

# Statistics 341

## Fall 2008 - Assignment #4 Solutions

### Due Wednesday, October 15

This assignment is worth a total of 134 points.

1. (13 pts) An individual claims to have extrasensory perception (ESP). As a test, a fair coin is flipped ten times, and he is asked to predict in advance the outcome.

- (a) (1 pt) If the individual does not have ESP, what is the probability he will correctly guess the outcome on any given coin flip?

This probability is 0.5.

- (b) (2 pts) If the individual does not have ESP, what is the probability that he will correctly guess 7 or more out of 10 flips?

Let  $Y$  equal the number of correct guesses.  $Y$  has a binomial distribution with  $n = 10$  and  $p = 0.5$ . We need to find  $P(Y \geq 7)$ . Using R, the code is `sum(dbinom(7:10, 10, 0.5))` which is 0.171875.

- (c) (2 pts) If the individual does not have ESP, what is the probability that he will correctly guess 9 or more out of 10 flips?

In this problem, we need to find  $P(Y \geq 9)$ . Using R, the code is `sum(dbinom(9:10, 10, 0.5))` which is 0.01074219.

- (d) (4 pts) The test is conducted and the individual guesses correctly on 6 out of 10 coin flips. What do you think about the individual's claim that he has ESP? Explain your answer.

The probability the individual would correctly guess on 6 or more coin flips if he did not have ESP is  $P(Y \geq 6)$ . Using R, the code is `sum(dbinom(6:10, 10, 0.5))` which is 0.3769531.

So, if the person did not have ESP, he would do just as well or better than he did on the test around 38% of the time. This evidence does not support his claim of ESP.

Note: If you just calculated  $P(Y = 6)$ , that is fine.

- (e) (4 pts) A new test is devised where the person is asked to predict the outcome in advance for 100 coin flips. The test is conducted and the individual guesses correctly on 60 out of 100 coin flips. What do you think about the individual's claim that he has ESP? Explain your answer.

Now,  $Y$  has a binomial distribution with  $n = 100$  and  $p = 0.5$  if he does not have ESP. The probability the individual would correctly guess on 60 or more coin flips if he did not have ESP is  $P(Y \geq 60)$ . Using R, the code is `sum(dbinom(60:100, 100, 0.5))` which is 0.02844397.

So, if the person did not have ESP, he would do just as well or better than he did on the test around 2.8% of the time. Since this is a fairly low probability, there is evidence to support his claim of ESP.

Again: If you just calculated  $P(Y = 60)$ , that is fine.

2. (27 pts) Go to the course webpage and click on the link **Using R to Investigate the Binomial Distribution**. In this problem, you will use the information in this file to investigate

the simulated observations from different binomial distributions. In doing this, you will be studying the effect of changes in both  $n$  and  $p$  on the binomial distribution.

- (a) (6 pts) Use R to simulate 10,000 observations each from binomial distributions with parameters  $n = 2, 10, 20$  and  $200$  and  $p = 0.5$ . Use R to calculate the means and variances and to make histograms of these 4 sets of observations. Use your output to answer the following questions.
- (2 pts) What happens to the values of the mean as  $n$  increases? What values should these means be close to?  
The values of the means increase. These observed mean values should be close to the theoretical mean value for the distribution which is  $np = 0.5n$ .
  - (2 pts) What happens to the values of the variance as  $n$  increases? What values should these variances be close to?  
The values of the variances increase. These observed variances should be close to the theoretical variance value for the distribution which is  $np(1-p) = n(0.5)(0.5) = 0.25n$ .
  - (2 pts) Describe the changes you see in the 4 histograms of the observations as  $n$  increases.  
The histogram of the simulated binomial values becomes more peaked as  $n$  increases. All of the histograms are symmetric and the shape of the histograms looks more like a bell-curve as  $n$  increases.
- (b) (6 pts) Use R to simulate 10,000 observations each from binomial distributions with parameters  $n = 4, 20, 40$  and  $400$  and  $p = 0.25$ . Use R to calculate the means and variances and to make histograms of these 4 sets of observations. Use your output to answer the following questions.
- (2 pts) What happens to the values of the mean as  $n$  increases? What values should these means be close to?  
The values of the means increase. These observed mean values should be close to the theoretical mean value for the distribution which is  $np = 0.25n$ .
  - (2 pts) What happens to the values of the variance as  $n$  increases? What values should these variances be close to?  
The values of the variances increase. These observed variances should be close to the theoretical variance value for the distribution which is  $np(1-p) = n(0.25)(0.75) = 0.1875n$ .
  - (2 pts) Describe the changes you see in the 4 histograms of the observations as  $n$  increases.  
The histogram of the simulated binomial values becomes more peaked as  $n$  increases. The histograms for smaller  $n$  are skewed to the right, but as  $n$  increases, these histograms become more symmetric and the shape of the histograms looks more like a bell-curve for the two larger values of  $n$ .
- (c) (6 pts) Use R to simulate 10,000 observations each from binomial distributions with parameters  $n = 10, 50, 100$  and  $1000$  and  $p = 0.1$ . Use R to calculate the means and variances and to make histograms of these 4 sets of observations. Use your output to answer the following questions.
- (2 pts) What happens to the values of the mean as  $n$  increases? What values should these means be close to?

The values of the means increase. These observed mean values should be close to the theoretical mean value for the distribution which is  $np = 0.1n$ .

- ii. (2 pts) What happens to the values of the variance as  $n$  increases? What values should these variances be close to?

The values of the variances increase. These observed variances should be close to the theoretical variance value for the distribution which is  $np(1-p) = n(0.1)(0.9) = 0.09n$ .

