CALD 10–602
Statistical Approaches to Learning and Discovery

Assignment 1

February 20, 2000

Sample Solutions

1. *Sufficient statistics.* A statistic $T(X_1, \ldots, X_n)$ is a *sufficient statistic* if $\theta \to T(\vec{X}) \to \vec{X}$ is a Markov chain for any distribution on the parameter θ, and a *minimal sufficient statistic* if $\theta \to T(\vec{X}) \to U(\vec{X}) \to \vec{X}$ is a Markov chain for every other sufficient statistic $U(\vec{X})$. Consider the following three statistics defined on real-valued random variables:

$$
\begin{aligned}
T_{\text{max}}(X_1, \ldots, X_n) &= \max(X_1, \ldots, X_n) \in \mathbb{R} \\
T_{\text{min,max}}(X_1, \ldots, X_n) &= (\min(X_1, \ldots, X_n), \max(X_1, \ldots, X_n)) \in \mathbb{R}^2 \\
T_{\text{min,med}}(X_1, \ldots, X_n) &= (\min(X_1, \ldots, X_n), \operatorname{median}(X_1, \ldots, X_n)) \in \mathbb{R}^2
\end{aligned}
$$

For each of the following models $\{f(x \,|\, \theta)\}_{\theta \in \mathbb{R}}$, which of the statistics T_{max},$T_{\text{min,max}}$, $T_{\text{min,med}}$ are sufficient? Which are minimal? Prove your answers are correct.

(a) $f(\cdot \,|\, \theta) = \text{Uniform}(0,\, \theta)$ with $\theta > 0$

(b) $f(\cdot \,|\, \theta) = \text{Uniform}(\theta_1,\, \theta_2)$ with $(\theta_1, \theta_2) \in \mathbb{R}^2$ and $\theta_2 > \theta_1$.

(c) $f(\cdot \,|\, \theta) = \text{Uniform}(\theta,\, \theta + 1)$ with $\theta > 0$.

Solution:

(a) The joint distribution is given by

$$
p(x_1, \ldots, x_n \,|\, \theta) = \begin{cases} \theta^{-n} & \text{if } 0 \le x_i \le \theta \\ 0 & \text{otherwise} \end{cases}
$$

Define

$$
v(T_{\text{max}}(X_1 \ldots, X_n), \theta) = \begin{cases} 1 & \text{if } T_{\text{max}}(X_1, \ldots, X_n) \le \theta \\ 0 & \text{otherwise} \end{cases}
$$

Then we have that

$$
p(X_1, \ldots, X_n \,|\, \theta) = \theta^{-n} v(T_{\text{max}}(X_1 \ldots, X_n), \theta)
$$

so that T_{max} is sufficient by the factorization theorem. Since T_{max} is a 1-dimensional statistic, it must also be minimal. Clearly $T_{\text{min,max}}$ is sufficient, but not minimal, since T_{max} is uniquely determined from it. Since $T_{\text{min,med}}$ does not uniquely determine T_{max}, it is not sufficient.

Note that n is treated as a parameter, or considered to be fixed. Otherwise the factorization theorem would not apply.

(b) In this case the joint distribution is factored as

$$p(X_1, \ldots, X_n \mid \theta) \;=\; (\theta_2 - \theta_1)^{-n} v(T_{\text{min,max}}(X_1 \ldots, X_n), \theta_1, \theta_2)$$

by defining

$$v(T_{\text{min,max}}(X_1, \ldots, X_n), \theta_1, \theta_2) \;=\; \begin{cases} 1 & \text{if } T_{\text{min}}(X_1, \ldots, X_n) \geq \theta_1 \text{ and } T_{\text{max}}(X_1, \ldots, X_n) \leq \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

Thus $T_{\text{min,max}}$ is sufficient. Since the distribution has parameter set $\mathbb{R}^2_+$, this statistic is also minimal. Thus T_{max} is not sufficient. Since $T_{\text{min,med}}$ does not uniquely determine T_{max}, it is not sufficient.

(c) For the case of $f(\cdot \mid \theta) = \text{Uniform}(\theta, \theta{+}1)$ with $\theta > 0$, the joint distribution is factored as

$$p(X_1, \ldots, X_n \mid \theta) \;=\; v(T_{\text{min,max}}(X_1 \ldots, X_n), \theta_1, \theta_2)$$

by defining

$$v(T_{\text{min,max}}(X_1, \ldots, X_n), \theta) \;=\; \begin{cases} 1 & \text{if } T_{\text{min}}(X_1, \ldots, X_n) \geq \theta \text{ and } T_{\text{max}}(X_1, \ldots, X_n) \leq \theta + 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus $T_{\text{min,max}}$ is sufficient and also minimal (though this is a little trickier to show). Neither T_{max} nor $T_{\text{min,med}}$ is sufficient.

