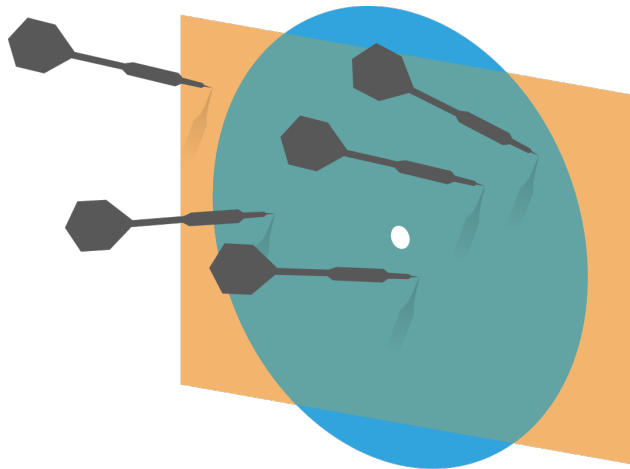


Probability and Statistics for Computer Science



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

Credit: wikipedia

- * Upon entry speakers of the students are muted for the quality of sound in Zoom room.
- * Please "raise up hand" to speak, the audio will be unmuted for you.
- * You can use "chat" to write private note to the instructor
- * Can you see the poll question? Take it if you can. Check piazza post #360
- * Don't share your screen during this lecture.

Last time

- ✱ Review of statistical inference
- ✱ Inferring probability model from data
- ✱ Maximum likelihood estimate (MLE)
- ✱ Confidence interval for MLE

Contents

- ✱ Review of Maximum likelihood Estimation (MLE)
- ✱ Bayesian Inference

Maximum likelihood estimation (MLE)

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

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

is probability
function

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

θ can be
a vector of
parameters

- ✱ The maximum likelihood estimate (MLE) of

θ is
why do we
do this?

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

$\hat{\theta}$ maximizes
 $L(\theta)$

Likelihood function: binomial example

- Suppose we have a coin with unknown probability of θ coming up heads

Model: binomial

- We toss it 10 times and observe 7 heads

- The likelihood function is:

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

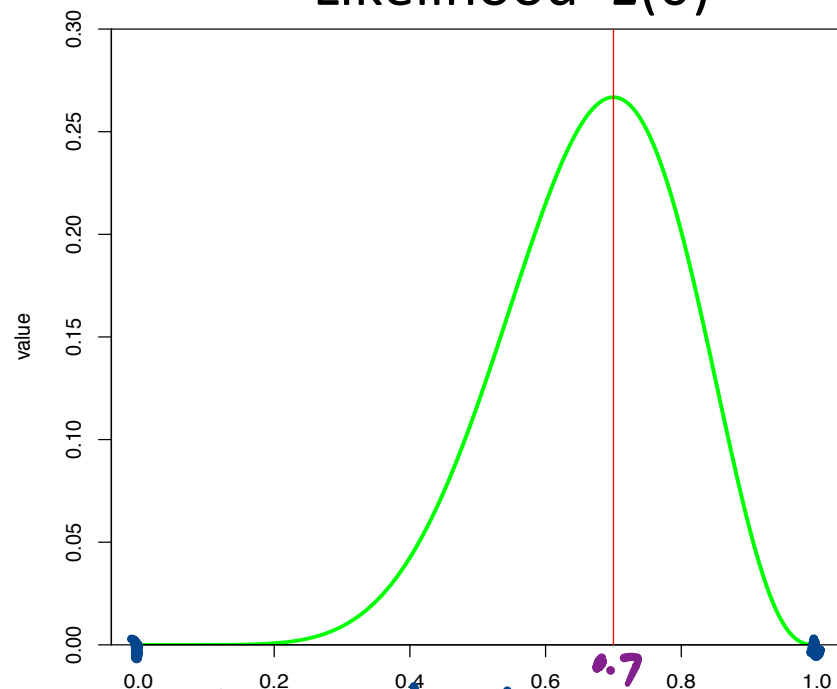
- The MLE is

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

$$\frac{dL(\theta)}{d\theta} = 0$$

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

Likelihood $L(\theta)$



generic soln. $\frac{k}{N} \theta = \frac{7}{10}$

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

A. $12!/7!/5!$

☒ B. $7/12$

C. $5/12$

D. $12/7$

$$\hat{\theta} = \frac{k}{n} \quad \begin{array}{l} \text{\# head} \\ \text{\# rolls} \end{array}$$
$$= \frac{7}{12}$$

