Likelihood and Bayesian Inference

Joe Felsenstein

Department of Genome Sciences and Department of Biology

Bayes' Theorem

Suppose we have related events, B and some other mutually exclusive events A_1, A_2, $A_3, \ldots, A_8$. The probability of B given A_3 (for example) is

$$\text{Prob}\,(A_3 \mid B) \;=\; \frac{\text{Prob}\,(A_3 \text{ and } B)}{\text{Prob}\,(B)}$$

and since it is also true that

$$\text{Prob}\,(B \mid A_3) \;=\; \frac{\text{Prob}\,(A_3 \text{ and } B)}{\text{Prob}\,(A_3)}$$

we can multiply by $\text{Prob}\,(A_3)$ and substitute for $\text{Prob}\,(A_3 \text{ and } B)$ to get

$$\text{Prob}\,(A_3 \mid B) \;=\; \frac{\text{Prob}\,(A_3)\;\text{Prob}\,(B \mid A_3)}{\text{Prob}\,(B)}$$

(Think of B as the data, and the A_i as different hypotheses).

Getting Bayes' Rule

Since the denominator can be rewritten as

$$\text{Prob}\,(B) \;=\; \text{Prob}\,(A_1)\,\text{Prob}\,(B \mid A_1) + \ldots + \;\text{Prob}\,(A_8)\,\text{Prob}\,(B \mid A_8)$$

We can substitute that in to get the final form of Bayes' Rule:

$$\text{Prob}\,(A_3|B) \;=\; \frac{\text{Prob}\,(A_3)\,\text{Prob}\,(B \mid A_3)}{\text{Prob}\,(A_1)\,\text{Prob}\,(B \mid A_1) + \ldots + \;\text{Prob}\,(A_8)\,\text{Prob}\,(B \mid A_8)}$$

What this does is compute the probability of A_3 given that we saw B from the prior probabilities of the A_i and the conditional probabilities of the observed data B given each A_i.

Odds ratio, Bayes' Theorem, maximum likelihood

We start with an "odds ratio" version of Bayes' Theorem: take the ratio of the numerators for two different hypotheses and we get:

D	**the data**
H_1	**Hypothesis 1**
H_2	**Hypothesis 2**
$\mid$	**the symbol for "given"**

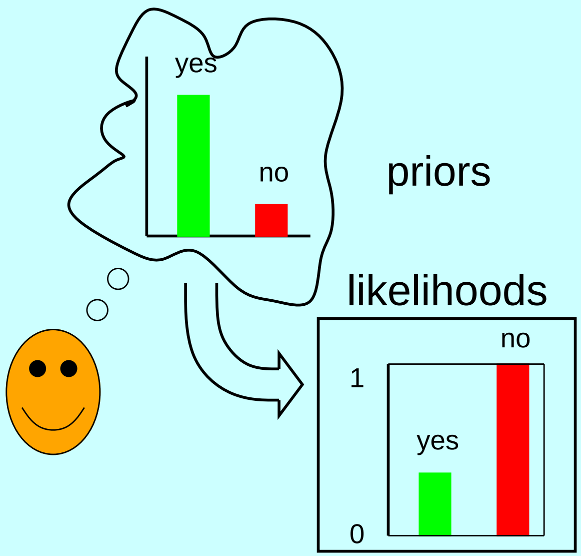
$$\underbrace{\frac{\mathrm{Prob}\,(H_1 \mid D)}{\mathrm{Prob}\,(H_2 \mid D)}}_{\text{Posterior odds ratio}} = \underbrace{\frac{\mathrm{Prob}\,(D \mid H_1)}{\mathrm{Prob}\,(D \mid H_2)}}_{\text{Likelihood ratio}} \; \underbrace{\frac{\mathrm{Prob}\,(H_1)}{\mathrm{Prob}\,(H_2)}}_{\text{Prior odds ratio}}$$

A simple example of Bayes Theorem

If a space probe finds no Little Green Men on Mars, when it would have a 1/3 chance of missing them if they were there:

A simple example of Bayes Theorem

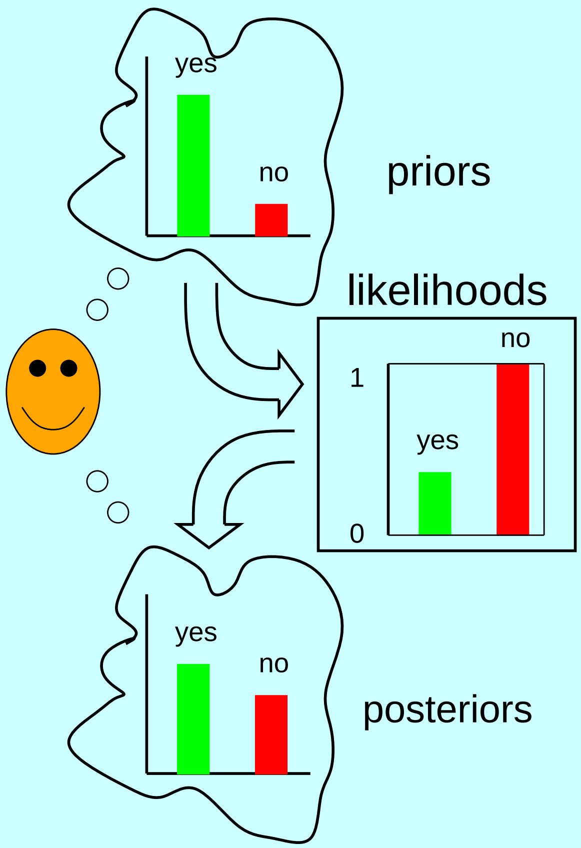
If a space probe finds no Little Green Men on Mars, when it would have a 1/3 chance of missing them if they were there:



$$\frac{4}{1} \times \frac{1/3}{1}$$

A simple example of Bayes Theorem

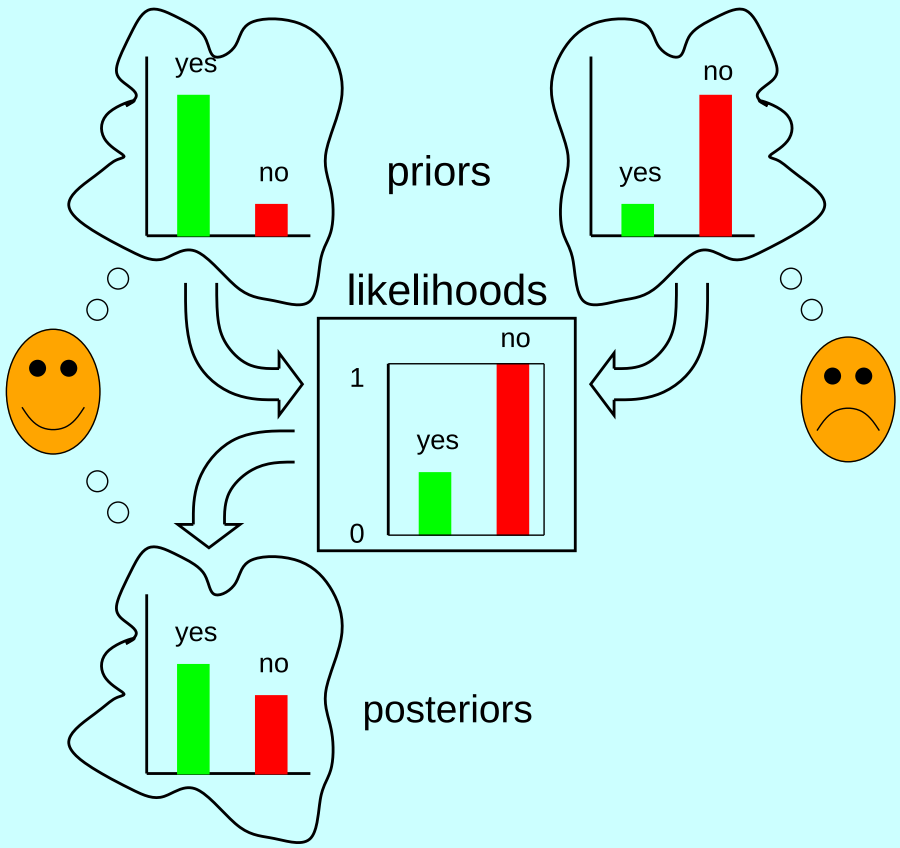
If a space probe finds no Little Green Men on Mars, when it would have a 1/3 chance of missing them if they were there:



