# Modeling Motif Using Positional Weight Matrix

Xiaohui S. Xie
University of California, Irvine

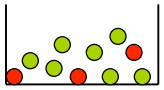
## Today's Goals

- Last lecture
  - Enumeration-based method
    - P-value
      - Binomial
      - Hypergeometric
- Probabilistic modeling
  - Position Weight Matrix
  - EM-algorithm

# Binomial distribution

**Experiment 1:** Sampling with replacement

A box with 3 red balls and 7 green balls:



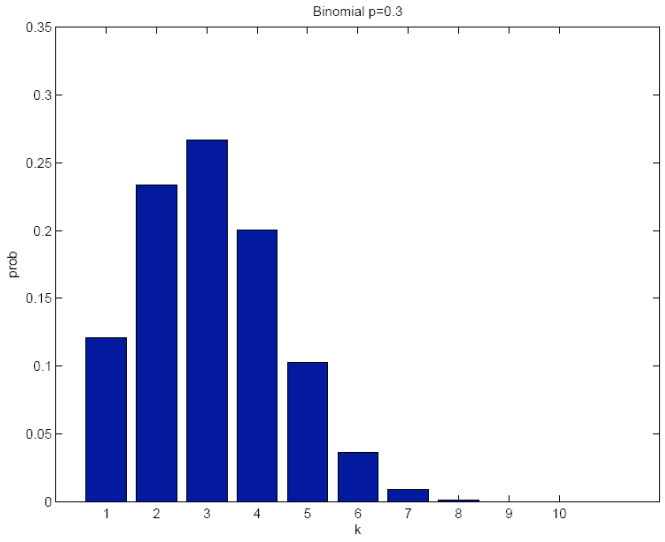
Q1: Randomly pick one ball. What's the chance that the ball is red?

**p = 3/10**

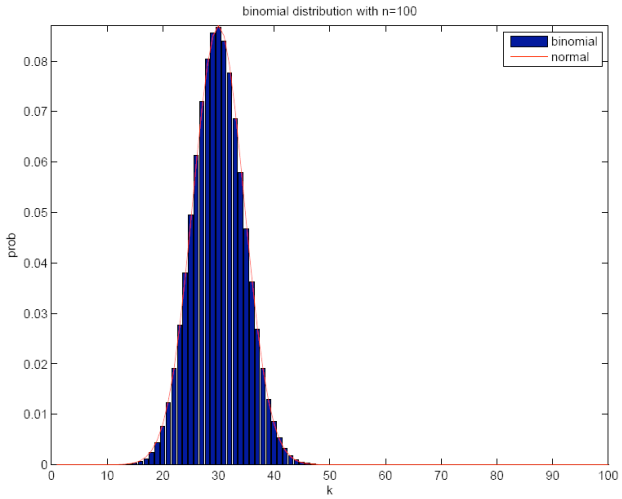
Q2: Randomly pick one ball.
Place back a ball with the same color.
Repeat *n* times.
What's the chance that the total number of red balls you pick is *k*?

**$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$**

# Binomial distribution



# Binomial distribution

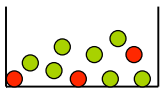



When n is large, the binomial distribution can be approximated by the normal distribution with
Mean: ***np***
Variance: ***np(1-p)***

# Hypergeometric distribution

**Experiment 2:** Sampling without replacement

A box with 30 red balls and 70 green balls:



Q1: Randomly pick one ball. What's the chance that the ball is red?

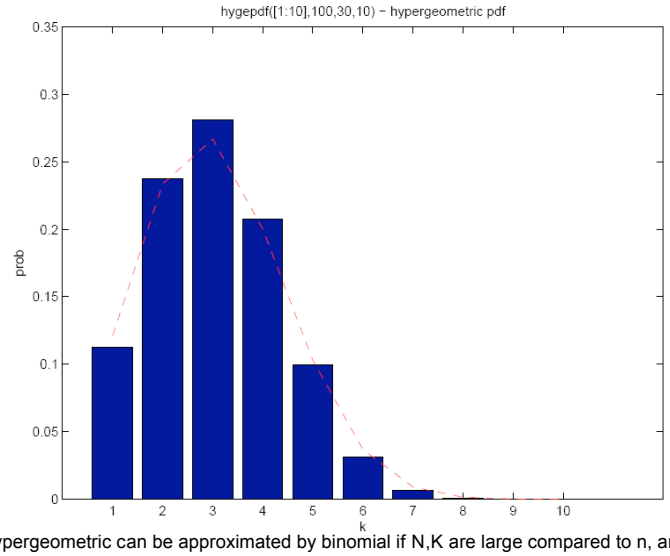
**p = 3/10**

Q2: Randomly pick one ball. Repeat *n* times.
What's the chance that the total number of red balls you pick is *k*?

