

# Debugging Intuition

- How to calculate the probability of at least  $k$  successes in  $n$  trials?

- $X$  is number of successes in  $n$  trials each with probability  $p$
- $P(X \geq k) =$

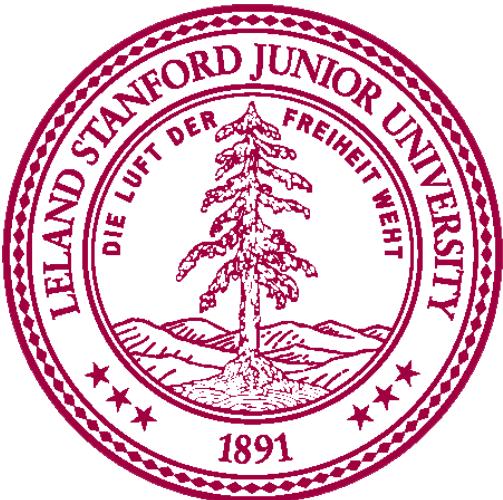
$$\binom{n}{k} p^k$$

# ways to choose slots for success      Don't care about the rest  
Probability that each is success

First clue that something is wrong.  
Think about  $p = 1$

Not mutually exclusive...

Correct: 
$$P(X \geq k) = \sum_{i=k}^n \binom{n}{i} p^i (1-p)^{n-i}$$



# Variance

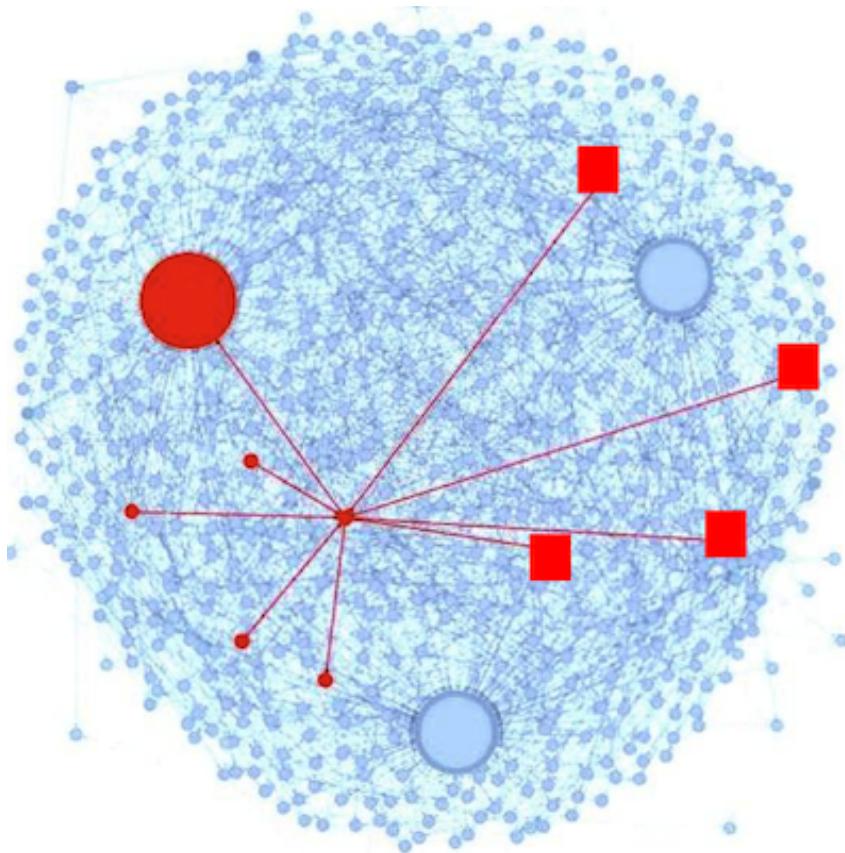
Chris Piech  
CS109, Stanford University

# Learning Goals

1. Be able to calculate variance for a random variable
2. Be able to recognize and use a Bernoulli Random Var
3. Be able to recognize and use a Binomial Random Var



# Is Peer Grading Accurate Enough?



Peer Grading on Coursera  
HCI.

31,067 peer grades for  
3,607 students.

# Review: Random Variables



A **random variable** takes on values probabilistically.

For example:  
 $X$  is the sum of two dice rolled.

$$P(X = 2) = \frac{1}{36}$$

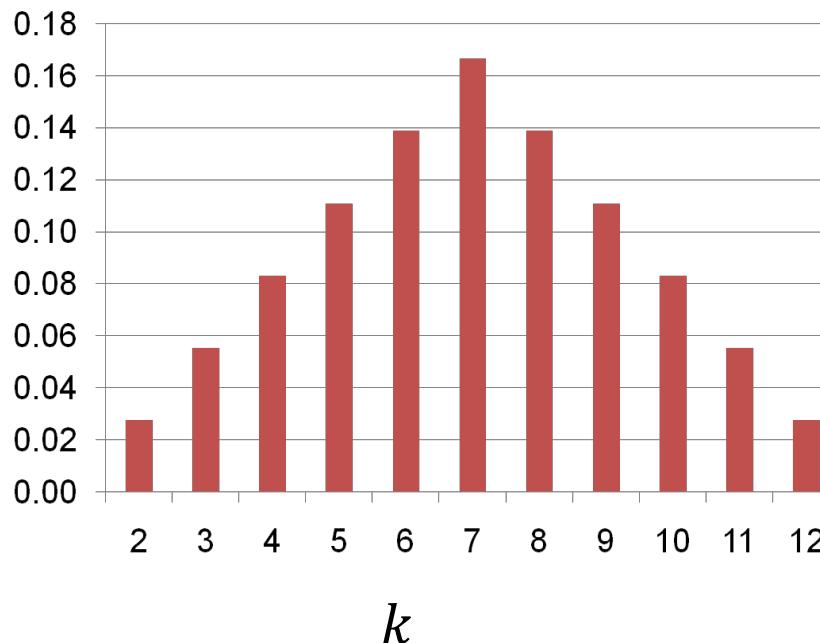
# Review: Probability Mass Function



The **probability mass function** (PMF) of a random variable is a function from values of the variable to probabilities.

$$p_Y(k) = P(Y = k)$$

$$P(Y = k)$$

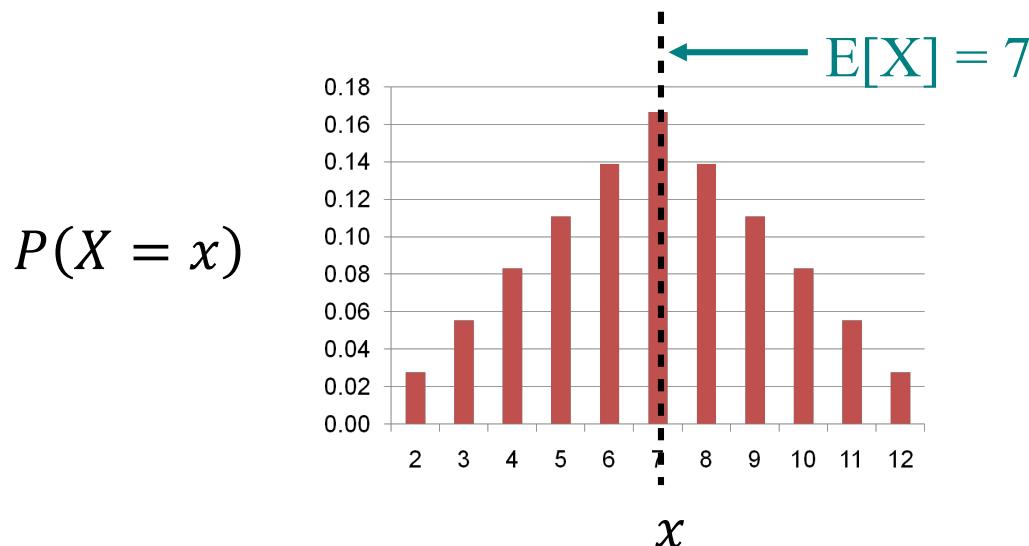


# Review: Expectation



The **expectation** of a random variable is the “**average**” value of the variable (weighted by probability).

$$E[X] = \sum_{x:p(x)>0} p(x) \cdot x$$

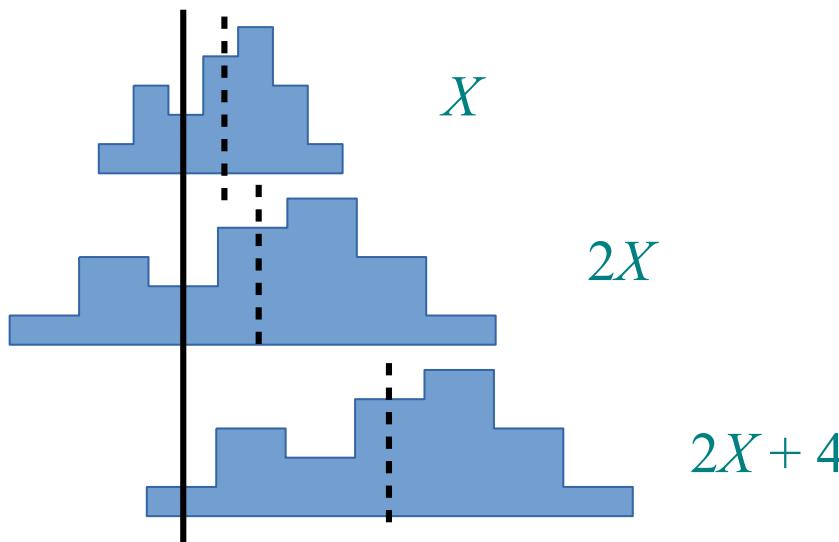


# Review: Linearity of Expectation

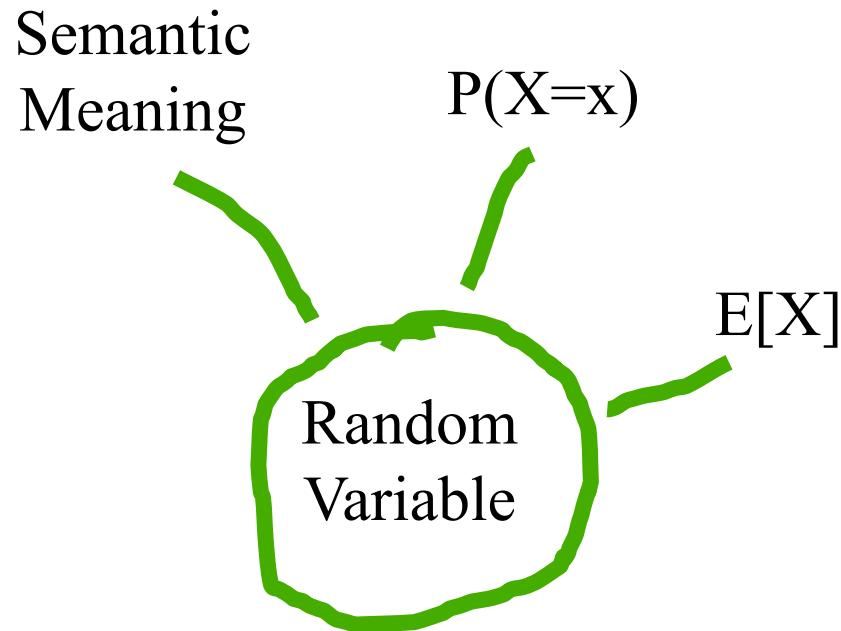


Adding random variables or constants? **Add** the expectations. Multiplying by a constant? **Multiply** the expectation by the constant.

$$E[aX + bY + c] = aE[X] + bE[Y] + c$$

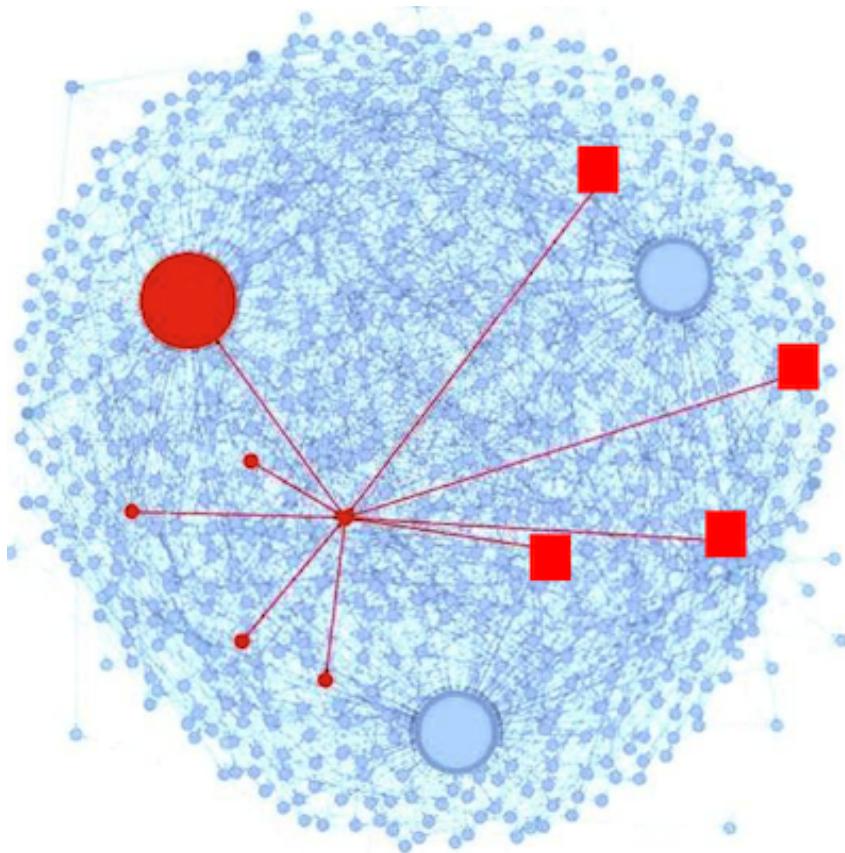


# Fundamental Properties



Is  $E[X]$  enough?

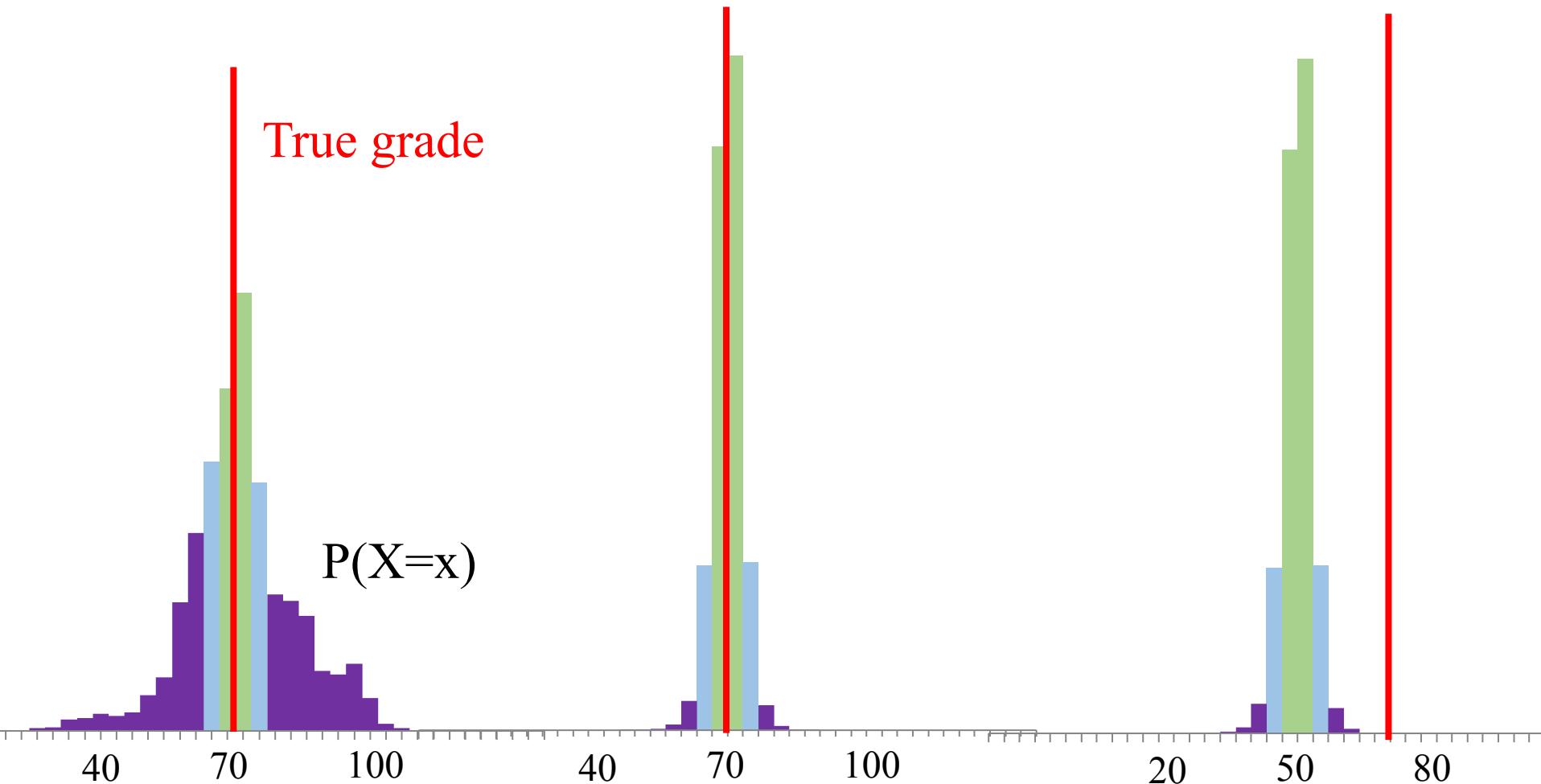
# Intuition



Peer Grading on Coursera  
HCI.

