4.1 Exercises

1. Calculate the result of Monty Hall problem assuming that there are N boxes instead of three.

2. What is the *prior predictive* distribution for the number of red balls in N draws, if you assume an 'infinite bag' with proportion of red balls $r \in [0, 1]$ and a $\mathrm{U}(0, 1)$ prior for r? You can use these results: beta function is $\mathrm{beta}(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathbf{d}t = \Gamma(x)\Gamma(y)/\Gamma(x+y)$, and $\Gamma(n) = \Gamma(n-1)(n-1)$, and $\Gamma(n) = (n-1)!$. (This prior predictive distribution was actually used by T. Bayes to justify the uniform prior).

3. Continue problem 2 in Exercise 3.1. by computing the approximate posterior density using the normal approximation. Plot both densities.
Hint: `plot(la,dgamma(la,a,b),type='l')`, `plot(la,dnorm(la,mu,sig),type='l')`.

4. Write out the *principle* (as a draft/pseudo code of the algorithm) for computing the HPD interval for a unimodal density, e.g. Beta-density. (You can implement the algorithm too if you like, e.g. using R). Hint: you can imagine the posterior density as a mountain under the ocean, and then the sea level is gradually lowered.

5. Using the R-function (shown in earlier lecture notes) for calculating posterior of N in a binomial(N, r) problem with known X and r, calculate the 95% 'HPD interval' (although the "interval" is a set of integer values). Hint: use `A<-sort(p,index.return=TRUE)`, then `attach(A)`, and then check what is `x` and `ix`. Operate with `p` and `ix`. Assume the data: `X<- 1;r<- 0.2;M<- 100`.

6. Disease monitoring. The unknown population prevalence of a disease is p. A random sample of N individuals is drawn from the population. Each individual is tested for the disease (resulting to '+' or '-'), but the test has sensitivity $p_1 = 0.8$, and specificity $p_2 = 0.9$. In other words: $p_1 = P(+ \mid \text{disease})$, $p_2 = P(- \mid \text{no disease})$. The data consist of the number of test positives, $X = 2$, among $N = 100$ tested. The probability of test positive is then $\theta = pp_1 + (1-p)(1-p_2)$. So, the conditional distribution of the data is Binomial(N, θ). Previously we solved the posterior of θ (assuming $\mathrm{U}(0,1)$-prior) as $\mathrm{Beta}(X+1, N-X+1)$, but now due to sensitivity and specificity: $0.1 = (1-p_2) \leq \theta \leq p_1 = 0.8$. Hence, we must assume $\mathrm{U}(1-p_2, p_1)$-prior. The posterior of θ is then

$$\pi(\theta \mid X, N, p_1, p_2) = \frac{\mathrm{Beta}(\theta \mid X+1, N-X+1)}{\int_{1-p_2}^{p_1} \mathrm{Beta}(\theta \mid X+1, N-X+1)\mathbf{d}\theta} \qquad , \theta \in [1-p_2, p_1]$$

Calculate the posterior density of p numerically in R. Hint: use the theorem of variable transform for $p = g(\theta)$. For evaluating the posterior of θ shown above, calculate the normalizing constant in R by `C<-integrate(dbeta,1-p2,p1,X+1,N-X+1)`, then `attach(C)`, and then apply `value` for the constant. Write the posterior of θ as a function in R, and use that for numerical calculation, applying the variable transform theorem.