Problem Set 2

C. Durso

February 04, 2019

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 work should be based on the data's being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots.

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

```
## Warning: package 'tidyverse' was built under R version 3.4.4
```

```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.2.0
```

```
## Warning: package 'stringr' was built under R version 3.4.4
```

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

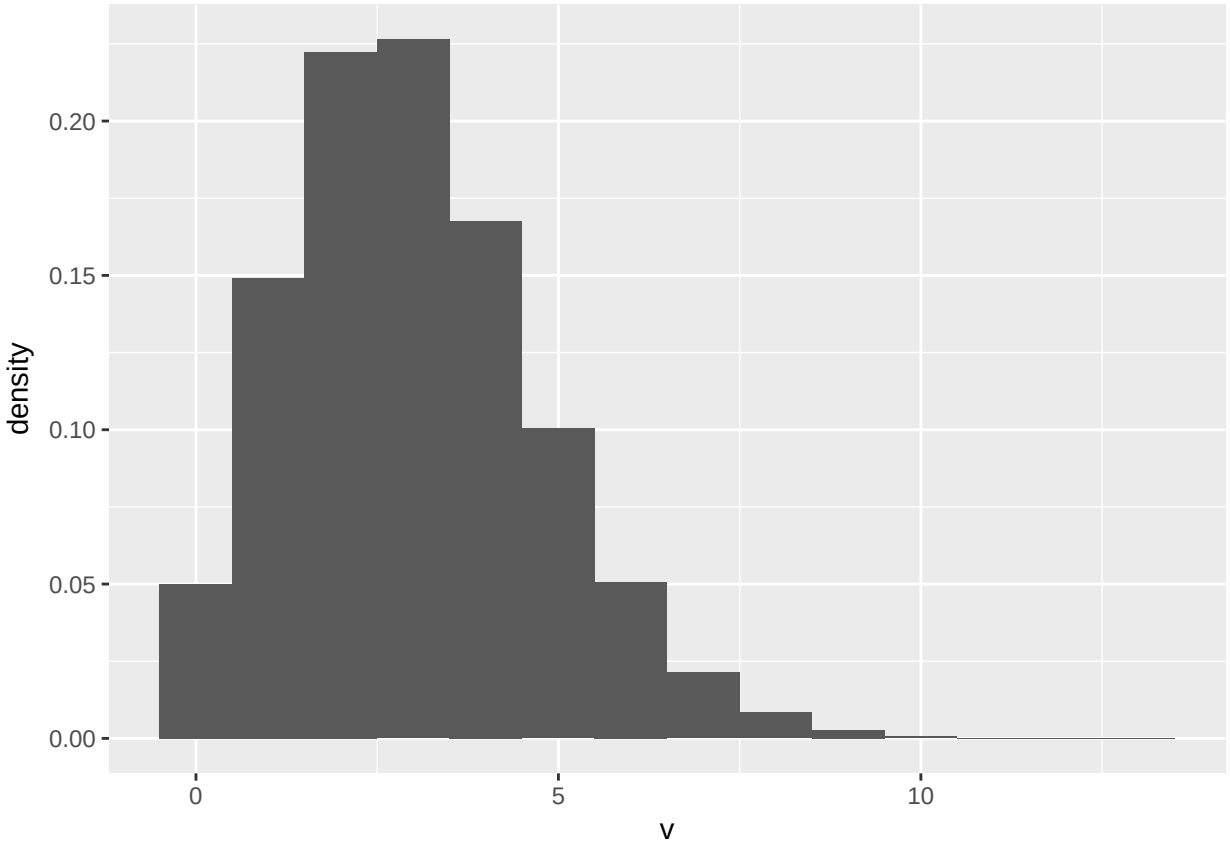
```
library(ggplot2)
library(knitr)
#Used to calculate column quantiles in question 4
library(matrixStats)
```

```
##
## Attaching package: 'matrixStats'
```

```
## The following object is masked from 'package:dplyr':
##
##     count
```

1. Please draw a random sample of 100,000 values from the Poisson distribution with $\lambda = 3$ using the random seed 7654321. Present a histogram of the results with a density scale using the techniques in distributions.Rmd. You may find a bin width of 1 helpful. (5 points)

