MSF100
Statistical Inference Principles - Lecture 2

Rebecka Jörnsten
Mathematical Statistics
University of Gothenburg/Chalmers University of Technology

January 19, 2012

1 Likelihood

The likelihood is proportional to the joint pdf of the sample. However, while the pdf has a probabilistic interpretation (data generating given a parameter value θ), the likelihood does not. We think of the likelihood as a function of θ *given* an observed data sequence.

- $L(\theta|\tilde{x}) = f(\tilde{x}|\theta)$
- If iid, $L(\theta|\tilde{x}) = \prod_{i=1}^{n} f(x_i|\theta)$
- Gives evidence of the plausability of θ given a sample.

Likelihood principle: If $L(\theta|\tilde{x}) = C(\tilde{x}, \tilde{y})L(\theta|\tilde{y})$ then $\tilde{x}$ and $\tilde{y}$ carry the same evidence about θ. We don't care about the absolute likelihood, only its relative value for different θ.

We can focus our attention on the scaled quantity

$$\frac{L(\theta|\tilde{x})}{\max_\theta L(\theta|\tilde{x})}$$

Invariance principle: Let's say we make a change of variables transformation, $Y = g(X)$. The model family $\{f(x|\theta), \theta \in \Theta\}$ is invariant w.r.t. transformations G if for all $\theta \in \Theta$ and all $g \in G$ there is a θ' such that $Y = g(X)$ has distribution $f(y|\theta')$ if X has distribution $f(X|\theta)$.
Then the likelihood inference is unaffected by the transformation.
This means simple data transformations like change-of-units etc should not alter the analysis conclusions.

Let's look at a few examples of likelihood functions.
In Figure 1 I depict the likelihood function of a binomial sample with different sample size n and different outcomes x. The more data we have, the "sharper" the likelihood function is. What does this mean?

Well, the maximizer of the likelihood function is a natural estimate for θ if we think of $L(\theta|\tilde{x})$ as measuring the plausability of θ as being capable of generating sample $\tilde{x}$. However, the *shape* of $L(\theta|\tilde{x})$ is also important. The shape tells us how easy or difficult it is to identify the maximizing θ. As we shall see later, this carries over the estimation uncertainty for θ.

The shape of the likelihood is even more important to consider when we have more than one parameter $\theta = (\theta_1, \ldots, \theta_r)$.
In Figure 2 I depict two examples when θ is two-dimensional. In the left panel is an example where the model is almost over-specified. That is, the two parameters "control" the distribution in almost the same

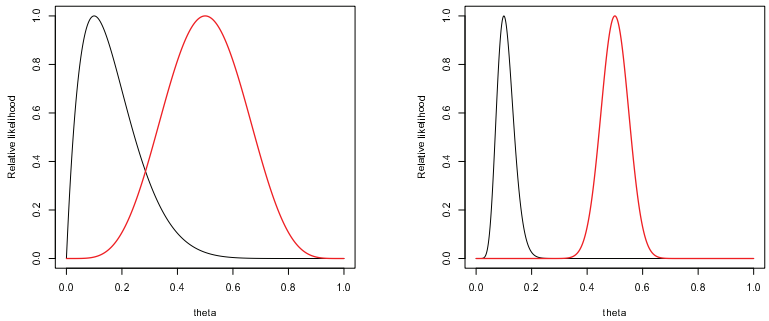

Figure 1: Likelihood functions for $X \sim Bin(n, \theta)$. (Left): $n = 10$ and $x = 1$ (black) and $x = 5$ (red). (Right): $n = 100$ and $x = 10$ (black) and $x = 50$ (red).

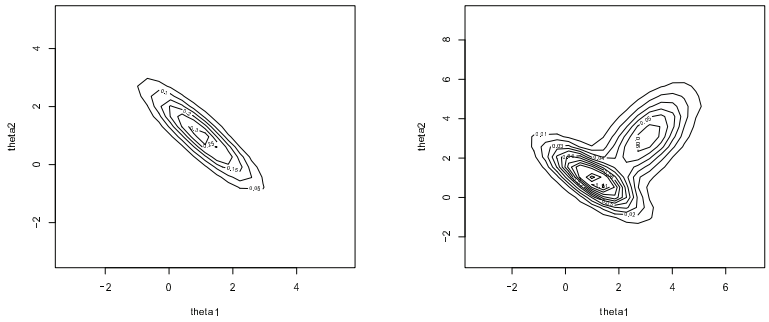

Figure 2: Likelihood functions for two-dimensional θ. (Left): An example with near-redundancy. (Right): A complex shape for the likelihood.

way and perhaps only one is needed (this can happen in regression models when you have too many predictors included). Notice that the likelihood is near flat along a ridge. The right panel is a likelihood from a mixture of distributions and as you can see, the likelihood can take on quite a complex shape. Here it is bimodal.

