Problem Set 2

Adam Ten Hoeve

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please complete the following tasks regarding the data in R. Please generate a solution document in R markdown and upload the .Rmd document and a rendered .doc, .docx, or .pdf document. Your solution document should have your answers to the questions and should display the requested plots.

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(knitr)
```

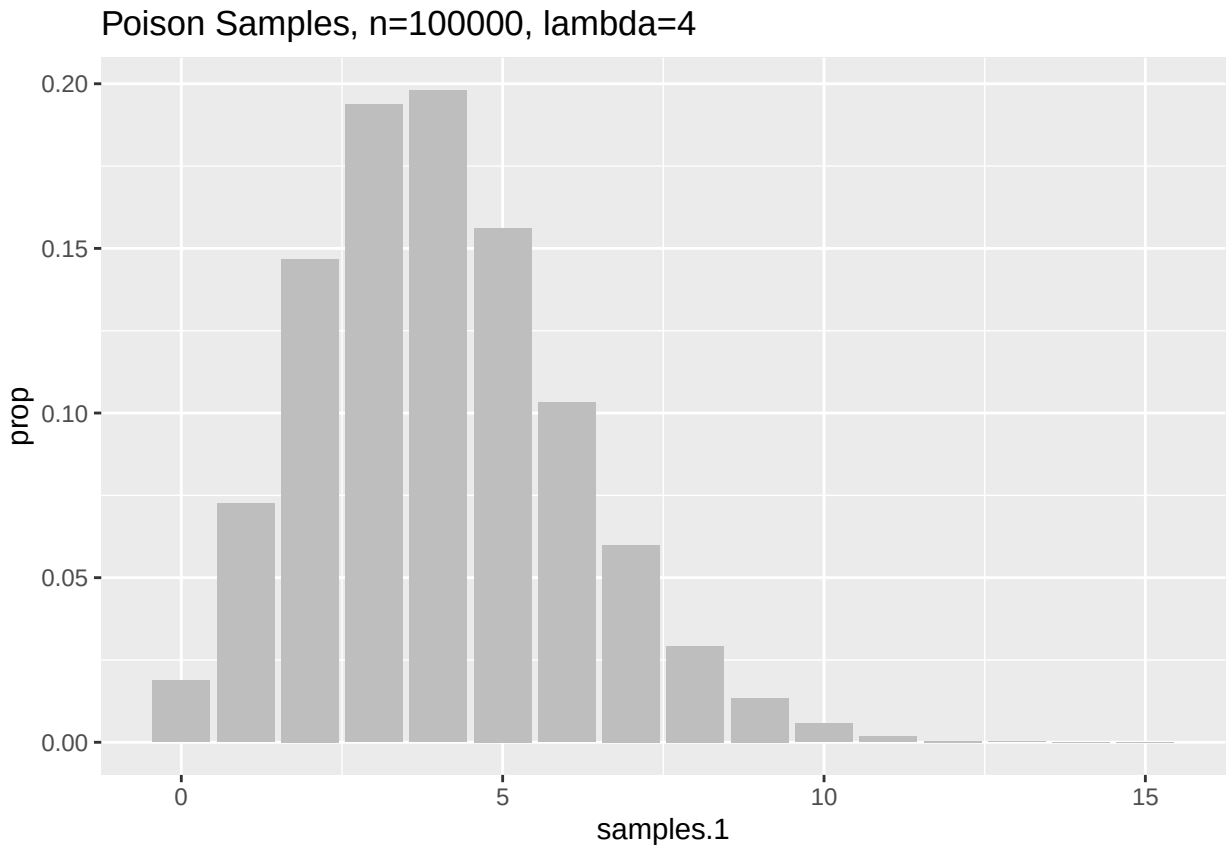
Question 1

Please draw a random sample of 100,000 values from the Poisson distribution with $\lambda = 4$ using the random seed 7654321. Present a histogram of the results with a density scale using the techniques in "Discrete_Probability_Distributions_2_3_3.Rmd" or "continuous_probability_distributions_2_4_2.Rmd". You may find a bin width of 1 helpful. Separately, please draw 100 values from the Poisson distribution with $\lambda = 4$ using the random seed 7654321 and present a histogram of the results with a density scale.(5 points)

```
set.seed(7654321)
# Draw 100000 values from the distribution
k.1 <- 100000
# Generate samples using the inbuilt poison function
samples.1 <- rpois(k.1, lambda=4)
prop.1 <- data.frame(samples.1)
prop.1 <- summarize(group_by(prop.1, samples.1), prop=n()/nrow(prop.1))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

