

# ANOVA: ANALYSIS OF VALUE

IS YOUR RESEARCH WORTH ANYTHING?

Developed in 1912 by geneticist R.A. Fisher, the Analysis of Value is a powerful statistical tool designed to test the significance of one's work.



am i  
wasting  
my time?

Significance is determined by comparing one's research with the **Dull Hypothesis**:

$$H_0: \mu_1 = \mu_2 ?$$

where,

$H_0$  : the Dull Hypothesis

$\mu_1$  : significance of your research

$\mu_2$  : significance of a monkey typing randomly on a typewriter in a forest where no one hears it.

The test involves computation of the  $F'd$  ratio:

$$F'd = \frac{\text{sum}(\text{people who care about your research})}{\text{world population}}$$

This ratio is compared to the F distribution with  $I-1$ ,  $N_T$  degrees of freedom to determine a  $p$ (in your pants) value. A low  $p$ (in your pants) value means you're on to something good (though statistically improbable).

## Type I/II Errors

The Analysis of Value must be used carefully to avoid the following two types of errors:

Type I: You incorrectly believe your research is not Dull.

Type II: No conclusions can be made. Good luck graduating.

Of course, this test assumes both Independence and Normality on your part, neither of which is likely true, which means *it's not your problem*.

# 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  $\mu$  and variance  $\sigma^2$  (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?

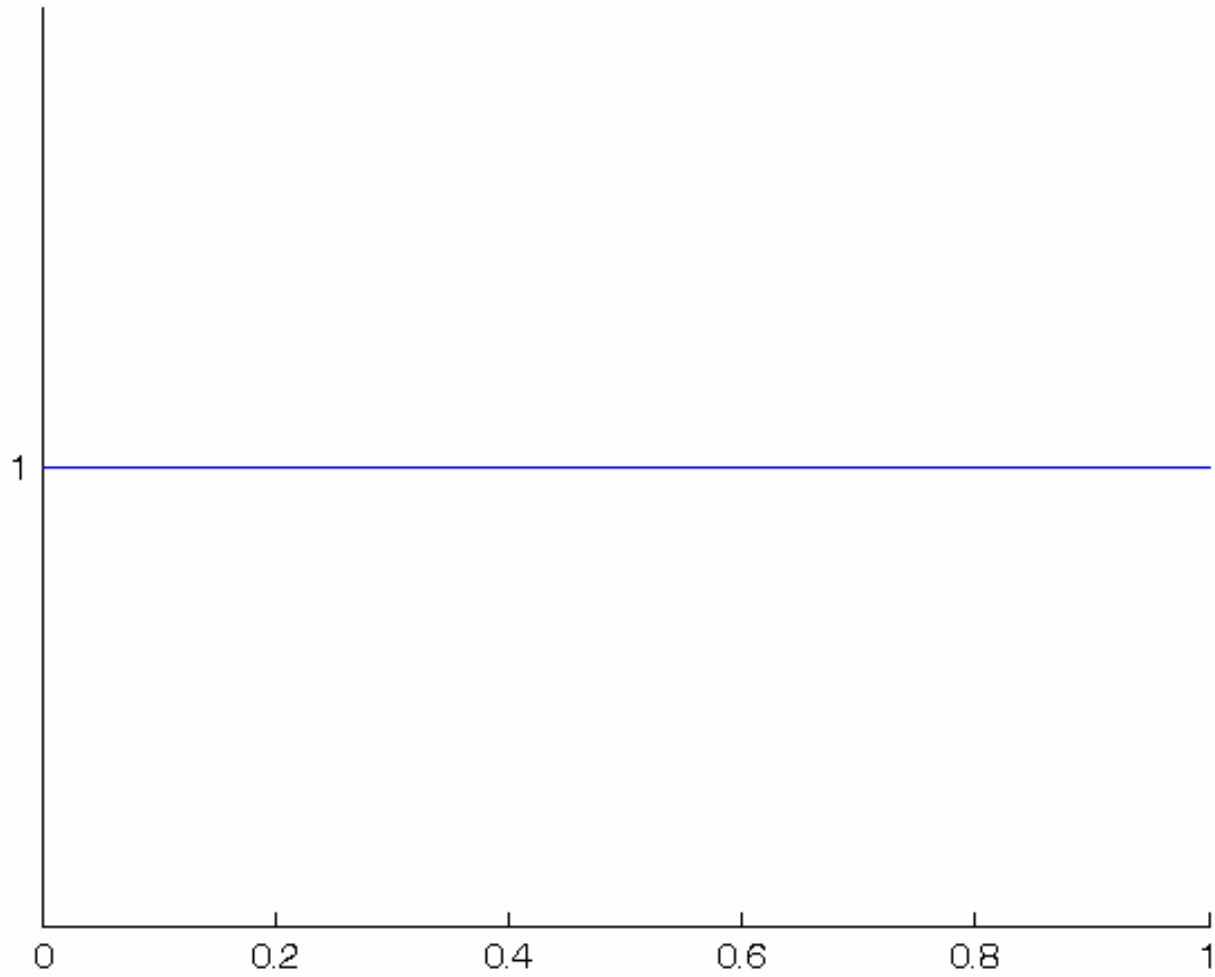
- Improper priors can (but not always) lead to proper posterior distributions
  - If likelihood dominates prior, no problem
- Proper priors *always* lead to proper posterior distributions

