Problem Set 2

C. Durso

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)  
#install.packages("cowplot")  
library(tidyverse)

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

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.1 v dplyr 1.0.0  
## v tidyr 1.1.0 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()

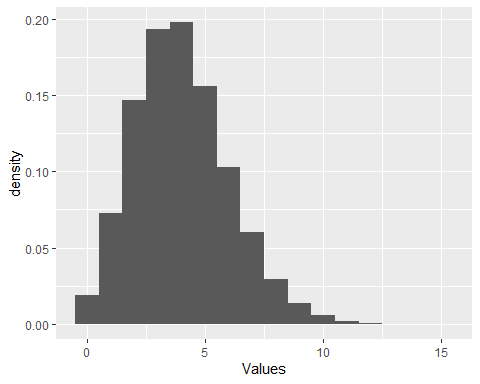
library(knitr)

## Warning: package 'knitr' was built under R version 4.0.2

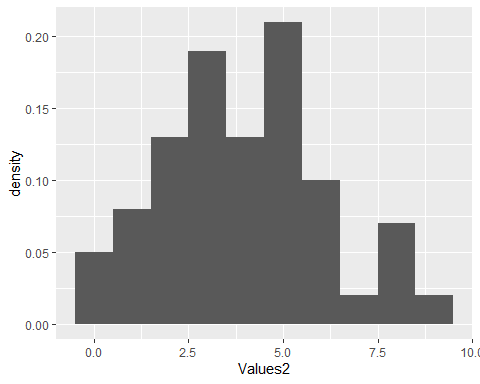
### Question 1

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)  
largeSample <- data.frame(Values=rpois(100000, 4))  
  
(histo1 <- ggplot(largeSample,aes(x=Values))+geom\_histogram(aes(y=..density..),binwidth=1))



set.seed(7654321)  
smallSample <- data.frame(Values2=rpois(100, 4))  
  
(histo2 <- ggplot(smallSample,aes(x=Values2))+geom\_histogram(aes(y=..density..),binwidth=1))



#+stat\_bin(geom="text", aes(label=round(..density..,2)))  
 #+stat\_bin(binwidth=1, geom='text', color='white', aes(label=..density..))

### Question 2

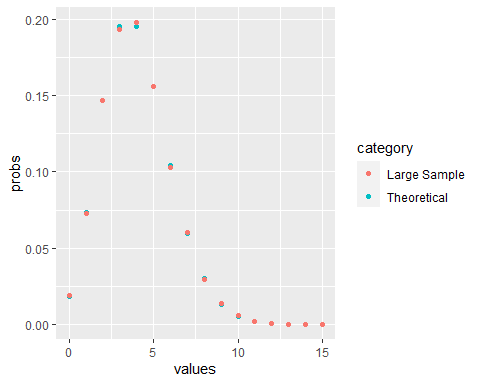
Please generate a visualization that compares the proportions each of the possible outcomes in the sample of size 100,000 above to the theoretical probabilities of each of the outcomes for the Poisson distribution with . Repeat for the sample of size 100. Discuss the appearance of the histograms in relation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

The histogram for the sample where n=100 has some significant differences with the theoretical probabilities of the Poisson distribution. On the other hand, the heights of the bars for the larger sample are almost indistinguishable from the values of the probability density function. This is to be expected. The Law of Large Numbers tells us that the sample densities should approach the theoretical probabilities in the long run.

largeSampDens <- ggplot\_build(histo1)$data[[1]]$density  
smallSampDens <- ggplot\_build(histo2)$data[[1]]$density  
smallSampDens2 <- c(smallSampDens,c(0,0,0,0,0,0))  
  
(theorPoissonVsLarge <-data.frame(values=rep(c(0:15),2),probs=c(dpois(0:15,4),largeSampDens),category=rep(c("Theoretical","Large Sample"),each=16)))

