**Midterm Solutions**
**MATH50011 Statistical Modelling 1**

1. Let $X_1, ..., X_n$ be a random sample, where $X_1$ has density $f_\theta(x) = \theta(x+1)^{-(\theta+1)}$ for $x \geq 0$ and unknown parameter $\theta > 0$. Denote by $F_\theta$ the cumulative distribution function of $X_1$. All the regularity conditions are satisfied in this case.

   (a) Show that $F_\theta(X_1) \sim Uniform(0,1)$. (2 marks)
   *Solution: For every $y \in (0,1)$ we have that*

   $$P(F_\theta(X_1) \leq y) = P(X_1 \leq F_\theta^{-1}(y)) = F_\theta(F_\theta^{-1}(y)) = y.$$

   *Hence, $F_\theta(X_1) \sim Uniform(0,1)$.*

   (b) Consider the random variable $Z = -2\log(1 - F_\theta(X_1))$. Show that $Z \sim \chi_2^2$. *(Hint: recall that the density of a $\chi_2^2$ is given by $f(z) = \frac{1}{2}e^{-\frac{z}{2}}$ for $z \geq 0$ and $f(z) = 0$ for $z < 0$)* (2 marks)
   *Solution:*

   $$P(Z \leq z) = P(-2\log(1 - F_\theta(X_1)) \leq z) = P(F_\theta(X_1) \leq 1 - e^{-\frac{z}{2}}) = 1 - e^{-\frac{z}{2}},$$

   *which is the distribution of a $\chi_2^2$.*

   (c) Using the random sample and the result in point (b) construct a 95% confidence interval for $\theta$. (For this question you do not need to write the critical values of the pivotal distribution explicitely). (4 marks)
   *Solution: First, we have that $-2\sum_{i=1}^n \log(1 - F_\theta(X_i)) \sim \chi_{2n}^2$. Moreover,*

   $$F(x) = \theta \int_0^x (t+1)^{-(\theta+1)} dt = \theta \int_1^{x+1} u^{-(\theta+1)} du = 1 - (x+1)^{-\theta}.$$

   *Hence, we have*

   $$-2\sum_{i=1}^n \log(1 - F_\theta(X_i)) = 2\theta \sum_{i=1}^n \log(X_i + 1) \sim \chi_{2n}^2.$$

   *Therefore, the (exact) 95% CI for $\theta$ is:*

   $$\left( \frac{k_{2n,0.025}}{2\sum_{i=1}^n \log(X_i+1)}, \frac{k_{2n,0.975}}{2\sum_{i=1}^n \log(X_i+1)} \right),$$

   *where $k_{2n,\alpha}$ indicate the critical values of $\chi_{2n}^2$, namely the $\alpha$ quantiles of $\chi_{2n}^2$.*

   (d) Compute the MLE for $\theta$. Is the computed MLE an unbiased estimator of $\theta$? (4 marks)
   *Solution: The likelihood and the log likelihood are is given by*

   $$L(\theta) = \theta^n \prod_{i=1}^n (x_i+1)^{-(\theta+1)}, \quad \text{and} \quad \ell(\theta) = n\log(\theta) - (\theta+1)\sum_{i=1}^n \log(x_i+1)$$

   *and so the first and second derivative of the log likelihood are given by*

   $$\frac{\partial}{\partial\theta}\ell(\theta) = \frac{n}{\theta} - \sum_{i=1}^n \log(x_i+1), \quad \text{and} \quad \frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\frac{n}{\theta^2}$$

*By equating the first derivative to zero and by observing that the second derivative is always strictly negative (recall the parameter space is $(0, \infty)$) we obtain that the MLE is given by*

$$\hat{\theta} = \frac{n}{\sum_{i=1}^n \log(X_i + 1)}.$$

*Using that the regularity conditions hold we obtain that $E_\theta[\frac{\partial}{\partial\theta} l(\theta, X)] = 0$, here $X$ stands for $(X_1, ..., X_n)$. This is a result we have seen in the lectures and is obtained as follows*

$$E_\theta[\frac{\partial}{\partial\theta} l(\theta, X)] = E[\frac{f'_\theta(X)}{f_\theta(X)}] = \int \frac{f'_\theta(x)}{f_\theta(x)} f_\theta(x) dx = \int f'_\theta(x) dx = \frac{\partial}{\partial\theta} \int f_\theta(x) dx = \frac{\partial}{\partial\theta} 1 = 0.$$

*Hence, we have that*

$$0 = E_\theta[\frac{\partial}{\partial\theta} l(\theta, X)] = E_\theta[\frac{n}{\theta} - \sum_{i=1}^n \log(X_i + 1)] = \frac{n}{\theta} - E_\theta[\sum_{i=1}^n \log(X_i + 1)],$$

*which implies that $E_\theta[\sum_{i=1}^n \log(X_i + 1)] = \frac{n}{\theta}$. Then, by Jensen inequality*

$$E_\theta[\hat{\theta}] = E_\theta\left[\frac{n}{\sum_{i=1}^n \log(X_i + 1)}\right] \neq \frac{n}{E_\theta[\sum_{i=1}^n \log(X_i + 1)]} = \theta.$$

*Hence, the MLE is biased.*

(e) Using the result in point (d) build an (approximate) rejection region for the $\alpha$ level test for $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$, for some $\theta_0 \in (0, \infty)$. (2 marks)

