$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

we call P(A) the "prior"

and P(A|B) the "posterior"

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning…*

Other Forms of Bayes Rule

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \sim A)P(\sim A)}$$

$$P(A|B \wedge X) = \frac{P(B \mid A \wedge X)P(A \wedge X)}{P(B \wedge X)}$$

Parameter Estimation

- How to estimate parameters from data?

Maximum Likelihood Principle:

Choose the parameters that maximize the probability of the observed data!

Maximum Likelihood Estimation Recipe

1. Use the log-likelihood
2. Differentiate with respect to the parameters
3. *Equate to zero and solve

*Often requires numerical approximation (no closed form solution)

An Example

- Let's start with the simplest possible case
  - Single observed variable
  - Flipping a bent coin

An Example

- Let's start with the simplest possible case
  - Single observed variable
  - Flipping a bent coin

- We Observe:
  - Sequence of heads or tails
  - HTTTTTHTHT
- Goal:
  - Estimate the probability that the next flip comes up heads

Assumptions

- Fixed parameter θ_H
  - Probability that a flip comes up heads
- Each flip is independent
  - Doesn't affect the outcome of other flips
- (IID) Independent and Identically Distributed

Example

- Let's assume we observe the sequence:
  - HTTTTTHTHT
- What is the **best** value of θ_H ?
  - Probability of heads

Example

- Let's assume we observe the sequence:
  - HTTTTTHTHT
- What is the **best** value of θ_H ?
  - Probability of heads
- Intuition: should be 0.3 (3 out of 10)
- Question: how do we justify this?

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

$$P(x_1 = H, x_2 = T, \ldots, x_m = T; \theta_H)$$

$$= P(x_1 = H; \theta_H) P(x_2 = T; \theta_H), \ldots P(x_m = T; \theta_H)$$

$$= \theta_H \times (1 - \theta_H), \times \ldots \times \theta_H$$

$$= \theta_H^3 \times (1 - \theta_H)^7$$

Maximum Likelihood Principle

- The value of θ_H which maximizes the probability of the observed data is best!
- Based on our assumptions, the probability of "HTTTTTHTHT" is:

