Event History Analysis in R

Day 1: Basics

Constantin Manuel Bosancianu

Doctoral School of Political Science
Central European University, Budapest
bosancianu@icloud.com

November 2, 2015

Outline

Today will be a bit more theoretical. I'll cover:

- ✓ Basic concepts in survival analysis: survivor function, hazard rate, risk set etc.
- ✓ Descriptive inspections of data: getting survivor curves for subgroups in the data
- ✓ Transition toward inferential survival analysis: tests for group differences in survival probability

As my contact with the method was not that long ago, I'll try to keep it as much as possible on an applied note.

The data

It comes from an academic team that is quite famous: C. Reinhart and K. Rogoff.

To the best of my knowledge, it's error free, unlike another data set from the same authors: `http://www.bloomberg.com/bw/articles/2013-04-18/faq-reinhart-rogoff-and-the-excel-error-that-changed-history.html`

It tracks 70 countries over 210 years of financial crises: stock market, banking, sovereign debt, inflation or currency.

The data

In the interest of a manageable task, I started tracking countries in the year 1980, after the Oil Crises.

The event that I am looking for is a sovereign debt crisis: when countries no longer have money with which to make payments on loans, and have to enter into default.

The question is: What are some of the factors that predict the event?

Foundational concepts

Things we need to cover before we start analyzing:

- ✓ Survivor function
- ✓ Hazard function
- ✓ Censoring
- ✓ Risk set

I'll present these in the context of the Reinhart–Rogoff data set.

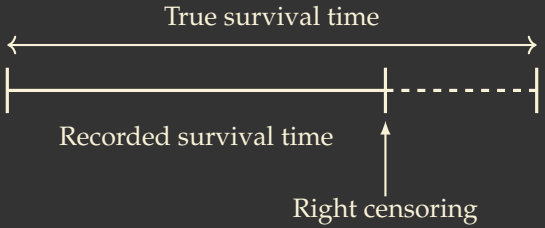
Censoring

A major source of hassle in analytic terms.

Part of the cases under analysis never get the event we're interested in studying: revolution, sovereign default, war.

These cases are said to be (right) censored. In the case of my sample, for example, 53 out of 70 countries were never in a sovereign debt crisis.

Censoring

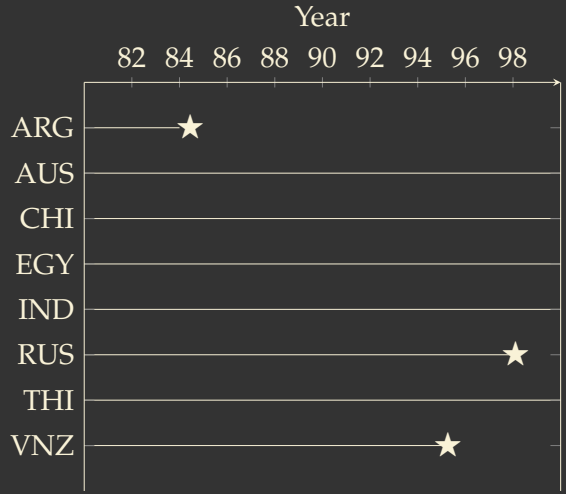


Censoring means we do not know the actual survival time.

Right censoring (hypothetical)

Study ended on the 20th week. ★ denotes event happened.

Censoring in my data



Study period: 1980–2000. ★ denotes event happened.

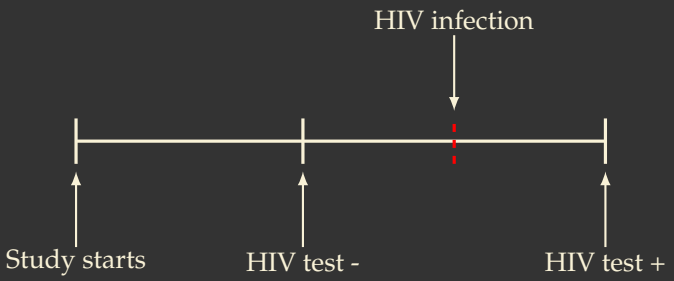
Censoring

There are other types of censoring as well, although these are less commonly encountered in the social sciences.

Left censoring: the actual time is shorter than the one observed by us.

