The Income Difference Change between Gender from 2014 to 2017

Ruixie Fang*

balloon0315@gmail.com

Department of Math and Stat, Georgia State University, Atlanta, GA, 30302

1 Introduction

Bayesian Analysis is a formal method for combining prior beliefs with observed information. It can fit very realistic but complicated models. In this project Markov Chain Monte Carlo method will be used to do Bayesian data analysis and to predict the relationship between male income and female income in the United State from 2014 to 2017. The data sets of this project are median earnings of full-time and year-round workers from 2014 to 2017 by gender and detailed occupation. The data sets are downloaded from www.census.gov. For each year, we use the difference (in thousand) between men's earnings and women's earnings for each occupation as our observations. And there are 60 observations for years respectively.

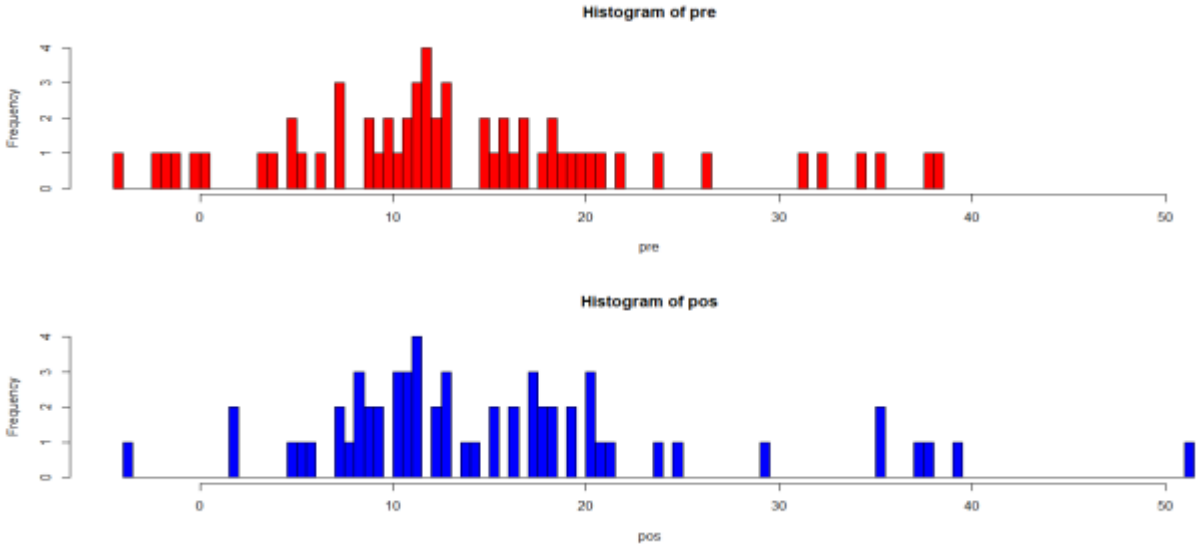
The question in this analysis is thus the following:

- Is the mean of differences in year 2017 greater than the mean of differences in year 2016? If so, how much?
- Is the mean of differences in year 2015 greater than the mean of differences in year 2014? If so, how much?

2 Data Analysis

2.1 Comparison of differences in 2016 and 2017

The histograms of differences in 2016 and 2017 are given below.



The conclusion can not be figured out directly from histograms. Further analysis is needed. So we may assume that the difference in each year follows normal distribution $N(\mu, \sigma^2)$. Let us assume that the difference in 2016 come from $N(\mu_1, \sigma_1^2)$ and the difference in 2017 come from $N(\mu_2, \sigma_2^2)$. So our parameter vector is $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$.

We will construct a likelihood function and a prior and then multiply them together to get posterior $P(\theta|Data) = P(Data|\theta) * P(\theta)$.

Let *x* denote the difference in 2016 and let *y* denote the difference in 2017. Then the likelihood is given by the formula

$$P(Data|\theta) = P(x|\theta)P(y|\theta) = \prod_{i=1}^{60} N(x_i|\mu_1, \sigma_1^2) \prod_{i=1}^{60} N(y_i|\mu_2, \sigma_2^2)$$

where the second equation holds since x_i and y_j are independent for $i = 1, 2, \ldots, 60$ and $j = 1, 2, \ldots, 60$. We can write the likelihood function according to the above formula.

