Introduction to Bayesian inference

We have data, y, and we want to make inferences (make statements of knowledge about) some unknown, call it θ.

$$y \rightarrow \theta$$

Given a model, we get $p(y|\theta)$, the likelihood (conditional probability of observing y given θ. We want get $p(\theta|y)$,

A probability distribution is a representation of uncertain knowledge. We want to learn about θ, not y (we also want to learn about y, but not conditional on a particular value of θ, since θ is unknown).

- We treat ***any*** unknown quantity in the same way as a parameter.

"One of the appeals of the Bayesian approach is that *all* unknowns are treated the same." [Rossi, Allenby &McCulloch (2005)]

We want to convert the likelihood, $p(y|\theta)$, into a probability for the unknown θ.

Bayes rule is the equation that gets us from $p(y|\theta)$ to $p(\theta|y)$. It can also be viewed as a learning engine.

Bayes rule: $p(\theta|y) \propto p(y|\theta)\, p(\theta) = c\, p(y|\theta)\, p(\theta)$

where $c = 1/\int p(y|\theta)\, p(\theta) d\theta$

c is the normalizing constant (to ensure $p(\theta|y)$ integrates to one, i.e. the probabilities 'sum' to one),

$p(\theta|y)$ is the posterior distribution for θ.

Notice in Bayes rule that, to get from $p(y|\theta)$ to $p(\theta|y)$, we need $p(\theta)$.

$p(\theta)$ is the prior distribution.

Prior probabilities do not have to be exact – we can specify intervals, etc.

For example, suppose we must choose a prior probability for some event occurring with two possibilities (e.g., unfair coin, male or female person entering store next, politician will/will not get elected, smoker/nonsmoker, gender of next baby born, rain or no rain tomorrow, pass or not pass a test, …). A lot of events we wish to make inferences and predictions about are of this form (two possible outcomes).

Instead of picking one number this probability, which may be very difficult (e.g. prob. of rain today, $p(\mathrm{R}) = 0.4$? 0.5?), we can use a **hierarchical** prior. We treat this probability as another unknown parameter, i.e. we let $p(\mathrm{R}) = \theta$ (so prob. of no rain is $p(\mathrm{NR})=1-p(\mathrm{R}) = 1 - \theta$).

Then we specify an entire prob. distribution for θ.

Some R code to plot some different priors (scaled to fit on one graph). See Bolstad ch.8 for plots of Beta distribution: (see priorforbinomial.R for a more detailed version)

```
sig21 = 10
sig22 = 1000
R = 1001
uniform = rep(1,R)
theta = rep(0,R)
normal1 = rep(0,R)
normal2 = rep(0,R)
for (i in 1:R)
{
theta[i] = i*0.001 - 0.001
normal1[i] = (1/sqrt(2*3.142*sig21))*exp(-((theta[i]-
0.5)^2)/(2*sig21))
normal2[i] = (1/sqrt(2*3.142*sig22))*exp(-((theta[i]-
0.5)^2)/(2*sig22))
}
# scale normal
normal1 = normal1/(0.125)
plot(theta,normal1,type='l')

normal2 = normal2/(0.013*0.97)
plot(theta,normal2,type='l')

yy = cbind(uniform,normal1,normal2)
matplot(theta,yy,type='l', col=1;3)
```

Binomial likelihood (applies when there are only two possible outcomes, "success/failure")

$$p(\theta|y) = \frac{n!}{y!(n-y)!}\theta^y(1-\theta)^{n-y}, \quad y = 0,1,\dots,n, \quad 0 \le \theta \le 1$$

Uniform prior:

$$p(\theta) = c \quad (= 1 \text{ with a binomial likelihood})$$

with *c* a constant. For the binomial, *c* = 1 since $0 \le \theta \le 1$ and $p(\theta)$ must integrate to one.

Beta prior:

$$p(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

mean(θ) = $E(\theta) = {}^{a}/_{a+b}$, variance($\theta$) = $\frac{ab}{(a+b)^2(a+b+1)}$

Note that any prior that puts nonzero weight on all values of θ between 0 and 1 is generally completely dominated by the likelihood even when the number of observations is fairly small.

