**Introduction to R II**

1. **Pre-work:**

* You will use **Google Cloud Platform (GCP)** for in-class work. At least **1 week before the course**, R-IT will add you to GCP. Follow their instructions to connect.
* **Optional (but recommended)**: Install R/Posit locally—especially useful after the course ends. <https://jacksonlaboratory.sharepoint.com/sites/IT/SitePages/R-Studio---Posit.aspx>
* Ensure R and RStudio are functional before Day 1.

1. **Pre-requisites:**

* Completion of “Intro to R I” (or equivalent competency) is a pre-requisite for taking this course.
* “Biostatistics I” (or equivalent competency). We will be focusing on R as a statistical tool so having familiarity with the common tests of biostatistics is highly recommended. In most modules, I will give a short background and/or links to online summaries to refresh your memory of the 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. 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!

1. **Why R:** <https://allisonhorst.com/everything-else>

* Like python, it is a “Table Stakes” language for Data Science
* Many libraries for genomics/bioinformatics and visualizations
* Statistical analysis of data sets is straightforward.

1. **Course Description:**

This workshop is the second half of the introduction to the R programming language course. We will build on the capabilities of R as a visualization tool, through ggplot and the Tidyverse, and spend most of our time exploring the power of R for statistical analysis of datasets. Topics covered will include: ggplot and the Tidyverse philosophy, hypothesis testing, goodness-of-fit tests (chi-squared, Fisher’s exact test), Student’s t test, normal distributions, correlation, Regression, ANOVA, and non-parametric versions of these parametric tests.

1. **Course Goals:**

Provide you with the tools to analyze and visualize data. By the end of the week, you will be able to:

* Compare the syntax and structure of R to the ideals of the tidyverse
* Use ggplot2 to plot your data
* Conduct and implement standard statistical tests, parametric and nonparametric hypothesis tests

1. **Locations of the course:**

* **See Module 0 on Canvas for zoom link, passcode and for any relevant classroom information for in-person courses**

1. **Instructor:** Danielle Presgraves, Ph.D. **email:** [Danielle.presgraves@jax.org](mailto:Danielle.presgraves@jax.org)

