p04 Background Lecture
Introduction to probabilistic thinking in bioinformatics (Part 2)

Rachel Karchin

BME 580.488, 580.688
Spring 2019

Lecture 3 topics

- Math
  - Probabilistic models applied to sequence motifs
  - Null models, statistical significance and P-values (including multiple testing correction “MTC”)
  - Bill Noble “Primer” from Nat. Biotechnology on MTC
- Biology
  - Binding sites and sequence motifs

Extra material

- Brief discussion of Bill Noble's MTC Primer
- Analytical null models and the binomial distribution
- Binomial Likelihood and maximum likelihood
- Multinomial distribution
- Multinomial Likelihood and maximum likelihood
- Very brief intro to hidden Markov models

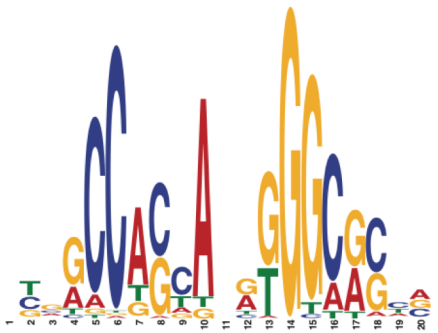
Nat. Biotechnology 27:12 2009

How does multiple testing correction work?

William S Noble

When prioritizing hits from a high-throughput experiment, it is important to correct for random events that falsely appear significant. How is this done and what methods should be used?

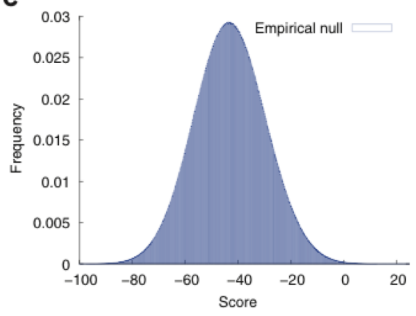
a



b

Position	Str	Sequence	Score
19390631	+	TTGACCAGCAGGGGGCGCCG	26.30
32420105	+	CTGGCCAGCAGAGGGCAGCA	26.30
27910537	–	CGGTGCCCCCTGCTGGTCAG	26.18
21968106	+	GTGACCACCAGGGGGCAGCA	25.81
31409358	+	CGGGCCTCCAGGGGGCGCTC	25.56
19129218	–	TGGCGCCACCTGCTGGTCAC	25.44
21854623	+	CTGGCCAGCAGAGGGCAGGG	24.95
12364895	+	CCCGCCAGCAGAGGGAGCCG	24.71
13406383	+	CTAGCCACCAGGTGGCGGTG	24.71
18613020	+	CCCGCCAGCAGAGGGAGCCG	24.71
31980801	+	ACGCCCAGCAGGGGGCGCCG	24.71
32909754	–	TGGCTCCCCCTGGCGGCCGG	24.71
25683654	+	TCGGCCACTAGGGGGCACTA	24.58
31116990	–	GGCCGCCACCTTGTGGCCAG	24.58
29615421	–	CTCTGCCCTCTGGTGGCTGC	24.46
6024389	+	GTTGCCACCAGAGGGCACTA	24.46
26610753	–	CACTGCCCTCTGCTGGCCCA	24.34
26912791	–	GGGCGCCACCTGGCGGTCAC	24.34
20446267	+	CTGCCCACCAGGGGGCAGCG	24.22
21872506	–	TGGCGCCACCTGGCGGCAGC	24.22

c



Model for CTCF binding site

Scores of 20-nt on chr 21

Distribution of scores on shuffled chr 21

Where do scores in Fig b) come from?

Example of a motif scoring method*

	A	A	T	T	G	A
	A	G	G	T	C	C
	A	G	G	A	T	G
	A	G	G	C	G	T

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

→

Weight Matrix

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	.96	.96	-1.6	.59	0
T	-1.6	-1.6	0	.59	0	0

How was the weight computed? → next slide

*Hertz and Stormo. Bioinformatics 1999. 15:7/8

	1	2	3	4	5	6
	j=1	j=2	j=3	j=4	j=5	j=6
N	A	A	T	T	G	A
	A	G	G	T	C	C
	A	G	G	A	T	G
	A	G	G	C	G	T

	1	2	3	4	5	6
A	4	1	0	1	0	1
C	0	0	0	1	1	1
G	0	3	3	0	2	1
T	0	0	1	2	1	1

$$\ln \frac{(n_{i,j} + p_i) / (N+1)}{p_i} \approx \ln \frac{f_{i,j}}{p_i}$$

$i = A, j = 1 \Rightarrow \ln\left(4 + \frac{1}{4}\right) / (4+1) \Big/ \frac{1}{4} = 1.1575$

	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	.96	.96	-1.6	.59	0
T	-1.6	-1.6	0	.59	0	0

n_{ij} = number times letter *i* seen in position *j*
A = alphabet size (rows in matrix)
p_i = background probability of letter *i*
N = number of samples in aligned training set
L = motif length (columns in matrix)

Using the weight matrix to score a putative motif

Position = *j*

Letter = *i*	1	2	3	4	5	6
A	1.2	0	-1.6	0	-1.6	0
C	-1.6	-1.6	-1.6	0	0	0
G	-1.6	.96	.96	-1.6	.59	0
T	-1.6	-1.6	0	.59	0	0

test sequence: A G G T G C

Each cell of the weight matrix contains this value

$$ln \frac{f_{ij}}{p_i}$$

$$I = \sum_{j=1}^{L} \sum_{i=1}^{A} ln \frac{f_{ij}}{p_i} = 4.3$$

$I = 1.2 + 0.96 + 0.96 + 0.59 + 0.59 + 0 = 4.3$

Hertz and Stormo. Bioinformatics 1999. 15:7/8

Key elements of probabilistic model paradigm from last week

Define X $X = \{X_1, X_2, \cdots, X_n\}$ Random variable(s) being modeled

Define M $\theta = \{\theta_1, \theta_2, \cdots, \theta_K\}$ Model parameters

Give an instantiation of X Data *n* observations

$$X = \{4,1,1,5,3,6,1,2,5,2\}$$

Compute P(X|M)

$$P(X \mid M) = \prod_{i=1}^{n} P(X_i = k \mid \theta_k) \quad k \in \{1, \cdots, K\}$$

Likelihood

Probability of X given M is true = $\frac{P(A \cap B)}{P(B)} = \frac{P(A \& B)}{P(B)}$

The analytical null model is a model

Examples of model families that are used as analytical null models

Binomial distribution family

Poisson distribution family

Normal distribution family

The analytical null model is a model

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}$$

Binomial distribution family $$p(X) = \binom{n}{X} p^X (1-p)^{n-X} \qquad X = 0, 1, 2 \cdots$$

Poisson distribution family $$p(X) = \frac{\lambda^X e^{-\lambda}}{X!} \qquad X = 0, 1, 2 \cdots$$

Normal distribution family $$p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad -\infty \le X \le \infty$$

The analytical null model is a model

Each family has a cumulative distribution function (CDF)*

Binomial distribution family $p(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$ $X = 0, 1, 2 \cdots$

Poisson distribution family $p(X \le x) = \frac{e^{-\lambda} \lambda^X}{X!}$ $X = 0, 1, 2 \cdots$

Normal distribution family $p(X \le x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{X - \mu}{\sigma\sqrt{2}}\right)\right]$ $-\infty \le X \le \infty$

*We will use the CDFs to compute p-values

The analytical null model is a model

Examples of model families that are used as analytical null models

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

What we will cover today

Poisson distribution family $p(X) = \frac{\lambda^X e^{-\lambda}}{X!}$

Normal distribution family $p(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(X-\mu)^2}{2\sigma^2}}$

What makes them null-ish?

- Members of these distribution families model events under certain assumptions, such as independence of events, mutual exclusivity of outcomes, and particular parameter values.
- If you choose the right parameters, members of these distribution families can make sense as null models.
- If the events you are looking for violate these assumptions, they may stand out as statistically significant.

The binomial model

Example of model family used as analytical null

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$ $X = 0, 1, 2 \cdots$

- X = number of “successes” in a sequence of *n* independent trials
- Each trial has two mutually exclusive possible outcomes (“0” or “1”)
- Each trial has the probability *p* of success

The binomial model

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}$$

Number of ways to get X "1s" in a sequence of *n* trials

Binomial distribution family $$p(X) = \binom{n}{X} p^X (1-p)^{n-X}$$

- X = number of "successes" in a sequence of *n* independent trials
- Each trial has two mutually exclusive possible outcomes ("0" or "1")
- Each trial has the probability *p* of success

The binomial model

Number of ways to get X "1s"
in a sequence of *n* trials

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Probability of getting X "1s"

for a given trial?

- X = number of "successes" in a sequence of *n* independent trials
- Each trial has two mutually exclusive possible outcomes ("0" or "1")
- Each trial has the probability *p* of success

The binomial model

Number of ways to get X "1s"
in a sequence of *n* trials

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

$(n-X)$?

