Introduction to Probability and Statistics Using R

Second Edition

G. Jay Kerns

February 2, 2013

Contents

IPSUR: Introduction to Probability and Statistics Using R
 ISBN: 978-0-557-24979-4

Date: February 2, 2013

Contents

Preface to the Second Edition

What is new in the Second Edition? Almost everything. I have addressed two of the goals from the first edition. I have now converted most of the plots to `ggplot2` format.

The Second Edition marks a departure from LyX to Emacs Org-Mode. I went with Org-Mode for many reasons. I liked LyX, and LyX is definitely a more user-friendly approach to writing a free book. My workflow, however, has radically changed over the last two years, and I've converted to using Org-Mode for (almost) everything. It truly is "Your Life in Plain Text".

An advantage of the Org-Mode approach is that I can generate an HTML version (that even looks good, still) with a few keystrokes. That means I can post an HTML version of IPSUR, which I've done.

The HTML version is **very** important, for **more** than the following reasons: 1) a person can read IPSUR without need to do anything else, period, 2) automatic full-text indexing by Google, Bing, YaCY, etc., and, most importantly to me, 3) **automatic translation to over 40 languages at the click of a button** (with Google Translate, which comes for free with Google Chrome/Chromium).

Acknowledgements

The success of the Second Edition (if any) would be due in no small part to the successes of the First Edition, so it would be apropos to copy-paste the acknowledgements from the earlier Preface here.

I think, though, that the *failures* of the First Edition have played an important role as well. I would like to extend gracious thanks to Mr. P.J.C. Dwarshuis (Hans), Statistician, from The Hague, Netherlands, and Jésus Juan, who, armed with a sharp eye, have pointed out mistakes, misstatements, and places where better discussion is warranted. It is the selfless contributions of people just like these gentlemen which make the hours spent polishing a FREE book all the more worthwhile.

Preface to the First Edition

This book was expanded from lecture materials I use in a one semester upper-division undergraduate course entitled *Probability and Statistics* at Youngstown State University. Those lecture materials, in turn, were based on notes that I transcribed as a graduate student at Bowling Green State University. The course for which the materials were written is 50-50 Probability and Statistics, and the attendees include mathematics, engineering, and computer science majors (among others). The catalog prerequisites for the course are a full year of calculus.

The book can be subdivided into three basic parts. The first part includes the introductions and elementary *descriptive statistics*; I want the students to be knee-deep in data right out of the gate. The second part is the study of *probability*, which begins at the basics of sets and the equally likely model, journeys past discrete/continuous random variables, and continues through to multivariate distributions. The chapter on sampling distributions paves the way to the third part, which is *inferential statistics*. This last part includes point and interval estimation, hypothesis testing, and finishes with introductions to selected topics in applied statistics.

I usually only have time in one semester to cover a small subset of this book. I cover the material in Chapter 2 in a class period that is supplemented by a take-home assignment for the students. I spend a lot of time on Data Description, Probability, Discrete, and Continuous Distributions. I mention selected facts from Multivariate Distributions in passing, and discuss the meaty parts of Sampling Distributions before moving right along to Estimation (which is another chapter I dwell on considerably). Hypothesis Testing goes faster after all of the previous work, and by that time the end of the semester is in sight. I normally choose one or two final chapters (sometimes three) from the remaining to survey, and regret at the end that I did not have the chance to cover more.

In an attempt to be correct I have included material in this book which I would normally not mention during the course of a standard lecture. For instance, I normally do not highlight the intricacies of measure theory or integrability conditions when speaking to the class. Moreover, I often stray from the matrix approach to multiple linear regression because many of my students have not yet been formally trained in linear algebra. That being said, it is important to me for the students to hold something in their hands which acknowledges the world of mathematics and statistics beyond the classroom, and which may be useful to them for many semesters to come. It also mirrors my own experience as a student.

The vision for this document is a more or less self contained, essentially complete,

correct, introductory textbook. There should be plenty of exercises for the student, with full solutions for some, and no solutions for others (so that the instructor may assign them for grading). By `Sweave`'s dynamic nature it is possible to write randomly generated exercises and I had planned to implement this idea already throughout the book. Alas, there are only 24 hours in a day. Look for more in future editions.

Seasoned readers will be able to detect my origins: *Probability and Statistical Inference* by Hogg and Tanis [**?**], *Statistical Inference* by Casella and Berger [**?**], and *Theory of Point Estimation* and *Testing Statistical Hypotheses* by Lehmann [**?**, **?**]. I highly recommend each of those books to every reader of this one. Some R books with "introductory" in the title that I recommend are *Introductory Statistics with R* by Dalgaard [**?**] and *Using R for Introductory Statistics* by Verzani [**?**]. Surely there are many, many other good introductory books about R, but frankly, I have tried to steer clear of them for the past year or so to avoid any undue influence on my own writing.