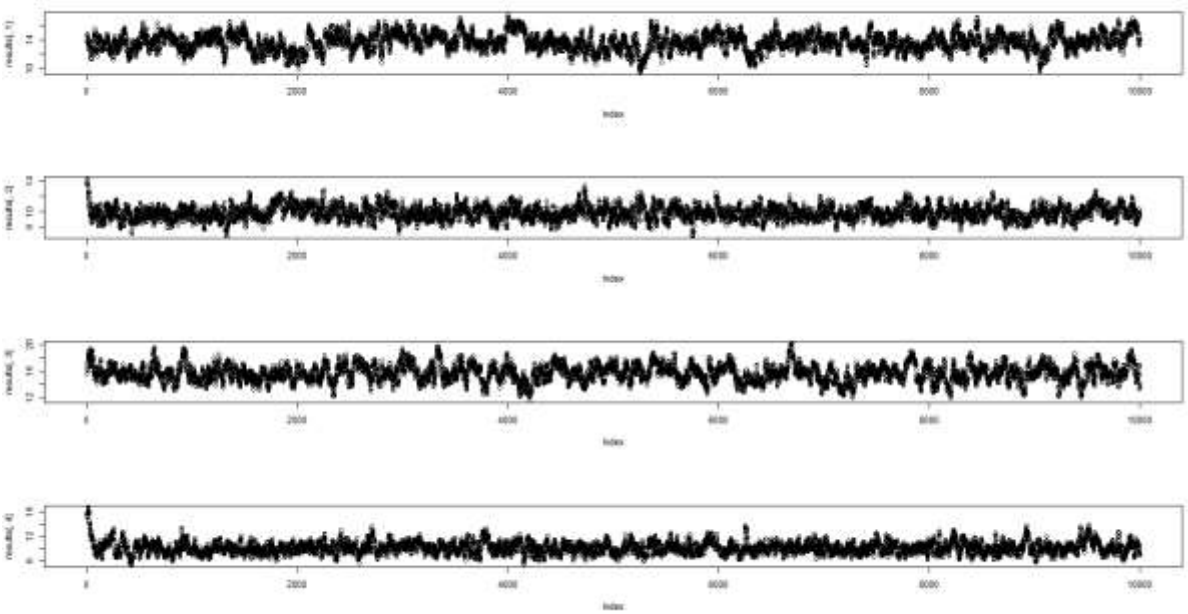
Then we need to decide on a prior density for $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$, and write it as a function. The basic idea is simply to choose priors for the μ_1 and μ_2 that do not seem to prejudge which is bigger, and to choose priors that are sort of uniformize over all even remotely plausible values, based on the purpose of letting the data produce any interesting features in the posterior distribution. We choose the distribution of μ_1 as $N(13.7, 13.7^2)$ and choose the distribution of μ_2 as $N(15.65, 15.65^2)$, where 13.7 is the sample mean of *x* and 15.65 is the sample mean for *y*. In order to decrease the influence of the prior and get an flatter distribution, we choose the distributions like this, otherwise the prior will control the

posterior, and the conclusion might be misleading. Then we choose the distribution of σ_1 as exponential with mean 13.7 and choose the distribution of σ_2 as exponential with mean 15.65. Thus all 4 variables are independent. Then the prior is given by

$$P(\mu_1, \sigma_1, \mu_2, \sigma_2) = P(\mu_1)P(\sigma_1)P(\mu_2)P(\sigma_2)$$
$$= \frac{1}{\sqrt{2\pi}}(13.7^2)^{-0.5} e^{-\frac{(\mu_1 - 13.7)^2}{2 * 13.7^2}} \frac{1}{13.7} e^{-\frac{\sigma_1}{13.7}} \frac{1}{\sqrt{2\pi}} (15.65^2)^{-0.5} e^{-\frac{(\mu_2 - 15.65)^2}{2 * 15.65^2}} \frac{1}{15.65} e^{-\frac{\sigma_2}{15.65}}$$

