Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: http://www.tandfonline.com/loi/utch20

Book Reviews

To cite this article: (2018) Book Reviews, Technometrics, 60:1, 124-132, DOI: 10.1080/00401706.2018.1424451

To link to this article: https://doi.org/10.1080/00401706.2018.1424451

Published online: 22 Feb 2018.

TECHNOMETRICS
2018, VOL. 60, NO. 1, 124–132
https://doi.org/10.1080/00401706.2018.1424451

Taylor & Francis
Taylor & Francis Group

BOOK REVIEWS

This section reviews those books whose content and level reflect the general editorial policy of *Technometrics*. Publishers should send books for review to Ejaz Ahmed, Department of Mathematics and Sciences, Brock University, St. Catharines, ON L2S 3A1 (dean.fms@brocku.ca).

The opinions expressed in this section are those of the reviewers. These opinions do not represent positions of the reviewer's organization and may not reflect those of the editors or the sponsoring societies. Listed prices reflect information provided by the publisher and may not be current.

The book purchase programs of the American Society for Quality can provide some of these books at reduced prices for members. For information, contact the American Society for Quality at 1-800-248-1946.

Community Structure of Complex Networks, by Hua-Wei Shen. Berlin, Heidelberg: Springer, 2013, xiv+117 pp., $ 129.00, ISBN: 978-3-642-31820-7.

The book belongs to the series of Springer Theses of the best Ph.D. works selected from around the world for their scientific excellence. In its five chapters the monograph presents the modern approaches to the community structure (CS) of networks (NW), including overlaps among communities, multiscale CS, relationship between CS and NW dynamics, and multiple types of structures beyond CS.

Chapter 1 introduces CS with a saying "the birds of a feather flock together" and indicating that communities are crucial for finding structures in NW, help to NW visualization and compression. Communities of NW can be loosely defined as groups of nodes more closely connected to each other than to the rest of NW. There are various local definitions of community based on distances and counts of nodes, but global definitions are preferable because of their focus on the properties of NW in whole. One popular global definition is the so-called modularity based on comparison of a real NW with simulated randomized reference NWs, another one uses the expected description length of a random walk on NWs, and probabilistic models are used as well. All of them give an insight to the community as a salient structural regularity of NW. The main aim of a NW study consists in community detection that can be achieved by graph clustering or cuts based on hierarchical clustering methods, both agglomerative and divisive. Community detection can be performed using modularity criteria Q and their optimization. One modularity criterion Q has the following interpretation: for a given partition, the modularity is the difference between the real fraction of edges within communities and their expected fraction when edges are placed at random. Another Q criterion is constructed rather on edges than on communities. Maximizing Q yields the optimum partitioning of a NW with each component of partition as one community. However, this optimization over all possible partitions is a NP-hard problem, so suboptimal solutions can be reached by numerical methods, such as simulated annealing and greedy algorithms, spectral and genetic methods, mathematical programming and tabu-search, extremal and multistep optimizations. Dynamics on NWs provide another ways of finding CS by methods of synchronization and diffusion processes, random walks and Laplacian matrices of NWs, map equations aka Infomap, and the fast converging label propagation technique.

Communities of the real world NWs are highly overlapped, for instance, in social NWs one person always belongs to multiple social circles by family, friends, profession, and other ties. To uncover overlapping CS, the so-called clique percolation method has been successfully applied in biological, social, and information NWs. NW can be specified by subgraphs such as a clique where all vertices are connected by edges. A maximal clique and weakened versions of clique are known too, such as a *k-core*, or a subgraph with all vertex degrees equal or more than a number *k*. Other approaches have also been developed, for instance, the extended label propagation technique called the Community Overlap PRopagation Algorithm (COPRA). Probabilistic models with maximum likelihood criteria are known as well, for instance, the models based on latent Dirichlet allocation, where the estimated parameters provide information on the overlapping CS. Communities' evolution in time, including the birth, growth, contraction, merge, split, and death is studied by the dynamic community detection methods focused on snapshots of NWs. Benchmark NWs with a priori known CS are used to measure the performance of different methods. Two well-known of them are the GN standard benchmark NW proposed by Girwan and Newman, and LFR benchmark NW proposed by Lancichinetti, Fortunato, and Radicchi. Measurements used for community detection include the normalized mutual information, variation of information, and normalized variation defining a distance in the space of partitions.