$$P(x_1 = H, x_2 = T, \ldots, x_m = T; \theta_H)$$

$$= P(x_1 = H; \theta_H) P(x_2 = T; \theta_H), \ldots P(x_m = T; \theta_H)$$

$$= \theta_H \times (1 - \theta_H), \times \ldots \times \theta_H$$

$$= \theta_H^3 \times (1 - \theta_H)^7$$

This is the Likelihood Function

Maximum Likelihood Principle

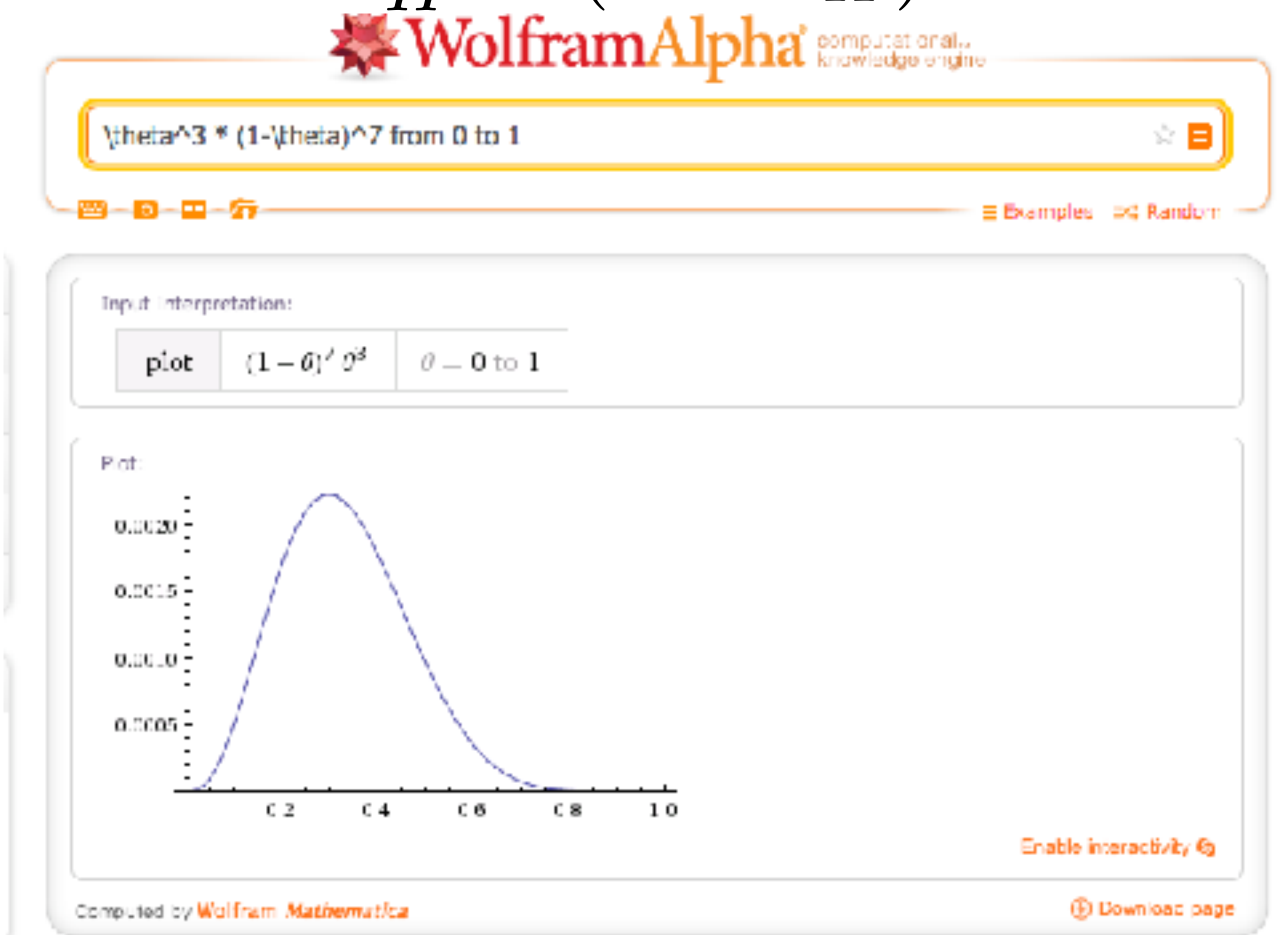
- Probability of "HTTTTTHTHT" as a function of θ_H