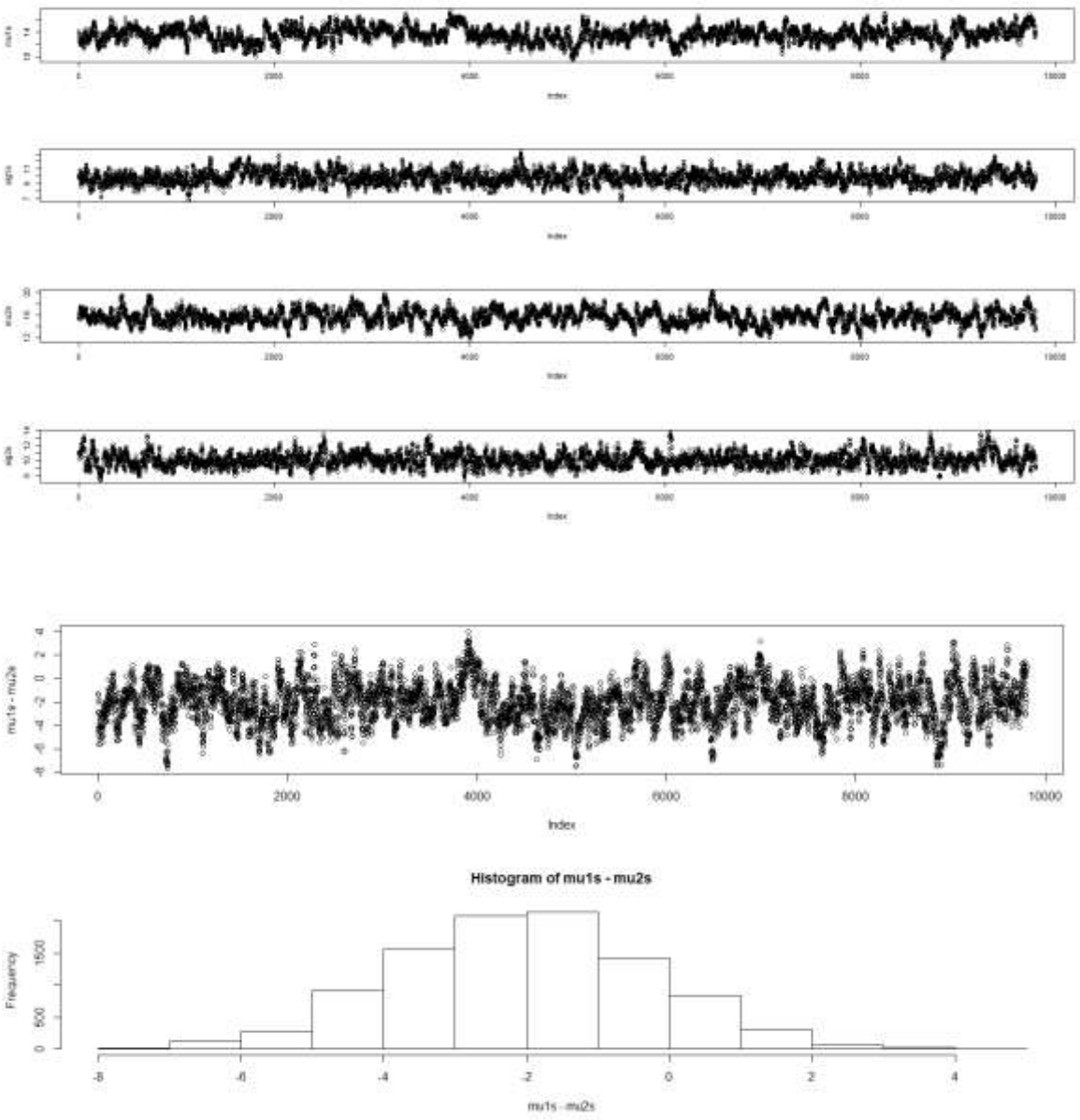
Next, we multiply prior by likelihood to get the posterior. With a large sample size, the accuracy of the conclusion might also be increased. Thus we will run a Markov chain using Metropolis for 10000 iterations to simulate a sample. First we choose a starting value $\theta_0 = (13.7, 13.7, 15.65, 15.65)$. Given a current state, we need to decide on a way to propose a "candidate" move, then evaluate the posterior of candidate move and the posterior of current state and then take the ratio. We will record our results in a big matrix with 4 columns and 10000 rows. The first row is the starting value and each successive row will record the next θ as we run the chain.

