

SISG Bayes: Exercise 4

Ken Rice

2024-06-03

Multinomial Data

Hardy-Weinberg via Fisher's exact test

```
library(hwde)
n1 <- 88
n2 <- 10
n3 <- 2
exact <- hwexact(n1,n2,n3)
exact
## [1] 0.06544427
```

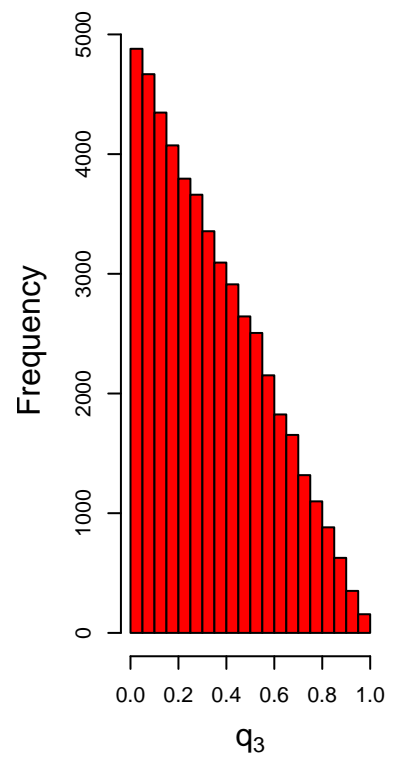
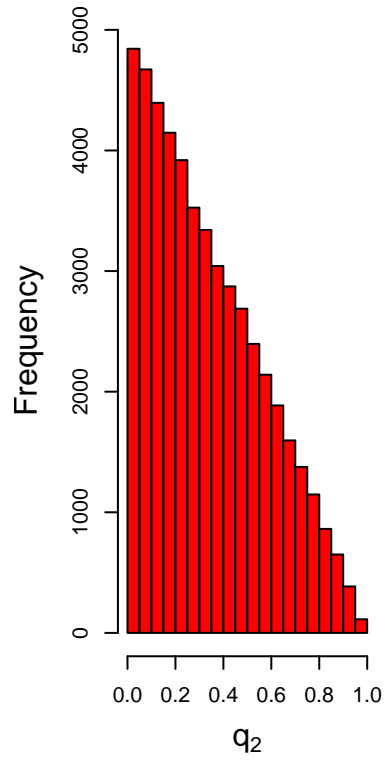
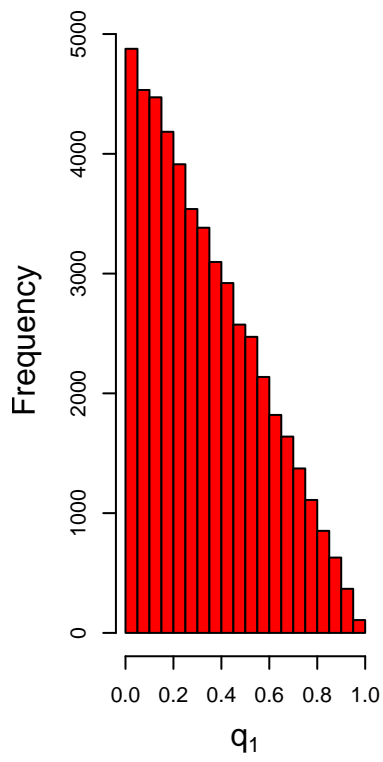
We obtain a p-value of 0.07

Displaying samples from a Dirichlet(1,1,1)

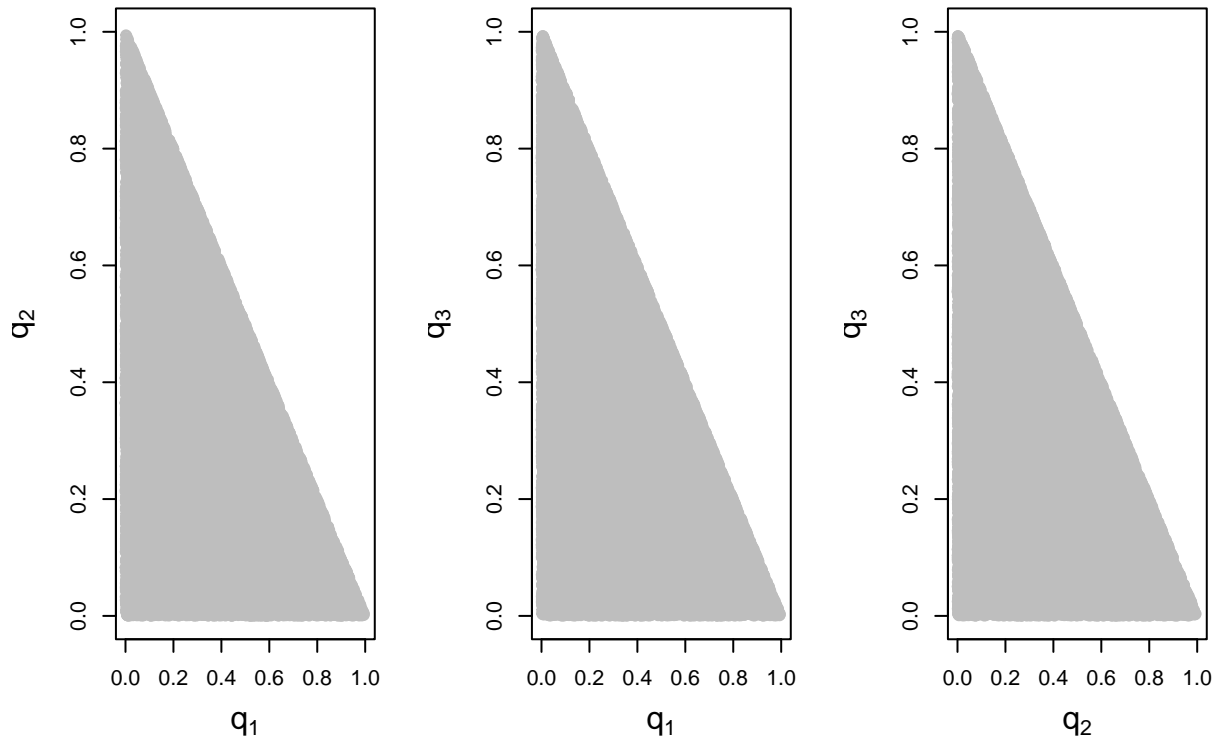
```
library(VGAM) # To access the rdiric function
nsim <- 50000
q <- rdiric(nsim,c(1,1,1))
#
# Univariate marginal representations
#
par(mfrow=c(1,3))
hist(q[,1],xlab=expression(q[1]),main="",cex.lab=1.5,xlim=c(0,1),col="red")
hist(q[,2],xlab=expression(q[2]),main="",cex.lab=1.5,xlim=c(0,1),col="red")
hist(q[,3],xlab=expression(q[3]),main="",cex.lab=1.5,xlim=c(0,1),col="red")
```

```
#
# Bivariate representations
#
plot(q[,1],q[,2],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[1]),
     ylab=expression(q[2]),cex.lab=1.5)
plot(q[,1],q[,3],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[1]),
     ylab=expression(q[3]),cex.lab=1.5)
plot(q[,2],q[,3],xlim=c(0,1),ylim=c(0,1),xlab=expression(q[2]),
     ylab=expression(q[3]),cex.lab=1.5)
```

