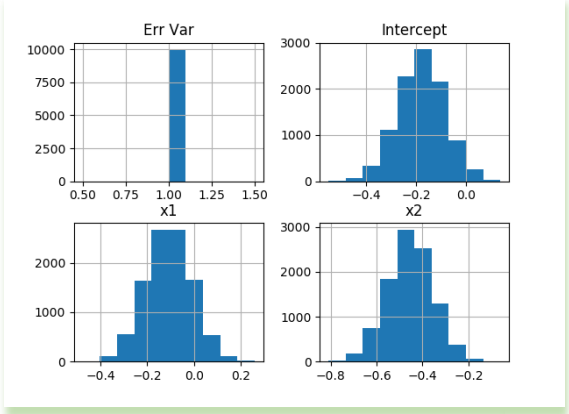
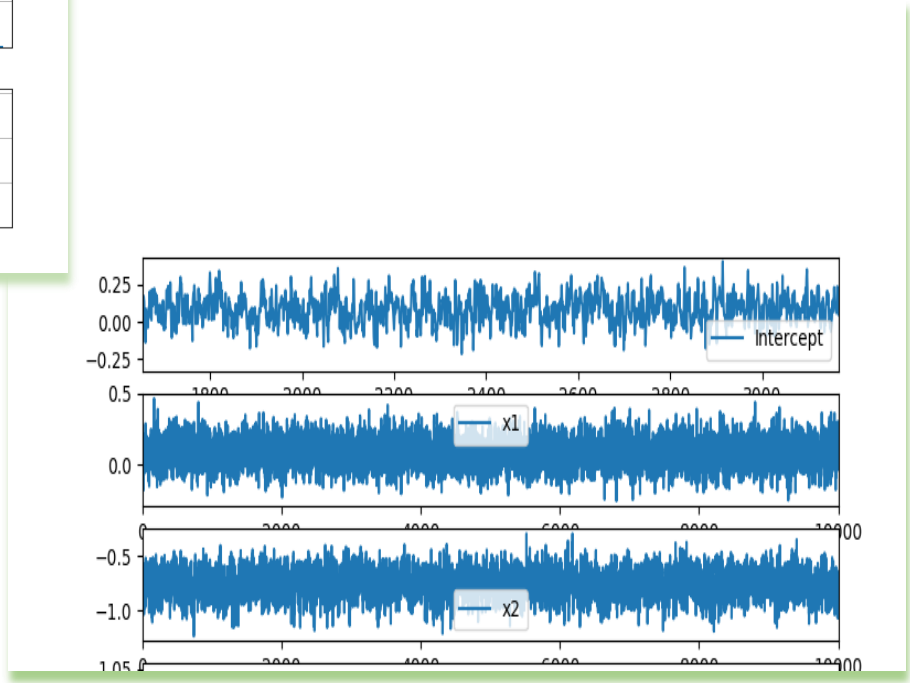
# Introduction to Bayesian Statistics

Kelvin Balcombe,
Applied Economics Marketing and Development Group
School of Agriculture Policy and Development
University of Reading

Thomas Bayes
1702-1761

Pierre-Simon Laplace
1749-1827




$$\underbrace{f(\beta|Y,M)}_{\text{Posterior}} = \frac{\overbrace{f(Y|\beta,M)}^{\text{Likelihood}} \times \overbrace{f(\beta|M)}^{\text{Prior}}}{\underbrace{f(Y|M)}_{\text{Marginal Likelihood}}}$$

# Some Do's

- Keep an open mind
  - What you've been told may not be 100% accurate
  - Bayesian statistics is not all about being subjective, and its not all about imposing prior beliefs on your results
- Believe that you are able to apply Bayesian methods
  - Bayesian modelling is not intrinsically hard
  - Bayesian modelling is getting easier
- Believe that Bayesian Methods can open up new avenues for your research

These days the statistician is often asked such questions as "Are you a Bayesian?" "Are you a frequentist?" "Are you a data analyst?" "Are you a designer of experiments?". I will argue that the appropriate answer to ALL of these questions can be (and preferably should be) "yes", and that we can see why this is so if we consider the scientific context for what statisticians do. **G.E.P. Box**

# Some Don’ts

- *Don’t Panic!, y*ou don’t have to  identify yourself as ‘being Bayesian’ if you use Bayesian methods
- Don’t feel that you have to do ONLY one or the other
  - Statistical paradigms need not be religious sects
  - As a referee please don’t tell people that they must only use one or the other!
- Don’t treat Classical v Bayesian debates too seriously
  - Though, read and enjoy them as they can make you reflect on the nature of statistical inference
- Don’t judge one by the tenants of the other. Not everything “translates” perfectly

> “It can be very dangerous to see things from somebody else's point of view without the proper training.”
> — **Douglas Adams**

# Please Remember !

- The terms Classical and Bayesian are “catch all” terms. There are debates and a variety of perspectives within each of these approaches.
  - Classical Statistics (as it is often practised) has been labelled an “uneasy alliance” between Neyman & Pearson and Fisherian approaches to inference.
  - There are Bayesian’s that have sought to be “objective” and those that are unapologetically “subjective”
- What bothers some Bayesian’s most is not the application of classical methods, but their misinterpretation.

> “I don't know what's the matter with people: they don't learn by understanding, they learn by some other way — by rote or something. Their knowledge is so fragile!”
> **— Richard Feynman**

# Part 1, Who is Bob’s Father?

Bob-Junior was reading the paper one morning when he came across…


# VINTAGE NEWSPAPER

**World - Sport - Business - Finance - Lifestyle**

One out of 25 chilidren falsely believe their Dad is theirs!

If you are worried, send two strands of hair, one of your own and one of your Dads', to Sherlock Laboratories, 221b Baker Street with only £10, and we tell you if its really your dad.

Surprised by the seemingly high possibility that the man he thought to be his father (Bob-Senior) may not be, Bob-Junior decided to send away two strands of hair (Bob-Junior's and Bob-Senior's) and awaited the results. Two weeks later, he received a letter

