Will Monroe
August 9, 2017

with materials by
Mehran Sahami
and Chris Piech

image: Aritio

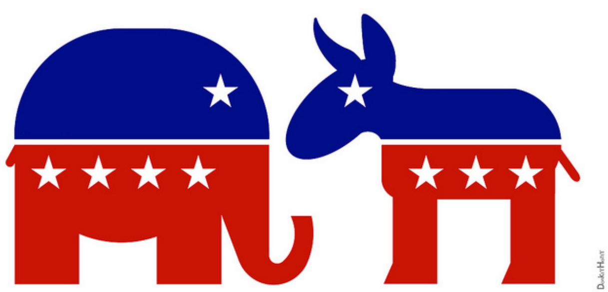
Parameter learning

Announcement: Problem Set #6

Goes out tonight.

Due the last day of class,
Wednesday, August 16
(before class).

Some serious coding!



Congressional voting

Heart disease diagnosis

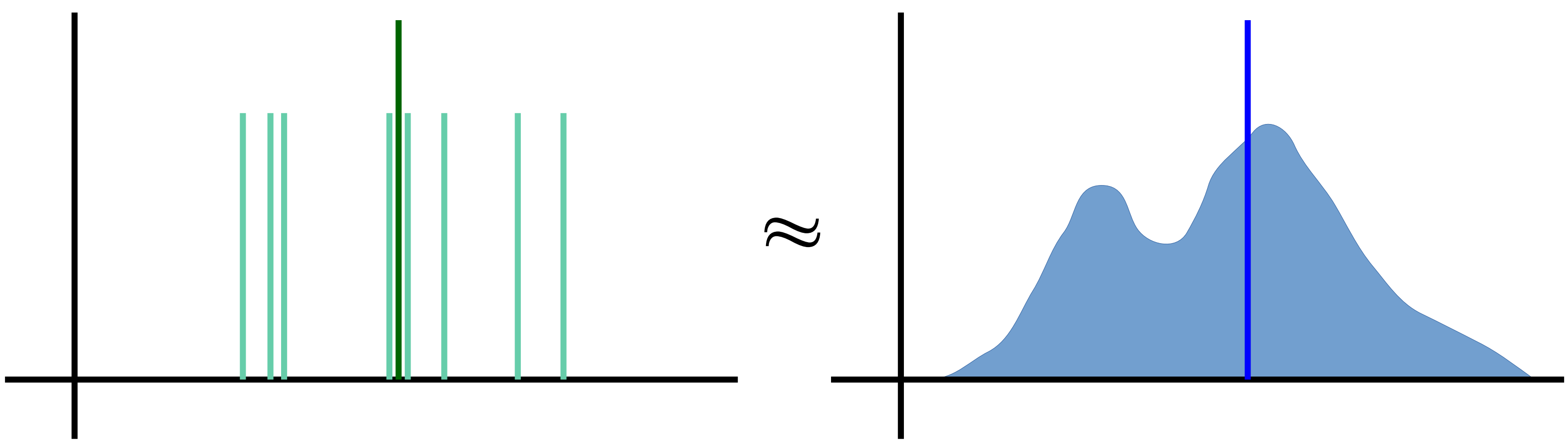
No late days!

Review: Parameter estimation

Sometimes we **don't know** things like the expectation and variance of a distribution; we have to **estimate** them from incomplete information.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$\hat{\theta} = \arg\max_{\theta} \mathrm{LL}(\theta)$$



Review: Central limit theorem

Sums and **averages** of IID random variables are **normally distributed**.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

