Lecture 7

From Information Theory to Bayesian Stats

Last Time:

- The Learning process
- Risk and Bayes Risk
- The KL Divergence and Deviance
- In-sample penalties: the AIC

Today

- Entropy
- Maximum Likelihood and Entropy
- Bayesian Stats
- Exponential Family

HW submissions

- are having too much entropy, SO
- put care into your submission, its a real-world document and must be well organized and neat.
- Dont leave stray code around document. Cite sources.
- use jupyter notebooks only. Not Colab, not additional python files.
- One notebook per submission please!
- learn how to use markdown and latex-in-markdown well.

Submission format

- only one per group
- **all names in group clearly at top of the document**
- **Name notebook** thus: `AM207_HWx.ipynb`
- Submit via a canvas group. Create a group in the people section. Then when one person submits you're all notified
- Please follow, or **TFs will start penalizing**

KL-Divergence

$$D_{KL}(p,q) = E_p[log(p) - log(q)] = E_p[log(p/q)]$$
$$= \sum_i p_i log(\frac{p_i}{q_i}) \; or \; \int dP log(\frac{p}{q})$$

$$D_{KL}(p,p) = 0$$

KL divergence measures distance/dissimilarity of the two distributions $p(x)$ and $q(x)$. Its >= 0.

Divergence:

The additional uncertainty induced by using probabilities from one distribution to describe another distribution

- McElreath page 179

MARS ATTACKS (Topps, 1962; Burton 1996)

$$Earth : q = \{0.7, 0.3\}, Mars : p = \{0.01, 0.99\}.$$

Earth to predict Mars, less surprise on landing: $D_{KL}(p, q) = 1.14, D_{KL}(q, p) = 2.62$.

PROBLEM: we dont know distribution p. If we did, why do inference?

SOLUTION: Use the empirical distribution

That is, approximate population expectations by sample averages.

$$\implies D_{KL}(p,q) = E_p[log(p/q)] = \frac{1}{N}\sum_i log(p_i/q_i)$$

Maximum Likelihood justification

$$D_{KL}(p, q) = E_p[log(p/q)] = \frac{1}{N}\sum_i (log(p_i) - log(q_i))$$

Minimizing KL-divergence $\implies$ maximizing $\sum_i log(q_i)$

Which is exactly the log likelihood! MLE!

Information and Uncertainty

- coin at 50% odds has maximal uncertainty
- reflects my lack of knowledge of the physics
- many ways for 50% heads.
- an election with $p = 0.99$ has a lot of Information

information is the reduction in uncertainty from learning an outcome

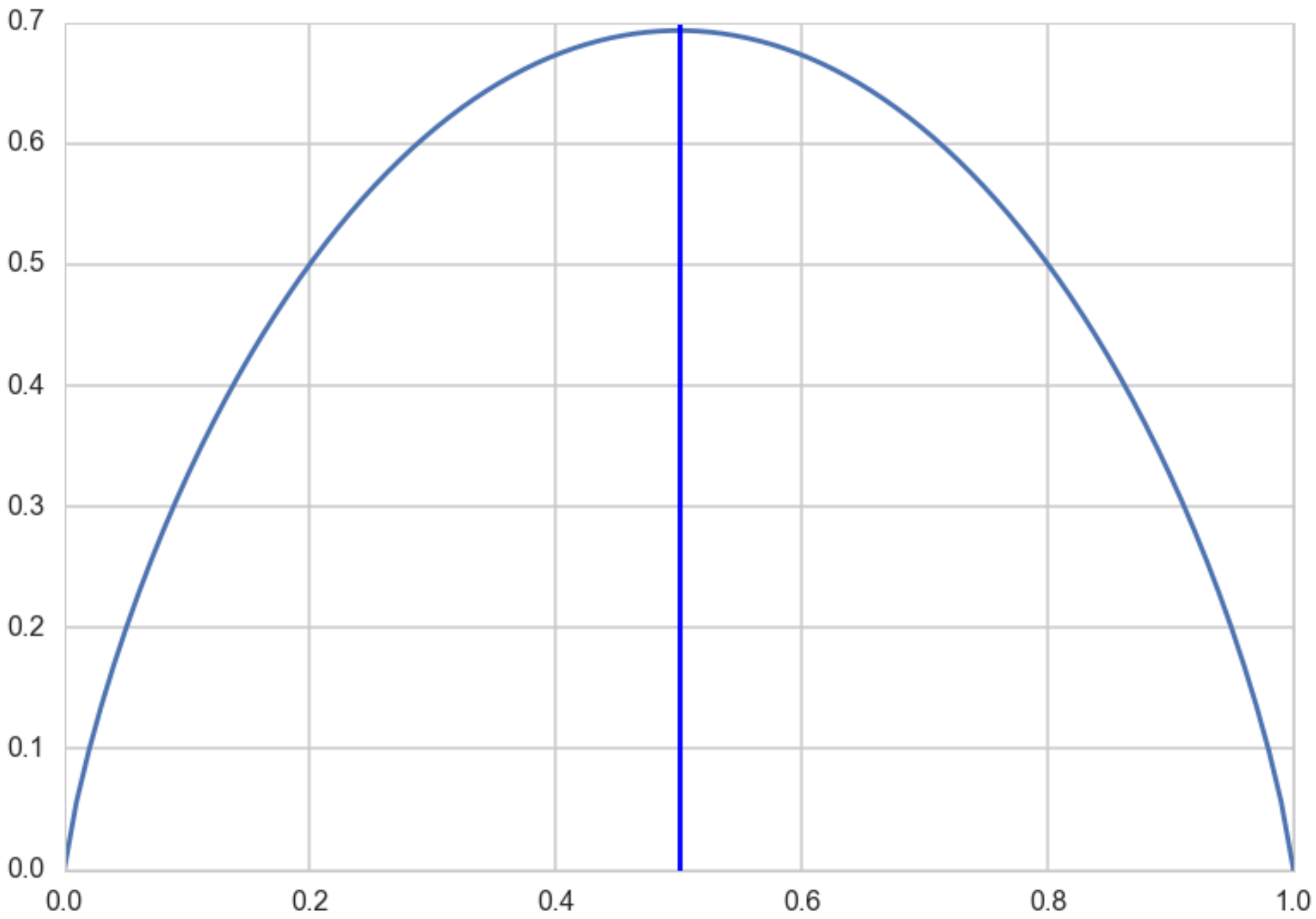
Information Entropy, a measure of uncertainty

Desiderata:

- must be continuous so that there are no jumps
- must be additive across events or states, and must increase as the number of events/states increases

$$H(p) = -E_p[log(p)] = -\int p(x)log(p(x))dx \;\; OR \; -\sum_i p_i log(p_i)$$

Entropy for coin fairness



$$H(p) = -E_p[log(p)] = -p * log(p) - (1-p) * log(1-p)$$

```
def h(p):
    if p==1.:
        ent = 0
    elif p==0.:
        ent = 0
    else:
        ent = - (p*math.log(p) + (1-p)* math.log(1-p))
```

Maximum Entropy (MAXENT)

- finding distributions consistent with constraints and the current state of our information
- what would be the least surprising distribution?
- The one with the least additional assumptions?

The distribution that can happen in the most ways is the one with the highest entropy

For a gaussian

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$H(p) = E_p[log(p)] = E_p[-\frac{1}{2} log(2\pi\sigma^2) - (x-\mu)^2/2\sigma^2]$$

$$= -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2\sigma^2} E_p[(x-\mu)^2] = -\frac{1}{2} log(2\pi\sigma^2) - \frac{1}{2} = \frac{1}{2} log(2\pi e\sigma^2)$$

Cross Entropy

$$H(p, q) = -E_p[log(q)]$$

Then one can write:

$$D_{KL}(p, q) = H(p, q) - H(p)$$

KL-Divergence is additional entropy introduced by using q instead of p.

We saw this for Logistic regression

- $H(p, q)$ and $D_{KL}(p, q)$ are not symmetric.
- if you use a unusual , low entropy distribution to approximate a usual one, you will be more surprised than if you used a high entropy, many choices one to approximate an unusual one.

Corollary: if we use a high entropy distribution to approximate the true one, we will incur lesser error.

Gaussian is MAXENT for fixed mean and variance

Consider
$D_{KL}(q,p) = E_q[log(q/p)] = H(q,p) - H(q) >= 0$

$$H(q,p) = E_q[log(p)] = -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}E_q[(x-\mu)^2]$$

$E_q[(x-\mu)^2]$ is CONSTRAINED to be σ^2.

$$H(q,p) = -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2} = -\frac{1}{2}log(2\pi e\sigma^2) = H(p) >= H(q)!!!$$

EXPONENTIAL FAMILY

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

Likelihood in 1D:

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

Example: Normal $f(y) = (1/\sigma\sqrt{2\pi})e^{-x^2/2\sigma^2}$, $u(y) = x/\sigma$, $g(\mu) = e^{-\mu^2/2\sigma^2}$, $\phi(\mu) = \mu/\sigma$

See wikipedia for more.

