UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4150:**
	Miljøstatistikk / Environmental statistics
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**2 June 2008 at 11:55 – 16 June 2008 at 14:00**

This is the exam set for STK 4150, spring semester 2008. It is made available on the course website as of *Monday 2 June 12:00*, and candidates must submit their written reports by *Monday 16 June 14:00* (or earlier, for those travelling to Vilnius), to the reception office at the Department of Mathematics. The supplementary oral examinations take place June 24 and 25 (practical details for these will be provided later).

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; 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. The full exam set *is* (admittedly) laborious, and candidates are 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 five exercises and comprises nine pages.

Exercise 1

EVERYBODY TALKS ABOUT THE WEATHER, but nobody *does* anything with it. The dataset `examdata101` is organised as (year, x_t, z_t), with minimum temperatures in two different geographical regions of Nevada, for each winter, from 1890 to 2000. (The temperatures given in the file are not absolute minima, per winter, but rather certain averages across cold days and across several neighbouring measurement positions. I have also converted Fahrenheit measurements to Celcius.)

(a) Read the dataset suitably into your computer, using

```
nevada = matrix(scan("examdata101", skip=6), byrow=T, ncol=3)
year = nevada[ ,1]
xt = nevada[ ,2]
zt = nevada[ ,3]
```

or something equivalent. Display a plot of both time series in the same diagram.

(b) Focus first on the first of these time series, say x_t for $t = 1, \ldots, n = 111$. The autoregressive model of order m, with a linear trend term for time, corresponds to

$$x_t = \beta_0 + \beta_1(t - t_0) + \eta_t \quad \text{for } t = 1, \ldots, n,$$

where t_0 is the average t value, and where

$$\eta_t = \phi_1 \eta_{t-1} + \cdots + \phi_m \eta_{y-m} + \varepsilon_t, \tag{1.1}$$

with $\varepsilon_1, \ldots, \varepsilon_n$ being i.i.d. $\mathrm{N}(0, \sigma^2)$ and $\phi_1, \ldots, \phi_m$ autoregressive parameters. We make (1.1) valid also for small t by setting $\eta_t = 0$ for $t = 0, -1, \ldots, -(m-1)$. – Fit these AR($m$) models to the x_t data, using maximum likelihood methodology, for $m = 0, 1, 2, 3$, where $m = 0$ corresponds to independence. Which model would you select among these, and why?

(c) Use the model you have selected in (b) to make a prediction of the (average) minimum winter temperature in this Nevada area, for the year 2010. Supplement your point prediction with a suitable prediction interval, and explain your arguments and your assumptions.

(d) Extend the model you have selected in (b) to include also the z_t series as an extra covariate. Would you say that the z_t process influences the x_t process?

Exercise 2

One man's interpolator is another man's extrapolator. A rather simple data set is given as follows, associated with a certain random field $Z = \{Z(x) : 0 \le x \le 10\}$, observed in just $n = 14$ point locations x (with x on the top line and $Z(x)$ on the second):

0.00	0.50	1.00	1.5	2.00	2.50	3.00	7.00	7.50	8.00	8.50	9.00	9.50	10.00
19.74	20.07	19.93	21.8	22.11	21.04	21.00	17.50	18.96	20.67	20.16	19.46	21.42	21.26

(a) Assume that $Z(\cdot)$ is a normal (Gaußian) process with constant mean m and covariance function

$$\mathrm{cov}\{Z(x), Z(x')\} = \sigma^2 \exp(-\lambda |x - x'|).$$

Give formulae for

$$\widehat{Z}(x) = \mathrm{E}\{Z(x) \,|\, \mathrm{data}\} \quad \text{and} \quad \mathrm{pe}(x) = \big[\mathrm{Var}\{Z(x) \,|\, \mathrm{data}\}\big]^{1/2}.$$

(b) Using $\lambda = 1.33$ and $\sigma = 0.98$, estimate the mean parameter m. How would you interpret the parameter λ and its value here? Also produce plots of the spatial interpolator $\widehat{Z}(x)$ (with the observed $Z(x_i)$ appearing in the plot) and of the prediction error $\mathrm{pe}(x)$.

(c) Explain briefly how the interpolator changes if one uses a linear trend function rather than a constant mean; if you have time, implement and display the modified interpolation curve.