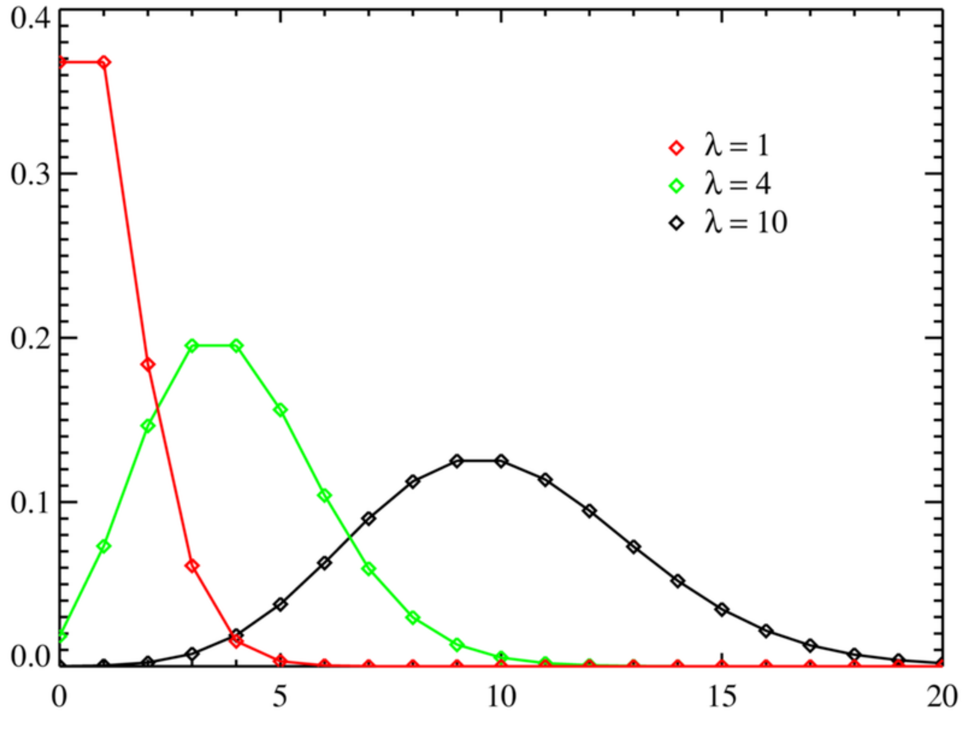
$$Y = n\bar{X} = \sum_{i=1}^{n} X_i \sim N\left(n\mu, n\sigma^2\right)$$

Easily-confused principles

Constant multiple of a normal	Sum of identical normals	CLT
$X \sim N(\mu, \sigma^2)$	$X_i \sim N(\mu, \sigma^2)$ (independent & identical)	$X_i \sim ???$ (independent & identical)
↓	↓	↓
$nX \sim N(n\mu, \boxed{n^2}\sigma^2)$	$\sum_{i=1}^{n} X_i \sim N(n\mu, \boxed{n}\sigma^2)$ (exactly)	$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$ (approximately, for large *n*)

Central limit theorem demo

Review: Approximating a Poisson with a normal



$$X \sim \text{Poi}(\lambda)$$

$$\approx$$

$$Y \sim N(\lambda, \lambda)$$

(for large λ)

Parameters

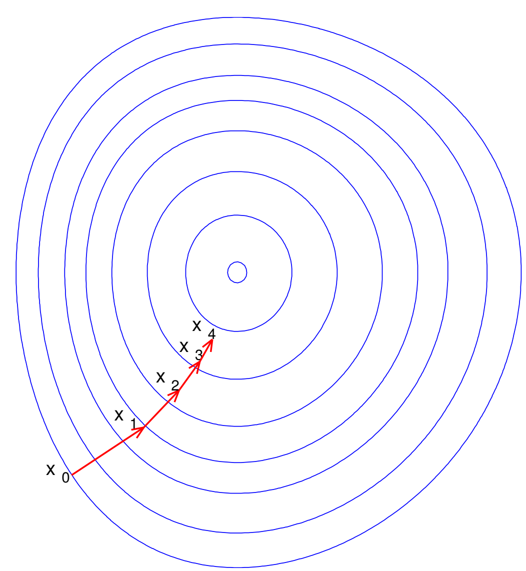
$$\theta$$

$$X \sim \begin{cases} \text{Ber}(p) & \theta = p \\ \text{Poi}(\lambda) & \theta = \lambda \\ \text{Uni}(a, b) & \theta = [a, b] \\ \text{N}(\mu, \sigma^2) & \theta = [\mu, \sigma^2] \end{cases}$$

Maximum likelihood estimation

Choose parameters that **maximize** the likelihood (**joint probability given parameters**) of the example data.

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



How to: MLE

1. Compute the likelihood.

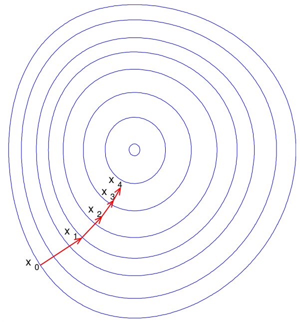
$$L(\theta) = P(X_1, \ldots, X_m | \theta)$$

2. Take its log.

$$LL(\theta) = \log L(\theta)$$

3. Maximize this as a function of the parameters.

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



Maximum likelihood for Bernoulli

The maximum likelihood *p* for **Bernoulli** random variables is the **sample mean**.

$$\hat{p} = \frac{1}{m} \sum_{i=1}^{m} X_i$$

Derivation: MLE for Bernoulli

1. Compute the likelihood.

$$\theta = p$$

$$L(\theta) = P(X_1, \ldots, X_m | \theta)$$

$$= \prod_{i=1}^{m} P(X_i | \theta)$$

don't forget: IID means independent!

$$= \prod_{i=1}^{m} \begin{cases} p & \text{if } X_i = 1 \\ (1 - p) & \text{if } X_i = 0 \end{cases}$$

