Lecture 11

Bayesian Stats

Last time: Metropolis, MH

- MH uses asymmetric proposals
- discrete version to generate poisson, for example
- tuning width up decreases acceptance, down increases acceptance
- want acceptance at about 30-40%
- want autocorrelation low, traceplots to look like white noise

Last time: Bayesian

- sample is the data fixed
- parameter is stochastic, has prior and posterior distribution
- posterior: $p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)}$, can summarize via MAP
- just bayes rule: $posterior = \frac{likelihood \times prior}{evidence}$

- evidence: $p(y) = E_{p(\theta)}[\mathcal{L}] = \int d\theta p(y|\theta)p(\theta)$ a normalization, irrelevant for sampling
- What if θ is multidimensional? Marginal posterior: $p(\theta_1|D) = \int d\theta_{-1} p(\theta|D)$.
- posterior predictive: the distribution of a future data point y^*:

$$p(y^*|D = \{y\}) = E_{p(\theta|D)}[p(y|\theta)] = \int d\theta p(y^*|\theta)p(\theta|\{y\}).$$

Today

- sufficient statistics, exchangeability and the poisson-gamma model
- globe toss beta binomial updating and posterior quantities
- normal-normal model and regularization of data
- selection of priors and weakly regularizing priors

Globe Toss Model

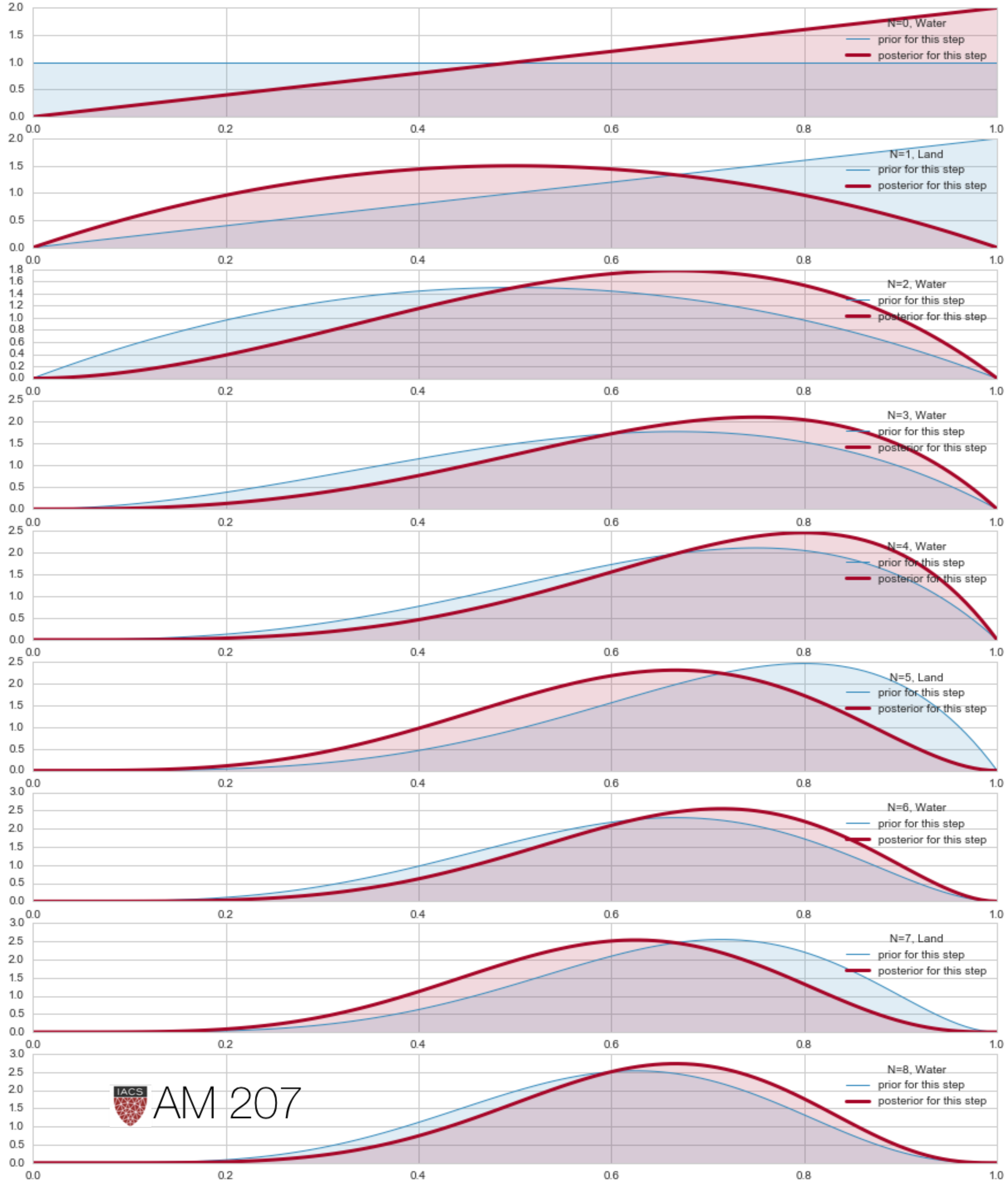
- Seal tosses globe θ is true water fraction
- The Beta distribution is conjugate to the Binomial distribution

$$p(\theta|y) \propto p(y|\theta)P(\theta) = Binom(n, y, \theta) \times Beta(\alpha, \beta)$$

- Because of the conjugacy, this turns out to be:

$$Beta(y + \alpha, n - y + \beta)$$

- a $Beta(1, 1)$ prior is equivalent to a uniform distribution.



