An Overview of the R Scripts

Introduction

The material at this website contains 6 files; four correspond to the four parts of RDASA3 (3rd edition of "Research Design and Statistical Analysis," Myers, Well, & Lorch, 2010). A fifth script contains analyses for mixed models and logistic regression. A sixth script, "Functions.R" contains commands for installing several packages from the cran.R project website, and functions that I have written or copied from other websites, and that are not included in the packages downloaded from the cran.R project website,

The scripts and data sets are at my R website,
https://udrive.oit.umass.edu/jlmyers/MyRWebsite/

To open or download the script files: At "MyRWebsite," click on "open web view." This allows access to a new screen from which files can be downloaded. In either Internet Explorer or Firefox,to download scripts or data files, open the relevant folder (double left-click) to access the list of files in the folder. Then, double-click on the file you want, and choose among the options provided (open or save). In these scripts, references to "my R website" indicate the above (udrive) URL My R website also contains a pdf file that contains an introduction to the R language, with references to sources providing further help.

One purpose of the scripts is to provide R functions to create graphics and do the appropriate statistical analyses. Thus, for individuals familiar with R, much of this may merely save some time and effort in generating the necessary code. A second purpose is to provide enough examples and commentary (comments begin with #) so that individuals who have not previously worked with R can begin to do so. However, R is so rich in functions that there are many ways to accomplish a given analysis and users may find more efficient alternatives. In some cases, several steps have been intentionally introduced where fewer would do in order to illustrate certain functions.

The Scripts

Functions.R

When R is downloaded from the CRAN website, a ;library of base packages containing many useful functions is included. However, in analyzing and plotting the data from RDASA3, I found many packages to be useful that were not in the base set. Commands to install these are in this script. In addition, I wrote functions and found some on the internet that were not in the base packages but proved useful. These also are contained in this script. Running this script before the others will expedite the use of the scripts we next describe.

The functions in "Functions.R" are organized by Session. With the R console open, select a function in the "Functions.R" script and run it. This stores it in the R workspace, making it available for execution by a single command. Or you can install all the functions into your workspace at once by opening the Functions.R script while the R console is open, click on the Edit menu, and then click on "Run all." Once these have been run, save the workspace so that the functions will be available for future sessions.

Part 1.R

Part 1.R provides scripts for the data plots, tables, and analyses for Chapters 2 - 7 of "Research Design and Statistical Analysis". Scripts are divided into sessions corresponding to chapters. *Session 2* contains data plots, including histograms, box, stem-andleaf, kernel density, and Q-Q plots, and bar and line graphs with standard error (or confidence interval) bars. There are also summary statistics, including mean, standard deviation, variance, percentiles, skewness and kurtosis with related hypothesis tests.

Session 3 contains commands to create 2 and 3-way frequency tables, and to calculate

marginal, joint, and conditional probabilities based on these tables.

Session 4 covers binomial probabilities, including generation of tables of probabilities, and hypothesis tests and power functions based on the binomial distribution. For all distributions in this and subsequent sessions, scripts contain commands for obtaining densities; cumulative probabilities; and values corresponding to those probabilities; and for random generation of data.

Session 5 enables analyses related to the normal distribution, including inferences based on large samples (z tests, power, and confidence intervals).

Session 6/7 illustrates functions relevant to the t distribution. the,2-independent-sample t test assuming homogeneous variances and not assuming homogeneous variances (Welch t test), the 1-sample (including correlated scores) t test, confidence limits for these cases, Cohen's *d* with confidence intervals, the power of the t test, and the trimmed t test.

Part 2,R

Part 2.R covers analyses in Chapters 8 - 12 of RDASA3. *Session 8* contains functions to perform the analysis of variance for between-subject one-factor designs, to calculate the power of the F test, to perform tests based on ranks, and also includes tests of homogeneity of variance, and power transformations of data when assumptions are violated.

Session 9 deals with the analysis of variance for multi-factor between-subjects designs, and includes functions to compare nested models.

Session 10 illustrates tests and confidence intervals for contrasts in between-subject designs, including cases where sample sizes and variances are unequal. The script also provides examples of how R functions construct contrast matrices and control of familywise error rates (e.g., Bonferroni, Tukey HSD, Scheffe's method). Functions are also illustrated that find Studentized Range (q) probabilities, or – given the probabilities – the corresponding critical value.

