

Week 1 Assignment: Basic R

Student Name; Z620: Quantitative Biodiversity, Indiana University

05 January, 2017

OVERVIEW

Week 1 Assignment introduces some of the basic features of the R computing environment (<http://www.r-project.org>). It is designed to be used along side your Week 1 Handout (hard copy). You will not be able to complete the exercise if you do not have your handout.

Directions:

1. Change “Student Name” on line 3 (above) with your name.
2. Complete as much of the assignment as possible during class; what you do not complete in class will need to be done on your own outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this assignment document. Space for your answers are provided in this document and is indicated by the “>” character. If you need a second paragraph be sure to start the first line with “>”. You should notice that the answer is highlighted in green by RStudio.
5. Before you leave the classroom today, it is *imperative* that you **push** this file to your GitHub repo.
6. When you have completed the assignment, **Knit** the text and code into a single PDF file. Basically, just press the **Knit** button in the RStudio scripting panel. This will save the PDF output in your Week1 folder.
7. After Knitting, please submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file (*Week1_Assignment.Rmd*; with all code blocks filled out and questions answered) and the PDF output of **Knitr** (*Week1_Assignment.pdf*).

The completed exercise is due on **Wednesday, January 18th, 2017 before 12:00 PM (noon)**.

1) HOW WE WILL BE USING R AND OTHER TOOLS

You are working in an RMarkdown file. It allows you to integrate text and R code into a single file. There are two major features to this document: 1) Markdown formatted text and 2) “chunks” of R code. Anything in an R code chunk will be interpreted by R when you *Knit* the document.

When you are done, you will *knit* your document together. However, if there are errors in the R code contained in your Markdown document, you will not be able to knit a PDF file. If this happens, you will need to review all of your code to make sure you do not have any errors. Even if you are able to knit without issue, it is good practice to review the PDF before you submit the assignment.

2) SETTING YOUR WORKING DIRECTORY

In the R code chunk below, please provide the code to: 1) clear your R environment, 2) print your current working directory, and 3) set your working directory to your Week1 folder.

3) USING R AS A CALCULATOR

To follow up on the Week 0 exercises, please calculate the following in the R code chunk below. Feel free to reference the Week 0 handout.

- 1) the volume of a cube with length, l , = 5.
- 2) the area of a circle with radius, r , = 2 (area = $\pi * r^2$).
- 3) the length of the opposite side of a right-triangle given that the angle, θ , = $\pi/4$. (radians, a.k.a. 45°) and with hypotenuse length $\sqrt{2}$ (remember: $\sin(\theta) = \text{opposite}/\text{hypotenuse}$).
- 4) the log (base e) of your favorite number.

4) WORKING WITH VECTORS

To follow up on the Week 0 exercises, please perform the requested operations in the Rcode chunks below. Feel free to reference the Week 0 handout.

Basic Features Of Vectors

In the R code chunk below, do the following: 1) Create a vector **x** consisting of any five numbers. 2) Create a new vector **w** by multiplying **x** by 14 (i.e., “scalar”). 3) Add **x** and **w** and divide by 15.

Now, do the following: 1) Create another vector (**k**) that is the same length as **w**. 2) Multiply **k** by **x**. 3) Use the combine function to create one more vector, **d** that consists of any three elements from **w** and any four elements of **k**.

Summary Statistics of Vectors

In the R code chunk below, calculate the **summary statistics** (ie., maximum, minimum, sum, mean, median, variance, standard deviation, and standard error of the mean) for the vector (**v**) provided.

```
v <- c(16.4, 16.0, 10.1, 16.8, 20.5, NA, 20.2, 13.1, 24.8, 20.2, 25.0, 20.5, 30.5, 31.4, 27.1)
```

5) WORKING WITH MATRICES

In the R code chunk below, do the following: Using a mixture of Approach 1 and 2 from the handout, create a matrix with two columns and five rows. Both columns should consist of random numbers. Make the mean of the first column equal to 8 with a standard deviation of 2 and the mean of the second column equal to 25 with a standard deviation of 10.

Question 1: What does the **rnorm** function do? What do the arguments in this function specify?

Answer 1:

In the R code chunk below, do the following: 1) Load **matrix.txt** from the Week1 data folder as matrix **m**. 2) Transpose this matrix. 3) Determine the dimensions of the transposed matrix.

Question 2: What are the dimensions of the matrix you just transposed?

Answer 2:

Indexing a Matrix

In the R code chunk below, do the following: 1) Index matrix `m` by selecting all but the third column. 2) Remove the last row of matrix `m`.

