Statistics Workshops 2017

Leibniz ScienceCampus "Primate Cognition"

primate cognition
Leibniz-WissenschaftsCampus

January 19, 2017

Contents

General Information 2

1 Basic Mathematics and Statistics 3

2 Introduction to R 4

3 Regression Modeling 5

4 Bayesian Statistics 6

5 Hierarchical Regression Models 7

General Information

For 2017, we revised the statistics workshops program and concept, to meet the needs of the members of the Leibniz ScienceCampus:

- Each course will run for more days, by simultaneously reducing the daily duration to four hours, giving participants the chance to digest the contents more easily.

- We aim to increase the number of practical applications presented in the workshops. Each workshop will have one additional day – a couple of days after the last day of the workshop – for participants to ask questions and work on tutorials. The participation in this additional practice/training day is optional.

Participation in the beginners' workshop *Basic Mathematics and Statistics* is not a formal prerequisite, but knowledge of its content is highly recommended. This is particularly true, as the mathematical notations are a high hurdle in understanding and learning statistical concepts, especially to understand more advanced statistical techniques. We therefore suggest that any participant interested in the courses *Regression Modeling*, *Bayesian Statistics*, and *Hierarchical Regression Models*, should check the contents of the *Basic Mathematics and Statistics* workshop. If most of the contents are unknown, we recommend to take part in the beginners' workshop.

The workshops build upon each other and assume that the contents of the other courses are known. Therefore, for those interested in all courses, we recommend participating in the following order: *Basic Mathematics and Statistics*, *Regression Modeling*, *Bayesian Statistics*, and *Hierarchical Regression Models*. Nevertheless, relevant key concepts from previous workshops are repeated and partial overlap between courses exists.

On the following pages, we provide a brief summary of the workshops, including organizational issues and a short description of the contents. These descriptions are subject to change and may be that might be updated throughout the year.

Registration procedure: For each workshop, registration via email is possible during a four week period. The available places are then allocated to members of the Science-Campus on a first come, first served basis. Potential remaining places will be allocated to non-members of the ScienceCampus after the registration window has expired. Open places due to cancellations will be distributed through a waiting list. Therefore, we strongly recommend signing up even if all places are currently taken. Expressing interest in the courses furthermore helps to assess the demand for future courses.

1 Basic Mathematics and Statistics

What is this workshop about? In this workshop, we will talk about fundamentals in mathematics and statistics that are mandatory in order to understand more complex statistical methods. In the first part of this workshop, we will focus on mathematical notation and concepts, and talk about basic ideas and general goals of statistics during the second part. Note that this is really a basic course that aims to provide a basic knowledge to anybody who has only a sparse training in mathematics and statistics, or who is just interested in hearing the basics again.

When? February 27 – March 1, 2017, with an additional practice/question day on March 3, 2017. 9.00h – 13.00h each day.

Where? German Primate Center (DPZ), seminar room E0.14.

What is the target group? PhD students, post-docs.

When can I register? January 16 – February 12, 2017.

How can I register? Please send an email to: hsennhenn-reulen(at)dpz.eu

What is the maximum number of participants? 15 (minimum number: 5).

What are the contents I can expect?

- **Basic Mathematics:** Function, (Co-)domain, Inverse function, and parameters (Absolute value function; Exponential function and logarithmic function; Sums; Indicator function; Products); Derivatives and integrals.

- **General Statistics:** Random variables (Discrete and continuous random variables; Probability; Cumulative distribution function and quantile function; Probability mass and density functions; important distributions); Expected value, variance, standard deviation, standard error, and covariance; Statistical inference (Point estimation; Statistical tests; Interval estimation).

2 Introduction to R

What is this workshop about? In this workshop, you will learn the basics and fundamentals of R in order to write R code that implements what you need for your data analyses. Our experience shows that statistics and R are each already quite demanding in isolation, and therefore the didactic strategy is to keep these two topics separated: R is a statistical programming language, and this workshop will consequentially also touch statistical concepts, but the main focus will be to learn how to work with R and understand the underlying work-flow.

When? March 27 – 29, 2017, with an additional practice/question day on March 31, 2017. 9.00h – 13.00h each day.