# Flat on One Scale not on Another

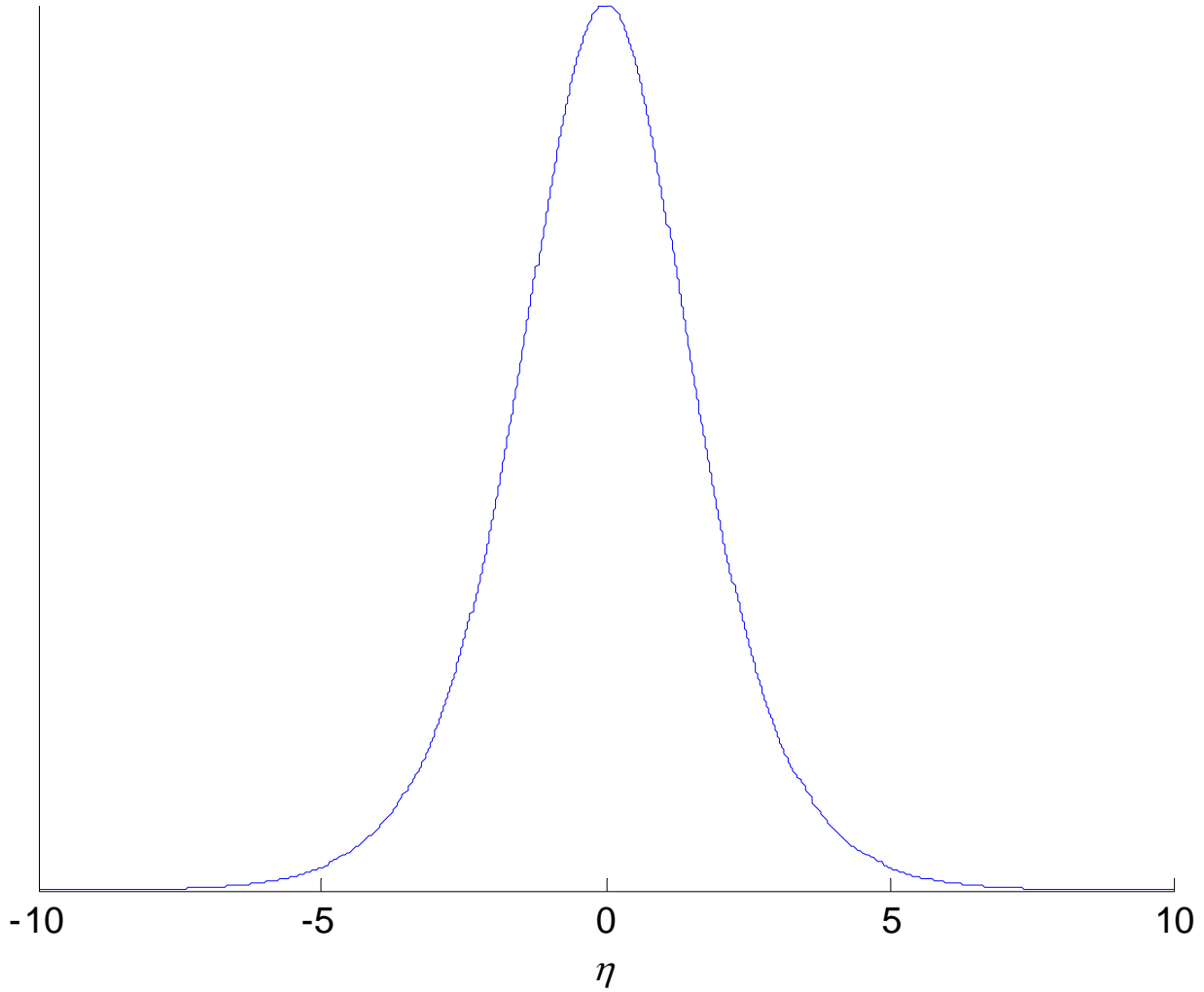
- For  $[Y] = B(n, p)$  the prior  $[p] = \text{Be}(1, 1)$  is constant
- Suppose we are interested in

