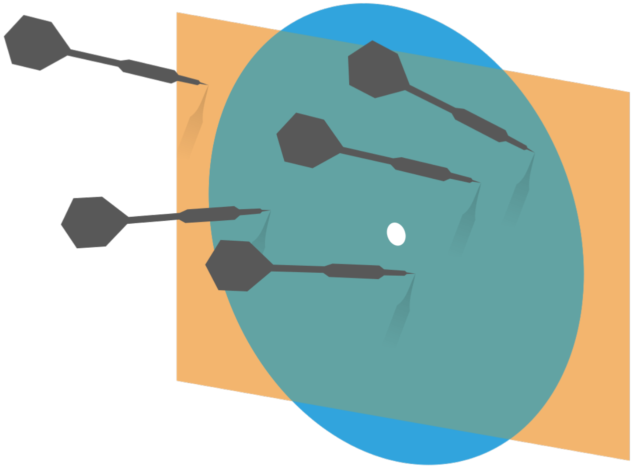
Probability and Statistics for Computer Science



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

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 10.20.2021

Last time

* Hypothesis test
* Chi-square test
* Maximum Likelihood Estimation (MLE) (I)

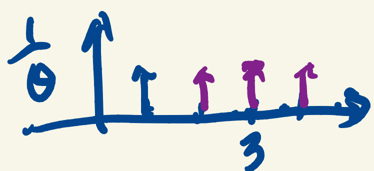
Objectives

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

If someone has a θ-sided die in a box, and tells you an outcome of 3 is observed, what is the likelihood function? what is the MLE of θ?

If someone has a θ-sided die in a box, and tells you an outcome of 3 is observed, what is the likelihood function? what is the MLE of θ?

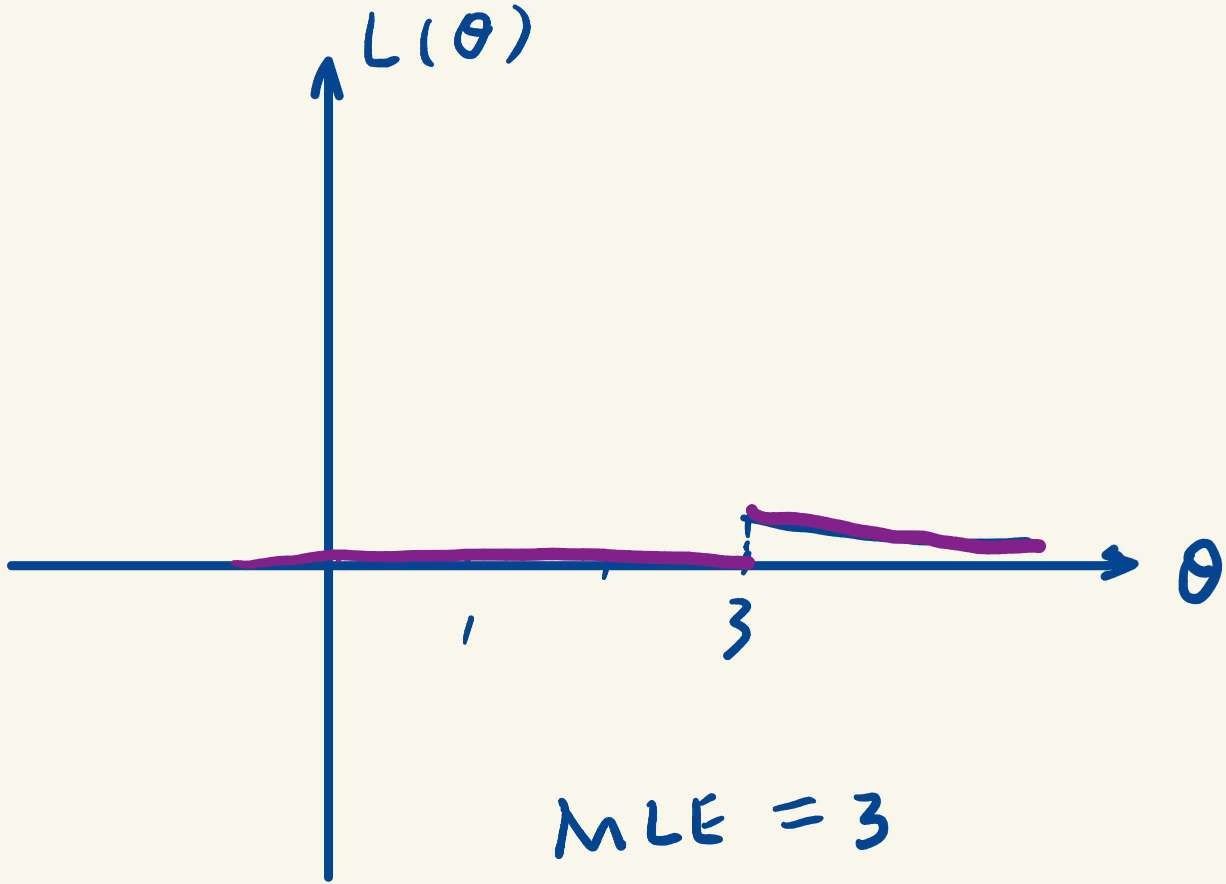
$$L(\theta) = P(D|\theta) = \begin{cases} 0 & \theta < 3 \\ \frac{1}{\theta} & \text{ow} \end{cases}$$



