2020 SISG Bayesian Statistics for Genetics R Notes: Binomial Sampling 2

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

2020-07-19

Introduction

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

- specifying a prior distribution
- prediction and
- testing.

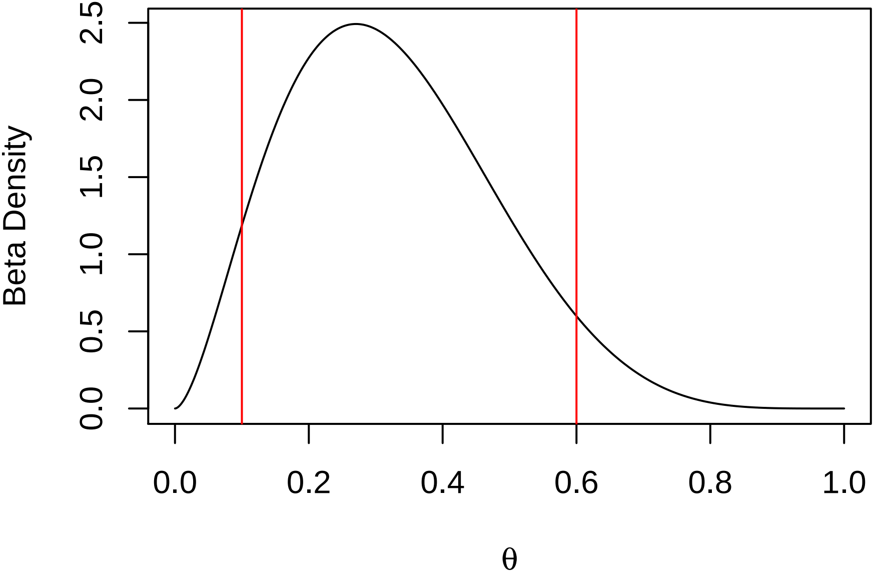
We also analyze allele specific expression (ASE) data

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.

```r
# 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$par, "\n")
## a and b are  2.730616 5.667462
probvals <- seq(0, 1, 0.001)
plot(probvals, dbeta(probvals, shape1 = opt$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



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 $z = 0, \ldots, M$.

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

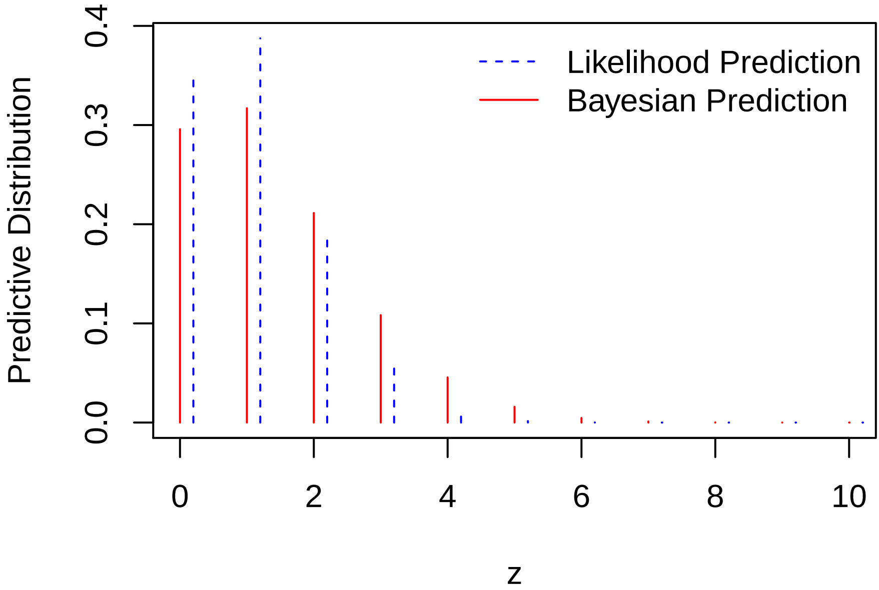
Along with the Bayesian predictive distribution, we also include a simple approach in which we assume simply take a `binomial(M,y/N)` distribution, i.e. assuming the probability is known to be the sample fraction.