Importance of MAXENT

- most common distributions used as likelihoods (and priors) are in the exponential family, MAXENT subject to different constraints.
- gamma: MAXENT all distributions with the same mean and same average logarithm.
- exponential: MAXENT all non-negative continuous distributions with the same average inter-event displacement

Importance of MAXENT

- Information entropy enumerates the number of ways a distribution can arise, after having fixed some assumptions.
- choosing a maxent distribution as a likelihood means that once the constraints has been met, no additional assumptions.

The most conservative distribution

Bayesian statistics

Frequentist Stats

- parameters are fixed, data is stochastic
- true parameter θ^* characterizes population
- we estimate $\hat{\theta}$ on sample
- we can use MLE $\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}$
- we obtain sampling distributions (using bootstrap)
- predictive distribution through the sampling distribution

Frequentist Bestiary

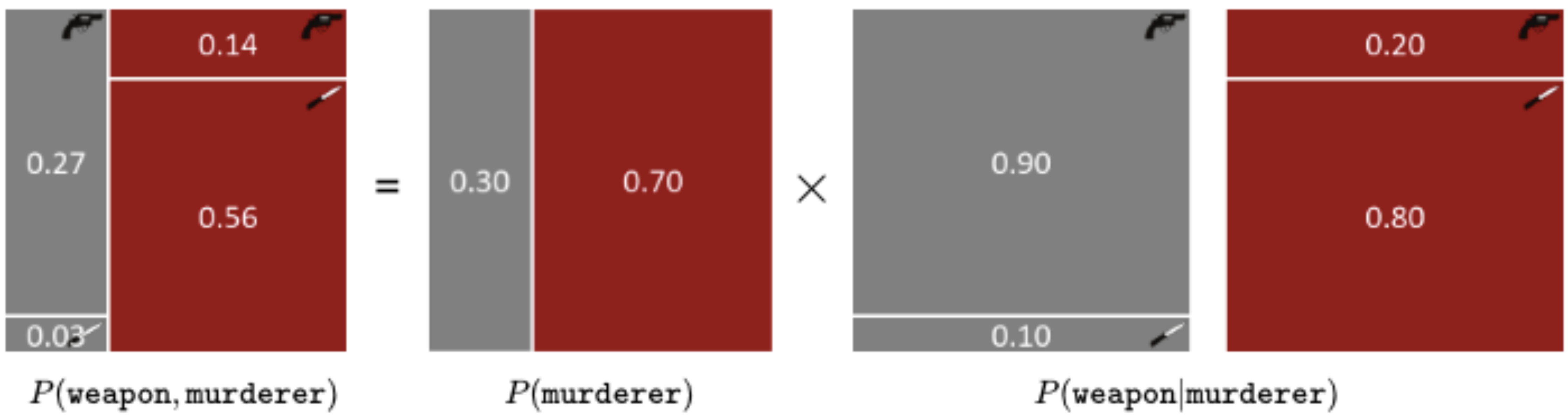
- Parameter sampling distribution
- predictive distribution
- MLE (or other point) estimate

Bayesian Stats

- assume sample IS the data, no stochasticity
- parameters θ are stochastic random variables
- associate the parameter θ with a prior distribution $p(\theta)$
- The prior distribution generally represents our belief on the parameter values when we have not observed any data yet (to be qualified later)
- obtain posterior distributions
- predictive distribution from the posterior

Basic Idea

Get the joint Probability distribution



Now we condition on some random variables and learn the values of others.

Rules

1. $P(A, B) = P(A \mid B)P(B)$
2. $P(A) = \sum_B P(A, B) = \sum_B P(A \mid B)P(B)$

$P(A)$ is called the **marginal** distribution of A, obtained by summing or marginalizing over B.

Posterior distribution from Bayes Rule

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)}$$

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{p(y)}$$

$$p(\theta|D=\{y\}) = \frac{p(D|\theta)\,p(\theta)}{p(D)}$$

$$p(\theta|D) \propto p(D|\theta)\,p(\theta)$$

Evidence

$p(D)$ or $p(y)$ (**marginal distribution of y**) the expected likelihood (on existing data points) over the prior $E_{p(\theta)}[\mathcal{L}]$:

$$p(y) = \int d\theta\, p(\theta, y) = \int d\theta\, p(y|\theta)p(\theta).$$

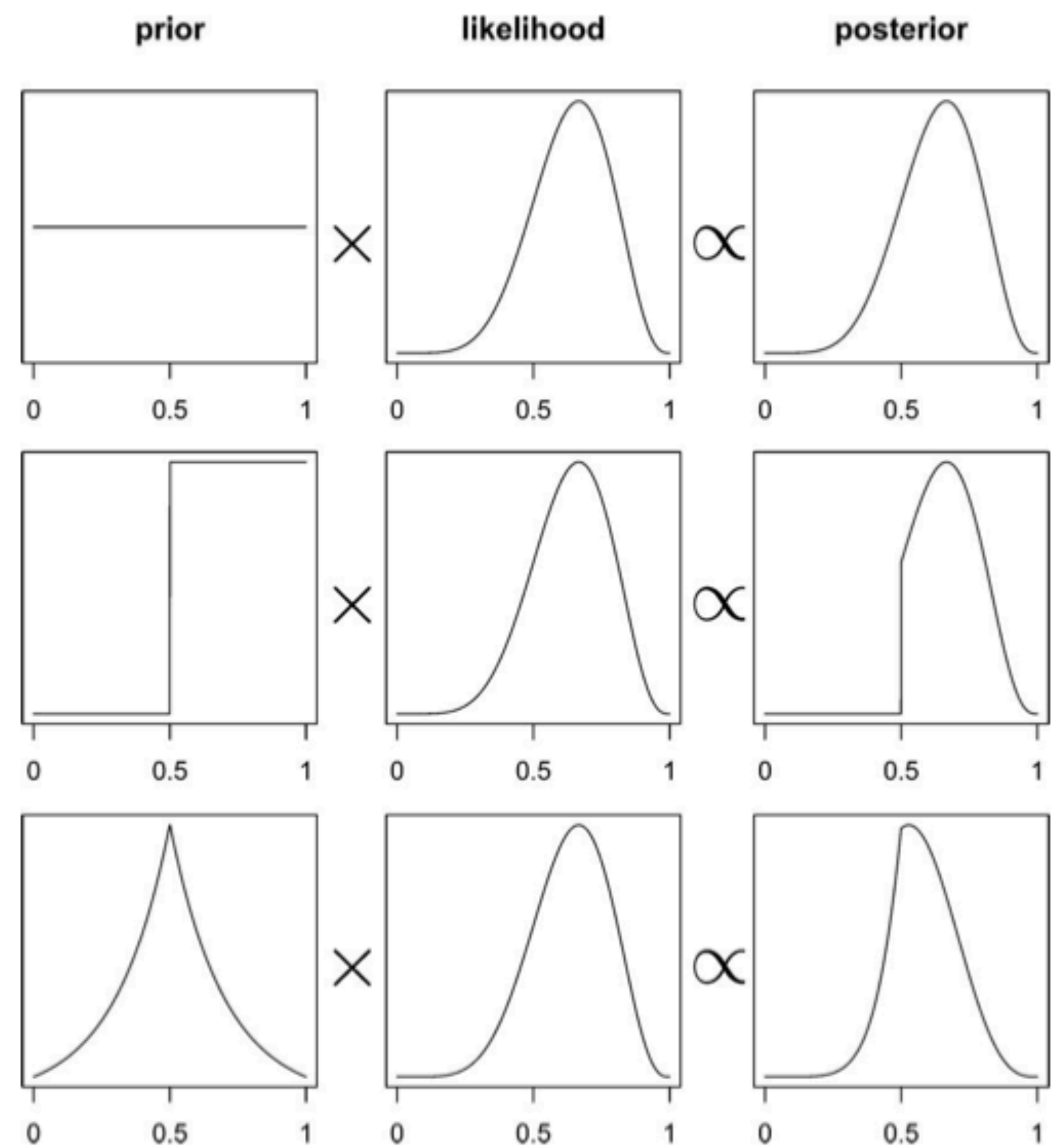
$$p(D = \{y\}) = \int d\theta\, p(\theta, D) = \int d\theta\, p(D|\theta)p(\theta).$$

Posterior

$$posterior = \frac{likelihood \times prior}{evidence}$$

$$posterior \propto likelihood \times prior$$

- evidence is just the normalization
- usually dont care about normalization (until model comparison), just pdf/pmf or samples



Marginalization

What if θ is multidimensional?