Where? German Primate Center (DPZ), seminar room E0.14.

What is the target group? PhD students, post-docs.

When can I register? February 13 – March 12, 2017.

How can I register? Please send an email to: hsennhenn-reulen(at)dpz.eu

What is the maximum number of participants? 15 (minimum number: 5).

What are the contents I can expect?

- **Basics:** Brief introduction to functions; Special characters; Mathematics operators; Mathematical functions; Logical comparisons; Your computer is a machine!
- **Workflow:** Tidy data-sets; Organize working session and workspace vocabulary; Coding style; Names.
- **Data Structures:** Vectors atomic vectors; Lists; Attributes; Names; Factors; Matrices and arrays; Data frames; Subsetting; Applications; Exercises.
- **Functions in Detail:** Why? A motivation to write functions; The three parts of a function; Scoping; Environments; Special function types; Applications; Exercises.
- **Programming:** `if{}`, `if{}else{}`, `for` and `while`-loops; Functionals: `apply`, `lapply`, `sapply`, ...; Exercises.

3 Regression Modeling

What is this workshop about? Regression models are the (!) statistical work-horse in empirical research, and this workshop aims to provide a foundation to this class of statistical approaches. While most projects in the ScienceCampus require the application of more complex model classes in order to adequately treat the underlying data generating processes – that is, more advanced model classes that go beyond the basic regression model classes introduced during this workshop – these basics are nevertheless of greatest importance: they directly transfer to the more complex model classes, and thus are essential to successfully perform any applied data analysis by the use of regression techniques.

When? May 8 – 12, 2017, with an additional practice/question day on May 16, 2017. 9.00h – 13.00h each day.

Where? German Primate Center (DPZ), seminar room E0.14.

What is the target group? PhD students, post-docs.

When can I register? March 27 – April 23, 2017.

How can I register? Please send an email to: hsennhenn-reulen(at)dpz.eu

What is the maximum number of participants? 15 (minimum number: 5).

What are the contents I can expect?

- **Simple Linear Models:** Basic model build-up for the simple linear regression model; Estimation of the regression parameters; Confidence intervals for regression parameters; Confidence interval for the regression line; Prediction intervals; Goodness of fit; Similarities and differences between the simple linear regression model and the concept of correlation.

- **Multiple Linear Models:** Modeling of different covariate scales (Non-linear effects; Categorical covariates; Interaction effects); Variable selection; Combined and overall F Tests; Diagnostics.

- **Generalized Linear Models:** Binomial distributed response; Poisson distributed response; Deviance; Offset.

4 Bayesian Statistics

What is this workshop about? As Andrew Gelman writes in his handy statistical lexicon (`http://andrewgelman.com/2008/10/03/bayes_bayesians/`):

> *Every statistician uses Bayesian inference when it is appropriate (that is, when there is a clear probability model for the sampling of parameters). A Bayesian statistician is someone who will use Bayesian inference for all problems, even when it is inappropriate. I am a Bayesian statistician myself (for the usual reason that, even when inappropriate, Bayesian methods seem to work well).*

Bayesian statistics is a powerful school for statistical inference, because it provides very natural solutions to problems that are very difficult to resolve in the classical framework of frequentist statistics (see the explanations in the description to the *Hierarchical Regression Models* workshop). Moreover, with the revolution that came with Markov chain Monte Carlo (MCMC) techniques, post-estimation calculations are practically much more easily feasible, which makes Bayesian inference very appealing in many applied scenarios where the classical framework is actually equally appropriate.

When? September 4 – 8, 2017, with an additional practice/question day on September 12, 2017. 9.00h – 13.00h each day.

Where? German Primate Center (DPZ), seminar room E0.14.

What is the target group? PhD students, post-docs.

When can I register? July 24 – August 20, 2017.

How can I register? Please send an email to: hsennhenn-reulen(at)dpz.eu

What is the maximum number of participants? 15 (minimum number: 5).

What are the contents I can expect?