Posterior distributions

For a **uniform prior**, we get the posterior,

$$p(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y}$$

For a **Beta prior** with parameters *a* and *b* (see Bolstad, p.144, Fig 8.1), we get the posterior,

$$p(\theta|y,a,b) = \frac{\Gamma(n+a+b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}$$

(compare this with the Beta prior above)

Robustness to prior: vary *a* and *b* and see the effect on the posterior.

Normal prior

The normal is not a particularly natural choice for the binomial situation. However, the Beta distribution converges to the normal as the parameters *a* and *b* increase, so the posterior will be approximately normal as the number of observations, *n*, increases, as will an informative prior with large *a* and/or *b*.

Since θ is bounded in the interval [0,1], the normal will be *truncated* at 0 and 1.

$$p(\theta|\mu_0, \sigma_0^2) = (2\pi\sigma_0^2)^{-\frac{1}{2}}\exp\left[-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right], \quad 0 \leq \theta \leq 1$$

mean(θ) = μ_0, variance(θ) = σ_0^2

The resulting posterior is

$$p(\theta|\mu_0, \sigma_0^2) = c\ (2\pi\sigma_0^2)^{-\frac{1}{2}}\exp\left[-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right]\theta^y(1-\theta)^{n-y}, \quad 0 \leq \theta \leq 1$$

Rearranging gives

$$p(\theta|\mu_0, \sigma_0^2) = c(\sigma_0^2)\exp\left[-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right]\theta^y(1-\theta)^{n-y}, \quad 0 \leq \theta \leq 1$$

Note that the normalizing constant, *c*, *is* a constant, though here it is written as a function of σ_0^2, $c(\sigma_0^2)$, to indicate its dependence on the prior variance.

Since the above posterior is not in any known distributional form, the easiest way to evaluate it is to use **simulation methods** to obtain draws from the posterior.

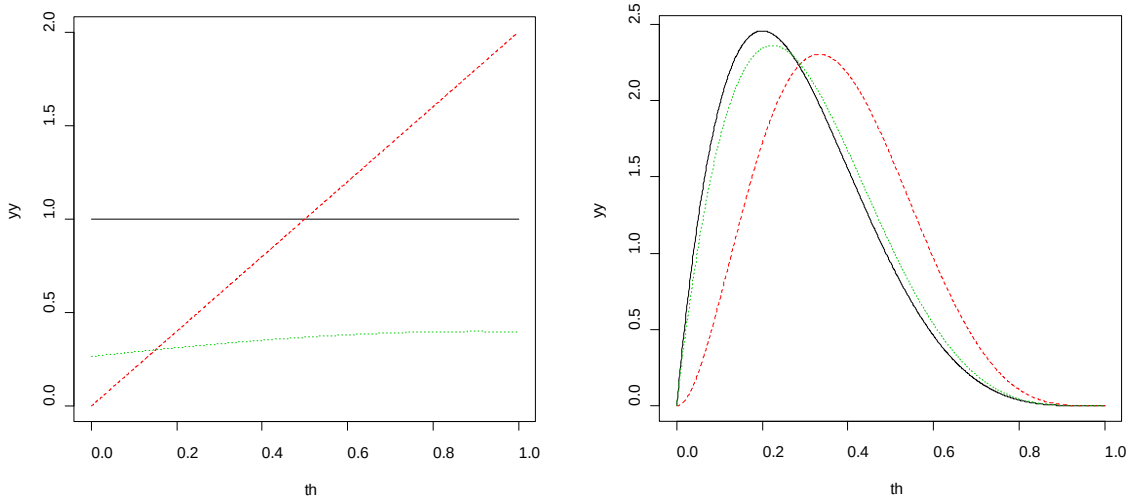
We can use a Gibbs MCMC algorithm to draw from the conditionals.

- Can we generate draws from the normbin example above (see also graphs below)?

