Statistical methods for linguistic research: Advanced Tools

Shravan Vasishth

Department of Linguistics
University of Potsdam, Germany

August 12, 2015

Today's goals

In this lecture, my goals are to

1. Give you a feeling for how Bayesian analysis works using five relatively simple examples.
2. Start thinking about priors for parameters in preparation for fitting linear mixed models.
3. Start fitting linear regression models in JAGS.

I will assign some homework which is designed to help you understand these concepts. Solutions are provided at the end of the exercise sheet so you can check them yourself.

Random variables

A random variable X is a function $X : S \to \mathbb{R}$ that associates with each outcome $\omega \in S$ exactly one number $X(\omega) = x$.
S_X is all the x's (all the possible values of X, the support of X).
I.e., $x \in S_X$.
Good example: number of coin tosses till H

- $X : \omega \to x$
- ω: H, TH, TTH,... (infinite)
- $x = 0, 1, 2, \ldots; x \in S_X$

Random variables

Every discrete (continuous) random variable X has associated with it a **probability mass (distribution) function (pmf, pdf)**.
PMF is used for discrete distributions and PDF for continuous.

$$p_X : S_X \to [0,1] \tag{1}$$

defined by

$$p_X(x) = P(X(\omega) = x), x \in S_X \tag{2}$$

Random variables

Probability density functions (continuous case) or probability mass functions (discrete case) are functions that assign probabilities or relative frequencies to all events in a sample space.
I will use the convention that the expression

$$X \sim f(\cdot) \tag{3}$$

means that the random variable X has pdf/pmf $f(\cdot)$. For example, if we say that $X \sim Normal(\mu, \sigma^2)$, we are assuming that the pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x-\mu)^2}{2\sigma^2}] \tag{4}$$

Random variables

We also need a **cumulative distribution function** or cdf because, in the continuous case, P(X=some point value) is zero and we therefore need a way to talk about P(X in a specific range). cdfs serve that purpose.
In the continuous case, the cdf or distribution function is defined as:

$$P(x < k) = F(x < k) = \text{The area under the curve to the left of k} \tag{5}$$

For example, suppose $X \sim Normal(600, 50)$.

Random variables

We can ask for $Prob(X < 600)$:

```
pnorm(600,mean=600,sd=sqrt(50))
```

```
## [1] 0.5
```

We can ask for the quantile that has 50% of the probability to the left of it:

```
qnorm(0.5,mean=600,sd=sqrt(50))
```

```
## [1] 600
```

. . . or to the right of it:

```
qnorm(0.5,mean=600,sd=sqrt(50),lower.tail=FALSE)
```

```
## [1] 600
```

Random variables

We can also calculate the probability that X lies between 590 and 610: $Prob(590 < X < 610)$:

```
pnorm(610,mean=600,sd=sqrt(50))-
  pnorm(490,mean=600,sd=sqrt(50))

## [1] 0.9213504
```

Random variables

Another way to compute the area under the curve is by simulation:

```
x<-rnorm(10000,mean=600,sd=sqrt(50))
## proportion of cases where
## x is less than 500:
mean(x<590)

## [1] 0.0829

## theoretical value:
pnorm(590,mean=600,sd=sqrt(50))

## [1] 0.0786496
```

We will be doing this a lot.

Random variables

E.g., in linguistics we take as continous random variables:

1. reading time: Here the random variable (RV) X has possible values ω ranging from 0 ms to some upper bound b ms (or maybe unbounded?), and the RV X maps each possible value ω to the corresponding number (0 to 0 ms, 1 to 1 ms, etc.).
2. acceptability ratings (technically not correct; but people generally treat ratings as continuous, at least in psycholinguistics)
3. EEG signals: measured in microvolts.

In this course, due to time constraints, we will focus almost exclusively on reading time data (eye-tracking and self-paced reading).

Normal random variables

We will also focus mostly on the normal distribution.

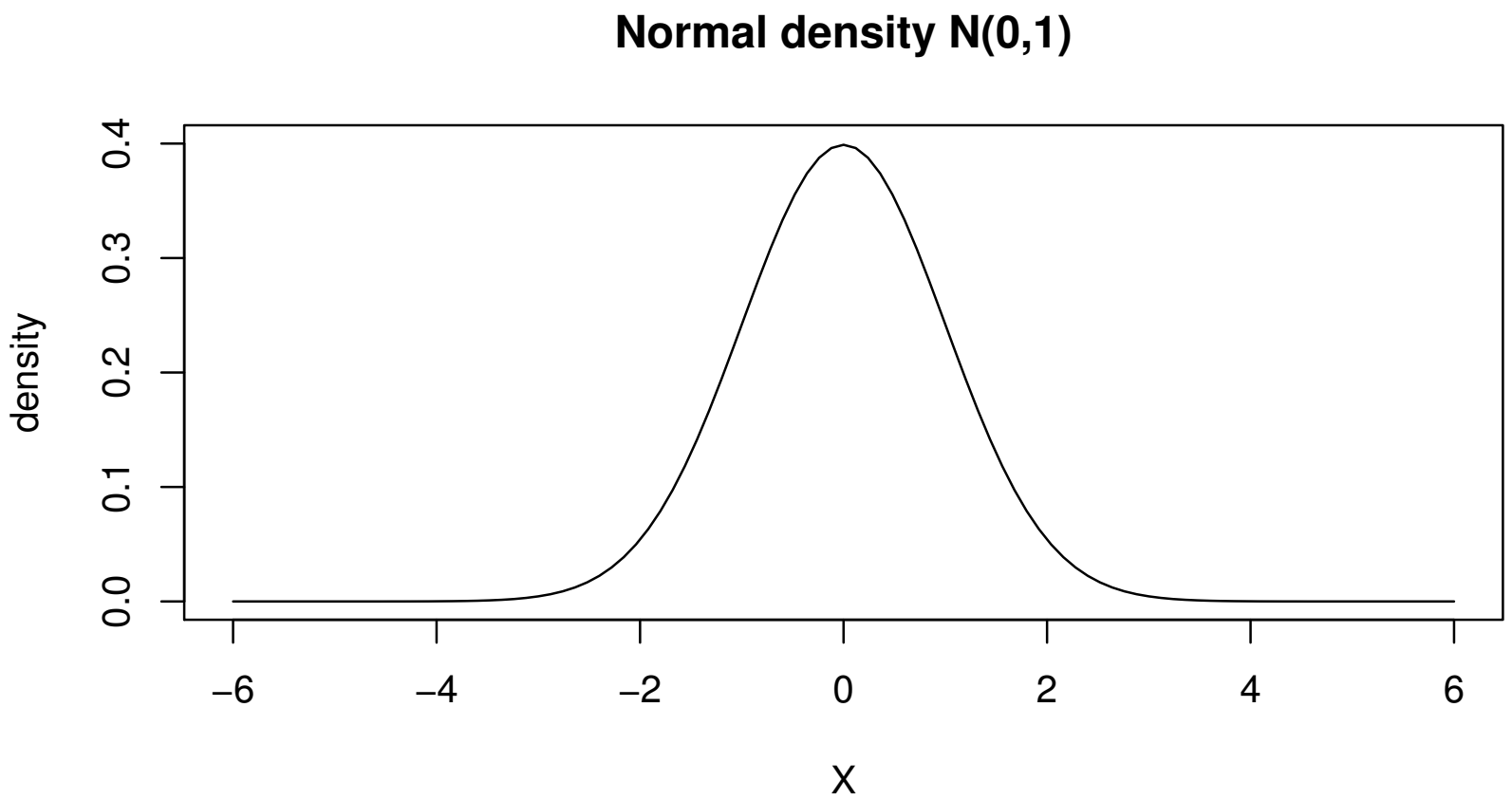
$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty. \tag{6}$$

It is conventional to write $X \sim N(\mu, \sigma^2)$.
Important note: The normal distribution is represented differently in different probabilistic programming languages:

1. R: `dnorm(mean,sigma)`
2. JAGS: `dnorm(mean,precision)` where precision $= 1/$variance
3. Stan: `normal(mean,sigma)`

Please be careful about this.

Normal random variables



Normal random variables

Standard or unit normal random variable:
If X is normally distributed with parameters μ and σ^2, then $Z = (X - \mu)/\sigma$ is normally distributed with parameters $0, 1$.
We conventionally write $\Phi(x)$ for the CDF:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy \quad \text{where } y = (x - \mu)/\sigma \tag{7}$$

In R, we can type `pnorm(x)`, to find out $\Phi(x)$. Suppose $x = -2$:

```
pnorm(-2)

## [1] 0.02275013
```

Normal random variables

If Z is a standard normal random variable (SNRV) then

$$p\{Z \leq -x\} = p\{Z > x\}, \quad -\infty < x < \infty \tag{8}$$

We can check this with R:

