Source code of this document:
https://github.com/cld4h/DASC_6510_Assignment/tree/master/STAT_5310_AST3

1 Question 1

Let $x_1, x_2, \ldots, x_m$ is a random sample from a Poisson(λ) distribution, where $\lambda \in \mathcal{R}$ is unknown. Determine/calculate/perform the:

1. Likelihood function,
2. Log-likelihood function,
3. Score equation,
4. Fisher's information,
5. MLE,
6. Second derivative test

$$
\begin{aligned}
L(\lambda) &= \prod_{i=1}^{m} P(X = x_i) \\
&= \prod_{i=1}^{m} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \propto e^{-m\lambda}\lambda^{\sum_{i=1}^{m} x_i} \\
l(\lambda) &= \sum_{i=1}^{m} \left(\ln\left(e^{-\lambda}\right) + \ln\left(\lambda^{x_i}\right) - \ln\left(x_i!\right) \right) \\
&= -m\lambda + \ln(\lambda) \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} \ln\left(x_i!\right) \\
U(\lambda) = \frac{\partial l(\lambda)}{\partial \lambda} &= -m + \frac{1}{\lambda} \sum_{i=1}^{m} x_i \\
I_n(\lambda) = \mathrm{Var}_\lambda\{U(\lambda)\} &= -\mathrm{E}_\lambda\{l''(\lambda)\} \\
&= -\mathrm{E}_\lambda\{-\lambda^{-2} \sum_{i=1}^{m} x_i\} \\
&= \lambda^{-2}\mathrm{E}_\lambda\left(\sum_{i=1}^{m} x_i\right) = \lambda^{-2} \cdot m\lambda \\
&= \frac{m}{\lambda} \\
\hat{\lambda}_{\mathrm{MLE}} &= \frac{\sum_{i=1}^{m} x_i}{m}
\end{aligned}
$$

$$
\forall \lambda \in \mathcal{R}\backslash\{0\}, x_i \in N, l''(\lambda) = -\lambda^2 \sum_{i=1}^{m} x_i < 0
$$

2 Question 2

If $x_1, x_2, \ldots, x_n$ is a random sample from an $N(\mu_0, \sigma^2)$ distribution, where $\sigma^2 > 0$ is unknown and μ_0 is known, then determine the MLE of σ^2.

$$
\begin{aligned}
L\left(\sigma^2|x_i,\mu_0\right) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{\left(x_i-\mu_0\right)^2}{2\sigma^2}\} \\
&= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\{-\sum_{i=1}^{n} \frac{\left(x_i-\mu_0\right)^2}{2\sigma^2}\} \\
l\left(\sigma^2|x_i,\mu_0\right) &= \sum_{i=1}^{n}\left(-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{\left(x_i-\mu_0\right)^2}{2\sigma^2}\right) \\
U\left(\sigma^2\right) = \frac{\partial l(\sigma^2,x_i,\mu_0)}{\partial\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu_0)^2 \\
\text{Solve } \frac{\partial l(\sigma^2,x_i,\mu_0)}{\partial\sigma^2} &= 0 \text{ for } \sigma^2 \\
\frac{n}{2\sigma^2} &= \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu_0)^2 \\
n\sigma^2 &= \sum_{i=1}^{n}(x_i-\mu_0)^2 \\
\therefore \hat{\sigma}^2_{\text{MLE}} &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2
\end{aligned}
$$

The second derivative of likelihood function is

$$
\begin{aligned}
\frac{\partial^2 l(\sigma^2,x_i,\mu_0)}{\partial^2(\sigma^2)} &= \frac{n}{2(\sigma^2)^2} - \frac{2}{2(\sigma^2)^3}\sum_{i=1}^{n}(x_i-\mu_0)^2 \\
&= \frac{1}{(\sigma^2)^3}(\frac{n\sigma^2}{2} - \sum_{i=1}^{n}(x_i-\mu_0)^2)
\end{aligned}
$$

when $\frac{n\sigma^2}{2} - \sum_{i=1}^{n}(x_i-\mu_0)^2 < 0 \Leftrightarrow \sigma^2 < \frac{2\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}$, the likelihood function gets local maximum value at $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu_0)^2$, when $\sigma^2 > \frac{2\sum_{i=1}^{n}(x_i-\mu_0)^2}{n}$, even though the $U(\sigma^2)$ is increasing, it is always a negative value (we have only 1 solution for the equation $U(\sigma^2) = 0$, so the likelihood function is monotonically decreasing when $\sigma^2 > \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2$, so the local maximum value is the global maximum value.

