Hierarchical models

Dr. Jarad Niemi

STAT 544 - Iowa State University

February 14, 2018

Outline

- Motivating example
  - Independent vs pooled estimates
- Hierarchical models
  - General structure
  - Posterior distribution
- Binomial hierarchical 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

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

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

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

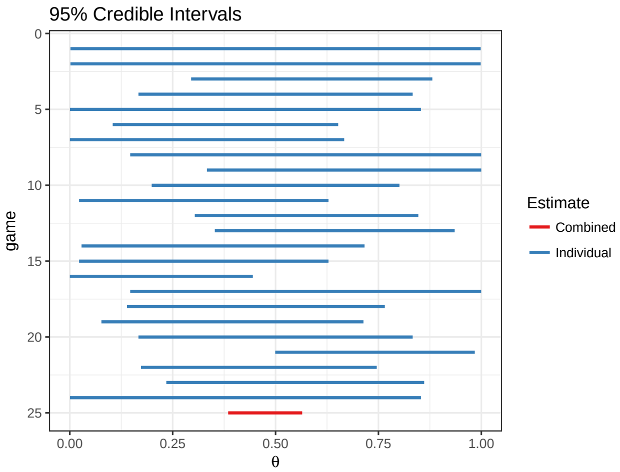
where

- 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	a	b	lcl	ucl	Estimate	game
1	11/8/13	davidson	0	0	0.50	0.50	0.00	1.00	Individual	1
2	11/12/13	kansas	0	0	0.50	0.50	0.00	1.00	Individual	2
3	11/15/13	florida atlantic	5	8	5.50	3.50	0.29	0.88	Individual	3
4	11/18/13	unc asheville	3	6	3.50	3.50	0.17	0.83	Individual	4
5	11/19/13	east carolina	0	1	0.50	1.50	0.00	0.85	Individual	5
6	11/24/13	vermont	3	9	3.50	6.50	0.10	0.65	Individual	6
7	11/27/13	alabama	0	2	0.50	2.50	0.00	0.67	Individual	7
8	11/29/13	arizona	1	1	1.50	0.50	0.15	1.00	Individual	8
9	12/3/13	michigan	2	2	2.50	0.50	0.33	1.00	Individual	9
10	12/16/13	gardner-webb	4	8	4.50	4.50	0.20	0.80	Individual	10
11	12/19/13	ucla	1	5	1.50	4.50	0.02	0.63	Individual	11
12	12/28/13	eastern michigan	6	10	6.50	4.50	0.30	0.85	Individual	12
13	12/31/13	elon	5	7	5.50	2.50	0.35	0.94	Individual	13
14	1/4/14	notre dame	1	4	1.50	3.50	0.03	0.72	Individual	14
15	1/7/14	georgia tech	1	5	1.50	4.50	0.02	0.63	Individual	15
16	1/11/14	clemson	0	4	0.50	4.50	0.00	0.44	Individual	16
17	1/13/14	virginia	1	1	1.50	0.50	0.15	1.00	Individual	17
18	1/18/14	nc state	3	7	3.50	4.50	0.14	0.77	Individual	18
19	1/22/14	miami	2	6	2.50	4.50	0.08	0.71	Individual	19
20	1/25/14	florida state	3	6	3.50	3.50	0.17	0.83	Individual	20
21	1/27/14	pitt	6	7	6.50	1.50	0.50	0.98	Individual	21
22	2/1/14	syracuse	4	9	4.50	5.50	0.17	0.75	Individual	22
23	2/4/14	wake forest	4	7	4.50	3.50	0.23	0.86	Individual	23
24	2/8/14	boston college	0	1	0.50	1.50	0.00	0.85	Individual	24
25		Total	55	116	55.50	61.50	0.38	0.56	Combined	25

Hierarchical models

Consider the following model

$$
\begin{array}{ll}
y_i & \stackrel{ind}{\sim} p(y|\theta_i) \\
\theta_i & \stackrel{ind}{\sim} p(\theta|\phi) \\
\phi & \sim p(\phi)
\end{array}
$$