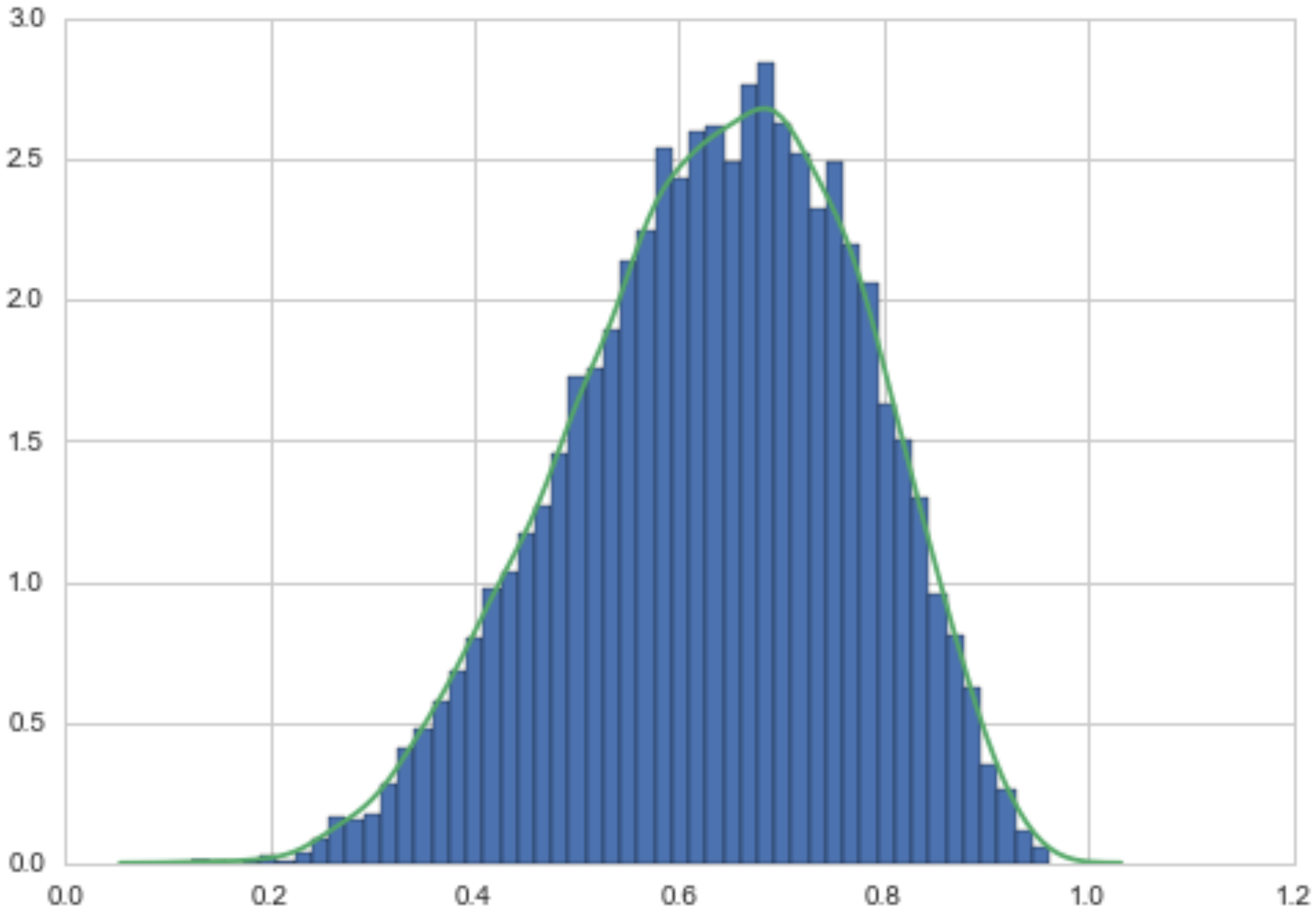
Bayesian Updating of globe

- data `WLWWWLWLW`
- notice how the posterior shifts left and right depending on new data

At each step:

$$Beta(y + \alpha, n - y + \beta)$$

Posterior



- The probability that the amount of water is less than 50%: `np.mean(samples < 0.5) = 0.173`
- Credible Interval: amount of probability mass. `np.percentile(samples, [10, 90]) = [ 0.44604094, 0.81516349]`
- `np.mean(samples), np.median(samples) = (0.63787343440335842, 0.6473143052303143)`

MAP

```
sampleshisto = np.histogram(samples, bins=50)
maxcountindex = np.argmax(sampleshisto[0])
mapvalue = sampleshisto[1][maxcountindex]
print(maxcountindex, mapvalue)
```

```
31 0.662578641304
```

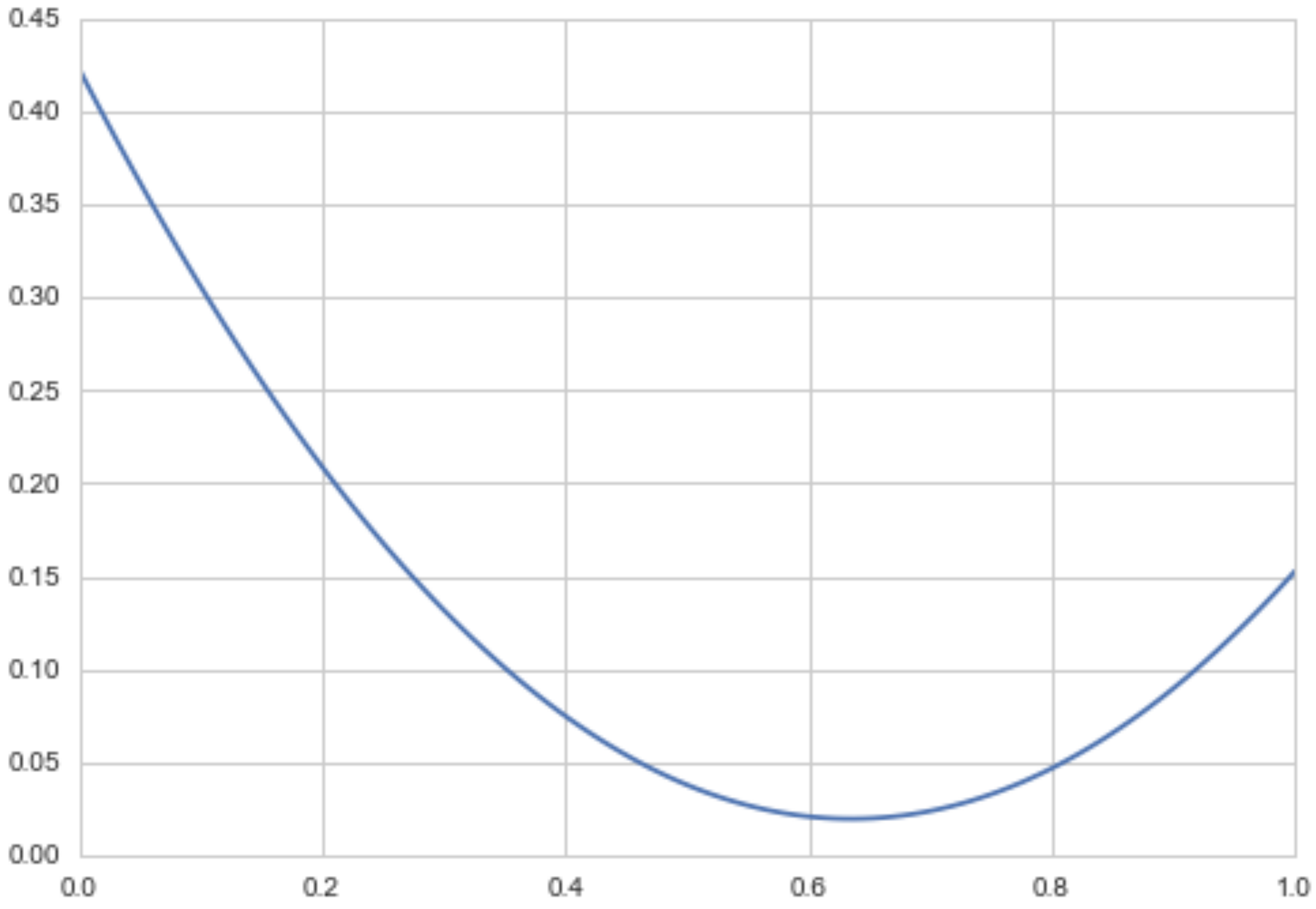
Posterior Mean minimizes squared loss

$$R(t) = E_{p(\theta|D)}[(\theta - t)^2] = \int d\theta(\theta - t)^2 p(\theta|D)$$

$$\frac{dR(t)}{dt} = 0 \implies t = \int d\theta\theta\, p(\theta|D)$$

```
mse = [np.mean((xi-samples)**2) for xi in x]
plt.plot(x, mse);
```

This is **Decision Theory**.



