Likelihood: Philosophical foundations

Patrick Breheny

August 23

Introduction

- This course is about developing a theoretical understanding of likelihood, the central concept in statistical modeling and inference
- Likelihood offers a systematic way of measuring agreement between unknown parameters and observable data; in so doing, it provides a unifying principle that all statisticians can agree is reasonable (which is not to say that it doesn't have any problems)
- This provides two enormous benefits in statistics:
  - A universal baseline of comparison for methods and estimators
  - A coherent and versatile method of summarizing and combining evidence of all sources without relying on arbitrary, ad hoc decisions

Likelihood: Definition

- Let X denote observable data, and suppose we have a probability model p that relates potential values of X to an unknown parameter θ
- **Definition:** Given observed data $X = x$, the *likelihood function* for θ is defined as

$$L(\theta|x) = p(x|\theta),$$

  although I will often just write $L(\theta)$
- Note that this is a function of θ, not x; once the data has been observed, x is fixed
- Also, note that a likelihood function is not a probability distribution – for example, it does not integrate to 1

Likelihood for continuous distributions

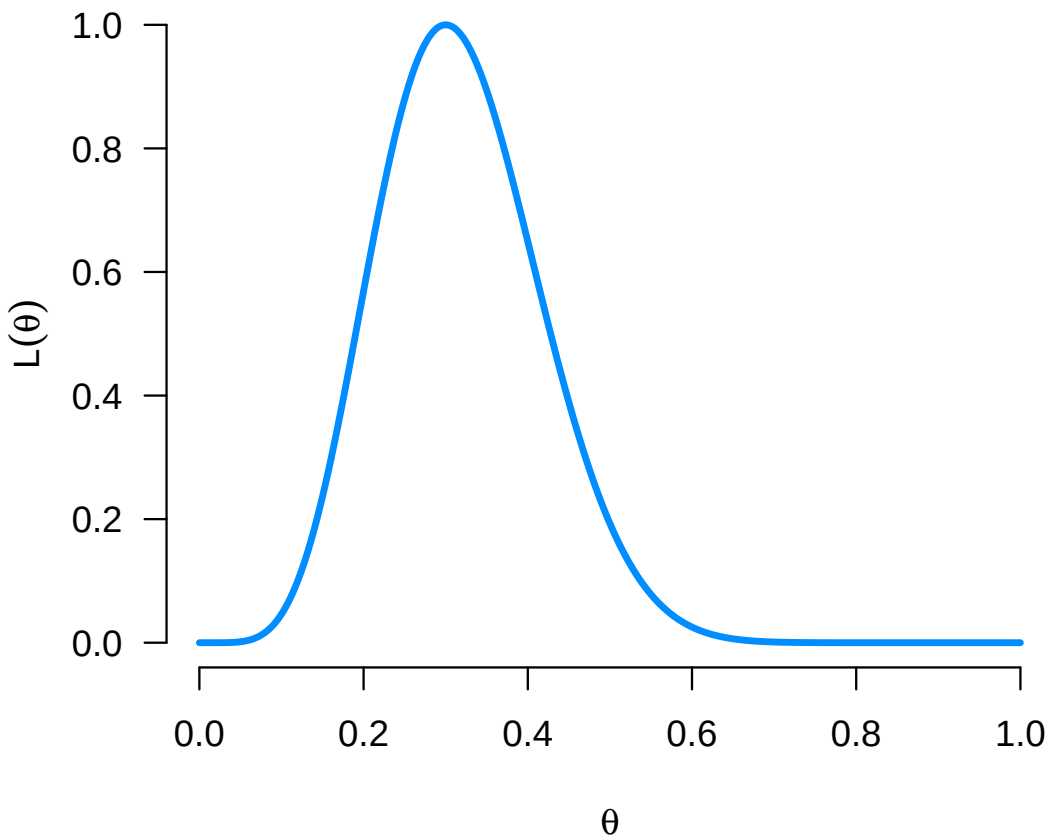
- The definition on the previous slide works regardless of whether p is a mass function, a density function, or even a mixture of the two
- Is it reasonable to mix probabilities and densities like this?
- Suppose we replace the density with the probability $\mathbb{P}\{X \in (x - \epsilon/2, x + \epsilon/2)\}$; then for small ϵ we have

$$\begin{aligned} L(\theta) &= \int_{x-\epsilon/2}^{x+\epsilon/2} p(u|\theta)du \\ &\approx \epsilon p(x|\theta) \end{aligned}$$

