Statistics X_400004 Midterm 27.10.17 Solutions

Exercise 1 [28 points]

Let $X_1, \ldots, X_n$ be independent random variables from the uniform distribution on the interval $[\theta, 2\theta]$, where $\theta > 0$ is an unknown parameter.

(a) [3+6 points] Give the moment estimator $\widehat{\theta}_{\mathrm{MM}}$ and the maximum likelihood estimator $\widehat{\theta}_{\mathrm{ML}}$ of the parameter θ.

Since $\mathbb{E}X_1 = \dfrac{3\theta}{2}$, we have $\theta = \dfrac{2}{3}\mathbb{E}X_1$ and the moment estimator is given by $\widehat{\theta}_{\mathrm{MM}} = \dfrac{2}{3}\overline{X}$.

The density of the uniform distribution on $[\theta, 2\theta]$ is $p_\theta(x) = \dfrac{1}{\theta}$, $\theta \le x \le 2\theta$, hence the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} p_\theta(X_i) = \frac{1}{\theta^n}.$$

The data implies an additional restriction on the possible values of the parameter θ: since $X_i \in [\theta, 2\theta]$ for all i, we have $\theta \le X_{(1)}$ and $X_{(n)} \le 2\theta$, i.e. $\dfrac{X_{(n)}}{2} \le \theta \le X_{(1)}$. The maximum likelihood estimator is given by

$$\widehat{\theta}_{\mathrm{ML}} = \operatorname{argmax}_{\frac{X_{(n)}}{2} \le \theta \le X_{(1)}} L(\theta) = \operatorname{argmax}_{\frac{X_{(n)}}{2} \le \theta \le X_{(1)}} \frac{1}{\theta^n}.$$

Since $L(\theta) = \dfrac{1}{\theta^n}$ is decreasing in θ, it is maximized by the smallest possible value for θ. Hence, $\widehat{\theta}_{\mathrm{ML}} = \dfrac{X_{(n)}}{2}$.

(b) [3+6 points] Transform the estimators $\widehat{\theta}_{\mathrm{MM}}$ and $\widehat{\theta}_{\mathrm{ML}}$ that you found in part *(a)* into unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of the parameter θ. (If $\widehat{\theta}_{\mathrm{MM}}$ is already unbiased, just put $\widehat{\theta}_1 = \widehat{\theta}_{\mathrm{MM}}$. If $\widehat{\theta}_{\mathrm{ML}}$ is already unbiased, just put $\widehat{\theta}_2 = \widehat{\theta}_{\mathrm{ML}}$.)

Hint 1: You can use without derivation the fact that the density of $X_{(n)}$ is given by $p_{X_{(n)}}(x) = n(x-\theta)^{n-1}/\theta^n$, $\theta \le x \le 2\theta$.

We have

$$\mathbb{E}\widehat{\theta}_{\mathrm{MM}} = \frac{2}{3}\mathbb{E}\overline{X} = \frac{2}{3}\mathbb{E}X_1 = \frac{2}{3}\cdot\frac{3\theta}{2} = \theta.$$

That is, $\widehat{\theta}_{\mathrm{MM}}$ is an unbiased estimator for θ and we put $\widehat{\theta}_1 = \dfrac{2}{3}\overline{X}$.

By Hint 1 and integration by parts,

$$\begin{aligned}
\mathbb{E}\widehat{\theta}_{\mathrm{ML}} &= \frac{1}{2}\mathbb{E}X_{(n)} = \frac{1}{2}\int_\theta^{2\theta} x p_{X_{(n)}}(x)dx = \frac{1}{2\theta^n}\int_\theta^{2\theta} x d(x-\theta)^n \\
&= \frac{1}{2\theta^n}\left[x(x-\theta)^n\Big|_\theta^{2\theta} - \int_\theta^{2\theta}(x-\theta)^n dx\right] = \frac{1}{2\theta^n}\left[2\theta^{n+1} - \frac{(x-\theta)^{n+1}}{n+1}\Big|_\theta^{2\theta}\right] \\
&= \frac{1}{2\theta^n}\left[2\theta^{n+1} - \frac{\theta^{n+1}}{n+1}\right] = \theta\left[1 - \frac{1}{2(n+1)}\right] = \theta\frac{2n+1}{2(n+1)}.
\end{aligned}$$

That is, $\widehat{\theta}_{\mathrm{MM}}$ is biased. It can be transformed into an unbiased estimator by scaling: put

$$\widehat{\theta}_2 = \widehat{\theta}_{\mathrm{ML}}\frac{2(n+1)}{2n+1} = X_{(n)}\frac{n+1}{2n+1},$$

then

$$\mathbb{E}\widehat{\theta}_2 = \frac{2(n+1)}{2n+1}\mathbb{E}\widehat{\theta}_{\mathrm{ML}} = \frac{2(n+1)}{2n+1}\theta\frac{2n+1}{2(n+1)} = \theta.$$

That is, $\widehat{\theta}_2 = X_{(n)}\dfrac{n+1}{2n+1}$ is an unbiased estimator for θ.