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

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



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



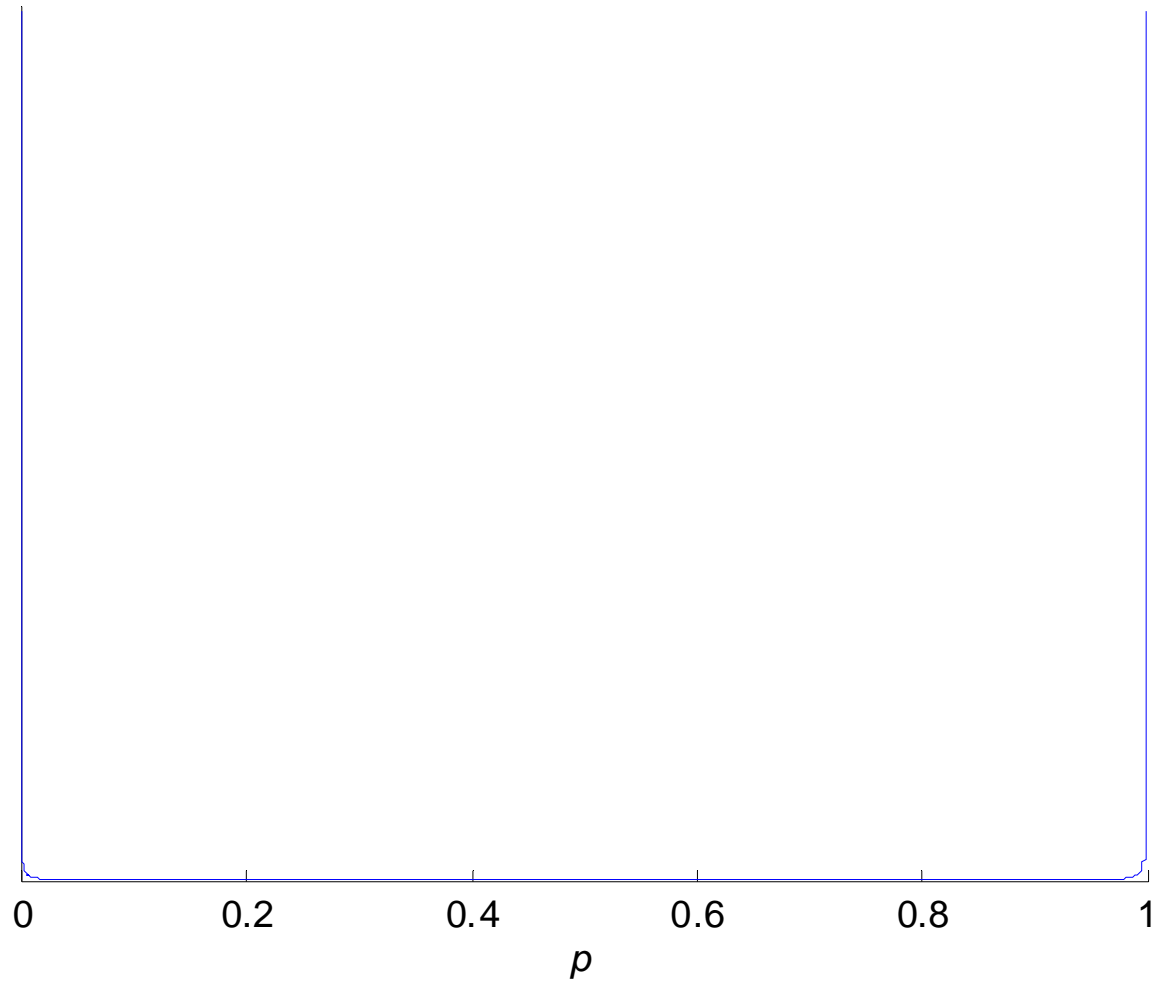
# Logit-normal

$$[\eta] = N(\mu, 1/\tau)$$

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

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

$[p = \text{expit}(\eta)]$  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
- Jeffreys' 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  $\theta$

