Bayes for beginners: **Likelihood**

Thom Baguley and Mark Andrews
Nottingham Trent University

The likelihood approach to inference

Classical, frequentist statistical inference treats observed data as random and parameters as fixed (but unknown).

The likelihood approach to inference

Classical, frequentist statistical inference treats observed data as random and parameters as fixed (but unknown).

The probability model is in effect one in which we model all the possible ways our data could have arisen in order to make inferences about the fixed parameters.

The likelihood approach to inference

Classical, frequentist statistical inference treats observed data as random and parameters as fixed (but unknown).

The probability model is in effect one in which we model all the possible ways our data could have arisen in order to make inferences about the fixed parameters.

The likelihood approach provides an alternative way to conceptualize inference

The likelihood approach to inference

Classical, frequentist statistical inference treats observed data as random and parameters as fixed (but unknown).

The probability model is in effect one in which we model all the possible ways our data could have arisen in order to make inferences about the fixed parameters.

The likelihood approach provides an alternative way to conceptualize inference

... one in which we model the uncertainty of of information about a parameter (treating it as a random variable) based on the observed data

The likelihood function

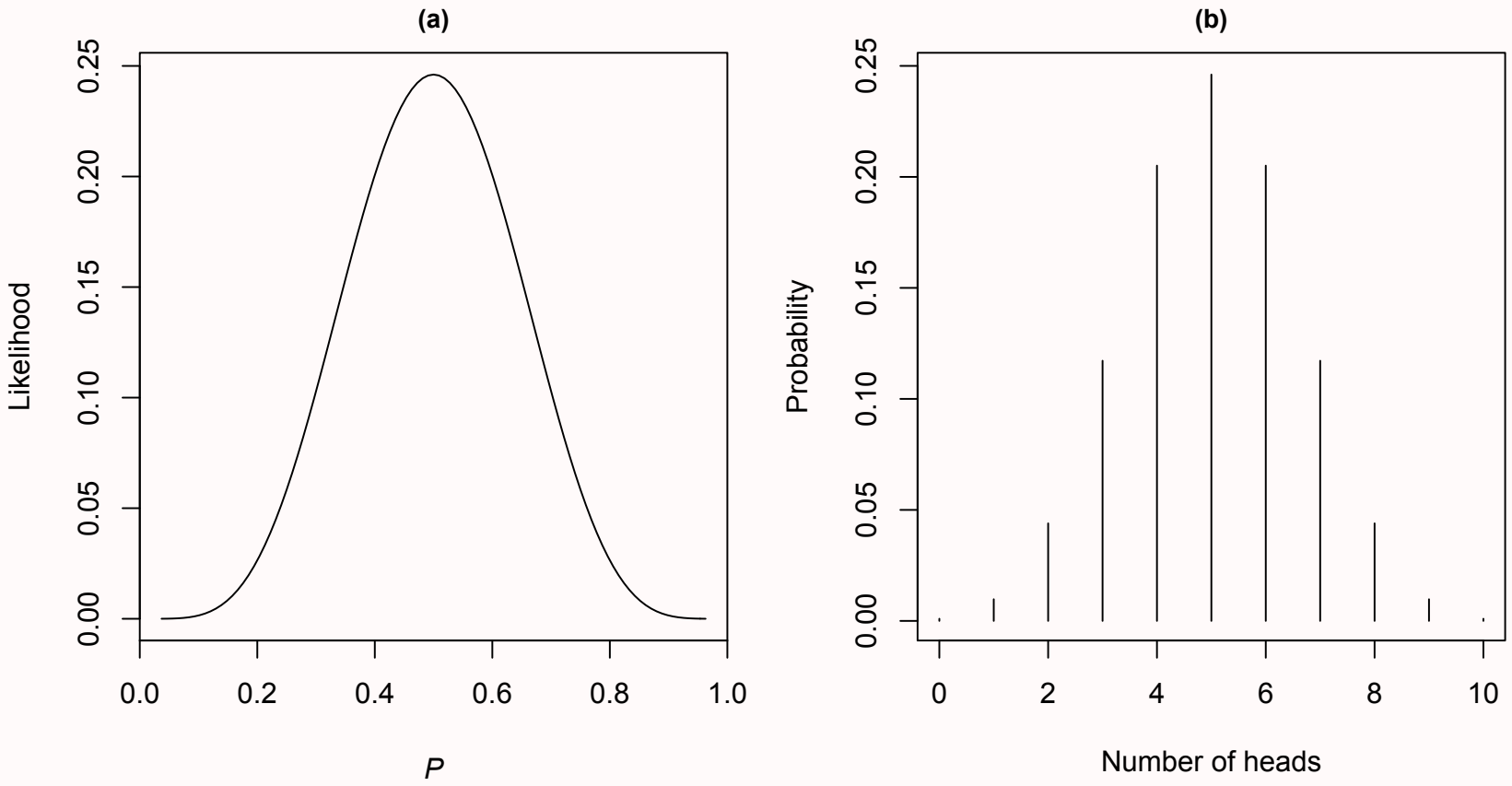
The likelihood is a mathematical function that is *proportional to* the probability of the observed data as a function of a parameter (e.g., a population mean) or a set of parameters

