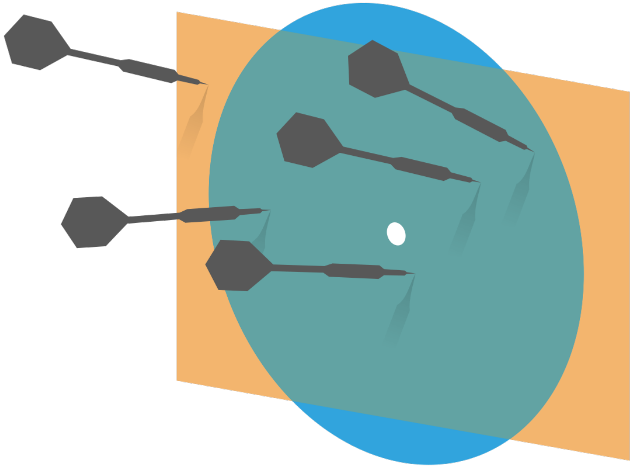
# Probability and Statistics for Computer Science



Credit: wikipedia

"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write." H. G. Wells

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.20.2020

# Last time

- Hypothesis test
- Chi-square test
- Maximum likelihood Estimation (MLE)

# Objectives

- More on Maximum likelihood Estimation (MLE)
- Bayesian Inference (MAP)

# Maximum likelihood estimation (MLE)

- We write the probability of seeing the data D given parameter θ

$$L(\theta) = P(D|\theta)$$

- The **likelihood function** $L(\theta)$ is **not** a probability distribution

- The **maximum likelihood estimate (MLE)** of θ is

$$\hat{\theta} = arg \max_{\theta} L(\theta)$$

# Likelihood function: binomial example

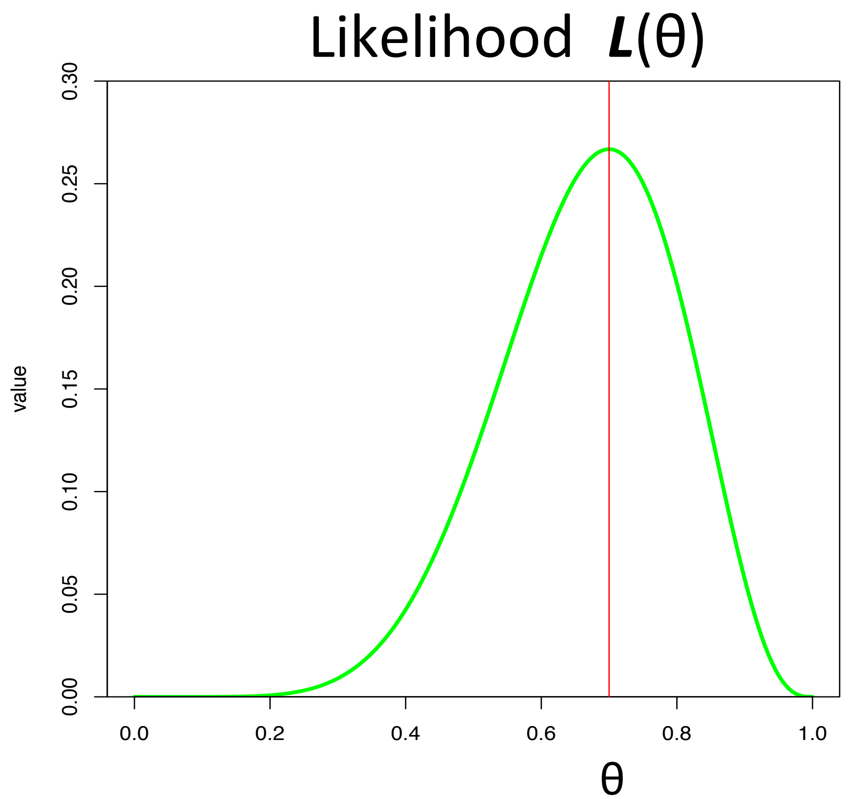
- Suppose we have a coin with unknown probability of θ coming up heads
- We toss it ***10*** times and