## values probs category  
## 1 0 1.831564e-02 Theoretical  
## 2 1 7.326256e-02 Theoretical  
## 3 2 1.465251e-01 Theoretical  
## 4 3 1.953668e-01 Theoretical  
## 5 4 1.953668e-01 Theoretical  
## 6 5 1.562935e-01 Theoretical  
## 7 6 1.041956e-01 Theoretical  
## 8 7 5.954036e-02 Theoretical  
## 9 8 2.977018e-02 Theoretical  
## 10 9 1.323119e-02 Theoretical  
## 11 10 5.292477e-03 Theoretical  
## 12 11 1.924537e-03 Theoretical  
## 13 12 6.415123e-04 Theoretical  
## 14 13 1.973884e-04 Theoretical  
## 15 14 5.639669e-05 Theoretical  
## 16 15 1.503912e-05 Theoretical  
## 17 0 1.871000e-02 Large Sample  
## 18 1 7.254000e-02 Large Sample  
## 19 2 1.467700e-01 Large Sample  
## 20 3 1.936700e-01 Large Sample  
## 21 4 1.980000e-01 Large Sample  
## 22 5 1.561700e-01 Large Sample  
## 23 6 1.031400e-01 Large Sample  
## 24 7 5.997000e-02 Large Sample  
## 25 8 2.927000e-02 Large Sample  
## 26 9 1.346000e-02 Large Sample  
## 27 10 5.730000e-03 Large Sample  
## 28 11 1.820000e-03 Large Sample  
## 29 12 4.800000e-04 Large Sample  
## 30 13 2.100000e-04 Large Sample  
## 31 14 2.000000e-05 Large Sample  
## 32 15 4.000000e-05 Large Sample

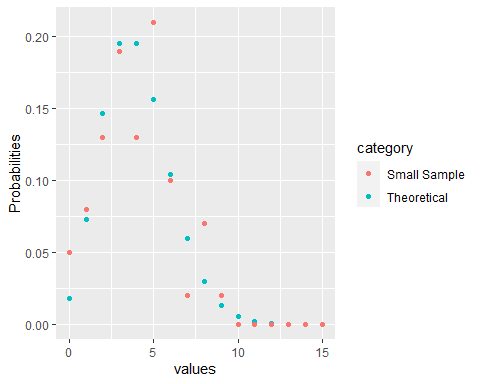
(theoPlot1 <- ggplot(theorPoissonVsLarge,aes(x=values,y=probs,colour=category))+geom\_point())



(theorPoissonVsSmall <- data.frame(values=rep(c(0:15),2),Probabilities=c(dpois(0:15,4),smallSampDens2),category=rep(c("Theoretical","Small Sample"),each=16)))

## values Probabilities category  
## 1 0 1.831564e-02 Theoretical  
## 2 1 7.326256e-02 Theoretical  
## 3 2 1.465251e-01 Theoretical  
## 4 3 1.953668e-01 Theoretical  
## 5 4 1.953668e-01 Theoretical  
## 6 5 1.562935e-01 Theoretical  
## 7 6 1.041956e-01 Theoretical  
## 8 7 5.954036e-02 Theoretical  
## 9 8 2.977018e-02 Theoretical  
## 10 9 1.323119e-02 Theoretical  
## 11 10 5.292477e-03 Theoretical  
## 12 11 1.924537e-03 Theoretical  
## 13 12 6.415123e-04 Theoretical  
## 14 13 1.973884e-04 Theoretical  
## 15 14 5.639669e-05 Theoretical  
## 16 15 1.503912e-05 Theoretical  
## 17 0 5.000000e-02 Small Sample  
## 18 1 8.000000e-02 Small Sample  
## 19 2 1.300000e-01 Small Sample  
## 20 3 1.900000e-01 Small Sample  
## 21 4 1.300000e-01 Small Sample  
## 22 5 2.100000e-01 Small Sample  
## 23 6 1.000000e-01 Small Sample  
## 24 7 2.000000e-02 Small Sample  
## 25 8 7.000000e-02 Small Sample  
## 26 9 2.000000e-02 Small Sample  
## 27 10 0.000000e+00 Small Sample  
## 28 11 0.000000e+00 Small Sample  
## 29 12 0.000000e+00 Small Sample  
## 30 13 0.000000e+00 Small Sample  
## 31 14 0.000000e+00 Small Sample  
## 32 15 0.000000e+00 Small Sample

(theoreticalPlot2 <- ggplot(theorPoissonVsSmall, aes(x=values,y=Probabilities,colour=category))+geom\_point())



