Problem Set 2

Megan Hoeksema

These questions were rendered in R markdown through RStudio (<<https://urldefense.com/v3/__https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf__>;!!NCZxaNi9jForCP\_SxBKJCA!Ca08b5NQZfvAu6v0RzkstRTPrXDXwDinjNj0uVYwQrXJdYd24oveKvHxwiWG6loMfYWG$ >, <<https://urldefense.com/v3/__http://rmarkdown.rstudio.com__>;!!NCZxaNi9jForCP\_SxBKJCA!Ca08b5NQZfvAu6v0RzkstRTPrXDXwDinjNj0uVYwQrXJdYd24oveKvHxwiWG6ms\_4QRX$ > ).

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)

## Warning: package 'tidyverse' was built under R version 3.6.3

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.1 v dplyr 1.0.6  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'purrr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(knitr)

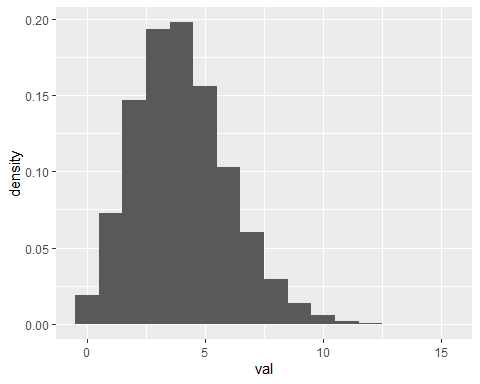
## Large and Small Samples

The relationship between a sample and the population from which it is drawn depends on the size of the sample. This feature underlies the need for sufficiently large samples in order to run meaningful statistical analyses.

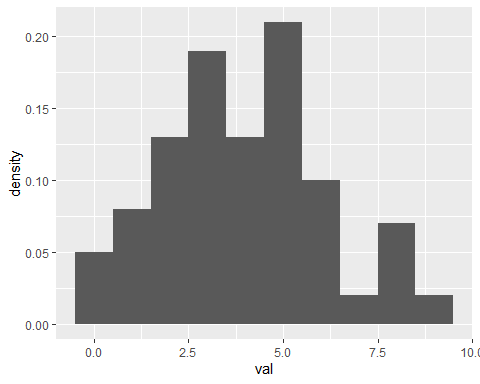
### Question 1

Please draw a random sample of 100,000 values from the Poisson distribution with 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 using the random seed 7654321 and present a histogram of the results with a density scale.(5 points)

set.seed(7654321)  
samp <- rpois(100000,4)   
dat<-data.frame((val=samp))  
myplot<-ggplot(data=dat,aes(x=val))+geom\_histogram(aes(y=stat(density)), binwidth = 1)  
myplot



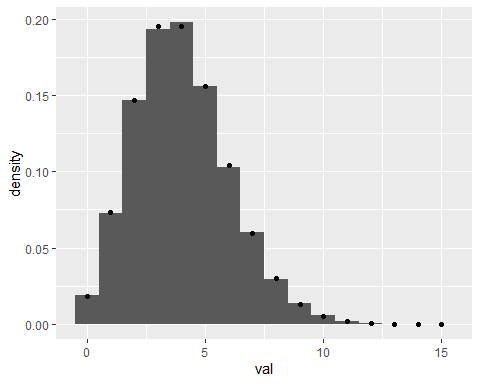
set.seed(7654321)  
samp.smaller<-rpois(100, 4)  
dat.smaller<-data.frame(val=samp.smaller)  
myplot.smaller<-ggplot(data=dat.smaller, aes(x=val)) + geom\_histogram(aes(y=stat(density)),binwidth = 1)  
myplot.smaller



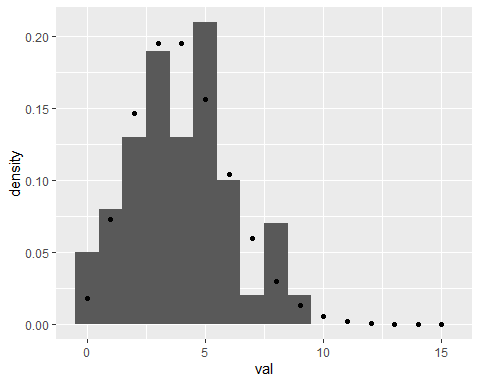
### Question 2

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 . Repeat for the sample of size 100. For which sample size are the proportions of each outcome more similar to the probabilities given by the density function of the Poisson distribution with ? (5 points)

in.theory<-data.frame(x=0:max(samp))  
in.theory$y<-dpois(in.theory$x,lambda = 4)  
  
my.plot.better<-myplot+geom\_point(data=in.theory,aes(x=x,y=y))  
my.plot.better



my.plot.smaller.better<-myplot.smaller+geom\_point(data=in.theory,aes(x=x,y=y))  
my.plot.smaller.better