Several other open problems have been considered in the author's research described in the next chapters of the monograph. Chapter 2 proposes detection of the overlapping and hierarchical CS in the algorithm EAGLE—the agglomerativE hierarchicAl clusterinG based on maxmaL cliquE. A maximum clique is a clique which is not a subset of any other clique. The algorithm generates dendrograms and chooses cuts which break them into communities, with technical details given and examples of application to the world association and the scientific collaboration NWs with many thousands of nodes and edges are considered. Chapter 3 addresses the problem of multiscale community detection in NWs with heterogeneous degree distributions. The identification of CS can be seen as finding the dimensionality reduction that captures the NW topology. The well-known Laplacian and modularity covariance matrices used for graph partitioning and a rescaling transformation of them to a kind of correlation matrices for handling heterogeneity of the nodes in NWs are considered by the principal component analysis applied for dimensionality reduction and CS detection. Experiments on various known and simulated NWs demonstrate that the better detection results can be achieved with the rescaled matrices. Chapter 4 focuses on the relation between CS and dynamics on complex NWs using diffusion process of a random walking between the neighboring nodes. Some local equilibrium states appear before stabilization of the process, and stability of these equilibria can be measured by their duration time. The intrinsic CS is revealed by these stable local equilibrium states, and the search for CS can be performed by optimizing the diffusion process conductance, or easiness of diffusion through different communities in the NW. Comparative analysis of the spectral methods for community detection is given by five matrices describing an undirected graph corresponding to various NWs: adjacency, standard and normalized Laplacian, modularity, and correlation matrices, and the number of communities can be determined by the gaps in their eigenvalues (eigengaps) and the related eigenvectors. Test results show that standard Laplacian and correlation matrices outperform the others, which can be explained by the heterogeneity normalization with respect to the degrees of nodes. Chapter 5 describes problems of exploratory analysis of the structural regularities in NWs based on probabilistic models viewing a NW structure as observed quantities. Then communities can be identified by fitting the model to the observed NW structure. Various approaches are known, for instance, fuzzy membership of nodes and latent Dirichlet allocation, hierarchical random graph and mixture models, degree-corrected, and mixed membership stochastic block models.

The latter can be presented by a matrix with elements of probability values that a randomly selected edge connects two latent groups of nodes, and in the standard block model the nodes in the same group are identical. A new model with this restriction relaxed is suggested, the objective of expected log-likelihood and expectation-maximization (EM) algorithm are considered for the estimation parameters of assigning to communities. Block models not only with positive but also with negative links (in social NWs those can correspond to friends and enemies, or war relations, respectively) are studied. Comparisons with other models and datasets were performed and they showed that without prior knowledge on the types of structures, this approach can detect both assortative and disassortative regularities in communities of NWs.

Each chapter presents multiple formulas, graphs and tables, references to numerous modern sources, gives clear exposition of the techniques and tests them on real NWs with known structure (e.g., the karate club studied by Zachary, or bottlenose dolphins studied by Lusseau, etc.) and artificially generated NWs. The monograph offers an exceptional set of methods of research on networks, and can be useful and interesting to researchers and students in various areas.

Stan Lipovetsky
GfK North America, Minneapolis

Biodemography of Aging: Determinants of Healthy Life Span and Longevity, by Anatoliy I. Yashin, Eric Stallard, and Kenneth C. Land. Dordrecht: Springer Science + Business Media B.V., 2016, xvii+463 pp., $ 99.99, ISBN: 978-94-017-7585-4.

The monograph presents the volume 40 in the Springer Series on Demographic Methods and Population Analysis, and covers work performed by the leading research team of the authors with contributions of 10 other scientists in the Biodemography of Aging Research Unit (BARU) at the Center for Population Health and Aging at Duke University, Durham, NC, USA. Biodemography is a relatively young multidisciplinary approach aiming to integrate biological knowledge and methods of traditional demographic analysis to study health and mortality in population and across individuals, and biodemography of aging is focusing on the aspects of health and life span. The goal of the research is determined by the questions stated in the first phrases of preface: "Some people live to age 100 years and more, others become sick and die prematurely. What factors determine human lifespan? Which biological mechanisms are involved in the regulation of health span and longevity? Which forces shape the age pattern and affect the time trend of human mortality?" (p. v). To answer such challenging questions, the researchers' team performs longitudinal studies on genetic and non-genetic supermassive *big data* collected in BARU and analyzed with advanced biodemographic models. Such models include a specific stochastic process model (SPM) and the grade of membership (GoM) model – they both had been originated by M.A. Woodbury in the 70s and developed by the BARU team in the last decades, particularly, GoM has been upgraded under the linear latent structure (LLS) paradigm.

The material is organized into three parts and 20 chapters, each with multiple sections and subsections. Chapter 1 introduces the main ideas of biodemographics in genetic analysis of human longevity, evolution of aging, health, and mortality, the frailty models, Strehler and Mildwan (SM) model of aging and mortality, and its development into SPM and other modern approaches. Part I in its nine chapters describes information on determinants of health and survival outcomes, and rather conventional statistical modeling of various types of data. Part II also contains nine chapters presenting more sophisticated analysis by the advanced statistical methods, and Part III in one chapter summarizes the obtained results.