Derivation: MLE for Bernoulli

1. Compute the likelihood.

$$\theta = p$$

$$L(\theta) = P(X_1, \ldots, X_m | \theta)$$

$$= \prod_{i=1}^{m} P(X_i | \theta)$$ don't forget: IID means independent!

$$= \prod_{i=1}^{m} \begin{cases} p & \text{if } X_i = 1 \\ (1-p) & \text{if } X_i = 0 \end{cases}$$

$$= \prod_{i=1}^{m} p^{X_i} (1-p)^{1-X_i}$$

Derivation: MLE for Bernoulli

2. Take its log.

$$\theta = p$$

$$L(\theta) = \prod_{i=1}^{m} \theta^{X_i}(1-\theta)^{1-X_i}$$

$$LL(\theta) = \log \prod_{i=1}^{m} \theta^{X_i}(1-\theta)^{1-X_i}$$

$$= \sum_{i=1}^{m} \log\left[\theta^{X_i}(1-\theta)^{1-X_i}\right]$$

$$= \sum_{i=1}^{m} \left[X_i \log \theta + (1-X_i)\log(1-\theta)\right]$$

Derivation: MLE for Bernoulli

3. Maximize this as a function of the parameters.

$$\theta = p$$

$$LL(\theta) = \sum_{i=1}^{m} \left[X_i \log \theta + (1 - X_i) \log(1 - \theta) \right]$$

$$\hat{\theta} = \hat{p} = \arg\max_{\theta} LL(\theta)$$

$$\frac{d}{d\theta} LL(\theta) = \sum_{i=1}^{m} \left[\frac{X_i}{\theta} - \frac{1 - X_i}{1 - \theta} \right]$$

$$= \frac{1}{\theta} \sum_{i=1}^{m} X_i - \frac{1}{1-\theta} \sum_{i=1}^{m} (1 - X_i)$$

$$= \left(\frac{1}{\theta} + \frac{1}{1-\theta} \right) \left(\sum_{i=1}^{m} X_i \right) - \frac{m}{1-\theta} = 0$$

$$\frac{1}{\theta} \left(\sum_{i=1}^{m} X_i \right) = m$$

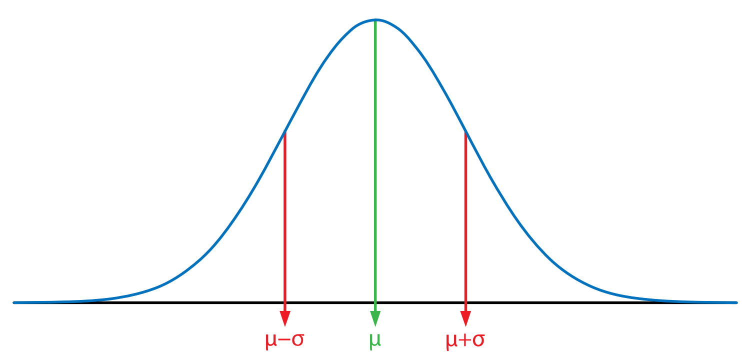
$$\left(\frac{1-\theta}{\theta} + 1 \right) \left(\sum_{i=1}^{m} X_i \right) = m$$

$$\theta = \frac{1}{m} \left(\sum_{i=1}^{m} X_i \right)$$

Maximum likelihood for normal

The maximum likelihood μ for **normal** random variables is the **sample mean**, and the maximum likelihood σ^2 is the "uncorrected" **mean square deviation**.

$$\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} X_i \qquad \hat{\sigma}^2 = \frac{1}{m} \sum_{i=1}^{m} (X_i - \hat{\mu})^2$$



Derivation: MLE for Normal

2. Take its log

$$\theta = [\mu, \sigma^2]$$

$$L(\theta) = \prod_{i=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$LL(\theta) = \sum_{i=1}^{m} \log\left[\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right]$$

$$= \sum_{i=1}^{m} -\log\sigma - \log\sqrt{2\pi} - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

Derivation: MLE for normal

3. Maximize this as a function of the parameters.

$$\theta = [\mu, \sigma^2]$$

$$LL(\theta) = \sum_{i=1}^{m} -\log \sigma - \log \sqrt{2\pi} - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

$$[\hat{\mu}, \hat{\sigma}^2] = \hat{\theta} = \arg\max_{\theta} LL(\theta)$$

$$\frac{\partial}{\partial \mu} LL(\theta) = \sum_{i=1}^{m} -\left(\frac{X_i - \mu}{\sigma}\right)\left(-\frac{1}{\sigma}\right)$$

$$= \sum_{i=1}^{m} \frac{X_i - \mu}{\sigma^2}$$

$$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{m} X_i\right) - \frac{m\mu}{\sigma^2} = 0$$

