



Estimation: The likelihood

Paul Hewson

Overview: There is a very good book in the library Yudi Pawitan (2001) “In all likelihood: Statistical Modelling and Inference Using Likelihood” Oxford (QA276.P286), particularly chapter 2. These notes are meant to be slightly interactive, mysterious green dots, squares and boxes appear which you can click on to answer questions and check solutions.

1. Overview

We have used long run frequency/symmetry arguments for some probability situations. We have also met a number of formulae that can be used as probability density functions, such as:

$$\text{Binomial } f(x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

$$\text{Exponential } f(x) = \theta e^{-\theta x}$$

Our interest now turns to the problem of parameter estimation. Given some data, we want a means of estimating the parameters such as π and θ) above. There are a range of methods that can be used, such as:

- Method of moments
- Method of least squares
- Method of maximum likelihood

We shall concentrate on the maximum likelihood method.



Back

◀ Doc

Doc ▶

The *likelihood* is a *function of the parameter*, given fixed data. Now imagine that a head appears nine times in ten throws:

- If the coin were unbiased,
$$P(\text{nine heads} | p = 0.5) = C_9^{10} \left(\frac{1}{2}\right)^{10} \approx 0.01.$$
- If we allowed our coin to be biased, and repeated this calculation for $p = 0.75$ we would find that
$$P(\text{nine heads} | p = 0.75) = C_9^{10} \left(\frac{3}{4}\right)^9 \left(\frac{1}{4}\right) \approx 0.19.$$
- We can repeat this calculation for other values of p .

Essentially, we carry out a this calculation to see how *likely* those data were, given different parameter values. We therefore call this calculation ($P(\text{Observed data} | \theta)$) the *likelihood function*. The *maximum likelihood estimator* is the value of θ which maximises this function.

Definition 1 *The likelihood function: let X_1, X_2, \dots, X_n be a random sample on a distribution characterised by parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. Let x_i be the observed value of X_i . The function:*

$$L = \prod_{i=1}^n P(X_i = x_i | \theta)$$

the likelihood function for θ given (x_1, x_2, \dots, x_n) . We wish to choose a value of θ that maximises L (or equivalently which maximises the log of L).



Back

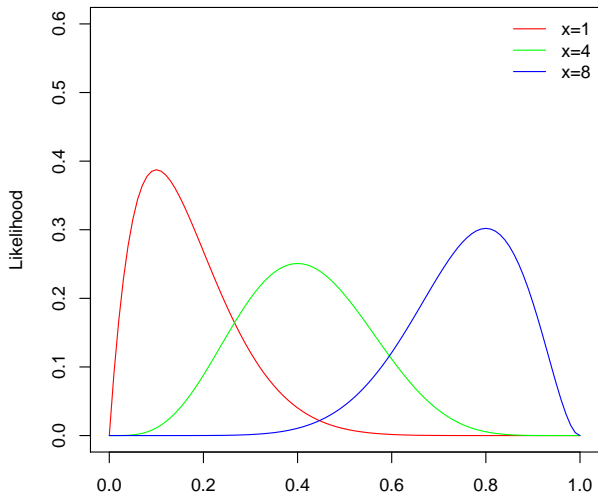
◀ Doc

Doc ▶

For a single observation from the Binomial distribution, the likelihood is simple:

$$L(\theta|x) = \binom{n}{x} \pi^x (1 - \pi)^{n-x}$$

For three separate experiments, each of size $n = 10$, we can plot three likelihoods having observed $x = 1$, $x = 4$ and $x = 8$:

 π

- The main idea behind maximum likelihood estimation is to take our estimate of π as the value that maximises this likelihood function.
- However, you should note immediately that there is other information here - specifically the curvature of the likelihood function

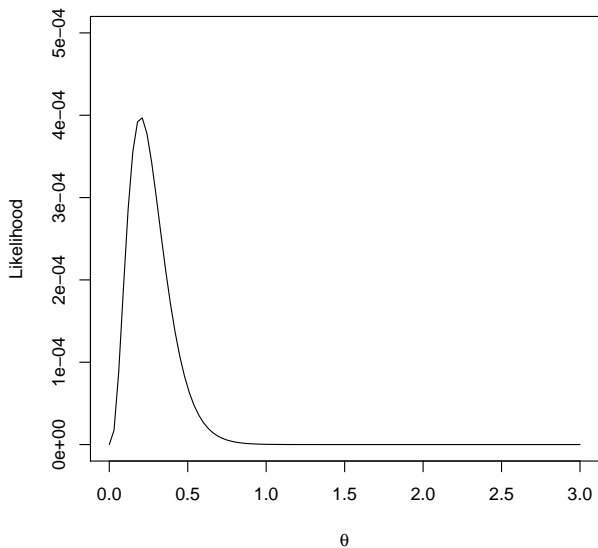
[Back](#)

This also works for continuous density functions. Consider an example with more than one observation, but also a continuous probability density function.