3 Question 3

Suppose that in a population of students in a course with a large enrollment, the mark, out of 100, on a final exam is approximately distributed $N(\mu, \sigma^2 = 9)$. The instructor places the prior $\mu \sim N(\mu_0 = 65, \tau_0^2 = 1)$ on the unknown parameter. A sample of 10 marks is obtained as given below.

46	68	34	86	75	56	77	73	53	64

(a) Determine the posterior mode and a 0.95-credible interval for μ. What does this interval tell you about the accuracy of the estimate?
(b) Use the 0.95-credible interval for μ to test the hypothesis $H_0 : \mu = 65$

Claim: The posterior distribution of $\{\mu|y_1, \ldots, y_n, \sigma^2\}$ follows a Normal distribution.
Proof:
The sampling distribution of $\{y_1, \ldots, y_n|\mu, \sigma^2\}$ is:

$$\begin{aligned} p(y_1, \ldots, y_n|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right). \end{aligned}$$

The prior distribution of $\{\mu|\sigma^2, \mu_0, \tau_0^2\}$ is:

$$p(\mu|\sigma^2, \mu_0, \tau_0^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\tau_0^2}\right).$$

The posterior distribution of $\{\mu|y_1, \ldots, y_n, \sigma^2, \mu_0, \tau_0^2\}$ is:

$$
\begin{aligned}
p(\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2) &= p(\mu|\sigma^2)\cdot p(y_1,\ldots,y_n|\mu,\sigma^2)\\
&\propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\tau_0^2}\right)\cdot\exp\left(-\frac{\sum_{i=1}^n(y_i-\mu)^2}{2\sigma^2}\right)\\
&= \exp\left[-\frac{1}{2}\left(\frac{(\mu-\mu_0)^2}{\tau_0^2}+\frac{\sum_{i=1}^n\left(y_i^2-2y_i\mu+\mu^2\right)}{\sigma^2}\right)\right]\\
&= \exp\left[-\frac{1}{2}\left(\frac{\left(\mu^2-2\mu\mu_0+\mu_0^2\right)}{\tau_0^2}+\frac{\left(\sum_{i=1}^n y_i^2-2\mu\sum_{i=1}^n y_i+n\mu^2\right)}{\sigma^2}\right)\right]\\
&= \exp\left[-\frac{1}{2}\left(\frac{\tau_0^2\left(\sum_{i=1}^n y_i^2-2\mu\sum_{i=1}^n y_i+n\mu^2\right)+\sigma^2\left(\mu^2-2\mu\mu_0+\mu_0^2\right)}{\sigma^2\tau_0^2}\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{(n\tau_0^2+\sigma^2)\mu^2-2\mu\left(\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0\right)}{\sigma^2\tau_0^2}\right)\right]\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{(n\tau_0^2+\sigma^2)(\mu-\frac{\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0}{n\tau_0^2+\sigma^2})^2}{\sigma^2\tau_0^2}\right)\right]
\end{aligned}
$$

Get the normalizing coefficient C for this normal distribution:

$$
\begin{aligned}
\because \int_{-\infty}^{\infty} p(\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2)d\mu &= \int_{-\infty}^{\infty} C\cdot\exp\left[-\frac{1}{2}\left(\frac{(n\tau_0^2+\sigma^2)(\mu-\frac{\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0}{n\tau_0^2+\sigma^2})^2}{\sigma^2\tau_0^2}\right)\right]d\mu = 1\\
\therefore C\sqrt{2\pi\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}}\cdot\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}}}&\exp\left[-\frac{1}{2}\left(\frac{(n\tau_0^2+\sigma^2)(\mu-\frac{\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0}{n\tau_0^2+\sigma^2})^2}{\sigma^2\tau_0^2}\right)\right]d\mu = 1\\
C &= \frac{1}{\sqrt{2\pi\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}}}
\end{aligned}
$$

