2017 SISG Bayesian Statistics for Genetics
R Notes: Multinomial Sampling

Jon Wakefield
Departments of Statistics and Biostatistics, University of Washington

2017-07-22

Hardy-Weinberg via Fisher's exact test

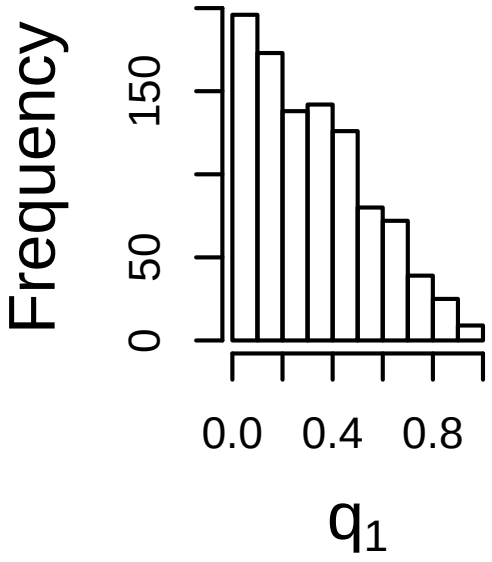
```
library(hwde)
n1 <- 88
n2 <- 10
n3 <- 2
exact <- hwexact(n1, n2, n3)
exact
## [1] 0.06544427
```

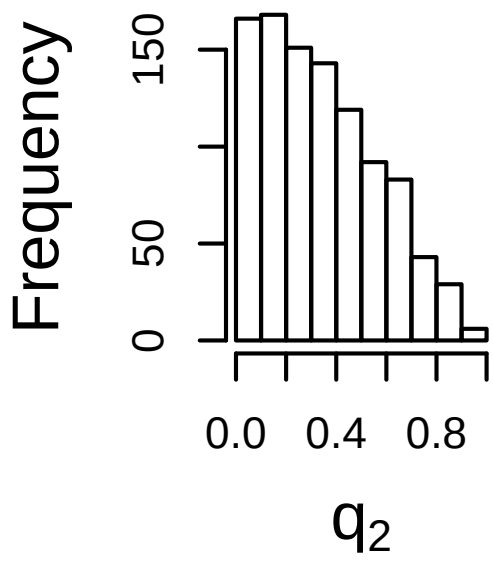
We obtain a p-value of 0.07

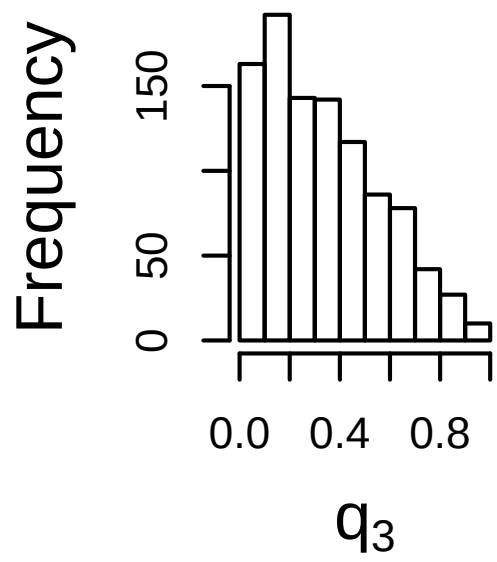
Displaying samples from a dirichlet(1,1,1)

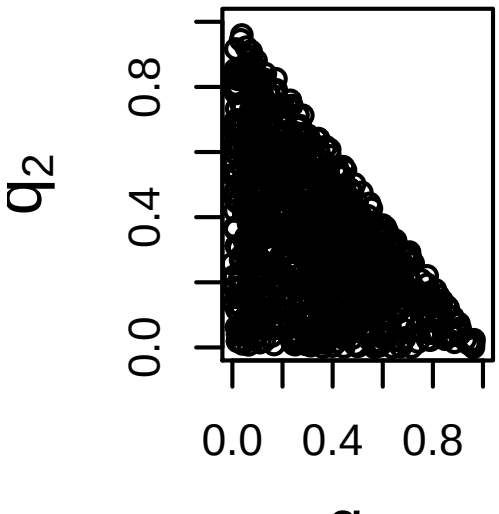
```
library(VGAM)  # To access the rdiric function
nsim <- 1000
q <- rdiric(nsim, c(1, 1, 1))
# Univariate marginal representations
par(mfrow = c(2, 3))
hist(q[, 1], xlab = expression(q[1]), main = "", cex.lab = 1.5,
    xlim = c(0, 1))
hist(q[, 2], xlab = expression(q[2]), main = "", cex.lab = 1.5,
    xlim = c(0, 1))
hist(q[, 3], xlab = expression(q[3]), main = "", cex.lab = 1.5,
    xlim = c(0, 1))
# 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)
```

