Computer Science

# CSC380: Principles of Data Science

## Bayesian Statistics and Inference

Prof. Jason Pacheco

TA: Enfa Rose George

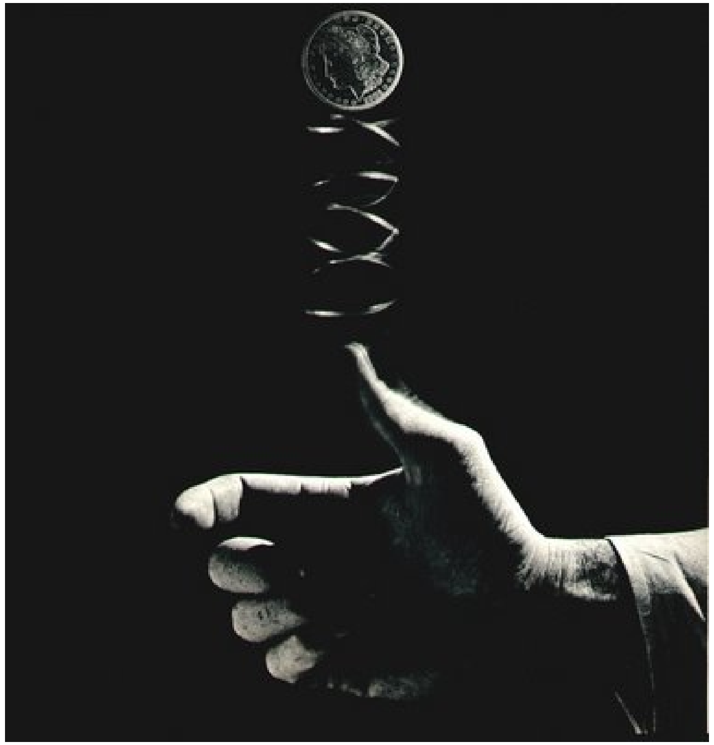
TA: Saiful Islam Salim

# Administrative Items

- Graders have been lenient on some items, but will be more strict
  - Attach code separately so TAs can run it
  - Answer questions clearly and in the order presented
  - Show your work (not just answers)
- It is not the TA’s responsibility to debug your code
- There are 3 official office hours

# What is Probability?

*What does it mean that the probability of heads is ½ ?*



*Two schools of thought…*

**Frequentist Perspective**
Proportion of successes (heads) in repeated trials (coin tosses)

**Bayesian Perspective**
Belief of outcomes based on assumptions about nature and the physics of coin flips

*Neither is better/worse, but we can compare interpretations…*

# Frequentist & Bayesian Modeling

*We will use the following notation throughout:*

$\theta$ - Unknown (e.g. coin bias)

$y$ - Data

**Frequentist**

(Conditional Model)

$$p(y;\theta)$$

- $\theta$ is a non-random unknown parameter
- $p(y;\theta)$ is the *sampling / data generating distribution*

**Bayesian**

(Generative Model)

$$\text{Prior Belief} \rightarrow p(\theta)p(y \mid \theta) \leftarrow \text{Likelihood}$$

- $\theta$ is a random variable (latent)
- Requires specifying $p(\theta)$ the prior belief

# Bayes' Rule

*Posterior distribution is complete representation of uncertainty*

$$p(\theta \mid y) = \frac{\overbrace{p(\theta)}^{\text{Prior Belief}}\overbrace{p(y \mid \theta)}^{\text{Likelihood}}}{\underbrace{p(y)}_{\text{Marginal Likelihood}}}$$

Just the definition of conditional probability

Contrary to the likelihood principle (likelihood contains all necessary information about a parameter)

If we don't care about $\theta$ we can just integrate (marginalize) it out,

$$p(y) = \int p(\theta)p(y \mid \theta)\, d\theta = \mathbf{E}_{p(\theta)}[p(y \mid \theta)]$$

# Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.


A recent home test states that you have high BP. Should you start medication?

An Assessment of the Accuracy of Home Blood Pressure Monitors When Used in Device Owners

Jennifer S. Ringrose,[1] Gina Polley,[1] Donna McLean,[2–4] Ann Thompson,[1,5] Fraulein Morales,[1] and Raj Padwal[1,4,6]

# Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.


- Latent quantity of interest is hypertension: $\theta \in \{true, false\}$
- Measurement of hypertension: $y \in \{true, false\}$
- Prior: $p(\theta = true) = 0.29$
- Likelihood: $p(y = true \mid \theta = false) = 0.30$

$$p(y = true \mid \theta = true) = 1.00$$

# Bayesian Inference Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.


Suppose we get a positive measurement, then posterior is:

$$p(\theta = true \mid y = true) = \frac{p(\theta = true)p(y = true \mid \theta = true)}{p(y = true)}$$

$$= \frac{0.29 * 1.00}{0.29 * 1.00 + 0.71 * 0.30} \approx 0.58$$

**What conclusions can be drawn from this calculation?**

# Bayesian Updating

What if we receive another test $y_2 = true$ ?

**Question** What is our belief about $\theta$ *before* seeing the second test?

$$p(\theta = true \mid y_1 = true) \approx 0.58$$

**Question** What is the likelihood that $y_2 = true$? Does it depend on $y_1$?

$$p(y_2 = true \mid \theta = true) = 1.0$$

Posterior over both tests is then:

$$p(\theta = true \mid y_1 = true, y_2 = true) \propto \underbrace{p(\theta = true \mid y_1 = true)}_{\text{Inference from first test}} p(y_2 = true \mid \theta)$$

Proportional to

# Bayesian Updating

Consider two *conditionally independent* observations $X_1$ and $X_2$, their joint distribution is:

Probability chain rule

$$p(\theta, X_1, X_2) = p(\theta)p(X_1 \mid \theta)p(X_2 \mid \theta) = p(\theta \mid X_1)p(X_1)p(X_2 \mid \theta)$$

So, conditioned on $X_1$:

Update prior belief after seeing $X_1$

$$p(\theta, X_2 \mid X_1) = p(\theta \mid X_1)p(X_2 \mid \theta)$$

This is proportional to the **full posterior** by Bayes' rule:

$$p(\theta \mid X_1, X_2) \propto p(\theta \mid X_1)p(X_2 \mid \theta)$$

Normalizer is $p(X_1,X_2)$

In general, given conditionally independent $X_1, \ldots, X_N$ :

$$p(\theta \mid X_1, \ldots, X_N) \propto p(\theta \mid X_1, \ldots, X_{N-1})p(X_N \mid \theta)$$

# Frequentist vs. Bayesian Inference

We have data $X_1, \ldots, X_N$ and want to infer unknown parameter $\theta$

**Frequentist Inference**

The data *uniquely determines* $\theta$*, e.g.* by the likelihood:

$$p(X_1, \ldots, X_N; \theta)$$

**How well it explains the data**

**Bayesian Inference**

The data *updates our belief* about $\theta$, which is random:

$$p(\theta \mid X_1, \ldots, X_N) \propto p(\theta \mid X_1, \ldots, X_{N-1}) p(X_N \mid \theta)$$

**Our belief changes with more data**

# Minimum Mean Squared Error (MMSE)

Posterior mean minimizes squared error,

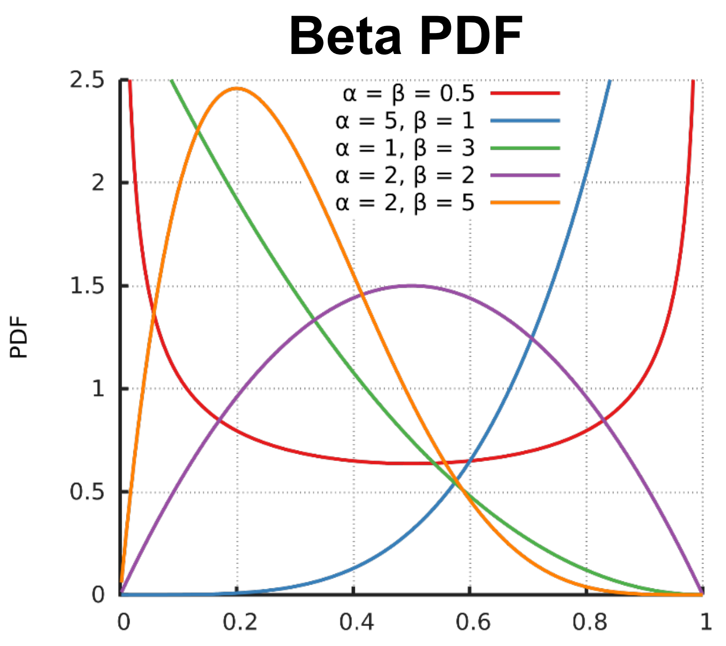
$$\hat{\theta}^{\mathrm{MMSE}} = \arg\min \mathbb{E}[(\hat{\theta} - \theta)^2 \mid y] = E[\theta \mid y]$$

- Minimizes error <u>conditioned on observed data</u>
- MMSE is an **unbiased estimator**
- MMSE is **asymptotically unbiased** and **asymptotically normal**,

$$\sqrt{N}(\hat{\theta}^{\mathrm{MMSE}} - \theta) \to \mathcal{N}(0, \sigma^2)$$

# Example: Beta-Bernoulli MMSE

Let $Y_1, \ldots, Y_N \sim \mathrm{Bernoulli}(\pi)$ and $\pi \sim \mathrm{Beta}(\alpha, \beta)$.




- Beta is a distribution on probabilities $\pi \in [0, 1]$
- *Shape parameters* $\alpha$ and $\beta$ with mean,

$$\mathbf{E}[\theta] = \frac{\alpha}{\alpha+\beta}$$

- Beta-Bernoulli has Beta posterior distribution,

$$p(\pi \mid X_1^N) = \mathrm{Beta}(\alpha + \textbf{number of heads},\ \beta + \textbf{number of tails})$$

MMSE given by posterior mean,

Prior belief (pseudo-heads)

$$\hat{\pi}^{\mathrm{MMSE}} = \frac{\alpha + \textbf{number of heads}}{\alpha + \beta + N}$$

**Q** What happens to MMSE when we have limited data?

**Q** What happens to MMSE when we have a lot of data?

# Bayes Estimators

Minimizes expected loss function,

$$\hat{\theta} = \arg\min_{\hat{\theta}} \mathbf{E}\left[L(\theta, \hat{\theta}) \mid y\right]$$

Expected loss referred to as *Bayes risk.*

**MMSE** minimizes squared-error loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$

**Minimum absolute error (MAE)** is posterior *median,*

$$\arg\min \mathbb{E}[|\hat{\theta} - \theta| \mid y] = \text{median}(\theta \mid y)$$

Note: Same answer for linear function: $L(\theta, \hat{\theta}) = c|\hat{\theta} - \theta|$

# Administrative Items

- HW3 Due tonight @ 11:59pm

- HW4 Out tomorrow (Due Thursday, 9/30)
  - Only 2 questions this time
  - MLE and MAP estimation