Integrate the posterior over all "other" or "nusisance" parameters.

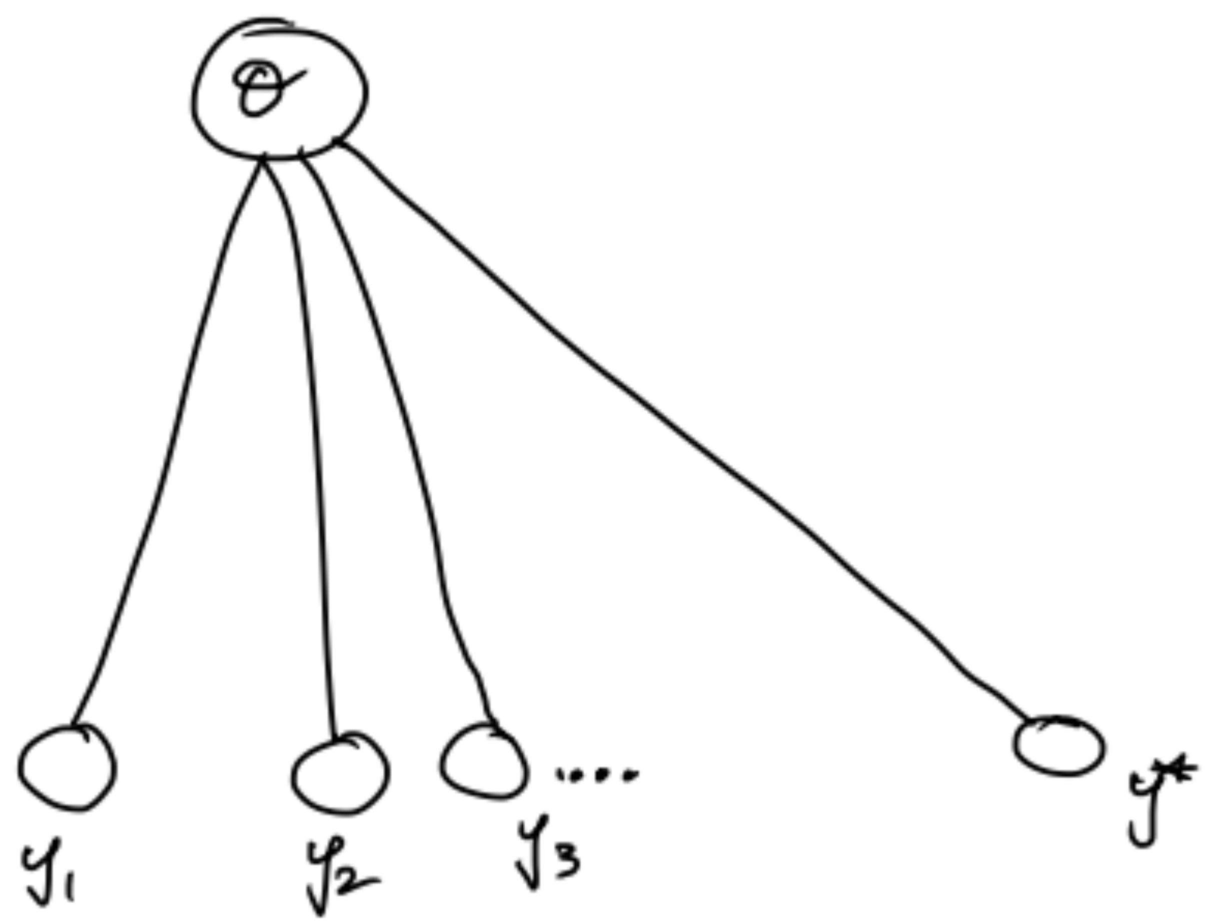
Marginal posterior: $p(\theta_1|D) = \int d\theta_{-1} p(\theta|D)$.

Basic Graph

$$p(\theta, y, y^*) = p(\theta)p(y|\theta)p(y^*|\theta)$$
$$= p(\theta|y)p(y)p(y^*|\theta)$$

$$p(y^*|y) = \int d\theta\, p(\theta, y^*|y)$$
$$= \int d\theta\, \frac{p(y^*, y, \theta)}{p(y)}$$

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



Posterior Predictive for predictions

The distribution of a future data point y^*:

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

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

Expectation of the likelihood at a new point(s) over the posterior $E_{p(\theta|D)}[p(y^*|\theta)]$.

Prior Predictive for simulations

The distribution of a data point y from the prior:

$$p(y) = \int d\theta \, p(\theta, y) = \int d\theta \, p(y|\theta)p(\theta).$$

the expected likelihood over the prior $E_{p(\theta)}[\mathcal{L}]$

(like the evidence, but not just at the data)

Summary via MAP (a point estimate)

$$\begin{aligned}
\theta_{\mathrm{MAP}} &= \arg\max_{\theta} \; p(\theta|D) \\
&= \arg\max_{\theta} \; \frac{\mathcal{L}\, p(\theta)}{p(D)} \\
&= \arg\max_{\theta} \; \mathcal{L}\, p(\theta)
\end{aligned}$$

Bayesian Bestiary

- Prior
- posterior
- evidence
- prior predictive
- posterior predictive
- MAP (or other point) estimate

Conjugate Prior

- A **conjugate prior** is one which, when multiplied with an appropriate likelihood, gives a posterior with the same functional form as the prior.
- Likelihoods in the exponential family have conjugate priors in the same family
- analytical tractability AND interpretability

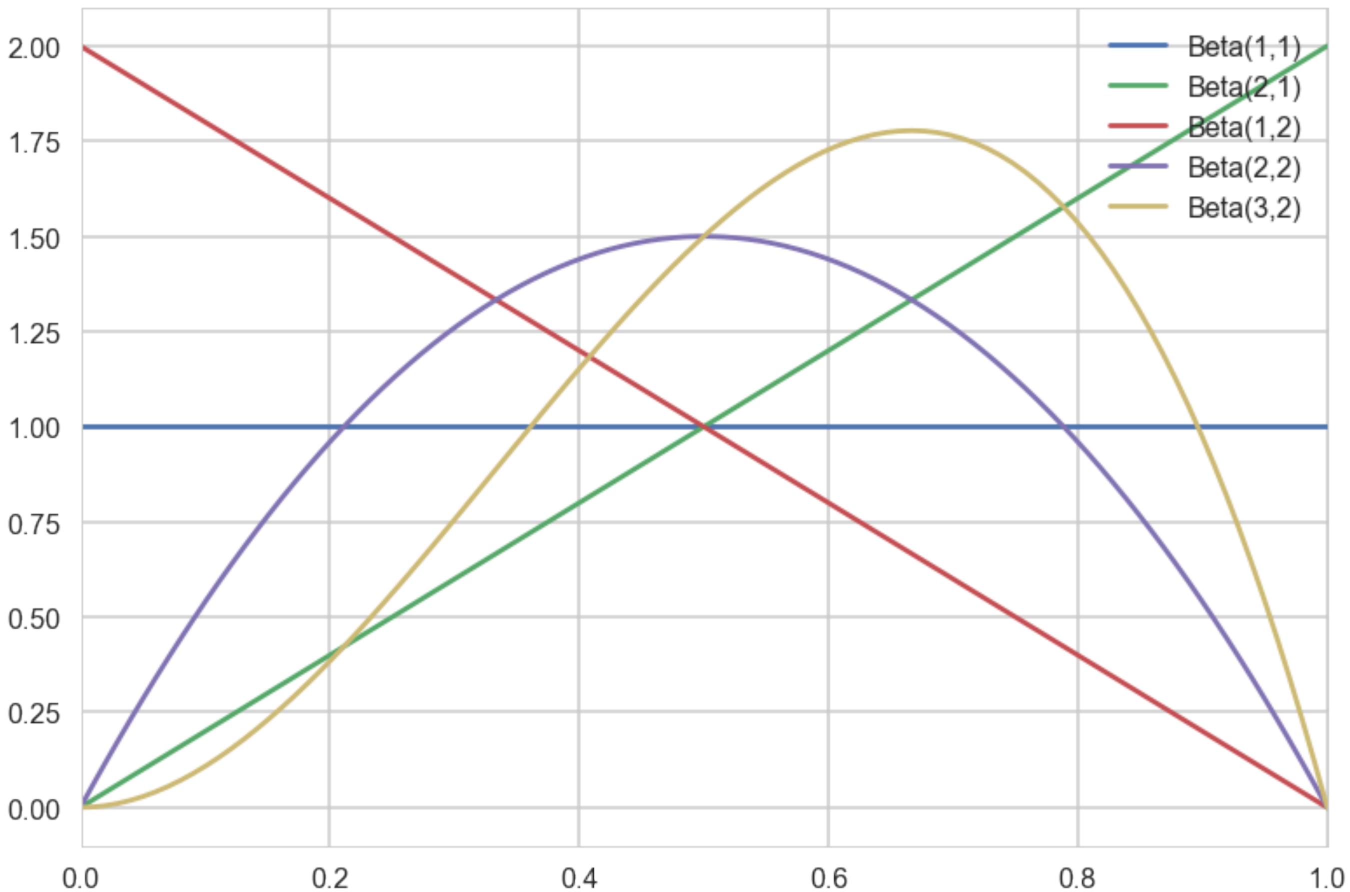
Coin Toss Model

- Coin tosses are modeled using the Binomial Distribution, which is the distribution of a set of Bernoulli random variables.
- The Beta distribution is conjugate to the Binomial distribution

