Problem Set 4

C. Durso

## 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 2018 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 9.0 [dataset]. Minneapolis, MN: IPUMS, 2019. <https://doi.org/10.18128/D010.V9.0>

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

### IPUMS Data

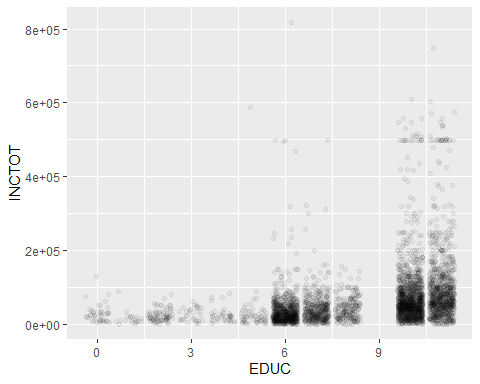
Read in the ipums data

dat<-read.csv("usa\_00016\_trim.csv")

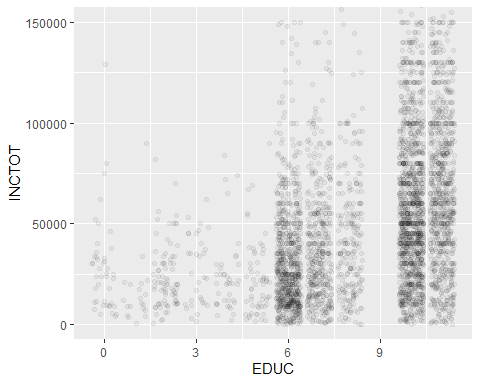
Trim to PUMAs that are predominantly in Denver. Restrict to respondents who are at least 25 years old with non-zero income. The variable “INCTOT” gives total annual individual income. The variable “EDUC” consists of ordered categories of amount of formal education, with 0 representing the least and 11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



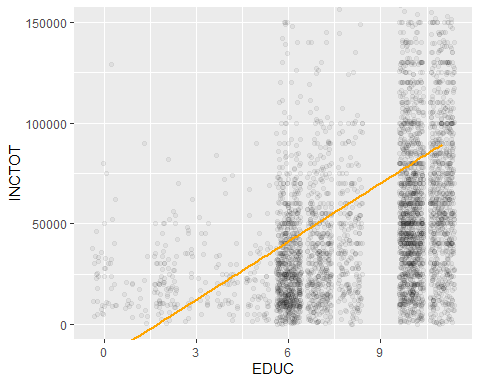
### Question 1.a (5 points)

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?

We would predict an increase of $9634 in the income category, if there were an increase of 1 in the EDUC category.

g <- g + geom\_smooth(method=lm,color="orange",se=FALSE)  
g

## `geom\_smooth()` using formula 'y ~ x'



lm(INCTOT~ EDUC, data = denver)

##   
## Call:  
## lm(formula = INCTOT ~ EDUC, data = denver)  
##   
## Coefficients:  
## (Intercept) EDUC   
## -16587 9634

### 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.

gd <- data.frame(denver%>%  
 group\_by(EDUC) %>%  
 summarise(INCTOT=mean(INCTOT)))

## `summarise()` ungrouping output (override with `.groups` argument)

gm <- data.frame(denver%>%  
 group\_by(EDUC) %>%  
 summarise(INCTOT=median(INCTOT)))

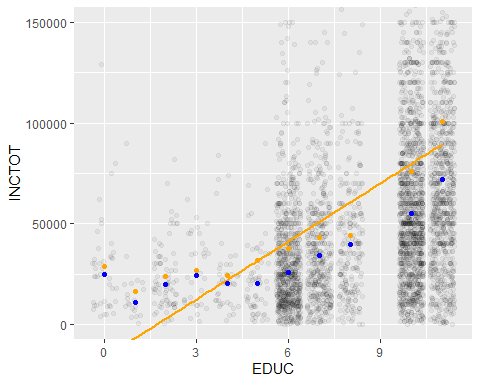
## `summarise()` ungrouping output (override with `.groups` argument)

gm

## EDUC INCTOT  
## 1 0 25000  
## 2 1 11000  
## 3 2 20000  
## 4 3 24400  
## 5 4 20400  
## 6 5 20410  
## 7 6 26000  
## 8 7 34050  
## 9 8 39850  
## 10 10 55000  
## 11 11 72000

g <- g + geom\_point(data=gd,color="orange")+geom\_point(data=gm,color="Blue")  
g

