Every dance is a kind of fever chart, a graph of the heart.

Martha Graham
Directed Acyclic Graphs (DAGs)

Nodes are constant, stochastic, or logical.

Constants must have values (data) and can’t have parents.

Stochastic nodes can be observed (data), unobserved (parameters), or missing values.

Logical nodes are functions of other nodes (not data).

Links into stochastic nodes are singles arrows $\rightarrow$, indicating stochastic dependence, and links into logical nodes are double arrows $\Rightarrow$.

Repeated parts of the graph are indicated by a plate and indexing $i = 1, \ldots, N$. 
Example: binomial

\[ \pi \sim \text{dunif}(0, 1) \]
\[ n \quad \text{constant; data} \]
\[ Y \sim \text{dbin}(p, n) \quad \text{data} \]

Here we know \( n \), observe \( y \), and want the posterior of \( \pi \) (which has a uniform prior).

The WinBUGS convention is to parameterise the binomial with the proportion first and number of trials second.
Example: normal population

\[ Y_i \sim \text{dnorm}(\mu, \tau) \quad \text{data} \]
\[ \mu \sim \text{dnorm}(\mu_{\mu}, \tau_{\mu}) \]
\[ \tau \sim \text{dgamma}(m_\tau, \lambda_\tau) \]
\[ \sigma = 1/\tau \]

The WinBUGS convention is to parameterise the normal with the mean and precision (1/variance), and to parameterise the gamma with the shape and rate (1/scale).

We observe \( y_1, \ldots, y_n \) and want the posteriors of \( \mu \) and \( \sigma \).

Here we assume that \( \mu_{\mu}, \tau_{\mu}, m_\tau, \lambda_\tau \) are given, and \( n \) is part of the data (together with the \( y_i \)).
Example: comparing two equivariant populations

\[ Y_i \sim \text{dnorm}(\mu_1, \tau) \quad \text{data} \]
\[ Z_i \sim \text{dnorm}(\mu_2, \tau) \quad \text{data} \]
\[ \tau \sim \text{dgamma}(0.1, 0.1) \]
\[ \mu_1 \sim \text{dnorm}(0, 0.001) \]
\[ \mu_2 \sim \text{dnorm}(0, 0.001) \]
\[ d = \mu_1 - \mu_2 \]

Here \( d \) is the difference between the two population means, and we want its posterior.
Example: expected survival time

For a particular group of patients undergoing surgery, the chance of surviving surgery is $p$ say, and for those who survive, the survival time has an $\exp(\theta)$ distribution (where we follow the WinBUGS convention, and $\theta$ is the rate).

We are interested in the expected survival time for a patient about to undergo surgery, that is, $L := p/\theta$.

Suppose that $p$ has a $\beta(1, 1)$ prior and $\theta$ has a $\Gamma(0.001, 0.001)$ prior, then we are interested in the posterior of $L$. 
\begin{align*}
n & \quad \text{constant; data} \\
p & \quad \sim \ dbeta(1, 1) \\
Y & \quad \sim \ \text{bin}(p, n) \quad \text{data} \\
\theta & \quad \sim \ \text{dgamma}(0.001, 0.001) \\
S_i & \quad \sim \ \text{dexp}(\theta) \quad \text{data} \\
L & = \ p/\theta
\end{align*}
Conditional independence

We say that $X$ and $Y$ are conditionally independent given $Z$, or independent conditioned on $Z$, if

$$p_{X,Y|Z}(x,y|z) = P(X = x, Y = y|Z = z)$$

$$= P(X = x|Z = z)P(Y = y|Z = z)$$

$$= p_{X|Z}(x|z)p_{Y|Z}(y|z)$$

These can be densities as well as probabilities.

We write $X \perp \perp Y|Z$.