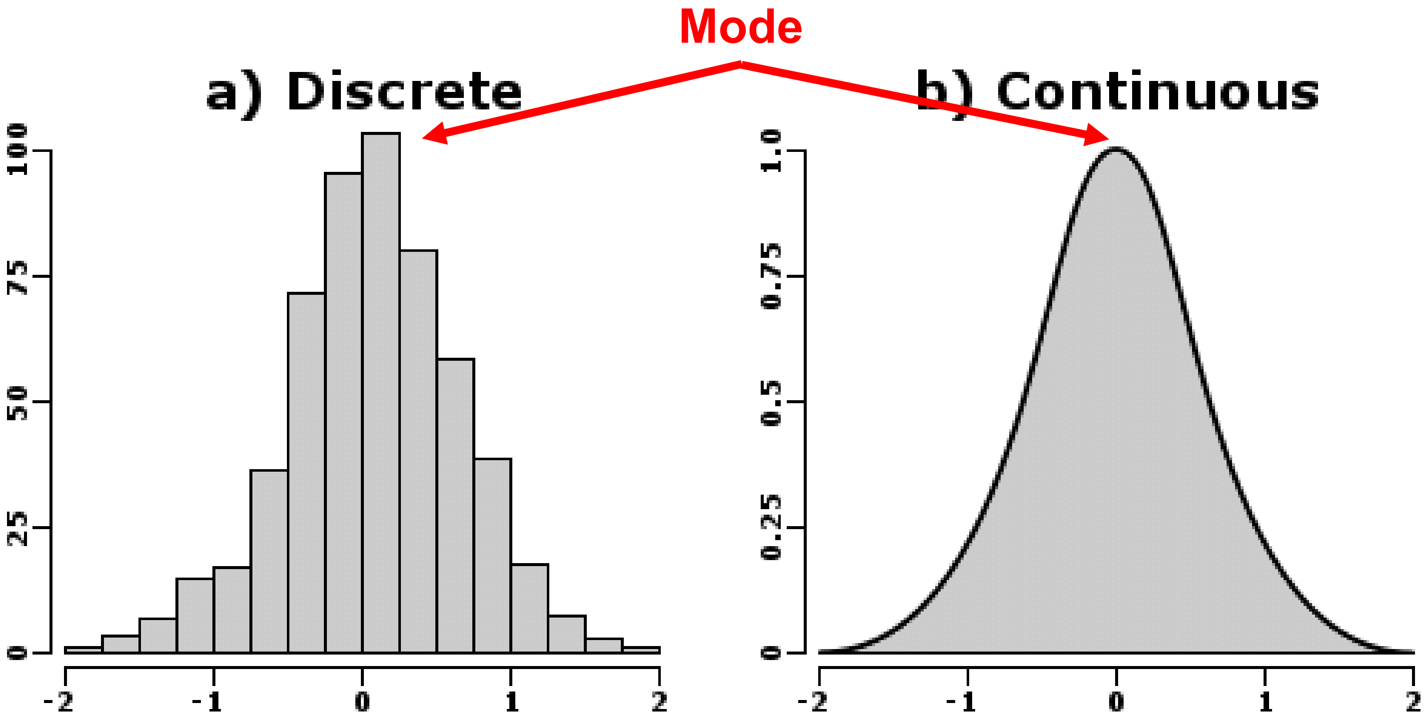
- Survey for early student feedback (see Piazza)
  - Please complete by next Monday
  - Should only take a couple minutes

# Maximum a Posteriori (MAP)

Very common to produce maximum probability estimates,

$$\hat{\theta}^{\mathrm{MAP}} = \arg max\, p(\theta \mid y)$$

*MAP is the* ***mode*** *( highest probability outcome ) of the posterior*



# Maximum a Posteriori (MAP)

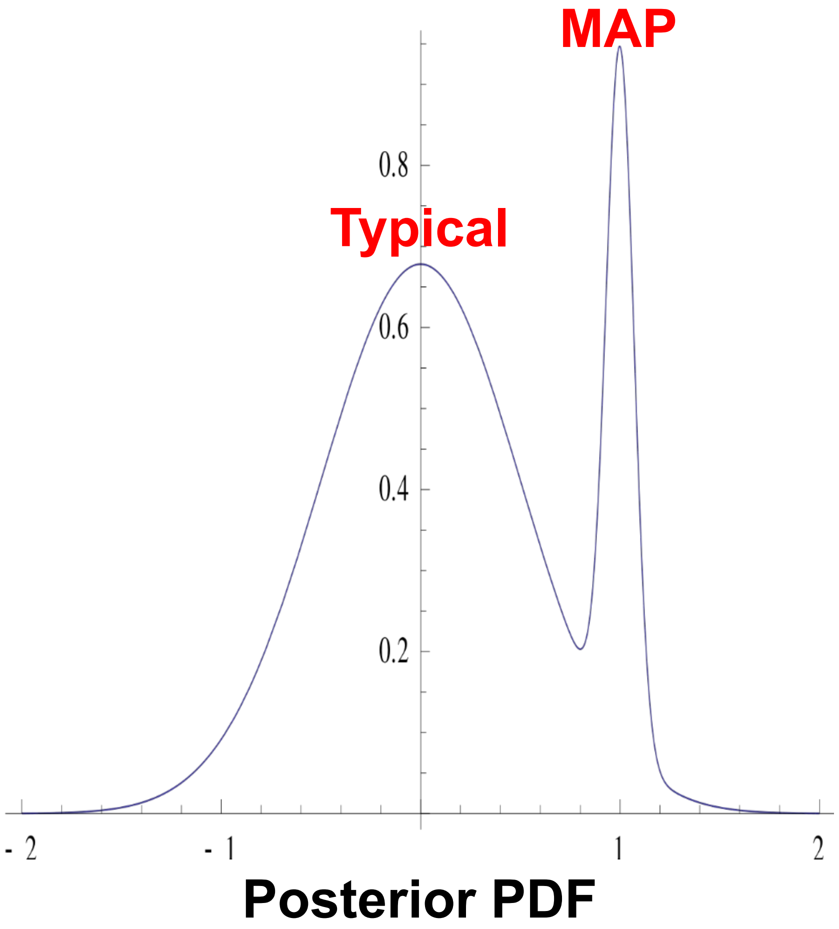
*MAP (mode) may not be representative of typical outcomes*

Also, not a Bayes estimator (unless discrete),

$$\lim_{c\to 0} L(\theta,\hat{\theta}) = \begin{cases} 0, & \text{if } |\hat{\theta}-\theta| < c \\ 1, & \text{otherwise} \end{cases}$$

**Degenerate loss function**

Despite its issues, MAP is frequently used in "Bayesian" inference and estimation



# Example: Beta-Bernoulli MAP

Let $Y_1, \ldots, Y_N \sim \text{Bernoulli}(\pi)$ and $\pi \sim \text{Beta}(\alpha, \beta)$ then posterior is,

$$p(\pi \mid X_1^N) = \text{Beta}(\alpha + \underbrace{\textbf{number of heads}}_{\mathbf{N_H}},\ \beta + \textbf{number of tails})$$