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

A. $C(12,7)$

B. $7/12$

C. $5/12$

D. $12/7$

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

A. 7

B. $1/7$

C. other

MLE of Geometric
with $k \sim D$

$$\hat{\theta} = \frac{1}{k}$$

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

A. 7

B. $1/7$

C. other

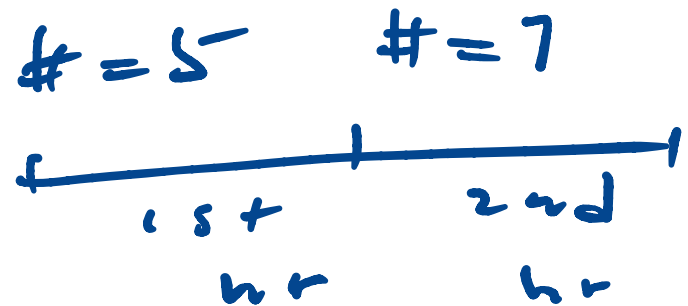
Q. What is the MLE of Poisson $k_1=5$, $k_2=7$, $n=2$

A. 6

B. $35/2$

C. 12

D. other



$$\text{MLE } \hat{\theta} = \frac{\sum k_i}{N}$$
$$\hat{\theta} = \frac{5+7}{2}$$

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 $L(\theta)$.

$L_1(\theta)$ Exp-1 12 times to check # of "5" Bino.

$L_2(\theta)$ Exp-2 ... 1st "5" Geom.

$$L(\theta) = \underline{L_1(\theta)} \cdot \underline{L_2(\theta)}$$

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 $L(\theta)$.

Exp-1 $\underline{N_1 = 12}$, $\underline{K = ?}^2$ } $N_1 + N_2 = 15$
 Exp-2 ... until a "5" $\underline{K=1}$ } $\Rightarrow \underline{N_2 = 3}$

\underline{K} (the "5"s in Exp-1) = $3 - 1 = 2$

\rightarrow Binomial for Exp-1 $\underline{L_1(\theta) = \binom{12}{2} \theta^2 (1-\theta)^{10}}$

Geometric for Exp-2 $L_2(\theta) = \underline{(1-\theta)^2 \theta}$
 $\rightarrow L(\theta) = L_1(\theta) L_2(\theta)$

$$\rightarrow L(\theta) = C \theta^3 (1-\theta)^{12}$$

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

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

$$\frac{3}{\theta} = \frac{12}{1-\theta}$$

$$12\theta = 3 - 3\theta$$

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

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 $p(\text{heads}) = 0.3$?
 - ✱ If we observe 0 heads in 2 coin tosses, should we accept that $p(\text{heads}) = 0$?

Bayesian inference

- ✱ In MLE, we maximized the likelihood function

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

why? ↓

↑
range of θ

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

$$P(\theta|D)$$

$P(\theta|D)$
A.I. (I) →
↑
 θ

- ✱ 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 $\hat{\theta}$

How are these two related?

The components of Bayesian Inference

- From Bayes rule $\overset{L(\theta)}{\uparrow}$

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

$\overset{P(\theta|D)}{\uparrow}$ $\overset{P(D|\theta)}{\uparrow}$ $\overset{P(\theta)}{\uparrow}$

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

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

- ✱ 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 $P(\text{heads}) = 0$
 - ✱ 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

$\theta \in [0, 1]$

- ✱ We see 7 heads in 10 tosses (D)

- ✱ We assume the prior about θ .

discrete prior

- ✱ We have this likelihood:

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

Binomial Model

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


- ✱ 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 .
$$P(\theta) = \begin{cases} \frac{2}{3} & \text{if } \theta = 0.5 \\ \frac{1}{3} & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$
- ✱ 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 .
$$P(\theta) = \begin{cases} \frac{2}{3} & \text{if } \theta = 0.5 \\ \frac{1}{3} & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$
- ✱ We have this likelihood:

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

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

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

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

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

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

$$\begin{aligned} P(D) &= \sum P(D|\theta_i) \cdot P(\theta_i) \\ &= \underbrace{\binom{10}{7} 0.5^7 \cdot 0.5^3}_{\theta=0.5} \cdot \underbrace{\frac{2}{3}}_{\theta=0.5} + \underbrace{\binom{10}{7} 0.6^7 \cdot 0.4^3}_{\theta=0.6} \cdot \underbrace{\frac{1}{3}}_{\theta=0.6} \end{aligned}$$

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

which θ maximize $P(\theta|D)$:

Bayesian Inference: a discrete prior

✱ We see 7 heads in 10 tosses (**D**)

✱ We assume the prior about θ .
$$P(\theta) = \begin{cases} \frac{2}{3} & \text{if } \theta = 0.5 \\ \frac{1}{3} & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$

✱ We have this likelihood:

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

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

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

MAP estimate=?

