Analysis of Time-to-Event Data: Study Sheet 4,

Submission by: René-Marcel Kruse

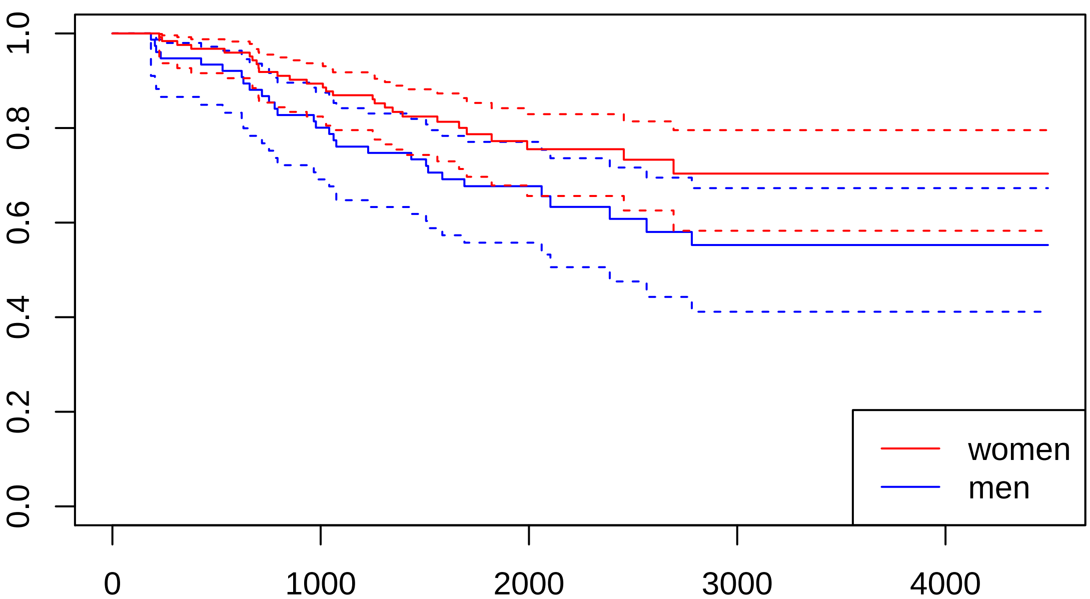
Exercise 1:

Recall the data file melanoma.dat that was analysed previously. In R, one can use the survdiff() function from the package survival to test whether male and female patients have the same survival function.

(a)

Create a graph of the Kaplan-Meier estimator stratified by gender along with 95% pointwise confidence bounds using the complementary log-log transformation (log-log). What do you conclude from this plot regarding the equality of the genderspecific survivor functions?

Answer:



(b)

Conduct both a log-rank test and a Wilcoxon test to test the null hypothesis that there is no difference in the survivor functions for men and women. Interpret the obtained output!

Answer:

```
survdiff(Surv(melanoma$time,delta) ~ melanoma$sex, rho=0)
```

```
## Call:
## survdiff(formula = Surv(melanoma$time, delta) ~ melanoma$sex,
##     rho = 0)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## melanoma$sex=0 126       28     37.1      2.25      6.47
## melanoma$sex=1  79       29     19.9      4.21      6.47
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.01
```

