**Bio 218P: Data Analysis in R**

Spring 2022: January 12-April 27th

**Instructor:** Danielle Presgraves, Ph.D.

**Email: danielle.presgraves@rochester.edu**

**Office Hours:** By appointment only. All of my appointments will be held on Zoom (see link on the Blackboard course page).

**Teaching Assistants:**

Please see the Blackboard page for office hours, as well as the Zoom link for virtual office hours.

**Text:**  There is no textbook. I will upload modules on Fridays onto Blackboard. Each module will have an interactive .rmd file (this is the document type that is produced from the "R Markdown language" from RStudio) for you to work through at your own pace, a video where I work through the .rmd file (in case you need extra support), and a relevant problem set, due at a scheduled date. There will also be an accompanying chapter in the "R book" that overlaps in content with each .rmd file but sometimes give additional background or examples.

**Course Description:** This one-credit hour course uses basic statistical concepts to introduce students to the **R** programming language. If you are considering a minor/major in computational biology or are going to undertake data analysis in your academic/professional career, it is in your own best interest to learn this ubiquitous, industry-standard programming language. Since it is an open-source (free!) language, if you take the time to learn **R** and **RStudio** now, you will be able to take them with you wherever you go after you graduate.For these reasons, it is a popular program in many statistical fields, including genomics and bioinformatics.

This course will be taught online as **12 modules** that will each contain an interactive .rmd file and relevant data sets. These 12 modules will also contain short explanatory videos. As the final activity for each module, participants will demonstrate competency with the presented material by completing and uploading a problem set.

The content of this course will focus on general data visualization (graphing and displaying uncertainty), basic statistical analysis (goodness of fit tests, student's t tests, ANOVA, correlation, regression and general linear models) and an introduction to programming in R (built-in functions, libraries and general strategies in problem solving). This course follows the expectations of the credit hour policy as required by the University of Rochester here: <https://www.rochester.edu/provost/assets/PDFs/UR-Credit-Hour-Policies-March2016.pdf>

**Course Goals:** You will learn to:

* Create markdown language files that produce attractive word documents containing their statistical data analysis (in R) and your interpretation (in R Markdown) of provided data sets (your future professors will be grateful for your skills)
* Explore and manipulate built-in functions, features and visualization in basic R
* Access and utilize appropriate libraries and packages in R
* Explore basic programming tools available in base R.
* Develop basic troubleshooting skills by learning to google error messages. This sounds silly, but it is an extremely valuable skill. Please **attribute** and explain any code snippets that you use from other locations (ex. stackoverflow etc).

*Note: If you are looking for an even deeper dive into the R language experience, you might also consider a four-credit course, Stat 276 "Statistical Computing in R", offered in the Spring semester.*

We will use **R** within the **RStudio** Markdown framework. You can find out all about R and download it at <http://www.r-project.org/>. You will need to download **RStudio** here**:** <https://www.rstudio.com/products/RStudio/>

For instructions on how to download and install RStudio and R, please see the document "How\_to\_Download\_Rstudio\_R.pdf" which has been uploaded with your syllabus.

**Pre-Requisites:** It is your responsibility to ensure that you have completed an introductory statistics course (Stat 212, 213, 214 etc) or are concurrently enrolled in Stat 212. The topics covered in Bio 218P are closely aligned to Bio/Stat 214 but any of the other introductory statistics courses offered at UR will provide sufficient knowledge for success in Bio 218P. Additionally, there will be some new topics (PCA, Kaplan-Meier Curves, ROC/AUC, and programming) that are not covered in most basic introductory statistics course so don't panic when you see these techniques listed in the schedule – you will soon appreciate why these topics are important in data analysis!

In most modules, I will give a short background and/or links to online summaries to refresh your memory of the particular statistical topic/technique covered. However, this course focuses on *using* **R** to apply statistical techniques not on teaching fundamental statistical techniques so previous knowledge is assumed and won't be taught in-depth.

**Website:** We will be using Blackboard at learn.rochester.edu. I will post the weekly module and accompanying problem set at **12pm (noon)** on the days indicated by the schedule. Problem sets will be uploaded to Blackboard and you should receive a 'receipt' that your PS has been uploaded correctly. No late problem sets will be accepted. You need to "knit" your .rmd file to create a word document. You can then convert it to a pdf document by saving it from word to upload. ***Knitting is equivalent to compiling your document and demonstrates that you have solved and successfully 'trouble-shot' the problem with working code.***

**Grades composition:** 100% of your grade comes from your problem sets. There are a total of **12** assignments but ONLY your **top 11** assignments (**which must include PS 1 and PS 12**) will contribute to your final grade so each assignment is worth ~9 % of your total grade. There are also some extra credit opportunities attached to some of the problem sets. This means that there are approx. 115-120 points available to you and you just need to get 93 of those points to get an "A"!

