Introduction

Patrick Breheny

January 10

Introductory example: Jane's twins

- Suppose you have a friend named Jane who is pregnant with twins
- At the ultrasound, she learns that her twins are the same sex
- It is unclear, however, whether the twins are identical or not, so she asks you to determine this probability
- Relevant background information: 1/3 of all twins are identical; all identical twins are the same sex; half of fraternal twins are the same sex

Bayes' rule

- The answer to this question can be calculated using Bayes' rule:

$$p(I|S) = \frac{p(I)p(S|I)}{p(S)},$$

where I is the event "having identical twins" and S is the event "twins are same sex"

- Thus,

$$\begin{aligned} p(I|S) &= \frac{\frac{1}{3} \cdot 1}{\frac{1}{3} \cdot 1 + \frac{2}{3} \cdot \frac{1}{2}} \\ &= \frac{1}{2} \end{aligned}$$

The frequentist interpretation

- There is, however, a 250-year-old debate in statistics over the meaning of this probability, and the two sides of this debate lead to fundamentally different ways of carrying out statistical inference
- The first – and no doubt more familiar – interpretation is called the *frequentist* interpretation, and says that the probability refers to the frequency with which the event happens in repeated, identical trials
- Thus, the frequentist interpretation of our earlier calculation is that, if a large number of women with same-sex twins gave birth, one-half of those twins would be identical

Frequentist interpretation of Jane's twins

- What about Jane's twins?
- Notably, the frequentist interpretation says *nothing* about the probability that *Jane's* twins will be identical
- Jane is just one person with one set of twins – there is no long-run series of repeated trials involved
- To put it another way, whether Jane's twins are identical or not is not a random variable – they either are, or are not, identical, and this is not randomly fluctuating inside Jane's womb

Bayesian interpretation

- A quite different interpretation (the *Bayesian* interpretation) is that the probability of one-half is a numerical representation of our *belief* (or *uncertainty*, depending on how you look at it) concerning Jane's twins
- The idea of "belief" sounds qualitative, but there are several ways of making it rigorous
- For example, suppose I offered you a bet: if Jane's twins are identical, you win \$6; if they are fraternal, you lose \$4
- If you believe that the probability that Jane's twins are identical is 1/2, then you would take the bet, since you expect to gain \$1
- However, if you don't know the results of the ultrasound, you wouldn't take the bet, since you expect to lose \$0.67

Contrasting the two interpretations

- Note that the Bayesian interpretation doesn't contradict the frequentist interpretation – either way, when faced with a large number of pregnant women with same-sex ultrasounds, we predict that half of them would give birth to identical twins
- However, the Bayesian interpretation goes farther, in asserting specific things about Jane, rather than merely a collection of women in a similar situation

The debate

- As alluded to earlier, intelligent people have argued for and against both interpretations for over two centuries
- To put it briefly, the basic arguments are:
  - *Frequentist:* The Bayesian interpretation is merely a subjective description of what is going on in someone's head; the frequentist interpretation actually corresponds to a concrete external reality
  - *Bayesian:* The frequentist interpretation is hollow and doesn't address the real question; if I am going to conclude something or make a decision in the presence of uncertainty, I need a coherent, rigorous way of quantifying my beliefs and attitudes about risks and benefits

Confidence intervals

- Anyone who has ever taken or taught an introductory statistics class on confidence intervals will instantly recognize the Bayesian criticism of Frequentist probabilities
- Suppose we calculate a frequentist 95% confidence interval for the amount by which Drug X will, on average, reduce a patient's blood pressure to be (5, 10)
- Does that mean that there is a 95% chance that the true average is between 5 and 10?
- "No", you explain to your students, "the 95% refers to the long-run proportion of confidence intervals calculated in a similar manner as this one that will contain the true parameter; we can't actually say anything about whether the true reduction is between 5 and 10 for this particular drug" (students are generally not thrilled with this explanation)

Additional remarks

- In addition to these fundamental philosophical differences, there are a number of practical advantages and disadvantages to frequentist and Bayesian approaches to analyzing data that we will see throughout the course
- For now, we will take the Bayesian interpretation as a given and see what it implies about analyzing data and learning from experiments, although we will occasionally return to the question of interpretation when we encounter differences in Bayesian and frequentist results