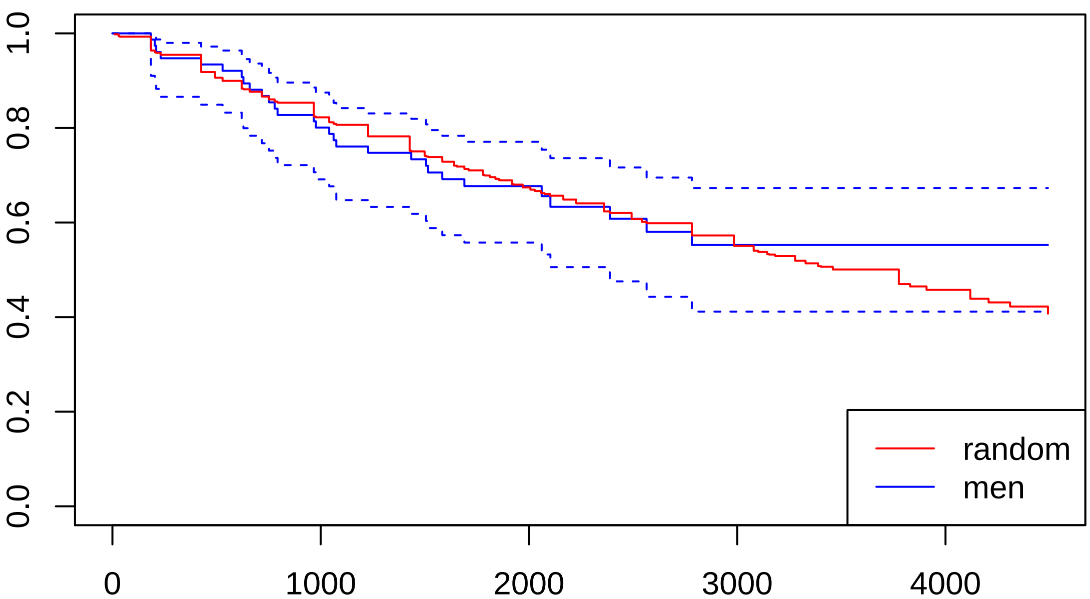
```r
survdiff(Surv(melanoma$time,delta)~ melanoma$sex, rho=1)
```

```
## Call:
## survdiff(formula = Surv(melanoma$time, delta) ~ melanoma$sex,
##     rho = 1)
##
##                  N Observed Expected (O-E)^2/E (O-E)^2/V
## melanoma$sex=0 126     23.4     31.6      2.14      7.09
## melanoma$sex=1  79     25.2     17.0      3.98      7.09
##
##  Chisq= 7.1  on 1 degrees of freedom, p= 0.008
```

(c)

Create a plot of the estimated Kaplan-Meier estimate of the survivor function for male patients with 95% pointwise confidence bounds using the complementary log-log transformation together with the survivor function of an exponentially distributed random variable $T \sim \mathcal{E}(0.0002)$. Use the log-rank test to test the null hypothesis that the distribution of the survival time of men is exponential with $\lambda = 0.0002$. Hint: Make use of `help(survdiff)` *to find out how a one sample test can be performed.*

Answer:




```
## Call:
## survdiff(formula = melaSurv.maenner ~ offset(exp(-2e-04 * km.maenner$time)),
##     rho = 0)
##
## Observed Expected        Z        p
##   29.000   30.742    0.314    0.753
```

Exercise 2:

An approximation to the log-rank test statistic $W_L = U_L^2 = V_L$ that was introduced in the lecture for comparing two survivor curves, and which avoids computing the variance V_L, is as follows:

$$X^2 = \sum_{k=1}^{2} \frac{(O_k - E_k)^2}{E_k} \sim \chi_1^2,$$

where $O_k = \sum_{j=1}^{r} d_{kj}$ and $E_k = \sum_{j=1}^{r} e_{kj}$ denotes, for group $k (k = 1, 2)$, the sum of the observed and expected counts over all r distinct failure times across the two groups, respectively. Using the approximation formula (1), carry out a log-rank test for the breast cancer data example on slide 2 of the set of slides "Nonparametric methods for comparing survival distributions". Hint: You may use some of the results given on slide 9. Compare your results with the ones obtained in the lecture using the test statistic W_L.

