Bayesian Data Analysis

Anthony R. Colombo, Dr. Paul Marjoram

2022-09-11

Contents

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 oltimate 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 $p(y_1, ..., y_n)$ 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 exchangability 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(θ), 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

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

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

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.2}$$

Note that $p(y|\theta)$ is taken as a function of θ, 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 θ 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 θ.

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 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 uncertainity 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)}{p(y)}= \frac{(1/2)N(1,4)}{(1/1
```

```
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(favorite wins | point spread =8) = 0.67.
   P(favorite wins by at least 8 | point spread =8) = 0.42.
   and P(fav. wins by at least 8 | spread =8, favorite team wins) = 0.62.

```
spread<-8

## outcome of the games favor score - underdog score
 games<-c(-7,-5,-3,-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( fav. wins >8 | fav. wins) = P(fav. wins > 8, fav. wins )/ P(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(wins by 8 | favorite team wins) = P(favorite team wins | wins by 8)P(wins by 8)/P(favorite team wins) = P(wins by 8)/P(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(Wins by 8 | Fav. wins) = P(Fav. wins | wins by 8)P(wins by 8) / P(fav. wins)
    ## P(Fav. wins | 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)))
```

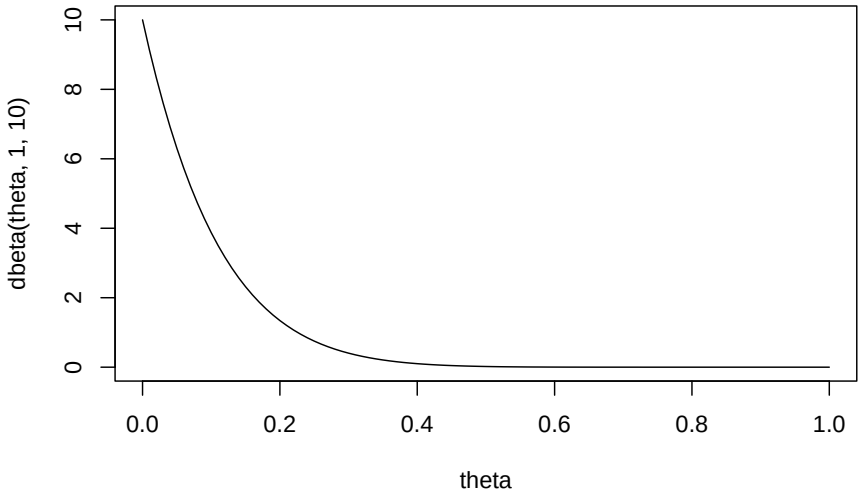
```
## (normal): fav wins by 8 given fav. wins: 0.6
```

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 θ_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 α, β are set to 1 for the uniform prior. For this case we set α, β equal to 1, 10 which has a prior mean of 0.09.

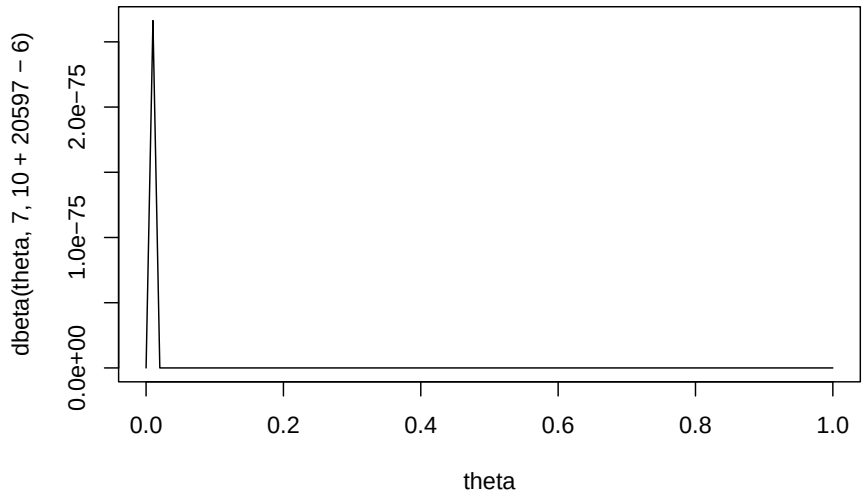
```
theta=seq(from=0,to=1,by=.01)
plot(theta,dbeta(theta,1,10),type='l')
```



(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 θ. 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 θ. 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 patitient 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,
    }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)){
```

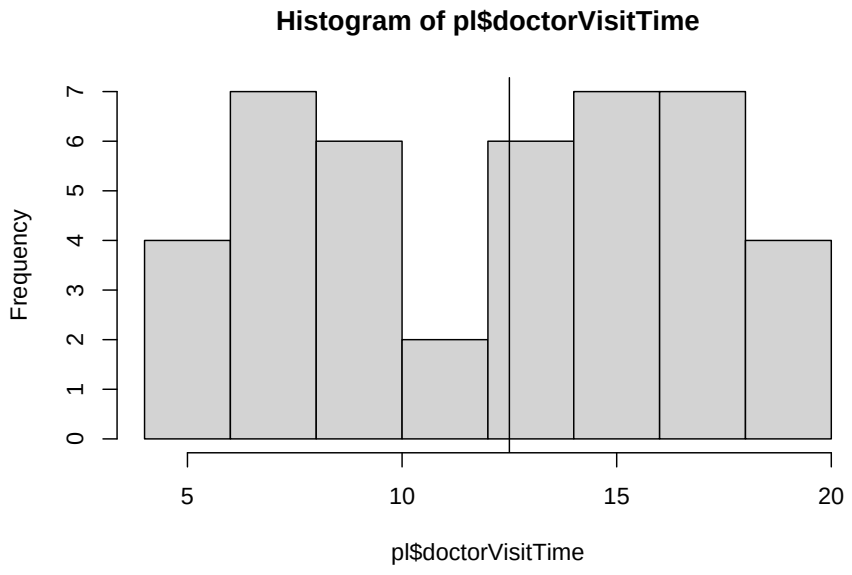
```
    waitTime=0
    assignedDr<-sample(doctors$dr[which(doctors$nextTimeAvail<currentTime)],1)
    ### current time + visitTime
    nextAvailTime<- visitTime+currentTime+waitTime
    ## completion time for patient exit (closing time).
  }else if(any(doctors$nextTimeAvail<currentTime)==FALSE){
    # all doctors are booked, no available doctors.
    # wait time is the difference between next available time (assuming all times are greater th
    waitTime<-min(doctors$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,ass
  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$patient==max(patientList$patie
 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.
```

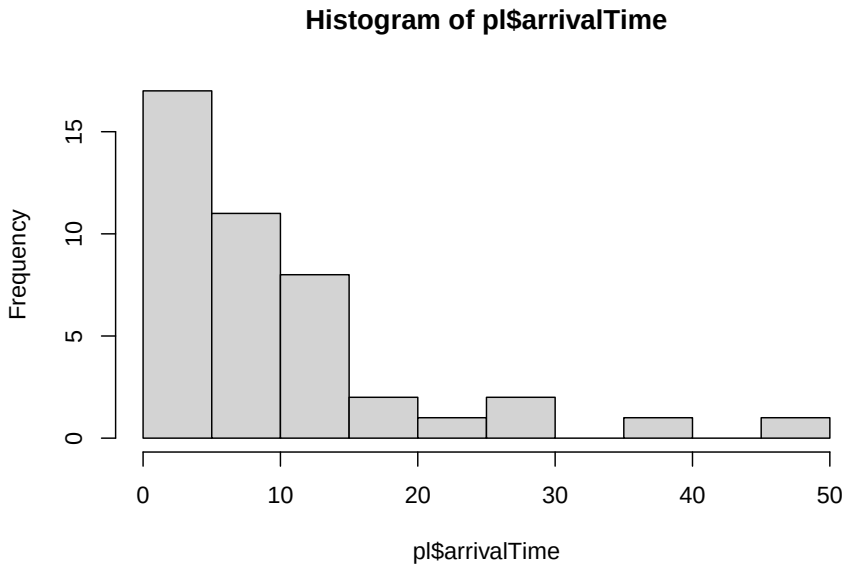
```
    if(any(doctors$nextTimeAvail<p1$currentTime)){
      doctors$avail[which(doctors$nextTimeAvail<p1$currentTime)]<-'yes'
    }

    ## 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[currentPatientN

    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
```

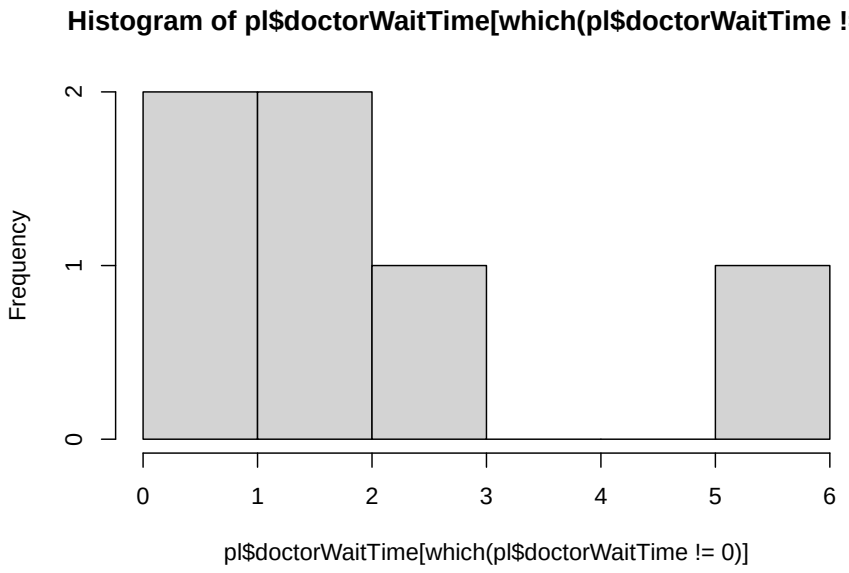
```
totalWait<-0
currentPatientNumber<-0
## clinic opens at 9am -4pm that is 7 hours (420 min.)
timeToClose<-7*60 ## stops admiting patienets in 420 minutes
## current time is 0
## this will be the running total of minutes.
currentTime<-0



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

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

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
```




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



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

```
## [1] -0.1422886
```

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

```
## [1] 43
```

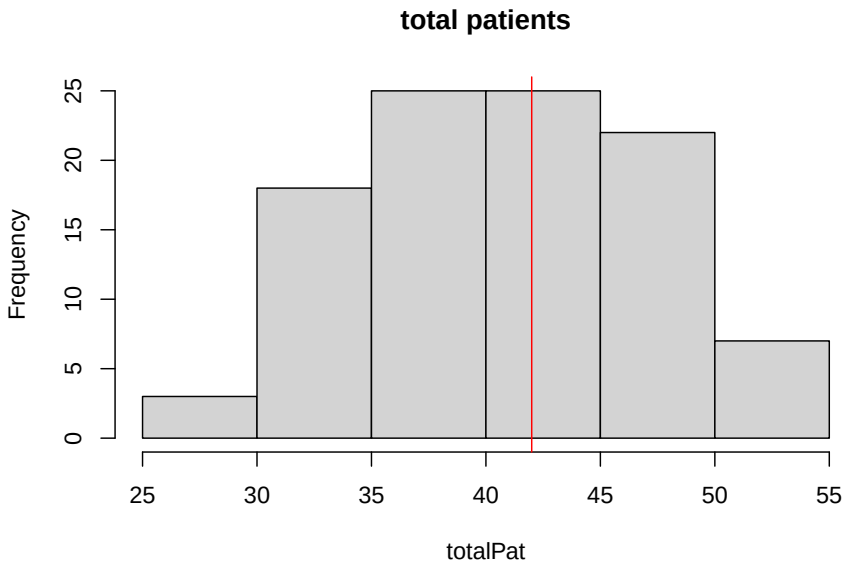
```
## (20-5)/6 + 10 this is about 12.5 minutes of arrival + visit time. which is approxima
## the arrival time is about 10 minutes.

## we should expect 42 patients
 #420/10

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

Simulation 100 times

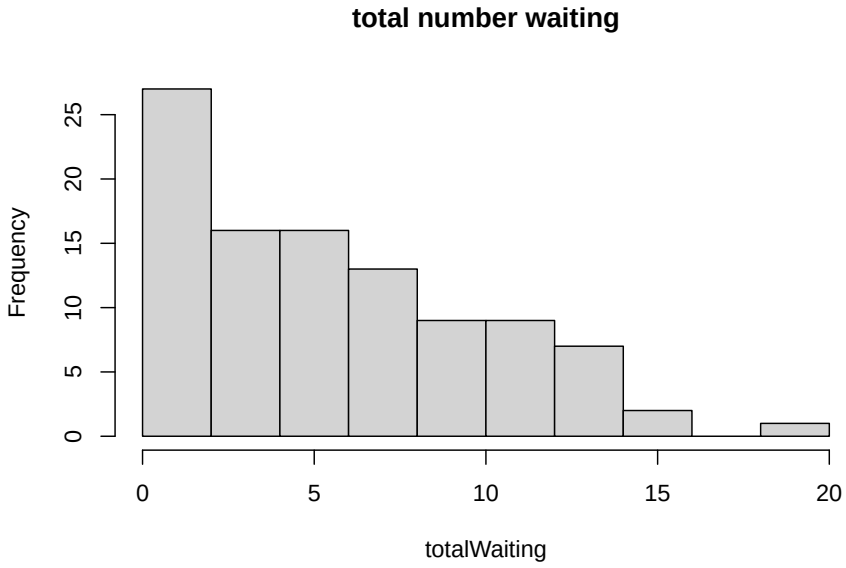
total number of patients was approximately 42, which we expect since the total 420/10. The total number waiting with 3 doctors is 6.61 for 1 day. the average waiting time was about 4-5 minutes. For 1 day, the average closing time was 5.32 minutes after 4 pm

```
 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
```

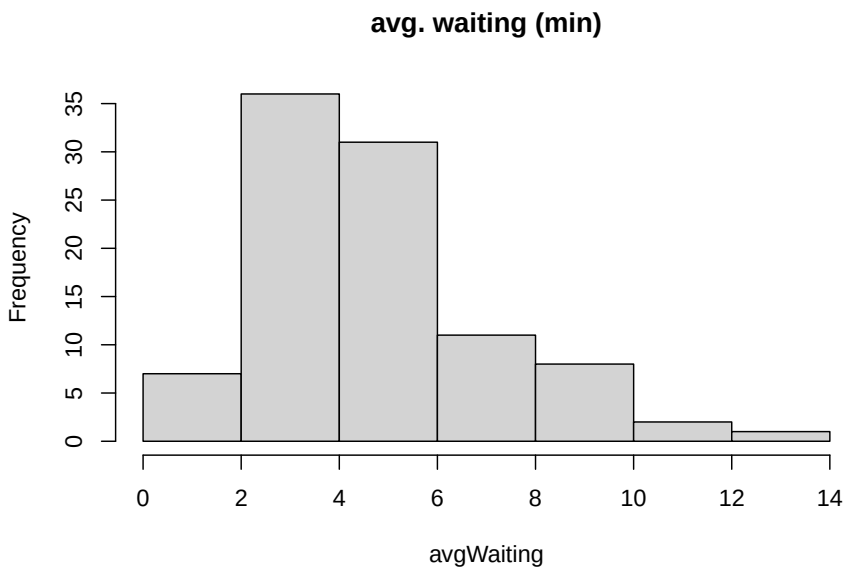
```
 currentPatientNumber<-0
 ## clinic opens at 9am -4pm that is 7 hours (420 min.)
 timeToClose<-7*60 ## stops admiting patienets in 420 minutes
 ## current time is 0
 ## this will be the running total of minutes.
 currentTime<-0


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')
```



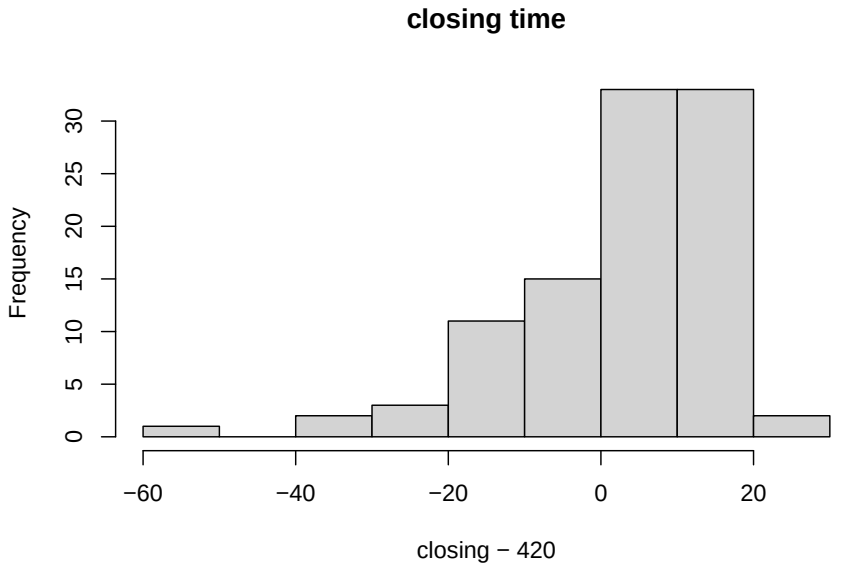

```
hist(totalWaiting,main="total number waiting")
```



```
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

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

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

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

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.

$$E(\theta) = E(E(\theta|y)) \tag{2.3}$$

$$V(\theta) = E(V(\theta|y)) + V(E(\theta|y)) \tag{2.4}$$

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 the 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 differences 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 $p(y|\theta) \propto \theta^a(1-\theta)^b$, 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 θ, then the class P is conjugate for F if $p(\theta|y) \in P$ for all $p(.|\theta) \in F$ and $p(.) \in P$.

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 θ. The $\phi(\theta)$ is called the *natural parameter* for the family (F). The likelihood of a sequence $y = (y_1, ..., y_n)$ iid is

$$p(y|\theta) = (\prod_{i=1}^{n} f(y_i))g(\theta)^n exp(\phi(\theta)^T \sum_{i=1}^{n} u(y_i))$$

$$\propto g(\theta)^n e^{\phi(\theta)^T t(y)}, t(y) = \sum_{y=1}^{n} u(y_i)$$

The *sufficient statistic* for θ is $t(y)$ because the likelihood for $ 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 θ and known variance σ^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 $\theta \sim N(\mu_0, \tau_0^2)$

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

WHere completing the square can find the posterior distribution

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

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

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

the future observations, x, does not depend on the past observations y given θ.

Normal model with multiple observations

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

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

Simplfying the algebra shows the posterior depends on y only through the sample mean (sufficient statistic), $\bar{y}$ is the sufficient statistic for θ, 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-χ^2 distribution with scale σ_0^2 and ν_0 degrees of freedom. The prior distribution for σ^2 is taken as $\sigma_0^2 \nu_0 / X$, where $X \sim \chi^2_{\nu_0}$ random variable.

The resulting posterior density for σ^2 is

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

Which is a scaled inverse-χ^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 (η, ν). 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 Gamma(α, β) equivalent to prior total count of $\alpha - 1$ in β prior observations. The posterior is

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

Negative Binomial distribution

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

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

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

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

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

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

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 θ is

$$p(y|\theta) = \theta exp(-y\theta), y > 0 \tag{2.8}$$

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

Constructing a prior distribution

In the text example, the prior distribution values were set using a Gamma distribution with α, β 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 couny, 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 NB$(\alpha, \beta/10n_j)$.

We then use a method of moments estimator to solve for α, β.

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

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 transofmraiton of variables, the prior density p(θ) is equivalent to the prior density on $p(\phi)$.

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

Jeffreys' general principle iss 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 θ.

$$J(\theta) = E((\frac{dlogp(y|\theta)}{d\theta})^2|\theta) = -E(\frac{d^2logp(y|\theta)}{d\theta^2}|\theta) \tag{2.11}$$

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

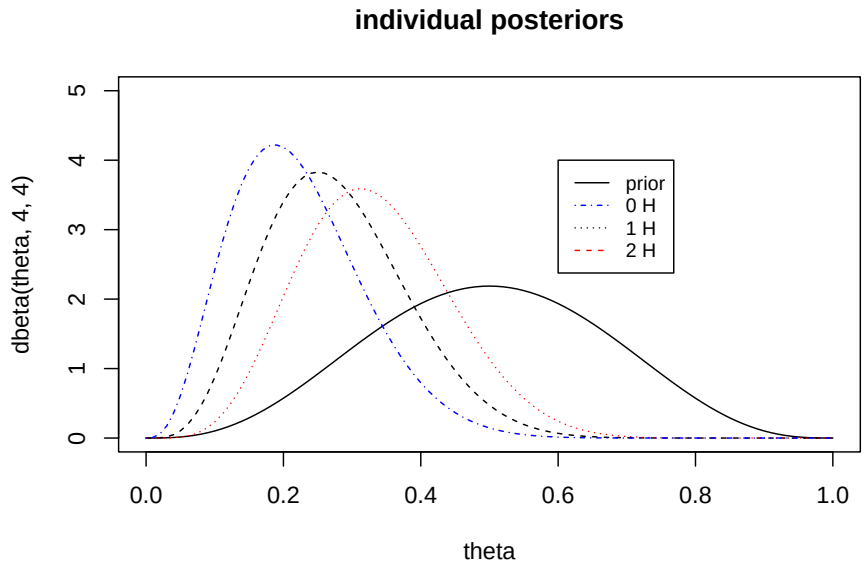
prior Beta(4,4), where a coin is tossed 10 times and heads appears fewer than 3 times. the exact posterior is Beta(4+y, 4+10-y) for y=0,1,2. Since we don't know the observed heads, but that $y < 3$ we plot the posterior distributions for each possibility. For 2 heads it is closer to the prior, with posterior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

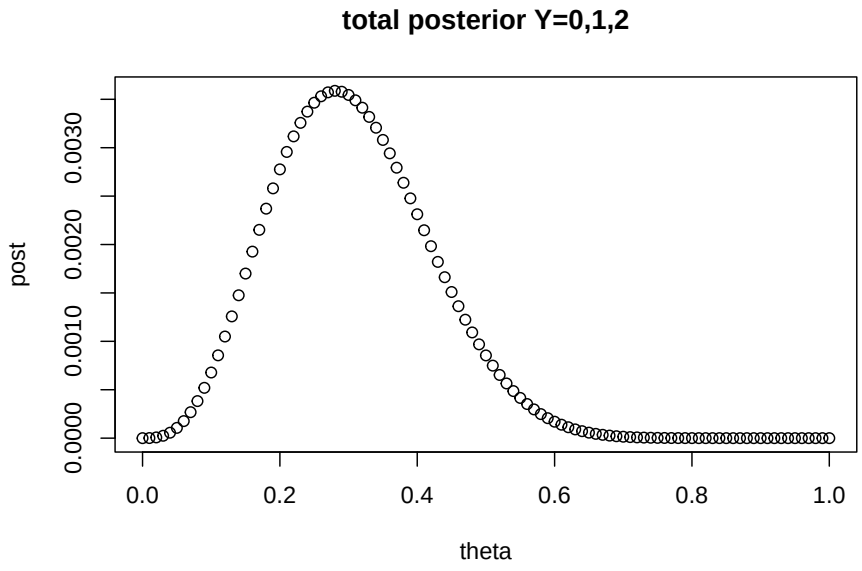
The likelihood of the total event is $\theta^0(1-\theta)^{10} + \binom{10}{1}\theta(1-\theta)^9 + \binom{10}{2}\theta^2(1-\theta)^8$

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 succe
  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)
```



