SISG Bayes: Exercise 3

Ken Rice

2024-06-03

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.

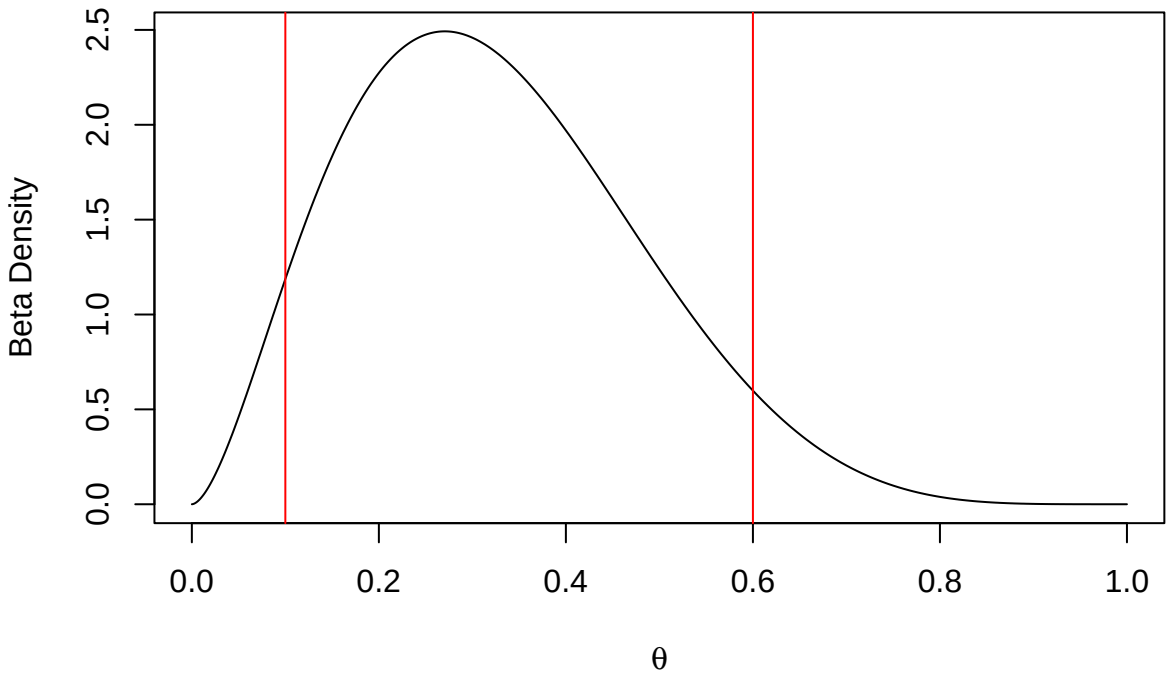
Specifying a prior distribution

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-8))
cat("a and b are ",opt$par,"\n")
## a and b are  2.730616 5.667462
```

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,.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")
```



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.

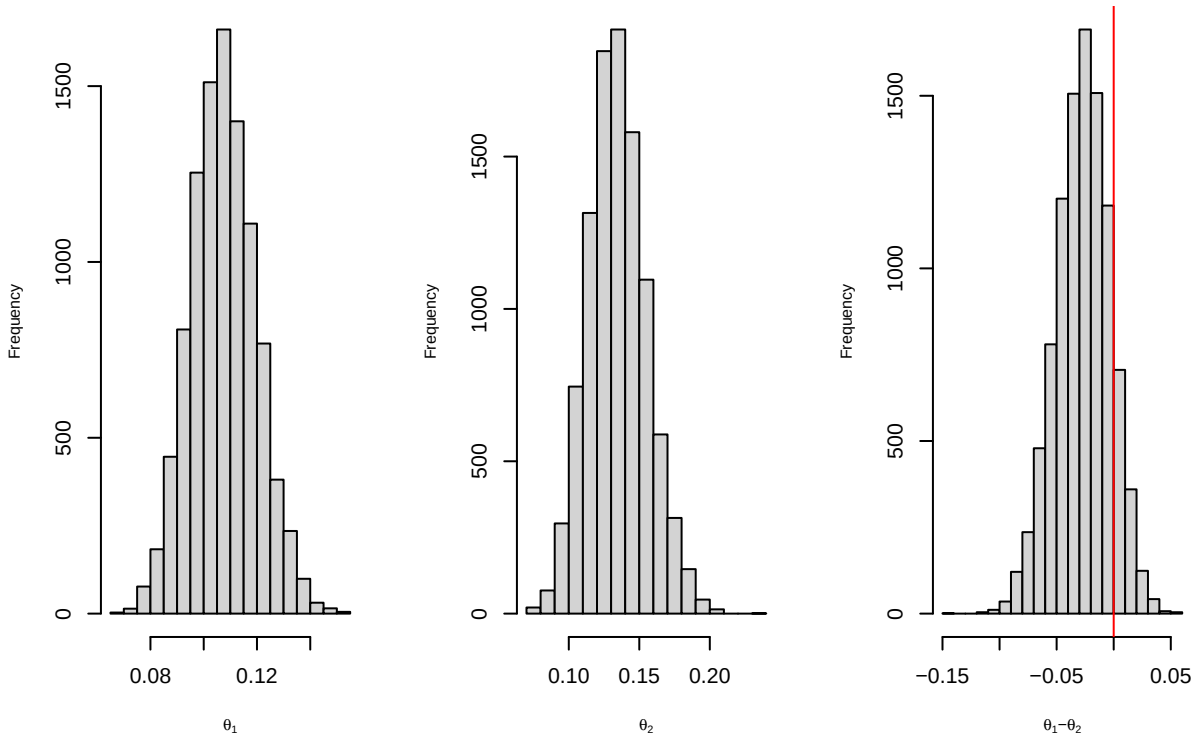
The `y1` and `y2` data (i.e. the numbers with the allele in the two populations) were reconstructed from figures in the original paper in which only the denominators and the frequencies were given, hence the `floor` function.

```
N1 <- 650
y1 <- floor(N1*.1069)
N2 <- 265
y2 <- floor(N2*.1321)
nsamp <- 10000
a <- b <- 1
post1 <- rbeta(nsamp,y1+a,N1-y1+b)
post2 <- rbeta(nsamp,y2+a,N2-y2+b)
```

