The Beta-Binomial Model

In the exchangeable binary case, we must have

$$p_{Y_i}(y_i;\theta) = \theta^{y_i}(1-\theta)^{1-y_i}$$

for each $0 \le \theta \le 1$; however different choices of $\pi_0(d\theta)$ lead to different models for the joint distribution of $Y_1, \ldots, Y_n$.

The Beta-Binomial Model

Suppose that

$$\pi_0(\theta) \equiv \mathit{Beta}(\alpha_0, \beta_0).$$

for $\alpha_0, \beta_0 > 0$. That is, the prior is *continuous*

$$\pi_0(\theta) = \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} \qquad 0 < \theta < 1.$$

and we are assuming that $0 < \theta < 1$ with probability 1.

The Beta-Binomial Model

This model presumes that

- $\theta \sim Beta(\alpha_0, \beta_0)$,
- $Y_1, \ldots, Y_n | \theta \sim Bernoulli(\theta)$,

with $Y_1, \ldots, Y_n$ conditionally independent. Let

$$S_n = \sum_{i=1}^{n} Y_i \qquad s_n = \sum_{i=1}^{n} y_i.$$

By a standard marginalization calculation, we may conclude that the prior predictive takes the form

$$
\begin{aligned}
p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n) &= \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i} \right\} \pi_0(\theta)\ d\theta \\
&= \int_0^1 \theta^{s_n} (1-\theta)^{n-s_n} \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \theta^{\alpha_0-1} (1-\theta)^{\beta_0-1}\ d\theta \\
&= \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \int_0^1 \theta^{s_n+\alpha_0-1} (1-\theta)^{n-s_n+\beta_0-1}\ d\theta \\
&= \frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_n)\Gamma(\beta_n)}{\Gamma(\alpha_n+\beta_n)}
\end{aligned}
$$

say.

This follows as the integrand is *proportional to a Beta pdf*, where

$$\alpha_n = s_n + \alpha_0 \qquad \beta_n = n - s_n + \beta_0$$

that is, in light of the data, the prior *hyperparameters* (α_0, β_0) are updated to (α_n, β_n) to define the prior predictive marginal model.

The Beta-Binomial Model

There are $n + 1$ distinct values for $p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$ defined by the possible values of s_n. We have for $s_n \in \{0, 1, \ldots, n\}$,

$$\Pr[S_n = s_n] = \binom{n}{s_n} p_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n)$$

where $(y_1, \ldots, y_n)$ is any binary sequence for which

$$\sum_{i=1}^{n} y_i = s_n.$$

This distribution for S_n is termed the *Beta-Binomial* model.

The Beta-Binomial Model

We have for the posterior density

$$\begin{aligned}\pi_n(\theta) &\propto \left\{\prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i}\right\}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \\ &= \theta^{s_n+\alpha_0-1}(1-\theta)^{n-s_n+\beta_0-1} \\ &= \theta^{\alpha_n-1}(1-\theta)^{\beta_n-1}\end{aligned}$$

Thus the posterior distribution is *also* a Beta distribution.

That is,

$$\pi_n(\theta) \equiv Beta(s_n + \alpha_0, n - s_n + \beta_0) \equiv Beta(\alpha_n, \beta_n)$$

say, so that

$$\pi_n(\theta) = \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \theta^{\alpha_n - 1}(1 - \theta)^{\beta_n - 1}.$$

The Beta-Binomial Model

Simulation: $n = 8$, $\alpha_0 = 2$, $\beta_0 = 1.5$.

```
set.seed(213)
n<-8
sim.exch01<-function(nv,a0v,b0v){  #Sample the exchangeable variables.
    thv<-rbeta(1,a0v,b0v)
    Yv<-rbinom(nv,1,thv)
}
al0<-2.0
be0<-1.5
nreps<-10000
Ymat<-t(replicate(nreps,sim.exch01(n,al0,be0)))  #2000 replicate draws of Y
```

We have that for $i = 1, \ldots, n$, by iterated expectation

$$\mathbb{E}_{Y_i}[Y_i] = \mathbb{E}_{\pi_0}[\mathbb{E}_{Y_i|\theta}[Y_i|\theta]] = \mathbb{E}_{\pi_0}[\theta] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

and

$$\mathbb{E}_{Y_i}[Y_i^2] = \mathbb{E}_{\pi_0}[\mathbb{E}_{Y_i|\theta}[Y_i^2|\theta]] = \mathbb{E}_{\pi_0}[\mathbb{E}_{Y_i|\theta}[Y_i|\theta]] = \mathbb{E}_{\pi_0}[\theta] = \frac{\alpha_0}{\alpha_0 + \beta_0}$$

so

$$\mathbb{V}ar_{Y_i}[Y_i] = \frac{\alpha_0}{\alpha_0 + \beta_0} - \left(\frac{\alpha_0}{\alpha_0 + \beta_0}\right)^2.$$

