# Statistical Analysis of Reliability and Survival Data: Rotterdam Dataset

**Authors**
Aksoy, Barış r0869901
Heller, Jack r0862809
Zdravković, Aleksandra r0869484

Leuven. May, 2022

# 1 Exploratory Data Analysis

These data sets are used in the paper by Royston and Altman that is referenced below. The Rotterdam data is used to create a fitted model, and the GBSG data for validation of the model. The paper gives references for the data source.

There are 43 subjects who have died without recurrence, but whose death time is greater than the censoring time for recurrence. A common way that this happens is that a death date is updated in the health record sometime after the research study ended, and said value is then picked up when a study data set is created. But it raises serious questions about censoring. For instance subject 40 is censored for recurrence at 4.2 years and died at 6.6 years; when creating the endpoint of recurrence free survival (earlier of recurrence or death), treating them as a death at 6.6 years implicitly assumes that they were recurrence free just before death. For this to be true we would have to assume that if they had progressed in the 2.4 year interval before death (while off study), that this information would also have been noted in their general medical record, and would also be captured in the study data set. However, that may be unlikely. Death information is often in a centralized location in electronic health records, easily accessed by a programmer and merged with the study data, while recurrence may require manual review. How best to address this is an open issue.

Table 1: Data description

| | |
|---|---|
| pid | Patient identifier |
| year | Year of surgery |
| age | Age at surgery |
| meno | Menopausal status (0 = premenopausal, 1 = postmenopausal) |
| size | Tumor size, a factor with levels <= 20, 20-25, > 50 |
| grade | Differentiation grade |
| nodes | Number of positive lymph nodes |
| pgr | Progesterone receptors (fmol/l) |
| er | Estrogen receptors (fmol/l) |
| hormon | Hormonal treatment (0=no, 1=yes) |
| chemo | Chemotherapy |
| rtime | Days to relapse or last follow-up |
| recur | 0 = no relapse, 1 = relapse |
| dtime | Days to death or last follow-up |
| death | 0 = alive, 1 = dead |

Table 1 explains the covariates in the Rotterdam dataset.

# 2 Further Analysis

Now the focus will be on the response variable, the censoring indicator, and the categorical variable.

## 2.1 Survival Distrubution by Levels of ...

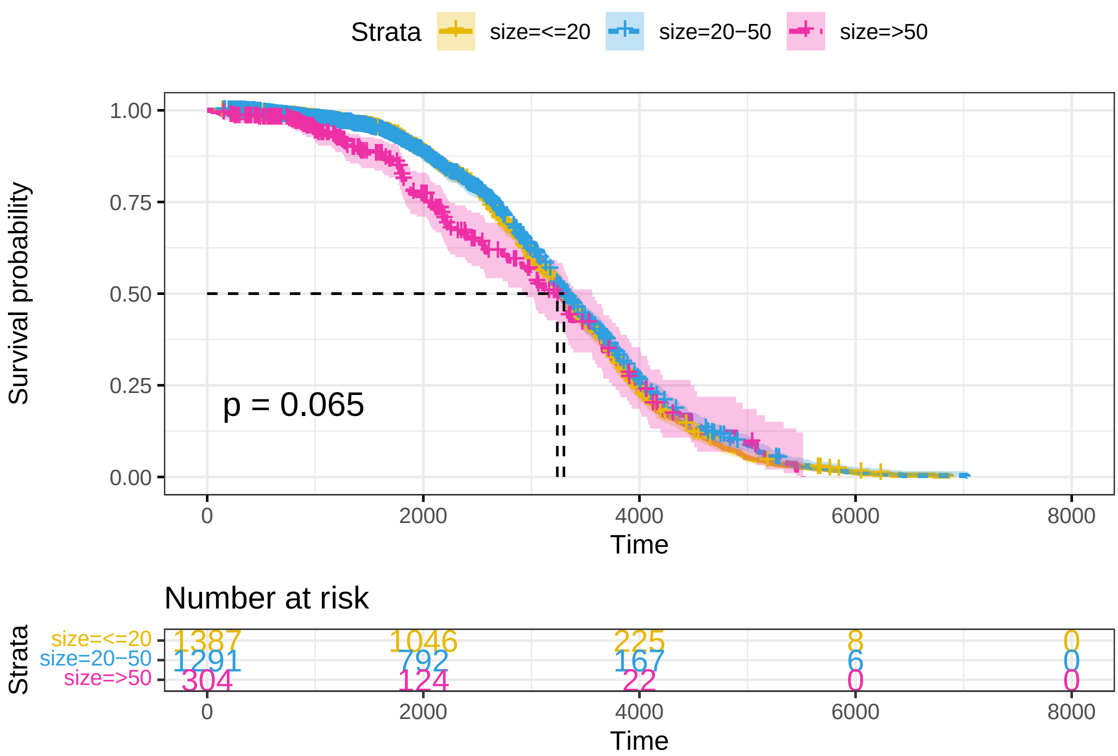
(a) For each of the levels of the categorical variable, compute the survival distribution. Plot them on the same graph. What do the graphs suggest?

