2019 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

Bayesian Learning

Introduction

- In this lecture we will first consider generic Bayesian learning.
- Background reading: Chapters 1 and 2 of Hoff (2009).
- A brief review of **probability theory** will be given.
- **Conjugate priors** will be introduced.
- In this lecture, I will state some results, and then these will be derived in later lectures.
- This lecture is the dryest!
- Sorry.... but we need to learn **THE RULES**.

Introduction

- In the next lecture, the binomial model will be studied in detail – motivating data is from a so-called allele specific expression (ASE) experiment.
- Think of N outcomes, each of which can be 0/1, with Y the total number of 1s observed – the unknown parameter we want to learn about is the probability of a 1, denoted θ.
- Context:
  - Experiment in yeast: 2 strains (BY and RM) are hybridized.
  - N is the total number of expression reads at a particular location in the genome.
  - Y is the number from BY.
  - θ is the probability of a read from BY.
  - If $\theta \neq 0.5$ we have allele specific expression.

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 use 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 principled method for updating uncertainty based on information.

Bayesian methods

Bayesian methods are data analysis tools that are derived from the principles of Bayesian inference and provide:

- parameter estimates with good statistical properties;
- parsimonious models that can describe observed data;
- inference for missing data and predictions for future data;
- a framework for estimation and model selection;
- 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 quantifies unknown population characteristics.

Data: y quantifies the outcome of a survey/experiment/...

Our goal is to make inference about θ given y.

In the ASE experiment, θ is the probability of a BY allele, and y is the observed BY count (out of N) .

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 ASE data at one location:

Sample space: $\mathcal{Y} = \{0, 1, \ldots, N\}$ is the set of all possible outcomes.

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

Ingredients of a Bayesian analysis

Quantifying information:

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

Sampling model: $p(y|\theta)$, defined for $\theta \in \Theta, y \in \mathcal{Y}$, describes our belief that y will be the 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 – probability of the data, given the model (likelihood and prior).

Ingredients of a Bayesian analysis

For the ASE data:

Prior distribution: $p(\theta)$ describes our beliefs about the unknown probability θ of a BY read, before we look at the data.

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

Posterior distribution: $p(\theta|y)$ describes our beliefs about the unknown probability θ, 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.

I view the Bayesian approach to statistical inference very pragmatically, as a means by which models for data can be constructed – I certainly use frequentist techniques for some problems.

ASE Example

A natural choice for the number of BY alleles is:

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

The maximum likelihood estimate (MLE) is

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

with standard error

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

which is estimated by

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

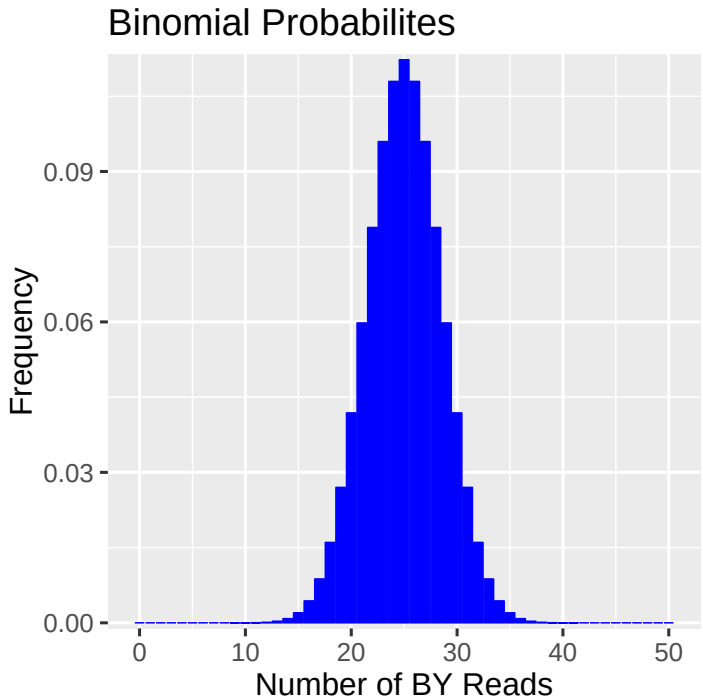


Figure: Probabilities of different binomial outcomes with $N = 50, \theta = 0.5$.