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

```r
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")
```

```r
sum(post1-post2>0)/nsamp
## [1] 0.1243
```



ASE Data

Analysis of ASE data

```r
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"]]
}
```

```
# 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) -lgam
       logBF <- logPrH0 - logPrH1
       list(logPrH0=logPrH0,logPrH1=logPrH1,logBF=logBF)
}
nsim <- 5000
a <- 1
b <- 1
p0 <- 0.5
```

```
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
```

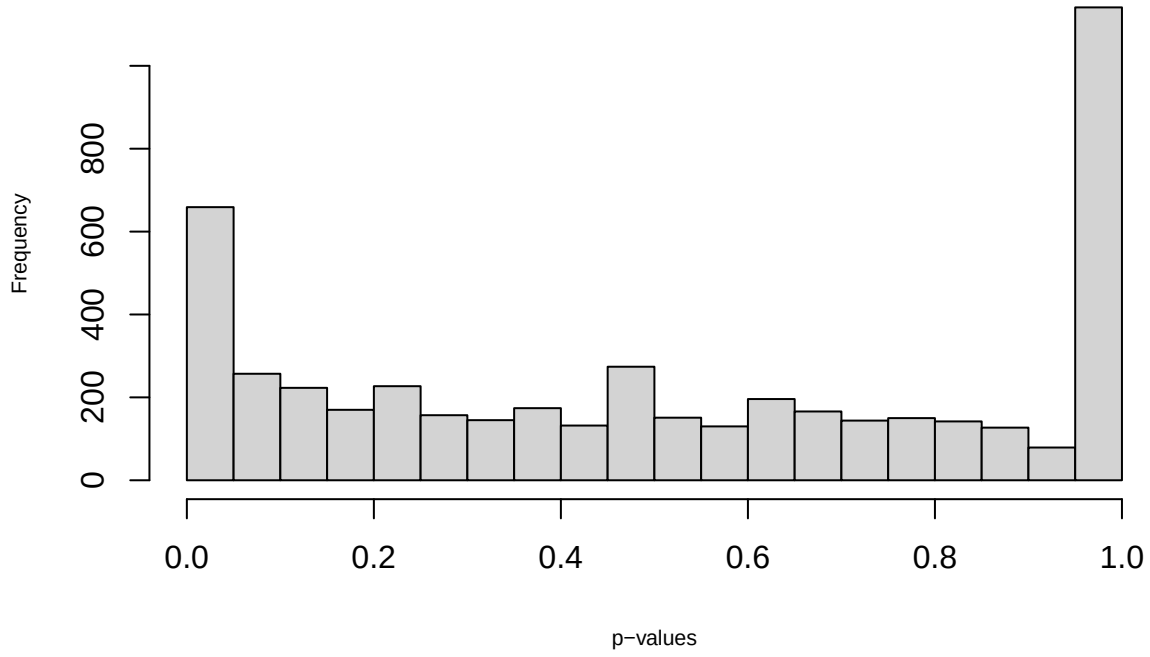
```
cat("log BFr > log(20) = ",sum(logBFr>log(20)),"\n")
## log BFr > log(20) =  359
```

```
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
```