Posterior predictive

$$p(y^*|D) = \int d\theta p(y^*|\theta)p(\theta|D)$$

Risk Minimization holds here too: $y_{minmse} = \int dy\, y\, p(y|D)$

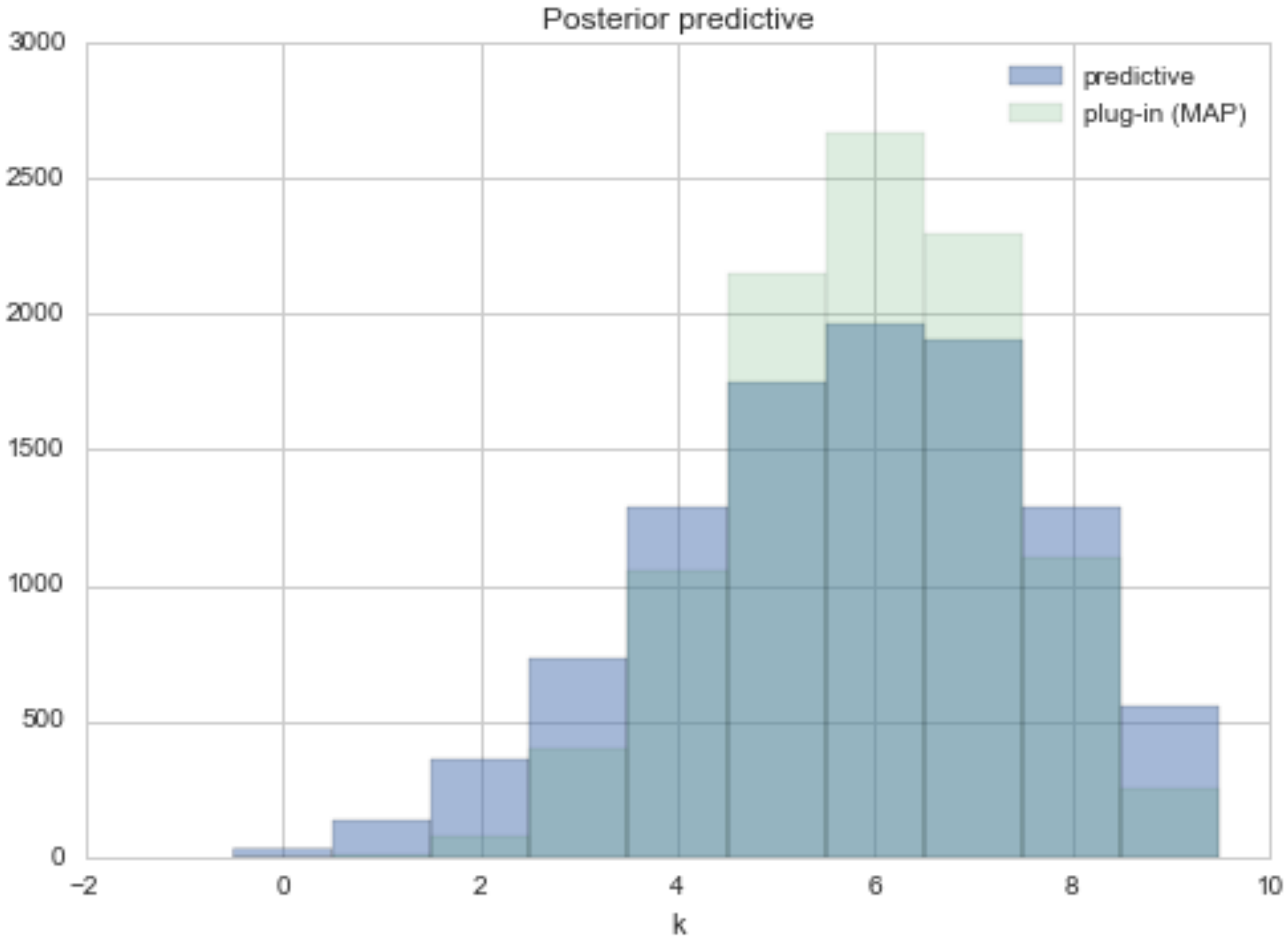
Plug-in Approximation: $p(\theta|D) = \delta(\theta - \theta_{MAP})$ and then draw

$p(y^*|D) = p(y^*|\theta_{MAP})$ a sampling distribution.

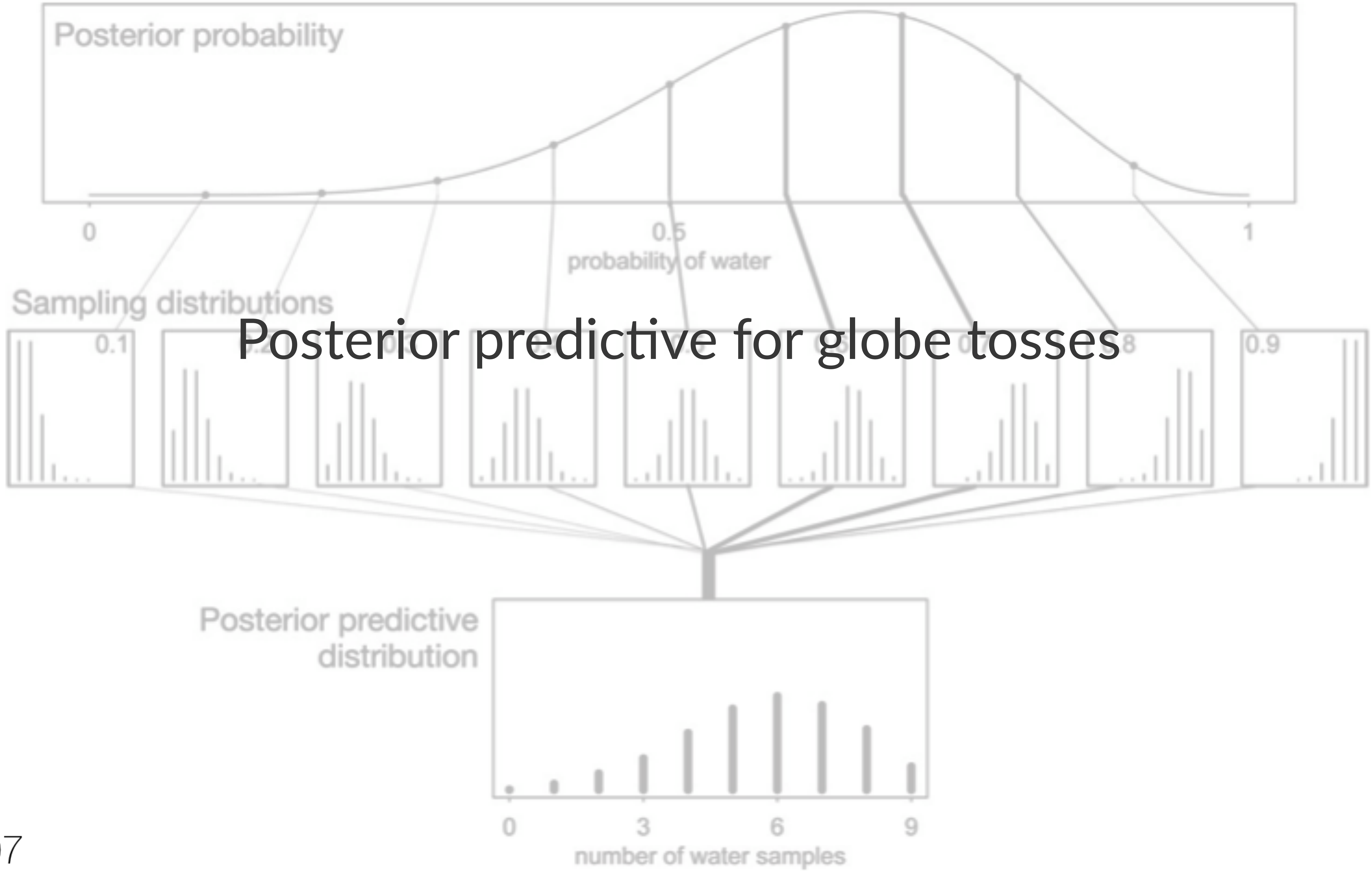
Posterior predictive from sampling

- first draw the thetas from the posterior
- then draw y's from the likelihood
- and histogram the likelihood
- these are draws from joint y, θ

```
postpred = np.random.binomial( len(data), samples);
```



Posterior predictive for globe tosses



Sufficient Statistics and the exponential family

$$p(y_i|\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}.$$

Likelihood: $p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) g(\theta)^n \ \exp\left(\phi(\theta) \sum_{i=1}^{n} u(y_i)\right)$

$\sum_{i=1}^{n} u(y_i)$ is said to be a **sufficient statistic** for θ

Poisson Gamma Example

The data consists of 155 women who were 40 years old. We are interested in the birth rate of women with a college degree and women without. We are told that 111 women without college degrees have 217 children, while 44 women with college degrees have 66 children.

