Technische Universiteit Delft — Van Mourik Broekmanweg 6
Faculteit Elektrotechniek, Wiskunde en Informatica — 2628 XE Delft

Exam Statistical Inference (WI4455)
January 24, 2019, 9.00–12.00

Using books or notes is not allowed at the exam.
Unless stated diferently, always add an explanation to your answer.

1. We say that the random variable X has the Shifted Geometric distribution with parameter $\theta \in (0,1)$, denoted by $X \sim \mathrm{SGeo}(\theta)$, if it has probability mass function

$$f_X(x \mid \theta) = \mathbb{P}_\theta(X = x) = (1-\theta)^x \theta, \qquad x = 0, 1, \ldots.$$

We have $\mathbb{E}X = \theta/(1-\theta)$. Suppose $X_1, \ldots, X_n$ are independent and identically distributed with $X_i \sim \mathrm{SGeo}(\theta)$. For notational convenience, we set $\underline{X}_n = (X_1, \ldots, X_n)$.

   (a) Give an unbiased estimator for θ. That is, state the estimator and prove it is unbiased for θ.

   (b) Prove that $\sum_{i=1}^n X_i$ is minimal sufficient for θ.

   (c) Derive the maximum likelihood estimator for $\tau = \theta/(1-\theta)$.

   (d) Suppose $n = 1$ and we wish to test the hypothesis $H_0 : \theta = 1/2$ versus $H_1 : \theta = \theta_1$ with $\theta_1 < 1/2$. Show that the optimal test leads to rejecting H_0 for large values of X_1. In case we obtain the realisation $x = 2$, compute the p-value of the test with X_1 as test statistic.

   (e) Consider the following hierarchical model

$$\begin{aligned} X_1, \ldots, X_n \mid \Theta = \theta &\overset{\text{ind}}{\sim} \mathrm{SGeo}(\theta) \\ \Theta &\sim \mathrm{Be}(\alpha, \beta), \end{aligned}$$

   where $\mathrm{Be}(\alpha, \beta)$ denotes the Beta-distribution with parameters α and β. That is,

$$f_\Theta(\theta) = \frac{1}{B(\alpha,\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} \mathbf{1}_{[0,1]}(\theta).$$

   The mean of a random variable with the $\mathrm{Be}(\alpha, \beta)$-distribution is known to be $\alpha/(\alpha+\beta)$. Derive the posterior mean estimator for θ. If you wish, you can use "Bayesian notation" in deriving your answer.

   (f) For the model under (e), show that Jeffreys' prior satisfies

$$f_\Theta(\theta) \propto \left(\frac{\theta^3 + (1-\theta)^3}{(1-\theta)^3 \theta^2} \right)^{1/2}.$$

   (g) Suppose next, that we parametrise the model in terms of $\tau = \theta/(1-\theta)$. Then the first line of the hierarchical model under (e) would change to

$$X_1, \ldots, X_n \mid T = \tau \overset{\text{ind}}{\sim} \mathrm{SGeo}(\tau/(1+\tau)).$$

Suppose that we wish to derive Jeffreys' prior for T in this model. Explain how an expression for this prior can be obtained from the result under (f). Note that you are not asked to actually compute this prior (once you know how to do it, this computation is just tedious calculus).

(h) Consider the hierarchical model where, conditional on $\Theta_1, \dots, \Theta_n$, $X_1, \dots, X_n$ are (conditionally) independent and furthermore

$$\begin{aligned} X_i \mid \Theta_i = \theta_i &\overset{\text{ind}}{\sim} \text{SGeo}(\theta_i) \\ \Theta_1, \dots, \Theta_n &\overset{\text{iid}}{\sim} \text{Be}(\alpha, \beta). \end{aligned}$$

Show that

$$f_{(X_1,\dots,X_n)}(x_1, \dots, x_n; \alpha, \beta) = \prod_{i=1}^{n} \frac{B(\alpha+1, \beta+x_i)}{B(\alpha, \beta)}.$$

Explain how this result can be used for obtaining empirical Bayes estimates for α and β.