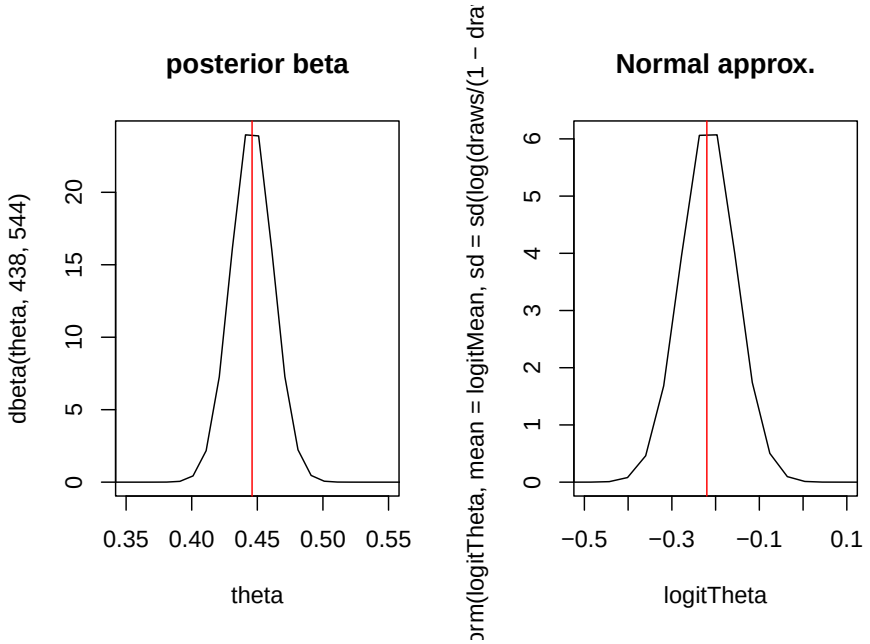

```
post<- theta^3*(1-theta)^(13)+choose(10,1)*theta^4*(1-theta)^12+choose(10,2)*theta^5*(1-theta)^11

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") a
  abline(v=0.446,col='red')
draws<-rbeta(1000,438,544)
plot(logitTheta,dnorm(logitTheta,mean =logitMean, sd=sd(log(draws/(1-draws))) ),type=
  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 approximateley 167, and 833 failures as the prior prediction. with probabiliy of success (1/6).

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

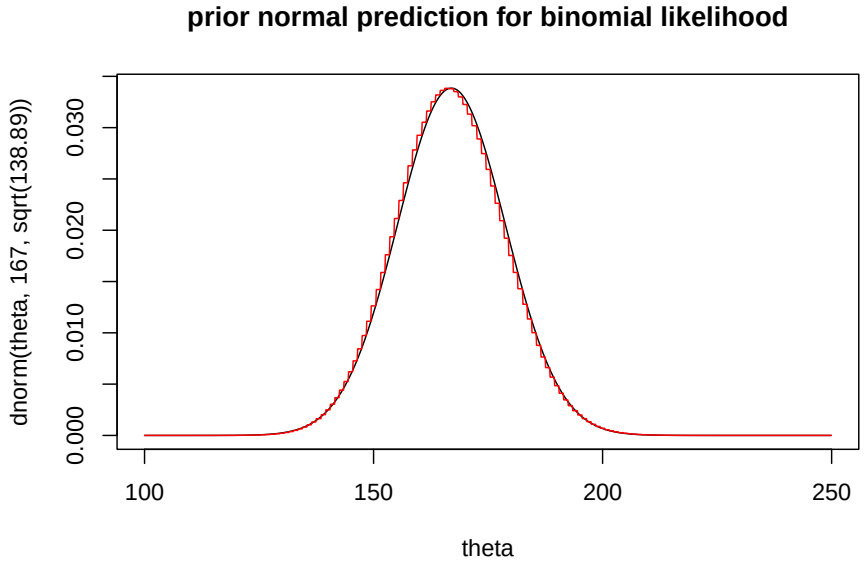
The normal approximation for the prior prediction uses the binomial distribution is $\mu = n * p = 167$ and $ ^2 = npq $= 138.89 $\sim N(167, 138.89)$

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 hte 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 fo
     lines(theta,dbinom(round(theta),size=1000,prob=(1/6)),main="prior binom predictio
```




```
  ## 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%
## 138.5098 156.0467 168.4335 178.5953 194.0984
```

```
     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

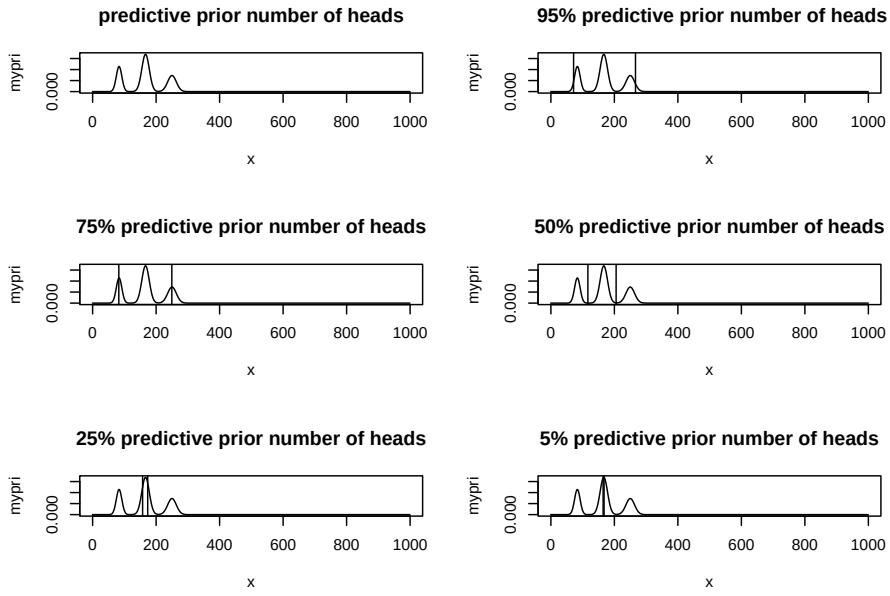
We have a mixture of 3 normal distributions, and show the central intervals for 5,25,50,75, and 95% predictive probabilities. The question gives θ as the probability of a 6 on a die, possibly unfair, in 1,000 tosses. we have $\theta = 1/12, 1/6, 1/4$ types of biased die. Using the normal approximation, 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 $\mathrm{p}(\theta) \propto \theta^{-}1(1-\theta)^{-}1$ 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.

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

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(ϕ) $\propto 1$. 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, θ, 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,sigm
    upper<-round(qnorm(c(0.025,0.975),mu_n(mu0,ybar,tau02,10,sigma2),sd=sqrt(taun2(tau

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(tau0

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,sigma
    upper<-round(qnorm(0.975,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(taun2(tau02,100,s

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(taun2(tau02,100,sigma2)+sigma2
    upper2<-round(qnorm(0.975,mu_n(mu0,ybar,tau02,100,sigma2),sd=sqrt(taun2(tau02,100,sigma2)+sig

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 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 $ 0.0471. The posterior mean and standard deviation is 279.09 a.nd 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) = 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 empirial 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] 279.9156
```

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

```
## [1] 80.39583
```

- (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.

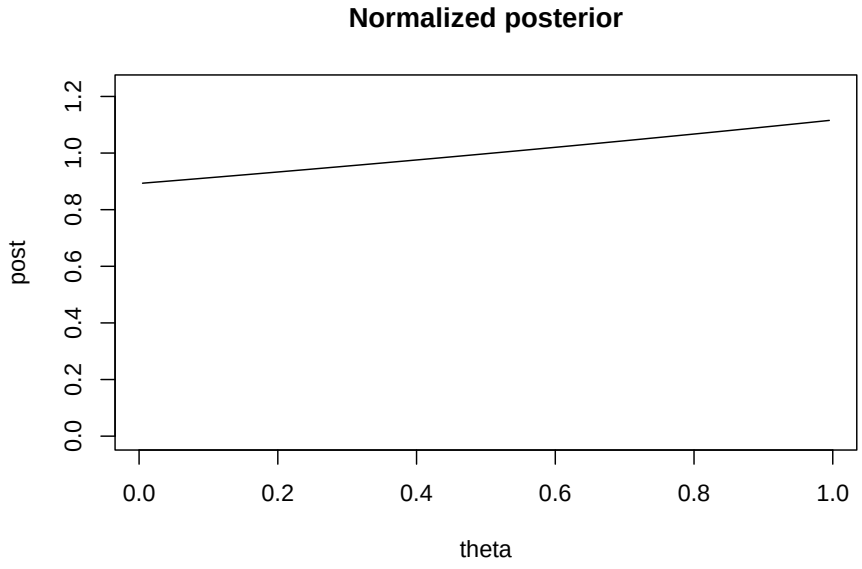
Question 11

suppose y1,...,y5 are iid 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 $\theta = 01/m, 2/m, ...100$. using the grid approximation, compute and plot the posterior density as a function of θ

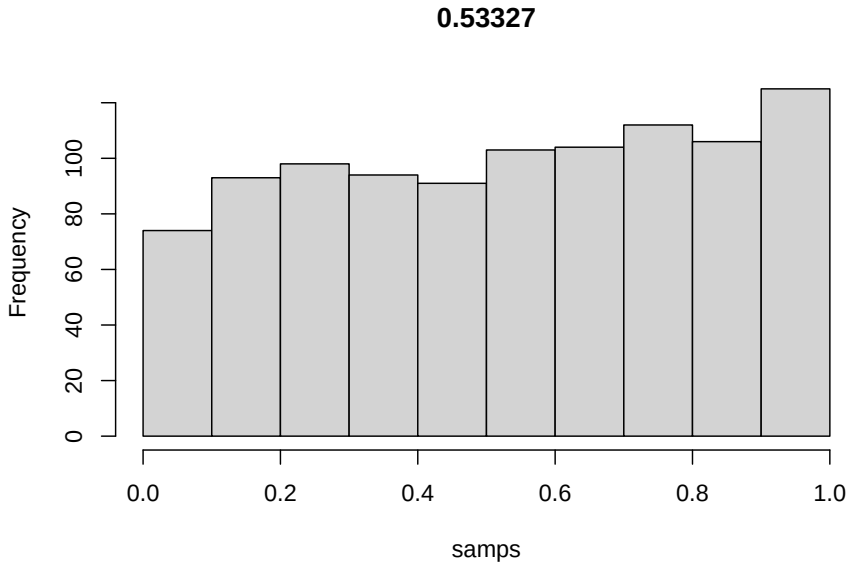
for the likelihood $L(\theta|y) = \prod_{i=1}^{n} f(y_i|\theta)$ requires the product of y for a given theta. this was a mistake i made in the first attempt. The posterior is $p(\theta|y) = L(\theta|y)p(\theta)$ wher $p(\theta) = 1/100$

```
 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| 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))
```



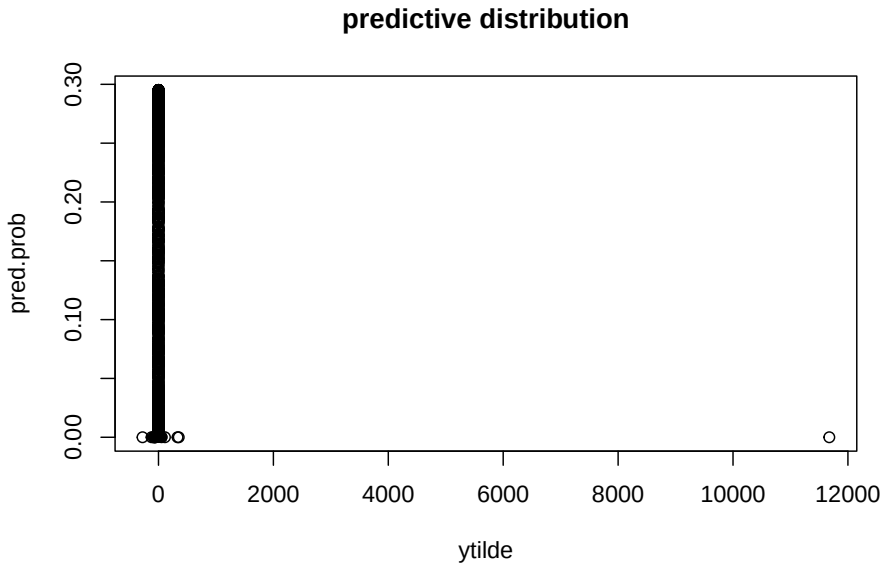

```
  ##sample mean is close to the mean of the observed.
```

- (c) Using the previous 1000 samples of θ 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.
##  -277.613    -0.733     0.454    11.857     1.386 11676.002
```

```
## 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, s
 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 0.516
```

Question 12

Suppose $y|\theta \sim Poisson(\theta)$, find Jeffreys' prior density for θ and then find α, β 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 closes 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 (β) 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 θ from the posterior distribution, and for each θ_j, we sampled a random variable from Poisson($\lambda = \theta_j$).

-Using Jeffreys' prior, the posteror is $p(\theta|y) \sim Gamma(238.5, 10.2)$. Using simulation the 95% predictive invterval 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 empi
#  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
```

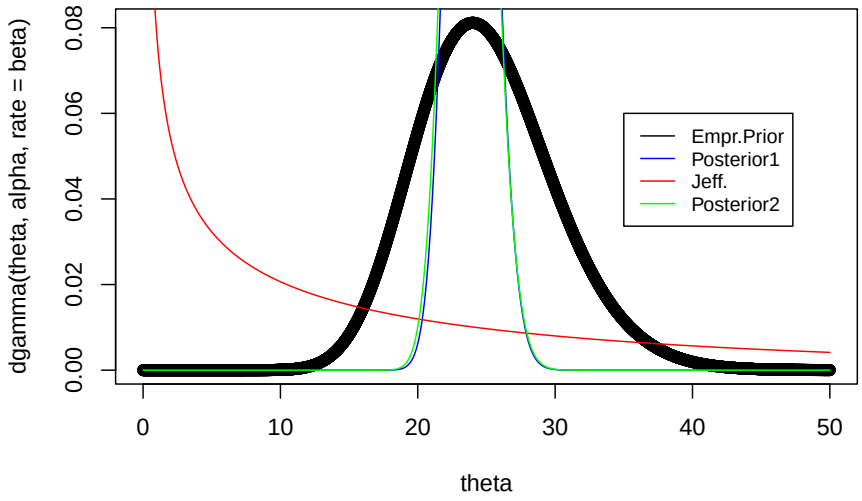
```
## non-informative prior using Jeffrey's prior
  Jalpha= 1/2
   Jbeta<- 0.02 ## we set close to 0
    ## prior mean is 25 which matches the observed data.
   ## the mean rate given the data is 0.126, so we find the beta parameter to match the
message("the prior interval:")
```

```
## 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 ##
```

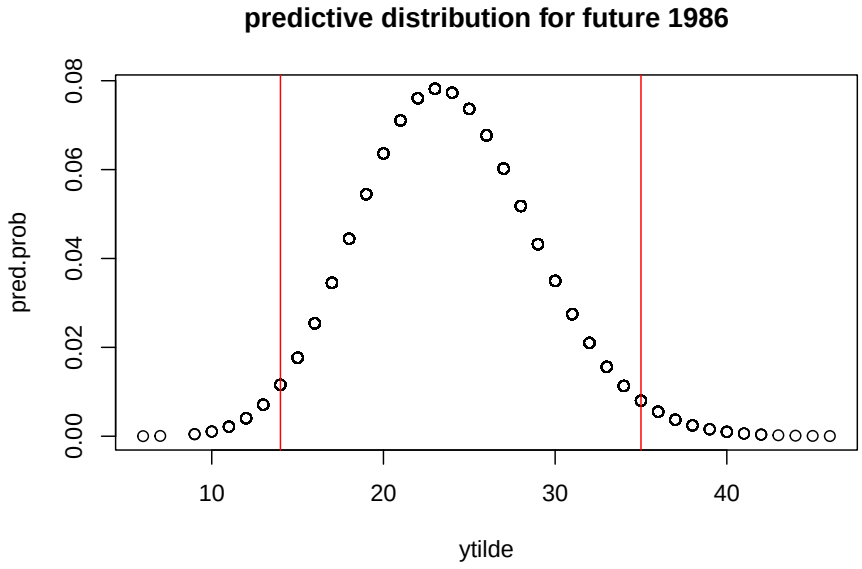
```
  sigma2<-10^2  ## we assume we know the variance of the fatalities (empirical variance


  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,sigm
  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.
##    6.00   20.00   24.00   23.85   27.00   46.00
```

```
## 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_p

 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    35
```

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.

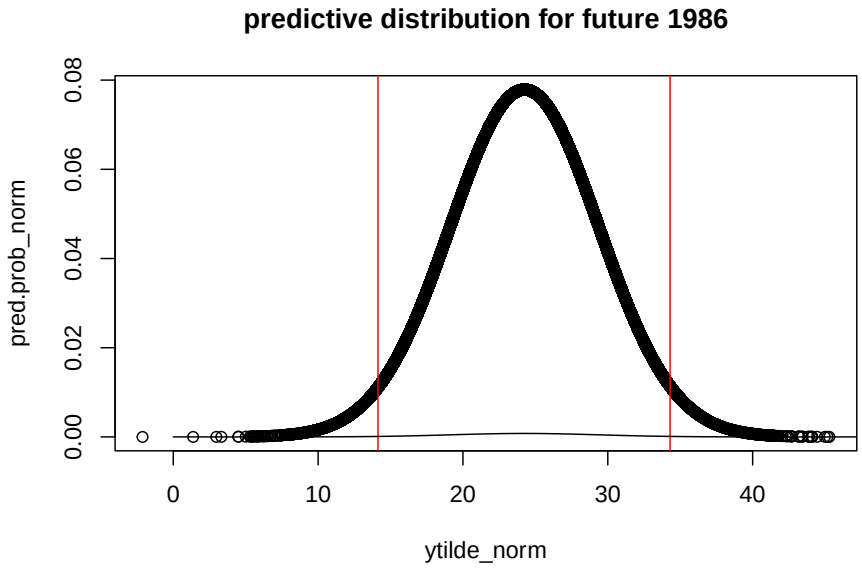
```
  ## Predictive with normal

## 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.
##  -2.122  20.813  24.255  24.245  27.694  45.292
```

```
## 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, s
pred.prob_norm<-sapply(ytilde_norm,function(x) sum(dnorm(x,24.25,sqrt(26.252))*post_pr

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%
## 14.13409 34.29907
```

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

```
## average predictive value from posterior:23.8522
```

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

let θ 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 θ = 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 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 θ from $p(\theta|y)$ Gamma posterior to generate θ. 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 $x \sim Pois(8e11 * \theta_j)$ 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)
```

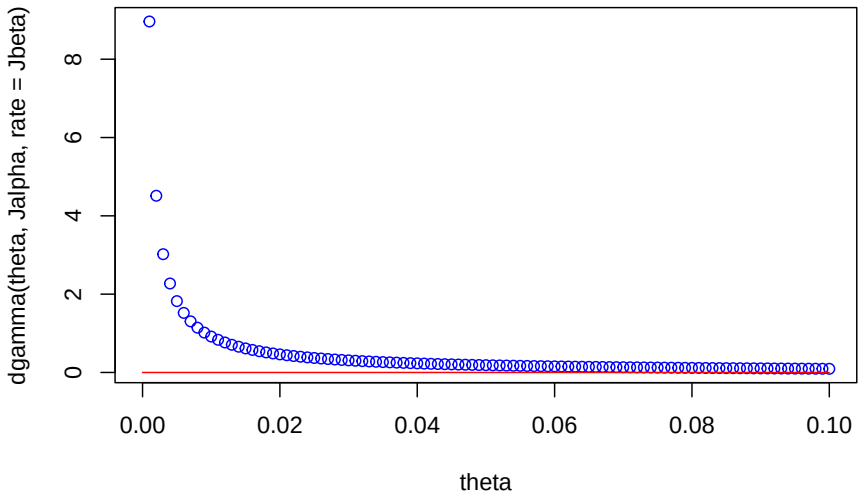
```
theta=seq(from=0,to=0.1,by=.001)
## empirical prior
 alpha= 5
  beta<- 1000
  ## the mean is 0.005 and quantiles 0.002 and 0.01 which is close to the observed ra
  #mean(acc_rate) ## fatal accidents per 100,000 million miles driven
  print("empirical prior 95% quantiles: ")
```

```
## [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 th
  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 approxim
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] 22 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
```

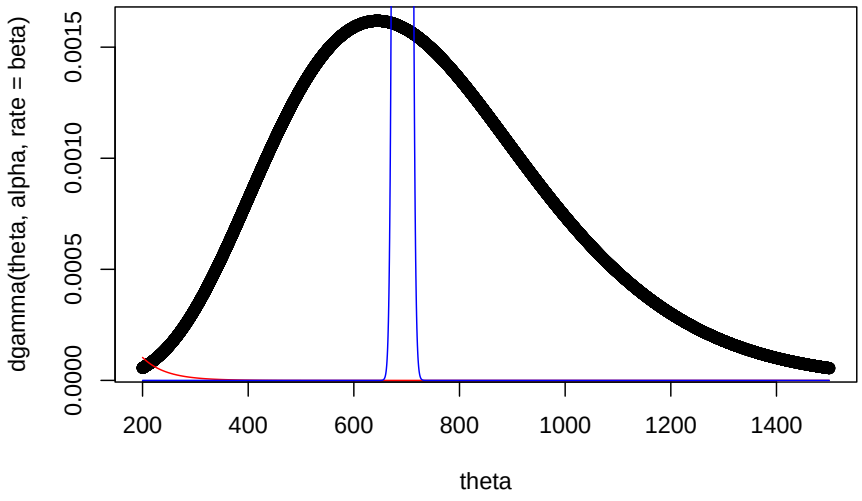
```
   ## Pois(a) ,  mean =a, var= a
    ## Pois(a) ~ N(a,sd= sqrt(a))

 ## 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)," "
```

```
## 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),'
```

```
## 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
```

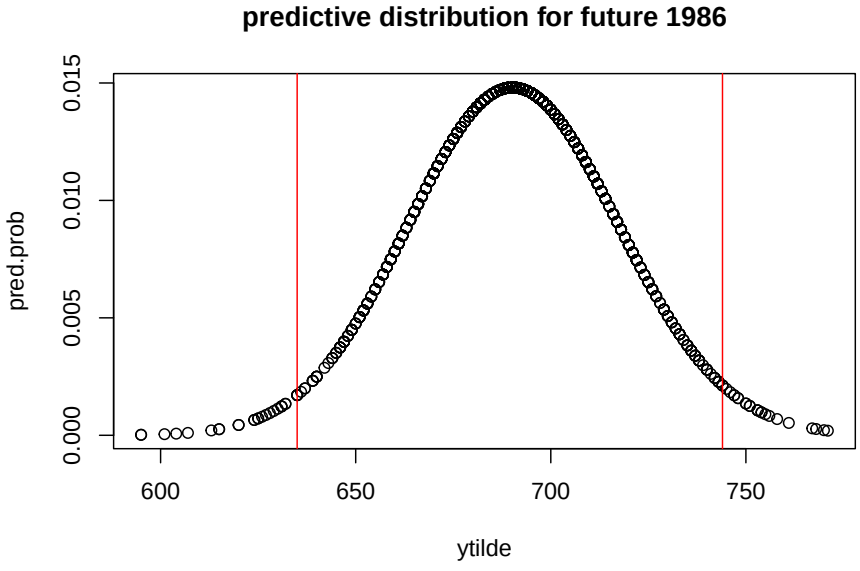
```
   quantile(ytilde_empiricalPrior,c(0.025,0.975))
```

```
##    2.5%   97.5%
## 638.975 746.025
```

```
## 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.
##   595.0   673.0   691.0   690.7   709.0   771.0
```

```
  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')
```



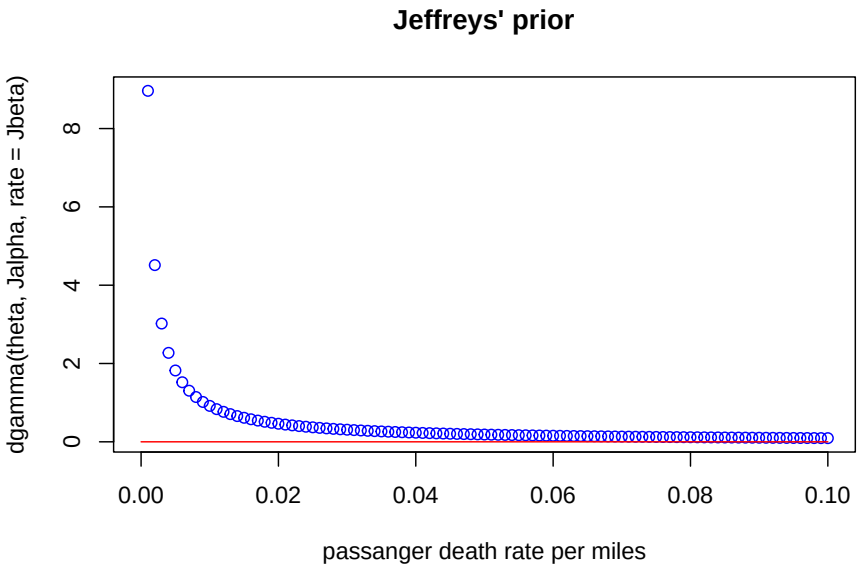
```
# lines(theta,post_pois)
 quantile(ytilde,c(0.025,0.975))
```

```
##  2.5% 97.5%
##   635   744
```

- (d) Using the rate the posterior rate mean is 0.12 estimated from sampling the posterior distribution. For 8 x $10^1$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')
```



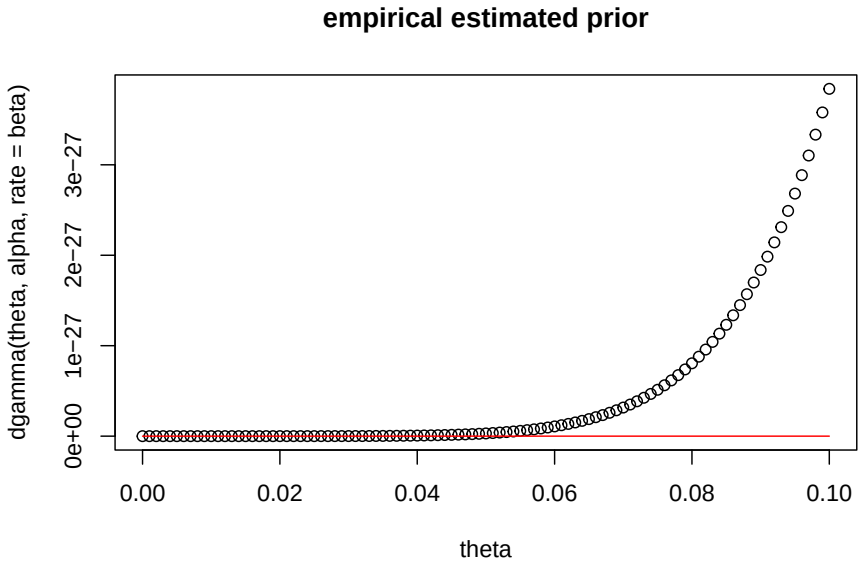

```
## credible interval of the posterior
pois_credible<-qgamma(c(.025,0.975),Jalpha+sum(death),rate=Jbeta+sum(miles))
message("Jeffreys' prior posterior gamma interval n=10: ")
```

```
## 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')
```




```
## credible interval of the posterior


theta_post<-rgamma(1000,Jalpha+sum(death),rate=Jbeta+sum(miles))
 y1986<-rpois(1000,8e+11*theta_post) ## use in the likelihood
 message(" 8,000 per 100 million miles flown has passanger death : Gamma posterior & .
```

```
##  8,000 per 100 million miles flown has passanger death : Gamma posterior & Jeffreys
```