31,067 peer grades for  
3,607 students.

$X$  is the score a peer grader gives to an assignment submission



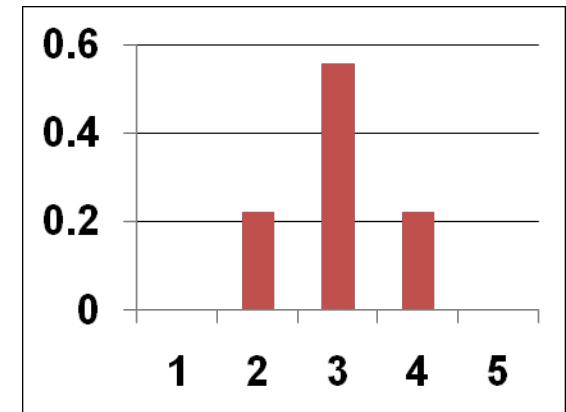
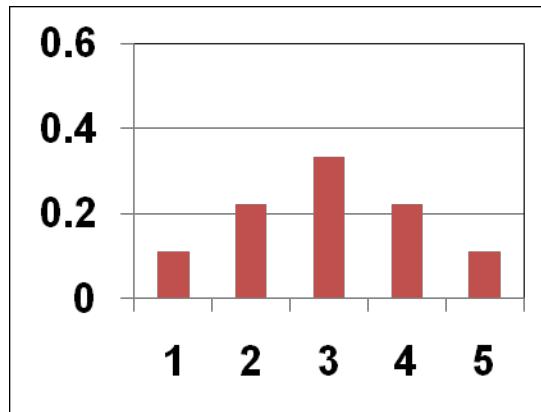
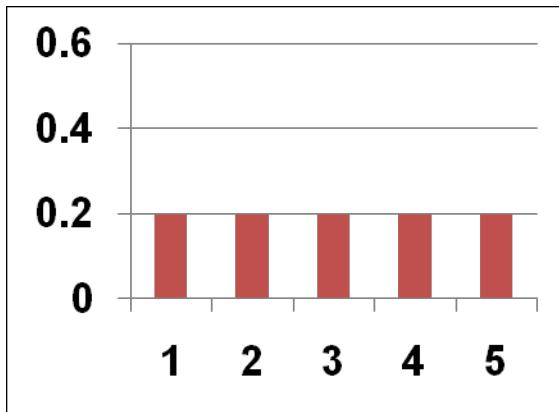
A

B

C

# Variance

- Consider the following 3 distributions (PMFs)



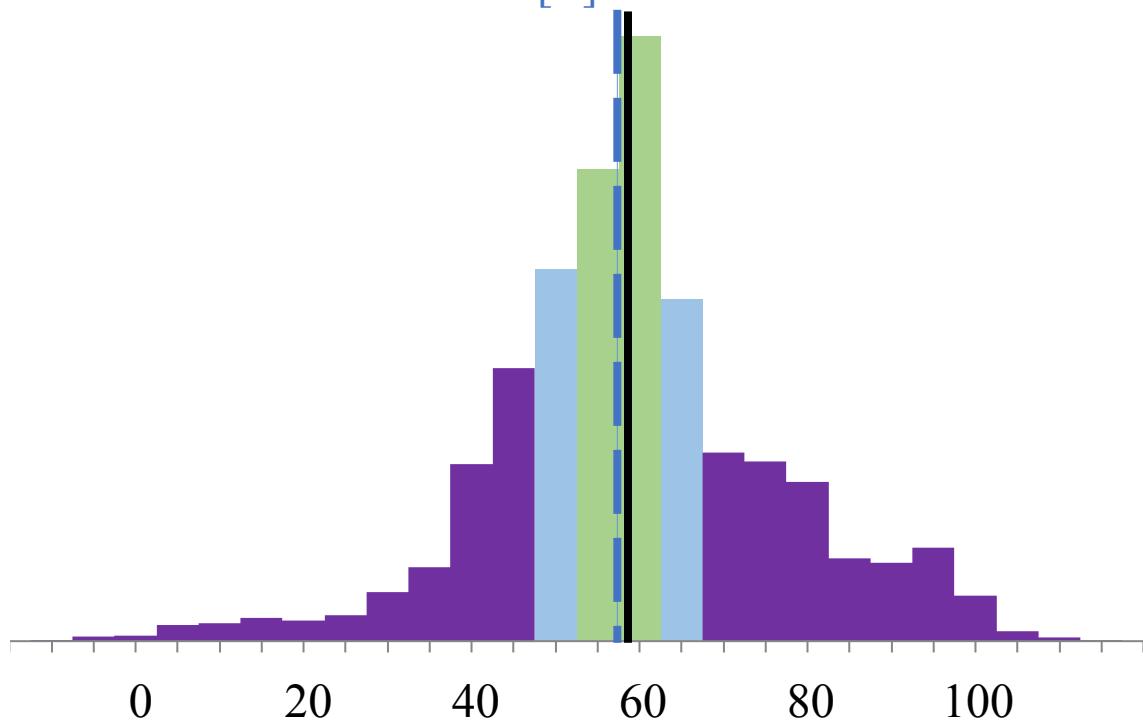
- All have the same expected value,  $E[X] = 3$
- But “spread” in distributions is different
- Variance = a formal quantification of “spread”

# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

True grade = 58

$$E[X] = 57.5$$



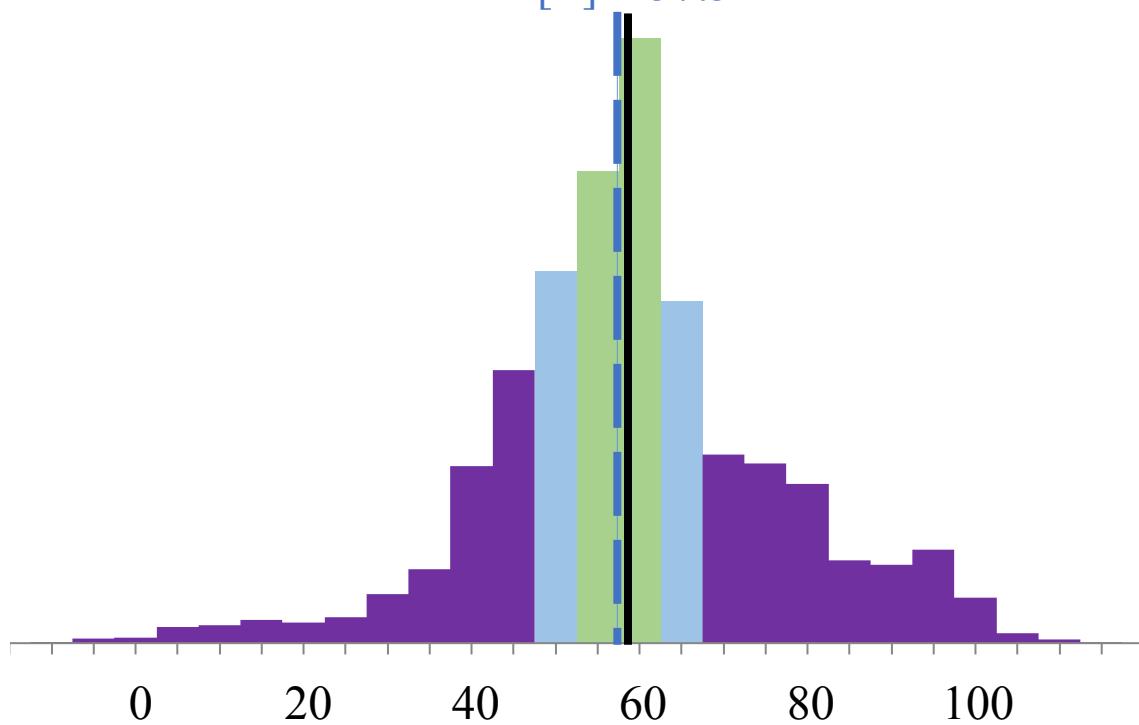
# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58

$$E[X] = 57.5$$



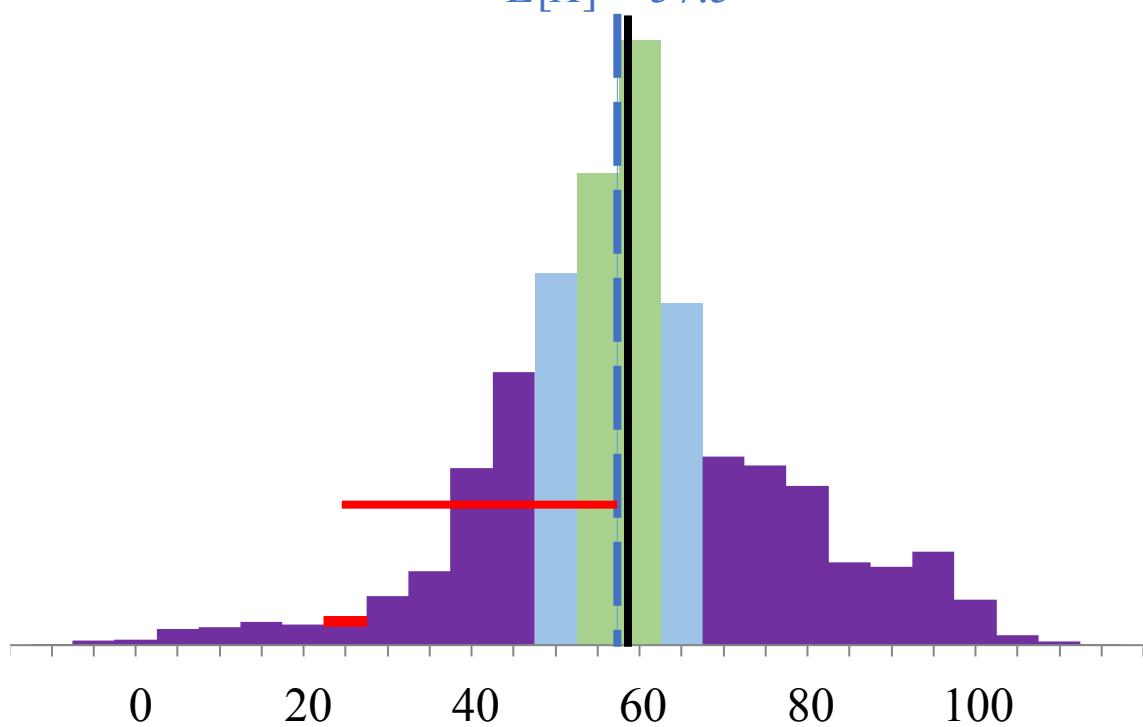
# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58  
 $E[X] = 57.5$

$X$	$(X - \mu)^2$
25 points	1056 points <sup>2</sup>

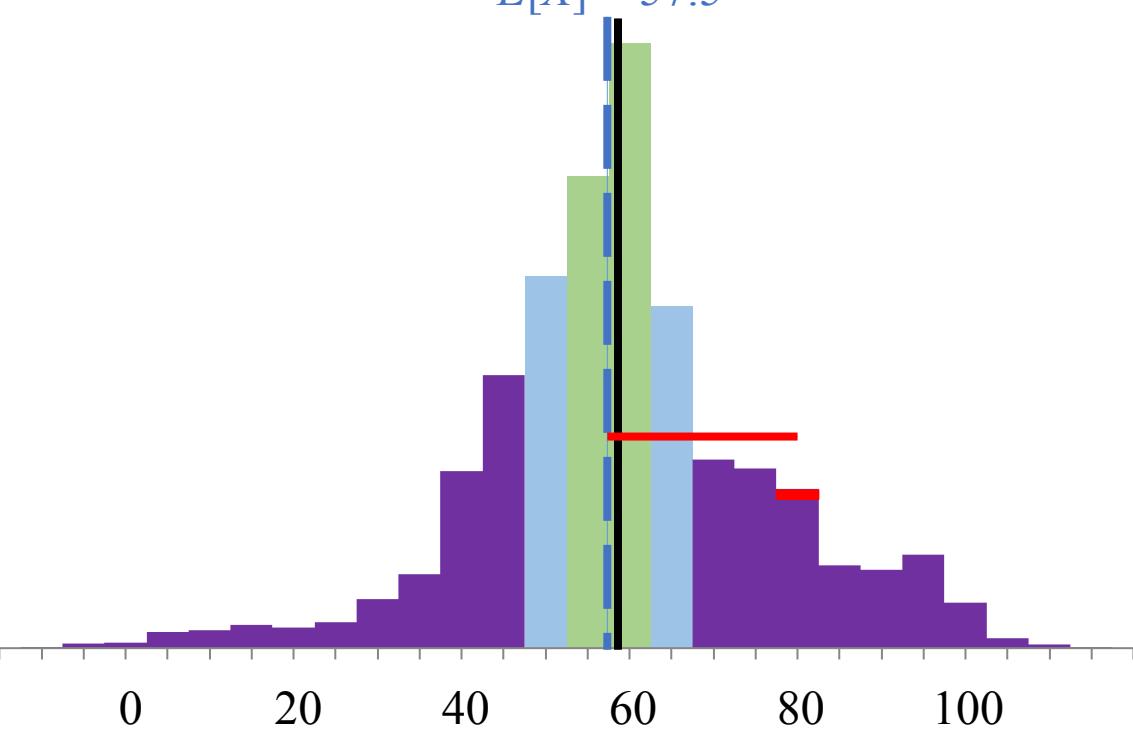


# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58  
 $E[X] = 57.5$



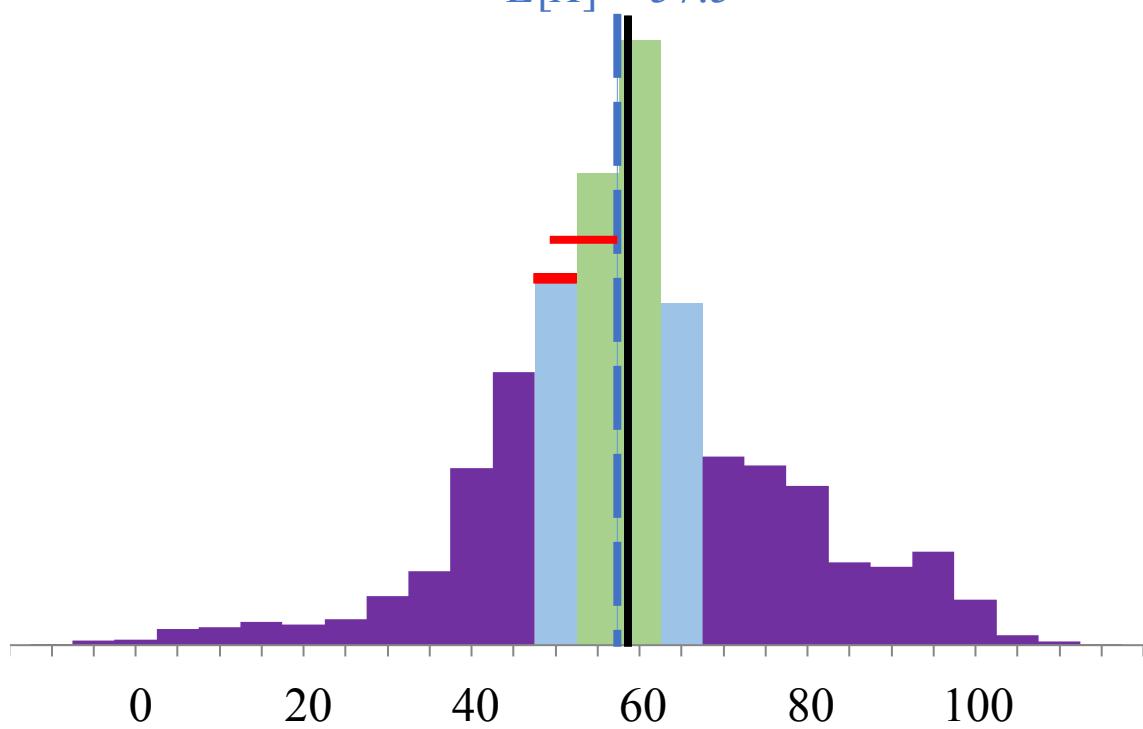
$X$	$(X - \mu)^2$
25 points	1056 points <sup>2</sup>
80 points	506 points <sup>2</sup>

# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58  
 $E[X] = 57.5$



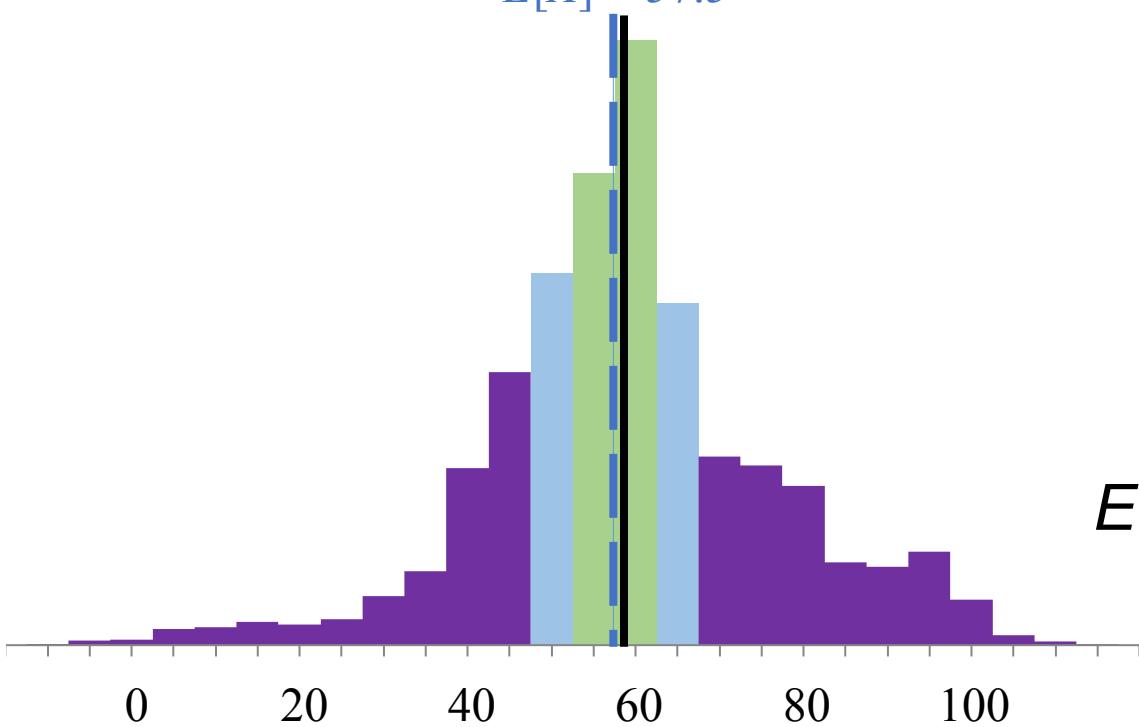
$X$	$(X - \mu)^2$
25 points	1056 points <sup>2</sup>
80 points	506 points <sup>2</sup>
50 points	56 points <sup>2</sup>

# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58  
 $E[X] = 57.5$



$X$	$(X - \mu)^2$
25 points	1056 points <sup>2</sup>
80 points	506 points <sup>2</sup>
50 points	56 points <sup>2</sup>
...	

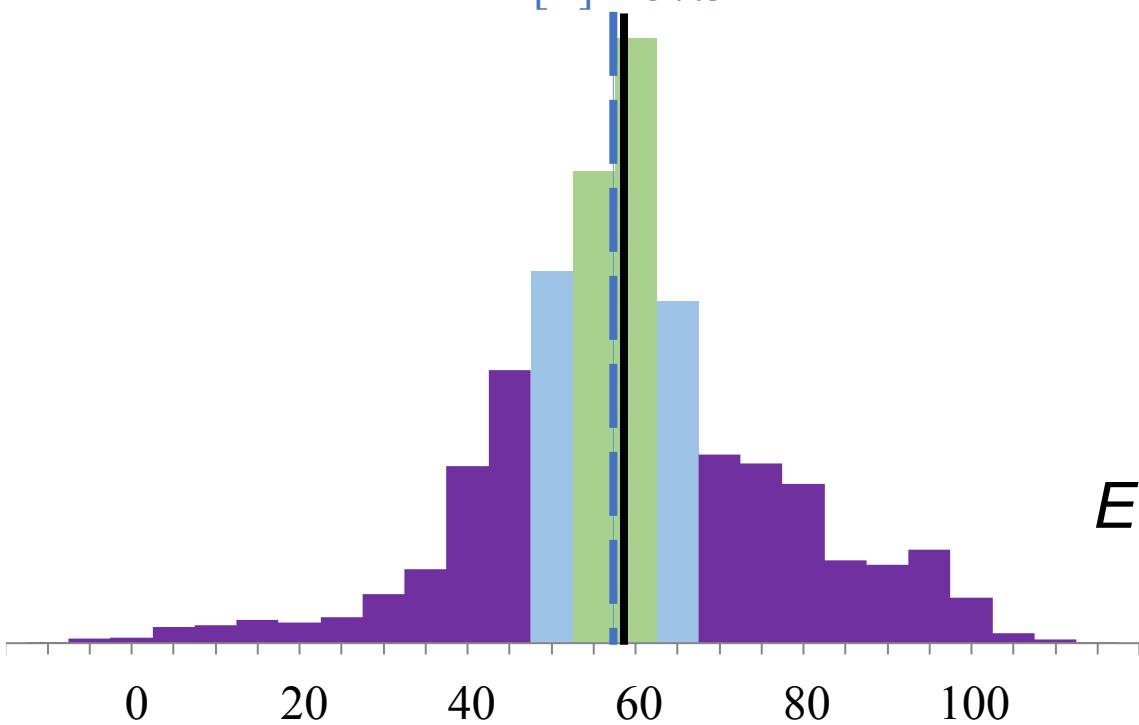
$$E [(X - \mu)^2] = 52 \text{ points}^2$$

# Peer Grades in Coursera HCI

Let  $X$  be a random variable that represents a peer grade

$$\text{Var}(X) = E[(X - \mu)^2]$$

True grade = 58  
 $E[X] = 57.5$



$X$	$(X - \mu)^2$
25 points	1056 points <sup>2</sup>
80 points	506 points <sup>2</sup>
50 points	56 points <sup>2</sup>
...	

$$E [(X - \mu)^2] = 52 \text{ points}^2$$
$$Std(X) = 7.2 \text{ points}$$

# Variance

- If  $X$  is a random variable with mean  $\mu$  then the **variance** of  $X$ , denoted  $\text{Var}(X)$ , is:

$$\text{Var}(X) = E[(X - \mu)^2]$$

- Note:  $\text{Var}(X) \geq 0$
- Also known as the 2nd **Central Moment**, or square of the Standard Deviation

# Computing Variance

$$\begin{aligned}\text{Var}(X) &= E[(X - \mu)^2] \\ &= \sum_x (x - \mu)^2 p(x)\end{aligned}$$

Note:  $\mu = E[X]$

$$\begin{aligned}&= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\ &= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x)\end{aligned}$$

$$= \boxed{E[X^2]} - 2\mu E[X] + \mu^2$$

$$= E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2$$

$$= \boxed{E[X^2] - (E[X])^2}$$

Ladies and gentlemen, please welcome the 2<sup>nd</sup> moment!

# Variance of a 6 sided dice

- Let  $X$  = value on roll of 6 sided die
- Recall that  $E[X] = 7/2$
- Compute  $E[X^2]$

$$E[X^2] = (1^2)\frac{1}{6} + (2^2)\frac{1}{6} + (3^2)\frac{1}{6} + (4^2)\frac{1}{6} + (5^2)\frac{1}{6} + (6^2)\frac{1}{6} = \frac{91}{6}$$

$$\text{Var}(X) = E[X^2] - (E[X])^2$$

$$= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

# Properties of Variance

- $\text{Var}(aX + b) = a^2 \text{Var}(X)$

- Proof:

$$\begin{aligned}\text{Var}(aX + b) &= E[(aX + b)^2] - (E[aX + b])^2 \\ &= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\ &= a^2E[X^2] + 2abE[X] + b^2 - (a^2(E[X])^2 + 2abE[X] + b^2) \\ &= a^2E[X^2] - a^2(E[X])^2 = a^2(E[X^2] - (E[X])^2) \\ &= a^2\text{Var}(X)\end{aligned}$$

- Standard Deviation of  $X$ , denoted  $\text{SD}(X)$ , is:

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

- $\text{Var}(X)$  is in units of  $X^2$
  - $\text{SD}(X)$  is in same units as  $X$

# Fundamental Properties

Semantic  
Meaning

$$P(X=x)$$

Random  
Variable

$$E[X]$$

$$\text{Var}(X)$$

$$\text{Std}(X)$$

$$E[X^2]$$

# Lots of fun with Random Variables

# Classics



# Jacob Bernoulli

- Jacob Bernoulli (1654-1705), also known as “James”, was a Swiss mathematician



- One of many mathematicians in Bernoulli family
- The Bernoulli Random Variable is named for him
- He is my *academic* great<sup>12</sup>-grandfather
- Same eyes as Ice Cube

# Bernoulli Random Variable

- Experiment results in “Success” or “Failure”
  - $X$  is random **indicator** variable ( $1 = \text{success}$ ,  $0 = \text{failure}$ )
  - $P(X = 1) = p(1) = p$        $P(X = 0) = p(0) = 1 - p$
  - $X$  is a Bernoulli Random Variable:  $X \sim \text{Ber}(p)$
  - $E[X] = p$
  - $\text{Var}(X) = p(1 - p)$
- Examples
  - coin flip
  - random binary digit
  - whether a disk drive crashed
  - whether someone likes a netflix movie

Feel the Bern!

# Does a Program Crash?



Run a program, crashes with prob.  $p$ , works with prob.  $(1 - p)$

$\mathbf{X}$ : 1 if program crashes

$$P(\mathbf{X} = 1) = p$$

$$P(\mathbf{X} = 0) = 1 - p$$

$$\mathbf{X} \sim \text{Ber}(p)$$

# Does a User Click an Ad?



Serve an ad, clicked with prob.  $p$ , ignored with prob.  $(1 - p)$

**C**: 1 if ad is clicked

$$P(\mathbf{C} = 1) = p$$

$$P(\mathbf{C} = 0) = 1 - p$$

$$\mathbf{C} \sim \text{Ber}(p)$$

More!

# Binomial Random Variable

- Consider  $n$  independent trials of  $\text{Ber}(p)$  rand. var.
  - $X$  is number of successes in  $n$  trials
  - $X$  is a **Binomial** Random Variable:  $X \sim \text{Bin}(n, p)$

$$P(X = i) = p(i) = \binom{n}{i} p^i (1-p)^{n-i} \quad i = 0, 1, \dots, n$$

- By Binomial Theorem, we know that  $\sum_{i=0}^{\infty} P(X = i) = 1$
- Examples
  - # of heads in  $n$  coin flips
  - # of 1's in randomly generated length  $n$  bit string
  - # of disk drives crashed in 1000 computer cluster
    - Assuming disks crash independently

# Bernoulli vs Binomial



Bernoulli is an indicator RV



Binomial is the sum of  $n$  Bernoullis

# Three Coin Flips

- Three fair (“heads” with  $p = 0.5$ ) coins are flipped
  - $X$  is number of heads
  - $X \sim \text{Bin}(n = 3, p = 0.5)$

$$P(X = 0) = \binom{3}{0} p^0 (1-p)^3 = \frac{1}{8}$$

$$P(X = 1) = \binom{3}{1} p^1 (1-p)^2 = \frac{3}{8}$$

$$P(X = 2) = \binom{3}{2} p^2 (1-p)^1 = \frac{3}{8}$$

$$P(X = 3) = \binom{3}{3} p^3 (1-p)^0 = \frac{1}{8}$$

# Properties of Bin(n, p)

Consider:  $X \sim \text{Bin}(n, p)$

- $P(X = i) = p(i) = \binom{n}{i} p^i (1 - p)^{n-i} \quad i = 0, 1, \dots, n$
- $E[X] = np$
- $\text{Var}(X) = np(1 - p)$
- Note:  $\text{Ber}(p) = \text{Bin}(1, p)$

# I Really Want the Proof of Var :)

$$\begin{aligned} E(X^2) &= \sum_{k \geq 0}^n k^2 \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=0}^n kn \binom{n-1}{k-1} p^k q^{n-k} \\ &= np \sum_{k=1}^n k \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \\ &= np \sum_{j=0}^m (j+1) \binom{m}{j} p^j q^{m-j} \\ &= np \left( \sum_{j=0}^m j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^m \binom{m}{j} p^j q^{m-j} \right) \\ &= np \left( \sum_{j=0}^m m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^m \binom{m}{j} p^j q^{m-j} \right) \\ &= np \left( (n-1)p \sum_{j=1}^m \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^m \binom{m}{j} p^j q^{m-j} \right) \\ &= np ((n-1)p(p+q)^{m-1} + (p+q)^m) \\ &= np ((n-1)p + 1) \\ &= n^2 p^2 + np(1-p) \end{aligned}$$

Definition of Binomial Distribution:  $p + q = 1$

Factors of Binomial Coefficient:  $k \binom{n}{k} = n \binom{n-1}{k-1}$

Change of limit: term is zero when  $k - 1 = 0$

putting  $j = k - 1, m = n - 1$

splitting sum up into two

Factors of Binomial Coefficient:  $j \binom{m}{j} = m \binom{m-1}{j-1}$

Change of limit: term is zero when  $j - 1 = 0$

Binomial Theorem

as  $p + q = 1$

by algebra

# How Many Program Crashes?



$n$  runs of program, each crashes with prob.  $p$ , works with prob.  $(1 - p)$

**H**: number of crashes

$$\mathbf{H} \sim \text{Bin}(n, p)$$

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

# How Many Ads Clicked?



1000 ads served, each clicked with  $p = 0.01$ , otherwise ignored.

