BIOS6611-Homework2

Randy

9/16/2019

### Exercise 1: Discrete Distributions: Binomial and Poisson

#### Question 1a

bino.p <- pbinom( 0.025, size = 120, prob = 0.01 ); bino.p

## [1] 0.2993804

pois.p <- ppois( 0.025, lambda = 120 \* 0.01 ); pois.p

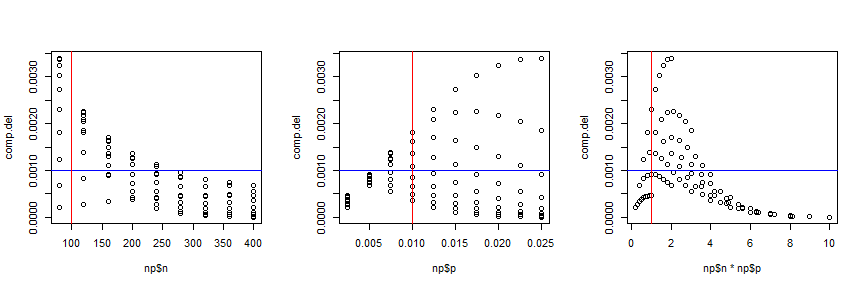
## [1] 0.3011942

pois.p - bino.p

## [1] 0.001813821

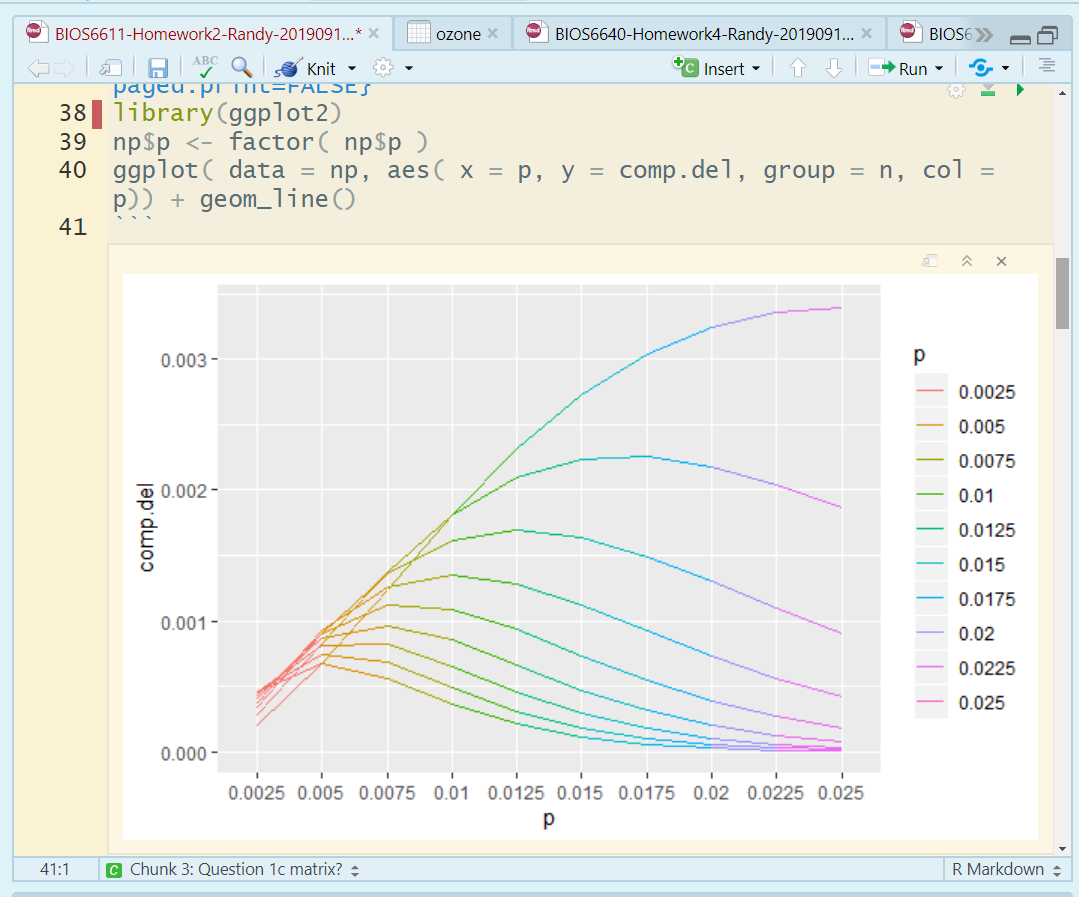
#### Question 1b

n <- seq( 80, 400, by = 40)  
p <- seq( 0.0025, 0.025, by = 0.0025)  
np <- expand.grid( n = n, p = p)  
bino.s <- pbinom( 0.025, np$n, np$p)  
pois.s <- ppois( 0.025, lambda = np$n \* np$p)  
comp.del <- pois.s - bino.s  
par( mfrow = c(1, 3) )  
plot( np$n, comp.del ); abline( h = 0.001, v = 100, col = c( "blue", "red" ) )  
plot( np$p, comp.del ); abline( h = 0.001, v = 0.01, col = c( "blue", "red" ) )  
plot( np$n \* np$p, comp.del ); abline( h = 0.001, v = 100 \* 0.01, col = c( "blue", "red" ) )



library(ggplot2)  
np$p <- factor( np$p )