Probability of getting X "1s" | Probability of getting n-X "0s"

- X = number of "successes" in a sequence of *n* independent trials
- Each trial has two mutually exclusive possible outcomes ("0" or "1")
- Each trial has the probability *p* of success

The binomial model

Applying our model notation from last week

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Random variable(s) being modeled

The binomial model

Applying our model notation from last week

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Random variable(s) being modeled

The binomial model

Applying our model notation from last week

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Random variable being modeled X

Model parameters $\theta = \{p, n\}$

The binomial model

Applying our model notation from last week

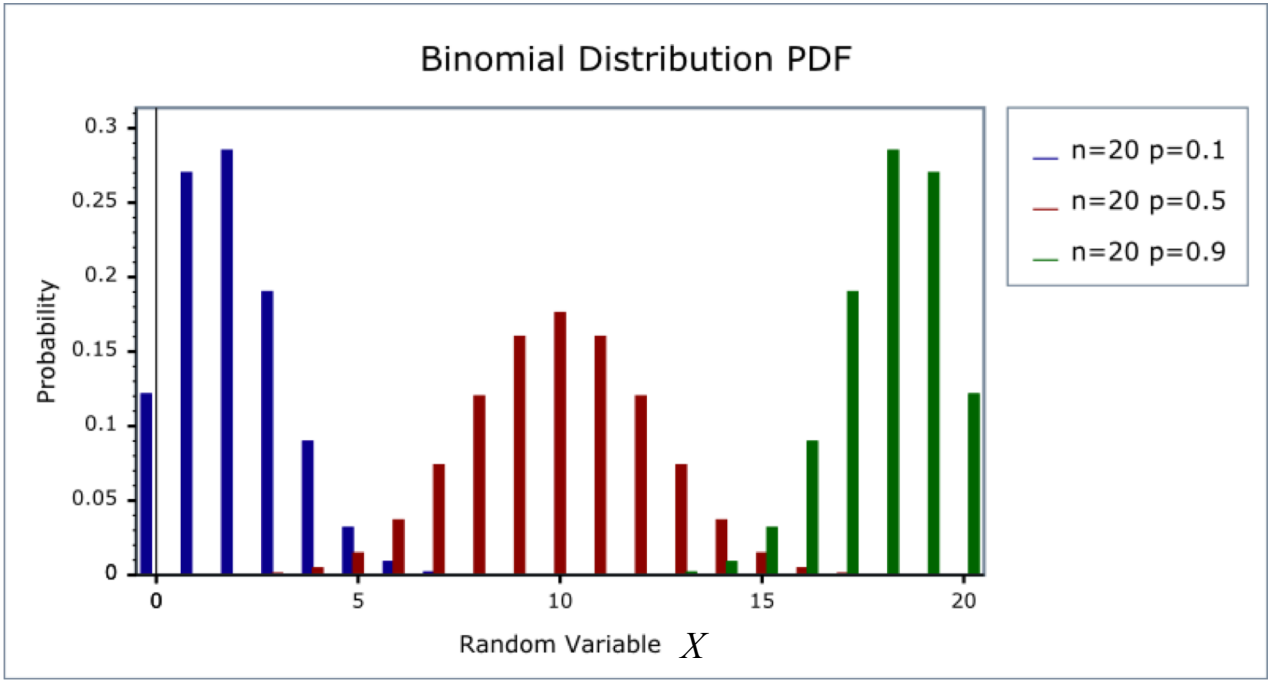
Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Random variable being modeled X

Model parameters $\theta = \{p, n\}$

Binomial distribution family

Each member of the family is defined by its parameter values



Expected value of X

$$E[X] = np$$

Variance of X

$$Var[X] = np(1-p)$$

Binomial null model

- What might be a good binomial distribution family / parameterization to use as an analytical null model?

Binomial null model

- Examples of using a binomial null model

Infer whether a coin is fair or not

- You see this sequence of coin throws: HHHHHHHHHHHHHHHHHHHHHTHHHHHHHHHH
- Can you conclude that the coin is not fair?

Infer whether a coin is fair or not

- Binomial null model
  - Fair coin Binomial $\theta = \{0.5, 30\} = \{p, n\}$
  - Each throw is independent
  - Two mutually exclusive outcomes for each throw
  - p is the same for each throw
- We pick a significance level/ confidence threshold $\alpha = 0.05$
- What is the p-value?
- We will reject the null if $\text{p-value} \leq 0.05$

First - intuition!

Data = HHHHHHHHHHHHHHHHHHHHHTHHHHHHHHH

30 Trials
29 H ("successes" or "1s")
1 T ("failure" or "0")

- If the null model is true (fair coin), how many H are expected?

$$E[X] = np = 30 * 0.5 = 15$$

$$\theta = \{0.5, 30\}$$

- How much variance should we expect to see around that value?

$$Var[X] = np(1-p) = 30 * 0.5 * 0.5 = 7.5$$

What is the p-value ?

Data = HHHHHHHHHHHHHHHHHHHHHHTHHHHHHHH

30 Trials

29 H ("successes" or "1s")

1 T ("failure" or "0")

- If the null model is true, what is the probability of seeing X>=29 (our "one-sided" definition of p-value)
- We know how to compute the probability that X<= x

$$\text{CDF} = p(X \le x) = \sum_{i=0}^{x} \binom{n}{i} p^i (1-p)^{n-i}$$

CDF(28) = 1 calculation?

- Some algebra shows us that the binomial p-value = 1 – CDF(x-1) = 1 – CDF(28) = 0
- So we reject the null hypothesis with $\text{p-value} \le 0.05$ The hypothesis that the coin is fair is rejected!

Is there an evolutionary relationship between two randomly selected short DNA segments in the human genome of the same length?

Segment 1 CCAGCGAACACTGCAATCTTGTTGGAAATTTAAATTAAGAAACATGCAGT
TTAAAGAACCTGGCTCTGAAAACAATTCATGTGGGGACCTTATTTAAGAA

Segment 2 CATGAGTAGGTCCTGCTGCCCAATAAATGAGCTTTGCAGGCACCAAAGCT
GAAAAAAGGGAGAAGAATAAATGTAATGTAGTTGTAATAAGGCTAAAATC

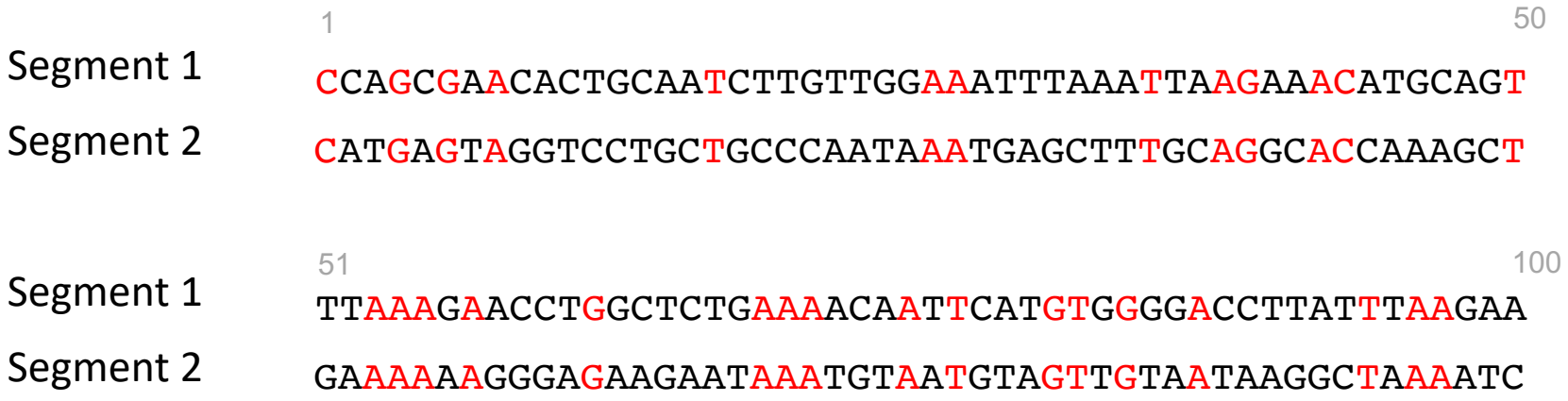
Each segment is 100 nt long

We can line them up and see if they are more identical than expected by chance

	1	50
Segment 1	CCAGCGAACACTGCAATCTTGTTGGAAATTTAAATTAAGAAACATGCAGT	
Segment 2	CATGAGTAGGTCCTGCTGCCCAATAAATGAGCTTTGCAGGCACCAAAGCT	

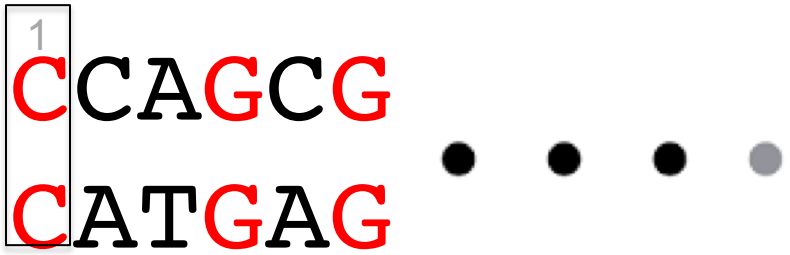
	51	100
Segment 1	TTAAAGAACCTGGCTCTGAAAACAATTCATGTGGGGACCTTATTTAAGAA	
Segment 2	GAAAAAAGGGAGAAGAATAAATGTAATGTAGTTGTAATAAGGCTAAAATC	

How many exact matches should we see by chance?



