2021 SISG Bayesian Statistics for Genetics R Notes: Binomial Sampling

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

2021-07-14

Binomial Data

Introduction

In these notes, in the context of binomial sampling, we look at:

- Specifying a prior distribution
- Prediction.
- Testing.
- Logistic regression.

We analyze allele specific expression (ASE) data, and low birth weight data.

Specifying a prior distribution

The code below finds the beta distribution, i.e. the a and the b, with 5% and 95% points of 0.1 and 0.6.

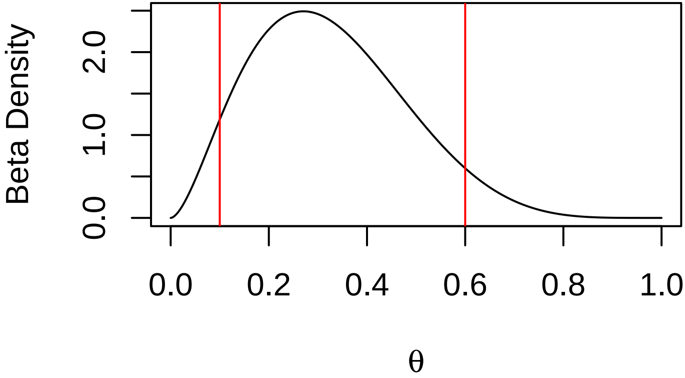
```
# Function to find a and b
priorch <- function(x, q1, q2, p1, p2) {
    (p1 - pbeta(q1, x[1], x[2]))^2 + (p2 - pbeta(q2,
        x[1], x[2]))^2
}
p1 <- 0.05
p2 <- 0.95
q1 <- 0.1
q2 <- 0.6
opt <- optim(par = c(1, 1), fn = priorch, q1 = q1,
    q2 = q2, p1 = p1, p2 = p2, control = list(abstol = 1e-08))
cat("a and b are ", opt$par, "\n")
## a and b are  2.730616 5.667462
```

Specifying a prior distribution

The code below finds the beta distribution, i.e. the a and the b, with 5% and 95% points of 0.1 and 0.6.

```
probvals <- seq(0, 1, 0.001)
plot(probvals, dbeta(probvals, shape1 = opt$par[1],
    shape2 = opt$par[2]), type = "l", xlab = expression(theta),
    ylab = "Beta Density")
abline(v = q1, col = "red")
abline(v = q2, col = "red")
```

Specifying a prior distribution



Differences in Binomial Proportions

We consider an example in which we wish to compare allele frequencies between two populations.

Let θ_1 and θ_2 be the allele frequencies in the NE and US population from which the samples were drawn, respectively.

The allele frequencies were 10.69% and 13.21% with sample sizes of 650 and 265, in the NE and US samples, respectively.

We assume independent Beta(1,1) priors on each of θ_1 and θ_2.

Differences in Binomial Proportions

The y1 and y2 data (i.e. the numbers with the allele in the two populations) were reconstructed from figures in the original paper in which only the denominators and the frequencies were given, hence the `floor` function.

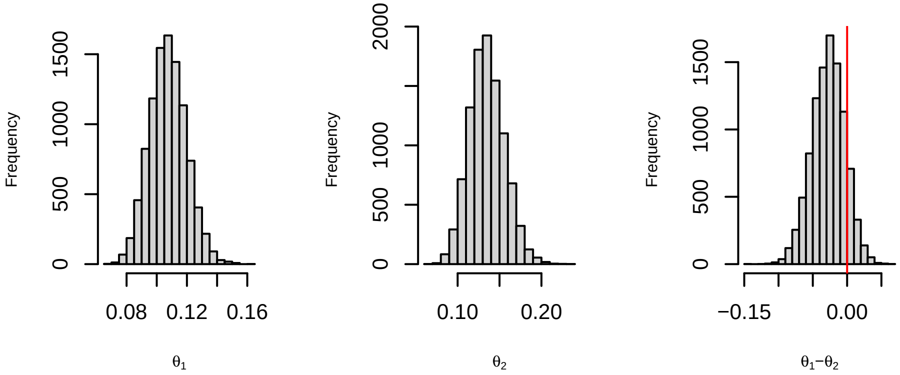
```
N1 <- 650
y1 <- floor(N1 * 0.1069)
N2 <- 265
y2 <- floor(N2 * 0.1321)
nsamp <- 10000
a <- b <- 1
post1 <- rbeta(nsamp, y1 + a, N1 - y1 + b)
post2 <- rbeta(nsamp, y2 + a, N2 - y2 + b)
```