If we think back to the previous lecture, we had that the pdf could be factorized as

$$f(\tilde{x}|\theta) = h(\tilde{x})g(T(\tilde{x})|\theta).$$

Now, with $L(\theta|\tilde{x})$ and its relative magnitude only playing a role, we can conclude that the likelihood is a function of the sufficient statistic $T(\tilde{X})$! It follows that our maximum likelihood estimator is also a function of the sufficient statistic $T(\tilde{X})$.

This also hold true for Bayesian inference. To see this, recall Bayes theorem:

$$\pi(\theta|\tilde{x}) = \frac{f(\theta|\tilde{x})\pi(\theta)}{f(\tilde{x})} = \frac{g(T(\tilde{x})|\theta)\pi(\theta)h(\tilde{x})}{\int g(T(\tilde{x})|\theta)\pi(\theta)h(\tilde{x})d\theta}$$

and the $h(\tilde{x})$ terms cancel. So, the part of the posterior distribution that involves θ, like the likelihood, is a function of the sufficient statistic.

2 Point estimation

Before we go into detail about maximum likelihood estimation, let us think about the estimation problem in more general terms.

The standard notation for an estimator is $\hat{\theta}$. It is a function of the data so $\hat{\theta} = W(\tilde{X})$ for some statistics W. Is W a good estimator? To answer this question we have to decide what we mean by "good".

- **Bias: bias**$(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$. Is there a tendency to over or underestimate θ?
  Traditionally, unbiased estimators have been very popular. However, recently there has been much more focus on estimators that relax this constraint in favor of reduced...
- **Estimation variance:** $Var_\theta(\hat{\theta})$. How much does each estimate $\hat{\theta}$ vary from sample to sample (from true distribution f_θ).
- We commonly combine bias and variance into the **MSE, Mean squared error** $= E_\theta[\hat{\theta} - \theta]^2 = (bias_\theta(\hat{\theta}))^2 + Var_\theta(\hat{\theta})$
- **Consistency**. Where bias addresses what happens on average across different samples (if we could repeat the experiments under the same conditions), consistency ask how the estimator behaves if we were able to increase the sample size in the same experiment (of course, if iid new samples can just be tagged on to old ones to form a bigger sample). An estimator is consistent if

$$\hat{\theta} \to \theta, \quad n \to \infty$$

- **Sampling distribution**. For interval estimation and testing, we need to know more than just the first and second order moments of θ, we need to know the distribution of $\hat{\theta}$. In many cases, we will appeal to large n results (asymptotics) which results in $\hat{\theta}$ normally distributed.

We will look at these quantities in more detail later. For now, let's look at some alternative estimators.

Method of Moments, MoM:
We can compute the moments $E_\theta[X^k]$ for different k and match to the sample moments $\bar{X}, \bar{X^2}, \ldots \bar{X^k}$. By equating the theoretical moments to the observed ones, we get an equation system that we can solve for θ. This means that there are many possible MoMs to choose from.
MoM tend to be easy to compute, but are usually not as good as say maximum likelihood estimators. However, they can provide sensible starting values for numerical searches in complex likelihood inference.

Maximum likelihood, MLE:
As we already discussed, maximizing the likelihood function w.r.t. θ seems a sensible way to go. The benefit of MLEs are many (as we shall see), but you should be a bit careful. First, MLEs may be difficult to compute. They can also be quite bad when the sample size is small. Be especially cautious when you derive MLEs in cases where the support of the pdf depends on θ and boundary values for θ come into play. You need to check that the estimator you obtain using standard optimization techniques (like zero-crossing of the derivative) really gives you a global maximum.

Examples:

- Let $X_i \sim i.i.d.\ Be(p)$. Now, $\hat{p} = \bar{X}$ is a intuitive estimator for p.
  It is unbiased and its variance is $V(\hat{p}) = \frac{p(1-p)}{n}$
  An alternative estimate for p is $\hat{p}_B = \frac{n\bar{X}+a}{n+a+b} = w\bar{X} + (1-w)\frac{a}{a+b}$, where $w = n/(n+a+b)$. That is, we can consider a weighted average of the intuitive estimate $\bar{X}$ and an *a priori* guess for $p = \frac{a}{a+b}$. The above is in fact the Bayes estimate when we assume a $Beta(a, b)$ prior for p (we will return to this later in the lectures).