Part I of mostly empirical studies starts with Chapter 2 discussing age-related changes in the analysis of average age trajectories for various physiological variables and biomarkers (e.g., a widely known BMI, or body mass index) by males and females, or smokers and nonsmokers, in Kaplan–Meier estimates of survival functions. Chapter 3 describes health deterioration in connection with different chronic diseases (e.g., cancer, heart diseases and stroke, diabetes, asthma, neurodegenerative diseases, etc.) related to common genetic and nongenetic risk factors, studied by observational data collected in the U.S. Medicare Files of Service Use(MFSU) that represents big data regression analysis of morbidity and mortality, disability, and comorbidity patterns of incidence rates in older population. The standard Charlson Comorbidity Index and a newly developed Adjusted for elderly population Multi-Morbidity Index were used in the specific cohort patterns, and predictive models for them were considered via logistic regressions by various covariates, with additional sensitivity–specificity analysis in receiver operating characteristic (ROC) curves. Chapter 4 focuses on the dependence among diseases for elderly, with interplay of ontogenetic changes, senescence processes of deterioration with age, and exposures to a hazardous environment leading to health pathologies that make the diseases mutually dependent. Correlation analysis of co-occurrence between different chronic diseases using multiple causes of death data was performed, which revealed negative temporal correlations between death due to cancer and other major diseases that could be explained by effects of apoptosis, or programmed cell death from different diseases, so the cancer survivors could be more resistant to other diseases and their longevity would grow. Chapter 5 describes the factors that could increase both cancer risk and longevity and go in parallel with economic growth in the developed countries. Those include a higher proportion of vulnerable individuals due to more surviving people with weaker immune system, exposures to novel medicine (such as hormone replacement therapy, oral contraceptives, and household chemicals), a Western life style impact (e.g., delayed childbirth and food enriched with growth factors) that can increase susceptibility to cancer. Chapter 6 studies the Medicare expenditures trajectories by the National Long Term Care Survey (NLTCS), MFSU, and other data bases. In 2009 more than 46 million people were covered by Medicare, and forecasting by cohort survival models for various diseases show that by 2031 there could be 77 million individuals, which is an important issue for the health care planning. Chapter 7 describes a special cumulative Deficits Indices (DI) for analysis of aging, health, and lifespan. Mitnitski and Rockwood proposed a frailty index incorporating multiple variables and accounting for degradation of neuroendocrine, immune, and other systems that yield a wide range of health disorders related also to accumulation of damage in cellular tissues. DI, phenotypic frailty index, physiological indices used in Framingham Heart Study (FHS) and described with age trajectories can be taken in estimating mortality rates for specific age patterns of decedents and survivors. Chapter 8 investigates dynamics of trajectories for age-related changes in FHS key physiological variables (BMI, systolic and diastolic pressure, pulse pressure and rate, blood glucose, hematocrit, and total cholesterol) in relation to the spans of life and health. The analysis of monotonicity of trajectory curves, their intercepts, and left and right slopes before and after the maximum, shows, for instance, that the slope as the rate of change within the range of 40–60 years old could be used for prediction of the total longevity and health span at the elder age. Chapter 9 describes some recent genome-wide association studies on the example of the Apolipoprotein E polymorphism, considering complex modes of gene actions, genetic trade-offs, antagonistic genetic effects on the same traits at different ages and on lifespan, and showing that studies of genetic signals are still far from reliable strategies aiming to improve population health. Chapter 10 concludes on the empirical patterns of health and longevity, indicating that age is a major risk factor of the phenotypes expressed via dysregulation of physiological functions, prevalence of diseases, and case fatality change that incorporates the occurrences happened with a human organism during the life to a given age.

Part II is devoted to the advanced statistical models of aging, health, and longevity, starting with Chapter 11 that reviews various approaches to such analyses of longitudinal data in biodemographic perspectives. The traditional Cox hazard model with time-dependent covariates yields underestimated impact of the longitudinal values on the event, and other standard techniques, such as mix-effect models or generalized estimating equations also lead to biased estimates. The appropriate to the aims of biodemography of aging methods are the so-called *joint models* that work with the combined longitudinal measurements and time-to-event data. The standard joint model consists of two submodels, one describing the dynamic of longitudinal data and another describing the survival or time-to-event process, formulated in a linear mix-effects model with normal errors and random effects. Various extensions are known, for instance, *change-point joint models*, nonlinear mix-effects and generalized linear mix-effects models, and many others. Chapter 12 describes the stochastic process models (SPMs) capable to incorporate variables of resistance to stresses, adaptive capacity, normal physiological states, allostatic adaptation and load, all of which help to link the aging-related changes in physiological indices with morbidity and mortality risks. Diffusion-type stochastic differential equations are used for describing the processes, and maximum likelihood (ML) objective yields the parameter estimates. Strehler and Mildwan (SM) model with its upgrading by Yashin and co-authors to two SPMs are considered as well. Chapter 13 continues with the latent class SPM and evaluation of hidden heterogeneity known in frailty models for longitudinal data. Such models are capable to account for biomarkers with genetic information that can become very important element in the interaction of genetic, biological, socioeconomics, and demographic characteristics for longitudinal aging problems. The multinomial-logit models are employed together with stochastic differential equations and ML parameters estimation, and simulations studies are performed as well. Chapter 14 describes results of simulation studies with longitudinal genetic-demographic SPMs for nongenotyped and genotyped data, and shows that including the latter data improve statistical power of hypotheses testing. Chapter 15 introduces integrative mortality models with parameters having biological interpretations. A diffusion-type continuous-time stochastic process describes evolution of the physiological states over the life time, and a finite-state continuous-time process describes changes in health during this course. Mathematically, such descriptions involve stochastic processes with continuous and jumping paths, and using Gaussian approximation facilitates calculations in ML optimization for parameter evaluation. These models help to understand how people lose health and functional capacities during the aging process. Chapter 16 continues description of the integrative mortality models for FHS and simulated data in various observational plans, and considers combined big data to reach a high quality of statistical estimation of dynamic characteristics in multidimensional models and of tests in statistical hypotheses. Chapter 17 presents a new longitudinal form of GoM model with time-varying covariates, and applies it to the analysis of dementia by NLTCS data. This progressive and generally fatal disease of developing cognitive impairment and deterioration in functioning in older people is modeled as a complex multidimensional process governed by a latent 3D bounded state-space process. ML objective is used in constrained iterative Newton–Raphson procedures, with model testing by AIC and BIC criteria, and ancillary analysis of acute and long-term-care, and mortality. An appendix describing consistency of the cross-sectional empirical GoM models is also provided. Chapter 18 considers latent class model and its recent development into the LLS analysis, also with application to the NLTCS data and comparison of the results with GoM model. This approach permits for data reduction from hundreds of actually and potentially observed variables to the main few latent variables representing “pure-type” individuals, e.g., healthy, disabled, having chronic diseases. Chapter 19 resumes on the problems of statistical modeling presented in Part II and their further development.

