YingHu\_Assignment5

Ying Hu

9/24/2019

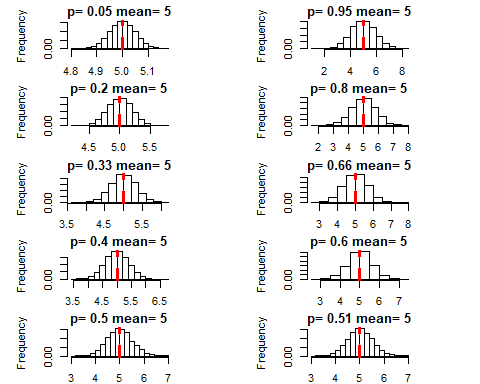
#short essay for comparison of two distributions:  
  
#In chapter 4, we have learned different distributions, the two particular distributions that I found interesting are Poisson and normal distributions.   
  
#Poisson distribution is a function for calculating mass which the result is an integer in statistics; it analyzes number and average probability of events that happened in a particular timeframe in data analysis, and the events are independent; besides, if the number is very large, the probability will get smaller or approach to zero.  According to those, it is commonly used in the rare events in real life, such as network failures; For example, if an office gets average of 12 failures per hours, the probability of having 20 failures in an hours is 2%. In this case, there is no necessity to hire more technicians since the rate is low. In addition, it is also used in traffic accidents, disease in a certain neighborhood and so on; we can figure out whether or not the data came from Poisson distribution by testing if the probability of an event occurs in a given time, distance, area or volume is the same. However, the limitation of Poisson distribution is that it is used mainly for rare events but not for the normal behaviors or events; in this case, we can group those probabilities and take a sum of results for normal distributions purpose.   
  
#Normal distribution is a function for calculating density in statistics; but its result is not just a number in data analysis, the data comes from the sum of all the independent trails or events. The larger the sample size the more closely fit a curve will generate; so there are not points but the area under the curve between two points, this is because it sums all the values into a categories to get the probabilities of every event; It is simple and for a large number of distributions and widely used in real life. For example, we can use it for weight: given a bag of cookies is underweight if it weighs 500 grams and the standard deviation is 4 grams, and the normal weight is 510 grams; from this information, we can get the probability of randomly getting an underweight bag of cookies to be (500-510)/4 =-2.5(Z<-2.5), and according to the z-table, the rate is 0.0062. Therefore, the store can stock more cookies since the risk of having underweight bags of cookies is very small. In addition, it is also used in stock market to forecast the risk and return, income distribution in economy for government and so on. However, the limitation of normal distribution is that there will be negative numbers of probability, caused by a mean and a symmetric range of variation sometimes. Like the example of the cookies, the probability can’t be negative. The log-normal distribution can handle this issues and it distributed random variable can never go below zero.  In addition, we can also find out if the data is from a normal distribution or not by checking if the curve has a bell shape, sample size is large for common events, and there is a range or not.   
  
#In conclusion, there is a connection between Poisson and normal distribution. Poisson distribution has interval of time and the events are independent; and normal distribution is a sum of those independent events and trails.

## R Markdown

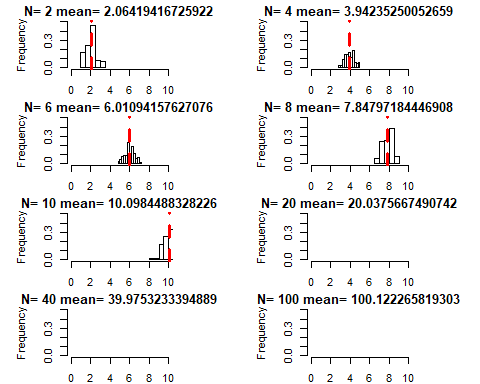
This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

# 1) Normal distribution:  
p <- c(0.05, 0.95, 0.2, 0.8, 0.33, 0.66, 0.4, 0.6, 0.5, 0.51)#set p as one of the parameter equal sd  
mean <- 5 #set mean is fix value   
df <- sapply(p, function(p) {rnorm(10000,mean=5, sd=p)})# when p is unstable , mean is fixed.  
  
  
colnames(df) <- c("p=0.05", "p=0.95", "p=0.2", "p=0.8", "p=1/3", "p=2/3", "p=0.4", "p=0.6", "p=0.5", "p=0.51")  
  
  
#pdf("normalProb1.pdf")  
par(mfrow=c(5,2))  
par(cex=0.7)  
par(mar=c(2,5,1.5,5))  
for (i in 1: dim(df)[2]) {  
 #dev.new()  
 h1 <- list()  
 h1[[i]] <- hist(df[,i], plot=FALSE)  
 h1[[i]]$counts <- h1[[i]]$counts/sum(h1[[i]]$counts)  
 plot(h1[[i]], main = paste("p=", p[i], "mean=", mean=5 ))  
 abline(v = mean, col="red", lwd=3, lty=2)# since the meant is fixed which is 5, so just put v=mean  
 #dev.off()  
}

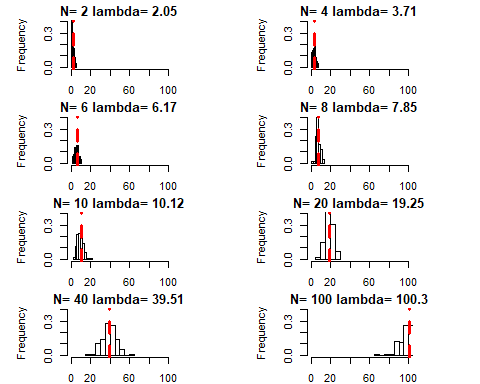


#2) Normal distribution:   
  