Now, clearly the estimator $\hat{p}_B$ is biased (since $\hat{p}$ is unbiased). It's variance is given as $w^2V(\hat{p})$. Combining the two we get

$$MSE(\hat{p_B}) = \frac{np(1-p)}{(n+a+b)^2} + (\frac{np+a}{n+a+b} - p)^2$$

Clearly, for some a, b and some n and p the MSE of $\hat{p}_B$ will be less than the MSE of $\hat{p}$.
What if we believe a priori that $p = .5$? This corresponds to $a = b$ in which case

$$MSE(\hat{p_B}) = \frac{p(1-p)(n-4a^2)+a^2}{(n+2a)^2}$$

This is still a function of the true p. What if we are interested in the performance of the estimator not being a function of p. This happens if we choose $a = \sqrt{n/4}$ for which the MSE is

$$MSE = \frac{n}{4(n+\sqrt{n})^2}$$

We can now compare $\hat{p}$ and $\hat{p_B}$ for this choice of $a = b = \sqrt{n/4}$. $\hat{p_B}$ is better than $\hat{p}$ (in a MSE sense) when

$$\frac{n}{4(n+\sqrt{n})} < \frac{p(1-p)}{n}$$

or

$$4(1+\frac{1}{\sqrt{n}})p(1-p) > 1$$

We can now clearly see that this happens when n is small, for almost all p, whereas when n gets larger and larger this only happens for p close to $1/2$. (Try this at home. Plot the MSEs.)
We shall return to this example when we talk about minimax estimation.

- Here's an example about the likelihood. Let's say X_i iid $U[0,\theta]$. The likelihood function is

$$L(\theta|\tilde{X}) = \prod_i \frac{1}{\theta} 1\{0 < x_i < \theta\} = \frac{1}{\theta^n} 1\{0 \le x_{(1)} \le x_{(n)} \le \theta\}$$

from which if follows that the likelihood is maximized for $\hat{\theta} = x_{(n)}$.

- Let's say X_i iid $f(x|\theta=0) = U[0,1]$, $f(x|\theta=1) = \frac{1}{2\sqrt{x}}, 0 < x < 1$. (Draw a picture of these two functions.)
If we have only one sample, it's easy to see that the MLE is $\hat{\theta} = 1\{x < 1/2\}$
What happens when we have n samples?

$$L(\theta=0|\tilde{x}) = 1, 0 < x_i < 1, \quad L(\theta=1|\tilde{x}) = \prod_i \frac{1}{2\sqrt{x_i}}$$

To determine if θ is 0 or 1 we compare the likelihood ratio to 1:

$$\frac{L(0|\tilde{x})}{L(1|\tilde{x})} = 2^n \prod_i \sqrt{x_i} > 1$$

If the above is true, the MLE is $\hat{\theta} = 0$ and otherwise it is 1.

3 Best unbiased estimator

As mentioned above, it has become much more common in recent years to focus on biased estimators that control estimation variance. However, we will first look more closely at unbiased estimators.

Is there such a thing as a "best" unbiased estimator? If MSE is our performance measure, the question can be rephrased as "among all unbiased estimators, is there a unique one with minimum variance?".

Best unbiased estimator: Let W^* be an unbiased estimator for θ. If $Var_\theta(W^*(\tilde{X}) \leq Var_\theta(W(\tilde{X}))$ for all unbiased estimators W, then W^* is UMVUE (uniform minimum variance unbiased estimator) for θ.

The above is not easy to check. You would have to be able to list all unbiased estimators W to check the statement. However, there is famous results by Cramer and Rao that help us out.

Cramer-Rao: Let $\tilde{X} \sim f(\tilde{x}|\theta)$ and $W(\tilde{X})$ be an estimator of θ. Further assume that $E_\theta W(\tilde{X})$ is differentiable in θ.
If $f(\tilde{x}|\theta)$ is such that

$$\frac{d}{d\theta}\int\cdots\int q(\tilde{x})f(\tilde{x}|\theta)dx_1\cdots dx_n = \int\cdots\int q(\tilde{x})\frac{d}{d\theta}f(\tilde{x}|\theta)dx_1\cdots dx_n$$

for all $q(\tilde{x})$ with $E_\theta|q(\tilde{x})|$, then

$$Var_\theta(W(\tilde{X})) \geq \frac{(\frac{d}{d\theta}E_\theta W(\tilde{X}))^2}{E_\theta[(\frac{d}{d\theta}\log f(\tilde{X}|\theta))^2]}$$

and iff X_i are iid $f(x_i|\theta)$ we can write

$$Var_\theta(W(\tilde{X})) \geq \frac{(\frac{d}{d\theta}E_\theta W(\tilde{X}))^2}{nE_\theta[(\frac{d}{d\theta}\log f(X|\theta))^2]}$$

Specifically, if $E_\theta W(\tilde{X}) = \theta$, then

$$Var_\theta(W(\tilde{X})) \geq \frac{1}{nE_\theta[(\frac{d}{d\theta}\log f(X|\theta))^2]}$$

The term $nE_\theta[(\frac{d}{d\theta}\log f(X|\theta))^2] = I_n(\theta) = nI(\theta)$ is called the *Information bound* or Cramer-Rao lower bound. Note that the bound in the iid case can be computed by a scaling factor n (sample size) and the *Information* $I(\theta)$ for a generic sample X from $f(x|\theta)$.

