

Introduction to Nonparametric Statistics for the Biological Sciences Using R

Thomas W. MacFarland • Jan M. Yates

Introduction to Nonparametric Statistics for the Biological Sciences Using R

 Springer

Thomas W. MacFarland
Office of Institutional Effectiveness
Nova Southeastern University
Fort Lauderdale, FL, USA

Jan M. Yates
Abraham S. Fischler College of Education
Nova Southeastern University
Fort Lauderdale, FL, USA

ISBN 978-3-319-30633-9 ISBN 978-3-319-30634-6 (eBook)
DOI 10.1007/978-3-319-30634-6

Library of Congress Control Number: 2016934853

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

This text is about the use of nonparametric statistics for the biological sciences and the use of R to support data organization, statistical analyses, and the production of both simple and publishable graphics. Nonparametric techniques have a role in the biological sciences, and R is uniquely positioned to support the actions needed to accommodate biological data and subsequent hypothesis-testing and graphical presentation.

Introduction to Nonparametric Statistics for the Biological Sciences Using R begins with a general discussion of data, specifically the four commonly listed data types: nominal, ordinal, interval, and ratio. This discussion is critical to this text given the frequent use of nominal and ordinal data using nonparametric statistics. The beginning presentation then moves to an introductory display of R, with a caution that far more detail in the use of R and specifically R syntax is covered in later chapters.

The remaining chapters are largely self-contained lessons that cover the following individual nonparametric tests, listed here in the order of presentation in the book:

- Sign Test
- Chi-square
- Mann-Whitney U Test
- Wilcoxon Matched-Pairs Signed-Ranks Test
- Kruskal-Wallis H-Test for Oneway Analysis of Variance (ANOVA) by Ranks
- Friedman Twoway Analysis of Variance (ANOVA) by Ranks
- Spearman's Rank-Difference Coefficient of Correlation
- Binomial Test
- Walsh Test for Two Related Samples of Interval Data
- Kolmogorov-Smirnov (K-S) Two-Sample Test
- Binomial Logistic Regression

A common approach is used for each nonparametric analysis, promoting a consistent and thorough attempt at analyses: background on the lesson, the importing of data into R, data organization and presentation of the Code Book, initial

visualization of the data, descriptive analysis of the data, the statistical analysis, and interpretation of outcomes in a formal summary. Most chapters have additional lessons, listed in an addendum, and many chapters have multiple addenda.

This text should help beginning students and researchers consider the use of nonparametric approaches to analyses in the biological sciences. With R used as a platform for presentation, the diligent reader will develop a reasonable level of expertise with the R language, aided by the clearly shown syntax in an easy-to-read fixed format font.

Additionally, all datasets are available on the publisher's Web page for this text. Each dataset is presented in .csv (i.e., comma-separated values) file format, facilitating simple use and universal availability, regardless of selected operating system and computing platform. The subject matter for these datasets is fairly general and should apply as useful examples to all disciplines in the biological sciences.

A parametric approach to biologically oriented statistical analyses is frequently seen in the literature. However, as presented throughout this text, a nonparametric approach should also receive consideration when there are concerns about scale, distribution, and representation. That is to say, nonparametric statistics provide a useful purpose for inferential analyses when data (1) do not meet the purported precision of an interval scale, (2) there are serious concerns about extreme deviation from normal distribution, and (3) there is considerable difference in the number of subjects for each breakout group.

Consider the importance of each condition from the three conditions listed above and why a nonparametric approach should be considered, either as an exploratory approach to statistical testing, a final approach to statistical testing, or at least as a confirming approach to statistical testing.

- **Scale:** Many nonparametric analyses are based on ranked data, where the scale used to define data may not be as precise as desired. Given the realities of field work in the biological sciences, there are many times when it is not possible to obtain a precise measure (i.e., a measure that uses a scale that is both reliable and valid). Instead, field staff may only be able to obtain measures such as (1) large, medium, or small; (2) successful or not successful; etc. When precise measures are lacking, data that are instead ranked can be applied to good effect through the use of nonparametric analyses.
- **Distribution:** As many biologically focused research projects are put into place, it often becomes only too evident that the sample in question not only does not follow normal distribution patterns for selected variables, but the measurements do not even begin to approximate any semblance of normal distribution. Nonparametric techniques are extremely valuable when distribution patterns come into question, since many nonparametric tests are based on the use of ranks and are distribution-free (i.e., selected nonparametric tests are often quite appropriate even when data from the sample do not meet expected distribution patterns typically associated with a normally distributed population).