You are required to hand-in **Module 1** and **Module 12** as these two modules cover important central concepts. If you want to drop any of the other modules, that is acceptable. You are able to earn an "A" even if you drop one module. You will be given 1 point for completing the Academic Honesty Policy (completing this "quiz" is also how you access the content of the course). It is your responsibility to note any date modifications on Blackboard. Please note: No late work will be accepted since answer keys will be automatically available two minutes after the listed due date for each PS. This means that you will get 0% for any assignments that you fail to hand in on time. To be clear: Except for truly exceptional circumstances, no late work will be accepted for credit.Plan to hand in your problem set well in advance of the due date since a computer crashing at 11:59 am for an assignment that is due at 12:00pm (noon) is NOT an exceptional circumstance. If you are struggling to hand in at *least eleven of the twelve* problem sets in this course on time, you need to talk with me.

**Grade cut-offs:**

**A/A-** >93 **/** 88 – 92.9

**B+/B/B-** 87.9-85 **/** 84.9-82.0 **/** 81.9-79.0

**C+/C/C-** 78.9 – 75.0 **/** 74.9-71 **/** 70.9 – 68.0

**D+/D/D-** 67.9 – 64 **/** 63.9 – 60.0 **/** 59.9 – 55.0

**E** < 55

**Academic Honesty:** You are required to fill in the academic honesty policy in order to gain access to the first module of the course.

All assignments and activities associated with this course must be performed in accordance with the University of Rochester’s Academic Honesty Policy. More information is available at: <http://www.rochester.edu/college/honesty/students.html>

On assignments, I encourage collaboration. It is acceptable (and often useful!) to work in groups but it is not acceptable to copy another’s work or to allow your work to be copied. When working in groups on assignments, you MUST include the names of members in your group on the assignment, and assignments MUST be individually completed. Copying of homework assignments is not only a reportable violation of our academic honesty policy, it also deprives the instructor of an important cue that a student is having trouble with course material.

Despite my warning to hand in your own work, I strongly encourage you to effectively – and liberally - utilize google. One of the goals of this course is to empower you to eventually be autonomous when using R. One of the best ways to achieve competence in R is to learn to google your error terms in R to refine your ability to trouble-shoot your code. I guarantee that other individuals have also had a similar problem to whatever one you are encountering in R and it is wonderful to have access to their knowledge about what might be wrong with your own code! This is one of the most glorious benefits of 'open access' languages like R. I also encourage you to use snippets of code found in places like stackoverflow in your work ***with attribution and explanation****.*Attribution means that you should clearly state that you used or heavily borrowed (with modification) code snippets and include a web link to where you found the link and explain what the code is doing. Not only will I appreciate this acknowledgment, you will fulfill the spirit of 'academic honesty' AND it will help 'future you' when you return to try to figure out a similar problem by re-reading your own code. This happens more than you might think!

Please remember that sharing of course material without permission is a copyright violation, and is prohibited under our academic honesty policy.

**Additional University Resources:**

**Disability Statement:** Your success in this course is important to me, and it is the policy and practice of the U of R to create inclusive learning environments consistent with federal and state law. If you have a documented disability, and need an accommodation, please contact me privately as soon as possible so that we can discuss with the relevant office how to meet your specific needs and the requirements of the course.

Undergraduates requiring accommodations should contact Disability Resources, located in Taylor Hall (https://www.rochester.edu/college/disability/about/staff.html) at: telephone:(585) 275 3424 email: disability@rochester.edu

**Care network:** Student success at the University of Rochester includes more than just academic performance. Please feel comfortable speaking with me about challenges you are experiencing within and outside of the classroom so that I may submit a CARE Referral on your behalf. A CARE Referral is submitted when the level of concern for a student necessitates inclusive, multi-layered support from the campus community. The CARE network administrator shares information only with staff who need to know it in order to help you. I care about your success and am committed to my role in helping you get connected to appropriate campus resources. Note: You can also submit a self-report to the CARE network if you are worried about yourself and want coordinated resources. https://www.rochester.edu/care/index.html

