UNIVERSITETET I OSLO

Matematisk Institutt

Exam in:	**STK 4180/9180 – Confidence Distributions**
	The project
With:	**Nils Lid Hjort**
Time for exam:	**8.–18.xii.2020**

This is the exam project set for STK 4180/9180, autumn semester 2020. It is made available on the course website as of *Tuesday 8 December 12:00*, and candidates must submit their written reports by *Friday 18 December 11:55* (or earlier), to the Inspera System with the Department of Mathematics. Reports may be written in nynorsk, bokmål, riksmål, English or Latin, and should be text-processed (TeX, LaTeX). Give your 'student 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, or perhaps briefly described, 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, each student needs to submit *one special extra page* with her or his report. This 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 project set contains four exercises and comprises five pages.

Exercise 1

This exercise looks into risk functions for and hence comparisons between confidence distributions, in simple prototype situations where calculations are easier than for general cases. We start out with $Y_1, \ldots, Y_n$ being i.i.d. from the $\mathrm{N}(\theta, 1)$ model. For a confidence distribution (CD) $C_n(\theta, y)$, where y denotes the full dataset, the risk function used is

$$\mathrm{risk}_n(C_n, \theta) = \mathrm{E}_\theta \int (\theta' - \theta)^2 \, \mathrm{d}C_n(\theta', Y) = \mathrm{E}_\theta(\theta_{\mathrm{cd}} - \theta)^2,$$

where θ_{cd} is the result of a two-stage random process: data Y lead to the CD $C_n(\theta, Y)$, and then θ_{cd} is drawn from this distribution.

(a) Show that the natural CD based on the observed sample mean $\bar{y}_{\mathrm{obs}} = n^{-1} \sum_{i=1}^n y_i$ is $C_n(\theta, y_{\mathrm{obs}}) = \Phi(\sqrt{n}(\theta - \bar{y}_{\mathrm{obs}}))$. Prove that its risk function is $\mathrm{risk}_n(C_n, \theta) = 2/n$.

(b) More generally, assume θ^* is some unbiased estimator of θ, with finite variance τ_n^2/n, with the property that $\widehat{\theta}^* - \theta$ has a distribution H_n symmetric around zero. Show that the associated CD becomes $C_n^*(\theta, y_{\mathrm{obs}}) = H_n(\theta - \theta^*_{\mathrm{obs}})$, and show that its risk function becomes $2\tau_n^2/n$. Exemplify with θ^* being the sample median.

(c) Relate the above results to the optimality theorem for CDs, in certain situations, from CLP's Chapter 5.

(d) Now we change gears a bit, by putting the a priori assumption $\theta \geq 0$ on the table. Show that the maximum likelihood estimator becomes $\widehat{\theta} = \max(0, \bar{y})$, i.e. the sample mean truncated, if necessary, to zero. Argue that this leads to the natural CD

$$\widetilde{C}_n(\theta, y) = \Phi(\sqrt{n}(\theta - \bar{y}_{\rm obs})), \quad \text{for } \theta \geq 0,$$

in particular having a positive point-mass at zero.

(e) Find an expression for the risk function for this $\widetilde{C}_n$, then compute and display it, for say $n = 10$. Comment on what we learn from this.

(f) There are various other estimators and CDs worth considering in this $\theta \geq 0$ setting. To simplify matters, take $n = 1$, and consider the Bayes estimator $\widehat{\theta}_B$, the conditional mean of $\theta \,|\, y$, with a flat prior on $(0, \infty)$. Show in fact that $\widehat{\theta}_B = y + \phi(y)/\Phi(y)$, and verify that this is positive even when y is negative. Work out an expression for the naturally associated CD $C_B(\theta) = \Pr_\theta\{\widehat{\theta}_B \geq \widehat{\theta}_{B,\rm obs}\}$, and comment.

Exercise 2

Here we consider a model sometimes called the truncated exponential model. We start with its simplest form, with data $Y_1, \ldots, Y_n$ i.i.d. from the density $\exp\{-(y-a)\}$ for $y \geq a$. The a is the unknown start point for the distribution.