- Thus, at least in the limit $\epsilon \to 0$, the value of ϵ is just an arbitrary multiplicative constant and may be ignored; we will come back to this point shortly

Illustration: Binomial success

To get a sense of how likelihood works, let's consider a simple, familiar situation in which a binary trial is independently repeated $n = 20$ times with $x = 6$ successes:



Remarks

- The likelihood is therefore

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

  up to a constant that does not involve θ
- Likelihoods provide only a relative measure of preference for one parameter value vs. another
- In other words, the actual value of $L(\theta)$ is not meaningful, but the relative quantity $L(\theta_1)/L(\theta_2)$ is meaningful
- For this reason, whenever I provide likelihood plots in this course, I will standardize L to have a maximum of 1
- Also, I will use the term "equivalent" to describe two likelihoods that are proportional to each other (i.e., for any meaningful purpose, they are identical)

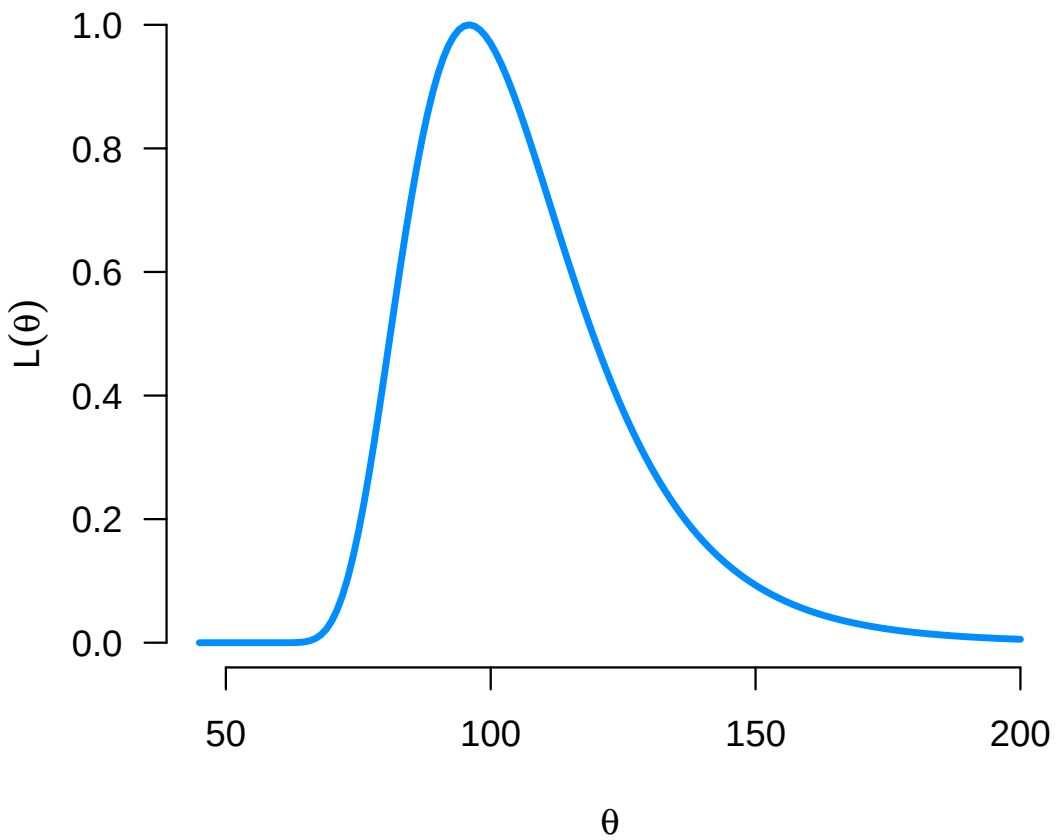
Another example: Mark-recapture

