Hierarchical models

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 20, 2019

Outline

- Motivating example
  - Independent vs pooled estimates
- Hierarchical models
  - General structure
  - Posterior distribution
- Binomial hierarchial model
  - Posterior distribution
  - Prior distributions
- Stan analysis of binomial hierarchical model
  - informative prior
  - default prior
  - integrating out θ
  - across seasons

Andre Dawkin's three-point percentage

Suppose Y_i are the number 3-pointers Andre Dawkin's makes in season i, and assume

$$Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$$

where

- n_i are the number of 3-pointers attempted and
- θ_i is the probability of making a 3-pointer in season i.

Do these models make sense?

- The 3-point percentage every season is the same, i.e. $\theta_i = \theta$.
- The 3-point percentage every season is independent of other seasons.
- The 3-point percentage every season should be similar to other seasons.

Andre Dawkin's three-point percentage

Suppose Y_i are the number of 3-pointers Andre Dawkin's makes in game i, and assume

$$Y_i \overset{ind}{\sim} Bin(n_i, \theta_i)$$

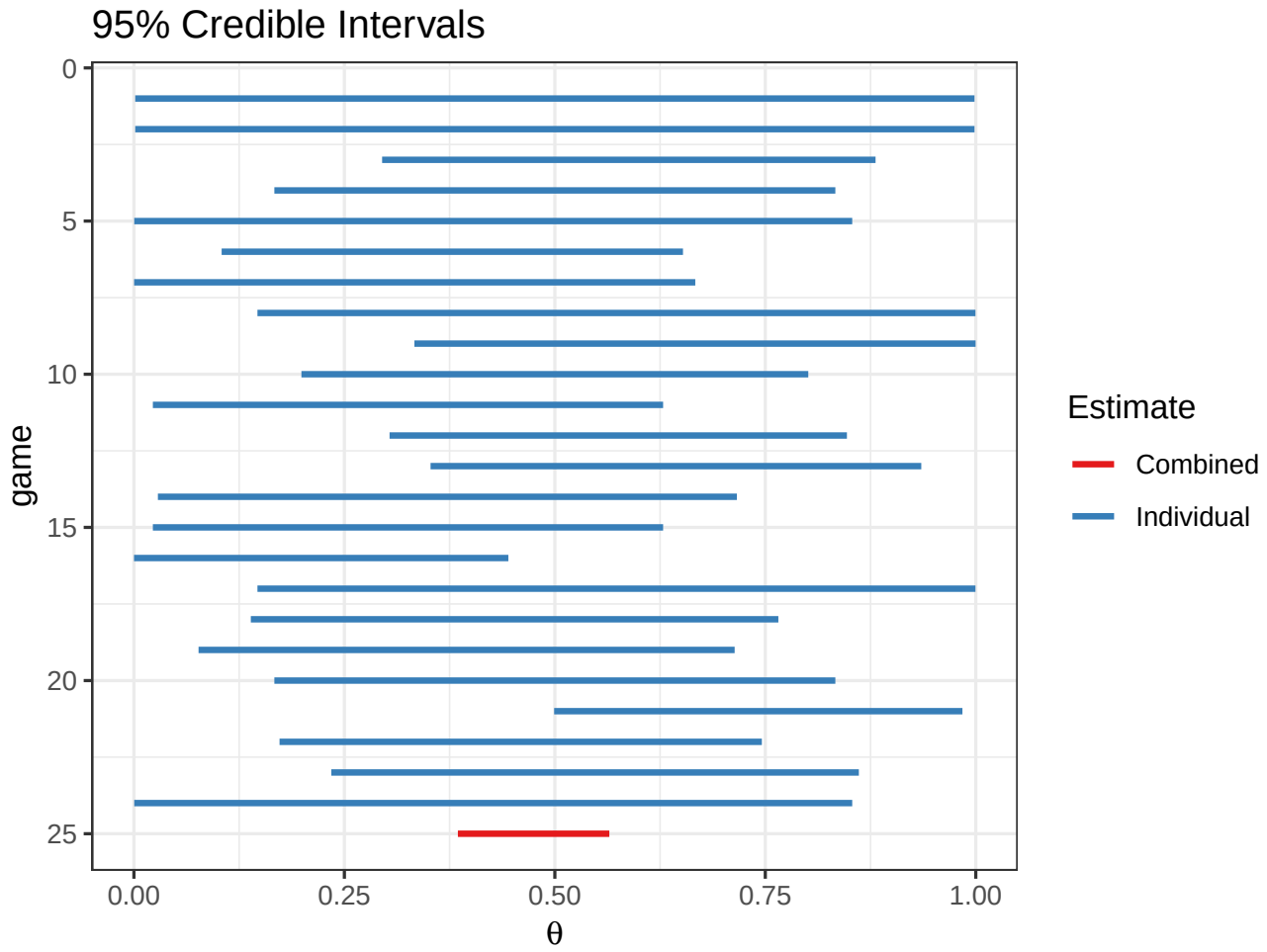
where

- n_i are the number of 3-pointers attempted in game i and
- θ_i is the probability of making a 3-pointer in game i.

Do these models make sense?

- The 3-point percentage every game is the same, i.e. $\theta_i = \theta$.
- The 3-point percentage every game is independent of other games.
- The 3-point percentage every game should be similar to other games.

Andre Dawkin's 3-point percentage



Andre Dawkin's 3-point percentage

	date	opponent	made	attempts
1	11/8/13	davidson	0	0
2	11/12/13	kansas	0	0
3	11/15/13	florida atlantic	5	8
4	11/18/13	unc asheville	3	6
5	11/19/13	east carolina	0	1
6	11/24/13	vermont	3	9
7	11/27/13	alabama	0	2
8	11/29/13	arizona	1	1
9	12/3/13	michigan	2	2
10	12/16/13	gardner-webb	4	8
11	12/19/13	ucla	1	5
12	12/28/13	eastern michigan	6	10
13	12/31/13	elon	5	7
14	1/4/14	notre dame	1	4
15	1/7/14	georgia tech	1	5
16	1/11/14	clemson	0	4
17	1/13/14	virginia	1	1
18	1/18/14	nc state	3	7
19	1/22/14	miami	2	6
20	1/25/14	florida state	3	6
21	1/27/14	pitt	6	7
22	2/1/14	syracuse	4	9
23	2/4/14	wake forest	4	7
24	2/8/14	boston college	0	1

Hierarchical models

Consider the following model

$$\begin{aligned} y_i &\overset{ind}{\sim} p(y|\theta_i) \\ \theta_i &\overset{ind}{\sim} p(\theta|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

where

- y_i is observed,
- $\theta = (\theta_1, \ldots, \theta_n)$ and ϕ are parameters, and
- only ϕ has a prior that is set.

This is a hierarchical or multilevel model.