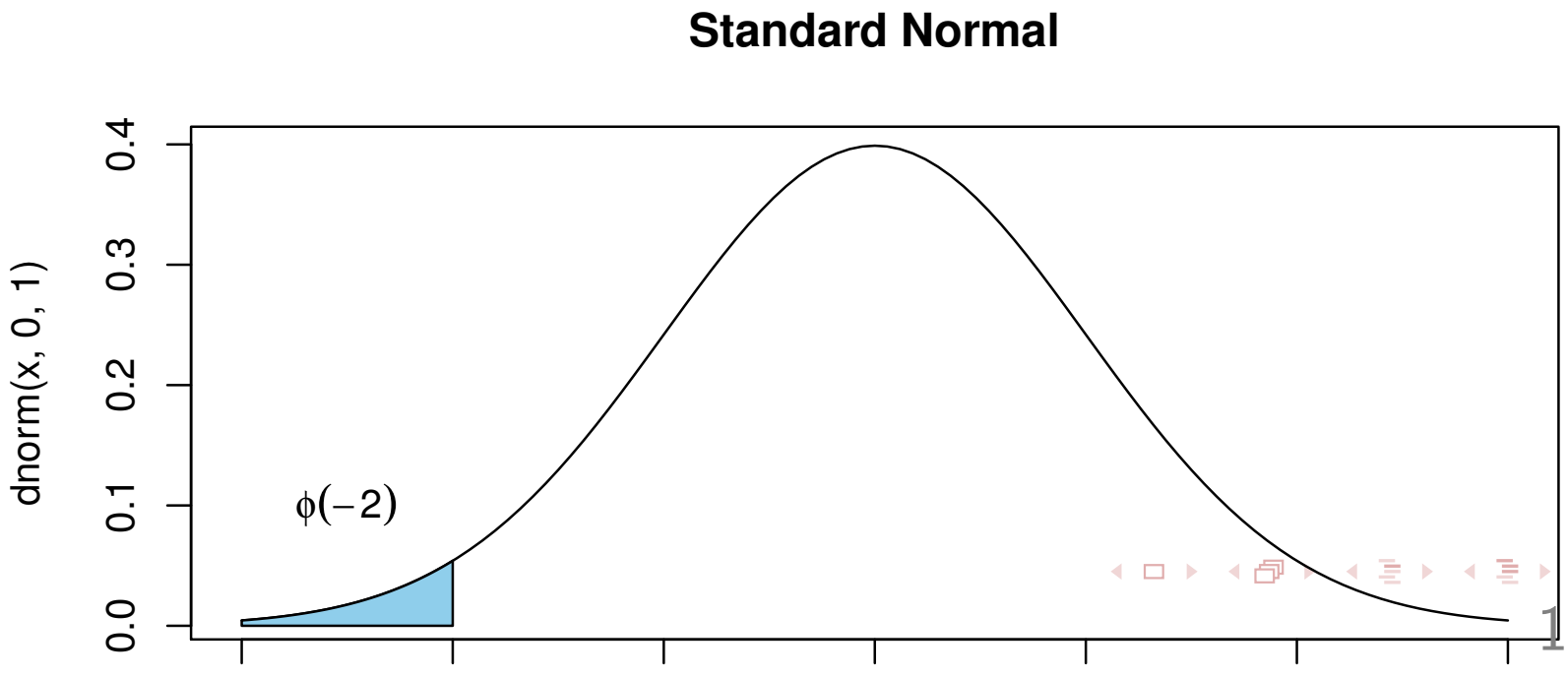
```
##P(Z < -x):
pnorm(-2)

## [1] 0.02275013

##P(Z > x):
pnorm(2,lower.tail=FALSE)

## [1] 0.02275013
```

Normal random variables

Although the following expression looks scary:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy \quad \text{where } y = (x - \mu)/\sigma \tag{9}$$

all it is saying is "find the area under the normal curve, ranging -Infinity to x. Always visualize! $\Phi(-2)$:



Normal random variables

Since $Z = ((X - \mu)/\sigma)$ is an SNRV whenever X is normally distributed with parameters μ and σ^2, then the CDF of X can be expressed as:

$$F_X(a) = P\{X \leq a\} = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right) \quad (10)$$

Practical application: Suppose you know that $X \sim N(\mu, \sigma^2)$, and you know μ but not σ. If you know that a 95% confidence interval is [-q,+q], then you can work out σ by

a. computing the $Z \sim N(0,1)$ that has 2.5% of the area to its right:

```
round(qnorm(0.025,lower.tail=FALSE),digits=2)

## [1] 1.96
```

b. Solve for σ in $Z = \frac{q-\mu}{\sigma}$.

Normal random variables

Summary of useful commands:

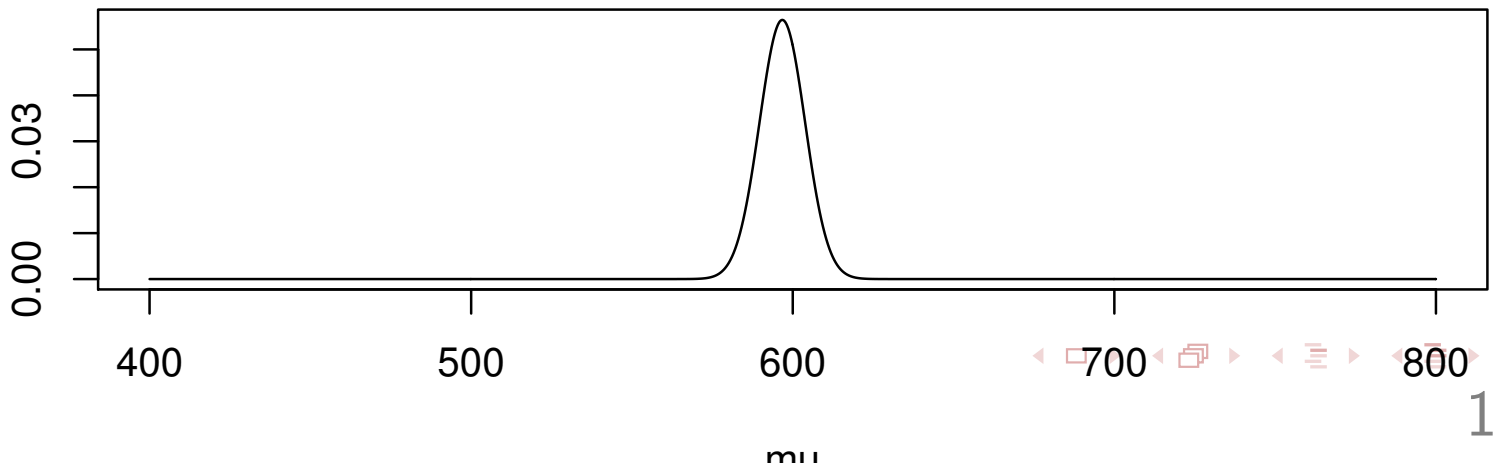
```
## pdf of normal:
dnorm(x, mean = 0, sd = 1)
## compute area under the curve:
pnorm(q, mean = 0, sd = 1)
## find out the quantile that has
## area (probability) p under the curve:
qnorm(p, mean = 0, sd = 1)
## generate normally distributed data of size n:
rnorm(n, mean = 0, sd = 1)
```

Likelihood function (Normal distribution)

Let's assume that we have generated a data point from a particular normal distribution:
$x \sim N(\mu = 600, \sigma^2 = 50)$.

```
x<-rnorm(1,mean=600,sd=sqrt(50))
```

Given x, and different values of μ, we can determine which μ is most likely to have generated x. You can eyeball the result:



Likelihood function

Suppose that we had generated **10** *independent* values of x:

```
x<-rnorm(10,mean=600,sd=sqrt(50))
```

We can plot the likelihood of *each* of the x's that the mean of the Normal distribution that generated the data is μ, for different values of μ:

```
## mu = 500
dnorm(x,mean=500,sd=sqrt(50))

##  [1] 5.760950e-46 3.781055e-52 9.775745e-46 5.929711e-43
##  [6] 2.181487e-58 3.693910e-48 6.075765e-65 1.051859e-42
```

Likelihood function

Since each of the x's are independently generated, the total likelihood of the 10 x's is:

$$f(x_1) \times f(x_2) \times \cdots \times f(x_1 0) \tag{11}$$

for some μ in $f(\cdot) = Normal(\mu, 50)$.

Likelihood function

It's computationally easier to just take logs and sum them up (**log likelihood**):

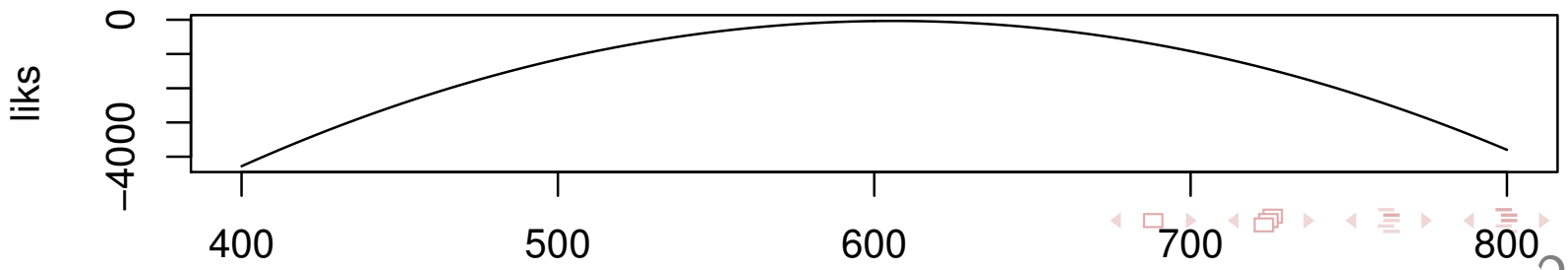
$$\log f(x_1) + \log f(x_2) + \cdots + \log f(x_1 0) \tag{12}$$

```
## mu = 500
sum(dnorm(x,mean=500,sd=sqrt(50),log=TRUE))

## [1] -1157.562
```

(Log) likelihood function

We can now plot, for different values of μ, the likelihood that each of the μ generated the 10 data points:

```
mu<-seq(400,800,by=0.1)
liks<-rep(NA,length(mu))
for(i in 1:length(mu)){
liks[i]<-sum(dnorm(x,mean=mu[i],sd=sqrt(50),log=TRUE))
}
plot(mu,liks,type="l")
```



(Log) likelihood function

1. It's intuitively clear that we'd probably want to declare the value of μ that brings us to the "highest" point in this figure.
2. This is the **maximum likelihood estimate**, MLE.
3. **Practical implication**: In frequentist statistics, our data vector x is assumed to be $X \sim N(\mu, \sigma^2)$, and we attempt to figure out the MLE, i.e., the estimates of μ and σ^2 that would maximize the likelihood.
4. In Bayesian models, when we assume a uniform prior, we will get an estimate of the parameters which coincides with the MLE (examples coming soon).

Bayesian modeling examples

Next, I will work through five relatively simple examples that use Bayes' Theorem.

1. Example 1: Proportions
2. Example 2: Normal distribution
3. Example 3: Linear regression with one predictor
4. Example 4: Linear regression with multiple predictors
5. Example 5: Generalized linear models example (binomial link).

