SpringerBriefs in Statistics

For further volumes:
http://www.springer.com/series/8921

Thomas W. MacFarland

Two-Way Analysis of Variance

Statistical Tests and Graphics Using R

Springer

Thomas W. MacFarland
Office for Institutional Effectiveness
Nova Southeastern University
Fort Lauderdale, FL, USA
tommac@nova.edu

ISSN 2191-544X e-ISSN 2191-5458
ISBN 978-1-4614-2133-7 e-ISBN 978-1-4614-2134-4
DOI 10.1007/978-1-4614-2134-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011943305

Contents

Chapter 1
Learn R with Sample Lessons in Education and the Social Sciences, Health, and the Biological Sciences

Abstract R was developed in the early-to-mid 1990s and it has developed into a robust open source software environment, used with multiple operating systems for the organization, statistical analysis, and graphical presentation of data. As presented in this chapter, this set of lessons provides an introduction to the use of R and emphasizes good programming practices for data import or data entry, data organization, development of a detailed code book, visual data checks, descriptive analyses, and selected inferential analyses. Two-way analysis of variance (ANOVA) is the selected inferential test emphasized in this set of lessons, given how two-way ANOVA is well-suited to investigations that examine factorial designs that model complex real-world problems with practical applications. This chapter also introduces the multiple ways datasets are organized for use by R: tab-separated ASCII files, fixed width–fixed column ASCII files, and comma-separated values ASCII files. Throughout these discussions, graphical displays, and other actions that relate to quality assurance practices receive continual reinforcement.

1.1 Purpose of These Lessons

This set of lessons has been developed to take advantage of R and the many features available through this extensive and expanding open source environment. The purpose of this set of lessons is to provide directional guidance, with graphical reinforcement, for those who wish to use R to examine data, conduct statistical analyses, and present findings in graphical format.

Along with instruction on the use of R and R syntax associated with Two-Way analysis of variance (ANOVA), these lessons will also reinforce the use of descriptive statistics and graphical figures, to complement outcomes from two-way ANOVA. With attention to these lessons as well as other available resources, an interested student or beginning researcher should be able to use R to conduct and graphically display simple to complex statistical analyses. The R user community is growing, for good reason. R works, R is free, and R is fairly easy to learn.

T.W. MacFarland, *Two-Way Analysis of Variance: Statistical Tests and Graphics Using R*, SpringerBriefs in Statistics 1, DOI 10.1007/978-1-4614-2134-4_1,

1.2 Background on R

R is an open source software environment that is used for the organization, statistical analysis, and graphical presentation of data. Because R is offered as freeware, it is available at no direct monetary cost. In contrast, many proprietary software packages that provide similar or even fewer features may have an annual license cost of `$1,000.00 USD` or more.

R shares a heritage with both S and Scheme, previously developed programming languages that have useful applications for data organization and statistical analysis. The first iteration of the R environment was made available in the early-to-mid 1990s. Since then, the R user community has seen steady growth, with regular improvements to R and the concurrent development of supplemental library packages, expanding R far beyond its initial form. Updates to R are made frequently and are available through access to The Comprehensive R Archive Network (CRAN, http://cran.r-project.org/).

R is available for multiple operating systems, including Linux, Mac, UNIX, and Windows, with much parity and portability for all platforms. A long and growing set of supplementary packages for R is also available as freeware, to meet specific needs that are not included in the base package. As of September 2011, there were more than 3,000 R-specific packages available at no direct cost, merely by using resources posted at the CRAN Contributed Packages Website (http://cran.r-project.org/web/packages/).

1.3 Background on Two-Way ANOVA

A common statistical technique to determine if differences exist between multiple groups (however, the concept of group is defined) is one-way ANOVA and the associated *F* test. The *F* test and subsequent one-way ANOVA methodology involve the determination of differences for: (1) one group with multiple (typically, three or more) variations, as well as (2) one variable, compared to multiple (typically, three or more) groups.

Research designs are often far more complex, however, than merely determining if differences to a measured object variable exist from the perspective of one factorial (e.g., categorical) object variable. Consider a concept as complex as weight and, more specifically, weight gain of lab rats. Is weight gain an outcome of food intake (grams of food intake) only, or do categorical variables such as food quality (e.g., high protein diet, moderate protein diet, and low protein diet) and food palatability (e.g., high palatability, moderate palatability, and low palatability), etc. also impact weight gain?