- Let's consider a less-familiar situation: trying to estimate the abundance of an animal species in a certain area
- A common way of doing this is via mark-recapture experiments
- Suppose a department of natural resources marks 30 mountain lions in an area, releases them into the wild, then recaptures 45 of them, of which 14 had been tagged previously
- Assuming the mountain lions are caught at random, the likelihood is given by the hypergeometric distribution:

$$L(\theta) = \frac{\binom{30}{14}\binom{\theta-30}{45-14}}{\binom{\theta}{45}}$$

Illustration: Mark-recapture

This leads to a relatively similar picture as far as the story likelihood tells us about which unknown values are consistent with the data and which are not:



The appeal of likelihood

These are just two examples of the appeal of likelihood: a universal method for obtaining:

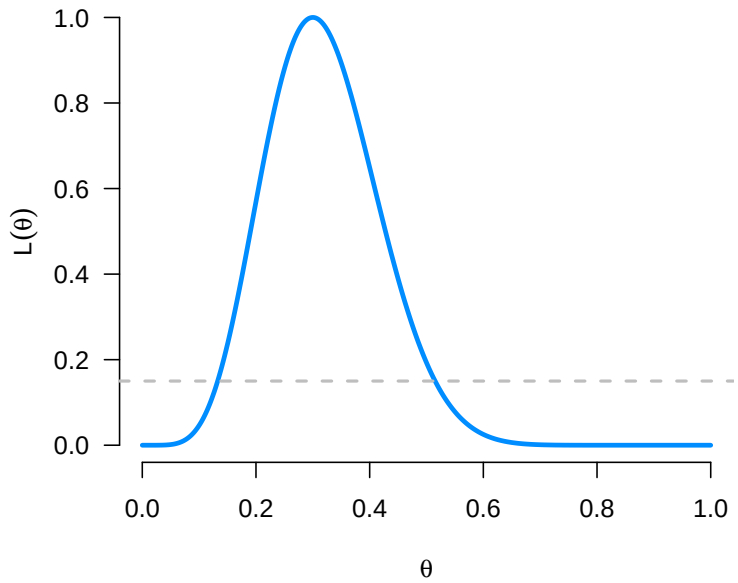
- Point estimates (maximum likelihood)
  - Binomial trials: $\hat{\theta} = 0.3$
  - Mark-recapture: $\hat{\theta} = 96$
- Intervals
  - Binomial trials: $(0.13, 0.51)$
  - Mark-recapture: $(75, 141)$