Question 3: Describe what we just did in the last series of indexing steps.

Answer 3:

6) BASIC DATA VISUALIZATION AND STATISTICAL ANALYSIS

Load Zooplankton Dataset

In the R code chunk below, do the following: 1) Load the zooplankton dataset from the Week1 data folder. 2) Display the structure of this data set.

Correlation

In the R code chunk below, do the following: 1) Create a matrix with the numerical data in the `meso` dataframe. 2) Visualize the pairwise **bi-plots** of the six numerical variables. 3) Conduct a simple **Pearson's correlation** analysis.

Question 4: Describe some of the general features based on the visualization and correlation analysis above?

Answer 4:

In the R code chunk below, do the following: 1) Redo the correlation analysis using the `corr.test()` function in the `psych` package with the following options: `method = "pearson"`, `adjust = "BH"`. 2) Now, redo this correlation analysis using a non-parametric method. 3) Use the print command from the handout to see the results of each correlation analysis.

Question 5: Describe what you learned from `corr.test`. Describe what you learned from `corr.test`. Specifically, are the results sensitive to whether you use parametric (i.e., Pearson's) or non-parametric methods? When should one use non-parametric methods instead of parametric methods? With the Pearson's method, is there evidence for false discovery rate due to multiple comparisons? Why is false discovery rate important?

Answer 5:

In the R code chunk below, use the `corrplot` function in the `corrplot` package to produce the ellipse correlation plot in the handout.

Linear Regression

In the R code chunk below, do the following: 1) Conduct a linear regression analysis to test the relationship between total nitrogen (TN) and zooplankton biomass (ZP). 2) Examine the output of the regression analysis. 3) Produce a plot of this regression analysis including the following: categorically labeled points, the predicted regression line with 95% confidence intervals, and the appropriate axis labels.

Question 6: Interpret the results from the regression model

Answer 6:

Question 7: Explain what the `predict()` function is doing in our analyses.

Answer 7:

Using the R code chunk below, use the code provided in the handout to determine if our data meet the assumptions of the linear regression analysis.

- Upper left: is there a random distribution of the residuals around zero (horizontal line)?
- Upper right: is there a reasonably linear relationship between standardized residuals and theoretical quantiles? Try `help(qqplot)`
- Bottom left: again, looking for a random distribution of `sqrt(standardized residuals)`
- Bottom right: leverage indicates the influence of points; contours correspond with Cook's distance, where values $> |1|$ are "suspicious"

ANALYSIS OF VARIANCE (ANOVA)

Using the R code chunk below, do the following: 1) Order the nutrient treatments from low to high (see handout). 2) Produce a barplot to visualize zooplankton biomass in each nutrient treatment. 3) Include error bars (± 1 sem) on your plot and label the axes appropriately. 4) Use a one-way analysis of variance (ANOVA) to test the null hypothesis that zooplankton biomass is affected by the nutrient treatment. 5) Use a Tukey's HSD to identify which treatments are different.

Question 8: How do you interpret the ANOVA results relative to the regression results? Do you have any concerns about this analysis?

Answer 8:

Using the R code chunk below, use the diagnostic code provided in the handout to determine if our data meet the assumptions of ANOVA (similar to regression).

SYNTHESIS: SITE-BY-SPECIES MATRIX

In the R code chunk below, load the `zoop.txt` dataset in your Week1 data folder. Create a site-by-species matrix (or dataframe) that does not include TANK or NUTS. The remaining columns of data refer to the biomass ($\mu\text{g/L}$) of different zooplankton taxa:

- CAL = calanoid copepods
- DIAP = *Diaphanasoma* sp.
- CYL = cyclopoid copepods
- BOSM = *Bosmina* sp.
- SIMO = *Simocephalus* sp.
- CERI = *Ceriodaphnia* sp.
- NAUP = naupuli (immature copepod)
- DLUM = *Daphnia lumholtzi*
- CHYD = *Chydorus* sp.

Question 9: With the visualization and statistical tools that we learned about in the Week 1 Handout, use the site-by-species matrix to assess whether and how different zooplankton taxa were responsible for the total biomass (ZP) response to nutrient enrichment. Describe what you learned below in the "Answer" section and include appropriate code in the R chunk.

SUBMITTING YOUR ASSIGNMENT

Use Knitr to create a PDF of your completed Week1_Assignment.Rmd document, push the repo to GitHub, and create a pull request. Please make sure your updated repo include both the PDF and RMarkdown files.

Unless otherwise noted, this assignment is due on **Wednesday, January 18th, 2015 at 12:00 PM (noon)**.