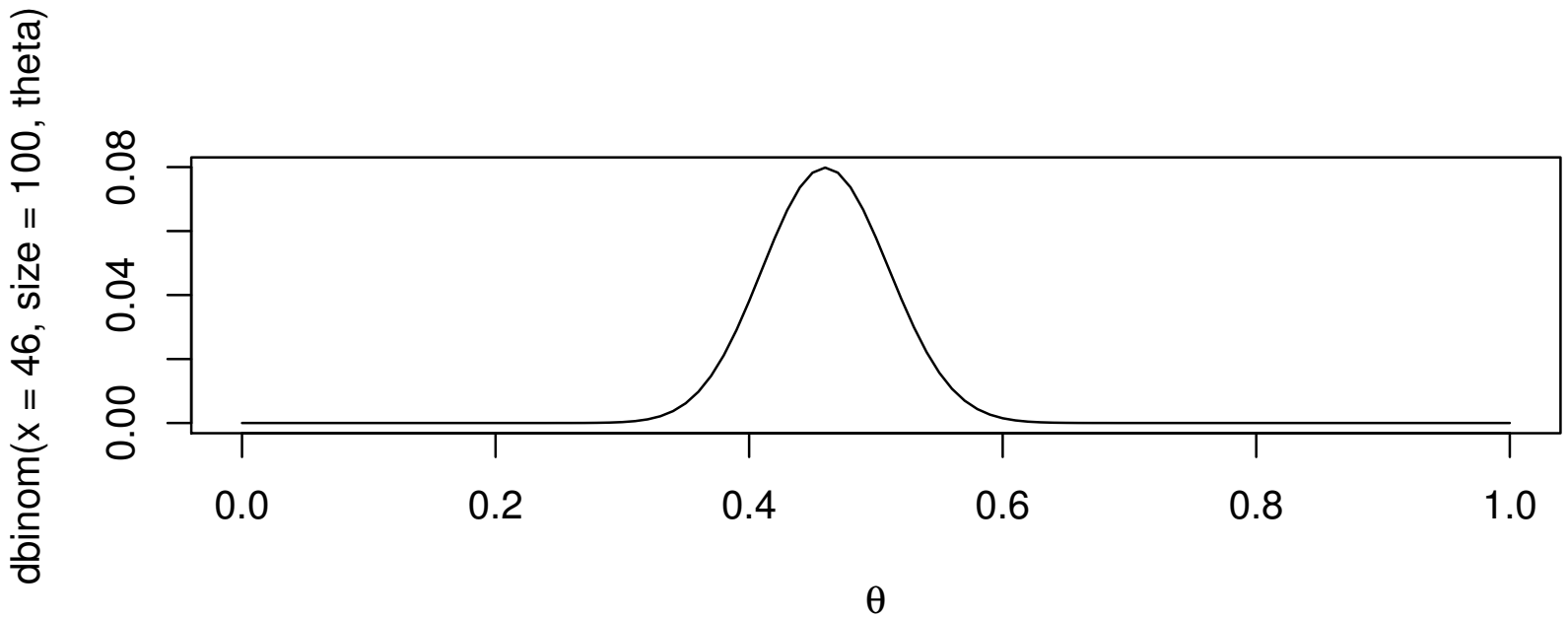
Proportions

1. Recall the binomial distribution: Let X: no. successes in n trials. We generally assume that $X \sim Binomial(n, \theta)$, θ unknown.
2. Suppose we have 46 successes out of 100. We generally use the empirically observed proportion 46/100 as our estimate of θ. I.e., we assume that the generating distribution is $X \sim Binomial(n = 100, \theta = .46)$.
3. This is because, for all possible values of θ, going from 0 to 1, 0.46 has the highest likelihood.

Proportions

```
dbinom(x=46,size=100,0.4)

## [1] 0.03811036

dbinom(x=46,size=100,0.46)

## [1] 0.07984344

dbinom(x=46,size=100,0.5)

## [1] 0.0579584

dbinom(x=46,size=100,0.6)

## [1] 0.001487007
```

Proportions



This is the **likelihood function** for the binomial distribution, and we will write it as $f(data \mid \theta)$. It is a function of θ.

Proportions

Since $Binomial(x, n, \theta) = \binom{n}{x}\theta^x\theta^{n-x}$, we can see that:

$$f(\text{data} \mid \theta) \propto \theta^{46}(1-\theta)^{54} \tag{13}$$

We are now going to use Bayes' theorem to work out the posterior distribution of θ given the data:

$$\underset{\text{posterior}}{f(\theta \mid \text{data})} \propto \underset{\text{likelihood}}{f(\text{data} \mid \theta)} \underset{\text{prior}}{f(\theta)} \tag{14}$$

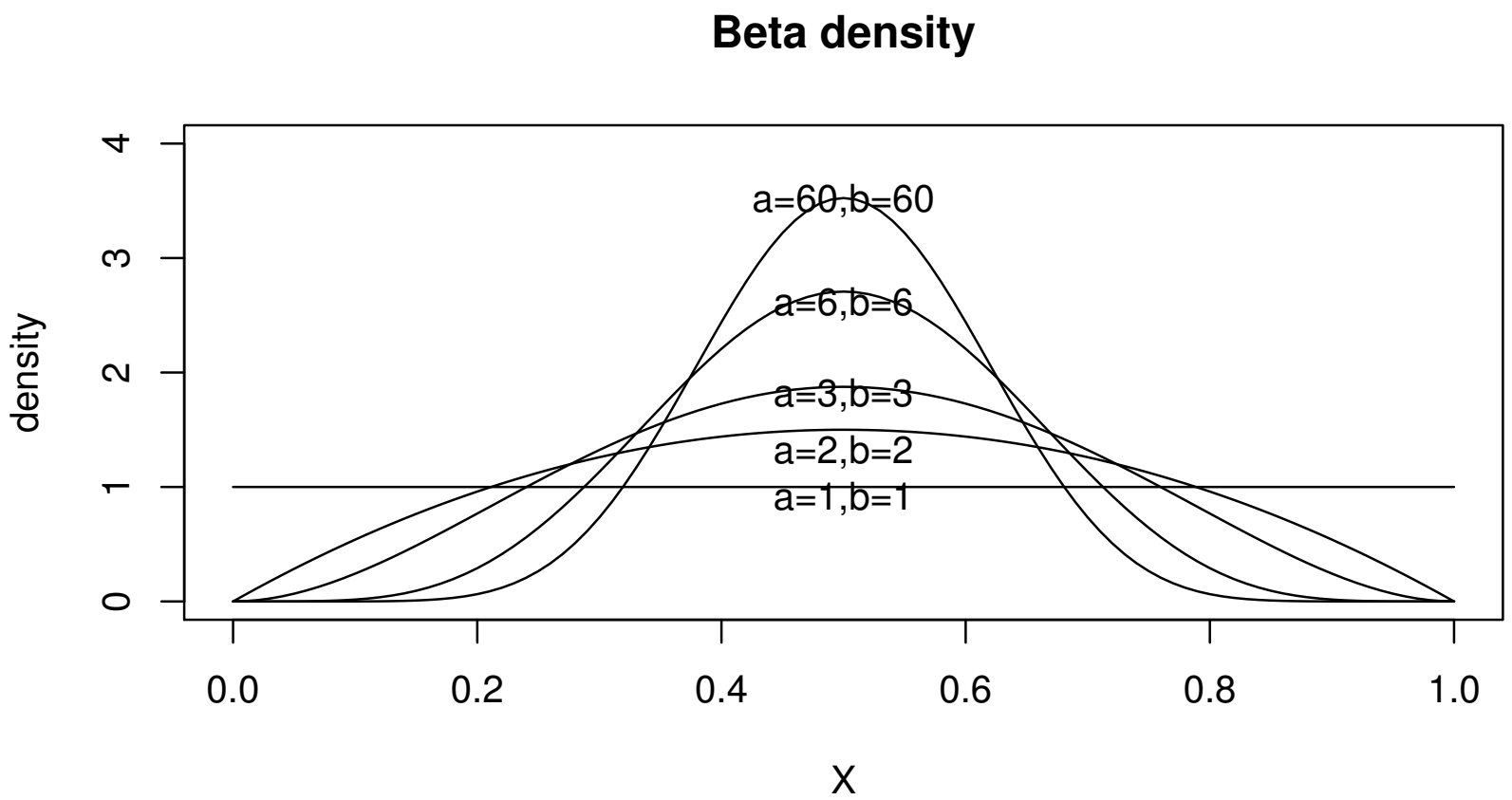
All that's missing here is the prior distribution $f(\theta)$. So let's try to define a prior for θ.

Proportions

To define a prior for θ, we will use a distribution called the Beta distribution, which takes two parameters, a and b.
We can plot some Beta distributions to get a feel for what these parameters do.
We are going to plot

1. Beta(a=1,b=1)
2. Beta(a=2,b=2)
3. Beta(a=3,b=3)
4. Beta(a=6,b=6)
5. Beta(a=10,b=10)

Proportions



Proportions

Each successive density expresses increasing certainty about θ being centered around 0.5; notice that the spread about 0.5 is decreasing as a and b increase.

1. Beta(a=1,b=1)
2. Beta(a=2,b=2)
3. Beta(a=3,b=3)
4. Beta(a=6,b=6)
5. Beta(a=10,b=10)

Proportions

1. If we don't have much prior information, we could use a=b=1; this gives us a uniform prior; we will call this a vague prior.
2. If we have a lot of prior knowledge and/or a strong belief that θ has a particular value, we can use a larger a,b to reflect our greater certainty about the parameter.
3. You can think of the parameter referring to the number of successes, and the parameter b to the number of failures.

So the beta distribution can be used to define the prior distribution of θ.

Proportions

Just for the sake of illustration, let's take four different beta priors, each reflecting increasing certainty.

1. Beta(a=2,b=2)
2. Beta(a=3,b=3)
3. Beta(a=6,b=6)
4. Beta(a=21,b=21)

Each reflects a belief that $\theta = 0.5$, with varying degrees of uncertainty.
Note an important fact: $Beta(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$.

This is because the Beta distribution is:
$f(\theta \mid a, b) = \frac{\Gamma(a,b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$

Proportions

Now we just need to plug in the likelihood and the prior to get the posterior:

$$f(\theta \mid \text{data}) \propto f(\text{data} \mid \theta) f(\theta) \tag{15}$$

The four corresponding posterior distributions would be as follows (I hope I got the sums right!).

