UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4050:**
	Statistiske simuleringer og numeriske beregninger
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**7 Dec 2007 at 11:55 – 17 Dec 2007 at 14:00**

This is the exam set for STK 4050, autumn 2007. It is made available on the course website as of *Friday 7 December 12:00*, and candidates must submit their written reports by *Monday 17 December 14:00*, to the reception office at the Department of Mathematics.

Reports may be written in Norwegian, English or German, 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. 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.

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 'Erklæring (exam, page A)'. *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 five exercises and comprises five pages.

Exercise 1

NUMERICAL INTEGRATION IS ESSENTIALLY EASY in the one-dimensional case but more troublesome in higher dimensions. Consider the integral

$$A = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \exp\big\{\cos(|xyzuv|)\sin(|x+y+z+u+v|)\big\}\,\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}z\,\mathrm{d}u\,\mathrm{d}v.$$

Use stochastic simulations to estimate A, and give a 99% confidence interval.

Exercise 2

ALMOST EXACTLY HUNDRED YEARS AGO Student (alias W. Gosset) introduced the famous t-test, in the landmark paper *The probable error of a mean* (*Biometrika*, 1980, pp. 1–25). In modern language, assume that $X_1, \ldots, X_n$ are independent and normal (μ, σ^2), and suppose we need to test the hypothesis $H_0\colon \mu = 0$ versus the alternative $\mu \neq 0$. Then the t statistic is

$$t_n = \frac{\sqrt{n}\bar{X}}{\widehat{\sigma}}, \quad \text{with } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ and } \widehat{\sigma} = \Big\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\Big\}^{1/2}.$$

Student's contribution was partly to find the correct null distribution, i.e. the distribution of t_n under H_0. As we know, it is a t_{n-1}, and the associated test with level say 0.05 is to reject when $|t_n| > t_{n-1,0.975}$, the upper 0.025 point in this distribution.

The point of the present exercise is to consider a robust competitor to Student's t. We shall use

$$Z_n = \frac{\sqrt{n}M_n}{\widetilde{\sigma}},$$

where M_n is the median of the n data points and

$$\widetilde{\sigma} = c\Big(\frac{1}{n}\sum_{i=1}^{n}|X_i - M_n|^{3/2}\Big)^{2/3}.$$

Here the constant c is chosen so that $\widetilde{\sigma}$ is consistent for the real σ in case the data really follow the normal distribution. The null hypothesis $\mu = 0$ is rejected when $|Z_n| > z_{0,n}$, with $z_{n,0}$ to be specified.

(a) Based on simulations for respectively $n = 100$ and $n = 1000$, find a numerical approximation for the c constant. Attempt also to find a formula for c mathematically. – The numerical value of c, using this formula, happens to be 1.10574. For the following points, use this value for c. (Since Z_n is used primarily as a test statistic, the value of c is of no serious consequence, but the 1.10574 figure secures a clear interpretation of the denominator.)

(b) For $n = 25$ and for significance level 0.05, establish the rejection limit for the $|Z_n|$ test (via simulations).

(c) Under normality assumptions, show that the power function

$$\pi_n(\mu, \sigma) = \Pr\{|Z_n| > z_{0,n} \,|\, \mu, \sigma\}$$

is a function of μ/σ alone. For $n = 25$, compute the power function (via simulations), and display it in a diagram, along with the power function of the t-test. Comment briefly on your results.

(d) One wishes to collect enough data in order for the $|Z_n|$ test to have probability at least 0.90 of detecting that H_0 is wrong, if the true state of affairs is $|\mu|/\sigma = 0.50$. How many data points are needed to achieve this?

Exercise 3

"Independence? That's middle class blasphemy. We are all dependent on one another, every soul of us on earth." Indeed too many statistical situations are modelled via independence, even if the context ought to or might indicate statistical dependence. This exercise is a little excursion into one particular way of modelling dependence between given distributions.

Suppose $f(x)$ and $g(y)$ are probability densities on $(0,\infty)$, and define

$$h(x,y) = f(x)g(y)\exp\{-\theta|x-y|\}/A(\theta) \quad \text{for } x>0,\ y>0,$$

with $A(\theta)$ the appropriate normalising constant. The task is to construct ways of simulating pairs (X_i, Y_i) from this distribution, for given start densities f and g and for given values of θ. For concreteness of illustration we shall take f and g to be unit exponential below, i.e. $f(x) = \exp(-x)$ for $x > 0$ and $g(y) = \exp(-y)$ for $y > 0$, but generalisations are not difficult to work through.

(a) What is the parameter region for the association parameter θ? Comment briefly on how realisations of h can be expected to behave, for cases $\theta = 0$, θ small, θ big.

(b) Before returning to the specific case at hand, prove the following version of the rejection-acceptance scheme: Suppose $p(z) = p_0(z)/a$ is the target density, and that $p_0(z) \le Kq(z)$ for all z, where q is another density that perhaps is easier to sample from. First generate $Z_1, Z_2, \ldots$ from q. For each Z_i, keep it, with probability $p_0(Z_i)/\{Kq(Z_i)\}$, and otherwise throw it away. Then the surviving Z_i (those that are accepted) follow the target density p.