e.g., If a fair coin was tossed 10 times and 5 heads were observed, the likelihood function could be defined as:

$$\ell(\theta) \propto f(5; 10, P) = \frac{10!}{5!5!} P^5 (1-P)^5$$

or (because the binomial coefficient is a constant for any observed set of data with 5 heads):

$$\ell(\theta) \propto= P^5 (1-P)^5$$



(a) likelihood function and (b) probability mass function for observing 5 heads from 10 fair coin tosses

The likelihood principle

Within the framework of a statistical model, all the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those two hypotheses on the data

Edwards (1972, p.30; see also Birnbaum, 1962)

The law of likelihood

Within the framework of a statistical model, a particular set of data supports one statistical hypothesis better than another if the likelihood of the first hypothesis, on the data, exceeds the likelihood of the second hypothesis.

Edwards (1972, p.30)

... the degree to which occurrence of the event supports A over B (the strength of evidence) is quantified by [the likelihood ratio of the two probabilities.

Royall (2004, p.123)

Likelihood ratios

A likelihood ratio quantifies the strength of evidence for one hypothesis relative to another.

Thus we can quantify evidence for a each hypothesis in the form of a likelihood:

$$\ell(H_1|D) = cPr(D|H_1) \propto Pr(D|H_1)$$

... and form a ratio of the two likelihoods for inference:

$$LR_{H_1|H_2} = \frac{\ell(H_1|D)}{\ell(H_2|D)} = \frac{cPr(D|H_1)}{cPr(D|H_2)} = \frac{Pr(D|H_1)}{Pr(D|H_2)}$$

Quantifying evidence

p values v. likelihood ratios I

Imagine a court case in which a mother is accused of a horrific crime: that of killing two infant children born several years apart.

Quantifying evidence

p values v. likelihood ratios I

Imagine a court case in which a mother is accused of a horrific crime: that of killing two infant children born several years apart.

H_0: Cause of death = SIDS H_1: Cause of death = murder

The prosecution expert argues that the probability of the data given H_0 is $(1/8543)^2 \approx 1.4 \times 10^{-8}$ or 1 in 73 million

Quantifying evidence

p values v. likelihood ratios I

Imagine a court case in which a mother is accused of a horrific crime: that of killing two infant children born several years apart.

H_0: Cause of death = SIDS H_1: Cause of death = murder

The prosecution expert argues that the probability of the data given H_0 is $(1/8543)^2 \approx 1.4 \times 10^{-8}$ or 1 in 73 million

... though a more accurate figure is $1/1300$ rather than $1/8543$

* This example is based loosely on the wrongful conviction of Sally Clark (see, Hill, 2004; Baguley, 2012, p. 365)

Quantifying evidence

p values v. likelihood ratios II

Although p for H_0 is very low - we can't quantity the evidence for or against H_0 with this information alone

Quantifying evidence

p values v. likelihood ratios II

Although *p* for H_0 is very low - we can't quantity the evidence for or against H_0 with this information alone

Child murder is rare and it turns out that the probability of is around $1/21700$

