Binomial data

Patrick Breheny

January 15

Introduction

- As our first substantive example of Bayesian inference, we will analyze binomial data
- This type of data is particularly amenable to Bayesian analysis, as it can be analyzed without MCMC sampling, and thus has played an important historical role in the field
- Our motivating example for today is a study which took place at Johns Hopkins to estimate the survival chances of infants born prematurely by surveying the records of babies born at their hospital in a three-year period
- In their study, they found 39 babies who were born at 25 weeks gestation, 31 of which survived at least 6 months

Uniform prior

- It seems reasonable to assume that the number of babies who survive (Y) follows a binomial distribution:

$$p(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

- Suppose we let $\theta \sim \text{Unif}(0,1)$ (we will have a much more thorough discussion of priors later); then

$$p(\theta|y) \propto \theta^y(1-\theta)^{n-y}$$

- We can recognize this as a beta distribution; in particular,

$$\theta|y \sim \text{Beta}(y+1, n-y+1)$$

Conjugacy and the beta-binomial model

- Suppose more generally that we had allowed θ to follow a more general beta distribution:

  $$\theta \sim \mathrm{Beta}(\alpha, \beta)$$

  (note that the uniform distribution is a special case of the beta distribution, with $\alpha = \beta = 1$)
- In this case, θ still follows a beta distributon:

  $$\theta | y \sim \mathrm{Beta}(y + \alpha, n - y + \beta)$$

- This phenomenon is referred to as *conjugacy*: the posterior distribution has the same parametric form as the prior distribution (in this case, the beta distribution is said to be the *conjugate prior* for the binomial likelihood)

Summarizing the posterior

The fact that $\theta|y$ follows a well-known distribution allows us to obtain closed-form expressions for quantities of interest, for example:

- Posterior mean:

$$\bar{\theta} = \mathrm{E}(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n}$$

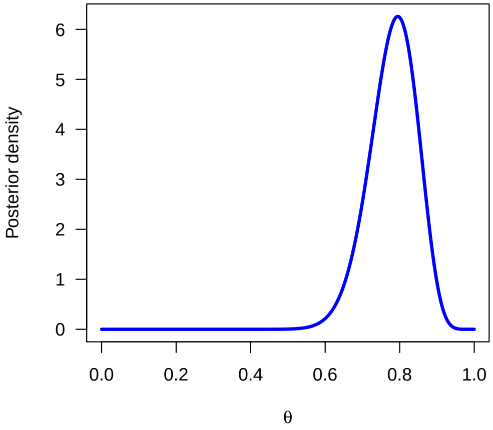
- Posterior mode:

$$\hat{\theta} = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}$$

- Posterior variance:

$$\mathrm{Var}(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

Premature birth example



$$\bar{\theta} = 0.780$$

$$\hat{\theta} = 0.795$$

$$\mathrm{SD}_{\mathrm{post}} = 0.064$$

Posterior intervals

- Although the quantiles of a beta distribution do not have a closed form, they are easily calculated by any statistical software program (*e.g.*, the `qbeta` function in `R`)
- Quantiles allow for the easy construction of intervals that have a specified probability of containing the outcome
- For example, we may construct a 90% *posterior interval* by finding the 5th and 95th percentiles of the posterior distribution
- For the premature birth example, this interval is $(0.668, 0.877)$

Interpretation of posterior intervals

- It is worth comparing the interpretation of this posterior interval (also referred to as a *credible interval*) with the frequentist interpretation of a confidence interval
- It is absolutely correct in Bayesian inference to say that "there is a 90% chance that the true probability of survival is between 66.8% and 87.7%"
- It is incorrect, however, to make a similar statement in frequentist statistics, where the properties of a confidence interval must be described in terms of the long-run frequency of coverage for confidence intervals constructed by the same method

