

Likelihood principle

- In making inferences about a parameter θ after observing data y , all information about θ is contained in $[y|\theta]$
- Example: flip a coin and get 9 heads and 3 tails
 - Predetermined 12 flips -- $\text{Bin}(12, \theta)$
 - Flip until get 3 tails – $\text{NB}(3, \theta)$
- Since these are proportional, contain same information about θ !

Therefore did not need to know how experiment was conducted!

Later we'll see how to include this information

Inference about Functions of Parameters

- Inference regarding functions of parameters straightforward in a Bayesian framework
- Suppose $Y \sim B(n, p)$ and based on $Y = y$ we wish to make inference about the log-odds:

$$\eta = \ln \left(\frac{p}{1-p} \right)$$

Transformation of Variables

- Given $[y]$ what is $[w]$ where $w = f(y)$?
 - If y is discrete and f invertible

$$[w] = [f^{-1}(w)]$$

- If y is continuous and f invertible

$$[w] = [f^{-1}(w)] | J |$$

where J is the Jacobian of $y = f^{-1}(w)$

Binomial example

Logit transform

$$[y] \sim B(n, p) \quad [y] \propto p^y (1-p)^{n-y}$$

$$[p | y] \propto p^{\alpha+y-1} (1-p)^{\beta+n-y-1}$$

?

Take $v = \log\left(\frac{p}{1-p}\right)$

what is $[v]$, $[v | y]$?

Analytical approach

$$[\nu] = [\text{expit}(\nu)]_p \frac{\partial \text{expit}(\nu)}{\partial \nu}$$

$$\text{expit}(\nu) = \frac{e^\nu}{1 + e^\nu}$$

$$[\nu] \propto \left(\frac{e^\nu}{1 + e^\nu} \right)^y \left(1 - \left(\frac{e^\nu}{1 + e^\nu} \right) \right)^{n-y} \frac{-1}{(1 + e^{-y})^2}$$

Not real friendly looking!

Instead

- Draw a sample $\{p_i\}$ from $[p]$ and for each p_i compute

$$v_i = \log\left(\frac{p_p}{1 - p_i}\right)$$

- Use the sample of $\{v_i\}$ to make inference on $[v]$
- Identical approach works for $[v | y]$

Excel example

- Get beta parameters for prior or posterior
- Simulate p
- Transform
- [binomial_sim.xls](#)

Suppose we have a function of >1 parameter?

- If the parameters are conditionally independent, then get the separate marginal distributions

Models with >1 parameter

- Virtually every problem in statistics has more than one unknown
 - Usually interested in a subset of parameters
 - Nuisance parameters are those needed in the model to ensure proper inference but that are of little interest
- Multiparameter problems highlight the advantages of Bayesian inference
- Require the *marginal* distribution of the parameters of interest.

Averaging over the ‘nuisance’ parameters

- We have a posterior distribution
 - $[\theta|y]$; θ has 2 parts (θ_1, θ_2)
 - Interest just centers on 1 parameter (e.g., θ_1)
- We need the posterior distribution of θ_1 , averaged over the values of θ_2
- Get from joint posterior

$$[\theta_1, \theta_2 | y] \propto [y | \theta_1, \theta_2][\theta_1, \theta_2]$$

or

$$[\theta_1 | y] = \int [\theta_1, \theta_2 | y] d\theta_2$$

$$[\theta_1 | y] = \int \underbrace{[\theta_1 | \theta_2, y]}_{\text{Conditional on nuisance parameter}} \underbrace{[\theta_2 | y]}_{\text{weights}} d\theta_2$$

Conditional on nuisance parameter

weights

In practice....