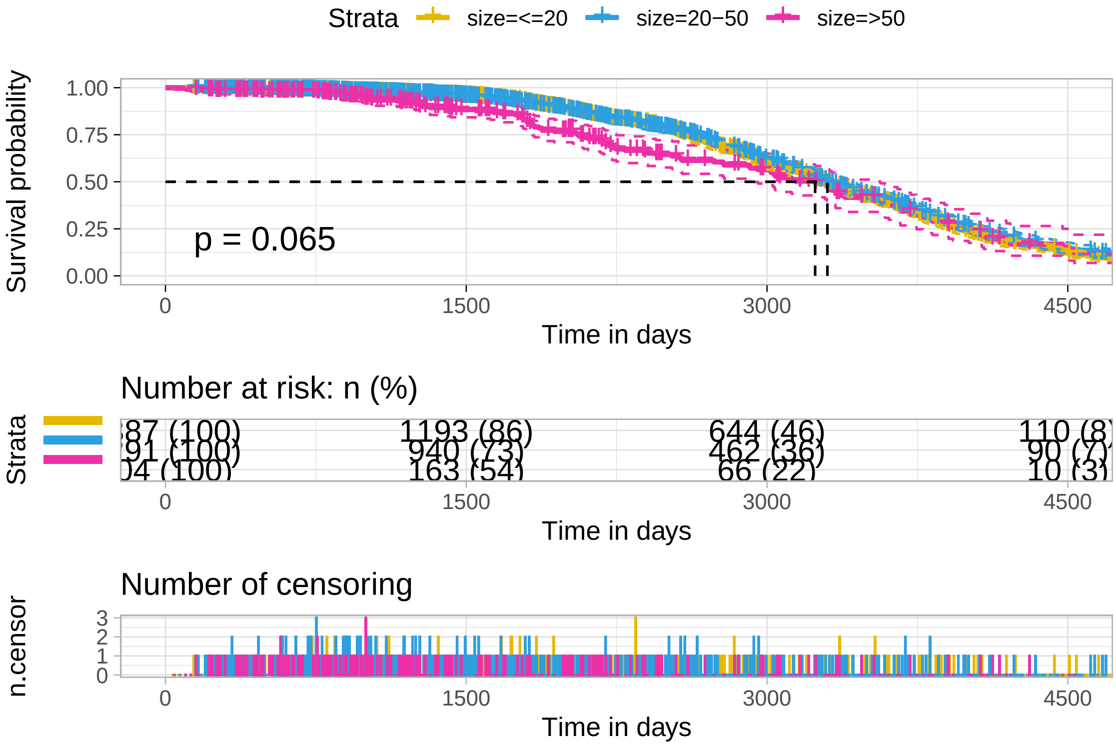
### 2.1.1 Size

```
## Call: survfit(formula = Surv(dtime, censored) ~ size, data = data)
##
##                n events median 0.95LCL 0.95UCL
## size=<=20   1387   1048   3289    3232    3370
## size=20-50  1291    734   3301    3230    3408
## size=>50     304    123   3240    2918    3565
```

```
##            records n.max n.start events    rmean se(rmean) median 0.95LCL
## size=<=20     1387  1387    1387   1048 3282.860  32.47471   3289    3232
## size=20-50    1291  1291    1291    734 3342.568  39.62310   3301    3230
## size=>50       304   304     304    123 3083.707 104.26023   3240    2918
##            0.95UCL
## size=<=20     3370
## size=20-50    3408
## size=>50      3565
```

```
##   time n.risk n.event n.censor      surv     upper     lower
## 1   36   1387       1        0 0.9992790 1.0000000 0.9978674
## 2   64   1386       1        0 0.9985580 1.0000000 0.9965631
## 3   97   1385       1        0 0.9978371 1.0000000 0.9953951
## 4  101   1384       1        0 0.9971161 0.9999422 0.9942980
## 5  129   1383       1        0 0.9963951 0.9995542 0.9932460
## 6  141   1382       0        1 0.9963951 0.9995542 0.9932460
```






