Prof. Alex Rogers
Department of Computer Science
University of Oxford
22nd April 2023

Probabilistic Programming with PyMC

Vaccine Efficacy

Pfizer: Over 43,000 participants got the two doses of the vaccine or placebo. The same number of subjects was assigned to the treatment and control group. An efficacy of 95% implies that among 170 confirmed cases of COVID-19, 8 of them observed in the vaccine group.

Moderna: The vaccine is being tested in 30,000 people. Half received two doses of the vaccine, and half received a placebo. Of the 95 cases of covid-19, 90 were in the group that received the placebo.

AstraZeneca regimen 1: Regimen 1(first a half dose and at least a month later a full dose) with 2742 participants showed 90% efficacy implies that among 37 confirmed cases of COVID-19, 3 of them observed in the vaccine group.

AstraZeneca regimen 2: Regimen 2 (two full doses at least one month apart) with 8896 participants showed 62% efficacy implies that among 94 confirmed cases of COVID-19, 26 of them observed in the vaccine group.

Vaccine Efficacy

Assume some unknown probability that trial participants exposed to infection will test positive during the trial.

Assume that those given the vaccine have the same risk of exposure but will only test positive if the vaccine fails to protect them.

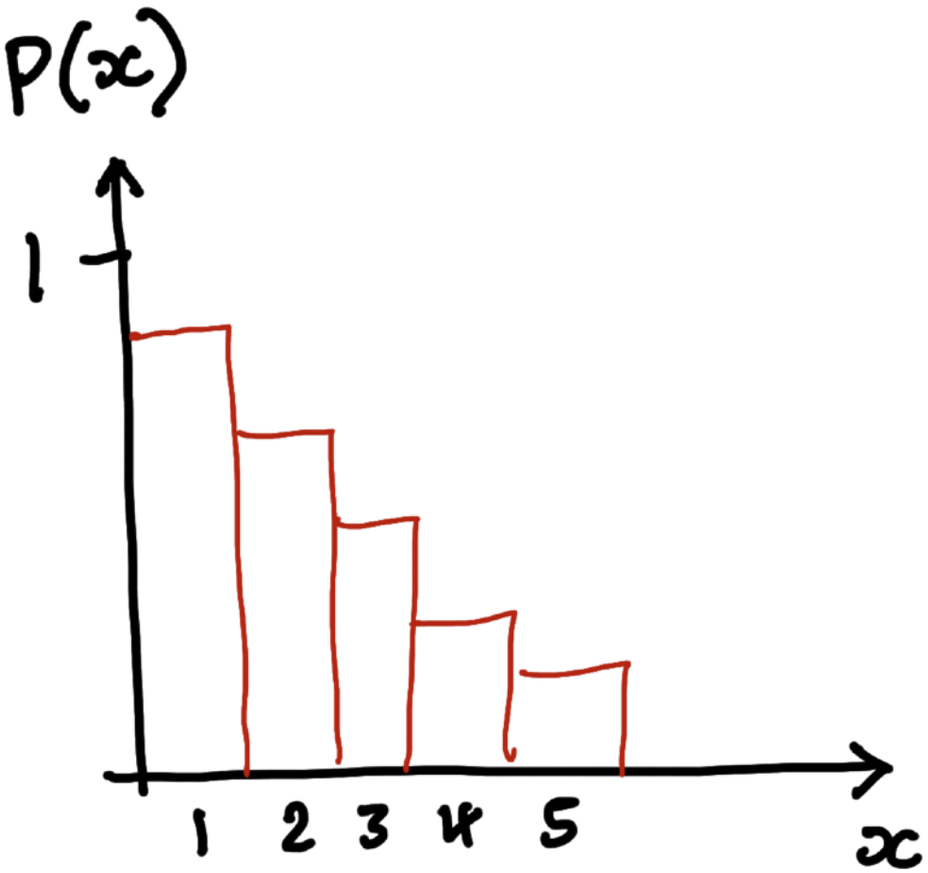
What is the probability of protection of each vaccine?

