1

Why R?

Package(s): `UsingR`
Dataset(s): `+AD1-9`

1.1 Why R?

Welcome to the world of Statistical Computing! During the first quartile of the previous century Statistics started growing at a great speed under the schools led by Sir R.A. Fisher and Karl Pearson. Statistical computing replicated similar growth during the last quartile of that century. The first part laid the foundations and the second part made the founders proud of their work. Interestingly, the beginning of this century is also witnessing a mini revolution of its own. The R Statistical Software, developed and maintained by the R Core Team, may be considered as a powerful tool for the statistical community. The software being a Free Open Source Software is simply icing on the cake.

R is evolving as the preferred companion of the Statistician. The reasons are aplenty. To begin with, this software has been developed by a team of Statisticians. Ross Ihaka and Robert Gentleman laid the basic framework for R, and later a group was formed who are responsible for the current growth and state of it. R is a command-line software and thus powerful with a lot of options for the user.

The legendary Prasanta Chandra Mahalanobis delivered one of the important essays in the annals of Statistics, namely, "Why Statistics?" It appears that Indian mathematicians were skeptical to the thought of including Statistics as a legitimate branch of science in general, and mathematics in particular. This essay addresses some of those concerns and establishes the scientific reasoning through the concepts of random samples, importance of random sampling, etc.

Naturally, we ask ourselves the question "Why R?" Of course, the magnitude of the question is oriented in a completely different and (probably) insignificant way, and we hope the reader will excuse us for this idiosyncrasy. The most important reason for the choice of R is that it is an open source software. This translates to the fact that the functioning of the software can be understood to the first line of code which steam rolls into powerful utilities. As an example, we can trace how exactly the important `mean` function works.

A Course in Statistics with R, First Edition. Prabhanjan Narayanachar Tattar, Suresh Ramaiah and B. G. Manjunath.

Companion Website: www.wiley.com/go/tattar/statistics

```
#  File src/library/base/R/mean.R
#  Part of the R package, http://www.R-project.org
#
#  A copy of the GNU General Public License is available at
#  http://www.r-project.org/Licenses/
mean <- function(x, ...) UseMethod("mean")
mean.default <- function(x, trim = 0, na.rm = FALSE, ...)
{
    if(!is.numeric(x) && !is.complex(x) && !is.logical(x)) {
        warning("argument is not numeric or logical: returning NA")
        return(NA_real_)
    }
    if (na.rm)
 x <- x[!is.na(x)]
    if(!is.numeric(trim) || length(trim) != 1)
        stop("'trim' must be numeric of length one")
    n <- length(x)
    if(trim > 0 && n > 0) {
 if(is.complex(x))
     stop("trimmed means are not defined for complex data")
 if(trim >= 0.5) return(stats::median(x, na.rm=FALSE))
 lo <- floor(n*trim)+1
 hi <- n+1-lo
 x <- sort.int(x, partial=unique(c(lo, hi)))[lo:hi]
    }
    .Internal(mean(x))
}
mean.data.frame <- function(x, ...) sapply(x, mean, ...)
```

Note that there is information about the address of the mean function, src/library/base/R/mean.R. The user can go to that address and open `mean.R` in any text editor. Now, if you find that the mean function does not work according to your requirement, modifications and new functions can be defined easily. For instance the default setting of the mean function is `na.rm=FALSE`, that is, if there are missing observations in a vector, see Section 2.3, the `mean` function will return `NA` as the answer. It is very simple to define a modified function whose default setting is `na.rm=TRUE`.

```
> x <- c(10,11,NA,13,14)
> mean(x)
[1] NA
> mean_new <- function(...,na.rm=TRUE) mean(...,na.rm=TRUE)
> mean_new(x)
[1] 12
> mean(x,na.rm=TRUE)
[1] 12
```

This is as simple as that. Thus, there are no restrictions imposed by the software on the user. The authors strongly believe that this freedom is priceless. If the decision to acquire the software is dictated by economic considerations, it is convenient that R comes freely.

