**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 uses. So, if the respondent got a double six, he would have to answer “yes” and if not, he would answer truthfully.

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 of users in the population is π).

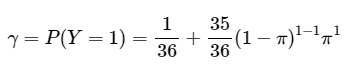
Y corresponds to the answer to the question “Have you smoked cannabis?” (yes=1) and D is the event that a participant “rolled a double 6”. From the law of total probability, we can write:

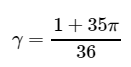
P(Y) = P(Y|D) x P(D) + P(Y|Dc) x P(Dc)

Where:

* P(D) = 1/36
* P(Dc) = 1-P(D) = 1/35
* P(X=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(X=0|D) = 0.
* Y|Dc ~ Ber(π)

The probability ϒ, the probability that an interviewee answered “yes”, is determined as follow:





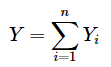
It is thus possible to conclude that X is following a Bernoulli distribution with parameter g:

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 Baye’s 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. Therefor, 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 ∝((1+35 π)/36)y (1-((1+35 π)/36))n-y

Where n is the number of people in the sample and y is the number of success (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:

P(p|Y) ∝((1+35 π)/36)y (1-((1+35 π)/36))n-y x 1

To simplify later calculation using the posterior distribution, the log-posterior is taken which is equal to

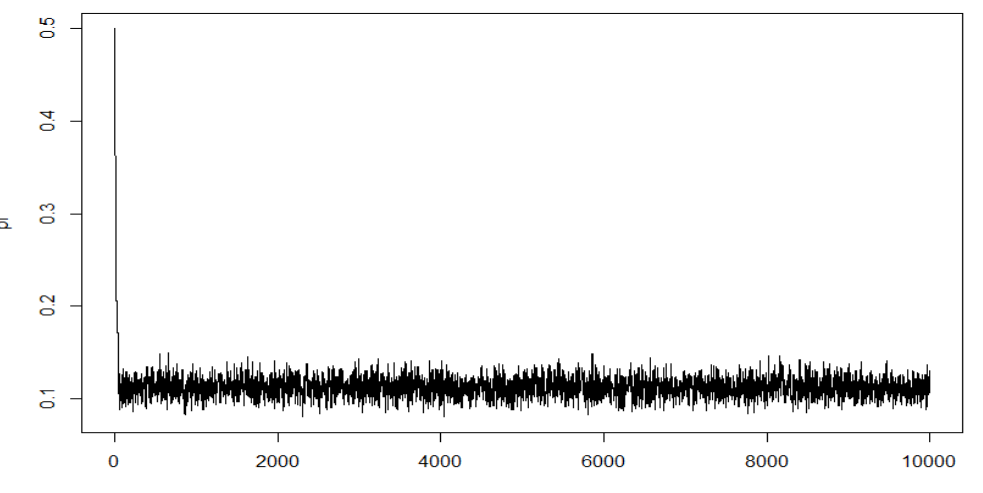
P(p|Y) ∝y log(g)+(n-y)log(1-g)

Where g is equal to (1+35p)/36

**Question 3**

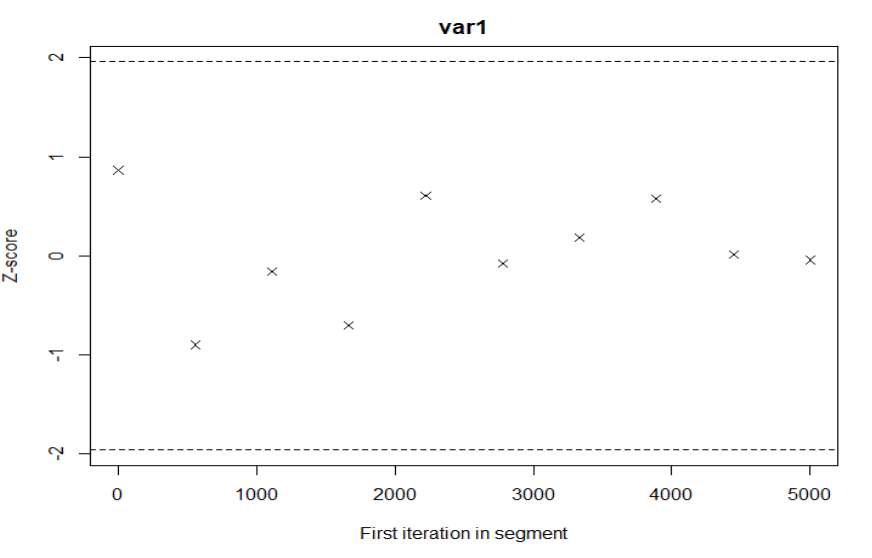
**A)**

The metropolis algorithm has been computed on the R software based on the posterior distribution found in the previous question. The objective of the algorithm is to produce a random sample of p. 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 0.028 which gives us the acceptance rate of 0.4. However, in this project, choosing the value for the standard deviation lead sometimes to produce an error in R which would indicate that NaN were produced. Using 0.028 was a value that did not create such a problem, and which led to a good value of acceptance rate. Having calculated the effective size, the information given by the random sample gives as much as the information of 295 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 consists of comparing the between and within variance of different chains. 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.