$$\frac{4}{1} \times \frac{1/3}{1} = \frac{4}{3}$$

A simple example of Bayes Theorem

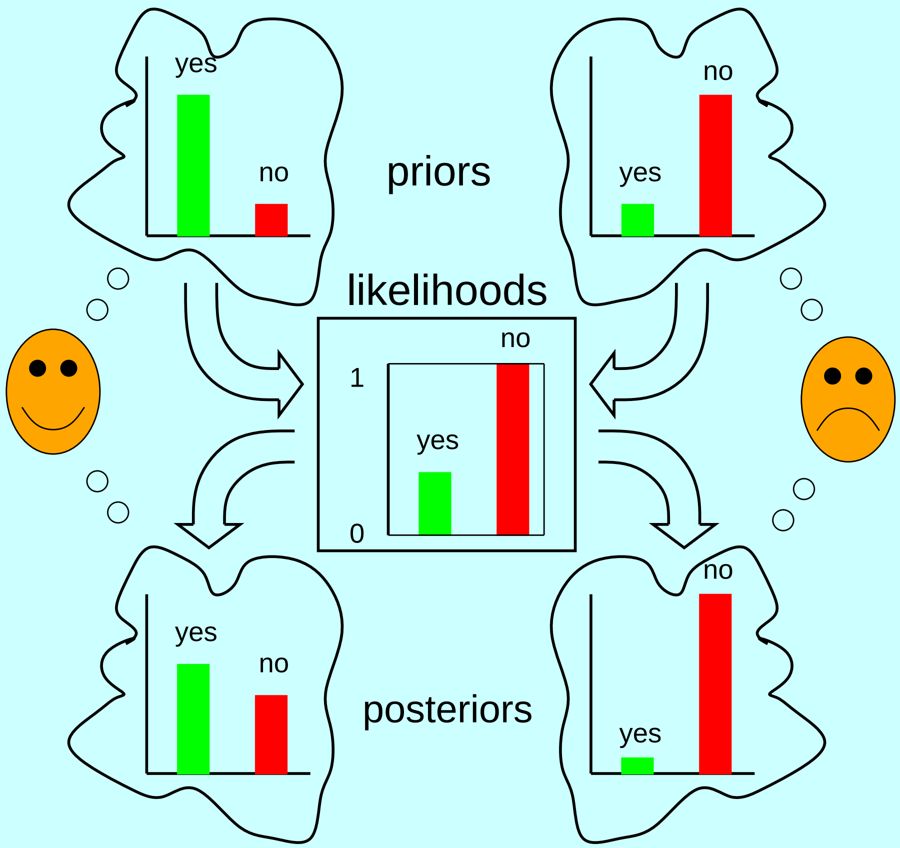
If a space probe finds no Little Green Men on Mars, when it would have a 1/3 chance of missing them if they were there:



$$\frac{4}{1} \times \frac{1/3}{1} = \frac{4}{3} \qquad\qquad \frac{1}{4} \times \frac{1/3}{1}$$

A simple example of Bayes Theorem

If a space probe finds no Little Green Men on Mars, when it would have a 1/3 chance of missing them if they were there:



$$\frac{4}{1} \times \frac{1/3}{1} = \frac{4}{3} \qquad \frac{1}{4} \times \frac{1/3}{1} = \frac{1}{12}$$

The likelihood ratio term ultimately dominates

If we see one Little Green Man, the likelihood calculation does the right thing:

$$\frac{\infty}{1} = \frac{2/3}{0} \times \frac{1}{4}$$

(put this way, this is OK but not mathematically kosher)

If after n missions, we keep seeing none, the likelihood ratio term is

$$\left(\frac{1}{3}\right)^n$$

It dominates the calculation, overwhelming the prior.
Thus even if we don't have a prior we can believe in, we may be interested in knowing which hypothesis the likelihood ratio is recommending ...

Likelihood in simple coin-tossing

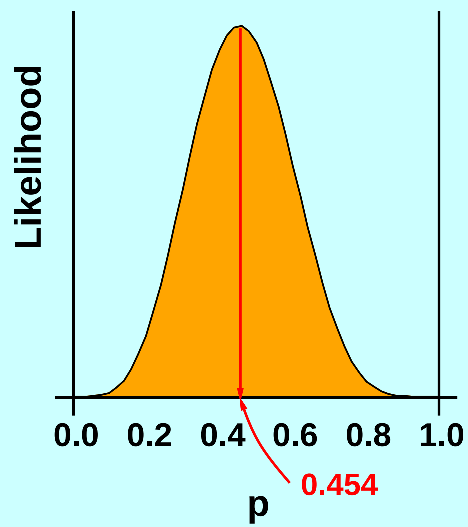
Tossing a coin n times, with probability p of heads, the probability of outcome HHTHTTTTHTTH is

$$pp(1-p)p(1-p)(1-p)(1-p)(1-p)p(1-p)(1-p)p$$

which is

$$L = p^5(1-p)^6$$

Plotting L against p to find its maximum:



Differentiating to find the maximum:

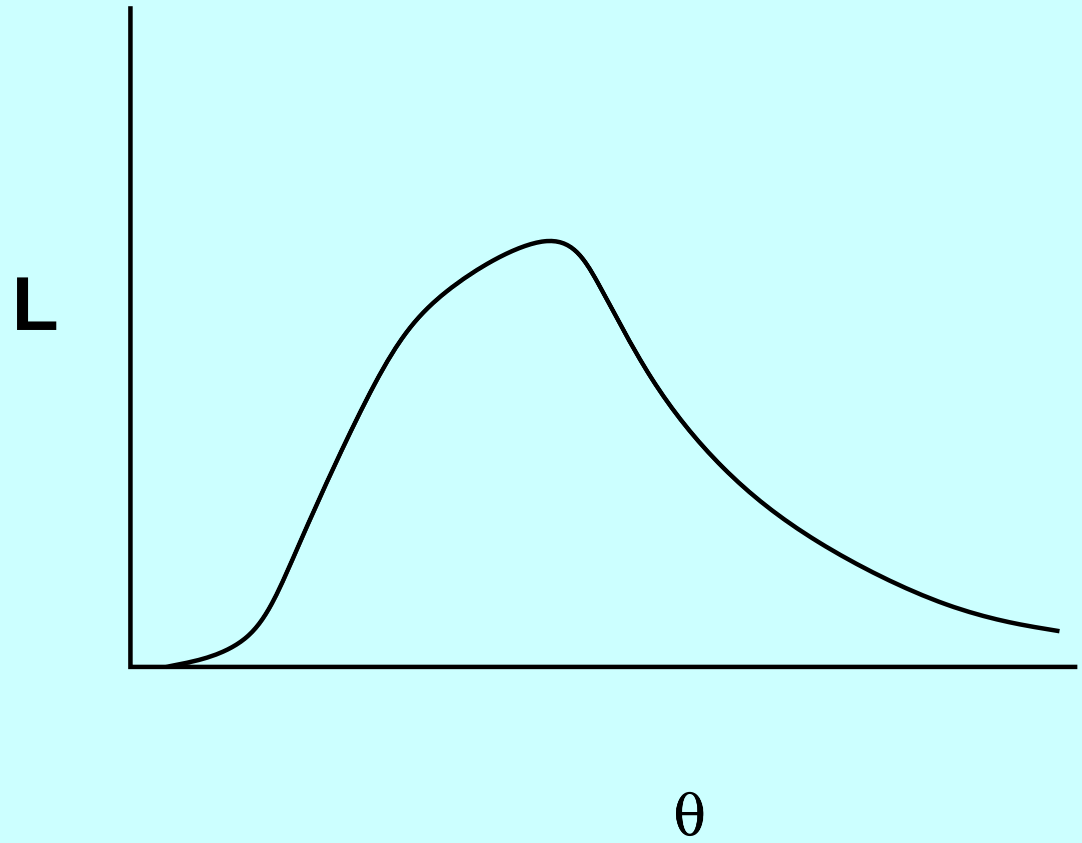
Differentiating the expression for L with respect to p and equating the derivative to 0, the value of p that is at the peak is found (not surprisingly) to be $\mathsf{p} = 5/11$:

$$\frac{\partial L}{\partial p} = \left(\frac{5}{p} - \frac{6}{1-p}\right) p^5 (1-p)^6 = 0$$

