Hierarchical models

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 19, 2024

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

Let $Y_s = \sum_{j=1}^{n_s} Y_{sj}$ be the number 3-pointers Andre Dawkin's makes in season s, and assume

$$Y_s \stackrel{ind}{\sim} Bin(n_s, \theta_s) \quad \text{or, equivalently,} \quad Y_{sj} \stackrel{ind}{\sim} Ber(\theta_s)\, j = 1, \ldots, n_s$$

where

- n_s are the number of 3-pointers attempted and
- θ_s 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. $\theta_s = \theta$.
- The 3-point percentage every season is independent of other seasons.
- The 3-point percentage a season should be similar to other seasons.

Andre Dawkin's three-point percentage

Let $Y_g = \sum_{j=1}^{n_g} Y_{gj}$ be the number of 3-pointers Andre Dawkin's makes in game g, and assume

$$Y_g \stackrel{ind}{\sim} Bin(n_g, \theta_g) \quad \text{or, equivalently,} \quad Y_{gj} \stackrel{ind}{\sim} Ber(\theta_g)\, j = 1, \ldots, n_g$$

where

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

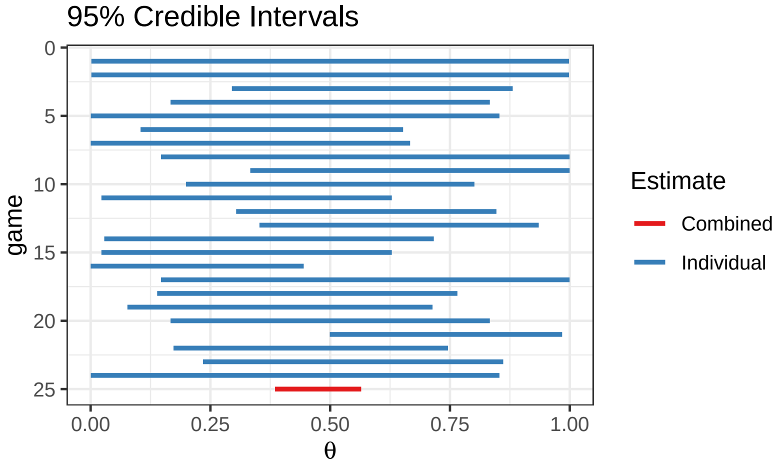
Do these models make sense?

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

Andre Dawkin's 3-point percentage

date	opponent	made	attempts	game
2013-11-08	davidson	0	0	1
2013-11-12	kansas	0	0	2
2013-11-15	florida atlantic	5	8	3
2013-11-18	unc asheville	3	6	4
2013-11-19	east carolina	0	1	5
2013-11-24	vermont	3	9	6
2013-11-27	alabama	0	2	7
2013-11-29	arizona	1	1	8
2013-12-03	michigan	2	2	9
2013-12-16	gardner-webb	4	8	10
2013-12-19	ucla	1	5	11
2013-12-28	eastern michigan	6	10	12
2013-12-31	elon	5	7	13
2014-01-04	notre dame	1	4	14
2014-01-07	georgia tech	1	5	15
2014-01-11	clemson	0	4	16
2014-01-13	virginia	1	1	17
2014-01-18	nc state	3	7	18
2014-01-22	miami	2	6	19
2014-01-25	florida state	3	6	20
2014-01-27	pitt	6	7	21
2014-02-01	syracuse	4	9	22
2014-02-04	wake forest	4	7	23
2014-02-08	boston college	0	1	24

Andre Dawkin's 3-point percentage



Hierarchical models

Consider the following model

$$\begin{aligned} y_{ij} &\stackrel{ind}{\sim} p(y|\theta_i) \\ \theta_i &\stackrel{ind}{\sim} p(\theta|\phi) \\ \phi &\sim p(\phi) \end{aligned}$$

where

- y_{ij} for $i = 1 \ldots, n_i$ are the observed data ,
- $\theta = (\theta_1, \ldots, \theta_n)$ 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

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

where $p(y_i|\theta_i) = \prod_{j=1}^{n_i} p(y_{ij}|\theta_i)$. The joint posterior distribution can be decomposed via

