UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4011/9011 – Statistical Inference Theory**
WITH:	**Nils Lid Hjort**
AUXILIA:	**One single sheet of paper with the candidate's own personal hand-written notes** **Calculator**
TIME FOR EXAM:	**Wednesday 6/xii/2023, 15:00–19:00**

This exam set contains three exercises and comprises four pages. Write your solutions in bokmål, nynorsk, riksmål, Danish, Swedish, English, or Latin.

Exercise 1: quantiles via sums

TRANSFORMATION CAN BE A TRANSFORMATIVE EXPERIENCE. Two well-known distributions from the course material are as follows. First, the Gamma$(a,1)$ distribution has density $\Gamma(a)^{-1}x^{a-1}\exp(-x)$ for x positive, and has mean a and variance a. Second, the Beta(a,b) has density $\{\Gamma(a+b)/(\Gamma(a)\Gamma(b))\}x^{a-1}(1-x)^{b-1}$ on the unit interval, with mean $\xi=a/(a+b)$ and variance $\xi(1-\xi)/(a+b+1)$.

(a) When $X\sim$ Gamma$(a,1)$, show that its moment-generating function $\mathrm{E}\exp(tX)$ takes the form $1/(1-t)^a$. Use this to show that a sum of independent Gamma$(a_i,1)$ variables, for $i=1,\ldots,k$, is a Gamma$(\sum_{i=1}^k a_i,1)$.

(b) Suppose X and Y are independent with the same Gamma$(a,1)$ distribution. Put up an expression for the joint density $f(x,y)$ for these. Then transform these to $R=X/(X+Y)$ and $Z=X+Y$. Find their joint density $g(r,z)$. Show from this that $R\sim$ Beta(a,a), and give its mean and variance.

(c) Let $U_1,\ldots,U_n$ be i.i.d. on the unit interval, with n odd, so we can write $n=2m+1$. With $M_n=U_{(m+1)}$ the median, show that $M_n\sim$ Beta$(m+1,m+1)$. Explain that this leads to the representation

$$M_n=\frac{X_1+\cdots+X_{m+1}}{X_1+\cdots+X_{m+1}+Y_1+\cdots+Y_{m+1}}=\frac{\bar X_{m+1}}{\bar X_{m+1}+\bar Y_{m+1}},$$

in which the X_i and the Y_i are all independent and standard exponentially distributed.

(d) Use the Central Limit Theorem to show that

$$\sqrt{m+1}(\bar X_{m+1}-1)\to_d U,\quad \sqrt{m+1}(\bar Y_{m+1}-1)\to_d V,$$

where U and V are independent standard normals. Then use the delta method to find the limit distribution for $\sqrt{m+1}(M_n-\frac12)$.

(e) For The Oblig you all showed that $\sqrt{n}(M_n-\frac12)\to_d \mathrm{N}(0,1/4)$, by working out an expression for its density and its limit. Now deduce this result from the above.

Exercise 2: exponential things

THE HUMAN MIND IS REALLY BAD AT THINKING about exponential things (says Baiju Bhatt), but let's try. Below you may find use for the facts that if $X \sim \chi^2_k$, then

$$\mathrm{E}\,X = k, \quad \mathrm{Var}\,X = 2k, \quad \mathrm{E}\,(1/X) = \frac{1}{k-2}, \qquad \mathrm{E}\,(1/X^2) = \frac{1}{(k-2)(k-4)},$$

the two last holding provided $k > 2$ and $k > 4$, respectively.

(a) Consider i.i.d. observations $Y_1, \ldots, Y_n$ from the Expo(θ) distribution, i.e. with density $\theta \exp(-\theta y)$ for y positive. Find the score function for this model, along with its mean and variance. Let's also honour Calyampudi Radhakrishna Rao (who died in August 2023, nearly 103 years old) by finding the Cramér–Rao lower bound for variances of unbiased estimators of θ.

(b) Write down the log-likelihood function $\ell_n(\theta)$, and show that the maximum likelihood estimator is $\widehat\theta = 1/\bar Y$, with as usual $\bar Y = (1/n)\sum_{i=1}^n Y_i$. Show via general maximum likelihood theory, without needing to re-think or re-do the mathematical details, that $\sqrt{n}(\widehat\theta - \theta_0) \to_d \mathrm{N}(0, \theta_0^2)$, as sample size increases, with θ_0 the true parameter value.

(c) It is easy to show that $2\theta Y_i \sim \chi^2_2$ (and you do not need to show it here). Use this to show that $\widehat\theta \sim \theta\, 2n/\chi^2_{2n}$. Find the mean and variance of $\widehat\theta$, and comment on how this compares with the Cramér–Rao lower bound.

