American Fisheries Society
Student Subsection of the Education Section

How To Analyze Your Data

Overview

Data analysis is the formal process of deriving scientific conclusions from quantitative and/or qualitative information collected during a study. In fisheries, sound data analysis is critical for reliable research, high-quality scientific publications, and science-based management. Students are often exposed to analytical procedures in courses (e.g., statistics, biometry, quantitative fisheries science, fisheries population analysis) that prepare them for graduate research, for which statistical acumen is indispensable. Here we review the process of data analysis and offer tips for success.

Analyzing your data

Proper data management and an organized database are critical for efficiently analyzing data in any modern statistical software program. Before you begin your research projects, visualize example data entries in a spreadsheet program like Microsoft Excel or Access to help you realize what you want to measure, how you want to group your data, and how you want to check for errors. Data can be entered in long-form, where each entry are defined separately over multiple rows. For example, in transect sampling, long-form data would have each new point listed as a new row. Alternatively, in wide-form, each transect would be listed as a single row and the associated variables are presented across multiple columns. Many analyses require a specific format and most data manipulation software provide simple means for converting between forms. Understanding how you want to analyze your data and entering the data appropriately will save a great deal of effort in the long run.

Long-Form

Transect	Species	Substrate
1	B.Crappie	Sand
1	B.Crappie	Sand
1	Bluegill	Gravel
2	L.Bass	Cobble
2	B.Crappie	Cobble
3	Bluegill	Sand
4	L.Bass	Silt
4	L.Bass	Silt
4	L.Bass	Silt

Wide-Form

Transect	B. Crappie	Bluegill	L. Bass	Sand	Gravel	Cobble	Silt
1	2	1	0	2	1	0	0
2	1	0	1	0	0	2	0
3	0	1	0	1	0	0	0
4	0	0	3	0	0	0	3

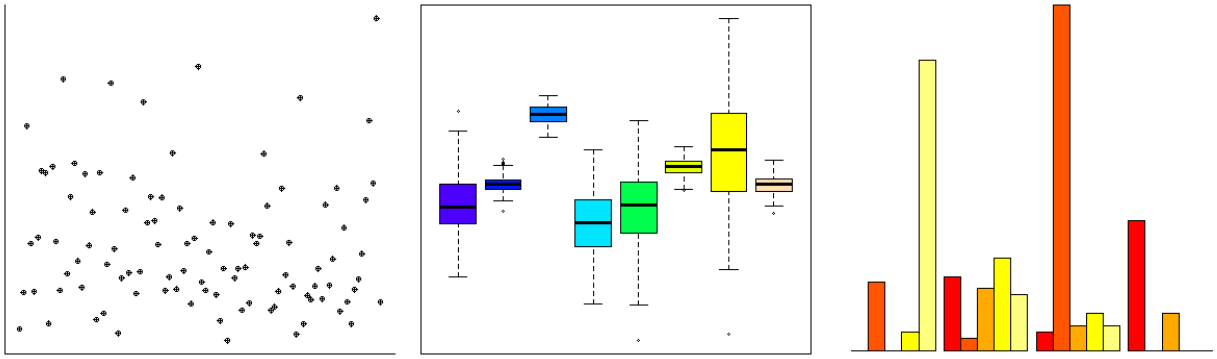
After projects are completed, decide on which statistical software program to use. Each program will have pros and cons. For example, SAS costs money but is verified by professional statisticians. JMP has a 'point-and-click' interface and may be easier to use for those disinclined to learn computer programming in addition to learning statistical and ecological concepts. R is

free, open-source, and uses an S based programming language. It is used for its wide-array of analytical and graphical capabilities (e.g., linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering). Unfortunately, not all of R's packages are verified by statisticians due to its open-source programming. The learning curve for R is steep, especially for users untrained in computer programming. Fortunately, R is commonly used and many experts offer free help online. Searching 'R help' along with any analytical problem or error will typically yield an expert's code and solution to surmounting your problem.
Here we provide an overview of common statistical programs used for fisheries data analysis:

1. R
    a. An open source statistical program with a wide variety of analytical and graphical capabilities (e.g., linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering)
    b. Users write codes to perform specific analyses (self-help resources are available: http://cran.r-project.org/doc/manuals/r-release/R-intro.html; http://www.statmethods.net/index.html)
    c. Available free online: http://www.r-project.org/
    d. Fisheries-specific applications: https://fishr.wordpress.com/
2. SAS (Statistical Analysis System)
    a. A software suite for data management, predictive analytics, and business intelligence with many applications for fisheries science
    b. Although it does not have the graphing capabilities of R, SAS features a user-friendly point-and-click interface
    c. Self-help resources available at: http://support.sas.com/techsup/
    d. Purchase online at: http://www.sas.com/en_us/software/how-to-buy.html
    e. Fisheries-specific applications: *Analysis and Interpretation of Freshwater Fisheries Data* (Christopher S. Guy and Michael L. Brown, Editors; https://fisheries.org/shop/55049c)
3. SPSS (Statistical Package for the Social Sciences)
    a. A statistical program widely used in the social sciences with applications for fisheries data analysis (e.g., descriptive statistics, linear regression, bivariate statistics, and classification methods)
    b. New mapping features add a geographic dimension to data analysis and reporting
    c. Self-help and purchasing information available at: http://www-01.ibm.com/software/analytics/spss/
4. PRIMER-E (Plymouth Routines In Multivariate Ecological Research),
    a. A program for analyzing multivariate (i.e., more than one variable) species or sample abundance (biomass) data
    b. Multivariate procedures include grouping, sorting, principle component identification, hypothesis testing, sample discrimination, and trend correlation
    c. Self-help and purchasing information available at: http://www.primer-e.com/
5. Excel
    a. A Microsoft spreadsheet program best suited for data organization, summary statistics (e.g., mean, standard deviation), and simple graphing
    b. Excel spreadsheets often represent data sets for import and analysis in other programs (e.g., R, SAS)

