2020 SISG Module 8: Bayesian Statistics for Genetics
Lecture 2: Review of Probability and Bayes Theorem

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Bayesian Learning

Bayesian Analysis of Binomial Data

Technical Appendices

Motivation

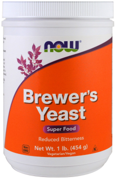
Introduction

- We will first briefly recap generic Bayesian learning.
- We will consider the Bayesian modeling of binomial data.
- Conjugate priors will be introduced.
- To motivate the binomial model we examine allele specific expression (ASE) experiment.
- Sampling from the posterior will be discussed as a method for flexible inference, including a non-conjugate example.

Motivating Example: Allele Specific Expression

- Gene expression variation is an important contribution to phenotypic variation within and between populations.
- Expression variation may be due to genetic or environmental sources.
- Genetic variation may be due to cis- (local) or trans (distant)-acting mechanisms.
- Polymorphisms that act in cis affect expression in an allele specific manner.
- RNA-Seq is a high throughput technology that allows allele-specific expression (ASE) to be measured.

- The data we consider is in yeast, and is a controlled experiment in which two strains, BY and RM, are hybridized.
- Consider a gene with one exon and five SNPs within that exon.
- Suppose the BY allele of the gene is expressed at a high level.
- In contrast, the RM allele has a mutation in a transcription factor binding site upstream of the gene that greatly reduces expression of this allele.
- Then, in the mRNA isolated from the yeast, when we look just at this gene, there are lots more BY mRNA molecules than RM mRNA molecules.



Example of ASE

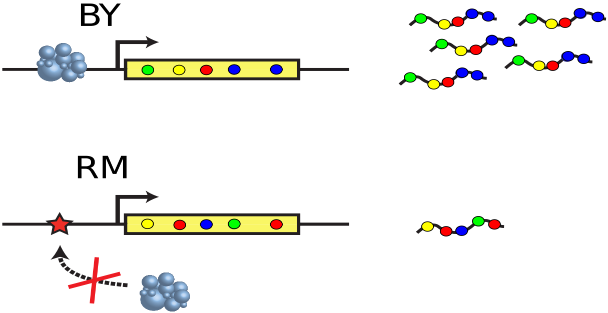


Figure 1: In the top figure the transcription factor (blue) leads to high transcription. In the bottom figure an upstream polymorphism (red star) prevents the transcription factor from binding.

Specifics of ASE Experiment

Details of the data:

- Two “individuals” from genetically divergent yeast strains, BY and RM, are mated to produce a diploid hybrid.
- Three replicate experiments: same individuals, but separate samples of cells.
- Two technologies: Illumina and ABI SOLiD.
- Each of a few trillion cells are processed.
- Pre- and post-processing steps are followed by fragmentation to give millions of 200–400 base pair long molecules, with short reads obtained by sequencing.
- Need SNPs since otherwise the reference sequence is identical and so we cannot tell which strain the read arises from.
- Strict criteria to call each read as a match are used, to reduce read-mapping bias.
- Data from 25,652 SNPs within 4,844 genes.
- More details in Skelly *et al.* (2011).

The Data

BY Count	Total Count	MLE $\widehat{\theta}$
62	107	0.58
33	59	0.56
658	1550	0.42
14	61	0.23
57	153	0.37
218	451	0.48
10	19	0.53
$\vdots$	$\vdots$	$\vdots$

Table 1: First few rows of ASE data.

How close to be 0.5 before we conclude no ASE?

Simple Approach to Testing for ASE

For a generic gene:

- Let N be the total number of counts at a particular gene, and Y the number of reads to the BY strain.
- Let θ be the probability of a map to BY.
- The first thing we need is a sampling model, aka the likelihood.
- A simple approach is to assume:

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

  and carry out a test of $H_0 : \theta = 0.5$, which corresponds to no allele specific expression.

Simple Approach to Testing for ASE

- A non-Bayesian approach might use an exact test, i.e. enumerate the probability, under the null, of all the outcomes that are equal to or more extreme than that observed.
- Issues:
  - *p*-values are not uniform under the null due to discreteness of Y.
  - How to pick a threshold? In general and when there are multiple tests.
  - Do we really want a point null, i.e. $\theta = 0.5$?
  - How would a Bayesian perform inference for this problem?

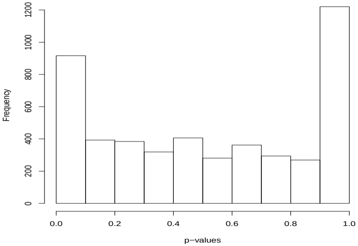


Figure 2: *p*-values from 4,844 exact tests.

