Multilevel statistical modeling

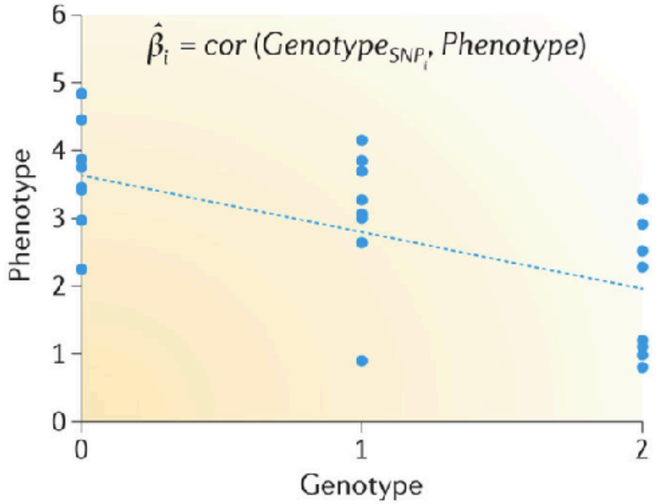
Manuel A. Rivas
Department of Biomedical Data Science
BMI217
Stanford University
rivaslab.stanford.edu

RIVASLAB

Motivating examples for today

- Dissecting the genetics of complex traits using summary association statistics
- Heritability in the genomics era

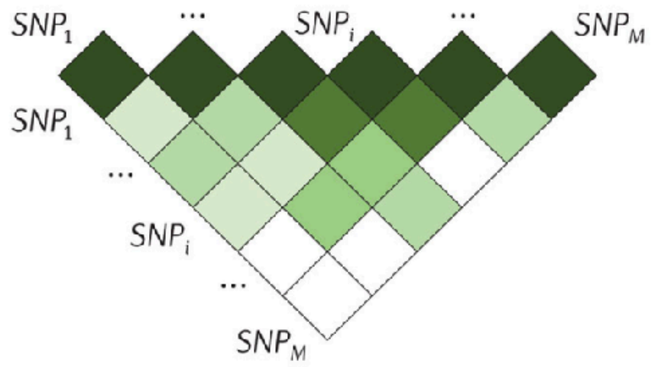
Effect size at SNP_i



z-scores

$$\frac{\hat{\beta}_1}{s.e(\hat{\beta}_1)}, \ldots, \frac{\hat{\beta}_i}{s.e(\hat{\beta}_i)}, \ldots, \frac{\hat{\beta}_M}{s.e(\hat{\beta}_M)} \sim MVN\,(0, V)$$

SNP LD (V)



Nature Reviews | Genetics

Pasaniuc and Price, 2017

RIVASLAB

In human genetics, summary statistics have been used for combining evidence for a particular variant across studies.

Meta-analysis model using Stan (http://mc-stan.org/)

Stan is a probabilistic programming language for statistical inference.

It can be used for statistical modeling across a variety of domains.

Eight schools example

School	Estimate	Standard Error
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Parameters of interest

1) the effects of coaching in each of the eight schools, and

2) the hyperparameter representing the variation of these effects in the modeled population.

Parameters of interest

1) the effects of coaching in each of the eight schools, and

2) the hyperparameter representing the variation of these effects in the modeled population.

- In a genetic association study this may be of immediate relevance when trying to identify variants with heterogeneous genetic effects across populations, for instance.
- See "Binary effects assumption" from Interpreting Meta-Analyses of Genome-Wide Association Studies from Han and Eskin(2012) http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002555.

Statistical model

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma_j\right), j = 1,\ldots,8;$$

$$\beta_j \sim \mathcal{N}\left(\mu, \tau\right), j = 1,\ldots,8;$$

$$p\left(\mu, \tau\right) \propto 1.$$

Where each σ_j is assumed known

Specify the data

```
data {
  int<lower=0> J;       // number of schools
  real betahat[J]; // estimated treatment effects
real<lower=0> sigma[J]; //standard error of effect estimates
}
```

The data block, specifies the data that is conditioned upon in Bayes Rule: the number of schools, J, the vector of estimates, $(\hat{\beta}_1 \ldots, \hat{\beta}_J)$, and the vector of standard errors of the estimates $(\sigma_1, \ldots, \sigma_j)$.

Specify the parameters

```
parameters {
  real mu;
  real<lower=0> tau;
  vector[J] eta;
}
```

The parameters block declares the parameters whose posterior distribution is sought. These are the the mean, μ, and standard deviation, τ, of the school effects, plus the standardized school-level effects η, which will be used to obtain β (for sampling purpose).

Transformed parameters

```
transformed parameters {
  vector[J] beta;
  beta = mu + tau * eta;
}
```

Model

```
model {
  target += normal_lpdf(eta | 0, 1);
  target += normal_lpdf(betahat | beta, sigma);
}
```

Evaluate the target density from which the sampler samples.

Drawing posterior samples

We prepare the data in R using lists:

In [1]:
```
schools_data <- list(
  J = 8,
  betahat = c(28,  8, -3,  7, -1,  1, 18, 12),
  sigma = c(15, 10, 16, 11,  9, 11, 10, 18)
)
```

Next, we call RStan to draw posterior samples:

In [32]:
```
options(warn=-1)
library(rstan)
fit1 <- stan(
  file = "schools.stan",  # Stan program
  data = schools_data,    # named list of data
  chains = 4,             # number of Markov chains
  warmup = 1000,          # number of warmup iterations per chain
  iter = 2000,            # total number of iterations per chain
  cores = 2,              # number of cores (using 2 just for the vignette)
  refresh = 1000          # show progress every 'refresh' iterations
  )
```

Summary of the parameters

```
print(fit1, pars=c("beta", "mu", "tau", "lp__"), probs=c(.1,.5,.9))
```

```
Inference for Stan model: schools.
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    10%    50%    90% n_eff Rhat
beta[1]  11.40    0.16 8.14   2.59  10.27  22.11  2471    1
beta[2]   8.00    0.10 6.14   0.71   7.95  15.16  3596    1
beta[3]   6.18    0.14 7.76  -3.25   6.81  14.82  2869    1
beta[4]   7.55    0.10 6.33  -0.02   7.53  15.21  3734    1
beta[5]   5.15    0.10 6.14  -2.87   5.64  12.45  4000    1
beta[6]   6.08    0.12 6.48  -2.05   6.46  13.71  3141    1
beta[7]  10.76    0.13 6.80   2.90  10.06  19.62  2895    1
beta[8]   8.63    0.14 7.78  -0.19   8.48  17.72  3056    1
mu        8.02    0.12 4.88   2.05   7.97  13.93  1761    1
tau       6.57    0.14 5.21   1.06   5.36  13.43  1457    1
lp__    -39.31    0.07 2.55 -42.67 -39.02 -36.26  1360    1

Samples were drawn using NUTS(diag_e) at Mon Apr 10 21:58:52 2017.
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).
```

Summary of the parameters

```
print(fit1, pars=c("beta", "mu", "tau", "lp__"), probs=c(.1,.5,.9))
```

```
Inference for Stan model: schools.
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     10%     50%     90% n_eff Rhat
beta[1]  11.40    0.16 8.14    2.59   10.27   22.11  2471    1
beta[2]   8.00    0.10 6.14    0.71    7.95   15.16  3596    1
beta[3]   6.18    0.14 7.76   -3.25    6.81   14.82  2869    1
beta[4]   7.55    0.10 6.33   -0.02    7.53   15.21  3734    1
beta[5]   5.15    0.10 6.14   -2.87    5.64   12.45  4000    1
beta[6]   6.08    0.12 6.48   -2.03    6.46   13.71  3141    1
beta[7]  10.76    0.13 6.80    2.90   10.06   19.62  2895    1
beta[8]   8.63    0.14 7.78   -0.19    8.48   17.72  3056    1
mu        8.02    0.12 4.88    2.05    7.97   13.93  1761    1
tau       6.57    0.14 5.21    1.06    5.36   13.43  1457    1
lp__    -39.31    0.07 2.55  -42.67  -39.02  -36.26  1360    1

Samples were drawn using NUTS(diag_e) at Mon Apr 10 21:58:52 2017.
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).
```