## Normal Approximations

Many statistical methods involve approximation of a distribution by a Normal distribution. Questions 3 through 5 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 value of “mean” and the value of “sd” to use in a Normal approximation will be handled by use of the interquartile range, defined below. (There are other approaches that that depend on sample mean and variance and the mean and variance of the Normal distribution, concepts covered later in the course.)

### Question 3

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

The **first quartile** of is the value .

The **third quartile** of is the value .

The **median** of is the value .

The **interquartile range** of is the value .

Given a data set with known values for the median and the interquartile range, one can calculate values of and such that the Normal distribution has the same median and interquartile range. This gives a way to select a Normal distribution approximating the data set.

The results of this exercise will be used in the next exercise identify Normal distributions that are reasonable approximations to simulated data sets.

#### Question 3.a

By analogy, for a random variable with cumulative density function , let satisfy .

The **first quartile** of the random variable is the value .

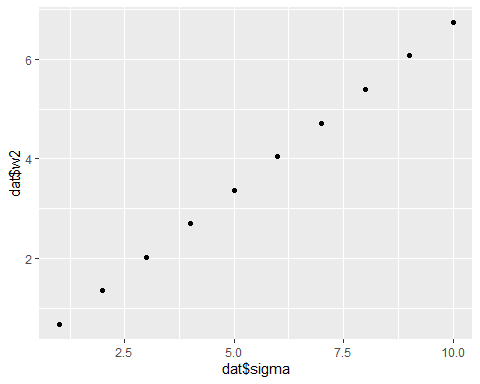
The **third quartile** of the random variable is the value .

The **median** of the random variable is the value .

The **interquartile range** of the random variable is the value .

Please calculate the values of for the Normal distributions with mean 0 and sd in and plot the points consisting of the value of the sd and the corresponding . This should give an indication of a simple function relating sd and .(5 points)

sigmas<-1:10  
w<-qnorm(0.75, sd=1:10)  
  
dat<-data.frame(sigma=sigmas, w2=qnorm(0.75, sd=1:10))  
qplot(dat$sigma, dat$w2)



qnorm(0.90)

## [1] 1.281552

qnorm(0.1)

## [1] -1.281552

qnorm(0.75, 10)

## [1] 10.67449

#### Question 3.b

Denote by the value such that . Denote by the value such that .

Thus

Note that the change of variable transforms the integrand of the integral on the right hand side of the equality to equal the integrand of the integral on the left hand side. Please use this change of variable applied carefully to the limits of integration to give a formula for in terms of , , and . (5 points)

#### Question 3.c

Please give a formula for in terms of , , and .

Please give a formula for in terms of , , and .

Please give a formula for in terms of .

Please give a formula for in terms of , and .

Please use the results of 3.b to address these questions. (5 points)

#### Question 3.d

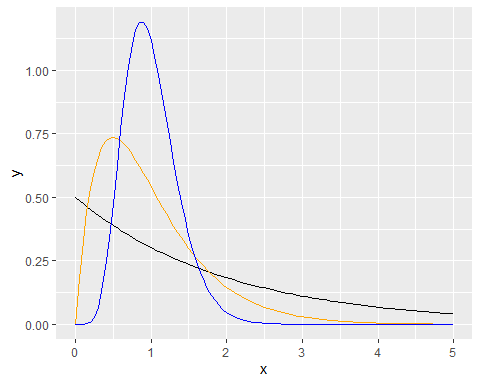
Suppose a data set has median equal to and interquartile range equal to . What values of and result in a random variable with the same median and interquartile range? Please use in your solution. The symmetry of the standard Normal density function implies that . (5 points)

### Question 4

Many methods in statistics assume that data are samples from a Normal distribution. This question and Question 5 investigate how (nearly) Normal distributions can arise from means of samples from non-Normal distributions.