We interpret links in DAGs as follows:

Conditional on its parent nodes $pa[X]$, node $X$ is independent of all other nodes except its descendents.
**Theorem** Let $G$ be a DAG with stochastic and logical nodes $\mathbf{X} = (X_1, \ldots, X_n)$, then it has likelihood

$$p_{\mathbf{X}}(\mathbf{x}) = \prod_i p_{X_i|\text{pa}[X_i]}(x_i|\text{pa}[x_i])$$

where $\text{pa}[x_i]$ are the observed values of $\text{pa}[X_i]$, the parents of $X_i$. That is, the joint distribution for all nodes is the product of the conditional distributions for each node conditioned on its parents.

Note that $\mathbf{X}$ includes observed and unobserved variables, so some elements of $\mathbf{x}$ will be fixed by observations, and others remain variable.
Proof (sketch) we consider two types of subgraph...
<table>
<thead>
<tr>
<th>DAGs</th>
<th>Likelihood</th>
<th>Bayesian Regression Models</th>
<th>Priors</th>
</tr>
</thead>
</table>

3. Bayesian models
Linear regression with Normal responses

\[ Y_i \sim \text{dnorm}(\mu_i, \tau) \]
\[ \mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \]
\[ \beta_j \sim \text{dnorm}(0, (0.01)^2) \]
\[ \tau \sim \text{dgamma}(0.001, 0.001) \]

Note that \( \beta_j \) has variance 10,000 and \( \tau \) has variance 1000, so our priors are uninformative.
Remarks:

1. WinBUGS does not require the $x_{ij}$ to be included in the model, just in the data.

2. For these models, sampling from the posterior using MCMC works better if the $x_{ij}$ are centered. That is, for $\bar{x}.j = \sum_i x_{ij} / n$ we put

$$\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j (x_{ij} - \bar{x}.j)$$
Example: average household income

Households in 284 Swedish municipalities were sampled. For each municipality we have the average household income (INCOME), the average age of the head of household (AVAGE), and a measure of how rural the municipality is (RURAL: 1 urban, 2 mixed, 3 rural).

income-model.txt
income-data2.txt
income-in2.txt

What is the model? Priors? Posterior means and credible intervals?

Note the use of double indexing for the categorical variable.
### 3. Bayesian models

<table>
<thead>
<tr>
<th>DAGs</th>
<th>Likelihood</th>
<th>Bayesian Regression Models</th>
<th>Priors</th>
</tr>
</thead>
</table>


But what about the outlier?

```r
> df <- data.frame(source("../scripts_data/income-data2.txt")$value)
> df$RURAL <- factor(df$RURAL)
> par(mfrow=c(2,2))
> plot(INCOME ~ AVAGE, data=df, col=RURAL, pch=as.character(RURAL))
> plot(INCOME ~ RURAL, data=df)
```
Rather than delete the outlier, we can allow for it by using a heavy tailed distribution for the responses, such as a $t_4$ distribution. This is called a *robust* model.

How does this change the fit?

Looking at credible intervals, should we drop the RURAL variable?
If $X \sim t_k$ then $\mathbb{E}X = 0$, $\text{Var} X = k/(k - 2)$ (for $k > 2$) and the density is:

$$f_X(x) \propto \frac{1}{(1 + t^2/k)^{(k+1)/2}}.$$

Put $Y = \mu + X/\sqrt{\tau}$ then $\mathbb{E}Y = \mu$, $\text{Var} Y = \tau^{-1}(k/(k - 2))$ (for $k > 2$) and the density is

$$f_Y(y) = f_X(x(y)) \frac{dx(y)}{dy} \propto \frac{1}{(1 + \tau(y - \mu)^2/k)^{(k+1)/2} \sqrt{\tau}}.$$

In WinBUGS notation $Y \sim \text{dt}(\mu, \tau, k)$.  