(d) Going back to the basic normal model with constant m and σ, give a formulae for

$$\operatorname{cov}\{Z(x), Z(x') \,|\, \text{data}\}.$$

Use this to simulate and display say ten realisations of $Z(\cdot)$, given the fourteen data points (i.e. the values of $Z(x_i)$ for $i = 1, \ldots, 14$). – For these simulations, use values $m = 20.25$, $\sigma = 0.98$, $\lambda = 1.33$.

(e) Still using the values of m, σ, λ of (d), simulate a large number of $Z(\cdot)$ realisations, given the data, and compute for each of these

$$Z_{\min} = \min\{Z(x) \colon 0 \le x \le 10\}.$$

Give a histogram of these simulated $Z_{\min}$ values, and estimate the probability that $Z_{\min} \le 17.00$.

(f) Finally, find estimates for the three model parameters, using the (admittedly small) dataset.

Exercise 3

"Independence? That's middle class blasphemy." This exercise is therefore concerned with chains exhibiting Markovian dependence.

(a) Let $X_1, \ldots, X_n$ form a (first order, i.e. one-step memory) Markov chain on some finite state space $\{1, \ldots, k\}$, with one-step transition probabilities

$$p(x_{i+1} \,|\, x_i) = \Pr\{X_{i+1} = x_{i+1} \,|\, X_i = x_i\}.$$

Show that

$$\Pr\{X_i = x_i \,|\, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}, X_{i+1} = x_{i+1}, \ldots, X_n = x_n\}$$

indeed depends only upon the neighbours $x_{i\pm1}$ of i. The resulting

$$p(b \,|\, a, c) = \Pr\{X_i = b \,|\, X_{i-1} = a, X_{i+1} = c\}$$

are called the local characteristics of the Markov model.

(b) Consider a stationary Markov chain on $\{1, 2, 3\}$ with transition probability matrix

$$\mathbf{P} = \begin{pmatrix} 1-3\theta & 2\theta & \theta \\ \theta & 1-2\theta & \theta \\ \theta & 2\theta & 1-3\theta \end{pmatrix}, \tag{3.1}$$

where θ is some parameter in $(0, \frac{1}{3})$. Your first task is to simulate a long chain $X_1, \ldots, X_n$ from this model, with

$$n = 2000 \quad \text{and} \quad \theta = 0.09,$$

using the *Gibbs Sampler* technique – i.e. by simulating a chain of chains, utilising the local characteristics, but not the 'direct foward way'. Check that the relative frequencies of visits to 1, 2, 3 match those predicted by Markov chain theory (the equilibrium distribution), and compute also the matrix of

$$\widehat{p}(b \,|\, a) = \frac{N_{a,b}}{N_{a,\cdot}} = \frac{N_{a,b}}{\sum_c N_{a,c}} \quad \text{for } a, b = 1, 2, 3,$$

where

$$N_{a,b} = \sum_{i=2}^{n} I\{(X_{i-1}, X_i) = (a, b)\} \quad \text{for } a, b = 1, 2, 3$$

is the number of observed 'from a to b' transitions. Comment on what you find. – To simplify the Gibbs sampling machinery you may fix $X_1 = 2$ and $X_n = 2$, only tending to the randomness of $X_2, \ldots, X_{n-1}$.

– We now assume that there is a real chain $x_1, \ldots, x_n$, with $n = 250$, that we may only observe in a blurred fashion, via

$$y_i = x_i + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i are assumed i.i.d. $\mathrm{N}(0, \sigma^2)$. In other words, the y_i are independent given the x sequence, and

$$y_i \,|\, x_i \sim f(y_i \,|\, x_i) = \mathrm{N}(x_i, \sigma^2)(y_i) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma} \exp\Big\{-\tfrac{1}{2} \frac{(y_i - x_i)^2}{\sigma^2}\Big\}, \tag{3.2}$$

with observed y_i but unknown x_i. The dataset `examdata103` contains these data (i, y_i) for $i = 1, \ldots, 250$. The restoration (or reconstruction) task is to estimate the x process from the observed y process.

(c) Using these data, find

$$\widehat{x}_i = \begin{cases} 1 & \text{if } y_i \text{ is closest to } 1, \\ 2 & \text{if } y_i \text{ is closest to } 2, \\ 3 & \text{if } y_i \text{ is closest to } 3. \end{cases}$$

This is the non-contextual classifier, set to work without any knowledge of or use of any spatial continuity of the x process. Display $(\widehat{x}, y)$ in a diagram.