I would like to make special mention of two other books: *Introduction to Statistical Thought* by Michael Lavine [**?**] and *Introduction to Probability* by Grinstead and Snell [**?**]. Both of these books are *free* and are what ultimately convinced me to release IPSUR under a free license, too.

Please bear in mind that the title of this book is "Introduction to Probability and Statistics Using R", and not "Introduction to R Using Probability and Statistics", nor even "Introduction to Probability and Statistics and R Using Words". The people at the party are Probability and Statistics; the handshake is R. There are several important topics about R which some individuals will feel are underdeveloped, glossed over, or wantonly omitted. Some will feel the same way about the probabilistic and/or statistical content. Still others will just want to learn R and skip all of the mathematics.

Despite any misgivings: here it is, warts and all. I humbly invite said individuals to take this book, with the GNU Free Documentation License (GNU-FDL) in hand, and make it better. In that spirit there are at least a few ways in my view in which this book could be improved.

Better data. The data analyzed in this book are almost entirely from the `datasets` package in base R, and here is why:

- I made a conscious effort to minimize dependence on contributed packages,
- The data are instantly available, already in the correct format, so we need not take time to manage them, and
- The data are *real*.

I made no attempt to choose data sets that would be interesting to the students; rather, data were chosen for their potential to convey a statistical point. Many of the data sets are decades old or more (for instance, the data used to introduce simple linear regression are the speeds and stopping distances of cars in the 1920's).

In a perfect world with infinite time I would research and contribute recent, *real* data in a context crafted to engage the students in *every* example. One day I hope to stumble over said time. In the meantime, I will add new data sets incrementally as time permits.

More proofs. I would like to include more proofs for the sake of completeness (I understand that some people would not consider more proofs to be improvement). Many proofs have been skipped entirely, and I am not aware of any rhyme or reason to the current omissions. I will add more when I get a chance.

More and better˜graphics. I have not used the `ggplot2` package [**?**] because I do not know how to use it yet. It is on my to-do list.

More and better exercises. There are only a few exercises in the first edition simply because I have not had time to write more. I have toyed with the `exams` package [**?**] and I believe that it is a right way to move forward. As I learn more about what the package can do I would like to incorporate it into later editions of this book.

About This Document

IPSURcontains many interrelated parts: the *Document*, the *Program*, the *Package*, and the *Ancillaries*. In short, the *Document* is what you are reading right now. The *Program* provides an efficient means to modify the Document. The *Package* is an R package that houses the Program and the Document. Finally, the *Ancillaries* are extra materials that reside in the Package and were produced by the Program to supplement use of the Document. We briefly describe each of them in turn.

The Document

The *Document* is that which you are reading right now – IPSUR's *raison d'être*. There are transparent copies (nonproprietary text files) and opaque copies (everything else). See the GNU-FDL in Appendix **??** for more precise language and details.

IPSUR.tex is a transparent copy of the Document to be typeset with a LaTeX distribution such as MikTeX or TeX Live. Any reader is free to modify the Document and release the modified version in accordance with the provisions of the GNU-FDL. Note that this file cannot be used to generate a randomized copy of the Document. Indeed, in its released form it is only capable of typesetting the exact version of IPSURwhich you are currently reading. Furthermore, the `.tex` file is unable to generate any of the ancillary materials.

IPSUR-xxx.eps, IPSUR-xxx.pdf are the image files for every graph in the Document. These are needed when typesetting with LaTeX.

IPSUR.pdf is an opaque copy of the Document. This is the file that instructors would likely want to distribute to students.

IPSUR.dvi is another opaque copy of the Document in a different file format.

The Program

The *Program* includes `IPSUR.lyx` and its nephew `IPSUR.Rnw`; the purpose of each is to give individuals a way to quickly customize the Document for their particular purpose(s).

IPSUR.lyx is the source LyX file for the Program, released under the GNU General Public License (GNU GPL) Version 3. This file is opened, modified, and compiled with LyX, a sophisticated open-source document processor, and may be used (together with `Sweave`) to generate a randomized, modified copy of the Document with brand new data sets for some of the exercises and the solution manuals (in the Second Edition). Additionally, LyX can easily activate/deactivate entire blocks of the document, *e.g.* the *proofs* of the theorems, the student *solutions* to the exercises, or the instructor *answers* to the problems, so that the new author may choose which sections (s)he would like to include in the final Document (again, Second Edition). The `IPSUR.lyx` file is all that a person needs (in addition to a properly configured system – see Appendix **??**) to generate/compile/export to all of the other formats described above and below, which includes the ancillary materials `IPSUR.Rdata` and `IPSUR.R`.