  observe ***7*** heads
- The likelihood function is:

$$P(D|\theta) = \binom{10}{7}\theta^7(1-\theta)^3$$

- The MLE is

$$\hat{\theta} = 0.7$$



# Q. What is the MLE of binomial N=12, k=7

A. 12!/7!/5!

B. 7/12

C. 5/12

D. 12/7

# Q. What is the MLE of Poisson $k_1=5$, $k_2=7$, $n=2$

A. 6

B. 35/2

C. 12

D. other

# MLE Example

You find a 5-sided die and want to estimate its probability θ of coming up 5, you decided to roll it 12 times and then roll it until it comes up 5. You rolled 15 times altogether and found there were 3 times when the die came up 5. Write down the likelihood function L(θ).

# Drawbacks of MLE

- Maximizing some likelihood or log-likelihood function is mathematically hard
- If there are few data items, the MLE estimate maybe very unreliable
  - If we observe 3 heads in 10 coin tosses, should we accept that p(heads)= 0.3 ?
  - If we observe 0 heads in 2 coin tosses, should we accept that p(heads)= 0 ?

# Bayesian inference

- In MLE, we maximized the likelihood function

$$L(\theta) = P(D|\theta)$$

- In Bayesian inference, we will maximize the **posterior**, which is the probability of the parameters **θ** given the observed data D.

$$P(\theta|D)$$

- Unlike $L(\theta)$, the posterior is a probability distribution

- The value of **θ** that maximizes $P(\theta|D)$ is called the **maximum a posterior (MAP)** estimate $\hat{\theta}$

# The components of Bayesian Inference

- From Bayes rule

# The components of Bayesian Inference

- From Bayes rule
  - Prior, assumed distribution of **θ** before seeing data **D**
  - Likelihood function of **θ** seeing **D**
  - Total Probability seeing **D** --- P(**D**)
  - **Posterior**, distribution of **θ** given **D**

# The usefulness of Bayesian inference

- From Bayes rule

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- Bayesian inference allows us to include prior beliefs about θ in the prior $P(\theta)$, which is useful
  - When we have reasonable beliefs, such as a coin can not have P(heads) = 0
  - When there isn't much data
  - We get a distribution of the posterior, not just one maxima

# Bayesian Inference: a discrete prior

- Suppose we have a coin of unknown probability θ of heads
  - We see 7 heads in 10 tosses (**D**)
  - We assume the prior about θ.

$$P(\theta) = \begin{cases} \frac{2}{3} & if\ \theta = 0.5 \\ \frac{1}{3} & if\ \theta = 0.6 \\ 0 & otherwise \end{cases}$$

  - We have this likelihood:

$$P(D|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$$

  - What is the posterior $P(\theta|D)$ ?

# Bayesian Inference: a discrete prior

- We see 7 heads in 10 tosses (**D**)
- We assume the prior about θ.

$$P(\theta) = \begin{cases} \frac{2}{3} & if \ \theta = 0.5 \\ \frac{1}{3} & if \ \theta = 0.6 \\ 0 & otherwise \end{cases}$$

- We have this likelihood:

$$P(D|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$$

- What is the posterior $P(\theta|D)$ ?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

# Bayesian Inference: a discrete prior

- We see 7 heads in 10 tosses (**D**)
- We assume the prior about θ.

$$P(\theta) = \begin{cases} \frac{2}{3} & if\ \theta = 0.5 \\ \frac{1}{3} & if\ \theta = 0.6 \\ 0 & otherwise \end{cases}$$

- We have this likelihood:

$$P(D|\theta) = \binom{10}{7}\theta^7(1-\theta)^3$$

- What is the posterior $P(\theta|D)$ ?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$$P(D) = \sum_{\theta_i \in \theta} P(D|\theta_i)P(\theta_i)$$

# Bayesian Inference: a discrete prior

- We see 7 heads in 10 tosses (**D**)
- We assume the prior about θ.

$$P(\theta) = \begin{cases} \frac{2}{3} & if\ \theta = 0.5 \\ \frac{1}{3} & if\ \theta = 0.6 \\ 0 & otherwise \end{cases}$$

- We have this likelihood:

$$P(D|\theta) = \binom{10}{7} \theta^7 (1-\theta)^3$$

- What is the posterior $P(\theta|D)$ ?