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

where

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

Three-pointer example

Let $Y_{i,g}$ be an indicator that 3-point attempt i in game g was successful for $i = 1, \ldots, n_g$ and $Y_g = \sum_{i=1}^{n_g} Y_{i,g}$. Assume

$$
\begin{array}{ll}
Y_{i,g} & \stackrel{ind}{\sim} Ber(\theta_g) \quad \text{or, equivalently} \quad Y_g \stackrel{ind}{\sim} Bin(n_g, \theta_g) \\
\theta_g & \stackrel{ind}{\sim} Be(\alpha, \beta) \\
\alpha, \beta & \sim p(\alpha, \beta)
\end{array}
$$

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

$$Y_g \stackrel{ind}{\sim} Bin(n_g, \theta_g) \quad \theta_g \stackrel{ind}{\sim} Be(\alpha, \beta) \quad \alpha, \beta \sim p(\alpha, \beta)$$

Conditional posterior for θ:

$$p(\theta|\alpha, \beta, y) = \prod_{i=1}^{n} p(\theta_g|\alpha, \beta, y_g) = \prod_{i=1}^{n} Be(\theta_g|\alpha + y_g, \beta + n_g - y_g)$$

Marginal posterior for (α, β):

$$\begin{array}{rl} p(\alpha, \beta|y) & \propto p(y|\alpha, \beta)p(\alpha, \beta) \\ p(y|\alpha, \beta) & = \prod_{i=1}^{n} p(y_g|\alpha, \beta) = \prod_{i=1}^{n} \int p(y_g|\theta_g)p(\theta_g|\alpha, \beta)d\theta_g \\ & = \prod_{i=1}^{n} \binom{n_g}{y_g} \frac{B(\alpha+y_g, \beta+n_g-y_g)}{B(\alpha, \beta)} \end{array}$$

Thus $y_g|\alpha, \beta \stackrel{ind}{\sim}$ Beta-binomial(n_g, 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 expectation (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> G;    // data
  int<lower=0> n[G];
  int<lower=0> y[G];
  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[G];
}
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);
}
"
```

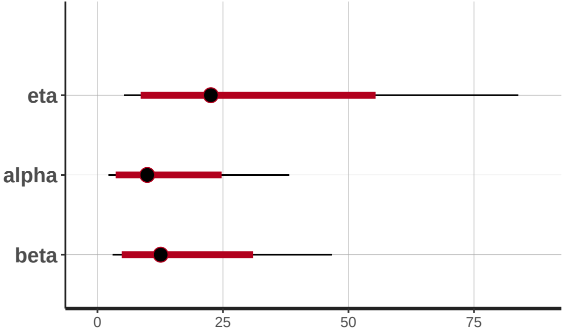
Stan

```
dat = list(y = d$made, n = d$attempts, G = 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)
```