$$P(D_1 = 2 \cap D_2 = 3 \mid \theta)$$

$$L(\theta) = P(D_1 = 2 \mid \theta)\, P(D_2 = 3 \mid \theta)$$

$$= \begin{cases} 0 & \theta < 2 \\ \frac{1}{\theta} & \theta \geq 2 \end{cases} \times \begin{cases} 0 & \theta < 3 \\ \frac{1}{\theta} & \theta \geq 3 \end{cases}$$



MLE = 3

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

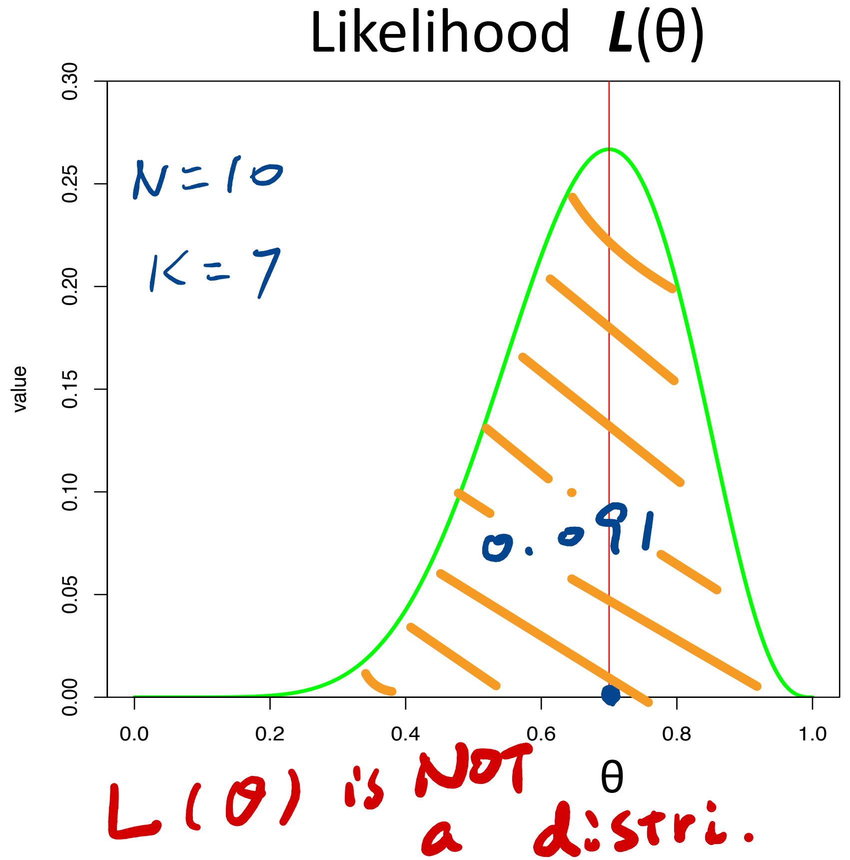
D: N, K

- The likelihood function is:

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

- The MLE is

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



Do it at home

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

A. 12!/7!/5!

B. 7/12

C. 5/12

D. 12/7

Do it at home

Q. What is the MLE of θ of geometric k=7

A. 7

B. 1/7

C. other

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

- Each x_i is one observed result from an IID trial

Q: MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

- Why is the above function defined by the product?