Computation complexity is a reason for the need of software. As the modern statistical methods are embedded with complexity, it becomes a challenge for the developers of the methodology to complement the applications with appropriate computer programs. It has been our observation that many statisticians tend to address this dimension with relevant R packages. Venables and Ripley (2002) developed a very useful package `MASS`, an abbreviation for the title of their book *Modern Applied Statistics with S*. This package is shipped along with the software and is "recommended" as a priority package. In Section 1.8 we will see how many statisticians have adopted R as the language of their statistical computations.

1.2 R Installation

The website http://cran.r-project.org/ consists of all versions of R available for a variety of Operating Systems. CRAN is an abbreviation for Comprehensive R Archive Network. An incidental fact is that R had been developed on the Internet only.

The R software can be installed on a variety of platforms such as Linux, Windows, and Macintosh, among others. There is also an option of choosing 32- or 64-bit versions of the software. For a Linuxian, under appropriate privileges, R may be easily installed from the terminal using the command `sudo apt-get install r-base`. Ubuntu operating system users can find more help regarding R installation at the link http://ubuntuforums.org/showthread.php?t=639710.

After the installation is complete, the user can start the software by simply keying in R at the terminal. If the user is a beginner and not too familiar with the Linux environments, it is a possibility that she may be disappointed with its appearance as she cannot find much help there. Furthermore, the Linux expert may find this too trivial to explain/help a beginner. Some help for the beginner is available at http://freshmeat.net/articles/view/2237/.

A user of Windows first needs to download the recent versions executable file, currently R-3.0.2-win32.exe, and then merely double-click her way to completing the installation process. Similarly, Macintosh users can easily find the related files and methods for installation. The web links "R MacOS X FAQ" and "R Windows FAQ" should further be useful to the reader. The authors have developed the R codes used in this book and verified them for Linux and Windows versions. We are confident that they will compile without errors on Macintosh too.

1.3 There is Nothing such as PRACTICALS

The reader is absolutely free to differ from our point of view that "There is nothing such as PRACTICALS" and may skip this section altogether. There are two points of view from the authors which will be put forward here. First, with the decreasing cost of computers and availability of Open Source Software, OSS, see Appendix A, there is no need for calculator-based practicals. Also within the purview of a computer lab, a Statistics student/expertise needs to be more familiar with software such as R and SAS among others. Our second point of view is that the integration of theory with applications can be seamlessly achieved using the software modules.

It is apparently clear with the exponential growth of technology that the days of separate sessions for practicals of are a bygone era, and it's not an intelligent proposition to hang onto

a weak rope, and blame it for our fall. It has been observed that in many of the developed Departments of the subject, calculator-based computations/practicals session have been done away with altogether. It is also noticed that many Statistical institutes do not teach C++/Fortran programming languages even at a graduate course, and a reason for this may be that statisticians need not necessarily be software programmers. There are many additional reasons for this reluctance. A practical reason is that computers have become very much cheaper, and if not within the financial reach of the students (especially in the developing countries), computing machines are easily available in most of their institutes. It is more often the case that the student has access to at least a couple of hours per week at her institute.

The availability of subject-specific interpretative software has also minimized the need of writing explicit programs for most of the standard practical methods in that subject. For example, in our Statistics subject, there are many software packages such as SAS, SYSTAT, STATISTICA, etc. Each of these contains inbuilt modules/menus which enable the user to perform most of these standard computations in a jiffy, and as such the user need not develop the programs for the statistical techniques in the applied area such as Linear Regression Analysis, Multivariate Statistics, among other topics of the subject.

It is true that one of the driving themes of this book is to convey as many ideas and concepts, both theoretical and practical, through a mixture of software programs and mathematical rigor. This aspect will become clear as the reader goes deeper into the book and especially through the asterisked sections or subsections. In short, this book provides a blend of theory and applications.