## `geom\_smooth()` using formula 'y ~ x'



### 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.

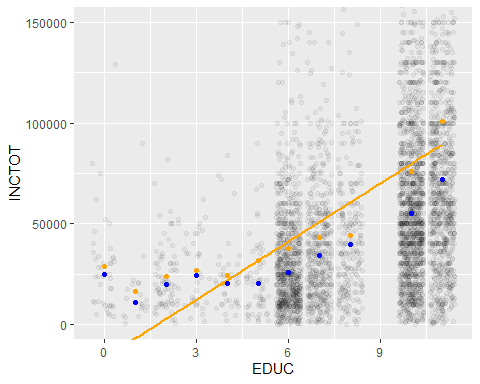
nlmFunc <- function(x){  
 return(sum(abs(gm-(9634\*x-16587))))  
}  
nlm(nlmFunc,p=9634)

## $minimum  
## [1] 348053  
##   
## $estimate  
## [1] 2.029352  
##   
## $gradient  
## [1] 0  
##   
## $code  
## [1] 1  
##   
## $iterations  
## [1] 11

g <- g + geom\_smooth(data=denver,method=nlm,color="blue",se=FALSE)  
g

## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Computation failed in `stat\_smooth()`:  
## argument "p" is missing, with no default



newFrame <- subset.data.frame(denver,EDUC<=5,select=c(EDUC,INCTOT))  
newFrame

## EDUC INCTOT  
## 2 2 10200  
## 4 3 45000  
## 18 2 10200  
## 23 5 4800  
## 37 5 2300  
## 47 1 4300  
## 57 5 4800  
## 73 5 4800  
## 84 5 18000  
## 88 5 8600  
## 90 3 45000  
## 99 2 10200  
## 104 5 4800  
## 110 2 10200  
## 111 5 2300  
## 115 2 10200  
## 116 3 45000  
## 161 1 600  
## 191 5 1600  
## 196 2 21100  
## 206 2 30400  
## 207 3 35400  
## 219 4 9000  
## 220 0 28000  
## 228 4 29600  
## 229 4 74000  
## 230 2 20400  
## 251 4 34000  
## 264 5 26000  
## 302 4 20000  
## 309 2 4800  
## 319 2 24000  
## 320 2 6000  
## 334 3 17000  
## 341 4 15500  
## 349 0 45900  
## 350 0 37500  
## 360 5 24000  
## 413 3 3000  
## 443 3 50100  
## 457 2 8400  
## 484 3 4800  
## 485 0 10800  
## 492 4 25000  
## 496 1 10300  
## 497 1 20800  
## 526 2 15000  
## 543 5 9000  
## 551 0 9500  
## 563 3 37300  
## 593 0 9000  
## 597 3 33600  
## 603 2 33700  
## 615 2 5700  
## 658 3 48040  
## 659 3 1500  
## 675 0 50000  
## 699 5 28000  
## 707 4 25000  
## 724 5 11200  
## 778 4 12000  
## 789 0 25000  
## 904 5 14000  
## 955 4 20100  
## 960 3 13200  
## 965 2 27300  
## 991 1 20000  
## 993 4 40000  
## 1031 4 21000  
## 1039 2 18000  
## 1042 5 36200  
## 1058 5 13100  
## 1064 0 33600  
## 1067 2 13800  
## 1068 2 27000  
## 1093 0 4800  
## 1101 4 15000  
## 1104 4 12000  
## 1121 2 12000  
## 1147 2 9600  
## 1173 2 24000  
## 1180 5 12300  
## 1185 4 10000  
## 1186 4 72000  
## 1231 5 14000  
## 1243 5 20820  
## 1244 5 3600  
## 1247 5 69000  
## 1263 5 54000  
## 1269 2 39000  
## 1358 2 15700  
## 1359 5 38000  
## 1405 5 9000  
## 1416 4 11700  
## 1419 5 10000  
## 1422 2 33700  
## 1441 2 50000  
## 1442 2 30000  
## 1455 0 75000  
## 1472 1 33800  
## 1508 0 33400  
## 1513 4 28000  
## 1542 5 14400  
## 1592 3 62000  
## 1593 2 20000  
## 1608 2 33600  
## 1636 2 56000  
## 1637 2 18000  
## 1639 1 15000  
## 1670 2 50000  
## 1683 2 30000  
## 1684 2 30000  
## 1693 1 9000  
## 1700 5 20000  
## 1726 2 10000  
## 1733 5 40000  
## 1747 2 17300  
## 1774 3 7900  
## 1775 3 34500  
## 1789 0 28000  
## 1797 5 37500  
## 1803 2 20000  
## 1823 2 40000  
## 1846 5 21800  