- We rarely evaluate these integrals explicitly, BUT
- Suggests a strategy
 - First draw θ_2 from its conditional posterior
 - Then draw θ_1 conditional on θ_2 and data

Bayes theorem for >1 Parameter

- If Y depends on $\theta_1, \dots, \theta_k$ then:

$$[\theta_1, \dots, \theta_k | Y = y] \propto [Y | \theta_1, \dots, \theta_k][\theta_1, \dots, \theta_k]$$

- $[\theta_1, \dots, \theta_k | Y = y]$ is the *joint posterior* distribution

Marginal Distribution

- If a random vector Z can be partitioned into X and Y the distribution of X can be found by integrating (summing) over the components of Y :

$$[X] = \int [X, Y] dY$$

- $[X] = f_X(x)$ is the marginal distribution of X

Joint Posterior Inference

- Inference can be made using joint posteriors
 - Contour plots for pairs of variables
- Can also examine parameters one at a time using marginal posterior distributions

$$[\theta_1, \theta_2 | y] \propto [y | \theta_1, \theta_2][\theta_1, \theta_2]$$

$$[\theta_1 | y] = \int [\theta_1, \theta_2 | y] d\theta$$

Example

- A random sample drawn from a $N(\mu, 1/\tau)$ with both parameters are unknown

- First step:

$$[\mu, \tau | y] = [y | \tau, \mu][\mu, \tau]$$

- Take $[\mu, \tau]$ as a vague (improper) prior

- Leads to

$$[\mu, \tau | y] = \tau^{\frac{n+2}{2}} \exp\left(-\frac{\tau}{2}[(n-1)s^2 + n(\bar{y} - \mu)^2]\right)$$

- Decompose as

$$[\mu, \tau | y] = [\mu | \tau, y][\tau | y] =$$
$$N(\bar{y}, 1/\tau) \chi^2(n-1)/(ns^2)$$

- So: to get joint distribution

– Draw τ from $[\tau ns^2 | y] = \chi^2(n-1)$

– Given τ draw μ from

$$[\mu | \tau, y] = N(\bar{y}, (n\tau)^{-1})$$

- Marginals

– Get by averaging over nuisance parameter(s)

Implementation in Excel

- [normal_simxls.xls](#)
- Simulation has advantages even when we know the posterior distributions
 - Transformations easier
 - Multiple parameter inference easy

Exercise 3

- Normal with prior $[\tau, y] = 0.02$
- Data: $\bar{y} = 112., s^2 = 400, n = 150$
- Find by simulation the mean, variance, and .025, .5, 0.975 quantiles, and graph, the following
 - The marginal posterior distributions $[\mu | y], [\tau | y]$
 - The conditional distributions

$$[\mu | \tau = 0.01], [\mu | \tau = 0.10], [\mu | \tau = 1.0]$$

Hints

- (1) The marginal distribution of $[x|y]$ is gotten by averaging over all the simulated values of $[x|y][y]$
- (2) The conditional distribution of $[x|y]$ is gotten by specifying values for y and then averaging over $[x|y=y_0]$

Prior Predictive Distribution

- Inference about an unknown observable is called *prediction*
- Before the data have been collected we can predict what a new value would be using the *prior predictive distribution* (also called the *marginal distribution*)

Prior predictive (marginal) distribution

$$[y] = \int [y, \lambda] d\lambda = \int [y | \lambda] [\lambda] d\lambda$$

Posterior Predictive Distribution

- Inference about an unknown observable is called *prediction*
- After the data have been collected we can predict what a new value would be using the *posterior predictive distribution*:

$$\begin{aligned} [y^{new} | y] &= \int [y^{new}, \theta | y] d\theta \\ &= \int [y^{new} | \theta][\theta | y] d\theta \end{aligned}$$

Example

- After carrying out a binomial experiment with y successes in n trials we wish to predict the outcome on one more trial
 - Assume a $\text{Be}(\alpha, \beta)$ prior for p

$$\Pr(y^{new} = 1 | y) = \int_0^1 \Pr(y^{new} = 1)[p | y]dp$$

- $\Pr(y^{new} = 0) = 1 - \Pr(y^{new} = 1 | y)$ completes our distribution

Generally get this by simulation!