1.4 Datasets in R and Internet

The `R` software consists of many datasets and more often than not each package, see Section 2.6 for more details about an R package, contains many datasets. The command `try(data(package= "\,"))` enlists all the datasets contained in that package. For example, if we need to find the datasets in the package, say `rpart` and `methods`, execute the following:

```
> try(data(package="rpart"))
car.test.frame           Automobile Data from 'Consumer
+ Reports' 1990
car90                    Automobile Data from 'Consumer
+ Reports' 1990
cu.summary               Automobile Data from 'Consumer
+ Reports' 1990
kyphosis                 Data on Children who have had Corrective
+ Spinal Surgery
solder                   Soldering of Components on Printed-Circuit
+ Boards
stagec                   Stage C Prostate Cancer
> try(data(package="methods"))
no data sets found
```

The function for loading these datasets will be given in the next chapter. It has been observed that authors of many books have created packages containing all the datasets from their book and released them for the benefit of the programmers. For example, Faraway (2002) and Everitt and Hothorn (2006) have created packages titled `faraway` and `HSAUR2` respectively, which may be easily downloaded from http://cran.r-project.org/web/packages/, see Section 2.6.

Another major reason for a student to familiarize herself with a software is that practical settings rarely have small datasets ($n < 100$, to be precise). It is a good exposition to deal with industrial datasets. Thus, we feel that the beginners must try their hand at as many datasets as they can. With this purpose in mind, we enlist in the next subsection a bunch of websites which contain large numbers of datasets. This era really requires the statistician to shy away from ordinary calculators and embrace realistic problems.

1.4.1 List of Web-sites containing DATASETS

Practical datasets are available aplenty on the worldwide web. For example, Professors A.P. Gore, S.A. Paranjape, and M.B. Kulkarni of the Department of Statistics, Poona University, India, have painstakingly collected 103 datasets for their book titled "100 Datasets for Statistics Education", and have made it available on the web. Most of these datasets are in the realm of real-life problems in the Indian context. The datasets are available in the `gpk` package. We will place much emphasis on the datasets from this package and use them appropriately in the context of this current book, and also thank them on behalf of the readers too.

Similarly, the website http://lib.stat.cmu.edu/datasets/ contains a large host of datasets. Especially, datasets that appear in many popular books have been compiled and hosted for the benefit of the netizens.

It is impossible for anybody to give an exhaustive list of all the websites containing datasets, and such an effort may not be fruitful. We have listed in the following what may be useful to a statistician. The list is not in any particular order of priorities.

- http://ces.iisc.ernet.in/hpg/nvjoshi/statspunedatabook/databook.html
- http://lib.stat.cmu.edu/data sets/
- http://onlinelibrary.wiley.com/journal/10.1111/%28ISSN%291467-985X/homepage/datasets_all_series.htm
- http://www.commondata set.org/
- https://datamarket.com/data/list/?q=provider:tsdl
- http://inforumweb.umd.edu/econdata/econdata.html
- http://www.ucsd.edu/portal/site/Libraries/
- http://www.amstat.org/publications/jse/information.html
- http://www.statsci.org/data sets.html
- http://archive.ics.uci.edu/ml/data sets.html
- http://www.sigkdd.org/kddcup/index.php

We are positive that this list will benefit the user and encourage them to find more such sites according to their requirements.

1.4.2 Antique Datasets

Datasets available on the web are without any doubt very valuable and useful for a learner as well as the expert. Apart from the complexity and dimensionality, the sources are updated regularly and thus we are almost guaranteed great data sources. In the beginning of statistical development though, such a luxury was not available and the data collection mechanism was severely restricted by costs and storage restrictions. In spite of such limitations, the experimenters really compensated for them by their foresight and innovation. We describe in the rest of this section a set of very useful and antique datasets. We will abbreviate "Antique Datasets" as "AD". All the datasets discussed here are available in the books associated with the `ACSWR` package.

Example 1.4.1. AD1. Galileo's Experiments. The famous scientist Galileo Galilei conducted this experiment four centuries ago. An end of a ramp is elevated to a certain height with the other end touching the floor. A ball is released from a set height on the ramp and allowed to roll down a long narrow channel set within the ramp. The release height and the distance traveled before landing are measured. The goal of the experiment is to understand the word should be split like this: relationship between the release height and distance traveled. Dickey and Arnold's (1995) paper reignited interest in the Galileo dataset in the statistical community. This paper is available online at http://www.amstat.org/publications/jse/v3n1/datasets.dickey.html#drake. □

