Stat 24400 Homework 5 Solution

Feb 18, 2016

(If you discover any errors, please notify byolkim@uchicago.edu.)

Total points: 100

1. **[10 pts]**

   (a) The likelihood $L(\theta \mid X)$ is

   $$L(\theta \mid X) = \binom{n}{X} \left(\frac{\theta}{1+\theta}\right)^X \left(1 - \frac{\theta}{1+\theta}\right)^{n-X},$$

   so the log-likelihood $\ell(\theta \mid X)$ is

   $$\ell(\theta \mid X) = X \log \theta - n \log(1+\theta) + \text{constant}.$$

   Differentiate with respect to θ, and set the derivative to zero to obtain

   $$\frac{d\ell}{d\theta} = \frac{X}{\theta} - \frac{n}{1+\theta} = 0.$$

   Provided that $X < n$, we have

   $$\hat{\theta} = \frac{X}{n-X}.$$

   In that case, $\hat{\theta}$ maximizes $L(\theta \mid X)$, because

   $$\left.\frac{d^2\ell}{d\theta^2}\right|_{\theta=\hat{\theta}} = -\frac{X}{\theta^2} + \left.\frac{n}{(1+\theta)^2}\right|_{\theta=\hat{\theta}} = -\frac{X}{\hat{\theta}^2} + \frac{n}{(1+\hat{\theta})^2} = -\frac{(n-X)^2}{X} + \frac{(n-X)^2}{n} < 0.$$

   When $X = n$,

   $$\frac{d\ell}{d\theta} = \frac{n}{\theta(1+\theta)} > 0 \quad \text{for all} \quad \theta \in (0, \infty)$$

   so that $\ell(\theta \mid X = n)$ is strictly increasing on $(0, \infty)$, the supremum is not attained, and the MLE does not exist.

   (b) Fisher's approximation theorem says that when n is large, $\hat{\theta}$ has an approximate $N(\theta, 1/I(\theta))$ distribution, where

   $$I(\theta) = -E\left[\frac{d^2}{d\theta} \log f(X \mid \theta)\right].$$

   Using the second derivative from part (a),

   $$\begin{aligned} I(\theta) = -E\left[\frac{d^2}{d\theta} \log f(X \mid \theta)\right] &= -E\left[-\frac{X}{\theta^2} + \frac{n}{(1+\theta)^2}\right] \\ &= \frac{EX}{\theta^2} - \frac{n}{(1+\theta)^2} = \frac{n\theta/(1+\theta)}{\theta^2} - \frac{1}{(1+\theta)^2} \\ &= \frac{n}{\theta(1+\theta)} - \frac{n}{(1+\theta)^2} = \frac{n}{\theta(1+\theta)^2}, \end{aligned}$$

so that

$$\frac{1}{I(\theta)} = \frac{\theta(1+\theta)^2}{n}.$$

Hence,

$$\hat{\theta} \text{ is approximately } N\left(\theta, \frac{\theta(1+\theta)^2}{n}\right) \text{ when } n \text{ is large.}$$

Grading Scheme: In part (a), 3 pts for the correct form of the MLE in the general case, 1 pt for the correct treatment of the special case when $X = n$, 1 pt for the check that the MLE maximizes the likelihood. In part (b), 4 pts for the accuracy of the steps, 1 pt for the correct form of the variance.

Remarks.

- In part (b), several people were confused about whether the variance should be divided by n as done in Prof Stigler's notes. I did not, at least not ostensibly, but Prof Stigler's formulation is also correct, if applied in the right way. The reason for this apparent discrepancy is that you can regard X both as the outcome of a *single* Binomial experiment or as the outcome of n independent Bernoulli experiments.
- Several people tried to verify their answers by deriving the variance using the more familiar parametrization p, the probability of success, and plugging in $p = \theta/(1+\theta)$ into the resulting expression. This will *not* result in the same expression as when the derivation is done with θ. This is because the Fisher information depends on the parametrization of the problem. In fact, if θ and η are two parametrizations of an estimation problem, and $\theta = \theta(\eta)$ is a continuously differentiable function of η, then

$$I_\eta(\eta) = I_\theta(\theta(\eta)) \left(\frac{d\theta}{d\eta}\right)^2.$$

- This is to convince people that the parametrization with θ is not as weird as it looks. Put $p = \theta/(1+\theta)$.

$$\frac{p}{1-p} = \frac{\theta/(1+\theta)}{1/(1+\theta)} = \theta.$$

θ is the *odds* of success, and this is fairly common in some contexts.