Let $Y_{1,1}, \ldots, Y_{n_1,1}$ children for the n_1 women without college degrees, and $Y_{1,2}, \ldots, Y_{n_2,2}$ for n_2 women with college degrees.

Exchangeability

Lets assume that the number of children of a women in any one of these classes can me modelled as coming from ONE birth rate.

The in-class likelihood for these women is invariant to a permutation of variables.

Poisson likelihood

$$Y_{i,1} \sim Poisson(\theta_1), Y_{i,2} \sim Poisson(\theta_2)$$

$$p(Y_{1,1}, \ldots, Y_{n_1,1}|\theta_1) = \prod_{i=1}^{n_1} p(Y_{i,1}|\theta_1) = \prod_{i=1}^{n_1} \frac{1}{Y_{i,1}!} \theta_1^{Y_{i,1}} e^{-\theta_1}$$

$$= c(Y_{1,1}, \ldots, Y_{n_1,1}) \; (n_1\theta_1)^{\sum Y_{i,1}} e^{-n_1\theta_1} \sim Poisson(n_1\theta_1)$$

$$Y_{1,2}, \ldots, Y_{n_1,2}|\theta_2 \sim Poisson(n_2\theta_2)$$

Posterior

$$c_1(n_1, y_1, \ldots, y_{n_1})\ (n_1\theta_1)^{\sum Y_{i,1}} e^{-n_1\theta_1}\ p(\theta_1) \times c_2(n_2, y_1, \ldots, y_{n_2})\ (n_2\theta_2)^{\sum Y_{i,2}} e^{-n_2\theta_2}\ p(\theta_2)$$

$\sum Y_i$, total number of children in each class of mom, is **sufficient statistics**

Conjugate prior

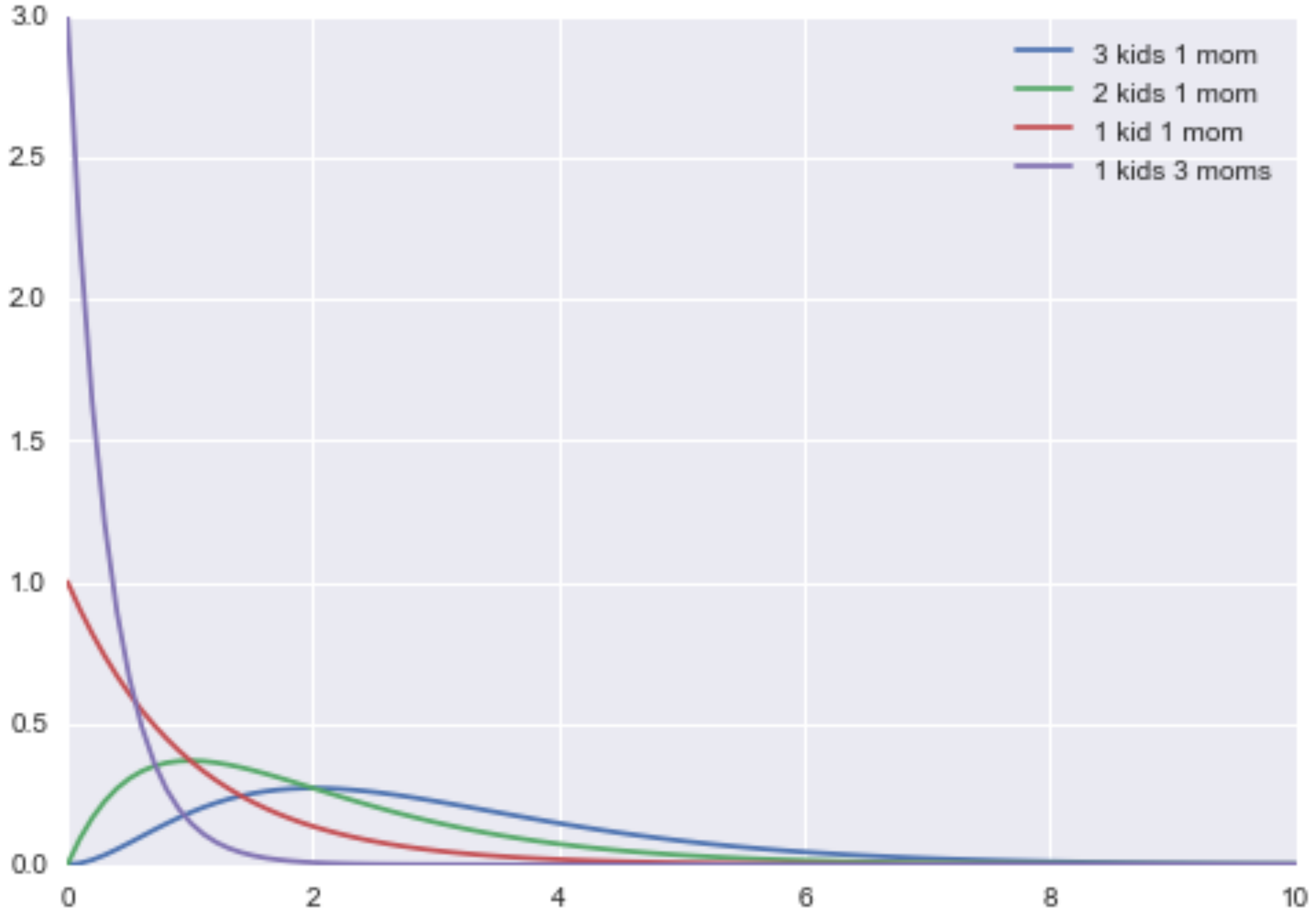
Sampling distribution for θ: $p(Y_1, \ldots, y_n|\theta) \sim \theta^{\sum Y_i} e^{-n\theta}$

Form is of $Gamma$. In shape-rate parametrization (wikipedia)

$$p(\theta) = \mathrm{Gamma}(\theta, \mathrm{a}, \mathrm{b}) = \frac{\mathrm{b}^{\mathrm{a}}}{\Gamma(\mathrm{a})} \theta^{\mathrm{a}-1} \mathrm{e}^{-\mathrm{b}\theta}$$

Posterior:
$p(\theta|Y_1, \ldots, Y_n) \propto p(Y_1, \ldots, y_n|\theta) p(\theta) \sim \mathrm{Gamma}(\theta, \mathrm{a} + \sum \mathrm{Y_i}, \mathrm{b} + \mathrm{n})$

Priors and Posteriors



We choose 2,1 as our prior.

$$p(\theta_1 | n_1, \sum_i^{n_1} Y_{i,1}) \sim \text{Gamma}(\theta_1, 219, 112)$$