Total number of red balls: K
Total number of balls: N

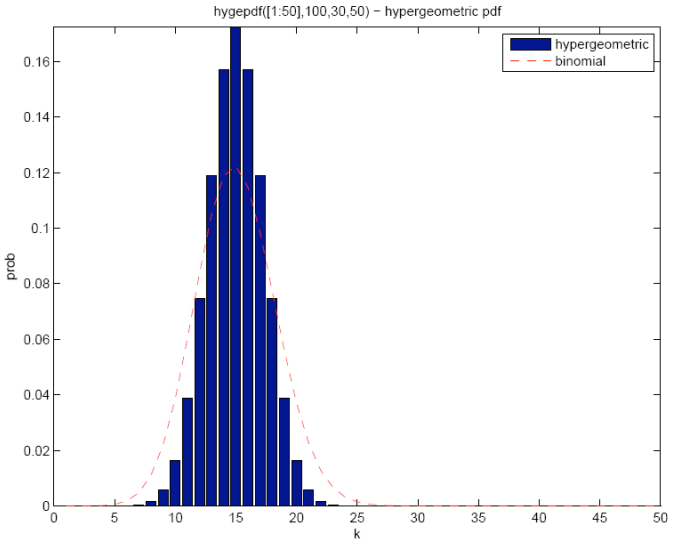
$$P(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

# Hypergeometric




Hypergeometric can be approximated by binomial if N,K are large compared to n, and K/N is not close to 0 or 1.

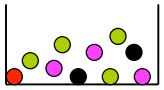
# Hypergeometric



# Multinomial distribution

**Experiment 3:** Sampling with replacement

A box with 1 red balls, 2 **black** balls, 3 pink balls, and 4 green balls:



Q1: Randomly pick one ball. What's the chance that the ball is red?

- $p_1$=1/10
- $p_2$=2/10
- $p_3$=3/10
- $p_4$=4/10

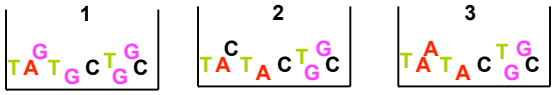
Q2: Sampling with replacement, what's the chance of picking 3 balls with colors in the order of:

$$p_1 * p_4 * p_2$$

# Position Weight Matrix

**Experiment 4:** Sampling with replacement from ***W=3*** boxes

Boxes with nucleotides: A, **C**, G, and T:




| | | | |
|---|---|---|---|
| **A** | $\theta_{11}$ | $\theta_{21}$ | $\theta_{31}$ |
| **C** | $\theta_{12}$ | $\theta_{22}$ | $\theta_{32}$ |
| **G** | $\theta_{13}$ | $\theta_{33}$ | $\theta_{33}$ |
| **T** | $\theta_{14}$ | $\theta_{44}$ | $\theta_{34}$ |

Q: Pick one letter from each box in the order of 1,2,3. What's the chance of picking: **ACT**?

$$\theta_{11}\,\theta_{22}\,\theta_{34}$$

# Positional weight matrix representation

```
1   GTATCACCGCCAGTGGTAT
2   ATACCACTGGCGGTGATAC
3   TCAACACCGCCAGAGATAA
4   TTATCTCTGGCGGTGTTGA
5   TTATCACCGCAGATGGTTA
6   TAACCATCTGCGGTGATAA
7   CTATCACCGCAAGGGATAA
8   TTATCCCTTGCGGTGATAG
9   CTAACACCGTGCGTGTTGA
10  TCAACACGCACGGTGTTAG
11  TTACCTCTGGCGGTGATAA
12  TTATCACCGCCAGAGGTAA
```

Lambda cI/cro binding sites

$W_{ij}$

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A: | 9 | 9 | 94 | 25 | 1 | 71 | 1 | 1 | 1 | 9 | 17 | 32 | 9 | 17 | 1 | 48 | 1 | 71 | 63 |
| C: | 17 | 17 | 1 | 25 | 94 | 9 | 86 | 55 | 9 | 40 | 71 | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 9 |
| G: | 9 | 1 | 1 | 1 | 1 | 1 | 1 | 9 | 71 | 40 | 9 | 55 | 86 | 9 | 94 | 25 | 1 | 17 | 17 |
| T: | 63 | 71 | 1 | 48 | 1 | 17 | 9 | 32 | 17 | 9 | 1 | 1 | 1 | 71 | 1 | 25 | 94 | 9 | 9 |

Sequence Logo