*Solution: Using the Fisher information identity, the Fisher information is given by $I(\theta) = -E[\frac{\partial^2}{\partial\theta^2} \ell(\theta)] = \frac{1}{\theta^2}$. Then, by the asymptotic normality of the MLE we have that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \theta_0^2)$ and by Slutsky lemma that $\sqrt{n}(\frac{\hat{\theta}}{\theta_0} - 1) \xrightarrow{d} N(0, 1)$. Thus, the approximate CI is*

$$\left(\frac{\hat{\theta}}{1 + \frac{c_{\alpha/2}}{\sqrt{n}}}, \frac{\hat{\theta}}{1 - \frac{c_{\alpha/2}}{\sqrt{n}}}\right), \text{ that is } \left(\frac{n}{(1 + \frac{c_{\alpha/2}}{\sqrt{n}}) \sum_{i=1}^n \log(x_i + 1)}, \frac{n}{(1 - \frac{c_{\alpha/2}}{\sqrt{n}}) \sum_{i=1}^n \log(x_i + 1)}\right)$$

*and so the approximate rejection region is*

$$\left\{(x_1, ..., x_n) \in [0, \infty)^n : \theta_0 \notin \left(\frac{n}{(1 + \frac{c_{\alpha/2}}{\sqrt{n}}) \sum_{i=1}^n \log(x_i + 1)}, \frac{n}{(1 - \frac{c_{\alpha/2}}{\sqrt{n}}) \sum_{i=1}^n \log(x_i + 1)}\right)\right\}.$$

*Alternatively, it is possible to use the consistency of the MLE to obtain that $\sqrt{n}(1 - \frac{\theta_0}{\hat{\theta}}) \xrightarrow{d} N(0, 1)$, which leads to the CI:*

$$\left(\hat{\theta}\Big(1 - \frac{c_{\alpha/2}}{\sqrt{n}}\Big), \hat{\theta}\Big(1 + \frac{c_{\alpha/2}}{\sqrt{n}}\Big)\right).$$

2. Let $X_1, ..., X_n$ and $Y_1, ..., Y_n$ be two random samples with $E[X_1] = E[Y_1] = 0$ and unknown variances. Assume that all moments exist.

   (a) Show that $\frac{1}{n} \sum_{i=1}^n X_i^2$ is an asymptotically normal estimator for $Var(X_1)$. (2 marks)

   *Solution: since $E[X_1] = 0$, we have $Var(X_1) = E[X_1^2]$. Then, by CLT we get that $\sqrt{n}(\frac{1}{n} \sum_{i=1}^n X_i^2 - Var(X_1)) \xrightarrow{d} N(0, Var(X_1^2))$.*

(b) Build an (approximate) rejection region for the $\alpha$ level test for $H_0 : (Var(X_1), Var(Y_1)) = (\theta_1, \theta_2)$ vs $H_1 : (Var(X_1), Var(Y_1)) \neq (\theta_1, \theta_2)$, for some $\theta_1, \theta_2 \in [0, \infty)$. (2 marks)

*Solution: Since* $Var(X_1^2) = E[X_1^4] - E[X_1^2]^2$, *a consistent estimator for* $Var(X_1^2)$ *is* $\frac{1}{n}\sum_{i=1}^n X_i^4 - (\frac{1}{n}\sum_{i=1}^n X_i^2)^2$, *which we denote by* $\hat{\tau}^2$. *Then, we have that*

$$\frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^n X_i^2 - Var(X_1))}{\hat{\tau}} \xrightarrow{d} N(0,1)$$

*and so the approximate* $1-\alpha$ *CI for* $Var(X_1)$ *is*

$$\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - \frac{c_{\alpha/2}\hat{\tau}}{\sqrt{n}}, \frac{1}{n}\sum_{i=1}^n X_i^2 + \frac{c_{\alpha/2}\hat{\tau}}{\sqrt{n}}\right),$$

*which we denote it by* $I_{\alpha/2}$. *The same applies to* $Var(Y_1)$ *and we denote its* $1-\alpha$ *CI by* $J_{\alpha/2}$. *By the Bonferroni correction we have that the* $1-\alpha$ *confidence region for* $(Var(X_1), Var(Y_1))$ *is* $I_{\alpha/4} \times J_{\alpha/4}$. *Thus, the approximate rejection region is*

$$\{(x_1, ..., x_n, y_1, ..., y_n) \in \mathbb{R}^{2n} : (\theta_1, \theta_2) \notin I_{\alpha/4} \times J_{\alpha/4}\}.$$

(c) How would your answer in point (b) change if we assume that $X_1, ..., X_n, Y_1, ..., Y_n$ are all independent? (2 marks)

*Solution: In this case the* $1-\alpha$ *confidence region for* $(Var(X_1), Var(Y_1))$ *is* $I_{\beta/2} \times J_{\beta/2}$, *where* $\beta$ *is such that* $(1-\beta)^2 = 1-\alpha$, *hence* $\beta = 1 - \sqrt{1-\alpha}$. *Thus, the rejection region becomes*

$$\{(x_1, ..., x_n, y_1, ..., y_n) \in \mathbb{R}^{2n} : (\theta_1, \theta_2) \notin I_{\beta/2} \times J_{\beta/2}\}.$$

*Since* $\beta > \alpha/2$, *the confidence region becomes smaller and so the rejection region becomes larger. Thus, we tend to reject the null hypothesis more often.*

(Total 20 marks)