The Beta-Binomial Model

```
muBetaBinomY<-al0/(al0+be0)
sigBetaBinomY<-sqrt(muBetaBinomY*(1-muBetaBinomY))
c(muBetaBinomY,sigBetaBinomY^2)

+ [1] 0.5714286 0.2448980

apply(Ymat,2,mean)

+ [1] 0.5673 0.5716 0.5723 0.5701 0.5749 0.5700 0.5674 0.5773

apply(Ymat,2,var)

+ [1] 0.2454953 0.2448979 0.2447972 0.2451105 0.2444144 0.2451245 0.2454818
+ [8] 0.2440491
```

The Beta-Binomial Model

For the covariances, using iterated expectation we have

$$\mathbb{E}_{Y_i,Y_j}[Y_iY_j] = \mathbb{E}_{\pi_0}\left[\mathbb{E}_{Y_i,Y_j|\theta}[Y_iY_j|\theta]\right]$$

$$= \mathbb{E}_{\pi_0}\left[\mathbb{E}_{Y_i|\theta}[Y_i|\theta]\mathbb{E}_{Y_j|\theta}[Y_j|\theta]\right]$$

as Y_i and Y_j are conditionally independent given θ.

Thus, as $\mathbb{E}_{Y_i|\theta}[Y_i|\theta] = \theta$ for each i, we have

$$\mathbb{E}_{Y_i,Y_j}[Y_i Y_j] = \mathbb{E}_{\pi_0}[\theta^2] = \mathbb{V}ar_{\pi_0}[\theta] + \{\mathbb{E}_{\pi_0}[\theta]\}^2$$

which, by properties of the Beta distribution, is equal to

$$\frac{\alpha_0\beta_0}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)} + \left(\frac{\alpha_0}{\alpha_0+\beta_0}\right)^2$$

The Beta-Binomial Model

Hence, the covariance is

$$\mathbb{C}ov_{Y_i,Y_j}[Y_i, Y_j] = \frac{\alpha_0\beta_0}{(\alpha_0+\beta_0)^2(\alpha_0+\beta_0+1)}$$

and the correlation

$$\frac{\mathbb{C}ov_{Y_i,Y_j}[Y_i, Y_j]}{\mathbb{V}ar_{Y_i}[Y_i]}$$

is computed using the previous variance expression.

```
covBetaBinomY<-al0*be0/((al0+be0)^2*(al0+be0+1))
corBetaBinomY<-covBetaBinomY/sigBetaBinomY^2
corBetaBinomY

+ [1] 0.2222222

round(cor(Ymat),3)

+       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
+ [1,] 1.000 0.221 0.227 0.245 0.225 0.242 0.233 0.224
+ [2,] 0.221 1.000 0.221 0.236 0.206 0.227 0.220 0.220
+ [3,] 0.227 0.221 1.000 0.228 0.223 0.225 0.213 0.227
+ [4,] 0.245 0.236 0.228 1.000 0.231 0.229 0.227 0.221
+ [5,] 0.225 0.206 0.223 0.231 1.000 0.224 0.218 0.228
+ [6,] 0.242 0.227 0.225 0.229 0.224 1.000 0.220 0.223
+ [7,] 0.233 0.220 0.213 0.227 0.218 0.220 1.000 0.217
+ [8,] 0.224 0.220 0.227 0.221 0.228 0.223 0.217 1.000
```

The Beta-Binomial Model

Note: the presumed *data generating model* here is the exchangeable form given by

$$p_{Y_1,\ldots,Y_n}(y_1,\ldots,y_n)$$

We can generate Ys from this model directly by using the Beta-Binomial model functions in `R`.

We generate using the `extraDistr` package

- S_n using the function `rbbinom`
- $Y_1,\ldots,Y_n|S_n = s_n$ by selecting uniformly from the set of binary sequences which have s_n ones.

