A Two Sample Problem

Suppose we have two infinite sequences

$$\{\mathbf{Y}_1\} = Y_{11}, Y_{12}, \ldots, Y_{1n}, \ldots$$
$$\{\mathbf{Y}_2\} = Y_{21}, Y_{22}, \ldots, Y_{2n}, \ldots$$

For two finite collections of sizes n_1 and n_2, we have the de Finetti representation

$$f_{\mathbf{Y}_1,\mathbf{Y}_2}(\mathbf{y}_1, \mathbf{y}_2) = \int\int \prod_{i=1}^{n_1} f_1(y_{1i}; \theta_1) \prod_{i=1}^{n_2} f_2(y_{2i}; \theta_2) \pi_0(d\theta_1, d\theta_2)$$

A Two Sample Problem

We can make different assumptions about a potential joint structure.

(a) exchangeability *within* each sequence and *independence* across sequences

$$\pi_0(d\theta_1, d\theta_2) = \pi_0(d\theta_1)\pi_0(d\theta_2)$$

that is, θ_1 and θ_2 are presumed independent *a priori*.

A Two Sample Problem

(b) exchangeability *within* each sequence

$$\pi_0(d\theta_1, d\theta_2)$$

has a general joint form; this is an example of *partial exchangeability*, based on the sequence labels 1 and 2.

A Two Sample Problem

(c) *complete* exchangeability amongst all observables

$$\pi_0(d\theta_1, d\theta_2)$$

is singular, and concentrates on the line $\theta_1 = \theta_2$. That is, there is in fact only one parameter

$$\theta = \theta_1 = \theta_2$$

and a prior π_0 on this parameter.

A Two Sample Problem

We focus on the case (a). In the binary observable case.

$$p_{Y_{ji}}(y;\theta_j) = \theta_j^y(1-\theta_j)^{1-y} \qquad j = 1,2$$

for each $0 \leq \theta \leq 1$. Suppose that, independently,

$$\pi_0(\theta_j) \equiv Beta(\alpha_0, \beta_0).$$

for $\alpha_0, \beta_0 > 0$.

A Two Sample Problem

We have for the posterior densities for $j = 1, 2$.

$$\pi_n(\theta_j) \equiv Beta(s_{jn} + \alpha_0, n_j - s_{jn} + \beta_0) \equiv Beta(\alpha_{jn}, \beta_{jn})$$

where

$$s_{1n} = \sum_{i=1}^{n_1} y_{1i} \qquad s_{2n} = \sum_{i=1}^{n_2} y_{2i}$$

A Two Sample Problem

Simulation: $n_1 = 8, n_2 = 12, \alpha_0 = 2, \beta_0 = 1.5$. Suppose that, in reality the true values of the parameters are

