UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4021/9021 – Applied Bayesian Analysis and Numerical Methods**
	Part II of two parts
WITH:	**Nils Lid Hjort**
AUXILIA:	**Calculator, plus one single sheet of paper with the candidate's own personal notes**
TIME FOR EXAM:	**Part I: home project, 30/xi1–12/xii/2017;**
	Part II: Wednesday 13/xii s.y., 9^{00}–13^{00}, written exam

This exam set contains four exercises and comprises four pages.

Note: Please write your **StudentWeb number** on the top of the first page of what you hand in today. In the marking process we need to connect your solutions of today with your project report.

Exercise 1

WE START WITH SOME BAYESIAN QUESTIONS related to the binomial distribution, which is fitting in that this is also where the Presbyterian minister Thomas Bayes (1702–1761) arguably started what is now known as Bayesian statistics, with a paper published two years after his death. Below you are free to utilise the formula

$$\int_0^1 x^c(1-x)^d \,\mathrm{d}x = \frac{c!\,d!}{(c+d+1)!},$$

valid for nonnegative integers c, d. Also, the Beta distribution density, with parameters (a, b), is

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad \text{for } x \in (0,1).$$

Its mean and variance can be expressed as

$$\mathrm{E}\, X = x_0 = \frac{a}{a+b}, \quad \mathrm{Var}\, X = \frac{x_0(1-x_0)}{a+b+1}.$$

(a) Assume that $y_1, \ldots, y_n$ is a sequence of 0–1 variables that for given probability θ are independent with the same probability $\Pr(y_i = 1 \,|\, \theta) = \theta$. Write down the conditional mean and variance of y_i as well as for $z = \sum_{i=1}^n y_i$, the observed number of events in n trials.

(b) The event probability θ is most typically an unknown quantity, however, and we shall now assume that θ has the Beta$(2, 2)$ prior distribution.

(i) Find the prior mean and prior variance of θ.

(ii) Find the mean and variance of y_i in its marginal distribution.

(iii) Find the covariance and correlation between y_i and y_j, where $i \neq j$.

(iv) Find the marginal mean and marginal variance of z.

(c) Find an explicit formula for the marginal distribution of z.

(d) Find the posterior distribution for θ, and show that

$$\widehat{\theta}_B = \mathrm{E}(\theta \,|\, z) = \frac{z+2}{n+4}.$$

(e) With the usual quadratic loss function for estimating θ, $L(\theta, \widetilde{\theta}) = (\widetilde{\theta} - \theta)^2$, find first the risk function for the standard estimator $\widetilde{\theta} = z/n$.

(f) Then find the risk function for the Bayes estimator found in (d) above. In what part of the parameter interval is the Bayes estimator better than the standard estimator?

Exercise 2

Blood on the Tracks: The blood classification system invented by another Nobel Prize winner, Karl Landsteiner, relates to certain blood substances of antigens 'a' and 'b'. There are four blood groups A, B, AB, O, where A corresponds to 'a' being present, B corresponds to 'b' being present, AB to both 'a' and 'b' being present, while O (for 'ohne', natürlich) is the case of neither 'a' nor 'b' present.

In this exercise we shall consider a historically important data set from 1924, used by Landsteiner, where Felix Bernstein had examined 502 Japanese living in Korea, and counted

$$N_A = 212, \quad N_B = 103, \quad N_{AB} = 39, \quad N_O = 148,$$

for the four categories. Assuming a multinomial model, the likelihood is proportional to

$$\theta_A^{212}\, \theta_B^{103}\, \theta_{AB}^{39}\, \theta_O^{148},$$

with $(\theta_A, \theta_B, \theta_{AB}, \theta_O)$ representing the four probabilities in question. Here we do not have the time to go into detailed examination of different theories for blood groups in man, but we consider *one of these*, the so-called two-loci theory, which for the present purposes can be seen to imply that

$$\theta_A = p(1-q), \quad \theta_B = (1-p)q, \quad \theta_{AB} = pq, \quad \theta_O = (1-p)(1-q),$$

for certain probabilities p and q.

(a) Write out a reasonably simplified expression for the likelihood function $L(p, q)$ under this theory.

(b) With a simple Bayesian analysis, take p and q independent and uniform, and derive the posterior distribution for the pair (p, q).

(c) Compute the posterior mean for $n\theta_{AB} = npq$. Compare with the observed number in the AB group. Briefly discuss what this means, and indicate further analyses that you could undertake (but after 13:00 today).

Exercise 3

SUPPOSE A CERTAIN COMPLICATED DATA SET has been carefully analysed, with model and a realistic prior for the model parameters, etc. Assume further that a nonnegative parameter θ of primary interest (a so-called focus parameter) after all this work is seen to have a posterior distribution with cumulative

$$P(\theta) = 1 - e^{-\theta}(1+\theta) \quad \text{for } \theta \geq 0.$$

(a) Find the posterior density function, say $p(\theta)$, and give a rough plot to indicate how it looks like. Compute the posterior mean. (Here you may use the fact that $\int_0^\infty x^m \exp(-x)\,\mathrm{d}x = m!$, for nonnegative numbers m.)

(b) A certain decision needs to be made, from the list of actions A, or B, or C. The loss function is

$$L(\theta, A) = \begin{cases} 0 & \text{if } \theta \leq 1.1, \\ 2 & \text{if } \theta > 1.1, \end{cases}$$

$$L(\theta, B) = \begin{cases} 0 & \text{if } \theta \in (1.1, 3.3), \\ 3 & \text{if } \theta \notin (1.1, 3.3), \end{cases}$$

$$L(\theta, C) = \begin{cases} 0 & \text{if } \theta > 3.3, \\ 4 & \text{if } \theta \leq 3.3. \end{cases}$$

Which action should be taken?

(c) One also wishes to produce a credibility interval for θ, say $[a, b]$. Rather than going for the traditional type of interval, with a fixed pre-determined credibility level like 95%, one decides to judge different candidate credibility intervals $[a, b]$ according to the loss function

$$L(\theta, [a,b]) = 0.10\,(b-a) + I\{\theta \notin [a,b]\} = \begin{cases} 0.10\,(b-a) & \text{if } \theta \in [a,b], \\ 0.10\,(b-a) + 1 & \text{if } \theta \notin [a,b]. \end{cases}$$

Write down an expression for the risk function associated with this loss function, say for an interval $[\widehat{a}(y), \widehat{b}(y)]$ constructed from data.

(d) What is the Bayes solution, i.e. the best interval $[a, b]$? Since this is a computer-free exam environment, you may give precise equations for determining a and b (these equations will then be easy to solve once you're having a couple of minutes with R).

Exercise 4

CONSIDER THE DISTRIBUTION for independent nonnegative observations $y_1, \ldots, y_n$ with density

$$f(y \,|\, \theta) = 3\theta y^2\, e^{-\theta y^3} \quad \text{for } y > 0,$$

with θ an unknown positive parameter.

(a) Write down the likelihood function for the observed data, and find a formula for the maximum likelihood estimator, say $\widehat{\theta}$.

(b) Assume θ is given a Gamma prior, with parameters (a, b), i.e. with prior density

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \quad \text{for } \theta > 0.$$

Find the posterior distribution for θ.

(c) Use theory from the course to give the approximate normal distribution of $\widehat{\theta}$. Also give a normal distribution approximation to the posterior distribution of θ.