B365 Homework 3

1. 20% of a population have trait, T. Of those that have trait T, 80% have trait S, while only 30% have S in the remaining population.

   (a) Are T and S indepedent? Explain your reasoning.

   (b) Suppose that a randomly-selected individual has trait S. What is the probability that the individual has trait T?

   (c) Using R, simulate 10,000 members of this population, assigning whether or not traits S and T occur according to the given model.

   (d) From your simulation above, estimate a 95% confidence interval for $P(T|S)$ using

   $$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Is this consistent with your earlier answer to the 2nd part of this problem?

2. Bayes' rule can be generalized to random variables X, Y where X takes values $x_1, \ldots, x_n$ and Y takes values $y_1, \ldots, y_m$. Specifically, if $p(x_i)$ gives the probability for $X = x_i$ and $p(y_j|x_i)$ gives the conditional probability for $Y = y_j$ *given* $X = x_i$, Bayes' rule says

   $$p(x_i|y_j) = \frac{p(x_i)p(y_j|x_i)}{p(x_1)p(y_j|x_1) + \ldots + p(x_n)p(y_j|x_n)}$$

   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 40% of the members of party A, 60% of the members of party B, and 80% of the members of party C. A randomly selected person is found to *not* favor the ballot measure.

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

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

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

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

   (e) Suppose we attend a wine festival in the state and meet a person at this event. 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.

3. 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. Are the department and the admit status independent?

   (b) Construct and print the two-way table of department and gender, as well as a mosiac plot of the table. Are the department and the gender of an applicant independent?

   (c) Construct a two-way table of gender and admit status for the applicants to deparment 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 Admitted students (without regard for department).

4. 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 this plot, which two of the four variables would be best choice for constructing such a classifier. There may be several reasonable choices, but be sure to justify your answer.

5. 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 even numbers and 1/9 to the odd numbers. Type C gives probability 2/9 to the numbers 1,2,3, and 1/9 to the numbers 4,5,6. The types, A,B,C are represented with proportions $\frac{1}{2}, \frac{1}{4}, \frac{1}{4}$ in the box.

   (a) Simulate this experiment of choosing a die (according to stated model) and rolling the die 3 times giving results x_1, x_2, x_3.

   (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?

   (c) Using a Bayes' classifier how would you classify this outcome?

   (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 best result? (If you have implemented your strategy correctly, there is no classifier that will give a better result on average.)