$$\theta_{10} = 0.5 \qquad \theta_{20} = 0.7$$

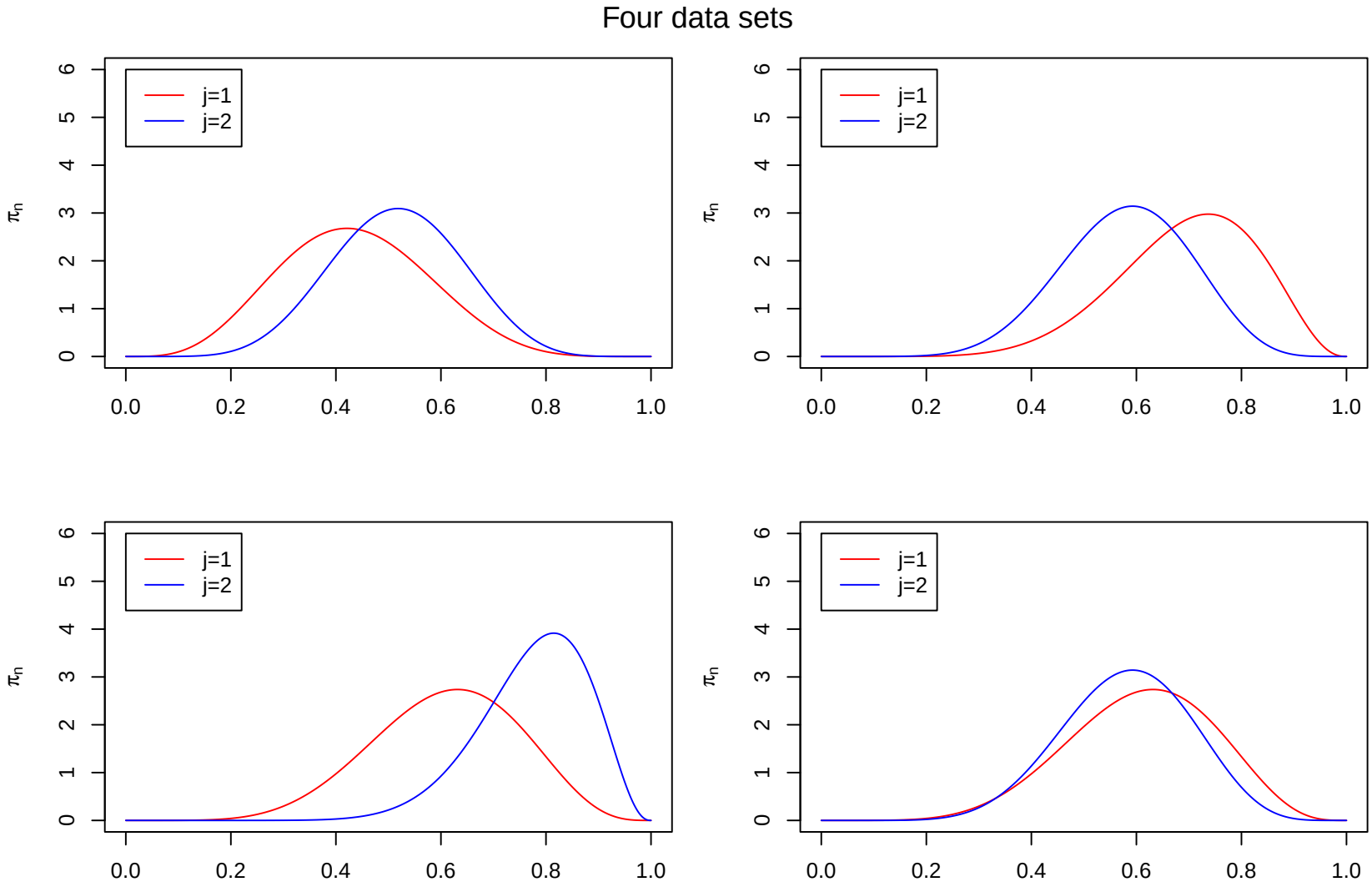
which are in the support of the prior.

```
set.seed(213)
al0<-2.0
be0<-1.5
n1<-8
n2<-12
nreps<-4
th10<-0.5
th20<-0.7
ymat1<-t(replicate(nreps,rbinom(n1,1,th10)))
ymat2<-t(replicate(nreps,rbinom(n2,1,th20)))
```

A Two Sample Problem

```
#Data
s1<-apply(ymat1,1,sum)
s2<-apply(ymat2,1,sum)
al.n1<-al0+s1
be.n1<-be0+n1-s1
al.n2<-al0+s2
be.n2<-be0+n2-s2
xv<-seq(0,1,length=1001)
par(mar=c(4,4,2,1),mfrow=c(2,2))
for(irep in 1:4){
    plot(xv,dbeta(xv,al.n1[irep],be.n1[irep]),type='l',col='red',
        ylim=range(0,6),xlab='',ylab=expression(pi[n]))
    lines(xv,dbeta(xv,al.n2[irep],be.n2[irep]),type='l',col='blue')
    legend(0,6,c('j=1','j=2'),col=c('red','blue'),lty=1)
}
mtext("Four data sets", side = 3, line = -1, cex=1.15, outer = TRUE)
```

A Two Sample Problem



A Two Sample Problem

We can report

1. Parameter Estimates:
   - Posterior Mean: $\alpha_n/(\alpha_n + \beta_n)$
   - Posterior Median: 50% quantile of $\pi_n(\theta)$
   - Posterior Mode: $(\alpha_n - 1)/(\alpha_n + \beta_n - 2)$

θ_1	1	2	3	4
Mean	0.4348	0.6957	0.6087	0.6087
Median	0.4309	0.7073	0.6152	0.6152
Mode	0.4211	0.7368	0.6316	0.6316

θ_2	1	2	3	4
Mean	0.5161	0.5806	0.7742	0.5806
Median	0.5168	0.5842	0.7862	0.5842
Mode	0.5185	0.5926	0.8148	0.5926

A Two Sample Problem

2. Posterior Credible Intervals: For $0 < \gamma < 1$, find $\mathcal{C}_\gamma$ such that

$$\int_{\mathcal{C}_\gamma} \pi_n(\theta)\, d\theta = \gamma$$

which records a region where θ is inferred to lie with posterior probability γ.