$$\theta_H^3 \times (1 - \theta_H)^7$$

Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

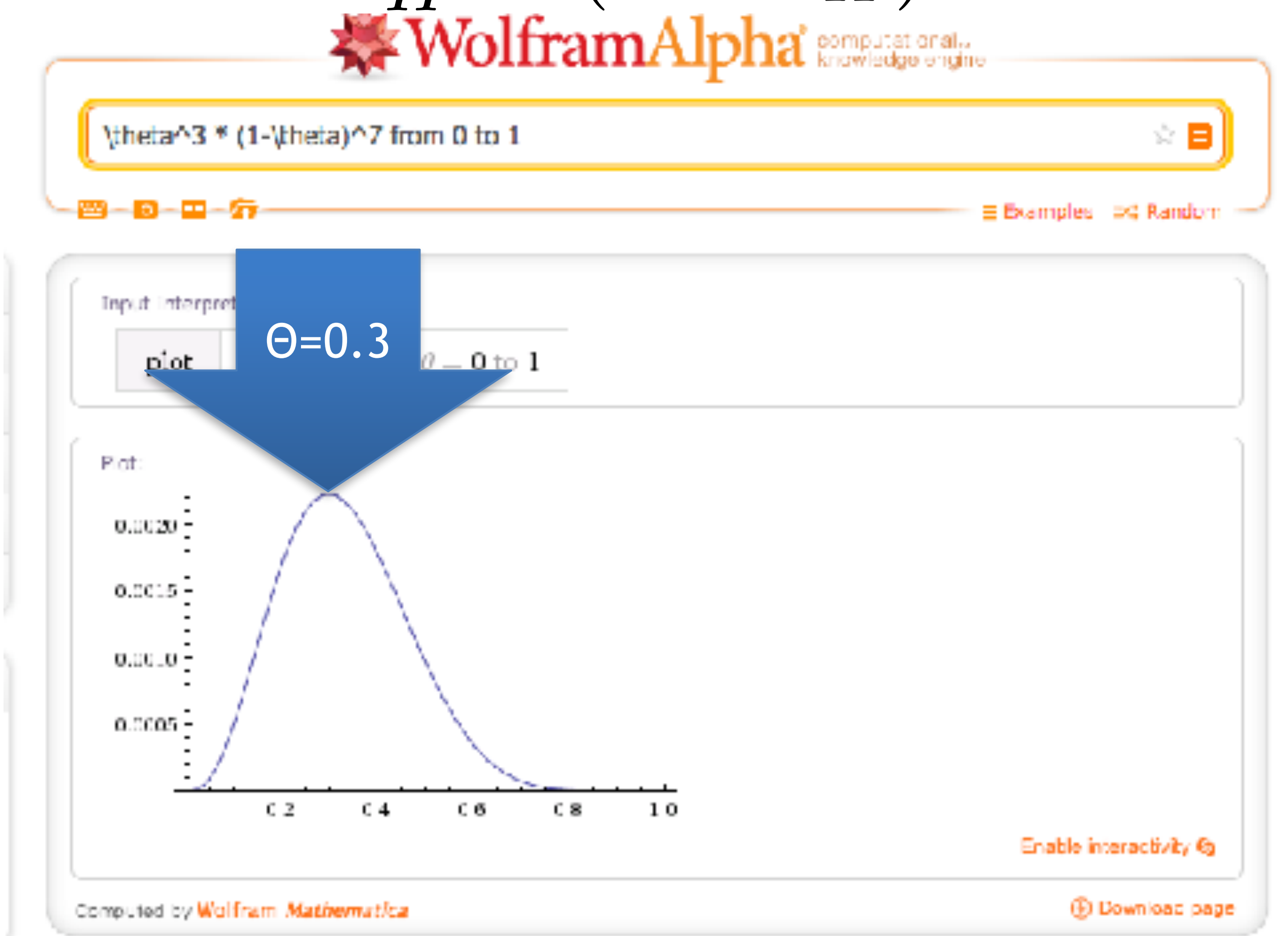
$$\theta_H^3 \times (1 - \theta_H)^7$$



Maximum Likelihood Principle

- Probability of “HTTTTTHTHT” as a function of θ_H

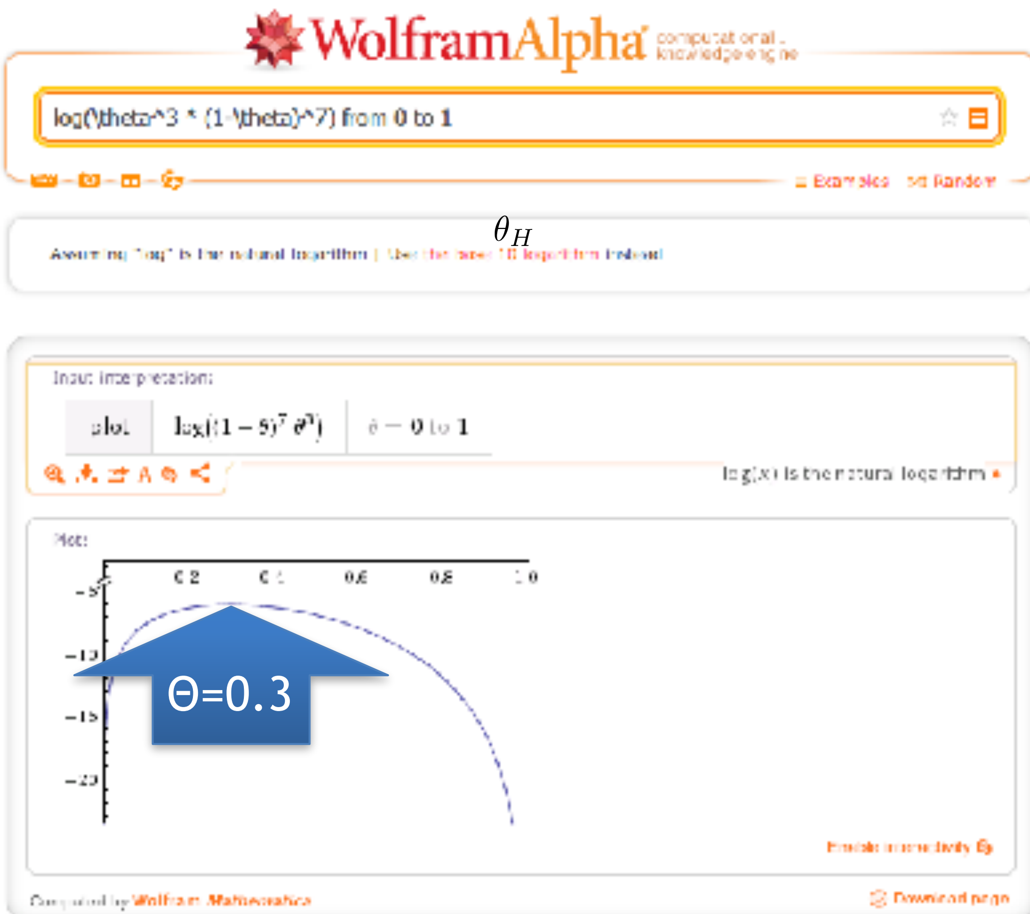
$$\theta_H^3 \times (1 - \theta_H)^7$$



Maximum Likelihood Principle

- Probability of "HTTTTTHTHT" as a function of θ_H

$$\log(\theta_H^3 \times (1-\theta_H)^7)$$



Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \log(\theta_H^{\#H}(1-\theta_H)^{\#T}) = 0$$

$$\frac{\partial}{\partial \theta_H} \log(\theta_H^{\#H}) + \log((1-\theta_H)^{\#T}) = 0$$

Log Identities

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1-\theta_H) = 0$$

Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1 - \theta_H) = 0$$

$$\frac{\#H}{\theta_H} - \frac{\#T}{1 - \theta_H} = 0$$

$$\hat{\theta} = \frac{\#H}{\#H + \#T}$$

Maximum Likelihood value of θ_H

$$\frac{\partial}{\partial \theta_H} \#H \log(\theta_H) + \#T \log(1 - \theta_H) = 0$$

$$\frac{\#H}{\theta_H} - \frac{\#T}{1 - \theta_H} = 0$$

$$\vdots$$

$$\hat{\theta} = \frac{\#H}{\#H + \#T}$$

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.

The problem with Maximum Likelihood

- What if the coin doesn’t look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.
