# Use R!

*Series Editors:*
Robert Gentleman  Kurt Hornik  Giovanni Parmigiani

More information about this series at http://www.springer.com/series/6991

# Use R!

*Albert:* Bayesian Computation with R (2nd ed. 2009)
*Bivand/Pebesma/Gómez-Rubio:* Applied Spatial Data Analysis with R (2nd ed. 2013)
*Cook/Swayne:* Interactive and Dynamic Graphics for Data Analysis: With R and GGobi
*Hahne/Huber/Gentleman/Falcon:* Bioconductor Case Studies
*Paradis:* Analysis of Phylogenetics and Evolution with R (2nd ed. 2012)
*Pfaff:* Analysis of Integrated and Cointegrated Time Series with R (2nd ed. 2008)
*Sarkar:* Lattice: Multivariate Data Visualization with R
*Spector:* Data Manipulation with R

Emilio L. Cano • Javier M. Moguerza
Mariano Prieto Corcoba

# Quality Control with R

## An ISO Standards Approach

Springer

Emilio L. Cano
Department of Computer Science
and Statistics
Rey Juan Carlos University
Madrid, Spain

Javier M. Moguerza
Department of Computer Science
and Statistics
Rey Juan Carlos University
Madrid, Spain

Statistics Area, DHEP
The University of Castilla-La Mancha
Ciudad Real, Spain

Mariano Prieto Corcoba
ENUSA Industrias Avanzadas
Madrid, Spain

ISSN 2197-5736 ISSN 2197-5744 (electronic)
Use R!
ISBN 978-3-319-24044-2 ISBN 978-3-319-24046-6 (eBook)
DOI 10.1007/978-3-319-24046-6

Library of Congress Control Number: 2015952314

Springer Cham Heidelberg New York Dordrecht London


Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

*To our families.*

# Foreword

Although it started almost two decades ago as a purely academic project, the R software has established itself as the leading language for statistical data analysis in many areas. The New York Times highlighted, in a 2009 article, this transition and pointed out how important companies, such as IBM, Google, and Pfizer, have embraced R for many of their data analysis tasks.

It is known that R is becoming ubiquitous in many other commercial areas, well beyond IT and big pharma companies. This is well described in this book, which focuses on many of the tools available for quality control (QC) in R and how they can be of use to the applied statistician working in an industrial environment. All products that we consume nowadays go through a strict quality protocol that requires a tight integration with data obtained from the production line.

The authors have put together a manual that makes Springer's use R! series become even more comprehensive as this topic has not been covered before. QC is an important field because it requires a specific set of statistical methodology that is often neglected in these times of the Big Data revolution. This volume could well serve as an accompanying textbook for a course on QC at different levels, as it provides a description of the main methods in QC and then illustrates their use by means of examples on real data sets with R.

But this book is not only about teaching QC. In fact, the authors combine an outstanding academic background with extensive expertise in the industry, including professional in-company training and an active involvement with the Spanish Association for Quality (AEC) and with the Spanish Association for Standardization (AENOR, member of ISO). Thus, the book will also be of use to researchers on QC and engineers who are willing to take R as their primary programming language. What makes QC different is that it is at the core of production and manufacturing. In this context, R provides a suitable environment for data analysis directly at the production lines. R has evolved in a way that it can be integrated with other software and tools to provide solutions and analysis as data (and goods) flow in the lines.

Furthermore, the authors have reviewed ISO standards on QC and how they have been implemented in R. This is important because it has serious implications in practice as production is often constrained to fulfill certain ISO standards. For this

reason, I believe that this book will play an important role to take R even further into the industrial sector.

Finally, I congratulate the authors for continuing the work that they started in their book on Six Sigma with R. These two books could well be used together not only to control for the quality of the products but also to improve the quality of the industrial production processes themselves. With R!

Albacete, Spain Virgilio Gómez-Rubio
July 2015

# Preface

## Why Quality Control with R?

Statistical quality control is a time-honored methodology extensively implemented in companies and organizations all over the world. This methodology allows to monitor processes so as to detect change and anticipate emerging problems. Moreover, it needs statistical methods as the building blocks of a successful quality control planning.

On the other hand, R is a software system that includes a programming language widely used in academic and research departments. It is currently becoming a real alternative within corporate environments. With R being a statistical software and a programming language at the same time, it provides a level of flexibility that allows to customize the statistical tools up to the sophistication that every company needs. At the same time, the software is designed to work with easy-to-use expressions, whose complexity can be scaled by users as they advance in learning.