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

Because of the conjugacy, this turns out to be:

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



BETA DISTRIBUTION

$$Beta(\alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$$

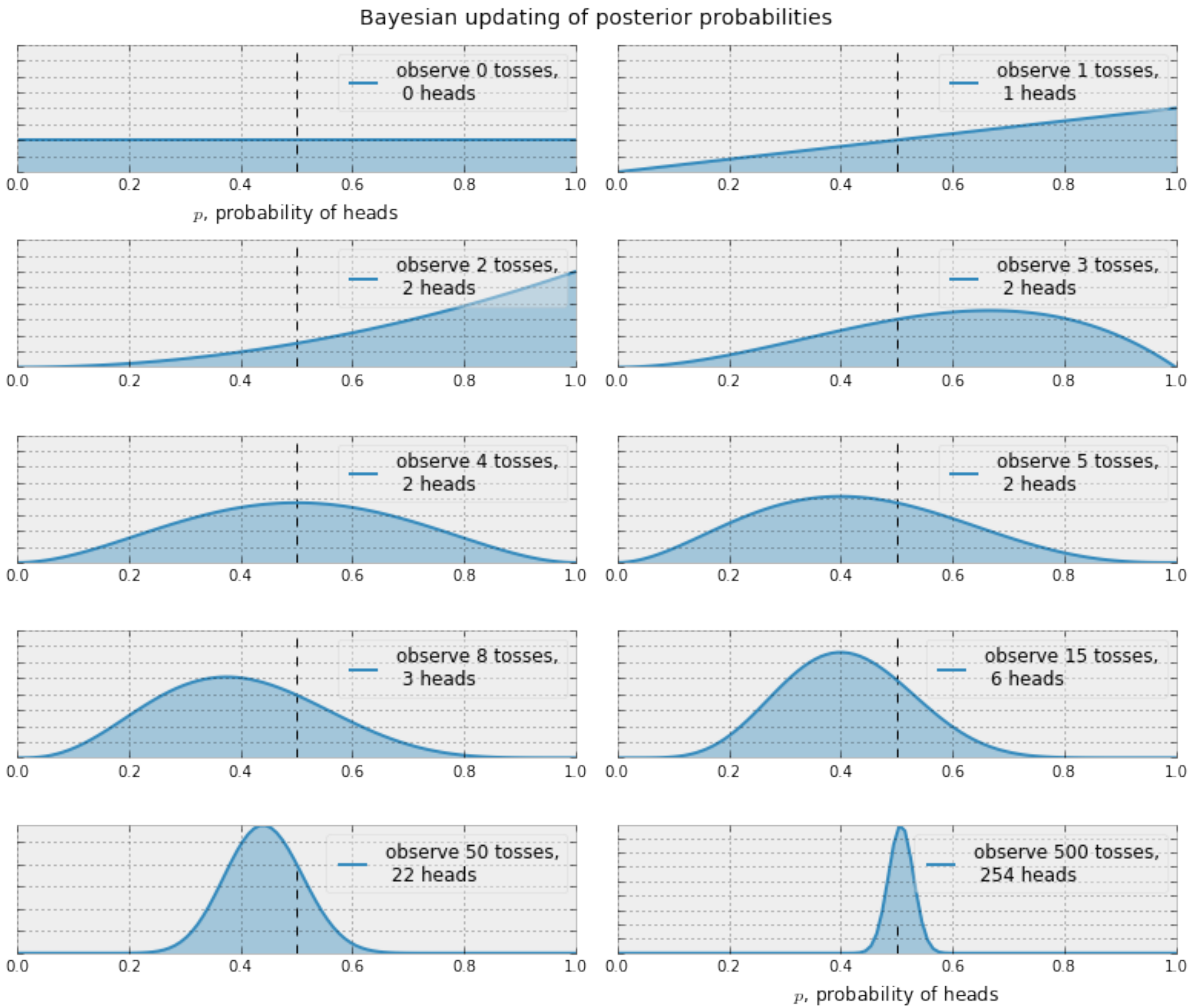
where

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1} dt$$

Prior heads: α, prior tails: β, so heads fraction is $\alpha/(\alpha+\beta)$.

Priors Regularize

- think of a prior as a regularizer.
- a $Beta(1,1)$ prior is equivalent to a uniform distribution.
- This is an **uninformative prior**. Here the prior adds one heads and one tails to the actual data, providing some "towards-center" regularization
- especially useful where in a few tosses you got all heads, clearly at odds with your beliefs.
- a $Beta(2,1)$ prior would bias you to more heads


Bayesian updating of posterior probabilities
observe 0 tosses, 0 heads
observe 1 tosses, 1 heads
observe 2 tosses, 2 heads
observe 3 tosses, 2 heads
observe 4 tosses, 2 heads
observe 5 tosses, 2 heads
observe 8 tosses, 3 heads
observe 15 tosses, 6 heads
observe 50 tosses, 22 heads
observe 500 tosses, 254 heads
p, probability of heads
0.0 0.2 0.4 0.6 0.8 1.0

Bayesian Updating "on-line"

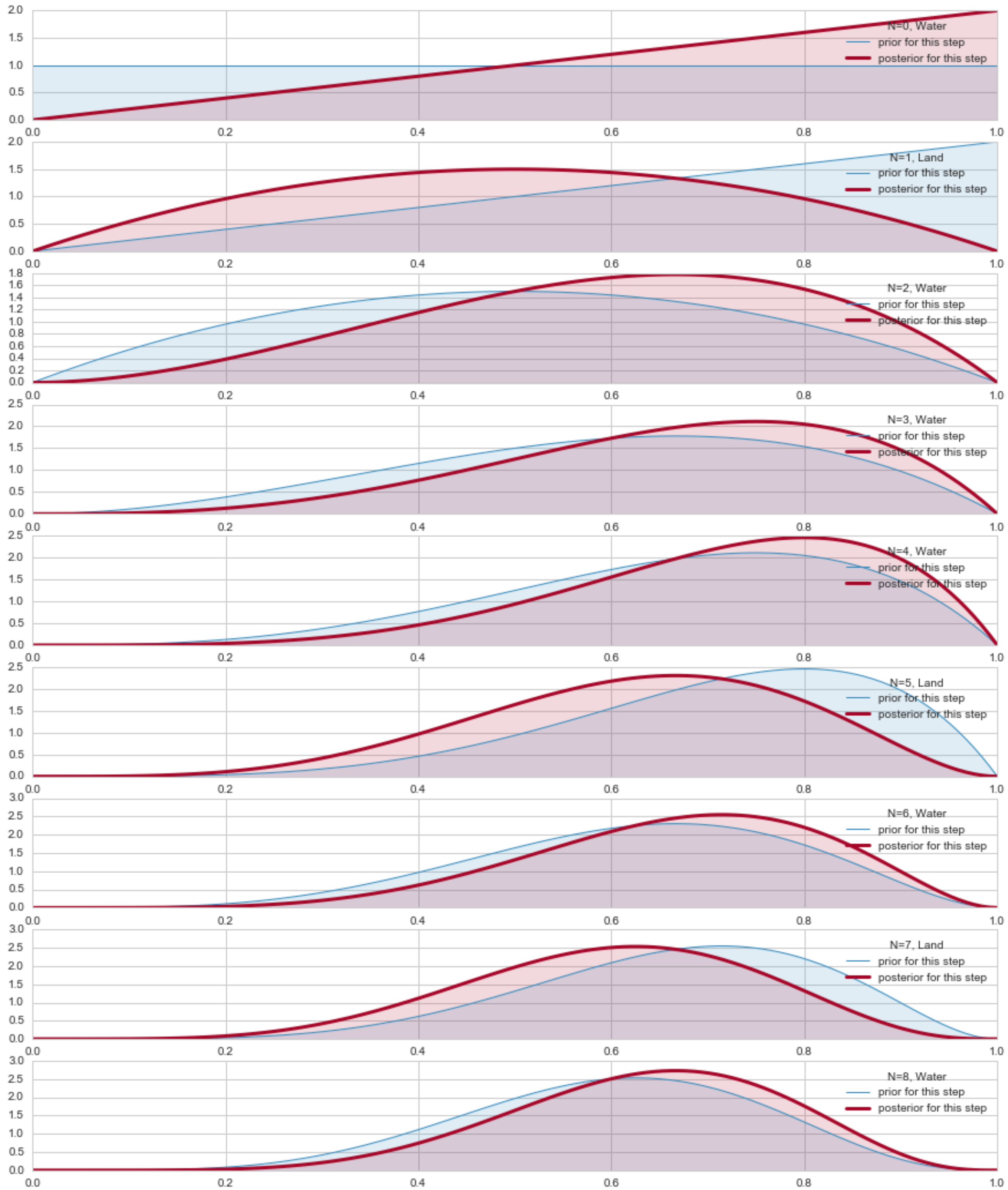
- can update prior to posterior all at once, or one by one
- as each piece of data comes in, you update the prior by multiplying by the one-point likelihood.
- the posterior you get becomes the prior for our next step

$$p(\theta \mid \{y_1, \ldots, y_{n+1}\}) \propto p(\{y_{n+1}\} \mid \theta) \times p(\theta \mid \{y_1, \ldots, y_n\})$$