Example 1.4.2. AD2. Fisher's Iris Dataset. Fisher illustrated the multivariate statistical technique of the *linear discriminant analysis* method through this dataset. It is important to note here that though there are only three species with four measurements of each observation, and 150 observations, this dataset is very much relevant today. Rao (1973) used this dataset for the hypothesis testing problem of equality of two vector means. Despite the availability of large datasets, the `iris` dataset is a benchmark example for the *machine learning community*. This dataset is available in the `datasets` package. □

Example 1.4.3. AD3. The Militiamen's Chest Dataset. Militia means an army composed of ordinary citizens and not of professional soldiers. This dataset is available in an 1846 book published by the Belgian statistician Adolphe Quetelet, and the data is believed to have been collected some 30 years before that. It would be interesting to know the distribution of the chest measurements of a militia which had 5738 militia men. Velleman and Hoaglin (1984), page 259, has more information about this data. We record here that though the dataset is not available, the summaries of frequency count is available, which serves our purpose in this book. □

Example 1.4.4. AD4. The Sleep Dataset – 107 Years of Student's *t*-Distribution. The statistical analysis of this dataset first appeared in the 1908 remarkable paper of William Gosset. The paper titled *The Probable Error of Mean* had been published in the *Biometrika* journal under the pen name Student. The purpose of the investigation had been identification of an effective soporific drug among two groups for more sleep. The experiment had been conducted on ten patients from each group and since the large sample Z-test cannot be applied here, Gosset solved the problem and provided the small-sample t-test which also led to the well-known Student's t-distribution. The default R package `datasets` contains this dataset. □

Example 1.4.5. AD5. The Galton's Dataset. Francis Galton is credited with the invention of the linear regression model and it is his careful observation of the phenomenon of *regression toward the mean* which forms the crux of most of regression analysis. This dataset is available in the `UsingR` package of Verzani (2005) as the `galton` dataset. It is also available in the companion `RSADBE` package of Tattar (2013). The dataset contains 928 pairs of height of `parent` and `child`. The average height of the parent is 68.31 inches, while that of the child is 68.09 inches. Furthermore, the correlation coefficient between the height of parent and child is 0.46. We will use this dataset in the rest of this book. □

Example 1.4.6. AD6. The Michelson-Morley Experiment for Detection of Ether. In the nineteenth century, a conjectured theory for the propagation of light was the existence of an ether medium. Michelson conducted a beautiful experiment in the year of 1881 in which the drift caused by ether on light was expected to be at 4%. What followed later, in collaboration with Morley, was one of the most famous *failed experiments* in that the setup ended by proving the non-existence of ether. We will use this dataset on multiple occasions in this book. In the `datasets` package, this data is available under `morley`, whereas another copy is available in the `MASS` package as `michelson`. □

Example 1.4.7. AD7. Boeing 720 Jet Plane Air Conditioning Systems. The time between failures of air conditioning systems in Boeing jet planes have been recorded. Here, the event of failure is recurring for a single plane. Additional information is available regarding the air conditioning undergoing a major overhaul during certain failures. This data has been popularized by Frank Proschan. This dataset is available in the `boot` package by the data frame `aircondit`. □

Example 1.4.8. AD8. US Air Passengers Dataset. Box and Jenkins (1976) used this dataset in their classic book on time series. The monthly totals of international airline passengers has been recorded for the period 1949–1960. This data consists of interesting patterns such as seasonal variation, yearly increment, etc. The performance of various time series models is compared and contrasted with respect to this dataset. The `ts` object `AirPassengers` from the `datasets` package contains the US air passengers dataset. □

Example 1.4.9. AD9. Youden and Beale's Data on Lesions of Half-Leaves of the Tobacco Plant. A simple and innovative design is often priceless. Youden and Beale (1934) sought to find the effect of two preparations of virus on tobacco plants. One half of a tobacco leaf was rubbed with cheesecloth soaked in one preparation of the virus extract and the second half was rubbed with the other virus extract. This experiment was replicated on just eight leaves, and the number of lesions on each half leaf was recorded. We will illustrate later if the small sample size is enough to deduce some inference. □

1.5 http://cran.r-project.org

We mentioned CRAN in Section 2. The worldwide web link of CRAN is the title of this Section. A lot of information about R and many other related utilities of the software are available from this web source. The "R FAQ" web page contains a lot of common queries and helps the beginner to fix many of the initial problems.