Now we have got a bunch of parameter vectors. Let's first take a look at how the chain ran.



It looks like the first few hundred iterations may be noticeably influenced by our starting values. Then we throw away the first 200 iterations and let the remaining be our new results.

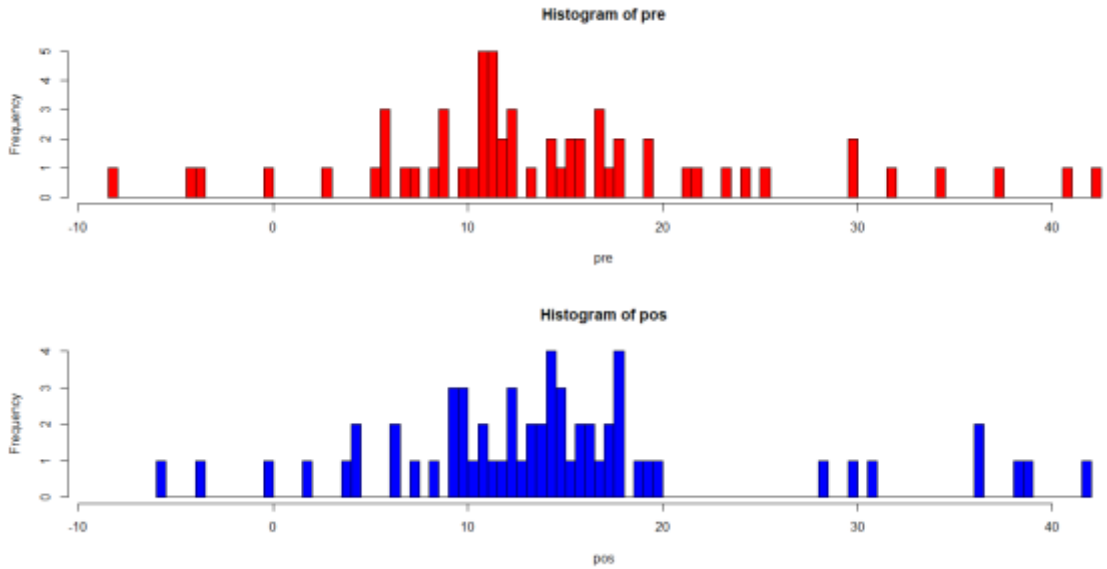
Then we could base our inferences on the new results instead of the whole results matrix.



Our estimated posterior probability that $\mu_1 - \mu_2 < 0$ is about 0.873. Thus with high probability, the mean of differences in year 2017 is greater than the mean of difference in year 2016.

