The Use of Likelihood Inference for Quantifying Statistical Evidence

Michael Lanier

May 10, 2016

1 Introduction to the Likelihood Function

The purpose of this paper is to introduce in broad strokes the use of the likelihood function as a tool for statistical inference and data evaluation. In this first section we will develop some definitions and theorems that will be more or less assumed as given for subsequent sections. This section is present to appease those who might otherwise hold issue with the lack of devlopment subsequent sections. One might consider passing over this section if the main point must be reached in haste.

Definition 1. Let $X_n = (X_1, ..., X_n)$ have joint density $p(x_n|\theta) = p(x_1, ..., x_n|\theta)$ where $\theta \in \Omega$. Ω is the parameter space and will be assumed compact for the duration of this paper. The likelihood function

$$L : \Omega \to [0, \infty)$$

is defined by

$$L(\theta) = L(\theta; x_n) = p(x_n|\theta)$$

where x_n is fixed and $\theta \in \Omega$.

Definition 2. The Maximum Likelihood Estimator is defined as follows:

$$MLE = \underset{\theta \in \Omega}{Argmax}(L(\theta; x_n))$$

The MLE does not always exist. The primary examples used in this paper to justify the use of likelihood inference rely on the Binomial Distribution, the Normal Distribution, and in one example the Negative Binomial distribution. These three distributions and their MLEs are to be examined.

J. Aragon (1991) proved the existance and uniqueness of the negative bionmial distibution MLEs and his proof will not be included here. The existance and uniqueness of the MLEs for the Binomial Distribution and the Normal Distribution will now be examined. We will present a definition and a theorem due to Johnson and Kotz (1969) in an effort to prove the existance and uniqueness of the MLE in the Bionomial case.

Theorem 3 (Johnson and Kotz (1969))**.** Let $\theta \in \Omega$ and let S be a random sample with sample mean $\bar{x}$ and sample variance s^2. If $s^2 > \bar{x}$, then the MLE for θ exists and is unique.[1]

Theorem 4. Let $X \sim BIN(n, \theta)$ with an observed sample of size n with x successes and likelihood $L(\theta; x)$ such that $n \neq x$.

$$MLE = \underset{\theta \in \Omega}{Argmax}(L(\theta; x_n))$$

exists and is unique.

Proof. It is sufficient to show that the sample variance is greater than the sample mean. Consider the following derivation.

$$n - x > 1, \text{since } n \neq x$$

so

$$1 > 1/(n - x)$$

$$n > n/(n - x)$$

leading to

$$n > \frac{1}{\frac{n-x}{n}}$$

$$n > \frac{1}{1 - x/n} = \frac{1}{1 - \hat{p}}$$

and finally

$$n\hat{q} > 1$$

this implies

$$n\hat{p}\hat{q} > \hat{p}$$

so,

$$s^2 > \bar{x}$$

Thus the sufficient condition is demonstrated and the MLE exists and is unique. □

Since the binomial distribution converges to the normal distribution for increasing n an argument can be proposed that since the MLE for the population proportion of a binomial distribution exists and is unique, then the MLE for the mean and varience of a normal distribution also exists and is unique. However, the existance and uniqueness of the MLE is not enough for our purposes here. We wish not only for their existance, but also for their identification. Next we will find the MLE of the population proportion in a binomial case, and the MLE of the population mean and varience in the normal case.

[1] Note though, this is not a necessary condition.

Claim 5. Let $X \sim BIN(n, \theta)$. The MLE of θ is $\hat{\theta} = x/n$ where x is the number of success observed in n Bernoulli trials.

Proof. We need to solve the following maximization problem

$$\underset{\theta}{Argmax}(L(\theta; x))$$

where

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

We will maxmize the log likelihood to find the MLE.[2]

$$ln(L(\theta; x)) = k + x ln(\theta) + (n-x) ln(1-\theta)$$

where k is a constant with respect to θ.

$$\frac{\partial}{\partial \theta} ln(L(\theta; x)) = \frac{x}{\theta} - \frac{n-x}{1-\theta}$$

Setting this to zero and solving for θ gives us

$$\theta = \frac{x}{n}$$

So the MLE for θ is $\hat{\theta} = \frac{x}{n}$. □

Claim 6. Consider a set of normal random variables $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with

$$ln(L(\mu, \sigma^2)) = (-\frac{n}{2} ln(2\pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_i - \mu)^2)$$

the MLE of the mean and varience are

$$\hat{\mu_n} = \frac{\sum_{j=1}^n x_j}{n}$$

and

$$\hat{\sigma_n^2} = \frac{\sum_{j=1}^n (x_j - \hat{\mu})^2}{n}$$

Proof. We need to solve the following maximization problem

$$\underset{\mu, \sigma^2}{Argmax}(L(\mu, \sigma^2; x_1, ... x_n))$$

The first order conditions for a maximum are

$$\frac{\partial}{\partial \mu} ln(L(\mu, \sigma^2; x_1, ..., x_n)) = \frac{\partial}{\partial \sigma^2} ln(L(\mu, \sigma^2; x_1, ... x_n)) = 0$$

[2] The natural log is monotonically increasing function so the logarithm of a function achieves its maximum value at the same value as the function itself.

The partial derivative of the log-likelihood with respect to the mean is

$$\frac{\partial}{\partial \mu} ln(L(\mu, \sigma^2; x_1, ..., x_n))$$
$$= \frac{\partial}{\partial \mu}(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_i - \mu)^2)$$
$$= \frac{\sum_{j=1}^{n}(x_j - \mu)}{\sigma^2}$$
$$= \frac{\sum_{j=1}^{n}(x_j) - n\mu}{\sigma^2}$$

which is equal to zero only if

$$\sum_{j=1}^{n}(x_j) - n\mu = 0$$

This implies an MLE

$$\hat{\mu_n} = \frac{\sum_{j=1}^{n} x_j}{n}$$

as the MLE for μ.

The partial derivative of the log-likelihood with respect to the variance is

$$\frac{\partial}{\partial \sigma^2} ln(L(\mu, \sigma^2; x_1, ..., x_n))$$
$$= \frac{\partial}{\partial \sigma^2}(-\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_i - \mu)^2)$$
$$= \frac{-n}{2\sigma^2} - [\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu)^2]\frac{d}{d\sigma^2}(\frac{1}{\sigma^2})$$
$$= \frac{-n}{2\sigma^2} - [\frac{1}{2}\sum_{j=1}^{n}(x_j - \mu)^2](\frac{-1}{\sigma^4})$$
$$= \frac{1}{2\sigma^2}[\frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j - \mu)^2 - n]$$

since $\sigma^2 > 0$, is equal to zero only if

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \mu)^2$$

This implies an MLE

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \hat{\mu})^2$$

for σ^2 $\square$

Theorem 7 (Fisher–Neyman Factorization Theorem)**.** The statistic T is sufficient for θ if and only if functions g and h can be found such that

$$f_X(x|\theta) = h(x)g(\theta, T(x))$$

A proof will be presented for the discrete case.

Proof. Since T is a function of X, we have $f_\theta(x,t) = f_\theta(x)$ (only when $t = T(x)$ and zero otherwise) and thus:

$$f_\theta(x) = f_\theta(x,t) = f_{\theta|t}(x)f_\theta(t)$$

with the last equality being true by the definition of conditional probability distributions. Therefore,

$$f_\theta(x) = a(x)b_\theta(t)$$

with

$$a(x) = f_{\theta|t}(x)$$

and

$$b_\theta(t) = f_\theta(t)$$

On the other hand, if $f_\theta(x) = a(x)b_\theta(t)$, we have

$$f_\theta(t) = \sum_{x:T(x)=t} f_\theta(x,t) = \sum_{x:T(x)=t} f_\theta(x) = \sum_{x:T(x)=t} a(x)b_\theta(t) = b_\theta(t) \sum_{x:T(x)=t} a(x)$$

The first equality is by the definition of pdf, the second by the remark above, the third by hypothesis, and the fourth because the summation is not over t. We find the conditional probability to then be,

$$\begin{aligned} &f_{\theta|t}(x) \\ &= \frac{f_\theta(x,t)}{f_\theta(t)} \\ &= \frac{f_\theta(x)}{f_\theta(t)} \\ &= \frac{a(x)b_\theta(t)}{b_\theta(t)(\sum_{x:T(x)=t} a(x))} \\ &= \frac{a(x)}{\sum_{x:T(x)=t} a(x)} \end{aligned}$$

With the first equality by definition of conditional probability density, the second by the remark above, the third by the equality proven above, and the fourth by algebraic simplification. This expression does not depend on θ and thus T is a sufficient statistic.

