Elementary Bayesian Statistics

Bayesian Statistics Seminar Series

S.Wotherspoon [1,2]   B. Raymond [1]

[1]Australian Antarctic Division

[2]Institute of Marine and Antartic Studies

Aug, 2013

Rough Schedule

- Introduction
- Simple mathematical examples
- WinBUGs
- Principles of Markov Chain Monte Carlo
- Advanced techniques
- Applications

This slide intentionally left blank.

Notations

$p(x, y)$ – the joint distribution of x and y describes how x and y vary together.

$p(x|y)$ – the conditional distribution of x given y describes how x varies for a given value of y.

$p(x)$ – the marginal distribution of x describes how x varies averaged over y.

Frequentist Probability – Probability is the 'long run frequency of occurrence'. This is the basis of classical statistics.

Subjective Probability – Probability is a measure of 'strength of belief'. As it based on beliefs, it is inherently subjective.

In classical statistics, parameters are fixed numbers. In the Bayesian paradigm, parameters are random.

Key Elements

Knowledge is represented as probability distributions.

Prior Distribution $p(\theta)$ – Represents our knowledge of the parameters θ before any data is observed.

Likelihood $p(y|\theta)$ – The distribution of the data y for given parameters θ – the likelihood is a probabilistic model of the data collection process.

Posterior Distribution $p(\theta|y)$ – Represents our knowledge of the parameters after the data is observed.

The posterior $p(\theta|y)$ is determined from the likelihood $p(y|\theta)$ and prior $p(\theta)$ by Bayes' rule

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

For fixed y the denominator is just a constant, so

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

Suppose we toss a coin N times and record the numbers of heads y and wish to estimate $\pi = \Pr(H)$. If we adopt a (conjugate) Beta prior