Censoring

Interval censoring: we have a guess as to the bounds of the actual time.



Most practical applications of survival analysis deal only with instances of right censoring.

Survivor function

A survivor function, $S(t)$, gives us the probability that a random observation will survive (not experience the event) past time t.

To present this in a more intuitive way I grouped my data for 35 of the 70 countries in 3-year periods.

With a 7-period division, we have 8 time points to evaluate: 1980, 1983, 1986, 1989, 1992, 1995, 1998, and 2001.

Survivor function

Time	Crisis	Total crisis	$S(t)$
t=1980	0	0	$35/35 = 1$
t=1983	5	5	$30/35 = 0.8571$
t=1986	3	8	$27/35 = 0.7714$
t=1989	2	10	$25/35 = 0.7143$
t=1992	0	10	$25/35 = 0.7143$
t=1995	1	11	$24/35 = 0.6857$
t=1998	3	14	$21/35 = 0.6$
t=2001	2	16	$19/35 = 0.5429$

Survivor function for sovereign debt crises

"Crises" denotes number of defaults in $[t-1, t)$ interval, while "Total crises" refers to crises in the $[0, t)$ interval.

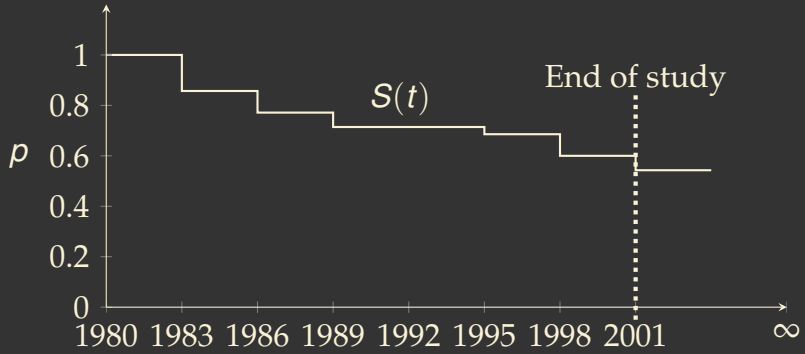
Survivor function

We can imagine that if we could let the study run to ∞, then the probability of survival would be 0.

$S(t)$ uses the information in the "Total crises" column. In this sense, it is an actual probability, since it will always range between 0 and 1.

At each time point, it uses *all* the information before it to give us the probability that the event will not happen to the remaining observations.

Survivor function



Survivor (step) function for sovereign debt crisis

$S(t)$ can also be thought as the proportion of units that survive past time t.

Hazard function

In a sense, the hazard function conveys the opposite information to the survivor function: the probability of the event *happening* per unit of time.

This is a *conditional* probability: probability of the event given that it hasn't happened prior to time t.[1]

[1] For now, because we are in the case of discrete periods, I will use the word "probability". As we move from the discrete to the continuous world, we will see that we're no longer working with probabilities, but with *rates* (which can be larger than 1).

Hazard function

Time	Crisis	Remaining obs.	$h(t)$
t=1980	0	35	$0/35 = 0$
t=1983	5	35	$5/35 = 0.1429$
t=1986	3	30	$3/30 = 0.10$
t=1989	2	27	$2/27 = 0.0741$
t=1992	0	25	$0/25 = 0$
t=1995	1	25	$1/25 = 0.04$
t=1998	3	24	$3/24 = 0.125$
t=2001	2	21	$2/21 = 0.0952$
TOTAL		16	

Hazard function for sovereign debt crises

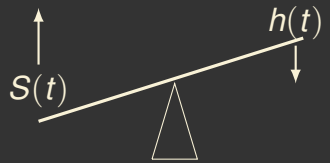
Hazard function

The "Remaining obs." column is called the *risk set* – it quantifies the number of observations which are still at risk to experience the event.

Unlike the survival probability, which could only decrease, the hazard probability (rate) can fluctuate from period to period in any direction.

> *[…] a hazard function $h(t)$ gives the instantaneous potential at time t for getting an event, like death or some disease of interest, given survival up to time t. (Kleinbaum & Klein, 2012)*

Hazard vs. survival



The two functions are connected. If the hazard rate is high, the survivor function will be low.