vaccine_efficacy.ipynb

Probability Distributions

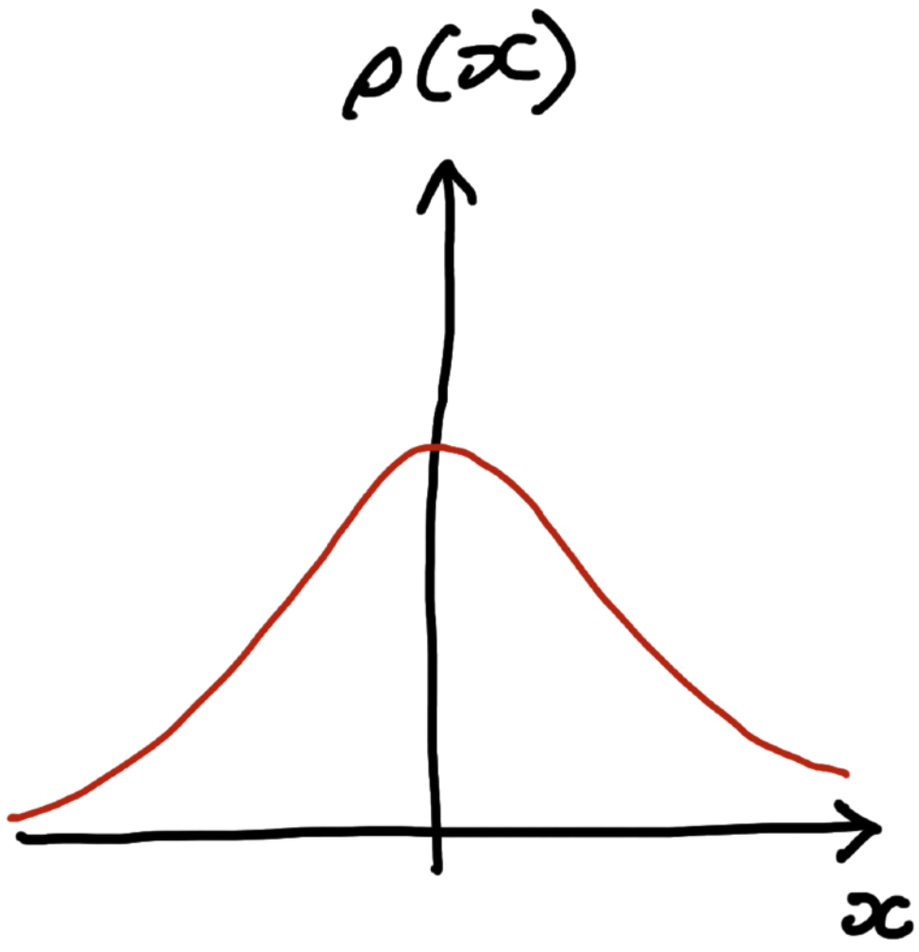
Discrete and Continuous Distributions

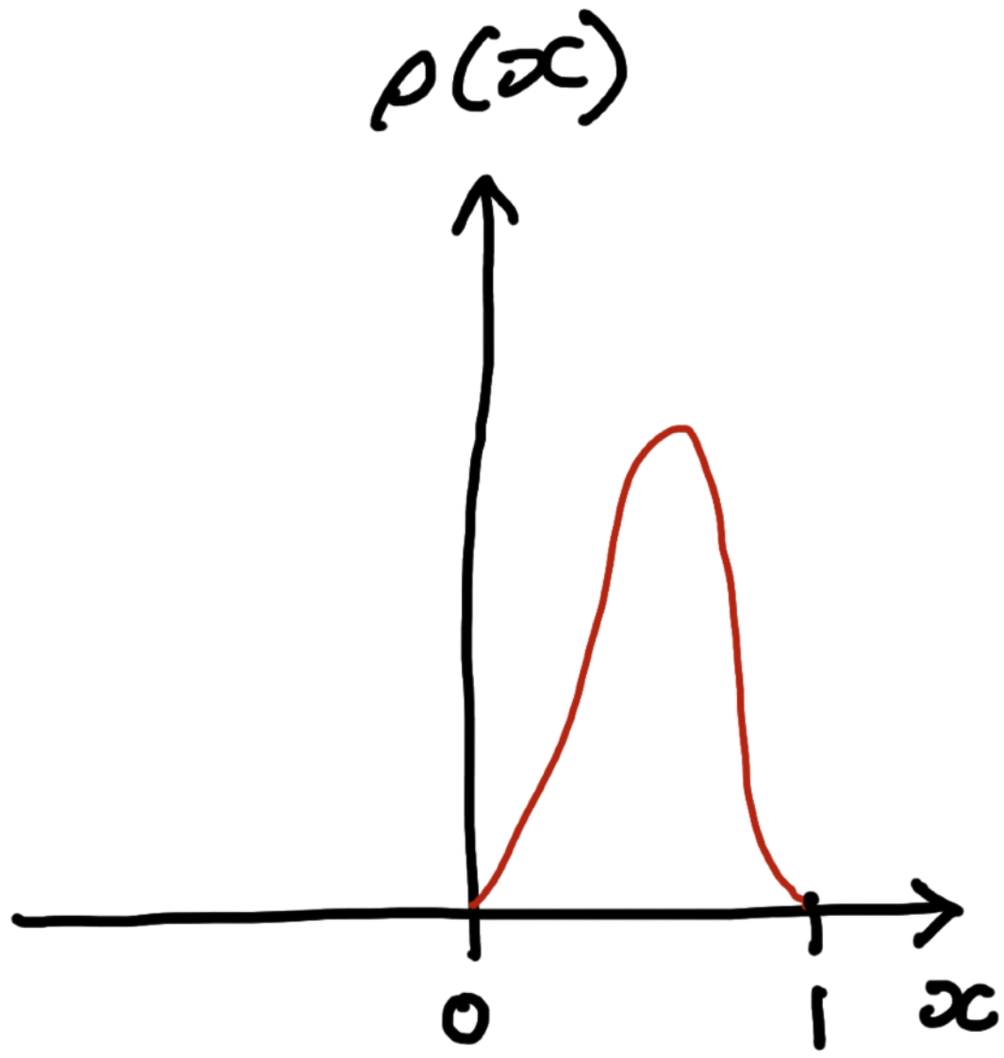
A key distinction between the probability distributions that we will consider is whether they are continuous or discrete.