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

#### Question 3.a

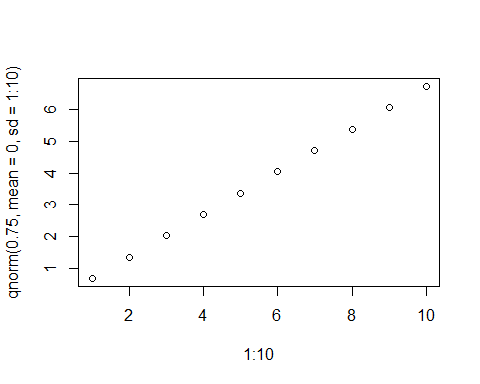
By analogy, for a random variable with cumulative density function , let satisfy . Let satisfy . Define to be the interquartile range for the random variable. Please calculate the values of for the Normal distributions with mean 0 and sd in and plot the points consisting of the value of the sd and the corresponding . This should give an indication of a simple function relating sd and . (5 points)

From the vector and corresponding plot, the function relating sd and for a normal distribution(mu=0) can be derived: .

(qnorm(0.75,mean=0,sd=1:10))

## [1] 0.6744898 1.3489795 2.0234693 2.6979590 3.3724488 4.0469385 4.7214283  
## [8] 5.3959180 6.0704078 6.7448975

plot(1:10,qnorm(0.75,mean=0,sd=1:10))



#### Question 3.b

Note that for mean=0, sd=, and and satisfying and , the value of is . By symmetry of the Normal family, . 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 and on theoretical grounds. You may find that defining by simplifies your argument. (5 points)

so and

#### 3.c

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

upper boundary of IQR for

By symmetry,

#### Question 3.d

In terms of , , and , what is the interquartile range for a Normal distribution with mean equal to and “sd” equal to ? (2 points)

IQR is a measure of spread. The mean does not affect the spread of the distribution.

### Question 4

Consider the probability space that models the observed proportion of successes in independent Bernoulli trials with probability of success equal to . For example, given and , the outcome in should have the same probability as the outcome of successes in independent Bernoulli trials with probability of success equal to . Give a formula for the density function for this probability space in terms of the density of the binomial distribution with size and probability . (5 points)

so for example:

### Question 5

For each pair with and , sample 100,000 values from the distribution of the observed proportion of successes in independent Bernoulli trials with probability of success equal to as in 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)  
  
dat.samp<-  
 data.frame(ns=rep(ns,times=rep(length(ps),length(ns))),  
 ps=rep(ps,times=length(ns)))  
  
binom10.5 <- data.frame(proportion=rbinom(100000,10,.5)/10)  
mean(binom10.5[["proportion"]])

## [1] 0.499983

IQR(binom10.5[["proportion"]])

## [1] 0.2

binom10.1 <- data.frame(proportion=rbinom(100000,10,.1)/10)  
mean(binom10.1[["proportion"]])

## [1] 0.100093

IQR(binom10.1[["proportion"]])

## [1] 0.2

binom10.01 <- data.frame(proportion=rbinom(100000,10,.01)/10)  
mean(binom10.01[["proportion"]])

## [1] 0.009931

IQR(binom10.01[["proportion"]])

## [1] 0

binom50.5 <- data.frame(proportion=rbinom(100000,50,.5)/50)  
mean(binom50.5[["proportion"]])

## [1] 0.4997226

IQR(binom50.5[["proportion"]])

## [1] 0.08

binom50.1 <- data.frame(proportion=rbinom(100000,50,.1)/50)  
mean(binom50.1[["proportion"]])

## [1] 0.0999388

IQR(binom50.1[["proportion"]])

## [1] 0.06

binom50.01 <- data.frame(proportion=rbinom(100000,50,.01)/50)  
mean(binom50.01[["proportion"]])

## [1] 0.0100626

IQR(binom50.01[["proportion"]])

## [1] 0.02

binom1000.5 <- data.frame(proportion=rbinom(100000,1000,.5)/1000)  
mean(binom1000.5[["proportion"]])

## [1] 0.4999404

IQR(binom1000.5[["proportion"]])

## [1] 0.022

binom1000.1<- data.frame(proportion=rbinom(100000,1000,.1)/1000)  
mean(binom1000.1[["proportion"]])

## [1] 0.1000322

IQR(binom1000.1[["proportion"]])