2. **[10 pts]**

Suppose $X \sim \text{Poisson}(\lambda_1)$, $Y \sim \text{Poisson}(\lambda_2)$, X and Y are independent. By direct

computation, the pmf of $Z = X + Y$ is

$$\begin{aligned}
\Pr(Z = n) = \Pr(X + Y = n) &= \sum_{k=0}^{n} \Pr(X = k) \Pr(Y = n - k) && \text{(independence)} \\
&= \sum_{k=0}^{n} \frac{\lambda_1^k e^{-\lambda_1}}{k!} \frac{\lambda_2^{n-k} e^{-\lambda_2}}{(n-k)!} \\
&= \frac{e^{-\lambda_1-\lambda_2}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} \\
&= \frac{(\lambda_1 + \lambda_2)^n e^{-(\lambda_1+\lambda_2)}}{n!}, && \text{(binomial theorem)}
\end{aligned}$$

which is the pmf of Poisson$(\lambda_1 + \lambda_2)$.

Alternatively, note that the moment-generating functions of X and Y are given by

$$M_X(t) = e^{\lambda_1(e^t-1)} \quad \text{and} \quad M_Y(t) = e^{\lambda_2(e^t-1)}.$$

This is because

$$E[e^{tX}] = \sum_{x=0}^{\infty} e^{tx} \cdot \Pr(X = x) = \sum_{x=0}^{\infty} e^{tx} \cdot \frac{\lambda_1^x e^{-\lambda_1}}{x!} = e^{-\lambda_1} \sum_{x=0}^{\infty} \frac{(\lambda_1 e^t)^x}{x!} = e^{-\lambda_1} e^{\lambda_1 e^t} = e^{\lambda_1(e^t-1)},$$

and likewise for Y. Now,

$$\begin{aligned}
M_Z(t) = M_{X+Y}(t) &= E[e^{t(X+Y)}] \\
&= E[e^{tX} e^{tY}] \\
&= E[e^{tX}] E[e^{tY}] && \text{(independence)} \\
&= e^{\lambda_1(e^t-1)} e^{\lambda_2(e^t-1)} \\
&= e^{(\lambda_1+\lambda_2)(e^t-1)},
\end{aligned}$$

which we recognize as the moment-generating function of Poisson$(\lambda_1 + \lambda_2)$.

Grading Scheme: Give partial credit as justified by partial progress.

3. **[20 pts]**

   (a) Let X_i denote the number of cavalry men kicked to death by horses in a single corps in a single year. The assumption is that

   $$\Pr(X_i = x_i) = \frac{\theta^{x_i} e^{-\theta}}{x_i!}.$$

   Since X_i's are independent, the likelihood $L(\theta \mid X_1, \ldots, X_n)$ is

   $$L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{\theta^{X_i} e^{-\theta}}{X_i!},$$

so that the log-likelihood $\ell(\theta \mid X_1, \ldots, X_n)$ is

$$\ell(\theta \mid X_1, \ldots, X_n) = \sum_{i=1}^{n} \Big(X_i \log\theta - \theta - \log X_i! \Big) = \log\theta \sum_{i=1}^{n} X_i - n\theta + \text{constant}.$$

Taking the derivative, and setting it to zero,

$$\frac{d\ell}{d\theta} = \frac{\sum_{i=1}^{n} X_i}{\theta} - n = 0 \quad \Longleftrightarrow \quad \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}.$$

We check that $\hat{\theta}$ indeed maximizes the likelihood:

$$\frac{d^2\ell}{d^2\theta} = -\frac{\sum_{i=1}^{n} X_i}{\theta^2} < 0.$$

(b) Since

$$E(\hat{\theta}) = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} EX_i = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta,$$

$\hat{\theta}$ is an unbiased estimator.

$$\begin{aligned} \text{Var}(\hat{\theta}) &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\,\text{Var}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\, X_i \qquad \text{(independence)} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \theta = \frac{\theta}{n}. \end{aligned}$$