Posterior distribution for hierarchical models

The joint posterior distribution of interest in hierarchical models is

$$p(\theta, \phi|y) \propto p(y|\theta, \phi)p(\theta, \phi) = p(y|\theta)p(\theta|\phi)p(\phi) = \left[\prod_{i=1}^{n} p(y_i|\theta_i)p(\theta_i|\phi)\right]p(\phi).$$

The joint posterior distribution can be decomposed via

$$p(\theta, \phi|y) = p(\theta|\phi, y)p(\phi|y)$$

where

$$\begin{aligned}
p(\theta|\phi, y) &\propto p(y|\theta)p(\theta|\phi) = \textstyle\prod_{i=1}^{n} p(y_i|\theta_i)p(\theta_i|\phi) \propto \prod_{i=1}^{n} p(\theta_i|\phi, y_i) \\
p(\phi|y) &\propto p(y|\phi)p(\phi) \\
p(y|\phi) &= \textstyle\int p(y|\theta)p(\theta|\phi)d\theta \\
&= \textstyle\int \cdots \int \prod_{i=1}^{n} \left[p(y_i|\theta_i)p(\theta_i|\phi)\right] d\theta_1 \cdots d\theta_n \\
&= \textstyle\prod_{i=1}^{n} \int p(y_i|\theta_i)p(\theta_i|\phi)d\theta_i \\
&= \textstyle\prod_{i=1}^{n} p(y_i|\phi)
\end{aligned}$$

Three-pointer example

Our statistical model

$$\begin{aligned} Y_i &\overset{ind}{\sim} Bin(n_i, \theta_i) \\ \theta_i &\overset{ind}{\sim} Be(\alpha, \beta) \\ \alpha, \beta &\sim p(\alpha, \beta) \end{aligned}$$

In this example,

- $\phi = (\alpha, \beta)$
- $Be(\alpha, \beta)$ describes the variability in 3-point percentage across games, and
- we are going to learn about this variability.

Decomposed posterior

$$Y_i \overset{ind}{\sim} Bin(n_i, \theta_i) \quad \theta_i \overset{ind}{\sim} Be(\alpha, \beta) \quad \alpha, \beta \sim p(\alpha, \beta)$$

Conditional posterior for θ:

$$p(\theta|\alpha, \beta, y) = \prod_{i=1}^{n} p(\theta_i|\alpha, \beta, y_i) = \prod_{i=1}^{n} Be(\theta_i|\alpha + y_i, \beta + n_i - y_i)$$

Marginal posterior for (α, β):

$$\begin{aligned}
p(\alpha, \beta|y) &\propto p(y|\alpha, \beta)p(\alpha, \beta) \\
p(y|\alpha, \beta) &= \textstyle\prod_{i=1}^{n} p(y_i|\alpha, \beta) = \prod_{i=1}^{n} \int p(y_i|\theta_i)p(\theta_i|\alpha, \beta)d\theta_i \\
&= \textstyle\prod_{i=1}^{n} \int Bin(y_i|n_i, \theta_i)Be(\theta_i|\alpha, \beta)d\theta_i \\
&= \textstyle\prod_{i=1}^{n} \int_0^1 \binom{n_i}{y_i} \theta_i^{y_i}(1-\theta_i)^{n_i-y_i} \frac{\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}}{B(\alpha,\beta)} d\theta_i \\
&= \textstyle\prod_{i=1}^{n} \binom{n_i}{y_i} \frac{1}{B(\alpha,\beta)} \int_0^1 \theta_i^{\alpha+y_i-1}(1-\theta_i)^{\beta+n_i-y_i-1} d\theta_i \\
&= \textstyle\prod_{i=1}^{n} \binom{n_i}{y_i} \frac{B(\alpha+y_i, \beta+n_i-y_i)}{B(\alpha,\beta)}
\end{aligned}$$

Thus $y_i|\alpha, \beta \overset{ind}{\sim}$ Beta-binomial(n_i, α, β).

A prior distribution for α and β

Recall the interpretation:

- α: prior successes
- β: prior failures

A more natural parameterization is

- prior expectation: $\mu = \frac{\alpha}{\alpha+\beta}$
- prior sample size: $\eta = \alpha + \beta$

Place priors on these parameters or transformed to the real line:

- logit $\mu = \log(\mu/[1-\mu]) = \log(\alpha/\beta)$
- $\log \eta$

A prior distribution for α and β

It seems reasonable to assume the mean (μ) and size (η) are independent *a priori*:

$$p(\mu, \eta) = p(\mu)p(\eta)$$

Let's construct a prior that has

- $P(0.1 < \mu < 0.5) \approx 0.95$ since most college basketball players have a three-point percentage between 10% and 50% and
- is somewhat diffuse for η but has more mass for smaller values.

Let's assume an informative prior for μ and η perhaps

- $\mu \sim Be(6, 14)$
- $\eta \sim Exp(0.05)$

```
a = 6
b = 14
e = 1/20
```

Prior draws

```
n = 1e4

prior_draws = data.frame(mu  = rbeta(n, a, b),
                         eta = rexp(n, e)) %>%
  mutate(alpha = eta*   mu,
         beta  = eta*(1-mu))

prior_draws %>%
  tidyr::gather(parameter, value) %>%
  group_by(parameter) %>%
  summarize(lower95 = quantile(value, prob = 0.025),
            median  = quantile(value, prob = 0.5),
            upper95 = quantile(value, prob = 0.975))


# A tibble: 4 x 4
  parameter lower95 median upper95
  <chr>       <dbl>  <dbl>   <dbl>
1 alpha       0.129  3.87    23.9
2 beta        0.359  9.61    51.4
3 eta         0.514 13.8     72.4
4 mu          0.124  0.292    0.511


cor(prior_draws$alpha, prior_draws$beta)


[1] 0.7951507
```

```
model_informative_prior = "
data {
  int<lower=0> N;    // data
  int<lower=0> n[N];
  int<lower=0> y[N];
  real<lower=0> a;   // prior
  real<lower=0> b;
  real<lower=0> e;
}
parameters {
  real<lower=0,upper=1> mu;
  real<lower=0> eta;
  real<lower=0,upper=1> theta[N];
}
transformed parameters {
  real<lower=0> alpha;
  real<lower=0> beta;

  alpha = eta*   mu ;
  beta  = eta*(1-mu);
}
model {
  mu    ~ beta(a,b);
  eta   ~ exponential(e);

  // implicit joint distributions
  theta ~ beta(alpha,beta);
  y     ~ binomial(n,theta);
}
"
```

Stan

```
dat = list(y = d$made, n = d$attempts, N = nrow(d),a = a, b = b, e = e)
m = stan_model(model_code = model_informative_prior)
r = sampling(m, dat, c("mu","eta","alpha","beta","theta"),
             iter = 10000)
```

