CS2810 Mathematics of Data Models	Gold
Spring 2019	March 11, 2019

Homework 07

Due: March 27, 9PM
Point total: 60

Instructions:

- Submit your PDF and/or .py file to Blackboard by the due date and time. Please do not zip your files together, as this interferes with Blackboard's preview functionality. Always show all your work, and for full credit, you must use the method that the problem instructs you to use (unless none is mentioned). Handwritten or typeset solutions are both acceptable, but unreadable submissions will be penalized. You may discuss problems with other students, but you may not write up solutions together, copy solutions from a common whiteboard, or otherwise share your written work or code. Do not use code or language that is copied from the Internet or other students; attribute the ideas *and* rephrase in your own words.
- *Note for this homework:* You should at least set up the initial equation for a distribution correctly for full credit, but you're welcome to use an online solver to get the value after that. For a Gaussian, you don't need to give the PDF, but do give the mean and standard deviation.

Problem 1 (12 points, 3 each)

i. Find the expectation and variance of an eight-sided die, numbered (1,2,3,4,5,6,7,8).

Solution: The expectation is just the average of the rolls, $\sum_i (1/8)i = 4.5$. The variance is $(1/8)(1-4.5)^2 + (1/8)(2-4.5)^2 + (1/8)(3-4.5)^2 + \ldots + (1/8)(8-4.5)^2 = 42/8 = 5.25$.

ii. Using the rule Var(X+Y) = Var(X) + Var(Y), find the variance of the sum of two eight-sided dice.

Solution: This is just $5.25 \times 2 = 10.5$.

iii. Compute the variance of the random variable $Z = 2X$, where X is the roll of a single eight-sided die. (So Z can take on the values 2, 4, 6, …, 16. The result isn't the same as the previous answer.)

Solution: The expectation is now $4.5 \times 2 = 9$. The variance is then $(1/8)(2-9)^2 + (1/8)(4-9)^2 + \ldots + (1/8)(16-9)^2 = ((1/8)7^2 + (1/8)5^2 + (1/8)3^2 + (1/8)1^2) \times 2$ (by symmetry) $= 168/8 = 21$. (This is also what we get using the rule $Var(aX) = a^2 Var(X)$.)

iv. Plot the two probability distributions on the same graph, but with different marks: on the one hand, the likelihood of the sum of two eight-sided dice, and on the other, the likelihood of each value of Z from the previous problem. How can we visually identify the distribution with the smaller variance?

Solution: The plot of the sum of two dice is a pyramid shape familiar from the discussion of a six-sided die, while the plot of the distribution of $Z = 2X$ is flat, even though it covers roughly the same range. The pyramid shape indicates a smaller variance, since the values are more concentrated toward the center.

Problem 2 (15 points [2, 3, 2, 3, 2, 3])

i. Suppose 1/10 the general population is left-handed. In a class of 30, what is the probability that exactly 3 turn out to be left-handed? (Use the binomial distribution.)

<u>*Solution:*</u> Using the binomial distribution, this is $\binom{30}{3}(1/10)^3(9/10)^7 = 0.236$.

ii. What is the probability that 2 *or fewer* students are left-handed? (Sum over the relevant values. Recall that $\binom{N}{0} = 1$.)

<u>*Solution:*</u> We can get this by summing over the three binomial distribution values, for 0, 1, and 2.
$\binom{30}{2}(1/10)^2(9/10)^8 = 0.228$
$\binom{30}{1}(1/10)^1(9/10)^9 = 0.141$
$\binom{30}{0}(1/10)^0(9/10)^{10} = 0.042$

The sum, the probability that it's any one of those values, is 0.411.

iii. What is the probability that 4 *or more* students are left-handed? (Use your previous results.)

<u>*Solution:*</u> This is just 1 - 0.411 - 0.236 = 0.353.

iv. Suppose we approximate the situation with a Poisson distribution instead, with $\lambda = 3$ as the expected number of left-handed students. Calculate the probability of exactly 3 left-handed students using this approximation.

<u>*Solution:*</u> The Poisson distribution would predict Pr (3 left-handers) $= e^{-3}\frac{3^3}{3!} = 0.224$.

v. Is a Gaussian a good approximation for the binomial here? Why or why not?

<u>*Solution:*</u> No, because the probability is small, making the distribution lopsided, and we're also pretty close to 0. So unlike a Gaussian, this distribution is not symmetric.

vi. If half the population were left-handed, would the variance be larger or smaller? And, would a Gaussian be a better or worse approximation?

<u>*Solution:*</u> The variance would be larger, which we can see using the binomial variance formula $np(1-p)$ $(1/2 \times 1/2 > 1/10 \times 9/10)$. The Gaussian would be a better approximation, because the distribution would no longer be skewed.