IPSUR.Rnw is another form of the source code for the Program, also released under the GNU GPL Version 3. It was produced by exporting `IPSUR.lyx` into R/Sweave format (`.Rnw`). This file may be processed with Sweave to generate a randomized copy of `IPSUR.tex` – a transparent copy of the Document – together with the ancillary materials `IPSUR.Rdata` and `IPSUR.R`. Please note, however, that `IPSUR.Rnw` is just a simple text file which does not support many of the extra features that LyX offers such as WYSIWYM editing, instantly (de)activating branches of the manuscript, and more.

The Package

There is a contributed package on `CRAN`, called `IPSUR`. The package affords many advantages, one being that it houses the Document in an easy-to-access medium. Indeed, a student can have the Document at his/her fingertips with only three commands:

Another advantage goes hand in hand with the Program's license; since IPSUR is free, the source code must be freely available to anyone that wants it. A package hosted on `CRAN` allows the author to obey the license by default.

A much more important advantage is that the excellent facilities at R-Forge are building and checking the package daily against patched and development versions of the absolute latest pre-release of R. If any problems surface then I will know about it within 24 hours.

And finally, suppose there is some sort of problem. The package structure makes it *incredibly* easy for me to distribute bug-fixes and corrected typographical errors. As an author I can make my corrections, upload them to the repository at R-Forge, and they will be reflected *worldwide* within hours. We aren't in Kansas anymore, Toto.

Ancillary Materials

These are extra materials that accompany IPSUR. They reside in the `/etc` subdirectory of the package source.

IPSUR.RData is a saved image of the R workspace at the completion of the Sweave processing of IPSUR. It can be loaded into memory with `File ▸ Load Workspace` or with the command `load("/path/to/IPSUR.Rdata")`. Either method will make every single object in the file immediately available and in memory. In particular, the data BLANK from Exercise BLANK in Chapter BLANK on page BLANK will be loaded. Type BLANK at the command line (after loading `IPSUR.RData`) to see for yourself.

IPSUR.R is the exported R code from `IPSUR.Rnw`. With this script, literally every R command from the entirety of IPSURcan be resubmitted at the command line.

Notation

We use the notation `x` or `stem.leaf` notation to denote objects, functions, *etc.*. The sequence `Statistics ▸ Summaries ▸ Active Dataset` means to click the `Statistics` menu item, next click the `Summaries` submenu item, and finally click `Active Dataset`.

Acknowledgements

This book would not have been possible without the firm mathematical and statistical foundation provided by the professors at Bowling Green State University, including Drs. Gábor Székely, Craig Zirbel, Arjun K. Gupta, Hanfeng Chen, Truc Nguyen, and James Albert. I would also like to thank Drs. Neal Carothers and Kit Chan.

I would also like to thank my colleagues at Youngstown State University for their support. In particular, I would like to thank Dr. G. Andy Chang for showing me what it means to be a statistician.

I would like to thank Richard Heiberger for his insightful comments and improvements to several points and displays in the manuscript.

Finally, and most importantly, I would like to thank my wife for her patience and understanding while I worked hours, days, months, and years on a *free book*. Looking back, I can't believe I ever got away with it.

List of Figures

List of Tables

1 An Introduction to Probability and Statistics

This chapter has proved to be the hardest to write, by far. The trouble is that there is so much to say – and so many people have already said it so much better than I could. When I get something I like I will release it here.

In the meantime, there is a lot of information already available to a person with an Internet connection. I recommend to start at Wikipedia, which is not a flawless resource but it has the main ideas with links to reputable sources.

In my lectures I usually tell stories about Fisher, Galton, Gauss, Laplace, Quetelet, and the Chevalier de Mere.

1.1 Probability

The common folklore is that probability has been around for millennia but did not gain the attention of mathematicians until approximately 1654 when the Chevalier de Mere had a question regarding the fair division of a game's payoff to the two players, supposing the game had to end prematurely.

1.2 Statistics

Statistics concerns data; their collection, analysis, and interpretation. In this book we distinguish between two types of statistics: descriptive and inferential.

Descriptive statistics concerns the summarization of data. We have a data set and we would like to describe the data set in multiple ways. Usually this entails calculating numbers from the data, called descriptive measures, such as percentages, sums, averages, and so forth.

Inferential statistics does more. There is an inference associated with the data set, a conclusion drawn about the population from which the data originated.

I would like to mention that there are two schools of thought of statistics: frequentist and bayesian. The difference between the schools is related to how the two groups interpret the underlying probability (see Section **??**). The frequentist school gained a lot of ground among statisticians due in large part to the work of Fisher, Neyman, and Pearson in the early twentieth century. That dominance lasted until inexpensive computing power

became widely available; nowadays the bayesian school is garnering more attention and at an increasing rate.

This book is devoted mostly to the frequentist viewpoint because that is how I was trained, with the conspicuous exception of Sections **??** and **??**. I plan to add more bayesian material in later editions of this book.

1.3 Exercises