Proportions

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{2-1}(1-\theta)^{2-1}] = \theta^{47}(1-\theta)^{55} \quad (16)$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{3-1}(1-\theta)^{3-1}] = \theta^{48}(1-\theta)^{56} \quad (17)$$

$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{6-1}(1-\theta)^{6-1}] = \theta^{51}(1-\theta)^{59} \quad (18)$$

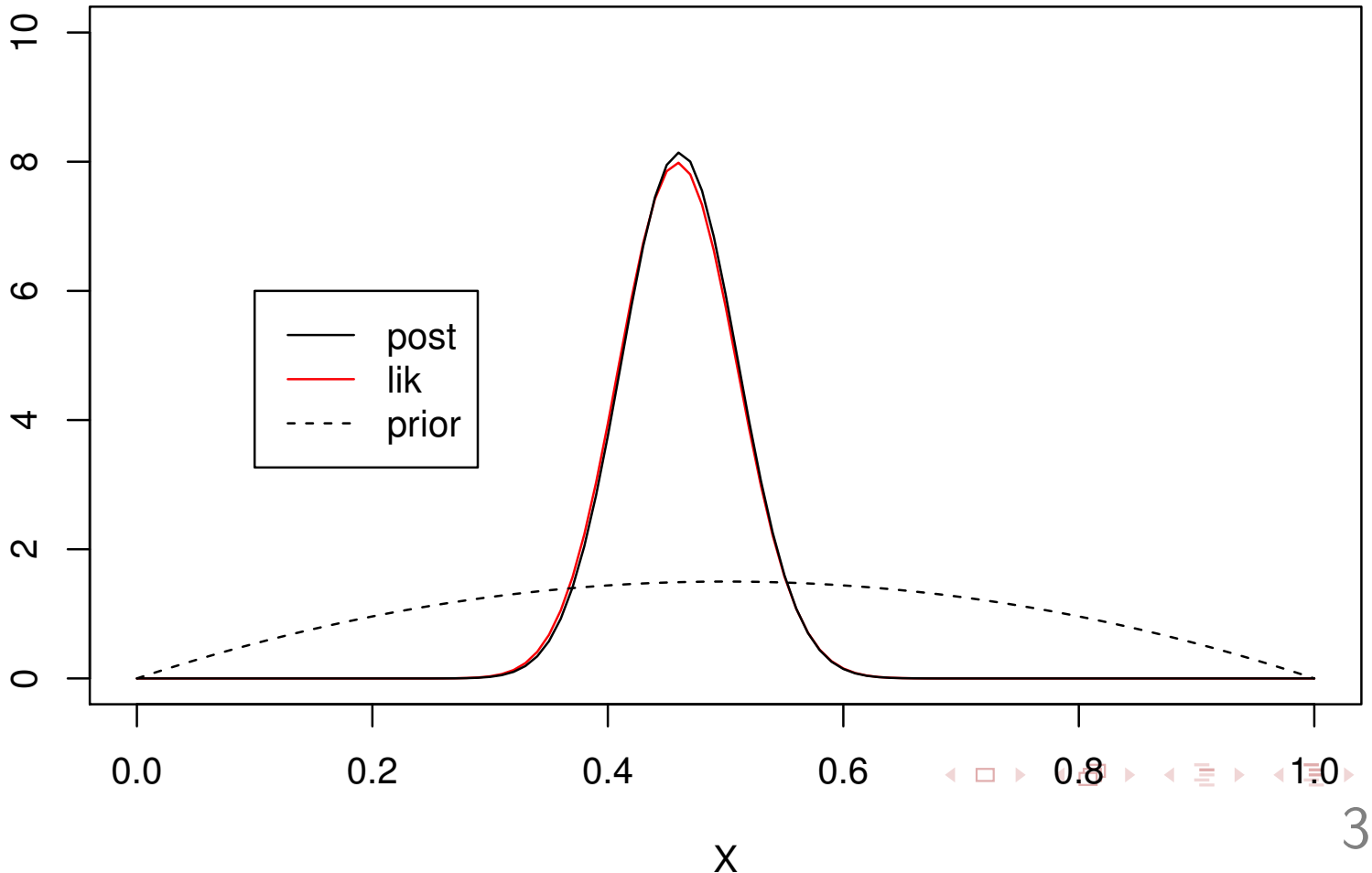
$$f(\theta \mid \text{data}) \propto [\theta^{46}(1-\theta)^{54}][\theta^{21-1}(1-\theta)^{21-1}] = \theta^{66}(1-\theta)^{74} \quad (19)$$

Proportions

1. We can now visualize each of these triplets of priors, likelihoods and posteriors.
2. Note that I use the beta to model the likelihood because this allows me to visualize all three (prior, lik., posterior) in the same plot.
3. I first show the plot just for the prior $\theta \sim Beta(a = 2, a = 2)$

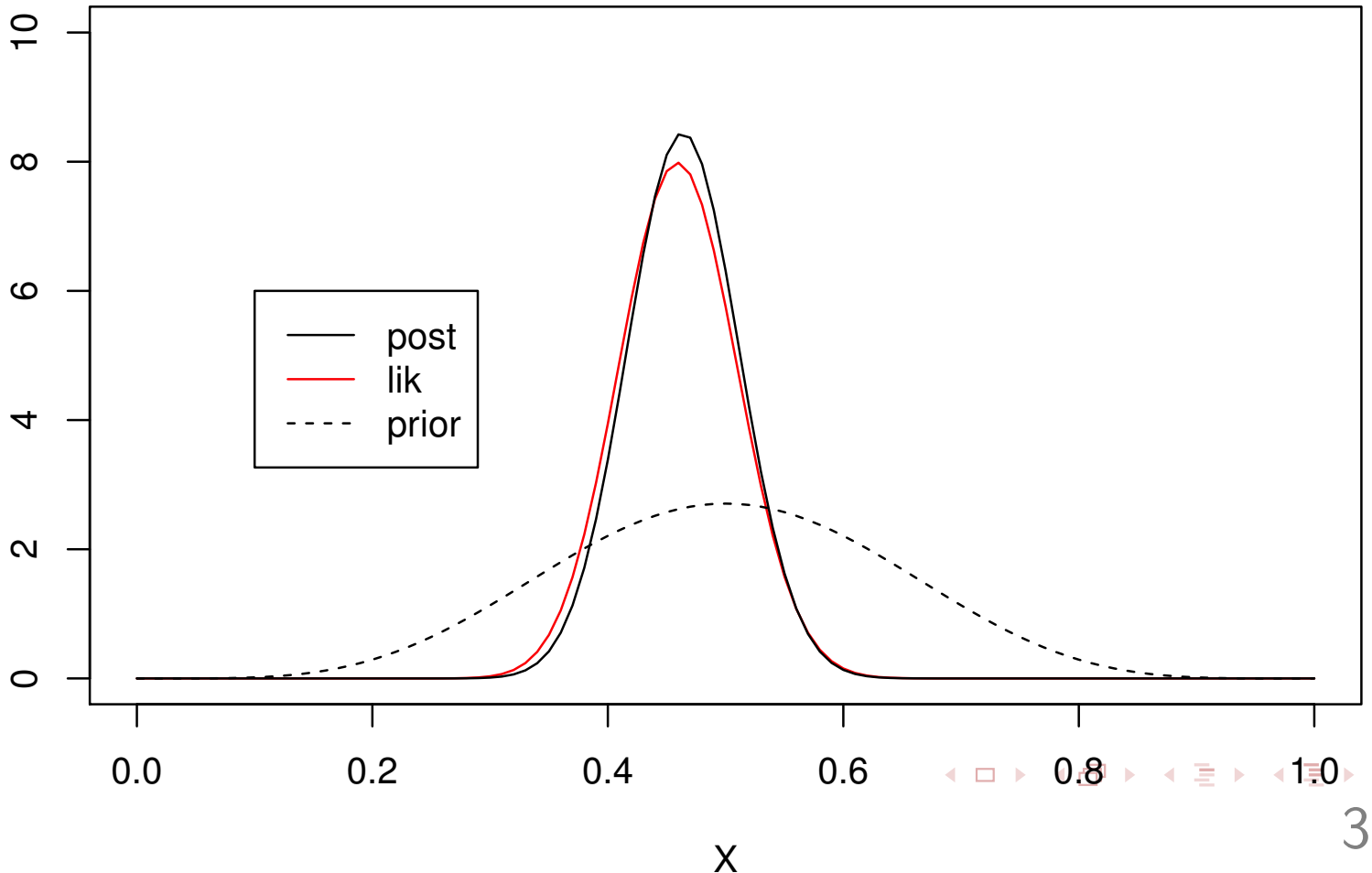
Proportions

Beta(2,2) prior: posterior is shifted just a bit to the right compared to the likelihood



