UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4011 / 9011– Statistical Inference Theory**
	Part I of two parts: The project
WITH:	**Nils Lid Hjort**
TIME FOR EXAM:	**2.–18.xii.2014**

This is the exam project set for STK 4011 / 9011, autumn semester 2014. It is made available on the course website as of *Tuesday 2 December 12:00*, and candidates must submit their written reports by *Thursday 18 December 14:00* (or earlier), to the reception office at the Department of Mathematics, in duplicate. The supplementary four-hour written examinations take place *Monday December 8* (practical details concerning this 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). They are graciously allowed not to despair if they do not manage to answer all questions well.

This exam set contains three plus one exercises and comprises five pages. The three first exercises are for both the STK 4011 and STK 9011 students, whereas the PhD students taking the STK 9011 version of the course also should do Exercise 4.

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.

Exercise 1

THE MEDIAN ISN'T THE MESSAGE – said Stephen Jay Gould, though that does not stop us from investigating some of the median issues. Consider independent and uniformly distributed variables $U_1, \ldots, U_n$ on the unit interval $[0,1]$, then sorted to $U_{(1)} < U_{(2)} < \cdots < U_{(n-1)} < U_{(n)}$.

(a) Show that the density of $U_{(i)}$ is

$$g_i(u) = \frac{n!}{(i-1)!(n-i)!} u^{i-1}(1-u)^{n-i} \quad \text{on } (0,1).$$

Find its mean and variance.

(b) For an order statistics sample $U_{(1)} < \cdots < U_{(100)}$ of size 100 from the uniform distribution, find the correlation between $U_{(33)}$ and $U_{(34)}$.

(c) Consider in particular the median M_n^0, and let us for convenience take $n = 2m + 1$ to be odd, so that $M_n^0 = U_{(m+1)}$. Give the density for $Z_n = \sqrt{n}(M_n^0 - \frac{1}{2})$, and show that it converges to the density of $\mathrm{N}(0, \frac{1}{4})$. You may use here Stirling's formula, which says that $k! \doteq k^k \exp(-k)\sqrt{2\pi k}$ for large k (and where the '$\doteq$' means that the ratio between the left and right hand sides converges to 1); also, $(1+c/k)^k \to e^c$ as $k \to \infty$.

(d) Assuming now that $X_1, \ldots, X_n$ are independent and identically distributed with positive density $f(x)$ and cumulative distribution function $F(x)$ on some interval, show that $X_{(i)}$ and $F^{-1}(U_{(i)})$ must have the same distribution.

(e) Let M_n be the sample median of the X_i, viewed as an estimator of the population median $\mu = F^{-1}(\frac{1}{2})$. Use the delta method to show that

$$\sqrt{n}(M_n - \mu) \to_d \mathrm{N}(0, \kappa^2), \quad \text{with } \kappa^2 = \tfrac{1}{4}\frac{1}{f(\mu)^2}.$$

(The delta method says that if $\sqrt{n}(Z_n - c) \to_d W$, then $\sqrt{n}\{h(Z_n) - h(c)\} \to_d h'(c)W$.)

(f) Suppose $X_1, \ldots, X_n$ are independent from a distribution with a density f symmetric around its centre point μ. Then both the sample average $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ and the sample median M_n can be used for estimating μ. Give a formula for the approximate variance of M_n and compare it to the variance formula for $\bar{X}_n$. Illustrate this for the case of the normal distribution. Then attempt to find a symmetric density where the sample median achieves a smaller variance than the sample mean.

Exercise 2

NEVER LET ANYONE DEFINE WHAT YOU ARE CAPABLE OF by using parameters that don't apply to you. Assume independent observations $X_1, \ldots, X_n$ follow the distribution with cumulative function

$$F(x) = (x/b)^a \quad \text{for } 0 \le x \le b.$$

Here a and b are two positive parameters. For questions (a)-(b)-(c) below we work with b known and a unknown and to be estimated; for questions (d)-(e)-(f) we work correspondingly with a known and b unknown. Then, finally, we work with questions (g)-(h)-(i)-(j) under the more natural but challenging assumption that both parameters are unknown.