HPD intervals

- Note that the quantile-based interval we calculated is not unique: there are many intervals that contain 90% of the posterior probability
- An alternative approach to constructing the interval is to find the region satisfying: (i) the region contains $(1-\alpha)\%$ of the posterior probability and (ii) every point in the region has higher posterior density than every point outside the region (note that this interval is unique)
- This interval is known as the *highest posterior density*, or *HPD* inveral; in contrast, the previous interval is known as the *central* interval, as it contains by construction the middle $(1-\alpha)\%$ of the posterior distribution

HPD vs. central intervals

- For the premature birth example, the two approaches give:

$$\begin{aligned} \text{Central} &: (66.8, 87.7) \\ \text{HPD} &: (67.8, 88.5) \end{aligned}$$

- The primary argument for the HPD interval is that it is guaranteed to be the shortest possible $(1-\alpha)\%$ posterior interval
- The primary argument for the central interval is that it is invariant to monotone transformations (it is also easier to calculate)

Similarities to Frequentist approaches

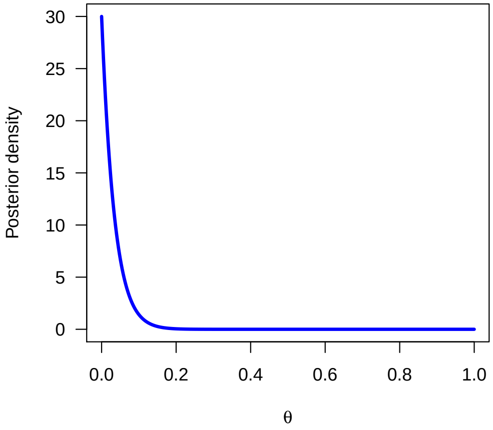
In many ways, the conclusions we arrive at with the Bayesian analysis are similar to those we would have obtained from a frequentist approach:

- The MLE, $\hat{\theta} = 31/39 = 0.795$, is exactly the same as the posterior mode and very close to the posterior mean (0.780)
- The standard error, $\sqrt{\hat{\theta}(1-\hat{\theta})/n} = 0.065$, is very close to the posterior standard deviation (0.064)
- The (Wald) confidence interval is (0.69, 0.90), close to the HPD interval (0.68, 0.89)

Proportions near 0 or 1

- However, this is not always the case
- For example, the Johns Hopkins researchers also found 29 infants born at 22 weeks gestation, none of which survived
- This sort of situation (and more generally, when the MLE occurs at the boundary of the parameter space) causes problems for frequentist approaches, but presents no problem for the Bayesian analysis

Premature birth example, 22 weeks



$$\bar{\theta} = 0.032$$

$$\hat{\theta} = 0$$

$$\text{SD}_{\text{post}} = 0.031$$

$$\text{HDI}_{95} : (0, 0.095)$$

Problems with frequentist approaches

- In contrast, the (Wald) frequentist approach produces nonsensical estimates for the standard error (0) and a 95% confidence interval (0, 0)
- To be fair, alternative frequentist approaches for constructing CIs exist for this problem, such as inverting the score test (0, 0.117) and inverting the exact binomial test (0, 0.119)
- However, these approaches do not always achieve correct coverage – in this particular case, they both produce a more conservative interval than the Bayesian approach

Informative vs. non-informative priors

- Thus far, we've focused on the prior $\theta \sim \mathrm{Unif}(0, 1)$ as a way of expressing a belief, before seeing any data, that all proportions are equally likely, or in other words a lack of any particular belief regarding θ
- A basic division may be made between "non-informative" priors such as these and "informative" priors explicitly intended to incorporate external information
- These priors, and the resulting posteriors, serve different purposes: the former is in some sense an attempt to convince everyone, while the latter may be intended only to allow one individual reach a conclusion

Informative priors

- Informative priors are well suited to research groups trying to use all of the information at their disposal to make the quickest possible progress
- For example, in trying to plan future studies and experiments about possible drugs to follow up on, a drug company may wish to take as much information (from animal studies, pilot studies, studies on related drugs, studies conducted by other groups, etc.) as possible into account when constructing a prior