□

Because of sufficiency no data information is lost from the MLE. We will be primarily be using the MLE to scale the likelihood ratio to make comparison across mutiple examples. However in example 2.5 we will see how the data evidence is not symmetric in the case of insufficiency.

Definition 8. Let the MLE exist and be unique, then the likelihood ratio is

$$LR(\theta) = \frac{L(\theta)}{L(\underset{\theta\in\Omega}{Argmax}(L(\theta; x_n))} = \frac{L(\theta)}{L(MLE)}$$

Definition 9. The Likelihood Principle states that all the information about the parameter is in its likelihood function.

Proposition 10. The Likelihood ratio is preserved by a reparametrization $\lambda = q(\theta)$ with a known function q. This means the Likelihood ratio is invarient under transformation, i.e. it has the invarience property.

2 Likelihood Inference

To motivate this paper, let us start with a simple binomial random variable, $X \sim BIN(n, \theta)$. Recall that the maximum likelihood estimator (MLE) is derived to be the sample proportion $\hat{\theta} = x/n$. Likelihood inference is based on the same reasoning, but rather than focus only on the maximum, focus is on how well the data supports parameters across the entire parameter space.

Example 2.1. Let $X \sim BIN(n = 10, \theta)$, with $x = 8$ successes observed. The likelihood function is given by

$$L(\theta) = \tbinom{n}{x}\theta^x(1-\theta)^{n-x}$$

The corresponding Likelihood ratio is given by

$$LR(\theta) = \frac{\tbinom{n}{x}\theta^x(1-\theta)^{n-x}}{\tbinom{n}{x}\hat{\theta}^x(1-\hat{\theta})^{n-x}}$$

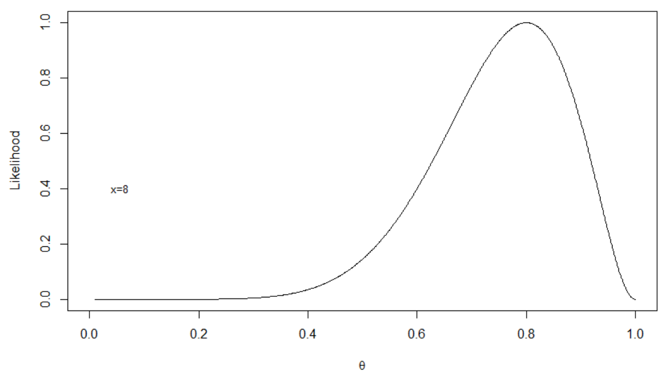
The MLE is $\hat{\theta} = .8$ ($x = 8$ successes in $n = 10$ tries) gives us

$$LR(\theta) = \frac{\theta^8(1-\theta)^2}{.8^8(.2)^2}$$

We can graph the likelihood ratio over its parameter space to note how the the evidence supports various values of the parameter.

Often the MLE is the focal point of the likelihood inference to such a degree that the MLE is thought to contain nearly everything we want to know about the evidence. This plot demonstrates this is not the case as there is a range

Figure 1: $LR(\theta; \hat{\theta} = .8)$



of parameters with likelihood nearly that of the maximum. Referring to the likelihood function[3] will allow us to quantify the amount of information in an observed sample. We will see this in the following examples.

Example 2.2. Consider a farmer planting 100 seeds of a particular species of corn. The farmer wants to estimate the probability that the seed germinates. Let the number of germinating seeds be a binomial random variable $X \sim BIN(n = 100, \theta)$ with θ being the probability of germination. Suppose he observes $x = 5$ successes. The likelihood ratio as a function of θ is shown as

$$LR(\theta) = \frac{\binom{100}{5}\theta^5(1-\theta)^{95}}{\binom{100}{5}\hat{\theta}^5(1-\hat{\theta})^{95}} = \frac{\theta^5(1-\theta)^{95}}{.05^5(.95)^{95}}$$

Example 2.3. Consider the farmer from Example 2.2. Every year seed is sown, in the following year some of the seed from the prior year germinates in the new year. Instead of observing $x = 5$ he has observed $x \leq 10$, because 10 seeds have germinated, but some of them could have been from the prior year. The farmer is interested in the number of seeds x sown this year. In this case, the likelihood ratio becomes

$$LR(\theta) = \frac{\sum_{x=0}^{10} \binom{100}{x}\theta^x(1-\theta)^{n-x}}{1}$$

Note that the denominator is 1. The likelihood function can be written as $P(X \leq 10|\theta)$. The largest a probability can be is 1, occurring when $\theta = 0$. Thus, we take $\hat{\theta}$ as the MLE, and $L(\hat{\theta}) = 1$ as the maximum likelihood. This is demonstrated in a graph of the likelihood over the parameter values.

[3]We will often scale the likelihood function as a factor of its maximum. We will use the terms "likelihood function" and "likelihood ratio" interchangeably.

Figure 2: $LR(\theta; \hat{\theta} = .05)$

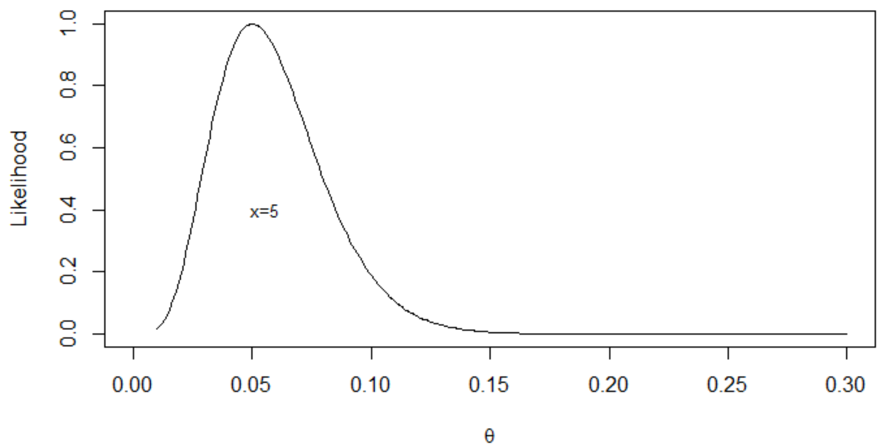
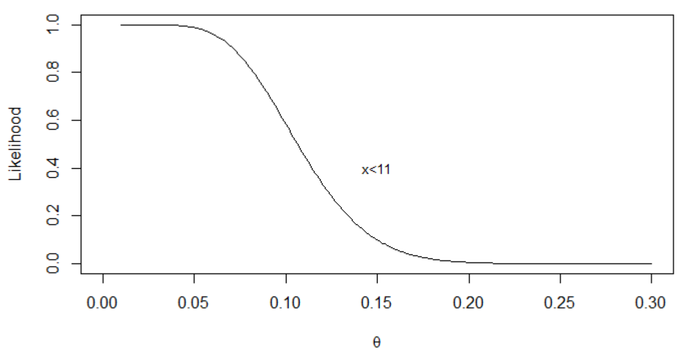


Figure 3: $LR(\theta; \hat{\theta} = 0)$



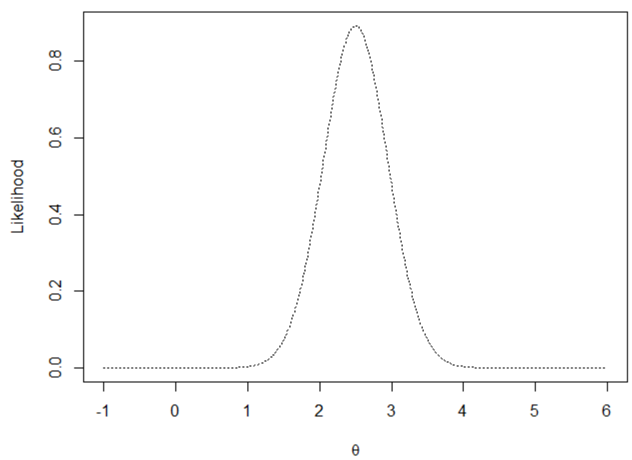
We have previously determined that the MLE occurs at $\hat{\theta} = 0$. This is also apprent from figure 3. We can compare this graph to that in example 1.2. Obviously there is more information when an exact value is observed, as in the first case, than when an interval is observed as in the second case. This conclusion could not be obtained from the MLE alone. Additionally, the MLE in the example 2.3, $\hat{\theta} = 0$, does not provide a summary of likely values of θ.