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

```
## [1]  905 1031
```

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
##              4535              1470              1380
```

```
 ## 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/2008El
ele<-read.csv("C:/Users/antho/Documents/Keck-graduate-school/PM590-Bayesian/bookdown-files/2008El
ele<-ele[which(ele$state!="Alaska"& ele$state!="Hawaii"),]
ele$state<-tolower(ele$state)
## graph poprotion 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
```

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

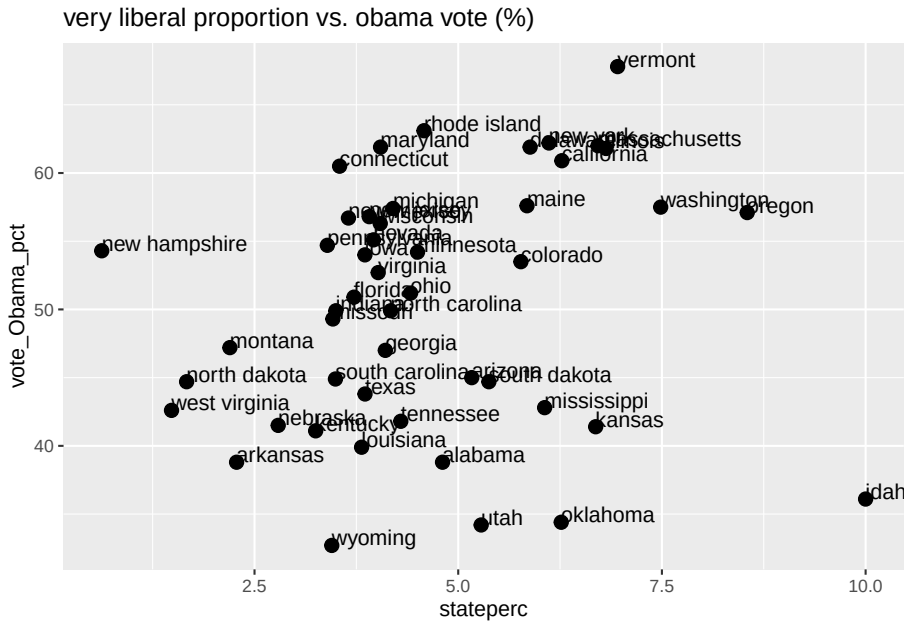
```
tally<-polls%>%group_by(state,ideo)%>%summarize(n=n())
```

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

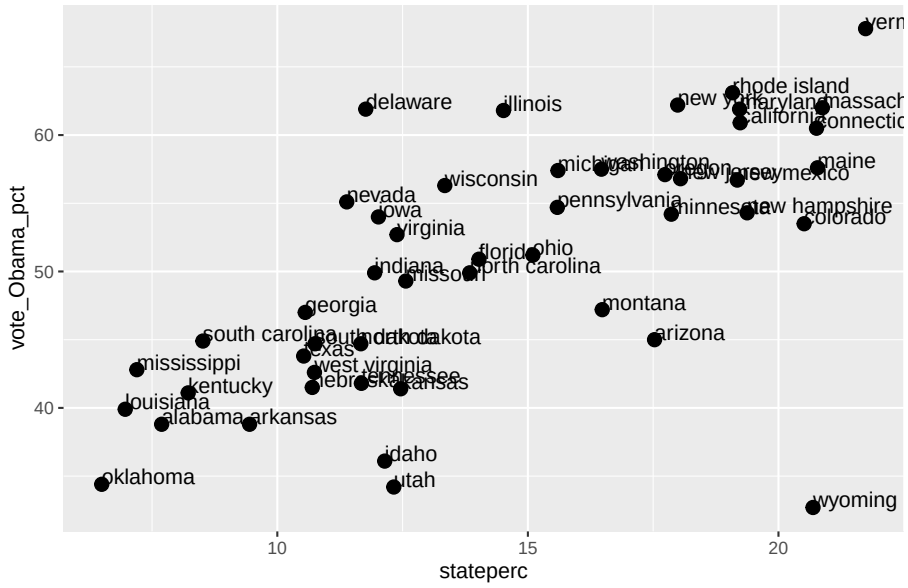
```
statetotal<-tally%>%group_by(state)%>%summarise(statetotal=sum(n))
tally<-left_join(tally,statetotal,by='state')
## average the density of the polling regions by state taking the density of each zip
densit<-polls%>%group_by(state)%>%summarize(density=mean(density))

# 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$density
   tally$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 sta
vl<-subset(tally,tally$ideo=='very liberal')
 hist(vl$stateperc, xlab="very lib. (%)", ylab="frequency")
```


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

  ## empircal 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)
```

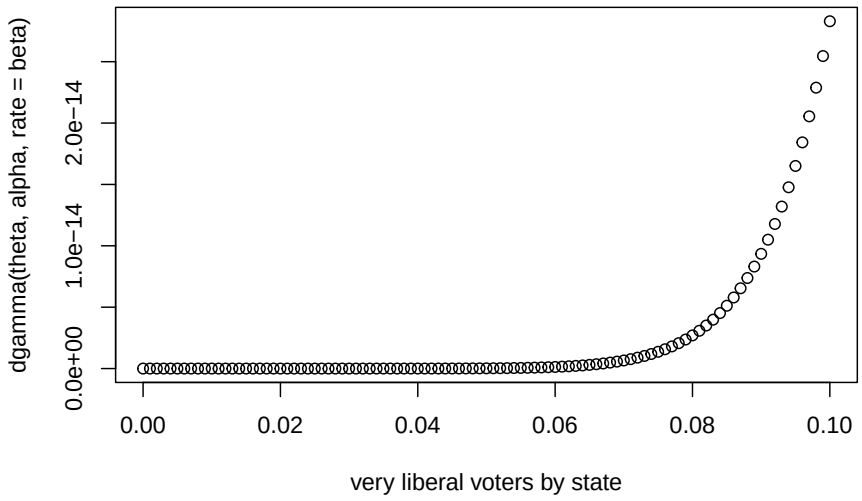
```
## [1] 2.422521 7.788166
```

```
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')
```

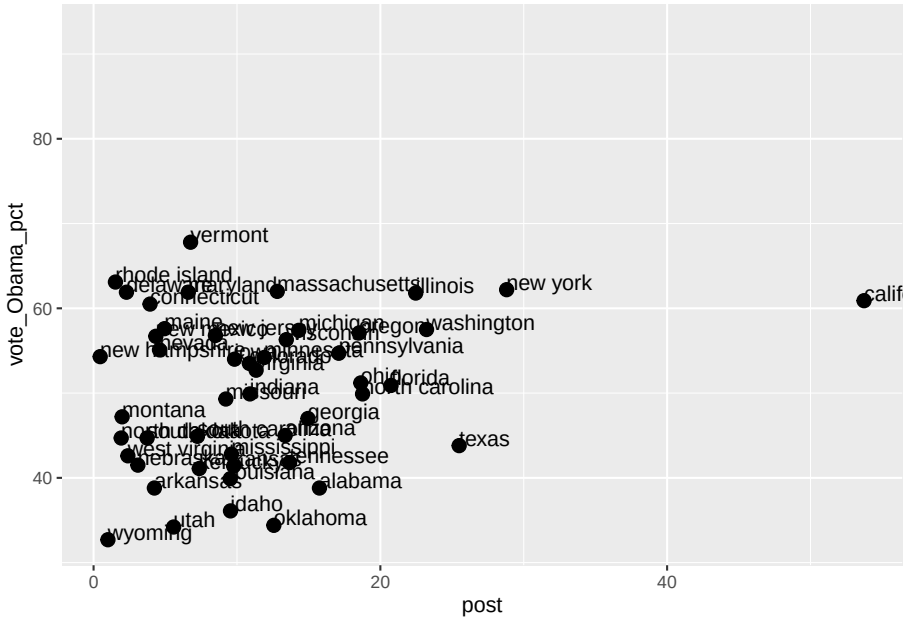



```
dat<-tt[which(tt$ideo=='very liberal'),]


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
```

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

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



Chapter 3

Chapter 3

Exercises

Question 1

- (a) find the marginal posterior of $\alpha = \theta_1/(\theta_1 + \theta_2)$ We need to find the joint posterior of θ_1, θ_2 and then perform a change of variables to (α, β). Let the prior $p(\theta) = Diri(a_1, ..., a_J)$ where y follows a multinomial likelihood.

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

  The posterior follows a Dirichlet distribution, but we are interested only in the θ_1, θ_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

$$(\theta_1, \theta_2, 1 - \theta_1 - \theta_2|y) \sim Diri(y_1 + a_1, y_2 + a_2, a_0 + y_0 - y_1 - y_2 - a_1 - a_2)$$

$$p(\theta_1, \theta_2, 1 - \theta_1 - \theta_2) \propto \theta_1^{y_1 + a_1 - 1} \theta_2^{y_2 + a_2 - 1} (1 - \theta_1 - \theta_2)^{a_0 + y_0 - y_1 - y_2 - a_1 - a_2 - 1}$$

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 $y_0' = \sum y_i - y_1 - y_2$

$$\begin{aligned} p(\alpha, \beta) &= (\alpha * \beta)^{y_1 + a_1 - 1} (\beta(1 - \alpha))^{y_2 + a_2 - 1} (1 - \alpha * \beta - \beta(1 - \alpha))^{a_0' + y_0' - 1} |\beta| \\ &= \alpha^{y_1 + a_1 - 1} (1 - \alpha)^{y_2 + a_2 - 1} \beta^{y_1 + a_1 + y_2 + a_2 - 1} (1 - \beta)^{a_0' + y_0' - 1} \\ &= Beta(y_1 + a_1, y_2 + a_2) Beta(y_1 + a_1 + y_2 + a_2, a_0' + y_0') \end{aligned}$$

By factorization we integrate out with respect β 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 $p(y|\alpha) \propto \alpha_1^y(1-\alpha)_2^y$ 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

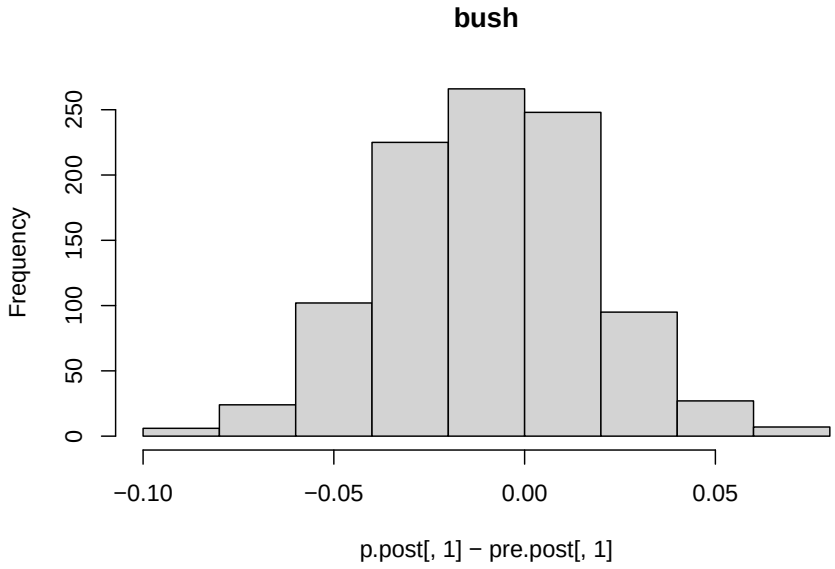
639 were polled before the debate and 639 different persons were polled afterward. we let j=1,2 let α_j be proprtion of voters who preferred Bush out of those who had a preference for bush or Dukakis at the time of the survey j. we need to model two different multinomial distributions, and find the posterior 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(3

# 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
```

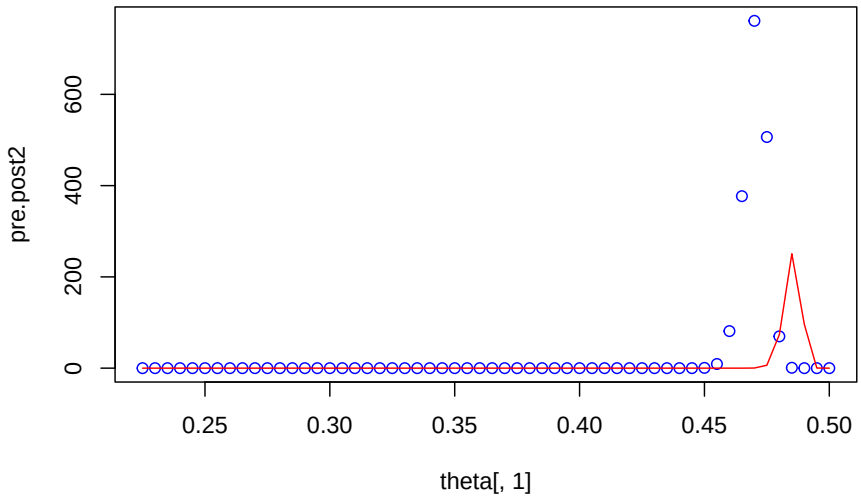
```
 theta<-data.frame(theta1,theta2,theta3)
 ## need the pre posterior
 ## the prior alpha1, alpha2 =0 for improper prior
 library(gtools)
  pre.post<-rdirichlet(1000, c(as.numeric(debate[1,2:4])))
   p.post<-rdirichlet(1000, c(as.numeric(debate[2,2:4])))
  ## differences between bush post .vs pre
   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
##   623   377
```

```
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')
```

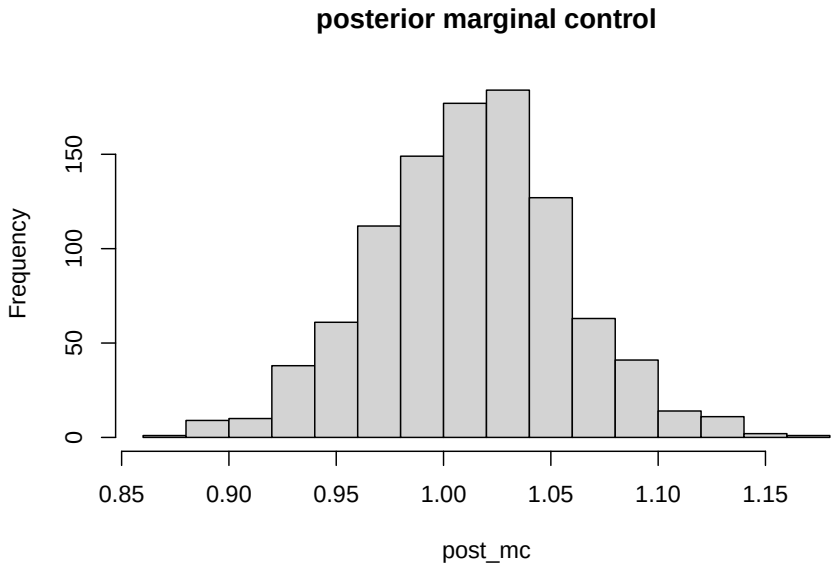


Question 3

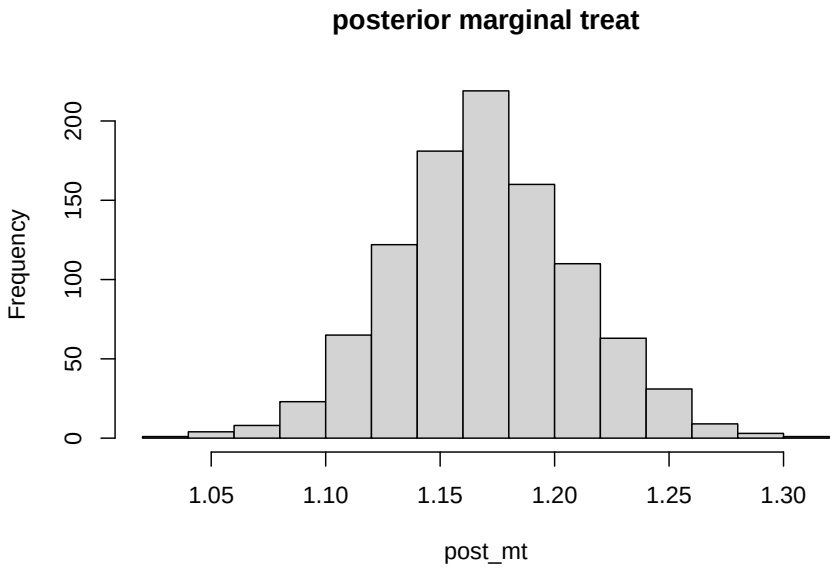
we are given two independent normal random variables with unknown variances and unknown true means. - (a) Assume a uniform prior on $(\mu_c, \mu_t, log(\sigma_c), log(\sigma_t))$ find the posterior of μ_c and μ_t.

```
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
```

```
## 95% posterior margin control:0.93 1.1
```

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

```
## 95% posterior margin control:1.09 1.25
```

```
        quantile(post_mc,c(0.025,0.975),na.rm=TRUE)
```

```
##      2.5%     97.5%
## 0.9219155 1.1027542
```

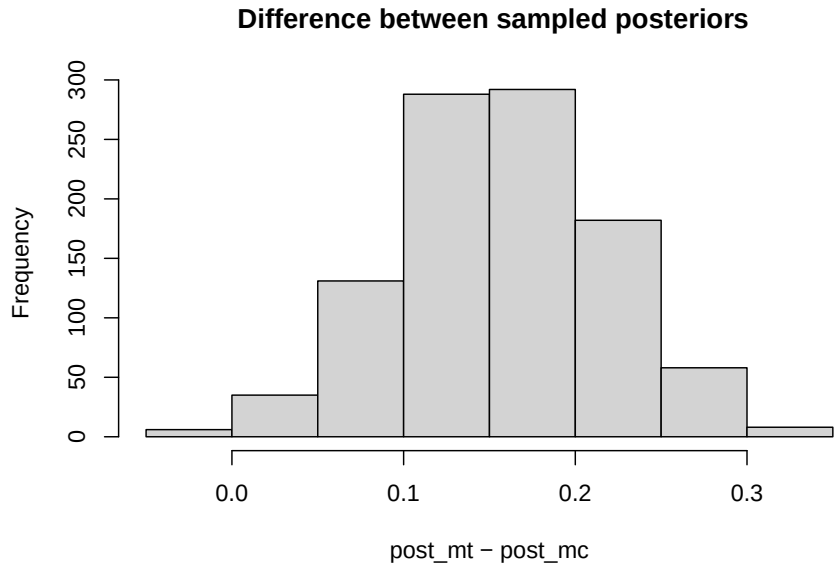
```
        quantile(post_mt,c(0.025,0.975),na.rm=TRUE)
```

```
##     2.5%    97.5%
## 1.089495 1.247000
```

- (b) what is $\mu_t - \mu_c$

The posterior interval (central) of the differences between 2 independent t-distributoins 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.

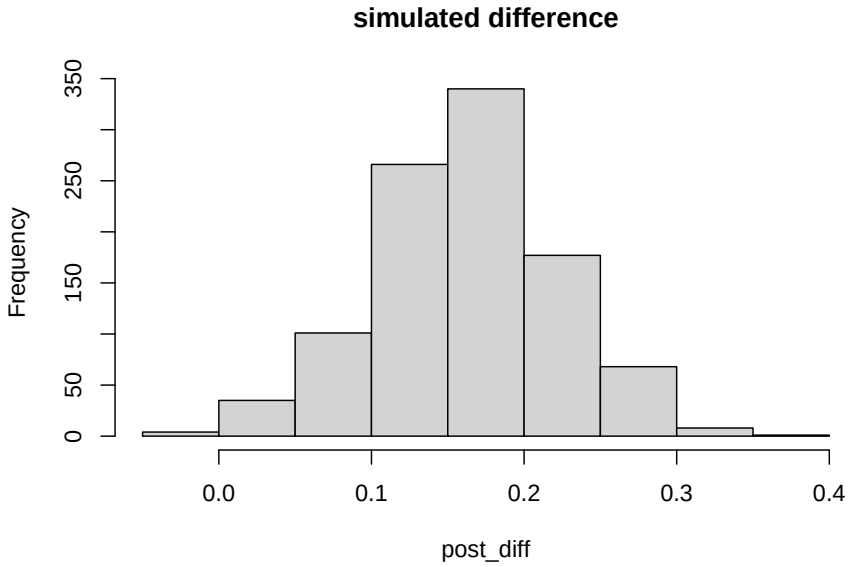
```
hist(post_mt-post_mc, main = ' Difference between sampled posteriors')
```




```
        quantile(post_mt-post_mc,c(0.025,0.975),na.rm=TRUE)
```

```
##       2.5%      97.5%
## 0.03927942 0.27312911
```

```
  ## 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),na.rm=TRUE)
```

```
##       2.5%      97.5%
## 0.04267863 0.27475540
```

```
## 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)
```

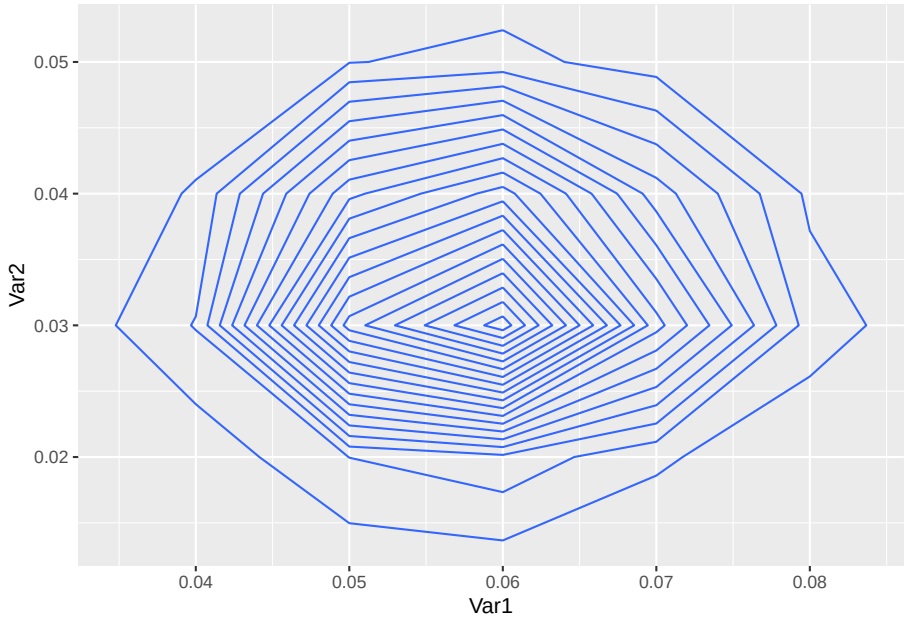
```
## [1] 0.04381525 0.27618475
```

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.

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

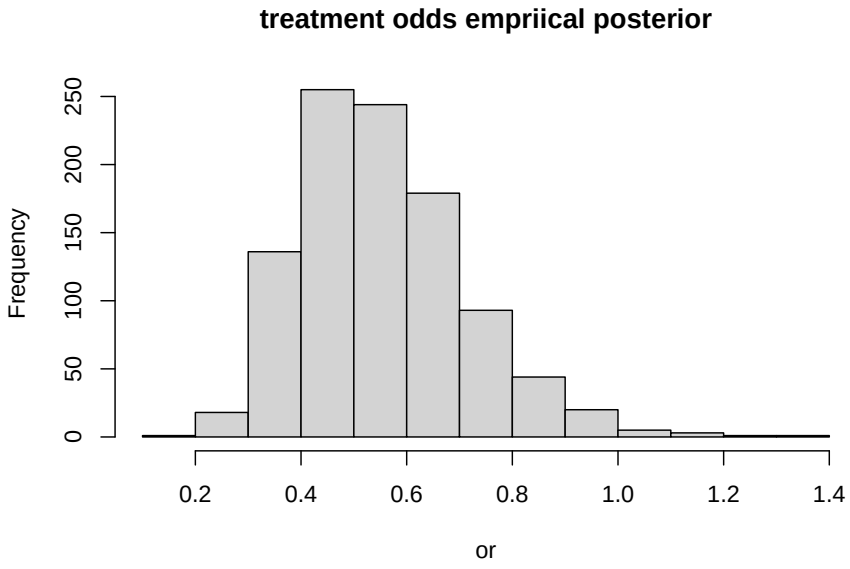
```
p0= seq(0.01,0.99,by=0.01)
p1= seq(0.01,0.99,by=0.01)

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 empriical posterior')
```




```
  summary(or)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1778  0.4421  0.5355  0.5559  0.6485  1.3261
```

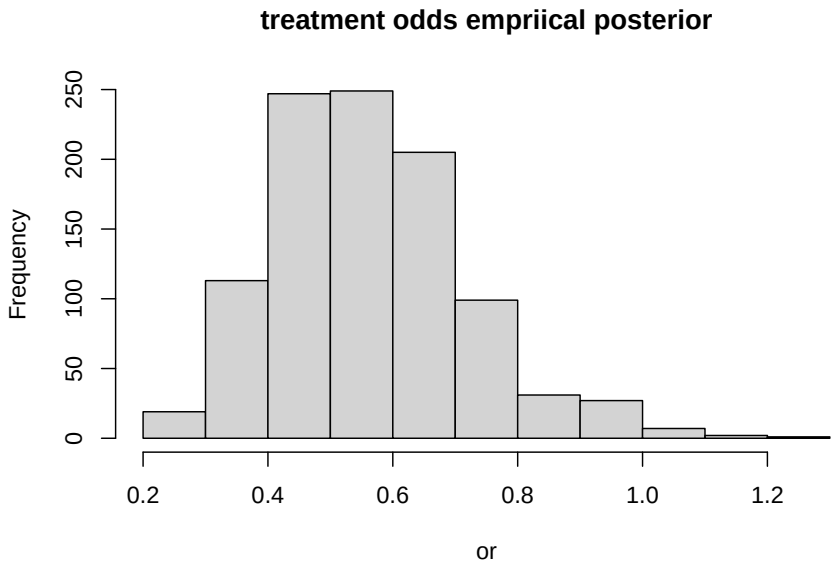
```
  message("the empricial OR: ", round((22*635)/(39*658),2))
```