## 1848 5 35000  
## 1869 5 4500  
## 1874 5 24000  
## 1885 4 25000  
## 1887 2 20800  
## 1908 4 20400  
## 1947 5 5300  
## 1989 0 4000  
## 1992 5 23000  
## 1993 5 45000  
## 2006 3 12000  
## 2018 1 90000  
## 2024 0 35700  
## 2025 2 16800  
## 2031 2 16000  
## 2032 2 50000  
## 2047 4 83700  
## 2056 5 25300  
## 2057 4 8000  
## 2058 5 9000  
## 2077 3 46600  
## 2085 0 24000  
## 2093 5 30000  
## 2101 4 20000  
## 2102 4 9000  
## 2132 2 32000  
## 2134 4 20300  
## 2135 3 10500  
## 2137 3 12400  
## 2195 0 52000  
## 2288 4 25000  
## 2299 5 67000  
## 2302 0 9460  
## 2331 2 28000  
## 2340 2 15000  
## 2366 2 9800  
## 2367 2 27000  
## 2379 2 200  
## 2395 2 8400  
## 2396 0 5400  
## 2422 4 24000  
## 2428 5 588000  
## 2447 1 17300  
## 2459 4 11000  
## 2469 1 22000  
## 2470 5 10300  
## 2471 3 10000  
## 2486 5 90000  
## 2487 5 21310  
## 2511 0 4800  
## 2512 0 32000  
## 2530 1 7800  
## 2540 2 20000  
## 2555 3 53000  
## 2560 0 11500  
## 2578 1 3600  
## 2588 3 35000  
## 2594 0 31210  
## 2601 4 10000  
## 2602 2 11400  
## 2624 2 54000  
## 2642 2 82000  
## 2665 3 10900  
## 2779 0 18000  
## 2790 2 2100  
## 2797 2 10000  
## 2798 3 9600  
## 2807 0 30000  
## 2838 2 17200  
## 2840 5 22000  
## 2853 2 18000  
## 2855 2 20000  
## 2889 1 20200  
## 2890 3 35000  
## 2934 2 42000  
## 2953 2 35000  
## 2954 2 70000  
## 2955 2 25000  
## 2956 4 4500  
## 2957 2 52000  
## 2962 5 22100  
## 2963 4 21600  
## 2964 5 17300  
## 2984 5 9600  
## 2997 0 11400  
## 3012 3 25400  
## 3017 4 29900  
## 3024 0 129000  
## 3025 0 9600  
## 3028 2 15500  
## 3039 5 39000  
## 3042 0 24000  
## 3043 5 30800  
## 3057 0 9000  
## 3062 3 40200  
## 3075 5 37000  
## 3076 5 20000  
## 3083 0 32900  
## 3091 4 33800  
## 3094 4 22300  
## 3124 4 29700  
## 3154 0 80000  
## 3155 0 30000  
## 3157 2 20800  
## 3175 0 25190  
## 3184 3 23000  
## 3194 4 2000  
## 3212 0 24000  
## 3213 0 10300  
## 3217 3 23400  
## 3247 2 23000  
## 3276 0 18000  
## 3287 5 22600  
## 3292 5 55000  
## 3296 5 21900  
## 3308 3 9200  
## 3309 0 61800  
## 3313 2 18700  
## 3332 0 20000  
## 3349 1 13000  
## 3350 2 9000  
## 3366 5 35000  
## 3422 2 2200  
## 3459 2 50000  
## 3567 0 30000  
## 3591 4 10300  
## 3614 2 30000  
## 3634 1 6700  
## 3635 1 11700  
## 3657 0 22600  
## 3664 5 12400  
## 3777 1 10000  
## 3778 1 9000  
## 3820 2 20000  
## 3838 0 7500  
## 3841 4 65000  
## 3860 2 30000  
## 3861 0 40000  
## 3888 2 36000  
## 3891 2 45000  
## 3892 2 12000  
## 3910 5 1700  
## 3967 3 21000  
## 3979 2 17000  
## 3995 2 8600  
## 4005 5 12400  
## 4105 4 4800  
## 4168 2 22000  
## 4169 1 4000

g2<-ggplot(newFrame,aes(x=EDUC,y=INCTOT))+geom\_jitter(alpha=.05)+coord\_cartesian(ylim =c(0,1e5))+ geom\_smooth(method=lm,color="orange",se=FALSE)  
  
lm(INCTOT~ EDUC, data = newFrame)

