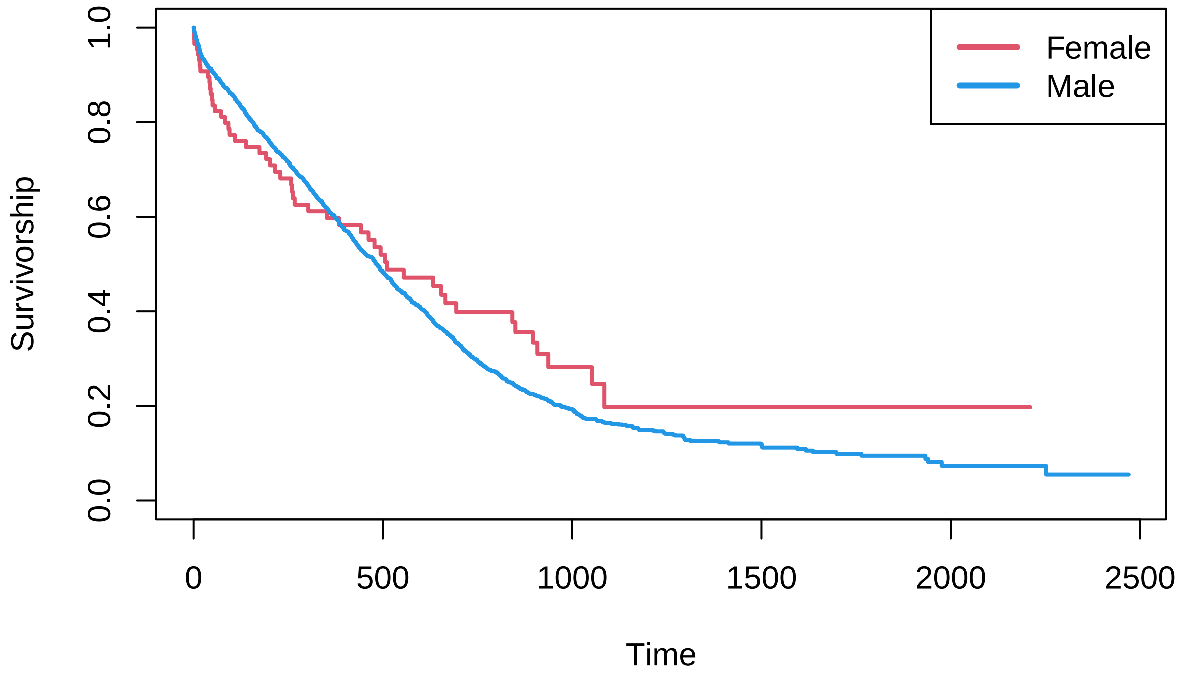
16. Calculate the expected survival time distribution using the formula from lecture (or using `dweibull()`). Overplot this curve on the (censorship-uncorrected) histogram from #4 (but with `freq=FALSE`). Why does the curve systematically exceed the histogram values at late times?

```
hist(times, freq=FALSE, main='')
lines(tplot, dweibull(tplot,shape=a,scale=s), lwd=3, col='brown')
```



17. Calculate and plot the Kaplan-Maier estimator of the survivorship curve for the male and female participants separately. Colour-code the two curves and add a legend.

```
kmsex = survfit(survival ~sex, data=df)
plot(kmsex,col=c(2,4),lwd=2, xlab='Time', ylab='Survivorship')
legend('topright',c('Female','Male'),col=c(2,4),lty=c(1,1),lwd=c(3,3))
```



18. Carry out a survival regression using `survreg()`. Consider as independent variables all the other fields in the model (state, sex, T.categ, age). You can ignore interactions. Which of these appear to have a significant impact on survival time?

```r
sr = survreg(survival~state+sex+T.categ+age,data=df)
summary(sr)
```

```
##
## Call:
## survreg(formula = survival ~ state + sex + T.categ + age, data = df)
##                  Value Std. Error     z       p
## (Intercept)    6.66986    0.21536 30.97 < 2e-16
## stateOther     0.12765    0.09534  1.34  0.1806
## stateQLD      -0.15559    0.09305 -1.67  0.0945
## stateVIC       0.02700    0.06495  0.42  0.6776
## sexM           0.08879    0.18831  0.47  0.6373
## T.categhaem   -0.08876    0.24596 -0.36  0.7182
## T.categhet     1.00931    0.29015  3.48  0.0005
## T.categhs      0.24256    0.14779  1.64  0.1008
## T.categhsid    0.31321    0.21601  1.45  0.1471
## T.categid      0.74303    0.27860  2.67  0.0077
## T.categmother  0.39979    0.76298  0.52  0.6003
## T.categother   0.16453    0.21168  0.78  0.4370
## age           -0.01350    0.00266 -5.07 3.9e-07
## Log(scale)     0.04952    0.01970  2.51  0.0120
```

```
##
## Scale= 1.05
##
## Weibull distribution
## Loglik(model)= -12964.8   Loglik(intercept only)= -12995.4
##  Chisq= 61.16 on 12 degrees of freedom, p= 1.4e-08
## Number of Newton-Raphson Iterations: 6
## n= 2814
```