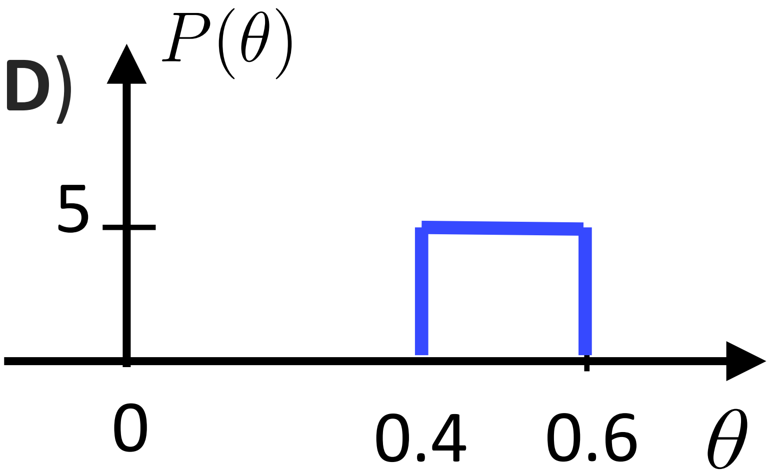
$$P(\theta|D) = \begin{cases} 0.52 & if\ \theta = 0.5 \\ 0.48 & if\ \theta = 0.6 \\ 0 & otherwise \end{cases}$$

MAP estimate=?

# Bayesian Inference: a continuous prior

- Suppose we have a coin of unknown probability θ of heads
- We see 7 heads in 10 tosses (**D**)
- We assume

