Outline of the Bayesian Statistics Course at the Faculty of Computer and Information Science, University of Ljubljana

Jure Demšar and Erik Štrumbelj

Course summary

Statistics is a pillar of modern science as it allows us to reason in the presence of uncertainty. We will start this course by going through some basics about how scientific studies should be prepared and executed. During this we will learn why statistics play such an important part in science and our everyday lives. The main goal of this course is for you to learn about the importance of statistics and how to execute modern state-of-the-art Bayesian statistics. To get there, we will learn how to do probabilistic programming and use it to build and apply Bayesian statistical models. With the gained knowledge you should be able to execute your own statistical analyses or provide support to other researchers and professionals.

Grading

The final grade is composed of two parts, the first 50% comes from the homework and the other 50% from an oral exam. We expect there will be around 5 ($\pm$ 1) homework assignments. Unless otherwise noted, all homework will be due in two weeks and should be submitted as a short pdf report (a page or two, depending on the homework). The submitted file should be named following this pattern: `surname_name_HW_number.pdf`. Therefore, Janez Novak should name the file `Novak_Janez_HW_4.pdf` when submitting homework 4. General feedback to your homework will be provided on the group level via the course forum. If some of you will want more detailed, individual feedback about your work, we will happily provide it on request. Each homework will be graded on a scale from 0 to 10:

- 10 – perfect or nearly perfect submissions,
- 9 – submissions of an above average quality,
- 8 – average submissions,
- 7 – submissions of below average quality,
- 6 – barely acceptable (try to do better in the long run),
- $\leq$ 5 – unacceptable.

To finish the course you need to take an oral exam at the end of the semester. Only students that gathered at least 50% of points from the homework are allowed to take the exam. The oral exam will be used to test your knowledge both in the theoretical and in the practical aspects of Bayesian statistics that we covered during lectures.

Reports

Even though the reports you are submitting are short, they should still follow conventions of good scientific and technical writing. A good practice is to follow the IMRAD format (https://en.wikipedia.org/wiki/IMRAD,):

- **Introduction** – In the introduction part of your report you should establish why you are doing what you are doing. In other words, you should answer what is the motivation behind the task you are about to take on and what you are trying to answer with your work? As you will see, all of our homework will be practically oriented, so you should be able to provide these kind of answers.
- **Methods** – Describe the methods, approaches and algorithms you used for tackling the task at hand. Describe them in a transparent and easy to understand manner so others can reproduce your work.
- **Results** – Present what you found out in a clear and concise fashion. If possible, use graphical visualizations. Good visualizations convey your story in a way that is more reader friendly than blocks of text or tables. Some libraries (e.g. ggplot in R) are designed in a way that will make your results compliant with modern standards of data visualization. With other libraries some additional work is usually required to get visualizations that look nice and also convey the information efficiently.

- **Discussion** – Wrap up your work by summarizing your results and explicitly answering the research questions you tackled. Outline any limitations and possible improvements to your work.

Note that the four sections (Introduction, Methods, Results and Discussion) do not need to be explicitly outlined and denoted, the important thing is that the flow of your reports follows the logic outlined above. For example, you can easily merge Methods and Results in a single section if this fits your project better.

Lectures

Below is the approximate outline of core lectures in this year's edition of the course. Note here that what is listed below is only an outline of lectures and will probably change slightly during the course of the year. The list contains only lectures executed by course organizers, we will probably also have some guest lectures which are not included in the list.

Introduction to Statistical Enquiry

The main goal of this lecture is to provide an introduction to the process of statistical enquiry. What is statistical enquiry, why is it important, and how do we approach it in a systematic way? We base our introduction on Wild and Pfannkuch's 4 dimensions of statistical enquiry, with special focus on the 1st dimension, the PPDAC investigative cycle (Problem, Plan, Data, Analysis, Conclusions).

Scientific Methodology and Probabilistic Thinking

The main goals of this lecture are to introduce you to good practices of scientific methodology and to illustrate how uncertainty is part of our everyday lives but in order to deal with it in a systematic way we require the rigor of probability theory. The language of probability theory allows us not only to express uncertainty but also to provide probabilistic interpretations of processes of interest, which then allows us to infer their properties from the data they generate. That is, probability theory is the very foundation of applied statistics.

Probabilistic Programming

Statistical (probabilistic) models are used for describing how certain data we are interested in was generated. Probabilistic programming languages offer both a methodological way for specifying statistical models and tools for performing automated inferences on these models.