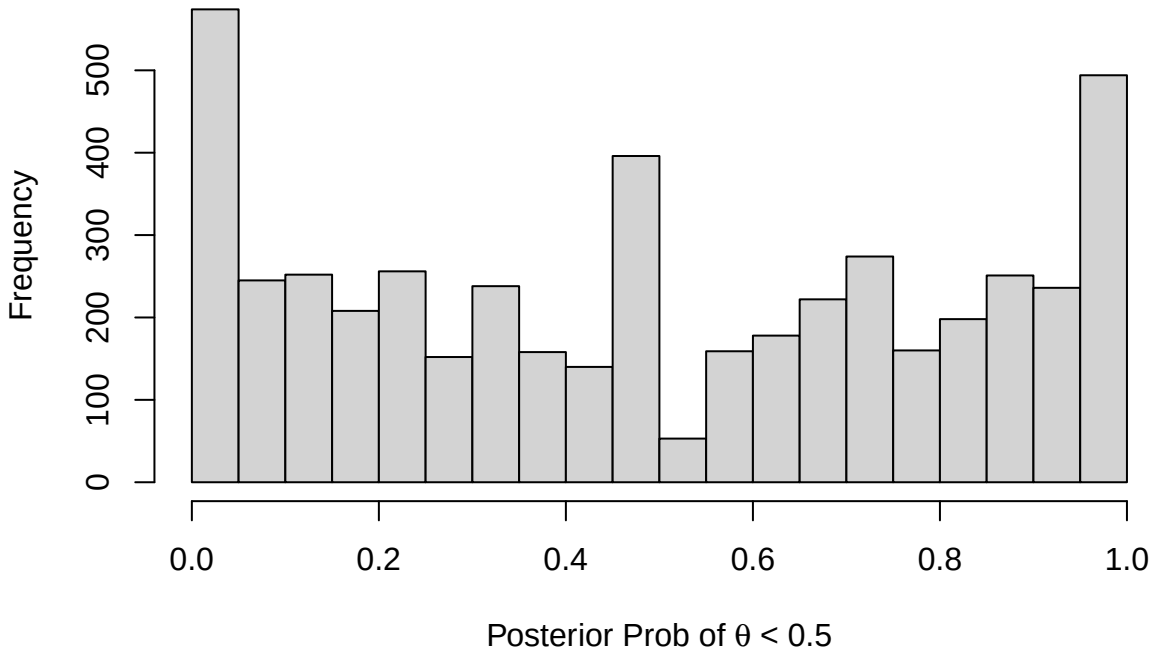
Histogram of p-values for ASE data

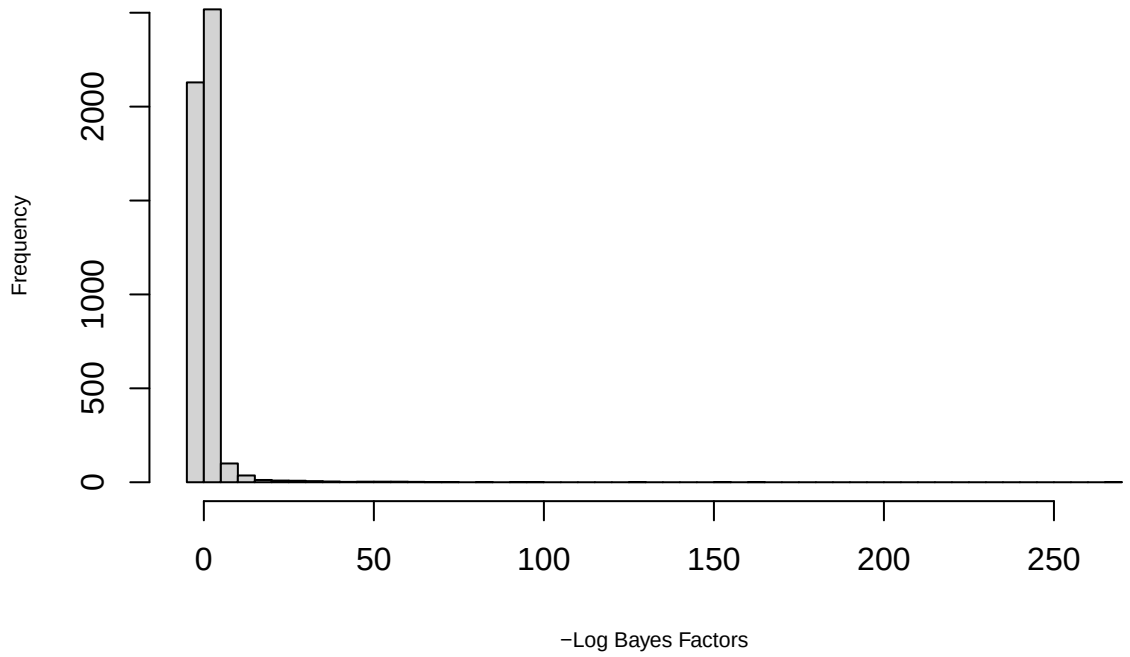
```
hist(pvals,xlab="p-values",main="",nclass=20,cex.lab=0.7,pch=16,cex=0.5)
```