  A. IID samples are independent

  B. Each trial has identical probability function

  C. Both.

MLE with data from IID trials

- If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

- The likelihood function is hard to differentiate in general, except for the binomial and geometric cases.
- Clever trick: take the (natural) log

Log-likelihood function

- Since log is a strictly increasing function

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

- So we can aim to maximize the **log-likelihood function**

$$logL(\theta) = logP(D|\theta) = log \prod_{x_i \in D} P(x_i|\theta) = \sum_{x_i \in D} logP(x_i|\theta)$$

- The log-likelihood function is usually much easier to differentiate

Log-likelihood function: Poisson example

- Suppose we have data on the number of babies born each hour in a large hospital

hour	**1**	**2**	**…**	*N*
# of babies	k_1	k_2	**…**	k_N

- We can assume the data comes from a Poisson distribution with parameter λ

- What is the log likelihood function $LogL(\theta)$?

Log-likelihood function: Poisson example

$$L(\theta) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}$$

$$log\ L(\theta) = log\ (\prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}) = \sum_{i=1}^{N} log(\frac{e^{-\theta}\theta^{k_i}}{k_i!})$$

$$= \sum_{i=1}^{N} (-\theta + k_i\ log\theta - log\ k_i!)$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \; log\theta - log \; k_i!)$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \; log\theta - log \; k_i!)$$

$$\frac{d}{d\theta} log \; L(\theta) = 0 \; \Rightarrow \sum_{i=1}^{N}(-1 + \frac{k_i}{\theta} - 0) = 0$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \ log\theta - log \ k_i!)$$

$$\frac{d}{d\theta}log \ L(\theta) = 0 \ \Rightarrow \sum_{i=1}^{N}(-1 + \frac{k_i}{\theta} - 0) = 0$$

$$-N + \frac{\sum_i^N k_i}{\theta} = 0$$

MLE : Poisson example

$$LogL(\theta) = \sum_{i=1}^{N}(-\theta + k_i \; log\theta - log \; k_i!)$$

$$\frac{d}{d\theta}log \; L(\theta) = 0 \;\Rightarrow \sum_{i=1}^{N}(-1 + \frac{k_i}{\theta} - 0) = 0$$

$$-N + \frac{\sum_i^N k_i}{\theta} = 0$$

$$\hat{\theta} = \frac{\sum_i^N k_i}{N}$$

The MLE of λ

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution?

  A. Yes B. No

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- What should be the likelihood function? Is the method of modeling the same as for the Poisson distribution? **Yes and No**. The idea is similar but the normal distribution is continuous, we need to use the **probability density** instead.

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed

- The likelihood function of a normal distribution:

$$L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(x_i - \mu)^2}{2\sigma^2})$$

$\theta_1 = \mu$ $\theta_2 = \sigma$

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- There are two parameters to estimate: $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$
  - If we fix $\boldsymbol{\sigma}$ and set θ= $\boldsymbol{\mu}$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - If we fix $\boldsymbol{\mu}$ and set θ= $\boldsymbol{\sigma}$

$$\hat{\theta} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

review at home

Confidence intervals for MLE estimates

- An MLE parameter estimate $\hat{\theta}$ depends on the data that was observed

- We can construct a confidence interval for $\hat{\theta}$ using the parametric bootstrap
  - Use the distribution with parameter $\hat{\theta}$ to generate a large number of bootstrap samples
  - From each "synthetic" dataset, re-estimate the parameter using MLE
  - Use the histogram of these re-estimates to construct a confidence interval