- iii. (2 pts) Describe the changes you see in the 4 histograms of the observations as  $n$  increases.

The histogram of the simulated binomial values becomes more peaked as  $n$  increases. The histograms for smaller  $n$  are skewed to the right, but as  $n$  increases, these histograms become more symmetric and the shape of the histograms looks more like a bell-curve for the two larger values of  $n$ .

- (d) (6 pts) Use R to simulate 10,000 observations each from binomial distributions with parameters  $n = 100, 500, 1000$  and  $10000$  and  $p = 0.01$ . Use R to calculate the means and variances and to make histograms of these 4 sets of observations. Use your output to answer the following questions.

- i. (2 pts) What happens to the values of the mean as  $n$  increases? What values should these means be close to?

The values of the means increase. These observed mean values should be close to the theoretical mean value for the distribution which is  $np = 0.01n$ .

- ii. (2 pts) What happens to the values of the variance as  $n$  increases? What values should these variances be close to?

The values of the variances increase. These observed variances should be close to the theoretical variance value for the distribution which is  $np(1-p) = n(0.01)(0.99) = 0.0099n$ .

- iii. (2 pts) Describe the changes you see in the 4 histograms of the observations as  $n$  increases.

The histogram of the simulated binomial values becomes more peaked as  $n$  increases. The histograms for smaller  $n$  are skewed to the right, but as  $n$  increases, these histograms become more symmetric and the shape of the histograms looks more like a bell-curve for the two larger values of  $n$ .

- (e) (3 pts) In each of the 4 problems above, there is the same relationship between the four values of  $n$  and  $p$ . What is this relationship?

If you look at the calculated mean values for each problem, you will see that the first value for  $n$  has a mean value of approximately 1 for each value of  $p$ , the second value for  $n$  has a mean value of approximately 5 for each value of  $p$ , the third value is 10 and the fourth value is 100. For these 4 problems,  $np = 1, 5, 10$  or  $100$ . As the values of  $p$  decrease, the values of  $n$  have to increase in order to keep the same values for  $np$ .

Note: This relationship between  $n$  and  $p$  becomes very important when we study something called statistical inference for  $p$ .

3. In a playlist of 10 songs, 3 songs are from the band Queen, 4 songs are from the band Poisson, and 3 songs are from Jimmy Buffett. The playlist is shuffled and the first three songs are played. Let  $Y$  denoted the number of Jimmy Buffett songs played in the first three songs played.

- (a) (3 pts) Find the probability distribution function for  $Y$ .

In the selection of songs from a playlist with  $N$  songs, the sampling is without replacement. If you keep listening to songs, you will hear all songs in the playlist, but if you only listen to the first  $n$  songs, the number of songs you will listen to from a particular artist has a hypergeometric distribution where  $r$  is the number of songs from that artist in the playlist,  $n$  is the number of songs from the playlist you listen to, and  $N$  is the number of songs in the playlist.

In this situation, there are  $r = 3$  songs from Jimmy Buffett out of the playlist of  $N = 10$  songs. You are listening to the first  $n = 3$  songs. So  $Y$  has a hypergeometric distribution with  $r = 3$ ,  $n = 3$  and  $N = 10$ .

- (b) (4 pts) Find the mean and variance for  $Y$ .

The mean of a hypergeometric distribution is

$$E(Y) = n \left( \frac{r}{N} \right) = 3 \left( \frac{3}{10} \right) = 0.9$$

The variance of a hypergeometric distribution is

$$V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) = 3 \left( \frac{3}{10} \right) \left( \frac{7}{10} \right) \left( \frac{7}{9} \right) = 441/900 = 0.49$$

4. (41 pts) An urn contains  $N$  marbles, of which 50% are green, 20% are blue and 30% are red. Three marbles are to be drawn from the urn. Let the random variable  $Y$  be the number of green marbles selected from the urn.

- (a) (9 pts) Assume  $N = 10$  and the marbles are drawn from the urn without replacement.

- i. (3 pts) What is the distribution of the random variable  $Y$ ?

The distribution is hypergeometric where  $r = 5$ ,  $n = 3$  and  $N = 10$ .

- ii. (2 pts) Use R to calculate the probability distribution function values for the random variable  $Y$ .

Here are the values for the probabilities of  $Y$ .

```
dhypcr(0:3, 5, 5, 3)
0.08333333 0.41666667 0.41666667 0.08333333
```

- iii. (2 pts) What is the expected value of the random variable  $Y$ ?

$$E(Y) = n \left( \frac{r}{N} \right) = 3 \left( \frac{5}{10} \right) = 1.5$$

- iv. (2 pts) What is the variance of the random variable  $Y$ ?

$$V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) = 3(5/10)(5/10)(7/9) = 0.5833$$

- (b) (9 pts) Still assume  $N = 10$  but that the marbles are drawn from the urn with replacement. This means that the first marble is drawn, its color noted and then replaced before the second marble is drawn from the urn.

- i. (3 pts) What is the distribution of the random variable  $Y$ ?

The distribution is binomial with  $n = 3$  and  $p = 5/10 = 0.5$

- ii. (2 pts) Use R to calculate the probability distribution function values for the random variable  $Y$ .