Session 11/12 illustrates functions to construct matrices of orthogonal polynomial coefficients, tests of polynomial trends for one- and multi-factor between-subjects designs, and commands to enable data fitting with polynomial functions.

Part 3,R

Part 3.R extends the developments of of the preceding sessions to more complex designs and analyses. *Session 13* introduces analyses for the simplest cases of several designs - treatments x blocks, repeated measures, and Latin squares – as well as the analysis of covariance (ANCOVA). The concept of design efficiency is illustrated through repeated sampling from a population. This script also contains a useful function for selecting a Latin square randomly from the population of squares of any given size.

Session 14 deals with repeated measures designs, including univariate and multivariate tests. The script first reviews methods for analysis of data when there is a single factor (other than subjects). It then illustrates a test for nonadditivity, a method of finding the best power transformation of nonadditive data, tests of contrasts in repeated measures designs, and several nonparametric tests.

Session 15 provides examples of analyses of repeated measures designs with two factors (besides subjects), one of which may be random. Quasi-F and minF analyses are illustrated. The script concludes with tests of contrasts as well as trend analyses within the context of the S x A x B design.

Session 16 extends these developments to nested (split-plot) and Latin square designs.

Part 4.R

Part 4.R deals with correlation and regression. *Session 18* provides an introduction to the topic, including functions for scatterplots and estimation of various statistics, such as regression and correlation coefficients. This session also illustrates functions to generate information about the distribution of residuals, and to calculate standard errors, as well as confidence intervals and significance tests. Influence plots are generated and leverage and Cook's distance are defined, providing ways of detecting data points that may have extreme influence on regression statistics.

Session 19 provides test of inferences about the population bivariate correlation, rho; tests of differences between two bivariate correlations, both independent and dependent cases; provides calculations of partial correlations, as well as significance tests and confidence intervals; calculates semipartial correlations; and calculates various other correlations (biserial, tetrachoric, point-biserial, phi, Spearman's rho, Kendall's tau), and illustrates tests of hypotheses.

Session 20 continues the presentation of R applications tin bivariate regression. The script illustrates R's generation of diagnostic plots to help assess the validity of assumptions (e.g., linearity, homoskedasticity), and to detect possible outliers. The effects of deleting such outliers are assessed. Tests of the assumptions of linearity, homoskedasticity, and normality are illustrated. Following evidence if heterosketasticity and nonnormality, a Box-Cox function is applied to obtain a data transformation. The script closes with an illustration of robust estimation of regression parameters, and tests of hypotheses about those parameters.

Session 21 introduces multiple regression. Using a multivariate data set, plots and calculations of summary statistics are illustrated. R functions to partition the regression sum of squares, and to calculate the variance-inflation factor and adjusted R-squared are applied to the data set.

Session 22 is concerned with inferences in multiple regression. Confidence intervals and tests of regression coefficients and R-squared are illustrated, as are tests comparing nested models. Predictions of individual scores are generated. This script also illustrates calculations of tests of differences between two multiple correlations, and power calculations in multiple regression, including determining sample size for various purposes. As in previous scripts, plots are generated to illustrate how R can enable checking the validity of assumptions.

Session 23 extends R analyses to polynomial regression and tests of interaction involving continuous predictors in multiple regression.

Session 24 focuses on categorical variables in multiple regression, showing how R can be applied to analyses of data from nonorthogonal designs. R functions are also applied to test interactions between categorical and continuous predictors.

Session 25 presents R functions for analysis of covariance (ANCOVA). The script includes coverage of adjusted group means, tests of contrasts, and tests following ANCOVA, as well as tests of assumptions underlying ANCOVA. Examples are provided of analyses of data sets with more than one covariate,

Part 5.R

Session 26 introduces packages and functions for mixed-effects model analyses. Such analyses employ maximum likelihood estimation, and use all of the data, negating the need for quasi-F tests or averaging over subjects or items. They are flexible and powerful. The script includes examples of alternative models for data sets, and the use of the *anova* function to compare the goodness of fit of competing models of the data.

Session 27 deals with logistic analysis. The script provides examples of R analyses of logistic regression, with both binary and multilevel outcomes; in the case of multilevel outcomes, both quantitative (e.g., numerical ratings) and qualitative (e.g., object names) outcomes are considered. There are also examples of analyses involving both categorical and continuous predictors. The script also contains examples of R functions to calculate odds ratios and their confidence bounds, and predicted response probabilities. Plots of predicted probabilities are generated.