Reference priors

- On the other hand, the drug company could not expect to convince the at-large research community (or the FDA) with such a prior
- Thus, even if the drug company did not actually believe in a uniform prior, they might still wish to conduct an analysis using this prior for the sake of arriving at more universally acceptable conclusions
- To emphasize this point, non-informative priors are often called *reference* priors, as their intent is to provide a universal reference point regardless of actual prior belief (note: the term "reference prior" is given a much more specific meaning in Bernardo (1979) and in later papers by Berger and Bernardo)
- Other names include "default" priors, "vague" priors, "diffuse" priors, and "objective" priors

Problems with labels

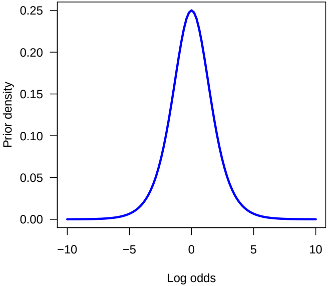
- The term "non-informative" is potentially somewhat misleading, as all priors contain some information in some sense of the word
- The term "objective" is potentially misleading when applied to any analysis (frequentist or Bayesian), as fundamentally subjective decisions must be made in terms of selecting the model, distributional assumptions, etc.
- A clear-cut distinction between informative and reference priors is potentially misleading, as it is often reasonable to use reference priors for parameters of interest, and informative priors for nuisance parameters

Information is not invariant

- As an example of the problem in saying a prior has "no information", consider transforming the variable of interest:

$$\psi = \log \frac{\theta}{1-\theta}$$

- Our uniform prior on θ, which stated that all probabilities were equally likely, is stating that log-odds values near 0 are more likely than others:



Jeffreys priors

- The statistician Harold Jeffreys developed a proposal for prior distributions which would make them invariant to such transformations, at least with respect to the Fisher information:

$$I(\theta) = \mathrm{E}\left\{\left(\frac{d}{d\theta}\log p(Y|\theta)\right)^2\right\}$$

- Thus, by the chain rule, the information for $\psi = f(\theta)$ satisfies

$$I(\theta) = I(\psi)\left|\frac{d\psi}{d\theta}\right|^2$$

Jeffreys priors (cont'd)

- Now, if we select priors according to the rule $p(\theta) \propto I(\theta)^{1/2}$, we have invariance with respect to the information:

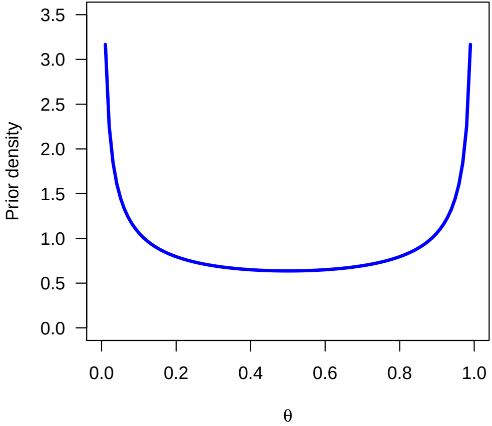
$$\begin{aligned} p(\psi) &= p(\theta) \left| \frac{d\theta}{d\psi} \right| \\ &= I(\theta)^{1/2} \left| \frac{d\theta}{d\psi} \right| \\ &= I(\psi)^{1/2}, \end{aligned}$$

  which is the same prior we would have if we parameterized the model in terms of ψ directly

- Various other formal rules for specifying automatic prior distributions have been proposed, although the Jeffreys (1946) approach is the most well-known