Histograms



Scatteplots

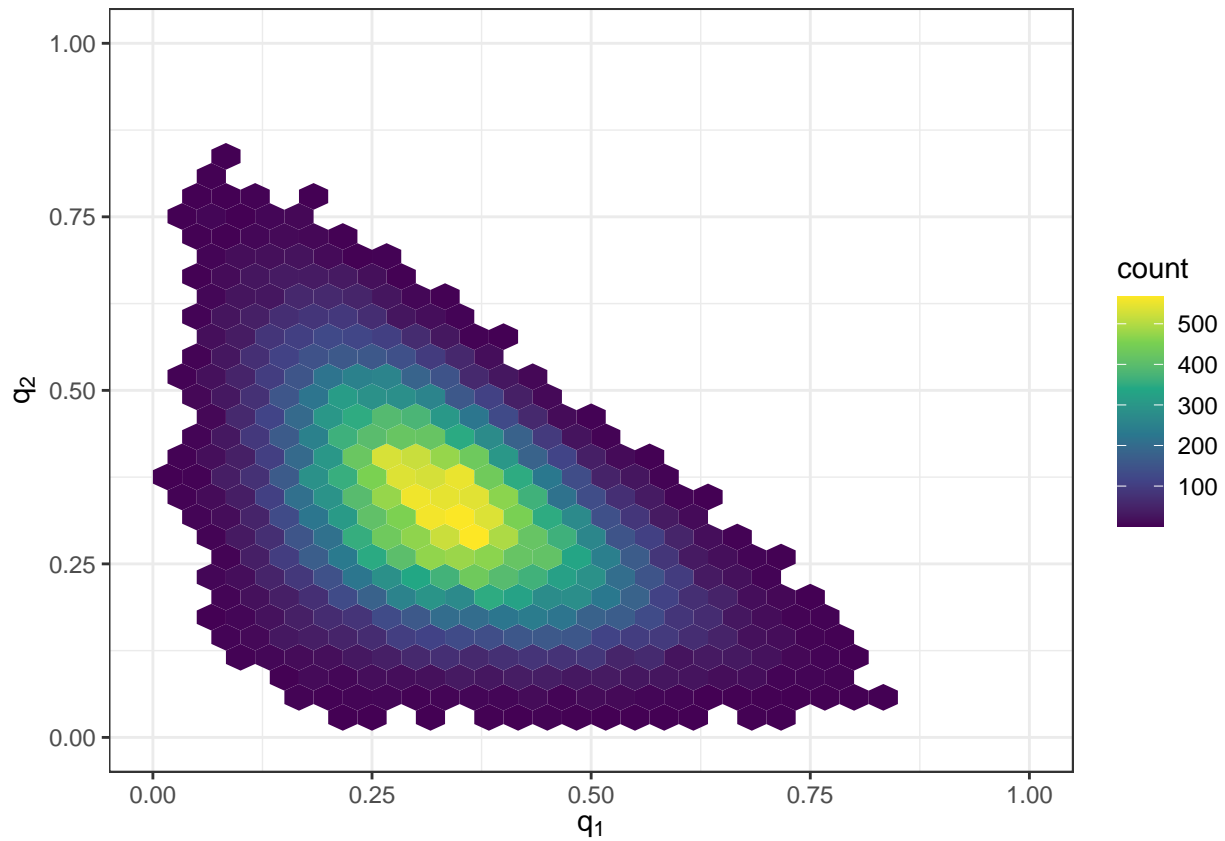


Displaying samples from Dirichlets

```
library(hexbin)
library(ggplot2)
nsim <- 50000
ex1 <- ex2 <- NULL
ex1[1] <- ex1[2] <- ex1[3] <- 5
qex1 <- rdiric(nsim,ex1)
ex2[1] <- 6; ex2[2] <- 4; ex2[3] <- 1
qex2 <- rdiric(nsim,ex2)
# See https://www.r-graph-gallery.com/2d-density-plot-with-ggplot2.html
```

