Problem Set 4

Megan Hoeksema

## Introduction

We discussed parameter fitting in class, and saw examples of modeling data with a model from a parametrized family. In these examples the model with the optimal parameters fit the corresponding data fairly well. This depends on the model family’s being well suited to the data. If it isn’t, even the best parameters won’t produce a model that closely reflects the data. Please be aware of this possibility as you work through the examples here.

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 work should be based on the data’s being in the same folder as the .Rmd file. Please turn in your work on Canvas. Your solution document should have your answers to the questions and should display the requested plots. Each part is worth 5 points.

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> ).

## Question 1

This question uses 2019 data primarily for Denver county accessed through IPUMS-USA, University of Minnesota, www.ipums.org ,

Steven Ruggles, Sarah Flood, Ronald Goeken, Josiah Grover, Erin Meyer, Jose Pacas and Matthew Sobek. IPUMS USA: Version 10.0 [dataset]. Minneapolis, MN: IPUMS, 2020. <https://doi.org/10.18128/D010.V10.0>

Detailed descriptions of the variable definitions are available from <https://usa.ipums.org/usa-action/variables/group> by selecting the variable name or “codes” from the household or person drop-down menus.

The PUMA-to-county restriction was done using MABLE, <http://mcdc.missouri.edu/websas/geocorr12.html>

### IPUMS Data

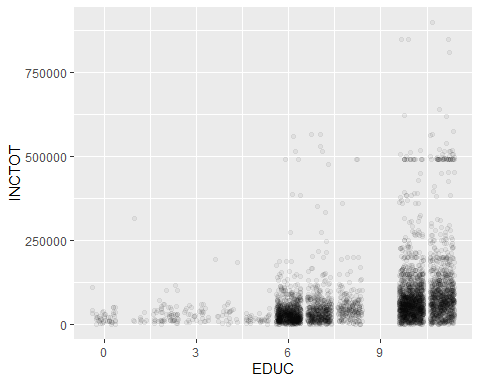
Read in the ipums data.

dat<-read.csv("usa\_00021.csv")  
dat\_trim <- select(dat, YEAR:INCTOT)  
  
write.csv(dat, file = "usa\_00016\_trim.csv")

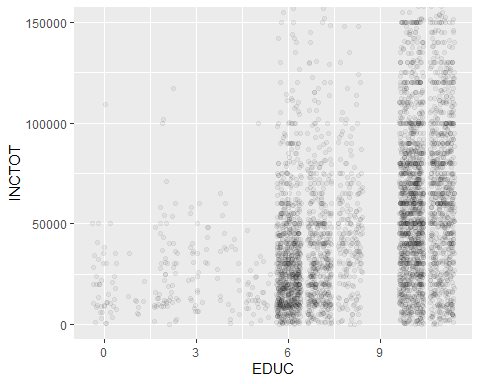
Trim to PUMAs that are predominantly in Denver. Restrict to respondents who are at least 25 years old with non-zero income. The variable “INCTOT” gives total annual individual income. The variable “EDUC” consists of ordered categories of amount of formal education, with 0 representing the least and 11 representing the most. Details are in the “educ\_codes.csv”.

The plot shows the “jitter” tool and the “alpha” approach for displaying large numbers of data points without losing information through plotting points on top of each other.

denver<-filter(dat,PUMA>=812 & PUMA<=816, INCTOT>0,AGE>=25)  
g<-ggplot(denver,aes(x=EDUC,y=INCTOT))+geom\_jitter(alpha=.05)  
g



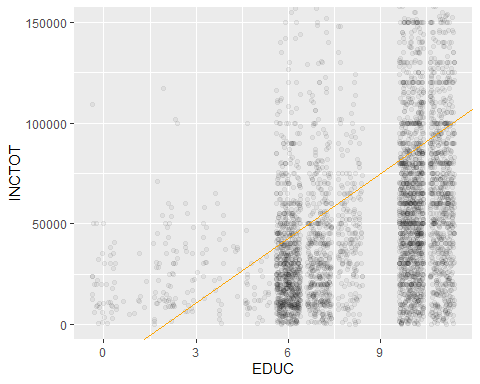
g<-g+coord\_cartesian(ylim =c(0,1.5e5))  
g



### Question 1.a (5 points)

Please add the least squares best fit line in orange to the plot “g” above, save the result as “g”, and display the result. What change in income is associated with an increase of 1 in the “EDUC” category according to the linear model?