N <- c(2, 4, 6, 8, 10, 20, 40, 100)##set N as one of the parameter equal mean  
p <- 0.5 # p is a fixed value which is equal sd  
df <- sapply(N, function(N) {rnorm(100,mean=N, sd=p)})  
  
colnames(df) <- c("n=2", "n=4", "n=6", "n=8", "n=10", "n=20", "n=40", "n=100")  
#head(df)  
#colnames(df)  
  
#pdf("binomProb1.pdf")  
par(mfrow=c(4,2))  
par(cex=0.7)  
#plot(title.adj=c(0,1))  
par(mar=c(2,5,1.5,5))  
for (i in 1: dim(df)[2]) {  
 #dev.new()  
 h1 <- list()  
 h1[[i]] <- hist(df[,i], plot=FALSE)  
 h1[[i]]$counts <- h1[[i]]$counts/sum(h1[[i]]$counts)  
 plot(h1[[i]], ylim=c(0,0.5), xlim=c(0,10), main = paste("N=", N[i], "mean=", mean(df[,i])))  
 abline(v = mean(df[,i]), col="red", lwd=3, lty=2)}



#dev.off()

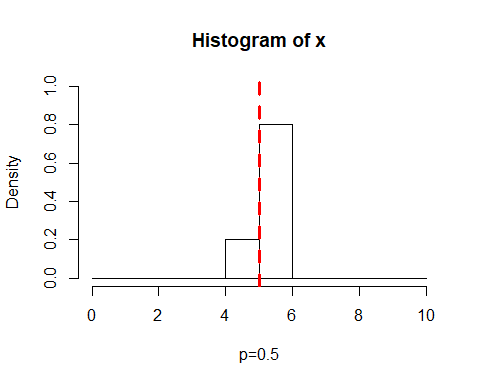
# 3) Poisson distribution:  
  
N <- c(2, 4, 6, 8, 10, 20, 40, 100) #since poisson only get one parameter lambda, we set lambada equal N.   
df <- sapply(N, function(N) {rpois(100,lambda=N)})  
  
colnames(df) <- c("n=2", "n=4", "n=6", "n=8", "n=10", "n=20", "n=40", "n=100")  
#head(df)  
#colnames(df)  
  
#pdf("binomProb1.pdf")  
par(mfrow=c(4,2))  
par(cex=0.7)  
#plot(title.adj=c(0,1))  
par(mar=c(2,5,1.5,5))  
for (i in 1: dim(df)[2]) {  
 #dev.new()  
 h1 <- list()  
 h1[[i]] <- hist(df[,i], plot=FALSE)  
 h1[[i]]$counts <- h1[[i]]$counts/sum(h1[[i]]$counts)  
 plot(h1[[i]], ylim=c(0,0.4), xlim=c(0,100), main = paste("N=", N[i], "lambda=", mean(df[,i]))) #lambda and mean has same value  
 abline(v = mean(df[,i]), col="red", lwd=3, lty=2)  
 #dev.off()  
}



#nomal distribution:  
  
p <- 0.5  
x <- rnorm(10, 5, 0.05)  
mean(x)

## [1] 5.019505

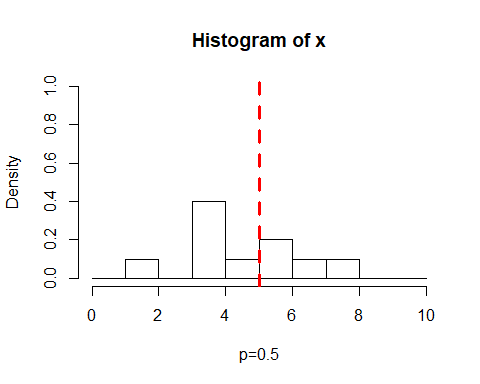
library(plyr)  
y <- count(x)  
  
y$relF <- y$freq/sum(y$freq)  
#barplot(y$relF ~ y$x, xlab="x for n=5, p=0.05, np=0.25", ylab="relative frequency")  
#axis(side = 2, pos = -0.2)  
#axis(side = 1, at = seq\_along(y$x) - 0.5, tick = FALSE, labels = y$x)  
hist(x, breaks=seq(0,10,1), prob=T, ylim=c(0,1), xlab=paste0("p=",p))  
abline(v = mean(x), col="red", lwd=3, lty=2)



#Poisson distribution:  
p<-0.5  
x <- rpois(10, 5)  
mean(x)

## [1] 5

library(plyr)  
y <- count(x)  
  
y$relF <- y$freq/sum(y$freq)  
#barplot(y$relF ~ y$x, xlab="x for n=5, p=0.05, np=0.25", ylab="relative frequency")  
#axis(side = 2, pos = -0.2)  
#axis(side = 1, at = seq\_along(y$x) - 0.5, tick = FALSE, labels = y$x)  
hist(x, breaks=seq(0,10,1), prob=T, ylim=c(0,1), xlab=paste0("p=",p))  
abline(v = mean(x), col="red", lwd=3, lty=2)



summary(cars)

## speed dist   
## Min. : 4.0 Min. : 2.00   
## 1st Qu.:12.0 1st Qu.: 26.00   
## Median :15.0 Median : 36.00   
## Mean :15.4 Mean : 42.98   
## 3rd Qu.:19.0 3rd Qu.: 56.00   
## Max. :25.0 Max. :120.00

## Including Plots

You can also embed plots, for example:



Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.