$$p(\theta_2 | n_2, \sum_i^{n_2} Y_{i,2}) \sim \text{Gamma}(\theta_2, 68, 45)$$

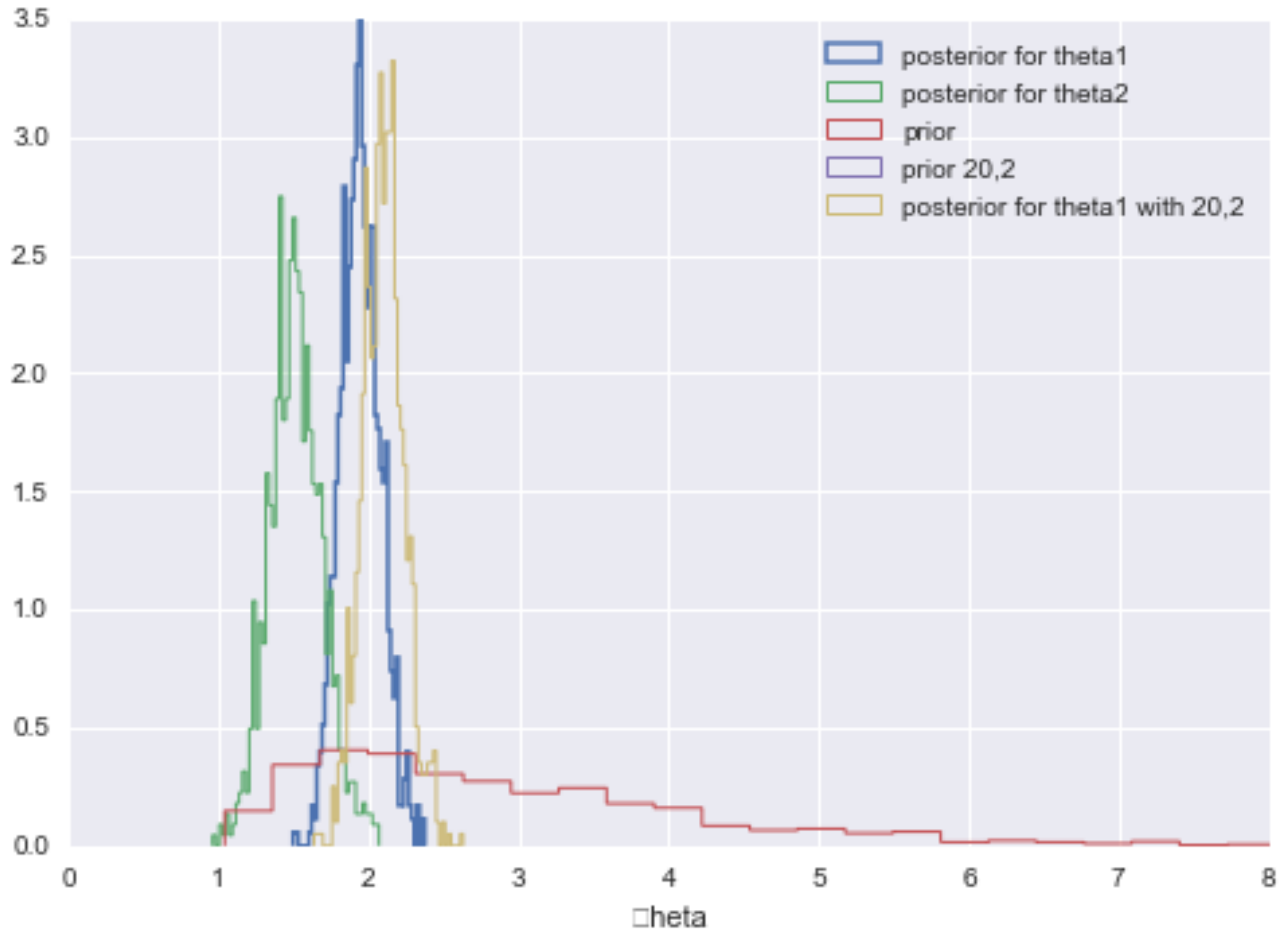
Prior mean, variance:
$E[\theta] = a/b, var[\theta] = a/b^2$.

Posteriors

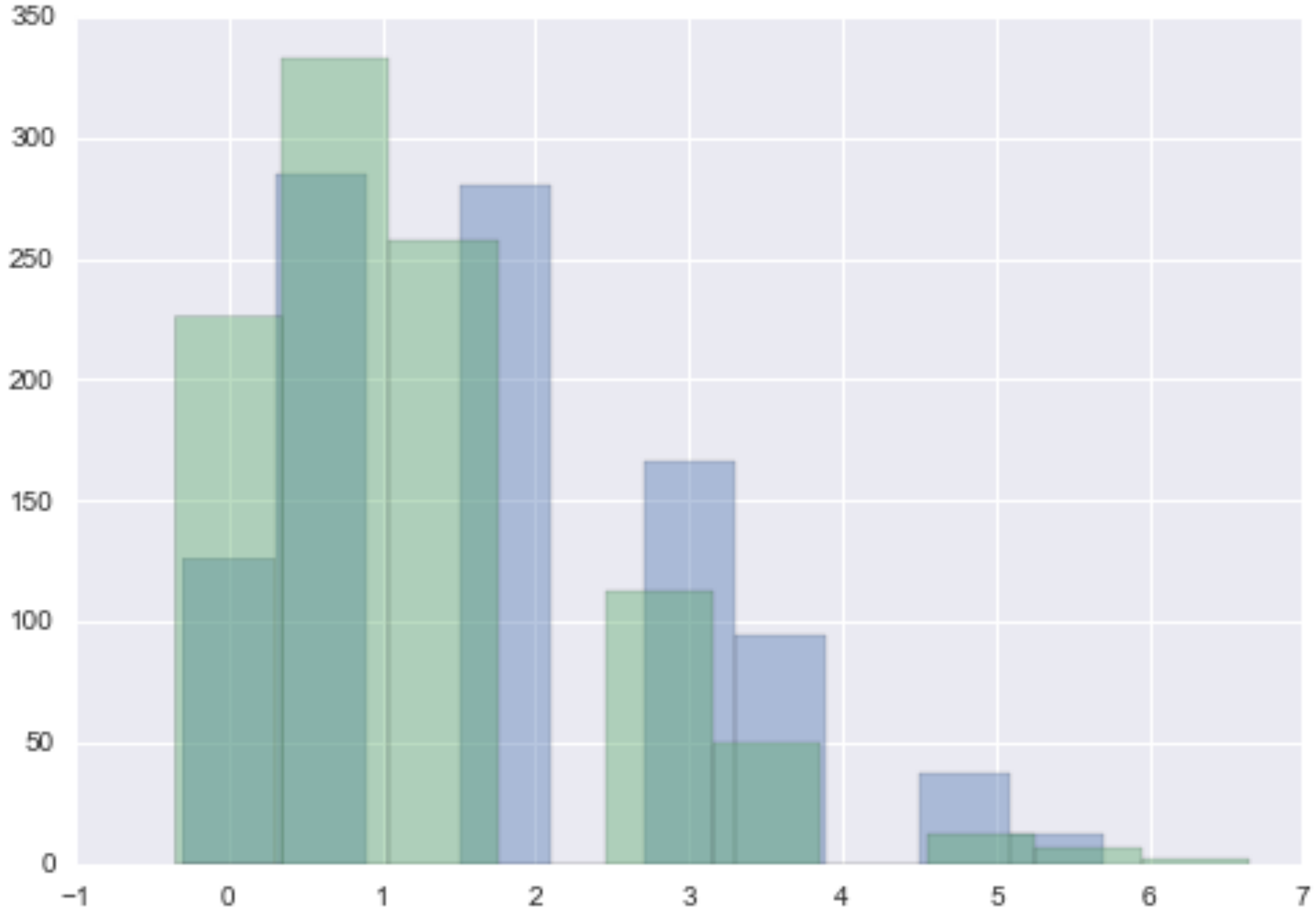
$$E[\theta] = (a + \sum y_i)/(b + N)$$
$$var[\theta] = (a + \sum y_i)/(b + N)^2.$$

```
np.mean(theta1), np.var(theta1)
= (1.9516881521791478,
0.018527204185785785)
```

```
np.mean(theta2), np.var(theta2)
= (1.5037252100213609,
0.034220717257786061)
```



Posterior Predictives



$$p(y^*|D) = \int d\theta p(y^*|\theta)p(\theta|D)$$

Sampling makes it easy:

```
postpred1 = poisson.rvs(theta1)
postpred2 = poisson.rvs(theta2)
```

Negative Binomial:

$$E[y^*] = \frac{(a + \sum y_i)}{(b+N)}$$

$$var[y^*] = \frac{(a + \sum y_i)}{(b+N)^2}(N+b+1).$$

But see width:

`np.mean(postpred1), np.var(postpred1)=(1.976, 1.8554239999999997)`

Posterior predictive smears out posterior error with sampling distribution

- use for making predictions
- use for model checking using cross-validation; also for data visualization

Normal-Normal Model

Posterior for a gaussian likelihood:

$$p(\mu, \sigma^2 | y_1, \ldots, y_n, \sigma^2) \propto \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} \sum (y_i - \mu)^2} \, p(\mu, \sigma^2)$$

What is the posterior of μ assuming we know σ^2?