(a) We shall first assume that b is a known constant, and shall investigate a coupl' o' estimation strategies for a. Show that the maximum likelihood estimator is

$$\widehat{a}_{\mathrm{ml}} = \Big\{\frac{1}{n}\sum_{i=1}^n \log\frac{b}{X_i}\Big\}^{-1}.$$

(b) Find also explicit formulae for the method of moments estimator $\widehat{a}_{\mathrm{mom}}$, found by equating the sample mean to the population mean, and the the method of quantiles estimator $\widehat{a}_{\mathrm{moq}}$, found by equating the sample median to the population median. These estimators are again allowed to depend on b.

(c) Find the limit distributions

$$\sqrt{n}(\widehat{a}_{\rm ml} - a) \to_d {\rm N}(0, \kappa_1^2),$$
$$\sqrt{n}(\widehat{a}_{\rm mom} - a) \to_d {\rm N}(0, \kappa_2^2),$$
$$\sqrt{n}(\widehat{a}_{\rm moq} - a) \to_d {\rm N}(0, \kappa_3^2),$$

with appropriate formulae for $\kappa_1, \kappa_2, \kappa_3$. You may find it useful to apply the delta method for some of these asymptotic calculations. Which of the three estimators considered here is best?

(d) Now assume that a is known and fixed, with b the unknown parameter to estimate from data. Show that the maximum likelihood estimator is $\widehat{b}_{\rm ml} = V_n = \max_{i \le n} X_i$.

(e) Find also formulae for the method of moments estimator $\widehat{b}_{\rm mom}$ and method of quantiles estimator $\widehat{b}_{\rm moq}$ (allowed to depend on a).

(f) Find the limit distribution of $W_n = n(1 - V_n/b)$. Use this to construct an approximate 95% confidence interval for b (still with a a known parameter).

(g) Time has finally come to examine the case where both parameters a and b are unknown and are to be estimated. Find the maximum likelihood estimators $a^*_{\rm ml}$ and $b^*_{\rm ml}$, and show that these are consistent for a and b.

(h) Find the limit distribution of $\sqrt{n}(a^*_{\rm ml} - a)$.

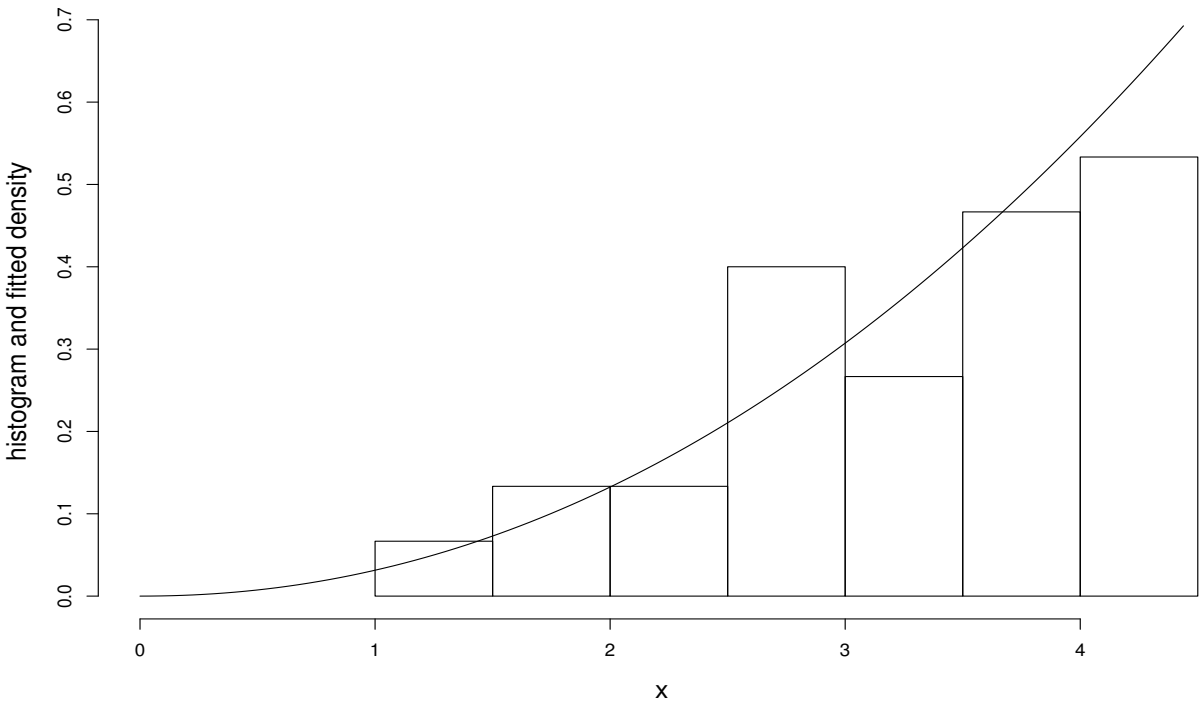

Figure: Histogram of the thirty data points of Exercise 2(j), along with the fitted density from the two-parameter model discussed there.