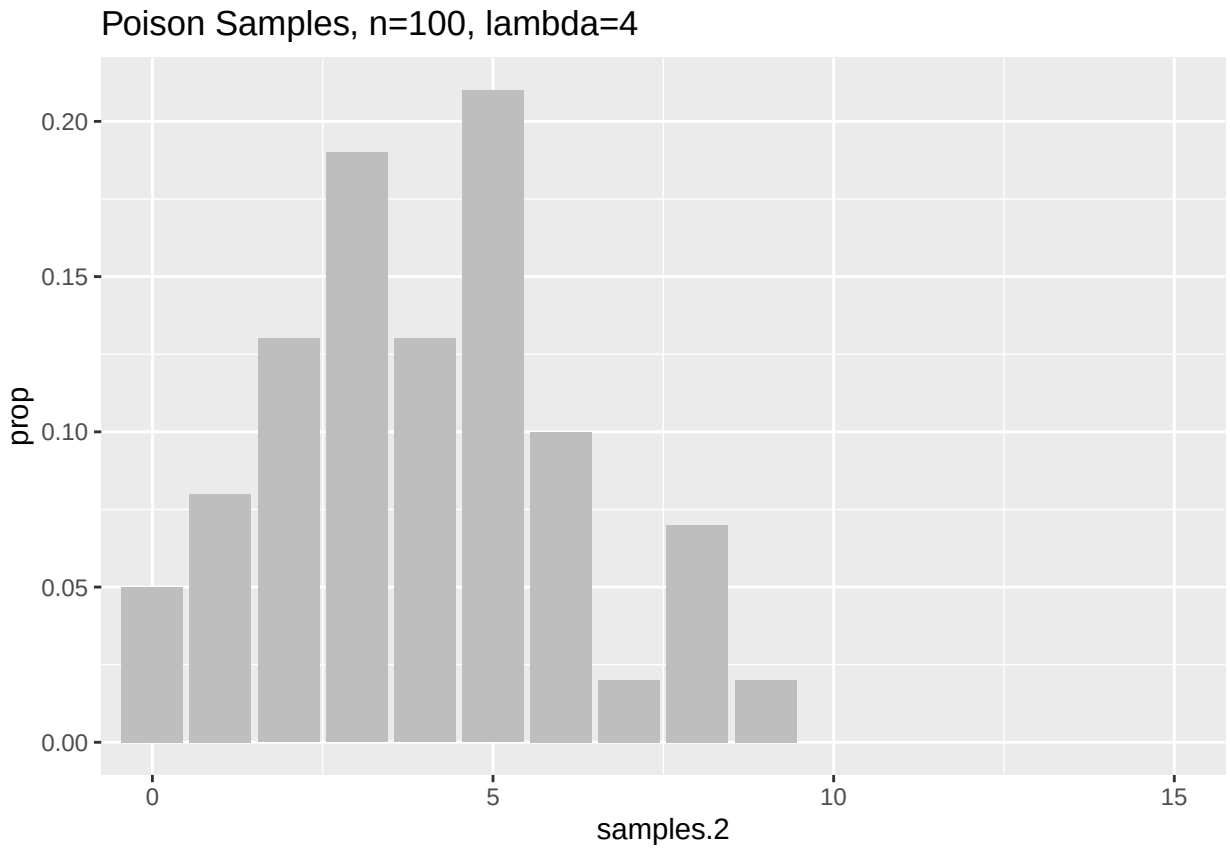
```
# Plot a histogram of the resulting proportion densities.
hist.1 <- ggplot(data=prop.1, aes(x=samples.1)) +
            geom_col(aes(y=prop), fill="gray") +
            labs(title="Poison Samples, n=100000, lambda=4")
hist.1
```




```
set.seed(7654321)
# Draw 100 values from the distribution
k.2 <- 100
# Generate samples using the inbuilt poison function
samples.2 <- rpois(k.2, lambda=4)
prop.2 <- data.frame(samples.2)
prop.2 <- summarize(group_by(prop.2, samples.2), prop=n()/nrow(prop.2))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

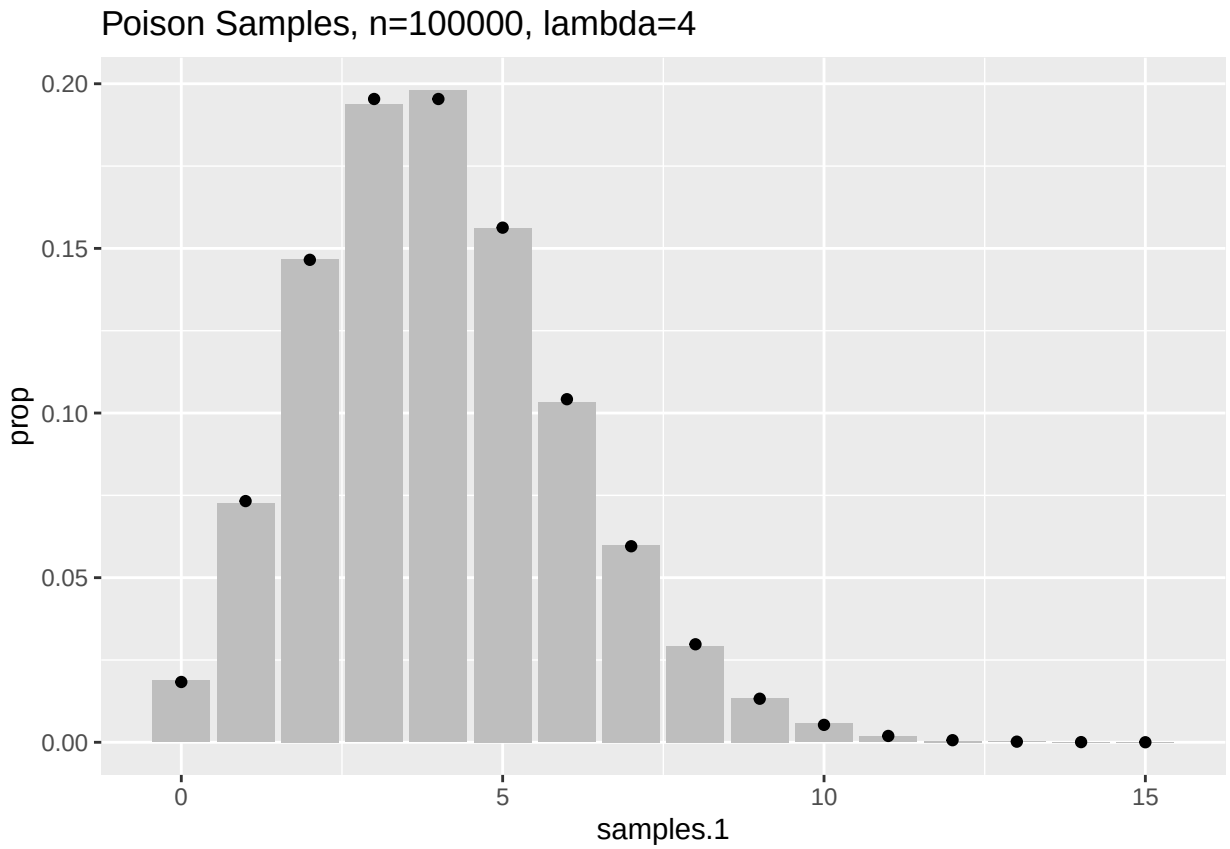
```
# Plot a histogram of the resulting proportion densities.
hist.2 <- ggplot(data=prop.2, aes(x=samples.2)) +
          geom_col(aes(y=prop), fill="gray") +
          labs(title="Poison Samples, n=100, lambda=4") +
          coord_cartesian(xlim=c(0,15))
hist.2
```



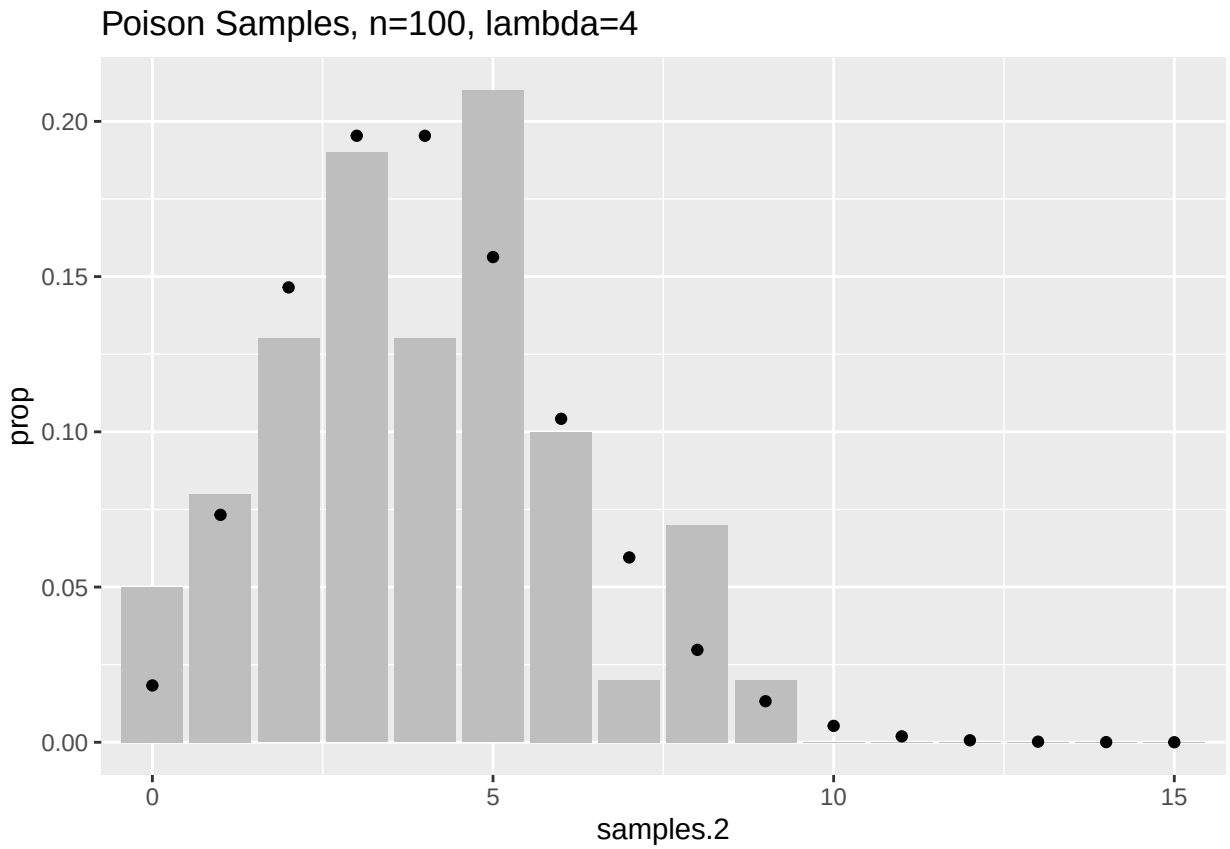
Question 2

Please generate a visualization that compares the proportions each of the possible outcomes in the sample of size 100,000 above to the theoretical probabilities of each of the outcomes for the Poisson distribution with $\lambda = 4$. Repeat for the sample of size 100. Discuss the appearance of the histograms in relation to the probability density function of the Poisson distribution with $\lambda = 4$. (5 points)

```
# Calculate the theoretical probabilities
dens <- dpois(0:max(c(samples.1, samples.2)), lambda=4)