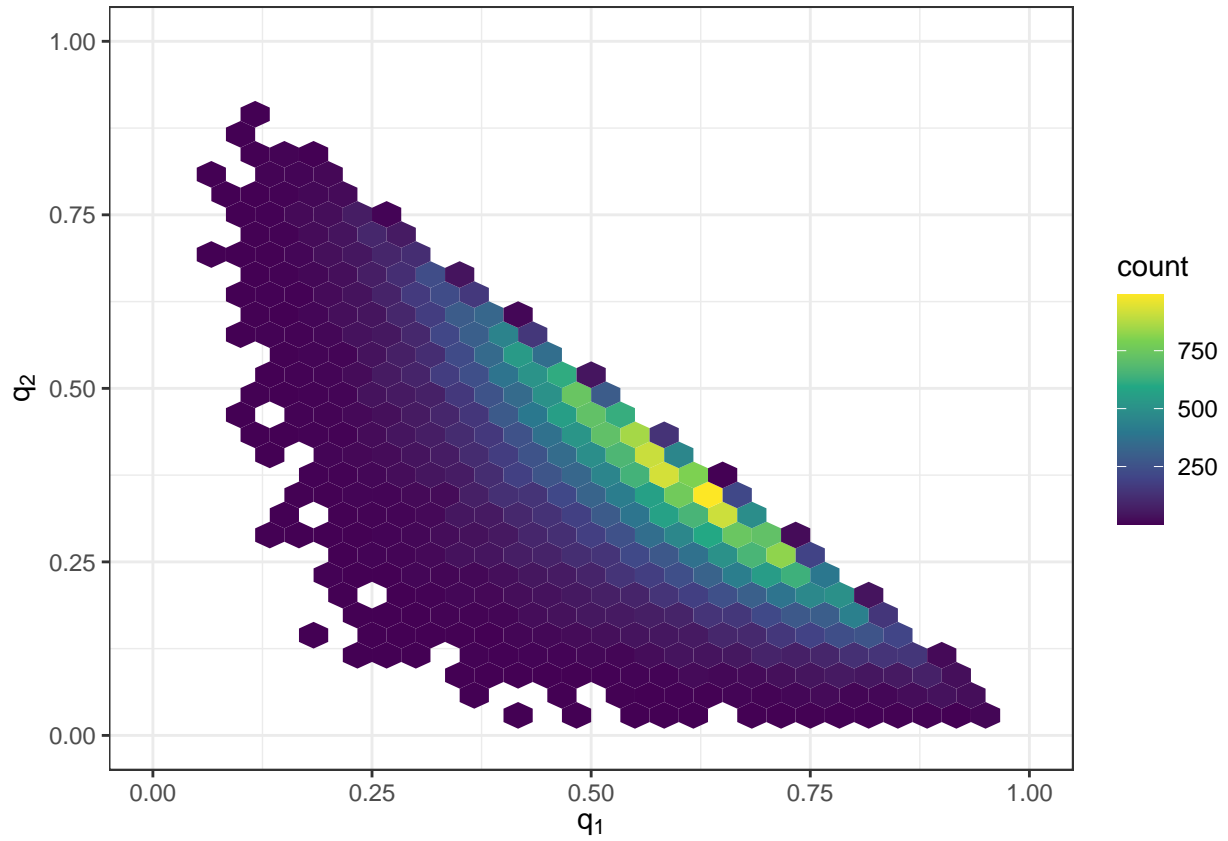
Displaying samples from a Dirichlet(5,5,5)

```
ggplot(data.frame(qex1[,1],qex1[,2]),aes(x=qex1[,1],y=qex1[,2]))+geom_hex(bins=30)+scale_fill_continuous
```



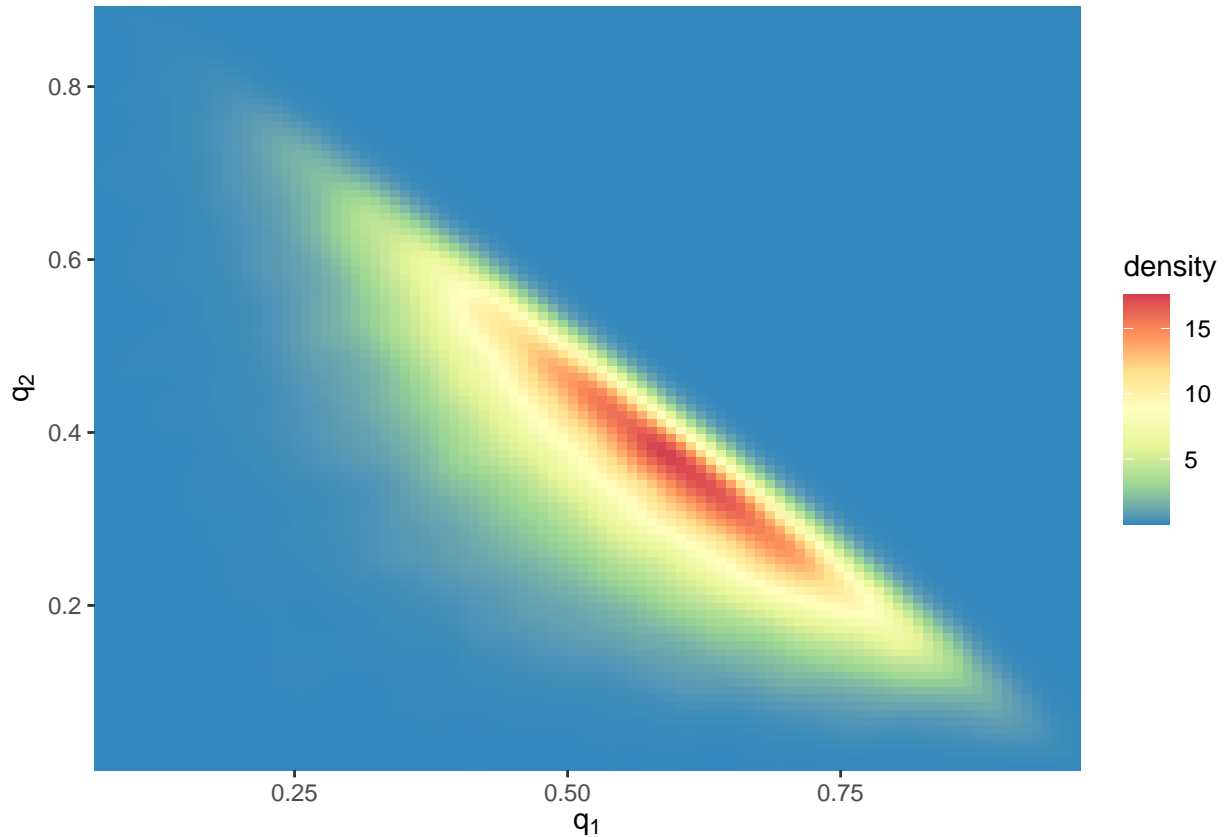
Displaying samples from a Dirichlet(6,4,1)

```
ggplot(data.frame(qex2[,1],qex2[,2]),aes(x=qex2[,1],y=qex2[,2]))+geom_hex(bins=30)+scale_fill_continuous
```

