Lecture 7: Frequent

Statistical modeling

So far, we have been focusing mainly on re-randomization techniques to assess significance in **study designs that employ some kind of explicit randomization**

In each case, we **defined a test statistic** that represented some aspect of the subjects we were interested in studying and then **created a sampling distribution** for this statistic **under the null hypothesis** that interventions (treatment and control) had no effect on the subjects in our study

The sampling distribution captures the variability present in our experiment under the null hypothesis -- We used this distribution **to judge the size of our observed effect**, deciding whether it was big enough (“extreme enough”) to be considered something other than noise

Statistical modeling

The fact that we employed randomization in making our intervention assignments, combined with the null hypothesis of homogeneity between treatment and control provide **a framework for conducting inference**

Random assignments and homogeneity tell us enough about **how the data were generated** (under the null hypothesis) to simulate draws from **the sampling distribution** (under the null hypothesis) and conduct a formal test

These assumptions are relatively weak as statistical assumptions go, and for the next few lectures we will start to add more, fleshing out a framework for **connecting inference (learning from data) to the stochastic (probabilistic) mechanism** that created the data

Statistical modeling

For the next few lectures, we are going to fit probability models to data -- In short, probability distributions will serve as a kind of **“origin story” for how the data were generated** -- There are many reasons for fitting probability laws to data

We might relate aspects of the distribution (features or parameters) to a scientific theory that reveals something about the **state of Nature**

Probability models can be used for purely descriptive purposes, acting as a kind of **data summary** or “compression”

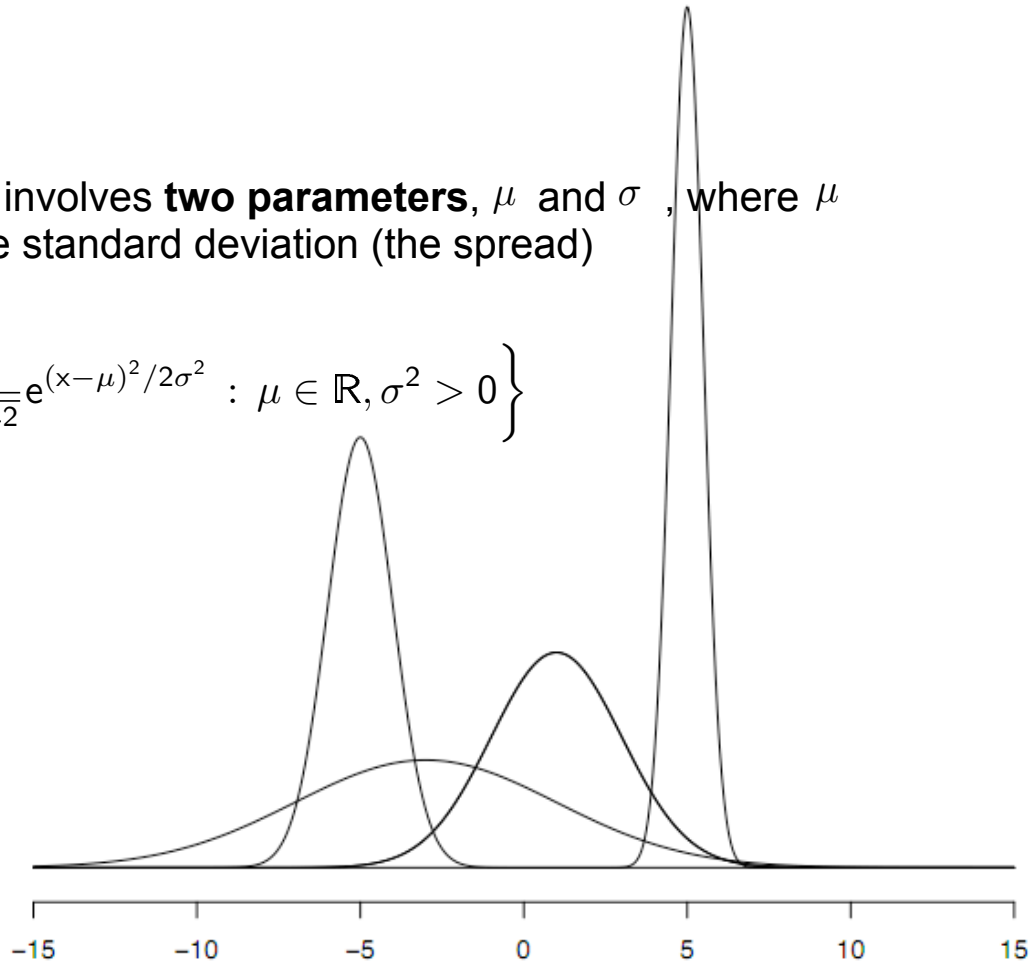
Finally, we are often interested in **simulation**, in using these models to make predictions or to generate new data that can be fed into a larger modeling exercise

Last time, we looked at two commonly used “families” of probability distributions -- In each case, the probability distributions were all of **a particular functional form and depended on one or more “parameters”**

The normal family

The normal or Gaussian distribution involves **two parameters**, μ and σ, where μ is the mean (the center) and σ is the standard deviation (the spread)

$$\mathcal{F} = \left\{ f(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{(x-\mu)^2/2\sigma^2} \; : \; \mu \in \mathbb{R}, \sigma^2 > 0 \right\}$$



The binomial family

Let X denote the number of successes in m independent trials, each with success probability p -- Then the probability function of X belongs to the family

$$\mathcal{F} = \left\{ f(k|p) = \binom{m}{k} p^k (1-p)^{m-k} \ : \ p \in [0,1] \right\}$$

Unlike the normal, the binomial is indexed by **a single parameter, p**, the success probability (we consider m, the number of trials, to be fixed and given)

Recall that if X has a binomial distribution (m,p), then **X has expected value mp** and **variance mp(1-p)**

Point estimation

To set notation a little, we will assume that $X_1, \ldots, X_n$ are a sample of n independent observations drawn from $f(x|\theta^*)$, a member of the parametric family of probability functions $f(x|\theta)$ indexed by a (possibly vector valued) parameter θ

For the next few slides, we will be interested in forming an estimate of θ^* based on data -- That is, we will examine ways to form $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ using the observations $X_1, \ldots, X_n$

Maximum likelihood

Recall that the joint distribution of independent random variables is given by the product of their individual distributions -- Therefore, the joint distribution of our n data points $X_1, \ldots, X_n$ is given by

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Maximum likelihood

We view this expression as a function of θ rather than the data $X_1, \ldots, X_n$ and define the likelihood function to be

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

with **the associated log-likelihood function**

$$l(\theta) = \log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f(x|\theta)$$

Likelihood

We can then examine different values of θ, comparing them based on how much support they offer to our observed data

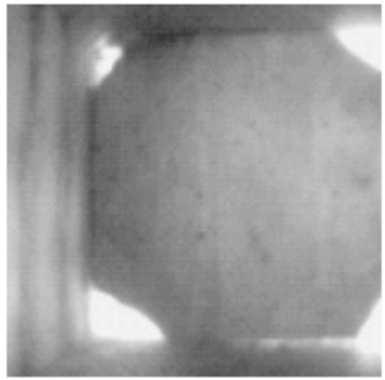
To be more precise, with each value of θ we define a probability model and can ask whether or not our data seem likely under the given model -- So, for example, we should probably avoid values of θ for which our data would have been a rare event

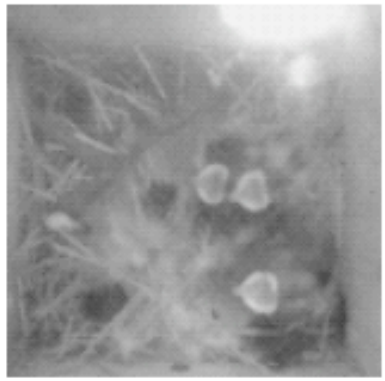
Nestbox 8

Images were taken of the inside of the nest box every 90 minutes for a period of two weeks (14 days) -- The images were manually tagged to indicate the presence or absence of an (adult) bird

For each 90 minute period, we can think of the m=14 trials as tossing a coin with probability p that we see a bird -- We then might think of the 15 counts as being observations of binomial random variables

	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13	d14
1	0	1	1	1	1	1	1	1	1	0	1	1	0	1
2	1	1	0	1	0	1	0	1	1	1	0	1	1	1
3	0	0	0	1	0	0	0	1	1	1	0	1	1	1
4	0	1	1	0	1	0	0	0	1	0	1	0	1	1
5	1	1	0	1	0	1	0	1	1	0	0	1	1	1
6	1	1	0	1	1	1	0	0	1	1	0	1	1	1
7	1	1	0	1	1	1	0	1	1	0	0	1	1	1
8	1	1	0	1	1	1	0	1	1	1	0	1	1	1
9	1	1	1	0	1	1	1	1	1	1	0	1	1	1
10	1	1	0	0	1	1	0	1	1	0	1	1	1	1
11	0	1	1	0	0	1	1	1	0	0	1	0	1	1
12	0	1	0	0	1	1	1	0	1	0	1	0	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	0	1	0	1	0	1	1	0	0	1	0	1
15	1	1	0	1	1	1	1	1	1	1	0	1	1	1





Nestbox 8

We can think of each row in the table on the previous slide as a series of 14 trials, each with success probability p -- That means that we have n=15 observations from the distribution binomial (14,p) for some value of p