3. Bayesian models
Generalised linear models

We model the mean response as for a linear model, but include a link function. For example, for a binomial regression:

\[
Y_i \sim \text{dbin}(p_i, m_i) \\
\quad i = 1, \ldots, n \\
\]

\[
p_i = g^{-1}(\beta_0 + \sum_{j=1}^{k} \beta_j (x_{ij} - \bar{x}_j)) \\
\]

\[
\beta_j \sim \text{dnorm}(0, (0.01)^2) \\
\quad j = 0, 1, \ldots, k
\]
Example: beetle mortality

Numbers of beetles killed $r$ out of number exposed $n$ to some concentration of carbon disulphide $x$ for five hours.

Fit using a logit link and use the posterior of $p_i$ to check the fit.

beetles2.txt

Note how WinBUGS specifies the link function. (The probit link can actually be specified in two ways in WinBUGS.)

WinBUGS allows the logit, probit and cloglog links. The cloglog link gives a better fit in this case. (Larger proportions are more likely under cloglog than logit.)
3. Bayesian models
Nonlinear regression

We illustrate with an example. We have data on the length $Y$ and age $x$ of 27 dugongs captured off the Queensland coast (Carlin and Gelfand 1991).
The von Bertalanffy growth model supposes that the mean length at age $x$ is

$$\mu(x) = L_\infty - (L_\infty - L_0)e^{-Kx}$$

We put

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \mu(x_i)$$

How to model $L_0$, $L_\infty$, and $K$?
Prior 1 \( L_0 \sim U(0, 100), \quad L_\infty \sim U(0, 100), \quad K \sim U(0, 100) \)

Problem with this is that we can’t guarantee \( L_\infty > L_0 \).

Prior 2 \( L_0 \sim U(0, 100), \quad L_\infty \sim L_0 + U(0, 100), \quad K \sim U(0, 100) \)

Better, but note that \( L_\infty \) now has a triangular rather than a uniform prior.

Prior 3 Put \( \alpha = L_0, \beta = L_\infty - L_0, \) and \( \gamma = e^{-K} \), then

\[ \mu_i = \alpha + \beta(1 - \gamma^{x_i}) \]

and we might take as priors

\[ \alpha \sim U(0, 100), \quad \beta \sim U(0, 100), \quad \gamma \sim U(0, 1) \]

Note that \( K = -\log \gamma \sim \exp(1) \).
<table>
<thead>
<tr>
<th>DAGs</th>
<th>Likelihood</th>
<th>Bayesian Regression Models</th>
<th>Priors</th>
</tr>
</thead>
</table>

3. Bayesian models
Convictions are more dangerous foes of truth than lies.

F. Nietzsche
Formally we like to distinguish between ‘noninformative’ and ‘informative’ priors. The former are used when we wish scientific objectivity and don’t want prior judgements to influence our conclusions.

The term noninformative is misleading however, as all priors contain some information. Many statisticians prefer the terms ‘vague’, ‘diffuse’, ‘reference’ or ’objective to noninformative. In general we should think of the choice of prior as part of the model building process, just as the choice of likelihood for the observed data requires a choice. Influential priors are not necessarily bad, but we need to pay attention to the magnitude of their influence.
Jeffreys priors

Jeffreys provided a sensible candidate for a vague/reference prior.

Suppose we have a parameter $\theta$ and we want the prior that says the least about $\theta$. Then this prior should also say the least about any transform of $\theta$, say $\phi = f(\theta)$. It follows that this prior can not be uniform, since

$$p_\theta(x) dx = \mathbb{P}(\theta \in (x, x + dx))$$
$$= \mathbb{P}(\phi \in (f(x), f(x) + f'(x) dx))$$
$$= p_\phi(f(x))|f'(x)|dx$$

**Example:** Suppose $\theta \sim U(0, 1)$ and $\phi = \theta^2$ then

$$p_\phi(y) = \frac{1}{2\sqrt{y}}$$

3. Bayesian models
Jeffreys priors are defined using a property that is retained under transformations. Suppose that $\theta$ is a parameter in some log-likelihood $l(x; \theta)$, then put

$$p_\theta(\theta) \propto \sqrt{\mathcal{I}(\theta)}$$

where $\mathcal{I}(\theta)$ is the Fisher information:

$$\mathcal{I}(\theta) = \mathbb{E} \left( \left( \frac{\partial l(X; \theta)}{\partial \theta} \right)^2 \right) = \mathbb{E} \left( \left( \frac{\partial l(X; \phi)}{\partial \phi} \right)^2 \left( \frac{d\phi}{d\theta} \right)^2 \right) = \mathcal{I}(\phi) \left( \frac{d\phi}{d\theta} \right)^2$$

