**Course Statistical Inference SIF5084**

Faglig kontakt under eksamen:

Nikolai Ushakov 735 93539

Hjelpemidler: (C):

K.Rottman: Mathematische Formelsamlung,

Statistiske tabeller og formler, Tapir,

Godkjent kalkulator med tomt minne,

Et gult håndskrevet A4-ark, stemplet av instituttet.

**1.** A sample $X_1, ..., X_n$ is drawn from a uniform $(\theta, 2\theta)$ distribution. Consider two estimators of $\theta$:

$$T_1 = \frac{1}{2}\max\{X_1, ..., X_n\},$$

$$T_2 = \frac{2}{3}\bar{X} = \frac{2}{3n}(X_1 + ... + X_n).$$

a) Prove that both estimators are consistent.

b) Which one should be preffered and why?

**Solution.** Find MSE of both estimators (this will be used for both (a) and (b)). The distribution function of $X_i$ is

$$F_{X_i}(x) = \begin{cases} 0 & \text{for } x < \theta, \\ \frac{1}{\theta}x - 1 & \text{for } \theta \le x \le 2\theta, \\ 1 & \text{for } x > 2\theta, \end{cases}$$

therefore

$$F_{T_1}(x) = [F_{X_i}(2x)]^n = \begin{cases} 0 & \text{for } x < \theta/2, \\ \left(\frac{2}{\theta}x - 1\right)^n & \text{for } \theta/2 \le x \le \theta, \\ 1 & \text{for } x > \theta, \end{cases}$$

i.e. probability density function of $T_1$ is

$$f_{T_1}(x) = n\frac{2}{\theta}\left(\frac{2}{\theta}x - 1\right)^{n-1}$$

on the interval $[\theta/2, \theta]$ and zero outside this interval. Using this representation, we get

$$ET_1 = \int x f_{T_1}(x)dx = \frac{2n+1}{2(n+1)}\theta,$$

the bias:

$$b(T_1) = ET_1 - \theta = -\frac{\theta}{2(n+1)},$$

$$ET_1^2 = \int x^2 f_{T_1}(x)dx = \frac{2n^2+4n+1}{2(n+1)(n+2)}\theta^2,$$

$$Var(T_1) = ET_1^2 - (ET_1)^2 = \frac{n}{4(n+1)^2(n+2)}\theta^2,$$

$$MSE(T_1) = Var(T_1) + [b(T_1)]^2 = \frac{\theta^2}{2(n+1)(n+2)}.$$

For $T_2$ moments and MSE are found directly:

$$ET_2 = \frac{2}{3}E\bar{X} = \theta,$$

$$Var(T_2) = \frac{4}{9n}Var(X_i) = \frac{\theta^2}{27n} = MSE(T_2).$$

a) We get from the Chebychev inequality

$$P(|T_i - \theta| \geq \varepsilon) \leq \frac{E(T_i-\theta)^2}{\varepsilon^2} = \frac{MSE(T_i)}{\varepsilon^2} \to 0, \quad n \to \infty, \quad i = 1, 2,$$

which implies

$$T_i \xrightarrow{P} \theta, \quad n \to \infty, \quad i = 1, 2,$$

i.e. both estimators are consistent.

b)

$$MSE(T_1) = O\left(\frac{1}{n^2}\right), \quad n \to \infty,$$

$$MSE(T_2) = O\left(\frac{1}{n}\right), \quad n \to \infty,$$

therefore the first estimator is better when the sample size is large. If $n$ is small, then $MSE(T_1) > MSE(T_2)$ for $n \leq 10$ ($T_2$ is better) and $MSE(T_1) < MSE(T_2)$ for $n > 10$ ($T_1$ is better).

**2.** Let $X_1, ..., X_n$ be a sample taken from a gamma distribution with parameters $(2, 1/\theta)$ i.e. distribution with pdf

$$f(x;\theta) = \theta^2 x e^{-\theta x}, \ x > 0, \ \theta > 0.$$

a) Prove that this family of distributions has a monotone likelihood ratio.

b) Suppose that $n$ is large enough so that the Central Limit Theorem can be used. For testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$ find the acceptance region of the significance level $\alpha$ a UMP (uniformly most powerful) test.

c) Find the $(1-\alpha)$ one-sided confidence interval that results from inverting the test of part (b).

**Solution.** a) The likelihood function is

$$L(\theta; X) = \theta^{2n} e^{-\theta \sum X_i} \prod_{i=1}^{n} X_i,$$

therefore, if $\theta' < \theta''$, then the ratio

$$\frac{L(\theta'; X)}{L(\theta''; X)} = \left(\frac{\theta'}{\theta''}\right)^{2n} e^{(\theta''-\theta')\sum X_i}$$

is a monotone (increasing) function of $T(X) = \sum X_i$.

b) Due to part (a) the UMP test has form

$$\sum_{i=1}^{n} X_i < c \Longrightarrow H_1$$

where $c$ is determined from condition

$$P_{\theta_0}(\sum X_i < c) = \alpha.$$

To find $c$ let us use CLT. We have $EX_i = 2/\theta$, $Var(X_i) = 2/\theta^2$ therefore

$$\alpha = P_{\theta_0}(\sum X_i < c) = P_{\theta_0}\left(\frac{\sum X_i - 2n/\theta_0}{\sqrt{2n}/\theta_0} < \frac{c - 2n/\theta_0}{\sqrt{2n}/\theta_0}\right) \approx \Phi\left(\frac{c - 2n/\theta_0}{\sqrt{2n}/\theta_0}\right)$$

and

$$c = \frac{2n + z_\alpha \sqrt{2n}}{\theta_0}.$$

Thus the acceptance region has form

$$\sum_{i=1}^{n} X_i \geq \frac{2n + z_\alpha \sqrt{2n}}{\theta_0}$$

or

$$\bar{X} \geq \frac{2 + z_\alpha \sqrt{2}/\sqrt{n}}{\theta_0}.$$

c) Inverting the test of part (b) we obtain the following $(1-\alpha)$ one-sided confidence interval:

$$\left[\frac{2}{\bar{X}} + \frac{z_\alpha\sqrt{2}}{\bar{X}\sqrt{n}}, \infty\right).$$

