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*
  - $P(X \geq k) =$

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

Don't care about the rest

\# ways to choose slots for success

Probability that each is success

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

Not mutually exclusive…

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

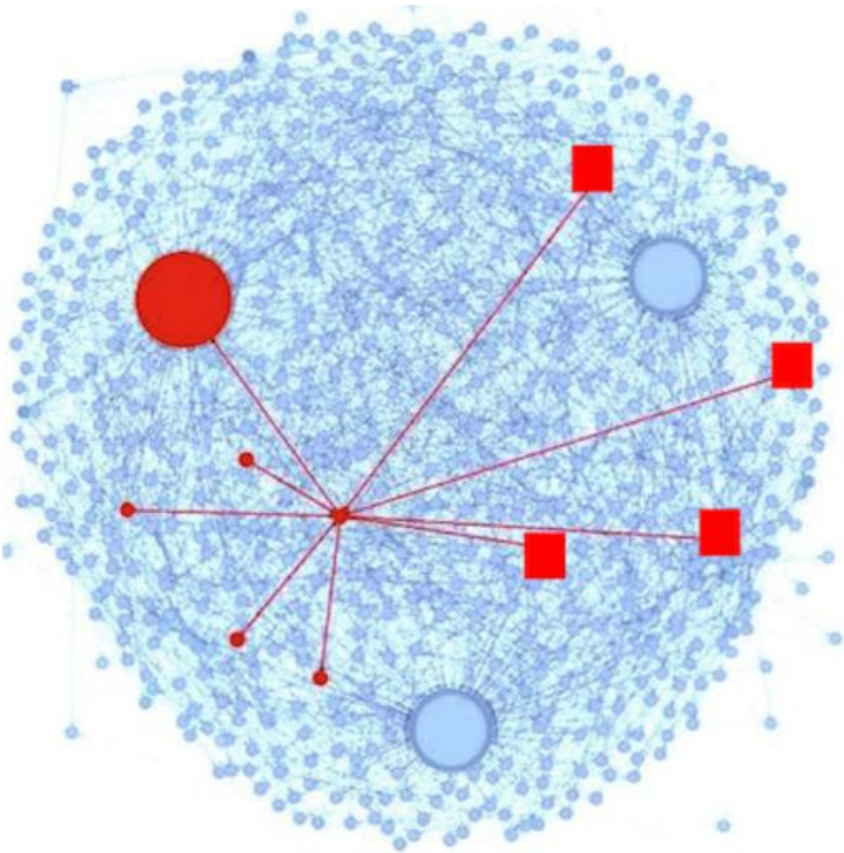
Variance

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.

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Review: Random Variables

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

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

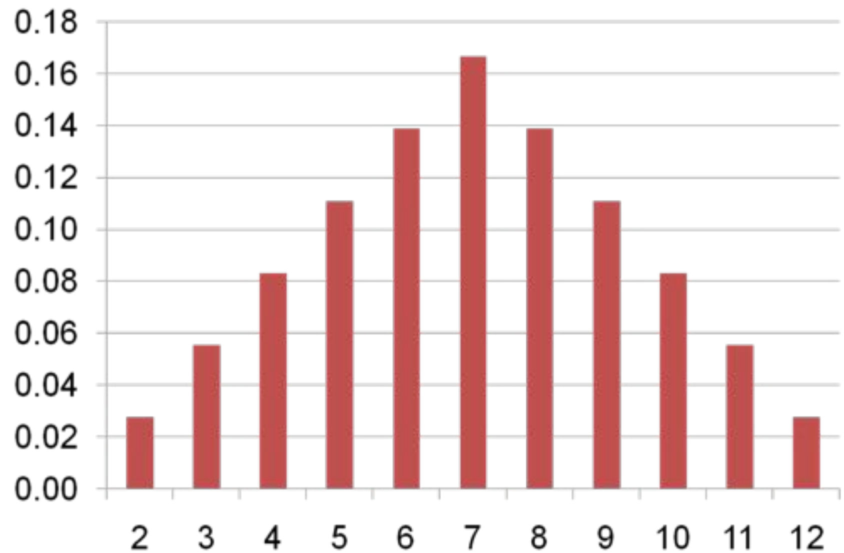
$$P(X = 2) = \frac{1}{36}$$

Review: Probability Mass Function

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

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

$P(Y = k)$

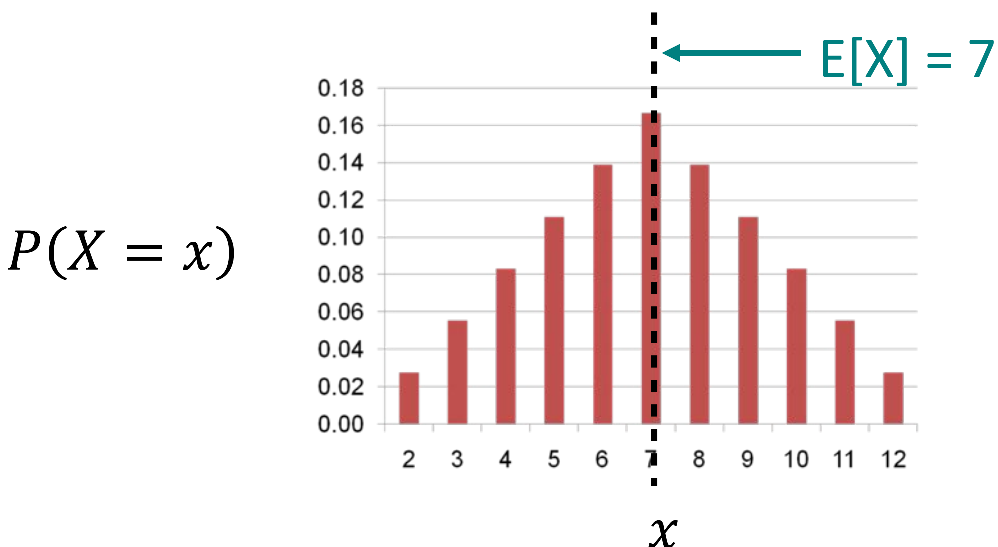


k

Review: Expectation

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

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



Properties of Expectation

- **Linearity**:

$$E[aX + b] = aE[X] + b$$

  - Consider X = 6-sided die roll, Y = 2X – 1.
  - E[X] = 3.5 E[Y] = 6

- **Expectation of a sum** is the sum of expectations

$$E[X + Y] = E[X] + E[Y]$$

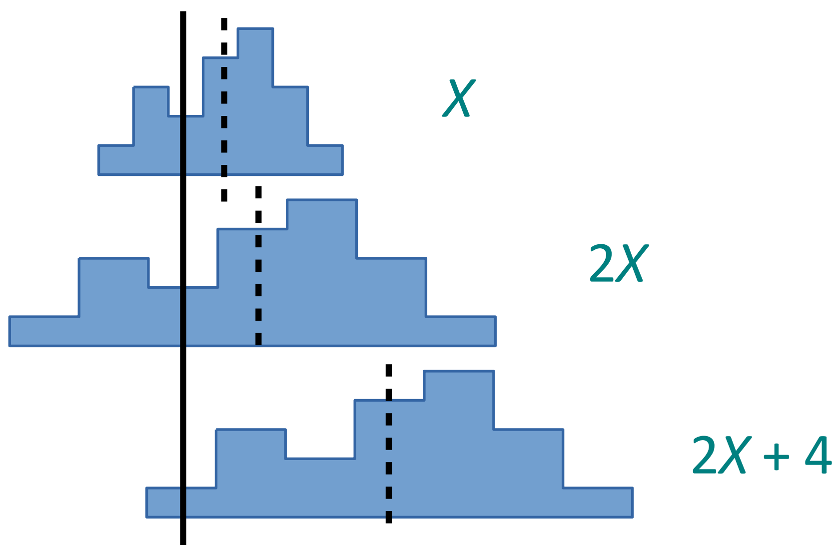
- **Unconscious statistician**:

$$E[g(X)] = \sum_x g(x)P(X = x)$$

Review: Linearity of Expectation

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

$$E[aX + b] = aE[X] + b$$



Review: Expectation of Sums

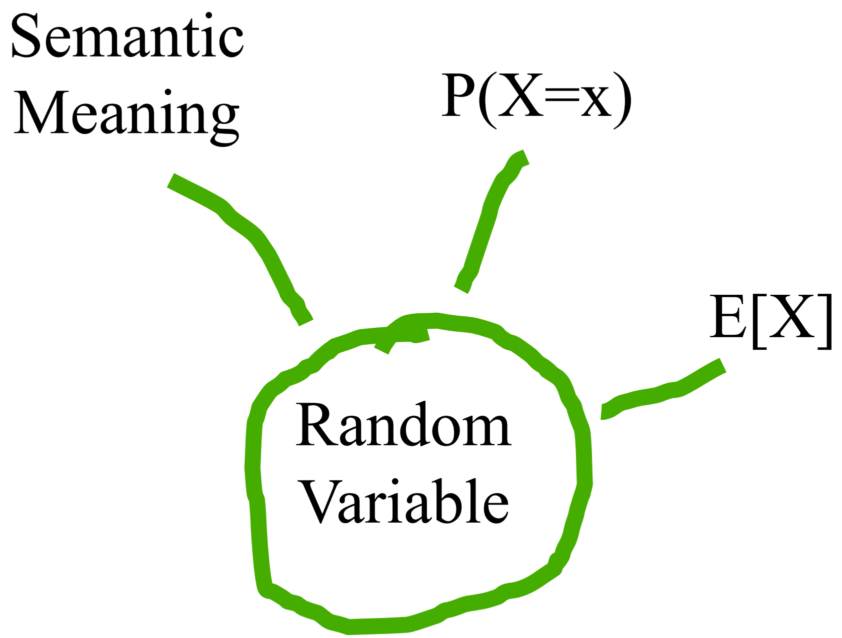
$$E[X + Y] = E[X] + E[Y]$$

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	-3	-2
1	0	1
5	9	14
4	1	5
...	...	...