ASE Example

Suppose for a particular gene $y = 0$, then

$$\widehat{\theta} = \frac{y}{N} = \overline{y} = 0$$

with standard error

$$\sqrt{\frac{\widehat{\theta}(1-\widehat{\theta})}{N}} = 0.$$

Both of these are clearly poor choices, the standard error in particular.

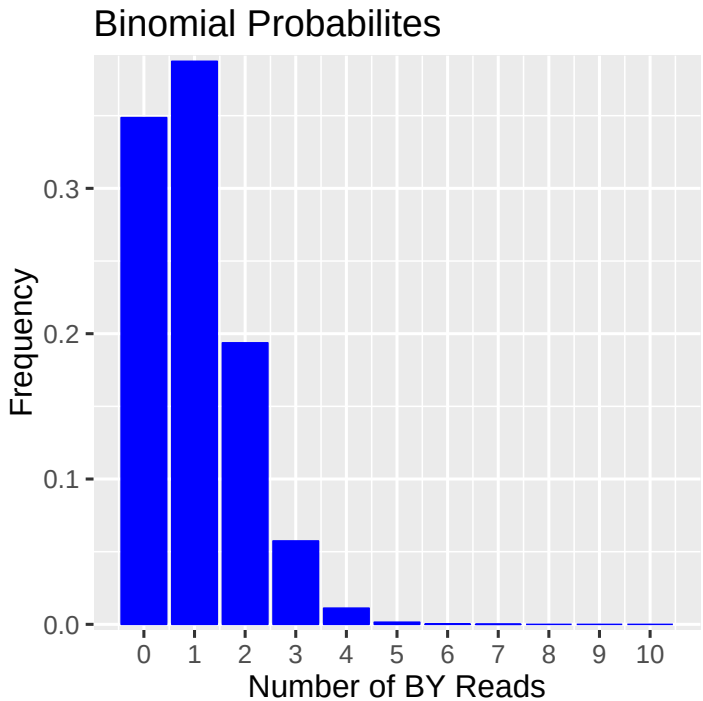


Figure: Probabilities of different binomial outcomes with $N = 10, \theta = 0.1$.

Bayes and frequentist estimates for binomial

"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).

Bayes and frequentist estimates for binomial

Example: $N = 20, y = 0$ 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

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

Can be seen as Bayesian procedure, with a Beta(2,2) prior for θ – Bayes 95% interval is (0.019,0.36) – see later for details.

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. 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, $i \neq j$ (A_i and A_j can't happen together).
- Ω is the sample space, and $\mathcal{F}$ be a suitable collection[1] 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 $A_1, A_2, \cdots \in \mathcal{F}$.

[1] 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: Suppose:

$$\{\text{ evidence = white tee-shirt }\}$$

and we know crime was committed by someone with a white tee-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 genotype is $\{bb, Bb, BB\}$ with probability $\{1/4, 1/2, 1/4\}$.
- $A_1 = \{\text{genotype is } bb\}$, $A_2 = \{\text{genotype is } Bb\}$
- A_1 and A_2 are disjoint, and so

$$\begin{aligned} \Pr(\text{one or more } b \text{ alleles}) &= \Pr(A_1 \cup A_2) \\ &= \Pr(A_1) + \Pr(A_2) \\ &= 1/4 + 1/2 \\ &= 3/4 \end{aligned}$$

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 $\mathcal{H}$, written 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.

Examples:

- $\mathcal{H}$=someone's number of children
  - {0, 1, 2, 3 or more};
  - {0, 1, 2, 3, 4, 5, 6, ...}.
- $\mathcal{H}$ = the relationship between a genotype and heart disease
  - {some relationship, no relationship};
  - Genotype is: { detrimental, not detrimental}.

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 is the probability 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.
- 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.

Monozygotic/dizygotic example

It is known that someone will have twins, e.g., from detection of two heartbeats.

A sonogram indicates there are twin girls.

What is the probability that the girls are monozygotic (single egg)?

Observed data: Twins are girls.

Prior information: Given twins, approximately one third of twins are monozygotic (from information in a particular population, remember the conditioning information, I).

Monozygotic/dizygotic example

$E_1 = \{GG\}$ is event of girl twins, H_1 is event of monozygotic, H_2 is event of dizygotic.