|  |  |  |
| --- | --- | --- |
| Module | Topic Covered | Enrichment Suggestions |
| **0** | **Learning Objectives:**   * If you took an introductory R course elsewhere, please download R/RStudio (now “Posit”) and ensure that everything is working prior to class. You might need support from IT so ensure you do this prior to the start of the course. | 1. **General R:**   https://r4ds.hadley.nz/   1. **Statistics Summary (flowchart):**   https://www.statsflowchart.co.uk |
| ***RStudio (Posit) Fundamentals Review*** | | |
| **1AB**  **2.5 hours** | **Title:** Reviewing the basics of R (from Intro to R I)  **Goals:** Competency in the following   1. **Basic R Functionality**    * uploading datasets; examination of datasets (e.g., summary())    * assignment operator; c()    * Slicing (with conditionals); subset; \*apply family 2. **Basic R graphing**    * Scatterplots, histograms, mosaic plots 3. **Graphing with the ggplot library**    * Loading library; using built-in data sets; ggplot() and geoms()   **Activities:**   * Produce graphs with manipulated datasets in each of the above categories. | **ggplot cheat sheet:**  <https://posit.co/wp-content/uploads/2022/10/data-visualization-1.pdf> |
| **2AB**  **2.5 hours** | **Title:** Introduction to thetidyverse  **Goals:**   1. Exploration of the tidyverse package (ggplot, dplyr) 2. Data in tidyverse – tibbles! 3. Using the foundational rules of dplyr to manipulate/clean datasets 4. Pipe, filter, arrange, distinct, count, mutate, select, group\_by, passing results to ggplot2   **Activities:**   * Manipulate and graph a built-in dataset using tidyverse | 1. **Another Philip Guo product:**   <https://tidydatatutor.com>   1. **Hadley Wickham explains:**   <https://r4ds.hadley.nz> |
| ***Statistics in R*** | | |
| **3AB**  **2.5 hours** | **Title:** Hypothesis Testing: a pipeline for your thinking  Goodness of fit Tests: Chi squared, Fisher’s Exact test  **Goals:**   1. Revisit the general outline of Hypothesis Testing (4 steps, null/alternate hypothesis, p-values, Type I, Type II error) 2. Build up a Binomial Distribution    1. The hard way: using sample, pbinom(), tidyverse etc.    2. The easy way: using binom.test() and prop.test() 3. Apply the Hypothesis Testing framework to the Bumpus dataset 4. Apply the chi-squared contingency test to the bumpus dataset 5. Work through an example of ROC/AUC using the pROC library 6. Fisher’s exact test (another type of Contingency test that is exact)   **Activities:**   * Formally test a hypothesis on the Bumpus data set * Recapitulate the ROC/AUC example | 1. <https://en.wikipedia.org/wiki/Statistical_hypothesis_test> 2. <http://www.intuitor.com/statistics/T1T2Errors.html> 3. <http://onlinestatbook.com/2/power/power_demo.html> 4. <http://onlinestatbook.com/2/power/power_demo2.html> |
| **4AB**  **2 hours** | **Title:** The Normal Distribution, testing assumptions and Student's t Tests  **Goals:**   1. Testing for Normality:    1. Visualization: q-q plots    2. shapiro.wilk test 2. You will simulate normally distributed data sets and non-normally distributed data sets (ex. Poisson Distributed data sets) 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 6. You will look at var.test to ensure that two distributions have similar distributions. 7. We will examine the non-parametric alternatives to these tests (Welch’s approximate t test, Wilcoxon test) 8. Kaplan Meier curves   **Activities:**   * 1. Use data sets to investigate normality, test a hypothesis with all appropriate examinations of the data (normality etc.), and produce a Kaplan-Meier curve for a built-in data set. |  |
| **5AB**  **2.5 hours** | **Title:** ANOVA and linear correlation/regression  **Goals:**   1. You will adequately test assumptions of the ANOVA test using bartlett to test for homogeneity of variances 2. You will use the nonparametric version of ANOVA, kruskal.test() 3. You will use tukeyHSD to identify which of the groups has a different mean value 4. You will use the tapply function to create a plot of the variances versus the mean values of the groups. 5. You will explore and contrast linear regression and linear correlation 6. You will test assumptions to ensure that they are using the appropriate tests 7. You will calculate Pearson's correlation, and nonparametric alternatives (Spearman, Kendall) 8. You will explore the residuals and calculate confidence intervals   **Activities:**   * Reproducing similar analysis on different data sets. | ANOVA simulations:  <https://demonstrations.wolfram.com/VisualANOVA>  <https://www.youtube.com/watch?v=ZpHCzpWHNYk>   1. Regression simulation: <http://onlinestatbook.com/2/regression/linear_fit_demo.html> 2. Correlation versus regression: <http://shiny.stat.calpoly.edu/Corr_Reg_Game/> |

**Extra material (if there is time):**

|  |  |
| --- | --- |
| **Title:** General linear Models (blocking, covariates, multivariate) and Principal Component Analysis, PCA  **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 |  |

**Assignments:**

**Module 1:**

Please explain your steps. Fully utilize the power of RStudio (Posit) by using Markdown cells AND R chunks (don't just use R chunks with hashed out comments). Use a new rmd file for your assignment and please use a format such as: Lastname\_Firstname\_PS1.rmd so that when I go through and mark/comment, it is straightforward for me to see that it is yours. Thank you!

--------------------------------

**Q1.**   Use one of the following built-in datasets (**iris** or **OrchardSprays**) to do the following:   
a.    Load in the data from the built-in R datasets.  
b.    Inspect the dataframe with at least three attributes (explain what you choose and how it is implemented).   
c.    Slice out rows and columns using any criteria that you wish, explain what the criteria is (and why), and then write it to an external file.   
d.    Slice out the rows or columns using Boolean conditions.   
e.    Subset the data in any way that you think is meaningful (explain why it is meaningful to you).   
f.    Use tapply to breakdown the mean values of one of the columns based on the factors in the dataset.

**Q2.** Use the **HairEyeColor** built-in dataset for basic R to create a mosaic plot. This will be a similar process to what we did in class. Ask and answer, using a graph, a question based on this table.

**Q3.** Use the built-in data set for ggplot2, **msleep**, to create a visualization using ggpplot and geoms (not base R) that answers questions about mammals sleeping.