Jeffreys prior for premature birth example

For the binomial likelihood, the Jeffreys prior is $\theta \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$:



$$\bar{\theta} = 0.788$$

$$\hat{\theta} = 0.803$$

$$\mathrm{SD}_{\mathrm{post}} = 0.064$$

$$\mathrm{HDI}_{90} : (0.686, 0.892)$$

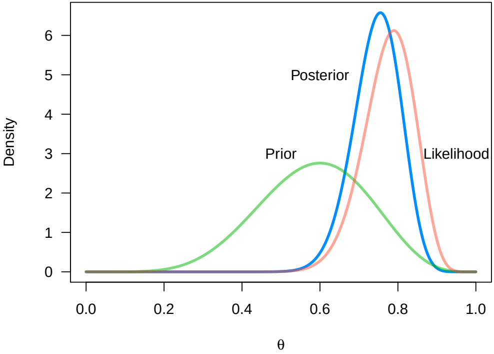
Prior sensitivity

- Note that the posterior didn't change much between the uniform and Jeffreys priors
- This is good; it would be an unattractive feature of Bayesian inference if two reasonable-sounding priors led to drastically different conclusions
- This is not always the case – certain models or data can lead to unstable inference in which the prior has a large influence on the posterior
- This phenomenon is called *sensitivity* to the prior, and good Bayesian analyses often consist of several priors to illustrate how robust the conclusions are to the specification of the prior

Informative prior for premature birth data

- Of course, an informative prior can exert greater influence over the posterior
- Let's analyze our premature birth data one more time: this time, let's suppose that there had been some previous studies that had suggested that the probability of survival was around 60%, and that it was rather unlikely to be close to 0% or 100%
- We might propose, in this situation, a $\theta \sim \text{Beta}(7, 5)$ prior
- Note that conjugacy is often helpful when thinking about priors: this is the same as the posterior we would obtain with a uniform prior after seeing 6 successes and 4 failures, thereby carrying an "effective prior sample size" of 12 (or 10 more than the reference prior)

Premature birth example, informative prior



$$\bar{\theta} = 0.745$$
$$\hat{\theta} = 0.755$$
$$\mathrm{SD}_{\mathrm{post}} = 0.060$$
$$\mathrm{HDI}_{90} : (0.647, 0.845)$$

Sequential updating

- Note that the posterior we obtain, $\theta|y \sim \mathrm{Beta}(38, 13)$, is the same as what we would obtain if we had started from a uniform prior, stopped and conducted an analysis after observing 6 successes and 4 failures, then used the posterior from that analysis as the prior for analyzing the rest of the data
- Indeed, we could have stopped and analyzed the data after each observation, with each posterior forming the prior for the next analysis, and it would not affect our conclusions; this is known as *sequential updating*

Posterior as compromise

- Note also that the posterior is in some sense a compromise between the prior and the likelihood
- This is true in a more formal sense as well:

  $$\frac{\alpha+y}{\alpha+\beta+n} = w\frac{\alpha}{\alpha+\beta} + (1-w)\frac{y}{n};$$

  in other words, the posterior mean is a weighted average of the prior mean and the sample mean
- This makes intuitive sense, as the posterior combines information from both sources

Posterior variance inequality

- Given that we are adding information from the data to the prior, one might expect that we have less uncertainty in the posterior than we had in the prior; this is also true, at least on average
- **Theorem:** $\mathrm{Var}(\theta) \geq \mathrm{E}(\mathrm{Var}(\theta|Y))$
- Note that this is true generally for Bayesian inference; nothing specific to the binomial distribution was used in the proof

Concentration of posterior mass

- Indeed, letting θ_0 denote the unobservable true value of θ, note that

$$\bar{\theta} \xrightarrow{\mathsf{P}} \theta_0$$

$$\mathrm{Var}(\theta|y) \xrightarrow{\mathsf{P}} 0$$

- Thus, $\theta|y$ converges in distribution to one which places a point mass of 1 on θ_0
- In other words, given enough data, the likelihood will overwhelm the prior and there will no longer be any uncertainty about θ (this is also usually true in Bayesian statistics)