Here are the probabilities of the values of  $Y$ .

```
dbinom(0:3, 3, 0.5)
0.125 0.375 0.375 0.125
```

- iii. (2 pts) What is the expected value of the random variable  $Y$ ?

$$E(Y) = np = 3(0.5) = 1.5$$

- iv. (2 pts) What is the variance of the random variable  $Y$ ?  
 $V(Y) = np(1 - p) = 3(0.5)(0.5) = 0.75$ .
- (c) (9 pts) Now assume  $N = 1000$  and the marbles are drawn from the urn without replacement.
- i. (3 pts) What is the distribution of the random variable  $Y$ ?  
 The distribution is hypergeometric where  $r = 500$ ,  $n = 3$  and  $N = 1000$ .
- ii. (2 pts) Use R to calculate the probability distribution function values for the random variable  $Y$ .  
 Here are the values for the probabilities of  $Y$ .  
`dhyper(0:3, 500, 500, 3)`  
 0.1246246 0.3753754 0.3753754 0.1246246
- iii. (2 pts) What is the expected value of the random variable  $Y$ ?  
 $E(Y) = n \left( \frac{r}{N} \right) = 3 \left( \frac{500}{1000} \right) = 1.5$
- iv. (2 pts) What is the variance of the random variable  $Y$ ?  
 $V(Y) = n \left( \frac{r}{N} \right) \left( \frac{N-r}{N} \right) \left( \frac{N-n}{N-1} \right) = 3(500/1000)(500/1000)(997/999) = 0.7485$
- (d) (9 pts) Still assume  $N = 1000$  but that the marbles are drawn from the urn with replacement.
- i. (3 pts) What is the distribution of the random variable  $Y$ ?  
 $Y$  is binomial with  $n = 3$  and  $p = 500/1000 = 0.5$ .
- ii. (2 pts) Use R to calculate the probability distribution function values for the random variable  $Y$ .  
 Here are the code and values.  
`dbinom(0:3, 3, 0.5)`  
 0.125 0.375 0.375 0.125  
 Note: the code and values are the same as in part (b).
- iii. (2 pts) What is the expected value of the random variable  $Y$ ?  
 $E(Y) = np = 3(0.5) = 1.5$
- iv. (2 pts) What is the variance of the random variable  $Y$ ?  
 $V(Y) = np(1 - p) = 3(0.5)(0.5) = 0.75$
- (e) (5 pts) Use your answers to above to discuss the effect the size of  $N$  and the selection method (without replacement and with replacement) have on the probability distribution function for  $Y$ , the probability  $Y = 3$  and the expected value and variance of the random variable  $Y$ .

In this problem, the probability of success on the first draw from the population is the same  $p = 0.5$ . When  $N$  is small and you draw without replacement, the probability of success on future draws changes greatly. This is why the probability distribution function for drawing without replacement is very different from the p.d.f. for drawing with replacement. While the expected values are the same for both selection methods, the variance is much smaller for drawing without replacement.

When  $N$  is large and you draw without replacement, the probability of success on future draws changes little. This is why the probability distribution function for drawing without replacement is not very different from the p.d.f. for drawing with replacement. The expected values are still the same for both selection methods, but the variances are much closer. This is seen through the similar p.d.f. of the two selection methods.

As the size of the population grows, sampling without replacement behaves more and more like sampling with replacement.

5. (14 pts) In a Newsweek article from January, 2005, Steven Levy reported on the perceived non-random behavior of the random shuffle feature on the iPod. As one example, he reported on the lack of favoritism for a particular song he purchased online. “Months after I bought *Wild Thing* from the iTunes store, I’m still waiting for my iPod to cue it up.” If you listen to an entire shuffled playlist, you will hear all songs on the playlist just once. However, this scenario almost never happens since people will listen only to the first  $n$  songs in an  $N$  song playlist before reshuffling the same playlist or choosing another.

- (a) (2 pts) For one shuffled playlist where you listen to the first  $n$  songs out of  $N$  total, what is the probability that one song will be played?

The song will be in the first  $n$  songs of an  $N$  song playlist with probability  $n/N$ .

- (b) (6 pts) Assume that successive shuffles of the playlist are independent. Let  $Y$  denote the number of shuffles required in order to hear a particular song once when you listen to the first  $n$  songs out of  $N$  total in each shuffle. What is the distribution of  $Y$ ? What is the mean of  $Y$ ? What is the variance of  $Y$ ?

Each shuffle is independent, and the probability of hearing the song in the first  $n$  songs of an  $N$  song playlist is the same for each shuffle  $n/N$ . This means that the number of shuffles required in order to hear the song when you listen to the first  $n$  songs out of  $N$  songs total in the playlist has a negative binomial distribution. We want to hear the song once, so the  $r$ th success is 1 and  $p$  is  $n/N$ .

The mean of  $Y$  is  $E(Y) = 1/p = N/n$  and the variance of  $Y$  is  $V(Y) = \frac{1-n/N}{(n/N)^2}$

- (c) (6 pts) Assume the playlist has  $N = 3000$  songs. What is the mean and variance of  $Y$  if you listen to the first  $n = 20, 50, 100$  songs before reshuffling the same playlist or choosing another?

For  $n = 20$  and  $N = 3000$ ;  $E(Y) = 3000/20 = 150$  and  $V(Y) = \frac{1-n/N}{(n/N)^2} = \frac{1-20/3000}{(20/3000)^2} = 22350$

For  $n = 50$  and  $N = 3000$ ;  $E(Y) = 3000/50 = 60$  and  $V(Y) = \frac{1-n/N}{(n/N)^2} = \frac{1-50/3000}{(50/3000)^2} = 3540$

For  $n = 100$  and  $N = 3000$ ;  $E(Y) = 3000/100 = 30$  and  $V(Y) = \frac{1-n/N}{(n/N)^2} = \frac{1-100/3000}{(100/3000)^2} = 870$

6. (14 pts) In a library of 3000 songs, 50 are from a particular group. The library is shuffled and played until the first song from this group appears. Let  $Y$  denote the number of the song in the shuffle where the first song from this group is played. For example, if the first song from this group is the 4th song played in the shuffle,  $Y = 4$ .