(a) Show that the maximum likelihood estimator is equal to $U_n = \min_{i \leq n} Y_i$, the smallest data point. Show that $n(U_n - a)$ has a unit exponential distribution. Build from this a natural CD for a.

(b) Construct a predictive CD for the next sample point Y_{n+1}. Illustrate by computing and displaying the confidence curve for the text sample point, after having observed the six data points `3.735, 3.338, 10.634, 3.839, 5.667, 5.808`.

(c) Then consider the more realistic two-parameter version of the model, with density

$$f(y_i, a, b) = \frac{1}{b} \exp\Big(-\frac{y_i - a}{b}\Big) \quad \text{for } y_i \geq a,$$

with a being the unknown start-point and b a scale parameter. Show that the maximum likelihood estimators become

$$\widehat{a} = U_n \quad \text{and} \quad \widehat{b} = (1/n) \sum_{i=1}^{n} (Y_i - U_n),$$

again with U_n being the smallest observation.

(d) Construct accurate CDs and confidence curves for a, for b, and for the next datapoint Y_{n+1}. If some of your formulae cannot be given very explicit mathematical forms, this is ok, as long as numerical solutions can be found via numerical integration or simulation. Give approximations for these CDs for large sample sizes n.

(e) Ignoring these large-sample approximations, compute and display confidence curves for a, b, Y_{n+1} with the simple $n = 6$ dataset above.

Exercise 3

The lifelength distribution for a certain type of technical components is considered exponential, i.e. with density $\theta \exp(-\theta t)$ for $t > 0$, on a priori grounds. To arrive at a point estimate and a confidence curve for θ, the firm producing these components sets in motion the simple experiment where n such items are set to work, under controlled natural conditions. One cannot wait until all components have died out, however, and the firm needs to report what can be said about the lifelength distribution, via θ, a certain time t_0 after project start.

(a) With data of the form observed t_i for the N of the items which have died within t_0, and the information $t_i > t_0$ for the $n - N$ which are still alive and well, show that the combined likelihood function may be expressed as

$$\theta^N \exp\Big[-\theta\Big\{\sum_{t_i \le t_0} t_i + (n-N)t_0\Big\}\Big].$$

(b) Show that the maximum likelihood estimator is

$$\widehat{\theta} = \frac{N}{R} = \frac{N}{\sum_{t_i \le t_0} t_i + (n-N)t_0}.$$

With increasing sample size, and fixed t_0, find expressions for the probability limits of N/n and R/n, and show that $\widehat{\theta}$ is consistent.

(c) Show in fact that there is a limiting normal distribution here, with $\sqrt{n}(\widehat{\theta} - \theta) \to_d \mathrm{N}(0, \tau(t_0, \theta)^2)$, and attempt to find an explicit (though not necessarily quick and simple) formula for the limit variance.

(d) Explain why the construction $C_n(\theta) = \Pr_\theta\{\widehat{\theta} \ge \widehat{\theta}_{\rm obs}\}$ yields a CD, and also how it can be computed in practice.

(e) Suppose the experiment described involves $n = 20$ such items, and that the lifelengths for the $N = 11$ of these that conk out before the deadline of $t_0 = 2.00$ years are

```
0.528 0.743 0.869 1.180 0.602 0.133 0.327 1.115 0.117 0.208 1.808
```

Compute and display perhaps as many as three (exact or approximate) confidence curves for θ, for this little experiment: the one described in (c); one based on the normal approximation to the distribution of the maximum likelihood estimator; and a t-bootstrap based version. Comment on your findings.

Exercise 4

One is interested in a particular parameter γ, the 0.75-quantile of a certain distribution. There are three independent sources of knowledge, related to this γ. Your statistical job is (1) to convert these sources of information into separate confidence curves, say $\text{cc}_1(\gamma)$, $\text{cc}_2(\gamma)$, $\text{cc}_3(\gamma)$; and (2) then to combine these into a single fused $\text{cc}^*(\gamma)$.