Quantifying evidence

p values v. likelihood ratios II

Although *p* for H_0 is very low - we can't quantity the evidence for or against H_0 with this information alone

Child murder is rare and it turns out that the probability of is around $1/21700$

...and also taking into account that a second murder or second SIDS case is not independent we arrive at:

$$LR_{H_1|H_0} = \frac{\ell(H_1|D)}{\ell(H_0|D)} \approx \frac{1/21700 \times 1/123}{1/1300 \times 1/128} \approx 1/16$$

The maximum likelihood estimator

The maximum likelihood estimator (MLE) $\hat{\theta}$ is the value that maximises the likelihood function:

$$\max\left[\ell\left(\theta\right)\right]$$

... it is an extremely useful quantity and widely used in statistics.

e.g., the mean is the MLE for a normal distribution
... though for the variance it is $1/n\sum_{i}^{n}(x_i-\bar{x})^2$ rather than $1/(n-1)\sum_{i}^{n}(x_i-\bar{x})^2$

Likelihood interval

... it is also possible to obtain an interval estimate around the point estimate given by the MLE for a likelihood function.

A likelihood interval for a parameter θ contains all values of θ consistent with the data.

Likelihood interval

... it is also possible to obtain an interval estimate around the point estimate given by the MLE for a likelihood function.

A likelihood interval for a parameter θ contains all values of θ consistent with the data.

This is defined by a likelihood ratio in favor of θ of at most $1/k$ relative to any other possible value of θ

Likelihood interval

... it is also possible to obtain an interval estimate around the point estimate given by the MLE for a likelihood function.

A likelihood interval for a parameter θ contains all values of θ consistent with the data.

This is defined by a likelihood ratio in favor of θ of at most $1/k$ relative to any other possible value of θ

... this can be illustrated graphically

i) standardise the likelihood (placing it on a 0 to 1 scale)

Likelihood interval

... it is also possible to obtain an interval estimate around the point estimate given by the MLE for a likelihood function.

A likelihood interval for a parameter θ contains all values of θ consistent with the data.

This is defined by a likelihood ratio in favor of θ of at most $1/k$ relative to any other possible value of θ

... this can be illustrated graphically

i) standardise the likelihood (placing it on a 0 to 1 scale)

ii) draw a line at $\ell(\theta) = 1/k$ (e.g., where $k = 8$ or $k = 32$)

Likelihood interval

... it is also possible to obtain an interval estimate around the point estimate given by the MLE for a likelihood function.

A likelihood interval for a parameter θ contains all values of θ consistent with the data.