Bayesian Inference: a discrete prior

✱ We see 7 heads in 10 tosses (**D**)

✱ We assume the prior about θ .
$$P(\theta) = \begin{cases} \frac{2}{3} & \text{if } \theta = 0.5 \\ \frac{1}{3} & \text{if } \theta = 0.6 \\ 0 & \text{otherwise} \end{cases}$$

✱ We have this likelihood:

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

MLE $\hat{\theta} = 0.7$

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

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

MAP $\hat{\theta} = 0.5$

Biased by the prior

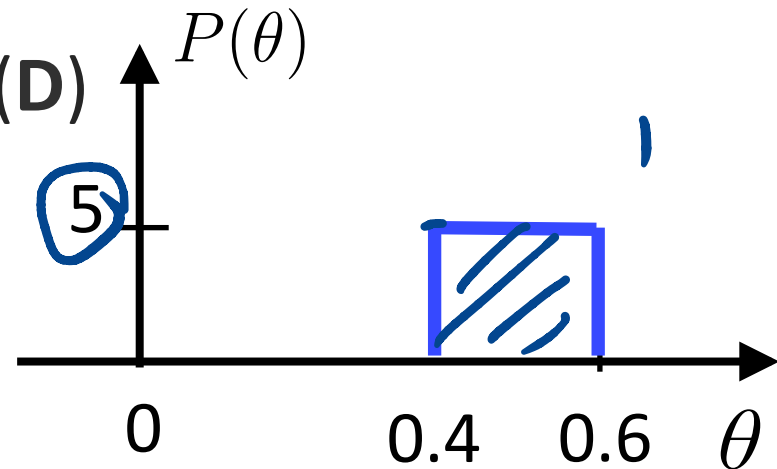
Bayesian Inference: a continuous prior

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

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

- ✱ We assume

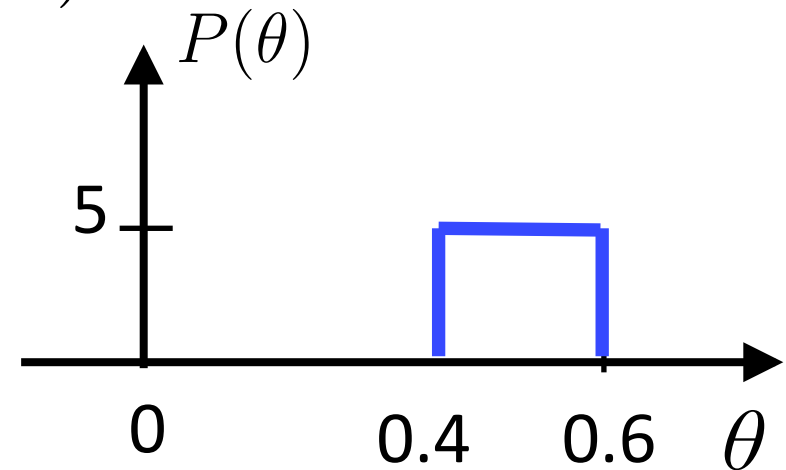
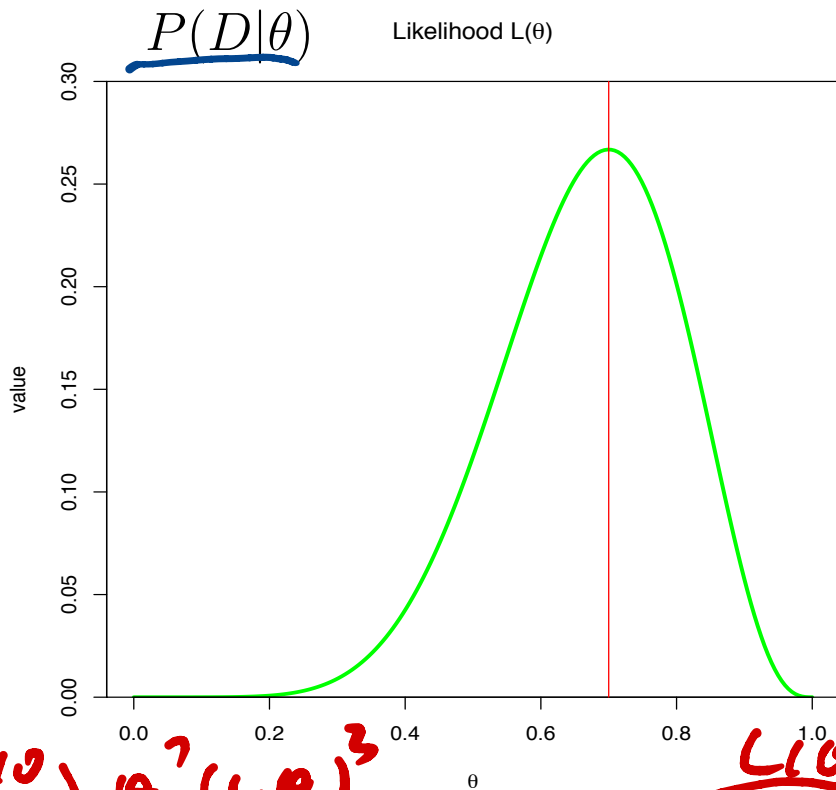
$$P(\theta) = \begin{cases} 5 & \text{if } \theta \in [0.4, 0.6] \\ 0 & \text{if } \theta \notin [0.4, 0.6] \end{cases}$$