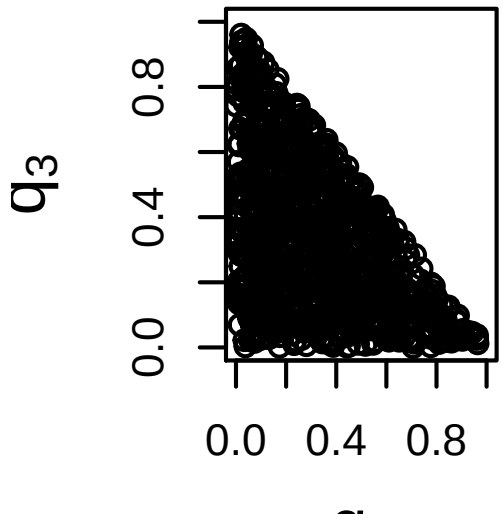
Displaying samples from a dirichlet(1,1,1)

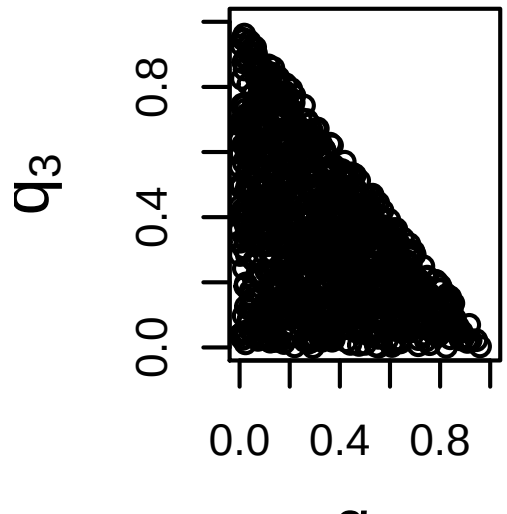










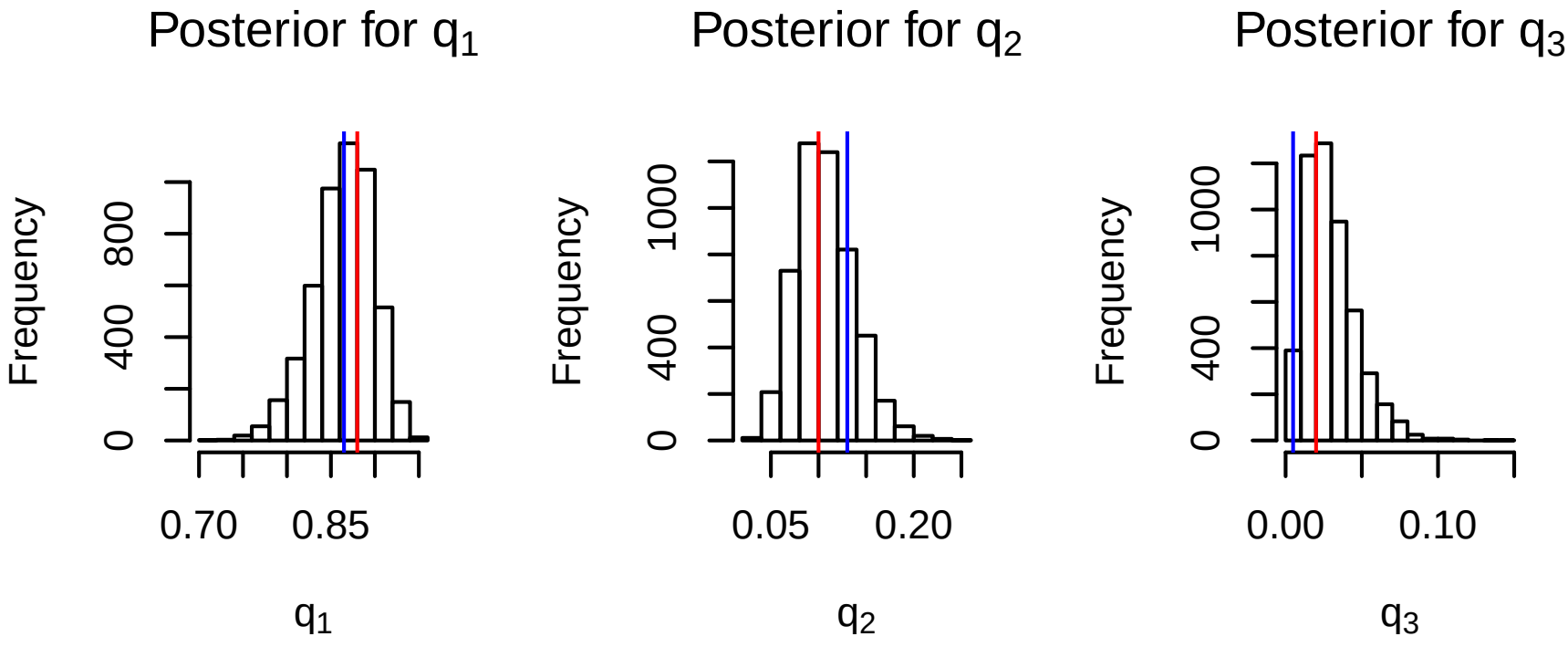


Bayes analysis of (88,10,2) data

```
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
nsim <- 5000
q <- rdiric(nsim, c(n1 + v1, n2 + v2, n3 + v3))  # The posterior
par(mfrow = c(1, 3))
hist(q[, 1], 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(q[, 2], 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(q[, 3], 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")