Motivating Example: Smoothing/Penalization

- When looking at estimates over space or time, we want to know if the differences we see are “real”, or simply reflecting sampling variability.
- In data sparse situations, when one expects similarity smoothing local patterns (in time, space, or both) this can be highly beneficial.
- This general approach can equivalently be thought of penalization, in which large deviations from “neighbors”, suitably defined, are discouraged.
- We will generically think of modeling prevalence over time.

Motivation for Smoothing: Temporal Case

- Temporal setting: Even if the underlying prevalence is the same over time, we will see differences in the empirical estimates.
- Figure 3 demonstrates: We sampled binomial data with $n = 10, 20, 200$ and $p = 0.2$ (shown in blue) in all cases.
- In the top plot in particular, we might conclude large temporal variation, but all we are seeing is sampling variation.
- Figure 4 summarizes estimates from a second simulation in which there is a real temporal pattern – here we would not want to oversmooth and remove the trend.
- Later (Lecture 5) we will apply temporal smoothing models to these two sets of data.

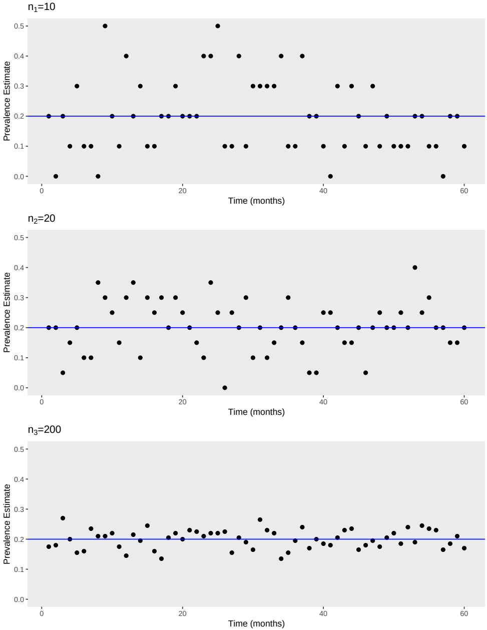


Figure 3: Prevalence estimates over time from simulated data with true prevalence of $p = 0.2$ (blue solid lines).

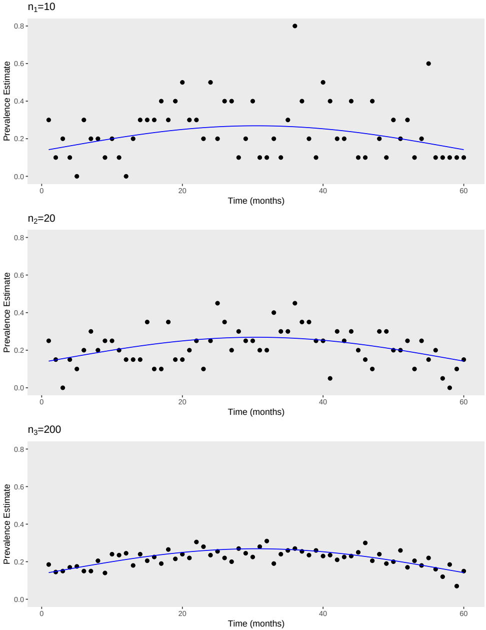


Figure 4: Prevalence estimates over time from simulated data, true prevalence corresponds to curved blue solid line.

Bayesian Learning

Bayes theorem

For a partition $\{H_1, \ldots, H_K\}$, the axioms of probability imply the following:

Rule of total probability : $\displaystyle\sum_{k=1}^{K} \Pr(H_k) = 1$

Rule of marginal probability :
$$\begin{aligned} \Pr(A) &= \sum_{k=1}^{K} \Pr(A \text{ and } H_k) \\ &= \sum_{k=1}^{K} \Pr(A|H_k) \Pr(H_k) \end{aligned}$$

Simple case: $K = 2$ with $H_1 = B$ and $H_2 = B^c$ (the complement of B):

$$\begin{aligned} \Pr(A) &= \Pr(A \text{ and } B) + \Pr(A \text{ and } B^c) \\ &= \Pr(A|B) \Pr(B) + \Pr(A|B^c) \Pr(B^c). \end{aligned}$$

Bayes theorem

$$\text{Bayes theorem}: \quad \Pr(H_j|E) = \frac{\overbrace{\Pr(E|H_j)}^{\text{“Likelihood”}}\,\overbrace{\Pr(H_j)}^{\text{“Prior”}}}{\underbrace{\Pr(E)}_{\text{Normalizing Constant}}} = \frac{\Pr(E|H_j)\Pr(H_j)}{\sum_{k=1}^{K}\Pr(E|H_k)\Pr(H_k)}$$

for $j = 1, \ldots, K$.

Anticipating Bayesian inference:

- One begins with (prior) beliefs about events H_j, $\Pr(H_j)$, and
- updates these to (posterior) beliefs $\Pr(H_j|E)$, given that an event E occurs.