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

Bayesian Inference: a continuous prior

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



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

$(\frac{10}{7}) \theta^7 (1-\theta)^3$

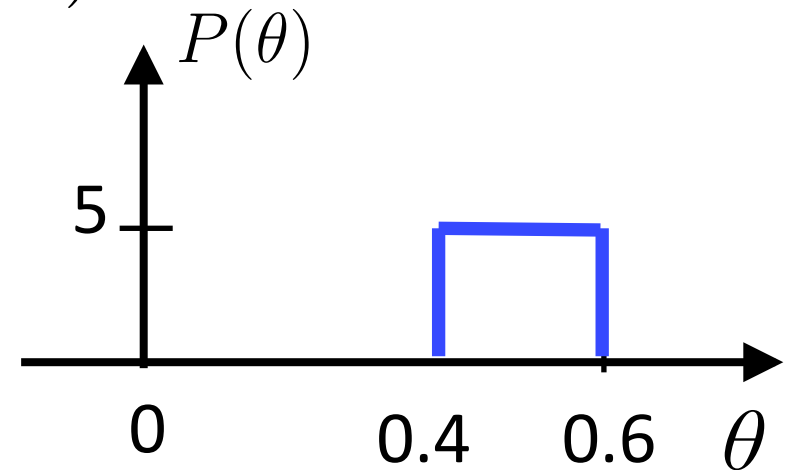
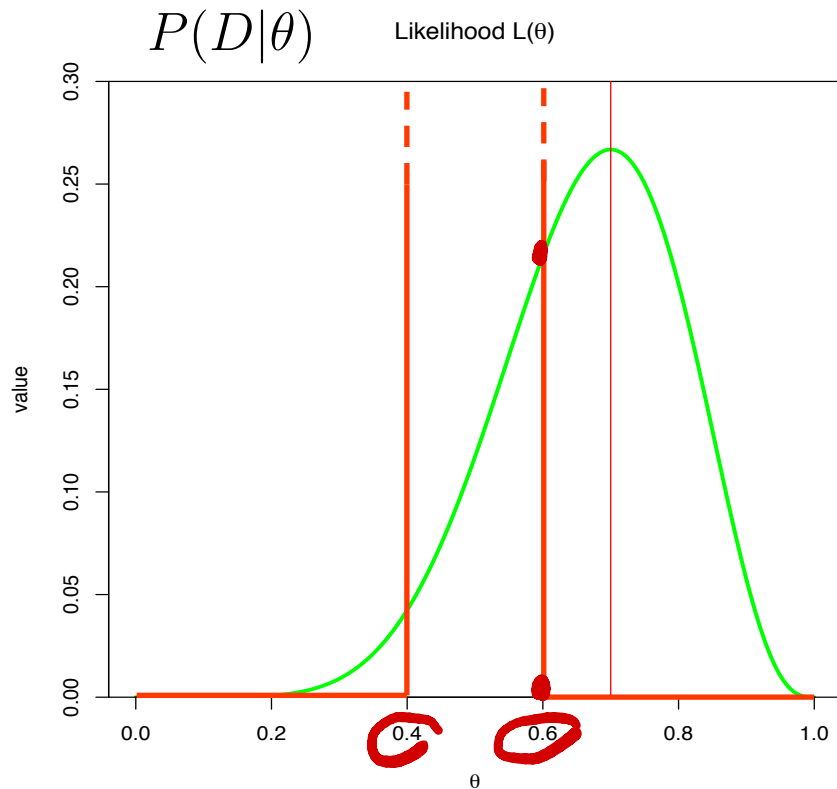
$$\underline{P(\theta|D)} \propto \underbrace{L(\theta)}_{P(D|\theta)} \underbrace{P(\theta)}$$

$\rightarrow \hat{\theta}$? some
 $P(D)$ is const.