```

Bayes analysis of (88,10,2) data

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

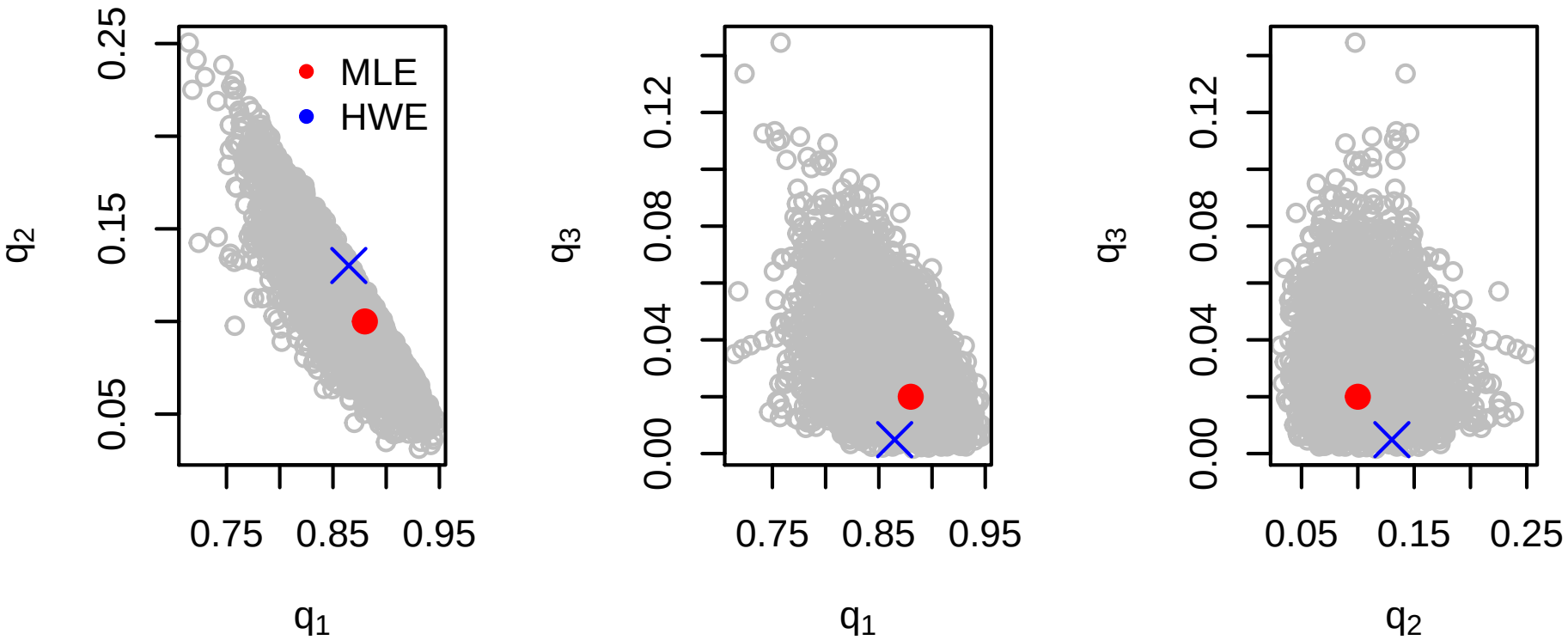


Bayes analysis of (88,10,2) data

```
par(mfrow = c(1, 3))
plot(q[, 2] ~ q[, 1], 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(q[, 3] ~ q[, 1], 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(q[, 3] ~ q[, 2], 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)
```

