**R Lessons for undergraduate interns**

This series of hands-on lectures is designed to get undergraduate interns familiar with R and facilitate data analysis and visualization for their research projects. At QBC, we expect to fit at least one R instruction session (2-3 hours working time) in each week, depending on weather, tides, and oyster farm responsibilities. We would like to front-load the teaching to June as much as possible, so that all interns feel capable of working independently on data analysis once they have selected their projects.

We have so little time and such varied experience levels that we cannot take on the responsibility of being both a full introductory environmental statistics course and a full introductory R course. The statistics we will touch on should be more of a refresher than a comprehensive college-level course, so that we can focus more on getting comfortable with application in R. Not every student will utilize the full breadth of statistical tests we will cover, and there is room for one-on-one instruction for more complex methods or extra help if desired.

We hope that our students have previously taken an environmental statistics course and will lean on that knowledge. Students with no quantitative analysis experience will have a heavier workload but should still be able to produce a research paper with appropriate analysis and high-quality data visualizations, if with more hands-on assistance from the instructor. All R lessons provide detailed narrative on the statistical theory behind what we are doing. The resources provided in the summer lectures should be a great foundation for individual learning after the internship or as supplemental materials for college courses.

**Learning outcomes:**

* Installation of R and RStudio,
  + Basic R syntax, file structure navigation and setting a working directory, installing and loading packages
* Utilization of RMarkdown to produce HTML notebooks
* Descriptive statistics
  + Types of data, measures of center and spread
* Inferential statistics
  + Populations and samples, sampling distributions, hypothesis testing, the null hypothesis, types of error, confidence intervals
  + Parametric vs. non-parametric modeling assumptions
  + Categorical inference: chi-square goodness of fit, chi square test of independence
  + Numerical inference: correlation, t-tests, ANOVA (and non-parametric versions of these tests)
* Linear regression
  + Simple and multiple linear regression
* Data handling
  + Tidyverse packages for filtering, merging, and grouping data
* Utilization of ggplot2 package to produce visualizations and maps
  + Handling spatial data using the sf package
* Bonus content
  + TeX-RMarkdown integration for producing PDFs
  + GitHub, version control, and R projects.