Hazard vs. survival

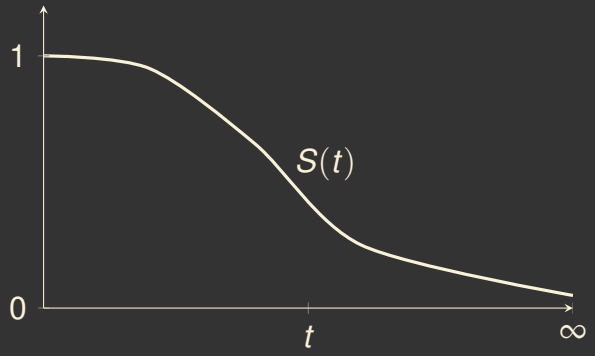
$$h(t) = \frac{f(t)}{S(t)} \tag{1}$$

$f(t)$ represents the density function of the survival times.

$$f(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq T \leq t + \Delta t)}{\Delta t} \tag{2}$$

$f(t)$ is the "*instantaneous* probability an event will occur (or a unit will fail) in the infinitesimally small area bounded by t and $t + \Delta t$."
(Box-Steffensmeier & Jones, 2004, p. 13)

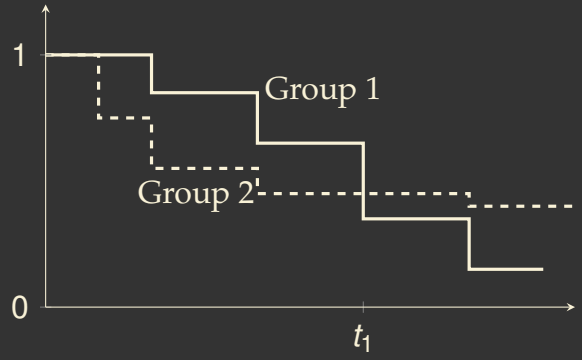
Continuous case



When time is continuous, these functions themselves are continuous. The formulas become a bit more complex, but the principles remain unchanged.

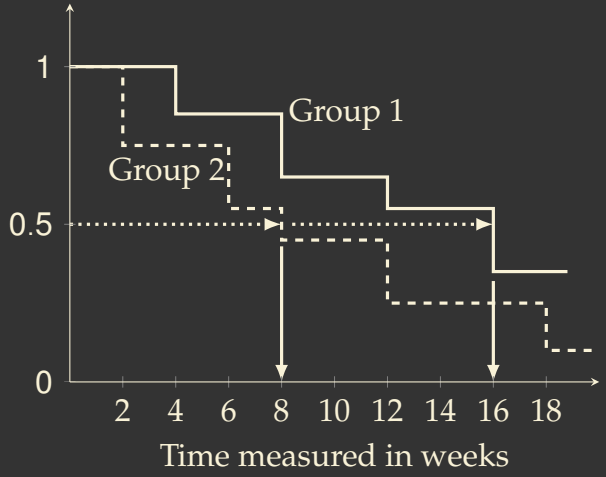
Descriptive survival analysis

Comparing survival times



There is clearly a difference between survival times before time t_1 and after this moment.

Comparing survival times



Finding out the median survival time is not difficult at all. Clearly, the first group is doing considerably better than the second one.

Examining data

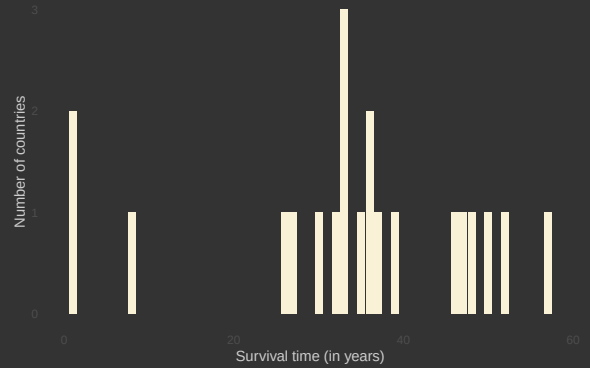
In a sense, I'm not treating the data fairly, because Argentina had multiple sovereign debt crises in the 1980s.

There are survival analysis techniques that allow you to study multiple occurrence, but they're not covered here.

In the 21 years tracked, Argentina was in a debt crisis in 3 years (two separate occasions), Russia in 2 years (one episode), and Venezuela in 4 years (1 episode).

