Probability Review - Bayes Introduction

Statistics 220

Spring 2005

Advantages of Bayesian Analysis

- Answers the questions that researchers are usually interested in, “What is the probability that ...”

- Formal method for combining prior beliefs with observed (quantitative) information.

- Natural way of combining information from multiple studies.

- General approach for model indentification

- Approach for comparing non-nested models (Bayes factors)

- Can fit very realistic but complicated models

Disadvantages of Bayesian Analysis

- Often computationally more demanding than classical (e.g. frequentist inference).

- Software availablility: no general purpose software packages like SAS, SPSS, S-Plus/R available. This is getting better though with programs like BUGS.

- Requires at least one of

  – Elicitation of a real subjective probability distributions of prior beliefs.
  – Sensitivity analysis to show that the choice of prior doesn't strongly affect inference.

- Can fit overly complicated, but realistic models.

Bayesian Analysis

- Want to make probabilistic statements about parameters, θ, functions of parameters, $g(\theta)$, processes (which I will throw into the parameters for now), given the probability model and the observed data, y.

- Need to determine the posterior distribution, $p(\theta|y)$.

- Information available: the prior, $p(\theta)$, and the data model, $p(y|\theta)$.

- This gives us the full probability model, describing the randomness in the parameters and the data.

- Want to go from $p(y|\theta)$ to $p(\theta|y)$.

Bayes Rule

$$\begin{aligned} p(\theta|y) &= \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \\ &= \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta} \end{aligned}$$

The above is written assuming that θ is a continuous random variable with a density. However it could be discrete, giving

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\sum_i p(y|\theta_i)p(\theta_i)}$$

Generally I'll just write the continuous version as the discrete version will be analogous (replace integration with summation).

Bayes's Rule is often written as

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

Note that $p(y|\theta)$, sometimes referred to as the measurement model, when treated as a function of θ for a fixed y, is just the likelihood $L(\theta|y)$. So Bayes' Rule can be thought of as

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood}$$

One consequence of this is that Bayesian analysis satisfies the Likelihood Principle, which states that two data sets, with the same likelihood function, should lead to the same inferences.

For example, suppose you had two sequences of independent $Ber(p)$ trials of length n. If those two sequences had the same number of successes, then you would want to make the same statements about p from both analyses.

Odds Ratios:

Bayes's rule has a nice form in terms of odds ratios

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{p(y|\theta_1)}{p(y|\theta_2)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{L(\theta_1|y)}{L(\theta_2|y)}$$

i.e. the posterior odds are the prior odds times the likelihood ratio.

What is Probability?

If we are going to use probability to describe our levels of belief about a parameter or a process, we need to have an idea what we mean by probability.

Example 1: I have two dice in my pocket, one yellow and one purple. What is

- The probability that the yellow one rolls a 6?
- The probability that the purple one rolls a 6?
- The sum of the two rolls is 12?

Here is a picture of the two dice for those in the back.

So the probabilities are

- $P[\text{Yellow} = 6] = 0$
- $P[\text{Purple} = 6] = \frac{1}{20}$
- $P[\text{Sum} = 12] = \frac{1}{20}$

When determining probabilities and probability model there are two things that need to be considered:

1. What assumptions are you making (e.g. each outcome equally likely for each die and the dice are independent)?
2. What information are you conditioning on? All probabilities are effectively conditional.

Example 2: I have another two dice in my pocket (Blue (die 1) and Yellow (die 2)). What is the probability that they both roll the same number?

Let $Y_i =$ Roll on die i. Then

$$\begin{aligned} P[\text{All the same}] &= P[\text{All } 1] + P[\text{All } 2] + \ldots \\ &= P[Y_1 = Y_2 = 1] + P[Y_1 = Y_2 = 2] + \ldots \end{aligned}$$

Considerations:

1. Any dependency between the rolls on each die? Lets assume not.

2. What should we condition on - “Fool me once, shame on you. Fool me twice, shame on me” or will he only try to fool us once?

3. What could each die look like? There are 5 different dice based on Platonic solids (4, 6, 8, 12, 20 sides). I've heard of somebody trying to develop a gambling game based on a 7 sided die.

