ECE/CS 498 DSU/DSG Spring 2020
In-Class Activity 1

Names: ____

NetIDs: ____

The purpose of the in-class activity is for you to:

(i) Review how to go from the word description of a problem to mathematical equations and probabilities.
(ii) Understand conditional probabilities and their applications
(iii) Understand hypothesis testing - in particular, the use of p-values for accepting or rejecting hypotheses

Problem 1:
PUBG is a well-known video game that many college students play. James and Haotian are two PUBG enthusiasts, and you, as a PUBG team manager, are picking one of them to play for you at the Champaign-Urbana tournament. The winner is rewarded with a chicken dinner. **You really want the chicken dinner, but you don't know who the better player is.** All you know is that one of the two players (either James or Haotian) has a PUBG winning rate of 80% (independent of previous rounds), while the other player has a PUBG winning rate of 50% (also independent of previous rounds). Therefore, you decide to let James and Haotian play a few individual rounds of PUBG to figure out who the better player is.

Let R_J denote the event that James is the one with a higher winning rate (i.e. the better player), and R_H denote the event that Haotian is the one with a higher winning rate. Since you do not have any *prior* information on who the better PUBG player is, we assume that both James and Haotian have an equal chance of being the better player (i.e. $P(R_j) = P(R_H) = 0.5$).

James plays the first round, and he loses. We are interested in the probability that James is the better PUBG player (i.e. the one with the higher winning rate).

Let $\bar{G}$ denote the event that the **James** did not win this round of PUBG. Assume that the outcomes of successive rounds are independent given that someone is the better player.

1. Which are the independent variables, and which are the dependent variables in the above scenario?
   Independent variable: R_J, R_H
   Dependent variable: $\bar{G}$

2. What are the conditional probabilities that you need to find $P(\bar{G})$? Express the probability using the theorem of total probability.

   By the theorem of total probability: $P(\bar{G}) = P(\bar{G}|R_H)P(R_H) + P(\bar{G}|R_J)P(R_J)$

   So we need the following conditional probabilities: $P(\bar{G}|R_H), P(\bar{G}|R_J)$

3. Find the value of $P(\bar{G})$, i.e., the probability that **James** did not win the round of PUBG.

$$P(\bar{G}) = P(\bar{G}|R_H)P(R_H) + P(\bar{G}|R_J)P(R_J)$$
$$= 0.5 * 0.5 + 0.2 * 0.5 = 0.35$$

4. What is the probability that James is a better PUBG player, given that he did not win the first round, i.e., $P(R_J|\bar{G})$?

$$P(R_J|\bar{G})$$
$$= \frac{P(\bar{G}|R_j)P(R_j)}{P(\bar{G})}$$
$$= \frac{0.2 * 0.5}{0.35} = \frac{2}{7}$$

5. The above exercise allowed you to update your *prior* about which PUBG player is a better player. As a result, it reduced your uncertainty of which player is the better one. To be a bit more certain, you chose **James** to play another round of PUBG. **This time, James wins the round!** Based on these two rounds of games, what is the probability that the James is the better PUBG player?
   a. For this question, you can denote $\bar{G}_1$as the event that James loses the first round and G_2 as the event that James wins the second round. Write down the conditional probability formula that models what you are trying to calculate.

$$P(R_J|G_2, \bar{G}_1) = \frac{P(R_j, G_2, \bar{G}_1)}{P(G_2, \bar{G}_1)} = \frac{P(G_2, \bar{G}_1|R_J)P(R_J)}{P(G_2, \bar{G}_1)}$$

   b. Using the information given, actually calculate the probability from part (a). Did the second attempt increase your uncertainty about which player has the higher scoring probability?

Method 1: Using conditional independence

$$P(G_2, \bar{G}_1) = P(G_2, \bar{G}_1|R_H)P(R_H) + P(G_2, \bar{G}_1|R_J)P(R_J)$$
$$= P(G_2|R_H)P(\bar{G}_1|R_H)P(R_H) + P(G_2|R_J)P(\bar{G}_1|R_J)P(R_J)$$

The second equality comes from our initial assumption where the outcomes of the two games are independent given that someone is the better player.

$$P(R_J|G_2, \bar{G}_1) = \frac{P(G_2, \bar{G}_1|R_J)P(R_J)}{P(G_2, \bar{G}_1)}$$
$$= \frac{P(G_2|R_J)P(\bar{G}_1|R_J)P(R_J)}{P(G_2, \bar{G}_1)}$$

$$= \frac{0.8 * 0.2 * 0.5}{0.5 * 0.5 * 0.5 + 0.8 * 0.2 * 0.5} = \mathbf{\frac{16}{41}}$$

Method 2: Using the updated prior based on the outcome of the first round

The outcome of the first round resulted in changing the *prior* for James having a higher probability of winning. As a result, we can think of the second round as being an attempt with the updated *prior*. We can write,

$$\begin{aligned} P(G_2|\bar{G}_1) &= P\left(G_2\middle|R_J, \bar{G}_1\right)P\left(R_J|\bar{G}_1\right) + P(G_2|R_H, \bar{G}_1)P(R_H|\bar{G}_1) \\ &= P\left(G_2\middle|R_J\right)P\left(R_J|\bar{G}_1\right) + \ P(G_2|R_H)P(R_H|\bar{G}_1) \\ &= (0.5) * \frac{5}{7} + (0.8) * \frac{2}{7} \\ &= \frac{41}{70} \end{aligned}$$

The second equality comes from our initial assumption where the outcomes of the two games are independent given that someone is the better player.

