Hierarchical models

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 21, 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

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

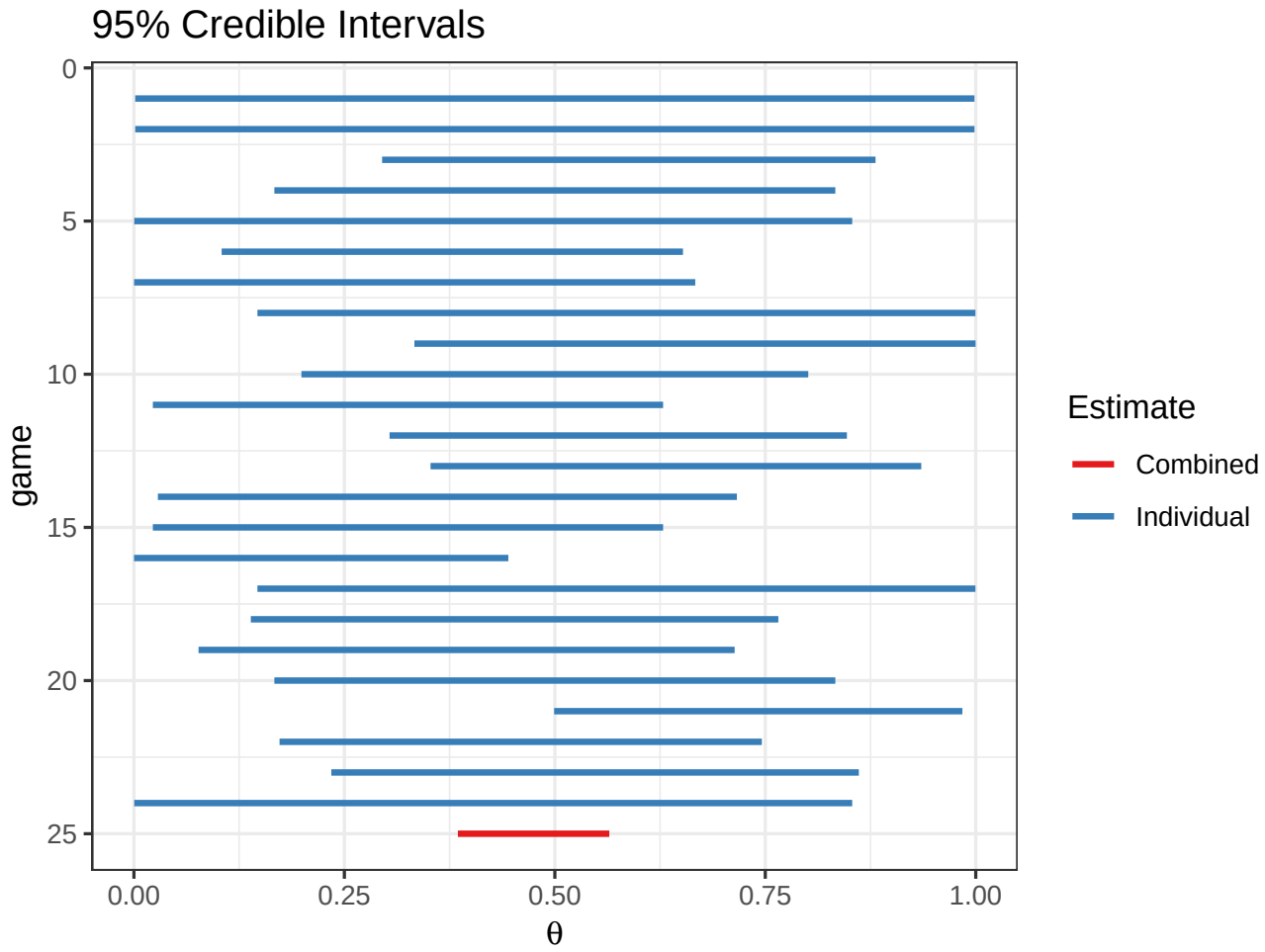
where

- 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	game
1	16017.00	davidson	0	0	1
2	16021.00	kansas	0	0	2
3	16024.00	florida atlantic	5	8	3
4	16027.00	unc asheville	3	6	4
5	16028.00	east carolina	0	1	5
6	16033.00	vermont	3	9	6
7	16036.00	alabama	0	2	7
8	16038.00	arizona	1	1	8
9	16042.00	michigan	2	2	9
10	16055.00	gardner-webb	4	8	10
11	16058.00	ucla	1	5	11
12	16067.00	eastern michigan	6	10	12
13	16070.00	elon	5	7	13
14	16074.00	notre dame	1	4	14
15	16077.00	georgia tech	1	5	15
16	16081.00	clemson	0	4	16
17	16083.00	virginia	1	1	17
18	16088.00	nc state	3	7	18
19	16092.00	miami	2	6	19
20	16095.00	florida state	3	6	20
21	16097.00	pitt	6	7	21
22	16102.00	syracuse	4	9	22
23	16105.00	wake forest	4	7	23
24	16109.00	boston college	0	1	24

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):