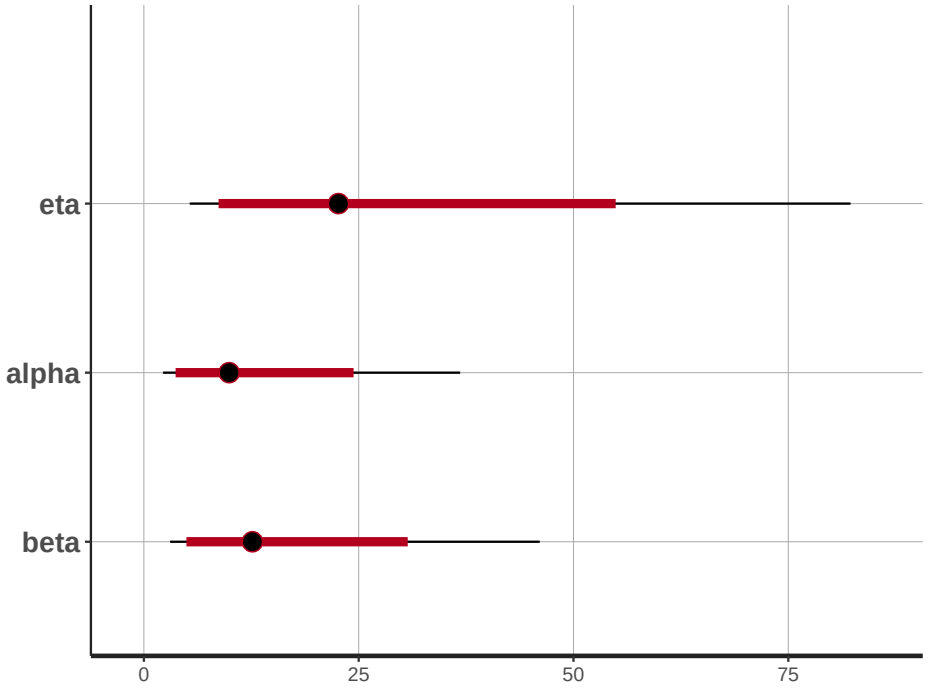
stan

```
r


Inference for Stan model: 81628daed98038e2857f7d139bbc17f5.
4 chains, each with iter=10000; warmup=5000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=20000.

             mean se_mean    sd    2.5%     25%     50%    75%  97.5% n_eff Rhat
mu           0.44    0.00  0.05    0.35    0.41    0.44   0.47   0.53  6217    1
eta         28.19    0.43 20.55    5.34   13.80   22.65  36.49  82.24  2326    1
alpha       12.42    0.19  9.23    2.24    5.93    9.92  16.22  36.81  2387    1
beta        15.76    0.24 11.55    3.05    7.74   12.63  20.30  46.07  2340    1
theta[1]     0.44    0.00  0.12    0.19    0.36    0.44   0.52   0.69 17667    1
theta[2]     0.44    0.00  0.12    0.19    0.36    0.44   0.52   0.69 16204    1
theta[3]     0.49    0.00  0.10    0.30    0.42    0.49   0.56   0.70 14599    1
theta[4]     0.45    0.00  0.10    0.26    0.38    0.45   0.52   0.66 18698    1
theta[5]     0.41    0.00  0.12    0.17    0.34    0.42   0.49   0.65 12542    1
theta[6]     0.41    0.00  0.10    0.22    0.34    0.41   0.47   0.59 14852    1
theta[7]     0.39    0.00  0.12    0.15    0.32    0.40   0.47   0.62 10969    1
theta[8]     0.47    0.00  0.12    0.24    0.39    0.46   0.54   0.72 17029    1
theta[9]     0.49    0.00  0.12    0.28    0.42    0.49   0.57   0.76 11266    1
theta[10]    0.46    0.00  0.10    0.27    0.39    0.45   0.52   0.66 18405    1
theta[11]    0.39    0.00  0.11    0.17    0.32    0.39   0.46   0.59 10145    1
theta[12]    0.49    0.00  0.10    0.31    0.43    0.49   0.55   0.69 14431    1
theta[13]    0.51    0.00  0.11    0.32    0.44    0.51   0.58   0.73  9785    1
theta[14]    0.41    0.00  0.11    0.18    0.34    0.41   0.48   0.62 12153    1
theta[15]    0.39    0.00  0.11    0.17    0.32    0.39   0.46   0.59 10404    1
theta[16]    0.36    0.00  0.11    0.12    0.29    0.37   0.44   0.57  6560    1
theta[17]    0.47    0.00  0.12    0.24    0.39    0.46   0.54   0.72 15400    1
theta[18]    0.44    0.00  0.10    0.24    0.37    0.44   0.50   0.64 18382    1
theta[19]    0.41    0.00  0.10    0.21    0.35    0.42   0.48   0.61 16023    1
```

stan

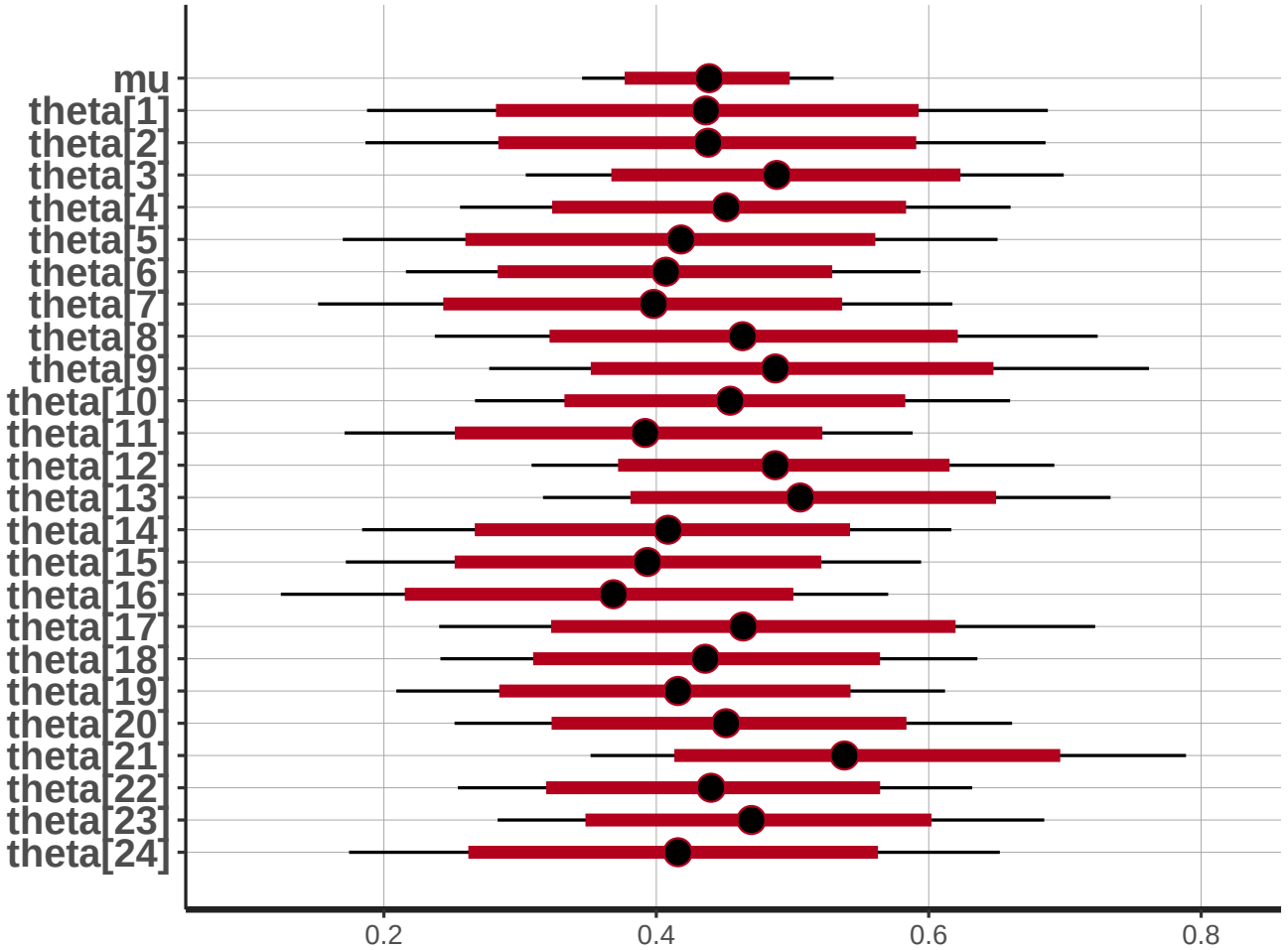
```
plot(r, pars=c('eta','alpha','beta'))
```

```
ci_level:  0.8 (80% intervals)
outer_level:  0.95 (95% intervals)
```