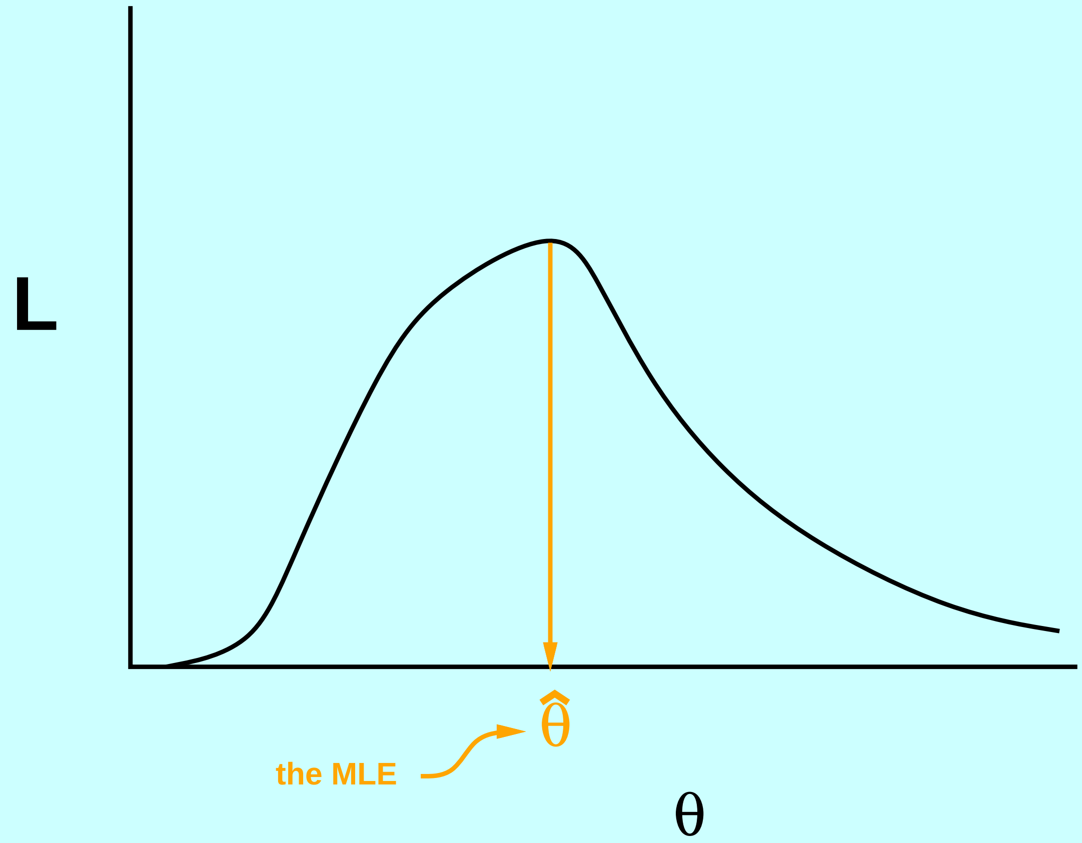
$$5 - 11\,p = 0$$

$$\hat{p} = \frac{5}{11}$$

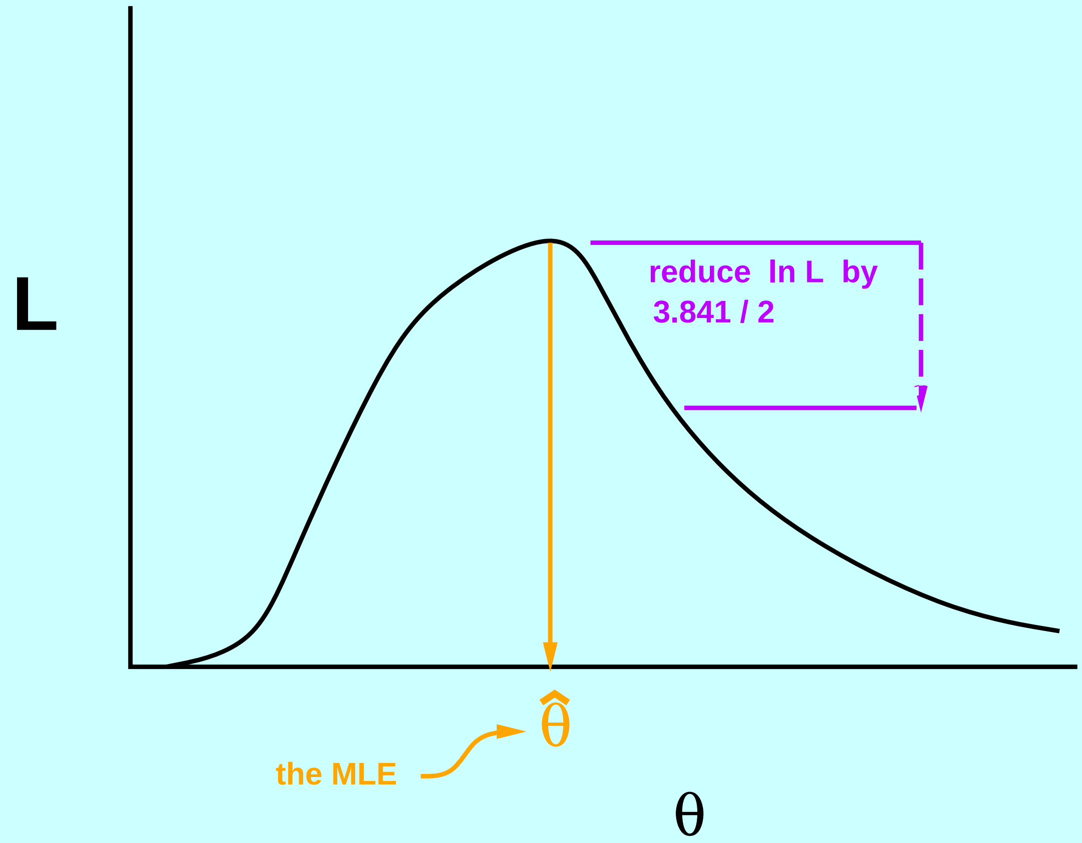
A likelihood curve



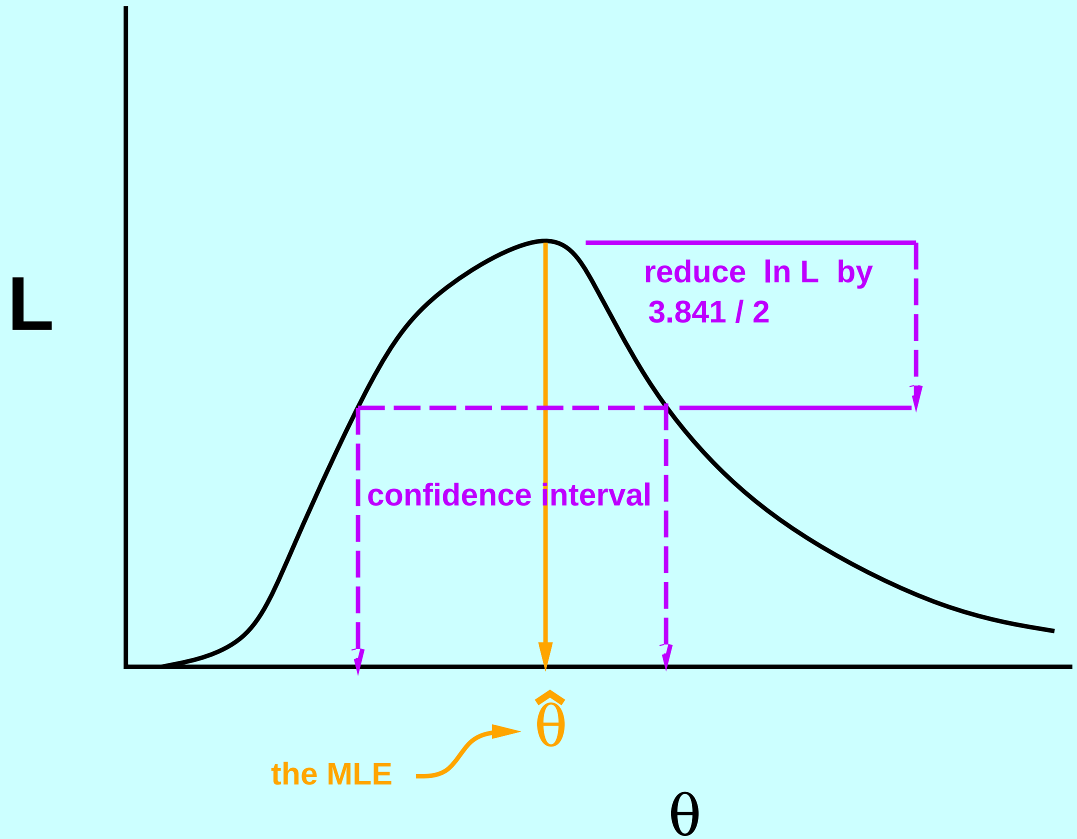
Its maximum likelihood estimate



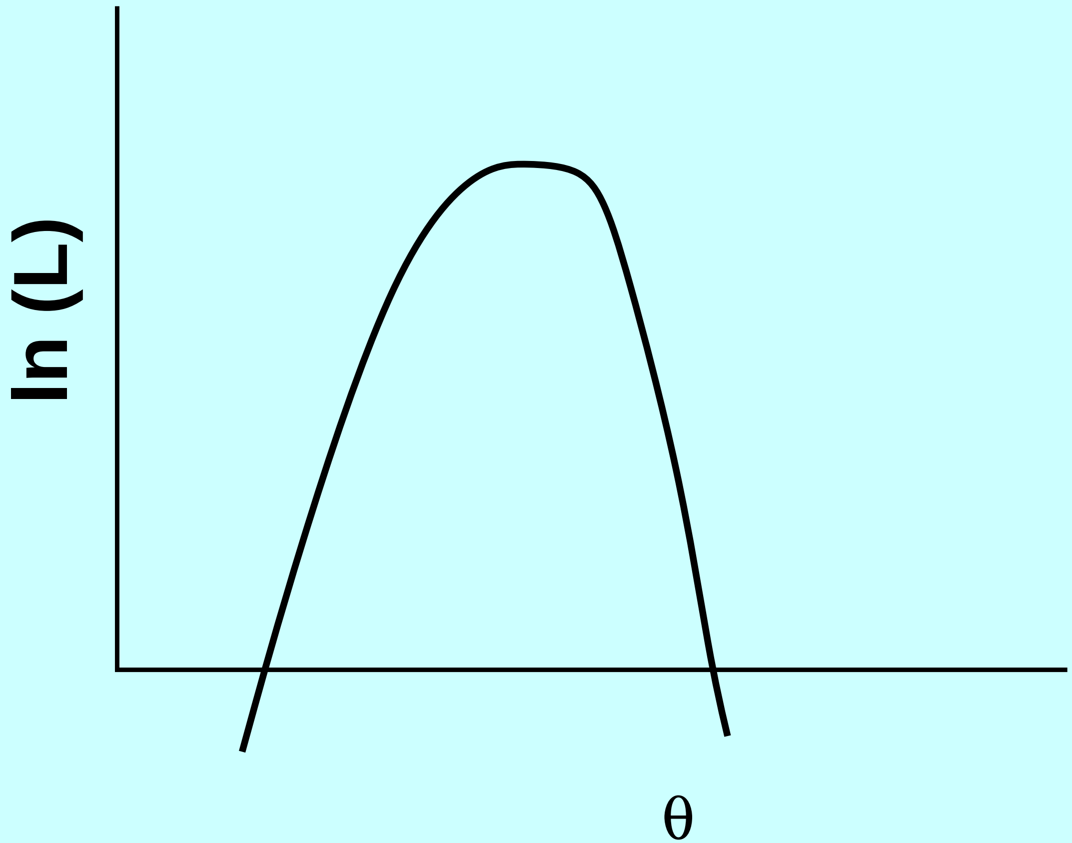
Using the Likelihood Ratio Test



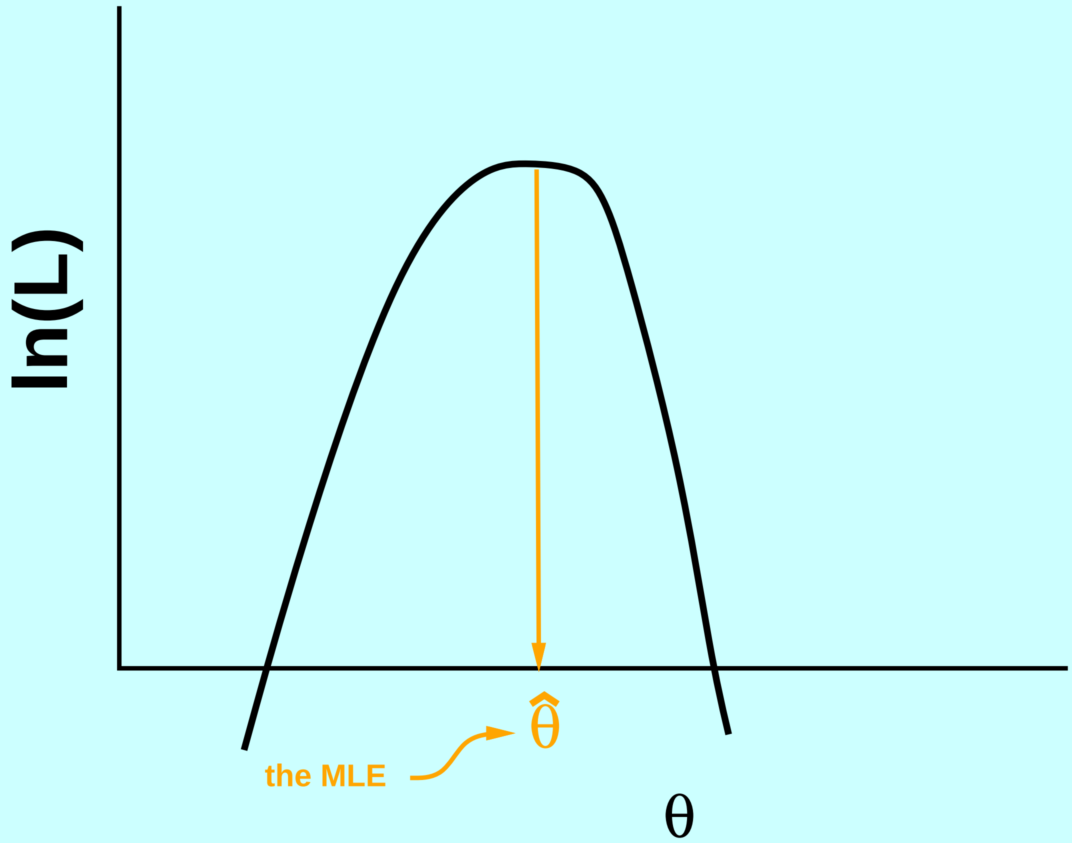
The (approximate, asymptotic) confidence interval