(c) Use rejection-acceptance sampling to generate 10,000 random pairs (X_i, Y_i) from the $h(x,y)$ density above, for $\theta = 1.3579$. Compute estimates of the means and standard deviations for X and Y, as well as their inter-correlation.

(d) For each of a string of nonnegative association parameter values θ, simulate random pairs (X_i, Y_i) as above, and display the (estimated) curve $\xi(\theta)$, where $\xi(\theta) = \mathrm{E}_\theta|X-Y|$. What is the exact value of $\xi(0)$?

(e) I have observed $n = 111$ pairs (X_i, Y_i) from a real data set, for which

$$\frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i| = 0.444.$$

Estimate the association parameter θ. Attempt also to estimate the standard deviation of your parameter estimate.

Exercise 4

PRIOR INFORMATION ABOUT PARAMETERS ought to be used in combination with data information to reach correct inference statements. This is precisely the Bayesian viewpoint. In some situations the Bayes method amounts to 'pushing back data in the direction of what is likely under the prior', as will be illustrated below.

We shall assume that data $X_1, \ldots, X_5$ are independent measurements of unknown parameters $\theta_1, \ldots, \theta_5$, each corresponding to standard normal error terms, i.e.

$$X_1 \sim \mathrm{N}(\theta_1, 1), \ldots, X_5 \sim \mathrm{N}(\theta_5, 1).$$

The observed data in this illustration are

$$(x_1, x_2, x_3, x_4, x_5) = (1.11, 2.22, 3.33, 4.44, 13.13).$$

Without prior information, these are automatically also the canonical point estimates of the five θ_i parameters. – We shall however assume here that there is prior information to the effect that the five θ_i parameters tend to be somewhat close to each other, and that this is modelled via the prior density

$$\pi(\theta_1,\ldots,\theta_5) = \exp\{-\lambda\, V(\theta_1,\ldots,\theta_5)\}/a(\lambda), \quad \text{where } V(\theta) = \sum_{i=1}^{5} |\theta_i - \bar\theta|,$$

and with $a(\lambda)$ the required normalisation constant; also, as usual, $\bar\theta$ denotes the average, $(1/5)\sum_{i=1}^5 \theta_i$. Here λ is a prior information parameter that dictates the degree to which the five θ_i are close to each other.

(a) Show that the posterior density for the five parameters takes the form

$$\pi(\theta_1,\ldots,\theta_5 \,|\, \text{data}) \propto \exp\Big[-\sum_{j=1}^{5}\{\lambda|\theta_i-\bar\theta| + \tfrac{1}{2}(\theta_i - x_i)^2\}\Big],$$

where '$\propto$' means 'proportional to'.

(b) Implement a Metropolis scheme for simulating vectors $(\theta_1,\ldots,\theta_5)$ from the posterior distribution, where you use $\lambda = 5.55$ for this illustration. You may use proposals of the form $\theta_i^{\text{new}} = \theta_i^{\text{old}} + \varepsilon_i$, where the ε_i are independent $\text{N}(0,\delta^2)$, for a suitable choice of δ. Display random pairs (θ_1,θ_5) (drawn from the posterior distribution) in a point cloud diagram. Explain briefly how you have decided on the δ parameter and the burn-in period.

(c) Compute the posterior mean and the posterior standard deviation for each of the five parameters, and give 90% credibility intervals (i.e. intervals that contain 90% of the posterior probability). Comment briefly on your results.

(d) In addition to the posterior means, which you found in point (c), Bayesians sometimes wish to find the 'MAP solution' (maximum a posteriori point), i.e. the max-point $(\theta_1^*,\ldots,\theta_5^*)$ in the posterior density. Attempt to find the MAP solution here, via simulations. *A reasonable approximation is sufficient here, full accuracy is not required.*

Exercise 5

How many sick days in three months? The answer to that question depends of course on the the illness and on the individual in question. We shall assume here that there is an underlying process $Z_1,\ldots,Z_{90}$ that determines a person's sickness or not; when Z_i exceeds a threshold value c, then the person is ill on day i. In this illustration, suppose that the Z_i follow a zero-mean Gaußian process with covariance function

$$k(i,j) = \text{cov}(Z_i,Z_j) = \exp\{-a|j-i|\} = \rho^{|j-i|}, \quad \text{where } \rho = e^{-a}.$$

In particular, each single Z_i is a standard normal.

(a) For $a = 0.20$, simulate and display three paths of $Z_1, \ldots, Z_{90}$. You may use the $Z = \Sigma^{1/2} \mathrm{N}_n(0, I_n)$ trick, involving the square root of the covariance matrix for the Z vector, via these lines:

```
# my "squareroot" function for covariance matrices:
squareroot <- function(K)
{rootL <- 0*K
diag(rootL) <- sqrt(eigen(K, symmetric = T)$values)
P <- eigen(K, symmetric = T)$vectors
P %*% rootL %*% t(P)}
```

(b) With sickness threshold level $c = 1.75$, let J be the number of days an individual is ill inside a ninety-day period, still using time dependence parameter $a = 0.20$. Find and display the distribution of J by simulation. What is the probability that the person is never ill, in the course of ninety days?