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

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)
```

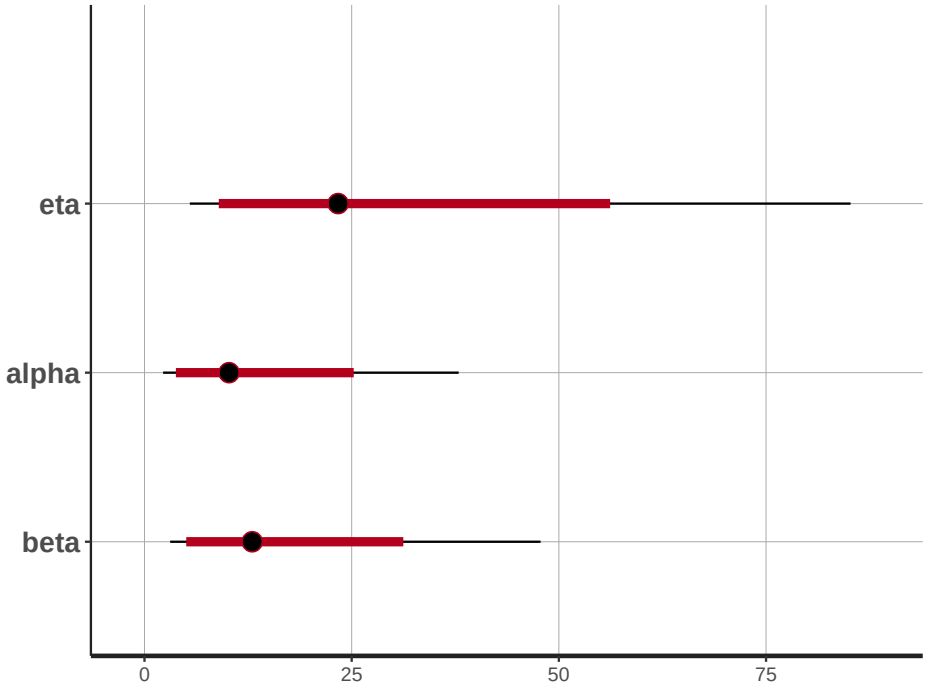
stan

