

# 1

## *Introduction*

The advent of a rigorous framework of studying probability and statistics dates back to the eighth century AD and is documented in the works of Al-Khalil who was an Arab philologist.<sup>1</sup> This branch of mathematics continues to be under development with major contributions from Soviet mathematician Andrey N. Kolmogorov who developed the modern foundations of probability and statistical theory from a measure theoretic standpoint in the twentieth century.<sup>2</sup>

### *1.1 Why should you study probability and statistics?*

Both theoretical as well as applied aspects of probability and statistics are subjects of broad intellectual and practical interest to engineers and scientists alike. This genre of mathematics is widely practised in diverse areas of engineering and sciences ranging from artificial intelligence, bio-informatics, wireless communications, aerodynamics, operations research, financial engineering, modern physical and chemical sciences to the social and humanitarian sciences, demography, weather forecasting, data analytics and so on. Over the past century, on the one hand, the scientific community has gradually come to accept the limitations of scientific measurements using precision instruments courtesy the theory of quantum mechanics, and to be precise, due to the unpredictability stemming from the consequences of the uncertainty principle; and on the other hand, engineers and scientists have been confronted by an explosion of information that has increased the dimensions of variability in data. These challenges have expanded the scope and dependence on probabilistic and statistical techniques to solve real world problems.<sup>3</sup>

In light of the above, there is a necessity and opportunity to train students in the practical aspects of probability and statistics so that they can readily apply their knowledge to find solutions to engineering and scientific problems. There is an urgent need to introduce the fundamentals and applications of this subject in a less intimidating manner by staying away from over reliance on theoretical and analytical derivations. An alternate approach which allows students to assimilate foundational concepts through projects and real data is often more helpful. This experiential mode of learning not only familiarises students with applications but also keeps them engrossed in the learning process by making it fun and free of abstraction.

<sup>1</sup> Lyle D. Broemeling. "An Account of Early Statistical Inference in Arab Cryptology". In: *The American Statistician* 65.4 (2011), pp. 255–257. DOI: DOI:10.1198/TAS.2011.10191

<sup>2</sup> A.N. Kolmogorov. *Foundations of the Theory of Probability*. Second. Chelsea Publishing Co., 1956. ISBN: 0828400237

<sup>3</sup> Uncertainty breeds risk and thereby rewards!

## 1.2 About the book

This book covers broad areas of Probability and Statistics starting with introductory preliminaries, probability distributions, introduction to stochastic processes through discrete and continuous time Markov chains, regression analysis, sampling distributions and their applications in statistical inference and hypothesis testing (including ANOVA), calculation of covariance functions, notion of *stationarity* and *ergodicity* of data, introduction to time series analysis of data (moving average and auto-regressive models), and basic introduction to multi-variate statistics and their application through principal component analysis. Almost all topics discussed in the book are presented using a practical (project based) as well as the theoretical framework. The book will be useful for advanced UG level students, beginning PG level students and practitioners from the engineering, basic and social sciences. The content of this book can be covered during a 4-5 month long semester. The computational/simulation instructions presented in this book are based on Matlab<sup>4</sup> tutorials and projects. The book has a nice appendix on how to get started with Matlab in twenty pages.

<sup>4</sup> <https://matlab.mathworks.com>

Each chapter has a project cum case-study<sup>5</sup> laced together with the concepts that is introduced in a systematic manner. Chapters begin with a brief introduction to the case-study (project) and the tools that are necessary to solve the problem. The concepts are then developed in a gradual manner and their relevance to the project are established systematically. In this way, as the reader sifts through the various sections of each chapter, he/she not only learns the concepts but also gets an opportunity to apply the knowledge towards solving the case-study project. For example, the chapter on “Statistical Experiments” has a case study on ‘post disaster reconstruction management’. In this project, students learn about some of the main issues that impede post disaster reconstruction of infrastructure by identifying the main bottlenecks like lack of availability of public funds, shortage of technical manpower to rebuild infrastructure, lack of community participation, etc. A detailed ANOVA hypothesis test is conducted for each of the aforementioned bottleneck areas across six different cities. The tests are based on numerical ratings registered by project managers overlooking the reconstruction projects in each city. This statistical analysis will enable the students to identify the bottleneck issues plaguing each city for which corrective measures are solicited from the authorities. Similar projects are designed for other topics throughout the book. These built-in projects will also enable practitioners to design their own projects and statistical experiments by referring to the design of similar experiments and case studies presented in this book. Another important feature of the presentation of the material in this book is the use of several illustrations of mathematical concepts, figures, pictures, additional references, and footnotes along the right margins. This feature is aimed at providing a visual imagery of the concepts being discussed in the corresponding section of the book. Such visual imagery often brings the more abstract mathematical formalism to life and enhances the learning experience of the reader.

<sup>5</sup> All project related sections in each chapter is typeset in **bold** and printed in **purple**.

The book has evolved through several rounds of lectures delivered to undergraduate and postgraduate students from diverse backgrounds both in the United States and in India. The students will find it useful to use as a textbook. Instructors of probability and statistics will find it very useful to consider this book to cater to both theoretical as well as laboratory teaching. The book is self sufficient for the most part. This means students with very elementary idea of probability and statistics (standard XI-XII level) can use this book easily. The only pre-requisites are elementary calculus (eg. single variable calculus - differentiation,

integration, series and sequences) and high school level familiarity with matrices. The book will serve as an invitation to students to further explore the fascinating area of probability and statistics through more advanced courses.

### 1.3 *A note on the choice of chapter topics*

An important unique feature of the book may become clear upon a closer inspection of the table of contents and the chronological order in which they have been presented. Typically, classic texts on applied probability and statistics for engineering and science students do not cover topics such as Markov chains, time series analysis, and principal component analysis at the introductory level. Instead they delve into statistical estimation theory, non-parametric tests or more theoretical topics from probability theory. However, I feel that there is a tremendous opportunity to introduce students to applied topics like Markov chains (discrete and continuous) immediately after the chapter on conditional probability. This has two advantages:

1. the area of Markov chains has wide applications in a variety of engineering and scientific fields ranging from queuing theory, weather modelling to financial engineering, natural language processing and AI, and engineering physics. This way we can get a grip over the students attention and affection for the subject at a very early stage of the learning process, and
2. two dedicated chapters on Markov chains also retains the alignment with the intended target audience of the book: engineering and science students and/or industry practitioners, who are perhaps more interested in the mathematical modelling aspect of the subject than the estimation theoretic approaches more suitable for pure statisticians.

Likewise, there is an opportunity to introduce techniques to analyse data through time series models and principal component analysis in an introductory text of this nature so that we are able to prepare students interested in pursuing data science as majors. In this way, all along, the need to cater to the target audience has been supreme in my mind while writing this textbook.

### 1.4 *Acknowledgements*

The book, in its present form, could not have been possible without the unrelenting support of my professional and personal family and friends. I am grateful to the university community for providing me with the right blend of creative and compassionate environment that is conducive to intellectual endeavours such as this one. I want to thank my wife Sasha for her patience and diligent proof reading of the material. Finally, I would like to dedicate this book to my five year old daughter Diya who never shies away from impressing upon me, through all her adventures, the power of imagination and the force of volition that is necessary to create a beautiful and a better future for humanity.

Amayo  
Island of Peace, 2021.