Examining data

How does the survival duration look for the entire sample?



53 out of 70 countries are censored in my data

The `survival` package

The most capable and established R package for conducting survival analysis is... `survival`.[2]

`Zelig` will soon have the capability of running these models (e.g., Cox proportional hazard) but it's too focused on regression for our purposes.

For the entire range of descriptive and inferential procedures, we'll use `survival`.

[2] Created by Thomas Lumley, also the author of the `survey` package.

Kaplan–Meier

$$S(t_{(f)}) = \frac{n_{\textit{surviving past } t}}{n_{\textit{sample}}} \tag{3}$$

At each point in time, we can compute this quantity, exactly as we did in one of the previous tables.

In my example $S(1989) = \frac{25}{35} = 0.7143$

Kaplan–Meier

Alternatively, it can also be computed as a product for individual period probabilities (these are essentially $1 - h(t)$).

$$S(t) = \prod_{i=1}^{t} \frac{n_{\textit{surviving past } i}}{n_{\textit{surviving past } i-1}} \tag{4}$$

In my example $S(1989) = \frac{30}{35} \times \frac{27}{30} \times \frac{25}{27} = \frac{25}{35} = 0.7143$

These two ways of computing the KM estimate will produce the same result.

Survival curves

```r
library(survival)
# First we have to define a "survival" object, which is then
# used by the package's functions.
Surv(df_fin$time, df_fin$event, type = "right")
```

```
 [1] 19+ 12   2  19+  2   6  19+ 19+ 19+ 19   1  19+  2  19+ 19+ 19+ 17  19+ 19+
[20] 19+  2   4   5  19+  8  19+  5  19+ 19+ 18  16  19+ 19+ 15  19+
```

```r
# The "type" argument asks what kind of censoring we have in
# our data, with 6 options. "right" is the default, but I
# added it here nonetheless. Other common options are "left"
# or "interval".
```

Right censored observations are marked with a + in the output.

Estimating KM curves

```
ObjSurv1 <- survfit(Surv(time, event) ~ 1, data = df_fin)
# ~1 means that the entire sample is used for 1 KM curve.
```

The `survfit()` function will do the actual estimation of the Kaplan-Meier survival curves.

```
round(summary(ObjSurv1)$surv, digits = 2)
```

```
 [1] 0.97 0.86 0.83 0.77 0.74 0.71 0.69 0.66 0.63 0.60 0.57 0.54
```

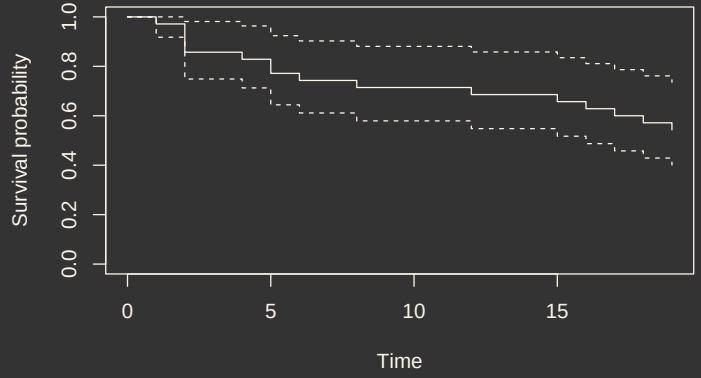
Lower and **upper** confidence bounds for these estimates are also produced and easily accessible.

```
round(summary(ObjSurv1)$upper, digits = 2)
```

```
 [1] 1.00 0.98 0.96 0.92 0.90 0.88 0.86 0.83 0.81 0.79 0.76 0.74
```

Estimating KM curves

```
plot(ObjSurv1, xlab = "Time", ylab = "Survival probability")
```