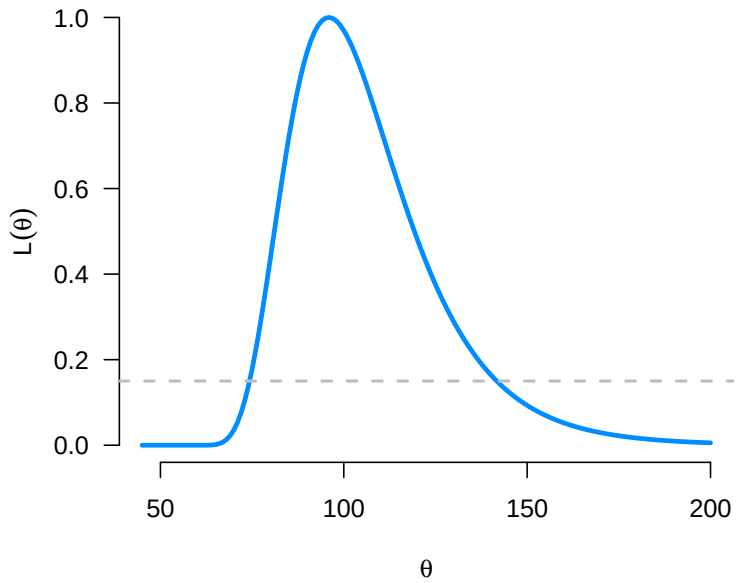
Likelihood ratios

- The intervals on the previous slide are based on

$$\left\{\theta : \frac{L(\theta)}{L(\hat{\theta})} > c\right\}$$

- In this case, I chose $c = 0.15$:





- The obvious question, though, is: how should we choose c and what does it mean?

Bayes' rule

- To answer this question, we have to connect likelihood to probability, and there are two schools of thought for doing so: Bayesian and frequentist
- Let us consider the Bayesian framework first, and treat θ as a random variable:

$$\begin{aligned} p(\theta|x) &= \frac{p(\theta)p(x|\theta)}{p(x)} \\ &\propto p(\theta)L(\theta|x), \end{aligned}$$

where

  - $p(\theta)$ is the *prior*: Our beliefs about the plausible values of our parameter before seeing any data
  - $p(\theta|x)$ is the *posterior*: Our updated beliefs about the plausible values for our parameter after seeing the data
  - $p(x)$ is a normalizing constant typically not of interest

Bayesian approach: Appealing aspects

- Indeed, if we consider θ a random variable, then Bayes' rule is the only mathematically correct way of relating likelihood and probability
- This is one very appealing aspect of the Bayesian approach: there is one universal, coherent approach to all statistical inference (decide on a prior, a model/likelihood, then use Bayes' rule to obtain a posterior)
- Another very appealing property is that we can make direct statements about θ based on what we have seen: assuming a uniform prior on θ in our binomial example, we can say that there is a 95% probability that θ is between 0.136 and 0.509

Bayesian approach: Concerns

The Bayesian approach is an enormously useful and successful approach for quantifying the uncertainty associated with likelihood; nevertheless, two meaningful concerns can be raised:

- Choosing a prior can be hard: A uniform prior might be reasonable for our binomial example, but what about the mark-recapture example? A uniform prior is convenient, but absurd, as it would imply a belief that there may be billions, or even trillions, of mountain lions in the area.
- What do we do with subjective belief? The Bayesian approach is beyond criticism in terms of quantifying one's subjective beliefs, but what can we do with those subjective beliefs? Suppose that a scientist at the FDA had a strong subjective belief that a COVID-19 vaccine was safe and effective . . . is that reasonable grounds for approval?

Frequentist approach: Appealing aspects

- The alternative approach uses long-run frequency to calibrate likelihood
- For example, suppose we could show that by constructing intervals via taking all the θ values such that $L(\theta)/L(\hat{\theta}) > 0.15$, then this contained the true value of θ 95% of the time
- This would also connect likelihood to probability, and provide an objective guarantee about the performance of our approach in a long-run sense

Frequentist approach: Concerns

The objective interpretation of the frequentist approach is appealing, but is subject to even more concerns:

- Easier said than done: It is quite difficult, even for simple problems, to obtain results like the claim made on the previous slide. Typically, approximations are involved – and may break down.
- Awkward interpretation: The approach tells us about our long-run error rates in, say, approving vaccines, but doesn't actually say anything about the specific COVID-19 vaccine that we are considering approving.