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad + \quad \frac{1}{n}\sum_{i=1}^{n} y_i \quad = \quad \frac{1}{n}\sum_{i=1}^{n} (x_i + y_i)$$

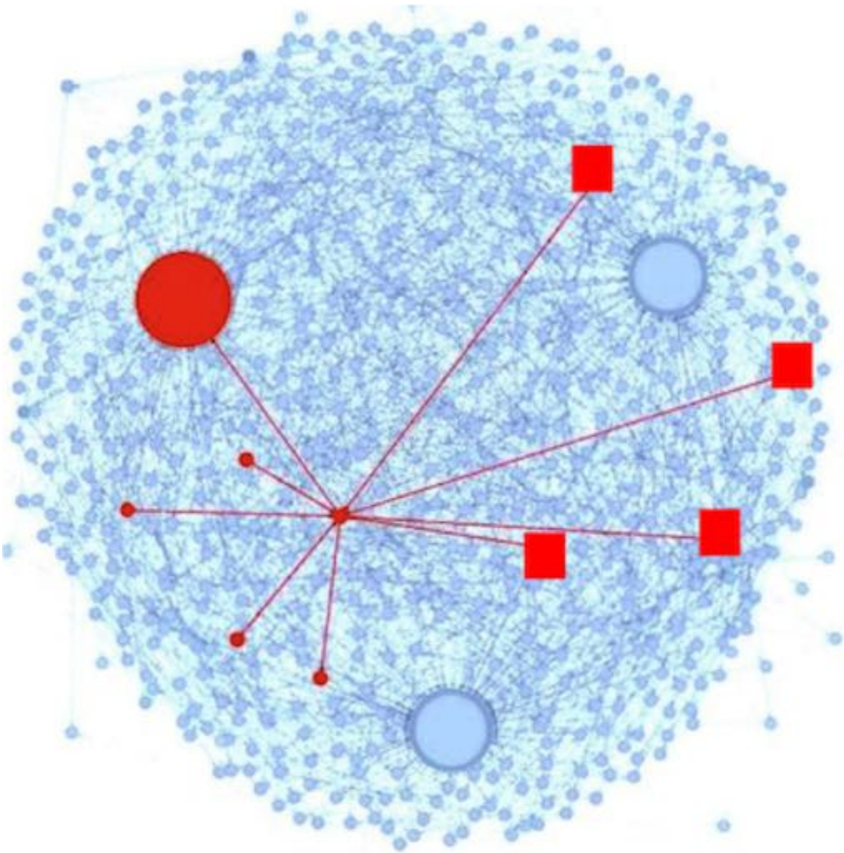
$$E(X) \quad + \quad E(Y) \quad = \quad E(X+Y)$$

Fundamental Properties



Is E[X] enough?

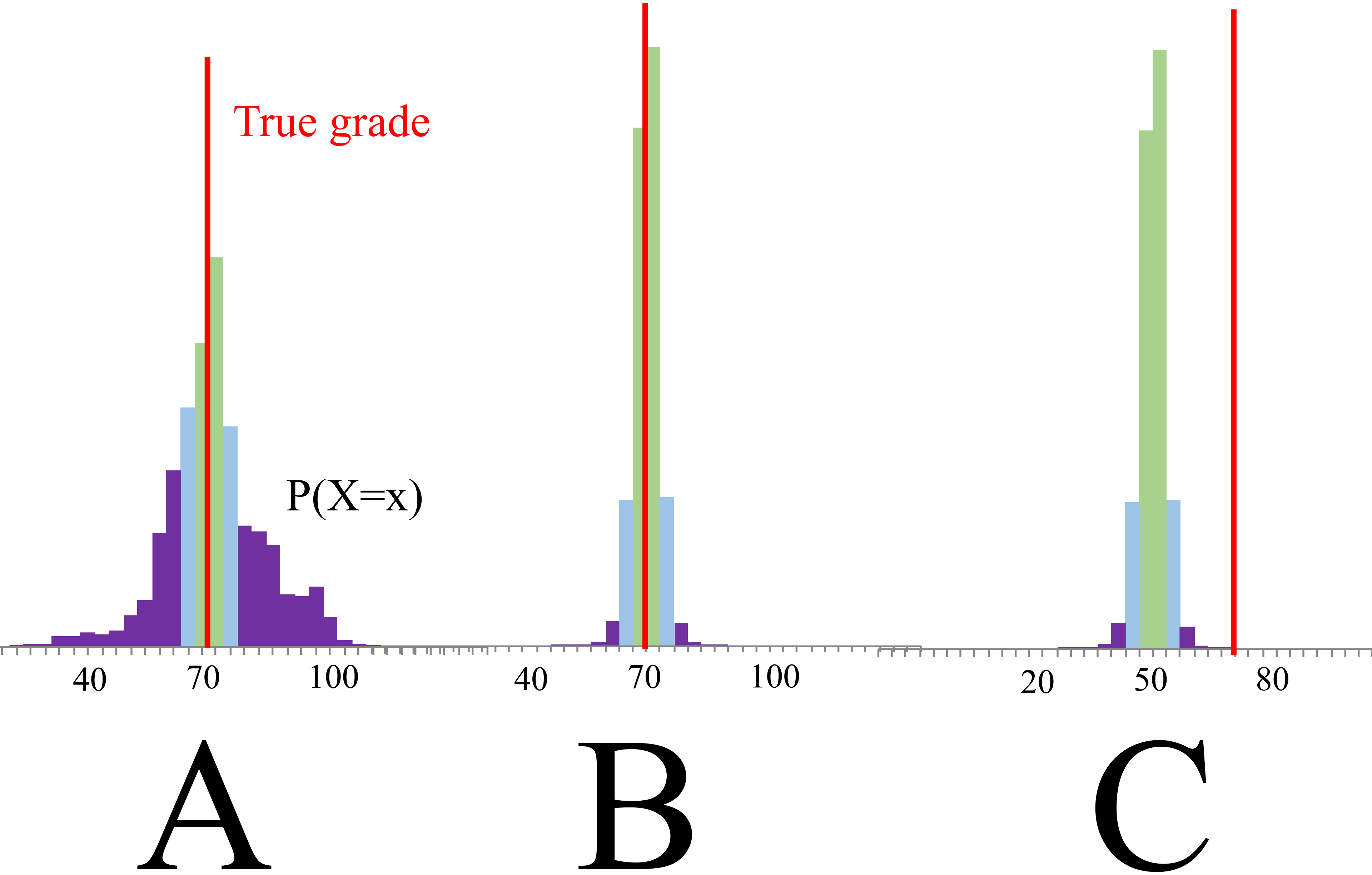
Intuition



Peer Grading on Coursera HCI.

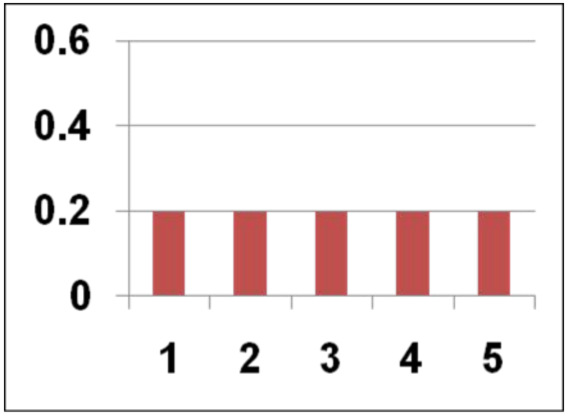
31,067 peer grades for 3,607 students.

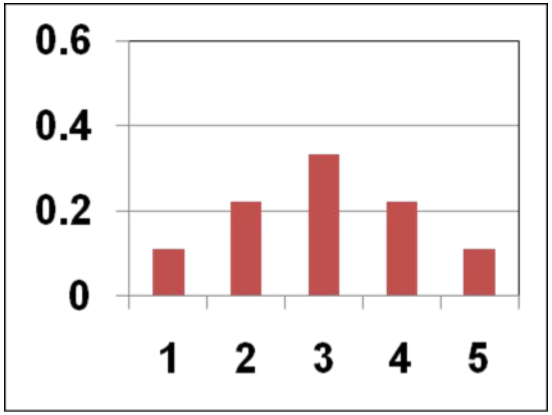
X is the score a peer grader gives to an assignment submission

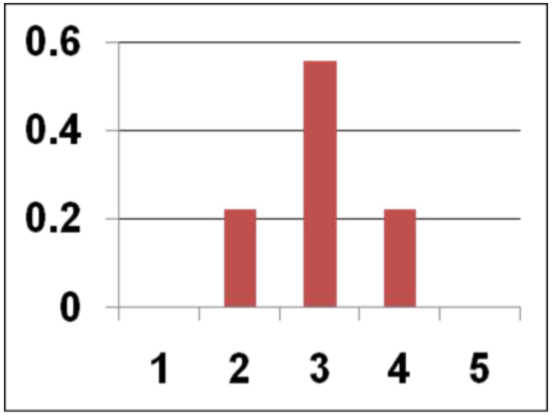


Variance

- Consider the following 3 distributions (PMFs)