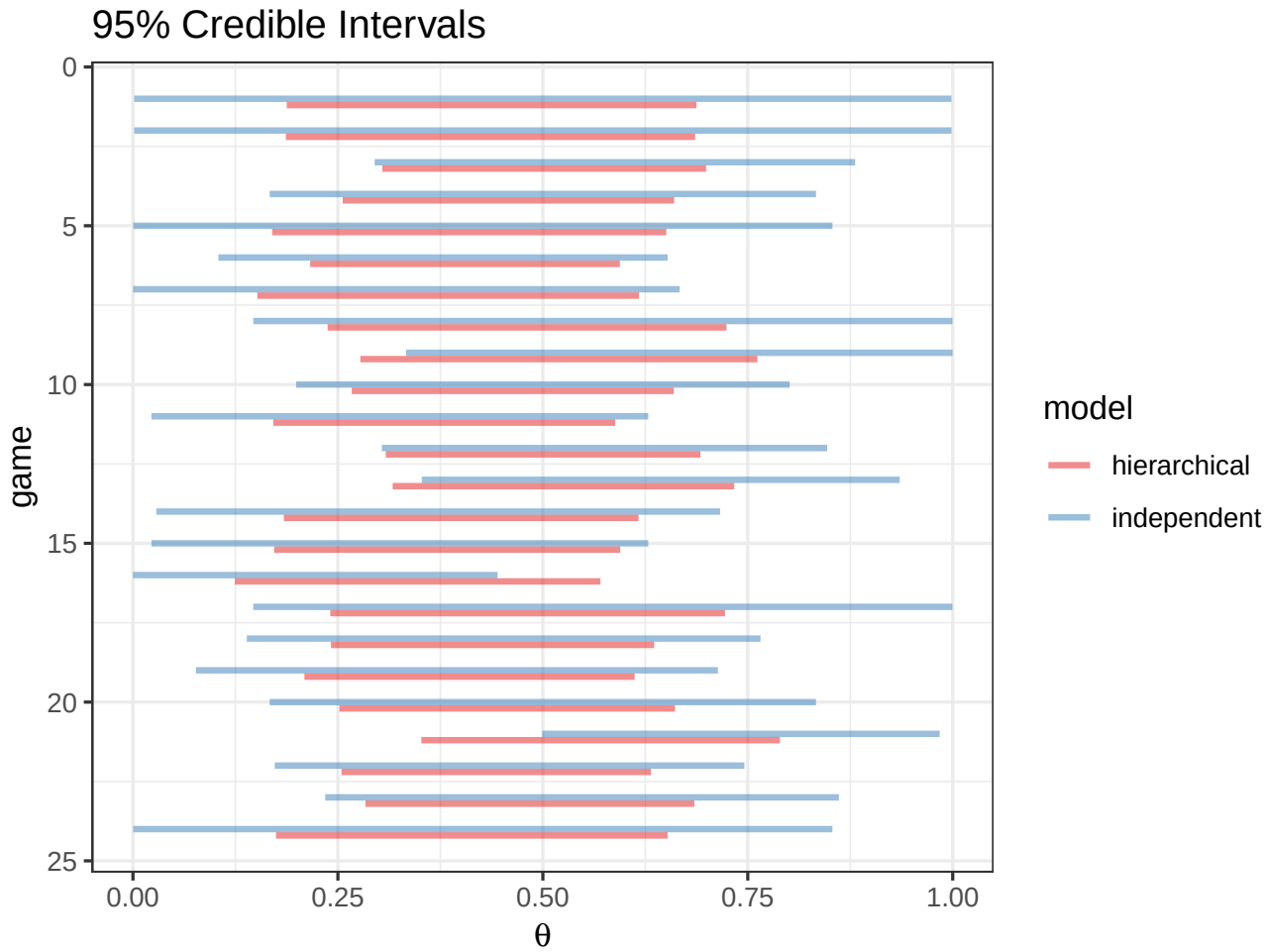


stan

```
plot(r, pars=c('mu','theta'))
```



Comparing independent and hierarchical models



A prior distribution for α and β

In Bayesian Data Analysis (3rd ed) page 110, several priors are discussed

- $(\log(\alpha/\beta), \log(\alpha+\beta)) \propto 1$ leads to an improper posterior.
- $(\log(\alpha/\beta), \log(\alpha+\beta)) \sim Unif([-10^{10}, 10^{10}] \times [-10^{10}, 10^{10}])$ while proper and seemingly vague is a very informative prior.
- $(\log(\alpha/\beta), \log(\alpha+\beta)) \propto \alpha\beta(\alpha+\beta)^{-5/2}$ which leads to a proper posterior and is equivalent to $p(\alpha, \beta) \propto (\alpha+\beta)^{-5/2}$.

