UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 9101SP, special curriculum: Probability, Measure, Large-Sample Theory**
FOR:	**Emil Aas Stoltenberg**
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**23.v.–6.vi.2016**

This is the exam project set for STK 9101SP, special curriculum on probability, measure and large-sample theory, spring semester 2016. It is made available for the candidate as of *Monday 23 May 12:00*, and he should submit his written report by *Monday 6 June 13:00* (or earlier), electronically as a pdf, to Nils Lid Hjort. There will also be a *conversation* with the candidate, Hjort and a colleague, with a blackboard nearby, on Tuesday 7 June; this conversation might touch both the exam project report and other aspects of the special curriculum.

The candidate is required to work by himself, i.e. independently of others. Importantly, in his report the candidate should also include a one-page summary of the work carried out, and this should also contain a brief self-assessment of its quality.

This exam set contains four exercises and comprises four pages.

Exercise 1

Consider independent observations $Y_1, \ldots, Y_n$ from the normal distribution with parameters (μ, σ).

(a) Write down the log-likelihood function $\ell_n(\mu, \sigma)$ and derive formulae for the maximum likelihood estimators, say $(\widehat{\mu}, \widehat{\sigma})$. Identify also the exact distributions for these.

(b) Show that

$$\begin{pmatrix} \sqrt{n}(\widehat{\mu} - \mu) \\ \sqrt{n}(\widehat{\sigma} - \sigma) \end{pmatrix} \to_d \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{1}{2}\sigma^2 \end{pmatrix}\Big).$$

Briefly explain or investigate how accurate the ensuing approximations are.

(c) Consider the parameter $\gamma = \mu/\sigma$, sometimes called the normalised mean. With $\widehat{\gamma} = \widehat{\mu}/\widehat{\sigma}$, show that

$$\sqrt{n}(\widehat{\gamma} - \gamma) \to_d \mathrm{N}(0, 1 + \tfrac{1}{2}\gamma^2).$$

(d) Let now $h(x) = \sqrt{2}\log(x/\sqrt{2} + \sqrt{1 + \frac{1}{2}x^2})$. Show that

$$\sqrt{n}\{h(\widehat{\gamma}) - h(\gamma)\} \to_d \mathrm{N}(0, 1).$$

(e) Suppose you observe $\widehat{\mu} = 3.333$ and $\widehat{\sigma} = 2.222$ from $n = 50$ observations. Find an approximate 90% confidence interval for γ based on (d). Compute and display a full confidence curve for γ.

(f) Suppose now that laboratories in the five Nordic countries carry out similar experiments, for simplicity taken to have the same sample size n for each, leading to maximum likelihood estimates $\widehat\gamma_1,\ldots,\widehat\gamma_5$. Devise a test for the null hypothesis that the underlying $\gamma_j=\mu_j/\sigma_j$ parameters are identical.

Exercise 2

Consider a sequence $X_1,X_2,\ldots$ of independent and identically distributed random variables, with some continuous distribution. We take an interest in the *records*, cases where an observation is bigger than all previous observations. Define therefore $Y_1=1$ and

$$Y_n=\begin{cases}1 & \text{if } Y_n>\max(Y_1,\ldots,Y_{n-1}),\\ 0 & \text{if else,}\end{cases}$$

for $n\ge 2$, so that $A_n=\sum_{i=1}^n Y_i$ is the number of records experienced in the course of the first n observations.

(a) Show that $Y_1,Y_2,\ldots$ are independent Bernoulli variables, with

$$\Pr\{Y_i=1\}=1/i \quad \text{for } i=1,2,\ldots.$$

(b) Study

$$Z_n=\frac{A_n-\xi_n}{\sigma_n}$$

where $\xi_n=\sum_{i=1}^n 1/i$ and $\sigma_n^2=\sum_{i=1}^n(1/i)(1-1/i)$. Find an expression for the skewness $\mathrm{E}\,Z_n^3$ and show that it converges to zero.

(c) Show that $Z_n\to_d \mathrm{N}(0,1)$. There might be easier ways to do this than by showing that the moment-generating function

$$M_n(t)=\mathrm{E}\,\exp(tZ_n)$$

tends to $\exp(\frac12 t^2)$, but please do attempt to show this too.

