Assignment 8

Sri Harsha Sudalagunta

2023-04-27

Problem 1.

(a) Given margin of error = 0.2 confidence level = 0.98 Significance level = 1- Confidence level = 1-0.98 = 0.02 Critical Value at alpha = 0.02 is 2.508 using conservative value p = 0.50

```
me = 0.02
p = 0.5
alpha = 0.02
z = qnorm(1-alpha/2)


n = ((z*z)*(p)*(1-p))/(me^2)
n
```

```
## [1] 3382.434
```

The sample size required is 3383

```
n = 3383
p = 20
total = n*p
total
```

```
## [1] 67660
```

The minimum amount required is 67660

(b)

If the researcher uses fewer subjects the confidence interval will be wider.

Rationale : as the sample size and margin of error are inversely proportional.The decrease in the sample size can increase the margin of error, which widens the confidence interval.

Problem 2.

The necessary changes made to the csv and uploaded to the r studio

```
# Load the dataset from a CSV file
mydata <- read.csv("C:\\Users\\SRI HARSHA S\\Downloads\\survival.csv")
mydata
```

```
##    Days smoke_code followup_code
## 1     2          1             0
## 2    10          1             1
## 3    15          1             1
## 4    20          0             0
## 5    30          0             0
## 6    50          1             1
## 7    60          0             0
## 8    70          1             1
## 9   120          1             1
## 10  120          1             1
## 11  120          0             0
## 12  140          0             0
## 13  250          0             1
## 14  150          1             1
## 15  160          1             1
## 16  160          0             0
## 17  160          1             1
## 18  180          1             1
## 19  200          0             1
## 20  250          0             0
## 21  250          0             0
## 22  250          0             0
## 23  300          0             0
## 24  300          1             0
## 25  350          0             0
## 26  350          0             0
## 27  350          1             0
## 28  500          1             0
## 29  500          0             1
## 30  600          0             0
```

Problem 3

The attached pdf have the kaplanmeier estimates made using the excel

```
kapdata <- read.csv("C:\\Users\\SRI HARSHA S\\OneDrive - Indiana University\\Documents\\R\\kpmeier_hand

kapdata
```

```
##    Separate.Kaplan.Meier..Estimates.for.Smokers.and.Non.Smokers      X
## 1                                                   Non-Smokers
## 2                                                          Days atrisk
## 3                                                            20     16
## 4                                                            30     16
## 5                                                            60     16
## 6                                                           120     16
## 7                                                           140     16
```

```
## 8                                                                 250     16
## 9                                                                 160     15
## 10                                                                200     15
## 11                                                                250     14
## 12                                                                250     14
## 13                                                                250     14
## 14                                                                300     14
## 15                                                                350     14
## 16                                                                350     14
## 17                                                                500     14
## 18                                                                600     13
##          X.1         X.2    X.3 X.4 X.5     X.6    X.7        X.8         X.9
## 1                                    NA  NA
## 2  dead_code           p   s(t)  NA  NA smokers
## 3          0           1      1  NA  NA    Days atrisk death_code           p
## 4          0           1      1  NA  NA       2     14          0           1
## 5          0           1      1  NA  NA      10     14          1 0.928571429
## 6          0           1      1  NA  NA      15     13          1 0.923076923
## 7          0           1      1  NA  NA      50     12          1 0.916666667
## 8          1      0.9375 0.9375  NA  NA      70     11          1 0.909090909
## 9          0           1 0.9375  NA  NA     120     10          1         0.9
## 10         1 0.933333333  0.875  NA  NA     120      9          1 0.888888889
## 11         0           1  0.875  NA  NA     150      8          1       0.875
## 12         0           1  0.875  NA  NA     160      7          1 0.857142857
## 13         0           1  0.875  NA  NA     160      6          1 0.833333333
## 14         0           1  0.875  NA  NA     180      5          1         0.8
## 15         0           1  0.875  NA  NA     300      4          0           1
## 16         0           1  0.875  NA  NA     350      4          0           1
## 17         1 0.928571429 0.8125  NA  NA     500      4          0           1
## 18         0           1 0.8125  NA  NA
##         X.10
## 1         NA
## 2         NA
## 3         NA
## 4  1.0000000
## 5  0.9285714
## 6  0.8571429
## 7  0.7857143
## 8  0.7142857
## 9  0.6428571
## 10 0.5714286
## 11 0.5000000
## 12 0.4285714
## 13 0.3571429
## 14 0.2857143
## 15 0.2857143
## 16 0.2857143
## 17 0.2857143
## 18        NA
```

Problem 4.