Stan - default prior

```
model_default_prior = "
data {
  int<lower=0> N;
  int<lower=0> n[N];
  int<lower=0> y[N];
}
parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
  real<lower=0,upper=1> theta[N];
}

model {
  // default prior
  target += -5*log(alpha+beta)/2;

  // implicit joint distributions
  theta ~ beta(alpha,beta);
  y     ~ binomial(n,theta);
}
"

m2 = stan_model(model_code=model_default_prior)
r2 = sampling(m2, dat, c("alpha","beta","theta"), iter=10000,
              control = list(adapt_delta = 0.9))
```

```
Warning:  There were 2145 divergent transitions after warmup.  Increasing adapt_delta above 0.9 may help.
See
http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
Warning:  There were 4 chains where the estimated Bayesian Fraction of Missing Information was low.  See
```

Marginal posterior for α, β

An alternative to jointly sampling θ, α, β is to

1. sample $\alpha, \beta \sim p(\alpha, \beta|y)$, and then
2. sample $\theta_i \overset{ind}{\sim} p(\theta_i|\alpha, \beta, y_i) \overset{d}{=} Be(\alpha + y_i, \beta + n_i - y_i)$.

The maginal posterior for α, β is

$$p(\alpha, \beta|y) \propto p(y|\alpha, \beta)p(\alpha, \beta) = \left[\prod_{i=1}^{n} \text{Beta-binomial}(y_i|n_i, \alpha, \beta)\right] p(\alpha, \beta)$$

Stan - beta-binomial

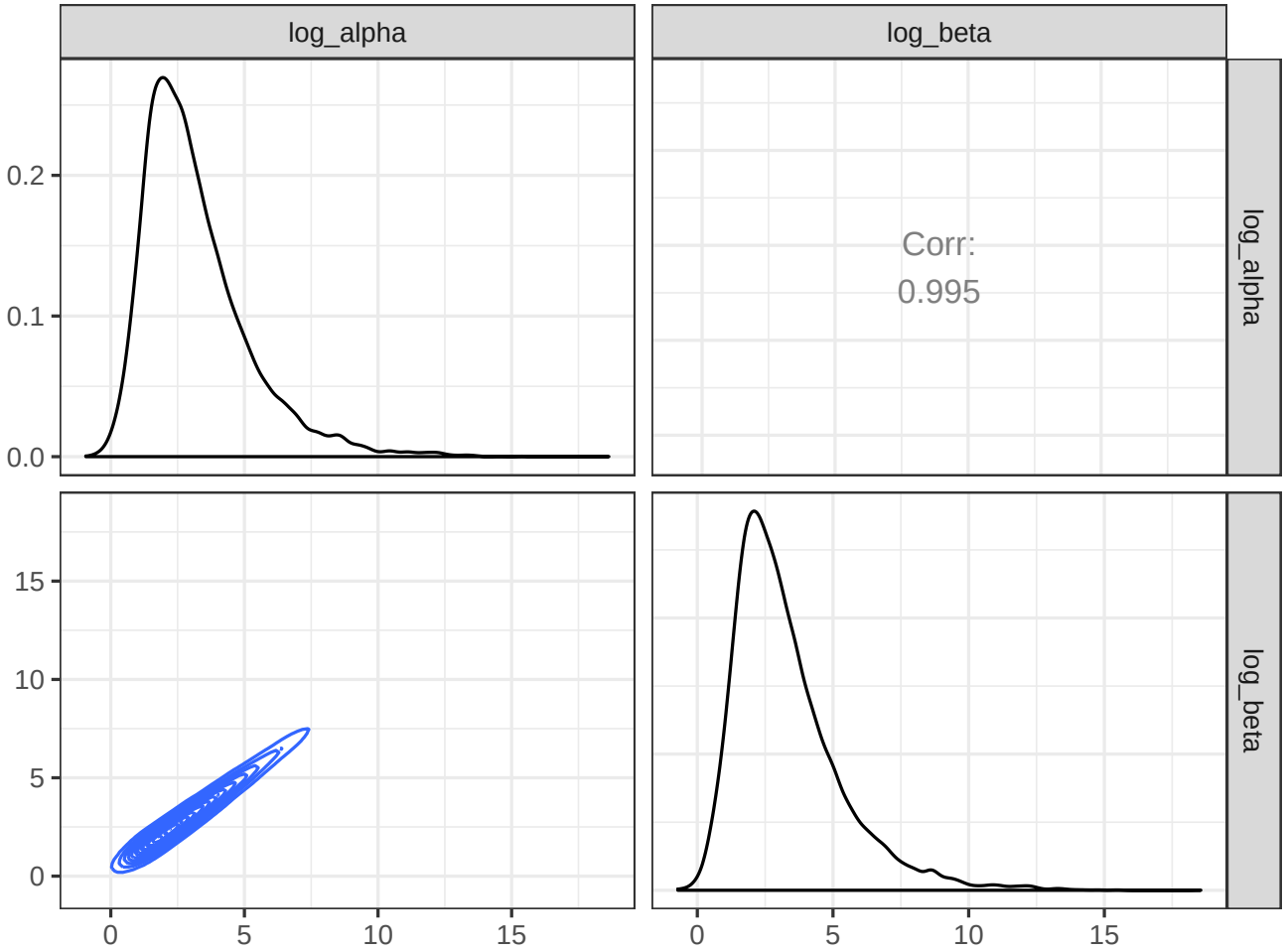
```
# Marginalized (integrated) theta out of the model
model_marginalized = "
data {
  int<lower=0> N;
  int<lower=0> n[N];
  int<lower=0> y[N];
}
parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
}
model {
  target += -5*log(alpha+beta)/2;
  y      ~ beta_binomial(n,alpha,beta);
}
generated quantities {
  real<lower=0,upper=1> theta[N];
  for (i in 1:N)
    theta[i] = beta_rng(alpha+y[i],beta+n[i]-y[i]);
}
"

m3 = stan_model(model_code=model_marginalized)
r3 = sampling(m3, dat, iter = 10000)
```

