

2016 SISG Module 17: Bayesian Statistics for Genetics

Lecture 2: Review of Probability and Bayes Theorem

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Introduction and Motivating Example

Bayesian learning

Probability and Bayes Theorem

Standard distributions and conjugacy

Introduction

- In this lecture we will first consider generic Bayesian learning.
- Background reading: Chapters 1 and 2 of Hoff (2009).
- This will include a review of probability theory.
- The analysis of **allele specific expression** data will be used to motivate the binomial model.
- **Conjugate priors** will be introduced.

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.

Motivating Example: An Example of ASE

- 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

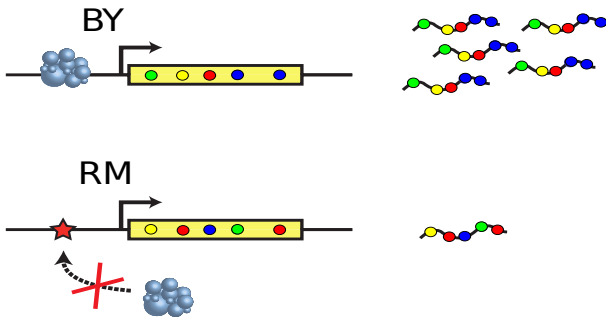


Figure 1 : In the top figure the transcription factor (blue) leads to high transcription. In the bottom figure an upstream polymorphism (red star) prevents the transcription factor from binding.

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).

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**.
- 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**.

- A non-Bayesian approach would 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?

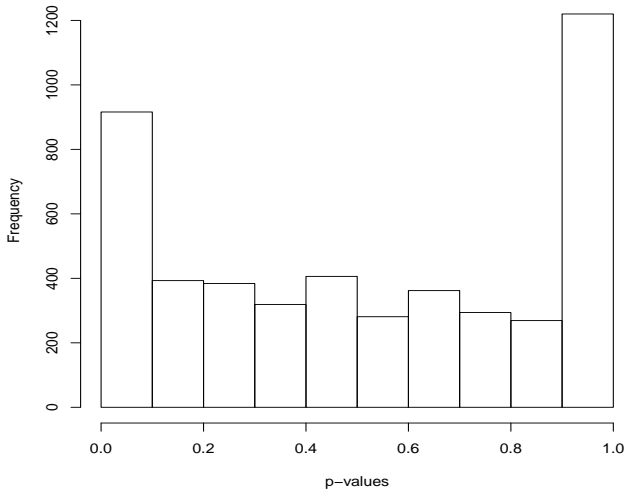


Figure 2 : p -values from 4,844 exact tests.

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 rational method for updating uncertainty based on information

Inductive learning via Bayes theorem is referred to as **Bayesian inference**.

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.

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 or 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 gene:

Sample space: $\mathcal{Y} = \{0, 1, \dots, N\}$ 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 experimental outcome, for each θ .

Updating information:

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

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta}) d\tilde{\theta}}.$$

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 $y = 0, 1, \dots, N$ **given** we (hypothetically) know the value of the probability θ . When viewed as a function of θ , 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

Recall Bayes theorem:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_{\Theta} p(y|\tilde{\theta})p(\tilde{\theta}) d\tilde{\theta}}.$$

The ratio of **posterior densities** at two points in the parameter space θ_a, θ_b is

$$\frac{p(\theta_a|y)}{p(\theta_b|y)} = \frac{p(y|\theta_a) p(\theta_a)}{p(y|\theta_b) p(\theta_b)},$$

and is the product of ratios of **likelihoods** and **priors**.

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.

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.

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

$$\hat{\theta} = \frac{y}{N} = \bar{y}$$

with standard error

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

which is estimated by

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

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

Comparison to non-Bayesian methods in the ASE 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:

$$\hat{\theta} \pm 1.96 \sqrt{\hat{\theta}(1 - \hat{\theta})/N}, \text{ where}$$

$$\hat{\theta} = \frac{4}{N+4} \frac{1}{2} + \frac{N}{N+4} \bar{y},$$

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).

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, \dots , are **mutually exclusive** (disjoint) events if $A_i \cap A_j = \emptyset$, for all pairs $i, j, i \neq j$.
- A^c is the complement of A , so that $A \cup A^c = \Omega$.
- Ω is the sample space, and \mathcal{F} be a suitable collection¹ of subsets of Ω .

