**Project**

**LSTAT2130 : Introduction to Bayesian Statistics**

Philippe Lambert

Hortense Doms

Year 2021-2022



Jessica De Rongé - 48081600

Didier Termont - 08741301

Antoine Visschers - 41922100

Table of Content

[Question 1: 1](#_Toc104500052)

[Question 2 2](#_Toc104500053)

[Question 3 4](#_Toc104500054)

[*Question 3.A* 4](#_Toc104500055)

[*Question 3.B* 5](#_Toc104500056)

[*Question 3.C* 5](#_Toc104500057)

[Question 4 6](#_Toc104500058)

[Question 5 9](#_Toc104500059)

[*Question 5.A* 9](#_Toc104500060)

[*Question 5.B* 9](#_Toc104500061)

[*Question 5.C* 10](#_Toc104500062)

[Appendix 11](#_Toc104500063)

[*Appendix A* 11](#_Toc104500064)

[*Appendix B* 13](#_Toc104500065)

[*Appendix C* 16](#_Toc104500066)

# Question 1:

The objective of the survey was to study the profile of cannabis’ users. Because the use of cannabis is not necessarily legal everywhere and the overall negative reputation of users in today’s society, researchers discover that asking the question “Have you recently used cannabis?” made people uneasy to answer. Thus, the interviewers created a process which aims to make people fill at ease with answering the question. The process consists of making the interviewees roll two dices and in function of the result they would either answer truthfully or answer “yes” regardless of their habits concerning cannabis’ use. So, if the respondent got a double six, he would have to answer “yes” and if not, he would answer truthfully. Before developing our answer to the first question, a precision should be made. When the word probability is used in this project, it actually refers to the degree of plausibility that the event studied is true.

Thus, the proportion of people who answered “yes” does not correspond to the proportion of people who do smoke cannabis in the population. To calculate g the probability of a random person would answer yes in the survey, the law of total probability is used (knowing that the probability of cannabis’ users in the population is π).

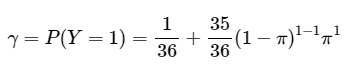
Y corresponds to the answer to the question “Have you smoked cannabis?” (yes=1) and D is the event that a participant “rolled a double 6”. From the law of total probability, we can write:

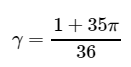
P(Y) = P(Y|D) x P(D) + P(Y|Dc) x P(Dc)

Where:

* P(D) = 1/36
* P(Dc) = 1-P(D) = 35/36
* P(Y=1|D) = 1 Indeed, if we know that the respondent got a double six, we can affirm that he will answer yes to the question and thus, we can deduce that P(Y=0|D) = 0.
* Y|Dc ~ Ber(π)

The probability g, the probability that an interviewee answered “yes”, is determined as follows:





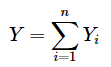
It is thus possible to conclude that ϒ is following a Bernoulli distribution with parameter g:

Une image contenant texte

Description générée automatiquement

# Question 2

Y is the random variable corresponding to the number of “yes” answered in the survey. We can find its distribution given what we have found about ϒ in the previous section.



Where

Une image contenant texte

Description générée automatiquement

Thus, given g, Y follows a Binomial distribution with parameters n (number of participants in the survey) and g.

Y ∼ Bin (n, g)

Having discovered the distribution of Y, the next step is to define the posterior distribution of p. The Bayes’ theorem is used to find this distribution:



∝Likelihood x Prior

Thus, to compute the posterior distribution, the prior distribution of p is needed, as well as the likelihood which is based on the observed data of the survey. Starting with the prior, as no prior information is given about the proportion of cannabis users in the population, a non-informative prior distribution has been chosen. Therefore, p ∼Uni(0,1) which is a special case of a beta distribution with parameter a=1 and b=1. Thus:

P(p) ∝ 1

The likelihood P(Y|p), as we know that Y, given g, follows a binomial distribution and that g is a function of p, it is possible to deduce that the likelihood is as follows:

P(Y|p) ∝gy (1-g)n-y

Une image contenant texte

Description générée automatiquement

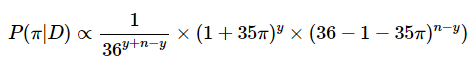
Where n is the number of people in the sample and y is the number of successes (here, understood as answering yes to the question “have you recently smoked cannabis?”).

