Bayesian SAE using Complex Survey Data
Lecture 1: Bayesian Statistics

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Bayesian Learning

Probability and Bayes Theorem

Standard Distributions and Conjugacy
- Binomial Distribution
- Normal Distribution

Technical Appendix: Details of Calculations for the Binomial Model

Motivation

Introduction

- In this lecture we will first consider generic Bayesian learning.
- Background reading: Chapters 1 and 2 of Hoff (2009)[1].
- Simulated continuous responses and count data will be used to motivate normal and binomial models, respectively.
- After introducing these examples, we give a brief review of probability theory.
- Conjugate priors will be introduced.

[1] Background Text on Bayes: P.D. Hoff (2009), *A First Course in Bayesian Statistical Methods*, Springer.

Motivating Example: Binomial Count Data

As a motivating example, consider the 48 health reporting areas (HRAs) of King County.

Later, we will analyze data from BRFSS (which uses a complex sampling design), but for now we look at simulated data in which in each of the HRAs, samples of size n_i in HRA i are taken using simple random sampling (SRS) from the total population N_i, $i = 1, \ldots, 48$.

For each sampled individual, let d represent their diabetes status and z their weight.

The objective is to estimate, in each HRA i, the:

- True number with diabetes, say D_i, and the true fractions with diabetes $q_i = D_i/N_i$.
- The average weight over the HRA, μ_i.

These are simple examples of small area estimation (SAE).

Motivating Examples: Normal and Binomial Data

To illustrate techniques, we simulate data in HRAs using simple random sampling in each area.

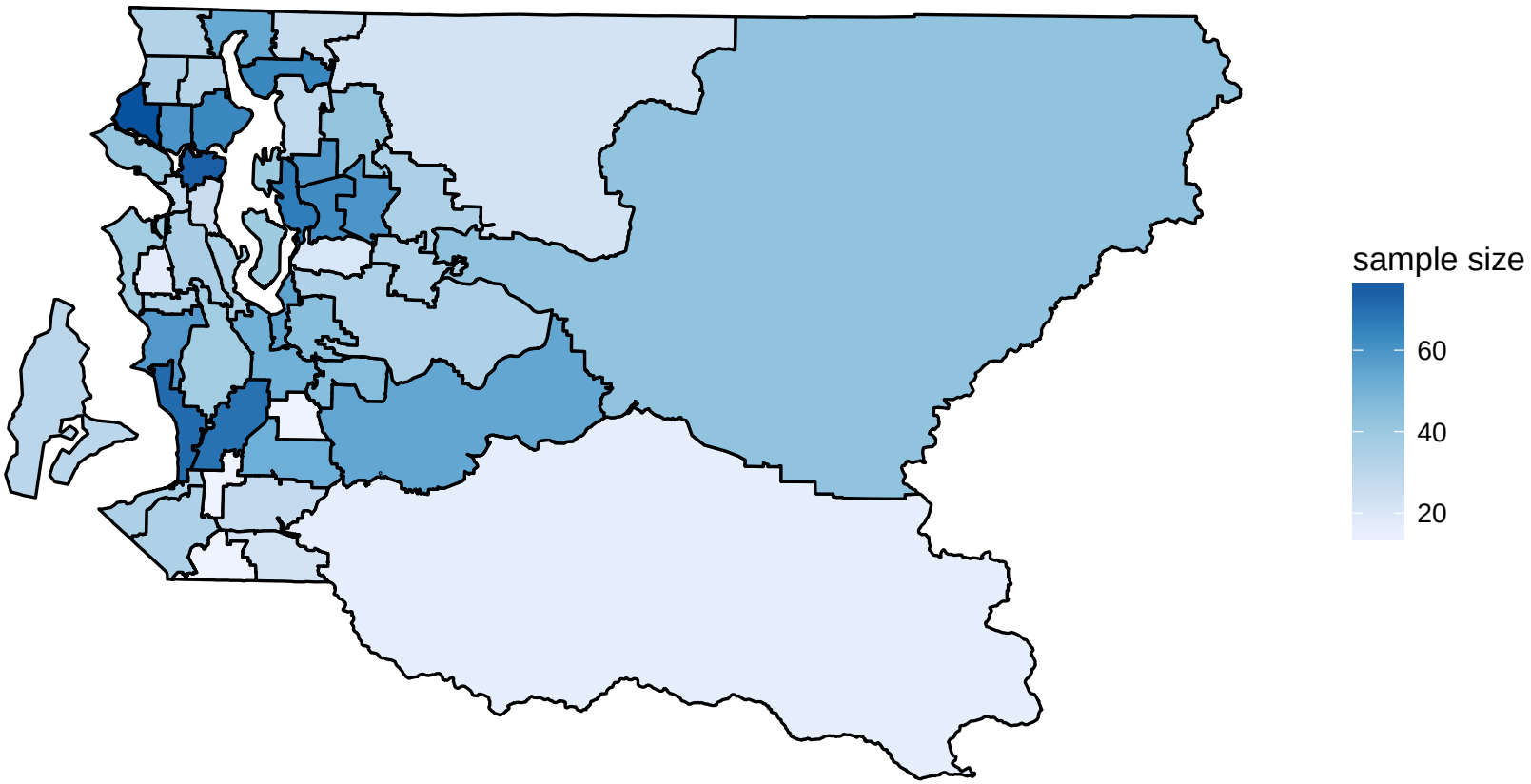


Figure 1: Sample sizes of simulated survey.

Motivating Examples: Normal and Binomial Data

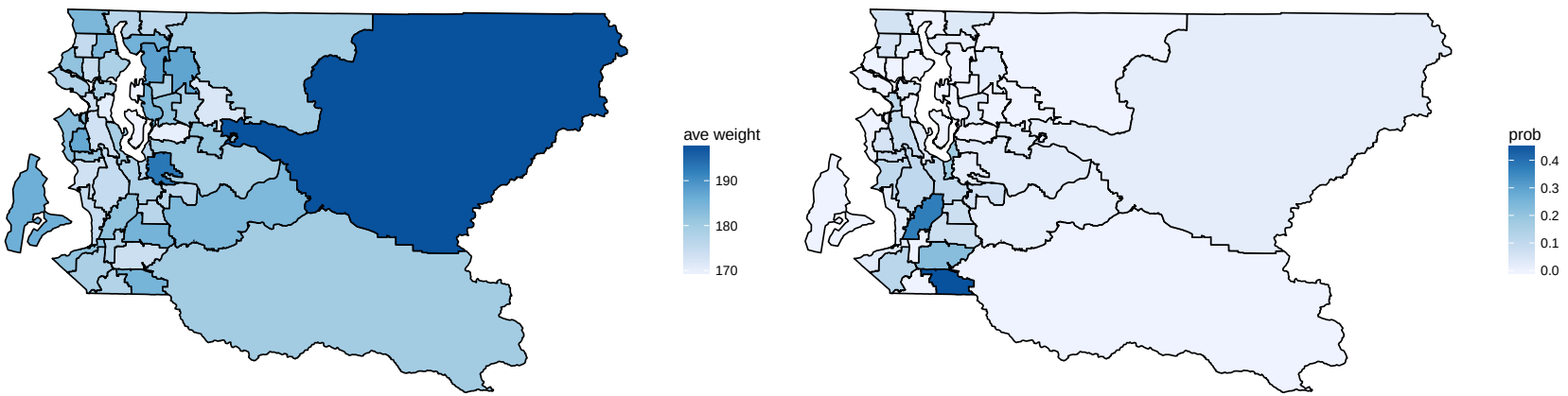


Figure 2: Sample mean weights (left) and fractions with diabetes (right).

Motivating Examples: Normal and Binomial Data

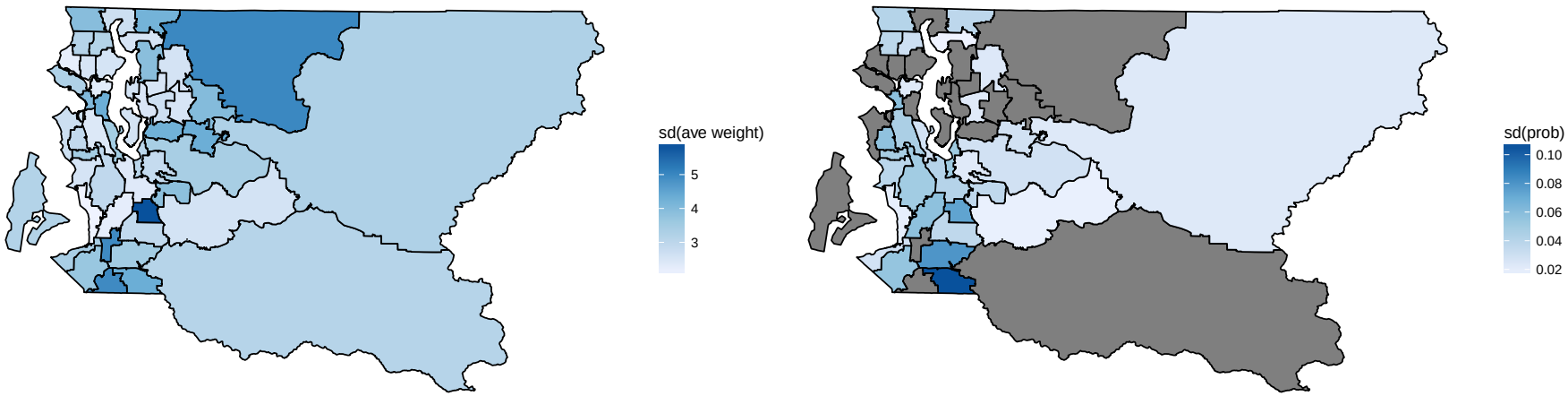


Figure 3: Standard errors of: mean weights (left) and fractions with diabetes (right). Gray areas in the right map correspond to areas with zero counts, and hence an estimated standard error of zero.

Motivating Examples: Normal and Binomial Data

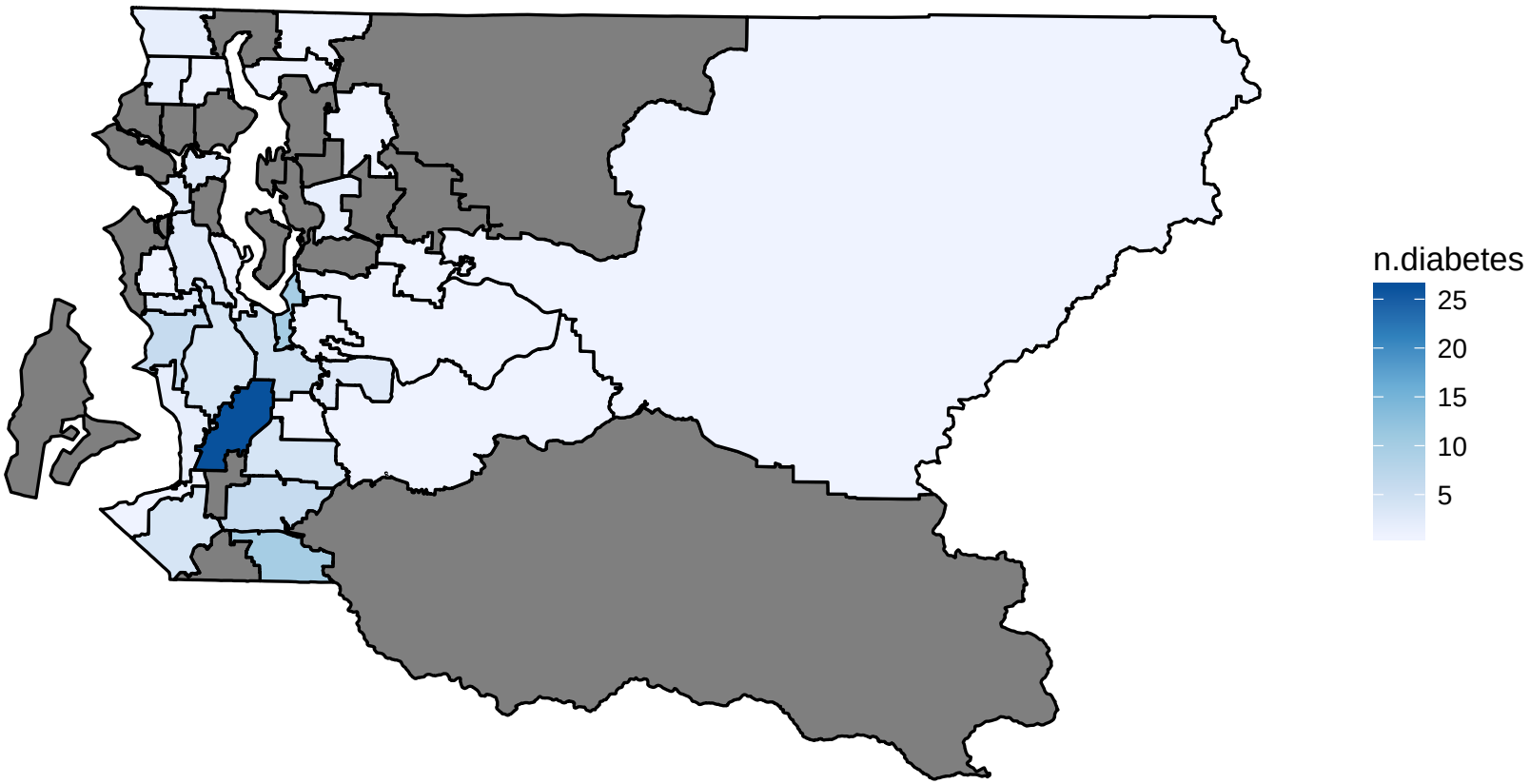


Figure 4: Number of individuals in the sample with diabetes; zero counts are indicated in gray.