## [1] 0.012

binom1000.01 <- data.frame(proportion=rbinom(100000,1000,.01)/1000)  
mean(binom1000.01[["proportion"]])

## [1] 0.01000057

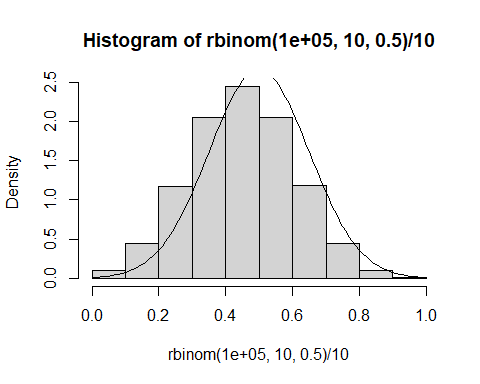
IQR(binom1000.01[["proportion"]])

## [1] 0.004

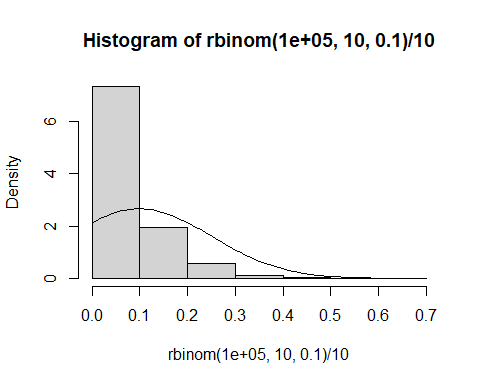
### Question 6

For each and 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 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 . (10 points)

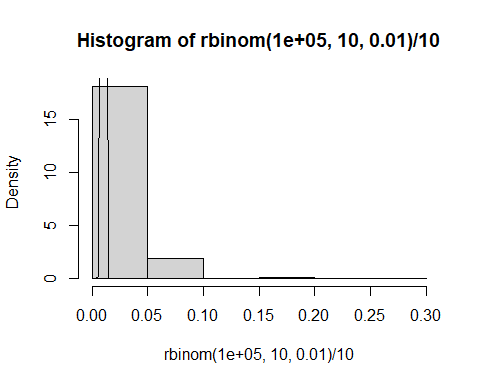
hist(rbinom(100000,10,.5)/10,breaks=11,freq = FALSE)  
x<-seq(0,10,by=0.02)  
curve(dnorm(x,mean=.5,sd=.15), add=TRUE)



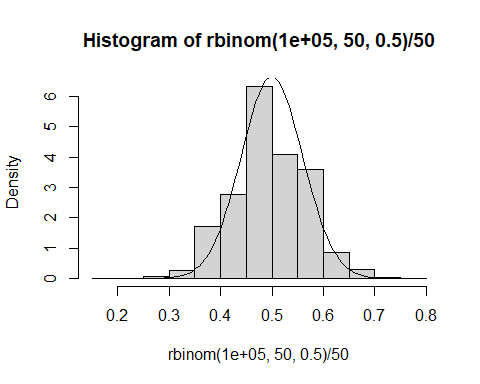
hist(rbinom(100000,10,.1)/10,breaks=5,freq = FALSE)  
x<-seq(0,10,by=0.02)  
curve(dnorm(x,mean=.1,sd=.15), add=TRUE)



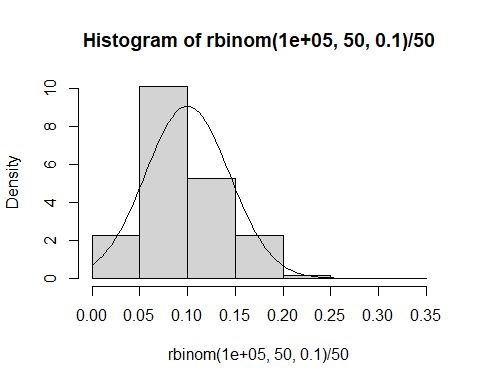
hist(rbinom(100000,10,.01)/10,breaks=10,freq = FALSE)  
x<-seq(0,10,by=0.02)  
curve(dnorm(x,mean=.01,sd=.001), add=TRUE)