$$\mu = \frac{1}{m}\left(\sum_{i=1}^{m} X_i\right) = \bar{X}$$

Derivation: MLE for normal

3. Maximize this as a function of the parameters.

$$\theta = [\mu, \sigma^2]$$

$$LL(\theta) = \sum_{i=1}^{m} -\log\sigma - \log\sqrt{2\pi} - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2$$

$$[\hat{\mu}, \hat{\sigma^2}] = \hat{\theta} = \arg\max_{\theta} LL(\theta)$$

$$\frac{\partial}{\partial\sigma} LL(\theta) = \sum_{i=1}^{m}\left[-\frac{1}{\sigma} - \left(\frac{X_i - \mu}{\sigma}\right)\left(-\frac{X_i - \mu}{\sigma^2}\right)\right]$$

$$= \sum_{i=1}^{m}\left[\frac{(X_i - \mu)^2}{\sigma^3} - \frac{1}{\sigma}\right]$$

$$= \frac{1}{\sigma^3}\sum_{i=1}^{m}(X_i - \mu)^2 - \frac{m}{\sigma} = 0$$

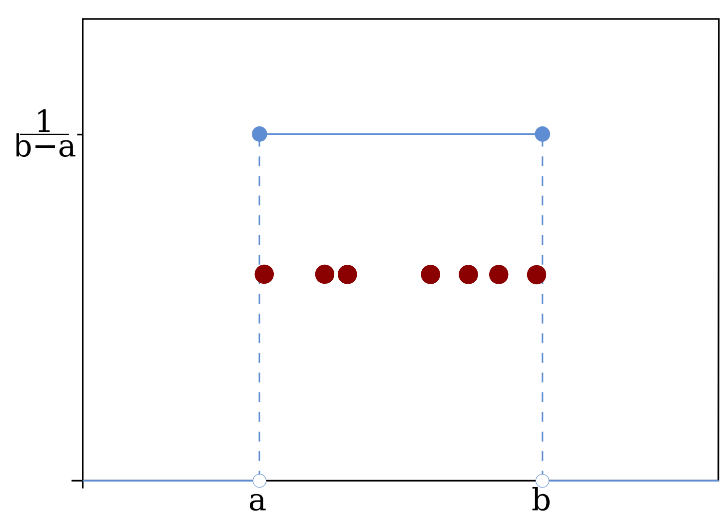
$$\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \mu)^2 = \frac{1}{m}\sum_{i=1}^{m}(X_i - \bar{X})^2$$

Break time!

Maximum likelihood for uniform

The maximum likelihood *a* and *b* for **uniform** random variables are the **minimum and maximum** of the data.

$$\hat{a} = \min_i X_i \qquad \hat{b} = \max_i X_i$$



Derivation: MLE for uniform

1. Compute the likelihood.

$$\theta = [a, b]$$

$$L(\theta) = \prod_{i=1}^{m} \begin{cases} \frac{1}{b-a} & \text{if } a \le X_i \le b \\ 0 & \text{otherwise} \end{cases}$$

2. Take its log.

$$LL(\theta) = \sum_{i=1}^{m} \begin{cases} -\log(b-a) & \text{if } a \le X_i \le b \\ -\infty & \text{otherwise} \end{cases}$$

3. Maximize this as a function of the parameters.

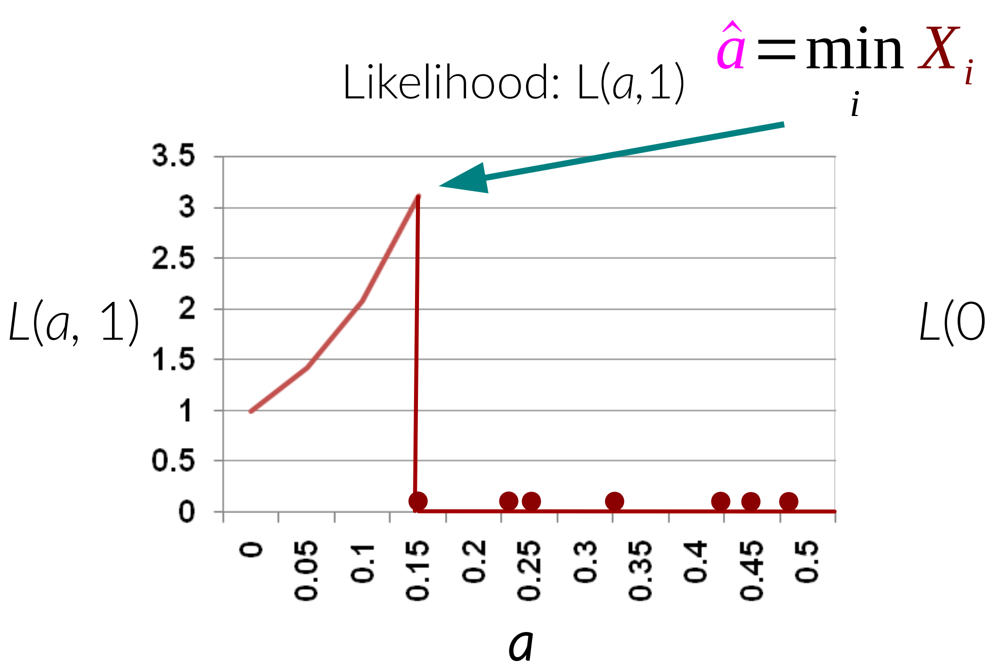
$$[\hat{a}, \hat{b}] = \hat{\theta} = \arg\max_{\theta} LL(\theta)$$