Problem 3 (8 points [4, 2, 2])

i. Suppose we model the number of times our server goes down per year as a Poisson distribution, with an expectation of going down twice a year. Calculate the probability that it goes down twice *or less* in a year. (Note that $0! = 1$.)

<u>*Solution:*</u> This is the sum of the probabilities of 0, 1, or 2 outages, with $\lambda = 2$.
$e^{-2}2^0/0! = 0.135$
$e^{-2}2^1/1! = 0.270$
$e^{-2}2^2/2! = 0.270$
The chance of going down 2 or fewer times is $0.135 + 0.270 + 0.270 = 0.675$.

ii. Use the previous probability to calculate the likelihood that it goes down more than twice in a year.

<u>Solution:</u> 1 - 0.675 = 0.325

iii. You hear that the software company down the road models their server's number of outages with a Gaussian instead of a Poisson distribution. If their model is reasonable, what does that say about the reliability of their server?

<u>Solution:</u> The reliability must be worse, since the Poisson only starts to look like a Gaussian when the expected value λ is large.

Problem 4 (13 points [4, 4, 5])

For each problem, identify the best distribution to use from among the Exact Binomial, Poisson, and Binomial Approximated as a Gaussian, then solve the problem. You should only use each distribution once, so be sure to use the *best* one for the problem setup. (If you approximate a binomial as a Gaussian, the standard deviation is $\sqrt{np(1-p)}$.)

i. On average, 1.75 hurricanes hit the East coast of the US every year. Calculate the likelihood that we see none in one year.

<u>Solution:</u> Poisson with $\lambda = 1.75$: $e^{-1.75}\frac{1.75^0}{0!} = 0.173$.

ii. There are about 15 stoplights on my way to work, each with an independent chance of 1/3 of being red when I reach it. Calculate the probability of *no more than* 2 reds on the commute.

<u>Solution:</u> Exact binomial, $n = 15, p = 1/3$.
$\binom{15}{2}(1/3)^2(2/3)^{13} = 0.060$
$\binom{15}{1}(1/3)^1(2/3)^{14} = 0.017$
$\binom{15}{0}(1/3)^0(2/3)^{15} = 0.002$
Sum is 0.079.

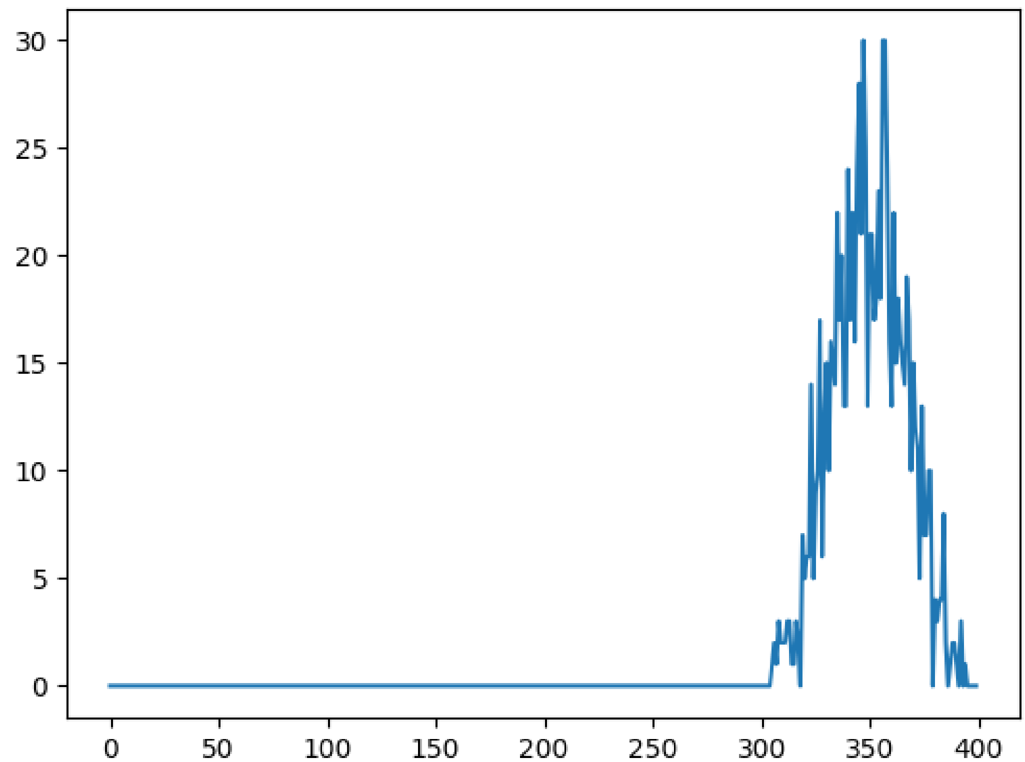
iii. A university is 50 percent female in its undergraduate population, but a particular hundred student class has only forty women; it's not immediately clear whether this is due to chance or some other effect. Calculate the likelihood of having 40 or fewer women in the class due to chance alone.