The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2250, the probability of survival is approximately 0.625 for size>=50, and 0.85 for size<50. The median survival is approximately 3300 for size>=50, and a bit more for other two groups, suggesting slightly worse survival for patients with tumor of larger size. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

The median survival times for each group can be seen from:

```
##             records n.max n.start events    rmean se(rmean) median 0.95LCL
## size=<=20      1387  1387    1387   1048 3282.860  32.47471   3289    3232
## size=20-50     1291  1291    1291    734 3342.568  39.62310   3301    3230
## size=>50        304   304     304    123 3083.707 104.26023   3240    2918
##             0.95UCL
## size=<=20      3370
## size=20-50     3408
## size=>50       3565
```

### 2.1.2 Menopause

```
## Call: survfit(formula = Surv(dtime, censored) ~ meno, data = data)
##
##           n events median 0.95LCL 0.95UCL
## meno=0 1312    855   3407    3318    3487
```

```
## meno=1 1670    1050    3200     3112     3262

##         records n.max n.start events    rmean se(rmean) median 0.95LCL 0.95UCL
## meno=0     1312  1312    1312    855 3451.798  35.59308   3407    3318    3487
## meno=1     1670  1670    1670   1050 3151.695  33.19115   3200    3112    3262
```

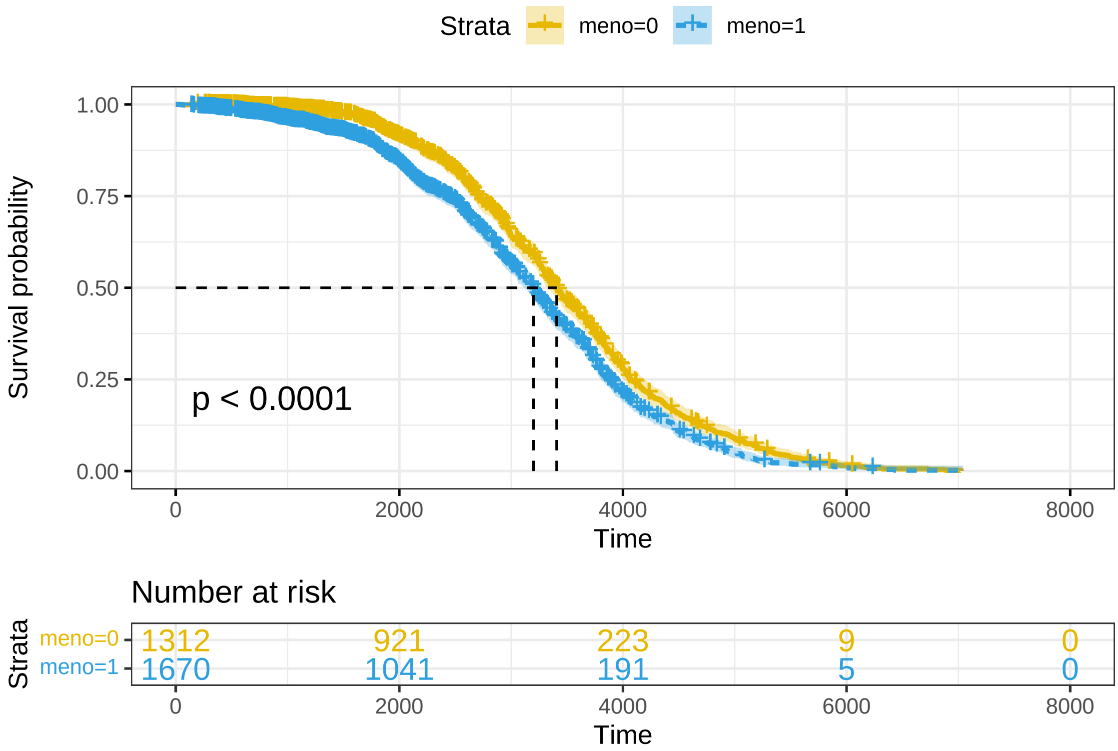



Table 2: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| meno=0 | 1312 | 1312 | 1312 | 855 | 3451.798 | 35.59308 | 3407 | 3318 | 3487 |
| meno=1 | 1670 | 1670 | 1670 | 1050 | 3151.695 | 33.19115 | 3200 | 3112 | 3262 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 2 can be seen that the median survival is 3407 for premenopausal patients, and 3200 for postmenopausal, suggesting slightly worse survival for patients that have gone through menopause. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

### 2.1.3 Hormonal Treatment

```
## Call: survfit(formula = Surv(dtime, censored) ~ hormon, data = data)
```

```
##
##              n events median 0.95LCL 0.95UCL
## hormon=0 2643   1701   3360    3307    3411
## hormon=1  339    204   2627    2340    2773
```