Prior for σ^2 is $p(\sigma^2) = \delta(\sigma^2 - \sigma_0^2)$

$$p(\mu|y_1,\ldots,y_n,\sigma^2=\sigma_0^2) \propto p(\mu|\sigma^2=\sigma_0^2)\, e^{-\frac{1}{2\sigma_0^2}\sum(y_i-\mu)^2}$$

The conjugate of the normal is the normal itself.

Say we have the prior

$$p(\mu|\sigma^2) = \exp\left\{-\frac{1}{2\tau^2}(\hat{\mu}-\mu)^2\right\}$$

posterior: $p(\mu|y_1,\ldots,y_n,\sigma^2) \propto \exp\left\{-\frac{a}{2}(\mu-b/a)^2\right\}$

Here

$$a = \frac{1}{\tau^2} + \frac{n}{\sigma_0^2}, \qquad b = \frac{\hat{\mu}}{\tau^2} + \frac{\sum y_i}{\sigma_0^2}$$

Define $\kappa = \sigma^2/\tau^2$

$$\mu_p = \frac{b}{a} = \frac{\kappa}{\kappa + n}\hat{\mu} + \frac{n}{\kappa + n}\bar{y}$$

which is a weighted average of prior mean and sampling mean.

The variance is

$$\tau_p^2 = \frac{1}{1/\tau^2 + n/\sigma^2}$$

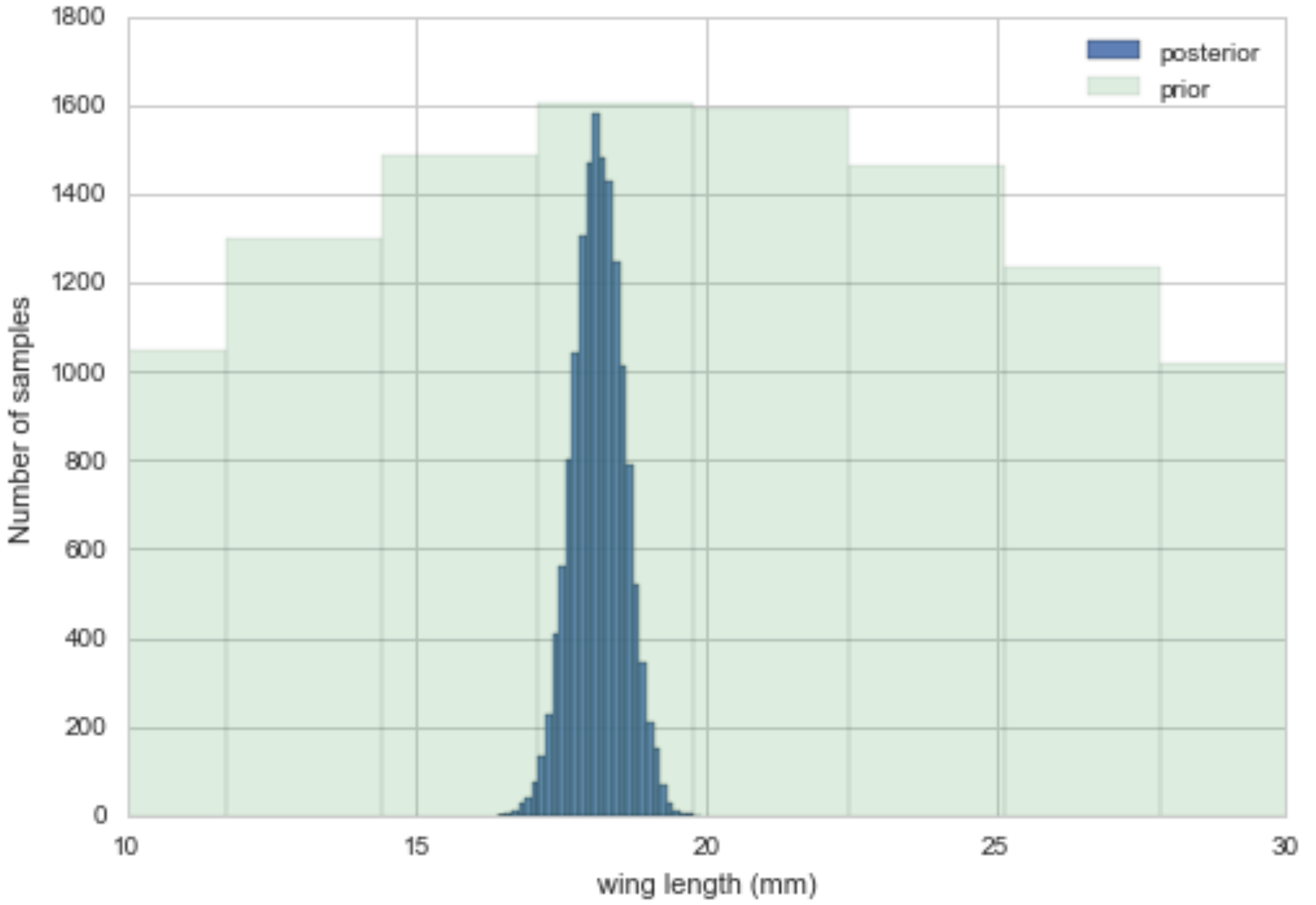
or better

$$\frac{1}{\tau_p^2} = \frac{1}{\tau^2} + \frac{n}{\sigma^2}.$$

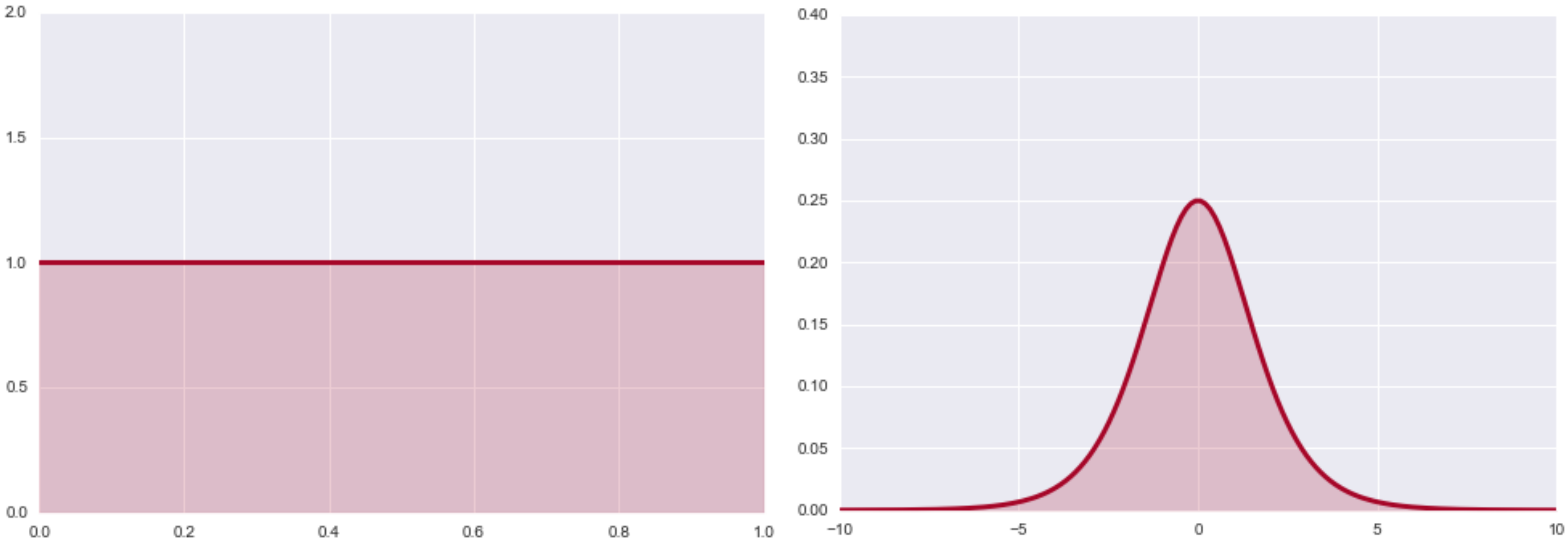
as n increases, the data dominates the prior and the posterior mean approaches the data mean, with the posterior distribution narrowing...