Let D_i = Number faces on die i

4. What numbers could be on each die? 1, 2, ... , D_i? All 5's? Is each side equally likely?

5. Any dependency between which dice are chosen? If the Blue die is 6 sided, can the other be 6 sided as well? Or are the D_i's independent?

6. If the D_i are independent, what are the probabilities?

Assumption 1 gives

$$\begin{aligned}
P[Y_1 = Y_2 = 1] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} P[Y_1 = 1, D_1 = i, Y_2 = 1, D_2 = j] \\
&= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} \{P[Y_1 = 1 | D_1 = i] P[Y_2 = 1 | D_2 = j] \\
&\qquad\qquad \times P[D_1 = i, D_2 = j]\}
\end{aligned}$$

If D_1 and D_2 are independent, then this reduces to

$$\begin{aligned} P[Y_1 = Y_2 = 1] &= \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\{P[Y_1 = 1|D_1 = i]P[Y_2 = 1|D_2 = j] \\ &\qquad\qquad \times P[D_1 = i, D_2 = j]\} \\ &= \left\{\sum_{i=1}^{\infty} P[Y_1 = 1|D_1 = i]P[D_1 = i]\right\} \\ &\qquad\qquad \times \left\{\sum_{j=1}^{\infty} P[Y_2 = 1|D_2 = j]P[D_2 = j]\right\} \\ &= P[Y_1 = 1]P[Y_2 = 1] \end{aligned}$$

So we can get an answer, lets assume that

1. $P[D_i = 4] = P[D_i = 6] = P[D_i = 8] = P[D_i = 12] = P[D_i = 20] = \frac{1}{5}$

2. $[Y_i = k|D_i] = \frac{1}{D_i}; \quad k = 1, \ldots, D_i$

Under these assumptions

k	1 - 4	5 - 6	7 - 8	9 - 12	13 - 20
$P[Y_i = k]$	$\frac{81}{600}$	$\frac{51}{600}$	$\frac{31}{600}$	$\frac{16}{600}$	$\frac{6}{600}$

$$P[\text{All the same}] = 0.0963$$

Where to probabilities come from?

- Long run relative frequencies

  If an experiment of independent trials is repeated over and over, the relative frequency of an event will converge to the probability of the event.

  Let A be the event of interest and let $p = P[A]$. Then for a sequence of independent trials, let

  $$Y_i = I(A \text{ occurs in trial } i); \qquad X_n = \sum_{i=1}^{n} Y_i$$

  Then by the law of large numbers, the sample proportion of successes

  $$\frac{X_n}{n} \longrightarrow p \quad \text{as } n \to \infty$$

For example, three different experiments looked at the probability of getting a head when flipping a coin.

- The French naturalist Count Buffon: 4040 tosses, 2048 heads ($\hat{p} = 0.5069$).
- While imprisoned during WWII, the South African statistician John Kerrich: 10000 tosses, 5067 heads ($\hat{p} = 0.5067$)
- Statistician Karl Pearson: 24000 tosses, 12012 heads ($\hat{p} = 0.5005$)

- Subjective beliefs

  Can be used for experiments that can't be repeated exactly, such as a sporting event. For example, what is the probability that the Patriots will win the Super Bowl next year. Can be done through comparison (i.e. is getting a head on a single flip of a coin more or less likely, getting a 6 when rolling a fair 6 sided die, etc). Can also be done by comparing different possible outcomes (1.5 times more likely than the Eagles, 10 more likely than than the Jets, 1,000,000 times more likely than the 49ers, etc).

  Often expressed in terms of odds

$$\text{Odds} = \frac{\text{Prob}}{1 - \text{Prob}}; \quad \text{Prob} = \frac{\text{Odds}}{1 + \text{Odds}}$$

The idea of subjective probability fits into the idea of a fair bet. Let $p \in [0, 1]$ be the amount that you are willing to bet for a return of \$1 if the event E occurs, i.e. gain \$(1-p) is E occurs, lose \$$p$ if the complement occurs. If you want this to be a fair bet ($E[\text{gain}] = 0$), then p is your subjective probability of the event E occuring.

