Homework 1

1. (a) First, we use the identity $\int p(x)dx = 1$.

$$\begin{aligned} &\int p(x|\eta)dx = 1 \\ \implies &\int h(x)\exp(\eta T(x) - A(\eta))dx = 1 \end{aligned} \tag{1}$$

Next, we differentiate Eq. 1 w.r.t. η.

$$\begin{aligned} \frac{d}{d\eta}\int h(x)\exp(\eta T(x) - A(\eta))dx = \frac{d}{d\eta}(1) &= 0 \\ \implies \int h(x)\ \exp(\eta T(x) - A(\eta))\ (T(x) - \nabla A(\eta))dx &= 0 \\ \implies \int h(x)\exp(\eta T(x) - A(\eta))T(x)dx - \int h(x)\exp(\eta T(x) - A(\eta))\nabla A(\eta)dx &= 0 \\ \implies \int p(x|\eta)T(x)dx - \left[\int p(x|\eta)dx\right]\nabla A(\eta) &= 0 \\ \implies E[T(X)|\eta] &= \nabla A(\eta) \end{aligned}$$

Therefore, $E[T(X|\eta)] = \nabla A(\eta) = \left(\frac{\partial A}{\partial \eta_1}, \frac{\partial A}{\partial \eta_2} \cdots \frac{\partial A}{\partial \eta_d}\right)$

(b)

$$\begin{aligned} \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) &= \tfrac{1}{\sqrt{2\pi}\sigma}\exp(-\tfrac{x^2}{2\sigma^2}) \cdot \exp(\mu \cdot \tfrac{x}{\sigma^2} - \tfrac{\mu^2}{2\sigma^2}) \\ &= \tfrac{1}{\sqrt{2\pi}\sigma}\exp(-\tfrac{x^2}{2\sigma^2}) \cdot \exp(\eta T(x) - A(\eta)) \end{aligned}$$

where, $\eta = \frac{\mu}{\sigma^2}$, $A(\eta) = \frac{\mu^2}{2\sigma^2} = \frac{\sigma^2\eta^2}{2}$, $T(x) = x$ and $h(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{x^2}{2\sigma^2})$.

Thus, we get $E[T(X|\eta)] = E[X|\eta] = \nabla A(\eta) = \eta\sigma^2 = \mu$

2. (a) The pmf $p(X = k; \lambda)$ is

$$\begin{aligned} p(X = k; \lambda) &= \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \frac{1}{k!} \cdot \exp(\ln \lambda^k) \cdot e^{-\lambda} \\ &= \frac{1}{k!} \cdot \exp(-\lambda + \ln \lambda^k) \\ &= \frac{1}{k!} \cdot \exp(\eta T(x) - A(\eta)) \end{aligned}$$

where, natural parameter $\eta = \ln \lambda$, sufficient statistic $T(x = k) = k$ and normalizing constant $A(\eta) = \lambda$.

(b) Two options for non-informative priors(probability mass functions) for $p(\lambda)$ are:

- Uniform prior $\rightarrow \lambda \sim Unif(0,1), p(\lambda) = 1$
- Jeffrey's prior $\rightarrow p(\lambda) \propto \sqrt{\frac{1}{\lambda}}$

(c) Let, $X = (x_1, x_2, \cdots x_n)$ be n observed values for x. Therefore, the posterior distribution for λ can be written as:

$$p(\lambda|x_1, \cdots, x_n) \propto L(\lambda; x_1, x_2, \cdots x_n)\ p(\lambda)$$

where, $L(\lambda; x_1, x_2, \cdots x_n)$ is the joint likelihood of the observed data. Considering that the data is i.i.d. and the uniform prior $p(\lambda) \propto 1$, we can write the following:

$$\begin{aligned} p(\lambda|x_1, \cdots, x_n) &\propto \prod_{k=1}^{n} p(x_k|\lambda)\ p(\lambda) \\ &\propto \prod_{k=1}^{n} \frac{\lambda^{x_k} e^{-\lambda}}{x_k!} \cdot 1 \\ &\propto \prod_{k=1}^{n} \lambda^{x_k} e^{-\lambda} \\ &\propto e^{-n\lambda} \lambda^{\bar{X}+1-1} \end{aligned} \tag{2}$$

where, $\bar{X} = \sum_{k=1}^{n} x_k$. Similarly considering Jeffrey's prior, we get

$$p(\lambda|x_1, \cdots, x_n) \propto e^{-n\lambda} \lambda^{\bar{X}+1/2-1} \tag{3}$$

Eqs. 2,3 are the *Gamma* distributions. Therefore, the posterior distributions for λ are

$$p(\lambda|X) = \begin{cases} Gamma(\lambda; \bar{X}+1, n), & p(\lambda) \sim Unif \\ Gamma(\lambda; \bar{X}+1/2, n), & p(\lambda) \sim Jeff \end{cases}$$

where $\Gamma(n) = (n-1)!$. The *Gamma* distribution is defined as :

$$Gamma(x; a, b) = \frac{a^b}{\Gamma(a)} \cdot x^{a-1} \cdot \exp(-bx) \tag{4}$$

To compute whether $p(\lambda|X)$ is proper or improper, we have to show that $\sum_{\lambda} p(\lambda|X) = 1$. For the uniform prior,

$$\begin{aligned} \sum_{\lambda=0}^{\infty} p(\lambda|X) &= \sum_{\lambda=0}^{\infty} \frac{n^{\bar{X}+1}}{\Gamma(\bar{X}+1)} \cdot \left[\lambda^{\bar{X}} \cdot \exp(-n\lambda)\right] \\ &= \frac{n^{\bar{X}+1}}{\Gamma(\bar{X}+1)} \cdot \left[\sum_{\lambda=0}^{\infty} \lambda^{\bar{X}} \cdot \exp(-n\lambda)\right] \\ &= \frac{n^{\bar{X}+1}}{(\bar{X})!} \cdot \Gamma(\bar{X}+1) \\ &= \frac{n^{\bar{X}+1}}{(\bar{X})!} \cdot (\bar{X})! = n^{\bar{X}+1} \end{aligned} \tag{5}$$

Similarly, using Jeffrey's prior, we get $\sum_{\lambda=0}^{\infty} p(\lambda|X) = n^{\bar{X}+1/2}$. This and eq. 5 are both finite. Therefore, the posterior distribution for both Uniform and Jeffrey's prior are proper.

3. The pdf for μ is,

$$p(\mu) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq \mu \leq b, \\ 0, & \text{otherwise} \end{cases}$$

According to Bayes' rule,

$$p(\mu|x_1, \cdots, x_n) = \frac{p(x_1, \cdots, x_n|\mu)p(\mu)}{\int_\mu p(x_1, \cdots, x_n, \mu)d\mu}$$

Since $p(\mu) = 0$, when $\mu \notin [a, b]$, we can modify the bounds of the integral. Thus, posterior pdf

$$\begin{aligned}
p(\mu|x_1, \cdots, x_n; \sigma^2, a, b) &= \frac{\left[\prod_{i=1}^{n} p(x_i|\mu, \sigma^2)\right] \cdot p(\mu|a, b)}{\int_{\mu=a}^{\mu=b} \left[\prod_{i=1}^{n} p(x_i|\mu, \sigma^2)\right] \cdot p(\mu|a, b) d\mu} \\
&= \frac{\left[\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})\right] \cdot \frac{1}{b-a}}{\int_{\mu=a}^{\mu=b} \left[\frac{1}{(\sqrt{2\pi}\sigma)^n} \exp(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})\right] \cdot \frac{1}{b-a} d\mu} \\
&= \frac{\exp(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2})}{\int_{\mu=a}^{\mu=b} \exp(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}) d\mu} \\
&= \frac{\exp(-\frac{n\mu^2 - 2n\bar{x}\mu}{2\sigma^2})}{\int_{\mu=a}^{\mu=b} \exp(-\frac{n\mu^2 - 2n\bar{x}\mu}{2\sigma^2}) d\mu}, \qquad (\text{where } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i)
\end{aligned}$$