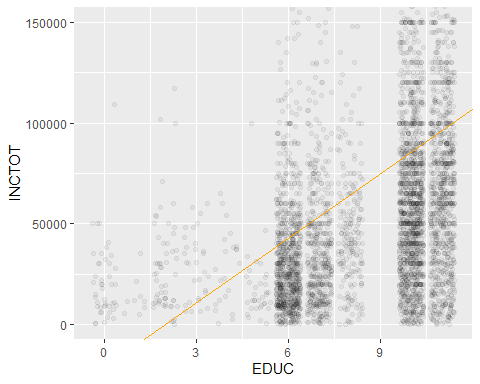
m.edu<-lm(INCTOT~EDUC,data=denver)  
g<-g+geom\_abline(slope=m.edu$coefficients[2],  
 intercept =m.edu$coefficients[1],color="orange" )  
g



g<-g+coord\_cartesian(ylim =c(0,1.5e5))

## Coordinate system already present. Adding new coordinate system, which will replace the existing one.

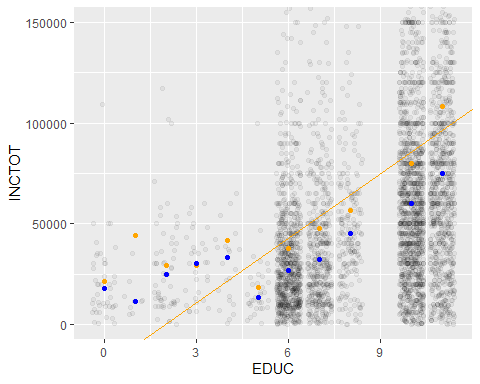
g



### Question 1.b (5 points)

Please add the income means for each education category in orange and the income medians (R function “median”) in each category in blue to the plot “g”, save the result as “g”, and display the result.

den.mean<-denver%>%group\_by(EDUC)%>%  
 summarize(inc.mean=mean(INCTOT),inc.med=median(INCTOT))  
g<-g+  
geom\_point(data=den.mean,aes(x=EDUC,y=inc.mean),color="orange")+  
 geom\_point(data=den.mean,aes(x=EDUC,y=inc.med),color="blue")  
  
g



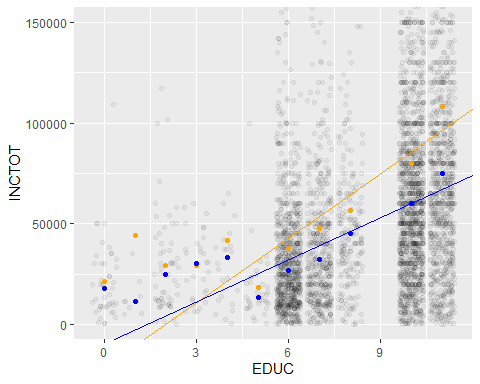
### Question 1.c (5 points)

The least squares criterion is one way of fitting a line to a collection of points .

One alternative is to fit the line that minimizes the sum of the absolute errors .

Please use “nlm” to fit this line for “INCTOT” as a function of “EDUC” and add this line in blue to “g”, save the result as “g”, and display the result. The slope and intercept from 1.a are option for the starting parameters. Other starting parameters may give other lines.

abs.error<-function(coeffs,x,y){  
 sum(abs(y-(coeffs[2]\*x+coeffs[1])))  
}  
  
m.abs<-nlm(abs.error,p=m.edu$coefficients, x=denver$EDUC,y=denver$INCTOT)  
  
g<-g+  
 geom\_abline(slope=m.abs$estimate[2],  
 intercept=m.abs$estimate[1],color="blue")  
g

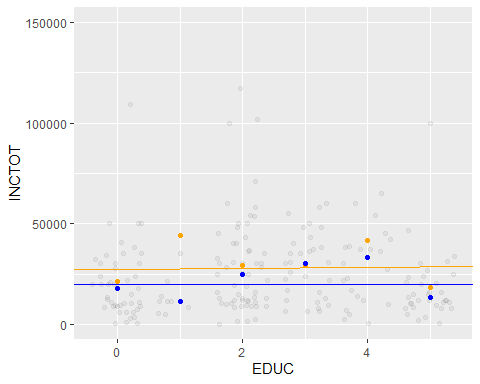


### Question 1.d (5 points)

Please redo 1.a-1.c restricting first to “EDUC” less than or equal to 5, then to “EDUC” greater than or equal to 5.