Aligned DNA segments

- Binomial null model
  - No evolutionary relationship
  - Each column is independent (not actually true)
  - Two mutually exclusive outcomes for each column
  - p is the same for each column
  - n columns (n=100)

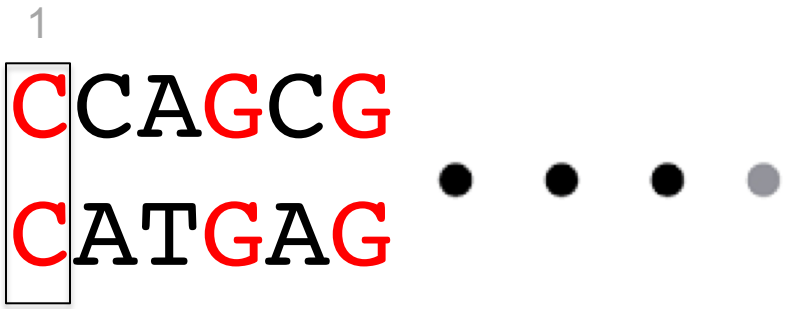


Aligned DNA segments

- What is p?
  - Assume that the nucleotide in the first segment is fixed, then what is the probability of a match under the null hypothesis?

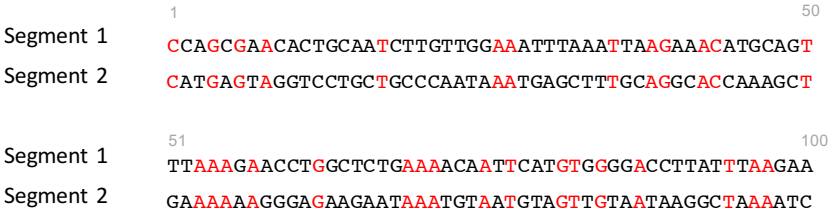
Aligned DNA segments

- What is p?
  - Assume that the nucleotide in the first segment is fixed, then what is the probability of a match under the null hypothesis?
- What is np?
  - How many matches expected under the null model?



What is the p-value

Data =



Binomial null model

$\theta = \{0.25, 100\}$

100 Trials
30 exact matches ("successes" or "1s")
70 mismatches ("failure" or "0")

p-value = 1 – CDF(x-1)

x-1 or n-1 ?

= 0.1495

Using the same significance level as with the fair coin, we can't reject the null hypothesis that the sequences have no evolutionary relationship

$\text{p-value} \leq 0.05$

Binomial likelihood

Applying our model notation from last week

Binomial distribution family $p(X) = \binom{n}{X} p^X (1-p)^{n-X}$

Random variable being modeled X

Model parameters $\theta = \{p, n\}$

Given an instantiation of X

HHHHHHHHHHHHHHHHHHHHHTHHHHHHHHH

Compute P(X|M)

Likelihood function $P(X \mid M) = \prod_{i=1}^{n} P(X_i = k \mid \theta_k) \quad k \in \{1, \cdots, K\}$

Binomial likelihood*

Likelihood function $$P(X \mid M) = \prod_{i=1}^{n} P(X_i = k \mid \theta_k) \quad k \in \{1, \cdots, K\}$$

Key ideas

We have a set of observed data

Example: instantiation of X n=30

	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	T	H	H	H	H	H	H	H	H	H
X_i i=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
success?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

We plug these observations "the X_i" into the PDF formula

$$p(X) = \binom{n}{X} p^X (1-p)^{n-X} \qquad \binom{30}{29} p^{29} (1-p)^{30-29}$$

*This is equivalent to 30 Bernoulli experiments

$$L(p|n, X) = \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i}$$

Binomial likelihood

$$L(p|n, X) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

- The binomial likelihood is a function of *p* given *n* and X
- Note that *n* is no longer a parameter, but is now fixed as part of the observed data how was n a parameter before?
- The prefactor is gone! $\binom{n}{X}$ prefactor
  - Typically we are just comparing likelihoods for different values of p so it cancels
  - Can be shown that the likelihood is only knowable up to a constant

Example: instantiation of X n=30

		H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	H	T	H	H	H	H	H	H	H	H	H
X_i	i=	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
success?		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1

Binomial likelihood

$$L(p|n, X) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$= p^{\sum_1^n X_i}(1-p)^{\sum_1^n 1-X_i}$$

$$L(p|n, s) = p^s(1-p)^{n-s}$$

Sum of successes

Binomial likelihood

$$L(p|n, X) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$

$$= p^{\sum_1^n X_i}(1-p)^{\sum_1^n 1-X_i}$$

$$L(p|n, s) = p^{s}(1-p)^{n-s}$$

Sum of successes

Binomial Likelihood

$$L(p|n,s) = p^s(1-p)^{n-s}$$

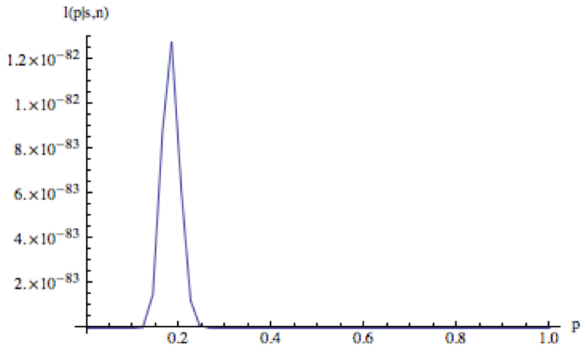
Let: n=400

s=72

We can plot the function over different values of the input p

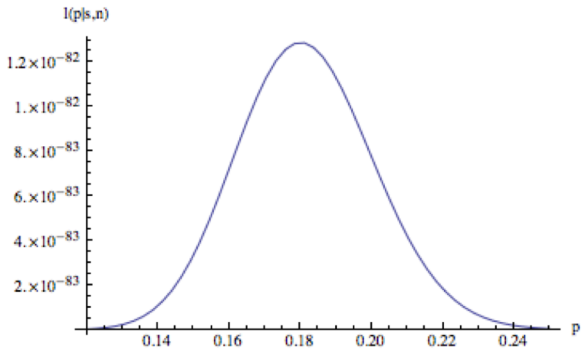
```
In[2]:= l[p_, s_, n_] := p^s * (1 - p)^(n - s)
In[9]:= Plot[l[p, 72, 400], {p, 0, 1}, PlotRange → Full, AxesLabel → {"p", "l(p|s,n)"}]
```

Out[9]=




```
In[10]:= Plot[l[p, 72, 400], {p, 0.12, 0.25}, PlotRange → Full, AxesLabel → {"p", "l(p|s,n)"}]
```

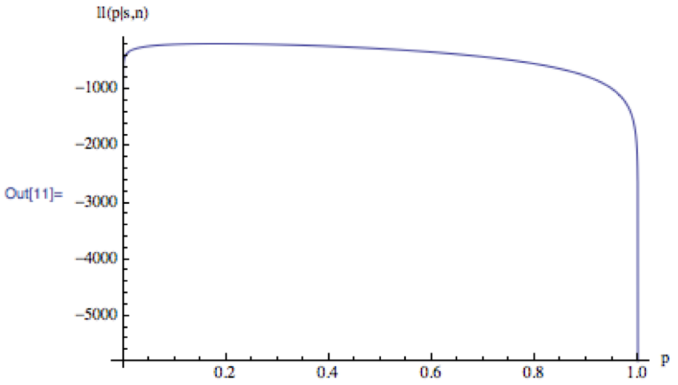
Out[10]=



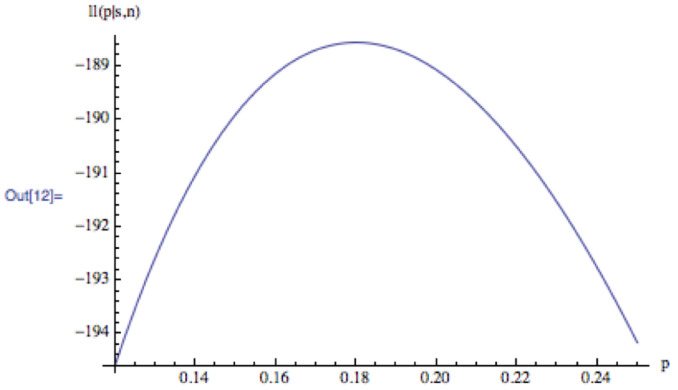
Binomial Log Likelihood

$$LL(p|n,s) = \text{Log}\left(p^s(1-p)^{n-s}\right)$$

```
In[5]:= ll[p_, s_, n_] := Log[l[p, s, n]]
In[11]:= Plot[ll[p, 72, 400], {p, 0, 1}, PlotRange → Full, AxesLabel → {"p", "ll(p|s,n)"}]
```