```
r


Inference for Stan model: 6ef4482c54275aff28bb77f1e2a57609.
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  5336 1.00
eta         29.07    0.43 21.28    5.47   14.08   23.36  37.93  85.19  2429 1.00
alpha       12.85    0.20  9.66    2.24    6.05   10.20  16.83  37.91  2439 1.00
beta        16.23    0.24 11.88    3.10    7.92   13.00  21.15  47.79  2473 1.00
theta[1]     0.44    0.00  0.12    0.19    0.36    0.44   0.52   0.69 15584 1.00
theta[2]     0.44    0.00  0.12    0.20    0.36    0.44   0.51   0.69 16099 1.00
theta[3]     0.49    0.00  0.10    0.31    0.42    0.49   0.55   0.70 11624 1.00
theta[4]     0.45    0.00  0.10    0.26    0.38    0.45   0.52   0.66 15187 1.00
theta[5]     0.42    0.00  0.12    0.17    0.34    0.42   0.49   0.65 12006 1.00
theta[6]     0.41    0.00  0.09    0.22    0.34    0.41   0.47   0.59 12370 1.00
theta[7]     0.40    0.00  0.12    0.15    0.32    0.40   0.47   0.61 10375 1.00
theta[8]     0.47    0.00  0.12    0.24    0.39    0.46   0.54   0.72 15045 1.00