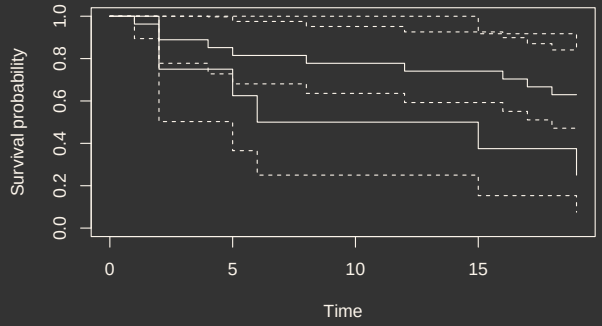
Estimating KM curves

```
rm(ObjSurv1)
ObjSurv2 <- survfit(Surv(time, event) ~ SouthA, data = df_fin)
# Compute separate curves for two groups. The formula can
# accommodate numerous situations, involving multiple
# group membership indicators.
```

The same command for plotting can be used in this instance as well.

```
plot(ObjSurv2, xlab = "Time", ylab = "Survival probability",
     conf.int = TRUE)
# The defaults with more than two groups is not to plot CIs.
# Here, though, I've asked it to plot them. When there is
# only one group, however, CIs are plotted by default.
```

Estimating KM curves



Should we be content with these bland graphics? I think we shouldn't.

Estimating KM curves

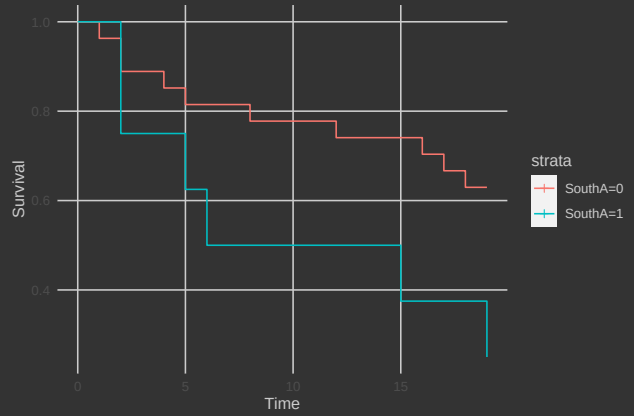
```
ObjSurvMod2 <- createSurvivalFrame(ObjSurv2)
rm(ObjSurv2)
qplot_survival(ObjSurvMod2) + xlab("Time") + ylab("Survival")
```

The functions `createSurvivalFrame()` and `qplot_survival()` were created by Ramon Saccilotto.

You can modify these yourselves, for better control of the graph (e.g. line thickness, proper legend), or just extract the quantities manually and use them in `ggplot2`.[3]

[3] Watch out for the `ggfortify` package, which will probably be launched soon on CRAN.

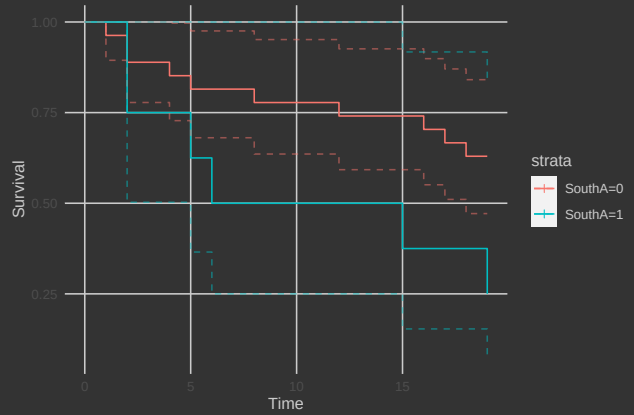
Estimating KM curves



By default, the functions plot the KM survival estimates without confidence intervals, as in the standard `plot()` function.

Estimating KM curves

```
qplot_survival(ObjSurvMod2, TRUE) + xlab("Time") +
    ylab("Survival")
```



Log-rank test

How do we truly know that the two curves are different from each other, when taking into account the uncertainty around them?

The `log-rank test` can be used for this.

It is essentially a χ^2 test, which relies on observed vs. expected[4] cell counts in the different outcome categories.

[4] What would be expected if the null hypothesis of no effect would be true.

Log-rank: theoretical

Assume we have a particular point in time, which we denote as t, and the two groups we're trying to compare, which I'll label as A and B.

The expected cell count for group A at time t is, in this situation:

$$e_A = \frac{n_{risk\ A}}{n_{risk\ A} + n_{risk\ B}} \times (n_{fail\ A} + n_{fail\ B}) \tag{5}$$

The first term in the equation is the proportion of elements in the risk set, which gets multiplied with the total number of events.