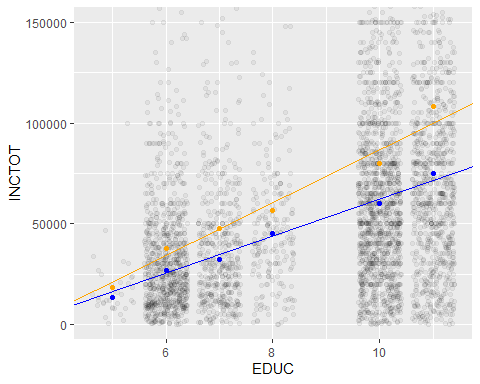
For the lower range of “EDUC”:

den.low<-filter(dat,PUMA>=812 & PUMA<=816, INCTOT>0,AGE>=25,  
 EDUC<=5)  
  
g<-ggplot(den.low,aes(x=EDUC,y=INCTOT))+geom\_jitter(alpha=.05)  
g<-g+coord\_cartesian(ylim =c(0,1.5e5))  
  
m.edu<-lm(INCTOT~EDUC,data=den.low)  
g<-g+geom\_abline(slope=m.edu$coefficients[2],  
 intercept =m.edu$coefficients[1],color="orange" )  
  
den.mean<-den.low%>%group\_by(EDUC)%>%  
 summarize(inc.mean=mean(INCTOT),inc.med=median(INCTOT))  
g<-g+  
geom\_point(data=den.mean,aes(x=EDUC,y=inc.mean),color="orange")+  
 geom\_point(data=den.mean,aes(x=EDUC,y=inc.med),color="blue")  
  
m.abs<-nlm(abs.error,p=m.edu$coefficients, x=den.low$EDUC,y=den.low$INCTOT)  
  
g<-g+  
 geom\_abline(slope=m.abs$estimate[2],  
 intercept=m.abs$estimate[1],color="blue")  
g



For the upper range of “EDUC”:

den.high<-filter(dat,PUMA>=812 & PUMA<=816, INCTOT>0,AGE>=25,  
 EDUC>=5)  
  
g<-ggplot(den.high,aes(x=EDUC,y=INCTOT))+geom\_jitter(alpha=.05)  
g<-g+coord\_cartesian(ylim =c(0,1.5e5))  
  
m.edu<-lm(INCTOT~EDUC,data=den.high)  
g<-g+geom\_abline(slope=m.edu$coefficients[2],  
 intercept =m.edu$coefficients[1],color="orange" )  
  
den.mean<-den.high%>%group\_by(EDUC)%>%  
 summarize(inc.mean=mean(INCTOT),inc.med=median(INCTOT))  
g<-g+  
geom\_point(data=den.mean,aes(x=EDUC,y=inc.mean),color="orange")+  
 geom\_point(data=den.mean,aes(x=EDUC,y=inc.med),color="blue")  
  
m.abs<-nlm(abs.error,p=m.edu$coefficients, x=den.high$EDUC,y=den.high$INCTOT)  
  
g<-g+  
 geom\_abline(slope=m.abs$estimate[2],  
 intercept=m.abs$estimate[1],color="blue")  
g



### Question 1.e (5 points)

Please comment on the quality of the models fitted in parts 1.a-1.d. In particular, please identify the cases in which a line appears to be an appropriate summary of the data, identify the cases in which a line does not appear to be an appropriate summary of the data, and explain your reasoning. Also, please comment on the apparent relationship between the two types of fitted line and the two statistics, mean and median, for location of center. Finally, please comment on the size of the estimated change in “INCTOT” associated with a change in “EDUC” relative to the sample standard deviations of “INCTOT” within “EDUC” category.

### MDH Comments: The fit isn’t great to start out with but then its very poor when you limit educ to 4 and therefore summarizes the data the least appropriately, however, the fit is much better when you limit the education to 5+ and therefore summarizes the data the most appropriately.. The median is a much better fit for location of center. For the standard deviation, the values are high to start out with but then do not get much higher as the level of education increases. As the level of education increases, the INCTOT increases as well giving allowing for a better standard deviation approximation.

denver%>%group\_by(EDUC)%>%summarize(stdev=sd(INCTOT))

## # A tibble: 11 x 2  
## EDUC stdev  
## <int> <dbl>  
## 1 0 19157.  
## 2 1 95998.  
## 3 2 24989.  
## 4 3 16512.  
## 5 4 48657.  
## 6 5 17284.  
## 7 6 47148.  
## 8 7 64210.  
## 9 8 59319.  
## 10 10 83982.  
## 11 11 118665.