##   
## Call:  
## lm(formula = INCTOT ~ EDUC, data = newFrame)  
##   
## Coefficients:  
## (Intercept) EDUC   
## 23358 1101

gd2 <- data.frame(newFrame%>%  
 group\_by(EDUC) %>%  
 summarise(INCTOT=mean(INCTOT)))

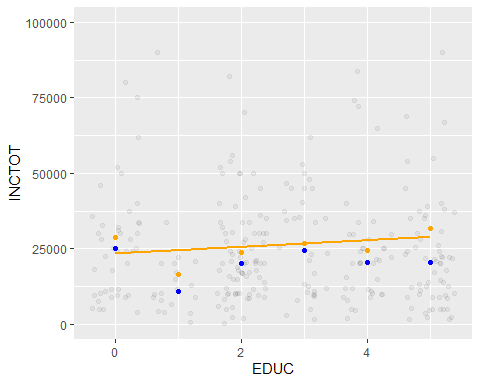
## `summarise()` ungrouping output (override with `.groups` argument)

gm2 <- data.frame(newFrame%>%  
 group\_by(EDUC) %>%  
 summarise(INCTOT=median(INCTOT)))

## `summarise()` ungrouping output (override with `.groups` argument)

g2 <- g2 + geom\_point(data=gd2,color="orange")+geom\_point(data=gm2,color="Blue")  
g2

## `geom\_smooth()` using formula 'y ~ x'



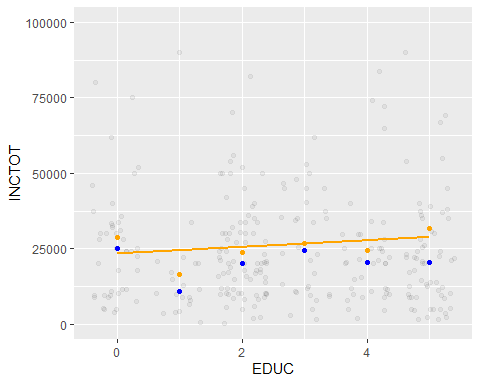
nlmFunc2 <- function(x){  
 return(sum(abs(gm2-(9634\*x-16587))))  
}  
nlm(nlmFunc,p=1101)

## $minimum  
## [1] 348053  
##   
## $estimate  
## [1] 1.800236  
##   
## $gradient  
## [1] 0  
##   
## $code  
## [1] 1  
##   
## $iterations  
## [1] 6

g2 <- g2 + geom\_smooth(data=newFrame,method=nlm,color="blue",se=FALSE)  
g2

## `geom\_smooth()` using formula 'y ~ x'  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Computation failed in `stat\_smooth()`:  
## argument "p" is missing, with no default



stddevs <- data.frame(denver%>%  
 group\_by(EDUC) %>%  
 summarise(INCTOT=sqrt(var(INCTOT))))

## `summarise()` ungrouping output (override with `.groups` argument)

stddevs

## EDUC INCTOT  
## 1 0 23843.50  
## 2 1 19061.66  
## 3 2 15866.89  
## 4 3 17163.22  
## 5 4 19182.59  
## 6 5 75194.12  
## 7 6 53121.88  
## 8 7 44333.22  
## 9 8 30585.55  
## 10 10 81118.58  
## 11 11 101827.94

### 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.

### 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.

It does not look like linear models are ideal in either situation. The residuals of the means and medians relative to the least squares line seem to follow a pattern, rather than be arbitrarily scattered. As far as the measures of center go, overall the medians are closer to the Least Squares line than the means, probably because of outlying values. The sample standard deviations within each category get progressively larger, generally speaking. At the same time, the predicted change in INCTOT increases also. The slope of the least squares line is when EDUC <= 5 is $1101 for each additional EDUC category. For all categories it is $9634.

