MAT7500 – Statistical Programming
Simulation Project

Purpose: The purpose of this project is to provide you with experience in solving a statistical problem using simulation techniques, including presentation of results and writing up your findings. Simulation is used increasingly in statistical analysis when theoretical derivations are too cumbersome, intractable, or difficult to define.

Requirement: A thorough simulation should be done to answer your problem. The project requires an element of randomness that you can simulate in SAS or R. For most of you, there will be a number of parameters that you can vary in order to determine the robustness of your results. You should determine what measure or outcome will be used to demonstrate relationships or success. Recall that simulation involves:

1. Describe the possible outcomes.
2. Link each outcome to one or more random numbers
3. Choose a source of random numbers.
4. Choose a random number.
5. Based on the random number, note the "simulated" outcome.
6. Repeat steps 4 and 5 multiple times; preferably, until the outcomes show a stable pattern.
7. Analyze the simulated outcomes and report results.

Project Proposal: **Due March 27, 2017**. For the proposal, you should include:

1. Your name,
2. The goal of your project (1-2 sentences),
3. A brief overview of the approach that you will take, including the way that you will measure the outcome.

I expect that some of you will change your proposals, either minor changes or a complete overhaul, which is fine. If you have multiple topics in mind, you are welcome to either speak with me about them ahead of time or submit two topics as your proposal, and we can discuss the relative merits of each one.

Project Meetings: I will ask each person to meet with me at least once after the project proposal is complete. You are welcome to meet with me more times, if you wish.

Final Project Presentations: Project presentations will take place on **May 1st and May 8th**. Each person will give a 15-20 minute presentation, including framing the problem, a brief description of your coding (no more than three slides), your results, and a discussion of limitations and what you could have done better. Others will be given the chance to ask questions. The date and time of presentations have been randomly assigned. If you need to switch your time, please coordinate with another student and let me know.

Final Submission: In addition to your presentation, you are asked to submit your code, so that I can review it, as well as a write-up of your presentation. This is where you can and should present more details about how and why the simulation was done..

Last Name	First Name	Date	Time		Last Name	First Name	Date	Time
Shi	Youhan	May 1	1:30		Fung	James	May 1	6:15
Reynolds	Claire	May 1	1:55		Rozran	Jacob	May 1	6:40
Zhuang	Bing	May 1	2:20		Boe	Lillian	May 1	7:05
Innocenti	Christopher	May 1	2:50		Lezis	Andreas	May 1	7:40
McKevitt	Erin	May 1	3:15		O'Brien	Jerome	May 1	8:05
Youn	Nara	May 1	3:40		Lake	Anna	May 1	8:30
Vangumalli	Dinesh	May 8	1:30		Diferdinando	Aaron	May 8	6:15
Cano	Andrew	May 8	1:55		Flynn	Kevin	May 8	6:40
Shukr	Bayan	May 8	2:20		Harder	Brett	May 8	7:05
Harder	Shane	May 8	2:50		Robbins	Katie	May 8	7:40
Tang	Chuanhai	May 8	3:15		Donohue	John	May 8	8:05
Kluge	Rob	May 8	3:40		Lin	Si Xue	May 8	8:30
					Smith	Ian	May 8	8:55

Project ideas:

- Softball team lineup. Goal: Using existing data on batting success of players, determine the optimal batting lineup. Use previous data on likelihood of each player hitting a single, double, triple, homerun, or out, permuted every possible batting order, and used random number generators to simulate 100 games for each batting order. There were a number of constraints placed on this, including (being a co-ed league) needing at least one person of each gender within the first three, second three, and third three batters as well as a cap of 12 runs per inning. In the end, I ranked the batting orders based on their average number of total runs for the seven inning game.
- Nonparametric efficiency. Goal: Can you determine how efficient a nonparametric test is compare to its parametric alternative based on the true underlying distribution and the sample size? You will have to define efficiency. Perhaps choose uniform distributions, normal distributions, and strongly skewed distributions and simulate random draws of data when the null hypothesis is true. This should help you determine the true alpha of the test and compare it under parametric tests and nonparametric tests.
- When is the normal approximation to the binomial appropriate? Some textbooks suggest verifying that np and n(1-p) are both at least 5 in order to know that the normal approximation to the binomial is accurate, while some suggest using 10 or even 15. Can you determine which one, and under what conditions, is best? The outcome measured might be how close the binomial is to the normal distribution or the difference between the exact binomial probability (with and without continuity correction) and the approximated normal. For varying values of n, p, and condition (5, 10, or 15), how accurate is the normal approximation to the binomial?
- Previous Year Projects
  - Simulating false positive and negatives as a function of clinician bias to examine misdiagnosis of mental illness
  - Driving vs. public transportation to work analysis through simulation of weather patterns
  - Identifying the best fantasy football team
  - When is estimating the mean and variance from the median and range valid
  - March Madness simulation
  - Golf shot analysis – simulation 100 rounds of golf, examining mastery via angle simulation (experts have smaller variation in angle of shot), using trigonometry
  - Blackjack – comparing strategies of card counting
  - Non-random mixing of Mike and Ike candies – how many mixes is optimal (like examining a perfect shuffle) – assumed local mixing, but not global mixing
- Other ideas:
  - What types of different inference would be made based on the MAD vs. the standard deviation?
  - What is the effect on measuring standard deviation using absolute deviations from the mean vs. squared deviations vs. absolute cubic deviations vs. quartic deviation?
  - How does the introduction of missing data affect your analysis? What if the likelihood of being missing is a function of the explanatory variables or the response variable?
  - How different are inferences from a Bayesian analysis with a Frequentist analysis based on the same sample size?
  - If you apply suboptimal sampling strategies, how much power are you losing?