 BAYESIAN DATA ANALYSIS MIDTERM 22 Oct 2020

NAME______________

This is a Take Home Exam
It is due to be emailed to knuthclass@gmail.com by Tuesday Oct 27, 2020 at 11:59 pm.

Do your own work.
You are not allowed to discuss the exam or problems herein with others.

Show all of your work!
Written solutions and all code are to be submitted. Just as with the homework, all figures are to be embedded in either a MSWord or a pdf document and described.

Copying and cheating will not be tolerated and will result in failure of the exam.
The exam (130 points total) consists of five problems, with sub-parts.

1. (15 points)
   Write Bayes' Theorem.
   Name each term in the equation and write a description (at least one sentence each) describing the meaning and/or use of each term. (15 points)

2. (20 points)
   You are playing a role-playing adventure game in which a player determines what happens in the game by rolling a die. You cannot see the die being rolled and have to trust the numbers that the person claims to roll.

   The person is supposed to be rolling a fair 20-sided die, but after rolling a 3 and a 9, you begin to suspect that the person might be rolling a 12-sided die by mistake. If there is an equal chance that the person is rolling either die, what is the probability that they are rolling the 12-sided die by mistake?

3. (35 points)

In World War II, when the Allies (mainly Britain and America) were planning an invasion of France to take Europe back from the Germans, they needed to know how many German tanks they would be up against.

From 20 captured German tanks, the Allies found this set of serial numbers on the engine blocks:

038839	019401	012031	020106	004802	006570	046893	047594	028633
002976	011687	017580	040877	000767	002142	008412	032312	036423
032243	022446							

In summary, the serial numbers range from 000767 to 047594.

Assuming that the Germans assigned sequential numbers to the engine blocks, and that the twenty captured tanks represent a uniformly-distributed sample from the set of tanks that the German's possess, how many tanks do the Germans have?

We will approach this problem in steps.

The aim is to estimate the number N of tanks possessed by the German military.

A. Assign a prior probability, $P(N \mid I)$, for the number of tanks.
(You can safely assume that they cannot possibly have more than a million tanks.)

Using Bayes Theorem, we can write the posterior probability as

$$P(N \mid d_1, d_2, d_3, \cdots, d_{20}, I) = \frac{1}{Z}\; P(\, N \mid I\,)\; P(d_1, d_2, d_3, \cdots, d_{20} \mid N, I)$$

B. The posterior probability depends on the joint likelihood $P(d_1, d_2, d_3, \cdots, d_{20} \mid N,\; I)$. Assuming that the serial numbers are independent, use the product rule to write the joint likelihood in terms of the individual likelihoods $P(\, d_1 \mid N, I)$, $P(\, d_2 \mid N, I)$, etc.

C. Given that there are N tanks (where $N \gg 20$) and the assumption that the captured tanks represent a uniformly-distributed sample, assign a likelihood for $P(\, d_i \mid N, I)$ by considering the two cases below (and filling in the blanks):

$$P(\, d_i \mid N, I) = \begin{cases} \_\_ & \text{for } N \;\geq\; d_i \\ \_\_ & \text{for } N \;<\; d_i. \end{cases}$$

D. Write a MATLAB program to evaluate the posterior probability for values of N ranging from 1 to 100,000. You can store the posterior probability values in a 1 x 100000 array called p.
To start, ignore the normalization factor (evidence) Z in the posterior probability. This means that the probabilities will not sum to one, as they should. You can handle this numerically by summing the probabilities later and then renormalizing them:

```
sumprobs = sum(p);   % sum the probabilities

posteriors = p / sumprobs;    % renormalize the probabilities
```

You can now check that the probabilities are now normalized by evaluating `sum(posteriors)` and verifying that the sum is 1.

Produce a plot of the normalized posterior probabilities and show that there is a most probable solution to the problem.

E. Find the most probable value of N.

F. Use MATLAB code to compute the average value, or expected value of N.

It may help to create an array `n = 1:100000`; and use this with the array of normalized posterior probabilities, `posteriors`, to compute this expected value.

G. How many German tanks should the Allies expect? <u>Explain</u> why you would recommend this solution.

4. (35 points)
Finding the minimum of the Rosenbrock's Banana Function

$$f(x, y) = 100\,(y - x^2)^2 + \,(1 - x)^2$$

is notoriously difficult using numerical optimization due to the slow convergence rates. (https://www.mathworks.com/help/optim/ug/banana-function-minimization.html)

We will work with this function as if it were a probability density function (I have multiplied $f(x, y)$ by -1 so that $p(x, y \mid I)$ has a maximum).

$$p(x, y \mid I) = \frac{-100\,(y - x^2)^2 - \,(1 - x)^2}{Z}$$

defined in the region $-c \le x \le c$ and $-d \le y \le d$ for some positive values of c and d.

A. Find the normalization constant Z.

B. Using calculus, find the (x, y) coordinates of the maximum of $p(x, y \mid I)$.

C. Find the covariance matrix about the peak of $p(x, y \mid I)$.

D. Find an expression for $p(x \mid I)$.

E. Find the peak of $p(x \mid I)$ and its associated uncertainty.

F. Is the peak of $p(x \mid I)$ at the same x-value as the peak of $p(x, y \mid I)$?
Discuss this.

G. Plot $p(x, y \mid I)$ in one figure and $p(x \mid I)$ in another.
You can concentrate on the range where $c = d = 5$.
You may want to look at the code included in the class notes from Oct 15.

5. (25 points)
   An archeologist is studying drug abuse in the times of the late Roman Empire (circa 250–450 A.D.). This is accomplished by testing pottery for traces of opium. Samples from one such vessel were sent to five different labs to determine the age of the artifact.

   Laboratory A returned a report stating that they dated the artifact at 327.1 ± 1.2 A.D.
   Laboratory B returned a report stating that they dated the artifact at 332.5 ± 3.2 A.D.
   Laboratory C returned a report stating that they dated the artifact at 321.2 ± 7.5 A.D.
   Laboratory D returned a report stating that they dated the artifact at 318.0 ± 2.2 A.D.
   Laboratory E returned a report stating that they dated the artifact at 325.3 ± 3.5 A.D.

   Assuming that the listed uncertainties represent a 1-σ estimate from a Gaussian distribution:

   A. Compute the most probable estimate for the age of the pottery and its uncertainty. You might consult with the Measuring Lengths lecture notes, but you will have to complete the notes by deriving the formula for estimating the uncertainty of the most probable estimate.

   B. Plot the data with error bars along with lines illustrating the most probable age along with the uncertainties.