11 Application: screening and diagnostic problems

11.1 Latent infection

Assume some individual (or group) is tested for infection at regular time intervals. This test has some sensitivity p_{sen}, and specificity p_{spe}. The true infection (at each time point) is thus a latent variable that can be either 1 or 0. The infection process is driven by two parameters describing the probability of infection and probability of recovery

$$p = P(\text{inf}_i \mid \text{not inf}_{i-1}) \quad \text{infection}$$

$$q = P(\text{not inf}_i \mid \text{inf}_{i-1}) \quad \text{recovery}$$

Probability of latent infection at the first time point can be assumed as p. Assume $p, q, p_{\text{sen}}, p_{\text{spe}}$ have given values, so that only the true infection status is unknown. Assume then that we have a series of N time steps, and at some given point of time the test is done, resulting to a positive outcome. Tests at other time points are not done (or are missing), therefore coded as `NA`. Compute the posterior probability of the latent states, and the predicted testing outcomes at other time points. Check from the simulations that the probability of infection in the distant future approaches $p/(p+q)$.

```
 model{
 state[1] ~ dbern(pinfection)
 obs[1] ~ dbern(pobs[1])
 pobs[1] <- state[1]*psen+(1-state[1])*(1-pspe)

 for(i in 2:n){
 state[i] ~ dbern(pr[i])
 pr[i] <- state[i-1]*(1-precovery)+(1-state[i-1])*pinfection
 obs[i] ~ dbern(pobs[i])
 pobs[i] <- state[i]*psen+(1-state[i])*(1-pspe)
 }
 psen <- 0.7
 pspe <- 0.9
 }
list(n=40,pinfection=0.1,precovery=0.02,
obs=c(NA,NA,NA,NA,NA,NA,1,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA,NA,NA,NA,NA))
```

This was an example of a discrete time stochastic process for infection, with imperfect testing method. A related problem in a real salmonella risk assessment has been implemented as a continuous time Markov process: `http://mox.polimi.it/sco2009/` , see 'papers', browse by author 'R'.

11.2 Two tests for diagnosis

(BSM, ch. 3.5). A patient either has a certain disease (status $=D$), or does not have it (status$=\bar{D}$). A diagnostic test can give a positive result, $(+)$, or a negative result, $(-)$. As an unknown parameter, we have the population prevalence of that disease, p. Hence, we may assume a sampling probability

$P(D) = p$. The capability of the testing method is described with sensitivity $P(+ \mid D) = \eta$, and specificity $P(- \mid \bar{D}) = \theta$, so that

$$P(+, D) = P(+ \mid D)P(D) = \eta p \quad \text{and} \quad P(-, \bar{D}) = P(- \mid \bar{D})P(\bar{D}) = \theta(1-p).$$

The *predictive value* of the test is given by the posterior probabilities $P(D \mid +)$ and $P(\bar{D} \mid -)$. A useful test should have high predictive values. Based on the testing result, the patient is diagnosed either correctly or incorrectly, and the decision has some consequences. Often, several tests are used and the decision is then made based on all test results. If two tests are used, then the (conditional) probabilities of test results are

$$\begin{array}{ll} P(+,+ \mid D) = \eta_{11} & P(+,+ \mid \bar{D}) = \theta_{11} \\ P(+,- \mid D) = \eta_{10} & P(+,- \mid \bar{D}) = \theta_{10} \\ P(-,+ \mid D) = \eta_{01} & P(-,+ \mid \bar{D}) = \theta_{01} \\ P(-,- \mid D) = \eta_{00} & P(-,- \mid \bar{D}) = \theta_{00}. \end{array}$$

Note that $\eta_{11} + \eta 10 + \eta 01 + \eta_{00} = 1$ and $\theta_{11} + \theta 10 + \theta 01 + \theta_{00} = 1$.

There are four possible rules for making the diagnosis that the patient has D:

R1. Diagnose D, if test 1 is positive.
R2. Diagnose D, if test 2 is positive.
R3. Diagnose D, if test 1 and 2 are positive.
R4. Diagnose D, if test 1 or 2 is positive.

Moreover, we could think of the following strategy: the 2nd test is done only if the 1st test is negative. In that case, diagnose D if the 2nd test is positive.