Bayesian Learning

Probability and information

We often use “probability” informally to express belief.

If we have strong belief that an event will occur, then we would assign a high probability to the event.

When probabilities are assigned in everyday life there is an implicit link with the information that the assigner has available to him/her.

This can be made mathematically formal via Bayesian theory:

- Probability can numerically quantify rational beliefs
- There is a relationship between probability and information
- Bayes theorem is a rational method for updating uncertainty based on information

Bayesian methods

Bayesian methods are data analysis tools that are derived from the principles of Bayesian inference.

Bayesian methods provide:

- parameter estimates with good statistical properties;
- parsimonious descriptions of observed data;
- predictions for missing data and forecasts of future data;
- a framework for model estimation, selection and validation;
- a means by which prior information can be incorporated.

Statistical induction

Induction: Reasoning from specific cases to a general principle.

Statistical induction: Using a data sample to infer population characteristics.

Notation:

Parameter: θ denotes unknown population characteristics.

Data: y is the outcome of a survey or experiment.

In the SAE experiments, our goal is to make inference about (in a generic area):

- Diabetes outcome: the unknown θ corresponds to q, the probability of diabetes, and the data y to d, and we want to learn about q given d.
- Weight outcome: the unknown θ corresponds to μ, the average weight, and the data y to z, and we want to learn about μ given z.

Ingredients of a Bayesian analysis

Parameter and sample spaces:

Sample space: $\mathcal{Y}$ is the set of all possible datasets.

Parameter space: Θ is the set of all possible θ-values

For the SAE examples, in one area:

Sample space for diabetes: $\mathcal{Y} = \{0, 1, \ldots, n\}$ is the set of all possible outcomes (=y).

Parameter space for diabetes: $\Theta = [0, 1]$ is the set of all possible values of the probability θ (=q).

Sample space for weight: $\mathcal{Y} = (0, \infty)$ is the set of all possible outcomes (=z).

Parameter space for weight: $\Theta = (0, \infty)$ is the set of all possible values of the probability θ (=μ).

Ingredients of a Bayesian analysis

Quantifying information:

Prior distribution: $p(\theta)$, defined for all $\theta \in \Theta$, describes our probabilistic beliefs about θ, the true value of the population parameter.

Sampling model: $p(y|\theta)$, defined for $\theta \in \Theta, y \in \mathcal{Y}$, describes our probabilistic beliefs that y will be the experimental outcome, for each θ.

Updating information:

Bayes theorem: After obtaining data y, the posterior distribution is

$$\underbrace{p(\theta|y)}_{\text{Posterior}} = \frac{p(y|\theta)p(\theta)}{p(y)} \propto \underbrace{p(y|\theta)}_{\text{Likelihood}} \underbrace{p(\theta)}_{\text{Prior}},$$

where

$$p(y) = \int_{\Theta} p(y|\theta)p(\theta)\, d\theta$$

is the normalizing constant (to ensure the posterior is legal probabilistically).

Ingredients of a Bayesian analysis

For the SAE diabetes data (in a generic area):

Prior distribution: $p(q)$ describes our beliefs about the unknown probability q of an individual having diabetes, before we look at the data.

Sampling model: $p(d|q)$, describes the probabilities of all of the possible outcomes $d = 0, 1, \ldots, n$ given we (hypothetically) know the value of the probability q. When viewed as a function of q, $p(d|q)$ is known as the likelihood.

Posterior distribution: $p(q|d)$ describes our beliefs about the unknown probability q, after we combine the data (via the sampling model) and the prior.

Role of prior information

There is a theoretical justification (e.g., Bernardo and Smith 1994) that tells us that probabilities should express uncertainties and how beliefs should change after seeing new information (via Bayes theorem!).

Bayes theorem does not tell us what our beliefs should be.

Adherents of frequentist inference might question the optimality of Bayesian inference, given the imperfect manner in which beliefs (in both the sampling model and the prior) are specified – clear we need to be careful in how we specify the model.

SAE Example

A natural choice for the number of individuals (out of n) with diabetes is:

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

The maximum likelihood estimate (MLE) is

$$\widehat{\theta} = \frac{y}{n} = \bar{y}$$

with standard error

$$\sqrt{\frac{\theta(1-\theta)}{n}}$$

which is estimated by

$$\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{n}}.$$

Suppose for a particular area $y = 0$, then $\widehat{\theta} = 0$ with standard error 0.

Comparison to non-Bayesian methods in the SAE setting

Non-Bayesian 95% confidence (Wald) interval:

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n}$$

If we have $y = 0$, then the interval is 0 ± 0, which is clearly unacceptable.

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

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

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

Can be seen as approximately Bayesian, with a beta(2,2) prior for θ (see later).

Probability and Bayes Theorem

The Big Picture

- Statistics: Probability models for data.
- Data: May be represented as real numbers.
- Probability Theory: Starting with sample spaces and events we consider a function (the probability) that measures “size” of the event. Mathematically, probabilities are measures of uncertainty obeying certain properties.
- Random Variables: Provide the link between sample spaces and data.

Basic Probability Review

Set notation:

- $A \cup B$ represents union, "A or B".
- $A \cap B$ represents intersection, "A and B".
- $\emptyset$ is the empty set.
- $A_1, A_2, \ldots,$ are mutually exclusive (disjoint) events if $A_i \cap A_j = \emptyset$, for all pairs i, j, $i \neq j$.
- Ω is the sample space, and $\mathcal{F}$ be a suitable collection[2] of subsets of Ω.
- A^c is the complement of A, so that $A \cup A^c = \Omega$.

Axioms of Probability:

P1 $\Pr(\Omega) = 1$,

P2 $\Pr(A) \geq 0$ for any event $A \in \mathcal{F}$,

P3 $\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$ for mutually exclusive events $A_1, A_2, \cdots \in \mathcal{F}$.

[2] Technically, a σ-algebra

Basic Probability Review

Definition: For events A and B in Ω, with $\Pr(A) > 0$ the conditional probability that B occurs, given that A occurs, is

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

Important point: $\Pr(\cdot|A)$ satisfies the axioms of probability, but $\Pr(B|\cdot)$ does not!

In particular, it is always true that: $\Pr(A|B) + \Pr(A^c|B) = 1.$

In contrast, in general: $\Pr(B|A) + \Pr(B|A^c) \neq 1.$

Basic Probability Review

Often confused, for example, the prosecutor's fallacy:

$$\Pr(\text{ evidence} \mid \text{guilt }) \neq \Pr(\text{ guilt} \mid \text{evidence }).$$

Example: {evidence = blue plaid shirt} and we know crime was committed by someone with a blue plaid shirt, so

$$\Pr(\text{ evidence} \mid \text{guilt }) = 1$$

but

$$\Pr(\text{ guilt} \mid \text{evidence }) < 1.$$

Example

P3 with two events: $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$ if $A_1 \cap A_2 = \emptyset$

Example:

- Suppose we have data on deaths by age, up to age 18, in years, in a certain population
- $A_1 = \{$ death in first year $\}$, $A_2 = \{$ death at ages 2–5 $\}$, $A_3 = \{$ death at ages 6–18$\}$
- $\Pr(A_1) = 0.04$, $\Pr(A_2) = 0.01$, $\Pr(A_3) = 0.003$.
- A_1 and A_2 are disjoint, and so the probability of death in first 5 years is

$$\begin{aligned} \Pr(\text{death in first 5 years}) &= \Pr(A_1 \cup A_2) \\ &= \Pr(A_1) + \Pr(A_2) \\ &= 0.04 + 0.01 \\ &= 0.05 \end{aligned}$$

Events and partitions

Definition: A collection of sets $\{H_1, \ldots, H_K\}$ is a partition of another set $\mathcal{H}$ if

1. the events are disjoint, which we write as $H_i \cap H_j = \emptyset$ for $i \neq j$;
2. the union of the sets is $\mathcal{H}$, which we write as $\cup_{k=1}^{K} H_k = \mathcal{H}$.

If $\mathcal{H}$ is the set of all possible truths (i.e., $\mathcal{H} = \Omega$) and $\{H_1, \ldots, H_K\}$ is a partition of $\mathcal{H}$, then exactly one out of $\{H_1, \ldots, H_K\}$ contains the truth.

Example: $\mathcal{H}$=someone's number of children

- $\{0, 1, 2, 3 \text{ or more}\}$;
- $\{0, 1, 2, 3, 4, 5, 6, \ldots\}$.

Events and partitions

Example: $\mathcal{H}$= the strata of a household in the Kenya 2008–2009 DHS:

1. { Western, Rural }
2. { Western, Urban }
3. { Nyanza, Rural }
4. { Nyana, Urban }
5. { Rift Valley, Rural }
6. { Rift Valley, Urban }
7. { Eastern, Rural }
8. { Eastern, Urban }
9. { North Eastern, Rural }
10. { North Eastern, Urban }
11. { Coast, Rural }
12. { Coast, Urban }
13. { Central, Rural }
14. { Central, Urban }
15. { Nairobi, Urban }

Bayes theorem

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

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

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

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: the classic example

Set up:

- 1% of people have a certain genetic defect.
- 90% of tests for the gene detect the defect (true positives).
- 5% of the tests are false positives.

If a person gets a positive test result, what are the odds they actually have the genetic defect?

First, define events and translate the above:

- A = event of having the defective gene, so that $\Pr(A) = 0.01$. A and A^c form a partition so the probability of not having the gene is $\Pr(A^c) = 0.99$.
- Y = event of a positive test result; this can happen in two ways, via either a true positive (for an A person) or a false positive (for an A^c person).

From the information above:

- $\Pr(Y|A) = 0.9$ is the chance of a positive test result given that the person actually has the gene.
- $\Pr(Y|A^c) = 0.05$ is the chance of a positive test if the person doesn't have the gene.

Bayes theorem: the classic example

Bayes theorem allows us to calculate the probability of the gene defect, given the test results:

$$\Pr(A|Y) = \frac{\Pr(Y|A)\Pr(A)}{\Pr(Y)}$$

First, let's consider the denominator, the probability of a positive test result:

$$\begin{aligned}
\Pr(Y) &= \Pr(Y|A)\Pr(A) + \Pr(Y|A^c)\Pr(A^c) \\
&= \underbrace{0.9 \times 0.01}_{\text{Positive and defective gene}} + \underbrace{0.05 \times 0.99}_{\text{Positive and non-defective gene}} \\
&= 0.009 + 0.0495 \\
&= 0.0585.
\end{aligned}$$

It is clear that the event of a positive test result is dominated by false positives.

Bayes theorem: the classic example

The (posterior) probability of interest is:

$$\Pr(A|Y) = \frac{0.9 \times 0.01}{0.0585} = \frac{0.009}{0.0585} = 0.154,$$

so there is a 15.4% chance that a person with a positive test result has the defective gene.

At first sight, this low probability may seem surprising but the posterior to prior odds is

$$\frac{\Pr(A|Y)}{\Pr(A)} = \frac{0.154}{0.01} = 15.4,$$

so that we have changed our beliefs by quite a large amount.

Bayes theorem

A more accurate representation acknowledges that all probabilities are also conditional on all current relevant knowledge/information, I.

$$\text{Bayes theorem}: \quad \Pr(H_j|E,I) = \frac{\Pr(E|H_j,I)\Pr(H_j|I)}{\Pr(E|I)}$$

$$= \frac{\Pr(E|H_j,I)\Pr(H_j|I)}{\sum_{k=1}^{K}\Pr(E|H_k,I)\Pr(H_k|I)}$$

Usually the conditioning on I is suppressed for notational ease, but one should always keep it in mind...

Different individuals, have different information, and so it should be no surprise that the required elements of Bayes theorem (likelihood and prior) may differ between individuals.

Note: all of the above is unambiguous, it's just a bunch of math, but it doesn't tell us how to assign prior probabilities or specify sampling models (likelihoods).