(d) Assume now that the x chain was generated via a first-order Markov chain of the type (3.1). Show that $x \,|\, y$ (the unknown, given the known) follows a first-order Markov chain over $\{1, 2, 3\}$, and provide formulae for

$$p(x_i = b \,|\, x_{\mathrm{rest}}, y) = \Pr\{X_i = b \,|\, X_{i-1} = a, X_{i+1} = c, y\} \quad \text{for } i = 2, \ldots, n-1.$$

(This data-conditional Markov chain is not stationary, since these local characteristics depend on y_i.)

(e) Implement and run a Gibbs Sampler to simulate many x chains, from this conditional distribution given the data. For this, use $\theta = 0.09$ in (3.1) and $\sigma = 0.80$ in (3.2), and again we avoid boundary problems by fixing $x_1 = 2$ and $x_n = 2$, focussing attention on $x_2, \ldots, x_{n-1}$. Compute

$$x_i^* = \text{the most probable } x_i \text{ given data,}$$

i.e. the state, among $1, 2, 3$, that has highest $p(x_i = a \,|\, y)$. Display (x^*, y) in a figure.

(f) Consider also

$$A_3 = \sum_{i=1}^{n} I\{x_i = 3\},$$

the number of times the real (but unknown) x chain has visited state 3. Please compute three different estimates of A_3: by plugging in the non-contextual $\hat{x}_i$ of (c); by plugging in the contextual x_i^* of (e); and by computing $\mathrm{E}(A_3 \,|\, \text{data})$. Which estimate would you consider to be the best?

(g) Above we took given values of θ and σ. Assume now that θ is unknown (but that the (3.1) model is in force), that $\sigma = 0.80$ is known, and attempt to find a good value of θ from the observed data.

Exercise 4

YET LOVE AND HATE ME TOO: SO THESE EXTREMES shall ne'er their office do, says John Donne (in words set to music by Ketil Bjørnstad, 2008, and performed for this year's Abel Prize winners).

(a) Assume first that $U_1, \ldots, U_n$ are an i.i.d. sample from the uniform distribution on $[0, 1]$. Find the explicit distribution for $M_n^u = \max_{i \le n} U_i$. Show that $\mathrm{E}\, M_n^u = 1 - 1/(n+1)$, and find the limit distribution of $n(1 - M_n^u)$.

(b) Let then $X_1, \ldots, X_n$ be an i.i.d. sample from some continuous distribution function F. Show that X_i and $F^{-1}(U_i)$ have the same distribution, and that this implies that

$$M_n = \max_{i \le n} X_i \text{ and } F^{-1}(M_n^u) \quad \text{have the same distribution.}$$

Explain why this indicates that $b_n^0 = F^{-1}(1 - 1/n)$ is a reasonable approximation for the mean value of M_n. Find also a good approximation to the median value of M_n, in terms of an appropriate high quantile of F.

(c) Suppose $X_1, \ldots, X_n$ are i.i.d. observations from the probability density

$$f(x) = \frac{\alpha}{x^{\alpha+1}} \quad \text{for } x \ge 1, \tag{4.1}$$

where α is some positive parameter. Show that its cumulative distribution function is $F(x) = 1 - 1/x^\alpha$ for $x \ge 1$, and that

$$\Pr\{M_n / b_n \le x\} \to H(x) = \exp(-1/x^\alpha) \quad \text{for all } x > 0,$$

where $b_n = n^{1/\alpha}$. Identify the parameters ξ, μ, ψ from the general GEV distribution formula

$$H(x) = \exp\Big\{-\Big(1 + \xi\frac{x-\mu}{\psi}\Big)^{-1/\xi}\Big\}.$$

(d) Consider exceedances $Y = X - u$ for those X that exceed a certain threshold $u > 1$. Show that such exceedances have distribution function

$$G_u(y) = 1 - (1 + y/u)^{-\alpha} \quad \text{for } y \geq 0.$$

The dataset `examdata104` (available at the course website) consists of a total of 117 observations, namely those from a rather larger data set that stems from the distribution (4.1) and that succeeded in exceeding the threshold 4.00. Estimate α from these data, and give a confidence interval for that parameter with confidence level approximately 95%.