## Question 2

(5 points)

Recall that Poisson distribution with parameter is a discrete probability distribution with non-negative integer outcomes. The probability of the outcome equals . Given the following computations, identify the mean and variance of the Poisson distribution with parameter .

and

and finally

### MDH Comment: In conclusion, both the mean and variance of the poisson distribution are equal to lambda.

### Question 3

(10 points)

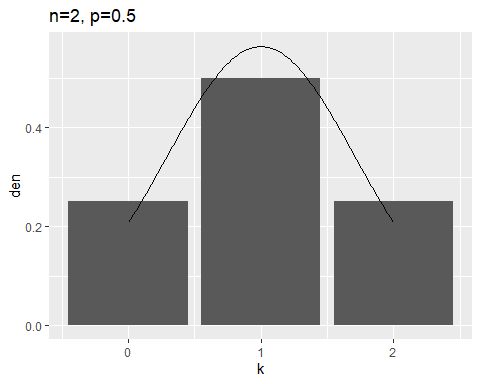
In this question, you will compare binomial distributions with Normal distributions having related means and variances.

For each pair with and , please create a column plot of the density function for over the non-negative integers less than or equal to for which for which is in the image of under the quantile function for . That is, only the probability of for those values of for which the cumulative distribution of at is in . The purpose of restricting to this set of ’s is to generate visualizations that emphasize the higher probability regions of the sample space.

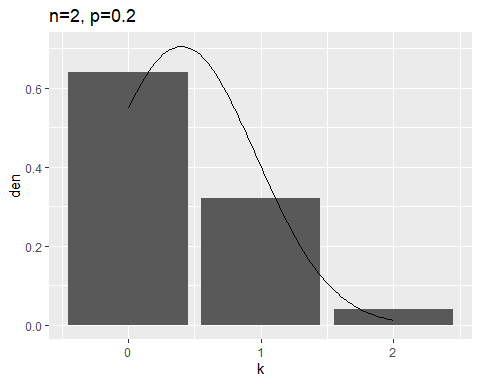
Set the height of the column over the value equal to the probability of under the distribution . On this plot, draw the density of the Normal distribution with the same mean and variance as . Please label the plots with the corresponding and .

ns<-c(2,5,20)  
ps<-c(.5,.2,.1)  
  
dat.samp<-  
 data.frame(ns=rep(ns,times=rep(length(ps),length(ns))),  
 ps=rep(ps,times=length(ns)))  
#dat.samp<-mutate(dat.samp,mn=ns\*ps,stdev=sqrt(ns\*ps\*(1-ps)))  
plot.make<-function(param){  
 n<-param[1]  
 p<-param[2]  
 k.lower<-qbinom(.01,n,p)  
 k.upper<-qbinom(.99,n,p)  
 k<-k.lower:k.upper  
 dat.this<-data.frame(k=k,den=dbinom(k,n,p))  
 g.this<-ggplot(data=dat.this,aes(x=k,y=den))+geom\_col()+  
 stat\_function(fun=dnorm,args=list(mean=n\*p,sd=sqrt(n\*p\*(1-p))))  
 g.this<-g.this+labs(title=str\_c("n=",n,", p=",p))  
}  
  
plots<-apply(dat.samp,1,plot.make)  
plots

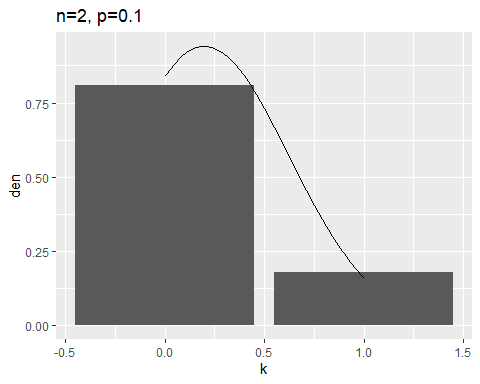
## [[1]]