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

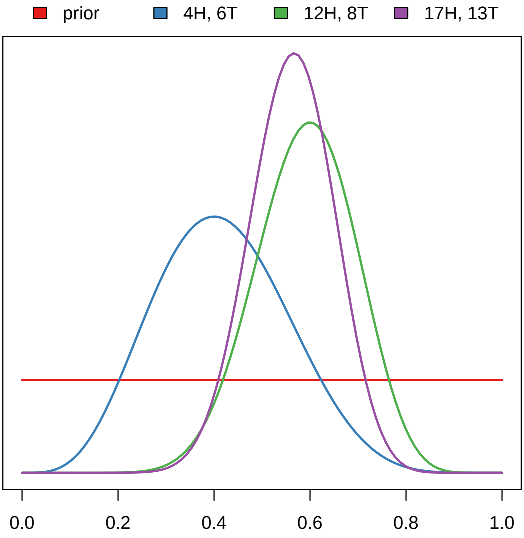
the likelihood is Binomial

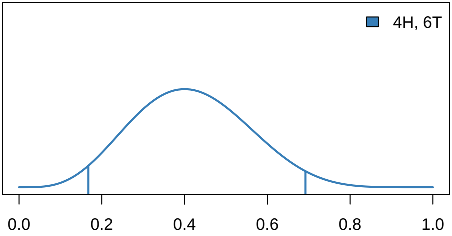
$$y|\pi \sim \text{Binomial}(N, \pi)$$

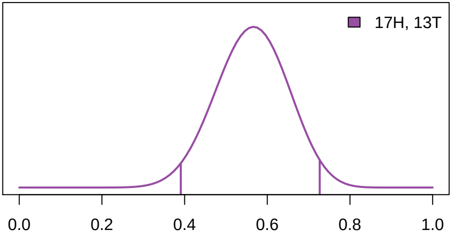
and by Bayes' rule

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

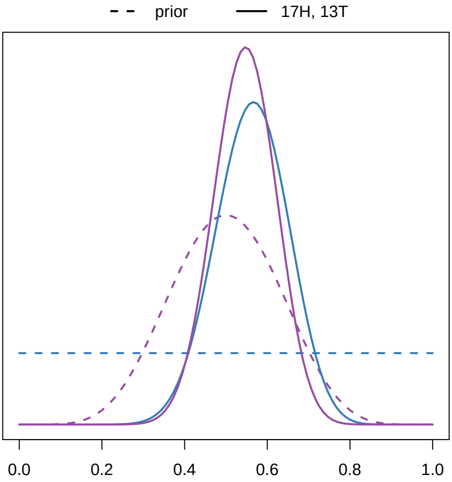
Coin Example





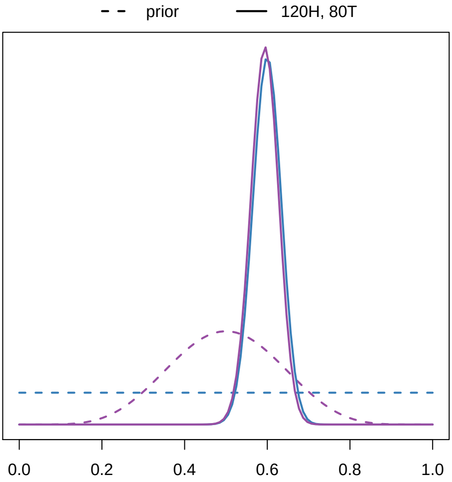


As the volume of data increases the posterior becomes more concentrated.



The choice of prior influences the posterior.

Large Samples



As the volume of data increases, the impact of the prior washes out.

Priors

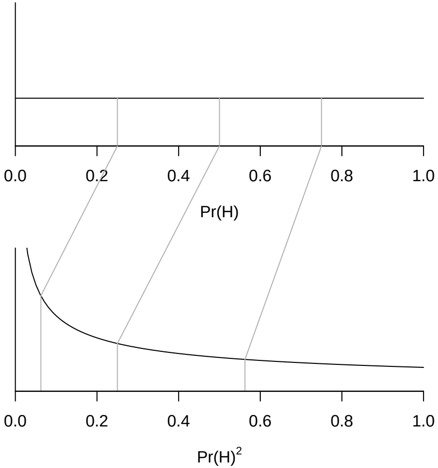
Informative Prior – Reflects the current state of knowledge of the model parameters.

Non-informative Prior – Intended to reflect a state of ignorance of the model parameters.

Conjugate Prior – A prior chosen for its mathematical expediency.

Improper Prior – A ‘prior’ that is not technically a distribution.

Jeffereys’ Prior – A prior that is (locally) invariant under a reparametrization of the model.



"Noninformative" is a slippery concept.

Improper Priors

For a typical regression problem, we would wish to adopt a non-informative priors for the coefficients of the form

$$\beta_i \sim \mathrm{U}(-\infty, \infty)$$

This prior is improper – there is no $\mathrm{U}(-\infty, \infty)$ distribution.

Instead we assume

$$p(\beta_i) \propto 1$$

and the missing constant of proportionality is eliminated by Bayes' rule.

In classical statistics, if $[L, U]$ is a 95% confidence interval for μ, we cannot write

$$\Pr(L < \mu < U) = 0.95$$

because none of L, U, or μ are random.

In the Bayesian paradigm, μ is random and confidence intervals have a natural interpretation.

