

# Bayesian Data Analysis

Anthony R. Colombo, Dr. Paul Marjoram

2022-08-11



# Contents

<b>About</b>	<b>5</b>
Independent study . . . . .	5
Supervised learning . . . . .	5
<b>1 Fundamentals of Bayesian Inference</b>	<b>7</b>
1.1 General notation for statistical inference . . . . .	7
1.2 Bayesian inference . . . . .	8
Prediction . . . . .	9
Likelihood . . . . .	9
Subjectivity and Objectivity . . . . .	9
1.3 Exercises . . . . .	10
Simulation 100 times . . . . .	20
<b>2 Single parameter models</b>	<b>25</b>
2.1 Estimating a probability from binomial data . . . . .	25
2.2 Posterior as a compromise between data and prior information .	25
2.3 Summarizing the posterior inference . . . . .	26
2.4 Informative prior distributions . . . . .	26
2.5 Normal distribution with known variance . . . . .	27
2.6 Other standard single-parameter models . . . . .	29
2.7 Noninformative prior distributions . . . . .	31
2.8 Exercises . . . . .	32
<b>3 Chapter 3</b>	<b>71</b>

<b>Exercises</b>	<b>73</b>
3.1 Question 1 . . . . .	73
Question 2 . . . . .	74
Question 3 . . . . .	76
Question 4 (independent binomial) . . . . .	80
Question 5 . . . . .	87
Question 7 . . . . .	97
question 8 . . . . .	98
question 9 . . . . .	100
Question 11 . . . . .	102
<b>4 Asymptotics and connections to non-Bayesian approaches</b>	<b>107</b>
4.1 Normal approximations to the posterior distribution . . . . .	107
4.2 Large-sample theory . . . . .	108
4.3 Frequency evaluations of Bayesian inferences . . . . .	109
4.4 Exercises . . . . .	111
<b>5 Blocks</b>	<b>115</b>
5.1 Equations . . . . .	115
5.2 Theorems and proofs . . . . .	115
5.3 Callout blocks . . . . .	115
<b>6 Sharing your book</b>	<b>117</b>
6.1 Publishing . . . . .	117
6.2 404 pages . . . . .	117
6.3 Metadata for sharing . . . . .	117
<b>7 Parts</b>	<b>119</b>

# About

This is an independent study of the text **Bayesian Data Analysis** by Andrew Gelman and the more introductory text **A First Course In Bayesian Statistical Methods** by Peter D. Hoff. We will read through most of the chapters and typeset the major definitions. The Gelman text can be difficult, and for difficult chapters, we will lean more on the Hoff textbook.

## Independent study

We will typeset each chapter of the Gelman text book, unless the chapter is too difficult, then we will use the introductory text **A First Course In Bayesian Statistical Methods** by Hoff. Each chapter will summarize the definitions, and attempt several problems selected.

## Supervised learning

Dr. Paul Marjoram will supervise the learning and have a general oversight to the learning process.



# Chapter 1

## Fundamentals of Bayesian Inference

The first few chapters of Gelman's text are introductory, and we attempt to highlight the key definitions and summarize each chapter. At the end of each chapter we attempt several problems. Probability and inference is defined using three steps

1. setting up the full probability model for a joint distribution for all observable and unobservable quantities.
2. Conditioning on observed data: computing the appropriate *posterior* distribution, the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
3. Evaluating the fit of the model.

### 1.1 General notation for statistical inference

There are two different kinds of estimands, the first are potentially observable quantities, such as future observations of a process, and the second are quantities that are not directly observable, namely the parameters that govern a process being investigated.

#### Exchangeability

One key assumption is that the  $n$  values  $y_i$  are regarded as *exchangeable*, meaning that the uncertainty can be expressed as a joint probability  $p(y_1, \dots, y_n)$  that is invariant to permutations of indexes. Often times the exchangeable distribution is modeled as *iid*.

## Explanatory variables

It is common to have observations on each unit which have non-random variables called *explanatory variables* or *covariates*. The explanatory variables are usually denoted by  $X$ . However treating  $X$  as random then exchangeability can be extended  $(x, y)_i$  which is invariant to permutations of the indexes. Further, it is always appropriate to assume exchangeability of  $y$ , conditioned on sufficient information of  $X$ , where the indexes can be thought of as randomly assigned. It follows that if two units have the same value of  $x$ , then the distributions of  $y$  are the same.

## Hierarchical modeling

for a model across patients across different cities, we can assume exchangeability to patients within a city. Further conditioned on the explanatory variables at the individual, the conditional distribution given these explanatory variables would be exchangeable.

## 1.2 Bayesian inference

The prior,  $p(\theta)$ , and the sampling distribution, or the *data distribution*,  $p(y|\theta)$  is related to the joint distribution by

$$p(\theta, y) = p(\theta)p(y|\theta)$$

Where using Bayes' rule the posterior distribution

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \quad (1.1)$$

Where  $p(y) = \int p(\theta)p(y|\theta)d\theta$ , or a sum in discrete case. An equivalent form of (1.1) is the *unnormalized posterior density* given as

$$p(\theta|y) \propto p(\theta)p(y|\theta) \quad (1.2)$$

Note that  $p(y|\theta)$  is taken as a function of  $\theta$ , not of  $y$ .

## Prediction

Inferences about an unknown *observable* variable, are called predictive inferences. Before the data  $y$  are considered, the distribution of the unknown, observable,  $y$  is

$$p(y) = \int p(y, \theta) d\theta = \int p(\theta)p(y|\theta)d\theta$$

this is defined as the marginal distribution of  $y$ , and also called *prior predictive distribution*. Prior refers that the data is not conditional on any previous observation, and predictive refers to the data being observable.

The *posterior predictive distribution* is conditional on the observed  $y$ , but is predictive because it is predicting observable values.

$$\begin{aligned} p(\hat{y}|y) &= \int p(\hat{y}, \theta|y) d\theta \\ &= \int p(\hat{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\hat{y}|\theta)p(\theta|y)d\theta \end{aligned} \tag{1.3}$$

## Likelihood

The data  $y$  affects the posterior inference only through (1.2) likelihood function  $p(y|\theta)$  which is regarded as a function of  $\theta$  for fixed  $y$ . The *likelihood function* is defined as  $p(y|\theta)$ , and the *likelihood principle* is for any given sample, and any two likelihood models  $p(y|\theta)$ , two models with the same likelihood will have the same inference for  $\theta$ .

## Subjectivity and Objectivity

The frequentist models, MLEs, have subjectivity in their assumptions because they rely on long sequence of identical trials, that are iid. The Bayesian model relies on the prior distribution. If any experiment is repeatable and can replicated, the prior distribution can be estimated from the data themselves and the analysis is more ‘objective’. Replication increases objectivity of a given model. However the Bayesian approach allows for (1) the ability to combine information from multiple sources (allowing for greater objectivity) and (2) more encompassing by accounting for uncertainty about the unknowns in a statistical problem.

It is important to include as much background information as possible

### 1.3 Exercises

1. Suppose for  $\theta = 1$ , then  $y \sim N(1, \sigma)$ , and if  $\theta = 2, y \sim N(2, \sigma)$ . Where  $P(\theta = 1) = P(\theta = 2) = 0.5$ .

- (a) For  $\sigma = 2$ ; we must write the formula for the pdf of  $y$ .  $p(y) = \sum_{\theta} p(y|\theta)p(\theta) = (1/2)N(1, \sigma^2) + (1/2)N(2, \sigma^2)$  as the marginal density.

```
fy<-function(y) {return(0.5*dnorm(y,mean=1,sd=2)+0.5*dnorm(y,2,SD=2))}
```

(b)  $P(\theta=1 | y=1) = \frac{p(\theta=1)p(y|\theta=1)}{\int p(y|\theta)p(\theta)d\theta} = \frac{(1/2)N(1, 4)}{\int (1/2)N(1, 4) + (1/2)N(2, 4)}$

```
dy<-function(y){ return( (1/2)*dnorm(y,mean=1,SD=2)/fy(y)) }
dy(1)
```

```
## [1] 0.5312094
```

4. twelve games with point spread of 8 points.

- (a) Using relative frequency,  $P(\text{favorite wins} | \text{point spread} = 8) = 0.67$ .  
 $P(\text{favorite wins by at least } 8 | \text{point spread} = 8) = 0.42$ .  
and  $P(\text{fav. wins by at least } 8 | \text{spread} = 8, \text{ favorite team wins}) = 0.62$ .

```
spread<-8
```

```
## outcome of the games favor score - underdog score
games<-c(-7,-5,-3,1,6,7,13,15,16,20,21)
## frequentist approach
fav.wins<- mean(games>0)
message(paste0("(frequentist): fav wins: ", round(fav.wins,2)))
```

```
## (frequentist): fav wins: 0.67
```

```
fav.by.8<- mean((games>8))
message(paste0("(frequentist): fav wins by 8: ", round(fav.by.8,2)))
```

```
## (frequentist): fav wins by 8: 0.42
```

```
##  $P(\text{fav. wins} > 8 | \text{fav. wins}) = P(\text{fav. wins} > 8, \text{fav. wins}) / P(\text{fav. wins})$ 
cond<-sum(games>8 & games>0)/sum(games>0)
c<-fav.by.8/fav.wins
message(paste0("(frequentist): fav wins by 8 given fav. wins: ", round(cond,2)))
```

```
## (frequentist): fav wins by 8 given fav. wins: 0.62
```

- (b) now we assume a normal distribution with  $d|x \sim N(-1.25, 10.10)$ . So  $P(d > -x) = P(Z\sigma + \mu > -x) = P(Z > -x - \mu/\sigma)$
- (c) Probablity fav team wins is 0.75
- (ii) fav team wins by 8 (beats the spread) is 0.45, we expect this to be 0.5 (the middle of the normal distribution because we centered on the spread)
- (iii)  $P(\text{wins by 8} | \text{favorite team wins}) = P(\text{favorite team wins} | \text{wins by 8})P(\text{wins by 8})/P(\text{favorite team wins}) = P(\text{wins by 8})/P(\text{fav. team wins})$  since the conditional prob. =1 given the favorite team wins. The prob. that they win by at least 8 is 0.6.

```
## part b
d<-games-8
sample.mean <-mean(d)
sample.sd<-sd(d)
## assume d/x ~ N(0,10.10)
fav.wins.norm<- 1-pnorm(-8,mean=sample.mean,sd=sample.sd)
message(paste0("(normal): fav wins: ", round(fav.wins.norm,2)))

## (normal): fav wins: 0.75

fav.by.8.norm<-1-pnorm(0,mean=sample.mean,sd=sample.sd)
message(paste0("(normal): fav wins by 8: ", round(fav.by.8.norm,2)))

## (normal): fav wins by 8: 0.45

##  $Pr(\text{Wins by 8} | \text{Fav. wins}) = P(\text{Fav. wins} | \text{wins by 8})P(\text{wins by 8}) / P(\text{fav. wins})$ 
##  $P(\text{Fav. wins} | \text{wins by 8}) = 1$ 
cond.norm<-fav.by.8.norm/fav.wins.norm
message(paste0("(normal): fav wins by 8 given fav. wins: ", round(cond.norm,2)))

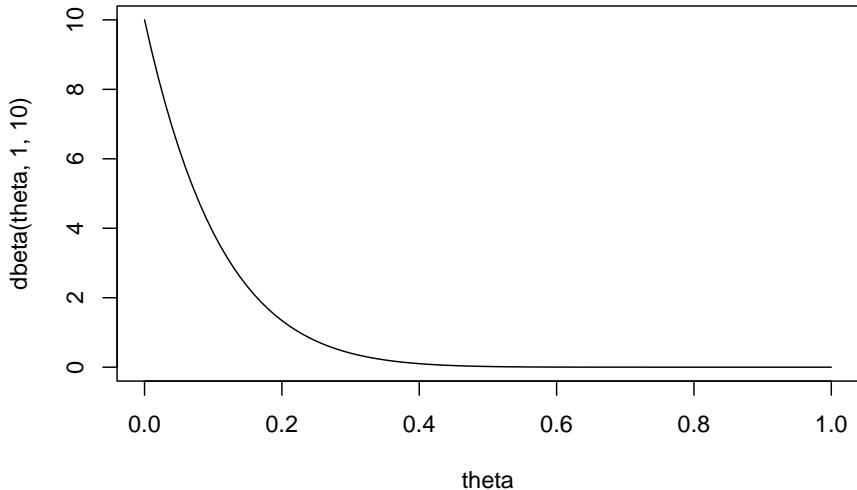
## (normal): fav wins by 8 given fav. wins: 0.6
```

5. We need to estimate the probability that there is at least one congressional election that is tied in the next U.S. election. There are 435 senate elections.

- (a) The parameters of interest are  $\theta_i$  the true probability that the election is tied. We can let the *prior*  $\theta \sim Beta(\alpha, \beta)$ . The *likelihood* is  $y|\theta_i \sim Binomial(435, \theta_i) = \theta^{\sum y_i} (1 - \theta)^{435 - \sum y_i}$  follows a Binomial distribution (ignoring the binomial coefficient) where we assume each election is independent. Hence the posterior for theta

$f(\theta|y) \sim Beta(\sum y_i + \alpha, n - \sum y_i + \beta)$ . where  $\alpha, \beta$  are set to 1 for the uniform prior. For this case we set  $\alpha, \beta$  equal to 1, 10 which has a prior mean of 0.09.

```
theta=seq(from=0,to=1,by=.01)
plot(theta,dbeta(theta,1,10),type='l')
```



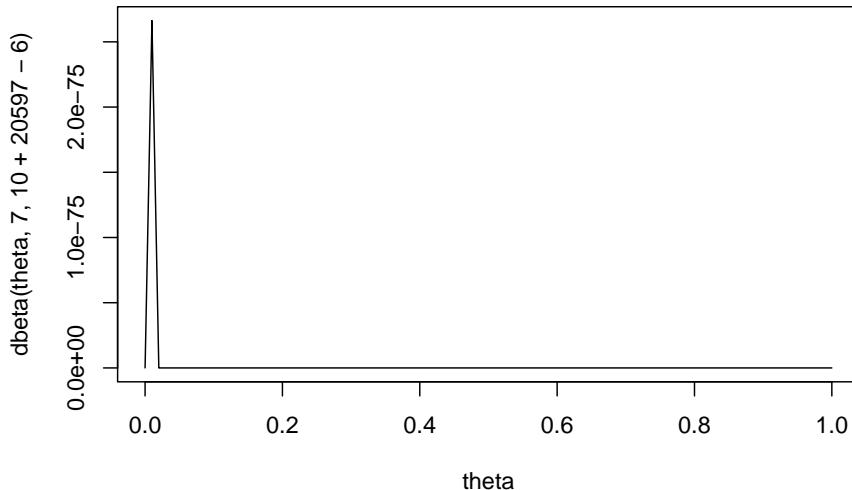
- (b) In the period of 1900-1992, there were 20,597 elections, out of which 6 were decided by less than 10 votes, and 49 were decided by less than 100 votes.

we can estimate the probability of a tie to be less than 6/20,597 and bounded by 49/20,597. So for the Binomial trials the sum of the successes is 6, and n=20,597, so the posterior could be  $\theta|y \sim Beta(1 + 6, 10 + 20,597 - 6)$  is the posterior for  $\theta$ . This assumes that 10 votes is within the neighborhood of an election tie.

The question asks to compute at least one election tie, from a total of 435 elections. This follows a  $Binomial(435, \hat{\theta})$ . Where we use the posterior mean to estimate  $\theta$ . The posterior mean using the  $Beta(7, 20601)$  yields a mean of  $\hat{\theta} = \frac{7}{20608} = 3.4e - 04$  as the posterior mean.

Then the probability that at least 1 election is tied, from 435 total elections will follow a  $Binomial(435, \hat{\theta})$ , where we can use the posterior distribution for  $\theta|y$  in the Binomial likelihood  $P(X \geq 1|\hat{\theta}) = 1 - P(X \leq 0|\hat{\theta})$  which has a probability of 0.14 of at least 1 election tie.

```
# the posterior for theta is Beta(1+6, 10+20597-6)
plot(theta, dbeta(theta, 7, 10+20597-6), type='l')
```



```
## posterior mean is 7/(20601)
## then P(X>=1) = 1-P(X<=0 | p)
1-pbinom(0, 435, prob=7/20608)
```

```
## [1] 0.1373819
```

9. A clinic has three doctors. Patients come into the clinic at random, starting at 9 a.m. according to a Poisson process, with a time parameter,  $t$ , of 10 minutes; that is after opening the first patient appears follows an exponential distribution with average waiting time of 10 minutes. Then the next patient arrives with a waiting time of an expected 10 minutes as iid exponential distribution. After a patient arrives, the patient waits until a doctor is available, and the doctor visits a patient uniformly between 5-20 minutes. The clinic stops admitting patients at 4 pm, and closes after the last patient is completed with the visit.

- (a) Simulate this process once. how many patients visited the office? how many had to wait for a doctor? what was the average wait? when did office close?

```

#### waiting time for a new patient to arrive in the clinic
#####
# patientList is the data frame of all patients
# closeTime is the time to stop admitting (420 minutes)
# currentPatient Number
# current time is the running total of time
newPatientArrival<-function(patientList,
                           closeTime=timeToClose,
                           waitTime,
                           visitTime,
                           currentPatientNumber=0,
                           currentTime,
                           assignedDoctor="none",
                           completionTime=0){
  # waiting time for next patient
  patientTime<-round(rexp(1,rate=1/10),2)
  # current time of existing patients
  current<-max(patientList$currentTime)
  ## the clinic stops admitting patients at 4pm
  if( (current+patientTime)<=closeTime){
    ## in minutes
    newPatient<-createPatientChart(currentPatientNumber,patientTime,waitTime,visitTime,
                                    closeTime)
  }else{
    newPatient<-createPatientChart(currentPatientNumber,patientTime,waitTime,visitTime)
  }
  return(newPatient)
}
#####

computeWaitTime<-function(doctors=NULL,
                            patientList=NULL,
                            patientID=1){
  ## need to compute visiting time (booked)
  ## next time available
  ## required input current time for a specific doctor/patient ?
  # patient time (minutes)

  ## FIX ME: it is grabbing 2 patient IDs?
  currentTime<-patientList$currentTime[which(patientList$patient==patientID)]
  visitTime<-runif(1,min=5,max=20) ## minutes

  if(any(doctors$nextTimeAvail<currentTime)){

```

```

waitTime=0
assignedDr<-sample(doctors$dr[which(doctors$nextTimeAvail<currentTime)],1)
### current time + visitTime
nextAvailTime<- visitTime+currentTime+waitTime
## completion time for patient exit (closing time).
}else if(any(doctors$nextTimeAvail<currentTime)==FALSE){
  # all doctors are booked, no available doctors.
  # wait time is the difference between next available time (assuming all times are greater than
  waitTime<-min(doctors$nextTimeAvail-currentTime)
  assignedDr<-doctors$dr[which( (doctors$nextTimeAvail-currentTime)==min(doctors$nextTimeAvail))]
  if(length(assignedDr)>1){
    assignedDr<-assignedDr[1]
  }
  nextAvailTime<- visitTime+currentTime+waitTime ## completion time for patient to exit
}## if all doctors unavail
#print(assignedDr)
#print(currentTime)
## update doctor list
doctors [which(doctors$dr==assignedDr), 'visitingPatient']<-patientID
doctors [which(doctors$dr==assignedDr), 'nextTimeAvail']<-nextAvailTime
doctors [which(doctors$dr==assignedDr), 'currentTime']<-currentTime ## patient time
doctors [which(doctors$dr==assignedDr), 'visitTimeLength']<-visitTime
# flag avail to no.
doctors [which(doctors$dr==assignedDr), 'avail']<-'no'
## update patient list
patientList [which(patientList$patient==patientID), 'doctorWaitTime']<-waitTime
patientList [which(patientList$patient==patientID), 'doctorVisitTime']<-visitTime
patientList [which(patientList$patient==patientID), 'assignedDoctor']<-assignedDr
patientList [which(patientList$patient==patientID), 'completionTime']<-nextAvailTime
return(list(patient=patientList,doctor=doctors))
}

## creates a patient object
createPatientChart<-function(currentPatientNumber,arrivalTime,waitTime,visitTime,currentTime,assignedDoctor)
  patientID<-data.frame(patient=currentPatientNumber+1,
                         arrivalTime=arrivalTime,
                         doctorWaitTime=waitTime,
                         doctorVisitTime=visitTime,
                         currentTime=currentTime,
                         assignedDoctor=assignedDoctor,
                         completionTime=0)
  return(patientID)
}

```

```

updatePatientList<-function(patientList,patientID){
  patientList<-rbind(patientList,patientID)
  return(patientList)
}

updateTime<-function(currentTime,newTime=NULL,p1){
  p1$currentTime<-currentTime+newTime
  return(p1)
}
totalPatients<-0
## this is the simulation
## first task : loop through the time update for patients
## second task : include the doctor assignment query.
simulateProcess<-function(doctors=NULL,
                           totalWait=NULL,
                           totalPatients=0,
                           timeToClose=420,
                           currentTime=NULL){

  ## initiate Patient List
  patientList<-data.frame(patient=0,
                           arrivalTime=0,
                           doctorWaitTime=0,
                           doctorVisitTime=0,
                           currentTime=0,
                           assignedDoctor='none',
                           completionTime=0)

  ## not sure what to put here.
  currentTime<-patientList$currentTime[which(patientList$patient==max(patientList$patient))]
  currentPatientNumber<-0

  ## timeToClose (minutes) is stopping to admit patients
  while(currentTime<timeToClose){
    ## patient enters after the (i-1) patient enters.
    p1<-newPatientArrival(patientList,
                           closeTime=timeToClose,
                           waitTime=0,
                           visitTime=0,
                           currentPatientNumber=currentPatientNumber,
                           currentTime)

    ## update time
    p1<-updateTime(p1$currentTime,newTime=p1$arrivalTime,p1)

    # given a patient time, switch the availability of any doctor
    # if a doctors next available time is less than the current time, switch him to av
    ## FIX ME: need to ensure this flag is correct.
}

```

```

if(any(doctors$nextTimeAvail<p1$currentTime)){
  doctors$avail[which(doctors$nextTimeAvail<p1$currentTime)]<- 'yes'
}

## create a patient list
if(currentPatientNumber==0){
  patientList<-p1
  # update patient number
  currentPatientNumber<-currentPatientNumber+1
} else{
  patientList<-rbind(patientList,p1)
  # update patient number
  currentPatientNumber<-currentPatientNumber+1
}

## task 2 assign a doctor
#### check for doctor availability
## compute wait time, and/or compute the next available time
## returns a list object.
clinicList<-computeWaitTime(doctors,patientList,patientID=patientList$patient[currentPatientNumber])

doctors<-clinicList[["doctor"]]
patientList<-clinicList[["patient"]]
## update flags
# update currentTime
## current time is cumulative sum of the arrival times.
currentTime<-patientList$currentTime[which(patientList$patient==max(patientList$patient))] ##

## fix me:
## reset doctor availability based on current patient time.
upID<-which(doctors$nextTimeAvail<currentTime)
doctors$nextTimeAvail[upID]<-currentTime
doctors$currentTime[upID]<-currentTime
doctors$visitTimeLength[upID]<-0
}## while loop
return(list(patient=patientList,doctors=doctors))
}

doctors<-data.frame(dr=c('a','b','c'),
                      visitingPatient=c(0,0,0), ## who is doctor seeing (patient ID)
                      visitTimeLength=c(0,0,0), # length of doctor visit U(5,20)
                      currentTime=c(0,0,0),    ## current Time
                      nextTimeAvail=c(0,0,0),  ## current time + visitTimeLength = next avail time
                      avail=c("yes","yes","yes"))
## initiate times

```

```

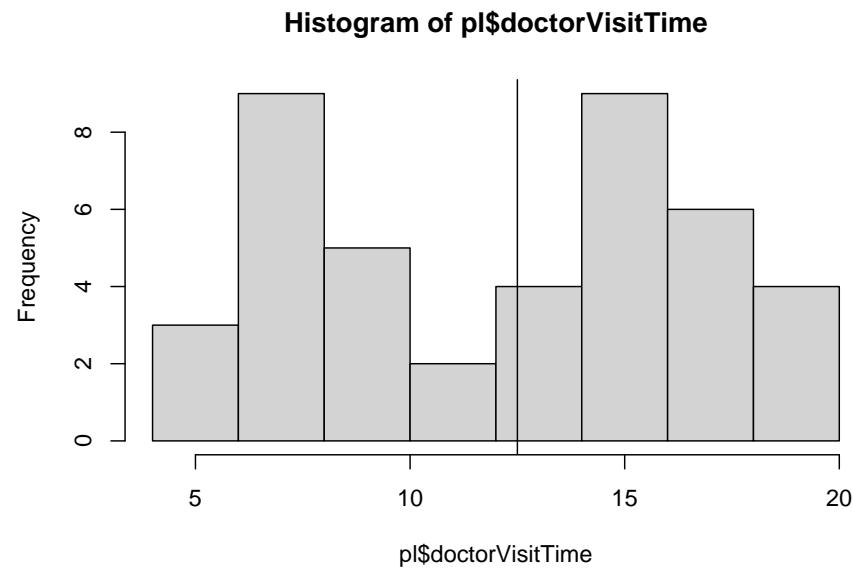
totalWait<-0
currentPatientNumber<-0
## clinic opens at 9am -4pm that is 7 hours (420 min.)
timeToClose<-7*60 ## stops admitting patients in 420 minutes
## current time is 0
## this will be the running total of minutes.
currentTime<-0

res<-simulateProcess(doctors,
                      totalWait,
                      totalPatients,
                      timeToClose,
                      currentTime)

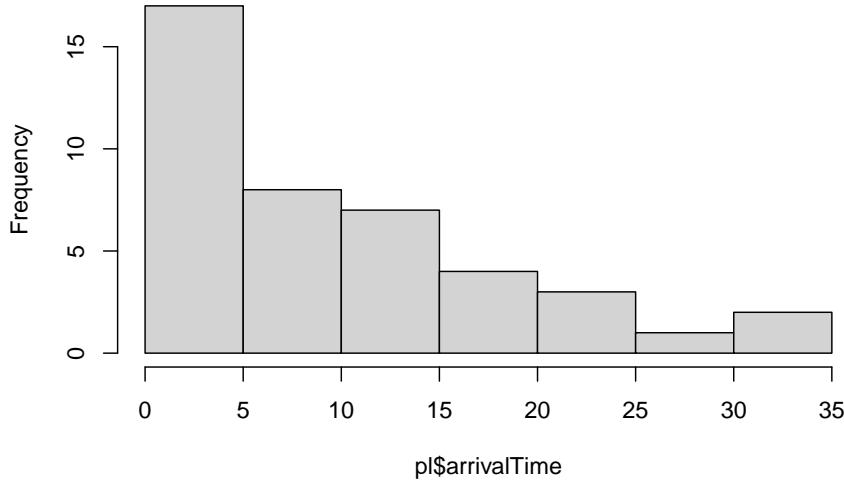
pl<-res$patient[which(res$patient$currentTime<=420),]

hist(pl$doctorVisitTime)
abline(v=(20+5)/2) ## should be ~12

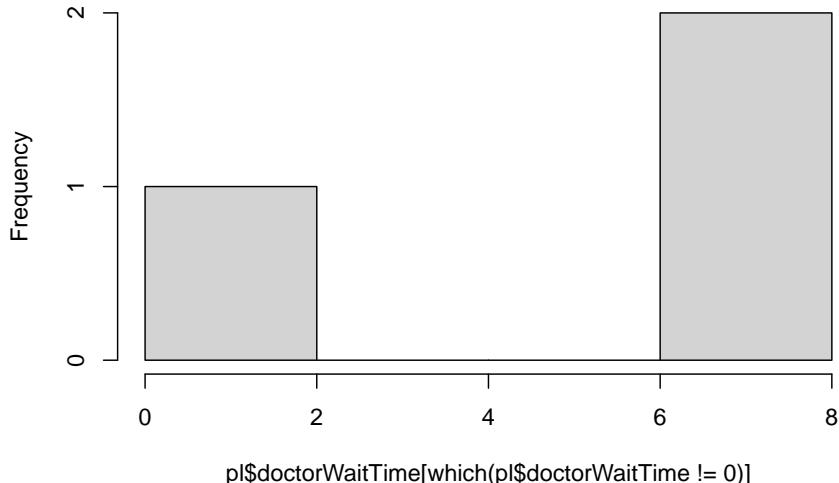
```



```
hist(pl$arrivalTime) ## should be close to 10 exp(1/10) has mean 10
```

**Histogram of pl\$arrivalTime**

```
hist(pl$doctorWaitTime[which(pl$doctorWaitTime!=0)]) ## about 2.41
```

**Histogram of pl\$doctorWaitTime[which(pl\$doctorWaitTime != 0)]**

```

print(max(pl$completionTime)-420) ## closing time

## [1] -4.370898

print(max(pl$patient)) ## total patient should be 42

## [1] 42

## (20-5)/6 + 10 this is about 12.5 minutes of arrival + visit time. which is approximated
## the arrival time is about 10 minutes.

## we should expect 42 patients
#420/10

## sanity check
#all(pl$currentTime+pl$doctorWaitTime+pl$doctorVisitTime-pl$completionTime==0)