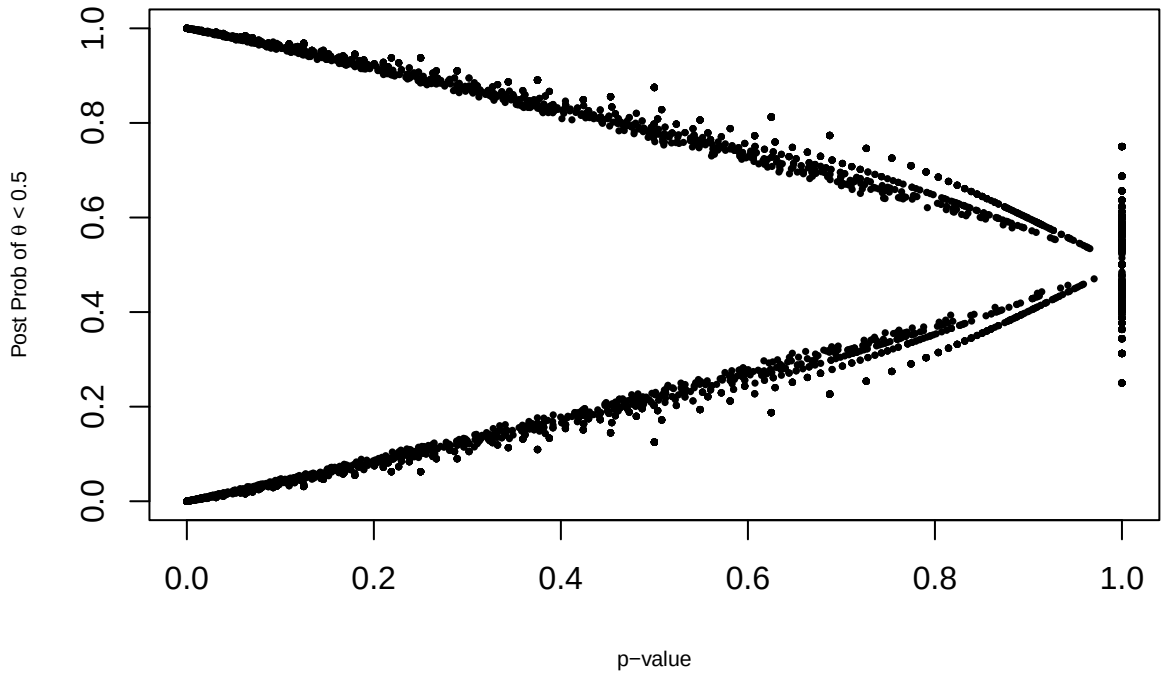


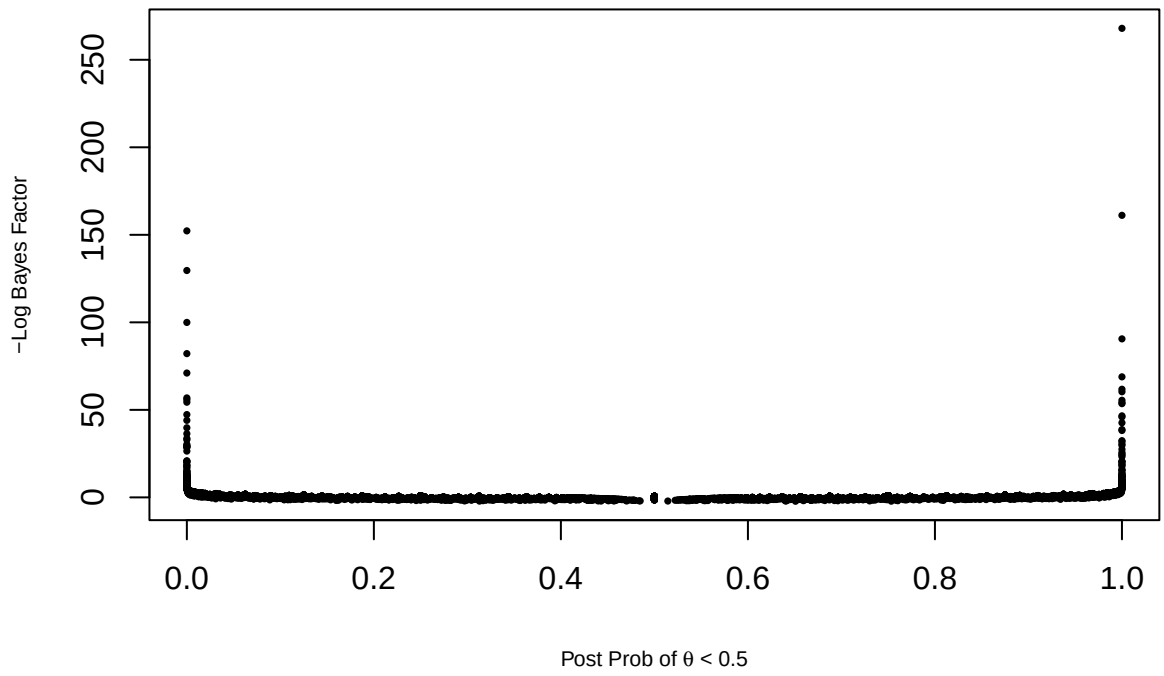
Histogram of posterior probabilities for ASE data

```
hist(postprob,nclass=20,xlab=expression(paste("Posterior Prob of ",theta," < 0.5")),main="",cex.label=0
```


Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Posterior Prob of θ < 0.5


Frequency
0
500
1000
2000
0
50
100
150
200
250
−Log Bayes Factors


Post Prob of θ < 0.5
0.0
0.2
0.4
0.6
0.8
1.0
p−value



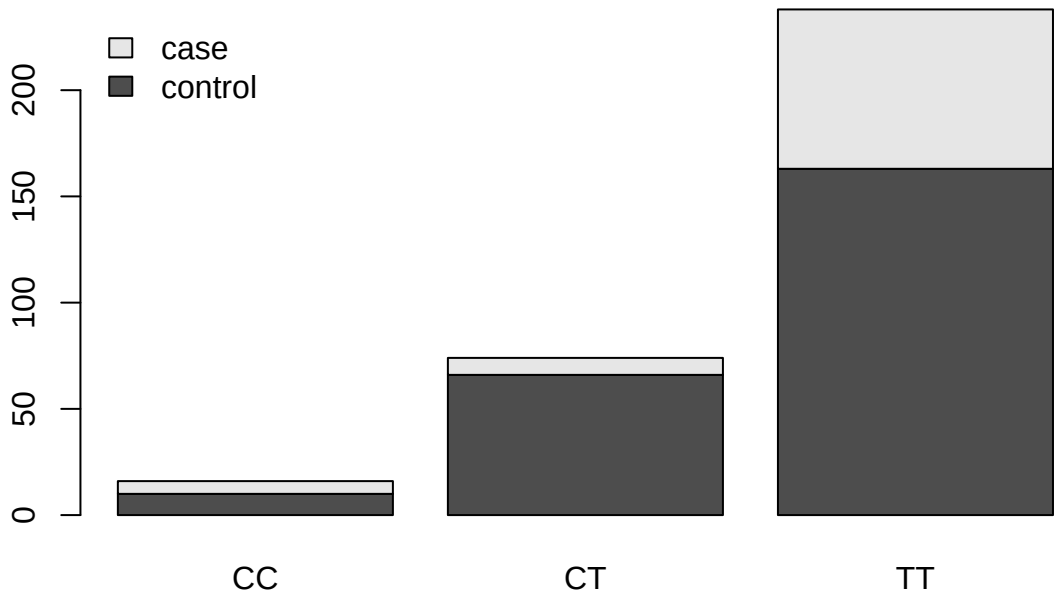
Logistic Regression

LHON example

We consider the LHON example from class

set up data and functions

```
library("ellipse")

lhon <- data.frame( y=c(1,1,1,0,0,0), x=c(0,1,2,0,1,2), n=c(6,8,75,10,66,163) )
dmat <- rbind(lhon$n[4:6], lhon$n[1:3])
names(dmat) <- c("CC","CT","TT")

expit <- function(x){exp(x)/(1+exp(x))}
lik <- function(beta){
    probs <- with(lhon, dbinom(y, 1, expit(beta[1] + beta[2]*x)))
    lik <- prod( probs^lhon$n )
    }

barplot(dmat, names.arg=c("CC","CT","TT"), legend.text=c("case","control")[2:1],
args.legend=list(x="topleft", bty="n"))
```

