# 2013 ECPR Summer School in Methods and Techniques
# University of Ljubljana
# Course Description Form *(2-week courses)*

**Course title**

## Event History and Survival Analysis

**Instructor details**

First name, last name: **Janez Stare**
Department/Unit: Institute of Biostatistics and Medical Informatics
Institution: Faculty of Medicine, University of Ljubljana
Full postal address for ECPR correspondence: Vrazov trg 2, SI-1000 Ljubljana, Slovenia
Phone: +386-1-543-7772
Fax: +386-1-543-7771
E-mail : janez.stare@mf.uni-lj.si

**Short Bio**

***Janez Stare*** graduated from the Faculty of Mathematics, University of Ljubljana, then held a Master Degree and Ph.D. in Biostatistics from the Faculty of Medicine, University of Ljubljana. He is currently full Professor of Biostatistics at the Faculty of Medicine, Ljubljana, and Head of the Institute of Biostatistics and Medical Informatics, Faculty of Medicine, Ljubljana. His research interests are explained variation in survival analysis, linear models in survival analysis, predictive ability of regression models in survival analysis, frailties, random effects in survival models, relative survival.

**Short outline**

In event history analysis (and survival analysis, which is the name used mostly in bio sciences, where the methods were first applied) we are interested in time intervals between successive state transitions or events. Typical examples are: duration of unemployment, duration of marriage, recidivism in criminology, duration of political systems, time from diagnosis to death, and so on. The most distinctive feature of time to event data is that the event is often not observed at the time of analysis. Applying standard statistical methods to such data leads to severe bias or loss of information. Special methods are therefore needed to extract information which we are used to get using standard methods (formally this means estimating the distribution function and incorporate predictive variables into such estimation). Further complications arise when covariates change in time, when times between recurring events are correlated, when there are competing risks, or when effects change in time.
In this course we will thoroughly study a situation when there is only one event per subject, but will also review the extensions to a sufficient degree for students to be able to continue their work in the area. Roughly half of the time will be devoted to practical exercises, for which the package R will be used. Familiarity with R is not assumed, but students will receive a short introductory material to the package before the summer school begins.
While it is impossible to avoid all formulas, I will focus on the concepts in my lectures, but will support the lectures with more rigorous written material.

**Long outline**

Say we are interested in how long people keep their first job. We start our study at some point in time and include a sample of people that obtained their first job after the study started. After some time, say a number of years, the study stops and we want to analyse data. Some people have lost their job in the meantime, some have changed it, but some are still working and we do not have complete data on their time at job. Further, if somebody has stopped working because of his inability to work (accident, death), we also don't know what his event time would have been had he still been able to work. When the event is not observed at the time of analysis we say that censoring has occurred. With such data we cannot even calculate the mean, or draw a histogram, let alone use linear regression or similar methods. Special methods are therefore needed, and most of them use the hazard (or intensity) function. Since this is defined via the conditional probability of event occurring in some time interval given it has not occurred before, the hazard can be estimated even in the presence of censored data. As we shall see, knowing the hazard function is equivalent to knowing the distribution function, which is the main goal of any analysis. In Survival analysis, and consequently in Event History Analysis, it has become customary to talk about the survival function, which is simply one minus the distribution function.

In the first part of the course we will deal with estimating the survival and the hazard function (parametrically and non-parametrically), some measures of central tendency commonly used, and learn how to write down the likelihood function in the presence of censoring.

To continue our example, we might then be interested if there are any differences in keeping the job between men and women, between people with different levels of education, among different working environments and so on, in short, do some covariates influence the time a person stays in her/his first job.

We will therefore learn about tests for comparing survival functions and discuss two most commonly used parametric models for inclusion of covariates.

The focus in the second week will be on the Cox proportional hazards model which is by far most often used in the analysis of time to event data. While the model is very simple, it is also very flexible, and an experienced statistician can make it fit to almost any data. We will learn the basics about the estimating procedure, interpretation, testing, checking the modelling assumptions and relaxing them, and some extensions like the stratified model and frailties.

For now we have not distinguished between a person losing his job, and a person changing his job. We also have not considered studying several spells for one person (one person can change jobs several times in the study period). These problems fall under the headings of competing risks and recurring events. The last two days will be devoted to these, as well as time varying effects, and, time permitting, some other models.

Here is a list of topics:

Univariate event history analysis

1. Censoring
2. Survival function
3. Hazard and cumulative hazard function
4. Mean time, mean residual time, median time
5. Likelihood function for censored time to event data
6. Parametric models for the survival function (exponential, Weibull)
7. Non parametric estimation of the survival function (Kaplan Meier and Nelson Aalen estimators)
8. Variance of the survival function, confidence intervals
9. Comparison of survival functions

Regression models for time to event data

10. Parametric models (exponential, Weibull)
11. Cox model (proportional hazards model):
    a. Estimation (partial likelihood)
    b. Interpretation
    c. Testing the null hypothesis (Wald, score, and likelihood ratio test)
    d. Some model fitting techniques
    e. Categorical variables in the model
    f. Relaxing the linearity assumption for continuous variables using splines
    g. Checking the model assumptions

h. Goodness-of-fit and explained variation
i. Stratified model
j. Time varying covariates
k. Frailties
l. Time varying effects
m. Competing risks
n. Recurring events