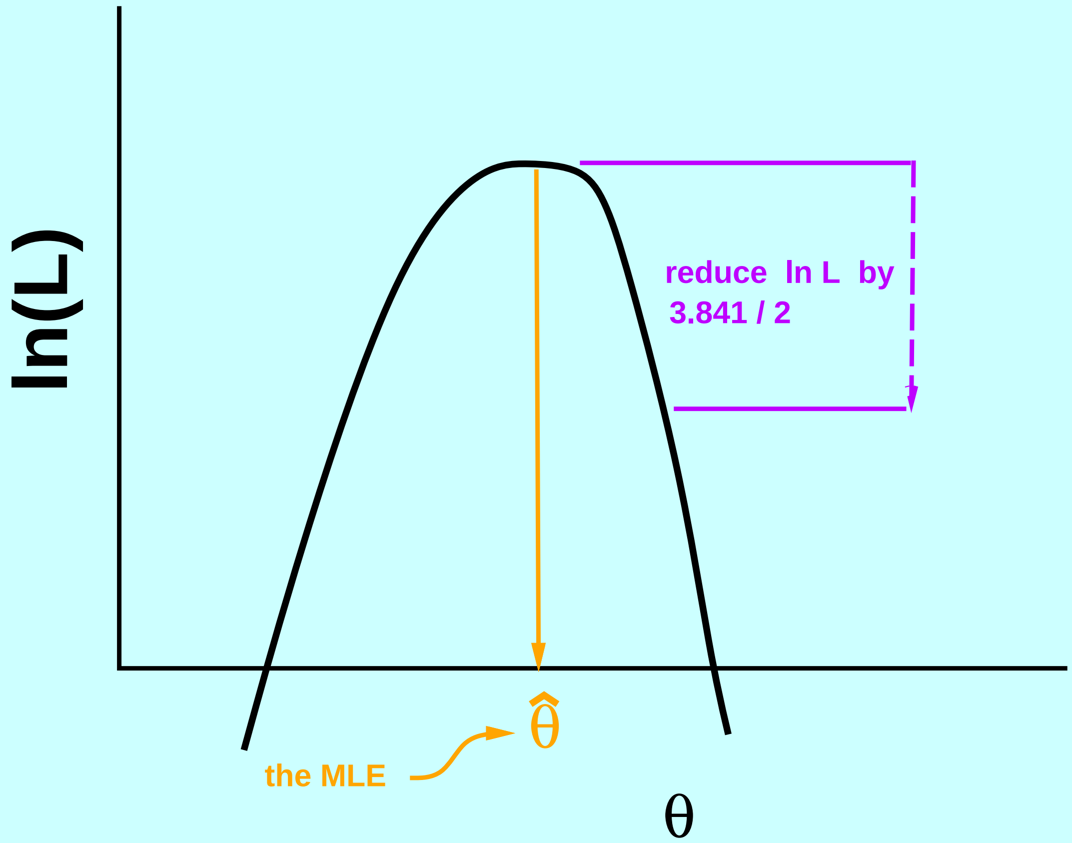
Better to plot log(L) rather than L



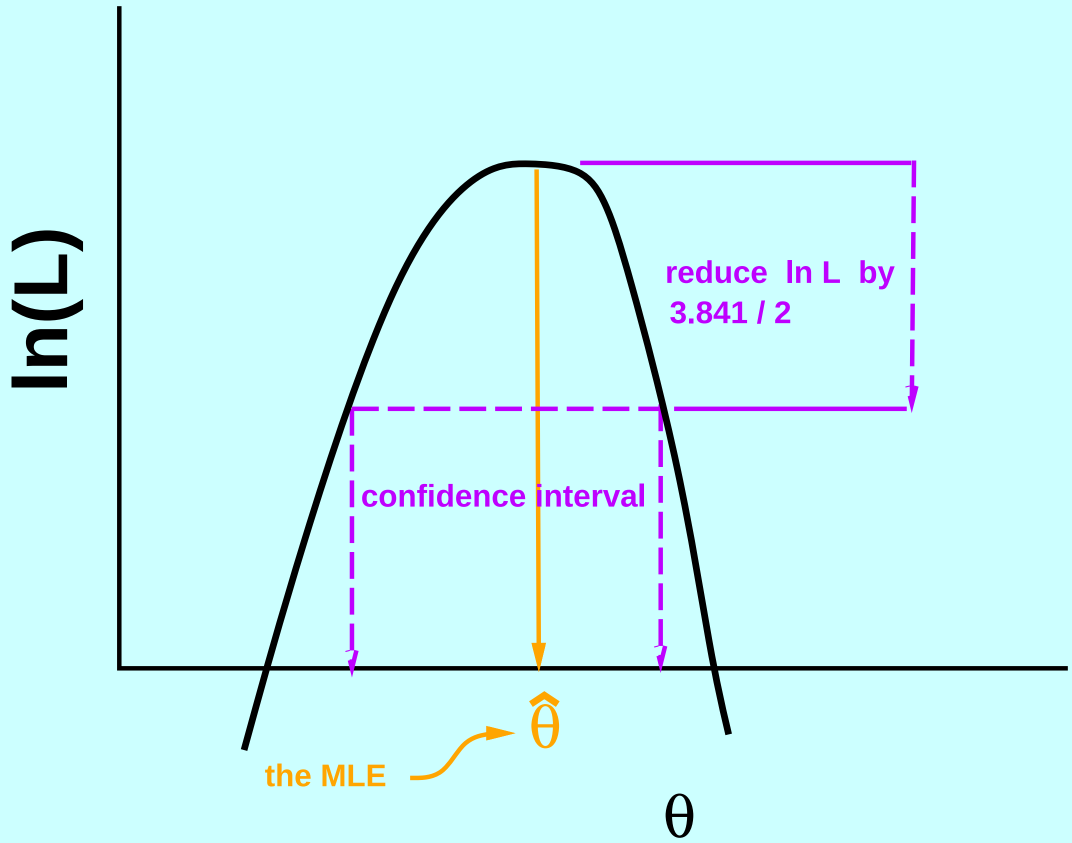
Better to plot log(L) rather than L



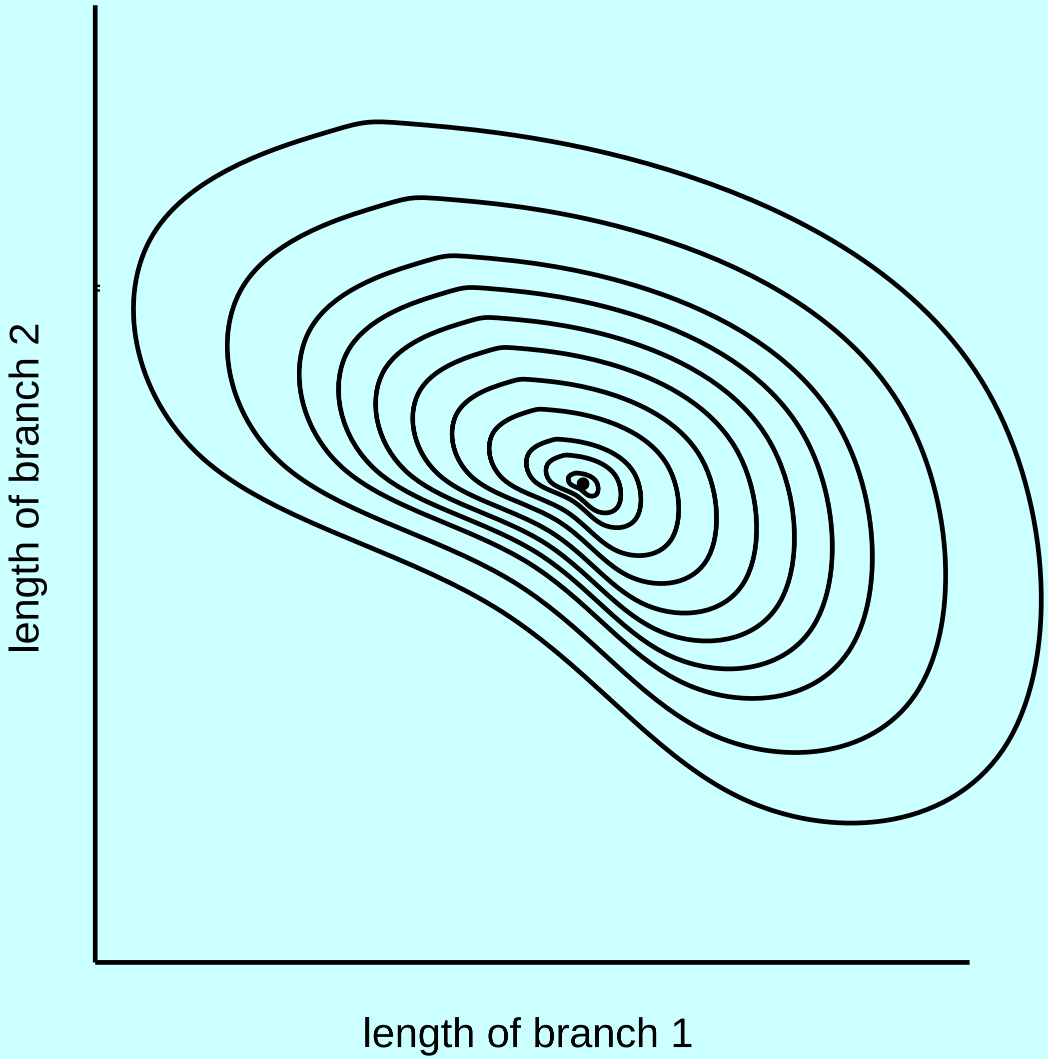