theta[9]     0.49    0.00  0.12    0.28    0.41    0.49   0.56   0.74 11403 1.00
theta[10]    0.46    0.00  0.10    0.27    0.39    0.45   0.52   0.66 15151 1.00
theta[11]    0.39    0.00  0.11    0.17    0.32    0.39   0.46   0.60 10054 1.00
theta[12]    0.49    0.00  0.10    0.31    0.42    0.49   0.55   0.69 13547 1.00
theta[13]    0.51    0.00  0.10    0.32    0.44    0.50   0.58   0.74  9692 1.00
theta[14]    0.41    0.00  0.11    0.19    0.34    0.41   0.48   0.62 12281 1.00
theta[15]    0.39    0.00  0.11    0.17    0.32    0.39   0.46   0.59 10039 1.00
theta[16]    0.37    0.00  0.11    0.13    0.29    0.37   0.44   0.57  7285 1.00
theta[17]    0.47    0.00  0.12    0.24    0.39    0.46   0.54   0.72 15088 1.00
theta[18]    0.44    0.00  0.10    0.24    0.37    0.44   0.50   0.63 14782 1.00
theta[19]    0.42    0.00  0.10    0.21    0.35    0.42   0.48   0.62 12728 1.00
```

stan

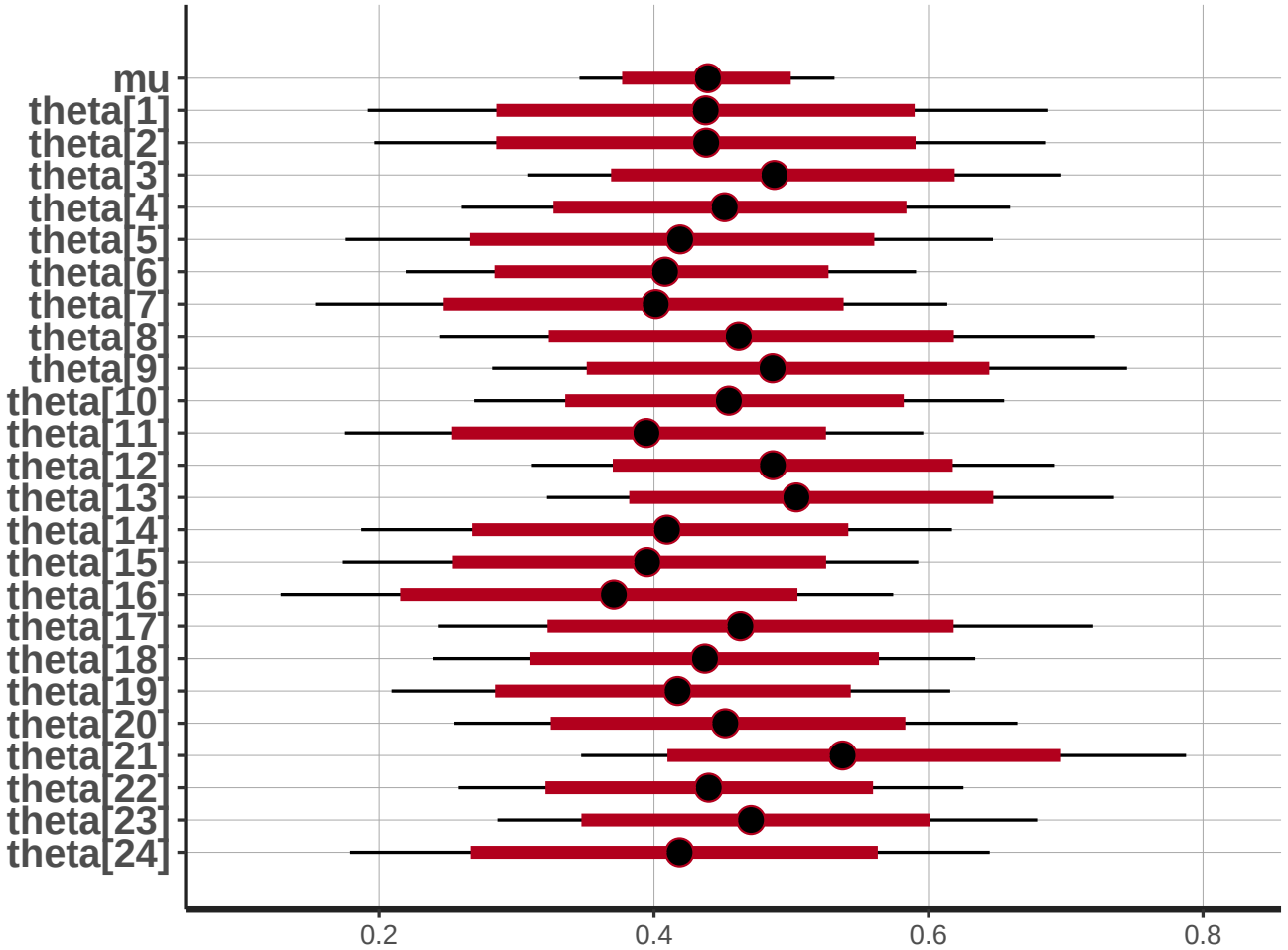
```
plot(r, pars=c('eta','alpha','beta'))
```

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

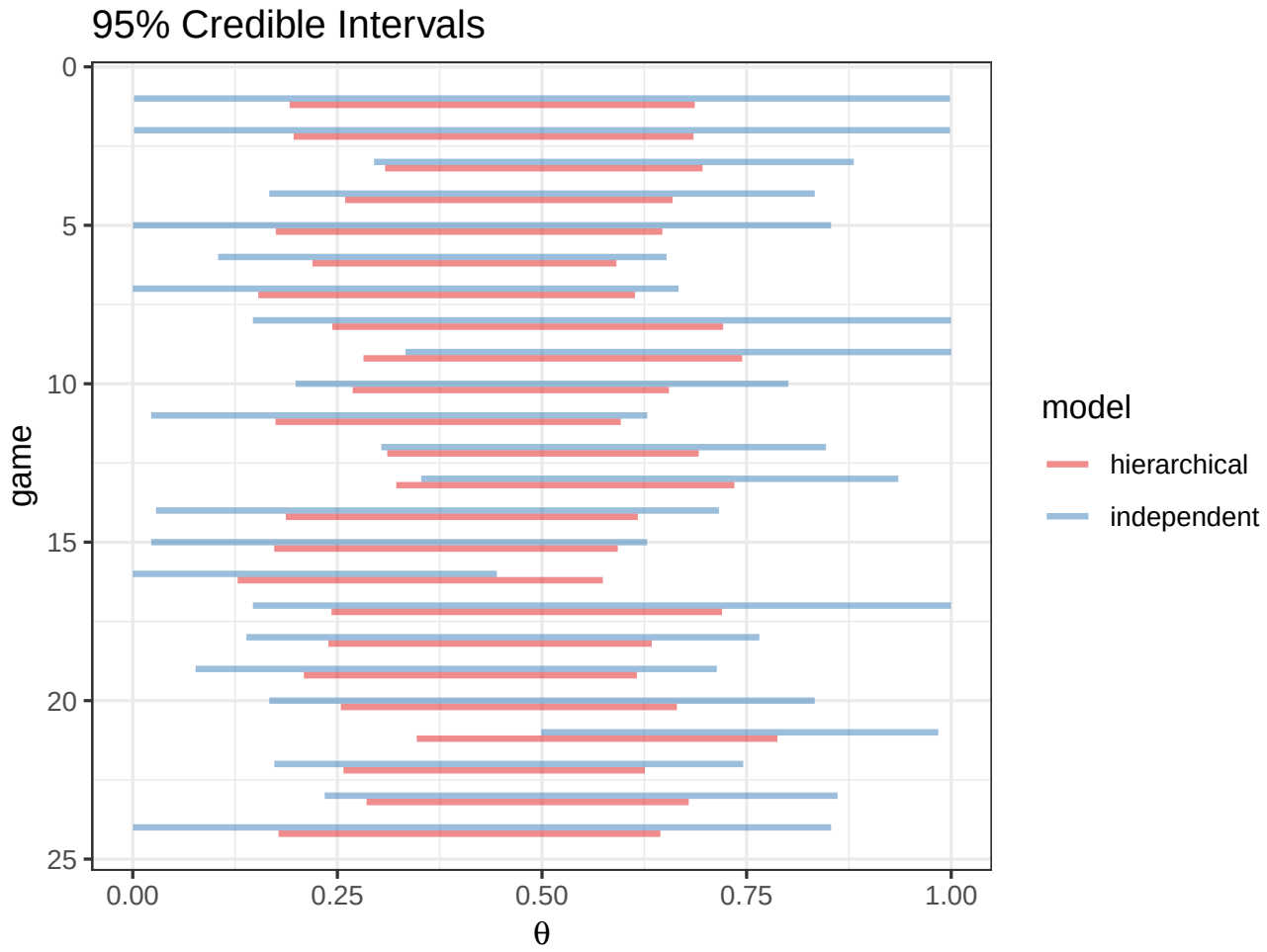


stan