(i) Consider the hierarchical model where, conditional on $\Theta_1, \dots, \Theta_n$, $X_1, \dots, X_n$ are (conditionally) independent and furthermore

$$\begin{aligned} X_i \mid \Theta_i = \theta_i &\overset{\text{ind}}{\sim} \text{SGeo}(\theta_i) \\ \Theta_1, \dots, \Theta_n \mid A = \alpha &\overset{\text{iid}}{\sim} \text{Be}(\alpha, \alpha) \\ A &\sim \mathit{Exp}\,(1)\,. \end{aligned}$$

That is, we assume both hyperparameters are the same, and equip this common parameter with a standard exponential distribution, i.e. $f_A(\alpha) = e^{-\alpha}\mathbf{1}_{[0,\infty)}(\alpha)$. For this model, the posterior cannot be obtained in closed form. However, the Gibbs sampler can be used. Show that the update step for A boils down to drawing from a density $p(\alpha)$ that satisfies

$$p(\alpha) \propto \left(\frac{c}{e}\right)^{\alpha} (B(\alpha, \alpha))^{-n}\, \mathbf{1}_{[0,\infty)}(\alpha),$$

with $c = \prod_{i=1}^{n} (\theta_i(1-\theta_i))$.

2. The Cramer-Rao theorem is as follows: Suppose X has density $f_X(\cdot \mid \theta)$ with respect to the measure ν.

Assume $\Omega \subset \mathbb{R}$ and let $\varphi(X)$ be a one-dimensional statistic with $\mathbb{E}_\theta|\varphi(X)| < \infty$ for all θ. Suppose the FI regularity conditions are satisfied, $I(\theta; X) > 0$ and also that $\int \varphi(x) f_X(x \mid \theta) d\nu(x)$ can be differentiated under the integral sign with respect to θ. Then

$$\text{Var}_\theta\, \varphi(X) \geq \frac{\left(\frac{\mathrm{d}}{\mathrm{d}\,\theta}\, \text{E}_\theta\, \varphi(X)\right)^2}{I(\theta; X)}.$$

Give a proof of this result.

Hint: First prove that

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta}\, \text{E}_\theta\, \varphi(X) = \mathbb{E}_\theta\left[\varphi(X) s(\theta \mid X)\right],$$

where $s(\theta \mid x) = \frac{\mathrm{d}}{\mathrm{d}\,\theta} \log f_X(x \mid \theta)$. *Next, apply the Cauchy-Schwarz inequality.*

3. Suppose X has density

$$f_X(x \mid \theta) = \frac{\theta}{x^{\theta+1}} \mathbf{1}_{[1,\infty)}(x).$$

Assume $\theta > 1$. Then $\mathrm{E}_\theta X = \theta/(\theta - 1)$.

Suppose $\theta_1 > \theta_0 > 0$ and we wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. Define the decision rules $d_c(X)$ by

$$d_c(X) = \begin{cases} a_0 = \{\text{accept } H_0\} & \text{if } X > c \\ a_1 = \{\text{accept } H_1\} & \text{if } X \leq c \end{cases},$$

where $c > 1$.

(a) Incorrectly accepting H_0 is considered twice as costly as incorrectly accepting H_1. Write down a loss function that reflects this.

(b) Using this loss function, derive an expression for the risk-function of the rule $d_c(X)$, both for $\theta = \theta_0$ and for $\theta = \theta_1$. Your answer should only depend on θ_0, θ_1 and c.

(c) Suppose $\theta_0 = 2$ and $\theta_1 = 3$. If a priori H_0 and H_1 are considered equally likely, show that the Bayes decision rule corresponds to $c = 3$.

(d) Suppose $\theta_0 = 2$ and $\theta_1 = 3$. Show that the minimax rule corresponds to $c = 1/y$ where $y \in (0, 1)$ solves $y^3 + 0.5y^2 - 0.5 = 0$.

4. Give an example where the maximum likelihood estimator (mle) is inadmissible. That is, specify the statistical model, the mle and the loss function. Note that you do not need to *prove* inadmissibility.

Solutions

1. (a) As

$$\mathrm{E}_\theta\, \mathbf{1}_{\{X_i=0\}} = \mathrm{P}_\theta(\mathrm{X}=0) = \theta$$

we have that $\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i=0\}$ is unbiased for θ. More explicitly:

$$\mathrm{E}_\theta\left[\frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i=0\}\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}_\theta\, \mathbf{1}\{X_i=0\} = \mathrm{P}_\theta(\mathrm{X}_1=0) = \theta.$$

(b) We have $f_{X_n}(x_n \mid \theta) = (1-\theta)^{\sum_{i=1}^n x_i}\theta^n$. This shows that $\sum_{i=1}^n X_i$ is sufficient (factorisation theorem). Now suppose that

$$\frac{f_{X_n}(x_n \mid \theta)}{f_{X_n}(y_n \mid \theta)} = (1-\theta)^{\sum_{i=1}^n x_i - \sum_{i=1}^n y_i}$$

does not depend on θ. This implies that $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, which in turn implies that $\sum_{i=1}^n X_i$ is minimal sufficient.