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

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

Three-pointer example

Our statistical model

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

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

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

Conditional posterior for θ:

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

Marginal posterior for (α, β):

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

Thus $y_i|\alpha, \beta \stackrel{ind}{\sim}$ 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 assume an informative prior for μ and η perhaps

- $\mu \sim Be(20, 30)$
- $\eta \sim LN(0, 3^2)$

where $LN(0, 3)$ is a log-normal distribution, i.e. $\log(\eta) \sim N(0, 3^2)$.

```
a = 20
b = 30
m = 0
C = 3
```

Prior draws

```
n = 1e4

prior_draws = data.frame(mu  = rbeta(n, a, b),
                         eta = rlnorm(n, m, C)) %>%
  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     0.00131  0.389 165
2 beta      0.00204  0.580 246
3 eta       0.00342  0.983 416
4 mu        0.270    0.398   0.539
```

```
cor(prior_draws$alpha, prior_draws$beta)
```

```
[1] 0.9451046
```

```
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> C;
  real m;
}
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   ~ lognormal(m,C);

  // 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, m = m, C = C)
m = stan_model(model_code = model_informative_prior)
r = sampling(m, dat, c("mu","eta","alpha","beta","theta"),
             iter = 10000)
```

```
Warning:  There were 178 divergent transitions after warmup.  Increasing adapt_delta above 0.8 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
http://mc-stan.org/misc/warnings.html#bfmi-low
Warning:  Examine the pairs() plot to diagnose sampling problems
```

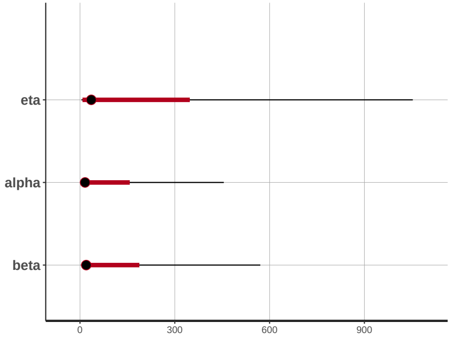
stan

```
r


Inference for Stan model: 72a83403796ce21af93650393c8e2ae4.
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          0.45    0.00   0.04    0.36    0.42    0.45    0.47    0.53  2871 1.00
eta       129.90   18.61 246.02    4.47   15.04   35.69  109.92 1053.19   175 1.03
alpha      58.55    8.42 111.50    1.90    6.61   16.03   49.53  455.05   175 1.03
beta       71.34   10.22 135.12    2.51    8.36   19.67   60.73  570.61   175 1.03
theta[1]    0.45    0.00   0.11    0.20    0.39    0.45    0.50    0.68 12075 1.00
theta[2]    0.45    0.00   0.11    0.21    0.39    0.45    0.50    0.69 20000 1.00
theta[3]    0.49    0.00   0.09    0.32    0.43    0.48    0.54    0.71  5228 1.00
theta[4]    0.46    0.00   0.09    0.27    0.40    0.46    0.51    0.66 20000 1.00
theta[5]    0.43    0.00   0.11    0.17    0.37    0.44    0.49    0.64 20000 1.00
theta[6]    0.42    0.00   0.09    0.22    0.37    0.43    0.48    0.58 20000 1.00
theta[7]    0.41    0.00   0.11    0.15    0.35    0.42    0.48    0.60  3457 1.00
theta[8]    0.47    0.00   0.11    0.26    0.41    0.46    0.52    0.73 20000 1.00
theta[9]    0.49    0.00   0.11    0.30    0.42    0.48    0.55    0.77  4549 1.00
theta[10]   0.46    0.00   0.09    0.28    0.40    0.46    0.51    0.65 20000 1.00
theta[11]   0.41    0.00   0.10    0.18    0.35    0.42    0.47    0.58  3041 1.00
theta[12]   0.49    0.00   0.09    0.33    0.43    0.48    0.54    0.69 20000 1.00
theta[13]   0.50    0.00   0.10    0.34    0.44    0.49    0.56    0.74  2402 1.00
theta[14]   0.42    0.00   0.10    0.20    0.37    0.43    0.48    0.60  5351 1.00
theta[15]   0.41    0.00   0.10    0.18    0.35    0.42    0.47    0.58  3279 1.00
theta[16]   0.38    0.00   0.11    0.12    0.32    0.40    0.46    0.56  1703 1.00
theta[17]   0.47    0.00   0.11    0.26    0.41    0.46    0.53    0.73 20000 1.00
theta[18]   0.44    0.00   0.09    0.26    0.39    0.45    0.50    0.63 20000 1.00
theta[19]   0.43    0.00   0.09    0.22    0.37    0.43    0.48    0.61 20000 1.00
```