# Add the theoretical densities to the first plot.
hist.1 <- ggplot(data=prop.1, aes(x=samples.1)) +
            geom_col(aes(y=prop), fill="gray") +
            labs(title="Poison Samples, n=100000, lambda=4") +
            # Add the theoretical points to the plot.
            geom_point(aes(y=dens))
hist.1
```




```
# Add zero proportion for extra unobserved values
unobserved.values <- nrow(prop.2):max(samples.1)
zeros <- rep(0, length(unobserved.values))

df.zeros <- data.frame(unobserved.values, zeros)
names(df.zeros) <- c("samples.2", "prop")
prop.2 <- rbind(prop.2, df.zeros)
# Add the theoretical densities to the second plot.
hist.2 <- ggplot(data=prop.2, aes(x=samples.2)) +
          geom_col(aes(y=prop), fill="gray") +
          labs(title="Poison Samples, n=100, lambda=4") +
          coord_cartesian(xlim=c(0,15)) +
          geom_point(aes(y=dens))
hist.2
```



From the histograms, we can see that the sample with 100,000 values was much closer to the theoretical density than the sample with 100 values. We can learn from this that the more samples we have, the closer our simulated values will match the theoretical distribution.

Normal Approximations

Many statistical methods involve approximation of a distribution by a Normal distribution. Questions 3 through 7 are intended to build intuition for when this is reasonable. The questions work toward visually assessing the quality of Normal approximations to several distributions.

The issue of the sd to use in a Normal approximation will be handled by use of the interquartile range, defined below.

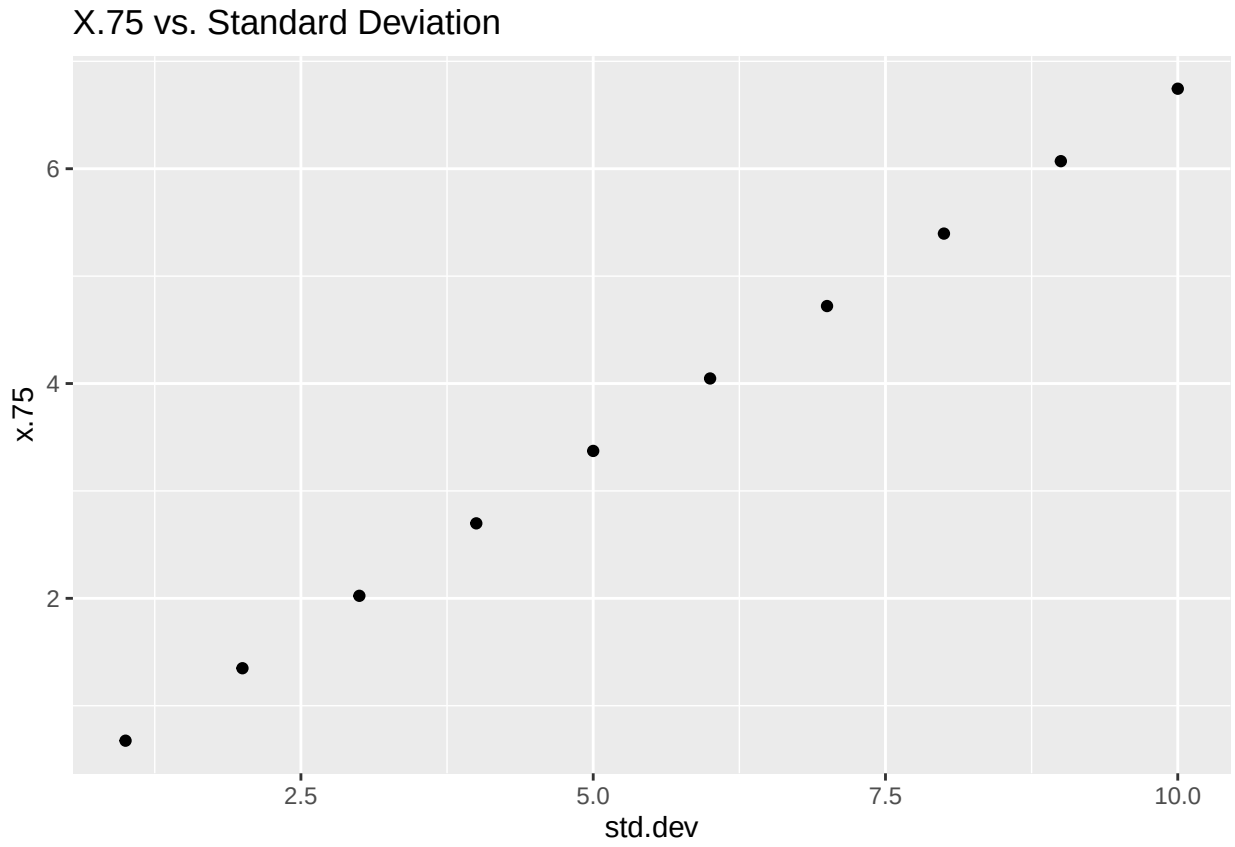
Question 3

Approximately, the p^{th} quantile of a set S of n numerical values is the value s_p such that the proportion of values in S that are less than or equal to s_p equals p. This concept may be familiar from the idea of the 95^{th} percentile. There are complications arising from the fact that there may not be a value s_p for which the proportion exactly equals p. The "quantile" function in R defaults to one approach to addressing this. The default is acceptable for these exercises.

The **interquartile range** of S is the value $s_{0.75} - s_{0.25}$.

Question 3.a By analogy, for a random variable with cumulative density function F, let $x_{0.25}$ satisfy $F(x_{0.25}) = 0.25$. Let $x_{0.75}$ satisfy $F(x_{0.75}) = 0.75$. Define $q = x_{0.75} - x_{0.25}$ to be the interquartile range for the random variable. Please calculate the values of $x_{0.75}$ for the Normal distributions with mean 0 and sd in $1, 2, 3, ...10$ and plot the points consisting of the value of the sd and the corresponding $x_{0.75}$. This should give an indication of a simple function relating sd and $x_{0.75}$. (5 points)

```r
std.dev = 1:10
x.75 = qnorm(0.75, mean=0, sd=std.dev)
d = data.frame(std.dev, x.75)

# Plot the points. Gives an inverse relationship.
ggplot(d, aes(x=std.dev, y=x.75)) + geom_point() + labs(title="X.75 vs. Standard Deviation")
```



Question 3.b Note that for mean=0, sd=σ, and $w1$ and $w2$ satisfying $0.25 = \int_{-\infty}^{w_1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx$ and $0.75 = \int_{-\infty}^{w_2} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx$, the value of q is $w2 - w1$. By symmetry of the Normal family, $w_1 = -w_2$. Please use this integration with a change of variable transforming the integrand to the density of a standard Normal distribution to calculate the function relating the value of σ and $w2$ on theoretical grounds. You may find that defining x_2 by $0.75 = \int_{-\infty}^{x_2} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$ simplifies your argument. (5 points)

Using change of variable, we let $z = \frac{x}{\sigma}$ and $dz = \frac{1}{\sigma} dx$. Then we plug these into the normal distribution to get:

$$0.75 = \int_{-\infty}^{w_2/\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(z/\sigma)^2}{2\sigma^2}} (\sigma dz) = \int_{-\infty}^{w_2/\sigma} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

From our bounds, we can see that $x_2 = w_2/\sigma \rightarrow \boxed{w_2 = x_2\sigma}$, where x_2 is the 3rd Quartile of the standard normal.