Easy case: uniform prior on p

- Posterior predictive distribution

$$[y^{new} = 1 | y] = \int_0^1 [y^{new} = 1 | p][p | y] dp$$

- But for uniform

$$[y^{new} = 1 | p] = p$$

- Leads to

$$[y^{new} = 1 | y] = \int_0^1 p[p | y] dp = E(p) = \frac{y+1}{n+2}$$

Harder cases: simulate

- Draw p from $[p | y]$
- Given p draw y^{new} from $[y^{new} | p]$

Exercise 4

- Prior $[p] = Be(20,20)$
- Data $n=100$ $y=60$ successes
- Get posterior distribution of getting 5 successes in 10 new trials
 - Hint 1: use beta to simulate $[p | y]$
 - Hint 2: $[y=5 \text{ successes} | p, n]$ follows a binomial distribution given p and n

More on priors

- Vague or non informative priors
- Proper versus improper
- Jeffrey's priors
- What is “vague” and why do we care?

Noninformative Priors

- Often (usually) want a prior that plays minimal role in inference
- “Reference prior distributions”
 - Density = “vague”, “flat”, “diffuse” or “noninformative”
 - Loosely: “flat”
- Let the data speak for themselves
 - Inference driven by data

Proper and Improper

- Suppose Y is normal with mean μ and variance σ^2 (known)

- $[\mu] = N(\mu_0, 1/\tau_0)$

$$[\mu | \bar{y}] = N\left(\frac{n\tau\bar{y} + \tau_0\mu_0}{n\tau + \tau_0}, \frac{1}{n\tau + \tau_0}\right)$$

- If prior precision is small relative to the data precision it is as if $\tau_0 = 0$ or $1/\tau_0 = \infty$

Proper and Improper

- If $[\mu | \bar{y}] \sim N(\bar{y}, 1./n\tau)$ this is same as saying:

$$[\mu | \bar{y}] = [\bar{y} | \mu] \times k$$

- But $[\mu] \sim N(\mu_0, \infty)$ has infinite integral
 - *Improper*
 - A prior is *proper* if (i) it does not depend on the data and (ii) it integrates to 1.0
 - If (i) holds but it integrates to $k \neq 1.0$ then $[\cdot]$ is said to be an *unnormalised density*

Are Improper Priors Bad?