Finally, Part III in its Chapter 20 concludes on the continuing research for determinants of human health, life span, and longevity. Many questions about the origin of chronic diseases and aging-related changes can be addressed by the evolutionary population genetics. The new science of the biodemography of aging is devoted to solving problems of the genetic structure of population cohorts with increasing age, leading to more informative analysis and decisions on healthy life and longevity for population and individuals.

The chapters are supplied with dozens and hundreds of references on the most recent works on the topics. The monograph presents a great compendium of modern views, developments, approaches, and achievements in numerous aspects of the biodemography of aging. Incredibly enormous amount of information on human biology on one side and mathematical and statistical complex modeling on another side are presented in a very dense form of this amazing book on the current research in aging and longevity problem. The book can be very useful not only for specialists and graduate students but also for readers interested in modern interdisciplinary sciences.

Stan Lipovetsky
GfK North America, Minneapolis

Causal Nets, Interventionism, and Mechanisms: Philosophical Foundations and Applications, by Alexander Gebharter. Cham, Switzerland: Springer, 2017, vii+184 pp., $ 99.99, ISBN: 978-3-319-49907-9.

The monograph belongs to the Springer series of Synthese library: Studies in epistemology, logic, methodology, and philosophy of science, vol. 381, and originally its content was presented as the author’s doctoral thesis at the University of Dusseldorf. The book consists of six chapters, and in Chapter 1 the author describes various aspects of this philosophical–statistical interscience study and the work structure. Chapter 2 presents a brief introduction to the main probability concepts and formulae, including Bayes theorem, graph theory, variables’ probabilistic dependence/independence, Markovian chain rules, parents, and conditions used in Bayesian networks (BNs). Chapter 3 continues with deeper description and axiomatization for the causal graphs and BNs framework, originated as early as 1921 by S. Wright, and developed by Reichenbach, Pearl, and many other authors. Chapter 4 starts with D. Hume question on any reason to belief in causation as ontological reality, not merely a subjective feature of human minds. The author argues that although the causation cannot be explicitly defined, it can be implicitly characterized by axioms connecting causality to observable phenomena, and continues with definitions in more complex scenarios describing empirical content of the theory of causal nets. Chapter 5 considers a recent Woodward’s intervention theory of causation, describes it in axioms of the BNs, and extends to an alternative theory of deterministic and stochastic interventions. Chapter 6 focuses on causal nets and mechanisms (latent complex systems or functions revealing in the observed phenomena), and debates with various authors within modern philosophy of science. The author introduces the recursive BN and multilevel causal model, applying them to description of mechanisms, particularly, to constitutive relevance relations in them, modeling causal cycles, solving static and dynamic problems.

The monograph presents an exceptionally clear exposition on Bayesian Nets and their extensions on one side, and on philosophical causality problems in various facets on the other, so it can be useful and interesting to researchers and students in different areas.

Stan Lipovetsky
GfK North America, Minneapolis

Applied Statistics for Agriculture, Veterinary, Fishery, Dairy, and Allied Fields, by Pradip Kumar Sahu. Springer India, 2016, xvi+533pp, $109 (eBook), $149 (Hardcover), ISBN: 978-81-322-2829-5; ISBN: 978-81-322-2831-8 (eBook).

This is a well-written book covering a number of introductory statistical topics. The book deploys real life examples from agriculture to allied health sciences. There are no end of chapter exercises. However, more than 165 worked out examples are presented throughout the book. Statistical software packages such as MS Excel, SPSS, and SAS are used throughout the book. Solutions along with screenshots of the results are very helpful in comprehending the topics covered in the book.

The book begins with an introduction to statistics and biostatistics covering the use and the scope of statistics, steps in statistical procedures, limitations of data, limitation of statistics, and tabulation of data.

The focus on introductory statistical topics continues in Chapter 2 covering concepts related to data including natural, experimental, primary, secondary, cross-sectional, and character data. Variables, and constants, grouping including cumulative and relative frequencies, and the frequency density are discussed in this chapter. The discussion on data ends with the presentation of data in textual, tabular, and diagrammatic forms.

In Chapter 3 the readers are introduced with additional introductory statistical topics including summary statistics, characteristics of good measures, measures of central tendency including arithmetic mean, geometric mean, harmonic mean, merit and demerit of each, the application of different types of mean and media. The author then focuses on partitioning of data values (percentiles, deciles, and quartiles), mode, relationship of mean, median, mode, midpoint range, selection of proper measure of central tendency. Dispersion and its measures including range, mean deviation, standard deviation, and merit and demerit of each, quartile deviation are presented along with worked out examples. Other related topics covered in this chapter include moments (conversion of moments, Sheppard's correction for moments (merits and demerits), relative measures of dispersion, skewness, and kurtosis along with going through several examples step-by-step with screenshots.