**H:** number of clicks

$$\mathbf{H} \sim \text{Bin}(n = 1000, p = 0.01)$$

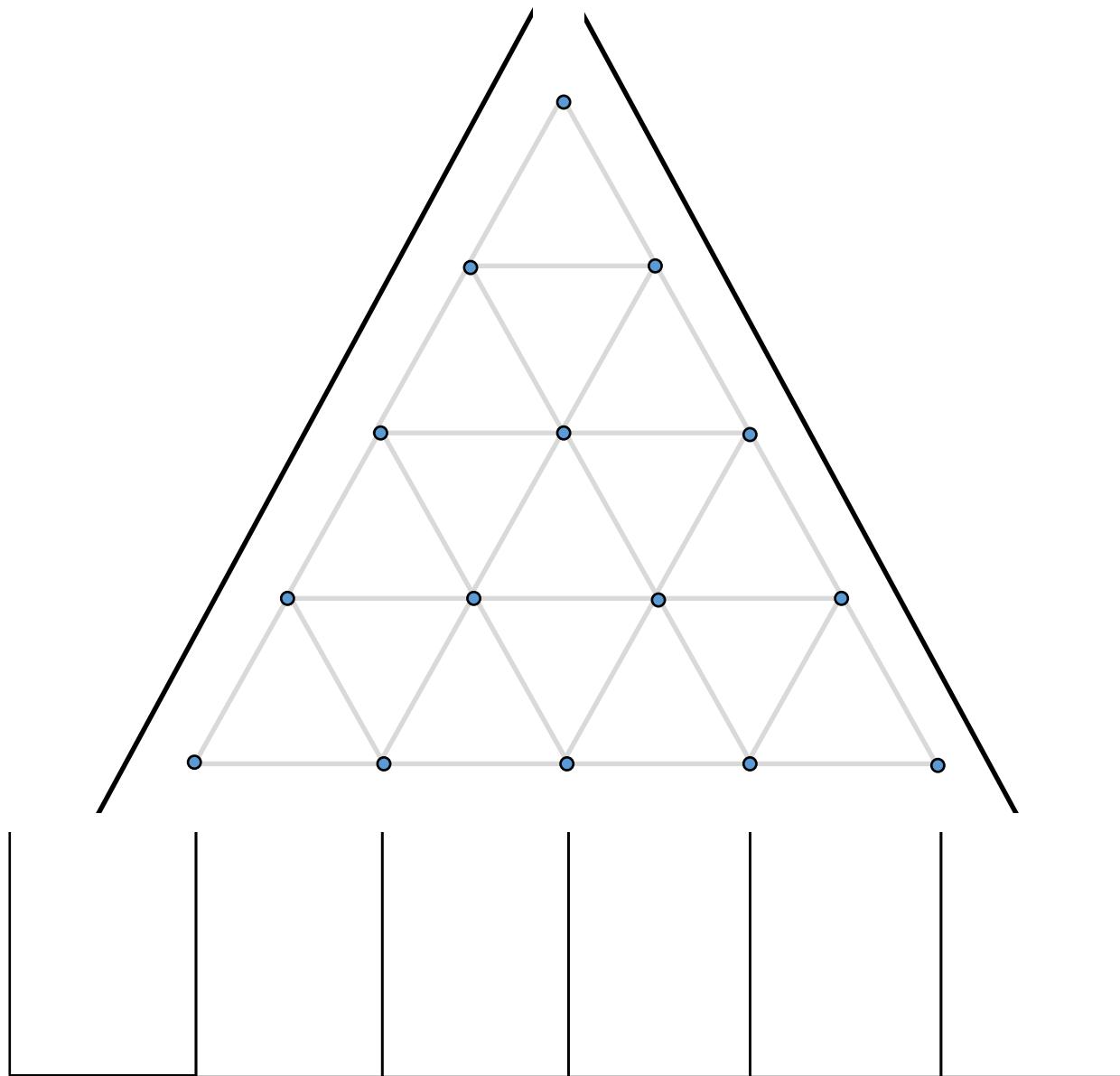
$$\mathbf{P}(\mathbf{H} = k) = \binom{1000}{k} (0.01)^k (0.99)^{1000-k}$$

Variance of number of ads clicked?

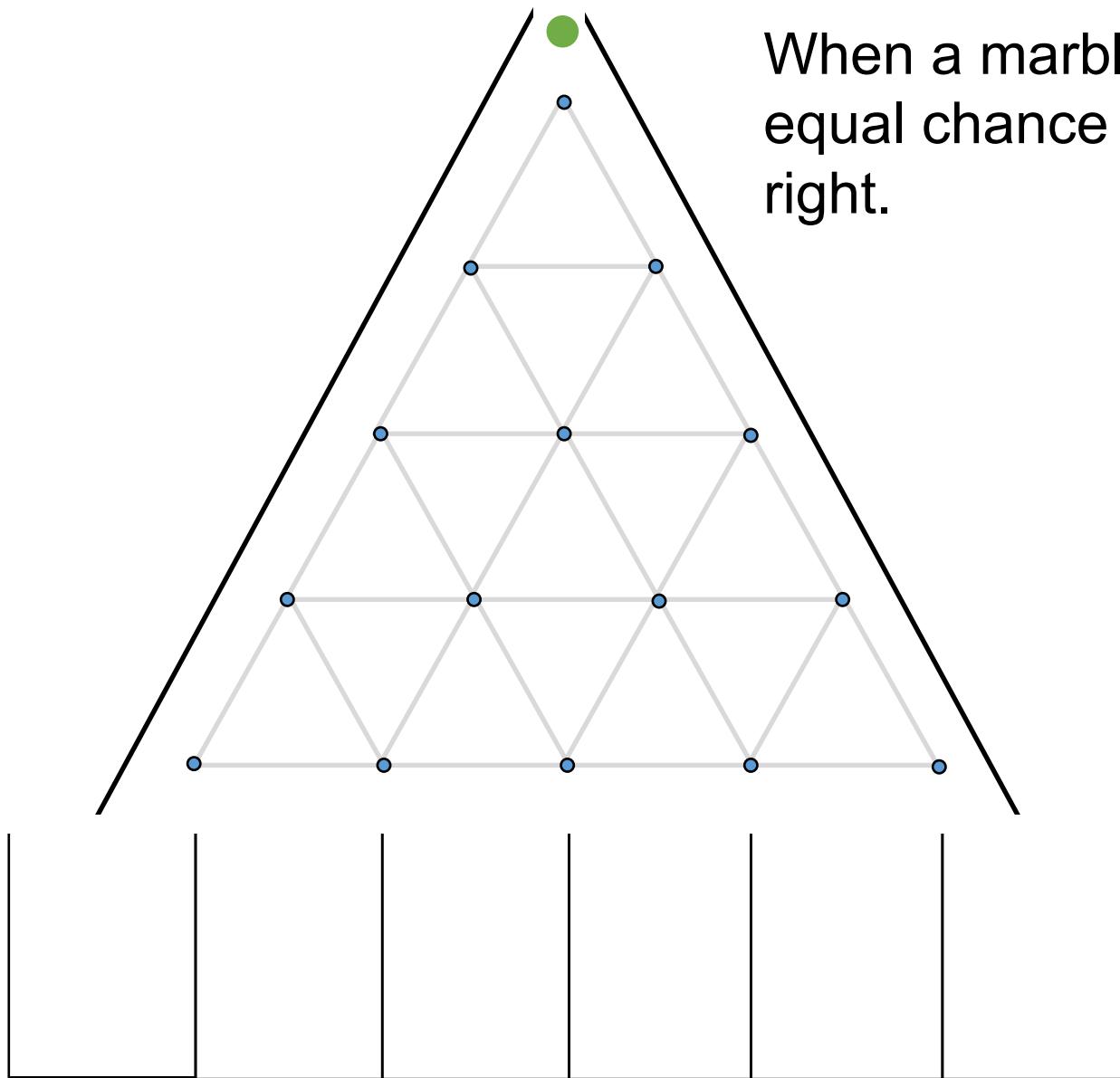
$$\mathbf{E}[\mathbf{H}] = np = 10$$

$$\text{Var}(\mathbf{H}) = np(1-p) = 9.9 \quad \text{Std}(\mathbf{H}) = 3.15$$

# Galton Board

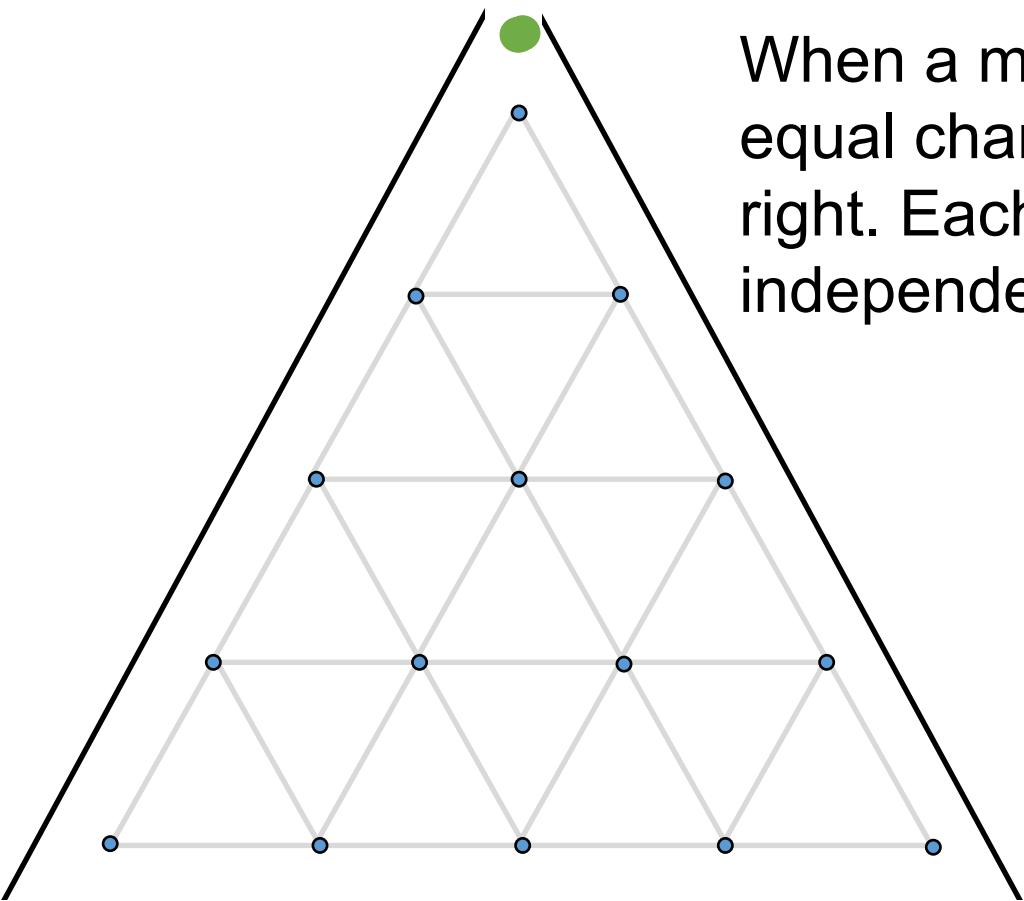


# Galton Board



When a marble hits a pin, it has equal chance of going left or right.

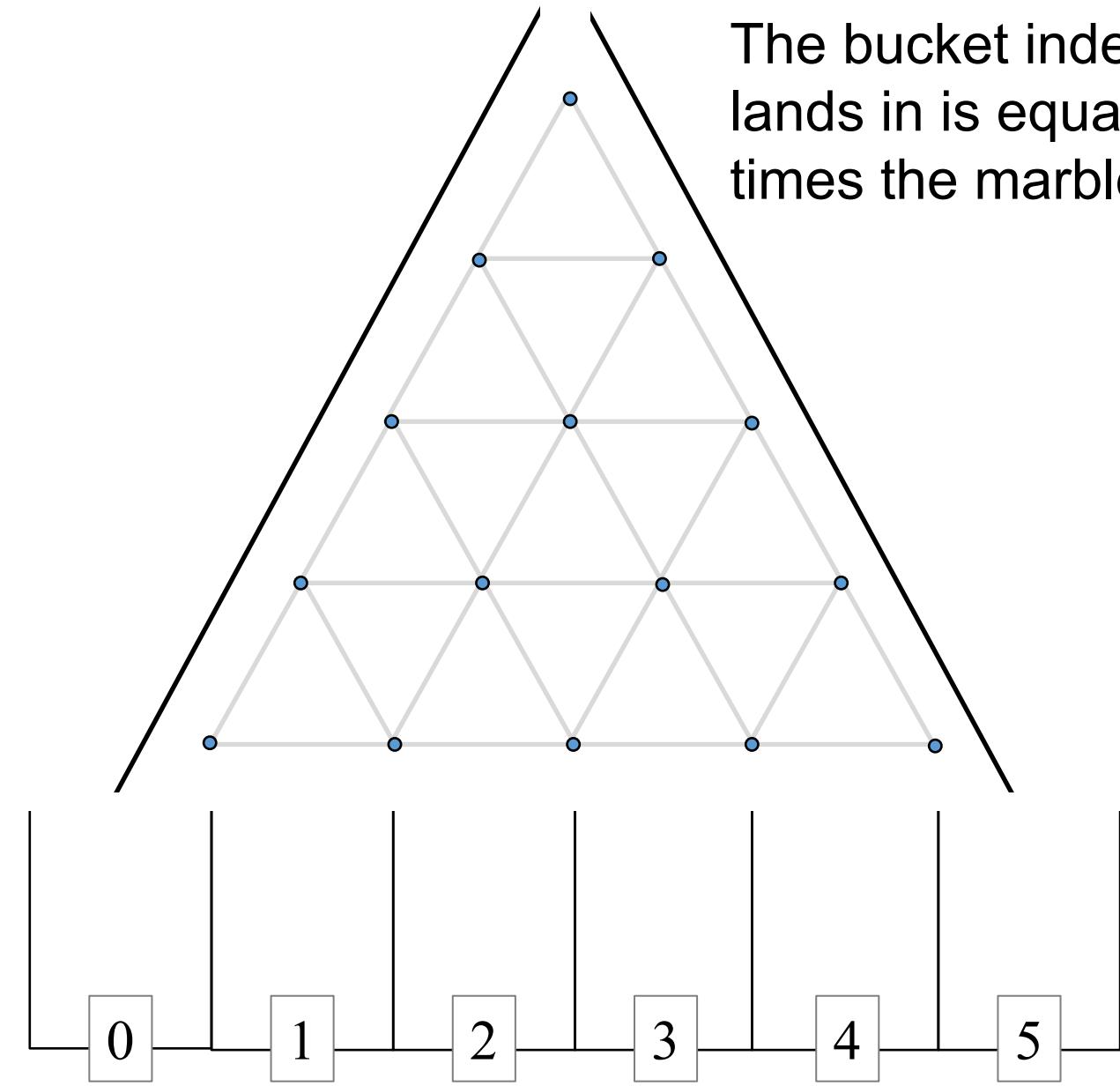
# Galton Board



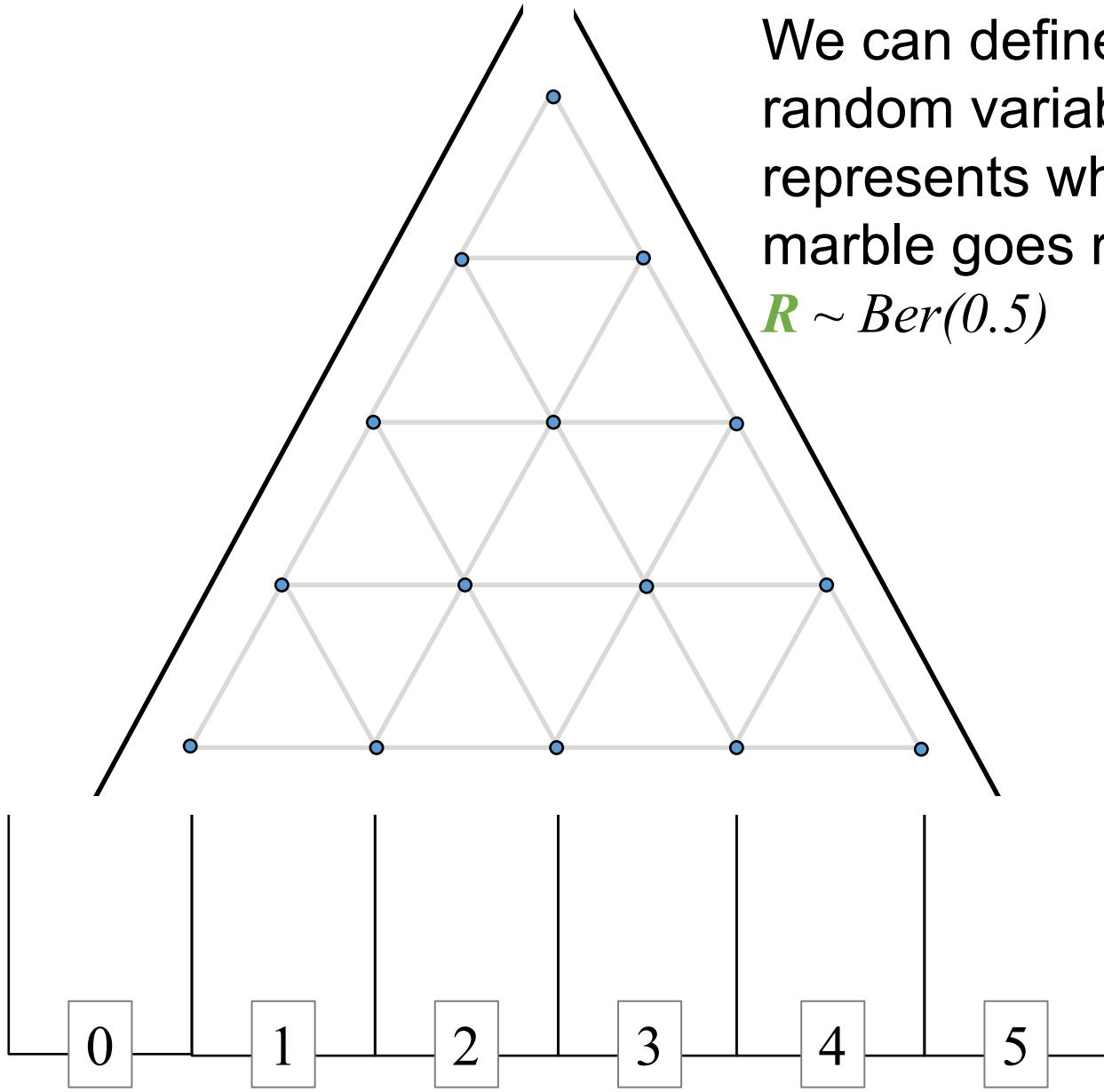
When a marble hits a pin, it has equal chance of going left or right. Each pin represents an independent event.