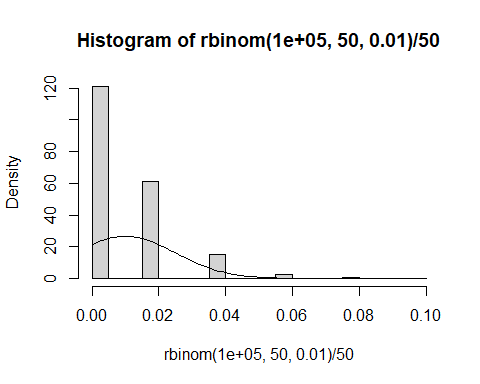
hist(rbinom(100000,50,.5)/50,breaks=20,freq = FALSE)  
x<-seq(0,50,by=0.02)  
curve(dnorm(x,mean=.5,sd=.06), add=TRUE)



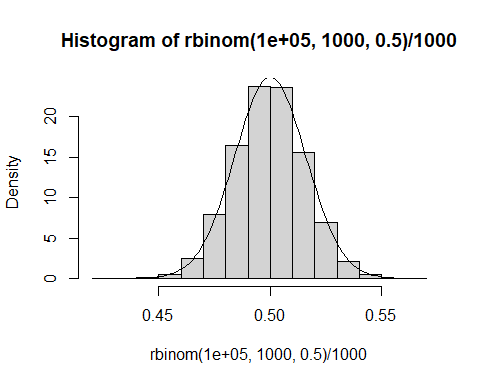
hist(rbinom(100000,50,.1)/50,breaks=5,freq = FALSE)  
x<-seq(0,50,by=0.02)  
curve(dnorm(x,mean=.1,sd=.044), add=TRUE)



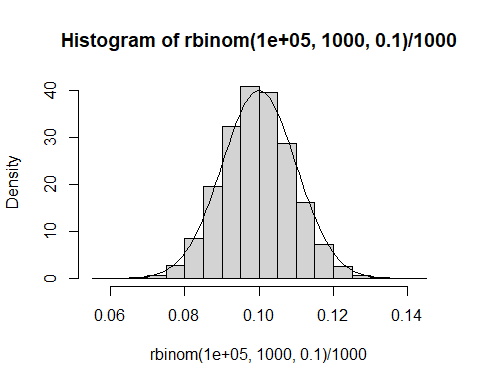
hist(rbinom(100000,50,.01)/50,breaks=20,freq = FALSE)  
x<-seq(0,50,by=0.02)  
curve(dnorm(x,mean=.01,sd=.015), add=TRUE)



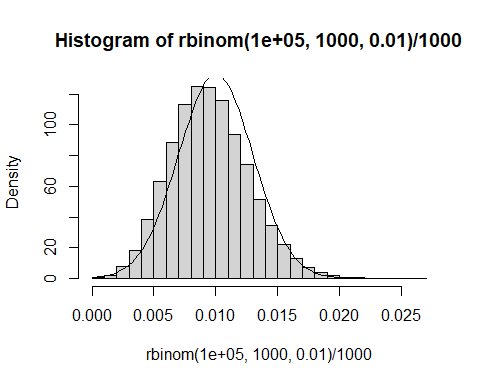
hist(rbinom(100000,1000,.5)/1000,breaks=20,freq = FALSE)  
x<-seq(0,1000,by=0.02)  
curve(dnorm(x,mean=.5,sd=.016), add=TRUE)



hist(rbinom(100000,1000,.1)/1000,breaks=20,freq = FALSE)  
x<-seq(0,1000,by=0.02)  
curve(dnorm(x,mean=.1,sd=.01), add=TRUE)



hist(rbinom(100000,1000,.01)/1000,breaks=20,freq = FALSE)  
x<-seq(0,1000,by=0.02)  
curve(dnorm(x,mean=.01,sd=.003), add=TRUE)



### Question 7

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

For binomial/Bernoulli situations the closer the p value is to .5, the more symmetric the distribution is, which therefore lends itself to approximation with a Normal distribution. That said, any binomial distribution regardless of the p value can eventually be approximated with a Normal curve, provided the sample size is high enough. The higher sample size shrinks the standard deviation of the sample distribution and it appears closer to a smooth, continuous distribution. The relevant example here is the final graph, where p=.01, but since n=100000, the natural skewness is overcome and the normal curve overlay fits reasonably well.