The meaning of probability

- Mathematically speaking probability is a function that obeys certain properties and, from this standpoint, one need not worry too much about the interpretation of probability.
- When it comes to statistical inference, however, we will see that the interpretation given to probabilities influences the criteria by which procedures are judged.
- In the frequentist view, probabilities are interpreted as limiting frequencies observed over (hypothetical) repetitions in identical situations (we will encounter this when we discuss design-based inference).
- In the subjective view, probabilities are purely personal. One way of assigning probabilities is the following.
  - The probability of an event E is the price one is just willing to pay to enter a game in which one can win a unit amount of money if E is true.
  - For example, if I believe a coin is fair and I am to win 1 unit if a head (the event E) arises, then I would pay $\frac{1}{2}$ a unit of money to enter the bet.

Standard Distributions and Conjugacy

Discrete random variables

Let Y be a random variable, an unknown numerical quantity.

Let $\mathcal{Y}$ be the set of all possible values of Y.

Y is discrete if the set of possible outcomes is countable, meaning that $\mathcal{Y}$ can be expressed as $\mathcal{Y} = \{y_1, y_2, \ldots\}$.

Examples

- $Y =$ number of people in a population with diabetes.
- $Y =$ number of children of a randomly sampled person.
- $Y =$ number of under-5 deaths in a particular area and time period.

Discrete random variables

For a discrete random variable Y, $\Pr(Y = y)$ is the probability that the outcome Y takes on the value y.

$\Pr(Y = y) = p(y)$ is often called the probability mass function or probability distribution of Y.

Requirements of a probability distribution to be probabilistically legal:

1. $0 \leq p(y) \leq 1$ for all $y \in \mathcal{Y}$;
2. $\sum_{y \in \mathcal{Y}} p(y) = 1$.

Discrete random variables

We can derive various probabilities from $p(y)$:

$$\Pr(Y \in A) = \sum_{y \in A} p(y)$$

If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} \Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \sum_{y \in A} p(y) + \sum_{y \in B} p(y). \end{aligned}$$

Continuous random variables

If (to a rough approximation) $\mathcal{Y} = \mathbb{R}$, then we cannot define $\Pr(Y \leq 5)$ as equal to $\sum_{y \leq 5} p(y)$ because the sum does not make sense.

Instead, we define a probability density function (pdf) $p(y)$ such that

$$\Pr(Y \in A) = \int_A p(y)\, dy$$

Example:

$$\Pr(Y \leq 5) = \int_{-\infty}^{5} p(y)\, dy.$$

Requirements of a pdf to be probabilistically legal:

1. $p(y) \geq 0$ for all $y \in \mathcal{Y}$;
2. $\int_{\mathbb{R}} p(y)\, dy = 1$.

Continuous random variables

If A and B are disjoint subsets of $\mathcal{Y}$, then

$$\begin{aligned} \Pr(Y \in A \text{ or } Y \in B) \equiv \Pr(Y \in A \cup B) &= \Pr(Y \in A) + \Pr(Y \in B) \\ &= \int_{y \in A} p(y)\, dy + \int_{y \in B} p(y)\, dy. \end{aligned}$$

Continuous random variables

Unlike the discrete case,

- $p(y)$ can be larger than 1;
- $p(y)$ is not "the probability that $Y = y$."

This is a bit weird, because we use pdfs as models for data. The rationale is that all "continuous" measurements are actually examples of discrete random variables (finite number of decimal places).

Suppose we observe $Y = y$:

$$\Pr(Y = y) \overset{\text{Actually}}{=} \Pr(Y \in (y - \epsilon, y + \epsilon)) = \int_{y-\epsilon}^{y+\epsilon} p(y)\, dy,$$

which is a probability.

We approximate these discrete distributions by pdfs.

Regardless, if $p(y_1) > p(y_2)$ we will sometimes informally say that y_1 "has a higher probability" than y_2.

The Bernoulli distribution

Let $\mathcal{Y} = \{0, 1\}$, so the outcome can be 0 or 1.

The outcome Y has a Bernoulli distribution with probability θ if

$$\Pr(Y = y|\theta) = p(y|\theta) = \begin{cases} \theta & \text{if } y = 1 \\ 1 - \theta & \text{if } y = 0 \end{cases}$$

Alternatively, we can write

$$\begin{aligned} \Pr(Y = y|\theta) &= p(y|\theta) \\ &= \theta^y (1 - \theta)^{1-y}. \end{aligned}$$

Conditionally independent binary outcomes

Suppose the prevalence of diabetes in a population is θ.

Let $Y_1, \ldots, Y_n$ indicate the presence of diabetes for n individuals randomly sampled from the population.

The probability of observing the sequence of n is:

$$\begin{aligned} \Pr(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= p(y_1, \ldots, y_n|\theta) \\ &= \theta^{y_1}(1-\theta)^{1-y_1} \times \cdots \times \theta^{y_n}(1-\theta)^{1-y_n} \\ &= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i} \end{aligned}$$

where we can go from lines 1 to 2 by conditional independence.

Note that $p(y_1, \ldots, y_n|\theta)$ depends only on $\sum_{i=1}^n y_i$.

Often, we only record n and the number of events: $y = \sum_{i=1}^n y_i$.

The binomial distribution

What is the probability that y people in a sample of size n will have diabetes?

Consider all n-sequences with y 1's:

$$\begin{aligned} \Pr(Y_1 = 0, Y_2 = 1, Y_3 = 0, \ldots, Y_n = 1|\theta) &= \theta^y(1-\theta)^{n-y} \\ &\vdots \quad\quad \vdots \\ \Pr(Y_1 = 1, Y_2 = 0, Y_3 = 1, \ldots, Y_n = 0|\theta) &= \theta^y(1-\theta)^{n-y} \end{aligned}$$

There are $\binom{n}{y}$ such sequences, so

$$\Pr\left(\sum_{i=1}^{n} Y_i = y|\theta\right) = \binom{n}{y}\theta^y(1-\theta)^{n-y}.$$

The binomial distribution

Let $\mathcal{Y} = \{0, 1, 2, \ldots, n\}$ for some positive integer n. The outcome $Y \in \mathcal{Y}$ has a binomial distribution with probability θ, denoted $Y|\theta \sim \text{Binomial}(n, \theta)$, if

$$\Pr(Y = y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

For example, if $\theta = 0.25$ and $n = 4$, we have 5 possibilities:

$$\begin{aligned}
\Pr(Y = 0|\theta = 0.25) &= \binom{4}{0} (0.25)^0 (0.75)^4 = 0.316 \\
\Pr(Y = 1|\theta = 0.25) &= \binom{4}{1} (0.25)^1 (0.75)^3 = 0.422 \\
\Pr(Y = 2|\theta = 0.25) &= \binom{4}{2} (0.25)^2 (0.75)^2 = 0.211 \\
\Pr(Y = 3|\theta = 0.25) &= \binom{4}{3} (0.25)^3 (0.75)^1 = 0.047 \\
\Pr(Y = 4|\theta = 0.25) &= \binom{4}{4} (0.25)^4 (0.75)^0 = 0.004 .
\end{aligned}$$

The beta posterior

What prior should we choose for θ?

The posterior is the normalized product of the prior times the likelihood:

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

And summarization of the posterior distribution requires integration, e.g., the mean of the posterior is

$$\mathrm{E}[\theta|y] = \int_{\theta} \theta p(\theta|y)\, d\theta.$$

In general, numerical, analytical or simulation techniques are required to carry out Bayesian inference.

We give 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.

The beta posterior

To carry out a Bayesian analysis with a binomial θ, we need a distribution on $[0, 1]$.

A beta prior fulfills these requirements, and has two parameters a and b.

A beta(a, b) prior has mean

$$\mathrm{E}[\theta] = \frac{a}{a+b}$$

and variance

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

Different choices of a and b lead to distributions with different locations and concentration.

The beta posterior

It can be shown (in detail in the Technical Appendix) that if:

- the prior is $\theta \sim \text{beta}(a, b)$; a and b are picked in advance to reflect our beliefs.
- the sampling model is $Y|\theta \sim \text{Binomial}(n, \theta)$
- then the posterior is

$$\theta|y \sim \text{beta}(a + y, b + n - y).$$

The posterior distribution is also beta, but the parameters have been updated to $a + y, b + n - y$.

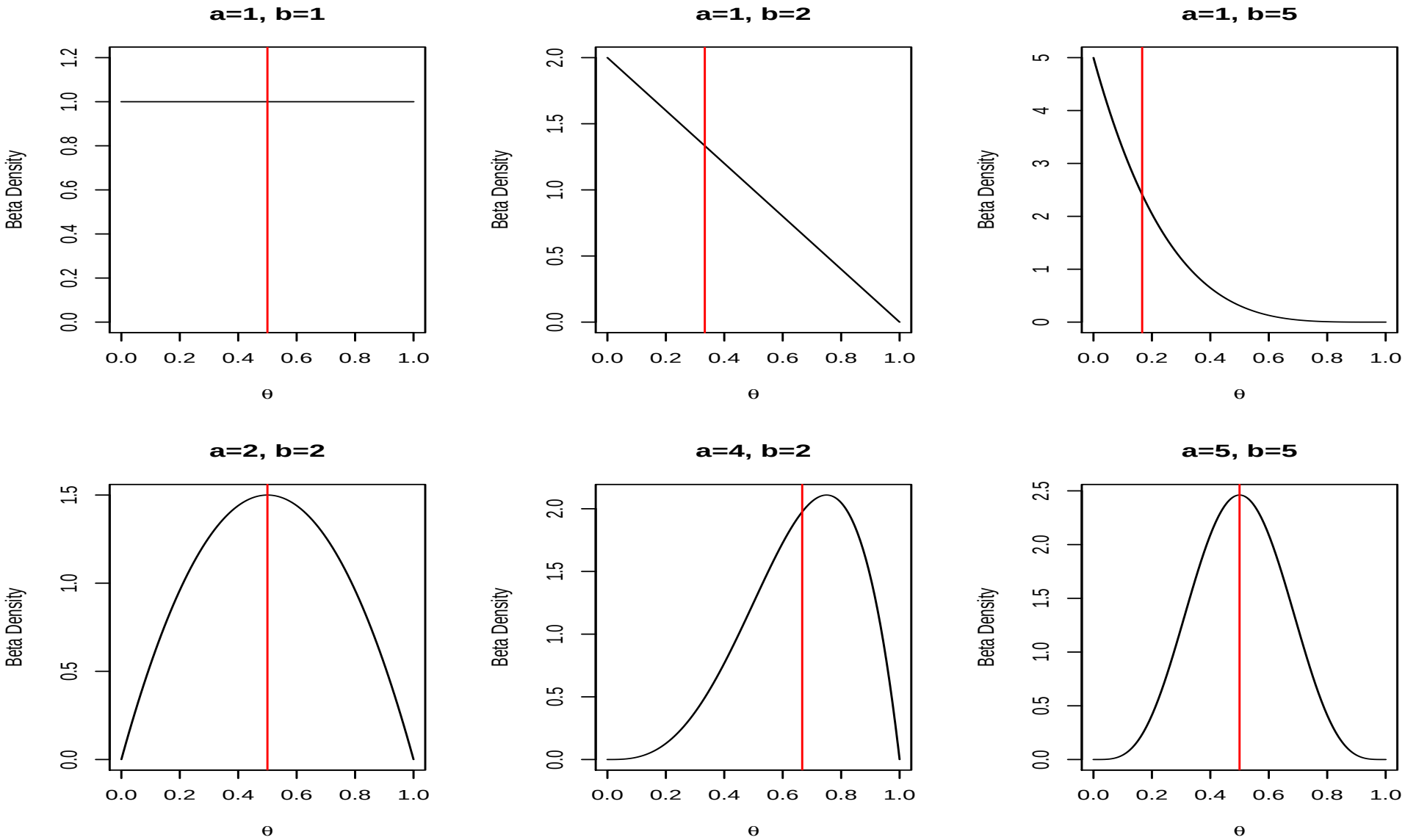


Figure 5: Different beta distributions, beta(a, b), the red lines indicate the means.

The beta posterior

To summarize the posterior distribution for θ, we could use the posterior mean:

$$\begin{aligned}
\mathrm{E}[\theta|y] &= \frac{a+y}{a+b+n} \\
&= \frac{a}{a+b}\left(\frac{a+b}{a+b+n}\right) + \frac{y}{n}\left(\frac{n}{a+b+n}\right) \\
&= \mathrm{E}[\theta]\underbrace{\left(\frac{a+b}{a+b+n}\right)}_{\text{Weight on Prior}} + \bar{y}\underbrace{\left(\frac{n}{a+b+n}\right)}_{\text{Weight on Data}},
\end{aligned}$$

a weighted combination of the prior mean and the sample mean.

The beta posterior

Recall, from earlier, the adjusted Wald interval:

$$\tilde{\theta} \pm 1.96\sqrt{\frac{\tilde{\theta}(1-\tilde{\theta})}{n}},$$

where

$$\tilde{\theta} = \frac{1}{2}\frac{4}{4+n} + \bar{y}\frac{n}{4+n}.$$

$\tilde{\theta}$ is equal to the posterior mean when we have a beta(2,2) prior (which has mean 1/2).