Bayes analysis of (88,10,2) data

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



Functions of interest: implied priors

We assume a "dirichlet(1,1,1)"' distribution

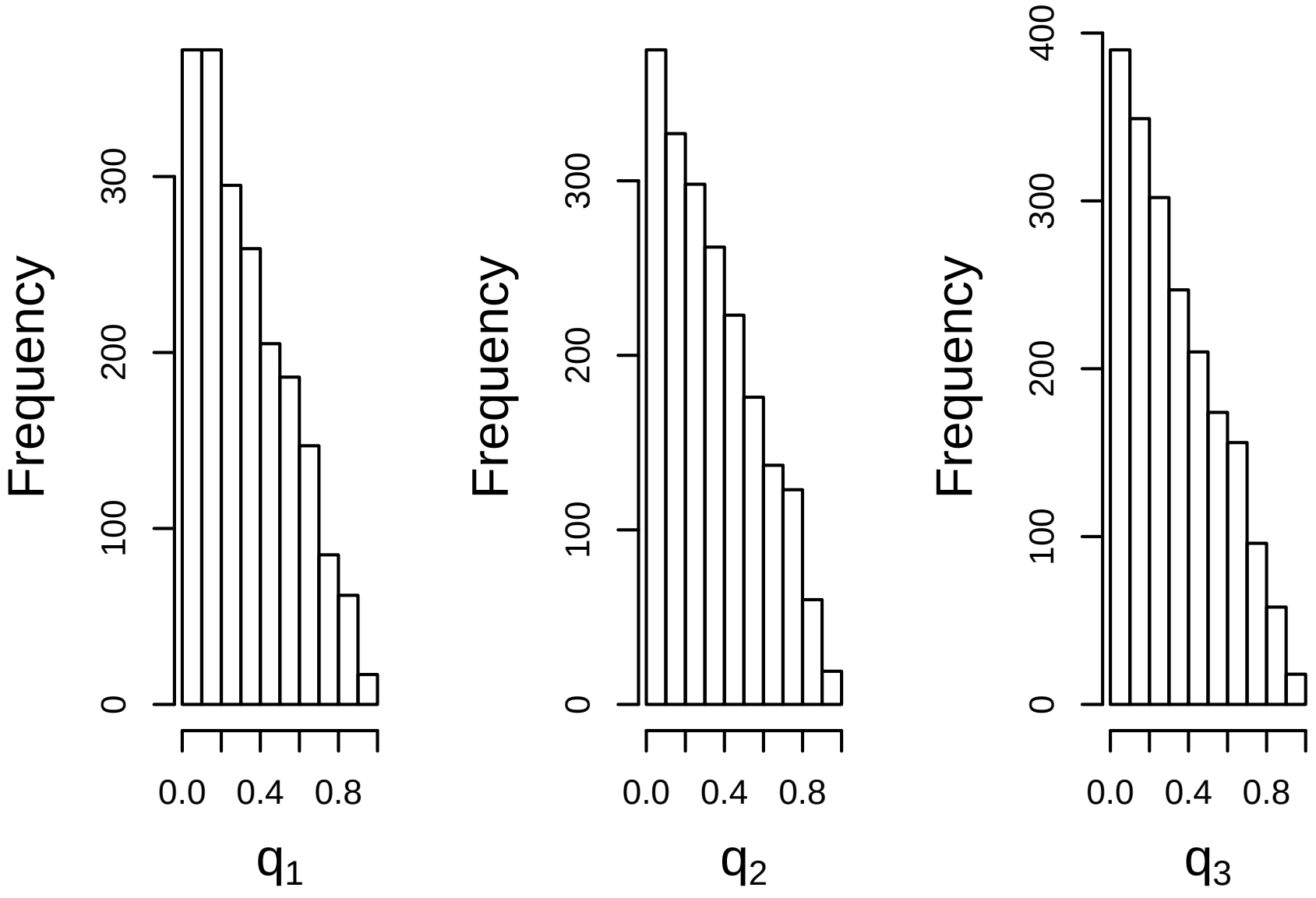
```
v1 <- v2 <- v3 <- 1
nsim <- 2000
samps <- rdiric(nsim, c(v1, v2, v3))
q1 <- samps[, 1]
q2 <- samps[, 2]
q3 <- samps[, 3]
p1 <- q1 + q2/2
p2 <- q3 + q2/2
f <- (q1 - p1^2)/(p1 * p2)
D <- q1 - p1^2
psi <- q2^2/(p1 * p2)
## Functions of interest
cat("Prior prob f>0: ", sum(f > 0)/nsim, "\n")
## Prior prob f>0:  0.6535
cat("Prior prob D>0: ", sum(D > 0)/nsim, "\n")
## Prior prob D>0:  0.6535
```