Girl twins can be either monozygotic or dizygotic:

$$\begin{aligned}
\Pr(E_1) &= \underbrace{\Pr(E_1|H_1)}_{\text{GG or BB}} \Pr(H_1) + \underbrace{\Pr(E_1|H_2)}_{\text{GG or GB or BB}} \Pr(H_2) \\
&= 1/2 \times 1/3 + 1/4 \times 2/3 \\
&= 1/6 + 1/6 = 1/3
\end{aligned}$$

Updated beliefs:

$$\begin{aligned}
\Pr(H_1|E_1) &= \frac{\Pr(E_1|H_1)\Pr(H_1)}{\Pr(E_1)} \\
&= \frac{\Pr(E_1|H_1)\Pr(H_1)}{\Pr(E_1|H_1)\Pr(H_1) + \Pr(E_1|H_2)\Pr(H_2)} \\
&= \frac{1/2 \times 1/3}{1/3} \\
&= 1/2 > 1/3 = \Pr(H_1)
\end{aligned}$$

Monozygotic/dizygotic example

Let $E_2 = \{BB\}$ be the event of knowing twin boys, and $E_3 = \{BG\}$ the event of knowing a boy and a girl, H_1 is again event of monozygotic.

Observed data (likelihood) calculations:

$$\begin{aligned} \Pr(E_1|H_1) &= \frac{1}{2} \qquad \Pr(E_2|H_1) = \frac{1}{2} \qquad \Pr(E_3|H_1) = 0, \\ \Pr(E_1|H_2) &= \frac{1}{4} \qquad \Pr(E_2|H_2) = \frac{1}{4} \qquad \Pr(E_3|H_2) = \frac{1}{2}. \end{aligned}$$

Show:

$$\begin{aligned} \Pr(E_2) &= \frac{1}{3} \\ \Pr(E_3) &= \frac{1}{3} \\ \Pr(H_1|E_2) &= \frac{1}{2} \\ \Pr(H_1|E_3) &= 0 \qquad (\textit{Implications?}) \end{aligned}$$

Bayesian inference

$\{H_1, \ldots, H_K\}$ often refer to disjoint hypotheses or states of nature

E refers to the the data.

Post-data evaluation of the relative evidence for hypotheses are via the posterior odds ratio:

$$
\begin{aligned}
\text{“posterior ratio”} &= \frac{\Pr(H_i|E)}{\Pr(H_j|E)} \\
&= \frac{\Pr(E|H_i) \times \Pr(H_i)/\cancel{\Pr(E)}}{\Pr(E|H_j) \times \Pr(H_j)/\cancel{\Pr(E)}} \\
&= \frac{\Pr(E|H_i) \times \Pr(H_i)}{\Pr(E|H_j) \times \Pr(H_j)} \\
&= \frac{\Pr(E|H_i)}{\Pr(E|H_j)} \times \frac{\Pr(H_i)}{\Pr(H_j)} \\
&= \text{“likelihood ratio”} \times \text{“prior ratio”}
\end{aligned}
$$

Later we will investigate this further, when we discuss Bayes factors.

Twin example

Prior odds:

$$\frac{\Pr(H_1)}{\Pr(H_2)} = \frac{1/3}{2/3} = 1/2$$

Prior favors H_2

Likelihood ratio:

$$\frac{\Pr(E_1|H_1)}{\Pr(E_1|H_2)} = \frac{1/2}{1/4} = 2$$

Data is more consistent with H_1

Posterior odds:

$$\frac{\Pr(H_1|E_1)}{\Pr(H_2|E_1)} = \frac{\Pr(H_1)}{\Pr(H_2)} \frac{\Pr(E_1|H_1)}{\Pr(E_1|H_2)} = 1$$

Independence

In general:

$$\Pr(F \cap G) = \Pr(F) \times \Pr(G|F).$$

F and G are independent if

$$\Pr(F \cap G) = \Pr(F) \times \Pr(G),$$

i.e.,

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

so that knowledge that F occurred does not alter our beliefs in G occurring.

Conditional Independence

Conditional independence is used far more than independence.

In general,

$$\Pr(F \cap G|H) = \Pr(F|H) \times \Pr(G|F \cap H).$$

F and G are conditionally independent given H, if

$$\Pr(F \cap G|H) = \Pr(F|H) \times \Pr(G|H).$$

i.e.,

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

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

Conditional Independence

Example of use in statistics:

$F = \{$ a patient will develop cancer $\}$
$G = \{$ the parents' genotypes $\}$
$H = \{$ a patient's genotype $\}$

$$\Pr(F|H) \stackrel{?}{=} \Pr(F|G,H)$$

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

Genomic imprinting is an epigenetic phenomenon that causes genes to be expressed in a parent-of-origin-specific manner, i.e., the expression of the gene depends upon the parent who passed on the gene.