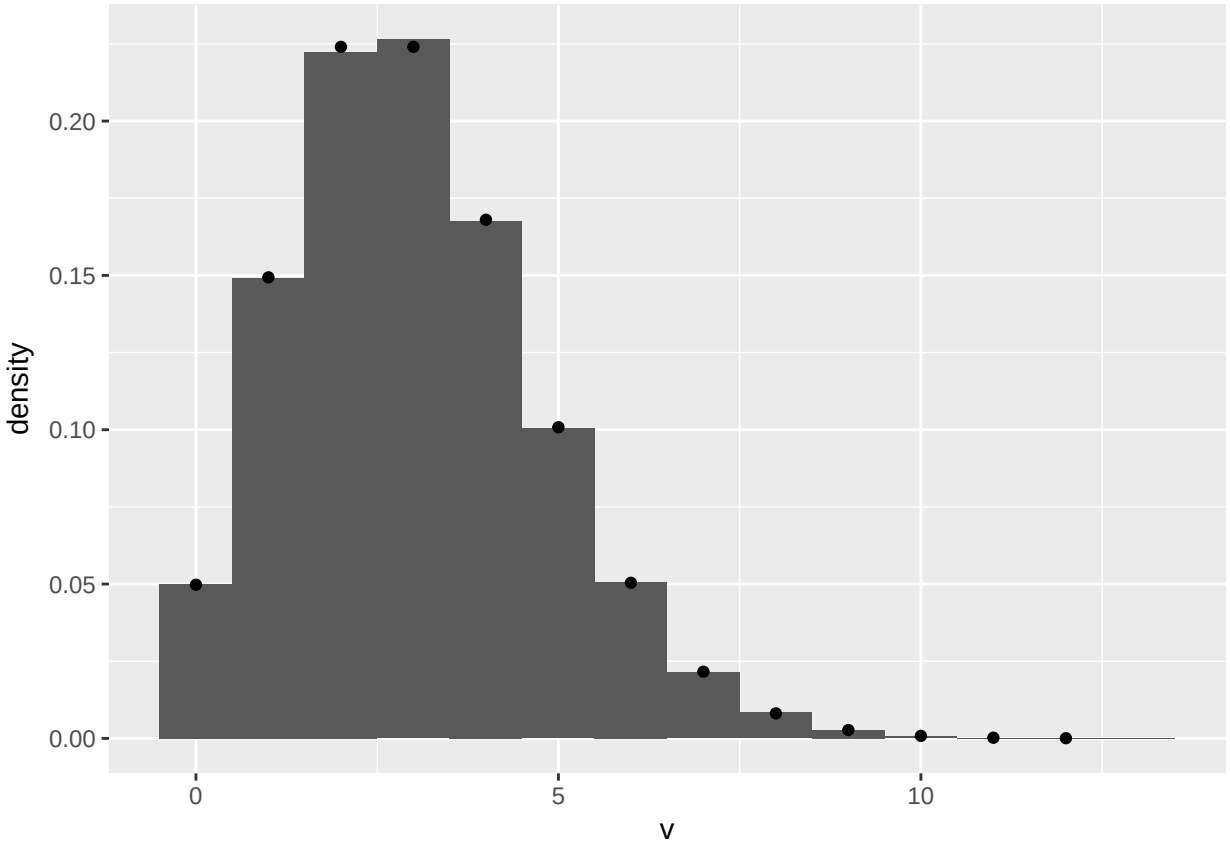
```
set.seed(7654321)
v<-rpois(100000,3)
dat<-data.frame(v=v)
g<-ggplot(dat,aes(x=v)) + geom_histogram(aes(y=..density..),binwidth=1)
g
```



2. Discuss the appearance of the histogram in relation to the probability density function of the Poisson distribution with $\lambda = 3$. (5 points)