Dear Bob

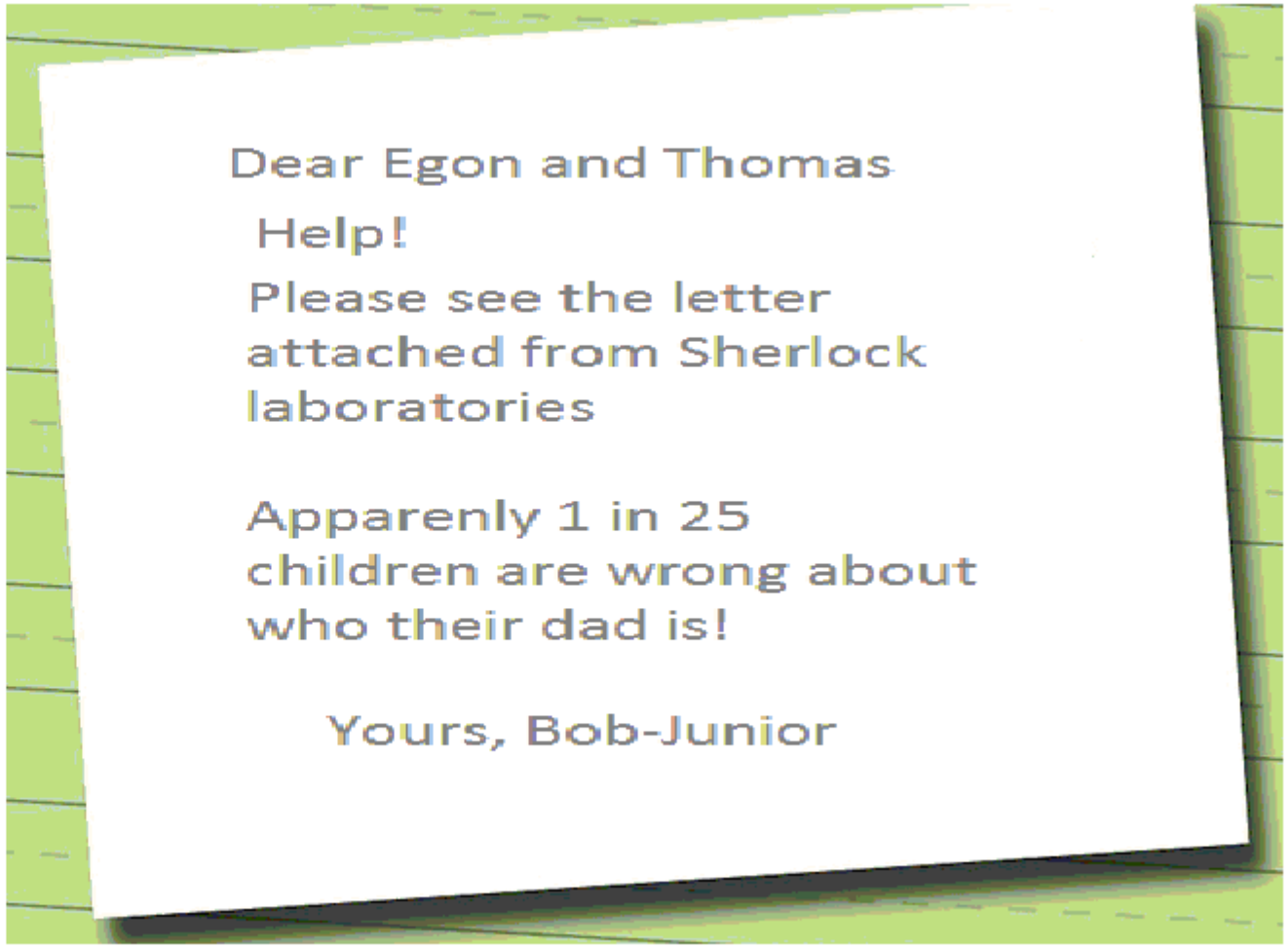
He's not your father!

"Please note, studies have shown that this test will identify a non-biological father in 99.99% of cases. However, in a little less than 1% of cases, a biological father will be falsely found not to be a father".

Sherlock Laboratories

Bob Junior asks for help from this two friends Egon  and Thomas 

Egon had done a course in Classical statistics whereas Thomas had done a course in Bayesian Statistics

University of Reading

# Dear Bob

If the information provided is correct, then one would normally consider this evidence to **be** rather strong. Specifically, the hypothesis that Bob-Senior is your father can be rejected at the 1% level of significance. Another way of looking at this is that, you can be 99% confident that Bob-Senior is not your father. On the other hand, how confident you need to be is up to you.

Cheers Egon


Dear Bob

If the information provided is correct, then there is still around a 20% chance that Bob-Senior is your father

Regards Thomas

# Bob-Junior sought clarification


Dear Egon and Thomas

You seem to disagree with each other !

Please explain yourselves

University of Reading

# Dear Bob

*If you randomly test 10000 children's matches with their potential fathers, we would expect that 400 (i.e. 10000/25) are not biological fathers, from which we would expect around 99.99% (399) of the non-biological fathers to be correctly identified. Of the other 9,600 we would expect to falsely identify around 96 who are not biological fathers. Therefore, of the 399+96=495 identified as not being biological fathers, 96/495 (19.3%) would be biological fathers. Hence my answer that there was 20% probability that Bob-Senior is not your father.*

*Thomas*

# Thomas's Argument

399/400 non-biological fathers identified

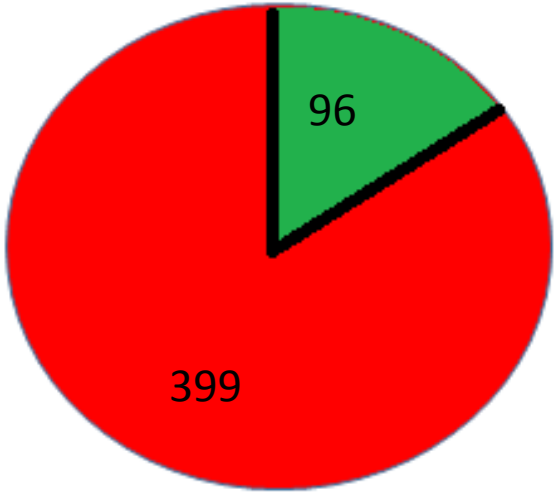
399

96

96 biological fathers wrongly identified as not being fathers

10,000

Approx 20%



# Or using Bayes' Theorem

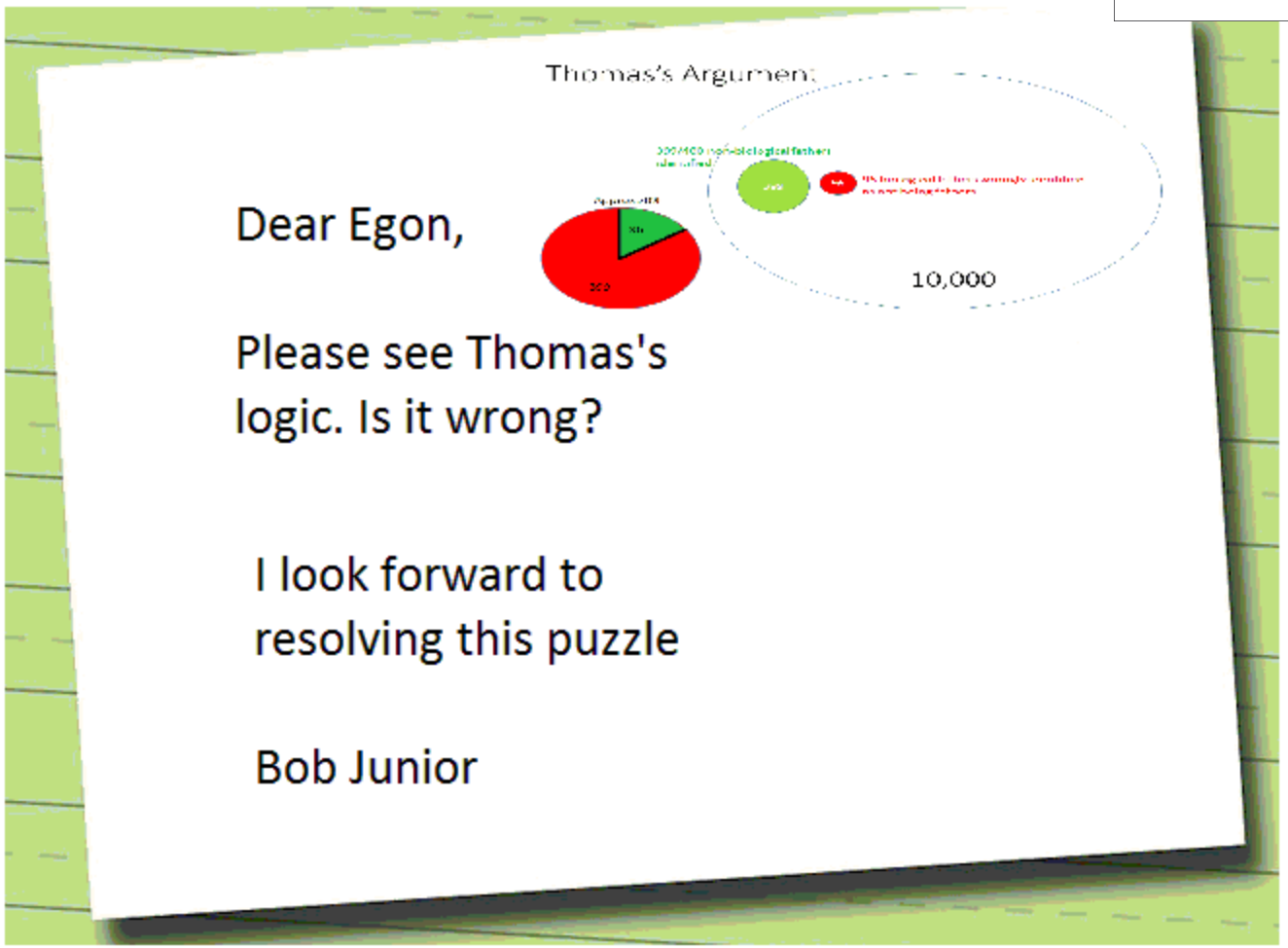
University of Reading

```
#  Denote        Reality=R, Test= T,       outcomes Father=F or Non Father=NF


prior_p=24.0/25.0          #prior_probability_bob_senior_is_bobs_juniors_father     P(R=F)
p_error1=0.01              #probability a father is identified as a non-father      P(T=NF|R=F)
p_error2=0.001              #probability a non-father is identified as  a father)    P(T=F|R=NF)

p_correct1=1.0-p_error1    #probability  a father is identified as a father         P(T=F|R=F)
p_correct2=1.0-p_error2    #probability a non-father is identified as a non-father  P(T=NF|R=NF)
```

```
#Bayes Theorem
#P(R=F|T=NF) = P(T=NF|R=F)P(R=F) /  (   P(T=NF|R=F)P(R=F)  +      P(T=NF|R=NF)P(R=NF)    )

posterior_p=   p_error1*prior_p/(p_error1*prior_p + p_correct2*(1-prior_p))

print('P(R=F|T=NF) = ',posterior_p)
print('Confidence = ', p_correct1)
print('Power = ', p_correct2)
```

```
('P(R=F|T=NF) = ', 0.19370460048426136)
('Confidence = ', 0.99)
('Power = ', 0.999)
```