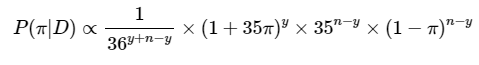
Since the prior distribution and the likelihood have been both computed, it is possible to find the posterior distribution:

Une image contenant texte

Description générée automatiquement

This expression is simplified by following those different steps which consist in bringing out a constant.





Thus, the following expression is the simplified version of the posterior distribution up to a multiplicative constant.

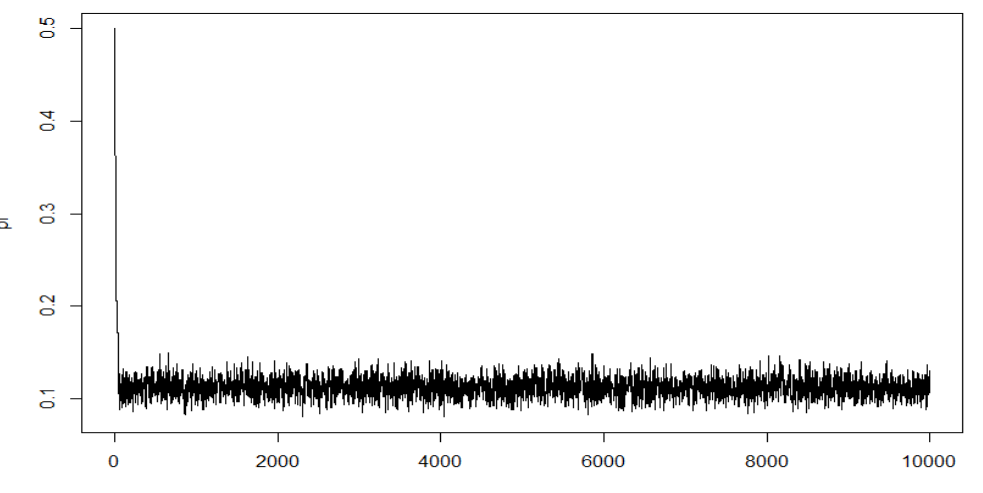


# Question 3

## *Question 3.A*

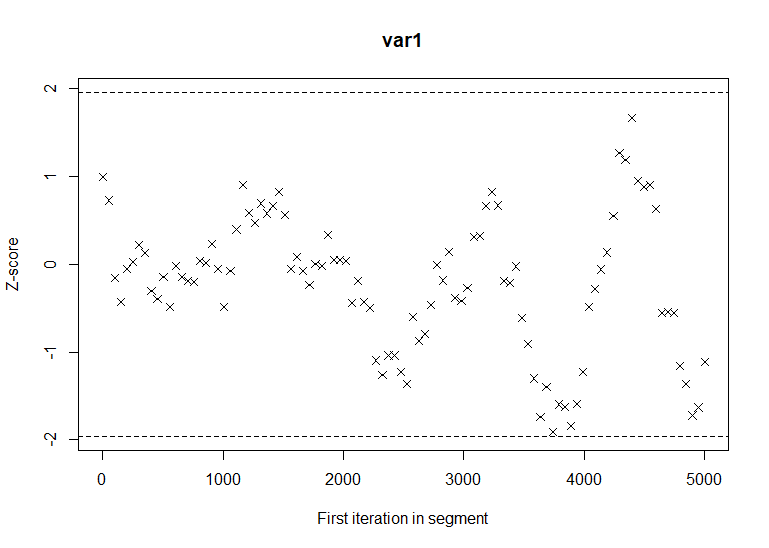
The Metropolis algorithm has been computed on the R software based on the logarithm of the posterior distribution found in the previous question (see appendix A). The objective of the algorithm is to produce a random sample of p. The starting value chosen for p is 0.5. Different values of standard deviation have been tested. The choice made on this value is based on the acceptance rate. The literature believes that a suitable standard deviation should be around 0.4 in a univariate case. However, usually a value of acceptance rate included in the range 0.3-0.5 is accepted. Thus, the chosen value is 0.02 which gives us the acceptance rate of 0.5. However, in this project, choosing the value for the standard deviation leads sometimes to produce an error in R which would indicate that NaN were produced. Using 0.02 was a value that did not create such a problem, and which led to an acceptable value of acceptance rate.