2.2 Comparison of differences in 2014 and 2015

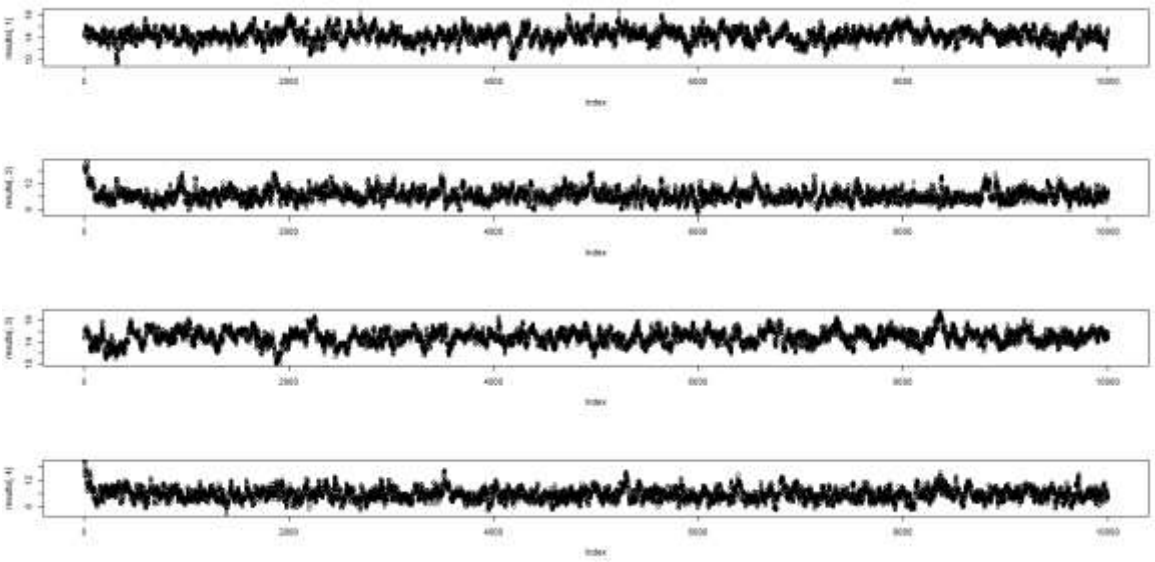
The histograms of differences in 2014 and 2015 are given below.



The conclusion is not so obvious from histograms. We may assume that the difference in each year follows normal distribution $N(\mu, \sigma^2)$. Let us assume that the difference in 2014 come from $N(\mu_1, \sigma_1^2)$ and the difference in 2015 come from $N(\mu_2, \sigma_2^2)$. So our parameter vector is $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$.

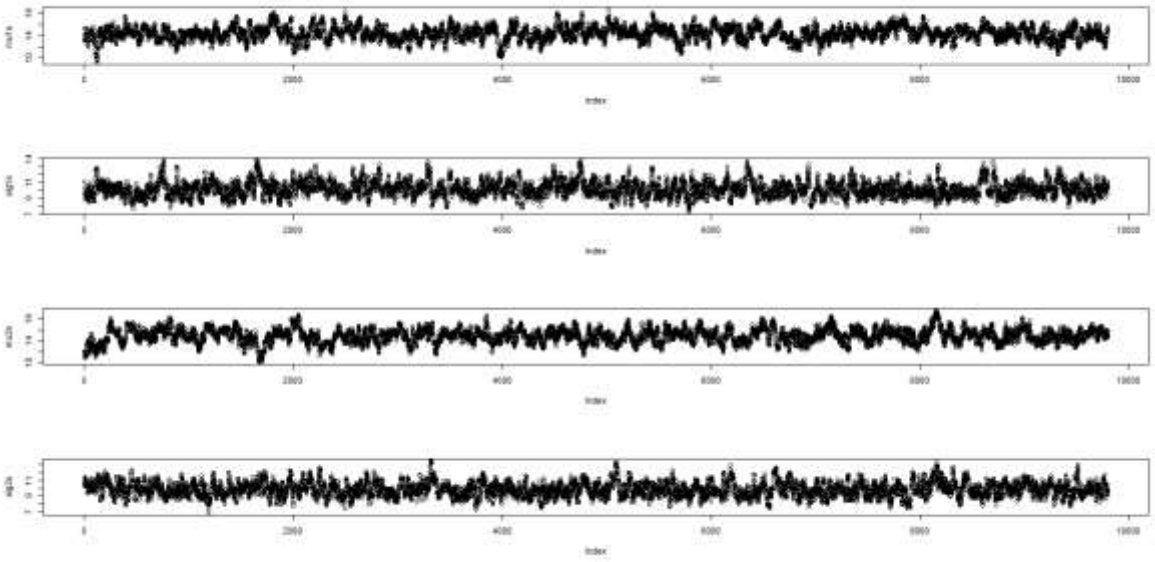
To get posterior, we still need to construct a likelihood function and a prior and then multiply them together. The likelihood function we use is exactly the same as the likelihood function in the previous subsection. To decide the prior density for $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2)$, We choose the distribution of μ_1 as $N(14.51, 14.51^2)$ and choose the distribution of μ_2 as $N(14.81, 14.81^2)$, where 14.51 is the sample mean in 2014 and 14.81 is the sample mean in 2015. Then we choose the distribution of σ_1 as exponential with mean 14.51 and choose the distribution of σ_2 as exponential with mean 14.81. Next, we multiply prior times likelihood to get the posterior .

