1 Examples

1.1 Detecting signal from noise

Reference: D.S.Sivia: Data analysis, a Baysian tutorial, 2nd ed. p.35-42.

Assume we have measurements (counts) y_i taken at points x_i. The simple model is Poisson with parameter e_i. This expected value is assumed to be a combination of background noise and the signal so that:

$$e_i = 100(ae^{-0.5(x_i-x_0)^2/s^2} + b)$$

And we assume the location and the width of the signal to be known, specified by constants $x_0 = 0, s = 1$. The unknown parameters to be estimated are the signal amplitude a and the level of noise b. Generate some data y_i by choosing values for a, b and x_i. Compute in BUGS and study the 2D-posterior distribution of (a, b), based on uniform priors over sufficient range, e.g. $U(0, 1000)$. You can extend the problem by trying to estimate x_0 and/or s too. Try different signals with different number of data points and study how the posterior behaves.

1.2 Estimating within herd prevalence

Reference:
Bollaerts KE et al: Development of a Quantitative Microbial Risk Assessment for Human Salmonellosis Through Household Consumption of Fresh Minced Pork Meat in Belgium. Risk Analysis, 29, (6), 2009. 820:840.

The data are given as number of pig herds with observed sample prevalence in intervals [0, 1/50, 2/50, ..., 49/50, 50/50] which define the bins in a histogram below. There were 48 intervals with nonzero counts. Below is a listing of observed 'values' (observed sample prevalence at the center of each bin can be calculated as value/50-1/100), and the counts (number of pig herds in each bin). The data are given separately for the non-zero bins and for all bins for ease of manipulation in the modeling to be done.

```
# data: interval [0,1] divided into 50 bins, of which 48 have nonzero counts.
list(Nall=50,N=48,
values=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,
26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,48,49,50),
counts=c(32, 20, 33,  3, 15, 14, 25,  2, 35, 12,  7, 25, 13, 17,  7, 15, 30,
3, 10, 10, 15,  2,  7, 12, 18,  2,  4, 15,  2, 10,  7,  7,  2, 16,  7,  2, 12,
4,  5,  2,  4,  4,  2,  2,  5,  2,  4,  7),
valuesall=c(1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,
26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50),
countsall=c(32, 20, 33,  3, 15, 14, 25,  2, 35, 12,  7, 25, 13, 17,  7, 15, 30,
3, 10, 10, 15,  2,  7, 12, 18,  2,  4, 15,  2, 10,  7,  7,  2, 16,  7,  2, 12,
4,  5,  2,  4,  4, 0, 2,  2,  5, 0, 2,  4,  7))
```

A simple model could be either based on normal model for logit-transformed observed prevalences or beta-distribution directly. Write the two BUGS models (any others?) for these data.

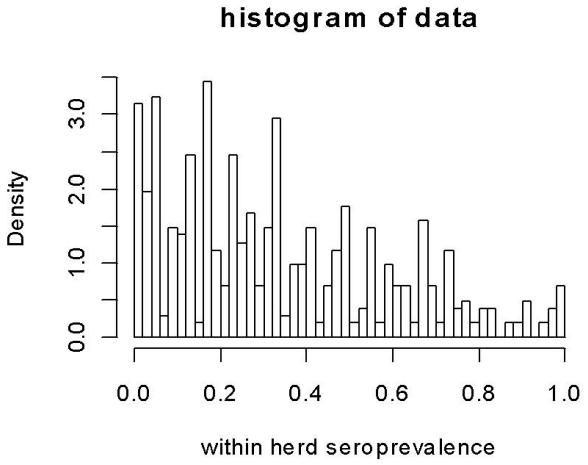

Figure 1: Density plot showing the probabilities of each bin. Visually approximated from original figure.

1.3 Estimating concentrations from censored data

Reference:
P Busschaert, AH Geeraerd, M Uyttendaele, JF Van Impe. Estimating distributions out of qualitatite and (semi)quantitative microbial contamination data for use in risk assessment. International Journal of Food Microbiology. 138 (2010), 260-269.

Number of samples	Concentration (CFU/g)
54	<0.04
2	<100
26	0.04-10
1	15
8	0.04 -100
2	>100
1	<1
1	>1
7	0.04 -1
1	1-100

Transform the CFU-values into $\log_{10}$-values. Assume a normal model for these. Write the BUGS-model for these data. Note: censoring is coded as "I()" in WinBUGS, but as "C()" in OpenBUGS. Study the sensitivity to e.g. different priors. Try to assess model fit.

1.4 Time series prevalence data

Reference:
J Ranta, D Matjushin, T Virtanen, M Kuusi, H Viljugrein, M Hofshagen, M Hakkinen. Bayesian temporal source attribution of foodborne zoonosis: Campylobacter in Finland and Norway. Risk Analysis.

To appear in 2011.