To address this complexity, the two-way ANOVA statistical test is often used to determine differences (and possible interactions) when variables are presented from

the perspective of two or more categories. When two-way ANOVA is used, it is possible to determine:

- Is there a difference because of variables acting independently of each other? In the above example, it would be useful to know if there is a difference in weight gain of lab rats because of (1) food quality or (2) food palatability.
- Is there a difference because of joint effects (e.g., interaction)? Again, in the above example, it would be useful to know if there is a difference in weight gain of lab rats because of interaction between (1) food quality and (2) food palatability.

Two-way ANOVA designs can become quite complex, not only to design but also to interpret. Yet, this highly useful methodology should not be avoided merely because it is not as simple as other statistical tests. On the contrary, two-way ANOVA should be used perhaps more than it is, due to the advantage of greater use of resources while modeling real-world scenarios.

Two-way ANOVA designs are often presented in a manner similar to other factorial analyses, such as the χ-square analysis. Like the χ-square analysis, two-way ANOVA uses a factorial organization with data placed in cells. The information within each cell provides the necessary data for later analysis. Thus, when using a two-way ANOVA, it is possible to examine three separate hypotheses: (1) Are the means for Variable *A* equal to the population? (2) Are the means for Variable *B* equal to the population? (3) Is there interaction between Variable *A* and Variable *B*?

To go back to the prior example of weight gain for lab rats, consider a scenario where (1) the measured object variable is weight gain (grams) of lab rats over an otherwise unspecified period of time, (2*A*) factorial object variable *A* represents the level of food quality (e.g., high protein diet, moderate protein diet, and low protein diet), and (2*B*) factorial object variable *B* represents the level of food palatability (e.g., high palatability, moderate palatability, and low palatability). In this example, a two-way ANOVA could be used to address the following questions:

- Is there a statistically significant difference in mean weight gain for the three protein levels? Intuitively, it would be reasonable to think that a high- protein diet would result in greater weight gain than would be experienced with a low-protein diet, but do the data support this assumption?
- Is there a statistically significant difference in mean weight gain for the three palatability levels? Intuitively, it would be reasonable to think that a highly palatable diet of easily digestible food would result in greater weight gain than would be experienced with a diet of low palatability, but do the data support this assumption?

Most importantly for the use of two-way ANOVA, is there any interaction between protein levels and palatability in terms of weight gain over time? Given the high cost of animal feed, it is best to recognize the possibility of interaction between food quality (high-protein feed is usually more expensive than low-protein feed) and food palatability (feed that is easily digestible is usually more expensive than feed that is of lower digestibility). With thousands of livestock typically kept

at any given time in a feeder lot, even small differences in feed costs v weight gain can have a major impact on financial return and two-way ANOVA can help support investigations into possible pricing models.

Given this background, two-way ANOVA is often used to help us explain real-world scenarios, where interaction is often found, or at least possible. These more complex designs are different from simplistic designs that can only explain scenarios designed for simplistic modeling. The decision to use a two-way ANOVA is the decision to see if complex issues can be understood, and possibly acted upon.

1.4 Organization of These Lessons

There will be three samples in this R-based lesson on the use of two-way ANOVA. With each sample building on the other, the general approach in this lesson is to present increasingly detailed examples that show:

1.4.1 Data Import or Data Entry

Typically, import the data into the R session. Data are often put into comma-separated values (.csv) file format, but R can also import tab-separated data and data that are also in fixed width–fixed column format. Or, enter the data, if the dataset is small, directly into the R session.

1.4.2 Data Organization

Are numerical data actually codes for various factors, such as 1 represents Female and 2 represents Male? If so, it may be necessary to coerce the numerical data into factor format. Are fairly cryptic codes used to identify data, such as 1 represents Alachua County, FL, up to 67 represents Washington County, FL, with the use of numerical codes for an otherwise alphabetical listing of Florida's 67 counties? If so, it may be necessary to either create new object variables or at least better identify the codes so that the meaning of the data is presented in plain language, providing a meaningful description of the data and object variables. R provides many possibilities on how data can and should be organized, with functions such as `as.numeric()` and `is.numeric()` used to support desired organizational outcomes.

