STAT370 Problem set #1 (Probability review)

Individual writeups due Thursday, January 25th at the start of class (please be sure to keep a copy of your work!)

Coaching session I in class on Thursday, January 25th and Friday January 26th

draft writeup (in Markdown, submitted by the first member of the group including R code and summary) due by midnight on Friday (automatic extension until noon on Saturday)

final group writeup (in Markdown, submitted by the first member of the group including R code and summary) due by 5pm on Monday, January 29th

Presentations (5 minutes per problem) on Tuesday, January 30th (I will give you a heads up on Monday regarding which problem or problems your group will be presenting)

For (almost all) problems, both an empirical (simulation study) solution plus an analytic (closed-form) solution are needed. See the sample solutions posted on github (`covxy.Rmd` and `rice3-11.Rmd`) for examples.

Everyone is to complete exercises 1 and 2. Group 1 should also complete problems 3, 4, 5; Group 2 should also complete problems 5, 6, 7; Group 3 should also complete problems 7, 8, 9; Group 4 should also complete problems 9, 10, 11; Group 5 should also complete problems 3, 5, 9; Group 6 should also complete 4, 6, 10; Group 7 should also complete problems 7, 8, 11; Group 8 should also complete 6, 10, 11.

Problems

1. Let X and Y be identically distributed independent random variables such that the moment generating function of X + Y is M(t)=$0.09e^{-2t} + 0.24e^{-t} + 0.34 + 0.24e^{t} + 0.09e^{2t}$, for $-\infty < t < \infty$. Calculate $P(X <= 0)$.

2. Assume that there are nine parking spaces next to one another in a parking lot. Nine cars need to be parked by an attendant. Three of the cars are hybrids, three are minivans, and three SUV's. Assuming that the attendant parks the cars at random, what is the probability that the three hybrids are parked adjacent to one another? Simulate to check your answer.

3. A machine produces bearings with mean diameter 3.00 inches and standard deviation 0.01 inches. Bearings with diameters in excess of 3.02 inches or less than 2.98 inches will be deemed to fail to meet quality specifications. Assume that the underlying process is normally distributed. Approximately what fraction of this machine's production will fail to meet specifications? Simulate from a normal distribution to check your analytic answer. Then assume that the underlying process is Gamma distributed with the same mean and sd. Simulate from an appropriate Gamma distribution to check your analytic answer. How important is the normality assumption?

4. Let $k \geq 1$. Show that, for any set of n measurements, the fraction included in the interval $\bar{y} - ks$ to $\bar{y} + ks$ is at least $(1 - 1/k^2)$. This result is known as Tchebysheff's theorem. Simulate data from a set of one hundred observations from two different distributions and compare the actual results to the bound. (Hint: to show the result, use the formula for the population standard deviation and replace all deviations for which $|y_i - \bar{y}| \geq ks$ with ks and simplify.)

5. A personnel manager for a certain industry has records of the number of employees absent per day. The average number absent is 5.5, and the standard deviation is 2.5. Because there are many days with zero, one or two absent and only a few with more than 10 absent, the frequency

distribution is highly skewed. The manager wants to publish an interval in which at least 75% of these values lie. Use the result from the previous problem to find such an interval. Generate data from a distribution that matches these parameters and compare your result to the true value.

6. A day's production (batch) of 2000 compact discs is inspected as follows. If an initial sample of 15 shows at most 2 defective discs, the batch is accepted and subject to no further sampling. If, however, the first sample shows 3 or more defective discs, then a second sample of 20 discs is chosen and the batch is accepted if the total number of defective discs in the two samples is not more than 4. Find the probability that the batch is accepted if the actual batch contains 100 defective discs. Then find the probability that the batch is accepted if it contains 200 defective discs.

7. The time a manager takes to interview a job applicant has an exponential distribution with mean of half an hour, and these times are independent of each other. The applicants are scheduled at quarter-hour intervals, beginning at 8:00am and all of the applicants arrive exactly on time (this is an excellent thing to do, by the way). When the applicant with an 8:15am appointment arrives at the manager's office office, what is the probability that she will have to wait before seeing the manager? Be sure to check the solution with an empirical simulation.

When the applicant with an 8:30 appointment arrives at the manager's office office, find an estimate of the probability that she will have to wait before seeing the manager?

8. A merchant stocks a certain perishable item. She knows that on any given day she will have a demand for either two, three or four of these items with probabilities 0.1, 0.4, or 0.5, respectively. She buys the items for $1.00 each and sells them for $1.50 each. If any are left at the end of the day, they represent a total loss. How many items should the merchant stock in order to maximize her expected daily profit?

9. Shuffle an ordinary deck of 52 playing cards containing four aces. Then turn up cards from the top until the first ace appears. On the average, how many cards are required to produce the first ace?

10. Demonstrate how to generate samples from a Binomial random variable with n=5 and p=0.4 using only a series of Uniform(0, 1) random variables. Write a function which implements this approach for arbitrary `n` and `p`. (You may not use the `rbinom()` function!)

11. It is known that 5% of the members of a population have disease A, which can be discovered by a blood test. Suppose that N (a large number) of people are to be tested. This can be done in two ways: (1) Each person is tested separately; or (2) the blood samples of k people are pooled together and analyzed. (Assume that $N = nk$ with n an integer). If the test is negative, all of them are healthy (that is, just this one test is needed). If the test is positive, each of the k persons must be tested separately (that is, a total of k+1 tests are needed). For fixed k what is the expected number of tests needed in (2)? Find the k that will minimize the expected number of tests in (2). If k is selected using that minimizer, on the average how many tests (2) save in comparison with (1)? Be sure to check your answer using an empirical simulation.