- the posterior predictive is the distribution of the next data point!

$$p(y_{n+1} \vert \{y_1, \ldots y_n\}) = E_{p(\theta \vert \{y_1, \ldots y_n\})}[p(y_{n+1} \vert \theta)] = \int d\theta \, p(y_{n+1} \vert \theta) p(\theta \vert \{y_1, \ldots y_n\})$$

.



Bayesian Updating of globe

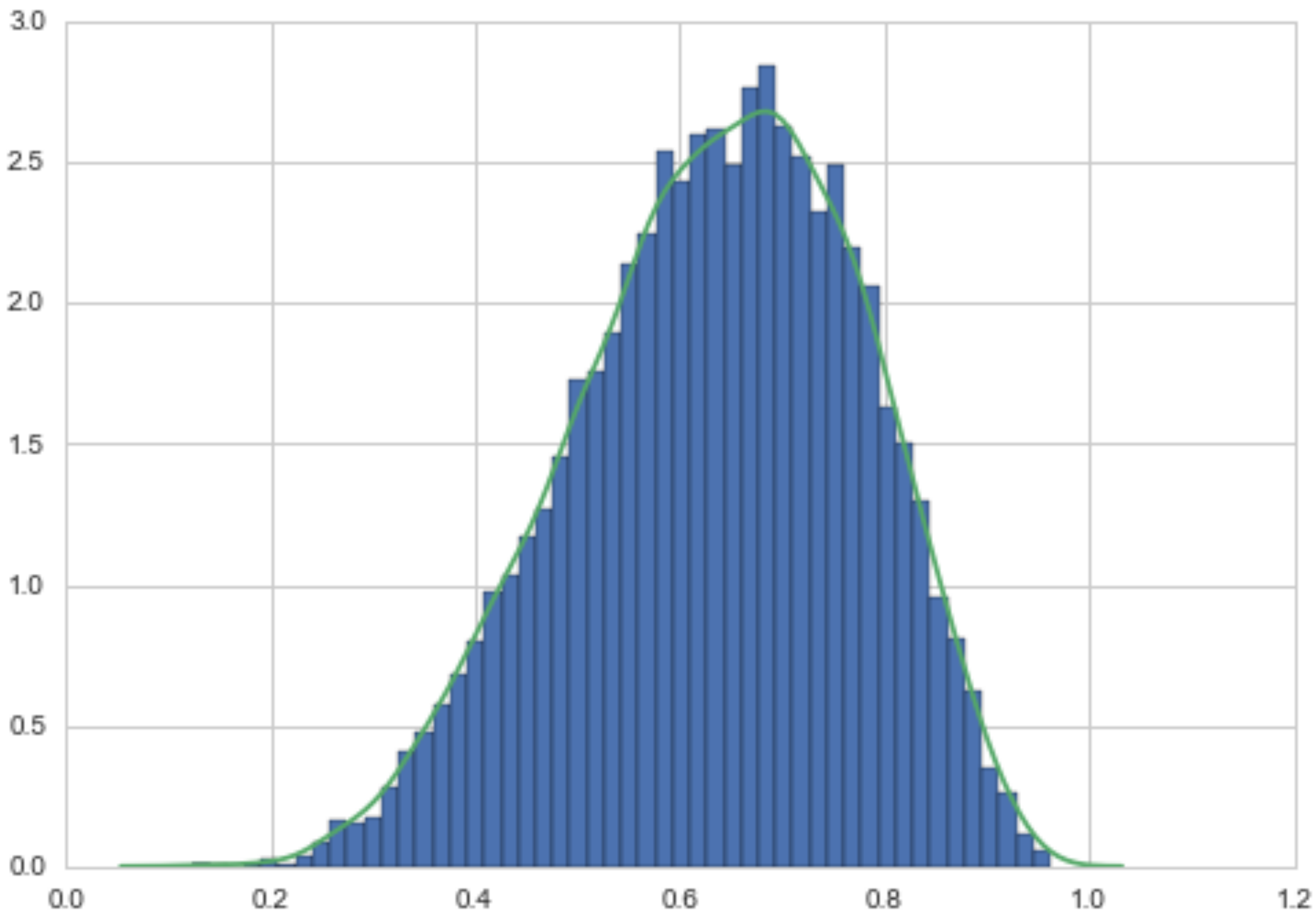
- Seal tosses globe, θ is true water fraction
- data `WLWWWLWLW`
- notice how the posterior shifts left and right depending on new data

At each step:

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

Samples, Samples, Samples

- for globe toss, simple use `scipy.stats` to sample from appropriate beta distribution. We then have our posterior
- what about the predictive distributions? They are Beta-Binomial distributions. Complicated.
- Sampling gives us an easier way!



Posterior properties

- 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, 0.64731430523031 43)`

MAP, a point estimate

$$
\begin{aligned}
\theta_{\text{MAP}} &= \arg\max_{\theta} \, p(\theta \vert D) \\
&= \arg\max_{\theta} \, \frac{\mathcal{L} \, p(\theta)}{p(D)} \\
&= \arg\max_{\theta} \, \mathcal{L} \, p(\theta)
\end{aligned}
$$

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

```
31 0.662578641304
```

OR Optimize!

Posterior Mean minimizes squared loss

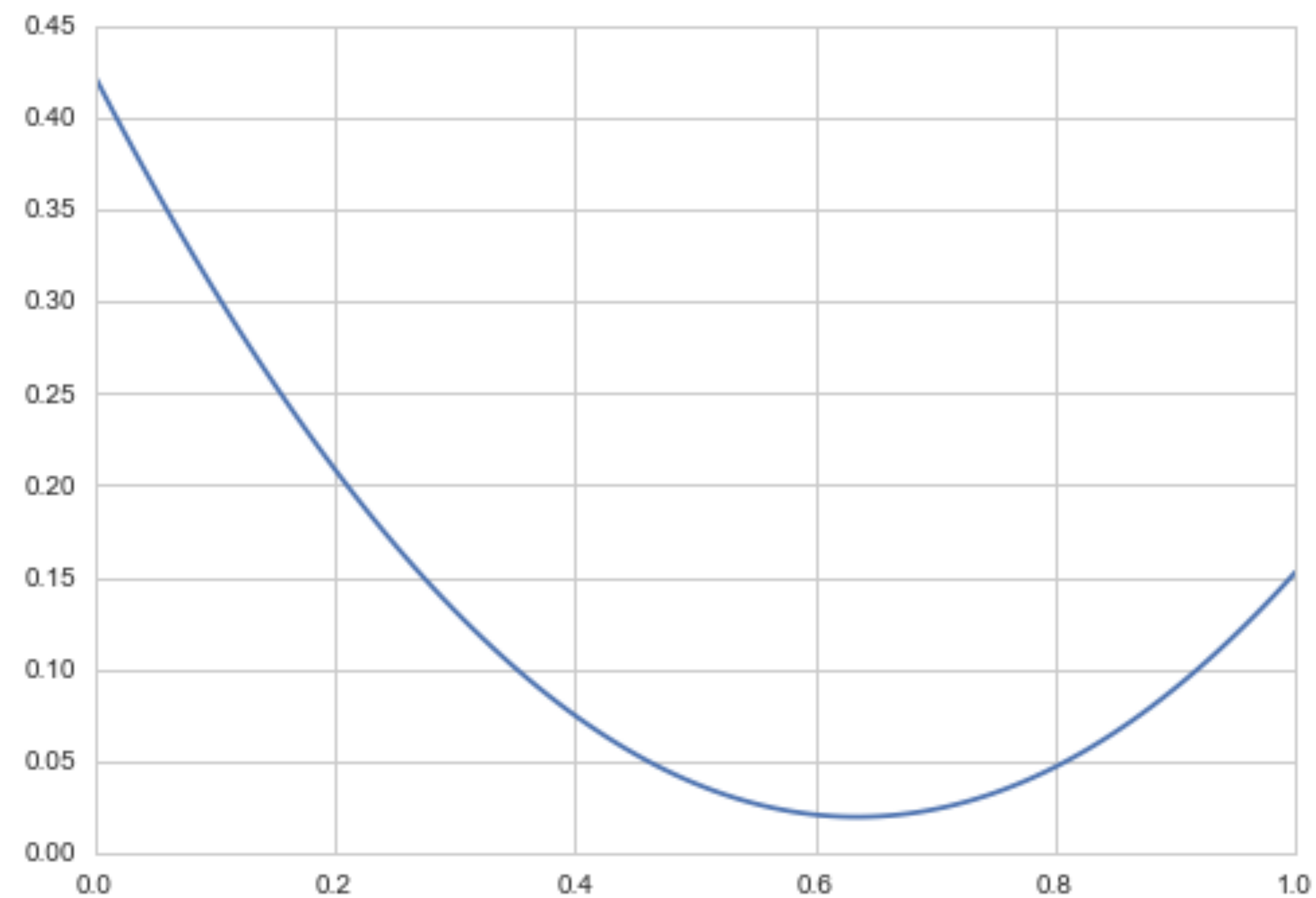
$$R(t) = E_{p(\theta|D)}[(\theta - t)^2] = \int d\theta (\theta - t)^2 p(\theta|D)$$