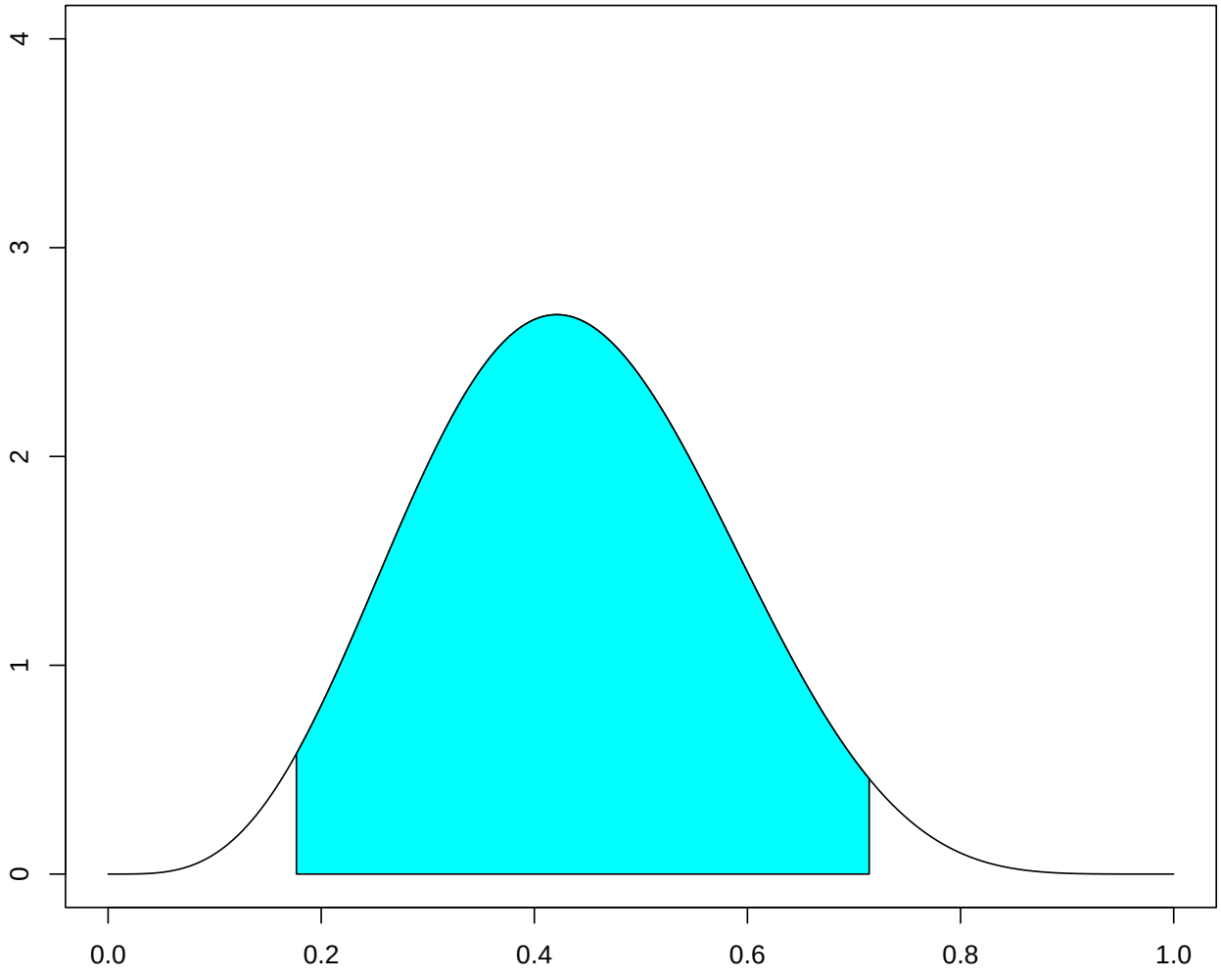
Typically we choose $\gamma = 0.95$. There are several ways to construct such an interval

