Survival Analysis in R

Mehmet Tevfik DORAK, MD PhD

School of Life Sciences, Pharmacy & Chemistry
Kingston University London

Kingston University London

Outline

Survival Analysis

(Kaplan-Meier Curve; Univariable-Log-rank test; Multivariable-Cox PH model)

Statistical Power in Survival Analysis

Survival Analysis

Background

REVIEW ARTICLE

Survival Analysis

Part 15 of a Series on Evaluation of Scientific Publications

Isabella Zwiener, Maria Blettner, Gerhard Hommel

SUMMARY

Background: Survival times are often used to compare treatments. Survival data are a special type of data, and therefore have to be analyzed with special methods.

Methods: We illustrate special techniques for analyzing survival times by applying them to a publication on the treatment of patients with brain tumors. The present article is based on textbooks of statistics, a selective review of the literature, and the authors' own experience.

Results: Survival times are analyzed with the Kaplan-Meier method, which yields two measures of interest: survival rates and the median survival time. The log-rank test is used to compare survival times across treatment groups. Cox regression is used in multivariable models. The hazard ratio, a descriptive measure for differences in survival times, is explained.

Conclusion: If survival times are analyzed without the use of special techniques, or if the underlying assumptions are not taken into account, faulty interpretation may result. Readers of scientific publications should know these pitfalls and be able to judge for themselves whether the chosen analytical method is correct.

▶ Cite this as:
Zwiener I, Blettner M, Hommel G: Survival analysis—part 15 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2011; 108(10): 163–9.
DOI: 10.3238/arztebl.2011.0163

In many areas of medicine, the primary target parameter is the time until an event occurs. Examples include the time from diagnosis of lung cancer to death, the time from fitting dentures to first repair, and the time from the beginning of treatment for urinary incontinence until successful treatment outcome. An "event" may be either success (cure) or failure (death). It is important that both the beginning of the period of time and the time of the event are clearly defined. The time between the two is generally called survival time, even when the event which ends it is not death.

Almost all specialized medical publications include articles in which survival analysis techniques are used. A recent example of this is a trial in patients with brain tumors. Von Hoff et al. (1) investigated 280 children and young people with medulloblastoma in the two-arm, randomized trial HIT '91 (HIT = *Hirntumor* [German for brain tumor]). Patients in arm 1 received chemotherapy before and after radiotherapy ("sandwich" chemotherapy), while patients in arm 2 first received radiotherapy and then chemotherapy (maintenance chemotherapy). The trial investigated whether one of the two types of treatment led to longer patient survival times.

In order to interpret the results and value of such publications correctly, readers should be familiar with the methods used to analyze survival times. This article provides a step-by-step introduction to survival analysis techniques based on the HIT '91 trial and enables readers to understand and interpret them themselves.

[**Supplement** An Overview of Study Design and Statistical Considerations]

CHEST

A Practical Overview and Reporting Strategies for Statistical Analysis of Survival Studies

Tanujit Dey, PhD; Anish Mukherjee, MS; and Sounak Chakraborty, PhD

Kingston University London

Survival Analysis

Background

https://statsandr.com/blog/what-is-survival-analysis

What is survival analysis? Examples by hand and in R

Antoine Soetewey 2022-12-22 26 minute read

- Introduction
- What is survival analysis?
- Why do we need special methods for survival analysis?
- Common functions in survival analysis
    - Survival function
    - Cumulative hazard function
    - Hazard function
- Estimation
    - By hand
    - In R
- Hypothesis testing
    - Log-rank test
        - By hand
        - In R
- To go further
- References

Survival Analysis

Survival analysis uses the survival function.

Survival function is a time to failure function that gives the probability that an individual survives past a time point. This probability depends on whether a subject is censored (due to dropout/withdrawal or loss to follow-up) or had the adverse outcome or event (like relapse, metastasis, death). All remaining subjects at the end of the follow-up period who have not yet had the event are censored.

The **Kaplan-Meier method** is used to estimate the survival function and to illustrate the survival curve/s (in comparative groups, if any).

The **log-rank test** checks whether the survival curves are statistically significantly different from one another. It is a univariate test and cannot be adjusted for potential confounders. This analysis yields a **hazard ratio** as the "*effect size*" and its confidence interval.

The **Cox proportional hazards model** is used for multivariable modelling to identify risk factors and to obtain an "**adjusted**" **hazard ratio**.

Survival Analysis

What is survival data?

Time-to-event data that consist of a distinct start time and end time.

Examples from cancer

- Time from surgery to death
- Time from start of treatment to progression
- Time from response to recurrence

Examples from other fields

Time-to-event data are common in many fields including, but not limited to

- Time from HIV infection to development of AIDS
- Time to heart attack
- Time to onset of substance abuse
- Time to initiation of sexual activity
- Time to machine malfunction

Aliases for survival analysis

Because survival analysis is common in many other fields, it also goes by other names

- Reliability analysis
- Duration analysis
- Event history analysis
- Time-to-event analysis

Survival Analysis

Box 1. Examples of when to use survival data

A. Blood pressure

In a trial comparing blood pressure reductions caused by two drugs, it is assumed that the changes in blood pressure of the subjects caused by the different drugs are normally distributed (this is 'the sample' from a population). Calculations to determine whether the differences between the interventions are statistically different (the probability of the difference having occurred by chance) are based on statistical methods which can be applied to continuous variables.

The mean of the blood pressure differences are calculated, and the variance (and standard deviation) or range of blood pressure changes can also be deduced. Using these measures a statistical test such as a Student's t-test or analysis of variance (ANOVA)[3] can be carried out to determine the probability of the differences observed having occurred by chance. Conventionally it is accepted that if this probability is less than 0.05 ($p<0.05$) then the differences are statistically significant and the null hypothesis can be rejected – the treatments are not the same.

B. Aspirin and mortality

In a trial designed to observe whether aspirin reduces mortality, patients who had sustained a myocardial infarction are randomised to aspirin or to placebo. After several years have elapsed the number who die in each treatment group is analysed and compared. The question to be answered here is whether there is a relationship between aspirin use and the risk of a patient dying, or whether the aspirin does not affect mortality (the null hypothesis). One way to determine this is using tests on categorical data (either the patient dies or does not).

In this example the Chi-squared test of association[3] can be used to determine whether to reject the null hypothesis of no association. The results show that the proportion of patients given aspirin who die is less than the proportion that dies when given placebo. If the Chi-squared test gives a p-value of <0.05, then it is unlikely that this result has occurred by chance.

C. Statins and cardiovascular events

In a trial examining whether a statin prevents a cardiovascular event in patients who have been admitted to hospital with unstable angina, patients are randomised to the statin or to placebo on admission. In this instance the focus of the study is examining the time between randomisation and a subsequent event. It is unlikely that these times are normally distributed. In this type of trial it is better, and possibly more ethical, if the study does not wait until events have occurred in all subjects. Also, some patients may leave the study early and become lost to follow-up, so that only the only information available regarding these patients will be that they were still without a further event at the last follow-up.

In this instance, it is preferential to analyse the data using a Kaplan–Meier analysis.[3] The basic idea is that the trial is split up into distinct time intervals. In each time interval the probability of 'surviving' that time interval without an event is calculated and these probabilities are multiplied to give the probability of 'survival' up to a given time point. Survival probability curves are plotted for those given the statin and those given placebo and the hazard ratio between these survival curves is calculated. The p-value for this hazard ratio is <0.05, so it is unlikely that this difference in time to an event has occurred by chance and, therefore, it is decided that statins do prevent and delay cardiovascular morbidity after admission for unstable angina.

NB In Example B it can be seen that if time-to-event data were available this could have been used as in Example C. Nowadays most studies of this nature are conducted this way. Analysing data in this way provides the added benefit of collecting information that allows assessment not just of whether a treatment prevents events but also by how much the time an event is delayed by treatment.

Kingston University London

Survival Analysis

Time-to-event/censoring

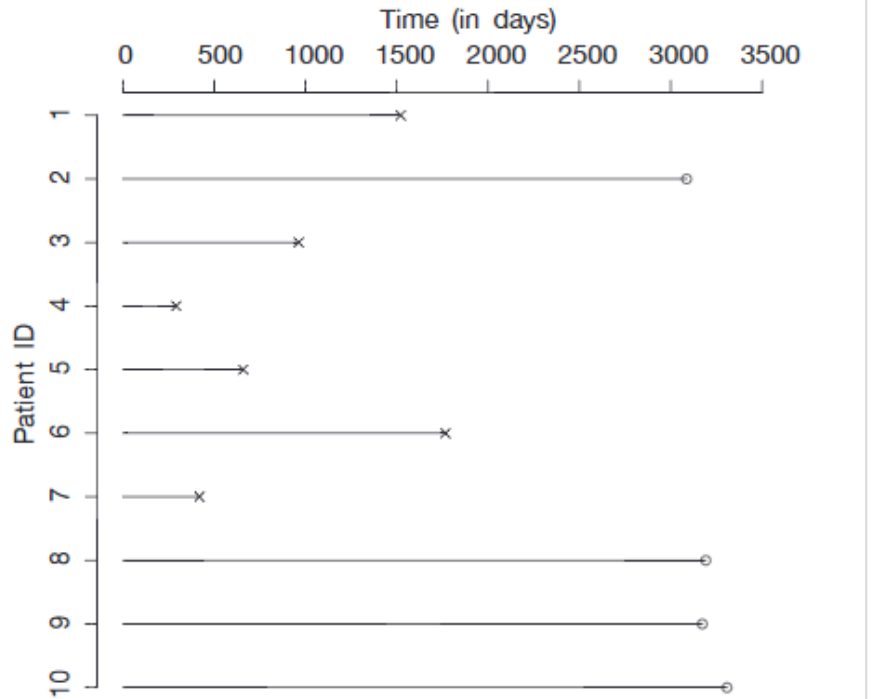


Figure 1 Survival times (×) and censoring times (○) in days.