Posterior vs prior

```
Y = [16.4, 17.0, 17.2, 17.4, 18.2, 18.2, 18.2, 19.9, 20.8]
#Data Quantities
sig = np.std(Y) # assume that is the value of KNOWN sigma (in the likelihood)
mu_data = np.mean(Y)
n = len(Y)
# Prior mean
mu_prior = 19.5
# prior std
tau = 10
# plug in formulas
kappa = sig**2 / tau**2
sig_post =np.sqrt(1./( 1./tau**2 + n/sig**2));
# posterior mean
mu_post = kappa / (kappa + n) *mu_prior + n/(kappa+n)* mu_data
#samples
N = 15000
theta_prior = np.random.normal(loc=mu_prior, scale=tau, size=N);
theta_post = np.random.normal(loc=mu_post, scale=sig_post, size=N);
```



Uninformative priors on location



- despite transformation change, flat priors still used for location priors
- may even be improper, ie integrate to ∞ as long as posterior integral is finite
- e.g. flat prior on mean in normal-normal model with strong likelihood.

Jeffreys prior

noninformative prior on scale variables $p_J(\theta) \propto \mathbf{I}(\theta)^{1/2}$

where

$$\mathbf{I}(\theta) = det(-E\left[\frac{d^2 \log p(X|\theta)}{d\theta_i\theta_j}\right])$$

is the Fisher Information, and expectation is with respect to the likelihood.

J for Normal Model

Known σ:

$I \propto E_{f|\sigma}\left[\frac{1}{\sigma^2}\right] = \frac{1}{\sigma^2};\; p_J(\mu) \propto 1/\sigma$: fixed σ improper uniform....

Known μ:

$$I = E_{f|\mu}\left[\frac{d^2}{d\sigma^2}(log(\sigma) + (x-\mu)^2/2\sigma^2)\right] = E_{f|\mu}\left[-\frac{1}{\sigma^2} + 3\frac{(x-\mu)^2}{\sigma^4}\right] = \frac{2}{\sigma^2}$$

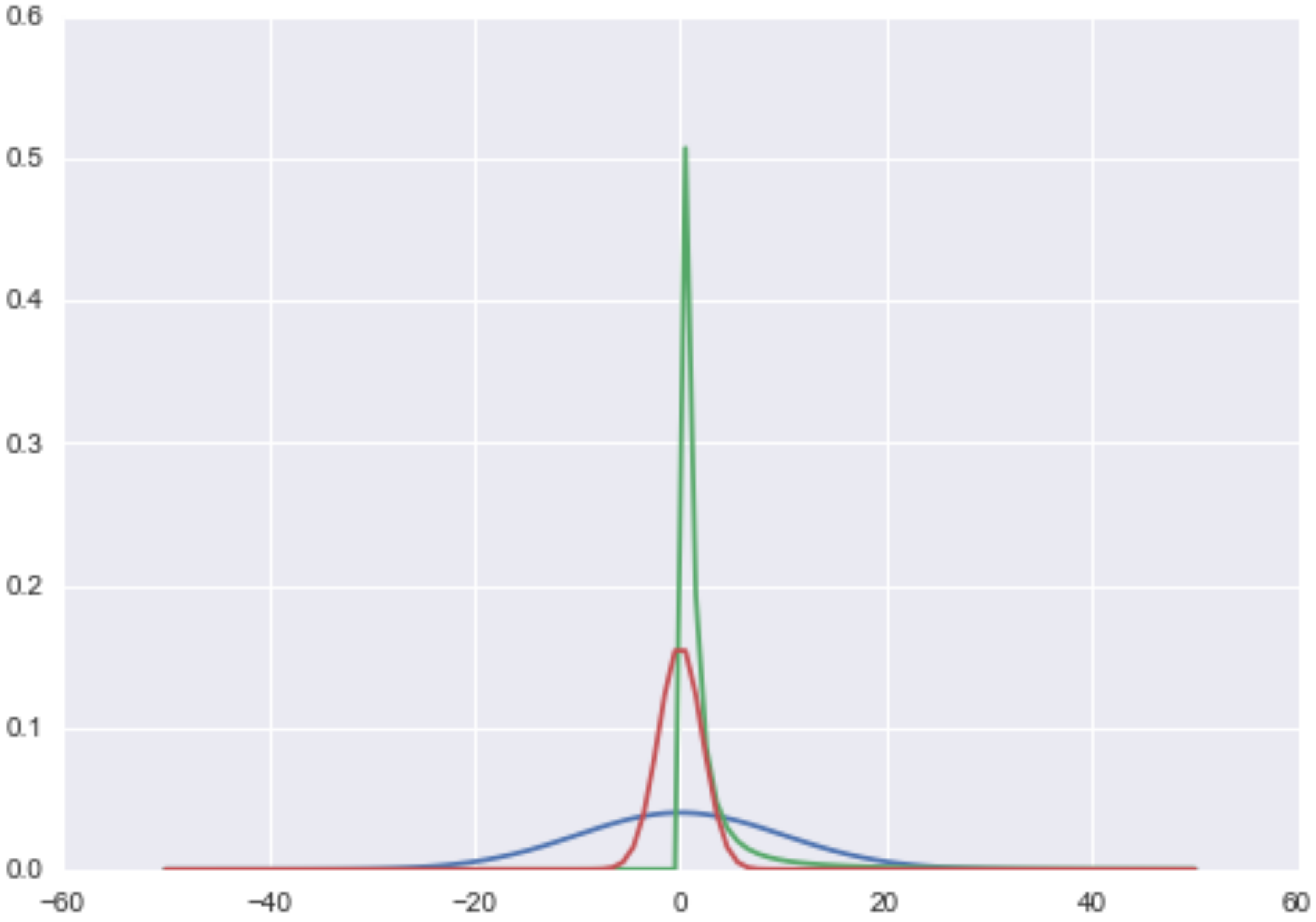
$$p_J(\sigma) \propto 1/\sigma$$

Weakly informative or regularizing priors

- these are the priors we will concern ourselves most with
- restrict parameter ranges
- help samplers