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

A. 6

B. 35/2

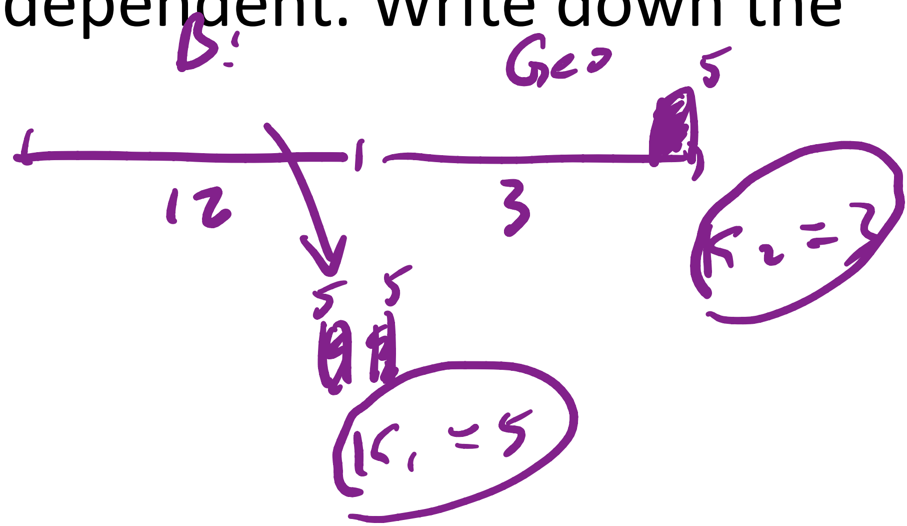
C. 12

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. All rolls are independent. Write down the likelihood function L(θ).

$P(D|\theta)$



MLE Example

$$L(\theta) = P(D|\theta) = P(D_1|\theta)\,P(D_2|\theta)$$

$$= \binom{N}{K_1} \theta^{K_1} (1-\theta)^{N-K_1} (1-\theta)^{K_2-1} \theta$$

$N = 12 \quad K_1 = 2 \quad K_2 = 3$

$$L(\theta) = \binom{12}{2} \theta^3 (1-\theta)^{12}$$

$$\log L(\theta) = \log c + 3 \log \theta + 12 \log(1-\theta)$$

$$\frac{d \log L}{d\theta} = 0 + \frac{3}{\theta} - \frac{12}{1-\theta} = 0$$

$$\hat{\theta} = \frac{3}{15} = \frac{1}{5}$$

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

$$L(\theta)$$

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

$$P(D) = \sum P(D \mid \theta_i)\, P(\theta_i)$$

The components of Bayesian Inference

- From Bayes rule

$$P(\theta \mid D) = \frac{P(D \mid \theta)\, P(\theta)}{P(D)}$$

  - 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=?

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \qquad P(\theta) = \begin{cases} \frac{2}{3} & \theta = 0.5 \\ \frac{1}{3} & \theta = 0.6 \\ 0 & \text{other} \end{cases}$$

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

if $\theta = 0.6$

$$P(D) = \sum P(D|\theta_i)\cdot P(\theta_i)$$

$$= \binom{10}{7}0.5^7 0.5^3 \cdot \frac{2}{3} + \binom{10}{7}0.6^7 0.4^3 \cdot \frac{1}{3}$$