(d) Show that the random interval $[\widehat\theta(1-1.96/\sqrt{n}), \widehat\theta(1+1.96/\sqrt{n})]$ covers the true θ with probability tending to 0.95.

Exercise 3: war and peace

WHY FUME'TH IN FIGHT: THE GENTILS SPITE / IN FURY RAGING STOUT?, sang Blindern Stunt- og Popupkor two days ago, at *Godt Hjort*. From the *Correlates of War* database I have extracted data (x_i, z_i) for the 96 major interstate wars, from 1823 to 2023, each with at least 1000 battle deaths. Here x_i is the onset date and z_i the battle death count for war i. Below is an attempt to address the famous question, 'Has the world become (somewhat) less brutal', translated here to assessments of the battle deaths numbers over time.

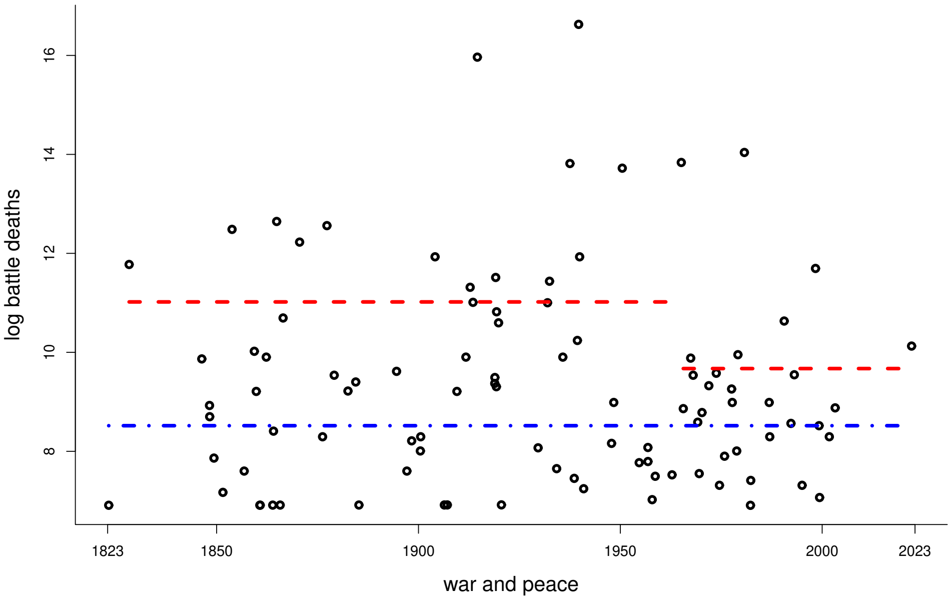
(a) There are theoretical reasons supporting the heavy-tailed power-law model for such data, formalised here by the assumption that

$$P(Z_i \geq z) = (z_0/z)^\theta \quad \text{for all } z \geq z_0,$$

for a suitably high threshold z_0, with the θ parameter dictating how quickly the tail goes to zero. Show that $Y_i = \log(Z_i/z_0)$, given $Z_i \geq z_0$, is Expo(θ). – Work by Cunen, Hjort, Nygård, in their 2020 paper *Statistical Sightings of Better Angels* in *Journal of Peace Research*, indicates that power-law behaviour is present for wars bigger than $z_0 = 5002$, which means that the $Y_i = \log(Z_i/z_0)$ being exponential is plausible for the $m = 56$ wars above that threshold. This is used below.

(b) The Cunen, Hjort, Nygård paper also argues that there is a changepoint in the battle deaths time series, with a before and up to Vietnam 1965 different from a somewhat less brutal world after Vietnam 1965, indicated with the lines in the figure below. With $\bar{Y}_L = 2.501$ the mean of the 38 Y_i before and up to Vietnam, and $\bar{Y}_R = 1.129$ the mean of the 18 Y_i after, use the setup of Exercise 2 (i) to find the maximum likelihood estimators $\widehat{\theta}_L$ and $\widehat{\theta}_R$, for the associated parameters θ_L and θ_R; and (ii) to write up 95 percent confidence intervals for these two parameters. Comment on what you find.

the Rus-Ukr war is not included in the analysis here, since one does not have clear data



The $(x_i, \log z_i)$ *for the 96 interstate wars over the last 200 years, those for which* $z_i \geq 1000$*. The 56 wars of size* $z_i \geq z_0 = 5002$*, those above the horizontal line at* $\log 5002$*, are assumed to follow the power-law distribution, with* $y_i = \log(z_i/z_0)$ *from the* $\mathrm{Expo}(\theta)$*. The lines before and up to Vietnam 1965 (38 wars), and after Vietnam 1965 (18 wars), are at* $\log z_0 + \bar{y}_L$ *and* $\log z_0 + \bar{y}_R$*.*