The random sample produced by the algorithm is not independent. There is auto-correlation. Thus, the sample size is calculated to find out how much information is included in the sample in terms of independent variables. Having calculated the effective size, the information given by the random sample gives as much as the information of 1585 independent values of p.



Concerning the convergence of the generated sample, the only way to ensure convergence is asymptotically through the Metropolis algorithm. As you can see on the graph, it seems that it converges quite quickly, only after 200 iterations it can be seen that there is convergence. However, there are different tools that are helpful to study convergence. However, none of these tools are able to prove convergence. They may be used to see if there is no convergence happening. The first one is the Gelman-Rubin diagnostic (also known as the R statistic) which is based on the study of the between and within variance of different chains. The R statistic is a ratio between the overall variance estimate and the within variance. Having a statistic close to 1 is a necessary condition to have convergence. Different chains with different starting values have been created to be able to conduct the Gelman-Rubin diagnostic and the statistic has been computed and gives us a R statistic equal to 1.

The second tool used to analyze convergence is the Geweke diagnostic which consists of testing if two means are equal. In the metropolis algorithm, the sample is divided in two samples to compute the p value and see if it is different. This division and test are done multiple times in the sample. Having all the different statistics being between -1.96 and 1.96 is again a necessary condition to have convergence. As seen in the graph, all the statistic observed are in between the two thresholds.



## *Question 3.B*

To find the plausible value of p, a 95% Bayesian credible interval, which will include 95% of the posterior probability for p, is measured on the sample for p. The interval chosen is the smallest interval. In this case, it is the Highest Probability Density interval (HPD interval) that is retained in this analysis (see appendix A). The burn-in part of the sample must be removed. The plausible value of p found through the previous result are 0.093 and 0.13. Thus, this interval does not include the value 0, so it can be said that p is different from 0.

lower upper

p 0.09302349 0.1336198

## *Question 3.C*

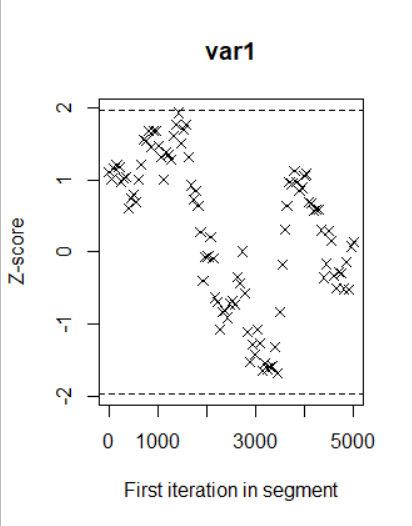
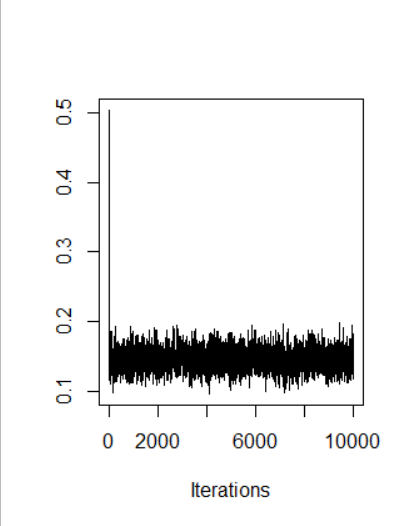
The posterior probability that among 20-59 year old, there is at least 10% of cannabis’ users is equal to 89.54% (see appendix A).

# Question 4

The metropolis algorithm has been calculated for both women and men separately (appendix B). The value of the standard deviation has again been chosen based on the acceptance rate. The same starting value as for the question 3 has been used. The answer to this question is separated in three different sections. The first one is consecrated to the study of p for the male population. The second one is dedicated to the study of the female population and the proportion p of cannabis’ users in this population. The third and last part focuses on studying the difference of proportion of cannabis’ users between men’s population and women’s population (d).