Displaying samples from a Dirichlet(6,4,1)

```
ggplot(data.frame(qex2[,1],qex2[,2]),aes(x=qex2[,1],y=qex2[,2]))+stat_density_2d(aes(fill=..density..),
```



Functions of interest: implied priors

We assume a “`dirichlet(1,1,1)`” distribution

```
q1 <- q[,1]
q2 <- q[,2]
q3 <- q[,3]
p1 <- q1+q2/2
p2 <- q3+q2/2;
f <- (q1-p1^2)/(p1*p2)
D <- q1-p1^2
psi <- q2^2/(p1*p2)
## Functions of interest
cat("Prior prob f>0: ",sum(f>0)/nsim,"\n")
## Prior prob f>0: 0.66576
```

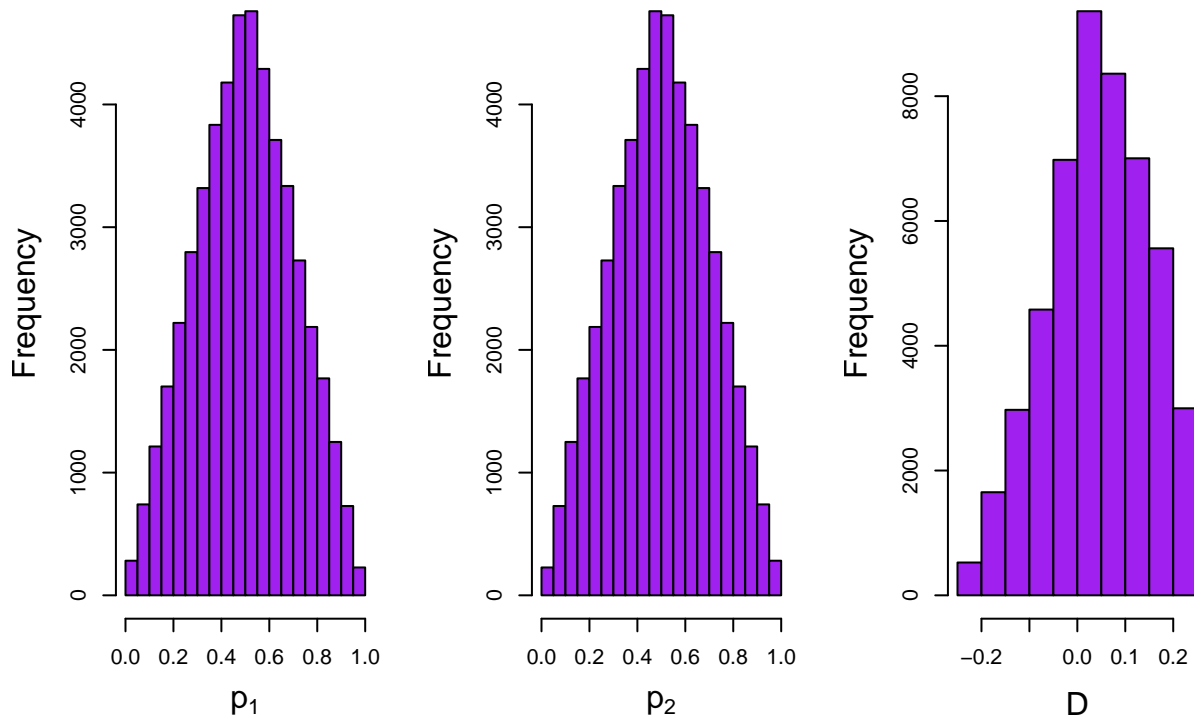
```
cat("Prior prob D>0: ",sum(D>0)/nsim,"\n")
## Prior prob D>0: 0.66576
```

Examine prior summaries for different functions of interest.