  - No way to incorporate our prior knowledge!

The problem with Maximum Likelihood

- What if the coin doesn't look very bent?
  - Should be somewhere around 0.5?
- What if we saw 3,000 heads and 7,000 tails?
  - Should this really be the same as 3 out of 10?
- Maximum Likelihood
  - No way to quantify our **uncertainty**.
  - No way to incorporate our prior knowledge!

Q: how to deal with this problem?

Bayesian Parameter Estimation

- Let's just treat θ_H like any other variable
- Put a prior on it!
  - Encode our prior knowledge about possible values of θ_H using a probability distribution
- Now consider two probability distributions:

$$P(x_i|\theta_H) = \begin{cases} \theta_H, & \text{if } x_i = H \\ 1 - \theta_H, & \text{otherwise} \end{cases}$$

$$P(\theta_H) = ?$$

Posterior Over θ_H

$$P(\theta | x_1 = H, x_2 = T, \ldots, x_m = T)$$

Posterior Over θ_H

$$P(\theta | x_1 = H, x_2 = T, \ldots, x_m = T)$$

$$= \frac{P(x_1 = H, x_2 = T, \ldots, x_m = T | \theta) P(\theta)}{P(x_1 = H, x_2 = T, \ldots, x_m = T)}$$

Posterior Over θ_H

$$P(\theta | x_1 = H, x_2 = T, \ldots, x_m = T)$$

$$= \frac{P(x_1 = H, x_2 = T, \ldots, x_m = T | \theta) P(\theta)}{P(x_1 = H, x_2 = T, \ldots, x_m = T)}$$

$$= \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

How can we encode prior knowledge?