```
r

Inference for Stan model: anon_model.
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  0.34  0.41  0.44  0.47  0.53  5429    1
eta       28.37    0.44 21.02  5.28 13.63 22.59 36.84 83.82  2315    1
alpha     12.55    0.20  9.54  2.18  5.86  9.92 16.31 38.22  2318    1
beta      15.82    0.24 11.73  3.02  7.62 12.60 20.47 46.71  2356    1
theta[1]   0.44    0.00  0.12  0.19  0.36  0.44  0.52  0.70 14481    1
theta[2]   0.44    0.00  0.12  0.19  0.36  0.44  0.51  0.69 14333    1
theta[3]   0.49    0.00  0.10  0.31  0.43  0.49  0.56  0.70 14108    1
theta[4]   0.45    0.00  0.10  0.26  0.39  0.45  0.52  0.66 17874    1
theta[5]   0.42    0.00  0.12  0.17  0.34  0.42  0.49  0.65 12842    1
theta[6]   0.41    0.00  0.10  0.22  0.34  0.41  0.47  0.60 13657    1
theta[7]   0.40    0.00  0.12  0.15  0.32  0.40  0.47  0.62 10358    1
theta[8]   0.47    0.00  0.12  0.24  0.39  0.47  0.54  0.73 15136    1
theta[9]   0.49    0.00  0.12  0.28  0.41  0.49  0.57  0.76 11804    1
theta[10]  0.46    0.00  0.10  0.27  0.39  0.46  0.52  0.66 16617    1
theta[11]  0.39    0.00  0.11  0.17  0.32  0.39  0.46  0.59  9644    1
theta[12]  0.49    0.00  0.10  0.31  0.43  0.49  0.55  0.69 14221    1
theta[13]  0.51    0.00  0.11  0.32  0.44  0.51  0.58  0.74 11588    1
theta[14]  0.41    0.00  0.11  0.18  0.34  0.41  0.48  0.62 11585    1
theta[15]  0.39    0.00  0.11  0.17  0.32  0.39  0.46  0.59 10164    1
theta[16]  0.36    0.00  0.11  0.12  0.29  0.37  0.44  0.57  6682    1
theta[17]  0.47    0.00  0.12  0.24  0.39  0.47  0.54  0.73 15593    1
theta[18]  0.44    0.00  0.10  0.24  0.37  0.44  0.50  0.64 15963    1
theta[19]  0.41    0.00  0.10  0.21  0.35  0.42  0.48  0.61 14077    1
theta[20]  0.45    0.00  0.10  0.26  0.39  0.45  0.52  0.66 17013    1
theta[21]  0.55    0.00  0.11  0.35  0.47  0.54  0.62  0.79  7677    1
theta[22]  0.44    0.00  0.10  0.26  0.38  0.44  0.50  0.63 18378    1
```

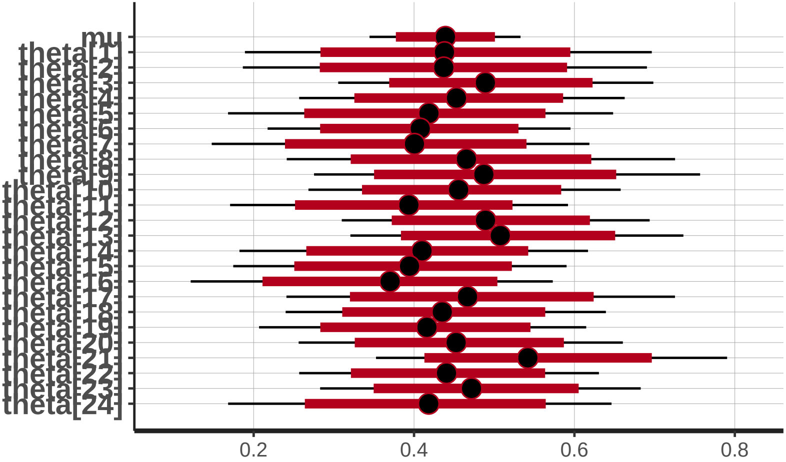
stan

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

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

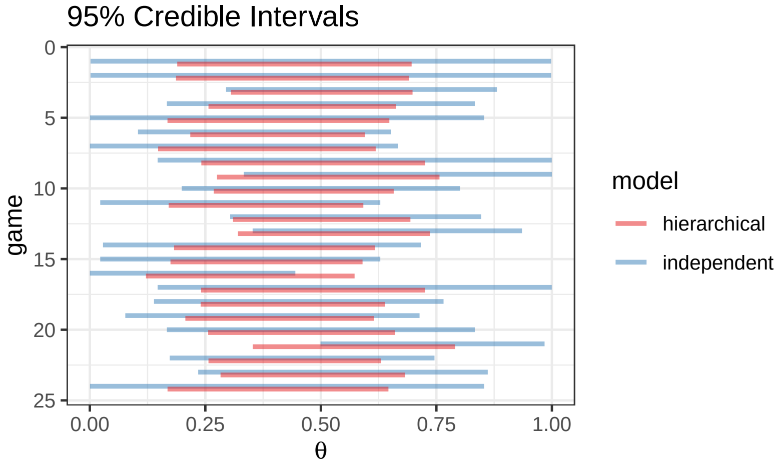


stan

```r
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> G;
  int<lower=0> n[G];
  int<lower=0> y[G];
}
parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
  real<lower=0,upper=1> theta[G];
}