We conclude

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) = \frac{\theta}{n}.$$

(c) Since $\hat{\theta}$ is the sample average of i.i.d. Poisson(θ) random variables, which have a finite variance, as a direct consequence of the central limit theorem,

$$\hat{\theta} \text{ is approximately } N(\theta, \theta/n) \text{ when } n \text{ is large.}$$

(d) Fisher's approximation theorem says that when n is large, $\hat{\theta}$ has an approximate $N(\theta, 1/nI(\theta))$ distribution, where

$$I(\theta) = -E\left[\frac{d^2}{d\theta}\log f(X_1 \mid \theta)\right].$$

Using the second derivative from part (a),

$$I(\theta) = -E\left[\frac{d^2}{d\theta}\log f(X_1 \mid \theta)\right] = -E\left[-\frac{X_1}{\theta^2}\right] = \frac{EX_1}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta},$$

so that

$$\frac{1}{nI(\theta)} = \frac{\theta}{n}.$$

Hence,

$$\hat{\theta} \text{ is approximately } N(\theta, \theta/n) \text{ when } n \text{ is large.}$$

We remark that this is the same distribution as that obtained in part (c).

(e)

$$\sum_{i=1}^{280} x_i = (0)(144) + (1)(91) + (2)(32) + (3)(11) + (4)(2) = 91 + 64 + 33 + 8 = 196.$$

Thus,

$$\hat{\theta} = \frac{196}{280} = \frac{7}{10}.$$

(f) By part (c) or (d), $\hat{\theta}$ has approximately $N(1, 1/280)$ distribution. Thus,

$$\Pr(\hat{\theta} < 0.85) = \Pr\left(\frac{\hat{\theta} - 1}{\sqrt{1/280}} < \frac{0.85 - 1}{\sqrt{1/280}}\right) \approx \Phi(-0.15\sqrt{280}) = 0.0060,$$

where Φ denotes the standard normal cdf.

Grading Scheme: 4 pts each for parts (a)-(d), 2 pts each for parts (e) and (f). For part (a), 3 pts for the correct derivation of the MLE, 1 pt for the check that the MLE maximizes the likelihood.

4. **[20 pts]**

Suppose X is observed. $L(\theta \mid X)$ is a line with slope $X/2$, so when $-1 \leq X < 0$, $L(\theta \mid X)$ is maximized at $\hat{\theta} = 0$, and when $0 < X \leq 1$, it is maximized at $\hat{\theta} = 1$. (When $X = 0$, $L(\theta \mid X)$ is a horizontal line, so $\hat{\theta}$ is any point in $[0, 1]$. But $\Pr(X = 0)$, and our derivation of the distribution of $\hat{\theta}$ should make it clear that we don't have to worry about this case.) So,

$$\hat{\theta} = \begin{cases} 1 & \text{if } X > 0, \\ 0 & \text{otherwise.} \end{cases}$$

This is a Bernoulli(p) random variable with

$$p = \Pr(X > 0) = \int_0^1 f(x \mid \theta)\, dx = \int_0^1 \frac{1 + x\theta}{2}\, dx = \left.\frac{(2 + x\theta)x}{4}\right|_0^1 = \frac{2 + \theta}{4}.$$

We have

$$E(\hat{\theta}) = p = \frac{2 + \theta}{4}.$$

This is zero if and only if $\theta = 2/3$, so $\hat{\theta}$ is unbiased if and only if $\theta = 2/3$. For other values of θ, $\hat{\theta}$ is biased. Since

$$\mathrm{Var}(\hat{\theta}) = p(1 - p) = \frac{2 + \theta}{4}\left(1 - \frac{2 + \theta}{4}\right) = \frac{(2 + \theta)(2 - \theta)}{16} = \frac{4 - \theta^2}{16}$$

and

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta = \frac{2+\theta}{4} - \theta = \frac{2-3\theta}{4},$$

we have

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) = \frac{4-\theta^2}{16} + \frac{(2-3\theta)^2}{16} = \frac{2-3\theta+2\theta^2}{4}.$$

Grading Scheme: 6 pts for the correct derivation of the MLE, 6 pts for the derivation of its distribution, 2 pts for noting when the MLE is unbiased, 3 pts each for the computation of the bias and the MSE.