- (a) (5 pts) What is the distribution of the random variable  $Y$ ? What values can the random variable  $Y$  take?

The shuffling behaves like sampling without replacement from a collection of  $N = 3000$  things where  $r = 50$  of them belong to one group. The number of shuffles to get the first song from this group is then negative hypergeometric where  $k = 1$ .

In this case, the values of  $Y$  can go from 1 to 2951.

- (b) (4 pts) Use R to calculate the value of  $p(y)$  for each possible value of  $y$ . What do you notice about the value of  $p(y)$  as  $y$  increases?

You will need the `dnhyper` function in R to do this problem.

```
dneghyper<- function(y,r,N,k){choose(y-1,k-1)*choose(N-y,r-k)/choose(N,r)}
dneghyper(1:2951, 50, 3000, 1)
```

These values might be easier to see with a plot. The plot is contained at the end of the assignment. The probabilities decrease as the value of  $y$  decreases.

- (c) (5 pts) Use R to find the theoretical value of the median of  $Y$ . Define what this value means in words.

The theoretical median of the random variable  $Y$  is the smallest value of  $y$  such that  $P(Y \leq y) \geq 0.5$ . We can use the cumsum command in R to find the values of  $P(Y \leq y)$  for  $y$  from 1 to 2951. The plot of these values is contained at the end of the assignment. You can see from the plot, that the median value occurs very early in the values of  $y$ . Looking specifically at the first 100 values of  $P(Y \leq y)$  we have that the probability is 0.5003 when  $y = 41$ .

This means that in 50% of all shuffles, we will hear the first song from this group at or before the 41st song in the shuffle.

7. (8 pts) The number of bacteria colonies of a certain type in sample of polluted water has a Poisson distribution with a mean of 2 per 1-cubic centimeter.

- (a) (3 pts) Find the probability that a 1-cubic centimeter sample taken from this water will have at least two bacteria colonies.

We need to find  $P(Y \geq 2)$ . This is  $1 - P(Y \leq 1)$  which can be found using the R code

```
1 - sum(dpois(0:1,2))
[1] 0.5939942
```

- (b) (5 pts) If 10 independent 1-cubic centimeter samples are taken from this water, find the probability that exactly 3 of these samples will contain two or more bacteria colonies.

So this means that 3 of them will contain two or more bacteria colonies and 7 of them will not. The number of bacteria colonies in the 10 samples are independent and so the number of samples that will contain two or more bacteria colonies has a binomial distribution with  $n = 10$  and  $p = 1 - P(Y \leq 1) = 0.5939942$ . We can determine the probability that exactly 3 of these sample will contain two or more bacteria colonies using the binomial distribution formula. In R, this is

```
dbinom(3,10,1 - sum(dpois(0:1,2)))
[1] 0.04573554
```

8. (10 pts) In class, we learned a Poisson distribution has the same value for the mean and variance. Use R to conduct an investigation of the shape of the Poisson distribution for different values of the parameter  $\lambda$ . What do you notice about the shape of the distribution relative to the parameter  $\lambda$ ?

Similar to the investigation of the Binomial distribution in problem 2, we will simulate Poisson values using R for different values of  $\lambda$ . I am going to use  $\lambda = 0.1, 0.5, 1, 5, 10, 100$ .

Here is the R code to simulate and graph the histograms of 10000 Poisson values for each value of  $\lambda$ .

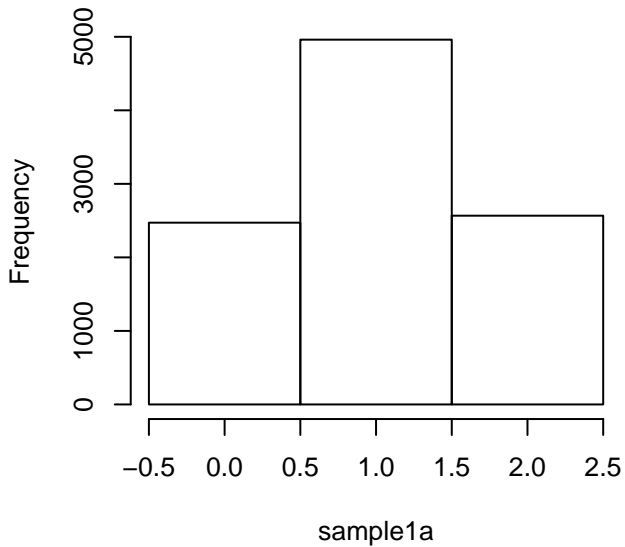
```
pois0.1<- rpois(10000,0.1) #simulating values
pois0.5<- rpois(10000,0.5)
```

```
pois1<- rpois(10000,1)
pois5<- rpois(10000,5)
pois10<- rpois(10000,10)
pois100<- rpois(10000,100)

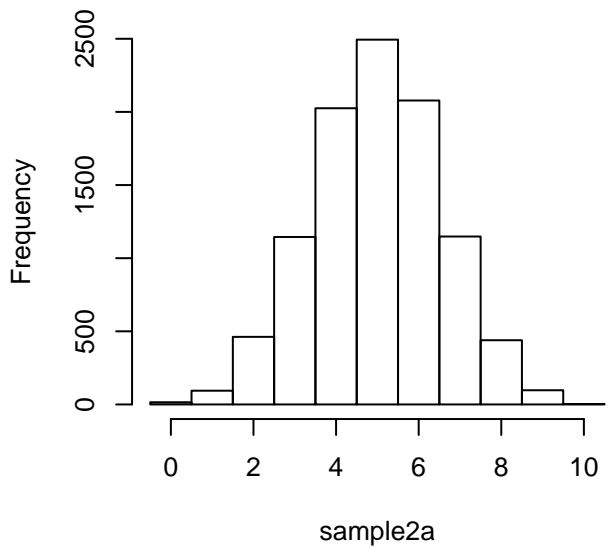
par(mfrow = c(2,3)) #puts 6 histograms on one page
hist(pois0.1, breaks = c(min((pois0.1) - 0.5):(max(pois0.1) + 0.5)))
hist(pois0.5, breaks = c(min((pois0.5) - 0.5):(max(pois0.5) + 0.5)))
hist(pois1, breaks = c(min((pois1) - 0.5):(max(pois1) + 0.5)))
hist(pois5, breaks = c(min((pois5) - 0.5):(max(pois5) + 0.5)))
hist(pois10, breaks = c(min((pois10) - 0.5):(max(pois10) + 0.5)))
hist(pois100, breaks = c(min((pois100) - 0.5):(max(pois100) + 0.5)))
```

