Data Analysis, Monte Carlo Methods & R- An Introduction

Shobhit Mahajan
shobhit.mahajan@gmail.com

Version 5.1
Last modified on October 27, 2020

Contents

A Note to the Reader

This set of notes are intended to serve as an introduction to Statistical Methods, Data Analysis and Monte Carlo Methods as well as an introduction to **R**. The initial motivation for this came from my students Nisha Rani and Akshay Rana who were using these techniques for their research. Since I was not very familiar with the techniques, I decided to learn them. This set of notes are a result of my own learning the subject.

Subsequently, another student, Sukhdeep Singh started working on Artificial Neural Networks and their use in cosmology. Once again, I was not familiar with the field and so decided that I should try to learn it with Sukhdeep's help. This version of the Notes includes an Introduction to some of the elementary concepts in Artificial Neural Networks and some applications. This field is very vast and there are many areas which we have not explored since this is meant only to be an introduction.

The Manual is organised into 11 Chapters. It is intended to be self-contained and so does not assume any prior knowledge of statistics or probability. **Chapter** 1 is intended to serve as an introduction to basic concepts in probability and statistics. This includes the concept of random variables, distributions, laws of probability etc. **Chapter** 2 focuses on statistical inference by using a Bayesian Approach. We discuss the Maximum Likelihood estimates as well as the use of Bayes' theorem to determine posterior probabilities given a prior. **Chapter** 3 is a discussion of an alternative approach to Statistical Inference by using the more common frequentist approach. The concept of Null Hypothesis testing as well as p-values are discussed in this. After this introduction to probability and statistics, we discuss some concepts in Error Analysis and Curve Fitting in **Chapter** 4. An important application of statistics is in Sampling to determine information about populations. **Chapter 5** discusses some very basic concepts of sampling theory. We then move on to discuss random numbers in detail in **Chapter** 6. Methods for generation of random numbers from various distributions are discussed in this. **Chapter** 7 is an introdution of Monte Carlo methods as well as their use in evaluating integrals, especially multi dimensional integrals. One of the major uses of Monte Carlo methods are in simulations of experiments. This is discussed in **Chapter** 8. We then introduce Markov Chains in **Chapter** 9 and discuss the use of Markov Chain Monte Carlo in **Chapter** 10**.** Artificial Neural Networks are introduced in **Chapter** 11.

Every chapter has many solved examples. In most of the examples, one needs computation tools. For this purpose, most programs are given in 3 languages- **C, Python & R.** The reason for this is that many readers might not yet be familiar with Python and R and so would find it easier to understand the logic by studying the C program. On the other hand, Python is fast becoming the language of

choice for most purposes. This is because there are many open source libraries available on the Internet which can be incorporated into our programs. Finally, R was until recently the language in which most data science was done and it continues to be very useful for statistical purposes. The reader can decide which language she is most comfortable with and then use that. **The programs are not written in the most efficient manner- instead, they are written in a way which might be easiest to understand and connect with the theoretical discussion about the topic. The reader is encouraged to program herself in a more efficient and succinct manner.**

In addition, there are 3 Appendices. The first is an introduction to using GNUPLOT for the purposes of visualising data. This assumes that you are using a Linux platform or a Linux emulator (like CYGWIN32/CYGWIN64) on a Windows machine. The second Appendix is an introduction to the use of the powerful tool MS Excel. This software is very versatile and can be used not only for Data Analysis and Visualisation but also for Statistical Analysis. Most of the functionalities of MS Excel would be available in Open Excel though I have used MS Excel and so the material is based on it.

There is also an appendix to provide a Quick Introduction to R. This introduction is by no means complete and is only meant to get the reader started and get familiar with the basic structure of R and its use in most circumstances that one would need. R is the language of choice nowadays for data analysis and statistics since it provides a lot of inbuilt routines and libraries. In addition, it is Open Source and there are many libraries available on the Internet which one can download and install to use. Throughout the Manual, we have used R where ever possible in the Examples so that the reader can get familiar with it. It is advisable to first go through the Appendix on R before one attempts to use it.

The notes are not intended to give the reader a detailed theoretical understanding of the concepts behind the topics. **Instead, it is meant to be a primer which we hope will equip the reader to become familiar enough with these important tools to be used in her work. Furthermore, we hope that this introduction will encourage the curious reader to delve more deeply into some of the topics discussed here.** With this background, detailed proofs of theorems are mostly avoided. Instead, we focus on discussing a lot of examples to illustrate the power of these tools.

We would very much like to get your suggestions regarding how to improve this Manual. In addition, if there are any errors or misprints that are spotted in the Manual, we would like to hear from you. Please send a mail with the suggestions/errors etc. to **shobhit.mahajan@gmail.com** making sure you quote the version number of the Manual as well as the Modification date of the Manual you are using. The version number and date are on the title page of the Manual.