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 818 divergent transitions after warmup.  See
https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
```

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 $\theta_g \stackrel{ind}{\sim} p(\theta_g|\alpha, \beta, y_g) \stackrel{d}{=} Be(\alpha + y_g, \beta + n_g - y_g)$.

The marginal posterior for α, β is

$$p(\alpha, \beta|y) \propto p(y|\alpha, \beta)p(\alpha, \beta) = \left[\prod_{g=1}^{G} \text{Beta-binomial}(y_g|n_g, \alpha, \beta)\right] p(\alpha, \beta)$$

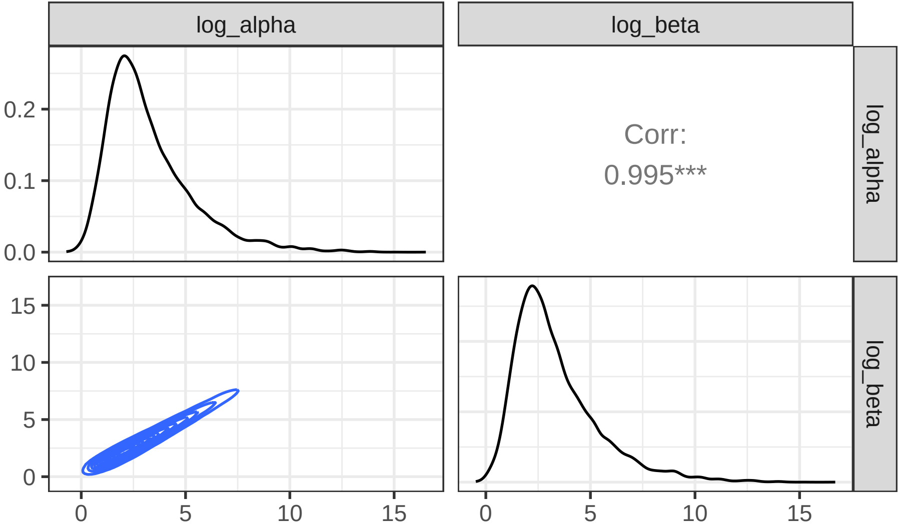
Stan - beta-binomial

```
# Marginalized (integrated) theta out of the model
model_marginalized <- "
data {
  int<lower=0> G;
  int<lower=0> n[G];
  int<lower=0> y[G];
}
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[G];
  for (i in 1:G)
    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)
```

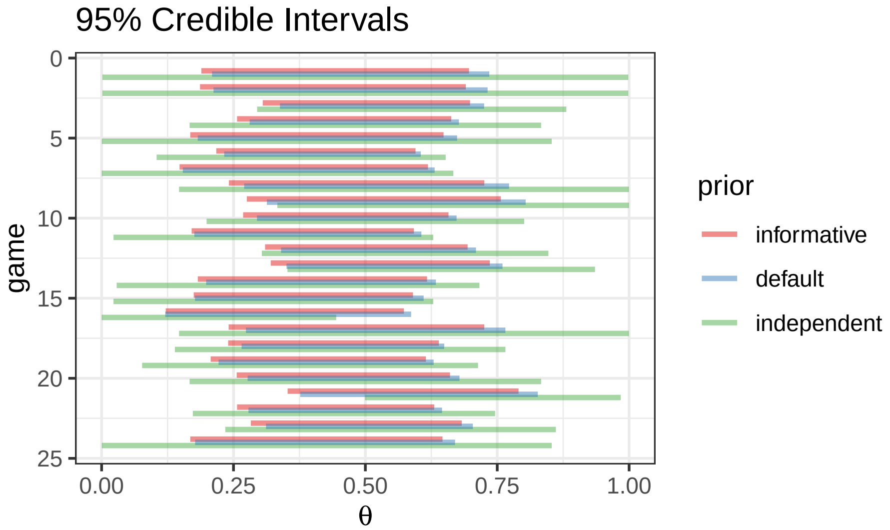
Stan - beta-binomial

