1. Emerson and Banks (1994) present data on survival of patients in a leukemia trial, at [http://lib.stat.cmu.edu/datasets/csb/ch14.dat](http://lib.stat.cmu.edu/datasets/csb/ch14.dat). Fields are described in [http://lib.stat.cmu.edu/datasets/csb/ch14.txt](http://lib.stat.cmu.edu/datasets/csb/ch14.txt). If you want to use SAS to analyze these data, assistance in reading dates can be found at [http://lib.stat.cmu.edu/datasets/csb/ch14.sas](http://lib.stat.cmu.edu/datasets/csb/ch14.sas). Entry time is the second field, treatment arm (D or I) is the third field, baseline white blood cell count is in field 7, date of last followup is field 15, and status (D for dead, A for alive) at last followup is field 16.

   a. Estimate the median time to death, and find a 95% confidence interval for the median survival time for both treatment groups separately. If this is impossible, explain why.

   ```r
   leukemia<-read.table("ch14.dat")[,c(2,3,15,16)]
   for(jj in c(1,3)){
     leukemia[[jj]][nchar(leukemia[[jj]]==5]<-paste("0",
     leukemia[[jj]][nchar(leukemia[[jj]]==5],sep="")
     leukemia[[jj]]<-as.Date(leukemia[[jj]],"%m%d%y")
   }
   leukemia$etime<-leukemia[[3]]-leukemia[[1]]
   leukemia$status<-leukemia$V16=="D"
   
   Now fit the survival curves, and invert CIs for the median confidence interval:
   
   library(survival)
   print(sf<-survfit(Surv(etime,status)~V3,data=leukemia),conf.int=.95)
   
   to obtain
   
   n events median 0.95LCL 0.95UCL
   V3=D 65 49 434 350 497
   V3=I 65 38 507 383 913
   
   b. Estimate the .75 quantile for time to death, and find a 95% confidence interval for this quantile of survival time for both treatment groups separately. If this is impossible, explain why.

   ```

   ```r
   quantile(sf,.75)
   to see
   ```
2. Consider the data set from McGilchrist and Aisbett, Biometrics 47, 461-66, 1991 at http://lib.stat.cmu.edu/datasets/kidney. The data set consists of times until infection or end of followup for two catheters for each of 38 subjects. The first column is subject number. For this question retain only the first observation, and delete the second, for each subject. Time and status (1 for infection, 0 for loss to followup) are in the second and third columns. Sex is in the fifth column. Use the log rank test to test for a difference in time until infection between men and women.

Read in the data:

```r
kidney<-as.data.frame(scan("kidney",skip=16,what=list(patient=0,
    time=0,status=0,age=0,sex=0,diseasetype=0,frailty=0))[-2*seq(38),]
```

Test for a difference by gender:

```r
clibrary(survival)
survdiff(Surv(time,status)~sex,data=kidney)
```

to observe

Chisq = 13.2 on 1 degrees of freedom, p = 0.0003.

Reject the null hypothesis of no difference in lifetime by gender.

3. The data set at http://lib.stat.cmu.edu/datasets/csb/ch15.dat contains results compiled by the Cooperative Early Lung Cancer Detection Program. A full description of the data set can be found at http://lib.stat.cmu.edu/datasets/csb/ch15.txt, and SAS code to read the data can be found at http://lib.stat.cmu.edu/datasets/csb/ch15.sas. Focus attention on three variables: cancer cell type (variable 5), stage (variable 6), survival time in days
Test the hypothesis of no difference in survival among individuals with the various cancer cell types.

*R commands are*

```r
fle<-as.data.frame(scan("ch15.dat",what=list(number=0,inst=0,gr=0,
  det=0,ct=0,op=0,a=0,b=0,c=0, surv=0, status=0),na.strings="."))
fle$delta<-fle$status==1
library(survival)
survdiff(Surv(surv,delta)~factor(ct),data=fle)
```

*to obtain*

*Chisq= 112 on 4 degrees of freedom, p= <2e-16*

Reject the null hypothesis of no difference. Note here that the SAS and R commands are calculating different statistics, explaining the difference in statistic values.

b. Estimate mean survival time by cancer cell type.

*R commands to do these calculations are*

```r
print(survfit(Surv(surv,delta)~factor(ct),data=fle),rmean="common")
```

*and important parts of the output are*

<table>
<thead>
<tr>
<th>factor(ct)</th>
<th>rmean</th>
<th>se(rmean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>160</td>
<td>1876</td>
</tr>
<tr>
<td>1</td>
<td>208</td>
<td>1641</td>
</tr>
<tr>
<td>2</td>
<td>126</td>
<td>1136</td>
</tr>
<tr>
<td>3</td>
<td>179</td>
<td>859</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>1447</td>
</tr>
</tbody>
</table>

By this measure, cell types 0 and 1 yeild similar lifetimes, 2 and 3 also yeild similar lifetimes, and 4 is intermediate between these. The mean time R estimates is truncated at a more or less arbitary value if the survival curves do not all hit zero; the syntax above takes this value as the maximum time value in the data set. Fortunately, this was not necessary, as all curves hit zero.

c. Draw the survival curves corresponding to patients with the five cell types. Do apparent differences appear to be of the type most easily detected by the log rank test?

*See Fig. 1, created by*

```r
plot(survfit(Surv(surv,delta)~factor(ct),data=fle),lty=1:5,
  main="Survival for different cell types",xlab="Days",ylab="Survival")
legend(2000,1,lty=1:5,legend=paste("Type",0:4))
```
The proportionality of hazards assumption looks OK except for type 4, but the estimate is based on few observations and we might ignore non-proportionality here.

d. Differences in the virulence of various cell types might be attributable to differential ease of detection. Perform the calculations of part (a) controlling for stage at detection.

Do the log rank test for ct stratified on stage.

\texttt{survdiff(Surv(surv,delta)~factor(ct)+strata(st),data=fle)}

The important part of the output is

\texttt{Chisq= 26.4 on 4 degrees of freedom, p= 3e-05}

\textit{Reject the null hypothesis of no effect of cell type.}