**Spring 2022 Schedule of topics, modules and problem set due dates:**

*The general heuristic: A new module will be released each Friday by noon (except for spring break) and the associated problem set will be due a week later by Friday at noon. I will release a few modules early as the semester progresses (usually modules 10, 11, and 12 will all be released at the same time).* ***Remember: completing modules 1 and 12 are required to pass this course.***

|  |  |  |  |
| --- | --- | --- | --- |
| Module | Date available | PS Due Date (by noon) | Topic Covered |
| **0** | **Jan 10** | **Prior to Jan**  **14** | **Learning Objectives:**   1. You will demonstrate awareness of course schedule (due dates) and expectations of participation   **Instructional and Assessment Activities:**   1. Complete the "academic honesty" statement in order to gain access to the next module (module 1) |
| **1** | **Jan 14** | **Jan 21** | **Title:** Introduction to R and RStudio and markdown  **Corresponding R Book Chapters:**  Chapter 1, Appendix  **Learning Objectives:**   1. You will demonstrate a minimum level of competency in the RStudio environment. Competency includes:  * You will create a new .rmd file * You will correctly incorporate the R Markdown language * You will appropriately use 'R chunks' * You will successfully 'Knit' (equivalent to 'compiling') their final .rmd document into a word document * You will examine the basic features of dataframes * You will manipulate a built-in data set as well as upload an external .csv file into the RStudio memory space   **Instructional and Assessment Activities:**   1. You will watch the video introduction to the RStudio environment 2. You will work through the module 1 .rmd file 3. You will hand in Problem Set 1 by posted date |
| **2** | **Jan 21** | **Jan 28** | **Title:** Graphics and Basic Statistics mosaic plots, scatterplots  **Corresponding R Book Chapters:**  Chapter 2  **Learning Objectives:**   1. You will assess what type of variable they are plotting 2. You will visualize data using the appropriate display format: scatterplot, histogram, stem and leaf plots   **Instructional and Assessment Activities:**   1. You will work through the module 2 .rmd file 2. You will hand in Problem Set 2 by posted date   **Additional resources:**   1. [**https://r4ds.had.co.nz**](https://r4ds.had.co.nz) 2. [**https://www.youtube.com/watch?v=sxQaBpKfDRk**](https://www.youtube.com/watch?v=sxQaBpKfDRk) 3. [**https://www.data-to-viz.com/#boxplot**](https://www.data-to-viz.com/#boxplot) |
| **3** | **Jan 28** | **Feb 4** | **Title:** The 'grammar of visualization': ggplot2  **Corresponding R Book Chapters:**  Chapter 2  **Learning Objectives:**   1. You will discover how to import libraries! 2. You will examine the basic model of ggplot2 3. You will compare and contrast the data visualizations in module 2 to the more attractive results of the ggplot2 package   **Instructional and Assessment Activities:**   1. You will work through the module 3 .rmd file 2. You will hand in Problem Set 3 by posted date |
| **4** | **Feb 4** | **Feb 11** | **Title:** Describing Uncertainty: Confidence intervals  **Corresponding R Book Chapters:**  Chapter 3  **Learning Objectives:**   1. You will compute common summary statistics commands 2. You will qualitatively interpret and quantify common descriptions of uncertainty such as confidence intervals and standard error 3. You will investigate the benefits of boxplots for visualizing uncertainty in data sets 4. You will simulate a bootstrap to calculate quantiles of data sets   **Instructional and Assessment Activities:**   1. You will work through the module 4 .rmd file 2. You will hand in Problem Set 4 by posted date |
| **5** | **Feb 11** | **Feb 18** | **Title:** Hypothesis Testing: a pipeline for refining your thinking  **Corresponding R Book Chapters:**  Chapter 4  **Learning Objectives:**   1. You will constructbinomial distributions. 2. You will demonstrate familiarity with Type I, Type II errors and power. 3. You will discover the sample() function and use it when writing a short program to simulate p-values. 4. You will clearly demonstrate the four components of hypothesis testing:  * You will justify use of hypothesis testing and will construct the appropriate null hypothesis * You will select the appropriate statistical test and will calculate the test value * You will either calculate a p-value (if using Binomial Distribution) or interpret the critical value from the appropriate table. * You will justify your conclusion and will offer support in the form of calculating a confidence interval  1. You will compare the results of the exact Binomial Test and the approximate proportion test 2. You will graph Receiver Operator Curves and calculate the Area under the curve (AUC).   **Instructional and Assessment Activities:**   1. You will work through the module 5 .rmd file   You will hand in Problem Set 5 by posted date |
| **6** | **Feb 18** | **Feb 25** | **Title:** Goodness of Fit Tests: Chi squared, Fisher's exact test  **Corresponding R Book Chapters:**  Chapter 5  **Learning Objectives:**   1. You will compare the χ2 contingency test (an approximate test) and Fisher's Exact test 2. You will manipulate dataframes and convert them into matrices and tables. 3. You will learn to directly read in data using the 'c()' function 4. You will create mosaic plots   **Instructional and Assessment Activities:**   1. You will work through the module 6 .rmd file 2. You will hand in Problem Set 6 by posted date |