Consider the gamma random variables plotted below.

dat.plot<-data.frame(x=c(0,5))  
ggplot(dat.plot,aes(x=x))+  
 stat\_function(fun=dgamma,args=list(shape=1,scale=2))+  
 stat\_function(fun=dgamma,args=list(shape=2,scale=.5),color="orange")+  
 stat\_function(fun=dgamma,args=list(shape=8,scale=.125),color="blue")



The following code generates 100,000 means of samples of size 2, 5, and 20 from each of the gamma distributions, plots the histograms of each set of 100,000 means, and overlays the histogram with the Normal distribution having the same median and interquartile range as the set of 100,000 means.

Create a data frame of the parameters for each set of 100,000 means.

ns<-rep(c(2,5,20),times=c(3,3,3))  
shapes<-rep(c(1,2,8),times=3)  
scales<-rep(c(2,.5,.125),times=3)  
  
dat.params<-data.frame(ns,shapes,scales)

Create a matrix of “val.mat” of NA’s with 100,000 rows and a column for the means of samples of size=2, shape=1, scale=2; size=5, shape=2, scale=.5;…size=20, shape=8, scale=0.125.

val.mat<-matrix(rep(NA,100000\*nrow(dat.params)),  
 ncol=nrow(dat.params))

Populate the matrix by stepping through the rows of the data frame of parameters for the 100,000 means.

set.seed(123456)  
for(i in 1:nrow(dat.params)){  
 # Generate a matrix with 100,000 rows of n   
 # random values from a gamma distribution where the value of n  
 # and of the shape and scale of the gamma distribution come from the   
 # ith row of dat.params.  
 samp.mat<-matrix(rgamma(dat.params$ns[i]\*100000,dat.params$shapes[i],  
 dat.params$scales[i]),nrow=100000)  
 # take the mean of each row and store the result in val.mat  
 val.mat[,i]=apply(samp.mat,1,mean)  
}

Write a function that takes a numeric vector x as its argument and returns the median and interquartile range of the values in x.

