Department of Mathematics and Statistics
Introduction to Bayesian methods and WinBUGS
Exam 13.10.2009

1. (6 points). Assume binary variables X_1 and X_2 are part of an infinite exchangeable sequence. Prove mathematically that $P(X_2 = 1 \mid X_1 = 1) > P(X_2 = 1)$.

$$P(X_2 = 1, X_1 = 1) = \int_0^1 \theta^2 \pi(\theta)\mathbf{d}\theta = E(\theta^2) = \mathrm{Var}(\theta) + E(\theta)^2$$

$$P(X_1 = 1) = P(X_2 = 1) = \int_0^1 \theta \pi(\theta)\mathbf{d}\theta = E(\theta)$$

$$P(X_2 = 1 \mid X_1 = 1) = \frac{P(X_2 = 1, X_1 = 1)}{P(X_1 = 1)} = \frac{\mathrm{Var}(\theta) + E(\theta)^2}{E(\theta)} > \frac{E(\theta)^2}{E(\theta)} = E(\theta) = P(X_2 = 1)$$

This shows that learning what the result from the 1st experiment was, leads to higher probability of the same result in the 2nd. (Assuming θ is not known, i.e. all that is assumed is prior exchangeability).

2. Explain the meaning of the following terms (1 point each):
Improper prior distribution
-A prior 'distribution' with an infinite integral, so that it cannot be normalized to make a proper distribution.
Marginal posterior distribution
-A lower dimensional distribution that is obtained from the full posterior distribution by integration. e.g. $\int \pi(x, y \mid \text{data})\mathbf{d}y = \pi(x \mid \text{data})$
Exchangeability (finite or infinite)
-A sequence of variables is said to be exchangeable, if the permutation of the variable indices does not affect the probability statement. e.g. finite exchangeability: $P(x_1, \ldots, x_n) = P(x_{r_1}, \ldots, x_{r_n})$ for all permutations r.
Full conditional distribution
-This is needed in Gibbs-sampling algorithms. Assume the full distribution is d-dimensional $\pi(x_1, \ldots, x_d)$. The full conditional, say for x_1, is $\pi(x_1 \mid x_2, \ldots, x_d)$, and likewise for all other components.
DAG
-A graphical (acyclic) representation of all conditional distributions in a Bayesian model, that will define (together with priors) all that is needed to compute a complete posterior distribution. (Directed Acyclic Graph).
Realized residual (bayesian)
-The realized residual for data point y_i is $y_i - E(y_i \mid \theta)$ where the expected value of y_i depends on the unknow model parameter θ for which the posterior is computed. Hence, for fixed data y, the realized residual will have a distribution that is implied by the distribution of θ. (In contrast to the 'classical' or fitted residual that is based on the (fixed) estimate $\hat{\theta}$).

3. Assume that the herd size distribution f (probability density) is a mixture distribution $f = \alpha f_1 + (1 - \alpha) f_2$ where the expected value and variance of the component distribution f_j are μ_j and σ_j^2. This can be e.g. a mixture of 'small' and 'large' herds so that the expected proportion of small herds is α. In other words, a randomly chosen herd has probability α to be 'small' in which case its

size follows distribution f_1. Otherwise, with probability $1-\alpha$, its size follows distribution f_2. Show that:

(A), (2 points): the expected number of animals within the chosen herd is then $\mu = \alpha\mu_1+(1-\alpha)\mu_2$, and:

(B), (2 points): the variance of the number is $\sigma^2 = \alpha(\sigma_1^2+\mu_1^2)+(1-\alpha)(\sigma_2^2+\mu_2^2)-(\alpha\mu_1+(1-\alpha)\mu_2)^2$.

(C), (2 points): assume that all parameters in these distributions would be known, and $f_j = \text{Normal}(\mu_j, \sigma_j^2)$, and that a herd is observed to have size X. Write the probability that it is a 'small' herd and explain this graphically.