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

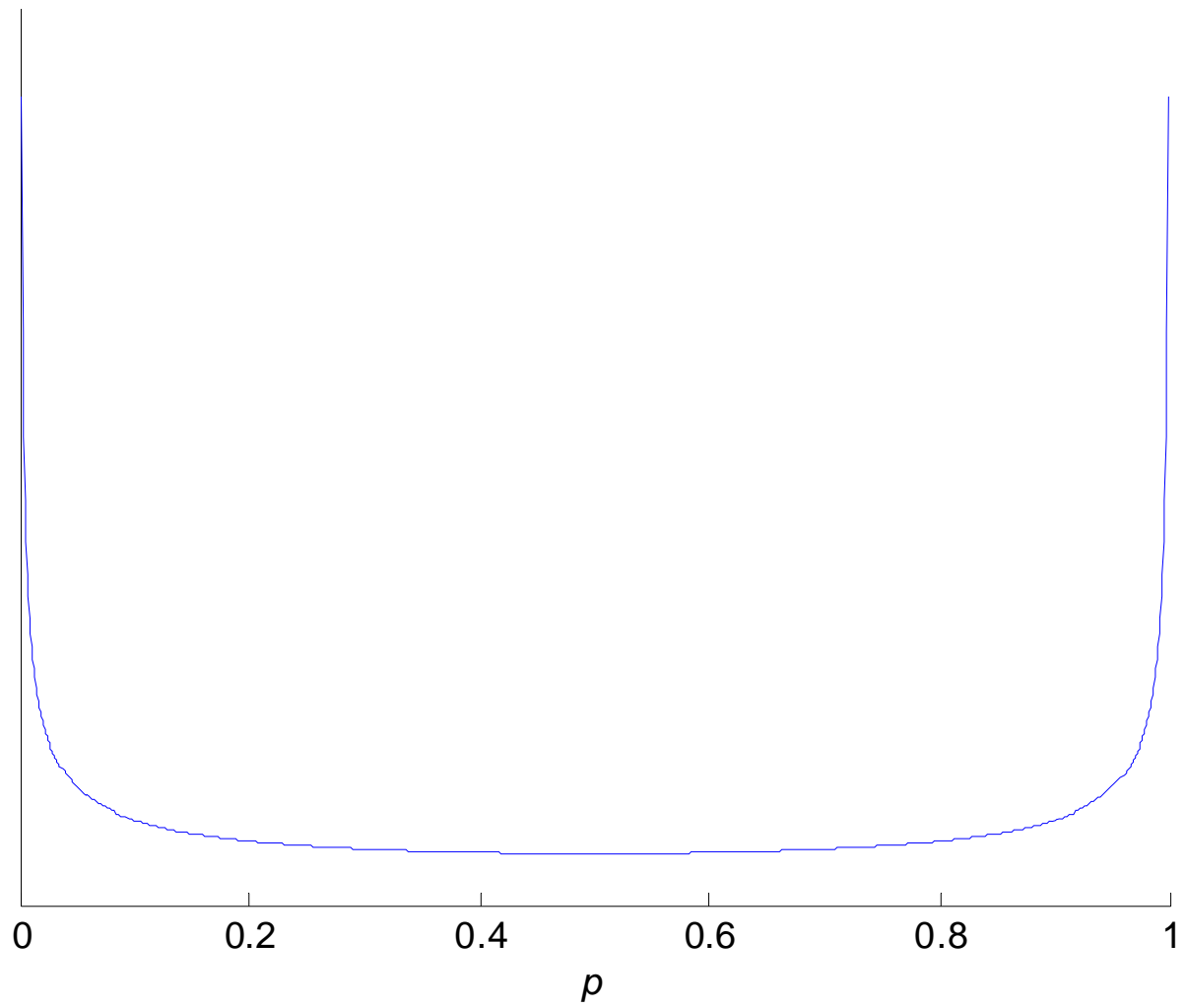
# Binomial Example

- 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

- Jeffreys' 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

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

# 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  $\tau$  from  $[\tau \mid y]$

---- Draw  $\mu$  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

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

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

# Posterior simulation

- So far we have been able to get posterior inference “directly” from the known distribution
- Sometime the distribution is not in known form
- Sometimes it is not convenient to sample from the known form, even if we have it
  - E.g., posterior prediction
  - E.g., multi-parameter models → conditioning
- Simulation is just a way to get random samples from a desired distribution

# Basic methods

- Direct methods

- Inverse CDF

- Requires that  $f(x)$  is normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Have (or can compute)  $F^{-1}(x)$

- Rejection

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^x f(x)dx$$

- Typically use for discrete distributions

- “Indirect”