Differences in Binomial Proportions

The key step is in constructing a sample estimate of the difference in probabilities $\theta_1 - \theta_2$.

```
par(mfrow = c(1, 3))
hist(post1, xlab = expression(theta[1]), main = "",
    cex.lab = 0.7)
hist(post2, xlab = expression(theta[2]), main = "",
    cex.lab = 0.7)
#
hist(post1 - post2, xlab = expression(paste(theta[1],
    "-", theta[2])), main = "", cex.lab = 0.7)
abline(v = 0, col = "red")
sum(post1 - post2 > 0)/nsamp
## [1] 0.1242
```

Differences in Binomial Proportions



ASE Data

Analysis of ASE data

```
download.file("http://faculty.washington.edu/kenrice/sisgbayes/ASEgene.txt",
    destfile = "ASEgene.txt")
ASEdat <- read.table("ASEgene.txt", header = TRUE)
head(ASEdat)
##     Y    N
## 1  62  107
## 2  33   59
## 3 658 1550
## 4  14   61
## 5  57  153
## 6 218  451
dim(ASEdat)
## [1] 4844    2
ngenes <- dim(ASEdat)[1]
pvals <- NULL
for (i in 1:ngenes) {
    pvals[i] <- binom.test(ASEdat$Y[i], ASEdat$N[i],
        p = 0.5, alternative = "two.sided")[["p.value"]]
}
```

Analysis of ASE data

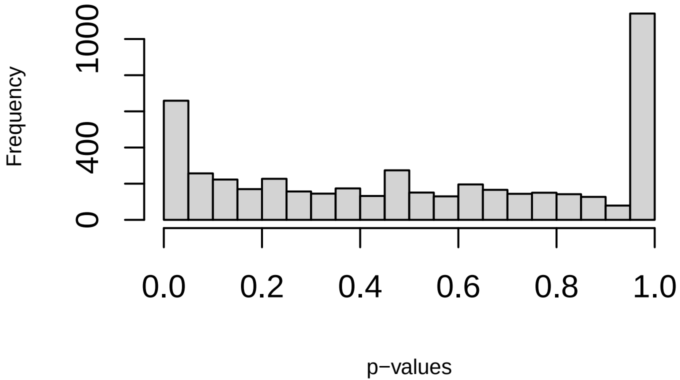
```
# Function to evaluate Bayes factors for a
# binomial likelihood and beta prior, and a point
# null at p0
BFbinomial <- function(N, y, a, b, p0) {
    logPrH0 <- lchoose(N, y) + y * log(p0) + (N - y) *
        log(1 - p0)
    logPrH1 <- lchoose(N, y) + gamma(a + b) - lgamma(a) -
        lgamma(b) + lgamma(y + a) + lgamma(N - y +
        b) - lgamma(N + a + b)
    logBF <- logPrH0 - logPrH1
    list(logPrH0 = logPrH0, logPrH1 = logPrH1, logBF = logBF)
}
nsim <- 5000
a <- 1
b <- 1
p0 <- 0.5
```

Analysis of ASE data

```r
postprob <- logBFr <- rep(0, ngenes)
pcutoff <- 0.05/length(pvals)
for (i in 1:ngenes) {
    BFcall <- BFbinomial(ASEdat$N[i], ASEdat$Y[i],
        a, b, p0)
    logBFr[i] <- -BFcall$logBF
    postprob[i] <- pbeta(0.5, a + ASEdat$Y[i], b +
        ASEdat$N[i] - ASEdat$Y[i])
}
cat("log BFr > log(150) = ", sum(logBFr > log(150)),
    "\n")
## log BFr > log(150) =  197
cat("log BFr > log(20) = ", sum(logBFr > log(20)),
    "\n")
## log BFr > log(20) =  359
cat("p-values >  ", pcutoff, sum(pvals < pcutoff),
    "\n")
## p-values >   1.032205e-05 111
cat("postprobs < 0.01 and > 0.99 ", sum(postprob <
    0.01), sum(postprob > 0.99), "\n")
## postprobs < 0.01 and > 0.99  278 242
```