```r
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

```r
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:

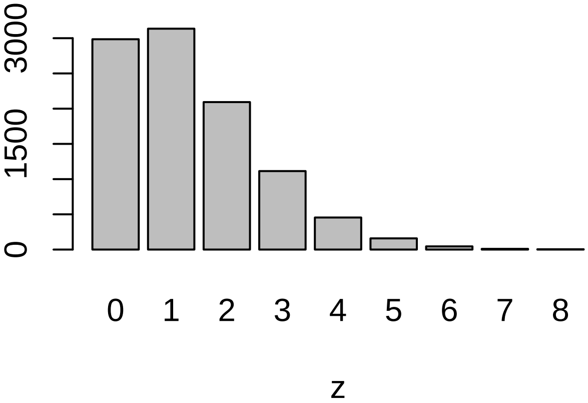
– Sampling from $\theta^{(s)} \sim p(\theta|y)$, $s = 1, \dots, S$.

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

```r
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")
```



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.

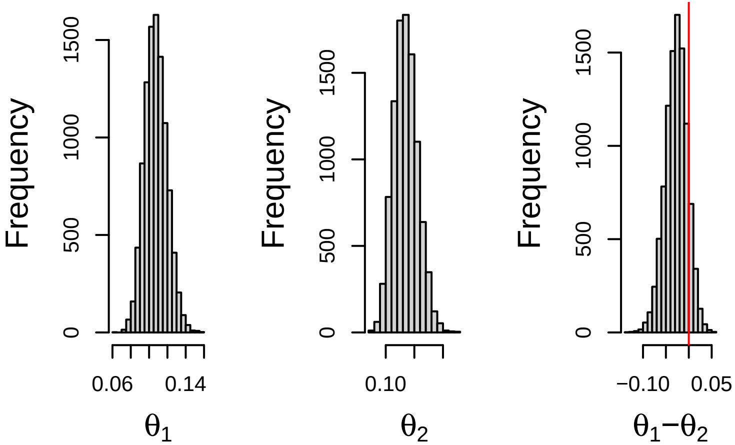
```
N1 <- 650
y1 <- floor(N1 * 0.1069)
N2 <- 265
y2 <- floor(N2 * 0.1321)
nsamp <- 10000
a <- b <- 1
post1 <- rbeta(nsamp, y1 + a, N1 - y1 + b)
post2 <- rbeta(nsamp, y2 + a, N2 - y2 + b)
```

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 = 1.5)
hist(post2, xlab = expression(theta[2]), main = "",
    cex.lab = 1.5)
#
hist(post1 - post2, xlab = expression(paste(theta[1],
    "-", theta[2])), main = "", cex.lab = 1.5)
abline(v = 0, col = "red")
sum(post1 - post2 > 0)/nsamp
## [1] 0.1217
```

Differences in Binomial Proportions



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

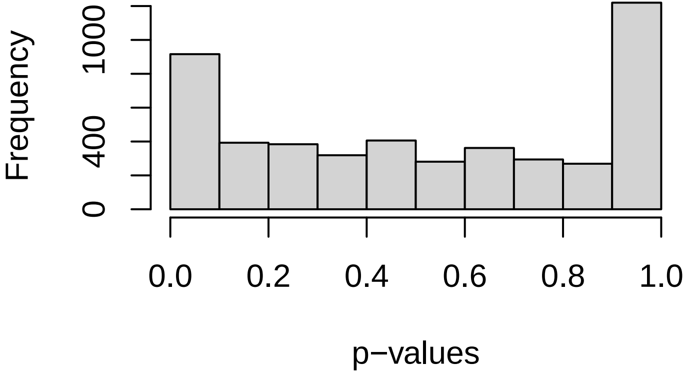
```r
# Function to evaluate Bayes factors for a binomial
# likelihood and beta prior, and a point null at p0
BFbinomial <- function(N, y, a, b, p0) {
    logPrH0 <- lchoose(N, y) + y * log(p0) + (N - y) *
        log(1 - p0)
    logPrH1 <- lchoose(N, y) + gamma(a + b) - lgamma(a) -
        lgamma(b)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

```
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
```

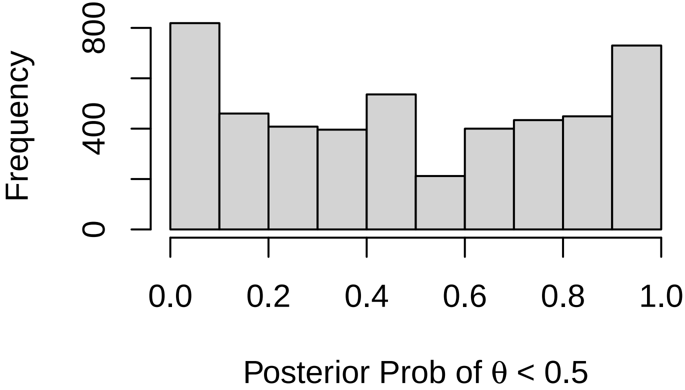
Histogram of *p*-values for ASE data