Example 2.4. Consider a set of normal random variables $X_1, ..., X_n \overset{iid}{\sim} N(\theta, 1)$ with $n = 5$ and an observed $\bar{x} = 2.5$. The MLE was earlier derived to be the sample mean $\bar{x}$. The corresponding likelihood ratio can be given by

$$LR(\theta) = \frac{\prod_{i=1}^{n} \frac{\exp\{\frac{-(x-\theta)^2}{2}\}}{\sigma\sqrt{2\pi}}}{\prod_{i=1}^{n} \frac{\exp\{\frac{-(x-2.5)^2}{2}\}}{\sigma\sqrt{2\pi}}}$$

Figure 4 provides a view of the evidence provided by the data.

Figure 4: $LR(\theta; \hat{\theta} = 2.5)$



In this example we know from sufficiency that no information is lost in observing just the sample mean. A consequence of sufficiency is that the likelihood function can be written based on the sample as

$$L(\theta) = \frac{\exp\{\frac{-n(\bar{x}-\theta)^2}{\sigma^2}\}}{\sigma\sqrt{2\pi}}$$

Let's observe a case where the observed data is not sufficient for the parameter. That is, the observed data does involve lost information. We can use likelihood inference to quantify the degree to which the evidence is less precise.

Example 2.5. Consider a set of normal random variables $X_1, ..., X_n \overset{iid}{\sim} N(\theta, 1)$ with $n = 5$ and an observed $x_{max} = 3.5$. Calculation of he likelihood ratio requires the probability distribution function of the largest order statistic. Begin with the cumulative distribution function.

$$G(x_{max}) = P(\text{all } X_i leqx) = F(x)^n$$

If F,f are the cumulative distibution function and probability distibution function respectively of the distribution $N(\theta, 1)$, we can write $F(x) = \Phi(x - \theta)$ and

$f(x) = \phi(x - \theta)$. The probability distribution on the largest order statistic is then,

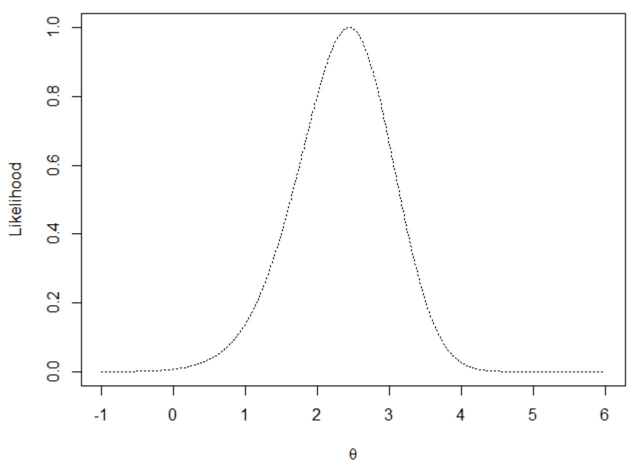
$$g(x_{max}) = n[F(x)]^{n-1}(f(x))$$

The likelihood function for normal data is then

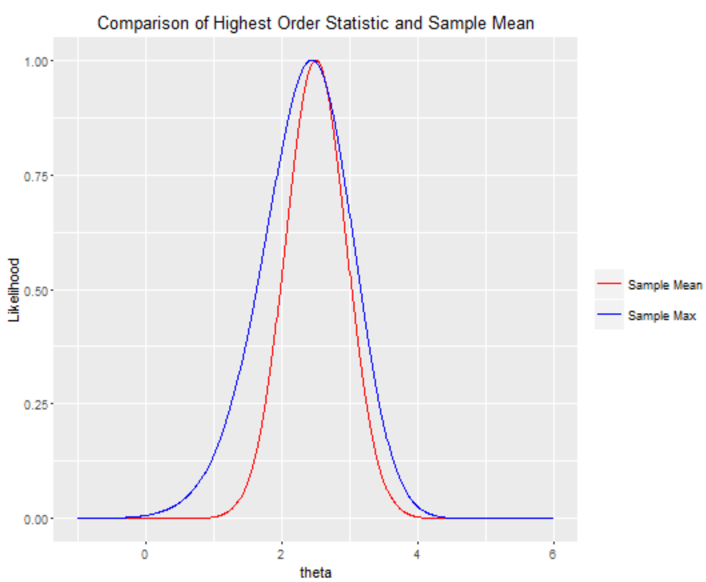
$$L(\theta) = n\phi(x_{max} - \theta)[\Phi(x_{max} - \theta)]^{n-1}$$

The MLE is found using numerical methods to be $\hat{\theta} = 2.44$. A graph of the likelihood function is as follows:

Figure 5: $LR(\theta; \hat{\theta} = 2.44)$



The likelihood ratios in the two cases are maximized near the same θ. But as an overlay shows us, the likelihood function in the second case is more disperse indicating a lower degree of available evidence. This is because of the loss of a sufficent statistic in the second example 2.5.



Comparison of Highest Order Statistic and Sample Mean
Likelihood
1.00
0.75
0.50
0.25
0.00
0
2
4
6
theta
Sample Mean
Sample Max

Not only does the likelihood function quantify evidence, it also has the property that it is not dependent on the intention of the experimenter. This is known as the likelihood principle and will be illustrated with the following examples.

Example 2.6. Consider a team of geneticists investigating the prevalence of a rare genotype. However, the geneticists' sampling scheme is not predetermined. If the sampling continues for a fixed number of trials, then the number of successes is a random variable $X \sim BIN(n, \theta)$. It may be that the sampling will continue until a certain number of successes. In this case, the number of trials is a random variable $N \sim NB(x, \theta)$. The likelihood ratio in the former case is given by

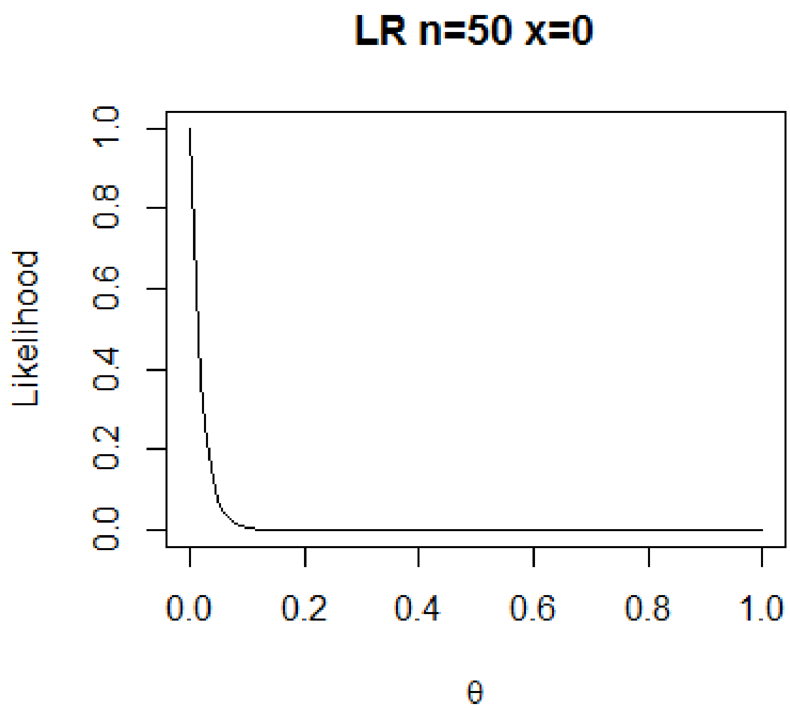
$$LR_1(\theta) = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{\binom{n}{x}\hat{\theta}^x(1-\hat{\theta})^{n-x}}$$