- Rejection sampling

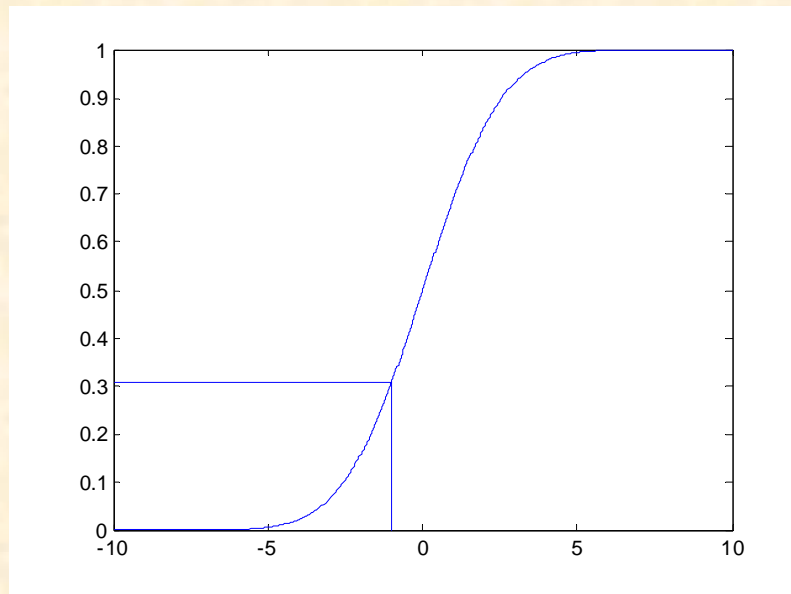
- MCMC methods

# Posterior simulation

- Simulations can be used to describe behavior of random variables
  - Use for sketching posterior distributions
- Useful software allows draws to be made from common distributions
- Helpful to know methods for sampling random variables

# Inverse CDF

- Also called rather grandly *probability integral transform approach*
- If you know the cdf then generate a  $U(0,1)$  and invert the CDF



# Generating an Exponential

- $f_Y(y) = \lambda e^{-\lambda y} \quad 0 < \lambda, 0 < y$

$$F_Y(y) = \int_0^{\infty} \lambda e^{-\lambda y} dy = 1 - e^{-\lambda y}$$

- Generate  $u \sim U(0, 1) \Leftrightarrow (1-u) \sim U(0,1)$

$$y = \frac{-\ln(u)}{\lambda}$$

# Exercise 5

- Simulate 1000 values from an exponential with parameter  $\lambda=0.56$ 
  - Note: ln in Excel is ln() not log()
- Compute the mean, median, and 95% credibility interval. Plot the distribution (scatter plot)

# Standard Distributions

- Most computer packages will draw uniform, normal, Poisson, beta, gamma deviates
  - Textbooks and internet good sources for algorithms
- Binomial
  - Draw  $u \sim U(0,1)$ : Assign  $y = \begin{cases} 0 & u > p \\ 1 & u < p \end{cases}$   
then  $y \sim B(1, p)$
  - Sum of  $n$  independent  $B(1, p)$  draws is  $B(n, p)$
  - [binomial\\_sim2.xls](#)

- For a  $k$ -cell Multinomial( $n, \{p_i\}$ ) compute  $c_i = \sum_{h=1}^i p_h$  the cumulative probabilities. Construct the partition:

$(0, c_1)$	1
$(c_1, c_2)$	2
$\vdots$	$\vdots$
$(c_{k-1}, 1)$	$k$

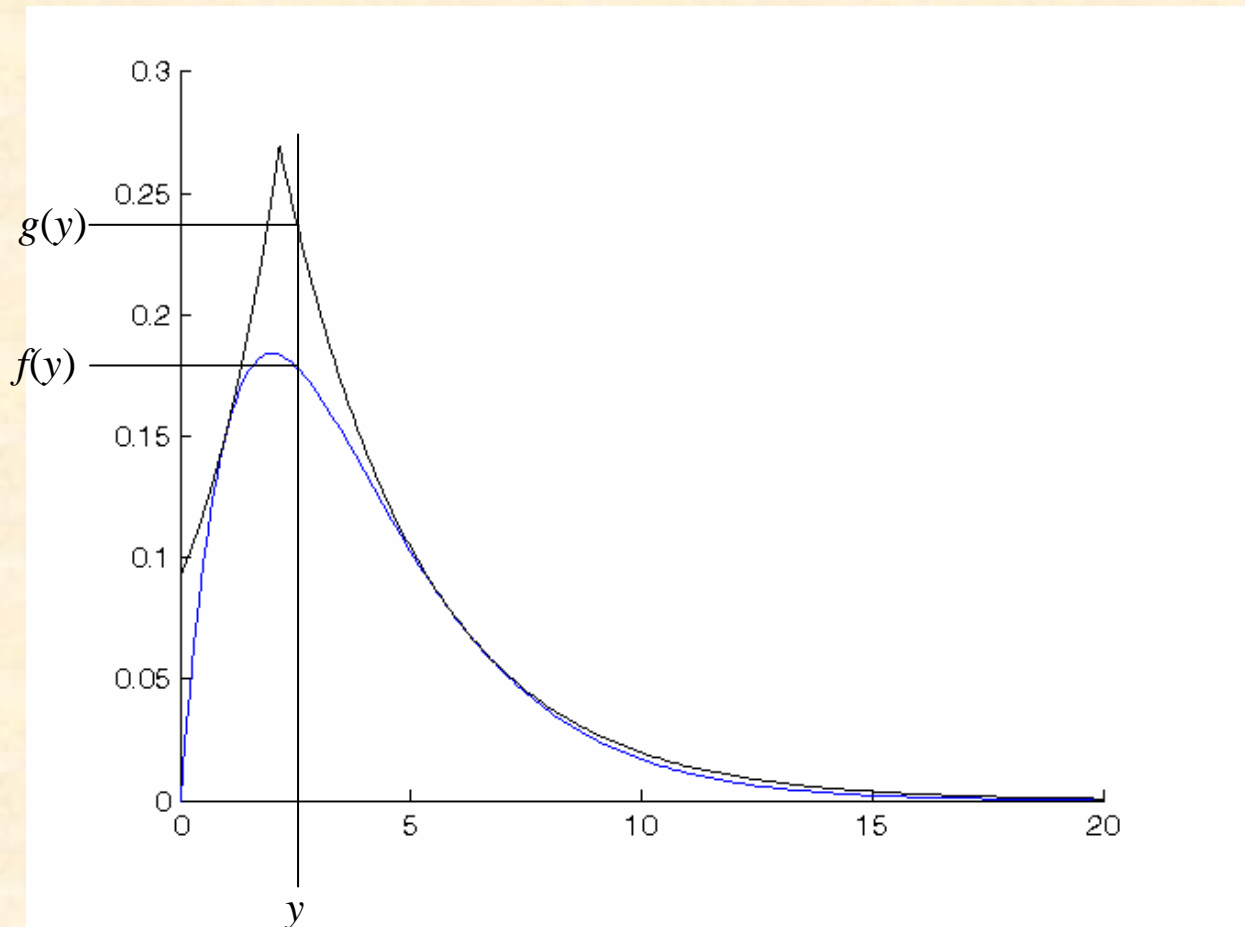
- Draw  $u \sim U(0, 1)$   $n$  times and tally the number of times  $u$  falls in each segment