As you can see from the graphs, the distribution of the Poisson values is skewed to the right for small values of  $\lambda$ . As  $\lambda$  increases, the distribution becomes more symmetric. At  $\lambda = 10$ , the graph is still a little right-skewed, but when  $\lambda = 100$ , the distribution looks very symmetric and bell-shaped.

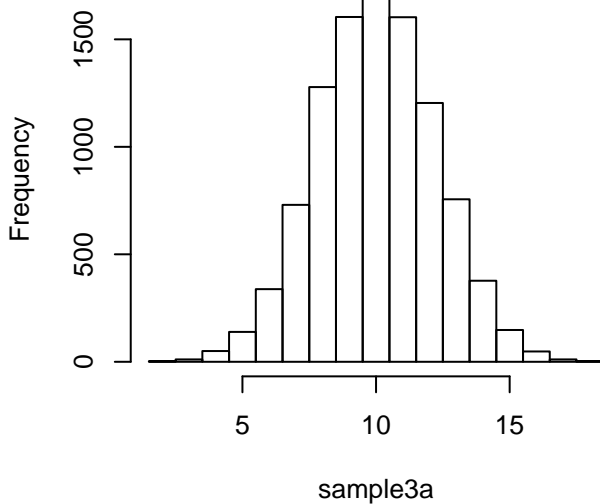
**Histogram of sample1a**



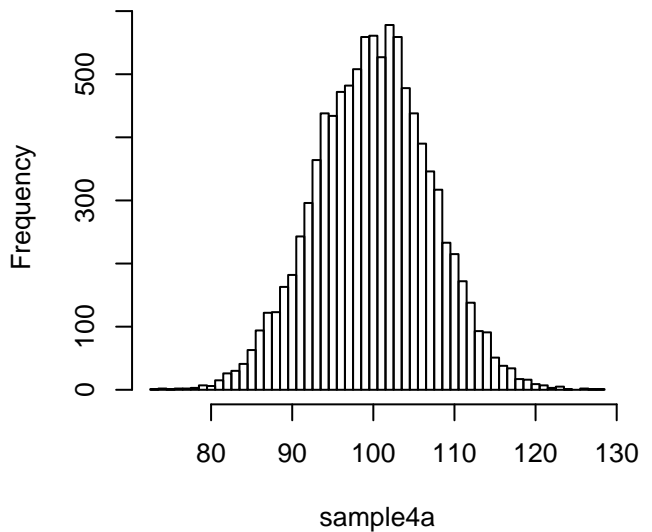
**Histogram of sample2a**



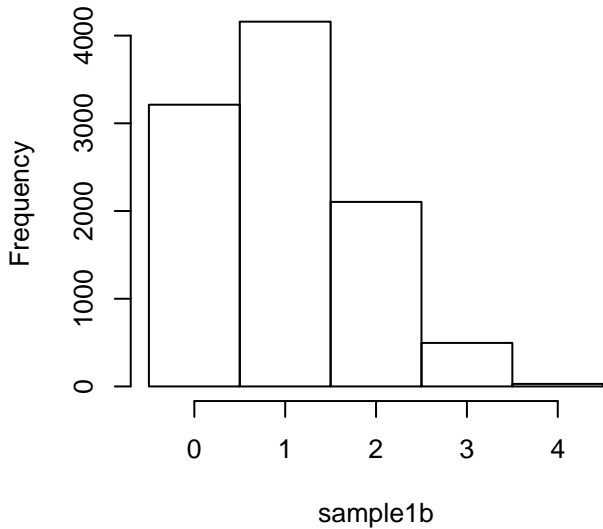