In the latter case the likelihood ratio is given by

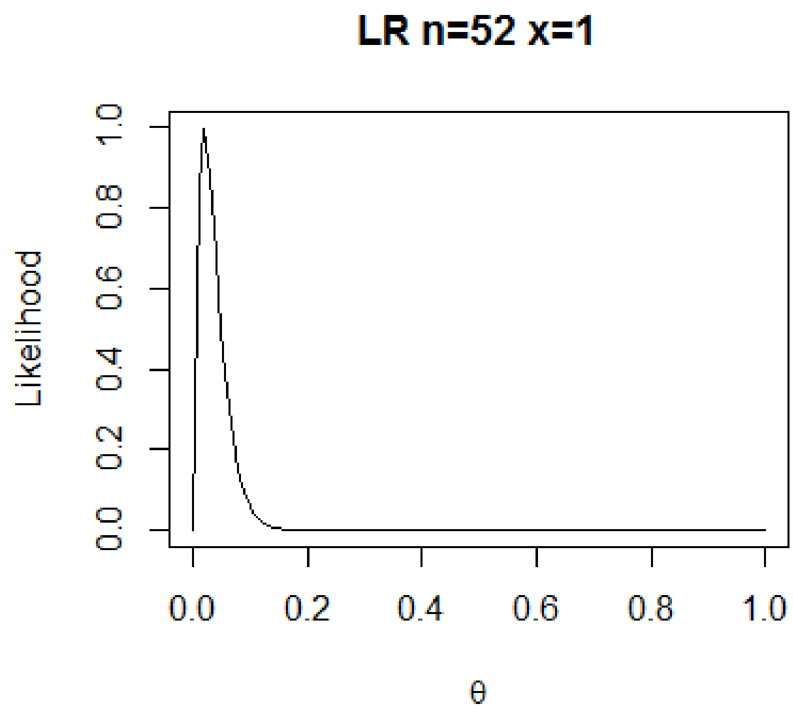
$$LR_2(\theta) = \frac{\binom{n-1}{x-1}\theta^x(1-\theta)^{n-x}}{\binom{n-1}{x-1}\hat{\theta}^x(1-\hat{\theta})^{n-x}}$$

More importantly $LR_1(\theta) = LR_2(\theta)$ because it is clear that $\hat{\theta} = \frac{x}{n}$ in both cases. The same outcomes lead to the same likelihood regardless of intended sampling scheme. Thus we have the same information about θ whether we observed n successes in a fixed number of trials or ran trials until we observed n successes. This is an illustration of the likelihood principle. Because of the likelihood principle there is no extra difficulty in quantifying data evidence at multiple points in the sampling. Let's illustrate the available evidence as the experiment progresses.

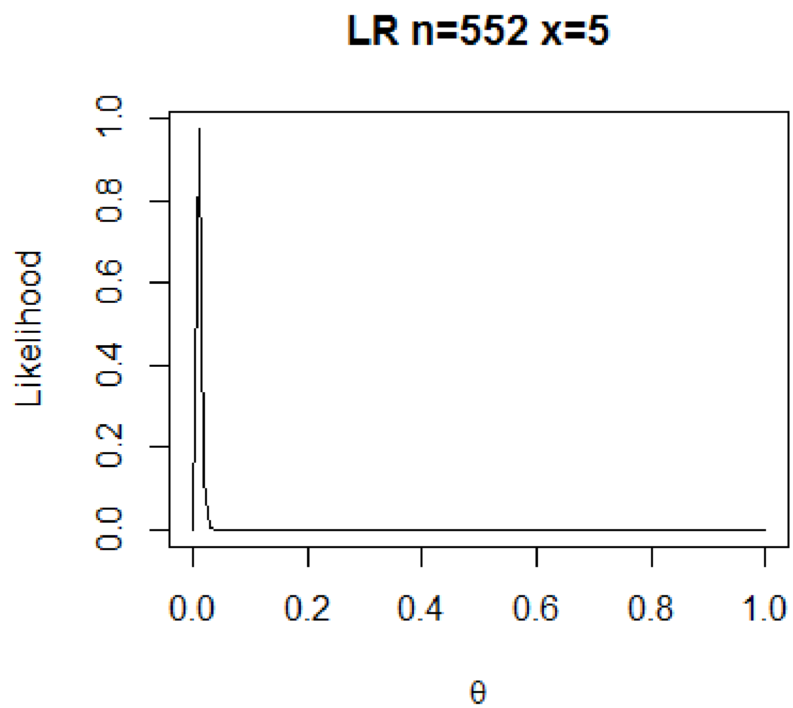
After the first 50 trials, no successes have been obtained. The likelihood function is as shown below. At this point, there is strong evidence pointing to a very small value for θ.



The first success is observed at trial 52. Now, the likelihood has its maximum at a positive value. The likelihood displays the data evidence shown as:



After 552 trials, we have observed 5 successes. At this point, the data evidence is very strong as shown in the likelihood function:



This case is a strong motivator that the likelihood function is an appropriate way to draw inferences from data. The inclusion of multiple views of the data is very difficult to handle under a frequentist framework, because it requires making further findings conditional on the previously observed data. Two experimenters running the same experiment can then get differing results due to the fact that one viewed the data half way through and the other did not. To frequentists, whether a result is significant may depend on the design of the experiment. Under the likelihood principle the inference is not influenced by design[4].

In the next example, we consider a two-parameter model.

Example 2.7. In this case our focus is on estimation of the mean of a normal distribution with an unknown variance σ^2. A parameter that is unknown, but not of interest, is called a nuisance parameter. Consider random variables $X_1, ..., X_n \overset{iid}{\sim} N(\theta, \sigma^2)$. The likelihood function becomes

[4]Note this is true for *most* designs. A design must be "ignorable" as Rubin defined in his 1976 and 1978 papers. A design is said to be ignorable if data not seem by the experimenter is missing at random.

$$L(\theta, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{\frac{-\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\}$$

Write

$$\sum_{i=1}^n (x_i-\theta)^2 = \sum_{i=1}^n (x_i - \bar{x} + (\bar{x}-\theta))^2 = \sum_{i=1}^n ((x_i-\bar{x})^2 + n(\bar{x}-\theta)^2)$$

So,

$$L(\theta, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{\frac{-\sum_{i=1}^n (x_i-\bar{x})^2}{2\sigma^2}\}\exp\{\frac{-\sum_{i=1}^n n(\bar{x}-\theta)^2}{2\sigma^2}\}$$

The MLEs are easily found to be

$$\hat{\theta} = \bar{x}$$
$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n}$$

We are looking for a way to display the likelihood function over the parameter space on the mean θ alone. The question is what to do with σ^2. First we will introduce the plug in method. In this method we will take the estimate $\hat{\sigma}^2$ as σ^2 itself. We will "plug in" $\hat{\sigma}^2$ for σ^2.