$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

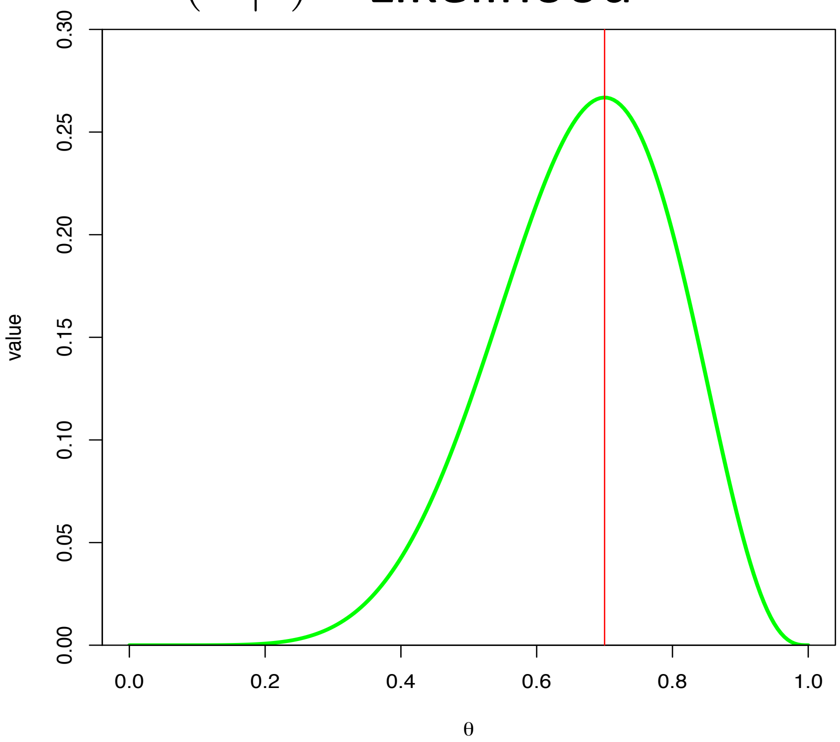


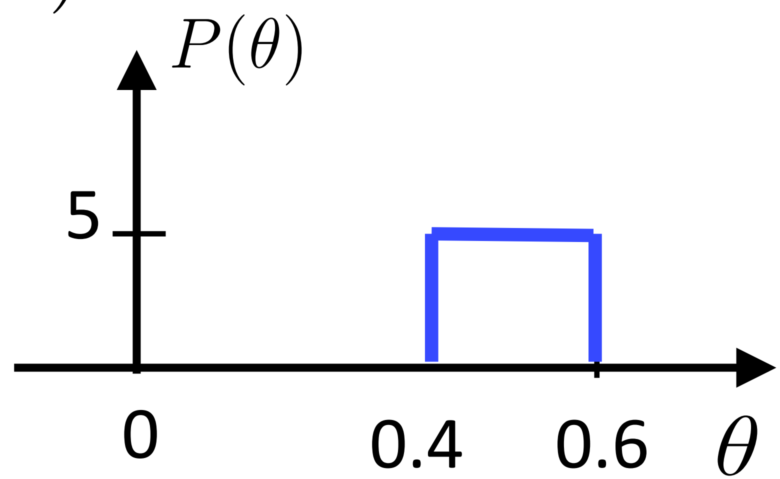

- What is the posterior $P(\theta|D)$ ?

# Bayesian Inference: a continuous prior

- What is the posterior $P(\theta|D)$ ?