```

## Simulation 100 times

total number of patients was approximately 42, which we expect since the total 420/10. The total number waiting with 3 doctors is 6.61 for 1 day. the average waiting time was about 4-5 minutes. For 1 day, the average closing time was 5.32 minutes after 4 pm

```

totalPat<-NULL
totalWaiting<-NULL
avgWaiting<-NULL
closing<-NULL
patientList<-NULL
p1<-NULL

for(i in 1:100){

  doctors<-data.frame(dr=c('a','b','c'),
                        visitingPatient=c(0,0,0), ## who is doctor seeing (patient ID)
                        visitTimeLength=c(0,0,0), # length of doctor visit U(5,20)
                        currentTime=c(0,0,0),    ## current Time
                        nextTimeAvail=c(0,0,0),   ## current time + visitTimeLength = next
                        avail=c("yes","yes","yes"))

  ## initiate times
  totalWait<-0
  totalPatients<-0
}

```

```
currentPatientNumber<-0
## clinic opens at 9am -4pm that is 7 hours (420 min.)
timeToClose<-7*60 ## stops admiting patienets in 420 minutes
## current time is 0
## this will be the running total of minutes.
currentTime<-0

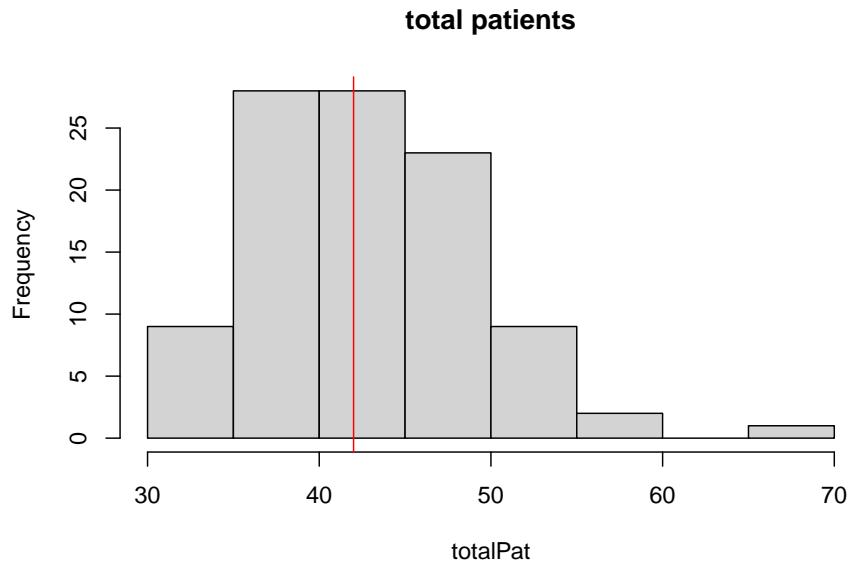
res<-simulateProcess(doctors,
                      totalWait,
                      totalPatients,
                      timeToClose,
                      currentTime)

pl<-res$patient[which(res$patient$currentTime<=420),]

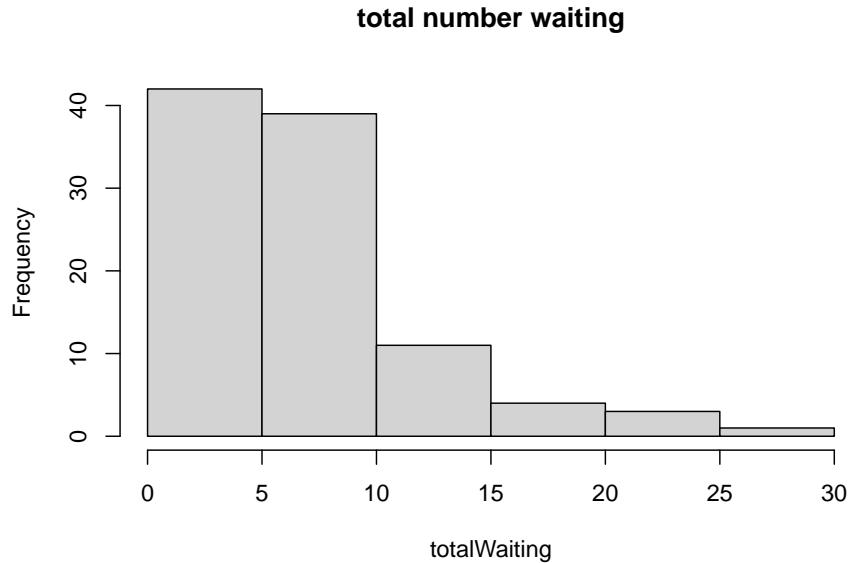
totalPat<-c(totalPat,max(pl$patient))
totalWaiting<-c(totalWaiting,nrow(pl[which(pl$doctorWaitTime!=0),]))
avgWaiting<-c(avgWaiting,mean(pl[which(pl$doctorWaitTime!=0),"doctorWaitTime"]))
closing<-c(closing,max(pl$completionTime))

}

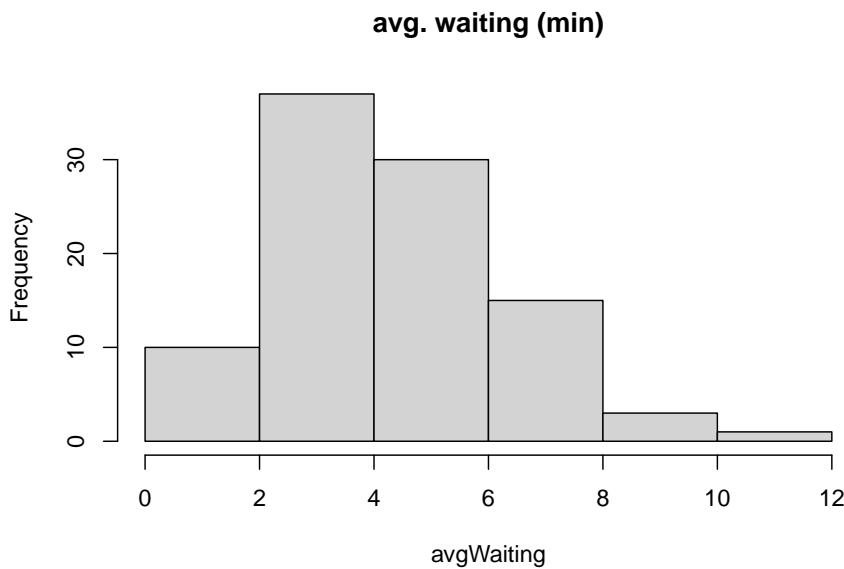
hist(totalPat,main="total patients")
abline(v=420/10,col='red')
```



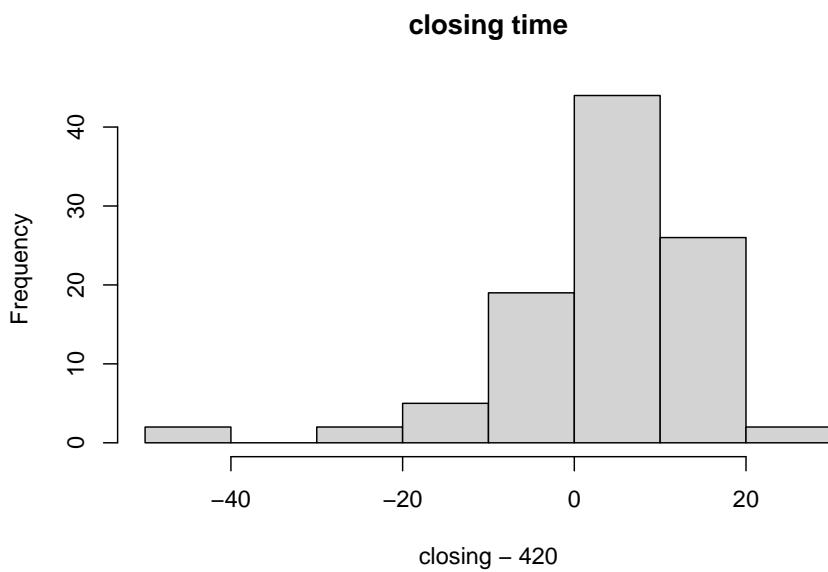
```
hist(totalWaiting,main="total number waiting")
```



```
hist(avgWaiting,main="avg. waiting (min)")
```



```
hist(closing-420,main="closing time")
```





# Chapter 2

## Single parameter models

### 2.1 Estimating a probability from binomial data

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \quad (2.1)$$

To perform Bayesian inference we assume  $\theta \sim U(0, 1)$  where the posterior is

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \quad (2.2)$$

which is the form of a **beta** distribution  $\theta|y \sim Beta(y+1, n-y+1)$

### 2.2 Posterior as a compromise between data and prior information

The posterior is less variable than the prior because it incorporates the information from the data.

$$E(\theta) = E(E(\theta|y)) \quad (2.3)$$

$$V(\theta) = E(V(\theta|y)) + V(E(\theta|y)) \quad (2.4)$$

where  $\theta|y$  is the posterior. So the average of the prior, is the average of the posterior means over the distribution of possible data. The variance of the prior (2.4) says the posterior variance is on average smaller than the prior variance.

## 2.3 Summarizing the posterior inference

The mean, median, mode, and standard deviation of the posterior distribution summarize all the current information about a model.

### Posterior quantiles and intervals

The posterior uncertainty can be reported by presenting the quantiles of the posterior distribution. The interval, a *central interval of posterior probability* corresponds to the case of  $100(1 - \alpha)\%$ , to the range of values above and below which lies exactly  $100(\alpha/2)\%$  of the posterior probability. The interval estimates are *posterior intervals*. This differs from the confidence interval because the confidence interval is not a probability interval, because either the parameter is within the region or it is not, but the confidence interval provides information in the long run over repeated experimentation as to how many experiments would contain the true parameter.

There is also the *highest posterior interval* which is a probabilistic interval that is not less than any region outside of the interval.

## 2.4 Informative prior distributions

the property that the posterior distribution follows the same parametric form as the prior distribution is called *conjugacy*. Where the beta prior distribution is a *conjugate family* for the binomial likelihood.

so given the binomial likelihood  $p(y|\theta) \propto \theta^y(1-\theta)^{n-y}$ , and a prior density  $p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$  the posterior is of the beta family.

$$\begin{aligned} p(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= Beta(\theta|\alpha + y, \beta + n - y) \end{aligned}$$

### Conjugate prior distributions

Conjugacy is formally defined as if  $F$  is a class of sampling distributions  $p(y|\theta)$ , and  $P$  is a class of prior distributions for  $\theta$ , then the class  $P$  is conjugate for  $F$  if  $p(\theta|y) \in P$  for all  $p(\cdot|\theta) \in F$  and  $p(\cdot) \in P$ .

### Conjugate prior, distributions, exponential families, and sufficient statistics

Posterior distributions can be derived using sufficient statistics from exponential families. The exponential family is defined as

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

Where  $\phi(\theta), u(y_i)$  are vectors of equal dimension to that of  $\theta$ . The  $\phi(\theta)$  is called the *natural parameter* for the family (F). The likelihood of a sequence  $y = (y_1, \dots, y_n)$  iid is

$$\begin{aligned} p(y|\theta) &= \left( \prod_{i=1}^n f(y_i) \right) g(\theta)^n \exp\left(\phi(\theta)^T \sum_{i=1}^n u(y_i)\right) \\ &\propto g(\theta)^n e^{\phi(\theta)^T t(y)}, t(y) = \sum_{y=1}^n u(y_i) \end{aligned}$$

The *sufficient statistic* for  $\theta$  is  $t(y)$  because the likelihood for  $\theta$  depends on the data,  $y$ , only through the value of  $t(y)$ .

Sufficient statistics benefit posterior distributions because if the prior density is specified as

$$p(\theta) \propto g(\theta)^\eta e^{\phi(\theta)^T \nu}$$

Then the posterior density using sufficient statistics is

$$p(\theta|y) \propto g(\theta)^{\eta+n} e^{\phi(\theta)^T (\nu+t(y))}$$

Exponential families are the only classes of distributions that have natural conjugate prior distributions.

## 2.5 Normal distribution with known variance

The normal distribution is foundational to statistical modeling, with the central limit theorem (CLT) allowing for the use of normal likelihood in many statistical problems which can approximate complex likelihoods. If the normal distribution does not provide a good model fit, finite mixtures of distributions can identify useful solutions.

### Likelihood of one data point

With mean  $\theta$  and known variance  $\sigma^2$  the sampling distribution of a given point is defined

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

### 2.5.1 Conjugate prior and posterior distributions

The prior has the exponential family form given as  $\theta \sim N(\mu_0, \tau_0^2)$

$$p(\theta) \propto \exp\left(\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right)$$

Where completing the square can find the posterior distribution

$$\begin{aligned} p(\theta|y) &\propto \exp\left(-1/2\left(\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right) \\ p(\theta|y) &\propto \exp\left(\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right), \theta|y \sim N(\mu_1, \tau_1^2) \end{aligned}$$

where  $\mu_1 = \frac{1/\tau_0^2\mu_0 + 1/\sigma^2 y}{1/\tau_0^2 + 1/\sigma^2}$  and  $1/\tau_1^2 = 1/\tau_0^2 + 1/\sigma^2$

In manipulating the distributions the inverse of the variance is defined as the *precision*. The posterior precision is equal to the prior precision plus the data precision. And the posterior mean is a weighted average of the prior mean and the observed value, y, proportional to the total precision.

### Posterior predictive distribution

the posterior predictive distribution of a future observation, x,  $p(x|y)$  can be calculated

$$p(x|y) = \int p(x|\theta)p(\theta|y)d\theta \propto \int \exp(-1/2\sigma^2(x-\theta)^2)\exp(-1/2\tau_1^2(\theta-\mu_1)^2)d\theta$$

the future observations, x, does not depend on the past observations y given  $\theta$ .

### Normal model with multiple observations

For multiple observations, y, the posterior density is formulated as:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &= p(\theta) \prod_i p(y_i|\theta) \\ &\propto \exp(-1/2\tau_0^2(\theta - \mu_0)^2) \prod_i \exp(-1/2\sigma^2(y_i - \theta)^2) \\ &\propto \exp(-1/2(1/\tau_0^2(\theta - \mu_0)^2 + 1/\sigma^2 \sum_i (y_i - \theta)^2)) \end{aligned}$$

Simplifying the algebra shows the posterior depends on  $y$  only through the sample mean (sufficient statistic),  $\bar{y}$  is the sufficient statistic for  $\theta$ , and the final model is  $\bar{y}|\theta, \sigma^2 \sim N(\theta, \sigma^2/n)$

## 2.6 Other standard single-parameter models

The binomial model is motivated by counting exchangeable outcomes. The normal distribution applies to a random variable that is the sum of many exchangeable- or independent terms. The poisson and exponential distribution applies to number of counts, rates, or waiting times for events modeled as occurring exchangeably in all time intervals, i.e. independently in time, and with a constant rate of occurrence.

### Normal distribution with known mean but unknown variance

For  $p(y|\theta, \sigma^2) = N(y|\theta, \sigma^2)$  with known mean and unknown variance the likelihood vector with iid observations follows

$$\begin{aligned} p(y|\sigma^2) &\propto \sigma^{-n} \exp(-1/2\sigma^2 \sum (y_i - \theta)^2) \\ &= (\sigma^2)^{-n/2} \exp(-n/2\sigma^2 \nu) \end{aligned}$$

The sufficient statistic  $\nu = -1/n \sum (y_i - \theta)^2$ . The corresponding conjugate prior density is the inverse-gamma

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

The convenient parameterization is as a scaled inverse- $\chi^2$  distribution with scale  $\sigma_0^2$  and  $\nu_0$  degrees of freedom. The prior distribution for  $\sigma^2$  is taken as  $\sigma_0^2 \nu_0 / X$ , where  $X \sim \chi_{\nu_0}^2$  random variable.

The resulting posterior density for  $\sigma^2$  is

$$\begin{aligned} p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\ &\propto \frac{\sigma_0^{2\nu_0/2+1}}{\sigma^2} \exp\left(\frac{\nu_0 \sigma_0^2}{2\sigma^2}\right) ((\sigma^2)^{-n/2} \exp(-n/2\frac{\nu}{\sigma^2})) \\ &\propto (\sigma^2)^{-(n+\nu_0)/(2+1)} \exp\left(\frac{-1}{2\sigma^2}(\nu_0 \sigma_0^2 + n\nu)\right) \\ &\text{then } \sigma^2|y \sim \text{Inv.}\chi^2(\nu_0 + n, \frac{\nu_0 \sigma_0^2 + n\nu}{\nu_0 + n}) \end{aligned}$$

Which is a scaled inverse- $\chi^2$  with scale equal to the degrees-of-freedom weighted average of the prior and data scales, and the d.f. equal to the sum of the prior and data degrees-of-freedom.

### Poisson model

The Poisson model arises in data in the form of counts; in epidemiology studies disease incidence is modeled using Poisson framework.

the likelihood is

$$\begin{aligned} p(y|\theta) &= \prod_i^n \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &= \theta^{t(y)} e^{-n\theta} \end{aligned}$$

Where the sufficient statistic is  $t(y) = \sum y_i$  and the likelihood in exponential family form is written as

$$p(y|\theta) \propto e^{-n\theta} e^{t(y)\log(\theta)}$$

where the natural parameter  $\phi(\theta) = \log(\theta)$  and the natural conjugate prior distribution is

$$p(\theta) \propto (e^{-\theta})^\eta e^{\nu * \log(\theta)}$$

indexed by hyperparameters  $(\eta, \nu)$ . We can rewrite the likelihood in the form  $\theta^a e^{-b\theta}$  and so the conjugate prior must be in the form of  $p(\theta) \propto \theta^A e^{-B\theta}$  which can be re-written to follow a Gamma density as

$$p(\theta) \propto e^{-\beta\theta} \theta^{\alpha-1}$$

and this follows a  $\text{Gamma}(\alpha, \beta)$  equivalent to prior total count of  $\alpha - 1$  in  $\beta$  prior observations. The posterior is

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$$

### Negative Binomial distribution

With conjugate families, with the known form of the prior and posterior densities can be used to find the marginal

$$p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)} \quad (2.5)$$

For a single Poisson model with 1 observation,  $y$ , this has a prior predictive distribution

$$p(y) = \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha, \beta)}{\text{Gamma}(\theta, \alpha + y, \beta + 1)} = \frac{\Gamma(\alpha + y)\beta^\alpha}{(\Gamma(\alpha)y!(1 + \beta)^{\alpha+y})} = \binom{\alpha + y - 1}{y} \left(\frac{\beta}{\beta + 1}\right)^\alpha \left(\frac{1}{\beta + 1}\right)^y y \sim NB(\alpha, \beta) \quad (2.6)$$

This shows the negative binomial distribution is a mixture of Poisson distribution with rates,  $\theta$  that follow the Gamma distribution

$$\text{Neg-bin}(y|\alpha, \beta) = \int \text{Pois}(y|\theta)\text{Gamma}(\theta|\alpha, \beta)d\theta \quad (2.7)$$

### Exponential distribution

The exponential distribution is used to model waiting times and other continuous positive real-valued random variables, often on a time scale. The sampling distribution of an outcome  $y$ , given parameter  $\theta$  is

$$p(y|\theta) = \theta \exp(-y\theta), y > 0 \quad (2.8)$$

and  $\theta = 1/E(y)$  is called the *rate*. The conjugate prior for the Exponential distribution is the  $\text{Gamma}(\theta|\alpha, \beta)$ , with corresponding posterior  $\text{Gamma}(\theta, \alpha + 1, \beta + y)$ . The Gamma prior  $\text{Gamma}(\alpha, \beta)$  for  $\theta$  can be viewed as  $\alpha - 1$  exponential observations with total waiting time  $\beta$ .

### Constructing a prior distribution

In the text example, the prior distribution values were set using a Gamma distribution with  $\alpha, \beta$  estimated from the data to match the distribution of the observed cancer death rates  $y_j/10n_j$ . It may seem inappropriate to use the data to set the prior distribution, but it is used in hierarchical modeling and is an approximation.

Under the model the observed count  $y_j$  for any county,  $j$ , comes from the *predictive distribution*  $p(y_j) = \int p(y_j|\theta_j)p(\theta_j)d\theta_j$ , which is the Negative binomial  $\text{NB}(\alpha, \beta/10n_j)$ .

We then use a method of moments estimator to solve for  $\alpha, \beta$ .

$$E(y_j) = 10n_j \frac{\alpha}{\beta} V(y_j) = 10n_j \frac{\alpha}{\beta} + (10n_j)^2 \frac{\alpha}{\beta^2} \quad (2.9)$$

Here we match the *observed* mean and variance to their expectations and solve for both parameters, yields the parameters of the prior.

## 2.7 Noninformative prior distributions

Reference prior distributions are described as vague, flat, diffuse or *noninformative*. This gives more weight to the data as opposed to prior beliefs. A related idea is *weakly informative* prior distribution, which contains some information enough to keep it within reasonable bounds.

### 2.7.1 Jeffreys' invariance principle

Noninformative prior distributions based on considering one-to-one transformations of the parameter:  $\phi = h(\theta)$ . By transformation of variables, the prior density  $p(\theta)$  is equivalent to the prior density on  $p(\phi)$ .

$$p(\phi) = p(\theta) \left| \frac{d\theta}{d\phi} \right| = p(\theta) |h'(\theta)|^{-1} \quad (2.10)$$

Jeffreys' general principle is that any rule for determining the prior density  $p(\theta)$  yields equivalent result by transforming to  $p(\phi)$ .

This principle leads to defining the noninformative prior density as  $p(\theta) \propto |J(\theta)|^{1/2}$ , where  $J(\theta)$  is the *Fisher information* for  $\theta$ .

$$J(\theta) = E\left(\left(\frac{d\log p(y|\theta)}{d\theta}\right)^2 | \theta\right) = -E\left(\frac{d^2 \log p(y|\theta)}{d\theta^2} | \theta\right) \quad (2.11)$$

### Difficulties with noninformative prior distributions

1. if the likelihood is truly dominant, then the prior densities can not matter, but using this as automatic assumption of the likelihood is mis-guided. 2. uniform or flat prior density is uniform for one parameterization, and will not be uniform in another parameterization. The scale of a given problem may change, and uniform prior may not be reasonable assumption in another parameterization. 3. averaging over a set of competing models that have improper prior distributions will lead to problems.

## 2.8 Exercises

### Question 1

prior Beta(4,4), where a coin is tossed 10 times and heads appears fewer than 3 times. the exact posterior is Beta( $4+y$ ,  $4+10-y$ ) for  $y=0,1,2$ . Since we don't know the observed heads, but that  $y < 3$  we plot the posterior distributions for each possibility. For 2 heads it is closer to the prior, with posterior mean of 0.33, which is closest to the prior mean of 1/2.

The random event is that  $Y = (0, 1, 2)$  and we sum all possibilities for the total likelihood.

The likelihood of the total event is  $\theta^0(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8$   
the posterior is  $\theta^3(1-\theta)^{13} + \binom{10}{1}\theta^4(1-\theta)^{12} + \binom{10}{2}\theta^5(1-\theta)^{11}$

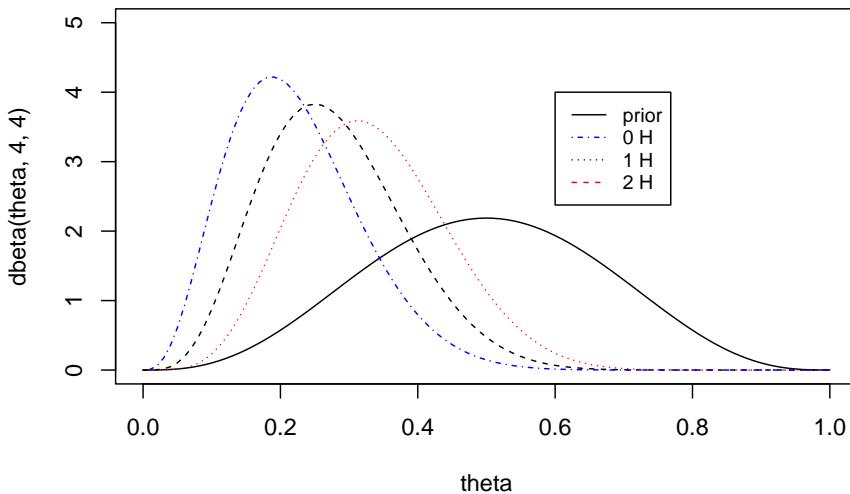
```

theta<-seq(from=0,to=1,by=0.01)

plot(theta,dbeta(theta,4,4),type='l',ylim=c(0,5),main='individual posteriors') ## prior
lines(theta,dbeta(theta,4+1,4+10-1),lty=2) ## 1 success
lines(theta,dbeta(theta,4+2,4+10-2),lty=3,col='red') ## 2 successes
lines(theta,dbeta(theta,4,4+10),lty=4,col='blue') ## 0 successes

legend(0.6, 4,
       legend=c("prior", "0 H", "1 H", "2 H"),
       col=c("black","blue","black", "red"), lty=c(1,4,3,2), cex=0.8)

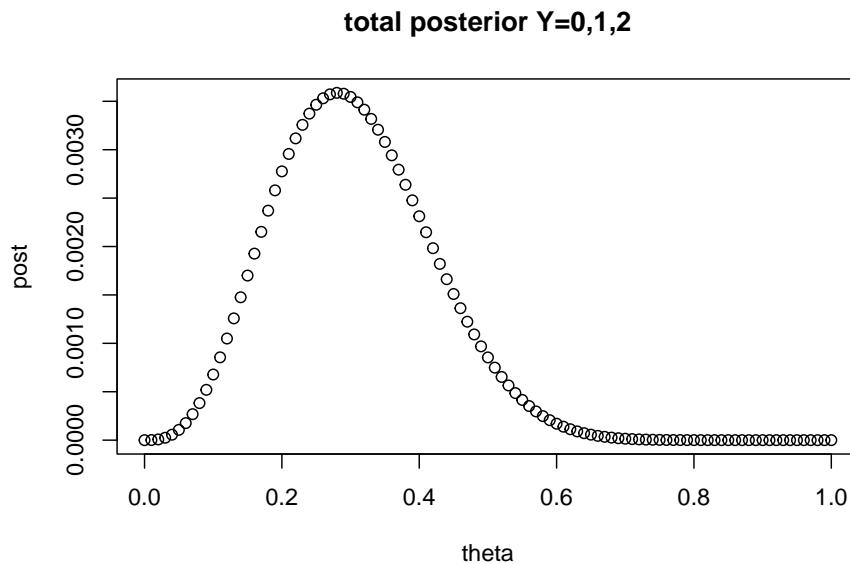
```

**individual posteriors**

```

post<- theta^3*(1-theta)^(13)+choose(10,1)*theta^4*(1-theta)^12+choose(10,2)*theta^5*(1-theta)^11
plot(theta,post,main='total posterior Y=0,1,2')

```



### Normal approximation example

For female births we have beta(438,544) we use the normal approximation. This replicates Gelman's Figure 2.3 (a, b)

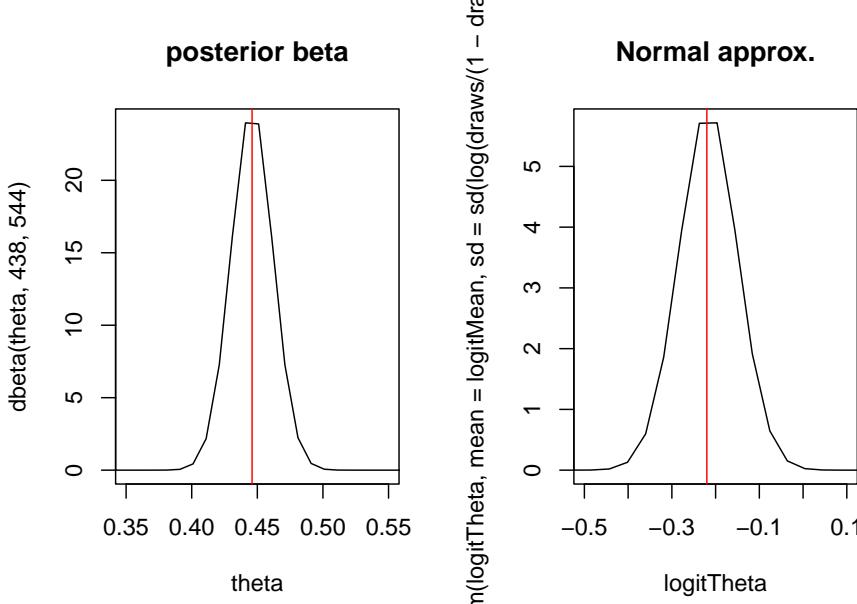
```

theta<-seq(from=0.001,to=1,by=0.01)
## example births
postMean <-function(alpha,beta,y,n){
  return( (alpha+y)/(alpha+beta+n))
}
postVar<-function(alpha,beta,y,n){
  return( ((alpha+y)*(beta+n-y))/((alpha+beta+n)^2*(alpha+beta+n+1)) )
}
sdnorm<-sqrt(postVar(438,544,0,0))
logitMean<-log( postMean(438,544,0,0)/(1-postMean(438,544,0,0)))

logitTheta<-log(theta/(1-theta))

par(mfrow=c(1,2))
plot(theta,dbeta(theta,438,544),type='l',xlim=c(0.35,0.55),main="posterior beta")
abline(v=0.446,col='red')
draws<-rbeta(1000,438,544)
plot(logitTheta,dnorm(logitTheta,mean =logitMean, sd=sd(log(draws/(1-draws)))) ,type="l")
abline(v=-0.22,col='red')

```



### Question 3

The prior predictive distributions for the number of 6's in a fair roll, tossed 1,000 times will follow a beta distribution. Let  $y$  be the number of 6's in 1000 rolls of fair die, the probability for a 6 is  $1/6$ , so the number of 6's (successes) in this trial is approximately 167, and 833 failures as the prior prediction. with probability of success ( $1/6$ ).