(c) [7 points] Which of the unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ that you found in part *(b)* is better for large data sets? (That is, as $n \to \infty$. You do not have to specify for which n exactly one estimator is better than the other).

Hint 2: You can use without derivation the fact that $\text{Var}X_{(n)} = \dfrac{n\theta^2}{(n+1)^2(n+2)}$.

Hint 3: If you have no answer to part *(b)*, compare $\widehat{\theta}_1 = \dfrac{2}{3}\overline{X}$ and $\widehat{\theta}_2 = \dfrac{n+1}{2n+1}X_{(n)}$ under the assumption that these are unbiased estimators for θ.

Since $\widehat{\theta}_1$ is unbiased, we have

$$\text{MSE}(\widehat{\theta}_1) = \text{Var}\widehat{\theta}_1 = \frac{4}{9}\text{Var}\overline{X} = \frac{4}{9} \cdot \frac{\text{Var}X_1}{n} = \frac{4}{9n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{27n}.$$

Since $\widehat{\theta}_2$ is unbiased as well and by Hint 2,

$$\text{MSE}(\widehat{\theta}_2) = \text{Var}\widehat{\theta}_2 = \frac{(n+1)^2}{(2n+1)^2}\text{Var}X_{(n)} = \frac{(n+1)^2}{(2n+1)^2} \cdot \frac{n\theta^2}{(n+1)^2(n+2)} = \frac{n\theta^2}{(2n+1)^2(n+2)}.$$

As $n \to \infty$, we have $\text{MSE}(\widehat{\theta}_2) < \text{MSE}(\widehat{\theta}_1)$, and hence the estimator $\widehat{\theta}_2$ is better for large data sets than $\widehat{\theta}_1$.

(d) [3 points] Assume now that the statistical model is parametrized differently: $X_1, \ldots, X_n$ are from the uniform distribution on the interval $[\frac{1}{\sqrt{a}}, \frac{2}{\sqrt{a}}]$, where $a > 0$ is an unknown parameter. Construct an estimator for the parameter a based on one of the estimators you constructed for the parameter θ in part *(a)*.

Put $\theta = \dfrac{1}{\sqrt{a}}$, then $a = \dfrac{1}{\theta^2}$. From part *(a)*, the moment and maximum likelihood estimators for θ are $\widehat{\theta}_{\text{MM}} = \dfrac{2}{3}\overline{X}$ and $\widehat{\theta}_{\text{ML}} = \dfrac{X_{(n)}}{2}$. Since $a = \dfrac{1}{\theta^2}$, the moment estimator for a is, by definition,

$$\widehat{a}_{\text{MM}} = \frac{1}{(\widehat{\theta}_{\text{MM}})^2} = \frac{9}{4(\overline{X})^2}.$$

Since $a = \dfrac{1}{\theta^2}$ is a 1-to-1 correspondence between $a > 0$ and $\theta > 0$, the maximum likelihood estimator for a is

$$\widehat{a}_{\text{ML}} = \frac{1}{(\widehat{\theta}_{\text{ML}})^2} = \frac{4}{(X_{(n)})^2}.$$

Exercise 2 [7 points]
Let $X_1, \dots, X_n$ be independent random variables from the following discrete distribution on $\{1, 2\}$:

$$p_\theta(k) := \mathbb{P}\{X_i = k\} = \frac{2}{2+\theta}\frac{\theta^{k-1}}{k}, \quad k = 1, 2,$$

where θ an unknown parameter. Give the Bayes estimator for θ under the prior belief about θ given by the density

$$\pi(\theta) = c\theta^a(2+\theta)^n e^{-\lambda\theta}, \quad \theta > 0,$$

where $a > 0$, $\lambda > 0$ are known and c is the normalizing constant.

Hint 4: The posterior parameter distribution is one of the standard distributions.