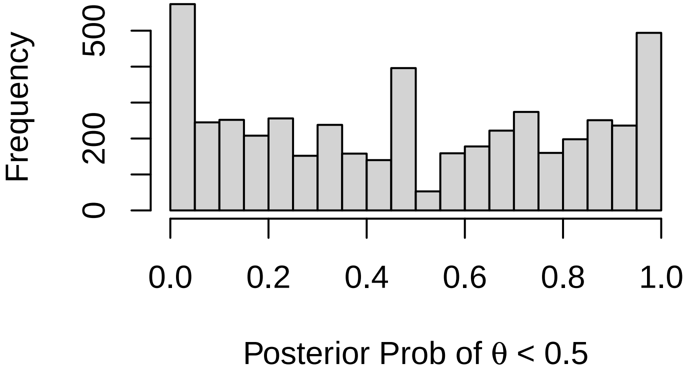
Histogram of *p*-values for ASE data

```
hist(pvals, xlab = "p-values", main = "", nclass = 20,
    cex.lab = 0.7, pch = 16, cex = 0.5)
```

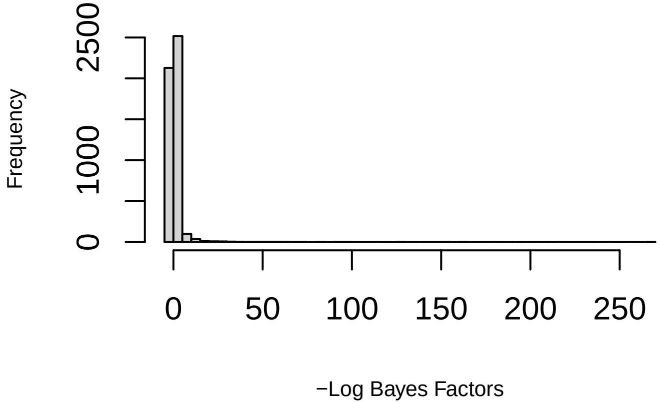


Histogram of posterior probabilities for ASE data

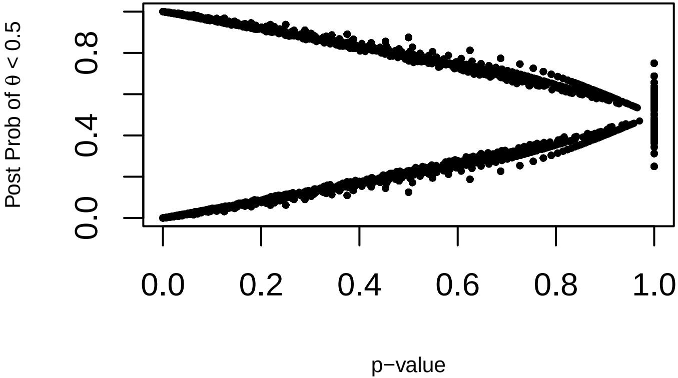
```
hist(postprob, nclass = 20, xlab = expression(paste("Posterior Prob of ",
    theta, " < 0.5")), main = "", cex.label = 0.7)
```



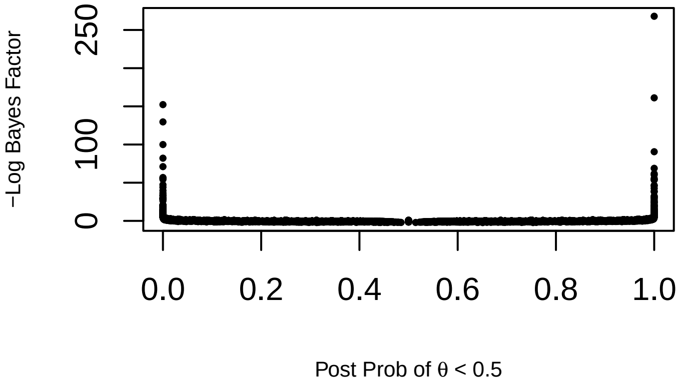
Analysis of ASE data



Analysis of ASE data



Analysis of ASE data



Logistic Regression

Birth weight analysis

Hosmer and Lemeshow (2004) present data on 189 births to women participating in a larger study at Baystate Medical Center (Springfield, MA), with information on the following variables:

- `LOW`: Low birth weight; ($0 = \geq 2500$g; $1 = < 2500$g)
- `AGE`: Mother's age
- `LWT`: Mother's weight
- `RACE`: Listed race of mother; (1 = white; 2 = black; 3 = other)
- `SMOKE`: Smoking status during pregnancy; (0 = no; 1 = yes)
- `HT`: History of hypertension; (0 = no; 1 = yes)
- `UI`: Presence of uterine irritability; (0 = no; 1 = yes)
- `FTV`: Number of physician visits during the first trimester.