# Galton Board

The bucket index that a marble lands in is equal to the number of times the marble went right

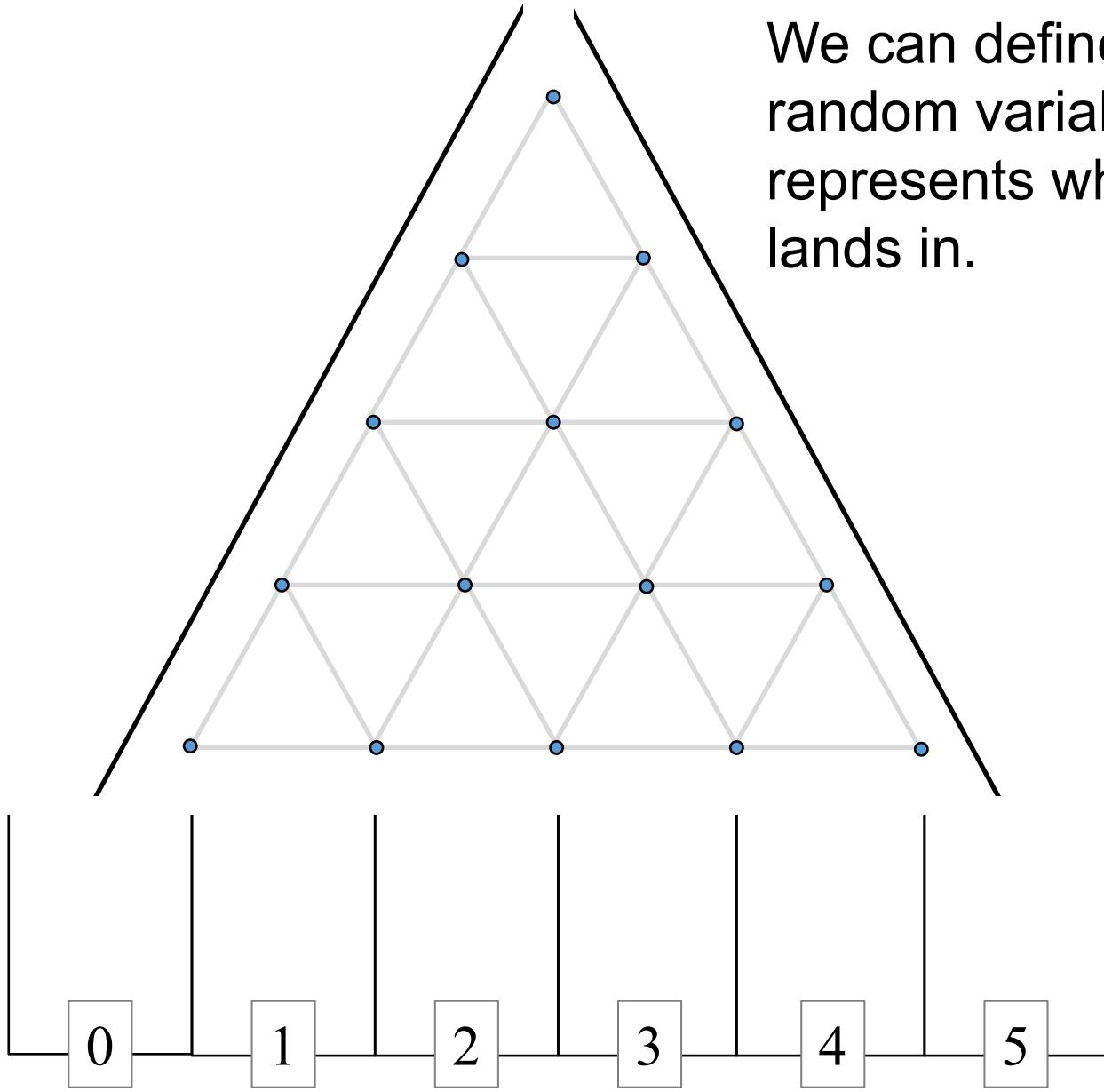


# Galton Board



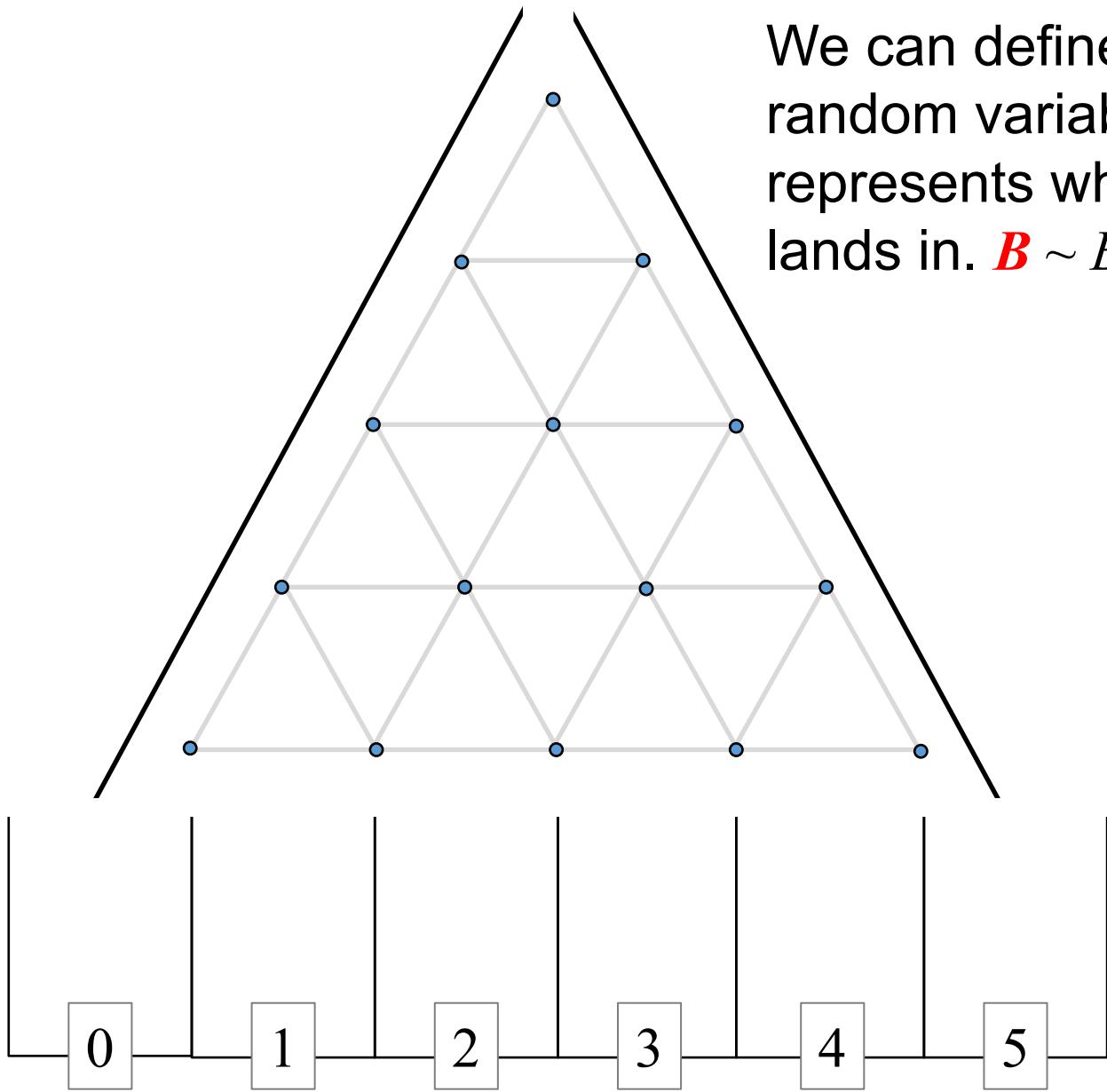
We can define an indicator random variable ( $R$ ) which represents whether a particular marble goes right as a Bernoulli  
 $R \sim Ber(0.5)$

# Galton Board



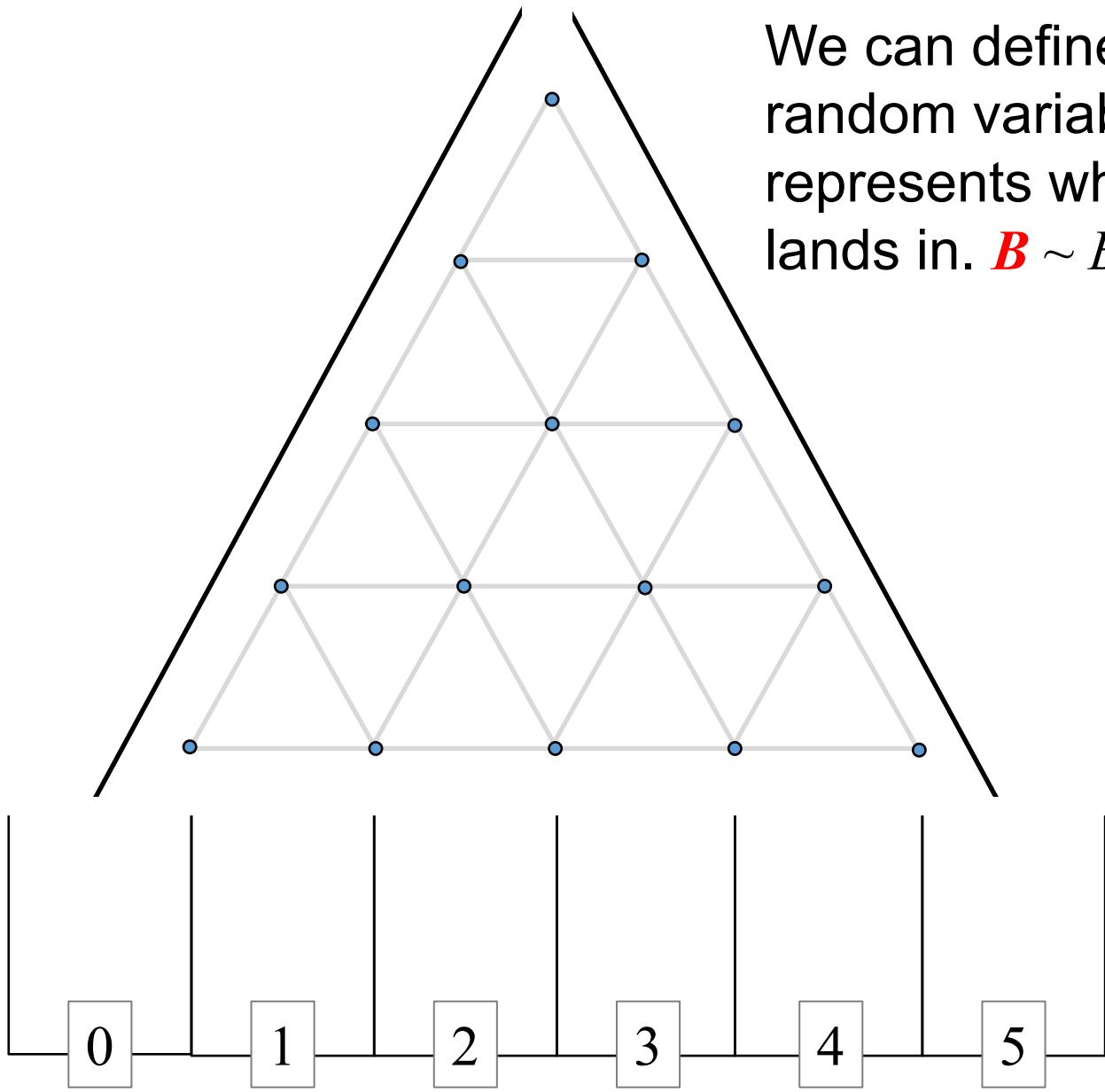
We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.

# Galton Board



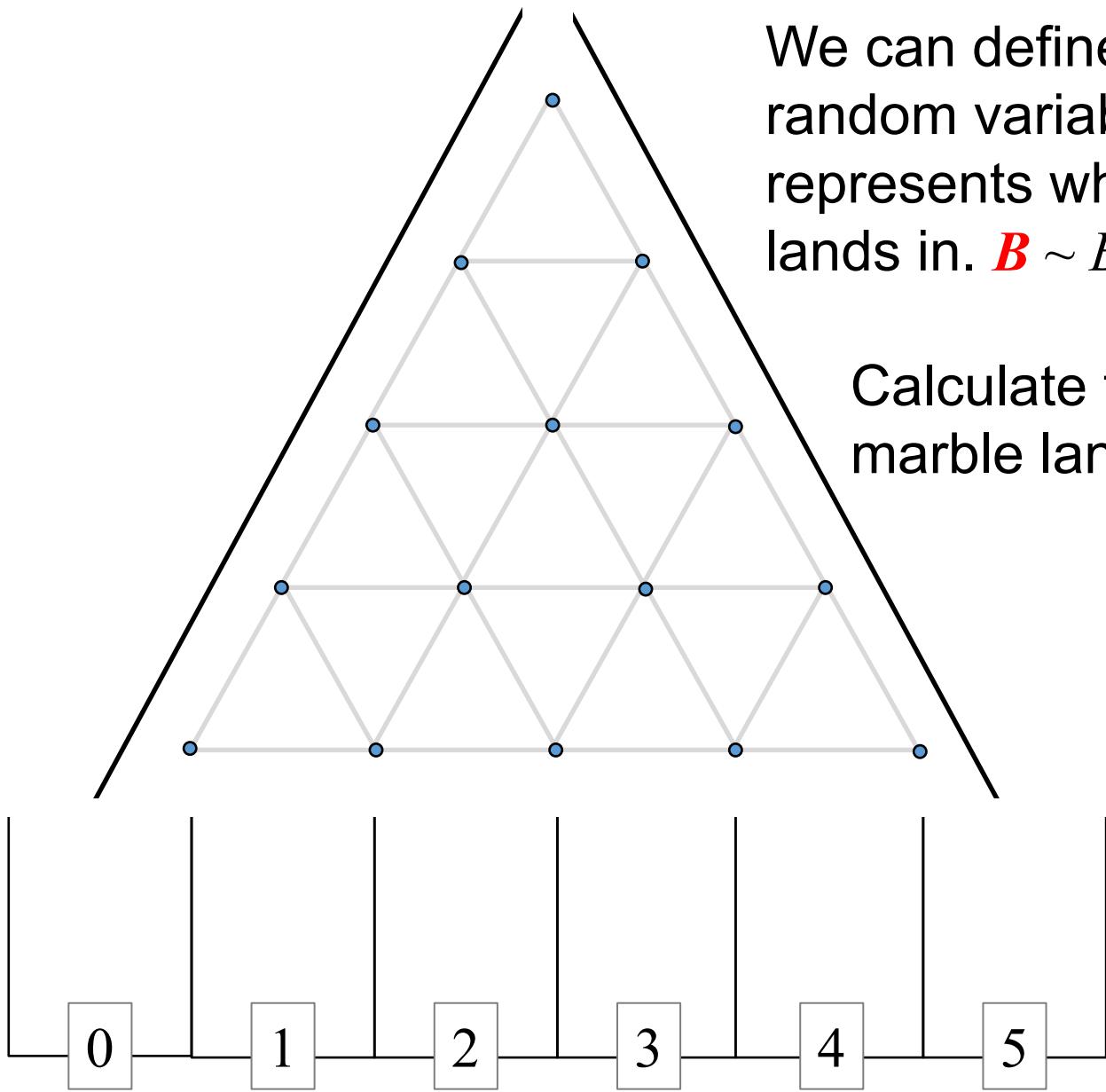
We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim \text{Bin}(\text{levels}, 0.5)$