$P(D|\theta)$ = Likelihood



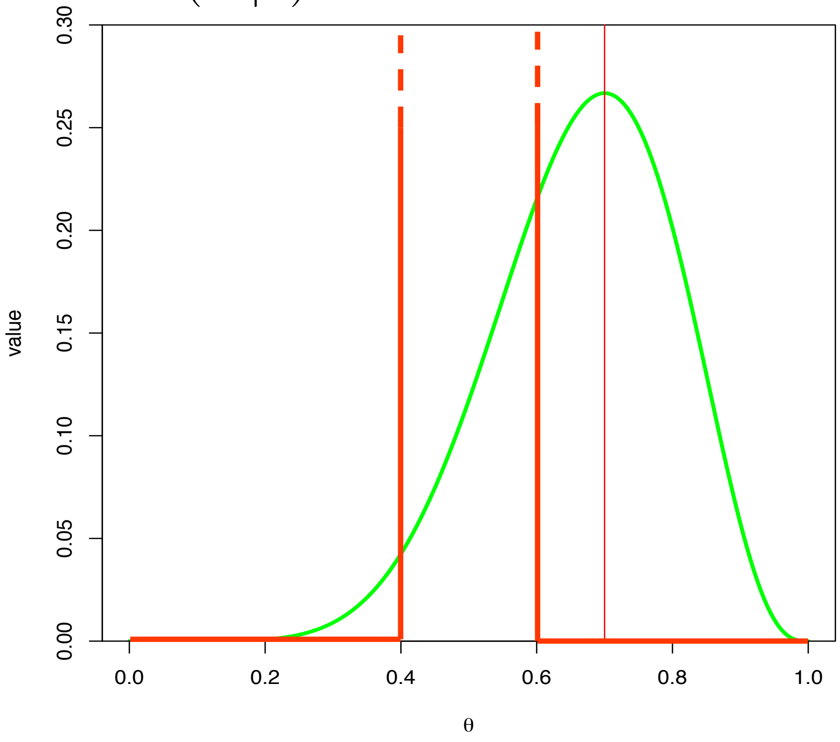

$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

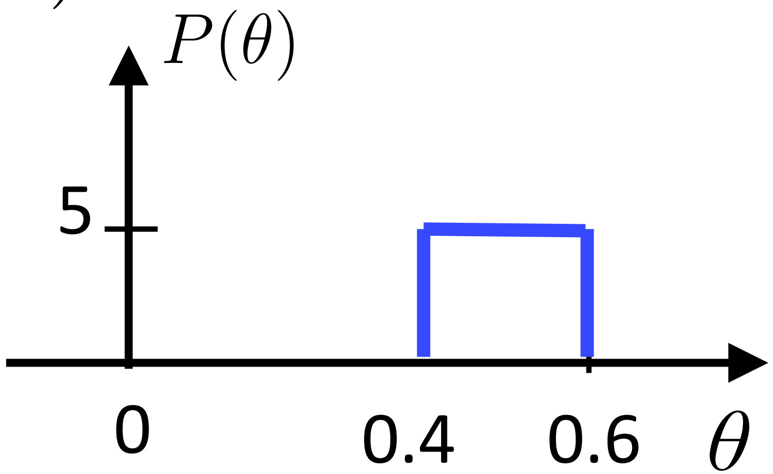
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

# Bayesian Inference: a continuous prior

- What is the posterior $P(\theta|D)$ ?

$P(D|\theta)$ = Likelihood



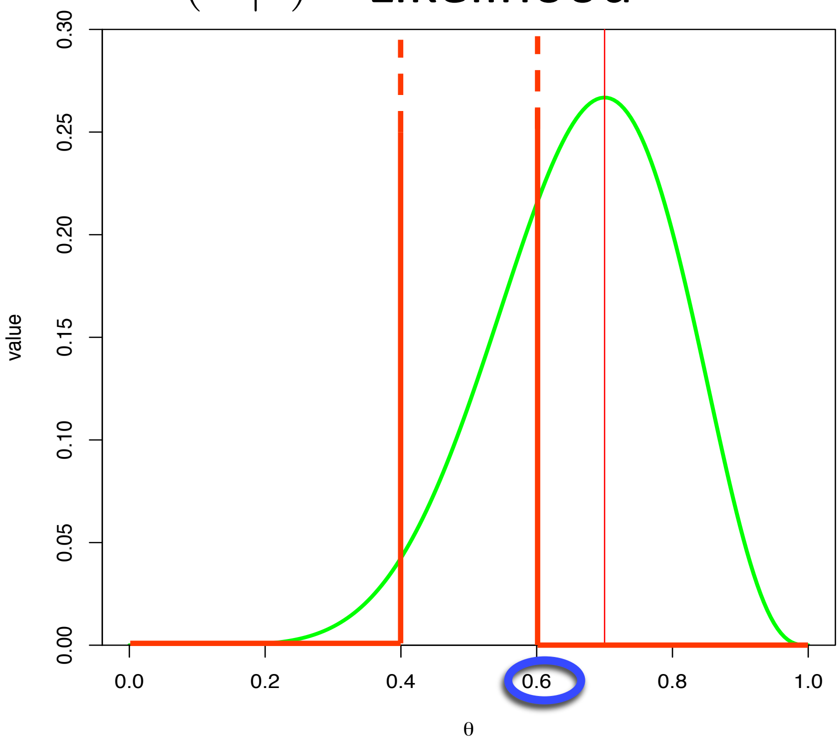

$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

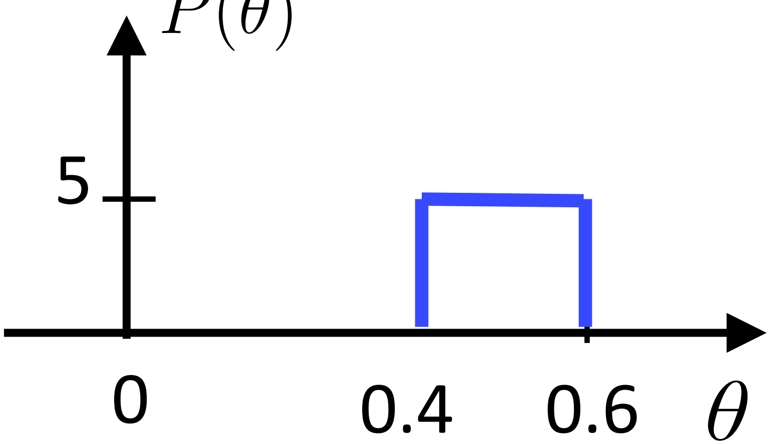
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

# Bayesian Inference: a continuous prior

- What is the posterior $P(\theta|D)$ ?

