UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4020 – Bayesian statistics**
	Part I of two parts
With:	**Nils Lid Hjort**
Time for exam:	**Part I: 1–13/xii/2010;**
	Part II: 16/xii s.y., written exam

This is the exam set for STK 4020, autumn semester 2010. It is made available on the course website as of *Wednesday 1 December 12:00*, and candidates must submit their written reports by *Monday 13 December 14:00* (or earlier), to the reception office at the Department of Mathematics, in duplicate. The supplementary written examinations take place Thursday December 16, 9:00–13:00 (practical details are provided elsewhere). Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should preferably be text-processed (TeX, LaTeX, Word), but may also be hand-processed. Give your name on the first page. Write concisely (in der Beschränkung zeigt sich erst der Meister; brevity is the soul of wit; краткость – сестра таланта). Relevant figures need to be included in the report. Copies of machine programmes used (in `R`, or `matlab`, or similar) are also to be included, perhaps as an Appendix to the report. Candidates are required to work on their own (i.e. without cooperation with any others), but are graciously allowed not to despair if they do not manage to answer all questions well.

Importantly, each student needs to submit *two special extra pages* with her or his report. *The first* (page A) is the 'erklæring' (self-declaration form), properly signed; it is available at the webpage as 'Exam Project, page A, declaration form'. *The second* (page B) is the student's one-page summary of the exam project report, which should also contain a brief self-assessment of its quality.

This exam set contains three exercises and comprises four pages.

Exercise 1

I'm trying out a new game of solitaire ('kabal'), the so-called 'Myshkin patience'. I wish to assess its associated probability θ of me 'winning' the solitaire game ('kabalen går opp'). My Experiment A consisted in playing the game repeatedly until I successfully 'won', and it turned out that I needed to play the game eleven times to achieve this (ten no-wins followed by a win). I then went through a second and similar Experiment B where it turned out that I this time needed seventeen games to win the first time (sixteen no-wins followed by a win).

(a) Formulate the above in Bayesian terms, starting with a prior for θ that is uniform on the unit interval, and briefly discuss any assumptions you need to make to get the formalism to work. Display two curves in a diagram, the posterior for θ after Experiment A and the posterior for θ after Experiments A and B. Also provide the 0.025, 0.50, 0.975 posterior quantiles, after Experiment A, and after Experiments A and B.

(b) Redo the analysis above, now starting with the Jeffreys prior for θ.

Exercise 2

A NUMBER OF FACTORS are involved in the processes that at the end determine the weight of a newborn child, and various studies are addressing questions related to factors that influence the chances of a child being either too small or too big. An important question in this regard is simply to know the full statistical distribution of birth weights. The focus of this exercise is Bayesian analysis of the statistical parameter

$$\kappa = F^{-1}(0.10),$$

the 0.10 quantile of the birth weight distribution in the normal population.

The data file `babies` (available at the course website) contains the birth weights, in gram, of babies associated with a certain North American study from the 1980ies (and we now identify F above as the distribution of birth weights associated with the greater population of mothers living in the state in which the study was conducted, in that time period). Access the file and convert in your computer the weights into the kilogram scale. We shall assume that the mothers recruited to this study represent a random sample from the population in question.

(a) We start out assuming that the birth weight distribution can be considered normal, i.e. that $y_1, \ldots, y_n$ can be seen as independent observations from the normal distribution $\mathrm{N}(\mu, \sigma^2)$. To carry out Bayesian analysis of κ we need a prior density for (μ, σ), which we construct as follows.

(i) Take $\lambda = 1/\sigma^2$ to be a Gamma (a_0, b_0), with parameters selected so that the 0.10 and 0.90 prior quantiles of σ are respectively 0.50 and 2.00. Find the Gamma parameters numerically (I find $(1.173, 0.649)$).

(ii) Take $\mu \,|\, \sigma$ to be a $\mathrm{N}(\mu_0, \sigma^2/\nu_0)$, with $\mu_0 = 2.500$ and $\nu_0 = 3.333$.

Use prior simulations or direct calculations to find the prior mean and the 0.99 prior quantile for κ.

(b) Give an adequate description of the posterior distribution of (μ, σ) – where you are allowed to use results obtained in the curriculum and in earlier exercises (i.e. it is not required to redo the algebra). Provide the posterior $0.05, 0.50, 0.95$ quantiles for κ under the present normality assumption.