3.c Please use the results of 3.b to derive the function relating the interquartile range q to the value of σ for Normal distributions with mean equal to 0. (3 points)

By combining the equations $q = w_2 - w_1$ and $w_2 = -w_1$, we get $q = w_2 - (-w_2) = 2w_2$. Pluggin this into our equation from part 2 we get $\boxed{q = 2(x_2\sigma)}$.

Question 3.d In terms of x_2, μ, and σ, what is the interquartile range for a Normal distribution with mean equal to μ and "sd" equal to σ? (2 points)

Because we are now including the mean μ, we need to redo our initial variable substitution. Instead of $z = \frac{x}{\sigma}$, we will use $z = \frac{x-\mu}{\sigma}$ and $dz = \frac{1}{\sigma}^2$. Following the same process as Part 2, we get:

$$0.75 = \int_{-\infty}^{\frac{w_2-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz$$

Again from our bounds we get $x_2 = \frac{w_2-\mu}{\sigma} \rightarrow w_2 = x_2\sigma + \mu$. Plugging this into $q = 2w_2$, we get $\boxed{q = 2(x_2\sigma + \mu)}$.

Question 4

Consider the probability space (S, M, P) that models the observed proportion of successes in n independent Bernoulli trials with probability of success equal to p. For example, given $n = 12$ and $p = 0.4$, the outcome $\frac{5}{12} = .41\bar{6}$ in (S, M, P) should have the same probability as the outcome of 5 successes in 12 independent Bernoulli trials with probability of success equal to 0.4. Give a formula for the density function $f(s)$ for this probability space in terms of the density $f_{n,p}$ of the binomial distribution with size n and probability p. (5 points)