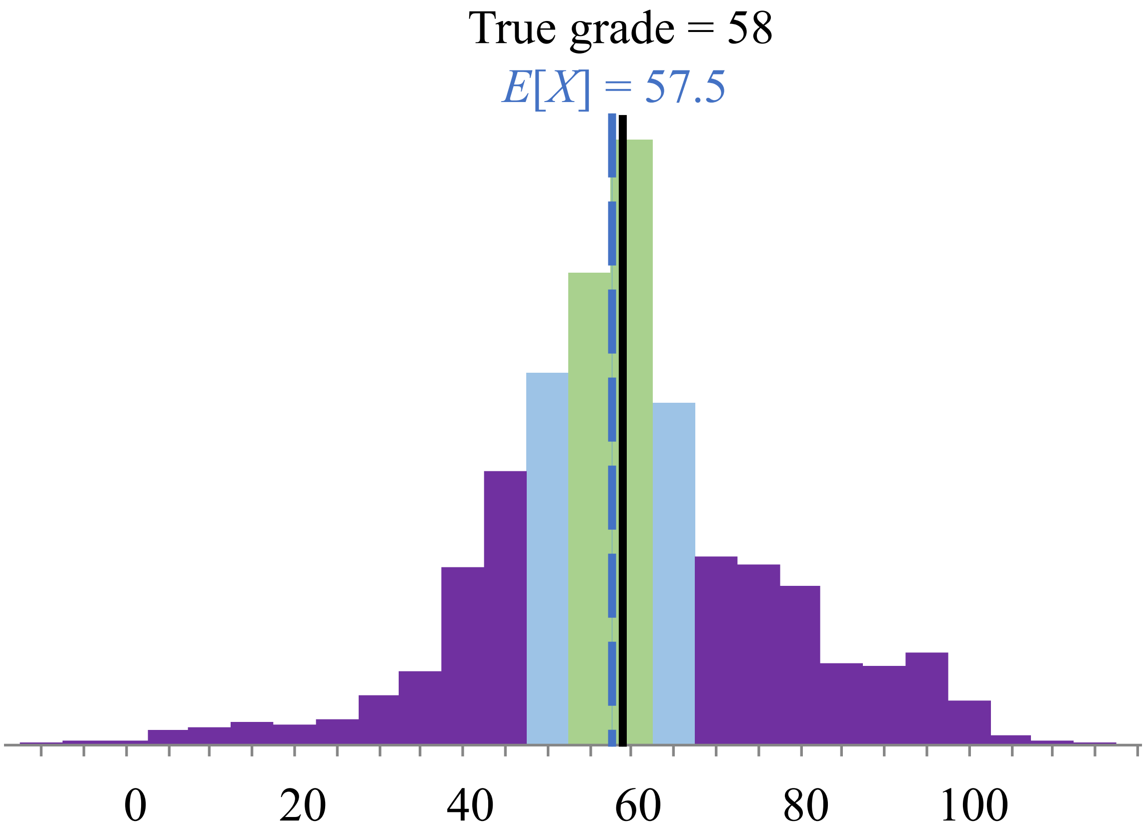




- All have the same expected value, E[X] = 3
- But "spread" in distributions is different
- Variance = a formal quantification of "spread"

Peer Grades in Coursera HCI

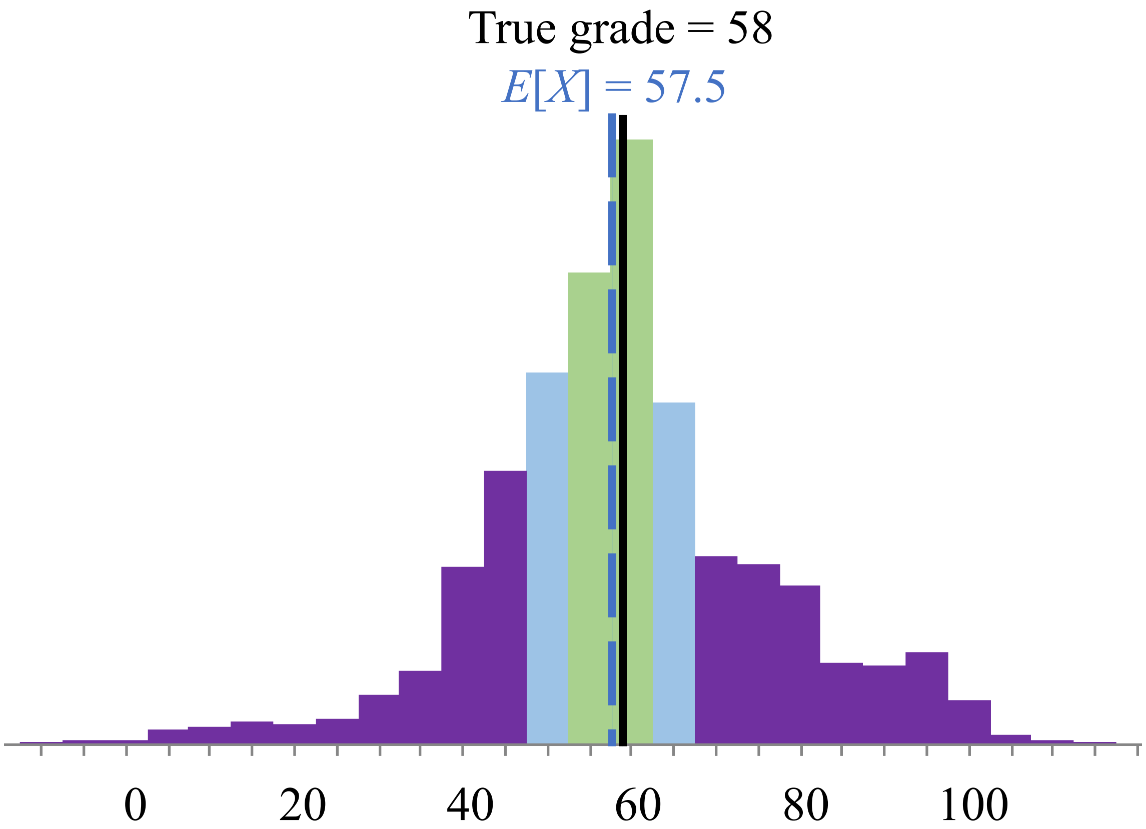
Let X be a random variable that represents a peer grade



Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

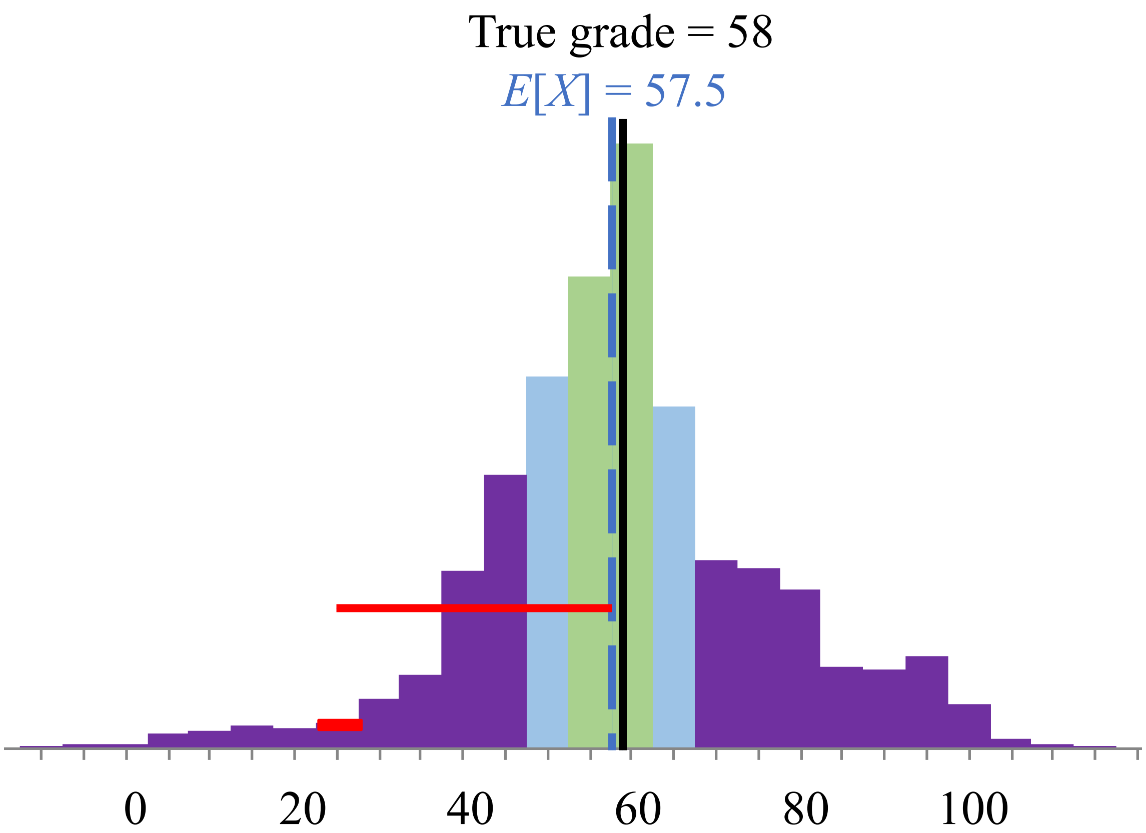
$$\mathrm{Var}(X) = E[(X - \mu)^2]$$



Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

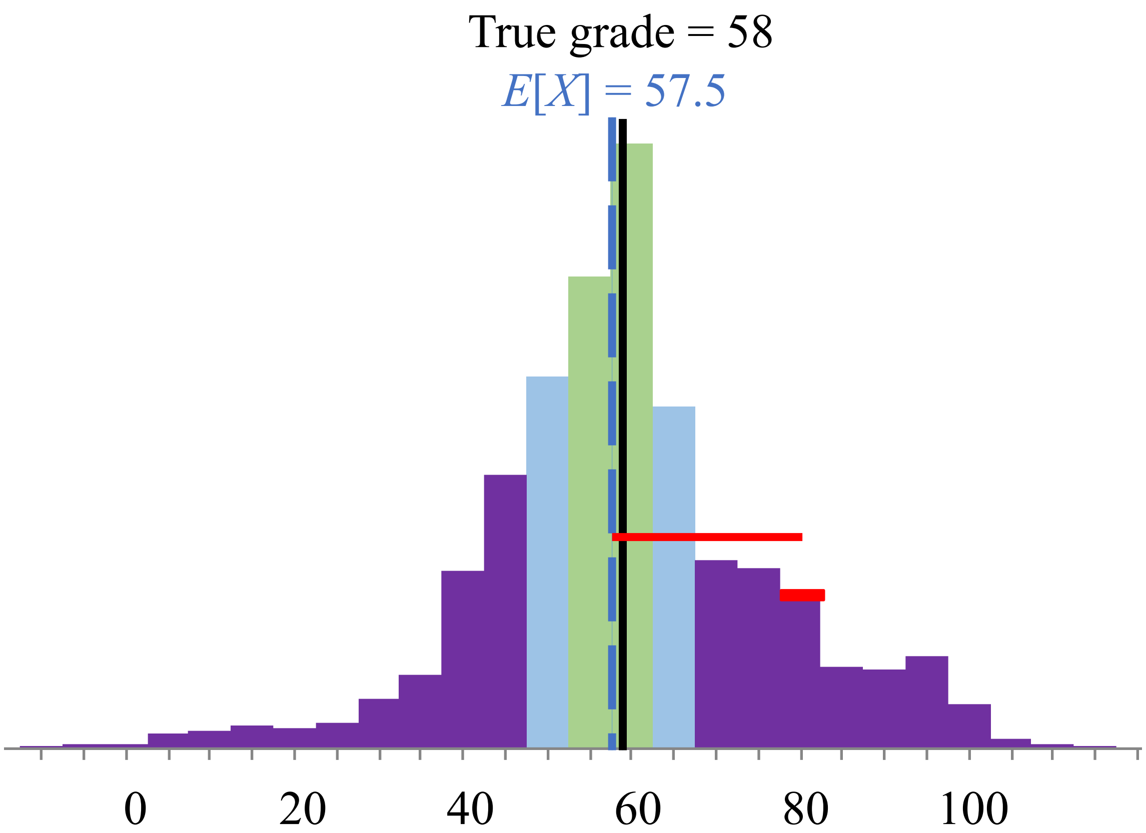


X	$(X - \mu)^2$
25 points	1056 points^2

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

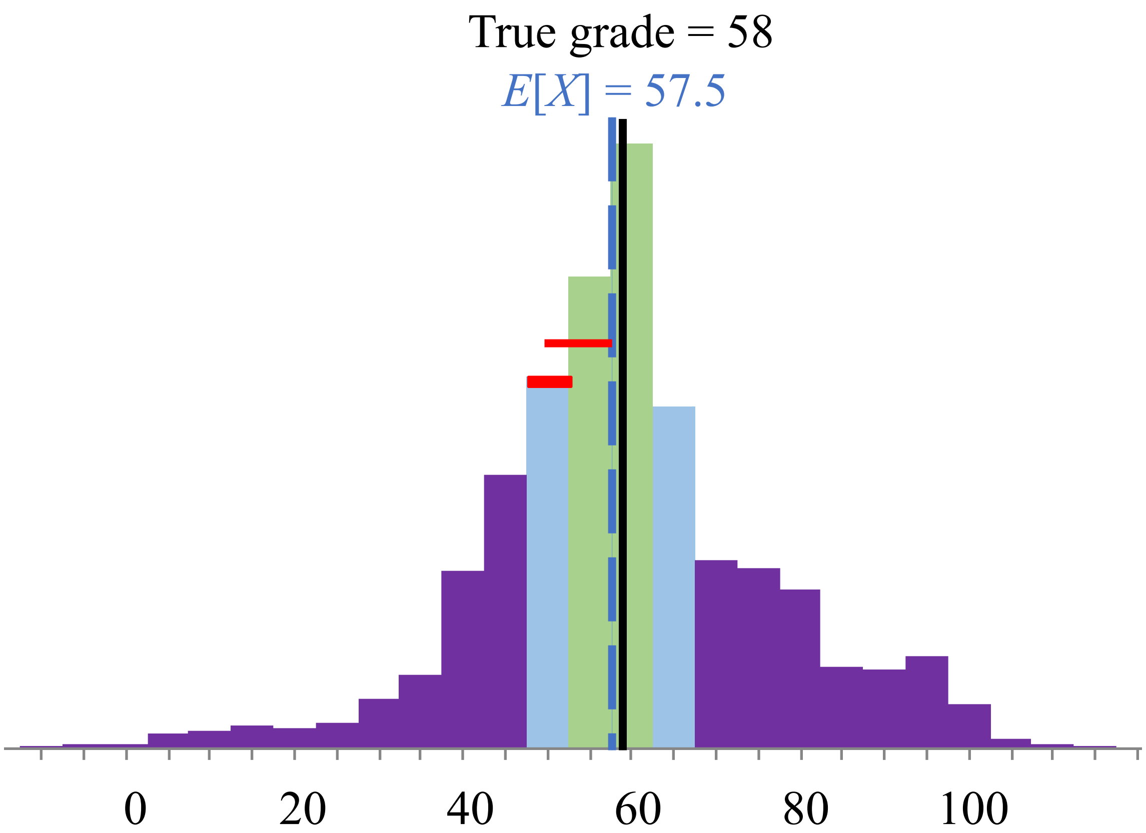


X	$(X - \mu)^2$
25 points	1056 points^2
80 points	506 points^2

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

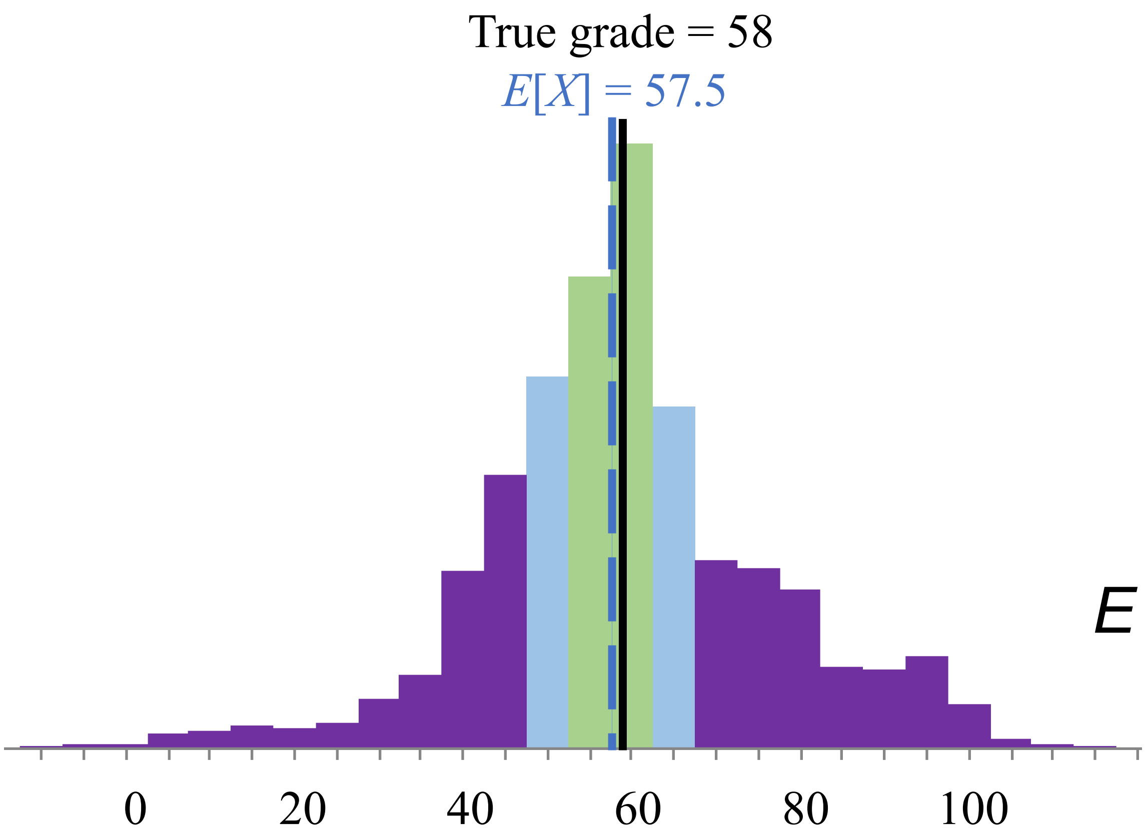


X	$(X - \mu)^2$
25 points	1056 points^2
80 points	506 points^2
50 points	56 points^2

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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



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

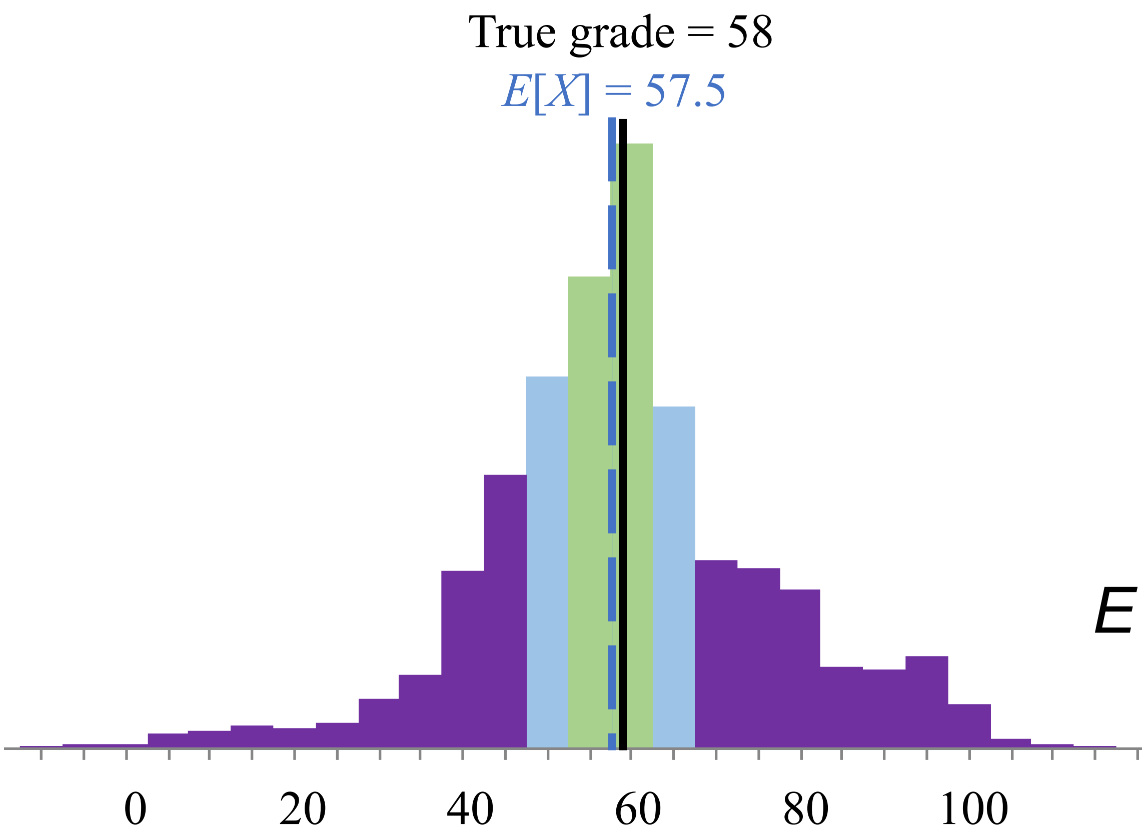
…

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

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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