$\therefore \{\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2\} \sim N(\frac{\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0}{n\tau_0^2+\sigma^2}, \frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2})$

The posterior mode (same as posterior mean) of $\{\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2\}$ is $\frac{\tau_0^2\sum_{i=1}^n y_i+\sigma^2\mu_0}{n\tau_0^2+\sigma^2} = \frac{1\times632+9\times65}{10\times1+9} = \frac{1217}{19} = 64.05263$

The posterior variance of $\{\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2\}$ is $\frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2} = \frac{9}{19} = 0.4736842$

The 0.95-credible interval for $\{\mu|y_1,\ldots,y_n,\sigma^2,\mu_0,\tau_0^2\}$ is:

$$
\begin{aligned}
&[64.05263+\text{qnorm}(0.025)\sqrt{0.4736842},\quad 64.05263+\text{qnorm}(0.975)\sqrt{0.4736842}]\\
=&[62.70369, 65.40157].
\end{aligned}
$$

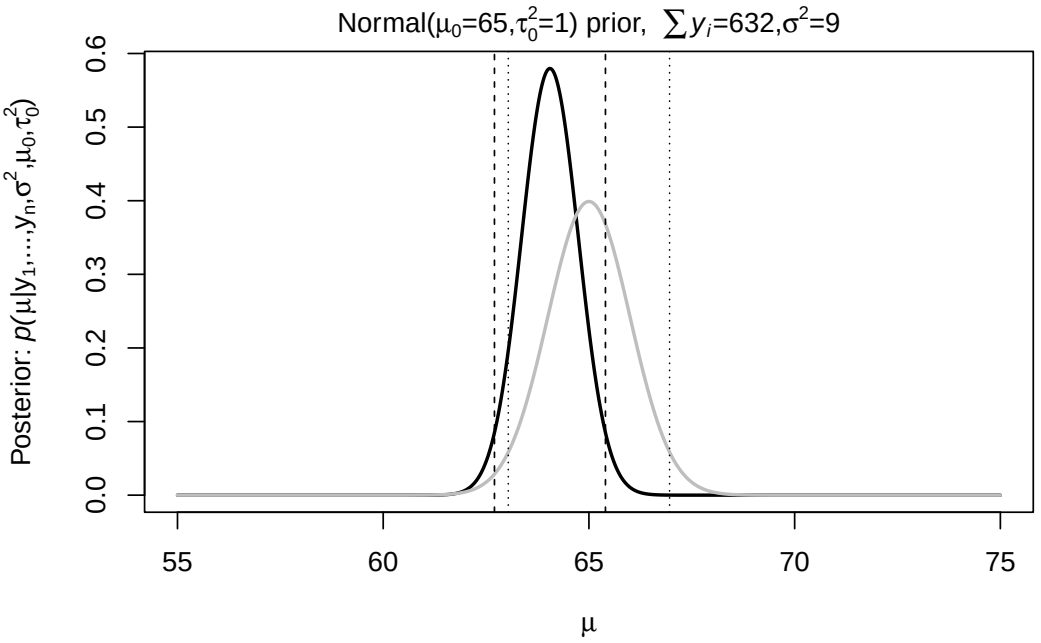

Figure 1: Prior (in gray) and Posterior (in black) distribution of μ.

As shown in Figure 1, the posterior distribution of μ is plotted in Black, and the 95% credible interval is ploted in dashed lines. The prior distribution of μ is plotted in gray, and the 95% credible interval of prior μ is plotted in dotted lines. We can see that the credible intervals is narrower for the posterior distribution compared to prior. Thus, we can say that the accuracy is improved.

