UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4090/9090:**
	Statistical Large-Sample Theory
	The project
With:	**Nils Lid Hjort**
Time for exam:	**12.–23.vi.2020**

This is the exam project set for STK 4090/9090, spring semester 2020. It is made available on the course website as of *Friday June 12, 11:11*, and candidates must submit their written reports by *Tuesday June 23, 12:12* (or earlier), to the Inspera System at the Department of Mathematics. Due to the korona situation there's no supplementary four-hour no-book exam this time. Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should be text-processed (TeX, LaTeX, Word). Give your student-web identification number 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 relevant parts 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 should they not manage to answer all questions well.

Importantly, by handing in your report to the Inspera system you guarantee that you've read, understood, and confirmed the points of the *self-declaration form* (available at the course website). Also, your report should contain *one separate extra page*, the student's one-page summary of the exam project report, which should briefly tell its readers about how the work has proceeded, and also contain a brief self-assessment of its quality. You may make this the very last page of your report.

This exam set contains four exercises and comprises six pages.

Exercise 1

Itj fårrå nålles. You need to access and upload the dataset `allwars-data`, available at the course website. It contains data pairs (x_i, z_i) for all $n_0 = 95$ gruesome interstate wars with at least 1000 battle deaths (as per well-maintained and publicly available databases for such matters, specifically the Correlates of War project), from the Franco-Spanish war in 1823 to the invasion of Iraque in 2003. Here x_i is the time where war i started, with dates transformed via months and days to decimals, so that the Korean war started at $x_{60} = 1950.483$, etc.; and z_i is the number of battle deaths. Figure A displays the $(x_i, \log z_i)$ data (one may exp the vertical scale to perhaps understand a bit better the deeper horror of these numbers).

In the present exercise we shall mostly work only with the x_i, not the battle deaths, and more specifically with *the between-times*

$$w_i = x_{i+1} - x_i \quad \text{for } i = 1, \ldots, n,$$

say, with $n = n_0 - 1 = 94$. Apart from a brief excursion in question (g) we leave work on battle deaths, and yet other features and connections and predictions and analyses of statistical sightings of better angels, to other occasions.

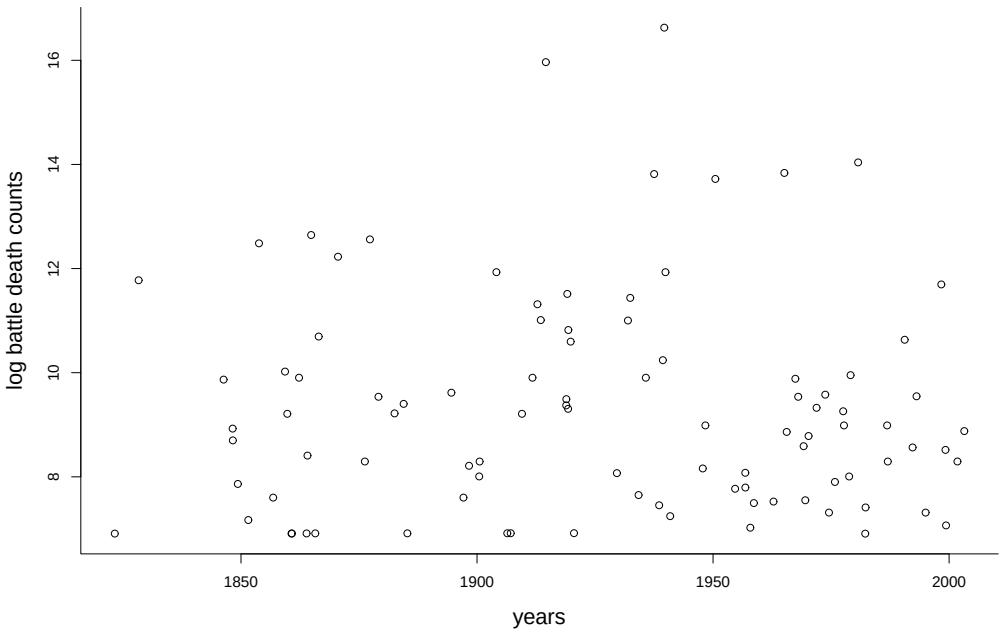

Figure A: The onset time x_i and log battle deaths count $\log z_i$ for the last 95 major interstate wars, those with $z_i \geq 1000$.