4. (a) The joint posterior $p(\mu_j, \sigma_j^2|y_{ij})$ is a normal inverse gamma distribution.

$$p(\mu_j, \sigma_j^2|y_{ij}) = NIG(\mu_n, n_n, \alpha_n, \beta_n),$$

where,

$$\begin{aligned}
\mu_{nj} &= \frac{n_0\mu_0 + n_j\bar{y}_j}{n_0 + n_j} \\
n_{nj} &= n_0 + n_j \\
\alpha_{nj} &= \alpha_0 + \frac{n_j}{2} \\
\beta_{nj} &= \beta_0 + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2 + \frac{n_j n_0}{n_j + n_0}\frac{(\bar{y}_j - \mu_0)^2}{2}
\end{aligned} \tag{6}$$

The marginal posterior $p(\sigma_j^2|y_{ij})$ is an Inverse-Gamma(IG) distribution with shape parameter as α_n and scale parameter as β_n.

I used the **"LaplacesDemon"** (`http://www.bayesian-inference.com/softwaredownload`) R package for the inverse-gamma and students-t distribtutions.

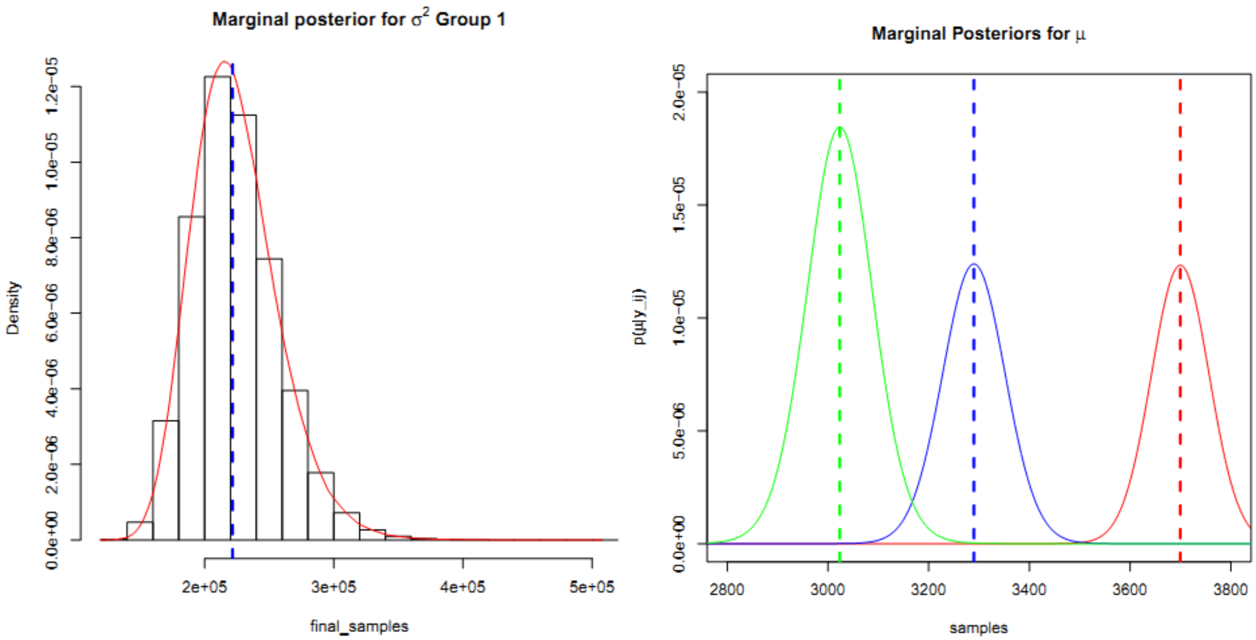

Figure 1: **(a):** Marginal posterior $p(\sigma_1^2|y_{ij})$. Dotted line shows the sample variance of the control group. **(b)** Marginal posterior $p(\mu_j|y_{ij})$. Control group is denoted by red, mild is denoted by blue and dementia group is denoted by green. Dotted lines denote the sample mean respectively.