12. Multistate models

**Day-to-day schedule**

- Week 1

| | **Topic(s)** | **Details [NB : incl. timing of lecture v/s lab or fieldwork etc. hours]** |
|---|---|---|
| Day 1 | Monday Mix (90 min general introduction to the topic) | 90' lecture<br>- topics of course<br>- course goals<br>- overview of course schedule |
| Day 2 | Event History and Social Science<br>Event history data structures<br>Basic definitions | Examples<br>Censoring<br>Survival function<br>Hazard and cumulative hazard function<br><br>2 hours lecture, 1 hour lab |
| Day 3 | Parametric and nonparametric descriptive methods | Mean survival time, mean residual time, median time<br>Exponential and Weibull distribution<br>Kaplan-Meier estimator<br>Life tables<br>2 hours lecture, 1 hour lab |
| Day 4 | Comparison of survival functions<br>Parametric regression models for single-spell duration data<br>Methods to check parametric assumptions | Log rank test<br><br>Exponential and Weibull regression model<br><br><br>90' lecture, 90' lab |
| Day 5 | Introduction to Cox model | Estimation (partial likelihood)<br>Interpretation<br><br>90' lecture, 90' lab |

- Week 2

| | | |
|---|---|---|
| Day 6 | Participants present their own data<br>Cox model (cont.) | <br><br>Some model fitting techniques<br>Categorical variables in the model<br>90' lecture (together with presentations), 90' lab |
| Day 7 | Cox model (cont.) | Relaxing the linearity assumption for continuous variables using splines<br>Checking the model assumptions<br>Stratified model<br><br>90' lecture, 90' lab |

| | | |
|---|---|---|
| Day 8 | Cox model (cont.) | Time varying covariates<br>Frailties<br>Time varying effects<br><br>90’ lecture, 90’ lab |
| Day 9 | Competing risks<br>Models for multiple events | Cox model for competing risks, repeated events and multistate models<br><br>90’ lecture, 90’ lab |
| Day 10 | Models for multiple events (labs only) | Cox model for repeated events and multistate models<br><br>3 hours lab |

(day 11, Saturday, 9:00-12:00: Exam)

**Day-to-day reading list**

I have chosen two books that could be used as a supplementary reading:

1. Janet M. Box-Steffensmeier, Bradford S. Jones. Event History Modeling. A Guide for Social Scientists]. Cambridge University Press 2004.
2. Habs-Peter Blossfeld, Götz Rohwer. Techniques of Event History Analysis. Lawrence Erlbaum Associates, London 2002.

The first is less, the second more technical. I am listing books that have Event History in the title, but there are books under the name Survival Analysis which are better. David Collett’s is one, Hosmer&Lemshov’s is another example (see point 9. below). Most of the material covered in the course will be contained in my own hand-outs, which the participants will be able to download from the web one month before the course begins.

- Week 1

| | **Readings (please read at least the compulsory reading for the scheduled day)** |
|---|---|
| Day 1 | Box-Steffensmeier 1,2 |
| Day 2 | Blossfeld 2 |
| Day 3 | Blossfeld 3.1, 3.2 |
| Day 4 | Box-Steffensmeier 1,2, Blossfeld 3.3 |
| Day 5 | Box-Steffensmeier 4 |

- Week 2

| | |
|---|---|
| Day 6 | Box-Steffensmeier 6 |
| Day 7 | Box-Steffensmeier 7, 9, Blossfeld 10.1 |
| Day 8 | Box-Steffensmeier 10 |
| Day 9 | Box-Steffensmeier 10 |
| Day 10 | |

**Requested prior knowledge**

Participants should have some working knowledge of linear regression models and be familiar with the basics of inferential statistics. If not, a crash course in inferential statistics is highly recommended. As for mathematics, it is understood that the participants will not have much mathematical skills, but for the exceptions the written material will contain more rigorous treatment of the subject. Still, I would suggest that the participants clear the dust from the mathematics lying buried in their memory, preferably with the notion of the integral included.

**Software used**

R will be used. R is freely available at http://cran.r-project.org/. It is not essential to have any experience with the package, but some familiarity is welcome, and a crash course, offered at the summer school just prior to this course, is recommended.

**Literature**


a) Collett D. Modelling Survival Data in Medical Research. Chapman and Hall/CRC; 2 edition (March 30, 2003)
b) Hosmer D.W., Lemeshow S., May S. Applied Survival Analysis: Regression Modeling of Time to Event Data (Wiley Series in Probability and Statistics). Wiley-Interscience; 2 edition (March 7, 2008)


**Lecture room requirement**

No need for a different room for lab exercises. See point 8. **Important: We accept participants to bring their own laptops (notebooks, . . .) with R installed**. In this way we do not depend on the availability of rooms with computers are much more flexible in carrying out the course.