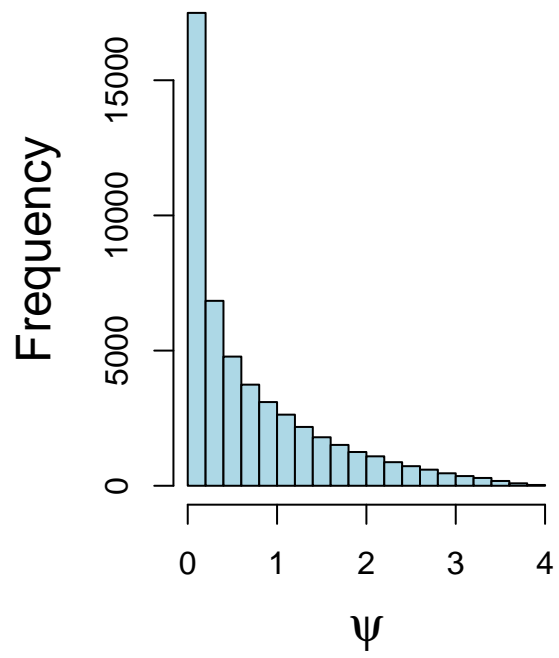
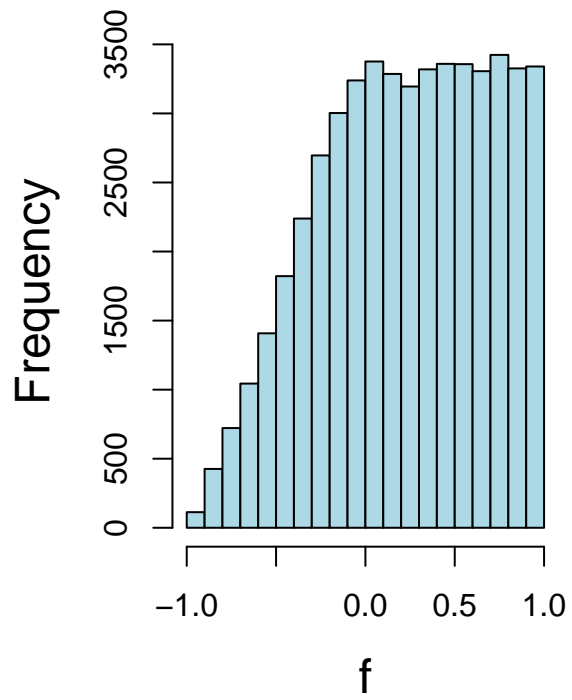
```
par(mfrow=c(1,3))
hist(p1,main="",xlab=expression(p[1]),cex.lab=1.5)
hist(p2,main="",xlab=expression(p[2]),cex.lab=1.5)
hist(D,main="",xlab=expression(D),cex.lab=1.5)
```

```
par(mfrow=c(1,2))
hist(f,main="",xlab="f",cex.lab=1.5)
hist(psi,main="",xlab=expression(psi),cex.lab=1.5)
```

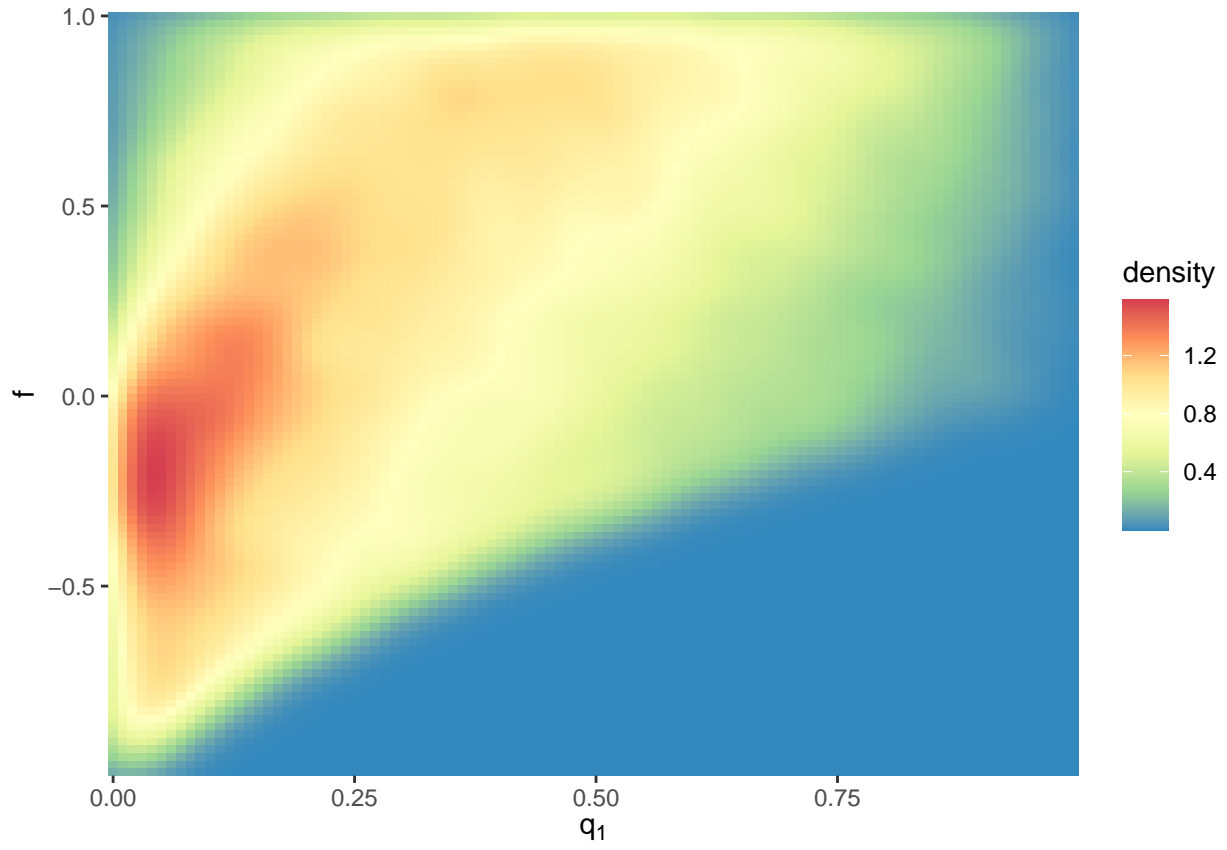
Functions of interest: priors on p_1, p_2, D