# Rejection Sampling

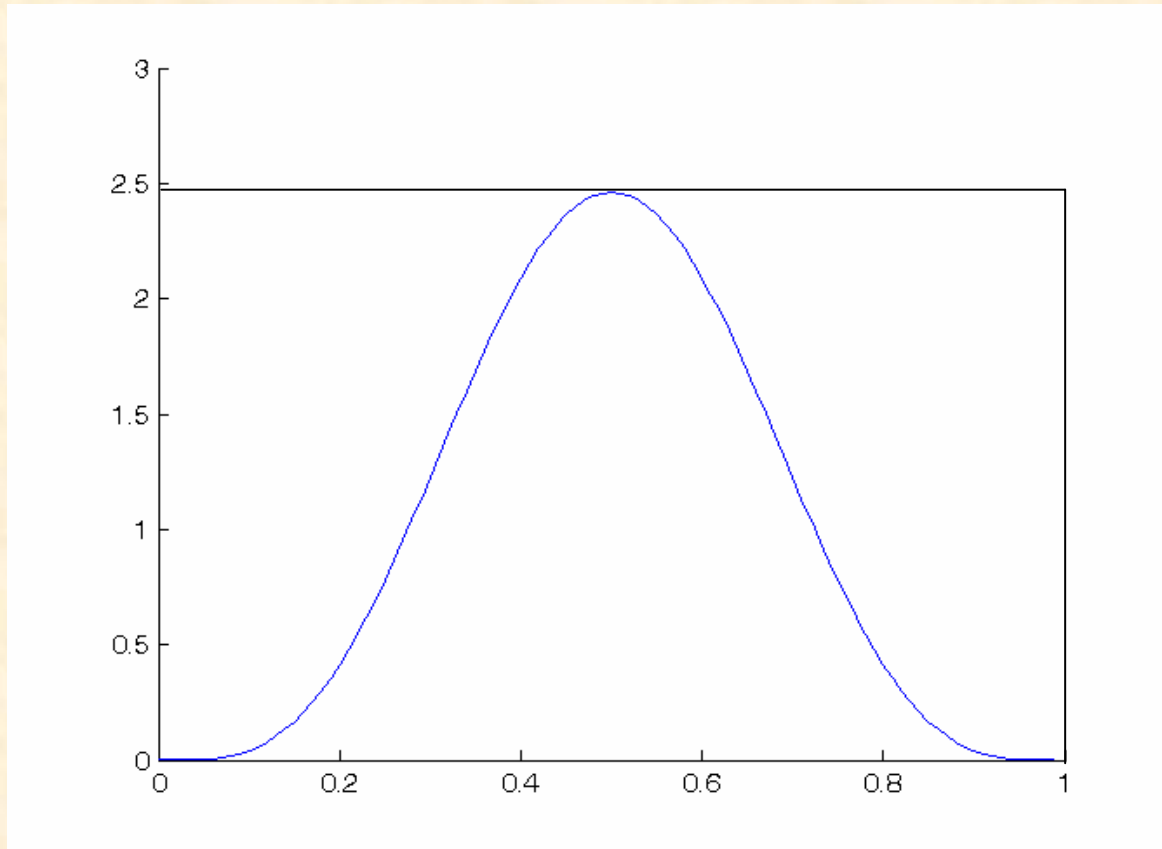
- Common in Bayesian inference for a posterior kernel to be specified
- If it is difficult to find the exact distribution several methods allow us to draw from the exact distribution using just the kernel.
- *Rejection sampling* allows us to draw a random sample from the exact target distribution
  - There is an efficiency cost

# Rejection Sampling

- Make use of another density  $g(y)$  that is easy to sample from that contains our target density  $f(y)$ 
  - Envelope
- Requires knowledge of  $f(y)$  that we use in constructing  $g(y)$
- Easy to construct an envelope if the target density is log-concave
  - Second derivative negative if scalar
  - Matrix of second partials of the log-density must be negative definite



- Generate a variate  $y$  from  $g(\cdot)$
- Accept with probability  $f(y)/g(y)$



- Often one can use a uniform envelope
  - Just need to find the mode
  - But this can be very inefficient

# Exercise 6

- Use a uniform envelope to sample ( $n=1000$ ) from a exponential with parameter  $=0.56$  and compute summary statistics
- Compare to earlier results (direct simulation)
- See if you can find a more efficient envelope distribution  $g(x)$ !

# Markov Chains

- Rejection sampling most useful when we have just one distribution to sample from
  - Usually interested in sampling from an often high-dimension posterior
- What if we are not smart enough to construct an envelope?
- Alternative is to take a non-random sample by constructing a Markov chain

# McMC

- Markov chain Monte Carlo
- Simulation based evaluation of posterior distributions
- Use dependent sequences of random variables  $\{X_t\}$ 
  - Usual positive autocorrelation means that the effective sample size is smaller than dimension of sequence

# First-order Markov chain

- $\{X_t\}$  is a first-order Markov chain if

$$\Pr(X_{t+1} | X_1, X_2, \dots, X_t) = \Pr(X_{t+1} | X_t)$$

- Corresponding to the *transition kernel*  $\Pr(\cdot | \cdot)$  is a unique stationary distribution satisfying:

$$\phi(X)dx = \lim_{t \rightarrow \infty} \Pr(X_{t+1} \in (x, x + dx] | X_t)$$

- MCMC involves constructing a Markov chain for which the stationary distribution is the target posterior distribution

# Stationary Distributions

- Major problem in Markov chain theory is to determine if an invariant distribution  $\phi(X)$  exists and conditions under which iterations of the transition kernel converge to  $\phi(X)$
- MCMC reverse this problem – given  $\phi(X)$  find a transition kernel that converges to  $\phi(X)$

# McMC

- Create a Markov process whose stationary distribution is the target distribution
  - Run the simulation long enough so that the distribution of current draws is close to the target
  - Need to decide how long
- Many ways to construct these Markov chains
- Need to check for convergence