19. Retain only the terms you found to be significant in #18 and re-run the model. Print out a summary.

```
srfinal = survreg(survival~T.categ+age,data=df)
summary(srfinal)
```

```
##
## Call:
## survreg(formula = survival ~ T.categ + age, data = df)
##                  Value Std. Error     z       p
## (Intercept)    6.72289    0.17909 37.54 < 2e-16
## T.categhaem   -0.04469    0.23679 -0.19 0.85031
## T.categhet     1.00325    0.28640  3.50 0.00046
## T.categhs      0.28652    0.13187  2.17 0.02980
## T.categhsid    0.35374    0.20559  1.72 0.08531
## T.categid      0.76165    0.27702  2.75 0.00597
## T.categmother  0.40811    0.76313  0.53 0.59280
## T.categother   0.19343    0.20907  0.93 0.35487
## age           -0.01359    0.00266 -5.11 3.1e-07
## Log(scale)     0.05009    0.01971  2.54 0.01103
##
## Scale= 1.05
##
## Weibull distribution
## Loglik(model)= -12967.6   Loglik(intercept only)= -12995.4
##  Chisq= 55.65 on 8 degrees of freedom, p= 3.3e-09
## Number of Newton-Raphson Iterations: 6
## n= 2814
```

20. By what factor is the survivorship behaviour “accelerated” (or decelerated) for a patient with transmission mode “het” versus a patient with transmission mode “hs”?

```
exp(srfinal$coeff['T.categhet'])/exp(srfinal$coeff['T.categhs'])
```

```
## T.categhet
##   2.047722
```

21 (optional). Use coxph() to fit a nonparametric proportional-hazard (Cox) model instead, using the same explanatory variables from your final parametric fit above. Produce a summary table.

```
cox = coxph(survival~T.categ+age,data=df)
summary(cox)
```

```
## Call:
## coxph(formula = survival ~ T.categ + age, data = df)
##
##   n= 2814, number of events= 1733
##
##                   coef exp(coef)  se(coef)      z Pr(>|z|)
## T.categhaem    0.049011  1.050232  0.225656  0.217 0.828058
## T.categhet    -1.007724  0.365049  0.272044 -3.704 0.000212 ***
## T.categhs     -0.310039  0.733418  0.125648 -2.468 0.013606 *
## T.categhsid   -0.378412  0.684948  0.195642 -1.934 0.053087 .
## T.categid     -0.780087  0.458366  0.263230 -2.964 0.003041 **
## T.categmother -0.415988  0.659688  0.725783 -0.573 0.566538
## T.categother  -0.211472  0.809392  0.199031 -1.063 0.288005
## age            0.013285  1.013373  0.002525  5.262 1.43e-07 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## T.categhaem      1.0502     0.9522    0.6748    1.6344
## T.categhet       0.3650     2.7394    0.2142    0.6222
## T.categhs        0.7334     1.3635    0.5733    0.9382
## T.categhsid      0.6849     1.4600    0.4668    1.0051
## T.categid        0.4584     2.1817    0.2736    0.7678
## T.categmother    0.6597     1.5159    0.1591    2.7361
## T.categother     0.8094     1.2355    0.5480    1.1956
## age              1.0134     0.9868    1.0084    1.0184
##
## Concordance= 0.562  (se = 0.008 )
## Likelihood ratio test= 60.63  on 8 df,   p=4e-10
## Wald test            = 60.68  on 8 df,   p=3e-10
## Score (logrank) test = 61.61  on 8 df,   p=2e-10
```

22 (optional). By what factor is hazard increased/decreased for a patient with transmission mode “het” versus a patient with transmission mode “hs”? (To calculate relative hazard, take the exponential of the appropriate model coefficient.)

```r
exp(cox$coefficients['T.categhet'])/exp(cox$coefficients['T.categhs'])
```

```
## T.categhet
##  0.4977364
```