Survival curves

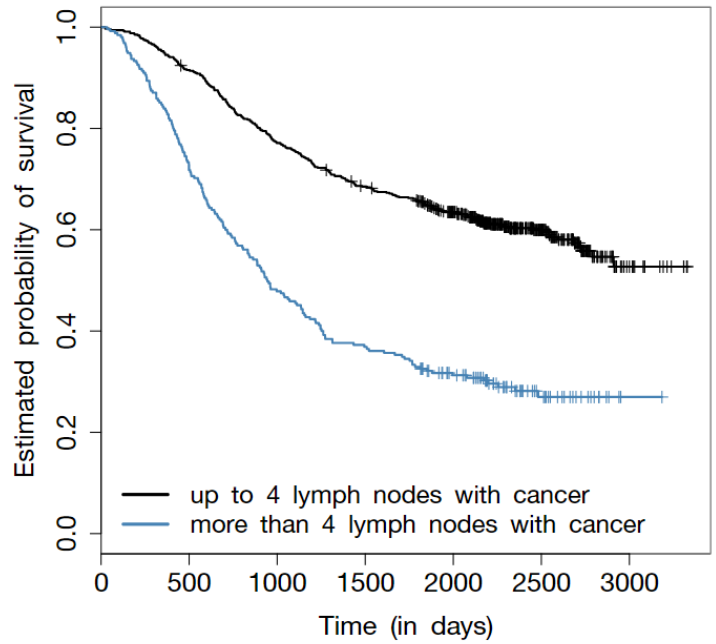


Figure 3 Kaplan–Meier curves for colon cancer patients with up to four lymph nodes with detectable cancer (black) and more than four lymph nodes with cancer (blue). + indicates censoring.

A censored subject is not deemed to have failed, but removed from the count that makes up the denominator in the estimation of survival function / hazard.

Survival analysis

Christiana Kartsonaki

Abstract
Survival analysis is the analysis of data involving times to some event of interest. The distinguishing features of survival, or time-to-event, data and the objectives of survival analysis are described. Some fundamental concepts of survival analysis are introduced and commonly used methods of analysis are described.

Keywords Cox proportional hazards model; failure times; hazard; Kaplan–Meier curve; survival data; time-to-event data

Kingston University London

Survival Analysis

Time-to-event/censoring

Survival Analysis in R

Survival Analysis with R

2017-09-25

by Joseph Rickert

With roots dating back to at least 1662 when John Graunt, a London merchant, published an extensive set of inferences based on mortality records, survival analysis is one of the oldest subfields of Statistics [1]. Basic life-table methods, including techniques for dealing with censored data, were discovered before 1700 [2], and in the early eighteenth century, the old masters - de Moivre working on annuities, and Daniel Bernoulli studying competing risks for the analysis of smallpox inoculation - developed the modern foundations of the field [2]. Today, survival analysis models are important in Engineering, Insurance, Marketing, Medicine, and many more application areas. So, it is not surprising that R should be rich in survival analysis functions. CRAN's Survival Analysis Task View, a curated list of the best relevant R survival analysis packages and functions, is indeed formidable. We all owe a great deal of gratitude to Arthur Allignol and Aurielien Latouche, the task view maintainers.

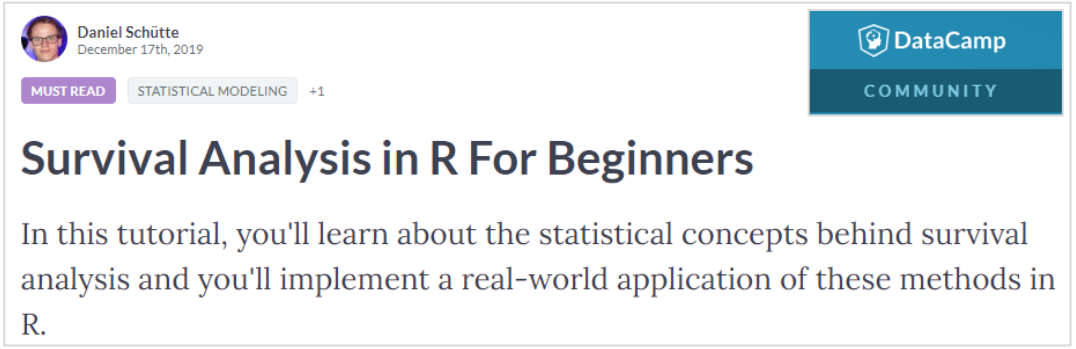
R Views

An R community blog edited by RStudio

Boston, MA

Kingston University London

Survival Analysis in R



Survival analysis involves:

- Time-to-event data and who has reached the event during follow-up (others are right-censored)

- Kaplan-Meier plot (survival curves)

- Log-rank statistics for the curve (no adjustment for covariates/potential confounders)

- Cox proportional hazard modeling (uses the hazard function with adjustments). It assumes that the hazards of the two groups being compared are constant over time.

Necessary variables:

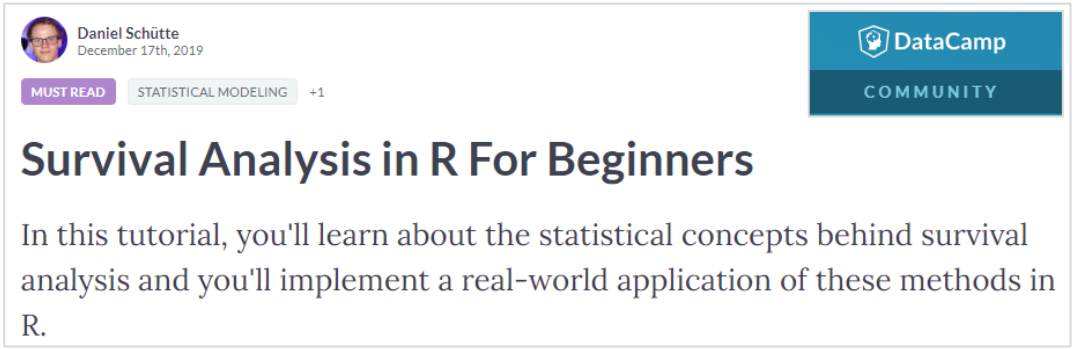
- time (time taken to reach the event, or time in follow-up until attrition)

- event (relapse, mortality, recovery)

- status (intervention/treatment vs control/placebo)

Kingston University London

Survival Analysis in R



futime: survival or censoring time
fustat: censoring status
age: in years
resid.ds: residual disease present (1=no,2=yes)
rx: treatment group
ecog.ps: ECOG performance status (1 is better, see reference)

time variable
(time-to-event/censoring)

event variable
(event/censored)

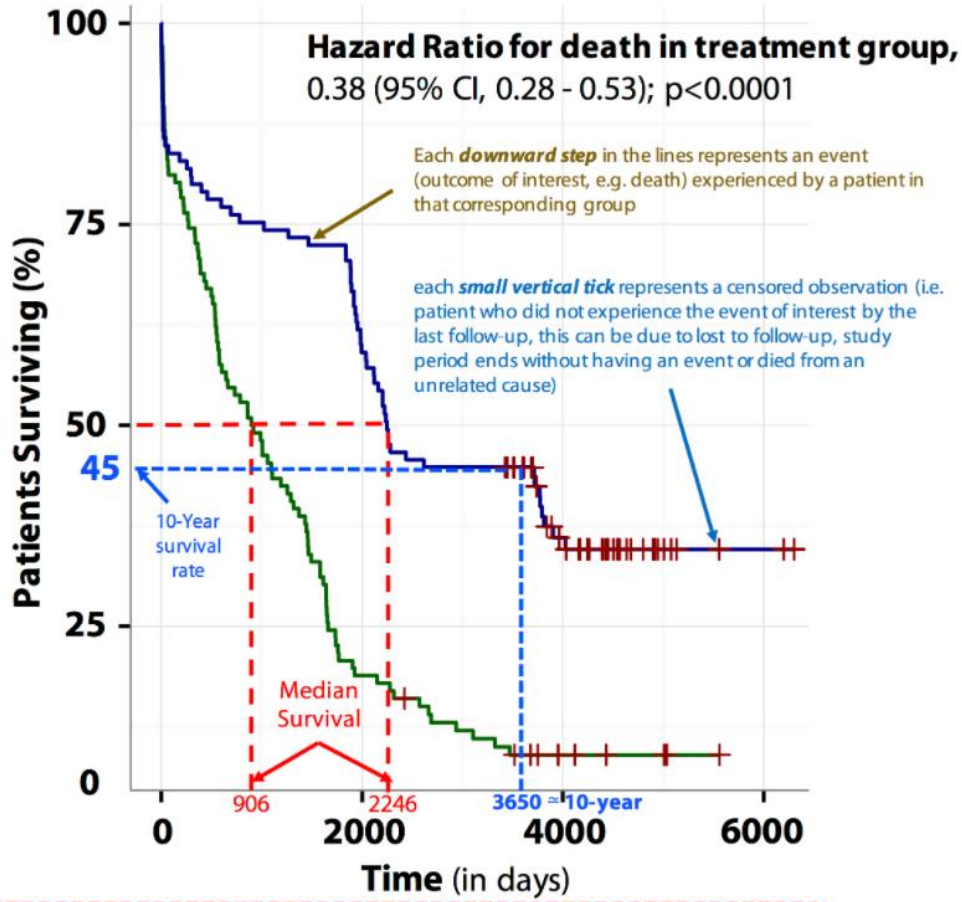
status variable
(treatment group)

```
> library(survival)
> data(ovarian)
> head(ovarian)
  futime fustat     age resid.ds rx ecog.ps
1     59      1 72.3315        2  1       1
2    115      1 74.4932        2  1       1
3    156      1 66.4658        2  1       2
4    421      0 53.3644        2  2       1
5    431      1 50.3397        2  1       1
6    448      0 56.4301        1  1       2
>
> str(ovarian)
'data.frame':   26 obs. of  6 variables:
 $ futime  : num  59 115 156 421 431 448 464 475 477 563 ...
 $ fustat  : num  1 1 1 0 1 0 1 1 0 1 ...
 $ age     : num  72.3 74.5 66.5 53.4 50.3 ...
 $ resid.ds: num  2 2 2 2 2 1 2 2 2 1 ...
 $ rx      : num  1 1 1 2 1 1 2 2 1 2 ...
 $ ecog.ps : num  1 1 2 1 1 2 2 2 1 2 ...
```

Kingston University London

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)



The slope of each curve represents the hazard.