1. **Men**

The chosen value is 0.033 which gives us the acceptance rate of 0.5. Having calculated the effective size, the information given by the random sample gives as much as the information of 2259 independent values of p. The traceplot shows that the value found by the algorithm converges quite quickly. The Geweke diagnostic has also been tested and as you can see in the plot, no value of z computed through the test is higher or lower than 1.96 and -1.96 respectively. The Geweke plot tells us that it does not detect any problem of convergence in the sample produced but it is not sufficient to say that convergence does occur. The traceplot shows us asymptotic convergence which is the only guarantee that convergence happens in this case.



A 95% Bayesian credible interval has been computed to find the plausible value of p for the men and gives us an interval going from 0.11 to 0.17. This is the result from the HPD interval.

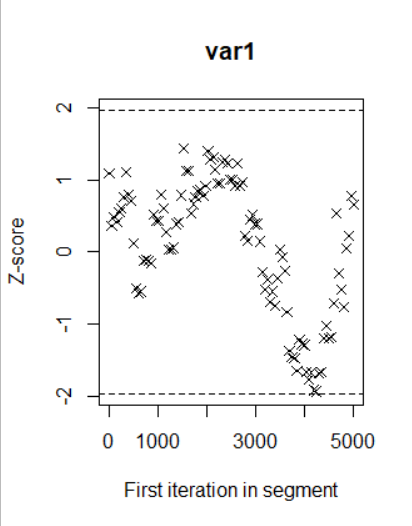
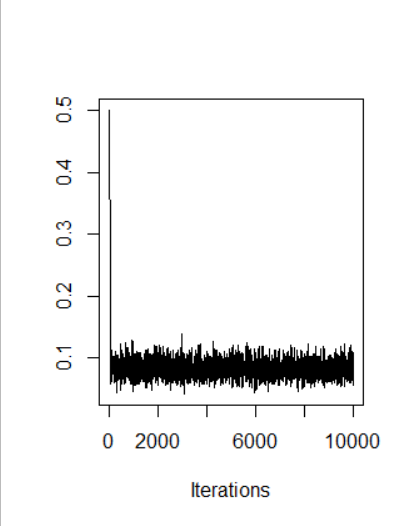
lower upper

p 0.1139749 0.1766558

The posterior probability that among 20-59 years old men, there is at least 10% of cannabis’ users is equal to 99.91%

1. **Women**

The value of the standard deviation chosen is 0.024 which gives us the acceptance rate of 0.53. The acceptance rate is a little higher than what is considered an acceptable rate. However, when the value was increased, there were some problems of convergence found in the Geweke plot. The traceplot shows that the value found by the sample converges quite quickly. This asymptotic convergence showed in the plot is what is supposed to ensure that convergence does happen. The Geweke diagnostic is used to see if there is no convergence, however it cannot prove that convergence does occur. The Geweke plot shows no value of z computed through the test is higher or lower than 1.96 and -1.96 respectively. Having calculated the effective size, the information given by the random sample gives as much as the information of 1987 independent values of p.



The plausible value has been computed through a 95% Bayesian credible interval. The HPD interval starts at 0.057 and finishes at 0.107.

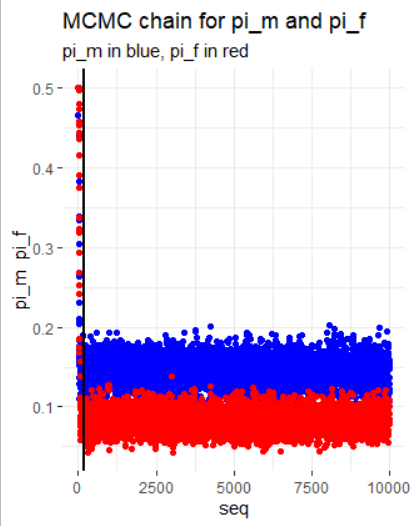
lower upper

p 0.05738883 0.1077418

The posterior probability that among 20-59 year old women, there is at least 10% of cannabis’ users is equal to 10.3%

1. **Delta (d)**

The graph below plots both the sample of the value of p for men (in blue) and the value of p for women (in red). The plot shows that both random samples are converging quite quickly (in the first 200 iterations, vertical black line). The graph also seems to inform us that there is a possibility that the men (in black in the graph) have higher value of p than women (in red). Indeed, the two chains are not covering the exact same area of possible value, the red chains covering lower values than the black chains.