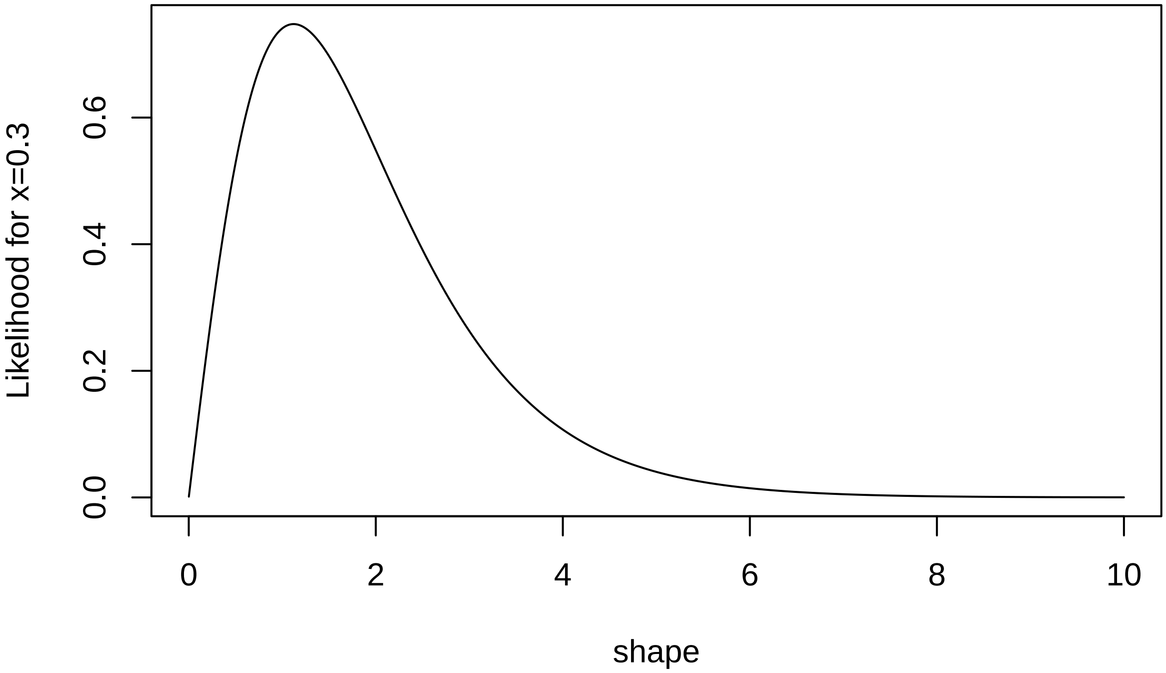
23 (optional). Plot the baseline cumulative hazard using `basehaz()`.

Bayesian Inference (1 parameter)

Suppose that some process is known to produce data that is Weibull-distributed, and you know in advance that the scale parameter is s=1.0. However, you do not know the shape parameter a. You obtain four measurements: 0.3, 2.1, 0.8, and 1.2.

