Bayesian Learning | **Instructor: Hedibert F. Lopes**
Professional Master in Economics | **Insper Institute of Education and Research**
Homework 1 | **Due date: May 12th 2025**

Question 1

A couple of yeras ago, researchers from the United Kingdom analyzed a series of published studies on commercial antigen tests. The study included 24,087 samples, of which 7,415 had tested positive for Sars-CoV-2 (`https://noticias.r7.com/saude/testes-rapidos-de-covid-tem-maior-risco-de-falso-negativo-08012022`). According to the study, in people with symptoms, the tests had an average of 72% accuracy in diagnosis. However, there was still a possibility, in this case, of 28% of patients infected with the coronavirus having a false negative. False positive cases are rare. The British researchers concluded that when the test confirms the infection, the result is correct 99.6% of the time. According to the research, consider the following probabilities:

$$\begin{aligned} P(X=1|\theta=0) &= 1\% \\ P(X=0|\theta=1) &= 28\% \\ P(X=1,Y=1|\theta=0) &= 0,01\% \\ P(X=1,Y=1|\theta=1) &= 51,84\% \end{aligned}$$

1.a

The prevalence of Covid in my mother's age group is 0.1571%, i.e., the accumulated scientific knowledge is that $P(\theta=1)$. What is the posterior probability of her having Covid after two independent tests, X and Y, come back positive, $P(\theta=1|X=1,Y=1)$?

1.b

Even after two independent tests came back positive, my mother is 90% sure she doesn't have Covid. With this level of posterior certainty, what implicit prior would she have to have for such a statement?

Question 2

Our word choices can reflect where we live. For example, suppose you're watching an interview of somebody that lives in the United States. Without knowing anything about this person, U.S. Census figures provide prior information about the region in which they might live: the Midwest (M), Northeast (N), South (S), or West (W):

$$\begin{aligned} Pr(region=M) &= 0.21 \\ Pr(region=N) &= 0.17 \\ Pr(region=S) &= 0.38 \\ Pr(region=W) &= 0.24 \end{aligned}$$

Notice that the South is the most populous region and the Northeast the least. Thus, based on population statistics alone, there's a 38% prior probability that the interviewee lives in the South. But then, you see the person point to a fizzy cola drink and say "please pass my pop." Though the country is united in its love of fizzy drinks, it's divided

in what they're called, with common regional terms including "pop," "soda," and "coke." This data, i.e., the person's use of "pop," provides further information about where they might live. To evaluate this data, we can examine the pop vs soda dataset which includes 374250 responses to a volunteer survey conducted at popvssoda.com.

$$\begin{aligned} Pr(say = pop|region = M) &= 0.6447 \\ Pr(say = pop|region = N) &= 0.2734 \\ Pr(say = pop|region = S) &= 0.0792 \\ Pr(say = pop|region = W) &= 0.2943 \end{aligned}$$

2.a

Show that there is a 28.26% chance that a person in the U.S. uses the word "pop".

2.b

Considering the fact that 38% of people live in the South but that "pop" is relatively rare to that region, what's the posterior probability that the interviewee lives in the South?

2.c

Update our understanding of the interviewee living in the Midwest, Northeast, or West.

Question 3

A 2017 Pew Research survey found that 10.2% of LGBT adults in the U.S. were married to a same-sex spouse. Now it's the 2024s, and Pamela guesses that π, the percent of LGBT adults in the U.S. who are married to a same-sex spouse, has most likely increased to about 15% but could reasonably range from 10% to 25%.

4.a Identify and plot a Beta model that reflects Palema's prior ideas about π.

4.b Pamela wants to update her prior, so she randomly selects 90 US LGBT adults and 30 of them are married to a same-sex partner. What is the posterior model for π?

4.c Calculate the posterior mean, mode, and standard deviation of π.

4.d Does the posterior model more closely reflect the prior information or the data? Explain your reasoning.

Question 4

Suppose you have the data $x_1, \ldots, x_n$ and that you want to entertain the adherence of the data to the following configurations of model and prior:

1. The data is iid from $N(\mu, \sigma^2)$, with the prior for μ being $N(0, 1)$ and $\sigma = 0.25$.
2. The data is iid from $N(\mu, \sigma^2)$, with the prior for μ being $t_3(0, 1)$ and $\sigma = 0.25$.
3. The data is iid from $N(\mu, \sigma^2)$, with the prior for σ^2 being $IG(0.1, 0.1)$ and $\mu = 0$.
4. The data is iid from $N(\mu, \sigma^2)$, with the prior for σ^2 being $U(0, 5)$ and $\sigma = 0.25$.

Here $z \sim IG(a, b)$ is the inverse-gamma distribution, such that $p(z) = \frac{b^a}{\Gamma(a)} z^{-(a+1)} e^{-b/z}$.

For 1) to 4), and the following $n = 15$ observations:

```
x = c(-0.13,1.46 ,0.37,0.26,0.86,0.68,0.29,1.51,-0.10,1.25,0.27,-0.26,0.64,0.73,0.48)
```

find the posterior mean for the unknown parameter analytically whenever possible, i.e.

1. $E(\mu|x_1, \ldots, x_n)$
2. $E(\mu|x_1, \ldots, x_n)$
3. $E(\sigma^2|x_1, \ldots, x_n)$
4. $E(\sigma^2|x_1, \ldots, x_n)$

If not possible, use simple deterministic grids. For example, a grid for μ could be between -5 and 5 of length $h = 0.01$, while for σ in 4) the grid could be between the range 0 to 1 of length $h = 0.01$.

Note: Later we will revisit this problem, but solving it via Monte Carlo integration and Monte Carlo simulations.