Mathematical Statistics **Exam 2017-12-21 at 8:00-13:00**
Centre for Mathematical Sciences **MASA01 B**
Lund University **Inference theory**

Complete solutions to all problems are required. Models should be specified, approximations and conclusions should be specified and motivated.

Problems 1, 2 and 3 can give a maximum of 10 points each. Problems 4, 5 and 6 can give a maximum of 20 points each. The maximum total on the exam is 90 points. The limit for Pass (P) is at least 45 points, and the limit for High Pass (HP) is at least 67 points.

Only the paper sheets provided by the institute in the exam room shall be used, both for solutions and for sketches/notes.

Each problem solution shall start at the top of a new paper.

Write your ID number on each paper that you hand in.

Red pencil or pen is not allowed.

Calculators are allowed. Only statistical tables provided by the institute are allowed. No other tables or formula sheets are allowed.

---

1. Biology students on a research trip have collected fish samples, of a certain species of fish, from a lake, for the purpose of studying the proportion of fish in the lake that are affected by a certain parasite. The lake is very large and the number of fish in the lake in question is also very large, and thus one can can with good approximation claim that when one samples one fish out of the lake, one does not change the proportion of fish that are affected by the parasite. The students sampled 132 fish and out of these 15 were affected by the parasite. Find the maximum likelihood estimate of the proportion of fish that are affected by the parasite. [10p]

2. One measures the acidity level in a lake several days in a row, before and after the construction of a thermonuclear power plant. The results of the measurements are (in unit pH)

$$\begin{aligned} \text{Levels before}: &\quad 9.32,\ 8.63,\ 6.22,\ 8.44,\ 7.02. \\ \text{Levels after}: &\quad 6.93,\ 6.62,\ 5.78,\ 6.90,\ 6.67,\ 7.11. \end{aligned}$$

   One can assume that the measurements of the acidity level (pH) are independent Gaussian, with the same variance but possibly different expectations, for the values before and after the construction. Test the null hypothesis of there being no difference in acidity level against the suspected hypothesis that the power plant has increased the acidity (i.e. has *lowered* the pH value) in the lake, on significance level 0.05. [10p]

3. The life lengths of electric components can to a good approximation be modeled as exponentially distributed with unknown expectation θ. In a very expensive experiment one has observed the life length of one electric component, and the observation is 6757 hours. One wants to make inference for the parameter θ. Construct a 95 % lower confidence interval for θ. [10p]

4. Suppose that X is $Po(\theta)$ distributed and suppose that we have one observation $x = 5$ of X.

   (a) Find the maximum likelihood estimator of the parameter θ. [5p]

(b) Perform an exact test, on significance level 0.05, of the hypotheses

$$\begin{aligned} H_0 : &\quad \theta \geq 25, \\ H_1 : &\quad \theta < 25. \end{aligned}$$

[5p]

(c) Construct instead an approximate test for the hypotheses above (motivate your approximation) on significance level 0.05. Calculate the (approximate) power of the constructed test at $\theta = 20$. [10p]

5. The r.v. X is continuous and has an unknown distribution function F_X. We want to make inference for the parameter $\gamma = E(e^{X^2})$. We have the following i.i.d. observations of X

$$0.96,\ 0.61,\ 0.61,\ 0.44,\ 0.087,\ 0.21,\ 0.20,\ 0.85,\ 0.63,\ 0.18,\ 0.42,\ 0.091,\ 0.58.$$

(a) Find the plug-in estimator $\hat{\gamma}$ of γ. [6p]

(b) Find an expression for the expectation and the variance of the estimator $\hat{\gamma}$. [6p]

(c) Construct a normal-based 95% (approximate) confidence interval for γ. If we would sample a new data sample, how large would n have to be in that sample in order for the width of the confidence interval to be 0.001 (you may assume the same value for the variance of $\hat{\gamma}$ as for the data you have in (a))? [8p]

6. A physicist wants to fit a linear model to data $(y_i, x_{1i}, x_{2i})), i = 1, \ldots, n$. The model the physicist wants to use is of the form

$$y_i = \alpha + \beta(x_{1i} - \bar{x}_{1\cdot}) + \gamma(x_{2i} - \bar{x}_{2\cdot}) + \epsilon_i,$$

for $i = 1, \ldots, n$, where one may assume that ϵ_i are i.i.d. $N(0, \sigma^2)$-distributed measurement errors. From previous experiments the physicist has concluded that the standard deviation for the measurement errors is $\sigma = 0.25$.

The physicist gathers $n = 100$ measurements and summarises the data, using the matrix formulation for the regression model, by

$$\begin{aligned} (X^tX)^{-1} &= \begin{pmatrix} 31.8733112 & -0.5720088 & -1.8509891 \\ -0.5720088 & 0.1081799 & -0.0017720 \\ -1.8509891 & -0.0017720 & 0.1200388 \end{pmatrix}, \\ X^tY &= \begin{pmatrix} 200.6948 \\ 1141.4747 \\ 3169.9606 \end{pmatrix}, \\ \bar{x}_{1\cdot} &= 5.54149, \\ \bar{x}_{2\cdot} &= 15.50173. \end{aligned}$$

(a) Construct individual 95 % confidence intervals for β and γ. [10p]

(b) A parameter of particular interest for the physicist is $\theta = \frac{1}{3}\alpha + \frac{1}{7}\gamma$. Construct a 95% confidence interval for θ. [10p]