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.

Answer:

When given a question in words, consider first introducing some notation, to then define the outcome of interest and consider what information is given.

- Let
  - C be the breast cancer outcome ($C = 1$ if the woman has breast cancer, 0 otherwise),
  - M the outcome of the mammogram ($M = 1$ if the mammogram is positive, 0 otherwise).
- What do we want to know?
  $P(C = 1|M = 1)$.
- What information do we have?
  - The probability a woman has breast cancer is 0.8%.
    $\Rightarrow P(C = 1) = 0.008$.

  - If she has breast cancer, the probability is 90% that she has a positive mammogram $\Rightarrow P(M = 1|C = 1) = 0.9$.
  - If she does not have breast cancer, the probability is 7% that she still has a positive mammogram $\Rightarrow P(M = 1|C = 0) = 0.07$.

- Putting it all together:

$$\begin{aligned} Prob(C = 1) &= 0.008, \\ Prob(M = 1|C = 1) &= 0.9, \\ Prob(M = 1|C = 0) &= 0.07. \end{aligned}$$

- Bayes' rule says that the prob. that the woman with a positive mammogram has breast cancer is

$$Prob(C = 1|M = 1) = \frac{Prob(M=1|C=1)Prob(C=1)}{Prob(M=1)},$$
$$= \frac{Prob(M=1|C=1)Prob(C=1)}{\sum_{c=0}^{1} Prob(M=1|C=c)Prob(C=c)} = \frac{0.9\times 0.008}{0.9\times 0.008+0.07\times 0.992} = 9.4\%.$$

  Note that here we carry out Bayesian updating of the prior probability that a woman has cancer based on evidence (a + mammogram).

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 posterior probability that μ is greater than $\bar{y}$? If yes, report it. If not, why not?

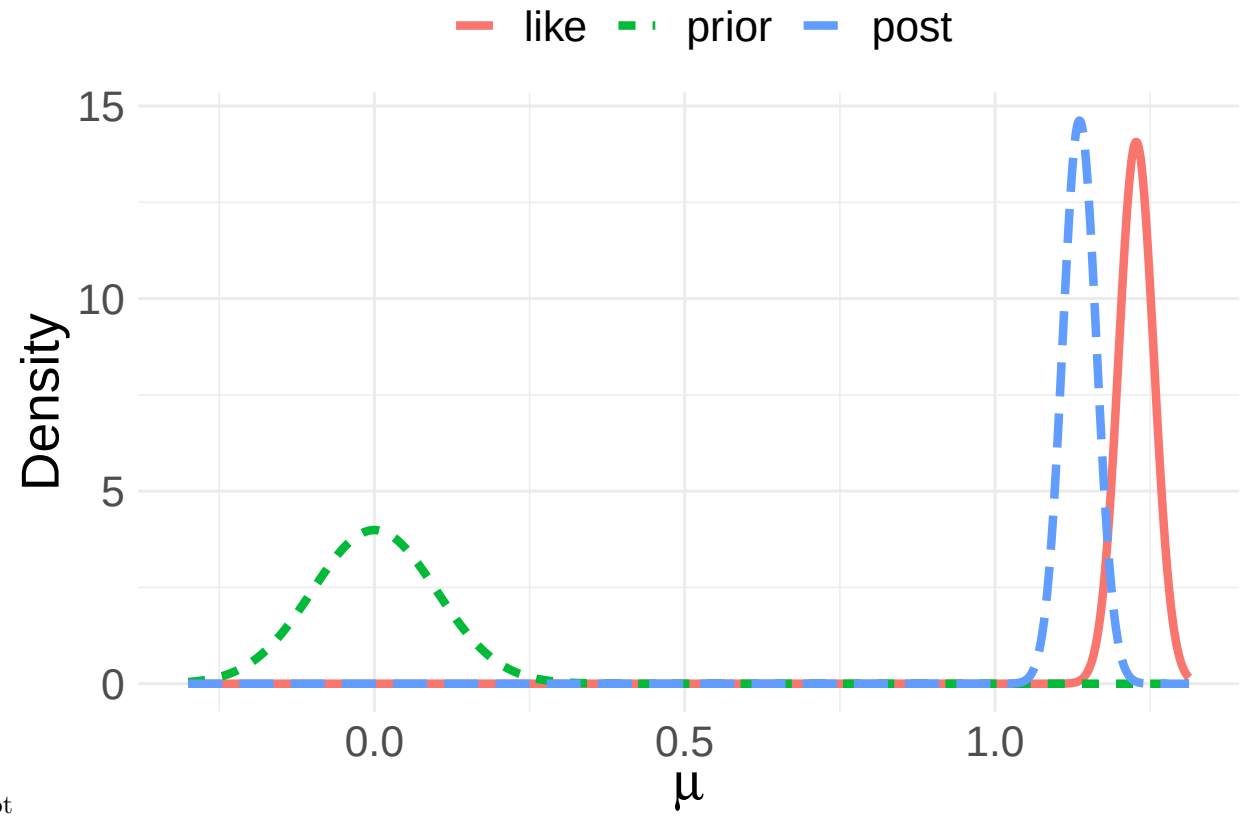
Answer: Fix prior mean and prior sd

```
mu0 <- 0 # prior mean
sigma.mu0 <- 0.1  # prior sd
```

Information that depends on the data

```r
ybar <- mean(y)
sd.y <- sd(y)
n <- length(y)
sigma <- sd.y
# sd for ybar follows from sigma
sd.ybar <- sigma/sqrt(n) # needed if you want to use the automated grid option
```

Posterior mean and variance:



Plot

Point estimate:

```r
mupost.mean # posterior mean
```

```
## [1] 1.136079
```

95% CI:

```r
qnorm(c(0.025, 0.975), mean = mupost.mean, sd = mupost.sd) # 95% quantile-based CI
```

```
## [1] 1.082604 1.189554
```

80% CI:

```r
qnorm(c(0.1, 0.9), mean = mupost.mean, sd = mupost.sd) # 80% quantile-based CI
```

```
## [1] 1.101113 1.171045
```

Interpretation of the 80%CI: there is an 80% probability that μ is between 1.10 and 1.17, where μ here refers to mean log-radon.