Better to plot log(L) rather than L



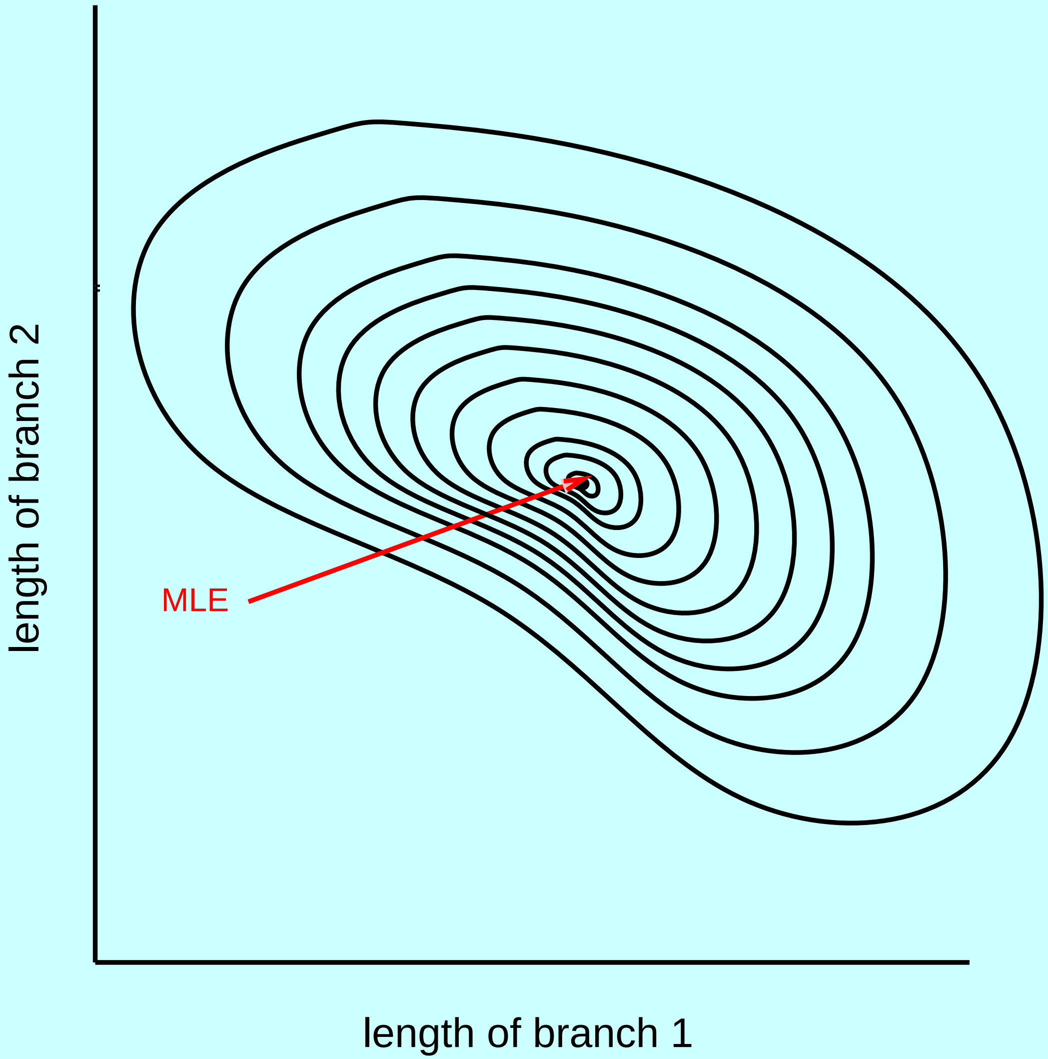
Better to plot log(L) rather than L



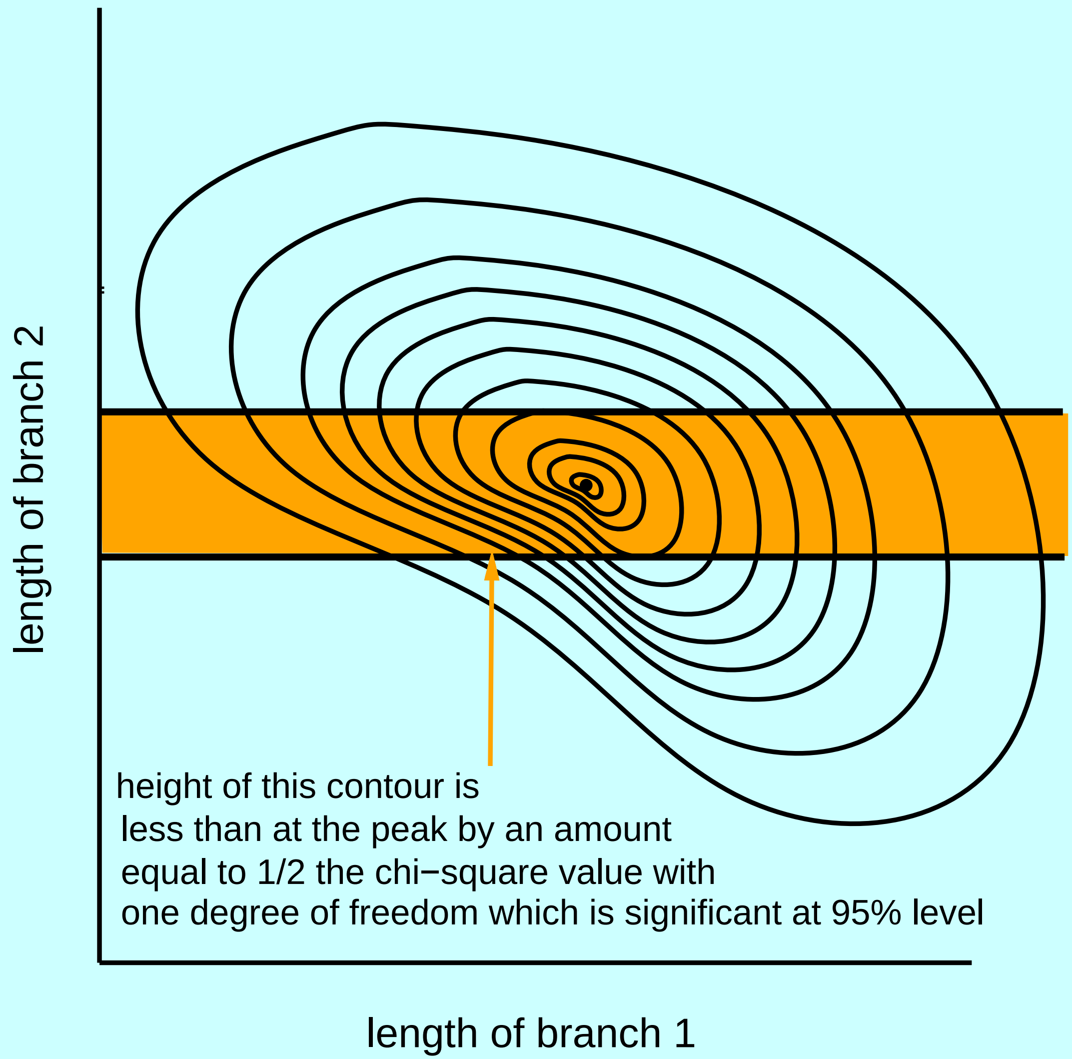
Contours of a likelihood surface in two dimensions