(d) In the following you may use

$$\sum_{i=1}^n\frac{1}{i}-\log n\to\gamma=0.5772\ldots \quad\text{and}\quad \sum_{i=1}^\infty\frac{1}{i^2}=\frac{\pi^2}{6}$$

(Euler's sensational finding from 1735). Show that

$$\frac{A_n-\log n}{\sqrt{\log n}}\to_d \mathrm{N}(0,1).$$

(e) I wonder: about how many new records will be set during the first million observations? Construct an interval that with probability approximately 90% contains $Y_{1\,000\,000}$. Attempt to compute and display the associated (approximate) full predictive confidence distribution.

(f) For $t > 1$, let $N(t)$ be the number of records set for indexes i with $n \le i \le [nt]$. Show that $N(t)$ tends in distribution to a Poisson with parameter $\log t$. Try to show that the full $N = \{N(t) \colon t \ge 1\}$ tends to the relevant Poisson process.

(g) We now forget the records for a while, and study the case where the Y_i are independent Bernoulli variables with probabilities $p_i = 1/i^2$. Then show that the number of times $Y_i = 1$ is finite, with probability 1. Attempt to find workable expressions for the probability distribution for K, the last n for which $Y_n = 1$, and display this distribution in a suitable diagram.

(h) Still with $p_i = 1/i^2$, investigate aspects of the limit distribution for $Z_n = (A_n - \xi_n)/\sigma_n$, where now $\xi_n = \sum_{i=1}^n 1/i^2$ and $\sigma_n^2 = \sum_{i=1}^n (1/i^2)(1 - 1/i^2)$. You may in particular find and numerically evaluate (or approximate) the limit of the log-moment-generating function, and display this limit alongside the 'normal limit', i.e. $\frac{1}{2}t^2$.

Exercise 3

Let's consider a very simple regression model, with just one covariate x_i associated with the response Y_i, and where

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

with the ε_i being i.i.d. from some distribution with zero mean and finite variance, say σ^2. We do *not* assume here that this error distribution is normal.

The ordinary least squares estimator is

$$\widehat{\beta}_n = \frac{\sum_{i=1}^n x_i Y_i}{M_n} \quad \text{with } M_n = \sum_{i=1}^n x_i^2.$$

Assume that $M_n \to \infty$ and that

$$\Delta_n = \max_{i \le n} |x_i|/M_n^{1/2} \to 0.$$

Show that then

$$M_n^{1/2}(\widehat{\beta}_n - \beta) \to_d \mathrm{N}(0, \sigma^2).$$

For the case of $x_i = i$, for example, find the limit distribution of $n^{3/2}(\widehat{\beta}_n - \beta)$.

Exercise 4

Consider independent observations $Y_1, \ldots, Y_n$ from a symmetric density on the real line, where the task is to estimate the unknown symmetry point, say θ. Thus

$$Y_i = \theta + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where the ε_i stem from a density $f_0(x)$ symmetric around zero. A standard estimator is of course $\widetilde{\theta} = \bar{Y}$, the mean of the data, which may also be seen as the solution to $\sum_{i=1}^n (Y_i - \theta) = 0$. That method is famously rather unrobust. As a more robust alternative, study $\widehat{\theta}$, the solution to

$$\sum_{i=1}^n \arctan(Y_i - \theta) = 0.$$

(a) Show that the $\widehat{\theta}$ estimator exists and is unique. Indicate why this method may be expected to be more robust than $\widetilde{\theta}$.

(b) Show that there is limiting normality,

$$\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathrm{N}(0, \kappa^2),$$

and identify the κ. In particular, in the standard case where the observations are from the $\mathrm{N}(\theta, 1)$ model, how much is lost in efficiency by using the arctan estimator? Also find the limiting correlation between $\bar{Y}$ and $\widehat{\theta}$.

(c) One may 'work backwards' from the estimating equation $\sum_{i=1}^{n}(Y_i - \theta) = 0$ and deduce that the estimator in question is the maximum likelihood estimator for the normal model. Attempt similarly to identify a model $f(y, \theta) = f_0(y - \theta)$ such that the arctan estimator is its maximum likelihood estimator.

(d) Generalise the setup to include also a scale parameter.