stan

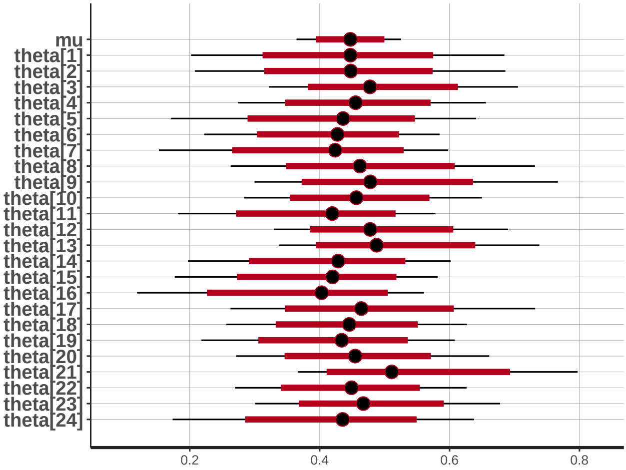
```
plot(r, pars=c('eta','alpha','beta'))
```

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

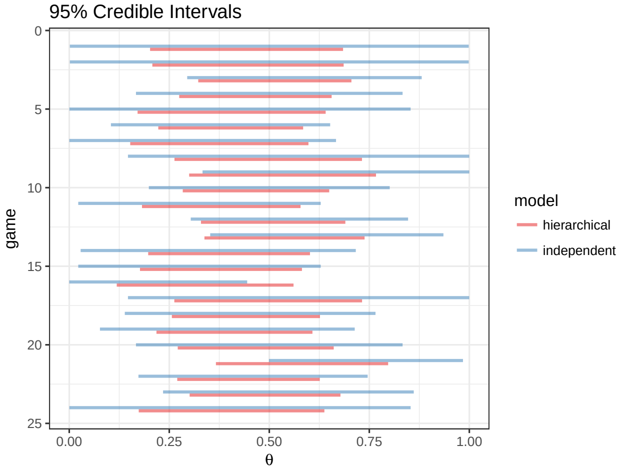


stan

```
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))
```

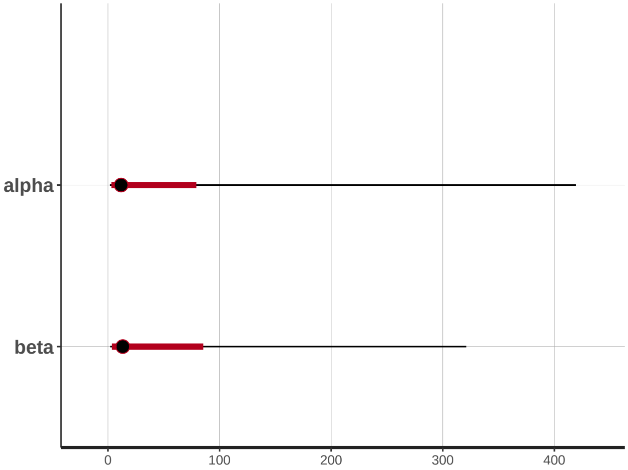
```
Warning:  There were 1991 divergent transitions after warmup.  Increasing adapt_delta above 0.9 may help.
See
http://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
Warning:  There were 4 chains where the estimated Bayesian Fraction of Missing Information was low.  See
```