Axioms of Probability:

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

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

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

¹Technically, 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!

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

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

Example

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

- 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.

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

Events and partitions

Definition (Partition)

A collection of sets $\{H_1, \dots, H_K\}$ 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 and $\{H_1, \dots, H_K\}$ is a partition of \mathcal{H} , then exactly one out of $\{H_1, \dots, H_K\}$ contains the truth.

Examples:

- \mathcal{H} = someone's number of children
 - $\{0, 1, 2, 3 \text{ or more}\}$;
 - $\{0, 1, 2, 3, 4, 5, 6, \dots\}$.
- \mathcal{H} = the relationship between a genotype and heart disease
 - $\{\text{some relationship, no relationship}\}$;
 - $\{\text{negative correlation, zero correlation, positive correlation}\}$.

Bayes theorem

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

Bayes theorem :
$$\Pr(H_j|E) = \frac{\Pr(E|H_j) \Pr(H_j)}{\Pr(E)}$$
$$= \frac{\Pr(E|H_j) \Pr(H_j)}{\sum_{k=1}^K \Pr(E|H_k) \Pr(H_k)}$$

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

One begins with (**prior**) beliefs about events H_j , $\Pr(H_j)$, and then updates these (**posterior**) beliefs to $\Pr(H_j|E)$, 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?

The first step in using Bayes theorem is to define events and translate the above:

- A = event of having the faulty gene, so that $\Pr(A) = 0.01$. That also means 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). A and A^c form a partition.

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:

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

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

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

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 .

$$\begin{aligned}
 \text{Bayes theorem : } \Pr(H_j|E, I) &= \frac{\Pr(E|H_j, I) \Pr(H_j|I)}{\Pr(E|I)} \\
 &= \frac{\Pr(E|H_j, I) \Pr(H_j|I)}{\sum_{k=1}^K \Pr(E|H_k, I) \Pr(H_k|I)}
 \end{aligned}$$

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 probabilities...

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 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 arises, then I would pay $\frac{1}{2}$ a unit of money to enter the bet.

Another 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, one third of twins are monozygotic, approximately, from information in a particular population (remember I).

Updated beliefs:

$$\begin{aligned}
 \Pr(H_1|E) &= \frac{\Pr(E|H_1) \Pr(H_1)}{\Pr(E)} \\
 &= \frac{\Pr(E|H_1) \Pr(H_1)}{\Pr(E|H_1) \Pr(H_1) + \Pr(E|H_2) \Pr(H_2)} \\
 &= \frac{1/2 \times 1/3}{1/2 \times 1/3 + 1/4 \times 2/3} = \frac{1/6}{1/6 + 1/6} = 1/2 > 1/3 = \Pr(H_1)
 \end{aligned}$$

Bayesian inference

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

E refers to the outcome of a survey, study or experiment.

Post-experimental evaluation of hypotheses are via the posterior odds ratio:

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

Later we will generalize this idea to 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:

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

Data is better explained by H_1

Posterior odds:

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

Independence and Conditional Independence

In general:

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

F and G are **independent** if

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

i.e., $\Pr(G|F) = \Pr(G)$ so that knowledge that F occur does not alter our beliefs in G occurring.

Conditional independence is appealed to far more than independence.

In general,

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

F and G are **conditionally independent** given H if

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

i.e., $\Pr(G|F \cap H) = \Pr(G|H)$ so that **given** H knowledge that F occur does not alter our beliefs in G occurring.

Conditional Independence

Example of use in Statistics:

$F = \{ \text{a patient will develop cancer} \}$

$G = \{ \text{the parents' genotypes} \}$

$H = \{ \text{a patient's genotype} \}$

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

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 $\mathcal{Y} = \{y_1, y_2, \dots\}$.

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

Probability distributions and densities

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

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

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

Probability densities

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

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

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

Probability densities

As in the discrete case, probability statements about Y can be derived from the pdf:

$$\Pr(Y \in A) = \int_{y \in A} p(y) dy,$$

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

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

Probability densities

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 $Y = y$:

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

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 $\mathcal{Y} = \{0, 1\}$.

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

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

Independent binary outcomes

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

Let Y_1, \dots, Y_n indicate the presence of the allele for n individuals randomly sampled from the population.