(b) The marginal posterior $p(\mu_j|y_{ij})$ is a students-t distribution.

$$p(\mu_j|y_{ij}) = t(\mu_j|\nu = 2\alpha_{nj}, \mu = \sqrt{\frac{2\beta_{nj}}{n_{nj}}}) \tag{7}$$

where, β_{nj} and n_{nj} are the posterior parameters from eq.6. In eq. 7, ν denotes the degrees of freedom and μ denotes the *non-centrality parameter*. Fig. 1(b) shows the marginal posterior for the three groups.

(c) The conditional distribution $p(d_{12}|\sigma_1^2, \sigma_2^2, y_{ij})$ is a normal distribution, with mean as $\bar{y}_1 - \bar{y}_2$ and variance as $(\frac{\sigma_{n_1}^2}{n_1} + \frac{\sigma_{n_2}^2}{n_2})$, where $\sigma_{n_1}^2$ and $\sigma_{n_2}^2$ are sampled from the posterior distributions $p(\sigma_{n_1}^2|y_{i1})$ and $p(\sigma_{n_2}^2|y_{i2})$ obtained in (a). The other distributions $p(d_{23}|\sigma_2^2, \sigma_3^2, y_{ij})$ and $p(d_{13}|\sigma_1^1, \sigma_3^2, y_{ij})$ are obtained similarly.

Probabilities $P(d_{12} < 0|y_{ij})$, $P(d_{13} < 0|y_{ij})$ and $P(d_{23} < 0|y_{ij})$ are $2 * 10^{-6}$, $5 * 10^{-6}$ and 0.03. These probabilities tell us that the mild and dementia groups overlap, as shown in Fig. 1, by 3%. On the other hand, the control group is quite separated from the other groups, using the NIG prior.

(d) Using the one-sided t-test with "less" as the null hypothesis, we get the highest p-value for d_{23} followed by d_{13} and d_{12}. This complies with our analysis that groups mild and dementia actually overlap. The p-value obtained from the t-test denotes the probability that the *alternative* hypothesis holds true. Thus, higher the p-value the event described in the alternative hypothesis is more likely.

In this case, "less" denotes the random event $\mu_1 - \mu_2 < 0$. Therefore, highest p-value of d_{23} denotes that this event is most likely.

5. (a) Using Jeffrey's non-informative prior $p(\mu, \sigma^2|y_{ij}) \propto \frac{1}{\sigma^2}$, the joint posterior $p(\mu_j, \sigma_j^2|y_{ij})$ is a normal distribution.

The marginal posterior $p(\sigma_j^2|y_{ij})$ is a scaled Inv-χ^2 distribution, with parameters ν(degrees of freedom) as $2\alpha_{nj} - 1$ and scale parameter is $\bar{\sigma_j^2}$, the sample variance for group j.