Stan - default prior

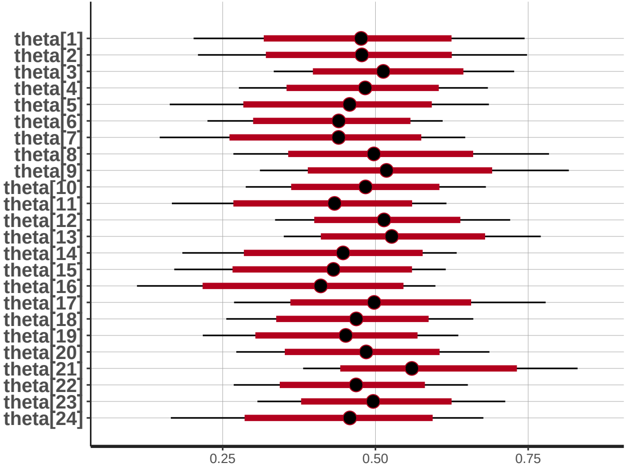
```
r2

Inference for Stan model: 5e03c866eb488d5c5da3d86e201810b1.
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     38.81   16.99 83.51   1.78   5.44  11.79  30.86 419.30    24 1.16
beta      37.61   14.33 68.17   2.02   6.16  13.27  33.88 321.13    23 1.16
theta[1]   0.47    0.00  0.13   0.20   0.40   0.48   0.55   0.74  1817 1.00
theta[2]   0.48    0.00  0.13   0.21   0.40   0.48   0.55   0.75  2400 1.00
theta[3]   0.52    0.00  0.10   0.33   0.45   0.51   0.57   0.73  4191 1.00
theta[4]   0.48    0.00  0.10   0.28   0.42   0.48   0.55   0.68  1244 1.00
theta[5]   0.45    0.00  0.13   0.16   0.38   0.46   0.53   0.69   996 1.01
theta[6]   0.43    0.00  0.10   0.23   0.37   0.44   0.50   0.61   438 1.01
theta[7]   0.43    0.01  0.12   0.15   0.36   0.44   0.51   0.65   439 1.01
theta[8]   0.51    0.00  0.12   0.27   0.43   0.50   0.58   0.78  4911 1.00
theta[9]   0.53    0.00  0.12   0.31   0.45   0.52   0.59   0.82  3797 1.00
theta[10]  0.48    0.00  0.10   0.29   0.42   0.48   0.55   0.68  1232 1.00
theta[11]  0.42    0.01  0.11   0.17   0.35   0.43   0.50   0.62   380 1.01
theta[12]  0.52    0.00  0.10   0.34   0.45   0.51   0.57   0.72  2677 1.00
theta[13]  0.54    0.00  0.11   0.35   0.46   0.53   0.60   0.77  1719 1.00
theta[14]  0.44    0.01  0.11   0.18   0.37   0.45   0.52   0.63   494 1.01
theta[15]  0.42    0.01  0.11   0.17   0.35   0.43   0.50   0.62   360 1.01
theta[16]  0.39    0.01  0.13   0.11   0.32   0.41   0.48   0.60   251 1.02
theta[17]  0.50    0.00  0.12   0.27   0.43   0.50   0.58   0.78  3032 1.00
theta[18]  0.47    0.00  0.10   0.26   0.40   0.47   0.53   0.66   678 1.01
theta[19]  0.44    0.00  0.10   0.22   0.38   0.45   0.51   0.64   622 1.01
theta[20]  0.48    0.00  0.10   0.27   0.42   0.48   0.55   0.69  1254 1.00
theta[21]  0.57    0.00  0.11   0.38   0.49   0.56   0.64   0.83  1371 1.00
```