A Two Sample Problem

Equal tail probability interval: find the $1-\gamma/2$ and $1-(1-\gamma/2)$ quantiles of the posterior.

```
gam<-0.95
par(mar=c(4,2,2,1))
plot(xv,dbeta(xv,al.n1[1],be.n1[1]),type='l',
        ylim=range(0,4),xlab='',ylab=expression(pi[n]))
eti.l<-qbeta((1-gam)/2,al.n1[1],be.n1[1])     #Lower tail quantile
eti.u<-qbeta(1-(1-gam)/2,al.n1[1],be.n1[1])   #Upper tail quantile
eti.x<-seq(eti.l,eti.u,length=1001)
eti.y<-dbeta(eti.x,al.n1[1],be.n1[1])
polygon(c(eti.x,eti.u,rev(eti.x),eti.l),
        c(rep(0,length(eti.x)),eti.y[length(eti.y)],rev(eti.y),0),
        col='cyan')
```

A Two Sample Problem



A Two Sample Problem

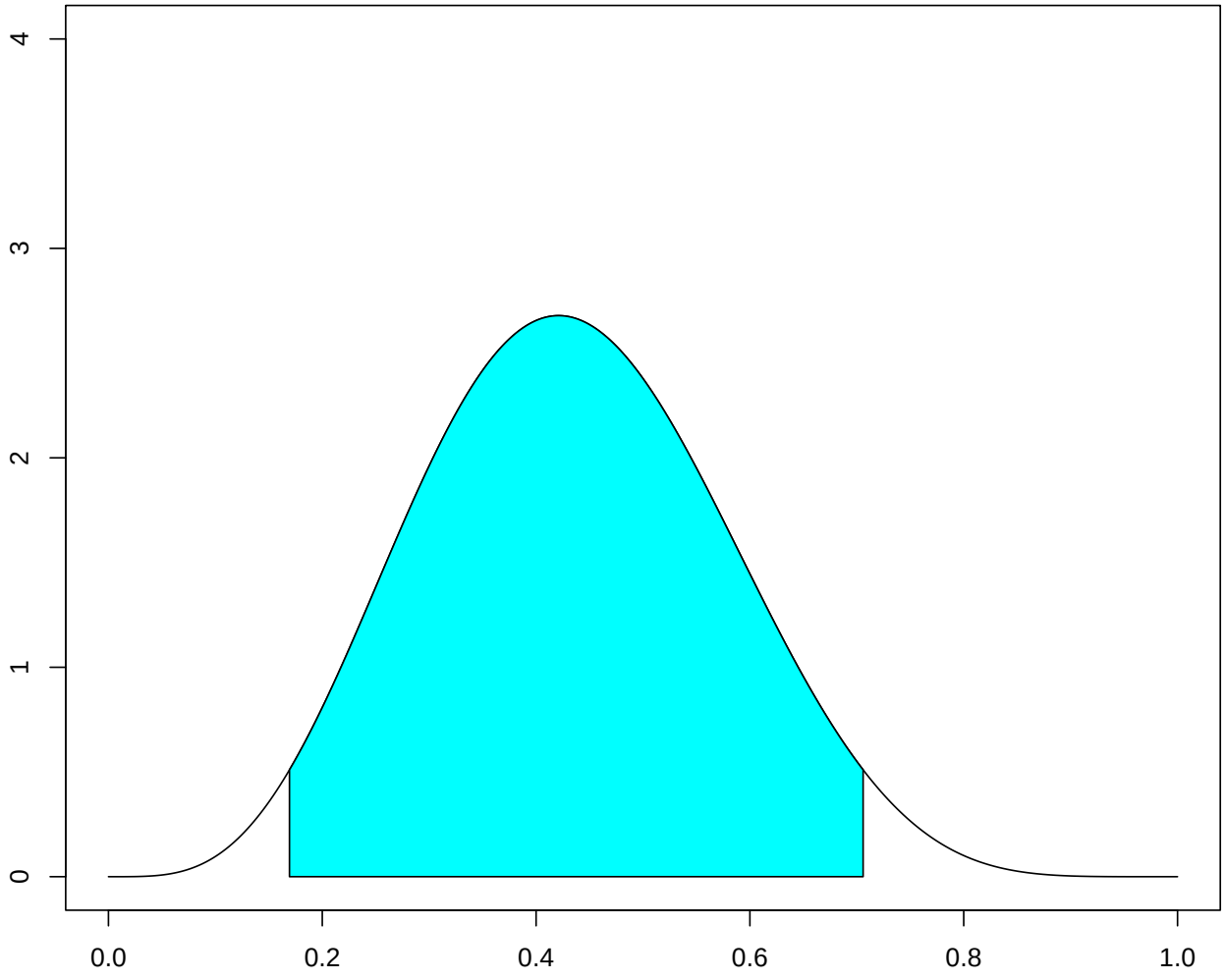
Highest posterior density (HPD) interval: find t_L and t_U such that $\pi_n(t_L) = \pi_n(t_U)$ with