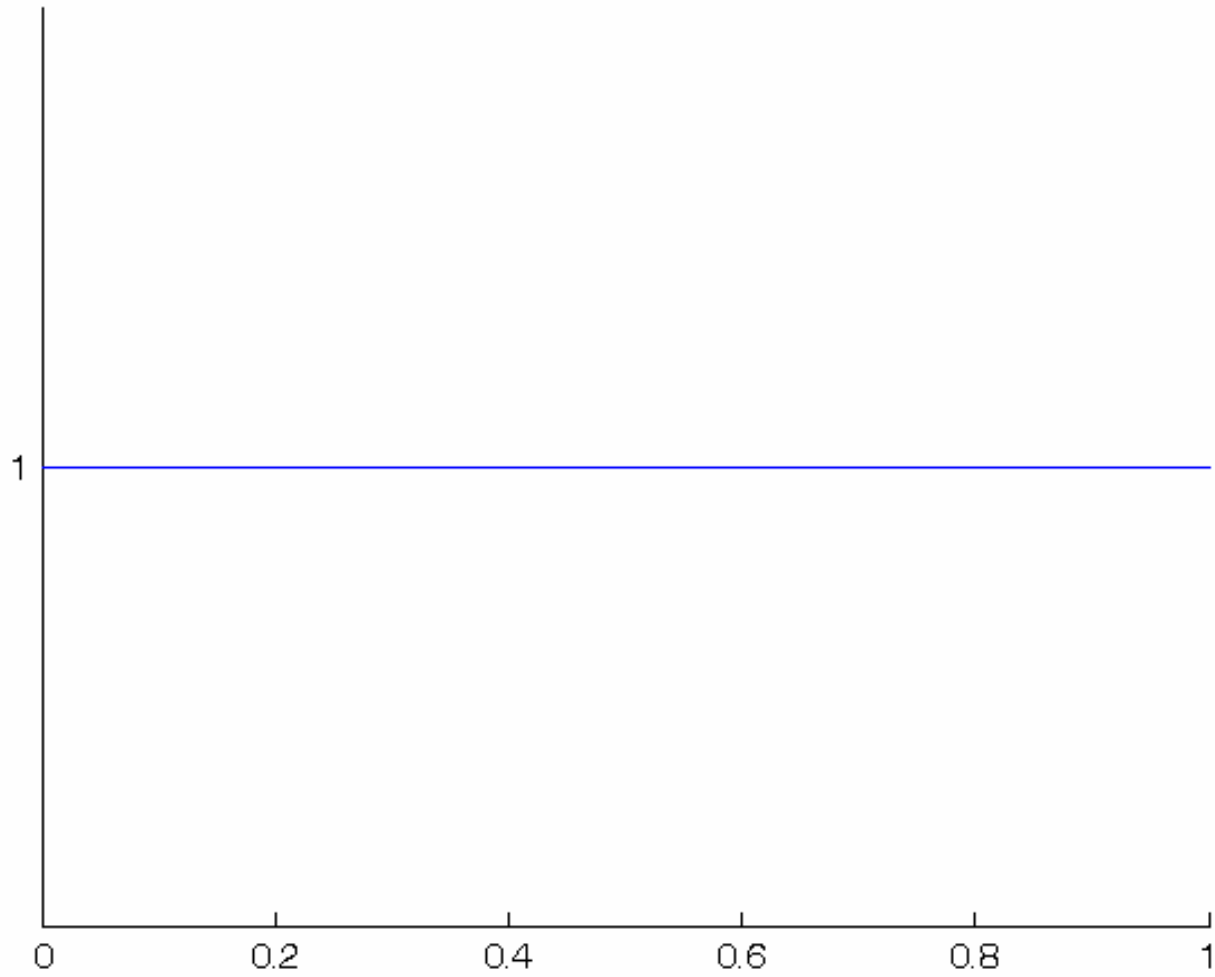
- Improper priors can (but not always) lead to proper posterior distributions
 - If likelihood dominates prior, no problem
- Proper priors *always* lead to proper posterior distributions