After determining the software of choice, read the data into your preferred software program using self-written code or a point-and-click interface. This can be done with many file types and most database management programs provide multiple options for exporting data. The most appropriate files depend on the software you are using. Text files (.txt) or comma delimited Excel files (.csv) are commonly used for many software programs. Understanding how the software program you are using is reading in data is extremely important to prevent costly mistakes later on.

We recommend conducting some exploratory analyses to calculate summary statistics (e.g., means, standard deviations, ranges) and use boxplots to visualize the data across a range of categories (e.g., plot animal densities across 10 different transects), or use X-Y scatter plots to visualize patterns in continuous data (e.g., plot animal densities across water temperature).



The next steps will be to use formal statistical procedures to achieve your project objectives. These objectives should be clearly delineated before you begin; this may require consulting with your research supervisors and collaborators. Common analyses are to use regression analysis to estimate the relationship between variables. Linear regressions, ANOVAs, logistic regression, and generalized linear models are all interrelated; the big differences are the assumed distribution of the dependent variables (e.g., binomial, normal, Poisson) or whether the predictor covariates are continuous (e.g., temperature), integers (e.g., number of animals), or categorical values (e.g., Lake X, Lake Y). Mixed-effects analyses are frequently used on the frontiers of fisheries and ecology because they help appropriately account for random error in your sampling design to draw meaningful inference on relationships between variables. This may be particularly helpful if you have a field project with an unbalanced sampling design, non-random sampling, or non-randomly chosen study sites. However, if your research is a well-controlled experiment or is truly random, you may not have much random error to account for.

Once your initial statistical tests are conducted, check whether the data (or corresponding residuals) violate assumptions of those tests. Commonly we assume our data is normal and residuals are homoscedastic (equal variance across the predictor covariates). Tests for normality include Shapiro-Wilk, skewness-kurtosis, *Q-Q* plots, stem-and-leaf plots. Homoscedasticity can be checked with Levene's test and Breusch-Pagan test, among others. If the data fulfill these assumptions or the analytical technique is robust to violations of these assumptions, you can conduct parametric statistical analyses. If not, you should perform analogous non-parametric procedures (see *Handbook of Parametric and Nonparametric Statistical Procedures* by David J.

Sheskin). If further problems emerge, please consult a graduate student or professor in your statistics department to discuss your difficulties. That is one of the many reasons we have statistics departments!

After formal analyses have been finished, interpret and visualize your data. A valuable self-help resource for data interpretation is *Analysis and Interpretation of Freshwater Fisheries Data* (Guy and Brown, 2007). The goal for this is to help focus your message to your intended audience to help them understand important patterns and processes in your data. This can include both descriptive patterns of the data (i.e., what did your results show within the bounds of your sampled population), but also prescriptive results or even model forecasting (e.g., extrapolating the relationships beyond the bounds of the sampled population into uncertain scenarios). Important questions to answer in data interpretation are: What were the interesting effects of your study? What does your audience *need* to know about your study? What results make your study important? Interpreting data according to these types of questions will help you identify relevant patterns, visualize results, and ultimately describe results with text, tables, and figures.

You can use a variety of programs to display data graphically. Communicating your analyzed data is extremely important to get across your findings through any medium (e.g. presentations, posters, publications). The book, *Grammer of Graphics* (Wilkinson, 2nd edition) provides a useful reference for visualizing statistical data and how it is interpreted by an audience. Sigma Plot is a point and click program for producing publication-quality figures (e.g., scatterplots, regressions, histograms, bar charts, box plots). Self-help, purchasing, and downloading information are available at http://www.sigmaplot.com/. Microsoft Excel allows for creating tables and simple graphics quickly. Finally, R offers an array of graphing capabilities within the base software and many external packages which allow for diverse and unique data visualization.

By:
Andrew Carlson, M.S. Candidate, South Dakota State University

Kyle Wilson, Ph.D. Student, University of Calgary

Nicholas Cole, Ph.D. Student, University of Nebraska

Image credits:

R: http://rprogramming.net/
SAS: http://en.wikipedia.org/wiki/SAS_Institute
SPSS: http://en.wikipedia.org/wiki/SPSS
Primer-E: http://www.primer-e.com/primer.htm
Excel: http://pubpages.unh.edu/~any45/excel.html