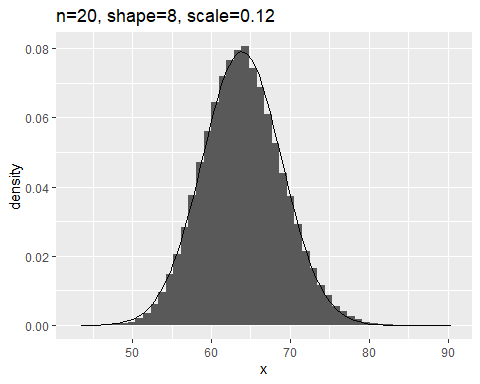
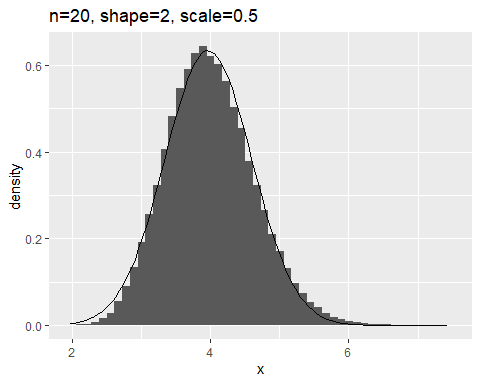
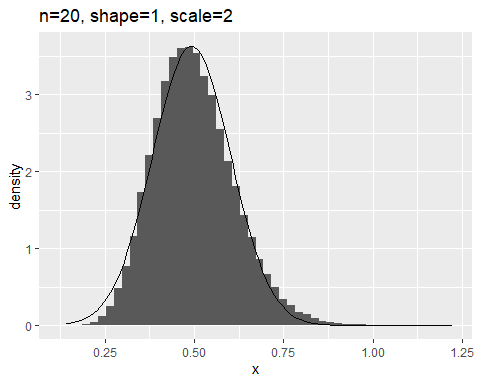
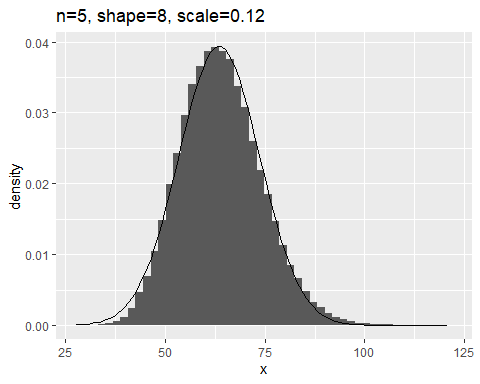
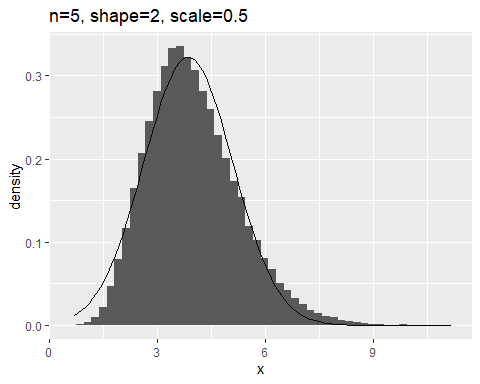
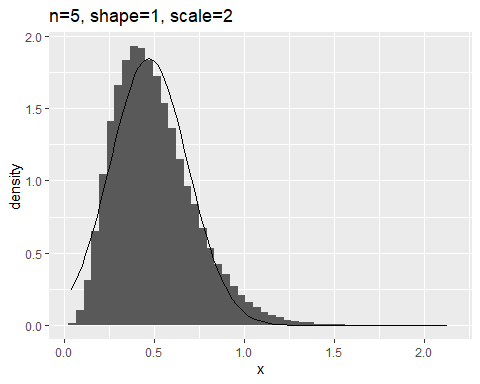
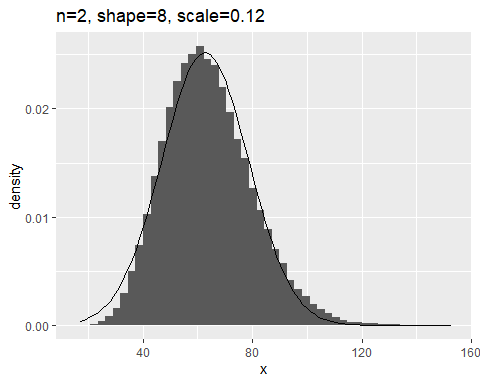
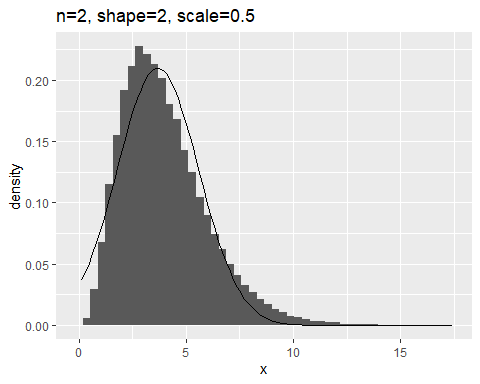
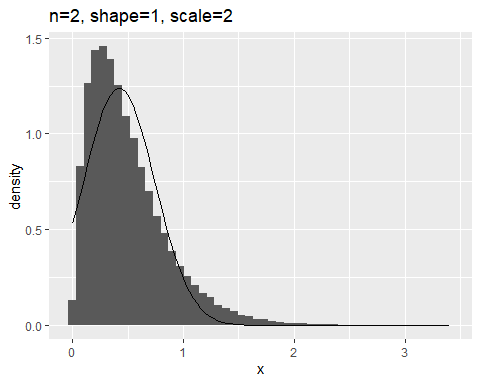
val.stats<-function(x){  
 return(c(median(x),quantile(x,.75)-quantile(x,.25)))  
}

Apply this to the columns of the matrix of means of samples to get a matrix “params” where the columns correspond to the columns of the matrix “val.mat”, the first row is the medians of the corresponding columns, and the second row is the interquartile ranges of the corresponding columns.

params<-apply(val.mat,2,val.stats)

Step through the columns of “val.mat” making a density histogram of the values in the column with the density curve of the Normal distribution having the same median and interquartile range as the values in the column. Note that the median aad interquartile range for the column are in the corresponding column of “params”.

for(i in 1:ncol(val.mat)){  
 dat.temp<-data.frame(x=val.mat[,i])  
 g<-ggplot(dat.temp,aes(x=x))+  
 geom\_histogram(aes(y=stat(density)),bins=50)+  
 labs(title=str\_c("n=",dat.params$ns[i],", shape=",  
 dat.params$shapes[i],  
 ", scale=",round(dat.params$scales[i],2)))+  
 stat\_function(fun=dnorm,args=list(mean=params[1,i],  
 sd=params[2,i]/(2\*qnorm(.75))))  
 print(g)  
}