We observe failure rates for 3 lightbulbs as 4,2,5 years.

$$\begin{aligned}L(\theta) &= L(\theta|x_1) \times L(\theta|x_2) \times L(\theta|x_3) \\ &= \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \times \theta e^{-\theta x_3} \\ &= \theta^n e^{-\theta \sum_{i=1}^n x_i}\end{aligned}$$

As before we can plot this function:



But we will now consider the log-Likelihood.

$$l(\theta) = n \log(\theta) - \theta \sum_{i=1}^n x_i$$

So the method for finding the maximum should be clear

$$\frac{dl(\theta)}{d\theta} = \frac{n}{\lambda} - \sum_{i=1}^n x_i$$

Set this equal to zero gives us:

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i} \quad (1)$$

Note the use of the hat symbol to denote that we have an estimate.



Back



It is also worth noting that many people use the parameterisation

$$f(x) = \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} \text{ where } x \geq 0, \lambda \geq 0$$

for this distribution which has $\hat{\lambda} = \bar{x}$.



Back



1.1. Sufficiency

Note that we only need information on $\sum x_i$, we don't actually need to know $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Definition 2 A statistic $T(X)$ is sufficient for θ if, in an experiment, the conditional distribution of $X|T = t$ is free of θ .

This can only be proven using Moments, which have not covered in any detail. However, a further definition is also important:

Definition 3 A statistic $T(X)$ is sufficient for θ in an experiment if and only if the model M can be factorised:

$$M_x = g(t(x), \theta)h(x)$$

where $h(x)$ is free of θ .

This is conventionally clarified by using the Normal distribution. If x_1, x_2, \dots, x_n are iid from $N(\mu, \sigma^2)$ (so $\theta = (\mu, \sigma^2)$) then:

$$\begin{aligned}M_x &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp \left\{ -\frac{\sum x^2}{2\sigma^2} + \frac{\mu \sum x_i}{\sigma^2} - \frac{n\mu^2}{2\sigma^2} \right\}\end{aligned}$$

If σ^2 is known then $\sum x_i$ is sufficient, if not $\sum x_i^2$ is also required.



Back



Doc



Doc

Quiz Earlier we gave the likelihood function for a $X \sim \text{Binomial}(n, p)$ as:

$$L = C_x^n p^x (1 - p)^{n-x}$$

and could have stated the log-likelihood is given by:

$$\log(L) = \log C_x^n + x \log p + (n - x) \log(1 - p)$$

Find an expression for the value which maximises the likelihood or log-likelihood

(a) $\hat{p} = \frac{x}{p}$

(b) $\hat{p} = \frac{x}{n}$

(c) $\hat{p} = \frac{n-x}{n}$

(d) $\hat{p} = \frac{n}{x}$



Back

◀ Doc

Doc ▶

1.2. Maximum and curvature

A single number is not a good summary of a function.

If we use the term Score, summarised by $S()$, we have just seen that:

$$S(\theta) = \frac{\partial \log L(\theta)}{\partial \theta}$$

So the MLE $\hat{\theta}$ is the solution of the score equation $S(\theta) = 0$

At the maximum, the second derivative is negative. The curvature at $\hat{\theta}$ is denoted by $I(\hat{\theta})$ where $I(\theta)$ is given by:

$$I(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2}$$

Note that this is a value evaluated at the MLE and is therefore a single number and not a function. $I(\hat{\theta})$ is called the observed Fisher information.

A large curvature at $I(\hat{\theta})$ denotes a tight peak, and a more certain estimate concerning θ .