$L_{pi}(\theta)$
$=\mathrm{L}(\theta, \hat{\sigma}^2)$
$=(2\pi)^{-n/2}(\hat{\sigma}^2)^{-n/2}\exp\{\frac{-\sum_{i=1}^n (x_i-\bar{x})^2}{2\hat{\sigma}^2}\}\exp\{\frac{\sum_{i=1}^n -n(\bar{x}-\theta)^2}{2\hat{\sigma}^2}\}$
$=(2\pi e)^{-n/2}(\hat{\sigma}^2)^{-n/2}e^{-n/2}\exp\{\frac{\sum_{i=1}^n -n(\bar{x}-\theta)^2}{2\hat{\sigma}^2}\}$

The plug-in likelihood ratio is then

$$LR_{pi}(\theta) = L_{pi}(\theta)/L_{pi}(\hat{\theta}) = \exp\{\frac{-n(\bar{x}-\theta)^2}{2\hat{\sigma}^2}\}$$

The plug-in method overstates evidence near $\hat{\theta}$ since the curvature is based as if the true variance were known. Let us examine an approach called the profile likelihood. In this method we will write the nuisance parameter as a function of the parameter of interest.

Define the profile likelihood as follows:

$$LR_{pr}(\theta) = L_{pr}(\theta, \sigma^2_\theta)$$

where for fixed θ we can maximize

$$L(\theta, \sigma^2) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\{\frac{-\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}\}$$

at

$$\sigma_\theta^2 = \frac{\sum_{i=1}^n (x_i - \theta)^2}{n}$$

denoted as

$$\sigma_\theta^2 = \underset{\sigma^2}{Argmax}(L(\theta, \sigma^2))$$

So,

$$L_{pr}(\theta) = (2\pi)^{-n/2}(\sigma_\theta^2)^{-n/2} \exp\{\frac{-\sum_{i=1}^n (x_i - \theta)^2}{2\sigma_\theta^2}\}$$

Since

$$\sum_{i=1}^n (x_i - \theta)^2 = n\sigma_\theta^2$$

we have

$$L_{pr}(\theta) = (2\pi e)^{-n/2}(\sigma_\theta^2)^{-n/2}$$

We can write

$$\sigma_\theta^2 = \frac{\sum_{i=1}^n ((x_i - \bar{x})^2 + (\bar{x} - \theta)^2)}{n}$$

$$= \hat{\sigma}^2 + (\bar{x} - \theta)^2$$

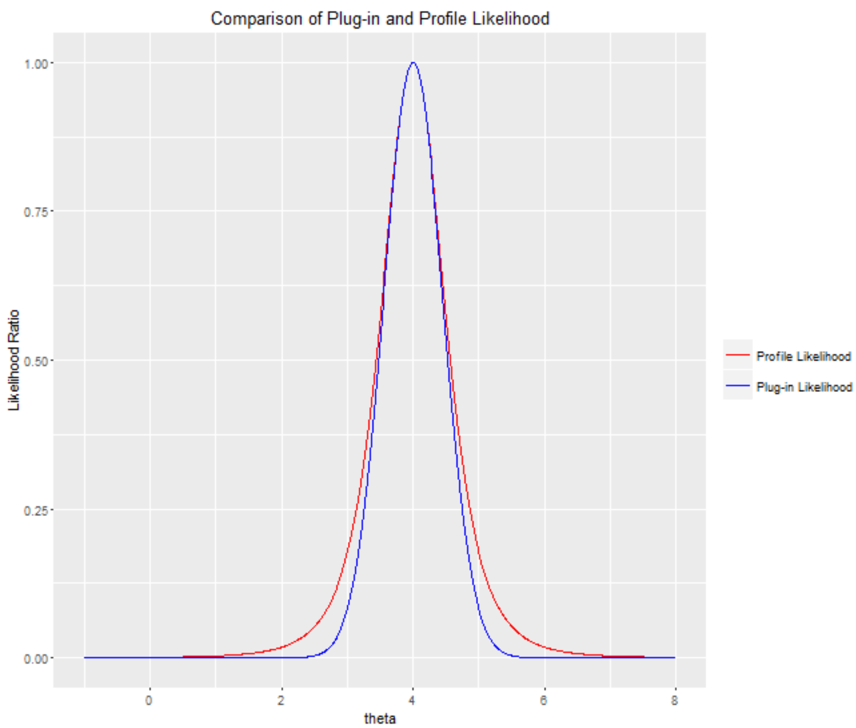
We can see $L_{pr}(\theta)$ is maximized when σ_θ^2 is minimized. By inspection, σ_θ^2 is minimized when $\theta = \hat{\theta} = \bar{x}$. Furthermore,

$$\sigma_{\hat{\theta}}^2 = \hat{\sigma}^2 + (\bar{x} - \hat{\theta})^2 = \hat{\sigma}^2$$

The profile likelihood ratio can be written as

$$LR_{pr}(\theta) = \frac{L_{pr}(\theta)}{L_{pr}(\hat{\theta})} = (\frac{\sigma_\theta^2}{\sigma_{\hat{\theta}}^2})^{-n/2} = (\frac{\hat{\sigma}^2 + (\bar{x} - \theta)^2}{\hat{\sigma}^2})^{-n/2} = (1 + (\frac{\bar{x} - \theta}{\hat{\sigma}})^2)^{-n/2}$$

Let us look at a short example of how the plug-in method overstates evidence near $\hat{\theta}$. Consider a sample $X_1, ..., X_n \overset{iid}{\sim} N(\theta, \sigma^2)$ resulting in $n = 10$, $\bar{x} = 4$, and $\hat{\sigma} = 1$.



This graph demonstrates that the the plug-in likelihood is assigning higher evidence near the MLE.

We will now establish an analytical connection between the plug-in likelihood and the profile likelihood.

Let

$$t(\theta) = \frac{\sqrt{n}(\bar{x} - \theta)}{\hat{\sigma}}.$$

Note that $t(\theta_0)$ is the standardized statistic for testing the null hypothesis $H_0 : \theta = \theta_0$. In general, $t(\theta)$ represents the difference between observed data and a parameter value θ. Therefore,

$$LR_{pi}(\theta) = \exp\{\frac{-t^2(\theta)}{2}\}$$

and

$$LR_{pr}(\theta) = (1 + \frac{t^2(\theta)}{n})^{-n/2}$$

So,

$$LR_{pr} \to LR_{pi}(\theta) \text{ as } n \to \infty$$

since

$$\lim_{n\to\infty}(1 + \frac{a}{n})^{-bt} = e^{-bt}$$

and by extension

$$\lim_{n\to\infty}(1+\frac{t^2}{n})^{-n/2}=e^{-t^2/2}$$

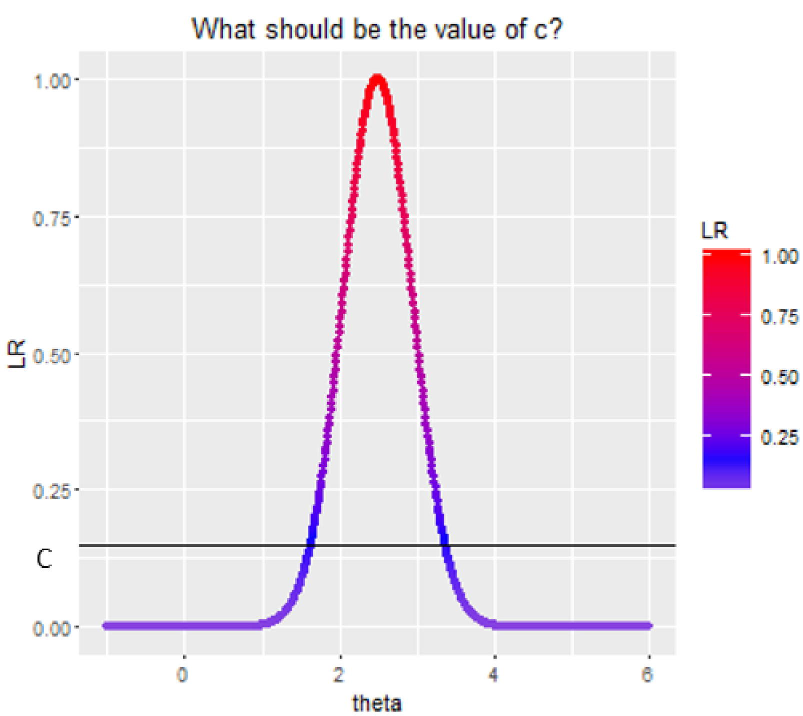
The profile likelihood takes the most conservative approach in defining likelihood across the nuisance parameter. Since $LR_{pr}(\theta)$ is maximized across σ^2, the level of evidence in the likelihood away from the MLE $\hat{\theta}$ decreases at the slowest rate. For large n, the overstatement of evidence in taking a nuisance parameter as previously known diminishes.