The counts in each row (our data points $X_1, \ldots, X_{15}$) are then

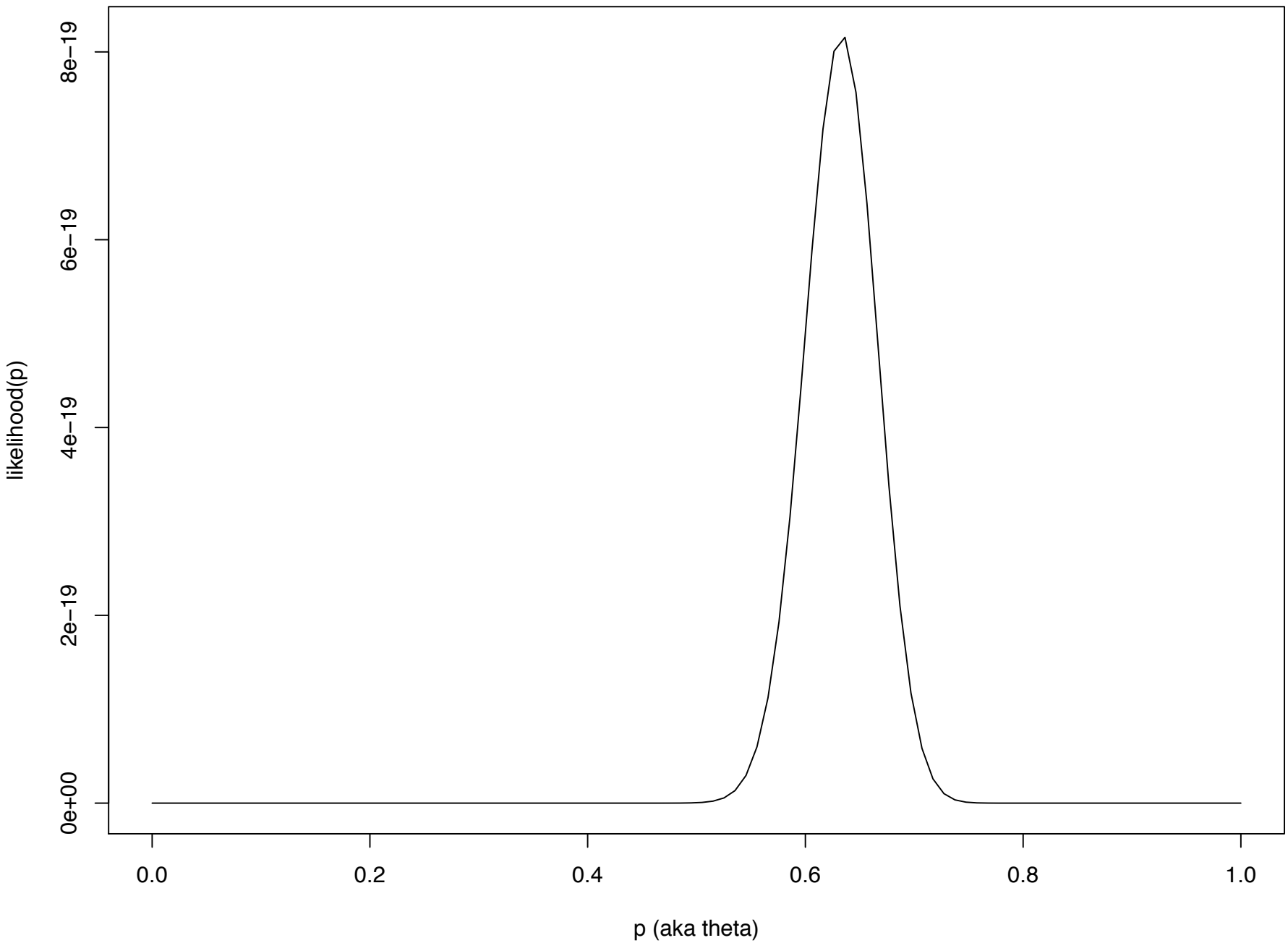
11 10 7 7 9 10 10 11 12 10 8 8 0 8 12

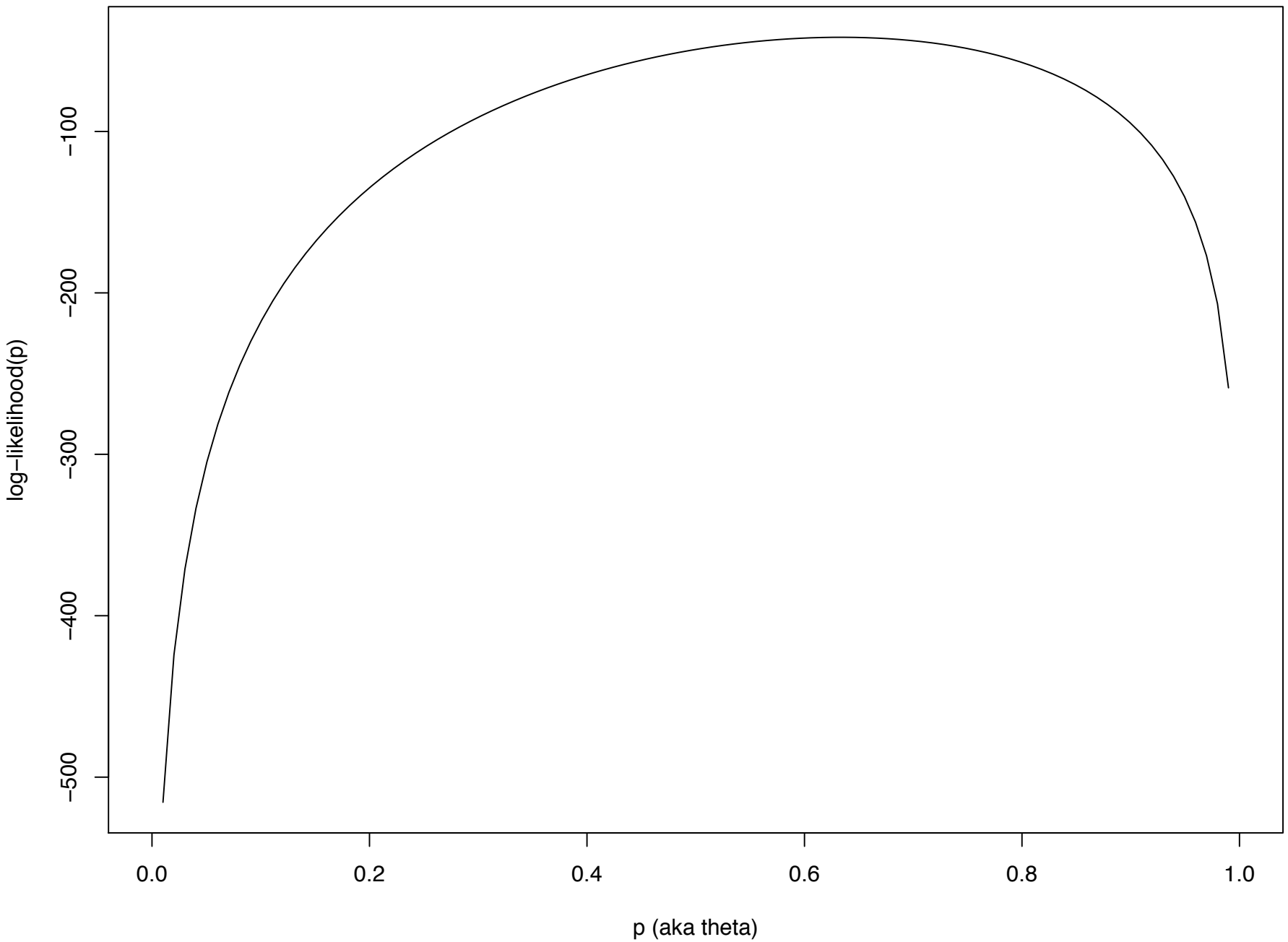
The likelihood is then

$$\prod_{i=1}^{15} \binom{14}{X_i} p^{X_i} (1-p)^{14-X_i}$$

which we can expand (ugh, let's only do this once)

$$\left[\binom{14}{11} p^{11}(1-p)^3\right]\left[\binom{14}{10} p^{10}(1-p)^4\right] \cdots \left[\binom{14}{12} p^{12}(1-p)^2\right] \propto p^{133}(1-p)^{77}$$


likelihood(p)
0e+00
2e−19
4e−19
6e−19
8e−19
0.0
0.2
0.4
0.6
0.8
1.0
p (aka theta)


log−likelihood(p)
−100
−200
−300
−400
−500
0.0
0.2
0.4
0.6
0.8
1.0
p (aka theta)

Maximum likelihood

As its name implies, as an estimation procedure, maximum likelihood suggests selecting a value for $\widehat{\theta}$ that makes the data the most likely or probable

Formally, this means $\widehat{\theta} = \text{argmax } \mathcal{L}(\theta)$

The binomial family

Let's apply ML to the binomial family -- Specifically we have a series of binomial observations which give rise to the likelihood

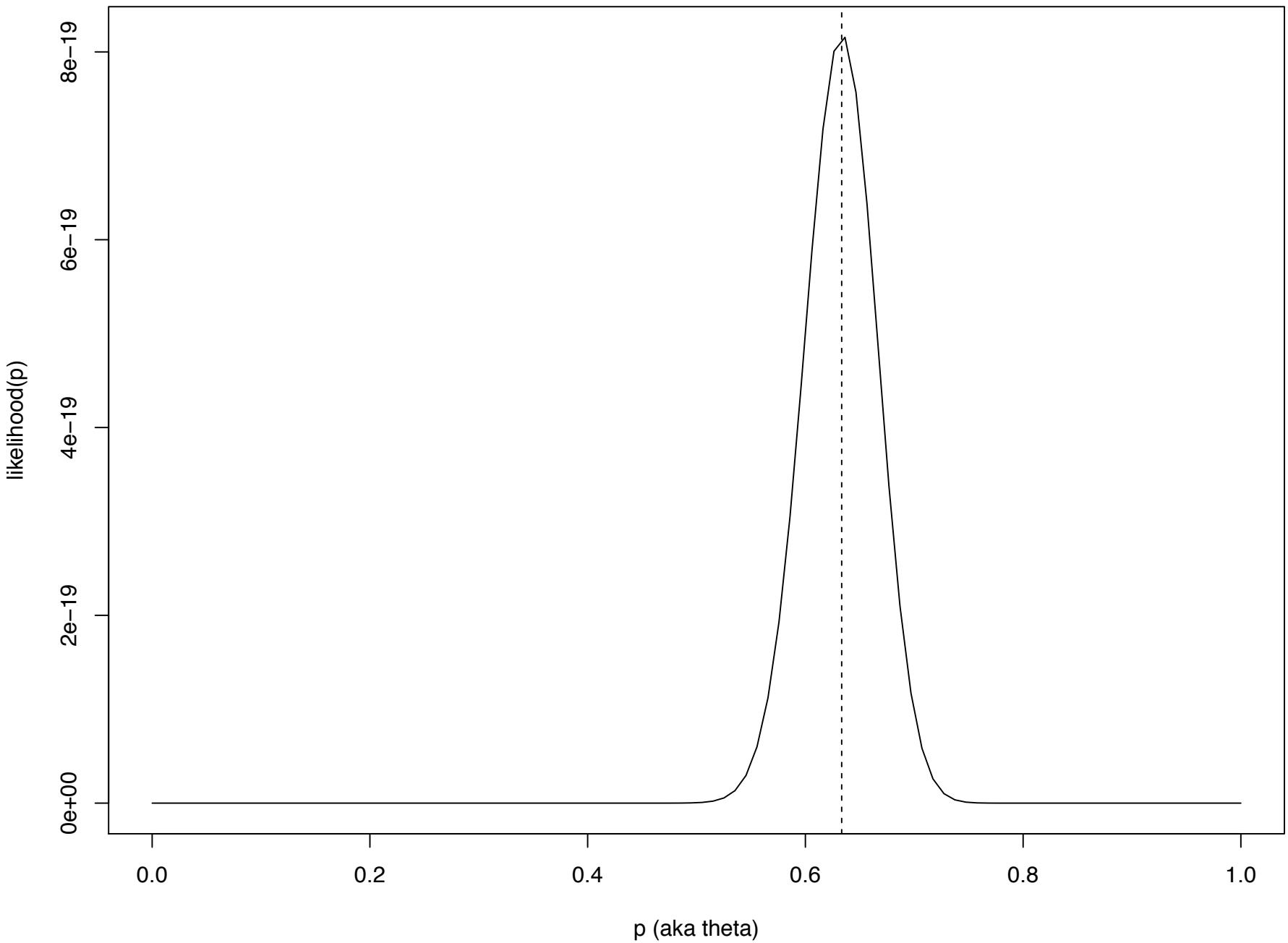
$$\begin{aligned}\mathcal{L}(p) &= \prod_{i=1}^{n} \binom{m}{X_i} p^{X_i}(1-p)^{m-X_i} \\ &= p^{U}(1-p)^{mn-U} \prod_{i=1}^{n} \binom{m}{X_i}\end{aligned}$$

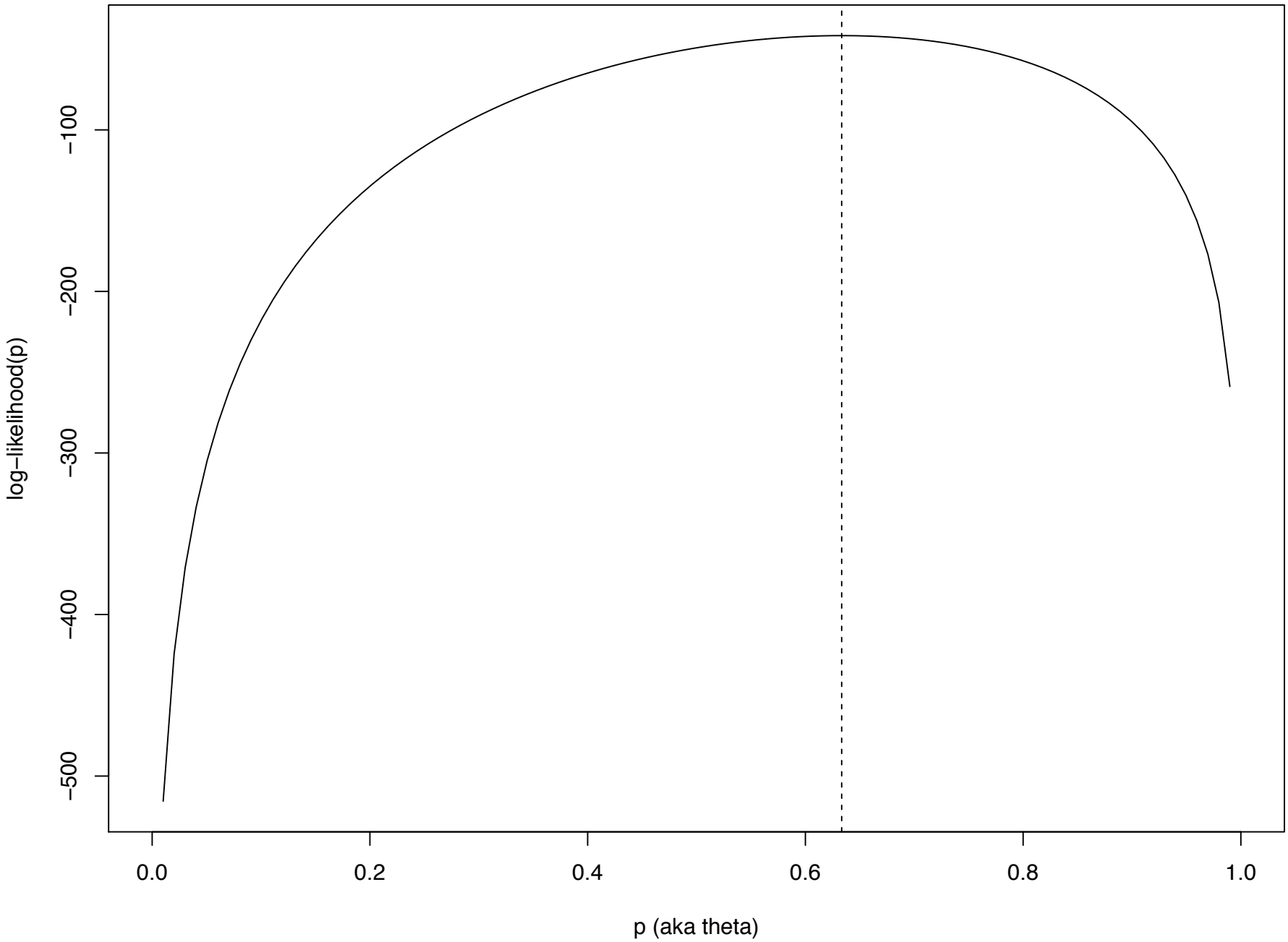
where we define $U = X_1 + \cdots + X_n$

This gives us a log-likelihood of the form (where C doesn't depend on p)

$$l(p) = U \log p + (nm - U) \log (1-p) + C$$

and after differentiating with respect to p, we find the MLE to be $\widehat{p} = U/nm$


likelihood(p)
0e+00
2e−19
4e−19
6e−19
8e−19
0.0
0.2
0.4
0.6
0.8
1.0
p (aka theta)


log−likelihood(p)
−100
−200
−300
−400
−500
0.0
0.2
0.4
0.6
0.8
1.0
p (aka theta)

The normal family

Let's derive the maximum likelihood estimates for observations coming from the normal family -- The likelihood is given by

$$\begin{aligned}\mathcal{L}(\mu,\sigma) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(X_i-\mu)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left[-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2\right]\end{aligned}$$

We can rewrite the term on the right using the fact that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}(X_i-\overline{X}+\overline{X}-\mu)^2 \\ &= \sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}(\overline{X}-\mu)^2 \\ &= nS^2+n(\overline{X}-\mu)^2\end{aligned}$$

where

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \quad \text{and} \quad S^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