Bayesian Inference: a continuous prior



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



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

$\hat{\theta} = ?$

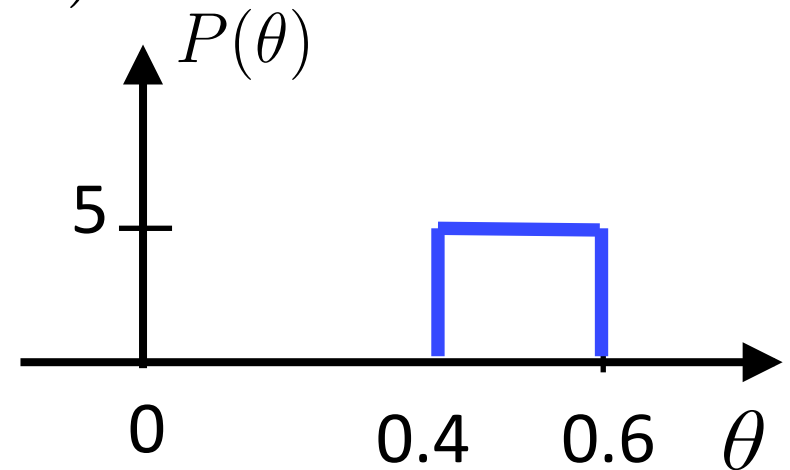
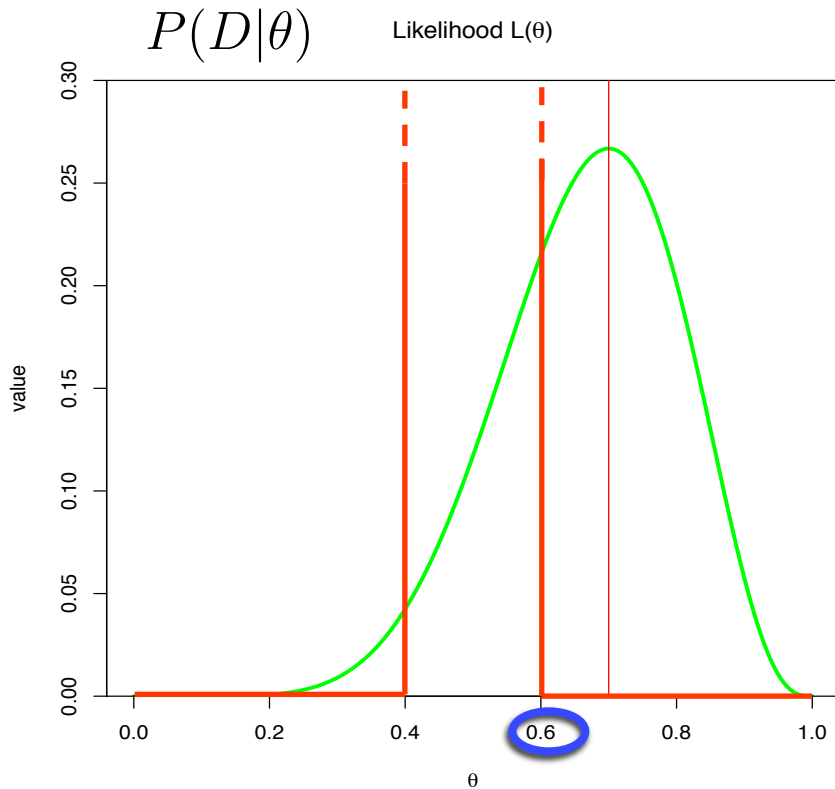
0.6