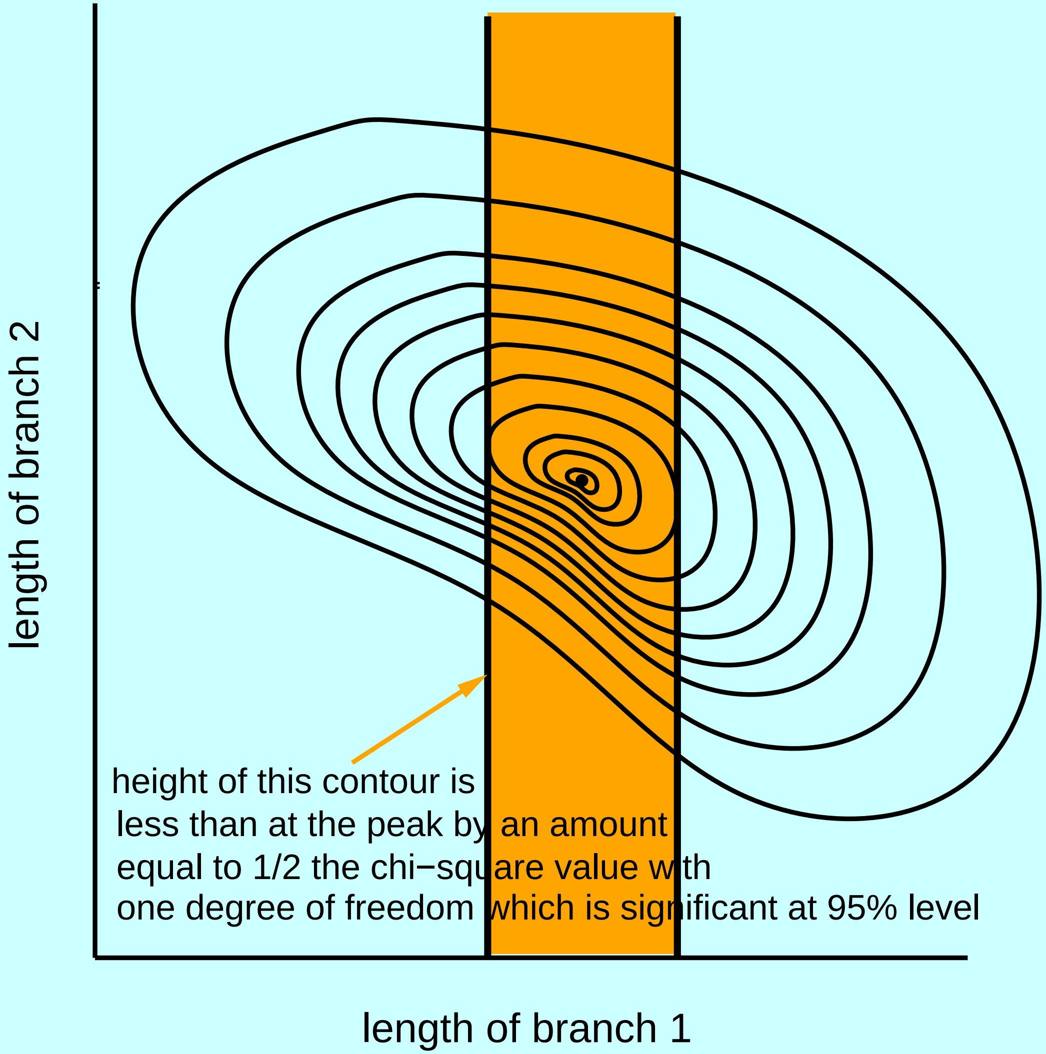
Where the maximum likelihood estimate is



Using the LRT to define a confidence interval

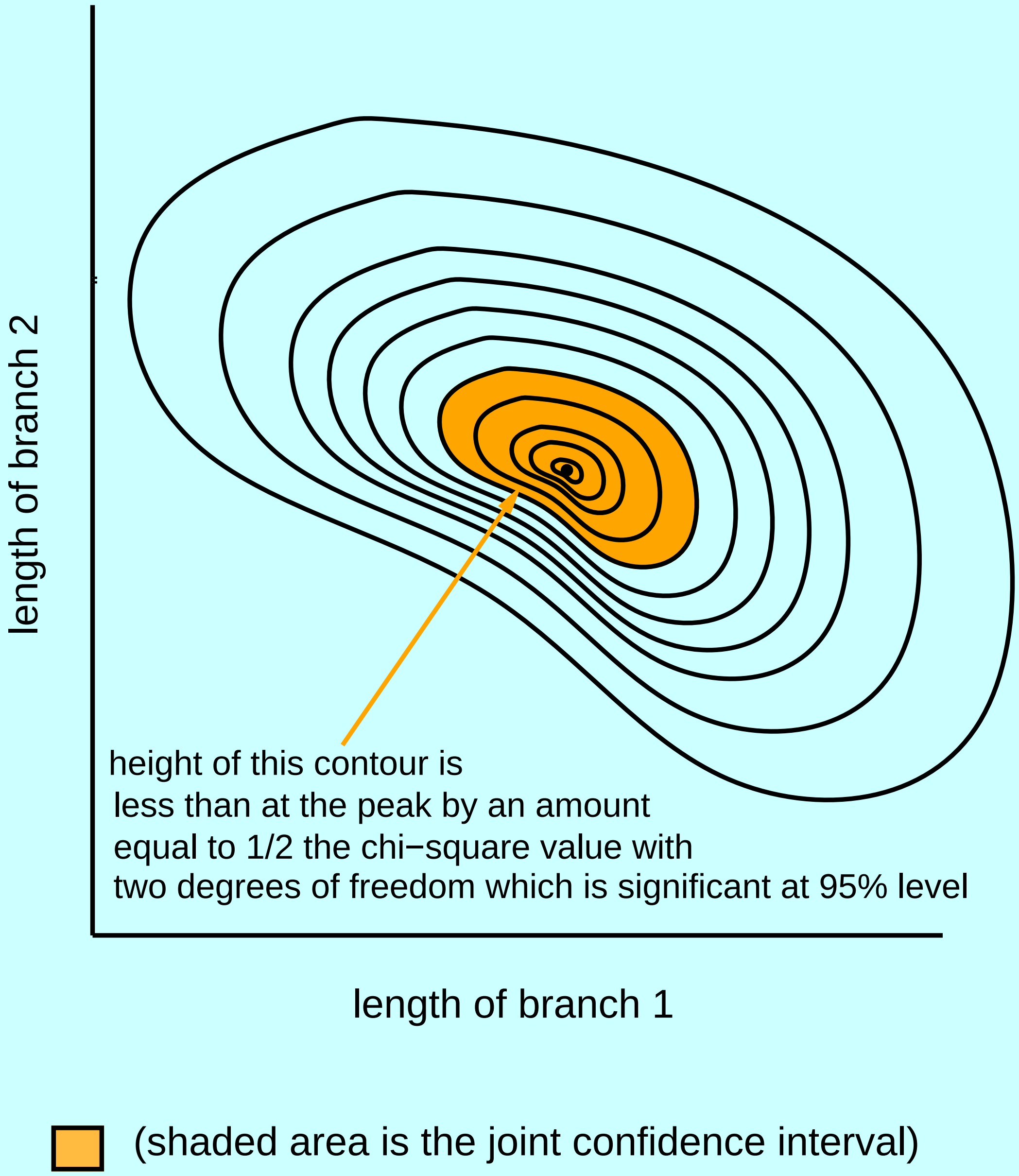


Ditto, in the other variable



(shaded area is the joint confidence interval)

A joint confidence region



The Likelihood Ratio Test

Remember that confidence intervals and tests are related: we test a null hypothesis by seeing whether the observed data's summary statistic is outside of the confidence interval around the parameter value for the null hypothesis.

The Likelihood Ratio Test invented by R. A. Fisher does this:

- Find the best overall parameter value and the likelihood, which is maximized there: $L(\theta_1)$.

The Likelihood Ratio Test

Remember that confidence intervals and tests are related: we test a null hypothesis by seeing whether the observed data's summary statistic is outside of the confidence interval around the parameter value for the null hypothesis.