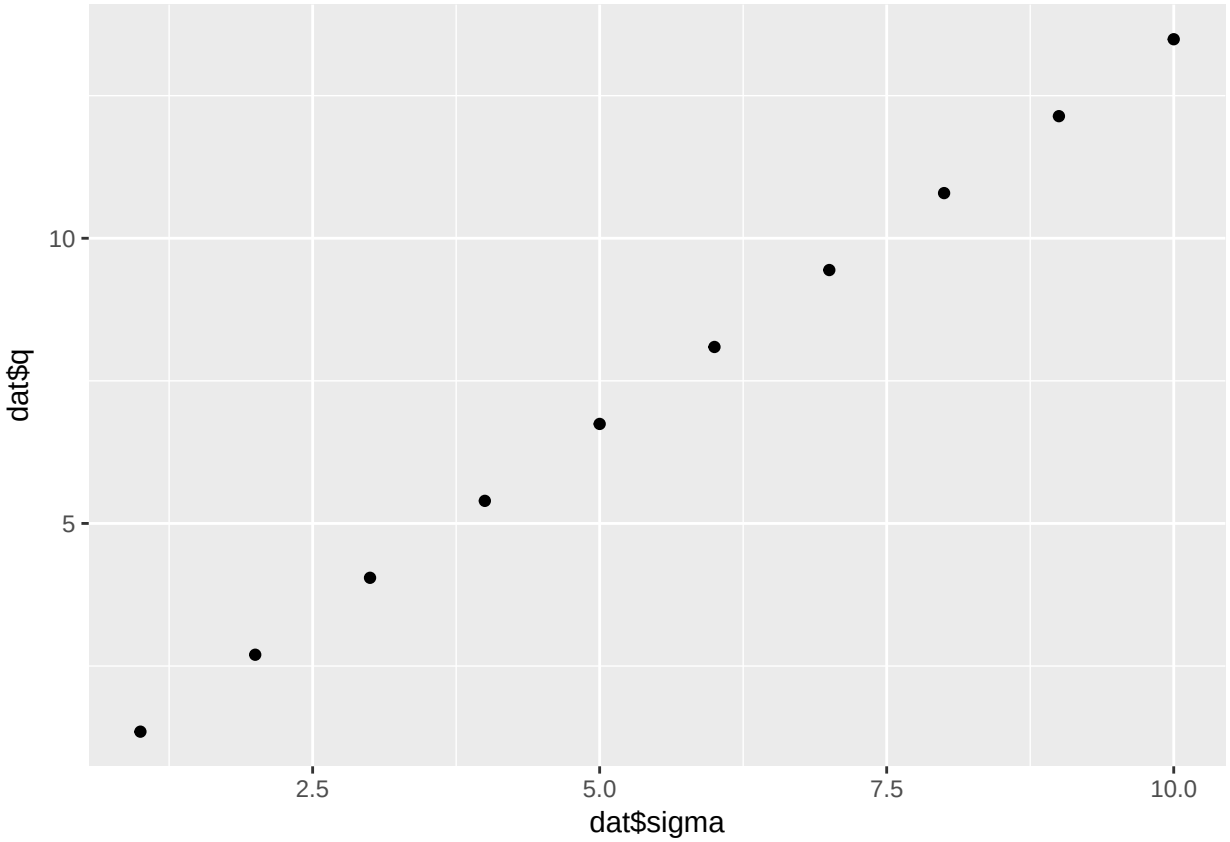
The observed frequencies match the theoretical probabilities pretty well. This is reasonable from the viewpoint that the value $f(k)$ is the proportion of times a random experiment modeled by Poisson(4) will produce k in the long run.

```
theoretical<-data.frame(x=0:12,y=dpois(0:12,3))
g<-g+geom_point(data=theoretical,aes(x=x,y=y))
g
```



3. The length of the interquartile range of data is an often used measure of its spread. By analogy, for a random variable with cumulative density function F, let x_1 satisfy $F(x_1) = 0.25$. Let x_2 satisfy $F(x_2) = 0.75$. Define $q = x_2 - x_1$. Please calculate the values of q 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 q. (5 points)

```
sigmas<-1:10
x1<-qnorm(.25,sd=1:10)
x2<-qnorm(.75,sd=1:10)
dat<-data.frame(sigma=sigmas,q=x2-x1)
qplot(dat$sigma,dat$q)
```



4. Use integration of the density function with a change of variable to calculate the function relating the value of sd and q. (10 points)

For a normal distribution centered at 0 we have the relationship between the inverses of the CDF : $\Phi^{-1}(0.25) = -\Phi^{-1}(0.75)$ which is $x_1 = -x_2$ (1)

We also have the relationship (assuming an sd = 1) : $0.25 = \int_{-\infty}^{x1} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$ (2)

In the general case we have: $0.25 = \int_{-\infty}^{w1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx$

Once again by making the following change of variable : $u = \frac{x}{\sigma}, x = \sigma u \; du = \frac{1}{\sigma} dx$

We rewrite our previous expression as : $0.25 = \int_{-\infty}^{\frac{w1}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$ (3)

We obtain the same expression in (3) as we did in (2). Thus $x_1 = \frac{w_1}{\sigma}$, or $\sigma x_1 = w_1$, demonstrating that σ and w_1 are linearly related. Since the IQR is $2w_1$, this completes the argument.