Highest probability (mode) of Beta given by,

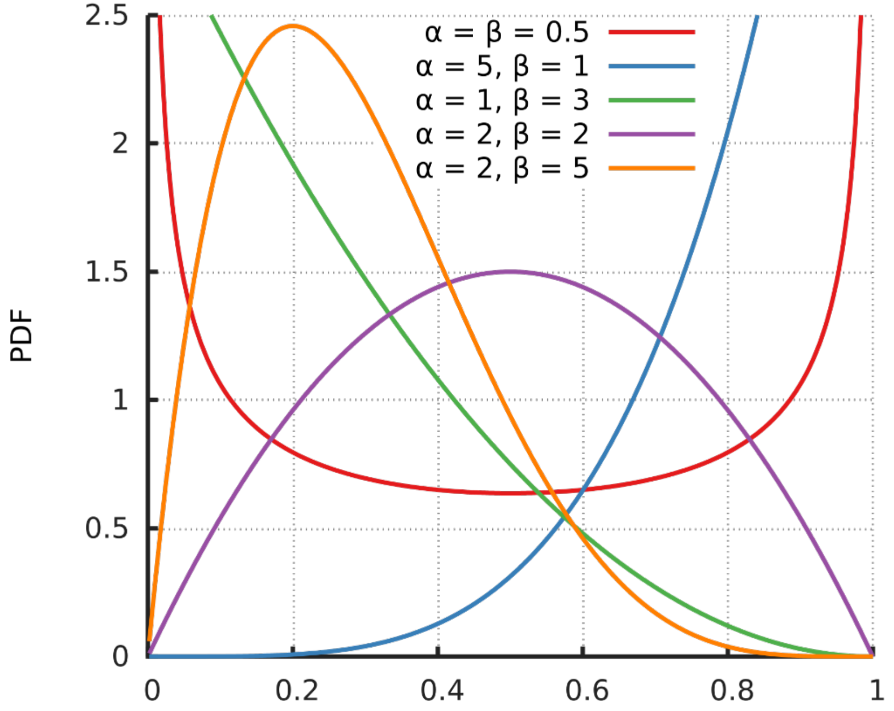
**Take derivative, set to zero, solve.**

$$\hat{\pi}^{\text{MAP}} = \frac{\alpha + N_H - 1}{\alpha + \beta + N - 2}$$

Beta distribution is not always convex!

- MAP is any value for $\alpha = \beta = 1$
- Two modes (bimodal) for $\alpha, \beta < 1$

**Beta Posterior PDF**



# Maximum a Posteriori (MAP)

Equivalent to maximizing joint probability,

$$\arg\max_{\theta} p(\theta \mid y) = \arg\max_{\theta} \frac{p(\theta, y)}{\underbrace{p(y)}_{\text{Constant}}} = \arg\max_{\theta} p(\theta, y)$$

For iid $y_1, \ldots, y_N$ solve in log-domain (like *maximum likelihood est.*),

$$\hat{\theta}^{\text{MAP}} = \arg\max_{\theta} \log p(\theta, y_1, \ldots, y_N) = \underbrace{\sum_i \log p(y_i \mid \theta)}_{\substack{\text{Log-Likelihood} \\ \text{(how well it fits data)}}} + \underbrace{\log p(\theta)}_{\substack{\text{Log-Prior} \\ \text{(how well it} \\ \text{agrees with prior)}}}$$

***Intuition*** *MAP is like MLE but with a “penalty” term (log-prior)*

# Bayes' Rule : Reminder

**prior** probability

**likelihood** function
for the parameters

$$p(\theta \mid D) = \frac{p(\theta)p(D \mid \theta)}{p(D)}$$

**posterior** probability

normalizer,
often is not of
interest

# Bayes' Rule : Reminder

**prior** probability

**likelihood** function
for the parameters

$$p(\theta \mid D) \propto p(\theta) p(D \mid \theta)$$

**posterior** probability

Posterior is **proportional**
**to** the joint

# Bayes' Rule : Reminder

**prior** probability

**likelihood** function
for the parameters

$$q(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

**posterior** probability

In general, distributions are
different functions

# Bayes' Rule : Reminder

**Prior** and **likelihood** chosen for model

$$q(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

**Posterior** determined by algebra

In general, distributions are different functions

# Conjugate Pairs

*For some special models the posterior takes a simple form*

$$p(\theta \mid D) \propto p(\theta)\omega(D \mid \theta)$$

**Prior and posterior are the same distribution (with different parameters)**

We have already seen one example, the Beta-Bernoulli conjugate pair:

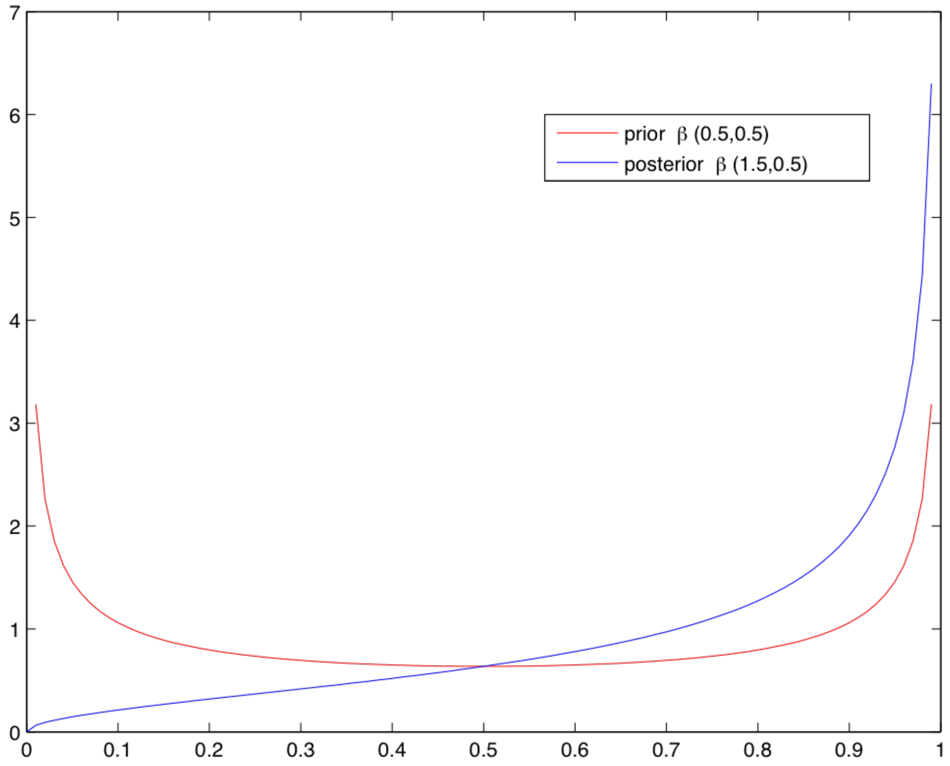
$$\mathrm{Beta}(\theta \mid \alpha + \textbf{num.-heads}, \beta + \textbf{num.-tails}) \propto \mathrm{Beta}(\theta \mid \alpha, \beta) \prod_i \mathrm{Bernoulli}(x_i \mid \theta)$$