We plot the beta distribution of the expected number of heads in 1000 tosses.

The normal approximation for the prior prediction uses the binomial distribution is  $\mu = n * p = 167$  and  $\sigma^2 = npq = 138.89 \sim N(167, 138.89)$

The normal approximation shows the probability of heads in a given 1000 tosses, using a non-informative prior beta( $167, 833$ ) which has a prior predictive mean of  $\exp(-1.79)/(1 + \exp(-1.79)) = 0.143$ . This is not the same for number of success in 1000 tosses.

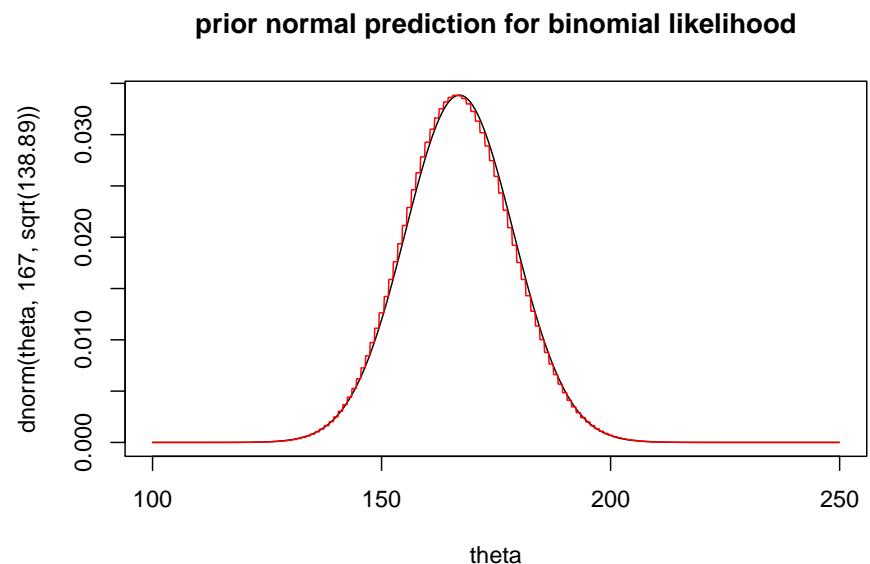
We find the probability distribution of a given success and the prior probability predictive interval follows a beta with 95% (0.12,0.17) for the probability of rolling a 6

```
## based on normal approximation sketch the distribution of y
## for normal we use the logit transform
## n = 1,000
## first lets construct a beta distribution.
```

```

## let the prior be beta(4,4) or even beta(1,1)
## prior prediction
theta<-seq(from=100,to=250,by=0.11)
plot(theta,dnorm(theta,167,sqrt(138.89)),type='l',main="prior normal prediction for binomial likelihood")
lines(theta,dbinom(round(theta),size=1000,prob=(1/6)),main="prior binom prediction for binomial likelihood")

```



```

## posterior prediction (normal).
## assume p=1/6 we expect 167 successes under likelihood.
## simulation
priordraws<-sample(theta,size=1000,replace=T,prob=dnorm(theta,167,sqrt(138.89)))
priorpred<-rnorm(1000,mean=priordraws,sd=sqrt(138.89))
## prior prediction interval
quantile(priorpred,c(0.05,0.25,0.5,0.75,0.95))

```

```

##      5%      25%      50%      75%      95%
## 141.6362 155.5543 167.2259 178.0557 195.0389

```

```
qnorm(c(0.05,0.25,0.5,0.75,0.95),167,sqrt(138.89))
```

```
## [1] 147.6151 159.0510 167.0000 174.9490 186.3849
```

## Question 4

We have a mixture of 3 normal distributions, and show the central intervals for 5,25,50,75, and 95% predictive probabilities. The question gives  $\theta$  as the probability of a 6 on a die, possibly unfair, in 1,000 tosses. we have  $\theta = 1/12, 1/6, 1/4$  types of biased die. Using the normal approximation, the predictive prior probability is  $\sum_{\theta} p(\theta)p(y|\theta)$  where the likelihood is approximated using a normal distribution  $\mu = n * \theta_i, \sigma = n * \theta_i(1 - \theta_i)$

```

x<-seq(0,1000)-0.5 # continuity correction.
theta<-c(1/12,1/6,1/4)
n=1000

a<-dnorm(x,mean=n*theta[1],sd=sqrt(n*theta[1]*(1-theta[1])))
b<-dnorm(x,mean=n*theta[2],sd=sqrt(n*theta[2]*(1-theta[2])))
c<-dnorm(x,mean=n*theta[3],sd=sqrt(n*theta[3]*(1-theta[3])))

## posterior computation
##p(y) = p(y| theta1 )*p(theta1)+ p(y| theta2 )*p(theta2)+ p(y| theta3 )*p(theta3)
# total law of probability
mypost<-a*0.25+b*0.5+c*0.25

## prior pred p(y) = p(y|theta)*p(theta) / p(theta| y)
mypri<-mypost

sum(mypri) ## sums to 1 it is a distribution

## [1] 1

par(mfrow=c(3,2))
plot(x,mypri,type='l',main='predictive prior number of heads')

data<-data.frame(x=x,p=mypri)

## highest probability interval
# 95%

plot(x,mypri,type='l',main=' 95% predictive prior number of heads')
abline(v=max(data[which(cumsum(data$p)<0.025),1]))
abline(v=max(data[which(1-cumsum(data$p)>0.025),1]))

plot(x,mypri,type='l',main='75% predictive prior number of heads')

```

```

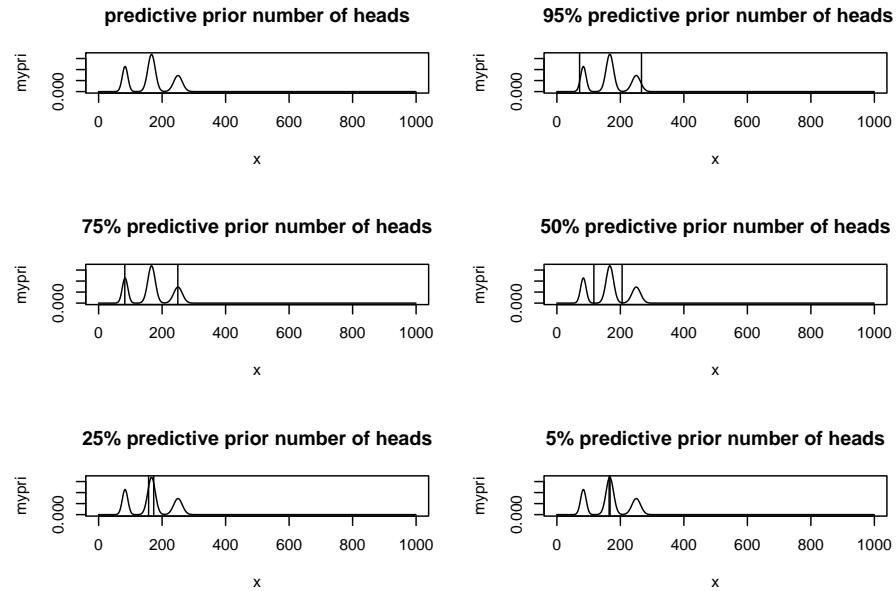
abline(v=max(data[which(cumsum(data$p)<0.125),1]))
abline(v=max(data[which(1-cumsum(data$p)>0.125),1]))
)

plot(x,mypri,type='l',main='50% predictive prior number of heads')
abline(v=max(data[which(cumsum(data$p)<0.25),1]))
abline(v=max(data[which(1-cumsum(data$p)>0.25),1]))
)

plot(x,mypri,type='l',main='25% predictive prior number of heads')
abline(v=max(data[which(cumsum(data$p)<0.375),1]))
abline(v=max(data[which(1-cumsum(data$p)>0.375),1]))
)

plot(x,mypri,type='l',main='5% predictive prior number of heads')
abline(v=max(data[which(cumsum(data$p)<0.475),1]))
abline(v=max(data[which(1-cumsum(data$p)>0.475),1]))
)

```



### Question 7

- (a) for the binomial likelihood  $y \sim Bin(n, \theta)$ , show that  $p(\theta) \propto \theta^{-1}(1 - \theta)^{-1}$  is the uniform prior for the natural parameter of the exponential

family.

- first we write the binomial likelihood in exponential family form, note that we are not deriving the posterior of any function, but just studying the likelihood.

$$\begin{aligned} p(y|\theta) &= \theta^y(1-\theta)^{n-y} \\ \implies \log(p) &= y\log(\theta/(1-\theta)) + n\log(1-\theta) \\ \implies p(y|\theta) &= \exp(y * \log(\theta/(1-\theta))) * g(\theta) \end{aligned}$$

hence  $\phi(\theta) = \log(\theta/1-\theta)$ , and we're given uniformity  $p(\phi(\theta)) \propto 1$

Inverting  $\phi(\theta) = \log(\theta/1-\theta)$  yields  $\theta = \frac{e^\phi}{1+e^\phi}$  and we know that  $p(\phi) \propto 1$ . Using the jacobian we can derive  $p(\theta) = p(\phi) \frac{d\phi}{d\theta}$  shown as

$$\begin{aligned} p(\theta) &= p(\phi) \frac{d\phi}{d\theta} \\ &= p(\phi) d/d\theta(\log(\theta/1-\theta)) \\ &\propto (1)(1-\theta/\theta)(1/(1-\theta)^2) \\ &= 1/\theta * (1/1-\theta) \end{aligned}$$

### Question 8 (Normal distribution with unknown mean)

A random sample of  $n$  students is drawn from a large population, and weights are measured. The average height of the  $n$  sampled students is  $\bar{y} = 150$  lbs. Assume the weights in the population are normally distribution with unknown mean,  $\theta$ , and known standard deviation 20 lbs. Suppose the prior for  $\theta \sim N(180, 40^2)$

- (a) For known variance, the limit of the posterior is  $p(\theta|y) \approx N(\theta|\bar{y}, \sigma^2/n)$ . And the direct formulation is  $p(\theta|\bar{y}) = N(\theta|\mu_n, \tau_n^2)$ . using equations 2.12.
- (b) For a posterior predictive interval, the marginal distribution for new data  $p(\tilde{y}|y) \sim N(\mu_n, \sigma^2 + \tau_n^2)$

```
mu_n<-function(mu0,ybar,n,tau02,sigma2){
  mun<- (mu0/tau02 + (ybar*n)/sigma2)/(1/tau02 + n/sigma2)
  return(mun)
}
taun2<-function(tau02,n,sigma2){
  inv.taun2<- 1/tau02 +n/sigma2
  return(1/inv.taun2)
}
```

- (c) Here we give the posterior interval and predictive interval for n=10

```
ybar=150
sigma2=20^2
mu0=180
tau02=40^2
n=10

## posterior interval n=10
#lower<-round(qnorm(0.025,mu_n(mu0,ybar,tau02,10,sigma2),sd=sqrt(taun2(tau02,10,sigma2)))
upper<-round(qnorm(c(0.025,0.975),mu_n(mu0,ybar,tau02,10,sigma2),sd=sqrt(taun2(tau02,10,sigma2)))

message("posterior interval n=10: ",upper[1]," ",upper[2])
```

```
## posterior interval n=10: 138.49 162.98
```

```
## posterior predictive interval
upper2<-round(qnorm(c(0.025,0.975),mu_n(mu0,ybar,tau02,10,sigma2),sd=sqrt(taun2(tau02,10,sigma2)))

message("posterior predictive interval n=10: ",upper2[1]," ",upper2[2])
```

```
## posterior predictive interval n=10: 109.66 191.8
```

- (d) the posterior and predictive for n=100

```
lower<-round(qnorm(0.025,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(taun2(tau02,100,sigma2)))
upper<-round(qnorm(0.975,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(taun2(tau02,100,sigma2)))

message("posterior interval n=100: ",lower," ",upper)
```

```
## posterior interval n=100: 146.16 153.99
```

```
## this approximately equals the limit
print("The asymptotic approximation")
```

```
## [1] "The asymptotic approximation"
```

```
qnorm(c(0.025,0.975),mean=150, sigma2/100)
```

```
## [1] 142.1601 157.8399
```

```

## posterior predictive interval
lower2<-round(qnorm(0.025,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(tau02(tau02,100,sigma2)+sigma2
upper2<-round(qnorm(0.975,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(tau02(tau02,100,sigma2)+sigma2

message("posterior predictive interval n=100: ",lower2," ",upper2)

## posterior predictive interval n=100: 110.68 189.47

```

## Question 10

Suppose there are  $N$  cable cars numbered sequentially 1 to  $N$ . You see a car at random labeled 203 and wish to estimate  $N$ . (a) assume the prior follows  $\text{Geo}(1/100)$  what is the posterior for  $N$ ?

- (a) The car  $X=203$  was observed, and the likelihood of this observation (data) is uniform  $p(X|N = 203) = 1/N$  which assumes each car is equally likely. So the posterior  $p(N|X) = (1/N)(1/100)(99/100)^{N-1}$ . So the posterior is proportional to  $(1/N)(99/100)^{N-1}$  with  $N \geq 203$ .
- what is the posterior mean and std. deviation for  $N$ ? we use Bayes' theorem  $p(N|X) = \frac{p(X|N)p(N)}{p(X)}$ . And use a computer to approximation  $p(X)$  which is approximately 0.0471 (ignoring constant). the infinite series  $\sum_{n=203}^{\infty} (1/n)(99/100)^{n-1}$  converges by using the ratio test with  $\rho < 1$  to  $\approx 0.0471$ . The posterior mean and standard deviation is 279.09 and 79.96, which was evaluated using the posterior distribution  $p(N|X)$  defined as

$$p(N|X) \propto \frac{(1/N)(99/100)^{N-1}}{p(X)}$$

```

totalSum=0

for(i in 203:30000){
  sum_i= (1/i)*(99/100)^(i-1)
  totalSum= totalSum+sum_i
}

print(totalSum)

```

```
## [1] 0.04705084
```

```

## approximate the distribution
px<-totalSum
X<-seq(203,300000)

```

```

dpx<-function(N,px){
  dd<- (1/px)*(1/N)*(99/100)^(N-1)

  return(dd)
}

## the posterior mean is
##  $E(X) = \bar{x}f(x)$ 
postMean<-function(N,px){
  fx<- dpx(N,px)
  ex<- N*fx
  return(ex)
}
EX<-sum(postMean(X,px))

## posterior variance
##  $VX = EX^2 - (EX)^2$ 
postMean.sq<-function(N,px){
  fx<- dpx(N,px)
  ex2<- N*N*fx
  return(ex2)
}
## the variance term has to be evaluated at smaller limit to avoid overflow
sdX<-sqrt(sum(postMean.sq(seq(203,35000),px))-(EX)^2)
message("posterior mean and sd is:", round(EX,2), " ", round(sdX,2))

```

```
## posterior mean and sd is:279.09 79.96
```

- (b) we use a simulation to estimate the posterior. We sample with replacement, with a distribution size equal to a large number ( $N$ ), with probability of observing a number equal to the posterior distribution. The empirical mean and std.dev are 280.5, and 80.9

```

N<-203:100000
Nsim<-10000
unnorm.post<-(1/N)*(99/100)^(N-10)
mean(sample(N,size=Nsim,prob=unnorm.post,replace=T))

```

```
## [1] 280.3206
```

```
sd(sample(N,size=Nsim,prob=unnorm.post,replace=T))
```

```
## [1] 80.18079
```

- (c) we can use a uniform prior  $p(N) = 1/N$  with a geometric likelihood  $p(X|N) = (1/100)(99/100)^{N-1}$  which will not change the result. We can also change the parameter p to be distributed under a binomial model.

## Question 11

suppose  $y_1, \dots, y_5$  are iid  $\text{Cauchy}(\theta, 1)$  r.vs. and the prior distribution for  $\theta \sim U[0, 100]$ . the given observations are  $y=43,44,45,46.5,47.5$ .

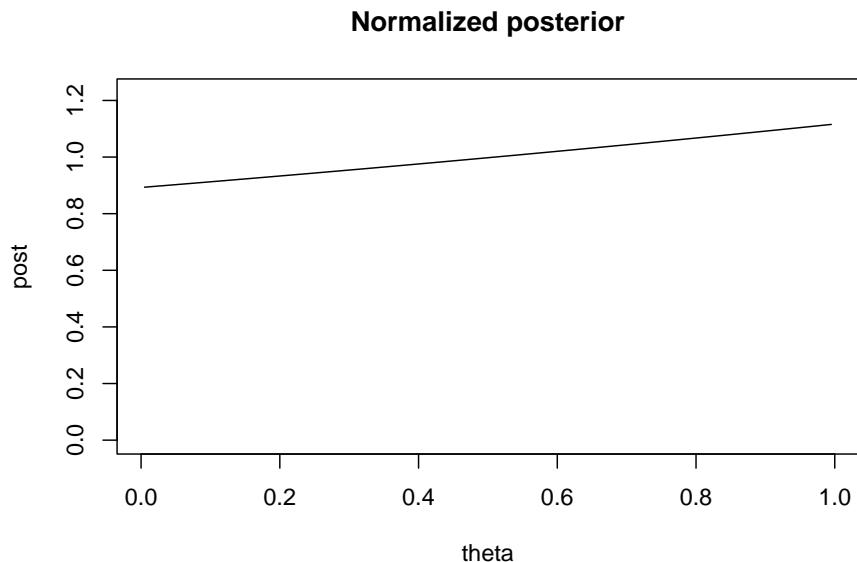
- (a) compute the unnormalized posterior density on a grid of points  $\theta = 01/m, 2/m, \dots, 100$ . using the grid approximation, compute and plot the posterior density as a function of  $\theta$

for the likelihood  $L(\theta|y) = \prod_{i=1}^n f(y_i|\theta)$  requires the product of  $y$  for a given theta. this was a mistake i made in the first attempt. The posterior is  $p(\theta|y) = L(\theta|y)p(\theta)$  wher  $p(\theta) = 1/100$

```

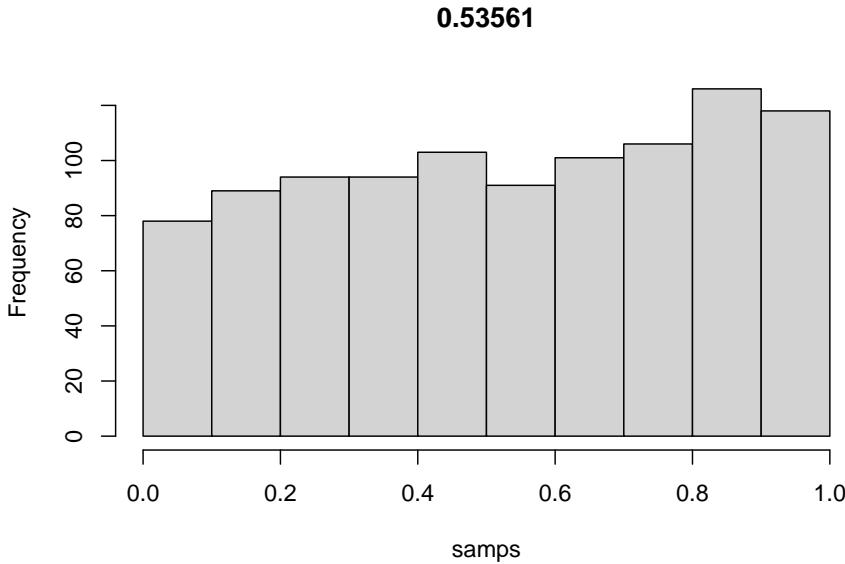
y= c(43,44,45,46.5,47.5)
## previous editions values for checking
#y=c(-2,-1,0,1.5,2.5)
step=0.01
#theta<-seq(from=0,to=100000)/m
theta<-seq(step/2, 1-step/2,step)
## p(theta | y) ~ p(y|theta)*p(theta)
dens<-function(y,th){
  dens0<-NULL
  for(i in 1:length(th)){
    dens0<-c(dens0, prod (dcauchy(y, th[i],1)))
  }
  dens0
}
#dens(y,theta)
#  $L(\theta | y) = \prod_{i=1}^n f(y_i | \theta)$  we need the product term here.
unnorm.post<-sapply(theta, function(x) prod(dcauchy(y,location=x,scale=1))) ## un norm post
##  $p(\theta | y) = p(y | \theta)p(\theta)$  where  $p(\theta)$  is  $U(0,100)$ 
post<-unnorm.post/(step*sum(unnorm.post))
plot(theta,post,type='l',main='Normalized posterior', ylim=c(0, 1.1*max(post)))

```



- (b) Sample 1000 draws from theta from posterior density and plot histogram we sample from theta [0,100] using the grid approximation, and the probability is from the posterior distribution.

```
samps<-sample(theta, 1000, prob=post*step, replace=T)
hist(samps, main=mean(samps))
```



**##sample mean is close to the mean of the observed.**

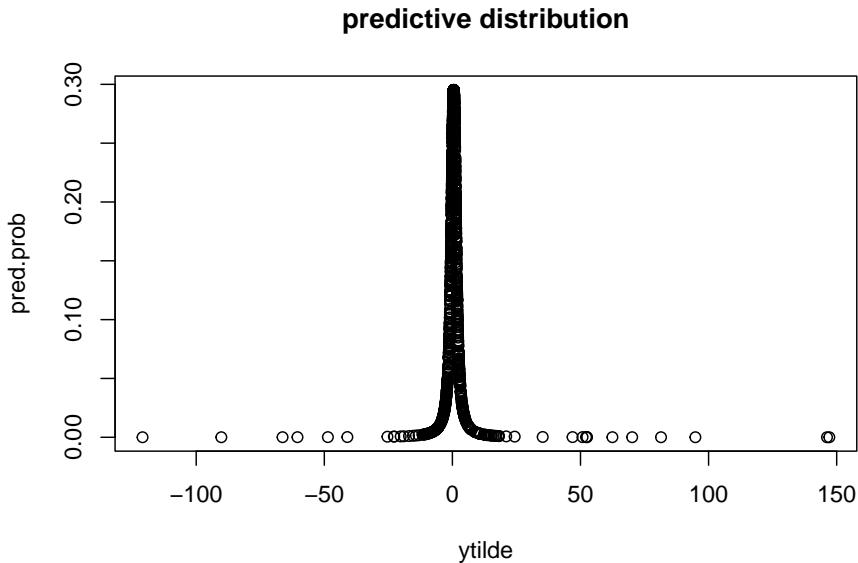
- (c) Using the previous 1000 samples of  $\theta$  to obtain 1000 samples from the predictive distribution of a future observation  $y_6$ , and plot the predictive draws. we use the sampled thetas from the posterior to sample from the Cauchy distribution. The predictive probability follows  $p(x|y) = \int p(x|\theta)p(\theta|y)d\theta$  where  $p(x|\theta)$  follows from the Cauchy distribution, given original sequence of thetas. We have the posterior values for each theta (given the uniform grid of thetas) and take the product. then for each predictive value, we sum the total probability across all thetas to compute the predictive probability. The maximum predictive value probability 0.52 with probability of 0.29.

```
## predictive distribution ? ??
# p(x | y) = int p(x|theta)*p(theta|y ) dtheta
## the posterior is p(theta|y)
# the likelihood p(x|theta) ## we use the sampled thetas using the posterior
ytilde<-rcauchy(1000,location=samps,scale=1)
## probability of the samples
#prob_samp<-post[match(samps,theta)]
summary(ytilde) ## we have a wide distribution of predictive values.
```

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
##	-120.9488	-0.5259	0.5074	0.9698	1.5091	147.1229

```
## for all predictive values, find the total probability
## int p(x(theta)*p(theta|y) d\theta note that p(x(theta) is a function of theta,
pred.prob<-sapply(ytilde,function(x) sum(dcauchy(x,location=theta,scale=1)*post*step))

plot(ytilde,pred.prob, main='predictive distribution')
```



```
## maximum predictive probability
message('max pred. prob ', round(ytilde[which(pred.prob==max(pred.prob))],3))

## max pred. prob 0.521
```

### Question 12

Suppose  $y|\theta \sim Poisson(\theta)$ , find Jeffreys' prior density for  $\theta$  and then find  $\alpha, \beta$  for which Gamma( $a, b$ ) density is a close match to Jeffreys' density.

$$\begin{aligned} \log(p(y|\theta)) &= y * \log(\theta) - \theta - \log(y!) \\ &\implies y' = y/\theta - 1, y'' = -y/\theta^2 \\ &\implies J(\theta) = E(-l'') = E(y/\theta^2) = 1/\theta \\ &= p(\theta) \propto |J(\theta)^{1/2}| = \sqrt{1/\theta} = \theta^{-1/2} \end{aligned}$$

So the closest prior is  $\alpha = 1/2, \beta = 0$ .

### Question 13

- (a) Use the normal approximation to gamma and poisson to determine a posterior for fatal accidents using table 2.2. Compute the 95% predictive interval.
- We set the empirical prior alpha =25 and beta =1 because the effective sample size ( $\beta$ ) is mean is approximately 25, and given these the quantiles of the prior distribution is 16.2 and 35.7 which contains the observed data well. and the prior mean is approximately 25 which is reasonably close to the table 2.2 values, although a sensitivity analysis is recommended.
- We use the conjugate prior for Gamma to find the posterior distribution, which closely is approximated by the normal distribution. The posterior distributions are similar. The normal approximation prior followed  $N(\alpha/\beta, \sqrt{\alpha}/\beta)$ . we set the variance to be known  $\sigma^2 = 100$ . Using the gamma conjugate prior the predictive interval is \$95% \$(21.1,26.9).
- The posterior predictive value was determined by sample values from the posterior, and using the parameters sampled from posterior to generate a Poisson random variable. The average predictive value is 23.9 fatal accidents.

-Using Jeffrey's prior we set alpha=1/2 and beta=0.02 which has a prior mean of 25 which matches the observed mean. The posterior using Jeffreys' prior credible interval 95% is (21.11, 26.88) which is similar to the empirical prior. Using the normal approximation to Jeffreys' prior has a 95% predictive interval of (14,34) with a predictive mean of 23.83. The posterior predictive value for 1986 was computed by sampling  $\theta$  from the posterior distribution, and for each  $\theta_j$ , we sampled a random variable from Poisson( $\lambda = \theta_j$ ).