```
## 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 empriical posterior')
```




```
summary(or)
```

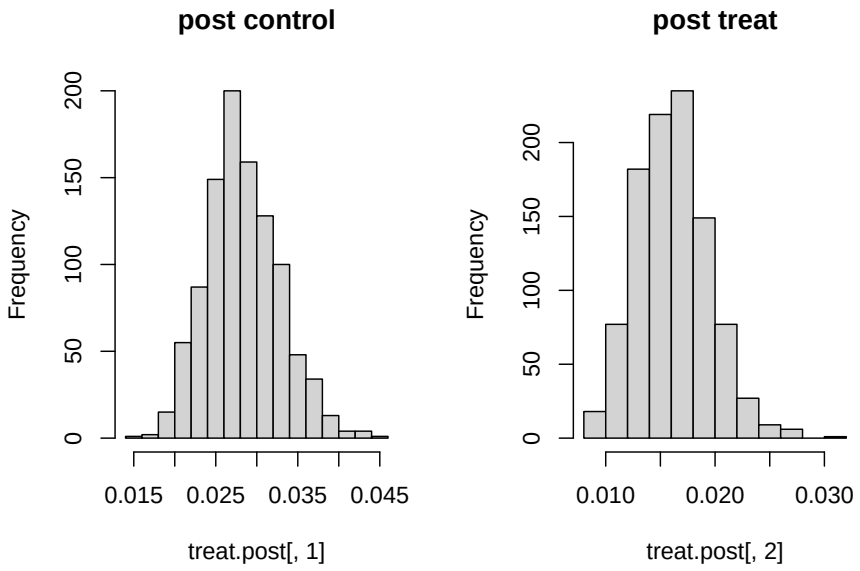
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2311  0.4509  0.5558  0.5635  0.6445  1.2841
```

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 Diri(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 independnet 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
```

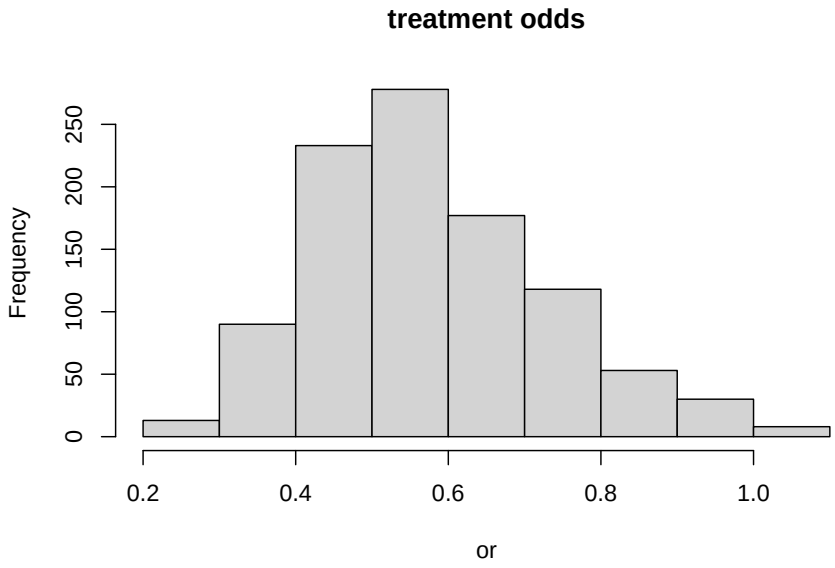
```
##                         NA
## odds ratio with 95% C.I.  estimate     lower     upper
```

```
##                     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.
##  0.2526  0.4688  0.5581  0.5784  0.6743  1.0987
```

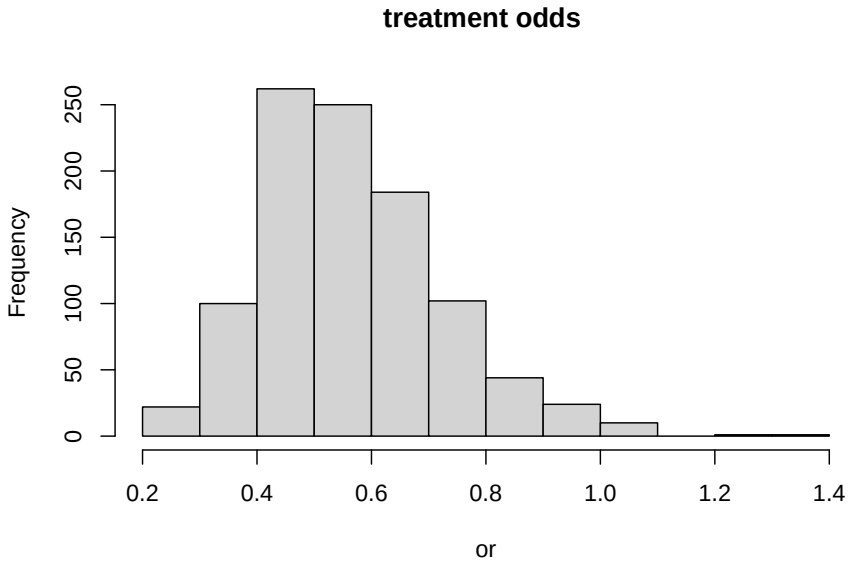
```
  hist(or,main='treatment odds')
```



- (4c) showing the improper prior (a=0), the posterior odds is 0.5793 which is very robust and similar to the noninformative. Using the emprical 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))
```

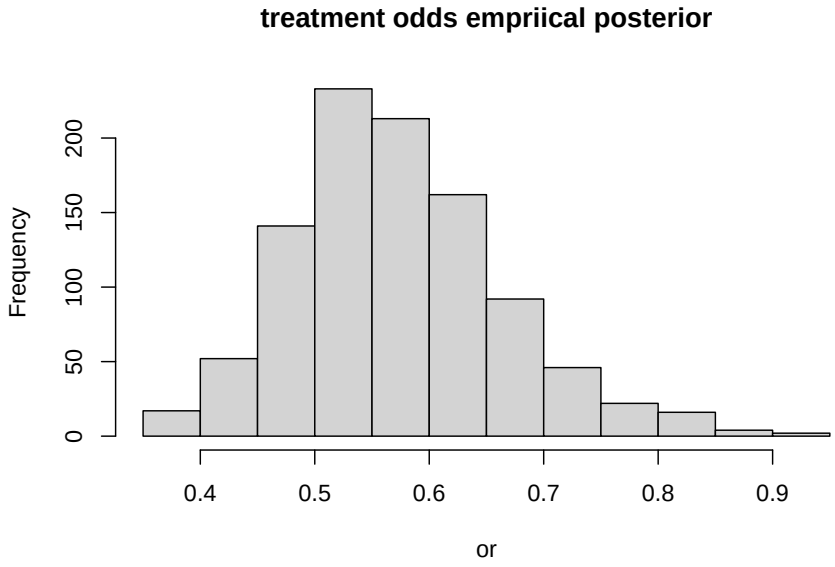
```
## b
or<-(treat.post[,2]/(1-treat.post[,2]))/(treat.post[,1]/(1-treat.post[,1]))
hist(or,main='treatment odds ')
```




```
summary(or)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2215  0.4557  0.5439  0.5658  0.6600  1.3556
```

```
## 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 empriical posterior')
```



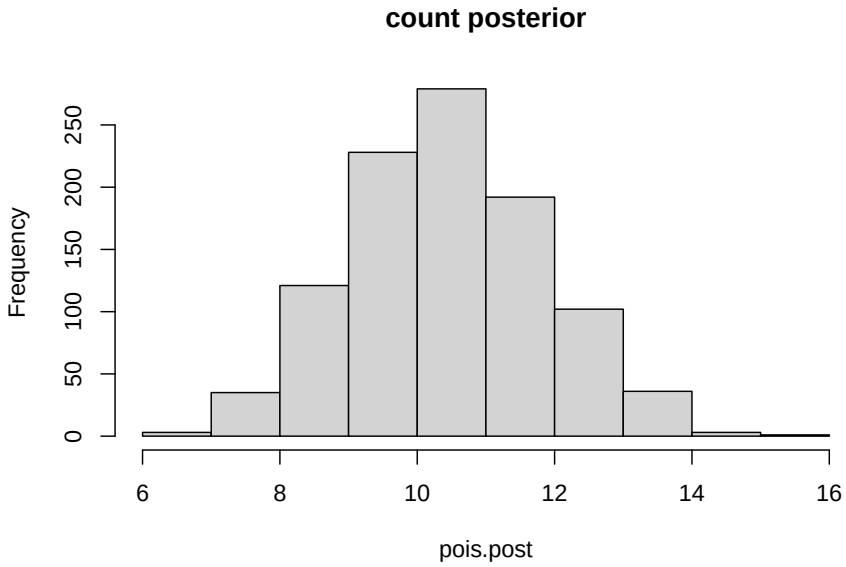

```
summary(or)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3556  0.5103  0.5629  0.5730  0.6268  0.9278
```

Question 5

- (a) Give the posterior distribution of μ, σ^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) th
# 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.
##   6.625   9.482  10.410  10.439  11.413  15.243
```

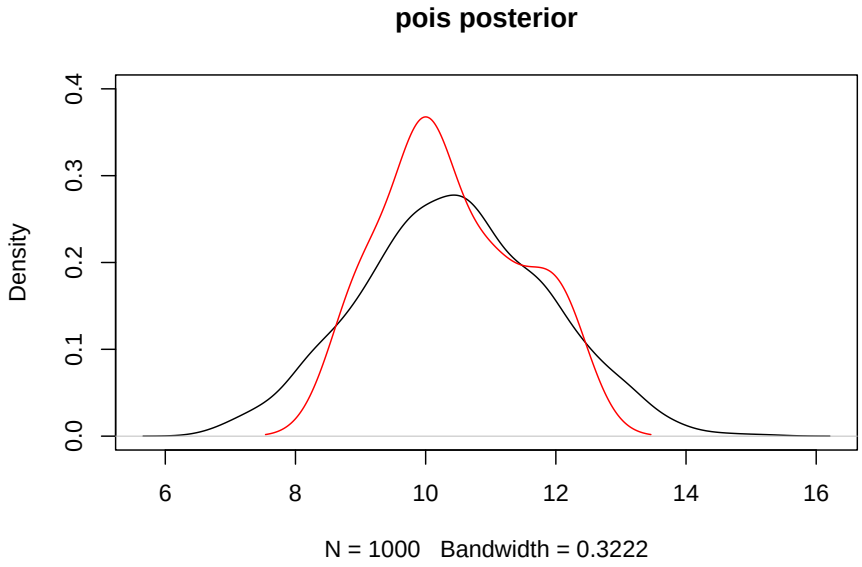
```
  mean(pois.post)
```

```
## [1] 10.43942
```

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

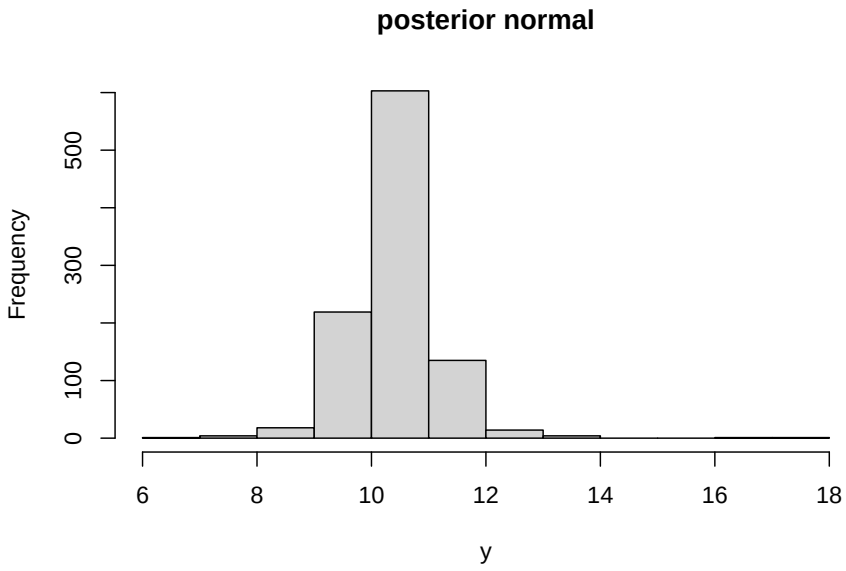
```
## [1] 2.030794
```

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

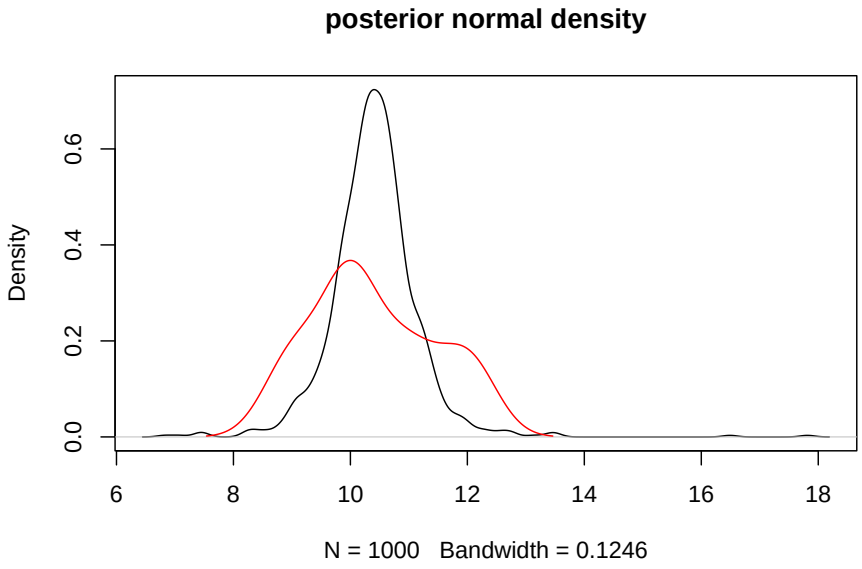


- (b) Give the correct posterior distribution treating the measurements as rounded Using non informative prior for both unknowns, the posterior mean is 10.4 and posterior variance is 0.433

```
## 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.

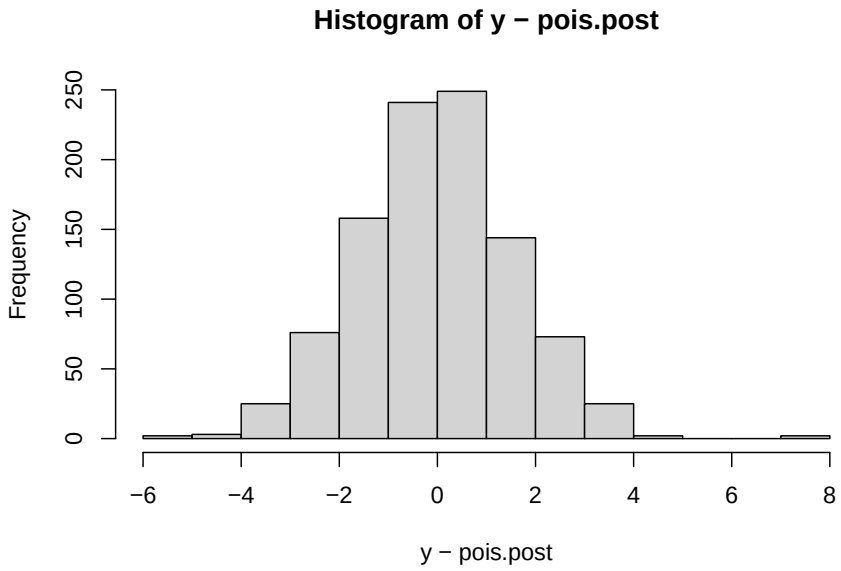
```
mean( y - pois.post) # 0.002926538
```

```
## [1] -0.02619315
```

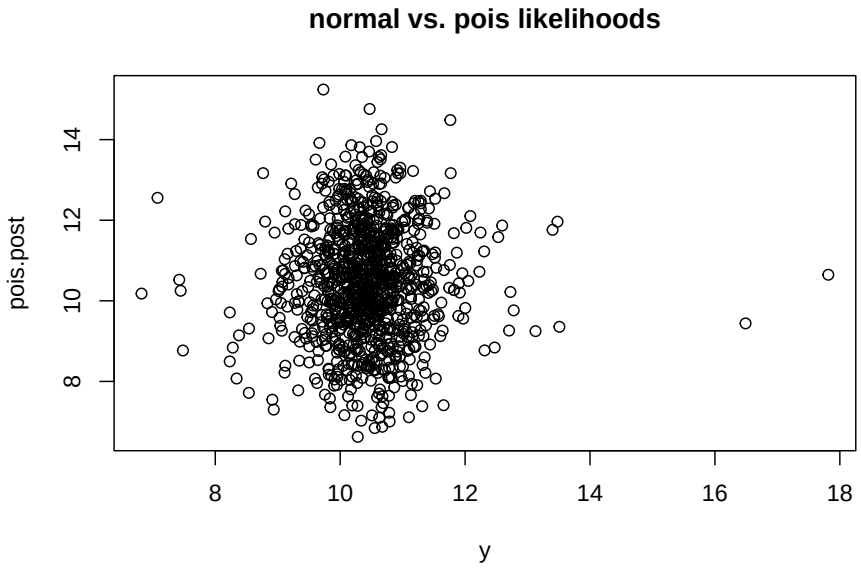
```
var( y - pois.post) # 2.7
```

```
## [1] 2.56876
```

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



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



- (d) Let z=(z1,z2,z3,z4,z5) be the original unrounded measurements corresponding to the five observations above. draw simulations of z and compare the posterior $(z1 - z2)^2$.

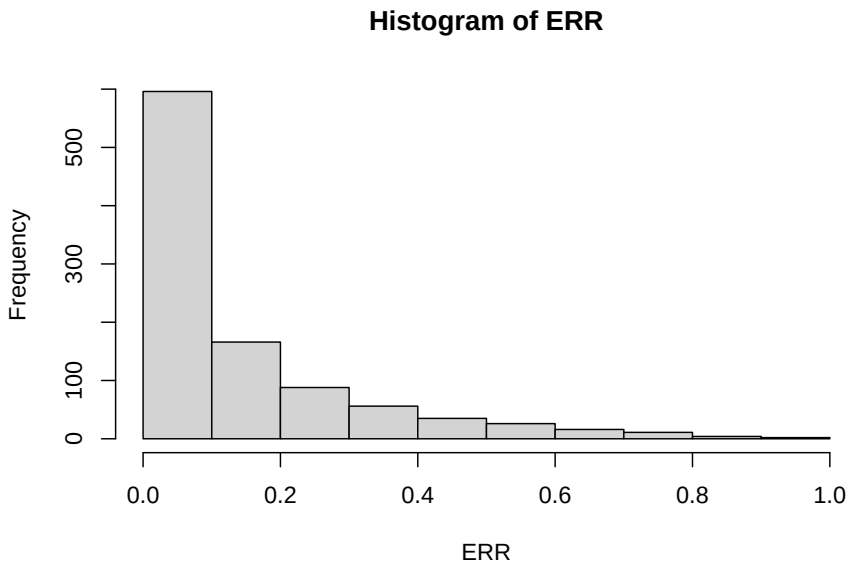
For this we sample from the normal posterior distribution, conditioned on the rounded values to match the observed. For each sample of the posterior, the rounded values are conditioned to match the observed data. The mean difference of $(z_1 - z_2)^2$ is 0.14, with variance of 0.03.

```
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:0.136664438057776
```

```
hist(ERR)
```



Question 6

Consider data y1,...,yn modeled as independent Bin(N,θ) with both N and θ unknown.

First we examine that $p(\lambda, \theta) \propto 1/\lambda$ described by Raftery (1988). We take the product of a vague prior for λ and the uniform prior for θ. let $\theta \sim U[0, 1]$ be uniform and use the uniform prior for λ as a vague prior. The motivation here

is to use a uniform vague prior on the hyperparameter. Alternatively we can use Jeffreys' Prior for Pois(λ) $= 1/\sqrt{\lambda}$ 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 Gamma(0,0) vague prior initially and independent of θ. where $\theta \sim U(0,1)$ as a uniform probability. Then $P(N, \theta) = P(N|\theta)p(\theta)$. Note this is an improprer 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 $N \sim Pois(\lambda)$.

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

So we can let N follow vague prior from the Gamma(0,0) distribution, or call N a uniform on (0,N) and independent of λ to derive p(N,θ).

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

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

Then we integrate with respect to θ 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.

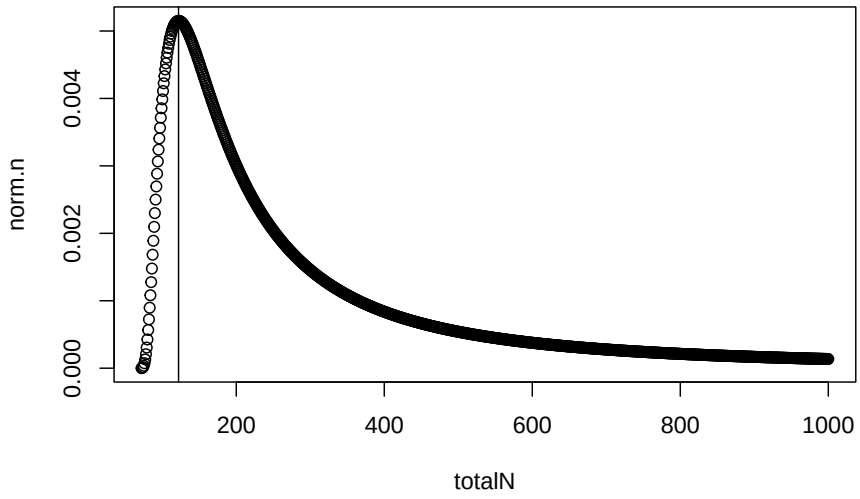
$$\begin{aligned} p(N|y) &= \int (1/N) \prod \binom{N}{y_i} \theta^S (1-\theta)^{nN-S} d\theta, S = \sum y_i \\ &= (1/N) \prod \binom{N}{y_i} \int \theta^S (1-\theta)^{nN-S} d\theta \\ &= (1/N) \prod \binom{N}{y_i} Beta(S+1, nN - S + 1) \end{aligned}$$

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))]
```

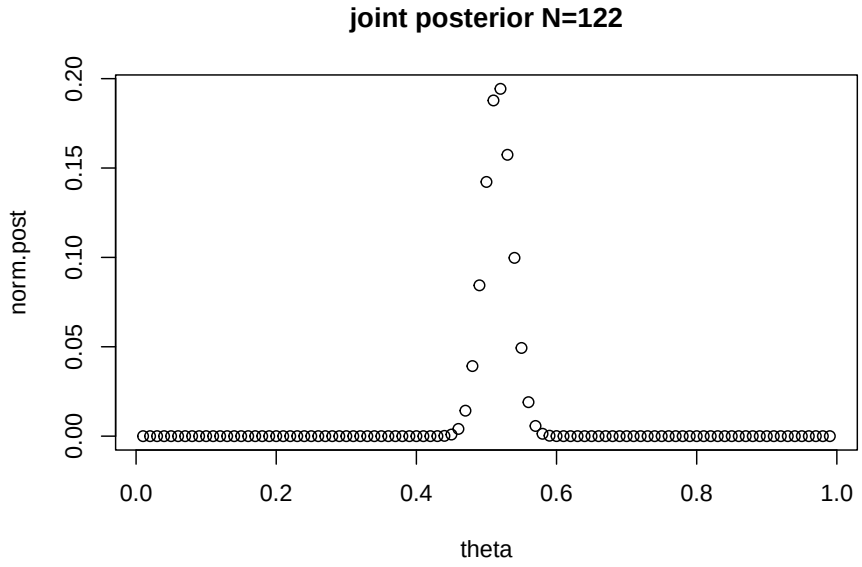
```
## [1] 122
```

```
  ##
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 $\mathrm{P}(N > 100) \approx 0.95$. 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 μ for N only gives the probability for N as a fixed realization, however we want to study the uncertainity over all possible N values.

Question 7

for 2 independent poisson distributions for x and y with λ_1, λ_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, $\lambda_1/(\lambda_1 + \lambda_2)$).

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

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, θ_y) independent of z$\sim Binomial(N_z, \theta_z)$.

We assume θ_y, θ_z are independent, so the posteriors are independent Beta models.