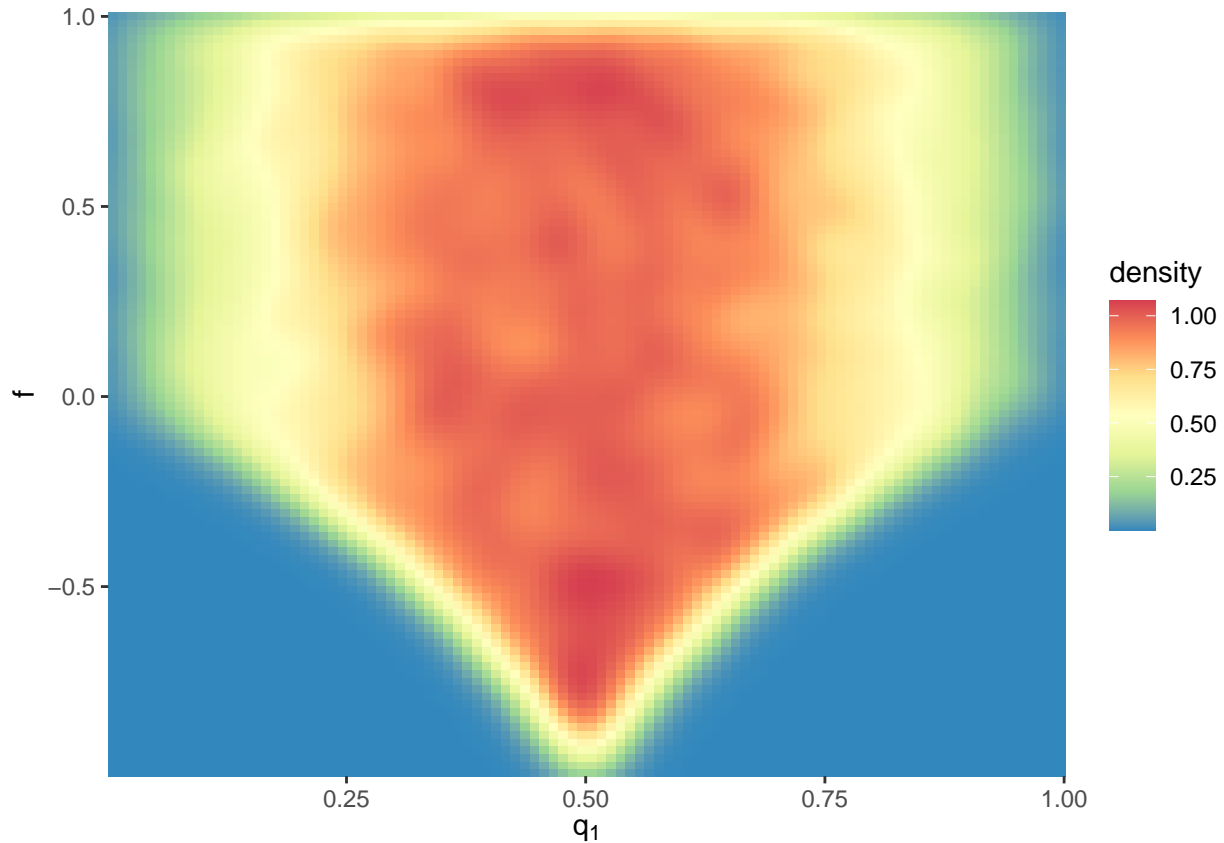


Functions of interest: priors on f, ψ .



Prior on f and q_1





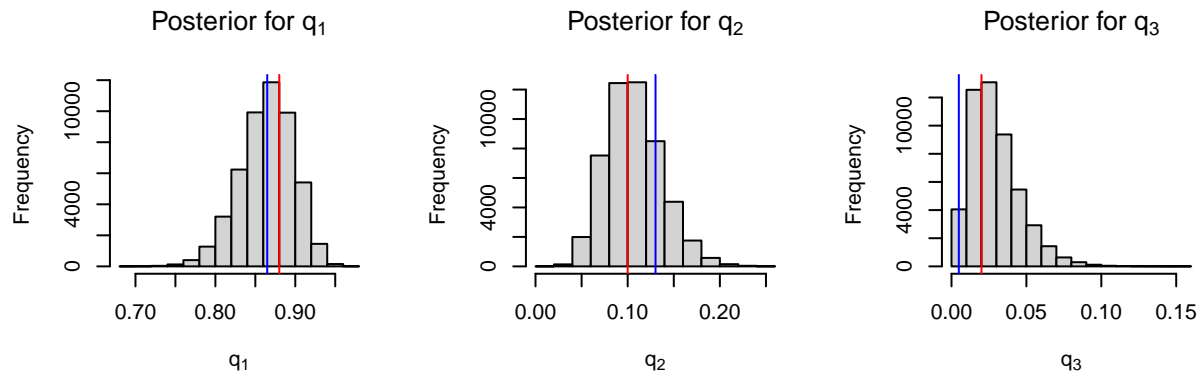
Bayes analysis of (88,10,2) data

```

n1 <- 88; n2 <- 10; n3 <- 2
p1 <- 88/100 + 0.5*10/100 # Estimated allele frequencies
p2 <- 2/100 + 0.5*10/100 # for A1 and A2
v1 <- v2 <- v3 <- 1
q <- rdiric(nsim,c(n1+v1,n2+v2,n3+v3)) # The posterior
q1 <- q[,1]; q2 <- q[,2]; q3 <- q[,3]
par(mfrow=c(1,3))
hist(q1,xlab=expression(q[1]),
     main=expression(paste("Posterior for ",q[1])))
abline(v=n1/(n1+n2+n3),col="red")
abline(v=p1^2,col="blue")
hist(q2,xlab=expression(q[2]),
     main=expression(paste("Posterior for ",q[2])))
abline(v=n2/(n1+n2+n3),col="red")
abline(v=2*p1*p2,col="blue")
hist(q3,xlab=expression(q[3]),
     main=expression(paste("Posterior for ",q[3])))
abline(v=n3/(n1+n2+n3),col="red")
abline(v=p2^2,col="blue")

