Statistiek (WISB263)

Retake Exam

April 18, 2018

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
(The exam is an *open–book* exam: notes and book are allowed. The scientific calculator is allowed as well).
The maximum number of points is 110 (10 bonus points!).
Grade=min(total points collected, 100).
Points distribution: 32-20-28-20-10

1. We assume that our data are sampled from a continuous random variable X with density function $f_X(x;\theta)$ given by:

$$f_X(x;\theta) := \begin{cases} \theta\, x\, e^{-\theta \frac{x^2}{2}} & x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta \in \Omega \equiv (0, \infty)$.

   (a) (6pt) We collect now a sample of size n of i.i.d. random variables distributed as X. Find a sufficient statistics for θ.

   (b) (8pt) Determine the maximum likelihood estimate of θ in case the collected sample $\mathbf{x}$ is:

$$\mathbf{x} = \{3, 1, 1, 2, 3, 5, 4, 4, 3, 4\}$$

   (c) (8pt) Give a general lower bound for the variance of an unbiased estimator of θ. Is the maximum likelihood estimator efficient?

   (d) (10pt) Determine now the maximum likelihood estimator of θ for a general sample of n i.i.d. random variables distributed as X, in case this time θ can only attain the values 1 and 2 (i.e. the parameter space is now $\Omega \equiv \{1, 2\}$).

2. Suppose that X_i are n i.i.d. normal random variables with mean μ and variance σ^2. We want to test the hypotheses $H_0 : \mu^2 = \sigma^2$ against $H_1 : \mu^2 \neq \sigma^2$.

   (a) (10pt) Find the generalized likelihood ratio statistic for testing the above hypotheses.

   (b) (10pt) What is the limiting distribution of log-likelihood ratio test statistic (under the null hypothesis)? Carefully explain your answer.

3. Mr. Thijs van Utrecht has a taxi company with 12 cabs. He is planning to buy 6 new tires of brand A and 6 new tires of brand B for the back wheels of the cabs. After every 500 km, he will check the wear of the tires. He can choose between the two following strategies:

   (1) put a single new back tire on each of the 12 cabs;

   (2) put a new back tire of each brand on 6 cabs.

   (a) [4pt] Which of the two strategies is preferable statistically? Try to justify your answer.

Mr. van Utrecht records now the following numbers of driven km when the 12 tires are worn:

km with brand A	51000	50500	61500	59000	64000	59000
km with brand B	55000	49500	62500	61500	65500	60000

(b) [10pt] In case these results are obtained using strategy (1), is there a significant difference between brand A and B? (take $\alpha = 0.1$). Try to justify the assumptions of the test used in the answer.

(c) [10pt] In case these results are obtained using strategy (2) and each column represents a different cab, is there a significant difference between brand A and B? (take $\alpha = 0.1$).

(d) [4pt] Comment on the results of points (b) and (c), try to find a statistical argument for the outcome.

4. A simple pendulum consists of a mass hanging at the end of a string of length ℓ. The period T of a pendulum is the time required for one complete cycle, that is, the time to go back and forth once. If the amplitude of motion of the swinging pendulum is small, then the pendulum behaves approximately as a simple harmonic oscillator, and the period T of the pendulum is given approximately by:

$$T = 2\pi\sqrt{\frac{\ell}{g}}$$

where g is the acceleration of gravity. From this expression, knowing ℓ, we can use measurements of T in order to estimate g.
Suppose now that we want to experimentally determine g by using the simple pendulum n times: we assume that we know exactly the lengths of the cords $\ell_1, \ldots, \ell_n$ and that we measure the periods of oscillations $T_1, \ldots, T_n$, so that we have the following observations:

$$(\ell_1, T_1), (\ell_2, T_2), \ldots, (\ell_n, T_n)$$

Furthermore, we assume that at each experiment we make small *measurements errors*, that can be reasonably modelled as realisations of i.i.d. random variables with zero expected value.

(a) [4pt] Describe a suitable *simple linear* regression model for estimating g.

(b) [6pt] Use the least squares principle in order to derive an estimator for g.

(c) [6pt] Determine the variance of the least square estimator for the regression parameter $2\pi/\sqrt{g}$ that represents the *slope*.

(d) [4pt] How would you choose the lengths of the cords in order to minimize the variance of point (c)?

BONUS Imagine that, for some reason, we want to estimate the number N of fishes in a small lake. We proceed as follow: we catch r fishes and *mark* them. We then release them back to the lake. Then, we wait some time and afterwards we catch n fishes (without putting them back). Let X_i be equal to 0 if the i–th fish we catch is *marked*, and 1 if it is not ($i \in \{1, \ldots, n\}$).

(a) [5pt] Determine the probability distribution of the random variable $Y := \sum_{i=1}^{n} X_i$, expressed in terms of N, r and n.

(b) [5pt] Find the maximum likelihood estimator of N (**Hint**: study the ratio $lik(N)/lik(N-1)$).