Bayes theorem simple case:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Independence and conditional independence

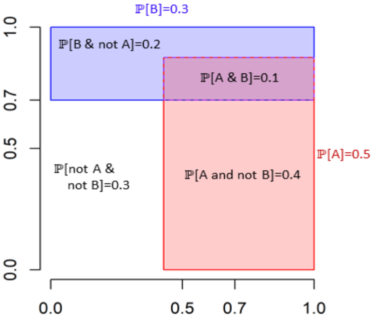
Conditional independence is a key concept when constructing statistical models – we start by describing independence.

For events A and B, it is always true that,

$$\Pr(\,A \text{ and } B\,) = \Pr(\,B\,) \times \Pr(\,A \,|\, B\,).$$

- The Venn diagram gives a representation of a set of probabilities over events A and B.
- In it we see,

  $\Pr(\,A|B\,) = 1/3 < 1/2 = \Pr(\,A\,).$
- Viewed in a Bayesian way, knowledge that B occurs has updated our beliefs about A.



Independence and conditional independence

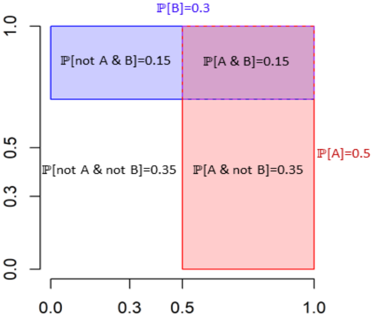
How about when we don’t learn anything from *B*’s occurrence?

Then

$$\Pr(A|B) = \Pr(A)$$

or equivalently

$$\Pr(A \text{ and } B) = \mathrm{F}$$



- The events *A* and *B* are said to be independent.
- Knowledge that *B* occurs does not affect our beliefs about *A*.
- Knowledge that *A* occurs does not affect our beliefs about *B*, i.e., this implies $\Pr(B|A) = \Pr(B)$.

Conditional Independence

In statistical modeling, independence is rarely relevant, but conditional independence is ubiquitous.

Extending this idea, events F and G are conditionally independent given H, if

$$\Pr(\, F \text{ and } G \,|\, H\,) = \Pr(\, F \,|\, H\,) \times \Pr(\, G \,|\, H\,),$$

or written another way:

$$\Pr(\, F \,|\, G, H\,) = \Pr(\, F \,|\, H\,).$$

Given H, knowledge that F occurred does not alter our beliefs in G occurring.

Conditional Independence: Example

Suppose we have the following events:

$F = \{$ a patient develops cancer $\}$
$G = \{$ patient's parent's genotype $\}$
$H = \{$ patient's genotype $\}$

Informal statement:

If we know the patient's genotype, does knowledge of the parents' genotype give any additional information?

Formal statement:

Does

$$\Pr(\, F \mid H\,) = \Pr(\, F \mid G, H\,)?$$

Answer: In genomic imprinting genes are expressed in a parent-of-origin-specific manner, i.e., the expression of the gene depends upon the parent who passed on the gene.

Bayes Theorem for Inference

- Unknown parameter $\boldsymbol{\theta}$, observed data $\boldsymbol{y}$.
- We derive the posterior distribution via Bayes theorem:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\Pr(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{\Pr(\boldsymbol{y})}. \quad (1)$$

- The denominator:

$$\Pr(\boldsymbol{y}) = \int \Pr(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathrm{E}[\Pr(\boldsymbol{y}|\boldsymbol{\theta})]$$

  is a normalizing constant to ensure the RHS of (1) integrates to 1 (we assume a continuous parameter $\boldsymbol{\theta}$).
- More colloquially:

$$\begin{aligned} \text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \\ &= \Pr(\boldsymbol{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \end{aligned}$$

  since in considering the posterior we only need to worry about terms that depend on the parameter $\boldsymbol{\theta}$.

Conditional Independence in Statistics

- Independence is rarely justified when constructing a sampling model – think about Bernoulli outcomes.
- Conditional independence assumption require more care in time series and spatial scenarios (in particular).
- Conditional independence is a key concept when building hierarchical models, as we will see – in this case prior distributions are fashioned using conditional independence.
- Markov random field models in particular are constructed from conditional independence assumptions (Rue and Held, 2005).
- Conditional independencies can be expressed through graphical models.

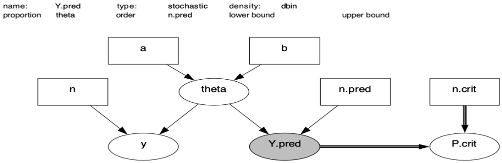


Figure 5: Example of a graphical model from Lunn *et al.* (2013).

Conditional Independence in Statistics

In likelihood-based inference, conditional independence is used all the time.

