COS302/SML305-Princeton University-Spring 2022
Assignment #6
Due: 11:59pm March 21, 2022

Upload at:

Remember to append your Colab PDF as explained in the first homework, with all outputs visible. When you print to PDF it may be helpful to scale at 95% or so to get everything on the page.

Problem 1 (15pts)

Consider a sample space Ω = {red, green, blue, orange, yellow}.

(A) What is the smallest possible event space that contains the set {blue, red}?

(B) What is the smallest possible event space that contains both {blue} and {red, green}?

(C) What is the smallest possible event space that contains both {blue} and {red, blue}?

Problem 2 (20pts)

Consider the following bivariate distribution $p(x, y)$ of two discrete random variables X and Y.

		x_1	x_2	x_3	x_4	x_5
	y_1	0.01	0.02	0.03	0.1	0.1
Y	y_2	0.05	0.1	0.05	0.07	0.2
	y_3	0.1	0.05	0.03	0.05	0.04

X

(A) What is the marginal distribution $p(x)$?

(B) What is the probability $p(X \neq x_1, Y \neq y_2)$?

(C) What is the conditional distribution $p(x \mid Y = y_1)$?

(D) What is the conditional distribution $p(y \mid X = x_3)$?

(E) What is the conditional distribution $p(x \mid Y \neq y_1)$?

Problem 3 (20pts)

In the early stage of the COVID-19 outbreak, COVID tests didn't always give the correct test results. The probability of being tested positive given that the citizen has COVID is 80%. The probability of being tested positive given that the citizen does not have COVID is 8%. We also know that the probability of having COVID for any given citizen is 2%. If a randomly-chosen citizen tests positive, what is that citizen's probability of actually having COVID?

Problem 4 (50pts)

In this problem you will do some mathematical calculations and also use Colab. Be sure to append your PDF and insert your link as usual. The following problems will be denoted as either **Calculation** or **Code** to indicate whether the problem requires a mathematical calculation (no code) or Colab (no mathematical calculation). .

(A) Import `numpy.random` (usually aliased to `npr`) and set the random seed.

(B) **Calculation**: Imagine drawing 1,000 independent Bernoulli variates $X_i \in \{0, 1\}$ with the probability $p(X_i = 1) = 0.35$, and computing their sum $Y = \sum_{i=1}^{1000} X_i$. What are the mean and variance of Y?

(C) **Code**: Generate 10,000 independent random variables Y as described above. That is, generate 10,000 sums of 1,000 independent Bernoulli variables. This is not as hard as it sounds, we list the steps required to do so below:

   (i) Use the `rand()` function only; do not use any functions from `scipy.stats`.

   (ii) Use `rand()` to generate a $10,000 \times 1,000$ matrix of independent uniform random variates in the interval $[0, 1]$, then threshold them appropriately to get 0 or 1.

   (iii) Finally, sum over the appropriate dimension to get 10,000 samples of Y above.

(D) **Code**: Use Matplotlib to make a histogram of these samples. Use at least 100 bins so you can see the structure in the distribution. Do you recognize the shape?

(E) **Code**: Compute the empirical mean and variance of the 10,000 samples you have drawn. Compare these results to your calculation from (B).

(F) **Calculation**: Now imagine drawing 1,000 independent (continuous) variates uniformly in the interval $[-1, 1]$. What are the mean and variance of their sum?

(G) **Code**: Generate 10,000 such sums using a variation of the procedure you performed for (C). As before, you'll only use `rand()` but you should scale and shift rather than threshold.

(H) **Code**: Create a histogram of these sums, as in (D). Describe the shape.

(I) **Code**: Compute the empirical mean and variance of the 10,000 samples and compare them to your computations in (F).

My notebook URL: https://colab.research.google.com/XXXXXXXXXXXXXXXXXXXXXXXXX