**Stat 301 -- Mini-Project 1**

**Designing and Analyzing a Study (Random Sampling)**

The goal of Mini-Project 1 is to apply methods from Ch 1. The assignment is to take a *random sample* (meaning “statistically random,” that is, using scientific sampling methods) from a well-defined population or process and to analyze one binomial variable on the selected observational units.  Before you collect the data you should state a research question about the process/population parameter (e.g., do a majority of words in Webster’s new pocket dictionary have Wikipedia come up as the first entry in a Google search, on average is Food For Less cheaper than Albertson’s, are more than 90% of recipes in a vegetarian cookbook are actually vegetarian; see the[previous topics page](https://canvas.calpoly.edu/courses/3945/pages/project-topics) for more student ideas. Note: You should be very careful in how you define your process or population. Keep in mind that the observational units do not have to be humans.  If your population is humans, please do not ask sensitive questions.  Also note, your focus is not on comparing groups, just on making and testing conjectures about a process or population based on an honest-to-goodness random sample. You are free to choose your own topic(s).  The topic may be related to your major or another topic of interest. ***I encourage you to find a topic where you are assessing some claim you have recently heard***(e.g., more than 50% of income cell phone calls are junk calls, a majority of customers will correct a barista returning too much money, less than 25% of passerbys will help out someone who has fallen if they believe that person is homeless – this last two are from a recent episode of *Brain Games*.) Make sure you choose a topic so that it is straightforward to gather the data or you have access to data from another class or professor.  Please start early so you have time to ask questions.

**Population vs. Process:** You could sample from a process (repeated observations under identical conditions, where the randomness is in the individual outcome, e.g., recording whether or not a bowler hits at least 6 pins, whether or not someone waits less than 5 minutes at Starbuck’s).  In this case, you need ways to ensure that the conditions are identical (e.g., bowler is not getting tired, take every 3rd person in Starbuck’s line so the outcomes can be considered independent).  Or you can sample from a population where the randomness is not in the individual outcomes but in which individuals you measure, for example, whether or not someone has blue eyes.  The outcome (blue eyes) does not change with different observations of the same individual, but varies across individuals.  In this case, you need to consider probability sampling methods (see below).

**Selecting a Random Sample:** I can’t emphasize enough that I want a real *random sample* for this study. It does not need to be a “simple random sample” but should employ probability sampling methods.  See p. 94-95 for how to use technology select a random sample of ID numbers. Click [here](https://canvas.calpoly.edu/courses/3945/pages/sampling) for a brief overview of other probability sampling methods; feel free to ask me for more details on implementing these.  Be sure to identify the method you use. *The sample size should be at least 30*. If your population is not at least 20 times the size of your sample, then you will want to apply a “finite population correction factor.”

**Project Proposal:**Teams (up to three people) should be formed and preliminary project topics (research question, population of interest, sampling plan) submitted in Canvas by **Jan. 27** (but the sooner the better - if you do submit sooner, please email me to take a look!). Once I approve your topic you can begin collecting your data. **Receiving your initial proposal on time is worth 5 pts.**  Creativity and feasibility of topic will also be considered.

**Final Report:**  Due **Feb. 3**, but you are highly encouraged to complete the project sooner as review for the first exam.  This should be a typed report with screen captures of relevant computer output, written collaboratively by all team members.  Your report should be written as if it will be read by other student researchers (so using the common terminology but not overly technical).  Make sure it includes at least the following (and *include these section headings* to improve readability of report):

*Title, student names, Stat 301 – section(s), Winter 2020*

*I. Introduction* (5 pts) – Introduce/motivate your topic and research question. (You should address the questions below but in a self-contained, well flowing paragraph – not just bullet by bullet responses)

What is your research question? Why did you choose this topic?  What did you expect to find?  Have similar studies been done elsewhere (Cite them)?  Why should the reader be interested in your results and continue reading?  In addition, you should describe the population parameter of interest in words/context, your initial conjecture of its value (that makes sense in the context), and whether you suspected (before you saw any data) the actual parameter value is higher or lower (or just different) than this conjectured value (matching your research question).

 II. *Data Collection Methods* (10 pts)

* How (e.g., when, where) did you collect the data?
* What were the observational units?
* If sampling from a
	+ Population: Which type of probability sampling method did you use (SRS, stratified, cluster, systematic)? What was your “sampling frame” (p. 94)? What was the response rate?  How often did you have to make repeat visits in order to obtain the observational units initially selected?  Discuss any other potential sampling errors. Estimate the population size.
	+ Random process: What steps did you take to ensure your sample is representative of the process? Did you examine the behavior of the observations over time? Are there any other reasons to believe your observations may not be independent with identical probabilities of success?
* What is your variable of interest? How was the variable measured? If the variable is binary, how are you defining success and failure? Any “operational definitions” (e.g., did you only observe people writing or did you take any behavior such as throwing a football as indication of handedness?)?
* Discuss any potential sources of non-sampling errors.  For example: If you designed a survey, are there any potential wording issues?  Did you “field-test” the questions first?  How did you ensure confidentiality or take other precautions to ensure honest responses?  Include supporting documentation. For example, if you did use a survey, include a copy of the questions.
* Did you have any problems or other unexpected results?  Did anything go wrong during the course of the study?

**Note:** You can never give me too much detail in this section!  In particular, there should be enough information that someone else could replicate your study on their own based only on your description (and hopefully improve upon it based on your suggestions in section 4). You are describing your study *protocol* where someone else could mimic exactly the same study that you carried out.

*III. Analysis of Results* (20 pts)

*Descriptive Statistics*(3 pts)

* Look at your sample data numerically and graphically

You will need to make choices as to which numerical and graphical summaries are most relevant.  Make sure you introduce and integrate the output into the body of the report.

* Discuss what you learn from this output about your sample

Include discussions of how you are interpreting the message in these summaries.  In your discussion you should fully describe your sample, sample size, and report the sample statistic and whether it supports your conjecture.

*Inferential Statistics* (17 pts)

* Discuss what you learn about the larger population/long-run process
	+ Define the population or process and parameter (again) in words
	+ State the null and alternative hypotheses in symbols and in words
	+ State what a type I and a type II error would represent in this setting and briefly describe the potential consequences of each type of error
	+ Assess the appropriateness of the binomial model
	+ Assess the appropriateness of the normal approximation to the binomial for your study
		- Mostly I’m asking about sample size but this also includes whether you think you have a reasonable random process (i.e., observations are independent and come from a stable process)
	+ Calculate the test statistic
	+ Calculate the exact p-value corresponding to your alternative hypothesis
		- Include a well-labeled graph representing your p-value
	+ Calculate (if appropriate) the theory-based p-value
		- Include a well-labeled graph representing your p-value
	+ Provide a one-sentence interpretation of the p-value in context
	+ Indicate what conclusion this p-value leads you to draw about the null hypothesis
	+ Calculate and interpret a confidence interval to describe the plausible values of your population parameter

           Make sure you include all relevant output in the body of your report.

*IV. Conclusion* (5 pts)

Summarize the results of your study (there will be some repetition, and you should cite your evidence).  You should tell a story: What did you learn?  Did the data behave as you expected?  Critique the methods you used to collect the data. Pay particular attention to whether or not it is reasonable to generalize your sample to the larger population or process. Is there anything you would do differently next time?  How might this affect the conclusions of the study?  What similar questions might someone chose to investigate in the future to build on your results?

*Presentation*(5 pts) - style, organization, layout, grammar, presentation of written report, creativity

*Initial proposal submission* (5 pts)

*Group member evaluation*

*Submit here:*

research question, population of interest, sampling plan