Functions of interest

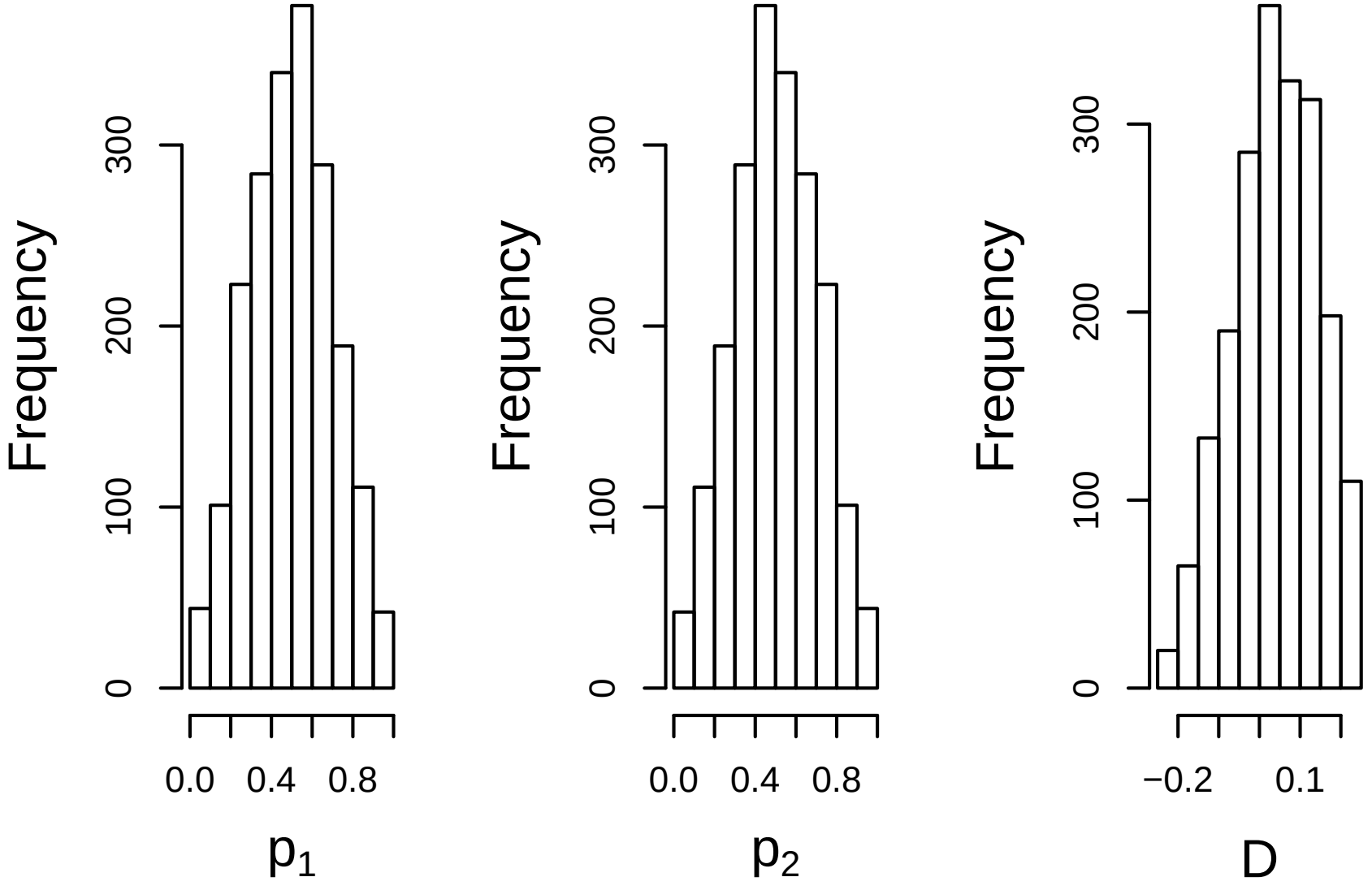
Examine prior summaries for different functions of interest.

```
par(mfrow = c(1, 3))
hist(q1, main = "", xlab = expression(q[1]), cex.lab = 1.5)
hist(q2, main = "", xlab = expression(q[2]), cex.lab = 1.5)
hist(q3, main = "", xlab = expression(q[3]), cex.lab = 1.5)
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)
```