**Same PDF**

# Example: Beta-Bernoulli

After a single coinflip of heads (x=1) the posterior is…

$$\mathrm{Beta}(\theta \mid \alpha + x, \beta + 1 - x)$$




The prior (red) is a fair coin,

$$\mathrm{Beta}(\theta \mid \alpha = 0.5, \beta = 0.5)$$

After observing one head, the posterior (blue) concentrates on heads,

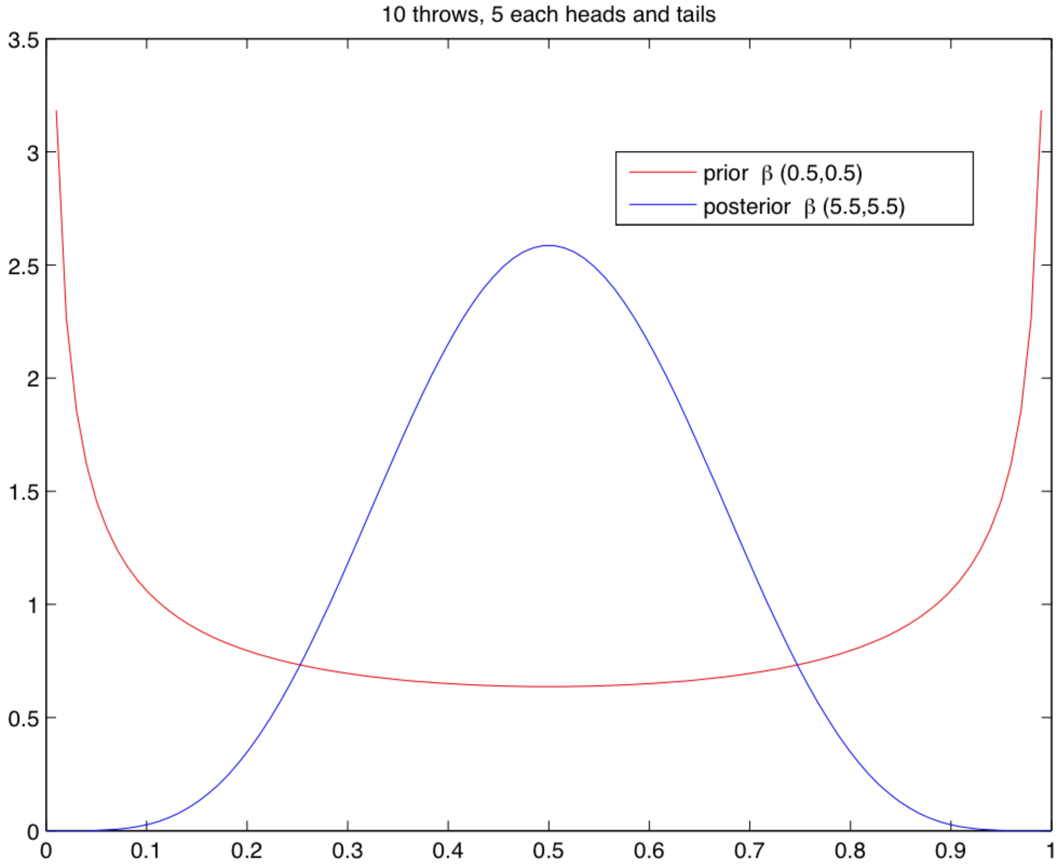
$$\mathrm{Beta}(\theta \mid 1.5, 0.5)$$

**What do you expect if we flip N=10 times with 5 heads and 5 tails?**

# Example: Beta-Bernoulli

After a N=10 flips (5 heads, 5 tails) we have…

$$\mathrm{Beta}(\theta \mid \alpha + 5, \beta + 5) = \mathrm{Beta}(\theta \mid 5.5, 5.5)$$