**Histogram of sample3a**



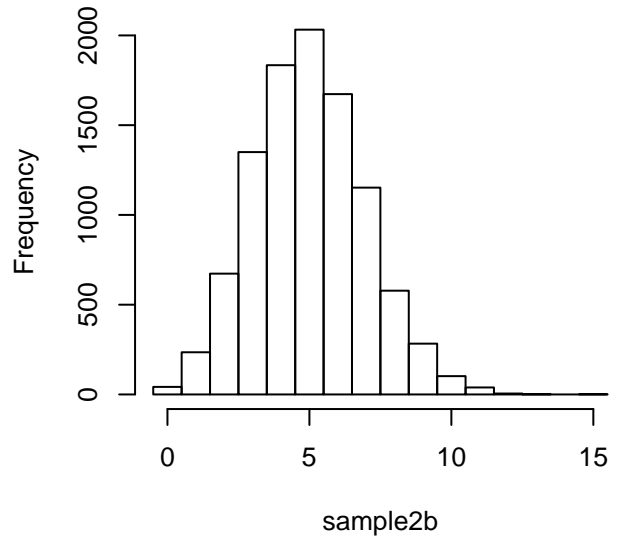
**Histogram of sample4a**



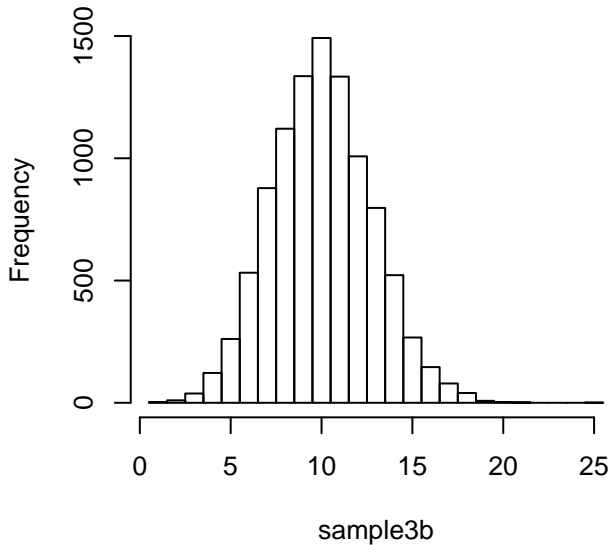
### Histogram of sample1b



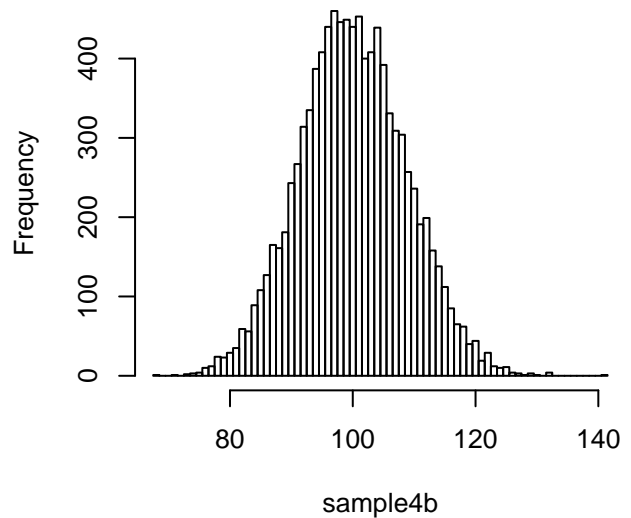
### Histogram of sample2b



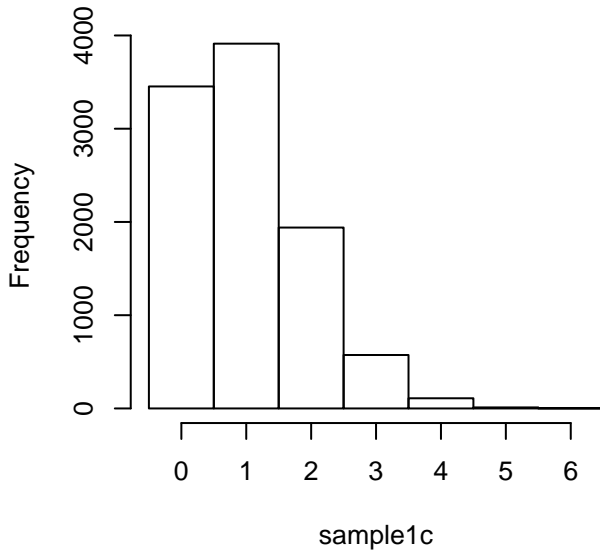
### Histogram of sample3b



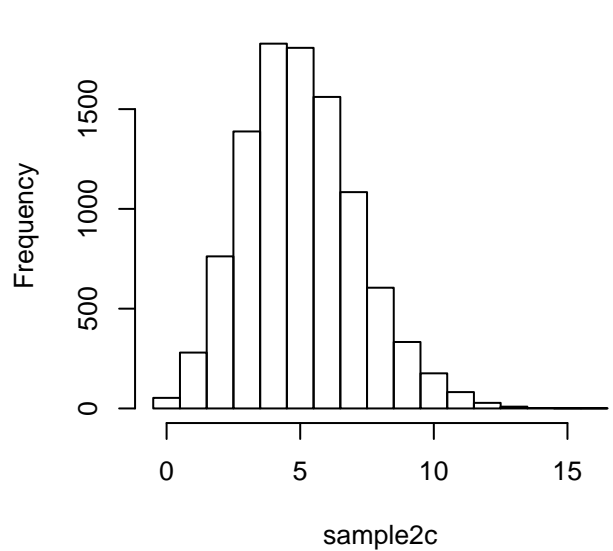
### Histogram of sample4b



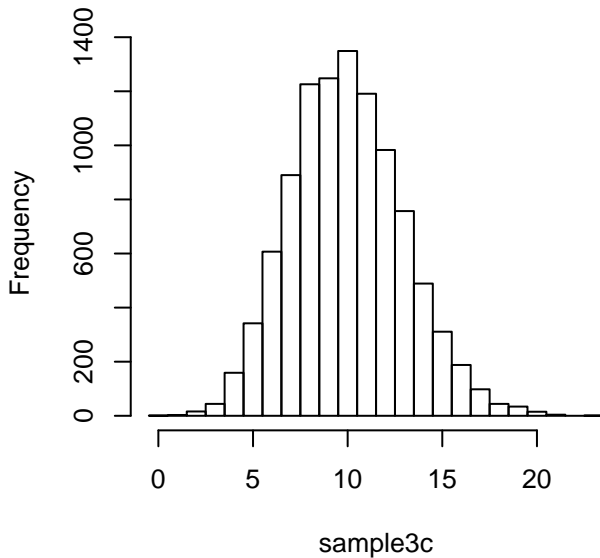
### Histogram of sample1c



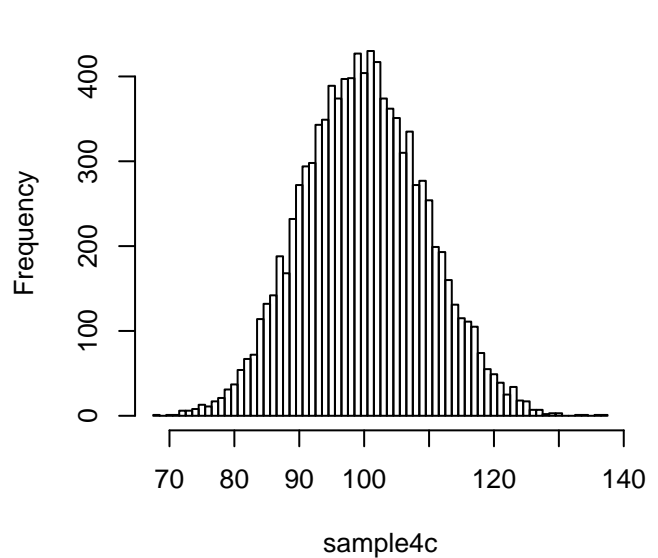
### Histogram of sample2c



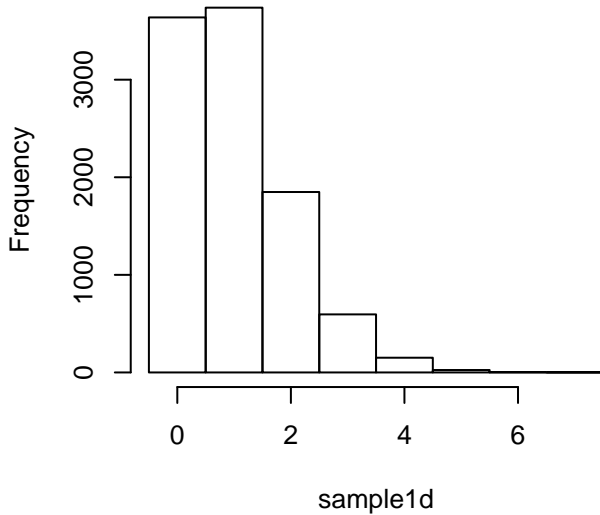