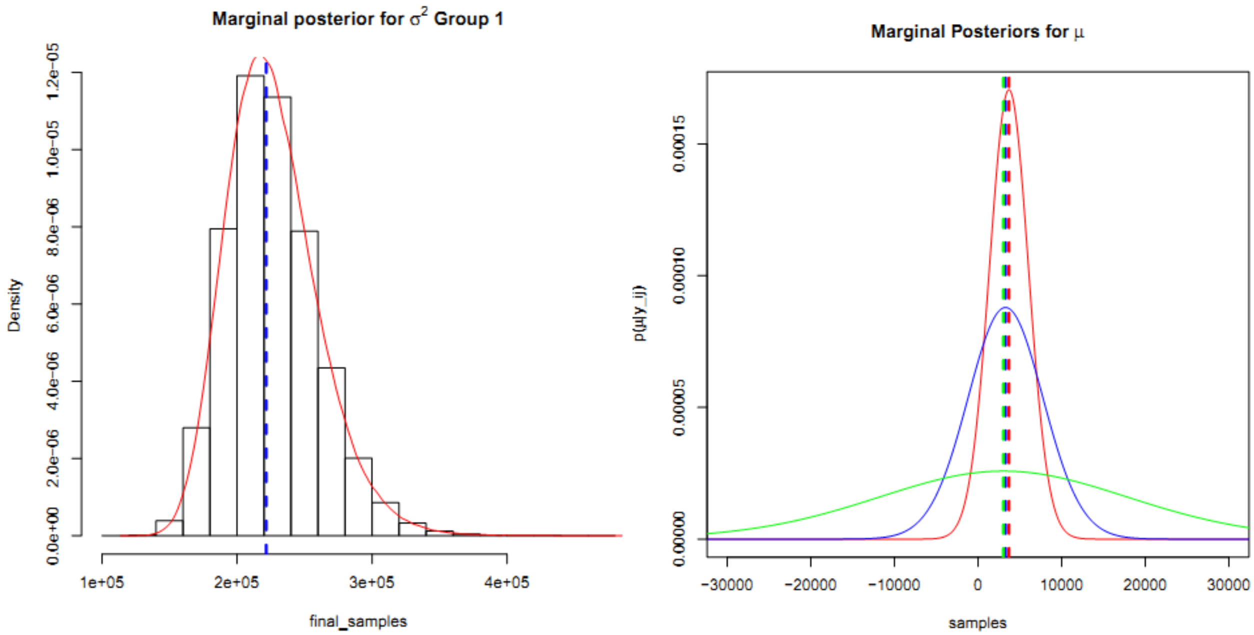

Figure 2: **(a):** Marginal posterior $p(\sigma_1^2|y_{ij})$. Dotted line shows the sample variance of the control group. **(b)** Marginal posterior $p(\mu_j|y_{ij})$. Control group is denoted by red, mild is denoted by blue and dementia group is denoted by green. Dotted lines denote the sample mean respectively.

(b) The marginal posterior $p(\mu_j|y_{ij})$ is a non-central students' t-distribution with parameters - degrees of freedom($\nu = n_j$) and *non-centrality parameter*(ncp $= \frac{\bar{\sigma_j^2}}{n_j}$). The plot of the marginal distributions for the three groups are shown in Fig. 2(b).

(c) The conditional distribution $p(d_{12}|\sigma_1^2, \sigma_2^2, y_{ij})$ is a normal distribution, with mean as $\bar{y_1} - \bar{y_2}$ and variance as $(\frac{\bar{\sigma_{n_1}^2}}{n_1} + \frac{\bar{\sigma_{n_2}^2}}{n_2})$, where $\bar{\sigma_{n_1}^2}$ and $\bar{\sigma_{n_2}^2}$ are the sample variances of groups 1 and 2 respectively. The other distributions $p(d_{23}|\sigma_2^2, \sigma_3^2, y_{ij})$ and $p(d_{13}|\sigma_1^1, \sigma_3^2, y_{ij})$ are obtained similarly.

Probabilities $P(d_{12} < 0|y_{ij})$, $P(d_{13} < 0|y_{ij})$ and $P(d_{23} < 0|y_{ij})$ are 10^{-6}, $1.5 * 10^{-5}$ and 0.034.

6. (a,b) Using the given "pseudo-observations", we get the following values of the initial parameters: $\mu_0 = 2133, \nu_0 = 127, \sigma_0 = 279, \alpha_0 = n_0/2$ and $\beta_0 = \frac{n_0-1}{2}\sigma_0^2$. We use these values for all three groups. Using these values, the joint posterior $p(\mu_j, \sigma_j^2|y_{ij})$ for each group j is a NIG distribution, whose parameters are same as those in eq. 6.

