(4) One-parameter models - Beta/binomial

ST440/550: Applied Bayesian Statistics
Estimating a proportion using the beta/binomial model

> A fundamental task in statistics is to estimate a proportion using a series of trials:
>   - What is the success probability of a new cancer treatment?
>   - What proportion of voters support my candidate?
>   - What proportion of the population has a rare gene?
>   - What proportion of trials can students correctly guess if the number if even or odd?

> Let $\theta \in [0, 1]$ be the proportion we are trying to estimate (e.g., the success probability).

> We conduct $n$ independent trial, each with success probability $\theta$, and observe $Y \in \{0, \ldots, n\}$ successes.

> We would like obtain the posterior of $\theta$, a 95% interval, and a test that $\theta$ equals some predetermined value $\theta_0$. 
Frequentist analysis

- The maximum likelihood estimate is the sample proportion
  \[ \hat{\theta} = \frac{Y}{n} \]

- For large \( Y \) and \( n - Y \), the sampling distribution of \( \hat{\theta} \) is approximately
  \[ \hat{\theta} \sim \text{Normal} \left( \theta, \frac{\theta(1 - \theta)}{n} \right) \]

- The standard error (standard deviation of the sampling distribution) is approximated as
  \[ \text{SE}(\hat{\theta}) \approx \sqrt{\frac{\hat{\theta}(1 - \hat{\theta})}{n}} \]

- A 95% CI is then
  \[ \hat{\theta} \pm 2\text{SE}(\hat{\theta}) \]
Bayesian analysis - Likelihood

- Since $Y$ is the number of successes in $n$ independent trials, each with success probability $\theta$, its distribution is

$$Y|\theta \sim \text{Binomial}(n, \theta)$$

- **PMF:** $P(Y = y|\theta) = \binom{n}{y} \theta^y (1 - \theta)^{n-y}$

- **Mean:** $E(Y|\theta) = n\theta$

- **Variance:** $V(Y|\theta) = n\theta(1 - \theta)$

- **Plots:** [Link](http://shiny.stat.ncsu.edu/bjreich/Binomial_PMF/)
Bayesian analysis - Prior

- The parameter $\theta$ is continuous and between 0 and 1, therefore a natural prior is

$$\theta \sim \text{Beta}(a, b)$$

- PDF: $f(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1 - \theta)^{b-1}$

- Mean: $E(\theta) = \frac{a}{a+b}$

- Variance: $V(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$

- Plots: [http://shiny.stat.ncsu.edu/bjreich/Beta_PDF/](http://shiny.stat.ncsu.edu/bjreich/Beta_PDF/)
Derivation of the posterior

- The posterior is $\theta | Y \sim \text{Beta}(a + Y, b + n - Y)$

Shrinkage

- The posterior mean is
  \[ \hat{\theta}_B = E(\theta|Y) = \frac{y + a}{n + a + b} \]

- The posterior mean is between the sample proportion \( Y/n \) and the prior mean \( a/(a + b) \):

- When (in terms of \( n, a \) and \( b \)) is the \( \hat{\theta}_B \) close to \( Y/n \)?

- When is the \( \hat{\theta}_B \) shrunk towards the prior mean \( a/(a + b) \)?
Selecting the prior

- The posterior is $\theta|Y \sim \text{Beta}(a + Y, b + n - Y)$

- Therefore, $a$ and $b$ can be interpreted as the “prior number of success and failures”

- This is useful for specifying the prior

- What prior to select if we have no information about $\theta$ before collecting data?

- What prior to select if historical data/expert opinion indicates that $\theta$ is likely between 0.6 and 0.8?
Summarizing the posterior

- A plot of the posterior contains all the information about the proportion
- [http://www4.stat.ncsu.edu/~reich/ABA/code/BetaBinom](http://www4.stat.ncsu.edu/~reich/ABA/code/BetaBinom)
- If the posterior is fairly symmetric, the posterior mean and standard deviation are good summaries
- The 90% posterior credible set is also a good summary.

- Interpretation of a credible interval:
One-sided hypothesis test

▶ How to test $\mathcal{H}_0 : \theta \leq \theta_0$ versus $\mathcal{H}_A : \theta > \theta_0$?
▶ We compute the posterior probability of each hypothesis
▶ $\text{Prob}(\mathcal{H}_0 | Y) = \text{Prob}(\theta \leq \theta_0 | Y)$

▶ This is found using the $\text{pbeta}$ function in $R$
▶ Reject if the probability of the null is “small” (we will discuss the threshold more later)
▶ How is this different than the p-value?
Two-sided hypothesis test

- How to test $\mathcal{H}_0 : \theta = \theta_0$ versus $\mathcal{H}_A : \theta \neq \theta_0$?

- Compute the posterior probability of each hypothesis?

- $\text{Prob}(\mathcal{H}_0|Y) = \text{Prob}(\theta = \theta_0|Y) = 0$ since the beta is a continuous distribution

- One approach is to compute the Bayes factor, which we will discuss later

- More common and simpler approach is to compute the posterior 95% interval and look to see if it includes $\theta_0$
Monte Carlo sampling

▶ For harder problems we won’t be able to compute posterior means or intervals mathematically

▶ We will have to resort to simulation

▶ Assume we (by “we”, I mean JAGS in R) can generate $S$ samples from the posterior distribution, denoted $\theta_1, \ldots, \theta_S$

▶ In this problem, the \texttt{rbeta} function in R can generate samples

▶ Analogy: The posterior is the population, and the samples are used to learn about the population
Monte Carlo sampling

How to use the posterior samples $\theta_1, \ldots, \theta_S$ to approximate summaries of the posterior?

- Posterior mean
- Posterior sd
- Posterior 95% interval
- Posterior probability that $\theta > 0.5$

http://www4.stat.ncsu.edu/~reich/ABA/code/BetaBinom
Related problem

Say $Y$ is the number of successes before we observe $n$ failures

Then $Y|\theta \sim \text{NegativeBinomial}(n, \theta)$ and

$$\text{Prob}(Y = y|\theta) = \binom{y + n + 1}{y} \theta^y (1 - \theta)^n$$

Assume the prior $\theta \sim \text{Beta}(a, b)$ and find the posterior

How does this compare to the beta-binomial posterior?