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

Note that $p(y_1, \dots, y_N | \theta)$ depends only on $\sum_{i=1}^N y_i$.

Often, we only record $y = \sum_{i=1}^N y_i$ and N .

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:

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

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 $\mathcal{Y} = \{0, 1, 2, \dots, N\}$ for some positive integer N . The outcome $Y \in \mathcal{Y}$ has a *binomial distribution with probability θ* if

$$\Pr(Y = y|\theta) = \text{dbinom}(y, N, \theta) = \binom{N}{y} \theta^y (1 - \theta)^{N-y}.$$

For example, if $\theta = 0.25$ and $N = 4$, we have:

$$\Pr(Y = 0|\theta = .25) = \binom{4}{0} (.25)^0 (.75)^4 = .316$$

$$\Pr(Y = 1|\theta = .25) = \binom{4}{1} (.25)^1 (.75)^3 = .422$$

$$\Pr(Y = 2|\theta = .25) = \binom{4}{2} (.25)^2 (.75)^2 = .211$$

$$\Pr(Y = 3|\theta = .25) = \binom{4}{3} (.25)^3 (.75)^1 = .047$$

$$\Pr(Y = 4|\theta = .25) = \binom{4}{4} (.25)^4 (.75)^0 = .004.$$

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:

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

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

Aside: notice the link with the adjusted Wald interval for the 0 successes case, the estimate is equal to the posterior mean with a $\text{beta}(a, b)$ prior with $a = b = 2$.

This 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 analyses are not possible.

The Poisson distribution

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

$$\Pr(Y = y|\theta) = \text{dpois}(y, \theta) = \theta^y e^{-\theta} / y!.$$

For example, if $\theta = 2.1$ (the 2006 U.S. fertility rate),

$$\Pr(Y = 0|\theta = 2.1) = (2.1)^0 e^{-2.1} / (0!) = .12$$

$$\Pr(Y = 1|\theta = 2.1) = (2.1)^1 e^{-2.1} / (1!) = .26$$

$$\Pr(Y = 2|\theta = 2.1) = (2.1)^2 e^{-2.1} / (2!) = .27$$

$$\Pr(Y = 3|\theta = 2.1) = (2.1)^3 e^{-2.1} / (3!) = .19$$

$$\vdots$$

$$\vdots$$

$$\vdots$$

The Poisson likelihood

Let Y_i be the number of intestinal tumors for mouse i , $i = 1, \dots, n$.

What is the mean tumor count in this population?

The likelihood:

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

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 Poisson likelihood

Suppose $n = 20$ and $y = \sum_{i=1}^n y_i = 324$ ($y/n = 16.2$).

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

A suitable prior distribution for θ which is consistent with this:

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

The gamma posterior distribution

It can be shown that if

- $\theta \sim \text{gamma}(a, b)$
- $Y_1, \dots, Y_n | \theta \sim \text{Poisson}(\theta)$ then the **posterior** is

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

Posterior mean:

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

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

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 if

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

The normal posterior distribution

It can be shown that if

- $\theta \sim \text{normal}(\mu_0, \tau_0^2)$ and
- $Y_1, \dots, Y_n | \theta \sim \text{normal}(\theta, \sigma^2)$

then

$$\theta | y_1, \dots, y_n \sim \text{normal}(\mu_n, \tau_n^2)$$

where

$$\begin{aligned} \text{Var}[\theta | y_1, \dots, y_n] = \tau_n^2 &= [1/\tau_0^2 + n/\sigma^2]^{-1} \\ 1/\tau_n^2 &= 1/\tau_0^2 + n/\sigma^2 \end{aligned}$$

(so precisions are additive) and

$$\begin{aligned} E[\theta | y_1, \dots, y_n] = \mu_n &= \frac{\mu_0/\tau_0^2 + \bar{y}n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \\ &= \mu_0 \left(\frac{1/\tau_0^2}{1/\tau_0^2 + n/\sigma^2} \right) + \bar{y} \left(\frac{n/\sigma^2}{1/\tau_0^2 + n/\sigma^2} \right) \end{aligned}$$

so the posterior mean is a weighted combination of the prior mean and the sample mean

Describing posterior location

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

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

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.

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

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

- **Conjugate analyses** are computationally convenient but rarely available in practice.

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.