<u>Solution:</u> Binomial distribution approximated as Gaussian. $\mu = 50, \sigma = \sqrt{100(1/2)(1/2)} = 5$. 40 just happens to be 2 standard deviations away, and we know both tails beyond 2 standard deviations contain 5 percent of the total probability. So a deviation at least this extreme in this direction has roughly a 2.5% chance of happening (2.3% if we actually calculate the CDF); we'd consider this significant deviation from chance in an experiment.

Problem 5 (12 points, 6 each)

For each part, run the requested simulation in Python and plot your results using matplotlib's plot() function. Then, identify the distribution that best fits the simulation results, including its parameters, which you should be able to figure out *exactly* from the problem description. (Hint: If the distribution is Gaussian, the standard deviation is the square root of the variance, which you could calculate through other means.) Turn in your simulation code along with your PDF.

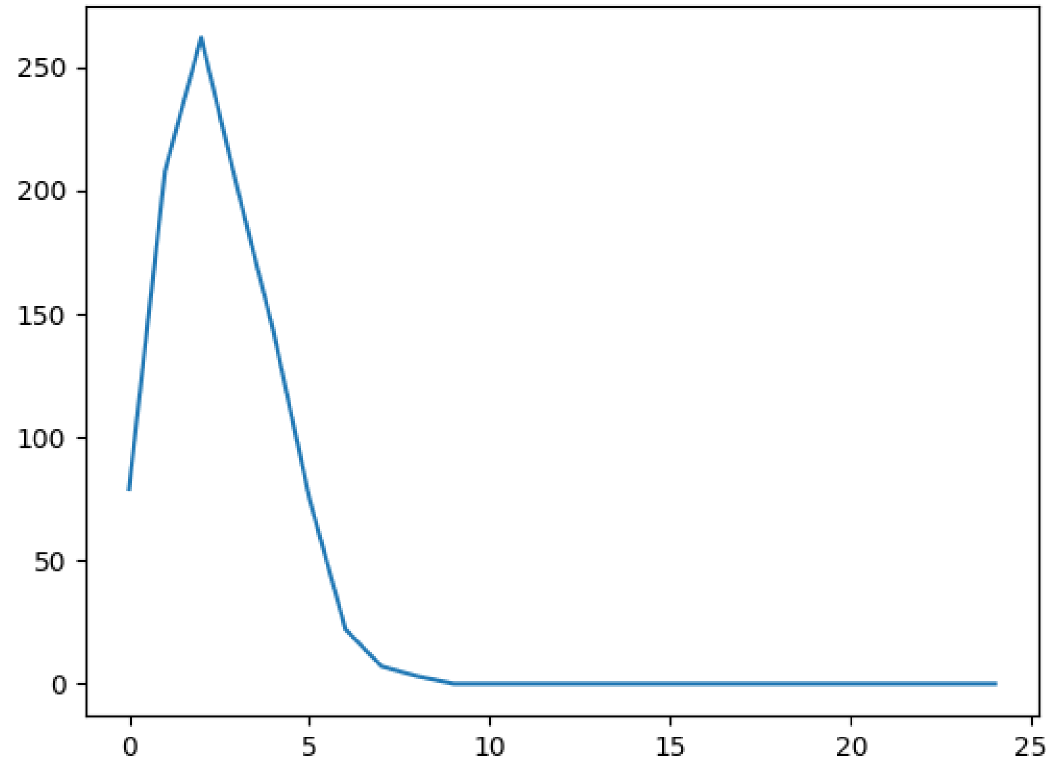
i. Sum 100 six-sided die rolls. Repeat 1000 times; plot result vs count of result.



Solution:

This is a Gaussian. If X is the random variable, the expectation is 100E[X] = 350. The variance is 100Var(X) = 292 or so, making the standard deviation 17. So we'll expect 95% of the simulation results to fall between 316 and 384 or so.

ii. Drop 1000 random points at random in the 2D square with x and y ranges both equal to [0,1), using continuous values from `random.random()`. Count how many landed in the square with x and y ranging from 0 to 0.05. Repeat 1000 times; plot the number of times each number of points landed in the smaller square.



Solution: [

This is a Poisson distribution, with expected number of arrivals $\lambda = (1/20)(1/20)1000 = 2.5$.