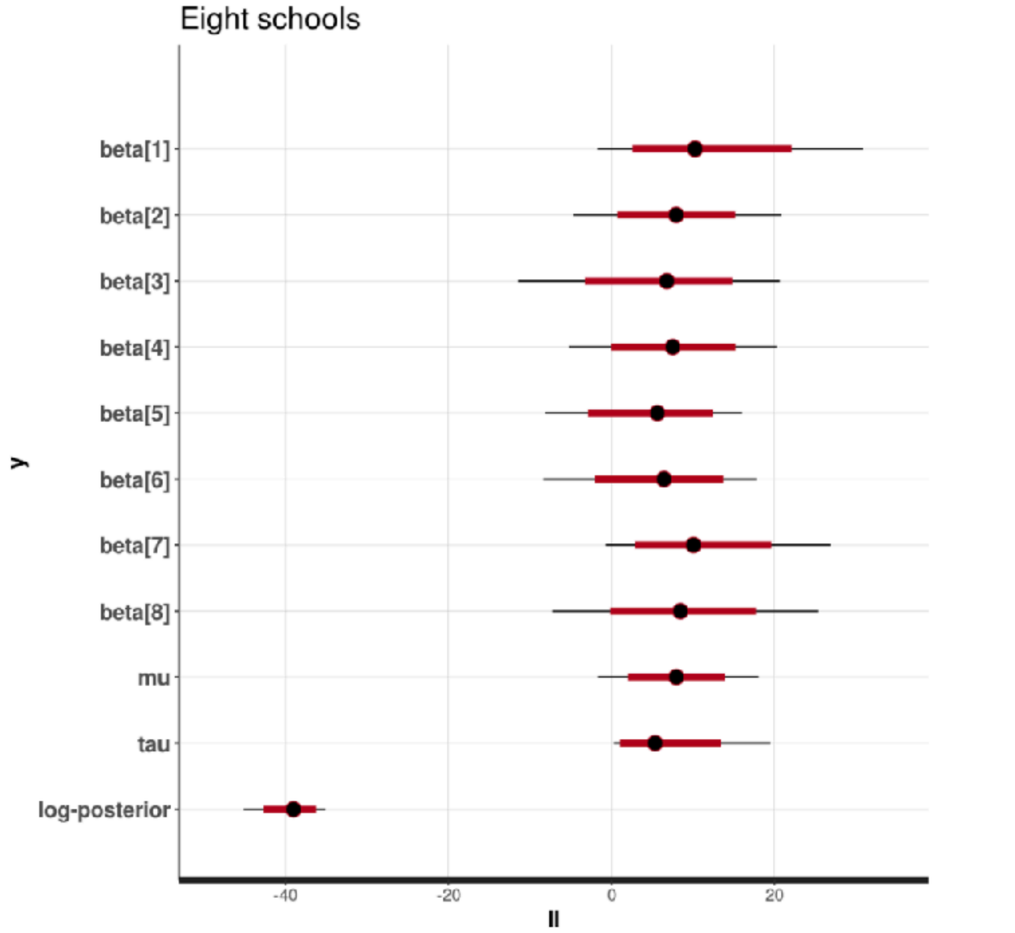
Mean estimate of beta corresponds to 9.02 with estimate of **variation** of the effects, estimated at 6.57

RIVASLAB

Plotting to visualize parameters

```
options(warn=-1)
options(repr.plot.width=8, repr.plot.height=8)
plot(fit1, pars=c("beta", "mu", "tau", face="bold"))
```

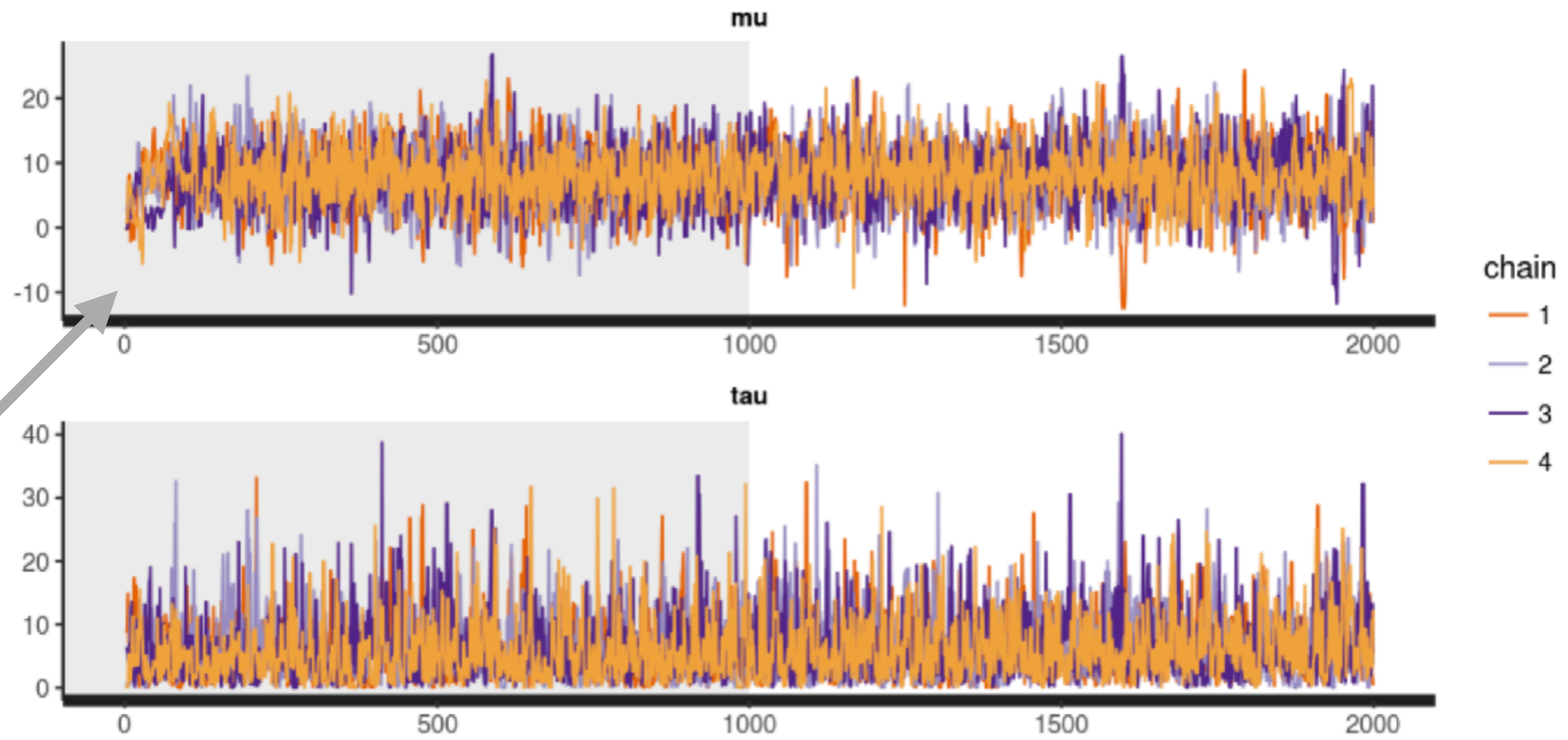
```
ci_level: 0.8 (80% intervals)
outer_level: 0.95 (95% intervals)
```



Traceplots

Used to plot the time series of the posterior draws.

```
options(repr.plot.width=8, repr.plot.height=4)
traceplot(fit1, pars = c("mu", "tau"), inc_warmup = TRUE, nrow = 2)
```



"**Warm-up**" stage of the sampler.

Convergence of Markov Chains

```
print(fit1, pars = c("mu", "tau"))
```

```
Inference for Stan model: schools.
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
mu  8.02    0.12 4.88 -1.71 5.0 7.97 11.06 18.07  1761    1
tau 6.57    0.14 5.21  0.25 2.6 5.36  9.39 19.53  1457    1

Samples were drawn using NUTS(diag_e) at Mon Apr 10 21:58:52 2017.
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).
```

To assess convergence of Markov Chains Gelman et al. propose the $\hat{R}$ statistic, which is presented in Rstan summary.

Motivating examples for today

- Dependencies on effects
- Tissues, cells, genes

Bernoulli Likelihood with hierarchical prior

1. A single coin from a single mint
2. Multiple coins from a single mint
3. Multiple coins from multiple mints

Goal: Assume you have two coins. Estimate each parameter θ_1 and θ_2 and estimate each bias in each coin.

Let's explore situations in which there are two or more parameters that do have meaningful dependencies.

