Assignment #7

Due: 12:00 noon November 9, 2020

Upload at: https://www.gradescope.com/courses/135869/assignments/795472/

Assignments in COS 302 should be done individually. See the course syllabus for the collaboration policy.

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 (10pts)

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

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

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

(B) What is the marginal distribution $p(y)$?

(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 2 (8pts)

In 2014-2016, West Africa experienced a massive outbreak of Ebola. We'll concentrate on Sierra Leone and imagine a mandatory screening of every citizen. The probability of being tested positive given that the citizen has Ebola is 84%. The probability of being tested positive given that the citizen does not have Ebola is 11%. We also know that the probability of contracting Ebola for any given citizen is 0.4%. If a randomly-chosen citizen tests positive, what is that citizen's probability of actually having Ebola?

Problem 3 (35pts)

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.

(A) 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 $Y = \sum_{i=1}^{1000} X_i$. Theoretically, what should the mean and variance of Y be?

(B) Now let's simulate this empirically. Start by importing `numpy.random` (usually aliased to `npr`) and set the random seed.

(C) 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. Use the `rand()` function only; do not use any functions from `scipy.stats`. 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. Finally, sum over the appropriate dimension to get 10,000 samples of Y above.

(D) Use Matplotlib to make a histogram of these samples. Use at least 100 bins so you can see the structure in the distribution. Describe the shape — do you recognize it?

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

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

(G) 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) Create a histogram of these sums, as in (D). Describe the shape.

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

Problem 4 (20pts)

Consider the following cumulative distribution function for a random variable X that takes values in $\mathbb{R}$:

$$F(x) = P(X \leq x) = \frac{1}{1 + e^{-x}}$$

(A) What is the probability density function for this random variable?

(B) Find the inverse distribution (quantile) function $F^{-1}(u)$ that maps from $(0, 1)$ to $\mathbb{R}$.

(C) In a Colab notebook, implement inversion sampling and draw 1000 samples from this distribution. Make a histogram of your results.

Problem 5 (25pts)

You're playing a game at a carnival in which there are three cups face down, and if you choose the one with a ball under it, you win a prize. This is sometimes called a *shell game*. At this carnival, there is a twist to the game: after you pick a cup, but before you're shown what is beneath it, the game operator reveals to you that one of the other cups (one of the two you did not choose) is empty. The operator now gives you the opportunity to switch your selection to the other unrevealed cup.

To clarify with an example: imagine there are cups A, B, and C. It is equally probable that the ball is beneath any of the three. You choose B. Before you see what is under B, the operator lifts A and shows you there is nothing under it. You are now presented with the option to keep your selection of B, or switch to the still-unrevealed cup C.

(A) Is it better, worse, or the same to switch to the other cup? Explain your reasoning in terms of probabilities. Assume that if you originally picked the correct cup, the operator will pick one of the other two with equal probability.

(B) In a Colab notebook, simulate this game. Run 1,000 games with the *stay* strategy and 1,000 games with the *switch* strategy. Report the win rate of each strategy and explain which one empirically seems better.

Problem 6 (2pts)
Approximately how many hours did this assignment take you to complete?

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

Changelog

- 26 October 2020 – Updated for fall, 2020
- 30 March 2020 – Initial version