Department of Mathematics and Statistics
Introduction to Bayesian methods and WinBUGS
Exam 17.12.2009

1. (6 points).
Assume binary variables X_1 and X_2 are part of an infinite exchangeable sequence.
Let $\theta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i$, as in de Finetti's theorem, with prior density $\pi(\theta)$. Which of the following statements are true (or false)? Explain why.

(A), (2p): $P(X_2 = 1 \mid X_1 = 1, \theta) > P(X_2 = 1 \mid \theta)$
(B), (2p): $P(X_2 = 1 \mid X_1 = 1) > P(X_2 = 1)$
(C), (2p): $P(X_2 = 1) > E(\theta)$

(A) False. Conditionally, given θ, X_i is independent of any other X_j. In other words, if θ would be known exactly, then there is nothing more to be learned about the probability of X_i from any other observables X_j. $P(X_2 = 1 \mid X_1 = 1, \theta) = P(X_2 = 1 \mid \theta) = \theta$.
(B) True. If θ is not given, then knowing about X_i is informative about another X_j. $P(X_2 = 1 \mid X_1 = 1) = \frac{P(X_2=1, X_1=1)}{P(X_1=1)} = \frac{\mathrm{Var}(\theta)+E(\theta)^2}{E(\theta)} > \frac{E(\theta)^2}{E(\theta)} = E(\theta) = P(X_2 = 1)$
(C) False. $P(X_2 = 1) = \int_0^1 P(X_2 = 1 \mid \theta)\pi(\theta)\mathbf{d}\theta = E(\theta) = P(X_i = 1), \forall i$.

2. Explain the meaning of the following terms (1 point each):
(A) Objective Bayesian methods, (B) Prior elicitation, (C) Posterior predictive distribution
(D) Highest posterior density interval, (E) One's trick, (F) Label switching (or aliasing)

(A): Bayesian methods where the prior is chosen to be 'uninformative' according to some principle. An example of such prior is the Jeffreys' prior according to Jeffreys' principle of transformation invariance.
(B): Methods of eliciting subjective priors to represent expert knowledge about some unknown quantity.
(C): Predictive distribution that is conditional on observed data, not conditional on unknown parameters. This is obtained by 'integrating out' the unknown parameter: $P(X \mid \text{data}) = \int_\Theta P(X \mid \theta)P(\theta \mid \text{data})\mathbf{d}\theta$ where $P(X \mid \theta)$ is the parametric model of X and $P(\theta \mid \text{data})$ is the posterior of θ.
(D): Shortest possible interval that contains the given probability. This can also be a collection of distinct intervals if the density is multimodal.
(E): Method of expressing the probability of an event (or logical statement) when this has no direct representation among the probability distributions available in WinBUGS. For example, $X \sim \mathrm{Bin}(p, N)$ could be directly written using `dbin` but also as

```
pr <- exp(logfact(n)-logfact(x)-logfact(n-x)+x*log(p)+(n-x)*log(1-p))
one ~ dbern(pr); one <- 1
```

(F): In mixture distribution models, the parameters and weights of the mixture components can be switched symmetrically so that the overall probability is not changed. This creates an identifiability problem which can also be seen by monitoring the Monte Carlo sample path of the parameters and weights. They can arbitrarily take over each other's positions which means there is no proper convergence in simulations. The problem is avoided by adding e.g. an ordering constraint for means $\mu_1 < \mu_2$ to 'fix the labels'.

3. (6 points):
A test for a rare disease (say, ≈ 0.001 prevalence) is so perfect that if a patient has the disease, the test will show positive result with 99% probability (=sensitivity). But the test also gives a positive result with 10% probability (=1-specificity) when the patient does not have the disease. A patient, Mr Illmore, was tested positive but the probability that he really has the disease is still quite low, why? What is required to have this probability larger than 50% and can we improve by increasing sensitivity? What are the implications for e.g. population-wide cancer screening? (e.g. prostata or breast cancer).