So for $\phi = f(\theta)$, $p_\theta(\theta) \propto \sqrt{\mathcal{I}(\theta)}$ implies

$$p_\phi(\phi) = p_\theta(\theta) \left| \frac{d\theta}{d\phi} \right| \propto \sqrt{\mathcal{I}(\phi)}$$
Jeffreys prior for a location parameter

A location parameter $\theta$ is such that the distribution of $X - \theta$ does not depend on $\theta$.

It follows that the likelihood $l(x; \theta)$ is a function of $x - \theta$, so if we put $g(x - \theta) = \frac{\partial l(x; \theta)}{\partial (x - \theta)}$, then $g$ also does not depend on $\theta$. Thus

$$I(\theta) = \mathbb{E} \left( \left( \frac{\partial l(X; \theta)}{\partial \theta} \right)^2 \right)$$

$$= \mathbb{E} g(X - \theta)^2 (-1)^2$$

does not depend on $\theta$

That is $p_\theta(\theta) \propto 1$. 

3. Bayesian models
An improper prior

Since $X - \theta$ does not depend on $\theta$, we must have $\theta \in (-\infty, \infty)$, which is a problem, because there is no (prior) distribution that is uniform on $(-\infty, \infty)$.

Sometimes we can actually just ignore this problem. For example suppose $X \sim N(\theta, 1)$ and $p_\theta(\theta) \propto 1$, then

$$p_{\theta|X}(\theta|x) \propto p_X(x|\theta)p_\theta(\theta) \propto p_X(x|\theta) \propto e^{-(x-\theta)/2}$$

That is $\theta|\{X = x\} \sim N(x, 1)$, which is a perfectly respectable posterior, even though the prior doesn’t exist. In cases like this we call the prior *improper*. 
In practice, for location parameters we can approximate the improper $U(-\infty, \infty)$ distribution with either a $U(-K, K)$ for a very large $K$ (large enough that you can not imagine $|\theta| > K$), or a $N(0, K^2)$ for $K$ large.

If we take $X \sim N(\theta, 1)$ and $\theta \sim U(-K, K)$ then the posterior is a truncated normal, $\theta|\{X = x\} \sim N(x, 1)1_{(-K,K)}(\theta)$, which converges to a $N(x, 1)$ as $K \to \infty$ (the posterior we got with the improper prior).

If we take $X \sim N(\theta, 1)$ and $\theta \sim N(0, K^2)$ then $\theta|\{X = x\} \sim N\left(\frac{x}{1+K^{-2}}, \frac{1}{1+K^{-2}}\right)$, which again converges to a $N(x, 1)$ as $K \to \infty$.

So (at least in this case), we can formally consider the improper $U(-\infty, \infty)$ prior as a limit of either a $U(-K, K)$ prior or a $N(0, K^2)$ prior, as $K \to \infty$. 
Jeffreys prior for a scale parameter

\( \sigma \) is a scale parameter for \( X \) if the distribution of \( X/\sigma \) does not depend on \( \sigma \). That is, \( p_{X|\sigma}(x) = \sigma^{-1}f(x/\sigma) \) for some \( f \).