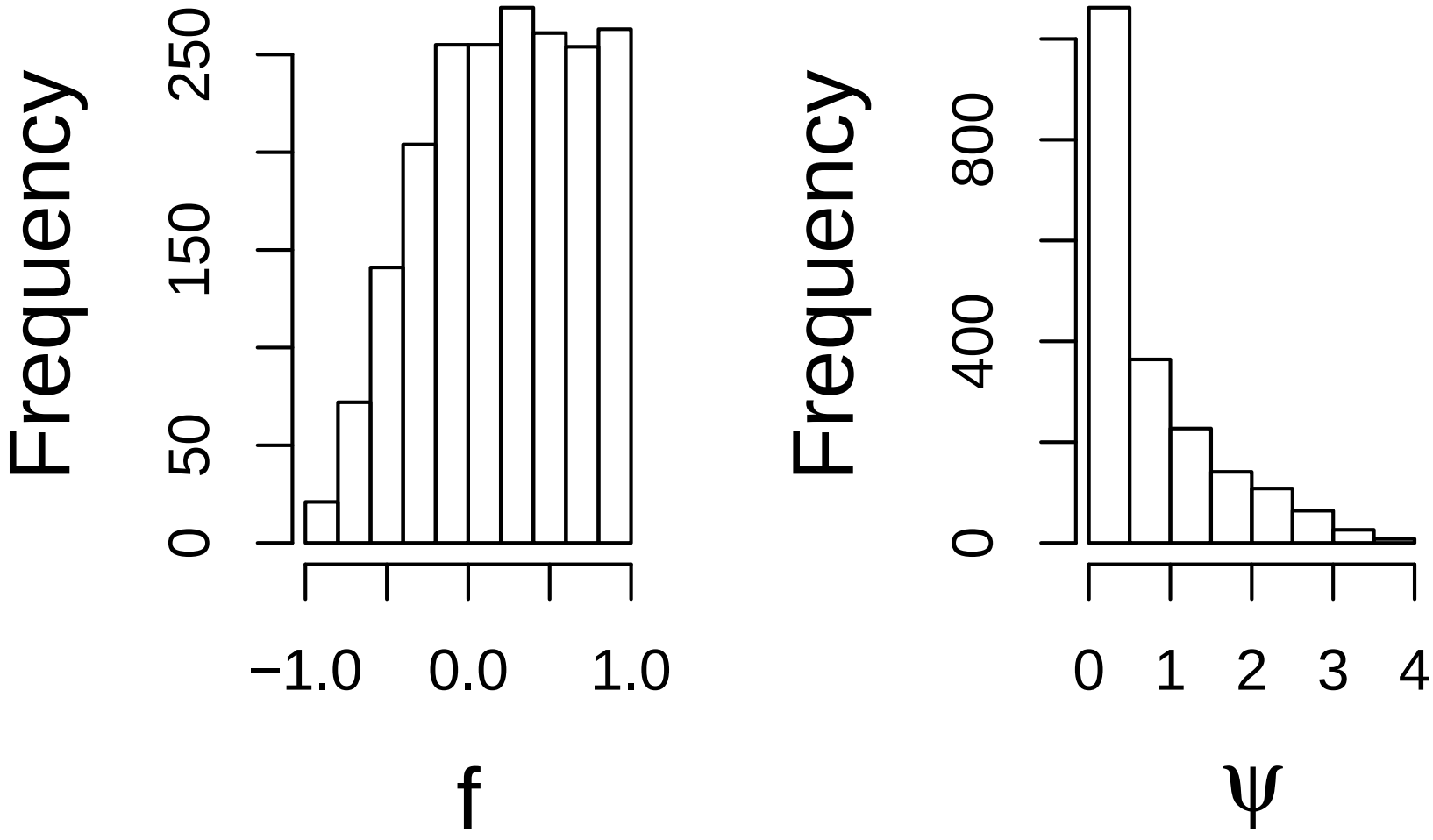
Functions of interest: prior margins on $q1, q_2, q3$.



