# STAT 447 Assignment 6

Caden Hewlett

2024-02-17

## Question 1: Efficacy of Vaccines

In this exercise we will model the effectiveness of COVID vaccines using clinical trials data. Each trial consists of two arms: vaccinated (i.e., treated) and control (i.e., not treated). For a typical trial we know

- The total number of patients in each arm: $t_v$ (vaccinated), $t_c$ (control).
- The number of patients that got infected with the SARS-CoV-2 virus in each arm: $n_v$ and $n_c$.

We model $n_v$ and $n_c$ as Binomial random variables. The unknown parameter for these distributions will depend on two numbers in $[0, 1]$

The first is Incidence (denoted $p$): the probability that a patient in the trial will become infected without being treated with the vaccine.

The second is Effectiveness (denoted $e$) the decrease in incidence that the vaccine provides.

We will take betaMP$(\mu, \lambda)$ distributions rather than beta$(\alpha, \beta)$. In this instance $\mu \in [0, 1]$ is the *mean* and $\lambda > 0$ is a *precision* parameter.

The following bijection holds regarding these parameters:

$$\mu = \frac{\alpha}{\alpha + \beta}, \text{ and } \lambda = \alpha + \beta \iff \alpha = \mu\lambda, \text{ and } \beta = (1 - \mu)\lambda$$

There are a few priors in this experiment.

Firstly, we have the mean and precision priors of Effectiveness, given by a Uniform and Exponential distribution, respectively:

$$\mu_e \sim \text{unif}(0, 1), \text{ and } \lambda_e \sim \exp(0.001)$$

This gives us the Effectiveness, as a Likelihood given these Beta Parameters.

$$\big(e \mid \{\mu_e, \lambda_e\}\big) \sim \text{betaMP}(\mu_e, \lambda_e)$$

We then arrive at the number of infected inoculated individuals $n_v$, which arises as a function of the total number of individuals (the known value $t_\text{v}$), the incidence $p$ (whose formulation will be discussed next) and $e$ (discussed above).

$$\big(n_\text{v} \mid \{e, p\}\big) \sim \text{binom}(t_\text{v}, p(1 - e))$$

Now, we discuss the Bayesian framework for *all* individuals (i.e. "Incidence.")

We have mean mean and precision assigned priors as follows:

$$\mu_\text{p} \sim \text{betaMP}(0.1, 10), \text{ and } \lambda_\text{p} \sim \exp(0.001)$$

Similarly, we have the likelihood for Incidence given these parameters:

$$\big(p \mid \{\mu_{\mathrm{p}}, \lambda_{\mathrm{p}}\}\big) \sim \mathrm{betaMP}(\mu_{\mathrm{p}}, \lambda_{\mathrm{p}})$$

Then, the number of infections in the control group:

$$\big(n_{\mathrm{c}} \mid p\big) \sim \mathrm{binom}(t_{\mathrm{c}}, p)$$

## Part 1

Expand the model in Equation 1 into a hierarchical model that covers both vaccines. The parameters $\{\mu_{\mathrm{e}}, \lambda_{\mathrm{e}}, \mu_{\mathrm{p}}, \lambda_{\mathrm{p}}\}$ must be shared across vaccines.

In contrast, each vaccine must have its own $(e, p)$ pair.

First, we note the hyper-parameter choice for each of these variables. For uniform $\mu_{\mathrm{e}}$, we declared parameters $\{a_{\mathrm{e}}, b_{\mathrm{e}}\}$ as $\{0, 1\}$. Similarly, we declared $\lambda_{\mathrm{e}}$ and $\lambda_{\mathrm{p}}$ as exponential random variables with parameters $\ell_{\mathrm{e}} = \ell_{\mathrm{p}} = 0.001$. Finally we declared $\mu_{\mathrm{p}}$ as a Beta MP random variable with parameters $\{\vartheta_{\mathrm{p}}, s_{\mathrm{p}}\}$ as $\{0.1, 10\}$. We include these hyper-parameters as constants in our Graphical model, to reflect the fact that these choices were made by us, the designers, and could be changed.