Beta Prior, Likelihood and Posterior

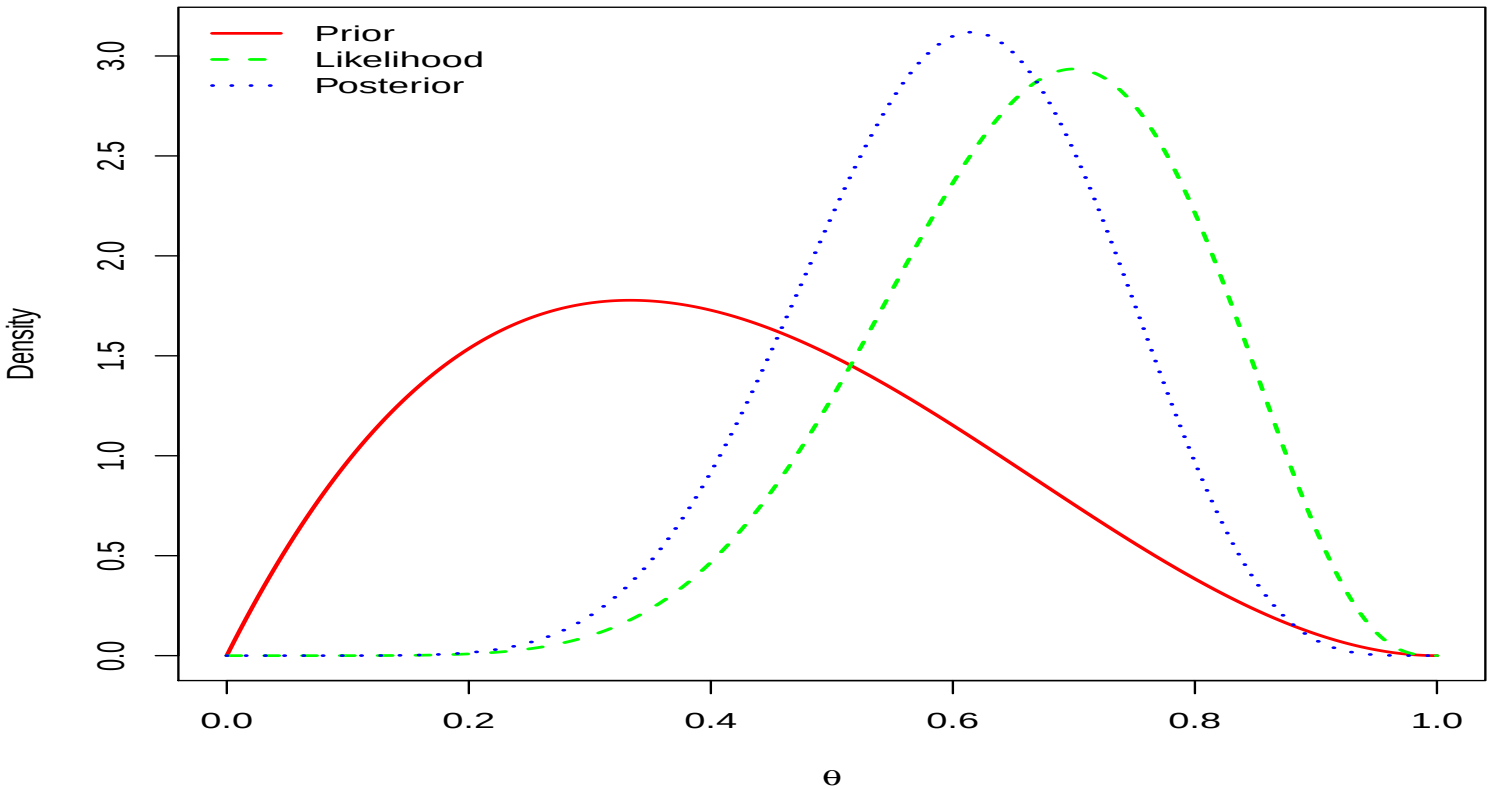


Figure 6: The prior is beta(2,3) the likelihood is binomial with $n = 10, y = 7$, and so the posterior is beta(7+2,3+3).

The normal distribution

Let the sample space be $\mathcal{Y} = (-\infty, \infty)$ and assume the outcome $Y \in \mathcal{Y}$ has a normal distribution with mean θ and variance σ^2, denoted

$$y|\theta \sim \mathrm{N}(\theta, \sigma),$$

if

$$p(y|\theta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right\}.$$

The normal posterior distribution with known variance σ^2

For a sample $Y_1, \ldots, Y_n$ from a normal distribution, the sampling model (likelihood) is

$$Y_1, \ldots, Y_n | \theta \sim \mathrm{N}(\theta, \sigma^2).$$

The MLE is

$$\widehat{\mu} = \overline{y},$$

and the variance of this estimator is

$$\mathrm{var}(\widehat{\mu}) = \frac{\sigma^2}{n}.$$

If:

- the prior on the mean is $\theta \sim \mathrm{N}(\mu_0, \tau_0^2)$ and
- the sampling model (likelihood) is again $Y_1, \ldots, Y_n | \theta \sim \mathrm{N}(\theta, \sigma^2)$.

Then, the posterior is also normal:

$$\theta | y_1, \ldots, y_n \sim \mathrm{N}(\mu_n, \tau_n^2).$$

The normal posterior distribution

The posterior mean is,

$$\begin{aligned} \mathrm{E}[\theta|y_1,\ldots,y_n] &= \mu_n \\ &= \mu_0(1-w)+\bar{y}w \end{aligned}$$

where the weight on the data is

$$w = \left(\frac{\tau_0^2}{\tau_0^2+\sigma^2/n}\right).$$

So the posterior mean is a weighted combination of the prior mean and the sample mean.

The posterior variance is,

$$\begin{aligned} \mathrm{var}(\theta|y_1,\ldots,y_n) &= \tau_n^2 \\ &= w\frac{\sigma^2}{n}\left(\leq \underbrace{\frac{\sigma^2}{n}}_{\text{Variance of MLE}} = \mathrm{var}(\widehat{\mu})\right) \end{aligned}$$

The normal posterior distribution

We see that the precisions (inverse variances) are additive:

$$\underbrace{1/\tau_n^2}_{\text{Posterior Precision}} = \underbrace{1/\tau_0^2}_{\text{Prior Precision}} + \underbrace{n/\sigma^2}_{\text{Data Precision}} .$$

so precision (or information) is additive.

We will consider the normal model for continuous responses for an area; in a generic area let y_k be the weight of sampled person k.

Then a starting model (the likelihood) is

$$y_k = \mu + \epsilon_k,$$

with

$$\epsilon_k \sim \mathrm{N}(0, \sigma_\epsilon^2),$$

for $k = 1, \ldots, n$.

A Bayesian analysis would put priors on μ and σ_ϵ^2.

Simple Normal Example

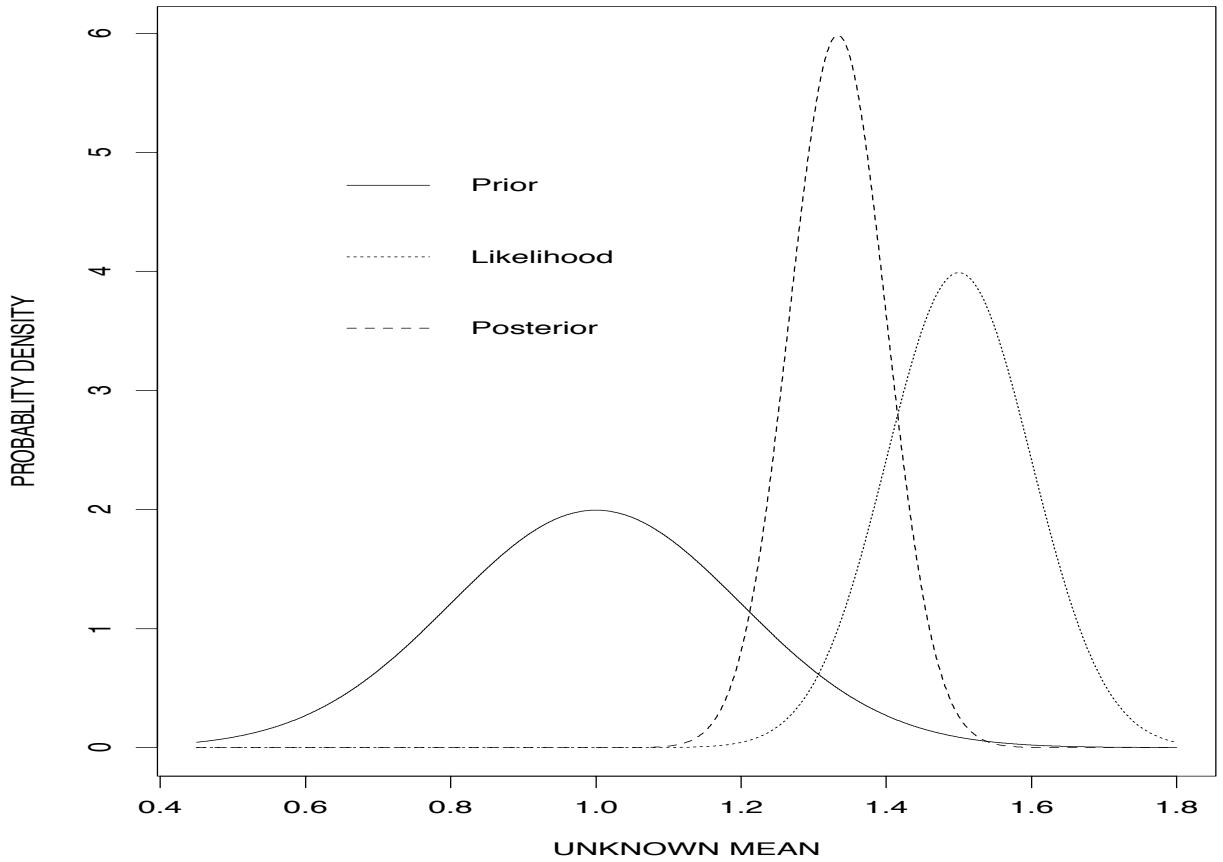


Figure 7: Normal likelihood ($\bar{y}$=1.5,n=10,σ=1), normal prior (m=1, k=5) and the resultant normal posterior.

Describing posterior location

When carrying out frequentist inference for a parameter θ, we may report the MLE as point estimate; in a Bayes analysis there are a number of ways of summarizing the posterior with a single number.

The posterior mean expectation of an unknown quantity θ is given by

$$\mathrm{E}[\theta|y] = \int_{\theta \in \Theta} \theta p(\theta|y)\, d\theta.$$

The mean is the center of mass of the distribution.

However, it is not in general equal to either of

- the mode: "the most probable value of θ," or
- the median: "the value of θ in the middle of the distribution."

For skewed distributions the mean can be far from a "typical" sample value.

If in doubt, use the posterior median!

Describing posterior uncertainty

In frequentist inference we might report a confidence interval.

What about expressing uncertainty? Posterior credible intervals!

For example, a 90% interval (θ_L, θ_U) can be reported by finding values

$$\int_{\theta_L}^{\infty} p(\theta|y)\, d\theta$$

$$\int_{-\infty}^{\theta_L} p(\theta|y)\, d\theta$$

The Bayesian analog of the standard error is the posterior standard deviation.

Summary

We have reviewed basic probability theory and began the discussion of how Bayes theorem can be used for statistical inference.

Probability distributions encapsulate information:

- $p(\theta)$ describes prior information
- $p(y|\theta)$ describes information about y for each θ
- $p(\theta|y)$ describes posterior information

Posterior distributions can be calculated via Bayes theorem

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

Conjugate analyses are computationally convenient but rarely available in practice.

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

Overview of Bayesian Inference

Simply put, to carry out a Bayesian analysis one must specify a likelihood (probability distribution for the data) and a prior (beliefs about the parameters of the model).

And then do some computation... and interpretation...

The approach is therefore model-based, in contrast to approaches in which only the mean and the variance of the data are specified (e.g., weighted least squares, quasi-likelihood).

Conclusions

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 the `INLA` method for carrying out Bayesian inference.
- 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.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley, New York.

Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York.

Savage, S. A., Gerstenblith, M. R., Goldstein, A., Mirabello, L., Fargnoli, M. C., Peris, K., and Landi, M. T. (2008). Nucleotide diversity and population differentiation of the melanocortin 1 receptor gene, MC1R. *BMC Genetics*, **9**, 31.

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

Technical Appendix: Details of Calculations for the Binomial Model

Elements of Bayes Theorem for a Binomial Model

We assume independent responses with a common "success" probability θ.

In this case, the contribution of the data is through the binomial probability distribution:

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

and tells us the probability of seeing $Y = y$, $y = 0, 1, \ldots, n$, given the probability θ.

For fixed y, we may view (1) as a function of θ – this is the likelihood function.