Likelihood and probability: Summary

To summarize the lecture so far:

- Likelihood alone can provide relative statements about which values of θ are more compatible with the observed data than others
- However, to make these statements absolute, we need to connect likelihood to probability
- Doing so involves choosing a framework (Bayesian or frequentist), which in turn introduces debate

Combining likelihoods

- It is straightforward to combine likelihood from independent sources:

$$L(\theta|\mathbf{x}_1, \mathbf{x}_2) = L_1(\theta|\mathbf{x}_1)L_2(\theta|\mathbf{x}_2),$$

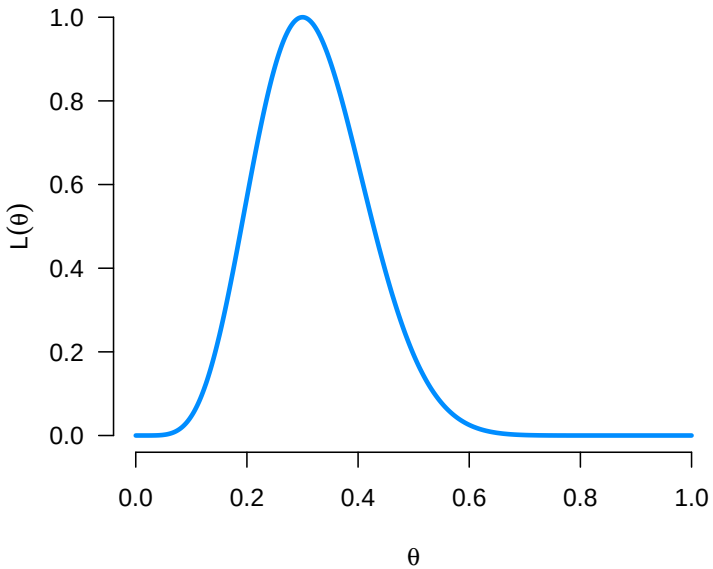
  where L_1 and L_2 can represent completely different models and $\mathbf{x}_1, \mathbf{x}_2$ completely different types of data
- This is even more conveniently represented on the log scale:

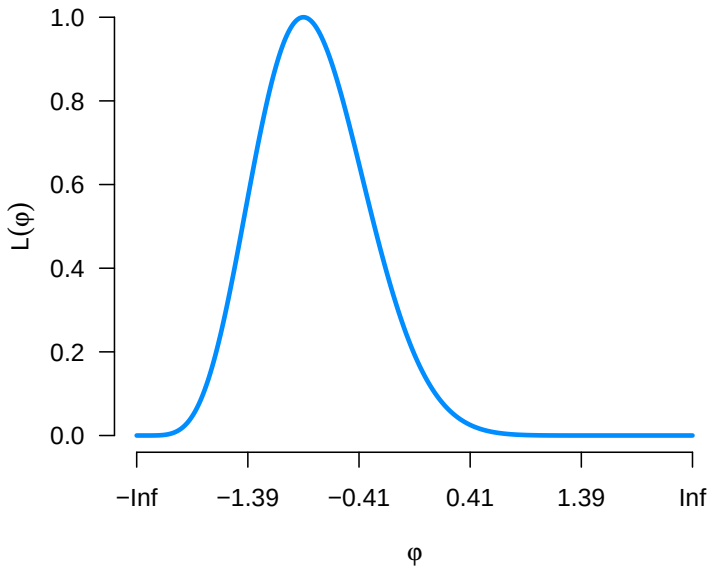
$$\ell(\theta|\mathbf{x}_1, \mathbf{x}_2) = \ell_1(\theta|\mathbf{x}_1) + \ell_2(\theta|\mathbf{x}_2),$$

  where $\ell(\theta) = \log L(\theta)$
- This is an extremely useful property

Invariance

- We now turn our attention to some appealing mathematical properties of likelihood
- One interesting result is that likelihood is *invariant* to transformations of the parameter
- For example, suppose we decide to parameterize our earlier binomial example in terms of the log-odds, $\phi = \log\{\theta/(1-\theta)\}$:





Practical implications

In particular, for any 1:1 transformation $\phi = g(\theta)$,

- $\hat{\phi} = g(\hat{\theta})$ ("invariance property of the MLE")
- If $[\theta_1, \theta_2]$ is a likelihood interval for θ, then $[g(\theta_1), g(\theta_2)]$ is a likelihood interval for ϕ, where *likelihood interval* means $[\theta_1, \theta_2] = \{\theta : L(\theta)/L(\hat{\theta}) > c\}$ for some c
- The argument can be extended to transformations that are not 1:1, although such transformations are of questionable relevance