"Manuals", "FAQs", and "Contributed" links on this website contains a wealth of information on documentation of the software. A journal called "The R Journal" is available at http://journal.r-project.org/, with the founders on the editorial board, who will help to keep track of developments in R.

1.5.1 http://r-project.org

This is the main website of the R software. The reader can keep track of the continuous stream of textbooks, monographs, etc., which use R as the computational vehicle and have been published in the recent past by checking on the link "Books". It needs to be mentioned here that this list is not comprehensive and there are many more books available in print.

1.5.2 http://www.cran.r-project.org/web/views/

The interest of a user may be in a particular area of Statistics. This web-link lists major areas of the subject and further directions to detailed available methods for such areas. Some of the major areas include Bayesian Inference, Probability Distributions, Design of Experiments, Machine Learning, Multivariate Statistics, Robust Statistical Methods, Spatial Analysis, Survival Analysis, and Time Series Analysis. Under each of the related links, we can find information about the problems which have been addressed in the R software. Information is also available on which additional package contains the related functions, etc.

As an example, we explain the link http://www.cran.r-project.org/web/views/Multivariate.html, which details the R package's availability for the broader area of multivariate statistics. This unit is maintained by Prof Paul Hewson. The main areas and methods in this page have been classified as (i) Visualizing Multivariate Data, (ii) Hypothesis Testing, (iii) Multivariate Distributions, (iv) Linear Models, (v) Projection Methods, (vi) Principal Coordinates/Scaling Methods, (vii) Unsupervised Classification, (viii) Supervised Classification and Discriminant Analysis, (ix) Correspondence Analysis, (x) Forward Search, (xi) Missing Data, (xii) Latent Variable Approaches, (xiii) Modeling Non-Gaussian Data, (xiv) Matrix Manipulations, and (xv) Miscellaneous utilities. Under each of the headings there will be a mention of the associated packages which will help in related computations and implementations.

In general, all the related web-pages end with a list of related "CRAN Packages" and "Related Links". Similarly, the url http://www.cran.r-project.org/web/packages/ lists all add-on packages available for download. As of April 10, 2015, the total number of packages was 6505.

1.5.3 Is subscribing to R-Mailing List useful?

Samuel Johnson long ago declared that "There are two types of knowledge. One is knowing a thing. The other is knowing where to find it." Subscribing to this list is the knowledge of the second type. We next explain how to join this club. As a first step, copy and paste the link www.r-project.org/mail.html into your web-browser. Next, find "web interface" and click on it, following which you will reach https://stat.ethz.ch/mailman/listinfo/r-announce. On this web-page, go to the section "Subscribing to R-announce". We believe that once you check the

URL http://www.r-project.org/contributors.html, you will not have any doubts regarding why we are pursuing you to join it.

1.6 R and its Interface with other Software

R has many strengths of its own, and is also true about many other software packages, statistics software or otherwise. However, it does happen that despite the best efforts and the intent to be as complete as possible, software packages have their limitations. The great Dennis Ritchie, for instance, had simply forgotten to include the power function when he developed one of the best languages in C. The reader should appreciate that if a software does not have some features, it is not necessarily a drawback. The missing features of a software may be available in some other package or it may not be as important as first perceived by the user. It then becomes useful if we have bridges across to the culturally different islands, with each of them rich in its own sense. Such bridges may be called *interfaces* in the software industry.

The interfaces also help the user in many other ways. A Bayesian who is well versed in the *Bayesian Inference Using Gibbs Samples* (BUGS) software may be interested in comparing some of the Bayesian models with their counterparts in the frequentist school. The BUGS software may not include many of the frequentist methods. However, if there is a mechanism to call, and frequentist methods of software such as R, SAS, SYSTAT, etc. are required, a great convenience is available for the user.

The bridge called interface is also useful in a different way. A statistician may have been working with BUGS software for many years, and now needs to use R. In such a scenario, if she requires some functions of BUGS, and if those codes can be called up from R and then fed into BUGS to get the desired result, it helps in a long way for the user. For example, a BUGS user can install the `R2WinBUGS` additional package in R and continue to enjoy the derived functions of BUGS. We will say more about such additional packages in the next chapter.