MLE $\hat{\theta} = 0.7$

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

MAP

which θ maximize $P(\theta|D)$: $\hat{\theta} = 0.5$

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 $\hat{\theta}$ **=0.5**

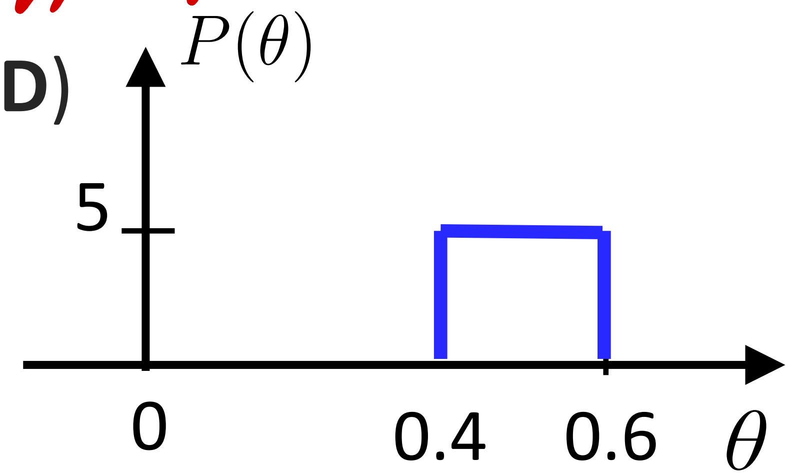
Biased by the prior

Bayesian Inference: a continuous prior

- Suppose we have a coin of unknown probability θ of heads

$P(D) = \int P(D|\theta) P(\theta)\, d\theta$

- We see 7 heads in 10 tosses (**D**)



- We assume

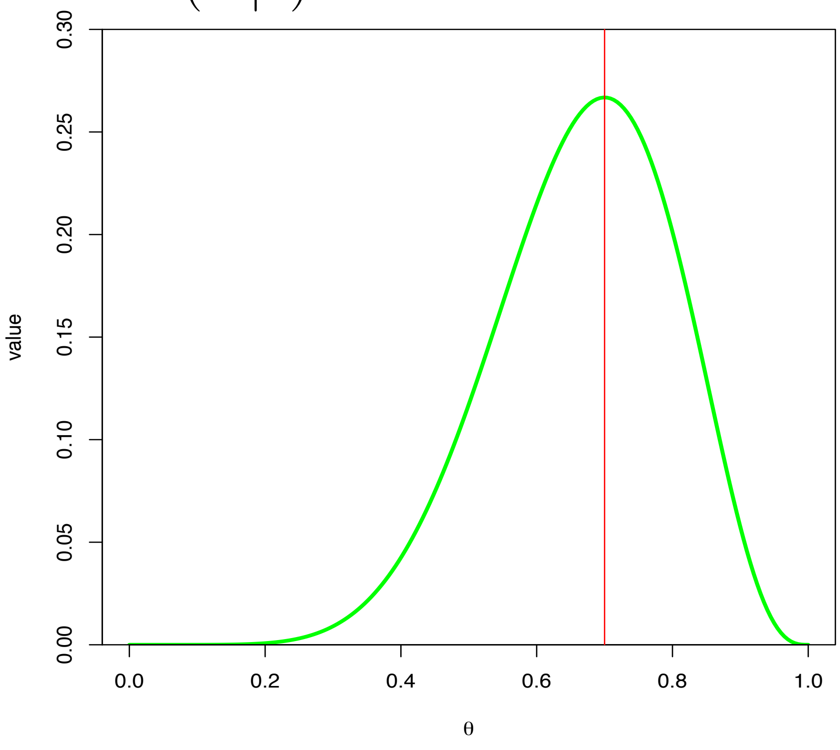
$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

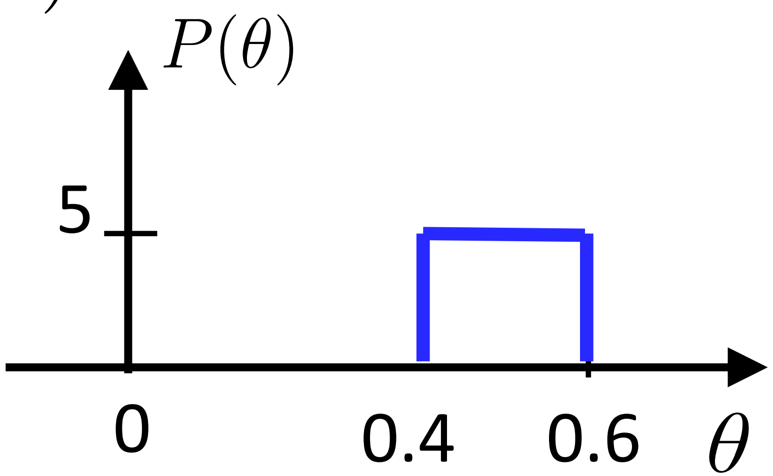
- What is the posterior $P(\theta|D)$?