Answer:

We do know that the observed are denoted as:

$$O_k = \sum_{j=1}^{r} d_{kj},$$

and the expected are denoted as:

$$E_k = \sum_{j=1}^{r} \frac{n_{kj} \cdot d_j}{n_j}.$$

The Log-Rank Test results for the underlying data is as follows:

$$O_1 = \sum_{j=1}^{r} d_{1,j} = d_{1,1} + d_{1,2} + \cdots + d_{1,25} = 5$$

$$O_2 = \sum_{j=1}^{r} d_{2,j} = d_{2,1} + d_{2,2} + \cdots + d_{2,25} = 21$$

$$E_1 = \sum_{j=1}^{r} \frac{n_{1j} \cdot d_j}{n_1} = \frac{n_{1,1} \cdot d_1}{n_1} + \frac{n_{1,2} \cdot d_2}{n_2} + \cdots + \frac{n_{1,25} \cdot d_{25}}{n_{25}} = 9.5652$$

We can calculate the value of E_2 based on the imposed relationship of $O_1 + O_2 = E_1 + E_2$. Following this notation, we can rearrange the relationship as follows:

$$\begin{aligned} O_1 + O_2 &= E_1 + E_2 \\ E_2 &= O_1 + O_2 - E_1 \\ E_2 &= 5 + 21 - 9.5652 \\ E_2 &= 16.4348 \end{aligned}$$

Now we only have to insert the calculated values into the original equation for X^2.

$$\begin{aligned} X^2 &= \sum_{k=1}^{2} \frac{(O_k - E_k)^2}{E_k} \\ X^2 &= \frac{(5 - 9.5652)^2}{9.5652} + \frac{(21 - 16)^2}{16.4348} \\ X^2 &= 3.4469 \end{aligned}$$

Resulting in a approximated p-value of: 0.0633.

Exercise 3:

Five hundred and ninety-five persons participate in a case control study of the association of cholesterol and coronary heart disease (CHD). Among them, 300 persons are known to have CHD and 295 are free of CHD. To find out if elevated cholesterol is significantly associated with CHD, the investigator decides to control the effects of smoking. The study subjects are then divided into two strata: smokers and nonsmokers. The following two tables provide the data.

Smokers

	with CHD	Without CHD	Total
Elevated Cholesterol			
Yes	120	20	140
No	80	60	140
Total	200	80	280

Nonsmokers

	with CHD	Without CHD	Total
Elevated Cholesterol			
Yes	30	60	90
No	70	155	225
Total	100	215	315

Conduct an appropriate test to judge whether elevated cholesterol is significantly associated with CHD after adjusting for the effects of smoking.

Answer:

We have a two strata problem, therefore, we have to use a Stratified Log-Rank Test. This test can be formulated as follows:

$$W_s = \frac{(\sum_{k=1}^{S} U_{L,K})^2}{\sum_{k=1}^{S} V_{L,K}} \sim \chi(1)^2.$$

Applied to the underlying question at hand we get:

$$W_s = \frac{(U_{L,smoker} + U_{L,nonsmoker})^2}{V_{L,smoker} + V_{L,nonsmoker}}.$$

Where as the variable U_L is defined as:

$$U_L = \sum_{j=1}^{r} = (d_{1,j} - e_{1,j}) = d_1 - e_1.$$

Hence we can assume that:

$$\begin{aligned}
U_{L,smoker} &= d_{1,smoker} - e_{1,smoker} \\
&= d_{CHD,smoker} - e_{CHD,smoker} \\
&= 120 - (\frac{200 \cdot 140}{280}) \\
&= 120 - 100 \\
&= 20.
\end{aligned}$$

Similary we can assume that:

$$U_{L,nonsmoker} = 30 - \frac{100 \cdot 90}{315} = 1.429$$

To calculate $V_{L,smoker}$ we do impose the following equation:

$$\begin{aligned} V_{L,smoker} &= \frac{n_{CHD,smoker} \cdot n_{NoCHD,smoker} \cdot d_{smoker} \cdot (n_{smoker} - d_{smoker})}{n_{smoker}^2 \cdot (n_{smoker} - 1)} \\ &= \frac{200 \cdot 80 \cdot 140 \cdot (280 - 140)}{280^2 \cdot (280 - 1)} \\ &= 14.337. \end{aligned}$$

