Module 4: Bayesian Methods
Lecture 1: Why Bayes?

Peter Hoff

Departments of Statistics and Biostatistics
University of Washington

Outline

Bayesian learning

Probability of a rare event

Predictive models

Probability and information

We often use "probability" informally to express belief or information.

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' rule is a rational method for updating information

Inductive learning via Bayes' rule 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 computational 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.

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

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' rule: After obtaining data y, the posterior distribution is calculated

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

Role of prior information

$$
\begin{aligned}
p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\tilde{\theta})p(\tilde{\theta})\ d\tilde{\theta}} \\
\frac{p(\theta_a|y)}{p(\theta_b|y)} &= \frac{p(y|\theta_a)}{p(y|\theta_b)}\frac{p(\theta_a)}{p(\theta_b)}.
\end{aligned}
$$

Bayes' rule does not tell us what our beliefs should be.

It tells us how they should change after seeing new information.

Probability of a rare event

Suppose we are interested in the prevalence of a rare genetic mutation in a human subpopulation.

A random sample of 20 individuals from the population are sampled and genotyped.

Parameter and sample space

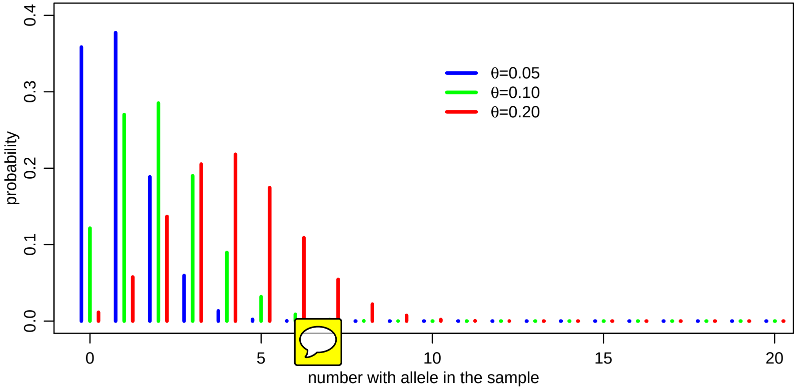
$$\Theta = [0, 1] \qquad \mathcal{Y} = \{0, 1, \ldots, 20\}\,.$$

Sampling model

Before the sample is obtained, the number of individuals with the mutation is unknown. Let Y denote this to-be-determined value.

Sampling model: If the value of θ were known, a reasonable sampling model for Y would be a binomial$(20, \theta)$ probability distribution:

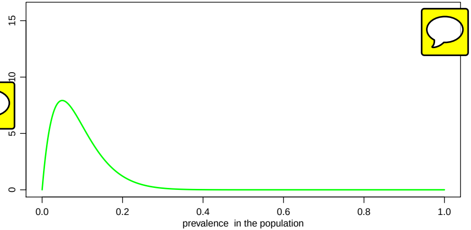
$$Y|\theta \sim \text{binomial}(20, \theta) \,.$$



Prior distribution

$$\theta \sim \text{beta}(2, 20)$$

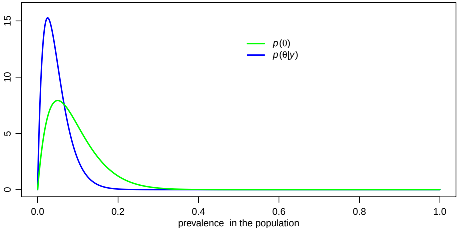
$$\begin{aligned} E[\theta] &= 0.09 \\ \text{mode}[\theta] &= 0.05 \\ \Pr(0.01 < \theta < 0.24) &= 0.95 \end{aligned}$$



Posterior distribution

$$\{\theta | Y = 0\} \sim \text{beta}(2, 40)$$

$$\begin{aligned} \text{E}[\theta | Y = 0] &= 0.048 \\ \text{mode}[\theta | Y = 0] &= 0.025 \\ \Pr(.006 < \theta < .129 | Y = 0) &= 0.95. \end{aligned}$$