(c) We first compute the MLE for θ. We have

$$\ell(\theta) = S\log(1-\theta) + n\log\theta,$$

where $S = \sum_{i=1}^n X_i$.

$$\ell'(\theta) = \frac{n(1-\theta)-\theta S}{\theta(1-\theta)}.$$

Setting $\ell'(\theta)=0$ gives $\theta = \frac{n}{n+S}$. It's easily verified that this stationary point indeed corresponds to a maximum. As $g\colon (0,1)\to\mathbb{R}$ with $g(\theta)=\theta/(1-\theta)$ is bijective, the MLE for τ is given by

$$g\left(\frac{n}{n+S}\right) = \frac{n}{S}.$$

(d) The optimal test is the Neyman-Peason test which rejects for large values of

$$\frac{L(\theta_1; X_1)}{L(1/2; X_1)} = \frac{(1-\theta_1)^{X_1}\theta_1}{(1-1/2)^{X_1}1/2} = \left(2(1-\theta_1)\right)^{X_1} 2\theta_1.$$

As $\theta_1 < 1/2$ we reject for large values of X. The p-value is given by

$$p = \mathrm{P}_{1/2}(\mathrm{X}\geq 2) = 1 - \mathrm{P}_{1/2}(\mathrm{X}=1) - \mathrm{P}_{1/2}(\mathrm{X}=0) = 1 - 1/2 - 1/4 = 1/4.$$

(e) We have

$$p(\theta \mid x_n) \propto \theta^{n+\alpha-1}(1-\theta)^{\beta+s-1},$$

with $s = \sum_{i=1}^n x_i$. Therefore, the posterior for θ has the $\mathrm{Be}(n+\alpha, \beta+s)$-distribution. This gives

$$\mathbb{E}[\Theta \mid X_n] = \frac{n+\alpha}{n+\alpha+\beta+\sum_{i=1}^n X_i}.$$

(f) We have

$$\ell''(\theta) = -\frac{S}{(1-\theta)^2} - \frac{n}{\theta^2}.$$

So

$$\begin{aligned} I(\theta; X_n) &= \frac{\mathrm{E}_\theta\, S}{(1-\theta)^2} + \frac{n}{\theta^2} \\ &= \frac{n\theta}{(1-\theta)^3} + \frac{n}{\theta^2} \propto \frac{\theta^3 + (1-\theta)^3}{(1-\theta)^3\theta^2} \end{aligned}$$

Now take the square root.

(g) Plug in $\theta = \tau/(1+\tau)$ and multiply with $\left|\frac{\mathrm{d}\,\theta}{\mathrm{d}\,\tau}\right|$. So

$$f_T(\tau) = f_\Theta\left(\frac{\tau}{1+\tau}\right)\left|\frac{\tau}{1+\tau}\right|.$$

(h)

$$\begin{aligned} f_X(x;\alpha,\beta) &= \int \prod_{i=1}^n \left[\frac{1}{B(\alpha,\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}(1-\theta_i)^{x_i}\theta_i\right]\mathrm{d}\,\theta_1\cdots\mathrm{d}\,\theta_n \\ &= \frac{1}{B(\alpha,\beta)^n}\prod_{i=1}^n \int \theta_i^\alpha (1-\theta_i)^{x_i+\beta-1}\mathrm{d}\,\theta_i \\ &= \prod_{i=1}^n \frac{B(\alpha+1,\beta+x_i)}{B(\alpha,\beta)} \end{aligned}$$

Empirical Bayes estimates can be obtained by maximising this expression with respect to (α,β).

(i) We need the "full conditional" of α, Now

$$p(\alpha \mid \text{all other rv}) \propto \frac{1}{B(\alpha,\alpha)^n}\left(\prod_{i=1}^n (\theta_i(1-\theta_i))^\alpha\right) e^{-\alpha}\mathbf{1}_{[0,\infty)}(\alpha).$$

This gives the required expression, since the term in brackets equals c^α.