$$\text{Note: density function of a normal}(\mu,\sigma^2) \text{ density: } \frac{1}{\sqrt{2\pi}\sigma}\exp(-0.5(x-\mu)^2/\sigma^2)$$

$$\text{and: } V(X) = E((X-E(X))^2).$$

(A,B): we calculate the mean and variance for the number, X, where this number X is a random variable from a mixture distribution:

$$\mu = \int_0^\infty xf(x)dx = \int_0^\infty x(\alpha f_1(x)+(1-\alpha)f_2(x))dx = \int_0^\infty (\alpha x f_1(x)+(1-\alpha)xf_2(x))dx = \alpha\mu_1+(1-\alpha)\mu_2. \tag{1}$$

$$\sigma^2 = E(x^2)-(E(x))^2 = E(x^2)-\mu^2. \tag{2}$$

To compute this we need to solve $E(x^2)$:

$$E(x^2) = \int_0^\infty x^2 f(x)dx$$

$$= \int_0^\infty x^2(\alpha f_1(x)+(1-\alpha)f_2(x))\,dy$$

$$= \int_0^\infty \alpha x^2 f_1(x)+(1-\alpha)x^2 f_2(x)\,dy$$

$$= \alpha E_1(x^2)+(1-\alpha)E_2(x^2)$$

$$= \alpha(\sigma_1^2+\mu_1^2)+(1-\alpha)(\sigma_2^2+\mu_2^2). \qquad (\text{because } \sigma_i^2 = E_i(x^2)-\mu_i^2)$$

Therefore, by substituting back to equation (2) we obtain:

$$\sigma^2 = E(x^2)-\mu^2$$

$$= E(x^2)-(\alpha\mu_1+(1-\alpha)\mu_2)^2$$

$$= \alpha(\sigma_1^2+\mu_1^2)+(1-\alpha)(\sigma_2^2+\mu_2^2)-(\alpha\mu_1+(1-\alpha)\mu_2)^2.$$

(C): the probability is

$$P(\text{small} \mid X) = \frac{N(X \mid \text{small})P(\text{small})}{N(X \mid \text{small})P(\text{small})+N(X \mid \text{large})P(\text{large})}$$

where $P(\text{small}) = \alpha$ and $P(X \mid \cdot)$ the normal density function (either for 'small' or 'large'). Graphically:

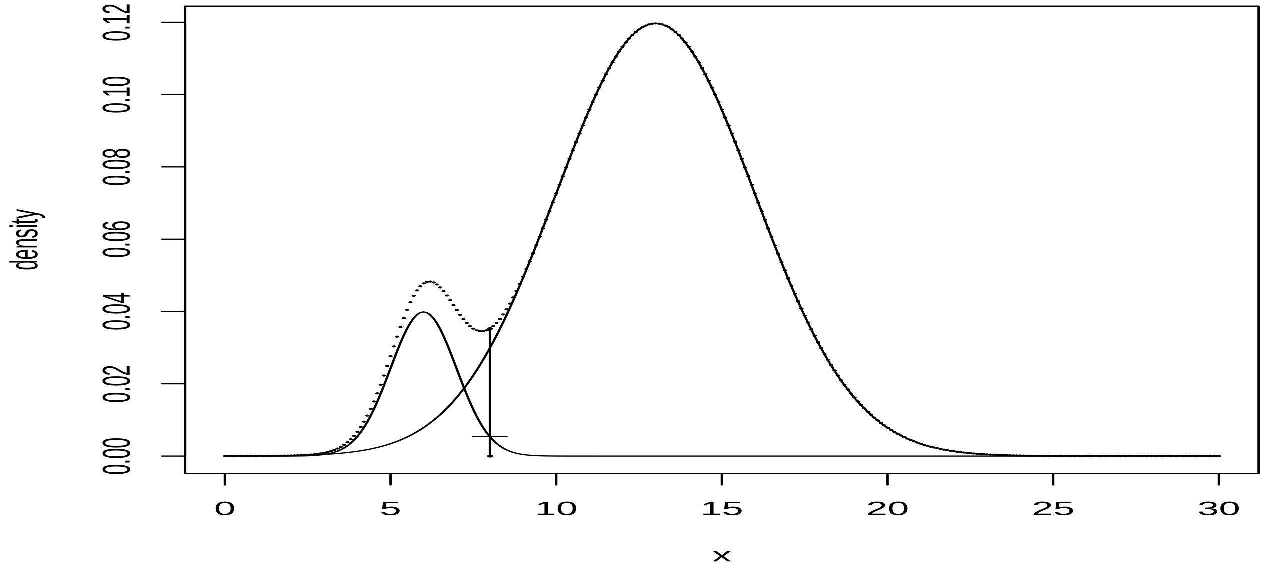

Figure: Mixture density (dotted line) and its two component densities multiplied by their weights $(\alpha f_1, (1-\alpha)f_2)$. Assume observation $X = 8$. The vertical line shows the probability of either class. The stick-length under the mark divided by the whole stick-length is the probability of 'small' class. This is the same as what is mathematically written in the above equation.