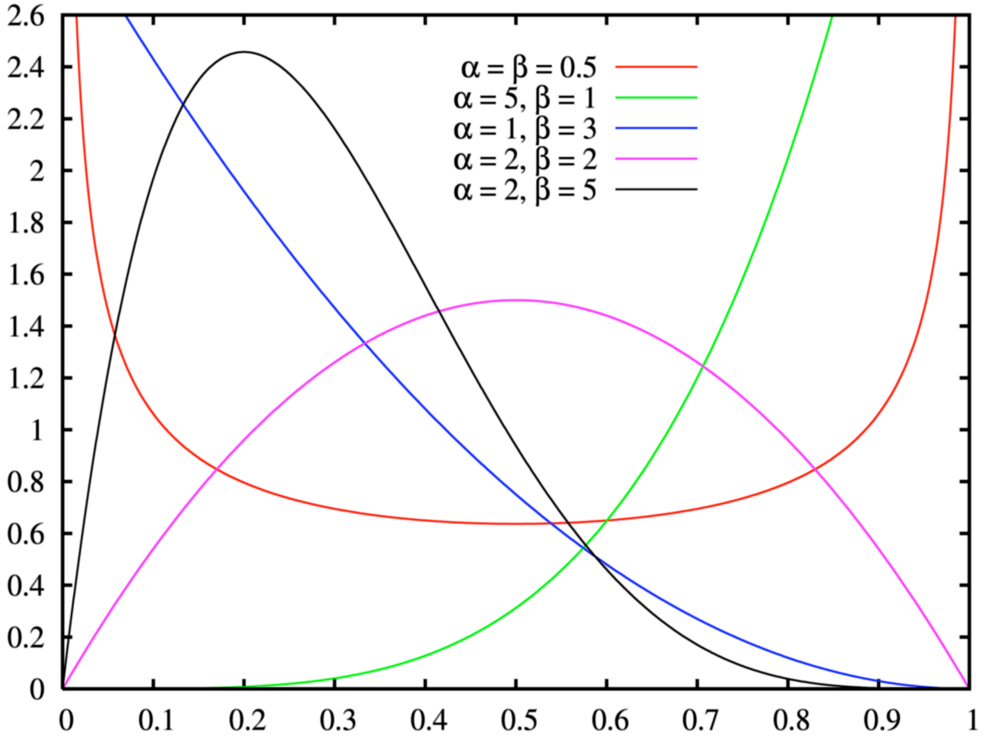

Posterior concentrates on fair coin $\theta = 0.5$

# Example: Beta-Bernoulli

**Bernoulli** *A.k.a. the* ***coinflip*** *distribution on* <u>*binary*</u> *RVs* $X \in \{0, 1\}$

$$\text{Bernoulli}(X \mid \theta) = \theta^X (1-\theta)^{(1-X)}$$

**Beta** *distribution on* $\theta \in (0,1)$ *with* $\alpha, \beta > 0$ *has PDF,*




$$\text{Beta}(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

For N coinflips $x_1, \ldots, x_N$ the posterior is,

$$\text{Beta}(\theta \mid \alpha + \sum_i x_i, \beta + N - \sum_i x_i)$$

# Example: Beta-Bernoulli

$$\mathrm{Beta}(\theta \mid \alpha, \beta) \prod_{i=1}^{N} \mathrm{Bernoulli}(x_i \mid \theta) \propto$$

$$\propto \theta^{\alpha-1}(1-\theta)^{\beta-1} \prod_i \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_i x_i}(1-\theta)^{\sum_i (1-x_i)}$$

$$= \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{\sum_i x_i}(1-\theta)^{\left(N-\sum_i x_i\right)}$$

$$= \theta^{\alpha-1+\sum_i x_i}(1-\theta)^{\beta-1+N-\sum_i x_i}$$

$$\propto \mathrm{Beta}(\theta \mid \alpha + \sum_i x_i, \beta + N - \sum_i x_i)$$

# Other Conjugate Pairs

| Likelihood | Model Parameters | Conjugate Prior |
|---|---|---|
| Normal | Mean | Normal |
| Normal | Mean / Variance | Normal-Inv-Gamma |
| Multivariate Normal | Mean / Variance | Normal-Inv-Wishart |
| Multinomial | Probability vector | Dirichlet |
| Gamma | Rate | Gamma |
| Poisson | Rate | Gamma |
| Exponential | Rate | Gamma |

Wikipedia has a nice list of standard conjugate forms…

https://en.wikipedia.org/wiki/Conjugate_prior

# Priors in AI / ML / Data Science

- Priors are often used as *regularizers* (promote smoothing)
  - Reduces overfitting as random noise is not smooth
  - Often regularizers can be of simple form, even conjugate

- Priors often house sophisticated domain knowledge
  - Possibly from earlier encounters with data
  - Possibly problem constraints (e.g. $\theta$ must be nonnegative)
  - World knowledge is complex, so good priors are often complex and **not conjugate**

# Choosing a Prior

- Conjugate priors can keep posteriors in closed form
  - This can speed up our codes (a lot!)
- The conjugate priors for standard distributions are fairly expressive
  - Often they can serve the purpose
- They are cool (better than doing nothing or the wrong thing)
- But they require that the likelihood is of a standard form
  - This is often a lot to hope for!
- Simply expressed functions may not be able to encode what you know
  - Constraints, non-local relationships

# Prediction

Can make predictions of unobserved $\tilde{y}$ before seeing any data,

$$p(\widetilde{y}) = \sum_k p(\theta = k) p(\widetilde{y} \mid \theta = k)$$

**Similar calculation to marginal likelihood**

*This is the* ***prior predictive*** *distribution*

For continuous parameters sum turns into integral,

$$p(\tilde{y}) = \int p(\theta) p(\tilde{y} \mid \theta) \, d\theta$$

*This is a prediction based on* ***no observed data***

# Prediction

When we observe $y$ we can predict future observations $\tilde{y}$,

$$p(\widetilde{y}) = \sum_k \underbrace{p(\theta = k \mid y)}_{\text{This is now the posterior}} p(\widetilde{\,} | \theta = k)$$

*This is the **posterior predictive** distribution*

Again, for continuous parameters sum turns into integral,

$$p(\tilde{y} \mid y) = \int p(\theta \mid y) p(\tilde{y} \mid \theta)\, d\theta$$

# Prediction Example

About 29% of American adults have high blood pressure (BP). Home test has 30% false positive rate and no false negative error.


What is the likelihood of *another* positive measurement?