3 Likelihood Intervals

In the previous section we made a case for the use of the likelihood function as a measure of statistical evidence beyond simply its MLE. One way we made this case was through the use of a graph as a summary of the known evidence. However, a numerical summary may be preferred. We will now investigate how we can summarize the information from the likelihood function as an interval. Define a set of parameter values as follows:

$$\{\theta : LR(\theta) > c\}$$

Such an interval includes all parameter values having data support above a specified cut-off c.



The question of how to choose the cut-off point is to be investigated. To find a solution we will look to cases where a frequentist confidence interval is seen to match a likelihood interval. Recall that a confidence interval can be derived through the inversion of a hypothesis test. We test $H_0 : \theta = \theta_0$ at level α based on Wilks' likelihood ratio statistic:

$$W(\theta_0) = -2log(\frac{L(\theta_0)}{L(\hat{\theta})}) = -2log(LR(\theta_0))$$

Large sample theory establishes $W \sim \chi^2$ when the null hypothesis is true. So our test accepts H_0 if and only if

$$W(\theta_0) \leq \chi^2_{\alpha,1}$$

where $\chi^2_{\alpha,1}$ is the upper α^{th} percentile of a χ^2_1.

By inverting the test to include all parameters θ accepted under a level α test, we find that a $(1 - \alpha)100\%$ confidence interval for θ is given by

$$\{\theta : W(\theta) \leq \chi^2_{\alpha,1}\}$$

The derivation of the likelihood ratio in terms of $\chi^2_{\alpha,1}$ follows directly. The inequality

$$W(\theta) \leq \chi^2_{\alpha,1}$$

along with

$$W(\theta) = -2log(LR(\theta)).$$

together imply

$$LR(\theta) \geq e^{-\chi^2_{\alpha,1}/2}$$

Thus a frequentist interval matches the likelihood interval when

$$c = e^{-\chi^2_{\alpha,1}/2}$$

For example, a 95% confidence interval would give a likelihood ratio cut-off of $c \approx .15$.

Example 3.1. Let $X \sim BIN(n, \theta)$ and observe $n = 100$, $x = 80$. The normal approximation of the binomial gives us

$$\hat{\theta} \approx N(\theta, \frac{\hat{\theta}(1 - \hat{\theta})}{n})$$

A 95% asymptotic confidence interval for θ is given by

$$\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}} = [.72, .88]$$

We can solve [5] $LR(\theta) = .15$ to find our likelihood interval. In our case we have

$$\frac{\theta^{80}(1-\theta)^{20}}{(.8)^{80}(1-.8)^{20}} = .15$$

This gives an interval of $[.72, .87]$.

As we can see our traditional confidence interval is very close to our likelihood interval. Likelihood intervals will closely match confidence intervals found with more exact methods as well. Consider the genotype example from example 1.6.

Example 3.2. Let $X \sim BIN(n, \theta)$. Observe $n = 552$ and $x = 5$. A 95% frequentist confidence interval can be found to be $[.0029, .0210]$ using the Clopper-Pearson method[6]. We find the likelihood interval $[\theta_l, \theta_u]$ by solving $LR(\theta) = .15$. Our likelihood interval becomes $[.0033, .0193]$.

In examples 2.1 and 2.2 we see that likelihood intervals are very close to the frequentst intervals.

However, the likelihood intervals can give a summary of data in a situation where only the rather unwieldy Clopper-Pearson formula would give a result as we will see in the following example.

Example 3.3. Let $X \sim BIN(n, \theta)$. Observe $n = 50$ and $x = 0$. Here we can find a likelihood interval in a situation where a normal approximation the frequentist interval would fail. Clearly, normal distribution asymptotics do not hold here. Solving

$$LR(\theta) = .15$$

gives us a likelihood interval of $[0, .037]$. An exact interval found by the Clopper-Pearson formula is $[0, .058]$.

In summary, the likelihood function allows us to closely approximate normal confidence intervals in cases where the normal approximation is appropriate. However, the likelihood approach works even in cases where frequentist intervals require complicated techniques such as the Clopper-Pearson method. Likelihood intervals are removed from the typical issues involving confidence intervals and are a good tool for summarizing data. Interval estimation is a meaningful form of statistical inference.

We will now present an example where a frequentist interval is computable, but easy to misinterpret.

Example 3.4. Consider random variables $X_1, ..., X_n \overset{iid}{\sim} N(\delta, 1)$. We are interested in a one-sided hypothesis test $H_0 : \delta = 0, H_a : \delta > 0$. Specifically, we

[5] Under a set of regularity conditions $LR(\theta)$ will be monotone increasing on an interval $[0, MLE]$ and monotone decreasing on an interval $[MLE, \infty]$. Given $c \in [0, MLE]$, the Intermediate Value Theorem guarantees the existence and uniqueness of $[\theta_l, \theta_u]$.

[6] Clopper-Pearson found the exact frequentist interval is

$$[(1 + \frac{n-x+1}{xF[1-.5\alpha; 2x, 2(n-x+1)]})^{-1}, (1 + \frac{n-x}{(x+1)F[.5\alpha; 2(x+1), 2(n-x)]})^{-1}]$$

want a $(1-\alpha)100\%$ confidence interval for δ as a companion result. Let $\hat{\delta} = \bar{x}$. Then $\hat{\delta} \sim N(\delta, \frac{1}{n})$. We can think of δ as representing an effect size for some comparison. For δ unrestricted, a 95% confidence interval for δ is seen to be

$$\bar{X} \pm 1.96 \frac{1}{\sqrt{n}}$$

Let $L = \bar{X} - 1.96\frac{1}{\sqrt{n}}$ and $U = \bar{X} + 1.96\frac{1}{\sqrt{n}}$. By the construction of confidence intervals

$$P_\delta[L \leq \delta \leq U] = .95$$

holds for all δ. For the problem where δ is restricted to non negative values, define the truncated interval as

$$CI^* = \begin{cases} [L, U] & \text{if } L > 0 \\ [0, U] & \text{if } L < 0 < U \\ \emptyset & \text{if } U < 0 \end{cases}$$

Then

$$P_\delta[\delta \in CI^*] = .95$$

holds for all $\delta \geq 0$, since CI^* truncates only negative values of δ. So, CI^* is a 95% confidence interval for δ satisfying frequentist properties.

The likelihood function for the data in this problem becomes

$$L(\delta) = (2\pi)^{\frac{-1}{2}} \sqrt{n} \exp\{\frac{-n}{2}(\bar{x} - \delta)^2\}, \delta \geq 0$$

If $\bar{x} > 0$, then $\hat{\delta} = \bar{x}$ maximizes the likelihood and

$$LR(\delta) = \exp\{\frac{-n}{2}(\delta - \bar{x})^2\}, \delta \geq 0$$

If $\bar{x} < 0$, then $\hat{\delta} = 0$ maximizes the likelihood and

$$LR(\delta) = \frac{\exp\{\frac{-n}{2}(\delta - \bar{x})^2\}}{\exp\{\frac{-n}{2}(\bar{x})^2\}}, \delta \geq 0$$

Let's look at how our likelihood ratio compares to the frequentist confidence interval. For simplicty we will take $n = 1$. First let's see what happens when our sample difference is positive.