### Histogram of sample3c



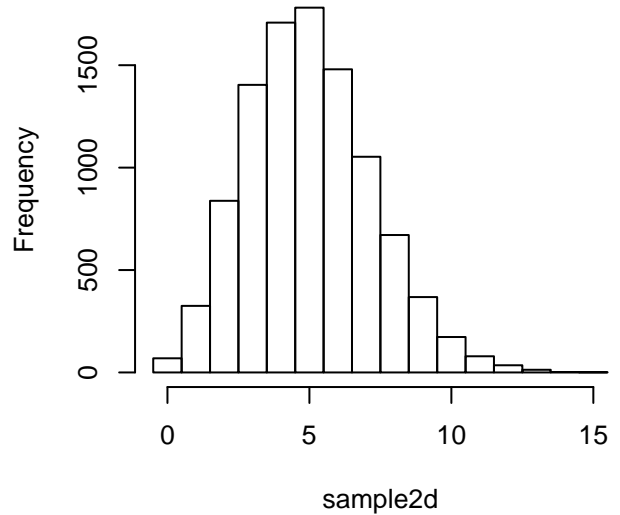
### Histogram of sample4c



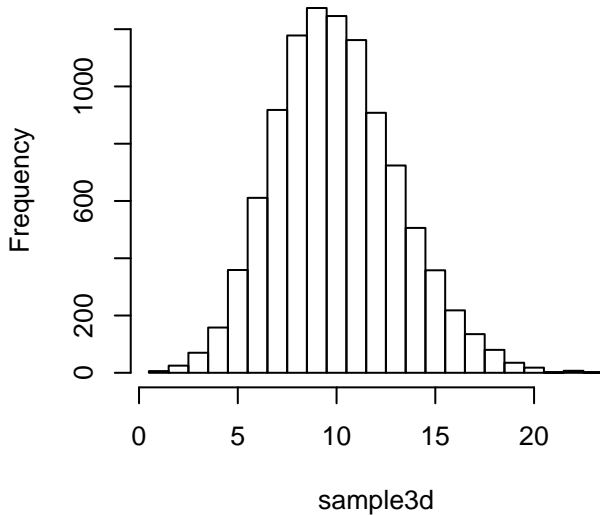
### Histogram of sample1d



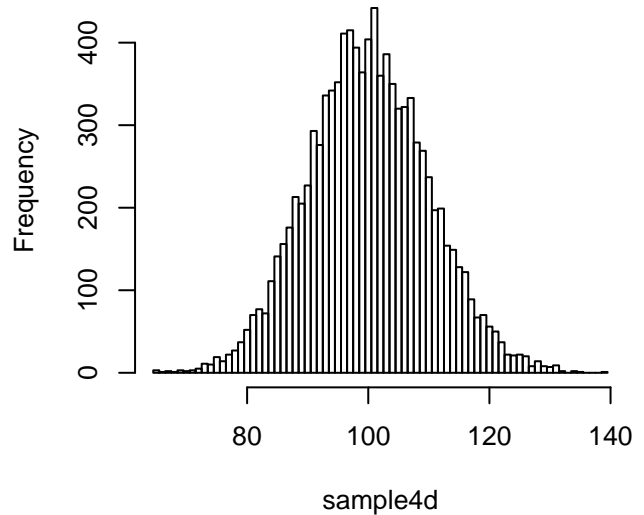
### Histogram of sample2d



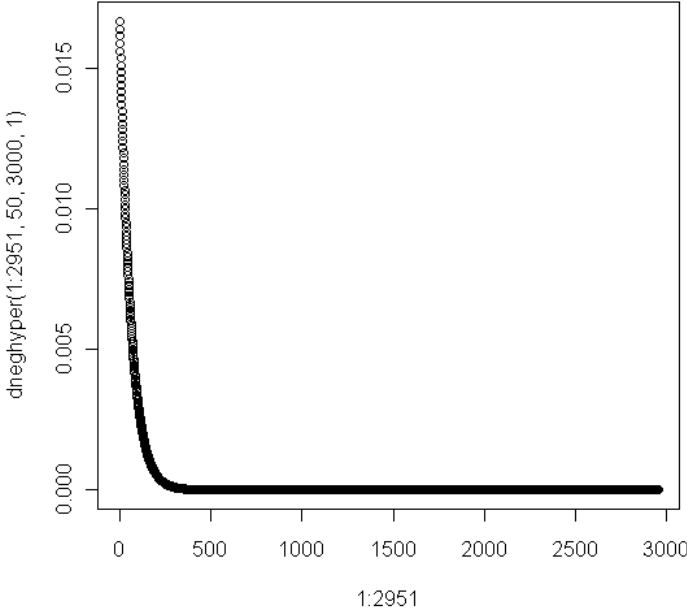
### Histogram of sample3d



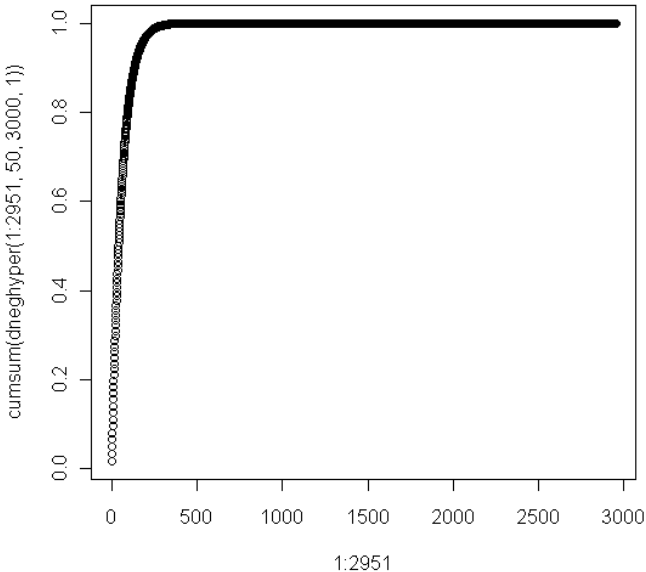
### Histogram of sample4d



Below is the graph for the negative hypergeometric distribution probabilities from problem #6.

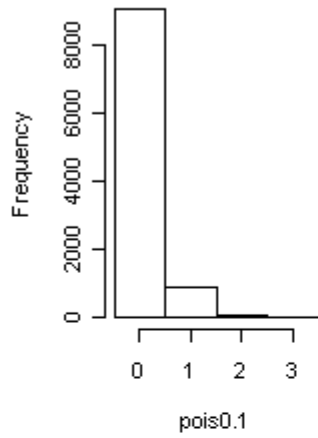