The ratio of hazards in each group should be consistent throughout follow-up (proportional hazard assumption).

The most obvious violation of the proportional hazard assumption would be crossing of the curves.

Kingston University London

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)

Otolaryngology–Head and Neck Surgery (2010) 143, 331-336

INVITED ARTICLE

A practical guide to understanding Kaplan-Meier curves

Jason T. Rich, MD, J. Gail Neely, MD, Randal C. Paniello, MD, Courtney C. J. Voelker, MD, DPhil, Brian Nussenbaum, MD, and **Eric W. Wang, MD,** St. Louis, MO

Table 3
Full accounting of data

Subject	Serial time (yrs) (serial time of event, = "event time")	Interval (ending at event occurrence)	Number "surviving" at risk in the interval (defines the denominator for the interval)	Event (defines end of interval)	Censored (removed from "surviving" in the interval)	Number "surviving" after event (defines the numerator)	Calc: interval "survival" rate after event	Interval "survival" rate after event	Calc: cumulative "survival" rate	Cumulative "survival" rate
Group 1	0	1								1.000
B	1	2	6	1	0	5	5 of 6	0.833	1.000 * 0.833	0.833
E	2	3	5	1	0	4	4 of 5	0.800	0.833 * 0.800	0.667
F	3	4	4	1	0	3	3 of 4	0.750	0.667 * 0.750	0.500
A	4	5	3	1	0	2	2 of 3	0.667	0.500 * 0.667	0.333
D	4.5	6	2	1	0	1	1 of 2	0.500	0.333 * 0.500	0.167
C	5			0	1					
Group 2	0	1								1.000
U	0.5	2	6	1	0	5	5 of 6	0.833	1.000 * 0.833	0.833
Z	0.75	3	5	1	0	4	4 of 5	0.800	0.833 * 0.800	0.667
W	1	4	4	1	0	3	3 of 4	0.750	0.667 * 0.750	0.500
V	1.5			0	1					
X	2	5	2	1	0	1	1 of 2	0.500	0.500 * 0.500	0.25
Y	3.5	6	1	1	0	0	0 of 1	0	0.25 * 0	0

Calc, calculation.

Kingston University London

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)

RDocumentation

R Enterprise Training | R package | Leaderboard

survfit

From survival v2.11-4
by Thomas Lumley

99th Percentile

Compute A Survival Curve For Censored Data

Computes an estimate of a survival curve for censored data using either the Kaplan-Meier or the Fleming-Harrington method or computes the predicted survivor function for a Cox proportional hazards model.

Usage

```
survfit(formula, data, weights, subset, na.action,
        newdata, individual=F, conf.int=.95, se.fit=T,
        type=c("kaplan-meier","fleming-harrington", "fh2"),
        error=c("greenwood","tsiatis"),
        conf.type=c("log","log-log","plain","none"),
        conf.lower=c("usual", "peto", "modified"))
basehaz(fit,centered=TRUE)
```

survfit: Create survival curves
In survival: Survival Analysis

Examples

```
#fit a Kaplan-Meier and plot it
data(aml)
fit <- survfit(Surv(time, status) ~ x, data=aml)
plot(fit)

# plot only 1 of the 2 curves from above
plot(fit[2])


#fit a cox proportional hazards model and plot the
#predicted survival curve
data(ovarian)
fit <- coxph( Surv(futime,fustat)~resid.ds+rx+ecog.ps,data=ovarian)
plot( survfit( fit))
```

Kingston University London

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)

survfit: Create survival curves
In survival: Survival Analysis

The usage of the function `survfit()`:
`survfit(Surv(survivaltime, event) ~ group, data = df)`

```
library("survival")                  # load the R package "survival"
data(leukemia)                       # load the dataset "leukemia"
df <- leukemia                       # rename the dataset "df" (optional)
df$survivaltime <- df$time           # rename the time variable as "survivaltime"
df$event <- df$status                # rename the event variable as event
df$group <- df$x                     # rename the treatment group as "group"

fit <- survfit(Surv(survivaltime, event) ~ group, data = df)
                                     # model the survival data
plot(fit)                            # plot the K-M plot based on the created model
summary(fit)                         # get the details of the model
```

Kingston University London

Script: `survival.R`

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)

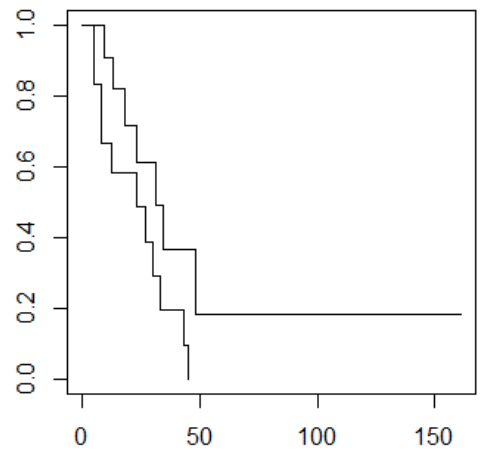
RGui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
>
> library("survival")# load the R package "survival"
> data(leukemia) # load the dataset "leukemia"
> df <- leukemia# rename the dataset "df" (optional)
> df$survivaltime <- df$time# rename the time variable as "survivaltime"
> df$event <- df$status# rename the event variable as event
> df$group <- df$x# rename the treatment group as "group"
>
> fit <- survfit(Surv(survivaltime, event) ~ group, data = df)
> # model the survival data
> plot(fit)# plot the K-M plot based on the created model
> summary(fit)
Call: survfit(formula = Surv(survivaltime, event) ~ group, data = df)

                group=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867       0.7541        1.000
   13     10       1    0.818  0.1163       0.6192        1.000
   18      8       1    0.716  0.1397       0.4884        1.000
   23      7       1    0.614  0.1526       0.3769        0.999
   31      5       1    0.491  0.1642       0.2549        0.946
   34      4       1    0.368  0.1627       0.1549        0.875
   48      2       1    0.184  0.1535       0.0359        0.944

                group=Nonmaintained
```

R Graphics: Device 2 (ACTIVE)



Kingston University London

Script: `survival.R`

Survival Analysis in R

Basic Approach
(Kaplan-Meier Curve)

survfit: Create survival curves
In survival: Survival Analysis

A little improvement on the basic plot:

```
plot(fit, lty = c(1,2), col = c("green", "red"), xlab = "Time
(days)", ylab = "Survival Probability", main = "An Example
Survival Curve")

legend(115, 1, legend = c("CT Maintained", "CT not Maintained"),
col = c("green", "red"), lty = 1:2, cex = 0.8)
```

OR

```
# legend("topright", legend = c("CT Maintained", "CT not
Maintained"), col = c("green", "red"), lty = 1:2, cex = 0.8)
```

Kingston University London

Script: `survival.R`

Survival Analysis in R

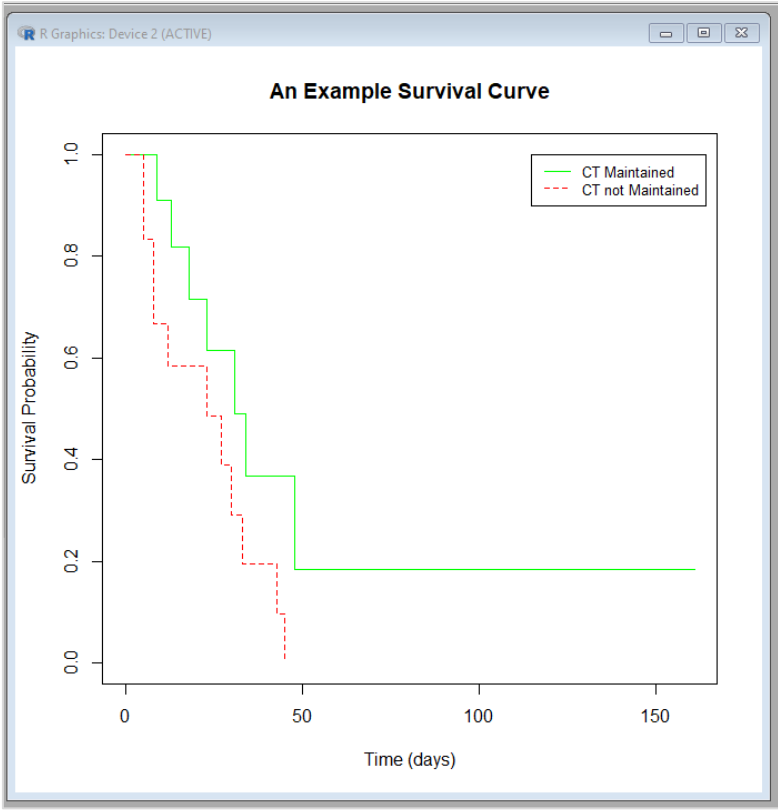
Basic Approach
(Kaplan-Meier Curve)

RGui (64-bit)

File History Resize Windows

R Console

```
>
>
> plot(fit, lty = c(1,2), col = c("green", "red"), xlab = "Time (days)", ylab = "Survival Probability", main = "An Example Survival Curve")
>
> legend(115, 1, legend = c("CT Maintained", "CT not Maintained"), col = c("green", "red"), lty = 1:2, cex = 0.8)
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
>
```



Kingston University London

Script: `survival.R`

Survival Analysis in R

Basic Approach
(log-rank test)

survdiff: Test Survival Curve Differences
In survival: Survival Analysis

Log-rank statistics for equality of survivor functions:

survdiff(Surv(time, status) ~ **x**)

Looks at the survival differences by the group/status variable "**x**" like "caco"

time: time-to-event
status: the event variable (yes/no)
x: group variable (intervention/placebo; male/female)

Kingston University London

Script: `survival.R`

Survival Analysis in R

Basic Approach
(log-rank test)

RDocumentation

Alternative log-rank test methods

coin (version 1.3-1)

SurvivalTests: Two- and \(K\)-Sample Tests for Censored Data

Description

Testing the equality of the survival distributions in two or more independent groups.