Example 3.5. Consider the case where $\bar{x} > 1.96$. Then

$$CI^* = [\bar{x} - 1.96, \bar{x} + 1.96]$$

The likelihood interval is found by solving

$$LR(\delta) = \exp\{-1/2(\delta - \bar{x})^2\} = c$$

Recall a previous result shows

$$c = e^{-\chi^2_{\alpha,1}/2}$$

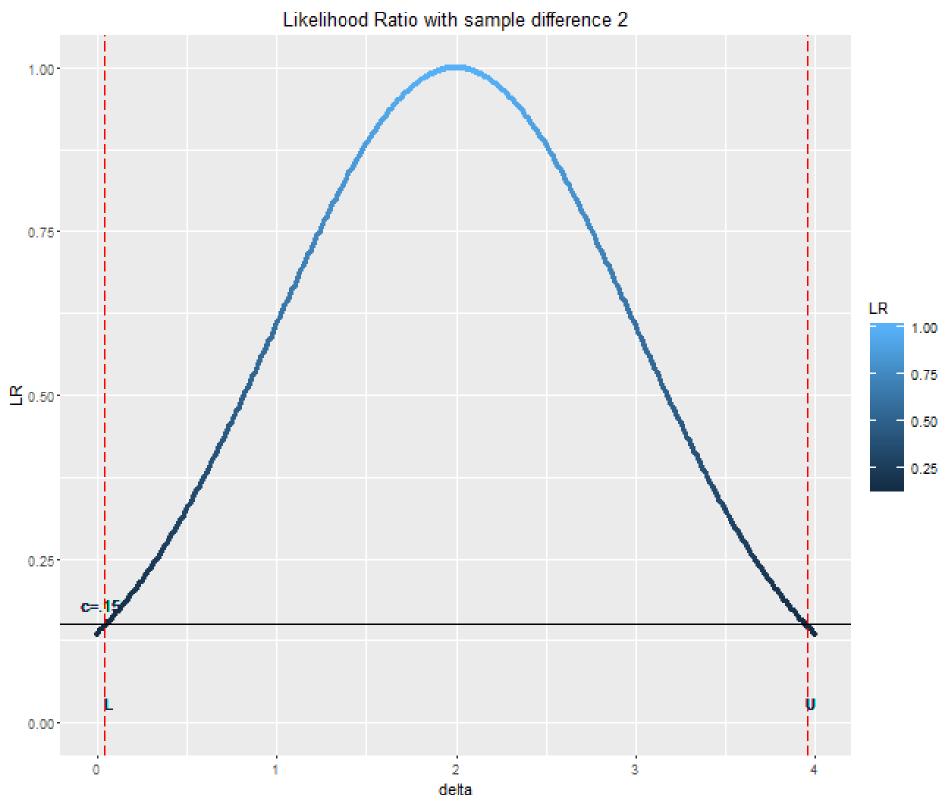
yields an equivalence between the frequentist interval and the likelihood interval. For our problem, solving $LR(\delta) = c$ at the 95% level gives

$$e^{-1/2(\delta-\bar{x})^2} = e^{-1/2(1.96^2)}$$

The likelihood interval is of the form

$$\bar{x}\pm 1.96$$

With $\bar{x} = 2$ we have $L = .4$ and $U = 3.96$.

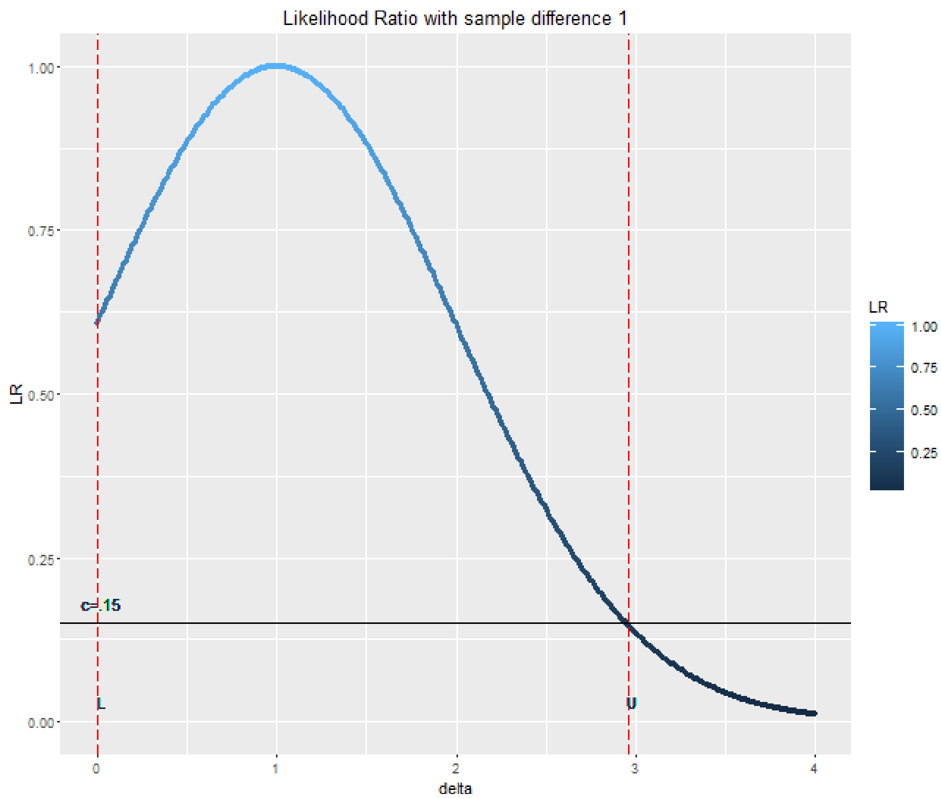


As we can see, the bounds of the frequentist interval denoted L and U respectively align with our companion 95% likelihood interval.

For $0 < \bar{x} < 1.96$, solving $LR(\delta) = c$ as

$$e^{-1/2(\delta-\bar{x})^2} = e^{-1/2(1.96)^2}$$

only gives a solution for the upper bound. The likelihood interval becomes truncated at a lower endpoint as $[0, \bar{x} + 1.96]$. For $\bar{x} = 1$, we have $L = 0$ and $U = 2.96$.



The frequentist CI^* and our likelihood interval are quanitifying the evidence identically.

Example 3.6. Now let's look at the case where $\bar{x} < 0$. For $\bar{x} = -1$, the truncated frequentist interval is given by $[0, .96]$. The likelihood interval is found by solving

$$LR(\delta) = \frac{\exp\{-1/2(\delta - \bar{x})^2\}}{\exp\{-1/2(\bar{x}^2)\}} = c$$