Finally, the authors wanted to provide the book with a new flavor, including the *ISO Standards Approach* in the subtitle. Standards are crucial in quality and are becoming more and more important also in academia. Moreover, statistical methods' standards are usually less known by practitioners, who will find in this book a nice starting point to get familiar with them.

## Who Is This Book For?

This book is not intended as a very advanced or technical reading. It is aimed at covering the interest of a wide range of readers, providing something interesting to everybody. To achieve this objective, we have tried to write the least possible mathematical equations and formulas. When necessary, we have used formulas followed by simple numerical examples in order to make them understandable.

The examples clarify the tools explained, using simple language and trying to transmit the principal ideas of quality control.

As far as the software is concerned, we have not used complicated programming structures. Most examples follow the structure `function(arguments)` → results. In this regard, the book is self-contained as it comprises all the necessary background. Nevertheless, we reference all the packages used and encourage the reader to consult their documentation. Furthermore, references both to generic and specific R books are also provided.

Quality control practitioners without previous experience in R will find useful the chapter with an introduction to the R system and the *cheat sheet* in the Appendix. Once the user has grasped the logic of the software, the results are increasingly satisfactory. For quality control beginners, the introductory chapter is an easy way to start through the comprehensive intuitive example.

Statistical software users and programmers working in organizations using quality control and related methodologies will find in this book a useful alternative way of doing things. Similarly, analysts and advisers of consulting firms will get new approaches for their businesses beyond the commercial software approach.

Statistics teachers have in a single book the essentials of both disciplines (quality control and R). Thus, the book can be used as a textbook or reference book for intermediate courses in engineering statistics, quality control, or related topics.

Finally, business managers who want to understand and get the background to encourage their teams to improve their business through quality control can read selected chapters or sections of the book, focusing on the examples.

## How to Read This Book

In this book, we present the main tools and methodologies used for quality control and how to implement them using R. Even though a sequential reading would help in understanding the whole thing, the chapters are written to be self-contained and to be read in any order. Thus, the reader might find parts of the contents repeated in more than one chapter, precisely to allow this self-contained feature. On the other hand, sometimes this repetition is avoided for the sake of clarity, but we provide a number of cross-references to other chapters. Finally, in some parts of the book, concepts that will be defined in subsequent chapters are intuitively used in advance, with a forward cross-reference.

We provide three indices for the book. In addition to the typical subject index, we include a functions and packages index and an ISO standards index. Thus, the reader can easily find examples of R code, and references to specific standards.

The book is organized in four parts. Part I contains four chapters with the fundamentals of the topics addressed in the book, namely: quality control (Chapters 1 and 3), R (Chapter 2), and ISO standards (Chapter 4). Part II contains two chapters devoted to the statistical background applied in quality control, i.e., descriptive statistics, probability, and inference (Chapter 5) and sampling (Chapter 6). Part III tackles the important task of assessing quality from two different

approaches: acceptance sampling (Chapter 7) and capability analysis (Chapter 8). Finally, Part IV covers the monitoring of processes via control charts: Chapter 9 for monitoring variables and attributes quality characteristics and Chapter 10 for monitoring so-called nonlinear profiles.

Three appendices complete the book. Appendix A provides the classical Shewhart constants used to compute control chart limits and the code to get them interactively with R; Appendix B provides the complete list of ISO standards published by the ISO Technical Committee ISO-TC 69 (Statistical Methods); and Appendix C is a *cheat sheet* for quality control with R, containing short examples of the most common tasks to be performed while applying quality control with R.

The chapters have a common structure with an introduction to the incumbent topic, followed by an explanation illustrated with straightforward and reproducible examples. The material used in these examples (data and code) and the results (output and graphics) are included sequentially as the concepts are explained. All figures include a brief explanation to enhance the understanding of the interpretation. The last section of each chapter includes a summary and references of the ISO standards relevant for the topics covered in the chapter.[1]

We are aware that the book does not cover all the topics concerning quality control. That was not the intention of the authors. The book paves the way to encourage readers to go into quality control and R in depth and maybe make them as enthusiastic as the authors in both topics. The reader can follow the references provided in each chapter to go into deeper detail on the methods, especially through the ISO standards.

Finally, if you read the Use R! series book entitled *Six Sigma with R*, co-authored by two of this book's authors, you may find very similar content in some topics. This is natural, as some techniques in quality control are shared with Six Sigma methodologies. In any case, we tried to provide a different approach, with different examples and the ISO standards extent.

## Conventions

We use a homogeneous typeset throughout the book so that elements can be easily identified by the reader. Text in Sans-Serif font is for software (e.g., R, Minitab). Text in `teletype` font within paragraphs is used for R components (packages, functions, arguments, objects, commands, variables, etc.).