**3.** A sample $X_1, ..., X_n$ is drawn from a beta distribution with parameters $\theta$ and $\theta + 5$ i. e. pdf of $X_i$ is

$$f(x;\theta) = \frac{\Gamma(2\theta+5)}{\Gamma(\theta)\Gamma(\theta+5)} x^{\theta-1}(1-x)^{\theta+4},\ 0 < x < 1,\ \theta > 0.$$

a) Find a one-dimensional sufficient statistic for $\theta$.

b) Find MME (the method of moments estimator) of $\theta$.

c) Show that MME differs from MLE (the maximum likelihood estimator) in the considered case. (Do not try to find MLE! Just use your result of part (a)).

**Solution.** a) The likelihood function is

$$L(\theta; X_1, ..., X_n) = \left[\frac{\Gamma(2\theta+5)}{\Gamma(\theta)\Gamma(\theta+5)}\right]^n \left[\prod_{i=1}^n X_i(1-X_i)\right]^{\theta-1} \left[\prod_{i=1}^n (1-X_i)\right]^5.$$

Put

$$T(X_1, ..., X_n) = \prod_{i=1}^n X_i(1-X_i),$$

$$g(T,\theta) = \left[\frac{\Gamma(2\theta+5)}{\Gamma(\theta)\Gamma(\theta+5)}\right]^n [T(X_1, ..., X_n)]^{\theta-1},$$

and

$$h(X_1, ..., X_n) = \left[\prod_{i=1}^n (1-X_i)\right]^5.$$

Then

$$L(\theta; X_1, ..., X_n) = g(T(X_1, ..., X_n), \theta) h(X_1, ..., X_n)$$

and hence, due to the factorization theorem, $T(X_1, ..., X_n)$ is a (univariate) sufficient statistic.

b)

$$\mu_1 = EX_i = \frac{\theta}{2\theta+5}$$

or

$$\theta = \frac{5\mu_1}{1-2\mu_1},$$

therefore

$$\hat{\theta}_{MME} = \frac{5\bar{X}}{1-2\bar{X}}.$$

c) $\hat{\theta}_{MLE}$ must be a function of any sufficient statistic, including in particular the sufficient statistic obtained in (a): $T(X) = \prod_{i=1}^n X_i(1-X_i)$. But $\hat{\theta}_{MME} = 5\bar{X}/(1-2\bar{X})$ is not a function of $T(X)$. Indeed, there exist two samples $X^{(1)}$ and $X^{(2)}$ such that

$$\frac{5\bar{X}^{(1)}}{1-2\bar{X}^{(1)}} \neq \frac{5\bar{X}^{(2)}}{1-2\bar{X}^{(2)}}$$

but

$$T(X^{(1)}) = T(X^{(2)}).$$

For example,

$$X_1^{(1)} = X_2^{(1)} = ... = X_n^{(1)} = 1/3$$

and

$$X_1^{(2)} = X_2^{(2)} = ... = X_n^{(2)} = 2/3.$$

**4.** Let $Y_1, ..., Y_n$ be observations. Consider the following two models:

$$Y_i = \tan(\beta x_i) + \epsilon_i \tag{1}$$

($\beta$ is the parameter to be estimated, $|\beta| \le 1$, $|x_i| < \pi/2$),

$$Y_i = \beta_0 + \beta_1^2 x_i + \epsilon_i \tag{2}$$

($\beta_0$ and $\beta_1$ are parameters to be estimated), where $\epsilon_1, ..., \epsilon_n$ are independent identically distributed random variables having normal $(0, \sigma^2)$ distribution ($\sigma^2$ is known).

a) Which of these two models is a generalized linear model and which is not? Why? What is the link function of the generalized linear model?

b) Suppose that $n$ is even: $n = 2m$, and $x_1 = ... = x_m = x \neq 0$, $x_{m+1} = ... = x_n = -x$. For model (1) find MLE (maximum likelihood estimator) of $\beta$.

**Solution.** a) Model (1) is a GLM, model (2) is not a GML (by definition). The link function in model (1) is $\arctan(\cdot)$.

b) We have

$$f_{Y_i}(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{y - \tan(\beta x_i)^2}{2\sigma^2}\right\}$$

hence the likelihood function is

$$L(\beta; Y) = (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{1}{2\sigma^2}\sum(Y_i - \tan(\beta x_i))^2\right\}$$

and

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^n (Y_i - \tan(\beta x_i))\frac{x_i}{\cos^2 \beta x_i} = \frac{x}{\sigma^2 \cos^2 \beta x}\left(-\sum_{i=1}^m Y_i + \sum_{i=m+1}^n Y_i + 2m\tan(\beta x)\right).$$

Solving equation

$$\frac{\partial \ln L}{\partial \beta} = 0$$

and taking into account that $|\beta| \leq 1$ we obtain

$$\hat{\beta}_{MLE}(x) = \begin{cases} b = \frac{1}{x} \arctan\left[\frac{1}{n}\left(\sum_{i=1}^{m} Y_i - \sum_{i=m+1}^{n} Y_i\right)\right] & \text{if } |b| \leq 1, \\ 1 & \text{if } b > 1, \\ -1 & \text{if } b < -1. \end{cases}$$