To see why, Put \( Y = X/\sigma \) then we can write formally

\[
p_Y(y)\,dy = \mathbb{P}(Y \in (y, y + dy)) = \mathbb{P}(X \in (\sigma y, \sigma y + \sigma dy)) = p_{X|\sigma}(\sigma y)\sigma \,dy
\]

As \( Y \) does not depend on \( \sigma \), we can write \( p_Y(y)\,dy = f(y)\,dy \), whence \( p_{X|\sigma}(x) = f(x/\sigma)/\sigma \).
Now

\[
\frac{\partial \log p_{X|\sigma}(x)}{\partial \sigma} = \frac{\partial \log \sigma^{-1}}{\partial \sigma} + \frac{\partial \log f(x/\sigma)}{\partial \sigma} \\
= -\frac{1}{\sigma} + \frac{\partial \log f(x/\sigma)}{\partial (x/\sigma)} \frac{\partial (x/\sigma)}{\partial \sigma} \\
= -\frac{1}{\sigma} + \frac{\partial \log f(x/\sigma)}{\partial (x/\sigma)} - \frac{x}{\sigma^2} \\
= -\frac{1}{\sigma} g(x/\sigma)
\]

where

\[
g(x/\sigma) = 1 + \frac{x}{\sigma} \frac{\partial \log f(x/\sigma)}{\partial (x/\sigma)}.
\]
Thus

\[
\frac{\partial^2 \log p_{X|\sigma}(x)}{\partial \sigma^2} = \frac{1}{\sigma^2} g(x/\sigma) - \frac{1}{\sigma} g'(x/\sigma) \frac{\partial(x/\sigma)}{\partial \sigma}
\]

\[
= \frac{1}{\sigma^2} \left( g(x/\sigma) + \frac{x}{\sigma} g'(x/\sigma) \right)
\]

and so the Fisher information is given by

\[
\mathcal{I}(\sigma) = -\mathbb{E} \frac{\partial^2 \log p_{X|\sigma}(X)}{\partial \sigma^2}
\]

\[
= -\frac{1}{\sigma^2} \mathbb{E} (g(Y) + Y g'(Y))
\]

\[
\propto \frac{1}{\sigma^2}
\]

since \( Y \) does not depend on \( \sigma \).

So the Jeffreys prior for \( \sigma \) is proportional to \( \sqrt{\mathcal{I}(\sigma)} \propto 1/\sigma \).
Put $\phi = \sigma^k$ then since $p_\phi(\phi) = p_\sigma(\sigma) |d\sigma/d\phi|$ we have that the Jeffreys prior for $\phi$ is proportional to

$$\frac{1}{\sigma} \frac{1}{k \sigma^{k-1}} \propto \frac{1}{\sigma^k} = \frac{1}{\phi}.$$ 

In particular, the Jeffreys prior for the precision, $\tau = 1/\sigma^2$, is proportional to $1/\tau$.

We must have $\phi \in (0, \infty)$, so the Jeffreys prior is proportional to $1/\phi$ for $\phi > 0$, which is again improper, since $\int_0^\infty x^{-1} dx = \infty$.

In practice we approximate the Jeffreys prior with a $\Gamma(\epsilon, \epsilon)$ distribution, for $\epsilon$ small. The $\Gamma(\epsilon, \epsilon)$ density is proportional to

$$e^{-\epsilon \phi} \phi^{\epsilon-1}.$$
Turning scale parameters into location parameters

An alternative approach to scale parameters is to take a log transform.

If $\sigma$ is a scale parameter for $X$ then $\log \sigma$ is a location parameter for $\log X$, so rather than take $\sigma \sim \Gamma(\epsilon, \epsilon)$ we could use

$$\log \sigma \sim U(-K, K) \text{ or } N(0, K^2).$$
Jeffreys prior for a proportion

If $X \sim \text{bin}(m, p)$ then the Jeffreys prior for $p$ is a $\beta(\frac{1}{2}, \frac{1}{2})$. The proof is left as an exercise.

The other common choice of prior for $p$ is the uniform (equivalent to a $\beta(1, 1)$). Compared to the uniform, the Jeffreys prior puts more weight near 0 and 1.

Recall that if $p_p \sim \beta(u, v)$ then $p_p|X \sim \beta(u + x, v + m - x)$, so the difference between a uniform and a Jeffreys prior is not great.