$P(D|\theta)$ = Likelihood




$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

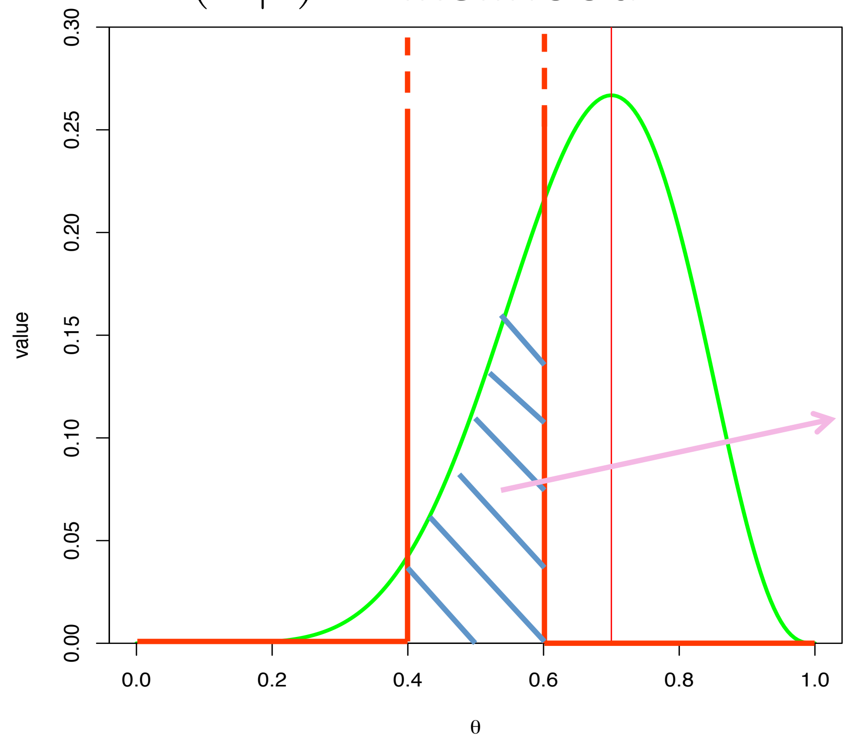
**MAP** $\hat{\theta}$ **=0.6**

# The constant in the Bayesian inference

$$P(D) = \int_\theta P(D|\theta)P(\theta)d\theta$$

- It's not always possible to calculating P(D) in closed form.
- There are a lot of approximation methods.

$P(D|\theta)$ = Likelihood




Scale by 5 for this example

# Drawbacks of Bayesian inference

- Maximizing some posteriors $P(\theta|D)$ is difficult
- Some choices of prior $P(\theta)$ can overwhelm any data observed.
- It's hard to justify a choice of prior

# The concept of conjugacy

- For a given likelihood function $P(D|\theta)$, a prior $P(\theta)$ is its conjugate prior if it has the following properties:
  - $P(\theta)$ belongs to a family of distributions that are expressive
  - The posterior $P(\theta|D) \propto P(D|\theta)P(\theta)$ belongs to the same family of distribution as the prior $P(\theta)$
  - The posterior $P(\theta|D)$ is easy to maximize
- For example, a conjugate prior for binomial likelihood function is Beta distribution

# Beta distribution

- A distribution is Beta distribution if it has the following pdf: $$P(\theta) = K(\alpha, \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