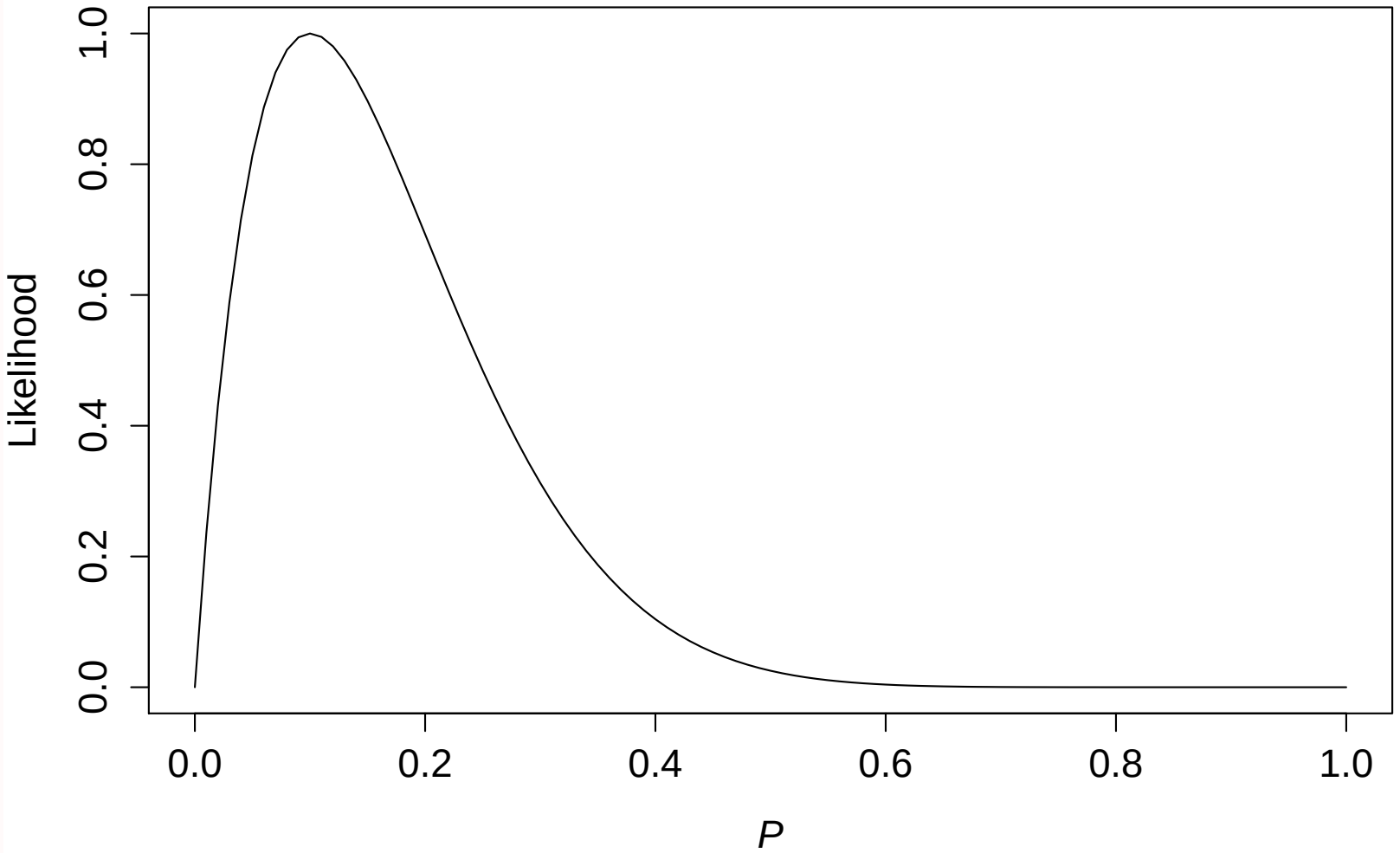
This is defined by a likelihood ratio in favor of θ of at most $1/k$ relative to any other possible value of θ

