Applied Bayesian modeling - HW1

Score: The maximum number of points in this HW is 15 points, with 3 points extra credit. For calculating a final HW grade, the points will be rescaled to a maximum score of (15+3)/15*100% = 120%.

What to hand in: For exercise 2, we need an Rmd and a knitted pdf. You may hand in the answer to exercise 1 in a different output form as long as it's legible (i.e., no difficult-to-read picture of handwritten notes).

Exercise 1: Breast cancer and mammogram screening [5 pts]

This exercise is about the material in module 2.

Background information: Gerd Gegerenzer explained to 24 physicians:

- For early detection of breast cancer, women are encouraged to have routine screening, even if they have no symptoms.
- Imagine you conduct such screening using mammography
- The following information is available about asymptomatic women aged 40 to 50 in your region who have mammography screening:
  - The probability an asymptomatic woman has breast cancer is 0.8%.
  - If she has breast cancer, the probability is 90% that she has a positive mammogram.
  - If she does not have breast cancer, the probability is 7% that she still has a positive mammogram.
- Suppose a woman has a positive mammogram: What is the probability she actually has breast cancer?
- Physicians' answers ranged from about 1% to about 90%.

Use Bayes' rule to obtain the probability that a woman with a postive mammogram has breast cancer, using the information provided above. Show working, meaning to write out how you obtained probabilities that are not given in the information.

Exercise 2: posteriors when everything's normal

This exercise is about module 3. You may find the R notebook with module 3 helpful.

We continue with the radon data set. (Note that I read in and process the data in the HW Rmd but don't print out the code).

We will carry out Bayesian inference assuming a normal likelihood and prior:

$y_i|\mu, \sigma^2 \sim N(\mu, \sigma^2)$, independent;
$\mu \sim N(m_0, s_{\mu 0}^2)$

with prior mean parameters $\mu_0 = 0$ and $s_{\mu 0} = 0.1$ and $\sigma = s\{y\}$.
$\bar{y}$ is given by the log-radon data.

Exercise 2a [5 pts, with additional 1 pt extra credit]

Use the radon data and obtain the posterior distribution $p(\mu|\mathbf{y})$, using the prior and likelihood as specified above. Use the posterior to produce the following outputs:
(i) one plot with the prior, posterior, and likelihood function;
(ii) a point estimate, 95% credible interval, and 80% credible interval.
(iii) Interpretation of the 80% credible interval.

Extra credit question (1 pt): Can you calculate the probability that μ is greater than $\bar{y}$? If yes, report it. If not, why not?

Exercise 2b [5pts]

Let's call the data set used so far data set 1. Suppose there is a second radon data set, referred to as data set 2, that has the same $\bar{y}$ and $s\{y\}$ as data set 1. What differs between the two data sets is that data set 2 has sample size $n = 4635$, which is 5 times the sample size of data set set 1 (with $n = 927$).

Obtain the posterior using data set 2, and produce the same outputs (i) and (ii) from exercise a. (No need to interpret the CI).

Exercise 2c [extra credit 2pts]

Briefly comment on the differences in posteriors between exercises a and b: in which setting is the posterior more data-driven, closer to $\bar{y}$? Is that what you expected?