Functions of interest: priors on p_1, p_2, D



Functions of interest: priors on f, ψ.



Inference for f

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

Hence, a 95% confidence interval is

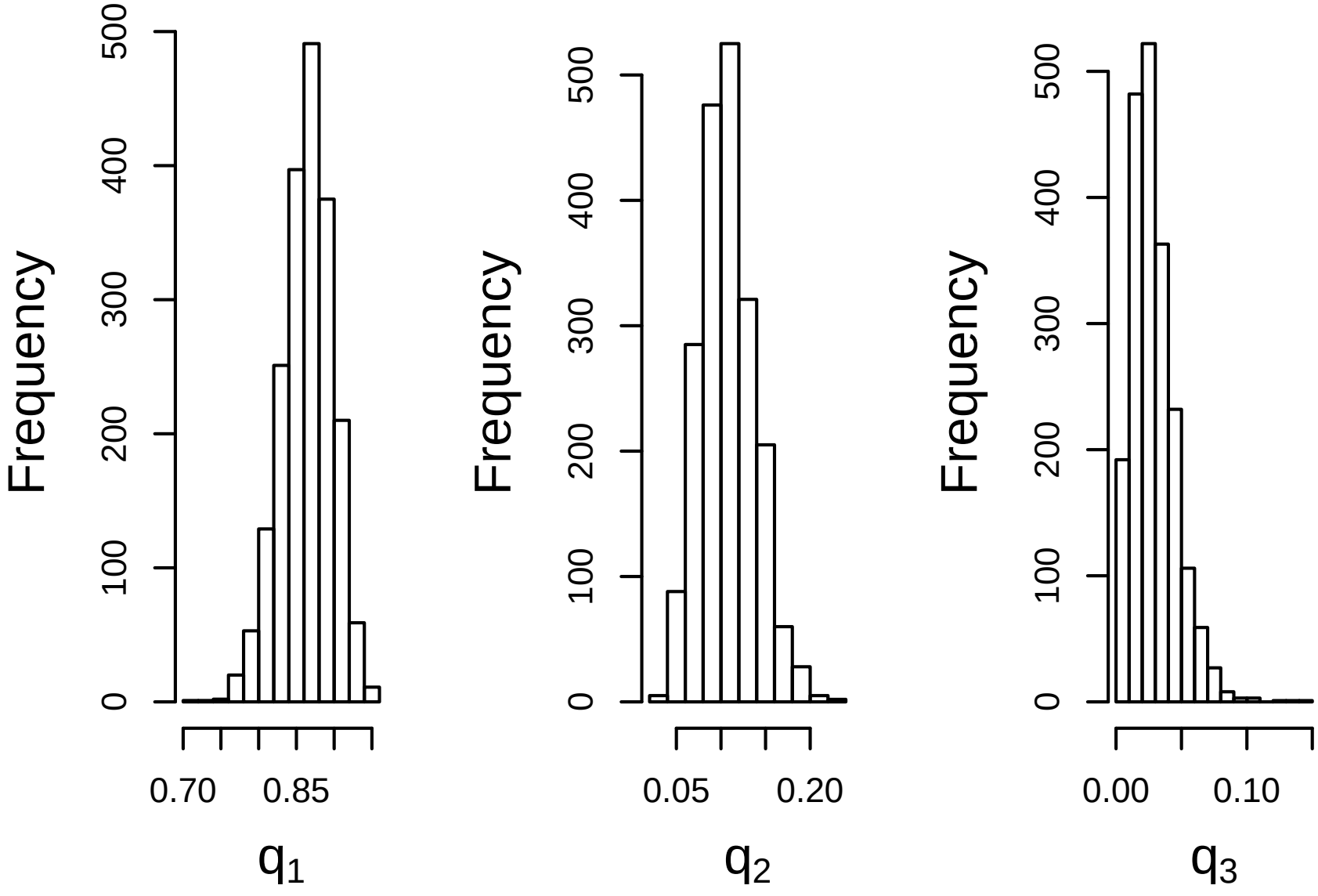
$$(0.23 - 1.96 \times 0.17, 0.23 + 1.96 \times 0.17) = (-0.1032, 0.5632).$$

The posterior median and 95% credible interval are given below.