... this can be illustrated graphically

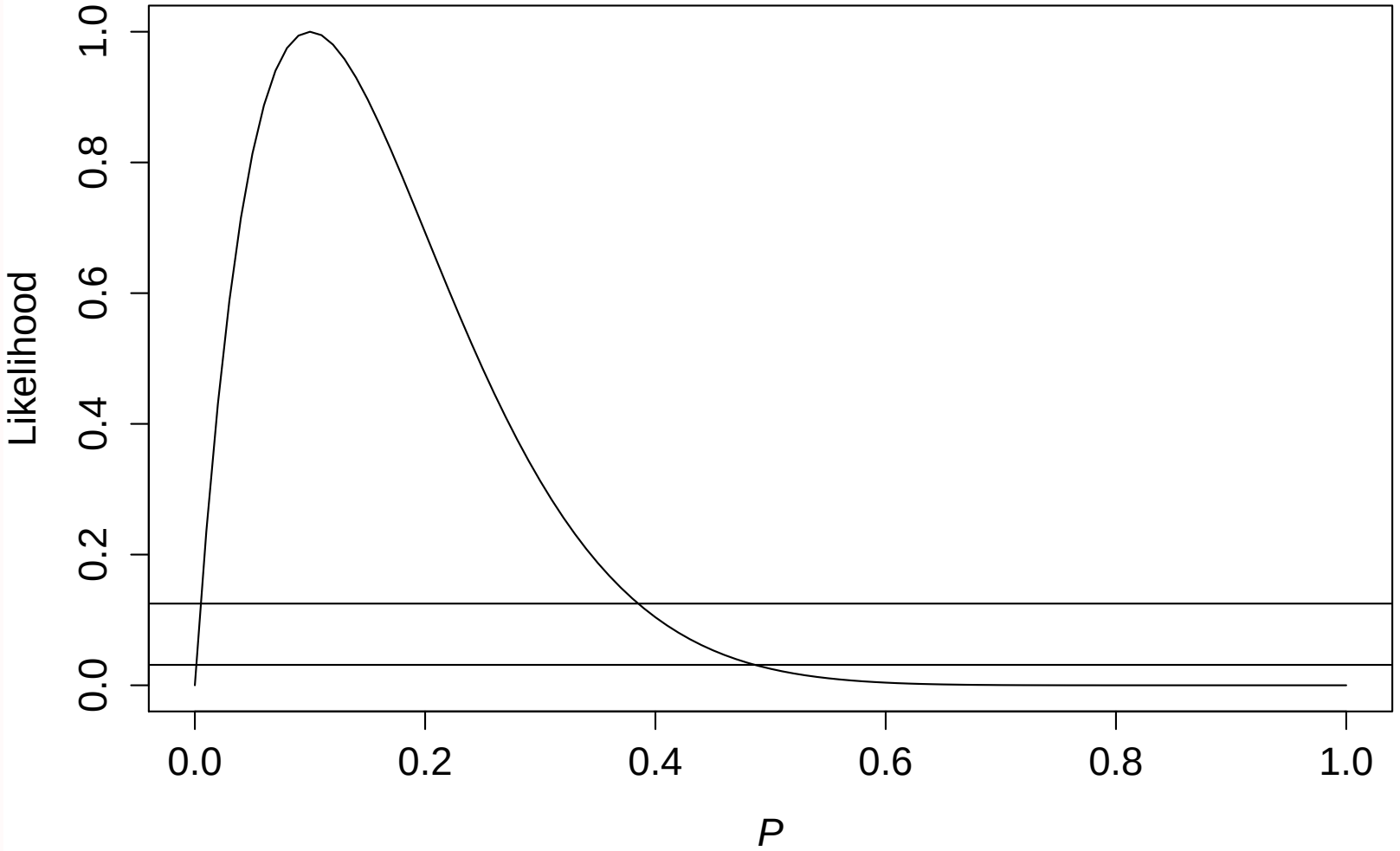
i) standardise the likelihood (placing it on a 0 to 1 scale)

ii) draw a line at $\ell(\theta) = 1/k$ (e.g., where $k = 8$ or $k = 32$)