Stan is probably the most widespread probabilistic programming language. It couples well with several popular programing languages and offers an intuitive framework for specifying statistical models along with algorithms for full Bayesian inference for continuous variable models through Markov chain Monte Carlo methods such as the No-U-Turn sampler, an adaptive form of Hamiltonian Monte Carlo sampling.

The Generalized Linear Model

The intercept and slope parameters in simple linear regression are easy to interpret and can as such provide key insight into our data generating process. When modeling the relationship between independent and dependent variables, simple linear regression often gives suboptimal results. We need a more powerful tool, especially so in cases where the dependent variable is not metric. Generalized linear model (GLM) is a powerful tool capable of working with dependent variables of various scale types (metric, ordinal, nominal and count) and error distributions other than normal. The simple linear model is actually the simplest of all GLMs. In the GLM, beta coefficients are usually not as easy to interpret as in the case of simple linear regression. Fortunately, we only need some relatively easy math to make them understandable!

Priors

Defining priors is an integral part in Bayesian statistics. The main purpose of priors is to introduce prior expert knowledge about the domain into our models. As we know by now, this is not the only usage of priors, they are also useful for regularization and can be of help during the sampling process which can lead to stabilization of inferences in certain models. Based on the amount of information priors provide they are commonly categorized into three groups: non-informative priors, weakly informative priors and informative priors. This document briefly explains the differences between these groups and provides some guidance about when to use certain priors. Since one of the main advantages of Bayesian statistics is its ability to facilitate existing knowledge to empower our models the second part of this document talks about approaches for eliciting relevant information from experts.

Fit Checking

Bayesian modeling is an iterative process, we start by picking an initial model and settings its priors. To investigate whether our priors have undesirable effects that contradict domain knowledge we start by performing prior checks. Once we are happy with our priors, we fit the model and diagnose the fitting process (traceplot, convergence, diagnostics ...). If all is good, we then execute posterior checks to evaluate whether our model is suitable for answering the questions we are asking.

Describe the Process

When developing statistical models we often fall into the habit of merely fitting some distributions onto our data. In this lecture we will show why this is bad an suboptimal. We should do our best to try and understand the actual data generating process and model it!

Bayesian Cross-Validation

Cross-validation is a widely spread technique for estimating how well our models generalize (how they fare when encountering unknown data). Since in Bayesian statistics we are

always working with probability distributions we can use log-score as our go-to model evaluation measure. This allows us to resort to information theory for calculating good cross-validation approximations without actually performing the actual (often very time consuming) cross-validation. An important point to emphasize here is that cross-validation and regularization are not mutually exclusive, they go hand in hand and traditionally we use both at the same time.

Hierarchical Models

In Bayesian modeling hierarchical models (also called multilevel models) are an extremely powerful tool. As already emphasized a couple of times now, when modeling we should try to describe the data generating process (and not fit some distributions to some data). Since data generating processes often have a hierarchical structure (e.g. groups of students, multiple repetitions of an experiment ...), hierarchical models enable us to efficiently describe such data generating processes.

Advanced GLMs

Data can often have more intricate structure and connections than what can be captured by simple GLMs or simple hierarchical models. In this lecture we will take a look at some more advanced GLMs that can help us model such data.

Questionnaires

Measuring things accurately and precisely is difficult to start with, but even more so when we try to measure peoples' psychological characteristics, opinions, preferences, etc. We will discuss the questionnaire as a measuring device. How do we design, test, and validate one. How do we select the type of question and scale. And what are the most common mistakes.

Sampling

This lecture will revolve around the basics of sampling in the context of survey sampling. We will cover three of the most common probability sampling approaches: simple random sampling, stratified sampling, and cluster sampling. We will also briefly discuss non-probability sampling methods: convenience sampling, judgment sampling, quota sampling, and snowball sampling.

Modelling time-series

We are often faced with data that is time sensitive in the form of time-series. Since the time component is usually of paramount importance in such data we have to be extra careful when handling and modelling our data. For this purpose, we will take a look at specialized Bayesian models for modelling time sensitive data.

Why skepticism is important in modern science and data analysis

We are constantly bombarded with miss information labeled as facts, be it in our everyday lives or in scientific literature. In this, more lightweight, lecture we will take a look on how to spot such dubious practices and thus keep them from spreading and succeeding.

Lab practice

The lab practice slots will be used as consultations for your homework. If you have any questions, if you are stuck, if you want to use Bayesian statsitics for some of your work outside of the course, lab practice is the place to go.