Another possibility would be to look at the IQR (interquartile range (1)) and write the integral (assuming sd=1): $0.5 = \int_{-x_q}^{x_q} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx$ (4)

In the general case we have : $0.5 = \int_{-w_1}^{w_1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}} dx$ (5)

By a change of variable of the form : $u = \frac{x}{\sigma}, du = \frac{1}{\sigma} dx \rightarrow dx = \sigma du$

We rewrite our previous expression as : $0.5 = \int_{-\frac{w_1}{\sigma}}^{\frac{w_1}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$ (7)

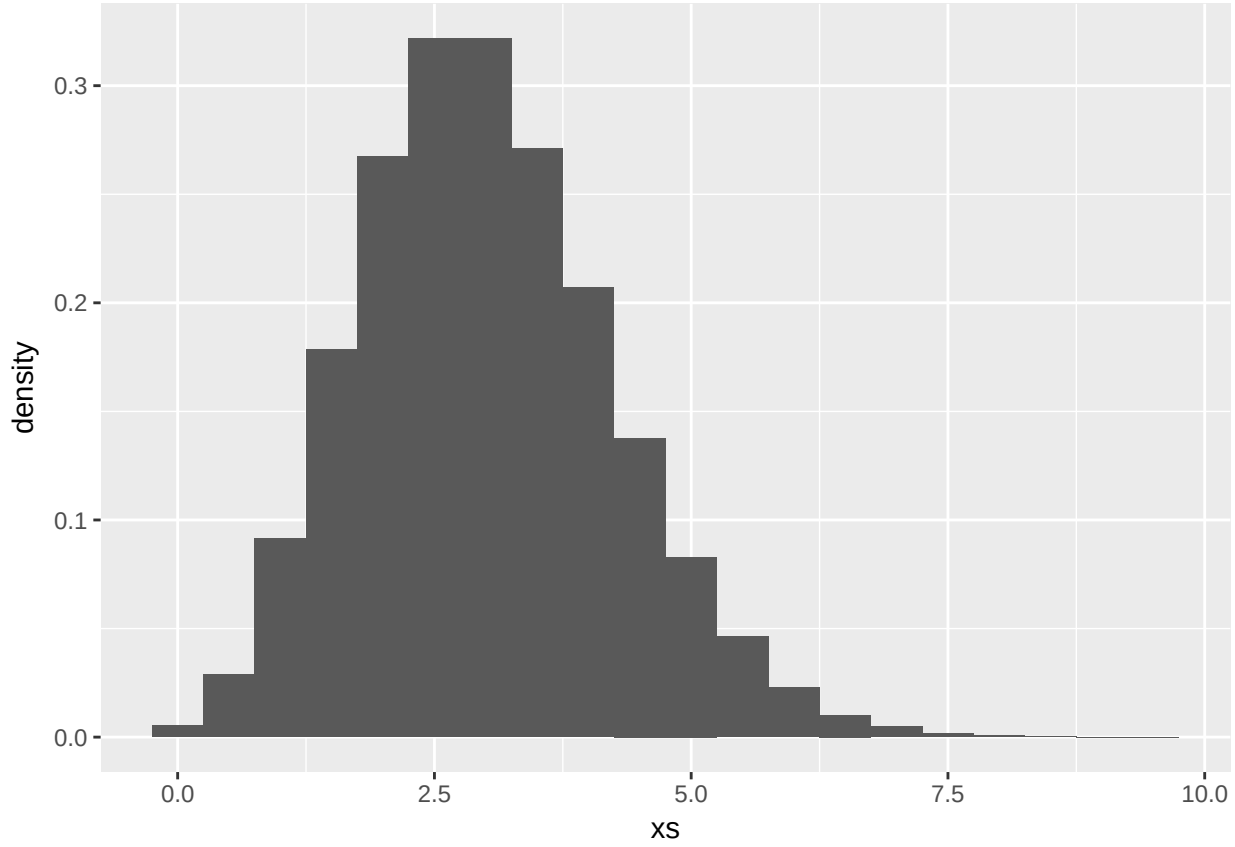
We obtain the same result in (7) as we did in (4).