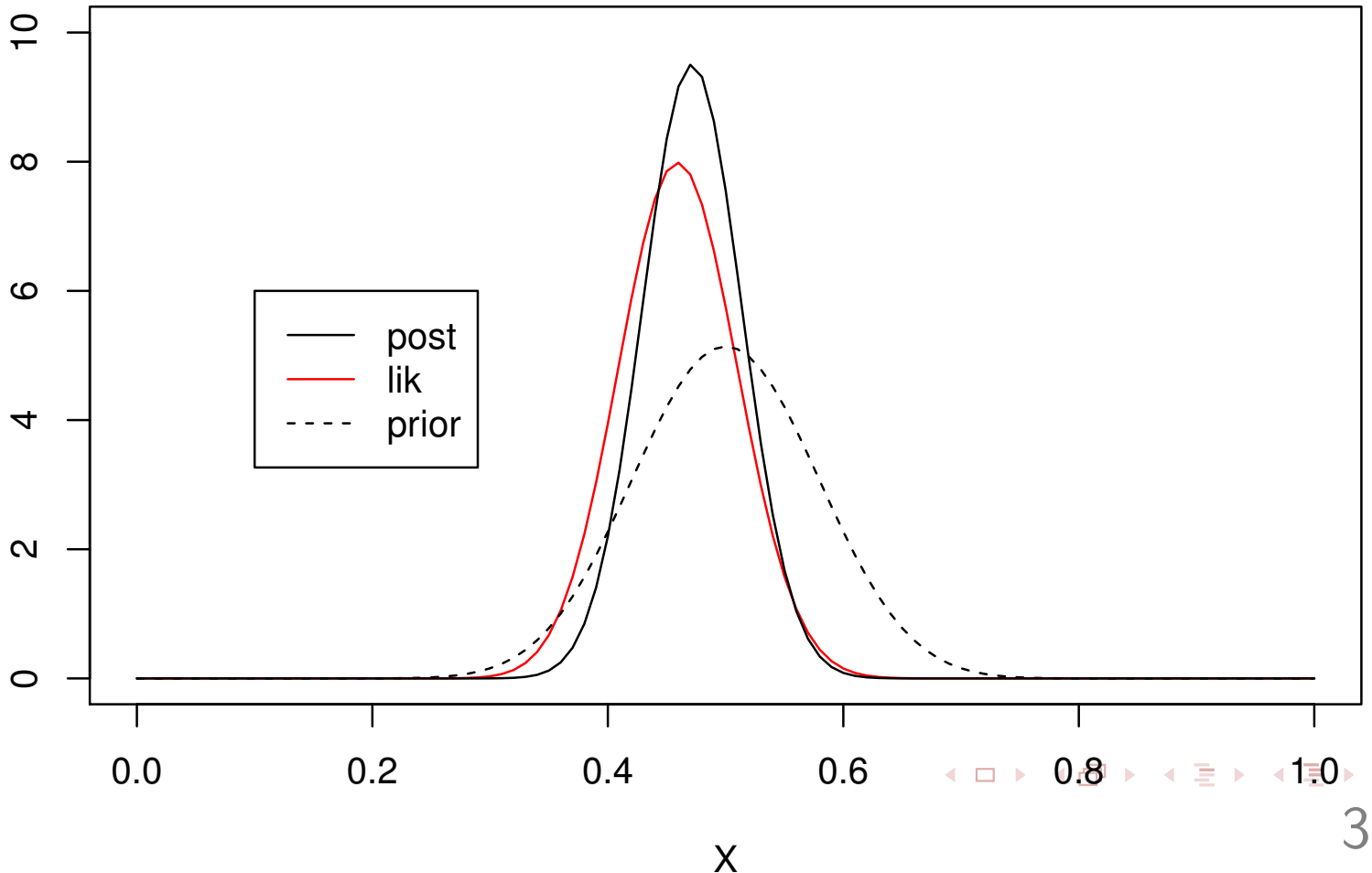
Proportions

Beta(6,6) prior: Posterior shifts even more towards the prior



Proportions

Beta(21,21) prior: Posterior shifts *even* more towards the prior



Proportions

In essence, the posterior is a compromise between the prior and the likelihood.

1. When the prior has high uncertainty or we have a lot of data, the likelihood will dominate.
2. When the prior has high certainty (like the Beta(21,21) case), then the prior will dominate, unless there is enough data for the likelihood to dominate.

So, Bayesian methods are particularly important when you have little data but a whole lot of expert knowledge.
But they are also useful for standard psycholinguistic research, as I hope to demonstrate.

An example application

"The French mathematician Pierre-Simon Laplace (1749-1827) was the first person to show definitively that the proportion of female births in the French population was less then 0.5, in the late 18th century, using a Bayesian analysis based on a uniform prior distribution. Suppose you were doing a similar analysis but you had more definite prior beliefs about the ratio of male to female births. In particular, if θ represents the proportion of female births in a given population, you are willing to place a Beta(100,100) prior distribution on θ.

1. Show that this means you are more than 95% sure that θ is between 0.4 and 0.6, although you are ambivalent as to whether it is greater or less than 0.5.
2. Now you observe that out of a random sample of 1,000 births, 511 are boys. What is your posterior probability that $\theta > 0.5$?"

An example application

Show that this means you are more than 95% sure that θ *is between* 0.4 *and* 0.6*, although you are ambivalent as to whether it is greater or less than* 0.5*.*

Prior: Beta(a=100,b=100)

```
round(qbeta(0.025,shape1=100,shape2=100),digits=1)

## [1] 0.4

round(qbeta(0.975,shape1=100,shape2=100),digits=1)

## [1] 0.6

## ambivalent as to whether theta <0.5 or not:
round(pbeta(0.5,shape1=100,shape2=100),digits=1)

## [1] 0.5
```

An example application

Now you observe that out of a random sample of 1,000 *births,* 511 *are boys. What is your posterior probability that* $\theta > 0.5$*?*

Prior: Beta(a=100,b=100)

Data: 489 girls out of 1000.

Posterior:

$$f(\theta \mid \text{data}) \propto [\theta^{489}(1-\theta)^{511}][\theta^{100-1}(1-\theta)^{100-1}] = \theta^{588}(1-\theta)^{610} \tag{20}$$

Since $Beta(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1}$, the posterior is Beta(a=589,b=611).

Therefore the posterior probability of $\theta > 0.5$ is:

```
qbeta(0.5,shape1=589,shape2=611,lower.tail=FALSE)

## [1] 0.4908282
```

Normal distribution

The normal distribution is the most frequently used probability model in psychology and linguistics.
If $\bar{x}$ is sample mean, sample size is n and sample variance is known to be σ^2, and if the prior on the mean μ is $Normal(m, v)$, then:
It is pretty easy to derive (see Lynch textbook) the posterior mean m^* and variance v^* analytically:

$$v^* = \frac{1}{\frac{1}{v} + \frac{n}{\sigma^2}} \quad m^* = v^* \left(\frac{m}{v} + \frac{n\bar{x}}{\sigma^2} \right) \tag{21}$$

$$E[\theta \mid x] = m \times \frac{w1}{w1 + w2} + \bar{x} \times \frac{w2}{w1 + w2} \quad w_1 = v^{-1}, w_2 = (\sigma^2/n)^{-1} \tag{22}$$

So: the posterior mean is a **weighted mean** of the prior mean and the sample mean.

Normal distribution

1. The weight w1 is determined by the inverse of the prior variance.
2. The weight w2 is determined by the inverse of the sample standard error.
3. It is common in Bayesian statistics to talk about $\textit{precision} = \frac{1}{\textit{variance}}$, so that
   1. $w_1 = v^{-1} = \textit{precision}_{\textit{prior}}$
   2. $w_2 = (\sigma^2/n)^{-1} = \textit{precision}_{\textit{data}}$

Normal distribution

If w1 is very large compared to w2, then the posterior mean will be determined mainly by the prior mean m:

$$E[\theta \mid x] = m \times \frac{\mathbf{w1}}{\mathbf{w1} + w2} + \bar{x} \times \frac{w2}{\mathbf{w1} + w2} \qquad w_1 = v^{-1}, w_2 = (\sigma^2/n)^{-1} \tag{23}$$

If w2 is very large compared to w1, then the posterior mean will be determined mainly by the sample mean $\bar{x}$:

$$E[\theta \mid x] = m \times \frac{w1}{w1 + \mathbf{w2}} + \bar{x} \times \frac{\mathbf{w2}}{w1 + \mathbf{w2}} \qquad w_1 = v^{-1}, w_2 = (\sigma^2/n)^{-1} \tag{24}$$

An example application

Let's say there is a hormone measurement test that yields a numerical value that can be positive or negative. We know the following:

- The doctor's prior: 75% interval ("patient healthy") is [-0.3,0.3].
- Data from patient: $x = 0.2$, known $\sigma = 0.15$.

Compute posterior N(m*,v*).
I'll leave this as Problem 1 (solution is provided with the exercise).
Hint: see slide 16.

Simple linear regression

We begin with a simple example. Let the response variable be $y_i, i = 1, \ldots, n$, and let there be p predictors, $x_{1i}, \ldots, x_{pi}$. Also, let

$$y_i \sim N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \sum \beta x_{ki} \tag{25}$$

(the summation is over the p predictors, i.e., $k = 1, \ldots, p$).
We need to specify a prior distribution for the parameters:

$$\beta_k \sim Normal(0, 100^2) \quad \log \sigma \sim Unif(-100, 100) \tag{26}$$

Better looking professors get better teaching evaluations

Source: Gelman and Hill 2007

```
beautydata<-read.table("data/beauty.txt",header=T)
## Note: beauty level is centered.
head(beautydata)

##       beauty evaluation
## 1  0.2015666        4.3
## 2 -0.8260813        4.5
## 3 -0.6603327        3.7
## 4 -0.7663125        4.3
## 5  1.4214450        4.4
## 6  0.5002196        4.2
```

Better looking professors get better teaching evaluations

```
## restate the data as a list for JAGS:
data<-list(x=beautydata$beauty,
           y=beautydata$evaluation)
```

Better looking professors get better teaching evaluations

JAGS model

We literally follow the specification of the linear model given above. We specify the model for the data frame row by row, using a for loop, so that for each dependent variable value y_i (the evaluation score) we specify how we believe it was generated.

$$y_i \sim Normal(\mu[i], \sigma^2) \quad i = 1, \ldots, 463 \tag{27}$$

$$\mu[i] \leftarrow \beta_0 + \beta_1 x_i \quad \text{Note: predictor is centered} \tag{28}$$

Define priors on the β and on σ:

$$\beta_0 \sim Uniform(-10, 10) \quad \beta_1 \sim Uniform(-10, 10) \tag{29}$$

$$\sigma \sim Uniform(0, 100) \tag{30}$$

Better looking professors get better teaching evaluations

Load rjags library:

```
library(rjags)

## Linked to JAGS 3.4.0
## Loaded modules:  basemod,bugs
```

Better looking professors get better teaching evaluations

```
cat("model{
    ## specify model for data:
    for(i in 1:463){
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta0 + beta1 * (x[i])
    }
    # priors:
    beta0 ~ dunif(-10,10)
    beta1 ~ dunif(-10,10)
    sigma ~ dunif(0,100)
    sigma2 <- pow(sigma,2)
    tau <- 1/sigma2
   }",
     file="JAGSmodels/beautyexample1.jag" )
```

Better looking professors get better teaching evaluations

Data from Gelman and Hill, 2007

Some things to note in JAGS syntax:

1. The normal distribution is defined in terms of precision, not variance.
2. $\sim$ can be read as "is generated by", or "is modeled by"
3. <-
   is a deterministic assignment, like $=$ in mathematics.
4. The model specification is declarative, order does not matter. For example, we would have written the following in any order:

```
sigma ~ dunif(0,100)
sigma2 <- pow(sigma,2)
tau <- 1/sigma2
```