```

Univariate posterior distributions: blue lines are the MLEs in the full model, red lines under the HWE model

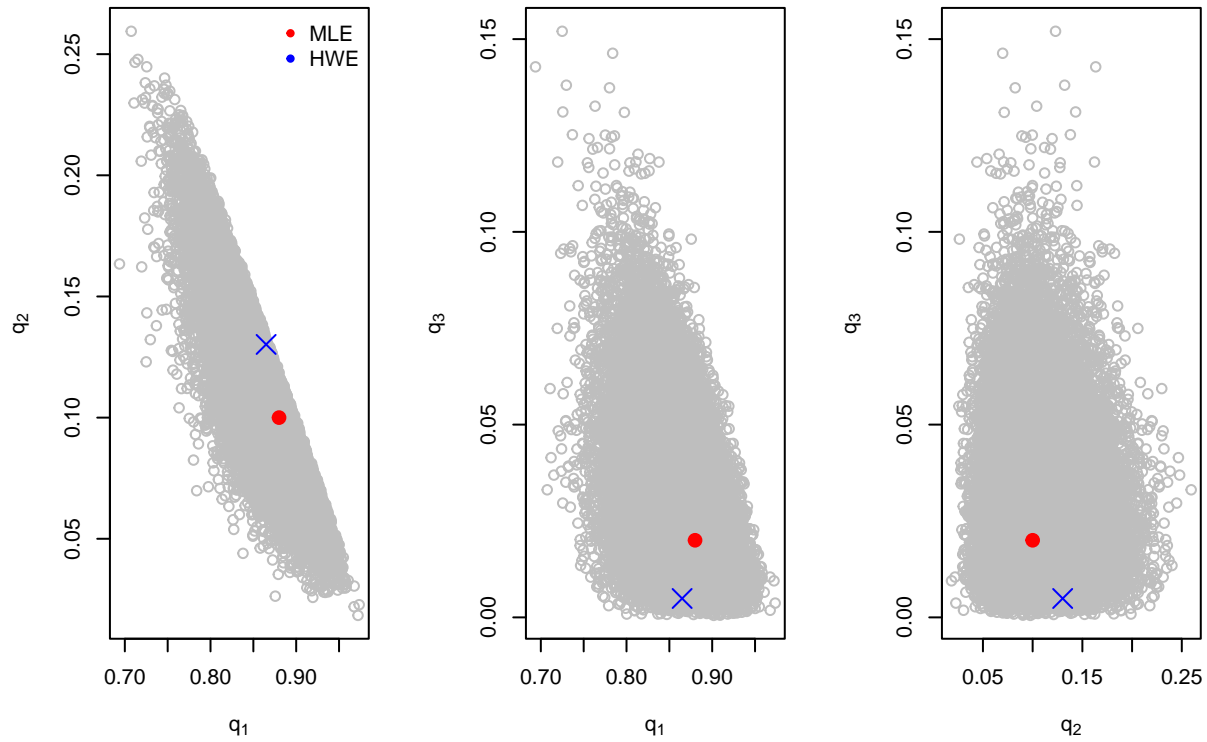


```

par(mfrow=c(1,3))
plot(q2~q1,xlab=expression(q[1]),ylab=expression(q[2]),
     col="grey")
points(n1/(n1+n2+n3),n2/(n1+n2+n3),col="red",pch=20,cex=2)
points(p1^2,2*p1*p2,col="blue",pch=4,cex=2)
legend("topright",legend=c("MLE","HWE"),col=c("red","blue"),
      pch=c(20,20),bty="n")
plot(q3~q1,xlab=expression(q[1]),ylab=expression(q[3]),
     col="grey")
points(n1/(n1+n2+n3),n3/(n1+n2+n3),col="red",pch=20,cex=2)
points(p1^2,p2^2,col="blue",pch=4,cex=2)
plot(q3~q2,xlab=expression(q[2]),ylab=expression(q[3]),
     col="grey")
points(n2/(n1+n2+n3),n3/(n1+n2+n3),col="red",pch=20,cex=2)
points(2*p1*p2,p2^2,col="blue",pch=4,cex=2)

```

Bivariate posterior distributions: blue lines are the MLEs in the full model, red lines under the HWE model