We can use $I(\theta)$ in order to calculate a confidence interval. Specifically, we can form a 95% “Wald” confidence interval using $\hat{\pi} \pm z_{1-\alpha/2} I(\hat{\pi})^{-1/2}$ where $z_{1-\alpha/2} = 1.96$. If we consider a binomial example (coins again) where $n = 10$ and $x = 8$ we have:

$$\hat{p} = \frac{8}{10}$$

Recall that, for the Binomial distribution with parameter π , if we observe x successes from n trials then:

$$\log L(\pi) = r \log \pi + (n - x) \log(1 - \pi)$$

$$S(\pi) = \frac{x}{\pi} - \frac{n - x}{1 - \pi}$$

$$I(\pi) = \frac{x}{\pi^2} + \frac{n - x}{(1 - \pi)^2}$$

where $\log L(\pi)$ is the log-likelihood function, $S(\pi)$ is the score equations and $I(\pi)$ is the observed information.

Given our expression for $I(\pi)$ as:

$$I(\pi) = \frac{x}{\pi^2} + \frac{n-x}{(1-\pi)^2}$$

our estimate of the standard error is given by

$$I(\pi)^{-\frac{1}{2}} = 1/\sqrt{\frac{8}{0.8^2} + \frac{2}{(0.2)^2}}$$

which is $1/\sqrt{62.5}$.

If we use the Wald formula for a confidence interval: $\pi \pm z_{\alpha/2} s.e.$ we have $0.8 \pm 1.96 \times 1/\sqrt{62.5}$

This gives the (somewhat implausible) C.I. as $0.55 < \pi < 1.05$.

Maybe we can understand what's going on a little better if we think about what we are doing.

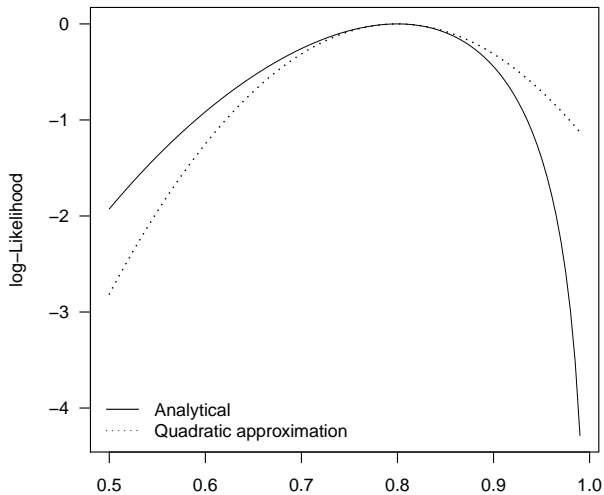
1.3. Quadratic approximation

If we take a second order Taylor's series expansion around $\hat{\theta}$ we get:

$$\log L(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\theta - \hat{\theta}) - \frac{1}{2}I(\hat{\theta})(\theta - \hat{\theta})^2$$

If we are interested in the log-likelihood of θ around $\hat{\theta}$ we can therefore use:

$$\log \frac{L(\theta)}{L(\hat{\theta})} = -\frac{1}{2}I(\hat{\theta})(\theta - \hat{\theta})^2$$

Binomial distribution with $n=10$, $x=8$  π 

Back

◀ Doc

Doc ▶

Plotting the likelihood and the quadratic approximation tells us something about how good our quadratic approximation is, and how reasonable it is to summarise our knowledge of the likelihood by two pieces of information.

2. Measures of Closeness

We need to say a little about how we determine whether a particular estimator (including the ones we get from maximum likelihood methods) are any good. It turns out that there are many (6 listed here) ways in which we can quantify how close an estimator is to the parameter of interest:

1. $P(|T - \theta| \leq |S - \theta|) = 1$
2. $E[g(T - \theta)] \leq E[g(S - \theta)]$ for every continuous function $g(\cdot)$ which is non-increasing for $x < 0$ and nondecreasing for $x > 0$.
3. $E[g(|T - \theta|)] \leq E[g(|S - \theta|)]$ for every continuous and non-decreasing function $g(\cdot)$ (universal dominance)
4. $P(|T - \theta| > \epsilon) \leq P(|S - \theta| > \epsilon)$ for every ϵ (stochastic dominance)
5. $E[(T - \theta)^2] \leq E[(S - \theta)^2]$ (mean squared error)
6. $P(|T - \theta| < |S - \theta|) \geq P(|T - \theta| > |S - \theta|)$



Here's a short game (intended to follow your mathematical logic). It's only intended as a way of seeing how well you understand what those six definitions mean (or alternatively whether you have a copy of Amemiya, 1994 to hand). They are all true or false questions about which property of closeness implies another property of closeness.