2. *Numerical computation of the MLE.* (Tanner, problem 2.9.) 197 animals are distributed into four categories: $Y = (y_1,\ y_2,\ y_3,\ y_4)$ according to the ***genetic linkage model*** $((2+\theta)/4,\ (1-\theta)/4,\ (1-\theta)/4,\ \theta/4)$.

(a) What is the likelihood for the data $Y = (125, 18, 20, 34)$?

(b) What is the likelihood for the data $Y = (14, 0, 1, 5)$?

(c) Use the Newton-Raphson algorithm to obtain the MLE $\widehat{\theta}$ of θ for $Y = (125,\ 18,\ 20,\ 34)$. Try starting your algorithm at $\theta = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.6$ and 0.8. How did you assess the convergence of the algorithm?

(d) Repeat (c) for $Y = (14,\ 0,\ 1,\ 5)$. Did the algorithm converge for all of the starting points?

(e) Plot the normalized likelihood and the associated normal approximation in the same figure for the data $Y = (125,\ 18,\ 20,\ 34)$. Discuss the adequacy of the normal approximation.

(f) Repeat (e) for $Y = (14,\ 0,\ 1,\ 5)$.

Solution:

(a) The likelihood function is given by

$$L(\theta \mid Y) \;\propto\; (1+\theta)^{y_1}\,(1-\theta)^{y_2+y_3}\,\theta^{y_4}$$

so that

$$L(\theta \mid (125, 18, 20, 34)) \;\propto\; (1+\theta)^{125}\,(1-\theta)^{38}\,\theta^{34}$$

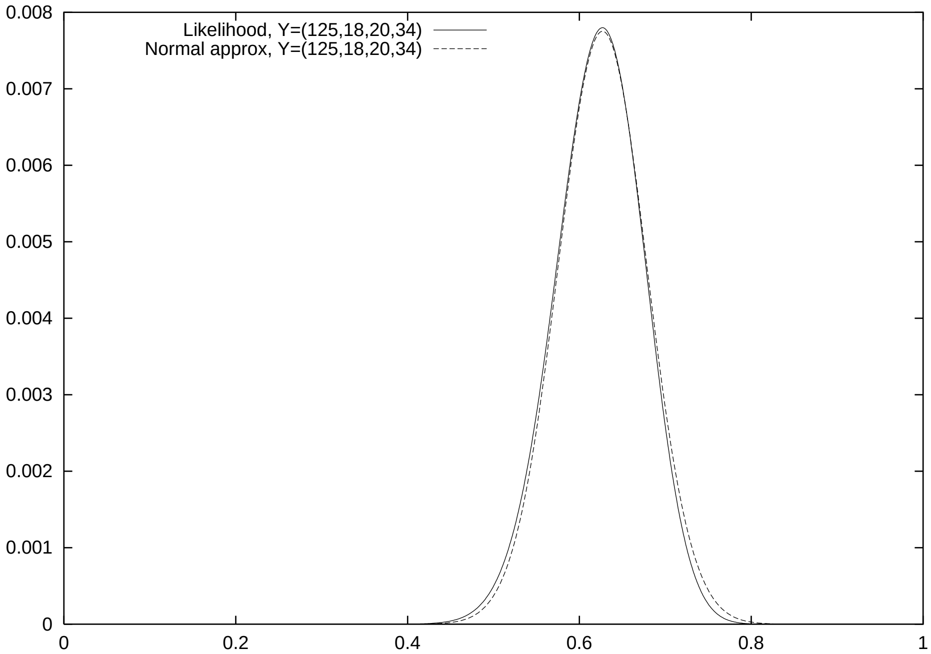
(b)

$$L(\theta \mid (14, 0, 1, 5)) \;\propto\; (1+\theta)^{14}\,(1-\theta)\,\theta^{5}$$

(c)