Example of Uniform (black), Beta (red) and Normal (green) priors and posteriors

with $n = 5$, $y = 1$ (see binomialposteriors.R)

parameters for beta prior $a = 2$ $b = 1$
n = number of obs, y = number of 'successes' in n trials $n = 5$ $y = 1$
prior parameters for normal prior $m0 = 0.9$ $s20 = 1$

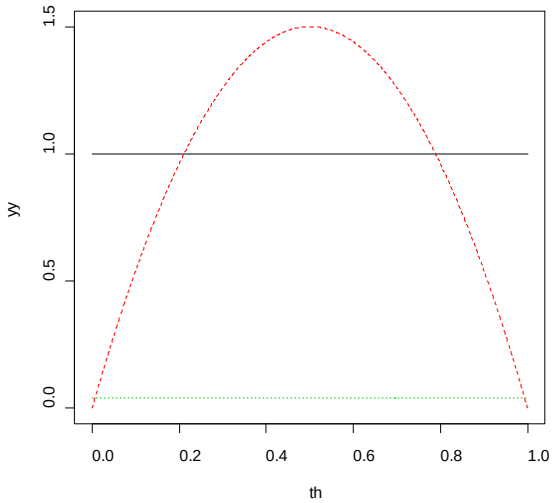


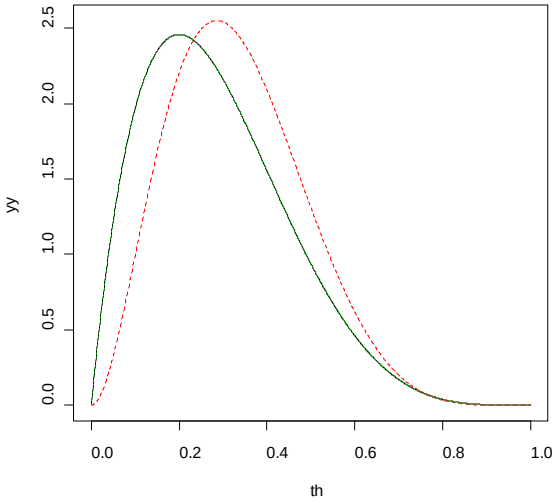
Note: the normal is not scaled properly (could integrate to get correct normalizing constant, but not necessary for posterior evaluation)

Note: parameters chosen to be fairly informative and incorrect (to illustrate robustness of posterior to variations in the prior). The normal prior really is not a good choice since it is not bounded between 0 and 1, and we use bad choices for the mean and variance. Despite this, we still get a posterior that is close to that from the uniform distribution, demonstrating that the posterior is reasonably robust to variations in the prior.

More uninformative priors:

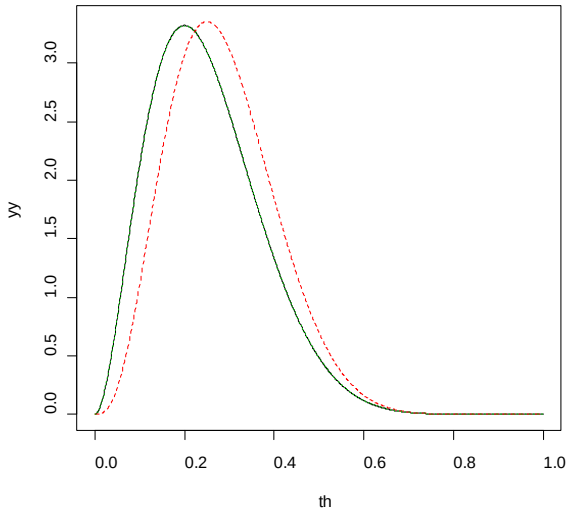
parameters for beta prior $a = 2$ $b = 2$
n = number of obs, y = number of 'successes' in n trials $n = 5$ $y = 1$
prior parameters for normal prior $m0 = 0.9$ $s20 = 100$