Quiz

- $(2) \implies (3)$
(a) True (b) False
- $(3) \implies (2)$
(a) True (b) False
- $(3) \implies (5)$
(a) True (b) False
- $(5) \implies (3)$
(a) True (b) False
- $(3) \implies (4)$
(a) True (b) False

[Back](#)[◀ Doc](#)[Doc ▶](#)

6. $(4) \implies (3)$

(a) True (b) False

7. $(1) \implies (3)$

(a) True (b) False

8. $(3) \implies (1)$

(a) True (b) False

9. $(1) \implies (6)$

(a) True (b) False

10. $(6) \implies (1)$

(a) True (b) False

What a lot of fun this could be. But it turns out convention suggests we most often use only one of these “measures” of closeness.



Back

◀ Doc

Doc ▶

2.1. Mean square error

There is no *a priori* reason to prefer any measure of closeness over another under all circumstances. Nevertheless, the use of the mean square error has tended to be the preferred measure of closeness used in assessing estimators. You should note this was the fourth measure listed above!

Let's illustrate this with an example that involves yet more coin tossing! Imagine a population of coin tosses. In the population, $X = 1$ with probability p , and $X = 0$ with probability $1 - p$. Given a sample of size $n = 2$, we could suggest three three estimators you might like to use for p :

- $T_2 = \frac{1}{2}(x_1 + x_2)$
- $S_2 = x_1$
- $W_2 = \frac{1}{2}$

Imagine that you have a biased coin and that $p = 0.75$. How do these three estimators behave. You might need to note for example that T_2 can take on three values $(0, \frac{1}{2}, 1)$ depending on whether 0, 1 or 2 heads were thrown, S_2 takes on two values and W_2 has one value regardless of whatever data you have. In order to work with expected values you need to determine the probability of obtaining these outcomes given $p = 0.75$.

Quiz

1. Estimate the *expected* mean square error for T_2 , i.e. $E[(T - 0.75)^2]$ over the three possible values of T_2 .

(a) $\frac{1}{32}$ (b) $\frac{1}{16}$ (c) $\frac{3}{32}$ (d) $\frac{1}{8}$

2. Estimate the expected mean square error for S_2 i.e. $E[(S_2 - 0.75)^2]$ over the two possible outcomes.

(a) $\frac{1}{16}$ (b) $\frac{1}{8}$ (c) $\frac{3}{16}$ (d) $\frac{1}{4}$

3. Estimate the expected mean square error for W_2 , i.e. $E[(W_2 - 0.75)^2]$ over the sole outcomes.

(a) $\frac{1}{16}$ (b) $\frac{1}{8}$ (c) $\frac{3}{16}$ (d) $\frac{1}{4}$



Back

◀ Doc

Doc ▶

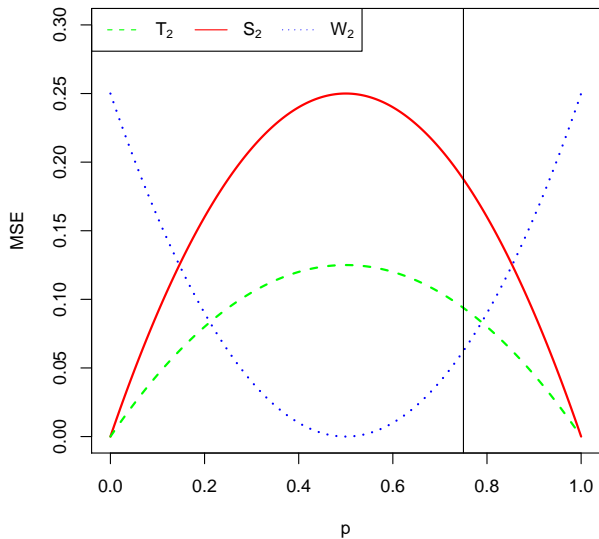
2.2. More generally

As it turns out, a little bit of algebra gives us a fairly general formula for each of these estimators, for any value of p .