This data (`lowbwt`) can be found in the `brinla` package, which contains code and data from *Bayesian Regression with INLA* by Wang et al. (2018).

Birth weight analysis: load packages and data

```r
# devtools::install_github('julianfaraway/brinla')
library(brinla)
library(dplyr)
library(tidyr)
data(lowbwt)
knitr::opts_chunk$set(dev = "pdf")
```

Frequentist logistic regression with glm

```
bwt.glm <- glm(LOW ~ AGE + LWT + RACE + SMOKE + HT +
    UI + FTV, data = lowbwt, family = binomial())
round(coef(summary(bwt.glm)), 3)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)    0.455      1.185   0.384    0.701
## AGE           -0.021      0.036  -0.570    0.568
## LWT           -0.017      0.007  -2.409    0.016
## RACE2          1.290      0.528   2.445    0.015
## RACE3          0.919      0.436   2.106    0.035
## SMOKE1         1.042      0.395   2.634    0.008
## HT1            1.885      0.695   2.713    0.007
## UI1            0.904      0.449   2.015    0.044
## FTV            0.059      0.172   0.344    0.731
```

Logistic regression with INLA (default prior)

```
bwt.inla <- inla(LOW ~ AGE + LWT + RACE + SMOKE + HT +
    UI + FTV, data = lowbwt, family = "binomial", Ntrials = 1,
    control.compute = list(dic = TRUE, cpo = TRUE))
round(bwt.inla$summary.fixed[, 1:5], 3)
##                mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)  0.567 1.186     -1.727    0.555      2.925
## AGE         -0.021 0.036     -0.092   -0.020      0.049
## LWT         -0.018 0.007     -0.032   -0.017     -0.005
## RACE2        1.340 0.528      0.316    1.336      2.386
## RACE3        0.946 0.436      0.104    0.941      1.816
## SMOKE1       1.075 0.395      0.315    1.070      1.867
## HT1          1.974 0.694      0.664    1.955      3.391
## UI1          0.933 0.449      0.052    0.933      1.813
## FTV          0.056 0.172     -0.289    0.059      0.386
```

Logistic regression example: comparison of estimates

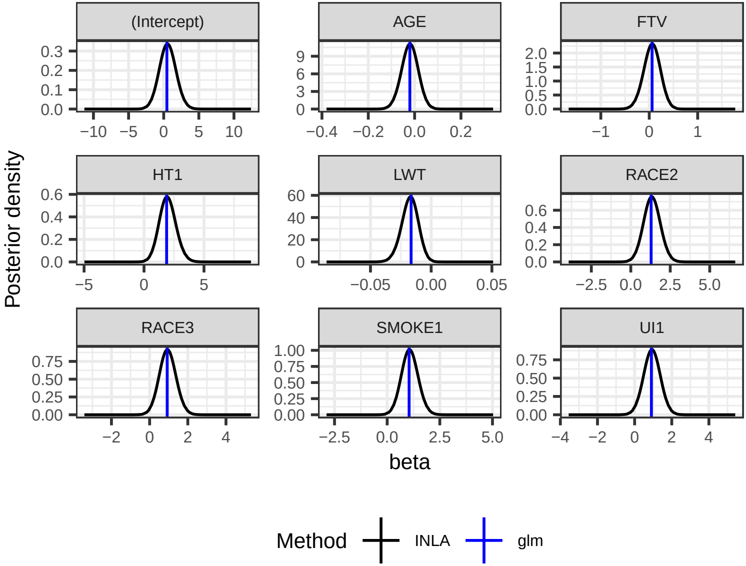
```
temp <- cbind(coef(summary(bwt.glm))[, 1:2], bwt.inla$summary.fixed[,
    1:2])
temp
##                Estimate  Std. Error        mean          sd
## (Intercept)  0.45481245 1.185407237  0.56676644 1.185562015
## AGE         -0.02050558 0.035952662 -0.02068527 0.035960548
## LWT         -0.01652405 0.006859658 -0.01760207 0.006853204
## RACE2        1.28975584 0.527614670  1.34031296 0.527672708
## RACE3        0.91906824 0.436301131  0.94558495 0.436260819
## SMOKE1       1.04159102 0.395478361  1.07504070 0.395394474
## HT1          1.88506202 0.694814773  1.97350609 0.694086432
## UI1          0.90414991 0.448602226  0.93278354 0.448573530
## FTV          0.05911925 0.171995642  0.05592564 0.171968043
```