| **7** | **Feb 25** | **Mar 4** | **Title:** The Normal Distribution, testing assumptions and Student's t Tests  **Corresponding R Book Chapters:**  Chapters 6, 7  **Learning Objectives:**   1. You will simulate normally distributed data sets and non-normally distributed data sets (ex. Poisson Distributed data sets) 2. You will investigate visual assessment of normality with q-q norm plots 3. You will compare the results of parametric tests (t-tests) under data that is normally distributed and non-normally distributed to investigate the robustness of the t-test against violations of assumptions. 4. You will use the one sample t-test, two sample t-test and paired t-test and see how the arguments change when specifying which type to t-test to use 5. You will look at the power of the t-test   **Instructional and Assessment Activities:**   1. You will work through the module 7 .rmd file 2. You will hand in Problem Set 7 by posted date |
| **Spring Break** | | | |
| **8** | **Mar 11** | **Mar 18** | **Title:** Non-Parametric Tests: what happens when the assumptions of a parametric test are not met?  **Corresponding R Book Chapters:**  Chapter 8  **Learning Objectives:**   1. You will review visual assessment for normal distribution 2. You will use simulations to investigate the effects on Type I, II error and Power when assumptions of a test are not met. 3. You will use the Shapiro-Wilk test to formally test for the normal distribution of data 4. You will assess homogeneity of variances between two populations using the var.test() function 5. You will investigatecommon data transformations 6. You will use a common non-parametric alternative to the t-test, the Wilcoxon test. 7. You will simulate the p values for data that aren't normal, can't be transformed and don't have a standard nonparametric test. 8. You will investigate a method survival curves when your data is incomplete, called a Kaplan-Meier Curve.   **Instructional and Assessment Activities:**   1. You will work through the module 8 .rmd file 2. You will hand in Problem Set 8 by posted date   **Additional Resources:**   1. <http://www.intuitor.com/statistics/T1T2Errors.html> 2. <http://onlinestatbook.com/2/power/power_demo.html> 3. <http://onlinestatbook.com/2/power/power_demo2.html> |
| **9** | **Mar 18** | **Mar 25** | **Title:** ANOVA  **Corresponding R Book Chapters:**  Chapter 9  **Learning Objectives:**   1. You will adequately test assumptions of the ANOVA test using bartlett to test for homogeneity of variances 2. You will explore ANOVA and the difference between the aov() function and the lm() function 3. You will use the nonparametric version of ANOVA, kruskal.test() 4. You will use tukeyHSD to identify which of the groups has a different mean value 5. You will use the tapply function to create a plot of the variances versus the mean values of the groups.   **Instructional and Assessment Activities:**   1. You will work through the module 9 .rmd file 2. You will hand in Problem Set 9 by posted date   **Additional resources:**   1. <https://www.youtube.com/watch?v=ZpHCzpWHNYk> 2. https://demonstrations.wolfram.com/VisualANOVA/ |
| **10** | **Mar 25** | **April 1** | **Title:** Correlation and Regression  **Corresponding R Book Chapters:**  Chapter 10  **Learning Objectives:**   1. You will explore and contrast linear regression and linear correlation 2. You will test assumptions to ensure that they are using the appropriate tests 3. You will calculate Pearson's correlation, and nonparametric alternatives (Spearman, Kendall) 4. You will explore the residuals and calculate confidence intervals   **Instructional and Assessment Activities:**   1. You will work through the module 10 .rmd file 2. You will hand in Problem Set 10 by posted date   **Additional Resources:**   1. Regression simulation: <http://onlinestatbook.com/2/regression/linear_fit_demo.html> 2. Correlation versus regression: <http://shiny.stat.calpoly.edu/Corr_Reg_Game/> |
| **11** | **April 1** | **April 15\***  **(two weeks since it is a long module)** | **Title:** General linear Models (blocking, covariates, multivariate) and Principle Component Analysis, PCA  **Corresponding R Book Chapters:**  Chapter 11  **Learning Objectives:**   1. You will explore three major types of general linear model: including a block, analyzing with two or more variables simultaneously (multi-factor design) and including a covariate in analysis (ANCOVA) 2. You will manipulate and conduct principle component analysis   **Instructional and Assessment Activities:**   1. You will work through the module 11 .rmd file 2. You will hand in Problem Set 11 by posted date |
| **12** | **April 1** | **April 22** | **Title:** Basic Programming in R: SWIRL() and more!  **Corresponding R Book Chapters:**  Chapter 12  **Learning Objectives:**   1. You will explore the interactive SWIRL() programming library 2. You will use the basic components of programming: loops, defining their own function, demonstrating proficiency with a subset of higher order functions such as 'filter'   **Instructional and Assessment Activities:**   1. You will work through the module 12 .rmd file 2. You will hand in Problem Set 12 by posted date |