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

…

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

$Std(X) = 7.2$ points

Variance

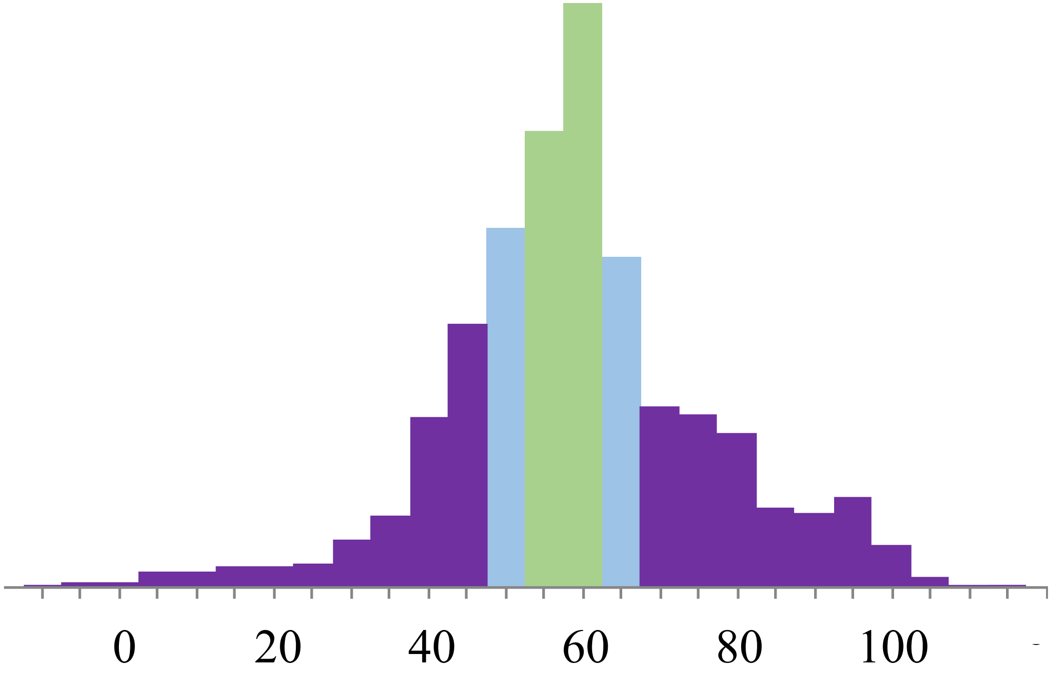
- If X is a random variable with mean μ then the **variance** of X, denoted $\mathrm{Var}(X)$, is:

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

- Note: $\mathrm{Var}(X) \geq 0$

- Also known as the 2nd **Central** Moment, or square of the Standard Deviation

Normalized **histograms** are approximations of **probability mass functions**



Computing Variance

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

Note: $\mu = E[X]$

Ladies and gentlemen, please welcome the 2nd 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]

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

$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - (E[X])^2 \\ &= \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12} \end{aligned}$$

Properties of Variance

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

    $\text{Var}(aX + b) = E[(aX + b)^2] - (E[aX + b])^2$

    $= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2$

    $= a^2E[X^2] + 2abE[X] + b^2 - (a^2(E[X])^2 + 2abE[X] + b^2)$

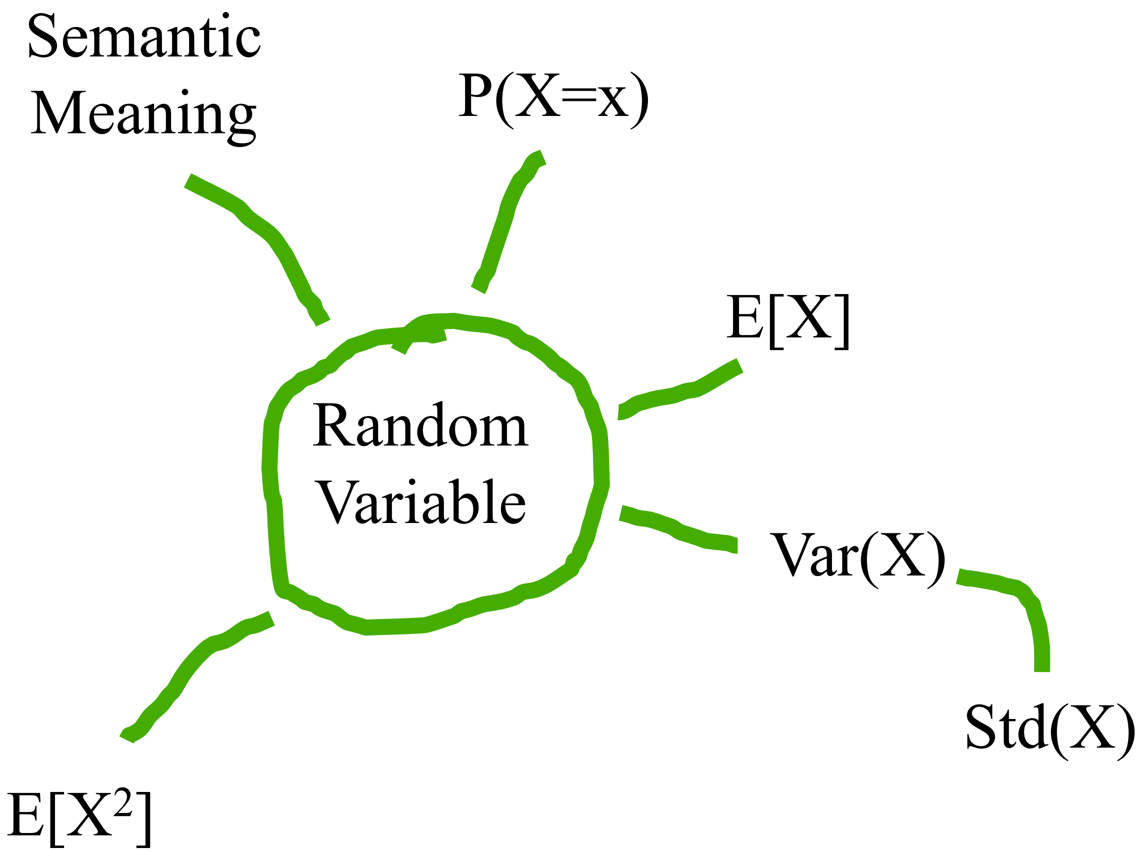
    $= a^2E[X^2] - a^2(E[X])^2 = a^2(E[X^2] - (E[X])^2)$

    $= a^2\text{Var}(X)$
- Standard Deviation of X, denoted SD(X), is:

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

  - Var(X) is in units of X^2
  - SD(X) is in same units as X

Fundamental Properties



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^{12}-grandfather
- Same eyes as Ice Cube

Bernoulli Random Variable

- Experiment results in "Success" or "Failure"
  - X is random **indicator** variable (1 = success, 0 = failure)
  - $P(X = 1) = p(1) = p$ $P(X = 0) = p(0) = 1 – p$
  - X is a **<u>Bernoulli</u>** Random Variable: $X \sim Ber(p)$
  - $E[X] = p$
  - $Var(X) = p(1 – p)$
- 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. $(1 – p)$

X: 1 if program crashes

$P(X = 1) = p$
$P(X = 0) = 1 - p$

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

Does a User Click an Ad?

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

C: 1 if ad is clicked

$P(C = 1) = p$

$P(C = 0) = 1 - p$

<u>$C \sim \mathrm{Ber}(p)$</u>

More!

Binomial Random Variable

- Consider *n* independent trials of Ber(p) rand. var.
  - X is number of successes in *n* trials
  - X is a **Binomial** Random Variable: X ~ Bin(n, p)

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

  - 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

Probability of success on each trial

Our random variable

Num trials

$$X \sim \mathrm{Bin}(n, p)$$

Is distributed as a

Binomial

With these parameters

If X is a binomial with parameters n and p

$$\underbrace{P(X = k)}_{\text{Probability that our variable takes on the value k}} = \overbrace{\binom{n}{k} p^k (1-p)^{n-k}}^{\text{Probability Mass Function for a Binomial}}$$

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 ~ 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 ~ Bin(n, p)

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

- $E[X] = np$

- $\mathrm{Var}(X) = np(1-p)$

- Note: $\mathrm{Ber}(p) = \mathrm{Bin}(1, p)$

I Really Want the Proof of Var :)

$$
\begin{aligned}
E\left(X^2\right) &= \sum_{k\geq 0}^{n} k^2 \binom{n}{k} p^k q^{n-k} && \text{Definition of Binomial Distribution: } p+q=1 \\
&= \sum_{k=0}^{n} kn \binom{n-1}{k-1} p^k q^{n-k} && \text{Factors of Binomial Coefficient: } k\binom{n}{k} = n\binom{n-1}{k-1} \\
&= np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} && \text{Change of limit: term is zero when } k-1=0 \\
&= np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j} && \text{putting } j=k-1, m=n-1 \\
&= np \left(\sum_{j=0}^{m} j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{splitting sum up into two} \\
&= np \left(\sum_{j=0}^{m} m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{Factors of Binomial Coefficient: } j\binom{m}{j} = m\binom{m-1}{j-1} \\
&= np \left((n-1)\, p \sum_{j=1}^{m} \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) && \text{Change of limit: term is zero when } j-1=0 \\
&= np \left((n-1)\, p (p+q)^{m-1} + (p+q)^m \right) && \text{Binomial Theorem} \\
&= np \left((n-1)\, p + 1 \right) && \text{as } p+q=1 \\
&= n^2 p^2 + np\,(1-p) && \text{by algebra}
\end{aligned}
$$

How Many Program Crashes?