(c) In general terms, let $F_{a,b}$ indicate a random variable having the F distribution with degrees of freedom (a, b); this stems from $F = (\chi^2_a/a)/(\chi^2_b/b)$, with numerator and denumerator being independent. To assess the degree to which the θ parameter has changed, consider $\rho = \theta_R/\theta_L$. Using results of Exercise 2, show now that

$$\widehat{\rho} = \widehat{\theta}_R/\widehat{\theta}_L \sim \rho F_{2\tau,2(m-\tau)},$$

where in this case $\tau = 38$ is the number of wars to the left of and up to Vietnam and $m - \tau = 56 - 38 = 18$ the number of wars to the right. From my R files I copy over

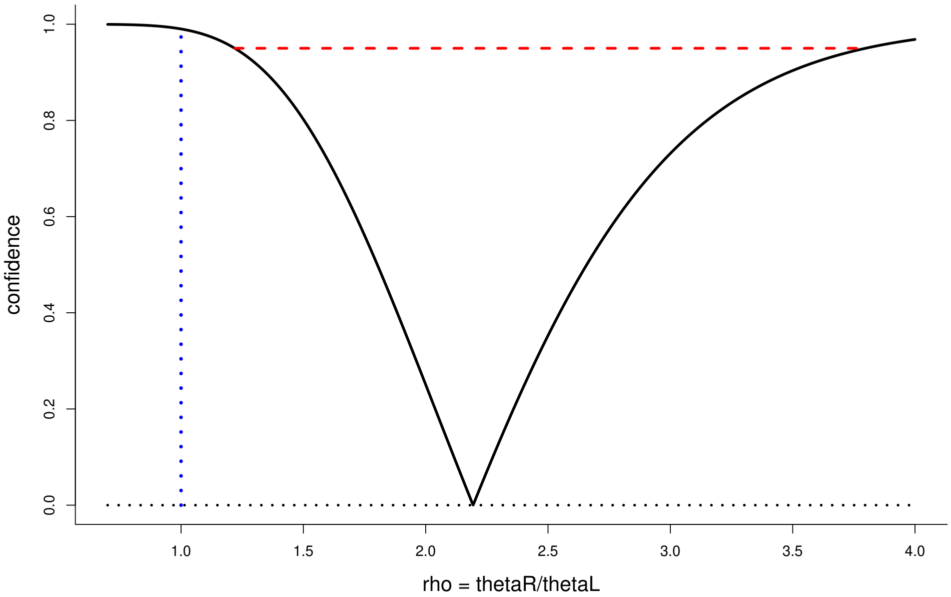
```
mm = 56
tau = 38
qf(0.99,2*tau,2*(mm-tau)) # is equal to 2.0402
```

Use this to test the hypothesis that $\theta_R = \theta_L$, against the alternative that $\theta_R > \theta_L$.

(d) The testing of $\theta_L = \theta_R$ above is already informative, but more may be done. Construct a full confidence distribution for the ρ parameter, via

$$C(\rho, \text{data}) = P_\rho(\widehat\rho \ge \widehat\rho_{\text{obs}}),$$

with your answer being given in terms of the c.d.f. for a F distribution. The figure below gives the confidence curve $\text{cc}(\rho, \text{data}) = |1 - 2\,C(\rho, \text{data})|$. Explain briefly what can be read off from this figure.



The confidence curve $\text{cc}(\rho, \text{data})$ *for the parameter* $\rho = \theta_R/\theta_L$*, with power-law parameters to the left of and to the right of Vietnam 1965. The horizontal line is at level 0.95.*

flower power against power-laws

(e) The above setup and analysis has implicitly taken for granted that Vietnam 1965 is a natural candidate for being a changepoint, viewing the underlying power-law parameters θ_i as a process over time. The assessments related to the change or not, and the size of that change, must be interpreted as 'given the information that the changepoint occurs in 1965'. The last question for this exercise concerns ways of estimating such a changepoint (which is then the start of further analysis). Suppose $Y_1, \ldots, Y_m$ are independent and exponentially distributed, with $Y_i \sim \text{Expo}(\theta_i)$, and furthermore that there is an unknown changepoint τ, for which we have $\theta_i = \theta_L$ for $i = 1, \ldots, \tau$ and $\theta_i = \theta_R$ for $i = \tau + 1, \ldots, m$. Find the log-likelihood function $\ell(\tau, \theta_L, \theta_R)$, and also an expression for the profiled

$$\ell_{\text{prof}}(\tau) = \max\{\ell(\tau, \theta_L, \theta_R)\colon \text{over all } \theta_L, \theta_R\}.$$

Explain how this may be used to estimate the changepoint position (which is how Cunen and Hjort discovered Vietnam).