2. We have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\,\theta}\mathrm{E}_\theta\,\varphi(X) &= \frac{\mathrm{d}}{\mathrm{d}\,\theta}\int \varphi(x) f_X(x\mid\theta)\mathrm{d}\,\nu(x) \\ &= \int \varphi(x)\frac{\mathrm{d}}{\mathrm{d}\,\theta} f_X(x\mid\theta)\mathrm{d}\,\nu(x) \\ &= \int \varphi(x) s(\theta\mid x) f_X(x\mid\theta)\mathrm{d}\,\nu(x) = \mathbb{E}_\theta\left[\varphi(X)s(\theta\mid X)\right], \end{aligned}$$

where $s(\theta \mid x) = \frac{\mathrm{d}}{\mathrm{d}\,\theta}\log f_X(x\mid\theta)$. Upon taking $\varphi \equiv 1$ we get $\mathrm{E}_\theta\, s(\theta\mid X) = 0$. Hence

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta}\mathrm{E}_\theta\,\varphi(X) = \mathrm{E}_\theta\left[(\varphi(X) - \mathrm{E}_\theta\,\varphi(X))s(\theta\mid X)\right].$$

The Cauchy-Schwarz inequality then gives

$$\left|\frac{\mathrm{d}}{\mathrm{d}\,\theta}\,\mathrm{E}_\theta\,\varphi(X)\right| \le \left(\mathrm{E}_\theta\left[(\varphi(X)-\mathrm{E}_\theta\,\varphi(X))^2\right]\mathrm{E}_\theta\left[s(\theta\mid X)^2\right]\right)^{1/2}.$$

The result now follows, since the square of the right-hand-side equals $\mathrm{Var}_\theta\varphi(X)\cdot I(\theta;X)$.

3. (a) There is no loss involved with a correct decision:

$$L(\theta_0,a_0)=L(\theta_1,a_1)=0.$$

We can further take $L(\theta_0,a_1)=1$. Then the loss of incorrectly accepting H_0, $L(\theta_1,a_0)$ can be taken equal to 2.

(b) We have

$$R(\theta,d_c(X))=\mathrm{E}_\theta\,L(\theta,d_c(X))=L(\theta,a_0)\mathrm{P}_\theta(\mathrm{d_c(X)=a_0})+\mathrm{L(\theta,a_1)P_\theta(d_c(X)=a_1)}$$

Hence

$$R(\theta_0,d_c(X))=L(\theta_0,a_1)\mathrm{P}_\theta(\mathrm{X}\le\mathrm{c})=1-\mathrm{c}^{-\theta_0}$$

and

$$R(\theta_1,d_c(X))=L(\theta_1,a_0)\mathrm{P}_\theta(\mathrm{X}>\mathrm{c})=2\mathrm{c}^{-\theta_1}.$$

(c) We need to find $c>1$ such that the Bayes risk is minimal. The Bayes risk is given by

$$\frac{1}{2}(1-c^{-\theta_0})+c^{-\theta_1}.$$

With $\theta_0=2$ and $\theta_1=3$ this becomes $\frac{1}{2}(1-c^{-2})+c^{-3}$. Define $y=c^{-1}$. Then equivalently we need to minimise

$$y\mapsto\frac{1}{2}(1-y^2)+y^3 \text{ over } y\in(0,1).$$

Taking the derivative with respect to y and equating to zero gives

$$-y+3y^2=y(3y-1)=0.$$

It is easily verified that $\hat{y}=1/3$ corresponds to a global minimum. Therefore, $c=1/\hat{y}=3$ gives the Bayes decision rule.

(d) By making a sketch, it is easy to see that

$$c\mapsto\max(0.5(1-c^{-2}),c^{-3})$$

has its minimum at the point where the two components cross. That is, $c>1$ for which

$$0.5(1-c^{-2})=c^{-3}.$$

Now set $c^{-1}=y$ and the result follows.

4. One can take Stein's example, where $X_i\overset{\text{ind}}{\sim}N(\theta_i,1)$, for $1\le i\le p$. Then the MLE for θ_i is simply X_i. Take the loss function

$$l(\theta,a)=\sum_{i=1}^{p}(\theta_i-a_i)^2.$$

Then for $p\ge3$ the MLE is inadmissible.