**Module 2:**

*Please hand in an .rmd file in the format: lastname\_firstinitial\_day2.rmd.****Include markdown chunks with information about the question and, if appropriate, a general outline of your approach to solving it. Also use comments in your R coding chunks.****Use a blank rmd notebook (it should not include the default chunks that are automatically rendered when you create a new .rmd file).*

***-------------------------------***

***NB: I failed to mention that you should use the built in data set iris to tackle this question.***

1. Is this data 'tidy'? Why or why not? How might you demonstrate that it follows the conventions of "tidy" data?
2. We want to do a selective breeding experiment using only the individuals with the largest 2 Petal.Lengths from each species. Create a tibble called "largestPetalSize" that contains the 6 individuals that fulfill this criterion. What do you notice about the output and what does this suggest about how tied values are resolved?
3. Now, we want to find the mean Sepal.Length of each of the iris species.
4. We want to add a new column to this dataframe that is the area of the petals. We will make some approximations and claim that petals are exactly rectangular shapes so that area is simply Petal.Length\*Petal.Width.
5. Finally, we want to graph out a boxplot, using ggplot, of only the smallest 10 Petal.Lengths of each species. This should be one command.

**Module 3:**

To answer question 3, write out the four steps of hypothesis testing. Recall the steps are:

1. *Stating the Ho/Ha*
2. *Stating and running the appropriate test and listing any assumptions mentioned so far*
3. *Stating the alpha and how the p-value compared to this alpha*
4. *Concluding – do you reject or fail to reject your null hypothesis*.

These steps should be included in your markdown chunks not your R programming chunks. Include interpretation for your answers.

1. Focus on only one trait in the Bumpus data set (you can choose whichever *numeric*trait appeals to you): is there any indication of stabilizing selection on your chosen trait among the house sparrows? To rephrase the previous question, more precisely: Do the survivors of the storm have less variance than the non-surviving? You don't need to conduct a test just provide graphical (hint: probably a boxplot) evidence and summary statistic evidence (mean, median etc.).
2. Is there is a difference in the variance of the trait between male and female sparrow? You don't need to conduct a formal test just provide graphical (boxplot) evidence with the males and female values of the trait separated and summary statistic evidence (mean, median etc.).
3. Using the **age** categorical variable test whether adult male and juvenile (young) male birds were equally likely to have survived the storm. In this data set, adult/youth information is only provided for the male birds; a=adult and y = young/juvenile. This question is like the hypothesis question we worked through in module 3 except that it is important to recognize that, for this question, **n** = total male survival count (where n=trials) and is not total survival (which would include females). Provide the answer using both the binom.test() and the prop.test() function. Which of these two tests had a smaller p-value? Explain why that is the case.
4. We explored multiple ways to answer the following question so use any two methods you want to get the answer: If you flipped a coin 1000 times, what is the probability of having *84 or fewer* heads? How do your two answers compare (for instance, are they exactly the same? Are they different? If so, why are they different?)
5. **ROC and AUC:**Repeat the example that was worked in your rmd file, but this time create one graph that includes gos6, WFNS, Age, s100b, and NDKA. Calculate the AUC for these factors. Based on the AUC value, which of these factors are good classifiers for outcome after aSAH?

**Module 4:**

1. Most animals are bilaterally symmetric, and among those that are there are very few systematic differences where the left side is larger or small than the right side. In part, this basic symmetry seems to be common because it is difficult to evolve asymmetry, but also for most things a symmetric condition seems to be most functional. In some cases, however, asymmetry may be lead to higher fitness. One such case is among snail-eating snakes. Snails are themselves asymmetrical (because they tend to coil to the right, rather than to the left), and for snakes who eat snails, it turns out to be more useful to have more teeth on the right side of the head, because the snakes grip the snail most effectively on the right to remove them from their shells. ***Given this, is asymmetry common among snail-eating snakes?***

The data file "R2\_AsymmetricalSnakes.csv" contains data on the asymmetry score (right side tooth count minus left side tooth count) for twelve species of snakes that eat snails. Follow the general visual procedure outlined in Module\_4AB and answer the above question using the qqnorm and qqline functions. Use a formal test to see if the data set is normally distributed. Could you run a t-test on this data? Explain why or why not.