The maximum likelihood estimate (MLE) is that value

$$\widehat{\theta} = y/n$$

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

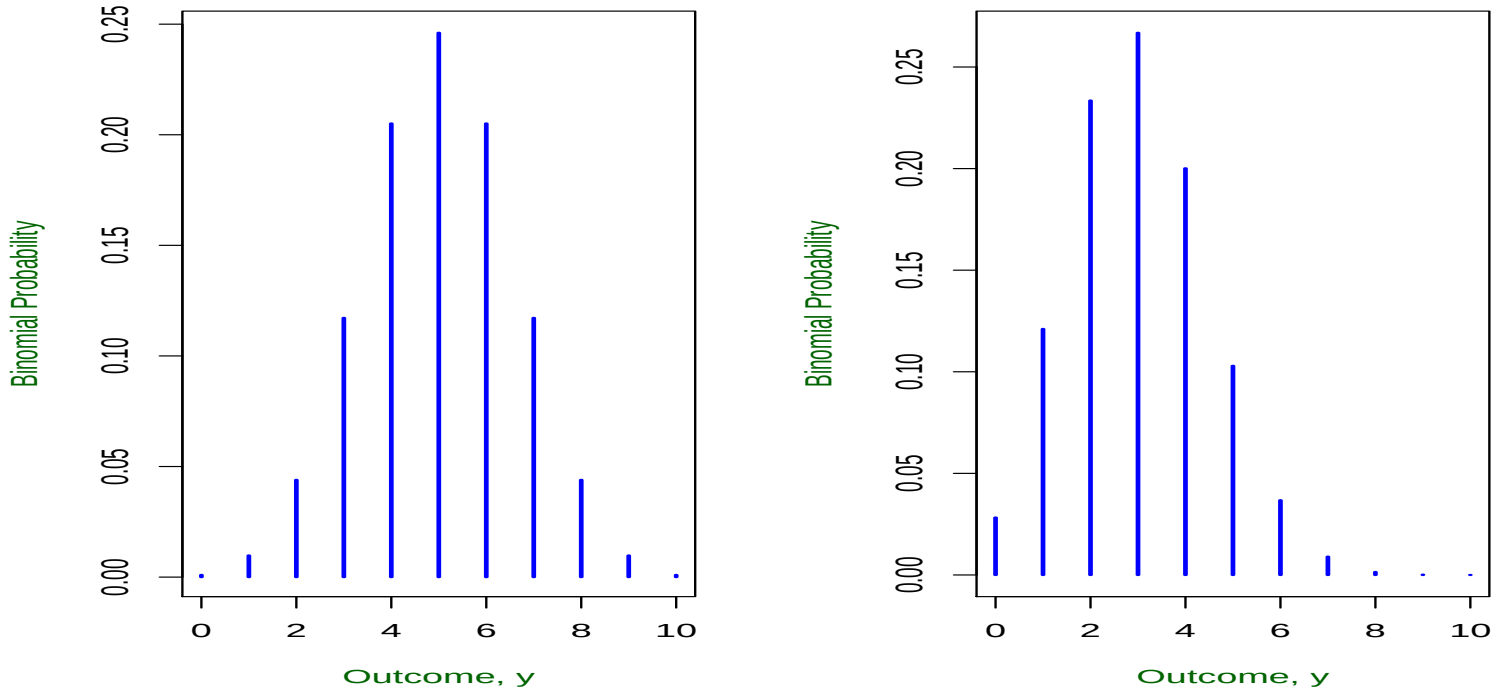
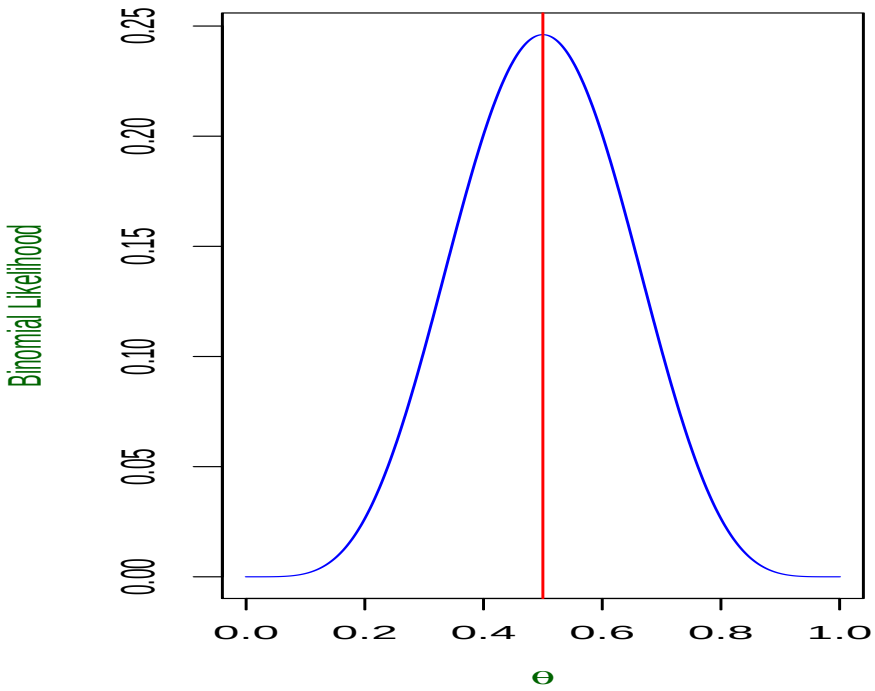


Figure 8: Binomial distributions for two values of θ with $n = 10$.



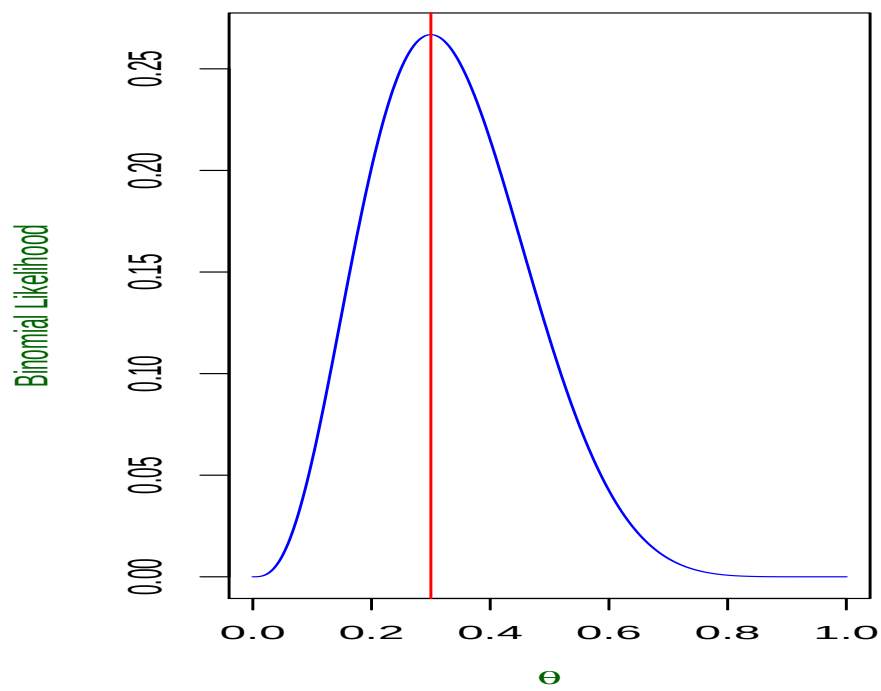


Figure 9: Binomial likelihoods for values of $y = 5$ (left) and $y = 10$ (right), with $n = 10$. The MLEs are indicated in red.

The Beta Distribution as a Prior Choice for Binomial θ

- Bayes theorem 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[3].
- The distribution is valid[4] for $a > 0, b > 0$.

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

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

The Beta Distribution as a Prior Choice for Binomial θ

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/or *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 such as `R` using the function `qbeta(p,a,b)`.

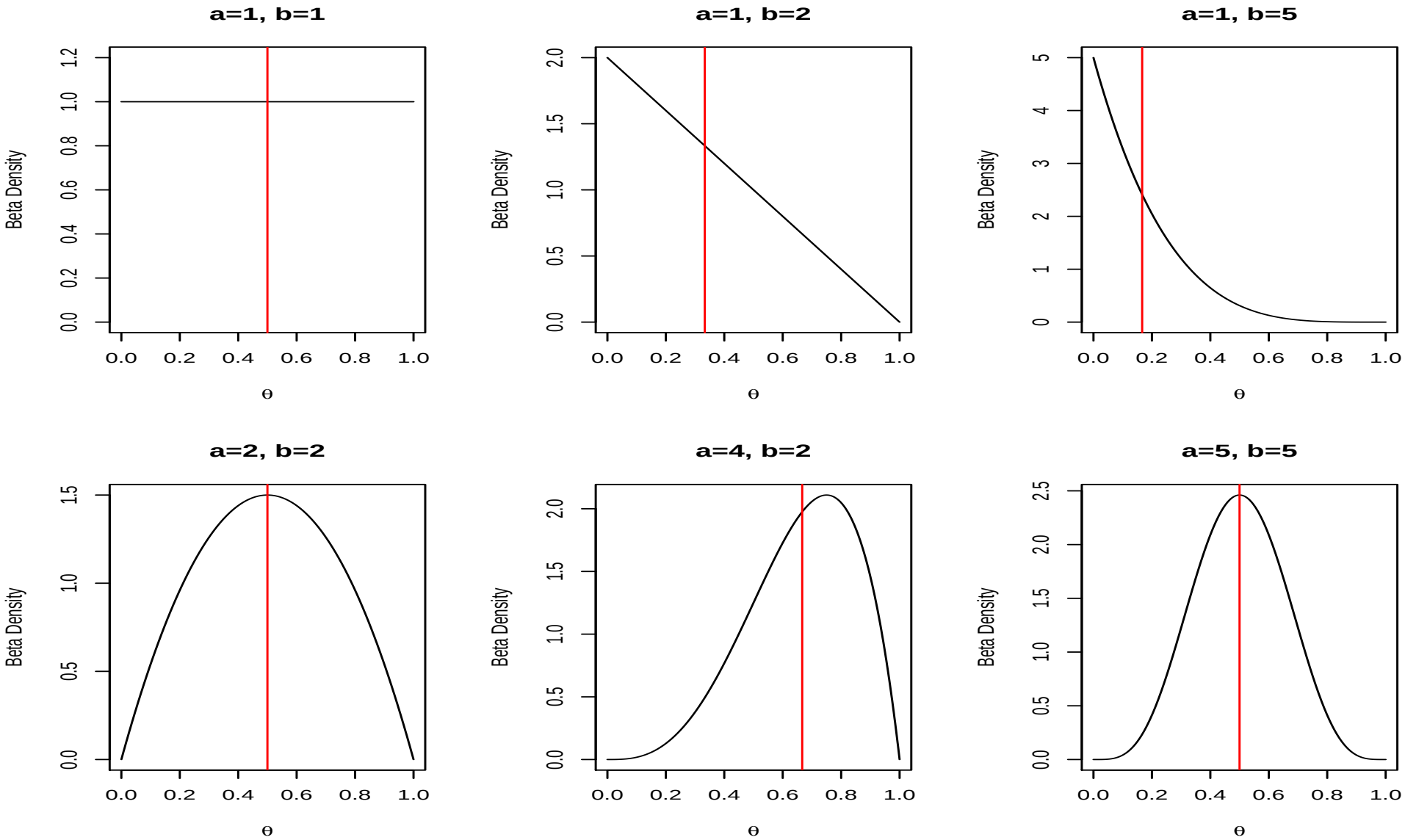


Figure 10: Beta distributions, beta(a, b), the red lines indicate the means.

Samples to Summarize Beta Distributions

Probability distributions can be investigated by generating samples and then examining histograms, moments and quantiles.

In Figure 11 we show histograms of beta distributions for different choices of a and b.