The marginal posterior $p(\sigma_j^2|y_{ij})$ is an *Inverse-Gamma* distribution, while $p(\mu_j|y_{ij})$ is a *students-t distribution*, with similar posterior hyperparameters as in 4.

Fig 3 shows the marginal posterior distributions on σ^2 and μ.

(c) $P(d_{12} < 0|y_{ij}), P(d_{13} < 0|y_{ij})$ and $P(d_{23} < 0|y_{ij})$ are now 0.0003, 0 and 0.0183 respectively.

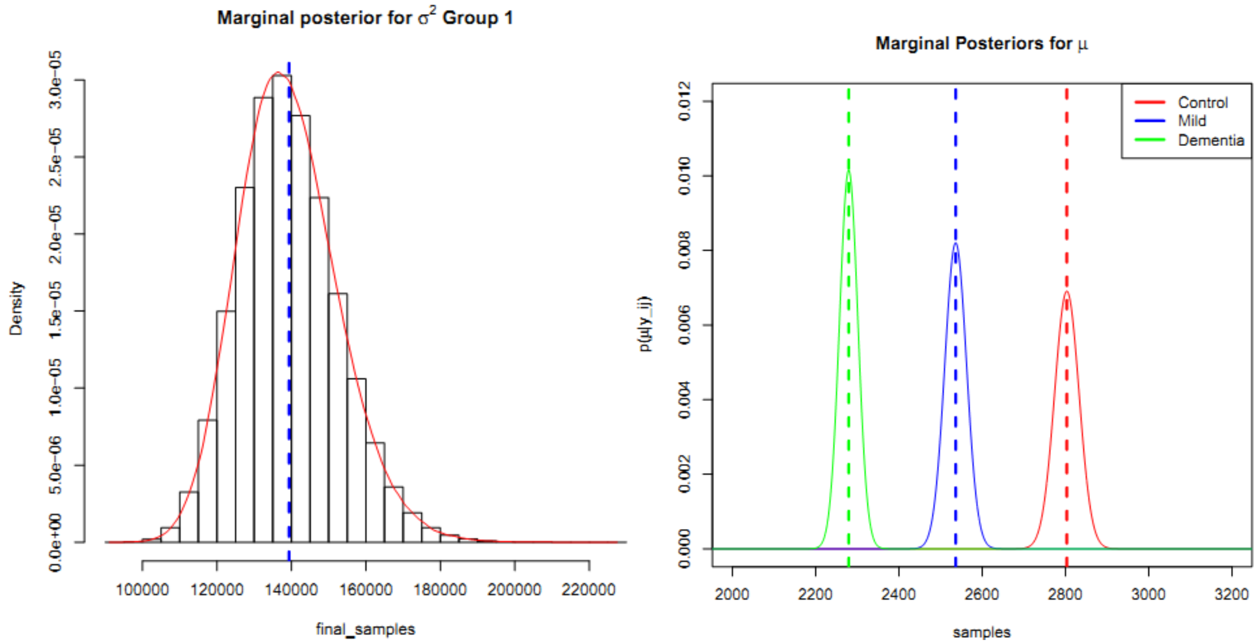

Figure 3: **(a):** Marginal posterior $p(\sigma_1^2|y_{ij})$. Dotted line shows the sample variance of the control group. **(b)** Marginal posterior $p(\mu_j|y_{ij})$. Control group is denoted by red, mild is denoted by blue and dementia group is denoted by green. Dotted lines denote the sample mean respectively.

From Fig. 3(b), it can be seen that the peaks are well-separated. Thus, the NIG conjugate prior based on the "pseudo-observations" actually help to separate the resulting posteriors of the three groups.

Thus, in this case, prior domain knowledge of the hippocampus data will result in better classification.