Chapter 4 is a short one focusing on the continuation of covering introductory topics in statistics including an introduction to set theory and its application, experiment, probability, important rules in probability, random variables, and their probability distributions, mean, variance, moments of random variable, and moment generating function. The chapter ends with some important continuous and discrete distributions along with a relatively long discussion on normal distribution.

In Chapter 5, the author discusses the population versus sample, parameter versus statistic, probability and nonprobability sampling techniques. A good number of these techniques are presented here that should be of interest to the audience of the book.

The next topic in sequence, which usually appears in a standard introductory book in statistics, is "statistical inference." The book addresses an extensive presentation on the estimation and hypothesis testing. The testing and interval estimates include mean, proportions, and the population variance. Discussions continue with Cochran's approximation to the Fisher–Behrens problem and Yate's correction in a 2 × 2 contingency tables where the frequency of a cell is less than 5.

In Chapter 7, we see a detailed presentation on the correlation coefficient along with several worked out examples. Several correlation coefficients such as correlation coefficient of bivariate frequency distribution, limitations, rank correlation, correlation ratio, properties of correlation radio, coefficient of concurrent deviation are discussed. SAS, MS Excel, and SPSS are used to demonstrate the calculation of the correlation coefficients through these packages.

Regression analysis is discussed in the next chapter where we see the importance of this topic to the statistics discipline. The equation, assumptions imposed on regression, linear regression along with the estimation of its parameters, relation between the correlation coefficient and linear regression are presented here. Linear regression is then extended to multiple regressions along with related topics such as estimation of the parameters. Following this topic, the author turns attention to multiple correlation coefficient, coefficient of determination along with some worked out examples related to the topics covered in the chapter.

Chapter 9 covers analysis of variance, linear ANOVA model, and assumptions imposed on this model. Some designs of experiments (DoE) are covered here including one-way ANOVA using MS Excel, two-way ANOVA with/without replicates. Additional topics includes violation of assumptions in ANOVA, followed by logarithmic, square root, angular transformation, and the effect of change in origin and scale on ANOVA.

Additional topics on DoE are covered in the following chapter where the author talks about principles of design, uniformity trial, optimum size and shape of experimental units, layout, and steps in DoE. Some examples of completely randomized design (CRD), advantages and disadvantages of this design randomization and layout, statistical model and analysis, merits and demerits of CRD are also covered here. Additionally, randomized block design/randomized complete block design (RBD/RCBD) and advantages versus disadvantages of this design are presented next. The chapter ends with Latin square design (LSD), advantages and disadvantages of this design and missing plot techniques in CRD, RBD, and LSD.

The topic of DoE is continued in the next chapter discussing factorial experiments, factor and its levels, type of factorial experiment, effects and notations used in this design followed by advantages and disadvantages of this design. The chapter concludes with two-factor asymmetrical ($m \times n$, $m \neq n$), factorial experiment, three factor experimental design, m ($m \times n$, $m \neq n$), asymmetrical factorial experiment, incomplete block design, and split plot design.

The last chapter (12) covering technical topics is devoted to additional presentations on DoE. Some special experiments and designs, comparison of factorial effects versus single control treatment, augmented designs for the evaluation of plant germplasms, augmented CRD, augmented RBD, combined experiment, analysis of experimental data measured over time are presented next. The last few technical topics discussed in the book include observations taken over time in RBD, in two-factor RBD, and in split plot design, experiments at farmers' field, major considerations during experimentations at farmers' fields.

The book ends with a discussion on the importance of statistics in research and misuse of this discipline. It also includes an advice on using statistical software packages. The author demonstrates a couple of examples on regression analysis and DoE run on SPSS and MS Excel and the different outputs generated by each software.

This is a well-written textbook covering a number of statistical topics suitable for a textbook in an undergraduate program for the areas listed in the title of the book. It is also a suitable textbook in a statistics curriculum. Special emphasis is placed on the topic of DoE. Examples are very helpful in comprehending the material covered in the book. Examples are carefully selected from the areas listed in the title of the book. Graphical presentations and screenshots are very helpful to the readers. With no doubt, the addition of some exercises at the conclusion of each chapter would make this an excellent textbook for the target audience.

Morteza Marzjarani
Saginaw Valley State University (retired)

Basic Experimental Strategies and Data Analysis for Science and Engineering, by John Lawson and John Erjavec. Boca Raton, FL: CRC Press (Chapman & Hall), 2016, 434 pp., $79.95 (H), ISBN: 978-1-4665-1217-7.

Overview

The purpose of this book is to educate scientists and engineers on statistical strategies for developing and improving products and processes, because: "Companies that use these strategies as standard operating procedures can expect large cost reductions in manufacturing, improved product quality, and reduced lead time for the introduction of new products and/or manufacturing methods" (Preface). The material covered in the book has been taught in a one-semester course and portions of the book have been taught in workshop settings. The book's organization and coverage are similar to that of *Statistics for Experimenters*, by Box, Hunter, and Hunter (1st and 2nd ed., 1978, 2005), and *Experimental Designs*, by Cochran and Cox (2^{nd} ed., 1957). An exception is the addition of a chapter on mixture experiments.

Chapter 1. Strategies for Experimentation with Multiple Factors

This chapter defines the terminology of experimental design and analysis. For example, the authors (section 1.2) define an experimental design as "the collection of experiments to be run."

