Homework 4

Not collected; Not graded; For practice only

1. Frequentist and Bayesian statisticians sometimes agree but sometimes disagree on what constitutes a well formed statistical situation. For each statement below explain why a frequentist and a Bayesian would agree or disagree that it poses a statistical question.

   a) The afternoon flight on Thursdays from O'Hare to Champaign often arrives early.
   Both would find this acceptable as the probability of early arrival can be viewed as a long-run average

   b) The probability of rain tomorrow is 20%.
   Bayesian only. This is a subjective probability. There is no underlying population of outcomes.

   c) After my wife won $500 on a $2 lottery ticket, I realized I should be playing the lottery more often myself.
   Neither. This is not a justified statistical conclusion.

2. Explain in your own words the difference between the weak and strong law of large numbers.
   For my words, see the lecture slides

3. In preparation for the upcoming CS440 midterm, you put in *h* hours of diligent studying. You then evaluate your knowledge on *r* independent practice problems. By this you estimate a probability *p* of knowing any particular correct answer. On March 1 you find that the CS440 midterm is comprised of *n* multiple choice questions each with *m* alternative answers (NB: this is convenient for the problem but will not hold in reality). During the test, when you do not know, you make a uniform random guess. Let *k* be the number of questions you answer correctly (either by knowing or guessing).

   a) What is the best estimate of *k* in terms of the other quantities?
   Assume you put down the right answer when you know it: np correct via knowing
   You probably won't know n(1-p) questions. For these you have a 1/m chance of being correct: n(1-p)/m correct via guessing
   So we would estimate k as np + n(1-p)/m

   b) What is the best estimate of *p* in terms of the other quantities?
   Solving the above for p instead of k: (mk-n) / (mn – n)

   c) You answer the first question correctly. What is the probability that you knew the answer to the first question?
   Use Bayes rule: Pr(knew | correct) = Pr(correct | knew) Pr(knew) / Pr(correct)
   Pr(correct | knew) = 1; Pr(knew) = p; Pr(correct) = k/n; then a little high school algebra to combine and eliminate k

   d) What distribution best describes *p*? Explain its arguments.
   It's our friend the Beta distribution. Its two shape arguments are successes+1 and failures+1.
   We estimated p over the r practice problems so successes = correct on practice = rp; failures = r(1-p)
   The desired distribution would be Beta(rp+1, r-rp+1)

4. Only six different kinds of fruit are found on an island. A random sample yields the numbers shown in the table for each kind of fruit.

RANDOM SAMPLE OF FRUIT FOUND ON THE ISLAND

SIZE	COLOR	TEXTURE	EDIBLE	Number
large	red	rough	FALSE	4
medium	green	smooth	FALSE	10
small	yellow	smooth	TRUE	4
medium	red	smooth	FALSE	8
medium	yellow	rough	FALSE	6
large	red	smooth	TRUE	8

a) Given the above information, how many numbers are there in the joint probability table for this problem?
The interesting facet of this problem is that although we can represent 36 different fruits (3 x 3 x 2 x 2) the support for the distribution is just 6 fruits. So I would accept either 6 or 36 as answers.

b) Given the above information, how many degrees of freedom are there in the joint probability table for this problem?
Given the support is just these 6 fruits, there are 5 degrees of freedom (at most 5 numerical choices before all the others are forced).

c) From this sample estimate the following:

i. Pr(SIZE) = 12/40, 24/40, 4/40 = 0.3, 0.6, 0.1 for (large, medium, small); we need the full prior (or unconditional) distribution for SIZE

ii. Pr(¬ yellow) = 30/40 = 0.75

iii. Pr(red | large) = 1

iv. Pr(EDIBLE | red) = 8/20 = 0.4

v. Pr(EDIBLE | ¬ red, ¬ small) = (1, 0) for EDIBLE = (False, True)

vi. Suppose we take a much larger sample of fruit. How would you expect each of these to change?
Parts (i) (ii) (iv) would likely change but parts (iii) (v) would not

d) & e) Suppose there are all different kinds of fruit on the island but no samples were found for most of them yielding the same table above. Repeat parts (a) & (b) with this information.
Without the constraint we need to allow for the possibility that the support is all 36 alternatives so
d) 36 numbers; e) 35 degrees of freedom

5. Explain the difference between *independence* and *conditional independence*. Which is stronger? Which is more useful in AI?
Again see lecture notes for the definition and verbal descriptions. Statistical independence is stronger. Conditional independence is more useful in AI.