Math 321 – Spring 2019 – R Activity 4

SUBMISSION INSTRUCTIONS:
- Submit this assignment via email. DUE DATE: Friday, 4/5
- Use the filename: `lastname_Ractivity4` (e.g. "stover_Ractivity4.pdf")
- Be sure to include your name in the actual document as well!
- Subject line of email: "R activity 4 submission"
- Include all R commands and output that is used to answer any questions or produce any requested graphics.

This R activity will have you simulate some of the random variable that we have studied. Here, simulating a random variable means creating simulated data that follows the probability rules of that random variable.

Binomial

Choose a number of trials, n (say between 20 and 50), and a probability of success, p (say between 0.2 and 0.8). Here is how you simulate a binomial random variable:

```
> rbinom(1, size=n, prob=p)
```

The output is the number of successes. Recall that we expect there should be $E(X) = np$ successes for $X \sim Bin(n, p)$. For example, 100 fair coin flips has $n = 100$ and $p = 0.5$, and we simulate this with `rbinom(1,size=100,prob=0.5)`. We expect that we get 50 successes in this example. Of course, you result is random and will therefor vary and generally will be different form what you expect. However, repeat the command `rbinom(1, size=`n`, prob=`p`)` a few times to convince your self that it is usually reasonably close to what you expect.

Now simulate N_{sim} repetitions of your binomial random variable (choose an N_{sim}):

```
> rbinom(N_sim, size=n, prob=p)
```

This will generate a list of N_{sim} numbers. Each number represents the number of successes in a particular n trials simulation. For example, if we do $n = 100$ fair coin flips (we expect 50 successes, $p = 0.5$), but repeat that $N_{sim} = 5$ times, we may get 44, 52, 56, 55, and 45 successes.

Question 1: Now set your n and p for your binomial random variable. Simulate this random variable N_{sim} times for $N_{sim} = 10, 100, 1000,$ and 10000. Plot a histogram of your data each time against the binomial pmf. Describe what happens as you increase N_{sim}

You can use the code provided below.

```
> n = n
  p = p
  Nsim = N_sim
  x = rbinom(Nsim,size=n,prob=p)
  h = hist(x,freq=FALSE,breaks=seq(-0.5,n+0.5,1))
  barplot(rbind(h$density,dbinom(0:n,size=n,prob=p)),beside=TRUE)
```

If we simulate 100 fair coin flips (for which we expect 50 successes), we know the actual outcome will vary. Since we know this is modeled by a binomial random variable, we can calculate the probability of the number of successes being between, say 27 and 42.

However, if I simulate my experiment of 100 fair coin flips over and over again, the proportion of those experiments that get between 27 and 42 successes will not reflect the theoretical probability perfectly. Let's do this! We'll use $N_{sim} = 500$ here. Again, it is important that you understand this means we are going to do 100 fair coin flips, 500 times.

```
> x = rbinom(500,size=100,prob=0.5)
```

The theoretical probability of a data point being between 27 and 42 for $X \sim Bin(n = 100, p = 0.5)$ is

$$P(27 \leq X \leq 42) = \sum_{i=27}^{42} \binom{100}{i} (0.5)^i (1 - 0.5)^{100-i}.$$

In R we can calculate this as

```
> sum(dbinom(27:42,size=100,prob=0.5))
```

Now let's compare it with the simulated probability from the data:

```
> sum(x>=27 & x<=42)/500
```

The above command checks each `x` value to see if it is 27 or larger AND 42 or smaller, counts the number of times that is TRUE, then divides by the total number of data points.

Now we will compare theoretical probabilities with those calculated from simulation instead of plotting a histogram and pmf.

Question 2: Choose an a and b (between 0 and n). Simulate your binomial RV with your chosen n and p for $N_{sim} = 100$ and $N_{sim} = 100000$. Run the code below for each N_{sim} value several times to see how your simulated probability and theoretical probability compare.

You can use the code provided below.

```
> n = n
  p = p
  Nsim = Nsim
  a = a
  b = b
  x = rbinom(Nsim,size=n,prob=p)
  simprob = sum( x>=a & x<=b)/Nsim
  theorprob = sum(dbinom(a:b,size=n,prob=p))
  cbind(simprob,theorprob)
```

Describe what happens to your simulated probability as you run the code above for $N_{sim} = 100$ a few times. Describe how the behavior is different when you use $N_{sim} = 100000$.

Poisson

Now we will simulate Poisson distributed data and compare it to the pmf.

Question 3: Let $\lambda = 5$ (Poisson rate parameter). Simulate $X \sim Pois(\lambda)$ for $N_{sim} = 10, 100, 1000, 10000$, and 100000, and plot the histogram of the simulated data next to the probability mass function. Describe what you observe. *(cont. on next page)*

You can use the code provided below.

```
> lam = 5
  Nsim = 10
  x = rpois(Nsim,lambda=lam)
  maxX = max(x)
  h = hist(x,freq=FALSE,breaks=seq(-0.5,maxX+0.5,1))
  barplot(rbind(h$density,dpois(0:maxX,lambda=lam)),beside=TRUE)
```

Exponential

Now we will simulate exponentially distributed data and compare it to the pdf. Note that this time it is a probability density function and not a mass function.

Question 4: Let $\lambda = 0.3$ (exponential rate parameter). Simulate $X \sim Exp(\lambda)$ for $N_{sim} = 100, 1000,$ and 10000 and plot the histogram of the simulated data next to the probability density function. Describe what you observe.

You can use the code provided below.

```
> lam = 0.3
  Nsim = 100
  x = rexp(Nsim,rate=lam)
  maxX = max(x)
  binwidth = 0.5
  h = hist(x,freq=FALSE,breaks=seq(0,maxX+binwidth,binwidth))
  lines(h$mids,dexp(h$mids,rate=lam))
```

Normal

Now we will simulate normally distributed data and compare it to the pdf.

Question 5: Let $\lambda = 0.3$ (exponential rate parameter). Simulate $X \sim N(\mu, \sigma^2)$ for $N_{sim} = 100, 1000,$ and 10000 and plot the histogram of the simulated data next to the probability density function. Describe what you observe.

You can use the code provided below.

```
> mu = 100
  sig = 5
  Nsim = 100
  x = rnorm(Nsim,mean=mu,sd=sig)
  minX = min(x)
  maxX = max(x)
  binwidth = 0.5
  h = hist(x,freq=FALSE,breaks=seq(minX-binwidth,maxX+binwidth,binwidth))
  lines(h$mids,dnorm(h$mids,mean=mu,sd=sig))
```