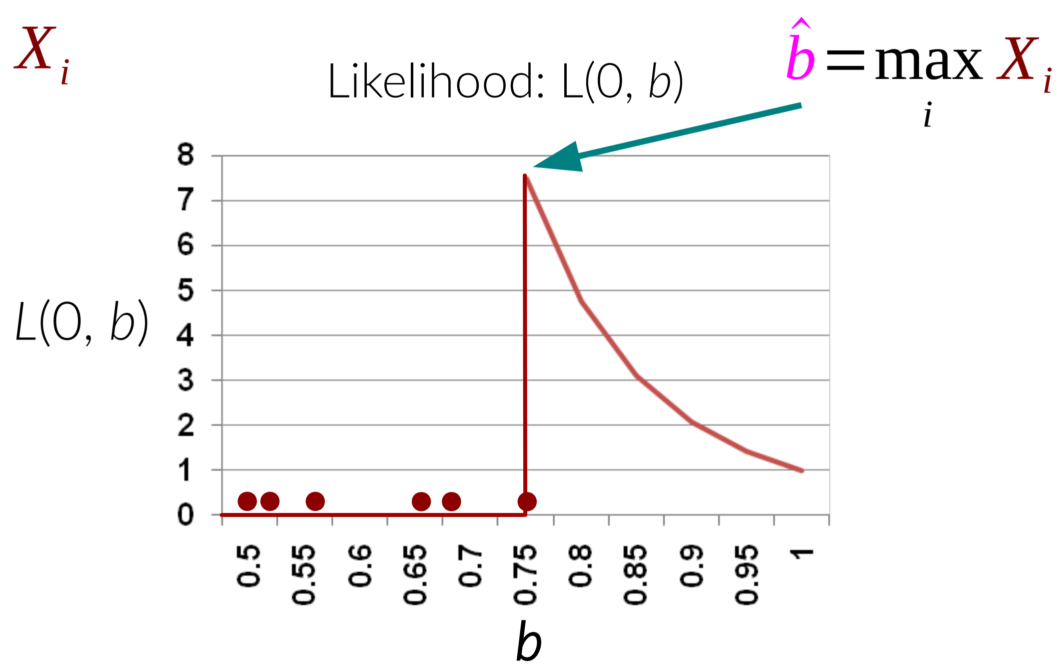
Derivation: MLE for uniform

$$\theta = [a, b]$$

$$L(\theta) = \prod_{i=1}^{m} \begin{cases} \frac{1}{b-a} & \text{if } a \le X_i \le b \\ 0 & \text{otherwise} \end{cases}$$

$$[\hat{a}, \hat{b}] = \hat{\theta} = \arg\max_{\theta} L(\theta)$$





Maximum a posteriori estimation

Choose the **most likely** parameters **given the example data**. You'll need a **prior probability** over the parameters.

$$\hat{\theta} = \arg\max_{\theta} P(\theta | X_1, \ldots, X_n)$$

$$= \arg\max_{\theta} \left[LL(\theta) + \log P(\theta) \right]$$

Review: Multinomial random variable

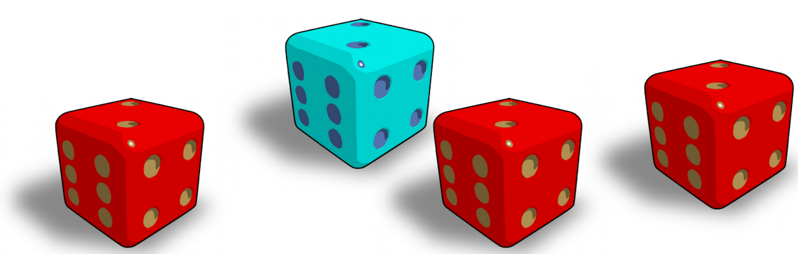
An **multinomial** random variable records the number of times each outcome occurs, when an experiment with multiple outcomes (e.g. die roll) is run multiple times.