Why is this an important results? Well, now we know what to compare the variance of an unbiased estimator W to. If it achieves the bound above, we know it is the UMVUE. It also turns out that the MLE can achieve this bound, at least asymptotically, which is one reason it plays such a large role in estimation.

For $E_\theta W(\tilde{X}) = \tau(\theta)$, the CR bound is

$$Var_\theta(W(\tilde{X})) \geq \frac{\tau'(\theta)^2}{nI(\theta)}$$

Example:

- Let's revisit the example $X_i \sim U[0,\theta]$.
  We have that $\frac{d\log f(x|\theta)}{d\theta} = \frac{d\log(1/\theta)}{d\theta} = -1/\theta$ and so from the CR theorem it follows that

  $$Var_\theta(W(\tilde{X})) \geq \theta^2/n$$

  for all unbiased estimators W.
- The MLE is $\hat{\theta} = x_{(n)}$ (see above). Is it unbiased?
- $P(X_{(n)} < z) = P(all\ X_i < z) = \prod_i P(X_i < z) = (z/\theta)^n$

- Which gives us $f_{X_{(n)}}(z|\theta) = nz^{n-1}/\theta^n,\ 0 < z < \theta$ as the pdf for the maximum of a sample of n observations from $U[0,\theta]$.
- We can now compute $E[X_{(n)}] = \int_0^\theta \frac{nz^n}{\theta^n} dz = \frac{n}{n+1}\theta$.
- That is, the MLE is *biased*. However, as n goes to infinity, the bias disappears. This asymptotic unbiasedness is a property of the MLE.
- We do a *bias correction* and form a new estimator $Y = \frac{n+1}{n} X_{(n)}$.
- What is its variance? $Var_\theta[\frac{n+1}{n} X_{(n)}] = (\frac{n+1}{n})^2 (E[X^2_{(n)}] - (\frac{n}{n+1}\theta)^2) = \frac{\theta^2}{n(n+2)}$
- Wait a minute! This is actually less than the CR bound θ^2/n!!!! How is that possible?
- This was actually an example when the CR did not apply. Check at home if you really can change the order of integration and expectation for this example.
- In general, be careful when the support of f depends on θ.

The CR bound is not a so-called *sharp* bound. That is, we have a lower limit, but we don't actually know that there is an estimator that *can* achieve this variance. If we add some additional constraints to the pdf of X we can get a sharp bound.

Cramer-Rao and the MLE: Let X_i iid $f(x|\theta)$, where f satisfies the conditions of the CR theorem. $L(\theta|\tilde{x}) = \prod_i f(x_i|\theta)$ is the likelihood function and we denote by $W(\tilde{X})$ an unbiased estimator for $\tau(\theta)$. W achieves the CR bound iff

$$a(\theta)[W(\tilde{X}) - \tau(\theta)] = \frac{d}{d\theta} \log L(\theta|\tilde{x})$$

for some $a(\theta)$.
That is, we hit the bound with the MLE under certain transformations of θ where $W(\tilde{X}) - \tau\theta$ is proportional to the derivative of the loglikelihood.

This solves our question about the best unbiased estimators in part, but not when we have a problem where the conditions for the CR do not hold and not when our natural estimators don't take the form as above. Here's a very nice example from Casella and Berger:

Example:

- The normal pdf $N(\mu, \sigma^2)$ satisfies the CR conditions. A natural estimate for σ^2 is $S^2 = \frac{1}{n-1}\sum_i (X_i - \bar{X})^2$. However, the MLE is actually $\hat{\sigma}^2 = \frac{(n-1)}{n} S^2$ which is biased (since S^2 is unbiased).
- The variance of S^2 is $\frac{2\sigma^4}{n-1}$
- However, the CR bound is $2\sigma^4/n$ so S^2 does not achieve it.
- Can one achieve this bound?
- Checking the above statement, we need to find $\frac{d}{d\sigma^2} \log L(\mu, \sigma^2|\tilde{x}) = \frac{n}{2\sigma^4}(\sum_i \frac{(x_i-\mu)^2}{n} - \sigma^2)$ and equate it to the components in the box above.
- We see that $a(\sigma^2) = \frac{n}{2\sigma^4}$ and the best estimator for σ^2 is $\sum_i (x_i - \mu)^2/n$
- Wait a minute... this involves the unknown μ, so we cannot achieve the CR lower bound in this case.