(a) There are both empirical studies and certain theoretical arguments, also for many other types of violence phenomena, pointing to the interesting and non-obvious supposition that the between-times ought to be approximately independent and identically exponentially distributed. In other words and terms, the w_i will behave as waiting times in a Poisson process with constant rate. Fit the model

$$f(w, \lambda) = \lambda \exp(-\lambda w) \quad \text{for } w > 0$$

to the $w_1, \ldots, w_n$ data, via maximum likelihood. Assuming the model holds, give a 90 percent confidence interval for λ.

(b) Broader models emerge by taking the w_i given λ to be exponential with this parameter λ, but to take the λ not as a single constant, but coming from a distribution of such rates. Assume that λ comes from a Gamma distribution with parameters (a, b), i.e. with density proportional to $\lambda^{a-1} \exp(-b\lambda)$. Show that this leads to the density

$$g(w, a, b) = \frac{ab^a}{(b + w)^{a+1}} \quad \text{for } w > 0.$$

(c) Fit also this two-parameter model to the $w_1, \ldots, w_n$ data, using maximum likelihood, and give approximate standard errors for the estimates $(\widehat{a}, \widehat{b})$. Give the estimated mean and standard deviation for this distribution of λ values. Construct a version of Figure B, which has the empirical cumulative distribution function along with the fitted parametric cumulatives for the one-parameter and two-parameter models.

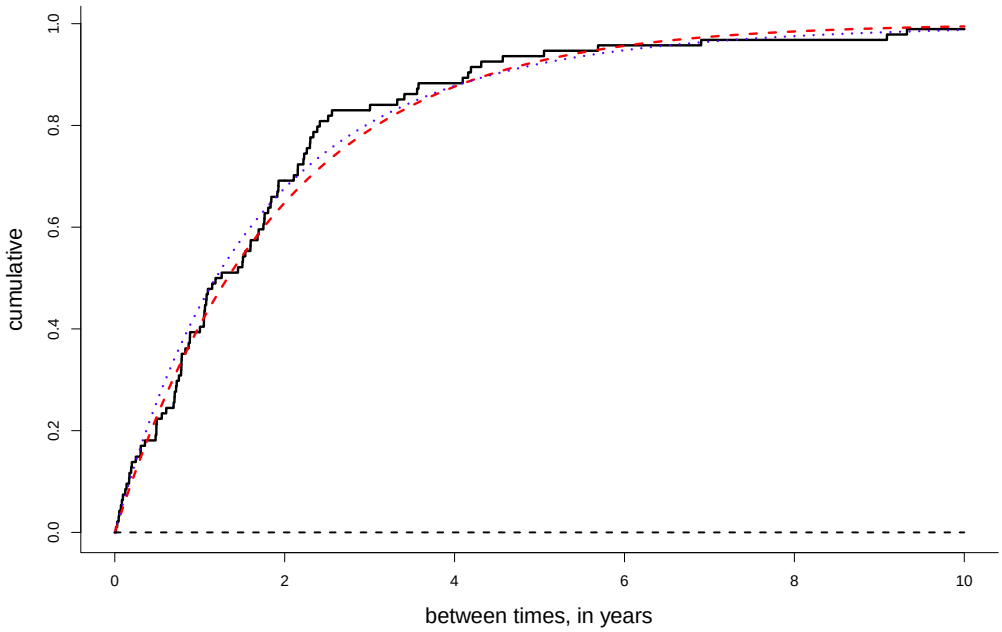

Figure B: The empirical distribution function for the between-wars time data (black curve), along with the two fitted parametric cumulatives $F(w, \widehat{\lambda})$ and $G(w, \widehat{a}, \widehat{b})$.

(d) For the one-parameter model, find a formula for the probability $p = p_1(\lambda)$ that the time between two consecutive wars is at least $w_0 = 3.00$ years. Construct and display a full confidence distribution for this probability parameter, say $\text{cc}_1(p)$, again assuming that the model holds.