***ggplot( data = np, aes( x = p, y = comp.del, group = n, col = p)) + geom\_line()*** the code works but there is always error on png() device..



#### Question 1c

The difference between the exact probability of binomial distribution and the approximation from possion distribution decreases with larger sample size and small prevalence. I would suggest her to use the approximation with solely p ≤ 0.005 or n ≥ 300; but combine the sample size and the prevalence together, I would recommond just as Rosner: a conservative rule to use the approximation when n ≥ 100 and p ≤ 0.01.

### Exercise 2: Expected Value and Variance for Exponential Distribution

#### Question 2a

Now we know that the , then ; Sally has to expect to wait for another 1/3 hour in the line.

#### Question 2b

So the variation around this estimation is

#### Question 2c

set.seed( seed = 555 )  
exp.r <- rexp( n = 100000, rate = 3)  
mean( exp.r ); var( exp.r )

## [1] 0.3330338

## [1] 0.1111126

The simulation mean is very near to the theoretical mean = 1/3; and the variation is similar to = 1/9.

#### Question 2d

I would assume Sally has to wait for another 1/3 hour. Because for exponential distribution, every instant is like the beginning of a new period. That means she has to go over the same distribution regardless of how much time has already elapsed.

### Exercise 3: Ozone Status for Normal Approximation to the Binomial

#### Question 3a

ozone <- read.csv("ozone.csv")  
ozone <- data.frame(ozone)  
n.ozone <- length( ozone$AQI.Category )  
ozone.good <- ozone[ozone$AQI.Category == "Good", ]   
n.ozone.good <- length( ozone.good$AQI.Category )  
n.ozone; n.ozone.good

## [1] 151

## [1] 80

p.ozone.good <- n.ozone.good / n.ozone; p.ozone.good

## [1] 0.5298013

So the daily probability of “good” ozone levels is 0.5298013

#### Question 3b

ozone.good.567 <- dbinom( x=5:7, size = 7, prob = p.ozone.good )  
sum(ozone.good.567)

## [1] 0.2783011

So the exact probability of at least 5 of the 7 days will have “good” ozone is 0.2783011

#### Question 3c

ozone.mu.ap <- 7 \* p.ozone.good; ozone.sigma.ap <- sqrt(7 \* p.ozone.good \* ( 1 - p.ozone.good ) )  
ozone.mu.ap; ozone.sigma.ap

## [1] 3.708609

## [1] 1.320524

z.ozone.5less <- (4.5 - ozone.mu.ap) / ozone.sigma.ap; p.ozone.5less.ap <- pnorm( z.ozone.5less ); p.ozone.567.ap <- 1 - p.ozone.5less.ap; p.ozone.567.ap

## [1] 0.2744862

So from the result, we can see the probability of normal distribution approximation 0.2744862 is pretty near to the probability of real binomial distribution 0.2783011.

#### Question 3d

let’s select the random 100 days for 100 times check whether the ozone “good” days mean and variation follow binomial distribution.

set.seed( seed = 555)  
ozone.good <- function( days )  
{  
 index <- sample( n.ozone, days , replace = F)  
 ozone.sp <- ozone[index, ]  
 ozone.sp.good <- ozone.sp[ ozone.sp$AQI.Category == "Good", ]  
 n.ozone.good <- length( ozone.sp.good$AQI.Category )  
 n.ozone.good  
}  
ozone.result <- sapply( rep(100, 100), FUN = ozone.good );  
summary(ozone.result)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 46.00 51.00 53.00 53.21 55.00 60.00

mean(ozone.result)

## [1] 53.21

ozone.bin.mean <- 100 \* p.ozone.good; ozone.bin.mean

## [1] 52.98013

var(ozone.result) \* 100

## [1] 798.5758

ozone.bin.var <- 100 \* p.ozone.good \* ( 1 - p.ozone.good); ozone.bin.var

## [1] 24.91119

According to the random samples, the mean of the “good” ozone days is pretty near to the theoretical binomial distribution mean, however the variance is not similar to the theoretical binomial distribution variance. So the “good” ozone days do not follow binomial distribution.