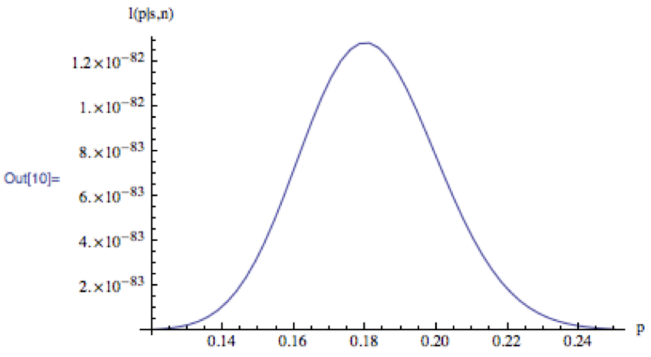

```
In[12]:= Plot[ll[p, 72, 400], {p, 0.12, 0.25}, PlotRange → Full, AxesLabel → {"p", "ll(p|s,n)"}]
```



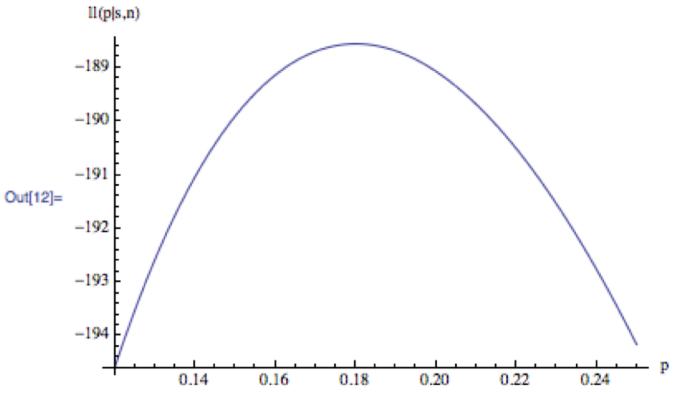
Easier to work with!

y-axis values are now integer-valued
ll(p) is locally quadratic

Maximum likelihood



Maximum of our binomial likelihood function along different values of p



Maximum of our binomial log likelihood function along different values of p

Maximum likelihood

- A primary use of likelihood is to estimate θ

Maximizing the log likelihood is often convenient

We use a standard approach for finding the maximum value of a variable with derivatives

$$\frac{d}{dp} Log\left(p^s \left(1-p\right)^{N-s} \right) = 0$$

We take the derivative of the function with respect to the parameter we are maximizing and set to 0

```
Simplify[D[Log[p^s * (1 - p)^(n - s)], p]]
```

$$\frac{s - n\,p}{p - p^2}$$

```
Solve[(s - n p)/(p - p^2) == 0, p]
```

$$\{\{p \to \frac{s}{n}\}\}$$

Solve for *p*

Definition of sample mean

$$m \equiv \frac{1}{n} \sum_{k=1}^{n} x_k .$$

The maximum likelihood estimate for *p* is the sample mean

Summary of binomial model material

- Binomial model can be a useful analytical null model
- The binomial model family is in fact many related distributions with different parameter values
- Using a binomial to compute a p-value
- The binomial likelihood
- Maximizing the binomial likelihood to get the most likely parameter value for a dataset

The multinomial model

- Generalizes the binomial model.
- Series of independent trials, but each trial can have more than two outcomes.
- Each outcome has its own probability.

The multinomial model

Multinomial distribution family $p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$

$X = 0, 1, 2 \cdots$

$\sum_{k=1}^{K} X_k = N$

$\sum_{k=1}^{K} \theta_k = 1$

- X_1, X_2, . . . = number of outcomes of type 1, type 2, . . . in a sequence of N independent trials
- Each event has k mutually exclusive possible outcomes (1,2, . . ., K)
- Each outcome has probability θ_K at each trial

different Θ or still $\{\varphi, n\}$?

The multinomial model

Number of ways to get X_1 Type 1 events, X_2 Type 2 events . . . X_K Type K events in a sequence of *N* trials

Multinomial distribution family $p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$ $X = 0, 1, 2 \cdots$

- X_1, X_2, . . . = number of outcomes of type 1, type 2, . . . in a sequence of *N* independent trials
- Each event has k mutually exclusive possible outcomes (1,2, . . ., K)
- Each outcome has probability θ_K at each trial

The multinomial model

Number of ways to get X_1 Type 1 events, X_2 Type 2 events . . . X_K Type K events in a sequence of *N* trials

Multinomial distribution family $$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k} \qquad X = 0, 1, 2 \cdots$$

Probability of getting X_k outcomes of type k

- X_1, X_2, . . . = number of outcomes of type 1, type 2, . . . in a sequence of *N* independent trials
- Each event has k mutually exclusive possible outcomes (1,2, . . ., K)
- Each outcome has probability θ_K at each trial

Binomial vs. multinomial trials

Trial $n = 4$	Outcome	
	X	
1	0	1
2	0	1
3	0	1
4	0	1

$$X = 3$$

Trial $N = 4$	Outcome					
	X_1	X_2	X_3	X_4	X_5	X_6
1	1	2	3	4	5	6
2	1	2	3	4	5	6
3	1	2	3	4	5	6
4	1	2	3	4	5	6

$$X_1 = 1$$
$$X_2 = 1$$
$$X_3 = 0$$
$$X_4 = 0$$
$$X_5 = 2$$
$$X_6 = 0$$

The multinomial model

Applying our model notation from last week

Multinomial distribution family $p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$

Random variable(s) being modeled

The multinomial model

Applying our model notation from last week

Multinomial distribution family $p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$

Random variable(s) being modeled

Model parameters $\theta = \{\theta_1, \theta_2, \cdots, \theta_K, N\}$

The multinomial model

Applying our model notation from last week

Multinomial distribution family $p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$

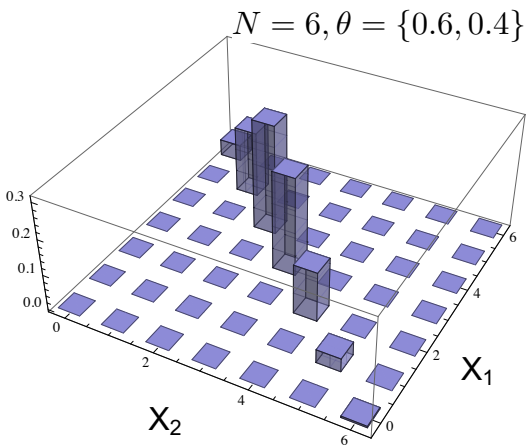
Random variable(s) being modeled

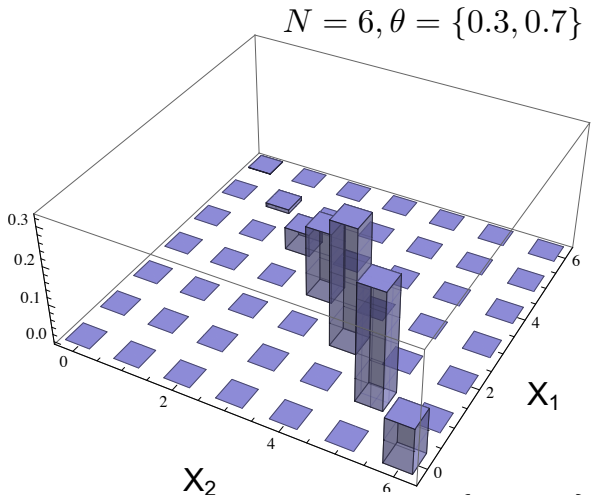
Model parameters $\theta = \{\theta_1, \theta_2, \cdots, \theta_K, N\}$

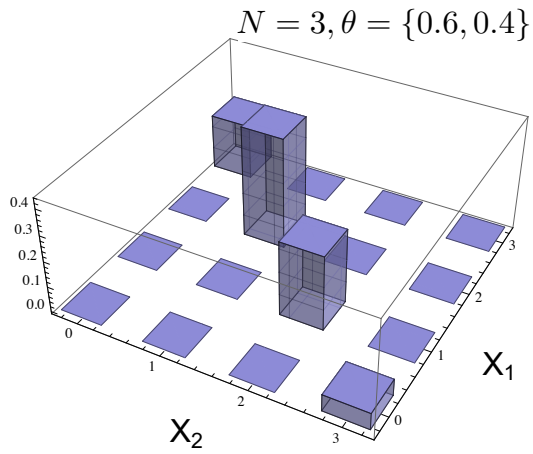
Multinomial distribution family

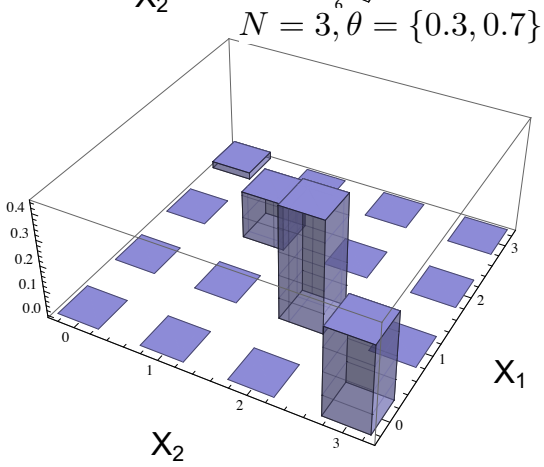
Each member of the family is defined by its parameter values