Also if we use classical frequentist method, the confidence interval we get is:

$$\begin{aligned}
&\left[\overline{y} - \frac{\sigma}{\sqrt{n_y}}, \overline{y} + \frac{\sigma}{\sqrt{n_y}}\right] \\
&= \left[63.2 - \frac{3}{\sqrt{10}}, 63.2 + \frac{3}{\sqrt{10}}\right] \\
&= [62.25132, 64.14868]
\end{aligned}$$

Therefore, we have a very good example of bayesian method NOT rejecting $H_0 : \mu = 65$ (because $65 \in [62.70369, 65.40157]$) but classical frequentist method will reject the $H_0 : \mu = 65$ ($65 \notin [62.25132, 64.14868]$).

4 Question 4

Tumor counts: A cancer laboratory is estimating the rate of tumorigenesis in two strains of mice, A and B. They have tumor count data for 10 mice in strain A and 13 mice in strain B. Type A mice have been well studied, and information from other laboratories suggests that type A mice have tumor counts that are approximately Poisson-distributed with a mean of 12. Tumor count rates for type B mice are unknown, but type B mice are related to type A mice. The observed tumor counts for the two populations are

$$\mathbf{y}_A = (12, 9, 12, 14, 13, 13, 15, 8, 15, 6);$$
$$\mathbf{y}_B = (11, 11, 10, 9, 9, 8, 7, 10, 6, 8, 8, 9, 7)$$

(a) Find the posterior distributions, means, variances and 95% quantile based confidence intervals for θ_A and θ_B , assuming a Poisson sampling distribution for each group and the following prior distribution: $\theta_A \sim Gamma(\alpha_A = 120, \beta_A = 10), \theta_B \sim Gamma(\alpha_B = 12, \beta_B = 1), \pi(\theta_A, \theta_B) = \pi(\theta_A) \times \pi(\theta_B)$.
(b) Compute and plot the posterior expectation of θ_B under the prior distribution $\theta_B \sim$ Gamma$(12 \times n_0, n_0)$ for each value of $n_0 \in \{1, 2, \ldots, 50\}$. Describe what sort of prior beliefs about θ_B would be necessary in order for the posterior expectation of θ_B to be close to that of θ_A.
(c)Should knowledge about population A tell us anything about population B? Discuss whether or not it makes sense to have $\pi(\theta_A, \theta_B) = \pi(\theta_A) \times \pi(\theta_A)$.

Claim: The posterior distribution of $\{\theta|y_1, \ldots, y_n, \alpha, \beta\} \sim$ Gamma (α, β)
Proof:
Prior distribution of $\{\theta|\alpha, \beta\}$ is:

$$p(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} e^{-\beta\theta}$$

The sampling distribution of $\{y_1, \ldots, y_n|\theta\}$ is:

$$\begin{aligned} p(y_1, \ldots, y_n|\theta) &= \prod_{i=1}^{n} \frac{1}{y_i!} \theta^{y_i} e^{-\theta} \\ &= c(y_1, \ldots, y_n) \cdot \theta^{\sum_{i=1}^n y_i} e^{-n\theta} \end{aligned}$$

The posterior distribution of $\{\theta|y_1, \ldots, y_n, \alpha, \beta\}$ is:

$$
\begin{aligned}
p(\theta|y_1,\ldots,y_n,\alpha,\beta) &= p(\theta|\alpha,\beta)\cdot p(y_1,\ldots,y_n|\theta) \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)}\cdot\theta^{\alpha-1}e^{-\beta\theta}\cdot c(y_1,\ldots,y_n)\cdot\theta^{\sum_{i=1}^n y_i}e^{-n\theta} \\
&\propto \theta^{\alpha+\sum_{i=1}^n y_i-1}e^{-(\beta+n)\theta}
\end{aligned}
$$

$\therefore \{\theta|y_1,\ldots,y_n,\alpha,\beta\} \sim \text{Gamma}(\alpha+\sum_{i=1}^n y_i, \beta+n)$

- The mean of a Gamma distribution $f(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$ is $\frac{\alpha}{\beta}$
- The variance of a Gamma distribution $f(\theta)=\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$ is $\frac{\alpha}{\beta^2}$

Table 1: part (a) answer

	Mean: $\frac{\alpha+\sum_{i=1}^n y_i}{\beta+n}$	Variance: $\frac{\alpha+\sum_{i=1}^n y_i}{(\beta+n)^2}$	95% quantile based credible interval
A	11.85	0.5925	[10.38924, 13.40545]
B	8.928571	0.6377551	[7.4320, 10.560308]

Figure 2 plots the relationship between θ_B and n_0. The larger the n_0, the stronger the belief, the more closer the posterior expectation of θ_B is to that of θ_A. When $n_0 = 274$, the two values are the same (the cross point of the two dashed line in Figure 2).