$$X_1, \ldots, X_m \sim \mathrm{MN}(n, p_1, p_2, \ldots, p_m)$$

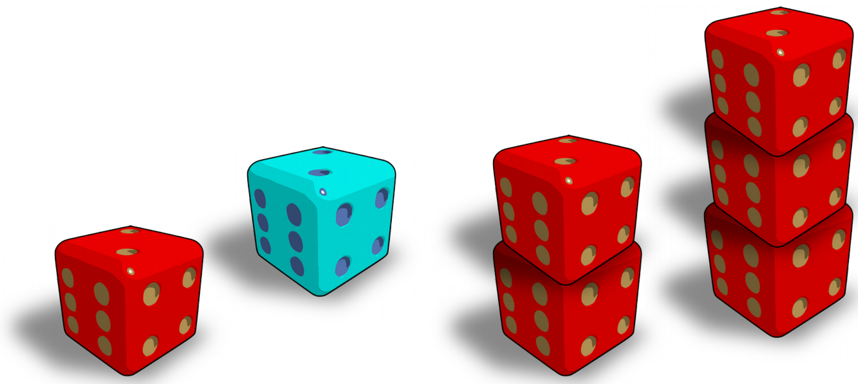
vector!

$$P(X_1 = c_1, X_2 = c_2, \ldots, X_m = c_m)$$

$$= \binom{n}{c_1, c_2, \ldots, c_m} p_1^{c_1} p_2^{c_2} \cdots p_m^{c_m}$$



Roll all of the dice!



A 6-sided die is rolled 7 times.

What is the probability we get:

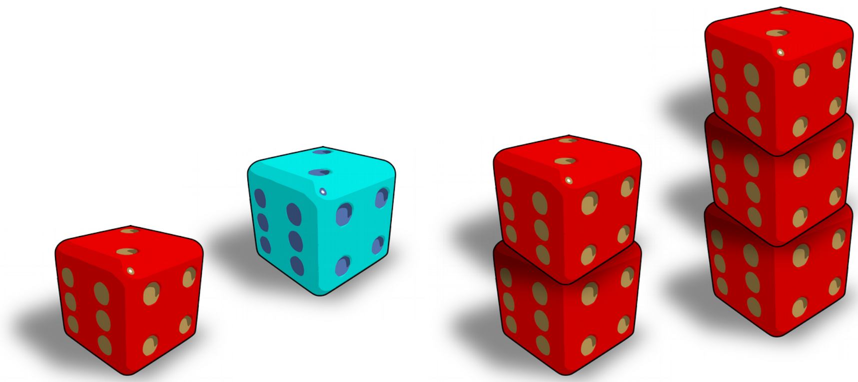
- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes?

$$X_1, \ldots, X_6 \sim \text{MN}\left(7, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$$

$$P\left(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 2, X_5 = 0, X_6 = 3\right)$$

$$= \binom{7}{1,1,0,2,0,3} \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^1 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^3 = 420 \left(\frac{1}{6}\right)^7$$

Maximum likelihood with multinomial



A 6-sided die is rolled 7 times. We get:

- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

What is the MLE for p_1, ..., p_6?

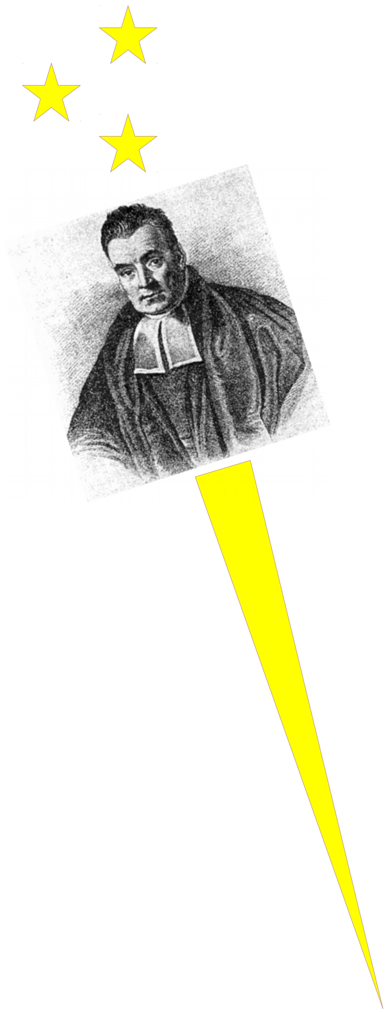
$$X_1, \ldots, X_6 \sim \mathrm{MN}\left(7, \frac{1}{7}, \frac{1}{7}, 0, \frac{2}{7}, 0, \frac{3}{7}\right)$$

you'll never roll a 3!
not in a million years!

