Bayesian Data Analysis Assignment 1

Benjamin Cox, S1621312

Question 1

a)

Our probability vector is $\theta = (\theta_1, \ldots, \theta_6)$ and our outcome vector is $c = (c_1, \ldots, c_6)$. We are drawing from a multinomial distribution (in the same way that 10 Bern(p) trials are the same as 1 Bin(10,p) trial distributionally), ie

$$c \sim \text{Multinomial}(120, \theta).$$

Therefore the likelihood of θ given c with n trials is the following:

$$L(\theta|c) = \frac{n!}{c_1!c_2!\cdots c_6!}\theta_1^{c_1}\cdots\theta_6^{c_6}.$$

A suitable conjugate prior for this would be the Dirichlet distribution (K is the number of possible outcomes, in our case 6),

$$f(x|\alpha, K) = \frac{\Gamma(\sum_{i=1}^{K}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{K}x_i^{\alpha_i-1}.$$

The Jeffrey's prior for the multinomial corresponds to a Dirichlet distribution with

$$\alpha_i = 1/2 \ \forall i \in \{1, \ldots, K\}.$$

b)

Our posterior distribution for θ is

$$\begin{aligned} p(\theta|c) &\propto \left(\frac{\Gamma(3)}{\Gamma(0.5)^6}\theta_1^{-1/2}\cdots\theta_6^{-1/2}\right)\left(\frac{n!}{c_1!c_2!\cdots c_6!}\theta_1^{c_1}\cdots\theta_6^{c_6}\right) \\ &= \text{Dirichlet}\left(\alpha = c + 0.5, K = 6\right). \end{aligned}$$

The expected value of the Dirichlet Distribution is given by $\mathbb{E}[X_i] = \frac{\alpha_i}{\sum\alpha_i}$, so in our case

$$\mathbb{E}[\theta_i|c] = \frac{c_i + \frac{1}{2}}{\sum c_i + 3}.$$

This corresponds to values of

$\mathbb{E}[\theta_1]$	$\mathbb{E}[\theta_2]$	$\mathbb{E}[\theta_3]$	$\mathbb{E}[\theta_4]$	$\mathbb{E}[\theta_5]$	$\mathbb{E}[\theta_6]$
0.142	0.199	0.183	0.142	0.199	0.134

We are going to compute symmetric 95% credible intervals for each θ_i, hence we must marginalise them. We could (theoretically) calculate a 95% credible region in the 6 dimensional parameter space, but this would get extremely complicated really quickly, and would also be hard to interpret.

Fortunately the marginal distributions of the Dirichlet are a lot easier, as they are beta distributions. Write $\alpha_0 = \sum \alpha_k$, then we have

$$\theta_i \sim \text{Beta}\left(\alpha_i, \alpha_0 - \alpha_i\right).$$

We can substitute in our expressions for α_i to obtain

$$\theta_i \sim \text{Beta}\left(c_i + 1/2, c_0 - c_i + 2.5\right).$$

Using this result we obtain the following 95% credible intervals:

θ	Lower	Upper
θ_1	0.08657456	0.20897331
θ_2	0.1337529	0.2738778
θ_3	0.1200040	0.2556043
θ_4	0.08657456	0.20897331
θ_5	0.1337529	0.2738778
θ_6	0.08007849	0.19945622

c)

We are going to simulate a large number of draws from our posterior for θ given our results. We are going to define a 'range of practical equivalence' around $\theta_i = 1/6$, with leeway of $1/30$. That is, we are going to test how many draws have all $1/6 - 1/30 < \theta_i < 1/6 + 1/30$.

We draw 10,000 times from our posterior and calculate that with probability 0.036 all probabilities are within our equivalence range. Therefore we can say that the dice is unlikely to be fair.

We chose $1/30$ as our range as it is a fifth of our fair probability, thus allowing for some discrepancy but not so much as to falsely not reject the null hypothesis. If we chose 0.05 as our range we would obtain a probability of 0.19372. However 0.05 is quite large compared to our probabilities, so is not a suitable choice.

d)

The posterior predictive distribution is the 'Dirichlet-Multinomial' distribution. The pmf for this is given by

$$f(x|n,\alpha) = \frac{n!\Gamma(\sum \alpha_i)}{\Gamma(n + \sum \alpha_i)} \prod_{k=1}^{K} \frac{\Gamma(x_k + \alpha_k)}{(x_k!)\Gamma(\alpha_k)}$$

for n the number of trials and $alpha_1, \ldots, \alpha_k > 0$.

Taken as our posterior predictive under the Jeffrey's prior we have

$$c_{\text{new}} \sim \text{DirMNom}(60, \text{c} + 1/2).$$

We can simulate from this. We draw 10,000 times from this distribution and find that with probability 0.737 we have more 5s than 6s in our next 60 trials.

e)

We incorporate these into our likelihood, denoting the new count vector as d. Our new posterior is

$$\theta \sim \text{Dirichlet}(c + d + 1/2, 6).$$

Our new posterior means are

$\mathbb{E}[\theta_1]$	$\mathbb{E}[\theta_2]$	$\mathbb{E}[\theta_3]$	$\mathbb{E}[\theta_4]$	$\mathbb{E}[\theta_5]$	$\mathbb{E}[\theta_6]$
0.163	0.183	0.216	0.138	0.167	0.138

with 95% marginal credible intervals given by

θ	Lower	Upper
θ_1	0.1189814	0.2113785
θ_2	0.1371556	0.2340370
θ_3	0.1667039	0.2698204
θ_4	0.0975285	0.1838316
θ_5	0.1225962	0.2159303
θ_6	0.0939960	0.1791973

The credible intervals have narrowed, as expected for more observations. It is of note that the new credible interval for θ_3 (barely) does not contain the value required for a 'fair' dice. This is good evidence that the dice is not fair.

Question 2