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

H: number of crashes

H ~ Bin(n, p)

P(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

H ~ Bin(n = 1000, p = 0.01)

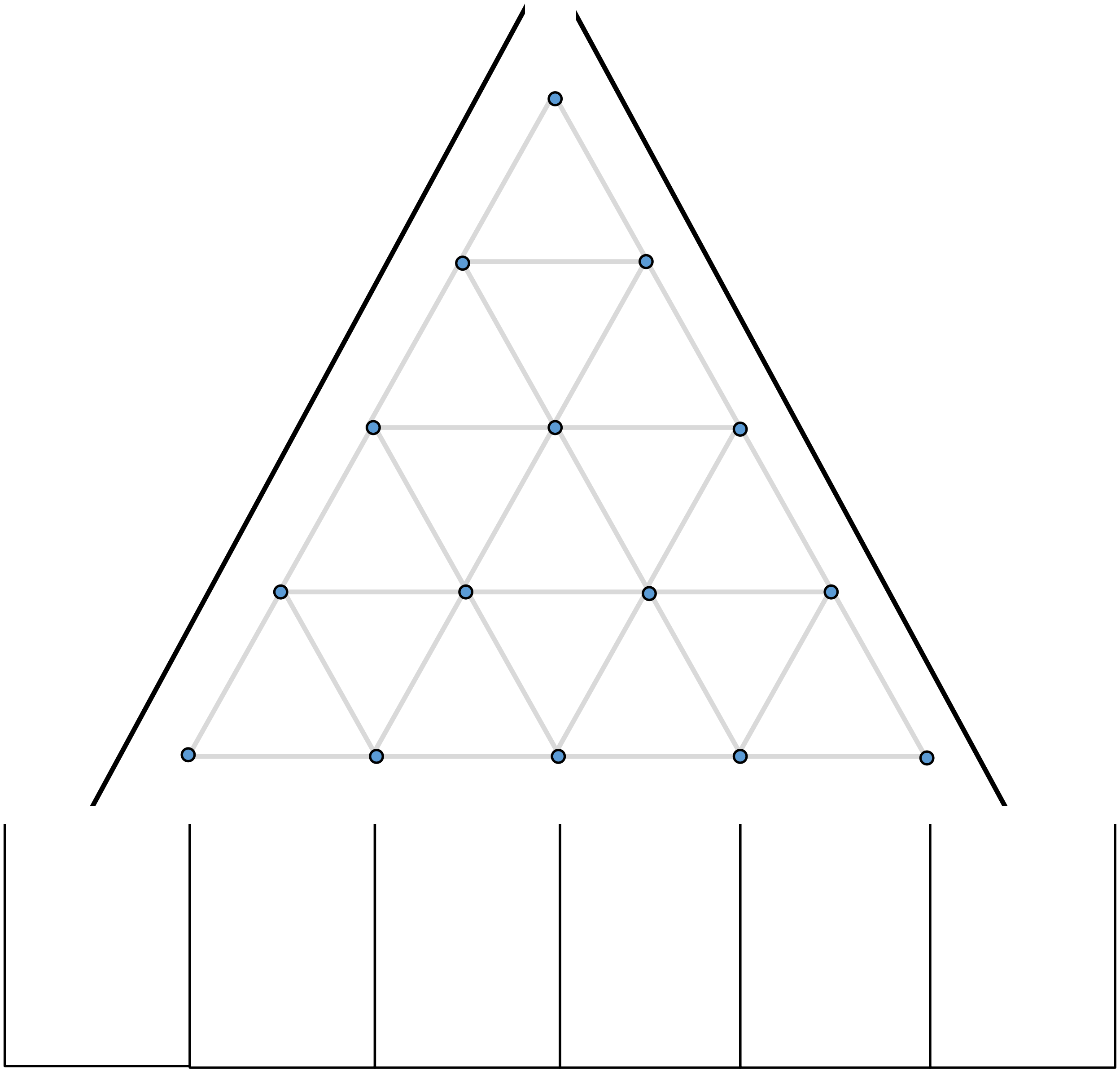
P(H = k**)** = $\binom{1000}{k}(0.01)^k(0.99)^{1000-k}$

Variance of number of ads clicked?