The data sources can be summarised as follows. *Source 1* is a classical normal sample, of size $n_1 = 20$, with unbiased estimates $\bar{y}_1 = 5.342$ and $\widehat{\sigma}_1 = 2.179$, with $\gamma = \xi_1 + 0.675\,\sigma_1$ the 0.75 quantile in the underlying normal distribution. *Source 2* is a different sample, of size $n_2 = 12$, set up under different circumstances, and where normality cannot be trusted (though γ retains its interpretation as the 0.75 quantile of the underlying distribution in question). In fact the data are (in ordered fashion)

`1.159 2.453 3.320 4.160 4.168 5.287 5.343 5.826 6.084 6.632 7.150 12.044`

Here there is hence only 'nonparametric information' about γ. *Source 3* stems from perhaps earlier and perhaps vaguely characterised data, expressed via an experienced Bayesian statistician, who after proper pondering says her prior for γ is a $\text{N}(7.50, 1.25^2)$.

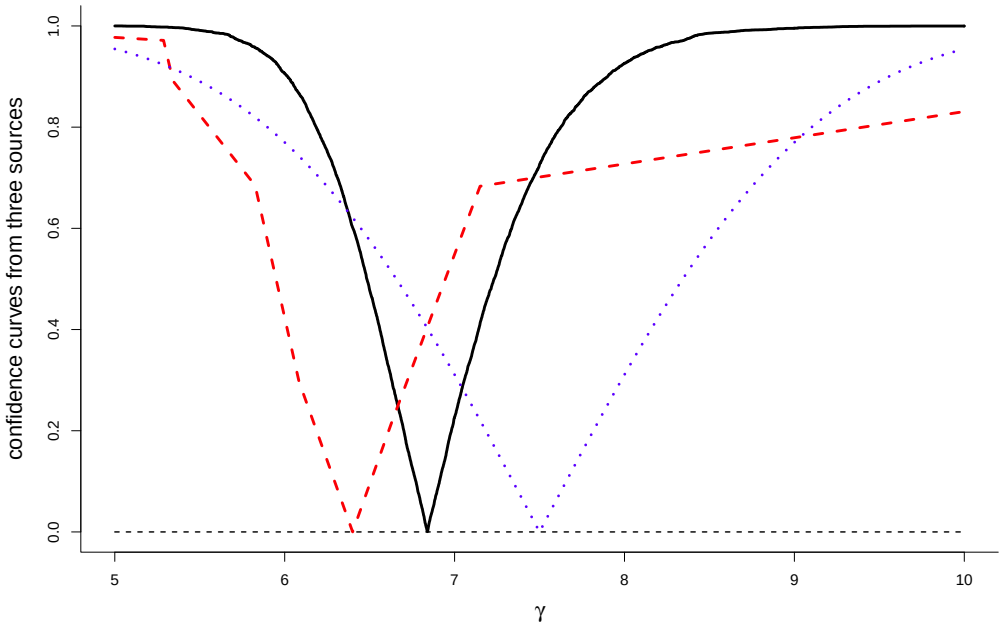

Figure A: Confidence curves from the three sources of information, for the same parameter $\gamma = F^{-1}(0.75)$ described in the text.

(a) Transform the pieces of information above to a construction of a version of Figure A, with the three confidence curves for γ. The most tricky of these is the nonparametric one, where you first might argue for the CD values $C_2(y_{2,(j)}) = 1 - G(0.75, j, n_2 - j + 1)$ at the observed ordered data $y_{2,(j)}$, where G is the relevant `pbeta` cumulative Beta distribution, and then use the convenient `approx` algorithm in `R`. You don't need to solve this just as I do it, but here are a few lines regarding the `approx` thing:

```
muval <- seq(5,10,by=0.01)
jj2 <- 1:nn2
C2valshort <- 1 - pbeta(0.75, jj2,nn2-jj2+1)
plot(yy2,C2valshort,type="p")
C2plusval <- approx(yy2,C2valshort,method="linear",xout=muval)$y
```

(b) Do the fusion. Comment on details and your findings.