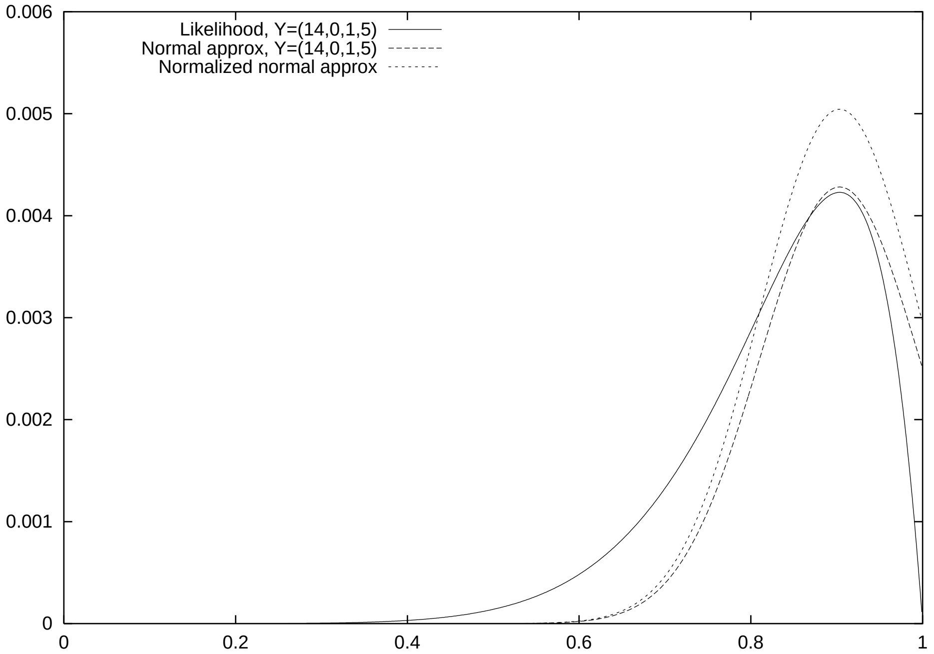
```
#include <math.h>
#include <stdio.h>
#include <stdlib.h>

/* Invariants of the following functions:
   theta in (0,1) */

/* Likelihood function. */
double f(int y1, int y2, int y3, int y4, double theta) {
  double l = pow((2+theta), y1) *
             pow((1-theta), y2) *
             pow((1-theta), y3) *
             pow((theta), y4);
  return l;
}

/* Derivative of the log-likelihood function.  */
double lp(int y1, int y2, int y3, int y4, double theta) {
  double l = y1/(2+theta) -
             y2/(1-theta) -
             y3/(1-theta) +
             y4/theta;
  return l;
}

/* Second derivative of the log-likelihood. */
double lpp(int y1, int y2, int y3, int y4, double theta) {
  double l = y1/pow(2+theta, 2) +
             y2/pow(1-theta, 2) +
             y3/pow(1-theta, 2) +
             y4/pow(theta, 2);
  return -l;
}

/* Run Newton's algorithm using the log-likelihood function
   defined above, with initial epsilon, until difference
   is less than epsilon */
double newton(int y1, int y2, int y3, int y4, double t0, double eps) {
  double theta=t0;
  double logprob;
  int iter=0;
```

```
  while (fabs(logprob=lp(y1, y2, y3, y4, theta)) > eps) {
    theta -= lp(y1, y2, y3, y4, theta) / lpp(y1, y2, y3, y4, theta);
  }
  return theta;
}
int main() {
  double theta0;
  double mode;
  mode=newton(14, 0, 1, 5, theta0=0.2, 1.0e-8);
  fprintf(stdout, "initial theta: %.4f  mode: %.4f\n", theta0, mode);
  return 0;
}
```



(e) For this data set the normal approximation is quite good. We made this and the next plot by generating the values numerically—interestingly, a package like Mathematica has problems with the large normalization constants required, even for this extremely simple statistical model.

(f) Here the normal approximation is poor. Note the difference when the normal approximation is normalized to be a density over the interval $\theta \in [0, 1]$.



3. *Biased estimates may be better.* Consider the problem of estimating μ and σ^2 from n samples of data drawn i.i.d. from a $\mathcal{N}(\mu, \sigma^2)$ distribution.

(a) Show that $\overline{X}$ is an unbiased estimator of μ.

(b) Show that the estimator

$$S_n^2 \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is biased and the estimator

$$S_{n-1}^2 \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is unbiased.

(c) Show that S_n^2 has a lower mean squared error than S_{n-1}^2. This illustrates the idea that a biased estimator may be "better" than an unbiased estimator.

Solution:

(a) Let $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then $E\overline{X_n} = \frac{1}{n}\sum EX_i = \mu$. Thus $\overline{X_n}$ is an unbiased estimator of μ.