Example: The bias of a coin depends on characteristics of the factory in which it is minted

Prior beliefs about parameters values of the mint

Dependence of the coin's bias on the minting parameter

Parameters that directly affect the data: **parameters**

Parameters that affect data indirectly by affecting beliefs about other parameters: **hyperparameters**

$$
\begin{aligned}
P(y|\theta) &= \mathrm{Ber}(y|\theta) \\
&= \theta^{y}(1-\theta)^{1-y}
\end{aligned}
$$

$y = 1$ for head

$y = 0$ for tail

Prior distribution

$$p(\theta) \sim \text{Beta}(a, b)$$

To make parameters of beta more intuitive, we express them in terms of the corresponding mean μ and sample size K.

To make parameters of beta more intuitive, we express them in terms of the corresponding mean w and sample size K.

If mean is w and confidence is reflected by prior sample size K

$$a = w(K-2)+1$$
$$b = (1-w)(K-2)+1$$

$$p(\theta|w) = \text{Beta}(\theta|w(K-2)+1,(1-w)(K-2)+1)$$

The magnitude of K is an expression of our prior certainty regarding the dependence of the bias on w.

To make parameters of beta more intuitive, we express them in terms of the corresponding mean W and sample size K.

If mean is w and confidence is reflected by prior sample size K

$$a = w(K-2)+1$$
$$b = (1-w)(K-2)+1$$

$$p(\theta|_{W}) = \text{Beta}(\theta|w(K-2)+1,(1-w)(K-2)+1)$$

The magnitude of K is an expression of our prior certainty regarding the dependence of the bias on w.

When K is large, the distribution of θ is very narrowly loaded over W

When K is small, the distribution of θ is very widely dispersed around w.

Thus, as K gets large, we are more and more certain about the form of the dependency of θ on w.

To make parameters of beta more intuitive, we express them in terms of the corresponding mear W and sample size K.

If mean is w and confidence is reflected by prior sample size K

$$a = w(K-2)+1$$
$$b = (1-w)(K-2)+1$$

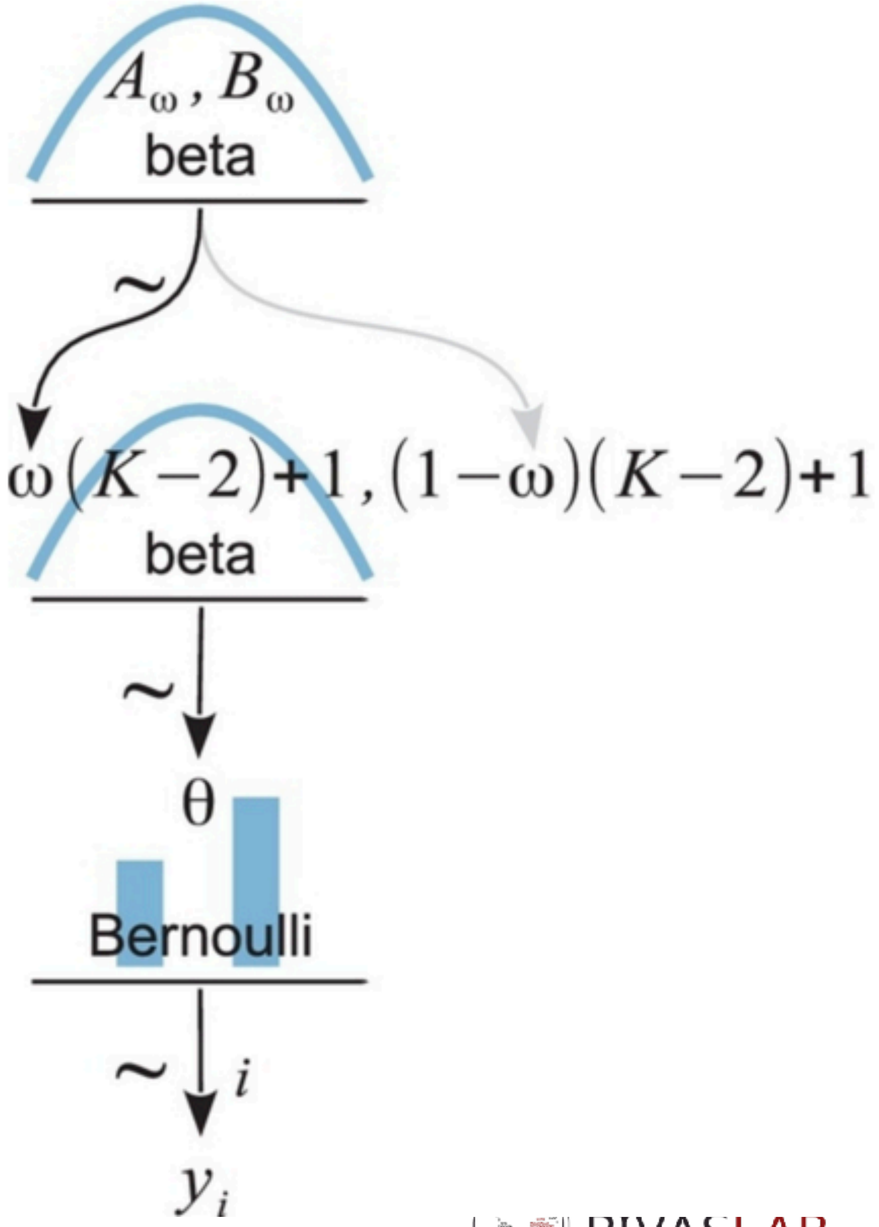
$$p(\theta|_{w}) = \text{Beta}(\theta|w(K-2)+1,(1-w)(K-2)+1)$$

The magnitude of K is an expression of our prior certainty regarding the dependence of the bias on w.

When K is large, the distribution of θ is very narrowly loaded over W

When K is small, the distribution of θ is very widely dispersed around w.

Thus, as K gets large, we are more and more certain about the form of the dependency of θ on w.



A single coin from a single mint

Expansion of the scenario into the realm of hierarchical models

Instead of specifying a single particular value for w

We think of w as taking on many possible values (from 0 to 1), and we specify a probability distribution over those values.

This distribution can be thought of as describing the uncertainty in our beliefs about the construction of the mint that manufactured the coin.

Expansion of the scenario into the realm of hierarchical models

When w is large, the mint tends to produce coins with large biases, and when w is small, the mint tends to produce coins with small biases.

Our prior distribution over w expresses what we believe about how mints are constructed. We suppose that the distribution on w is again a beta distribution, $p\left(w\right) = \text{Beta}\left(w|A_w, B_w\right)$ where A_w and B_w are constants. In this case, we believe that w is typically near $A_w/(A_w + B_w)$.

Expansion of the scenario into the realm of hierarchical models

When w is large, the mint tends to produce coins with large biases, and when w is small, the mint tends to produce coins with small biases.

Our prior distribution over w expresses what we believe about how mints are constructed. We suppose that the distribution on w is again a beta distribution, $p(w) = \text{Beta}(w|A_w, B_w)$ where A_w and B_w are constants. In this case, we believe that w is typically near $A_w/(A_w + B_w)$.

The hyperparameter w expresses the bias of the mint that created the coin, and w depends on a beta distribution with parameters A_w and B_w, which are set by prior beliefs

This form of model is referred to as a **hierarchical model** because of the layers of dependencies.

Expansio

When w is la
biases.

Our prior dist
beta distribut
.