Egon


Thomas's Argument:
10,000
Dear Egon,
Please see Thomas's logic. Is it wrong?
I look forward to resolving this puzzle
Bob Junior

University of Reading

# Dear Bob

*If I accepted that 1 out of 25 children falsely believe who their biological father is, and if I agreed to attach a probability to the event that Bob-Senior is Bob-Junior's father, then I entirely accept Thomas's calculations! However, Bob-Senior either is or isn't Bob-Junior's father. I did not assign a probability to this event, nor do I have no clear basis to do so (the article is mere hearsay). Even If I were to accept the 1 in 25 ratio, then out of 9600 biological fathers, we would correctly identify 9504 out of 9600 (99%) as biological fathers. Hence, my advice remains that you can be 99% confident of your result. I see no reason to retract my original advice.*

*Regards Egon*

# Having copied this letter to Thomas, Bob got the final response from Thomas

Dear Bob

*It seems odd to me that Egon will not attach a probability to Bob-Senior being your father. It seems a perfectly reasonable action to me. Moreover, being 99% confident seems very much like assigning a 99% probability to me!*

*I can easily generate a 1%Probability that Bob-Senior is your father by assuming that prior to the test there was a 50% probability that he was not your father.*

*However, unless you had some previous suspicions this does not seem a plausible belief*

*Regards Thomas*

# Some points to take away

- Egon's statement that Bob-Junior can be 99% confident, did not mean that Bob-Junior can attach a 1% probability to the event that Bob - Senior is his biological father. Egon never attached a probability to this event at all!
- Thomas does not deny that Bob-Senior either is or isn' t Bob-Junior's father. Putting a probability of this event does not deny there is an objective truth ( it is "epistemic" uncertainty)
- Egon would not attach a probability to an event. Many Classical statisticians might therefore agree with Thomas in this case. However, they draw the line at assigning probabilities to parameters within Models….

"A common mistake that people make when trying to design something completely foolproof is to underestimate the ingenuity of complete fools."
— **Douglas Adams**

# Part 2: Bayesian Modelling Basics

# What's a Model?

- I this talk, when we talk of  model I imagine dependent/or response variable y, that needs to be explained

$$y_i \sim f(x_i, \beta) \qquad i = 1, \ldots\ldots n$$

- We think of the x's as being fixed values as if being set in an experiment.
- By "model" (M) we mean the nature of $f(x_i, \beta)$ that is  a probability distribution that depends on$(x_i, \beta)$ where $\beta$ are the parameters that characterise that distribution.

Essentially, all models are wrong, but some are useful". **G.E.P Box**

# The Bayesian approach

- In Classical statistics a probability distribution is never assigned to β
  - not knowing what β is, does not mean it has a distribution
- In Classical statistics there is no formal place for "Prior Beliefs" about parameters or models
  - Though, prior beliefs play a large role in any approach to modelling
- Classical's do not always obey the likelihood principle

- Bayesians formally attach probability distributions to the parameters β that characterise models
- Bayesians even attach probability distributions to the models M
- The attach them both before they observe data, and after
- Bayesians obey the "likelihood principle".

"We demand rigidly defined areas of doubt and uncertainty!"
— **Douglas Adams**,

# The likelihood

$$Y = Dependent\ Data$$
$$\beta = Parameters$$
$$M = Model$$
$$X = Exogenous\ Variables$$

- f(Y|β,X,M) AKA $L(\beta; X, Y)$

- The likelihood is the same quantity that is maximised in Maximum Likelihood Estimation (MLE). It is the density of the data for given values of β. It is the probability mass or density of the observed data.

- Note the difference in how Bayesians tend to write it (on the left) compared to commonly used Classical Notation (on the right). This is because the Bayesian notation explicitly acknowledges the stochastic nature of β

# Bayes' Theorem Applied to Models

Posterior

$$f(\beta|Y,X,M) = \frac{\text{Likelihood } f(Y|\beta,X,M) \times \text{Prior } f(\beta|M)}{\text{Marginal Likelihood } f(Y|X,M)}$$

$$f(\beta|M) = f(\beta|X,M)$$

$$f(Y|X,M) = \int f(Y|\beta,X,M) \times f(\beta|M) d\beta$$

i.e. The prior expected likelihood

$Y$ = *Dependent Data*

$\beta$ = *Parameters*

$M$ = *Model*

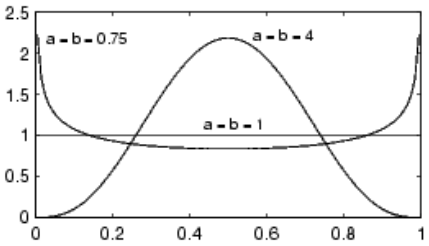
$X$ = *Exogenous Variables*

# The Prior Distribution $f(\beta|M)$

- Is the distribution of the parameters before observing the data
- Can be ‘strongly-informative’ or ‘weakly-informative’
- Can be ‘Proper’ or ‘Improper’.
  - An example of an improper prior that is commonly used is f(β)∝1 (without bounds) – sometimes referred to as a flat prior
- Is ‘Conjugate’ if they generate a posterior that is of the same form as the prior.
  - Conjugate Priors are nice, but not essential
- Are “Empirically Bayesian” if they use data in the likelihood
- Can be shaped by using a subset of the data as a “training sample” to form the prior
- Can be “hierarchical” in the sense of putting priors on the parameters in the prior distribution for other parameters.

Prior distributions are usually not “subjective” and do not represent “belief” Andrew Gelman

# What is a non-informative prior?




University of Reading

- Non-informativeness is actually hard to define and has been the source of much debate
- A uniform prior is not necessarily a non-informative prior.
- Often, non-informativeness is taken to mean a very diffuse proper prior, or the limiting case of that proper prior.
- Jeffrey's priors (the prior is proportional to the square root of the determinant of the Fisher information matrix) are priors give invariant results under parameter transformations.
  - However, these are not usually used in applied work, not least because they are often not proper.