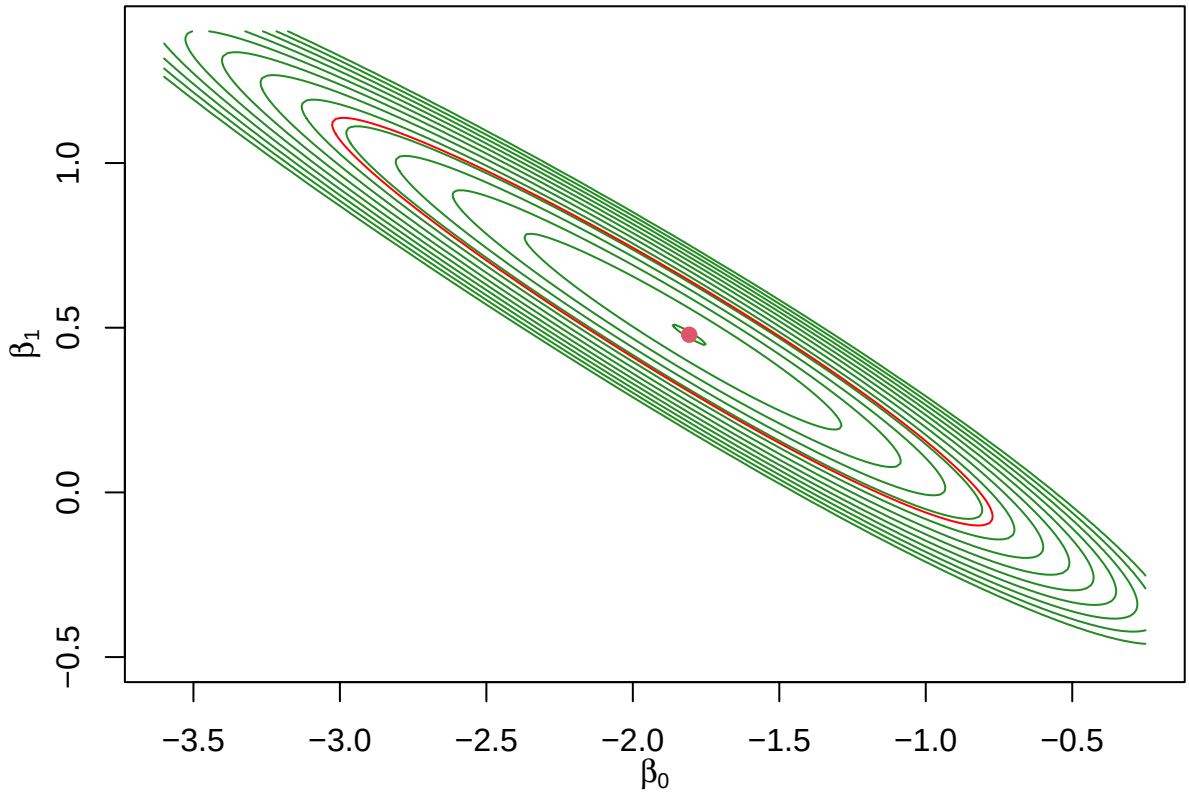



```
knitr::opts_chunk$set(dev = 'pdf')
```

Frequentist logistic regression with `glm`

```
glm1 <- glm( y~x, family="binomial", weights=n, data=lhon)
round(coef(summary(glm1)), 3)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.808      0.455  -3.970    0.000
## x              0.479      0.250   1.911    0.056
```

```
par(mar=c(3,3,1.5,1.5)+0.2)
b0vals <- seq(-3.6,-0.25,l=201)
b1vals <- seq(-0.5,1.4,l=201)
gg <- expand.grid(b0=b0vals, b1=b1vals)
gg$lik <- apply(gg, 1, lik)
contour(x=b0vals, y=b1vals, z=matrix(gg$lik, 201,201),
        xlab="", ylab="", levels=0.99*lik(coef(glm1))*0.5^(0:12), drawlabels=FALSE, col="forestgreen")
contour(x=b0vals, y=b1vals, z=matrix(gg$lik, 201,201),
        xlab="", ylab="", levels=lik(coef(glm1))/20, drawlabels=FALSE, add=TRUE, col="red")
points(coef(glm1)[1], coef(glm1)[2], pch=19, col=2)
mtext(side=1, expression(beta[0]), line=2)
mtext(side=2, expression(beta[1]), line=2)
```