Please comment on the visual resemblance between the histograms and the corresponding Normal distribution. Qualitatively, how does the degree of resemblance depend on the number of terms in the sample? Qualitatively, how does degree of resemblance depend on distribution of terms in the sample? (5 points)

### MDH Comment: All of the histograms seem to represent a normal distribution however less so for both the n=2, shape = 1 ,scale = 1 and n=2, shape = 2, scale = 0.5. The degree of resemblence is increasingly better as the number of terms in the sample increases. The degree of resemblence improves as the distribution of terms becomes more equivalent.

### Question 5

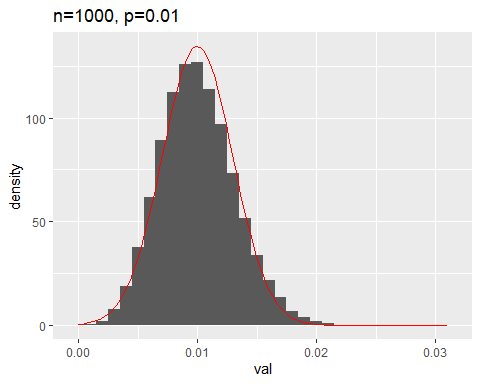
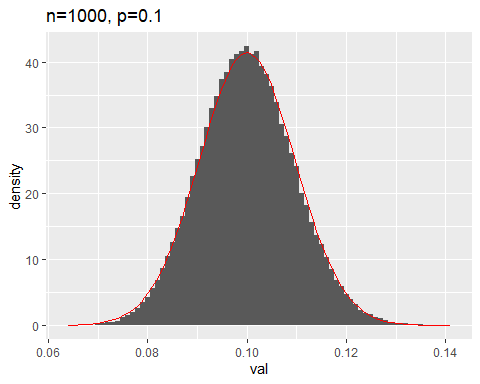
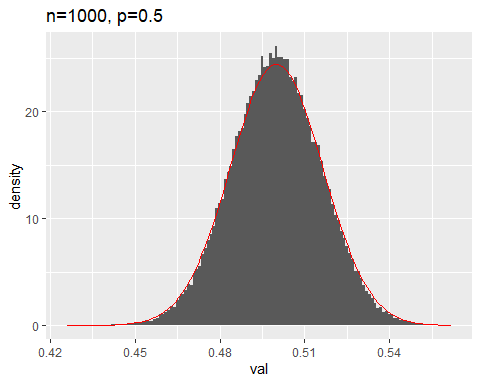
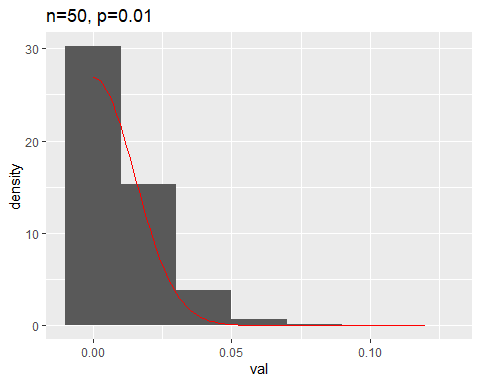
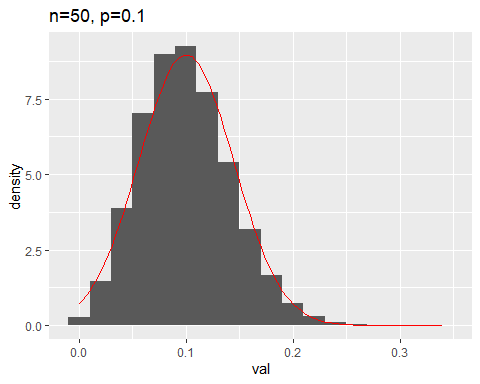
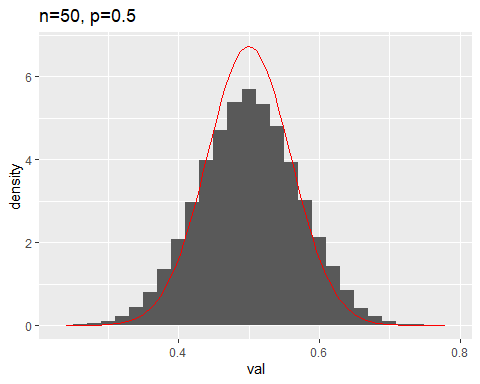
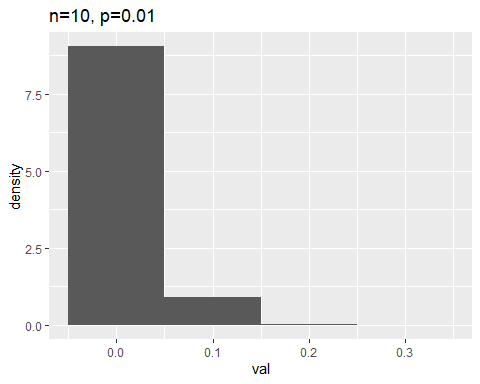
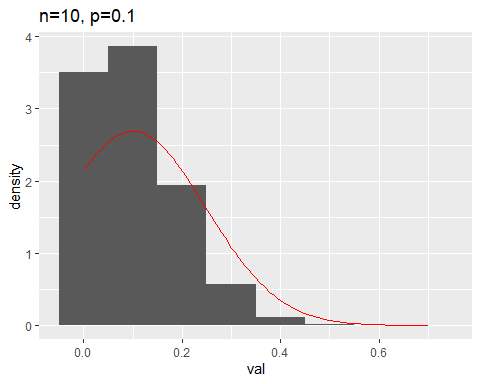
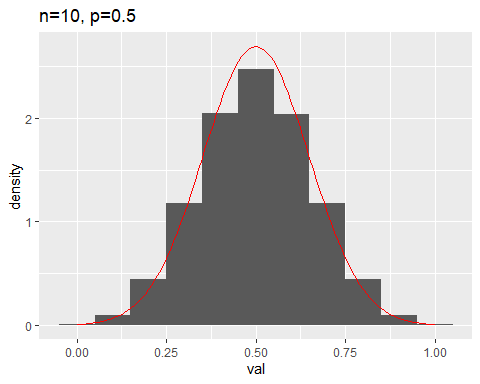
This question revisits the issue of how (nearly) Normal distributions can arise from means of samples from non-Normal distributions.