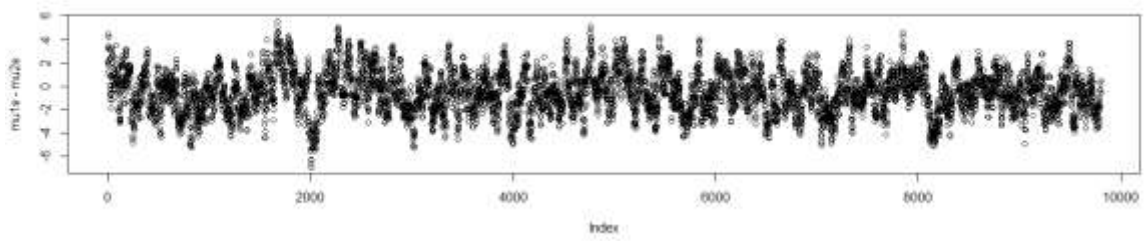
We still run a Markov chain using Metropolis for 10000 iterations to simulate a sample. This time we choose a starting value θ_0 = (14.51, 14.51, 14.81, 14.81). The results are shown below.

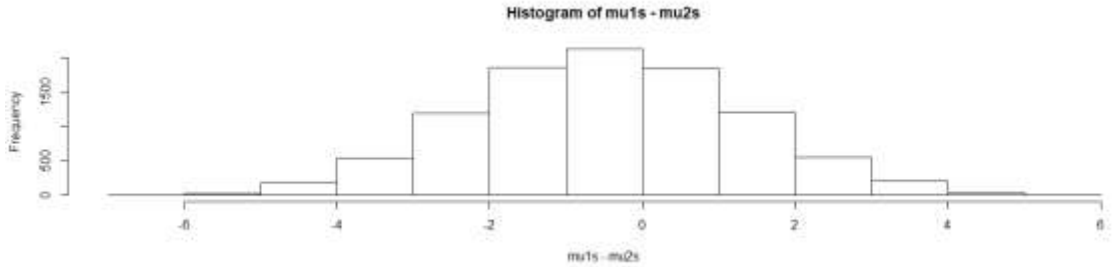


It looks like the first few hundred iterations may be noticeably influenced by our starting values. Then we throw away the first 200 iterations and let the remaining be our new results.

Then we could base our inferences on the new results instead of the whole results matrix.







Our estimated posterior probability that $\mu_1 - \mu_2 < 0$ is about 0.61. Thus with high probability, the mean of differences in year 2015 is greater than the mean of difference in year 2014.

3 Conclusion

According to our discussion in section 2, we get the conclusion that

- The mean of differences in year 2017 is greater than the mean of differences in year 2016.
- The mean of differences in year 2015 is greater than the mean of differences in year 2014.

Thus overall, the difference between male income and female income was increasing.

4 Reference

Dr. Jing Zhang's lecture notes.