**Modules / Lesson plans:**

1. Module 1: Introduction to R
   1. File: 1.1\_Intro\_to\_R.R
      1. Purpose: Install R and RStudio. Get interns comfortable with layout and basic syntax of writing and running a script in R. Explain how to access data outside of R (may require brief instruction on the basics of file systems, previous cohorts needed help navigating their systems). Describe data structures and types.
      2. Quantifiable goals: Functioning and updated installations of R and RStudio. Understanding that R is a tool and has no intelligence to infer and execute your intentions without explicit instruction.
      3. Estimated instruction time: 2 hours, expecting 1-on-1 installation troubleshooting.
   2. File: 1.2\_Packages.R
      1. Purpose: Emphasis on R’s nature as a flexible and open-source tool with modular ability to complete tasks built on open-source packages. Ensure that students understand that many of our most-used packages are external, only need to be installed once, but need to be loaded into the library every session. Use this opportunity to install two packages that we will use in every lesson, tidyverse and here. Familiarize students with setting a working directory using the here package.
      2. Quantifiable goals: Comprehension of working directories and package installation/loading process.
      3. Estimated instruction time: 1 hour
      4. BONUS: contains instruction on creating a GitHub account, forking the repository that contains all R lecture information, setting up a local project folder for real-time version control.
   3. File: Intro R Lab 1.doc
      1. Purpose: self-guided lab to build confidence in R. Will not be required. Basically a carbon copy of lecture 1, but with a different data set.
      2. Estimated completion time: 30 minutes
   4. Supplementary Data files
      1. baptisms.csv (Lesson 1.1 data)
      2. us\_baptisms.csv (Self-guided lab data)
      3. ANSWERS\_R\_Lab\_1.R (Self-guided lab answers)
2. Module 2: Introduction to R Markdown
   1. File: 2.1\_Intro\_to\_RMarkdown.RMD
      1. Purpose: Move from narrating in comment chunks in a script (unreadable) to narrating in text chunks in an RMarkdown html notebook (legible, incorporates narrative, input, and output all in one document).
      2. Quantifiable goals: Understand how to write both narrative and code chunks in an R Markdown html notebook. Remember to run all code chunks prior to saving for the last time.
      3. Estimated instruction time: 1 hour
      4. BONUS: Previous attempts at this lesson included installation of MikTeX/ MacTeX to knit results to PDF. This was generally unsuccessful because of the massive file size for MacTeX and the high usage rate of Macs among undergrad interns. We will no longer attempt to get this to work for all students, as outputting code as an html notebook is quicker and just as legible. Instruction on how to install MikTeX/ MacTeX/ BasicTeX (smaller version of MacTeX) has been updated and included at the bottom of the lesson document for students to view on their own.
   2. Supplementary Data files
      1. cdc.csv (Lesson 2.1 data)
3. Module 3: Statistical Methods
   1. File: 3.1\_Descriptive\_Statistics.RMD
      1. Purpose: Introduce types of data (categorical, numeric). Understand, describe, and visualize measures of center and spread. Save output to RData file. Introduction of piping as a data handling method. Introduction of ggplot as visualization method.
      2. Quantifiable goals: Write successful data cleaning code using piping and successful visualization code using ggplot2.
      3. Estimated instruction time: 1.5 hours
   2. File: 3.2\_Hypothesis\_testing.RMD
      1. Purpose: Understand the difference between populations and samples, and how we use samples to predict what’s going on at the population level via random sampling and the central limit theorem. Description of basic hypothesis testing, types of error, and assumptions. Run a t-test and generate a 95% confidence interval around the true difference in means of two independent populations. Exposure to loops (will not be used again in written lessons, may be necessary for individual research projects).
      2. Quantifiable goals: Working knowledge of process to perform a hypothesis test: state null and alternative hypotheses, set confidence level, check assumptions, and run test.
      3. Estimated instruction time: 2 hours
   3. File: 3.3\_Regression.RMD
      1. Purpose: Identify linear relationships between quantitative variables. Assess the assumptions of and then calculate Pearson’s correlation coefficient. Assess the assumptions of and then calculate a linear regression. Compare models using AIC. Interpret results.
      2. Quantifiable goals: Working knowledge of process to assess relationships between two quantitative variables.
      3. Estimated instruction time: 3 hours
   4. File: 3.4\_Chi\_Square.RMD
      1. Purpose: Assess the assumptions of and then run chi-square tests of goodness of fit and independence. Interpret results.
      2. Quantifiable goals: Working knowledge of process to assess relationships between two categorical variables.
      3. Estimated instruction time: 2 hours
   5. File: 3.5\_ANOVA.RMD
      1. Purpose: Assess the assumptions of and then run ANOVA and post-hoc Tukey HSD tests. Interpret results.
      2. Quantifiable goals: Working knowledge of process to assess relationships of quantitative and categorical variables.
      3. Estimated instruction time:
   6. File: 3.6\_Non-Parametrics.RMD
      1. Purpose: Introduce non-parametric methods of calculating correlation, comparison of means of two independent samples, comparisons of means of three or more independent samples, and Tukey’s HSD post-hoc test for datasets that violate assumptions of normality and heteroskedasticity.
      2. Quantifiable goals: Assess normality and equality of variance as appropriate for statistical test. Working knowledge of Spearman rank correlation, Mann-Whitney U test, Kruskal-Wallis, and Dunn tests.
      3. Estimated instruction time: 2 hours
   7. Supplementary Data files
      1. descriptive\_stats\_results.R (Lesson 3.1 output, used in Lesson 3.2)
      2. correlation\_rules.png (Illustration of Pearson correlation rules of thumb)
      3. pearson\_correlation.png (Mathematical formula of Pearson correlation)
      4. credit\_backup.csv (Lessons 3.1 and 3.2 data in case hosting website fails)
4. Module 4: Data Handling
   1. File: 4.1\_Data\_Handling\_and\_Analysis.RMD
      1. Purpose: Real-world application of statistical methods we have learned. Advanced data cleaning. From a research question, make hypotheses, select the appropriate statistical test to address that question, assess the assumptions, run the test, interpret the results, and discuss ecological significance.
      2. Quantifiable goals: A miniature walk-through of developing their own summer research project. Should stimulate thinking about what statistical methods they can use for their chosen topic.
      3. Estimated instruction time: 2 hours
   2. Supplementary Data files
      1. trips\_cleaned.csv (Lesson 4.1 Data)
      2. bio\_inf\_cleaned.csv (Lesson 4.1 Data)
      3. abundance\_cleaned.csv (Lesson 4.1 Data)
5. Module 5: Visualization
   1. File: 5.1\_GGplot\_Livecode\_2023.R
      1. Purpose: We will do a live-coding demo of ggplot functionality with the palmerpenguins dataset. Interns can come up with visualization tweaks and we will code them on the fly. Final result may or may not be similar to what we did in 2023.
      2. Quantifiable goals: Familiarity with ggplot functions (geom\_boxplot, geom\_histogram, geom\_point, etc) and themes to make customized, high-quality data visualizations.
      3. Estimated instruction time: 1 hour
   2. File: 5.2\_Mapping\_in\_R.RMD
      1. Purpose: A brief introduction to creating maps in R. Most students have familiarity with GIS applications and may not need a detailed description of mapping concepts (coordinate reference system, extent and resolution, projects, etc). Lesson focuses on the functions needed to manipulate and plot spatial data in R, with use of the sf and ggplot2 packages.
      2. Quantifiable goals: Handle projects and transformations in R, plot spatial locations with ggplot2 and sf packages.
      3. Estimated instruction time: 1.5 hours
   3. Supplementary data files
      1. ggplot\_cheat\_sheet.pdf (Handout with ggplot functions)
      2. us\_coarse\_coastline\_poly.shp/. proj/ etc (Data for Lesson 5.2)
      3. us\_medium\_shoreline\_poly.shp/ .proj/ etc (Data for Lesson 5.2)
      4. oyster\_lease\_map.png (Output from Lesson 5.2)