The hyperpa
B_w, which a

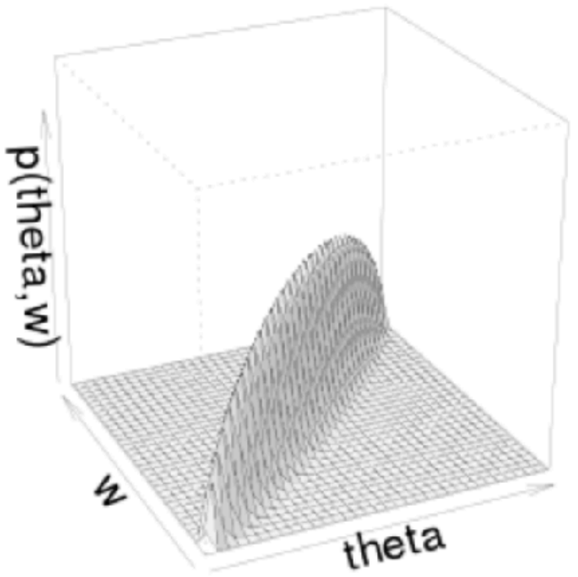
This form of

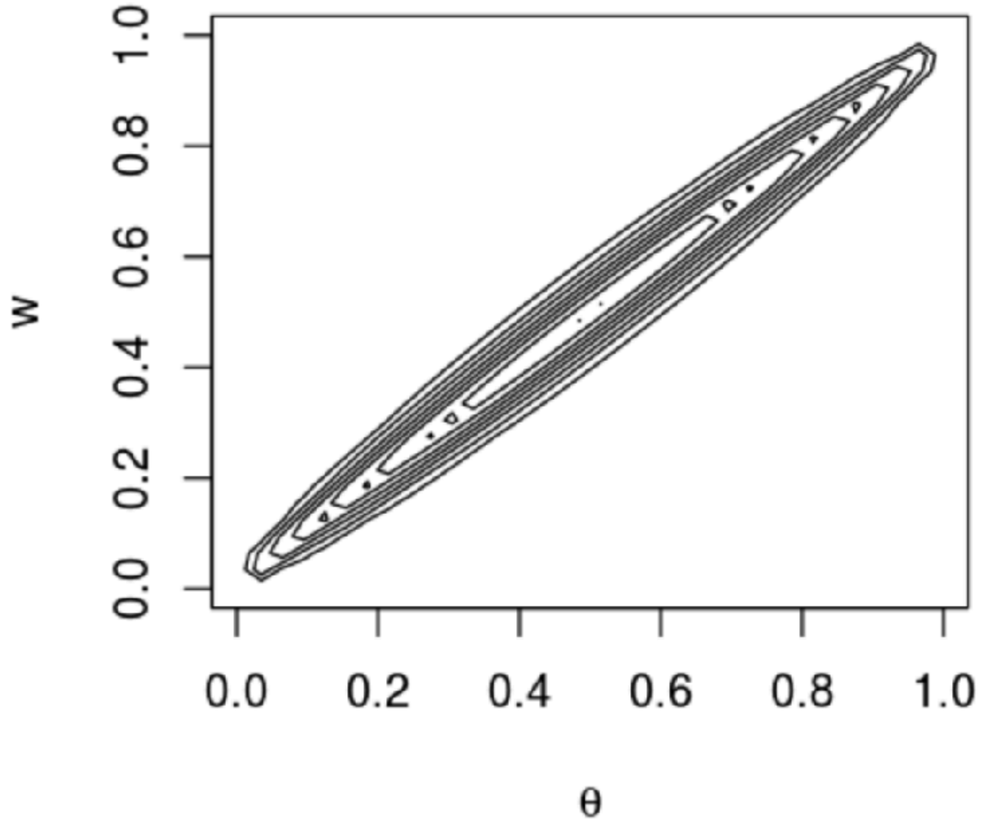
nall

is again a
$(A_w + B_w)$

ers A_w and

Prior Distribution





RIVASLAB

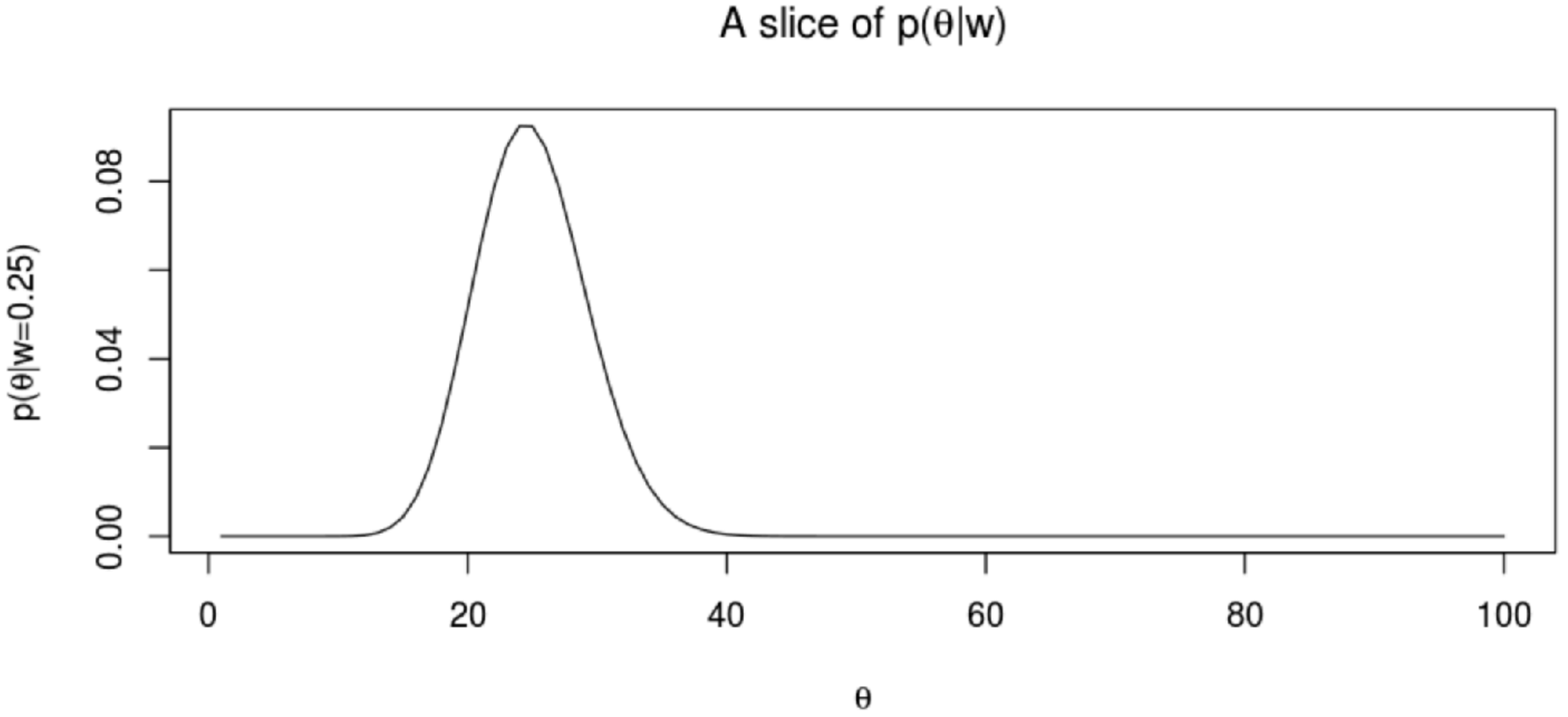
Expansion of the scenario into the realm of hierarchical models

When w is large, the mint tends to produce coins with large biases, and when w is small, the mint tends to produce coins with small biases.

Our prior di gain a beta distrib $+ B_w)$.