The Likelihood Ratio Test invented by R. A. Fisher does this:

- Find the best overall parameter value and the likelihood, which is maximized there: $L(\theta_1)$.
- Find the best parameter value, and its likelihood, under constraint that the null hypothesis is true: $L(\theta_0)$.

The Likelihood Ratio Test

Remember that confidence intervals and tests are related: we test a null hypothesis by seeing whether the observed data's summary statistic is outside of the confidence interval around the parameter value for the null hypothesis.

The Likelihood Ratio Test invented by R. A. Fisher does this:

- Find the best overall parameter value and the likelihood, which is maximized there: $L(\theta_1)$.
- Find the best parameter value, and its likelihood, under constraint that the null hypothesis is true: $L(\theta_0)$.
- The degrees of freedom is the difference of the number of parameters in these two models, $p_1 - p_0$. The null hyothesis model must be a subcase of the general hypothesis, and must be within its parameter space, not on the boundary.

The Likelihood Ratio Test

Remember that confidence intervals and tests are related: we test a null hypothesis by seeing whether the observed data's summary statistic is outside of the confidence interval around the parameter value for the null hypothesis.

The Likelihood Ratio Test invented by R. A. Fisher does this:

- Find the best overall parameter value and the likelihood, which is maximized there: $L(\theta_1)$.
- Find the best parameter value, and its likelihood, under constraint that the null hypothesis is true: $L(\theta_0)$.
- The degrees of freedom is the difference of the number of parameters in these two models, $p_1 - p_0$. The null hyothesis model must be a subcase of the general hypothesis, and must be within its parameter space, not on the boundary.
- Take the log of the ratio of these likelihoods, or (what is the same), the difference of the logs of these two likelihoods: $\ln(L(\theta_1)/L(\theta_0))$.

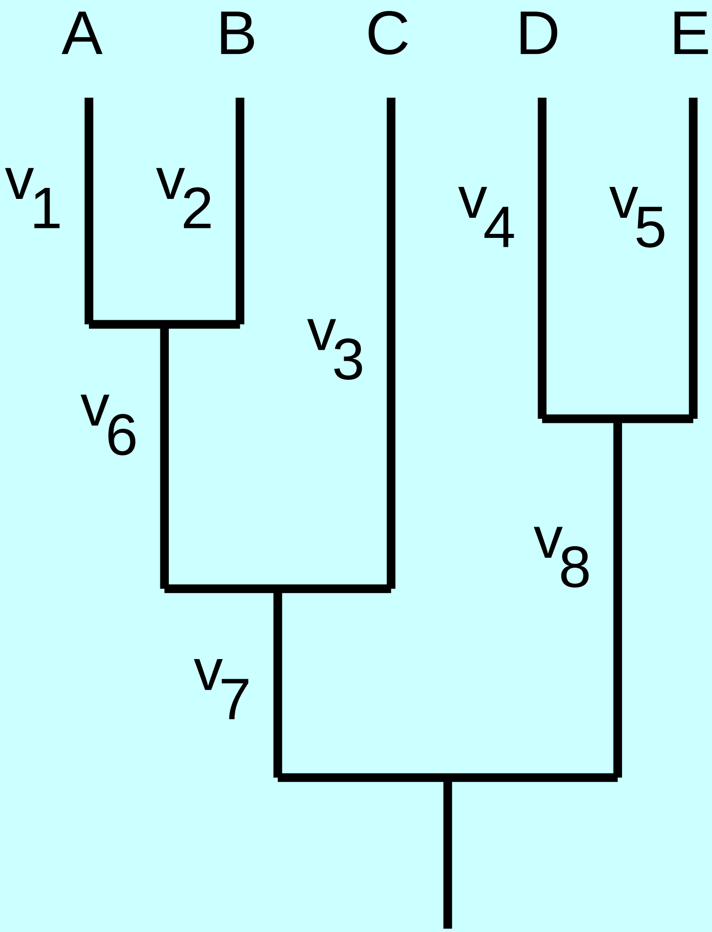
The Likelihood Ratio Test

Remember that confidence intervals and tests are related: we test a null hypothesis by seeing whether the observed data's summary statistic is outside of the confidence interval around the parameter value for the null hypothesis.

The Likelihood Ratio Test invented by R. A. Fisher does this:

- Find the best overall parameter value and the likelihood, which is maximized there: $L(\theta_1)$.
- Find the best parameter value, and its likelihood, under constraint that the null hypothesis is true: $L(\theta_0)$.
- The degrees of freedom is the difference of the number of parameters in these two models, $p_1 - p_0$. The null hyothesis model must be a subcase of the general hypothesis, and must be within its parameter space, not on the boundary.
- Take the log of the ratio of these likelihoods, or (what is the same), the difference of the logs of these two likelihoods: $\ln(L(\theta_1)/L(\theta_0))$.
- Double it, and look it up on a chi-square distribution with $p_1 - p_0$ degrees of freedom.

An example with phylogenies: molecular clock?



Constraints for a clock

$$v_1 = v_2$$

$$v_4 = v_5$$

$$v_1 + v_6 = v_3$$

$$v_3 + v_7 = v_4 + v_8$$

Testing for a molecular clock

To test for a molecular clock:

- Obtain the likelihood with no constraint of a molecular clock (For primates data with $T_s/T_n = 30$ we get $\ln L_1 = -2616.86$)
- Obtain the highest likelihood for a tree which is constrained to have a molecular clock: $\ln L_0 = -2679.0$
- Look up $2(\ln L_1 - \ln L_0) = 2 \times 62.14 = 124.28$ on a χ^2 distribution with $n - 2 = 12$ degrees of freedom (in this case the result is significant)

An example – samples from a Poisson distribution

Suppose we have m samples from a Poisson distribution whose (unknown) mean parameter is λ. Suppose the numbers of events we see are $n_1, n_2, \ldots, n_m$. The likelihood is

$$L \;=\; \frac{e^{-\lambda}\lambda^{n_1}}{n_1!} \times \frac{e^{-\lambda}\lambda^{n_2}}{n_2!} \times \ldots \times \frac{e^{-\lambda}\lambda^{n_m}}{n_m!}$$

collecting powers and exponentials, this becomes

$$L \;=\; e^{-m\lambda}\lambda^{n_1+n_2+\ldots+n_m}/(\text{lots of factorials})$$

Taking logarithms, which makes it easier

$$\ln L \;=\; -m\lambda \;+\; \left(\sum n_i\right)\,\ln\lambda \;+\; (\text{stuff not involving } \lambda)$$

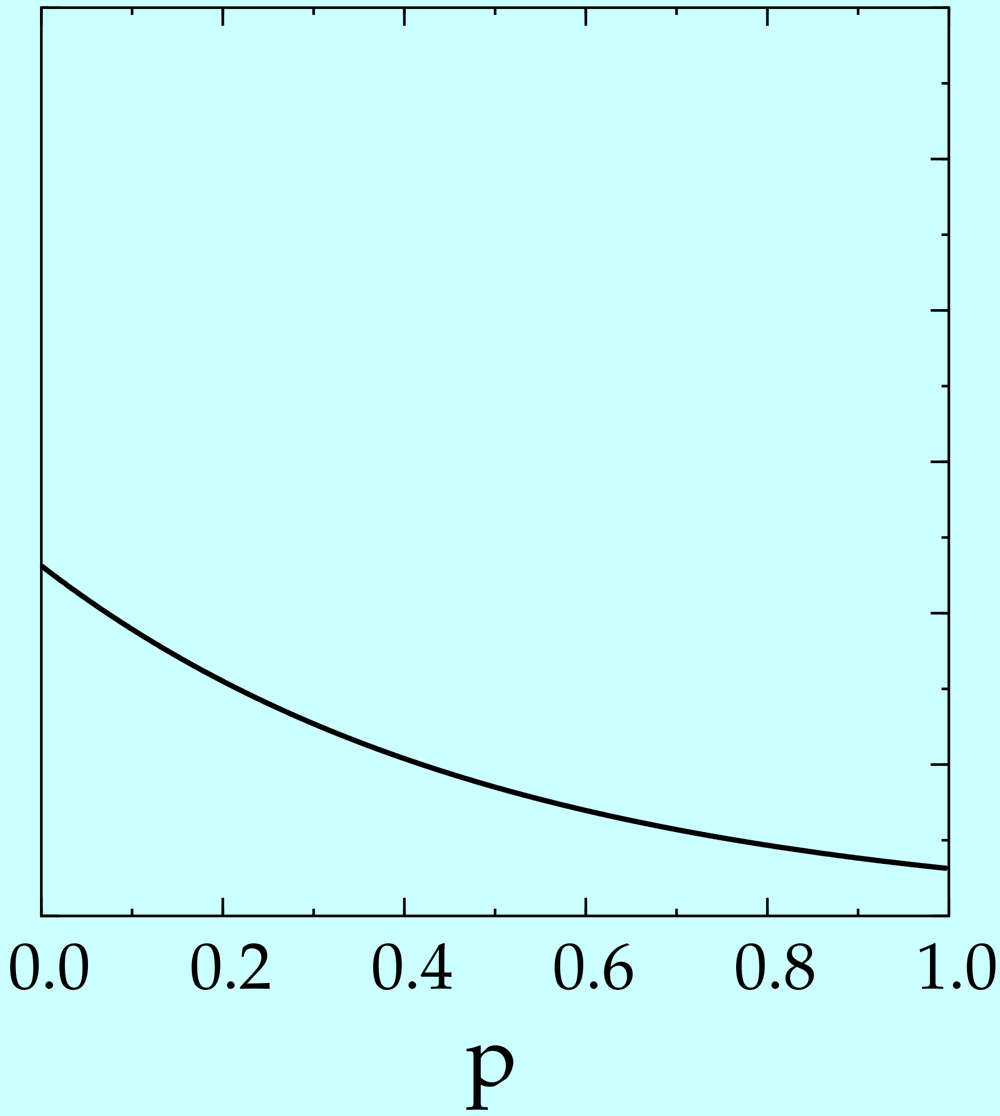
Differentiate this, set to zero:

$$\frac{\partial \ln L}{\partial\lambda} \;=\; -m \;+\; \left(\sum n_i\right)\frac{1}{\lambda} \;+\; 0 \;=\; 0$$

When you solve this for λ, you find that the MLE of λ is just the average number of events.