```
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 1298 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 2 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)
```

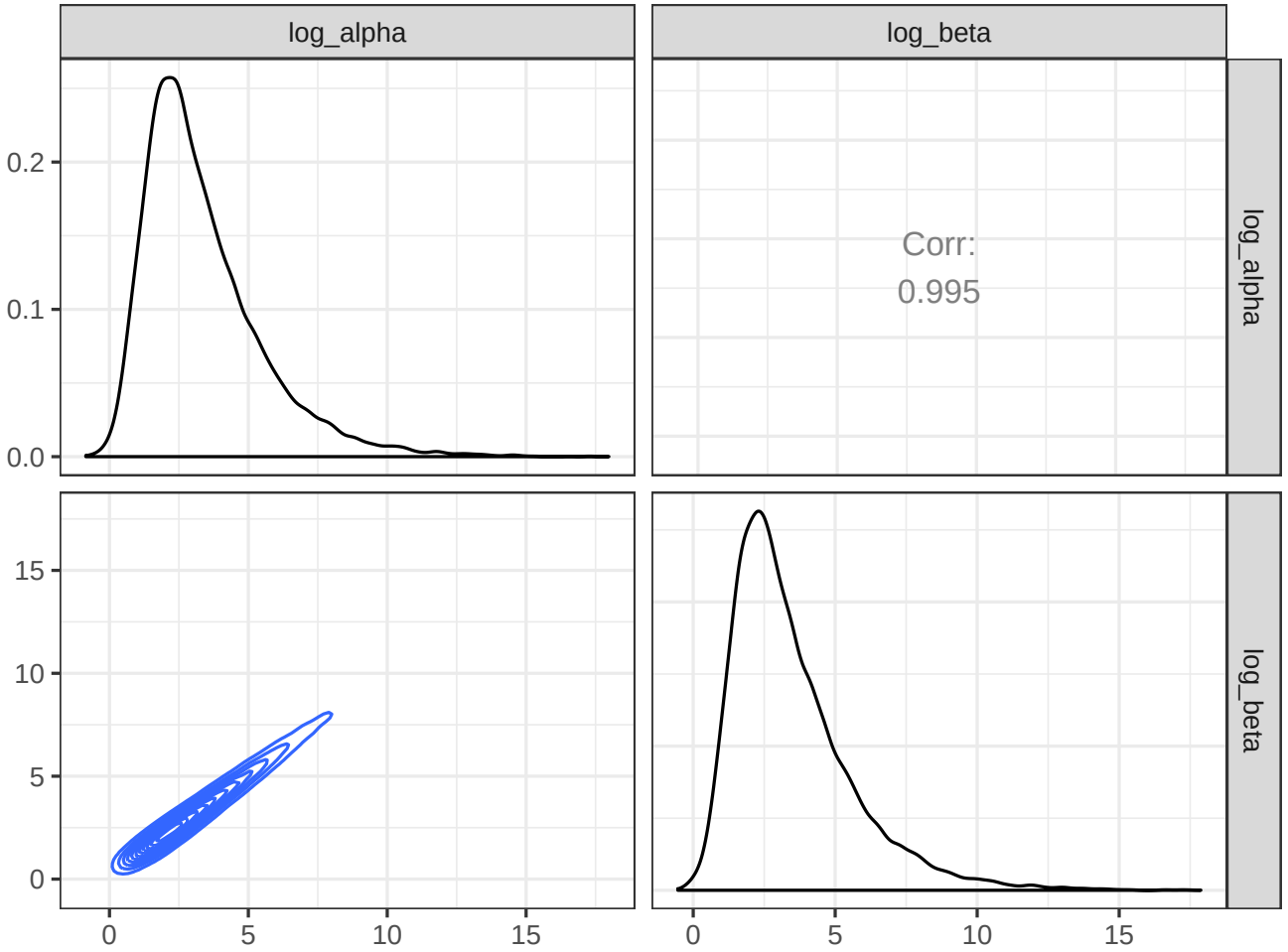
Stan - beta-binomial

