Survival Data Analysis & Lab.
Assignment #2

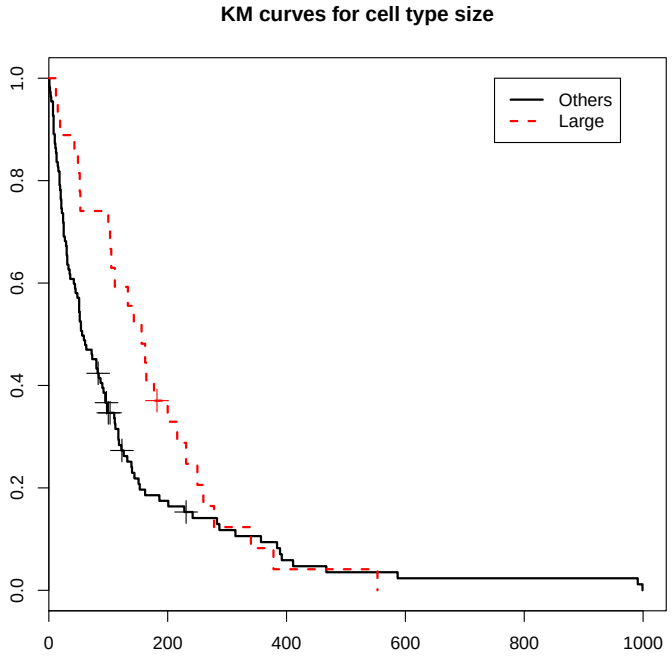
To answer the questions below, you will need to use an R program that computes and plots KM curves and computes the log-rank test. **Provide your R program as well as your solution.**

1. For the vets.dat data set described in the lecture note:

   (a) Obtain KM plots for the two categories of the variable cell type 1 (1=large, 0=other). Comment on how the two curves compare with each other. Carry out the log-rank, and draw conclusions from the test.

   (Solution)

```
library(survival)
vet.dat<-read.table('vet.dat',sep='\t',header=TRUE)
vet<-Surv(vet.dat[,6],vet.dat[,11])
grp<-vet.dat[,2]
res<-survfit(vet~grp)
plot(res,main='KM curves for cell type size', lty=1:2, col=1:2, lwd=2, cex=2)
legend(750,1,c('Others', 'Large'), lty=1:2, col=1:2, lwd=2)
survdiff(vet~grp)
```



From the above plots, group 'Large' shows lower survivor function estimate than group 'Others' up to about 300 days from the study start, and then both of estimates becomes similar with each other until the study ends.

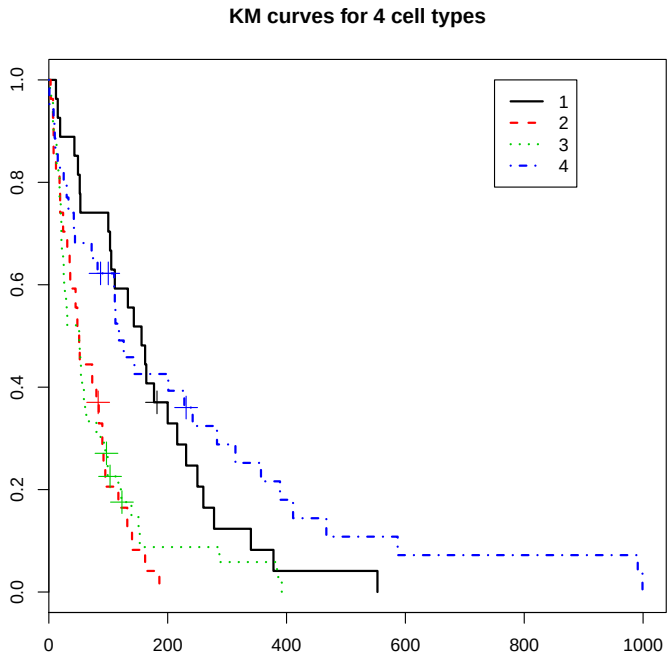
```
> survdiff(vet~grp)
Call:
```

```
survdiff(formula = vet ~ grp)

        N Observed Expected (O-E)^2/E (O-E)^2/V
grp=0 110      102     93.5     0.782      3.02
grp=1  27       26     34.5     2.116      3.02

 Chisq= 3  on 1 degrees of freedom, p= 0.0822
```

The log-rank test statistic is obtained as 3 (or 3.02) and its p-value is given as 0.0822, which is larger than $\alpha = 0.05$. We conclude that there is no statistical evidence that two groups have different survival time distributions under the significance level under $\alpha = 0.05$.