Logistic regression with INLA (default prior)

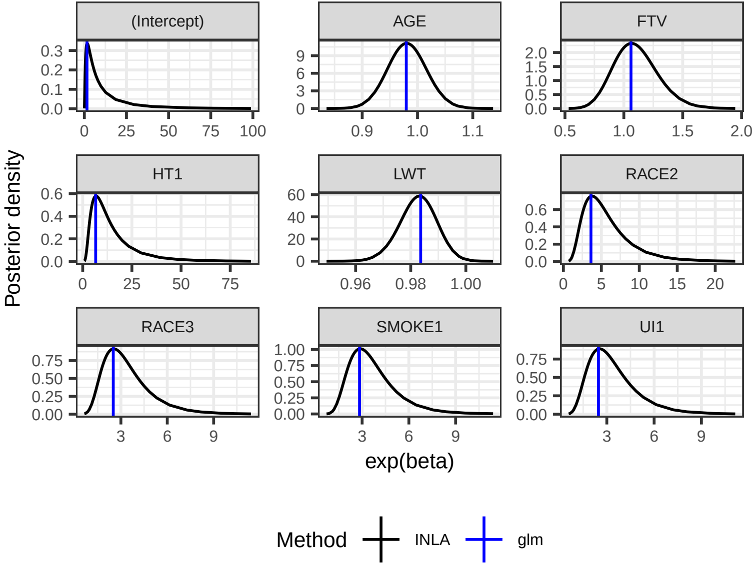
We calculate $\Pr(\beta_j > 0|y)$, $j = 0, \ldots, 8$.

```
round(unlist(lapply(bwt.inla$marginals.fixed, function(x) 1 -
    inla.pmarginal(0, x))), 3)
## (Intercept)         AGE         LWT       RACE2       RACE3      SMOKE1
##       0.680       0.283       0.003       0.995       0.986       0.998
##         HT1         UI1         FTV
##       0.999       0.981       0.632
```

Logistic regression with INLA (default prior)



Logistic regression with INLA (default prior)



Logistic regression with INLA (alternate prior)

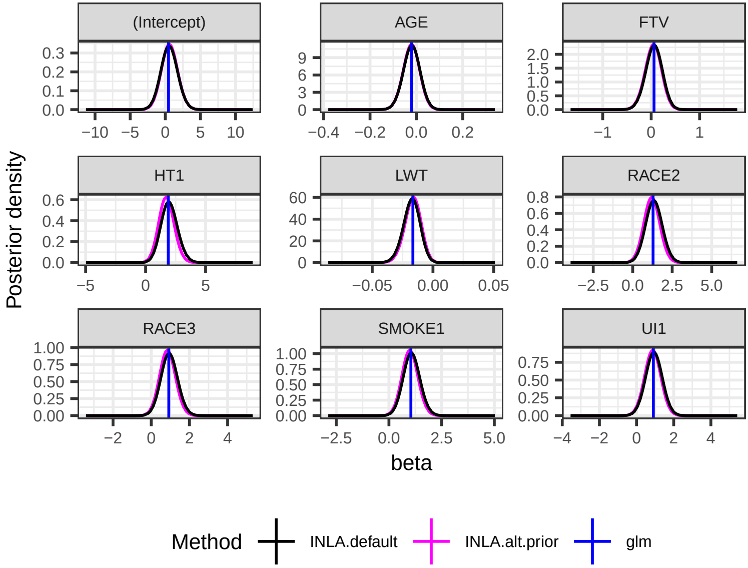
```
alt.prior.fixed <- list(mean.intercept = 0, prec.intercept = 1/10^2,
    mean = 0, prec = 1/2^2)
bwt.inla.alt <- inla(LOW ~ AGE + LWT + RACE + SMOKE +
    HT + UI + FTV, data = lowbwt, family = "binomial",
    Ntrials = 1, control.compute = list(dic = TRUE,
        cpo = TRUE), control.fixed = alt.prior.fixed)
round(bwt.inla.alt$summary.fixed, 3)
##                mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)  0.607 1.157     -1.632    0.595      2.909  0.573   0
## AGE         -0.022 0.035     -0.092   -0.021      0.047 -0.021   0
## LWT         -0.017 0.007     -0.030   -0.016     -0.004 -0.016   0
## RACE2        1.210 0.503      0.229    1.208      2.201  1.204   0
## RACE3        0.854 0.416      0.047    0.850      1.679  0.844   0
## SMOKE1       0.999 0.379      0.267    0.995      1.756  0.987   0
## HT1          1.740 0.643      0.511    1.729      3.036  1.706   0
## UI1          0.869 0.435      0.013    0.869      1.721  0.870   0
## FTV          0.043 0.170     -0.298    0.046      0.369  0.051   0
```