For example, the sampling model for data $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ is often taken to be:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}) &= p(y_1, \ldots, y_n|\boldsymbol{\theta}) \\ &= p(y_1|\boldsymbol{\theta}) \times p(y_2|y_1, \boldsymbol{\theta}) \times \cdots p(y_n|y_{n-1}, \ldots, y_1, \boldsymbol{\theta}) \\ &= p(y_1|\boldsymbol{\theta}) \times p(y_2|\boldsymbol{\theta}) \times \cdots p(y_n|\boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta}) \end{aligned}$$

where we have assumed conditional independence, i.e., given $\boldsymbol{\theta}$, the observations are independent.

Overview of Bayesian Inference

To carry out inference, integration is required, and a large fraction of the Bayesian research literature focusses on this aspect. Bayesian approaches to:

1. Estimation: marginal posterior distributions on parameters of interest.
2. Hypothesis Testing: Bayes factors give the evidence in the data with respect to two or more hypotheses, and provide one approach.
3. Prediction: via the predictive distribution.

These three endeavors will now be described in the context of a binomial model.

Bayesian Analysis of Binomial Data

Elements of Bayes Theorem for a Binomial Model

Suppose the data consist of N Bernoulli (0/1) responses y_i, $i = 1, \ldots, N$.

We may assume these responses are conditionally independent, given a common “success” probability θ.

Under this conditional independence assumption, the distribution of the total $y = \sum_{i=1}^{N} y_i$ is binomial:

$$\Pr(Y = y|\theta) = \left(\begin{array}{c} N \\ y \end{array} \right) \theta^y (1 - \theta)^{N-y} \tag{2}$$

and tells us the probability of seeing $Y = y$, for the permissible values $y = 0, 1, \ldots, N$ given the probability θ.

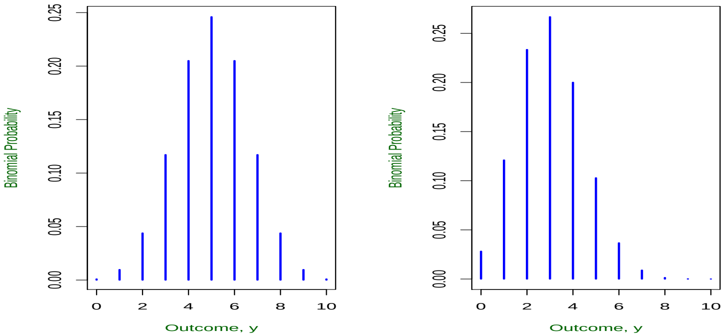


Figure 6: Binomial distributions for two values of θ with $N = 10$.

Elements of Bayes Theorem for a Binomial Model

For fixed y, we may view (2) as a function of θ – this is the likelihood function:

$$L(\theta) = \theta^y (1 - \theta)^{N-y}.$$

The maximum likelihood estimate (MLE) is the average number of successes:

$$\widehat{\theta} = \frac{y}{N} = \overline{y},$$

and gives the highest probability to the observed data, i.e. maximizes the likelihood function.

The standard error of this estimate is $\sqrt{\widehat{\theta}(1 - \widehat{\theta})/N}$.

Bayes and frequentist estimates for binomial

If $y = 0$ ($y = N$), we have estimate $\widehat{\theta} = 0$ (=1) and a standard error of 0, which is clearly problematic.

"Adjusted Wald interval": Agresti and Coull (1998) discuss the use of the alternative estimator:

$$\tilde{\theta} = \frac{4}{N+4}\frac{1}{2} + \frac{N}{N+4}\overline{y},$$

to give the interval:

$$\tilde{\theta} \quad \pm \quad 1.96\sqrt{\tilde{\theta}(1-\tilde{\theta})/N}$$

as an approximation to an earlier suggestion of Wilson (1927).

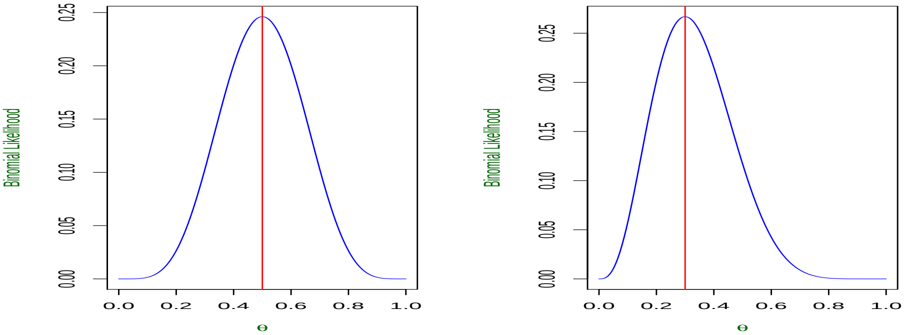


Figure 7: Binomial likelihoods for values of $y = 5$ (left) and $y = 10$ (right), with $N = 10$. The MLEs are indicated in red.