# The Marginal Likelihood

$$f(Y|X,M) = \int f(Y|\beta,X,M) \times f(\beta|M) d\beta$$

- Is very useful if you want to test hypotheses, evaluate models or to average over models – to be discussed later.
- It can often be ignored when estimating the posterior distribution since

$$f(\beta|Y,X,M) \propto f(Y|\beta,X,M) \times f(\beta|M)$$

- As we shall see, being able to calculate the posterior up to a proportional constant is all that is needed.

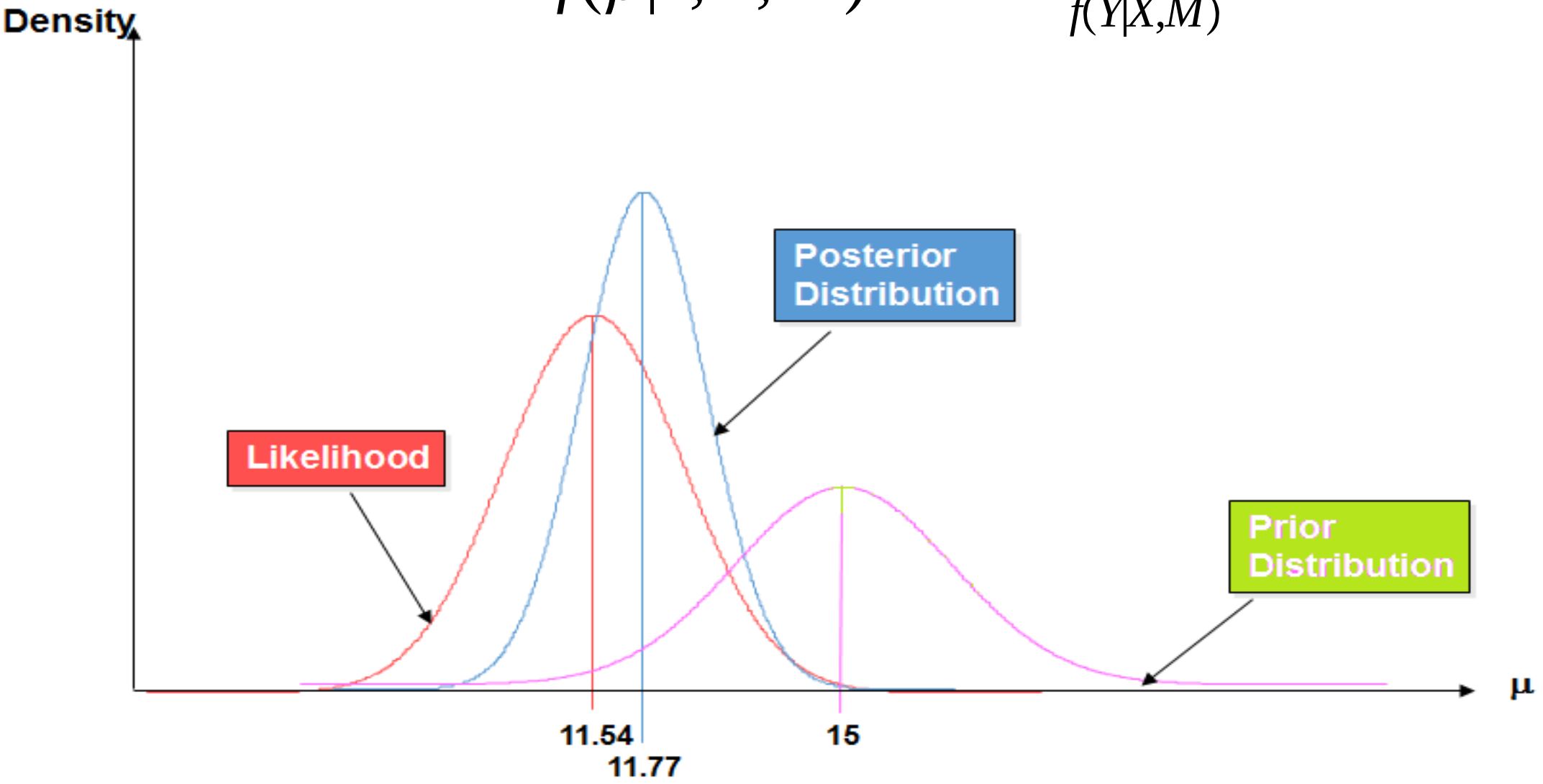
"Taking a model too seriously is really just another way of not taking it seriously at all.“ Andrew Gelman

# The Posterior $f(\beta|Y,X,M)$

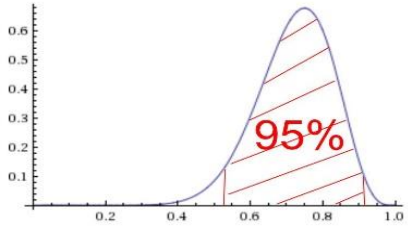
- The Posterior Distribution is the “updated” version of the prior distribution when new information is observed
- If more information came to light, it can serve as the prior in the next round of updating
- Posteriors become more densely packed around a point as you add more and more data, and become less dependent on the Prior.
- Many Posteriors become approximately Normal, even if the priors are not normal.

# The Tri-Plot

$$f(\beta|Y,X,M) = \frac{f(Y|\beta,X,M)\times f(\beta|M)}{f(Y|X,M)}$$



# Bayesian Reporting




- Bayesian's can report the entire posterior visually (or in some cases analytically)
- Given the posterior, one can report the mean, median, mode, and/or standard deviations of the posterior
- The generally preferred Bayesian approach is to report the mean of the posterior, along with quantiles +standard deviations and or "credible intervals" for parameters.

- Note: These can be numerically very similar to Classical estimates in some cases

# Contrast, “Throwing a Coin” Bayesian vs Classical Inference


- Let a computer choose a computerised coin from
  - A fair Coin
  - A Coin with both Sides Heads
  - A Coin with both Sides Tails
  - A biased coin with an unknown Probability of Heads/Tails

- If wanted to find out about which of the coins had been chosen……let’s contrast a Bayesian and Classical Approach.

# Model Comparison and Hypothesis Testing

Let us assume that we two models, $M_0, M_1$ (I will drop the X for simplicity!)

$$Posterior\ Odds = \frac{f(M_0|Y)}{f(M_1|Y)} = \frac{f(Y|M_0)}{f(Y|M_1)}\frac{f(M_0)}{f(M_1)}$$

*in favour of* $M_0$

The Bayes Ratio

The Prior Odds

- In Principle this Approach to Model Comparison Applies to any Models, Nested or Non-Nested
- Needs Proper Priors
- Can be dubious if priors are not common to all models – i.e. your model selection may simply be a function of your priors

# Hypothesis Testing

- $H_0$:β=0 v $H_1$:β ≠ 0, can be tested using model comparison through Bayes Posterior odds,
  - Where the prior probability of $H_0$ is p the imposition of which leads to $M_0$ and the encompassing Model is $M_1$

$$f(M_0|y) = \frac{f(y|\beta=\beta_0)p}{f(y|M_1)(1-p)+f(y|\beta=\beta_0)p}$$

- The Bayes Ratio collapses to a density ratio referred to as the Savage-Dickey Density Ratio. If the encompassing model is $M_1$, then

$$\frac{f(Y|M_1,\beta=\beta_0)}{f(Y|M_1)} = \frac{f(\beta=\beta_0|Y,M_1)}{f(\beta=\beta_0|M_1)}$$