Because type B mice are related to type A mice, so knowledge about ppopulation A should tell us something about population B. Having $\pi(\theta_A,\theta_B)=\pi(\theta_A)\times\pi(\theta_B)$ is assuming the prior distribution of θ_A annd θ_B to be independent, which prevent us from borrowing information from population A to population B. So it doesn't make sence.

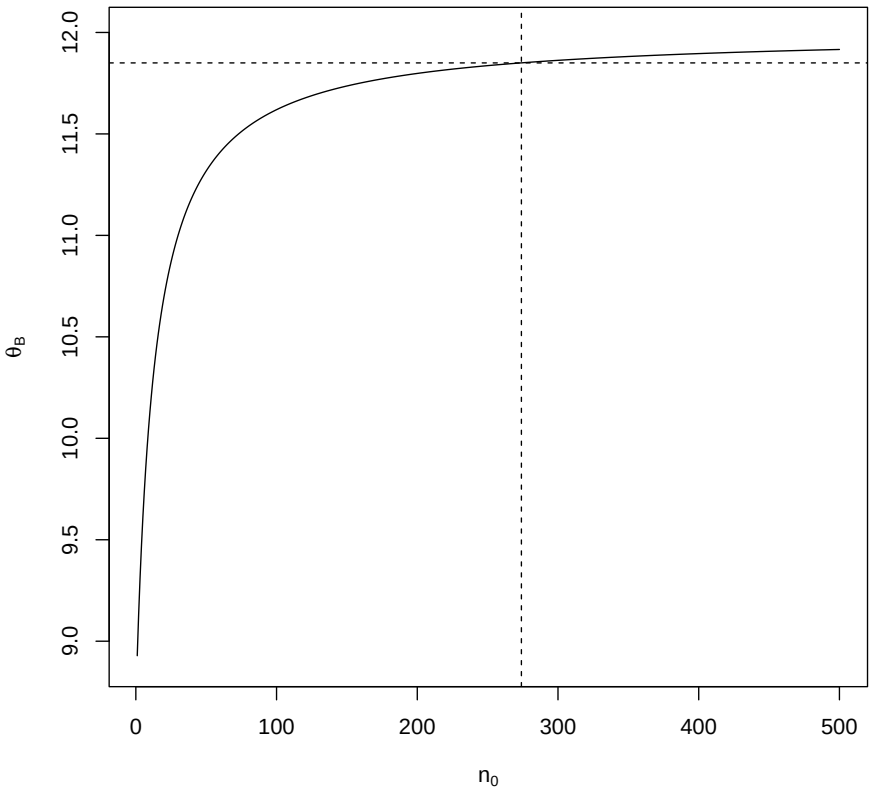

Figure 2: Mean of θ_B with respect to different n_0

5 Question 5

Let $f(x) = \exp(-x^4 - x^6 - x^8)$ for $x \in \mathcal{R}^1$ and let $K = \frac{1}{\int_{-\infty}^{\infty} f(x)dx}$. Describe in detail a Metrpolis-Hastings algorithm for the distribution having density $Kf(x)$, which uses proposal distributions (a)$N(\mu = x, \sigma^2 = 1)$ and (b)N $(\mu = x, \sigma^2 = 10)$.
(a) By producing traceplots, compare and contrast your experience in (a) and (b).
(b) Calculate $E(X)$ and $\mathrm{Var}(X)$ for the distribution of $X \sim Kf(x)$