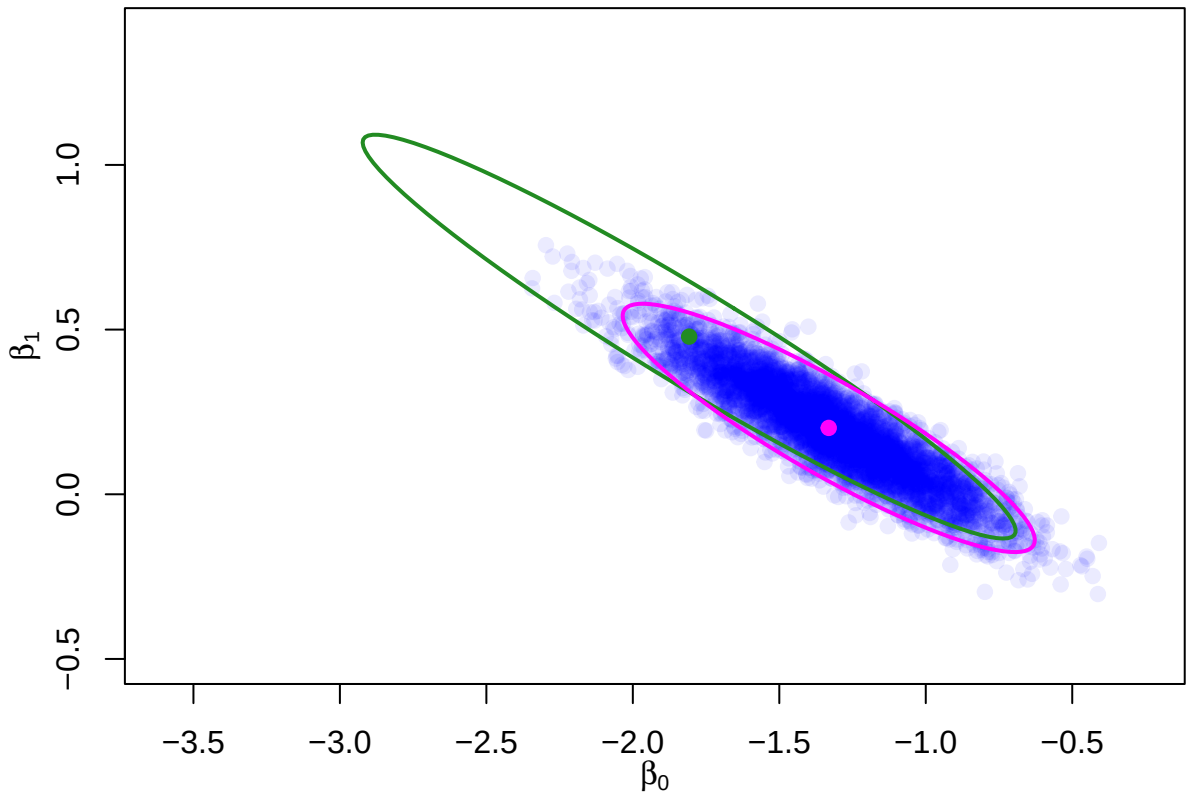
Logistic regression via rejection sampling (diffuse prior)

```
betahat <- coef(glm( y~x, family="binomial", weights=n, data=lhon))

do.many <- function(bigB){
#bigB <- 5
    beta.sample <- matrix(NA,bigB,2)
    i <- 1
    while(i<= bigB){
        beta.try <- c(rnorm(1, 0, sqrt(10)), rnorm(1, 0, sqrt(beta1pvar)))
        u <- runif(1)
        if(u < lik(beta.try)/lik(betahat)){
            beta.sample[i,] <- beta.try
            i <- i+1}
        }
    beta.sample
}

set.seed(4)
#informative prior selection
beta1pvar <- uniroot(function(w){ qnorm(0.975, 0, sqrt(w)) - log(1.5)}, c(0.01, 1))$root
beta.post <- as.data.frame(do.many(5000))
names(beta.post) <- c("beta0","beta1")
cbind( apply(beta.post, 2, mean), apply(beta.post, 2, sd))
```

```
##             [,1]      [,2]
## beta0 -1.331261 0.2874245
## beta1  0.201543 0.1540337
```

```
quantile(beta.post[,2], c(0.5, 0.025, 0.975))
##         50%        2.5%       97.5%
##  0.20241657 -0.09472991  0.50796330
```

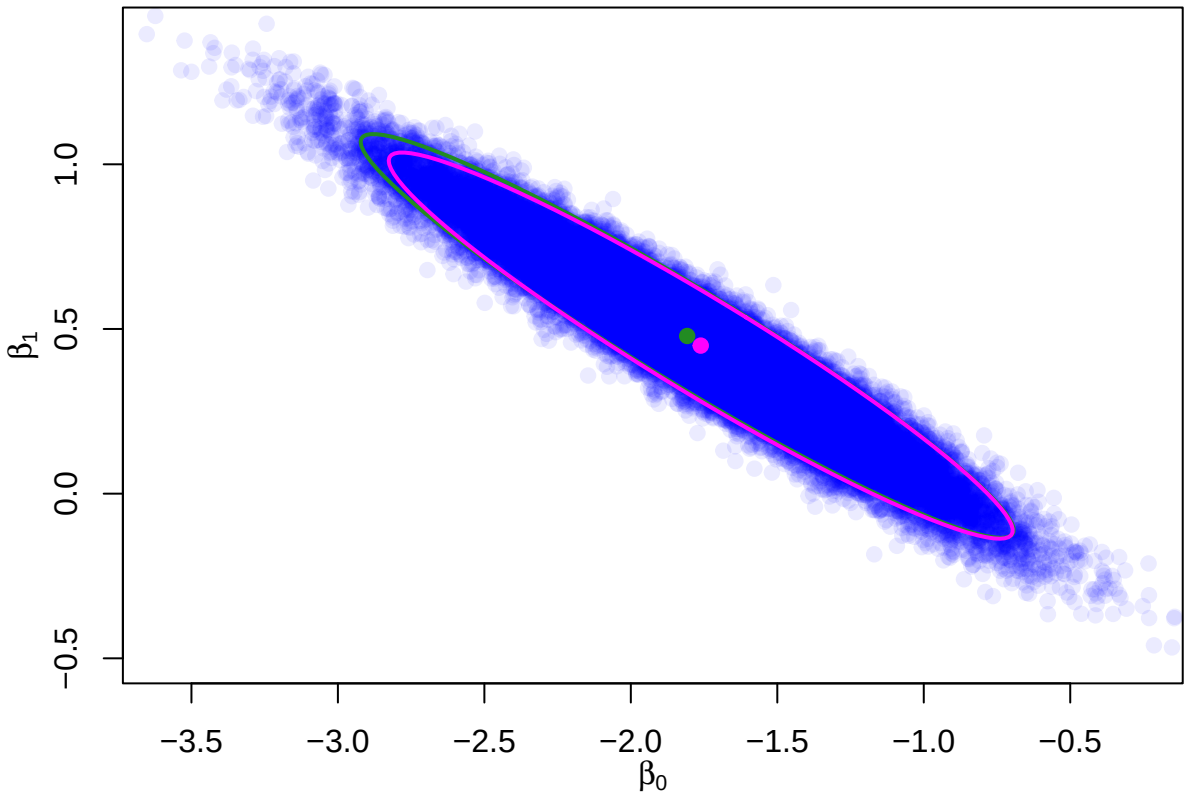
```
par(mar=c(3,3,1.5,1.5)+0.2)
plot(beta1~beta0, data=beta.post, pch=19, col="#0000FF15",
xlim=c(-3.6,-0.25), ylim=c(-0.5,1.4), xlab="", ylab="")
mtext(side=1, expression(beta[0]), line=2)
mtext(side=2, expression(beta[1]), line=2)

lines(ellipse(vcov(glm1), centre=coef(glm1)), col="forestgreen", lwd=2)
points(x=coef(glm1)[1], y=coef(glm1)[2], pch=19, col="forestgreen")
lines(ellipse(var(beta.post), centre=colMeans(beta.post)), col="magenta", lwd=2)
points(x=mean(beta.post$beta0), y=mean(beta.post$beta1), pch=19, col="magenta")
```

