Project B

Exercise 1.1. Under the same setting as the SAT lecture slides, consider the SAT dataset :

```
## SAT Data
SATy <- c(28,8,-3,7,-1,1,18,12) # this is \bar y_j for j = 1,..., 8
SATsigma <- c(15,10,16,11,9,11,10,18) # this is \sigma_j for j = 1,..., 8
```

Remember that :

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$
$$\sigma_j^2 = \sigma^2/n_j$$
$$\bar{y}_j|\theta_j \sim N(\theta_j, \sigma_j^2)$$

and since $\theta_1, \dots, \theta_J$ are exchangeable we have :

$$p(\theta_1, \dots, \theta_J|\mu, \tau) = \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2)$$
$$p(\theta_1, \dots, \theta_J) = \int \prod_{j=1}^{J} N(\theta_j|\mu, \tau^2) p(\mu, \tau) d(\mu, \tau)$$
$$p(\mu, \tau) = p(\mu|\tau)p(\tau) \propto p(\tau).$$

The goal of this project is to implement in `R` the normal hierarchical model, as described in the slides, to sample from $\theta|y$ and $\theta|(\tau, y)$. In the following always consider samples of size $n = 10^4$.

1. Assuming a uniform prior over τ, $p(\tau) \propto 1$, write code to sample from $\tau|y$ knowing that :

$$p(\tau|y) = \frac{p(\mu, \tau|y)}{p(\mu|\tau, y)} \propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} (\sigma_j^2 + \tau^2)^{-1/2} \exp\left(\frac{-(\bar{y}_j - \hat{\mu})^2}{2(\sigma_j^2 + \tau^2)}\right)$$
$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{\bar{y}_j}{\sigma_j^2 + \tau^2}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}}, \qquad V_\mu = \left(\sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}\right)^{-1}$$

   you should use some techniques that you have already seen in practical B. Plot the histogram of the obtained sample superimposing the density $p(\tau|y)$ for $\tau \in (0, 30]$. If you notice something unexpected in the histogram, discuss it.

2. Write code to sample from $\mu|\tau, y$ and $\theta|\mu, \tau, y$:

$$\mu|\tau, y \sim N(\hat{\mu}, V_\mu)$$
$$\theta_j|\mu, \tau, y \sim N(\hat{\theta}_j, V_j)$$
$$\hat{\theta}_j = \frac{\bar{y}_j/\sigma_j^2 + \mu/\tau^2}{1/\sigma_j^2 + 1/\tau^2}, \qquad V_j = 1/(1/\sigma_j^2 + 1/\tau^2)$$

3. With the functions you have, sample from $\theta|y$ and plot the histogram of the sample.
4. With the functions you have, sample from $\theta|\tau, y$ and plot the histogram of the sample for $\tau = 5$.
5. Produce a plot of the conditional expectation of $\theta|(\tau, y)$ and a plot of the conditional standard deviation of $\theta|(\tau, y)$ as a function of τ, for $\tau \in (0, 30]$.
6. Approximate the distribution of $\max_j\{\theta_j\}|y$ using a histogram and approximate $\mathbb{P}(\min_j\{\theta_j\} < 0|y)$ and $\mathbb{P}(\theta_C < \theta_E|y)$ using your simulations.

Instructions

— This project is worth 20% of your final mark.
— The deadline for this project is Sunday 16th of April 2023 at midnight.
— You can work on this in groups of 3 students. You can decide the groups. Make sure that the names of each person in the group appears on the report.
— You should submit both the `.rmd` (with the material needed to knit it) and the `.html` files on Moodle. Make sure your results are reproducible.
— Late submissions are allowed but you will lose 0.5 pt for each 12 hours period of delay.
— No external package is allowed except for plotting, or unless specifically announced.
— The length of the report should be reasonable. Reports that are too long will be penalized.
— You will receive group feedbacks for your work but no general solution will be released.
— The evaluation will be based on the marking grid, see Table **??**.

Marking scheme

E/R	**2** pts
CA	**2** pts
R/C	**1** pt
P	**1** pt

Table 1 – Marking grid.

— **E/R** = Clearness of **E**xplanations and **R**easoning. You should provide a neat description of the thought process you followed to arrive to the solution you present.
— **CA** = **C**orrect **A**nswer. Arriving to a correct solution is important, however even if the solution is incorrect, but the reasoning you followed is sound, you would still get points from **E/R**.
— **R/C** = **R**eproducibility and clearness of **C**ode. You should make sure that your code is understandable and tidy. Make comments and explain what you are doing. If you do simulations, set the seed for reproducibility of your results.
— **P** = **P**resentation. The report should be well structured and it should have a logic development and present all your results. The quality of the graphs, tables and figures you present will be evaluated. They should help to understand your results in a compact and smart way. Choose carefully what to include and not include in your reports.