A Binomial distribution is the same as the sum of n Bernoulli trials. Normally, the random variable for this distribution is the total number of successes x in those n trials. Therefor, if we want a function for the random variable $s = x/n$ to give them same probability as a Binomial(x), then we need to multiply s by n in the input to our Binomial function. $\boxed{f(s) = f_{n,p}(s \cdot n)}$.

Question 5

For each pair (n, p) with $n \in \{10, 50, 1000\}$ and $p \in \{.5, .1, .01\}$, sample 100,000 values from the distribution of the observed proportion of successes in n independent Bernoulli trials with probability of success equal to p as in (S, M, P) above and compute the mean and interquartile range of the sample. The data frame and vectors provided may be helpful.(5 points)

```
ns<-c(10,50,1000)
ps<-c(.5,.1,.01)
k<-100000

dat.samp<-
  data.frame(ns=rep(ns,times=rep(length(ps),length(ns))),
             ps=rep(ps,times=length(ns)))
# For each (n,p), sample from the binomial distribution to find the number of success.
set.seed(0)
means <- numeric(nrow(dat.samp))
IQRs <- numeric(nrow(dat.samp))
# Create matrix to store each drawn sample
samples <- matrix(0, nrow=nrow(dat.samp), ncol=k)
i <- 1
for (n in ns){
    for (p in ps){
        # Sample from the binomial distribution
        successes <- rbinom(k, size=n, prob=p)
```

```
        # save the mean, IQR and sample.