Flat on One Scale not on Another

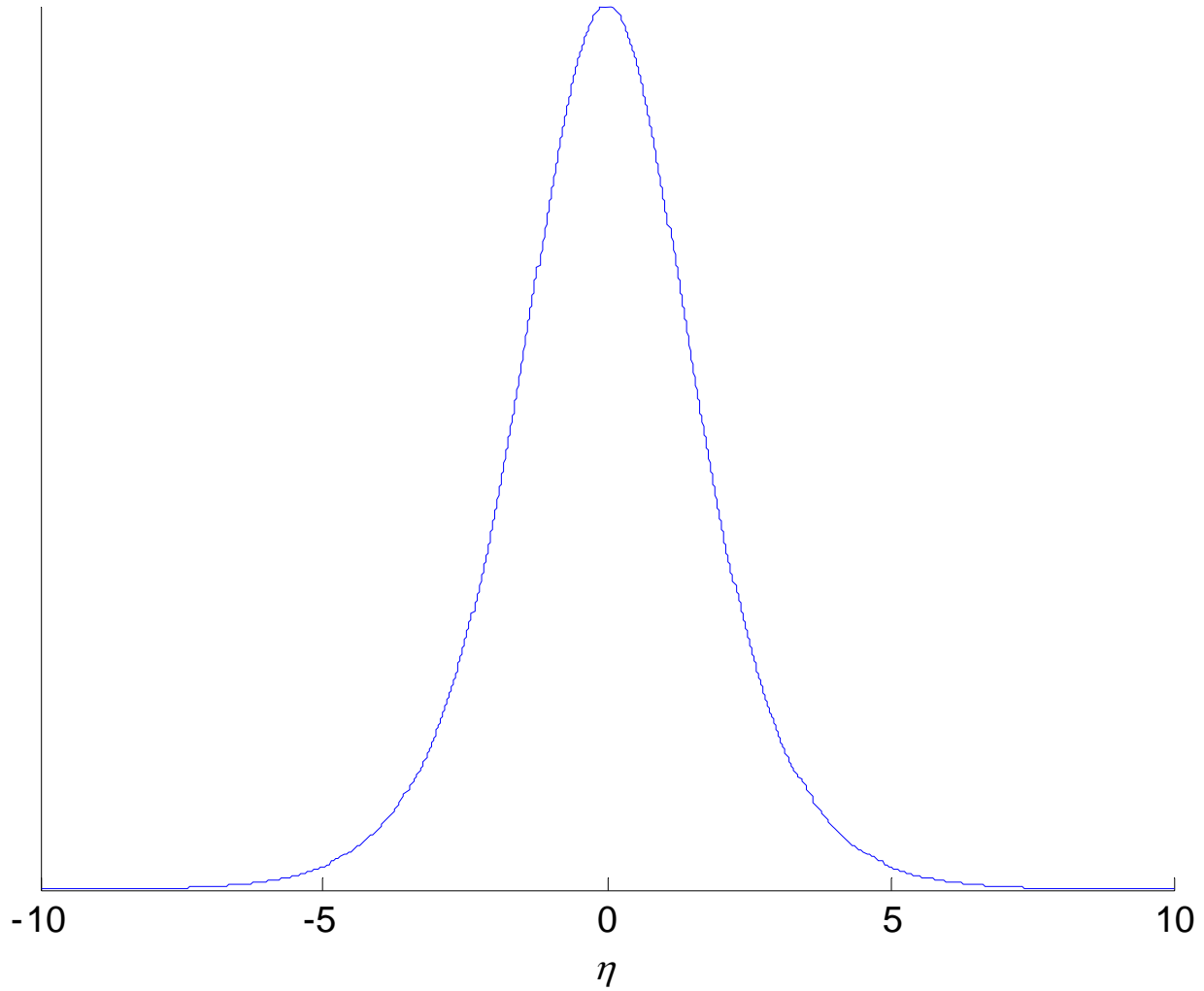
- For $[Y] = B(n, p)$ the prior $[p] = \text{Be}(1, 1)$ is constant
- Suppose we are interested in