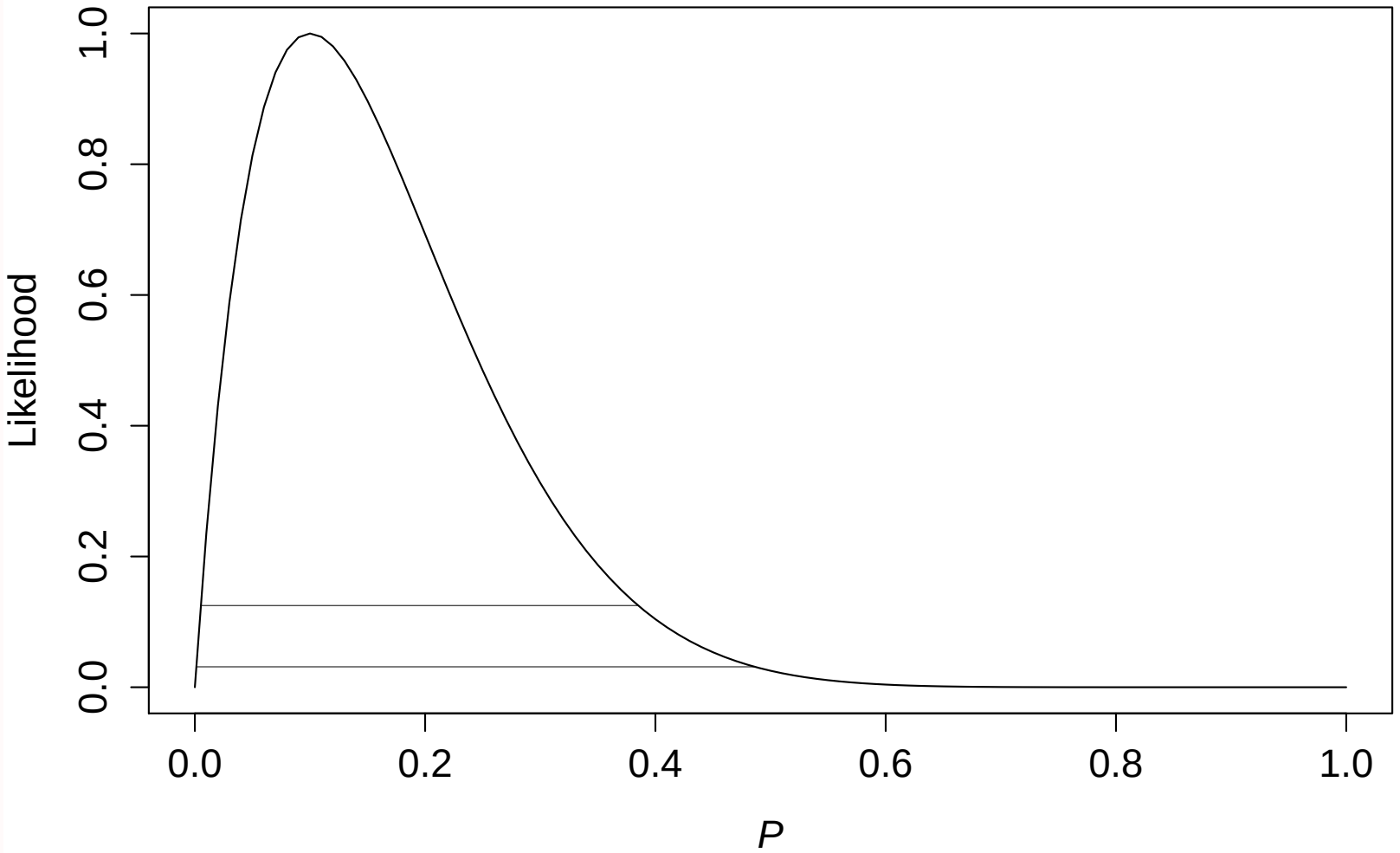
iii) where this line intersects the likelihood curve marks the bounds of the interval