The normal family

In the end, we can rewrite the likelihood as

$$\left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\sum_{i=1}^{n}(X_i-\mu)^2/2\sigma^2\right] = \left(\frac{1}{2\pi}\right)^{n/2} \sigma^{-n} \exp\left[-\frac{nS^2}{2\sigma^2}\right] \exp\left[-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right]$$

In terms of the log-likelihood function, this becomes (up to a constant C that doesn't depend on μ or σ)

$$l(\mu,\sigma) = C - n\log\sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}$$

Now, we take partial derivatives with respect to the parameters μ and σ to find the maximum likelihood estimates (MLEs) $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$

The normal family

I have to admit that running through that (albeit not that horrible math) and coming up with the sample mean and standard deviation as the MLEs is **a little bit of a let-down**

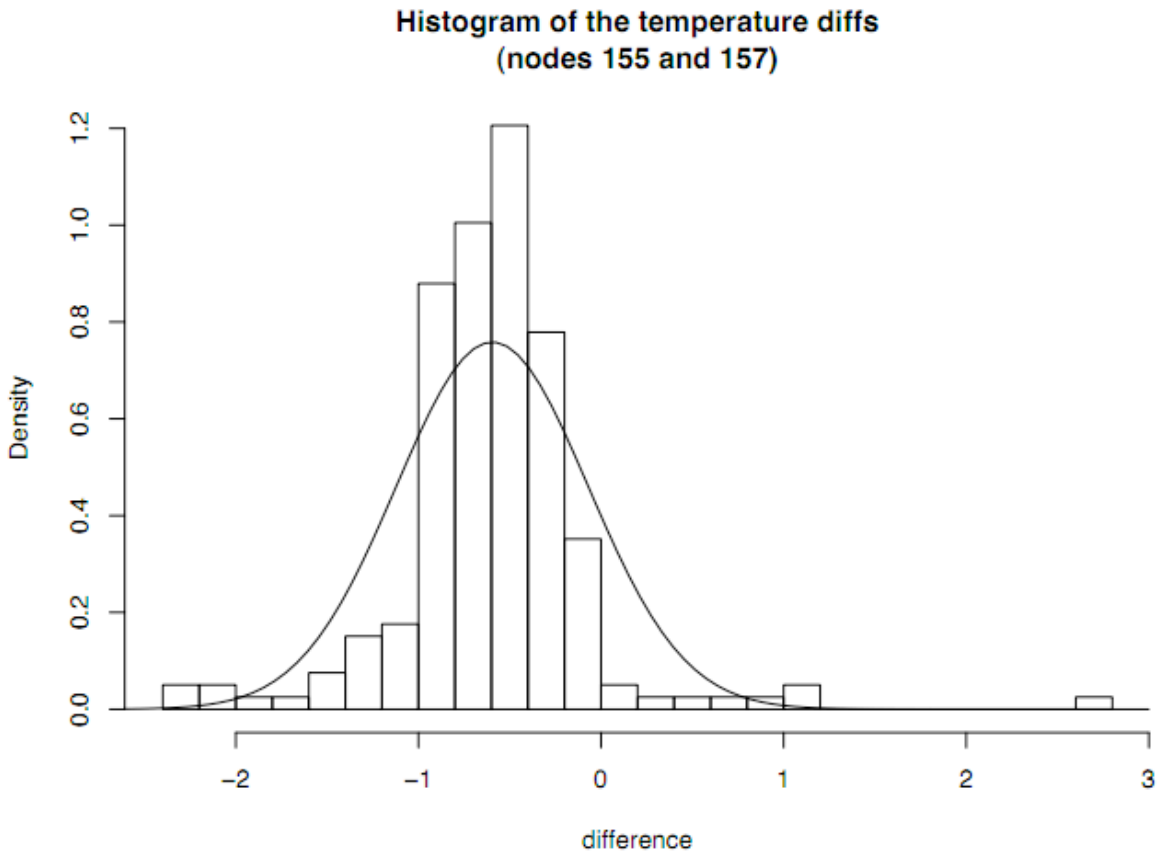
Granted, the estimates make intuitive sense and as we will see, being an MLE provides you with certain properties you might not otherwise expect

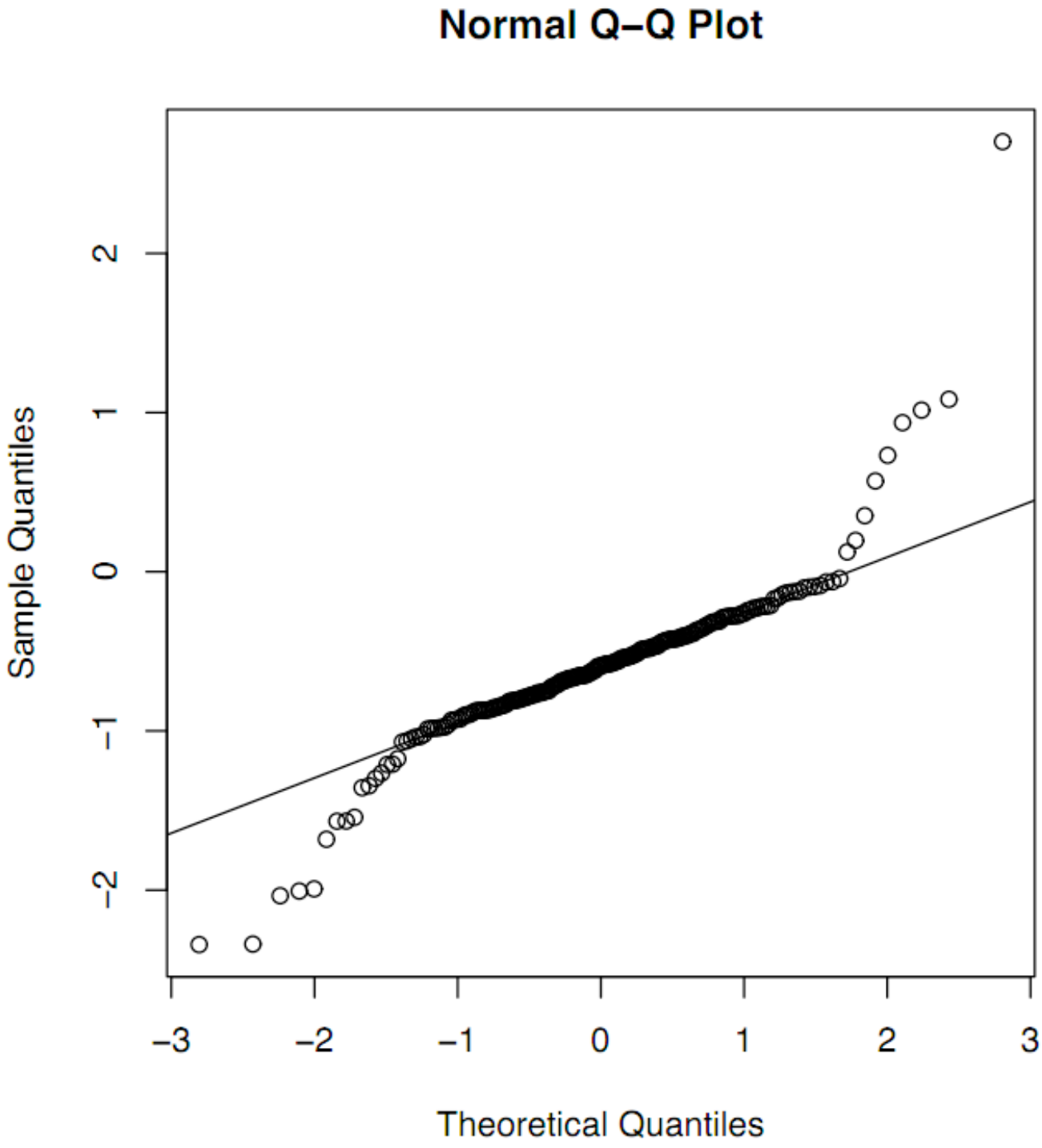
However, this does beg the question, if all we're doing here is computing the sample mean and variance, well, we could do that for any set of data -- **Fitting a model doesn't mean it, um, fits**

Example: CAD

At the right we have one of our histograms of the difference data with an overlay of our fitted normal (here, $\widehat{\mu} = -0.6$ and $\widehat{\sigma} = 0.53$)

How well have we done? Is this the best display to assess the fit?




Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
2
1
0
−1
−2
−3
−2
−1
0
1
2
3

Model checking

Whenever we fit a model, we should spend some time assessing whether or not the result seems reasonable given the data -- In the case of the “normal” data from the CAD transect, we see that the tails are probably heavier than we would expect from a normal and this parametric model isn’t right

That doesn’t stop us from considering statistics like the mean and standard deviation, it suggests that **any inferences we make that rely on assuming our probability model is correct are suspect**

In short, **probability modeling implies making assumptions about our data**, providing it with an **“origin story”** -- These stories can be pretty precise in what they say about patterns we should see in our data, patterns we need to verify

Model checking will be a big part of what we do in this class...

The score function and Fisher information

Given a likelihood function $\mathcal{L}(\theta)$ and a log-likelihood $\mathsf{l}(\theta) = \log \mathcal{L}(\theta)$, the score function is just the derivative

$$\mathsf{S}(\theta) = \mathsf{l}'(\theta)$$

Using this definition, the MLE is just the solution to $\mathsf{S}(\theta) = 0$

The (observed) Fisher information is defined to be the second derivative of the log-likelihood $\mathrm{I}(\theta) = -\mathsf{l}''(\theta)$ -- In what sense does this represent information?

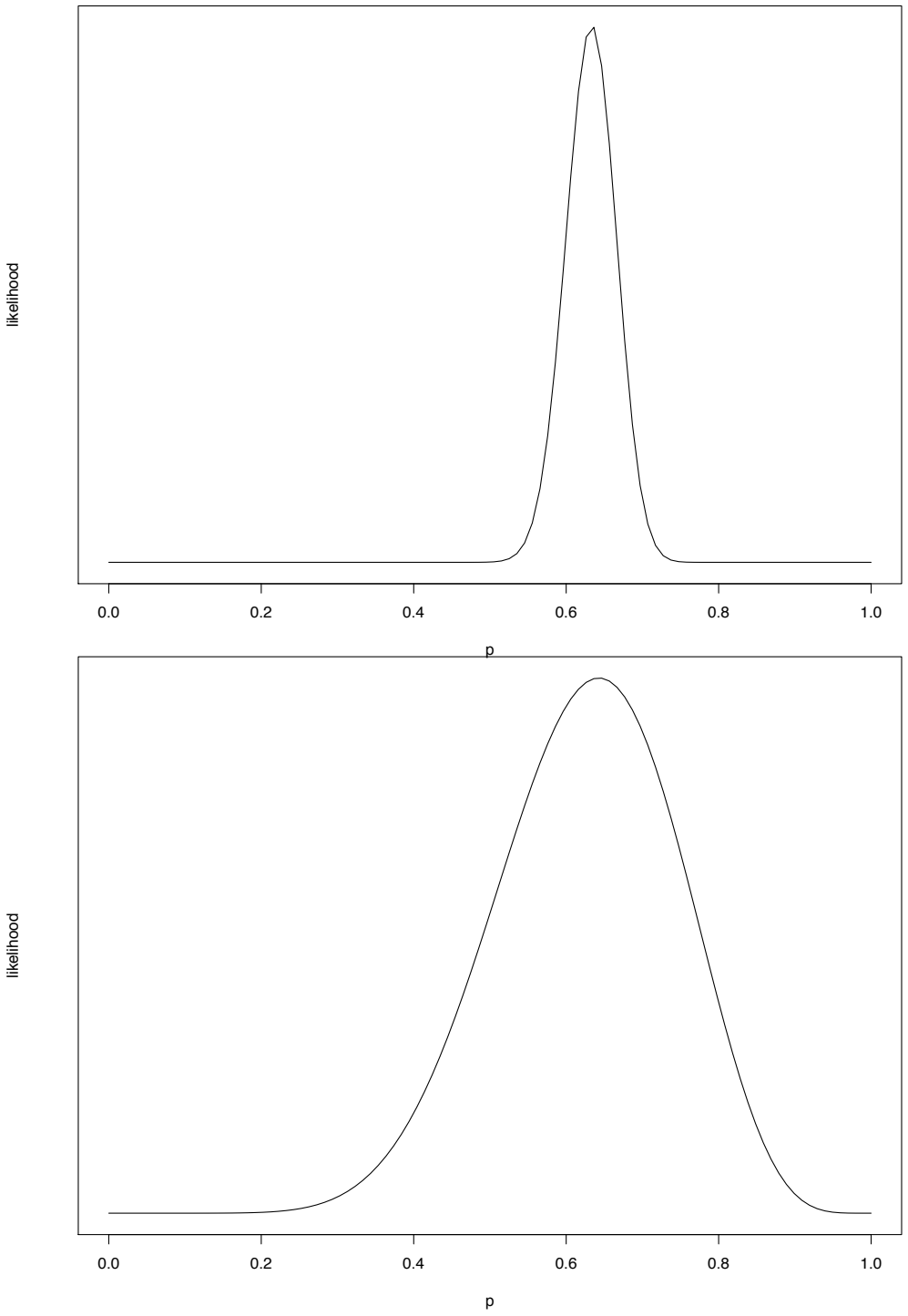
Returning to the binomial

At the top right we have the likelihood associated with our n=15 observations from binomial (14,p)

Below we have the likelihood associated with just a single of our 15 observations, $X_5 = 9$

So the top plot represents a “learning” problem when we have n=15 observations and the bottom is the same problem but using just one observation

What do you notice?