The central role of Bayes rule

- In the words of Brad Efron, "Using Bayes rule doesn't make one a Bayesian. *Always* using Bayes rule does"
- This is an accurate summary of the approaches to statistical inference resulting from the two interpretations of probability
- Although there are techniques and principles that tend to be used repeatedly, there is no end to the different tools that one can use in frequentist statistics – any method that satisfies long-run frequency considerations is allowable
- With Bayesian analysis, however, there is no ambiguity: Bayes rule is used to carry out all inference

Paradigm for Bayesian inference

Letting θ denote an unknown parameter of interest and y observed data, the basic approach to Bayesian inference can be represented as follows:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)},$$

where

- $p(\theta)$ is the *prior*: Our beliefs about the plausible values of our parameter before seeing any data
- $p(y|\theta) = L(\theta; y)$ is the *likelihood*: The sampling distribution for how the data depends on the unknown parameters
- $p(\theta|y)$ is the *posterior*: Our updated beliefs about the plausible values for our parameter after seeing the data
- $p(y)$ is a normalizing constant which is often not of interest

Notation

The following notation is used fairly consistently in the Bayesian literature, including our two textbooks:

- $p(\cdot)$ is a probability distribution function (or mass function; $p(\cdot)$ is used for both continuous and discrete distributions)
- Upper-case Roman letters (Y) are used to denote observable random variables
- Lower-case Roman letters (y) are used to denote realizations of random variables; having been observed, they are now fixed
- Greek letters (θ) are used for unobservable quantities

Posterior distributions

- Bayesian inference rests entirely on the posterior distribution $p(\theta|y)$, an explicit quantitative representation of everything we could possibly want to know about θ in light of all the available information
- For example, we might want to know the central tendency of θ: $\mathrm{E}(\theta|y) = \int \theta p(\theta|y) d\theta$
- We might want a numerical measure of the uncertainty about θ: this can be summarized using $\mathrm{Var}(\theta|y) = \int \{\theta - \mathrm{E}(\theta)\}^2 p(\theta|y) d\theta$
- We might want a range of likely values of values for θ: we can calculate the probability of θ falling into any range (θ_1, θ_2) by $\int_{\theta_1}^{\theta_2} p(\theta|y) d\theta$

Integration

- It seems clear that Bayesian inference is going to involve a lot of integration to obtain the quantities we are interested in
- Occasionally, it is possible to work out these integrals in closed form
- In most situations, however, this is not possible

Monte Carlo integration

- Historically, this was a major impediment to the use of Bayesian methods
- However, Bayesian inference has seen a dramatic resurgence in the computer era, as numerical methods for integration have enabled researchers to calculate otherwise intractable integrals
- With *Monte Carlo integration*, instead of actually evaluating an integral, we approximate it numerically by drawing random samples $\{\theta^{(1)}, \ldots, \theta^{(T)}\}$ from $p(\theta|y)$
- By the law of large numbers, this approximation will converge to the value of the integral as the number of random samples that we draw goes to infinity:

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta^{(t)}) \xrightarrow{\text{P}} \text{E}g(\theta) = \int g(\theta)p(\theta|y)d\theta$$

Markov chain Monte Carlo

- The most powerful and widely used class of algorithms for carrying out Monte Carlo integration is known as *Markov chain Monte Carlo*, or *MCMC*, which we discuss in greater detail later on
- MCMC methods are central to the practical application of the Bayesian methods, and we will use MCMC methods throughout this course
- To provide a brief glimpse into the practice of Bayesian statistics and the use of MCMC methods, we are going to conduct a brief preview, taking a superficial look at a basic problem (two samples, normally distributed outcomes)

Example: lead exposure and IQ

- Our example for today deals with a study concerning lead exposure and neurological development for a group of children in El Paso, Texas
- The study compared the IQ levels of 57 children who lived within 1 mile of a lead smelter and a control group of 67 children who lived at least 1 mile away from the smelter
- In the study, the average IQ of the children who lived near the smelter was 89.2, while the average IQ of the control group was 92.7

Components of a complete model

- A Bayesian model has two necessary components:

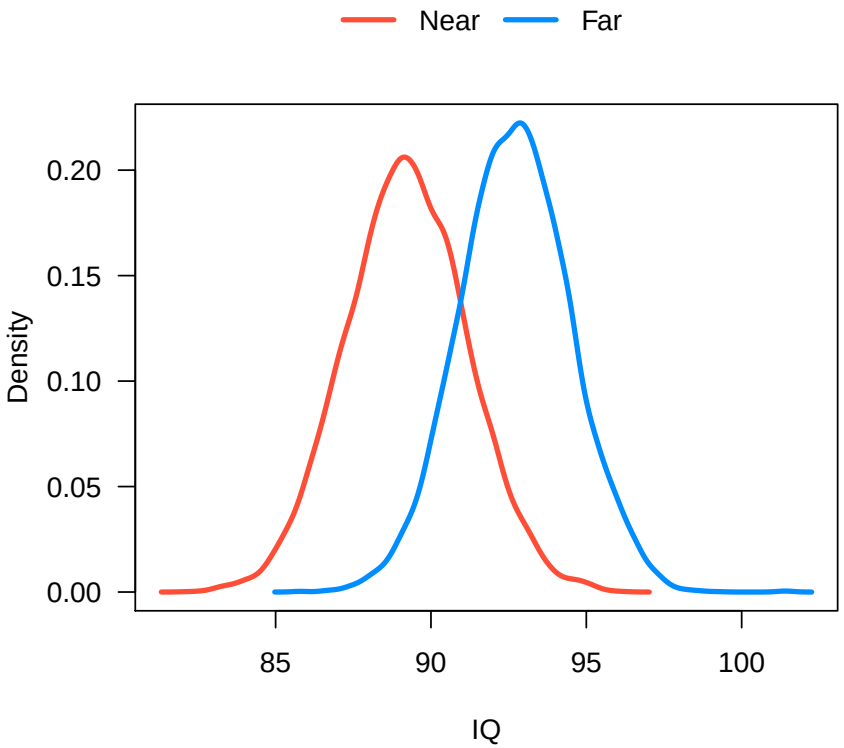
$$\text{Likelihood: } p(y|\theta)$$
$$\text{Prior: } p(\theta)$$

- Remark: In general, we may be interested in a quantity ω that depends on θ, in which case we require an additional component $p(\omega|\theta, y)$

MCMC software

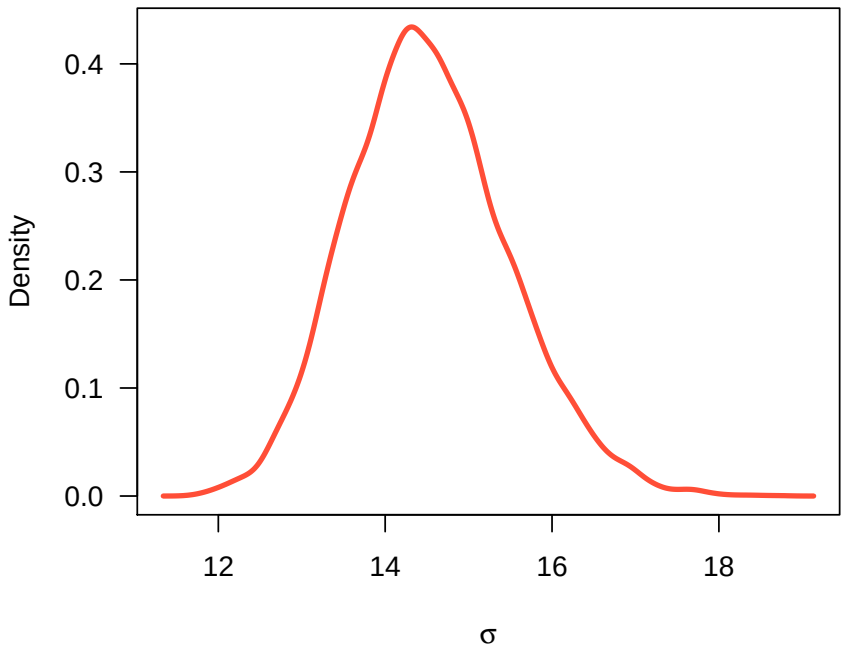
- Various programs are available for automatically carrying out the MCMC process, and have similar syntax
- We will focus on two of these in this class: BUGS (Bayesian inference Using Gibbs Sampling) and JAGS (Just Another Gibbs Sampler)
- To use either JAGS or BUGS, you must provide:
  - A complete model (which, recall, includes both likelihood and prior)
  - The data (y)
- My preferred way to provide these is via `R` using the `R2OpenBUGS` and `R2jags` (see code), although there are many other options

