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

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

List of Figures