The probability is low because the prevalence of the disease is low. This can be seen by calculating the conditional probability $P(\text{disease} \mid \text{test}+)$ (by application of Bayes' formula). This cannot be much improved even if the sensitivity would be 100%. Additionally, the chance to get positive result for a healthy should be reduced to about 0.001. Thus, only increasing specificity could improve, or targeting the testing to high risk groups where the prevalence is known to be greater. This is the practical implication for e.g. targeting at specific age groups instead of whole population.

4. (6 points):
Assume we want to compute approximately the expected value $E_\pi(h(\theta))$ of some $h(\theta)$, with respect to some density π of the variable θ. What is the mathematical principle of rejection sampling (2p), importance sampling (2p), direct Monte Carlo sampling (1p), and MCMC sampling (1p) for doing that? What are the differences of these approaches?

In rejection sampling, an easy-to-sample distribution g is used for generating proposed random variables. There has to exist a constant M such that $\pi(x) < Mg(x), \forall x$. Generated values are accepted with probability $\pi(x)/(Mg(x))$, otherwise they are rejected. Finally, the expected value $E(h(x))$ is approximated by the average of the accepted draws $\sum_{i=1}^n h(x_i)/n$. In importance sampling, all draws are accepted but they are weighted. An easy-to-sample distribution g is again used, but the expected value $E_\pi(h(x)) = E_g\Big(h(x)\frac{\pi(x)}{g(x)}\Big)$ is approximated by $\frac{1}{n}\sum_{i=1}^n h(x_i)\frac{\pi(x_i)}{g(x_i)}$. In both approaches, the choice of g is important for the efficiency of the method. In direct Monte Carlo sampling of π, independent random draws are produced from π so that no rejections or weighting are needed. A method such as inverse cdf might be used. Again, the expected value is approximated as the average of the draws $h(x_i)$. In MCMC sampling, independent draws are replaced by dependent draws such that the long run average $\sum_{i=1}^n h(x_i)/n$ will approximate $E(h(x))$ but the efficiency will depend on the MCMC method. No draws are rejected but the algorithm can get stuck with current draw (that will be used unless a jump to another point is accepted) for some time. Also, the starting value can have an effect if the run is not long enough. MCMC can be based on Gibbs sampling method where all new draws are accepted, or more general methods where a proposal distribution is used. In the latter case, the efficiency naturally depends on the choice of proposal density.

5. (6 points):
What posterior distribution is computed in the following WinBUGS implementation and how could the code be simplified? (2p). In addition to that, what research question may be addressed by the use of the step function? (2p). What does the Monte Carlo sample mean of 'I' approximate and why is that so? Interpret the result (2p).

```
model{
x ~ dbin(p,n); p ~ dbeta(1,1)
```

```
y ~ dbin(q,m); q ~ dbeta(1,1)
I <- step(p-q)
}
list(x=1,y=3,n=20,m=100)
```

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
p	0.09059	0.05987	6.638E-4	0.01233	0.07814	0.2399	1001	10000
q	0.03925	0.01906	1.998E-4	0.01109	0.03611	0.08409	1001	10000
I	0.7935	0.4048	0.003831	0.0	1.0	1.0	1001	10000

The code will compute two posterior distributions $\pi(p \mid x, n)$ and $\pi(q \mid y, m)$. The code could be slightly simplified by writing the posterior density directly `p ~ dbeta(a,b); a<-x+1; b<-n-x+1`. The research question here could be the comparison of prevalences in two populations based on a sample from both. The Monte Carlo sample mean of `I` is an approximation of the expected value of the indicator variable $1_{\{p \geq q\}}(p, q)$, which is equivalent to the probability that this indicator is one, which in this example is equivalent to the probability of $p \geq q$. In the result, we see that the 95% probability intervals overlap, but there is considerable probability (79%) that $p \geq q$.