The function `dbbinom` computes the Beta-Binomial pmf

$$\Pr[S_n = s_n] = \binom{n}{s_n} \frac{\Gamma(\alpha_0 + \beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)} \frac{\Gamma(\alpha_0 + s_n)\Gamma(\beta_0 + n - s_n)}{\Gamma(\alpha_0 + \beta_0 + n)}$$

for $s_n \in \{0, 1, \ldots, n\}$.

- n is the *size* parameter,
- (α_0, β_0) are the *shape* parameters.

The Beta-Binomial Model

```
library(extraDistr)
Sn<-rbbinom(nreps,size=n,alpha=al0,beta=be0)
table(Sn)/nreps

+ Sn
+      0      1      2      3      4      5      6      7      8
+ 0.0351 0.0689 0.0937 0.1203 0.1431 0.1464 0.1499 0.1393 0.1033

round(dbbinom(0:n,size=n,alpha=al0,beta=be0),4)

+ [1] 0.0376 0.0708 0.0991 0.1219 0.1386 0.1478 0.1478 0.1351 0.1013
```

The Beta-Binomial Model

```
Ymat2<-matrix(0,nrow=nreps,ncol=n)
for(i in 1:nreps){
    Ymat2[i,sample(1:n,size=Sn[i],rep=FALSE)]<-1
}
apply(Ymat2,2,mean)

+ [1] 0.5728 0.5760 0.5797 0.5823 0.5707 0.5808 0.5809 0.5793

apply(Ymat2,2,var)

+ [1] 0.2447246 0.2442484 0.2436723 0.2432510 0.2450260 0.2434957 0.2434795
+ [8] 0.2437359

round(cor(Ymat2),3)

+       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
+ [1,] 1.000 0.214 0.236 0.231 0.215 0.224 0.216 0.231
+ [2,] 0.214 1.000 0.223 0.214 0.223 0.233 0.214 0.216
+ [3,] 0.236 0.223 1.000 0.213 0.212 0.230 0.207 0.222
+ [4,] 0.231 0.214 0.213 1.000 0.225 0.205 0.199 0.213
+ [5,] 0.215 0.223 0.212 0.225 1.000 0.228 0.209 0.226
+ [6,] 0.224 0.233 0.230 0.205 0.228 1.000 0.232 0.206
+ [7,] 0.216 0.214 0.207 0.199 0.209 0.232 1.000 0.213
+ [8,] 0.231 0.216 0.222 0.213 0.226 0.206 0.213 1.000
```

The Beta-Binomial Model

Let

$$R_n = \frac{S_n}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Then for $r \in \{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$

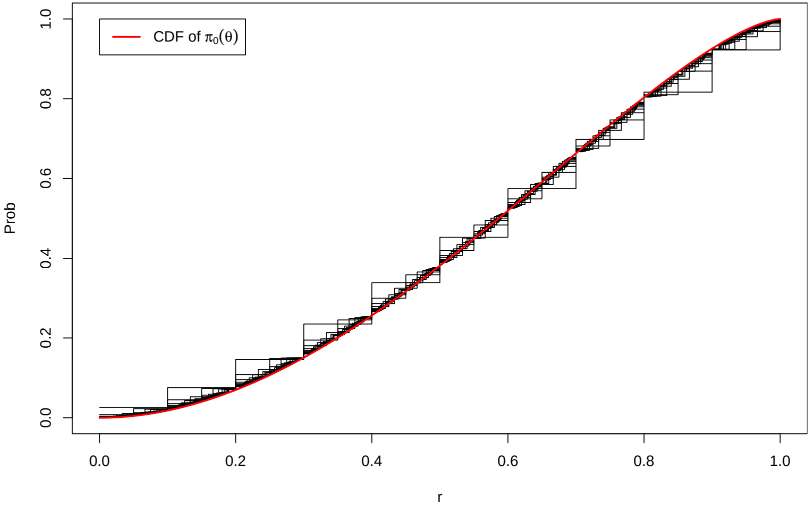
$$\begin{aligned}\Pr[R_n = r] &\equiv \Pr[S_n = nr] \\ &= \binom{n}{nr}\frac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\frac{\Gamma(\alpha_0+nr)\Gamma(\beta_0+n-nr)}{\Gamma(\alpha_0+\beta_0+n)}\end{aligned}$$

We can compute this distribution for $n = 10, 20, \ldots, 100$, and plot the cdf.