```r
hist(pvals, xlab = "p-values", main = "")
```

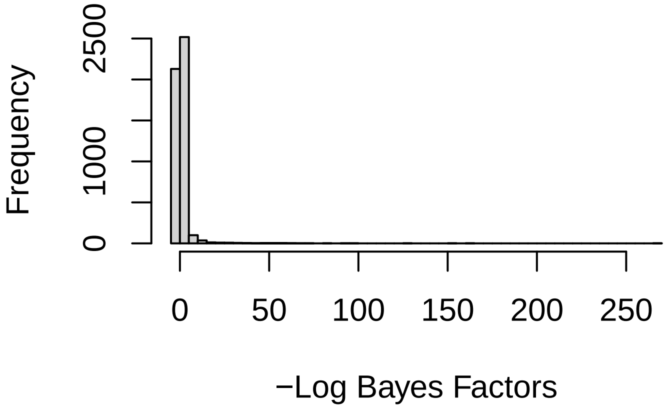


Histogram of posterior probabilities for ASE data

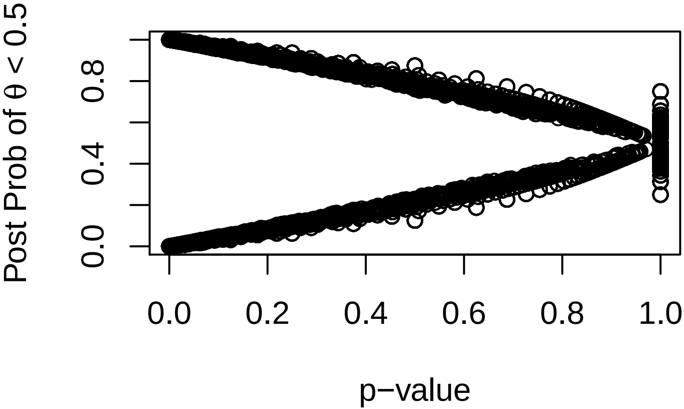
```
hist(postprob, xlab = expression(paste("Posterior Prob of ",
    theta, " < 0.5")), main = "")
```



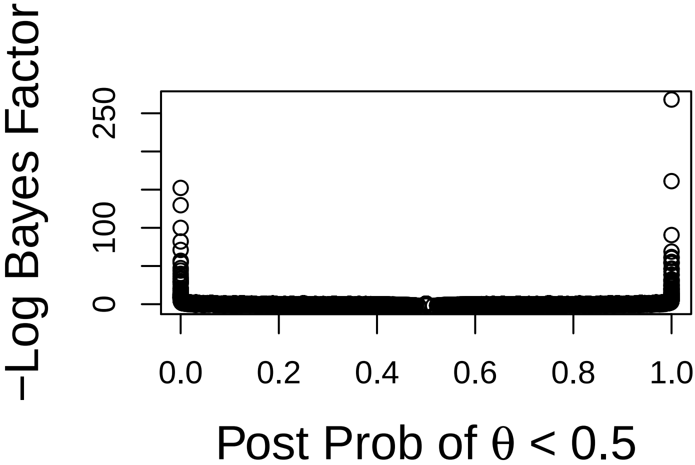
Analysis of ASE data



Analysis of ASE data



Analysis of ASE data



Exercises

- Redo the seroprevalence example from the first lecture with a Beta prior on the prevalence which has $\Pr(\theta < 0.01) = 0.05$ and $\Pr(\theta > 0.03) = 0.05$.

- Two populations are sampled to learn about the frequencies of a particular allele. The observed data are $N_1 = 100, y_1 = 30$ in population 1 and $N_2 = 150, y_2 = 60$ in population 2. Let the unobserved true frequencies in the two complete populations be θ_1 and θ_2.
  - With Beta(1,2) priors on θ_1 and θ_2 what are the Beta posterior distributions $\theta_1|y_1$ and $\theta_2|y_2$?
  - Obtain samples from the posteriors for θ_1 and θ_2 and find the posterior medians and 90% intervals for each.
  - Obtain samples for $\theta_1 - \theta_2|y$ and estimate the posterior probability that $\theta_1 > \theta_2$.

Exercises

- 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 $1/\text{BF} > 150$?