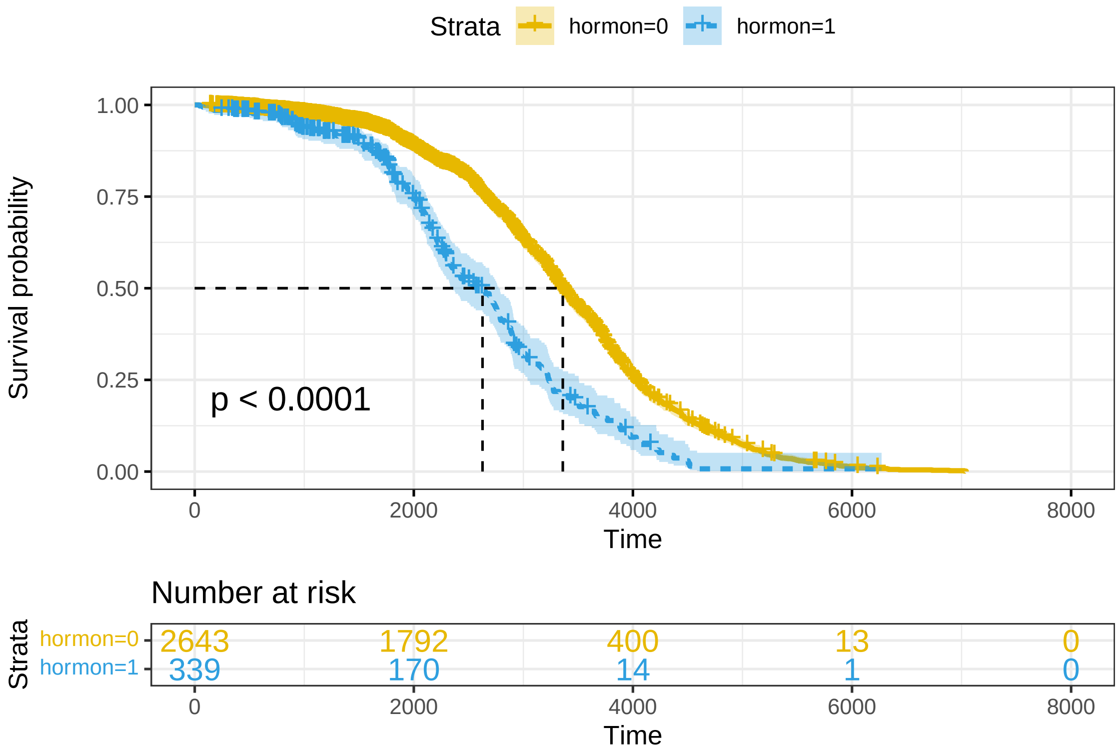



Table 3: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| hormon=0 | 2643 | 2643 | 2643 | 1701 | 3361.878 | 25.67140 | 3360 | 3307 | 3411 |
| hormon=1 | 339 | 339 | 339 | 204 | 2619.358 | 66.01373 | 2627 | 2340 | 2773 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 3 can be seen that the median survival is 3407 for premenopausal patients, and 3200 for postmenopausal, suggesting slightly worse survival for patients that have gone through menopause. