CSE 3521

Homework 4 - Probability

Due 11/3 at 11:59pm

Learning Goals What I want to evaluate with this assignment is:

- Understanding of Bayes' Rule and its application to statistical inference.
- Ability to calculate single-variable, joint, and conditional probabilities from given information.
- Ability to formulate and evaluate well-formed, testable hypotheses for statistical hypothesis testing.

General Instructions Show your work for all questions. Even if you get the final answer wrong, you can get partial credit if you show me that you're taking the right approach.

Submission Instructions You can either bring answers on paper to class on 11/3 or submit electronic answers to these questions (in PDF form) via Carmen.

1. (1 point) Bayes' Rule states that $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

   Explain what this theorem is saying, using the terms *prior* and *posterior*. (1-2 sentences)

2. Your doctor has recommended that you get a particular test done to see if you have tuberculosis. You are given the following information about the test:

   - The base likelihood of anyone having tuberculosis is 0.01%.
   - The recommended test has a false positive rate of 1%. In other words, if you have tuberculosis, the test will be positive 99% of the time, and if you don't have tuberculosis, the test will be negative 99% of the time.

   You take the test and test positive for tuberculosis.

   (a) (2 points) Given the positive test, what is the likelihood that you actually have tuberculosis?

   (b) (Extra Credit: 1 point) You want to make the test better. What is the maximum false positive rate that would mean the probability that you have tuberculosis given a positive test is at least 50%?

3. You want to construct a probability distribution over three binary random variables that describe my lunch: *Cheese*, *Bread*, and *Relish*.

   You know the following information:

   - $P(cheese) = 0.4$
   - $P(bread|cheese) = 0.8$

- $P(\neg bread|\neg cheese) = 0.3$
- $Relish$ is conditionally independent of $Bread$, given $Cheese$
- $P(relish, cheese) = 0.2$

(a) (1 point) Do you have enough information to compute the full joint probability table over $Cheese$, $Relish$, and $Bread$?

(b) (1 point) If so, calculate $P(bread, \neg relish, \neg cheese)$.
If not, what information is missing to determine all the joint probabilities?

4. For this question, use the joint probability table given below. It involves three random variables: $A \in \{a_1, a_2, a_3\}$, $B \in \{b, \neg b\}$, and $C \in \{c, \neg c\}$.

	c		$\neg c$	
	b	$\neg b$	b	$\neg b$
a_1	0.012	0.09	0.049	0.063
a_2	0.024	0.072	0.098	0.084
a_3	0.084	0.018	0.343	0.063

(a) (1 point) Calculate $P(c)$.

(b) (1 point) Reduce this to calculate the joint probability table over A and B only.

(c) (1 point) Calculate $P(\neg b|\neg c)$.

(d) (1 point) Is C independent of A or B?

(e) (2 points) Is A conditionally independent of C, given B?

5. Let S_1 and S_2 be two sets of random samples from the OSU student population, where the random variable we're sampling for is current year in the undergrad program. We want to formulate a **null hypothesis** (a particular relationship is not present) and an **alternative hypothesis** (the relationship is present) about these data.

(a) (2 points) Consider the following hypotheses:

Null hypothesis S_1 has more seniors than S_2.
Alternative hypothesis S_2 has more seniors than S_1.

Are these well-formed, testable hypotheses? If not, how could you change them to be well-formed? (1-2 sentences)

(b) (2 points) Consider the following new hypotheses:

Null hypothesis S_2 and S_1 give equally good estimates of the true probability distribution over current year in college.
Alternative hypothesis S_2 gives a better estimate of the true probability distribution than S_1.

Are these hypotheses testable? If so, how? If not, why not? (1-2 sentences)

6. You are working on developing a food delivery robot, that needs to be able to navigate through foot traffic on campus. You have two test robots, one using old software and the other with updated algorithms.

   Over 3 days, you've taken every order you get and assigned it randomly to one of the robots. In each delivery, you measure:

   - The amount of time the delivery takes (in minutes)
   - The number of people your robot bumps into on the delivery
   - The customer's satisfaction rating after the completed delivery (on a 5-point scale)

   These data are given in the table below; S_1 corresponds to the robot with old software, and S_2 is the robot with new software.

	S_1			S_2	
Time	Bumps	Rating	Time	Bumps	Rating
28	26	3	20	27	4
18	14	5	14	3	5
6	10	4	44	30	4
6	3	4	34	18	2
10	12	2	3	4	5
53	30	2	27	19	3
77	44	1	33	9	5
22	15	4	9	8	1
61	37	3	5	2	4
28	32	5	11	13	3
			89	50	1
			30	19	4

   You want to see if the robot with the new algorithm (S_2) does better than the one with the old routines (S_1). You can consider any or all of the three things we're measuring ($Time$, $Bumps$, and $Rating$).

   Answer the following questions with 1-2 sentences each.

   (a) (2 points) Formulate a testable <u>null hypothesis</u> and <u>alternative hypothesis</u> about the performance of the new robot compared to the old robot.

   (b) (1 point) You can choose Student's t test or Pearson's χ^2 test as your test statistic. Which one would you use for testing your hypotheses? Justify your answer.

   (c) (2 points) Use the data given above and the test statistic you chose in part (b) to evaluate your hypotheses. Use a significance level of $p < 0.05$. Is your alternative hypothesis supported?

   Tip: the Scipy scientific library for Python includes methods for calculating both Student's t test and Pearson's χ^2. Scipy is not installed on stdlinux, but you can install it locally.