N = # of trials









Expected value of X_k

$$E[X_k] = N\theta_k$$

Variance of X_k

$$Var[X_k] = N\theta_k(1 - \theta_k)$$

Multinomial model: understanding the pre-factor

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Multinomial model: understanding the pre-factor

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Given our input, the number of possible orderings

Multinomial model: understanding the pre-factor

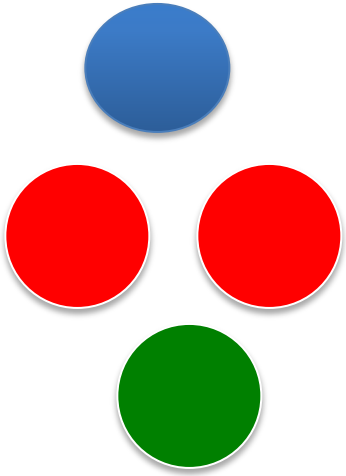
$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Drawing balls from an urn that has three colors of balls

N = 4 = Number of observed trials (Balls drawn from the urn)

$$X_1 = 1$$
$$X_2 = 2$$
$$X_3 = 1$$

Observations (unordered)



1 ball / trial

K = 3 = Number of possible outcomes for each trial (blue red green)

Multinomial model: understanding the pre-factor

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Drawing balls from an urn that has three colors of balls

N = 4 = Number of observed trials (Balls drawn from the urn)
X_1 (blue) = 1, X_2 (red) =2, X_3 (green) = 1

Many different possible orderings!

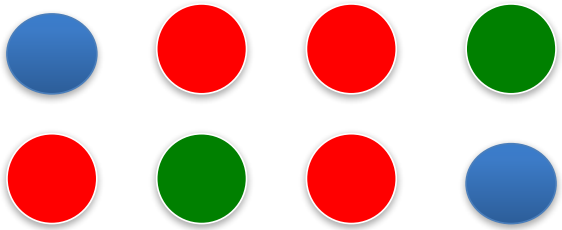
Multinomial model: understanding the pre-factor

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Drawing balls from an urn that has three colors of balls

N = 4 = Number of observed trials (Balls drawn from the urn)
X_1 (blue) = 1, X_2 (red) =2, X_3 (green) = 1

Many different possible orderings!



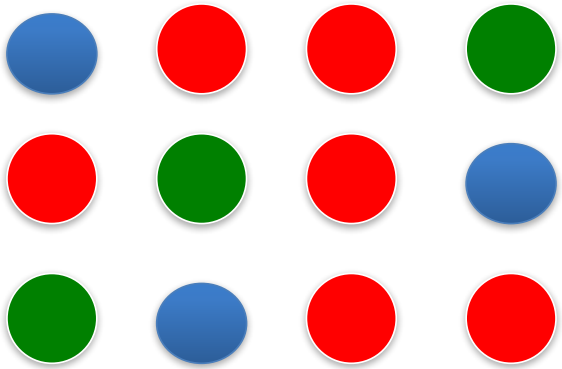
Multinomial model: understanding the pre-factor

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Drawing balls from an urn that has three colors of balls

N = 4 = Number of observed trials (Balls drawn from the urn)
X_1 (blue) = 1, X_2 (red) =2, X_3 (green) = 1

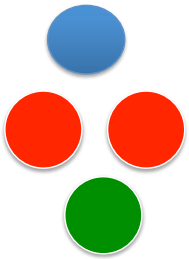
Many different possible orderings!

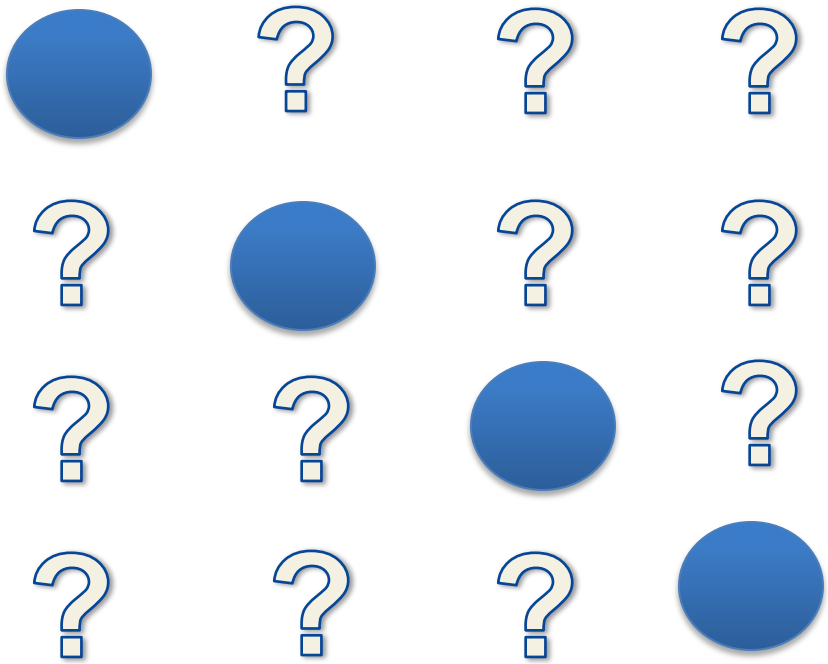


$N = 4$

How many orderings?

$X_1 = 1$
$X_2 = 2$
$X_3 = 1$





We have N (four) "slots". X_1 is the number of times we observed a blue circle How many ways could this have happened?

$$\binom{N}{X_1} = \binom{4}{1} = 4$$

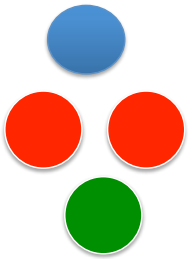
$$\binom{N}{X_1}$$

4

First term in the prefactor

$N = 4$

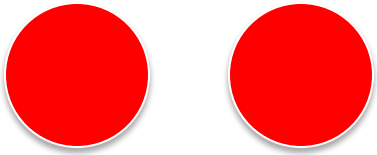
$X_1 = 1$
$X_2 = 2$
$X_3 = 1$



How many orderings for the two red balls?

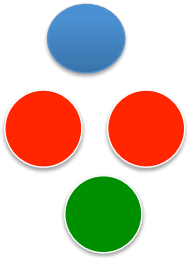
Is it $\binom{N}{X_2}$?

No!

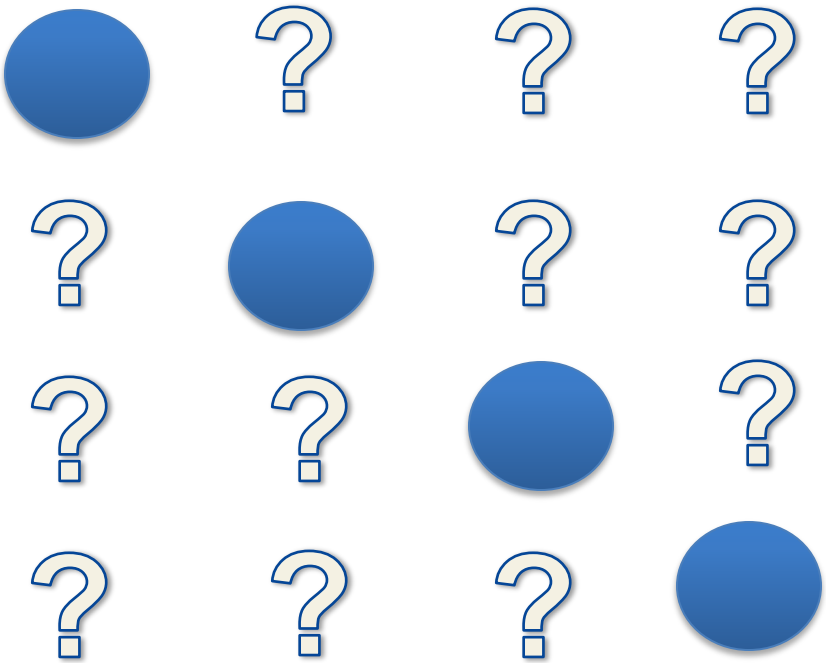


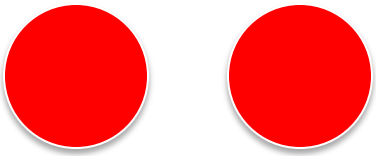
$$N = 4$$

$X_1 = 1$
$X_2 = 2$
$X_3 = 1$



We don't have N slots anymore. How many do we have for the red balls? We have N-X_1 (three) left because we used up X_1 slots available in each possible ordering already