- **What is Statistical Inference?** Statistical inference and models; Four general tasks in statistical inference; Two general approaches: Frequentist (exclusively likelihood-based), and Bayesian inference.
- **Bayesian and Frequentist Perspectives on Statistical Inference**
- **Bayesian Inference** Posterior distribution; Bayesian point estimates; Credible regions; Bayesian tests; Choice of the prior distribution; Numerical methods for Bayesian inference.
- **Bayesian Inference using MCMC Sampling** Metropolis-Hastings-algorithm and Gibbs-sampler.
- **Bayesian Regression** Linear regression; Logit regression.
- **Bayesian Model Choice**

5 Hierarchical Regression Models

What is this workshop about? Hierarchical regression models are particularly suitable for research designs in which data are organized in more than one observation level (which is the case in many research designs within the ScienceCampus): The primary observation units are usually individuals who are nested within higher order units, such as groups, or when repeated measurements of individuals are examined.

When? November 6 – 10, 2017, with an additional practice/question day on November 14, 2017. 9.00h – 13.00h each day.

Where? German Primate Center (DPZ), seminar room E0.14.

What is the target group? PhD students, post-docs.

When can I register? September 25 – October 22, 2017.

How can I register? Please send an email to: hsennhenn-reulen(at)dpz.eu

What is the maximum number of participants? 15 (minimum number: 5).

What are the contents I can expect? A hierarchical regression model does not only include model terms that explain variation in the (expectation of) the response by products of covariates x_k and regression coefficients β_k (this is supposed to describe the data generating mechanism across observation units, often denoted as fixed model terms), but also coefficients γ_i that explain variation between the observation units $i \in \{1, \ldots, n\}$ (this is often denoted as a random term of the model). Other terminologies for hierarchical regression models are mixed effects regression model (fixed and random model terms), or multilevel regression model (referring to primary observation units being nested within higher order units). During this workshop, we will lay the emphasis on the Bayesian approach to this class of regression models. This is why this workshop is called Hierarchical Regression instead of Mixed Effects Regression Models: by the possibility to naturally incorporate the assumption that $\gamma_i \sim \mathrm{N}\left(0, \sigma_\gamma^2\right)$ into a Bayesian framework, there is no need anymore to distinct between fixed and random model terms.

But why should I learn how to use a Bayesian inference approach to hierarchical regression models? In the classical statistical inference framework, mixed model regression parameters do not have nice asymptotic distributions to test against (this is in contrast to ordinary least squares and generalized linear models parameters, which asymptotically converge to known distributions), which complicates the inferences that can be made from mixed models in this classical framework. The main source of this added complexity is a shrinkage factor that is applied to the random effects by the usual assumption $\gamma_i \sim \mathrm{N}\left(0, \sigma_\gamma^2\right)$, leading to complications in the determination of degrees of freedom associated with this model term. In an applied example, the variance parameter σ_γ^2 may be estimated from n levels of a variable, and a design matrix used to estimate the parameters of this variable incorporates n indicator variables for these n levels. If we would include this variable in the usual way (taking it as a fixed coefficients variable), we would associate one degree of freedom with one estimated value, and so we would usually associate n degrees of freedom with these n indicators. But since these n indicators have a shrinkage factor applied to them (this results in the so-called *partial pooling*), we do not really need n degrees of freedom. So what would be the correct degrees of freedom to use for the cost to estimate this random effects model term? Is it one (we

only estimate one variance parameter), or n (we explain variation in the expectation of the response by the use of n coefficients), or something in between (partial pooling)? The latter option must be correct, but, unfortunately, there is no generally accepted theory that can provide us with an exact value to answer this question. Moreover, assuming we can find a good value for the degrees of freedom, we still can not count on our test statistic (from likelihood ratio tests and the like) to be F or χ^2 distributed, now that we added this shrinkage part to the model. However, if we now move on towards a Bayesian framework in order to estimate this model, we see that the shrinkage is just a very natural consequence of the model assumptions – here seen as prior formulations. This is a major benefit that comes with the use of a Bayesian approach, and it can be assumed that Bayesian approaches to multilevel/hierarchical/mixed models will become the standard in the next years to come, moreover since recent great improvements in software solutions made this approach much easier applicable now (see `STAN` based `R` add-ons `rstanarm` and `brms`).