$$p(\tilde{y} = true \mid y = true) = \sum_{\theta \in \{true, false\}} p(\theta \mid y = true) p(\tilde{y} = true \mid \theta)$$

$$= 0.42 * 0.30 + 0.58 * 1.00 \approx 0.71$$

**What conclusions can be drawn from this calculation?**

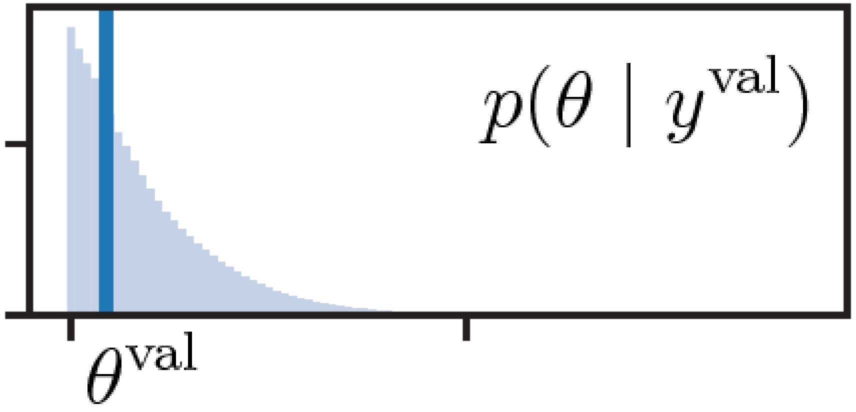
# Model Validation

*How do we know if the model* $p(\theta, y)$ *is* <u>*good*</u>*?*

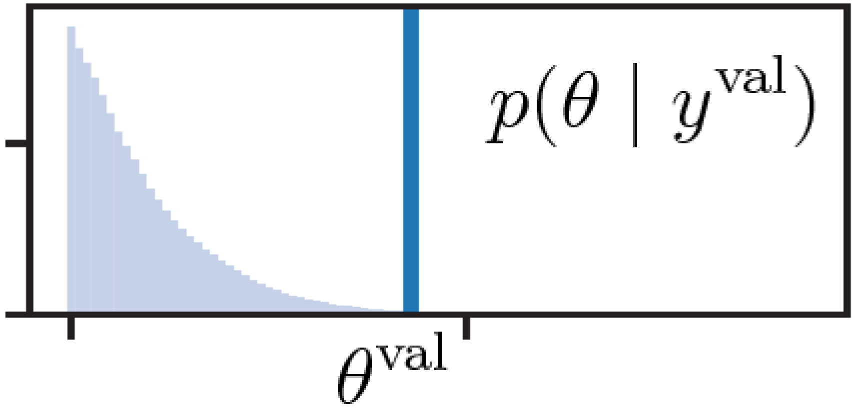
**Supervised Learning**

Validation set $\{(\theta^{\text{val}}, y^{\text{val}})\}$ consists of known $\theta^{\text{val}}$. Are true values typically preferred under the posterior?

Good (maybe lucky)




Not Good (maybe unlucky)




Repeat trials over validation set for more certainty

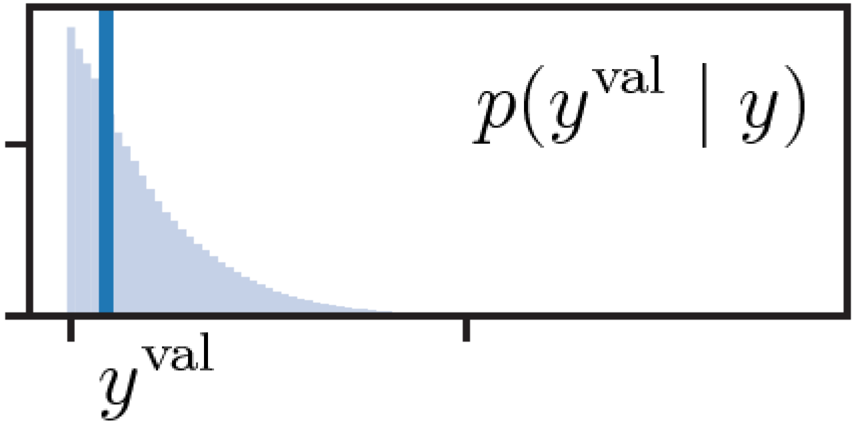
# Model Validation

*How do we know if the model* $p(\theta, y)$ *is <u>good</u>?*

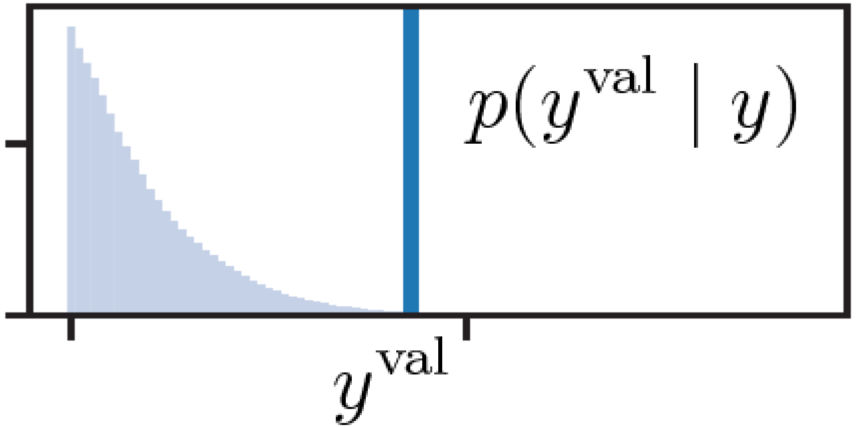
**Unsupervised Learning**

Validation set $\{y^{\text{val}}\}$ only contains observable data. Check validation data against posterior-predictive distribution.

Good (maybe lucky)




Not Good (maybe unlucky)




Repeat trials over validation set for more certainty

# Likelihood and Odds Ratios

Which parameter value $\theta_1$ or $\theta_2$ is more likely to have generated the observed data $y$ ?

The **posterior odds ratio** is:

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \underbrace{\frac{p(\theta_1)}{p(\theta_2)}}_{\text{Prior Odds Ratio}} \underbrace{\frac{p(y \mid \theta_1)}{p(y \mid \theta_2)}}_{\text{Likelihood Ratio}} \cancel{\frac{p(y)}{p(y)}}$$

**Observe:** the marginal likelihood $p(y)$ cancels!

# Posterior Summarization

*Ideally we would report the <u>full posterior distribution</u> as the result of inference…but this is not always possible*

**Summary of Posterior Location:**

Point estimates: mean (MMSE), mode, median (min. absolute error)