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo (MCMC) method used for sampling from a probability distribution that is difficult to sample directly. It is particularly useful for sampling from high-dimensional distributions, such as posterior distributions in Bayesian inference.
Let $\pi(x)$ be the target distribution from which we want to sample, and let $q(x' \mid x)$ be a proposal distribution that generates a candidate sample x' given the current sample x. The algorithm proceeds as follows:

1. Initialize the Markov chain at some initial state x_0.
2. For each iteration t:
   (a) Generate a candidate sample x' from the proposal distribution $q(x' \mid x_t)$.
   (b) Calculate the acceptance probability:

$$\alpha(x_t, x') = \min\left\{1, \frac{\pi(x')\, q(x_t \mid x')}{\pi(x_t)\, q(x' \mid x_t)}\right\}$$

   (c) Generate a uniform random number u from the interval [0, 1].
   (d) If $u \leq \alpha(x_t, x')$, accept the candidate sample x' by setting $x_{t+1} = x'$; otherwise, set $x_{t+1} = x_t$.

The resulting sequence of samples $x_0, x_1, x_2, \ldots$ (often correlated) forms a Markov chain whose stationary distribution is $\pi(x)$. By discarding an appropriate "burn-in" period and thinning the samples, one can obtain samples approximately drawn from the target distribution.

Metropolis Algorithm

When q is symmetric proposal distribution, $q\left(x|y\right) = q\left(y|x\right)$, the expression for the acceptance ratio $\alpha\left(x_t, x'\right)$ simplifies to:

$$\alpha\left(x_t, x'\right) = \min\left\{1, \frac{\pi(x')}{\pi(x_t)}\right\}.$$

Then Metropolis-Hastings algorithm simplifies to Metropolis algorithm. A common choice for $q(y|x)$ is $N(x, \tau^2)$ for some $\tau > 0$. This means that the proposal is drawn from a Normal distribution centered at the current value.

The histograms of the M-H sampled data are shown in Figure 3; The traceplots are shown in Figure 4; I also plotted the Stationarity boxplots and Autocorrelation functions values in Figure 5 and Figure 6

From the program output, we know that the accept rate for proposal a is 0.517, and for proposal b it is 0.1897. We say that the accept rate for proposal a is nice because it's neither too high (which would result in slow mixing speed) or too low (wich might result in inefficient exploration). However, for proposal b, we can see from the traceplot it's more sparsely distributed compared to proposal a. This indicate that the exploration of probability space is inefficient.

We can also see from the stationarity plot (Figure 5 that chain a is more stable than chain b. And from the autocorrelation function plot (Figure 6 we can see that proposal b has a stronger autocorrelation between successive samples. And we also need to do a thinning of every 10^{th} sample for proposal a to remove autocorrelation between adjacent samplings.

For part (b), we redo the Metropolis Hastings Algorithm from proposal a, apply a thinning process for every 10^{th} to calculate $E(X)$ and $\text{Var}(X)$:

$E(X) : -0.009059417$

$\text{Var}(X) : 0.2174496$

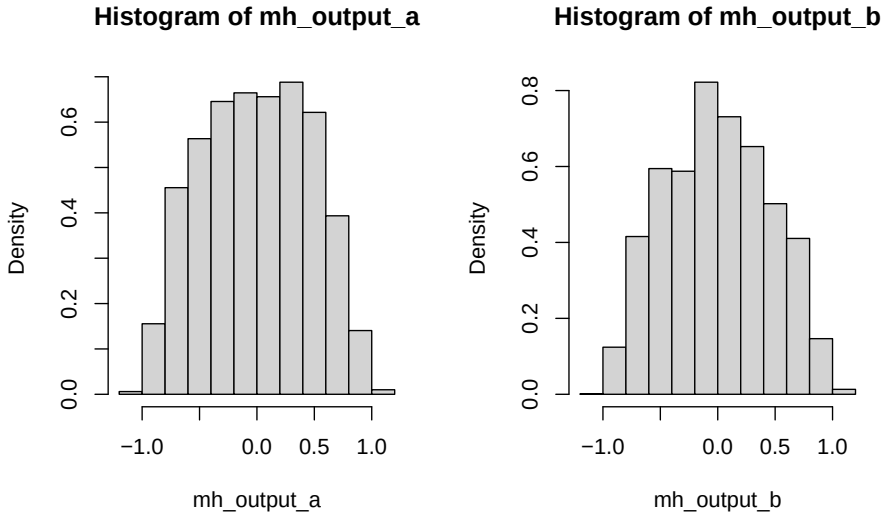

Figure 3: Histogram of the M-H sampled data

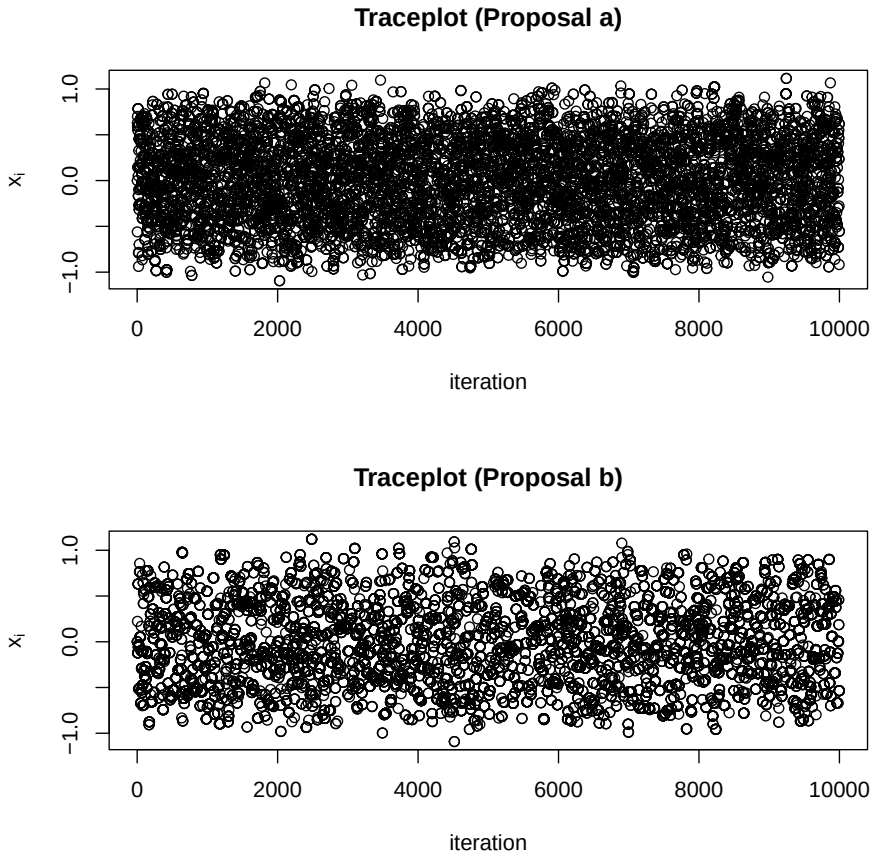

Figure 4: Traceplot of the M-H sampled data

