

STAT3016/4116/7016

Introduction to Bayesian Data Analysis

RSFAS, College of Business and Economics, ANU

Discussion

- ▶ What is the Bayesian approach to statistics?
- ▶ What are the key features of Bayesian methods?
- ▶ Why would you choose to use Bayesian methods (vs frequentist methods)?

↓ . PRIOR \times LIKELIHOOD \propto POSTERIOR
(of data)

$P(\theta) \times p(y|\theta) \propto P(\theta|y) \rightarrow$ goal of Bayesian inference

Thinking like a Bayesian ^{each team} ^{likelihood} $p(y|\theta)$ $\rightarrow p(\theta|y)$ $(y|\theta \sim \text{Pois}(\theta))$

Frequentist: the prob is fixed.

1. Manchester United is playing Chelsea in their next Premier League match. How likely is it that Man U will beat Chelsea by at least 2 goals? $p(\tilde{y}_{mu} - \tilde{y}_c > 2)$ posterior predictive dist. $p(\tilde{y}|y)$
2. A patient requires surgery and needs to select a hospital to go to for the operation. There are 5 hospitals to choose from. The following surgical mortality data is available from each hospital.

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

No. surgeries	521	878	676	421	553
No. Deaths	0	0	10	18	9

Rank the hospitals based on surgical mortality performance.

But there is also no guarantee that the death rate is fixed.

Bayesian vs Frequentist

Example 1 : Construct a 95% interval estimate for the population mean μ given observed data y_1, \dots, y_n

- ▶ Frequentist confidence interval: $\bar{y} \pm 1.96\sqrt{s^2/n} \rightarrow$
Interpretation?? "repeated sampling"/not a probability statement
- ▶ Bayesian probability interval: $Pr(\mu_L < \mu < \mu_u | \mathbf{y}) = 0.95$
Explicit use of probability to quantify uncertainty in statistical inference.
sig. level α vs. p-value.

Example 2: Hypothesis testing -what are the decision rules in a frequentist analysis? How do these compare to the decision rules in a Bayesian analysis?

can quantify uncertainty.


Bayesian Learning - fundamentals and notation

Statistics is all about learning the characteristics of a population from a subset of units of that population.

Statistical inference - draw conclusions from numerical data about quantities that are not observed (both future observations of a process; and parameters that govern the hypothetical process which generates the observed data).

- ▶ Let θ be the unknown parameters of the population
- ▶ Let y be the data we collect on the population
- ▶ Let \tilde{y} be predictions of unobserved data.

Once we have collected y , we can use this information to decrease our uncertainty on θ .

 Quantifying uncertainty on θ and \tilde{y} is the purpose of Bayesian inference.

Bayesian Learning - fundamentals and notation

- ▶ The **sample space** \mathcal{Y} is the set of all possible data sets from which a single data set y will result.
- ▶ The **parameter space** Θ is the set of all possible parameter values

What values of $\theta \in \Theta$ are most likely??? \rightarrow let's try and make some probability statements. *derive*

- ▶ *Prior distribution* : $p(\theta)$
- ▶ *Sampling model*: $p(y|\theta)$
- ▶ *Posterior distribution*: $p(\theta|y)$

our target: ① analytic form
or
② simulate the data

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

The posterior distribution enables us to see how our beliefs change after seeing new information. That is, we can decrease our uncertainty on the true value of θ .

Bayesian Learning - fundamentals and notation

The primary task of Bayesian inference is to develop the model $p(\theta, y)$ (a full probability model) and perform the computations to summarize $p(\theta|y)$ in appropriate ways. We also need to evaluate the model fit, that is, the assumptions underlying $p(\theta, y)$.

Example I - estimating the probability of a rare event

Suppose we are interested in the prevalence of an infectious disease in a small city. A small random sample of 20 individuals from the city will be checked for infection.

- ▶ What is θ ? What is Θ ?
- ▶ What is y ? What is \mathcal{Y} ?

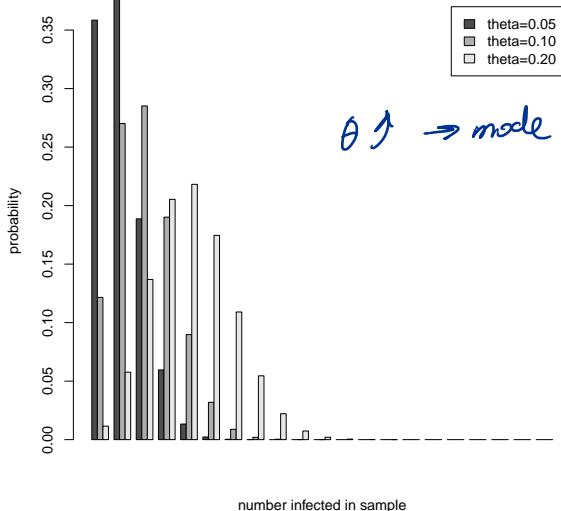
. infection rate , 0 to 1

. 20 people , 0 to 20.

