STAT 520 Midterm 1

This midterm has 5 questions, some of which have multiple parts. All questions concern a study to estimate the probability that a planted cucumber seed produces at least one cucumber. The study is described on page 2 of this packet. The questions begin on page 3.

Please write your answers on separate sheets of paper. When you are finished, please post a scan or photograph of your answers to the Midterm 1 Assignment on the course Canvas page. I think that the midterm is self-explanatory, but if you have questions, please message your proctor through the private chat.

Cucumber Study Description:

A horticulturalist would like to estimate the probability that a planted cucumber seed survives. A seed is declared to have survived if it grows into a cucumber plant that produces at least one cucumber. A seed may fail to survive for many reasons. The seed itself may be defective, or the surrounding conditions may not be conducive to survival. The horticulturalist uses the information on the survival probabilities of cucumber seeds for several purposes. For example, the estimated survival probability determines the number of seeds to include in a package for sale. The horticulturalist's primary objective is to conduct inference for the survival probability of a cucumber seed.

The horticulturalist has two varieties of cucumber seeds: variety A and variety B. The horticulturalist has 2 seeds from variety A and 3 seeds from variety B. One specific goal of the study is to estimate the survival probability of cucumbers from each of the two varieties. The horticulturalist would also like to understand if the survival probabilities from the two varieties are the same or different.

The horticulturalist conducts a study to estimate the survival probabilities for the two varieties. The horticulturalist acquires 5 pots that are identical in terms of their size and general design. The horticulturalist divides soil from a batch of a single soil type among the 5 pots. Each pot has exactly the same amount of soil. The horticulturalist plants each of the 5 seeds in a randomly selected pot. The horticulturalist monitors the growth of the seeds over an 8-week period. During the growing period, the pots are subjected to identical growing conditions. The plants experience the same level of lighting. The horticulturalist applies the same amount of water to each pot at the same time of day. At the end of the 8-week growing period, the horticulturalist is able to determine which plants survived.

The horticulturalist records a binary indicator of whether or not the plant in a pot survived. The binary indicator is 1 if the plant survived and is 0 otherwise. Table 1 summarizes the collected data.

		Survival Indicator
Variety	Plant	(1 = survived; 0 = did not survive)
A	1	1
A	2	1
B	1	1
B	2	0
B	3	1

Table 1: Data from cucumber study.

Questions:

1. Parts 1(a)-1(c) will guide you do discuss how the study relates to the concepts of scientific and statistical abstraction that we discussed in Chapter 1. (Note: You only need to answer 1(a)-1(c); you do not need to write anything between 1 and 1(a).)

   (a) Name 1 measure that the horticulturalist took to exercise control over the study conditions.

   (b) The horticulturalist naively observes that 100% of the plants from variety A survived, while 66% of the plants from variety B survived. Does that mean that planting cucumbers from variety A *causes* the plants to have a higher survival rate? Explain why or why not.

   (c) Define appropriate random variables and associated sample spaces for this study.

2. Name 1 limitation of using a normal distribution as a model for the data collected in the study.

3. Now, we will consider only the data for variety B. Assume that random variables associated with the survival indicators for the 3 plants have independent Bernoulli distributions with success probability p_B. Recall that the pmf of a random variable having a Bernoulli distribution with success probability p_B is of the form

$$f_Y(y \mid p_B) = p_B^y(1-p_B)^{1-y}, \quad p_B \in [0,1]; y \in \{0,1\}. \tag{1}$$

Answer questions 3(a)-3(k) below.

(a) Let $\ell(p_B)$ be the log likelihood function of p_B based on the data from variety B. Give a mathematical expression for $\ell(p_B)$.

(b) Let $U(p_B)$ be the score function for p_B. Give a mathematical expression for $U(p_B)$.

(c) What is the maximum likelihood estimate of p_B?

(d) Give a mathematical expression for the expected information in a random sample of size 3 from the probability density function (1).

(e) Construct a 95% Wald interval for p_B.

(f) The log odds of survival is defined as

$$\theta = \log\left(\frac{p_B}{1-p_B}\right).$$

Estimate θ and provide a corresponding standard error for your estimator.

(g) Express the pmf (1) in canonical form. What is the canonical parameter? What statistic is sufficient for the canonical parameter?

(h) Use your canonical parametrization to derive an expression for $E[Y]$ as a function of the canonical parameter, where Y is a Bernoulli random variable with pmf (1). Show the steps of your derivation.

(i) Is it true that the probability density function (1) is a member of the natural exponential family? You only need to write “Yes” or “No.”

(j) Is it true that the probability density function (1) is a member of the exponential dispersion family? You only need to write “Yes” or “No.”

(k) Is it true that the probability density function (1) forms a location/scale family? You only need to write “Yes” or “No.”

4. We will now consider the problem of comparing the survival probabilities for the two varieties. A full model allows the survival probabilities for the two varieties to differ. Formally, the full model allows for the possibility that $p_A \neq p_B$, where p_A and p_B are the survival probabilities for varieties A and B, respectively. A reduced model restricts

the two varieties to share a common survival probability. Under the reduced model, $p_A = p_B$, where p_A and p_B are the survival probabilities for the two varieties. Conduct a likelihood ratio test of the null hypothesis $H_o : p_A = p_B$ against the alternative hypothesis, $H_1 : p_A \neq p_B$. (Hint: For the full model, you may wish to first evaluate the likelihood and then take the log. Remember that $0^0 = 1$. To calculate your p-value from the standard normal table, remember that the square of a standard normal random variable has a chi-square distribution with 1 degree of freedom.)

5. The horticulturalist conducts a separate study to assess the effect of soil type on the probability that a plant survives. The three soil types of interest have complex names that the horticulturalist abbreviates as Loam (L), Silt (S), and Clay (C). The horticulturalist assigns 1 plant to each of the three soil types in alphabetical order. All of the plants are from variety B. Note that exactly three plants are used in the study. What is the most important problem with the study design?