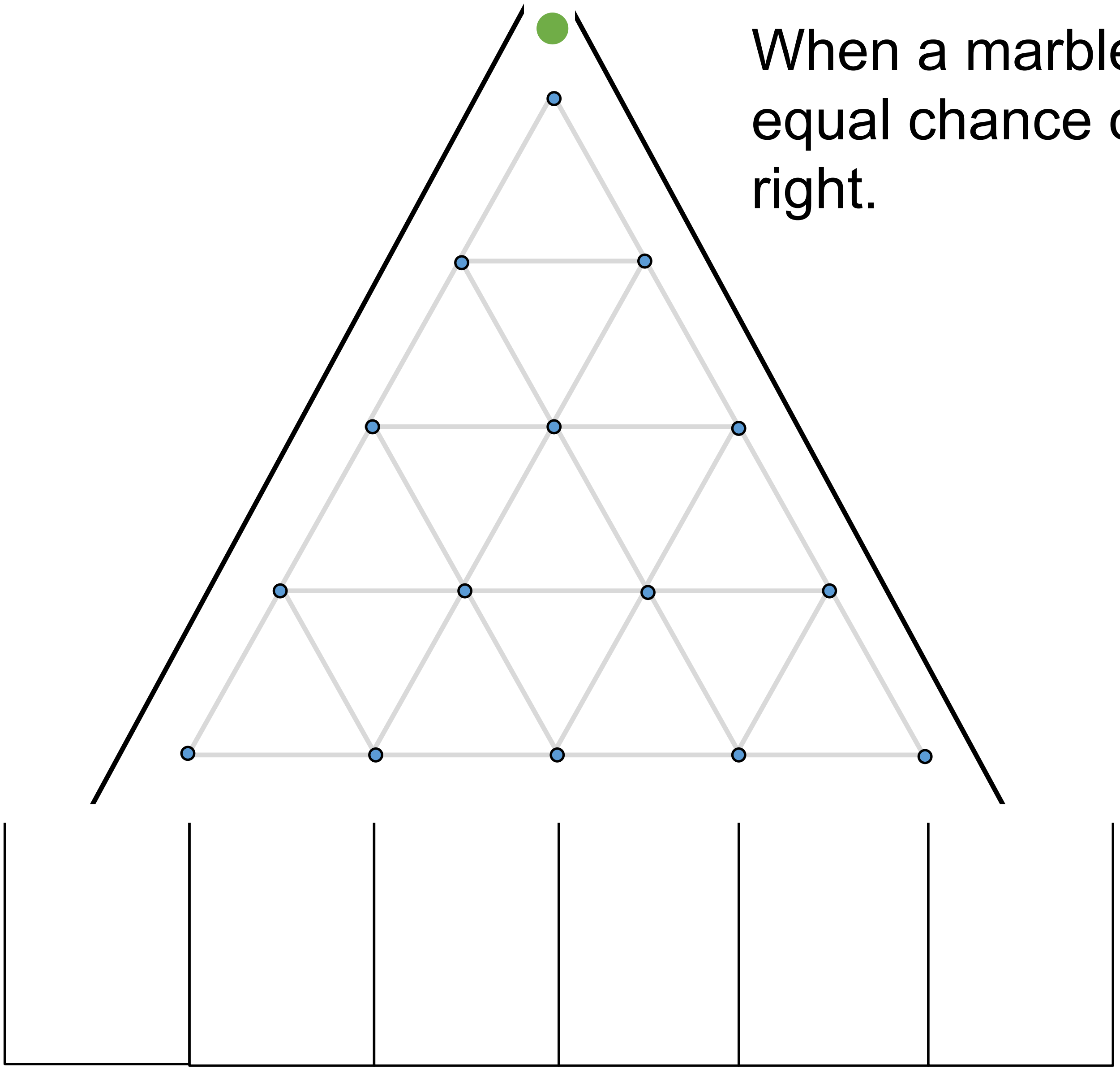
E[**H**] = np = 10

Var(**H**) = $np(1-p)$ = 9.9 Std(**H**) = 3.15

Galton Board

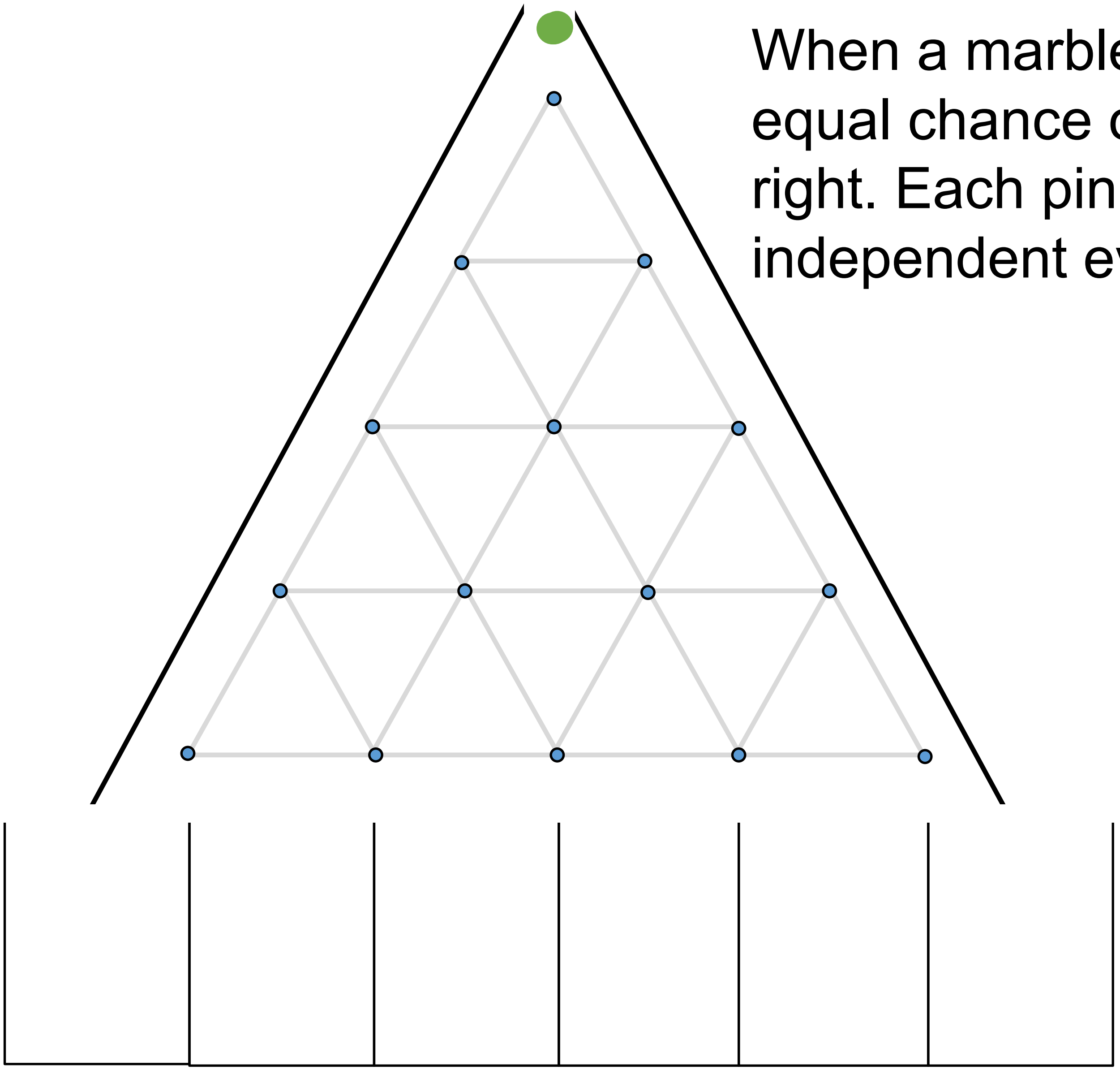


Galton Board



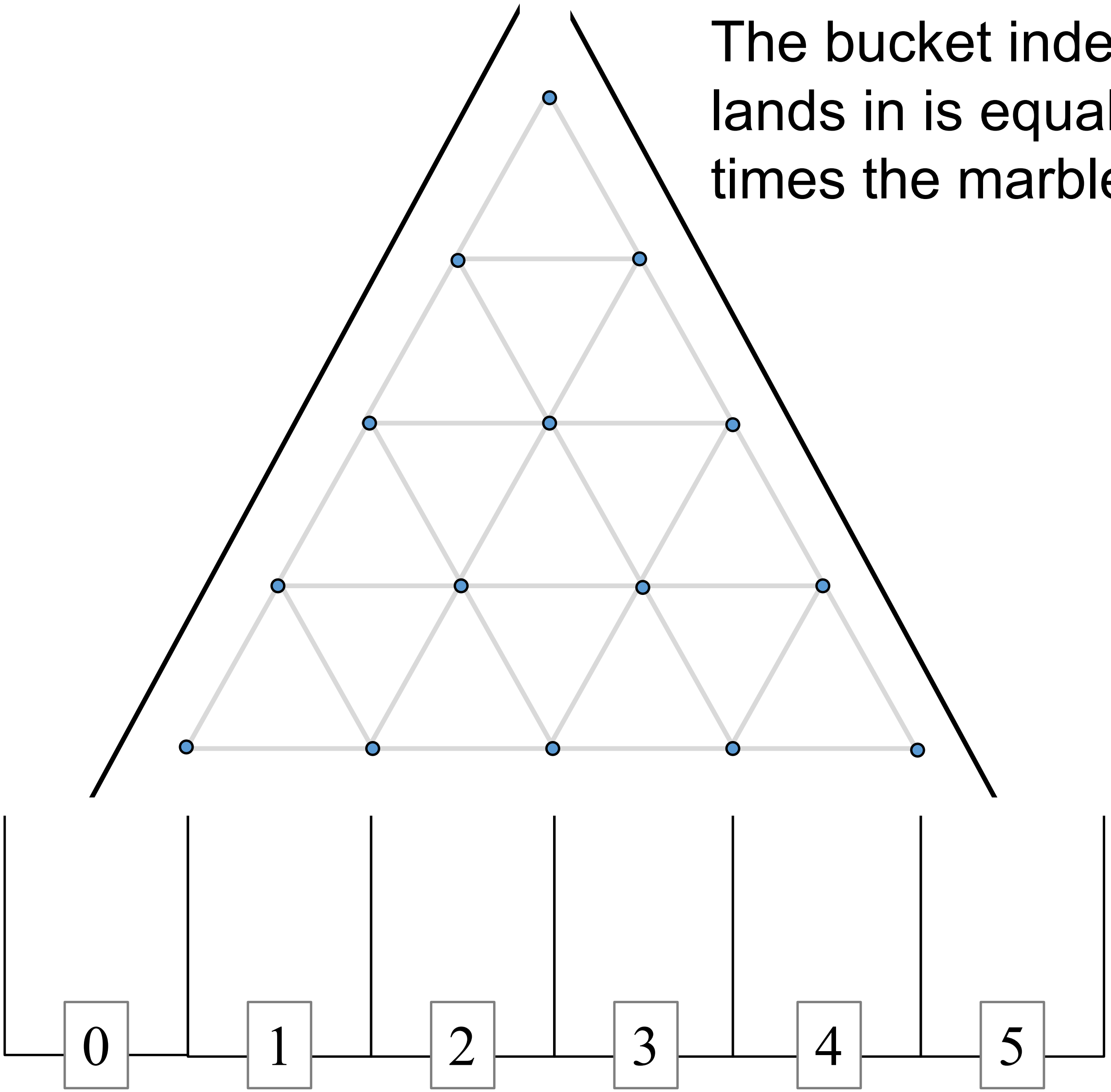
When a marble hits a pin, it has equal chance of going left or right.

Galton Board



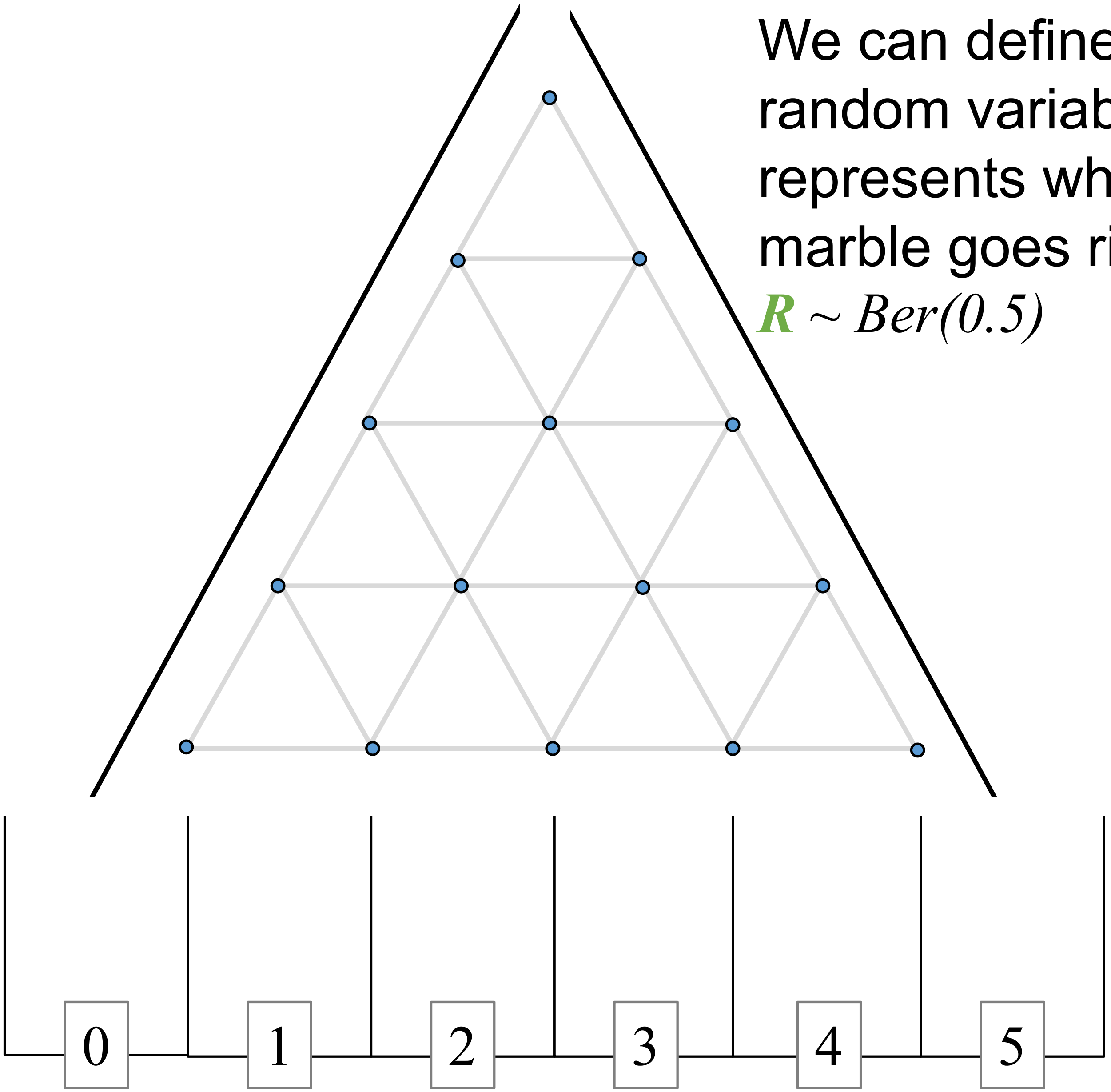
When a marble hits a pin, it has equal chance of going left or right. Each pin represents an independent event.