4. Normal model $X_i \sim \mathrm{N}(\mu, \sigma^2)$, with $i = 1, \ldots, n$ observations. Assume σ^2 is known, μ is unknown, with prior $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$. Assume that σ_0^2 is **very large** ($\approx \infty$).

(A, 3 points): what is posterior distribution $\pi(\mu \mid X_1, \ldots, X_n, \sigma^2, \mu_0, \sigma_0^2)$ approximately? What is its mean and variance?

Approximately, the posterior is the same as with improper uniform prior, resulting to: $\mathrm{N}(\bar{X}, \sigma^2/N)$, with mean $\bar{X}$ and variance σ^2/N. This can also be derived from the posterior mean and variance (remember what they were, or calculate again) by taking the limit $\sigma_0 \to \infty$. (See lecture notes).

(B, 3 points): what is (approximately) posterior predictive distribution $\pi(X^* \mid X_1, \ldots, X_n, \sigma^2, \mu_0, \sigma_0^2)$ of a new observation X^*? What is its mean and variance?

The posterior predictive distribution results from the model $N(\mu, \sigma^2)$ where the uncertainty about μ is described by the posterior distribution $N(\bar{X}, \sigma^2/N)$. That is: $\pi(X^* \mid X) = \int_{-\infty}^{\infty} \pi(X^* \mid \mu, \sigma)\pi(\mu \mid \bar{X}, \sigma^2/N)\mathbf{d}\mu$. The predictive distribution is again a normal distribution. It suffices to recall this from the lectures. Alternatively, one could find this result as follows: $X^* = (X^* - \mu) + \mu$, so that $\pi(X^* - \mu \mid \mu, \sigma, \bar{X}) = N(0, \sigma^2)$ and $\pi(\mu \mid \sigma, \bar{X}) = N(\bar{X}, \sigma^2/N)$. Now $(X^* - \mu)$ and μ are independent, given $\bar{X}$. Hence the mean $E(X^*) = E(X^* - \mu) + E(\mu) = 0 + \bar{X}$ and variance $V(X^*) = V(X^* - \mu) + V(\mu) = \sigma^2 + \sigma^2/N$. Thus the posterior predictive distribution is normal with mean $\bar{X}$ and variance $\sigma^2 + \sigma^2/N$. Moreover, it would be possible to use here the theorem of conditional expectations and variances, $E(X^*) = E(E(X^* \mid \mu))$ and $V(X^*) = V(E(X^* \mid \mu)) + E(V(X^* \mid \sigma))$, to find out the mean and variance.

Note: density function of a normal(μ, σ^2) density: $\frac{1}{\sqrt{2\pi}\sigma} \exp(-0.5(x-\mu)^2/\sigma^2)$

5. (4+2 points):
(A, 4 points): assume the model is $X_i \sim$ Bernoulli(p) with prior $p \sim$ U$(0,1)$. Write a WinBUGS code for assessing the model fit, $P(T(x^{\text{pred}}) > T(x) \mid x)$, based on a discrepancy T that counts the number of occurrences of *three* consecutive results that have the same value (either 000, or 111). Assume the data:

```
x=c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0)
```

It was not required that the sequences should be distinct; they could be overlapping. (Then '0000' counts as two '000's). The solution for the latter is given below. Essentially, whatever the chosen counting rule, the same rule should be applied to the actual data as well as to the simulated replicate data.

```
model{
p  ~ dunif(0,1)
for(i in 1:n){ x[i] ~ dbern(p); xrep[i]~ dbern(p) }
for(i in 3:n){
test[i-2] <- equals(xrep[i],xrep[i-1])*equals(xrep[i-1],xrep[i-2])
}
T <- sum(test[1:n-2])
P <- 1-step(Tobs-T) # compare with the observed
}
list(n=20,Tobs=12,x=c(1,1,0,0,0,0,0,1,1,1,1,1,0,0,0,0,0,0,0,0))
```

(B, 2 points): Assume a is some fixed value and U is a variable for which a distribution will be computed in WinBUGS. We want to calculate the following probabilities: $P(U \leq a)$, $P(U < a)$, $P(U \geq a)$, $P(U > a)$. What lines are needed to be written in WinBUGS to simulate these probabilities?

Step-function gives 1 when the argument is positive or zero. It gives 0 when the argument is negative.

```
P1 <- step(a-U)
P2 <- 1-step(U-a)
P3 <- step(U-a)
P4 <- 1-step(a-U)
```