Every test has a cost. Also, wrong diagnosis has a cost. The exact specification of all these costs can be difficult. What is the cost of life? The cost of a test can be relatively easy to quantify. But what is the cost of diagnosing someone to have a disease, when he/she does not really have it? How are the direct and indirect costs counted? Careful specification of costs can involve a lot of research. In common examples of decision theory, it is assumed that the costs are specified, so that we can focus on the probabilities and decisions, based on the pre-defined costs. For example BSM, p. 82 shows a cost function in the form of a table:

	D	$\bar{D}$
$+$	L_{11}	L_{10}
$-$	L_{01}	L_{00}.

From this table, we can calculate the expected total cost:

$$pS_i(L_{11} - L_{01}) + (1-p)T_i(L_{00} - L_{10}) + pL_{01} + (1-p)L_{10} + \text{AC},$$

where AC = "administrative costs" (including cost of the tests), i is the index of the decision rule ($i = 1, 2, 3, 4$), S_i and T_i are the sensitivity and specificity of the decision rule i:

Rule	S_i	T_i
$R1$	$P(+ \mid D) = \eta_{11} + \eta_{10}$	$P(- \mid \bar{D}) = \theta_{00} + \theta_{01}$
$R2$	$P(+ \mid D) = \eta_{11} + \eta_{01}$	$P(- \mid \bar{D}) = \theta_{00} + \theta_{10}$
$R3$	$P(+ \mid D) = \eta_{11}$	$P(- \mid \bar{D}) = \theta_{00} + \theta_{01} + \theta_{10}$
$R4$	$P(+ \mid D) = \eta_{11} + \eta_{10} + \eta_{01}$	$P(- \mid \bar{D}) = \theta_{00}$

Assuming that the losses L are given, the remaining unknown parameters in this problem are p, θ and η. For bayesian analysis of decisions, we need to compute the posterior probability of these. (Other parameterizations could be possible, but this was used in an example in BSM ch. 3.5).

11.3 Example: HIV testing from donated blood, two tests

BSM, p. 83, example 3.12. Unknown parameter: true HIV prevalence p in blood samples. As background data, there is an earlier study where 14 positive blood samples were found among 94496 samples (Canada, 1986). The performance of the two test is described by the following table:

	HIV	
	2nd test	
1st test	+	−
+	92	0
−	1	0

	no HIV	
	2nd test	
1st test	+	−
+	8	9
−	23	370.

The cost of a wrong negative result is that some patient will get HIV positive blood transfusion. This cost is assumed fairly high, $L_{01} = \$100000$. The cost of a wrong positive result is just that the donated blood is thrown away. This has a fairly low cost, $L_{10} = \$25$. In other situations the losses are assumed to be zero. Additionally, each test has a cost of \$1. Therefore, $A_1 = A_2 = 1$ and $A_{12} = 2$. As a prior of unknown parameters, we use

$$\eta_{11}, \eta_{10}, \eta_{01}, \eta_{00} \sim \text{Dir}(3.9, 0.5, 0.5, 0.1)$$
$$\theta_{11}, \theta_{10}, \theta_{01}, \theta_{00} \sim \text{Dir}(0.1, 0.5, 0.5, 3.9).$$

The prior distribution reflects the assumption that both tests work similarly so that they both are more likely to give a right result than wrong. (Results, BSM p. 84). Decision $R2$ becomes chosen since it minimizes the expected costs.

This example may be oversimplifying. Real decision problems can be much more complicated because many different costs and benefits can be involved. It may be difficult to describe them all, and the cost functions may quickly change over time.

Stangl DK: Bridging the gap between statistical analysis and decision making in public health research. Statistics in Mecidine. 2005. 24: 503-511:

"...Methods for providing such summaries [statistical estimates] are highly formalized and constantly evolving. While decision making is the incentive for nearly all such efforts, the process that transforms statistical summaries into decisions usually remains informal and ad hoc. Statisticians have not eagerly accepted the role of promoting formalized decision-theoretic techniques... ...the gap between statistical synthesis and decision making is an unnatural and undesirable one, because it undermines the impact of quantitative information."

11.4 Example: diagnosing strongyloides infection

Available data: test results of 2 different tests from 162 individuals. Test positives: 40/162 and 125/162.

Goal: to estimate test sensitivity η, specificity θ and population prevalence p based on either of the test results, or both results.

Background knowledge: expert knowledge about η and θ. No prior knowledge about p.

Assume the prior knowledge can be formalized as:

$$\eta \sim \text{Beta}(s,t)$$
$$\theta \sim \text{Beta}(c,d).$$

For prevalence p, we choose the prior U(0,1), i.e. Beta(1,1).

Data of testing results:

	1st test	
2nd test	+	−
+	38	87
−	2	35.