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

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

Mean is at 0.638.

This is **Decision Theory**.



Posterior predictive

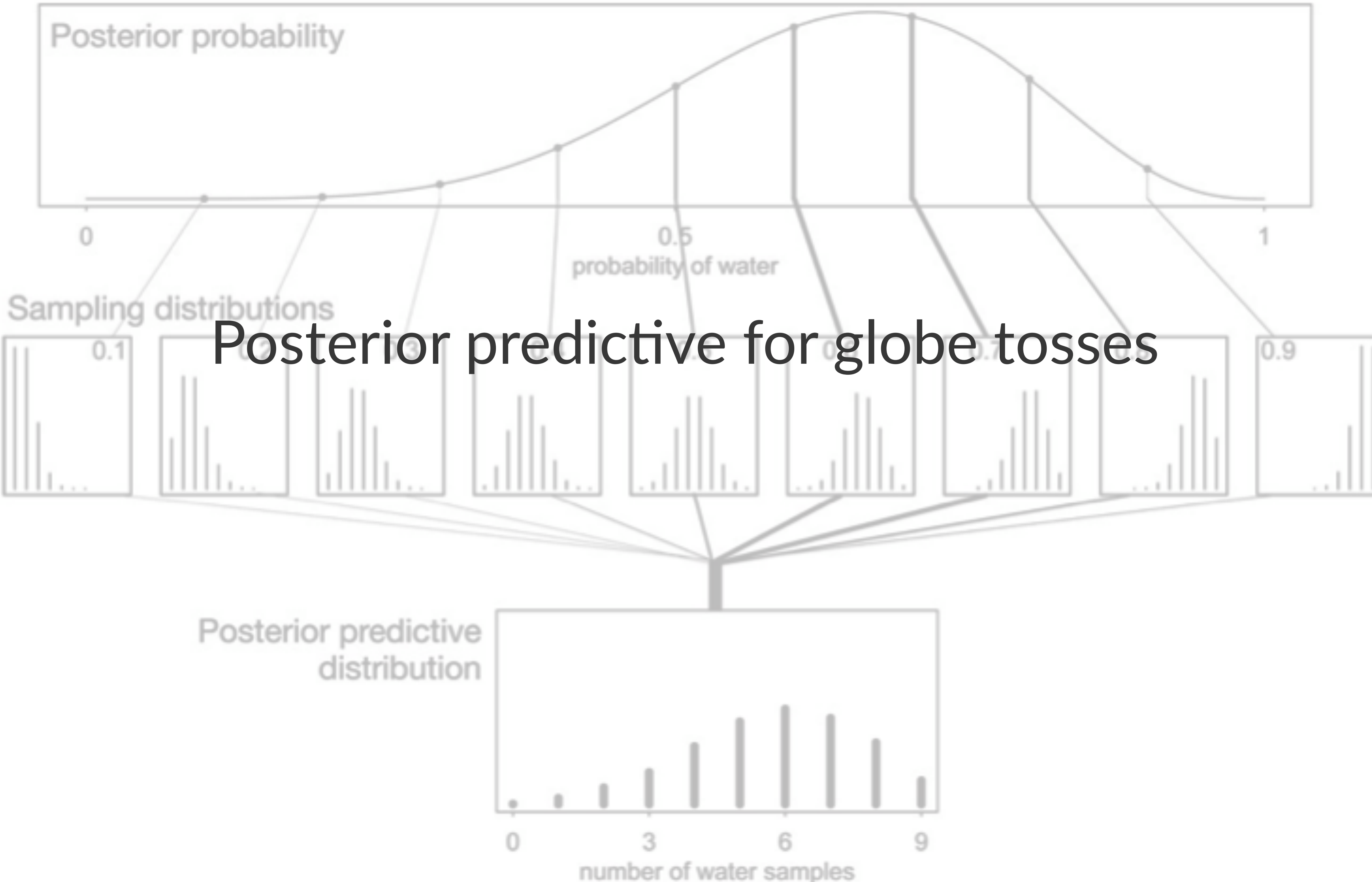
$$p(y^* | D) = \int d\theta p(y^* | \theta) p(\theta | D)$$

Its a Beta-Binomial distribution.

Risk Minimization holds here too:

$$y_{minmse} = \int dy \, y \, p(y|D)$$

Posterior predictive for globe tosses



Plug-in Approximations

θ_{MAP} is a point estimate.

Consider $p(\theta|D) = \delta(\theta - \theta_{MAP})$ and then draw

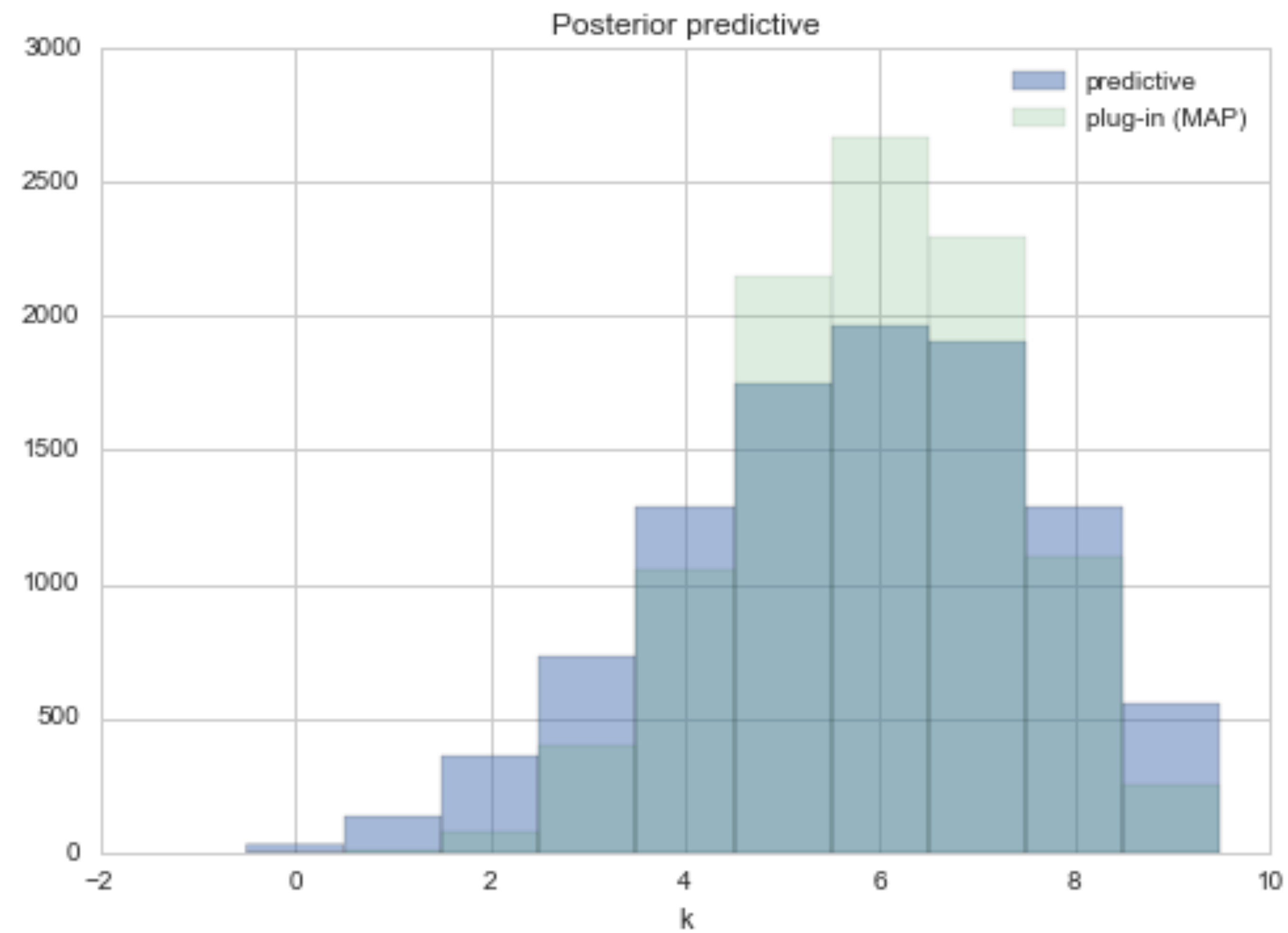
$$p(y^*|D) = p(y^*|\theta_{MAP})$$ a sampling distribution.