The commands and scripts are formatted in blocks, using `teletype` font with gray background. Moreover, the syntax is highlighted, so the function names, character strings, and function arguments are colored (in the electronic version) or

[1] ISO Standards are continuously evolving. All references to standards throughout the book are specific for a given point in time. In particular, this point in time is end of June 2015.

with different grayscales (printed version). Thus, an input block of code will look like this:

```
#This is an input code example
my.var <- rnorm(n = 10, mean = 2, sd = 0.5)
summary(my.var)
```

The text output appears just below the command that produces it, and with a gray background. Each line of the output is preceded by two *hashes* (##):

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.262   1.806   2.040   2.063   2.527   2.642
```

There are quite a lot of examples in the book. They are numbered and start with the string *Example (Brief title for the example)* and finish with a square (□) at the end of the example. In the subsequent evolution of the example within the chapter, the string *(cont.)* is added to the example title.

Throughout the book, when we talk about products, it will be very often suitable for services. Likewise, we use in a general manner the term *customer* when referring to customers and/or clients.

## The Production

The book has been written in `.Rnw` files. Both Eclipse + StatET IDE and RStudio have been used as both editor and interface with R. Notice that if you have a different version of R or updated version of the packages, you may not get exactly the same outputs. The session info of the machine where the code has been run is:

- R version 3.2.1 (2015-06-18), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=es_ES.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=es_ES.UTF-8`, `LC_COLLATE=es_ES.UTF-8`, `LC_MONETARY=es_ES.UTF-8`, `LC_MESSAGES=es_ES.UTF-8`, `LC_PAPER=es_ES.UTF-8`, `LC_NAME=es_ES.UTF-8`, `LC_ADDRESS=es_ES.UTF-8`, `LC_TELEPHONE=es_ES.UTF-8`, `LC_MEASUREMENT=es_ES.UTF-8`, `LC_IDENTIFICATION=es_ES.UTF-8`
- Base packages: base, datasets, graphics, grDevices, grid, methods, stats, utils
- Other packages: AcceptanceSampling 1.0-3, car 2.0-25, ctv 0.8-1, downloader 0.3, e1071 1.6-4, Formula 1.2-1, ggplot2 1.0.1, Hmisc 3.16-0, ISOweek 0.6-2, knitr 1.10.5, lattice 0.20-31, MASS 7.3-42, nortest 1.0-3, qcc 2.6, qicharts 0.2.0, qualityTools 1.54, rj 2.0.3-1, rvest 0.2.0, scales 0.2.5, SixSigma 0.8-1, spc 0.5.1, survival 2.38-3, XML 3.98-1.3, xtable 1.7-4

- Loaded via a namespace (and not attached): acepack 1.3-3.3, class 7.3-13, cluster 2.0.2, colorspace 1.2-6, crayon 1.3.0, curl 0.9.1, digest 0.6.8, evaluate 0.7, foreign 0.8-64, formatR 1.2, gridExtra 0.9.1, gtable 0.1.2, highr 0.5, httr 1.0.0, labeling 0.3, latticeExtra 0.6-26, lme4 1.1-8, magrittr 1.5, Matrix 1.2-0, memoise 0.2.1, mgcv 1.8-6, minqa 1.2.4, munsell 0.4.2, nlme 3.1-121, nloptr 1.0.4, nnet 7.3-10, parallel 3.2.1, pbkrtest 0.4-2, plyr 1.8.3, proto 0.3-10, quantreg 5.11, R6 2.1.0, RColorBrewer 1.1-2, Rcpp 0.11.6, reshape2 1.4.1, rj.gd 2.0.0-1, rpart 4.1-10, selectr 0.2-3, SparseM 1.6, splines 3.2.1, stringi 0.5-5, stringr 1.0.0, tcltk 3.2.1, testthat 0.10.0, tools 3.2.1

## Resources

The code and the figures included in this book are available at the book companion website: http://www.qualitycontrolwithr.com. The data sets used in the examples are available in the `SixSigma` package. Links and materials will be updated in a regular basis.

## About the Authors

The authors are members of the technical subcommittee AEN CTN66/SC3 at AENOR (Spanish member of ISO), with Mariano Prieto as the president of such committee.

**Emilio L. Cano** is Adjunct Lecturer at the University of Castilla-La Mancha and Research Assistant Professor at Rey Juan Carlos University. He also collaborates with the Spanish Association for Quality (AEC) as trainer for in-company courses. He has more than 14 years of experience in the private sector as statistician.

**Javier M. Moguerza** is Associate Professor in Statistics and Operations Research at Rey Juan Carlos University. He publishes mainly in the fields of mathematical programming and machine learning. Currently, he is leading national and international research ICT projects funded by public and private organizations. He belongs to the Global Young Academy since 2010.

**Mariano Prieto Corcoba** is Continuous Improvement Manager at ENUSA Industrias Avanzadas. He has 30 years of experience in the fields of nuclear engineering and quality. He collaborates with the Spanish Association for Quality (AEC) as trainer in Six Sigma methodology. Currently, he is president of the Subcommittee of Statistical Methods in AENOR.

Albacete, Spain — Emilio L. Cano