# Models and Hypotheses can also be evaluated Using

- Deviance Information Criteria (DIC)
- Watanabe–Akaike criterion  (WAIC)
  - https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12062
- Bayesian “Leave One Out “Cross Validation (LOO) https://arxiv.org/pdf/1507.04544.pdf

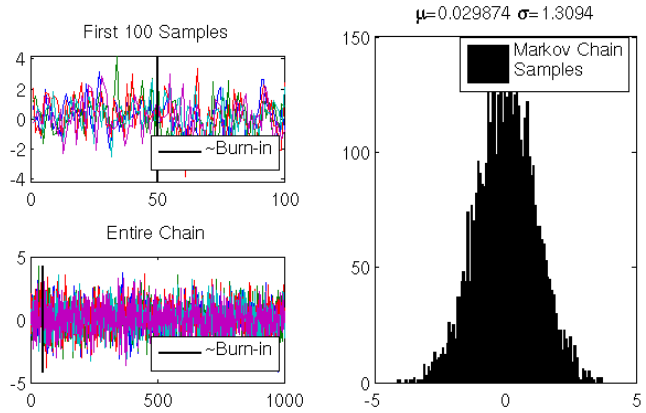
# Diagnostics

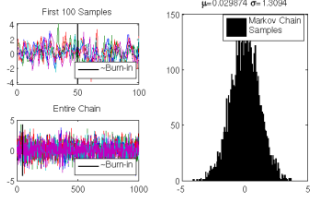
## Can be done using the posterior predictive distribution

$$p(g(Y^*)|Y) = \int p(g(Y^*)|\beta, Y) f(\beta|Y) d\beta$$

- Draw Parameters from the Posterior
- Draw Data given the Parameters
- Calculate a simulated statistic (e.g. the Skew)
- Repeat many times to obtain a distribution for the simulate statistic
- Compare the distribution of the simulated statistic with the real statistic
- If the real statistic is the in the upper or lower tails of the simulated distribution?
    - you have a problem

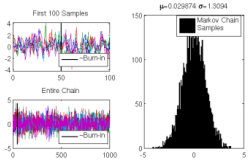
# Part 3: Bayesian Estimation






# Estimation

- Sometimes the posterior can derived analytically and is of a “well known” form
  - e.g. The Normal Linear Model with Normal-Gamma priors.
- Multiplying the Likelihood by the Prior then maximising the Posterior (MAP) instead of the Likelihood (as in penalised Maximum Likelihood)
- Monte-Carlo Techniques including:
  - Importance Sampling
  - Gibbs Sampling
  - Monte Carlo Markov Chain (MCMC) Sampling (of which there are many variants including Gibbs Sampling which is a special case)
  - Hamiltonian MCMC is becoming the preferred choice (though not in all circumstances)




# Gibbs Sampling

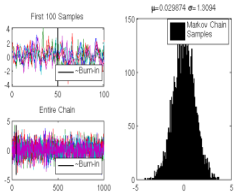
- In many cases we can deduce conditional posterior distributions that are of a known form but not the joint posterior distribution
- We can sample from the joint posterior by sequentially drawing from the conditionals

For example, if $\beta = (\alpha, \sigma^2)$,
and we know $f(\alpha|Y, \sigma^2)$ and $f(\sigma^2|Y, \alpha)$ but not $f(\alpha, \sigma^2|Y)$
then we can

- Start from an arbitrary point $\sigma_0^2$ with $n = 0$
- While $n \leq N$
  a) sample $\alpha_n$ from $f(\alpha|Y, \sigma_n^2)$
  b) sample $\sigma_{n+1}^2$ from $f(\sigma^2|Y, \alpha_n)$
  c) $n = n + 1$ and return to a)
- Dump the first $N^*$ and only record the next $N - N^*$ ($N^*$ is known as a burn in)

$$\{\alpha_n, \sigma_n^2\}_{n=N^*}^{N}$$

are draws from $f(\alpha, \sigma^2|Y)$



# Metropolis Algorithms

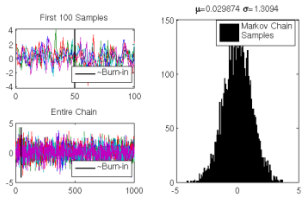
- This can be used more generally than the Gibbs Sampler.
- The form the joint or marginal posteriors may not be of a recognisable form but deduced up to a proportional constant

$$f(\beta|Y,M) \propto f(Y|\beta,M) \times f(\beta|M)$$

## The Metropolis-Hastings (with Normal Random Walk Proposal)

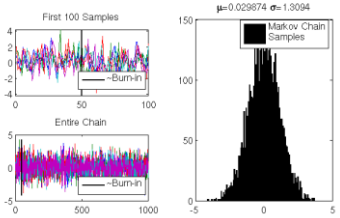
- Start from an arbitrary point $\beta_0$ with $n = 0$
- While $n \leq N$
  a) "Propose" (draw) $\beta^*$ from $N(\beta_n, \Omega)$
  b) Draw a uniform u
  b) Set $\beta_{n+1} = \beta^*$ if $\frac{f(Y|\beta^*,M)\times f(\beta^*|M)}{f(Y|\beta_n,M)\times f(\beta_n|M)} > u$ else $\beta_{n+1} = \beta_n$
  c) $n = n + 1$ and return to a)
- Dump the first $N^*$ and only record the next $N - N^*$ ($N^*$ is known as a burn in)

Note it is usual to "tune this covariance" so that there is around 40% acceptance rates



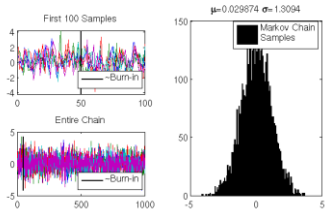

# “MCMC Samplers”

- Some samplers do not operate very well given low density starting points
- Some programs tend not to choose arbitrary starting points but first maximise the posterior, and use that as the starting point.
  - This can involve extra complexity for the user in the start up phase.
- MCMC delivers auto-correlated sequences. Therefore they are often “thinned”
- MCMC cannot be “parallelised” in the sense that each chain cannot be sped up by parallel processing. However, it can be parallel in the sense of running multiple chains simultaneously




# Convergence of MCMC “Samplers”

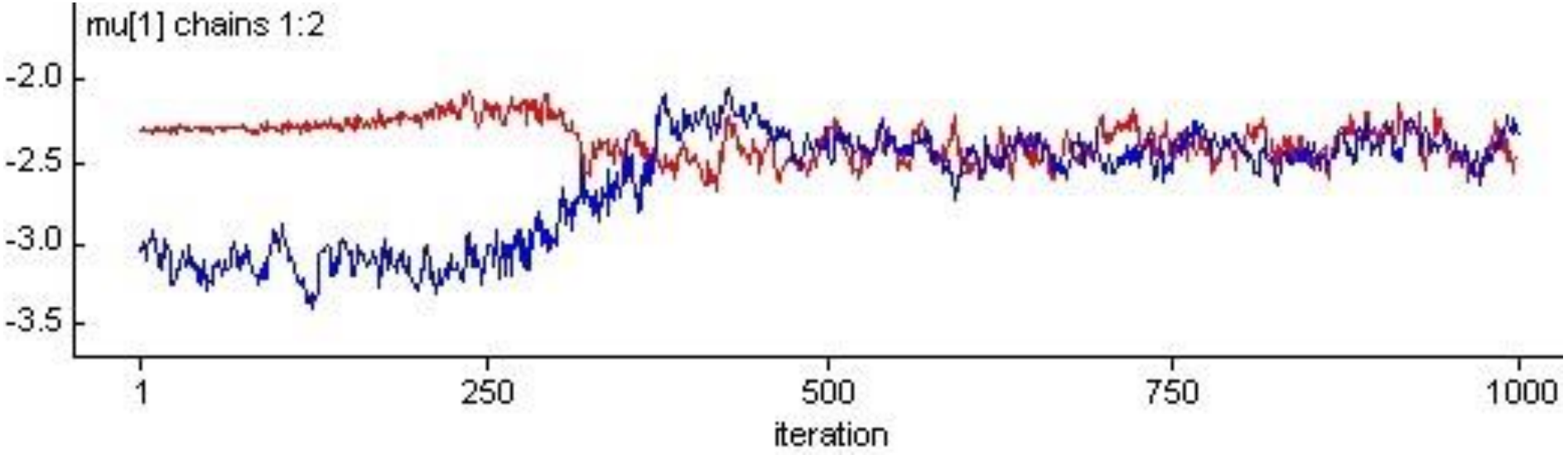
- Different from Classical notions of convergence
- Chain should be (nearly) independent of the starting points at the point where samples are recorded.
  - This first phase is called the burn-in or warmup
- Any parameters that “tune” the sampler should have been optimised.
- Convergence is also used in the sense that the posteriors that have been mapped in the post warmup phase accurately reflect the entire posterior.
  - i.e. a sampler can be working perfectly, yet need billions of iterations to map a posterior well

