Exercises IV

1. Let $\pi(X \mid n, p) = \text{Bin}(n, p)$ and the prior of p is $U(0, 1)$. Show that the prior predictive distribution of X is discrete uniform distribution: P(X=i)=1/(n+1) for all $i \in \{0, 1, 2, \dots, n\}$. This shows that the flat Bayes-Laplace prior implies a flat predictive distribution too.

2. Observed monthly number of sunspots in 2013 were `list(spots = c(63,38,58,72,79,52,57,66,37,86,78,90))`. (Actually, these are smoothed monthly values http://www.ips.gov.au/Solar/1/6). Assuming Poisson model for the monthly numbers, $P(X_i \mid \lambda) = \frac{\lambda^{X_i}}{X_i!}\exp(-\lambda)$, and Gamma prior $\pi(\lambda) = \text{Gamma}(\alpha, \beta)$, (with $\alpha = \beta \approx 0$), solve the posterior density for the mean of sunspots: $\pi(\lambda \mid X_1, \dots, X_{12})$, and compute the 95% Credible Interval for it (hint: in R: `qgamma(c(0.025,0.975),,)`). Note that the posterior is not proper density if "Gamma(0,0)" improper prior is chosen and if the observed count happens to be zero. To prepare for all possible scenarios, including zero counts, a prior could also be chosen as $\pi(\lambda) = U(0, L)$ with very large $L \approx \infty$. Show that the posterior then exists even with a single data point $X = 0$. With the sunspot data, show that the posterior becomes nearly the same with either prior: "$U(0,\infty)$" or "Gamma(0,0)". Finally, solve the mean and variance for the predicted number of sunspots for "next month", based on posterior predictive distribution. (Hint: lecture slides on 'Predictive distributions', last page).

3. In estimating population prevalence p, a large sample of individuals is collected (without replacement) so that the sample size is nearly as large as the population size. In such situation, binomial model may not be appropriate. Instead, a hypergeometric distribution is often recommended for the number of positives x in the sample of size n, from population of size N, where $x < n < N$, and n nearly as large as N. Regardless of what prior is chosen, the posterior is not as easy to compute as with the binomial sampling model. Assuming a uniform prior for unknown population prevalence p, compute the posterior $\pi(p \mid x_{\text{obs}}, n, N)$ using Monte Carlo method with rejection sampling applied for the joint distribution of p, x. That is: use method of composition to sample 'in tandem' first p and then x conditionally on p, then selecting those outcomes where x matches the observed data value x_{obs}. This can be done in R with the code below. Investigate how large the Monte Carlo sample should be to get reasonably accurate distribution that would be based on enough many accepted points (where $x = x_{\text{obs}}$). Make the prior more narrow by adjusting its length. This could be justified by prior knowledge telling us that prevalence is surely below e.g. 50%. Report 95% posterior CI calculated from the sampled values of p. (in R: `quantile(...)`)

```
N <- 500  # population size
n <- 400  # sample size
q <- 0.1  # true prevalence
k <- round(q*N) # actual true positives
xobs <- rhyper(1,k,N-k,n) # generate observation, check its value!

mc <- ....  # number of monte carlo samples, CHOOSE BIG ENOUGH
p <- numeric(); x<-numeric() # vectors for p, x
 # generate a sample, (size mc), of p drawn from U(0,1):
 p <- runif(mc,0,1)
 # generate vector of x, drawn from hyperg:
  # one element in x per each value of p:
  x <- rhyper(mc,round(p*N),N-round(p*N),n)
  # select those p where x=xobs and plot the posterior:
  plot(density(p[x==xobs]),main="posterior density of p")
```

4. Are the following predictive models for X_1, X_2, X_3 equivalent or is there a difference? Explain your findings. You can either (1) run the code in BUGS to see it empirically and explain in words, or (2) you can analyze mathematically by solving e.g. predictive mean and predictive variance (for the analytic approach, see slides on 'Predictive distributions', last page, and also the 'preliminary pages' of the longer text, and the expressions of mean and variance of beta-distribution, bernoulli-distribution and binomial distribution).

```
model{
# model 1:
x[1] ~ dbin(p,n)
p ~ dbeta(a,b)
# model 2:
x[2] <- sum(x0[1:n])
for(i in 1:n){ q[i] ~ dbeta(a,b); x0[i] ~ dbern(q[i]) }
# model 3:
x[3] ~ dbin(mp,n); mp <- a/(a+b)
}
# data:
list(n=20,a=1,b=1)
```