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 a specific allele
- $Y =$ number of children of a randomly sampled person
- $Y =$ number of years of education of a randomly sampled person

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:

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

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:

1. $0 \leq p(y)$ 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 $Y = y$." (the probability of any y is zero for a continuous random variable).

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 $Y = y$:

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

for $\epsilon > 0$, which is a non-zero 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 $\mathcal{Y} = \{0, 1\}$.

For a random variable that can only take 2 values, there is only one possible distribution – *obvious?*

Namely, 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

$$\Pr(Y = y|\theta) = p(y|\theta) = \theta^y (1 - \theta)^{1-y}$$

Mean is θ, variance is $\theta(1 - \theta)$ (so greatest uncertainty when $\theta = 0.5$).

Conditionally independent binary outcomes

Suppose the prevalence of an allele in a population is θ.

Let $Y_1, \ldots, Y_N$ indicate the presence of the allele for N individuals randomly sampled from the population.

Due to **conditional independence**:

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

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 the allele?

Consider all N-sequences with y 1's:

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

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 , written Binomial(N, θ) 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:

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

Bayes theorem in statistics

Bayes theorem:

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

So choices of the likelihood (sampling model) $p(y|\theta)$ and prior $p(\theta)$ lead to particular posterior distribution.

Once we obtain the posterior we might display the complete distribution, or report summaries.

The denominator is obtained as (continuous version of rule of marginal probability we saw earlier):

$$p(y) = \int p(y|\theta)p(\theta)\, d\theta,$$

– the parameter θ is giving the partition, which makes it clear that the data y are assumed to arise from $p(y|\theta)$ for some θ, so with probability 1 we are saying the data had to arise from whatever form $p(y|\theta)$ we assume for the data.

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

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

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:

$$\sqrt{\mathrm{E}[(\theta - \mathrm{E}[\theta|y])^2]} = \sqrt{\int_{\Theta} (\theta - \mathrm{E}[\theta|y])^2 p(\theta|y)\, d\theta}.$$

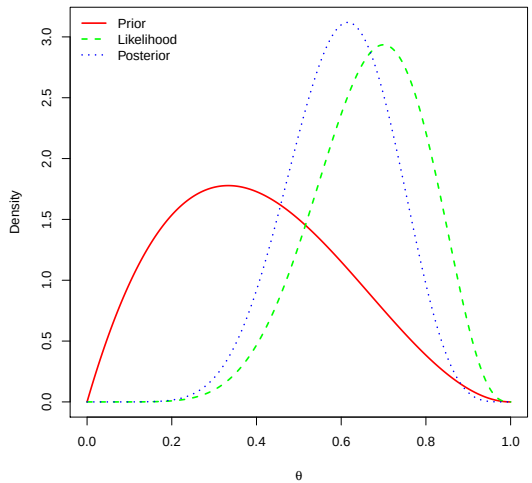
Not so useful for posterior distributions that are not normal-like in shape.

The beta posterior

It can be shown (in detail next lecture!) that if:

- $\theta \sim \text{Beta}(a, b)$
- $Y|\theta \sim \text{Binomial}(N, \theta)$ then the posterior is

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



Posterior mean:

$$\begin{aligned}
\text{E}[\theta|y] &= \frac{a+y}{a+b+N} \\
&= \frac{a}{a+b}\left(\frac{a+b}{a+b+N}\right) + \frac{y}{N}\left(\frac{N}{a+b+N}\right) \\
&= \text{E}[\theta]\left(\frac{a+b}{a+b+N}\right) + \overline{y}\left(\frac{N}{a+b+N}\right)
\end{aligned}$$

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

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 $a = b = 2$.