The Beta-Binomial Model

```
par(mar=c(4,4,2,0))
plot(c(0,1),c(0,1),type='n',ylab='Prob',xlab='r')
for(j in 1:10){
    n<-10*j
    rv<-c(0:n)/n
    pv<-pbbinom(0:n,size=n,alpha=al0,beta=be0)
    lines(rv,pv,type='s')
}
xv<-seq(0,1,length=1000)
lines(xv,pbeta(xv,al0,be0),col='red',lwd=2)
legend(0,1,c(expression(paste('CDF of ',pi[0](theta)))),col='red',lty=1,lwd=2)
```

The Beta-Binomial Model



The Beta-Binomial Model

For the posterior predictive

$$p_{Y_{n+1},\dots,Y_{n+m}|Y_1,\dots,Y_n}\left(y_{n+1},\dots,y_{n+m}|y_1,\dots,y_n\right)$$

$$= \frac{p_{Y_1,\dots,Y_{n+m}}\left(y_1,\dots,y_{n+m}\right)}{p_{Y_1,\dots,Y_n}\left(y_1,\dots,y_n\right)}$$

$$= \frac{\dfrac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\dfrac{\Gamma(\alpha_{n+m})\Gamma(\beta_{n+m})}{\Gamma(\alpha_{n+m}+\beta_{n+m})}}{\dfrac{\Gamma(\alpha_0+\beta_0)}{\Gamma(\alpha_0)\Gamma(\beta_0)}\dfrac{\Gamma(\alpha_n)\Gamma(\beta_n)}{\Gamma(\alpha_n+\beta_n)}}$$

where

$$\alpha_{n+m} = s_{n+m} + \alpha_0 \qquad \beta_{n+m} = (n+m) - s_{n+m} + \beta_0$$

After cancellation and rearrangement, we have that

$$p_{Y_{n+1},\dots,Y_{n+m}|Y_1,\dots,Y_n}(y_{n+1},\dots,y_{n+m}|y_1,\dots,y_n)$$

$$= \frac{\Gamma(\alpha_n+\beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)}\frac{\Gamma(\alpha_{n+m})\Gamma(\beta_{n+m})}{\Gamma(\alpha_{n+m}+\beta_{n+m})}$$

$$= \frac{\Gamma(\alpha_n+\beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)}\frac{\Gamma(\alpha_n+s_{n+1,n+m})\Gamma(\beta_n+m-s_{n+1,n+m})}{\Gamma(\alpha_n+\beta_n+m)}$$

where

$$s_{n+1,n+m} = \sum_{i=n+1}^{n+m} y_i.$$

If we examine the distribution of the random variable

$$S_{n+1,n+m} = \sum_{i=n+1}^{n+m} Y_i$$

conditional on $Y_1, \ldots, Y_n$, we see that

$$\begin{aligned}\Pr[S_{n+1,n+m} &= s | Y_1 = y_n, \ldots, Y_n = y_n] \\ &= \binom{m}{s} \frac{\Gamma(\alpha_n + \beta_n)}{\Gamma(\alpha_n)\Gamma(\beta_n)} \frac{\Gamma(\alpha_n + s)\Gamma(\beta_n + m - s)}{\Gamma(\alpha_n + \beta_n + m)}\end{aligned}$$

which we identify as the Beta-Binomial distribution with size m and shape parameters (α_n, β_n).

The Beta-Binomial Model

We can study the conditional distribution of

$$R_{n+1,n+m} = \frac{S_{n+1,n+m}}{m} = \frac{1}{m}\sum_{i=n+1}^{n+m} Y_i$$

given $Y_1 = y_1, \ldots, Y_n = y_n$ as $m = 10, 20, \ldots, 100$ for any fixed $(y_1, \ldots, y_n)$.

Note that α_n and β_n only depend on s_n, so we may simulate that directly using the `rbbinom` function.

The Beta-Binomial Model

```
#Data
n<-10
Sn<-rbbinom(1,size=n,alpha=al0,beta=be0)
aln<-al0+Sn
ben<-be0+n-Sn
print(Sn)

+ [1] 7

par(mar=c(4,4,2,0))
plot(c(0,1),c(0,1),type='n',ylab='Prob',xlab='r')
for(j in 1:10){
    m<-10*j
    rv<-c(0:m)/m
    pv<-pbbinom(0:m,size=m,alpha=aln,beta=ben)
    lines(rv,pv,type='s')
}
xv<-seq(0,1,length=1000)
lines(xv,pbeta(xv,aln,ben),col='red',lwd=2)
legend(0,1,c(expression(paste('CDF of ',pi[n](theta)))),col='red',lty=1,lwd=2)
```

The Beta-Binomial Model