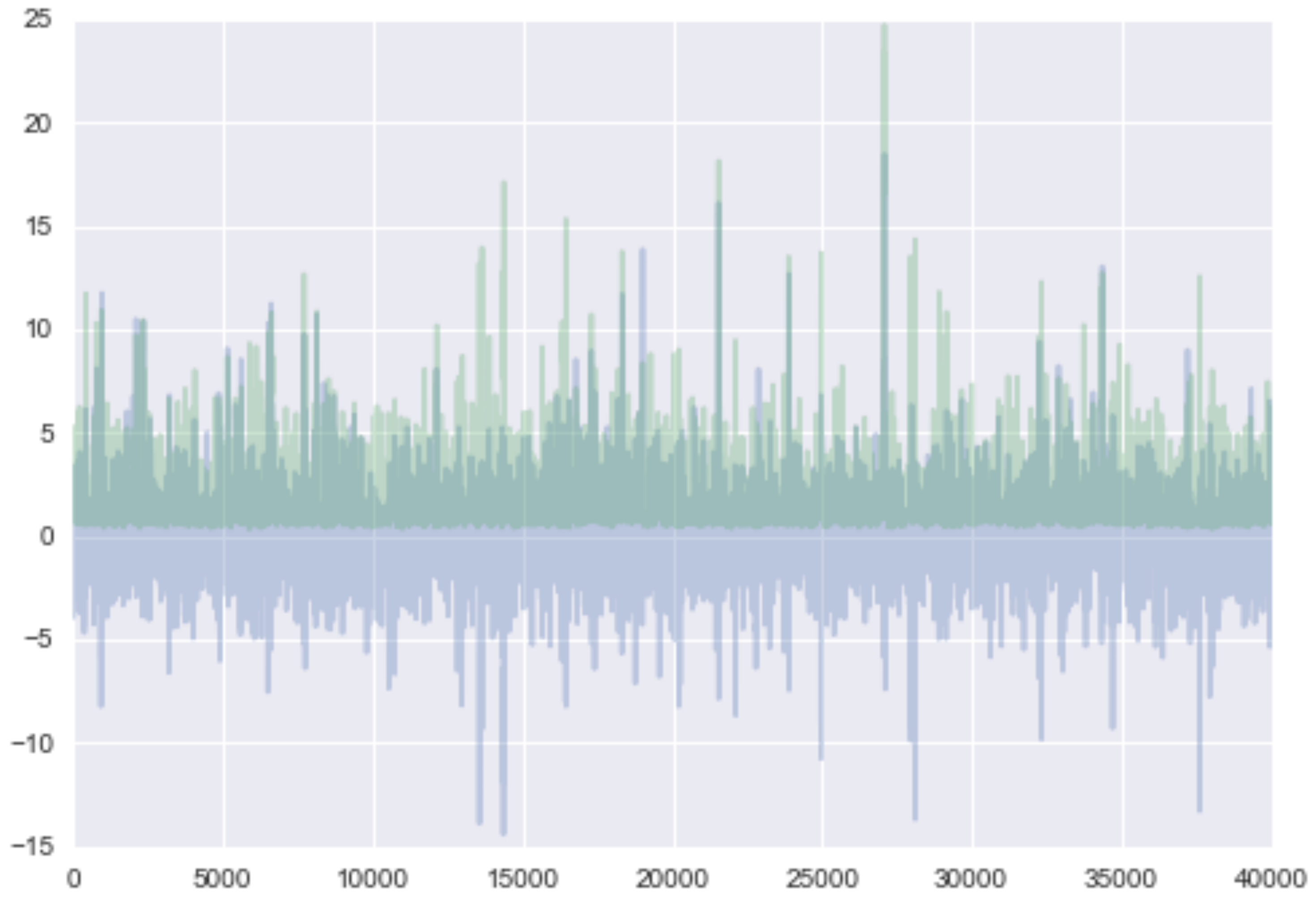
Normal model Example

- two data points 1 and -1
- flat improper priors on $\mu, \sigma > 0$
- model drifts wildly as less data
- flat priors say extreme implausible values quite likely
- extreme drifts overwhelm chain

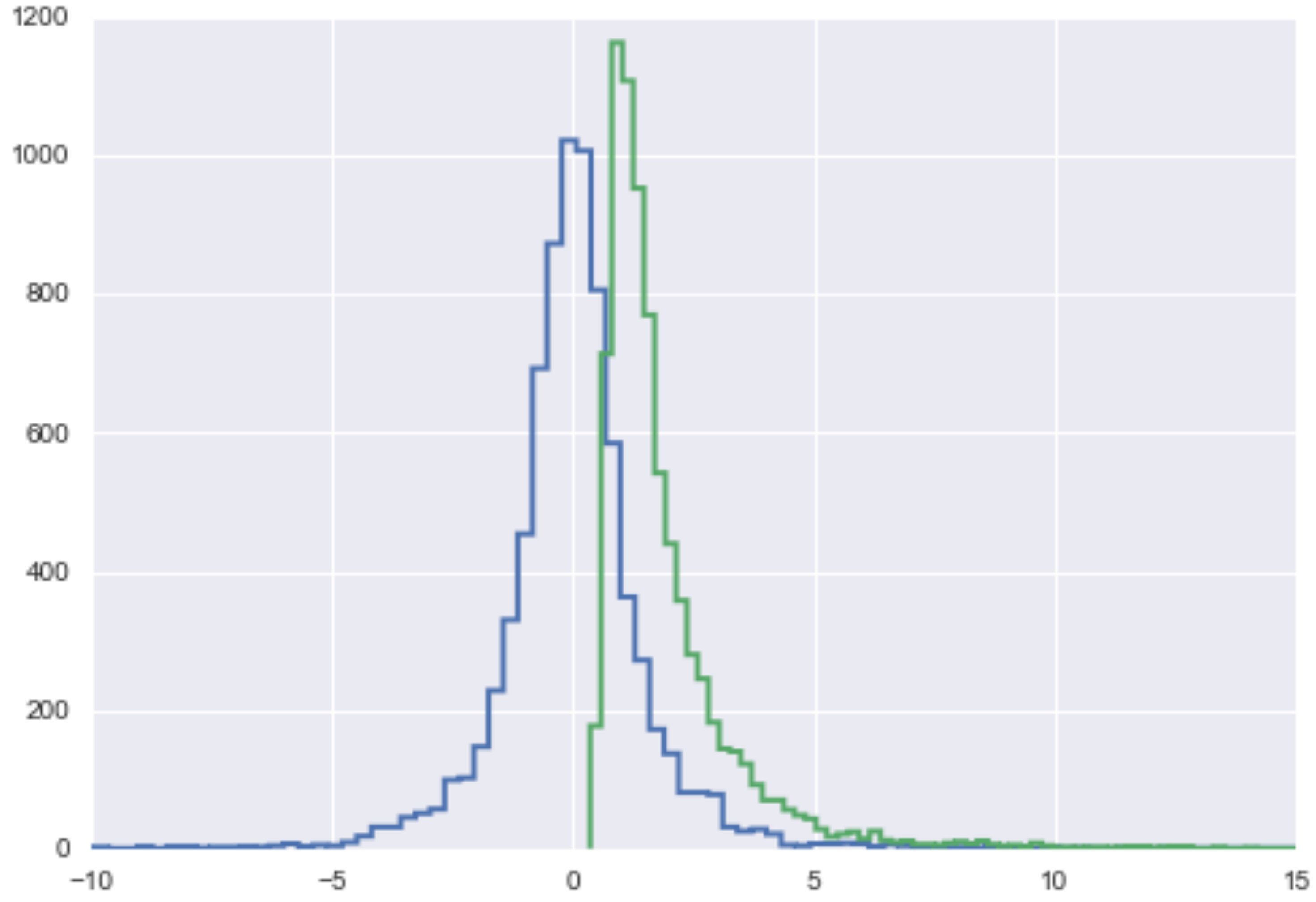


weakly regularizing priors

- choose $\mu \sum N(0,10)$
- choose $\sigma \sim HalfCauchy(0,1)$
- lets mean vary widely but not crazily
- HalfCauchy lets variance be positive and occasionally can have high value samples



25
20
15
10
5
0
−5
−10
−15
0
5000
10000
15000
20000
25000
30000
35000
40000



1200
1000
800
600
400
200
0
−10
−5
0
5
10
15

Other priors

- KL Maximization non-informative prior by Bernardo
- Maximum Entropy prior when some assumptions but no more..
- Empirical bayes prior: usee data! in hierarchical models

Data overwhelms prior