$$\eta = \ln \left(\frac{p}{1-p} \right)$$

$$[p] = \text{Be}(1, 1) = \text{U}(0, 1)$$



$[\eta = \text{logit}(p) \text{ when } [p] = \text{Be}(1, 1)]$



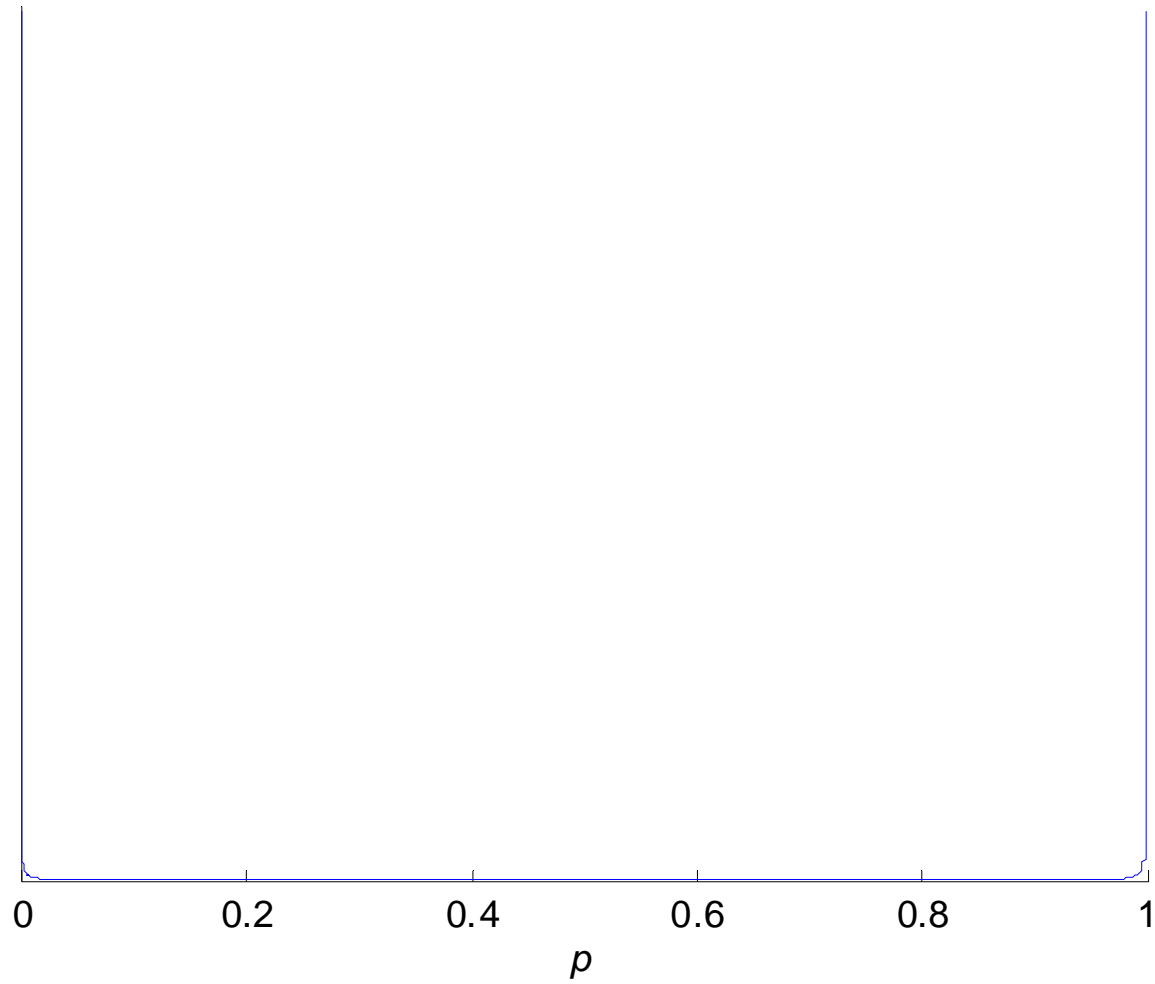
- Logit-Normal: if $[\eta] = \mathbf{N}(\psi, 1/\tau)$

$$[p] = \sqrt{\frac{\tau}{2\pi}} \exp\left(\frac{-\tau}{2} (\text{logit}(p) - \psi)^2\right) \frac{1}{p(1-p)}$$

(verify as exercise)

where $p = \frac{e^\eta}{1 + e^\eta}$

$[p = \text{expit}(\eta)]$ when $[\eta] = N(0, 1/0.00001)$



Jeffrey's Prior

- Jeffrey's principle: *any rule for determining the prior density should yield an equivalent result if applied to the transformed parameter*
- Suppose $Y \sim B(n, p)$ and based on $Y = y$ we wish to make inference about the log-odds. Two approaches:
 - (1) Find $[p | y]$ then use transformation of variables method to find $[\text{logit}(p) | y]$

(2) Use:

$$[\text{logit}(p) | y] \propto [y | \text{logit}(p)] \times [\text{logit}(p)]$$

- The two methods do not usually agree
- Jeffrey's method is to find the prior for which they do agree:

– Jeffrey's prior $[\theta] \propto (I(\theta))^{1/2}$ where $I(\theta)$ is the *Fisher information* for θ

$$I(\theta) = -E \left[\frac{d^2 \ln(L(\theta; y))}{d\theta^2} \right]$$

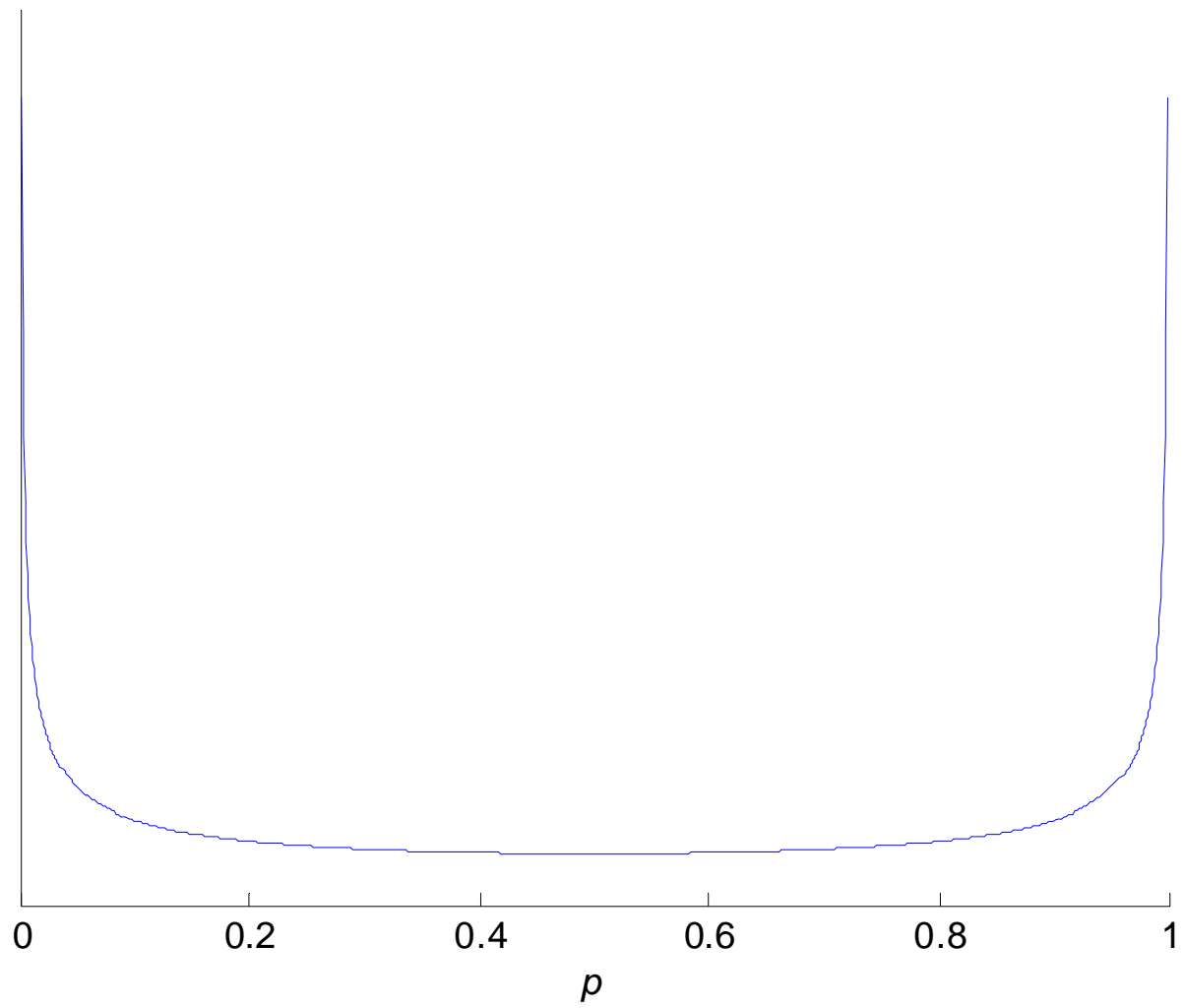
Binomial Example

- From math-stats $I(\theta) = \frac{n}{p(1-p)}$

hence $[p] \propto p^{-1/2}(1-p)^{-1/2}$

– Prior is $\text{Be}(1/2, 1/2)$

$[\rho] = \text{Be}(1/2, 1/2)$



Multivariate Problems

- Jeffrey's principle can be extended to multiparameter problems
 - Controversial as assuming independent priors for the components can give different results than are obtained by Jeffrey's principle
- If the number of parameters is large better to use hierarchical models than vague priors for the components (later in course)

General Comments

- Searching for a prior that is always vague is misguided
 - If the prior matters you don't have enough data
- If unwilling to use an informative prior, then use vague priors that are convenient but check:
 - That posterior is proper
 - Sensitivity to choice of prior. If sample is large then no problem.
- WinBUGS does not allow improper priors

More multi-parameter models

- (Back to the) Normal
- Multinomial
- Multivariate Normal

Normal

- Conjugate prior

$$\begin{aligned} [\mu, \tau] &= [\mu \mid \tau][\tau] \\ &= N\left(\mu_0, \frac{1}{\tau K_0}\right) \times \frac{\chi^2(v_0)}{v_0 \tau_0^{-1}} \end{aligned}$$

Joint posterior

$$[\mu, \tau | y] = N - Inv\chi^2(\mu_1, (\kappa_1\tau_1)^{-1}, \nu_1\tau_1^{-1})$$