I may be (actually am) old-school, but I think of an experimental design (or "strategy," the term of choice by the authors) as the protocol by which "treatments" are assigned to "experimental units." For example, the family of randomized block designs is defined by a particular protocol for assigning treatments to experimental units (or "runs" in the context of a manufacturing process, often the situation addressed by the authors) in multiple blocks of experimental units. The treatments may be multiple levels of a single factor or combinations of multiple levels of multiple factors. Other important aspects of the protocol are randomization, replication, and blocking, all of which shape the collection of experiments to run.

Another sometimes-troubling term is *Experimental Error*. The authors define experimental error as "the difference between any given observed response, Y, and the long-run average (or "true" value of Y)." I would avoid the morass of explaining a "true" value.

I think of experimental error as a term for the variability of responses among experimental units that receive the same treatment. In the statistical models we use we write $Y = f(X) + e$, where Y is the measured response, X represents the controlled variables in an experiment, and e, for error, is often modeled as a random observation from a Normal distribution with mean zero and unknown standard deviation. This term in the model represents the response variation among similar experimental units that receive the same treatment.

Section 1.3 compares the "classical" *one-at-a-time* multifactor experimental strategy with statistical strategies for simultaneously studying the effects and interactions of many factors. It also discusses the perils of using historical data to estimate the effect of X on Y.

Experimental design texts typically present and illustrate design families in stand-alone fashion—one design at a time. In practice, though, manufacturers, R&D labs, government agencies, etc. conduct *experimental programs*—series of experiments that, it is hoped, lead to better processes and products. That's the basis of "continuous improvement;" that's the scientific method in action. However, the process is sometimes not very attractive, for example, planned by committee, or determined by a *Goldilocks* strategy: Try this, try that, ….

The authors take a more structured approach. They present a table (Fig. 1.4 and repeated elsewhere) depicting a sequence of experiments leading to "increased knowledge." The stages in this sequence are: i. Preliminary exploration, ii. Screening factors, iii. Effect estimation, iv. Optimization, v. Mechanistic modeling. This strategy is illustrated with a chemical process example. Chapter 11 is devoted to this strategy for experimental programs. The authors are to be commended for taking this programmatic view of a suite of experiments.

Chapter 2. Statistics and Probability

Chapter 2 is a condensed Stat101. It starts, as I think it should, with data plots. The first illustrative data, Table 2.1, are the bore diameters of "nine consecutive parts" (off a production line, say), but, unfortunately, the first plot is premature: a dot plot of the measured bore diameters. The dot plot ignores a possibly important "lurking variable," the production sequence. Machinery can wear, temperatures can rise, … and the responses of sequentially produced items can reflect such effects. Before the data are plotted as though there are no such effects, that assumption should be examined.

A scatterplot of the last two digits of measured bore diameter (they all are measured to 0.57yy) versus part number shows an increasing trend over the first eight consecutive parts, then the ninth is appreciably less than observations 1–8 (an outlier? a tool adjustment? a production pause? just random? …). All these parts may be well within functional spec limits for bore diameter, so this erratic pattern may be of no practical concern. Nevertheless, when "sequence" is a lurking, observable variable, it should be considered before doing an analysis or data-plot that assumes sequence has no effect. A subsequent section on box-and-whiskers plots has the same problem—the lurking sequence effect is overlooked.

[An Aside: Lawson and Erjavec are not alone. The Box, Hunter, and Hunter (1978, 2005) text has a similar example. The experiment is a comparison of the yield of tomatoes for a row of plants treated by either of two fertilizers. A plot of yield versus row-position, by fertilizer, shows a distinct trend (soil fertility?) along the row of plants, but little difference between fertilizers. A plot or analysis (e.g., ANOVA) that ignores the yield trend along the row of tomatoes misses the message in the data.]

In the subsequent section on histograms, 84 pH measurements are presented in a 6 × 14 array. The reader is not told whether this structure is just for the sake of displaying the data on the page or because the rows and columns correspond to structural variables in the suite of measurements, such as technician and day. More than once in my consulting life I've been presented with a table of data and only by questioning have found out that the table structure represents pertinent variables.

In Sec. 2.3, Table 2.5 is a set of data from Johnson and Leone (1964) in which different levels of stress were applied to steel specimens and the time to crack-initiation (termed *incubation time*) was measured. The authors use these data to illustrate a scatterplot and the correlation coefficient.

First, though, I'd like to know about the *strategy* behind this experiment. I'd like to know why the selected stress levels and replication pattern were chosen. Might there be regulatory limits on what is an acceptable "stress/strength" relationship in order to qualify a particular steel for use in a bridge, say? Also, why was the response variable selected for tabulating and plotting the "log-incubation-time," rather than incubation time? It's difficult to exponentiate numbers in one's head in order to translate log-times in a table or scatterplot to time in hours, days, or the pertinent time unit of interest. Sometimes a log-linear relationship is an empirical choice that linearizes a nonlinear relationship or homogenizes the variance, or it might be physics-based. Here, because of the range of the incubation times, the scatterplot patterns of time versus stress and log-time versus stress are very similar. As one would expect, the higher the stress, the shorter the crack-incubation time.

Throughout Chapter 2 there is digititis—too many digits in various tables and statistics. Statistics is about separating signal and noise. Excess digits in a table is noise that obscures the signal. *P*-values do not need to be calculated to 0.00894 and an F-statistic does not need to be reported as 3.117153.