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

### 2.1.4 Chemotherapy

```
## Call: survfit(formula = Surv(dtime, censored) ~ chemo, data = data,
##     type = "kaplan-meier")
```

```
##
##              n events median 0.95LCL 0.95UCL
## chemo=0 2402   1576   3269    3220    3318
## chemo=1  580    329   3445    3288    3723
```

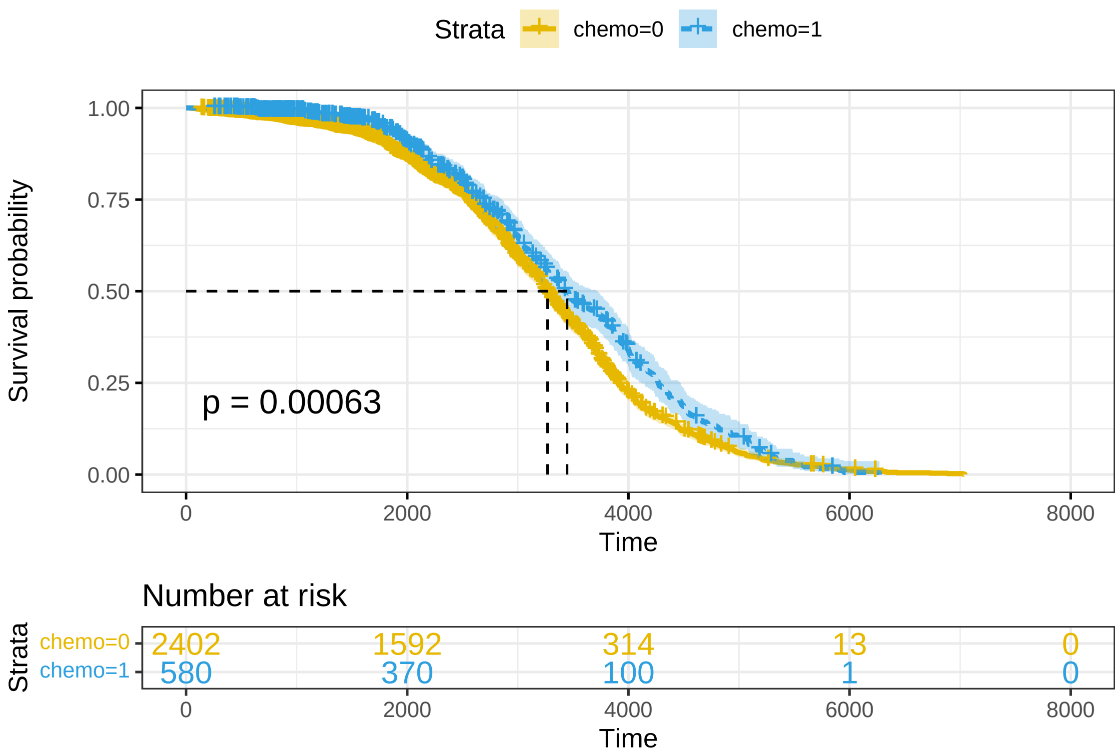



Table 4: Median survival times for each group.