Fisher information

The observed information measures t**he curvature of the log-likelihood at its maximum** -- The more peaked the likelihood, the greater the information in the data about the parameter

It will turn out that this quantity (or, rather, its expected value) will provide us with a bound on **the precision with which we can estimate a parameter** -- More information means greater precision

Computing the MLE

In the examples we have seen so far, we could come up with closed-form expressions for the MLEs (and they were usually sensible things like means and variances) -- In many cases, however, we cannot simply write down the form of our estimator and instead we have to perform **some kind of numerical optimization**

Thankfully, the likelihood function is often well behaved, or rather, is concave in the parameters we're trying to estimate; that means our objective function has just one peak -- **Simple Newton-Raphson iterations perform quite well...**

Newton-Raphson

Suppose we have just a single parameter θ that we're trying to estimate, and let's consider the log-likelihood

Given an initial guess θ_0 for our MLE, Newton-Raphson starts by forming a quadratic approximation (Taylor's theorem!) to the log-likelihood at

$$l(\theta) = l(\theta_0) + l'(\theta_0)(\theta - \theta_0) + \frac{1}{2} l''(\theta_0)(\theta - \theta_0)^2 + \cdots$$

Substituting in for the (observed) Fisher information and the score function, we can write this expression as

$$l(\theta) = l(\theta_0) + S(\theta_0)(\theta - \theta_0) - \frac{1}{2} I(\theta_0)(\theta - \theta_0)^2 + \cdots$$

Newton-Raphson

The maximum of this quadratic occurs at the point

$$\theta_1 = \theta_0 + \frac{\mathrm{S}(\theta)}{\mathrm{I}(\theta)}$$

This suggests a simple updating rule -- Starting from an initial guess, form the quadratic approximation and move to its maximum

The binomial

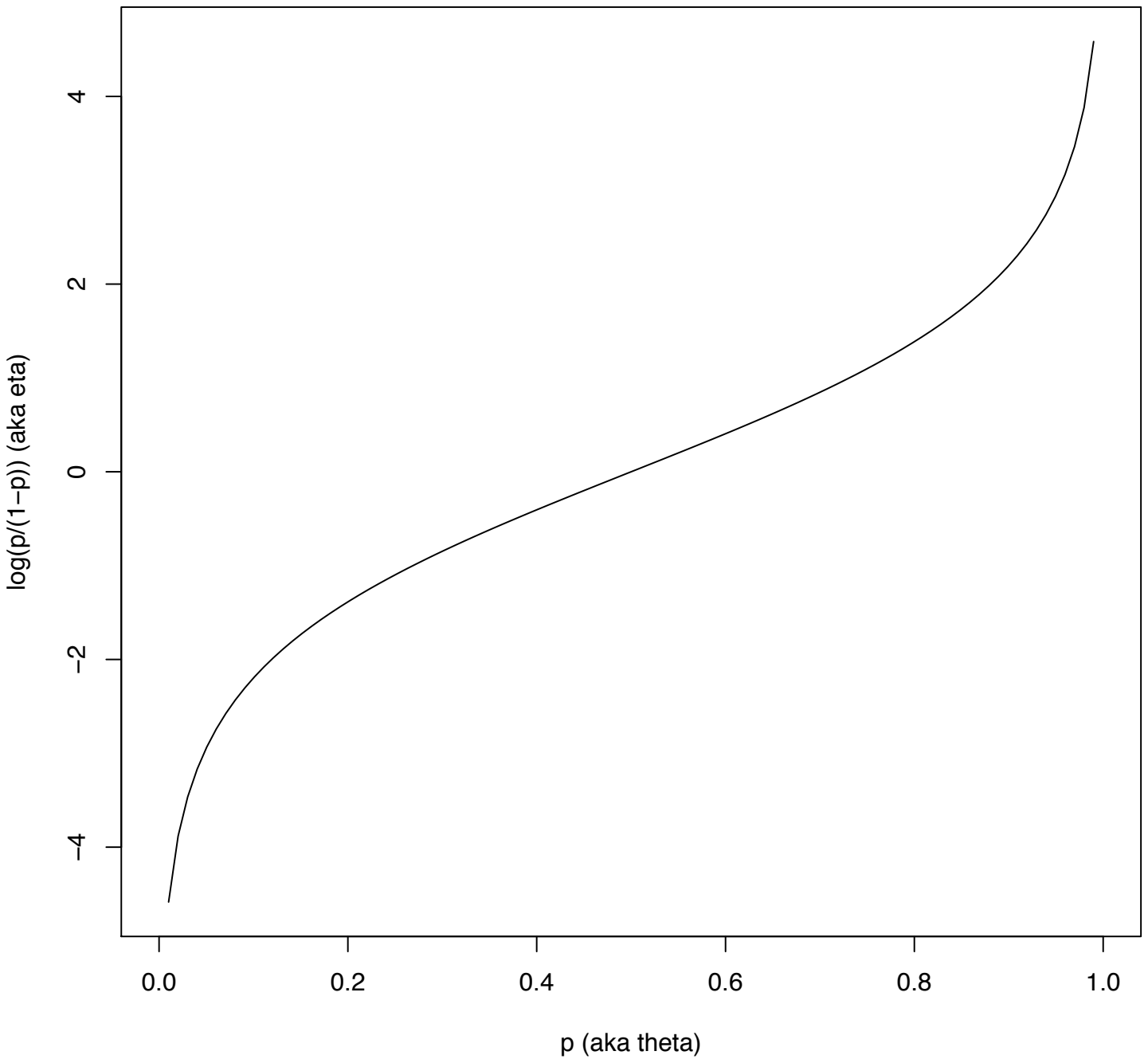
The binomial probability function looks like

$$\begin{aligned}\binom{m}{x} p^x (1-p)^{m-x} &= \binom{m}{X} \exp\left[x \log \frac{p}{1-p} + m \log(1-p)\right] \\ &= \binom{m}{x} \exp\left[x\,\eta + m \log(1+e^{\eta})\right]\end{aligned}$$

where we have rewritten things in terms of the so-called "natural parameter" η

$$\eta = \log \frac{p}{1-p} \quad \text{where} \quad p = \frac{e^{\eta}}{1+e^{\eta}}$$

Why might we prefer to perform optimization over η rather than p?


log(p/(1−p)) (aka eta)
4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0
p (aka theta)

Aside: Reparametrizing

In this case, we have reparametrized the problem and instead of using p to index the binomial family, we are using log p/(1-p)

We refer to p/(1-p) as the odds ratio -- If the odds in favor of an event is greater than one, it means the probability associated with that event is bigger than 1/2

The binomial

If we have just one observation from a binomial, our log-likelihood function will look like (where C doesn't depend on η)

$$\mathsf{l}(\eta) = \mathsf{X}\eta - \mathsf{m}\log(1 + \mathsf{e}^{\eta}) + \mathsf{C}$$

with score and (observed) Fisher information

$$\mathsf{l}'(\eta) = \mathsf{X} \quad - \mathsf{m}\,\frac{\mathsf{e}^{\eta}}{1 + \mathsf{e}^{\eta}} \quad \text{and} \quad -\mathsf{l}''(\eta) = \mathsf{m}\frac{\mathsf{e}^{\eta}}{1 + \mathsf{e}^{\eta}}\,\frac{1}{1 + \mathsf{e}^{\eta}}$$

which we can write as $\mathsf{X} - \mathsf{mp}$ and $\mathsf{mp}(1 - \mathsf{p})$, respectively

The binomial

This means that the Newton-Raphson iterations take the form

$$\eta_1 = \eta_0 + \frac{X - mp_0}{mp_0(1 - p_0)}$$

Can you tell me something about this?

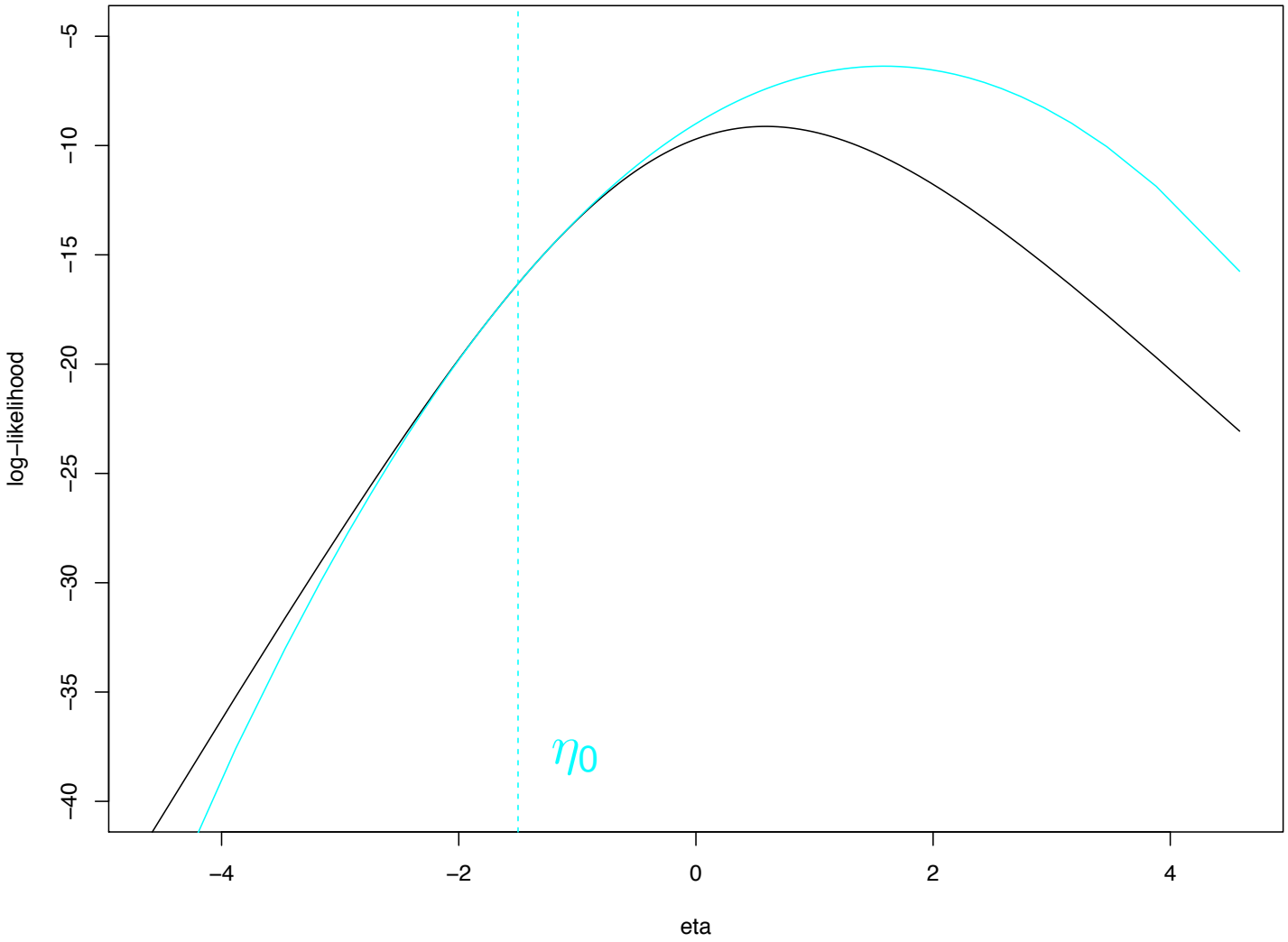
Example

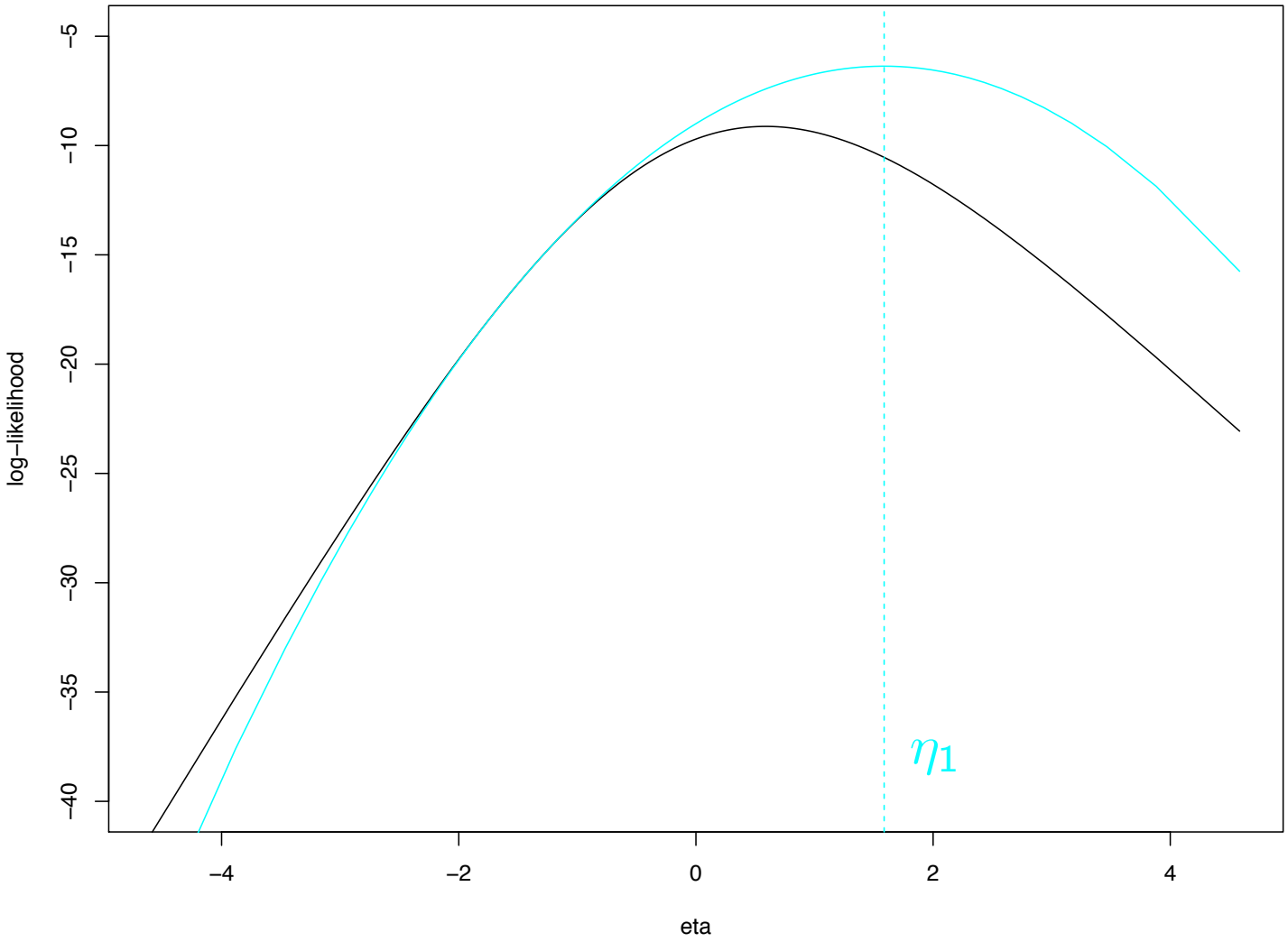
Let's consider the fifth data point again for Nestbox 8 -- Recall that the number of times we saw the bird (out of 14 days) in this time period was 9