$$\int_{t_L}^{t_U} \pi_n(\theta)\, d\theta = \gamma$$

```
library(HDInterval)
par(mar=c(4,2,2,1))
xv<-seq(0,1,length=1001)
yv<-dbeta(xv,al.n1[1],be.n1[1])

hpd.int<-hdi(qbeta, 0.95, shape1=al.n1[1], shape2=be.n1[1])
hpd.int
+     lower     upper
+ 0.1693075 0.7059122
+ attr(,"credMass")
+ [1] 0.95
plot(xv,yv,type='l',ylim=range(0,4),xlab='',ylab=expression(pi[n]))
hpd.l<-hpd.int[1]     #Lower tail value
hpd.u<-hpd.int[2]     #Upper tail value
hpd.x<-seq(hpd.l,hpd.u,length=1001)
hpd.y<-dbeta(hpd.x,al.n1[1],be.n1[1])
polygon(c(hpd.x,hpd.u,rev(hpd.x),hpd.l),
        c(rep(0,length(hpd.x)),hpd.y[length(hpd.y)],rev(hpd.y),0),
        col='cyan')
```

A Two Sample Problem



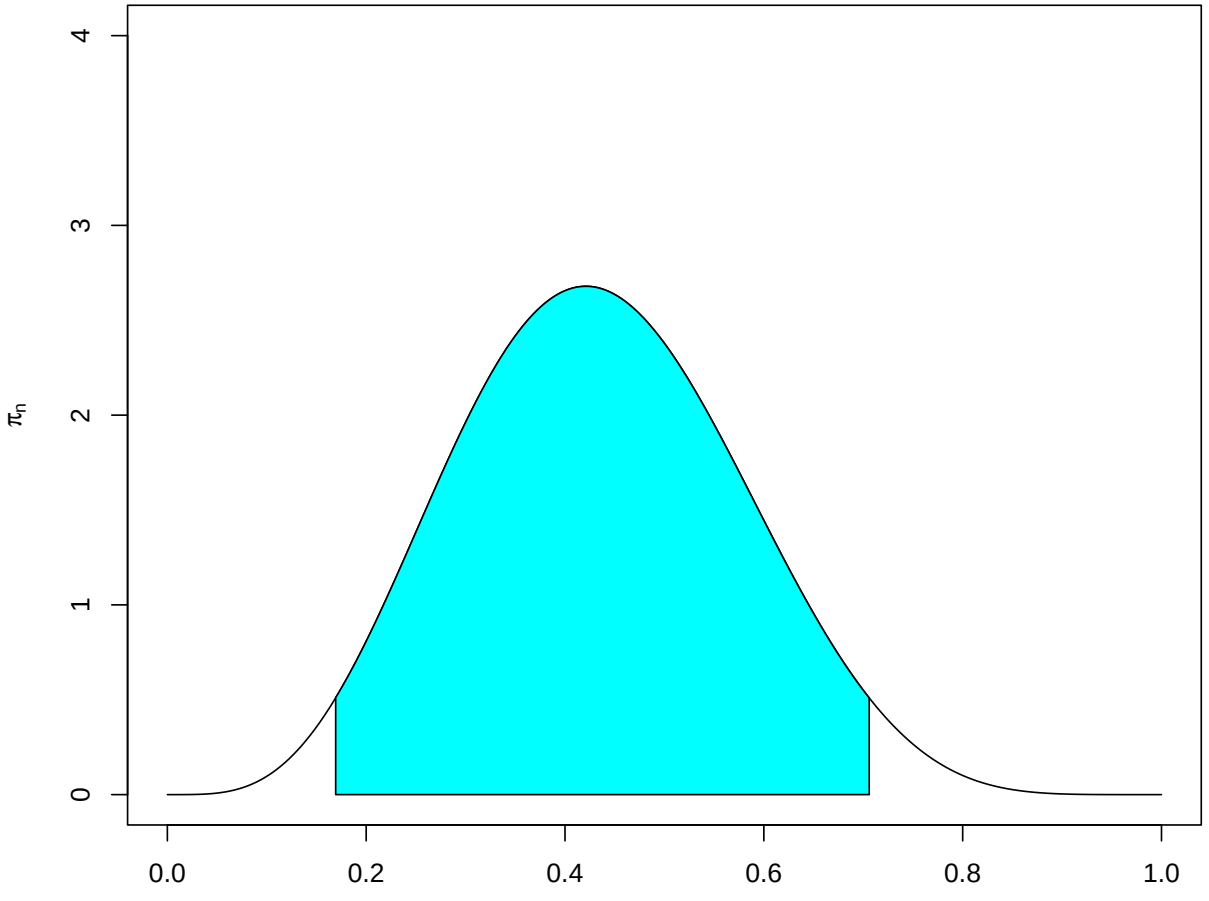
A Two Sample Problem

Shortest interval:

```
lowp<-seq(0.001,1-gam-0.001,by=0.001)
lowq<-qbeta(lowp,al.n1[1],be.n1[1])
upq<-qbeta(lowp+gam,al.n1[1],be.n1[1])
minval<-which.min(upq-lowq)

plot(xv,yv,type='l',ylim=range(0,4),xlab='',ylab=expression(pi[n]))
spd.l<-lowq[minval]     #Lower tail value
spd.u<-upq[minval]      #Upper tail value
spd.x<-seq(spd.l,spd.u,length=1001)
spd.y<-dbeta(spd.x,al.n1[1],be.n1[1])
polygon(c(spd.x,spd.u,rev(hpd.x),hpd.l),
        c(rep(0,length(spd.x)),spd.y[length(hpd.y)],rev(spd.y),0),
        col='cyan')
```

A Two Sample Problem



A Two Sample Problem

```
#Equal-tailed