5. Consider the random variable that is the mean of 2 independent draws from the Poisson distribution with $\lambda = 3$. Construct a vector of 100,000 samples from this random variable. Present a histogram of the results with a density scale using the techniques in distributions.Rmd. You may find a bin width of 0.5 helpful. (5 points)

```
xs<-rpois(200000,3)
xmat<-matrix(xs,ncol=2)
xs<-apply(xmat,1,mean)

dat.2<-data.frame(xs=xs)

g<-ggplot(dat.2,aes(x=xs))+ geom_histogram(aes(y=..density..),binwidth=.5)
g
```



6. Find the mean and interquartile range of your vector from question 5. Plot the histogram from question 5 with the density curve of the Normal distribution with that mean and q equal to the interquartile range of the data. (10 points)

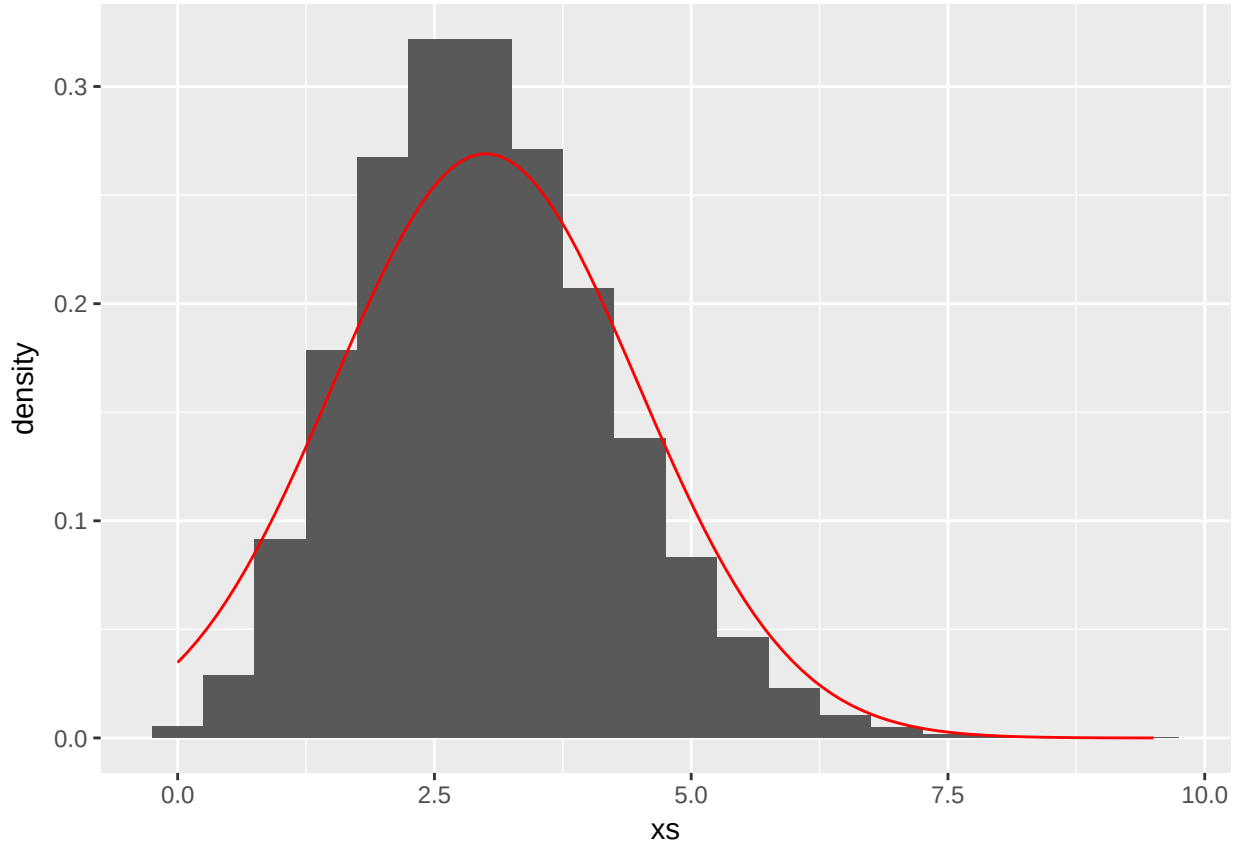
```
m<-mean(xs)
q<-quantile(xs,c(.25,.75))
q<-q[2]-q[1]


slope=qnorm(.75)-qnorm(.25)
stdev<-q/slope

qnorm(.75, sd=stdev)-qnorm(.25, sd=stdev)
```

```
## [1] 2
```

```
g <- ggplot(dat.2, aes(x=xs)) +
        geom_histogram(aes(y = ..density..),binwidth=.5) +
        stat_function(fun = dnorm, colour = "red",args = list(mean =m, sd = stdev))
g
```