The Beta Distribution as a Prior Choice for Binomial θ

Bayes Theorem:

$$p(\theta|y) \propto p(y|\theta) \times p(\theta).$$

- Bayes theorem requires the likelihood, which we have already specified as binomial, and the prior.
- For a probability $0 < \theta < 1$ an obvious candidate prior is the uniform distribution on (0,1): but this is too restrictive in general.
- The beta distribution, Beta(a, b), is more flexible and so may be used for θ, with a and b specified in advance, i.e., *a priori*. The uniform distribution is a special case with $a = b = 1$.
- The form of the beta distribution is

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

for $0 < \theta < 1$, where $\Gamma(\cdot)$ is the gamma function[1].

[1] $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$

The Beta Distribution as a Prior Choice for Binomial θ

- The Beta(a, b) distribution is valid[2] for $a > 0, b > 0$.
- How can we think about specifying a and b?
- For the normal distribution the parameters μ and σ^2 are just the mean and variance, but for the beta distribution a and b have no such simple interpretation.
- The mean and variance are:

$$\begin{aligned} \mathrm{E}[\theta] &= \frac{a}{a+b} \\ \mathrm{var}(\theta) &= \frac{\mathrm{E}[\theta](1-\mathrm{E}[\theta])}{a+b+1}. \end{aligned}$$

  Hence, increasing a and b concentrates the distribution about the mean.
- The quantiles, e.g. the median or the 10% and 90% points, are not available as a simple formula, but are easily obtained within software – in R we use the function `qbeta(p,a,b)`.

[2] A distribution is valid if it is non-negative and integrates to 1

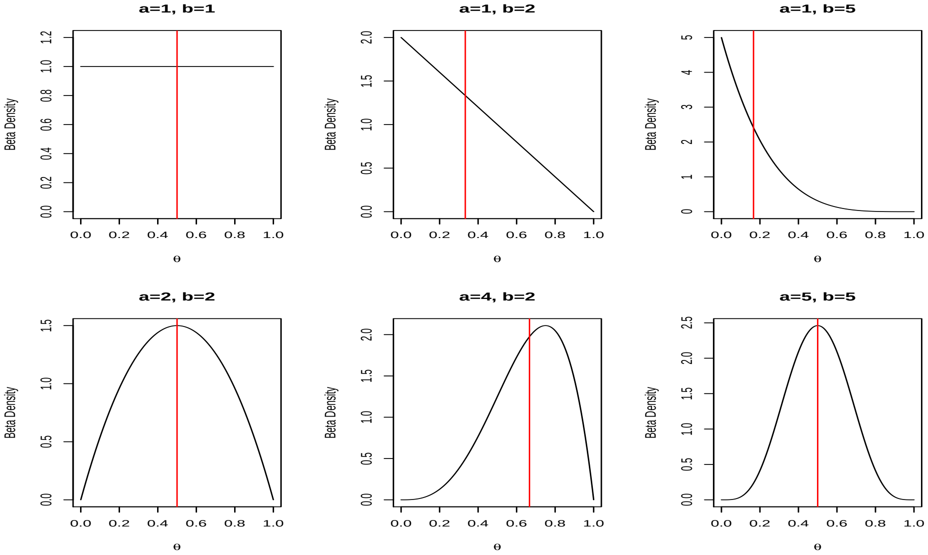


Figure 8: Beta distributions, Beta(a, b), the red lines indicate the means.

Samples to Summarize Beta Distributions

- In general, there is a duality between probability distributions and samples from distributions: given a probability distribution we can generate a sample, and given a sample, we can construct the probability distribution from which they arose — this is key to the direct sampling and Markov chain Monte Carlo (MCMC) Bayesian implementation methods.
- Probability distributions can be investigated by generating samples and then examining histograms, moments and quantiles.
- In Figure 9 we show histograms of beta distributions for different choices of a and b.

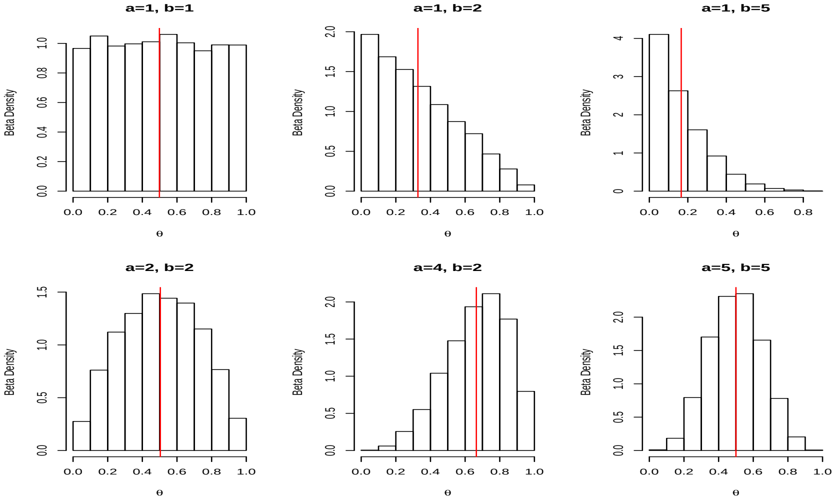


Figure 9: Random samples from beta distributions; sample means as red lines.