(e) Then supplement the above $\text{cc}_1(p)$ with the similar confidence curve $\text{cc}_2(p)$, now using the two-parameter model, starting with a formula $p = p_2(a, b)$ for the probability that the waiting time between two wars is at least $w_0 = 3.00$ years. Read off (approximate) 90 percent confidence intervals for p, under the one-parameter and the two-parameter models.

(f) Which of the two models might be best?

(g) *Why do the nations so furiously rage together? / Why do the people imagine a vain thing?* Perhaps the size of a war influences the eagerness with which cohorts of humankind again decide to embark on the next war? Fit the model where $w_i = x_{i+1} - x_i$ is an exponential with parameter $\lambda_i = \lambda_0 \exp(\beta v_i)$, where $v_i = \log z_i$, and comment on your findings. (And sing along, in C major.)

(h) Above various analyses have been based on the observed between-war times, up to $w_{94} = x_{95} - x_{94}$. There is also information in the fact that since onset time $x_{95} = 2003.219$, there have gudsigforbyde as of June 1, 2020, been no further interstate wars (well, according to the operative definitions of the Correlates of War project). Explain how this may be used to modify or update your previous analyses.

Exercise 2

Shall I compare thee to a, well, better estimator? For this exercise, assume $Y_1, \ldots, Y_n$ are i.i.d. on the unit interval, from a positive density. The parametric model to be worked with has density $f(y, a) = ay^{a-1}$, with a an unknown positive parameter.

(a) Write down the log-likelihood function, and find a formula for the maximum likelihood (ML) estimator, $\widehat{a}$. Assume first that the parametric model is correct, so there is a value a_0 such that $f(y, a_0)$ has generated the data. Find the limiting distribution of $\sqrt{n}(\widehat{a} - a_0)$.

(b) We now take an interest in the median of the distribution, say $\mu = F^{-1}(\frac{1}{2}, a)$, in terms of the cumulative $F(y, a)$. Find the ML estimator $\widehat{\mu}$ for μ, and also the limiting distribution for $\sqrt{n}(\widehat{\mu} - \mu_0)$, with $\mu_0 = F^{-1}(\frac{1}{2}, a_0)$ the true median.

(c) An alternative median estimator is of course the sample median itself, say M_n. Find the limiting distribution for $\sqrt{n}(M_n - \mu_0)$, still under parametric model conditions.

(d) Discuss how much might be lost by being 'statistically conservative', using the non-parametric sample median, instead of the ML estimator, provided the model is correct. In which cases might the sample median nevertheless be the better method?

(e) Suppose now that the true density g generating the data is a Beta density with parameters $(a, b) = (0.333, 1.222)$, rather than being of the parametric form above. Explain what the ML estimator $\widehat{a}$ is aiming for, and find the associated limit distribution.

(f) Suppose you have a dataset $y_1, \ldots, y_n$ on the unit interval, with say $n = 100$, assumed to follow a Beta distribution, but with parameters (a, b) unknown. How can you test whether the simpler model studied above, with density ay^{a-1}, holds?

Exercise 3

Brownian motion virrer as a virrevandring. Let $W = \{W(t): t \in [0, 1]\}$ be a standard Brownian motion on the unit interval – it starts at $W(0) = 0$, and its increments are independent over disjoint intervals, with $W(t) - W(s) \sim \mathrm{N}(0, t - s)$ for $s < t$.

(a) Simulate ten paths of this process, and display them in a diagram. For your ten paths, compute the empirical mean and standard deviation of the ten values of $W(1)$, and comment briefly on this.

(b) For $t \in (0, 1)$, find the distribution of $W(t)$ given that $W(1) = 0$.

(c) Consider the process

$$U(t) = \frac{W(t)}{\sqrt{t}} \quad \text{for } t > 0.$$

Find the correlation between $U(s)$ and $U(t)$, for $s \le t$.

(d) Consider the variable

$$X_m = \frac{1}{m} \sum_{i=1}^{m} W(i/m).$$

Find the distribution of this X_m, and show that $X_m \to_d \mathrm{N}(0, 1/3)$ as m grows.

(e) Show that X_m also has the limit $X = \int_0^1 W(t)\,\mathrm{d}t$.