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

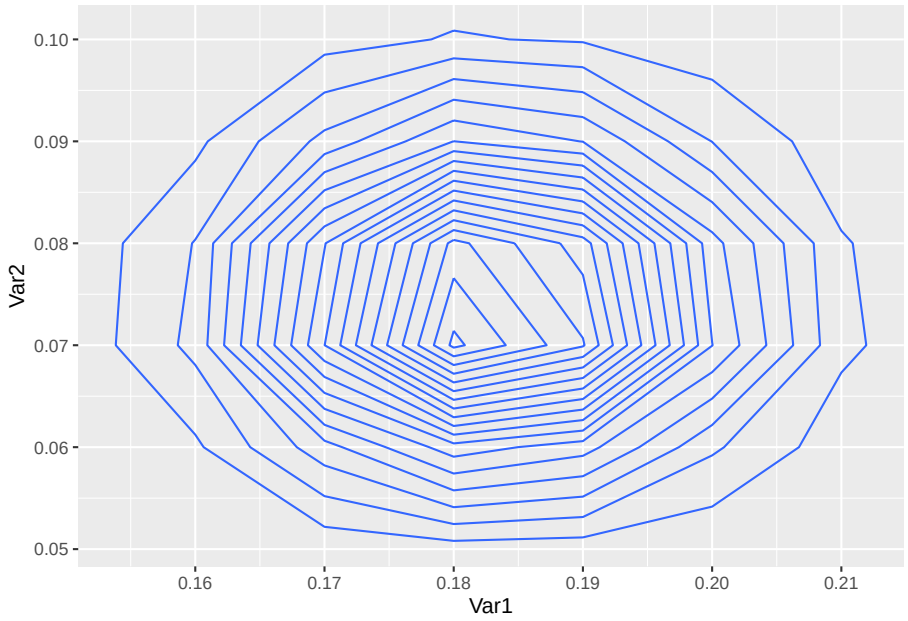
- (b) the prior uses Jeffreys' Beta(1/2,1/2) for each parameter independently.

```
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 sampline independently. Using a grid a pproach we see the contour map centered around 0.18 and 0.09 which is equivalent to the MLEs.

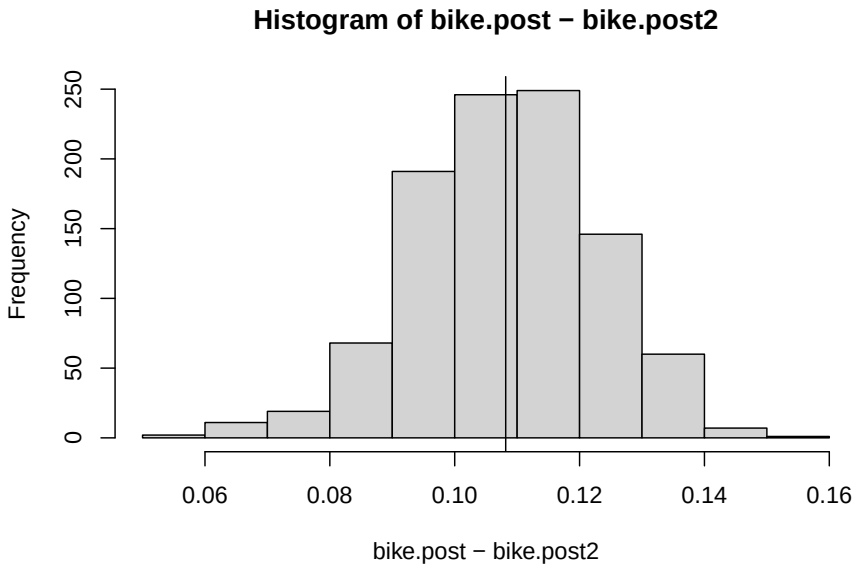
```
bike.post<-rbeta(1000, (1/2 + sum(rez.lane.bike)), (1/2+sum(rez.lane.other)))
bike.post2<-rbeta(1000, (1/2 + sum(rez.nolane.bike)), (1/2+sum(rez.nolane.other)))



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.
```

question 9

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

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

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

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

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

here we see that σ_n^2, μ_n are sufficient statistics for μ, σ^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.

$$
\begin{aligned}
f(z,w) &= f(u)f(v)|J| \\
&= \frac{1}{\Gamma(p/2)2^{p/2}}(pzw)^{p/2-1}e^{-pzw/2}/frac1\Gamma(q/2)2^{q/2}(wq)^{q/2-1}e^{-wq/2}|pwq| \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)2^{p+q/2}\Gamma(q/2)}(p)^{p/2}q^{q/2}w^{(p+q/2-1)}e^{-w/2(pz+q/2)} \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)2^{p+q/2}\Gamma(q/2)}(p)^{p/2}q^{q/2}[\frac{\Gamma(p+q/2)}{(1/2(pz+q))^{p+q/2}}] \\
&= \frac{z^{p/2-1}}{\Gamma(p/2)\Gamma(q/2)}(p)^{p/2}q^{q/2}[\frac{\Gamma(p+q/2)}{(q^{p+q/2}(pz/q+1))^{p+q/2}}] \\
&= \frac{z^{p/2-1}\Gamma(p+q/2)}{\Gamma(p/2)\Gamma(q/2)}(p/q)^{p/2}\frac{1}{(pz/q+1)^{p+q/2}}
\end{aligned}
$$

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 $=\frac{(n-1)S^2}{\chi^2_{n-1}}$. 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

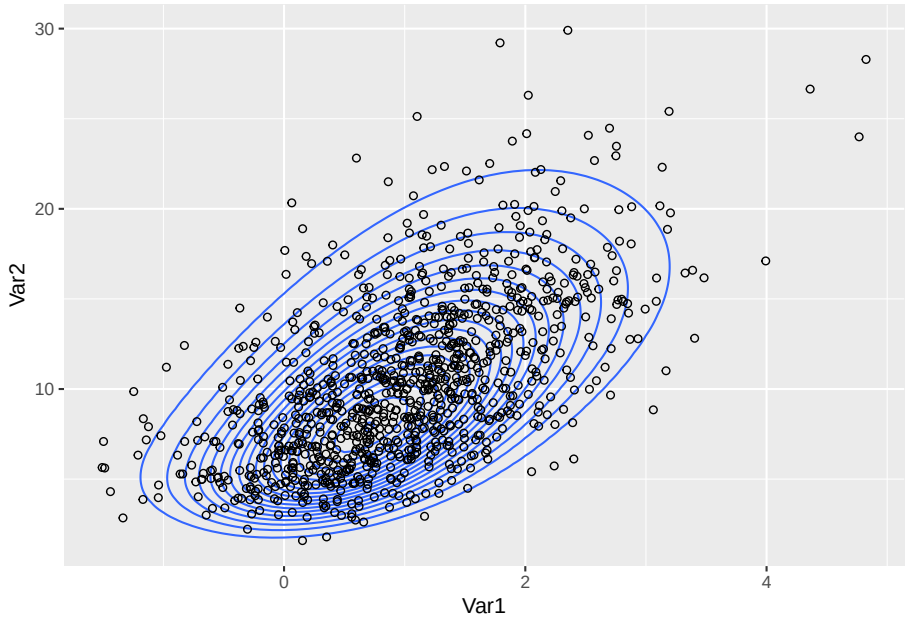
the model is defined as $p(\alpha,\beta,y,n,x) \propto p(\alpha,\beta)\prod_{i=1}^{4}p(y_i|\alpha,\beta,n_i,x_i)$, where (α,β) follows a joint normal prior with $\alpha \sim N(0,2^2)$ and $\beta \sim N(10,10^2)$ and $\text{corr}(\alpha,\beta)=0.5$.

- (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){
```

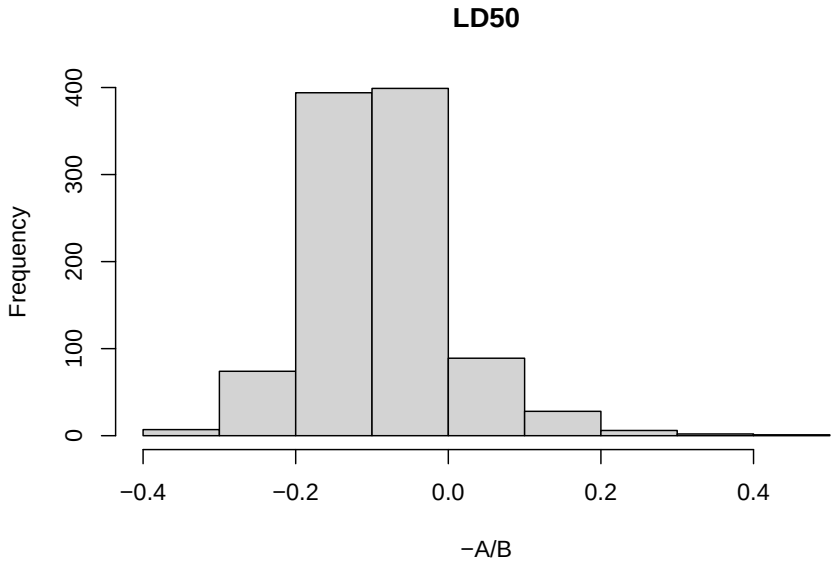
```
  return( log(theta/(1-theta)))
}

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

## alpha, beta come from multivariate normal.
 ## \prod_(i=1)^5 p(y|theta) ~ \theta^y (1-theta)^n-y ## 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 θ.

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

$$logp(\theta|y) = logp(\widehat{\theta|y}) + (1/2)(\theta - \widehat{\theta}))^T[d^2/d\theta^2 logp(\theta|y)]_{\theta=\hat{\theta}} + ... \tag{4.1}$$

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.

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

Now to show the normal approximation

$$\begin{aligned} logp(\theta|y) &= logp(\widehat{\theta|y}) + (1/2)(\theta - \hat{\theta}))^T[d^2/d\theta^2 logp(\theta|y)]_{\theta=\hat{\theta}} + ... \\ &= logp(\widehat{\theta|y}) + (1/2)(\theta - \hat{\theta}))^T(d^2/d\theta^2 logp(\hat{\theta}) + \sum_{i=1}^{n} d^2/d\theta^2 logp(y_i|\theta)_{\theta=\hat{\theta}}) \end{aligned}$$

$$logp(\theta|y) - logp(\widehat{\theta|y}) = +(1/2)(\theta - \hat{\theta}))^T(c - n * J(\theta_0))$$

$$p(\theta|y) - p(\hat{\theta}|y) \propto exp(-\frac{1}{2(nJ(\theta_0))^{-1}}(\theta - \hat{\theta})^T)$$

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

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

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

Where I(θ) is the *observed information*

$$I(\theta) = -d^2/d\theta^2 logp(\theta|y) \tag{4.4}$$

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(θ), 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, θ_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 θ_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 θ, and that θ_0 is not a boundary point), as $n \rightarrow \infty$, the posterior distribution of θ approaches normality with mean θ_0 and variance $(nJ(\theta_0))^{-1}.WhereJ(\ )\$$ is the *FIsher information* in context fo Jeffreys' Prior.

The posterior mode is consistent for θ_0, as $n \rightarrow \infty$, so the mass of the posterior $p(\theta|y)$ becomes concentrated in small neighborhoods of θ_0 and the distance of $|\theta - \theta_0| \rightarrow 0$.

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

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

This is considered a function of θ, 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 θ.

In summary, as the limit of n, in the context of a family of models posterior mode, $\hat{\theta}$, approaches the truth θ_0, and the curvature approaches $nJ(\hat{\theta})$ or nJ(θ_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 θ arrives, the posterior distribution converges to the true value of θ.

Large sample correspondence

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

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

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 $\hat{\theta} \to \theta_0$ we can write the approximation using $\mathrm{I}(\theta_0)$. 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 θ, as $n \to \infty$ then

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

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 θ will cover the true value 95% of the time under repeated sampling with any fixed true θ.

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 θ is consistent if its sampling distribution converges to a point mass at θ_0 for $n \to \infty$.

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 θ with lower mean squared error, that is if the expression $\mathrm{E}((\hat{\theta} - \theta_0)^2|\theta_0)$ 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 θ_0 at least $100(1-\alpha)\%$ of the time, then C(y) is called the $100(1-\alpha)\%$ *confidence region* for parameter θ. We saw previously that asymptotically a $100(1-\alpha)\%$ central posterior interval for θ has the property that , in repeated samples of y, $100(1-\alpha)\%$ of the intervals include θ_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))
```

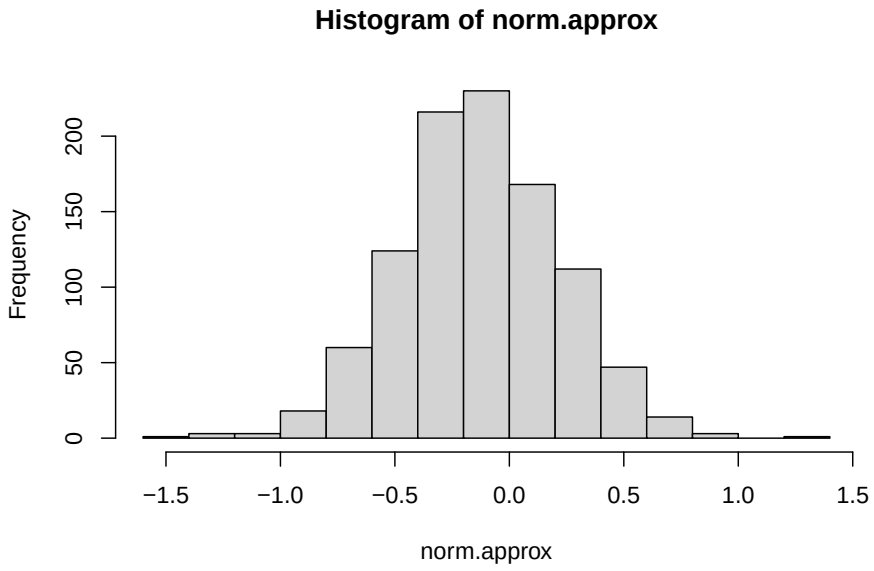
```
## $minimum
## [1] -0.1376593
##
## $objective
## [1] 11.17292
```

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

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

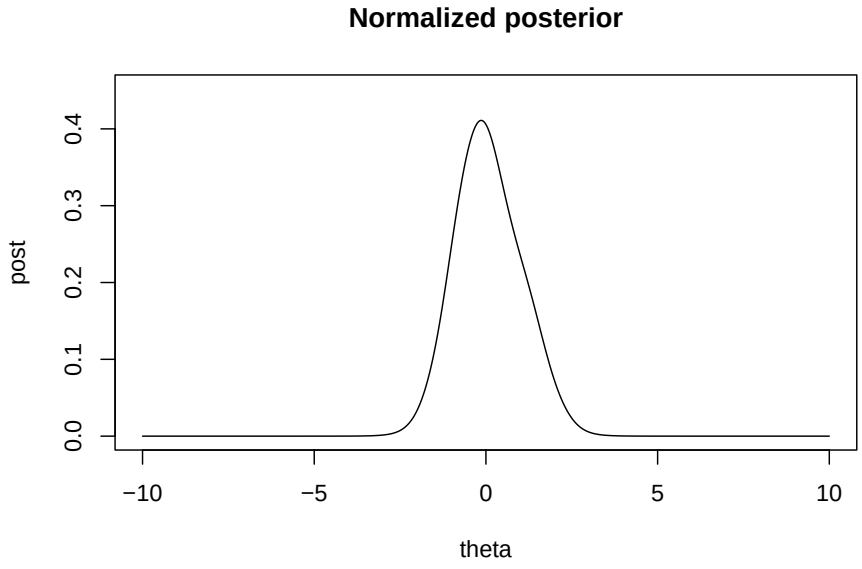
$p(\theta|y) \approx N(\hat{\theta}, \sigma^2 = 1/(n * I(\theta)))$

```
# 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)
```

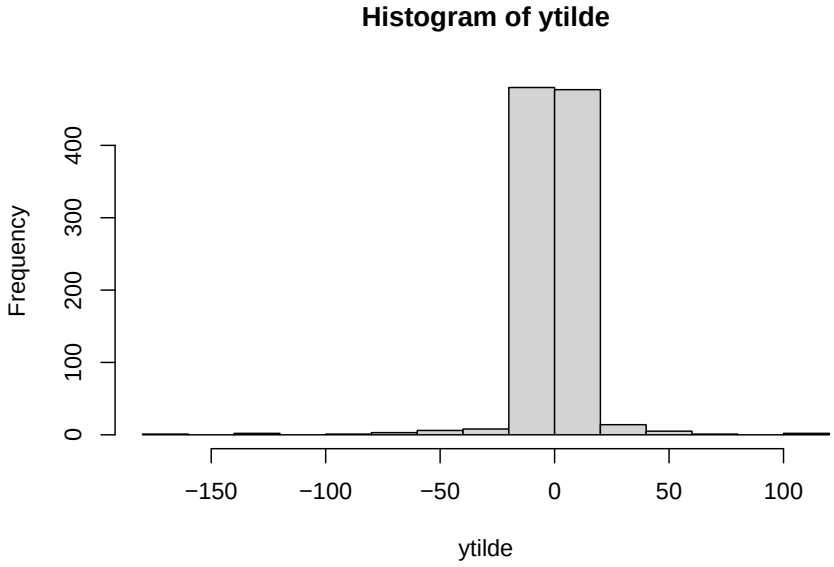


compare with 2.11 grid approach using the Cauchy distribution, we see that normal approximation has much wider tails compared to the exact distribution, this is because of the small sample size.

```
  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| 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)
```



Exercise 2

To show the analytic information matrix of the bioassay example.

The logistic function has a derivative in the form

$$\partial/\partial\alpha logit^{-1}(\alpha+\beta*x_i) = \frac{exp(\alpha+\beta*x_i)}{(1+exp(\alpha+\beta*x_i))^2} = \frac{logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)} \quad (4.8)$$

$$\partial/\partial\beta logit^{-1}(\alpha+\beta*x_i) = \frac{exp(\alpha+\beta*x_i)x_i}{(1+exp(\alpha+\beta*x_i))^2} = \frac{logit^{-1}(\alpha+\beta*x_i)x_i}{1+exp(\alpha+\beta*x_i)} \quad (4.9)$$

Also by algebra

$$\frac{1}{1-logit^{-1}(\alpha+\beta*x_i)} = 1+exp(\alpha+\beta*x_i) \quad (4.10)$$

Normal approximation using the information

$$\begin{aligned}
p(\alpha,\beta|y_i,n_i,x_i) &\propto p(\alpha,\beta)p(y_i|\alpha,\beta,n_i,x_i)\\
&\propto [logit^{-1}(\alpha+\beta*x_i)]^{y_i}[1-logit^{-1}(\alpha+\beta*x_i)]^{n_i-y_y}\\
lp(\alpha,\beta|y_i,n_i,x_i) &\propto y_i log([logit^{-1}(\alpha+\beta*x_i)]) + (n_i-y_i)log([1-logit^{-1}(\alpha+\beta*x_i)])
\end{aligned}$$

$$\begin{aligned}
\frac{\partial lp}{\partial\alpha} &\propto y_i\frac{1}{logit^{-1}(\alpha+\beta*x_i)}\frac{logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)} - \frac{n_i-y_i}{1-logit^{-1}(\alpha+\beta*x_i)}\frac{logit^{-1}(\alpha+\beta*x_i)x_i}{1+exp(\alpha+\beta*x_i)}\\
&= \frac{y_i}{1+exp(\alpha+\beta*x_i)} - (n_i-y_i)logit^{-1}(\alpha+\beta*x_i)\\
&= y_i(1-logit^{-1}(\alpha+\beta*x_i) - (n_i-y_i)logit^{-1}(\alpha+\beta*x_i)\\
&= y_i - n_i*logit^{-1}(\alpha+\beta*x_i)
\end{aligned}$$

note that

$$\frac{1}{1+exp(\alpha+\beta*x_i)} = 1-logit^{-1}(\alpha+\beta*x_i) \quad (4.11)$$

$$\begin{aligned}
\frac{\partial^2 lp}{\partial\alpha^2} &\propto \frac{-n_i logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)}\\
&= logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))n_i\\
\frac{\partial^2 lp}{\partial\alpha\partial\beta} &\propto \frac{-x_i n_i*logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)}\\
&= -x_i*n_i*logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))
\end{aligned}$$

$$\begin{aligned}
\frac{\partial lp}{\partial\beta} &\propto \frac{y_i}{1+exp(\alpha+\beta*x_i)}\frac{logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)} + \frac{n_i-y_i}{1-logit^{-1}(\alpha+\beta*x_i)}\frac{-x_i*logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)}\\
&\propto y_i*x_i(1-logit^{-1}(\alpha+\beta*x_i)) - x_i(n_i-y_i)logit^{-1}(\alpha+\beta*x_i)
\end{aligned}$$

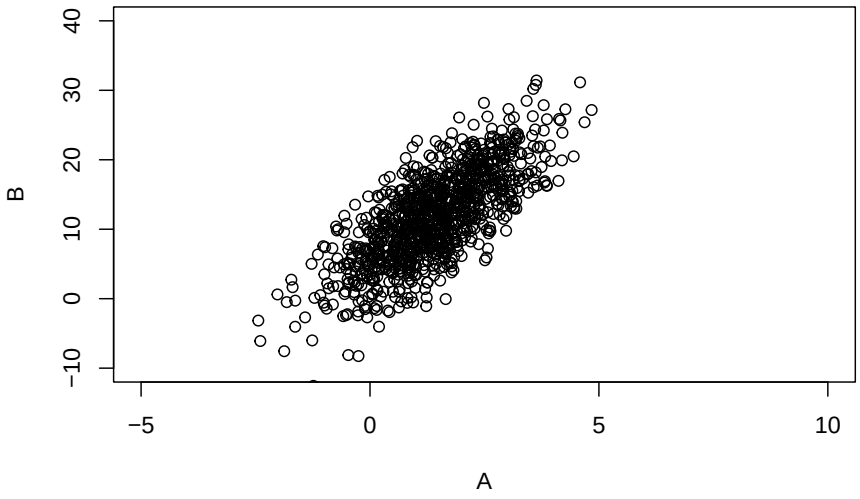
$$\begin{aligned}\frac{\partial^2 lp}{\partial \beta^2} &\propto x_i y_i(\frac{-x_i logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)}) - x_i(n_i-y_i)\frac{logit^{-1}(\alpha+\beta*x_i)}{1+exp(\alpha+\beta*x_i)} \\ &= logit^{-1}(\alpha+\beta*x_i)x_i^2 - y_i(1-logit^{-1}(\alpha+\beta*x_i)) - (n_i-y_i)(1-logit^{-1}(\alpha+\beta*x_i)) \\ &= logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))(-n_i x_i^2)\end{aligned}$$

Now the information matrix is evaluated at the posterior mode $\hat{\alpha}$

$$I(\alpha,\beta) = \begin{bmatrix} \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))n_i & \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta \\ \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))x_i n_i & \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta \end{bmatrix}$$

$$p(\theta|y) \approx N(\hat{\alpha},\hat{\beta},I(\alpha,\beta)^{-1}) \tag{4.12}$$

```
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))

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

### normal approximation
 alpha_hat<-1.353
 beta_hat<-11.634


I_hat_theta<-matrix(c(sum(inv.logit(alpha_hat+beta_hat*assay$x)*(1-inv.logit(alpha_hat-
                      sum(inv.logit(alpha_hat+beta_hat*assay$x)*(1-inv.logit(alpha_hat-
                      sum(inv.logit(alpha_hat+beta_hat*assay$x)*(1-inv.logit(alpha_hat-
                      sum(inv.logit(alpha_hat+beta_hat*assay$x)*(1-inv.logit(alpha_hat-

## asymptotic variance (a,b)
invI<-matlib::inv(I_hat_theta)

norm.approx<-rmvnorm(1000,mean=c(alpha_hat,beta_hat),sigma=invI)
 colnames(norm.approx)<-c("A","B")
plot(norm.approx,xlim=c(-5,10),ylim=c(-10,40)) # figure 4.1(b)
```



Exercise 3

A reasonable estimate for the posterior mode and standard deviation as proxy to the asymptotic values will utilize a GLM, where $\theta = \frac{-\alpha}{\beta}$= -0.109 (-0.15,0.37) as the limit values.

```
 dr<-data.frame(xi=c(rep(-0.86,5),
                     rep(-0.30,5),
                     rep(-0.05,5),
                     rep(0.73,5)),
                yi=c(rep(0,5),
                     1,rep(0,4),
                     1,1,1,rep(0,2),
                     rep(1,5)))

bioassay<-glm(yi~xi,family='binomial',dr)
#summary(bioassay)
 ab<-coef(bioassay)
 ab<-cbind(ab,confint(bioassay))
```

```
## Waiting for profiling to be done...
```

```
ld50<--ab[1,]/ab[2,]
print(ld50)
```

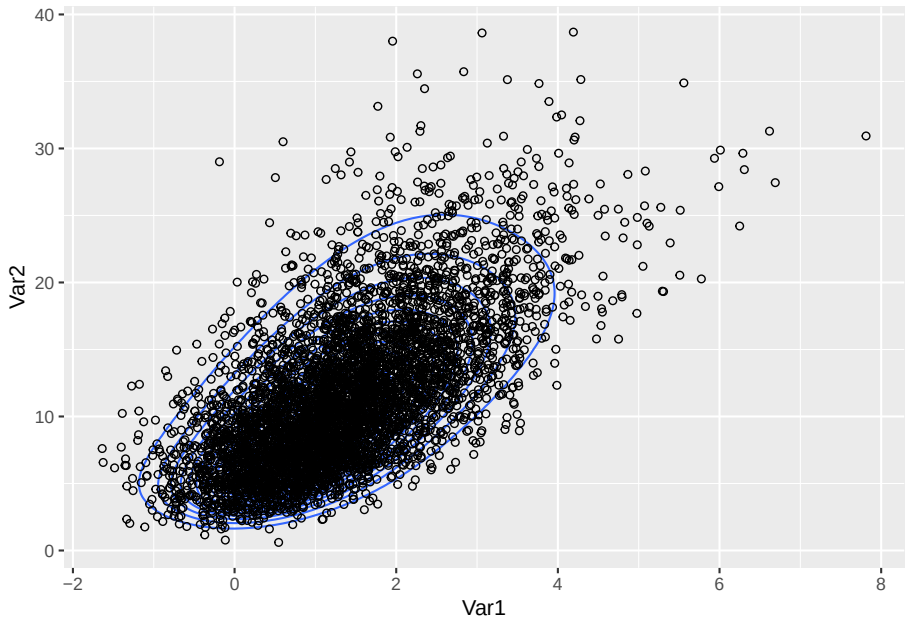
```
##         ab      2.5 %     97.5 %
## -0.1092528  0.3663445 -0.1560259
```