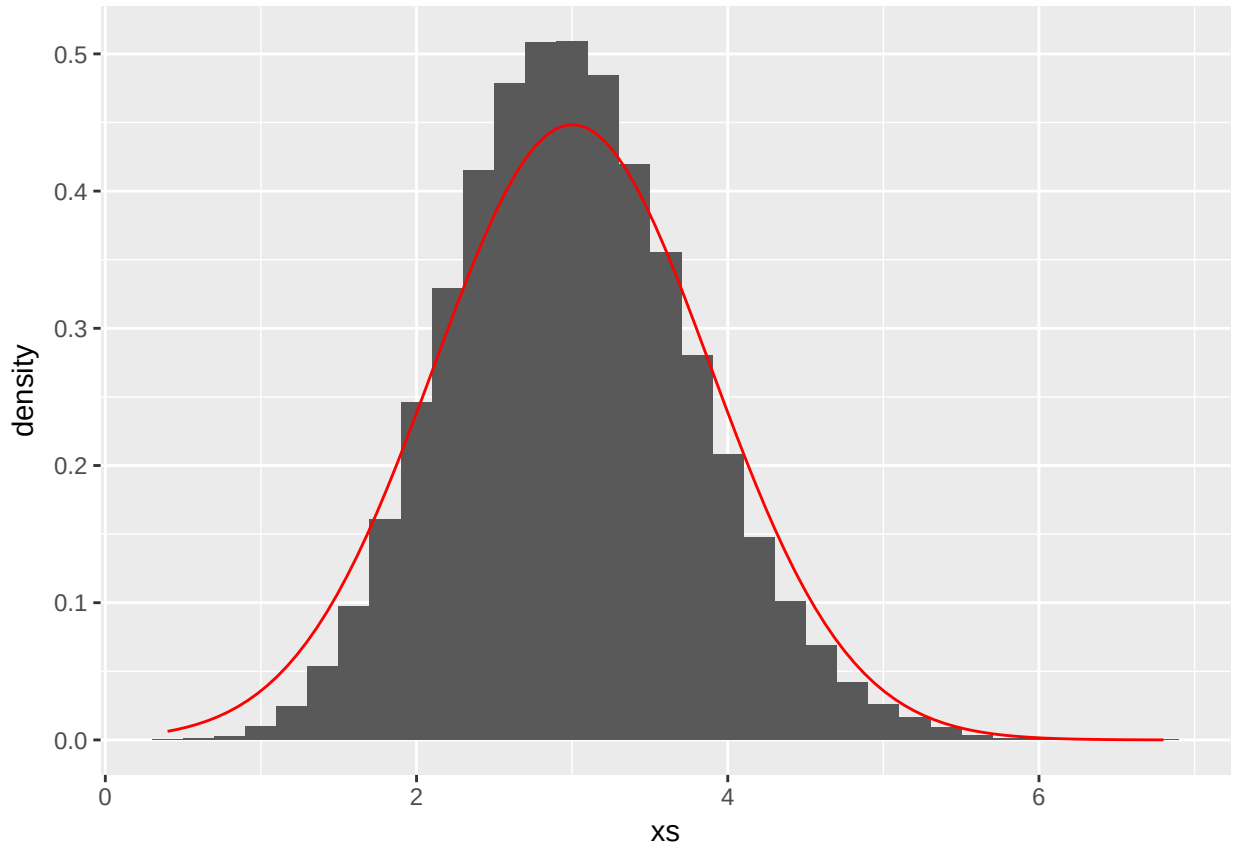
7. Repeat questions 5 and 6 for means of 5 draws amd means of 20 draws. Please comment on the relation of the Normal distribution to the density histogram for means of 2, 5, and 20 draws.(10 points)

```
xs<-rpois(500000,3)
xmat<-matrix(xs,ncol=5)
xs<-apply(xmat,1,mean)

dat.5<-data.frame(xs=xs)
m<-mean(xs)
q<-quantile(xs,c(.25,.75))
q<-q[2]-q[1]


slope=qnorm(.75)-qnorm(.25)
stdev<-q/slope

g.5<- ggplot(dat.5, aes(x=xs)) +
        geom_histogram(aes(y = ..density..),binwidth=.2) +
        stat_function(fun = dnorm, colour = "red",args = list(mean =m, sd = stdev))
g.5
```