The following is an example of sample data from broiler production flocks monthly. From the samples, *Campylobacter* was analyzed (below only results for species C. *jejuni* shown). Hence the original data represent the number of positive samples x_i among all analyzed samples n_i, from the i^{th} month. To simplify, first we can take the estimated prevalence $\hat{p}_i = x_i/n_i$ as our observed value, and write a model for this, ignoring uncertainty due to sampling. These observed prevalences are usually low, sometimes zero. To transform from interval $[0, 1]$ to $\mathbb{R}$, we could use logit-transform. However, the zero-values cause a problem because $\text{logit}(0) = \log(0/(1-0)) = -\infty$. To avoid this with brute force, we could replace the zeros by a very small positive value 0.000000001.

```
list(N=36,phat=c(
0.000000001,0.000000001,0.000000001,0.000000001,0.000000001,0.026415094,
0.116279070,0.127208481,0.064638783,0.031872510,0.052631579,0.000000001,
0.000000001,0.000000001,0.000000001,0.000000001,0.000000001,0.029661017,
0.089147287,0.074349442,0.101214575,0.009836066,0.000000001,0.000000001,
0.000000001,0.000000001,0.000000001,0.000000001,0.000000001,0.018796992,
0.128919861,0.137931034,0.049122807,0.025157233,0.000000001,0.000000001,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA))
```

The autoregressive logit-model 'M1' could be: $\text{logit}(\hat{p}_i) \sim \text{N}(\mu_i, \tau)$ with $\mu_i = \alpha_0 + \alpha_1 \text{logit}(\hat{p}_{i-1})$. Try this model in BUGS and plot the results, also the predictions for `NA`s. Try some alternative model, e.g. Hidden Markov Model (HMM). In this model 'M2', the definition is $\text{logit}(\hat{p}_i) \sim \text{N}(\mu_i, \tau)$ with $\mu_i = \theta_i$ so that $\theta_i \sim \text{N}(\theta_{i-1}, \tau_\theta)$.

You can also use the sample data (bposje/bn) and specify a binomial model with logit-link. Similar data concerning turkey (tposje/tn) is also available. Make a model for both.

```
list(bn=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,258,269,268,271,250,
17,14,15,11,14,11,15,265,258,283,263,251,19,19,14,19,11,18,13,
236,258,269,247,305,25,23,15,17,11,13,9,266,287,290,285,318,16,17),
bposje=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,6,22,
25,12,9,0,0,0,0,0,0,0,6,30,35,17,8,1,0,0,0,0,0,0,6,20,18,21,2,0,0,
0,0,0,0,0,1,37,38,13,8,0,0),
tn=c(1,1,1,1,1,1,1,1,1,1,1,1,21,24,31,22,22,29,18,21,26,22,31,23,
20,23,24,28,26,28,25,28,23,21,25,26,27,22,34,24,27,18,16,20,23,24,
15,19,27,22,34,24,27,18,16,20,23,24,15,19),
tposje=c(NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,1,0,0,0,0,0,0,1,0,1,0,
0,1,2,1,0,0,1,2,5,2,1,1,5,2,1,0,2,0,0,1,3,1,1,0,0,1,0,0,0,1,2,5,1,
4,1,1,2))
```

1.5 Zero inflated model for microbial counts

Reference:
Aarnisalo K, Vihavainen E, Rantala L, Maijala R, Suihko M-L, Hielm S, Tuominen P, Ranta J, Raaska L. Use of results of microbiological analyses for risk-based control of Listeria monocytogenes in marinated broiler legs. International Journal of Food Microbiology, 2008, 121: 275-284.

The data set represents 165 observations of Listeria concentrations in broiler legs (CFU/g, transformed to counts per 10 gram). Most observations are zeros, and only a few are fairly high. Any standard distribution is unlikely to describe the data well. A zero inflated model might be reasonable.

```
list(N=165,
cfu10g=c(0,0,0,0,0,0,20,0,0,0,3,0,3,0,0,8,0,0,0,0,0,0,0,5,10,13,
3,0,3,28,14,5,211,5,0,0,3,18,0,0,5,8,0,10,100,0,0,0,0,0,0,0,0,0,
0,64,0,55,0,70,0,35,0,0,0,0,0,0,5,27,0,0,20,84,0,68,0,52,0,0,34,
0,0,5,0,0,148,0,1470,0,0,0,0,0,0,5,8,5,0,0,0,5,0,8,5,103,0,0,25,
0,35,0,0,0,10,8,0,0,0,13,0,0,13,0,0,18,0,0,0,5,18,8,0,27,0,9,0,0,
0,0,26,0,0,3,0,0,297,0,5,0,5,0,0,0,0,27,0,0,15,0,0,0,0,0,0),
company=c(1,9,1,1,1,1,1,9,9,9,9,9,9,9,1,8,8,1,8,9,1,1,8,9,8,8,8,
8,9,8,8,8,8,8,9,9,1,1,8,1,1,1,9,9,8,8,1,1,8,1,1,1,1,1,8,1,8,9,8,1,
8,9,8,1,1,9,9,1,1,8,8,9,9,8,8,1,8,9,8,1,8,8,9,9,8,9,9,8,9,8,9,9,
9,1,9,9,8,8,8,1,1,9,1,1,9,8,1,1,1,9,9,8,1,8,8,9,9,8,9,1,8,8,9,8,
9,1,1,9,1,1,9,8,9,8,9,1,8,9,8,1,9,8,1,1,1,1,1,1,1,9,1,1,8,9,8,8,
8,8,9,8,8,9,9,9,9,9))
```