c(eti.l,eti.u)                    #Ends of interval
+ [1] 0.1767363 0.7142649
pbeta(eti.l,al.n1[1],be.n1[1]) #Left-tail probability
+ [1] 0.025
#Highest posterior density
c(hpd.l,hpd.u)                    #Ends of interval
+     lower     upper
+ 0.1693075 0.7059122
pbeta(hpd.l,al.n1[1],be.n1[1]) #Left-tail probability
+      lower
+ 0.02096575
#Shortest
c(spd.l,spd.u)                    #Ends of interval
+ [1] 0.1693746 0.7059793
pbeta(spd.l,al.n1[1],be.n1[1]) #Left-tail probability
+ [1] 0.021
```

A Two Sample Problem

We may also examine parameters that *compare* the two samples:

- Difference
$$\delta = \theta_2 - \theta_1$$
- Ratio
$$\lambda = \frac{\theta_2}{\theta_1}$$
- Odds Ratio
$$\phi = \frac{\theta_2/(1-\theta_2)}{\theta_1/(1-\theta_1)}$$

In principle the posteriors for each of these derived parameters can be computed directly from

$$\pi_n(\theta_1) \qquad \pi_n(\theta_2)$$

using standard transformation methods.

A Two Sample Problem

For example, for $t > 0$,

$$\Pr[\lambda \le t] = \iint_{\mathcal{A}_t} \pi_n(\theta_1)\pi_n(\theta_2)\, d\theta_2\, d\theta_1$$

where

$$\mathcal{A}_t = \{(x_1, x_2) : 0 < x_1, x_2 < 1, x_2 \le t x_1\}.$$

that is

$$\Pr[\lambda \le t] = \int_0^1 \int_0^{t\theta_1} \pi_n(\theta_1)\pi_n(\theta_2)\, d\theta_2\, d\theta_1$$

This can be computed numerically.

