UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4011 – Statistical Inference Theory**
WITH:	**Nils Lid Hjort**
AUXILIA:	**One single sheet of paper with the candidate's own personal hand-written notes**
TIME FOR EXAM:	**Tuesday 26/xi/2019, 14:30–18:30**

This exam set contains four exercises and comprises four pages.

Exercise 1

LIFE ITSELF IS EXPONENTIAL, says Jeff Rich. We say that Y has the exponential distribution with parameter α, and write $Y \sim \text{Expo}(\alpha)$ to indicate this, if its density is $\alpha \exp(-\alpha y)$ for $y > 0$. It is well known that the mean and variance become $1/\alpha$ and $1/\alpha^2$. For the following questions, you may encounter the partial sums

$$a_n = 1 + 1/2 + 1/3 + \cdots + 1/n,$$
$$b_n = 1 + 1/2^2 + 1/3^2 + \cdots + 1/n^2.$$

Here the first is slowly divergent, with $a_n \doteq \log n + 0.5772$, and the second is convergent, with $b_n \to \pi^2/6$, as shown by Euler in 1734, bringing him instant world fame.

(a) In a certain game of learning a player needs to complete tasks $1, 2, \ldots, n$, which become increasingly simpler with each passing of a new level. Assume that the time needed to complete these tasks are $V_1, \ldots, V_n$, with these being independent with $V_i \sim \text{Expo}(i/\theta)$, where θ is an unknown parameter. Find expressions for the mean and variance of $T_n = V_1 + \cdots + V_n$, the time it takes the player to complete all tasks. In particular, show that T_n has mean $a_n\theta$.

(b) Put up the unbiased estimator based on T_n, say $\widehat{\theta}$. Find its variance, and show that the estimator is consistent.

(c) Work out a formula for the log-likelihood function, based on having observed not merely the total time T_n, but the individual waiting times $V_1, \ldots, V_n$. Find the maximum likelihood estimator, say θ^*.

(d) Show that also this estimator θ^* is unbiased, and compare its variance to that of $\widehat{\theta}$.

(e) Find also the Cramér–Rao lower bound for variances of unbiased estimators for θ, and comment.

(f) Assume the game goes on, up to level $2n$, and consider the time a player needs to pass the last half of these levels, i.e. $T_n^* = T_{2n} - T_n$. Show that T_n^* tends in probability to a certain limit as n grows.

Exercise 2

THERE ARE 10 TYPES OF PEOPLE IN THE WORLD – those who understand binary models and those who do not. Consider the standard setup where X given the probability parameter p is a binomial (n, p). Below you may use that if V is a Beta distributed random variable, with parameters (a, b), i.e. with density

$$g(v) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} v^{a-1}(1-v)^{b-1} \quad \text{for } v \in (0,1),$$

then its mean and variance are

$$\mathrm{E}\, V = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}\, V = \frac{1}{a+b+1}\frac{a}{a+b}\frac{b}{a+b}.$$

(a) The classic estimator for p is of course $\widehat{p} = X/n$. Find its risk function, under squared error loss, i.e. the function

$$r(\widehat{p}, p) = \mathrm{E}_p\, (\widehat{p} - p)^2.$$

Where is the risk at its largest, and where at its smallest?

(b) Then do as Thomas Bayes did in his most famous publication (i.e. not *Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures*, 1731, but the other one), place a uniform prior distribution on the unknown p. What is the prior mean and prior variance? Show that if one observes that the random variable X is equal to x, then $p \,|\, x$ is a Beta distribution with parameters $(1 + x, 1 + n - x)$.

(c) Show that

$$\widehat{p}_B = \mathrm{E}\,(p \,|\, x) = \frac{x+1}{n+2},$$

and work out a formula for the risk estimator of this $(X+1)/(n+2)$ estimator.

(d) From a minimax perspective, which estimator is best, the X/n or $(X+1)/(n+2)$?

(e) You're a Bayesian finding yourself on a new planet, with no particular knowledge of the physics of the planet or its surroundings. You've observed that its sun has risen every morning, for each of the n consecutive days you've been there, which you in view of your ignorance of the astrophysics translate into meaning that with X the number of mornings the sun has risen, out of its n chances to do so, it must be a binomial (n, p), and that X has been observed to be n. What is then your updated belief about p?

(f) What is your probability that the sun will rise tomorrow morning, too?

Exercise 3

CORRELATION DOES NOT IMPLY CAUSATION, so here we start with something else before this causes you to work with correlations after all. Suppose $V_1, \ldots, V_5$ are independent normals with variance 1 and with means $\theta_1, \ldots, \theta_5$. A natural test for the null hypothesis that the five means are equal is based on $Z = \sum_{j=1}^5 (V_j - \bar{V})^2$, where $\bar{V}$ as usual is the average $(1/5)\sum_{j=1}^5 V_j$.

(a) With significance level 0.05, how large must Z be, before you reject the null hypothesis? You do not have access to chi-squared tables during exam hours (I believe), so describe your answer using the appropriate quantile language of things.

(b) To help explain why Z should become larger when the null hypothesis is not correct, compared to what can be reckoned with if it holds, work out a formula for the mean of Z, as a function of the five mean parameters.