Underestimates spread.

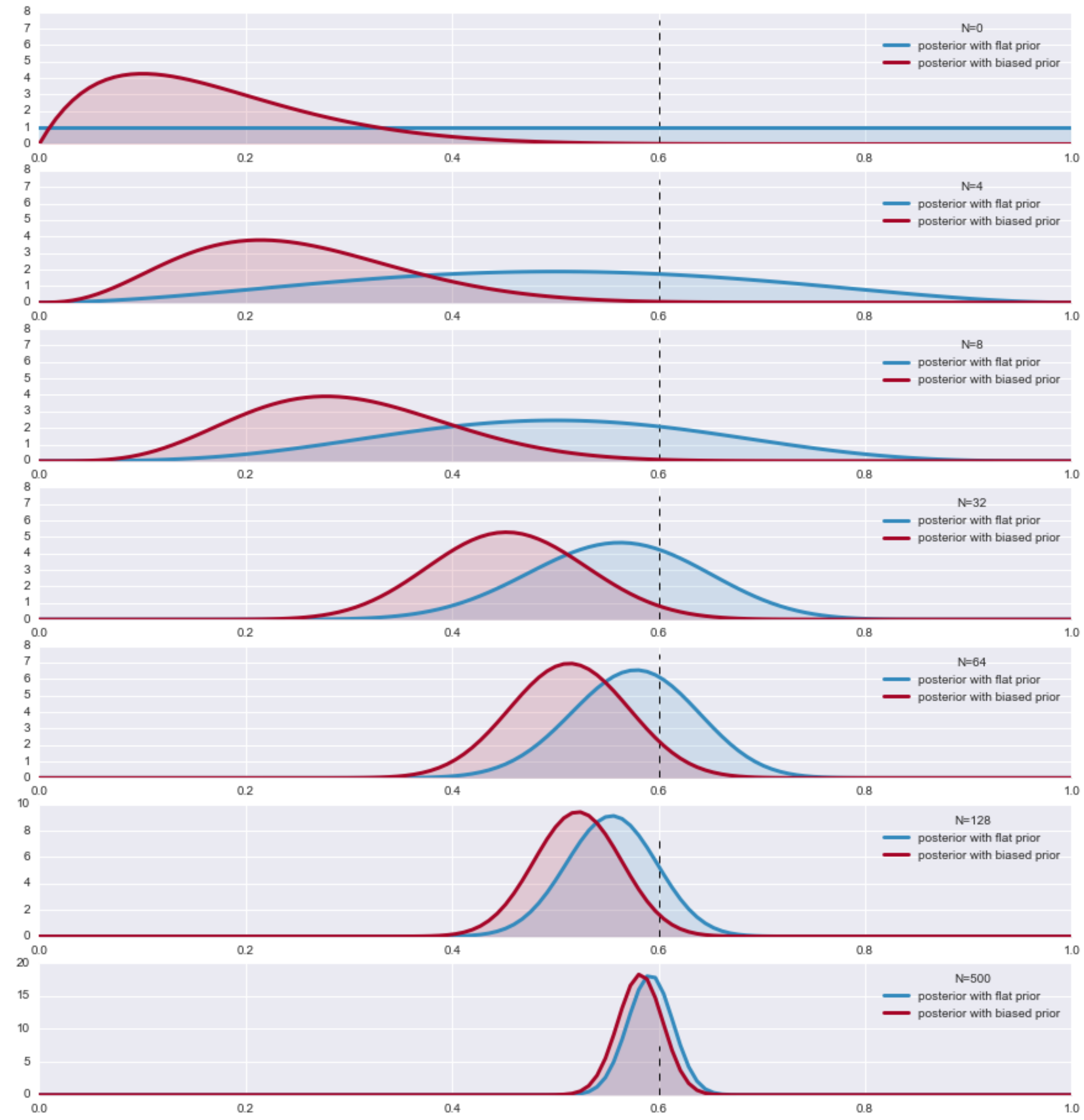
Posterior predictive from sampling

- draw the thetas from posterior
- then draw y's from the sampling distribution
- and histogram it
- these are draws from joint y, θ

```
postpred = np.random.binomial(n,samples)
```



Data overwhelms prior eventually



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.

This is really a statement about what is IID and what is not.

It depends on how much knowledge you have...

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

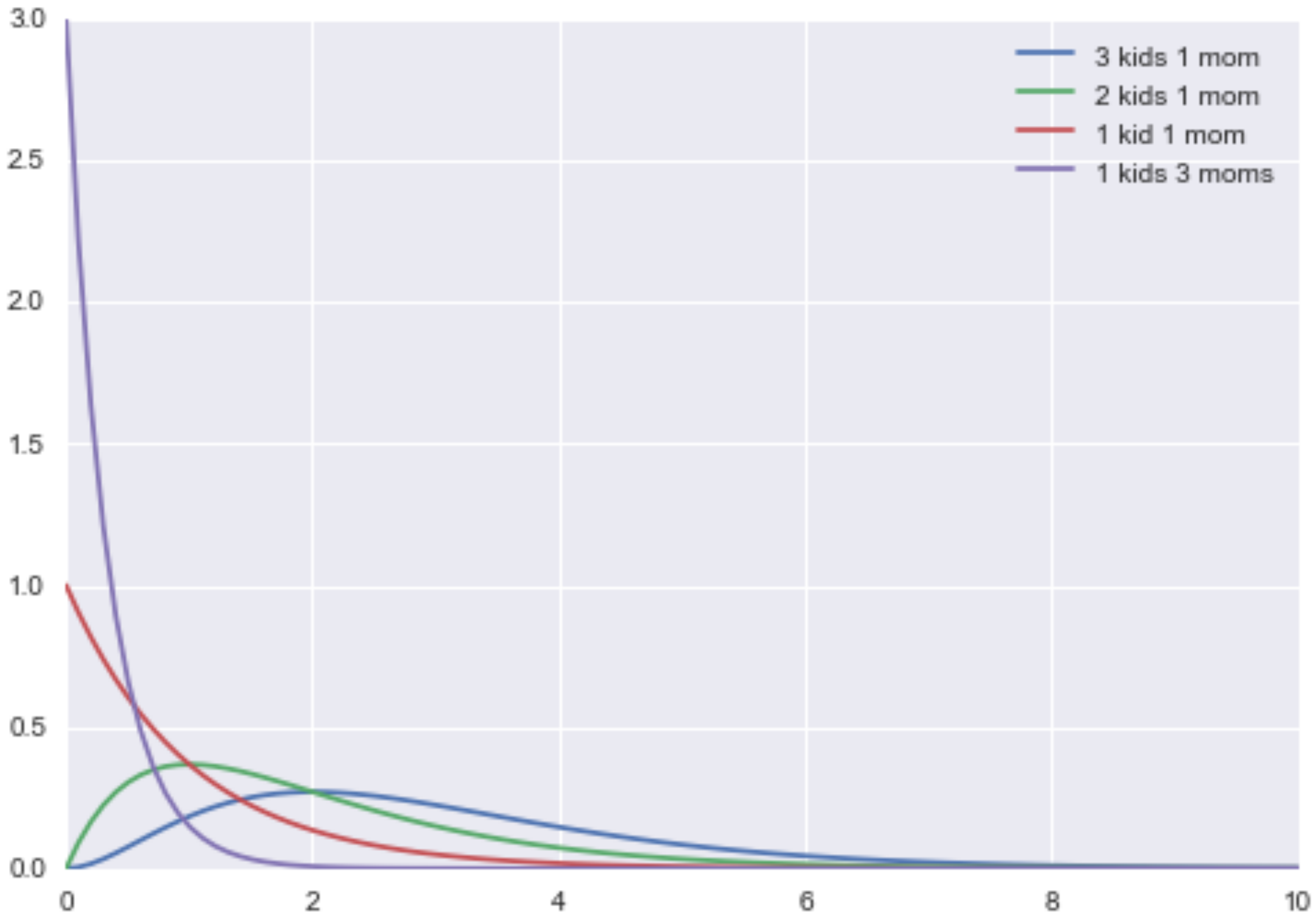
Form is of $Gamma$. In shape-rate parametrization (wikipedia)

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

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.