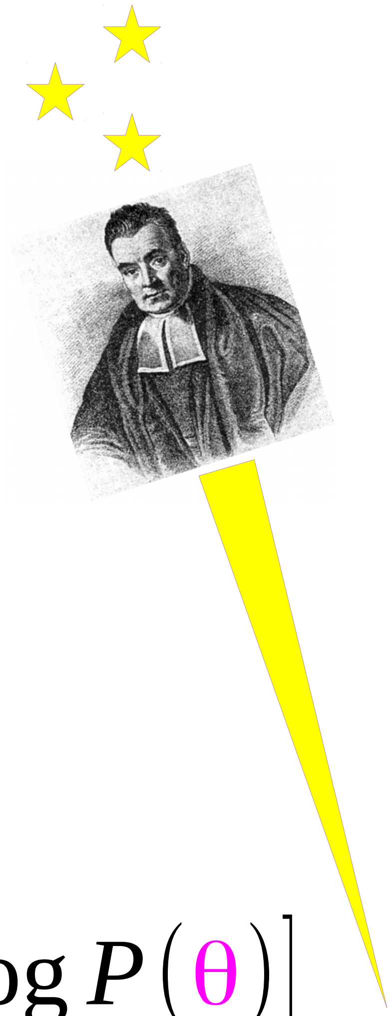
Are we doing this backwards?

$$\hat{\theta} = \arg\max_{\theta} P(X_1, \ldots, X_n | \theta)$$

$$\hat{\theta} = \arg\max_{\theta} P(\theta | X_1, \ldots, X_n)$$



Bayes to the rescue



$$\hat{\theta} = \arg\max_{\theta} P(\theta | X_1, \ldots, X_n)$$

$$= \arg\max_{\theta} \frac{P(X_1, \ldots, X_n | \theta) P(\theta)}{P(X_1, \ldots, X_n)}$$

$$= \arg\max_{\theta} P(X_1, \ldots, X_n | \theta) P(\theta)$$

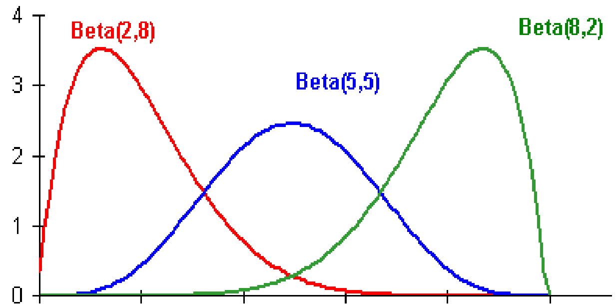
$$= \arg\max_{\theta} \left[\log P(X_1, \ldots, X_n | \theta) + \log P(\theta) \right]$$

Review: Beta random variable

An **beta** random variable models the **probability** of a trial's success, given previous trials. The PDF/CDF let you compute **probabilities of probabilities**!

$$X \sim \text{Beta}(a, b)$$

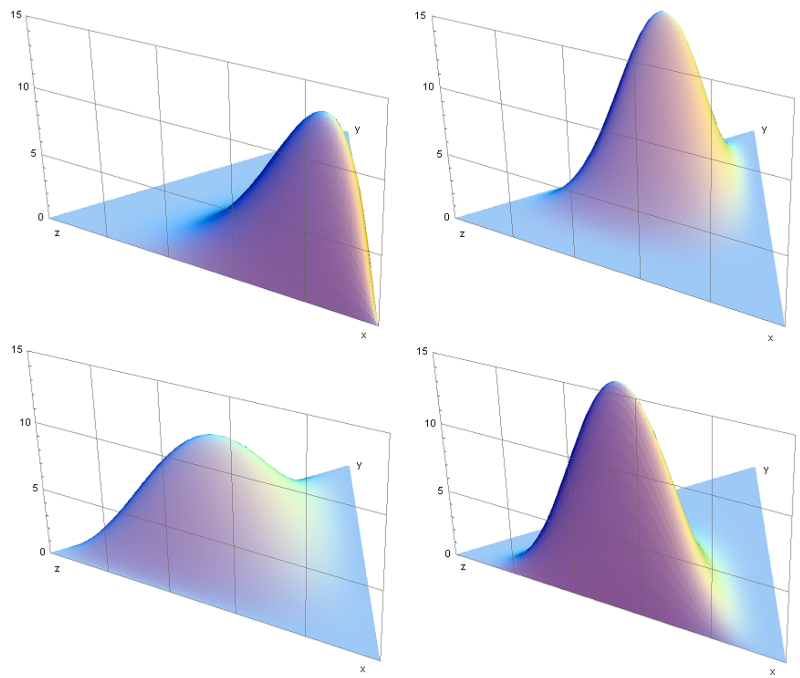
$$f_X(x) = \begin{cases} C\, x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$



Review: Dirichlet distribution

Beta is the distribution (“conjugate prior”) for the p in the **Bernoulli** and **binomial**.