- Example: The coin doesn’t look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha,\beta) = \frac{1}{B(\alpha,\beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

Hyper-Parameters

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

Hyper-Parameters

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

How can we encode prior knowledge?

- Example: The coin doesn't look very bent
  - Assign higher probability to values of θ_H near 0.5
- Solution: The **Beta Distribution**

$$P(\theta_H|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta_H^{\alpha-1}(1-\theta_H)^{\beta-1}$$

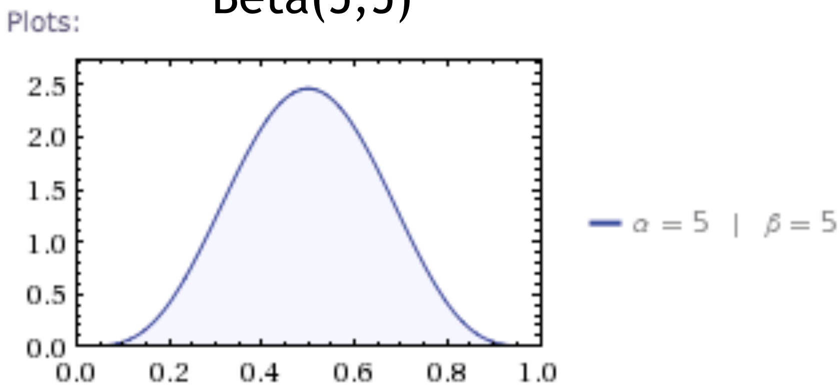
Hyper-Parameters

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

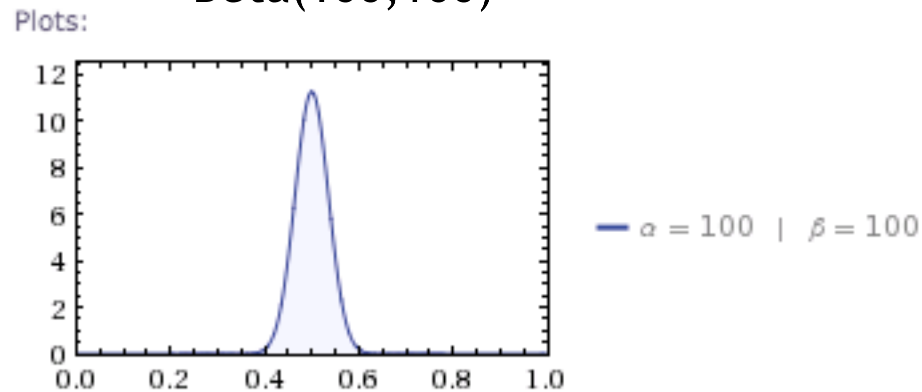
Gamma is a continuous generalization of the Factorial Function

Beta Distribution

Beta(5,5)

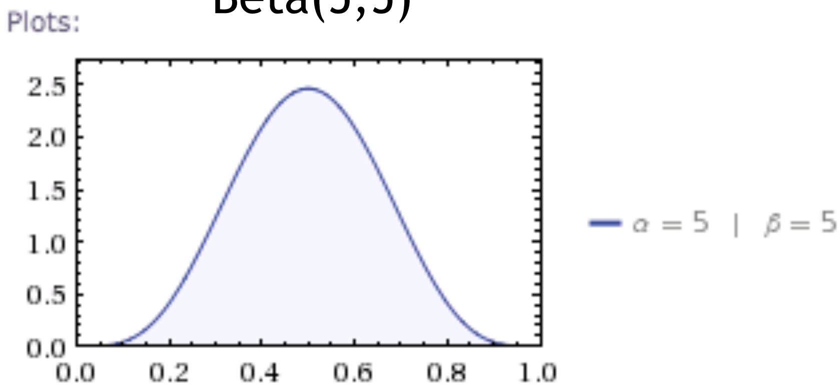


Beta(100,100)