1.7 help and/or ?

Help is indispensable! Let us straightaway get started with the help in R. Suppose we need details of the `t.test` function. A simple way out is to enter `help(t.test)` at the R terminal. This will open up a new page in the R Windows version. The same command when executed in UNIX systems leads to a different screen. The Windows user can simply close the new screen using either "Alt+F4" or by using the mouse. If such a process is replicated in the UNIX system, the entire R session is closed without any saving of the current R session. This is because the screen is opened in the same window. The UNIX user can return to the terminal by pressing the letter `q` at any time. The R code `?t.test` is another way of obtaining the help on `t.test`.

Help on a topic, say `t.test`, can be obtained using `help(t.test)` or `?t.test`

Programming skills and the ability to solve mathematical problems share a common feature. If it is not practiced for even a short period of time, as little as two months after years of experience, it undoes a lot of the razor sharpness and a lot of the program syntax is then forgotten. It may be likely that the expert in *Survival Analysis* has forgotten that the call function

of the famous Cox Proportional Hazards model is `coxph` and not `coxprop`. A course of retrieval is certainly referred to in the related R books. Another way is using the help feature in a different manner `??cox`.

> ??, equivalently "help.search", helps you when ? fails

A search can also be made according to some keyword function, and we can also restrict it to a certain package in light of appropriate information.

```
help.search(keyword = "character", package = "base")
```

In the rest of this book, whenever help files give more information, we provide the related help at the right-hand end of the section in a box. For instance, the help page for the `beta` function is in the main help page `Special` and inquiring for `?beta` actually loads the `Special` help file.

1.8 R Books

Thanks to the user-friendliness of the software, many books are available with an "R-specific" focus. The purpose of this section is to indicate how R has been a useful software in various facets of the subject, although it will not be comprehensive. The first manual that deserves a mention is the notes of Venables and Smith (2014), the first version of which probably came out in 1997. Such is the importance of these notes that it comes with the R software and may be easily assessed. It is very readable and lucid in flow and covers many core R topics. Dalgaard (2002–9) is probably the first exclusive book on the software and it helps the reader to gain a firm footing and confidence in using the software. Crawley's (2007–13) book on R covers many topics and will be very useful on the deck of an R programmer. Purohit, et al. (2008) is a good introductory book and explains the preliminary applications quite well. Zuur, et al. (2009) is another nice book to start learning about the R software.

Dobrow (2013) and Horgan (2008) provide an exposition of probability with the software. Iacus (2008) deals with solving a certain class of "Stochastic Differential Equations" through the R software. Ugarte, et al. (2008) provides a comprehensive treatment of essential mathematical statistics and inference. Albert and Rizzo (2012) is another useful book to familiarize with R and Statistics. A useful reference for Bayesian analysis can be found in Albert (2007–9). It is important to note here that though Nolan and Speed (2000) have not written in the R-text book mold, they have developed very many R programs.

R produces some of the excellent graphics and the related development can be seen in Sarkar (2008), and Murrel (2006).

Freely circulated notes on Regression and ANOVA using R is due to Faraway (2002). Faraway has promptly followed these sets of notes with two books, Faraway (2006) and Faraway (2006). Nonlinear statistical model building in R is illustrated in Ritz and Streibig (2008). Maindonald and Braun (2010) is an early exposition to data analysis methods and graphics. Multivariate data analysis details can be found in Everitt and Hothorn (2011). Categorical data analysis in-depth treatment is found in Bilder and Loughin (2015).

The goal of this section is not to introduce all R books, but to give a glimpse into the various areas in which it can be aptly used. Appropriate references will be found in later chapters.

1.9 A Road Map

The preliminary R introduction is the content of Chapter 2. In this chapter we ensure that the user can do many of the basic and essential computations in R. Simple algebra, trigonometry, reading data in various formats, and other fundamentals are introduced in an incremental phase. Chapter 3 contains enhanced details on manipulation of data, as the data source may not be in a ready-to-use format. Its content will also be very useful to practitioners.