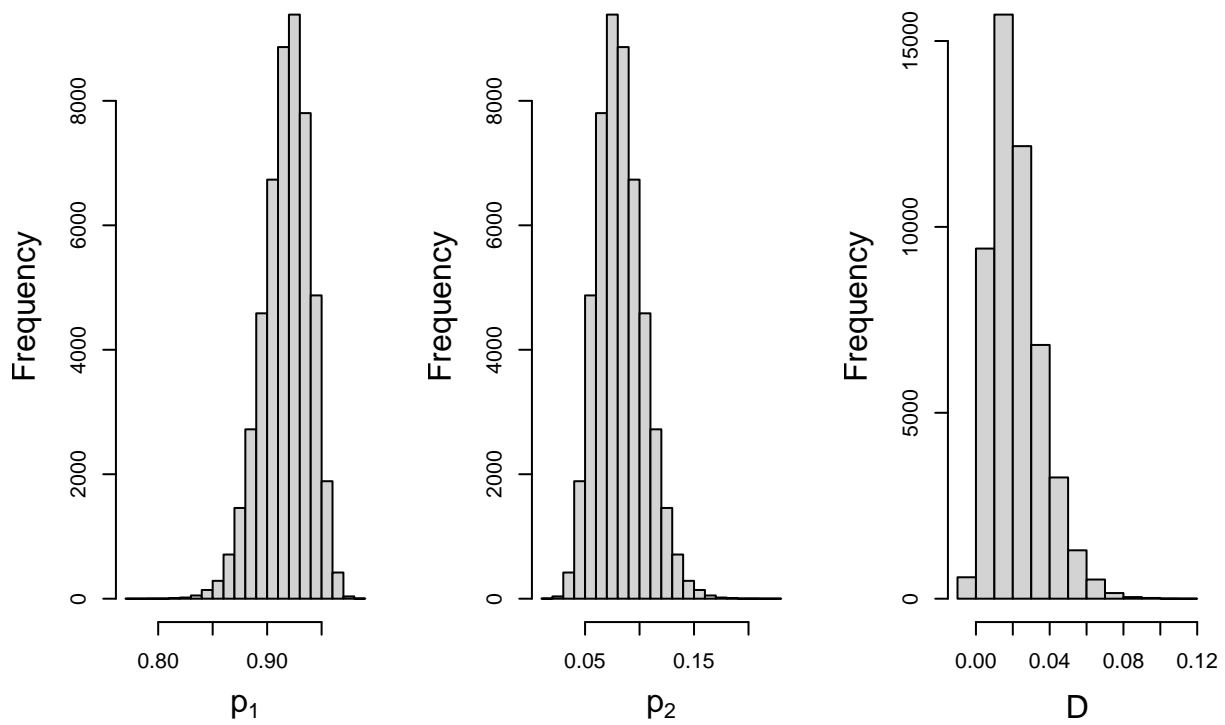
Inference for f

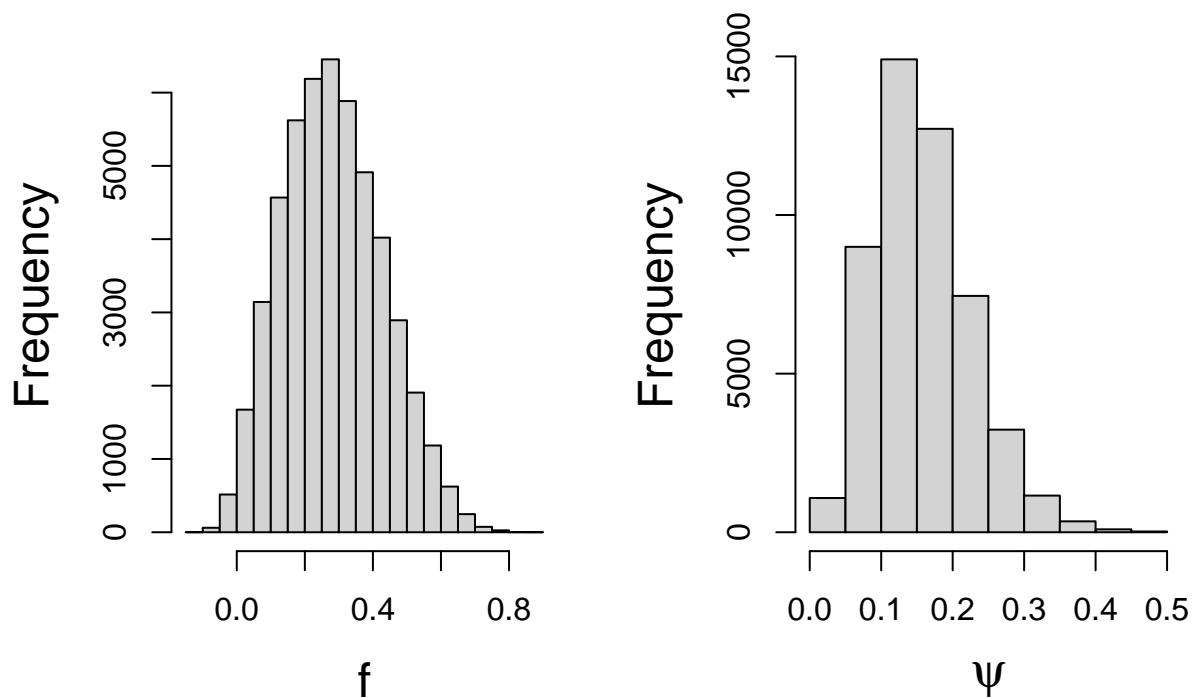
The MLE is $\hat{f} = 0.23$ with asymptotic standard error 0.17.

Hence, a 95% asymptotic confidence interval is

$$(0.23 - 1.96 \times 0.17, 0.23 + 1.96 \times 0.17) = (-0.1032, 0.5632).$$

Posterior summaries





Exercises

- Repeat the analyses in these notes for the Lidicker et al data discussed in class, ie with $n_1 = 37$, $n_2 = 20$, $n_3 = 7$. In particular, with a Dirichlet(1,1,1) prior on the 3 probabilities, generate samples, and examine the various summaries, including marginal allele frequencies and the inbreeding coefficient f .
- Download and install the INLA package from the website in the course notes. Run the LHON example using the commands from the slides. If you get this to work, repeat the Session 3 exercise switching the informative prior from β_1 to β_0 .