To test this difference of value, delta d (corresponding to p male minus p female) has been calculated. The plausible values of delta have been computed through the computation of the HPD interval. The 95% Bayesian credible interval lower born is equal to 0.025 and the higher born is equal to 0.1. As the fact that the interval does not include 0 and is positive, it would mean that men have a higher proportion of cannabis’ smokers than woman. Moreover, if we compare the plausible values found for men and women, the interval for men contained higher values than the one for the women.

lower upper

d 0.02111185 0.1016545

To further confirm this result, the probability that delta is bigger than 0 has been calculated and is equal to 99.9%. Thus, it is possible to conclude that proportionally less women than men have recently smoked cannabis.

# Question 5

## *Question 5.A*

Two different logistics regressions have been computed using JAGS through R (see appendix C). The first one evaluates the influence of the age of men on the probability of smocking cannabis, the two estimated coefficient are a0 and a1 and the model is the following:

Une image contenant texte, horloge, antenne

Description générée automatiquement

Where hx is equal to a0 + a1 (x - 40) and where x is the variable age and px is the probability of being a cannabis’ user.

The second model analyzes the same relationships between using cannabis and the age. However, this time only the women’s population is taken into account, the two estimated coefficient are thus b0 and b1 and the model is the same as previously. The only difference is that hx is now equal to b0 + b1 (x - 40), where x is the variable age and px is the probability of being a cannabis’ user.

Notice that in the code, only one logistic regression has been computed. However, the logistic regression in the code is composed of the two different expressions, the first one in terms of a for the men and the second one in terms of b for the women. The code has been made so that, if it is a man, only the first expression will be computed and the second one will be canceled by being null and if it is a woman, only the second expression will be computed and the first one will be equal to zero. Thus, in practice, it is as if two different regressions have been computed.

## *Question 5.B*

The value of a1, b1 as well as d (which correspond to a1-b1) have been computed:

Mean SD

alpha1 -0.05573 0.01325

beta1 -0.02111 0.01531

delta -0.03463 0.02026

The plausible values for the parameters have also been calculated through a 95% Bayesian credible interval for both chains. The HPD interval gives the following results:

lower upper

alpha1 -0.08197535 -0.030092314

beta1 -0.05080021 0.009399152

delta -0.07424458 0.005889047

lower upper

alpha1 -0.08183165 -0.030493625

beta1 -0.05162156 0.008117775

delta -0.07440068 0.005093041

The difference between the two chains is different starting values. However, the plausible values for both chains are very similar which gives more confidence in the study of the different parameters.

Concerning a1, both intervals do not include zero and are similar. Thus, it can be said that the logit probability that a man older than 40 years old smokes will decrease. However, when he is younger than 40 years old, the logit probability will rise as he grows older until he is 40.

Concerning b1, as both intervals include the value 0. It is not possible to reject the null hypothesis, that b1 is equal to 0. Thus, we cannot conclude that for the women population, the age plays a role to explain if a person is a cannabis’ user.

Regarding d, the plausible value of both intervals includes 0. Thus, the null hypothesis cannot be rejected. Thus, in conclusion the data does not permit to say that d is different from 0. Following this conclusion, it is not possible to say that b1 and a1 are different.