```
Inference for Stan model: anon_model.
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     3272.33  984.91 114326.51 1.80  6.22 15.68 67.45 6192.47 13474    1
beta      3684.92 1145.66 135896.11 2.13  7.09 17.69 75.14 7073.72 14070    1
theta[1]     0.47    0.00      0.12 0.21  0.41  0.47  0.53    0.73 19405    1
theta[2]     0.47    0.00      0.12 0.21  0.41  0.47  0.53    0.73 19501    1
theta[3]     0.51    0.00      0.10 0.34  0.45  0.50  0.56    0.73 13524    1
theta[4]     0.48    0.00      0.10 0.28  0.42  0.48  0.53    0.68 20289    1
theta[5]     0.45    0.00      0.12 0.18  0.39  0.46  0.52    0.67 16614    1
theta[6]     0.44    0.00      0.09 0.23  0.38  0.45  0.50    0.60 12029    1
theta[7]     0.43    0.00      0.12 0.15  0.37  0.44  0.50    0.63 11106    1
theta[8]     0.50    0.00      0.12 0.27  0.43  0.49  0.55    0.77 17402    1
theta[9]     0.52    0.00      0.12 0.31  0.44  0.50  0.57    0.80 10324    1
theta[10]    0.48    0.00      0.09 0.29  0.42  0.48  0.53    0.67 18925    1
theta[11]    0.42    0.00      0.11 0.18  0.36  0.44  0.49    0.61  9252    1
theta[12]    0.51    0.00      0.09 0.34  0.45  0.50  0.56    0.71 14135    1
theta[13]    0.52    0.00      0.10 0.35  0.45  0.51  0.58    0.76 10134    1
theta[14]    0.44    0.00      0.11 0.20  0.38  0.45  0.51    0.63 12469    1
theta[15]    0.42    0.00      0.11 0.18  0.37  0.44  0.50    0.61  8810    1
theta[16]    0.40    0.00      0.12 0.12  0.33  0.42  0.49    0.59  6684    1
theta[17]    0.50    0.00      0.12 0.27  0.43  0.49  0.55    0.77 17733    1
theta[18]    0.46    0.00      0.09 0.27  0.41  0.47  0.52    0.65 18477    1
theta[19]    0.44    0.00      0.10 0.22  0.39  0.45  0.51    0.63 15034    1
theta[20]    0.48    0.00      0.10 0.28  0.42  0.48  0.53    0.68 20435    1
theta[21]    0.55    0.00      0.12 0.38  0.47  0.53  0.62    0.83  6096    1
theta[22]    0.46    0.00      0.09 0.28  0.41  0.47  0.52    0.65 19494    1
```

Posterior samples for α and β

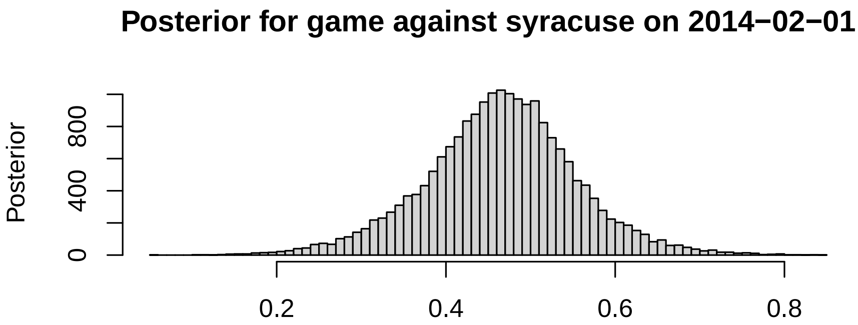


Comparing all models

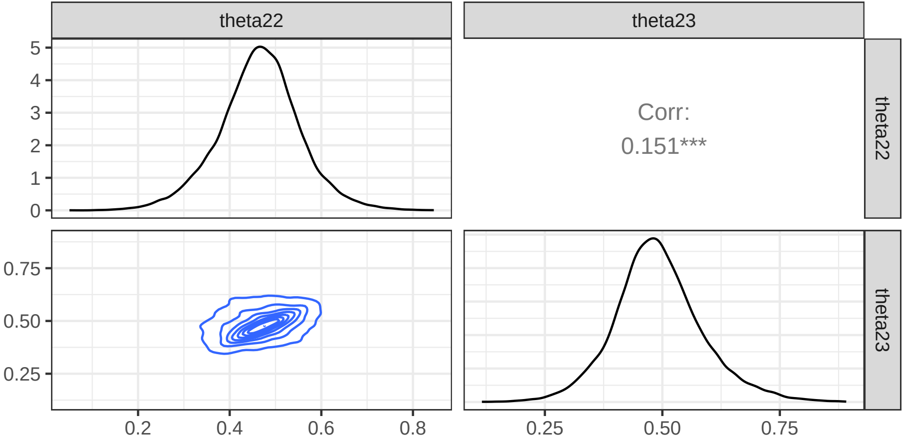


Posterior sample for θ_{22}

