SPE-2024: postmortem

Next year's SPE? Tartu is no more the cultural capital of Europe in 2025. What about 2nd last week of May, 2025, starting on Thursday 22 ending Tuesday 27? Combination with the planned Lexis Symposium? – Krista explores.

General experiences from the course

Great heterogeneity again in the participants' preknowledge and skills in R. What to do in order not to bore those who already are familiar with basics in R?

Exercises of the first day in particular – again – need rethinking.

General points for improvement (some already in pm of 2019!)

A lot of time in the beginning of the first practical was spent to introduce the github environment of the practicals; which was new to the faculty, too. This must be taken into account next year.

Creation of metapackage that would load all the packages that are needed?

Better explanation and motivation for using certain wrap-up functions, which are our favourites, such as `ci.lin()` vs. `confint()`. Motivation for function `I()` should be given, too.

Recruit external reviewers (who?) for the exercises, especially those of the first two days. Other exercises good to be reviewed among us.

Simplify the code as much as possible. Don't use negative indices. Develop a wrapper for `cut()` having defaults that are conventional in epidemiology.

Continue encouraging the use of explicit naming of arguments. Emphasize this also in the exercises – modify them as necessary. Avoid using excessively dense code which would be readable only to seasoned R user.

Change all tre remaining `T` → `TRUE` and `F` → `FALSE`, whenever these logical constants are needed.

Remember to have some motivational lines in the beginning of lectures and practicals; why we think these things are important.

Put all `library()` statements needed in a practical to the beginning.

Use `round()` to cut unnecessary decimal points – or put `options(digits=3)`, or often even `digits=2`. Encourage students to do this as well. – Bendix develops a `print` method for e.g. `ci.lin()` nd `ci.exp()`. See e.g. function `publish` in package `publish`, which provides nice printing e.g. of estimated hazard ratios from Cox models.

Continue to instruct, how to efficiently write, check and run the R code. Follow a proper style guide e.g. `https://google.github.io/styleguide/Rguide.html`

Continue to further develop the idea of having enough explanation about what the next lines of cryptic looking script are supposed to do for you.

Choice of colours: Try to develop the exercises including graphics to comply with recent recommendations on colours. Avoid at least bright red and green in the same plot. Consult package `colorspace` and the papers by Zeileis et al. (2009) *Computational Statistics & Data Analysis*, 53, 3259-3270. and Zeileis et al. (2020) in *Journal of Statistical Software*. See also `colorBlindness` package vignette. – Martyn shares his notes on graphics.

Recap presentations remain more visible when using black on white (old default of R console and script windows) rather than white on black format. The font size should be 20 pt.

Day 1

9:15–10:00 Lecture on history and ecology of R (Martyn) was OK. Even though 45 min was allocated to it, Martyn could do it in 30 min. Should this be taken into account in the program next year?

Lecture on `dplyr/tidyverse` (Damien): Not given in 2024. In 2023 it was concise and informative. 15 min enough?

Use of pipe all the time in practicals?

Exercises 1.1 to 1.3 were not done in 2024. The comments below are from postmortem 2023.

Exercises 1.1: Use R first interactively but soon turn to using a script. In section 1.1.10 reduce the number of variants for selecting variables and observations from a data frame. For the latter, recommend just `subset()`. At least sections 1.1.8 on creating data frames could be moved towards the end?

Exercise 1.2: It continues to be important. We should make sure that students have time to do this. Many students did not get to Ex 1.2 at all. This means that Ex 1.1 must be still shorter, or optional items moved after current 1.2. – Martyn will take care of these two exercises in future, too?

Advanced participants: Exercise 1.3 on `dplyr` and `tidyverse` tools for those already familiar with basic R: Those people ($n \approx 7$) who did this exercise mostly could do everything. – Damien prepares.

10:30–11.30: Exercise 1.4 on tabulation: One hour was a bit too short for covering both 1.4 and 1.5. Another problem: The TV screen at IARC was too small. – Krista continues be in charge of 1.4 & 1.5?

Exercise 1.5 on graphics quite good, too, but the time allocated to this together with 1.4 was again dramatically shorter than last time; too short. The program must be further planned and organized such that people have time to do the essential items also in 1.5. – Krista works with this? A lot of details in graphics will be illustrated in later exercises, but then it will be good to explain there, what is being done.

Combine and integrate better 1.4 & 1.5 into one single section 1.4?

11:30-12:00 Lecture on Poisson & binary regression (Janne): Improved from last year! Perhaps there is still too much R code on some slides. Add a slide with the same population and constant rate but contrast cumulative risk e.g. by 5 y vs. by 1 y of follow-up? Could both the single rate item and that comparing two rates be based on the same data like those from Thorotrast?

12:00-13:00 Exercise 1.6 on Poisson rates and binary proportions: Should the numbers be based on more real-looking data, like Thorotrast? Perhaps some items and technicalities could still be dropped. As to simple analysis of proportions, one item on `twoby2()` would be good, and another one illustrating the use of all three links: identity, log and logit with binomial for RD, RR and OR, respectively. (steal from Esa's practical, section 1.7.8 of the afternoon). Drop the other modelling tasks for `lowbw`. – Janne?

14:30-15:15: Lecture on linear and generalized linear models (Esa): Motivation to use `I()`! Think about the order of this lecture and Janne's on simple Poisson and binomial!