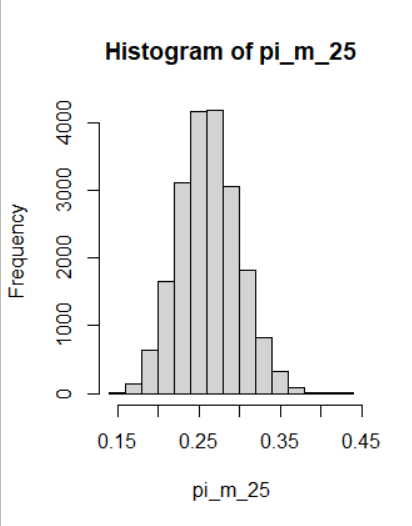
## *Question 5.C*

The value of h found for a 25 year old man is

Mean SD

etha\_m\_25 -1.043304 0.1899575

Having discovered h, the probability than a 25 year old man being a recent cannabis’ user can be computed using the inverse of the logit function which gives us the result that the proportion of users is 26.23%. The distribution of the posterior probability that a 25 year old man is a recent cannabis’ user can be visualised as follows:



# Appendix

## *Appendix A*

################################################################################

################################################################################

# Question 3

################################################################################

################################################################################

# Question 3.A

##The logarithm of the posterior distribution found in the previous question

log.f.2 <- function(pi,N,yplus) {

(N-yplus)\*log(1-pi)+yplus\*log(1+35\*pi)

}

## The Metropolis algorithm

metropolis <- function(pi0,M, sd.prop, lpost, N, yplus) {

pi = numeric(M)

pi[1]=pi0

n.accept = 0

for(i in 2:M) {

pi.prop = pi[i-1]+rnorm(1,0,sd.prop)

prop = min(1,exp(lpost(pi.prop,N,yplus) - lpost(pi[i-1],N,yplus)))

accept = (runif(1) <= prop)

if(accept) {

n.accept = n.accept+1

pi[i] = pi.prop

} else {

pi[i] = pi[i-1]

}

}

accept.rate = round(n.accept/(M-1),2)

cat("Acceptance rate: ",accept.rate,"\n")

return(pi)

}

## Running of the metropolis algo on the full population

## acceptance rate in the range 0.3-0.5 is accepted

## first chain

M <- 10000

pi = metropolis(

pi0=0.5,M,sd=0.02,

lpost=log.f.2,N=nrow(cannabis),yplus=sum(cannabis$y))

## second chain

pi\_2 = metropolis(

pi0=0.75,M,sd=0.02,

lpost=log.f.2,N=nrow(cannabis),yplus=sum(cannabis$y))

## The effectif size

effectiveSize(mcmc(pi[200:M]))

## Convergence

traceplot(mcmc(pi))

##1. the Gelman-Rubin diagnostic

gelman.diag(list(mcmc(pi),mcmc(pi\_2)))

##2. the Geweke diagnostic

geweke.diag(mcmc(pi))

geweke.plot(mcmc(pi),nbins=100)

# Question 3.B

quantile(pi[200:M],c(0.025,0.975))

HPDinterval(mcmc(pi[200:M]))

# Question 3.C

nrow(subset(data.frame(pi[200:M]),pi>=0.1))/(M-200)

## *Appendix B*

################################################################################

################################################################################

# Question 4

################################################################################

################################################################################

# 1. Men

## Metropolis for the male subset

can\_m =subset(cannabis,male==1)

pi\_m = metropolis(

pi0=0.5,M,sd.prop=0.033,

lpost=log.f.2,N=nrow(can\_m),yplus=sum(can\_m$y))

## The effectif size

effectiveSize(mcmc(pi\_m[200:M]))

## Convergence

traceplot(mcmc(pi\_m))

## The Geweke diagnostic

geweke.diag(mcmc(pi\_m))

geweke.plot(mcmc(pi\_m),nbins=100)

## 95% credible interval for pi\_m

quantile(pi\_m[200:M],c(0.025,0.975))

HPDinterval(mcmc(pi\_m[200:M]))

## The posterior probability that among 20-59 years old men,

##there is at least 10% of cannabis’ users

mean(pi\_m>0.1)

#2. Women

## Metropolis for the female subset

can\_f=subset(cannabis,male==0)

pi\_f = metropolis(

pi0=0.5,M,sd.prop=0.024,

lpost=log.f.2,N=nrow(can\_f),yplus=sum(can\_f$y))

## The effectif size

effectiveSize(mcmc(pi\_f[200:M]))

## Convergence

traceplot(mcmc(pi\_f))