```r
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 season-specific 3-point percentage. The model is exactly the same, but the data changes.

season	y	n
1	36	95
2	64	150
3	67	171
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> G; int<lower=0> n[G]; int<lower=0> y[G];
  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[G];
  for (g in 1:G) theta[g] = beta_rng(alpha+y[g], beta+n[g]-y[g]);
}
"
```

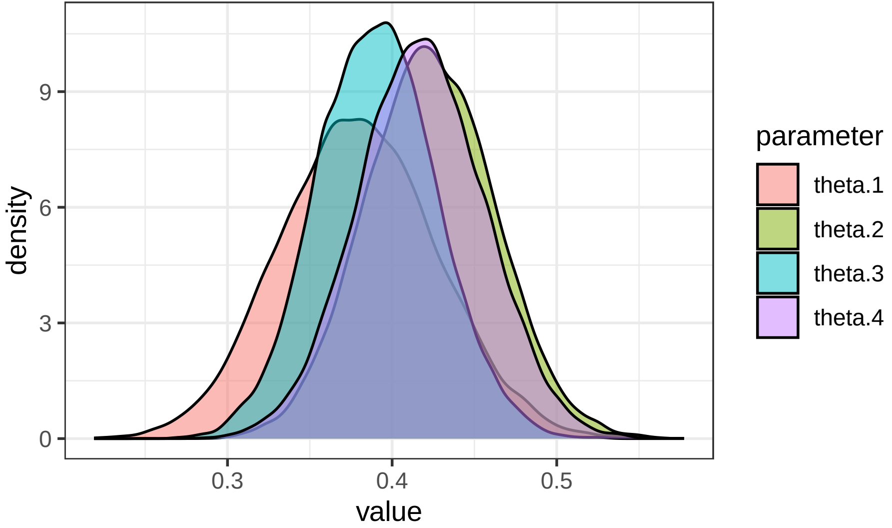
Run stan

```
dat       <- list(G = 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, iter = 10000,
                      c("alpha","beta","mu","eta","theta"))
```

Stan - hierarchical model for seasons

```
Inference for Stan model: anon_model.
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       4.90    0.03 3.09    0.92    2.61    4.23    6.45   12.54 11353    1
beta        7.97    0.04 4.63    1.81    4.57    7.05   10.38   19.41 11997    1
mu          0.38    0.00 0.06    0.25    0.33    0.38    0.42    0.50 10904    1
eta        12.87    0.07 7.56    2.88    7.31   11.35   16.72   31.53 11712    1
theta[1]    0.38    0.00 0.05    0.29    0.35    0.38    0.41    0.47 19424    1
theta[2]    0.42    0.00 0.04    0.35    0.40    0.42    0.45    0.50 19340    1
theta[3]    0.39    0.00 0.04    0.32    0.37    0.39    0.41    0.46 20016    1
theta[4]    0.42    0.00 0.04    0.34    0.39    0.42    0.44    0.49 19726    1
lp__     -402.07    0.01 1.07 -404.87 -402.49 -401.74 -401.30 -401.02  7343    1

Samples were drawn using NUTS(diag_e) at Fri Feb 16 09:00:17 2024.
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] 0.73465

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

[1] 0.45475

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

[1] 0.699
```

Summary - hierarchical models

Two-level hierarchical model:

$$y_{ij} \stackrel{ind}{\sim} p(y|\theta_i) \qquad \theta_i \stackrel{ind}{\sim} p(\theta|\phi) \qquad \phi \sim p(\phi)$$

Conditional independencies:

- $y_{ij} \perp\!\!\!\perp y_{ij'}|\theta$ for $j \neq j'$
- $y_{ij} \perp\!\!\!\perp y_{i'j'}|\theta$ for $i \neq i'$ and any j, j'
- $\theta_i \perp\!\!\!\perp \theta_{i'}|\phi$ for $i \neq i'$
- $y_{ij} \perp\!\!\!\perp \phi|\theta$ for any i, j
- $y_{ij} \perp\!\!\!\perp y_{i'j'}|\phi$ for $i \neq i'$ and any j, j'
- $\theta_i \perp\!\!\!\perp \theta_{i'}|\phi, y$ for $i \neq i'$

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.

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