(b) Obtain KM plots for the four categories of cell type-large, adeno, small, and squamous. Note that you will need to recode the data to define a single variable which numerically distinguishes the four categories (e.g., 1=large, 2=adeno, 3=small, and 4=squamous). As in part a, compare the four KM curves. Also, carry out the log-rank for the equality of the four curves and draw conclusions. **(Solution)**

```
library(survival)
vet.dat<-read.table('vet.dat',sep='\t',header=TRUE)
vet<-Surv(vet.dat[,6],vet.dat[,11])
grp<-rep(0,nrow(vet.dat))
grp[vet.dat[,2]==1]<-1 #Large group
grp[vet.dat[,3]==1]<-2 #Adeno group
grp[vet.dat[,4]==1]<-3 #Small group
grp[vet.dat[,5]==1]<-4 #Squamous group
res<-survfit(vet~grp)
plot(res,main='KM curves for 4 cell types', lty=1:4, col=1:4, lwd=2, cex=2)
legend(750,1,1:4, lty=1:4, col=1:4, lwd=2)
survdiff(vet~grp)
```




```
> survdiff(vet~grp)
Call:
```

```
survdiff(formula = vet ~ grp)

       N Observed Expected (O-E)^2/E (O-E)^2/V
grp=1 27       26     34.5      2.12      3.02
grp=2 27       26     15.7      6.77      8.19
grp=3 48       45     30.1      7.37     10.20
grp=4 35       31     47.7      5.82     10.53

 Chisq= 25.4  on 3 degrees of freedom, p= 1.27e-05
```

The log-rank test statistic is obtained as 25.4 and its p-value is given as 1.27e-05, which is smaller than $\alpha = 0.05$. We conclude that there is a strong statistical evidence that four groups have different survival time distributions under the significance level under $\alpha = 0.05$.

2. The following data set consists of remission survival times on 42 leukemia patients, half of whom get a certain new treatment therapy and the other half of whom get a standard treatment therapy. The exposure variable of interest is treatment status ($Rx = 0$ if new treatment, $Rx = 1$ if standard treatment). Two other variables for control as potential confounders are log white blood cell count (i.e., logWBC) and sex. Failure status is defined by the relapse variable (0 if censored, 1 if failure). The data set is listed as follows:

Subj	Survt	Relapse	Sex	log WBC	Rx
1	35	0	1	1.45	0
2	34	0	1	1.47	0
3	32	0	1	2.20	0
4	32	0	1	2.53	0
5	25	0	1	1.78	0
6	23	1	1	2.57	0
7	22	1	1	2.32	0
8	20	0	1	2.01	0
9	19	0	0	2.05	0
10	17	0	0	2.16	0
11	16	1	1	3.60	0
12	13	1	0	2.88	0
13	11	0	0	2.60	0
14	10	0	0	2.70	0
15	10	1	0	2.96	0
16	9	0	0	2.80	0
17	7	1	0	4.43	0
18	6	0	0	3.20	0
19	6	1	0	2.31	0
20	6	1	1	4.06	0
21	6	1	0	3.28	0
22	23	1	1	1.97	1
23	22	1	0	2.73	1
24	17	1	0	2.95	1
25	15	1	0	2.30	1
26	12	1	0	1.50	1
27	12	1	0	3.06	1

28	11	1	0	3.49	1
29	11	1	0	2.12	1
30	8	1	0	3.52	1
31	8	1	0	3.05	1
32	8	1	0	2.32	1
33	8	1	1	3.26	1
34	5	1	1	3.49	1
35	5	1	0	3.97	1
36	4	1	1	4.36	1
37	4	1	1	2.42	1
38	3	1	1	4.01	1
39	2	1	1	4.91	1
40	2	1	1	4.48	1
41	1	1	1	2.80	1
42	1	1	1	5.00	1