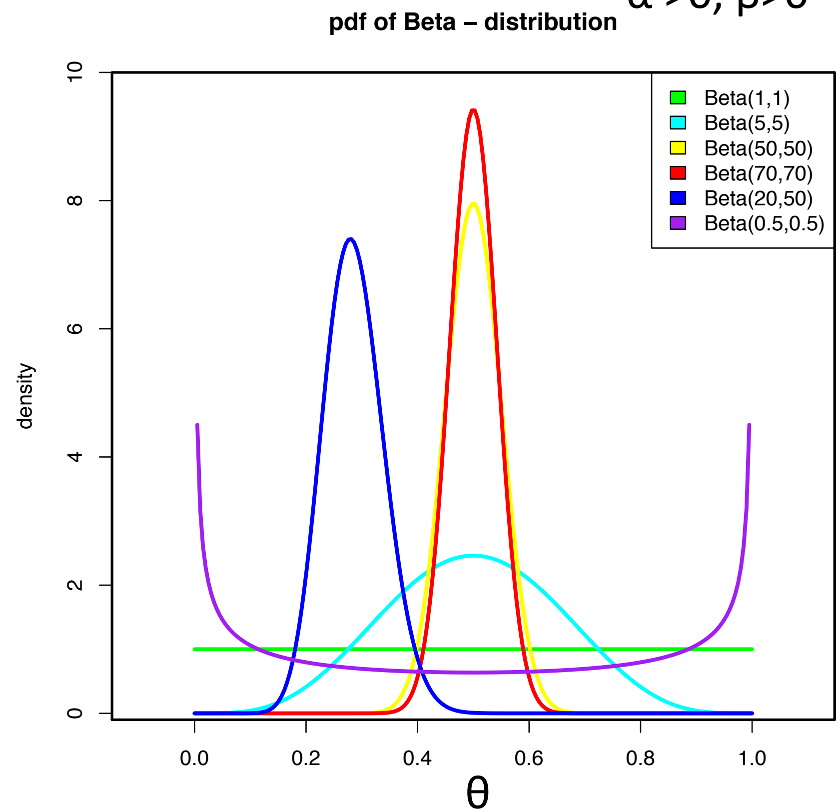
  = 0 O.W.

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

$0 \leq \Theta \leq 1$

$\alpha > 0, \beta > 0$

- Is an expressive family of distributions
- $Beta(\alpha = 1, \beta = 1)$ is uniform



# Beta distribution as the conjugate prior for Binomial likelihood

- The likelihood is Binomial (*N*, *k*)

$$P(D|\theta) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

- The Beta distribution is used as the prior

$$P(\theta) = K(\alpha, \beta)\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- So $P(\theta|D) \propto \theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$

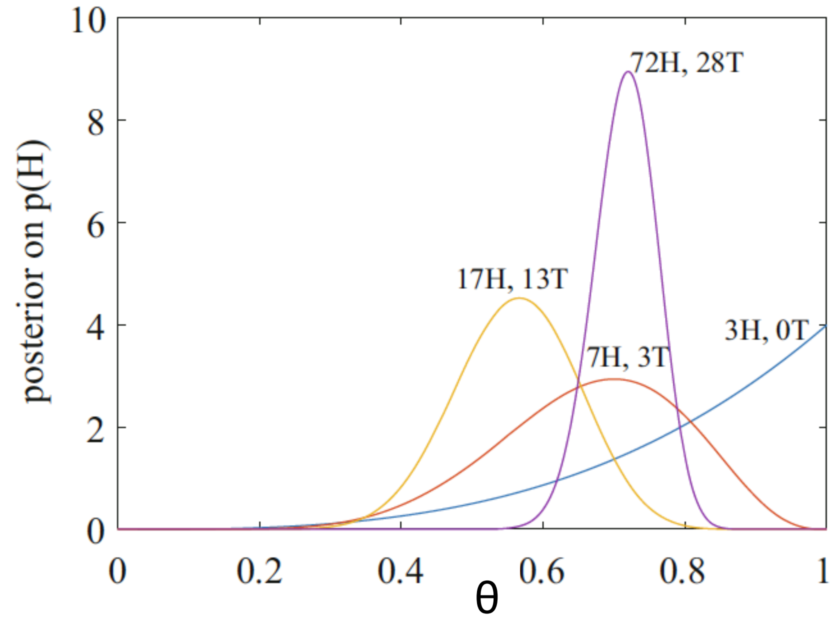
- Then the posterior is $Beta(\alpha+k, \beta+N-k)$

$$P(\theta|D) = K(\alpha+k, \beta+N-k)\theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$$

# The update of Bayesian posterior

- Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed.