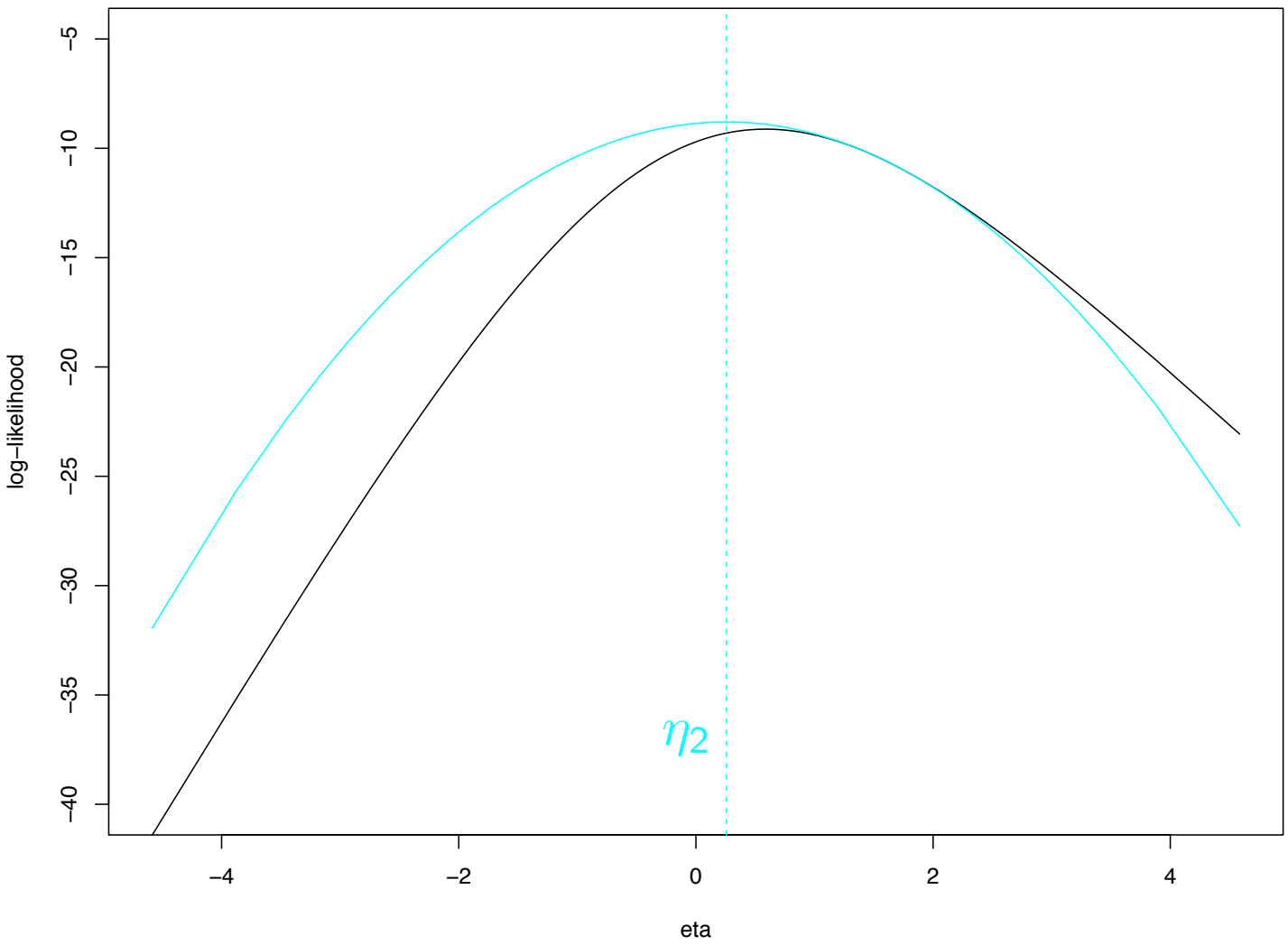
Here is what we get if we iterate 10 times, using a value of -1.5 as our starting point (recall that we know the MLE is 9/14 = 0.64)

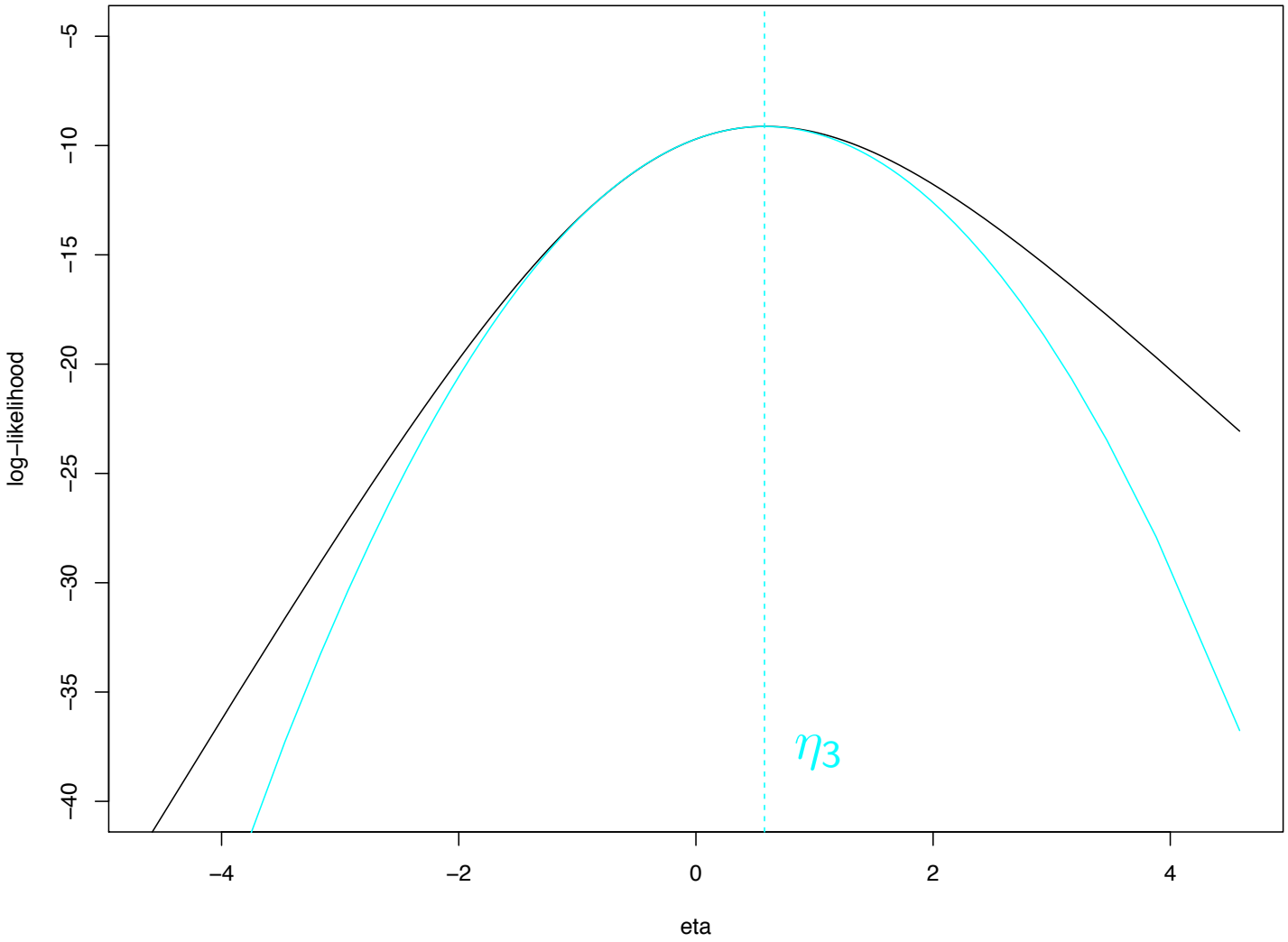
```
eta0 <- -1.5
p0 <- exp(eta0)/(1+exp(eta0))
print(c(eta0,p0))

for(i in 1:10){

  eta1 <- eta0 + (9/14 - p0)/(p0*(1-p0))
  p1 <- exp(eta1)/(1+exp(eta1))

  print(c(eta1,p1))

  p0 <- p1
  eta0 <- eta1
}
```

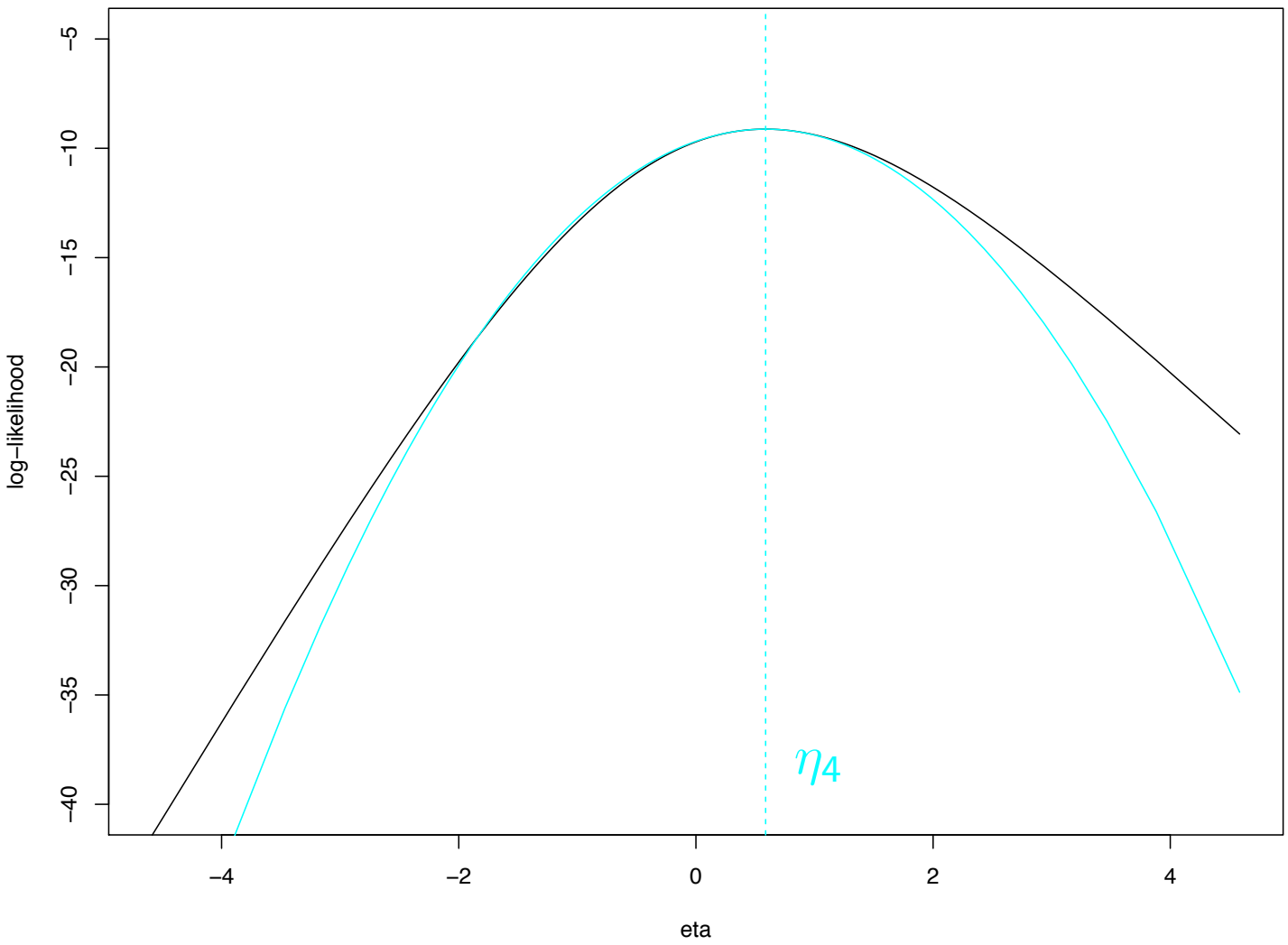
```
[1] -1.500000 0.1824255
[1] 1.5871108 0.8302092
[1] 0.2580132 0.5641478
[1] 0.5781193 0.6406345
[1] 0.5877735 0.6428541
[1] 0.5877867 0.6428571
[1] 0.5877867 0.6428571
[1] 0.5877867 0.6428571
[1] 0.5877867 0.6428571
[1] 0.5877867 0.6428571
[1] 0.5877867 0.6428571
```