Bayesian Inference: a continuous prior

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

$P(D|\theta)$ = Likelihood





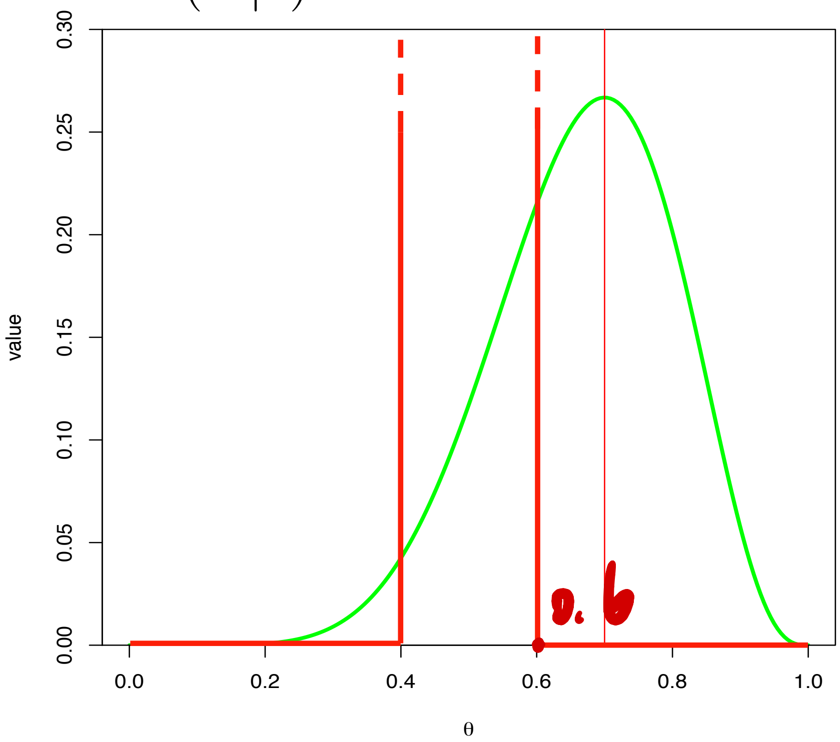
$$P(\theta) = \begin{cases} 5 & if\ \theta \in [0.4, 0.6] \\ 0 & if\ \theta \notin [0.4, 0.6] \end{cases}$$

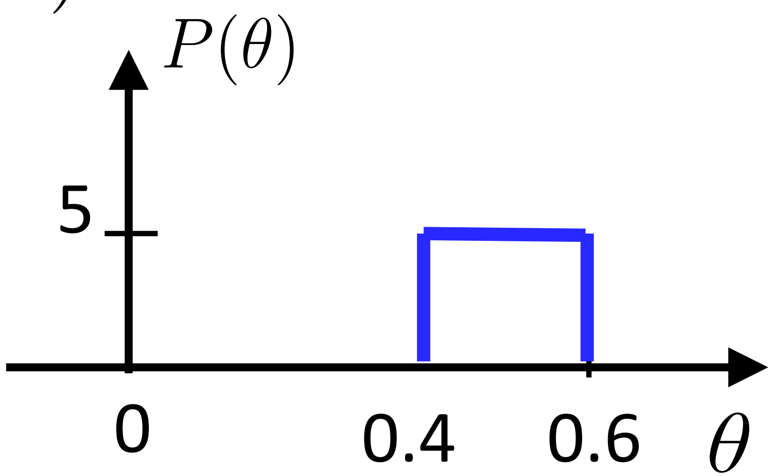
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