Chapter 4 on *Exploratory Data Analysis* will be the first statistical chapter. This chapter serves as an early level of analyses on the dataset and provides a rich insight. As the natural intent is to obtain an initial insight into the dataset, a lot of graphical techniques are introduced here. It may be noted that most of the graphical methods are suitable for continuous variables and we have introduced a slew of other graphical methods for discrete data in Chapter 16 on Categorical Data Analysis. The first four chapters forms Part I of this book.

The purpose of this book is to complement data analysis with a sound footing in the theoretical aspects of the subject. To proceed in this direction, we begin with *Probability Theory* in Chapter 5. A clear discussion of probability theory is attempted, which begins with set theory and concludes with the important *Central Limit Theorem*. We have enriched this chapter with a clear discussion of the challenging problems in probability, combinatorics, inequalities, and limit theorems. It may be noted that many of the problems and discussions have been demonstrated with figures and R programs.

Probability models and their corresponding distributions are discussed in Chapter 6. Sections 2 to 4 deal with univariate and multivariate probability distributions and also consider discrete and continuous variants. *Sampling Distributions* forms a bridge between probability and statistical inference. Bayesian sampling distributions are also dealt with in this chapter and we are now prepared for inference.

The Estimation, Testing Hypotheses, and Confidence Intervals trilogy is integrated with computations and programs in Chapter 7. The concept of families of distribution is important and the chapter begins with this and explores the role of loss functions as a measure which can be used to access the accuracy of the proposed estimators. The role of sufficient statistics and related topics are discussed, followed by the importance of the likelihood function and construction of the maximum likelihood estimators. The EM algorithm is developed in a step-by-step manner and we believe that our coverage of the EM algorithm is one of the pedagogical ones available in the books. Testing statistical hypotheses is comprehensively developed in Sections 7.9–7.15. The development begins with Type I and II errors of statistical tests and slowly builds up to multiple comparison tests.

Distribution-free statistical inference is carried out in Chapter 8 on *Nonparametric Inference*. The empirical distribution function plays a central role in non-parametrics and is also useful for estimation of statistical functions. Jackknife and bootstrap methods are essentially non-parametric techniques which have gained a lot of traction since the 1980s. Smoothing through the use of kernels is also dealt with, while popular and important non-parametric tests are used for hypotheses problems to conclude the chapter.

The problems of the frequentist school are parallelly conveyed in Chapter 9 titled *Bayesian Inference*. This chapter begins with the idea of Bayesian probabilities and demonstrates how the choice of an appropriate prior is critically important. The posterior distribution gives a unified answer in the Bayesian paradigm for all three problems of estimation, confidence intervals

(known as credible intervals in the Bayesian domain), and hypotheses testing. Examples have been presented for each set of the problems.

Bayesian theory has seen enormous growth in its applications to various fields. A reason for this is that the (complex) posterior distributions were difficult to evaluate before the unprecedented growth in computational power of modern machines. With the advent of modern computational machines, a phenomenal growth has been witnessed in the Bayesian paradigm thanks to the Monte Carlo/Markov Chain methods inclusive of two powerful techniques known as the Metropolis-Hastings algorithm and Gibbs sampler. Part III starts by developing the required underlying theory of Markov Chains in Chapter 10. The Monte Carlo aspects are then treated, developed, and applied in Chapter 11.

Part IV titled "Linear Models" is the lengthiest part of the book. *Linear Regression Models* begins with a simple linear model. The multiple regression model, diagnostics, and model selection, among other topics, are detailed with examples, figures, and programs. *Experimental Designs* have found many applications in agricultural studies and industry too. Chapter 13 discusses the more popular designs, such as completely randomized design, blocked designs, and factorial designs.

Multivariate Statistical Analysis is split into two chapters, 14 and 15. The first of these two chapters forms the *core* aspects of multivariate analysis. Classification, Canonical Correlations, Principal Component Analysis, and Factor Analysis concludes Chapter 15.

If the regressand is a discrete variable, it requires special handling and we describe graphical methods and preliminary methods in Chapter 16 titled *Categorical Data Analysis*. The chapter begins with exploratory techniques useful for dealing with categorical data, and then takes the necessary route to chi-square goodness-of-fit tests. The regression problem for discrete data is handled in Chapter 17. The proceedings of statistical modeling in the final chapter parallels Chapter 12 and further considers probit and Poisson regression models.