(e) Based on these data, that all exceeded 4.00, one is interested in estimating the probability $p = \Pr\{X \geq 3.00\}$. Provide such an estimate, along with a confidence interval, and make clear which assumptions are being used.

(f) An estimate of α was found above utilising only data points exceeding a certain threshold. Explain why restricting tail estimation to such high-valued data is often sensible, even if the full dataset (i.e. all $X \geq 1$) is available.

Exercise 5

CITIUS, ALTIUS, FORTIUS indeed, and yesterday [as Nils writes this] Usain Bolt of Jamaica set a new World Record on the 100 metres sprint, completing the dash in 9.72 seconds. The point of this exercise is to discuss 'how unlikely' this event is, based on top sprint data from the eight previous seasons 2000–2007; cf. the figure below.

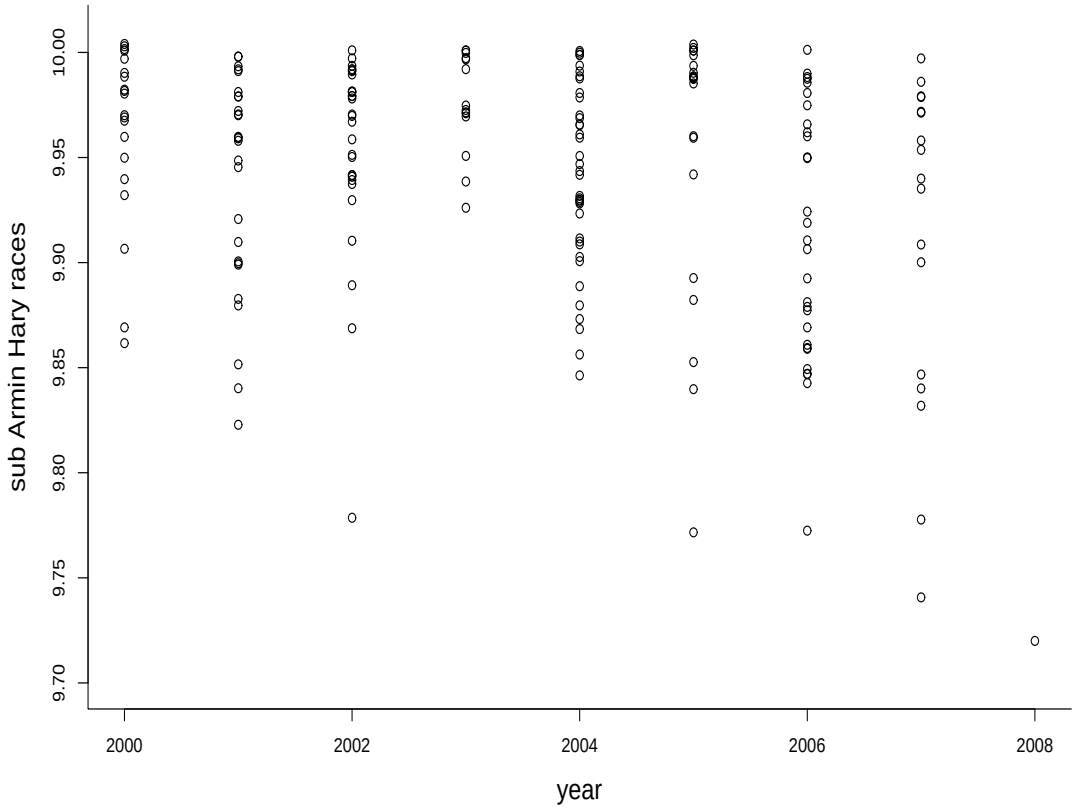
The dataset `examdata105` contains all the $n = 195$ top results achieved by humankind over the last eight seasons, defined here as 10.00 or better. I include all such races, for each season, not only one race for each top sprinter of that year (thus Asafa Powell, the previous World Record holder, accounts for 8 of the 17 sub-ten races during 2007, for example, and each of these eight races are inside my dataset). Results achieved by athletes taken in doping have been forcefully removed from my dataset (Ben Johnsen's 9.79 from 1988, Tim Montgomery's 9.78 of 2002, etc.). Similarly results achieved with too strong tail wind (more than 2.0 m/s) are not taken on board here (cf. the 9.69 of Obadele Thompson of Barbados in 1996, with tail wind in excess of 5 m/s).