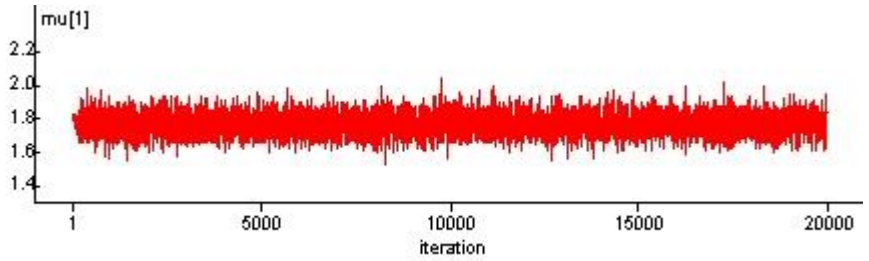




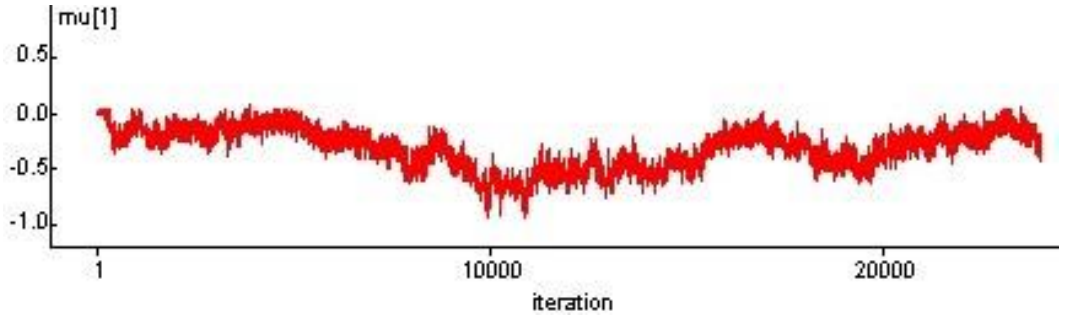


# Visual Convergence (Source Robert Grants talk (2013))

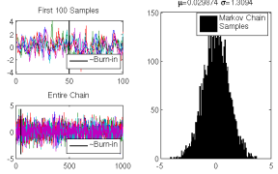



Good Mixing




Slow Mixing



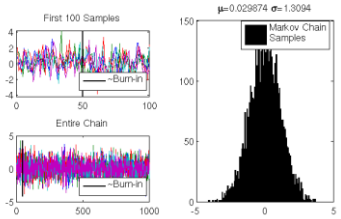

# Some Software (ALL freely available)

University of Reading

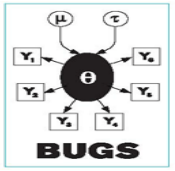
- **Winbugs**, Open Bugs, can also be run within R using several libraries and also can be called in Stata
- **JAGS** (Just Another Gibbs Sampler) – also runs within R (Rjags)
- **BayesM** (Runs on R)
- **PyMC** (2,3) Runs on Python
  - Note PyMC3 is a very different animal to PyMC2– uses Hamiltonian MCMC
- **Emcee** (Ensemble Sampler, runs on Python)
- **Stan** (Runs on CmdStan, R, Python, Stata, Julia, Matlab)

Of course many practitioners write their own code in Matlab, Gauss, Stata etc.

Programming today is a race between software engineers striving to build bigger and better idiot-proof programs, and the Universe trying to produce bigger and better idiots. So far, the Universe is winning."
— **Douglas Adams**




# Example 1. WinBUGS Mean and Variance




University of Reading

## Model

$$y_i = \mu + e_i$$

$$e_i \sim N(0, \sigma^2)$$

$$\tau = \sigma^{-2}$$

```
model
{
        for(i in 1:N)
        {
                y[i] ~ dnorm(mu,tau)
        }
        mu ~ dnorm(150, 0.001)
        tau ~ dgamma(0.001,0.001)
}
#----Initial values file-------------------------
list(mu =161,tau=1)

#----Data File-------------------------------
list(N= 10,
y=c(169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3))
```

Precision not Variance

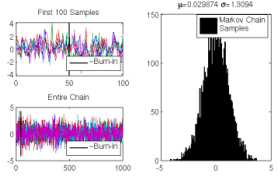
## Priors

$$\mu \sim N(150, 10^3)$$

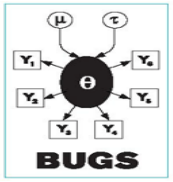
$$\tau \sim gamma(10^{-3}, 10^{-3})$$

See winbugs the movie…https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/winbugsthemovie.swf

And https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/



# Example 2. WinBUGS
# Bivariate Regression




University of Reading

```
model
{

for(i in 1:N) {
y[i] ~ dnorm(mu[i],tau)
mu[i] <- beta0 + beta1* x[i]
}
beta0 ~ dnorm(0,1.0E-6)
beta1 ~ dnorm(0,1.0E-6)
tau ~ dgamma(1.0E-3,1.0E-3)
}
#----Initial values file-------------------------
list(beta0 = 0, beta1 = 0, tau = 1)

#----Data File---------------------------------

list(N= 10,
y=c(169.6,166.8,157.1,181.1,158.4,165.6,166.7,156.5,168.1,165.3),
x =c(71.2,58.2,56.0,64.5,53.0,52.4,56.8,49.2,55.6,77.8))
```

## Model

$$y_i = \beta_0 + \beta_1 x + e_i$$

$$e_i \sim N(0, \sigma^2)$$

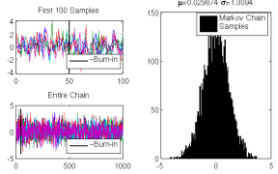
$$\tau = \sigma^{-2}$$

## Priors

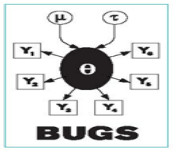
$$\beta_0 \sim N(0, 10^6)$$

$$\beta_1 \sim N(0, 10^6)$$

$$\tau \sim gamma(10^{-3}, 10^{-3})$$



# Example 3. WinBUGS Probit Regression




University of Reading

```
model
{
for(i in 1:N)
        {
                p[i] <-  phi(beta0 + beta1*(x[i]-mean(x[])))
                y[i] ~ dbern(p[i])
        }
        beta0 ~ dnorm(0,1.0E-0)
        beta1 ~ dnorm(0,1.0E-0)
}
#----Initial values file-------------------------
list(beta0 = 0, beta1 = 0)

#----Data File-----------------------------------

list(N= 25, y=c(1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,1,1,1,1,1,1,0,0,0,0,1),
        x=c(1,3,2,1,3,2,0,1,0,2,1,2,1,0,0,1,2,3,2,1,0,0,0,0,1))
```