The Poisson distribution

Let $\mathcal{Y} = \{0, 1, 2, \ldots\}$. The outcome $Y \in \mathcal{Y}$ has a Poisson distribution with mean θ, written Poisson(θ), if

$$\Pr(Y = y|\theta) = \frac{\theta^y e^{-\theta}}{y!}.$$

For example, suppose Y is the number of children of a randomly selected couple; $\theta = 2.1$ (the 2006 U.S. fertility rate),

$$\begin{array}{lll}
\Pr(Y = 0|\theta = 2.1) = & (2.1)^0 e^{-2.1}/(0!) = & 0.122 \\
\Pr(Y = 1|\theta = 2.1) = & (2.1)^1 e^{-2.1}/(1!) = & 0.257 \\
\Pr(Y = 2|\theta = 2.1) = & (2.1)^2 e^{-2.1}/(2!) = & 0.271 \\
\Pr(Y = 3|\theta = 2.1) = & (2.1)^3 e^{-2.1}/(3!) = & 0.189 \\
\Pr(Y = 4|\theta = 2.1) = & (2.1)^4 e^{-2.1}/(4!) = & 0.099 \\
\Pr(Y = 5|\theta = 2.1) = & (2.1)^5 e^{-2.1}/(5!) = & 0.042 \\
\Pr(Y = 6|\theta = 2.1) = & (2.1)^6 e^{-2.1}/(6!) = & 0.015 \\
\vdots & \vdots & \vdots
\end{array}$$

Another example: tumor counts in mice.

The Poisson likelihood

Let Y_i be the number of tumor counts in experiment i, $i = 1, \ldots, n$.

What is the mean tumor count in this population?

The likelihood: Again assuming conditional independence:

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

Simplification: Let $Y = \sum_{i=1}^{n} Y_i$. Then $Y|\theta \sim \text{Poisson}(n\theta)$ and so

$$\Pr(Y = y|\theta) = \theta^y e^{-n\theta} \times (n^y/y!)$$

The "business end" of the likelihood in both cases is $\theta^y e^{-n\theta}$.

The gamma posterior distribution

It can be shown that if

- The likelihood is $Y_1, \ldots, Y_n | \theta \sim \text{Poisson}(\theta)$,
- $\theta \sim \text{Gamma}(a, b)$ (the conjugate prior),
- then the posterior is

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

Posterior mean:

$$\begin{aligned} \text{E}[\theta | y] &= \frac{a + y}{b + n} \\ &= \frac{a}{b}\left(\frac{b}{b+n}\right) + \frac{y}{n}\left(\frac{n}{b+n}\right) \\ &= \text{E}[\theta]\left(\frac{b}{b+n}\right) + \overline{y}\left(\frac{n}{b+n}\right), \end{aligned}$$

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

The Poisson likelihood

Suppose $n = 20$ mice and $y = \sum_{i=1}^{n} y_i = 324$ is the total tumor count ($y/n = 16.2$).

Similar populations of mice suggest $\theta \approx 10$.

A prior distribution for θ which is consistent with this (though we would need to think about whether the spread of this prior is appropriate) is:

$$\begin{aligned} \theta &\sim \text{gamma}(10, 1) \\ \text{E}[\theta] &= 10 \\ \text{SD}[\theta] &= \sqrt{10} \approx 3.16 \end{aligned}$$

The posterior is Gamma$(10 + 324, 1 + 20)$ and the posterior mean for the rate is

$$\text{E}[\theta|y] = \frac{a + y}{b + n} = \frac{334}{21} = 15.9.$$

The normal distribution

Let $\mathcal{Y} = (-\infty, \infty)$.

The outcome $Y \in \mathcal{Y}$ has a normal distribution with mean θ and variance σ^2, written $\mathrm{N}(\theta, \sigma^2)$, if

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

The normal posterior distribution

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

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

and the variance of this estimator is

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

If:

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

Then, the posterior is also normal.

The normal posterior distribution

The posterior mean is,

$$\mathrm{E}[\theta|y_1,\ldots,y_n] \quad = \quad m(1-w)+\bar{y}w$$

where the weight on the data is

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

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

The posterior variance is,

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

Bayes Example

- Likelihood:

$$\bar{y} \mid \theta \sim \mathrm{N}(\theta, \sigma^2/n),$$

  where σ^2/n is assumed known ($\sigma/\sqrt{n}$ is the standard error).
- We also imagine the prior is normal:

$$\theta \sim \mathrm{N}(m, v),$$

  so that values of the mean θ that are (relatively) far from m are penalized.
- The log posterior is:

$$\underbrace{\log p(\theta \mid y\ )}_{\textbf{Updated Beliefs}} = \underbrace{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2}_{\textbf{Data Model}} \underbrace{-\frac{1}{2v}(\theta-m)^2}_{\textbf{Penalization}}.$$

