Problem Set 1

1. Consider the following two-player game. Each player rolls a six-sided die (and the outcomes of the rolls are independent and uniformly distributed). They cannot see their own outcome but they can see the outcome of the other player. They must submit a guess of their own roll and they do so in such a way that absolutely no information is passed between players. If they are both correct, they win. If either is wrong, they lose. What is the maximum probability of victory?

2. This question is about using the `rand` or `random` function to simulation random variables. Let `rand` denote a function that generates a random number uniformly distributed on $[0, 1]$.

   a. Use `rand` to select one of $8,094$ students uniformly at random.

   b. Write a code to select k students uniformly at random without replacement.

   c. Write a code to generate a bivariate random variable according to the distribution below

```
0.0267    0.0697    0.0775    0.0313    0.0101
0.0283    0.0761    0.0739    0.0566    0.0362
0.0109    0.0337    0.0552    0.0780    0.0740
0.0014    0.0093    0.0309    0.0750    0.0698
0.0001    0.0028    0.0124    0.0318    0.0283
```

3. Consider the movie preference prediction example from class. The two arrays below denote student ratings for *Star Wars* (SW), 1-5, left to right. and *Sleepless in Seattle* (SS), 1-5 bottom to top. The array on the left are the counts of students that did not like *Guardians of the Galaxy* (GG), and the array in the right is those that did.

```
216   560   606   216    42          0     4    21    37    40
228   604   527   247    35          1    12    71   211   258
 83   236   231    78    21          5    37   216   553   578
  7    39    35    18     3          4    36   215   589   562
  1     2     3     1     0          0    21    97   256   231
```

didn't like GG　　　　　　liked GG

   (a) What is the probability that a randomly selected student (RSS) will like GG and give SW a 5-star rating?

   (b) What is the probability that an RSS who likes GG, will give SW 5-stars and SS 2-stars?

   (c) What is the probability that an RSS will give SW *at least* 3 stars and SW *at most* 2 stars? item Suppose a new student gave SW a 3-star rating and SS a 2-star rating. Will she like GG?

(d) Suppose a new student gave SW a 3-star rating. Predict her rating for SS.

(e) Suppose you know that a new student gave SS a 2 star rating and also that she liked GG, what is your estimate of her (unknown) rating of SW?

(f) Suppose all you know is that a new student didn't like GG. Make predictions of her ratings for SW and SS?

4. **a.** Imagine that you have been hired by a biotechnology start-up company to help them identify whether certain genes may be associated with a form of cancer. They are currently interested in a particular gene, because they have developed a very cost-effective screening procedure. The procedure generates a number for each person that is screened. A study conducted on a large group of people showed that the numbers produced by the procedure can be modeled as outcomes of a $\mathcal{N}(0,1)$ random variable for healthy people and $\mathcal{N}(1,1)$ for cancer patients; $\mathcal{N}(\mu,\sigma^2)$ denotes the Gaussian distribution with mean μ and variance σ^2. Pose the screening process as a signal detection problem and explain your answer mathematically. Do you think the company should market the procedure as a good test for this cancer? Why or why not?

   b. After further analysis, it turns out that if you repeat the screening procedure on the same person then you get a different number each time. The values that are produced by repeating the screening procedure multiple times can also be modeled as independent realizations of a $\mathcal{N}(0,1)$ or $\mathcal{N}(1,1)$ random variable, for healthy people and cancer patients, respectively. Can you construct a more robust test for cancer based on this observation?

   c. Suppose that you've discovered another procedure that is more accurate, but also more expensive. Specifically, the outcomes of the new procedure are realizations of a $\mathcal{N}(0,0.25)$ or $\mathcal{N}(1,0.25)$ random variable, for healthy people and cancer patients, respectively. The cost of the new procedure is 5 times greater than the cost of each repeat of procedure above. Which procedure would you recommend?

5. Suppose that scientists are studying how the brain performs a certain information-processing tasks. Three regions of the brain are involved, denoted A, B and C. There is prior evidence that there are direct neural connections between regions A and B and regions B and C. However, it is uncertain whether regions A and C are directly connected. The scientists design an experiment to test this. The activity in human subject's brains is measured while they perform the information-processing tasks. The activity level in each region is a binary-valued variable, indicating whether the region is significantly active. Let x_A, x_B, and x_C denote the activity level in each region, which we will model as sequences of random variables. If there is no direct connection between regions A and C, then we conjecture that x_A and x_C will be *conditionally independent* given x_B.

   Many measurements of these variables, for repeated trials of the task and different human subjects, are recorded. The dataset is $\{(x_A^{(i)}, x_B^{(i)}, x_C^{(i)})\}_{i=1}^n$, where n is the total number of measurements. We can model each triple $(x_A^{(i)}, x_B^{(i)}, x_C^{(i)})$ as independently and identically distributed (i.e., each triple is an independent realization from the same multivariate distribution, but $x_A^{(i)}$, $x_B^{(i)}$, and $x_C^{(i)}$) may be correlated). How would you use the data to check for whether x_A and x_C are conditionally independent given x_B?

This problem can be simulated in Matlab. I have generated two datasets, `brain_data1.mat` and `brain_data2.mat`, which simulate two different information-processing tasks. Use these data to determine whether x_A and x_C are conditionally independent given x_B. Your answer may be different in the two cases. Implement your procedure in Matlab or Python or R (or some other language) and discuss your conclusions.