Standardized likelihood for 1 head from 10 fair coin tosses



... adding a horizontal line at $k = 8$ and $k = 32$

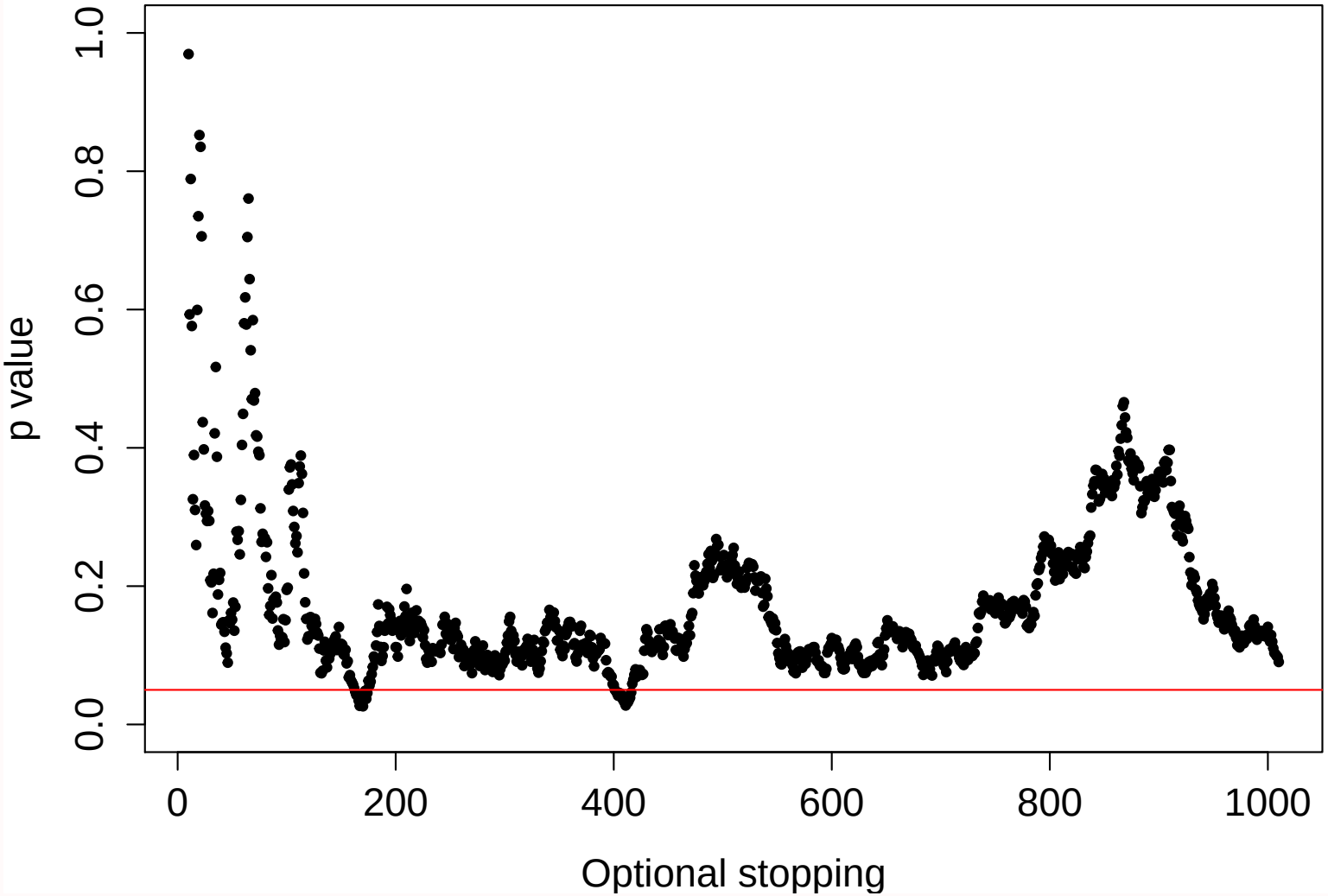


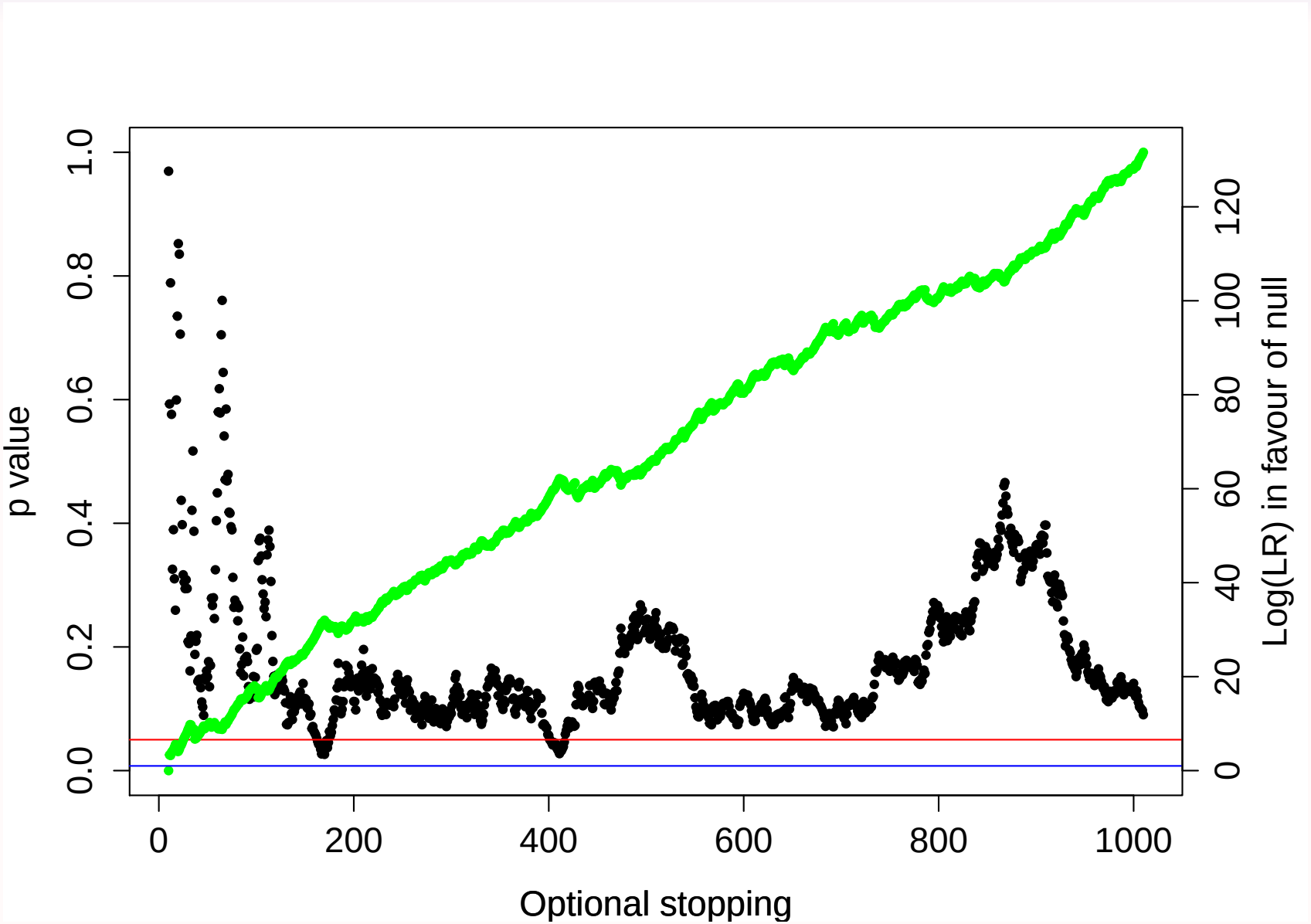
Intervals are: [.0051,.3850] and [.0012, .4863]