## Question 2

This question uses the distribution means and variances to compare Poisson distributions and binomial distributions to corresponding Normal distributions.

### Question 2.a (5 points)

The 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

Finally:

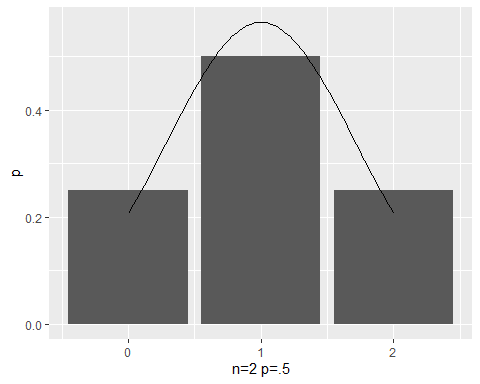
, which is the variance of the Poisson disrtibution.

### Question 2.b (5 points)

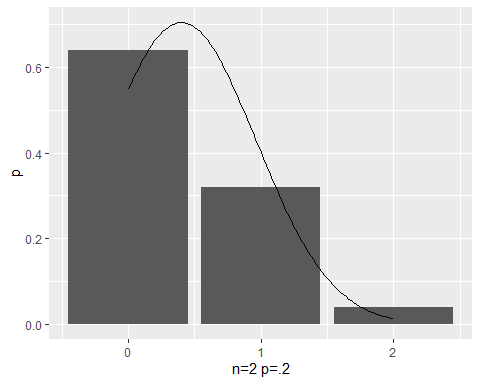
In this question, you will compare binomial distributions with Normal distributions having related means and variances.

For each pair with and , please create a column plot over the non-negative integers less than or equal to for which the quantile function of the distribution is in . 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

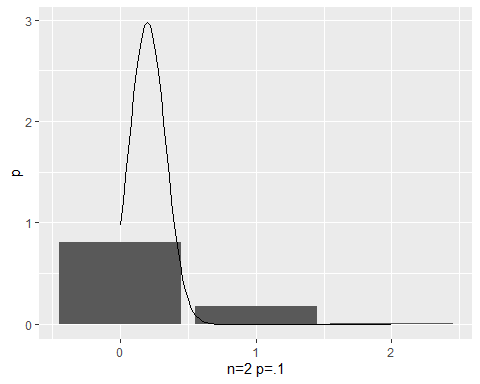
binom1 <- data.frame(x=0:2,y=dbinom(0:2,2,.5))  
ggplot(data=binom1, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=2 p=.5",y="p")+stat\_function(fun=dnorm,args=list(mean=1,sd=.707))



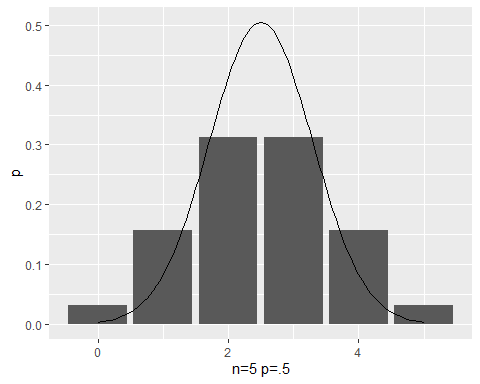
binom2 <- data.frame(x=0:2,y=dbinom(0:2,2,.2))  
ggplot(data=binom2, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=2 p=.2",y="p")+stat\_function(fun=dnorm,args=list(mean=.4,sd=.566))



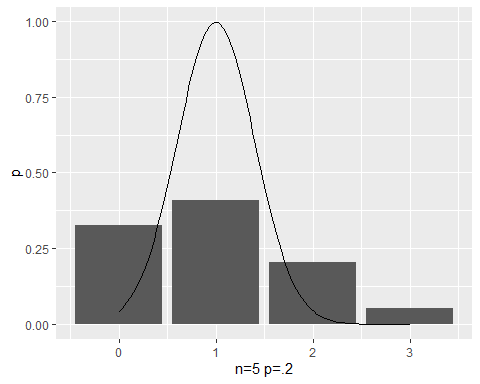
binom3 <- data.frame(x=0:2,y=dbinom(0:2,2,.1))  
  
ggplot(data=binom3, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=2 p=.1",y="p")+stat\_function(fun=dnorm,args=list(mean=.2,sd=.134))