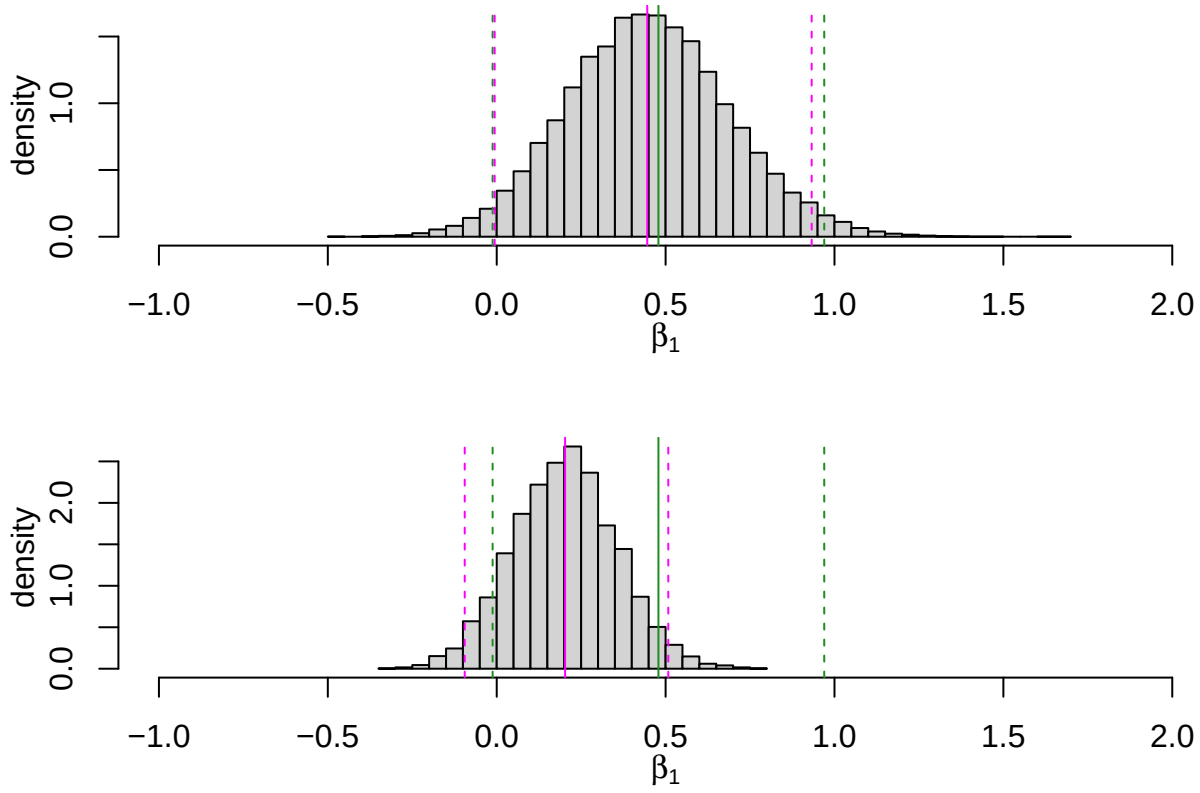


Logistic regression example: comparison of estimates

For the more-diffuse prior:

```
##             [,1]      [,2]
## beta0 -1.7613319 0.4351593
## beta1  0.4495068 0.2395200
##          50%         2.5%        97.5%
##  0.445647721 -0.006067766  0.932485055
```






```
## null device
##           1
```

We calculate $\Pr(\beta_1 > 0|y)$.

```
round(mean(beta.post[,2]>0),3)
## [1] 0.973
```

```
round(mean(beta.post.inf[,2]>0),3)
## [1] 0.905
```

Prediction

Predictions from a Binomial Distribution

We now consider prediction.