Stan - beta-binomial

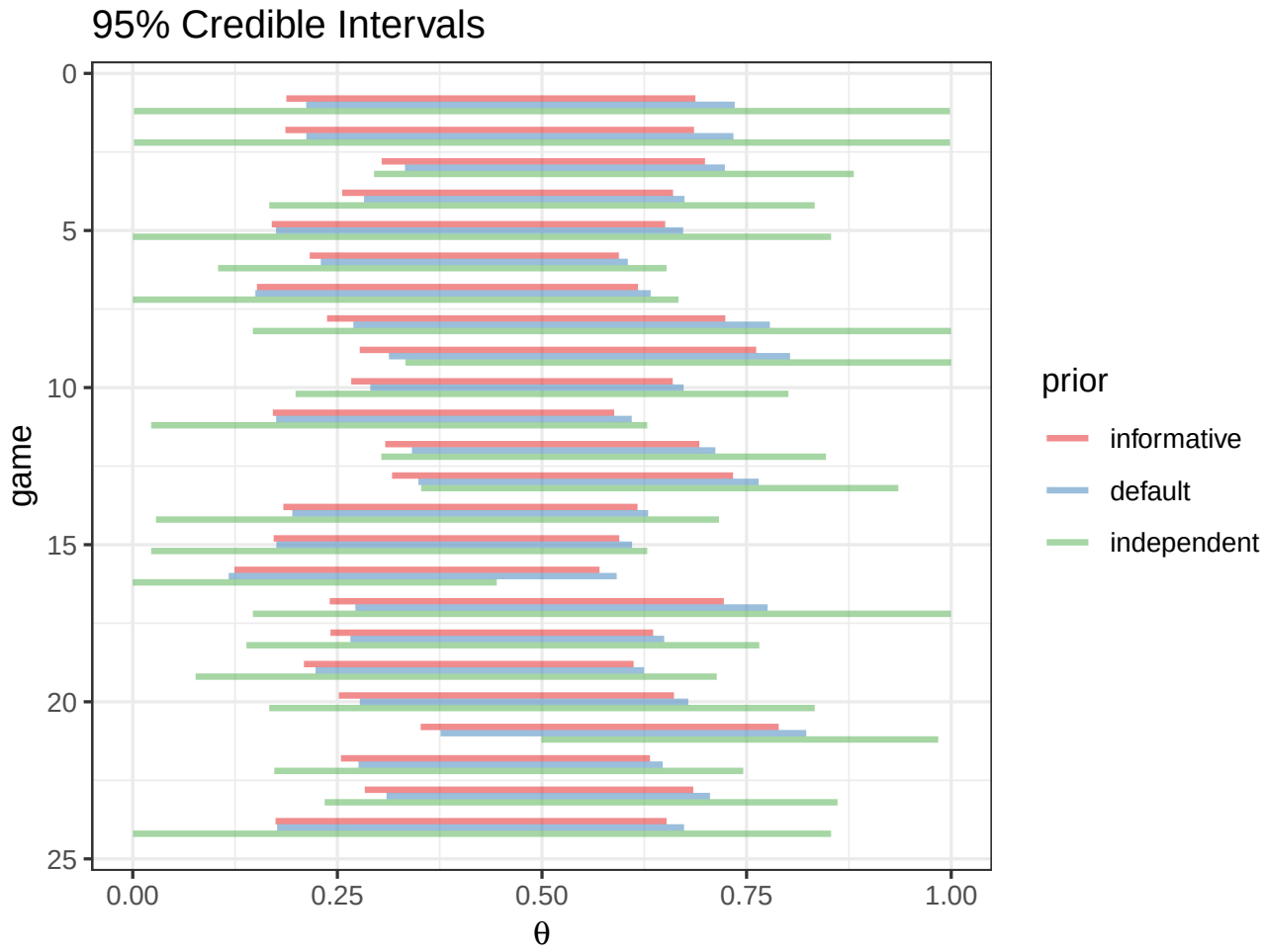
```
Inference for Stan model: 690888f74d999a6a8ba05f22d06a00df.
4 chains, each with iter=10000; warmup=5000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=20000.

             mean se_mean        sd   2.5%    25%    50%    75%   97.5% n_eff Rhat
alpha     9292.69 6582.63 883453.80   1.80   6.04  15.69  59.93 4440.68 18012    1
beta      8972.73 6099.40 813300.64   2.05   6.93  17.57  66.53 5011.78 17780    1
theta[1]     0.47    0.00      0.12   0.21   0.41   0.47   0.54    0.74 18590    1
theta[2]     0.47    0.00      0.12   0.21   0.41   0.47   0.53    0.73 18123    1
theta[3]     0.51    0.00      0.10   0.33   0.44   0.50   0.56    0.72 15106    1
theta[4]     0.48    0.00      0.10   0.28   0.42   0.47   0.53    0.67 19252    1
theta[5]     0.45    0.00      0.12   0.18   0.39   0.46   0.52    0.67 14645    1
theta[6]     0.44    0.00      0.09   0.23   0.38   0.44   0.50    0.60 12404    1
theta[7]     0.43    0.00      0.12   0.15   0.37   0.44   0.50    0.63  9920    1
theta[8]     0.50    0.00      0.12   0.27   0.43   0.49   0.55    0.78 15303    1
theta[9]     0.52    0.00      0.12   0.31   0.44   0.50   0.58    0.80 11289    1
theta[10]    0.48    0.00      0.09   0.29   0.42   0.48   0.53    0.67 19366    1
theta[11]    0.42    0.00      0.11   0.18   0.36   0.44   0.49    0.61  9856    1
theta[12]    0.51    0.00      0.09   0.34   0.45   0.50   0.56    0.71 13818    1
theta[13]    0.52    0.00      0.10   0.35   0.46   0.51   0.58    0.76  8746    1
theta[14]    0.44    0.00      0.11   0.19   0.38   0.45   0.50    0.63 12348    1
theta[15]    0.42    0.00      0.11   0.18   0.36   0.44   0.49    0.61  9765    1
theta[16]    0.40    0.00      0.12   0.12   0.33   0.42   0.48    0.59  6552    1
theta[17]    0.50    0.00      0.12   0.27   0.43   0.49   0.56    0.78 15239    1
theta[18]    0.46    0.00      0.09   0.27   0.41   0.46   0.52    0.65 18245    1
theta[19]    0.44    0.00      0.10   0.22   0.39   0.45   0.51    0.62 13366    1
theta[20]    0.48    0.00      0.10   0.28   0.42   0.48   0.53    0.68 19259    1
theta[21]    0.56    0.00      0.12   0.38   0.47   0.53   0.62    0.82  6119    1
theta[22]    0.46    0.00      0.09   0.28   0.41   0.47   0.52    0.65 18251    1
theta[23]    0.49    0.00      0.10   0.31   0.43   0.49   0.55    0.71 16955    1
theta[24]    0.45    0.00      0.12   0.18   0.39   0.46   0.52    0.67 13191    1
```

Posterior samples for α and β

```
samples = extract(r3, c("alpha","beta"))

ggpairs(data.frame(log_alpha = log(as.numeric(samples$alpha)),
                   log_beta  = log(as.numeric(samples$beta))),
        lower  = list(continuous='density')) + theme_bw()
```