- $E[(T_2 - p)^2] = \frac{1}{2}p(1 - p)$
- $E[(S_2 - p)^2] = p(1 - p)$
- $E[(W_2 - p)^2] = \left(\frac{1}{2} - p\right)^2$

We can even graph these functions as a value of p :

Expected mean square error for three estimators



You should see from the figure for example that for any value of p , T_2 has lower mean square error than S_2 and in some sense to be defined is clearly a better estimator. A vertical line also illustrates the calculations we have just done assuming $p = 0.75$. This leads us into one desirable property of estimators:

Definition 4 *Efficiency*: consider S and T (in general, not in relation to the two estimators we have just been working with); two estimators of θ . T is said to be more efficient than S if

$$E[(T - \theta)^2] \leq E[(S - \theta)^2] \text{ for all } \theta \in \Theta$$

and

$$E[(T - \theta)^2] < E[(S - \theta)^2] \text{ for at least one value of } \theta \in \Theta$$

So *efficiency* of an estimator is one way of deciding whether it is any good. Note that the definition is all about the goodness of the estimator in terms of mean square!

Now, what about the terminology we use to compare estimators:

Definition 5 *Admissibility*: let $\hat{\theta}$ be an estimator of θ . We say that it is inadmissible if there is a more efficient estimator (and we have just defined efficiency). An admissible estimator is one that is not inadmissible.

Quiz In our example above, S_2 is said to be inadmissible

(a) True

(b) False



Back



3. Biased estimators

Now that we have made comparisons in terms of mean square. But we need to look at a component of this, the way which our estimator could systematically depart from the true value.

Definition 6 *Unbiased estimators:* θ is said to be an unbiased estimator of θ if $E[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. The quantity $E[\hat{\theta} - \theta]$ is referred to as the bias.

Quiz

1. In our example above, T_2 is unbiased
 - (a) True
 - (b) False
2. Likewise, S_2 is unbiased
 - (a) True
 - (b) False
3. And finally, W_2 is unbiased
 - (a) True
 - (b) False



Back

◀ Doc

Doc ▶

However, when considering estimators in general, how important is bias. Specifically:

Quiz Is it more important to have an unbiased estimator at the expense of higher mean square error than it is to reduce mean square error at the expense of bias

(a) Yes (b) No



Back

◀ Doc

Doc ▶

3.1. Decomposition of mean square error

Finally, it is valuable to know that we can “decompose” the mean square error of an estimator into these two constituent parts:

Definition 7 *Decomposition of estimator variance: the mean square error is the sum of the variance and the bias squared, i.e.*

$$E[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + (E[\hat{\theta} - \theta])^2$$

So a large mean square error could mean either a large bias, or a large variance, or a combination.

4. Asymptotic properties

4.1. Cramér-Rao Lower Bound

It turns out that there is a well developed method for assessing the quality of our estimators in terms of their theoretical variance. Specifically, the term known as the Cramér-Rao Lower Bound places a limit on the variance.

Definition 8 *Cramér-Rao Lower Bound*: let $L(X_1, X_2, \dots, X_n | \theta)$ be the likelihood function, and let $\theta(X_1, X_2, \dots, X_n)$ be an unbiased estimator of θ . Under some general conditions, we have:

$$\text{Var}(\hat{\theta}) \geq -\frac{1}{E \left[\frac{d^2 \log L}{d\theta^2} \right]}$$

The phrase “under general conditions” provides me with an excuse to say nothing further about this formula. If you like worrying about missing details, you may be reassured to note that these conditions include *regularity conditions* such as the support of L (the domain over which it is positive) must not depend on θ .



Back

◀ Doc

Doc ▶

Mean square error, bias and so on are features of the finite sample properties of the estimators we are working with. It is commonly quite difficult to obtain the exact moments, never mind the exact distribution of estimators. Under these circumstances we work with an approximation to the distribution of the moments. Such *asymptotic approximation* relies on us considering what happens in the limit of the sample size going to infinity. Our last session considered the techniques necessary to be able to do this. In order to do this, we require our estimators to have the following properties:

Definition 9 *Consistency*: $\hat{\theta}$ is a consistent estimator of θ if $\hat{\theta} \xrightarrow{P} \theta$. Under general conditions, the maximum likelihood estimators can be shown to be consistent estimators of their population values

Definition 10 *Asymptotically Normal*: let $L(X_1, X_2, \dots, X_n | \theta)$ be the likelihood function. Under general conditions, the maximum likelihood estimator $\hat{\theta}$ is asymptotically distributed as:



Back



Doc



Doc

$$\hat{\theta} \stackrel{A}{\sim} \text{Normal} \left(\theta, -E \left[\frac{d^2 \log L}{d\theta^2} \right]^{-1} \right)$$

You should recognise the variance term here! The asymptotic variance of the maximum likelihood estimator is the same as the Cramér-Rao Lower Bound. We can therefore say:

Definition 11 *Asymptotic efficiency*: If the asymptotic distribution is indeed given as above, then we can regard a consistent estimator as being asymptotically efficient.