```
Inference for Stan model: ac767356ee5840533dc9fd89cf9c2a24.
4 chains, each with iter=10000; warmup=5000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=20000.
```

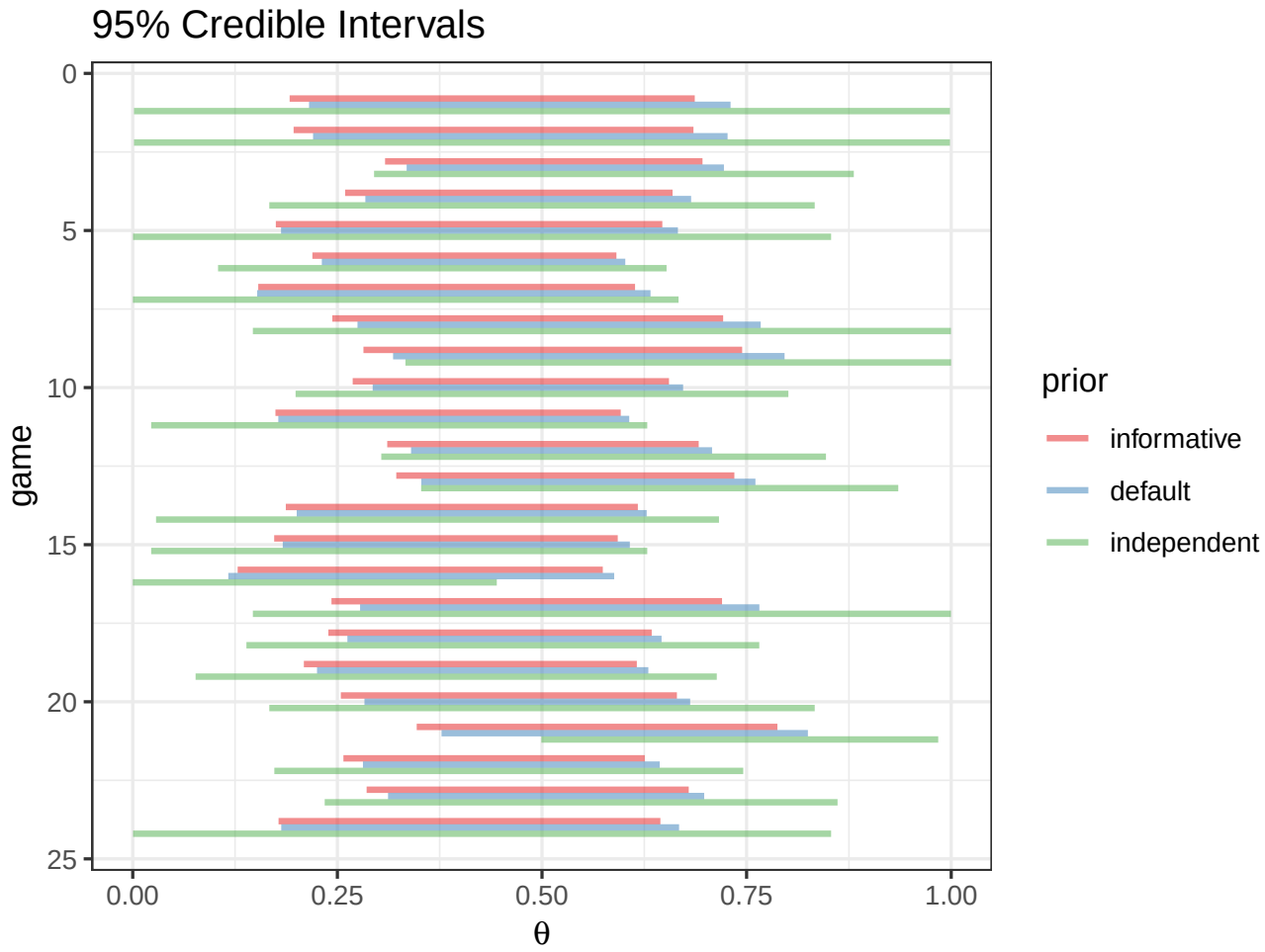
	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
alpha	17745.09	12699.07	689851.55	1.85	6.47	17.63	77.59	6787.77	2951	1
beta	17934.93	12603.38	664367.11	2.15	7.34	19.61	86.32	7616.62	2779	1
theta[1]	0.47	0.00	0.12	0.22	0.41	0.47	0.53	0.73	19930	1
theta[2]	0.47	0.00	0.12	0.22	0.41	0.47	0.53	0.73	19708	1
theta[3]	0.51	0.00	0.09	0.33	0.44	0.50	0.56	0.72	12260	1
theta[4]	0.48	0.00	0.10	0.28	0.42	0.47	0.53	0.68	19717	1
theta[5]	0.45	0.00	0.11	0.18	0.39	0.46	0.52	0.67	13765	1
theta[6]	0.44	0.00	0.09	0.23	0.38	0.45	0.50	0.60	11960	1
theta[7]	0.43	0.00	0.11	0.15	0.37	0.45	0.50	0.63	10039	1
theta[8]	0.49	0.00	0.12	0.27	0.43	0.48	0.55	0.77	14971	1
theta[9]	0.51	0.00	0.12	0.32	0.44	0.50	0.57	0.80	10661	1
theta[10]	0.48	0.00	0.09	0.29	0.42	0.48	0.53	0.67	19633	1
theta[11]	0.43	0.00	0.11	0.18	0.37	0.44	0.50	0.61	9353	1
theta[12]	0.50	0.00	0.09	0.34	0.44	0.50	0.56	0.71	13105	1
theta[13]	0.52	0.00	0.10	0.35	0.45	0.51	0.58	0.76	9044	1
theta[14]	0.44	0.00	0.10	0.20	0.38	0.45	0.51	0.63	12829	1
theta[15]	0.43	0.00	0.10	0.18	0.37	0.44	0.50	0.61	9772	1
theta[16]	0.40	0.00	0.12	0.12	0.34	0.43	0.48	0.59	6242	1
theta[17]	0.50	0.00	0.11	0.28	0.43	0.49	0.55	0.77	16551	1
theta[18]	0.46	0.00	0.09	0.26	0.41	0.46	0.52	0.65	18959	1
theta[19]	0.45	0.00	0.10	0.23	0.39	0.45	0.51	0.63	14710	1
theta[20]	0.48	0.00	0.09	0.28	0.42	0.48	0.53	0.68	19886	1
theta[21]	0.55	0.00	0.12	0.38	0.47	0.53	0.62	0.82	5331	1
theta[22]	0.47	0.00	0.09	0.28	0.41	0.47	0.52	0.64	18780	1
theta[23]	0.49	0.00	0.09	0.31	0.43	0.49	0.54	0.70	17797	1
theta[24]	0.45	0.00	0.11	0.18	0.39	0.46	0.52	0.67	13735	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()
```