-Using Jeffreys' prior, the posteror is  $p(\theta|y) \sim \text{Gamma}(238.5, 10.2)$ . Using simulation the 95% predictive invtrerval is (14,34). Using the normal approximation - Using the posterior predictiction equation  $E(x|y) = E(\theta|y) = \mu$ . and the  $V(x|y) = E(\theta|y) + V(\theta|y)$ . Using the normal approximation  $\mu_n = 23.8$  and  $\tau_n^2 = 9.9$ . Then the posterior variance is  $23.8+9.9 = 33.7$ . Then using  $\mu \pm 1.96\sigma_n = (12.4335,19)$ . The true observed value for 1986 is 22 fatal accidents.

```
# data entry
year<-seq(1976,1985)
fatal<-c(24,25,31,31,22,21,26,20,16,22)
death<-c(734,516,754,877,814,362,764,809,223,1066)
rate<-c(0.19,0.12,0.15,0.16,0.14,0.06,0.13,0.13,0.03,0.15)
data<-data.frame(year=year,fatal=fatal,pass.deaths=death,death.rate=rate)

## gamma prior on theta>0
```

```

theta=seq(from=0,to=50,by=.01)

## normal approximation to gamma
## Gamma(a,b) , mean = a/b, var = a/b^2 ## where beta is the inverse-scale
## gamma(a,b) ~ N (a/b, sd= sqrt(a)/b)
## normal approx to Pois
## Pois(a) , mean =a, var= a
## Pois(a) ~ N(a,sd= sqrt(a))

## gamma prior of total count a-1, in b observations
## empirical prior
## we approximate a prior to have an average of 25 deaths across the decade. the emp
# the beta parameter is set rate is set such that gamma(25,1) has quantiles 16.2, 3
alpha= 25
beta<- 1
## the mean rate given the data is 0.126, so we find the beta parameter to match the
n<-nrow(data)
message("the prior interval:",qgamma(c(0.025,0.975),alpha,beta))

## the prior interval:16.178681847829335.7100975937532

## non-informative prior using Jeffrey's prior
Jalpha= 1/2
Jbeta<- 0.02 ## we set close to 0
## prior mean is 25 which matches the observed data.
## the mean rate given the data is 0.126, so we find the beta parameter to match the
message("the prior interval:")

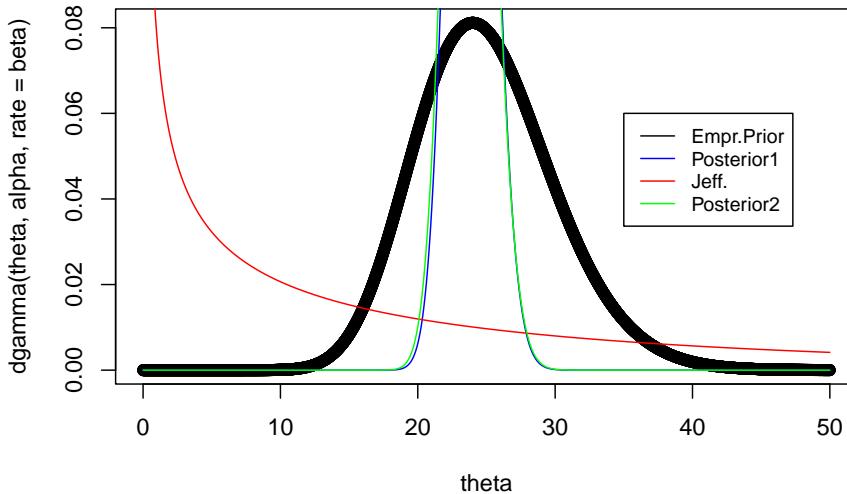
## the prior interval:

print(qgamma(c(0.025,0.975),Jalpha,Jbeta))

## [1] 0.02455173 125.59715468

plot(theta,dgamma(theta,alpha,rate=beta),lty=2)
lines(theta,dgamma(theta,alpha+n*mean(fatal),rate=beta+n),col='blue')
lines(theta,dgamma(theta,Jalpha,rate=Jbeta),col='red')
lines(theta,dgamma(theta,Jalpha+n*mean(fatal),rate=Jbeta+n),col='green')
legend(35, 0.06, legend=c("Empr.Prior", "Posterior1","Jeff.", "Posterior2"),
       col=c("black", "blue", "red", "green"), lty=1, cex=0.8)

```



```

## credible interval of the posterior
pois_credible<-qgamma(c(.025,0.975),alpha+n*mean(fatal),rate=beta+n)
message("posterior gamma interval n=10: ",pois_credible[1]," ",pois_credible[2]) ##

## posterior gamma interval n=10: 21.1065649790459 26.883779321636

## jeffrey's prior
pois_credible_jp<-qgamma(c(.025,0.975),Jalpha+n*mean(fatal),rate=Jbeta+n)
message("posterior gamma interval n=10: ",pois_credible[1]," ",pois_credible[2]) ##

## posterior gamma interval n=10: 21.1065649790459 26.883779321636

post_pois<-dgamma(theta,Jalpha+n*mean(fatal),rate=Jbeta+n)
## posterior mean is 23.8
sum(theta*post_pois/sum(post_pois))

## [1] 23.8024

# normal approximation using Jeffrey's Prior
mu_0<- Jalpha/Jbeta ## average of 12.6 fatal accidents
tau2_0<- Jalpha/Jbeta^2 ##

```

```

sigma2<-10^2 ## we assume we know the variance of the fatalities (empirical variance)

mu_n<-function(mu0,ybar,n,tau02,sigma2){
  mun<- (mu0/tau02 + (ybar*n)/sigma2)/(1/tau02 + n/sigma2)
  return(mun)
}

taun2<-function(tau02,n,sigma2){
  inv.taun2<- 1/tau02 +n/sigma2
  return(1/inv.taun2)
}

ybar=mean(fatal)

mn<-mu_n(mu_0,ybar,tau2_0,n,sigma2)
tn<-taun2(tau2_0,n,sigma2)
## posterior interval n=10
#lower<-round(qnorm(0.025,mu_n(mu0,ybar,tau02,10,sigma2),sd=sqrt(taun2(tau02,10,sigma2)))
print(c(mn-1.96*sqrt(mn+tn), mn+1.96*sqrt(mn+tn)))

## [1] 12.42630 35.19275

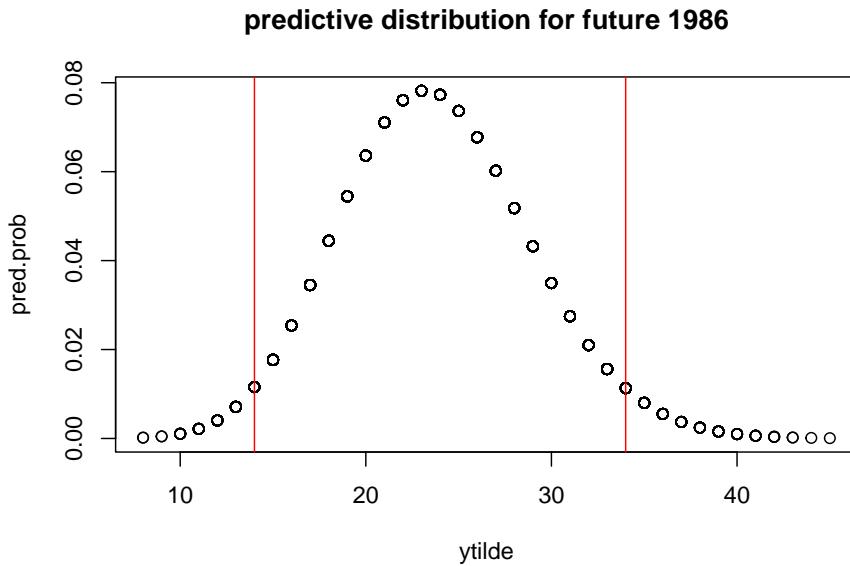
## posterior predictive value
## check this with the grid approach in the Gamma framework.
theta_samps<-sample(theta,10000,prob=post_pois,replace=T)
ytilde<-rpois(10000,lambda=theta_samps)
## probability of the samples
#prob_samp<-post[match(samps,theta)]
summary(ytilde) ## we have a wide distribution of predictive values.

##      Min. 1st Qu. Median Mean 3rd Qu. Max.
##     8.00 20.00 24.00 23.79 27.00 45.00

## for all predictive values, find the total probability
## int p(x|theta)*p(theta|y) d\theta note that p(x|theta) is a function of theta,
pred.prob<-sapply(ytilde,function(x) sum(dpois(x,lambda = theta)*post_pois/sum(post_pois)))

plot(ytilde,pred.prob, main='predictive distribution for future 1986')
abline(v=quantile(ytilde,c(0.025,0.975)),col='red')

```



```
#lines(theta,post_pois)
quantile(ytilde,c(0.025,0.975))

## 2.5% 97.5%
##      14      34
```

From the gamma posterior  $E(\theta|y) = 238.5/10.02 = 24.25$  and the variance  $V(\theta|y) = 238.5/10.02^2 = 2.87$ . From the Poisson likelihood  $E(x|\theta) = \theta$  and  $V(x|\theta) = \theta$ .  $V(x|y) = E(V(x|\theta,y)|y) + V(E(x|\theta,y)|y) = E(\theta|y) + V(\theta|y) = 23.382 + 2.87 = 26.252$

Using the normal approximation on the posterior values, the variance of the predictive value and is a larger interval (14.21, 34.3).

The average predictive value from 1986 is 24.3 which is close to the true value 22 fatal accidents.

```
## Predictive with normal

## using the normal approximation
## predictive interval must use a grid approach.
pred<-round(qnorm(c(0.025,0.975),24.25,sd=sqrt(26.252)),2)
message("predictive interval n=10: ",pred[1]," ",pred[2])

## predictive interval n=10: 14.21 34.29
```

```

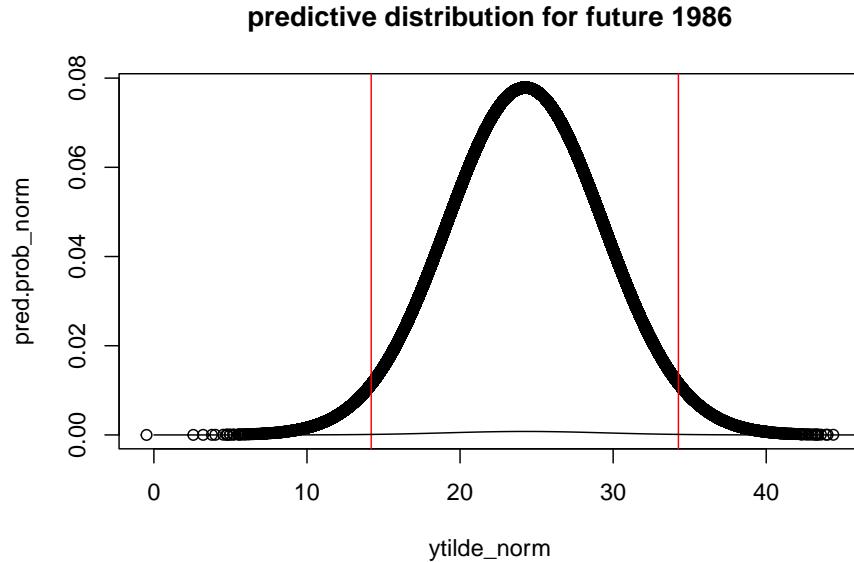
## check this with the grid approach in the normal framework.
post_pred_norm<-dnorm(theta,24.25,sqrt(26.252))/sum(dnorm(theta,24.25,sqrt(26.252)))
ytilde_norm<-rnorm(100000,mean=24.25,sd=sqrt(26.252))
## probability of the samples
#prob_samp<-post[match(samps,theta)]
summary(ytilde_norm) ## we have a wide distribution of predictive values.

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## -0.4808 20.7783 24.2597 24.2506 27.6996 44.3823

## for all predictive values, find the total probability
## int p(x|theta)*p(theta|y) d\theta note that p(x|theta) is a function of theta,
pred.prob_norm<-sapply(ytilde_norm,function(x) sum(dnorm(x,24.25,sqrt(26.252))*post_prob_norm))

plot(ytilde_norm,pred.prob_norm, main='predictive distribution for future 1986')
abline(v=quantile(ytilde_norm,c(0.025,0.975)),col='red')
lines(theta,post_pred_norm)

```



```
quantile(ytilde_norm,c(0.025,0.975))
```

```

##      2.5%    97.5%
## 14.19498 34.26337

```

```
message("average predictive value from posterior:", mean(ytilde))

## average predictive value from posterior:23.7858
```

- (b) Assume fatal accidents follow a Poisson distribution with a constant rate and an exposure in each year proportional to the number of passenger miles flown. Set a prior and determine a posterior and predictive distribution. estimate the number of passenger miles flown in each year by dividing

let  $\theta$  be the fatal accidents *rate* per 100,000 million miles flown. and let  $x_i$  be the number per 100,000 million miles flown. We compute the miles by dividing the passenger deaths divided by the death rate (deaths per 100,000 million miles flown).

We then compute the fatal accident rate by computing the (passanger deaths per 100,000 million miles)\*(number of fatal accidents / number of passenger deaths). We let  $\theta$  = fatal accidents per 100 million miles flown.

The mean accident rate is 0.004, so we set alpha =5 and beta = 1000, so the prior mean is 0.005 (95%: 0.002,0.01) which is similar to the observed ranges. This is using the observed data to create an empirically driven prior.

However, we must use the Jeffreys' prior for a non-informative prior setting  $\alpha = 0.01, \beta = 0.01$ .

Using Jeffreys' prior, the posterior is  $\text{Gamma}(238.01, 5.716e12)$  with a 95 credible interval of (3.65e-11, 4.71e-11).

Now for the posterior predictive value, because the parameters space is so small, we sample  $\theta$  from  $p(\theta|y)$  Gamma posterior to generate  $\theta$ . In the previous exercise, we used theta from a uniform sequence, but these values are so small due to the scaling it is easier to sample directly from the posterior using `rgamma`. THen we find the predictive value  $x \sim \text{Pois}(8e11 * \theta_j)$  by random sampling observation variables from the posterior, and sorting them for the 95 predictive interval.

We can also use the normal approximation, but the equations for the mean and variance must be scaled by the factor 8e11.

```
miles100<-death/rate ## 100 million miles
miles = miles100*1e+08

## likelihood
## y | miles, theta ~ Pois( miles*theta)
```

```

theta=seq(from=0,to=0.1,by=.001)
## empirical prior
alpha= 5
beta<- 1000
## the mean is 0.005 and quantiles 0.002 and 0.01 which is close to the observed rate
#mean(acc_rate) ## fatal accidents per 100,000 million miles driven
print("empirical prior 95% quantiles: ")

## [1] "empirical prior 95% quantiles: "

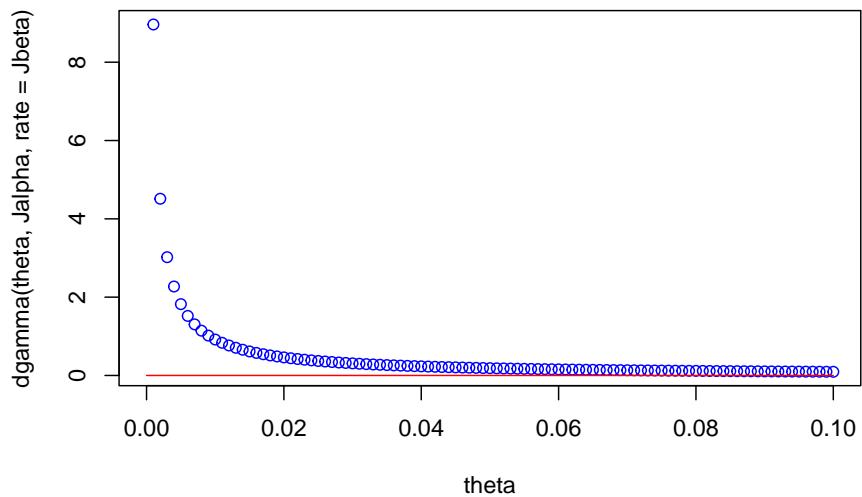
qgamma(c(0.025,0.975),alpha,rate=beta)

## [1] 0.001623486 0.010241589

## the mean rate given the data is 0.126, so we find the beta parameter to match the
n<-nrow(data)

## we must use Jeffreys' prior!!
Jalpha=0.01
Jbeta<-0.01
plot(theta,dgamma(theta,Jalpha,rate=Jbeta),col='blue')
lines(theta,dgamma(theta,Jalpha+sum(fatal),rate=Jbeta+sum(miles)),col='red')

```



```

## credible interval of the posterior
pois_credible<-qgamma(c(.025,0.975),Jalpha+sum(fatal),rate=Jbeta+sum(miles))
message("Jeffreys' prior posterior gamma interval n=10: ")

## Jeffreys' prior posterior gamma interval n=10:

print(pois_credible) ##

## [1] 3.651772e-11 4.709401e-11

## generate parameters from the posterior we sample thetas.
theta_post<-rgamma(1000,Jalpha+sum(fatal),rate=Jbeta+sum(miles))
y1986<-rpois(1000,8e+11*theta_post) ## use in the likelihood
# message("predictive 1986 interval: ")
# sort(y1986)[c(25,976)]

### the original sequenced theta do not produce parameters from the posterior and this fails.
# post_pois<-dgamma(theta,Jalpha+sum(fatal),rate=Jbeta+sum(miles))
# theta_samps<-sample(theta,100000,prob=post_pois,replace=T)
# ytilde<-rpois(100000,lambda=8e+11*theta_samps)

post_a= Jalpha+sum(fatal)
post_b= Jbeta+sum(miles)
## use the normal approximation
##  $E = a/b$   $V = a/b^2$ 
mn_a = post_a/post_b
tn_a = post_a/post_b^2
##  $xi * E(\theta | y) = xi * \mu_1$ 
mu_approx = mn_a*8e+11
##  $V(x|y) = xi * E(\theta/y) + xi^2 * V(\theta/y)$  using the poisson likelihood and xi is the const
sigma2_approx<-mn_a*8e+11+(8e+11)^2*tn_a

## normal approx
upper<- mu_approx+1.96*sqrt(sigma2_approx) ## 46
lower<- mu_approx-1.96*sqrt(sigma2_approx) ## 22
message("normal approx: 95% pred. interval ")

## normal approx: 95% pred. interval

```

```

print(c(lower,upper))

## [1] 21.23396 45.39038

message(" 8,000 per 100 million miles flown has fatality rate: Gamma posterior")

## 8,000 per 100 million miles flown has fatality rate: Gamma posterior

print(sort(y1986)[c(25,976)])

## [1] 21 46

message(" 8,000 per 100 million miles flown has fatality rate: normal approx")

## 8,000 per 100 million miles flown has fatality rate: normal approx

print(c(lower,upper))

## [1] 21.23396 45.39038

```

- (c) we repeat part (a) but for passenger deaths. we need to find a prior and determine the posterior distribution with a predictive interval.

We used an empirically derived prior, which does not match the solution, and so we include Jeffreys' prior setting  $\alpha = 1/2, \beta = 0.01$ .

Note that the death rate mean is not equal to the variance so the performance is not as good compared to the fatal accidents.

The 1986 prediction mean is 692 95% (638, 750) using the empirical derived priors estimated from the data.

If we use Jeffreys' prior the prediction interval is (638,741.05).

which is not the expected prediction of 546 passenger deaths. This is likely because Poisson is not a good fit because it is constrained.

```

theta=seq(from=200,to=1500,by=.1)

## normal approximation to gamma
## Gamma(a,b) , mean = a/b, var = a/b^2 ## where beta is the inverse-scale
## gamma(a,b) ~ N (a/b, sd= sqrt(a)/b)
## normal approx to Pois

```

```

## Pois(a) , mean =a, var= a
## Pois(a) ~ N(a, sd= sqrt(a))

## gamma prior of total count a-1, in b observations
##
## we approximate a prior to have an average of 600 passenger deaths across the decade.
#the empirical mean is 691 and variance is 63700.32
# a/b = 691 and a/b^2 = 63700.32, we solve for a and b
# the beta parameter is set rate is set such that 404, 832 which matches well the observed da
alpha= 8
beta<- 0.0108618
## the pr
n<-nrow(data)
message("empricial prior: ")

## empricial prior:

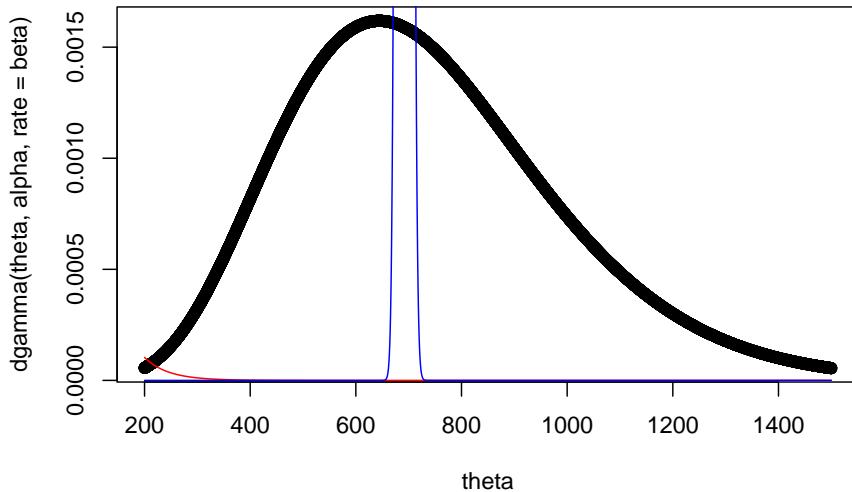
print(qgamma(c(0.025,0.795),alpha,beta))

## [1] 317.9797 936.6278

Jalpha<- 1/2
Jbeta<- 0.02

plot(theta,dgamma(theta,alpha,rate=beta),lty=2)
lines(theta,dgamma(theta,Jalpha,rate=Jbeta),col='red')
lines(theta,dgamma(theta,alpha+n*mean(death),rate=beta+n),col='blue')

```



```

## credible interval of the posterior
pois_credible<-qgamma(c(.025,0.975),alpha+n*mean(death),rate=beta+n)
message("posterior for empirical gamma interval n=10: ",round(pois_credible[1],2),"\n")

## posterior for empirical gamma interval n=10: 675.75 708.338

jp_pois_credible<-qgamma(c(.025,0.975),Jalpha+n*mean(death),rate=Jbeta+n)
message("posterior for Jeffreys' gamma interval n=10: ",round(jp_pois_credible[1],2),"\n")

## posterior for Jeffreys' gamma interval n=10: 674.39 706.934

## produces too small probabilities from Jeffrey's and we can sample here
post_pois<-dgamma(theta,alpha+n*mean(death),rate=beta+n) ## prob
## prediction: with the grid approach in the Gamma framework does not work for small
theta_samps<-sample(theta,1000,prob=post_pois,replace=T)
ytilde_empiricalPrior<-rpois(1000,lambda=theta_samps)
message("using the empirical prior, predictive posterior interval")

## using the empirical prior, predictive posterior interval

quantile(ytilde_empiricalPrior,c(0.025,0.975))

```

```

## 2.5% 97.5%
## 636    746

## direct sampling from posterior.
theta_post<-rgamma(1000,Jalpha+n*mean(death),rate=Jbeta+n)

ytilde<-rpois(1000,lambda=theta_post)
## probability of the samples
#prob_samp<-post[match(samps,theta)]
summary(ytilde) ## we have a wide distribution of predictive values.

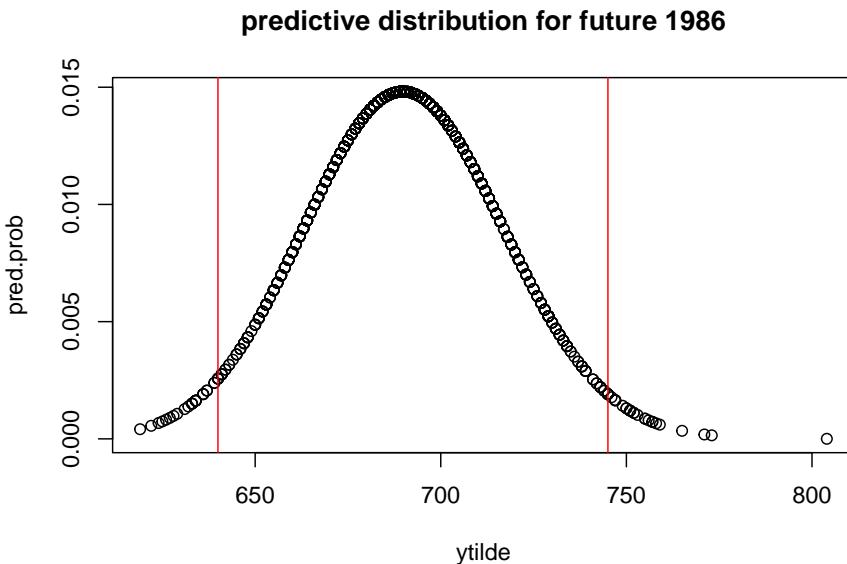
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
## 619.0   671.0   690.0   690.6   708.0   804.0

post_pois<-dgamma(theta_post,Jalpha+n*mean(death),rate=Jbeta+n)

## for all predictive values, find the total probability
## int p(x|theta)*p(theta|y) d\theta note that p(x|theta) is a function of theta, so we input
pred.prob<-sapply(ytilde,function(x) sum(dpois(x,lambda = theta_post)*post_pois/sum(post_pois)))

plot(ytilde,pred.prob, main='predictive distribution for future 1986')
abline(v=quantile(ytilde,c(0.025,0.975)),col='red')

```



```
# lines(theta,post_pois)
quantile(ytilde,c(0.025,0.975))
```

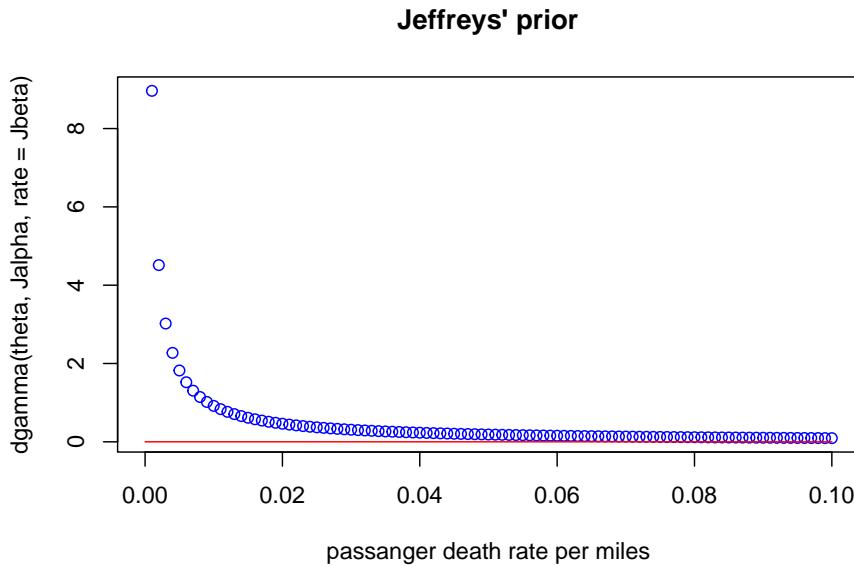
```
##      2.5%    97.5%
## 639.975 745.025
```

- (d) Using the rate the posterior rate mean is 0.12 estimated from sampling the posterior distribution. For  $8 \times 10^1$  miles flown the deaths are approximately 970 (902, 1039) this is higher than the 1986 observed value.

Note we used Jeffreys' prior here, but an empirical prior is okay

```
theta=seq(from=0,to=0.1,by=.001)
miles100<-death/rate ## 100 million miles
miles<-miles100*100000000

## we must use Jeffreys' prior
Jalpha=0.01
Jbeta<-0.01
plot(theta,dgamma(theta,Jalpha,rate=Jbeta),col='blue',xlab='passanger death rate per m')
lines(theta,dgamma(theta,Jalpha+sum(death),rate=Jbeta+sum(miles)),col='red')
```



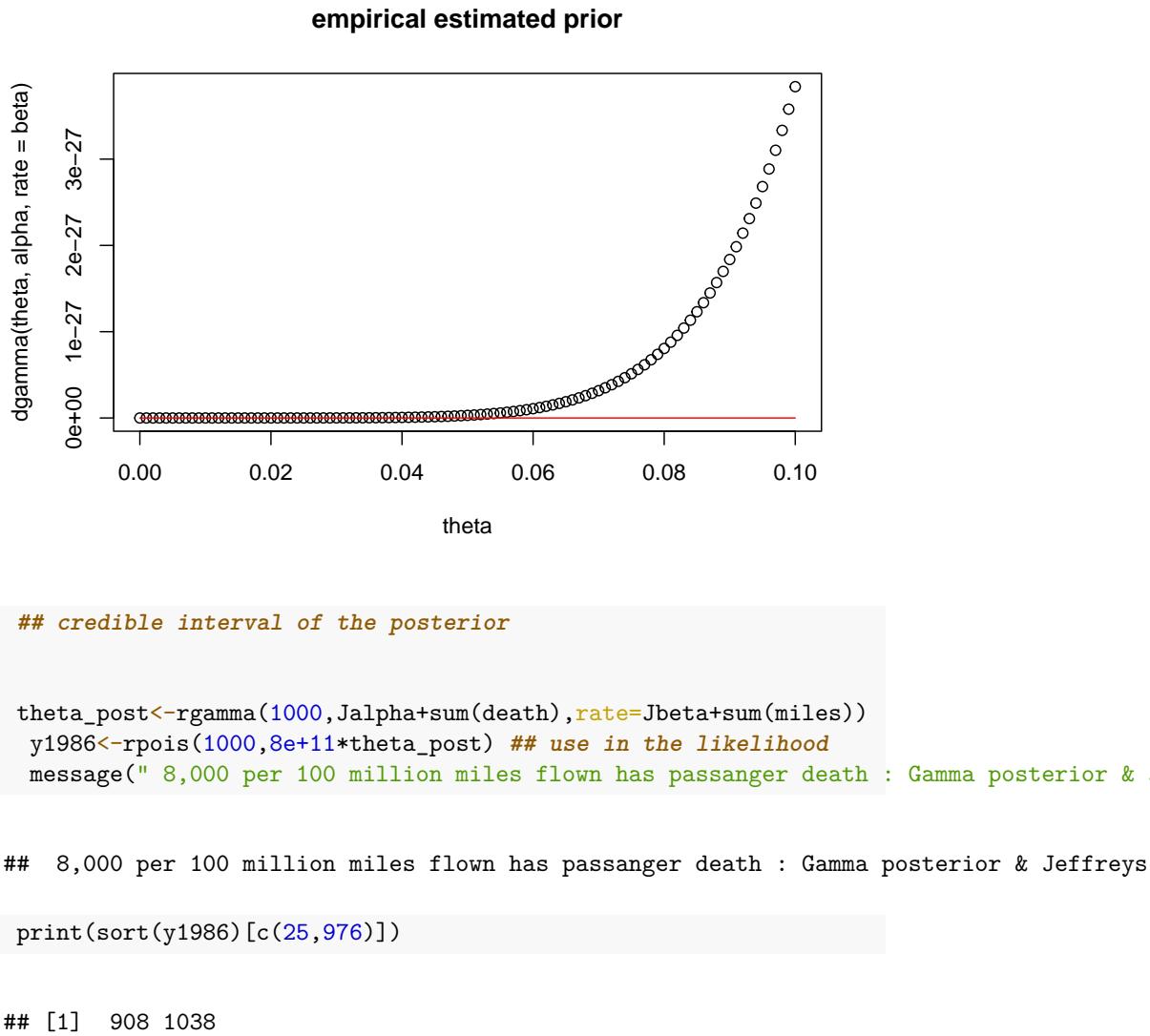
```
## credible interval of the posterior
pois_credible<-qgamma(c(.025,0.975),Jalpha+sum(death),rate=Jbeta+sum(miles))
message("Jeffreys' prior posterior gamma interval n=10: ")
```

```
## Jeffreys' prior posterior gamma interval n=10:
```

```
print(pois_credible) ##
```

```
## [1] 1.182135e-09 1.239179e-09
```

```
plot(theta,dgamma(theta,alpha,rate=beta),lty=2, main='empirical estimated prior')
lines(theta,dgamma(theta,alpha+sum(death),rate=beta+sum(miles)),col='red')
```



### 2.8.1 Question 21

estimate percentage of adult population in each state (excluding alaska and hawaii) who label themselves as very liberal. plot estimate vs. obama's vote share in 2008.

