• If we now also want to estimate $\hat{\sigma}$ at the same
time, we want that pair $(\hat{\mu}, \hat{\sigma})$ that maximizes $L$.
• With $\mu = \bar{X}$, which $\sigma$ maximizes $L$?
  Setting $\frac{\partial}{\partial \sigma} \hat{\sigma} = 0$,
  $-\sum_{j=1}^{n} (\frac{1}{2}(X_j - \bar{X})^2\hat{\sigma}^{-3} \times -2 - n/\hat{\sigma} = 0$, or
  $\hat{\sigma} = \sqrt{\sum_{j=1}^{n} (X_j - \bar{X})^2/n}$.
iii. Exponential:
  • $f(\lambda; X) = -\lambda X + \ln(\lambda) \Rightarrow$ likelihood
  arising from an ind. sample $X_1, \cdots, X_n$ is
  $l(\lambda; X_1, \cdots, X_n) = -\lambda \sum_{j=1}^{n} X_j + n \ln(\lambda)$.
  • Setting the first derivative $= 0$,
    $-\sum_{j=1}^{n} X_j + n/\lambda = 0$, or
    $\hat{\lambda} = 1/(\sum_{j=1}^{n} X_j/n) = 1/\bar{X}$.
  • Do we have a maximum?
    $l''(\lambda; X_1, \cdots, X_n) = -n/\lambda^2$; always negative, and
    so $\hat{\lambda}$ is a global maximizer.
  • Recall that this is not an unbiased estimator; in
    fact, its expectation is infinite.
  • mean is $\mu = 1/\lambda$
  > Similar calculations say $\hat{\mu} = \bar{X}$.
iv. Harder m.l.e. example: Cauchy distr.
  Take $X_1, \cdots, X_n \sim$ Cauchy $\mu$;
  $l(\mu, X_1, \cdots, X_n) = -\sum_{j=1}^{n} \log(1 + (X_j - \mu)^2)$.
  • Likelihood equation is
    $-\sum_{j=1}^{n} \frac{\mu - X_j}{1 + (X_j - \mu)^2} = 0$.
  • See Fig. 6.
v. Uniform Example:
  • Invariance property: If $\tau = g(\theta)$, for $g$ onto, then
    $\hat{\tau} = g(\hat{\theta})$.
  • Often easier to consider this function’s log $l(\theta)$.
    i. $\theta$ shows up in the exponents of the normal,
      exponential, and Poisson dists, and
    ii. In the above-mentioned dists, and in the binomial
      distribution, for any value of $X$, $L(\theta) > 0 \forall \theta$
      (sound familiar)?
g. Relaxed definition:
  i. Since the log likelihood is concerned with relative
    comparisons of potential parameter values, we can
    eliminate any terms not containing $\theta$.
  ii. Hence we’ll also call a log-likelihood function to be
    that defined above, plus any function of the data
    not containing $\theta$.
WMS: 9.4

N. Sufficiency: How much of information do we have to
consider, and how much can we toss away as not giving
information about the quantity of interest?
1. Example:
  a. $X_1, \cdots, X_n \sim$ Bin$(m, \theta)$ an ind. sample.
  \begin{itemize}
  \item $X_1, \cdots, X_n \sim \mathcal{U}[0, \theta]$. 
  \item Product of densities is
    $$\prod_{i=1}^{n} \begin{cases} 1/\theta & \text{if } X_i \leq \theta \\ 0 & \text{otherwise} \end{cases} \prod_{i=1}^{n} \begin{cases} 1/\theta & \text{if } \theta \geq X_i \\ 0 & \text{otherwise} \end{cases}$$
  \end{itemize}
  e. Hence the joint p.m.f. is
    $$p_{X_1, \cdots, X_n}(x_1, \cdots, x_n; \tau) = \prod_{i=1}^{n} \begin{cases} m \pi^{x_i}(1-\pi)^{m-x_i} & \text{if } \tau = \theta \\ 0 & \text{otherwise} \end{cases} \prod_{i=1}^{n} \begin{cases} m \pi^{x_i}(1-\pi)^{m-x_i} & \text{if } \tau = \theta \\ 0 & \text{otherwise} \end{cases}$$

    \begin{align*}
    p(\hat{\mu}, \hat{\sigma}) &= \left( \frac{m}{n} \right) \pi^{\hat{\mu}} (1-\pi)^{m-n\hat{\sigma}}; \\
    \end{align*}

    Hence
    $$p_{X_1, \cdots, X_n}(x_1, \cdots, x_n; \hat{\theta}; \pi) = \prod_{i=1}^{n} \left( \frac{m}{\sum_{i}^{n} x_i} \right).$$

    Hence the additional information given by the $X_i$, after
    we know their total tells us nothing about $\pi$.
2. Definition: $T(X_1, \cdots, X_n)$ is sufficient for $\theta$ if the
distr of $X_1, \cdots, X_n$ conditional on $T$ doesn’t depend
O. Rao Blackwell Theorem:
Reduce the variance of an unbiased estimator by conditioning on a sufficient statistic.

1. Suppose
   a. $\hat{\theta}$ unbiased for $\theta$
   b. $U$ sufficient for $\theta$
2. Let $\hat{\theta} = E[\hat{\theta}|U]$.
   a. Then $\text{Var}[\hat{\theta}] = \text{Var}[E[\hat{\theta}|U]] + E[\text{Var}[\hat{\theta}|U]] \geq \text{Var}[\hat{\theta}]$.
3. Hence can find another estimator with often smaller variance.
4. Example: $X_1, \cdots, X_n \sim \text{U}[0, \theta]$.

5. Example $X, Y \sim P(\theta)$
   a. $\hat{\mu} = \frac{1}{4} X + \frac{2}{3} Y$
   i. $\hat{\mu} = \frac{1}{3} \Rightarrow X = 2$ and $Y = 0$ or $X = 0$ and $Y = 1$
   ii. $P [X = 2|\hat{\mu} = \frac{1}{3}] = \frac{\exp(-\mu)\mu^2/2!\exp(-\mu) + \exp(-\mu)\exp(-\mu)\mu^1/1!}{\mu^2 + 2\mu}$
   iii. does not depend on $\mu$; $\hat{\mu}$ not sufficient
   b. $\hat{\mu} = \frac{1}{3} X + \frac{2}{3} Y$
   i. $P [X = x|\hat{\mu} = u] = \frac{\exp(-2u\mu^2/2!\exp(-2u\mu^2)\mu^{2u-x}/(2u-x)!}{x!(2u-x)!}$
    ii. does not depend on $\mu$; sufficient
6. Hence entire data set $X_1, \cdots, X_n$ is sufficient.
   a. For independent data, so is ordered data set.
7. Example where sufficient statistic doesn’t tell the whole story:
   a. A collection of cars is inspected for defective wheels
   b. Estimate the proportion $\pi$ of wheels which are defective.
   c. Under the binomial model, the sample proportion is sufficient for inference on $\pi$.
   d. Consider two scenarios:

<table>
<thead>
<tr>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td># of wheels</td>
<td># of times</td>
</tr>
<tr>
<td>defective</td>
<td>observed</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>19</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
</tr>
</tbody>
</table>

   i. Both scenarios give the same estimate of $\pi$
   ii. The second case gives strong evidence that the binomial model is wrong.
   iii. This demonstrates that the sufficient statistic tells about the parameters in the model; remainder tells about the suitability of the model itself.

WMS: 9.5