Example I - estimating the probability of a rare event

What would be a reasonable sampling model for Y ??

Binomial sampling



$\theta \uparrow \rightarrow$ mode shifts to right

Example I - estimating the probability of a rare event

- Beta prior*
- ▶ What would be a reasonable choice for the prior distribution $p(\theta)$?

Suppose other studies suggest that the infection rate is between 0.05 and 0.20, with an average prevalence rate of 0.10. Let's incorporate this prior information into our prior distribution.

A mathematically convenient choice of prior (we will see why later) is a beta distribution.

For this case study, let $\theta \sim \text{beta}(2, 20)$. \rightarrow

$$E[\theta] = 0.09; \text{Mode}[\theta] = 0.05; \Pr(\theta < 0.10) = 0.64;$$

$$\Pr(0.05 < \theta < 0.20) = 0.66;$$

- ▶ **What is the posterior distribution of θ ?**
- ▶ Suppose $Y=0$. How are our beliefs on the prevalence rate updated?

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

$$X \sim \text{Beta}(a, b)$$

$$EX = \frac{a}{a+b}$$

$$\text{Var} X = \frac{ab}{(a+b)^2(a+b+1)}$$

$$\theta \sim \text{Beta}(a, b)$$

$$P(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \theta^{a-1} (1-\theta)^{b-1}$$

$$P(\theta|y) = P(\theta) P(y|\theta) / P(y) \sim P(\theta) P(y|\theta)$$

$$\sim \theta^{a-1} (1-\theta)^{b+n-\sum y_i-1}$$

$$\sim \text{Beta}(a + \sum y_i, b + n - \sum y_i)$$



prior parameters

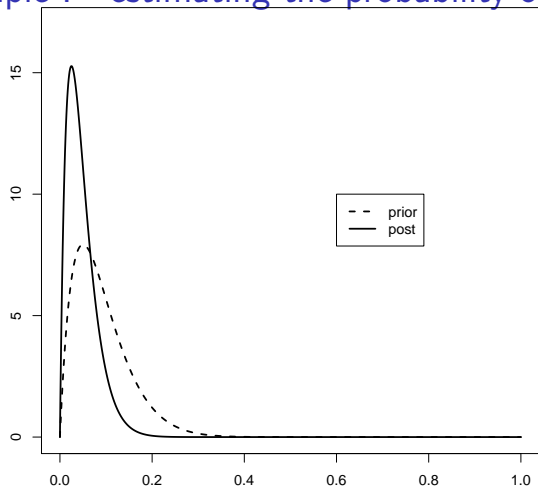
Example:

if $\theta \sim U(0.05, 0.20)$

$P(\theta|y) \sim \text{truncated}$

$\text{Beta}(1 + \sum y_i, 1 + n - \sum y_i)$

Example I - estimating the probability of a rare event



$$E[\theta] = 0.09; \text{Mode}[\theta] = 0.05; Pr(\theta < 0.10) = 0.64;$$

$$Pr(0.05 < \theta < 0.20) = 0.66;$$

$$E[\theta|Y = 0] = 0.048; \text{Mode}[\theta|Y = 0] = 0.025;$$

$$Pr(\theta < 0.10|Y = 0) = 0.93; Pr(0.05 < \theta < 0.20|Y = 0) = 0.38;$$

Example I - estimating the probability of a rare event

Exercise 1 - what is your posterior distribution on θ if you assume the prior distribution on θ is a uniform distribution on the interval $(0.05, 0.20)$.

Example I - estimating the probability of a rare event

Note the following:

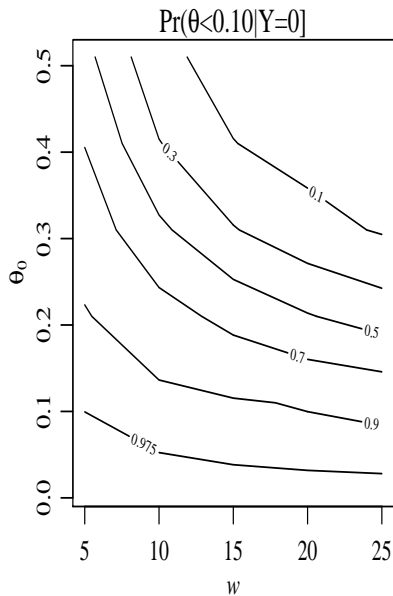
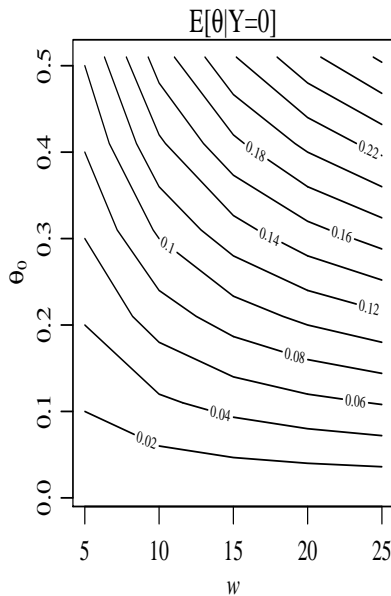
posterior mean is weighted sum of sample mean & prior guess of prior mean