Using a uniform prior, we compute the Bayesian posterior for $\theta = -\alpha/\beta$ using posterior distribution for $p(\alpha, \beta|y)$ and then compute θ with interval -0.1049 (-0.27, 0.09) and the standard deviation is 0.0988. This reproduces figure 3.3 from section 3.7

```
## compare with the Bayesian approximation using 3.16 from section 3.7
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)))  )
}

## equation 3.16 using the uniform
## alpha, beta come from uniform distribution
## \prod_(i=1)^5 p(y|theta) ~ \theta^y (1-theta)^n-y ## 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 pri
## uniform prior p(a,b) ~ 1
joint.prior<-function(alpha,beta){
```

```
  prob.ab<-1
  return(prob.ab)
 }

 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 1:5000){
    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.
    Ajit<-A+runif(length(A),min=0,max=0.1)
    Bjit<-B+runif(length(B),min=0,max=0.1)
   ab.post<-data.frame(A=Ajit,B=Bjit)
 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)
```




```
## posterior means (1.35,11.6)
message('posterior mean for A :',round(mean(A),3),' B ',round(mean(B),3))
```

```
## posterior mean for A :1.293 B 11.482
```

```
quantile(-A/B,c(0.025,0.975),na.rm=TRUE)
```

```
##       2.5%      97.5%
## -0.2771114  0.1000510
```

```
message('posterio std.dev -a/b: ',round(sd(-A/B),3))
```

```
## posterio std.dev -a/b: 0.095
```

Q3 Delta method approximation

Using the *delta-method* and the correlation matrix computed from Q2. The delta method can be used to derive a normal approximation.

$$\sqrt{n}(\hat{\alpha}/\hat{\beta} - \alpha/\beta) \rightarrow N(0, \nabla\frac{\alpha}{\beta}^T \Sigma \nabla\frac{\alpha}{\beta}) \tag{4.13}$$

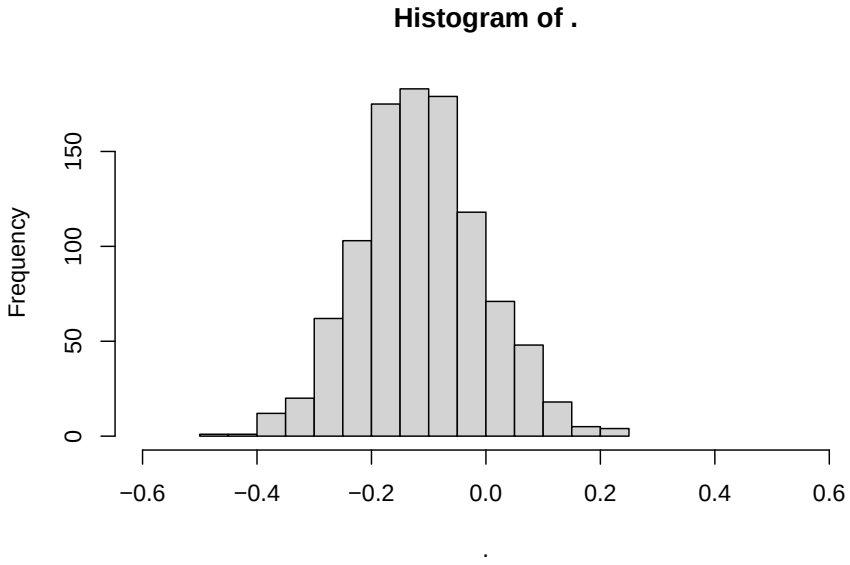
$$\begin{aligned}
V(\alpha/\beta) &= (\frac{1}{\beta}, \frac{-\alpha}{\beta^2}) \begin{bmatrix} \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))n_i & \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta* \\ \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))x_i n_i & \sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta* \end{bmatrix} \\
&= \begin{pmatrix} \sum_i \frac{1}{\beta} logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))n_i - \frac{-\alpha}{\beta^2}\sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i)) \\ \sum_i \frac{1}{\beta} logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))x_i n_i - \frac{-\alpha}{\beta^2}\sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i) \end{pmatrix} \\
&= \sum_i \frac{1}{\beta^2} logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))n_i - \frac{-2\alpha}{\beta^3}\sum_i logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i) \\
&= \sum_i \frac{n_i * logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))}{\beta^2}(\frac{\alpha^2 x_i^2}{\beta^2} - \frac{2\alpha}{\beta}x_i + 1) \\
&= \sum_i \frac{n_i * logit^{-1}(\alpha+\beta*x_i)(1-logit^{-1}(\alpha+\beta*x_i))}{\beta^2}(\frac{\alpha x_i}{\beta} - 1)^2_{\alpha,\beta=\hat{\alpha},\hat{\beta}}
\end{aligned}$$

Using the derivation of the delta method shown above, we input the posterior means into the equation and derive the variance to the normal approximation. the posterior mean is approximately 0.109 with $\sigma = 0.107$. The solution from the author derived the $\sigma = 0.096$ so we are very close.

```
theor.var<-sum((assay$n/beta_hat^2)*inv.logit(alpha_hat+beta_hat*assay$x)*(1-inv.logit(alpha_hat
 message('the theoretical std dev:', round(sqrt(theor.var),3))
```

```
## the theoretical std dev:0.109
```

```
 emp.var<-var(-A/B)
ld50_normApprox<- rnorm(1000,mean(-A/B),sd=sqrt(theor.var))
 ld50_normApprox%>%hist(xlim=c(-0.6,0.6))
```




```
ld50_hat<-mean(ld50_normApprox)  ## poster mean (should use the mode but close enough

message('normal approx posterior mean ', round(mean(ld50_normApprox),3),' std. dev ',r
```

```
## normal approx posterior mean -0.113 std. dev 0.108
```

4.4.1 Question 4.2 using the posterior sampling via grid approach

We derive the posterior marginal using a grid approach and the fact that the jacobian $= |\nu|$ and we let $\theta = -\alpha/\beta$ and $\nu = \beta$ to solve for the change of variables.

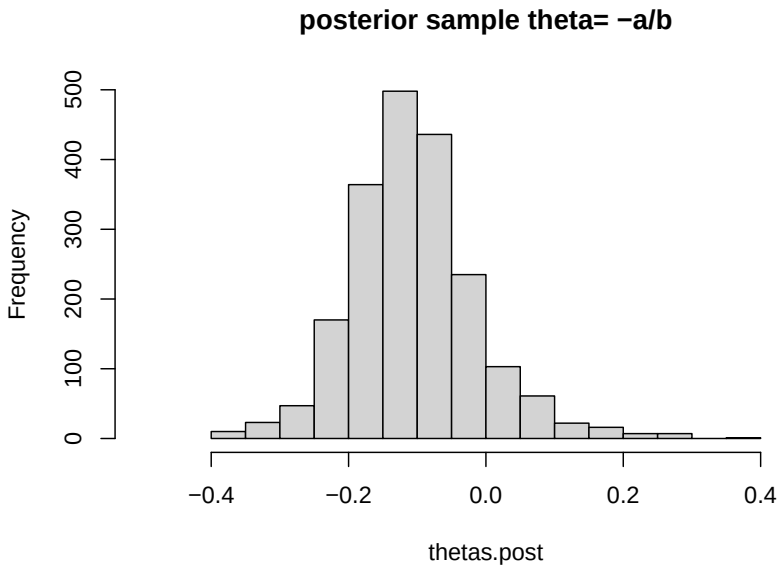
$$p(\theta|y) = \int p(\theta,\nu|y)d\nu = \int p(\alpha,\beta|y)|\nu|d\nu = \int \prod_i logit^{-1}(-\theta*\nu+\nu*x_i)(1-logit^{-1}(-\theta*\nu+\nu*x_i))|\nu|d\nu$$

We use the grid approach described in section 3.7 to sample from the posterior of $p(\theta,\nu)$ where θ =LD50 term.

The posterior mean is -0.1055, with marginal posterior standard deviation of 0.092.

```
 ## using change of variable equation -\theta*nu +nu*xi in the likelihood
 logit.likelihood.variableChange<-function(alpha,beta,x,y,n){
   theta_approx<- -alpha*beta+beta*x
   bin.like<-(inv.logit(theta_approx))^y*(1-inv.logit(theta_approx))^(n-y)
   return(bin.like)
}

### re-doing based on sampling
 posterior_density_Jacobian<-function(alpha,beta,assay,jacobian){
   joint_density<-joint.prior(alpha,beta)
   probs<-NULL
   for(k in 1:nrow(assay)){
     p<-logit.likelihood.variableChange(alpha,beta,assay$x[k],assay$y[k],assay$n[k])
     probs<-c(probs,p)
   }
   totalLikelihood<-prod(probs)*abs(jacobian)
   return( joint_density*totalLikelihood)
 }

theta_seq= seq(-0.4,0.4,by=0.001) ## alpha
### numerical integration
 nu_seq<- seq(-10,70,by=0.05) ## beta

  thetaNu<- expand.grid(theta_seq, nu_seq)
  posts<-NULL
for(i in 1:nrow(thetaNu)){
  posts[i]<-posterior_density_Jacobian(thetaNu[i,"Var1"],thetaNu[i,"Var2"],assay,jacobian = theta
}

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

 ## grid sampling procedure
 ## marginal post of alpha by summing over nu terms
  marginal.post_theta<- thetaNu%>%group_by(Var1)%>%summarise(p=sum(joint.prob))%>%data.frame
    thetas.post<-nu.post<-NULL
 for(s in 1:2000){
    a_s<-sample(marginal.post_theta$Var1,1,prob=marginal.post_theta$p)
    p.a_s<-marginal.post_theta[which(marginal.post_theta$Var1==a_s),'p']
    #
    marginal.post_b<-thetaNu[which(thetaNu$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)
    thetas.post<-c(thetas.post,a_s)
    nu.post<-c(nu.post,b_s)
```

```
}
 hist(thetas.post,xlim=c(-0.5,0.5),main='posterior sample theta= -a/b')
```




```
message('sample posterior mean: ', mean(thetas.post))
```

```
## sample posterior mean: -0.1048945
```

```
message('sample posterior variance: ', round(var(thetas.post),3))
```

```
## sample posterior variance: 0.009
```

```
message('sample posterior std.dev: ', round(sd(thetas.post),3))
```

```
## sample posterior std.dev: 0.092
```

4.4.2 Taylor series approximation of the information matrix

Taylor series expansion to approximate $d^2 logp(\hat{\theta}|y)/d\theta^2 \approx \frac{f(\hat{\theta}+h)+f(\hat{\theta}-h)-2f(\hat{\theta})}{h^2}$ centered on the posterior mean $\hat{\theta}$ within a neighborhood h=0.002.

$$f(a+h) \approx f(a) + h'f(a) + 1/2h^2 f''(a)$$
$$f(a-h) \approx f(a) - h'f(a) + 1/2h^2 f''(a)$$
$$\rightarrow f''(a) \approx \frac{f(a+h) + f(a-h) - 2f(a)}{h^2}$$

Using a numerical approximation of $p(\theta|y)$ from the grid approach had $\sigma = 0.064$ which is lower than we would expect from the theoretical delta method.

```
library(DescTools)
theta_approx<- mean(thetas.post)
h=0.002
a<-marginal.post_theta[which(abs(marginal.post_theta$Var1-theta_approx)==min(abs(marginal.post_t
ah<-marginal.post_theta[which(abs(marginal.post_theta$Var1-theta_approx-h)==min(abs(marginal.pos
adh<-marginal.post_theta[which(abs(marginal.post_theta$Var1-theta_approx+h)==min(abs(marginal.po

tapprox<-(log(ah$p)+log(adh$p)-2*log(a$p))/h^2

std.approx<-(-tapprox)^(-0.5)

message('taylor approximation of std.dev: ',round(std.approx,3))
```

```
## taylor approximation of std.dev: 0.064
```

If we would not able to solve for the theoretical value from the *delta-method* in closed form we can use the standard deviation from marginal posterior sampled from the grid approach.

```
theoretical.approx<-(dnorm(theta_approx+h,mean=theta_approx,sd=sd(thetas.post),log=TRUE)+  dnorm(

std.approx2<-(-theoretical.approx)^(-0.5)

message('taylor approximation of std.dev using posterior sample variance: ',round(std.approx2,3)
```

```
## taylor approximation of std.dev using posterior sample variance: 0.092
```

Question 4

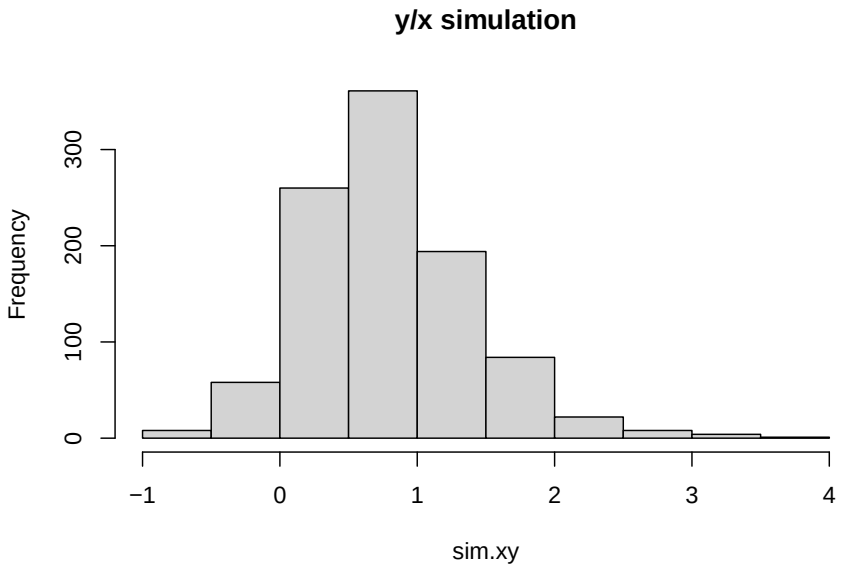
For $n \rightarrow \infty p(\theta|y)$ approaches normality, and any 1-1 transformation approaches normality $\phi = f(\theta), p(\phi|y)$ is also asymptotically normal. A nonlinear transformation we can use the delta method such that $g'(\theta) \neq 0$ and given that $p(\theta|y)$ is normal. Then for any function, the Taylor series in the neighborhood of the true parameter can converge to a normal approximation.

This is given by Slutsky's theorem to show that a sequence of random variables converges in distribution, and also if $lim_{n\to\infty}V(S_n^2) \to 0$ where the sample variance goes to 0 in the limit. Hence the normal approximation in the limit has 0 variance, and converges almost certaintly to the true distribution.

Question 5

Using a simulation, we simply take the ratio of 2 random normal variables from their samples

```
x= rnorm(1000,mean=4,sd=1)
y= rnorm(1000,mean=3,sd=2)
sim.xy<-y/x

hist(sim.xy,main='y/x simulation')
```




```
message('simulated mean/std.dev: ', round(mean(sim.xy),3),' ',round(sd(sim.xy),3))
```

```
## simulated mean/std.dev: 0.791 0.611
```

```
print(var(sim.xy))
```

```
## [1] 0.3727425
```

```
## theoretical mean  E(Y/X)=E(Y)E(1/X) = 3/4

## taylor series for Var(Y/X) =
```

-(b) without using a simulation we use the delta method

$$\begin{aligned} V(Y/X) &= (\frac{-Y}{X^2}, \frac{1}{X}) \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \begin{pmatrix} -Y/X^2 \\ 1/X \end{pmatrix} \\ &= (-Y/X^2 \quad 4/X) \begin{pmatrix} -Y/X^2 \\ 1/X \end{pmatrix} \\ &= Y^2/X^4 + 4/X^2|_{X,Y=\hat{X},\hat{Y}} \\ &= \frac{\mu_x^2\sigma_y^2 + \sigma_x^2\mu_y^2}{\mu_y^4} \\ &= \frac{3^2 + 4^3}{4^4} = 0.28516 \end{aligned}$$

```
 xy<-rnorm(1000,mean=3/4,sd=sqrt( (9+4^3)/4^4))
```

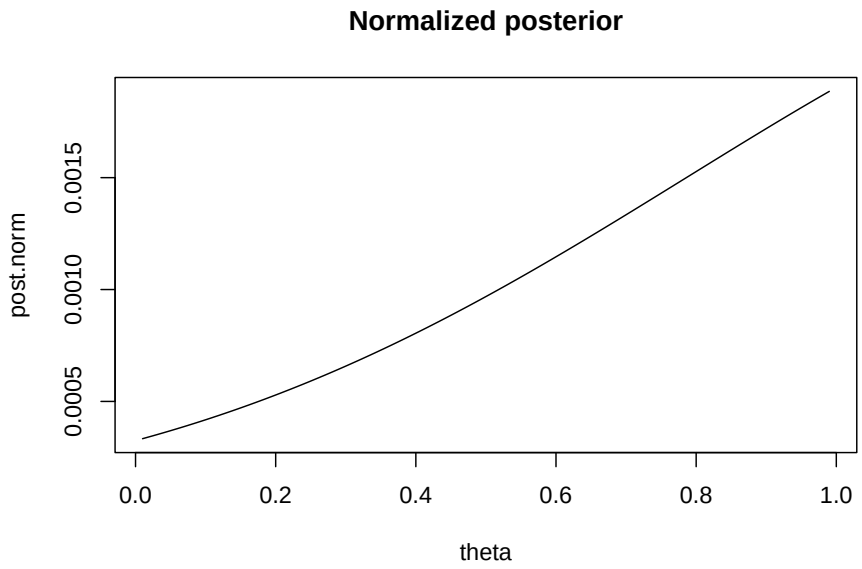
-(c) we must assume that $x \neq 0$, for the posterior. note that for the ratio of 2 standard normal distributions

Question 9

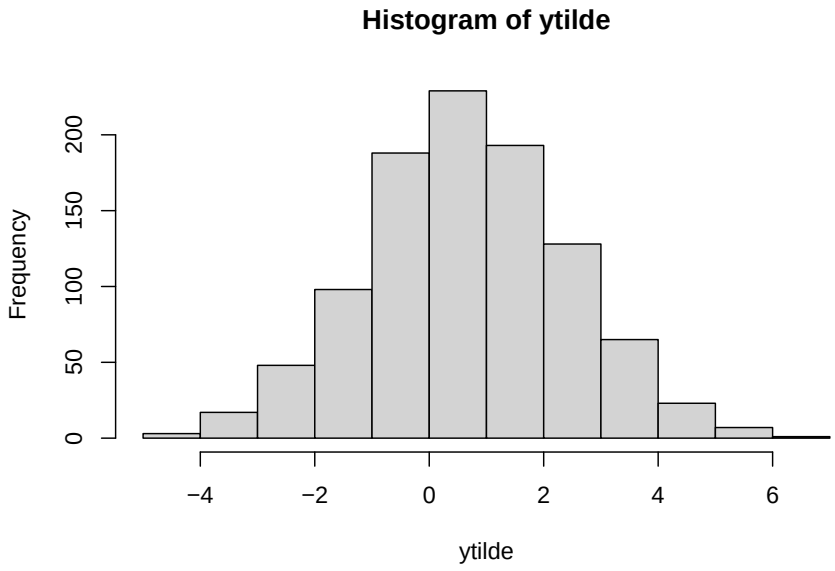
- we set the known variance to 3 and select 5 observed values of y $\sim N(\mu, \sigma)$ where $\sigma = 3$ is known and $\mu = 0.720925$ is the population parameter that we make inferences for.
- the MLE is the sample mean of 1.56, with MSE of 0.71 with observed sample size of 5. Note that if the sample size is larger, n=1000, then the MLE is more powerful.
- the posterior mean is used to compute the MSE which is 0.001.
- note that if $\sigma = 0.01$ the MLE is far more powerful.

```
 set.seed(8182022)
true_mu<-0.720925
true_var<-3

#observed_y<-sapply(true_var,function(x) rnorm(1,true_mu,sd=sqrt(x)))
#Y<-data.frame(Y=y,var=true_var)

##observed point
observed_y<-rnorm(5,mean=true_mu,sd=sqrt(true_var))
```

```
MLE = mean(observed_y) ## point estimate for n=1
# MSE for 1 data point is just itself
 mse1=(MLE-true_mu)^2


 ## posterior mean
 ## known variance, uniform prior for theta
  # theta| y ~ p(theta)*l(y|theta)
 theta<-seq(0.01,0.99,by=0.001)
 post.norm<-NULL
 for(i in theta){
 post_theta<-dunif(i)*prod(dnorm(observed_y,mean=i,sd=sqrt(true_var)))
 post.norm<-c(post.norm,post_theta)
 }
  post.norm<-post.norm/sum(post.norm)
 ## need to sample from the posterior

plot(theta,post.norm,type='l',main='Normalized posterior')
```




```
 samps<-(sample(theta,1000,prob=post.norm,replace=T))

ytilde<-rnorm(1000,mean=samps,sd=sqrt(true_var))
 hist(ytilde)
```



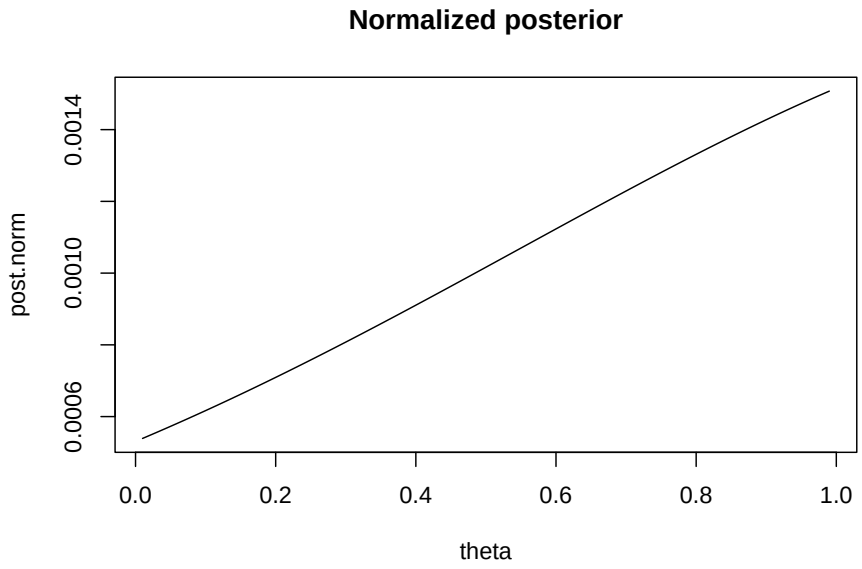

```
mse2<-(mean(ytilde)-true_mu)^2


message('MSE by MLE: ',round(mse1,4),' MSE by Bayesian:',round(mse2,4))
```

```
## MSE by MLE: 0.7051 MSE by Bayesian:0.0013
```

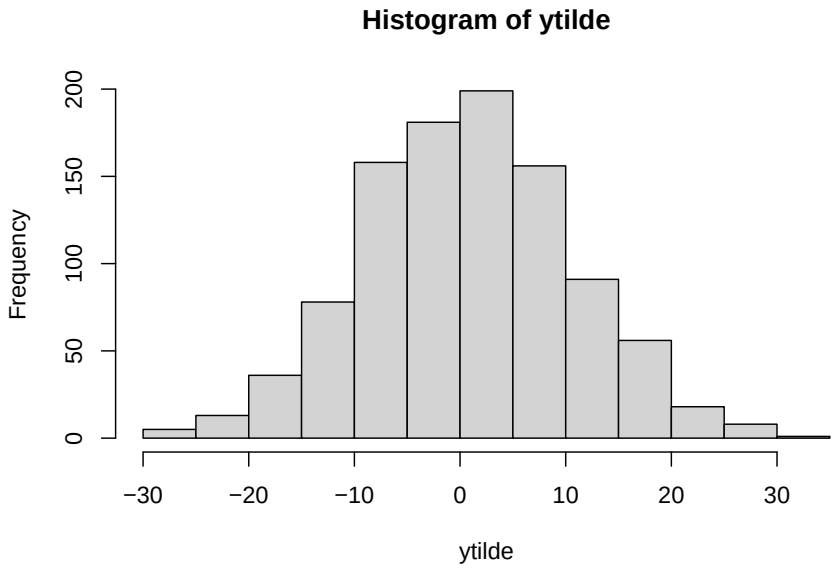
- note that if we increase the sample size to n=100, and keep the variance low, $\sigma = 3$, then the MLE has smaller MSE. if we increase the sample size to 100, but increase the $\sigma = 100$, then the Bayesian estimate has smaller MSE.

```
 set.seed(8182022)
true_mu<-0.720925
true_var<-100

#observed_y<-sapply(true_var,function(x) rnorm(1,true_mu,sd=sqrt(x)))
#Y<-data.frame(Y=y,var=true_var)

##observed point
observed_y<-rnorm(100,mean=true_mu,sd=sqrt(true_var))
MLE = mean(observed_y) ## point estimate for n=1
# MSE for 1 data point is just itself
```

```
 mse1=(MLE-true_mu)^2


 ## posterior mean
 ## known variance, uniform prior for theta
  # theta| y ~ p(theta)*l(y|theta)
 theta<-seq(0.01,0.99,by=0.001)
 post.norm<-NULL
 for(i in theta){
 post_theta<-dunif(i)*prod(dnorm(observed_y,mean=i,sd=sqrt(true_var)))
 post.norm<-c(post.norm,post_theta)
 }
  post.norm<-post.norm/sum(post.norm)
 ## need to sample from the posterior

plot(theta,post.norm,type='l',main='Normalized posterior')
```




```
 samps<-(sample(theta,1000,prob=post.norm,replace=T))

ytilde<-rnorm(1000,mean=samps,sd=sqrt(true_var))
 hist(ytilde)
```




```
mse2<-(mean(ytilde)-true_mu)^2


message('MSE by MLE: ',round(mse1,4),' MSE by Bayesian:',round(mse2,4))
```

```
## MSE by MLE: 0.6861 MSE by Bayesian:0.0136
```

Chapter 5

Hierarchical Models

Many problems involve multiple parameters that are related somehow by the structure of the problem. Each θ_j are viewed as a sample from a common *population distribution*, that is governed by *hyperparameters*.