Stan - default prior

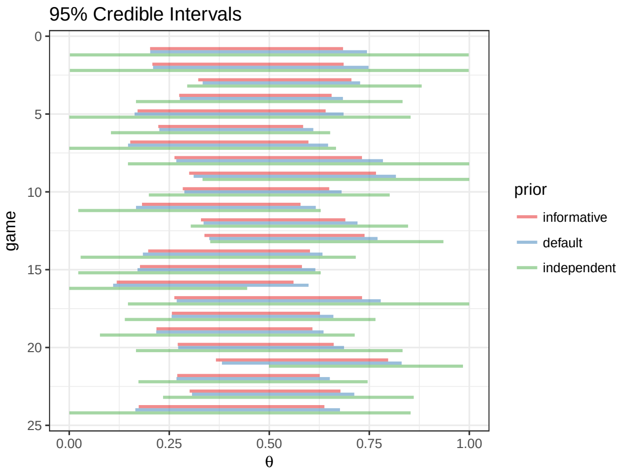
```
plot(r2, pars=c('alpha','beta'))
```



Stan - default prior



Comparing all models



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 $\theta_i \stackrel{ind}{\sim} p(\theta_i|\alpha, \beta, y_i) \stackrel{d}{=} Be(\alpha + y_i, \beta + n_i - y_i)$.

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

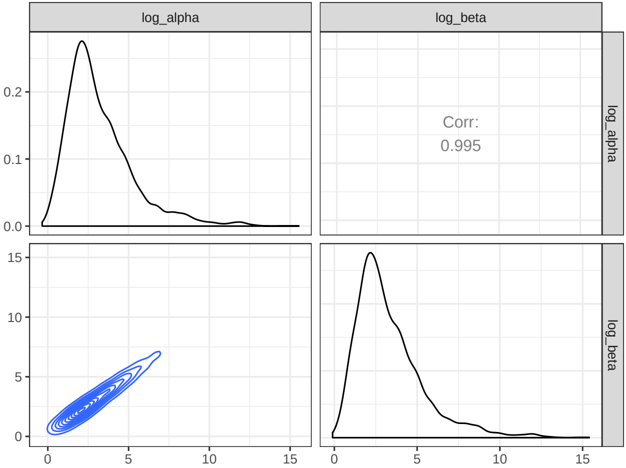
```
# Marginalized (integrated) theta out of the model
model_marginalized = "
data {
  int<lower=0> N;
  int<lower=0> n[N];
  int<lower=0> y[N];
}
parameters {
  real<lower=0> alpha;
  real<lower=0> beta;
}
model {
  target += -5*log(alpha+beta)/2;
  y     ~ beta_binomial(n,alpha,beta);
}
"

m3 = stan_model(model_code=model_marginalized)
r3 = sampling(m3, dat, c("alpha","beta"))
```

Stan - beta-binomial

```
Inference for Stan model: e43d085e5efc74fdcaa9b1ceb76cdc65.
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 3894.94 2310.00 98697.42   1.68   6.11  14.98  63.17 5789.49  1826    1
beta  3945.95 2217.53 89677.10   2.05   6.81  16.98  70.05 6477.70  1635    1
lp__   -84.60    0.04     1.07 -87.43 -85.05 -84.29 -83.82  -83.50   603    1

Samples were drawn using NUTS(diag_e) at Wed Feb  7 21:14:56 2018.
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).
```

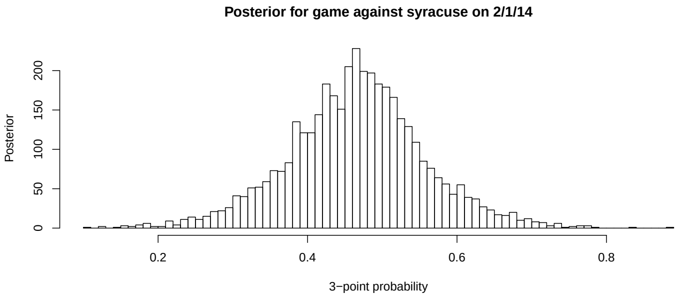
Posterior samples for α and β

```
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()
```