5. **[20 pts]**

   (a) The likelihood $L(\theta \mid X)$ is

   $$L(\theta \mid X) = \frac{1}{\theta}\mathbb{1}\{0 < X \le \theta\}.$$

   As a function of θ, this is the reciprocal function restricted to $[X, \infty)$. It is strictly decreasing on $[X, \infty)$, and hence attains the maximum at $\hat{\theta} = X$.

   (b)

   $$\text{MSE}(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = E[(X-\theta)^2] = EX^2 - 2\theta EX + \theta^2.$$

   Now,

   $$EX = \int_0^\theta \frac{x}{\theta}\,dx = \frac{x^2}{2\theta}\bigg|_0^\theta = \frac{\theta}{2} \quad \text{and} \quad EX^2 = \int_0^\theta \frac{x^2}{\theta}\,dx = \frac{x^3}{3\theta}\bigg|_0^\theta = \frac{\theta^2}{3},$$

   so that

   $$\text{MSE}(\hat{\theta}) = \frac{\theta^2}{3} - 2\theta\cdot\frac{\theta}{2} + \theta^2 = \frac{\theta^2}{3}.$$

   (c) We seek a constant c such that $E[cX] = \theta$:

   $$E[cX] = c\,EX = \frac{c\,\theta}{2} = \theta \quad \iff \quad c = 2.$$

   Since $2X$ is an unbiased estimator of θ, its MSE is equal to its variance:

   $$\text{MSE}(2X) = \text{Var}(2X) = 4\,\text{Var}\,X = 4(EX^2 - (EX)^2) = 4\left(\frac{\theta^2}{3} - \frac{\theta^2}{4}\right) = 4\cdot\frac{\theta^2}{12} = \frac{\theta^2}{3}.$$

   (d) In general,

   $$\text{Bias}(cX) = \frac{c\,\theta}{2} - \theta = \frac{(c-2)\theta}{2} \quad \text{and} \quad \text{Var}(cX) = c^2\,\text{Var}\,X = \frac{c^2\theta^2}{12},$$

   so that

   $$\text{MSE}(cX) = \text{Var}(cX) + \text{Bias}^2(cX) = \frac{c^2\theta^2}{12} + \frac{(c-2)^2\theta^2}{4} = \frac{(c^2-3c+3)\theta^2}{3}.$$

   Now, $c^2 - 3c + 3$ is a quadratic passing through $(0, 3)$ with discriminant $3^2 - (4)(1)(3) = -3 < 0$, so it is minimized at $c = 3/2$. Since $\theta^2/3$ is a positive constant, the MSE, as a function of c, is minimized at $c = 3/2$ also.

Grading Scheme: 5 pts for each part.

6. **[20 pts]**

Let $X_1, \ldots, X_n$ be an i.i.d. sample from the double exponential distribution. The likelihood of the sample $L(\theta \mid X_1, \ldots, X_n)$ is given by

$$L(\theta \mid X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{2} e^{-|X_i - \theta|} = \frac{1}{2^n} e^{-\sum_{i=1}^{n} |X_i - \theta|},$$

and this is maximized when $S(\theta) := \sum_{i=1}^{n} |X_i - \theta|$ is minimized.

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ denote the *order statistics* of the sample. Note that $\tilde{X} = X_{(m+1)}$ is the median. We shall show that $\tilde{X}$ is the MLE by showing that $S(\theta) > S(\tilde{X})$ whenever $\theta \neq \tilde{X}$.

$\theta \neq \tilde{X}$ if and only if $\theta < \tilde{X}$ or $\theta > \tilde{X}$. We treat the $\theta < \tilde{X}$ case; the proof for the $\theta < \tilde{X}$ case is analogous.

Assume, in addition, that $\theta \geq X_{(1)}$. Then, $\theta \in [X_{(k)}, X_{(k+1)})$ for some $k \in \{1, \ldots, m\}$. (This step is for figuring out where to break up the sum into partial sums so that we don't have to deal with pesky absolute values.) Put

$$s_L(\theta) = \sum_{i=1}^{k} |X_{(i)} - \theta|, \; s_M(\theta) = \sum_{i=k+1}^{m+1} |X_{(i)} - \theta|, \; s_R(\theta) = \sum_{i=m+2}^{2m+1} |X_{(i)} - \theta|,$$