In the attached pdf, we have the log rank test done using the excel. Based on the test statistic i.e., 10.7 we can reject the null hypothesis i.e., two survival curves are equal.

```
logdata <- read.csv("C:\\Users\\SRI HARSHA S\\OneDrive - Indiana University\\Documents\\R\\logrank_hand
logdata
```

```
##    Non.Smokers      X         X.1            X.2      X.3 X.4 smokers     X.5
## 1         Days atrisk   dead_code            exp       NA  NA    Days  atrisk
## 2           20     16           0              0       NA  NA       2      14
## 3           30     16           0              0       NA  NA      10      14
## 4           60     16           0              0       NA  NA      15      13
## 5          120     16           0              0       NA  NA      50      12
## 6          140     16           0              0       NA  NA      70      11
## 7          250     16           1    0.615384615       NA  NA     120      10
## 8          160     15           0              0       NA  NA     120       9
## 9          200     15           1    0.652173913       NA  NA     150       8
## 10         250     14           0              0       NA  NA     160       7
## 11         250     14           0              0       NA  NA     160       6
## 12         250     14           0              0       NA  NA     180       5
## 13         300     14           0              0       NA  NA     300       4
## 14         350     14           0              0       NA  NA     350       4
## 15         350     14           0              0       NA  NA     500       4
## 16         500     14           1              1       NA  NA
## 17         600     13           0              0       NA  NA
## 18                             3    2.267558528       NA  NA
## 19                                                    NA  NA         Totals
## 20     Totals     O 2          3                     NA  NA
## 21                E 2 2.267558528                     NA  NA
## 22                                                    NA  NA
## 23                                                    NA  NA
## 24                                 Test Statistic 10.62366  NA
##           X.6         X.7
## 1  death_code         exp
## 2           0           0
## 3           1 0.466666667
## 4           1 0.448275862
## 5           1 0.428571429
## 6           1 0.407407407
## 7           1 0.384615385
## 8           1       0.375
## 9           1 0.347826087
## 10          1 0.333333333
## 11          1         0.3
## 12          1 0.263157895
## 13          0           0
## 14          0           0
## 15          0           0
## 16         10 3.754854064
## 17
## 18
## 19         O1          10
## 20         E1     3.75485
```

```
## 21
## 22
## 23
## 24
```

Problem 5.

The log rank test using the r studio

```
library(survival)
```

```
## Warning: package 'survival' was built under R version 4.2.3
```

```
survdiff(Surv(Days, followup_code) ~ smoke_code, data = mydata)
```

```
## Call:
## survdiff(formula = Surv(Days, followup_code) ~ smoke_code, data = mydata)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## smoke_code=0 16        3     8.24      3.33      9.45
## smoke_code=1 14       10     4.76      5.75      9.45
##
##  Chisq= 9.4  on 1 degrees of freedom, p= 0.002
```

Based on the test statistic and p value we can reject the null hypothesis i.e., equal survival curves for the smokers and non smokers, as the statistic value is greater than the critical value.

Problem 6.

The Cox proportional hazards regression model from the data

```
coxph(Surv(Days, followup_code) ~ smoke_code, data = mydata)
```

```
## Call:
## coxph(formula = Surv(Days, followup_code) ~ smoke_code, data = mydata)
##
##              coef exp(coef) se(coef)     z       p
## smoke_code 1.8167    6.1514   0.6648 2.733 0.00628
##
## Likelihood ratio test=9.23  on 1 df, p=0.002381
## n= 30, number of events= 13
```

```
summary(coxph(Surv(Days, followup_code) ~ smoke_code, data = mydata))
```

```
## Call:
## coxph(formula = Surv(Days, followup_code) ~ smoke_code, data = mydata)
##
##   n= 30, number of events= 13
##
```

```
##              coef exp(coef) se(coef)     z Pr(>|z|)
## smoke_code 1.8167    6.1514   0.6648 2.733  0.00628 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##            exp(coef) exp(-coef) lower .95 upper .95
## smoke_code     6.151     0.1626     1.671     22.64
##
## Concordance= 0.768  (se = 0.045 )
## Likelihood ratio test= 9.23  on 1 df,   p=0.002
## Wald test            = 7.47  on 1 df,   p=0.006
## Score (logrank) test = 9.59  on 1 df,   p=0.002
```

In the cox proportional we can get the hazard ratio which we cannot get in the log rank test.