## The Geweke diagnostic

geweke.diag(mcmc(pi\_f))

geweke.plot(mcmc(pi\_f),nbins=100)

## 95% credible interval for pi\_f

quantile(pi\_f[200:M],c(0.025,0.975))

HPDinterval(mcmc(pi\_f[200:M]))

## The posterior probability that among 20-59 years old women,

##there is at least 10% of cannabis’ users

mean(pi\_f>0.1)

# 3.Delta

##the value of pi for men

f.pi\_m = data.frame(pi\_m)

f.pi\_m$seq = seq(1,length(pi\_m))

##the value of pi for women

f.pi\_f = data.frame(pi\_f)

f.pi\_f$seq = seq(1,length(pi\_f))

##plots both the sample of the value of pi for men

## and the value of pi for women

ggplot() +

theme\_bw() +

theme(panel.border = element\_blank()) +

geom\_point(data=f.pi\_m,aes(x=seq,y=pi\_m),color="blue") +

geom\_point(data=f.pi\_f,aes(x=seq,y=pi\_f),color="red") +

geom\_vline(aes(xintercept=200),color="black",size=1) +

labs(

title="MCMC chain for pi\_m and pi\_f",

subtitle="pi\_m in blue, pi\_f in red",

y="pi\_m \ pi\_f")

## The plausible values of delta

delta = pi\_m[200:M] - pi\_f[200:M]

## Convergence

traceplot(mcmc(delta))

## The Geweke diagnostic

geweke.diag(mcmc(delta))

geweke.plot(mcmc(delta))

## The density of delta

densplot(mcmc(delta))

## Histogram of delta

plot(pi\_m,ylim=c(0,0.5))

points(pi\_f,col="red")

hist(delta)

## 95% credible interval for delta

HPDinterval(mcmc(delta))

## The probability that delta is bigger than 0

nrow(subset(data.frame(delta[200:M]),delta>0))/(M-200)

## *Appendix C*

################################################################################

################################################################################

# Question 5

################################################################################

################################################################################

# Question 5.A

## Logistic Regression

model <- function() {

#sample

for(i in 1:n) {

y[i] ~ dbin((1+35\*pi[i])/36,1)

logit(pi[i]) <- ((alpha0 + alpha1\*(age[i]-40))\*male[i]) + ((beta0 + beta1\*(age[i]-40))\*(1-male[i]))

#logit(pi[i]) <- alpha0 + alpha1\*(age[i]-40)

}

# prior

alpha0 ~ dnorm(0,1E-6)

alpha1 ~ dnorm(0,1E-6)

beta0 ~ dnorm(0,1E-6)

beta1 ~ dnorm(0,1E-6)

# function to monitor

delta <- alpha1 - beta1

# predict the odds for males who are 25 year old

etha\_m\_25 <- alpha0 + alpha1\*(25-40)

}

write.model(model,"cannabis\_2.bug")

cannabis.list <- with(cannabis,

list(y=y,

age=age,

male=male,

n=nrow(cannabis)))

# Create two chains with differences values

inits.list <- list(list(

alpha0=0,

alpha1=0,

beta0=0,

beta1 =0),

list(

alpha0=-4,

alpha1=-1,

beta0=-5,

beta1=-1

)

)

params <- c("alpha0","alpha1","beta0","beta1","delta","etha\_m\_25")

cannabis.model <- jags.model(

file="cannabis\_2.bug",

data=cannabis.list,

inits = inits.list ,

n.chains = length(inits.list))

update(cannabis.model,1000)

out = coda.samples(model=cannabis.model,

variable.names = params,

n.iter=20000)

out.matrix = as.matrix(out)

## Question 5.B

summary(out)

HPDinterval(out)

## Question 5.C

## mean etha\_m\_25

mean(out.matrix[,"etha\_m\_25"])

## sd etha\_m\_25

sd(out.matrix[,"etha\_m\_25"])

## The distribution of the posterior probability that

##a 25 year old man is a recent cannabis users

pi\_m\_25 = 1/(1 + exp(-out.matrix[,"etha\_m\_25"]))

hist(pi\_m\_25,freq=F)