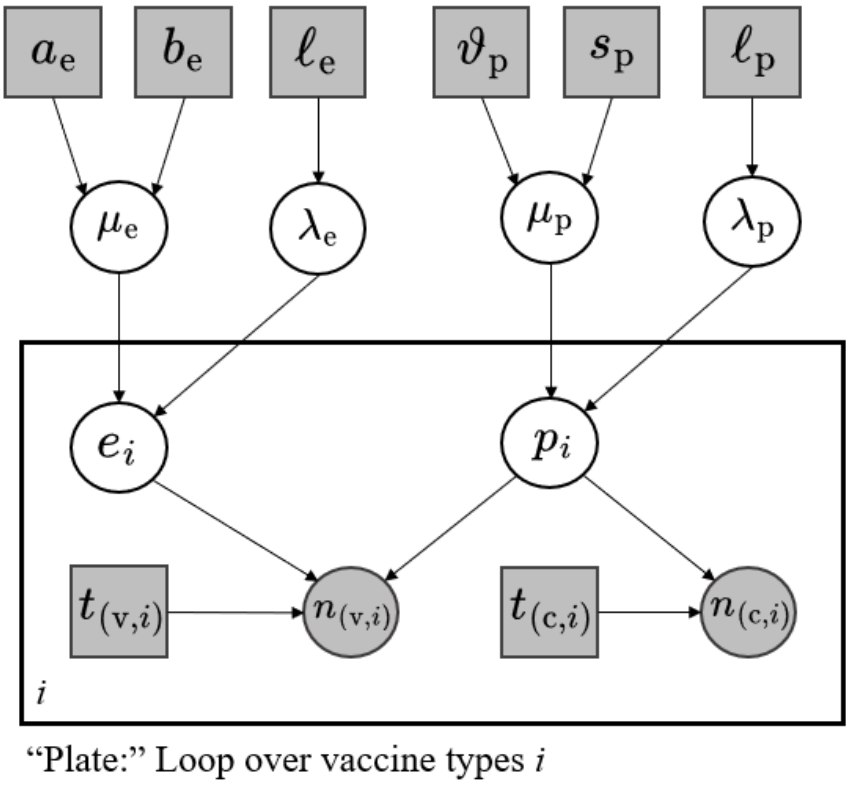


Figure 1: Graph Model of Vaccine Hierarchy

## Part 2

Now, with this framework presented, we will inspect the data before implementing Hierarchical Model to describe the setting.

```
df = present = read.csv("vaccines.csv")
present$IR = round(df$numbersOfCases / df$groupSizes, 4)
colnames(present) = c("Trials", "Arms", "Group Sizes",
                      "Number of Cases", "Infection Rate")
```

We need to do a `simPPLe` implementation for this, too. The PPL function should return the indicator that Moderna is more effective than Pfizer. We include a hidden code cell that defines the `BetaMP` Distribution.

```
# Get each Vaccine Name for Iteration
vaccines = unique(df$trials)

# define PPL
vaccinePPL = function(){
  # generate prior realizations
  mu_e = simulate(Unif(0, 1))
  lambda_e = simulate(distr::Exp(0.001))
  # using BetaMP function
  mu_p = simulate(BetaMP(0.1, 10))
  lambda_p = simulate(distr::Exp(0.001))
  # create an effectiveness vector
  effectiveness = c()
  # iterate through plates
  for (i in vaccines){
    e = simulate(BetaMP(mu_e, lambda_e))
    p = simulate(BetaMP(mu_p, lambda_p))
    # get totals for this vaccine group
    t_v = df[df$trials == i & df$arms == "vaccinated", "groupSizes"]
    # get the total number of infections for this group
    n_v = df[df$trials == i & df$arms == "vaccinated", "numbersOfCases"]
    # observe the Binomial Random Variables
    observe(as.numeric(n_v), Binom(size = as.numeric(t_v), prob = p*(1 - e)))
    effectiveness[i] = e
  }
  # fetch values from vector
  moderna = effectiveness[2]
  pfizer = effectiveness[1]
  # return the indicator : a posteriori, this is P(e_Moderna > e_Pfizer)
  return(as.double(moderna > pfizer))
}
vaccinePPL()
```

```
## [1] 0
```

## Part 3

We have already loaded the data frame, and we coerce the values to the correct type within `vaccinePPL`.

Here is the loaded data frame:

```
kable(present)
```

| Trials | Arms | Group Sizes | Number of Cases | Infection Rate |
|---|---|---|---|---|
| Pfizer-BioNTech | vaccinated | 18198 | 8 | 0.0004 |
| Pfizer-BioNTech | control | 18325 | 162 | 0.0088 |
| Moderna-NIH | vaccinated | 14134 | 11 | 0.0008 |
| Moderna-NIH | control | 14073 | 185 | 0.0131 |

## Part 4

Run your model in `simPPLe` using posterior with $M = 20,000$ iterations, and report the estimated posterior probability that Moderna is more effective than Pfizer. Is there strong evidence for this claim? (Note: it takes a couple of minutes to run 20,000 iterations. Therefore, for prototyping it may be easier to test your implementation with fewer iterations.)

Note that all we need to do is run our simulator, and then make our conclusions from the result. Why is this? Let $\mathcal{M}_e$ be Moderna effectiveness and $\mathcal{P}_e$ be Pfizer effectiveness. Let $Y$ be a short-hand for the data and hence $y$ is the realization of our data. The PPL returns the indicator $\mathbb{1}(\mathcal{M}_e > \mathcal{P}_e \mid Y = y)$. This, in effect, is our function $g(\mathcal{M}_e, \mathcal{P}_e)$ for the forward simulator.

Hence, we know that the following is the result of `posterior` called on `vaccinePPL`:

$$\hat{G}_M \overset{\text{appx.}}{\longrightarrow} \mathbb{E}_\pi\big[g(\mathcal{M}_e, \mathcal{P}_e)\big] = \mathbb{E}\big[\mathbb{1}(\mathcal{M}_e > \mathcal{P}_e \mid Y = y)\big] = \mathbb{P}(\mathcal{M}_e > \mathcal{P}_e)$$

So, let's run some simulations! We will test for $M \in \{1000, 5000, 10000, 20000\}$:

```
set.seed(42)
posterior(vaccinePPL, 1000)
```

```
## [1] 0.4231814
```

```
set.seed(42)
posterior(vaccinePPL, 5000)
```

```
## [1] 0.3069552
```

```
set.seed(42)
posterior(vaccinePPL, 10000)
```

```
## [1] 0.3413241
```

```
set.seed(42)
posterior(vaccinePPL, 20000)
```

```
## [1] 0.3523511
```

**Comments**: Is there strong evidence for the claim that Moderna has better effectiveness than Pfizer?

From these trials, our estimated posterior probability that Moderna is better than Pfizer seems to be converging to approximately 0.35. In other words, given these data, there's roughly a 35% chance that the Moderna Vaccine is better at preventing COVID-19 infection that the Pfizer Vaccine. This doesn't give strong evidence for the claim that Moderna is better. Notably, this doesn't mean we can automatically conclude that Pfizer is better either. We would have to conduct additional tests and simulations to deem whether or not there's evidence to suggest inferior or equivalent effectiveness. All we can conclude from these simulation results is that there is not strong evidence suggesting that the Moderna vaccine is better than the Pfizer vaccine (given these trials.)