The maximum likelihood estimator is therefore asymptotically efficient by definition.



Back



Doc



Doc

5. Invariance principle

Definition 12 *Invariance principle*: if $\hat{\theta}$ is a maximum likelihood estimator of θ , and $g(\theta)$ is a function of θ then $g(\hat{\theta})$ is a maximum likelihood estimator of $g(\theta)$.

Imagine $n = 10$ and $x = 8$. Until now we've discussed the idea of a proportion $\frac{x}{n}$. But we might also be interested in odds: $\frac{x}{n-x}$

$$\frac{\text{Number of successes}}{\text{Number of failures}}$$

Let's consider odds from the point of view of the parameters. If π is the proportion, then odds are $\frac{\pi}{1-\pi}$, and log odds are given as:

$$\psi = \log\left(\frac{\pi}{1-\pi}\right)$$

[Back](#)[◀ Doc](#)[Doc ▶](#)

So if we have an m.l.e. for π such that $\hat{\pi} = \frac{8}{10}$, and if we have $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ then the m.l.e. for $g(\pi)$ is

$$\hat{\psi} = \log\left(\frac{0.8}{1-0.8}\right) = 1.39$$



Back



5.1. Transforming intervals

Instead of deriving an interval for π directly, we can derive an interval for a **transformation** of π . In order to form an interval estimate for ψ , we need to use:

$$\text{se}[g(\pi)] = \text{se}(\pi) \left| \frac{\delta g}{\delta \hat{\pi}} \right|.$$

We know that $\frac{\delta g}{\delta \hat{\pi}} = \frac{1}{\hat{\pi}} + \frac{1}{1-\hat{\pi}}$. Given that we earlier noted $s.e.(\pi) = 1/\sqrt{62.5}$

$$s.e.[g(\pi)] = 1/\sqrt{62.5} \times \left(\frac{1}{0.8} + \frac{1}{1-0.8} \right) = 0.79$$

We therefore obtain a CI using $1.39 \pm 1.96 \times 0.79$ i.e., $-0.16 < \psi < 2.94$

Noting that $g^{-1}(\psi) = \frac{\exp(\psi)}{1+\exp(\psi)}$ we can transform this interval to give the 95% Wald Interval for π :

$$0.46 < \pi < 0.95$$



Back

◀ Doc

Doc ▶

- What do you think the quadratic approximation to the logit transformed likelihood might look like?



Back



5.2. Bayesian analysis

As we shall find out, in practice, we shall use Bayes theorem to give us:

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

We shall essentially use the formula:

$$p[A_i|B] = \frac{p[B|A_i]p[A_i]}{\sum_{j=1}^n p[B|A_j]p[A_j]}, i = 1, 2, \dots, n$$

Which can be applied to discrete and continuous density functions:

$$f(\theta|x) = \frac{L(\theta)f(\theta)}{\int L(\theta)f(\theta)}$$

But, given we are working with a likelihood (rather than a pdf which must sum/integrate to 1) we will often ignore the denominator and simply use

$$f(\theta|x) \propto L(\theta)f(\theta)$$



It follows that properties of the Likelihood are important in both Bayesian and Frequentist statistics.

But even if we assume that the prior $f(\pi) = 1$, then the posterior:

$$post(\psi|x) = f(\theta(\psi)|x) \times \left| \frac{\partial \theta}{\partial \psi} \right|$$

Which might or might not be a problem. For the logit transformation,

$$\left| \frac{\partial \theta}{\partial \psi} \right| = \frac{e^\psi}{(1+e^\psi)^2}.$$

If we examined the likelihood ratio of $\pi_1 = 0.8$ against $\pi_2 = 0.3$:

$$\frac{L(\pi_1 = 0.8)}{L(\pi_2 = 0.3)} = \frac{\pi_1^8(1 - \pi_1)^2}{\pi_2^8(1 - \pi_2)^2} = 208.7$$

But if we evaluate the Jacobian when comparing $\psi = 1.386$ with $\psi = -0.847$ we find the likelihood ratio is reduced by 0.81.