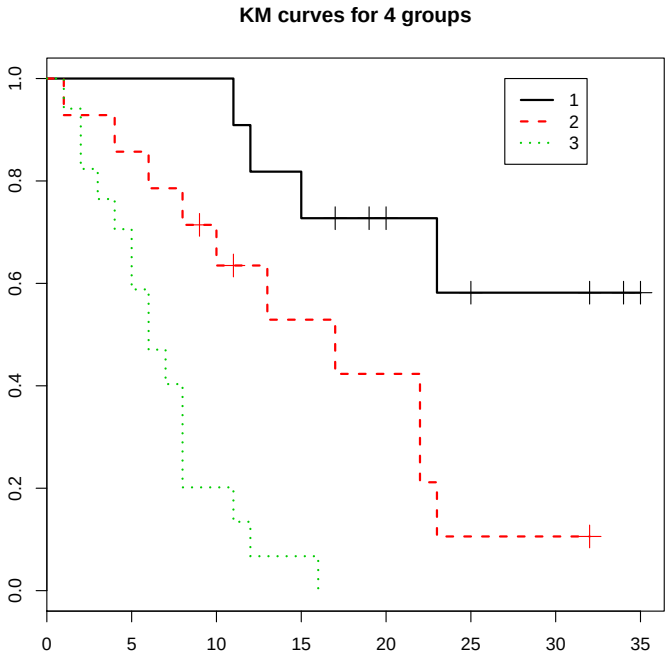
(a) Suppose we wish to describe KM curves for the variable log WBC. Because log WBC is continuous, we need to categorize this variable before we compute KM curves. Suppose we categorize log WBC into three categories–low, medium, and high–as follows:

low (0–2.30), $n = 11$;
medium (2.31–3.00), $n = 14$;
high ($>$3.00), $n = 17$.

Base on this categorization, compute and graph KM curves for each of the three categories of log WBC.

(Solution)

```
library(survival)
survt<-c(35,34,32,32,25,23,22,20,19,17,16,13,11,10,10,9,7,6,6,6,6,23,22,17,15,12,12,11,11,8,8,8,
8,5,5,4,4,3,2,2,1,1)
relapse<-c(0,0,0,0,0,1,1,0,0,0,1,1,0,0,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
logWBC<-c(1.45,1.47,2.20,2.53,1.78,2.57,2.32,2.01,2.05,2.16,3.60,2.88,2.60,2.70,2.96,2.80,4.43,
3.20,2.31,4.06,3.28,1.97,2.73,2.95,2.30,1.50,3.06,3.49,2.12,3.52,3.05,2.32,3.26,3.49,3.97,4.36,
2.42,4.01,4.91,4.48,2.80,5.00)
time<-Surv(survt,relapse)
grp<-rep(0,length(logWBC))
grp[logWBC<=2.30]<-1 #low
grp[logWBC>=2.31 & logWBC<=3.00]<-2 #medium
grp[logWBC>3.00]<-3 #high
res<-survfit(time~grp)
plot(res,main='KM curves for 4 groups', lty=1:3, col=1:3, lwd=2, cex=2)
legend(27,1,1:3, lty=1:3, col=1:3, lwd=2)
survdiff(time~grp)
```



(b) Compare the three KM plots you obtained in part a. How are they different?

(Solution) Group 1 has consistently higher survivor function than Group 2 and Group 3 has consistently lower survivor function than Group 2 across all survival time.

(c) Below is an edited printout of the log-rank test comparing the three groups.

Group	Events observed	Events expected
1	4	13.06
2	10	10.72
3	16	6.21
Total	30	30.00

Log-rank=26.39 and P-value=0.0000

Fill out the blanks in the table. What do you conclude about whether or not the three survival curves are the same?

(Solution) Since p-value is less than $\alpha = 0.05$, we statistically rejects that three survival curves are the same under $\alpha = 0.05$.

```
> survdiff(time~grp)
Call:
survdiff(formula = time ~ grp)

       N Observed Expected (O-E)^2/E (O-E)^2/V
grp=1 11        4    13.06    6.2880    12.769
```

```
grp=2 14       10    10.72    0.0489     0.081
grp=3 17       16     6.21   15.4173    23.104

 Chisq= 26.4  on 2 degrees of freedom, p= 1.86e-06
```