Samples for Describing Weird Parameters

- So far the samples we have generated have produced summaries we can easily obtain anyway.
- But what about functions of the probability θ, such as the odds $\theta/(1-\theta)$?
- Once we have samples for θ we can simply transform the samples to the functions of interest.
- We may have clearer prior opinions about the odds, than the probability.

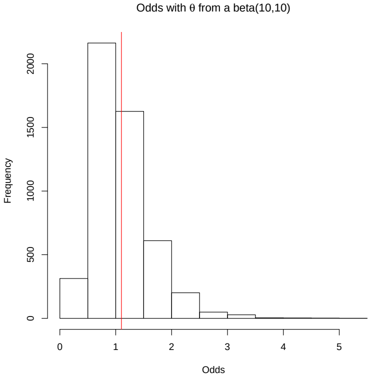


Figure 10: Samples from the prior on the odds $\theta/(1-\theta)$ with $\theta \sim \text{Beta}(10, 10)$, the red line indicates the sample mean.

Issues with Uniform Priors

We might think that if we have little prior opinion about a parameter then we can simply assign a uniform prior, i.e. a prior

$$p(\theta) \propto \text{ const.}$$

There are two problems with this strategy:

1. We can't be uniform on all scales since, if $\phi = g(\theta)$:

$$\underbrace{p_\phi(\phi)}_{\text{Prior for } \phi} = \underbrace{p_\theta(g^{-1}(\phi))}_{\text{Prior for } \theta} \times \underbrace{\left|\frac{d\theta}{d\phi}\right|}_{\text{Jacobian}}$$

   and so if $g(\cdot)$ is a nonlinear function, the Jacobian will be a function of ϕ and hence not uniform.

2. If the parameter is not on a finite range, an improper distribution will result (that is, the form will not integrate to 1). This can lead to an improper posterior distribution, and without a proper posterior we can't do inference.

Are Priors Really Uniform?

- We illustrate the first (non-uniform on all scales) point.
- In the binomial example a uniform prior for θ seems a natural choice.
- But suppose we are going to model on the logistic scale so that

$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

  is a quantity of interest.
- A uniform prior on θ produces the very non-uniform distribution on ϕ in the figure.
- Not being uniform on all scales is not necessarily a problem, and is correct probabilistically, but one should be aware of this characteristic.

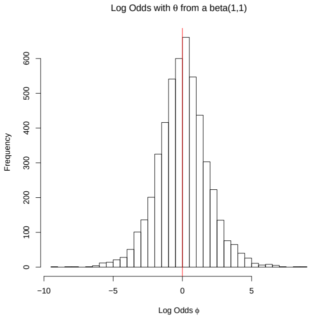


Figure 11: Samples from the prior on the odds $\phi = \log[\theta/(1-\theta)]$ with $\theta \sim \text{Beta}(1,1)$, the red line indicates the sample mean.

Posterior Derivation: The Quick Way

- When we want to identify a particular probability distribution we only need to concentrate on terms that involve the random variable.
- For example, if the random variable is X and we see a density of the form

  $$p(x) \propto \exp(c_1 x^2 + c_2 x),$$

  for constants c_1 and c_2, then we know that the random variable X must have a normal distribution.

Posterior Derivation: The Quick Way

- For the binomial-beta model we concentrate on terms that only involve θ.
- The posterior is

$$\begin{aligned} p(\theta|y) &\propto \Pr(y|\theta) \times p(\theta) \\ &\propto \theta^{y}(1-\theta)^{N-y} \times \theta^{a-1}(1-\theta)^{b-1} \\ &= \theta^{y+a-1}(1-\theta)^{N-y+b-1} \end{aligned}$$

- We recognize this as the important part of a

$$\text{Beta}(y+a, N-y+b)$$

  distribution.
- We know what the normalizing constant must be, because we have a function which must integrate to 1.

The beta posterior

The above is an example of a conjugate Bayesian analysis in which the prior is in the same family as the posterior, unfortunately for most models such computationally convenient analyses are not possible.

Recall, from earlier, the adjusted Wald interval:

$$\begin{aligned} \tilde{\theta} &\pm& 1.96\sqrt{\tilde{\theta}(1-\tilde{\theta})/N}\,,\ \text{where} \\ \tilde{\theta} &=& \frac{1}{2}\frac{4}{N+4}+\overline{y}\frac{N}{N+4}. \end{aligned}$$

Notice the link with the adjusted Wald interval for the 0 successes case, the estimate is equal to the posterior mean with a Beta(a, b) prior with $a = b = 2$.