Log-rank: theoretical

Identically, the expected cell count at time t for group B is:

$$e_B = \frac{n_{risk\ B}}{n_{risk\ A} + n_{risk\ B}} \times (n_{fail\ A} + n_{fail\ B}) \tag{6}$$

For each group, we can compute the sum of the difference between observed and expected cell counts at each time point. For example, for A,

$$O_A - E_A = \sum_{1}^{n_{fail\ A}} (n_{fail\ A} - e_A) \tag{7}$$

Log-rank: theoretical

Finally, the log-rank test statistic value for any one of the two groups is

$$Log - rank = \frac{(O_A - E_A)^2}{Var(O_A - E_A)} \tag{8}$$

$$Var(O_A - E_A) = \frac{n_{risk\ A} \times n_{risk\ B} \times (n_{fail\ A} + n_{fail\ B}) \times (n_{risk\ A} + n_{risk\ B} - n_{fail\ A} - n_{fail\ B})}{(n_{risk\ A} + n_{risk\ B})^2 (n_{risk\ A} + n_{risk\ B} - 1)} \tag{9}$$

Log-rank: practical

This will have a χ^2 distribution with 1 degree of freedom (in the situation when we're comparing only 2 groups).[5]

I tried doing all these calculations by hand, but I must have made a mistake somewhere because I got a slightly different result at the decimal points.

So I switched to a pre-worked example, on a different data set.

[5] When comparing N groups, it will be N-1 degrees of freedom.

Log-rank: practical

EXAMPLE

The data: remission times (weeks) for two groups of leukemia patients

Group 1 ($n = 21$) treatment	Group 2 ($n = 21$) placebo
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+,	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

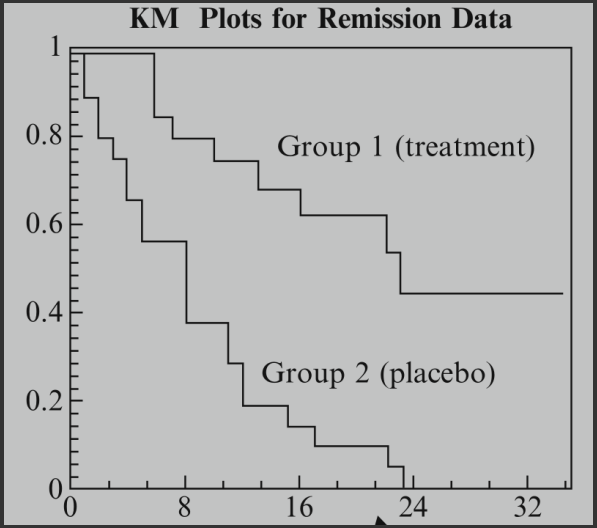
Note: + denotes censored

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

Descriptive statistics:

$\bar{T}_1(\text{ignoring} + \text{'s}) = 17.1,\ \bar{T}_2 = 8.6$

$\bar{h}_1 = .025,\ \bar{h}_2 = .115,\ \dfrac{\bar{h}_2}{\bar{h}_1} = 4.6$



Log-rank: practical

EXAMPLE

Expanded Table (Remission Data)

		# failures		# in risk set		# expected		Observed-expected	
f	$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	e_{1f}	e_{2f}	$m_{1f}-e_{1f}$	$m_{2f}-e_{2f}$
1	1	0	2	21	21	$(21/42) \times 2$	$(21/42) \times 2$	−1.00	1.00
2	2	0	2	21	19	$(21/40) \times 2$	$(19/40) \times 2$	−1.05	1.05
3	3	0	1	21	17	$(21/38) \times 1$	$(17/38) \times 1$	−0.55	0.55
4	4	0	2	21	16	$(21/37) \times 2$	$(16/37) \times 2$	−1.14	1.14
5	5	0	2	21	14	$(21/35) \times 2$	$(14/35) \times 2$	−1.20	1.20
6	6	3	0	21	12	$(21/33) \times 3$	$(12/33) \times 3$	1.09	−1.09
7	7	1	0	17	12	$(17/29) \times 1$	$(12/29) \times 1$	0.41	−0.41
8	8	0	4	16	12	$(16/28) \times 4$	$(12/28) \times 4$	−2.29	2.29
9	10	1	0	15	8	$(15/23) \times 1$	$(8/23) \times 1$	0.35	−0.35
10	11	0	2	13	8	$(13/21) \times 2$	$(8/21) \times 2$	−1.24	1.24
11	12	0	2	12	6	$(12/18) \times 2$	$(6/18) \times 2$	−1.33	1.33
12	13	1	0	12	4	$(12/16) \times 1$	$(4/16) \times 1$	0.25	−0.25
13	15	0	1	11	4	$(11/15) \times 1$	$(4/15) \times 1$	−0.73	0.73
14	16	1	0	11	3	$(11/14) \times 1$	$(3/14) \times 1$	0.21	−0.21
15	17	0	1	10	3	$(10/13) \times 1$	$(3/13) \times 1$	−0.77	0.77
16	22	1	1	7	2	$(7/9) \times 2$	$(2/9) \times 2$	−0.56	0.56
17	23	1	1	6	1	$(6/7) \times 2$	$(1/7) \times 2$	−0.71	0.71
Totals		9	21			19.26	10.74	−10.26	−10.26