5 R Code

[2.1]

```
dat=read.table(file="earning1.csv", header=T, sep=',')
summary(dat)
dif=(dat[,2]-dat[,3])/1000
pre=dif[dat[,1]==2016]
pos=dif[dat[,1]==2017]
mean(pre)
mean(pos)

###draw a picture
xlim=c(min(dif),max(dif))
par(mfrow=c(2,1))
hist(pre,100,col="red",xlim=xlim)
hist(pos,100,col="blue",xlim=xlim)

###likelihoodfunction
lik=function(th){
  mu1=th[1];sig1=th[2];mu2=th[3];sig2=th[4]
  prod(dnorm(pre,mean=mu1,sd=sig1))*prod(dnorm(pos,mean=mu2,sd=sig2))
}

###priorfunction
prior=function(th){
  mu1=th[1];sig1=th[2];mu2=th[3];sig2=th[4]
  if(sig1<=0|sig2<=0) return(0)
dnorm(mu1,13.7,13.7)*dnorm(mu2,15.65,15.65)*dexp(sig1,rate=1/13.7)*dexp(sig2,rate=1/15.65)
}
```

```
###posteriorfunction
post=function(th){prior(th)*lik(th)}

#Startingvalues
mu1=13.7;sig1=13.7;mu2=15.65;sig2=15.65
th0=c(mu1,sig1,mu2,sig2)

#Here is what does the MCMC (Metropolis method):
nit=10000
results=matrix(0,nrow=nit,ncol=4)
th=th0
results[1,]=th0
for(it in 2:nit){
  cand=th+rnorm(4,sd=.5)
  ratio=post(cand)/post(th)
  if(runif(1)<ratio)th=cand
  results[it,]=th
}

#Take a peek at what we got
edit(results)
par(mfrow=c(4,1))
plot(results[,1])
plot(results[,2])
plot(results[,3])
plot(results[,4])

res=results[201:10000,]
mu1s=res[,1]
sig1s=res[,2]
mu2s=res[,3]
sig2s=res[,4]
```

```
par(mfrow=c(4,1))
plot(mu1s)
plot(sig1s)
plot(mu2s)
plot(sig2s)
par(mfrow=c(2,1))
plot(mu1s-mu2s)
hist(mu1s-mu2s)
mean(mu1s-mu2s<0)
```

[2.2]

```
 dat=read.table(file="earning2.csv", header=T, sep=',')
summary(dat)
dif=(dat[,2]-dat[,3])/1000
pre=dif[dat[,1]==2014]
pos=dif[dat[,1]==2015]
mean(pre)
mean(pos)

###draw a picture
xlim=c(min(dif),max(dif))
par(mfrow=c(2,1))
hist(pre,100,col="red",xlim=xlim)
hist(pos,100,col="blue",xlim=xlim)

###likelihoodfunction
lik=function(th){
  mu1=th[1];sig1=th[2];mu2=th[3];sig2=th[4]
  prod(dnorm(pre,mean=mu1,sd=sig1))*prod(dnorm(pos,mean=mu2,sd=sig2))
}

###priorfunction
prior=function(th){
```

```
 mu1=th[1];sig1=th[2];mu2=th[3];sig2=th[4]
 if(sig1<=0|sig2<=0) return(0)

dnorm(mu1,14.51,14.51)*dnorm(mu2,14.81,14.81)*dexp(sig1,rate=1/14.51)*dexp(sig2,rate=
1/14.81)
}

###posteriorfunction
post=function(th){prior(th)*lik(th)}

#Startingvalues
mu1=14.51;sig1=14.51;mu2=14.81;sig2=14.81
th0=c(mu1,sig1,mu2,sig2)

#Here is what does the MCMC (Metropolis method):
nit=10000
results=matrix(0,nrow=nit,ncol=4)
th=th0
results[1,]=th0
for(it in 2:nit){
 cand=th+rnorm(4,sd=.5)
 ratio=post(cand)/post(th)
 if(runif(1)<ratio)th=cand
 results[it,]=th
}

#Take a peek at what we got
edit(results)
par(mfrow=c(4,1))
plot(results[,1])
plot(results[,2])
plot(results[,3])
plot(results[,4])
```

```
res=results[201:10000,]
mu1s=res[,1]
sig1s=res[,2]
mu2s=res[,3]
sig2s=res[,4]

par(mfrow=c(4,1))
plot(mu1s)
plot(sig1s)
plot(mu2s)
plot(sig2s)
par(mfrow=c(2,1))
plot(mu1s-mu2s)
hist(mu1s-mu2s)
mean(mu1s-mu2s<0)
```