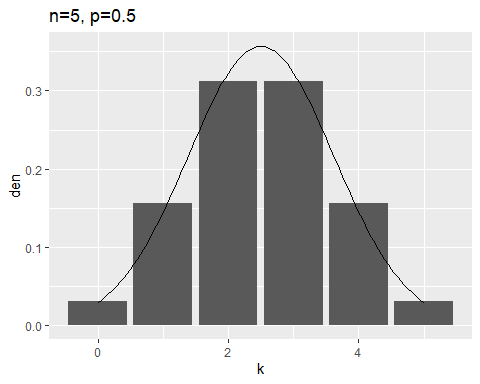
##   
## [[2]]



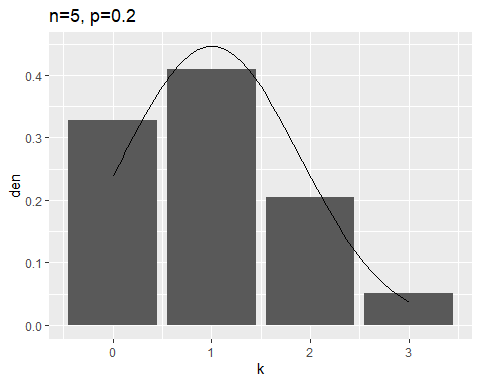
##   
## [[3]]



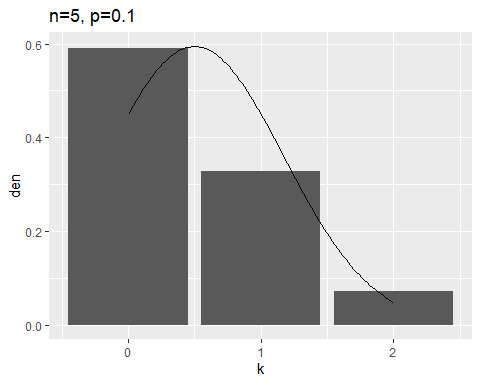
##   
## [[4]]



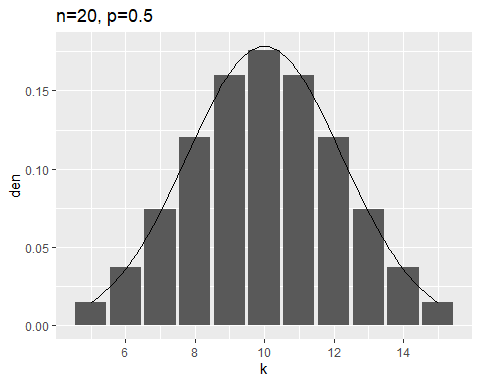
##   
## [[5]]



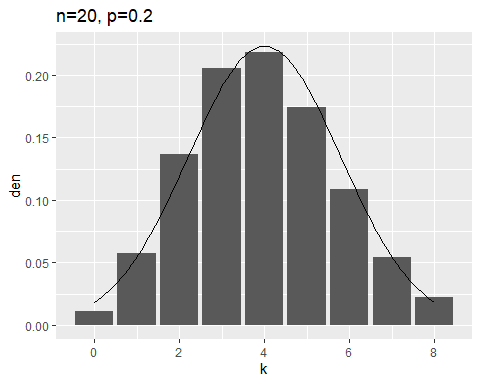
##   
## [[6]]



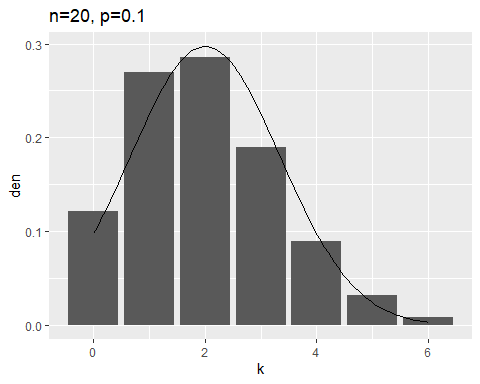
##   
## [[7]]



##   
## [[8]]



##   
## [[9]]



## Question 4

One concern regarding research publications that report p-values is that there is a publication bias toward papers that report p-values under a threshold, often . If this standard is applied to many papers investigating similar phenomena, papers with large p-values may be suppressed while papers with small p-values are published. This would create an inaccurate impression of the statistical significance.

This problem simulates the situation in which a large number of researchers each draw samples of a specified size from a population and count the number of successes in the sample. Each researcher tests the hypothesis that the number of successes in their sample is consistent with the null model that the number of successes has a binomial distribution with size equal to the sample size and the probability of success equal to a specified value .