(i) Try to find the exact distributions of $\widehat{a}_{\rm ml}$ (given by the formula in (a), where it was allowed to depend on b) and also the exact distribution of $a^*_{\rm ml}$. How much is lost in precision, when passing from $\widehat{a}_{\rm ml}$ to $a^*_{\rm ml}$, by not knowing the value of b?

(j) The two-parameter model above is being used in various contexts, e.g. in connection with situations where the individual data point suitably represents the 'best achievement' and where there is an underlying upper threshold for such achievements. I have used the model to model log-counts for Google Scholar profiles, listing the most frequently cited publications of different scholars. The following simple data set stems via a suitable transformation from such an analysis, with thirty data points. Take these to be independent from the two-parameter model above. Find estimates and (exact or approximate) 90% confidence intervals for a and b. Try to reproduce the figure shown on page 3.

```
1.402  1.583  1.728  2.043  2.424  2.740  2.811  2.826
2.953  2.961  2.965  3.258  3.481  3.481  3.481  3.520
3.632  3.643  3.718  3.755  3.878  3.879  4.121  4.146
4.148  4.268  4.269  4.270  4.301  4.439
```

Exercise 3

In order to shake a hypothesis, it is sometimes not necessary to do anything more than push it as far as it will go. Suppose θ is some parameter of interest, associated with observations $X_1, \ldots, X_n$. The null hypotheses traditionally dealt with in statistical testing are of the type $\theta = \theta_0$, or $\theta \leq \theta_0$, or $\theta \geq \theta_0$, for some pre-specified value θ_0, or even $|\theta - \theta_0| \leq \varepsilon$ for some small positive ε. On this occasion we turn things slightly around, however, and wish to test $H_0\colon |\theta - \theta_0| \geq \varepsilon$ versus the alternative that $|\theta - \theta_0| < \varepsilon$.

(a) Describe a situation where such a scenario would be fruitful.

(b) To give an illustration of more general constructions of the type pointed to above, suppose now that observations $X_1, \ldots, X_n$ are independent and normal $\mathrm{N}(\theta, 1)$, and assume for simplicity that $\theta \geq 0$ a priori. We shall test the hypothesis H_0 that $\theta \geq \varepsilon$, versus the alternative that $\theta < \varepsilon$, where we for concreteness set $\varepsilon = \frac{1}{4}$. Consider the test which rejects H_0 if $\bar{X}_n \leq c_n$, where $\bar{X}_n$ as usual is the average $n^{-1}\sum_{i=1}^n X_i$ of the observations. Find c_n such that this test has significance level ('type I error') 0.05.

(c) Find the power function for this test, i.e. the probability that H_0 will be rejected, as a function of the parameter. Give a plot of this power function for $n = 100$. Comment on the size of the maximal power.

(d) How big must the sample size be, in order for the above power probability to be above 0.95, if in fact the true θ is equal to $\frac{1}{2}\varepsilon$ (i.e. $\frac{1}{8}$)?

(e) Show that the test worked with here, rejecting $H_0\colon \theta \geq \frac{1}{4}$ vs. the alternative that $\theta < \frac{1}{4}$ when $\bar{X}_n \leq c_n$, is uniformly most powerful, among all tests with significance level 0.05.

(f) Suppose that θ is not restricted to be nonnegative a priori, and that one needs a test for $H_0: |\theta| \geq \frac{1}{4}$ versus the alternative that $|\theta| < \frac{1}{4}$. Construct a test for this situation, again with significance level 0.05, and draw its power function alongside the one from point (c).

Exercise 4: for the PhD students taking the STK 9011 exam

Find and read the paper *Robust and efficient estimation by minimising a density power divergence* by A. Basu, I.R. Harris, N.L. Hjort and M.C. Jones (Biometrika, 1998). Give a brief summary of the methods developed in that paper, and apply them to the analysis of a data set of your own choice, along with a discussion of your findings.