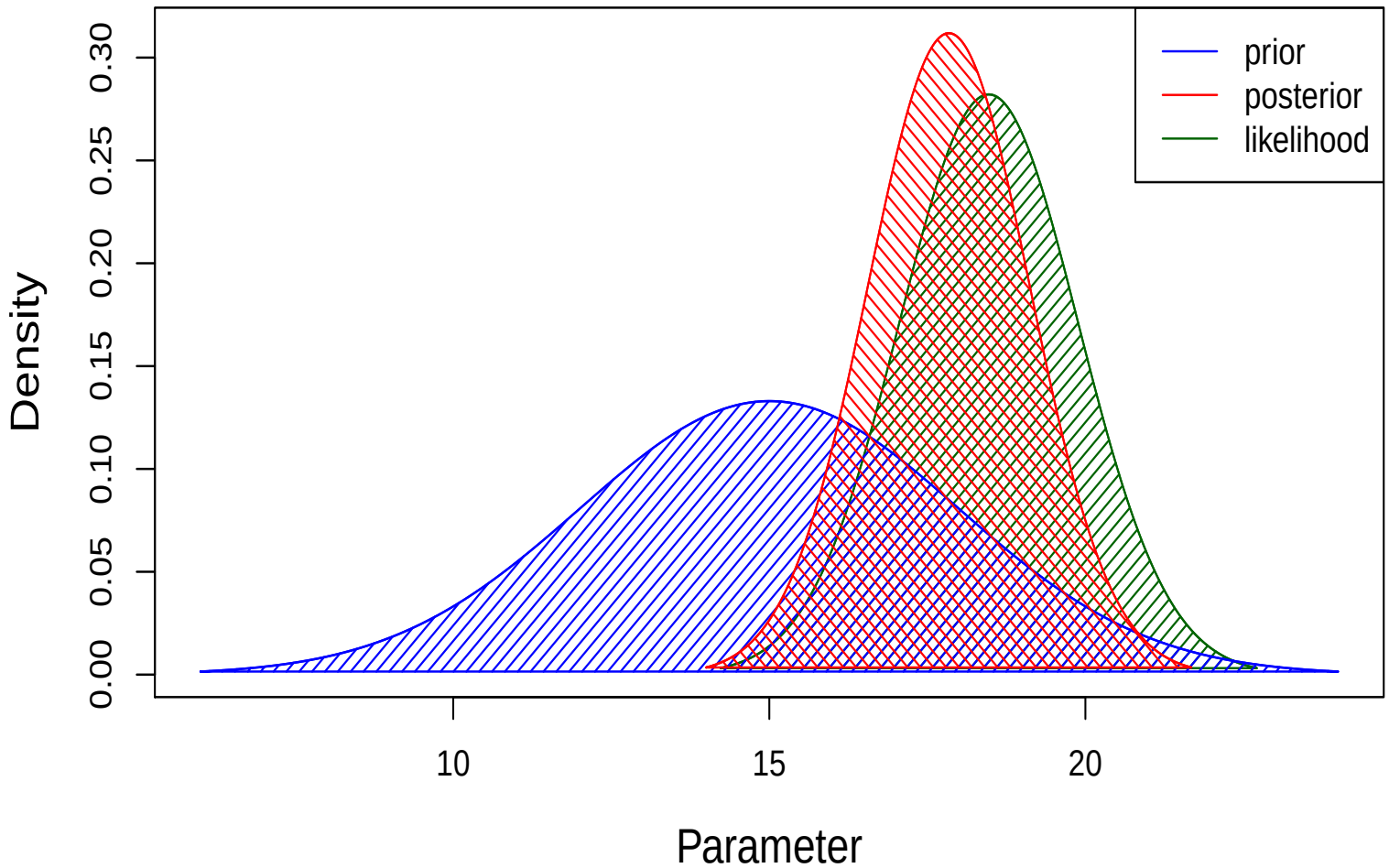


Figure: Normal data model with $n = 10$, $\overline{y} = 19.3$ and standard error 1.41. The prior for θ has mean $m = 15$ and $v = 3^2$. The posterior for the parameter θ is a compromise between the two sources of information: the posterior mean is 18.5 and the posterior standard deviation is 1.28.

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.

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.

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

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