- **Representation:** There are many situations when there are extreme differences in the number and corresponding percent of total for breakout groups when samples are drawn from a population. Consider the representation of blood types. In the United States, there is extreme variation in the expected representation of blood type, such that O-positive is an expected blood type for nearly 40 % of the population, whereas AB-negative is a rare blood type and is observed for only 1 %, or less, of the population. This difference in representation by blood type is so extreme that comparisons of some measured variable by the two blood types would be greatly compromised in most cases, unless a nonparametric approach was used for later inferential analyses.

Although many nonparametric analyses were developed back when nearly all analyses were attempted using paper and pencil, it is now common to use a computer-mediated approach with contemporary statistical analysis software. This text is based on the use of R for this purpose. The R programming language is freely available open source software that it is now among the top 10 programs for worldwide use. R has gained wide acceptance due to its flexibility for data organization and data management, statistical analysis, and production of graphical images portraying relationships between and among data.

The comparative advantage of R is not only its functionality, which is also found to a degree in other computer-based programs; but, instead, the comparative advantage of R is the user community, where interested individuals can develop and use functions that operate on data for specific purposes and these actions are self-initiated, with no interference by a manager-led development team or marketing staff members. With R, a researcher has control over the data in ways that cannot be equaled when using commercial software that can be limiting to the imagination.

However, a limited degree of functionality is available when R is first downloaded. The extreme functionality comes from the more than 5000 packages available to the worldwide R community, with many packages having 25, 50, 100, or more functions. Again, the R data-centric environment is free and the R software is open source, such that the use of R is only limited by vision and skills. Functions developed by others are made freely available and the functions can be modified as desired.

Fort Lauderdale, FL, USA

Thomas W. MacFarland
Jan M. Yates

Contents

1	Nonparametric Statistics for the Biological Sciences	1
1.1	Background on This Lesson	1
1.2	Data Types	2
1.2.1	Nominal Data	3
1.2.2	Ordinal Data	4
1.2.3	Interval Data	4
1.2.4	Ratio Data	5
1.3	How R Syntax, R Output, and Graphics Show in This Text	5
1.4	Graphical Presentation of Populations	6
1.4.1	Samples that Exhibit Normal Distribution	7
1.4.2	Samples That Fail to Exhibit Normal Distribution	9
1.5	R and Nonparametric Analyses	11
1.5.1	Precision of Scales: Ordinal vs Interval	11
1.5.2	Deviation from Normal Distribution	12
1.5.3	Sample Size and Possible Issues with Representation	17
1.6	Definition of Nonparametric Analysis	23
1.7	Statistical Tests and Graphics Associated with Normal Distribution	25
1.8	Addendum: Data Distribution and Sampling	30
1.9	Prepare to Exit, Save, and Later Retrieve This R Session	50
2	Sign Test	51
2.1	Background on This Lesson	51
2.1.1	Description of the Data	51
2.1.2	Null Hypothesis (Ho)	54
2.2	Data Entry by Copying Directly into a R Session	54
2.3	Organize the Data and Display the Code Book	57
2.4	Conduct a Visual Data Check	60
2.5	Descriptive Analysis of the Data	63
2.6	Conduct the Statistical Analysis	73
2.7	Summary	74

- 2.8 Prepare to Exit, Save, and Later Retrieve This R Session 76
- 3 Chi-Square** 77
 - 3.1 Background on This Lesson..... 77
 - 3.1.1 Description of the Data..... 78
 - 3.1.2 Null Hypothesis (Ho) 80
 - 3.2 Data Import of a .csv Spreadsheet-Type Data File into R 80
 - 3.3 Organize the Data and Display the Code Book 82
 - 3.4 Conduct a Visual Data Check 84
 - 3.5 Descriptive Analysis of the Data..... 90
 - 3.6 Conduct the Statistical Analysis 92
 - 3.7 Summary 97
 - 3.8 Addendum: Calculate the Chi-Square Statistic
from Contingency Tables..... 100
 - 3.9 Prepare to Exit, Save, and Later Retrieve This R Session 102
- 4 Mann–Whitney U Test** 103
 - 4.1 Background on this Lesson..... 103
 - 4.1.1 Description of the Data..... 104
 - 4.1.2 Null Hypothesis (Ho) 106
 - 4.2 Data Import of a .csv Spreadsheet-Type Data File into R 106
 - 4.3 Organize the Data and Display the Code Book 108
 - 4.4 Conduct a Visual Data Check 111
 - 4.5 Descriptive Analysis of the Data..... 118
 - 4.6 Conduct the Statistical Analysis 125
 - 4.7 Summary 128
 - 4.8 Addendum: Stacked Data vs Unstacked Data 129
 - 4.9 Prepare to Exit, Save, and Later Retrieve this R Session 132
- 5 Wilcoxon Matched-Pairs Signed-Ranks Test** 133
 - 5.1 Background on this Lesson..... 134
 - 5.1.1 Description of the Data..... 134
 - 5.1.2 Null Hypothesis (Ho) 136
 - 5.2 Data Import of a .csv Spreadsheet-Type Data File into R 137
 - 5.3 Organize the Data and Display the Code Book 139
 - 5.4 Conduct a Visual Data Check 141
 - 5.5 Descriptive Analysis of the Data..... 150
 - 5.6 Conduct the Statistical Analysis 158
 - 5.7 Summary 160
 - 5.8 Addendum 1: Stacked Data and the Wilcoxon
Matched-Pairs Signed-Ranks Test 163
 - 5.9 Addendum 2: Similar Functions from Different Packages 167
 - 5.10 Addendum 3: Nonparametric vs Parametric
Confirmation of Outcomes 172
 - 5.11 Prepare to Exit, Save, and Later Retrieve this R Session 174