Section 2.5 addresses the two-sample *t*-test for comparing two means and the ANOVA for comparing seven means. I would suggest that when teaching the examples in this section, an instructor should first plot the data so that the student can see how the patterns in the data are reflected in the findings in the quantitative analyses. (As stated in Chapter 3: "It always helps to get a visual appreciation of the data before tackling the numerical analysis.")

In their discussion of significance testing the authors follow the convention of characterizing a *P*-value of 0.05 or less as a demonstration that the difference of interest is "statistically significant," while $P = 0.01$ or less means the difference is "highly (statistically) significant." Confidence intervals are not introduced in this introductory chapter, but arise later in the context of confidence intervals for linear model coefficients.

Chapter 3. Basic Two-Level Factorial Experiments

Section 3.7 presents an interesting example of a completely randomized design with eight treatments and two replicates of each. The eight treatments are the 2^3 combinations of three two-level factors of interest. Pardon me as I embellish the story:

Fly-ash is a residue of coal-burning. If captured it can possibly be profitably used by adding it to the mix of materials in concrete.

A cement company has two potential suppliers of fly ash. The company's engineering staff wants to compare the quality (as measured by the compressive strength) of concrete made with the two ash sources, and they want to make that comparison over a range of operating conditions. Two pertinent operating variables are (i.) the water/cement ratio and (ii.) the set-up temperature. It is known that these characteristics of the environment in which cement is mixed can affect the strength of the resulting cement.

After consulting with a friendly local statistician at the nearest university, the experiment above is run and the measured strengths of cement samples for each of the 16 runs are tabulated. The three treatment factors are: Supplier, water/cement ratio, and set-up temperature.

The experiment's results are summarized in a cube plot, Fig. 3.13, which shows that the average strengths for Supplier B are appreciably greater than those for Supplier A at all four combinations of water/cement ratio and set–up temperature. Also, Supplier B strengths are less sensitive to the operational variables than they are for Supplier A: For A the average strengths for the four combinations of operational variables range from 24.6 to 43.1 mPa, while for B the range is 41.8 to 49.8 mPa. Bottom Line: *Fly ash from Supplier B provides stronger cement and is more robust to environmental conditions than is the case for Supplier A.*

These eyeball observations are supported by *t*-tests pertaining to effects and interactions of the three factors in the experiment and by the corresponding analysis of variance. The authors, for reasons not given, pooled the ANOVA sums of squares for the three main effects and the three two-way interactions, rather than separate them.

Numerical calculations of (estimated) "effects" and "interactions" are developed heuristically; for example, the effect of X_1 is defined as the difference between the average responses of all the data points at the high value of X_1 versus the average response at the low value of X_1. Later chapters address linear models for multifactor experiments in which coefficients in the model represent effects and interactions. It might have been better to introduce these bilinear models as a basis for estimating effects and interactions for multiple two-level factors.

Chapters 4–10, 12

Many variations on the theme of multifactor experiments have been developed over the years, mostly to deal with experimental constraints or particular objectives. Blocked experiments, split-plots, fractional factorial experiments, response surface fitting, variance component estimation, and mixture experiments are examples. This book provides a good compendium of such situations and pertinent analyses, though I was surprised not to see the Latin square design—which is itself a fractional factorial collection of experiments—mentioned and illustrated. It shows the richness of multifactor experimental designs and analyses. As such, it should encourage students or professionals in science and engineering to use these tools to help understand and improve the products and processes for which they are responsible.

Chapter 11

Chapter 11 illustrates how to construct a sequence of experiments, or analyses, that move knowledge in a positive direction. For the most part this involves augmenting fractional factorial results by additional runs that fill in the factorial to a desired degree, e.g., to separate confounded effects or interactions.

Another approach is to fit different models to experimental results for a design. An example given is to run an experiment according to what are called the *definitive screening design* (Jones and Nachtsheim 2011, 2013) which is a three-level design. In the authors' example, a main effects model for five factors said first that only X_1 was "significant." (This is *screening* knowledge.) That model was followed by a quadratic model in X_1 only. (This knowledge can be used for *optimization*.) (I think I'd like to see some intermediate models and some comparison of lack-of-fit statistics before I conclude that my quest for knowledge is complete.)

Chapter 13

The authors conclude with a list of *Points to Remember*. The first is: Do not use one-factor-at-a-time designs for multifactor experiments. The others can be read at Amazon.com.

References

Box, E. P., Hunter, J. S., and Hunter, W. G. (1978, 2005), *Statistics for Experimenters* (1st and 2nd ed.), New York: John Wiley & Sons. [129,130]

Cochran, W. G., and Cox, G. M. (1957), *Experimental Designs* (2nd ed., 1957), New York: John Wiley & Sons. [129]

Johnson, N. L., and Leone, F. C. (1964), *Statistics and Experimental Design in Engineering and the Physical Sciences, volume II*, New York: John Wiley & Sons. [xxxx]

Jones, B., and Nachtsheim, C. (2011), "A Class of Three-Level Designs for definitive Screening in the Presence of Second-Order Effects," *Journal of Quality Technology*, 43, 1–15. [131]

——— (2013), "Definitive Screening Designs with Added Two-Level Factors," *Journal of Quality Technology*, 45, 121–129. [131]

Robert Easterling
Sandia National Laboratories (retired)

The Essentials of Data Science: Knowledge Discovery Using R, by Graham J. Williams. Boca Raton, FL: CRC Press, 2017, 322 pp., $69.95 (softcover), ISBN: 978-1-138-08863-4.

