Question 2

Joseph Froelicher

9/17/2020

# Part A

N = 120  
P = 0.01  
  
binomial\_prob <- pbinom(0.025, N, P)  
binomial\_prob

## [1] 0.2993804

poisson\_prob <- ppois(0.025, N \* P)  
poisson\_prob

## [1] 0.3011942

I would call the difference between the binomial estimation technique and the poisson approximation to the binomial estimation negligible. The difference between the two is about 1%. As we increase the sample size this difference will decrease.

# Part B

library(plot.matrix)

## Warning: package 'plot.matrix' was built under R version 3.6.3

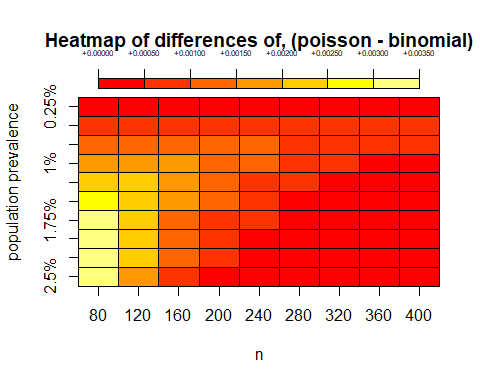
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

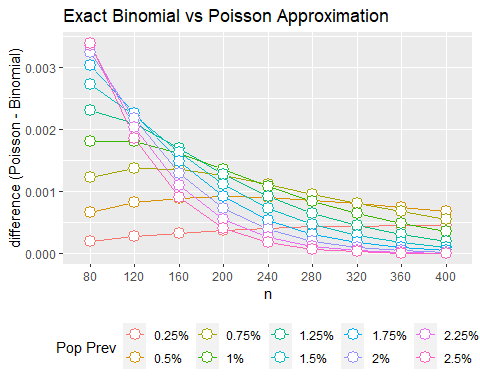
library(reshape2)

## Warning: package 'reshape2' was built under R version 3.6.3

size\_start <- 80  
size\_stop <- 400  
size\_step <- 40  
  
prevalence\_start <- 0.0025  
prevalence\_stop <- 0.025  
prevalence\_step <- 0.0025  
  
PREVALENCE <- 0.025  
  
difference.mat <- matrix(  
 NA,  
 nrow = prevalence\_stop / prevalence\_step,  
 ncol = (size\_stop - size\_step) / size\_step  
)  
  
i <- 0  
  
for ( n in seq(size\_start, size\_stop, size\_step) ) {  
 i <- i + 1  
 j <- 0  
   
 for ( p in seq(prevalence\_start, prevalence\_stop, prevalence\_step)) {  
 j <- j + 1  
   
 difference.mat[j, i] <- ppois(PREVALENCE, n \* p) - pbinom(PREVALENCE, n, p)   
   
 }  
}  
  
colnames(difference.mat) <- c(seq(80,400,40))  
rownames(difference.mat) <- c('0.25%','0.5%','0.75%','1%','1.25%','1.5%','1.75%','2%','2.25%','2.5%')  
  
plot(  
 difference.mat,  
 main = "Heatmap of differences of, (poisson - binomial)",  
 xlab = "n",  
 ylab = "population prevalence",  
 key = list(side = 3, cex.axis = 0.5)  
)



difference.df <- data.frame(  
 'pop\_prev' = c('0.25%','0.5%','0.75%','1%','1.25%','1.5%','1.75%','2%','2.25%','2.5%'),  
 difference.mat  
)  
  
colnames(difference.df) <- c('pop\_prev','80','120','160','200','240','280','320','360','400')  
  
difference.df <- melt(difference.df, id.vars="pop\_prev", value.name="difference", variable.name="n")  
  
ggplot(difference.df, aes(n, difference, group = pop\_prev, colour = pop\_prev)) +  
 geom\_line() +  
 geom\_point(size = 4, shape = 21, fill="white") +  
 ggtitle("Exact Binomial vs Poisson Approximation") +  
 scale\_colour\_discrete("Pop Prev") +  
 labs(y = "difference (Poisson - Binomial)") +  
 theme(legend.position = "bottom")



# Part C

In general, because sampling is so expensive, the most cost effective option is to use a Poisson approximation with a sample size of 80, and assuming a population prevalence of 0.25%. However, we can always dream about the world where we have unlimited money. If we were to have unlimited money, we could sample with n = 400, and assume a population prevalence of 2.5%. The difference between a binomial probability and the Poisson approximation to the binomial probability is negligible at n = 400, p = 0.025.

Rosner’s recommendation is to use the Poisson approximation when n >= 100 and p <= 0.01. At least from the plot provided above, I am gathering that while the difference is again negligible when n = 80 (we are talking about a maximum difference of 0.0035, or 0.35%), our best simulations are coming at n >= 320 (difference of sample p < 0.001). Note that the plot provides evidence of convergence as n increases, which while the cut off suggested by Rosner is 100, our plot still supports the theory that as n get large, the difference is no longer significant. I would also argue that in this simulation, population prevalence becomes irrelevant as n gets large.