Better looking professors get better teaching evaluations

```
## specify which variables you want to examine
## the posterior distribution of:
track.variables<-c("beta0","beta1","sigma")

## define model:
beauty.mod <- jags.model(
  file = "JAGSmodels/beautyexample1.jag",
                         data=data,
                         n.chains = 2,
                         n.adapt =2000,
                         quiet=T)

## sample from posterior:
beauty.res <- coda.samples(beauty.mod,
                           var = track.variables,
                           n.iter = 2000,
                           thin = 1 )
```

Better looking professors get better teaching evaluations

```
round(summary(beauty.res)$statistics[,1:2],digits=2)

##        Mean   SD
## beta0 4.01 0.02
## beta1 0.13 0.03
## sigma 0.55 0.02

round(summary(beauty.res)$quantiles[,c(1,3,5)],digits=2)

##        2.5%  50% 97.5%
## beta0 3.96 4.01  4.06
## beta1 0.07 0.13  0.19
## sigma 0.51 0.55  0.58
```

Better looking professors get better teaching evaluations

Compare with standard `lm` fit:

```
lm_summary<-summary(lm(evaluation~beauty,
                       beautydata))

round(lm_summary$coef,digits=2)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.01       0.03  157.21        0
## beauty          0.13       0.03    4.13        0

round(lm_summary$sigma,digits=2)

## [1] 0.55
```

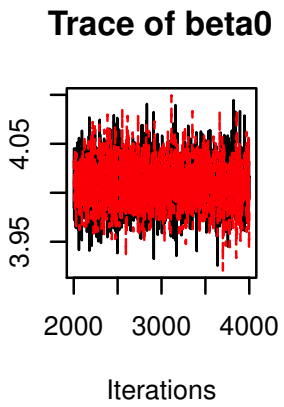
Better looking professors get better teaching evaluations

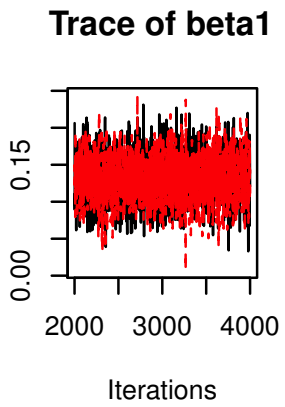
Note that: (a) with uniform priors, we get a Bayesian estimate equal to the MLE, (b) we get uncertainty estimates for σ in the Bayesian model.

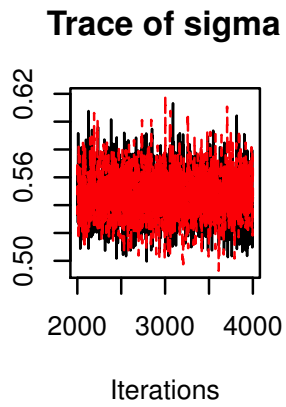
Better looking professors get better teaching evaluations

Posterior distributions of parameters

```
op<-par(mfrow=c(1,3),pty="s")
library(coda)
traceplot(beauty.res)
```

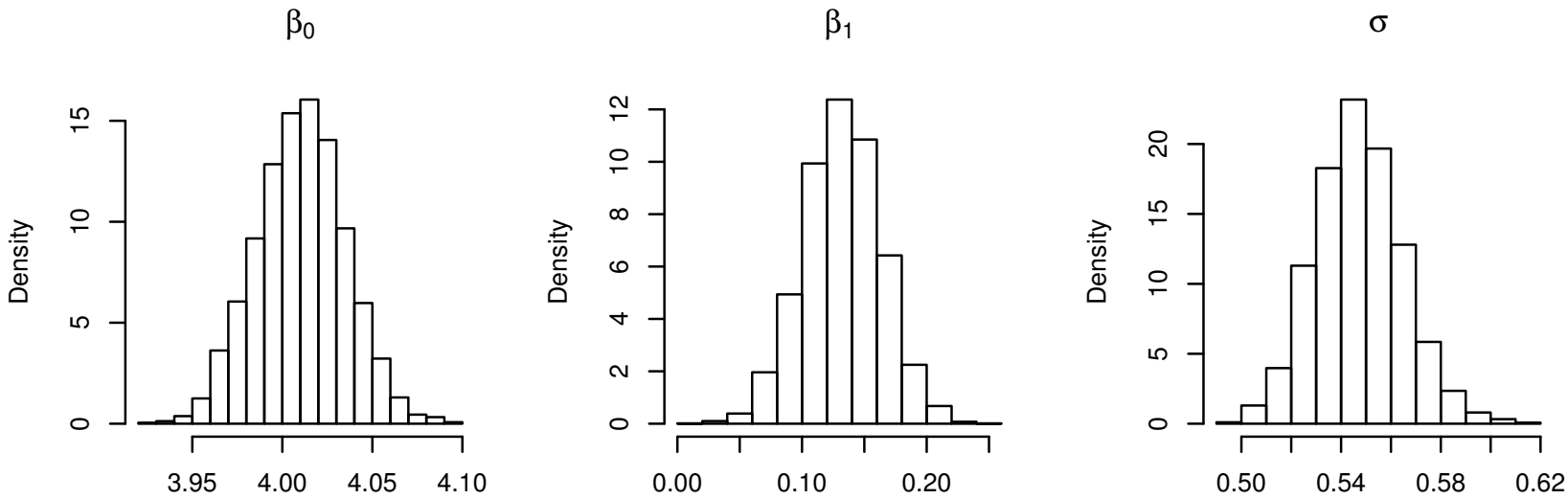






Better looking professors get better teaching evaluations

Posterior distributions of parameters



Problem 2: Rats' weights

Source: Lunn et al 2012

Five measurements of a rat's weight, in grams, as a function of some x (say some nutrition-based variable). Note that here we will center the predictor in the model code.
First we load/enter the data:

```
data<-list(x=c(8,15,22,29,36),
           y=c(177,236,285,350,376))
```

Problem 2: Rats' weights

Source: Lunn et al 2012

Then we fit the linear model using `lm`, for comparison with the Bayesian model:

```
lm_summary_rats<-summary(fm<-lm(y~x,data))
round(lm_summary_rats$coef,digits=3)

##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  123.886     11.879  10.429    0.002
## x              7.314      0.492  14.855    0.001
```

Fit this linear model using JAGS.

Problem 3: Rats' weights (solution in next class)

Fit the following model in JAGS:

```
cat("
model
   {
    ## specify model for data:
    for(i in 1:5){
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta0 + beta1 * (x[i]-mean(x[]))
    }
    # priors:
    beta0 ~ dunif(-500,500)
    beta1 ~ dunif(-500,500)
    tau <- 1/sigma2
    sigma2 <-pow(sigma,2)
    sigma ~ dunif(0,200)
   }",
     file="JAGSmodels/ratsexample2.jag" )
```

Review of the Likelihood Function

Discrete case: Suppose the observed sample values (binomially distributed) are $x_1, x_2, \ldots, x_n$. The joint probability of getting them is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta) \tag{31}$$

i.e., the function f is the value of the joint probability **distribution** of the random variables $X_1, \ldots, X_n$ at $X_1 = x_1, \ldots, X_n = x_n$. Since the sample values have been observed and are fixed, $f(x_1, \ldots, x_n; \theta)$ is a function of θ. The function f is called a **likelihood function**.

Review of the Likelihood Function

Continuous case

Here, f is the joint probability **density**, the rest is the same as above.

If $x_1, x_2, \ldots, x_n$ are the values of a random sample from a population with parameter θ, the **likelihood function** of the sample is given by

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta) \tag{32}$$

for values of θ within a given domain. Here, $f(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta)$ is the joint probability distribution or density of the random variables $X_1, \ldots, X_n$ at $X_1 = x_1, \ldots, X_n = x_n$.

Review of the Likelihood Function

So, the method of maximum likelihood consists of maximizing the likelihood function with respect to θ. The value of θ that maximizes the likelihood function is the **MLE** (maximum likelihood estimate) of θ.

Finding the MLE by hand

The likelihood function in the binomial case:

$$L(\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} \tag{33}$$

Log likelihood:

$$\ell(\theta) = \log \binom{n}{x} + x \log \theta + (n-x) \log(1-\theta) \tag{34}$$

Finding the MLE by hand

Differentiating and equating to zero to get the maximum:

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \tag{35}$$

How to get the second term: let $u = 1-\theta$. Then, $du/d\theta = -1$.
Now, $y = (n-x)\log(1-\theta)$ can be rewritten in terms of u:
$y = (n-x)\log(u)$. So, $dy/du = \frac{n-x}{u}$. Now
$dy/d\theta = dy/du \times du/d\theta = \frac{n-x}{u} \times (-1) = -\frac{n-x}{1-\theta}$.

Finding the MLE by hand

Rearranging terms we get:

$$\frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \Leftrightarrow \frac{x}{\theta} = \frac{n-x}{1-\theta} \Leftrightarrow \hat{\theta} = \frac{x}{n} \tag{36}$$

The Likelihood Function

The point of the MLE method is to find out the most likely value(s) of the parameter(s) that generated the data (assuming some probability model for the data).