For each state we measure  $y_j \sim Pois(n_j\theta_j)$  from section 2.7 where  $y$  is the number of adult population who identify themselves as *very liberal* in ideology for a given state.

we do not have population of each state, but we do have the number of respondents of each state, and denote  $n_j$  as the integer representing the total respondents.

Alternatively, we do have census density category for each state (1-5) which can account for the populaiton density.

- (a) graph the proportion of liberal in each state vs obama vote share.  
here the proportion of liberal votes was weighted by the total respondents

we see that the highest proportion of very liberal polls is associated with the states with the highest obama percentage of votes in CA, NY, WA, and IL. Conversely we see that WY UT, OK are very red states with low polls in very liberal voters, with the lowest support for obama.

For those that identify as *liberal* we see a much linear trend.

```
library(foreign)
library(dplyr)

## 
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
## 
##     filter, lag

## The following objects are masked from 'package:base':
## 
##     intersect, setdiff, setequal, union

library(magrittr)
library(ggplot2)
#polls<-read.dta("C:/Users/UOSC/Documents/Keck-graduate-school/PM590-Bayesian/bookdown-files/pew_"
polls<-read.dta("C:/Users/antho/Documents/Keck-graduate-school/PM590-Bayesian/bookdown-files/pew_"
  #proportion of very liberal
  table(polls$ideo)

## 
## missing/not asked very conservative      conservative      moderate
##          0             2417              9795            11197
##     liberal      very liberal      dk/refused      1380
##          4535            1470
```

```

## adult pop
table(polls$age)

## 
##  18   19   20   21   22   23   24   25   26   27   28   29   30   31   32   33   34   35   36   37
## 628  431  363  341  316  374  355  343  406  337  391  398  446  366  375  378  334  398  417  482
## 38   39   40   41   42   43   44   45   46   47   48   49   50   51   52   53   54   55   56   57
## 509  441  617  464  542  470  529  582  571  615  639  593  891  598  727  618  639  731  600  541
## 58   59   60   61   62   63   64   65   66   67   68   69   70   71   72   73   74   75   76   77
## 662  561  799  584  590  419  489  596  422  426  373  368  494  294  358  266  256  323  260  269
## 78   79   80   81   82   83   84   85   86   87   88   89   90   91   92   93   94   95   96   97
## 233  220  337  195  179  143  126  132  106  97   81   73   62   25   21   13   9    7   2   18
## 99
## 517

table(polls$age2)

## 
##      18-29       30-49       50-64      65+ dk\\refused
##      4683        9768        9449        6784        517

## each state
table(polls$state)

## 
##      alabama      alaska      arizona      arkansas      california
##          624           0          542          307          2854
##      colorado connecticut      delaware  washington dc      florida
##          468         395          119          62          1747
##      georgia      hawaii      idaho      illinois      indiana
##          1023           1          140         1130          829
##      iowa      kansas      kentucky      louisiana      maine
##          441         329          523          603          154
##      maryland massachusetts      michigan      minnesota mississippi
##          593         685          1000          711          264
##      missouri      montana      nebraska      nevada new hampshire
##          780           91          215          202          160
##      new jersey      new mexico      new york north carolina      north dakota
##          870           219          1701         1055          120
##      ohio      oklahoma      oregon      pennsylvania      rhode island
##          1404          431          468          1591          131
##      south carolina      south dakota      tennessee      texas      utah
##          458            93          745         1919          284

```

```

##      vermont      virginia      washington      west virginia      wisconsin
##          115           896            668             270            742
##      wyoming
##          29

# drop AL and HA
polls<-polls[which(polls$state!="alaska" & polls$state!="hawaii"),]
# average the points that are missing.
polls$density[which(is.na(polls$density))]<-mean(polls$density[which(!is.na(polls$density))])

#ele<-read.csv("C:/Users/UOSC/Documents/Keck-graduate-school/PM590-Bayesian/bookdown-files/2008E")
ele<-read.csv("C:/Users/antho/Documents/Keck-graduate-school/PM590-Bayesian/bookdown-files/2008E")
ele<-ele[which(ele$state!="Alaska"& ele$state!="Hawaii"),]
ele$state<-tolower(ele$state)
## graph proportion liberal in each state vs obama vote share scatter.

yj<-as.data.frame.matrix(table(polls$state,polls$ideo))
nj<-as.data.frame.matrix(table(polls$state,polls$density))
## ideology
head(yj)

##      missing/not asked      very conservative      conservative      moderate      liberal
## alabama          0              82                254               165               48
## alaska           0                 0                  0                  0                  0
## arizona          0              33                152               201               95
## arkansas         0              30                115               101               29
## california      0              178                728              1054              549
## colorado         0                 31                134               165               96
##      very liberal      dk/refused
## alabama         30              38
## alaska           0                 0
## arizona          28              24
## arkansas         7                22
## california     179              121
## colorado        27                12

## density
head(nj)

##      1      2 2.8302567571921      3      4      5
## alabama  247  196                   9 170      2      0
## alaska    0    0                   0  0      0      0
## arizona   247   1                   0 294      0      0
## arkansas  148 120                   4  35      0      0

```

```

##  califonia 455 496          19 438 566 880
##  colorado   123 157          7 101   1  79

tally<-polls%>%group_by(state,ideo)%>%summarize(n=n())

## `summarise()` has grouped output by 'state'. You can override using the
## ` `.groups` argument.

statetotal<-tally%>%group_by(state)%>%summarise(statetotal=sum(n))
tally<-left_join(tally,statetotal,by='state')
## average the density of the polling regions by state taking the density of each zip
densit<-polls%>%group_by(state)%>%summarize(density=mean(density))

# highest states with highest density
densit[order(densit$density,decreasing = T),]

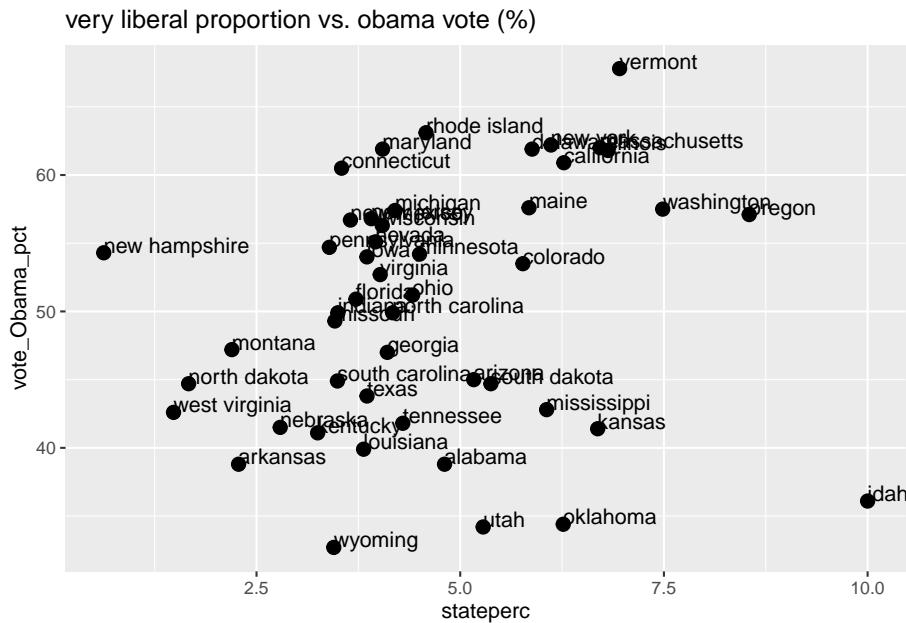
## # A tibble: 49 x 2
##       state      density
##   <fct>     <dbl>
## 1 washington dc    4.95
## 2 new jersey      3.99
## 3 rhode island    3.93
## 4 maryland        3.62
## 5 new york        3.60
## 6 massachusetts   3.58
## 7 connecticut     3.55
## 8 illinois        3.42
## 9 california      3.32
## 10 ohio           3.32
## # ... with 39 more rows
## # i Use `print(n = ...)` to see more rows

tally<-left_join(tally,densit,by='state')
tally$popdensity<-tally$n/tally$statetotal
tally$stateperc<-100*tally$n/tally$statetotal
tally$stateperc_density<-100*tally$n/(tally$statetotal*tally$density)

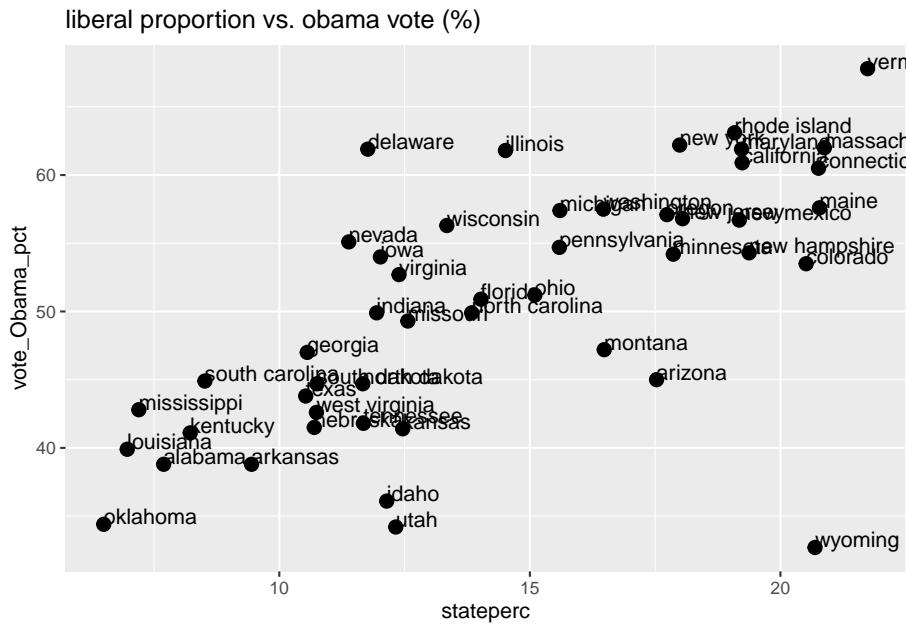
tt<-left_join(ele,tally,by="state")

ggplot(tt[which(tt$ideo=='very liberal'),],aes(x=stateperc,y=vote_Obama_pct,label=state)

```



```
ggplot(tt[which(tt$ideo=='liberal')],aes(x=stateperc,y=vote_Obama_pct,label=state))+geom_point(s
```



- (b) Graph the posterior Bayes mean vs. the Obama vote share. In order to construct the prior we use equations (2.17, 2.18) where we

match the moments from the prior predictive distribution to the observed data for each state.  $p(\tilde{y}) \sim Neg-Bin(\alpha, \beta)$  follows a negative binomial distribution.

```
## we need to understand the underlying rates in a prior for gamma given multiple states
vl<-subset(tally,tally$ideo=='very liberal')
hist(vl$stateperc, xlab="very lib. (%)", ylab="frequency")
```



```
## for each state we must identify the estimate

## empirical bayesian prior for percentage of respondents
emp.mean<-mean(vl$stateperc)
emp.var<-var(vl$stateperc)

beta=beta<-1/((var(vl$stateperc)-(mean(1/vl$density))*emp.mean)/(emp.mean))
alpha = beta*emp.mean

alpha_unif= 0.01
beta_unif = 0.01

qgamma(c(0.025,0.975),alpha,beta)

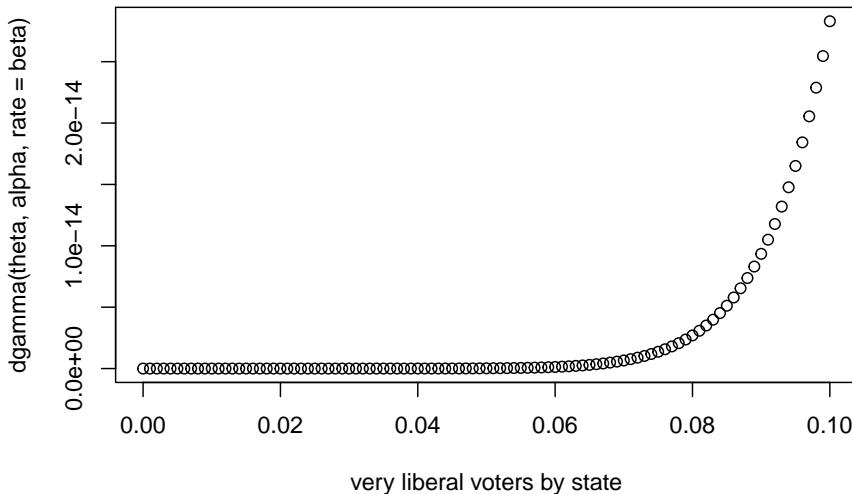
## [1] 2.422521 7.788166
```

```
quantile(tally$prop,c(0.025,0.975))

## Warning: Unknown or uninitialised column: `prop`.

## 2.5% 97.5%
##      NA      NA

theta=seq(from=0,to=0.1,by=.001)
plot(theta,dgamma(theta,alpha,rate=beta),lty=2, xlab='very liberal voters by state')
lines(theta,dgamma(theta,alpha_unif,rate=beta_unif),col='red')
```



```
dat<-tt[which(tt$ideo=='very liberal'),]

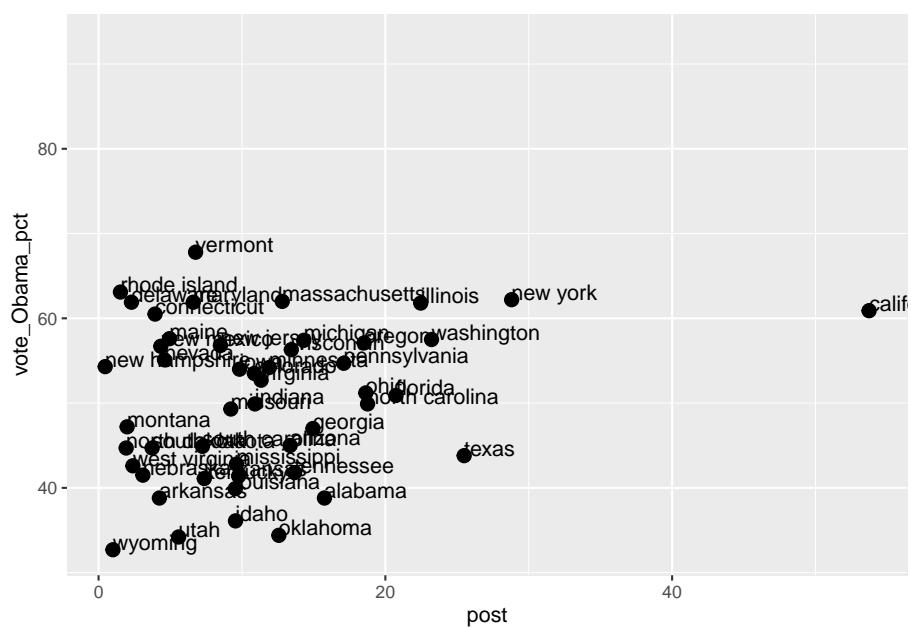
post<-function(theta,yj,nj){
  dd<-dgamma(theta,alpha+yj,rate=beta+nj)
  return(dd)
}
postMean<-function(alpha,yj,beta,nj){
  (alpha+yj)/(beta+nj)
}

post_thetaj<- postMean(alpha_unif,dat$n,beta_unif,dat$density)
```

```
pp<-data.frame(state=dat$state,post=postMean(alpha_unif,dat$n,beta_unif,dat$density))
tt2<-left_join(ele,pp,by="state")
ggplot(tt2,aes(x=post,y=vote_Obama_pct,label=state))+geom_point(size=3)+geom_text(hjust=0)

## Warning: Removed 1 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_text).
```



**Chapter 3**

**Chapter 3**



# Exercises

## 3.1 Question 1

- (a) find the marginal posterior of  $\alpha = \theta_1/(\theta_1 + \theta_2)$  We need to find the joint posterior of  $\theta_1, \theta_2$  and then perform a change of variables to  $(\alpha, \beta)$ . Let the prior  $p(\theta) = \text{Diri}(a_1, \dots, a_J)$  where  $y$  follows a multinomial likelihood.

$$\begin{aligned} p(\theta|y) &= \prod \theta_j^{y_j} \theta_j^{a_j-1} \\ &= \prod \theta_j^{y_j+a_j-1} \\ &\sim \text{Diri}(y_j + a_j) \end{aligned}$$

The posterior follows a Dirichlet distribution, but we are interested only in the  $\theta_1, \theta_2$  parameters and need to find the marginal posterior of these subvectors. The prior marginal follows for  $a_0 = \sum a_i$  and  $y_0 = \sum y_i$

From the Appendix A the joint posterior of the sub vector is

$$\begin{aligned} (\theta_1, \theta_2, 1 - \theta_1 - \theta_2 | y) &\sim \text{Diri}(y_1 + a_1, y_2 + a_2, a_0 + y_0 - y_1 - y_2 - a_1 - a_2) \\ p(\theta_1, \theta_2, 1 - \theta_1 - \theta_2) &\propto \theta_1^{y_1+a_1-1} \theta_2^{y_2+a_2-1} (1 - \theta_1 - \theta_2)^{a_0+y_0-y_1-y_2-a_1-a_2-1} \end{aligned}$$

In order to find the distribution of  $\theta_1/(\theta_1 + \theta_2)$  we need the jacobian, let  $(\alpha, \beta) = (\theta_1/(\theta_1 + \theta_2), \theta_1 + \theta_2)$ . Rearranging the terms we have  $\theta_1 = \alpha * \beta$ , and  $\theta_2 = \beta(1 - \alpha)$

$$\begin{aligned} |J| &= \begin{bmatrix} \beta & \alpha \\ -\beta & (1 - \alpha) \end{bmatrix} \\ &= |\beta| \end{aligned}$$

where  $a'_0 = \sum a_i - a_1 - a_2$ , and  $y'_0 = \sum y_i - y_1 - y_2$

$$\begin{aligned} p(\alpha, \beta) &= (\alpha * \beta)^{y_1+a_1-1} (\beta(1 - \alpha))^{y_2+a_2-1} (1 - \alpha * \beta - \beta(1 - \alpha))^{a'_0+y'_0-1} |\beta| \\ &= \alpha^{y_1+a_1-1} (1 - \alpha)^{y_2+a_2-1} \beta^{y_1+a_1+y_2+a_2-1} (1 - \beta)^{a'_0+y'_0-1} \\ &= \text{Beta}(y_1 + a_1, y_2 + a_2) \text{Beta}(y_1 + a_1 + y_2 + a_2, a'_0 + y'_0) \end{aligned}$$

By factorization we integrate out with respect  $\beta$  which yields the marginal posterior for  $\alpha \sim Beta(y_1 + a_1, y_2 + a_2)$

- (b) To show the relation with binomial let the likelihood  $p(y|\alpha) \propto \alpha^y(1-\alpha)^{y_2}$  and a  $p(\alpha) = Beta(a_1, a_2)$  then the posterior is

$$\begin{aligned} p(\alpha|y) &= p(y|\alpha)p(\alpha) \\ &= \alpha^{y_1+a_1-1}(1-\alpha)^{y_2+a_2-1} \\ &= Beta(y_1 + a_1, y_2 + a_2) \end{aligned}$$

## Question 2

639 were polled before the debate and 639 different persons were polled afterward. we let  $j=1,2$  let  $\alpha_j$  be proportion of voters who preferred Bush out of those who had a preference for bush or Dukakis at the time of the survey  $j$ . we need to model two different multinomial distributions, and find the posterior probability in  $\alpha_2 - \alpha_1$ . what was the posterior probability for support of Bush?

There was a 0.355 probability of a shift toward Bush after both debates.

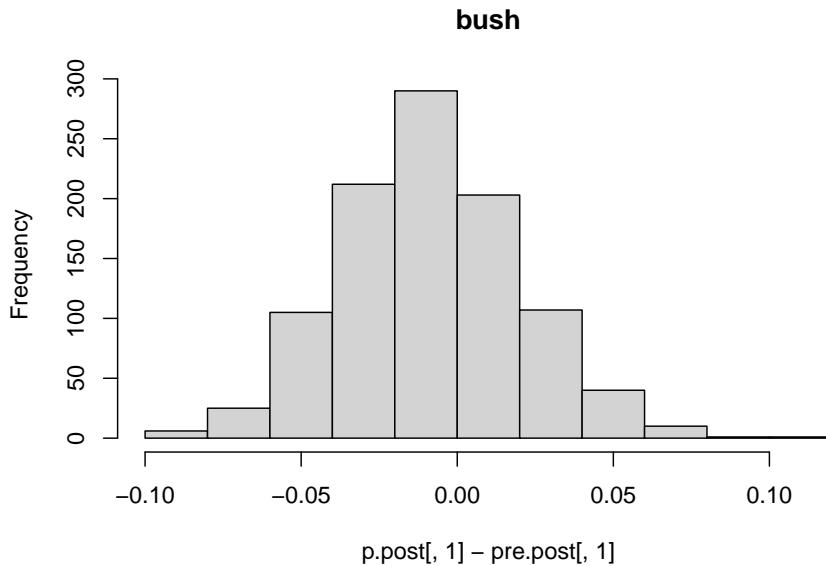
```
library(ggplot2)
library(magrittr)
library(dplyr)
debate<-data.frame(survey=c("pre","post"), bush=c(294,288), dukakis=c(307,332),none=c(338,313))

# we have 3 outcomes bush,duk, none
## we examine the proportion.

# need to set a theta parameter with sum theta =1
theta1<-seq(0.45,1,by=0.01)/2
theta2<-seq(0.45,1,by=0.01)/2
theta3<-1-(theta1+theta2)
all(theta1+theta2+theta3 ==1) ## sums to 1 for each j

## [1] TRUE

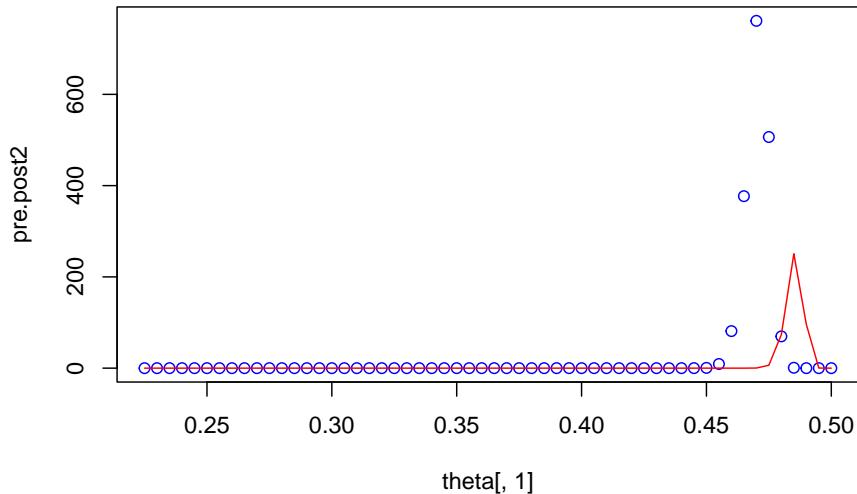
theta<-data.frame(theta1,theta2,theta3)
## need the pre posterior
## the prior alpha1, alpha2 =0 for improper prior
library(gtools)
pre.post<-rdirichlet(1000, c(as.numeric(debate[1,2:4])))
p.post<-rdirichlet(1000, c(as.numeric(debate[2,2:4])))
## differences between bush post .vs pre
hist(p.post[,1]-pre.post[,1],main='bush')
```



```
table((p.post[, 1]-pre.post[, 1])>0) ## 0.355% supported bush post debate.
```

```
##  
## FALSE TRUE  
## 638 362
```

```
pre.post2<-apply(theta[,1], function(x) ddirichlet(x,as.numeric(debate[1,2:4])))  
post.post2<-apply(theta[,1], function(x) ddirichlet(x,as.numeric(debate[2,2:4])))  
  
plot(theta[,1],pre.post2,col='blue',lty=2)  
lines(theta[,1],post.post2,col='red')
```



### Question 3

we are given two independent normal random variables with unknown variances and unknown true means. - (a) Assume a uniform prior on  $(\mu_c, \mu_t, \log(\sigma_c), \log(\sigma_t))$  find the posterior of  $\mu_c$  and  $\mu_t$ .

```

nc =32
nt= 36
mc = 1.013
sdc = 0.24
mt = 1.173
sdt = 0.24

### unknow true mean/variances we only have sample data.
## we have two unknowns and need to find the joint posterior distribution.

# the marginal posterior of mc follows a t-dist.
mc_range<-c(mc-sdc/sqrt(nc)*qt(0.975,df=nc-1),mc+sdc/sqrt(nc)*qt(0.975,df=nc-1))
mt_range<-c(mt-sdt/sqrt(nt)*qt(0.975,df=nt-1),mt+sdt/sqrt(nt)*qt(0.975,df=nt-1))

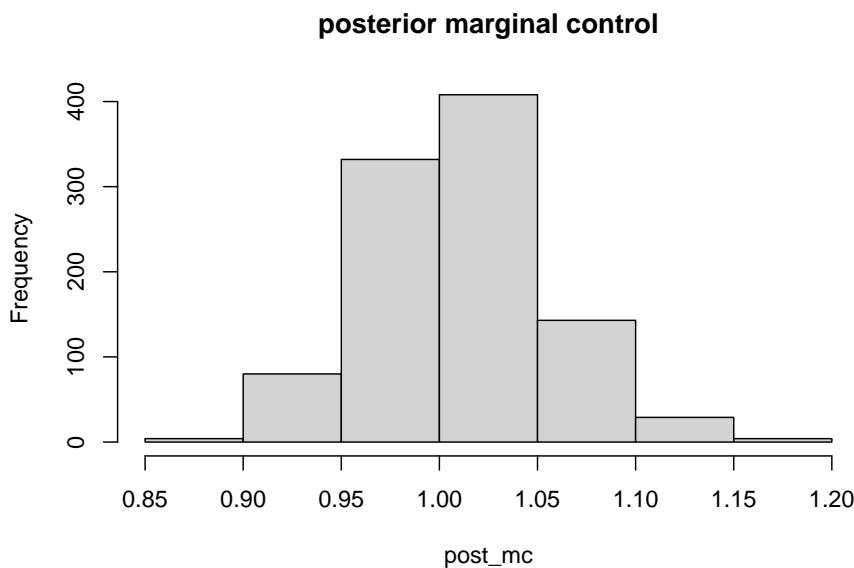
## to sample from the posterior we use 3.5 and 3.3 equations
posterior_sample<-function( mc=1, sdc=1, nc=10){
  invx2<-((nc-1)*sdc^2)/rchisq(1,nc-1)
}

```

```

post.mu<-rnorm(1,mean=mc,sd=sqrt(invx2/nc))
return(post.mu)
}
post_mc<-sapply(1:1000,function(x) posterior_sample(mc=mc,sdc=sdc,nc=nc))
hist(post_mc,main='posterior marginal control')

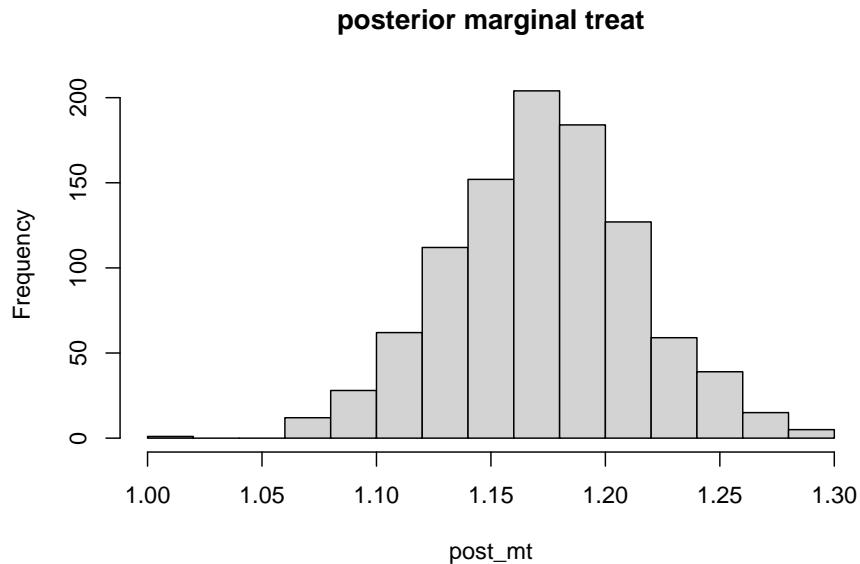
```



```

post_mt<-sapply(1:1000,function(x) posterior_sample(mc=mt,sdc=sdt,nc=nt))
hist(post_mt,main='posterior marginal treat')

```



```

message("95% posterior margin control:", round(mc_range[1],2)," ",round(mc_range[2],2))

## 95% posterior margin control:0.93 1.1

message("95% posterior margin control:", round(mt_range[1],2)," ",round(mt_range[2],2))

## 95% posterior margin control:1.09 1.25

quantile(post_mc,c(0.025,0.975))

##      2.5%      97.5%
## 0.9259494 1.1042008

quantile(post_mt,c(0.025,0.975))

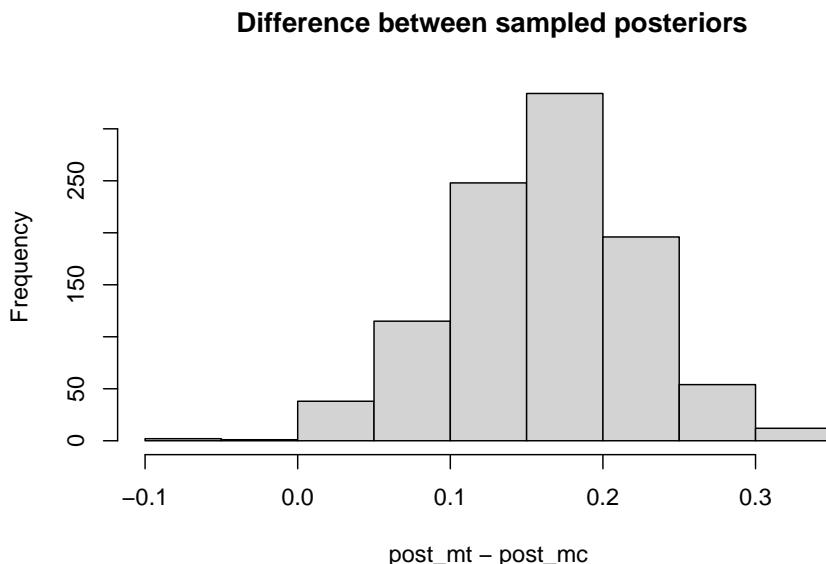
##      2.5%      97.5%
## 1.090623 1.255044

• (b) what is  $\mu_t - \mu_c$ 

```

The posterior interval (central) of the differences between 2 independent t-distributions is 0.16 (0.04, 0.28), which closely matches the posterior simulation interval (0.0338, 0.281). Here since both are independent we use  $X - Y \sim N(\mu_x - \mu_y, sd_x + sd_y)$  as the sampling distribution.