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

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

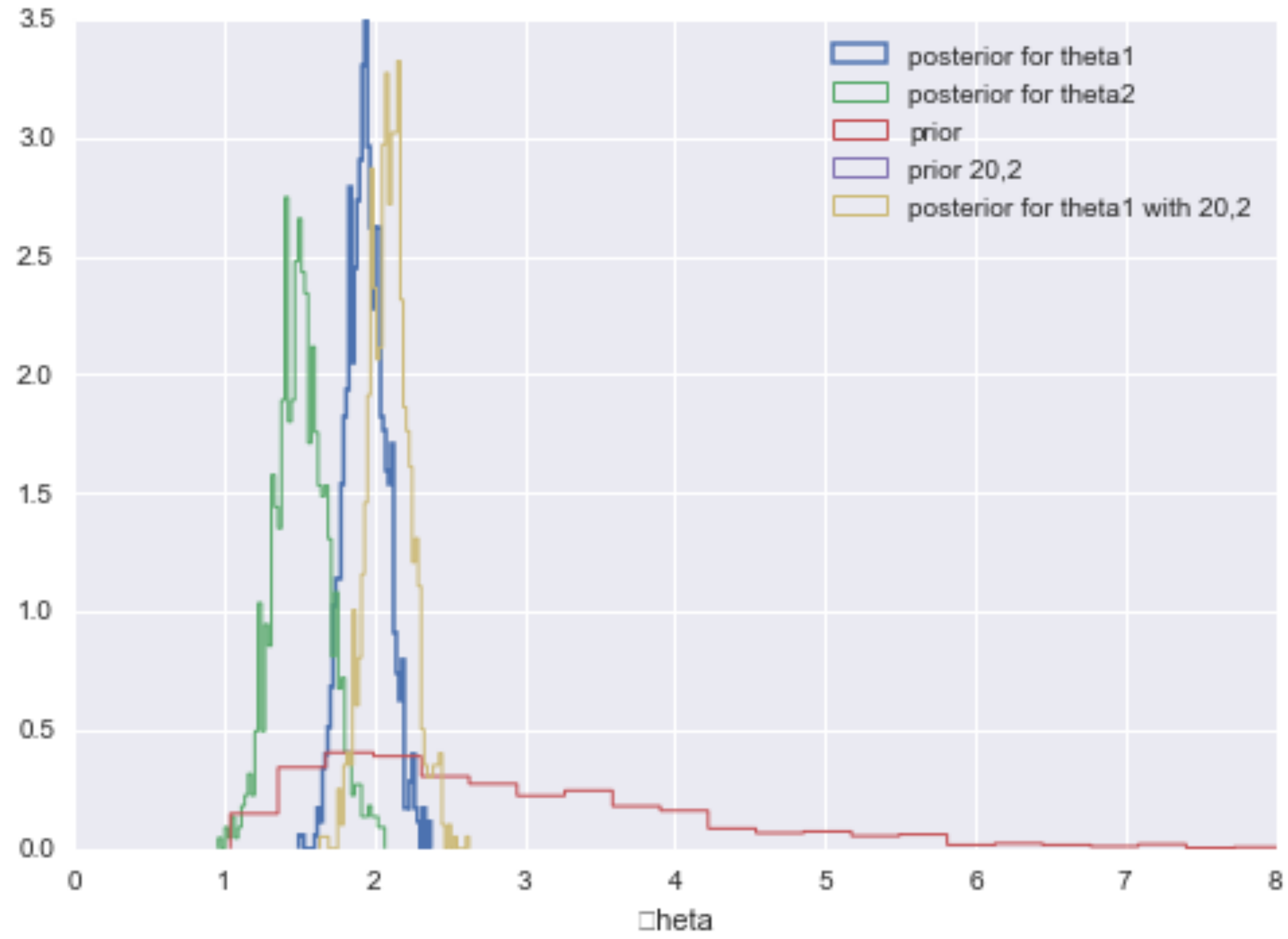
Prior mean, variance:
$E[\theta] = a/b, var[\theta] = a/b^2.$

Posteriors

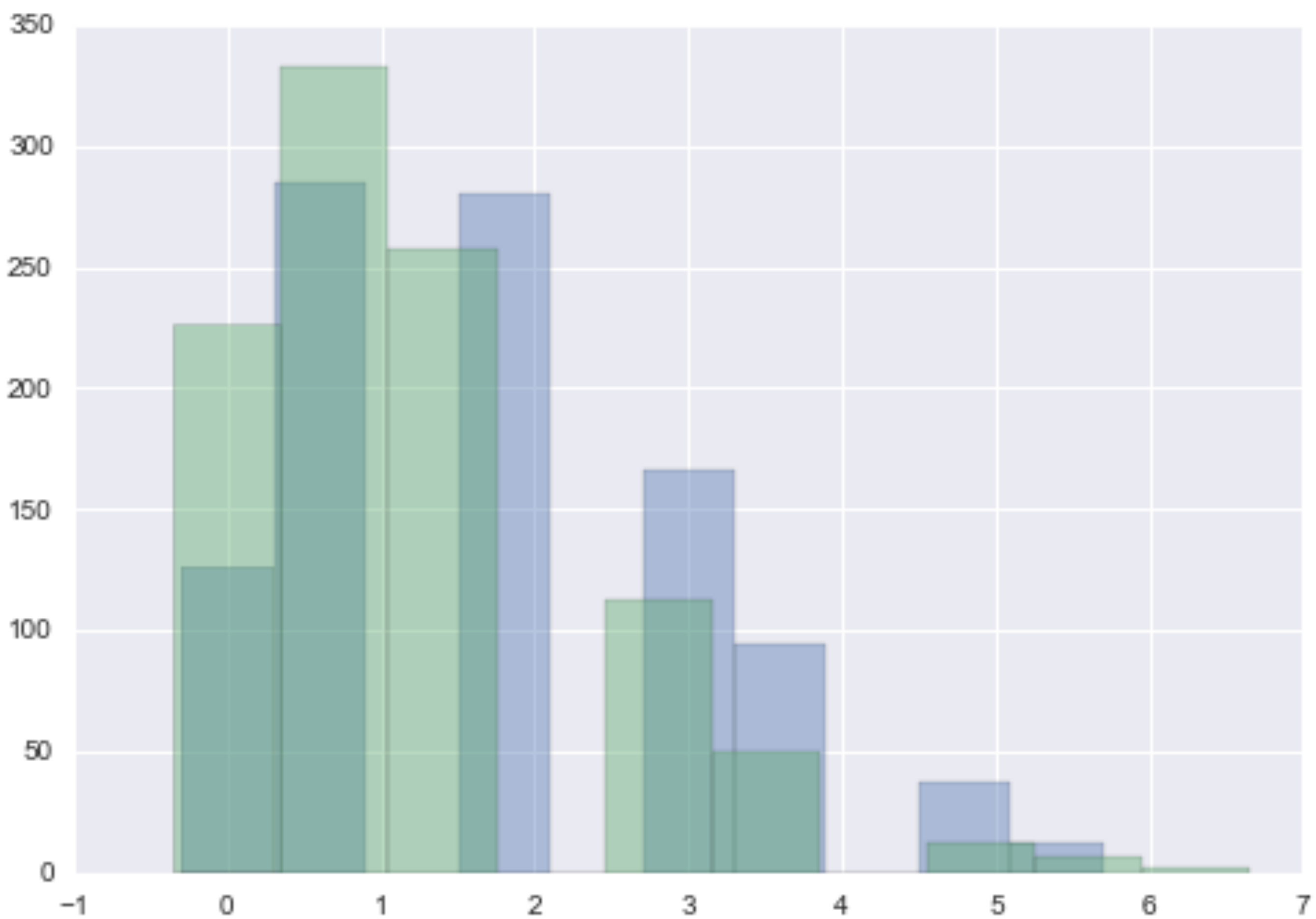
$$E[\theta] = (a + \sum y_i)/(b+N)$$
$$var[\theta] = (a + \sum y_i)/(b+N)^2.$$

```
np.mean(theta1),
np.var(theta1) =
(1.951688152179l478,
0.01852720418578578S)
```

```
np.mean(theta2),
np.var(theta2) =
(1.503725210021360 9,
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