## Model

$$y_i^* = \beta_0 + \beta_1(x - \bar{x}) + e_i$$

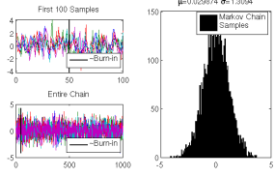
$$e_i \sim N(0,1)$$

$y_1 = 1$ if $y_i^* > 0$ and 0 otherwise

## Priors

$$\beta_0 \sim N(0,1)$$

$$\beta_1 \sim N(0,1)$$



# WinBUGS has a host of other examples you can follow

University of Reading

WinBUGS14

File Tools Edit Attributes Info Model Inference Options Doodle Map Text Window Help

Examples Volume I

BUGS

## Examples Volume 1

Rats: Normal hierarchical model

Pump: conjugate gamma-Poisson hierarchical model

Dogs: log linear binary model

Seeds: random effects logistic regression

Surgical: institutional ranking

Salm: extra-Poisson variation in dose-response study

Equiv: bioequivalence in a cross-over trial

Dyes: variance components model

Stacks: robust and ridge regression

Epil: repeated measures on Poisson counts

Blocker: random effects meta-analysis of clinical trials

Oxford: smooth fit to log-odds ratios in case control studies

LSAT: latent variable models for item-response data

Bones: latent trait model for multiple ordered catagorical responses

Inhalers: random effects model for ordinal responses from a cross-over trial

Examples Volume II

BUGS

## Examples Volume 2

Dugongs: a nonconjugate, nonlinear model

Orange trees: a hierarchical, nonlinear model

Multivariate Orange trees: a hierarchical, nonlinear model

Biopsies: latent class model

Eyes: normal mixture model

Hearts: a mixture model for count data

Air: covariate measurement error

Cervix: case-control study with errors in covariates

Jaw: repeated measures analysis of variance

Birats: a bivariate Normal hierarchical model

Schools: muliivariate hierarchical model of examination results

Ice: non-parametric smoothing in an age-cohort model

Beetles: logistic, probit and extreme value models

Alli: multinomial_logistic model

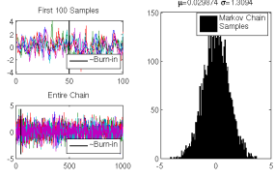
Endo: conditional inference in case-control studies

Stagnant: a change point problem

Asia: an expert system

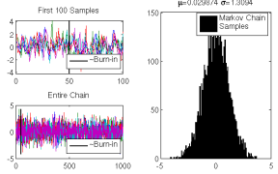
winbugs linear reg | Distributions | Distributions | WinBUGS User Manu... | probit

09:09
22/09/2017

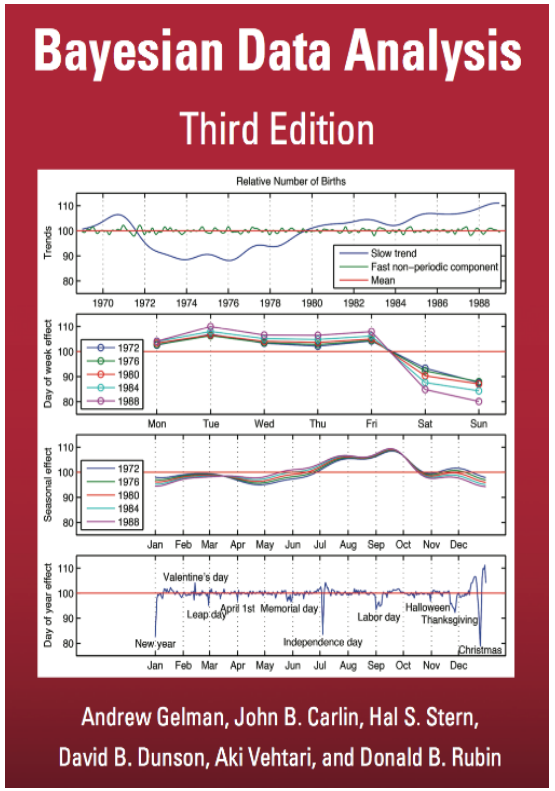



# Back to the Future…

- **Winbugs** is a great place to start, but not to finish
- To get good estimates Gibbs+ Metropolis can sometimes take days to run for some models.
- **PYMC3** and **Stan** use Hamiltonian Monte Carlo (HMC) which run using C (or Cython) which is very-very fast.
- These programs tune the parameters using something called the “NUTS” sampler see https://arxiv.org/pdf/1111.4246.pdf for a description of both HMC and NUTs
- Unlike Gibbs or Metropolis, HMC is not something that the average researcher would be able to write for themselves, but the PYMC3 and Stan platforms are very flexible.




# Pystan Example in Jupyter Notebook




dit View Insert Cell Kernel Help Trusted

Run Code

n [1]:
```
from prepare import *
```
prepared

[18]:
```
lin_reg_code = """
data { int k; int n; matrix[n,k] x;  vector[n] y; }

transformed data {}

parameters { vector[k] b; real sigma;}

transformed parameters { vector[n] mu;
                         for (i in 1:n) { mu[i] = x[i,:]*b;}    }

model { b ~ uniform(-10,10);
        sigma ~ uniform(0, 20);
        y ~ normal(mu, sigma);  }

generated quantities { vector[n] ys;
                       for (i in 1:n) {ys[i]= normal_rng(x[i,:]*b, sigma);}   }
"""
```

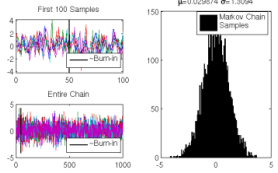
[19]:
```
sm = pystan.StanModel(model_code=lin_reg_code)
```
INFO:pystan:COMPILING THE C++ CODE FOR MODEL anon_model_2786f3f8d8656b3548b17d3d682e9592 NOW.

[20]:
```
savemodel(sm,'model3')
```

[21]:
```
sm = loadmodel('model3')
```

[31]:
```
linregdata=load('/home/kelvin/python/linregdata.xlsx')
y=linregdata['Y']
x=linregdata[['int','X1','X2']]
Y=frame(y)
```

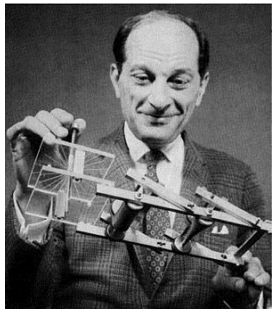
n [8]:
```
lin_reg_dat = {'k': cols(x),'n': rows(x),'x': x,'y': y}
```




# STAN – A regression example

http://mc-stan.org/

Andrew Gelman

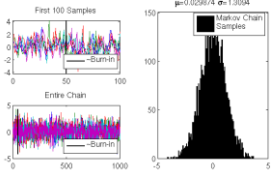


Stan Ulam
1909-1984

University of Reading

In [18]:
```
lin_reg_code = """
data { int k; int n; matrix[n,k] x;  vector[n] y; }

transformed data {}

parameters { vector[k] b; real sigma;}

transformed parameters { vector[n] mu;
                         for (i in 1:n) { mu[i] = x[i,:]*b;}    }

model { b ~ uniform(-10,10);
        sigma ~ uniform(0, 20);
        y ~ normal(mu, sigma);  }

generated quantities { vector[n] ys;
                       for (i in 1:n) {ys[i]= normal_rng(x[i,:]*b, sigma);}   }
"""
```