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

Bayesian Inference: a continuous prior



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



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

better than

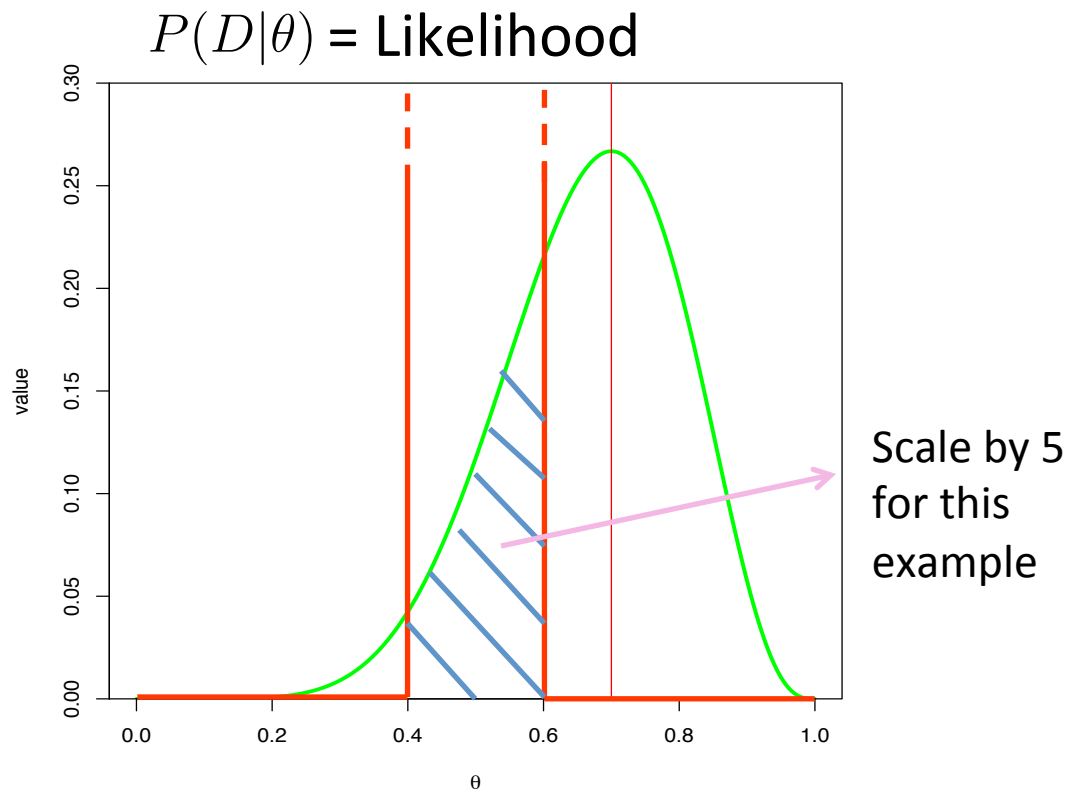
$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

MAP $\hat{\theta} = 0.6$ *0.5*

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.



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:

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

\uparrow \uparrow \uparrow \uparrow
expressive!

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

pdf of Beta - distribution

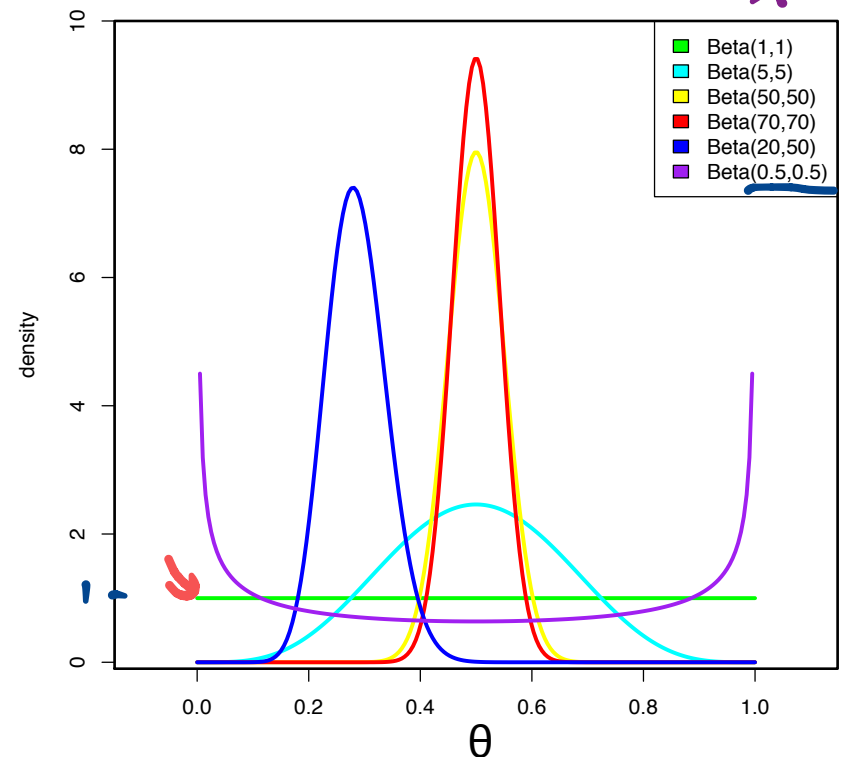
$\theta \in [0, 1]$
 $\alpha > 0, \beta > 0$