(c) The normal model for the birth weight data is not necessarily adequate, however. There are various three-parameter models that in different ways extend the two-parameter normal, and here we concentrate on the one with cumulative distribution function

$$G(y,\mu,\sigma,a)=\Phi\Big(\frac{y-\mu}{\sigma}\Big)^{\exp(a)},$$

where Φ as usual denotes the standard normal cumulative distribution function. Explain why this provides a bona fide probability distribution; give a formula for $G^{-1}(q,\mu,\sigma,a)$, the q quantile of the distribution; and briefly explain the role of the a parameter.

(d) To pursue a Bayesian analysis here one needs a prior for (μ,σ,a), and for the present purposes we shall use the one that takes (μ,σ) and a independent, with (μ,σ) having the same prior as given above and a a uniform on $[-c,c]$ with $c=4.00$. Briefly discuss why this may or may not be a sensible idea, and explain how you perhaps could construct alternative priors.

(e) Fit the three-parameter model to the data using maximum likelihood (e.g. via numerical optimisation and the `nlm` algorithm in `R`). Display a histogram of the data along with two fitted densities, the normal one and that based on the three-parameter model. Comment on what you see from such a diagram.

(f) Carry out Bayesian analysis using the three-parameter model with the prior given above, providing in particular the posterior $0.05, 0.50, 0.95$ quantiles for κ. Use both the 'lazy Bayes' strategy, via normal approximations, and the more careful analysis that utilises simulations from the exact posterior distribution.

(g) A Bayesian way in which to select among the two models used above is to give each of them prior probability $\frac12$ and then compute the posterior model probabilities. Do this (with numerical approximations, if necessary), and comment on your findings.

Exercise 3

This exercise concerns estimation of a normal mean parameter, for a given data set, but with somewhat different prior distributions. We model the data points $y_1,\ldots,y_n$ as being independent from the same $\mathrm{N}(\theta,1)$ distribution, for given θ. For the following illustrations we take $n=10$ and mean value $\bar y=0.444$.

(a) Show that the likelihood function is proportional to $\exp\{-\frac12 n(\theta-\bar y)^2\}$. Display the maximum likelihood estimate and an ordinary 95% confidence interval for θ.

(b) For estimating θ here, consider the loss function

$$L_\varepsilon(\theta,\widetilde\theta)=\begin{cases}0 & \text{if } |\widetilde\theta-\theta|\le\varepsilon,\\ 1 & \text{if } |\widetilde\theta-\theta|>\varepsilon,\end{cases}$$

with ε being a small positive number. Characterise the Bayes solution when this loss function is employed, along with its limit as ε shrinks to zero.

(c) Consider the double exponential prior for θ, say DE(λ), with density $\frac{1}{2}\lambda\exp(-\lambda|\theta|)$. Here λ is a positive parameter characterising the density (a so-called hyperparameter) and the range of θ is $(-\infty,\infty)$. For each of the choices $\lambda = 2.5, 4.5, 6.5$, compute the posterior distribution of θ, and display the three posterior densities in a diagram. You may use numerical integration, if fruitful, and you are also free to use the formula

$$\begin{aligned} g(a,c) &= \int \phi(x)\exp(-c|x+a|)\,\mathrm{d}x \\ &= \{1-\Phi(a+c)\}\exp(ac+\tfrac{1}{2}c^2) + \Phi(a-c)\exp(-ac+\tfrac{1}{2}c^2). \end{aligned}$$

(d) For the three DE(λ) priors used above, find the posterior mode, and comment on the general phenomenon at work here. Furthermore, for each of these three priors, display the 0.025, 0.50, 0.975 posterior quantiles, and compare your results to the confidence interval found in (a). Find these posterior quantiles via posterior simulation, e.g. using an acceptance-rejection strategy.

(e) Use the data to infer a sensible value for λ here. There is perhaps no uniquely superior way of doing this, so 'sensible' is interpreted broadly. Explain your reasoning.

(f) Rather than relying on an empirical Bayes type approach, which as above requires a reasonable value of the hyperparameter λ to be found from data, a more full-fledged Bayesian solution is to employ also a background prior $\pi(\lambda)$ for this. Attempt to follow this idea through, starting with a uniform prior on $[0,10]$ for λ. If you succeed, display simulated (λ_j,θ_j) pairs from the joint posterior distribution, and compute, in the end, the 0.025, 0.50, 0.975 posterior quantiles for θ in this wider framework.