Session 3 - Probabilistic Distributions and PyMC3 - Exercises

1. Use the knowledge that the Dirichlet distribution is identical to the beta distribution in the special case that $k = 2$ to derive a relationship between a Gamma function and a factorial.

2. For the example of measurement noise in the notes, confirm that a Gamma distribution can be used as a prior for both the value of the physical property being measured and also the standard deviation of measurement noise. Propose suitable values for the parameters of these Gamma distributions.

3. A supermarket is trying to estimate the number of customers who will visit a store each day and uses a Poisson process to model the arrival of customers. If 40 people visit the store on the first day, how many are expected to visit the store on any day? What is the uncertainty in this estimate? If on subsequent days 30, 60, and 50 customers actually do visit the store, how does the estimate of the expected number of customers change?

4. A common probabilistic model is a Gaussian mixture model where there are two possible Gaussian distributions from which a measurement is made and a binary latent variable is used to indicate which one any measurement actually comes from. This is similar to the example from Session 1 where rather than two independent Gaussians, one Gaussian and an offset was used to explain the height measurements observed.

   (a) Modify the example from Session 1 to be a true mixture of two independent Gaussians, each with their own mean and variance. Generate some suitable data and show that your model is able to infer the correct properties of the two Gaussian distributions.

   (b) The mixture model in Session 1 used the following Bernoulli distribution to describe whether a measurement fell into each category:

   ```
   male_or_female = pm.Bernoulli('male_or_female', 0.5, shape=N)
   ```

   What does `shape=N` do here? What happens if it is removed? Why might this be an appropriate thing to do in some models?

Session 3 - Probabilistic Distributions and PyMC3 - Solutions

1. We know from session 2 that if we do N Bernoulli trials and see n successes then the two parameters of the beta distribution are given by $a = n + 1$ and $b = N - n + 1$. Now, in session 3 we saw that the multinomial distribution is given by:

$$\rho(n_1, \ldots, n_k | p_1, \ldots, p_k) = \frac{N!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i} \tag{1}$$

   and that the Dirichlet distribution is given by:

$$\rho(x_1, \ldots, x_k | a_1, \ldots, a_k) = \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k} \Gamma(a_i)} \prod_{i=1}^{k} x_i^{a_i - 1} \tag{2}$$

   It follows then that $a_i = n_i + 1$ and thus for the factors to be equal:

$$\Gamma(x) = (x - 1)! \tag{3}$$

2. We can replace the exponential probability density functions with gamma probability density functions.

```
mu = pm.Gamma("mu", 15.0, 1.0)
sigma = pm.Gamma('sigma', 2.0, 2.0)
```

3. After the first observation the estimate is 39.9 +/- 6.2 customers per day. After three more days, the estimate is updated to 45.0 +/- 3.4 customers per day.

4. See notebook.