Logistic regression with INLA (alternate prior)

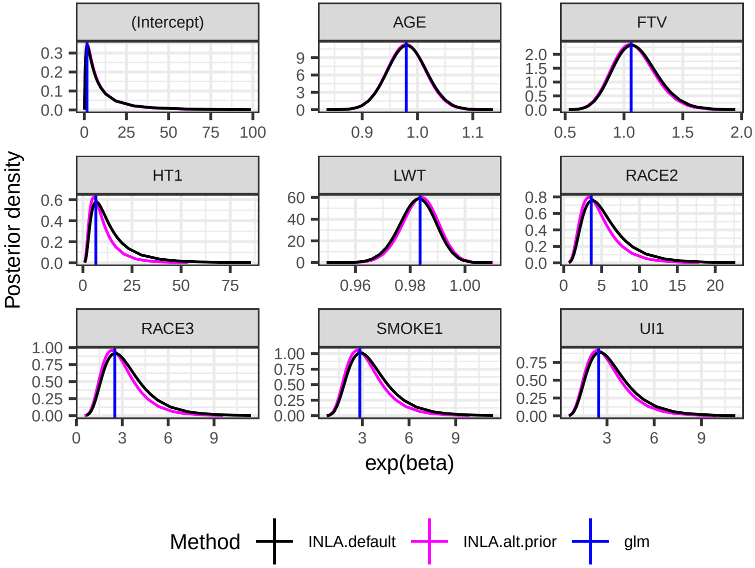
We calculate $\Pr(\beta_j > 0|y)$, $j = 0, \ldots, 8$.

```
round(unlist(lapply(bwt.inla.alt$marginals.fixed, function(x) 1 -
    inla.pmarginal(0, x))), 3)
## (Intercept)         AGE         LWT       RACE2       RACE3      SMOKE1
##       0.696       0.270       0.004       0.992       0.981       0.997
##         HT1         UI1         FTV
##       0.998       0.977       0.605
```

Logistic regression with INLA (alternate prior)



Logistic regression with INLA (alternate prior)



Prediction

Predictions from a Binomial Distribution

We now consider prediction.

Assume $y|\theta \sim \text{binomial}(N, \theta)$ and $\theta \sim \text{beta}(a, b)$.

We suppose we wish to predict the number of successes Z from M trials.

The predictive distribution is

$$\Pr(z|y) = \left(\begin{array}{c} M \\ z \end{array} \right) \frac{\Gamma(N+a+b)}{\Gamma(y+a)\Gamma(N-y+b)} \frac{\Gamma(a+y+z)\Gamma(b+N-y+M-z)}{\Gamma(a+b+N+M)}$$

for $z = 0, \dots, M$.