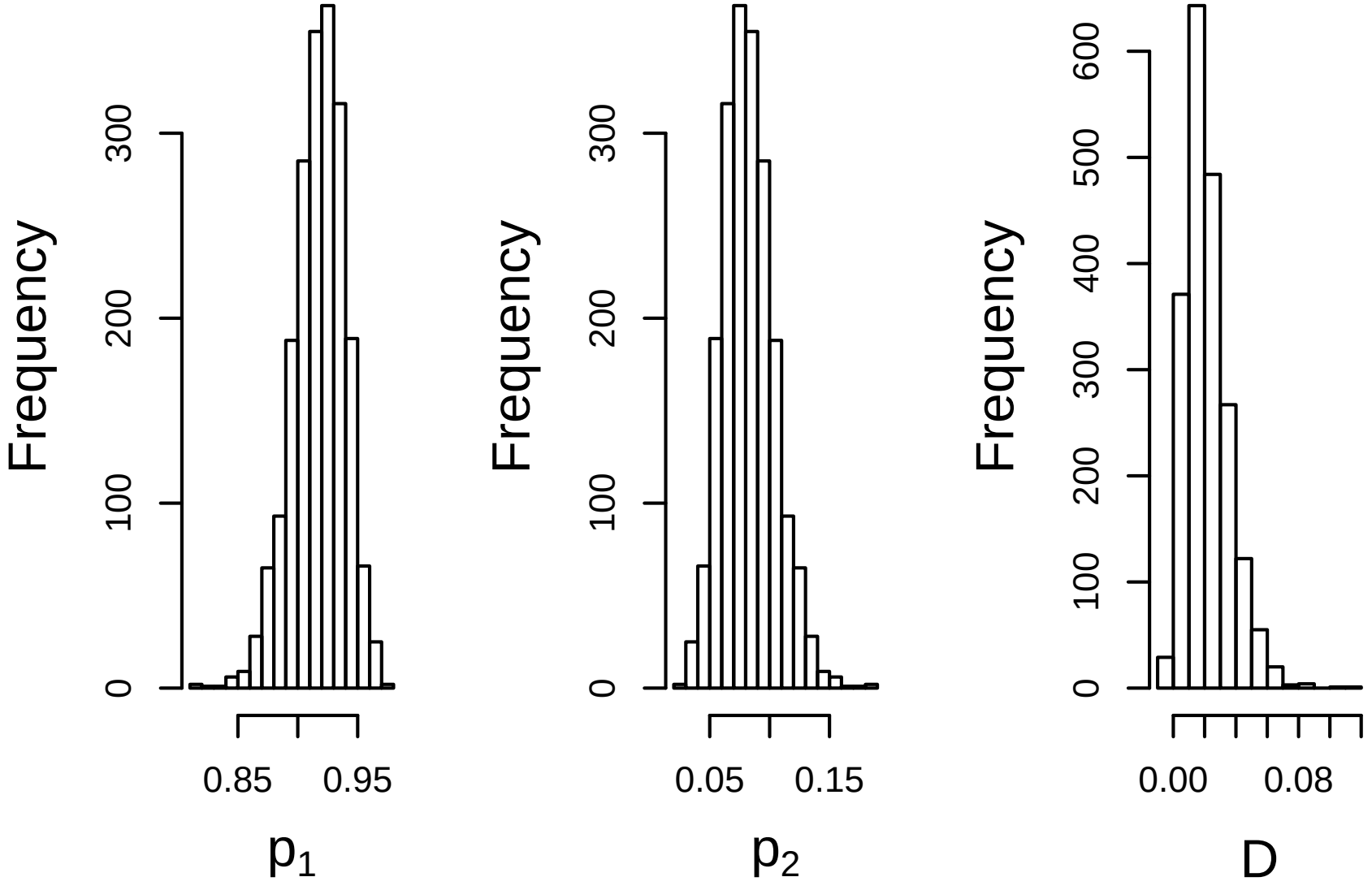
```
# Bayesian posterior quantiles are
quantile(f, c(0.025, 0.5, 0.975))
##       2.5%        50%      97.5%
## -0.6922894  0.2414654  0.9588643
```

Subsequent figures give posterior distributions on functions of interest.

Dirichlet Posterior Distribution



Posterior summaries



Posterior summaries