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

where $\theta_0 = \frac{a}{a+b}$ is the prior expectation of θ , and $w = a + b$ is like the amount of prior information.

Changing θ_0 and w means changing your prior beliefs and we can assess how the posterior information is affected by differences in prior opinion.

Example I - estimating the probability of a rare event



Bayesian inference

Likelihood and odds ratio

- ▶ $p(y|\theta)$: likelihood function (data y only affects posterior through the likelihood function)
- ▶ Likelihood principle: any two probability models $p(y|\theta)$ that have the same likelihood function yield the same inference for θ .
- ▶ Posterior odds:

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

Interpretation?

θ is the prob of some event A composite hypothesis
 $H_0: \theta > 0.5$ vs $H_A: \theta \leq 0.5$
$$\frac{P(\theta > 0.5 | y)}{P(\theta \leq 0.5 | y)}$$

Bayesian inference

Prediction - We can also make probability statements using $p(\tilde{y}|y)$

Prior predictive/marginal distribution:

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

(before the data y are observed)

After the data is observed, predict an unknown but observable quantity \tilde{y} from the same process. What will it be? We can describe the distribution of \tilde{y} with the posterior predictive distribution:

test if we have a correct sampling distribution.

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y) d\theta \\ &= \int p(\tilde{y}|\theta, y) p(\theta|y) d\theta \\ &= \int p(\tilde{y}|\theta) p(\theta|y) d\theta \end{aligned}$$

Example - estimating the probability of a rare event

For the infectious disease example, do the following: **Exercise 2** - compare the prior predictive distribution from assuming the prior is $\theta \sim \text{Beta}(2, 20)$ versus $\theta \sim \text{Unif}(0.05, 0.20)$

Exercise 3 - compare the posterior predictive distribution probability that $Y=0$ from assuming the prior is $\theta \sim \text{Beta}(2, 20)$ versus $\theta \sim \text{Unif}(0.05, 0.20)$

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

$$\begin{aligned} p(\tilde{y}=1|y=0) &= \int_0^1 p(\tilde{y}|\theta)p(\theta|y) d\theta \\ &= K \int_0^1 \theta^{\tilde{y}} (1-\theta)^{1-\tilde{y}} \theta^{a-1} (1-\theta)^{b+n-1} d\theta \end{aligned}$$

Example I - estimating the probability of a rare event

Comparison to non-Bayesian methods

- ▶ What would be a standard estimate of θ using the sampled data y ?
- ▶ Provide a 95% confidence interval for θ . $\hat{\theta} \pm 1.96 \sqrt{\hat{\theta}(1-\hat{\theta})/n}$
- ▶ How does your interval estimate perform when $y=0$? What if n is small???

where

$$\hat{\theta} = \frac{n}{n+4} \bar{y} + \frac{4}{n+4} \cdot \frac{1}{2}$$

General estimation of a population mean

- ▶ If the sample size is large enough, the sample mean \bar{y} is generally a reliable estimate of the population mean. What about for small n ? Can we still use the sample mean to obtain a precise estimate of the population mean θ ?
- ▶ Consider the following estimator:

$$\hat{\theta} = \frac{n}{n+w}\bar{y} + \frac{w}{n+w}\theta_0$$

θ_0 = “best guess”; w = degree of confidence in the guess. How does this estimator perform when n is large and when n is small??

Key Point: When the sample size is small, Bayesian methods allow us to combine the data with prior information to stabilize our estimation of θ .

Building a predictive model - diabetes case study

Case Study: Our task is to build a predictive model of diabetes progression as a function of 64 baseline exploratory variables such as age, sex, and body mass index.

We will first estimate the parameters in a regression model using a “training” dataset consisting of measurements from 342 patients. We will then evaluate the predictive performance of the estimated regression model using a separate “test” dataset of 100 patients.