```
## Exact logrank test
logrank_test(Surv(time, event) ~ group, data = g3,
             distribution = "exact")
```

When survival times are tied, use the exact test (data = "glioma" in this example). Package = coin

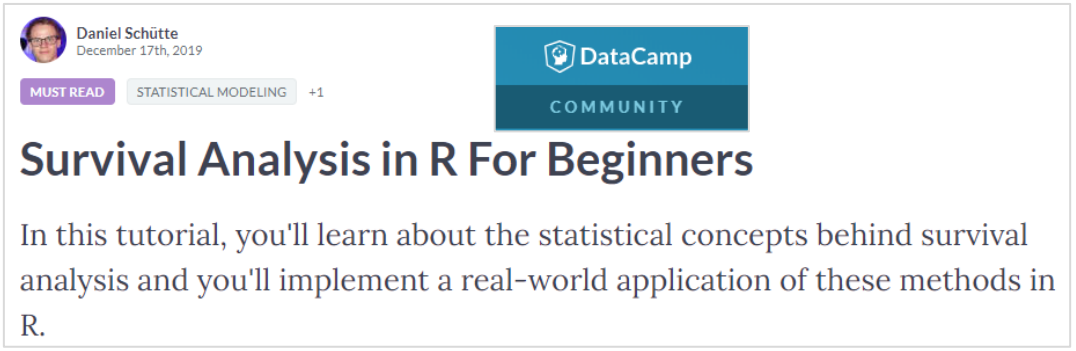
Usage

```
logrank_test(object, ties.method = c("mid-ranks", "Hothorn-Lausen",
                                     "average-scores"),
             type = c("logrank", "Gehan-Breslow", "Tarone-Ware",
                      "Peto-Peto", "Prentice", "Prentice-Marek",
                      "Andersen-Borgan-Gill-Keiding",
                      "Fleming-Harrington", "Gaugler-Kim-Liao", "Self"),
             rho = NULL, gamma = NULL, ...)
```

Kingston University London

Script: `survival.R`

Survival Analysis in R



Create a survival object using time and event variables:

```
> surv_object <- Surv(time = ovarian$futime, event = ovarian$fustat)
> surv_object
 [1]   59   115   156   421+  431   448+  464   475   477+  563   638   744+
[13]  769+  770+  803+  855+ 1040+ 1106+ 1129+ 1206+ 1227+  268   329   353
[25]  365   377+
```

> These are times to the event or los to follow-up. A (+) behind survival times indicates censored data points.

This survival object is used in coxph() function to fit Cox proportional hazard model when covariates are present

The `survival` package is the cornerstone of the entire R survival analysis edifice. Not only is the package itself rich in features, but the object created by the `Surv()` function, which contains failure time and censoring information, is the basic survival analysis data structure in R. Dr. Terry Therneau, the package author, began working on the survival package in 1986. The first public release, in late 1989, used the Statlib service hosted by Carnegie Mellon University. Thereafter, the package was incorporated directly into Splus, and subsequently into R.

Script: `survival.R`

Kingston University London

Survival Analysis in R

Cox PH Modelling
(for multivariable analysis)

coxph: Fit Proportional Hazards Regression Model
In survival: Survival Analysis

Cox Proportional Hazard modelling for equality of survivor functions with adjustments for potential confounders:

coxph(Surv(time, status) ~ **x + age + sex + eth + ses**)

Examines the survival differences by the group/status variable "**x**" like "caco" with the possibility of including potential confounders in the model

time: time-to-event
status: the event variable (yes/no)
x: group variable (intervention/placebo)
age, sex, eth, ses (etc.): potential confounders

Kingston University London

Script: `survival.R`

Survival Analysis in R

Cox PH Modelling

(for multivariable analysis)

Usage of coxph() with covariates:

```
coxph (survival_object ~ group_variable + covariate1 + covariate2 + … )
```

EXAMPLE:

```
library(survival)
data(veteran)
cox <- coxph(Surv(time, status) ~ trt + celltype + karno + diagtime + age
                                                  + prior , data = veteran)
summary(cox)

# OR

library(survival)
data(veteran)
survival_object = Surv(veteran$time, veteran$status)
cox <- coxph(survival_object ~ trt + celltype + karno + diagtime + age
                                                  + prior , data = veteran)
summary(cox)
```

Kingston University London

Script:
survival.R and survival_coxph.R

Survival Analysis in R

Cox PH Modelling
(for multivariable analysis)

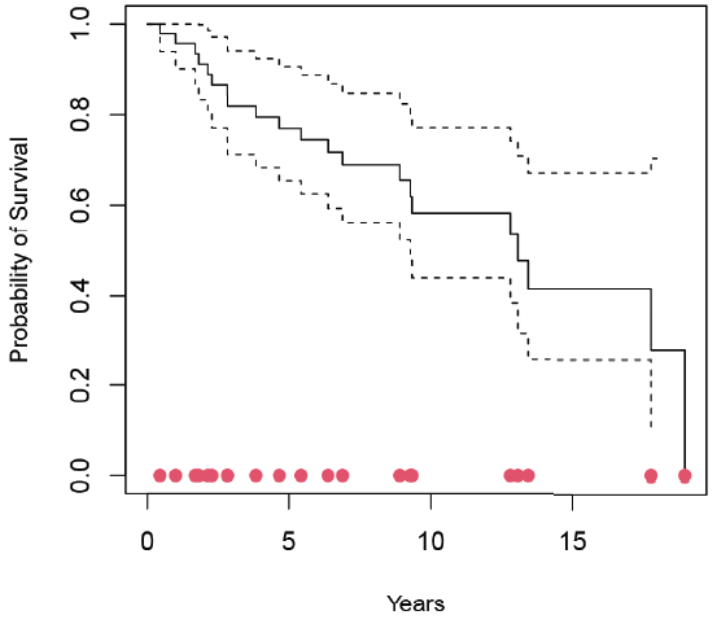


Figure 1: Kaplan–Meier survival curve (with 95% confidence bands) for a set of 50 patients. Red dots indicate the times at which a patient died.

Cox Proportional Hazard model is for equality of survivor functions with **adjustments** for potential confounders.

SIGNIFICANCE
ROYAL STATISTICAL SOCIETY
ASA AMERICAN STATISTICAL ASSOCIATION
Statistical Society of Australia

Profiles | Full Access

What is Cox's proportional hazards model?

Robert Tibshirani

First published: 29 March 2022 | https://doi.org/10.1111/1740-9713.01633

Kingston University London

Survival Analysis in R

Cox PH Modelling
(for multivariable analysis)

coxph() function of the survival package is used to fit a Cox proportional hazard (PH) model

coxph

Fit Proportional Hazards Regression Model

survival_object = `Surv(time = time-to-event variable, event = event/status variable)`

Usage of coxph():

```
coxph (survival_object ~ group_variable)
```

>>> equivalent to univariate log-rank test <<<

Usage of coxph() with covariates:

```
coxph (survival_object ~ group_variable + covariate1 + covariate2 + … )
```

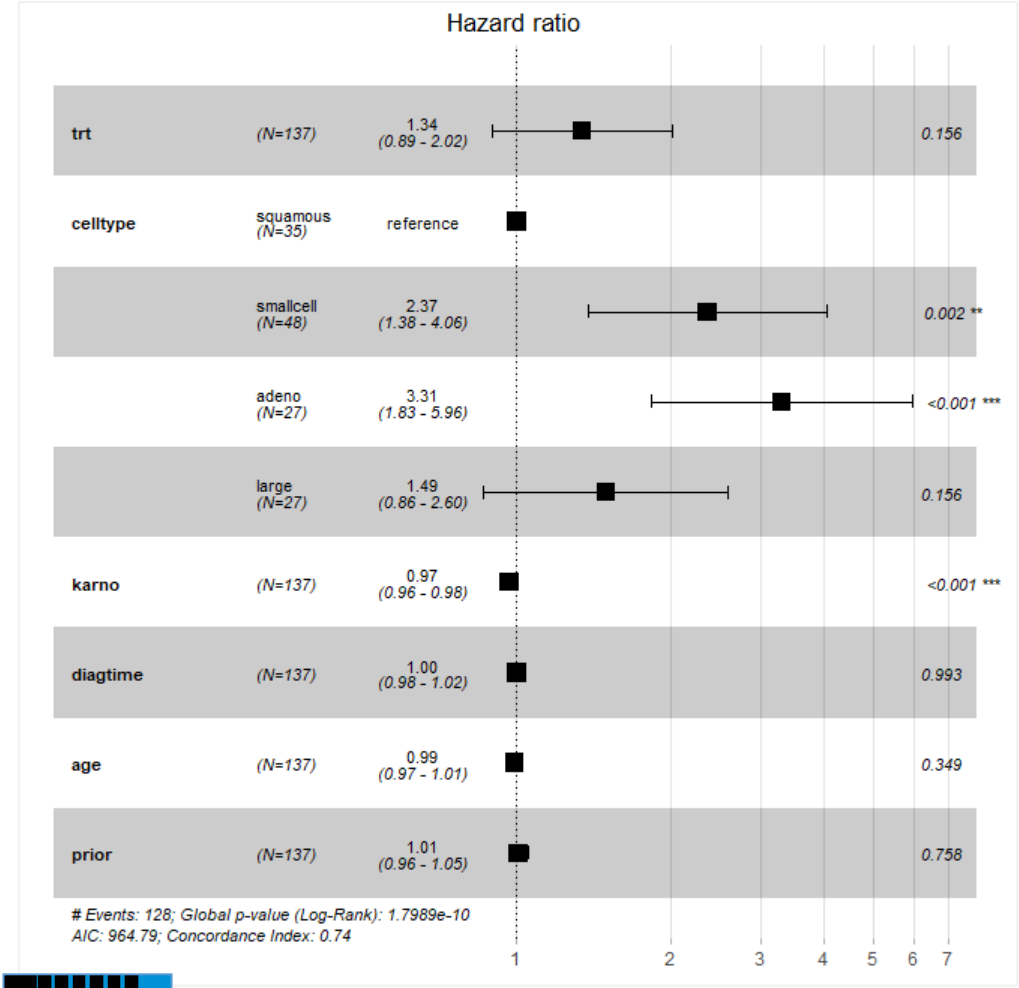
Kingston University London

Script: `survival.R`

Survival Analysis in R

Cox PH Modelling

(visualising the multivariable analysis)



RDocumentation

Search all packages and functions

survminer (version 0.4.9)

ggforest: Forest Plot for Cox Proportional Hazards Model

Description

Drawing Forest Plot for Cox proportional hazards model. In two panels the model structure is presented.