Probability that μ is greater than $\bar{y}$ is 0.04%.

```
(1 - pnorm(ybar, mean = mupost.mean, sd = mupost.sd) )
```

```
## [1] 0.0004055814
```

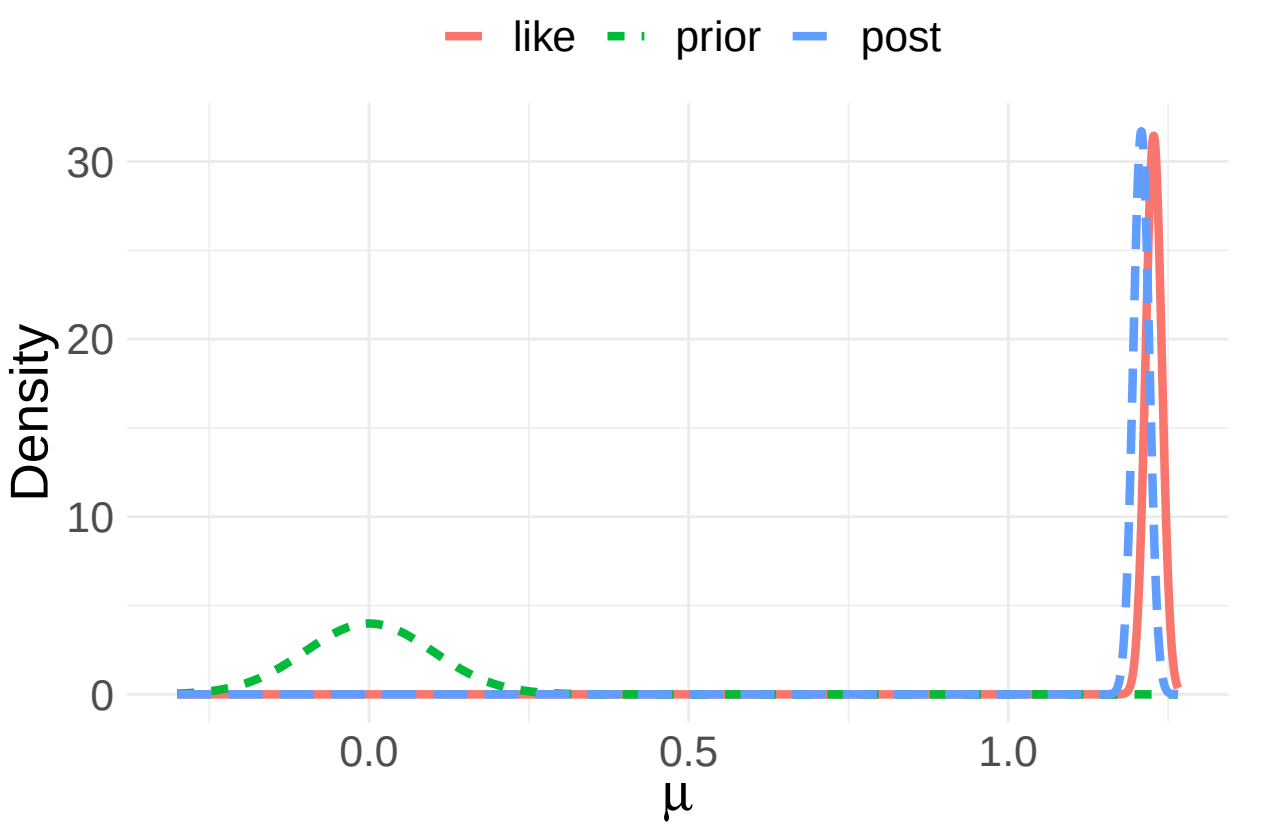
Exercise 2b [5pts]

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).

Answer Updating what changes:

```
n <- 5*length(y)
sd.ybar <- sigma/sqrt(n)
mupost.mean <- (mu0/(sigma.mu0^2) + n*ybar/(sigma^2))/(1/(sigma.mu0^2) + n/(sigma^2))
mupost.sd <- sqrt(1/(1/(sigma.mu0^2)+n/(sigma^2)))
```



Point estimate:

```r
mupost.mean # posterior mean
```

```
## [1] 1.20802
```

```r
qnorm(c(0.025, 0.975), mean = mupost.mean, sd = mupost.sd) # 95% quantile-based CI
```

```
## [1] 1.183359 1.232680
```

```r
qnorm(c(0.1, 0.9), mean = mupost.mean, sd = mupost.sd) # 80% quantile-based CI
```

```
## [1] 1.191895 1.224144
```

```r
(1 - pnorm(ybar, mean = mupost.mean, sd = mupost.sd)) # prob that mu is greater than ybar
```

```
## [1] 0.06124678
```

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?

Answer The posterior is closer to $\bar{y}$, more data driven, in (b). That's what I expected because there is more data to inform the posterior in (b).