| | records | n.max | n.start | events | rmean | se(rmean) | median | 0.95LCL | 0.95UCL |
|---|---|---|---|---|---|---|---|---|---|
| hormon=0 | 2643 | 2643 | 2643 | 1701 | 3361.878 | 25.67140 | 3360 | 3307 | 3411 |
| hormon=1 | 339 | 339 | 339 | 204 | 2619.358 | 66.01373 | 2627 | 2340 | 2773 |

The horizontal axis represents time in days, and the vertical axis shows the probability of surviving, or the proportion of people surviving. The lines represent survival curves of the three groups. A vertical drop in the curves indicates an event. The vertical tick mark on the curves means that a patient was censored at this time.

At time zero, the survival probability is 1.0 (or 100% of the participants are alive). At time 2200, the probability of survival is approximately 0.75 for premenopausal patients, and 0.825 for postmenopausal patients.

From Table 3 can be seen that the median survival is 3407 for premenopausal patients, and 3200 for postmenopausal, suggesting slightly worse survival for patients that have gone through menopause. However, to evaluate whether this difference is statistically significant requires a formal statistical test, a subject that is discussed in the next sections.

## 2.2 Confidence Intervals and Estimators by Levels of ...

(b) For each level obtain an appropriate estimator and confidence interval for the 3 quartiles of the survival curves. Interpret the results.

### 2.2.1 Size

```
## $quantile
##                25   50   75
## size=<=20  2591 3289 3950
## size=20-50 2640 3301 4028
## size=>50   2058 3240 4006
##
## $lower
##                25   50   75
## size=<=20  2541 3232 3887
## size=20-50 2544 3230 3934
## size=>50   1829 2918 3746
##
## $upper
##                25   50   75
## size=<=20  2660 3370 4031
## size=20-50 2725 3408 4176
## size=>50   2248 3565 4474
```

### 2.2.2 Menopause

```
## $quantile
##            25   50   75
## meno=0 2703 3407 4075
## meno=1 2416 3200 3885
##
## $lower
##            25   50   75
## meno=0 2643 3318 4005
## meno=1 2278 3112 3804
##
## $upper
##            25   50   75
## meno=0 2818 3487 4193
## meno=1 2541 3262 3965
```

### 2.2.3 Hormonal Treatment

```
## $quantile
##              25   50   75
## hormon=0 2660 3360 4027
## hormon=1 2004 2627 3234
##
## $lower
##              25   50   75
## hormon=0 2596 3307 3976
## hormon=1 1834 2340 3038
```

```
##
## $upper
##             25   50   75
## hormon=0 2712 3411 4089
## hormon=1 2116 2773 3506
```

#### 2.2.4 Chemotherapy

```
## $quantile
##           25   50   75
## chemo=0 2565 3269 3919
## chemo=1 2690 3445 4272
##
## $lower
##           25   50   75
## chemo=0 2502 3220 3860
## chemo=1 2526 3288 4112
##
## $upper
##           25   50   75
## chemo=0 2616 3318 3983
## chemo=1 2854 3723 4465
```

### 2.3 Test of Differences Between the Survival Curves

(c) Conduct a single test of differences between the survival curves. Justify your choice of test.