We adopt the view here that the world is 'essentially stable', from 2000 to 2008, regarding top results on the 100-m dash. This is at least not unreasonable: Armin Hary did 10.0 already in 1960 (and his autobiography is proudly and appropriately titled '10,0'); Carl Lewis raced 9.86 as early as in 1991; and new World Records have been set only eleven times since electronic timing was made mandatory in 1968 (followed later by accurate tail-wind measurement devices). Geir Moen managed 10.08 in 1996, and Jaysuma Saidy Ndure (Gambian born, but Norwegian since December 2006) did 10.01 three weeks ago.

To approach problems associated with extreme value probabilities it is here convenient to translate race results to 'amount of time better than 10.00'. Viewing 'better than 10.00' as equivalent to running time less than or equal to 10.005 (since times are given down to hundredths of second), we are led to

$$Y = 10.00 - \text{race result} + 0.005 = 10.005 - \text{race result} \tag{5.1}$$

(so a very fast race is equivalent to a high Y; no Scandinavian has ever managed a positive Y, but Ndure is only 1 centimetre away). These are found in the third column of the datafile `examdata105`.



Usain Bolt *set a new World Record yesterday [as Nils writes this] with 9.72. His achievement may be compared to the 195 previous occasions, during seasons 2000 to 2007, where sprinters have raced 10.00 or faster (found here, there & everywhere and stitched together from some hours of book reading and internet trawling). How 'unlikely' is 9.72, is he a Lightning Bolt from Heaven?*

(a) General extreme value statistics theory may be used to claim that such Y data must follow a distribution that may be approximated by

$$G(y) = 1 - (1 + \xi y/\sigma)^{-1/\xi} \quad \text{for } y > 0. \tag{5.2}$$

Discuss the underlying assumptions for this statement, in view of the character of the present data.

(b) Regardless of any critical points you might have enlisted for point (a), we now view the $n = 195$ data points $Y_1, \ldots, Y_n$ (each of the form '10.005 minus race result') as a sample of independent data from the distribution (5.2). Estimate the parameters (ξ, σ), using maximum likelihood. Display the estimated probability density curve, say $g(y, \widehat{\xi}, \widehat{\sigma})$, along with the histogram of the Y data. Also provide approximate standard errors (square roots of estimated variances) for these two parameter estimates.

(c) Probability calculations in this point will be carried out pretending that today is 1 January 2008, i.e. conditional on data of 2007 and earlier only. From this perspective, consider

$$W = \text{best result during 2008} = \max\{Y_1', \ldots, Y_N'\}$$

(on the Y scale of (5.1)), where N is the (unknown) number of 10.00-or-better races during 2008 (if any at all) and $Y_1', \ldots, Y_N'$ given N are viewed as i.i.d. with distribution (5.2). In addition, take N to be Poisson with parameter λ (such an approximation is implied by general extreme value statistics theory). Show that

$$\Pr\{W \geq w\} = 1 - \exp\{-\lambda(1 + \xi w/\sigma)^{-1/\xi}\}.$$

(d) Argue that $\lambda = 195/8 = 24.375$ is a reasonable value of the Poisson intensity λ in the present context. Compute the probability, again pretending that today is 1 January 2008 (i.e. we do not yet know any race results for 2008, and we still believe that Usain Bolt's personal best is 10.03, from July 2007), that someone on the planet will do a 100-m race in 9.72 seconds or less, in the course of 2008. – Discuss my claim that this is the natural 'surprise level probability' that may be associated with the Bolt 9.72 news of 31 May 2008 (cf. again the figure). How surprised are you?

(e) Letting q denote the probability defined in (d), supplement your point estimate $\widehat{q}$ with an approximate 95% confidence interval (you may take the sub-ten Poisson rate $\lambda = 195/8$ as a given number here).

(f) Finally, and once more computing probabilities from the perspective of 1 January 2008, estimate and display the probability distribution of

$$R = \text{best race result during 2008–2011} = \min\{X_1', \ldots, X_M'\},$$

where $X_1', \ldots, X_M'$ are all race results at 10.00 or better set during 2008–2011 (i.e. in the present Olympic period, from pre-Beijing up to pre-London).

Saturday May 31, 2008, at Icahn Stadium in New York: 9.72