Final Exam

STAT 343: Mathematical Statistics

- This exam is **due to Gradescope at 11:59 PM ET on Tuesday, May 11, 2021.** This is a firm deadline.
- There are three exam questions. You must show all your work to receive full credit.
- This is a **open** notes, **open** book, **open** Moodle exam. Other resources, besides R, **are not permitted.** Make sure to show all your work and include all your code for full credit.
- Once you have accessed the exam, you may only ask for clarification on questions from the instructor. Do not communicate with any other students about the exam until after the exams have been returned. Any such communication is a violation of the Honor Code.
- If you become aware of any honor code violations, you have a duty to report them to the Honor Code Council.

Problem 1

It is spring time, so it is time for gardening! I usually like to a get a jump start on the season and start my plants from seed. I planted 20 tomato seeds, and I am interested in modeling how many seeds germinate (sprout from seeds and become plants). I don't want to give any seeds an unfair advantage, so I make sure that they all come from the same seed company and the same packaging year, are planted in the same kind of soil, and have the same temperature, amount of light and water, and are spaced far enough apart that they do not compete for space and nutrients. This makes me reasonably confident that whether a seed germinates is independent of another and has the same chance of germinating.

(a) What is an appropriate statistical model for X, where X is the number of seeds that germinate? Justify your answer in 1-2 sentences to receive full credit.

(b) Explain in 1-3 sentences why it is appropriate to think about X as a random variable in this problem.

(c) My seed package tells me that I should expect 70% of my seeds to germinate, on average. However, I am an experienced gardener and believe that I can get 80% of the seeds to germinate. Write down a set of suitable hypotheses for this test. To get full credit, please express your hypotheses in terms of Ω_0 and Ω_A, where these sets are defined in the context of the problem.

(d) Are the hypotheses expressed in (c) simple or composite?

(e) Consider a test based on the statistic $W = X$, where X is still the number of seeds that germinate. What is the distribution of W under the null? What is the distribution under the alternative? Make plots (histograms) of each of these distributions using R and make it clear which plot corresponds to which hypothesis.

(f) Suppose 15 of my seeds germinate. Carry out the test specified in (c) using this observed value at a significance level of 0.1. Is there enough evidence to reject the null hypothesis? In carrying out your test, be sure to specify what you are considering to be "at least as extreme as" when defining your p-value. You will need R to answer this question – please include all of your code for full credit on this problem.

(g) Suppose a classmate decides to conduct this test (with the same hypotheses) based on the likelihood ratio test, still at a significance level of 0.1. Citing the Neyman Pearson lemma, this classmate claims that their test is **more powerful** than the one you conducted in (c). Show why this is **not** true.

Problem 2

Recall that the size of a test is

$$P(\text{Reject } H_0 | H_0 \text{ is true}) = \alpha.$$

(a) One criticism of tests based on discrete distributions is that in many cases it is difficult to specify a test of exactly size α. Consider the test in Problem 1 based on $W = X$. What is the true size of that test?

(b) A potential remedy to the problem presented in (a) is to conduct the test based on a *normal distribution approximation.* In this case, a test statistic that is often used is

$$W = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{\theta}\left(1 - \hat{\theta}\right)/n}},$$

where $\hat{\theta} = \frac{X}{n}$ is the maximum likelihood estimator.

(i) In the context of Problem 1 (same n, etc.), do you expect the size to be **exactly** α? Justify your answer briefly.

(ii) Suppose n is really large (e.g., 1,000,000,000) - should the size be α in this case (or at least very very close)?

Problem 3

Take one homework problem you have worked on this semester that you struggled to understand and solve and explain how the struggle itself was valuable. In the context of this question, describe the struggle and how you overcame the struggle. You might also discuss whether struggling built aspects of character in you (e.g., endurance, self-confidence, competence to solve new problems) and how these virtues might benefit you in later ventures.