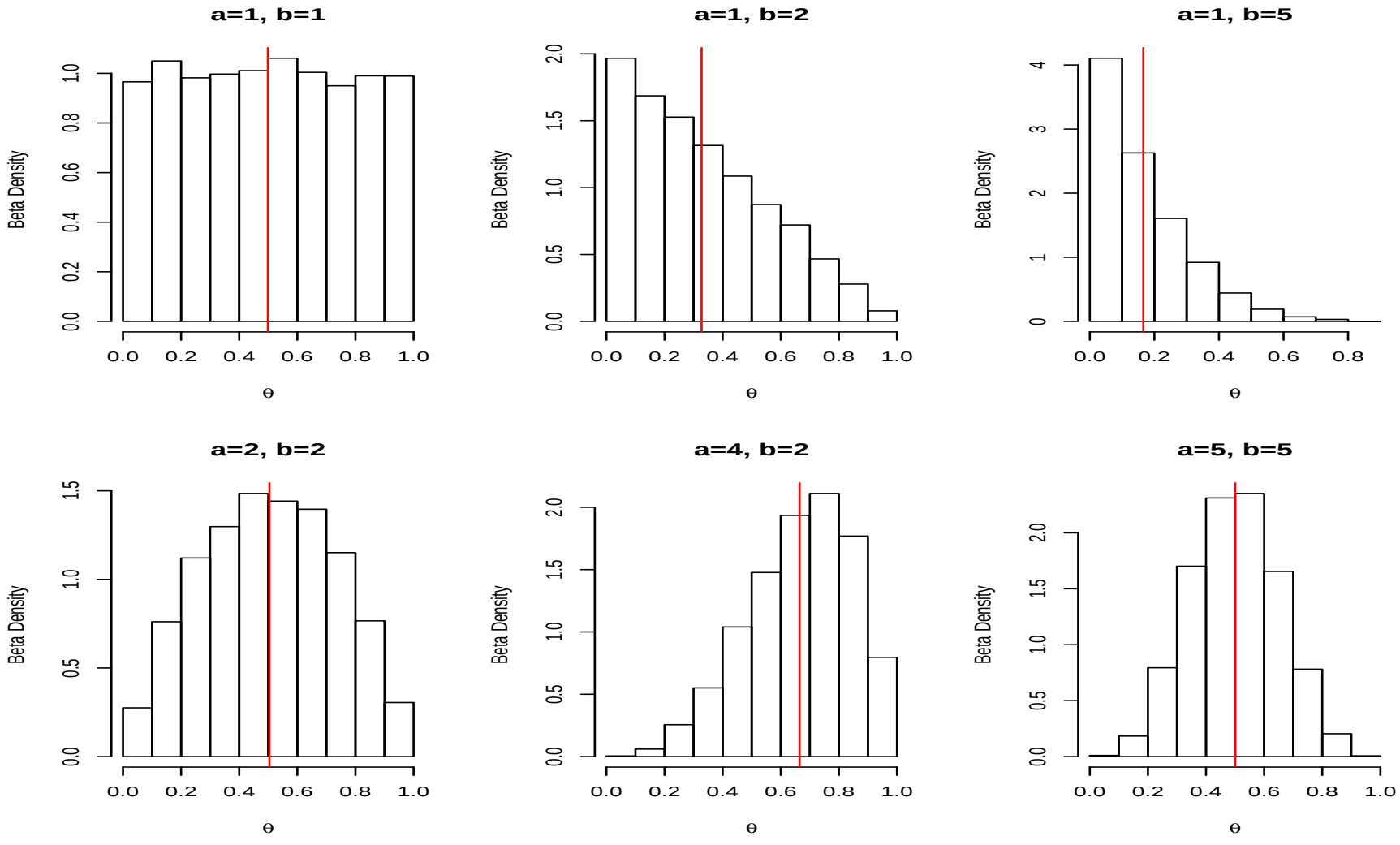


Figure 11: Random samples from beta distributions; sample means as red lines.

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.
- The histogram representation of the prior on the odds $\theta/(1-\theta)$ when θ is beta(10,10).

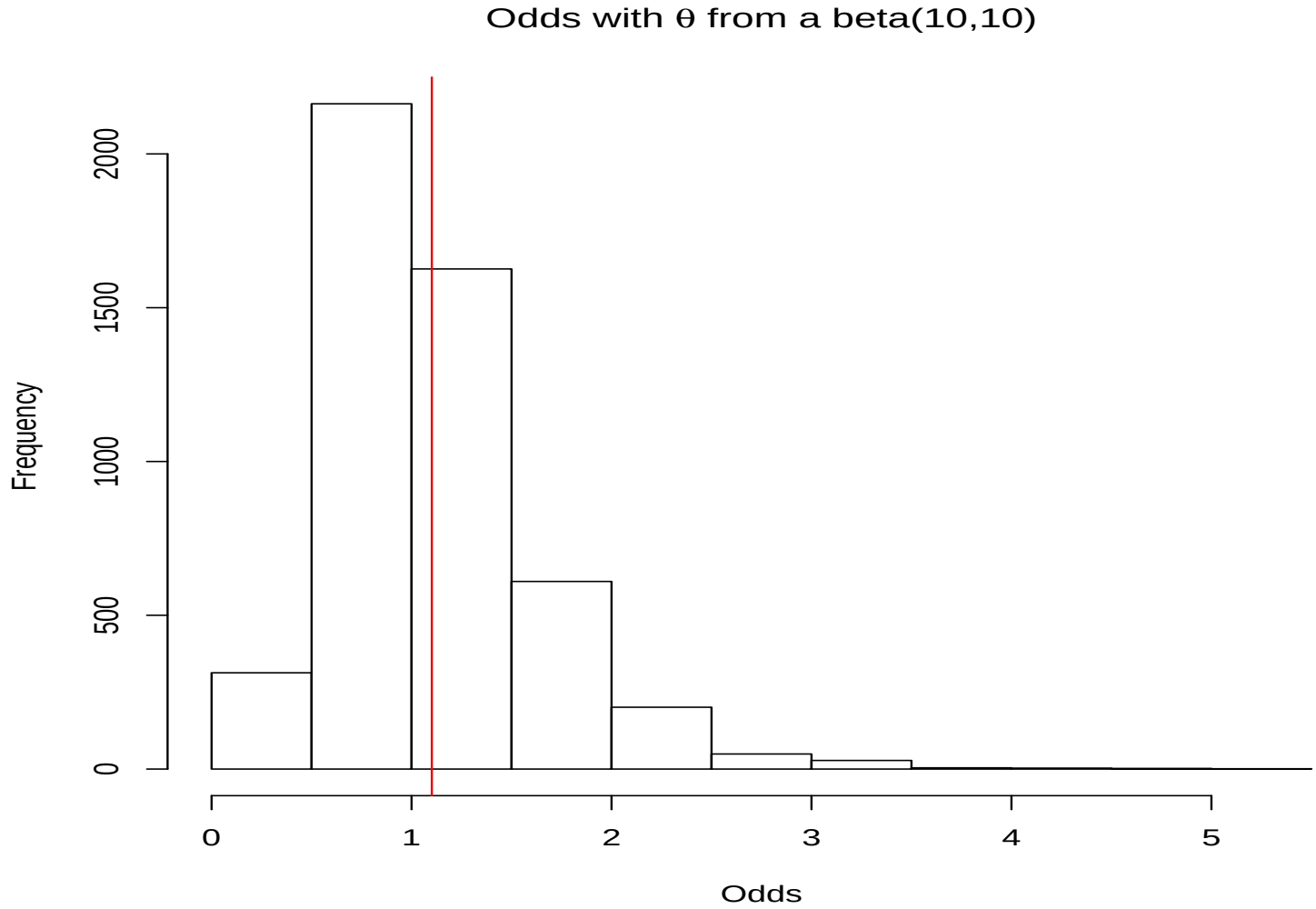


Figure 12: Samples from the prior on the odds $\theta/(1-\theta)$ with $\theta \sim$ beta$(10, 10)$, the red line indicates the sample mean.

Issues with Uniformity

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:

- 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.
- 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 Figure 13.
- Not being uniform on all scales is not necessarily a problem, and is correct probabilistically, but one should be aware of this characteristic.

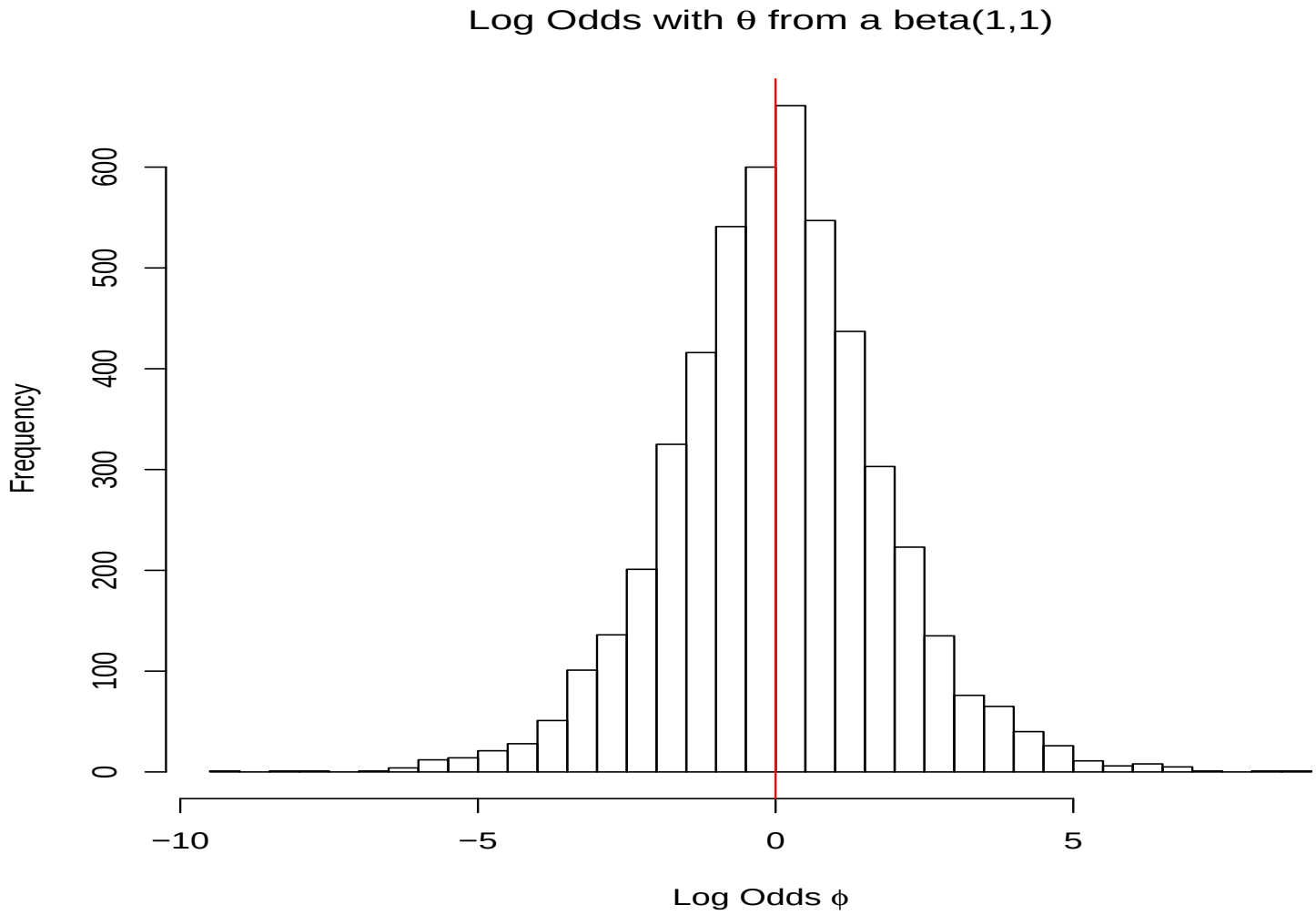


Figure 13: Samples from the prior on the odds $\phi = \log[\theta/(1-\theta)]$ with $\theta \sim$ beta$(1, 1)$, the red line indicates the sample mean.

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

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

- We recognize this as the important part of a beta$(y+a, n-y+b)$ distribution.
- We know what the normalizing constant must be, because we have a distribution which must integrate to 1.

Posterior Derivation: The Long (Unnecessary) Way

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

$$\begin{aligned} p(\theta|y) &= \frac{\Pr(y|\theta) \times p(\theta)}{\Pr(y)} \\ &= \frac{1}{\Pr(y)} \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1} (1-\theta)^{b-1}. \quad (2) \end{aligned}$$

- 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:

$$\Pr(y) = \binom{n}{y} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(y+a)\Gamma(n-y+b)}{\Gamma(n+a+b)}$$

- This is a probability distribution, i.e. $\sum_{y=0}^{n} \Pr(y) = 1$ with $\Pr(y) > 0$.
- For 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 (2) 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 beta$(y+a, n-y+b)$.

The Posterior Mean: A Summary of the Posterior

- Recall the mean of a beta(a, b) is $a/(a+b)$.
- The posterior mean of a beta$(y+a, n-y+b)$ is therefore

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

- 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

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

The Posterior Mode

- First, note that the mode of a beta(a, b) is

$$\text{mode}(\theta) = \frac{a-1}{a+b-2}.$$

- As with the posterior mean, the posterior mode takes a weighted form:

$$\begin{aligned}
\text{mode}(\theta|y) &= \frac{y+a-1}{n+a+b-2} \\
&= \frac{y}{n} \times \frac{n}{n+a+b-2} + \frac{a-1}{a+b-2} \times \frac{a+b-2}{n+a+b-2} \\
&= \text{MLE} \times \text{W}^\star + \text{Prior Mode} \times (1\text{-W}^\star).
\end{aligned}$$

- The weight W* is

$$\text{W}^\star = \frac{n}{n+a+b-2}.$$

- Notice that the uniform prior $a = b = 1$ gives a posterior mode of

$$\text{mode}(\theta|y) = \frac{y}{n},$$

the MLE. Which makes sense, right?