The hyperp $_w$ and B_w, which

This form c



RIVASLAB

Exp

Wher

biase

Our p

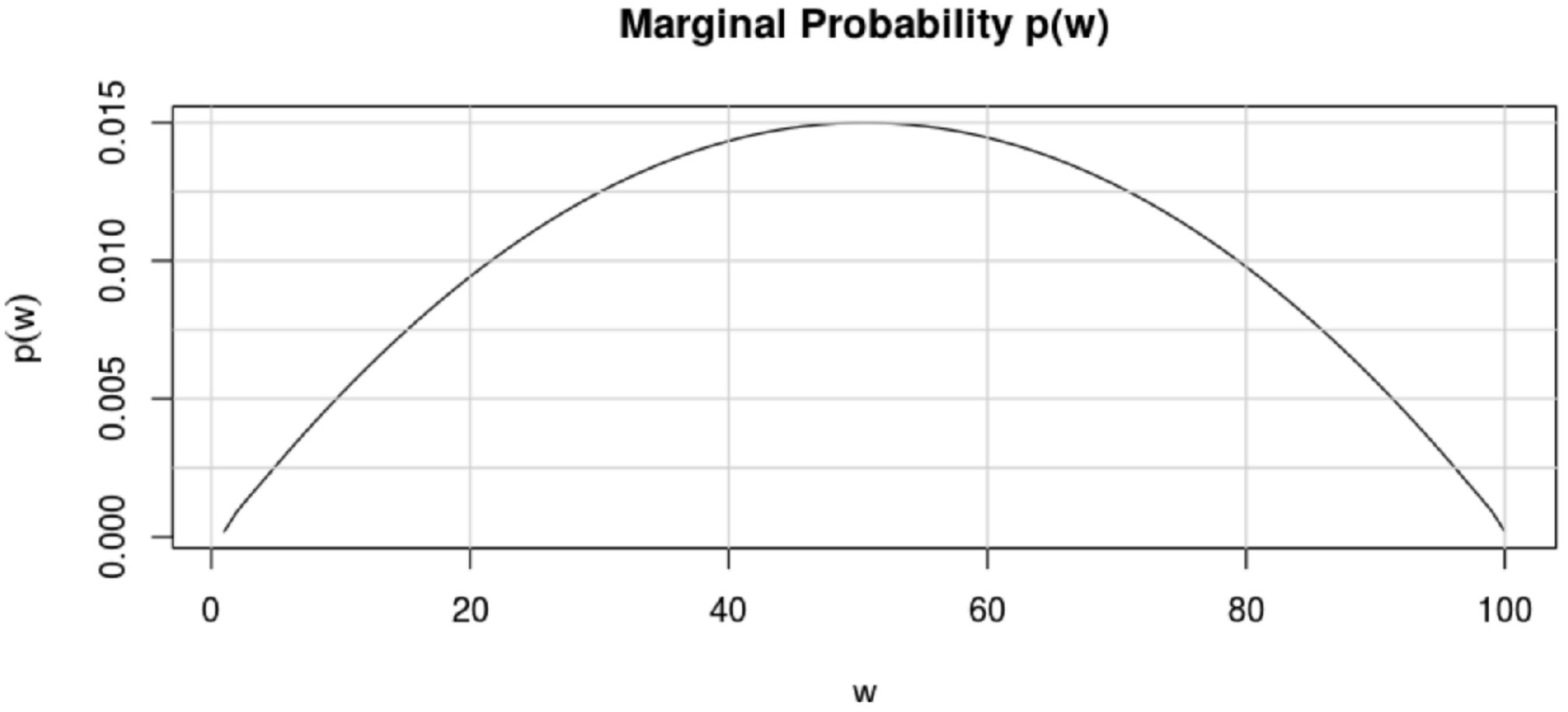
beta (

.

The l

B_w, \

This



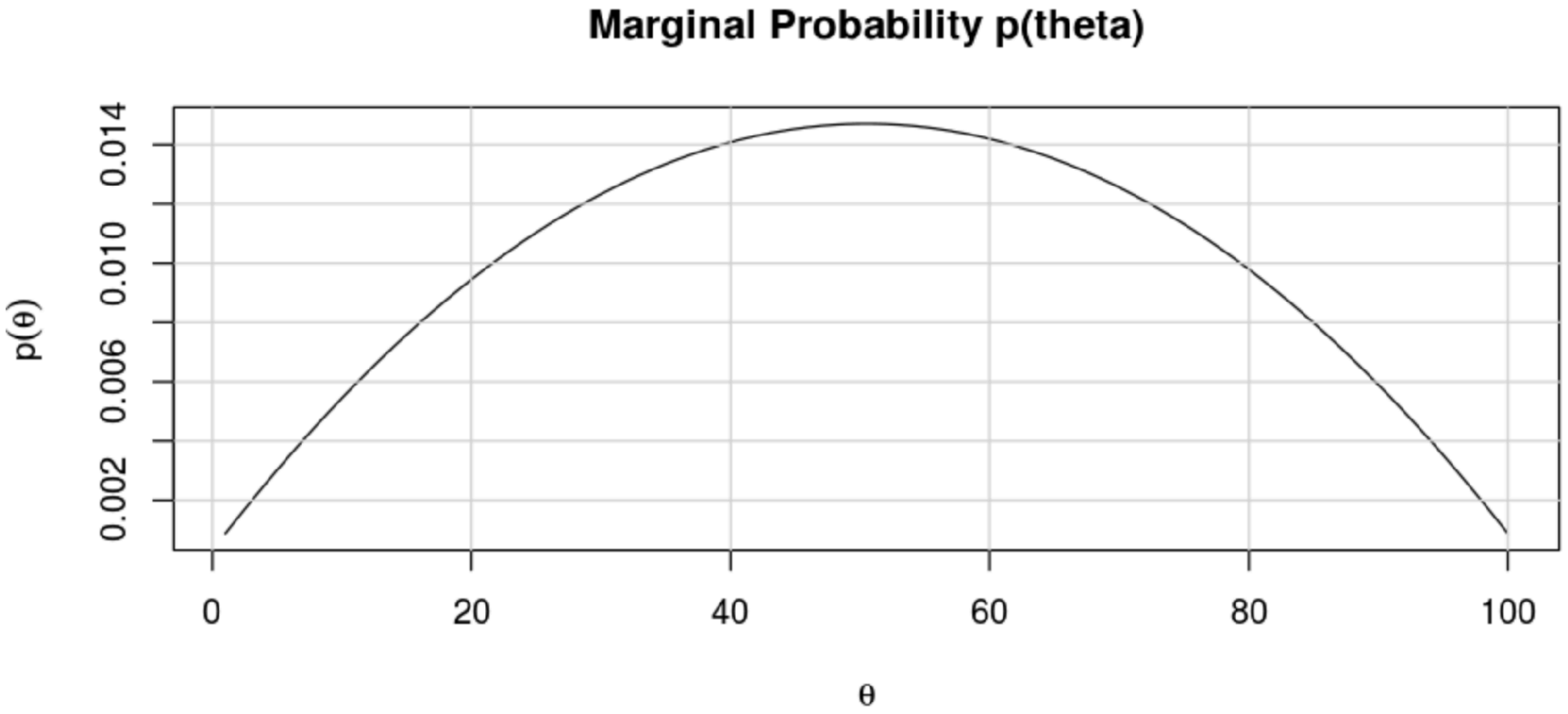
all

is again a

$A_w + B_w$)

ers A_w and

RIVASLAB


Marginal Probability p(theta)
p(θ)
0.002
0.006
0.010
0.014
0
20
40
60
80
100
θ

n a
B_w)

and

Expansion of the sce

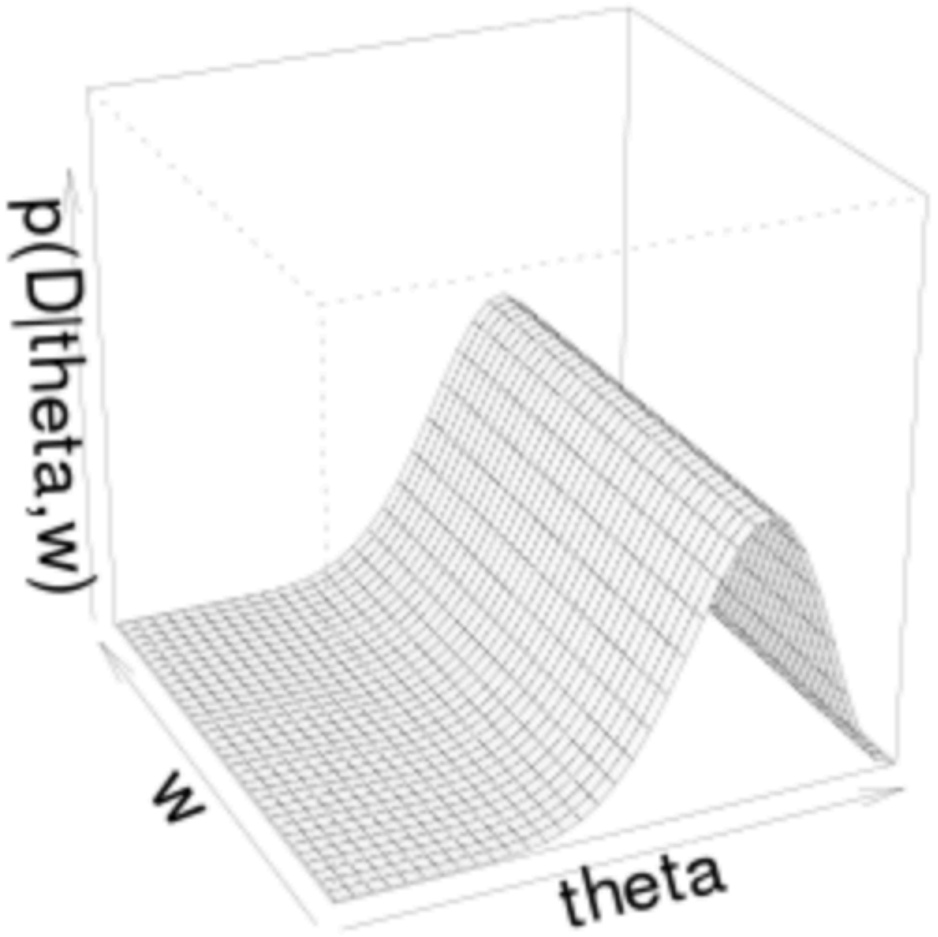
When w is large, the mint ter
biases.