$V_{L,Non-smoker}$ can be calculated via:

$$\begin{aligned} V_{L,non-smoker} &= \frac{90 \cdot 225 \cdot 100(315 - 100)}{315^2 \cdot (315 - 1)} \\ &= 13.974. \end{aligned}$$

Now plugging in the calculated values into W_S gives:

$$W_S = \frac{(20 + 1.429)^2}{14.337 + 13.974} = 16.22 \sim \chi(1)^2$$

Using this result we arrive at a approximated p-value of $5e^{-}5$

Exercise 4:

Consider two groups with sizes $n1 = n2 = 100$ from a population of size $n = n1 + n2$. Suppose that the true survival times of the first group are distributed as $T_1 \sim \mathcal{WB}(3, 0.928)$ and that the second group has true survival times $T2$ with hazard rate $h_2(t) = 3t^2$. For both groups we assume that censoring times are independent and identically distributed. To begin with, execute the command `set.seed(1234)`.

(a)

Generate right censored survival times for both groups separately. Assume that the censoring times in both groups are exponentially distributed with parameter $\lambda = 2 = 3$. Use the inverse transform sampling method (exercise 5, study sheet 1) to generate true survival times of group 2. Combine your data (`time` = observed survival times, `delta` = censoring indicator, `group` = group membership) into a data frame.

Answer:

```
##          time delta group
## 1 0.796687161     1     1
## 2 0.370138325     0     1
## 3 0.009872935     0     1
## 4 1.261347033     1     1
## 5 0.580773875     0     1
## 6 0.134924507     0     1
```

(b)

Use a two-sample test of your choice to test whether the survivor functions of both groups are identical. Give an interpretation of the test result ($\alpha = 0.05$)

Answer:

```
## Call:
## survdiff(formula = Surv(time, delta) ~ group, data = data, rho = 0)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## group=1 100       59     56.4     0.118     0.225
## group=2 100       63     65.6     0.102     0.225
##
##  Chisq= 0.2  on 1 degrees of freedom, p= 0.6
```

```
## Call:
## survdiff(formula = Surv(time, delta) ~ group, data = data, rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## group=1 100     32.5     29.9     0.220     0.576
## group=2 100     34.2     36.8     0.179     0.576
##
##  Chisq= 0.6  on 1 degrees of freedom, p= 0.4
```

```
## [1] 122
```

(c)

Use a one-sample test of your choice to test whether the distribution of the survival times T of the whole population are distributed as $T \sim \mathcal{WB}(3, 0.928)$. Give an interpretation of the test result ($\alpha = 0.05$).

Answer:

```
## Call:
## survdiff(formula = Surv(time, delta) ~ offset(exp(-(0.928 * time)^3)),
##     data = data, rho = 0)
##
## Observed Expected        Z        p
## 122.0000  96.7913  -2.5623   0.0104
```