We can test

$$\mathrm{H}_0 : \pi < \frac{1}{2}$$

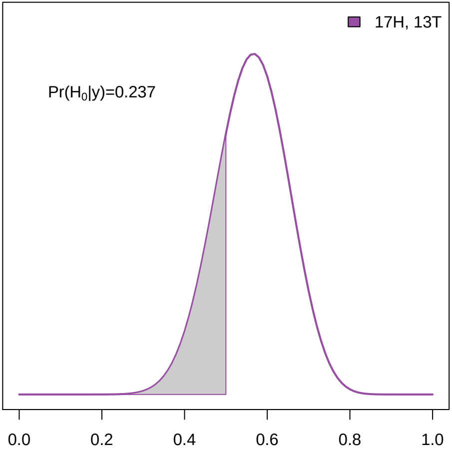
against the alternative

$$\mathrm{H}_1 : \pi > \frac{1}{2}$$

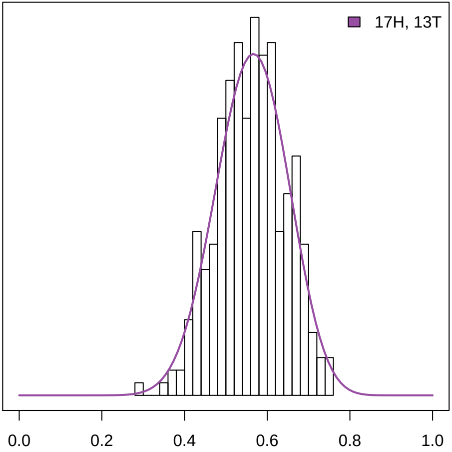
by computing the posterior probability

$$\Pr(\mathrm{H}_0|y) = \int_0^{\frac{1}{2}} p(\pi|y) d\pi.$$

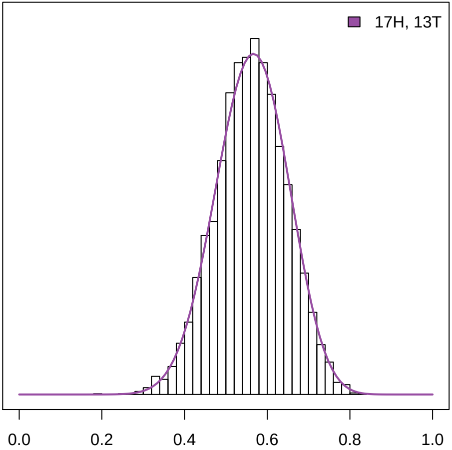
Moreover, this provides evidence in support of H_0, in contrast to the classical hypothesis test which is phrased in terms of evidence against H_0.



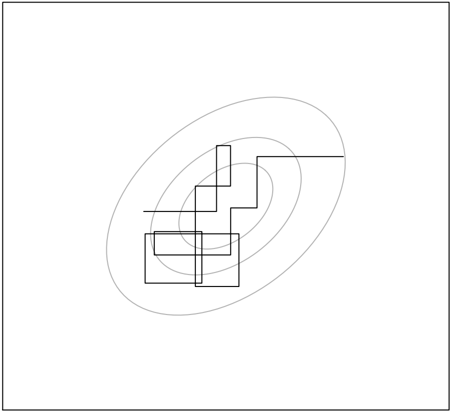
Composite tests correspond to simple probability statements.



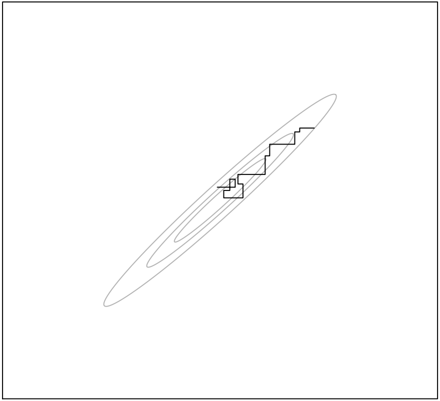
MCMC draws a random sample from the posterior. The properties of the sample approximate the properties of the posterior.



Increasing the size of the sample improves the accuracy of the approximation.



Sample from the conditional distribution to update parameters in blocks, one block at a time.



Strong correlation leads to poor mixing.

Binomial Problem

```
model {
  ## Likelihood
  for(i in 1:N) {
    y[i] ~ dbin(pi,n[i])
  }
  ## Prior
  pi ~ dbeta(a,b)
}
```

Observations are Binomially distributed

$$y_i|\pi \sim \text{Binomial}(n_i, \pi)$$

and we adopt a Beta prior

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

```
model {
  ## Likelihood
  for(i in 1:N) {
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- b0+b1*x1[i]+b2*x2[i]
  }

  ## Prior
  b0 ~ dnorm(0,0.01)
  b1 ~ dnorm(0,0.01)
  b2 ~ dnorm(0,0.01)
  tau ~ dgamma(0.01,0.01)
}
```

Observations are Normally distributed about a mean that is a function of covariates

$$y_i|\beta_i,\tau \sim \mathrm{N}(\mu_i,\tau)$$
$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

and we adopt diffuse priors

$$\beta_i \sim \mathrm{N}(0, 0.01)$$
$$\tau \sim \mathrm{Gamma}(0.01, 0.01)$$

Poisson GLM

```
model {
  ## Likelihood
  for(i in 1:N) {
    y[i] ~ dpois(mu[i])
    log(mu[i]) <- b0+b1*x[i]
  }

  ## Prior
  b0 ~ dnorm(0,0.01)
  b1 ~ dnorm(0,0.01)
}
```

Observations are Poisson distributed about a mean that is related to covariates through a link function

$$y_i|\beta_i \sim \text{Poisson}(\lambda_i)$$
$$\lambda_i = \beta_0 + \beta_1 x_i$$

and we adopt diffuse priors

$$\beta_i \sim \text{N}(0, 0.01)$$