        means[i] <- mean(successes)
        IQRs[i] <- IQR(successes)
        samples[i, ] <- successes
        # Iterate
        i <- i + 1
    }
}
dat.samp$mean <- means
dat.samp$IQR <- IQRs
dat.samp
```

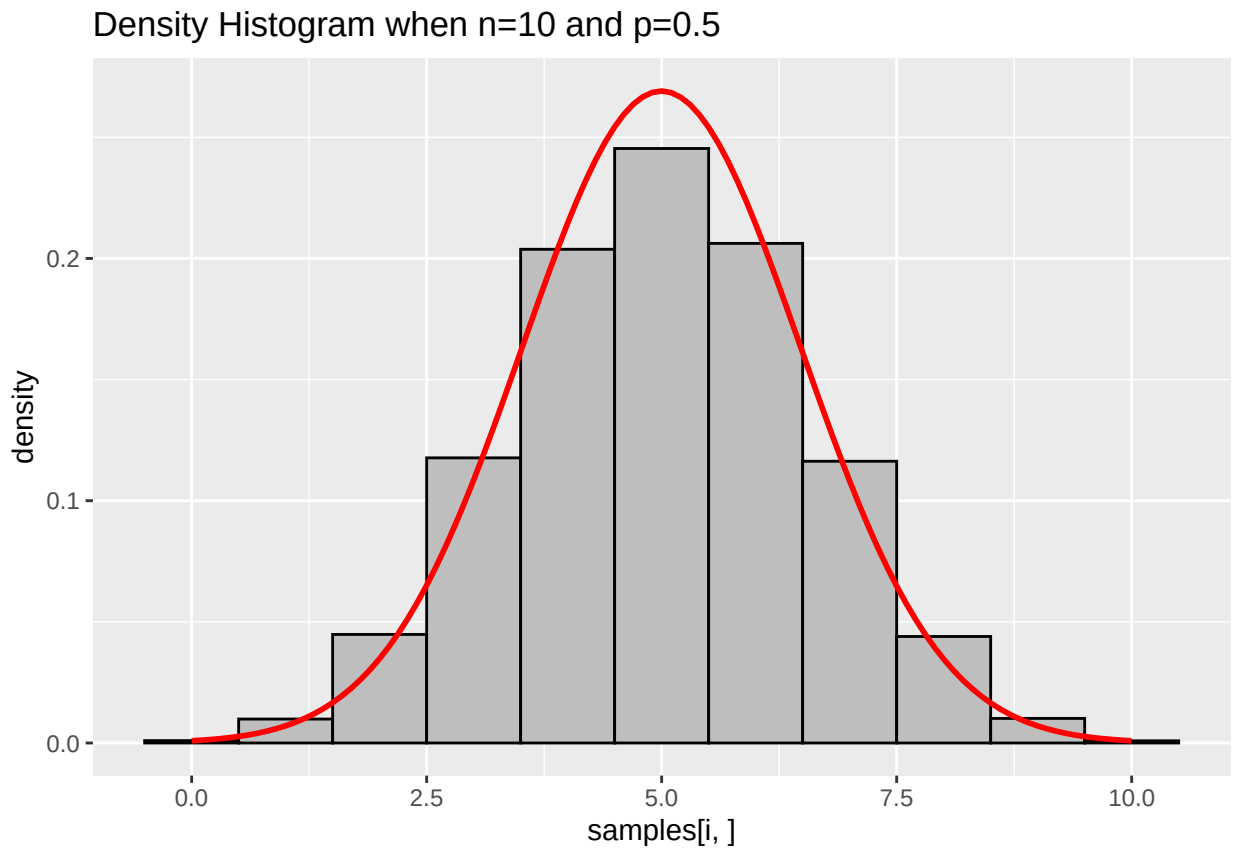
```
##     ns   ps      mean IQR
## 1   10 0.50   4.99778   2
## 2   10 0.10   1.00155   2
## 3   10 0.01   0.09947   0
## 4   50 0.50  24.98796   4
## 5   50 0.10   4.99042   3
## 6   50 0.01   0.50175   1
## 7 1000 0.50 500.01836  22
## 8 1000 0.10  99.99269  13
## 9 1000 0.01   9.99988   4
```

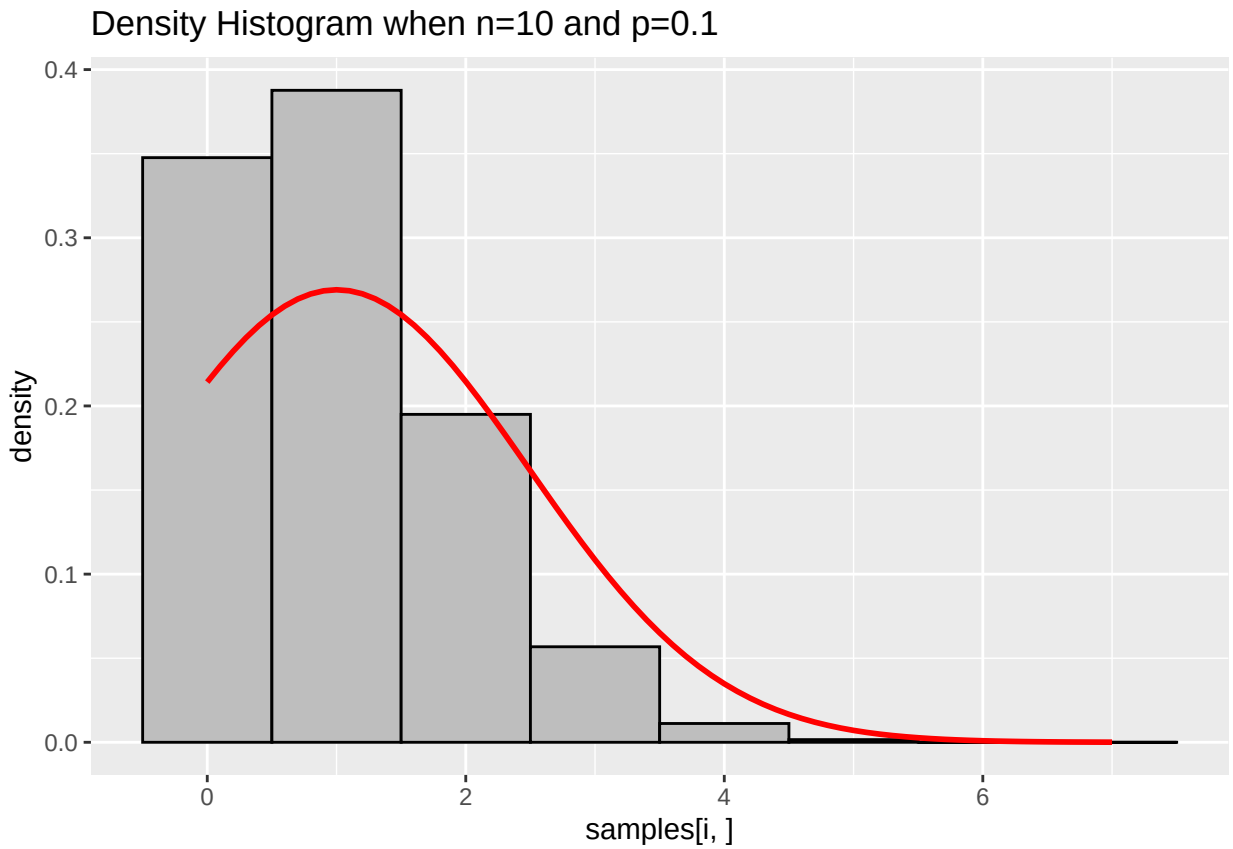
Question 6

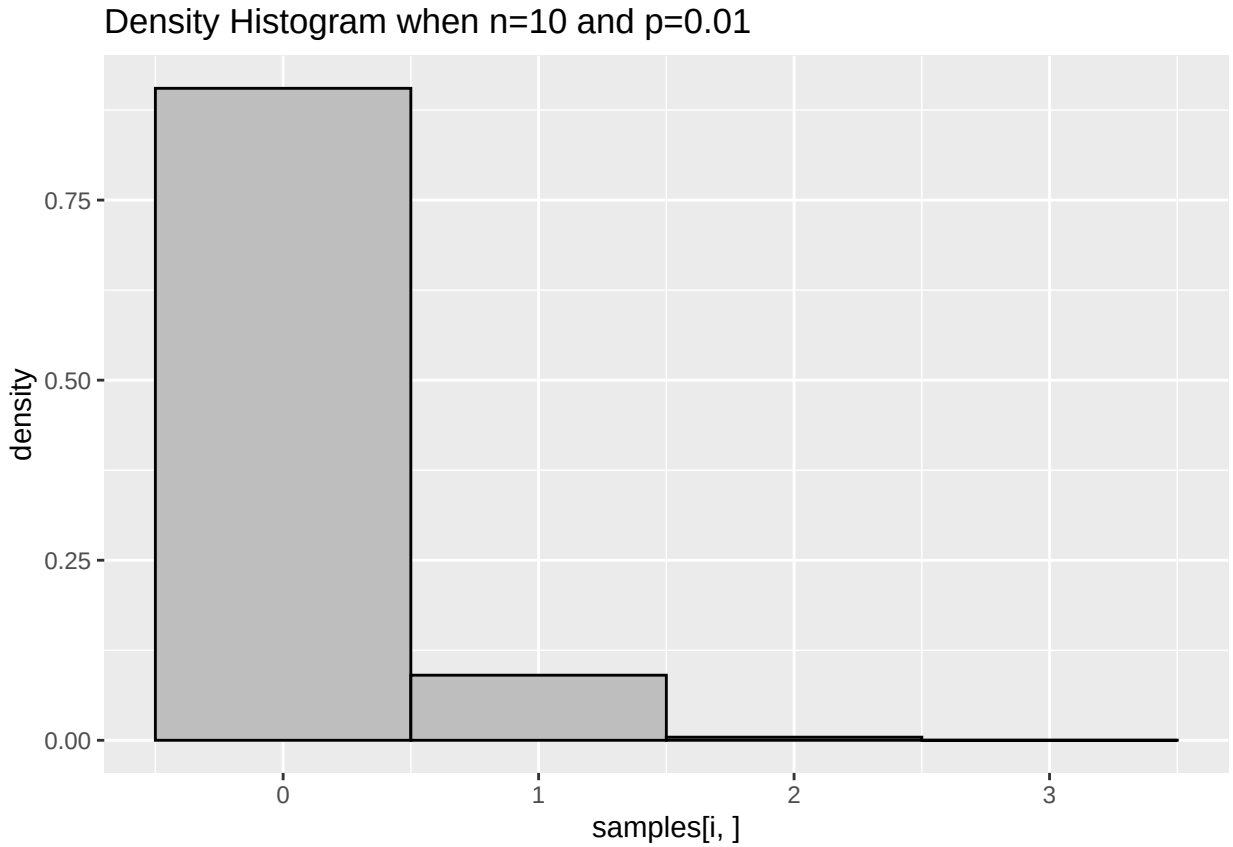
For each n and p in question 5, plot the density histogram of the sample drawn in question 5. If the interquartile range is non-zero, superimpose the density curve of the Normal distribution with the same mean as the sample and q equal to the interquartile range of the sample. You may use the theoretical relationship derived in Question 3 or the observed relationship in Question 3 to find the Normal distribution with the required value of q. (10 points)

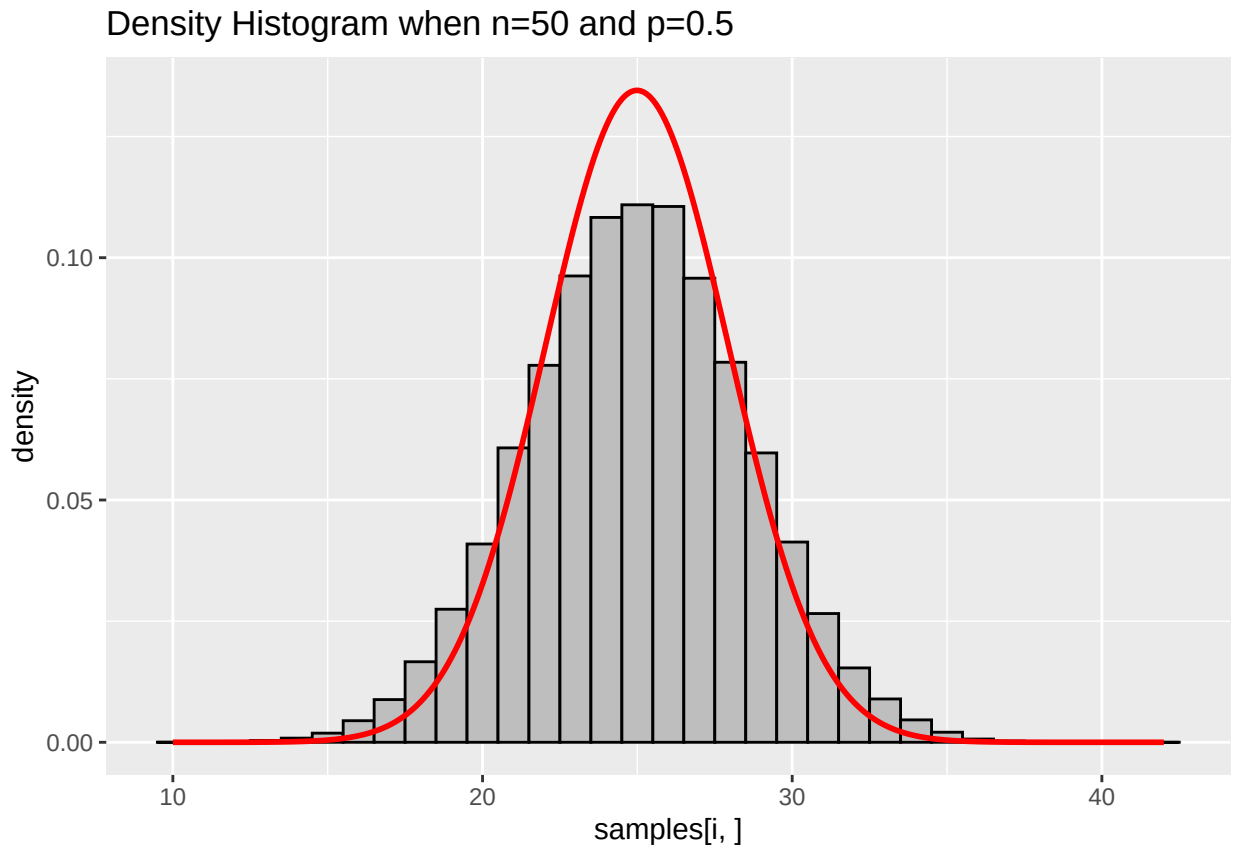
```
# Loop through all 9 pairs to plot their results.
for (i in 1:nrow(samples)){
    # Pull out important varaibles for use later
    n = dat.samp$ns[i]
    p = dat.samp$ps[i]
    iqr = dat.samp$IQR[i]
    mean = dat.samp$mean[i]
    range = seq(min(samples[i, ]), max(samples[i, ]), 0.1)
    # q = 2x_2 sigma + 2mu