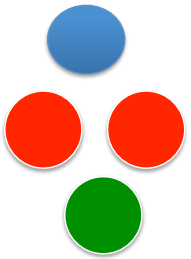


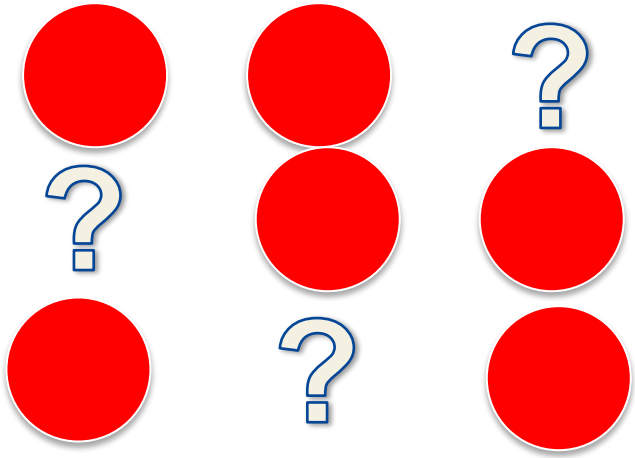


$$N = 4$$

How many ways can the X_2 (two) red balls go into the N-X_1 (three) slots?

$X_1 = 1$
$X_2 = 2$
$X_3 = 1$





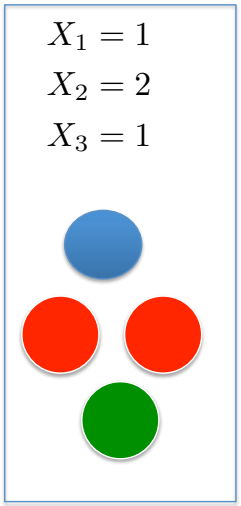
$$\binom{N}{X_1}\binom{N - X_1}{X_2}$$

4 * 3

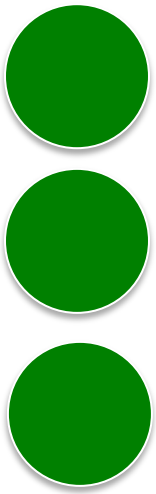
$$\binom{N - X_1}{X_2} = \binom{3}{2} = 3$$

First and second term in the prefactor

$$N = 4$$



Now we have N – X_1 – X_2 slots left
How many do we have for the green ball?
One!
How many ways can it go? The number of ways we can put X_3 (one) thing into N – X_1 – X_2 (one) slot.



$$\binom{N}{X_1}\binom{N - X_1}{X_2}\binom{N - X_1 - X_2}{X_3}$$

4 * 3 * 1 = 12

First and second and third term in the prefactor

$$\binom{N}{X_1}\binom{N-X_1}{X_2}\binom{N-X_1-X_2}{X_3}$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

binomial coefficient

$$= \frac{N!}{X_1!X_2!X_3!}$$

Multinomial Distribution

- We roll a fair 100-sided die N times
- Probability of getting each of the 100 outcomes is $\theta_1, \theta_2, \ldots, \theta_{100}$ where $\theta_1=\theta_2=\ldots=\theta_{100}=0.01$
- What is the probability of rolling 200 times and getting each outcome twice?

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100} \right)^{200}$$

$$\frac{N!}{X_1! X_2! X_3! \cdots X_{100}!}$$

$$\theta_1 \cdot \theta_2 \cdot \theta_3 \cdots \theta_{100} \cdot \theta_1 \cdot \theta_2 \cdot \theta_3 \cdots \theta_{100}$$

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100} \right)^{200}$$

Configuration of events we observed in terms of our random variables

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100} \right)^{200}$$

Configuration of events we observed in terms of our random variables

Number of ways the observed configuration could occur

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

Probability of getting outcome k on a single trial

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100} \right)^{200}$$

Configuration of events we observed in terms of our random variables

Number of ways the observed configuration could occur

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

Probability of getting outcome k on a single trial

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100}\right)^{200}$$

Configuration of events we observed in terms of our random variables

Number of ways the observed configuration could occur

Multinomial Distribution

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

Probability of getting outcome k on a single trial

$$p(X_1 = 2, X_2 = 2, \ldots, X_{100} = 2) = \frac{200!}{2!^{100}} \left(\frac{1}{100} \right)^{200}$$

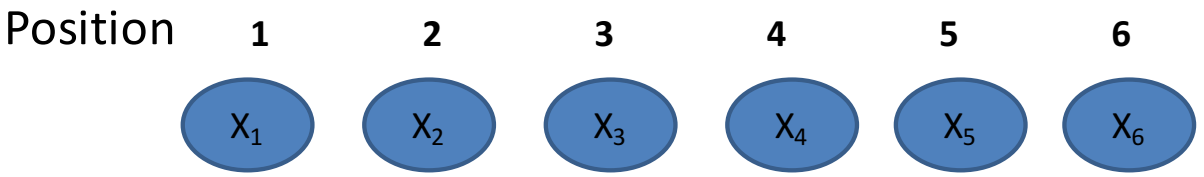
$2*2*2* \ldots * 2$ (100 terms)

Configuration of events we observed in terms of our random variables

Number of ways the observed configuration could occur

Multinomial distribution

AGAAGA
ACGACU
GAUCAA
AAGCCA
GAAACA
CAGAUC
AGGAAA
CAAUCA
UGGAAC
GAUGAA
AAGGAU
GAAGGA
CAGAGG
GAAACG
GAAAGC
AAAUCC
AGAAAC
AUCCAA
UGAAGU
AAUGAC
GUCAAG
GACAAA



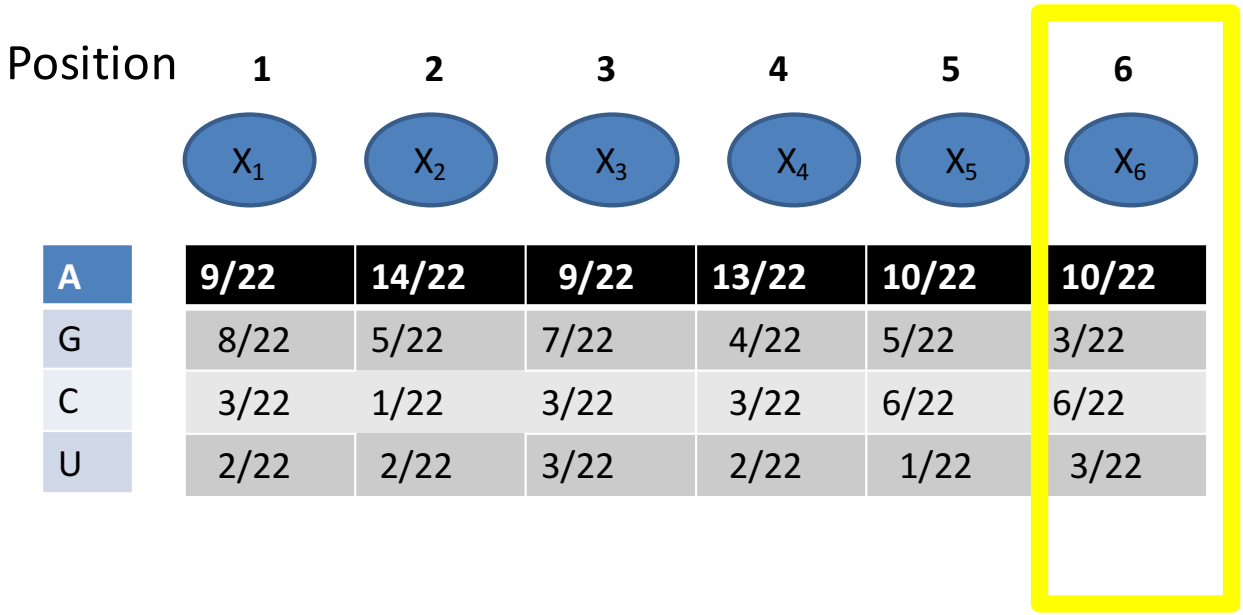
A	9/22	14/22	9/22	13/22	10/22	10/22
G	8/22	5/22	7/22	4/22	5/22	3/22
C	3/22	1/22	3/22	3/22	6/22	6/22
U	2/22	2/22	3/22	2/22	1/22	3/22

- For each position in the motif, we can generate a multinomial model

Exonic Splicing Enhancers

Multinomial distribution

AGAAGA
ACGACU
GAUCAA
AAGCCA
GAAACA
CAGAUC
AGGAAA
CAAUCA
UGGAAC
GAUGAA
AAGGAU
GAAGGA
CAGAGG
GAAACG
GAAAGC
AAAUCC
AGAAAC
AUCCAA
UGAAGU
AAUGAC
GUCAAG
GACAAA