From Kleinbaum and Klein (2012, p. 69).

Log-rank: practical

EXAMPLE

$$O_2 - E_2 = 10.26$$
$$\text{Var}(O_2 - E_2) = 6.2685$$

$$\text{Log - rank statistic} = \frac{(O_2 - E_2)^2}{\widehat{\text{Var}}(O_2 - E_2)}$$
$$= \frac{(10.26)^2}{6.2685} = 16.793$$

From Kleinbaum and Klein (2012, p. 71).

But how would this test work in my case?

Log-rank: practical

```
survdiff(Surv(time, event) ~ SouthA, data = df_fin)

Call:
survdiff(formula = Surv(time, event) ~ SouthA, data = df_fin)

            N Observed Expected (O-E)^2/E (O-E)^2/V
SouthA=0 27       10    12.99     0.689      3.82
SouthA=1  8        6     3.01     2.973      3.82

 Chisq= 3.8  on 1 degrees of freedom, p= 0.05
```

It's technically not statistically significant. It's anyone's guess, but my personal opinion is that with a slightly larger sample size for one of the groups, we could have found a significant difference.

Log-rank: theoretical

For more groups, the formulas get considerably more complicated, but the software cares little about this.

However, there are numerous variants of the log-rank test, which differ in how they weight the time points.

As you could see from the formula, Kaplan–Meier weights each point equally when computing the final test statistic.

Log-rank: theoretical

Test statistic	$w(t)$
Log-rank	1
Wilcoxon	n_{risk}
Tarone-Ware	$\sqrt{n_{risk}}$
Peto	$\widetilde{s}(t)$
Flemington-Harrington[6]	$\widehat{S}(t)^p \times [1 - \widehat{S}(t)]^q$

There might be valid situations when you want to assign more influence to earlier time points, e.g. if you believe that a drug would have a more dramatic short-term effect than a long-term one.

[6] Offers maximum flexibility because the user supplies the values for p and q.

Log-rank: theoretical

Only the Peto variant is implemented in the `survival` package.

```
survdiff(Surv(time, event) ~ SouthA, data = df_fin, rho = 1)

Call:
survdiff(formula = Surv(time, event) ~ SouthA, data = df_fin,
    rho = 1)

           N Observed Expected (O-E)^2/E (O-E)^2/V
SouthA=0 27     7.97    10.29     0.521      3.34
SouthA=1  8     4.80     2.49     2.155      3.34

 Chisq= 3.3  on 1 degrees of freedom, p= 0.07
```

The Wilcoxon test is available in the `stats` package, the Flemington–Harrington in `FHtest`, and the Tarone–Ware in `EnvStats`.

Conclusion

We had to go through these terms because they are quite unique to the entire modeling approach.

So far we've mainly been concerned with the descriptive part of survival analysis. We also made small steps into the inferential area.

Tomorrow we go full-inferential with the Cox proportional-hazards model.

Thank you for the kind attention!

References

Box-Steffensmeier, J. M., & Jones, B. S. (2004). *Event History Modeling: A Guide for Social Scientists*. New York, NY: Cambridge University Press.

Kleinbaum, D. G., & Klein, M. (2012). *Survival Analysis: A Self-Learning Text*. New York, NY: Springer.