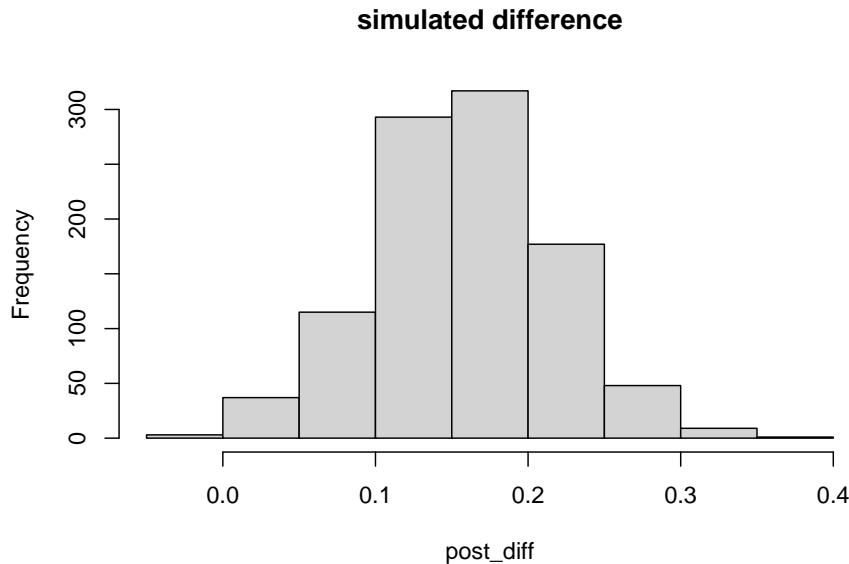
```
hist(post_mt-post_mc, main = ' Difference between sampled posteriors')
```



```
quantile(post_mt-post_mc,c(0.025,0.975))
```

```
##      2.5%      97.5%
## 0.03648754 0.27771995
```

```
## diff
post_diff<-sapply(1:1000,function(x) posterior_sample(mc=mt-mc,sdc=sdt+sdc,nc=nt+nc))
hist(post_diff, main='simulated difference')
```



```
quantile(post_diff,c(0.025,0.975))
```

```
##      2.5%      97.5%
## 0.03531293 0.27131692

## theoretical (two independent t distributions)
md = mt-mc
sdd = sdt+sdc
nd = nt+nc
d_range<-c(md-sdd/sqrt(nd)*qt(0.975,df=nd-1),md+sdd/sqrt(nd)*qt(0.975,df=nd-1))
print(d_range)

## [1] 0.04381525 0.27618475
```

## Question 4 (independent binomial)

so using the multinomial is not correct because we do not have k outcomes, instead we have 2 independent binomial processes (as the question states!) so we modeled the independent likelihoods, which is a product of independent beta distributions.

$$p(p_0, p_1) \propto p(x|p_0, p_1)p(p_0, p_1) = p_0^{38.5}(1-p_0)^{634.5}p_1^{21.5}(1-p_1)^{657.5}$$

```

p0= seq(0.01,0.99,by=0.01)
p1= seq(0.01,0.99,by=0.01)

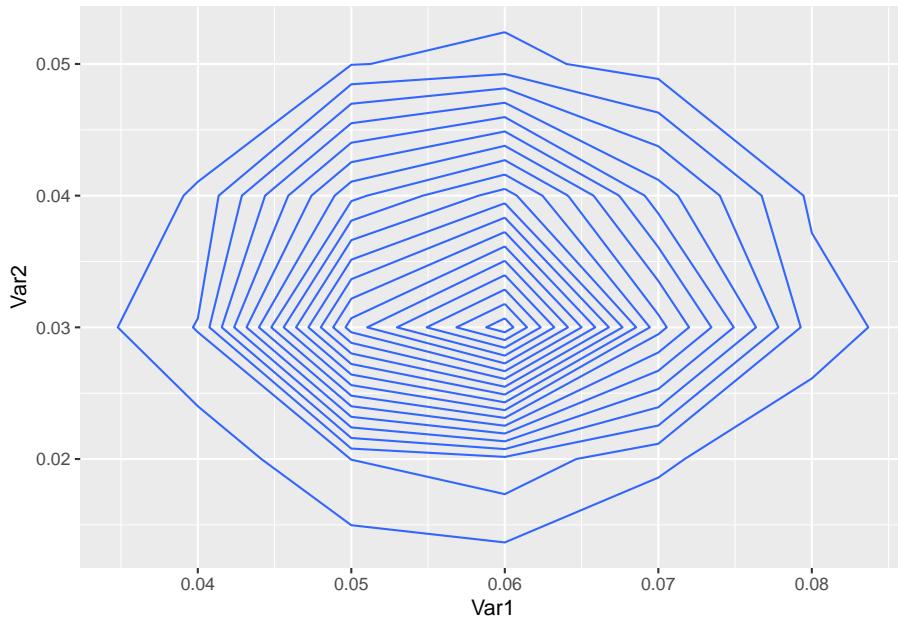
post.bin<-function(p0,p1){
  pos<- p0^(38.5)*(1-p0)^(634.5)*p1^(21.5)*(1-p1)^(657.5)
  return(pos)
}
posts<-NULL

p0p1<- expand.grid(p0, p1)

for(i in 1:nrow(p0p1)){
  posts[i]<-post.bin(p0p1[i,"Var1"],p0p1[i,"Var2"])
}

ggplot(p0p1)+
  geom_contour(mapping = aes(x = Var1, y = Var2, z = posts), bins = 20)

```



- (b) summarize the odds ratio Note that using a 2x2 table the OR is 0.54 and the independent binomials do approximate well.

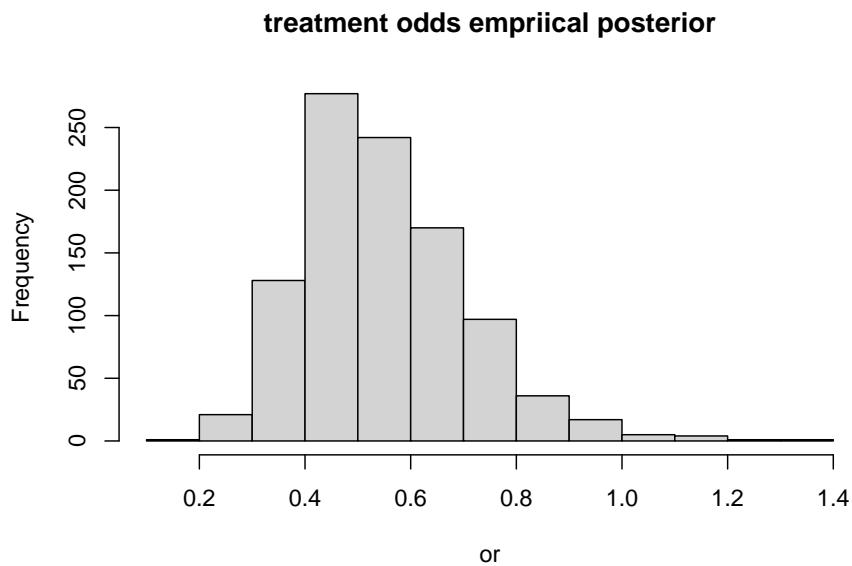
```

b1<-sapply(p0,function(x) pbeta(x,38.5,634.5))
post.control<-rbeta(1000, 38.5,634.5)

```

```
post.treat<-rbeta(1000, 21.5,657.5)

or<-(post.treat/(1-post.treat))/(post.control/(1-post.control))
hist(or,main='treatment odds emprical posterior')
```



```
summary(or)

##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.1861  0.4362  0.5323  0.5514  0.6503  1.3336

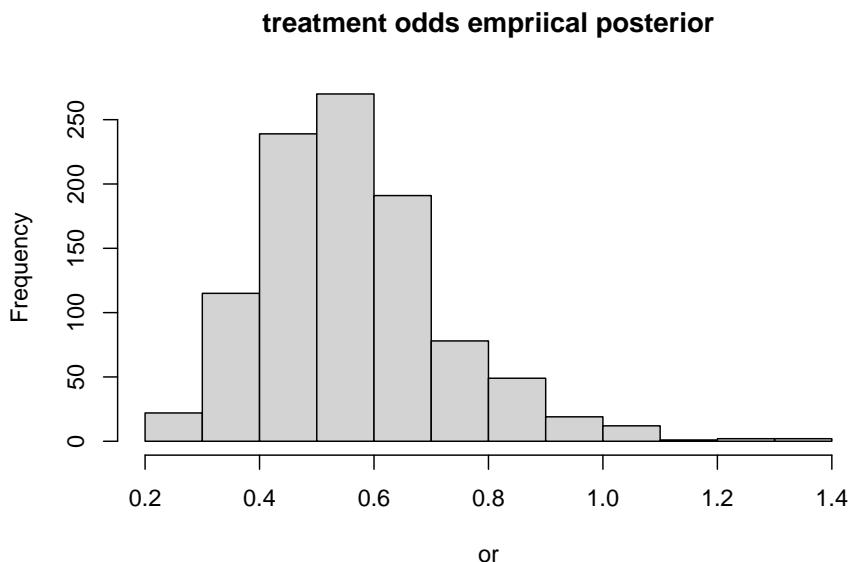
message("the empricial OR: ", round((22*635)/(39*658),2))
```

```
## the empricial OR: 0.54
```

- (c) the sensitivity of the noninformative prior density If we use the prior Beta(1,1) as the uniform prior we do see marginal changes to the posterior mean 0.5639.

```
b1<-sapply(p0,function(x) pbeta(x,39.5,635.5))
post.control<-rbeta(1000, 39.5,635.5)
post.treat<-rbeta(1000, 22.5,658.5)
```

```
or<- (post.treat/(1-post.treat))/(post.control/(1-post.control))
hist(or,main='treatment odds empirical posterior')
```



```
summary(or)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.2564  0.4556  0.5468  0.5654  0.6480  1.3812
```

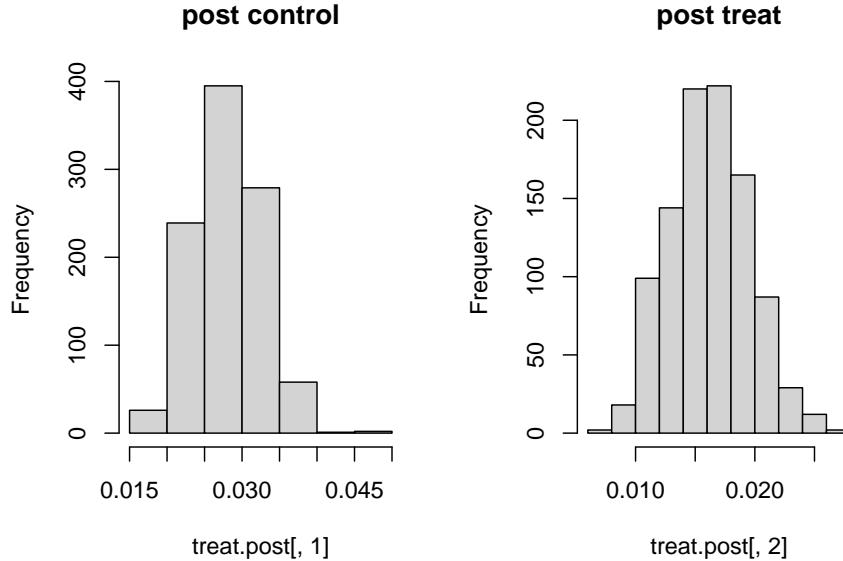
#### Question 4 (dirichlet)

- (4a) set up a noninformative prior on  $(p_0, p_1)$  and obtain posterior simulations. Assume outcomes are independent and binomially distributed. The posterior for probability of death is  $\text{Dir}(39+1, 22+1, 675, 681)$
- the dirichlet is the event of rolling a 4 sided die, whereas this problem is rolling independent 2 sided die, however the dirichlet does approximate independent binomial processes well.

```
nc= 674
dc= 39
nt = 680
dt = 22
```

```
## we have two categories Control ~Bin(p0) and Treat~Bin(p1) we use multinomial
## non-informative prior Diri(a1=1, a2=1)
library(gtools)
## p( p0, p1 ) = Diri( 1,1 ) noninfom prior
# p( p0,p1 | y ) = Diri( 39+1, 22+1 ) ## posterior

treat.post<-rdirichlet(1000, c(39+1,22+1, 674+1,680+1))
par(mfrow=c(1,2))
hist(treat.post[,1],main="post control")
hist(treat.post[,2],main='post treat')
```



- (4b) The posterior odds has a mean of 0.58 using non-informative prior. Using the empirical odds ratio we see the odds of death is 0.546 comparing treatment to controls.

```
## data table
dd<-data.frame(none=c(674-39,680-22),outcome=c(39,22),row.names=c('unexp','exp'))
library(epitools)
oddsratio(as.matrix(dd))$measure

## NA
## odds ratio with 95% C.I. estimate      lower      upper
```

```

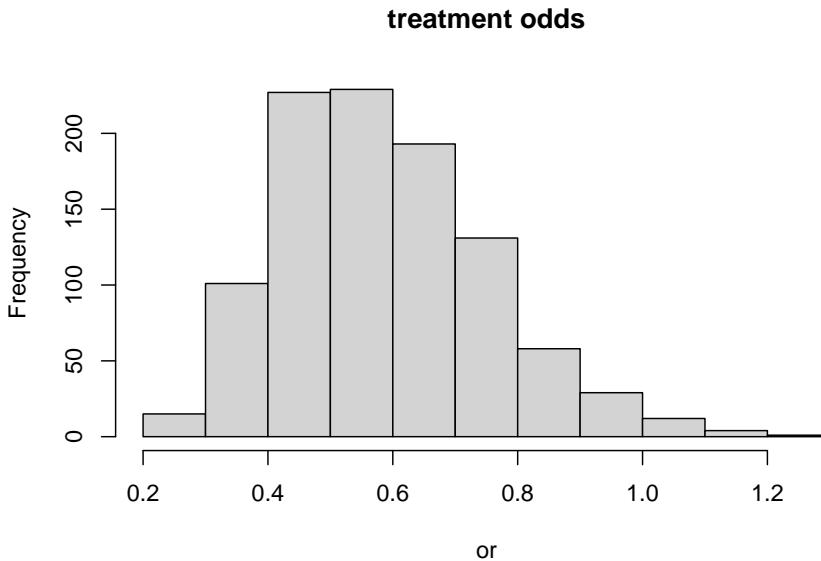
##          unexp 1.0000000      NA      NA
##          exp   0.5463245 0.3147296 0.9249902

## b
or<-treat.post[,2]/(1-treat.post[,2])/treat.post[,1]/(1-treat.post[,1])
summary(or)

##    Min. 1st Qu. Median    Mean 3rd Qu.    Max.
## 0.2448  0.4614  0.5643  0.5856  0.6845  1.2522

hist(or,main='treatment odds')

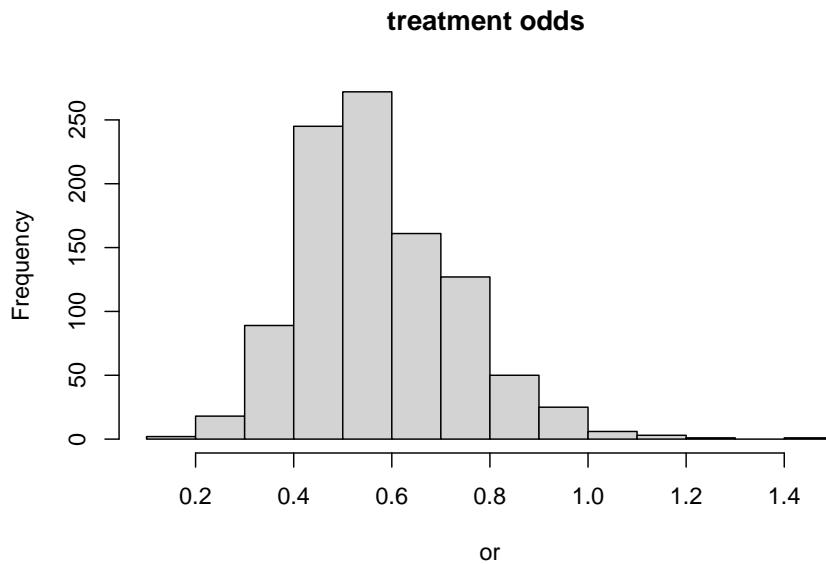
```



- (4c) showing the improper prior ( $a=0$ ), the posterior odds is 0.5793 which is very robust and similar to the noninformative. Using the empirical prior that sets the prior hyperparameters to the cohort expected number of deaths  $(39/674)*Total = 79$ , and  $(22/658)*Total = 45$  and a prior cohort size of  $677 = (674+680)/2$ , yields a similar posterior mean of approximately 0.57 for the odds ratio. Hence the posterior is robust to the choice of prior but does not match well with the empirical non-parametric odds.

```
treat.post<-rdirichlet(1000, c(39,22,674,680))
```

```
## b
or<-treat.post[,2]/(1-treat.post[,2])/((treat.post[,1]/(1-treat.post[,1]))
hist(or,main='treatment odds ')
```

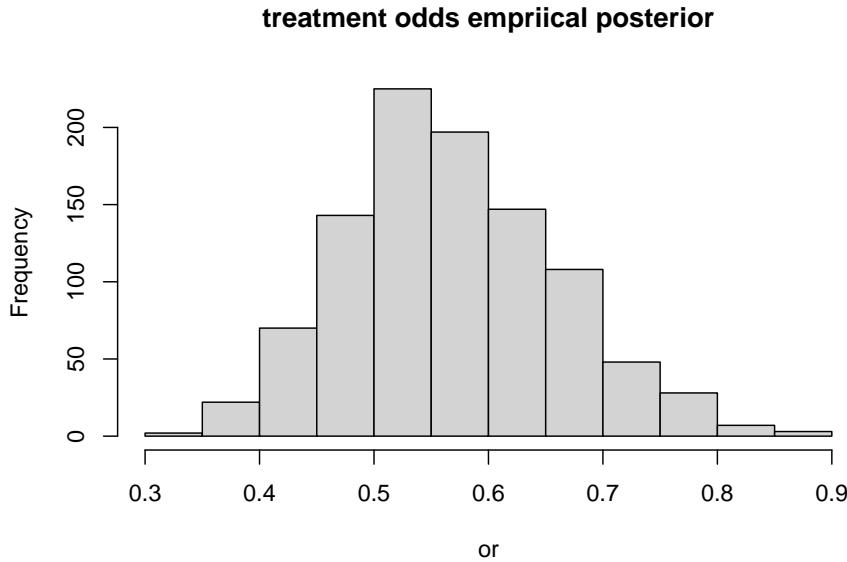


```
summary(or)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.1643  0.4595  0.5532  0.5752  0.6759  1.4929
```

```
## empirical
# control = (39/674)*1354 = 78.34718
# treat = (22/658)*1354 = 45.27052
treat.post<-rdirichlet(1000, c(39+79, 22+46, 674+677, 680+677))

## b
or<-treat.post[,2]/(1-treat.post[,2])/((treat.post[,1]/(1-treat.post[,1]))
hist(or,main='treatment odds empirical posterior')
```



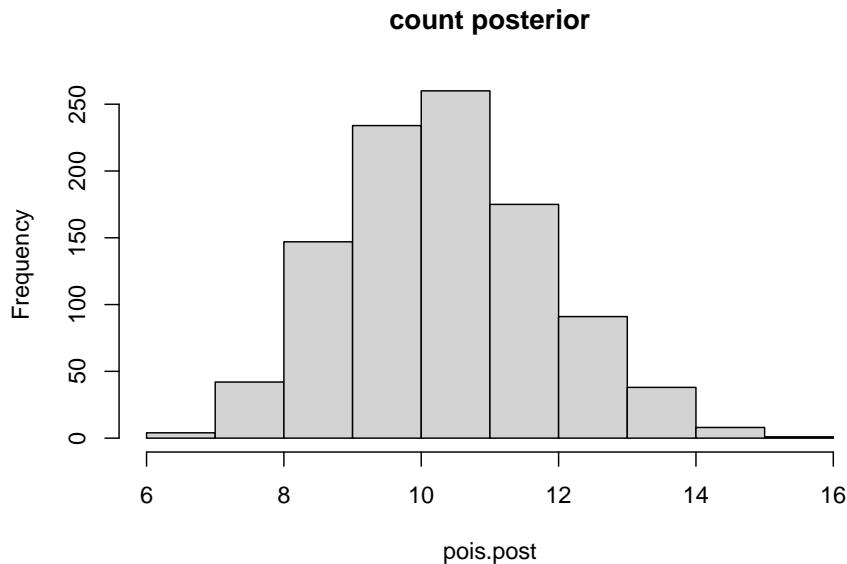
```
summary(or)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##  0.3176  0.5037  0.5570  0.5681  0.6301  0.8848
```

## Question 5

- (a) Give the posterior distribution of  $\mu, \sigma^2$  obtained by pretending the observations are exact, unrounded, measurements. we choose a prior for  $\theta \sim Gamma(0.1, 0.01)$  which has a mean and variance of 10.4 and 1000. Using a Poisson likelihood, the mean is 10.4 and variance 2.2 from the posterior.

```
obs=c(10,10,12,11,9)
### under the exact model, we can not use a binomial because we do not have a proportion (p0) the
# we can use the Poisson distribution
## not sure how to use jeffreys prior
a=0.1
b=0.01
pois.post<-rgamma(1000,a+5*mean(obs),b+5)
hist(pois.post,main='count posterior')
```



```
# post summary mean : 8.8, var= 1.47
summary(pois.post)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##    6.310   9.251  10.248  10.330  11.279  15.249
```

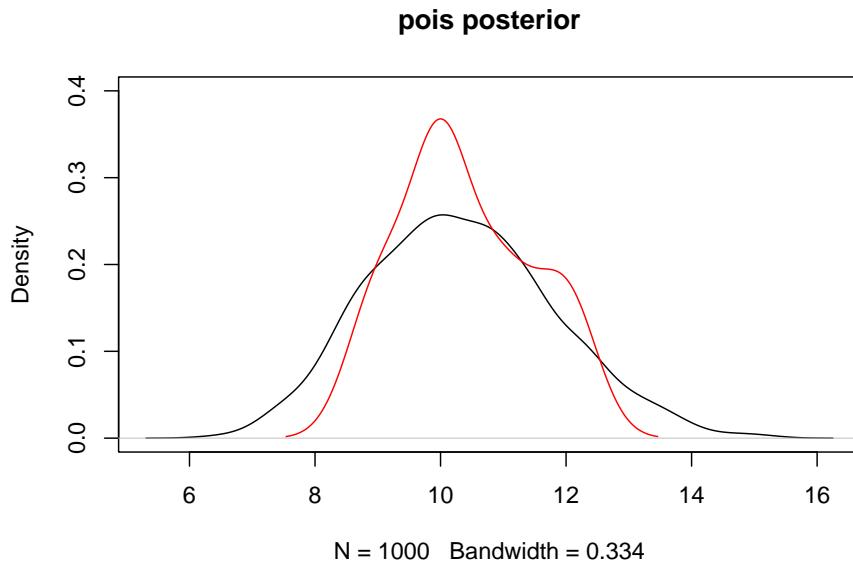
```
mean(pois.post)
```

```
## [1] 10.33018
```

```
var(pois.post)
```

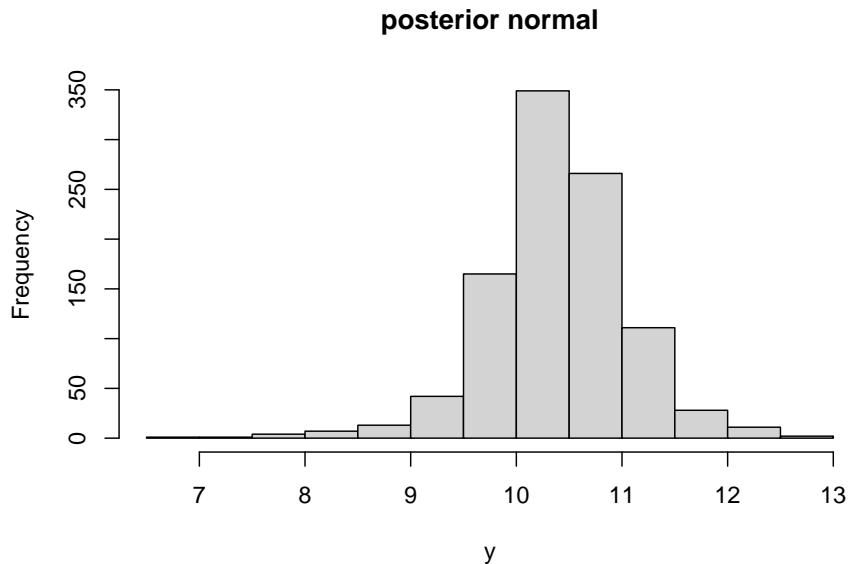
```
## [1] 2.18227
```

```
plot(density(pois.post), ylim=c(0,0.4), main='pois posterior')
lines(density(obs), col='red')
```

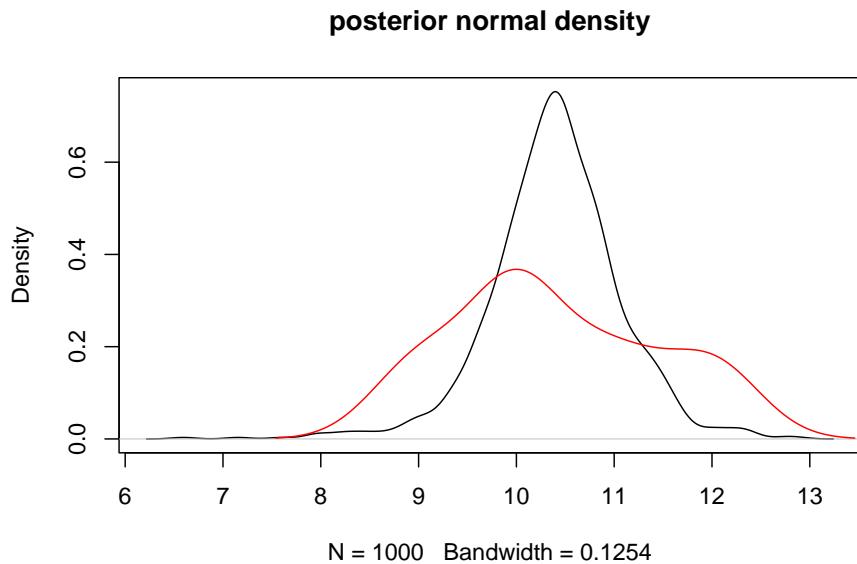


- (b) Give the correct posterior distribution treating the measurements as rounded. Using non informative prior for both unknowns, the posterior mean is 10.4 and posterior variance is 0.433

```
## normal with 2 unknowns.
## joint poisterior
## need the sigma2 | y ~ inv-x2(n-1,s2)
nu = length(obs)-1
s2 = var(obs)
chi2= rchisq(1000, nu)
## inv-chi2 v*s^2/X (Appendix A)
sigma2 = nu*s2/chi2
## posterior prob = p(m | sigma2, y)p(sigma2|y)
## for each variance term draw N( ybar, sd= sqrt(sigma2/n))
y<-sapply(sigma2,function(x) rnorm(1, mean=mean(obs), sd=sqrt(x)/sqrt(5)))
hist(y,main='posterior normal')
```



```
plot(density(y), main='posterior normal density')
lines(density(obs), col='red')
```



- (c) how do the correct/incorrect differ, compare means/variances

The difference in means is approximately 0, with a variance (residual) of 2.7.