(f) Consider now i.i.d. variables $Y_1, Y_2, \ldots$ with mean zero and variance one, and with partial sums $S_1 = Y_1$, $S_2 = Y_1 + Y_2$, etc. Show using the Donsker theorem that

$$A_n = \frac{1}{n^{3/2}} \sum_{i=1}^{n} S_i \to_d A \sim \mathrm{N}(0, 1/3).$$

(g) Show that

$$\sum_{i=1}^{n} S_i = nY_1 + (n-1)Y_2 + \cdots + 2Y_{n-1} + Y_n,$$

and use the Lindeberg theorem to give another proof of $A_n \to_d \mathrm{N}(0, 1/3)$.

Exercise 4

Normally limits are normal, but not always. Here we shall indeed work with variables with mean zero and variance one, where the sample averages have nonnormal limits. The basic construction is as follows. Let $U_1, U_2, \ldots$ be i.i.d., with mean zero and variance one, and with moment-generating function $M_0(s) = \mathrm{E}\exp(sU_i)$ finite in a neighbourhood around zero; in particular, all moments for the U_i are finite. Let independently of these $J_1, J_2, \ldots$ be independent Bernoulli variables with $\Pr\{J_i = 1\} = 1/i$, $\Pr\{J_i = 0\} = 1 - 1/i$. Then form

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J_i \sqrt{i}\, U_i = \sum_{i=1}^{n} J_i \sqrt{i/n}\, U_i.$$

A picture to have in mind is that most of the terms will be zero, with non-zero contributions becoming both more rare and more big as time proceeds.

(a) Show that there will with probability one be infinitely many $J_i = 1$, i.e. non-zero terms in the Z_n sum as n grows.

(b) Show that the terms $J_i\sqrt{i}U_i$ have mean zero and variance one; hence also the normalised sample average Z_n has mean zero and variance one. Find also an expression for the kurtosis

$$\kappa_n = \mathrm{E}\, Z_n^4 - 3$$

of Z_n, and show that $\kappa_n \to \frac{1}{2}a_4$, where $a_4 = \mathrm{E}\, U_i^4$. Compare this to what we are 'used to' from the Lindeberg theorem.

(c) We already know from point (b) that if Z_n has a limit distribution, it can't be normal. Working with the moment-generating function, show that

$$M_n(t) = \mathrm{E}\,\exp(tZ_n) = \prod_{i=1}^{n}\Big[1 + \frac{1}{i}\{M_0(t\sqrt{i/n}) - 1\}\Big],$$

for all t around zero for which $M_0(t)$ is finite.

(d) Here one may show that

$$\prod_{i=1}^{n}\Big[1 + \frac{1}{i}\{M_0(t\sqrt{i/n}) - 1\}\Big] \to \exp\Big\{\int_0^1 \frac{M_0(t\sqrt{x}) - 1}{x}\,\mathrm{d}x\Big\}. \qquad (*)$$

Work first with Special Case One, where we let U_i have the simple symmetric two-point distribution $\Pr\{U_i = 1\} = \Pr\{U_i = -1\} = \frac{1}{2}$. Find the limiting kurtosis for Z_n in this case. Show that

$$M_0(s) = \tfrac{1}{2}e^s + \tfrac{1}{2}e^{-s} = 1 + (1/2!)s^2 + (1/4!)s^4 + \cdots,$$

and use this to find an infinite-sum expression for the limit of $M_n(t)$. Have you now proved that Z_n has a limit distribution?

(e) Then work with Special Case Two, where the U_i have a double exponential distribution, of the form

$$f(u) = \tfrac{1}{2}\sqrt{2}\exp(-\sqrt{2}|u|)$$

on the real line (the $\sqrt{2}$ factor is there to ensure variance one). Find the moment-generating function $M_0(s)$ for the U_i, and then the moment-generating function $M(t)$ for the limit distribution of Z_n.

(f) For most cases, regarding the distribution for the U_i, it is hard to learn the explicit distribution for Z_n (even in cases where there might be a clear distribution for its limit). For Special Case Two, however, attempt to find the explicit distribution for Z_n, for any given n.

(g) I do not define this question as the most important one, on this occasion, but please attempt to prove the limit result of (*).