Let $\tilde{p}$ be the probability of success. Then

$$E[\text{gain}] = (1 - p)\tilde{p} - p(1 - \tilde{p})$$

For this to be 0, $\tilde{p} = p$.

- Models, physical understanding, etc.

  The structure of the problem will often suggest a probability model. For example, the physics of rolling a die suggest that no one side should be favoured (equally likely outcomes), giving the uniform model used in the earlier example. However this could be verified by looking at the long run frequencies.

  Example: Genetics - Mendel's breeding experiments.

  The expected fraction of observed phenotypes in one of Mendel's experiments is given by the following model

Round/Yellow	$\frac{2+(1-\theta)^2}{4}$
Round/Green	$\frac{\theta(1-\theta)}{4}$
Wrinkled/Yellow	$\frac{\theta(1-\theta)}{4}$
Wrinkled/Green	$\frac{(1-\theta)^2}{4}$

The value θ, known as the recombination fraction, is a measure of distance between the two genes which regulate the two traits. θ must satisfy $0 \leq \theta \leq 0.5$ (under some assumptions about the process of meiosis) and when the two genes are on different chromosomes, $\theta = 0.5$.

The probabilities used to describe Mendel's experiments come from current beliefs of the underlying processes of meiosis (actually approximations to the processes), and the relationship between genes (genotype) and expressed traits (phenotypes).

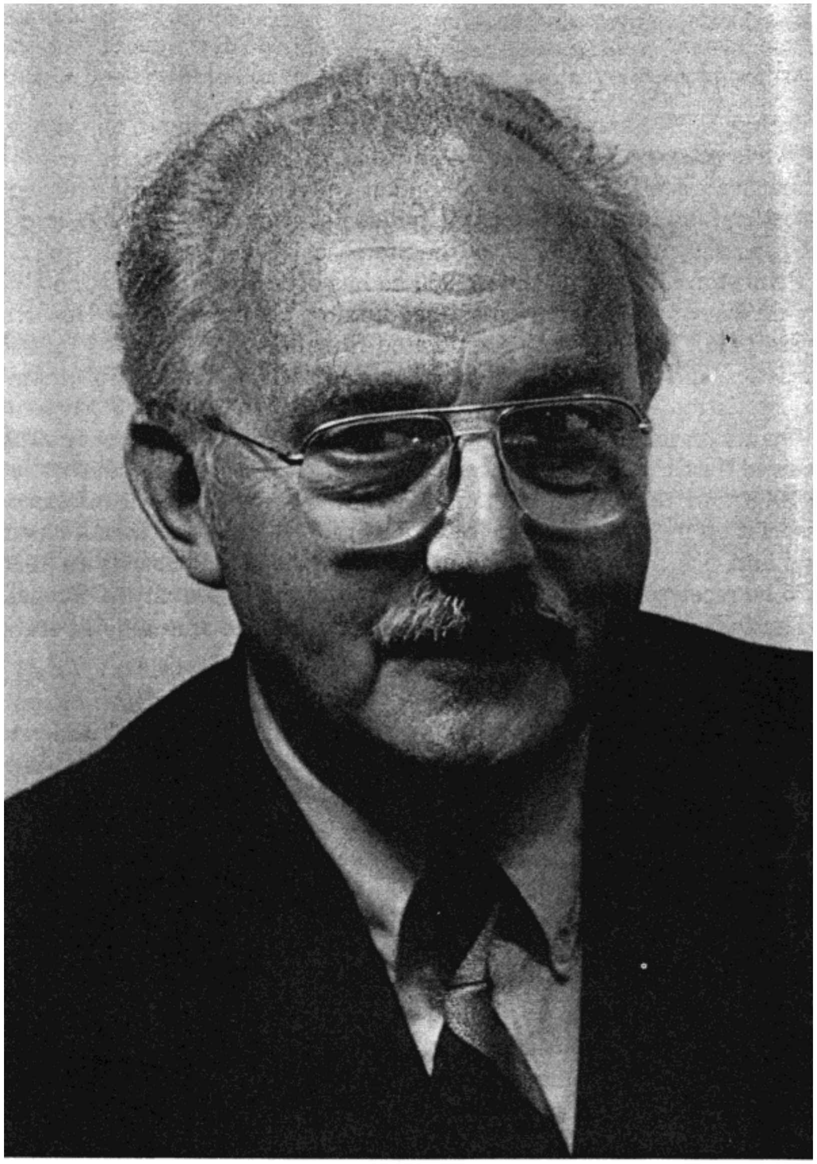
Actually the model is not right if we analyze Mendel's data. The model assumes that Mendel knew the genotypes of the plants he was crossing in his experiment. However his mechanism for determining this was not perfect.