- Is an expressive family of distributions

$$K(1, 1) = 1$$

- Beta($\alpha = 1, \beta = 1$) is uniform

cf. $\rightarrow \binom{n}{k} \theta^k (1-\theta)^{n-k}$



Q. Beta distribution is a continuous probability distribution

A. TRUE

B. FALSE

Q. Is this true?

$$\int_{-\infty}^{\infty} K(\alpha, \beta) \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = 1$$

A. YES

B. No

$$\int_0^1 K(\alpha, \beta) \theta^{\alpha-1} (1-\theta)^{\beta-1} d\theta = 1$$

Beta distribution as the conjugate prior for Binomial likelihood

- * The likelihood is Binomial (N, k)

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

$N, k \rightarrow \text{integer}$

$\alpha, \beta \rightarrow \text{real}$

- * The Beta distribution is used as the prior

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

k $N-k$
 θ $(1-\theta)$

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

C_1 C_2
 $(\theta)^{\alpha+k-1} (1-\theta)^{\beta+N-k-1}$

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

$\alpha' = \alpha + k$ $\beta' = \beta + N - k$

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

? $P(D)$

$$\int_0^1 P(\theta|D) d\theta = 1$$

The update of Bayesian posterior

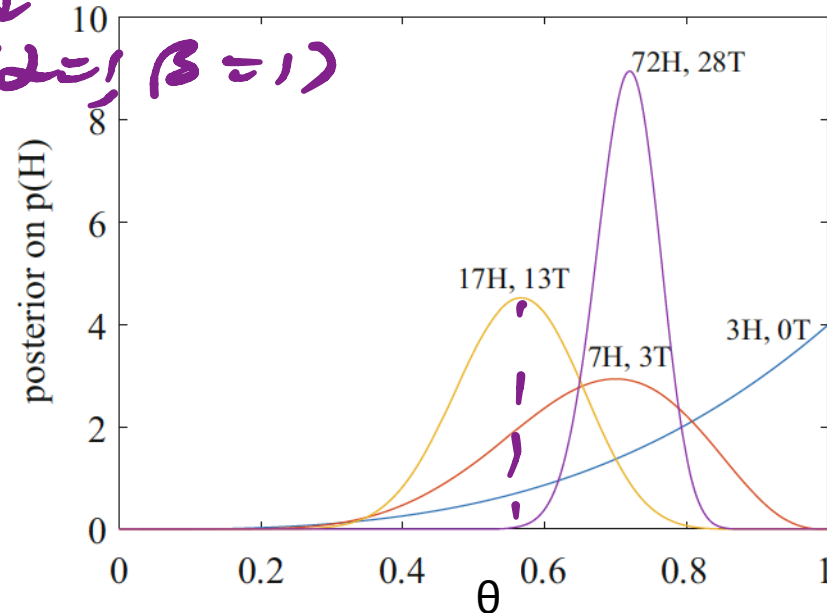
- Since the posterior is in the same family as the conjugate prior, the posterior can be used as a new prior if more data is observed. $P(\theta | D_1) \rightarrow P(\theta^*)$

- Suppose we start with a uniform prior on the probability θ of heads $P(\theta | D_0)$

- 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

$\hat{\theta}$

$\text{Beta}(\alpha=1, \beta=1)$



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.

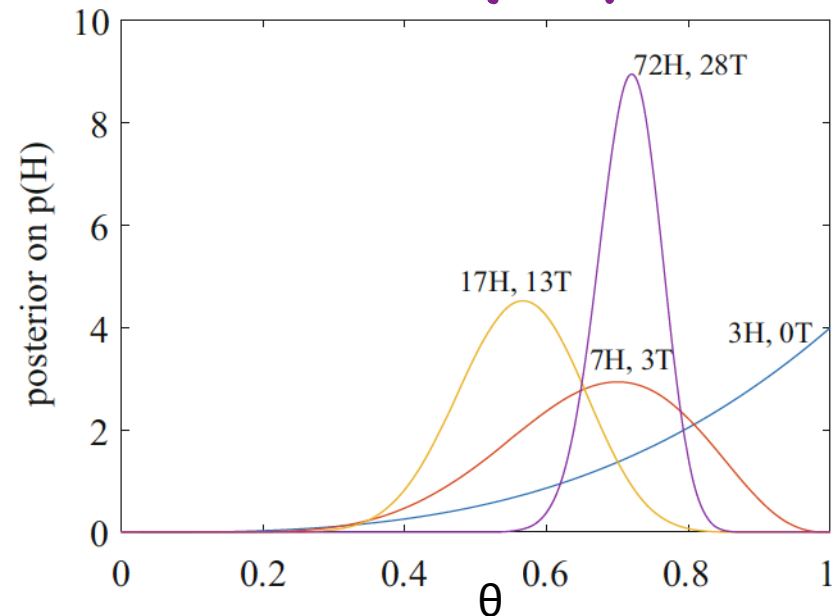
- Suppose we start with a uniform prior on the probability θ of heads