Why use the likelihood approach?

Likelihood ratios:

- quantify evidence provided by the data
- minimise both Type I and Type II errors
- are not vulnerable to optional stopping
- less vulnerable to problems such as multiple testing


p value
0.0
0.2
0.4
0.6
0.8
1.0
0
200
400
600
800
1000
Optional stopping


p value
0.0
0.2
0.4
0.6
0.8
1.0
Log(LR) in favour of null
0
20
40
60
80
100
120
0
200
400
600
800
1000
Optional stopping

Nuisance parameters

A parameter that you are not interested in estimating that needs to be accounted for to estimate the parameter of interest

e.g., σ^2 when estimating a normal mean

Nuisance parameters

A parameter that you are not interested in estimating that needs to be accounted for to estimate the parameter of interest

e.g., σ^2 when estimating a normal mean

Methods for dealing with nuisance parameters include:

- profile likelihood (replace nuisance parameter with MLE)

Nuisance parameters

A parameter that you are not interested in estimating that needs to be accounted for to estimate the parameter of interest

e.g., σ^2 when estimating a normal mean

Methods for dealing with nuisance parameters include:

- profile likelihood (replace nuisance parameter with MLE)
- marginal likelihood (integrate out nuisance parameters)

e.g., the profile likelihood for a normal mean with unknown variance is proportional to the t distribution

Information criteria I

Information criteria such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (AIC) are widely used in model comparison

They are computed from the loglikelihood $\ln(\ell)$ of a model but include a penalty for the number of parameters (q):

$$AIC = -2\ln(\ell) + 2q$$

$$AIC_C = -2\ln(\ell) + 2q + \frac{2q(q+1)}{N-q-1}$$

$$BIC = -2\ln(\ell) + q\ln(N)$$

Information criteria II

A more informative model has a lower AIC or BIC *for the same data*.

Taking the difference strips out the arbitrary constant terms in the likelihood:

$$e.g., \Delta_{BIC} = BIC_{M0} - BIC_{M1}$$

This difference can be transformed into a form of likelihood ratio:

$$LR_{BIC} = e^{1/2\Delta_{BIC}}$$

... or a probability:

$$Pr(M1/M0) = LR_{BIC}/(1 + LR_{BIC})$$