You will simulate the results in the case in which the samples are in fact drawn from the null distribution, the binomial distribution with size equal to the sample size and the probability of success equal to .

### Question 4.a (5 points)

Assume each researcher uses the null model that their sample is a sample from a binomial distribution with the size parameter equal to the number of items in the sample and the parameter giving the probability success equal to the value . One way to define a p-value for this null hypothesis is to calculate the twice probability of a value less than or equal to the observed number of successes under the null hypothesis, to calculate twice the probability of a value greater than or equal to the observed number of successes under the null hypothesis, and to take the smaller of these to be the p-value.

Please complete the function template below to make a p-value calculator for the researchers. The function should return the p-value of this test of the null hypothesis if the argument “obs” is the observed number of successes, the parameter “size” is the sample size, and the parameter “pr” is the probability of success under the null hypothesis. For future use, you may wish to implement this so that, given a vector of observations, the function returns a vector of the corresponding p-values.

Please apply your function to the case of an observation of 25 successes in a sample of size 100 with hypothesized probability of success equal to .3 and to an observation of 30 successes in a sample of size 100 with hypothesized probability of success equal to .3

# p-value calculator  
  
p.get<-function(obs,size,pr){  
 p.low<-2\*pbinom(obs,size,pr)  
 p.high<-2\*pbinom(obs-1,size,pr,lower.tail = FALSE)  
 p.mat<-cbind(p.low,p.high)  
 return(apply(p.mat,1,min))  
# return(min(p.low,p.high))  
}  
  
#with the lower.tail=TRUE, the equality is accounted for  
pbinom(0,10,.5)

## [1] 0.0009765625

#with the lower.tail=FALSE, the equality is not accounted for  
pbinom(10,10,.5, lower.tail = FALSE)

## [1] 0

#also notice this:  
pbinom(1,10,.5, lower.tail = FALSE)+pbinom(1,10,.5)

## [1] 1

p.get(c(25,30),100,.3)

## [1] 0.3262602 1.0753205

pbinom(30,100,.3)+pbinom(30,100,.3, lower.tail = FALSE)

## [1] 1

pbinom(30,100,.3)+pbinom(29,100,.3, lower.tail = FALSE)

## [1] 1.086784

pbinom(30,100,.3, lower.tail = FALSE)

## [1] 0.4508764

pbinom(29,100,.3, lower.tail = FALSE)

## [1] 0.5376603

#Refined version so values do not exceed 1  
p.get<-function(obs,size,pr){  
 probs.m<-cbind(2\*pbinom(obs,size,pr),  
 2\*pbinom(obs-1,size,pr,lower.tail=F))  
 return(apply(probs.m,1,function(a) min(c(1,a))))   
}  
  
p.get(c(25,30),100,.3)

## [1] 0.3262602 1.0000000

### Question 4.b (5 points)

Write a function with the arguments “n” for the number of researchers, “size” for the sample size, “pr” for the probability of success, and “p” for the p-value. The function should draw “n” samples from the binomial distribution with size equal “size” and probability of success equal to “pr”. It should return the number of p-values less that or equal to “p” for the test above of the null hypothesis that the sample comes from a binomial distribution with size equal “size” and probability of success equal to “pr”. (These are called *type 1 errors* at the significance level “p”.) Apply “replicate” to this function 1000 times and calculate the mean number of type 1 errors for the sets of values

n=200,size=100,pr=.3,p=.05

n=200,size=500,pr=.3,p=.05

n=400,size=500,pr=.3,p=.01 n=400,size=1000,pr=.3,p=.01

What do the numbers computed represent in terms of the conclusion drawn by each of the 200 or 400 research teams? In particular, does a larger sample size protect a research team from a type 1 error?

### MDH Comment: p- values become more significant when there are larger sample sizes which makes p-values controversial. We can increase the sample size as much as we want to create significance that may not be there. True significant p-values should be able to be found on small sample sets.

#Function Definition  
sim<-function(n,size,pr,p){  
 samp<-rbinom(n,size,pr)  
 return(sum(p.get(samp,size,pr)<p))  
}  
  
set.seed(1234)  
mean(replicate(1000,sim(200,100,.3,.05)))

## [1] 7.532

mean(replicate(1000,sim(200,500,.3,.05)))

## [1] 9.012

mean(replicate(1000,sim(400,500,.3,.01)))

## [1] 3.376

mean(replicate(1000,sim(400,1000,.3,.01)))

## [1] 3.421