**B)**

To find the plausible value of p, a 95% credible interval is measured on the sample for pi. The burn-in part of the sample must be removed. The plausible value of p found through the previous result are 0.094 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

0.0937084 0.1333216

**C)**

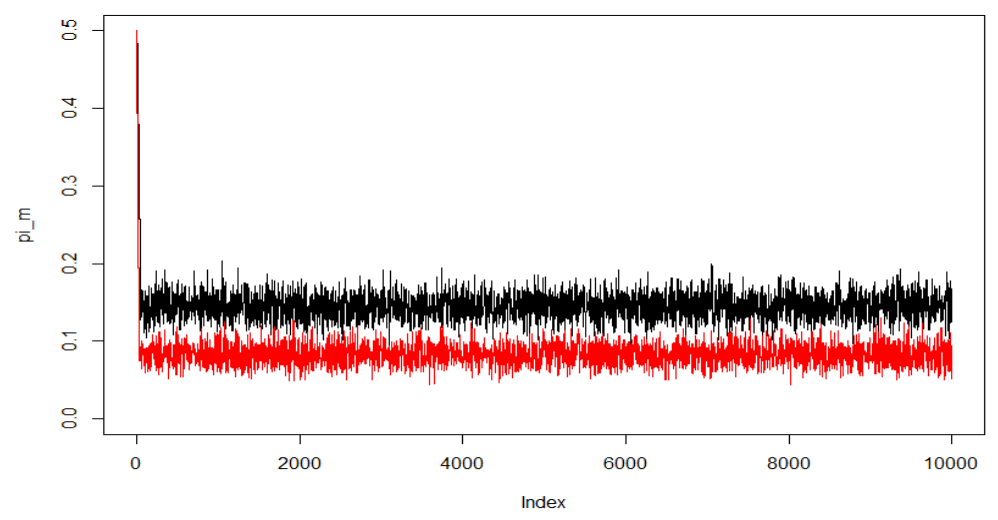
The posterior probability that among 20-59 year olds, there is at least 10% of cannabis’ users is equal to 92.62%.



* Les résultats changent en fonction de l’écart-type défini

**Question 4**

The metropolis algorithm has been calculated for both women and men separately. The value of the standard deviation has agains been chosen based on the acceptance rate. The plot shows that both random samples are converging quite quickly (probably in the first 200 iterations). 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. Then, the possible value of delta has been determined. Thus a 95% credible interval has been measured and the lower born is equal to 0.024 and the higher born is 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.

Une image contenant texte

Description générée automatiquement

To further confirm this result, the probability that delta is bigger than 0 has been calculated and is equal to 99.97%. Thus, it is possible to conclude that proportionally less women than men have recently smoked cannabis.



**Question 5**

**A)**

Two different logistics regressions have been computed using JAGS through R. The first one evaluates the influence of the age of men and 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 analyze the same relationships between using cannabis and the age. However, this time only the women 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.

**B)**

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

Mean SD

alpha1 -0.05798 0.01104

beta1 -0.01894 0.01196

delta -0.03904 0.01822

and the plausible values for the parameters have also been computed through a 95% credible interval for both chains that were computed (difference entre les deux chaines ?) are :

lower upper

alpha1 -0.06766081 -0.039818087

beta1 -0.03663591 -0.009718917

delta -0.05514141 -0.014377487

lower upper

alpha1 -0.07755705 -0.038803242

beta1 -0.03408882 0.003073788

delta -0.07348159 -0.017999085

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 smoked will decrease. However, when he is younger than 40 years old, the logit probability with rise as he grows older until he is 40.

Concerning b1, problème deux intervalles ne disent pas la même chose

Regarding d, the credible interval does not include zero which means that there exists a significant difference between men and women users of cannabis.

**C)**

The posterior distribution of the probability a 25 year old man being a recent cannabis users can be visualised in this graph. (Il faut ajouter un plot mais j’arrive pas à le faire sur R) :

An 95% credible interval has been computed to find the plausible value of such a probability:

* Il y a un truc que je comprend pas c’est pq on a des valeurs négatives pour une probabilité ?