In particular, this question looks at probability spaces for where , and is defined by the density by and .

#### Question 5.a

For each pair with and , sample 100,000 values of the mean of samples from and plot the density histogram of the 100,000 values. If the interquartile range for the 100,000 values for that is nonzero, also include the density function of the Normal distribution having the same median and interquartile range as the 100,000 values in the plot. You may gain efficiency by relating the means of the samples of values from to the binomial distribution with size and probability of success . Also, a bin width equal to is appropriate because the means take on values (10 points)

ns<-c(10,50,1000)  
ps<-c(.5,.1,.01)  
dat.samp<-data.frame(ns=rep(ns,times=rep(length(ps), length(ns))),ps=rep(ps,times=length(ns)))  
  
set.seed(7899)  
val.mat2<-matrix(rep(NA,100000\*nrow(dat.samp)),ncol=nrow(dat.samp))  
  
for(i in 1:nrow(dat.samp)){  
 val.mat2[,i]<-rbinom(100000,dat.samp$ns[i],dat.samp$ps[i])/dat.samp$ns[i]  
}  
  
m.iqr<-apply(val.mat2,2,val.stats)  
  
for(i in 1:ncol(m.iqr)){  
 dat.temp<-data.frame(val=val.mat2[,i])  
   
 g <- ggplot(dat.temp,aes(x=val))  
 g<-g+geom\_histogram(aes(y=stat(density)), binwidth = 1/dat.samp$ns[i])  
 g<-g+labs(title=str\_c("n=",dat.samp$ns[i],", p=",dat.samp$ps[i]))  
 if(m.iqr[2,i]>0){  
 g<-g+stat\_function(fun=dnorm, colour='red', args = list(mean = m.iqr[1,i], sd = m.iqr[2,i]/(2\*qnorm(0.75))))  
 }  
 print(g)  
 }



#### Question 5.b

Please comment on the visual resemblance between the histograms and the corresponding Normal distribution. Qualitatively, how does the degree of resemblance depend on the number of terms in the sample? Qualitatively, how does degree of resemblance depend on ? (5 points)

### MDH Comment: The degree of resemblence gets better as the number of terms increases, however, as p gets smaller, the degree of resemblence gets worse.