```
library(survminer)
ggforest(cox, data = veteran)
# cox is the coxph object generated in the previous slide
```

Script:
`survival.R and survival_coxph.R`

Kingston University London

Survival Analysis in R

Cox PH Modelling
(assumptions)

The easiest way is the visual examination of the univariate Kaplan-Meier curves. **If during the follow-up, the curves of the univariate analysis cross each other or have sharp turns, it is most likely that the preconditions for the proportional hazard assumption are not fulfilled**. [Unterrainer *et al*, *HLA* 2018]

A major assumption of the Cox proportional hazards model is that the effect of a given covariate does not change over time. If this assumption is violated, the simple Cox model is invalid, and more sophisticated analyses are required. **Graphical methods are available for detecting violations of the proportional hazards assumption**. Smoothed plots of the scaled Schoenfeld residuals are recommended for assessing PH violations because they provide information about the time dependence of the covariate effects. [Hess KR. *Stat Med* 1995]

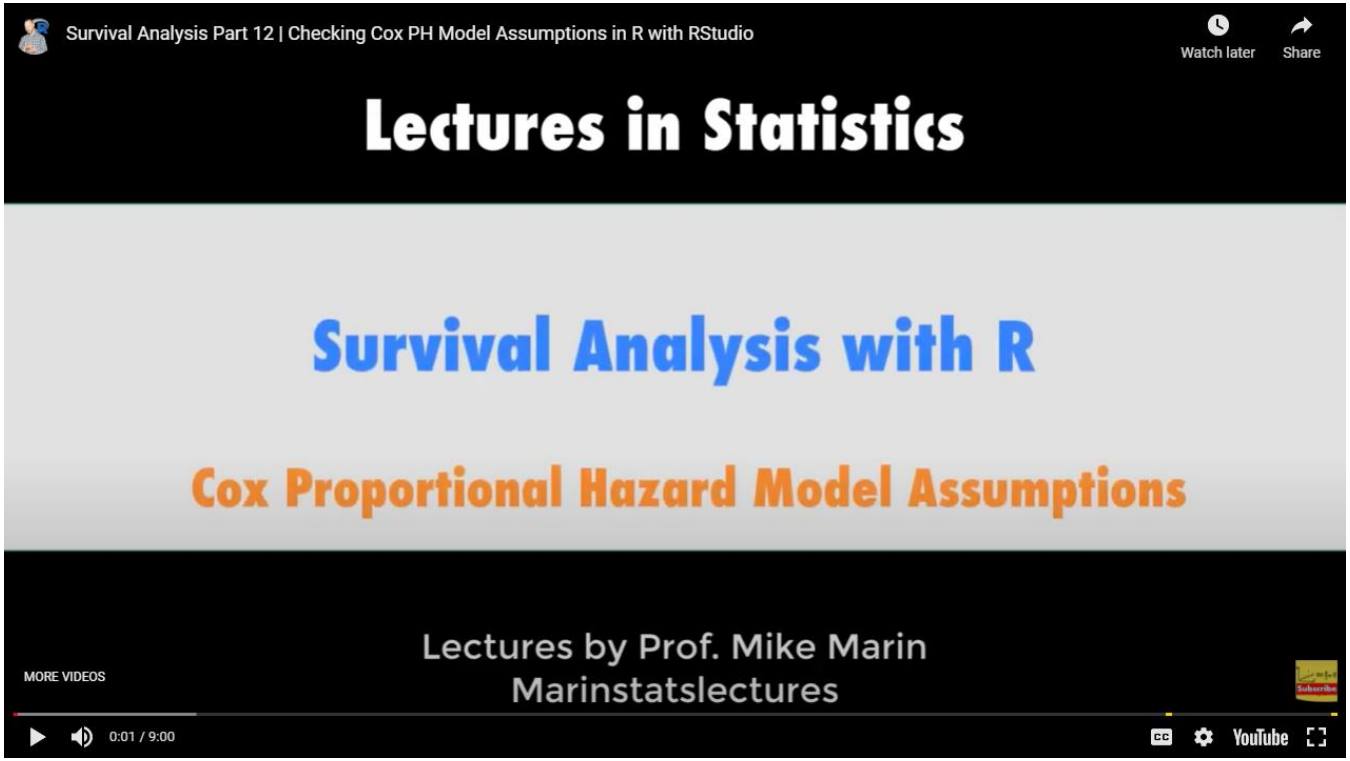
See also:

3.6 How to evaluate the PH assumption?

Survival Analysis

Cox PH Modelling
(assumptions)

For a review of checking Cox PH model assumptions in R, see this video:



Kingston University London

Script: `survival_coxph.R`

Survival Analysis in R

Visualising survival analysis results

Package 'survminer'

May 28, 2020

Type Package

Title Drawing Survival Curves using 'ggplot2'

Version 0.4.7

Date 2020-05-28

Description Contains the function 'ggsurvplot()' for drawing easily beautiful and 'ready-to-publish' survival curves with the 'number at risk' table and 'censoring count plot'. Other functions are also available to plot adjusted curves for `Cox` model and to visually examine 'Cox' model assumptions.

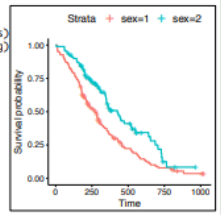
Creating Survival Plots

Informative and Elegant with *survminer*

Survival Curves

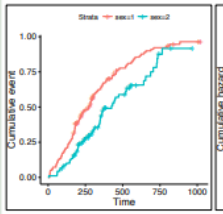
The **ggsurvplot()** function creates **ggplot2** plots from **survfit** objects.

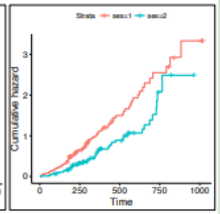
```
library("survival")
fit <- survfit(Surv(time,status)
               ~ sex, data = lung)

class(fit)
## [1] "survfit"

library("survminer")
ggsurvplot(fit, data = lung)
```



Use the **fun** argument to set the transformation of the survival curve. E.g. **"event"** for cumulative events, **"cumhaz"** for the cumulative hazard function or **"pct"** for survival probability in percentage.

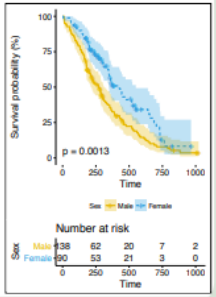
```
ggsurvplot(fit, data = lung, fun = "event")
ggsurvplot(fit, data = lung, fun = "cumhaz")
```





With lots of graphical parameters you have full control over look and feel of the survival plots; position and content of the legend; additional annotations like p-value, title, subtitle.

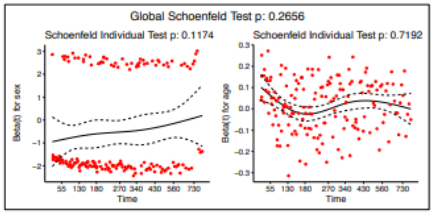
```
ggsurvplot(fit, data = lung,
    conf.int = TRUE,
    pval = TRUE,
    fun = "pct",
    risk.table = TRUE,
    size = 1,
    linetype = "strata",
    palette = c("#E7B800",
                "#2E9FDF"),
    legend = "bottom",
    legend.title = "Sex",
    legend.labs = c("Male",
                    "Female"))
```



Diagnostics of Cox Model

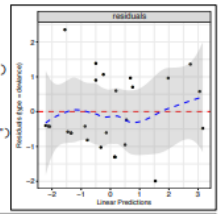
The function **cox.zph()** from **survival** package may be used to test the proportional hazards assumption for a Cox regression model fit. The graphical verification of this assumption may be performed with the function **ggcoxzph()** from the **survminer** package. For each covariate it produces plots with scaled Schoenfeld residuals against the time.

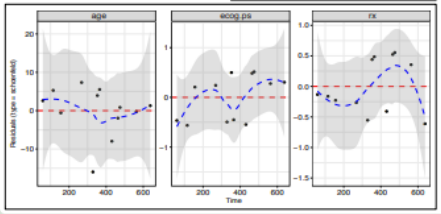
```
library("survival")
fit <- coxph(Surv(time, status) ~ sex + age, data = lung)
ftest <- cox.zph(fit)
ftest
##             rho chisq     p
## sex     0.1236 2.452 0.117
## age    -0.0275 0.129 0.719
## GLOBAL      NA 2.651 0.266
library("survminer")
ggcoxzph(ftest)
```



The function **ggcoxdiagnostics()** plots different types of residuals as a function of time, linear predictor or observation id. The type of residual is selected with **type** argument. Possible values are "martingale", "deviance", "score", "schoenfeld", "dfbeta", "dfbetas", and "scaledsch".
The **ox.scale** argument defines what shall be plotted on the OX axis. Possible values are "linear.predictions", "observation.id", "time".
Logical arguments **hline** and **sline** may be used to add horizontal line or smooth line to the plot.