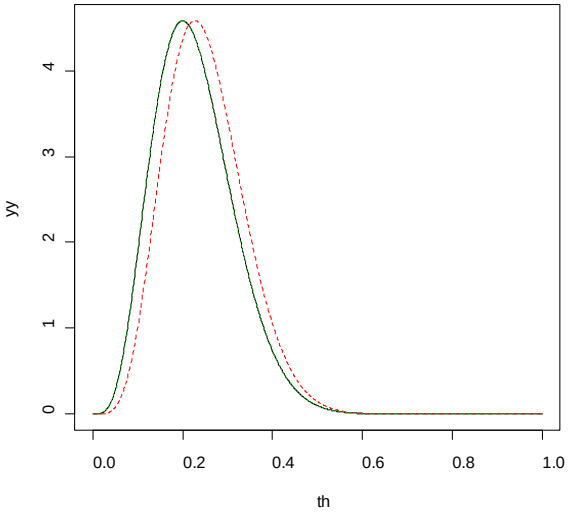




Effect of increasing *n*

n = number of obs, y = number of 'successes' in n trials $n = 10$ $y = 2$ and $n = 20$ $y = 4$
(same priors as above, normal prior: $m0 = 0.9$ $s20 = 100$, beta prior: $a = 2$ $b = 2$)





$n = 10$

$n = 20$

R code to generate above graphs (in binomialposterior.R)

```
# binomialposteriors.R
#
# Evaluates and plots posteriors for binomial experiment with
# three different priors: normal, beta and uniform

# set vectors to store values
R = 1001
th = rep(0,R)
normbin = rep(0,R)
betaprior = rep(0,R)
betapost = rep(0,R)
unifpost = rep(0,R)
normprior = rep(0,R)

# parameters for beta prior
a = 2
b = 2

# n = number of obs, y = number of 'successes' in n trials
n = 20
y = 4

# prior parameters for normal prior
m0 = 0.9
s20 = 100

# univariate numerical integration to determine normalizing
constant for normbin spec
int1 <- function(z) (exp(-((z-m0)^2)/(2*s20)))*(z^y)*(1-z)^(n-y)
cth = integrate(int1, lower = 0, upper = 1.0)
# c is the normalizing constant for use with normbin
c = cth$value
c


# Evaluate each posterior from 0 to 1, and the beta and normal
priors
# normbin = posterior with normal prior
# betapost = posterior with beta prior
# unifpost = posterior with uniform prior
# betaprior = beta prior
# normprior = normal prior
```

```
for (i in 1:R)
{
th[i] = i*0.001 - 0.001
normbin[i] = (1/c)*(exp(-((th[i]-m0)^2)/(2*s20)))*(th[i]^y)*(1-
th[i])^(n-y)
betaprior[i] = (gamma(a+b)/(gamma(a)*gamma(b)))*(th[i]^(a-
1))*(1-th[i])^(b-1)
betapost[i] = (gamma(n+a+b)/(gamma(y+a)*gamma(n-
y+b)))*(th[i]^(y+a-1))*(1-th[i])^(n-y+b-1)
unifpost[i] = (gamma(n+2)/(gamma(y+1)*gamma(n-
y+1)))*(th[i]^y)*(1-th[i])^(n-y)
normprior[i] = (1/sqrt(2*3.142*s20))*exp(-((th[i]-
m0)^2)/(2*s20))

}

# plot of everything on one graph
yy = cbind(normbin,betaprior,betapost,unifpost,normprior)
matplot(th,yy,type='l', col= 1:5)

# plot of priors
# uniform prior = 1 for all 0 < theta < 1
unif = rep(1,R)
yy = cbind(unif,betaprior,normprior)
matplot(th,yy,type='l', col= 1:3)

# plot of posteriors
yy = cbind(unifpost,betapost,normbin)
matplot(th,yy,type='l', col= 1:3)
```

Prediction is defined as making probability statements about the distribution of as yet unobserved data, denoted by y_f. The only real distinction between parameters and unobserved data is that y_f is potentially observable.

Predictive distribution of y_f given y, $p(y_f|y)$:

$$p(y_f|y) = \int p(y_f,\theta|y)\, d(\theta) = \int p(y_f|\theta,y)\, p(\theta|y)\, d\theta.\text{"}$$

[Rossi, Allenby &McCulloch (2005)]

$p(y_f|\theta,y)$ is the predictive distribution for y conditional on θ and the data,

$p(\theta|y)$ is the posterior distribution for θ.