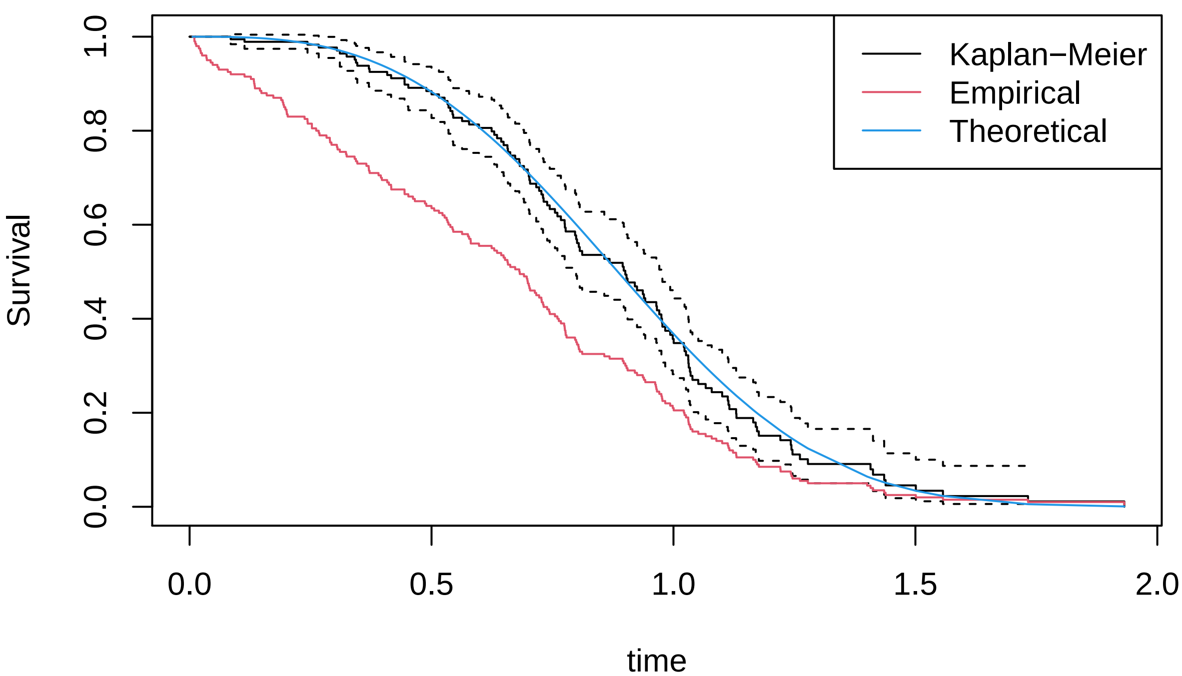
```
## Call:
## survdiff(formula = Surv(time, delta) ~ offset(exp(-(0.928 * time)^3)),
##     data = data, rho = 1)
##
##  Observed  Expected         Z         p
## 122.00000  96.79127  -2.69379   0.00706
```

(d)

Assume that the theoretical event times of the whole population are distributed as $T \sim \mathcal{WB}(3,1)$. Compute the Kaplan-Meier estimator, the theoretical survivor function and the empirical survivor function for the whole population and visualise them in a single plot. How do the results obtained in (c) change if the empirical survivor function is used as an offset?

Answer:

```
##          time delta group
## 1 0.796687161     1     1
## 2 0.370138325     0     1
## 3 0.009872935     0     1
## 4 1.261347033     1     1
## 5 0.580773875     0     1
## 6 0.134924507     0     1
```




```
## Call:
```

```
## survdiff(formula = Surv(time, delta) ~ offset(emp.Survival),
##     data = data, rho = 0)
##
## Observed Expected        Z        p
## 1.22e+02 1.96e+02 5.31e+00 1.09e-07
```

```
## Call:
## survdiff(formula = Surv(time, delta) ~ offset(exp(-(1 * time)^3)),
##     data = data, rho = 0)
##
## Observed Expected        Z        p
## 122.0000 121.1134  -0.0806   0.9360
```

(e)

For group 2, compute the mean and median lifetime and the probability to survive longer than the mean and median lifetime. Carry out calculations for both the theoretical data and the censored data.

Answer:

Mean:

```
## Empirical: 0.7024507
```

```
## Theoretical: 0.8929795
```

```
## Kaplan-Meier: 0.9018
```

Median:

```
## Empirical: 0.7409372
```

```
## Theoretical: 0.884997
```

```
## Kaplan-Meier: 0.939
```

Survival times:

```
## Theoretical survival times
```

```
## Mean: 0.4906261
```

```
## Median: 0.5
```

```
## Empirical estimator
```

```
## Mean: 0.7207001
```

```
## Median: 0.7065687
```

```
## Kaplan-Meier
```

```
## Mean: 0.5540986
```

```
## Median: 0.490773
```