15:15-15:45: Lecture on splines (Martyn): A good lecture. Smooth transition from Esa's lecture was a nice change from last version.

16:15-18:00: Exercise 1.7 on estimation of effects: Shortened from last year and `effx()` was dropped out. Some motivation and explanations could be added on some items. – Esa improves.

Exercise 1.8 on estimation and reporting curved effects: The code was simplified from last year. – Esa develops further.

Day 2

9:30–10:30 Lecture on causal inference (Krista): Much improved from last time. The concrete examples from Estonian Biobank in the beginning provided nice motivation. p. 11; the symbol for independence could be opened in words. p. 15; would a causal graph be useful in showing the effect of randomization? p. 16; could the symbol for marginal non-independence be opened in words? Would "conditional exchangeability" here be written "conditional independence"? p. 18-20; there is no definition or illustration of the concept of path in a DAG. p. 19; Could the notation in the terminology box be changed to X:s with subscripts, as they are in the diagram? – Same thing on p. 20? p. 21; Could the sentence "Also called as backdoor path between X and Y" be changed to "This path from X to Y via Z is called as backdoor path ...". p. 24; Could "Here the paths collide at Z" be changed: "Here the arrows collide ..."? p. 28; Could the exposure have letter X and outcome Y instead of C and D? Comment also that with linear structural models for continuous Y many things are easier (such that IV estimation), but with other types of outcome complications arise (e.g. with direct and indirect effects).

11:00-13:00 Exercise 1.9 on causal effects: Again improved from last year. Item 1.9.2; Could the exposure have letter X and outcome Y instead of C and D (see above)? Item 1.9.3; A nice exercise but perhaps some more introduction and explanation could be given in the beginning, so that people would better catch the idea. – Krista.

14:30–15:30 Lecture on graphics (Martyn): As good as always. What about saying something on `ggplot()` on the slides?

16:00-18:00 Exercise 1.10: Can't be improved?!

Day 3

09:30–10:30 Lecture on survival (Janne): Improved from last time. p. 10-11: Should we finally give up from recommending to save `Surv(time, event)` as a separate object within the original data frame (like `sobj` on slide 10 etc.), if it in some further applications will cause trouble? p. 17; Use `\href{}` when referring to Andersen & Keiding 2012. p. 22; Colour codes are not explicitly opened. Could use solid lines for both sexes because colours are used to distinguish the sexes. p. 25; Could perhaps add a brief note on hazards of hazard ratio, i.e. problems with HR when trying to interpret it as causal parameter and refer to the later lecture on causal inference 2. p. 29; Is this slide on SMR needed?

11:00-13:00 Exercise 1.11 on survival methods: Item 1.11.2: The specification of the new colour palette is a nice novelty. Item 1.11.3: Should we get rid of `suob`? Change `T` and `F` in the script finally into `TRUE` and `FALSE`. – Confidence intervals for cumulative risks. – Janne.

14:30–15:30 Lecture on representation of follow-up, Lexis & SMR (Bendix): Generally OK. Change colouring (especially red & green) into more colour-blind friendly! See the

solution in Janne's lecture and practical. When starting to use Thorotrast data, give some introduction to the background and content of it also mentioning that it was already used by Janne especially some aspects of the simple analyses were already present in Janne's lecture slides. As to SMR analysis, a reference to the notion of causal effect of the exposure among those exposed could be made. SMR is estimating such a hazard ratio. Comment the issue of incomparability of two different SMRs, because they are based on different standard populations.

16:00-18:00 Exercise 1.12: Refer to the specific slides in the lecture handouts when asking to compute and tabulate summary measures like `D, Y, rate, E, SMR`. Mark SMR modeling as optional. – Bendix

Day 4

09:30–10:30 Lecture on NCC & CC (Krista): At the end, when comparing NCC and CC designs, perhaps could add a comment that with heavy censoring case-cohort design may become less efficient than NCC.

11:00-13:00 Exercise 1.13 on NCC & CC: `cal.yr()` works for data frames, so simplify the code accordingly.

14:30–15:30 Second lecture on causal inference (Esa): Heavy stuff. Notation is complicated; could this be simplified? It was good to leave time-to-event outcomes out. Should talk more about confidence intervals and how to get them?

16:00-18:00 Exercise 1.14 on causal inference 2: Could the motivation to the exercise be expressed more clearly? Computation of confidence intervals should perhaps be given more attention? Analysis of real data. Check an example from Hernan's homepage for the book H & R. Instead of clever covariates have a tabular summary of different approaches with varying accuracy vs. misspecification of the exposure and/outcome model.

Day 5

09:30–10:30 Lecture on multistate models (Bendix): Colour coding! Try to get rid of red and green. Is it possible to shorten the lecture to give more time for the practical? The number of slides is still 49, which is a lot for an hour.

11:00-12:15 Exercise 1.16 on multistate model: This is heavy stuff considering that it is on the last day, when everybody are already quite tired. Drop Cox models? (comment from 2019). Reduce the still placewise somewhat exaggerated number of decimal points when printing the results. Motivation to using `Ns`? Avoid red and green in the graphs; change into a more colour-blind friendly palette. Give more time to the practical. Also, can the practical be somewhat shortened and simplified? It would be interesting to know, how many participants could still follow the recap.

Could the whole business with multi-state models moved to an earlier phase of the programme, and have a less complicated topic on the last day?