We can now compute the updated posterior based on the updated prior and the outcome of the second round.

$$P\left(R_J\middle|G_2, \bar{G}_1\right) = \frac{P\left(R_j, G_2, \bar{G}_1\right)}{P(G_2, \bar{G}_1)}$$

$$= \frac{P(G_2|R, \bar{G}_1)P\left(R_J|\bar{G}_1\right)P(\bar{G}_1)}{P(G_2|\bar{G}_1)P(\bar{G}_1)}$$

$$= \frac{P(G_2|R, \bar{G}_1)P\left(R_J|\bar{G}_1\right)}{P(G_2|\bar{G}_1)}$$

$$= \frac{(0.8) * \frac{2}{7}}{\frac{41}{70}} = \mathbf{\frac{16}{41}}$$

Your posterior for James being the one with higher probability of scoring went from 2/7 after the first attempt to 16/41 after the second attempt. Therefore, as his posterior became closer to 50%, the uncertainty that he is a better player increased.

Problem 2:
The San Diego Comic Convention (usually referred to as Comic-Con) is an international convention that takes place over a period of four days in July every year. It is one of the world's most widely attended conferences, with attendance exceeding 130,000 people.[1]

You are working as an administrator at Comic-Con. Suppose that there is an exclusive event for Spider-Man fans, where there will be a special trivia contest and the first-time screening of an upcoming movie. Of course, many of the people at Comic-Con will want to get into the event to get a sneak peek at the new movie. Your manager entrusts you with the responsibility of ensuring that people who are admitted into this event are actually Spider-Man fans who are knowledgeable about the web-slinging superhero.

In order to distinguish the true and fake fans, you recall the hypothesis testing techniques taught in your data science course ☺ You decide to administer a 100-question multiple choice test to those who try to get into this event. Each question has 2 possible answer choices. Vikram is the first person to take this exam. If Vikram was a true Spider-Man fan, then he would be able to answer the questions with an accuracy better than chance. Otherwise, he would just have to randomly guess which answer choice is correct. Let p be the probability that Vikram answers an individual question correctly (assume this probability is constant). Of the 100 questions, Vikram answers 34 of them correctly.

[1]: https://www.comic-con.org/about

1. Using mathematical notation, write down null and alternative hypotheses for a one-sided test with the claim that Vikram is a true Spider-Man fan.

Vikram can be said to be a true Spider-Man fan if $p > 0.5$ (better than chance level). This gives us the following hypothesis test:
H_0: $p \leq 0.5$
H_1: $p > 0.5$

2. If the test statistic is the number of correct answers (34) from 100 questions, write down the formula for calculating the p-value. [Hint: Think about the probability for a binomial random variable X...]

Recall that the p-value tells us the probability of the observed test statistic or a more extreme one <u>assuming the null hypothesis is true</u>. Since we are performing a one-sided test, we want to check the probability of Vikram correctly answering 34 or more questions under the null hypothesis distribution (i.e. binomial distribution with probability of 0.5 and 100 trials). Let $X \sim Binomial(100, 0.5)$.

$$\text{p-value} = P(X \geq 34) = 1 - P(X \leq 33)$$

3. Continued from part (b). If you only know that $P(34 \leq X \leq 66) = 0.9991$, how would you find the exact p-value of the test? [Hint: use the fact that a binomial distribution with an individual trial success probability of 0.5 is symmetric…]

Again, let $X \sim Binomial(100, 0.5)$.

$$1 = P(X \leq 33) + P(34 \leq X \leq 66) + P(X \geq 67)$$

From symmetry of this binomial distribution (P(X=k) = P(X=100-k)), we get $P(X \leq 33) = P(X \geq 67)$. Substituting this, we get

$$P(X \leq 33) = \frac{1 - P(34 \leq X \leq 66)}{2} = \frac{1 - 0.9991}{2} = 0.00045$$

p-value $= 1 - P(X \leq 33) = 1 - 0.00045 = 0.99955$

4. State your conclusion about Vikram's Spider-Man fandom for a significance level of 0.05. Does this agree with your intuition about Vikram's performance?

A high p-value means that the chance of answering 34 out of 100 questions correctly is high even for a fake fan. For a significance level of $\alpha = 0.05$, since the p-value=0.99955 $> \alpha$, we fail to reject the null hypothesis. Therefore, we cannot accept the claim that Vikram is a true Spider-Man fan, and he should be denied admission into the event.

Intuitively, this makes sense since the expected score with truly random guessing would be 50/100, and Vikram's performance of 34/100 is actually worse than that. We've now provided statistical justification for not admitting Vikram into the event.

5. What if Vikram had correctly answered 67 out of the 100 questions? Assuming a significance level of 0.05, again state your conclusion about Vikram's Spider-Man fandom.

Our null and alternate hypotheses still remain the same in this case. Moreover, it still holds that $X \sim Binomial(100, 0.5)$.

Since Vikram correctly answered 67 questions, we can formulate the p-value as p-value= $P(X \geq 67)$.

As we described in question 3 above, $P(X \geq 67) = P(X \leq 33)$, which we already calculated. Thus, p-value= $P(X \geq 67) = 0.00045$.

Since the p-value= $0.00045 < \alpha = 0.05$, we reject the null hypothesis. Therefore, we accept the claim that Vikram is a true Spider-Man fan, and he should be granted admission into the event.