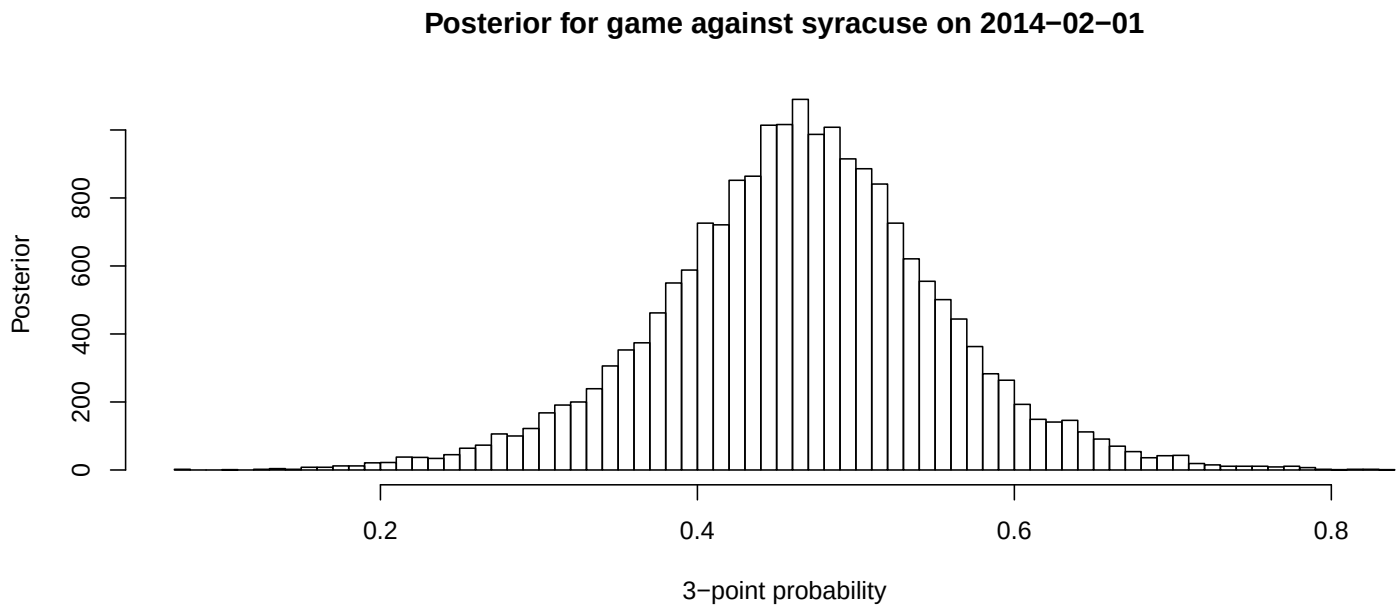


Comparing all models

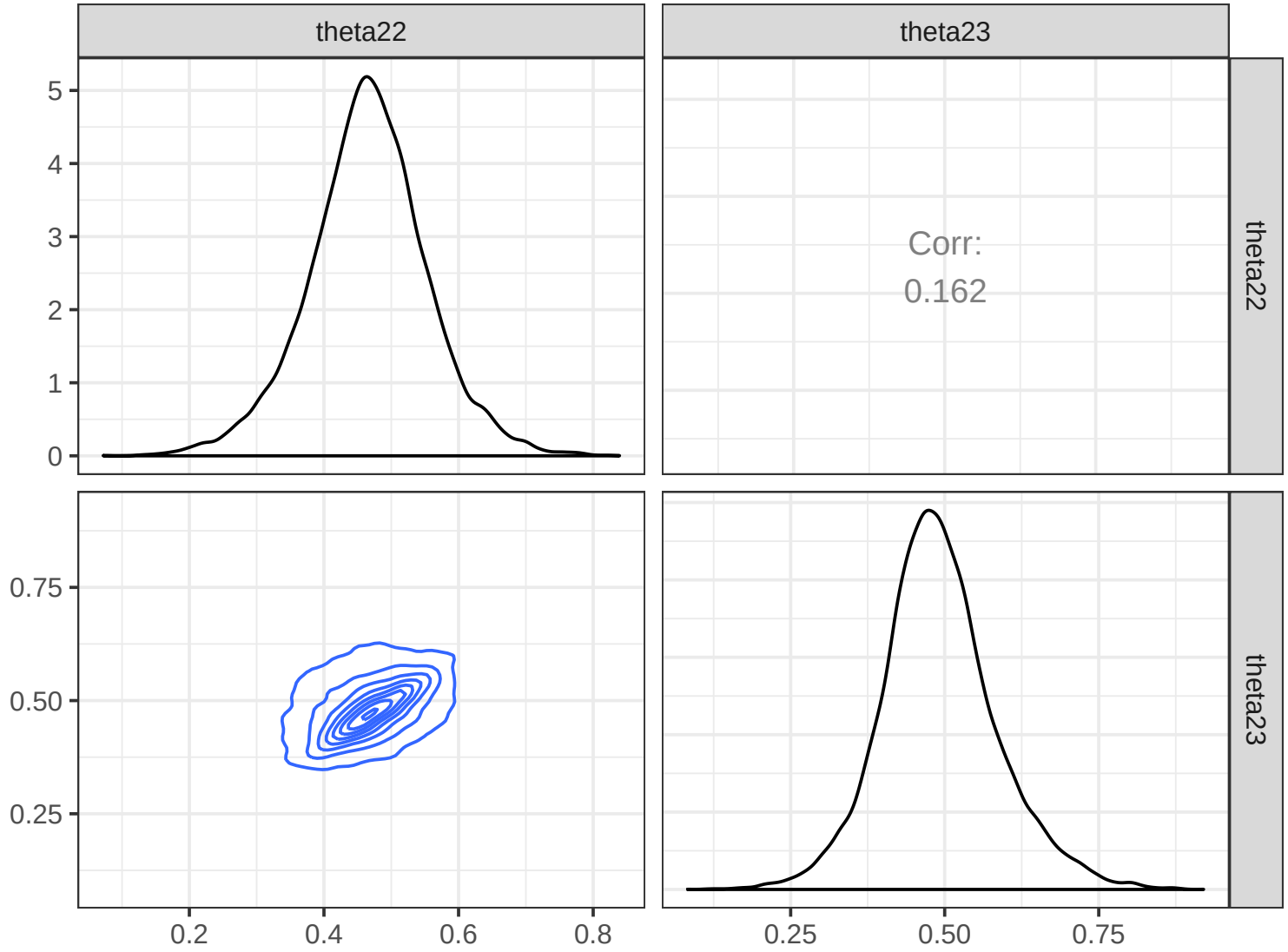


Posterior sample for θ_{22}

```
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.

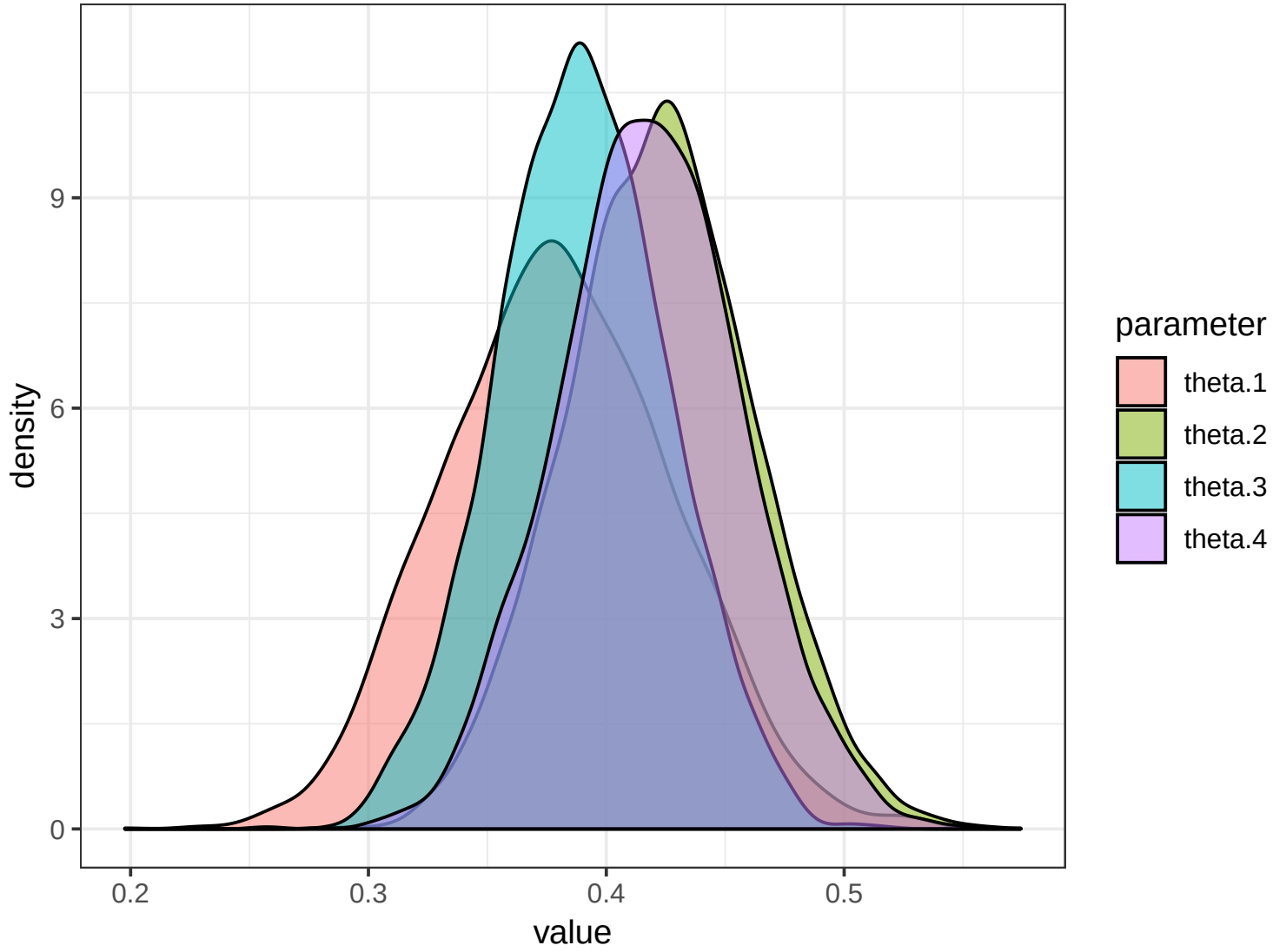
Stan - beta-binomial

```
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: 9c3c3b21e436bf9a6e70ee168d740430.
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      4.97    0.06 3.07    0.95    2.73    4.31    6.51   12.49  2390    1
beta       8.04    0.09 4.58    1.76    4.65    7.19   10.49   19.40  2470    1
mu         0.38    0.00 0.06    0.26    0.34    0.38    0.42    0.50  2772    1
eta       13.01    0.15 7.48    2.91    7.48   11.63   16.93   30.86  2411    1
theta[1]   0.38    0.00 0.05    0.29    0.35    0.38    0.41    0.47  4095    1
theta[2]   0.42    0.00 0.04    0.35    0.40    0.42    0.45    0.50  3780    1
theta[3]   0.39    0.00 0.04    0.32    0.37    0.39    0.41    0.46  3882    1
theta[4]   0.42    0.00 0.04    0.34    0.39    0.42    0.44    0.49  3906    1
lp__    -402.06    0.03 1.07 -404.99 -402.46 -401.73 -401.30 -401.03  1610    1

Samples were drawn using NUTS(diag_e) at Thu Feb 21 07:57:01 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] 0.73475
```

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

```
[1] 0.47
```

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

```
[1] 0.71125
```

Summary - hierarchical models

Two-level hierarchical model:

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

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.

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