$$\mu_1 = \frac{\kappa_0}{\kappa_0 + n} \mu_0 + \frac{n}{\kappa_0 + n} \bar{y}$$

$$\kappa_1 = \kappa_0 + n,$$

$$\nu_1 = \nu_0 + n,$$

$$\frac{\nu_1}{\tau_1} = \frac{\nu_0}{\tau_0} + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n} (\bar{y} - \mu_0)^2$$

Marginals

$$[\mu \mid \tau, y] = N(\mu_1, (\tau \kappa_1)^{-1})$$

$$[\tau \mid y] = \frac{\chi^2(v_1)}{v_1 \tau_1^{-1}}$$

For joint inference:

-- Draw τ from $[\tau \mid y]$

---- Draw μ from $[\mu \mid \tau, y]$

Multinomial distribution

- Generalize binomial to $k > 2$ possible outcomes on each trial
- Now are $k-1$ probabilities

$$[y \mid p_1, p_2, \dots, p_k] = \prod_{i=1}^k p_i^{y_i}$$

$$p_k = 1 - \sum_{i=1}^{k-1} p_i \quad \sum_{i=1}^k y_i = n$$

Conjugate prior

- Multivariate generalization of Beta called Dirichlet distribution

$$[p_1, p_2, \dots, p_k] \propto \prod_{i=1}^k p_i^{\alpha_i - 1}$$

- When $k=2$ reduces to Beta

$$p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} = p^{\alpha - 1} (1 - p)^{\beta - 1}$$

Posterior

$$[p_1, p_2, \dots, p_k \mid y] \propto \prod_{i=1}^k p_i^{y_i} \prod_{i=1}^k p_i^{\alpha_i - 1} = p_i^{\alpha_i + y_i - 1}$$

- Dirichlet with parameters $\alpha_i + y_i$

Multivariate normal

- Vector of k components

$$\underline{y} = (y_1, y_2, \dots, y_k)$$

$$[\underline{y} \mid \underline{\mu}, \Sigma] = MVN(\underline{\mu}, \Sigma)$$

Priors- posteriors

- Conjugate priors/posteriors

$[\Sigma], [\Sigma | y] - \textit{Wishart}$

$[\underline{\mu} | \Sigma], [\underline{\mu} | \Sigma, y] - \textit{MVnormal}$