Multiple predictors

Source: Baayen's book

We fit log reading time to Trial id (centered), Native Language, and Sex. The categorical variables are centered as well. Note: This model is incorrect! We should have fit a linear mixed model, but we are going to fit a multiple regression using `lm` anyway.

```
lexdec<-read.table("data/lexdec.txt",header=TRUE)
data<-lexdec[,c(1,2,3,4,5)]

contrasts(data$NativeLanguage)<-contr.sum(2)
contrasts(data$Sex)<-contr.sum(2)
```

Multiple predictors

Source: Baayen's book

```
lm_summary_lexdec<-summary(fm<-lm(RT~
  scale(Trial,scale=F)+
  NativeLanguage+Sex,data))

round(lm_summary_lexdec$coef[,1:2],digits=2)

##                          Estimate Std. Error
## (Intercept)                  6.41       0.01
## scale(Trial, scale = F)      0.00       0.00
## NativeLanguage1             -0.08       0.01
## Sex1                        -0.03       0.01
```

Multiple predictors

Preparing the data for JAGS:

```
contrasts(data$NativeLanguage)

##         [,1]
## English    1
## Other     -1

contrasts(data$Sex)

##   [,1]
## F    1
## M   -1
```

Multiple predictors

```
## redo contrasts as vectors:
eng<-ifelse(data$NativeLanguage=="English",1,-1)
sex<-ifelse(data$Sex=="F",1,-1)
```

Multiple predictors

```
dat<-list(y=data$RT,
          Trial=(data$Trial-mean(data$Trial)),
          Lang=eng,
          Sex=sex)
```

Multiple predictors

The JAGS model:

```
cat("
model
   {
    ## specify model for data:
    for(i in 1:1659){
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta0 +
             beta1 * Trial[i]+
             beta2 * Lang[i] +  beta3 * Sex[i]
    }
    # priors:
    beta0 ~ dunif(-10,10)
    beta1 ~ dunif(-5,5)
    beta2 ~ dunif(-5,5)
    beta3 ~ dunif(-5,5)
    tau <- 1/sigma2
    sigma2 <-pow(sigma,2)
    sigma ~ dunif(0,200)
   }",
     file="JAGSmodels/multregexample1.jag" )
```

Multiple predictors

```
track.variables<-c("beta0","beta1",
                   "beta2","beta3","sigma")
library(rjags)

lexdec.mod <- jags.model(
  file = "JAGSmodels/multregexample1.jag",
                    data=dat,
                    n.chains = 2,
                    n.adapt =2000,
                     quiet=T)

lexdec.res <- coda.samples( lexdec.mod,
                                 var = track.variables,
                               n.iter = 3000)
```

Multiple predictors

```
round(summary(lexdec.res)$statistics[,1:2],
      digits=2)

##         Mean   SD
## beta0   6.41 0.01
## beta1   0.00 0.00
## beta2  -0.08 0.01
## beta3  -0.03 0.01
## sigma   0.23 0.00
```

Multiple predictors

```
round(summary(lexdec.res)$quantiles[,c(1,3,5)],
      digits=2)
```

```
##         2.5%   50% 97.5%
## beta0   6.40  6.41  6.42
## beta1   0.00  0.00  0.00
## beta2  -0.09 -0.08 -0.07
## beta3  -0.04 -0.03 -0.02
## sigma   0.22  0.23  0.23
```

Multiple predictors

As an exercise, compare the above model's results with the output of the lm function.
Note: We should have fit a linear mixed model here; I will return to this later.

GLMs

We have considered linear models like

$$E[Y_i] = \mu_i = x_i^T \beta \quad y_i \sim N(\mu_i, \sigma^2) \tag{37}$$

GLMs allow us to stay within the linear modeling framework, even if the relationship betwen response and explanatory variable is not linear.

GLMs

There is a wider class of distributions beyond the two we have seen (normal, binomial), that are called the **exponential family of distributions**; the normal and binomial fall within this family. The likelihood function of the exponential family's distributions can be written in very general terms as follows:

$$f(y;\theta_i,\phi) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi/w} + c(y,\phi)\right] \tag{38}$$

GLMs

Consider the normal distribution. We can write it in the general form of equation 38.

$$
\begin{aligned}
f(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{(y-\mu)}{\sigma}\right)^2\right] \\
&= \exp\left[\log 1 - \log\sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{(y-\mu)}{\sigma}\right)^2\right] \qquad (39) \\
&= \exp\left[-\frac{1}{2}\left(\frac{y^2+\mu^2-2y\mu}{\sigma^2}\right) - \log\sigma\sqrt{2\pi}\right]
\end{aligned}
$$

GLMs

A little bit of algebraic manipulation (exercise) will now give us:

$$
\begin{aligned}
&= \exp\left[\frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{y^2}{2\sigma^2} + \frac{\log\sigma\sqrt{2\pi}}{2}\right] \\
&= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} + c(y,\phi)\right] \qquad \text{i.e., } c(y,\phi) = -\frac{y^2}{2\sigma^2} + \frac{\log\sigma\sqrt{2\pi}}{2} \\
&= \exp\left[\frac{y\theta - b(\theta)}{\phi/w} + c(y,\phi)\right]
\end{aligned}
\tag{40}
$$

Here, $\theta = \mu$, $\phi = \sigma^2$, $w = 1$, and we have $b(\theta) = \mu^2/2$, $c(y,\phi) = -\frac{y^2}{2\sigma^2} + \frac{\log\sigma\sqrt{2\pi}}{2}$.

GLMs

This general formulation gives us two useful results:

1. The first derivative of $b(\theta) = \frac{\mu^2}{2}$, is $b'(\theta) = \mu$. This is a general result for the exponential family: $E[y] = b'(\theta) = \mu$
2. The variance of Y is $Var(Y) = \frac{\phi}{w} b''(\theta)$. So, here, we'd get $Var(Y) = \frac{\sigma^2}{1} 1 = \sigma^2$

GLMs

- Let's look at another example of how we can write an exponential family distribution in this general form.
- Consider the binomial distribution, which we will start by writing as below.
- Here, n is the total number of trials, and y is the proportion of successes.

GLMs

For example, n=10, y=7/10, gives us 7 successes out of 10. This is just another way to parameterize the binomial distribution, although it is not one that you have seen before.

$$ny \sim Binomial\left(n, \frac{\exp(\theta)}{1+\exp(\theta)}\right) \quad \text{i.e., } p = \frac{\exp(\theta)}{1+\exp(\theta)} \tag{41}$$

GLMs

$$
\begin{aligned}
f(ny;\theta,\phi) &= \binom{n}{ny} p^{ny}(1-p)^{n-ny} \\
&= \exp\left[\log\binom{n}{ny} + ny\log p + (n-ny)\log(1-p)\right] \quad (42) \\
&= \exp\left[ny\log\frac{p}{1-p} + n\log(1-p) + c(y,\phi)\right]
\end{aligned}
$$

[i.e., $c(y,\phi) = \log\binom{n}{ny}$]

GLMs

Since $p = \frac{\exp(\theta)}{1+\exp(\theta)}$, we can write

$$n\log(1-p) = n\log\frac{1}{1+\exp(\theta)} = -n\log(1+\exp(\theta)) \tag{43}$$

Also, let $\theta = \log\frac{p}{1-p}$.
Then, we can continue as follows:

$$\begin{aligned} f(ny;\theta,\phi) &= \exp\left[ny\log\frac{p}{1-p} + n\log(1-p) + c(y,\phi)\right] \\ &= \exp\left[ny\theta - n\log(1+\exp(\theta)) + c(y,\phi)\right] \\ &= \exp\left[\frac{y\theta - b(\theta)}{\phi/n} + c(y,\phi)\right] \qquad \text{i.e., } b(\theta) = n\log(1+\exp(\theta)) \end{aligned} \tag{44}$$

GLMs

For each data point Y_i from a distribution that's a member of the exponential family, the general form of the likelihood function is:

$$f(y;\theta_i,\phi) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi/w} + c(y,\phi)\right] \tag{45}$$

where $E[Y_i] = \mu_i = h(x_i^T\beta)$. Since we know that $E[Y_i] = b'(\theta_i)$, we can write

$$E(Y_i) = \mu_i = h(x_i^T\beta) = b'(\theta) \tag{46}$$

GLMs

Now, if we want to get $x_i^T \beta$, we just take the inverse of the function $h(\cdot)$, call it $g(\cdot)$. This gives us something called the canonical link function:

$$x_i^T \beta = h^{-1}(b'(\theta)) = \underset{\text{canonical link}}{g} b'(\theta) \tag{47}$$

GLMs

For different distributions in the exponential family, the canonical link functions are as follows:

Distribution	$h(x_i^T \beta) = \mu_i$	$g(\mu_i) = \theta_i$
Binomial logit link	$\frac{\exp[\theta_i]}{1+\exp[\theta_i]}$	$\log \frac{y}{1-y}$
Normal identity	θ	$g = h$
Poisson log	$\exp[\theta]$	$\log[\mu]$
Gamma inverse	$-\frac{1}{\theta}$	$-\frac{1}{\mu_i}$
Cloglog cloglog	$1 - \exp[-\exp[\theta_i]]$	$\log(-\log(1-\mu_i))$
Probit probit	$\Phi(\theta)$	$\Phi^{-1}(\theta)$ (qnorm)

GLMs

The big thing about the canonical link is that is expresses θ_i as a linear combination of the parameters: $x_i^T \beta$. You can decide which link to use by plotting $g(\mu_i)$ against the predictor (in case we have only a single predictor x).

GLMs

We consider the model

$$y_i \sim Binomial(p_i, n_i) \quad logit(p_i) = \beta_0 + \beta_1(x_i - \bar{x}) \qquad (48)$$

GLMs

A simple example is the beetle data from Dobson et al 2010:

```
beetledata<-read.table("data/beetle.txt",header=T)
head(beetledata)

##      dose number killed
## 1 1.6907     59      6
## 2 1.7242     60     13
## 3 1.7552     62     18
## 4 1.7842     56     28
## 5 1.8113     63     52
## 6 1.8369     59     53
```

GLMs

Prepare data for JAGS:

```
dat<-list(x=beetledata$dose-mean(beetledata$dose),
          n=beetledata$number,
          y=beetledata$killed)
```

GLMs

```
cat("
model
   {
for(i in 1:8){
    y[i] ~ dbin(p[i],n[i])
    logit(p[i]) <- beta0 + beta1 * x[i]
}
    # priors:
     beta0 ~ dunif(0,100)
     beta1 ~ dunif(0,100)
   }",
     file="JAGSmodels/glmexample1.jag" )
```

GLMs

Notice use of initial values:

```
track.variables<-c("beta0","beta1")
## new:
inits <- list (list(beta0=0,
                    beta1=0))

glm.mod <- jags.model(
  file = "JAGSmodels/glmexample1.jag",
                     data=dat,
                    ## new:
                    inits=inits,
                     n.chains = 1,
                      n.adapt =2000, quiet=T)
```

GLMs

```
glm.res <- coda.samples( glm.mod,
                         var = track.variables,
                             n.iter = 2000)
```

GLMs

```
round(summary(glm.res)$statistics[,1:2],
      digits=2)

##         Mean   SD
## beta0   0.74 0.14
## beta1  34.44 2.88

round(summary(glm.res)$quantiles[,c(1,3,5)],
      digits=2)

##         2.5%   50% 97.5%
## beta0   0.47  0.74  1.02
## beta1  28.98 34.38 39.96
```