(b) By the independence of the X_i's, we have

$$\mathrm{var}(\overline{X_n}) = \frac{1}{n^2}\sum_i \mathrm{var}(X_i) = \frac{\sigma^2}{n}.$$

Also,

$$
\begin{aligned}
E(X_i-\mu)(\overline{X_n}-\mu) &= E(X_i-\mu)\left(\frac{1}{n}\sum_j (X_j-\mu)\right)\\
&= \frac{1}{n}E(X_i-\mu)^2+\frac{1}{n}\sum_{j\neq i}E(X_i-\mu)(X_j-\mu)\\
&= \frac{\sigma^2}{n},
\end{aligned}
$$

since by independence, $E(X_i-\mu)(X_j-\mu)=0$.

Letting

$$W=\sum_i (X_i-\overline{X_n})^2=\sum_i \left((X_i-\mu)-(\overline{X_n}-\mu)\right)^2,$$

we have

$$
\begin{aligned}
EW &= \sum_i E(X_i-\mu)^2-2\sum_i E(X_i-\mu)(\overline{X_n}-\mu)+nE(\overline{X_n}-\mu)^2\\
&= n\sigma^2-2n\frac{\sigma^2}{n}+n\frac{\sigma^2}{n}\\
&= (n-1)\sigma^2
\end{aligned}
$$

Thus,

$$S_n^2=\frac{1}{n}\sum_{i=1}^n (X_i-\overline{X}_n)^2=\frac{W}{n}$$

has $ES_n^2=\frac{n-1}{n}\sigma^2$, and it is therefore a biased estimator of σ^2. However,

$$S_{n-1}^2=\frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X}_n)^2=\frac{W}{n-1}$$

has expected value σ^2 and is therefore an unbiased estimator of σ^2.

(c) We have that

$$
\begin{aligned}
E(S_{n-1}^2-\sigma^2)^2-E(S_n^2-\sigma^2)^2 &= E(S_{n-1}^4-2S_{n-1}^2\sigma^2+\sigma^4-S_n^4+2S_n^2\sigma^2-\sigma^4)\\
&= E(S_{n-1}^4-2S_{n-1}^2\sigma^2-\frac{(n-1)^2}{n^2}S_{n-1}^4+\frac{2(n-1)}{n}S_n^2\sigma^2\\
&= E(\frac{2n-1}{n^2}S_{n-1}^4-\frac{2}{n}S_{n-1}^2\sigma^2)\\
&= \frac{2n-1}{n^2}E(S_{n-1}^4)-\frac{2}{n}\sigma^2
\end{aligned}
$$

Now, using the fact (from Problem 4(a)) that

$$\mathrm{Var}(S_{n-1}^2)\geq\frac{1}{nJ(\sigma^2)}=\frac{2\sigma^4}{n}$$

we conclude that

$$\begin{aligned}
E(S_{n-1}^4) &= E^2(S_{n-1}^2) + \mathrm{Var}(S_{n-1}^2) \\
&= \sigma^4 + \mathrm{Var}(S_{n-1}^2) \\
&\geq \sigma^4 + \frac{2\sigma^4}{n}
\end{aligned}$$

Plugging into the above, we get

$$\begin{aligned}
E(S_{n-1}^2 - \sigma^2)^2 - E(S_n^2 - \sigma^2)^2 &\geq \frac{2n-1}{n^2}(\sigma^4 + \frac{2\sigma^4}{n}) - \frac{2}{n}\sigma^4 \\
&= \frac{3n-2}{n^3}\sigma^4 \\
&> 0
\end{aligned}$$

4. *Examples of Fisher information.* The Fisher information $J(\theta)$ for the family $f_\theta(x), \theta \in \mathbb{R}$ is defined by

$$J(\theta) = E_\theta\left(\frac{\partial f_\theta(X)/\partial\theta}{f_\theta(X)}\right)^2 = \int \frac{(\frac{\partial f_\theta(x)}{\partial\theta})^2}{f_\theta(x)}\, dx$$

Find the Fisher information for the following families:

(a) $f_\theta(x) = \mathcal{N}(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-x^2/2\theta}$.

(b) $f_\theta(x) = \theta e^{-\theta x}$, $x \geq 0$

(c) What is the Cramér-Rao lower bound on $E_\theta(\hat{\theta}(X) - \theta)^2$, where $\hat{\theta}(X)$ is an unbiased estimator of θ for (a) and (b)?