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_{64} x_{i,64} + \epsilon_i$$

$$\epsilon_i \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Building a predictive model - diabetes case study

Prior distribution

- ▶ Prior on what? $\beta \& \sigma$
- ▶ What would be your prior beliefs on the parameters in the model? How would you specify the prior distribution to capture your prior beliefs? $\beta=0$

Posterior distribution

- ▶ Given data $\mathbf{y} = (y_1, y_2, \dots, y_{342})$ and $\mathbf{X} = (x_1, x_2, \dots, x_{342})$, the joint posterior distribution $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$ can be computed.
- ▶ What posterior quantities may we be interested in??

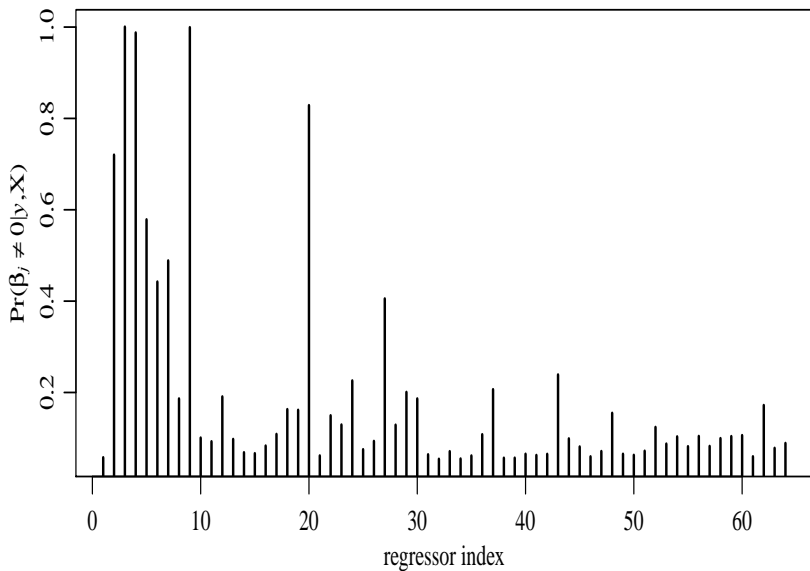
$$p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) \propto p(\beta, \sigma^2) p(\mathbf{y} | \mathbf{X}, \beta, \sigma^2)$$
$$p(\beta_j \neq 0 | \mathbf{y}) \quad \#$$

Building a predictive model - diabetes case study

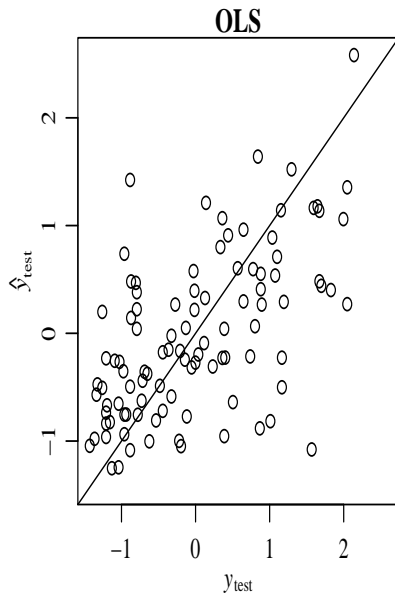
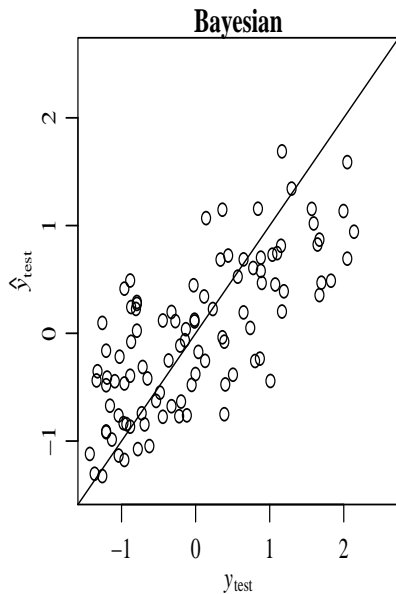
Predictive performance and comparison to non-Bayesian methods

- ▶ Using $p(\beta, \sigma^2 | \mathbf{y}, \mathbf{X})$ how might we evaluate the performance of the model?
- ▶ How can we compare the Bayesian approach to Ordinary Least Squares regression? (Recall: $\hat{\beta}_{ols} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$)
- ▶ In this example, the OLS methods performs poorly due to sparsity for some explanatory variable values. The Bayesian approach is one way to get around this. Another method is the “lasso” which minimises the sum of squared residuals with a penalty term to make some element of $\beta = 0$.

Building a predictive model - diabetes case study



Building a predictive model - diabetes case study



Summary

Bayesian approach provides:

- ▶ estimators that work for small and large sample sizes
- ▶ structure to incorporate prior information and quantify uncertainty in posterior beliefs
- ▶ methods for generating statistical procedures to complicated problems eg dealing with **sparse data sets**.