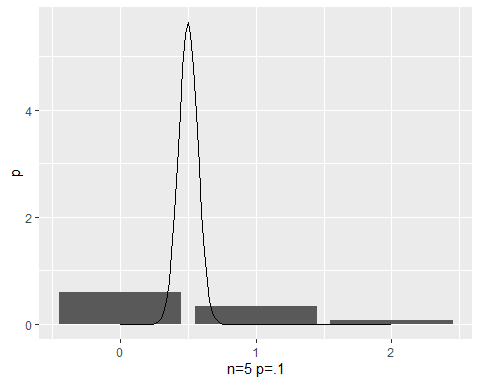
binom4 <- data.frame(x=0:5,y=dbinom(0:5,5,.5))  
ggplot(data=binom4, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=5 p=.5",y="p")+stat\_function(fun=dnorm,args=list(mean=2.5,sd=.79))



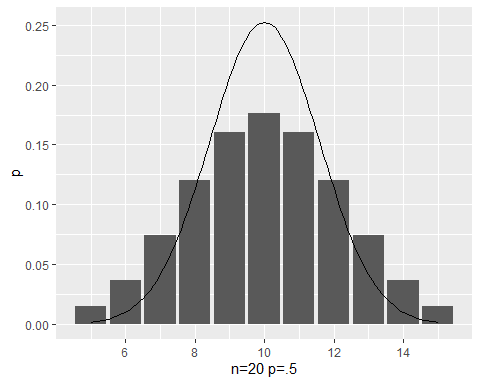
binom5 <- data.frame(x=0:3,y=dbinom(0:3,5,.2))  
ggplot(data=binom5, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=5 p=.2",y="p")+stat\_function(fun=dnorm,args=list(mean=1,sd=.4))



binom6 <- data.frame(x=0:2,y=dbinom(0:2,5,.1))  
ggplot(data=binom6, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=5 p=.1",y="p")+stat\_function(fun=dnorm,args=list(mean=.5,sd=.0707))



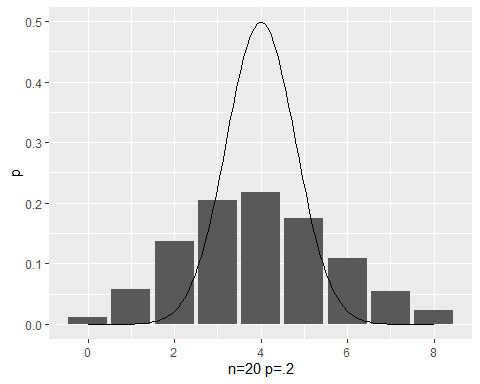
binom7 <- data.frame(x=5:15,y=dbinom(5:15,20,.5))  
ggplot(data=binom7, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=20 p=.5",y="p")+stat\_function(fun=dnorm,args=list(mean=10,sd=1.58))



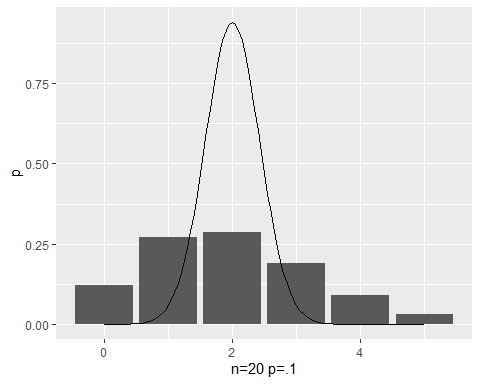
binom8 <- data.frame(x=0:8,y=dbinom(0:8,20,.2))  
binom8

## x y  
## 1 0 0.01152922  
## 2 1 0.05764608  
## 3 2 0.13690943  
## 4 3 0.20536414  
## 5 4 0.21819940  
## 6 5 0.17455952  
## 7 6 0.10909970  
## 8 7 0.05454985  
## 9 8 0.02216088

ggplot(data=binom8, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=20 p=.2",y="p")+stat\_function(fun=dnorm,args=list(mean=4,sd=.8))



binom9 <- data.frame(x=0:5,y=dbinom(0:5,20,.1))  
ggplot(data=binom9, aes(x=x,y=y)) + geom\_bar(stat="identity")+labs(x="n=20 p=.1",y="p")+stat\_function(fun=dnorm,args=list(mean=2,sd=.424))