Our prior distribution over w
beta distribution, $p\,(w) = \mathrm{B}$
.

The hyperparameter w expr
B_w, which are set by prior b

This form of model is referre

Likelihood Distribution



nds to produce coins with small

se that the distribution on w is again a
e that w is typically near $A_w/(A_w + B_w)$

eta distribution with parameters A_w and

RIVASLAB

Expansion of the scenari

When w is large, the mint tends to

biases.

Our prior distribution over w expre

beta distribution, $p\,(w) = \mathrm{Beta}\,(w$

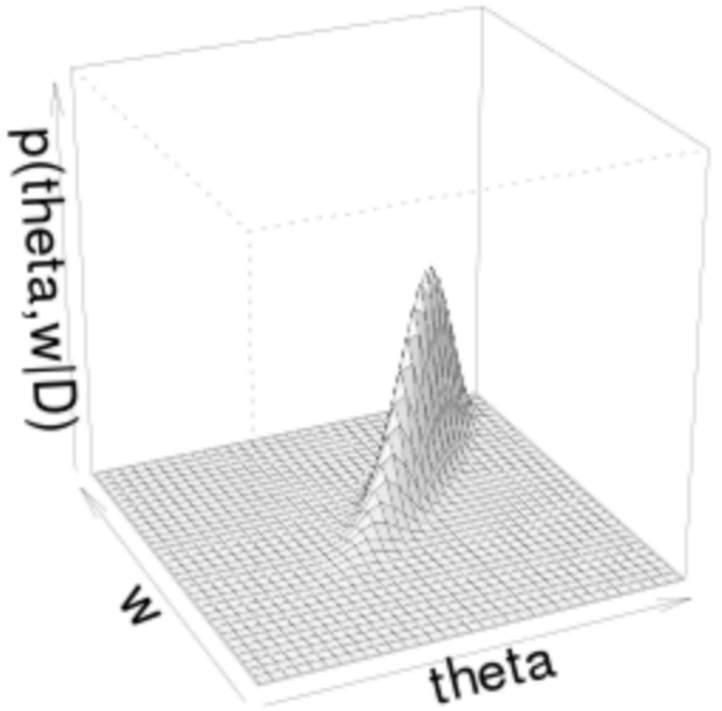
.

The hyperparameter w expresses

B_w, which are set by prior beliefs

This form of model is referred to a

Posterior Distribution



Evidence, p(D) = 0.000397

ıe mint tends to produce coins with small

We suppose that the distribution on w is again a

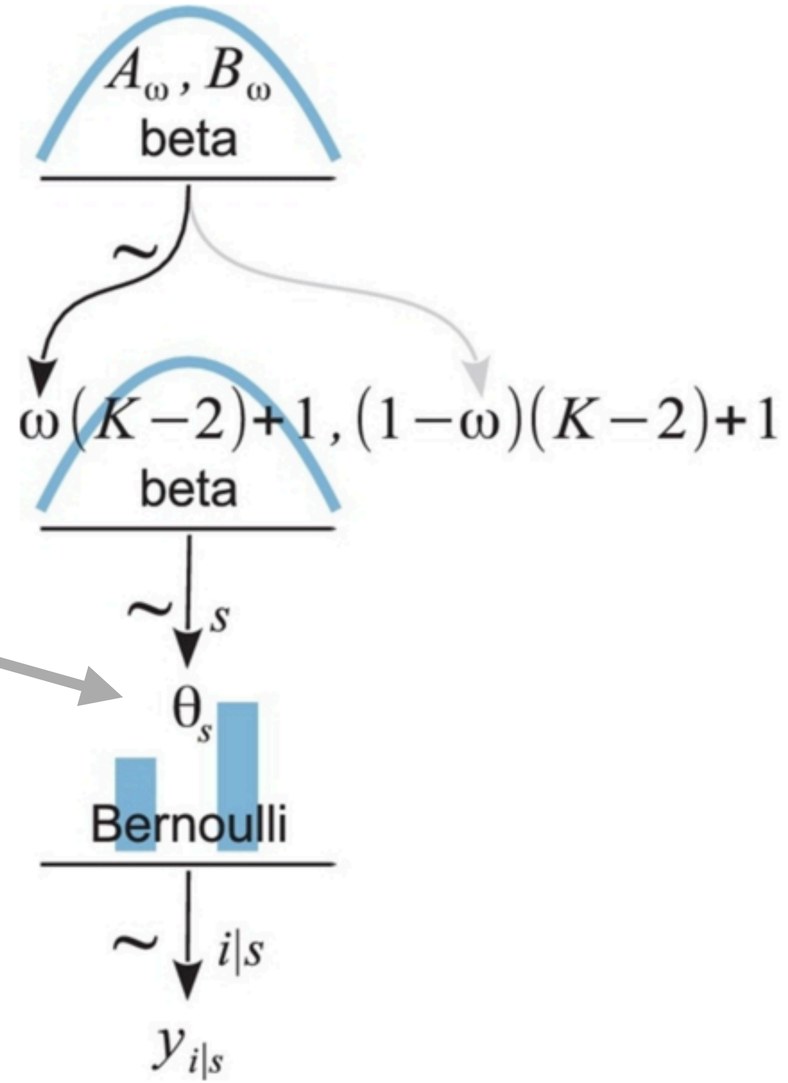
we believe that w is typically near $A_w/(A_w + B_w)$

ds on a beta distribution with parameters A_w and

encies.

