What STAT 1300 students should know (besides the learning outcomes)

A couple years ago I was asked what students coming out of a 1818 STAT/MATH 1300 course and entering college should know. I am finally getting around to attempting to answer that.

We want to empower our students to use data for good. A statistics course has the immediate application of helping the graduate formulate cogent data-based arguments on myriad important topics and communicate those with anyone. While math courses such as calculus are valuable and serve as prerequisites, this is one difference between statistics courses and basic math courses.

We want to impress on our students that they can make good sense of large datasets using data visualization and statistical tools in combination with a professional statistical software package. Students should know a professional statistical package (here R) is a powerful tool which they will have mastered enough to generate reports for, say, their subsequent neuroscience, sociology, or data driven history course. Further, their ability to work with that package will translate to their handling other software, such as MATLAB, for engineering courses. They should be aware that the goal is not to memorize the R commands but to be able to modify given ones to their data and purpose. They should be aware they can research and experiment with more nuanced aspects of R. But in this course, those should not be needed.

For more specifics, next are the learning outcomes and a few lines for each about why it's important- or other random thoughts. I encourage you to come up with your own dialog about why these outcomes are valuable and communicate that to your students. After the learning outcomes come R tasks.

SLU Course Learning Outcomes:

1. **Students will be able to create and interpret statistical graphs.**

I'd call this visual/data literacy. We want students to be curious about graphs and to relate them back to the source in a humanistic manner. What does that long tail indicate? If it is data about people, who is in the long tail? What does that mean for them? What do those two clusters in the scatter plot mean?

2. **Students will be able to compute and understand descriptive statistics.**

What is the central tendency and how much variability is there? What does that mean in context? How linearly related are these two variables? Can we use a line to model and predict?

3. **Students will be able to graph, compute and interpret regression lines.**

This is an example of using data to create a model (one of the simplest models we can make) to predict. We discuss the strength of that model and its limitations.

4. **Students will understand basic aspects of statistical sampling and experiments.**

If the method for collecting our data has flaws, its analysis shouldn't be trusted. But can data collection be perfect? What all methods are used for sampling? How do they compare? How are experiments designed? These are also topics students will see in later courses.

5. **Students will be able to perform probability calculations**.

Students often think in binary, deterministic terms and probability gives them more tools to think beyond those. Even in a topic such as history, considering what the odds were for an outcome leads to critical analysis.

And then probability is the basis for inference. Students should be aware of some basic probability models-at least binomial and normal- and how those two models often arise. They should see how frequently conditional probability occurs and be able to work with and understand Bayes' formula.

6. **Students will understand the sampling distribution**.

This is key to helping a student understand p-values. It's important for students to see simulations creating multiple samples and a partial sampling distribution.

7. **Students will demonstrate an ability to construct and interpret confidence intervals for one and two sample means and proportions**.

It's important to understand the parts in a confidence interval, both as part of the formula and in context of the application. They also should know there is one basic model for a confidence interval.

8. **Students will demonstrate an ability to construct and interpret hypothesis tests.**

They will see p-values in later courses and should know the full story-not just p-value <0.05 is of merit.

9. **Students will demonstrate an ability to apply R to generate graphs, statistics and reports from datasets for all covered material**.

Our departments that require this course for prerequisites or majors expect students to be able to do this, perhaps with a tad of coaching.

R tasks: Note: generally, most students will become adept at these. Working in teams, especially at first, may be useful. Note that for almost all of these you want to give students the command in an example where they copy and paste it and have them modify it to another example. You don't want them to be bogged down in memorizing or typing these commands.

(Just approximate order of difficulty)

1. (exploring home turf) Navigate the rstudio environment. If they are working with Posit, the online version, navigate that.
2. basic computations in the console. Basic operations on a dataset
3. Use built in datasets (see datasets in packages) to: create barplots, histograms, boxplots, scatterplots.

4.Use built-in datasets to compute mean, standard deviation, five number summary, correlation coefficient, regression line. Compare with graphs.
5.Import a larger dataset and analyze it using the above tools. (You rarely want students typing in data as hopefully they won't do that much in another class.)
6.Create some graphs where there is also a categorical variable. Examples: side by side boxplots of income based on region. Scatterplots of countries' literacy rate and life expectancy colored by region. Such comparisons show us even more the power of these graphs.
7. Do more computations where a categorical variable is involved. Example: medians of countries' literacy rates by region. Compare those to the graphs. Here we are getting closer to the detail a student in a subsequent course might want to have in a data analysis report.
8. Probability calculations for the normal distribution, for other distributions.
9. Use these probability calculations to calculate a z-confidence interval, or a p-value. Use these to do the same for a t-confidence interval where the sample mean and standard deviation are given.
10.Use t.test to calculate a p-value and a confidence interval from data.

Basic R is not meant to be hard. What is difficult is the choices one sometimes has with it. Sometimes there are nuances with data. But know you can ask me or there are other resources.

Also, here is a spreadsheet with basic commands which your students can paste into r. Other resources are posted in our 1818 collaboration folder.