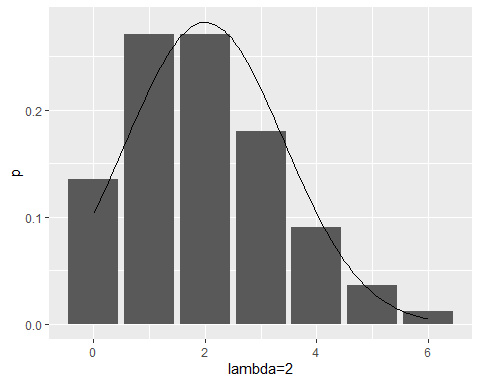


### Question 2.c (5 points)

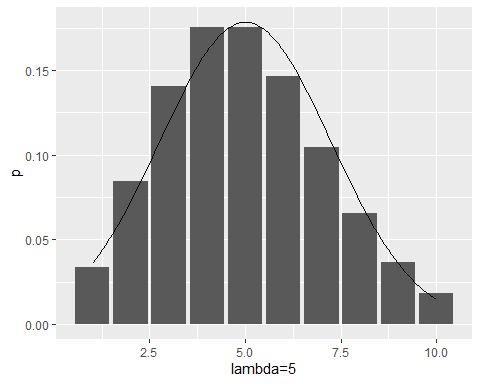
In this question, you will compare Poisson distributions with Normal distributions having related means and variances.

For values , please create a column plot over the non-negative integers for which the quantile function of the distribution is in . 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 mean and variance both equal to . Please label the plots with the corresponding value of .

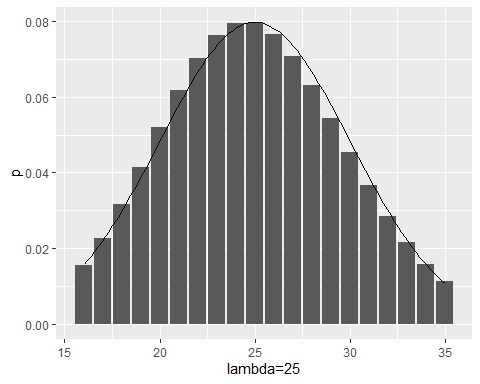
poisson1 <- data.frame(x=0:6,y=dpois(0:6,2))  
ggplot(data=poisson1, aes(x=x,y=y))+ geom\_bar(stat="identity")+labs(x="lambda=2",y="p")+stat\_function(fun=dnorm,args=list(mean=2,sd=1.414))



poisson2 <- data.frame(x=1:10,y=dpois(1:10,5))  
ggplot(data=poisson2, aes(x=x,y=y))+ geom\_bar(stat="identity")+labs(x="lambda=5",y="p")+stat\_function(fun=dnorm,args=list(mean=5,sd=2.236))



poisson3 <- data.frame(x=16:35,y=dpois(16:35,25))  
ggplot(data=poisson3, aes(x=x,y=y))+ geom\_bar(stat="identity")+labs(x="lambda=25",y="p")+stat\_function(fun=dnorm,args=list(mean=25,sd=5))



## Question 3

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 3.a (5 points)

Assume each researcher uses the null model that their sample is 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){min((2\*pbinom(obs,size,pr)),(2\*(1-pbinom(obs-1,size,pr))))  
}  
  
  
p.get(25,100,.3)

## [1] 0.3262602

p.get(30,100,.3)

## [1] 1.075321

### Question 3.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?

The larger the sample size, the less frequently, proportionally speaking, that Type 1 errors will occur.

#p.get(40,100,.3)  
#p.get(20,100,.3)  
  
n=200  
size=100  
pr=.3  
p=.05  
  
pvals1 <- c(replicate(n,p.get(rbinom(1,size,pr),size,pr),simplify=FALSE))  
length(which(pvals1<=p))

## [1] 7

size2=500  
pvals2 <- c(replicate(n,p.get(rbinom(1,size2,pr),size2,pr),simplify=FALSE))  
length(which(pvals2<=p))

## [1] 5

n2=400  
p2=.01  
pvals3 <- c(replicate(n2,p.get(rbinom(1,size2,pr),size2,pr),simplify=FALSE))  
length(which(pvals3<=p2))

## [1] 1

size3=1000  
pvals4 <- c(replicate(n2,p.get(rbinom(1,size3,pr),size3,pr),simplify=FALSE))  
length(which(pvals4<=p2))

## [1] 3