1. It is discovered that the data discussed in the recitation were all male mooses (let's not discuss how this was ascertained). Female mooses from the same region, the north region, were all weighed: 301, 320, 382, 401, 349, 318, 342 and 369. Do northern female mooses weight less than their male counterparts? **(include ALL four hypothesis testing steps and code with answer).**
2. The data file "R2\_Sleep contains information on 10 individuals who took two different sleep aids. Each row represents a different individual, measured in each of the two ways. Do both measurements give the same result? **(include ALL four hypothesis testing steps and code with answer)**
3. If we wanted to pick up a difference of 1 between Drug A and Drug B and we wanted a power level of 0.90, what sample size would we need?
4. Produce a Kaplan-Meier curve for the built-in Leukemia data set ('leukemia'). This dataset is similar to the lung dataset, but instead of male/female column, there is a column, **x**, that is "Maintained" or "Nonmaintained". Under the Status columns, the Censored individuals are given 1 and dead individuals are assigned the number 2.
   1. Begin by producing a ggplot histogram that distinguishes between censored and uncensored data.
   2. Create a survival object and peek at the columns.
   3. Use ggplot2 to draw an attractive KM survival curve with the median indicated on it.
   4. Use basic R to draw the survival curves for maintained and non-maintained individuals.

**Module 5:**

***These questions mimic the analysis in the module5AB notebook so you might want to complete all of them if you are interested in correlation/regression/ANOVA in R since there is one question per each of these analyses. I have included the datasets in the module.***

1. The European cuckoo does not look after its own eggs but instead lays them in the nests of birds of other species (note: this seems like an excellent strategy in my opinion). We have seen previously that cuckoos sometimes have evolved to lay eggs that are colored similarly to the host birds' eggs. Is the same true of size -- do cuckoos lay eggs of different sizes in nests of different hosts? The data file "cuckooeggs.csv" contains data on the lengths of cuckoo eggs laid in a variety of other species' nests.
   1. Test for a difference between the host species in the size of the cuckoo eggs. Remember: You need to include all four steps when conducting a test (including testing and investigating the assumptions of the test you wish to use).
   2. Using a Tukey-Kramer test, identify which pairs of host species are significantly different from each other? If your result in **part a** failed to reject the null hypothesis, still conduct a Tukey-Kramer test and comment on its appropriateness and conclusions (are the conclusions consistent with the conclusion in part a?).
2. Animals that are infected with a pathogen often have disturbed circadian rhythms. (A circadian rhythm is a daily cycle in a behavior or physiological trait that persists in the absence of time cues.) Shirasu-Hiza et al. (2007) wanted to know whether it was possible that the circadian timing mechanism itself could have an effect on disease. They used three groups of fruit flies: one "normal", one with a mutation in the timing gene ***tim01*** (**heterozygous**= 1 copy of allele), and one group that had the **tim01** mutant in a **homozygous** state (2 copies). They exposed these flies to a dangerous bacteria, Streptococcus pneumoniae, and measured how long the flies lived afterwards, in days. The date file "circadian mutant health.csv" shows some of their data.
   1. Plot a boxplot that includes each of the three groups (not separate boxplots, all three groups should be on the same boxplot). Do these data match the assumptions of an ANOVA?
   2. Do an appropriate test to ask whether lifespan differs between the three groups of flies. (Hint: use a nonparametric test in this case – don't worry about transformations since you have already worked through the transformations in the notebook and they are seldom worth the time....).
3. Larger animals tend to have larger brains. But is the increase in brain size proportional to the increase in body size? A set of measurements on mammal body and brain size was collated by Allison and Cicchetti (1976), and these data are to be found in the data set “mammals.csv.” The file contains columns giving the species name, the body mass (in kg) and brain size (in g) of 62 different mammal species
4. Plot brain size against body size. Is the relationship linear?
5. Find a transformation (for either or both variables) that makes the relationship between these two variables linear. Is there statistical evidence that brain size is correlated with body size?
6. Can you use a parametric test or do you need to use the parametric version of the test? Why? Check the assumptions of the test. Run both the parametric and the nonparametric tests and compare their results. Do you have any explanation for any discrepancies?
7. What line best predicts (transformed) brain size from (transformed) body size. Superimpose this line onto your scatterplot.
8. Make a residual plot. Which species has the highest brain size for their body size? Which species has the smallest brain relative to that predicted by its body size? You will want to superimpose the names of the animals from the downloaded file onto the graph of the residuals. You can do this with the function >text(). To see how this works, an example is given below:

i.e. > plot(test\_mamamal.resid, col='purple', pch=19,cex=1,lty=”solid”,lwd=2)

> text(test\_mam.resid,labels=name,srt=60, cex=0.3,pos=3)

1. Are humans significantly different from the brain size predicted by our body size? To break this (general) question down into component: Conduct the appropriate test based on the assumptions that are met. Are there other outliers present? If so, give the names of two other mammals that are quite different.
2. Is there a clear correlation between brain size and body size?
3. Are there outliers present in this data set? Can you identify them