# Galton Board



We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

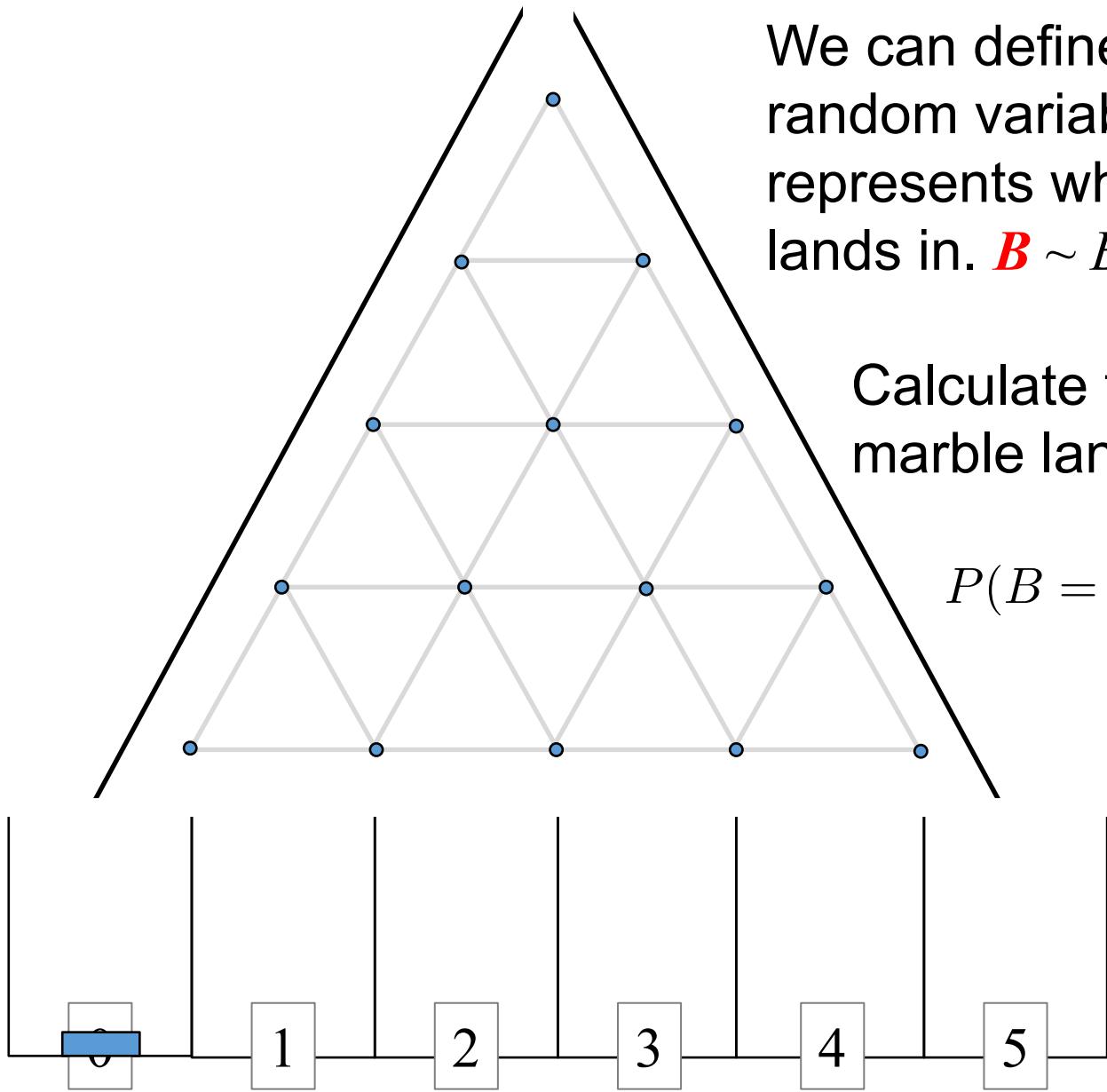
# Galton Board



We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

# Galton Board

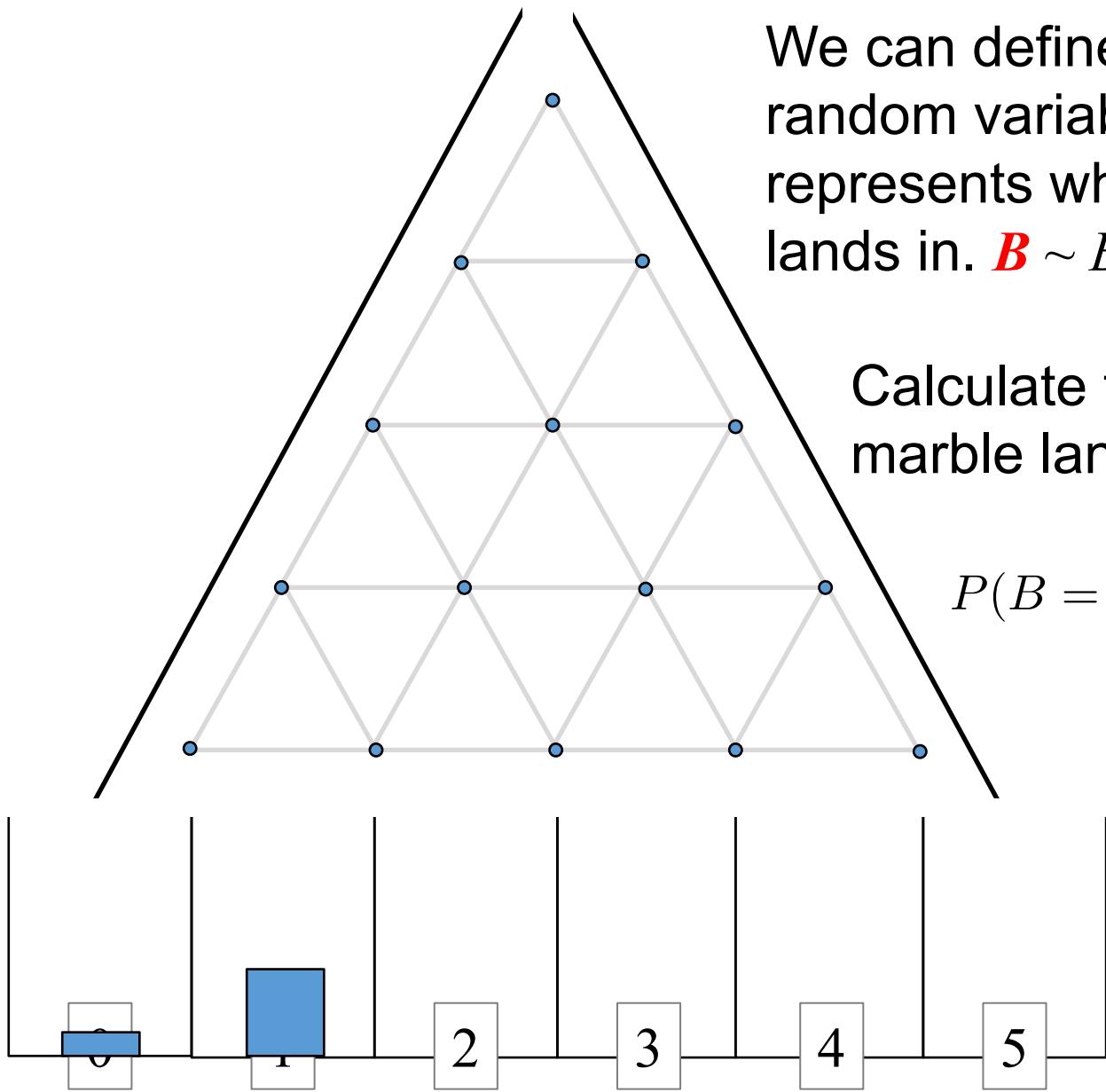


We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

$$P(B = 0) = \binom{5}{0} \frac{1}{2}^5 \approx 0.03$$

# Galton Board

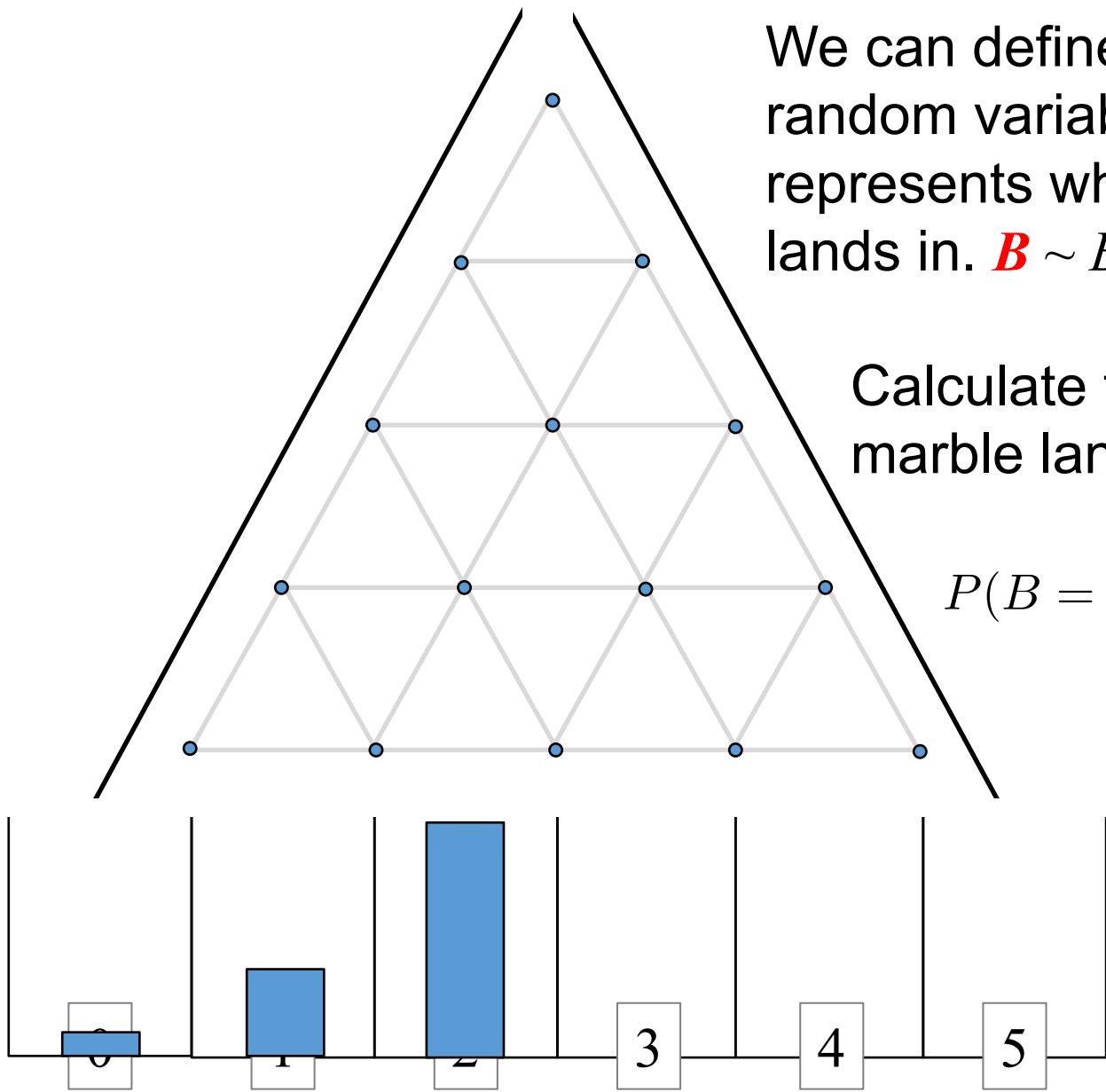


We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

$$P(B = 1) = \binom{5}{1} \frac{1}{2}^5 \approx 0.16$$

# Galton Board

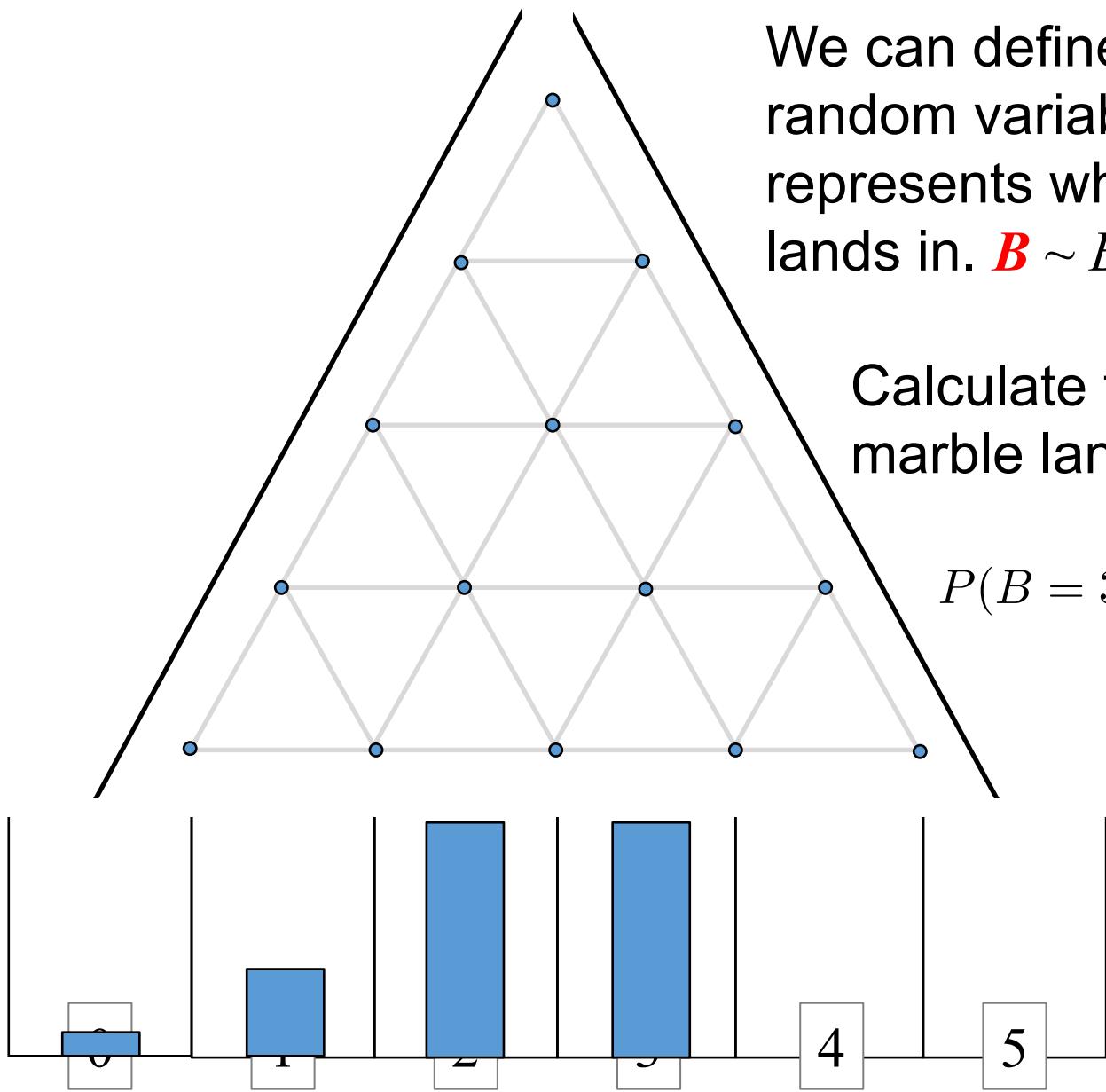


We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

$$P(B = 2) = \binom{5}{2} \frac{1}{2}^5 \approx 0.31$$

# Galton Board

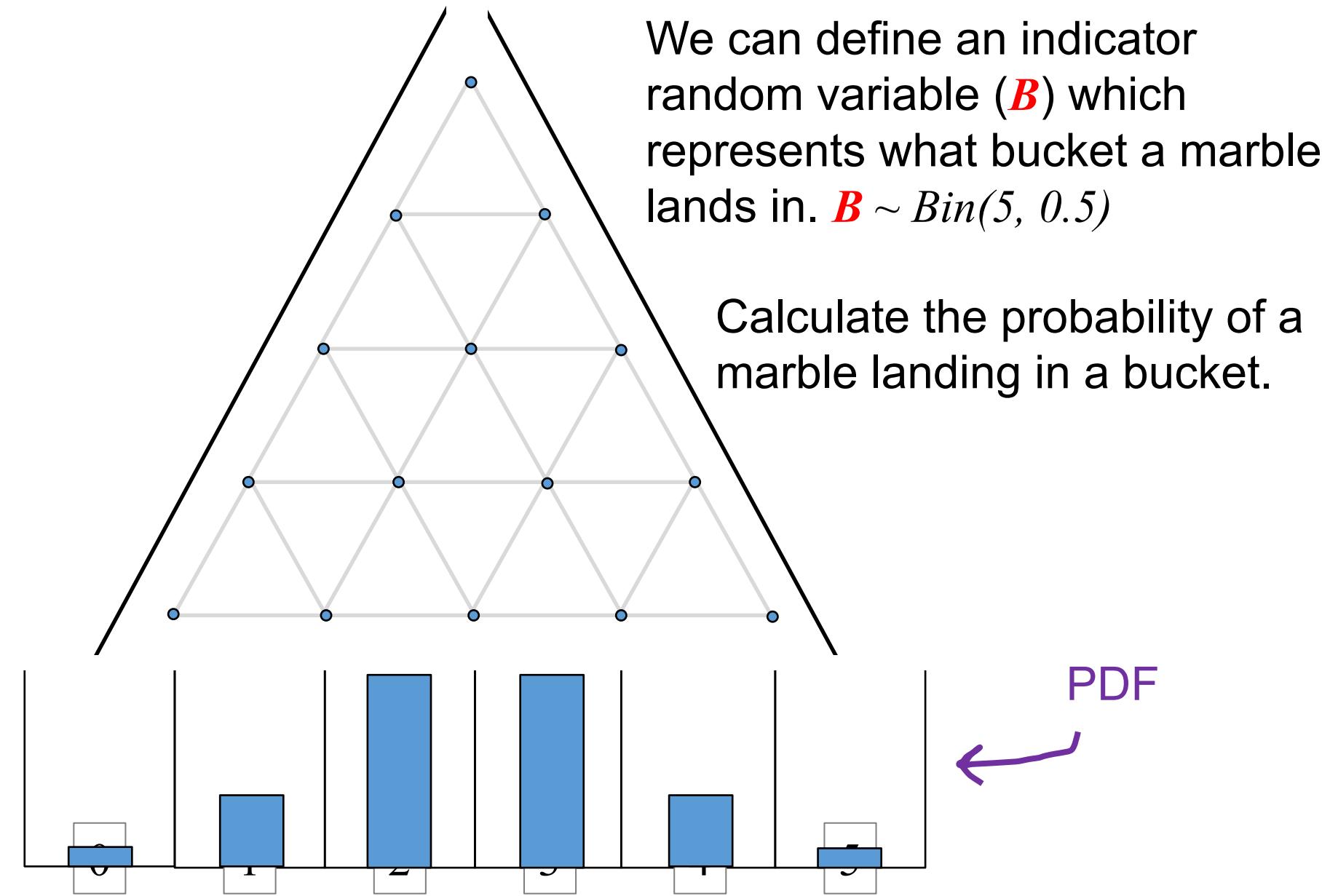


We can define an indicator random variable ( $B$ ) which represents what bucket a marble lands in.  $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

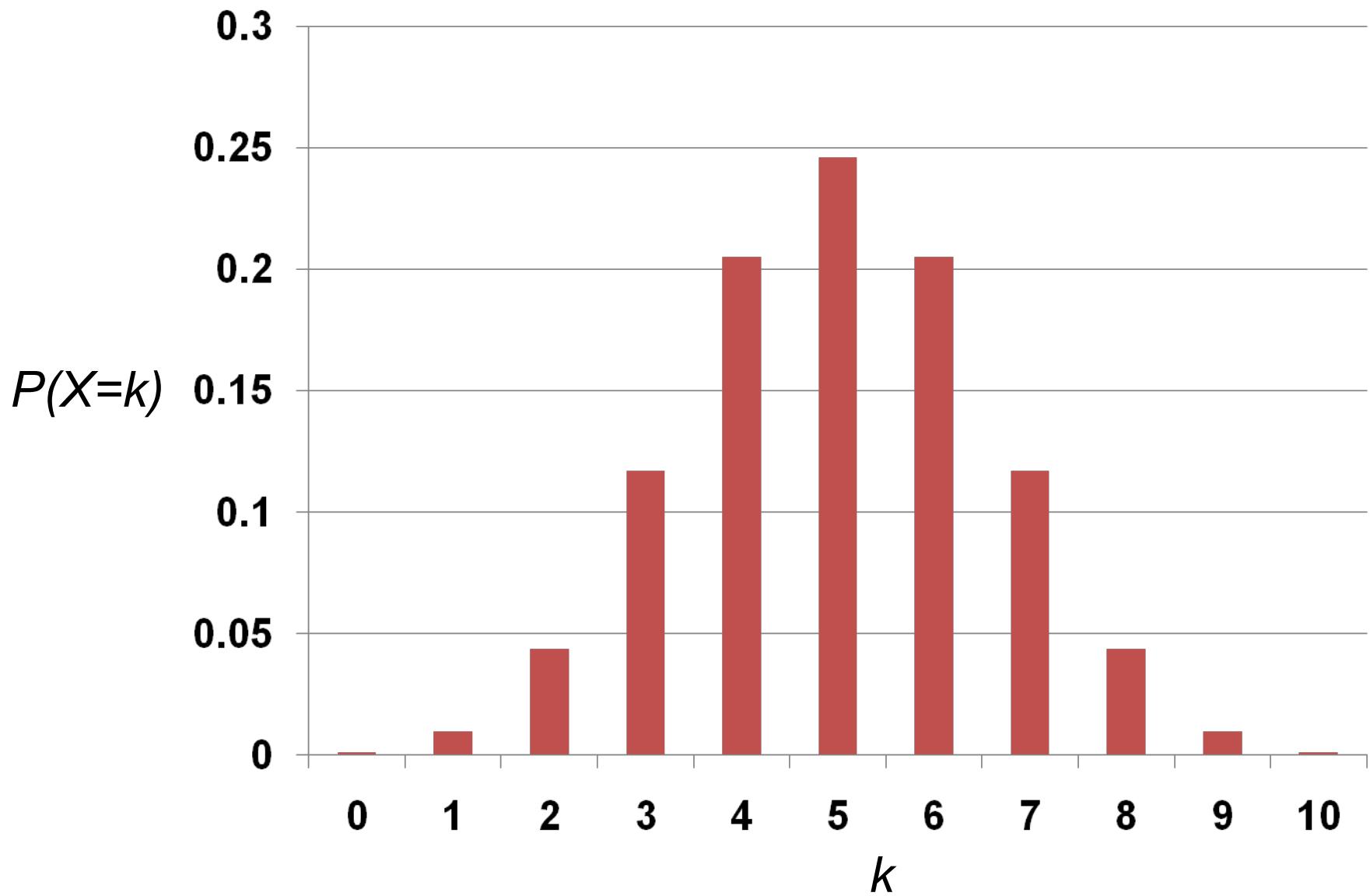
$$P(B = 3) = \binom{5}{2} \frac{1}{2}^5 \approx 0.31$$

# Galton Board

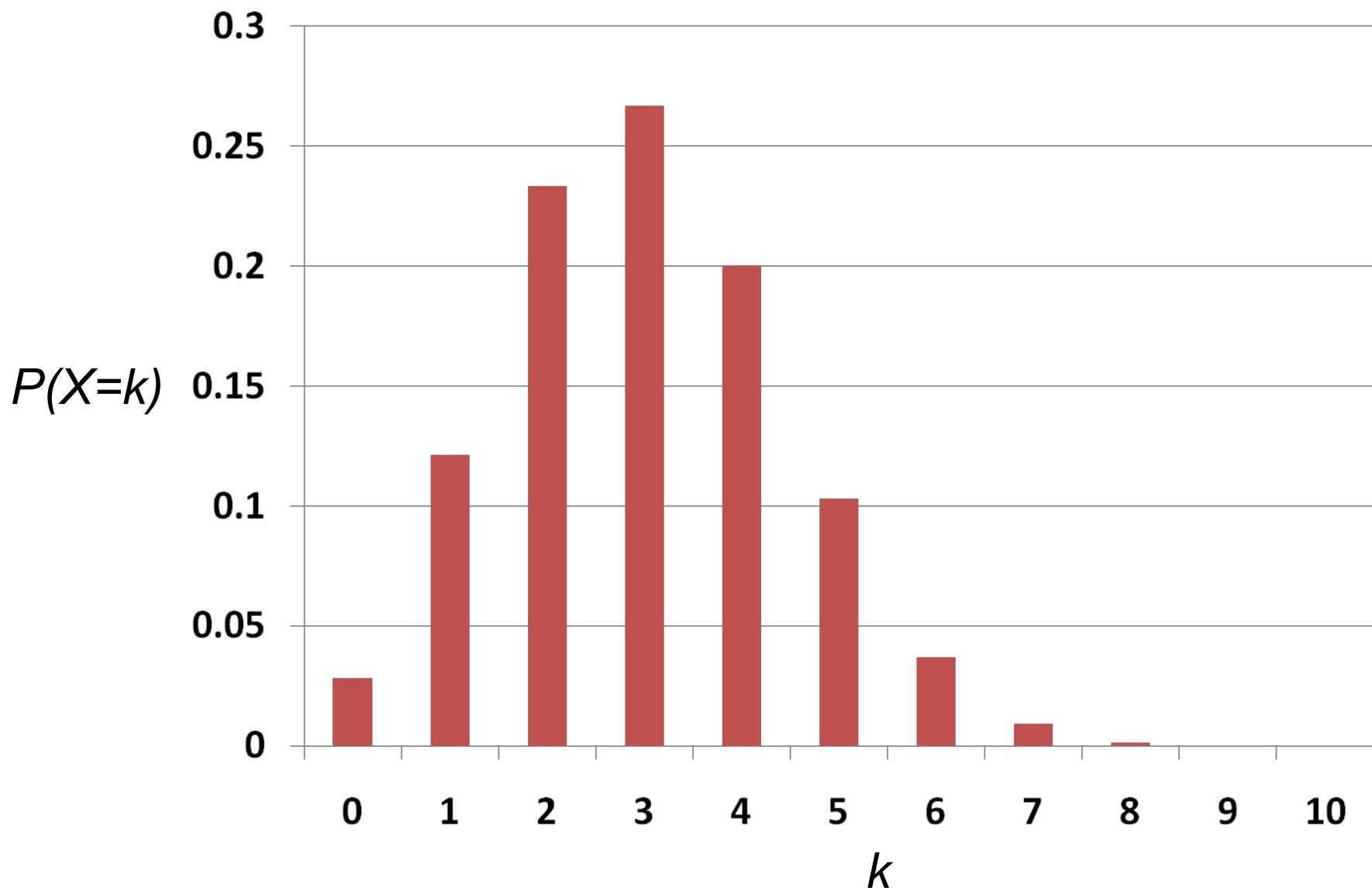


*FROM CHAOS TO ORDER*

# PMF for $X \sim \text{Bin}(10, 0.5)$



# PMF for $X \sim \text{Bin}(10, 0.3)$



# Genetic Inheritance

- Person has 2 genes for trait (eye color)