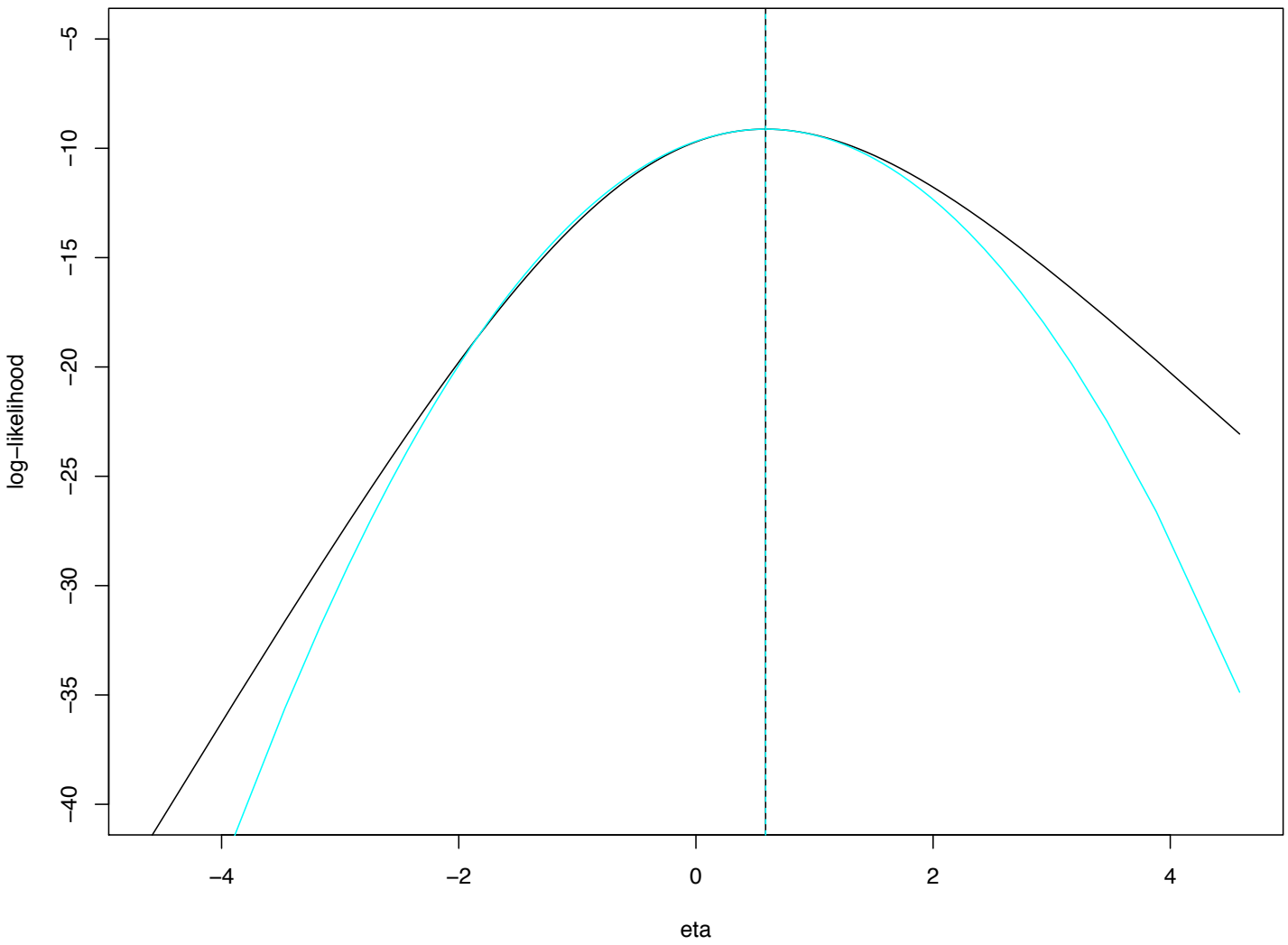

log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_0


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_1


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_2


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_3


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta
η_4


log−likelihood
−5
−10
−15
−20
−25
−30
−35
−40
−4
−2
0
2
4
eta

A generalization

Both the normal and the binomial are examples of a larger class known as **an exponential family** of probability distributions

For a single parameter, members of this family take the form

$$f(x|\theta) = h(x)\, e^{\eta(\theta)T(x) - B(\theta)}$$

and for more than one parameter

$$f(x|\theta) = h(x)\, \exp\left[\sum_{k=1}^{K} \eta_k(\theta) T_k(x) - B(\theta)\right]$$

The binomial (take 2)

Rewriting the binomial we see

$$\binom{m}{x} p^x (1-p)^{m-x} = \binom{m}{x} \exp\left[x \ \log p/(1-p) + m \log(1-p)\right]$$

which is in the required form (match terms!) once we recall that the parameter of interest in this model is p

$$\underbrace{\binom{m}{x}}_{h(x)} \exp\Big[\underbrace{x}_{T(x)} \ \underbrace{\log p/(1-p)}_{\eta(p)} + \underbrace{m \log(1-p)}_{B(p)}\Big]$$

Exponential families

A number of the probability distributions in your textbook can be rewritten in this way -- Here are a few more

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$ Normal

$$\frac{e^{-\lambda}\lambda^x}{x!}$$ Poisson

$$\binom{m}{x} p^x (1-p)^{m-x}$$ Binomial (with known m, the number of trials)

$$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$ Gamma

$$\binom{x+r-1}{x} (1-p)^r p^x$$ Negative binomial (where r, the number of failures until a success, is known)

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$ Beta

Exponential families

If $f(x|\theta)$ belongs to a single-parameter exponential family, our likelihood inherits the same special form -- That is,

$$\begin{aligned}\mathcal{L}(\theta) &= \prod_{i=1}^{n} h(X_i) e^{\eta(\theta) T(X_i) - B(\theta)} \\ &= e^{\eta(\theta) \sum_i T(X_i) - nB(\theta)} \prod_i h(X_i)\end{aligned}$$

Exponential families

The log-likelihood can be written as follows (where C doesn't involve θ)

$$l(\theta) = \eta(\theta)T - nB(\theta) + C$$

where we let $T = \sum_{i=1}^{n} T(X_i)$

Given an exponential family, we call T the **natural sufficient statistic** for θ -- From the point of view of model fitting, we see that **maximizing the log-likelihood will involve the data only through T**

Sufficiency

The concept of sufficiency arises more generally as an attempt to answer the question **"Is there a statistic, some function of the data, that contains all the information in the sample about the parameter of interest?"**

In general, a statistic T is said to be sufficient for θ if the conditional distribution of $X_1, \ldots, X_n$ given T=t does not depend on θ for any value of t

In words, it means given some value of T we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, \ldots, X_n$

Sufficiency

What we've done here is replace our data $X_1, \ldots, X_n$ with a single value T that depends on the data -- If we were to repeat our experiment, we would have a new set of observations and an new statistic T

With the binomial family, for example, we wrote

$$\underbrace{\binom{m}{x}}_{h(x)} \exp\Big[\underbrace{x}_{T(x)} \underbrace{\log p/(1-p)}_{\eta(p)} + \underbrace{m \log(1-p)}_{B(p)}\Big]$$

Sufficiency

And so our likelihood is

$$\exp\left[\left(\log\frac{p}{1-p}\right)\sum_{i=1}^{n}X_i - nm\log(1-p)\right]\prod_{i=1}^{n}\binom{m}{X_i}$$

where the sufficient statistic is just $T = \sum_{i=1}^{n} X_i$

Aside: Natural exponential families

Recall the form for an exponential family

$$f(x|\theta) = h(x)\, e^{\eta(\theta)T(x) - B(\theta)}$$

It is often the case that η is a one-to-one function of θ, meaning we can drop θ altogether and work with η instead -- In the Binomial example, we have that

$$\eta = \log \frac{p}{1-p}$$

which is invertible,

$$p = \frac{e^{\eta}}{1 + e^{\eta}}$$

Aside: Natural exponential families

If we can make this reduction and write our family in the form

$$f(x|\eta) = h(x)e^{\eta T(x) - B(\eta)}$$

we refer to f as being part of **a natural exponential family**

In this case, we can prove that

1. $E\,T(X) = B'(\eta)$
2. $\text{var}\,T(X) = B''(\eta)$

Aside: Natural exponential families

In this case, maximum likelihood takes on a simple form -- For a natural exponential family where

$$f(x|\eta) = h(x)e^{\eta T(x) - B(\eta)}$$

the likelihood for n observations $X_1, \ldots, X_n$ is given by

$$\prod_{i=1}^{n} f(X_i|\eta) = e^{\eta \sum_{i=1}^{n} T(X_i) - nB(\eta)} \prod_{i=1}^{n} h(X_i)$$

and the log-likelihood is

$$l(\eta) = \eta T - nB(\eta) + C$$

where $T = \sum_{i=1}^{n} T(X_i)$ is the natural sufficient statistic and C doesn't depend on η

Aside: Natural exponential families

Taking a derivative with respect to η yields

$$l'(\eta) = T - nB'(\eta)$$

Setting the derivative to zero we find

$$\frac{1}{n}\sum_{i=1}^{n} T(X_i) = B'(\eta)$$

which means the MLE is the value $\widehat{\eta}$ that equates the empirical mean (based on our data points $X_1, \ldots, X_n$ of T(X) with its theoretical mean!

Rationale

We present this material because maximum likelihood, as a methodology has had quite a history in (and out of) statistics -- With a little bit of rigor, you can start to appreciate some of that effort and have access to some of the fundamental concepts in the field

Next time we will examine how to fit MLEs and the use of an exponential family will add a layer of interpretability to the process

Another estimation approach

While maximum likelihood has a lot to recommend it, it is certainly not the only game in town

The method of moments is another estimation procedure that has at least as long a history as maximum likelihood (going back to another Bernoulli)

Method of moments

The $k\text{th}$ moment of a probability distribution is defined to be (as one might expect)

$$\mu_k = EX^k$$

Given a sample of data $X_1, \ldots, X_n$, we can also form the $k\text{th}$ sample moment

$$\widehat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

The method of moments

Suppose we have a two parameter family with (θ_1, θ_2); suppose we could write these parameters in terms of the first two moments

$$\theta_1 = g_1(\mu_1, \mu_2) \quad \text{and} \quad \theta_2 = g_2(\mu_1, \mu_2)$$

Then, the method of moments estimate simply replaces the theoretical quantities with their sample counterparts

$$\widehat{\theta}_1 = g_1(\widehat{\mu}_1, \widehat{\mu}_2) \quad \text{and} \quad \widehat{\theta}_2 = g_2(\widehat{\mu}_1, \widehat{\mu}_2)$$

The method of moments

The first few moments of the normal distribution are given as follows; we'll use them in a moment so they're worth recording

$$\mu_1 = \mu$$

$$\mu_2 = \mu^2 + \sigma^2$$

$$\mu_3 = \mu(\mu^2 + 3\sigma^2)$$

$$\mu_4 = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$$

The normal family

The first and second moments of the normal distribution are given by

$$\mu_1 = EX = \mu \quad \text{and} \quad \mu_2 = EX^2 = \mu^2 + \sigma^2$$

which we can rewrite in terms of the parameters

$$\mu = \mu_1 \quad \text{and} \quad \sigma^2 = \mu_2 - \mu_1^2$$

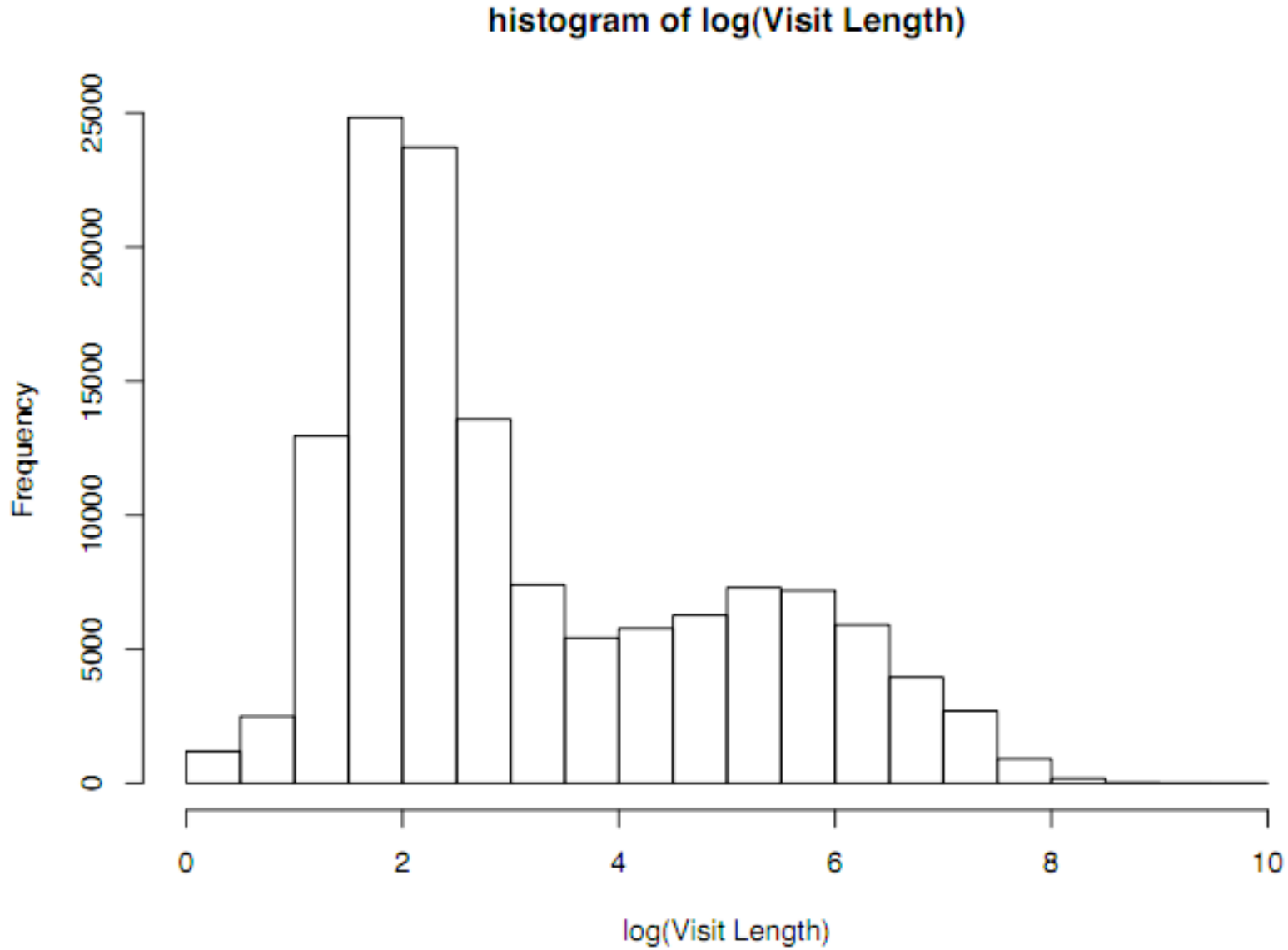
yielding estimates of the form

$$\widehat{\mu} = \overline{X} \quad \text{and} \quad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = S^2$$