#### 2.3.1 Rephrase all of this text, because it's been copy-pasted!!!

Now, the questions that arises is if these two curves are statistically equivalent. For answering it, we can use the log-rank test (Mantel 1966; Peto and Peto 1972). This is the most well-known and widely used method to test the null hypothesis of no difference in survival between two or more independent groups. It is a large-sample chi-square test that is obtained by constructing a two by two contingency table at each distinct event time, and comparing the failure rates between the two groups, conditional on the number at risk in each group. The test compares the entire survival experience between groups and can be thought of as a test of whether the survival curves are identical or not. When we state that two KM curves are statistically equivalent, we mean that, based on a testing procedure that compares the two curves in some overall sense, we do not have evidence to indicate that the true (population) survival curves are different. The null hypothesis of the testing procedure is that there is no overall difference between the two (or k) survival curves. Under this, the log–rank statistic is approximately a chi-square with k−1 degree of freedom. Thus, tables of the chi-square distribution are used to determine the p-value. This test is the one with most power to test differences that fit the proportional hazards model - so works well as a set-up for subsequent Cox regression. It gives equal weight to early and late failures. An alternative test that is often used is the Peto & Peto (Peto and Peto 1972) modification of the Gehan-Wilcoxon test (Gehan 1965). This last one is a variation of the log-rank test statistic and is derived by applying different weights at the f−th failure time. This approach is most sensitive to early differences (or earlier time points) between survival. This type of weighting may be used to assess whether the effect of a treatment/marketing campaign on survival is strongest in the earlier phases of administration/contacto and tends to be less effective over time. (Marta Sestelo 2017)

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ size, data = data)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## size=<=20  1387     1048     1028     0.395     0.862
## size=20-50 1291      734      772     1.904     3.214
## size=>50    304      123      105     3.156     3.352
##
##  Chisq= 5.5  on 2 degrees of freedom, p= 0.06
```

We fail to reject the null hypothesis, hence we do not have evidence to indicate that the three survival curves are different.

### 2.3.2 Log-rank Test

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ meno, data = data,
##     rho = 0)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## meno=0 1312      855      965      12.6      25.8
## meno=1 1670     1050      940      13.0      25.8
##
##  Chisq= 25.8  on 1 degrees of freedom, p= 4e-07
```

Using the log-rank test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through menopause, and those who have not.

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ hormon, data = data,
##     rho = 0)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 2643     1701     1802      5.71       108
## hormon=1  339      204      103    100.35       108
##
##  Chisq= 108  on 1 degrees of freedom, p= <2e-16
```

Using the log-rank test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through hormonal therapy, and those who have not.

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ chemo, data = data,
```

```
##     rho = 0)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## chemo=0 2402     1576     1516      2.37      11.7
## chemo=1  580      329      389      9.24      11.7
##
##  Chisq= 11.7  on 1 degrees of freedom, p= 6e-04
```

Using the log-rank test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through chemotherapy, and those who have not.

### 2.3.3 Peto & Peto Test

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ meno, data = data,
##     rho = 1)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## meno=0 1312      426      505      12.1      33.9
## meno=1 1670      613      535      11.4      33.9
##
##  Chisq= 33.9  on 1 degrees of freedom, p= 6e-09
```

Using the Peto & Peto test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through menopause, and those who have not.

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ hormon, data = data,
##     rho = 1)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## hormon=0 2643      892    970.2      6.32       123
## hormon=1  339      148     69.6     88.17       123
##
##  Chisq= 124  on 1 degrees of freedom, p= <2e-16
```

Using the Peto & Peto test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through hormonal therapy, and those who have not.

```
## Call:
## survdiff(formula = Surv(dtime, censored) ~ chemo, data = data,
##     rho = 1)
##
##            N Observed Expected (O-E)^2/E (O-E)^2/V
## chemo=0 2402      874      837      1.69      12.6
```

```
## chemo=1  580       165       203       6.99       12.6
##
##  Chisq= 12.6  on 1 degrees of freedom, p= 4e-04
```

Using the Peto & Peto test, we reject the null hypothesis. Hence, we it is concluded that there is statistically significant difference in survival curves between patients that have gone through chemotherapy, and those who have not.

# 3 References

Marta Sestelo. 2017. *A Short Course on Survival Analysis.* https://bookdown.org/sestelo/sa_financial/comparing-survival-curves.html.