- 6 Kruskal–Wallis H-Test for Oneway Analysis of Variance (ANOVA) by Ranks** 177
 - 6.1 Background on this Lesson..... 178
 - 6.1.1 Description of the Data..... 178
 - 6.1.2 Null Hypothesis (Ho) 181
 - 6.2 Data Import of a .csv Spreadsheet-Type Data File into R 181
 - 6.3 Organize the Data and Display the Code Book 183
 - 6.4 Conduct a Visual Data Check 190
 - 6.5 Descriptive Analysis of the Data..... 197
 - 6.6 Conduct the Statistical Analysis 206
 - 6.7 Summary 207
 - 6.8 Addendum: Comparison of Kruskal–Wallis Test Differences by Multiple Breakout Groups..... 208
 - 6.9 Prepare to Exit, Save, and Later Retrieve this R Session 211
- 7 Friedman Twoway Analysis of Variance (ANOVA) by Ranks** 213
 - 7.1 Background on This Lesson..... 214
 - 7.1.1 Description of the Data..... 214
 - 7.1.2 Null Hypothesis (Ho) 218
 - 7.2 Data Import of a .csv Spreadsheet-Type Data File into R 218
 - 7.3 Organize the Data and Display the Code Book 220
 - 7.4 Conduct a Visual Data Check 223
 - 7.5 Descriptive Analysis of the Data..... 230
 - 7.6 Conduct the Statistical Analysis 236
 - 7.7 Summary 239
 - 7.8 Addendum: Similar Functions from External Packages 240
 - 7.9 Prepare to Exit, Save, and Later Retrieve This R Session 247
- 8 Spearman’s Rank-Difference Coefficient of Correlation** 249
 - 8.1 Background on This Lesson..... 250
 - 8.1.1 Description of the Data..... 250
 - 8.1.2 Null Hypothesis (Ho) 253
 - 8.2 Data Import of a .csv Spreadsheet-Type Data File into R 253
 - 8.3 Organize the Data and Display the Code Book 254
 - 8.4 Conduct a Visual Data Check 261
 - 8.4.1 Use of the Graphics Package 262
 - 8.4.2 Use of the Lattice Package 269
 - 8.4.3 Use of the ggplot2 Package 272
 - 8.5 Descriptive Analysis of the Data..... 275
 - 8.6 Conduct the Statistical Analysis 282
 - 8.7 Summary 294
 - 8.8 Addendum: Kendall’s Tau..... 295
 - 8.9 Prepare to Exit, Save, and Later Retrieve This R Session 297

- 9 Other Nonparametric Tests for the Biological Sciences 299**
 - 9.1 Binomial Test 300
 - 9.2 Walsh Test for Two Related Samples of Interval Data..... 303
 - 9.3 Kolmogorov-Smirnov (K-S) Two-Sample Test 308
 - 9.4 Binomial Logistic Regression..... 312
 - 9.5 Prepare to Exit, Save, and Later Retrieve This R Session 324
 - 9.6 Future Applications of Nonparametric Statistics..... 325
 - 9.7 Contact the Authors 326