1.4.3 Display the Code Book

Are object variable names self-documenting? The use of 1 or 2 for Gender may be open to confusion without a code book. Does 1 represent Female and does 2 represent Male, or does 1 represent Male and does 2 represent Female? Codes such as F and M, for gender, may be a better attempt at self-documentation, but even with these codes there may be a misunderstanding without an explicit set of declarations in a code book. Will lowercase f be accepted as an alternate presentation of uppercase F and will lowercase m be accepted as an alternate presentation of uppercase M? However, the code book is constructed, the codes should always be clearly detailed. It is also reasonable to identify expected ranges for *high* and *low* values, when appropriate, to help identify later values that may be either illogical or out-of-range. Even when the data seem self-explanatory, remember that data are often shared with others and as such, others may not have sufficient background knowledge to know all subtleties about the data and background surrounding the data. Consider the term Washington County as a datum. Does the term Washington County refer to Washington County in Alabama, Arkansas, Colorado, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Maine, Maryland, Minnesota, Mississippi, Missouri, Nebraska, New York, North Carolina, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, Tennessee, Texas, Utah, Vermont, Virginia, Wisconsin, or Washington Parish, Louisiana? Self-documentation is useful, but explicit documentation is more prudent.

1.4.4 Conduct a Visual Data Check

Graphics are an excellent way to check the data for values that are either illogical or out-of-range. All graphical images do not have to be of publishable quality. It is common, and encouraged, to use simple graphical images to visually review outcomes and to add another component to data quality assurance. Initially, a simple graphic in black and white with minimal detail may be all that is needed to visually check the data for illogical or out-of-range data. Then, if appropriate, a more detailed graphic of publishable quality can always be generated.

1.4.5 Descriptive Analysis of the Data

Descriptive statistics such as *N*, Mean, Standard Deviation (SD), Median, Range, etc., provide an excellent first view of the data. It would be the rare attempt at data analysis that did not include a full set of descriptive analyses, with the emphasis on statistics associated with measures of central tendency. Summary descriptive statistics provide an overall view of the data, with quantitative description of the

data at the broadest level. Breakout descriptive statistics provide even greater detail than summary descriptive statistics. As a typical example, when working with livestock and other biological specimens (including humans), it is often important to consider differences between the two genders, Female and Male. Breakout descriptive statistics are used to provide this level of detail.

Because these many activities require a fair amount of planning and time-on-task, it is essential to have a dataset that is correctly organized and understood. The outcomes of later statistical analyses can only be accepted if there is prior acceptance and understanding of the dataset. Indeed, it is not at all uncommon for experienced researchers to devote more time to dataset quality assurance tasks than the time devoted to analytics. Quality outcomes are best gained through attention to quality inputs.

1.5 Details of the Three Sample Datasets

1.5.1 Tab-Separated ASCII File

The first dataset is a collection of Software Engineering Final Examination scores (0–100) by (A) teaching method (1 = Lecture, 2 = CBT (Computer-based training), 3 = Video, 4 = IDS (independent study) and by (B) status as a Community College graduate (1 = is a Community College graduate, 2 = is not a Community College graduate). The dataset consists of fewer than 100 subjects and there are no missing data. The first dataset was prepared as a tab-separated ASCII file.

1.5.2 Fixed Width–Fixed Column ASCII File

The second dataset is a collection of measured data (e.g., height, weight, blood pressure, etc.) and responses to a Wellness Inventory (e.g., smoking habits, drinking habits, etc.). The dataset consists of more than 100 subjects and typical to the realities of a field-based activity where possibly intrusive questions and measurements are involved, data for a few individual responses are missing. The second dataset was prepared as a fixed width–fixed column ASCII file.

1.5.3 Comma-Separated Values ASCII File

The third dataset is a collection of data taken from an agricultural integrated pest management (IPM) study. Data are all numeric and represent the three object variables of interest to this study: formulation of the agricultural chemical (three

factors), time-of-day for chemical application (two factors), and the number of larvae for a specific insect per square meter at random locations one week after chemical application. The dataset has no missing data and there is an equal number of measurements for each cell. The third dataset was prepared in .csv format, as a comma-separated values ASCII file.

Presentation of how R is used against these three datasets goes from simple to detailed. The first dataset, based on a student learning outcomes assessment in Education, is examined in a fairly simple manner. Complete analyses and useful graphics are provided, but in an attempt to provide simple examples, complexity is kept to a minimum. The second dataset, based on a Health Science Wellness Inventory, introduces a moderate degree of complexity. There are some challenges to the dataset and graphical presentations increase in detail and complexity. The third dataset, using data from the Biological Sciences, is examined in complete detail. Details are examined closely and graphical presentation approaches publishable quality. By following along with this measured level of increased detail, self-confidence in the use of R is developed and new R functions are introduced with each lesson. Be sure to examine all three lessons.