```
library("survival")
library("survminer")
fit <- coxph(Surv(time, status) ~
             sex + age, data = lung)

ggcoxdiagnostics(fit,
    type = "deviance",
    ox.scale = "linear.predictions")

ggcoxdiagnostics(fit,
    type = "schoenfeld",
    ox.scale = "time")
```

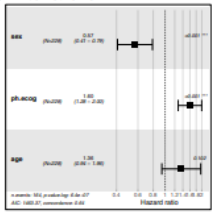




Summary of Cox Model

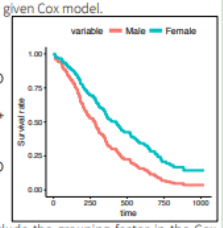
The function **ggforest()** from the **survminer** package creates a forest plot for a Cox regression model fit. Hazard ratio estimates along with confidence intervals and p-values are plotter for each variable.

```
library("survival")
library("survminer")
lung$age <- ifelse(lung$age > 70, ">70","<= 70")
fit <- coxph( Surv(time, status) ~ sex + ph.ecog + age, data = lung)
fit

## Call:
## coxph(formula = Surv(time, status) ~ sex+ph.ecog+age, data=lung)
##
##              coef exp(coef) se(coef)     z       p
## sex      -0.567     0.567    0.168 -3.37 0.00075
## ph.ecog  0.470     1.600    0.113  4.16 3.1e-05
## age>70   0.307     1.359    0.187  1.64 0.10175
##
## Likelihood ratio test=31.6  on
## n= 227, number of events= 164

ggforest(fit)
```



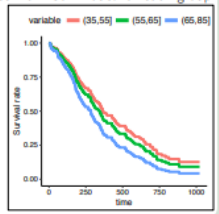
The function **ggadjustedcurves()** from the **survminer** package plots Adjusted Survival Curves for Cox Proportional Hazards Model. Adjusted Survival Curves show how a selected factor influences survival estimated from a Cox model.
Note that these curves differ from Kaplan Meier estimates since they present expected ssurvival based on given Cox model.

```
library("survival")
library("survminer")

lung$sex <- ifelse(lung$sex == 1,
                   "Male", "Female")

fit <- coxph(Surv(time, status) ~
             sex + ph.ecog + age +
             strata(sex),
             data = lung)

ggcoxadjustedcurves(fit, data=lung)
```



Note that it is not necessary to include the grouping factor in the Cox model. Survival curves are estimated from Cox model for each group defined by the factor independently.

```
lung$age3 <- cut(lung$age,
                 c(35,55,65,85))

ggcoxadjustedcurves(fit, data=lung,
                    variable="lage3")
```



This onepager presents the **survminer** package [Alboukadel Kassambara, Marcin Kosinski 2017] in version 0.4.0.999
See *https://github.com/kassambara/survminer/* for more details.

Kingston University London

Survival Analysis in R

For a more detailed survival analysis using the dataset "lung",
follow the annotated script "survival_survminer.R"
@ http://www.dorak.info/r/survival_survminer.R

Go to: http://www.dorak.info/r
locate: survival_survminer.R
and right click to choose Save link as… to download the script file

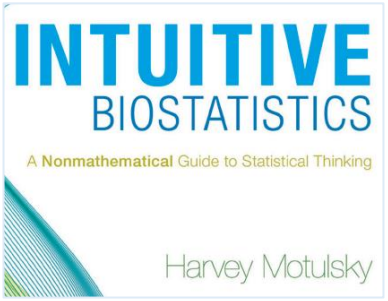
Kingston University London

Script: `survival_survminer.R`

Survival Analysis

CHAPTER 29

Comparing Survival Curves



A Tour of Survival Analysis

By Alexander Moreno — 3 Comments

BU
Boston University School of Public Health

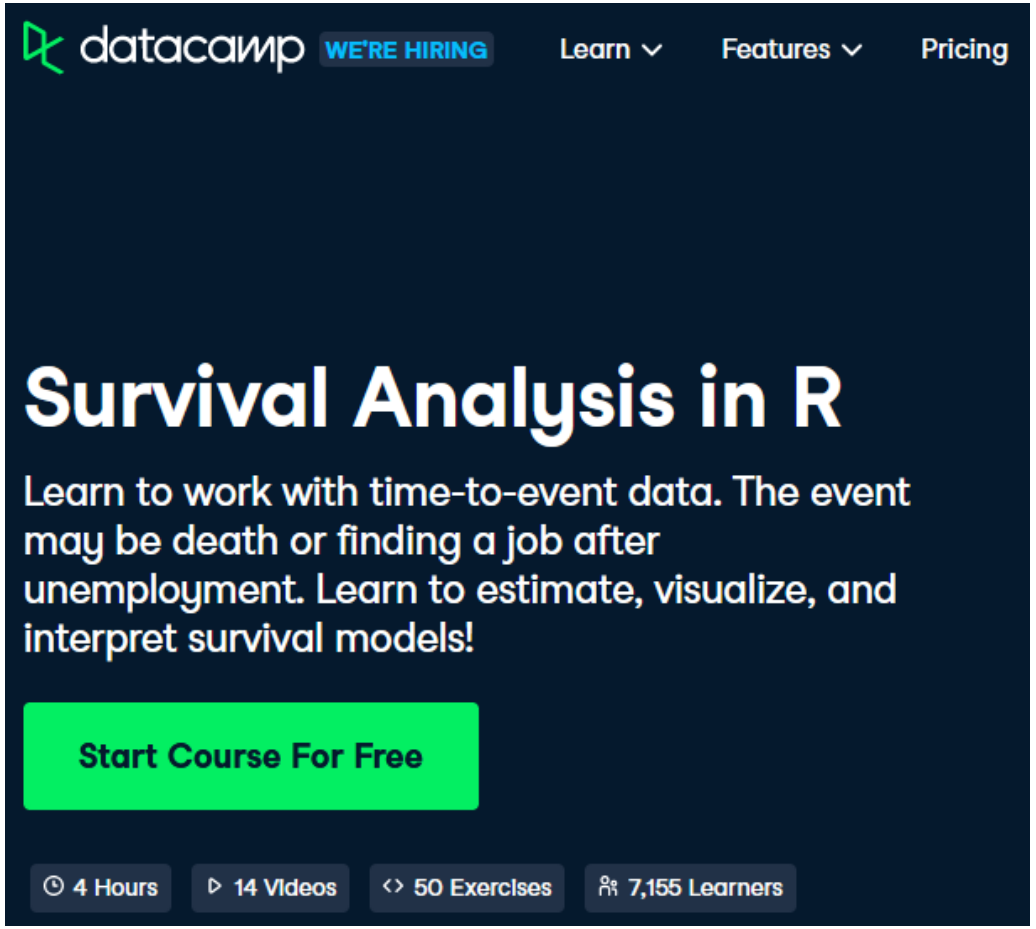
Basic Statistical Analysis Using the R Statistical Package

Survival Analysis

Survival Analysis with R: Exercises

Kingston University London

Survival Analysis



1 What is Survival Analysis? FREE

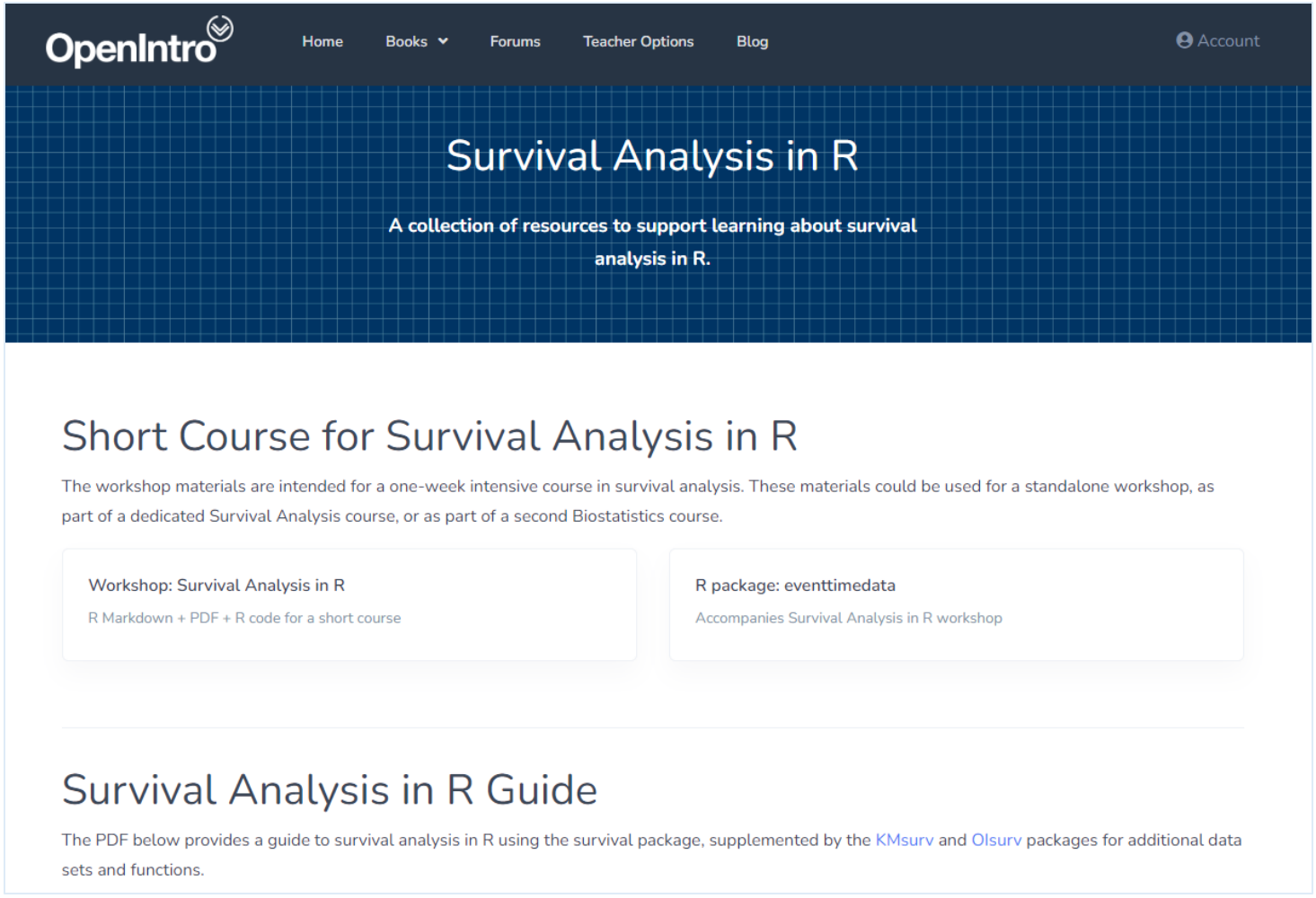
In the first chapter, we introduce the concept of survival analysis, explain the importance of this topic, and provide a quick introduction to the theory behind survival curves. We discuss why special methods are needed when dealing with time-to-event data and introduce the concept of censoring. We also discuss how we describe the distribution of the elapsed time until an event.

View chapter details +

Play Chapter Now

Kingston University London

Survival Analysis



Statistical Power

Background

In Survival Analysis

Statistical Power

Background

[**Supplement** An Overview of Study Design and Statistical Considerations]

CHEST

Sample Size Estimation in Clinical Research

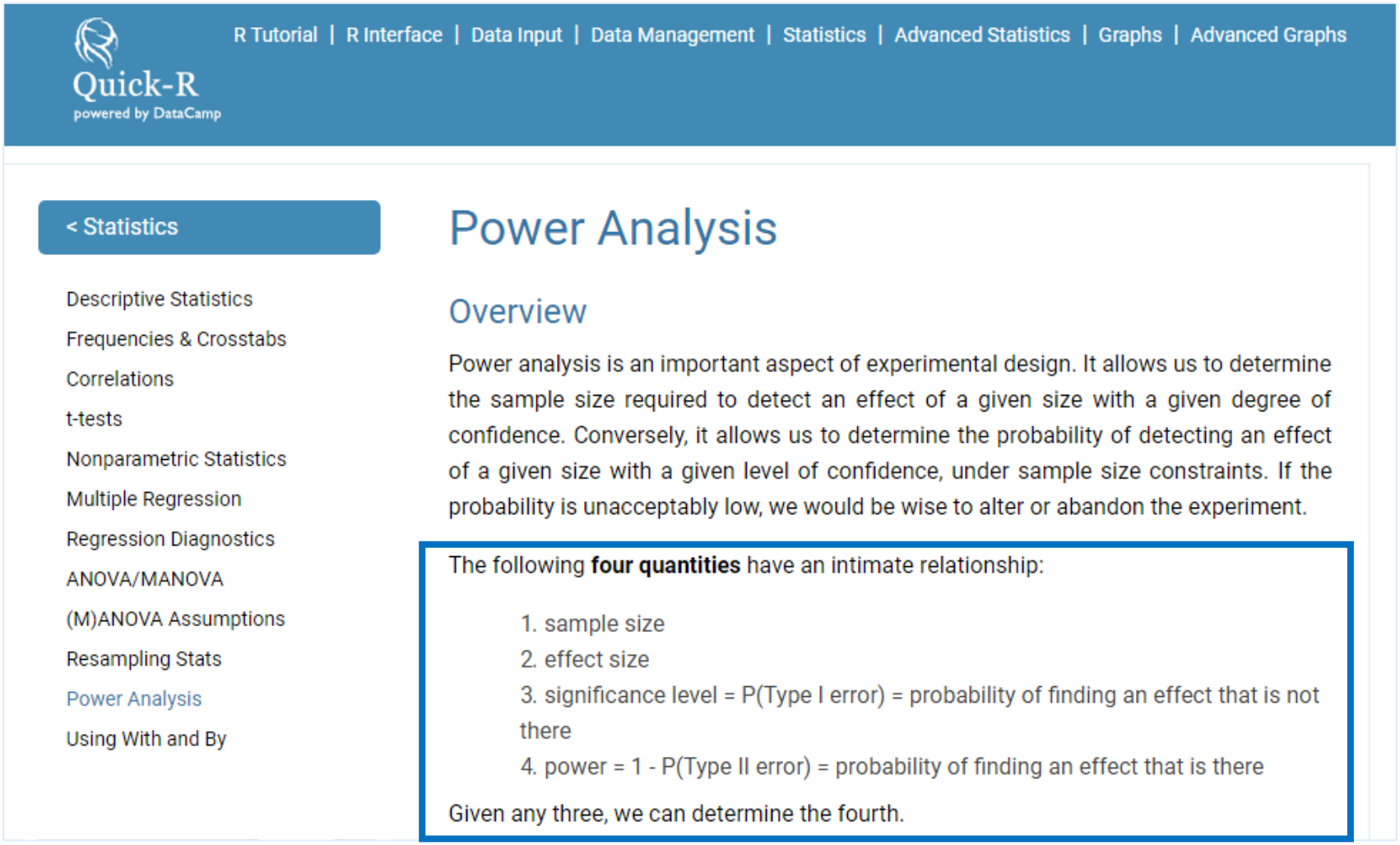
From Randomized Controlled Trials to Observational Studies

Xiaofeng Wang, PhD; and Xinge Ji, MS

Check for updates

Kingston University London

Statistical Power



Kingston University London

Statistical Power

pwr v1.2-2

by Helios De Rosario

Functions in pwr

Search

Name	Description
pwr.2p.test	Power calculation for two proportions (same sample sizes)
pwr.anova.test	Power calculations for balanced one-way analysis of variance tests
pwr.chisq.test	power calculations for chi-squared tests
pwr.t.test	Power calculations for t-tests of means (one sample, two samples and paired samples)
pwr.t2n.test	Power calculations for two samples (different sizes) t-tests of means
ES.h	Effect size calculation for proportions
ES.w1	Effect size calculation in the chi-squared test for goodness of fit
pwr.f2.test	Power calculations for the general linear model
pwr.norm.test	Power calculations for the mean of a normal distribution (known variance)
plot.power.htest	Plot diagram of sample size vs. test power
pwr-package	Basic Functions for Power Analysis pwr
ES.w2	Effect size calculation in the chi-squared test for association
cohen.ES	Conventional effects size
pwr.2p2n.test	Power calculation for two proportions (different sample sizes)
pwr.p.test	Power calculations for proportion tests (one sample)
pwr.r.test	Power calculations for correlation test

Kingston University London

pwr v1.2-2

by Helios De Rosario

Statistical Power

The following **four quantities** have an intimate relationship:

1. sample size
2. effect size
3. significance level = P(Type I error) = probability of finding an effect that is not there
4. power = 1 - P(Type II error) = probability of finding an effect that is there

Given any three, we can determine the fourth.

Power analysis using "pwr":