Remarks

A few remarks on invariance:

- There are two kinds of invariance:
  - Measurement invariance, which deals with transformations of an observable quantity (e.g., measuring height in inches vs centimeters)
  - Parameter invariance, which is the kind we have been discussing;

  the first kind of invariance is universally agreed upon as reasonable
- The second kind, however, is more debatable – for example, Bayesian inference does not satisfy it
- The reason is that if θ is a random variable, then a transformation will introduce a Jacobian term, which must also be taken into account

Sufficiency

- Another important property possessed by likelihood is *sufficiency*
- **Definition:** A statistic $T(X)$ is *sufficient* for θ if the conditional distribution $X|T(X)$ does not depend on θ
- For example, suppose $X_1, \ldots, X_n \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(\theta, 1)$:
  - $\bar{x}$ is sufficient for θ
  - This means that if I were to, say, take a set of data and simulated another set with the same mean, then both sets of data would be equally informative about θ – nothing can possibly be learned about θ from the real data that I couldn't learn from the simulated version
- Note that sufficiency is entirely dependent on the assumed model

Factorization theorem

- Proving sufficiency based on the definition is often tedious; it is usually much easier to show using the following theorem
- **Theorem (Factorization):** The statistic $T(X)$ is sufficient for θ if and only if the model $p(x|\theta)$ can be factorized as follows:

$$p(x|\theta) = g(t(x), \theta)h(x)$$

- Here, $p(x|\theta)$ can be continuous, discrete, or mixed; we will discuss these situations more in a future lecture
- **Corollary:** The likelihood based on a sufficient statistic is equivalent to the likelihood based on the entire data.

Minimal sufficiency

- Sufficiency seems like a nice property to have, although there are lots of sufficient statistics – in particular, the full set of observations $\{x_1, \ldots, x_n\}$ is always a sufficient statistic
- There is a (partial) ordering here, though: from $\{x_1, \ldots, x_n\}$, we could calculate $\bar{x}$, but not the other way around; this leads to the following refinement of sufficiency
- **Definition:** A sufficient statistic $T(X)$ is *minimal sufficient* if it is a function of any other sufficient statistic.
- In our $\mathrm{N}(\theta, 1)$ example, $\bar{x}$ is minimal sufficient, as is $\sum_i x_i$, but $\{x_1, \ldots, x_n\}$ is not
- The general idea is that a minimal sufficient statistic offers the maximum amount of data reduction – any additional reduction would introduce loss of information

Likelihood and minimal sufficiency

- A rather intriguing result is that the likelihood function itself is minimal sufficient
- This perhaps requires some explanation: for any given θ, the expression $L(\theta)$ is simply a real number – this is not the sufficient statistic we're talking about
- Instead, the statistic we're talking about is $L(\cdot)$, the function over all possible values of θ; note that we do not have to know the true value of θ in order to construct this curve (i.e., it is a statistic)
- **Theorem:** The function $L(\cdot|\mathbf{x})$ is minimal sufficient.

Likelihood = perfect?

- As I said, this result is intriguing – if likelihood is minimal sufficient, then it always summarizes everything we need to know about the data . . . nothing else matters and there is no reason to use anything other than likelihood for inference
- However, this begins to head into territory that is not agreed upon by all statisticians
- To clarify the controversies involved, let us first distinguish between two concepts: the strong likelihood principle and the weak likelihood principle

The weak likelihood principle

- **Weak likelihood principle:** Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are two sets of observed data coming from the same experiment. If $L(\cdot|\mathbf{x}_1)$ is equivalent to $L(\cdot|\mathbf{x}_2)$, then any conclusions drawn from observing $\mathbf{x}_1$ and observing $\mathbf{x}_2$ should be identical.
- In reality, statisticians routinely use data to check model assumptions, so it is possible that observing, say, $\mathbf{x}_2$ could lead us to use a different model, which would violate the above principle
- However, if we assume the model is known, then the above principle is entirely reasonable; it is hard to mount any meaningful objection to it