Multiple coins from a single mint

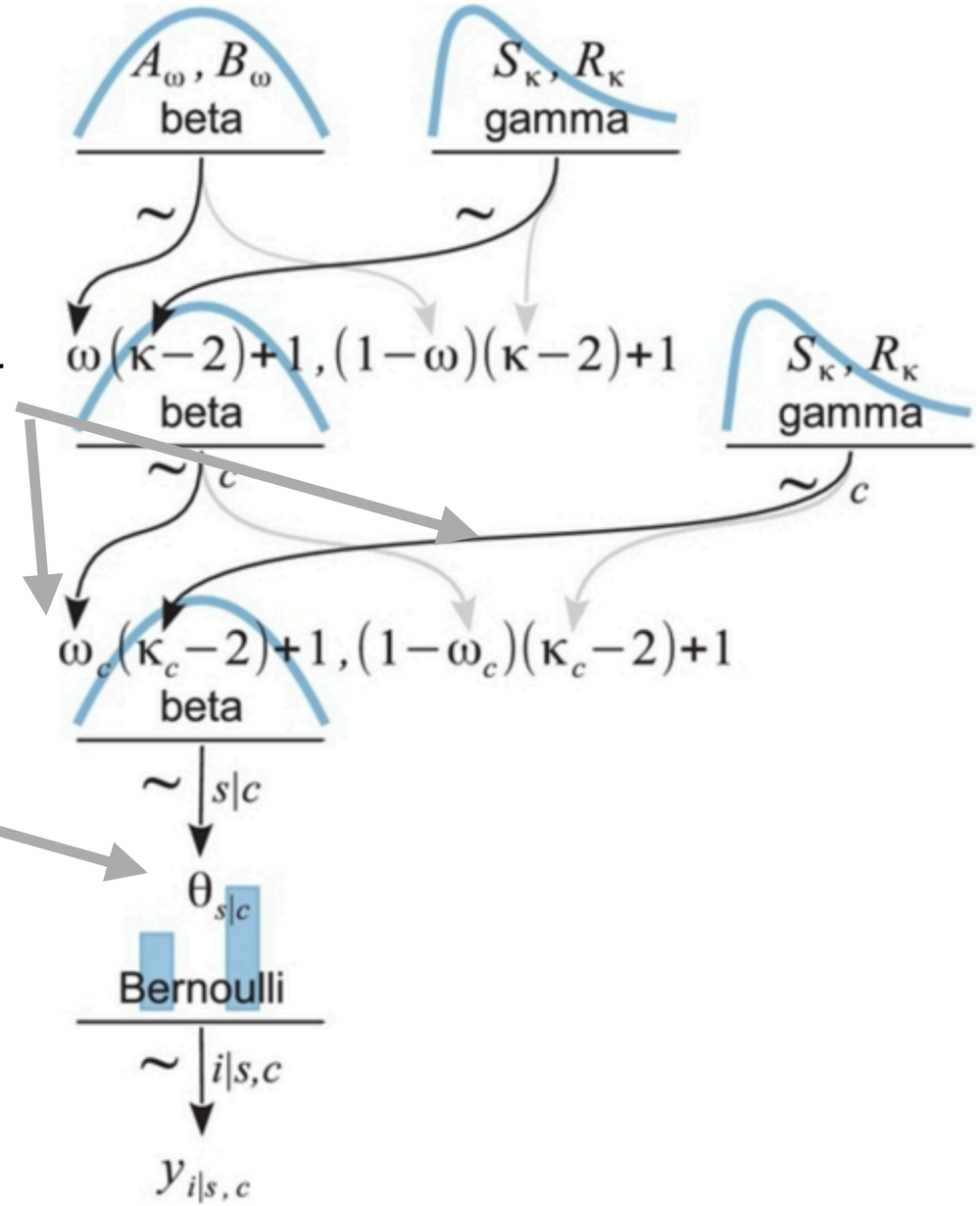
Each coin has its own parameter



Multiple coins from multiple mints

Each mint has its own parameter

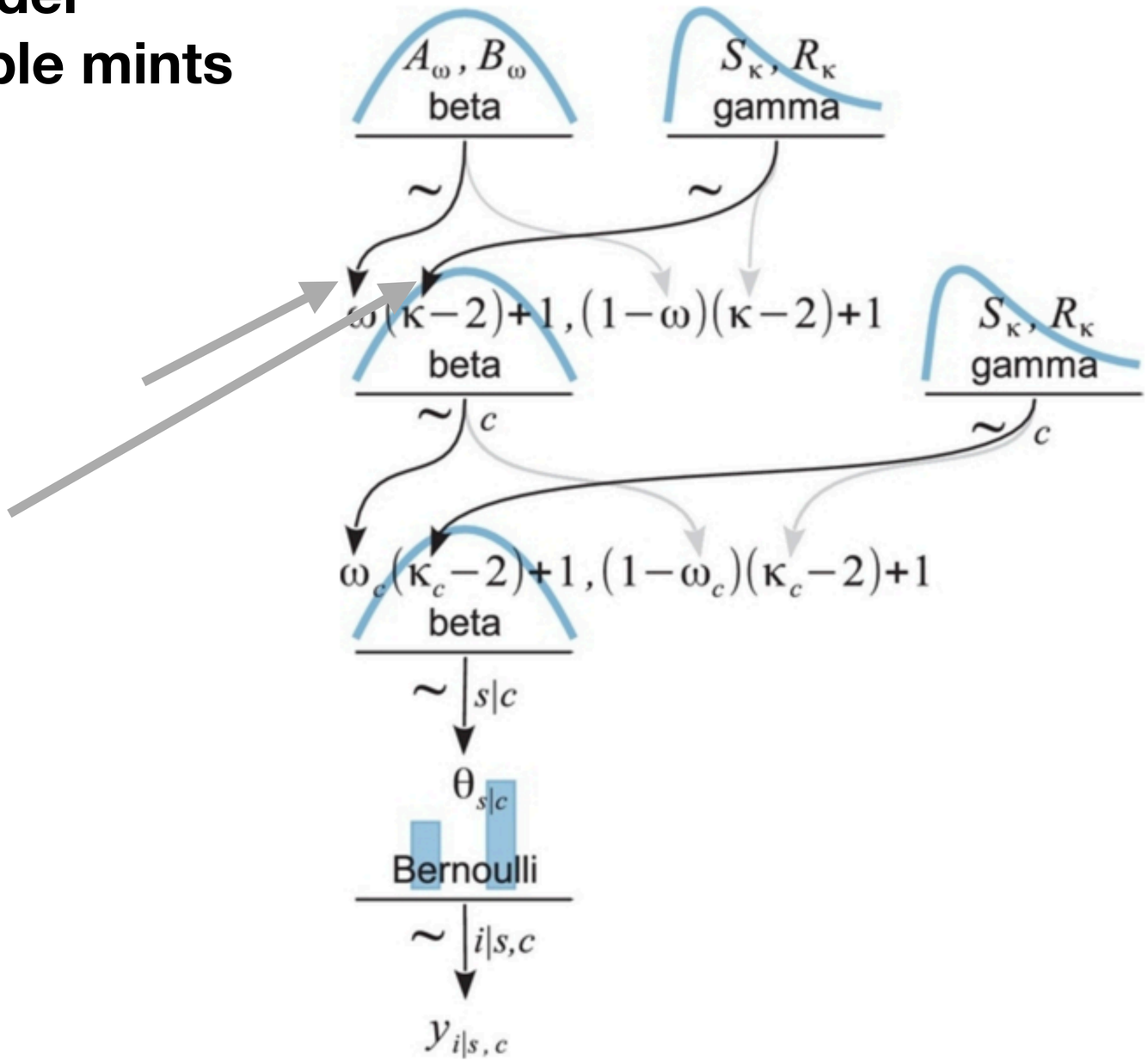
Each coin has its own parameter and dependent on mint



LAB

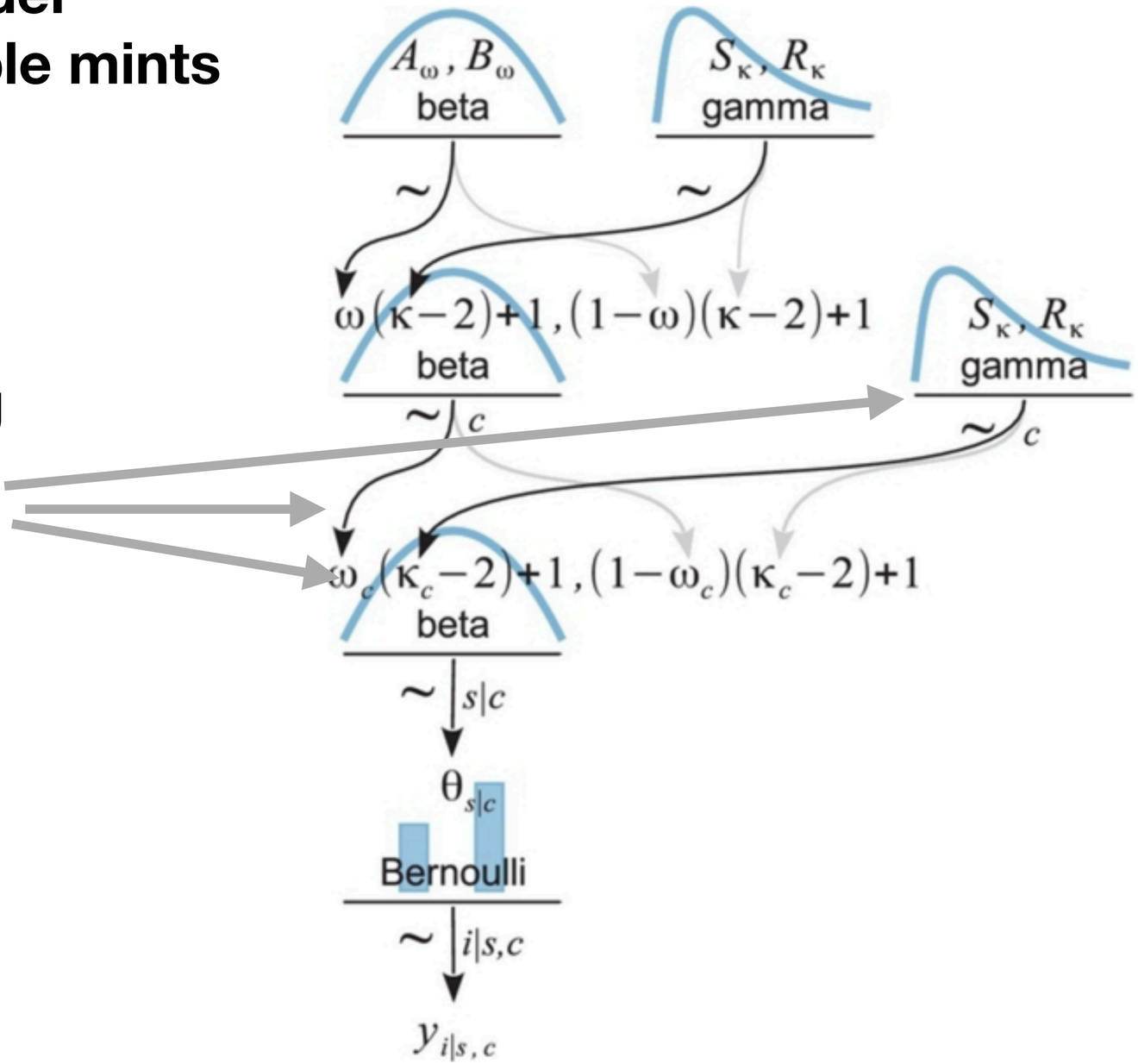
Three levels for statistical model for multiple points from multiple mints

Level 3: Hyperpriors for w0 and kappa0, the parameters of the hyper distribution from which the condition (mint) parameters, wc and kappa_c are sampled.