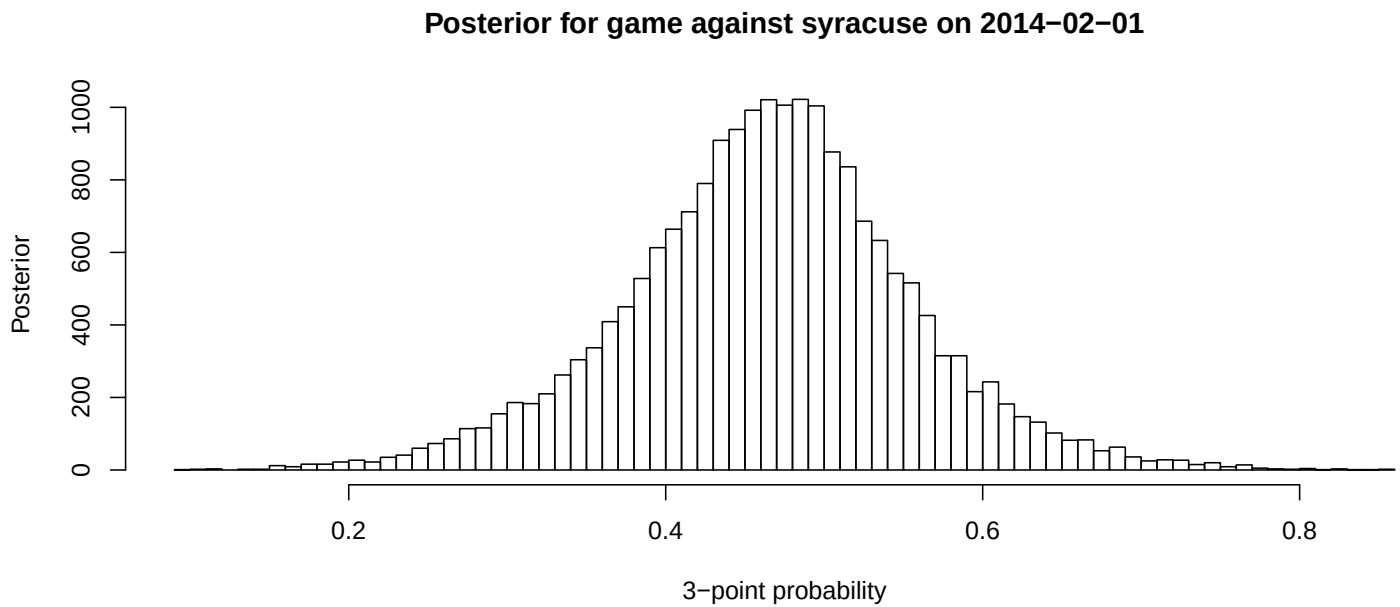


Comparing all models

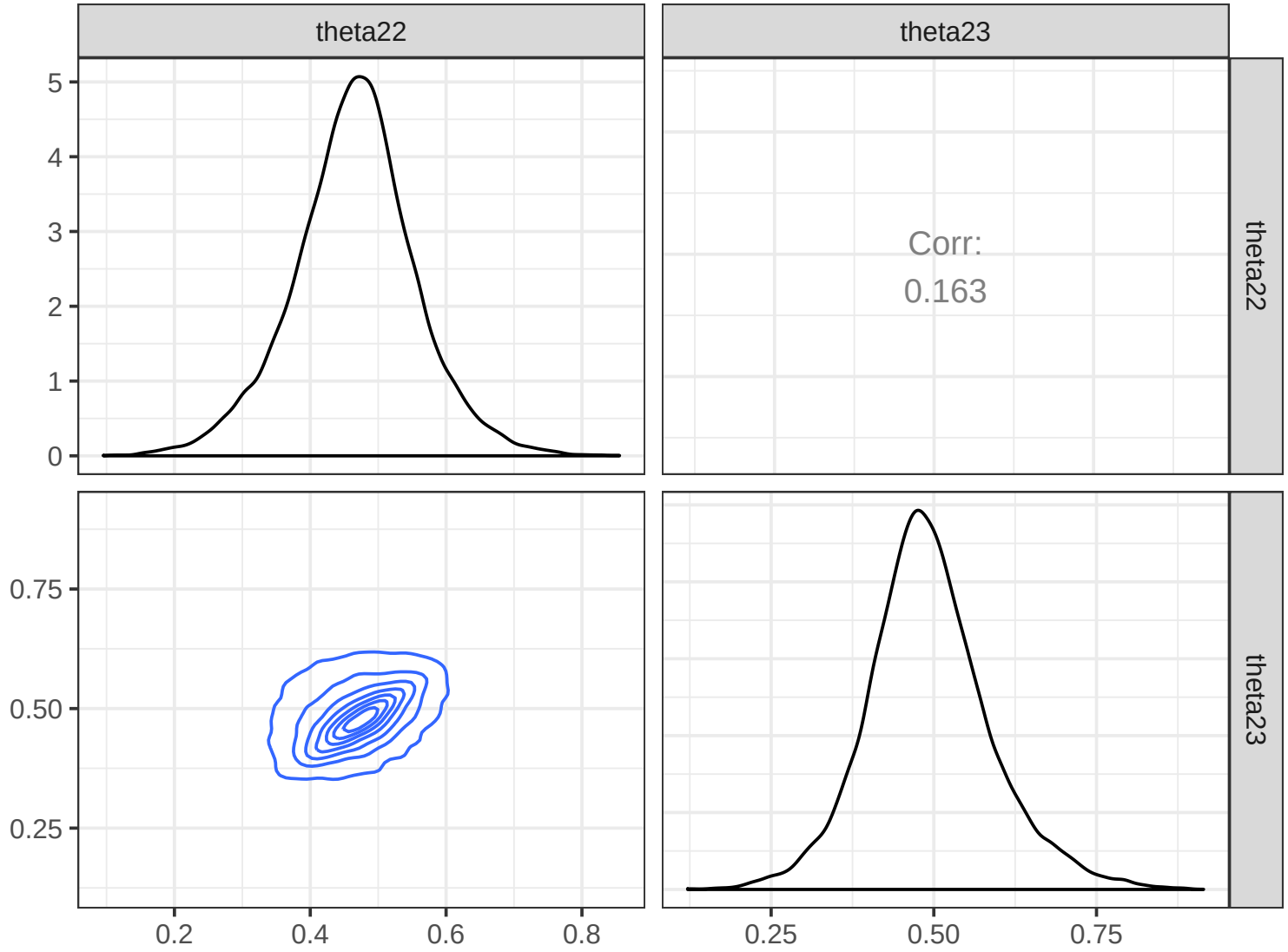


Posterior sample for θ_{22}

```
game = 22
theta22 = extract(r3, "theta")$theta[,game]

hist(theta22, 100,
     main=paste("Posterior for game against", d$opponent[game], "on", d$date[game]),
     xlab="3-point probability",
     ylab="Posterior")
```



θs are not independent in the posterior



3-point percentage across seasons

An alternative to modeling game-specific 3-point percentage is to model 3-point percentage in a season. The model is exactly the same, but the data changes.

	season	y	n
1	1	36	95
2	2	64	150
3	3	67	171
4	4	64	152

Due to the low number of seasons (observations), we will use an informative prior for α and β.