Predictions from a Binomial Distribution

We demonstrate with a uniform prior and observing $y = 2$ successes from $N = 20$ trials, and suppose we wish to predict the number of successes we will see in 10 additional trials.

```
# User written function
binomialpred <- function(a, b, y, N, z, M) {
    lchoose(M, z) + lgamma(a + b + N) - lgamma(a +
        y) - lgamma(b + N - y) + lgamma(a + y + z) +
        lgamma(b + N - y + M - z) - lgamma(a + b +
        N + M)
}
# Set up the prior and data
a <- b <- 1
y <- 2
N <- 20
M <- 10
```

Predictions from a Binomial Distribution

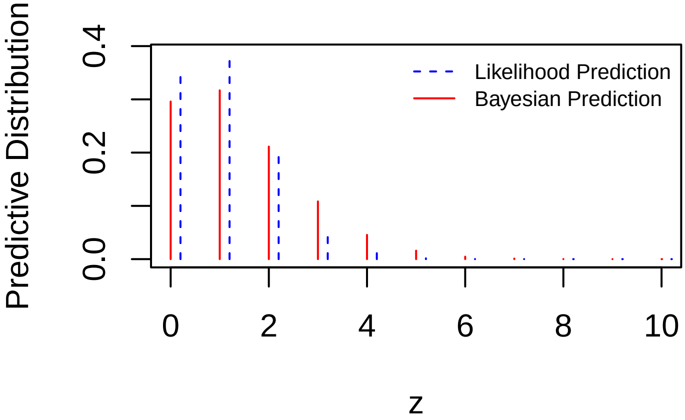
Along with the Bayesian predictive distribution, we also include a simple approach in which we assume simply take a `binomial(M,y/N)` distribution, i.e. assuming the probability is known to be the sample fraction.

```
binpred <- NULL
z <- seq(0, M)
sumcheck <- 0
for (i in 1:(M + 1)) {
    binpred[i] <- exp(binomialpred(a, b, y, N, z[i],
        M))
    sumcheck <- sumcheck + binpred[i]
}
likpred <- dbinom(z, M, prob = y/N)
cat("Sum of probs = ", sumcheck, "\n")
## Sum of probs =  1
```

Predictions from a Binomial Distribution

```
plot(binpred ~ z, type = "h", col = "red", ylim = c(0,
    max(likpred, binpred)), ylab = "Predictive Distribution")
points(z + 0.2, likpred, type = "h", col = "blue",
    lty = 2)
legend("topright", legend = c("Likelihood Prediction",
    "Bayesian Prediction"), lty = 2:1, col = c("blue",
    "red"), bty = "n")
```

Predictions from a Binomial Distribution



Predictions with a Binomial Distribution

We now simulate directly via:

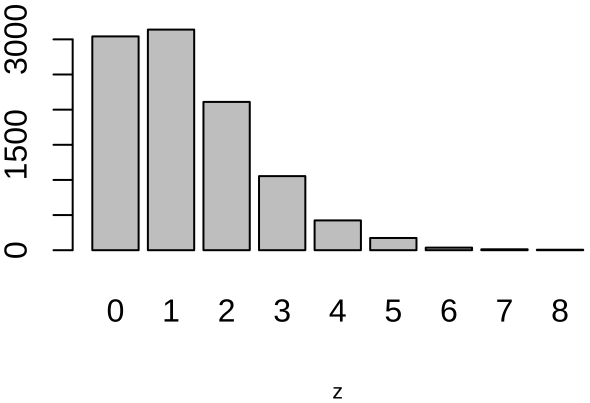
– Sampling from $\theta^{(s)} \sim p(\theta|y)$, $s = 1, \dots, S$.

– Sampling from $z^{(s)} \sim p(z|\theta)$, $s = 1, \dots, S$.

```
a <- b <- 1
y <- 2
N <- 20
M <- 10
nsim <- 10000
theta <- z <- NULL  # This is inefficient but makes method clear
for (s in 1:nsim) {
    theta[s] <- rbeta(1, a + y, b + N - y)
    z[s] <- rbinom(1, M, theta[s])
}
```

Predictions with a Binomial Distribution

```
barplot(table(z), xlab = "z", cex.lab = 0.7)
```



Exercises

1. Experiment with the priors Beta(a, a) for the ASE example. In particular, for $a = 2$:
    - Obtain a histogram of the posterior probabilities $\Pr(\theta < 0.5|y)$, across genes.
    - Plot these posterior probabilities versus the versions under $a = 1$, and comment.
    - How sensitive are the (log) Bayes factors to the prior specification?
    - For how many genes would we reject $H_0 : \theta = 0.5$ if we use a rule of $1/\text{BF} > 150$?
2. Redo the logistic regression birthweight example, using the default priors, but with the smoking variable only in the model. Compare with a frequentist smoking only example with the `glm` function.