Posterior distribution: μ



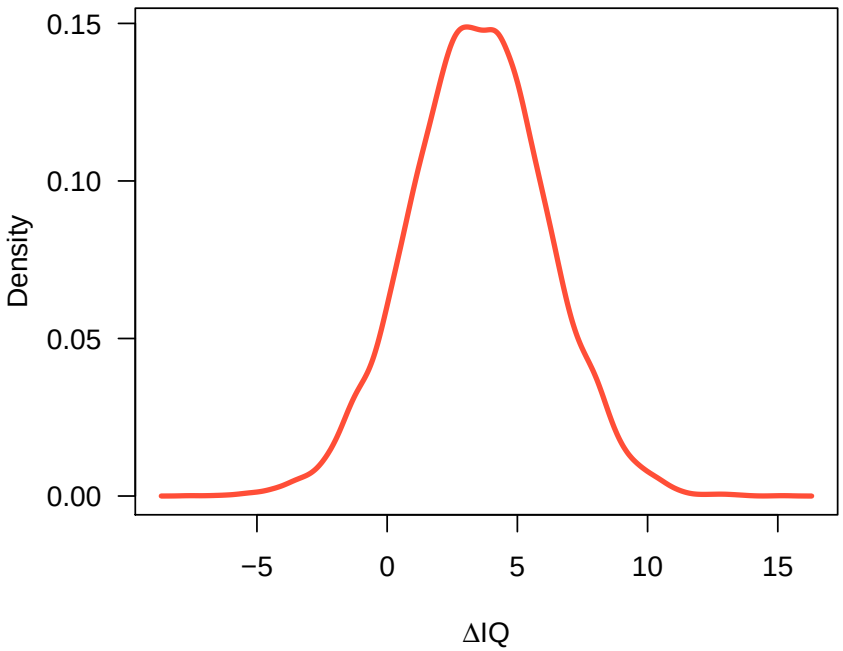
	Mean	SD	2.5%	97.5%
Near	89.2	2.0	85.4	93.1
Far	92.7	1.8	89.2	96.2

Posterior distribution: σ



	Mean	SD	2.5%	97.5%
σ	14.5	1.0	12.8	16.5

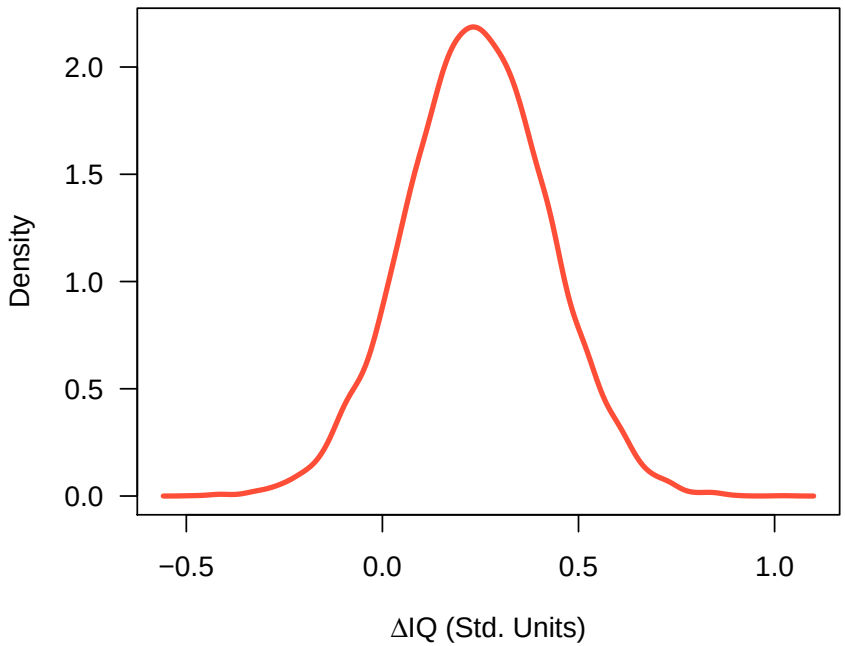
Posterior distribution: Difference



	Mean	SD	2.5%	97.5%
Δ	3.5	2.6	-1.7	8.6

$$\Pr(\Delta < 0) = .09$$

Posterior distribution: Signal-to-noise ratio



	Mean	SD	2.5%	97.5%
SNR	0.24	0.18	-0.11	0.60

$$\Pr(\text{SNR} < 0) = .09$$

Final remarks

In summary:

- Bayesian inference requires a complete model in the form a likelihood $p(y|\theta)$ and prior $p(\theta)$
- All inference centers around the posterior distribution $p(\theta|y)$, which is specified by the model via Bayes rule
- This distribution is relatively straightforward to describe and summarize, provided that we can sample from it using MCMC methods