i.e. $s_L(\theta)$ is the partial sum over the data points satisfying $X_{(i)} \leq \theta < \tilde{X}$, $s_M(\theta)$ is the partial sum over the data points satisfying $\theta < X_{(i)} \leq \tilde{X}$, and $s_R(\theta)$ is the partial sum over the data points satisfying $\theta < \tilde{X} \leq X_{(i)}$. We use this to remove the absolute

values from the summands:

$$\begin{aligned}
s_L(\theta) &= \sum_{i=1}^{k} |X_{(i)} - \theta| = \sum_{i=1}^{k} (\theta - X_{(i)}) \\
&= \sum_{i=1}^{k} [(\tilde{X} - X_{(i)}) - (\tilde{X} - \theta)] \\
&= \sum_{i=1}^{k} |X_{(i)} - \tilde{X}| - \sum_{i=1}^{k} (\tilde{X} - \theta) \\
&= s_L(\tilde{X}) - k\,(\tilde{X} - \theta), \\
s_M(\theta) &= \sum_{i=k+1}^{m+1} |X_{(i)} - \theta| = \sum_{i=k+1}^{m+1} (X_{(i)} - \theta) \\
&= \sum_{i=k+1}^{m+1} [(\tilde{X} - \theta) - (\tilde{X} - X_{(i)})] \\
&= \sum_{i=k+1}^{m+1} (\tilde{X} - \theta) - \sum_{i=k+1}^{m+1} |X_{(i)} - \tilde{X}| \\
&= (m - k + 1)(\tilde{X} - \theta) - s_M(\tilde{X}), \\
s_R(\theta) &= \sum_{i=m+2}^{2m+1} |X_{(i)} - \theta| = \sum_{i=m+2}^{2m+1} (X_{(i)} - \theta) \\
&= \sum_{i=m+2}^{2m+1} [(X_{(i)} - \tilde{X}) + (\tilde{X} - \theta)] \\
&= \sum_{i=m+2}^{2m+1} |X_{(i)} - \tilde{X}| + \sum_{i=m+2}^{2m+1} (\tilde{X} - \theta) \\
&= s_R(\tilde{X}) + m\,(\tilde{X} - \theta).
\end{aligned}$$

Clearly,

$$S(\theta) = s_L(\theta) + s_M(\theta) + s_R(\theta).$$

Now,

$$\begin{aligned}
S(\theta) &= s_L(\theta) + s_M(\theta) + s_R(\theta) \\
&= s_L(\tilde{X}) - k\,(\tilde{X} - \theta) + (m - k + 1)(\tilde{X} - \theta) - s_M(\tilde{X}) + s_R(\tilde{X}) + m\,(\tilde{X} - \theta) \\
&= S(\tilde{X}) + (2(m - k) + 1)(\tilde{X} - \theta) - 2s_M(\tilde{X}).
\end{aligned}$$

We make two observations. First, $\tilde{X} = X_{(m+1)}$, so

$$s_M(\tilde{X}) = \sum_{i=k+1}^{m+1} |X_{(i)} - \tilde{X}| = \sum_{i=k+1}^{m} |X_{(i)} - \tilde{X}|.$$

Second, $\theta < X_{(i)}$ for $i = k+1, \ldots, m$, so

$$\tilde{X} - X_{(i)} < \tilde{X} - \theta \quad \Longrightarrow \quad s_M(\tilde{X}) = \sum_{i=k+1}^{m} |X_{(i)} - \tilde{X}| < (m-k)(\tilde{X} - \theta).$$

Therefore,

$$\begin{aligned} S(\theta) &= S(\tilde{X}) + (2(m-k)+1)(\tilde{X} - \theta) - 2s_M(\tilde{X}) \\ &> S(\tilde{X}) + (2(m-k)+1)(\tilde{X} - \theta) - 2(m-k)(\tilde{X} - \theta) \\ &= S(\tilde{X}) + (\tilde{X} - \theta) \\ &> S(\tilde{X}). \end{aligned}$$

When $\theta < X_{(1)}$, we only have two partial sums $s_M(\theta)$ and $s_R(\theta)$. The computations are similar, and hence omitted.

Grading Scheme: Be generous. Attempts illustrating most of the intuition should receive full credit. Suggested points division is 10 pts for graphs or other evidence that the student has understood the gist of the problem, 10 pts for the correct *idea* of the proof using the rank-ordered data.