$$\beta \rightarrow \alpha(1, 1)$$


$$\alpha' = \alpha + k$$

$$\beta' = \beta + N - k$$

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

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

$$\frac{dP(\theta|D)}{d\theta} = 0$$

- ✱ Differentiating and setting to 0 gives the MAP estimate

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

$$= \frac{1 - 1 + 96}{1 + 1 - 2 + 143} = 0.67$$

prior
 α, β
 $\alpha = 1$
 $\beta = 1$

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

standard normal

Conjugate prior for other likelihood functions

- ✱ If the likelihood is Bernoulli or geometric, the conjugate prior is Beta $N=1$
- ✱ 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

Gamma Distribution

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} \quad \begin{matrix} \alpha > 0 \\ \beta > 0 \end{matrix}$$

If $\alpha = 1$, it's the exponential distr.

$$\beta e^{-\beta\theta} \text{ - exponential}$$

Poisson

$$\frac{e^{-\lambda} \lambda^k}{k!}$$

Assignments

- ✱ Finish Chapter 9 of the textbook
- ✱ Next time: PCA

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!*



A spinner example for MLE

- ✱ If I have a spinner that has been divided into three sections, 1, 2 and 3. The probability of the spin ending in these sections are p^2 , $2p(1-p)$, $(1-p)^2$ respectively. Suppose I spin N times and find it n_1 times in section 1, n_2 times in section 2, and n_3 times in section 3. What is the MLE estimate of \mathbf{p} ?

A spinner example

- ✱ Find the Likelihood function: The probability of seeing the data given $\theta = p$;

This is a multinomial distribution, $k=3$, $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$, $p_3 = (1-\theta)^2$

$$P(X_1 = n_1, X_2 = n_2, \dots, X_k = n_k) = \frac{N!}{n_1!n_2!\dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

where $N = n_1 + n_2 + \dots + n_k$

A spinner example

- ✱ Find the Likelihood function: The probability of seeing the data given $\theta = p$;

This is a multinomial distribution, $k=3$, $p_1 = \theta^2$, $p_2=2\theta(1-\theta)$, $p_3 = (1-\theta)^2$

$$P(D|\theta) = \frac{N!}{n_1!n_2!n_3!}(\theta^2)^{n_1}(2\theta(1-\theta))^{n_2}((1-\theta)^2)^{n_3}$$

$$L(\theta) = P(D|\theta) = \frac{n!2^{n_2}}{n_1!n_2!n_3!}\theta^{2n_1+n_2}(1-\theta)^{n_2+2n_3}$$

$$\text{Log}L(\theta) = \log C + (2n_1 + n_2)\log\theta + (n_2 + 2n_3)\log(1 - \theta)$$

A spinner example

- ✱ Find the Likelihood function: The probability of seeing the data given $\theta = p$;

This is a multinomial distribution, $k=3$, $p_1 = \theta^2$, $p_2 = 2\theta(1-\theta)$, $p_3 = (1-\theta)^2$

$$\text{Log}L(\theta) = \log C + (2n_1 + n_2)\log\theta + (n_2 + 2n_3)\log(1 - \theta)$$

$$\frac{d}{d\theta}L(\theta) = \frac{2n_1 + n_2}{\theta} - \frac{n_2 + 2n_3}{1 - \theta} = 0$$

$$\hat{\theta} = \frac{2n_1 + n_2}{2n_1 + 2n_2 + 2n_3} = \frac{2n_1 + n_2}{2N}$$

$$L(\theta) = C (1-\theta)^{x_1} \theta (1-\theta)^{x_2} \cdot \theta$$

$$\frac{dL_{\theta}}{d\theta} = (\log C + (x_1 + x_2) \log(1-\theta) + 2 \log \theta)' = 0$$

$$- (x_1 + x_2) \frac{1}{1-\theta} + 2 \frac{1}{\theta} = 0$$

$$\frac{x_1 + x_2}{1-\theta} = \frac{2}{\theta}$$

$$(x_1 + x_2)\theta = 2 - 2\theta$$

$$\theta = \frac{2}{x_1 + x_2 + 2} = \frac{1}{\frac{(x_1 + x_2)}{2} + 1}$$