    # sigma = (q - 2mu) / (2x_2)
    std.norm.quantile = qnorm(0.75)
    sigma = iqr / (2 * std.norm.quantile)
    dist.theo = dnorm(range, mean=mean, sd=sigma)
    # Generate the title of the dataframe
    title = sprintf("Density Histogram when n=%s and p=%s", n, p)

    # Create the histogram
    g <- ggplot() + aes(x=samples[i, ], y=..density..) +
            geom_histogram(binwidth=1, fill="grey", color="black") +
            labs(title=title)
    # Add the density curve if IQR is not 0.
    if (dat.samp$IQR[i] != 0){
        g <- g + geom_line(aes(x=range, y=dist.theo), size=1, color="red")
    }
```

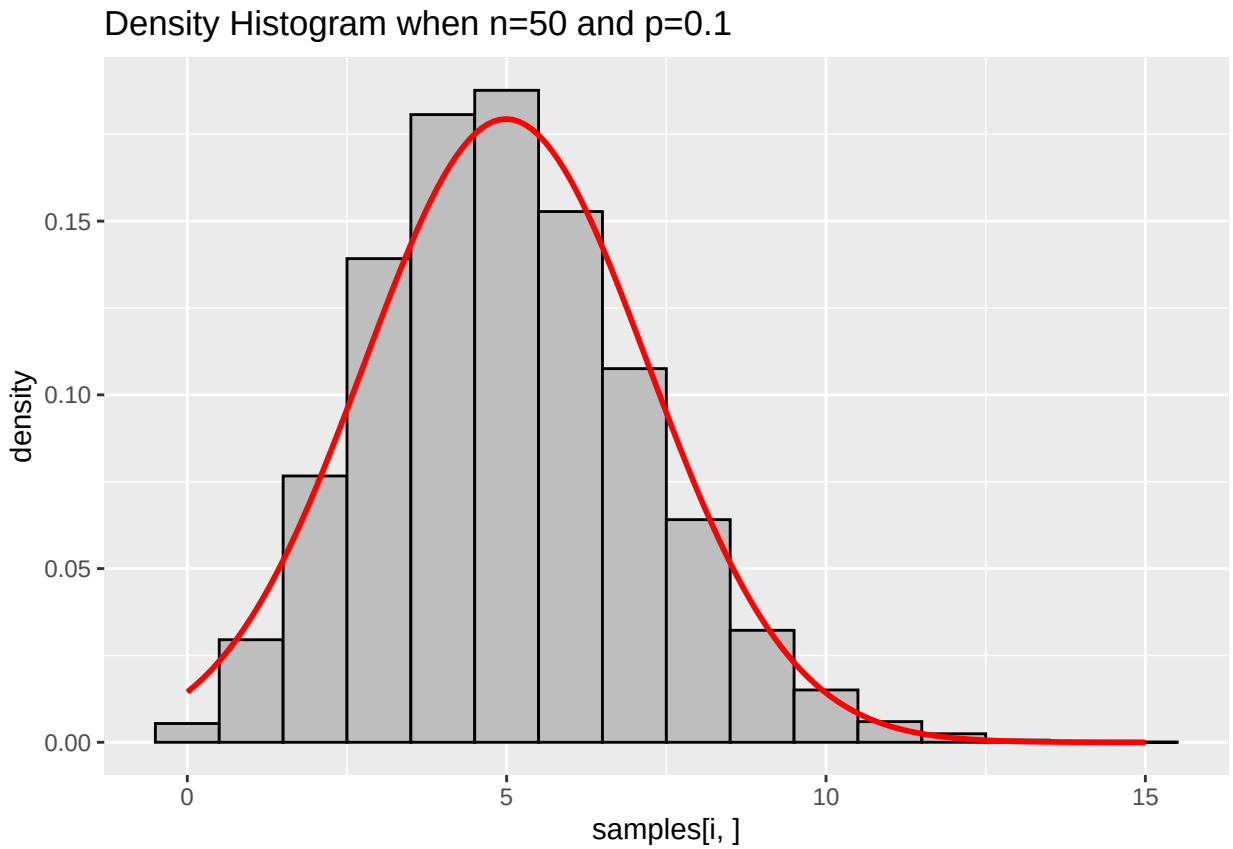
```
    print(g)
}
```

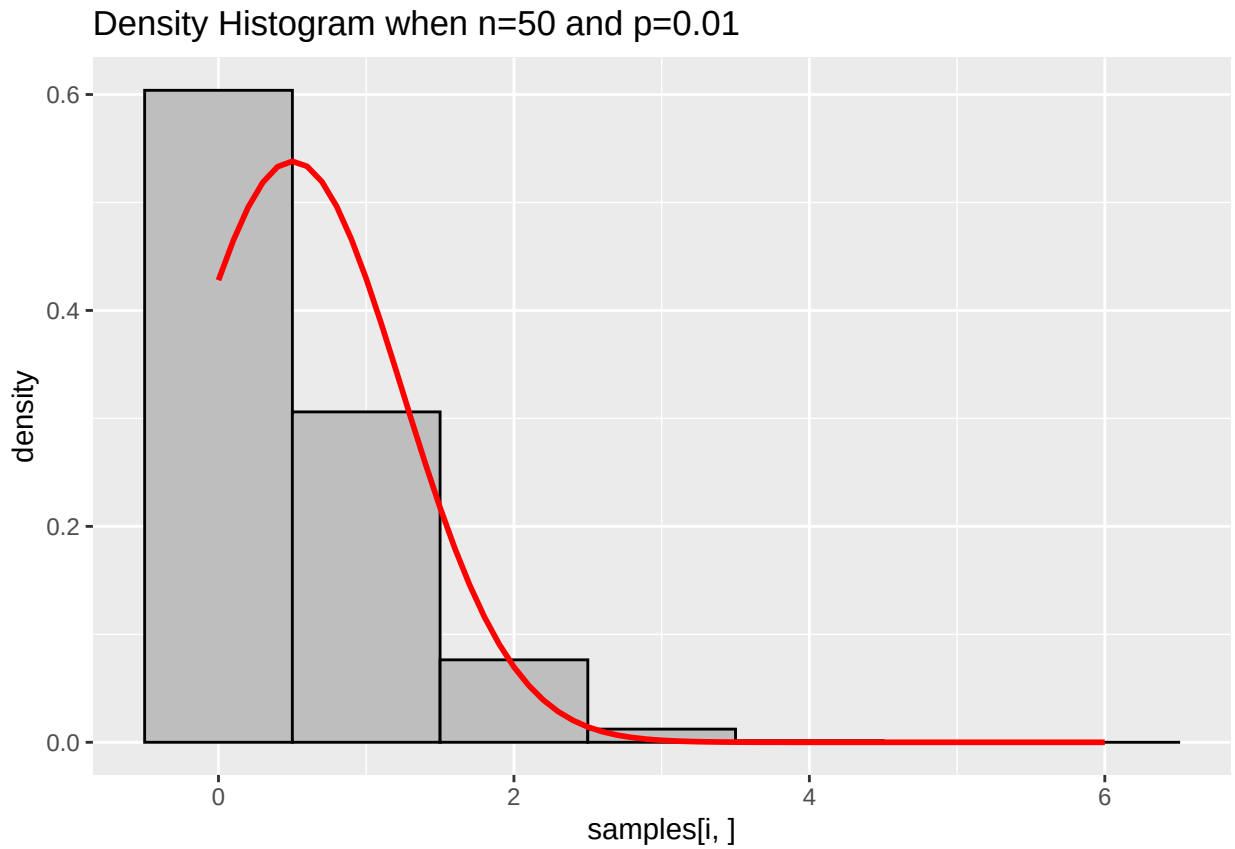

Density Histogram when n=10 and p=0.5
density
0.0
0.1
0.2
0.0
2.5
5.0
7.5
10.0
samples[i,]

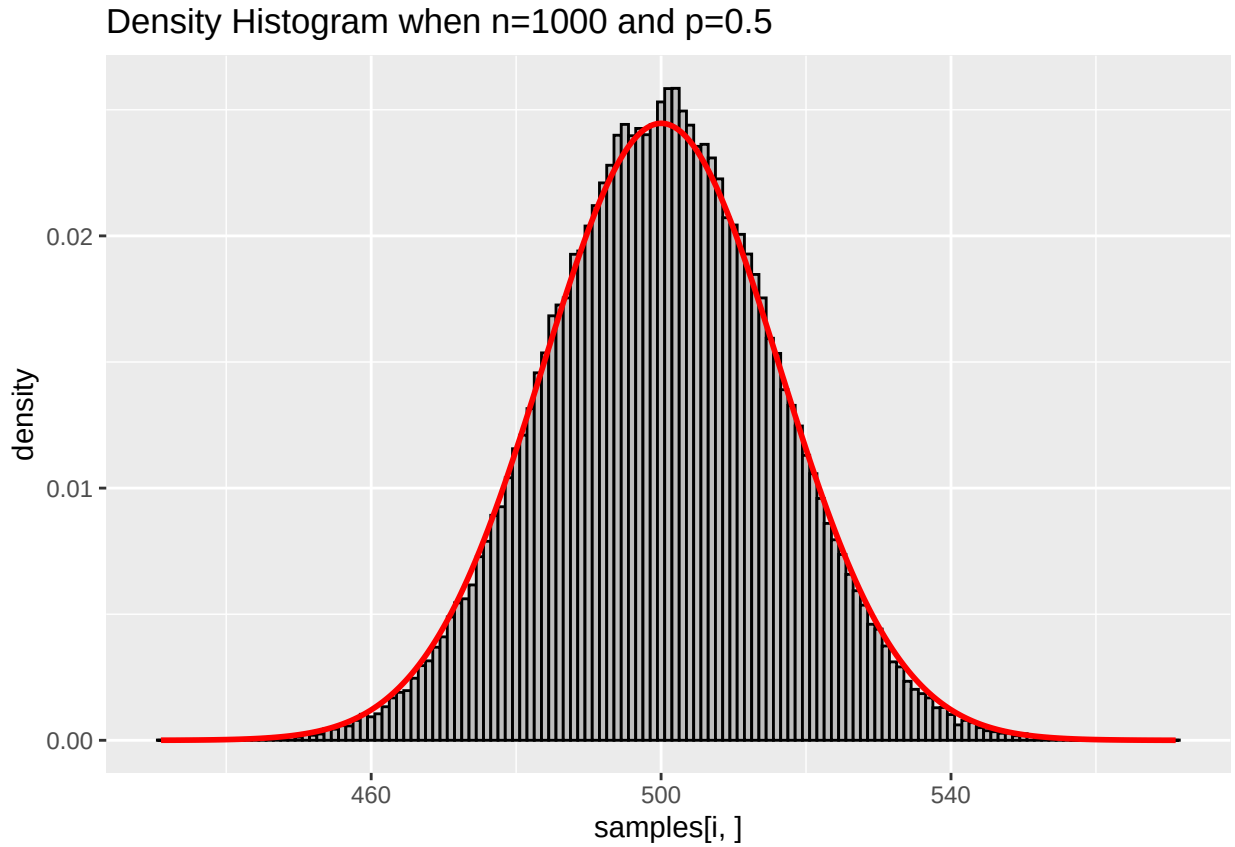

Density Histogram when n=10 and p=0.1
density
0.4
0.3
0.2
0.1
0.0
0
2
4
6
samples[i,]

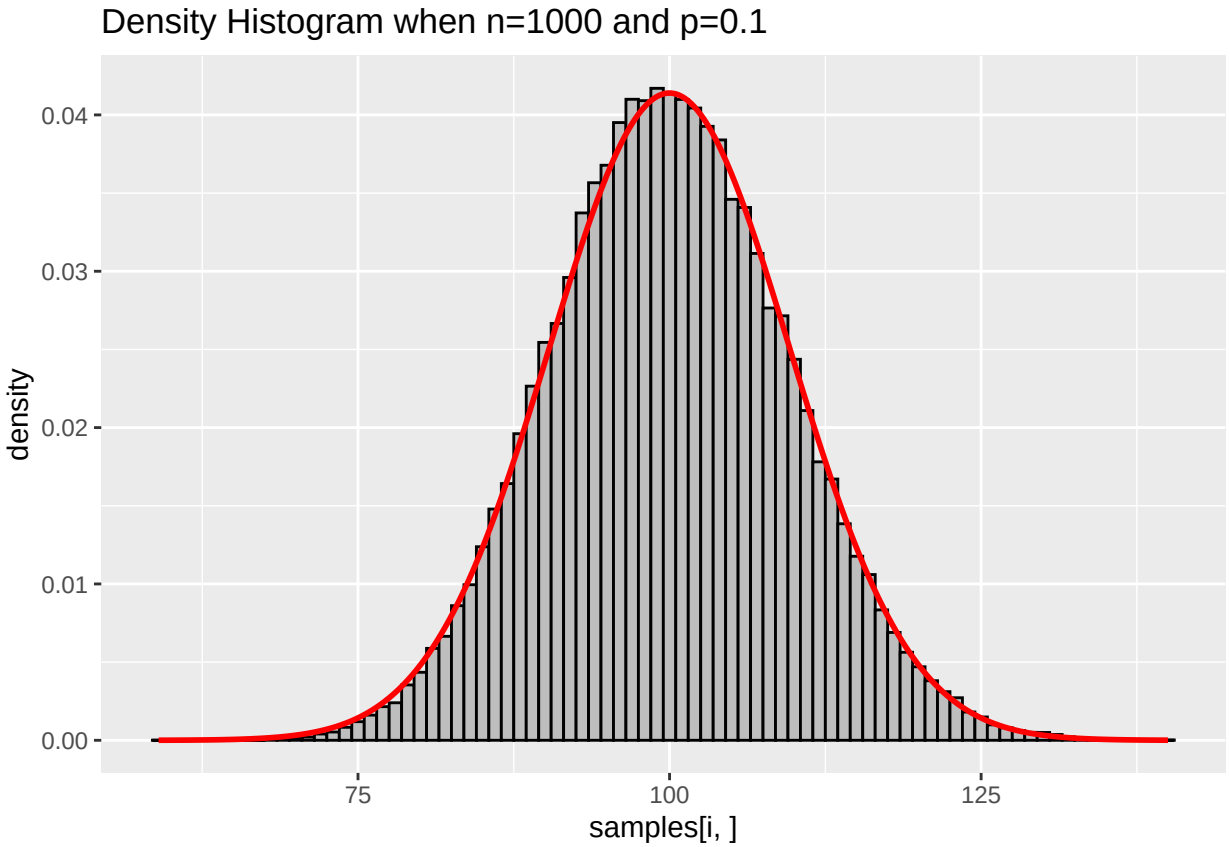

Density Histogram when n=10 and p=0.01
density
0.75
0.50
0.25
0.00
0
1
2
3
samples[i,]

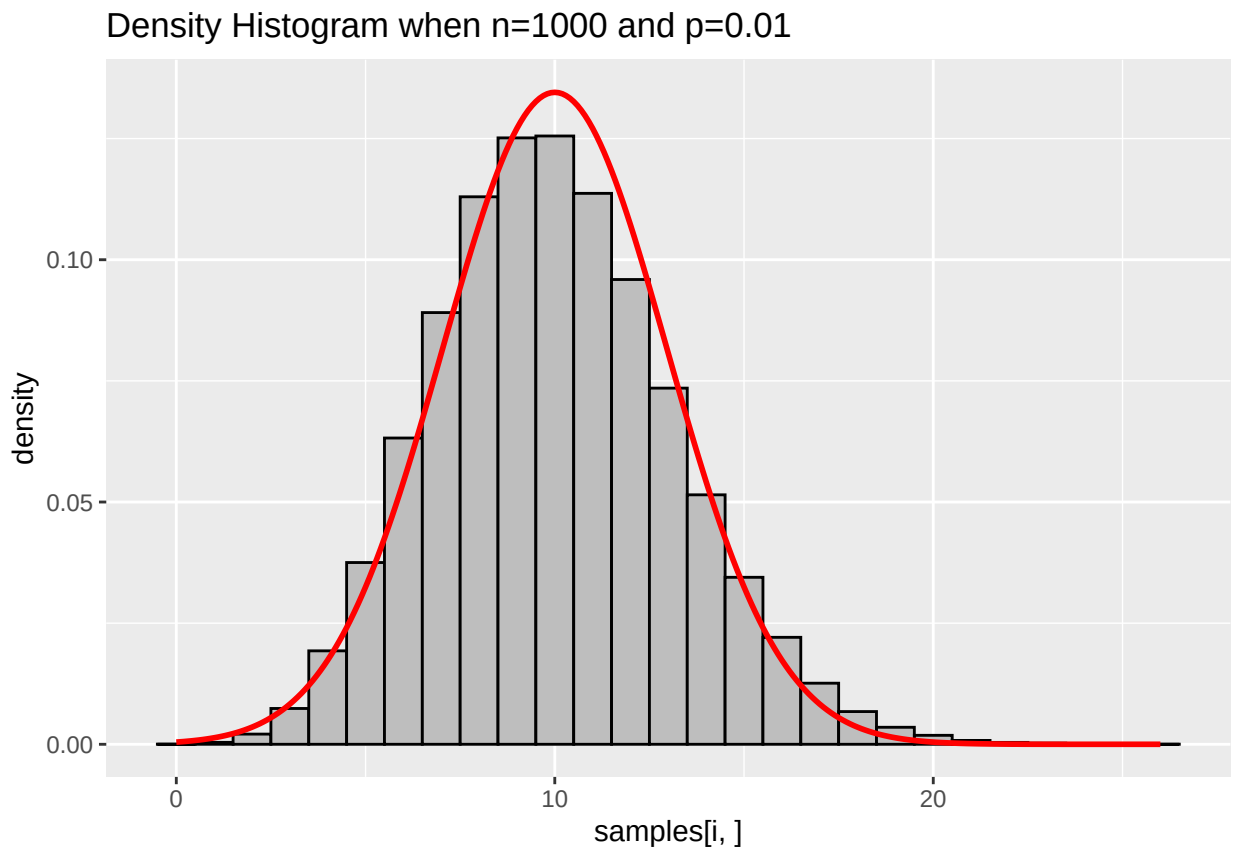

Density Histogram when n=50 and p=0.5
density
0.10
0.05
0.00
10
20
30
40
samples[i,]


Density Histogram when n=50 and p=0.1
density
0.15
0.10
0.05
0.00
0
5
10
15
samples[i,]


Density Histogram when n=50 and p=0.01
density
0.6
0.4
0.2
0.0
0
2
4
6
samples[i,]


Density Histogram when n=1000 and p=0.5
density
0.02
0.01
0.00
460
500
540
samples[i,]


Density Histogram when n=1000 and p=0.1
density
0.04
0.03
0.02
0.01
0.00
75
100
125
samples[i,]



Question 7

Based on the plots in question 6, how does the value of p affect the extent to which the sample distributions are approximately Normally distributed? How does the number n of independent Bernoulli trials affect the extent to which the sample distributions are approximately Normally distributed? (5 points)

To determine the effects of p on the approximate distribution, we should hold n constant. Let's look at the three plots where $n = 10$. When $p = 0.5$, the sampled data appears to follow the normal distribution. When we send the probability towards the extremes, we observe that the sample follows the normal to a less and less degree. This pattern is repeated with the other sets of plots when $n = 50$ and $n = 1000$. Therefor we can conclude that having a "balance" value of p (i.e. near 0.5) will lead to a more normal distribution than a more extreme version of p.

To determine the effects of n on the approximate distribution, we should hold the probability constant. Let's look at the three plots where $p = 0.1$. We can see that when $n = 10$, the sampled data looks very different than the theoretical normal and as we increase n, the samples begin to tend towards the theoretical distribution. This trend is repeated with the sets of plots when $p = 0.5$ and $p = 0.01$. This means that as n increases, the overall distribution becomes closer to matching the normal approximation.