The normal family

In many situation, we use the normal not as an endpoint, but as an ingredient in a more complex model -- We have already seen one instance in which a single normal seems implausible as an “origin story” but some combination of normals might work out just right

Recall the logarithm of visit lengths from the Travel Section experiment...

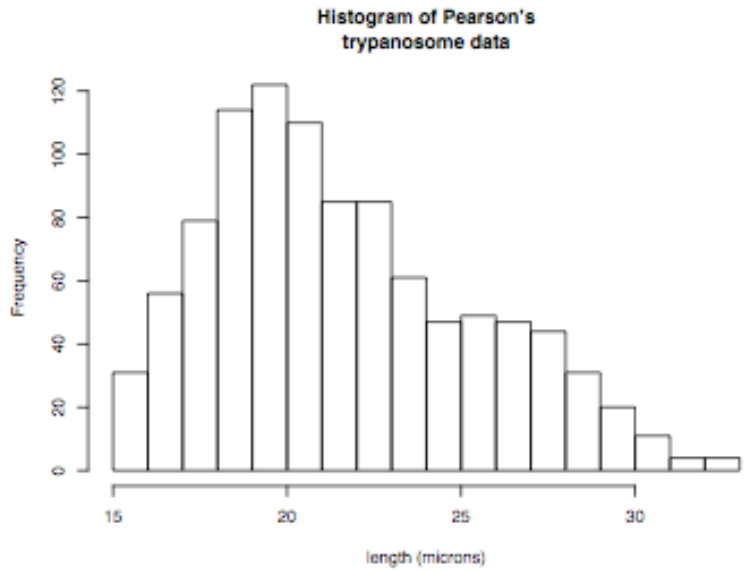

histogram of log(Visit Length)
Frequency
0
5000
10000
15000
20000
25000
0
2
4
6
8
10
log(Visit Length)

Normal mixtures

Biologists were among the first to need methods for analyzing data from mixtures; “a biologist would often have a sample of measurements of some physical characteristic that would contain observations from more than one species, or a sample from one species that would contain observations from more than one age group”

At the same time, there was often not a way to classify these items further, properly separating species or age groups

At the right is the data that started it all; Pearson used these to develop a method to fit a normal mixture



When a series of measurements gives rise to a normal curve, we may probably assume something approaching a stable condition; there is a production and destruction impartially around the mean. In the case of certain biological, sociological, and economic measurements there is, however, a well-marked deviation from this normal shape, and it becomes important to determine the direction and amount of such deviation. The asymmetry may arise from the fact that the units gruped together in the measured material are not really homogeneous. It may happen that we have a mixture of 2, 3, ... n homogeneous groups, each of which deviates about its own mean symmetrically and in a manner represented with sufficient accuracy by the normal curve... Even where the material is really homogeneous, but gives and abnormal frequency-curve, the amount and direction of the abnormality will be indicated if this frequency-curve can be split up into normals.

Pearson (1894)

Normal mixtures

We can think of data from a mixture model as being generated in two stages, starting with a coin toss

At the first stage, we define Y such that

$$\Pr(Y = 1) = \alpha \quad \text{and} \quad \Pr(Y = 0) = 1 - \alpha$$

Then, depending on the value of Y we draw a sample from one of two normal distributions

$$f(x|y) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)/2\sigma_1^2}, & y = 0 \\ \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-(x-\mu_2)/2\sigma_2^2}, & y = 1 \end{cases}$$

Normal mixtures

Dusting off our probability, we can write down the marginal density for X

$$
\begin{aligned}
f(x) &= \sum_{y \in \{0,1\}} f(x,y) \\
&= \sum_{y \in \{0,1\}} f(x|y) f(y) \\
&= \alpha \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(x-\mu_1)/2\sigma_1^2} + (1-\alpha) \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-(x-\mu_2)/2\sigma_2^2}
\end{aligned}
$$

Normal mixtures

It turns out that a handful of normal distributions “mixed” in this way can be extremely flexible, approximating a number of different data “shapes”...

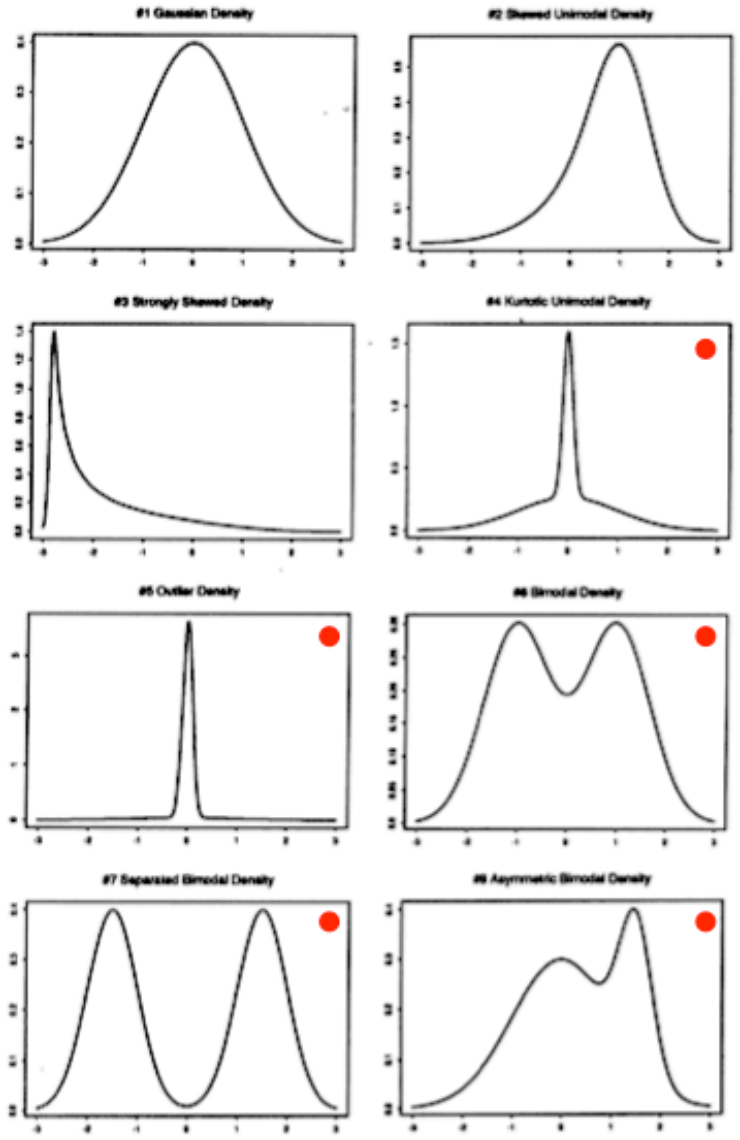


FIG. 1. *Normal mixture densities.*

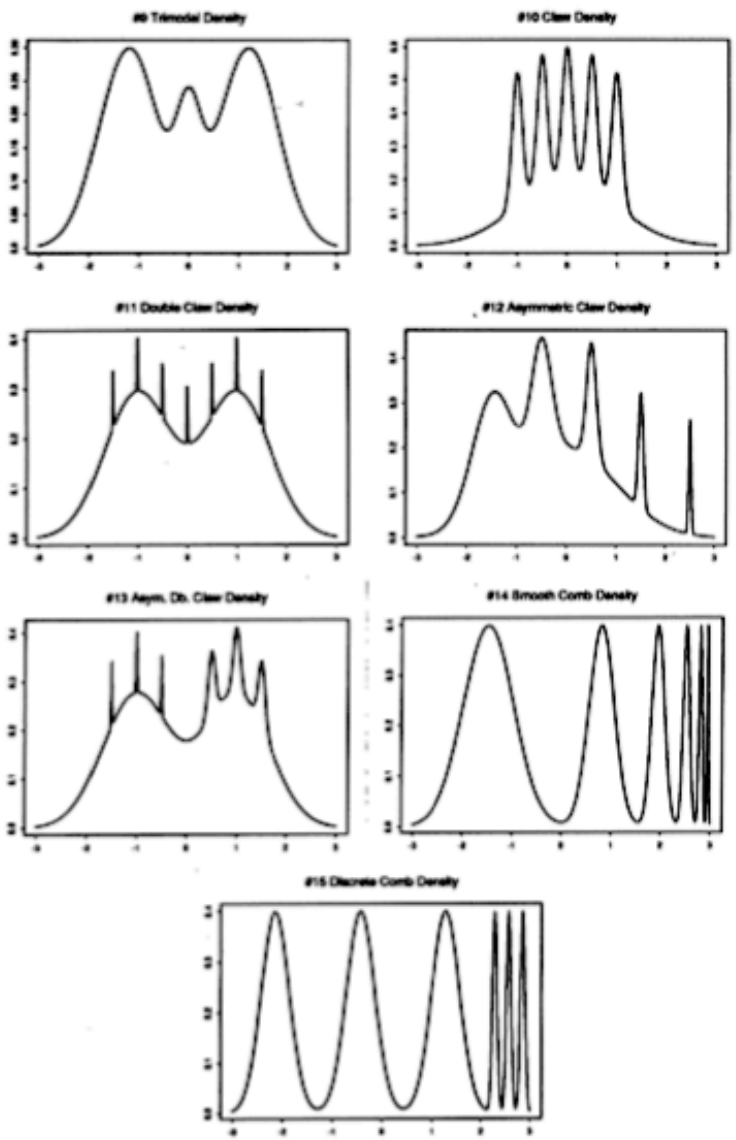


FIG. 1. *Continued.*

Mixtures

General mixture models have been applied in a variety of settings -- Here, for example, a two-component mixture is used to model a manufacturing process

One component represents “zero” defects meaning encountering a problem with a manufactured device is rare -- The second represents a process that is out of control, spitting up a Poisson count of defects

TECHNOMETRICS, FEBRUARY 1992, VOL. 34, NO. 1

Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing

Diane Lambert

AT&T Bell Laboratories
Murray Hill, NJ 07974

Zero-inflated Poisson (ZIP) regression is a model for count data with excess zeros. It assumes that with probability p the only possible observation is 0, and with probability $1 - p$, a Poisson(λ) random variable is observed. For example, when manufacturing equipment is properly aligned, defects may be nearly impossible. But when it is misaligned, defects may occur according to a Poisson(λ) distribution. Both the probability p of the perfect, zero defect state and the mean number of defects λ in the imperfect state may depend on covariates. Sometimes p and λ are unrelated; other times p is a simple function of λ such as $p = 1/(1 + \lambda^{\tau})$ for an unknown constant τ. In either case, ZIP regression models are easy to fit. The maximum likelihood estimates (MLE's) are approximately normal in large samples, and confidence intervals can be constructed by inverting likelihood ratio tests or using the approximate normality of the MLE's. Simulations suggest that the confidence intervals based on likelihood ratio tests are better, however. Finally, ZIP regression models are not only easy to interpret, but they can also lead to more refined data analyses. For example, in an experiment concerning soldering defects on printed wiring boards, two sets of conditions gave about the same mean number of defects, but the perfect state was more likely under one set of conditions and the mean number of defects in the imperfect state was smaller under the other set of conditions; that is, ZIP regression can show not only which conditions give lower mean number of defects but also why the means are lower.

KEY WORDS: EM algorithm; Negative binomial; Overdispersion; Positive Poisson; Quality control.

Standard arguments suggest that, when a reliable manufacturing process is in control, the number of defects on an item should be Poisson distributed. If the Poisson mean is λ, a large sample of n items should have about $ne^{-\lambda}$ items with no defects. Sometimes, however, there are many more items without defects than would be predicted from the numbers of defects on imperfect items (an example is given in Sec. 1). One interpretation is that slight, unobserved changes in the environment cause the process to move randomly back and forth between a perfect state in which defects are extremely rare and an imperfect state in which defects are possible but not inevitable. The transient perfect state, or existence of items that are unusually resistant to defects, increases the number of zeros in the data.