Sensitivity analysis

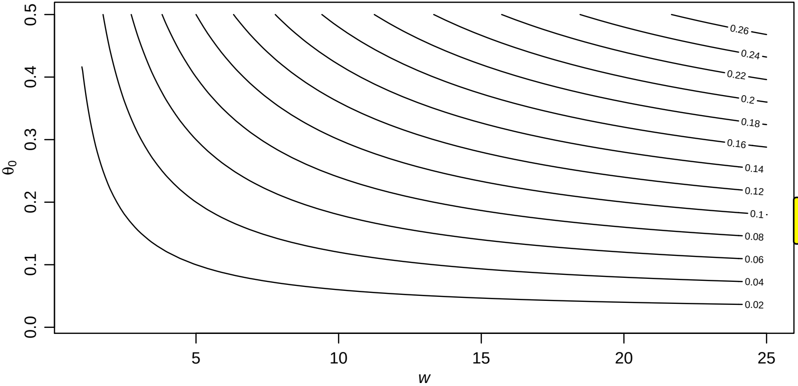
$$\theta \sim \text{beta}(a, b) \Rightarrow \{\theta | Y = y\} \sim \text{beta}(a + y, b + n - y)$$

$$\begin{aligned} \mathrm{E}[\theta | Y = y] &= \frac{a + y}{a + b + n} \\ &= \frac{n}{w + n}\bar{y} + \frac{w}{w + n}\theta_0 \end{aligned}$$

where $\theta_0 = a/(a + b)$ is the prior expectation of θ and $w = a + b$.

Sensitivity analysis

$$
\begin{aligned}
\mathrm{E}[\theta | Y = 0] &= \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0 \\
&= \frac{20}{w+20} \times 0 + \frac{w}{w+20} \times \theta_0 \\
&= \frac{w}{w+20} \times \theta_0
\end{aligned}
$$



Comparison to non-Bayesian methods

Non-Bayesian estimate:

$$\hat{\theta} = \bar{y} = y/n$$

For our data, $\hat{\theta} = 0$.

Non-Bayesian confidence interval:

$$\bar{y} \pm 1.96\sqrt{\bar{y}(1-\bar{y})/n} \quad \text{(Wald interval)}$$

For our data, 0 ± 0.

"Adjusted Wald interval" :

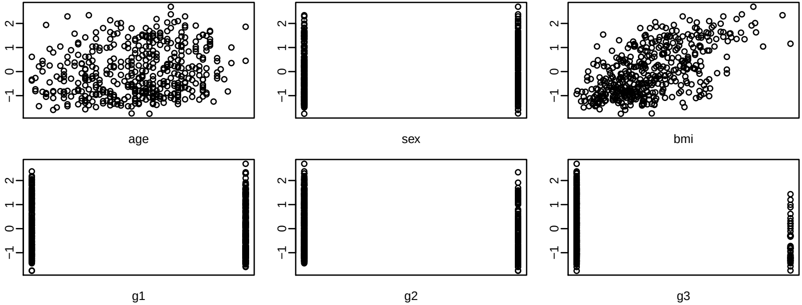
$$\begin{aligned} \hat{\theta} \quad &\pm \quad 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}\ , \text{ where} \\ \hat{\theta} \quad &= \quad \frac{n}{n+4}\bar{y} + \frac{4}{n+4}\frac{1}{2}. \end{aligned}$$

Can be seen as approximately Bayesian.

Example: diabetes progression

Goal:
Predict diabetes progression Y as a function of explanatory variables $\mathbf{x}$.

- Y is a subject's yearly diabetes progression index
- $\mathbf{x}$ is a 100-dimensional vector, including sex, age, genotype, ...



Building a predictive model

We need to

- build a predictive model,
- evaluate its predictive accuracy.

We will do this using out-of-sample validation.

Data on 442 subjects. We divide these into

- $n = 342$ training cases, with which to fit the model;
- $n_{\text{test}} = 100$ test cases, with which to evaluate the model.

Sampling model and parameter space

Data:

- item Y_i is a subject i's progression;
- $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,100})$ are i's explanatory variables.