– Consider now independent random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the binormal distribution, with mean parameters ξ_1, ξ_2, standard deviation parameter σ_1, σ_2, and correlation parameter ρ, assumed to lie safely inside $(-1, 1)$. Consider the sample correlation coefficient

$$R_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\}^{1/2}},$$

where $\bar{X}$ and $\bar{Y}$ are the means of the first and second components. The exact distribution for R_n can be worked out, but is rather complicated (and shall not be worked with here). But the large-sample limit distribution takes a reasonably simple form; one can with appropriate efforts (after exam hours) show that

$$\sqrt{n}(R_n - \rho) \to_d \mathrm{N}(0, (1-\rho^2)^2) \quad \text{as } n \to \infty.$$

(c) And now for the next question: use the delta method to show that with the transformation

$$h(\rho) = \tfrac{1}{2} \log \frac{1+\rho}{1-\rho},$$

we have $\sqrt{n}(h(R_n) - h(\rho)) \to_d \mathrm{N}(0, 1)$.

(d) For my $n = 100$ observed pairs I compute $R_n = 0.666$. Use the above to find a 95 percent confidence interval for ρ. (This is a no-calculator exam; if you should need $e^{0.666}$, or similar, just write it like this.)

(e) I have actually gathered such interesting correlation information for each of five groups, each based on sample size $n = 100$. I wish to test the null hypothesis H_0 that $\rho_1 = \cdots = \rho_5$. How can I do this?

Exercise 4

THE TRUE TRANSFORMATION TAKES PLACE WITHIN, my local psychologist claims – in the present exercise we shall transform from one thing to another but then perhaps transform partly back the other way again.

(a) We start with (X, Y) being independent standard normals, so that their joint density may be written

$$f_0(x, y) = \frac{1}{2\pi} \exp\{-\tfrac{1}{2}(x^2 + y^2)\}$$

for x, y on the real line. We then transform to so-called polar coordinates,

$$X = R\cos\theta \quad \text{and} \quad Y = R\sin\theta,$$

with $\theta \in [-\pi, \pi]$. Show that (R, θ) has the density

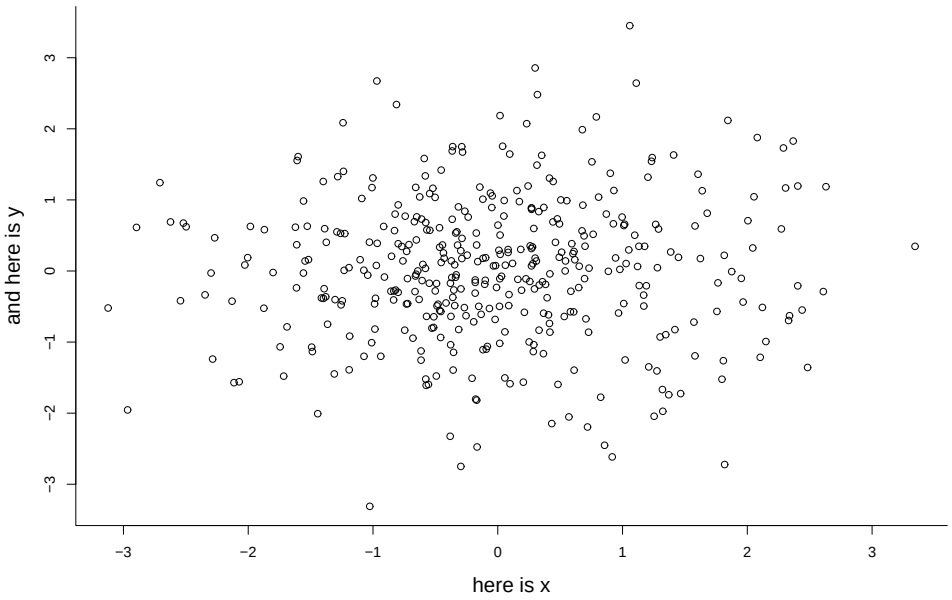
$$g_0(r, \theta) = h_0(r)\frac{1}{2\pi} = r\exp(-\tfrac{1}{2}r^2)\frac{1}{2\pi} \quad \text{for } r > 0 \text{ and } \theta \in [-\pi, \pi].$$

– The density $h_0(r) = r\exp(-\frac{1}{2}r^2)$ reached above is sometimes called the Raleigh distribution (after Lord Raleigh, who won the Nobel in physics in 1904), though a separate name might not really be needed in that it is simply a $\chi_2 = (\chi_2^2)^{1/2}$ – the density of the square-root of a chi-squared with two degrees of freedom. But the representation above opens the door for generalising the binormal distribution we started out with, by inventing a more general density than $h_0(r)$.

(b) Suppose the random radius R has density $h(r)$, rather than the $h_0(r)$, and keep θ uniform on $[-\pi, \pi]$, independent of R. Show that $(X, Y) = (R\cos\theta, R\sin\theta)$ then must have density

$$f(x, y) = h(\sqrt{x^2 + y^2})\frac{1}{\sqrt{x^2 + y^2}}\frac{1}{2\pi}.$$

You may find a need for the mathematical fact that the derivative of $A(u) = \arctan u$ is $A'(u) = 1/(1 + u^2)$.



(c) Consider one such generalisation, namely

$$h(r) = \tfrac{1}{2}\gamma r^{\gamma-1}\exp(-\tfrac{1}{2}r^\gamma) \quad \text{for } r > 0,$$

where $\gamma > 0$ is such an extra parameter. We let this be the new distribution for the random radius R, and again keep the random angle θ distributed independently and uniform on $[-\pi, \pi]$. In the figure above I have simulated $n = 400$ pairs from the associated distribution $f(x, y)$, with a certain value of γ known so far only to me. Explain how you can test whether the data are from the simple binormal density, i.e. from the $f_0(x, y)$ of (a), or not.