Note that effectively, all probabilities and probability models are subjective. We must make assumptions about independence, exchangability, models whenever we analyze data.

It also I feel makes moot the idea of an objective analysis. We are always making assumptions when building models. Even if we agree on a model, say linear regression, different people may proceed differently. Two people could look at a residual plot and one may decide that it support the standard linear regression model and the other may decide that it supports some curvature and non-homogeneity of variance.

“Since all models are wrong the scientist cannot obtain a ‘correct’ one by excessive elaboration. ... Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity. Since all models are wrong the scientist must be alert to what is importantly wrong. It is inappropriate to be concerned about mice when there are tigers abroad.” – George E. P. Box, 1976

”All models are wrong but some are useful” – George E. P. Box



Single Parameter Models

- Binomial model
- Prior choice - conjugate vs non-conjugate priors
- Summarizing the posterior
- Sensitivity analysis
- Prediction
- Normal, Poisson, and Exponential models

Binomial Model

- Punxsutawney Phil and Wiarton Willie from first class

- Observe Willie for n years and observe y, the number of times the Willie correctly predicts winters finish.

- Assume that each year is independent and the outcome each year is a Bernoulli trial with success probability π.

- This implies that $y|\pi \sim Bin(n, \pi)$

- The sample size n is fixed and the success probability π can be any number in $[0, 1]$.

- The measurement model is

$$p(y|\pi) = \binom{n}{y} \pi^y (1-\pi)^{n-y}; \quad \pi \in [0,1]$$

  Since n is fixed in this analysis, we'll drop it as a conditioning argument for ease of notation.

- Want to make inference on π given y and n.

- Need a prior distribution $p(\pi)$ for π.

- Bayes' choice: $\pi \sim U(0,1)$

$$p(\pi) = \begin{cases} 1 & 0 \leq \pi \leq 1 \\ 0 & \text{Otherwise} \end{cases}$$

- So the Bayes' Rule gives

$$p(y, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

$$p(y) = \int_0^1 \binom{n}{y} \pi^y (1 - \pi)^{n-y} d\pi = \frac{1}{n+1}$$

$$p(\pi|y) = (n+1) \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

In the Wiarton Willie case, $n = 41$ and $y = 37$. So the posterior density is

$$p(\pi|y) = 42 \binom{41}{37} \pi^{37} (1 - \pi)^4 = \frac{42!}{37!4!} \pi^{37} (1 - \pi)^4$$

- Where did $p(y) = \frac{1}{n+1}$ come from?

- It is often easier to deal with the proportional form of Bayes' Rule

$$p(\pi|y) \propto p(\pi)p(y|\pi)$$

so

$$p(\pi|y) \propto 1 \times \pi^y(1-\pi)^{n-y}$$

This is proportional to the density of a $Beta(y+1, n-y+1)$ distribution.

- The PDF for the for the beta distribution ($Beta(\alpha, \beta)$) is

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

- In the example last class the properties of the posterior distribution used the fact that for a $Beta(\alpha, \beta)$ RV

$$
\begin{aligned}
E[X] &= \frac{\alpha}{\alpha + \beta} \\
\mathrm{Var}(X) &= \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \\
\mathrm{Mode}(X) &= \frac{\alpha - 1}{\alpha + \beta - 2}
\end{aligned}
$$