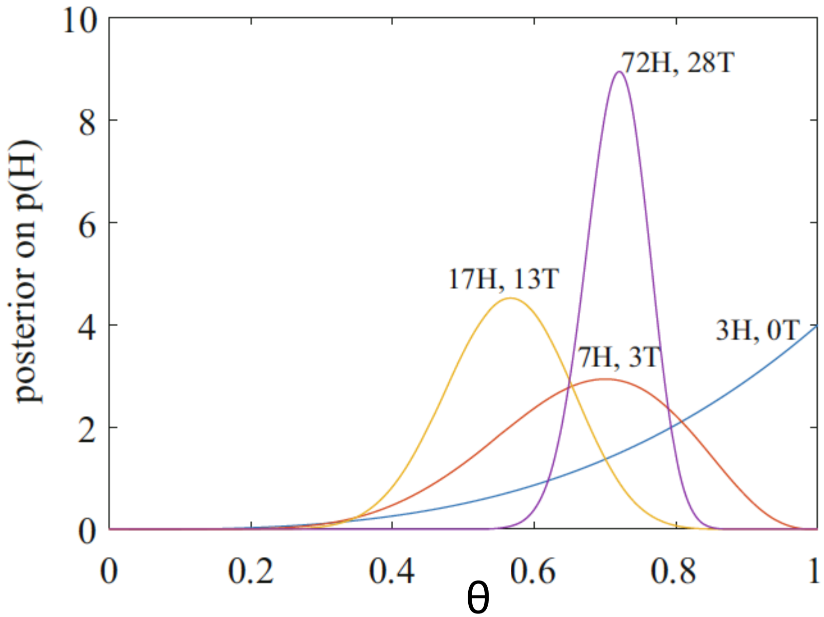
- Suppose we start with a uniform prior on the probability θ of heads
  - Then we see 3H 0T
  - Then we see 4H 3T for 7H 3T in total
  - Then we see 10H 10T for 17H 13T in total
  - Then we see 55H 15T for 72H 28T in total



# The update of Bayesian posterior

- Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed.
- Suppose we start with a uniform prior on the probability θ of heads

| N | k | α | β |
|---|---|---|---|
| | | 1 | 1 |
| 3 | 0 | 1 | 4 |
| 10 | 7 | 8 | 7 |
| 30 | 17 | 25 | 20 |
| 100 | 72 | 97 | 48 |



# Simulation of the update of Bayesian posterior

https://seeing-theory.brown.edu/bayesian-inference/index.html

# Maximize the Bayesian posterior (MAP)

- The posterior of the previous example is

$$P(\theta|D) = K(\alpha + k, \beta + N - k)\theta^{\alpha+k-1}(1-\theta)^{\beta+N-k-1}$$

- Differentiating and setting to 0 gives the MAP estimate

$$\hat{\theta} = \frac{\alpha - 1 + k}{\alpha + \beta - 2 + N}$$

θ

# Conjugate prior for other likelihood functions

- If the likelihood is Bernoulli or geometric, the conjugate prior is Beta
- If the likelihood is Poisson or Exponential, the conjugate prior is Gamma
- If the likelihood is normal with known variance, the conjugate prior is normal

$\theta$

# Assignments

- Finish Chapter 9 of the textbook
- Next time: Covariance matrix

# Additional References


- Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- Morris H. Degroot and Mark J. Schervish "Probability and Statistics”

# See you next time

*See You!*