The strong likelihood principle

- However, what about likelihoods coming from *different* experiments?
- **Strong likelihood principle:** Suppose $\mathbf{x}_1$ and $\mathbf{x}_2$ are two sets of observed data from different experiments involving the same unknown parameter. If $L(\cdot|\mathbf{x}_1)$ is equivalent to $L(\cdot|\mathbf{x}_2)$, then any conclusions drawn from observing $\mathbf{x}_1$ in experiment 1 should be identical to conclusions drawn from observing $\mathbf{x}_2$ in experiment 2.
- This is quite a bit stronger, as it is saying: not only do I not need the rest of the data, I also don't need to know anything about the experimental design – just give me the likelihood, everything else is irrelevant to inference

Binomial vs. Negative binomial

- Unlike the weak likelihood principle, which is relatively uncontroversial, the strong likelihood principle is directly contradicted by the standard practice of frequentist statistics
- For example, consider Experiment #1: a trial is repeated $n = 20$ times and $x = 6$ successes are observed:

$$L(\theta|x) = \binom{20}{6}\theta^6(1-\theta)^{14}$$

- Now consider Experiment #2: a trial is repeated until $x = 6$ successes are observed and this requires $n = 20$ trials:

$$L(\theta|x) = \binom{19}{5}\theta^6(1-\theta)^{14}$$

Hypothesis testing

- Clearly, the likelihoods are equivalent, so the strong likelihood principle tells us to draw the same conclusions from both experiments
- However, calculating frequentist p-values involves the sampling design as well, and we get different results for the two experiments (p-values are two-sided for $H_0 : \theta = 0.5$):
  - Experiment #1: $p = 0.12$
  - Experiment #2: $p = 0.03$
- Thus, we might conclude that $\theta \neq 0.5$ after observing experiment 2, but not from experiment 1, even though the likelihoods are the same

Is this absurd?

- This might seem reasonable at first, but it can take us into potentially absurd territory
- Our conclusions must now depend on the internal thoughts and intentions of the researcher?
- What if these aren't known?
- It is worth noting that Bayesian inference, by contrast, obeys the strong likelihood principle and draws the same conclusion from these two experiments (assuming the priors are the same, of course)

Sequential testing

- On the other hand, consider the idea of sequential testing: based on the existing data $x_1, \ldots, x_n$, we decide to either stop where we are or continue to collect more data
- Our motivation for collecting data does not enter the likelihood, only the data we collect, so the strong likelihood principle would tell us to ignore the sequential aspect of the testing and treat the data as if the number of observations was prespecified

Problematic?

- However, this would seem to open the door for abuse, as it allows us to stop collecting data at a point where the data looks good
- In particular, if we are calculating p-values, we could decide to stop collecting data as soon as $p < 0.05$, which will happen eventually with probability 1 (we will prove this fact later in the course)
- This would of course be a big problem for frequentist inference (our type I error rate would be 100%); seems a bit fishy for a Bayesian to ignore this as well, although it is difficult to make an objective case against it

Other challenges to the likelihood principle

- One final concern that is often raised against the likelihood principle is how to handle multi-model or multi-parameter uncertainty
- For example, perhaps we have a large number of potential covariates to include in a model, and only include some of them in the final model; the likelihood itself does not reflect this
- The Bayesian approach can address this issue by introducing priors over the potential space of models; we will discuss some more direct modifications and extensions to the likelihood later in the course

Summary

- Likelihood has many attractive properties and is a natural quantity to focus inference upon
- Likelihood is important to both Bayesian and Frequentist statistical paradigms
- Likelihood is not without challenges, however – there are controversies over its use as well as situations where likelihood alone may be misleading