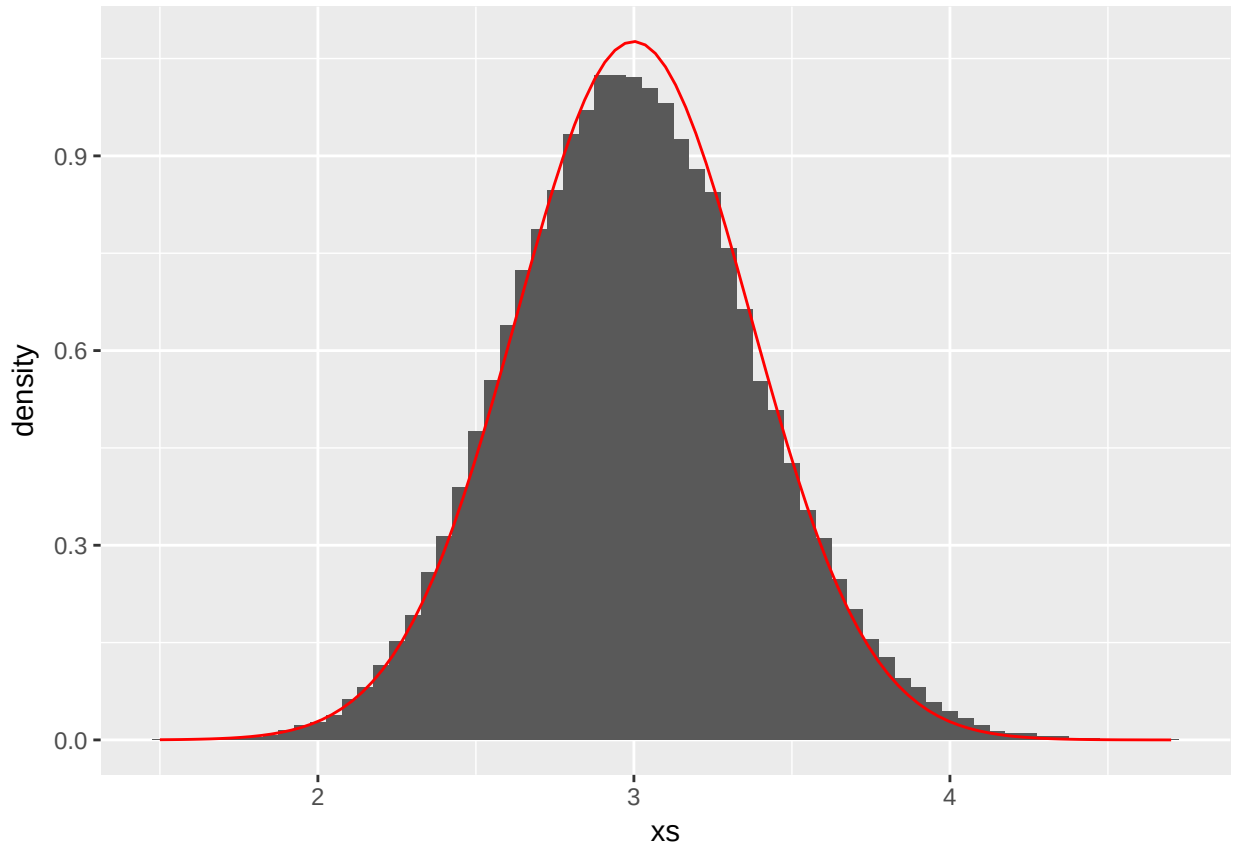



```
xs<-rpois(2000000,3)
xmat<-matrix(xs,ncol=20)
xs<-apply(xmat,1,mean)

dat.20<-data.frame(xs=xs)
m<-mean(xs)
q<-quantile(xs,c(.25,.75))
q<-q[2]-q[1]


slope=qnorm(.75)-qnorm(.25)
stdev<-q/slope

g.20<- ggplot(dat.20, aes(x=xs)) +
        geom_histogram(aes(y = ..density..),binwidth=0.05) +
        stat_function(fun = dnorm, colour = "red",args = list(mean =m, sd = stdev))
g.20
```



The Normal approximation seems to improve as the number of observation contributing to the mean increases.