  - Child receives 1 gene (equally likely) from each parent
  - Child has brown eyes if either (or both) genes brown
  - Child only has blue eyes if both genes blue
  - Brown is “dominant” (d) , Blue is “recessive” (r)
  - Parents each have 1 brown and 1 blue gene
- 4 children, what is  $P(3 \text{ children with brown eyes})$ ?
  - Child has blue eyes:  $p = (\frac{1}{2})(\frac{1}{2}) = \frac{1}{4}$  (2 blue genes)
  - $P(\text{child has brown eyes}) = 1 - (\frac{1}{4}) = 0.75$
  - $X = \# \text{ of children with brown eyes. } X \sim \text{Bin}(4, 0.75)$

$$P(X = 3) = \binom{4}{3} (0.75)^3 (0.25)^1 \approx 0.4219$$

# Power of Your Vote

- Is it better to vote in small or large state?
  - Small: more likely your vote changes outcome
  - Large: larger outcome (electoral votes) if state swings
  - $a (= 2n)$  voters equally likely to vote for either candidate
  - You are deciding  $(a + 1)^{\text{st}}$  vote

$$P(2n \text{ voters tie}) = \binom{2n}{n} \left(\frac{1}{2}\right)^n \left(\frac{1}{2}\right)^n = \frac{(2n)!}{n! n! 2^{2n}}$$

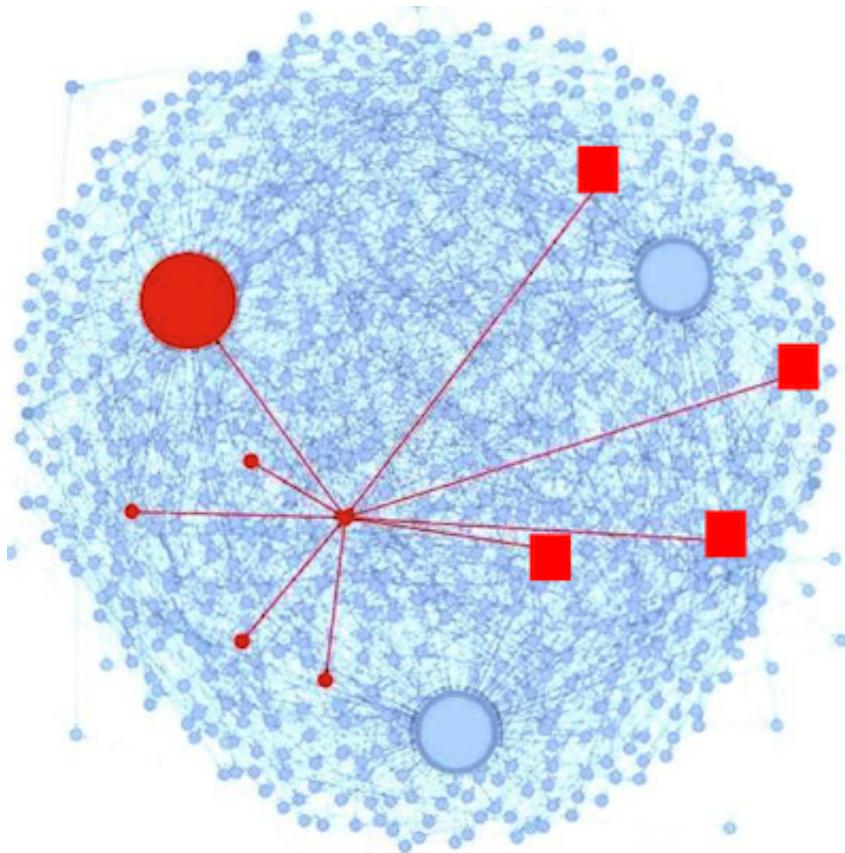
- Use Stirling's Approximation:  $n! \approx n^{n+1/2} e^{-n} \sqrt{2\pi}$

$$P(2n \text{ voters tie}) \approx \frac{(2n)^{2n+1/2} e^{-2n} \sqrt{2\pi}}{n^{2n+1} e^{-2n} 2\pi 2^{2n}} = \frac{1}{\sqrt{n\pi}}$$