Bayesian Inference: a continuous prior

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

$P(D|\theta)$ = Likelihood





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

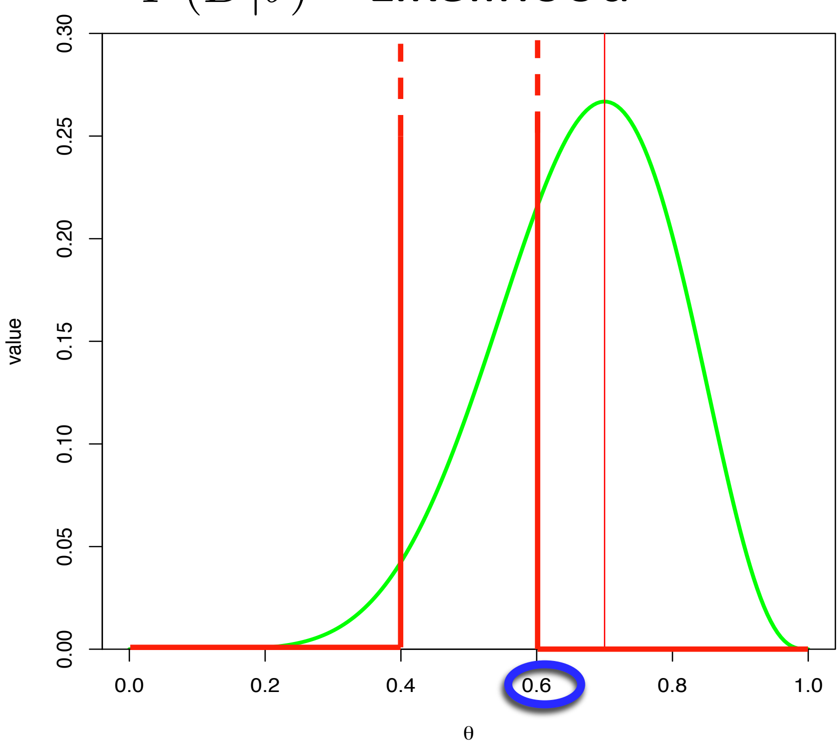
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