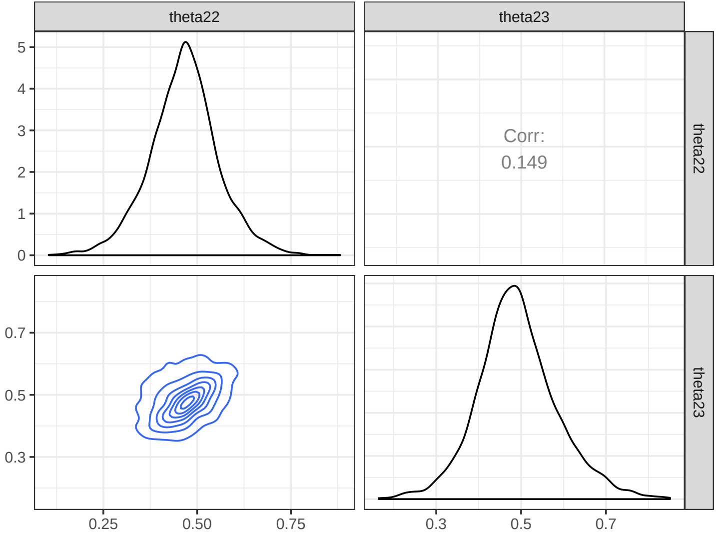


Posterior sample for θ_{22}

```
samples = extract(r3, c("alpha","beta"))
game = 22
theta22 = rbeta(length(samples$alpha),
                samples$alpha + d$made[game],
                samples$beta  + d$attempts[game] - d$made[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.

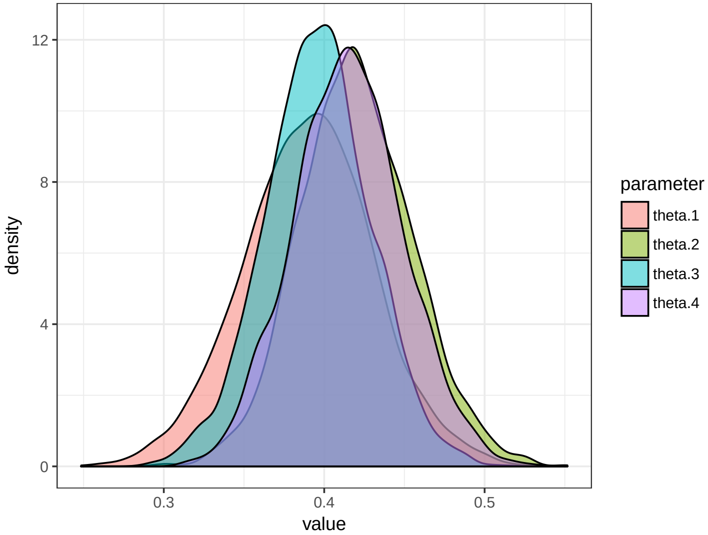
Stan - beta-binomial

```
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> C; real m;
}
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 ~ lognormal(m,C);
  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 = 20, b = 30, m = 0, C = 2)
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: 24d4f28c4da8aec87d2181da8fb225b4.
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      58.84    3.16 134.65    1.83   10.73   25.01   57.00  327.58  1812    1
beta       86.14    4.58 196.86    2.80   15.80   36.72   83.56  463.47  1844    1
mu          0.41    0.00   0.04    0.34    0.38    0.41    0.43    0.48  1779    1
eta       144.97    7.74 331.11    4.73   26.99   61.66  140.79  793.69  1829    1
theta[1]    0.39    0.00   0.04    0.31    0.36    0.39    0.42    0.47  3670    1
theta[2]    0.42    0.00   0.04    0.35    0.40    0.42    0.44    0.49  3342    1
theta[3]    0.40    0.00   0.03    0.33    0.37    0.40    0.42    0.46  3912    1
theta[4]    0.42    0.00   0.03    0.35    0.39    0.41    0.44    0.48  3624    1
lp__     -422.68    0.03   1.07 -425.58 -423.13 -422.35 -421.89 -421.61  1365    1

Samples were drawn using NUTS(diag_e) at Wed Feb  7 21:18:12 2018.
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] 0.68575

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

[1] 0.45075

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

[1] 0.66075
```

Summary - hierarchical models

Two-level hierarchical model:

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

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.

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