Non-hierarchical models have inferior performance for prediction because they overfit the training data that only fit the current data well, this is because the uncertainty of the prior parameters are not included within the model. Hierarchical models avoid this problem because the model is structured around the dependence within the parameters.

5.1 Constructing a parameterized prior distribution

The example from 70 lab rat can use a fixed prior distribution under the binomial likelihood, and a beta prior. One can choose fixed estimates for the prior parameters α, β, or can use the moments to estimate the prior parameters. This is *not* a Bayesian model, and is considered an empirical Bayesian model. Given 70 historical experiments, the sample mean and sample variance are used to solve for the beta mean/variance α/β and α/β^2. Note that this is not a fully Bayesian approach.

Logic of combining information

Given 70 historical values, we combine all the parameters θ_j, for $j = 1, ..., 71$ with the addition of a *newly* performed experiment. Combining these into a data set makes sense because there are some dependency of the parameters which is reflected into a full joint prior probability model.

5.2 Exchangeability and hierarchical models

Consider a set of experiments j=1,...,J in which j has the data y_j and parameter θ_j with likelihood $p(y_j|\theta_j)$. Let $\theta_j \sim N(\mu_j, \sigma^2)$ come from a superpopulation with common fixed variance.

Exchangeability

If no information, other than the data y_j is availble to distinguish any of the θ_j from any others, and no ordering or grouping of parameters can be made, then by symmetry among the parameters called *exchangeability*. The joint distribution $p(\theta_1, ..., \theta_J)$ is invariant to permutations of the indices (1,...,J). Generally, the less we know about a problem the more we can rely on the exchangeability assumption.

The form of exchangeabile distribution each parameter θ_j as an independent sample from a prior population distribution governed by some unknown ϕ is written as

$$p(\theta|\phi) = \prod_j p(\theta_j|\phi) \tag{5.1}$$

in general ϕ is unknown so we average over the uncertainty of ϕ

$$p(\theta) = \int \prod_j p(\theta_j|\phi) * p(\phi) d\phi \tag{5.2}$$

This form is the mixture of independent identical distributions.

A related result *de Finetti's theorem* states that as the limit of $J \to \infty$ any suitably exchangeable distribution on $(\theta_1, ..., \theta_J)$ can be expressed as a mixture of identical and independent distributions. This theorem connects exchangeability to IID assumptions routinely practiced.

Exchangeability when additional information is available on the units

observations often are not exchangeable, but are partially or *conditionally* exchangeable - if observations can be grouped, we may make a hierarchical model where each group has its own sub-model, but the group properties are unknown. if we assume that the group properties are exchangeable, we can use a common prior distribution for the group properties. - if y_i has additional information x_i so that y_i are not exchangeable, but (y_i, x_i) still are exchangeable, then we can make a joint prior for (y_i, x_i) or a condtional model $y_i|x_i$.

For the rat example, if we knoew specific batches of experiments were made in different laboratories we could assume partial exchangeability and use two level hierarchical model to account for variation within each laboratory and between laboratories.

In general the usual way to model exchangeability with covariates is through conditional independence : $p(_1,..., _J) = (_j p(_j| ,x_j))p(| x)d

Objections to exchangeable models

it is natural to object to exchangeability on the grounds that the units actually differ. For the rat example, one can argue that different rats were studied at different times across different laboratories, and therefore not exchangeable. However this information does **not** invalidate exchangeability, these differences in experimental conditions imply that θ_j differ, but can still arise from a common parent distribution. If we had no information to distinguish them, we have a logical choice to assume exchangeability.

The full Bayesian treatment of the hierarchical model

let ϕ be unknown with a prior $p(\phi)$ then the joint prior is

$$p(\phi, \theta) = p(\phi)p(\theta|\phi) \tag{5.3}$$

and the joint posterior is

$$p(\phi, \theta|y) \propto p(\phi, \theta)p(y|\phi, \theta) = p(\phi, \theta)p(y|\theta) \tag{5.4}$$

The $p(y|\phi, \theta)$ depends only on θ the hyperparameters ϕ only affects y through θ. This model includes the uncertainty about ϕ in the model.

The hyperprior distribution

In order to create a joint hyperprior for (ϕ, θ) we must assign a prior distribution to ϕ. If little is known about ϕ we can choose a diffuse prior, or noninformative prior, ensuring that the posterior is proper. Further assumptions about the prior must be checked for sensitivity in the hyperprior choice.

The rat example the hyperparameters are (α, β) which determine the beta parameter θ.

Posterior predictive distribution

Hierarchical models are characterized by hyperparameters ϕ and model parameters θ. There are two posterior predictive distributions that might be an interest - (1) the distribution of future observation $\tilde{y}$ corresponding to an existing θ_j. - (2) the distribution of observations $\tilde{y}$ corresponding to future θ_j, $\tilde{\theta}$, drawn from the same superpopulation, common, population

5.3 Bayesian analysis of conjugate hierarchical models

For large number of parameters is it difficult to plot contour plots of the numerous parameters simulated from the joint posterior (ϕ, θ). For the rat example, we obtain simulations of the posterior distribution $p(\theta, \phi|y)$ for the beta-binomial model for the rat-tumor for which the **population distribution** $p(\theta|\phi)$ is conjugate to the likelihood $p(y|\theta)$. For non-conjugate models , we must use MCMC.

Analytic derivation of conditional and marginal distributions

- (1) write the joint posterior density $p(\theta, \phi|y)$ in unnormalized form as a product of the hyperprior density $p(\phi)$, the **population distribution** $p(\theta|\phi)$ and the likelihood $p(y|\theta)$.
- (2) determine analytically the conditional posterior density of θ given the hyperparameter ϕ, for fixed observed y this is a function of $\phi, p(\theta|\phi, y)$.
- (3) estimate ϕ using the Bayesian paradigm, that is obtain its marginal posterior distribution $p(\phi|y)$.

The first step is immediate, and the second step is easy for conjugate models because conditional on ϕ the population distribution for θ is just the independent and identical distribution model (5.1). The conditional posterior density is a product of conjugate posterior densities for the components θ_j.

The marginal posterior $p(\phi|y)$ can be obtained using an integral

$$p(\phi|y) = \int p(\theta, \phi|y) d\theta \tag{5.5}$$

For many models the marginal computed using albegra and conditional probability formula

$$p(\phi|y) = \frac{p(\theta, \phi|y)}{p(\theta|\phi|y)} \tag{5.6}$$

This expression is useful because the numerator is just the joint posterior, and the denominator is the posterior distribution for θ if ϕ where known. So the denominator $p(\theta|\phi, y)$ is regarded as a function of both θ for fixed ϕ, y.

Drawing simulations from the posterior distribution

The strategy for simulating posterior distribution for $p(\theta, \phi|y)$ for simple hierarchical models follows

-(1) draw the vector of hyperparameters ϕ from its marginal posterior distribution $p(\phi|y)$. If ϕ is low-dimensional, the methods in Chapter 3 can be used, and for high dimensional problems, MCMC must be used. -(2) draw the parameter vector θ from its conditional posterior distribution $p(\theta|\phi, y)$ given the drawn value for ϕ. for simple examples $p(\theta|\phi, y) = \prod p(\theta_j|\phi, y)$ the components θ_j can be drawn independently, one at a time. -(3) if desired, draw the predictive values $\tilde{y}$ from the posterior predictive distribution given the drawn θ. Depending on the problem it might be necessary to first draw $\tilde{\theta}$ given ϕ, and then draw a future value from the distribution.

Application to the model for rat tumors

Exercises

Q1

A box has 1 black and 1 white ball - a i) pick 1 ball, y1, return it, and draw another y2. This is exchangeable, because other than y1, we have no information about what the draw y2 could be.
- a ii) we have independence by replacement - a iii) since the draws are identical, then y1 and y2 are both independent draws and the pair of draws can be independent.

-b i) drawing without replacement, y1 is not exchangeable because given y1 color, we know what y2 color is. -b ii) not independent -b iii) as a pair independent (RB) independent of a second pair (BR) with order then the order RB is independent of BR if order matters.

-c i) if there were a million balls, then due to large numbers this is an exchangeable and independent process.

Q2

For unknown model parameters (a) we have total of n black and white balls, then this is an exchangeable and independent process. (b) due to ignorance we

have exchangeability, but we do not have independence since this is without replacement. (c) if we know how many colors balls there are then this is not an exchangeable process and not independent process for finite n. For large n, then we can assume independence and exchangeability.

Q3

all 8 parameters are estimated from sampling from the posterior $p(\tau|y)$, for ease of data wrangling, we obtain the quantiles for each theta separately. but in each posterior estimation, all points are simulataneously estimated.

This reproduces Table 5.3 from the text.

```
school<-data.frame(school=LETTERS[1:8],
                   yj =c(28,8,-3,7,-1,1,18,12),
                   sigmaj = c(15,10,16,11,9,11,10,18))
school$sigma2j<-school$sigmaj^2

 ## reproducing section 5.5 on pooling

## pooled
 ybar_pool<-sum(school$yj/school$sigmaj^2)/sum(1/school$sigmaj^2)
 pool_var<-1/sum(1/school$sigmaj^2)
  ybar_pool+2*sqrt(pool_var) # 15.82946
```

```
## [1] 15.82946
```

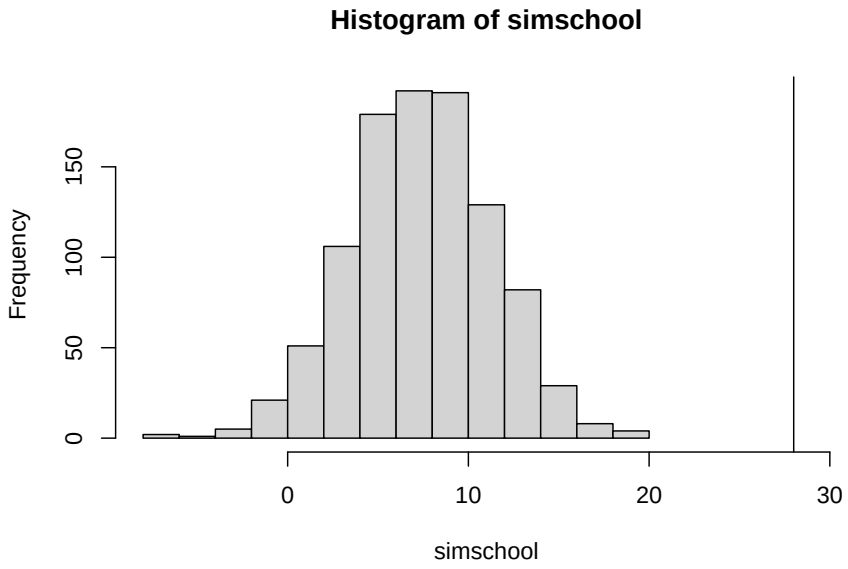
```
  ybar_pool-2*sqrt(pool_var) ## -0.5
```

```
## [1] -0.4582216
```

```
 ## classical test
   sum( (school$yj-mean(school$yj))^2/school$sigmaj^2)  ## 4.7 which is less than the
```

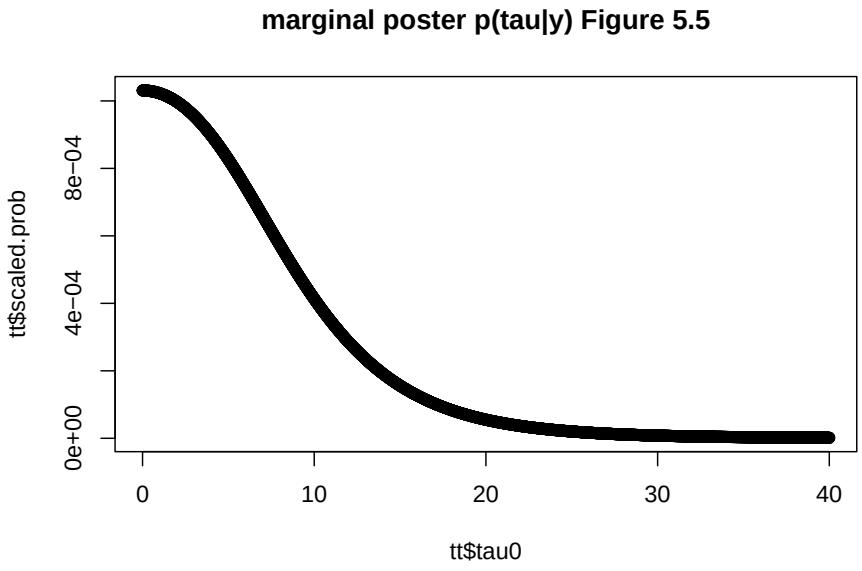
```
## [1] 4.775407
```

```
   ## checking to see if 28 is from the pooled population

   simschool <-rnorm(1000, mean= ybar_pool, sd=sqrt(pool_var))
   hist(simschool,xlim=c(-8,30))
   abline(v=max(school$yj ))
```




```
 ### the maximum observation  School A = 28 points  is not with the population, so pooling does
 ###

 ## posterior simulation under a normal model
 ## yij | thetaj ~ N(thetaj, sigma^2)
 ##  likelihood using sufficient statistics
  ## ybarj | thetaj ~ N(thetaj, sigmaj^2)

 ### prior on tau can follow a uniform distribution
 ## using a grid approach for tau
    tau0= seq(0,40,by=0.01)
    stepsize=0.01
# 5.20 total precision
vmu.inverse<-function(tau2,sigma2j){
  sum(1/(sigma2j+tau2))
}
vmu<-sapply(tau0^2,function(x) vmu.inverse(x,school$sigmaj^2))

# total mean effect
muhat<-function(ybarj, sigma2j,tau2){
  numer<- sum(ybarj/(sigma2j+tau2))
  denom<- sum( 1/(sigma2j+tau2))
  hat<- numer/denom
  return(hat)
}
```

```
  ## marginal posterior distribution p(tau| y)  5.21
marginal.posterior.tau<-function(tau,sigma2j,ybarj){
  tau2<-tau^2
  prob.tau<-dunif(tau,min=0,max=40) ## assuming unif prior
  vmu.inv<-vmu.inverse(tau2,sigma2j)
  vmu<-sqrt(1/vmu.inv)
  total.precision<-1/(sigma2j+tau2)
  group.precision.term<- sqrt(total.precision)
  group.mu<-muhat(ybarj,sigma2j,tau2)
  exp.term<- exp(-1*(ybarj -group.mu )^2/(2*(sigma2j+tau2)))

  group.prod<-prod( group.precision.term*exp.term)

  final<- prob.tau*vmu*group.prod
  return(final)
}
tt<-data.frame(tau0,pt=sapply(tau0,function(x) marginal.posterior.tau(x,school$sigma
tt$scaled.prob<-tt$pt/sum(tt$pt)
plot(tt$tau0,tt$scaled.prob, main='marginal poster p(tau|y) Figure 5.5')
```




```
  ## prior on tau can follow a scaled inverse-X2(n-1,s^2) distribution
## alternatively we can use scaled inverse

 ## now that we have the distribution of tau,  sample from it
```

```
 ## marginal posterior 5.19
muhat_given_tau.y<-sapply(tau0,function(x) muhat(school$yj,school$sigma2j,x^2))
  vmu_given_tau.y<-1/(sapply(tau0^2,function(x) vmu.inverse(x,school$sigmaj^2)))


   ### FIX ME:  sample mu
# equation 5.20
 marginal.post.mu_given_tau.y<-function(yj,sigma2j,tau2){
   muhat<-muhat(yj,sigma2j,tau2)
   ## precision term
   vmu<-vmu.inverse(tau2,sigma2j)
  mu<- rnorm(100,mean=muhat,sd=sqrt(1/vmu)) ## how many points per posterior tau
  ## we take the posterior mean given 1 value of tau
  prob.mu<-dnorm(mean(mu),mean=muhat,sd=sqrt(1/vmu))
  return( data.frame(mu=mean(mu),prob.mu=prob.mu))
 }

 ## given mu, sample theta j
 ## conditional posterior for each theta j
 ## eq 5.17
conditional.posterior.thetaj<-function(yj,sigma2j,tau2,mu){
   thetajhat<-(yj/sigma2j + mu/tau2)/(1/sigma2j+1/tau2)
   vj<-1/((1/sigma2j)+(1/tau2))
   nj<-length(thetajhat)
  thetaj<- sapply(seq(1,nj),function(x) rnorm(1, thetajhat[x], sd=sqrt(vj)[x]))
  prob.thetaj<-numeric(nj)

  prob.thetaj<- sapply(seq(1,nj),function(x) dnorm(thetaj[x], mean=thetajhat[x], sd=sqrt(vj)[x]


  return( data.frame(thetaj=(thetaj),prob.thetaj=prob.thetaj))
 }

posterior<-function(tau,sigma2j,yj){
  tau2<-tau^2
  post.tau<-marginal.posterior.tau(tau,sigma2j,yj)
  post.mu<-marginal.post.mu_given_tau.y(yj,sigma2j,tau2)
  post.theta<-conditional.posterior.thetaj(yj,sigma2j,tau2,post.mu$mu)

    full.post<-data.frame(theta=post.theta$thetaj,pdf=post.tau*post.mu$prob.mu*post.theta$prob.
    return(full.post)
}

## simulate the effects of thetaj
## closely reproduces table 5.3 in section 5
```

```
theta1<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[1,1])
theta1[is.nan(theta1)]<-mean(theta1[!is.nan(theta1)])

t1q<-quantile(theta1,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta2<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[2,1])
theta2[is.nan(theta2)]<-mean(theta2[!is.nan(theta2)])

t2q<-quantile(theta2,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta3<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[3,1])
   theta3[is.nan(theta3)]<-mean(theta3[!is.nan(theta3)])


t3q<-quantile(theta3,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta4<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[4,1])
    theta4[is.nan(theta4)]<-mean(theta4[!is.nan(theta4)])


t4q<- quantile(theta4,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta5<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[5,1])
  theta5[is.nan(theta5)]<-mean(theta5[!is.nan(theta5)])


t5q<- quantile(theta5,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta6<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[6,1])
  theta6[is.nan(theta6)]<-mean(theta6[!is.nan(theta6)])


t6q<-quantile(theta6,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 theta7<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                function(x) posterior(x,school$sigma2j,school$yj)[7,1])
  theta7[is.nan(theta7)]<-mean(theta7[!is.nan(theta7)])
```

```
  t7q<-quantile(theta7,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)


   theta8<- sapply(sample(tt$tau0,1000,prob=tt$scaled.prob),
                   function(x) posterior(x,school$sigma2j,school$yj)[8,1])
    theta8[is.nan(theta8)]<-mean(theta8[!is.nan(theta8)])

  t8q<-quantile(theta8,c(0.025,0.25,0.5,0.75,0.975),na.rm=TRUE)

 allq<-rbind(t1q,t2q,t3q,t4q,t5q,t6q,t7q,t8q)
 print(allq)
```

```
##            2.5%      25%       50%       75%    97.5%
## t1q   0.4457383 7.326460 10.017273 16.091238 33.37148
## t2q  -3.9503168 4.831116  7.726473 10.619370 20.25554
## t3q -14.8226207 2.859082  6.858540  9.538431 18.57518
## t4q  -7.3115433 5.009563  7.783060 10.934833 21.04213
## t5q  -9.4629862 1.658382  6.024443  8.293584 14.69150
## t6q -10.7831533 2.029477  6.586762  8.797824 17.50866
## t7q   0.9487190 7.244206  9.623687 14.367122 25.85331
## t8q  -7.3113309 4.974897  7.888110 11.282060 27.24911
```

- i could not figure out how to take the margin across μ for 5.6 and 5.7.
- we found that our posterior effects had 4% greater than the maximum observed value 28.4 whereas the text found 22/200 which is slightly larger from 200 simulations, we used 1,000.

```
  abest<-table(theta1>(theta2) & theta1>theta3 & theta1>theta4 & theta1>theta5 & theta1>theta6 &
  bbest<-table(theta2>(theta1) & theta2>theta3 & theta2>theta4 & theta2>theta5 & theta2>theta6 &
  cbest<-table(theta3>(theta1) & theta3>theta2 & theta3>theta4 & theta3>theta5 & theta3>theta6 &
  dbest<-table(theta4>(theta1) & theta4>theta2 & theta4>theta3 & theta4>theta5 & theta4>theta6 &
  ebest<-table(theta5>(theta1) & theta5>theta2 & theta5>theta3 & theta5>theta4 & theta5>theta6 &
  fbest<-table(theta6>(theta1) & theta6>theta2 & theta6>theta3 & theta6>theta4 & theta6>theta5 &
  gbest<-table(theta7>(theta1) & theta7>theta2 & theta7>theta3 & theta7>theta4 & theta7>theta6 &
  hbest<-table(theta8>(theta1) & theta8>theta2 & theta8>theta3 & theta8>theta4 & theta8>theta5 &

  a<-abest['TRUE']/sum(abest)
  b<-bbest['TRUE']/sum(bbest)
  c<-cbest['TRUE']/sum(cbest)
  d<-dbest['TRUE']/sum(dbest)
  e<-ebest['TRUE']/sum(ebest)
  f<-fbest['TRUE']/sum(fbest)
  g<-gbest['TRUE']/sum(gbest)
  h<-hbest['TRUE']/sum(hbest)
```

5.3.0.1 Q3 part a

- 5.3a this is correct and is very close to section 5 results.

```
better<-matrix(0,nrow=8,ncol=8)

fillIn<-function(better,theta1,theta2,i=1,j=2){
 pij<-table(theta1>theta2)/sum(table(theta1>theta2))
 better[i,j]<-pij['TRUE']
return(better)
}
better<-fillIn(better,theta1,theta2,i=1,j=2)
better<-fillIn(better,theta1,theta3,i=1,j=3)
better<-fillIn(better,theta1,theta4,i=1,j=4)
better<-fillIn(better,theta1,theta5,i=1,j=5)
better<-fillIn(better,theta1,theta6,i=1,j=6)
better<-fillIn(better,theta1,theta7,i=1,j=7)
better<-fillIn(better,theta1,theta8,i=1,j=8)

better<-fillIn(better,theta2,theta1,i=2,j=1)
better<-fillIn(better,theta2,theta3,i=2,j=3)
better<-fillIn(better,theta2,theta4,i=2,j=4)
better<-fillIn(better,theta2,theta5,i=2,j=5)
better<-fillIn(better,theta2,theta6,i=2,j=6)
better<-fillIn(better,theta2,theta7,i=2,j=7)
better<-fillIn(better,theta2,theta8,i=2,j=8)

better<-fillIn(better,theta3,theta1,i=3,j=1)
better<-fillIn(better,theta3,theta2,i=3,j=2)
better<-fillIn(better,theta3,theta4,i=3,j=4)
better<-fillIn(better,theta3,theta5,i=3,j=5)
better<-fillIn(better,theta3,theta6,i=3,j=6)
better<-fillIn(better,theta3,theta7,i=3,j=7)
better<-fillIn(better,theta3,theta8,i=3,j=8)


better<-fillIn(better,theta4,theta1,i=4,j=1)
better<-fillIn(better,theta4,theta2,i=4,j=2)
better<-fillIn(better,theta4,theta3,i=4,j=3)
better<-fillIn(better,theta4,theta5,i=4,j=5)
better<-fillIn(better,theta4,theta6,i=4,j=6)
better<-fillIn(better,theta4,theta7,i=4,j=7)
better<-fillIn(better,theta4,theta8,i=4,j=8)

 better<-fillIn(better,theta5,theta1,i=5,j=1)
```