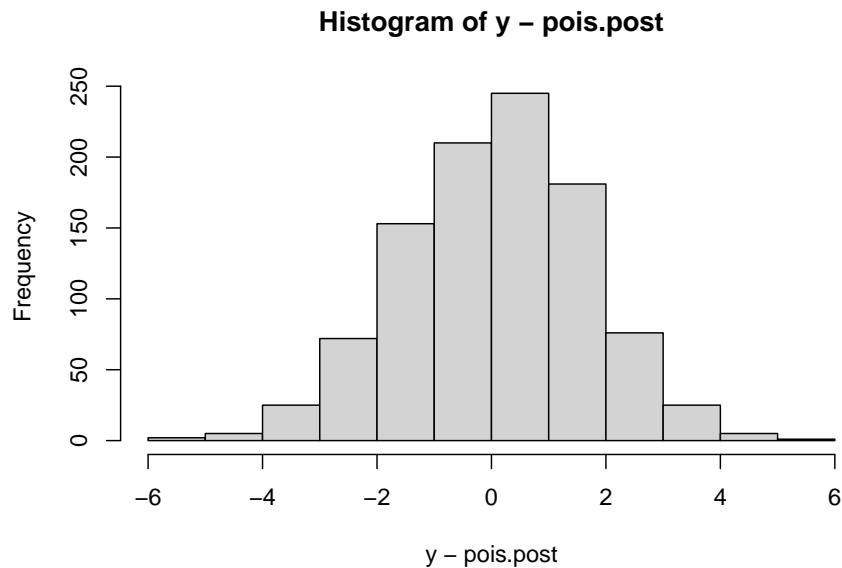
```
mean( y - pois.post) # 0.002926538
```

```
## [1] 0.06352079
```

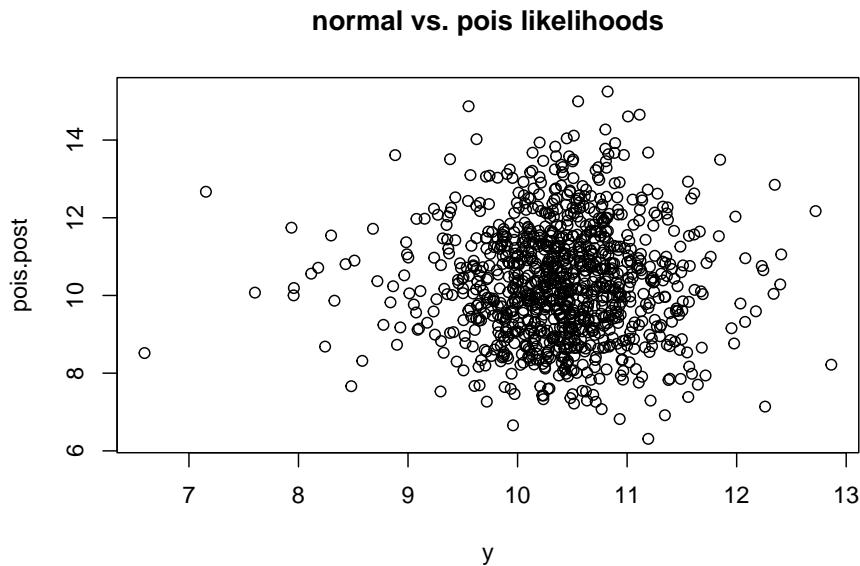
```
var( y - pois.post) # 2.7
```

```
## [1] 2.63661
```

```
hist(y-pois.post)
```



```
plot(y,pois.post,main="normal vs. pois likelihoods")
```



- (d) Let  $z=(z_1, z_2, z_3, z_4, z_5)$  be the original unrounded measurements corresponding to the five observations above. draw simulations of  $z$  and compare the posterior  $(z_1 - z_2)^2$ .

For this we sample from the normal posterior distribution, conditioned on the rounded values to match the observed. For each sample of the posterior, the rounded values are conditioned to match the observed data. The mean difference of  $(z_1 - z_2)^2$  is 0.14, with variance of 0.03.

```
generateObs<-function(obs){
  z<-NULL
  for(i in 1:5){
    zi= sample(y,1)
    while( round(zi)!=obs[i]){
      zi=sample(y,1)
    }
    z<-c(z,zi)
  }
  stopifnot(all(round(z)==obs))
  return(z)
}
```

```

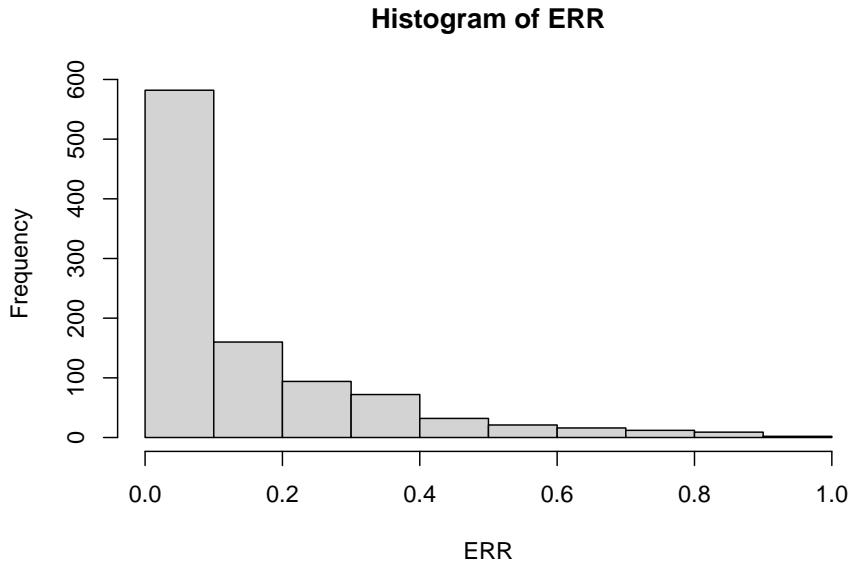
ERR<-NULL

for(j in 1:1000){
  ## generate z samples such that round(z) == obs
  z<-generateObs(obs)
  err<- (z[1]-z[2])^2
  ERR<-c(ERR,err)
}
message("mean error:",mean(ERR))

## mean error:0.141103017760277

hist(ERR)

```



## Question 6{-} Consider data  $y_1, \dots, y_n$  modeled as independent  $\text{Bin}(N, \theta)$  with both  $N$  and  $\theta$  unknown.

First we examine that  $p(\lambda, \theta) \propto 1/\lambda$  described by Raftery (1988). We take the product of a vague prior for  $\lambda$  and the uniform prior for  $\theta$ . let  $\theta \sim U[0, 1]$  be uniform and use the uniform prior for  $\lambda$  as a vague prior. The motivation here is to use a uniform vague prior on the hyperparameter. Alternatively we can use Jeffreys' Prior for  $\text{Pois}(\lambda) = 1/\sqrt{\lambda}$  where the  $-E(\partial l^2 / \partial \lambda^2) = 1/\lambda$  is known for the Poisson.

$$\begin{aligned} p(\lambda, \theta) &\propto (1)1/\lambda \\ &\propto \lambda^{-1} \end{aligned}$$

- (a) to write the prior  $N \sim P(\lambda)$  as a sampling distribution, but we can let  $N$  follow a Gamma, with a  $\text{Gamma}(0,0)$  vague prior initially and independent of  $\theta$ . where  $\theta \sim U(0, 1)$  as a uniform probability. Then  $P(N, \theta) = P(N|\theta)p(\theta)$ . Note this is an improper prior.

$$\begin{aligned} p(N, \theta) &\propto (1)e^{-\beta * N} N^{\alpha-1} \\ &\propto N^{-1}, \text{ for } \alpha, \beta = 0 \end{aligned}$$

We can solve for  $p(N, \theta)$  using the prior  $p(\lambda, \theta) \propto \lambda^{-1}$ . and  $N \sim \text{Pois}(\lambda)$ .

$$\begin{aligned} p(N, \lambda, \theta) &= p(N|\lambda, \theta)p(\lambda, \theta) \rightarrow p(N, \theta) = \int p(N, \lambda, \theta)d\lambda \\ &= \int \frac{(\lambda/N)^N e^{-\lambda/\theta}}{N!} \left(\frac{1}{\lambda}\right) d\lambda \\ &= \frac{1}{N!\theta^N} \int \lambda^{N-1} e^{-\lambda/\theta} d\lambda \\ &= \frac{1}{N!\theta^N} \frac{\Gamma(N)}{1/\theta^N} = 1/N \end{aligned}$$

So we can let  $N$  follow vague prior from the  $\text{Gamma}(0,0)$  distribution, or call  $N$  a uniform on  $(0, N)$  and independent of  $\lambda$  to derive  $p(N, \theta)$ .

- (b) we need to find the marginal posterior of  $N$ , first we find the joint posterior which follows a Beta distribution

$$\begin{aligned} p(\theta, N|y) &= p(y|N, \theta)p(N, \theta) \\ &\propto \theta^{\sum y_i} (1-\theta)^{nN - \sum y_i} (1/N) \\ &\propto (1/N) \text{Beta}(\sum y_i + 1, nN - \sum y_i + 1) \end{aligned}$$

Then we integrate with respect to  $\theta$  to find the marginal posterior, where the integrand is the kernel for the Beta distribution, and we have the beta coefficient that remains in the marginal equation.

$$\begin{aligned} p(N|y) &= \int (1/N) \prod \binom{N}{y_i} \theta^S (1-\theta)^{nN-S} d\theta, S = \sum y_i \\ &= (1/N) \prod \binom{N}{y_i} \int \theta^S (1-\theta)^{nN-S} d\theta \\ &= (1/N) \prod \binom{N}{y_i} \text{Beta}(S+1, nN-S+1) \end{aligned}$$

The maximum probability is 122 trials with the highest marginal,

```

x<-c(53,57,66,67,72)
totalN=seq(max(x),1000)

marginal.post<-function(N,x){
  n=length(x)
  S=sum(x)
  # choose on the log scale
  a<-sum( lchoose(N,x))
  b = (-1)*log(N)
  # b=0
  c = lbeta( S+1, n*N- S +1)

  #rez<-(a+b)/c
  rez<-a+b+c
  return(exp(rez))
} ## problems here

## can we condition we can derive the analytic form of the joint posterior.
## 
marg.n<-sapply(totalN,function(y) marginal.post(y,x))
## we use a grid approach to theta
totalN[which.max(exp(marg.n))]

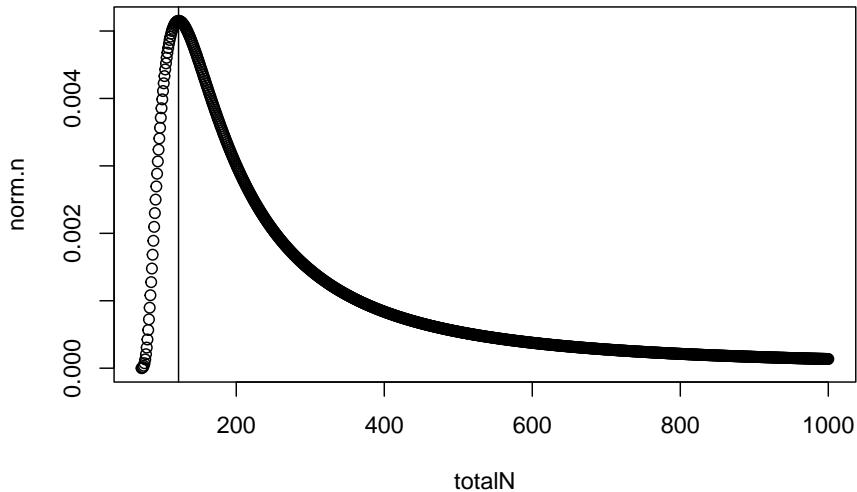
## [1] 122

```

```

## 
norm.n<-marg.n/sum(marg.n)
plot(totalN,norm.n)
abline(v=totalN[which.max(exp(marg.n))])

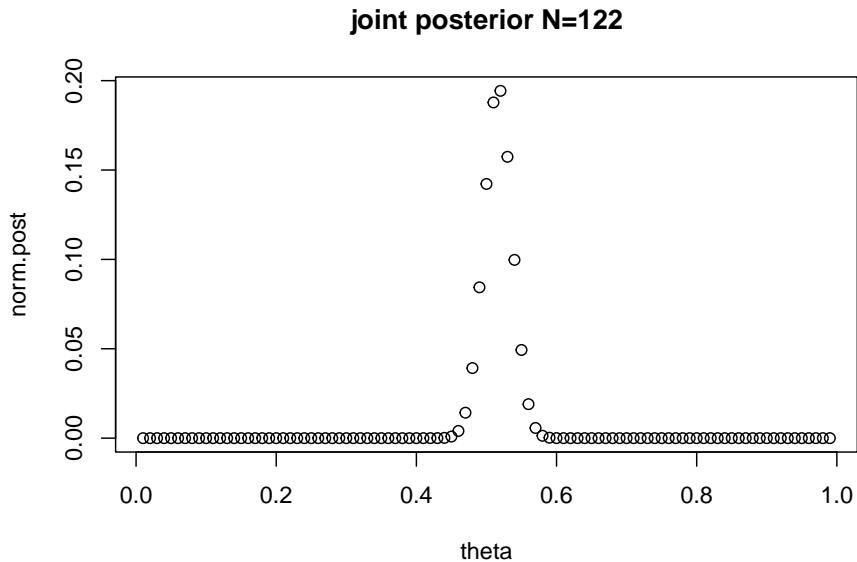
```



The joint posterior distribution given the data and  $N=122$  (which was maximal marginal) has the maximum  $\theta = 0.52$ .

```
theta=seq(0.01,0.99,by=0.01)

posterior.joint<-dbeta(theta, sum(x)+1, 5*122-sum(x)+1)/122
norm.post<-posterior.joint/sum(posterior.joint)
plot(theta,norm.post, main='joint posterior N=122')
```



```
theta[which.max(norm.post)]
```

```
## [1] 0.52
```

- (b ii) and for  $P(N > 100) \approx 0.95$ . We had to use the normalized marginal posterior distribution to correctly compute the probability.

```
1-sum(norm.n[totalN<=100])
```

```
## [1] 0.9518952
```

- (d) the poisson distribution using  $\mu$  for  $N$  only gives the probability for  $N$  as a fixed realization, however we want to study the uncertainty over all possible  $N$  values.

## Question 7

for 2 independent poisson distributions for  $x$  and  $y$  with  $\lambda_1, \lambda_2$ . And the outcome  $X$  has a binomial distribution with  $N=x+y$  and unknown  $p$ . show that the two models have the same likelihood.

We need to solve for the conditional probability of X given X+Y using the Poisson likelihoods. Here if you input the Poisson likelihoods, and use  $X+Y \sim Poi(\lambda_1 + \lambda_2)$  this returns to a  $Binomial(n, \lambda_1/(\lambda_1 + \lambda_2))$ .

$$\begin{aligned} p(X = k | X + Y = n) &= \frac{P(X + Y = n | X = k)P(X = k)}{P(X + Y = n)} \\ &= \frac{P(Y = n - k)P(X = k)}{P(X + Y = n)} \end{aligned}$$

## question 8

notes: dirichlet distribution? - (a) For a given block there are only two outcomes, bicycles, and other vehicles. So a binomial distribution is appropriate (2 outcomes only). We model  $y$  and binomial  $(N_y, \theta_y)$  independent of  $z \sim Binomial(N_z, \theta_z)$ .

We assume  $\theta_y, \theta_z$  are independent, so the posteriors are independent Beta models.

$$\begin{aligned} p(\theta_y, \theta_z | y, z) &= p(\theta_y | y)p(\theta_z | z) \\ &= Beta(\theta_y | 1/2 + \sum y_i, 1/2 + N_y - \sum y_i) * Beta(\theta_z | 1/2 + \sum z_i, 1/2 + N_z - \sum z_i) \end{aligned}$$

- (b) the prior uses Jeffreys' Beta(1/2,1/2) for each parameter independently.

```
rez.lane.bike<-c(16,9,10,13,19,20,18,17,35,55)
rez.lane.other<-c(58,90,48,57,103,57,86,112,273,64)
NY<-(rez.lane.bike+rez.lane.other)
rez.lane.p<-rez.lane.bike/(rez.lane.bike+rez.lane.other)

rez.nolane.bike<-c(12,1,2,4,9,7,9,8)
rez.nolane.other<-c(113,18,14,44,208,67,29,154)
NZ<-(rez.nolane.bike+rez.nolane.other)

rez.nolane.p<-rez.nolane.bike/(rez.nolane.bike+rez.nolane.other)
```

- (c) we determine the posterior distribution using the beta distribution by sampling independently. Using a grid approach we see the contour map centered around 0.18 and 0.09 which is equivalent to the MLEs.

```
bike.post<-rbeta(1000, (1/2 + sum(rez.lane.bike)), (1/2+sum(rez.lane.other)))
bike.post2<-rbeta(1000, (1/2 + sum(rez.nolane.bike)), (1/2+sum(rez.nolane.other)))
```

```

p0= seq(0.01,0.99,by=0.01)
p1= seq(0.01,0.99,by=0.01)

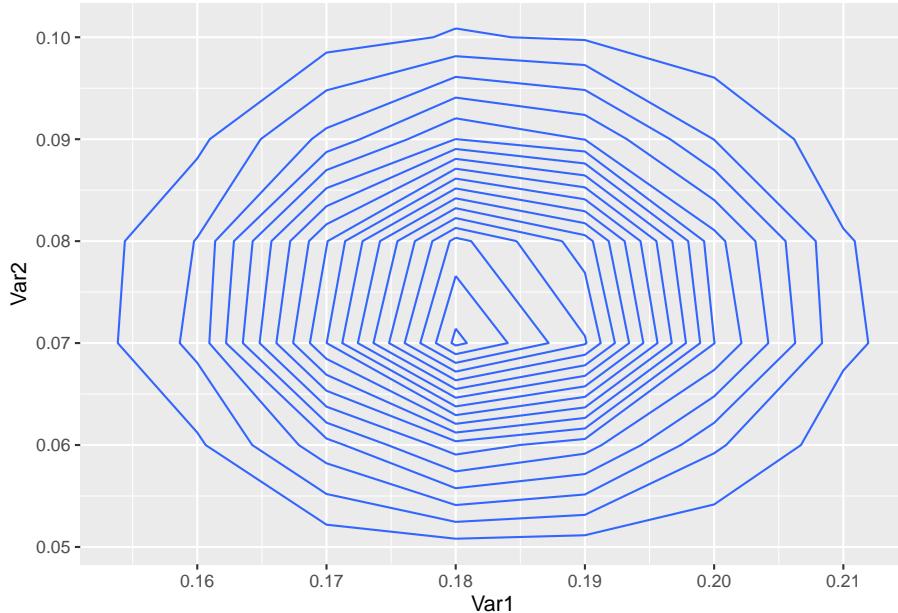
post.bike<-function(p0,p1,y,z,ny,nz){
  pos<- p0^(sum(y))*(1-p0)^(sum(ny)-sum(y))*p1^(sum(z))*(1-p1)^(sum(nz)-sum(z))*dbeta(p0,1/2,1/2)
  return(pos)
}
posts<-NULL

p0p1<- expand.grid(p0, p1)

for(i in 1:nrow(p0p1)){
  posts[i]<-post.bike(p0p1[i,"Var1"],p0p1[i,"Var2"], rez.lane.bike,rez.nolane.bike, NY,NZ)
}

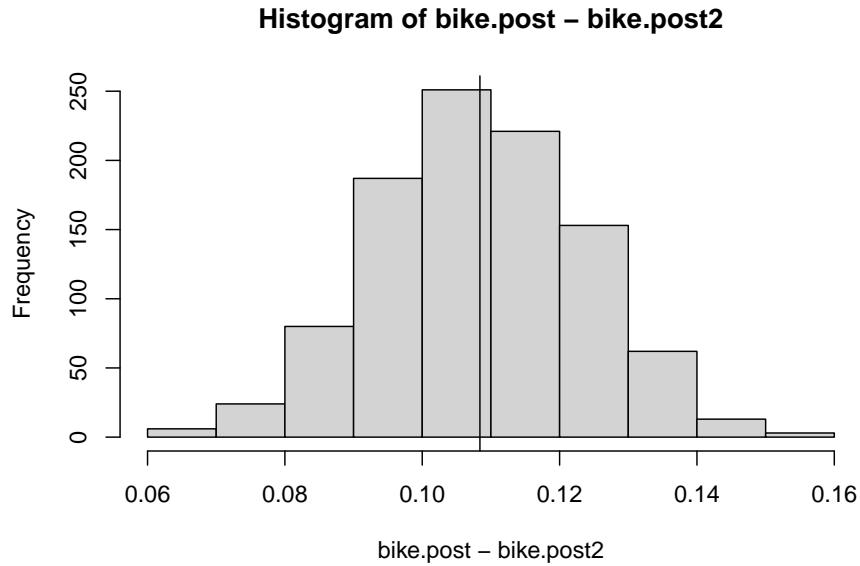
library(ggplot2)
ggplot(p0p1)+
  geom_contour(mapping = aes(x = Var1, y = Var2, z = posts), bins = 20)

```



- (d) expected difference and the posterior distribution we find the difference in the posterior distribution which is approximately 0.11 higher proportions in bike routes.

```
hist(bike.post-bike.post2)
abline(v=mean(bike.post)-mean(bike.post2))
```



```
meany =(1/2 + sum(rez.lane.bike))/(1/2 + sum(rez.lane.bike)+(1/2+sum(rez.lane.other)))
meanz =(1/2 + sum(rez.nolane.bike))/(1/2 + sum(rez.nolane.bike)+(1/2+sum(rez.nolane.othe
```

## question 9

We need to derive the posterior of  $p(\mu^2, \sigma^2|y)$ .  $p(\mu, \sigma^2|y) = p(y|\mu, \sigma^2) * p(\mu, \sigma^2)$

$$\begin{aligned}
 p(\mu, \sigma^2|y) &= \sigma^{-1}(\sigma^2)^{-(\nu_0/2+1)} \exp(-1/2\sigma^2(\nu_0\sigma_0^2 + k_0(\mu - \mu_0)^2)) \times (\sigma^2)^{-n/2} \exp(-1/2\sigma^2((n-1)s^2 + n(\mu - \bar{y})^2)) \\
 &= \sigma^{-1}(\sigma^2)^{-(\nu_0+n)/2+1} \exp(-1/2\sigma^2(\nu_0\sigma_0^2 + (n-1)s^2 + \mu^2(k_0 + n) + \mu(-2\mu_0k_0 - 2\bar{y}n) + \mu_0^2k_0))
 \end{aligned}$$

Where we factor the inside using  $\mu_n = k_0\mu_0/(k_0 + n) + n\bar{y}/(k_0 + n)$

$$\begin{aligned}
 & \rightarrow \mu^2(k_0 + n) + \mu(-2\mu_0 k_0 - 2\bar{y}n) + \mu_0^2 k_0 + n\bar{y}^2 \\
 &= (k_0 + n)(\mu^2 - 2\mu(\mu_0 k_0/(k_0 + n) + \bar{y}n/(k_0 + n)) + \mu_0^2 k_0/(k_0 + n) + n\bar{y}^2/(k_0 + n)) \\
 &= (k_0 + n)(\mu^2 - 2\mu(k_0\mu_0/(k_0 + n) + n\bar{y}/(k_0 + n)) + \frac{\mu_0^2 k_0(k_0 + n)}{(k_0 + n)^2} + \frac{n\bar{y}^2(k_0 + n)}{(k_0 + n)^2} + \frac{2k_0\mu_0 n\bar{y}}{(k_0 + n)^2} - \frac{2k_0\mu_0 n\bar{y}}{(k_0 + n)^2}) \\
 &= (k_0 + n)(\mu^2 - 2\mu(k_0\mu_0/(k_0 + n) + n\bar{y}/(k_0 + n)) + \frac{\mu_0^2 k_0^2}{(k_0 + n)^2} + \frac{\mu_0^2 k_0 * n}{(k_0 + n)^2} + \frac{n^2 \bar{y}^2}{(k_0 + n)^2} + \frac{n\bar{y}^2 k_0}{(k_0 + n)^2} + \frac{2k_0\mu_0 n\bar{y}}{(k_0 + n)^2} - \\
 &\quad \frac{2k_0\mu_0 n\bar{y}}{(k_0 + n)^2}) \\
 &= (k_0 + n)(\mu^2 + -2\mu\mu_n + \mu_n^2 + \frac{\mu_0^2 k_0 * n}{(k_0 + n)^2} + \frac{n\bar{y}^2 k_0}{(k_0 + n)^2} - \frac{2k_0\mu_0 n\bar{y}}{(k_0 + n)^2}) \\
 &= (k_0 + n)((\mu - \mu_n)^2 + (k_0 * n)/(k_0 + n)^2(\mu_0^2 + \bar{y}^2 - 2\mu_0\bar{y})) \\
 &= (k_0 + n)((\mu - \mu_n)^2 + (k_0 * n)/(k_0 + n)(\mu_0 - \bar{y})^2)
 \end{aligned}$$

putting the factorization into the posterior and derived the formula for each updated term.

here we see that  $\sigma_n^2, \mu_n$  are sufficient statistics for  $\mu, \sigma^2$  by the factorization theorem because the posterior depends on parameters only through the summary statistics.

$$\begin{aligned}
 p(\mu, \sigma^2 | y) &= \sigma^{-1}(\sigma^2)^{-((\nu_0+n)/2+1)} \exp(-1/2\sigma^2 + (\nu_0\sigma_0^2 + (n-1)s^2 + (k_0+n)((\mu - \mu_n)^2 + (k_0 * n)/(k_0 + n)(\mu_0 - \bar{y})^2)) \\
 &= \sigma^{-1}(\sigma^2)^{-((\nu_0+n)/2+1)} \exp(-1/2\sigma^2 + (\nu_n\sigma_n^2 + (k_n)((\mu - \mu_n)^2)) \\
 &= \sigma^{-1}(\sigma^2)^{-((\nu_n)/2+1)} \exp(-1/2\sigma^2(\nu_n\sigma_n^2 + (k_n)((\mu - \mu_n)^2))
 \end{aligned}$$

# Question 10{-} To show the F-distribution we prove it directly, using a change of variables. Although there is a theorem (definition) that for  $X \sim N(a, b)$  and  $Y \sim N(c, d)$  then  $F = S_X^2/\sigma_X^2 / (S_Y^2/\sigma_Y^2)$  follows an F-distribution.

For  $(n-1)S^2/\sigma^2 \sim X_{n-1}^2$  we let  $U \sim \chi_p^2, V \sim \chi_q^2$ . We let  $Z = (U/p)/(V/q)$  and  $W = V/q$ .

$$J = \begin{bmatrix} \partial u / \partial z & \partial u / \partial w \\ \partial v / \partial z & \partial v / \partial w \end{bmatrix} = \begin{bmatrix} q & 0 \\ zp & pw \end{bmatrix} = qpw$$

For  $V = wq$ , and  $U = zpw$ . Note that after combining terms we have the Gamma kernel with  $\alpha = p + q/2 - 1$  and  $\beta = 1/2(pz + q)$  which is the F-distribution.

$$\begin{aligned}
f(z, w) &= f(u)f(v)|J| \\
&= \frac{1}{\Gamma(p/2)2^{p/2}}(pzw)^{p/2-1}e^{-pzw/2}/\frac{1}{\Gamma(q/2)}2^{q/2}(wq)^{q/2-1}e^{-wq/2}|pwq| \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)2^{p+q/2}\Gamma(q/2)}(p)^{p/2}q^{q/2}w^{(p+q/2-1)}e^{-w/2(pz+q/2)} \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)2^{p+q/2}\Gamma(q/2)}(p)^{p/2}q^{q/2}\left[\frac{\Gamma(p+q/2)}{(1/2(pz+q))^{p+q/2}}\right] \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)\Gamma(q/2)}(p)^{p/2}q^{q/2}\left[\frac{\Gamma(p+q/2)}{(q^{p+q/2}(pz+q+1))^{p+q/2}}\right] \\
&= \frac{z^{p/2-1}\Gamma(p+q/2)}{\Gamma(p/2)\Gamma(q/2)}(p/q)^{p/2}\frac{1}{(pz/q+1)^{p+q/2}}
\end{aligned}$$

We showed that for U, V as Chi-square variables, their ratio is the F-distribution, In order to show that the posterior distribution  $F = S_X^2/\sigma_X^2/(S_Y^2/\sigma_Y^2)$  is F. The posterior  $Y \sim \sigma^2 | y \sim Inv - \chi^2(\nu, \sigma^2)$ . where  $Y = \frac{(n-1)S^2}{\chi_{n-1}^2}$ . So The inverse chi-square distribution can be re-arranged to follow a Chi-square. Then the variable Z in terms of Chi-square variables can be expressed using Inv-Chisquare variables.