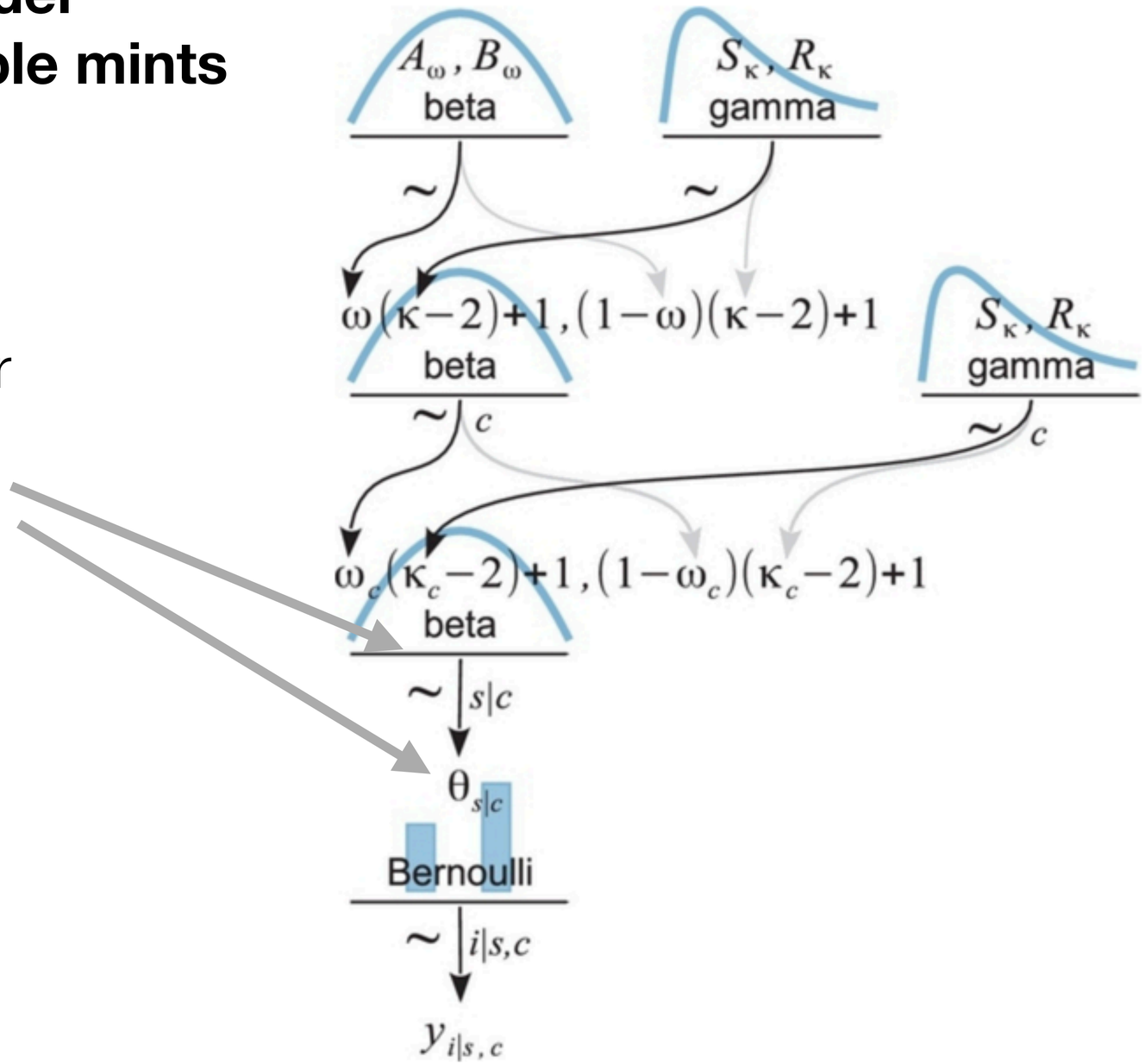
Three levels for statistical model for multiple points from multiple mints

Level 2: Hyperpriors for the condition parameters, wc and kappa_c, that govern the sampling of the individual parameters theta, for subjects within condition c.



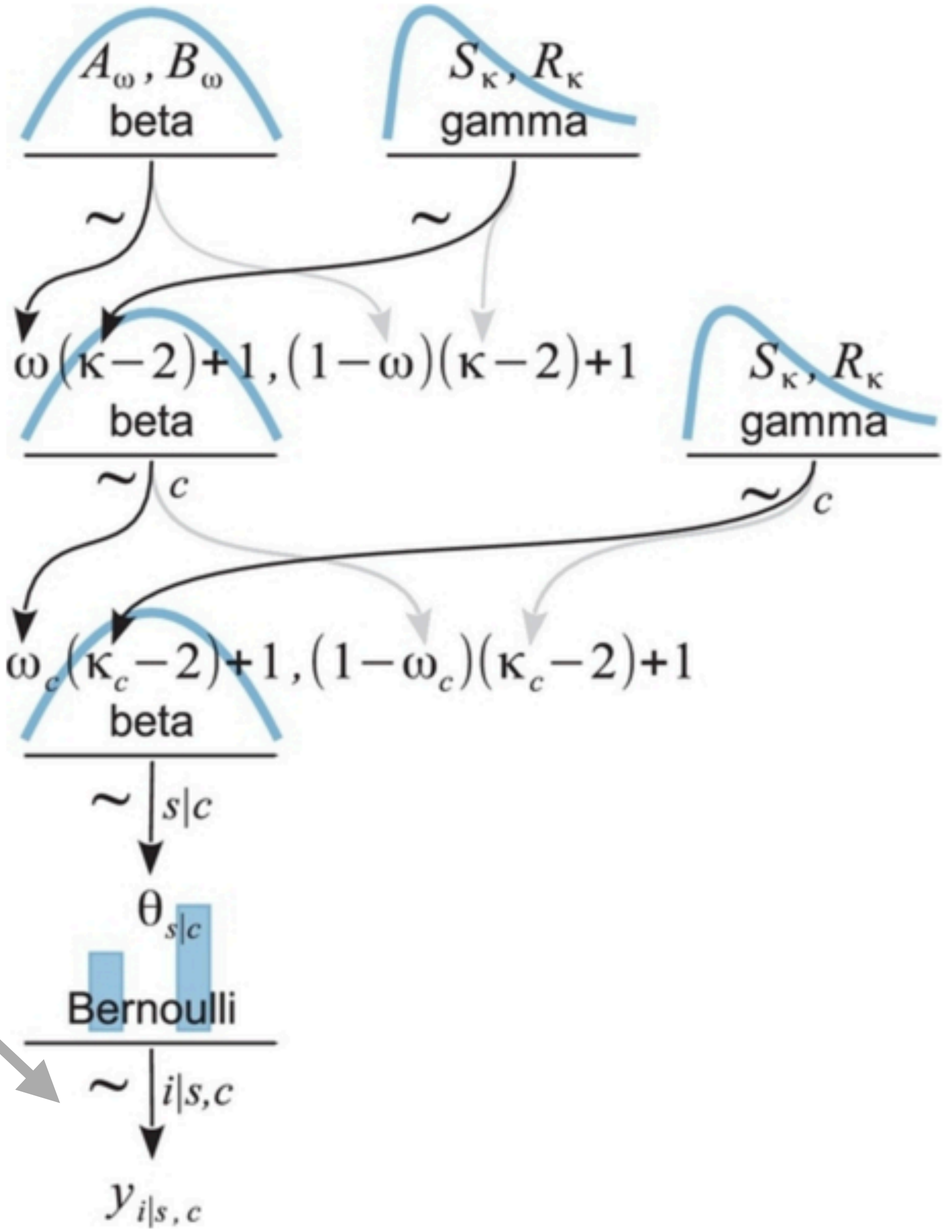
Three levels for statistical model for multiple points from multiple mints

Level 1: Prior for the sampling of individual parameters Theta_s|c for every coin s for the condition to which it was assigned.



Three levels for statistical model for multiple points from multiple mints

Likelihood: Sampling distribution of the observed number of successes yi|s, c



LAB

References

1. Doing Bayesian Data Analysis Second Edition (https://sites.google.com/site/doingbayesiandataanalysis)

2. RStan (https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started)