This article describes a new technique, called *zero-inflated Poisson* (ZIP) regression, for handling zero-inflated count data. ZIP models without covariates have been discussed by others (for example, see Cohen 1963; Johnson and Kotz 1969), but here both the probability p of the perfect state and the mean λ of the imperfect state can depend on covariates. In particular, $\log(\lambda)$ and $\text{logit}(p) = \log(p/(1 - p))$ are assumed to be linear functions of some covariates. The same or different covariates might affect p and λ, and p and λ might or might not be functionally related. When p is a decreasing function of λ, the probability of the perfect state and the mean of the imperfect state improve or deteriorate together.

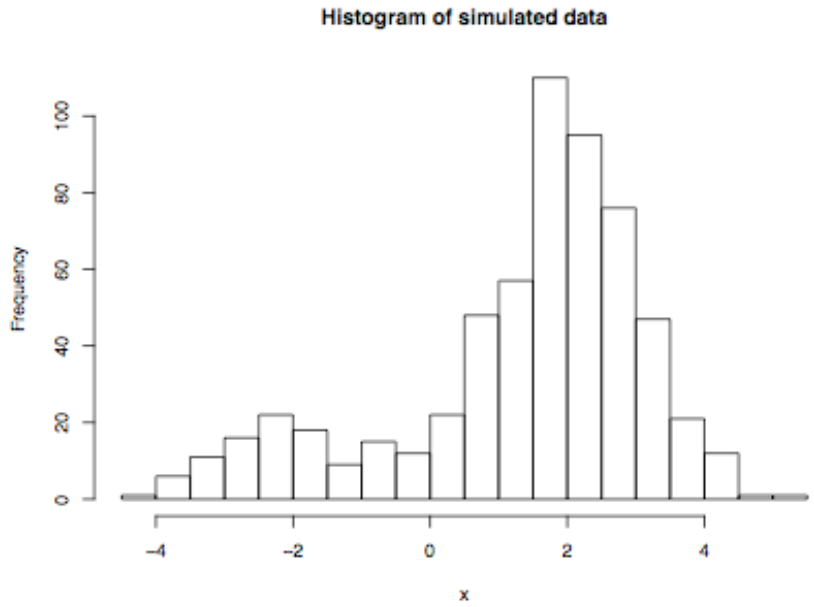
Heilbron (1989) concurrently proposed similar zero-altered Poisson and negative binomial regression models with different parameterizations of p and applied them to data on high-risk behavior in gay men. (Although the models were developed independently, the acronym ZIP is just an apt modification of Heilbron's acronym ZAP for zero-altered Poisson.) He also considered models with an arbitrary probability of 0. Arbitrary zeros are introduced by mixing point mass at 0 with a positive Poisson that assigns no mass to 0 rather than a standard Poisson.

Other authors have previously considered mixing a distribution degenerate at 0 with distributions other than the Poisson or negative binomial. For example, Feuerverger (1979) mixed zeros with a gamma distribution and coupled the probability of 0 with the mean of the gamma to model rainfall data. Farewell (1986) and Meeker (1987) mixed zeros with right-censored continuous distributions to model survival data when some items are indestructible and testing

An example

Before looking at the visit data, we'll consider a toy example -- This one will let us kick the tires on our fitting routines, on graphical displays and so on

Here we have a mixture of two normals; tell me about what you see in the histogram

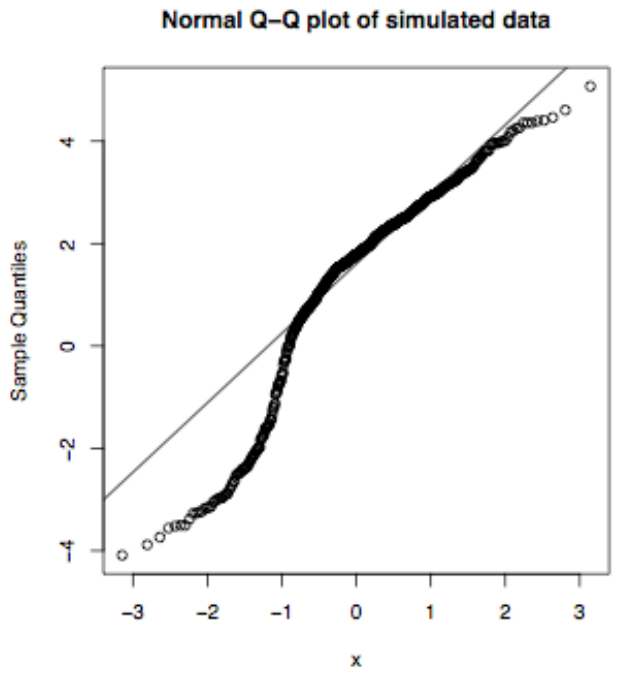


An example

For simplicity, we set $\sigma_1 = \sigma_2$ in our simulations; the general case being treated a little later

At the right we present a normal Q-Q plot of the simulated data; tell me something about the structures in this plot

Did we expect this?



Fitting normal mixtures

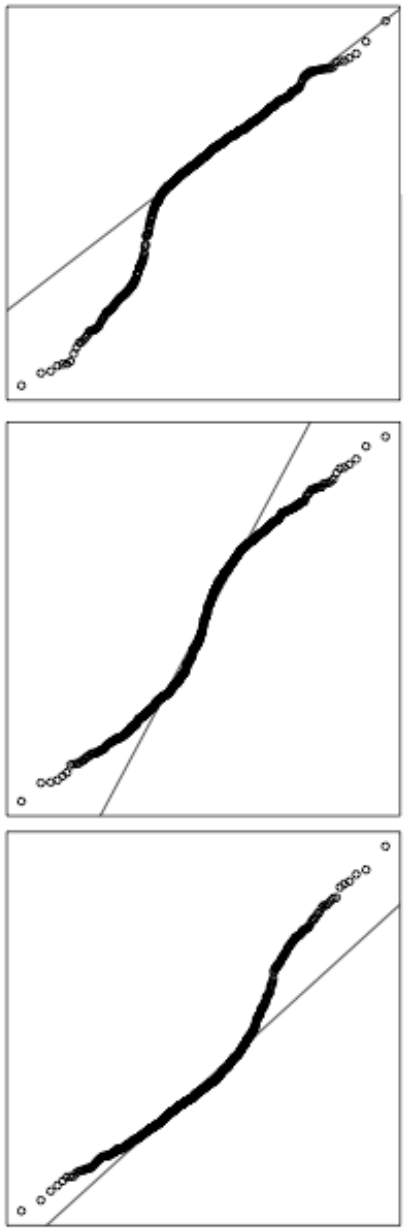
Pearson (1894) was the first to propose a computational scheme for fitting normal mixtures

Much of the subsequent work in this area, however, was graphical or semi-graphical; it would remain this way until computer power crossed a threshold

Graphical methods

For example, Harding (1948) shows that the inflection point x_I of these curves can be used to approximate the mixing proportion

Specifically, $\Phi(x_I)$ is a good estimate, assuming $\sigma_1 = \sigma_2$ and α is not too extreme (say between 0.3 and 0.7)



Method of moments

Pearson's original approach to fitting a normal mixture involved the method of moments

We'll repeat his work, but will assume that $\sigma_1 = \sigma_2$; the reason will be clear in a little while; for those who are interested, this discussion borrows heavily from Cohen (1967)

That leaves us with four parameters to estimate, meaning we have to work with four theoretical/empirical moments

Method of moments

Cohen (1967) starts by computing the first moment of the mixture distribution (notice we've changed notation here and we're putting a * on the theoretical moments to distinguish them from the terms in the normal mixture)

$$\begin{aligned}
\mu_1^* &= \int x f(x) dx \\
&= \alpha \int x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_1)^2/2\sigma^2} dx + (1-\alpha) \int x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu_2)^2/2\sigma^2} dx \\
&= \alpha\mu_1 + (1-\alpha)\mu_2
\end{aligned}$$

He then uses this to define the central moments (central in the sense that they're computed about the mean)

$$\mu_k^* = \int (x - \mu_1^*)^k f(x) dx$$

Method of moments

These will then be equated with their sample equivalents, $\overline{X}$ and the central moments

$$v_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^k, \quad k \geq 2$$

Finally, we introduce the variables

$$m_1 = \mu_1 - \mu_1^* \quad \text{and} \quad m_2 = \mu_2 - \mu_1^*$$

With the assumption that $\mu_1 < \mu_2$, we have $m_1 \leq 0 \leq m_2$

Method of moments

We can use the facts we gave earlier for the (non-central) moments of the normal distribution to derive the following for equations

1. $\alpha m_1 + (1-\alpha)m_2 = 0$
2. $\alpha[\sigma^2 + m_1^2 - v_2] + (1-\alpha)[\sigma^2 + m_2^2 - v_2] = 0$
3. $\alpha[3\sigma^2 m_1 + m_1^3 - v_3] + (1-\alpha)[3\sigma^2 m_2 + m_2^3 - v_3] = 0$
4. $\alpha[3\sigma^4 + 6m_1^2\sigma^2 + m_1^4 - v_4] + (1-\alpha)[3\sigma^4 + 6m_2^2\sigma^2 + m_2^4 - v_4]$

Four equations, four unknowns, let's see what kind of trouble we can get into!

Method of moments

Following Cohen (1967), we introduce a couple of auxiliary variables; namely

$$r = m_1 + m_2 \quad \text{and} \quad v = m_1 m_2$$

With these, we can re-express our equations as follows

1. $\alpha = m_2/(m_2 - m_1)$
2. $\sigma^2 = v + v_2$
3. $vr = -v_3$
4. $2v^2 + vr^2 = -(v_4 - 3v_2^2)$

Method of moments

Cohen (1967) spells out how to solve these equations; starting with the auxiliary variables, he notes that the equation

$$2v^3 - (v_4 - 3v_2^2)v + v_3^2 = 0$$

has only one negative root (what we're after; recall that $m_1 m_2 < 0$ by construction)

Having found that, we can unwind the remaining equations using the recipe at the right

$$\widehat{r} = -v_3/\widehat{v}$$

$$\widehat{\sigma}^2 = \widehat{v} + v_2$$

$$\widehat{m}_1 = \tfrac{1}{2}[\widehat{r} - \sqrt{\widehat{r}^2 - 4\widehat{v}}]$$

$$\widehat{m}_2 = \tfrac{1}{2}[\widehat{r} + \sqrt{\widehat{r}^2 - 4\widehat{v}}]$$

$$\widehat{\mu}_1 = \widehat{m}_1 + \overline{X}$$

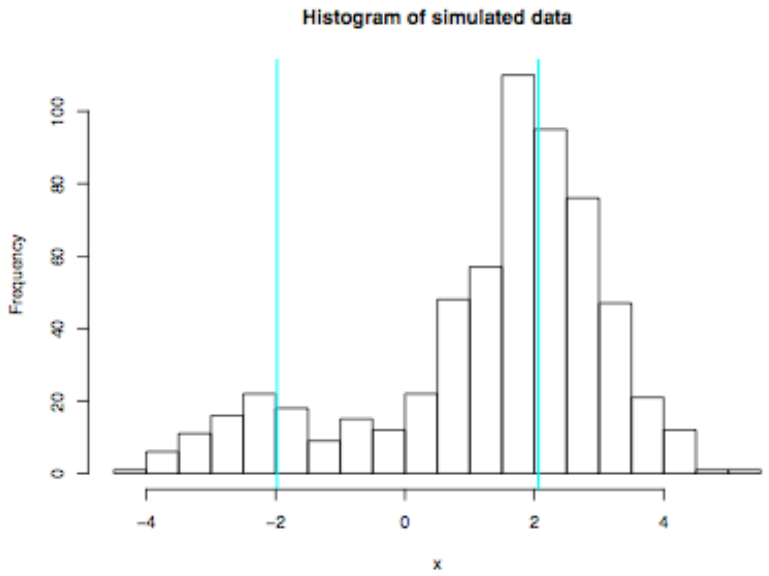
$$\widehat{\mu}_2 = \widehat{m}_2 + \overline{X}$$

$$\widehat{\alpha} = \widehat{m}_2/(\widehat{m}_2 - \widehat{m}_1)$$

Method of moments

Right, so this seems a little tedious, but the end is pretty simple to implement

The histogram is our toy simulated data ($\sigma = 1$, $\mu_1 = -2$, $\mu_2 = 2$, and $\alpha = 0.15$) and the vertical bars mark our estimates of the means




```
# compute moments
xb = mean(x)
v2 = mean((x-xb)^2)
v3 = mean((x-xb)^3)
v4 = mean((x-xb)^4)
k4 = v4-3*v2^2

# form estimates
roots = polyroot(c(v3^2,k4,0,2))
v = Re(roots[2])  # here, 2nd root is negative
r = -v3/v
sig2 = v+v2
m1 = 0.5*(r-sqrt(r^2-4*v))
m2 = 0.5*(r+sqrt(r^2-4*v))
mu1 = m1+xb
mu2 = m2+xb
alpha = m2/(m2-m1)

# look at fit (!)
print(c(mu1,mu2,sig2,alpha))
hist(x)
abline(v=c(mu1,mu2),col=5,lwd=2)
```

Method of moments

Keep in mind that we've only tackled a reduced problem; namely, we've assumed that $\sigma_1 = \sigma_2$

If we drop this assumption, then the first step in our solution is not a cubic polynomial, but a ninth-degree polynomial!

The solution of an equation of the ninth degree, where almost all powers, to the ninth, of the unknown quantity are existing, is, however, a very laborious task. Mr. Pearson has indeed possessed the energy to perform his heroic task... But I fear he will have few successors...

Charlier (1906)