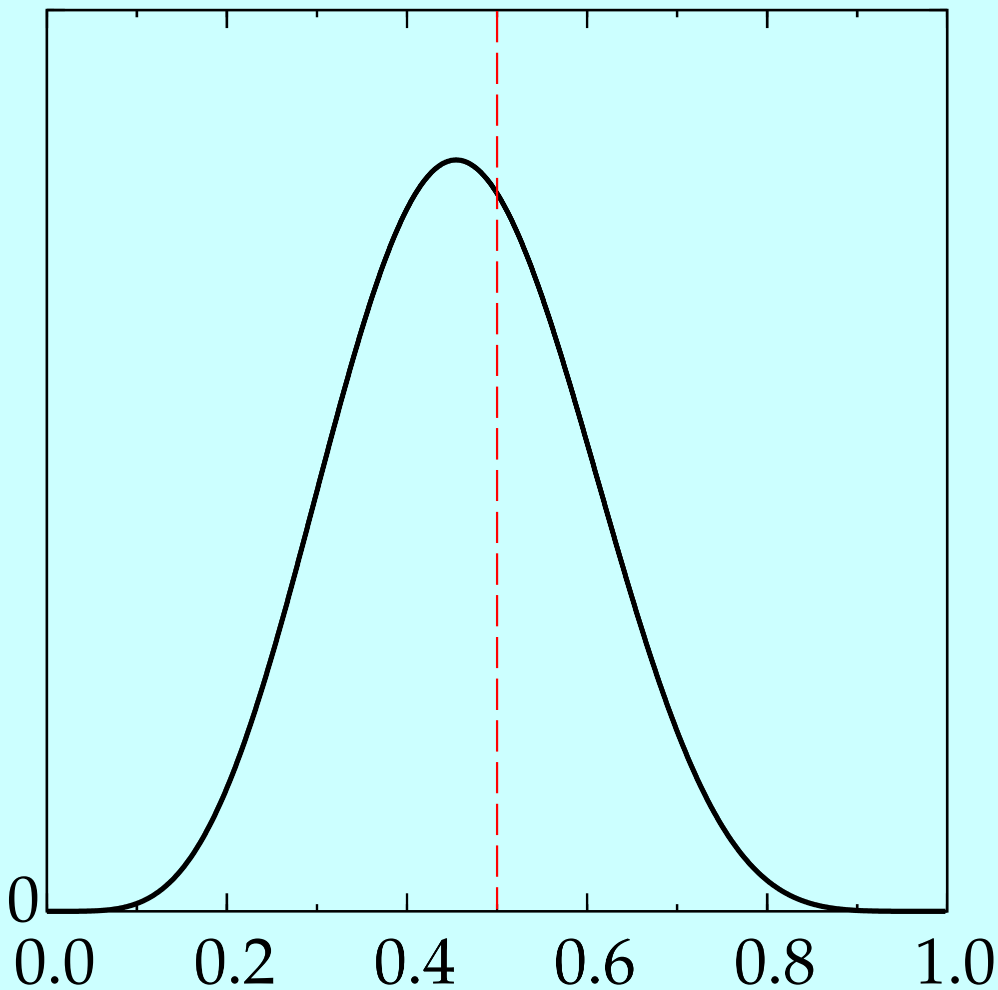
$$\hat{\lambda} \;=\; \frac{\sum n_i}{m}$$

An example of Bayesian inference with coins



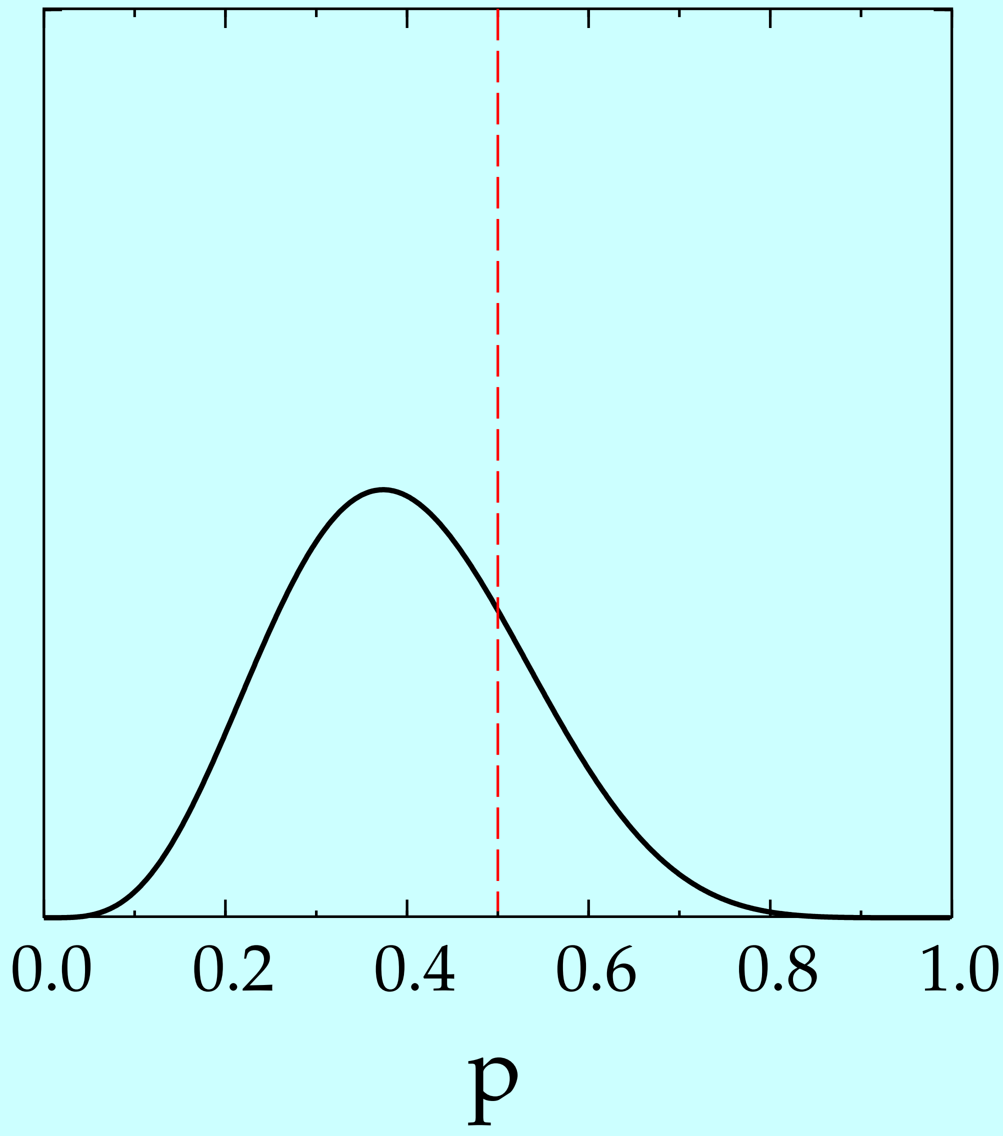
The prior on Heads probability – a truncated exponential distribution

An example of Bayesian inference with coins



The likelihood curve for 11 tosses with 5 heads appearing.

An example of Bayesian inference with coins



The resulting posterior on Heads probability

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,
- ... then it takes the resulting curve and restandardizes it so the area under it is 1.

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,
- ... then it takes the resulting curve and restandardizes it so the area under it is 1.
- That is the posterior, the very thing we need.

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,
- ... then it takes the resulting curve and restandardizes it so the area under it is 1.
- That is the posterior, the very thing we need.
- In practice, for complex models, Markov Chain Monte Carlo (MCMC) methods are used to wander in the parameter space and take a large enough sample from the posterior.

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,
- ... then it takes the resulting curve and restandardizes it so the area under it is 1.
- That is the posterior, the very thing we need.
- In practice, for complex models, Markov Chain Monte Carlo (MCMC) methods are used to wander in the parameter space and take a large enough sample from the posterior.
- The controversy between Bayesians and non-Bayesians is really over just one thing – whether assuming you know the prior is justified.

Bayesian inference

Bayesian inference uses likelihoods, but has a prior distribution on the unknown parameters.

- In theory it just multiplies the prior density by the likelihood curve,
- ... then it takes the resulting curve and restandardizes it so the area under it is 1.
- That is the posterior, the very thing we need.
- In practice, for complex models, Markov Chain Monte Carlo (MCMC) methods are used to wander in the parameter space and take a large enough sample from the posterior.
- The controversy between Bayesians and non-Bayesians is really over just one thing – whether assuming you know the prior is justified.
- If the prior is flat in that region, the highest point on the likelihood curve (i.e., the MLE) is also the peak of the posterior density.