Other 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 running example, a 90% posterior credible interval (θ_L, θ_U) results from the points

$$0.05 = \int_0^{\theta_L} p(\theta|y)\, d\theta$$

$$0.95 = \int_0^{\theta_U} p(\theta|y)\, d\theta$$

- 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.

Prior Sensitivity

- For small datasets in particular it is a good idea to examine the sensitivity of inference to the prior choice, particularly for those parameters for which there is little information in the data.
- An obvious way to determine the latter is to compare the prior with the posterior, but experience often aids the process.
- Sometimes one may specify a prior that reduces the impact of the prior.
- In some situations, priors can be found that produce point and interval estimates that mimic a standard non-Bayesian analysis, i.e. have good frequentist properties.
- Such priors provide a baseline to compare analyses with more substantive priors.
- Other names for such priors are objective, reference and non-subjective.
- We now describe another approach to specification, via subjective priors.

Choosing a Prior, Approach One

- To select a beta, we need to specify two quantities, a and b.
- The posterior mean is

$$\mathrm{E}[\theta|y] = \frac{y+a}{n+a+b}.$$

- Viewing the denominator as a sample size suggests a method for choosing a and b within the prior.
- We need to specify two numbers, but rather than a and b, which are difficult to interpret, we may specify the mean $m_{\text{prior}} = a/(a+b)$ and the prior sample size $n_{\text{prior}} = a+b$
- We then solve for a and b via

$$\begin{aligned} a &= n_{\text{prior}} \times m_{\text{prior}} \\ b &= n_{\text{prior}} \times (1 - m_{\text{prior}}). \end{aligned}$$

- Intuition: a is like a prior number of successes and b like the prior number of failures.

An Example

- Suppose we set $n_{\text{prior}} = 5$ and $m_{\text{prior}} = \frac{2}{5}$.
- It is as if we saw 2 successes out of 5.
- Suppose we obtain data with $N = 10$ and $\frac{y}{n} = \frac{7}{10}$.
- Hence $W = 10/(10+5)$ and

$$\begin{aligned} \mathrm{E}[\theta|y] &= \frac{7}{10} \times \frac{10}{10+5} + \frac{2}{5} \times \frac{5}{10+5} \\ &= \frac{9}{15} = \frac{3}{5}. \end{aligned}$$

- Solving:

$$\begin{aligned} a &= n_{\text{prior}} \times m_{\text{prior}} = 5 \times \frac{2}{5} = 2 \\ b &= n_{\text{prior}} \times (1 - m_{\text{prior}}) = 5 \times \frac{3}{5} = 3 \end{aligned}$$

- This gives a beta$(y+a, n-y+b)$ = beta$(7+2, 3+3)$ posterior.

Beta Prior, Likelihood and Posterior

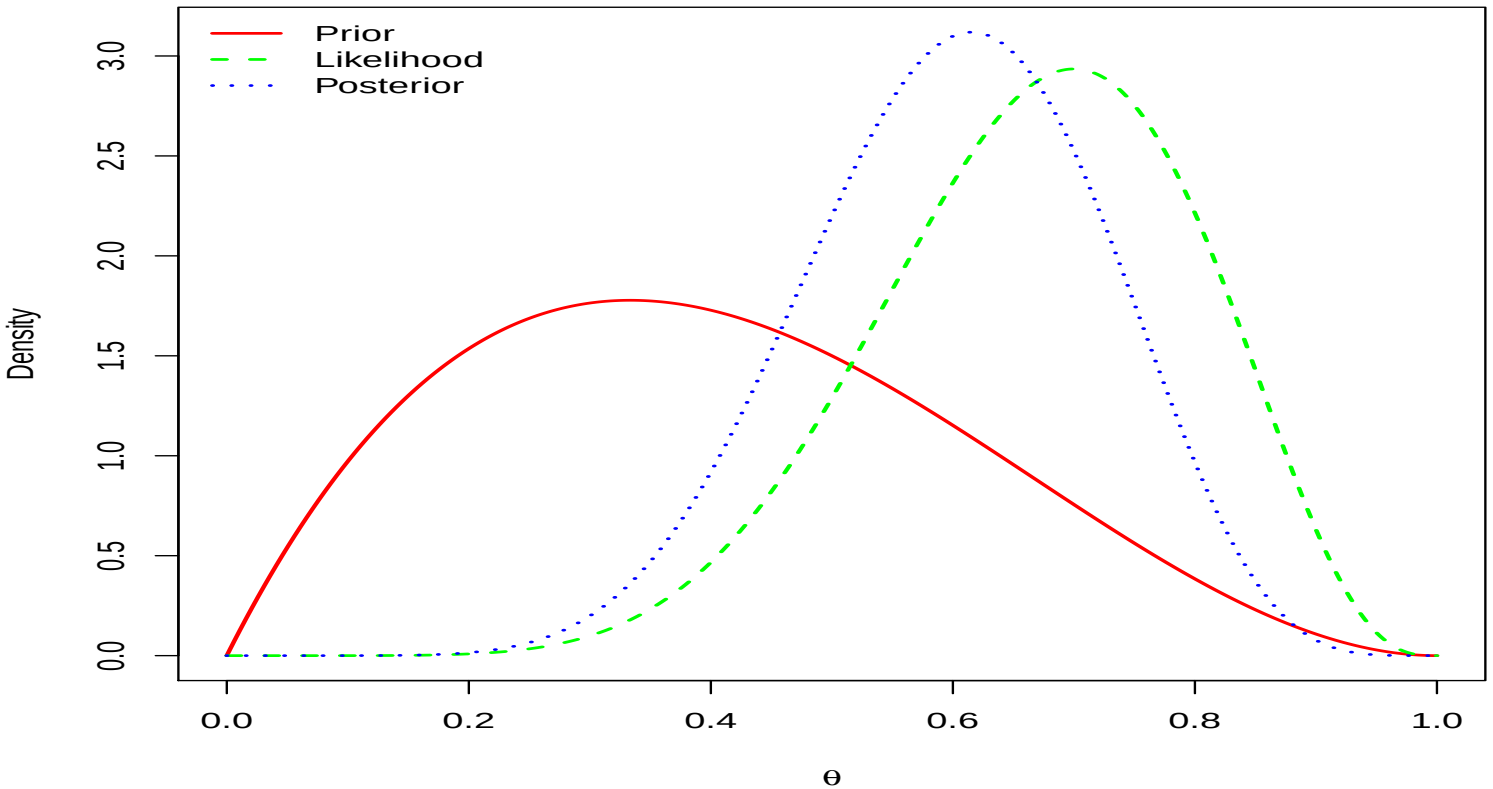


Figure 14: The prior is beta(2,3) the likelihood is proportional to a beta(7,3) and the posterior is beta(7+2,3+3).

Choosing a Prior, Approach Two

- An alternative convenient way of choosing a and b is by specifying two quantiles for θ with associated (prior) probabilities.
- For example, we may wish $\Pr(\theta < 0.1) = 0.05$ and $\Pr(\theta > 0.6) = 0.05$.
- The values of a and b may be found numerically.
- For example, we may solve

$$[p_1 - \Pr(\theta < q_1 | a, b)]^2 + [p_2 - \Pr(\theta < q_2 | a, b)]^2 = 0 \qquad (3)$$

for a, b.

Beta Prior Choice via Quantile Specification

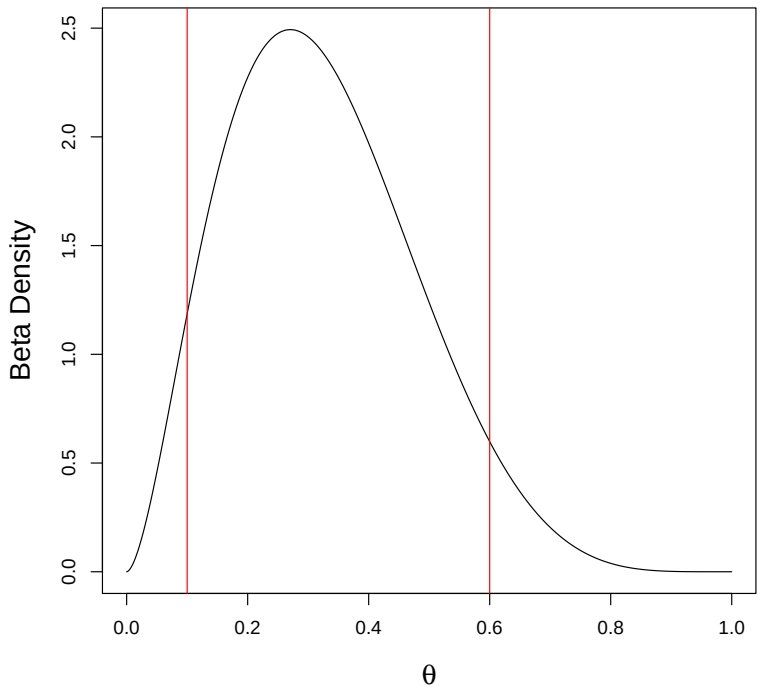


Figure 15: beta(2.73,5.67) prior with 5% and 95% quantiles highlighted.

Bayesian Sequential Updating

- We show how probabilistic beliefs are updated as we receive more data.
- Suppose the data arrives sequentially via two experiments:
  1. Experiment 1: (y_1, n_1).
  2. Experiment 2: (y_2, n_2).
- Prior 1: $\theta \sim \text{beta}(a, b)$.
- Likelihood 1: $y_1|\theta \sim \text{binomial}(n_1, \theta)$.
- Posterior 1: $\theta|y_1 \sim \text{beta}(a + y_1, b + n_1 - y_1)$.
- This posterior forms the prior for experiment 2.
- Prior 2: $\theta \sim \text{beta}(a^\star, b^\star)$ where $a^\star = a + y_1$, $b^\star = b + n_1 - y_1$.
- Likelihood 2: $y_2|\theta \sim \text{binomial}(n_2, \theta)$.
- Posterior 2: $\theta|y_1, y_2 \sim \text{beta}(a^\star + y_2, b^\star + n_2 - y_2)$.
- Substituting for $a^\star, b^\star$:

$$\theta|y_1, y_2 \sim \text{beta}(a + y_1 + y_2, b + n_1 - y_1 + n_2 - y_2).$$

Bayesian Sequential Updating

- Schematically:

  $$(a, b) \to (a + y_1, b + n_1 - y_1) \to (a + y_1 + y_2, b + n_1 - y_1 + n_2 - y_2)$$

- Suppose we obtain the data in one go as $y^\star = y_1 + y_2$ successes from $n^\star = n_1 + n_2$ trials.
- The posterior is

  $$\theta | y^\star \sim \text{beta}(a + y^\star, b + n^\star - y^\star),$$

  which is the same as when we receive in two separate instances.

Predictive Distribution

- Suppose we see y successes out of N trials, and now wish to obtain a predictive distribution for a future experiment with M trials.
- Let $Z = 0, 1, \ldots, M$ be the number of successes.
- Predictive distribution:

$$\begin{aligned}
\Pr(z|y) &= \int_0^1 p(z, \theta|y) d\theta \\
&= \int_0^1 \Pr(z|\theta, y) p(\theta|y) d\theta \\
&= \int_0^1 \underbrace{\Pr(z|\theta)}_{\text{binomial}} \times \underbrace{p(\theta|y)}_{\text{posterior}} d\theta
\end{aligned}$$

where we move between lines 2 and 3 because z is conditionally independent of y given θ.

Predictive Distribution

Continuing with the calculation:

$$
\begin{aligned}
\Pr(z|y) &= \int_0^1 \Pr(z|\theta) \times p(\theta|y) d\theta \\
&= \int_0^1 \binom{M}{z} \theta^z (1-\theta)^{M-z} \\
&\times \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)} \theta^{y+a-1} (1-\theta)^{n-y+b-1} d\theta \\
&= \binom{M}{z} \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)} \int_0^1 \theta^{y+a+z-1} (1-\theta)^{n-y+b+M-z-1} d\theta \\
&= \binom{M}{z} \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)} \frac{\Gamma(a+y+z)\Gamma(b+n-y+M-z)}{\Gamma(a+b+n+M)}
\end{aligned}
$$

for $z = 0, 1, \ldots, M$.

A likelihood approach would take the predictive distribution as binomial$(M, \widehat{\theta})$ with $\widehat{\theta} = y/n$: this does not account for estimation uncertainty.

In general, we have sampling uncertainty (which we can't get away from) and estimation uncertainty.