- Power =  $P(\text{tie}) * \text{Elec. Votes} = \frac{1}{\sqrt{(a/2)\pi}} (ac) = \frac{c\sqrt{2a}}{\sqrt{\pi}}$
- Larger state = more power

# Is Peer Grading Accurate Enough?

*Looking ahead*

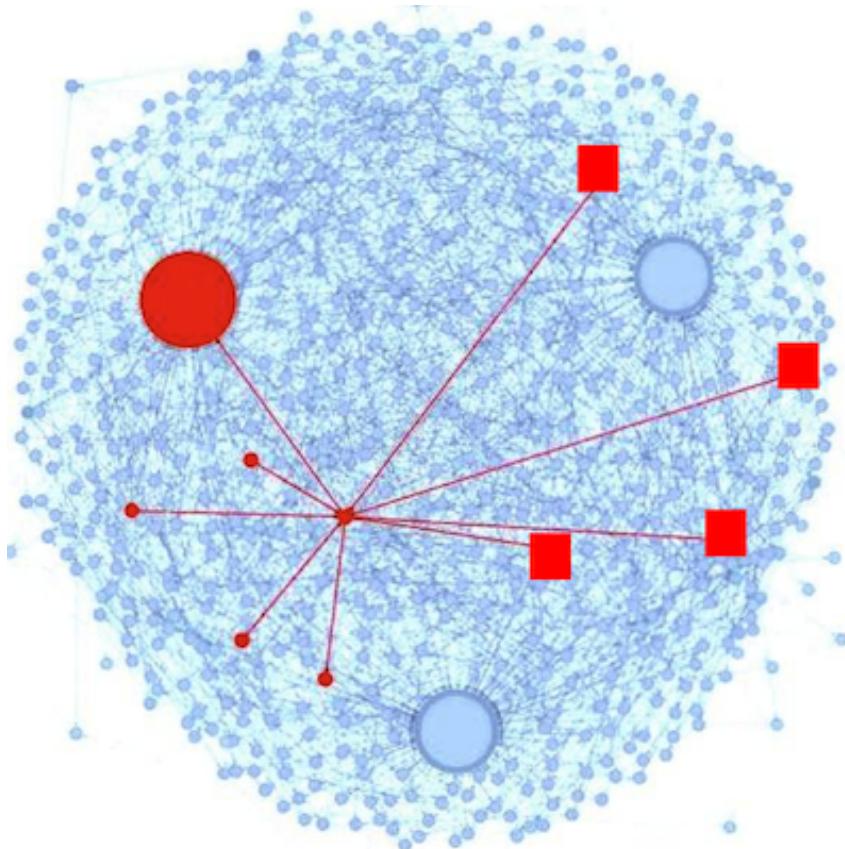


Peer Grading on Coursera  
HCI.

31,067 peer grades for  
3,607 students.

# Is Peer Grading Accurate Enough?

*Looking ahead*



1. Defined random variables for:
  - True grade ( $s_i$ ) for assignment  $i$
  - Observed ( $z_i^j$ ) score for assign  $i$
  - Bias ( $b_j$ ) for each grader  $j$
  - Variance ( $r_j$ ) for each grader  $j$
2. Designed a probabilistic model that defined the distributions for all random variables

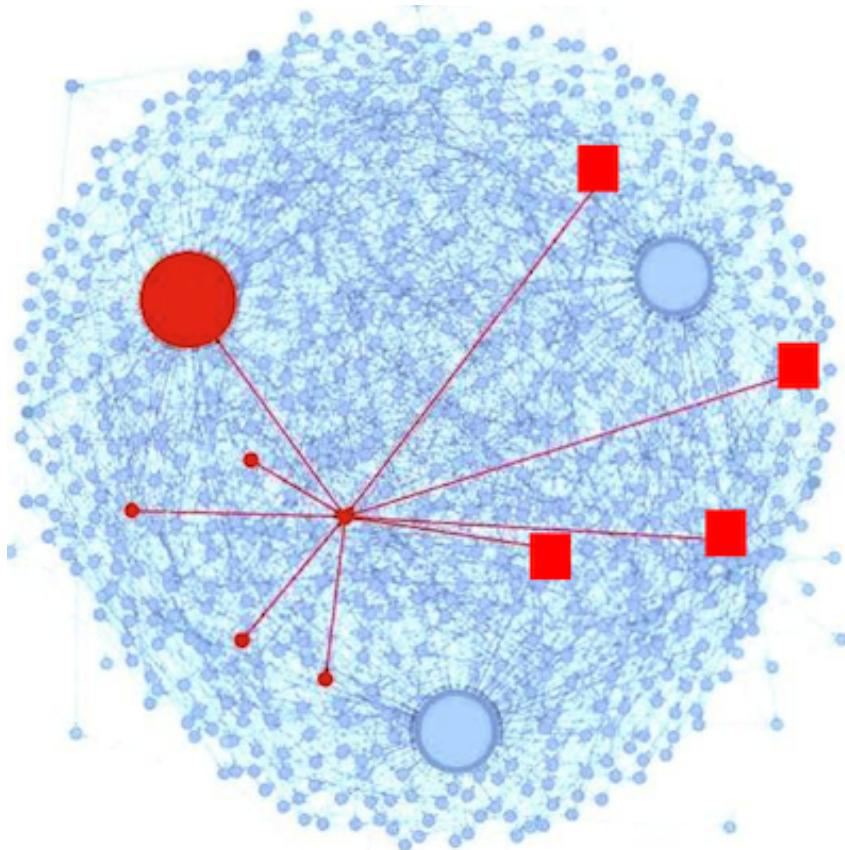
$$s_i \sim \text{Bin}(\text{points}, \theta)$$

*Problem param*

$$z_i^j \sim \mathcal{N}(\mu = s_i + b_j, \sigma = \sqrt{r_j})$$

# Is Peer Grading Accurate Enough?

*Looking ahead*

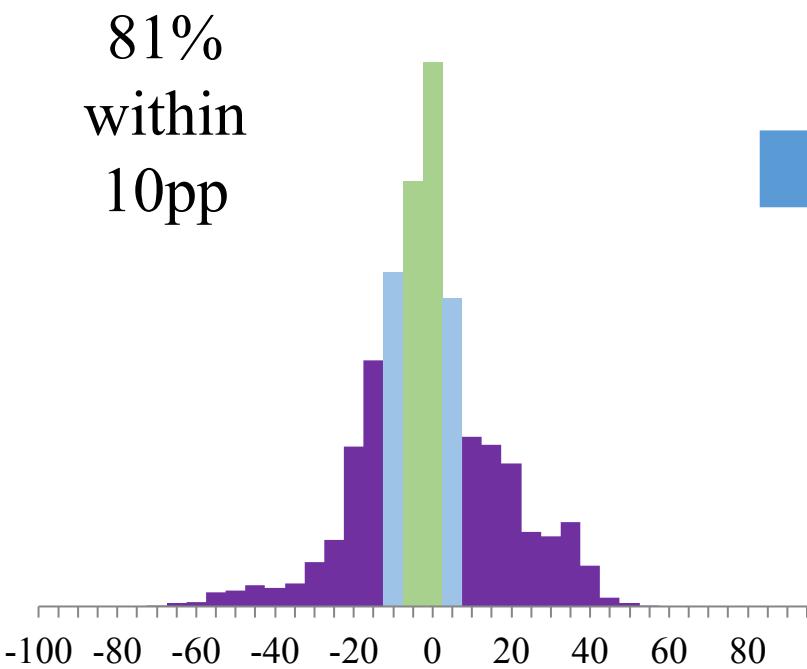


1. Defined random variables for:
  - True grade ( $s_i$ ) for assignment  $i$
  - Observed ( $z_i^j$ ) score for assign  $i$
  - Bias ( $b_j$ ) for each grader  $j$
  - Variance ( $r_j$ ) for each grader  $j$
2. Designed a probabilistic model that defined the distributions for all random variables
3. Found the variable assignments that maximized the probability of our observed data

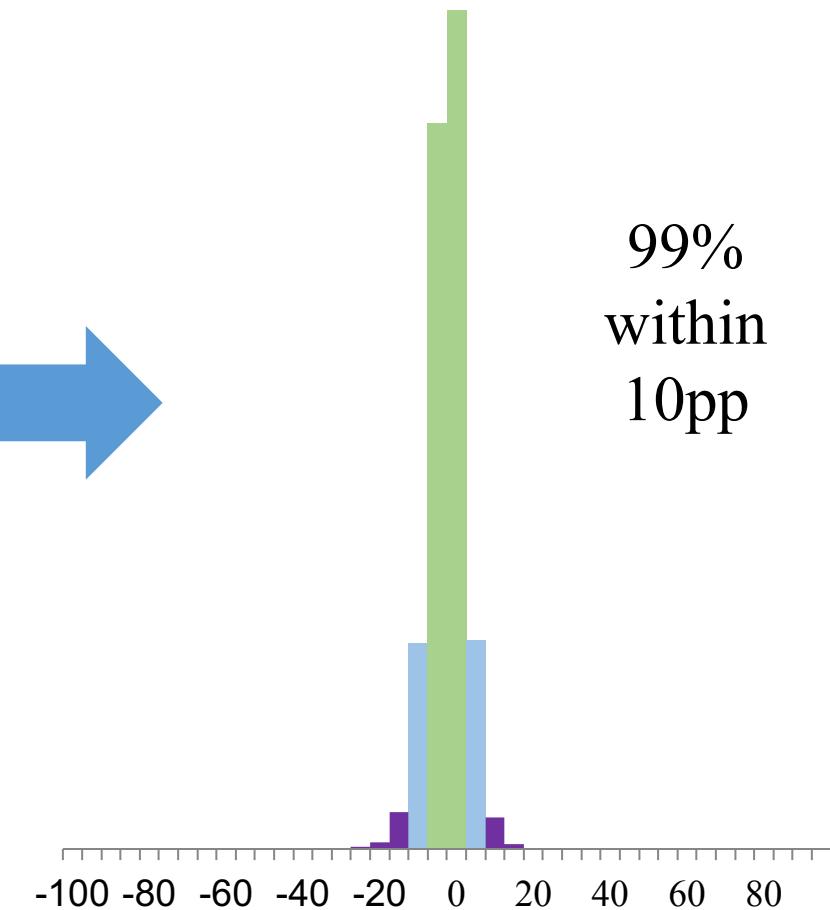
Inference or Machine Learning

# Yes, With Probabilistic Modelling

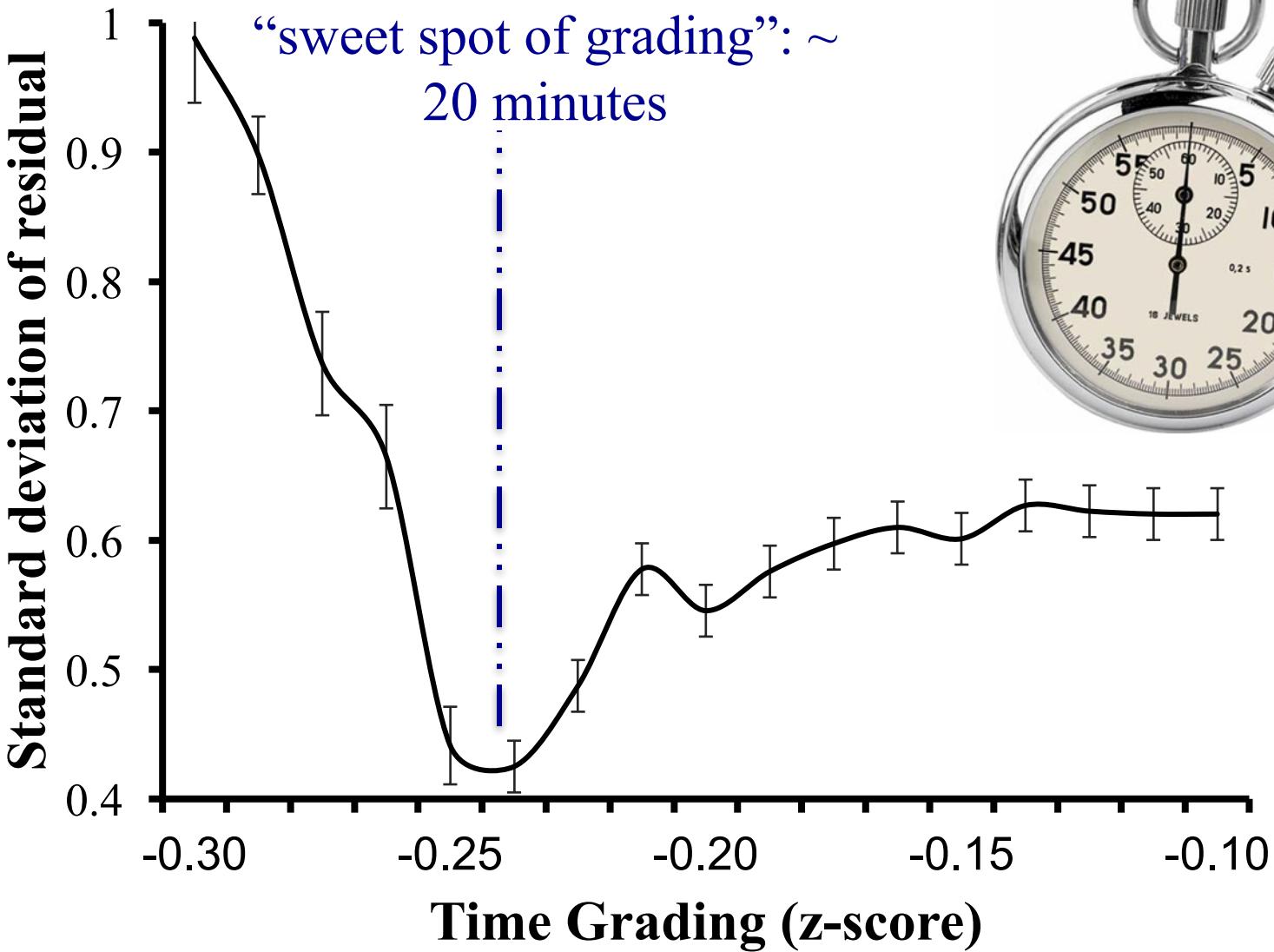
Before:



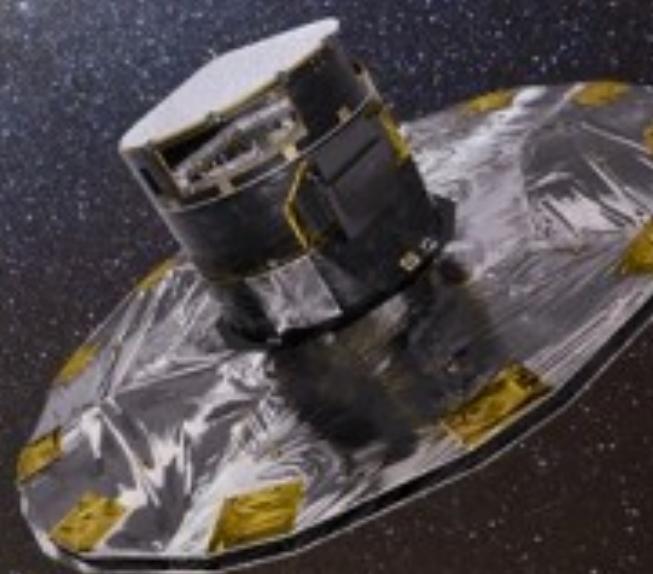
After:



# Grading Sweet Spot



1001



Voilà, c'est tout