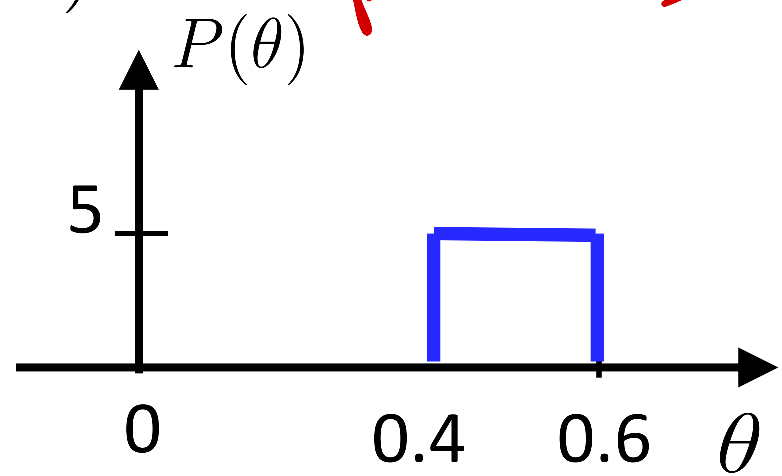
Bayesian Inference: a continuous prior

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

$P(\theta|D)$

$P(D|\theta)$ = Likelihood





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

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

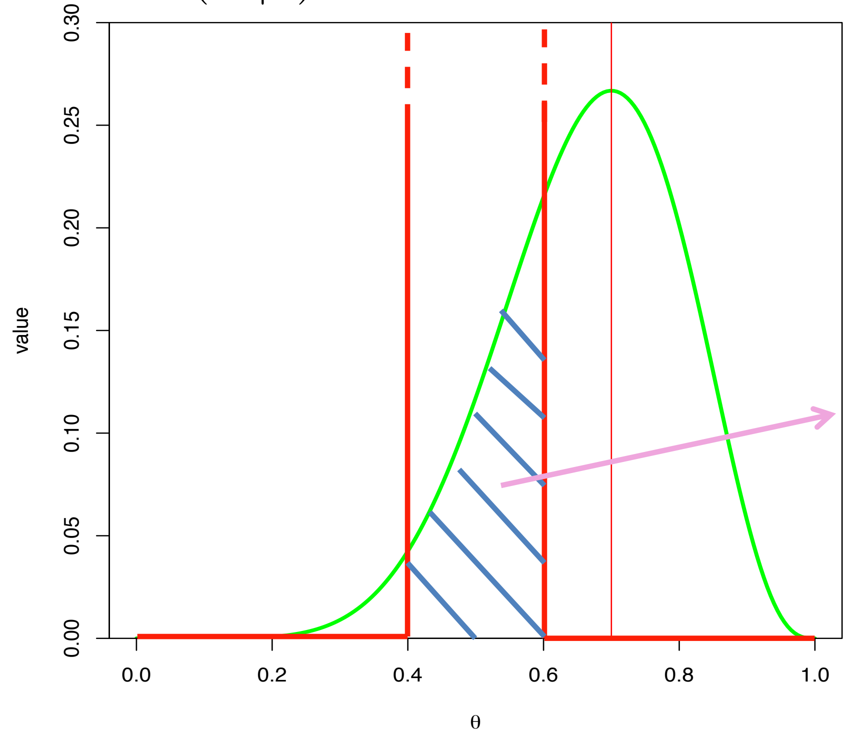
MAP $\hat{\theta}$ **=0.6**

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.

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

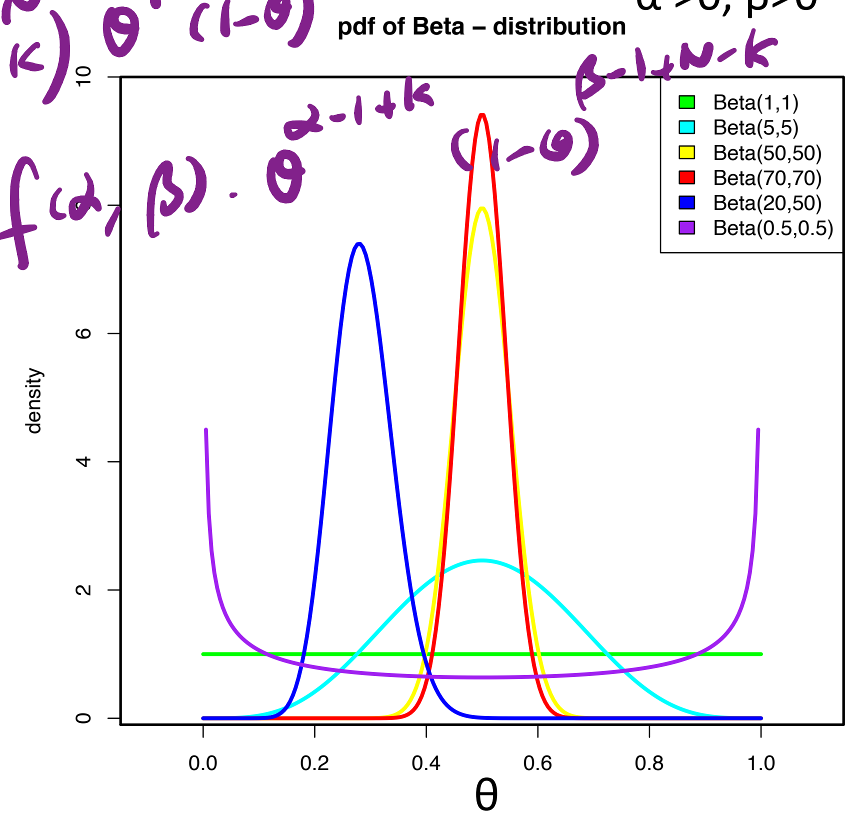
- Is an expressive family of distributions
- $Beta(\alpha = 1, \beta = 1)$ is uniform

0≤ Θ ≤1

α >0, β>0

$\binom{N}{K}\theta^{K}(1-\theta)^{N-K}$

$f(\alpha,\beta)\cdot\theta^{\alpha-1+k}(1-\theta)^{\beta-1+N-K}$



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!