Below is the graph for the negative hypergeometric distribution values of  $P(Y \leq y)$ .

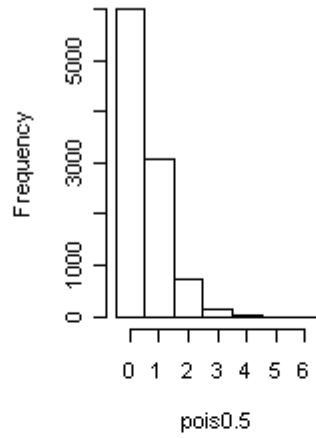


Below is the graph of the simulated Poisson values for 6 different values of  $\lambda$ .

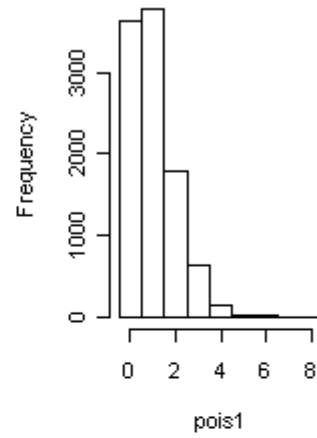
**Histogram of pois0.1**



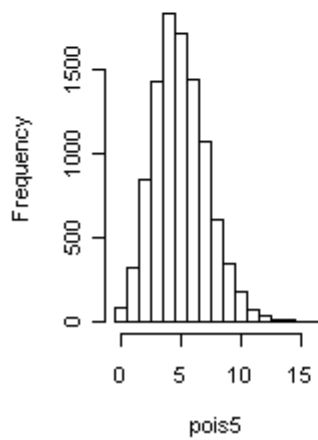
**Histogram of pois0.5**



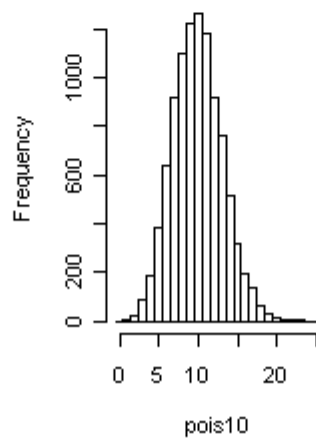
**Histogram of pois1**



**Histogram of pois5**



**Histogram of pois10**



**Histogram of pois100**