```
library(pwr)
pwr.t2n.test(sig.level = 0.05, n1 = 32, n2 = 23, power = 0.8,
alternative = "two.sided")     # 1,2,3 are provided
```

To obtain the effect size (d) or # 4 above

Result:

```
     t test power calculation
             n1 = 32
             n2 = 23
              d = 0.7800906          # d is Cohen's d value;
                                      a generic effect size measure
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

Kingston University London

Statistical Power

Now, run Script: `pwr.R`

Kingston University London

pwr v1.2-2

by Helios De Rosario

Statistical Power in Survival Analysis

Sample Size for Survival Analysis

Menu location: **Analysis_Sample Size_Survival Times**.

This function gives you the minimum number of subjects that you require to detect a true ratio of median survival times (hr) with power POWER and two sided type I error probability ALPHA (Dupont, 1990; Schoenfeld and Richter, 1982).

The method used here is suitable for calculating sample sizes for studies that will be analysed by the log-rank test.

Information required

- POWER: probability of detecting a real effect.
- ALPHA: probability of detecting a false effect (two sided: double this if you need one sided).
- A: accrual time during which subjects are recruited to the study.
- F: additional follow-up time after the end of recruitment.
- *: input either (C and r) or (C and E), where r=E/C.
- C: median survival time for control group.
- E: median survival time for experimental group.
- r: hazard ratio or ratio of median survival times.
- M: number of controls per experimental subject.

Practical issues

- Usual values for POWER are 80%, 85% and 90%; try several in order to explore/scope.
- 5% is the usual choice for ALPHA.
- C is usually estimated from previous studies.
- If possible, choose a range of hazard ratios that you want have the statistical power to detect.

Technical validation

The estimated sample size per group n is calculated as:

$$n = \left(z_{\alpha/2} + z_{\beta}\right)^2 \left(\frac{(1+1/m)/p}{\ln(r)^2}\right)$$