# Example

- 2 states – at  $t$  make a decision about state at  $t+1$ .  
This depends on the state at  $t$
- $\mathbf{p}_t = (p_{1t} \quad p_{2t})$   $p_{it} = \text{Pr}(\text{in state } i \text{ at time } t)$
- *Transition matrix:*  $\Psi = \begin{pmatrix} \psi_{11} & 1 - \psi_{11} \\ \psi_{21} & 1 - \psi_{21} \end{pmatrix}$

$$(p_{1t} \quad p_{2t}) \times \begin{pmatrix} \psi_{11} & 1 - \psi_{11} \\ \psi_{21} & 1 - \psi_{21} \end{pmatrix} = (p_{1t+1} \quad p_{2t+1})$$

# Example

- 2-stage transition:  $\mathbf{p}_{t+2} = \mathbf{p}_t \Psi \Psi$
- $n$ -stage transition:  $\mathbf{p}_{t+n} = \mathbf{p}_t \Psi^n$  ( $\Psi^n = \prod_{i=1}^n \Psi$ )

but as  $n \rightarrow \infty$   $\mathbf{p}_{t+n} \rightarrow \begin{pmatrix} \frac{\psi_{21}}{\psi_{21} + \psi_{12}} & \frac{\psi_{12}}{\psi_{21} + \psi_{12}} \end{pmatrix}$

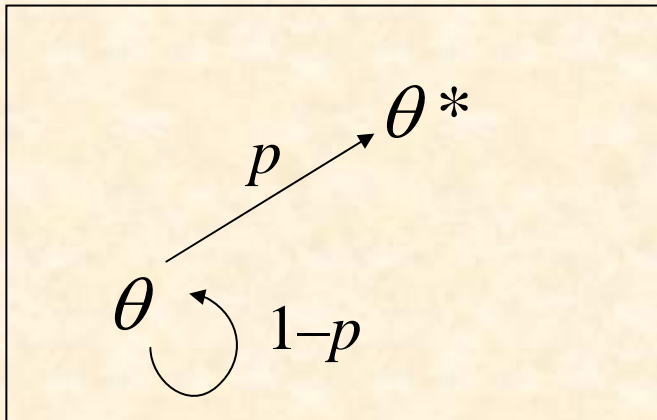
- If Markov chain ‘well-behaved’ final state is the same regardless of where you start
  - *Stationary distribution*

# When Does it Work?

- (1) The sequence must be a Markov chain with a unique stationary distribution
  - Chain must be irreducible, aperiodic, not transient
  - Except for trivial conditions any random walk is aperiodic and not transient
  - Irreducible if there is a non-zero probability of reaching any state from any other state
- (2) The unique stationary distribution must be the target distribution

# Metropolis-Hastings

- The Metropolis-Hastings algorithm ( and Gibbs sampling as a special case) satisfy these properties provided the chain that is generated is irreducible
- Currently at  $\theta$  - propose a new value  $\theta^*$  that is drawn from a candidate generating distribution  $J(\theta^*/\theta)$



- Accept the proposal with

$$p = \frac{f(\theta^*)J(\theta | \theta^*)}{f(\theta)J(\theta^* | \theta)}$$

# Metropolis-Hastings

- Acceptance probability has two parts:

$$(1) \frac{f(\theta^*)}{f(\theta)} \text{ - model} \quad (2) \frac{J(\theta | \theta^*)}{J(\theta^* | \theta)} \text{ - candidate generation}$$

- Useful for posterior simulation as only the kernel is needed

# Example

- Excel demo [mh.xls](#)
- Issues:
  - Specification of jumping (proposal) distribution
  - Starting values
  - Convergence diagnostics

# Gibbs Sampling

- Used in multidimensional applications
- *Alternating conditional sampling*
- Suppose  $\theta$  has  $k$  components  $(\theta_1, \dots, \theta_k)$
- At each iteration, cycle through the components drawing  $\theta_i$  from  $p(\theta_i | \theta_{i-})$ 
  - $\theta_{i-}$  is all the components of  $\theta$  except  $\theta_i$
- $p(\theta_i | \theta_{i-}) =$  Full conditional distribution for  $\theta_i$

# Gibbs Sampling

- In iteration  $t$  element  $\theta_i$  is updated conditional on the latest components of  $\theta$  which are iteration  $t$  values if they have already been updated, iteration  $t-1$  values otherwise
- For many problems the full conditional distributions can be found as known distributions
  - If not use M-H

# Example

- $y_1, \dots, y_n$ , a random sample from a  $N(\mu, 1/\tau)$  with both parameters unknown

$$[\mu | \tau] = N(\psi, \kappa\tau) \quad [\tau] = Ga(\alpha, \beta)$$

$$[\mu | y_1, \dots, y_n, \tau] = N(m, t)$$

$$[\tau | y_1, \dots, y_n, \mu] = Ga(a, b)$$