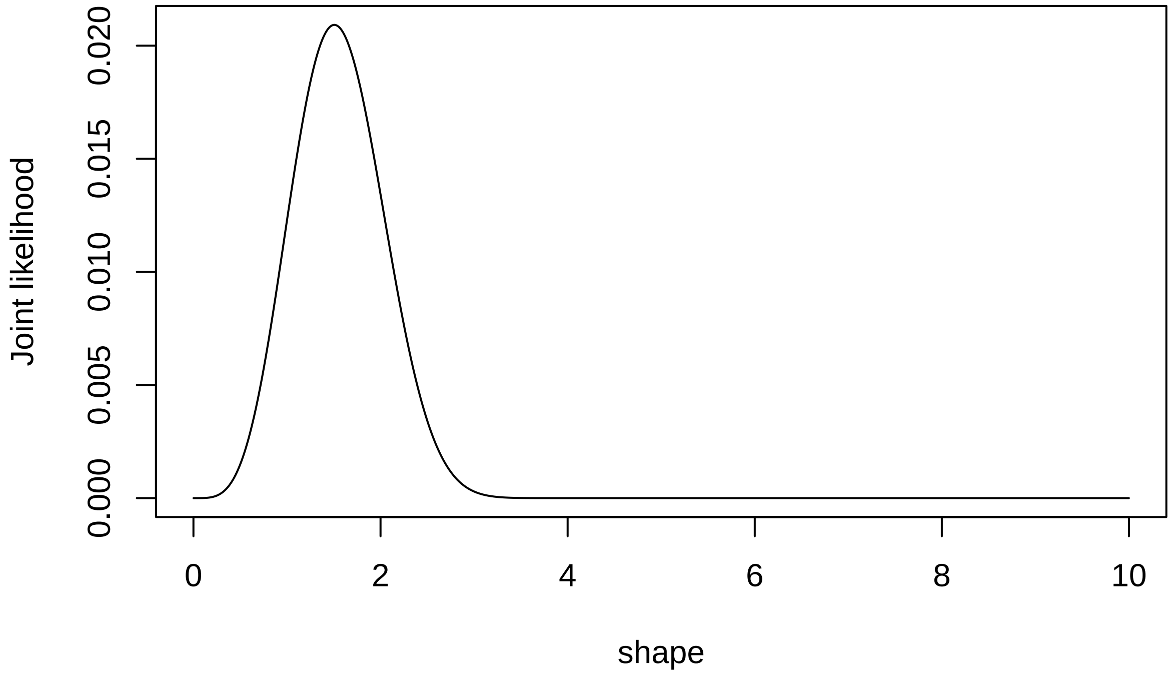
24. Plot the likelihood function $L(a)$ for this situation given the first measurement only (x=0.3). Make sure you set the upper limit for the shape parameter high enough to see the full distribution, and use a finely-spaced grid (0.01 or less). Note that a=0 is not a mathematically valid value.

```
x = 0.3
shape = seq(0.001,10,0.001)
plot(shape, dweibull(x, shape, 1), type='l', ylab='Likelihood for x=0.3')
```



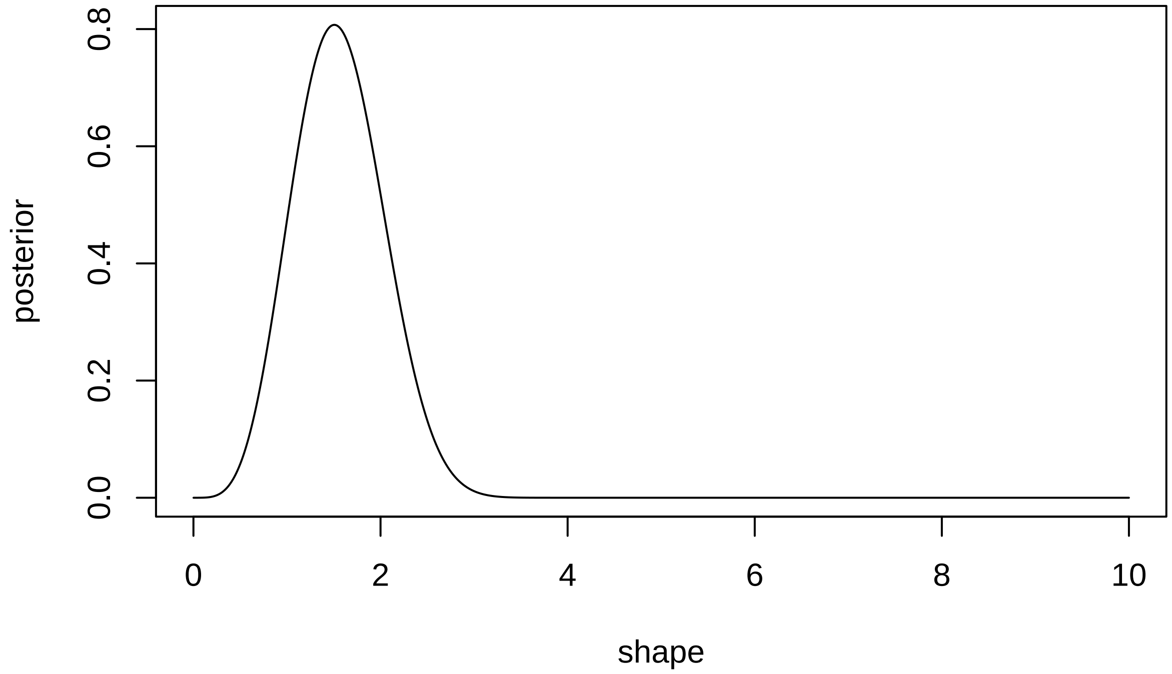
25. Calculate the joint likelihood function given all four data points by taking the product of the individual likelihood functions for each individual data point. Then make a plot of this function. (You may want to adjust the x axis range.)

```
x = c(0.3, 2.1, 0.8, 1.2)
jointL = rep(1.0, length(shape))
for (i in 1:length(x)) jointL = jointL * dweibull(x[i], shape, 1)
plot(shape, jointL, type='l', xlim=c(0,10), ylab='Joint likelihood')
```