Lots More Examples here



# Stan Output

```
fit = sm.sampling(data=lin_reg_dat, iter=6000,
warmup=1000, chains=2)
```

```
fit
```

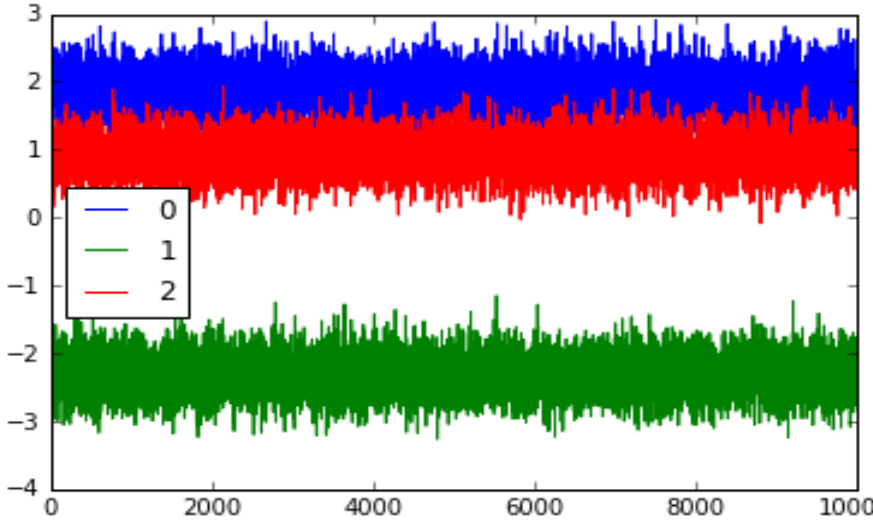
```
Inference for Stan model: anon_model_2786f3f8d8656b3548b17d3d682e9592.
2 chains, each with iter=6000; warmup=1000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=10000.
```

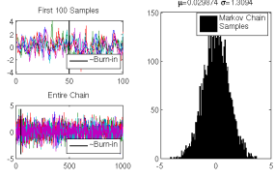
| | mean | se_mean | sd | 2.5% | 25% | 50% | 75% | 97.5% | n_eff | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| b[0] | 1.87 | 2.9e-3 | 0.29 | 1.29 | 1.67 | 1.87 | 2.06 | 2.45 | 10000 | 1.0 |
| b[1] | -2.32 | 2.8e-3 | 0.28 | -2.86 | -2.51 | -2.33 | -2.14 | -1.77 | 10000 | 1.0 |
| b[2] | 0.93 | 2.7e-3 | 0.27 | 0.39 | 0.74 | 0.93 | 1.11 | 1.47 | 10000 | 1.0 |
| sigma | 2.9 | 2.1e-3 | 0.21 | 2.53 | 2.75 | 2.89 | 3.04 | 3.35 | 10000 | 1.0 |
| mu[0] | -1.25 | 4.9e-3 | 0.49 | -2.19 | -1.58 | -1.25 | -0.93 | -0.28 | 10000 | 1.0 |
| mu[1] | 0.82 | 6.9e-3 | 0.68 | -0.54 | 0.36 | 0.81 | 1.27 | 2.14 | 9774.0 | 1.0 |
| mu[2] | 3.25 | 4.9e-3 | 0.49 | 2.3 | 2.92 | 3.25 | 3.58 | 4.22 | 10000 | 1.0 |
| mu[3] | 1.19 | 3.3e-3 | 0.33 | 0.54 | 0.97 | 1.19 | 1.41 | 1.84 | 10000 | 1.0 |
| mu[4] | 0.55 | 4.2e-3 | 0.42 | -0.29 | 0.27 | 0.56 | 0.83 | 1.38 | 10000 | 1.0 |
| mu[5] | 3.55 | 3.8e-3 | 0.38 | 2.81 | 3.3 | 3.55 | 3.8 | 4.29 | 10000 | 1.0 |
| mu[6] | 1.57 | 3.5e-3 | 0.35 | 0.88 | 1.34 | 1.57 | 1.8 | 2.25 | 10000 | 1.0 |
| mu[7] | 1.07 | 5.3e-3 | 0.52 | 9.0e-3 | 0.72 | 1.07 | 1.41 | 2.07 | 9932.0 | 1.0 |
| mu[8] | 0.91 | 3.5e-3 | 0.35 | 0.21 | 0.68 | 0.92 | 1.14 | 1.59 | 10000 | 1.0 |
| mu[9] | -0.09 | 4.0e-3 | 0.4 | -0.9 | -0.36 | -0.09 | 0.17 | 0.68 | 10000 | 1.0 |

Effective Sample Size

Evaluating Convergence
Rhat should be <1.1

Trace Plot for bs






# Calculating the Marginal Likelihood

- Sometimes the Marginal Likelihood can be deduced analytically
  - (e.g. in the case of the normal linear model with normal-gamma priors)
- Not done automatically in most software. Popular methods are:
  - Gelfand-Dey Method
  - Chib + Chib and Jeliazkov Method
  - Bridge Sampling (which can be very difficult) but required for mixture models
- Warning: The Marginal Likelihood can be particularly sensitive to Priors (more sensitive than the parameters themselves).

# Part 4: Some Frequently Asked Questions

# Doesn’t the use of the Prior undermine the objectivity of my results?

- …it it is misleading to think that the required subjectivity always takes the form of prior belief. …….first, prior distributions are not necessarily any more subjective than other aspects of a statistical model; indeed, in many applications priors can and are estimated from data ……..second, somewhat arbitrary choices come into many aspects of statistical models, Bayesian and otherwise, …….. we think it is a mistake to consider the prior distribution as the exclusive gate at which subjectivity enters a statistical procedure.

- **Beyond subjective and objective in statistics, Gelman and Henning, 2015**

# Are Bayesian Models Different to Classical Models?

- For the most part NO, e.g. Linear Models, Logits, Probits, Tobits, Heckman, Stochastic Frontier, Random effects models etc…
- While ‘philosophically’ the estimates represent different things, they can often look practically identical (but not always)
- The similarity can be deceptive, however, because of the different interpretation of the outputs

# When will Bayesian and Classical Estimates of the same models be very different?

University of Reading

- When the Prior is highly informative
- When inequality parameter restrictions are imposed – especially when the unconstrained classical estimates would not satisfy the bounds
- When the likelihood function is not `well behaved'.
  - where the likelihood has multiple modes
  - where the likelihood is nearly flat over some regions

"Bayesianism assumes: (a) Either a weak or uniform prior, in which case why bother?, (b) Or a strong prior, in which case why collect new data?, (c) Or more realistically, something in between, in which case Bayesianism always seems to duck the issue." -Ehrenberg

# Are there “Classical Models” that do not have a precise Bayesian analogue?

- Yes
- Many “classical approaches” do not specify a likelihood function.
- But:
  - Many non-likelihood approaches can be approximated using likelihood ones
  - there are variants of Bayesian models that have Bayesian Roots, but do not require exact likelihood functions.
    - Approximate Bayesian computation (ABC))
    - “Empirical likelihood” approaches e.g. Bayesian Quantile regression.

# When might a Bayesian Approach Help Me?

University of Reading

- When you have reliable prior information that you want to embody in your estimates
- Maximum Likelihood may fail to converge in practice. ‘Weak’ prior information can often give a well defined posterior.
- High dimensionality can be a curse for Classical Methods, but often do not present such an obstacle to Bayesian Approaches
- Inequality restrictions in a classical setting can be tricky when they are binding. Bayesian Settings will give an estimate inside the interval, and classical will give ones on the edge.

# What sort of thing could I do in a Bayesian setting that I could not do using a Classical approach?

University of Reading

- Model Averaging (properly)
- MCMC which “jumps” from model to model (depending on the model worth –including variable selection)
- Infinite Mixture Models (number of mixtures treated like an additional parameter – Dirichlet Process Models )
- Non-Linear State Space models using sequential MC\particle filters