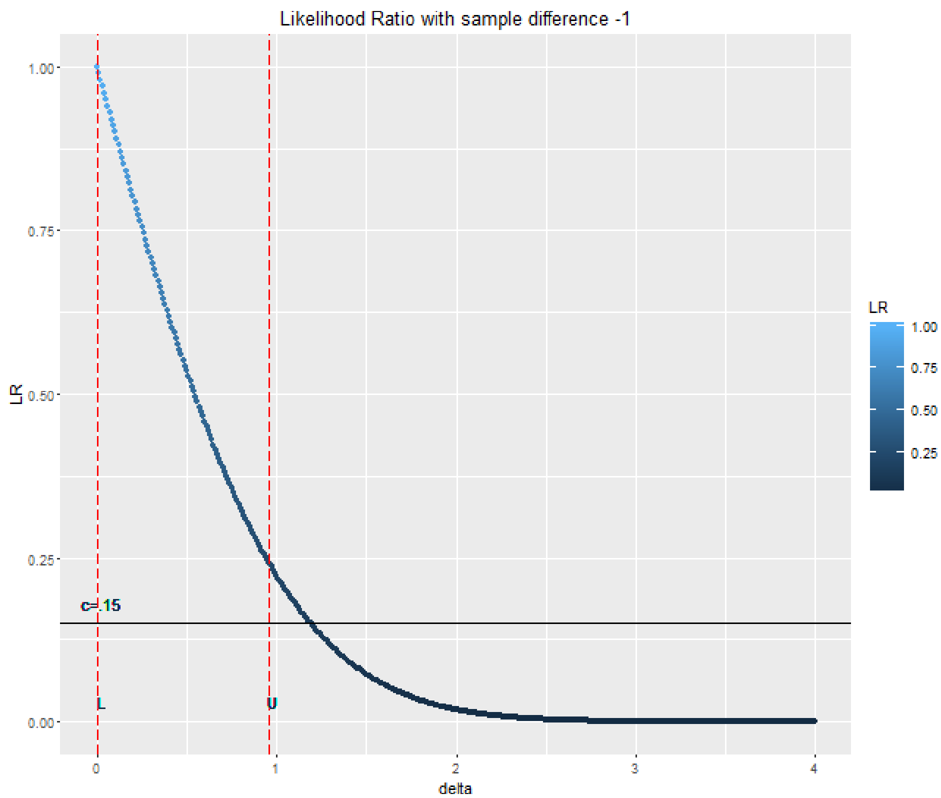
This gives

$$e^{-1/2(\delta - \bar{x})^2} = e^{-1/2(1.96^2 + \bar{x}^2)}.$$

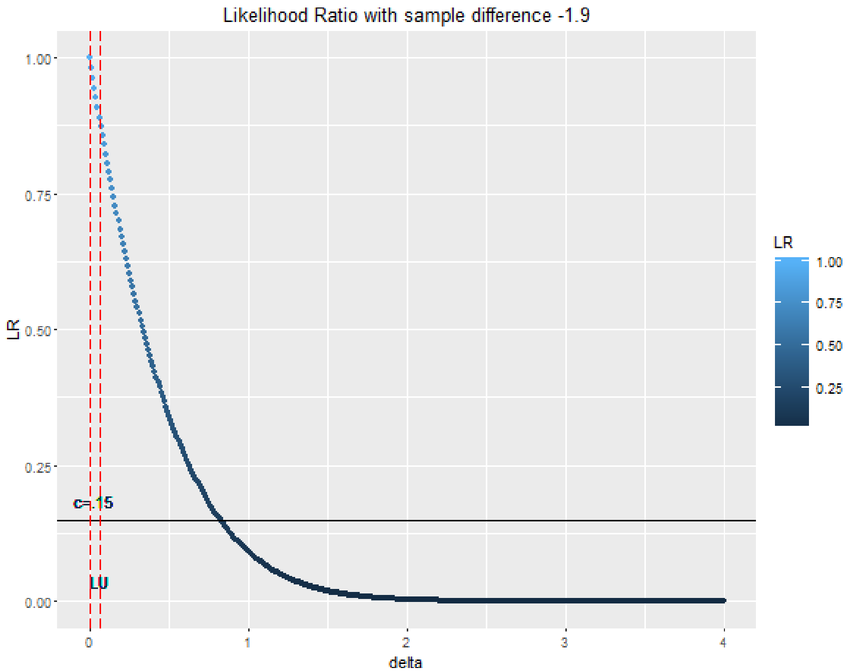
The likelihood interval is of the form

$$[0, \bar{x} + \sqrt{1.96^2 + \bar{x}^2}]$$

For $\bar{x} = -1$, the likelihood interval becomes $[0, 1.2]$. From the following graph we can see the freqentist upper bound U is closer to 0, resulting in an overstatement of evidence near 0.



This situation becomes worse the farther the observed data is in the negative direction. Consider $\bar{x} = -1.9$. Following the same steps as above we arrive at a likelihood interval of $[0, .83]$. The frequentist interval is $[0, .06]$



Here the likelihood approach supports values 13 times larger than the frequentist approach suggests. To summarize, an observed $\bar{x} = -1.9$ is improbable under all values of δ in the parameter space, but not that much more improbable under larger values of δ than for the MLE $\hat{\delta} = 0$. Treating a frequentist based CI as providing a measure of evidence is a fallacy in logic as demonstrated in this example.

4 Bayesian Inference

We have looked at likelihood inference as a method for quantifying data evidence. Unlike a frequentist approach to inference, the likelihood approach is not based on probability. As we have seen, there are examples where a frequentist based approach is misleading as a representation of data evidence. In this final section, we compare likelihood inference to a Bayesian approach.

Let's recall a Bayesian prior,

$$p(\theta),$$

and a sampling distribution

$$L(\theta) = f(x_1, ..., x_n|\theta)$$

leading to a posterior distribution

$$p(\theta|\ \underline{x}) = kp(\theta)L(\theta)$$

where k is a constant of integration depending on $\underline{x} = (x_1, ..., x_n)$. A Bayesian interpretation of the posterior distribution is that of a subjective probability on the location of the parameter θ. The posterior distribution is a combination of historical evidence specified through $p(\theta)$, and data evidence, specified through the likelihood $L(\theta)$.

There is a notable difference between the likelihood approach and the Bayesian approach. Likelihood inference follows the Invariance property. To demonstrate the Invariance property, consider the following example.

Example 4.1. Let $X \sim BIN(n, \theta)$ and observe $n = 100$ and $x = 80$. Suppose that we are interested in the log odds of θ,

$$\Phi = log[\frac{\theta}{1-\theta}]$$

Consider the question, how much more likely is $\theta = .8$ than $\theta = .2$? Since we can solve back for θ as a function of Φ, define

$$L^*(\Phi) = L(\frac{e^{\Phi}}{1+e^{\Phi}})$$

where parameters Φ_1 and Φ_2 correspond to the transformed values of θ_1 and θ_2, respectively. Examine the relative support for θ_1 compared to θ_2.

$$L^*(\Phi_1)/L^*(\Phi_2) = L(\theta_1 = .8)/L(\theta_2 = .2) = \frac{\theta_1^{80}(1-\theta_1)^{20}}{\theta_2^{80}(1-\theta_2)^{20}} = 13.1$$

This illustrates that our evidence didn't change, just our parametrization. This is not so under Bayesian inference. Take a uniform prior on θ

$$p(\theta) = 1$$

Then the prior on Φ is found by transformation methods as

$$p^*(\Phi) = p(\frac{e^{\Phi}}{1+e^{\Phi}})\frac{d}{d\Phi}(\frac{e^{\Phi}}{1+e^{\Phi}}) = \frac{e^{\Phi}}{(1+e^{\Phi})^2}$$

So, a uniform prior on the probability θ transforms to a logistic prior on the log odds. To answer the question with a Bayesian approach,

$$\frac{p(\Phi_1|\ \underline{x})}{p(\Phi_2|\ \underline{x})} = \frac{L(\theta_1)}{L(\theta_2)}\frac{\frac{e^{\Phi_1}}{(1+e^{\Phi_1})^2}}{\frac{e^{\Phi_2}}{(1+e^{\Phi_2})^2}} = \frac{L(\theta_1)}{L(\theta_2)}\frac{e^{\Phi_1}(1+e^{\Phi_2})^2}{e^{\Phi_2}(1+e^{\Phi_1})^2} = 9.98$$

Clearly, the Bayesian approach does not adhere to the Invariance principle.

Under the likelihood approach, parametrization does not affect the evidence provided by the data. This fits our intuition. Although a complete examination of Bayesian inference is beyond the scope of this paper, the result here is that likelihood inference is not simply a subset of Bayesian inference.

5 Conclusion

"We have the duty of formulating, of summarizing, and of communicating our conclusions, in intelligible form, in recognition of the right of other free minds to utilize them in making their own decisions."[7] The aim here has been to demonstrate an alternative to Bayesian and Frequentist inference. The likelihood approach allows for the quantification of evidence in situations where traditional approaches fail or are otherwise unwieldy. The likelihood approach offers a way to view data not contingent on easily misinterpreted notions of confidence, while not going so far as to admit Bayesian priors or lose the Invariance property.

References

- Pawitan, Y. (2013), "In All Likelihood: Statistical Modeling and Inference using Likelihood" Oxford.

- J. Aragon. (1991), "Existance and Uniqueness of the Maximum Likelihood Estimator For The Two-Parameter Negative Binomial Distribution. Stanford University.

Johnson, N.L. and Kotz, S. (1969) Discrete distributions. Houghton Mifflin Company, Boston.

[7]Ronald Fisher