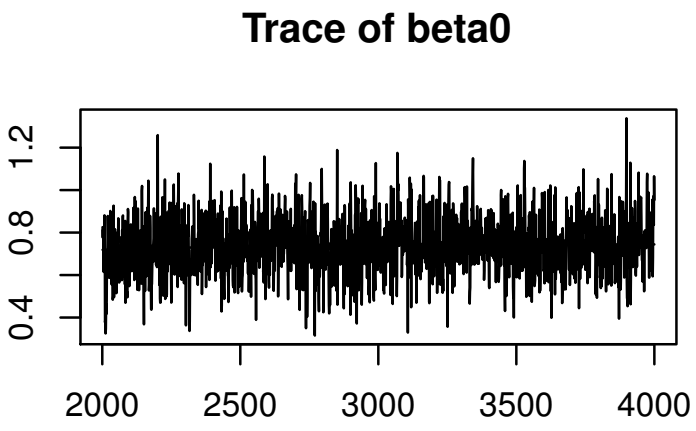
GLMs

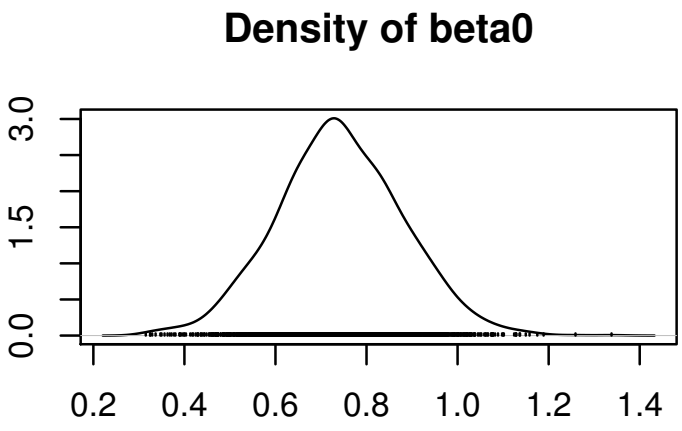
The values match up with glm output:

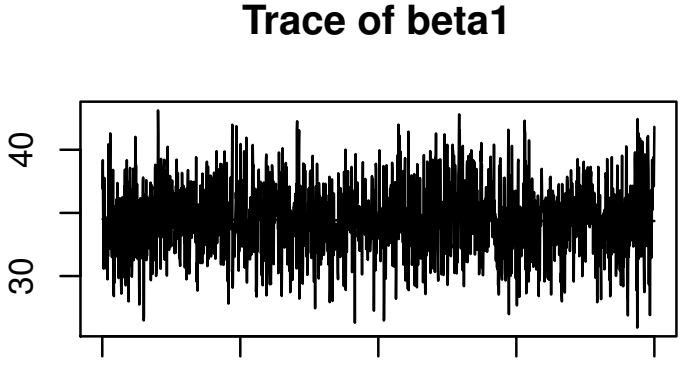
```
round(coef(glm(killed/number~scale(dose,scale=F),
          weights=number,
            family=binomial(),beetledata)),
       digits=2)

##                (Intercept) scale(dose, scale = F)
##                       0.74                  34.27
```

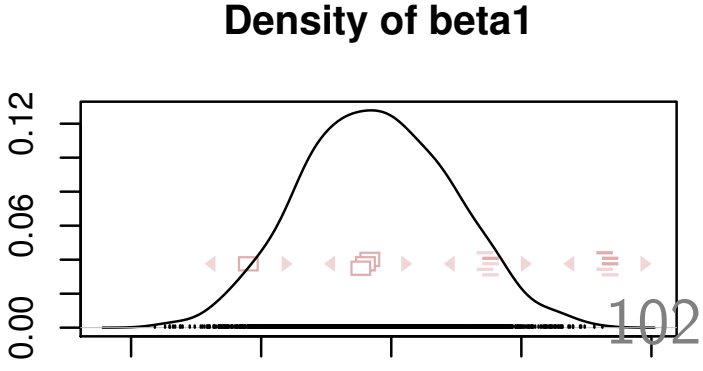
GLMs

```
plot(glm.res)
```









Homework: GLMs

We fit uniform priors to the coefficients β:

```
# priors:
beta0 ~ dunif(0,100)
beta1 ~ dunif(0,100)
```

Fit the beetle data again, using some suitable normal distribution priors for the coefficients beta0 and beta1. Does the posterior distribution depend on the prior?

GLMs: Predicting future/missing data

One important thing we can do is to predict the posterior distribution of future or missing data.
One easy to way to do this is to define how we expect the predicted data to be generated.
This example revisits the earlier toy example from Lunn et al. on rat data (slide 61).

```
data<-list(x=c(8,15,22,29,36),
           y=c(177,236,285,350,376))
```

GLMs: Predicting future/missing data

```
cat("model{
    ## specify model for data:
    for(i in 1:5){
    y[i] ~ dnorm(mu[i],tau)
    mu[i] <- beta0 + beta1 * (x[i]-mean(x[]))
    }
    ## prediction
    mu45 <- beta0+beta1 * (45-mean(x[]))
    y45 ~ dnorm(mu45,tau)
    # priors:
    beta0 ~ dunif(-500,500)
    beta1 ~ dunif(-500,500)
    tau <- 1/sigma2
    sigma2 <-pow(sigma,2)
    sigma ~ dunif(0,200)
   }",
     file="JAGSmodels/ratsexample2pred.jag" )
```

GLMs: Predicting future/missing data

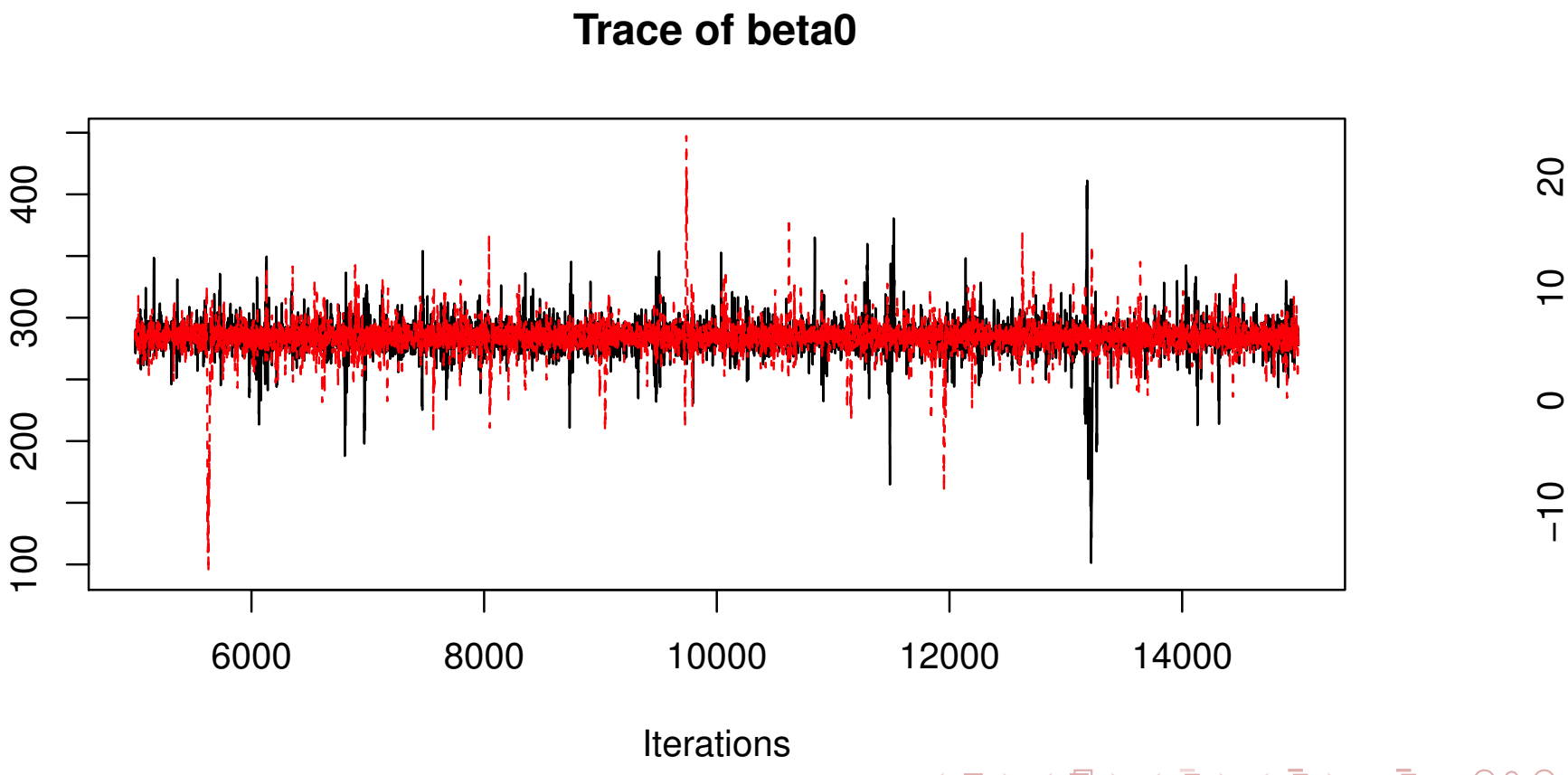
```
track.variables<-c("beta0","beta1","sigma","y45")

rats.mod <- jags.model(
  file = "JAGSmodels/ratsexample2pred.jag",
                        data=data,
                        n.chains = 2,
                         n.adapt =5000, quiet=T)

rats.res <- coda.samples( rats.mod,
                             var = track.variables,
                                 n.iter = 10000,
                                   thin = 1)
```

GLMs: Predicting future/missing data

```
round(summary(rats.res)$statistics[,1:2],
      digits=2)

##          Mean    SD
## beta0  284.54 14.04
## beta1    7.30  1.40
## sigma   22.47 20.36
## y45    452.38 46.01
```

GLMs: Predicting future/missing data

```
traceplot(rats.res)
```



Summing up

1. In some cases Bayes' Theorem can be used analytically (Examples 1, 2)
2. It is relatively easy to define different kinds of Bayesian models using programming languages like JAGS.
3. We saw some examples from linear models (Examples 3-5).
4. Coming up next: MCMC sampling and then Linear Mixed Models.