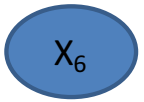
A	9/22	14/22	9/22	13/22	10/22	10/22
G	8/22	5/22	7/22	4/22	5/22	3/22
C	3/22	1/22	3/22	3/22	6/22	6/22
U	2/22	2/22	3/22	2/22	1/22	3/22

- We will model position 6

Exonic Splicing Enhancers

Multinomial model of position 6

- 22 sequence motifs
- The column has 10 A, 3 G, 6 C, 3 U



10/22
3/22
6/22
3/22

$$p(X_1, X_2, \ldots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{x_k}$$

N = 22 # of trials

K = 4 # of possible outcomes

θ = $\left\{\frac{10}{22}, \frac{3}{22}, \frac{6}{22}, \frac{3}{22}\right\}$

X = $\{10, 3, 6, 3\}$

$$p(X_1, X_2, X_3, X_4) = \frac{22!}{X_1! X_2! X_3! X_4!} \left(\frac{10}{22}^{X_1} \times \frac{3}{22}^{X_2} \times \frac{6}{22}^{X_3} \times \frac{3}{22}^{X_4}\right)$$

Multinomial model θ estimates are maximum likelihood estimates

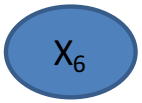
N = 22

K = 4

θ = $\left\{\frac{10}{22}, \frac{3}{22}, \frac{6}{22}, \frac{3}{22}\right\}$

X = $\{10, 3, 6, 3\}$

$$\hat{\theta_k} = \frac{X_k}{N}$$



10/22
3/22
6/22
3/22

Multinomial maximum likelihood

$$L(\theta_1, \theta_2, \cdots, \theta_K | N, X_1, \cdots, X_K) = \frac{N!}{\prod_{k=1}^{K} X_k!} \prod_{k=1}^{K} \theta_k^{X_k}$$

Log of the likelihood

$$LL(\theta_1, \theta_2, \cdots, \theta_K | N, X_1, \cdots, X_K) = Log(N! - \sum_{k=1}^{K} Log(X_k!) + \sum_{k=1}^{K} X_k Log(\theta_k)$$

Constrain with Lagrange multipliers to ensure that $\sum_{k=1}^{K} \theta_k = 1$

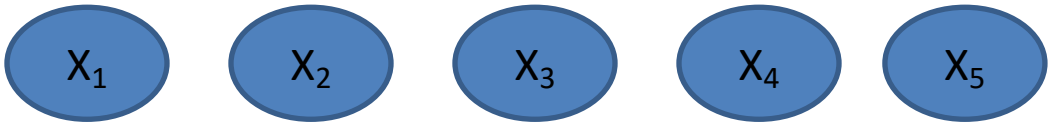
Solution is $\hat{\theta}_k = \frac{X_k}{N}$

Summary of multinomial model material

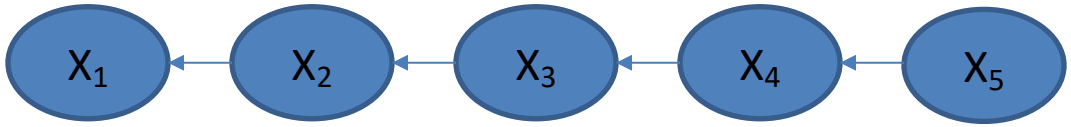
- How multinomial model can incorporate more than one outcome for each trial in a series of trials
- Applying the multinomial model to our ESE sequence data set
- Maximum likelihood estimates of multinomial parameters

Markov chains and Markov models

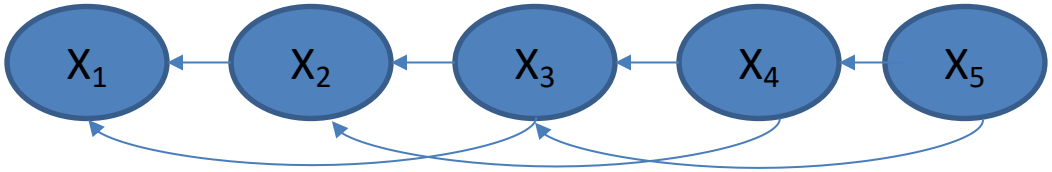
Markov chain (zero-order)



Markov chain (first-order)



Markov chain (second-order)



Order 1 Markov Chain

- $X_i = \{A,C\}$
- Observed sequence: X = ACCCACA

Model:

Prev	Next	Prob
A	A	0.7
A	C	0.3
C	A	0.5
C	C	0.5

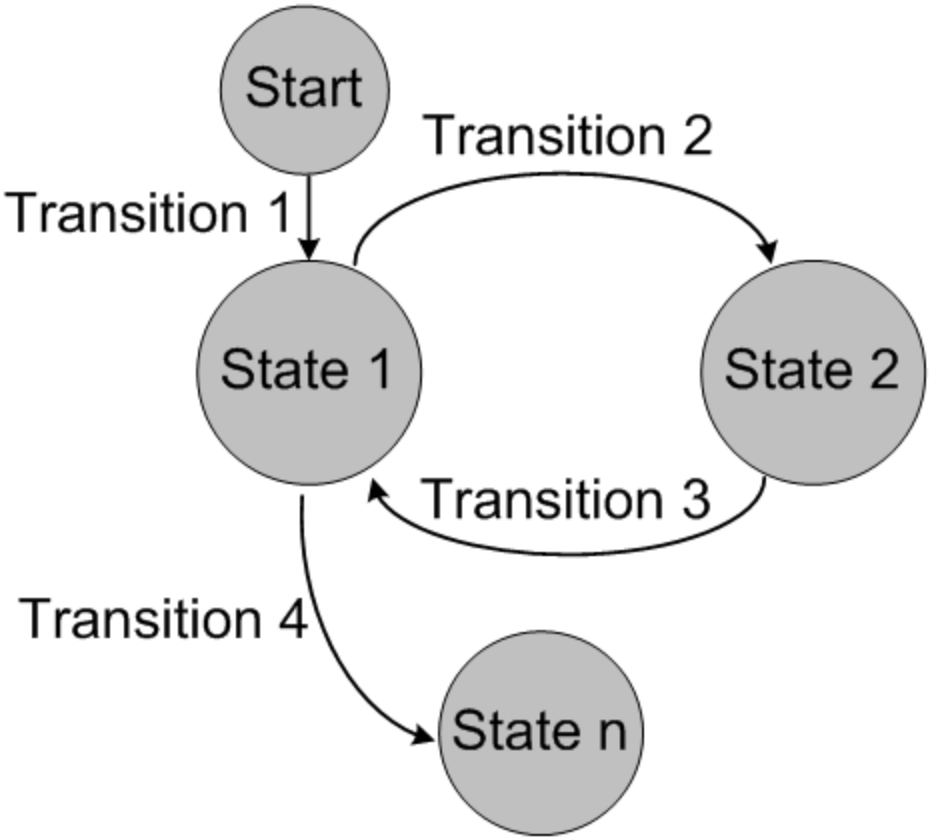
start probs	A	0.5
	C	0.5

P(X) =0.5 * 0.3* 0.5 * 0.5* 0.5 * 0.3 * 0.5

Hidden Markov model

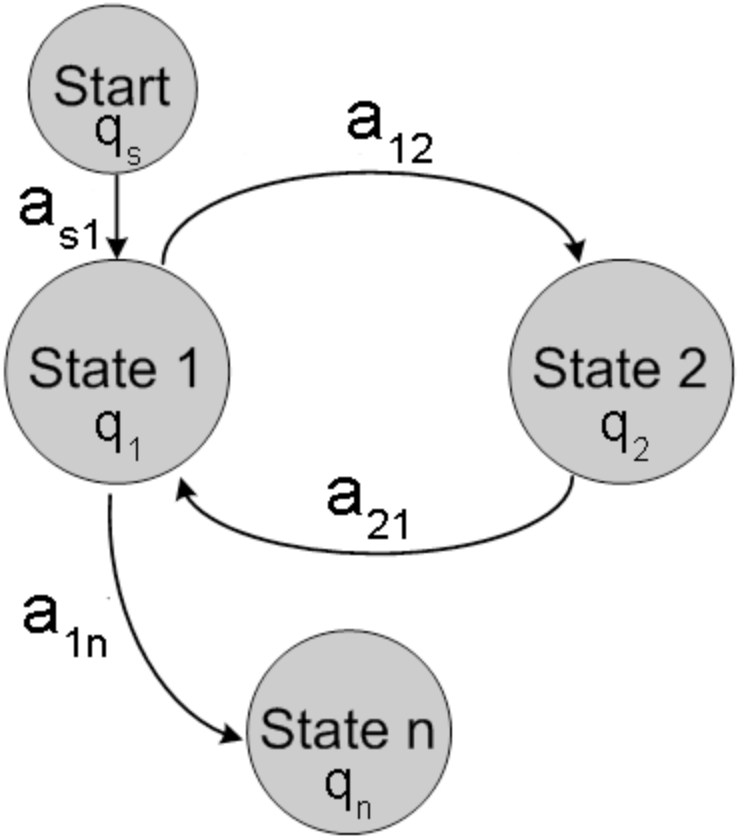
- You have a sequence of observations
- That are the result of a sequence of hidden states
- The hidden states are random variables that form a Markov chain
- Algorithms have been developed to infer the sequence of hidden states from the sequence of observations.