Posterior Summaries

- We will rarely want to report a point estimate alone, whether it be a posterior mean or posterior median.
- Interval estimates are obtained in the obvious way.
- A simple way of performing testing of particular parameter values of interest is via examination of interval estimates.
- For example, does a 95% interval contain the value $\theta_0 = 0.5$?

Other Posterior Summaries

- In our beta-binomial example, a 90% posterior credible interval (θ_L, θ_U) results from the points

$$
\begin{aligned}
0.05 &= \int_0^{\theta_L} p(\theta|y)\, d\theta \\
0.95 &= \int_0^{\theta_U} p(\theta|y)\, d\theta
\end{aligned}
$$

- The quantiles of a beta are not available in closed form, but easy to evaluate in `R`:

```
y <- 7; N <- 10; a <- b <- 1
qbeta(c(0.05,0.5,0.95),y+a,N-y+b)
[1] 0.4356258 0.6761955 0.8649245
```

- The 90% credible interval is (0.44,0.86) and the posterior median is 0.68.

Bayes and frequentist estimates for binomial

Example: $N = 10, y = 0$ gives

$$\tilde{\theta} = \frac{4}{10+4}\frac{1}{2} + \frac{10}{10+4}\overline{y} = \frac{4}{28} = 0.14$$

with adjusted standard error

$$\sqrt{\tilde{\theta}(1-\tilde{\theta})/10} = \sqrt{\frac{4}{28}\left(1-\frac{24}{28}\right)/10} = 0.11$$

Under the Bayesian interpretation Bayesian procedure, with a Beta(2,2) prior for θ:

```
> y <- 0;N <- 10;a <- b <- 2;apost <- a+y;bpost <- b+(N-y)
> qbeta(p=c(0.025,0.975),apost,bpost)
[1] 0.01920667 0.36029744
```

So Bayes 95% interval is (0.019,0.36).

A More Interesting Example

Suppose a seroprevalence test is carried out with sensitivity

$$\delta = \text{Pr(+ve test} \mid \text{disease)}$$

and specificity,

$$\gamma = \text{Pr(-ve test} \mid \text{no disease)}.$$

Let π be the true prevalence.

We test *n* people and *y* are recorded as having the disease, and a starting model is

$$y|p \sim \text{Binomial}(N, p)$$

where *p* is the probability of a +ve test result. with

$$\begin{aligned} p &= \text{Pr(+ve test)} \\ &= \text{Pr(+ve test} \mid \text{disease)} \, \text{Pr(disease)} \\ &+ \text{Pr(+ve test} \mid \text{no disease)} \, \text{Pr(no disease)} \\ &= \delta\pi + (1-\gamma)(1-\pi) \\ &= \pi(\delta+\gamma-1) + (1-\gamma) \end{aligned}$$

Suppose for simplicity the sensitivity and specificity are known and we want to estimate π.

A More Interesting Example

With this binomial model the MLE is (exercise!):

$$\widehat{\pi} = \frac{y - N(1-\gamma)}{N(\delta + \gamma - 1)}$$

A Bayesian model is

$$\begin{aligned} y|\pi &\sim \text{Binomial}(N, \pi(\delta + \gamma - 1) + (1 - \gamma)) \\ \pi &\sim \text{Beta}(a, b) \end{aligned}$$

Not conjugate!

However, a simple rejection algorithm (Smith and Gelfand, 1992) can be implemented that simulates samples from the posterior $p(\pi|y)$.

Direct Sampling

We briefly describe the rejection algorithm that can be used to generate samples from the posterior.

Let $\boldsymbol{\theta}$ denote the unknown parameters and assume that we can evaluate the maximized likelihood

$$M = \sup_{\boldsymbol{\theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y} \mid \widehat{\boldsymbol{\theta}})$$

where $\widehat{\boldsymbol{\theta}}$ is the MLE. The algorithm then proceeds as follows:

1. Generate $U \sim U(0,1)$ and, independently, $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.
2. Accept $\boldsymbol{\theta}$ if
$$U < \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})}{M},$$
otherwise return to 1.

The probability that a point is accepted is given by

$$p_a = \frac{\int p(\boldsymbol{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{M} = \frac{p(\boldsymbol{y})}{M}.$$

COVID-19 Prevalence Estimate

- In early April, 2020, Bendavid *et al.* (2020) recruited 3330 residents of Santa Clara County, California and tested them for COVID-19 antibodies. 50 people tested positive, yielding a raw estimate of 1.50%.
- We take the sensitivity as 0.8 and the specificity as 0.995 and the prior parameters as $a = b = 1$.
- See Gelman and Carpenter (2020) for a more comprehensive Bayesian analysis.