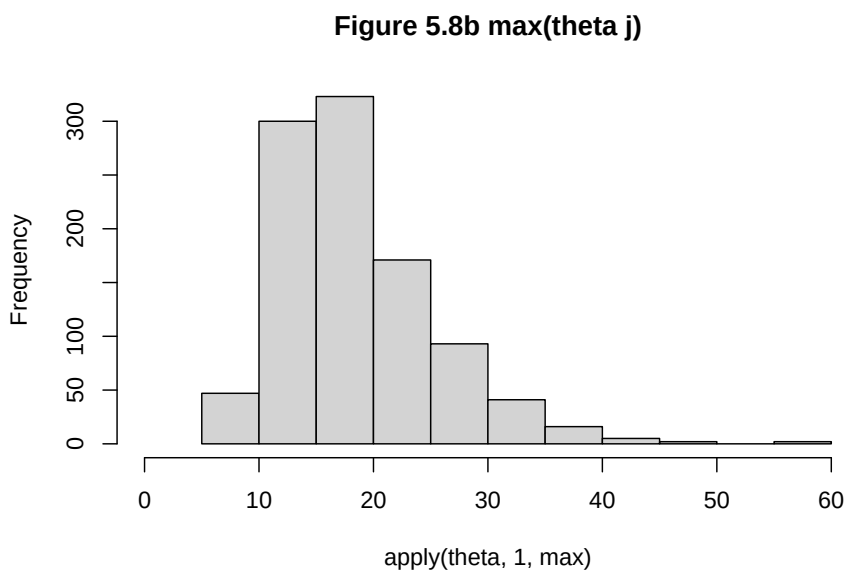
```
  better<-fillIn(better,theta5,theta2,i=5,j=2)
  better<-fillIn(better,theta5,theta3,i=5,j=3)
  better<-fillIn(better,theta5,theta4,i=5,j=4)
  better<-fillIn(better,theta5,theta6,i=5,j=6)
  better<-fillIn(better,theta5,theta7,i=5,j=7)
  better<-fillIn(better,theta5,theta8,i=5,j=8)


  better<-fillIn(better,theta6,theta1,i=6,j=1)
  better<-fillIn(better,theta6,theta2,i=6,j=2)
  better<-fillIn(better,theta6,theta3,i=6,j=3)
  better<-fillIn(better,theta6,theta4,i=6,j=4)
  better<-fillIn(better,theta6,theta5,i=6,j=5)
  better<-fillIn(better,theta6,theta7,i=6,j=7)
  better<-fillIn(better,theta6,theta8,i=6,j=8)

    better<-fillIn(better,theta7,theta1,i=7,j=1)
  better<-fillIn(better,theta7,theta2,i=7,j=2)
  better<-fillIn(better,theta7,theta3,i=7,j=3)
  better<-fillIn(better,theta7,theta4,i=7,j=4)
  better<-fillIn(better,theta7,theta5,i=7,j=5)
  better<-fillIn(better,theta7,theta6,i=7,j=6)
  better<-fillIn(better,theta7,theta8,i=7,j=8)

   better<-fillIn(better,theta8,theta1,i=8,j=1)
  better<-fillIn(better,theta8,theta2,i=8,j=2)
  better<-fillIn(better,theta8,theta3,i=8,j=3)
  better<-fillIn(better,theta8,theta4,i=8,j=4)
  better<-fillIn(better,theta8,theta5,i=8,j=5)
  better<-fillIn(better,theta8,theta6,i=8,j=6)
  better<-fillIn(better,theta8,theta7,i=8,j=7)

  better<-cbind(c(a,b,c,d,e,f,g,h),better)
  rownames(better)<-LETTERS[1:8]
  colnames(better)<-c("best",LETTERS[1:8])

  print(better)
```

```
##    best     A     B     C     D     E     F     G     H
## A 0.290 0.000 0.671 0.714 0.662 0.783 0.752 0.533 0.633
## B 0.102 0.329 0.000 0.570 0.503 0.634 0.620 0.343 0.487
## C 0.074 0.286 0.430 0.000 0.425 0.556 0.526 0.306 0.417
## D 0.100 0.338 0.497 0.575 0.000 0.636 0.601 0.358 0.482
## E 0.040 0.217 0.366 0.444 0.364 0.000 0.467 0.220 0.346
## F 0.047 0.248 0.380 0.474 0.399 0.533 0.000 0.249 0.377
```

```
## G 0.229 0.467 0.657 0.694 0.642 0.780 0.751 0.000 0.609
## H 0.124 0.367 0.513 0.583 0.518 0.654 0.623 0.391 0.000
```

Inference on the $P(max(\theta_j) > 28.4)$ in our data is 93/1000 which is 0.093. from the text 22/200 which is roughly 0.11

```
theta<-data.frame(theta1,theta2,theta3,theta4,theta5,theta6,theta7,theta8)
 hist(apply(theta,1,max),xlim=c(0,60),main='Figure 5.8b max(theta j)')
```




```
 table(apply(theta,1,max)>28.4)['TRUE']/sum(table(apply(theta,1,max)>28.4))
```

```
##  TRUE
## 0.088
```

5.3.0.2 Q 5.3 part b

- 5.3.b taking $\tau \to \infty$ we have $\theta_j \sim N(y_j, \sigma_j^2)$

For $P(_i > _j) = P(\theta_i - \theta_j > 0)$

```
 x1<-rnorm(1000,mean=school$yj[1],sd=school$sigmaj[1])
 x2<-rnorm(1000,mean=school$yj[2],sd=school$sigmaj[2])
 x3<-rnorm(1000,mean=school$yj[3],sd=school$sigmaj[3])
 x4<-rnorm(1000,mean=school$yj[4],sd=school$sigmaj[4])
 x5<-rnorm(1000,mean=school$yj[5],sd=school$sigmaj[5])
 x6<-rnorm(1000,mean=school$yj[6],sd=school$sigmaj[6])
 x7<-rnorm(1000,mean=school$yj[7],sd=school$sigmaj[7])
 x8<-rnorm(1000,mean=school$yj[8],sd=school$sigmaj[8])

 allQ<-rbind(  quantile(x1,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x2,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x3,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x4,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x5,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x6,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x7,c(0.025,0.25,0.5,0.75,0.975)),
               quantile(x8,c(0.025,0.25,0.5,0.75,0.975))
               )
 rownames(allQ)<-LETTERS[1:8]

 ## pair wise
 better2<-matrix(0,nrow=8,ncol=8)
 for(i in 1:8){
   for(j in 1:8){
   better2[i,j]<- pnorm( (school$yj[i]- school$yj[j])/sqrt( school$sigma2j[i]+school$sigma2j[j]))
   }
 }

allind<-cbind(x1,x2,x3,x4,x5,x6,x7,x8)

 x1b<-table(x1>=apply(allind,1,max))['TRUE']/1000
 x2b<-table(x2>=apply(allind,1,max))['TRUE']/1000
 x3b<-table(x3>=apply(allind,1,max))['TRUE']/1000
 x4b<-table(x4>=apply(allind,1,max))['TRUE']/1000
 x5b<-table(x5>=apply(allind,1,max))['TRUE']/1000
 x6b<-table(x6>=apply(allind,1,max))['TRUE']/1000
 x7b<-table(x7>=apply(allind,1,max))['TRUE']/1000
 x8b<-table(x8>=apply(allind,1,max))['TRUE']/1000

 sep.prob<- cbind(c(x1b,x2b,x3b,x4b,x5b,x6b,x7b,x8b),better2)
 rownames(sep.prob)<-LETTERS[1:8]
   colnames(sep.prob)<-c("best",LETTERS[1:8])

   print(sep.prob)
```

```
##    best          A         B         C         D         E         F          G
## A 0.554 0.50000000 0.8663713 0.9212424 0.8705441 0.9513231 0.9266837 0.71045013
## B 0.034 0.13362875 0.5000000 0.7200530 0.5268155 0.7482410 0.6811336 0.23975006
## C 0.023 0.07875756 0.2799470 0.5000000 0.3032674 0.4566223 0.4183914 0.13285469
## D 0.029 0.12945588 0.4731845 0.6967326 0.5000000 0.7132410 0.6501386 0.22966818
## E 0.004 0.04867694 0.2517590 0.5433777 0.2867590 0.5000000 0.4440458 0.07893687
## F 0.017 0.07331631 0.3188664 0.5816086 0.3498614 0.5559542 0.5000000 0.12640645
## G 0.174 0.28954987 0.7602499 0.8671453 0.7703318 0.9210631 0.8735935 0.50000000
## H 0.165 0.24734659 0.5770127 0.7333055 0.5936804 0.7408523 0.6989733 0.38537815
##           H
## A 0.7526534
## B 0.4229873
## C 0.2666945
## D 0.4063196
## E 0.2591477
## F 0.3010267
## G 0.6146218
## H 0.5000000
```

5.3.0.3 Q 5.3.c

- 5.3.c Discussing the differences the intervals are much wider in the separate models for θ_j compared to the Bayesian model. This is because the Bayesian model borrows strenghts of associaton across other school information.

Under separate model, the probability that A is best is approxmiately 0.521, whereas in Bayesian model it is 0.283. These inferences are more conservative in the Bayesian model because it incorporates uncertainty in the model. The Bayesian model assumes a uniform prior for τ which is the deviation/variability of the school coaching effectiveness, whereas in the separate model operates under the assumption that there is large variability of coaching effectiveness $\tau = \infty$.

```
print(allQ)
```

```
##          2.5%         25%        50%       75%    97.5%
## A  -1.6671674  16.8008160 28.4378866 38.292573 56.64421
## B -11.7881530   1.4666970  7.8338828 15.011281 25.71342
## C -32.9503177 -13.1854338 -2.8801841  8.302523 27.71221
## D -13.2036155   0.6344413  7.8860750 14.677863 28.52870
## E -18.8775183  -6.4175582 -0.3387703  5.609494 16.85795
## F -20.2917461  -6.9974197  0.9482539  8.851298 22.31686
## G  -0.9984314  11.3323360 17.5468144 24.429637 35.98980
## H -21.8486938  -0.4152265 12.2354066 23.914510 45.65002
```

```
print(allq)
```

```
##            2.5%      25%       50%       75%    97.5%
## t1q   0.4457383 7.326460 10.017273 16.091238 33.37148
## t2q  -3.9503168 4.831116  7.726473 10.619370 20.25554
## t3q -14.8226207 2.859082  6.858540  9.538431 18.57518
## t4q  -7.3115433 5.009563  7.783060 10.934833 21.04213
## t5q  -9.4629862 1.658382  6.024443  8.293584 14.69150
## t6q -10.7831533 2.029477  6.586762  8.797824 17.50866
## t7q   0.9487190 7.244206  9.623687 14.367122 25.85331
## t8q  -7.3113309 4.974897  7.888110 11.282060 27.24911
```

5.3.0.4 Q 5.3 part d

- 5.3 d setting $\tau = 0$ creates $\theta_j|\mu, \tau, y \sim N(\infty, 0)$ because of the equation for $\hat{\theta_j} = \infty/\infty$ and $V_j = 1/\infty = 0$. the inferences for posterior for $\mu|\tau, y$ and $\tau|y$ are not degenerate. but for the parameters for θ these are degenerate. Setting $\tau = 0$ means that the variance parameter goes to 0, which reduces all points to a point mass, and sets them equal (0 variability).

Q4

- (a) by definition of exchangability, other than the data there can be no information that can distinguish θ_j, and by having a distinct grouping of parameters in either N(1,1) or N(-1,1) this violates exchangability. However, the problem states that *we have not observed* which parameters come from any distribution. This is similar to the divorce problem, such that over the 8 mid-western states, we know that Utah and Nevada will have lower/higher divorce rates, however after observing 7 values, we do not have information about the 8th state that was not yet observed; this is an exchangeable process. However if we *had information* that the last state was Nevada or Utah, then this is *not* exchangeable. Since we have information that there will be a grouping, but currently did not observe the parameters, this is exchangeable.

Using equation 5.2 the prior is $p(\theta) = \int \prod p(\theta_j|\phi)p(\phi)d\phi$

$$p(\theta_1, ...\theta_{2J}) \int \prod N(\theta_j|\mu, \sigma^2)p(\mu, \sigma^2) = \sum_p \prod_{i=1}^{J} N(\theta_j(p)|1, 1) \prod_{k=J+1}^{2J} N(\theta_{k(p)}|-1, 1) \frac{1}{\binom{2J}{J}} \tag{5.7}$$

Equation 5.14 we take the sum over the posterior under the prior. the probability $p(\mu, \sigma^2)$ we have 2J total groups and choose J.

-(b) given two distributions N(1,1) and N(-1,1) the estimates will have a negative covariance so we do not have an identically distributed mixture, we can assume independence, but for large values observed, these are likely to be originated from N(1,1), and for smaller , negative, values these are likely originated from N(-1,1). Thus these are not identically distributed.

-(c) for $J \to \infty$ the covariance between these two groups reduces to 0 as the number of groups approaches infinity. De-finitti's theorem for infinite groups indicates that the distinction between exactly 2 groups disappears, and also assigning exactly half of the parameters to the correct half also disappears. So for infinite groupings, the distinction between groups is not possible and there is exchangability in the limit. However for finite groups we can not apply Definitti's theorem.

Q7

-(a) we use Adam's law to find the mean/variance of the prior y, s.t. $y|\theta \sim Poi(\theta)$, and a gamma prior

$$\begin{aligned} E[y] &= E[E(y|\theta)] = E[\theta] \\ &= \alpha/\beta \\ V[y] &= E[V(y|\theta)] + V[E(y|\theta)] = E[\theta] + V[\theta] = \alpha/\beta + \alpha/\beta^2 = \frac{\alpha(\beta+1)}{\beta^2} \end{aligned}$$

-(b) in the normal model we have marginally t= $\frac{\sqrt{n}(\mu-\bar{y})}{s}$ that is t_{n-1}. Deriving the first 2 moments, where $\mu \sim N(\bar{y}, \sigma^2/n)$. where $\sigma^2|y \sim$ Scaled-Inv-Chi2(n-1,s^2).

Note we have to condition on y because the formula for t_{n-1} includes $\bar{y}$

$$\begin{aligned} E[t_{n-1}|y] &= E[E(\frac{\sqrt{n}(\mu-\bar{y})}{s})|y] = 0 \\ V[t_{n-1}|y] &= E[V(t|\sigma^2, y)|y] + V[E(t|y)] \\ &= E[\frac{n}{s^2}V((\mu-\bar{y})|\sigma^2, y)|y] + 0 \\ &= E[\frac{n\sigma^2}{n*s^2}] = \frac{s^2(n-1)}{s^2(n-3)} = (n-1)/(n-3); n > 3 \end{aligned}$$

Q10

-(a) if the hyperprior $p(\mu, \tau) \propto \tau^{-1}$ show that the posterior is improper the posterior is written as $p(\theta, \mu, \tau|y) = p(\theta, \mu|\tau, y)p(\tau|y)$. This is proper if both terms are proper, and it is improper if at least 1 term is improper. we show

that $p(\tau|y)$ is improper for $p(\tau) \propto \tau^{-1}$. Using equation 5.21

$$p(\tau|y) \propto \int p(\tau) V_\mu^{-1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} exp(\frac{-(\bar{y_{.j}} - \hat{\mu})^2)}{2(\sigma_j^2 + \tau^2)}) d\tau$$
$$\int_0^\infty \frac{1}{\tau} V_\mu^{-1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} exp(\frac{-(\bar{y_{.j}} - \hat{\mu})^2)}{2(\sigma_j^2 + \tau^2)}) d\tau$$

so this is undefined for $\tau = 0$ and hence the term $p(\tau|y)$ is undefined and causes the posterior to be undefined.

-(b) for $p(\mu, \tau) \propto 1$ equation 5.16 has 2 proper distributions that are in the normal family distribution and are well-defined. $N(\theta_j|\mu, \tau^2)$ and $N(\bar{y_{.j}}|\theta_j, \sigma_j^2)$ so these are proper distributions.

Another way to see this is to take the limits of equation 5.21 where $p(\tau) \propto 1$ As the limit goes to 0 the term is well defined.

$$\lim_{\tau \to 0} p(\tau) V_\mu^{-1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} exp(\frac{-(\bar{y_{.j}} - \hat{\mu})^2)}{2(\sigma_j^2 + \tau^2)}) =$$
$$\frac{1}{\sum \frac{1}{\sigma_j^2}}^{1/2} \prod (\sigma_j^2)^{-1/2} exp(\frac{-(\bar{y_{.j}} - \hat{\mu})^2)}{2\sigma_j^2})$$

As the limit $\tau \to \infty$ we the exponential term goes to 0 because τ is in the denominator, so the entire term will go to 0. So the posterior is proper because the limits are defined.

-(c) if i had data from J=2 schools only, i would test for between variance and within variance to see is MS justify pooling the estimates for both schools, or require separate models for each. if within each school there were enough students to justify modeling school-specific parameters, I would use equation 5.16 and model the posterior distribution.
Another option is to model each school independently as separate models with known variance and i would use non-informative priors.

Q11

For non conjugate models suppose that in the rat tumor example we had $logit(\theta_j) \sim N(\mu, \tau^2)$ for j=1,...,J and assume a non-informative prior for the hyperparameters. -(a) the joint posterior can be written as

$$p(\theta, \mu, \tau|y) = p(\theta, \mu, \tau)p(y|\theta, \mu, \tau) = p(\theta|\mu, \tau)p(\mu, \tau)p(y|\theta, \mu, \tau) \tag{5.8}$$

Note that $p(\theta, \mu, \tau)$ is in terms of θ and not on the logit scale, so we need the jacobian for change of variables.

$$p(\theta|\mu,\tau)p(\mu,\tau)p(y|\theta,\mu,\tau) = p(\mu,\tau)\prod_j \frac{1}{\sqrt{2\pi\tau}} exp(\frac{-1}{2\tau^2}(logit(\theta_j-\mu)^2))\prod_j(\theta_j)^{y_i}(1-\theta_j)^{n_i-y_i}|dlogit(\theta)/d\theta$$

$$p(\theta|\mu,\tau)p(\mu,\tau)p(y|\theta,\mu,\tau) = p(\mu,\tau)\prod_j \frac{1}{\sqrt{2\pi\tau}} exp(\frac{-1}{2\tau^2}(logit(\theta_j-\mu)^2))\prod_j(\theta_j)^{y_i}(1-\theta_j)^{n_i-y_i}|\frac{1}{\theta(1-\theta)}$$

-(b) we can not integrate this equation fully because it is not in a known family. Although we have separate factors in the equation, we can not analytically derive the integral in closed form. -(c) Using equation 5.5 $p(\mu,\tau|y) = \frac{p(\theta,\mu,\tau|y)}{p(\theta|\mu,\tau,y)}$ we can not integrate the denominator $p(\theta|mu,\tau,y)$ in closed form as a function of μ,τ which doesn't have a closed form solution.

Q12

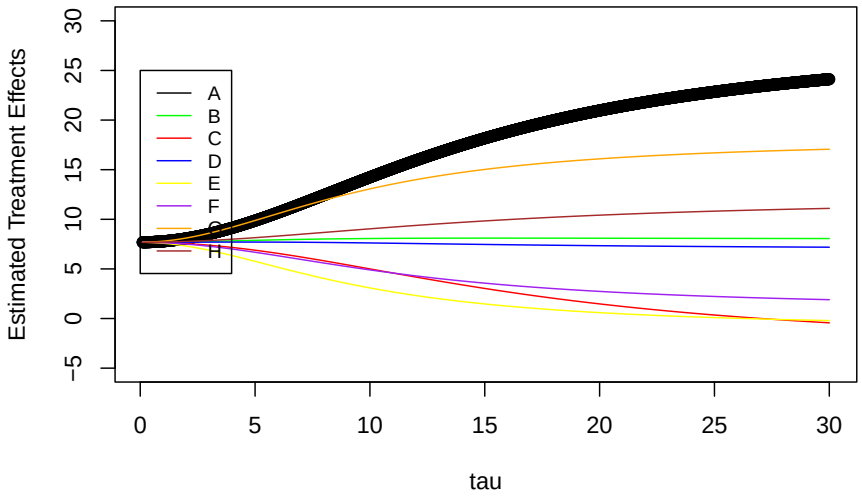
To find the conditional expection we use equation 5.20 and 5.17 from the text

$$\begin{aligned} E[\theta_j|\tau,y] &= E[E(\theta_j|\mu,\tau,y)|\tau,y] \\ &= E[\hat{\theta_j}|\tau,y] = \frac{1/\sigma_j^2\bar{y_{.j}} + 1/\tau^2 E[\mu]}{1/\sigma_j^2+1/\tau^2} \\ &= \frac{1/\sigma_j^2\bar{y_{.j}} + 1/\tau^2\hat{\mu}}{1/\sigma_j^2+1/\tau^2} \end{aligned}$$

where $\hat{\mu} = \frac{\sum_j \frac{1}{\sigma_j^2+\tau^2}\bar{y_{.j}}}{\sum_j \frac{1}{\sigma_j^2+\tau^2}}$

```
school<-data.frame(school=LETTERS[1:8],
                   yj =c(28,8,-3,7,-1,1,18,12),
                   sigmaj = c(15,10,16,11,9,11,10,18))
school$sigma2j<-school$sigmaj^2

## conditional expectation

mu_hat<-function(allsigma2j,tau2,allybarj){
   numer<-sum( (1/(allsigma2j+tau2))*allybarj)
   denom<-sum(1/(allsigma2j+tau2))
   muhat= numer/denom
   return(muhat)
}

conditionalExpectation<-function(sigma2j,ybarj,tau2,allsigma2j,allybarj){
  muhat<-mu_hat(allsigma2j, tau2,allybarj)
```

```
  a<-((1/sigma2j)*ybarj)/(1/sigma2j + 1/tau2)
  b<- ((1/tau2)*muhat)/(1/sigma2j + 1/tau2)
  cond.mean<-a+b
  return(cond.mean)
}

tau_seq = seq(0.1,30,by=0.01)

  allmeans<-matrix(0, nrow=length(tau_seq),ncol=nrow(school))
for(i in 1:nrow(school)){
 aa<-sapply(tau_seq,function(x) conditionalExpectation(school$sigma2j[i],school$yj[i],tau2=x^2,sc
 allmeans[,i]<-aa
}

plot(tau_seq,allmeans[,1],ylim=c(-5,30),ylab='Estimated Treatment Effects', xlab='tau')
lines(tau_seq,allmeans[,2],col='green')
lines(tau_seq,allmeans[,3],col='red')
lines(tau_seq,allmeans[,4],col='blue')
lines(tau_seq,allmeans[,5],col='yellow')
lines(tau_seq,allmeans[,6],col='purple')
lines(tau_seq,allmeans[,7],col='orange')
lines(tau_seq,allmeans[,8],col='brown')

legend(0, 25,
       legend=c("A", "B","C","D","E","F","G","H"),
       col=c("black", "green","red","blue","yellow","purple","orange","brown"), lty=1, cex=0.8)
```

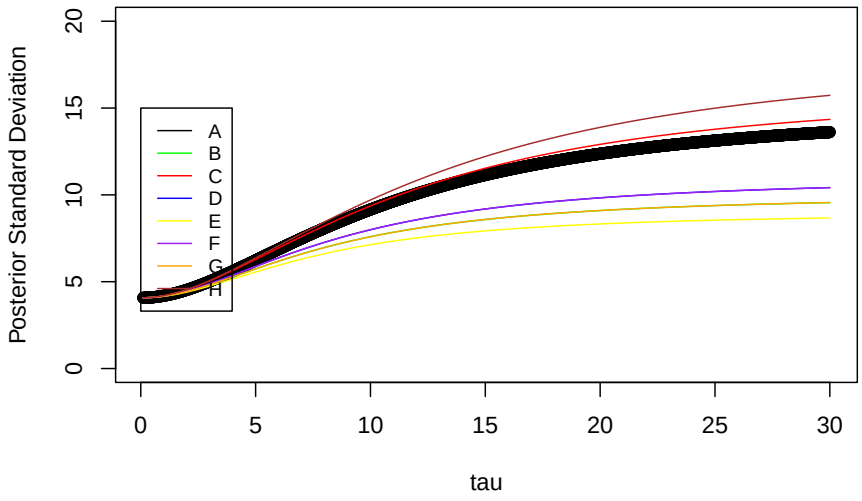


For the $V(\theta_j|\tau, y) = E[V(\theta_j|\mu, \tau, y)|\tau, y] + V[E(\theta_j|\mu, \tau, y)|\tau, y)]$ Using the equations for V_j, V_μ from section 5 we derived the conditional variance.

$$\begin{aligned} E[V(\theta_j|\mu, \tau, y)|\tau, y] + V[E(\theta_j|\mu, \tau, y)|\tau, y)] &= E[V_j|\tau, y] + V[\widehat{\theta}_j|\tau, y] \\ &= V_j + Vj^2(1/\tau^2)^2 V_\mu \end{aligned}$$

```
 V_mu<-function(allsigma2j,tau2){
   vinv<-sum( 1/(allsigma2j+tau2))
   vmu=1/vinv
   return(vmu)
 }

conditionalVariance<-function(sigma2j, tau2,allsigma2j){
  vmu<-V_mu(allsigma2j,tau2)
  Vj = 1/(1/sigma2j+1/tau2)
  cond.var<-Vj+Vj^2*(1/tau2)^2*vmu
  return(cond.var)
}

tau_seq = seq(0.1,30,by=0.01)

  allsd<-matrix(0, nrow=length(tau_seq),ncol=nrow(school))
for(i in 1:nrow(school)){
 aa<-sapply(tau_seq,function(x) conditionalVariance(school$sigma2j[i],tau2=x^2,school$s
 allsd[,i]<-sqrt(aa)
```

```
}

plot(tau_seq,allsd[,1],ylim=c(0,20),ylab='Posterior Standard Deviation', xlab='tau')
lines(tau_seq,allsd[,2],col='green')
lines(tau_seq,allsd[,3],col='red')
lines(tau_seq,allsd[,4],col='blue')
lines(tau_seq,allsd[,5],col='yellow')
lines(tau_seq,allsd[,6],col='purple')
lines(tau_seq,allsd[,7],col='orange')
lines(tau_seq,allsd[,8],col='brown')

legend(0, 15,
       legend=c("A", "B","C","D","E","F","G","H"),
       col=c("black", "green","red","blue","yellow","purple","orange","brown"), lty=1, cex=0.8)
```



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.

Chapter 8

Blocks

8.1 Equations

Here is an equation.

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{8.1}$$

You may refer to using `\@ref(eq:binom)`, like see Equation (8.1).

8.2 Theorems and proofs

Labeled theorems can be referenced in text using `\@ref(thm:tri)`, for example, check out this smart theorem 8.1.

Theorem 8.1. *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.

8.3 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

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.