1.6 Nonparametric Poisson intensity model

The change point estimation of coal mine accidents in Britain, 1851-1962, can be generalized to multiple change points of the Poisson intensity, and finally with enough change points, the model can be seen as a Poisson model with piece-wise constant intensity. This can be used as an approximately nonparametric model where the intensity function is not a simple parametric function. Implement the model in BUGS and study model fit and sensitivity analysis. Plot the posterior median intensity as a function of time. ('Compare' tool from 'Inference' menu). How could you generalize to predictions?

Decade	Number of accidents									
1851-1860	4	5	4	1	0	4	3	4	0	6
1861-1870	3	3	4	0	2	6	3	3	5	4
1871-1880	5	3	1	4	4	1	5	5	3	4
1881-1890	2	5	2	2	3	4	2	1	3	2
1890-1900	1	1	1	1	1	3	0	0	1	0
1901-1910	1	1	0	0	3	1	0	3	2	2
1911-1920	0	1	1	1	0	1	0	1	0	0
1921-1930	0	2	1	0	0	0	1	1	0	2
1931-1940	2	3	1	1	2	1	1	1	1	2
1941-1950	4	2	0	0	0	1	4	0	0	0
1951-1960	1	0	0	0	0	0	1	0	0	1
1961-1962	0	0								

1.7 Estimating within flock prevalence for egg laying flocks

Reference:

`http://www.fsis.usda.gov/Science/Risk_Assessments/index.asp`

Risk Assessment for Salmonella Enteritidis in Shell Eggs and Salmonella spp. in Egg Products (Oct 2005).

In this example, it is assumed that the flocks are either completely free of salmonella (within flock prevalence zero), or that there is some positive prevalence in them. Furthermore, a false negative rate of $h = 15\%$ is assumed for the testing. 58 pooled samples are analyzed from each flock. Each pooled sample consists of 5 individual sub-samples combined. The unknown fraction of completely clean flocks is θ. The within flock prevalence of the ith flock, if nonzero, is p_i.

Below, `x` is the number of observed positive pooled samples (out of 58), and `nf` is the number of flocks with the corresponding result.

```
# US data:  h=false negative rate (assumption)
list(K=32, h = 0.15, S=58, ss=5,
x=c(0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,
21,22,23,24,25,26,27,28,36,39,42,44),
nf=c(464,77,39,23,18,9,6,8,7,8,4, 6,4,4,2,2,6,1,3,3,
2,3,1,1,1,2,2,1,1,1,1,1))
```

Write hierarchical WinBUGS model for these data. Compute a predicted within flock prevalence from the hierarchical model. Study sensitivity to choice of priors. Compare the result with the Weibull-model suggested in the US report. Note the parametrization of the Weibull (b=0.43,c=0.0054) which has to be changed to a different parametrization for the Weibull in WinBUGS. What's logically slightly peculiar with the Weibull-model? Compute sufficiently many Monte Carlo draws to observe.

1.8 Other examples

You can suggest your own example and write a report from that. As long as it fits to max 10 pages and has the structure described below.

1.9 Structure of the report

Homework assignment

In the homework (10 points), a short applied problem is described, a bayesian model defined, a WinBUGS code implemented and results discussed. Suitable example problems and data are given during the course, but you can also suggest your own. As long as it is not too large problem! The written report should be less than 10 pages long (including figures and tables). If you use Word, could you please save and return the file readable in MSOffice2003 (I don't have 2007) or as a pdf-file. If you use LaTeX, please return it as pdf.

The report should include the main sections:

1. Introduction to the application problem: goal of analysis (what do we want to know in this example?), available data (what do we have for that?), assumptions & limitations,... (is that enough, under

what assumption? Is the final problem a limited version of the original larger problem?)

2. Specification of the probability model (or models) for the data. Mathematical expression and its written explanation. Rationale for choosing this model. Specification of the prior used for the unknown parameters. Explain why you use either informative or uninformative prior. Possible hierarchical structures explained.

3. Commented WinBUGS/OpenBUGS code for running the computations. This should be running free of syntax-bugs or runtime traps!

4. Check of MCMC convergence and possible remedies done if needed.

5. Presentation and discussion of the results, model assessment. Results can consist of posterior marginal or joint posterior distributions, or predictive distributions, according to the problem goals. This should answer the questions given in section 1. Possible reanalysis with a better model - if possible. Discussing the effects of model and/or prior choice.

6. References. E.g. the source of data, previous publications analyzing the same data, if a real publication is known. (No need to make reference to lecture memo of this course).