26. Calculate the posterior probability density given our data and assuming a flat prior. To do this, calculate the the numerical integral of the joint likelihood (= sum(L*stepsize)) and divide the joint likelihood by this quantity. Then, make a plot of the posterior probability density.

```
posterior = jointL / sum(jointL*0.001)
plot(shape, posterior, type='l', xlim=c(0,10))
```



27. Find the maximum likelihood value of the shape parameter a by finding the position (value of a) where the posterior density is maximized.

```
shape[posterior==max(posterior)]
```

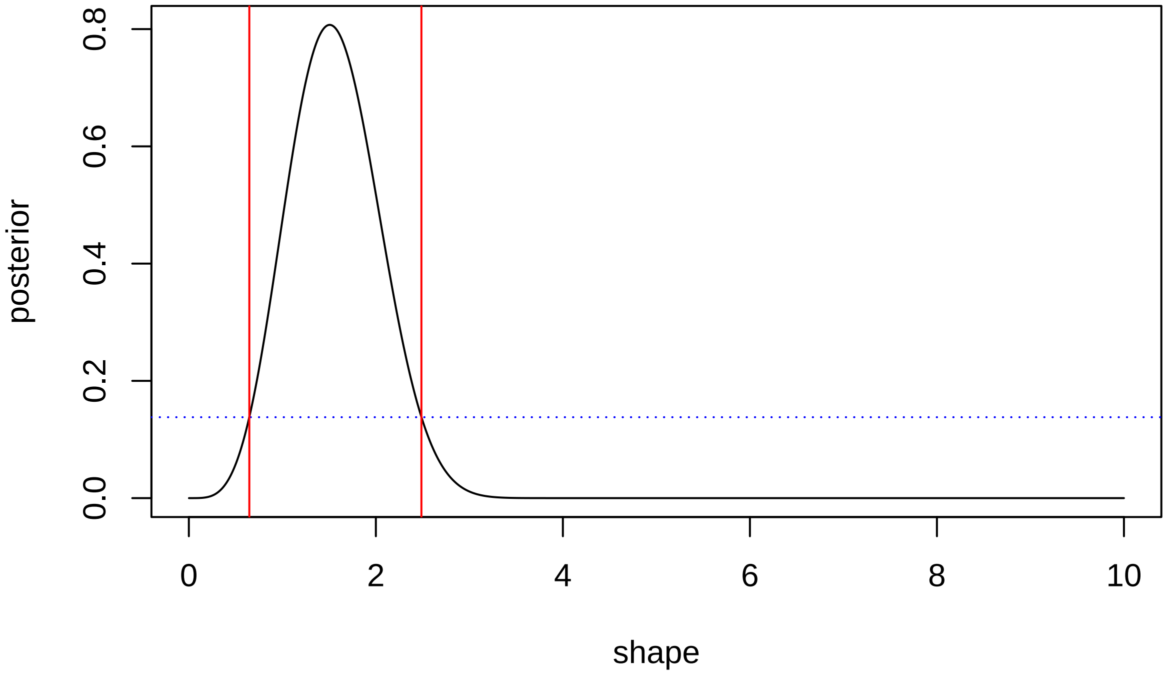
```
## [1] 1.506
```

28. Calculate an approximate 95% credible interval on the shape a, by finding two values of a such that the integral of the posterior density between these points is 0.95. (If possible, do this by finding a value of P such that the integral of the values of the posterior density >= P is 0.95. Then report the maximum and minimum values of a for which P is above this value.)

```
for (P in seq(0.0,max(posterior),0.001)) {
probenclosed = sum(posterior[posterior > P]*0.001)
if (probenclosed < 0.95) break
}
CI = range(shape[posterior > P])
CI
```

```
## [1] 0.647 2.487
```

```
plot(shape, posterior, type='l', xlim=c(0,10))
abline(h=P, col='blue', lty=3)
abline(v=CI[1], col='red')
abline(v=CI[2], col='red')
```


posterior
0.0
0.2
0.4
0.6
0.8
0
2
4
6
8
10
shape