Linear regression model:

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{100} x_{i,100} + \sigma \epsilon_i.$$

Parameters to estimate include

- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{100})$
- σ

Prior distribution

The role of the Bayesian prior:

- Idealistic: $p(\boldsymbol{\beta})$ should exactly describe your prior information about $\boldsymbol{\beta}$.
- Realistic: $p(\boldsymbol{\beta})$ should capture the gross features of our information about $\boldsymbol{\beta}$.

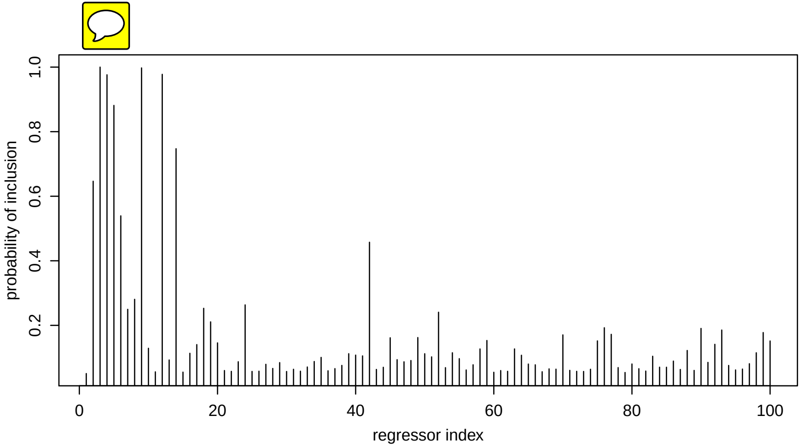
In our example we believe that $\beta_j \approx 0$ for many j.

We will use a prior such that

$$\Pr(\beta_j = 0) = 1/2 \text{ for each } j$$

Posterior distribution

$$\begin{aligned} \Pr(\beta_j \neq 0) &= 1/2 \text{ for each } j \in \{1, \ldots 100\} \\ \Pr(\beta_j \neq 0|\mathbf{Y}, \mathbf{X}) &\geq 1/2 \text{ for only six } j \end{aligned}$$



Out of sample validation

How well does the model perform?

Out of sample predictive performance: Compare $y_{i,\text{test}}$ to $\hat{y}_{i,\text{test}}$, where

$$\hat{y}_{i,\text{test}} = \hat{\boldsymbol{\beta}}^T \mathbf{x}_{i,\text{test}}$$

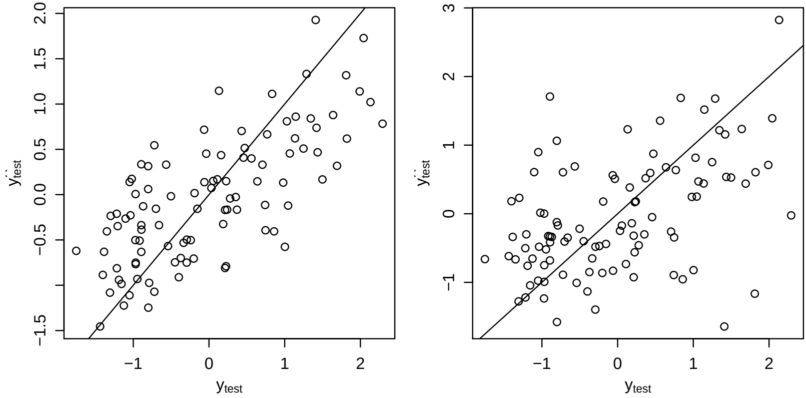
Important! $\hat{\beta}$ is estimated from the 342 training subjects, not the test subjects.

Prediction error

$$SPE(\hat{\beta}) = \frac{1}{100} \sum (y_{i,\text{test}} - \hat{y}_{i,\text{test}})^2$$

Out of sample validation

$SPE(\hat{\beta}_{\text{bayes}}) = 0.4853$ $SPE(\hat{\beta}_{\text{ols}}) = 0.9263$



Summary

The Bayesian approach provides

- models for rational, quantitative learning;
- estimators that work for small and large sample sizes;
- methods for generating statistical procedures in complicated problems.