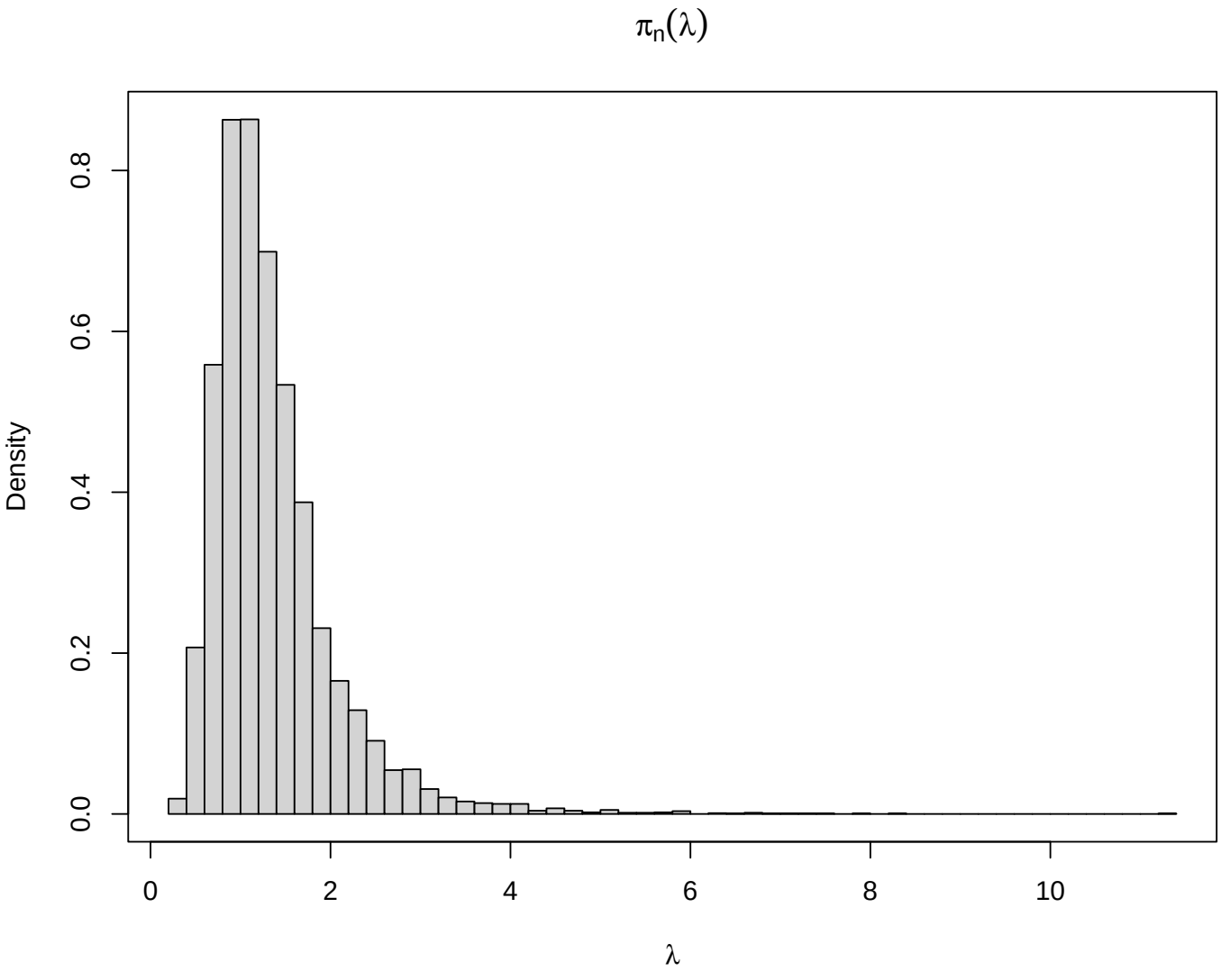
A Two Sample Problem

However, it is very straightforward to use a *sampling* approach:

- sample 10000 values from $\pi_n(\theta_1)$
- sample 10000 values from $\pi_n(\theta_2)$
- transform samples to obtain samples of λ

```
nsamp<-10000
th1.samp<-rbeta(nsamp,al.n1[1],be.n1[1])
th2.samp<-rbeta(nsamp,al.n2[1],be.n2[1])
lam.samp<-th2.samp/th1.samp
hist(lam.samp,freq=FALSE,nclass=50,xlab=expression(lambda),
        main=expression(pi[n](lambda)))
box()
```

A Two Sample Problem



A Two Sample Problem

We can compute a HPD credible interval from the sample using the `hdi` function:

```
hdi(lam.samp, 0.95)

+     lower     upper
+ 0.4239076 2.6913475
+ attr(,"credMass")
+ [1] 0.95
```