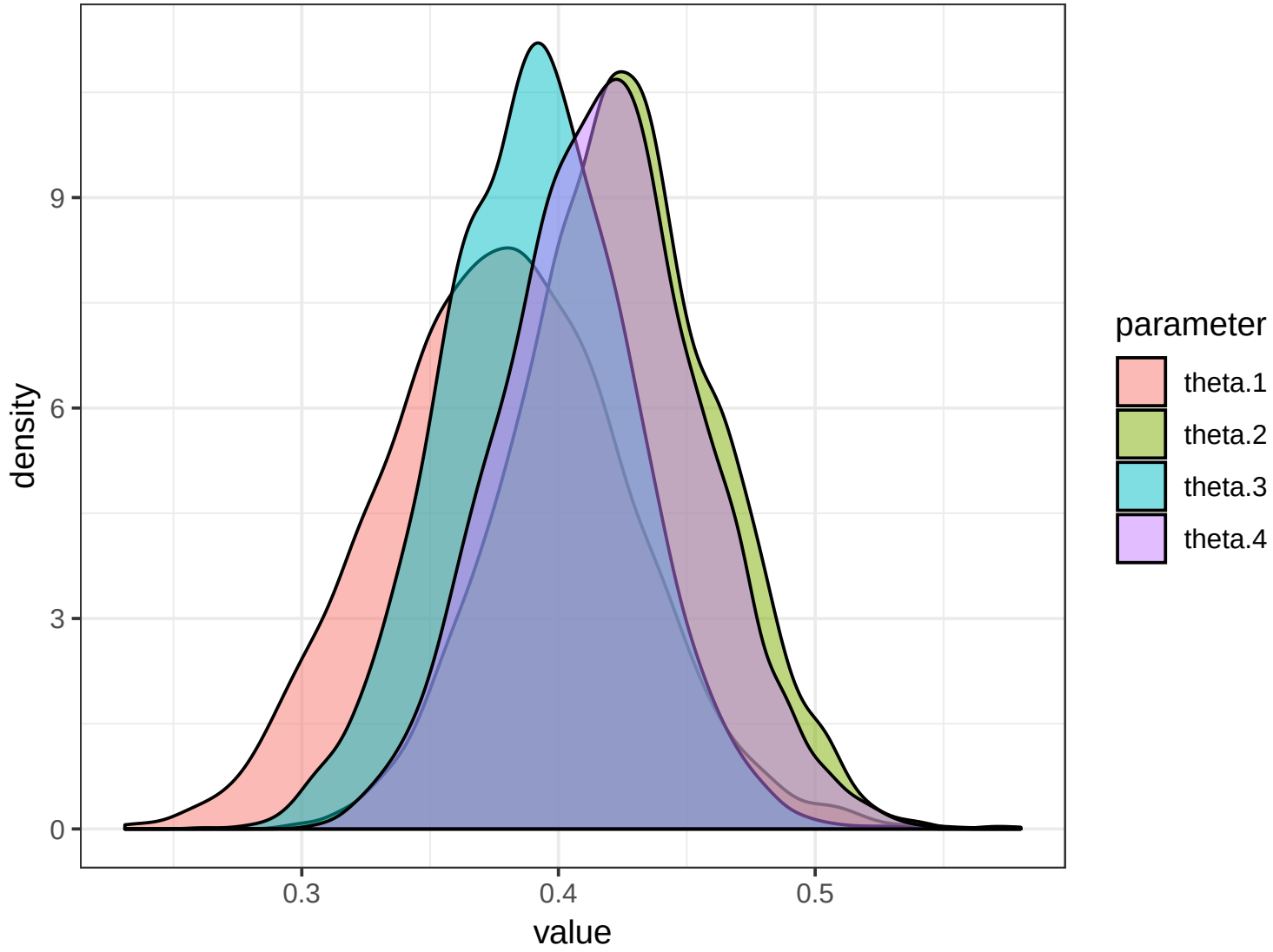
Stan - beta-binomial

```
model_seasons = "
data {
  int<lower=0> N; int<lower=0> n[N]; int<lower=0> y[N];
  real<lower=0> a; real<lower=0> b; real<lower=0> e;
}
parameters {
  real<lower=0,upper=1> mu;
  real<lower=0> eta;
}
transformed parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
  alpha = eta *    mu;
  beta  = eta * (1-mu);
}
model {
  mu  ~ beta(a,b);
  eta ~ exponential(e);
  y   ~ beta_binomial(n,alpha,beta);
}
generated quantities {
  real<lower=0,upper=1> theta[N];
  for (i in 1:N) theta[i] = beta_rng(alpha+y[i], beta+n[i]-y[i]);
}
"

dat  = list(N = nrow(d), y = d$y, n = d$n, a = a, b = b, e = e)
m4 = stan_model(model_code = model_seasons)
r_seasons = sampling(m4, dat,
                     c("alpha","beta","mu","eta","theta"))
```

Stan - hierarchical model for seasons

```
Inference for Stan model: add4563cfb5d1f9dead00bcc0fd9c410.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

           mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
alpha      4.87    0.06 3.00    0.96    2.72    4.25    6.34   12.30  2286    1
beta       7.89    0.08 4.40    1.76    4.69    7.04   10.23   18.68  2740    1
mu         0.38    0.00 0.06    0.25    0.34    0.38    0.42    0.50  2312    1
eta       12.76    0.14 7.23    2.84    7.53   11.33   16.44   30.86  2524    1
theta[1]   0.38    0.00 0.05    0.29    0.35    0.38    0.41    0.47  4117    1
theta[2]   0.42    0.00 0.04    0.35    0.40    0.42    0.45    0.50  4105    1
theta[3]   0.39    0.00 0.04    0.32    0.37    0.39    0.42    0.46  3655    1
theta[4]   0.42    0.00 0.04    0.35    0.39    0.42    0.44    0.49  3886    1
lp__    -402.05    0.03 1.06 -404.97 -402.48 -401.71 -401.29 -401.02  1430    1

Samples were drawn using NUTS(diag_e) at Wed Feb 20 11:32:10 2019.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

Stan - hierarchical model for seasons



Stan - hierarchical model for seasons

Probabilities that 3-point percentage is greater in season 4 than in the other seasons:

```
theta = extract(r_seasons, "theta")[[1]]
mean(theta[,4] > theta[,1])
```

```
[1] 0.7415
```

```
mean(theta[,4] > theta[,2])
```

```
[1] 0.461
```

```
mean(theta[,4] > theta[,3])
```

```
[1] 0.69775
```

Summary - hierarchical models

Two-level hierarchical model:

$$y_i \overset{ind}{\sim} p(y|\theta) \qquad \theta_i \overset{ind}{\sim} p(\theta|\phi) \qquad \phi \sim p(\phi)$$

Conditional independencies:

- $y_i \perp\!\!\!\perp y_j|\theta$ for $i \neq j$
- $\theta_i \perp\!\!\!\perp \theta_j|\phi$ for $i \neq j$
- $y \perp\!\!\!\perp \phi|\theta$
- $y_i \perp\!\!\!\perp y_j|\phi$ for $i \neq j$
- $\theta_i \perp\!\!\!\perp \theta_j|\phi, y$ for $i \neq j$

Summary - extension to more levels

Three-level hierarchical model:

$$y \sim p(y|\theta) \qquad \theta \sim p(\theta|\phi) \qquad \phi \sim p(\phi|\psi) \qquad \psi \sim p(\psi)$$

When deriving posteriors, remember the conditional independence structure, e.g.

$$p(\theta, \phi, \psi|y) \propto p(y|\theta)p(\theta|\phi)p(\phi|\psi)p(\psi)$$