B365 Homework 3

1. Bayes' rule can be generalized to random variables X, Y. Specifically, suppose if $P(X = x)$ gives the *prior* probability for the event $X = x$ for generic outcome x, and $P(Y = y|X = x)$ gives the conditional probability for $Y = y$ *given* $X = x$. Suppose X can take outcomes $\{x_1, \ldots, x_n\}$ and we observe $Y = y$. Bayes' rule generalizes to this case, giving the probability that $P(X = x_i|Y = y)$ by

$$\begin{aligned} P(X = x_i|Y = y) &= \frac{P(X = x_i, Y = y)}{P(Y = y)} \\ &= \frac{P(X = x_i)P(Y = y|X = x_i)}{P(X = x_1)p(Y = y|X = x_1) + \ldots + P(X = x_n)p(Y = y|X = x_n)} \end{aligned}$$

Suppose a state has voters from three parties, A, B, C with proportions 20%, 30%, and 50% respectively. A certain ballot measure is favored by 30% of the members of party A, 50% of the members of party B, and 90% of the members of party C. A randomly selected person is found to *not* favor the ballot measure.

   (a) Using the above formula, write an expression for $P(A|\text{not favor})$ using

$$P(A), P(B), P(C), P(\text{favor}|A), P(\text{favor}|B), P(\text{favor}|C)$$

   (You may not need all of these).

   (b) What is the probability this person is a member of party A?

   (c) What is the probability this person is a member of party B?

   (d) What is the probability this person is a member of party C?

   (e) Simulate this experiment in R with 10,000 individuals and, using your experiment, give an estimate of the probability of political party A for individuals who favor the ballot measure.

   (f) Suppose we attend the Rocky Horror Picture Show and meet a person at this movie. In conversation it comes up that the person does not favor the measure. Does the probability computed above in part 1 apply to *this* individual? Be sure to explain your reasoning in either case.

2. Consider the University of California, Berkeley (UCB) admissions data discussed in class, and explored in the simpsons_paradox.r example.

   (a) Construct and print the two-way table of the department and the admit status, as well as a mosiac plot of the table. Do department and the admit status appear independent? Justify your answer.

   (b) Construct and print the two-way table of department and gender, as well as a mosiac plot of the table. Do these variable appear independent? Justify your answer.

   (c) Construct a two-way table of gender and admit status for the applicants to department F and create its mosaic plot. For the applicants to department F do gender and admit status appear to be independent? (This question could be rephrased to ask if admit status and gender *conditionally independent* given department F).

   (d) Create a 1-way table of the gender of Rejected students (without regard for department).

3. This problem deals with Fisher's iris data discussed in class.

   (a) Create a pairs plot of the Fisher iris data, using a different plot character for each variety of iris.

   (b) Suppose we want to build a classifier that identifies the type of iris: setosa, versicolor, and virginica. Reasoning from your plot, which single variable from the given four would be best choice for constructing such a classifier. There may be several reasonable choices, but be sure to justify your answer.

4. There are three types of dice in a box, A,B, and C. Type A are fair (all numbers equally likely). Type B gives probability 2/9 to the odd numbers and 1/9 to the even numbers. Type C gives probability 2/9 to the numbers 4,5,6, and 1/9 to the numbers 1,2,3. The types, A,B,C are represented with proportions $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ in the box.

(a) Simulate the experiment of choosing a die (according to stated model) and rolling the die 3 times giving results x_1, x_2, x_3. Submit code for doing the simulation 1000 times.

(b) Since the rolls are (conditionally) independent given the choice of die, we can compute the probability of the outcome as

$$\begin{aligned} P(x_1, x_2, x_3|A) &= P(x_1|A)P(x_2|A)P(x_3|A) \\ P(x_1, x_2, x_3|B) &= P(x_1|B)P(x_2|B)P(x_3|B) \\ P(x_1, x_2, x_3|C) &= P(x_1|C)P(x_2|C)P(x_3|C) \end{aligned}$$

Suppose the outcome of the experiment gives $x_1 = 2, x_2 = 4, x_3 = 5$. What are the probabilities of the three types of dice, given this outcome? In other words, what are $P(A|x_1, x_2, x_3)$, $P(B|x_1, x_2, x_3)$, $P(C|x_1, x_2, x_3)$?

(c) Using the result from the previous part, how would a Bayes' classifier classify this particular outcome of x_1, x_2, x_3?

(d) Simulate the experiment 1000 times and calculate the classification for each case using the Bayes' method being sure to "remember" the true class. What is the proportion of times your classifier gives the true result? (If you have implemented your strategy correctly, there is no classifier that will give a better result on average.)

5. Consider the "three_related_vars.csv" dataset on Canvas summarizing a computer experiment involving three binary variables, A, B, C. The dataset can be read into R with the command

```
X = read.csv("three_related_vars.csv")
```

producing a matrix X whose columns are the three variables. Don't worry that there are no commas in the .csv file.

(a) Show mosiac plots of each pair of variables, identifying which, if any, pairs of variables appear independent, justifying your answers.

(b) Similarly, consider conditioning on each of the variables, asking of the remaining two variables are *conditionally* independent. Are there any conditional independencies in the data. Justify your answers.

(c) Write an R program that could have been used to generate these data. Hint: All probabilities in the actual program used only a single digit after the decimal point.