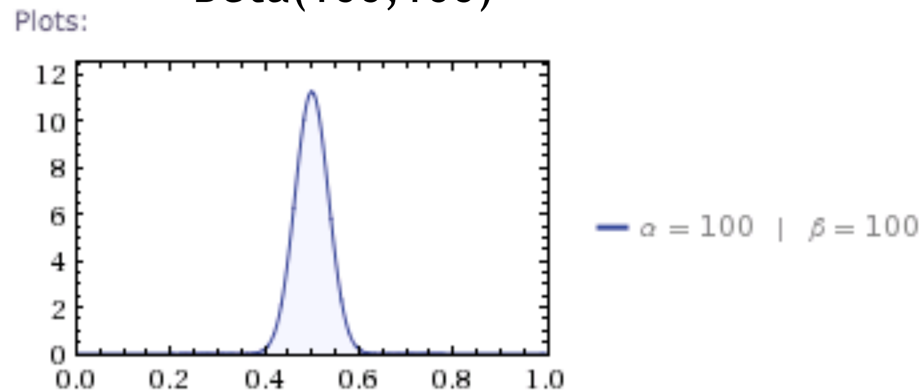


Beta Distribution

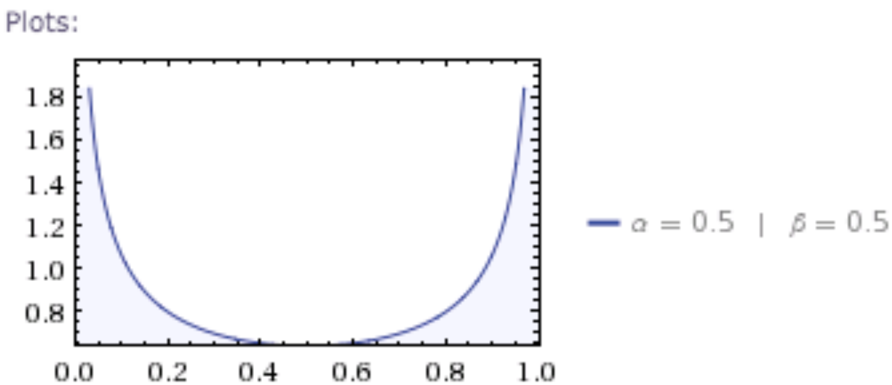
Beta(5,5)



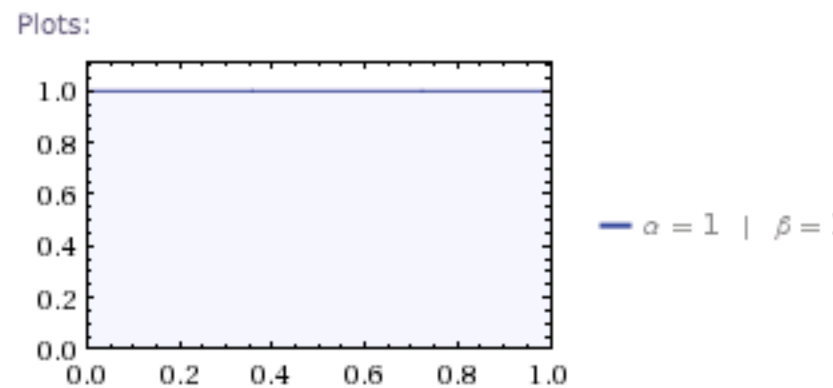
Beta(100,100)



Beta(0.5,0.5)



Beta(1,1)



MAP Estimate

$$\theta^{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \frac{\#H + \alpha - 1}{\#T + \#H + \alpha + \beta - 2}$$

MAP Estimate

$$\theta^{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \frac{\#H + \alpha - 1}{\#T + \#H + \alpha + \beta - 2}$$

-Add-N smoothing
-Pseudo-counts