Madrid, Spain — Javier M. Moguerza
Madrid, Spain — Mariano Prieto Corcoba
July 2015

# Acknowledgments

We wish to thank Virgilio Gómez-Rubio for his kind foreword and the time devoted to reading the manuscript. We appreciate the gentle review of Iván Moya Alcón from AENOR on the ISO topics. We thank the Springer staff (Mark Strauss, Hannah Bracken, Veronika Rosteck, Eve Mayer, Michael Penn, Jay Popham) for their support and encouragement. A debt of gratitude must be paid to R contributors, particularly to the R core group (http://www.r-project.org/contributors.html), for their huge work in developing and maintaining the R project. We also acknowledge projects OPTIMOS 3 (MTM2012-36163-C06-06), PPI (RTC-2015-3580-7), and UNIKO (RTC-2015-3521-7), Content & Inteligence (IPT-2012-0912-430000) in which the methodology described in this book has been applied.

Last but not least, we are eternally grateful to our families for their patience, forgiving us for the stolen time. Thanks Alicia, Angela, Manuela, Beatriz, Helena, Isabel, Lucía, Pablo, and Sonia.

# Contents

# List of Figures

# List of Tables

# Acronyms

| | |
|---|---|
| AEC | Asociación Española para la Calidad |
| AENOR | Asociación Española de NORmalización y certificación |
| AHG | Ad Hoc Group |
| ANOVA | ANalysis Of VAriance |
| ANSI | American National Standards Institute |
| AQL | Acceptable (or Acceptability) Quality Level |
| ARL | Average Run Length |
| AWI | Approved Work Item |
| BSI | British Standards Institution |
| CAG | Chairman Advisory Group |
| CD | Committee Draft |
| CL | Center Line |
| CLI | Command Line Interface |
| CRAN | The Comprehensive R Archive Network |
| DBMS | DataBase Management System |
| DFSS | Design for Six Sigma |
| DIS | Draft International Standard |
| DoE | Design of Experiments |
| DPMO | Defects Per Million Opportunities |
| ESS | Emacs Speaks Statistics |
| EWMA | Exponentially Weighted Moving Average |
| FAQs | Frequently Asked Questions |
| FDA | Federal Drug Administration |
| FDIS | Final Draft International Standard |
| FOSS | Free and Open Source Software |
| GUI | Graphical User Interface |
| ICS | International Classification for Standards |
| IDE | Integrated Development Environment |
| IEC | International Electrotechnical Council |
| IQR | Interquartile range |
| ISO | International Standards Organization |

| | |
|---|---|
| JTC | Joint Technical Committee |
| LCL | Lower Control Limit |
| LL | Lower Limit |
| LSL | Lower Specification Limit |
| LT | Long Term |
| MAD | Median Absolute Deviation |
| MDB | Menus and Dialog Boxes |
| MR | Moving Range |
| NCD | Normal Cumulative Distribution |
| OBP | Online Browse Platform (by ISO) |
| OC | Operating Characteristic (curve) |
| ODBC | Open Database Connectivity |
| OS | Operating System |
| PAS | Publicly Available Specification |
| PLC | Programmable Logic Controller |
| PMBoK | Project Management Base of Knowledge |
| QC | Quality Control |
| QFD | Quality Function Deployment |
| RCA | Root Cause Analysis |
| RNG | Random Number Generation |
| RPD | Robust Parameter Design |
| RSS | Really Simple Syndication |
| RUG | R User Group |
| SC | Subcommittee |
| SDLC | Software Development Life Cycle |
| SME | Small and Medium-sized Enterprise |
| SPC | Statistical Process Control |
| ST | Short Term |
| SVM | Support Vector Machine |
| TC | Technical Committee |
| TMB | Technical Management Board |
| TR | Technical Report |
| TS | Technical Specification |
| UCL | Upper Control Limit |
| UL | Upper Limit |
| URL | Uniform Resource Locator |
| USL | Upper Specification Limit |
| VoC | Voice of the Customer |
| VoP | Voice of the Process |
| VoS | Voice of Stakeholders |
| WD | Working Draft |
| XML | eXtended Markup Language |