Distributions with Infinite or Bounded Support

We also need to consider the support of a distribution. Continuous distributions may have infinite or bounded support.





Random Variables

When we define variables in PyMC3 we are really defining random variables. The random variable has a distribution from which its actual value will be derived at some future time.

The notation for random variables is often confusing but in general a tilde is used to indicate that the random variable derives its value from the given distribution:

$$X \sim \text{Beta}(3, 5)$$

We may also have multiple such values and use a subscript to indicate this:

$$x_i \sim \text{Beta}(3, 5)$$

random_variables.ipynb

Measurement Noise

We may also make actual observations of realised values of random variables. It is this feature that allows us to describe likelihood functions and to calculate the posterior probability density function of prior distributions:

Assume we make repeated measurements of some physical parameter. Each time we do so, we make small measurement errors. We normally assume that these measurement errors are Gaussian.

$$x_i \sim \text{Normal}\left(\mu, \sigma^2\right)$$

There is some true value of the physical parameter that we are trying to measure, given by μ. Our noisy measurement process is described by the standard deviation σ.

Normal Distribution

The normal distribution is a symmetrical continuous infinite probability density function described by its mean and variance:

$$\rho(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

https://en.wikipedia.org/wiki/Normal_distribution

https://docs.pymc.io/api/distributions/continuous.html

continuous_distributions.ipynb

Prior Distributions

To complete our model we need to describe the prior distributions for the two unknown parameters.

We know that the standard deviation has to be positive. Depending on what we are measuring the expected value may also be strictly positive.

We may know something about the measurement process and have a good intuition for suitable ranges. We should try to incorporate all our domain knowledge into the model and the prior distributions.

We should be careful about forcing low probability outcomes to be zero probability!

Exponential Distribution

The exponential distribution is described by a continuous probability density function with positive support defined by a single parameter.

$$\rho(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x \leq 0 \end{cases}$$

https://en.wikipedia.org/wiki/Exponential_distribution

https://docs.pymc.io/api/distributions/continuous.html

Gamma Distribution

The gamma distribution is described by a continuous probability density function with positive support defined by two parameters.

$$\rho(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

https://en.wikipedia.org/wiki/Gamma_distribution

https://docs.pymc.io/api/distributions/continuous.html

continuous_distributions.ipynb

Measurement Noise

Putting it all together we can declare our model with a normal distribution describing our measurement process and continuous distributions representing our prior beliefs about the parameters of this normal distribution.

Given some observations we can use Markov chain Monte Carlo to infer the posterior probability distribution of these parameters.

measurement_noise.ipynb

Bayesian Linear Regression

In standard linear regression we minimise the squared vertical distance between the data points and a straight line.

When we do this we are assuming random Gaussian noise on each vertical axis measurement:

$$y = \alpha + \beta x$$

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

The normal calculation for least-squares linear regression corresponds to finding the maximum likelihood values of the parameters of the straight line.

We can describe this explicitly, and consider more complex cases, using probabilistic programming.

linear_regression.ipynb

Model Building Methodology

1. Understand the setting and the scientific question being asked of the data.

2. What are the sources of noise and imprecision in the process being considered.
    - What likelihood function is appropriate?
        - Are we making error-prone continuous measurements or are our observations discrete?
        - Do we know anything about the expected accuracy of the measurements?

3. Think about what prior information is available.
    - How should that be represented within the model?
        - What prior probability distributions should we use?
        - Are random variable discrete or continuous?

Next Time

Applying Probabilistic Programming to Help Reduce Energy Consumption in Homes