## Question 11

the model is defined as  $p(\alpha, \beta, y, n, x) \propto p(\alpha, \beta) \prod_{i=1}^4 p(y_i | \alpha, \beta, n_i, x_i)$ , where  $(\alpha, \beta)$  follows a joint normal prior with  $\alpha \sim N(0, 2^2)$  and  $\beta \sim N(10, 10^2)$  and  $\text{corr}(\alpha, \beta) = 0.5$ .

- (a) we repeat the computations of this data
- (b) and overlay the contour with the posterior distribution

```

library(mvtnorm)
library(magrittr)
library(dplyr)
library(ggplot2)
assay<-data.frame(x=c(-0.86,-0.30,-0.05,0.73), n=c(5,5,5,5), y=c(0,1,3,5))

## the point a,b by MLE is (0.8,7.7) so we grid around these solution.
a0= seq(-5,10,by=0.1)
b1= seq(-10,40,by=0.1)
a0b1<- expand.grid(a0, b1)

```

```

logit<-function(theta){
  return( log(theta/(1-theta)))
}

inv.logit<-function(theta){
  return( (exp(theta)/(1+exp(theta)))  )
}

## alpha, beta come from multivariate normal.
## \prod_{i=1}^5 p(y_i|\theta) ~ \theta^{y_i} (1-\theta)^{n-y_i} ## kth dose.
### logit(theta) = alpha + beta*x
logit.likelihood<-function(alpha,beta,x,y,n){
  theta_approx<- alpha+beta*x
  bin.like<-(inv.logit(theta_approx))^y*(1-inv.logit(theta_approx))^(n-y)
  return(bin.like)
}

## the alpha,beta stem from the grid, and we return the probability mass from the prior.
joint.prior<-function(alpha,beta){
  cormat<-matrix(c(4,0.5*2*10,0.5*2*10,100),2)
  return(dmvnorm(c(alpha,beta), mean=c(0,10),sigma=cormat))
}

posterior_density<-function(alpha,beta,assay){
  joint_density<-joint.prior(alpha,beta)
  probs<-NULL
  for(k in 1:nrow(assay)){
    p<-logit.likelihood(alpha,beta,assay$x[k],assay$y[k],assay$n[k])
    probs<-c(probs,p)
  }
  totalLikelihood<-prod(probs)
  return( joint_density*totalLikelihood)
}

posts<-NULL

## unnormalized posterior over grid which covers (a,b)
for(i in 1:nrow(a0b1)){
  posts[i]<-posterior_density(a0b1[i,"Var1"],a0b1[i,"Var2"],assay)
}

posts_norm<-posts/sum(posts)
a0b1$joint.prob<-posts_norm

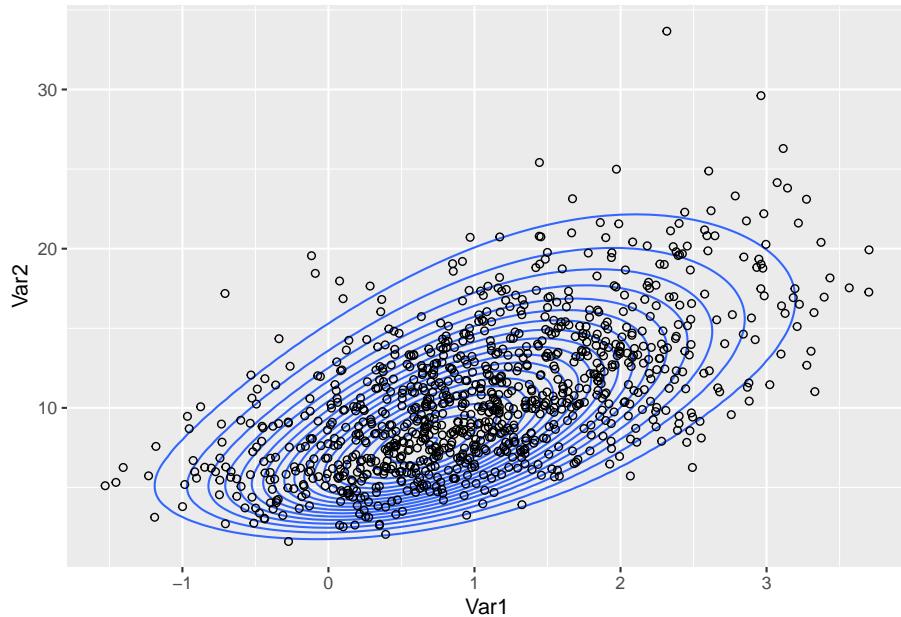
```

```

## grid sampling procedure
marginal.post_a<- a0b1%>%group_by(Var1)%>%summarise(p=sum(joint.prob))%>%data.frame
A<-B<-NULL
for(s in 1:1000){
  a_s<-sample(marginal.post_a$Var1,1,prob=marginal.post_a$p)
  p.a_s<-marginal.post_a[which(marginal.post_a$Var1==a_s), 'p']
  marginal.post_b<-a0b1[which(a0b1$Var1==a_s),]
  marginal.post_b$cond.prob<-marginal.post_b$joint.prob/p.a_s
  b_s<-sample(marginal.post_b$Var2,1,prob=marginal.post_b$cond.prob)
  A<-c(A,a_s)
  B<-c(B,b_s)
}

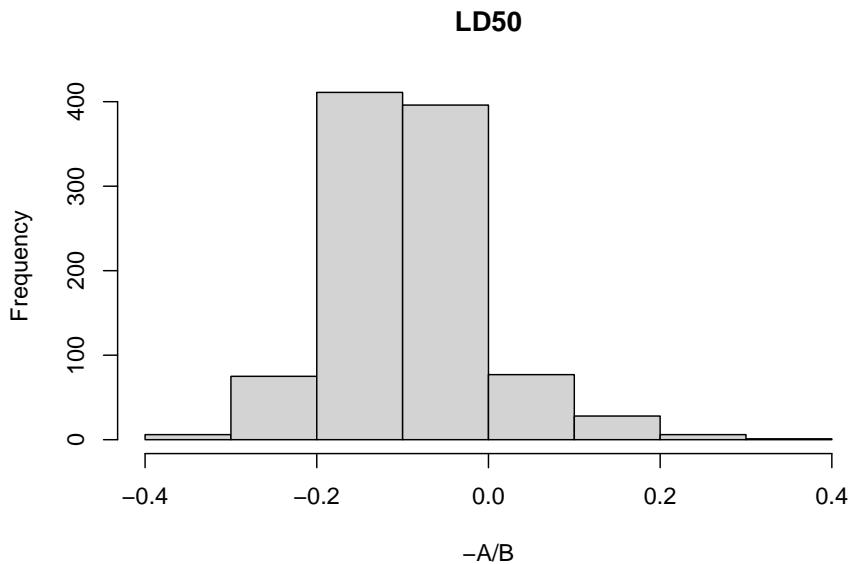
#### need to add a random jitter around each point.
A<-A+runif(length(A),min=0,max=0.1)
B<-B+runif(length(B),min=0,max=0.1)
ab.post<-data.frame(A=A,B=B)
ggplot(a0b1)+
  geom_contour(mapping = aes(x = Var1, y = Var2, z = posts), bins = 20)+
  geom_point(data = ab.post, aes(x = A, y = B), pch = 21)

```



- (bii) the LD50 is summarized ensuring the  $B>0$  for a positive dose-response.

```
hist(-A/B, main='LD50')
```



```
## probability of dose-response being responsive to the dose. negative means dose-response has  
table(B>0)
```

```
##  
## TRUE  
## 1000
```



## Chapter 4

# Asymptotics and connections to non-Bayesian approaches

### 4.1 Normal approximations to the posterior distribution

If the posterior distribution is unimodal and roughly symmetric, it can be approximated by a normal distribution, such that the logarithm of the posterior density is approximated by a quadratic function via the Taylor series expansion of  $\theta$ .

Consider a quadratic approximation to the log-posterior centered on the posterior mode. where the linear term goes to 0.

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + (1/2)(\theta - \hat{\theta})^T [d^2/d\theta^2 \log p(\theta|y)]_{\theta=\hat{\theta}} + \dots \quad (4.1)$$

The first term is a constant and the second term is proportional to the normal density yielding the approximation and we can expand the posterior density second derivative in terms of the prior and likelihood.

$$[d^2/d\theta^2 \log p(\theta|y)]_{\theta=\hat{\theta}} = [d^2/d\theta^2 \log p(\theta)]_{\theta=\hat{\theta}} + \sum_{i=1}^n [d^2/d\theta^2 \log p(y_i|\theta)]_{\theta=\hat{\theta}} \quad (4.2)$$

Now to show the normal approximation

$$\begin{aligned}
\log p(\theta|y) &= \log p(\hat{\theta}|y) + (1/2)(\theta - \hat{\theta})^T [d^2/d\theta^2 \log p(\theta|y)]_{\theta=\hat{\theta}} + \dots \\
&= \log p(\hat{\theta}|y) + (1/2)(\theta - \hat{\theta})^T (d^2/d\theta^2 \log p(\hat{\theta})) + \sum_{i=1}^n d^2/d\theta^2 \log p(y_i|\theta)_{\theta=\hat{\theta}} \\
\log p(\theta|y) - \log p(\hat{\theta}|y) &= +(1/2)(\theta - \hat{\theta})^T (c - n * J(\theta_0)) \\
p(\theta|y) - p(\hat{\theta}|y) &\propto \exp(-\frac{1}{2(nJ(\theta_0))^{-1}}(\theta - \hat{\theta})^T)
\end{aligned}$$

Which taking the limit as  $|\theta - \hat{\theta}| \rightarrow 0$ , then the posterior converges to 0 written as  $p(\theta|y) - p(\hat{\theta}|y) \rightarrow 0$  as  $n \rightarrow \infty$ , and we have normality.

In discussing large-sample properties, the concept of Fisher Information ,  $J(\theta)$ , in the context of Jeffreys' prior is used.

$$p(\theta|y) \approx N(\hat{\theta}, [I(\hat{\theta})]^{-1}) \quad (4.3)$$

Where  $I(\theta)$  is the *observed information*

$$I(\theta) = -d^2/d\theta^2 \log p(\theta|y) \quad (4.4)$$

Where is the mode  $\hat{\theta}$  is in the interior of the parameter space, then the information is positive definite.

### Summarizing posterior distributions by point estimates and standard errors

From the asymptotic theory, if  $n$  is large enough, a posterior distribution is approximated by the normal distribution. A standard inferential summary is the 95% interval obtained by computing a point estimate  $\hat{\theta}$  such as the MLE (which is the posterior under a uniform prior density), plus or minus two standard errors, with the standard error estimated from the information at the estimate  $I(\hat{\theta})$ .

## 4.2 Large-sample theory

The basic tool of Bayesian inference is asymptotic normality of the posterior distribution, as more data arrive from the same underlying process, the posterior distribution of the parameter vector approaches multivariate normality. Suppose the data are modeled by a parametric family,  $p(y|\theta)$ , and a prior  $p(\theta)$ , and suppose that the true distribution is included in the parametric family (i.e. if

$f(y) = p(y|\theta_0)$ ) then the property of asymptotic normality and *consistency* holds.

Consistency is defined as the posterior distribution converges to a point mass at the true parameter,  $\theta_0$  as  $n \rightarrow \infty$ .

Note, that if the true distribution is not included in the parametric family, then there is no longer a true parameter to converge to. One must use the *Kullback-Leibler divergence* to determine the value  $\theta_0$  that makes the model distribution closest to the true distribution.

### Asymptotic normality and consistency

Under regularity conditions (the likelihood is a continuous function of  $\theta$ , and that  $\theta_0$  is not a boundary point), as  $n \rightarrow \infty$ , the posterior distribution of  $\theta$  approaches normality with mean  $\theta_0$  and variance  $(nJ(\theta_0))^{-1}$ . Where  $J(\cdot)$  is the *Fisher information* in context of Jeffreys' Prior.

The posterior mode is consistent for  $\theta_0$ , as  $n \rightarrow \infty$ , so the mass of the posterior  $p(\theta|y)$  becomes concentrated in small neighborhoods of  $\theta_0$  and the distance of  $|\theta - \theta_0| \rightarrow 0$ .

Further, we can write the coefficient of the quadratic term in (4.1).

$$[d^2/d\theta^2 \log p(\theta|y)]_{\theta=\hat{\theta}} = [d^2/d\theta^2 \log p(\theta)]_{\theta=\hat{\theta}} + \sum_{i=1}^n [d^2/d\theta^2 \log p(y_i|\theta)]_{\theta=\hat{\theta}} \quad (4.5)$$

This is considered a function of  $\theta$ , as a constant term plus the sum of  $n$  terms whose expected value under the true sampling distribution  $p(y_i|\theta_0)$ , is approximately  $-J(\theta_0)$ , assuming  $\hat{\theta}$  is close to  $\theta$ .

In summary, as the limit of  $n$ , in the context of a family of models posterior mode,  $\hat{\theta}$ , approaches the truth  $\theta_0$ , and the curvature approaches  $nJ(\hat{\theta})$  or  $nJ(\theta_0)$ . Interesting, as  $n \rightarrow \infty$  the prior term is a constant, and the likelihood dominates the posterior because the likelihood alone is used to obtain the mode and curvature for the normal approximation.

## 4.3 Frequency evaluations of Bayesian inferences

The notion of *stable estimation* which says that for a fixed model, the posterior approaches a point as more data arrive, leading, in the limit, to inferential certainty, is based on the concepts of repeated sampling. It is certainly appealing that if the hypothesized family of probability models contain the true distribution, then as more information about  $\theta$  arrives, the posterior distribution converges to the true value of  $\theta$ .

### Large sample correspondence

Suppose that the normal approximation holds (4.3) for the posterior distribution for  $\theta$ , then we can transform to the *standard* normal multivariate normal

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta})|y \sim N(0, I) \quad (4.6)$$

Where  $\hat{\theta}$  is the posterior mode and  $[I(\hat{\theta})]^{1/2}$  is any matrix square root of the observed fisher information. In addition to  $\hat{\theta} \rightarrow \theta_0$  we can write the approximation using  $I(\theta_0)$ . If the true data distribution is included in the class of models, so that  $f(y) = p(y|\theta)$ , then under *repeated sampling* with fixed  $\theta$ , as  $n \rightarrow \infty$  then

$$[I(\hat{\theta})]^{1/2}(\theta - \hat{\theta})|\theta \sim N(0, I) \quad (4.7)$$

This is generally proven for the MLE, but can be extended for the posterior mode  $\hat{\theta}$ . This results suggest that for any function  $(\theta - \hat{\theta})$  the posterior distribution derived from (4.6) is asymptotically the same as the repeated sampling distribution from (4.7). Thus for a 95% central posterior interval for  $\theta$  will cover the true value 95% of the time under repeated sampling with any fixed true  $\theta$ .

### Point estimation, consistency, and efficiency

For large samples, obtaining an estimate- makes most sense when the posterior mode  $\hat{\theta}$  is the obvious center and the  $nI(\theta_0)$  is small and practically unimportant. However in smaller samples, one can define optimal point estimates, but it is better to show the full representation of the full posterior distribution. In most problems, the point estimate and the standard error are adequate to summarize the posterior inference. We interpret the estimate as an inferential summary, not as a decision solution / classification.

A point estimate is said to be *consistent* as the samples get larger, it converges to the true value parameter. Thus if  $f(y) = p(y|\theta_0)$ , then a point estimate  $\hat{\theta}$  of  $\theta$  is consistent if its sampling distribution converges to a point mass at  $\theta_0$  for  $n \rightarrow \infty$ .

*Asymptotic unbiasedness* is defined as  $(E(\hat{\theta}|\theta_0 - \theta_0))/sd(\hat{\theta}|\theta_0)$  converges to 0 as sample size increases.

*Efficiency* for a point estimate is if there is no other function of  $y$  that estimates  $\theta$  with lower mean squared error, that is if the expression  $E((\hat{\theta} - \theta_0)^2|\theta_0)$  is at its optiomal lowest value. An estimate is asymptotically efficient if its efficiency approaches 1 as the sample size  $n$ , increases to infinity.

### Confidence coverage

If a region  $C(y)$  includes  $\theta_0$  at least  $100(1 - \alpha)\%$  of the time, then  $C(y)$  is called the  $100(1 - \alpha)\%$  *confidence region* for parameter  $\theta$ . We saw previously that asymptotically a  $100(1 - \alpha)\%$  central posterior interval for  $\theta$  has the property that, in repeated samples of  $y$ ,  $100(1 - \alpha)\%$  of the intervals include  $\theta_0$ .

## 4.4 Exercises

### Question 1

- a using simple calculus we found  $l' = \frac{2 * \sum_{i=1}^5 (y_i - \theta)}{1 + (y_i - \theta)^2}$  and  $l'' = \frac{-2 \sum_{i=1}^5 ((y_i - \theta)^2 - 1)}{(1 + (y_i - \theta)^2)^2}$

```
mleCauchy<-function(x,tolerance=0.001){
  startvalue<-median(x)
  n=length(x)
  theta_current<-startvalue
  first_deriv<- 2*sum((x-theta_current)/(1+(x-theta_current)^2))

  while( abs(first_deriv)>tolerance){
    second_deriv<- 2*sum(((x-theta_current)^2-1)/(1+(x-theta_current)^2)^2)

    theta_new<- theta_current - first_deriv/second_deriv
    theta_current<-theta_new
    first_deriv<- 2*sum((x-theta_current)/(1+(x-theta_current)^2))
  }
  return(theta_current)
}
x<-c(-1.94,0.59,-5.98,-0.08,-0.77)
mleCauchy(x,0.0001)
```

```
## [1] -0.5343968
```

- b the MLE for theta is -0.138. Using newton-raphson method for  $\theta^1 = \theta^0 - l'(\theta^0)/l''(\theta^0)$

```
x<-c(-2,-1,0,1.5,2.5)
posterior_mode<-mleCauchy(x,0.0001)
print(posterior_mode)
```

```
## [1] -0.1376488
```

```

## optimize function
optimize(function(theta) -sum(dcauchy(x, location=theta, log=TRUE)), c(-100,100))

## $minimum
## [1] -0.1376593
##
## $objective
## [1] 11.17292

```

-c for the normal approximation we use  $p(\hat{\theta}|y) \approx N(\hat{\theta}, nI(\theta_0)^{-1})$

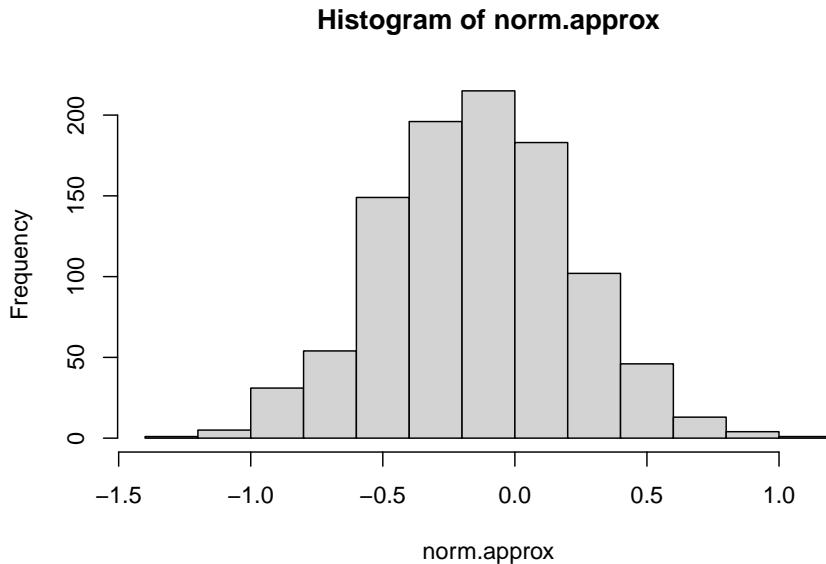
$$\begin{aligned}
I(\theta) &= E(d/d\theta \log p(\theta|y))^2 \\
E[(\frac{2(y-\theta)}{1+(y-\theta)^2})^2] &= \int_{-\infty}^{\infty} (\frac{2(y-\theta)}{1+(y-\theta)^2})^2 \frac{1}{1+(y-\theta)^2} d\theta \\
&= 8 \int_0^{\infty} \frac{u^2}{(1+u^2)^2} du, u = y-\theta, du = -d\theta \\
&\text{2nd substitution } x = \frac{1}{1+u^2} \implies u^2 = \frac{1}{x}-1, du = (1/2)(1/x-1)^{-1/2}(-1/x^2)dx \\
&= -4 \int (1-x)x^2[1/2(\frac{1}{x}-1)^{-1/2}(-1/x^2)dx] \\
&= 2 \int_0^{\infty} (1-x)(1/x-1)^{-1/2} dx \\
&= 2 \int_0^{\infty} x^{1/2}(1-x)^{1/2} \sim Beta(\alpha = 3/2, \beta = 3/2) \\
I(\theta) &= \frac{4\Gamma(3/2)\Gamma(3/2)}{\Gamma(3)} \approx 1.5708
\end{aligned}$$

$$p(\theta|y) \approx N(\hat{\theta}, \sigma^2 = 1/(n * I(\theta)))$$

```

# need to derived the Fisher information for Cauchy
fishers.n<- (4*gamma(3/2)*gamma(3/2)/(gamma(3)))
var.approx<- 1/(length(x)*fishers.n)
norm.approx<- rnorm(1000, mean=posterior_mode, sd=sqrt(var.approx))
hist(norm.approx)

```



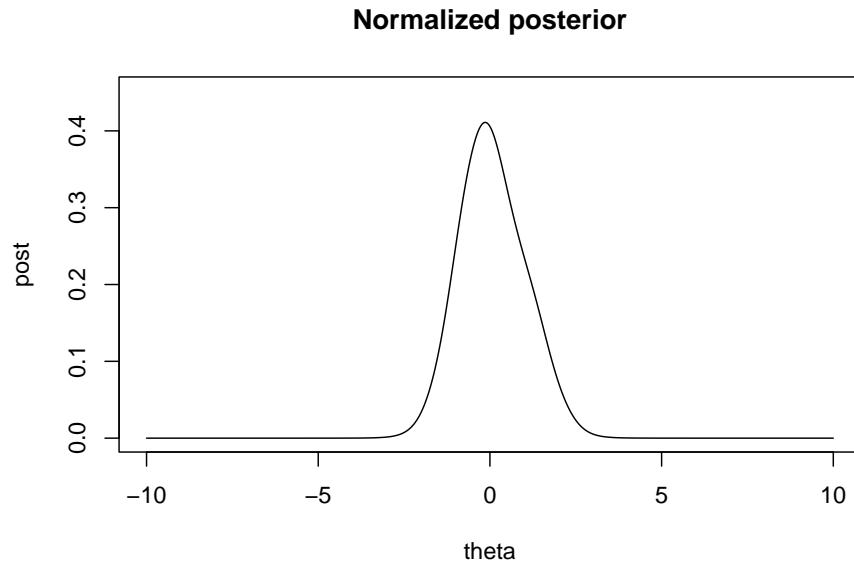
compare with 2.11 grid approach using the Cauchy distribution, we see that normal approximation has much wider tails compared to the exact distribution, this is because of the small sample size.

```

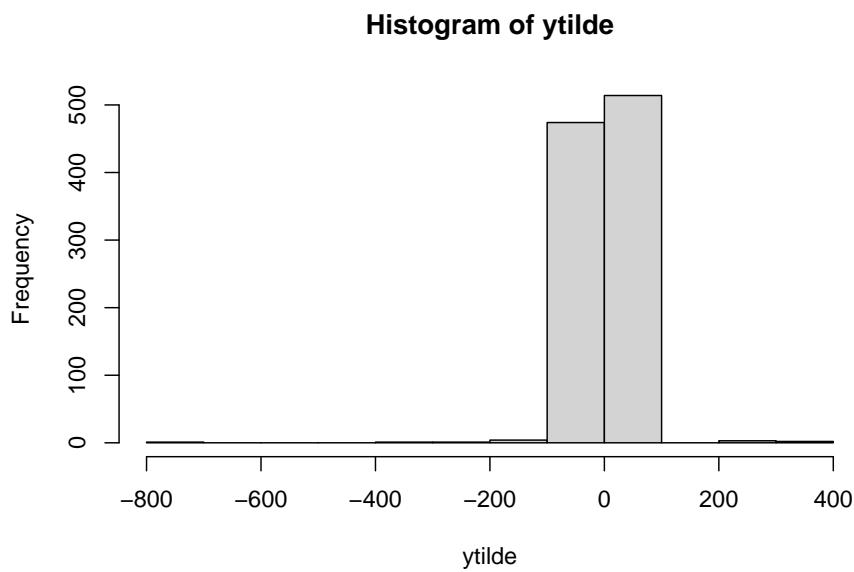
y<-x<-c(-2,-1,0,1.5,2.5)

step=0.01
#theta<-seq(from=0,to=100000)/m
theta<-seq(-10, 10,by=step)
## p(theta | y) ~ p(y|theta)*p(theta)
dens<-function(y,th){
  dens0<-NULL
  for(i in 1:length(th)){
    dens0<-c(dens0, prod (dcauchy(y, th[i],1)))
  }
  dens0
}
#dens(y,theta)
#  $L(\theta | y) = \prod_{i=1}^n f(y_i | \theta)$  we need the product term here.
unnorm.post<-sapply(theta, function(x) prod(dcauchy(y,location=x,scale=1))) ## un norm post
##  $p(\theta | y) = p(y | \theta)p(\theta)$  where  $p(\theta)$  is  $U(0,100)$ 
post<-unnorm.post/(step*sum(unnorm.post))
plot(theta,post,type='l',main='Normalized posterior', ylim=c(0, 1.1*max(post)))

```



```
samps<-sample(theta,1000,prob=post*step,replace=T)
ytilde<-rcauchy(1000,location=samps,scale=1)
hist(ytilde)
```



# Chapter 5

## Blocks

### 5.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad (5.1)$$

You may refer to using `\@ref(eq:binom)`, like see Equation (5.1).

### 5.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 5.1.

**Theorem 5.1.** *For a right triangle, if  $c$  denotes the length of the hypotenuse and  $a$  and  $b$  denote the lengths of the **other** two sides, we have*

$$a^2 + b^2 = c^2$$

Read more here <https://bookdown.org/yihui/bookdown/markdown-extensions-by-bookdown.html>.

### 5.3 Callout blocks

The R Markdown Cookbook provides more help on how to use custom blocks to design your own callouts: <https://bookdown.org/yihui/rmarkdown-cookbook/custom-blocks.html>



# Chapter 6

## Sharing your book

### 6.1 Publishing

HTML books can be published online, see: <https://bookdown.org/yihui/bookdown/publishing.html>

### 6.2 404 pages

By default, users will be directed to a 404 page if they try to access a webpage that cannot be found. If you'd like to customize your 404 page instead of using the default, you may add either a `_404.Rmd` or `_404.md` file to your project root and use code and/or Markdown syntax.

### 6.3 Metadata for sharing

Bookdown HTML books will provide HTML metadata for social sharing on platforms like Twitter, Facebook, and LinkedIn, using information you provide in the `index.Rmd` YAML. To setup, set the `url` for your book and the path to your `cover-image` file. Your book's `title` and `description` are also used.

This `gitbook` uses the same social sharing data across all chapters in your book—all links shared will look the same.

Specify your book's source repository on GitHub using the `edit` key under the configuration options in the `_output.yml` file, which allows users to suggest an edit by linking to a chapter's source file.

Read more about the features of this output format here:

<https://pkgs.rstudio.com/bookdown/reference/gitbook.html>

Or use:

```
?bookdown::gitbook
```

# Chapter 7

## Parts

You can add parts to organize one or more book chapters together. Parts can be inserted at the top of an .Rmd file, before the first-level chapter heading in that same file.

Add a numbered part: `# (PART) Act one {-} (followed by # A chapter)`

Add an unnumbered part: `# (PART\*) Act one {-} (followed by # A chapter)`

Add an appendix as a special kind of un-numbered part: `# (APPENDIX) Other stuff {-} (followed by # A chapter)`. Chapters in an appendix are prepended with letters instead of numbers.



# Bibliography

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL <http://yihui.org/knitr/>. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2022. URL <https://CRAN.R-project.org/package=bookdown>. R package version 0.27.