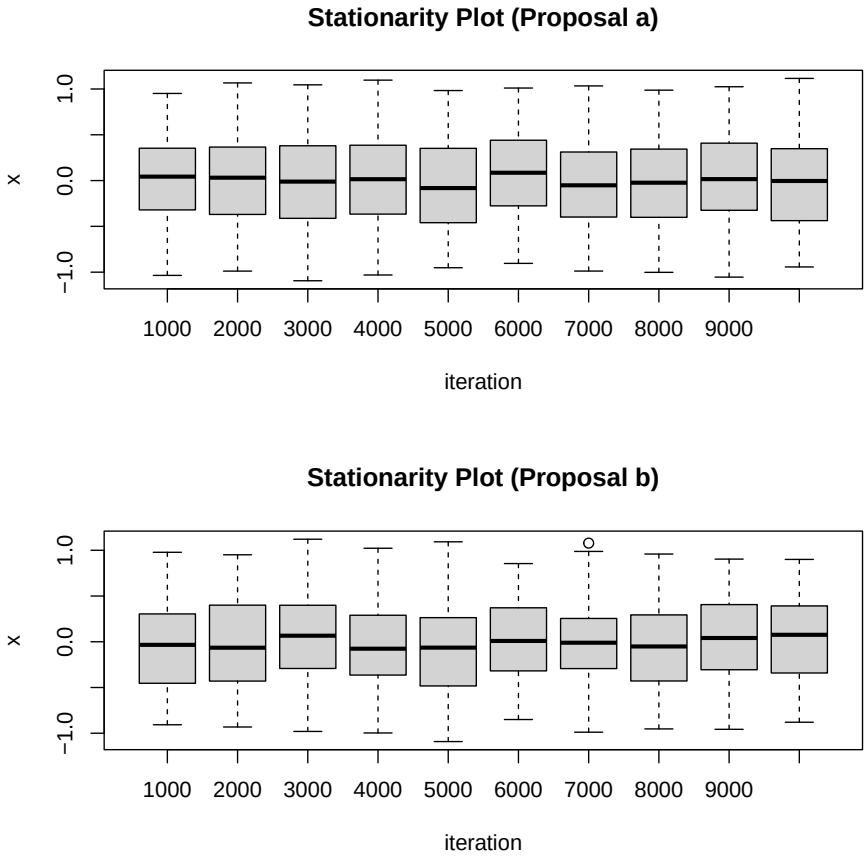

Figure 5: Stationarity plots of the M-H sampled data

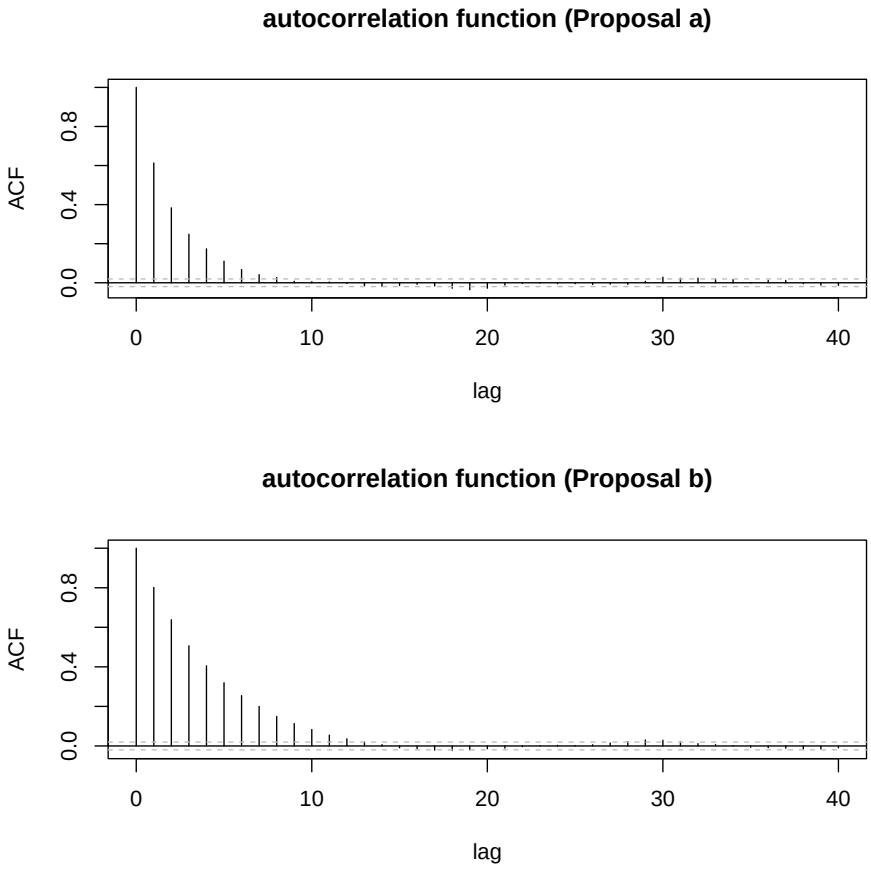

Figure 6: Autocorrelation plots of the M-H sampled data