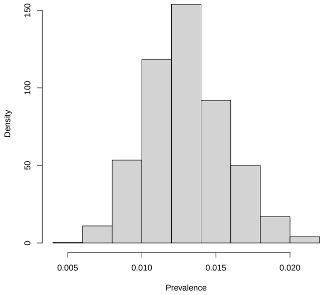


Figure 12: Histogram representation of the posterior distribution for the prevalence π. The posterior median is 1.28% and a 90% interval is (0.88%,1.77%).

Summary

Conjugate analyses are computationally convenient but rarely available in practice.

Historically, the philosophical standpoint of Bayesian statistics was emphasized, now pragmatism is taking over.

Summary

Benefits of a Bayesian approach:

- Inference is based on probability and output is very intuitive.
- Framework is flexible, and so complex models can be built.
- Can incorporate prior knowledge!
- If the sample size is large, prior choice is less crucial.

Challenges of a Bayesian analysis:

- Require a likelihood and a prior, and inference is only as good as the appropriateness of these choices.
- Computation can be daunting, though software is becoming more user friendly and flexible; later we will describe and illustrate a number of approaches including `INLA` and `Stan`.
- One should be wary of model becoming too complex – we have the technology to contemplate complicated models, but do the data support complexity?

References

Agresti, A. and Coull, B. A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra, R., Tedrow, J., *et al.* (2020). Covid-19 antibody seroprevalence in Santa Clara county, California. *MedRxiv*.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society, Series A*. To appear.

Lunn, D., Jackson, C., Best, N., Spiegelhalter, D., and Thomas, A. (2013). *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press, Boca Raton.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome research*, **21**, 1728–1737.

Smith, A. and Gelfand, A. (1992). Bayesian statistics without tears: a sampling-resampling prespective. *American Statistician*, **46**, 84–88.

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, **22**, 209–212.

Technical Appendices

Posterior Derivation: The Long (Unnecessary) Way

- The posterior can also be calculated by keeping in all the normalizing constants:

$$\begin{aligned} p(\theta|y) &= \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)} \\ &= \frac{1}{\Pr(y)} \left(\begin{array}{c} N \\ y \end{array} \right) \theta^y (1-\theta)^{N-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}. \quad (3) \end{aligned}$$

- The normalizing constant is

$$\begin{aligned} \Pr(y) &= \int_0^1 \Pr(y|\theta) \times p(\theta) d\theta \\ &= \left(\begin{array}{c} N \\ y \end{array} \right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{y+a-1} (1-\theta)^{N-y+b-1} d\theta \\ &= \left(\begin{array}{c} N \\ y \end{array} \right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)} \end{aligned}$$

- The integrand on line 2 is a Beta$(y+a, N-y+b)$ distribution, up to a normalizing constant, and so we know what this constant has to be.

Posterior Derivation: The Long (and Unnecessary) Way

- The normalizing constant is therefore:

$$\Pr(y) = \left(\begin{array}{c} N \\ y \end{array} \right) \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(N-y+b)}{\Gamma(N+a+b)}$$

- This is a probability distribution, i.e. $\sum_{y=0}^{N} \Pr(y) = 1$ with $\Pr(y) > 0$, for $y = 0, 1, \ldots, N$.
- For a particular y value, this expression tells us the probability of that value given the model, i.e. the likelihood and prior we have selected: this will reappear later in the context of hypothesis testing.
- Substitution of $\Pr(y)$ into (3) and canceling the terms that appear in the numerator and denominator gives the posterior:

$$p(\theta|y) = \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \theta^{y+a-1}(1-\theta)^{N-y+b-1}$$

which is a $\text{Beta}(y+a, N-y+b)$.

The Posterior Mean: A Summary of the Posterior

- Recall the mean of a $\text{Beta}(a, b)$ is $a/(a+b)$.
- The posterior mean of a $\text{Beta}(y+a, N-y+b)$ is therefore

$$\begin{aligned}
\text{E}[\theta|y] &= \frac{y+a}{N+a+b} \\
&= \frac{y}{N+a+b} + \frac{a}{N+a+b} \\
&= \frac{y}{N} \times \frac{N}{N+a+b} + \frac{a}{a+b} \times \frac{a+b}{N+a+b} \\
&= \text{MLE} \times \text{W} + \text{Prior Mean} \times \text{(1-W)}.
\end{aligned}$$

- The weight W is

$$\text{W} = \frac{N}{N+a+b}.$$

- As N increases, the weight tends to 1, so that the posterior mean gets closer and closer to the MLE.
- Notice that the uniform prior $a = b = 1$ gives a posterior mean of

$$\text{E}[\theta|y] = \frac{y+1}{N+2}.$$