Assume $y|\theta \sim$ binomial(N, θ) and $\theta \sim$ 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 $z = 0, \ldots, M$.

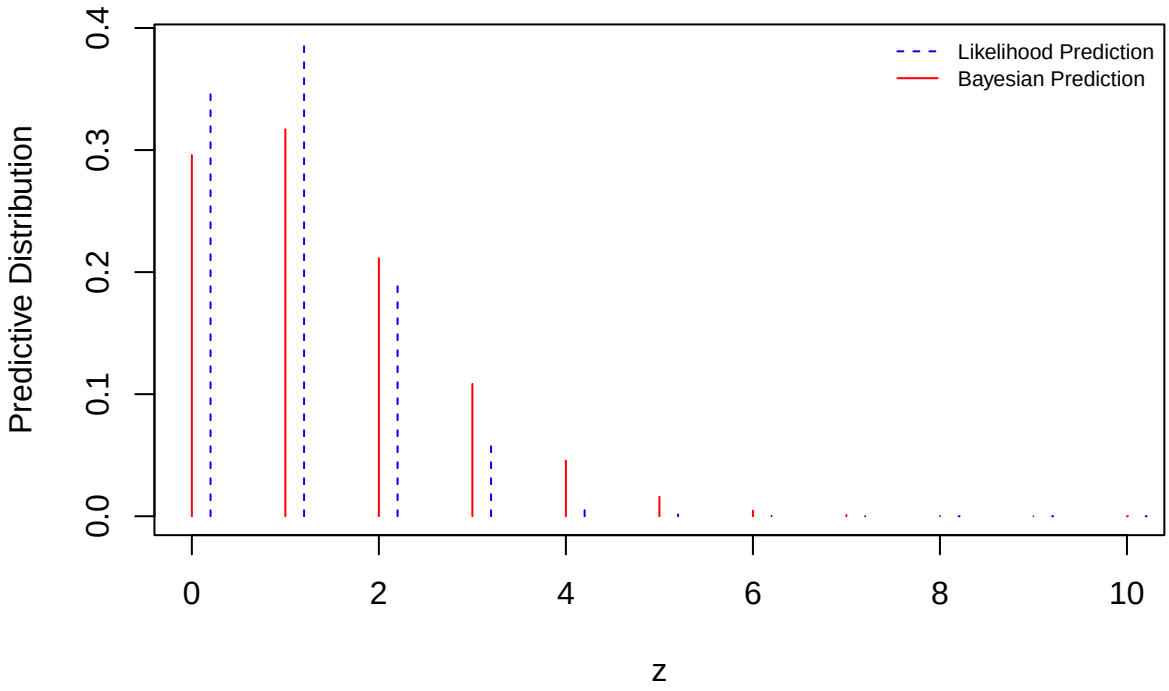
We demonstrate with a uniform prior and observing $y = 2$ successes from $N = 20$ trials, and suppose we wish to predict the number of successes we will see in 10 additional trials.

```
# 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
```

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
```

```
plot(binpred~z,type="h",col="red",ylim=c(0,max(likpred,binpred)),
     ylab="Predictive Distribution")
points(z+.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")
```



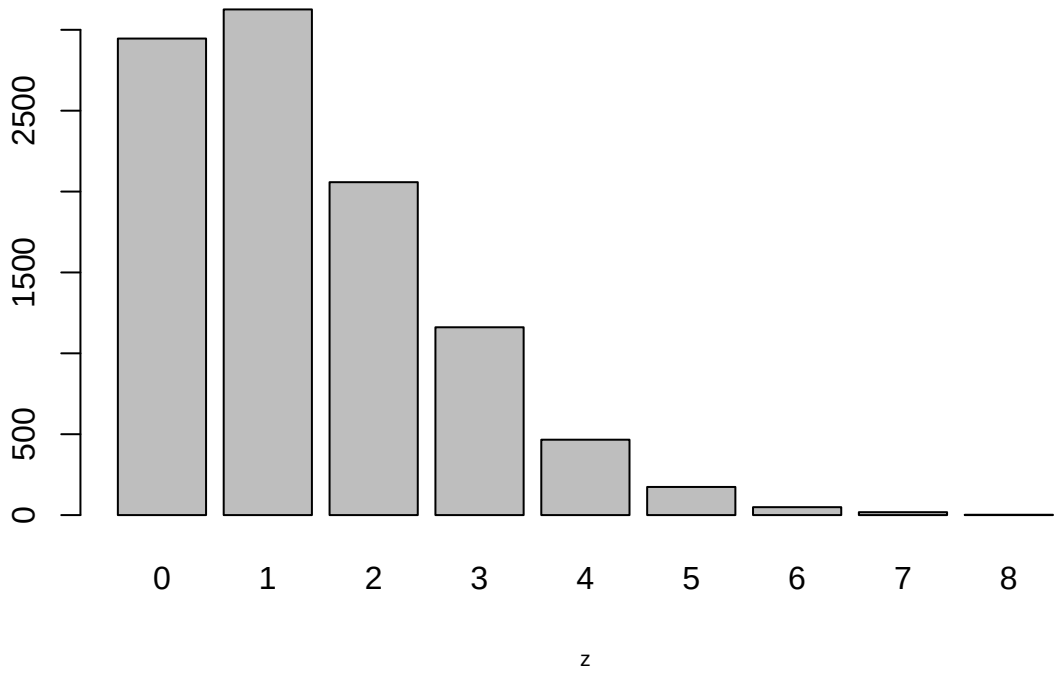
We now simulate directly via:

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

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

```
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])
}
```

```
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 LHON example, but use the less-diffuse prior on (only) the intercept β_0, not the log odds ratio β_1. How do the results compare to using the informative prior on β_1?