According to the author, this book "introduces the essentials of data analysis and machine learning as the foundations of data science. It uses the free and open source software R which is freely available to anyone. All are permitted, and indeed encouraged, to read the source code to learn, understand, verify, and extend it." The author further describes the book's key feature as its "focus on the hands-on end-to-end process" of performing data science on data. This book "covers data analysis beginning with loading the data into R, wrangling the data to improve its quality and utility, visualising (author's spelling not mine) the data to gain understanding and insight, and, importantly, using machine learning to discover knowledge from the data." The author claims that the book brings together 30 years of experience in order to present a "programming-by-example approach that allows students to quickly achieve outcomes whilst building a skill set and knowledge base, without getting sidetracked into the details of programming." The book also focuses on creating templates to support the end-to-end process flow. The R code as well as the templates are available from the book's website at *https://essentials.togaware.com*. The graphics and code examples in the book are colorized making them easier to read and interpret.

The book has 11 chapters. References are provided at the end of the book in a bibliography.

Chapter 1: Data Science
Chapter 2: Introducing R
Chapter 3: Data Wrangling
Chapter 4: Visualising Data
Chapter 5: Australian Ports
Chapter 6: Case Study: Web Analytics
Chapter 7: A Pattern for Predictive Modeling
Chapter 8: Ensemble of Predictive Models
Chapter 9: Writing Functions in R
Chapter 10: Literate Data Science
Chapter 11: R With Style

Chapter 1 covers an overview of data science, the data scientist, and both closed and open source software. It is very well written and provides a great introduction to the rest of the book.

Chapter 2 introduces the reader to the R programming language and the RStudio software interface. Packages and libraries are also discussed, as well as functions, commands, and operators in the R language. Weather data from 50 stations across Australia over an 8-year period is used to demonstrate the ideas in this chapter.

Chapter 3 discusses how to prepare the data and shows the reader how to build a template that can be used and reused to prepare the datasets in the book as well as your own. The weather data introduced in Chapter 2 is used as an example to demonstrate the use of this template. The chapter shows the reader how to examine the data as well as how to clean it and prepare it for analysis. Feature selection and creation are also discussed in the context of data wrangling. Finally, the data are prepared for model building and the data are saved. All in all it is an excellent chapter on how to get your data ready for data visualization (Chapter 4).

The R package ggplot2 is prominently featured in Chapter 4 which covers the topic of data visualization. Common graphics such as the scatterplot, bar chart, and box plot are illustrated with the weather data. Chapter 5 covers a case study of an analysis of data from the Australian government pertaining to the potential of sea ports across the country for future development and considered the resulting impact on jobs in the surrounding regions of those ports. This data provides the reader an opportunity to practice loading, wrangling, exploring, and visualizing the data as covered in the prior chapters. The reader is instructed on how to download the data so they can work along with the book.

Chapter 6 affords another case study on web analytics for the reader to practice the methods discussed in this text. In this chapter, the reader learns how to access data through the web rather than downloading a file for analysis (as shown in Chapter 5).

Chapter 7 presents a template for building analytic or machine learning models. Just as the concept of a template was used in earlier chapters for the ingesting, processing, reviewing, transforming, and cleaning of data, in this chapter the reader develops a pattern for building models using R. The weather data are again used to demonstrate multiple approaches, such as building a decision tree model, assessing model performance through accuracy and error rate using the confusion matrix, and construction of a ROC curve and risk chart. Finally, the model parameters are tuned in order to explore the particular model that best fits the data. This involves searching through a range of parameter values, building a model for each setting, and then identifying the particular combination of parameter settings that yields the best model. The process of model tuning is shown using a variety of R packages. A comparison of performance measures is discussed as a means of identifying the best model. Then the model is saved to a file for future use.

This process culminates in Chapter 8 that discusses the use of a model building template to illustrate the use of alternative machine learning algorithms through ensemble modeling. This is a popular approach that has resulted in the winning of several Kaggle and Data Mining Cup competitions. This chapter focuses on two popular and effective machine learning algorithms that exhibit an ensemble approach—random forest (one of my preferred methods) and extreme gradient boosting (another decision tree approach).

The remaining chapters deal with more programming. In Chapter 9, writing functions in R is covered. Chapter 10 is dedicated to the topic of literate programming. In this chapter, LaTeX is discussed for the formatting of programs, Kable and XTable for the formatting of tables, and the use of the knitr package for the inclusion of figures. Finally, in Chapter 11, the reader is presented with how to program R with style. Conventions for naming, comments, layout, functions, and assignments are shown in terms of what is preferred, what an alternative might look like, and what form is discouraged.

I have several books on data science and R, as well as other similar subjects and programming languages, in my personal library. However, this book is a great blend of important data science topics and R programming that will make it a great reference for anyone working in this important and immensely popular area. I highly recommend this book for college students learning what it takes to start their career in data science or even current professionals wanting to make a career change or who just want to know more about the subject (and do some R programming as well). I should also mention that the author builds this book on an earlier book of his entitled *Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery* (Williams, 2011). While I will not review that book in detail here, it serves as a great introduction and companion text to the book I have reviewed here. I would highly recommend to the interested reader that they consider buying both and use them as a set.

Reference

Williams, G. (2011), *Data Mining with Rattle and R*, New York: Springer. [132]

Dean V. Neubauer
Corning Incorporated