Predictive Distribution

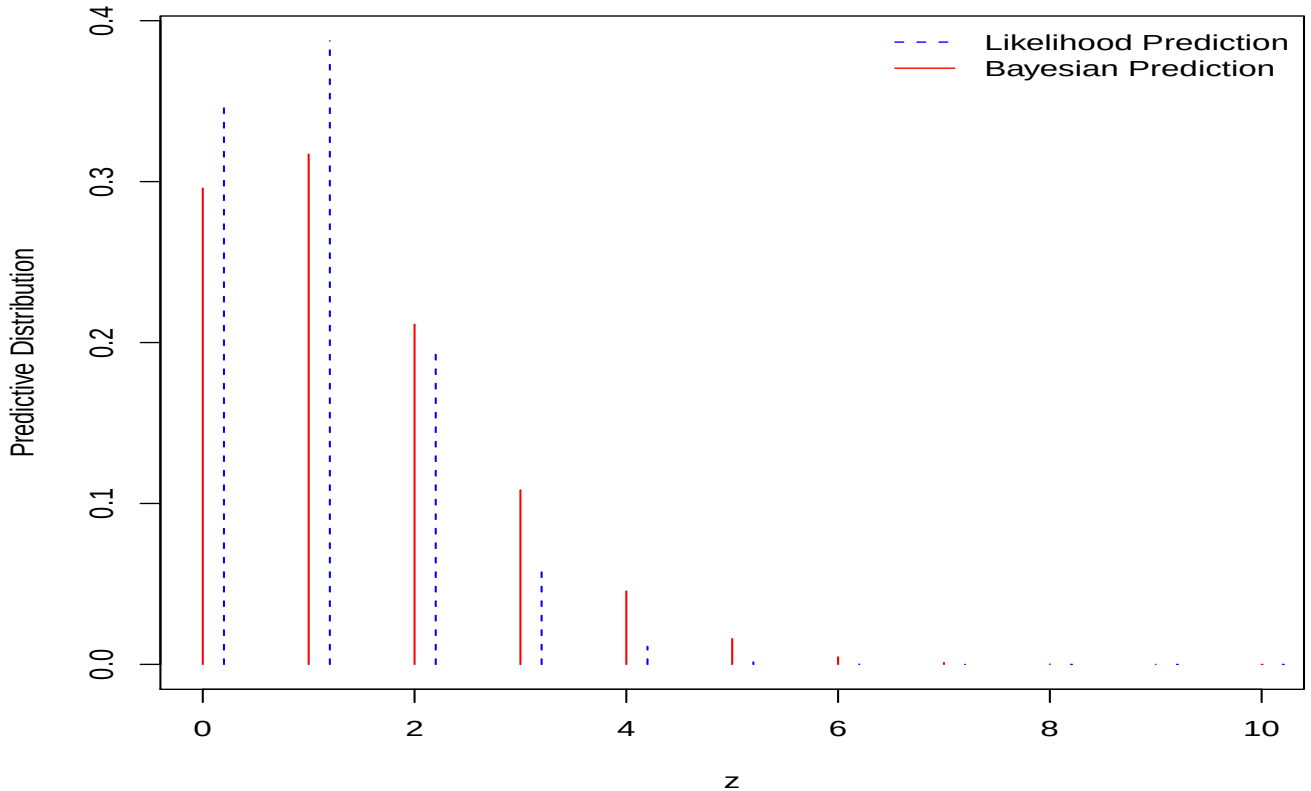


Figure 16: Likelihood and Bayesian predictive distribution of seeing $z = 0, 1, \ldots, M = 10$ successes, after observing $y = 2$ out of $n = 20$ successes (with $a = b = 1$).

Predictive Distribution

The posterior and sampling distributions won't usually combine so conveniently.

In general, we may form a Monte Carlo estimate of the predictive distribution:

$$\begin{aligned} p(z|y) &= \int p(z|\theta)p(\theta|y)d\theta \\ &= \mathrm{E}_{\theta|y}[p(z|\theta)] \\ &\approx \frac{1}{S}\sum_{s=1}^{S} p(z|\theta^{(s)}) \end{aligned}$$

where $\theta^{(s)} \sim p(\theta|y)$, $s = 1, \ldots, S$, is a sample from the posterior.

This provides an estimate of the predictive distribution at the point z.

Predictive Distribution

Alternatively, we may sample from $p(z|\theta^{(s)})$ a large number of times to reconstruct the predictive distribution:

$$\begin{aligned} \theta^{(s)}|y &\sim p(\theta|y),\ s = 1, \ldots, S && \text{Sample from posterior} \\ z^{(s)}|\theta^{(s)} &\sim p(z|\theta^{(s)}),\ s = 1, \ldots, S && \text{Sample from predictive} \end{aligned}$$

To give a sample $z^{(s)}$ from the posterior, this is illustrated in Figure 17.

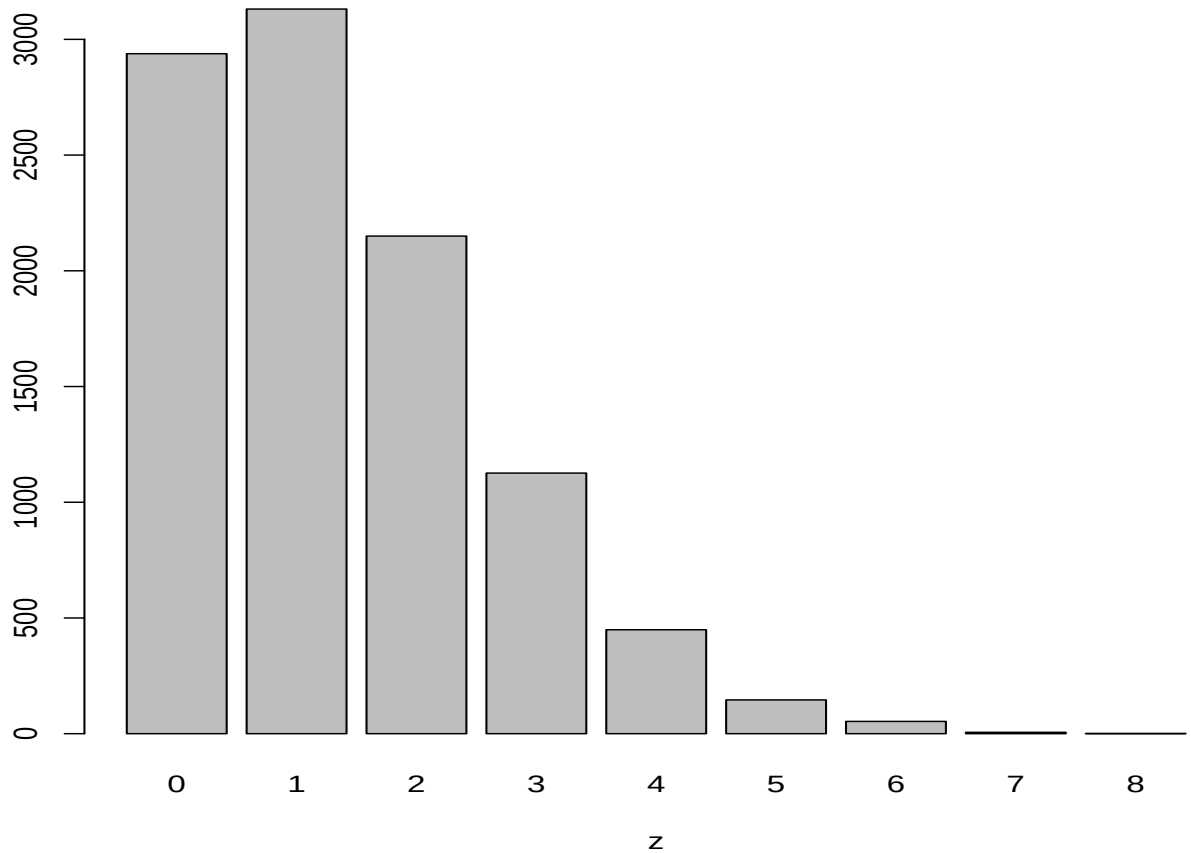


Figure 17: Sampling version of prediction in Figure 16, based on $S = 10,000$ samples.

Difference in Binomial Proportions

- It is straightforward to extend the methods presented for a single binomial sample to a pair of samples.
- Suppose we carry out two binomial experiments:

$$\begin{aligned} Y_1|\theta_1 &\sim \text{binomial}(n_1, \theta_1) \quad \text{for sample 1} \\ Y_2|\theta_2 &\sim \text{binomial}(n_2, \theta_2) \quad \text{for sample 2} \end{aligned}$$

- Interest focuses on $\theta_1 - \theta_2$, and often in examing the possibitlity that $\theta_1 = \theta_2$.
- With a sampling-based methodology, and independent beta priors on θ_1 and θ_2, it is straightforward to examine the posterior $p(\theta_1 - \theta_1|y_1, y_2)$.

Difference in Binomial Proportions

- Savage *et al.* (2008) give data on allele frequencies within a gene that has been linked with skin cancer.
- It is interest to examine differences in allele frequencies between populations.
- We examine one SNP and extract data on Northern European (NE) and United States (US) populations.
- Let θ_1 and θ_2 be the allele frequencies in the NE and US population from which the samples were drawn, respectively.
- The allele frequencies were 10.69% and 13.21% with sample sizes of 650 and 265, in the NE and US samples, respectively.
- We assume independent beta(1,1) priors on each of θ_1 and θ_2.
- The posterior probability that $\theta_1 - \theta_2$ is greater than 0 is 0.12 (computed as the proportion of the samples $\theta_1^{(s)} - \theta_2^{(s)}$ that are greater than 0), so there is little evidence of a difference in allele frequencies between the NE and US samples.

Binomial Two Sample Example

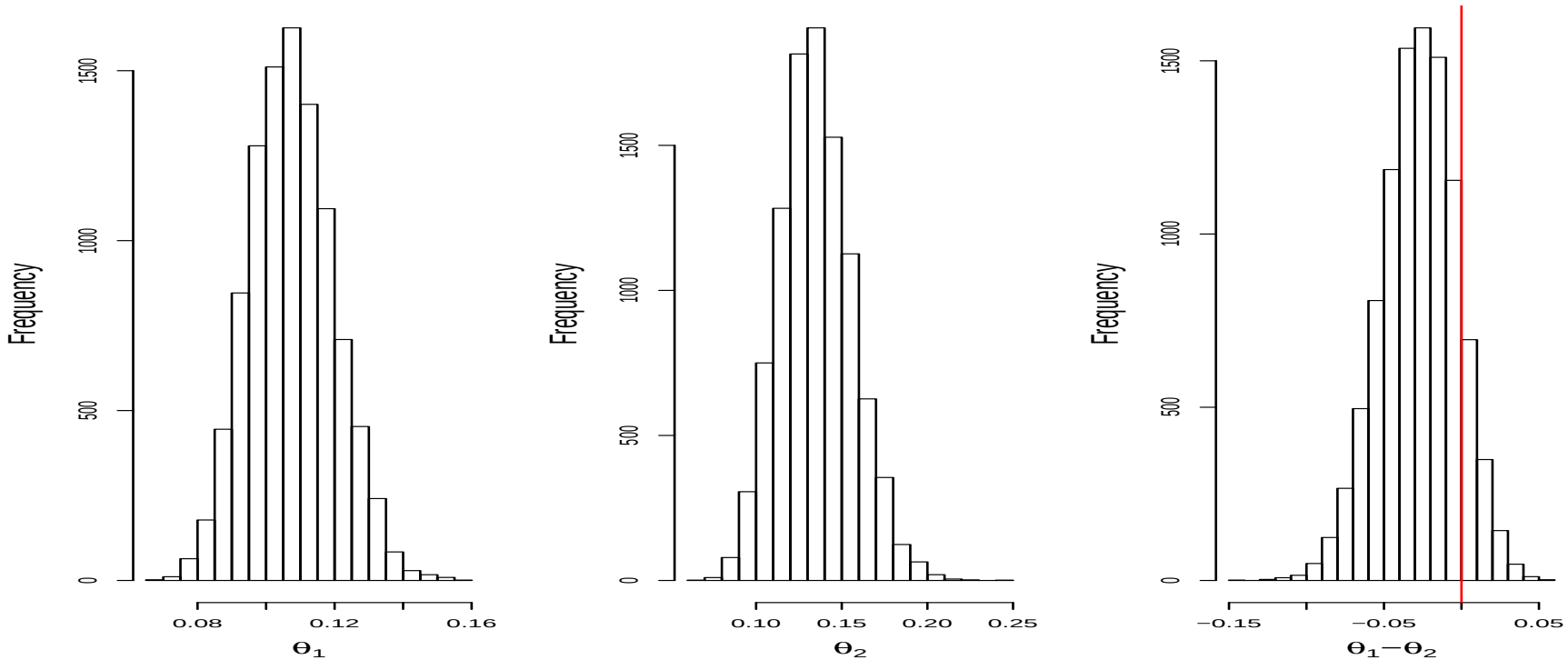


Figure 18: Histogram representations of $p(\theta_1|y_1)$, $p(\theta_2|y_2)$ and $p(\theta_1 - \theta_2|y_1, y_2)$. The red line in the right plot is at the reference point of zero.