The posterior density is proportional to

$$\begin{aligned} p_{\overline{\Theta}|X_1,\dots,X_n}(\theta) &\propto \pi(\theta)\prod_{i=1}^n p_\theta(X_i) \propto \theta^a(2+\theta)^n e^{-\lambda\theta}\prod_{i=1}^n \frac{1}{2+\theta}\theta^{X_i-1} \\ &= \theta^a(2+\theta)^n e^{-\lambda\theta}\frac{1}{(2+\theta)^n}\theta^{\sum_{i=1}^n X_i - n} = \theta^{a+\sum_{i=1}^n X_i - n}e^{-\lambda\theta}, \quad \theta > 0. \end{aligned}$$

Hence, the posterior distribution is $\text{Gamma}(1 + a + \sum_{i=1}^n X_i - n, \lambda)$ and the Bayes estimator for θ is

$$\widehat{\theta}_{\mathrm{B}} = \mathbb{E}[\overline{\Theta}|X_1,\dots,X_n] = \mathbb{E}\text{Gamma}(1 + a + \sum_{i=1}^n X_i - n, \lambda) = \frac{1 + a + \sum_{i=1}^n X_i - n}{\lambda}.$$

Exercise 3 [15 points]
The number of cars on a highway per day in 2016 was normally distributed with expectation 4000 and standard deviation 50. The highway feels more congested in 2017 and there is a need to analyze statistically whether the number of cars on the highway has indeed increased in 2017 compared to 2016. The number of cars on the highway has been measured for 25 consecutive days in 2017. The observed numbers are given in the table below.

day 1	4009	day 6	4006	day 11	4009	day 16	4026	day 21	4021
day 2	4010	day 7	4005	day 12	4021	day 17	3998	day 22	4010
day 3	4011	day 8	4009	day 13	4019	day 18	4030	day 23	3999
day 4	3994	day 9	4025	day 14	4035	day 19	4025	day 24	4020
day 5	4000	day 10	4023	day 15	4030	day 20	4026	day 25	4021

In particular, the total number of cars observed in the 25 days is 100382.

(a) [2 points] Formulate a statistical model and a null and alternative hypotheses that are suitable for the analysis of the above problem.

Numbers of cars on different days are independent random variables $X_1, \dots, X_{25} \sim N(\mu, 50^2)$, where the parameter μ is unknown.
We have to test $H_0 : \mu \leq 4000$ VS $H_1 : \mu > 4000$.

(b) [4 points] Compute the p-value.

The observed sample mean is $\overline{X} = 100382/25 = 4015.28$. Hence,

$$p\text{-value} = \mathbb{P}\{N(4000, \frac{50^2}{25}) \geq 4015.28\} = \mathbb{P}\{N(0,1) \geq \frac{4015.28 - 4000}{50/5}\}$$
$$= \mathbb{P}\{N(0,1) \geq 1.53\} = 1 - \Phi(1.53) = 1 - 0.937 = 0.063.$$

(c) [2 points] Test the hypotheses that you formulated in part *(a)* so that the maximum chance of type-1 error is 2.5%.

The hypotheses should be tested at the significance level $\alpha_0 = 0.025$. Since p-value $= 0.063 > \alpha_0$, we fail to reject H_0.

(d) [7 points] For how many days should the highway be observed to ensure that, in case $\mu = 4100$, the Gauss test of significance level $\alpha_0 = 0.025$ rejects the wrong null hypothesis with probability of at least 95%?

By the standard test for $X_1, \dots, X_n \sim N(\mu, 50^2)$ and $H_0 : \mu \leq 4000$ VS $H_1 : \mu > 4000$, with significance level $\alpha_0 = 0.025$, the null hypothesis H_0 should be rejected if

$$\overline{X} \geq 4000 + z_{0.025}\frac{50}{\sqrt{n}} = 4000 + 1.96\frac{50}{\sqrt{n}} = 4000 + \frac{98}{\sqrt{n}}.$$

Since

$$\mathbb{P}_{\mu=4100}\{\text{reject } H_0\} = \mathbb{P}_{\mu=4100}\{\overline{X} \geq 4000 + \frac{98}{\sqrt{n}}\} = \mathbb{P}\{N(4100, \frac{50^2}{n}) \geq 4000 + \frac{98}{\sqrt{n}}\}$$
$$= \mathbb{P}\{N(0,1) \geq \frac{4000 + 98/\sqrt{n} - 4100}{50/\sqrt{n}}\} = \mathbb{P}\{N(0,1) \geq 1.96 - 2\sqrt{n}\},$$

we have $\mathbb{P}_{\mu=4100}\{\text{reject } H_0\} = \mathbb{P}\{N(0,1) \geq 1.96 - 2\sqrt{n}\} \geq 0.95$ if and only if $1.96 - 2\sqrt{n} \leq -z_{0.05} = -1.645$, which is if and only if $n \geq 4$.