Solution:

(a) $f_\theta(x) = N(0,\theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}}$, and therefore

$$\frac{\partial f_\theta}{\partial\theta} = -\frac{1}{2}\frac{1}{\sqrt{2\pi\theta^3}} e^{-\frac{x^2}{2\theta}} + \frac{x^2}{2\theta^2}\frac{1}{\sqrt{2\pi\theta}} e^{-\frac{x^2}{2\theta}},$$

and

$$\frac{\frac{\partial f_\theta}{\partial\theta}}{f_\theta} = \left(-\frac{1}{2\theta} + \frac{x^2}{2\theta^2}\right).$$

Therefore the Fisher information is given by

$$\begin{aligned}
J(\theta) &= E_\theta\left(\frac{\frac{\partial f_\theta}{\partial\theta}}{f_\theta}\right)^2 \\
&= E_\theta\left(\frac{1}{4\theta^2} - 2\frac{1}{2\theta}\frac{x^2}{2\theta^2} + \frac{x^4}{4\theta^4}\right) \\
&= \frac{1}{4\theta^2} - 2\frac{1}{\theta}\frac{\theta}{2\theta^2} + \frac{3\theta^2}{4\theta^4} \\
&= \frac{1}{2\theta^2},
\end{aligned}$$

using the easily verified fact that a normal $\mathcal{N}(0,\theta)$ distribution has fourth moment $3\theta^2$.

(b) $f_\theta(x) = \theta e^{-\theta x}, x \geq 0$, and therefore $\ln f_\theta = \ln \theta - \theta x$, and

$$\frac{\partial \ln f_\theta}{\partial \theta} = \frac{1}{\theta} - x,$$

and therefore

$$\begin{aligned} J(\theta) &= E_\theta \left(\frac{\partial \ln f_\theta}{\partial \theta} \right)^2 \\ &= E_\theta \left(\frac{1}{\theta^2} - 2\frac{1}{\theta} x + x^2 \right) \\ &= \frac{1}{\theta^2} - 2\frac{1}{\theta}\frac{1}{\theta} + \frac{2}{\theta^2} \\ &= \frac{1}{\theta^2} \end{aligned}$$

using the fact that for an exponential distribution, $EX = \frac{1}{\theta}$, and $EX^2 = \frac{2}{\theta^2}$.

(c) The Cramer-Rao lower bound is the reciprocal of the Fisher information, and is therefore $2\theta^2$ and θ^2 for parts (a) and (b) respectively.

5. *Cramér-Rao inequality with a bias term.* Let $X \sim f(x \mid \theta)$ and let $T(X)$ be an estimator for θ. Let $b_T(\theta) = E_\theta(T) - \theta$ be the bias of the estimator. Show that

$$E_\theta (T - \theta)^2 \geq \frac{\left(1 + b_T'(\theta)\right)^2}{J(\theta)} + b_T^2(\theta)$$

Solution:

The proof parallels the proof without the bias term. We first calculate $E_\theta(VT)$, where V is the score function and T is the estimator.

$$\begin{aligned} E(VT) &= \int \frac{\frac{\partial}{\partial \theta} f(x|\theta)}{f(x|\theta)} T(x) f(x|\theta)\, dx \\ &= \int \frac{\partial}{\partial \theta} f(x|\theta) T(x)\, dx \\ &= \frac{\partial}{\partial \theta} \int f(x|\theta) T(x)\, dx \\ &= \frac{\partial}{\partial \theta} E_\theta T \\ &= \frac{\partial}{\partial \theta} (b_T(\theta) + \theta) \\ &= b_T'(\theta) + 1 \end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$\left(E\left[(V - EV)(T - ET) \right] \right)^2 \leq E(V - EV)^2 E(T - ET)^2.$$

Also, $EV = 0$ and therefore $E(V-EV)(T-ET) = E(VT)$, and by definition, $\mathrm{var}(V) = J(\theta)$. Thus we have

$$(b_T'(\theta) + 1)^2 \leq J(\theta) E(T - \theta - b_T(\theta))^2.$$

Finally,

$$\begin{aligned} E(T - \theta - b_T(\theta))^2 &= E(T-\theta)^2 + b_T^2(\theta) - 2E(T-\theta)b_T(\theta) \\ &= E(T-\theta)^2 + b_T^2(\theta) - 2b_T^2(\theta) \\ &= E(T-\theta)^2 - b_T^2(\theta). \end{aligned}$$

Substituting this in the Cauchy-Schwarz inequality, we get

$$E(T-\theta)^2 \geq \frac{[1 + b_T'(\theta)]^2}{J(\theta)} + b_T^2(\theta).$$