

Data Analysis, Monte Carlo Methods & R- An Introduction

Shobhit Mahajan
shobhit.mahajan@gmail.com

**ONLY FOR INTERNAL CIRCULATION. NOT TO BE SHARED OR
CIRCULATED WITHOUT PERMISSION.**

Version 5.1

Last modified on October 27, 2020

Contents

Note to the Reader	4
	4
1 Introduction to Probability & Statistics	7
1.1 Some Elementary Concepts	7
1.2 Conditional Probability	15
1.3 Law of Total Probability	17
1.4 Independent Events	20
1.5 Bayes' Theorem	21
1.6 Random Variables	25
1.7 Discrete Distributions	27
1.7.1 Bernoulli Distribution	28
1.7.2 Binomial distribution	28
1.7.3 Geometric Distribution	32
1.7.4 Uniform Distribution	33
1.7.5 Poisson Distribution	34
1.8 Mean or Expectation Values	37
1.9 Variance & Standard Deviation	46
1.10 Continuous Random Variables	50
1.10.1 Uniform Distribution	51
1.10.2 Exponential Distribution	52
1.10.3 Normal Distribution	52
1.10.4 Expectation value	55
1.10.5 Variance	56
1.11 Joint Probability Distributions	59
1.12 Covariance & Correlation	60
2 Statistical Inference- Bayesian Approach	67
2.1 Introduction	67
2.2 Maximum Likelihood Estimate (MLE)	68
2.3 Bayes' Theorem & Posterior Probability	76
2.3.1 Discrete Variables	76
2.3.2 Odds	82
2.3.3 Continuous Hypothesis & Discrete Data	87
2.3.4 Beta distribution	94
2.3.5 Continuous hypothesis & Continuous Data	100
3 Statistical Inference- Frequentist Approach	102
3.1 Null Hypothesis Significance Testing (NHST)	102
3.2 p-values	111
3.3 Student t distribution	118
3.4 Chi Squared Test for Goodness of Fit	124

4	Error Analysis & Fitting	132
4.1	Types of Errors	132
4.2	Uncertainty in Measurement	133
4.2.1	Uncertainty, Accuracy & Precision	133
4.2.2	Systematic & Random Errors	135
4.3	Error Estimation	136
4.3.1	Propagation of Errors	138
4.4	Estimation and Error of the Mean	156
4.4.1	Method of Maximum Likelihood	156
4.4.2	Estimated Error in the Mean	158
4.5	Method of Least Squares	161
5	An Introduction to Sampling	165
5.1	Some Basic Concepts	165
5.2	A review of some concepts	166
5.2.1	Chebyshev's Inequality	167
5.2.2	Measurement Errors	168
5.3	Random & Independent Sampling	168
5.4	Sampling Distributions	170
5.4.1	Z or Standard Normal Distribution	171
5.4.2	t distribution	174
5.4.3	Use of t distribution vs Normal distribution	176
5.4.4	Chi Square Distribution	176
5.5	Model Building & Testing	177
5.6	Testing the Difference between Means	186
5.7	Sampling Distribution of Proportions	189
5.8	Testing of Difference of Proportions	193
5.9	Types of Errors	197
5.10	Estimation	199
5.11	Connection with Bayesian Statistics	205
5.12	Sample Size	207
6	Random Variables & Numbers	214
6.1	Introduction	214
6.2	Random Variables	215
6.2.1	Discrete Random Variables	215
6.2.2	Continuous Random Variables	218
6.2.3	Sums of Independent Random Variables	221
6.3	Random Numbers	226
6.3.1	von Neumann Rejection Method	229
6.3.2	Inverse Transformation Method	236
6.3.3	Analytical Rejection Method	238
6.4	Distributions	239
6.4.1	Exponential Distribution	239
6.4.2	Poisson Distribution	240
6.4.3	Gaussian Distribution	246
7	Integration using Monte Carlo Methods	256
8	Monte Carlo Simulations	285
9	Markov Chains	341
9.1	Kinds of Markov Chains	347
10	Markov Chain Monte Carlo	351
10.1	Metropolis Algorithm	352

10.2	Metropolis-Hastings Algorithm	377
10.3	Multivariate Distributions	397
10.4	Metropolis Algorithm for multivariate distributions	398
10.5	Gibbs Sampling	410
11	Artificial Neural Networks- A Primer	425
11.1	Introduction	425
11.2	Perceptrons	426
11.2.1	Comparison with a NAND gate	428
11.3	Learning Networks	432
11.4	Sigmoid Neuron	432
11.5	Structure of Neural Networks	435
11.6	Training a Neural Network	441
11.6.1	Gradient Descent	442
11.6.2	Backpropagation Algorithm	443
11.7	Use of ANN in Regression	452
11.8	Use of ANN in Classification	463
11.8.1	Binary Classification	467
11.9	Using R for Artificial Neural Networks	499
11.9.1	Classification Problems	500
11.9.2	Regression	506
11.9.3	Cross-validation	510
11.10	Improvements in Learning	526
11.10.1	Learning Slowdown	526
11.10.2	Softmax	529
11.10.3	ReLU	532
11.11	Regularization	533
11.12	An introduction to some Advanced Topics	542
11.12.1	Convolutional Networks	543
11.12.2	Recurrent Neural Networks (RNNs)	547
11.12.3	Bayesian Networks	552
A	Using Gnuplot	562
A.1	Introduction	562
A.2	Plotting with inbuilt functions of GNUPLOT	563
A.2.1	Interactive plotting	563
A.3	Saving Plots	565
A.3.1	Customization	566
A.4	Plotting using data from a file	567
A.5	Plotting using data from file and fitting to a smooth curve	569
A.5.1	Curve Fitting & Interpolation	572
B	Using MS Excel	576
B.1	Simple calculations	577
B.2	Plotting Data	582
B.2.1	Error Bars	583
B.2.2	Formatting Graphs & Plots	585
B.3	Fitting Data	585
B.4	Statistical Analysis	589
C	A Quick Introduction to R	597
C.1	Introduction	597
C.2	Installing R	597
C.3	Getting Started	598
C.4	Inputting Data from Files	601
C.5	Writing Data into Files	603

C.6	R Decision Making & Loops	604
C.7	Functions in R	607
	C.7.1 In-built functions	607
	C.7.2 Integration	607
	C.7.3 User Defined Functions	610
C.8	Programming in R	611
C.9	Plotting in R	613
	C.9.1 3D Graphics	618
C.10	Matrices	619
	C.10.1 Constructing a Matrix	619
	C.10.2 Basic Matrix operations	621
	C.10.3 Eigenvalues & Eigenvectors, Inverse of a Matrix	623
C.11	Basic Statistics	627
	C.11.1 Mean & Variance	627
	C.11.2 Distributions	628
	C.11.3 Confidence Limits	637
	C.11.4 Histograms	640
	C.11.5 Probability of Type II Error	642
C.12	Markov Chain Monte Carlo	645
	C.12.1 Metropolis Algorithm	645
	C.12.2 Metropolis-Hastings Algorithm	649
	C.12.3 Sampling from Multivariate Distributions	652
	C.12.4 Gibbs Sampling	654
C.13	Regression	660
	C.13.1 Linear Regression	660
	C.13.2 Curve Fitting	664

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 introduction 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.