Number of positive and negative results from each test: $(A_1, A_2) = (40, 125)$ (posit.), $(B_1, B_2) = (122, 37)$ (negat.).

• Consider using only the results from a single test, for example $A = 40$ (posit.), $B = 122$ (negat.), in a total of $N = A + B = 162$ tested. Define a *latent variable* T_1 and T_2:

$$T_1 = \text{unknown number of truly positives among } A \text{ test positives}$$
$$T_2 = \text{unknown number of false negatives among } B \text{ test negatives}$$

Note: $T_1 + T_2$ is the number of all truly positives among all tested $N = A + B$. Using these latent variables, we can write the results of the single test as a 2×2 table, in which only the row sums are uniquely identified from data:

	True		
Test	+	−	
+	T_1	$A - T_1$	A
−	T_2	$B - T_2$	B.

For this table, we can write the cell probabilities according to the model parameters p, η, θ:

	True		
Test	+	−	
+	$p\eta$	$(1-p)(1-\theta)$	
−	$p(1-\eta)$	$(1-p)\theta$	.

Note: bayesian model is the joint distribution $\pi(\text{parameters}, \text{data})$, from which the posterior is obtained as $\pi(\text{parameters} \mid \text{data})$. In this example, the joint distribution of all unknown parameters, latent variables, and data can be written using the cell probabilities as:

$$\pi(p, \eta, \theta, T_1, T_2, A, B) = \pi(T_1, T_2, A, B \mid p, \eta, \theta)\pi(p, \eta, \theta)$$

$$\propto (p\eta)^{T_1}[(1-p)(1-\theta)]^{A-T_1}[p(1-\eta)]^{T_2}[(1-p)\theta]^{B-T_2}\underbrace{\pi(p)}_{=1}\eta^{s-1}(1-\eta)^{t-1}\theta^{c-1}(1-\theta)^{d-1}$$

By re-arranging the terms, we can write it as:

$$\underbrace{p^{(T_1+T_2+1)-1}(1-p)^{(A-T_1+B-T_2+1)-1}}_{\text{Beta}(T_1+T_2+1,\,A+B-T_1-T_2+1)} \times \underbrace{\eta^{T_1+s-1}(1-\eta)^{T_2+t-1}}_{\text{Beta}(T_1+s,\,T_2+t)} \times \underbrace{\theta^{B-T_2+c-1}(1-\theta)^{A-T_1+d-1}}_{\text{Beta}(B-T_2+c,\,A-T_1+d)},$$

It is now easy to recognize beta-distributions (up to normalizing constant) for parameters p, η, θ. By re-arranging the terms slightly differently, the same joint probability can be written also in the form:

$$\Big[\frac{p\eta}{p\eta+(1-p)(1-\theta)}\Big]^{T_1}\Big[\frac{(1-p)(1-\theta)}{p\eta+(1-p)(1-\theta)}\Big]^{A-T_1} \times \underbrace{[p\eta+(1-p)(1-\theta)]^{A}}_{\text{constant with respect to } T_1 \text{ and } T_2}$$

$$\times\Big[\frac{p(1-\eta)}{p(1-\eta)+(1-p)\theta}\Big]^{T_2}\Big[\frac{(1-p)\theta}{p(1-\eta)+(1-p)\theta}\Big]^{B-T_2} \times \underbrace{[p(1-\eta)+(1-p)\theta]^{B}}_{\text{constant with respect to } T_1 \text{ and } T_2}$$

from which the binomial distributions can be recognized for T_1 and T_2.

Lesson: since the joint distribution can be written in a form that shows what the conditional distributions $\pi(p, \eta, \theta \mid T_1, T_2, A, B)$ and $\pi(T_1, T_2 \mid p, \eta, \theta, A, B)$ are, this provides us the necessary full conditional distributions for Gibbs sampling. This has been implemented as WinBUGS code in BSM Example 3.13. This example is slightly peculiar in a sense that it uses the full conditional distributions directly in WinBUGS. Therefore, the user actually implements the Gibbs algorithm, instead of implementing the *model definition* as a DAG in WinBUGS. (Compare with the example of sampling from 2D normal density).

The model definition as a DAG would correspond to:

```
x ~ dbin(pr,n)
pr <- p*eta+(1-p)*(1-theta)
p ~ dbeta(1,1)
eta ~ dbeta(s,t)
theta ~ dbeta(c,d)
```

from which the model is much easier to understand than from the corresponding Gibbs sampler with additional latent variables. The implementation by writing the full conditionals merely demonstrates how this model could be simulated, whereas the DAG representation demonstrates what the model is.