Hidden Markov model as finite state machine



Hidden Markov model as finite state machine

$$\left.\begin{matrix}\{a_{kl}\} \\ \{e_k(b)\}\end{matrix}\right\} \theta$$

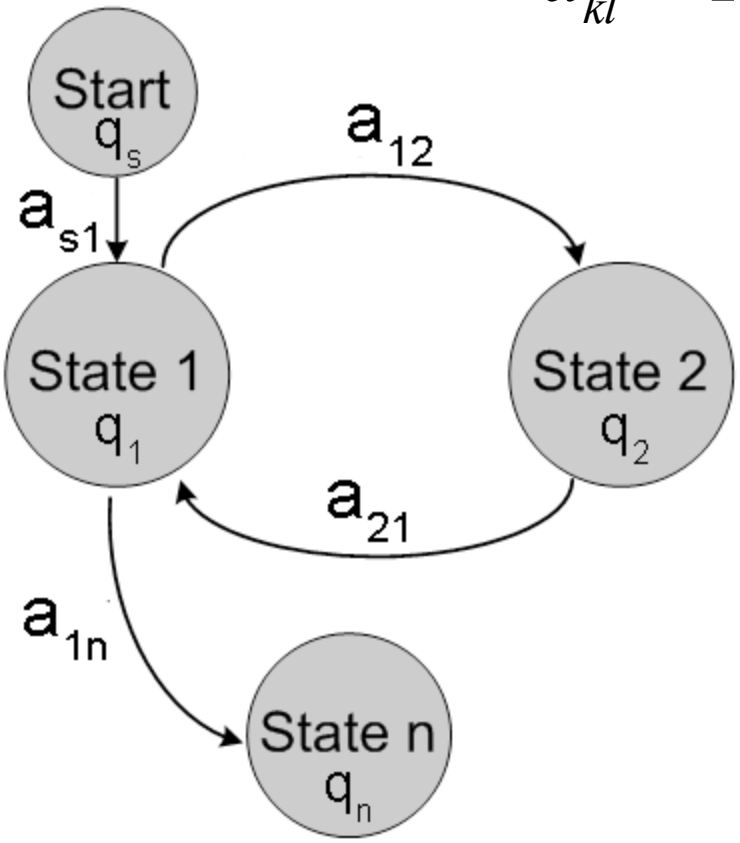


- State transition probabilities
- State emission probabilities

Hidden Markov model as finite state machine

$$\left.\begin{matrix}\{a_{kl}\} \\ \{e_k(b)\}\end{matrix}\right\} \theta$$

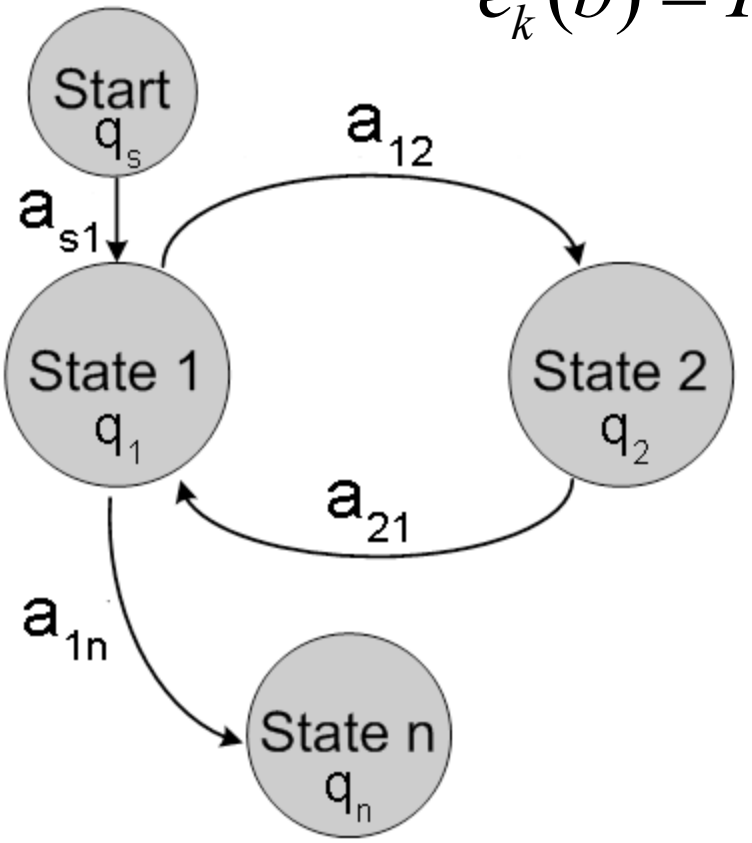
$$a_{kl} = P(q_i = l \mid q_{i-1} = k)$$



Hidden Markov model as finite state machine

$$\left.\begin{matrix}\{a_{kl}\} \\ \{e_k(b)\}\end{matrix}\right\} \theta$$

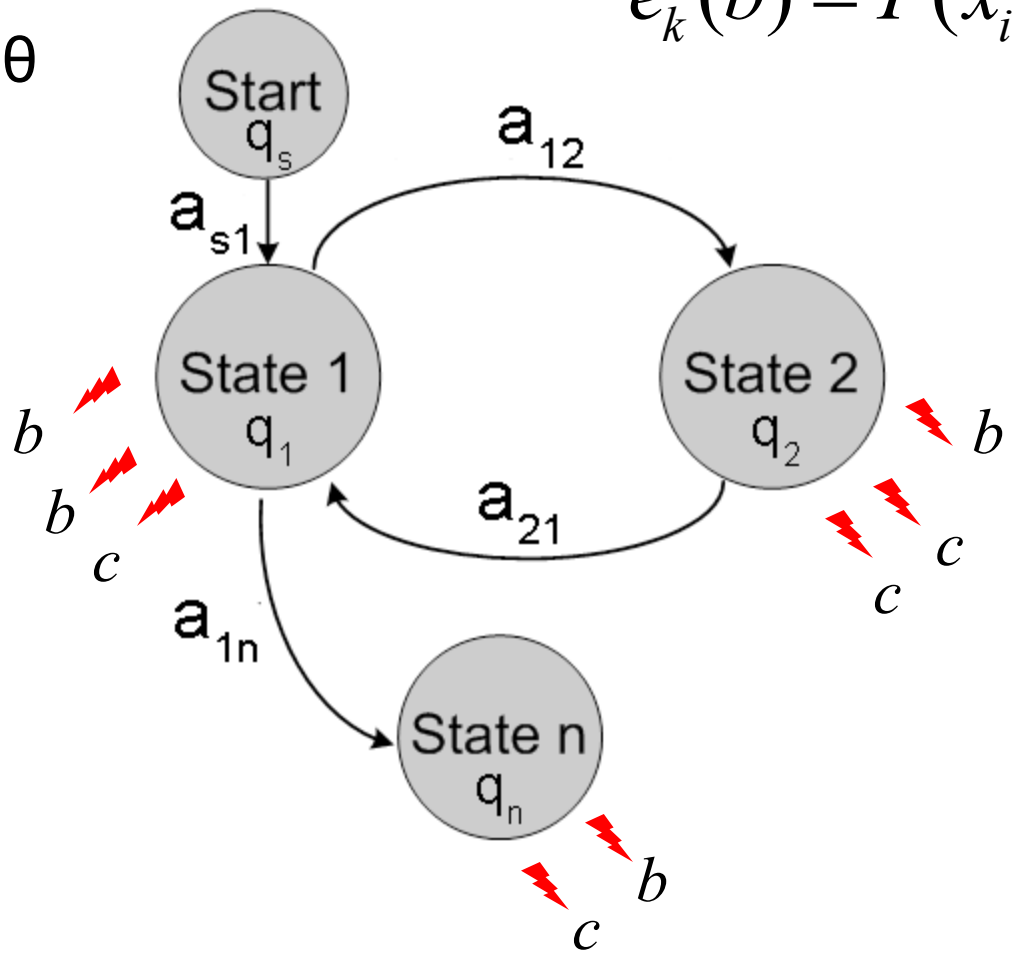
$$e_k(b) = P(x_i = b \mid q_i = k)$$



Hidden Markov model as finite state machine

$$\left.\begin{matrix}\{a_{kl}\} \\ \{e_k(b)\}\end{matrix}\right\} \theta$$

$$e_k(b) = P(x_i = b \mid q_i = k)$$



Summary

- Motif scoring matrices
- Binomial models as analytical null
- Binomial likelihood and maximum likelihood
- Multinomial models
- Multinomial likelihood and maximum likelihood
- Brief intro to Markov chains and hidden Markov models