- Index..... 327**

List of Figures

Fig. 1.1	Histogram and density plot: normal distribution	8
Fig. 1.2	Histogram and density plot: failure to meet normal distribution ...	10
Fig. 1.3	Stacked bar plot of two object variables	14
Fig. 1.4	Multiple density plots	19
Fig. 1.5	Histogram, density plot, and Quantile-Quantile plot: normal distribution.....	29
Fig. 1.6	Throwaway histogram	32
Fig. 1.7	Throwaway histograms showing multiple nclass declarations	33
Fig. 1.8	Histogram showing a rug along the X axis	34
Fig. 1.9	Density plot	35
Fig. 1.10	Multiple graphing curves in one figure	36
Fig. 1.11	Boxplot and violin plot in one figure	36
Fig. 1.12	Histogram and normal curve overlay	38
Fig. 1.13	Embellished histogram and normal curve overlay	39
Fig. 1.14	Quantile-Quantile (i.e., QQ or Q-Q) plot	40
Fig. 1.15	Histogram and Quantile-Quantile plot	43
Fig. 1.16	Detailed histograms.....	45
Fig. 1.17	Embellished histogram with multiple legends.....	47
Fig. 1.18	Quantile-Quantile plot with noise showing in the tails	48
Fig. 1.19	Multiple embellished histograms	50
Fig. 2.1	Bar chart using the <code>epicalc::tab1()</code> function	63
Fig. 2.2	Sorted dotplot using the <code>epicalc::summ()</code> function.....	69
Fig. 2.3	QQ plots comparing two separate object variables.....	73
Fig. 3.1	Mosaic plot using the <code>ved::mosaic()</code> function	85
Fig. 3.2	Side-by-side bar plot of two separate object variables	89
Fig. 4.1	Boxplot using the <code>lattice::bwplot()</code> function.....	113
Fig. 4.2	Comparative density plots using the <code>lattice::densityplot()</code> function	116

Fig. 4.3 Comparative density plots using the `sm::sm.density.compare()` function 117

Fig. 5.1 Comparative boxplots of separate object variables in one common graphic 145

Fig. 5.2 Comparative density plots of separate object variables in one common graphic 147

Fig. 5.3 Comparative histograms, normal curves, and density curves of separate object variables using the `descr::histkdnc()` function placed into one common graphic 148

Fig. 5.4 Comparative QQ plots with QQ lines 158

Fig. 6.1 Frequency distribution of four breakout groups using the `epicalc::tab1()` function 188

Fig. 6.2 Multiple (two rows by two columns) density plots using the `which()` function for Boolean selection 190

Fig. 6.3 Multiple (one row by two columns) density plots using the `which()` function for Boolean selection 191

Fig. 6.4 Boxplots of four breakout groups using the `lattice::bwplot()` function with emphasis on outliers 194

Fig. 6.5 Boxplots of two breakout groups using the `lattice::bwplot()` function with emphasis on outlines 194

Fig. 6.6 Color-coded sorted dot plots of four breakout groups using the `epicalc::summ()` function 199

Fig. 6.7 Multiple bar plots in one graphic based on enumerated values 202

Fig. 6.8 Multiple side-by-side QQ plots based on use of the `with()` function for Boolean selection 205

Fig. 7.1 Simple density plot of a single object variable 225

Fig. 7.2 Box plot with descriptive enumerated legends 225

Fig. 7.3 Multiple violin plots using the `UsingR::simple.violinplot()` function 228

Fig. 7.4 Color-coded sorted dot plots of five breakout groups using the `epicalc::summ()` function 232

Fig. 7.5 Interaction plot of median values for multiple object variables 239

Fig. 7.6 Sum of ranks comparison bar plots of breakout groups using the `agricolae::bar.group()` function 243

Fig. 7.7 Boxplot of breakout groups using the `descr::compmeans()` function 247

Fig. 8.1 Comparative box plots of separate object variables 266

Fig. 8.2 Multiple scatter plots of separate object variables placed into one graphical figure 268

Fig. 8.3 Box plots of two breakout groups using the `lattice::bwplot()` function 271

Fig. 8.4 Scatter plot of two continuous object variables using the `ggplot2::ggplot()` function 275

Fig. 8.5 Multiple QQ plots in one graphic, to compare distribution patterns 283

Fig. 8.6 Scatter plot of two continuous object variables with a legend showing Spearman’s rho statistic 285

Fig. 8.7 Scatter plot matrix (SPLOM) showing only the lower panel 287

Fig. 8.8 Color-gradient correlation plot of four continuous object variables using the `psych::cor.plot()` function 289

Fig. 8.9 Bagplot of two continuous object variables using the `aplpack::bagplot()` function..... 290

Fig. 9.1 Histogram of binomial probability 302

Fig. 9.2 Comparative density plots with color-coded legend 306

Fig. 9.3 Simple comparison of two side-by-side density plots..... 310

Fig. 9.4 Simple frequency distribution of two breakout groups 316

Fig. 9.5 Density plot of M1: original scale 100–200 316

Fig. 9.6 Density plot of M2: original scale 2.00–4.00..... 317

Fig. 9.7 Scatter plot of M1 and M2 317

Fig. 9.8 Scatter plot with box plots on X axis and Y axis using the `car::scatterplot()` function..... 318

Fig. 9.9 Cumulative probability (0.0–1.0) plot 318

Fig. 9.10 Conditional density plot 319