This matter needs some consideration!

Solutions to Quizzes

Solution to Quiz: The log-likelihood was given by:

$$\log(L) = \log C_x^n + x \log p + (n - x) \log(1 - p)$$

we want to find $\frac{d \log L}{dp} = 0$, i.e.

$$\frac{d \log L}{dp} = \frac{x}{p} - \frac{n - 1}{1 - p}$$

set this equal to zero and solve:

$$\hat{p} = \frac{x}{n}$$

To be very tidy, one perhaps should check that $\frac{d^2 \log L}{dp^2}$ is negative. Why?



Solution to Quiz:

We need to calculate $E[(T-0.75)^2]$. We have four possible outcomes, tabulated below:

Head	$\frac{3}{4}$	\nearrow	Head	$\frac{3}{4} \times \frac{3}{4}$	$\frac{9}{16}$
		\searrow	Tail	$\frac{3}{4} \times \frac{1}{4}$	$\frac{3}{16}$
		\nearrow	Head	$\frac{1}{4} \times \frac{3}{4}$	$\frac{3}{16}$
Tail	$\frac{1}{4}$	\searrow	Tail	$\frac{1}{4} \times \frac{1}{4}$	$\frac{1}{16}$

The expected mean square is therefore:

$$E\left[\left(T - \frac{3}{4}\right)^2\right] = \left(0 - \frac{3}{4}\right)^2 \left(\frac{1}{16}\right) + \left(\frac{1}{2} - \frac{3}{4}\right)^2 \left(\frac{3}{8}\right) + \left(1 - \frac{3}{4}\right)^2 \left(\frac{9}{16}\right)$$

which gives

$$\frac{9}{16} \left(\frac{1}{16}\right) + \frac{1}{16} \left(\frac{6}{16}\right) + \frac{1}{16} \left(\frac{9}{16}\right) = \frac{9}{256} + \frac{6}{256} + \frac{9}{256} = \frac{24}{256} = \frac{3}{32}$$



Solution to Quiz: Here we are only interested in X_1 . Hence if $X_1 = 1$ (an event that occurs with probability 0.75) our estimate $\hat{p} = 1$, if $X_1 = 0$ (an event that occurs with probability 0.25), our estimate $\hat{p} = 0$. Hence

$$E[(T - p)^2] = \left(1 - \frac{3}{4}\right)^2 \left(\frac{3}{4}\right) + \left(0 - \frac{3}{4}\right)^2 \left(\frac{1}{4}\right)$$

which is


$$\frac{1}{16} \left(\frac{3}{4}\right) + \frac{9}{16} \left(\frac{1}{4}\right) = \frac{3}{16}.$$



Solution to Quiz: This is the equivalent of the person who was offered one of two watches, one that was completely fixed and didn't move, one that was behind the time by an hour. This person chose the first watch on the basis that it would be correct twice a day.

So, we have to calculate only:

$$E\left[\left(W_2 - \frac{3}{4}\right)\right] = \left(\frac{1}{2} - \frac{3}{4}\right)^2 = \frac{1}{16}$$

Hey, this is the lowest mean square error! And we didn't need any data to form our estimate. 

Solution to Quiz: S_2 is said to be inadmissible because it is dominated by T_2 , i.e. T_2 is a more efficient estimator than S_2 ■

Solution to Quiz: We need to note that $E[\hat{\theta}] = 1 \left(\frac{9}{16}\right) + \frac{1}{2} \left(\frac{3}{8}\right) + 0 \left(\frac{1}{16}\right) = \frac{12}{16}$ hence the estimator is unbiased. ■

Solution to Quiz: Here we only need to note that $E[\hat{\theta}] = 1 \left(\frac{3}{4}\right) + 0 \left(\frac{1}{4}\right)$ hence the estimator is unbiased. ■

Solution to Quiz: Here, I'm not sure whether $E[\hat{\theta}]$ is really defined, but if we follow the logic above it is $\frac{1}{2}(1)$ which looks pretty much biased to me. ■

Solution to Quiz: Well, the convention has been to prefer to keep the mean square error as small as possible, even if that means we have some bias in our estimator. But do note that this is only a convention!