$$p = 1 - p_a \exp(-\ln(2)F/m)$$

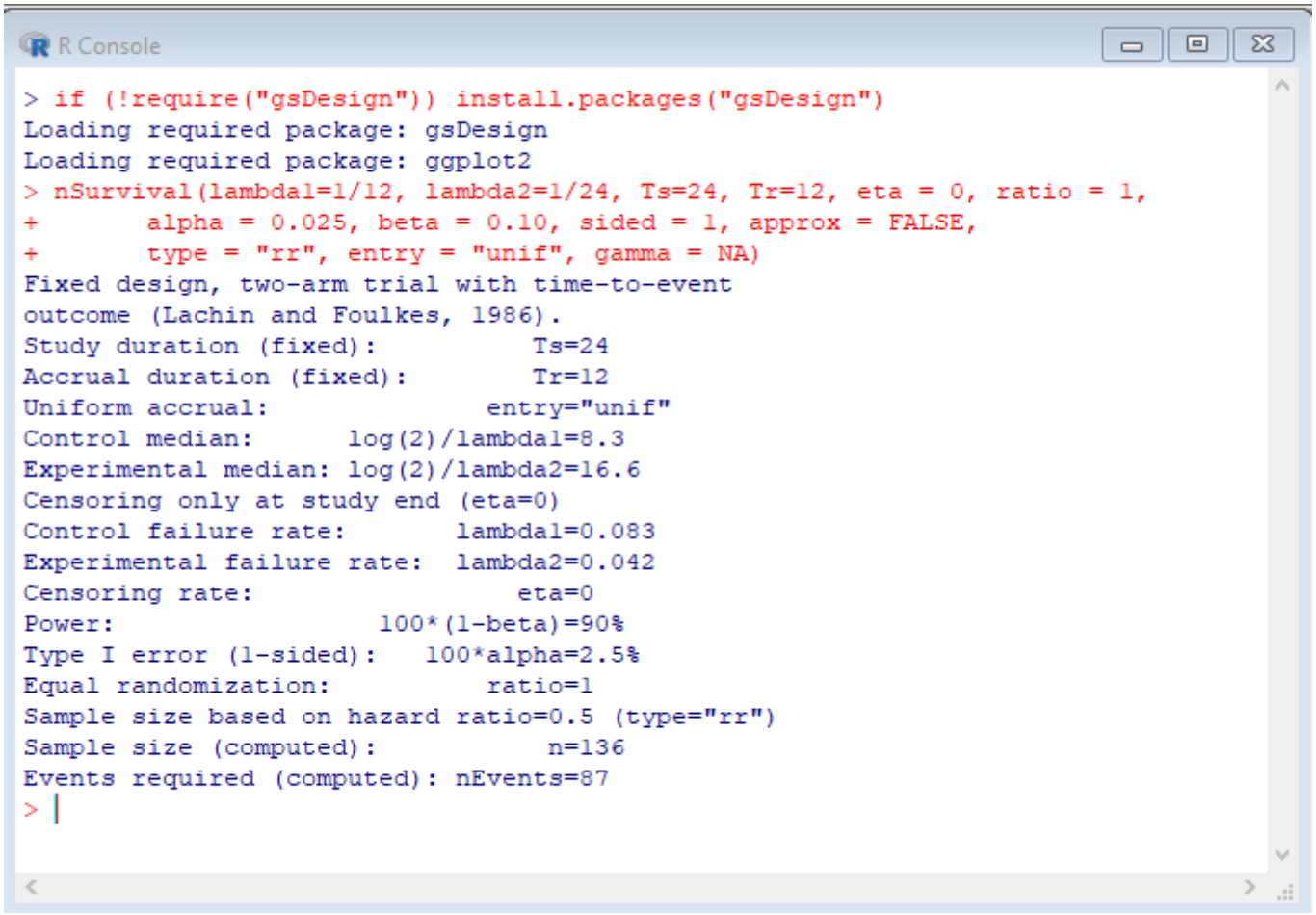
$$p_a = 1 - \frac{\exp(-\ln(2)A/m)}{\ln(2)A/m}$$

$$m = (C + E)/2$$

- where α = alpha, β = 1 - power and z_p is the standard normal deviate for probability p. n is rounded up to the closest integer. (1+1/m)/p is equivalent to 2/p in the first equation if the experimental and control group sizes are unequal.

Kingston University London

Statistical Power in Survival Analysis



Script: `Time-to-event Sample Size Calculation.R`

Kingston University London

Statistical Power in Survival Analysis

nSurvival: 3.4: Time-to-event sample size calculation (Lachin-Foulkes)

In gsDesign: Group Sequential Design

Description | Usage | Arguments | Details | Value | Author(s) | References | See Also | Examples

Description

`nSurvival()` is used to calculate the sample size for a clinical trial with a time-to-event endpoint. The Lachin and Foulkes (1986) method is used. `nEvents` uses the Schoenfeld (1981) approximation to provide sample size and power in terms of the underlying hazard ratio and the number of events observed in a survival analysis. The functions `hrz2n()`, `hrn2z()` and `zn2hr()` also use the Schoenfeld approximation to provide simple translations between hazard ratios, z-values and the number of events in an analysis; input variables can be given as vectors.

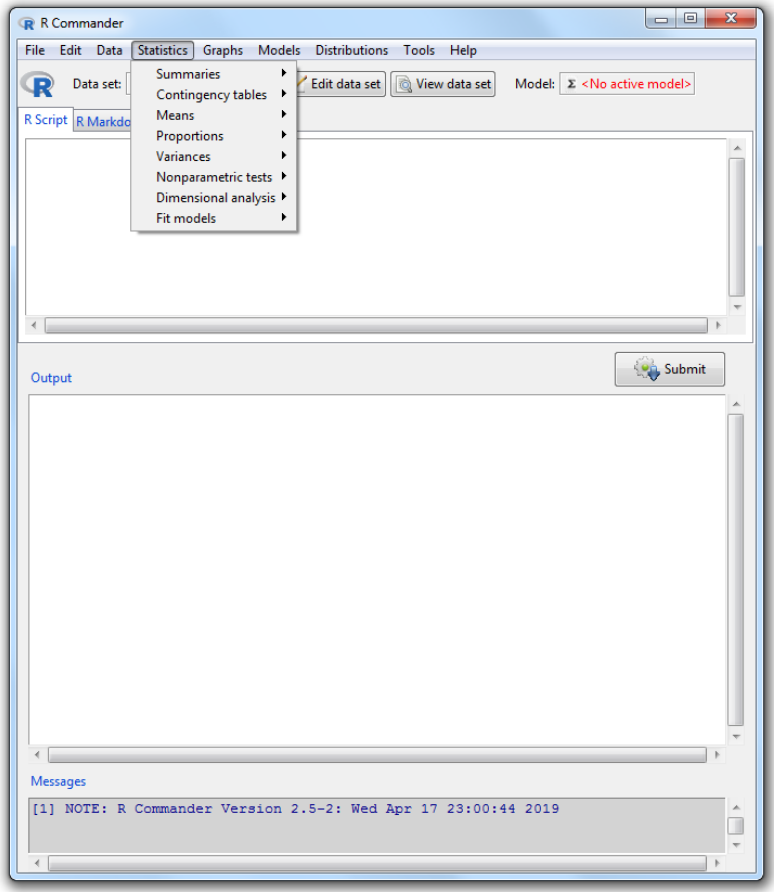
Usage

```
nSurvival(lambda1=1/12, lambda2=1/24, Ts=24, Tr=12, eta = 0, ratio = 1,
      alpha = 0.025, beta = 0.10, sided = 1, approx = FALSE,
      type = c("rr", "rd"), entry = c("unif", "expo"), gamma = NA)
## S3 method for class 'nSurvival'
print(x,...)
nEvents(hr = .6, alpha = .025, beta = .1, ratio = 1, sided = 1,
        hr0 =  1, n = 0, tbl = FALSE)
hrn2z(hr, n, ratio=1, hr0=1, hr1=.7)
hrz2n(hr, z, ratio=1, hr0=1)
zn2hr(z, n, ratio=1, hr0=1, hr1=.7)
```

Kingston University London

R Commander

For those of you who do not like to type a few lines of code…..



… there is always R Commander!

R Commander for Survival Analysis

The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis

John Fox
McMaster University

Marilia Sá Carvalho
Escola Nacional de Saúde Pública

PDF R Commander an introduction
https://cran.r-project.org/doc/contrib/Karp-Rcommander-intro2.pdf
R Commander an introduction Natasha A. Karp nk3@sanger.ac.uk Jan 2014 Preface This material is intended as an introductory guide to data analysis with R Commander. It was written as part of an applied statistics course, given at the Wellcome Trust Sanger Institute, Hinxton, UK.

A Guide to Data Analysis in R Commander

Viann Nguyen-Feng, M.A.
Mark A. Stellmack, Ph.D.
University of Minnesota

A Comparative Review of the R Commander GUI for R

by Robert A. Muenchen

Kingston University London

R Commander for Survival Analysis

R Commander for Survival Analysis

Install Survival Package

To do survival analysis using R Commander, first you have to install **RcmdrPlugin.survival** package. The process is similar to installing R Commander. To do this, just like installing R Commander, in R Console click **Packages** select **Install package(s)…**, you will be asked to choose CRAN mirror site. I would choose a site that is close to my current location. And then, select **RcmdrPlugin.survival**. The package will be installed. To run it, you can first run **R Commander** and then load the **RcmdrPlugin.survival** package.

Use Software R to do Survival Analysis and Simulation. A tutorial

Mai Zhou

The RcmdrPlugin.survival Package: Extending the R Commander Interface to Survival Analysis

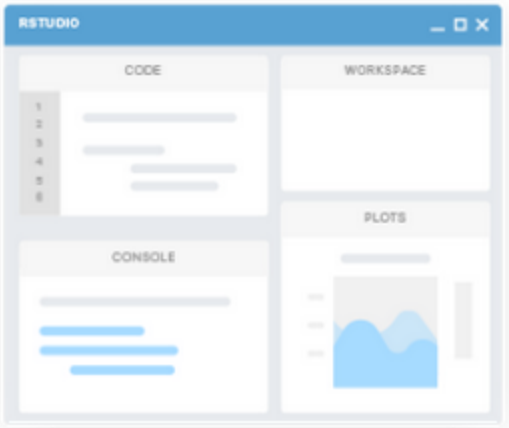
John Fox
McMaster University

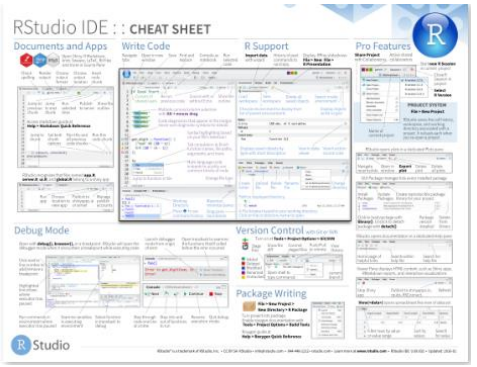
Marilia Sá Carvalho
Escola Nacional de Saúde Pública

Kingston University London

R Studio

Want to make life easier?
Try R Studio, the most popular integrated development environment (IDE).





[Rstudio Cheet Sheet]

Kingston University London

Welcome to Mehmet Tevfik DORAK's Website

Epidemiology
Biostatistics & Statistical Literacy
Statistics & Graphics with R
Critical Skills Critical Thinking
Leadership & Equality Gender Equality
Essentials of Genetics
Genome Biology for Genetic Epidemiologists
Immunogenetics
Evolutionary Biology Notes
Music Notes

Mehmet Tevfik DORAK, BA (Hons), MD, PhD

http://www.dorak.info


Good Luck