**Summary of Posterior Uncertainty:**

Credible intervals / regions, posterior entropy, variance

**Bayesian analysis should report uncertainty when possible**

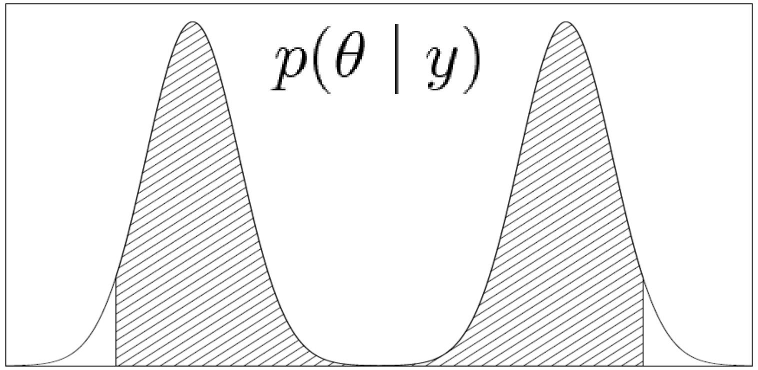
# Credible Interval

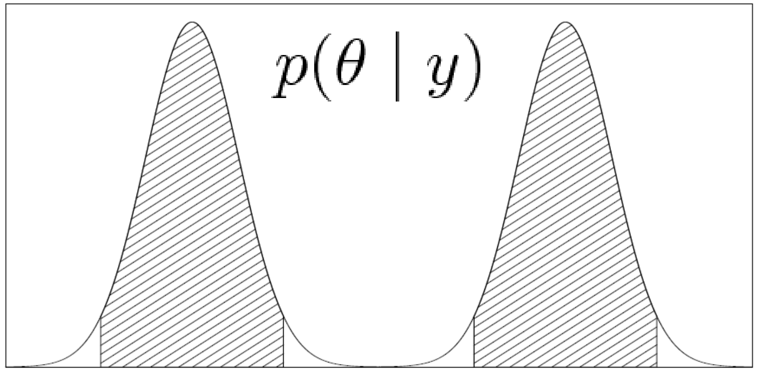
**Def.** For parameter $0 < \alpha < 1$ the $100(1-\alpha)\%$ credible interval $(L(y), U(y))$ satisfies,

$$p(L(y) < \theta < U(y) \mid y) = \int_{L(y)}^{U(y)} p(\theta \mid y) = 1 - \alpha$$

**Interval containing fixed percentage of posterior probability density.**

**Note:** This is not unique -- consider the 95% intervals below:




[Source: Gelman et al., “Bayesian Data Analysis”]

# Frequentist Inference

**Example:** Suppose we observe the outcome of N coin flips. $y = \{y_1, \ldots, y_N\}$ . What is the probability of heads $\theta$ (coin bias)?

- Coin bias $\theta$ is <u>not random</u> (e.g. there is some *true* value)
- Uncertainty reported as <u>confidence interval</u> (typically 95%)

Correct Interpretation: On repeated trials of N coin flips $\theta$ will fall inside the confidence interval 95% of the time (in the limit)

- Inferences are valid for multiple trials, **never on single trials**

Wrong Interpretation: For *this trial* there is a 95% chance $\theta$ falls in the confidence interval

# Bayesian Inference

*Posterior distribution is complete representation of uncertainty*

**Prior Belief**

**Likelihood**

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

**Marginal Likelihood (more on this later)**

- Must specify a prior belief $p(\theta)$ about coin bias
- Coin bias $\theta$ is a random quantity
- Interval $p(l(y) < \theta < u(y) \mid y) = 0.95$ can be reported in lieu of full posterior, and takes intuitive interpretation for a single trial

Interval Interpretation: For this experiment there is a 95% chance that $\theta$ lies in the interval

# Summary

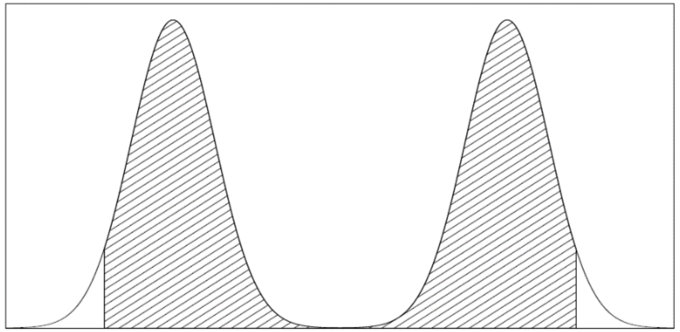
- Bayesian statistics interprets probability differently than classical stats
  - Frequentist: Probability → Long run odds in repeated trials
  - Bayesian: Probability → Belief of outcome that captures all uncertainty
- Bayesian models treat unknown parameter as random, with a prior
- Bayesian inference via the *posterior distribution* using Bayes' rule

$$p(\theta \mid y) = \frac{p(\theta)p(y \mid \theta)}{p(y)}$$

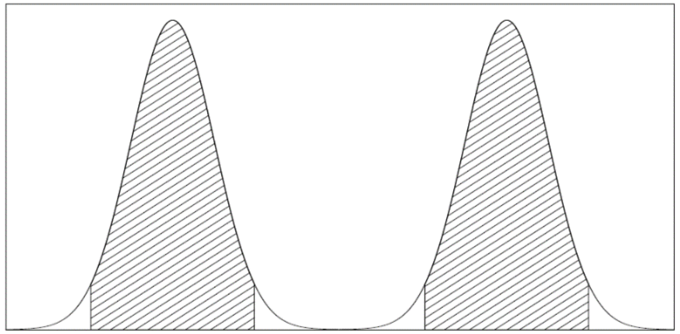
- Bayesian estimators minimize expected risk (e.g. MMSE)
- Maximum a posteriori (MAP) estimate maximizes posterior probability

# Summary

- Conjugate prior-posterior pairs ensure closed-form posterior inference
- Posterior uncertainty can be characterized by credible intervals



**Not necessarily unique**



- Selecting models can be done via posterior odds ratio

$$\frac{p(\theta_1 \mid y)}{p(\theta_2 \mid y)} = \frac{p(\theta_1)}{p(\theta_2)} \frac{p(y \mid \theta_1)}{p(y \mid \theta_2)} \cancel{\frac{p(y)}{p(y)}}$$

- Parameter can be marginalized out via prior/posterior predictive dist’n