Project Proposal For **Justin Papagelis** Stat 495

Answer the questions below to help you craft a proposal/synopsis.

What is the new statistical topic you will be exploring for the project? If you have a large topic area in mind (like “clustering” or “imputation”), you should work (with me!) to narrow it down to one or several sub-topics (particular methods) within the area.

The statistical topic that I will be exploring for the project is comparing multiple bootstrap methods in creating confidence intervals and how well they can perform.

Consider where you have encountered this topic. The audience for the paper is a peer in the class, who won’t necessarily have had the same electives as you. Besides details about the statistical topic, what other topics will your peers need information about / an introduction to in order to understand the topic? (For example, if you want to study a particular method of imputation, you are going to need to describe the types of missing data and imputation methods in general in order for it to be understandable to a classmate).

We have covered basic bootstrapping in this class and previous statistic classes so there does not need to be an introduction to the actual procedure, but perhaps there can be some explanation about how there are multiple methods of constructing bootstraps. Some of the individual bootstraps will surely need some background, but I have not really looked into that too much in depth.

Are you currently leaning more towards an application or simulation for the required second component? Provide a bit of your thinking about why this is your current thinking. You may or may not have very specific ideas here yet.

I am leaning towards a simulation for the required second component because it would make sense to simulate the different bootstrap methods and then compare them to each other and see how they perform.

If you have any other thoughts on planned applications or particular things you’d want to do in your simulation, feel free to list those here.

I might include the basic bootstrap in the simulation as a standard that the other bootstraps can be compared to as well.

A main part of the project is for you to prepare an expository review / exposition (an extremely solid Literature / Background / Introduction to the topic section, however you want to call it) of the topic. This may include additional information about the general statistical topic area. Thinking of your statistical and writing skills (and goals), what do you expect will be the most challenging aspect of preparing this exposition for you? (Expositions may have a code component too, with toy examples.) What skills will this section help you demonstrate?

I expect the most challenging part of preparing the exposition will be talking about some of the statistical theory in a way that is accessible and not confusing for a reader. This section will help me demonstrate my understanding of bootstraps because I will be able to communicate the theory to a different audience in a concise manner.

The application / simulation (or both, etc.) that you include will mean some coding in R (or another software). What do you expect will be the most challenging aspect of preparing this section and reporting the findings from your work? (This includes code, statistical, and writing skills). What skills will this section help you demonstrate?

I expect the most challenging aspect of the preparing this section is doing the actual code because the simulation that I will be performing might be code heavy. Reporting the findings will help me show my data analysis skills.

Now, let's synthesize all of this into a complete proposal via making it a project synopsis – e.g. you lay out what you plan to do. Here is an example assuming that random forests were the topic (this is not "new" for any of you though!), so I can demonstrate how all of these responses (from the previous pages) go together. See if you can see how each sentence relates to one of the questions above. Space for you to write your own synopsis is below the example.

Example Proposed Project Title: Using Random Forests to Predict Company Bankruptcy

Example Project Synopsis:

For my project, I plan to explore the use of random forests in supervised learning (classification) problems. My paper will introduce my peers to the classification setting, review the concept of decision trees in order to motivate forests, and then delve into the details of random forests, as well as touching on the pre-cursor of bagging. I will demonstrate how random forests work with a toy example such as predicting Species in iris in the exposition to help make points clear for my peers. I will also include several examples of applications of random forests in the literature to demonstrate recent use of the method. For my application, I intend to apply random forests to a larger data set with a binary response that indicates whether a company is bankrupt or not. I intend to use a new R package for fitting the random forest. Wrangling the data set will let me demonstrate appropriate wrangling skills and reporting my results will show my ability to communicate my data analysis process. I will include a comparison to logistic regression to showcase my modeling and variable selection skills in a regression setting.

Note that in the synopsis, I tried to write about intentions (if some things were uncertain) or write about what I "will" do. This helps provide a road map for you – you know you need to do those things. Your

synopsis may not have some of these details. You can write that you intend to find a data set that allows you to demonstrate skills X, Y, and Z, or that you intend to write a simulation, etc.

Your turn:

Proposed Project Title: A Performance Comparison of Bootstrap Methods

Project Synopsis:

For my project, I plan to explore different bootstrap methods for creating confidence intervals and then perform comparisons between a couple of the methods. My paper will re-introduce the simple bootstrap to my peers and then go deeper into the theory behind the construction of different bootstrap methods and the cases in which they are used. These can include the percentile bootstrap, bias-corrected bootstrap, accelerated bootstrap, and studentized bootstrap, as well as others. I will demonstrate how these bootstrap methods work using "toy examples," which will be datasets in which a specific bootstrap method is appropriate. To demonstrate my understanding of the methods, I will write a simulation to compare a few of them and determine how they perform against each other. I will include the basic bootstrap in the simulation to demonstrate how these other bootstrap methods compare to the standard. I will demonstrate my understanding of the different bootstrap methods by communicating the statistical theory in a concise and accessible way to my peers. Writing the simulation and sharing conclusions will demonstrate my ability to implement statistical methods in practice as well as my ability to analyze the results of the simulation.