Dirichlet is the distribution for the p_1, p_2, ... in the **multinomial**.



$$X_1, X_2, \ldots \sim \mathrm{Dir}(a_1, a_2, \ldots)$$

$$f_{X_1, X_2, \ldots}(x_1, x_2, \ldots) = C\, x_1^{a_1 - 1}\, x_2^{b_2 - 1} \ldots$$

if $0 < \{x_1, x_2, \ldots\} < 1$,

$x_1 + x_2 + \cdots = 1$

(0 otherwise)

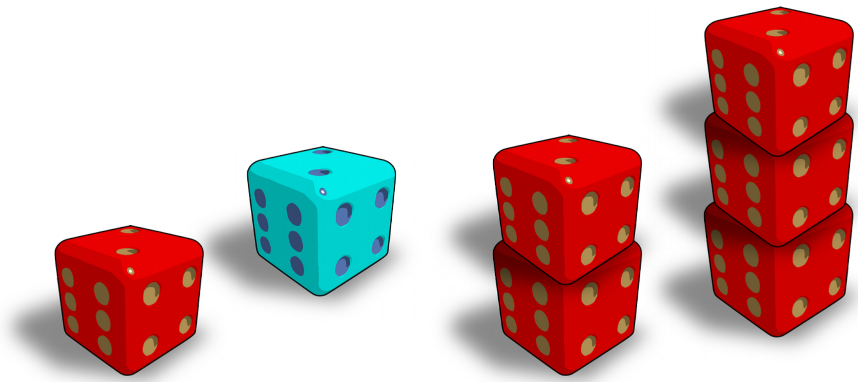
Laplace smoothing

Also known as **add-one** smoothing: assume you've seen one "imaginary" occurrence of each possible outcome.

$$p_i = \frac{\#(X=i)+k}{n+mk}$$

$$p_i = \frac{\#(X=i)+1}{n+m}$$

Maximum likelihood with multinomial



A 6-sided die is rolled 7 times. We get:

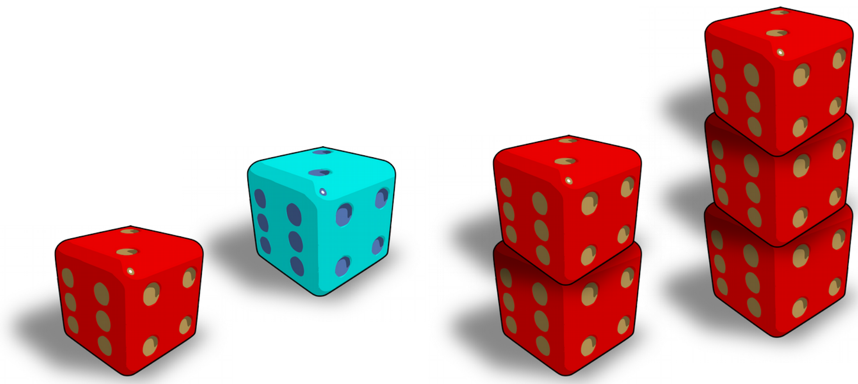
- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

What is the MLE for $p_1, \ldots, p_6$?

$$X_1, \ldots, X_6 \sim \text{MN}\left(7, \frac{1}{7}, \frac{1}{7}, 0, \frac{2}{7}, 0, \frac{3}{7}\right)$$

you'll never roll a 3!
not in a million years!

Laplace with multinomial



A 6-sided die is rolled 7 times. We get:
- 1 one
- 1 two
- 0 threes
- 2 fours
- 0 fives
- 3 sixes

What is the Laplace estimate for $p_1, \ldots, p_6$?

$$X_1, \ldots, X_6 \sim \mathrm{MN}\left(7, \frac{2}{13}, \frac{2}{13}, \frac{1}{13}, \frac{3}{13}, \frac{1}{13}, \frac{4}{13}\right)$$

still a chance!

Parameter priors

$$X \sim \begin{cases} \text{Ber}(p) & p \sim \text{Beta}(a, b) \\ \text{Bin}(n, p) & p \sim \text{Beta}(a, b) \\ \text{MN}(\boldsymbol{p}) & \boldsymbol{p} \sim \text{Dir}(\boldsymbol{a}) \\ \text{Poi}(\lambda) & \lambda \sim \text{Gamma}(k, \theta) \\ \text{Exp}(\lambda) & \lambda \sim \text{Gamma}(k, \theta) \\ \text{N}(\mu, \sigma^2) & \mu \sim \text{N}(\mu', \sigma'^2) \\ & \sigma^2 \sim \text{InvGamma}(\alpha, \beta) \end{cases}$$