Galton Board



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

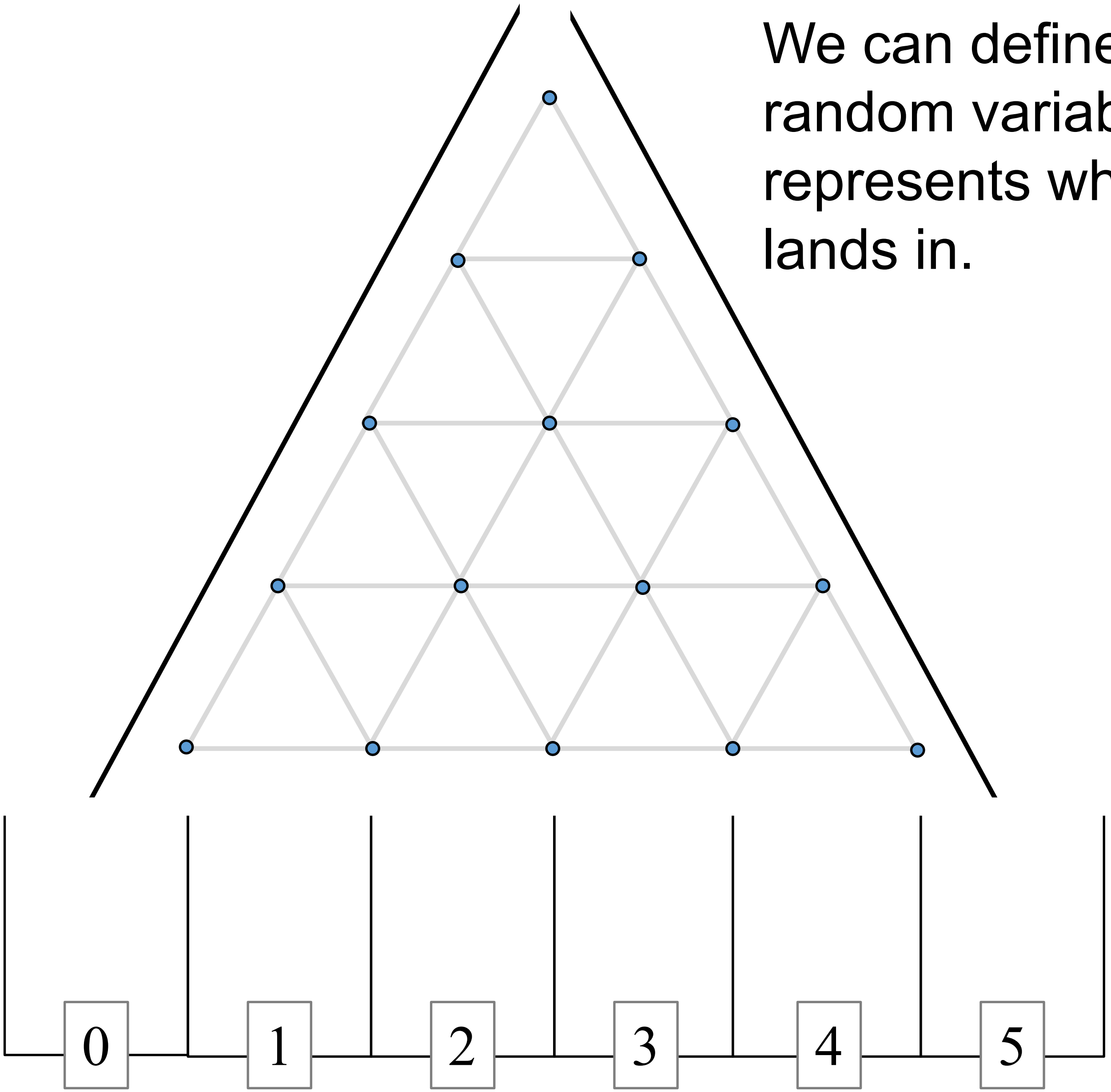
Galton Board



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

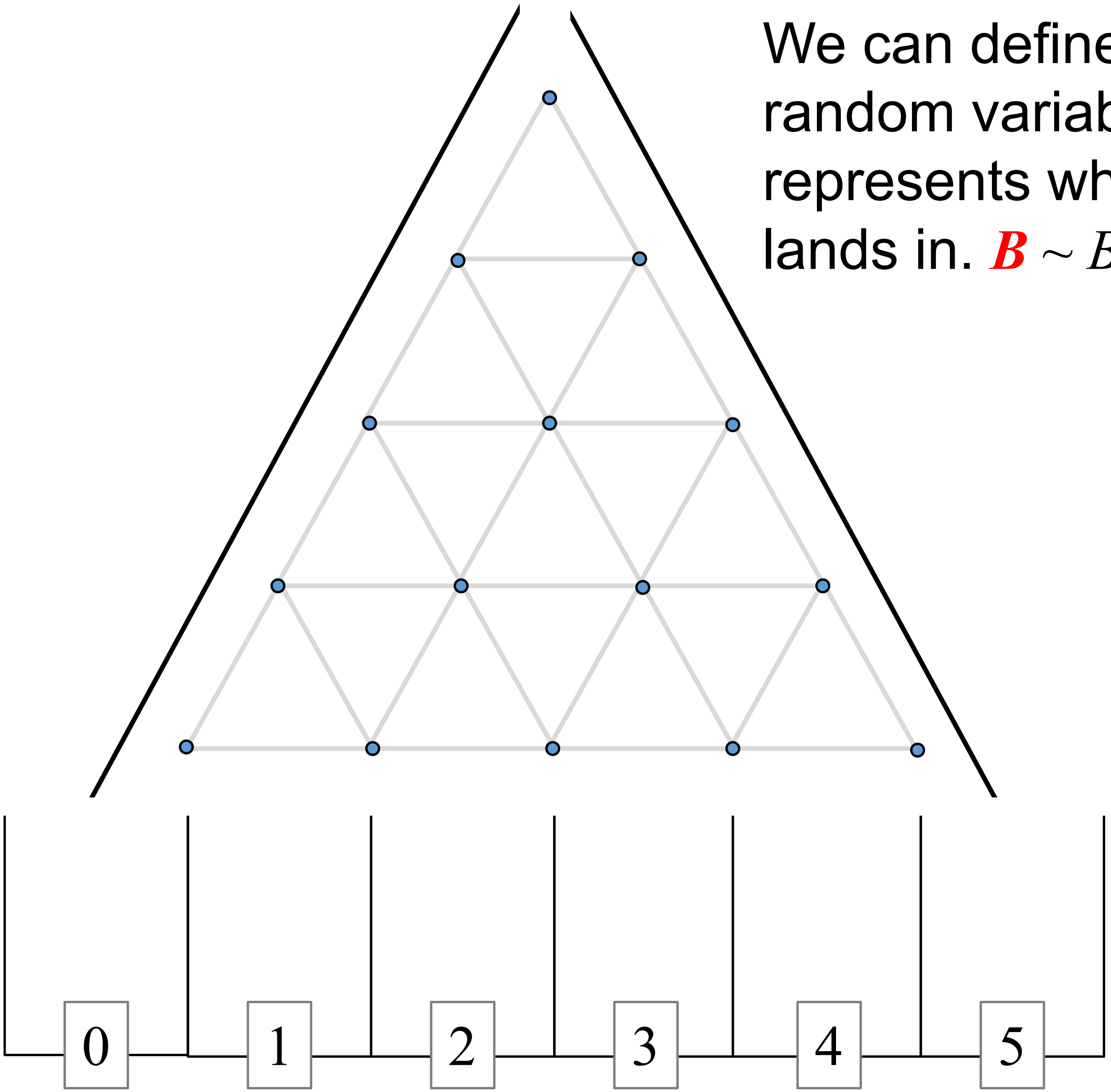
$R \sim Ber(0.5)$

Galton Board



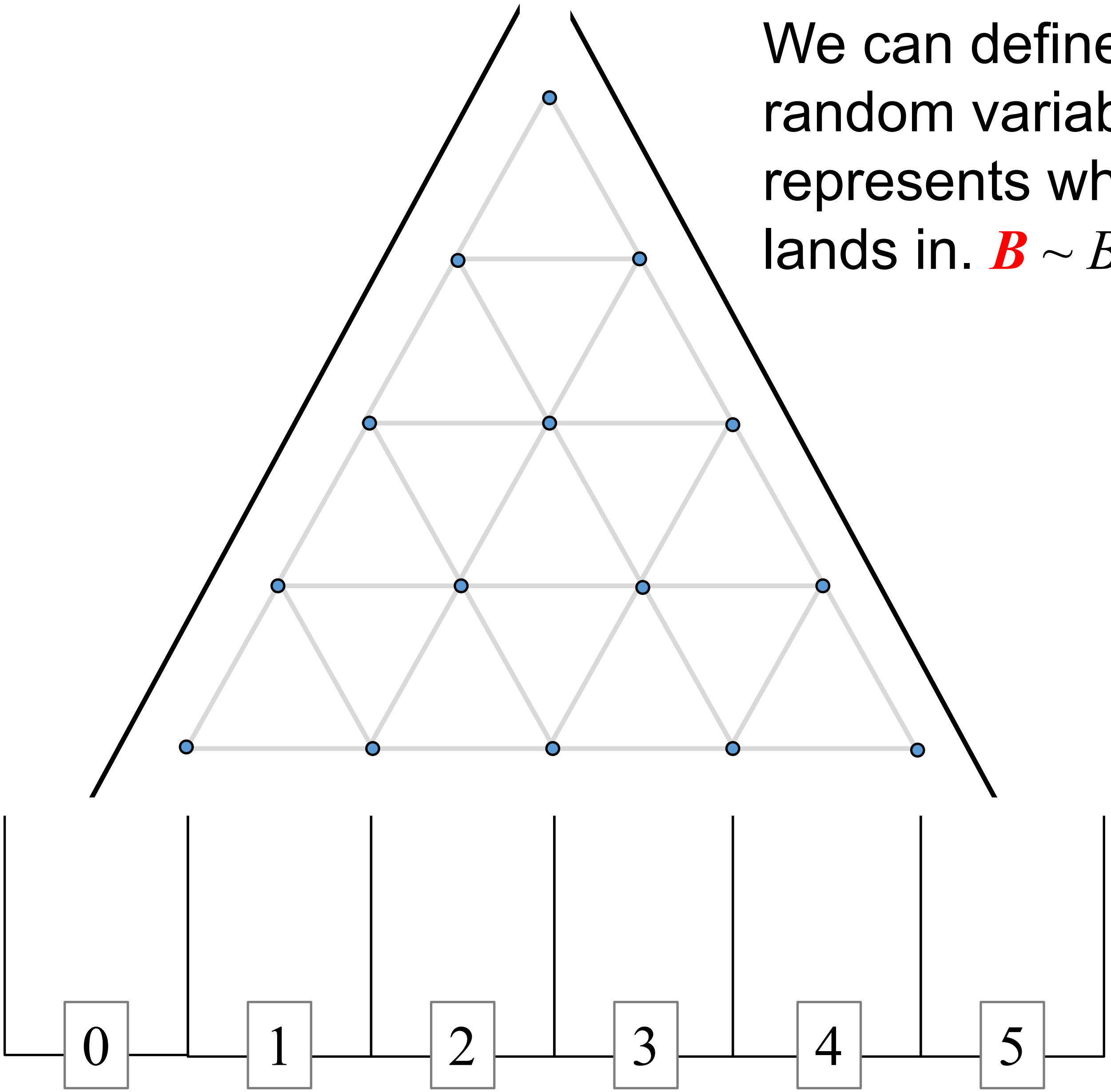
We can define an indicator random variable (***B***) which represents what bucket a marble lands in.

Galton Board



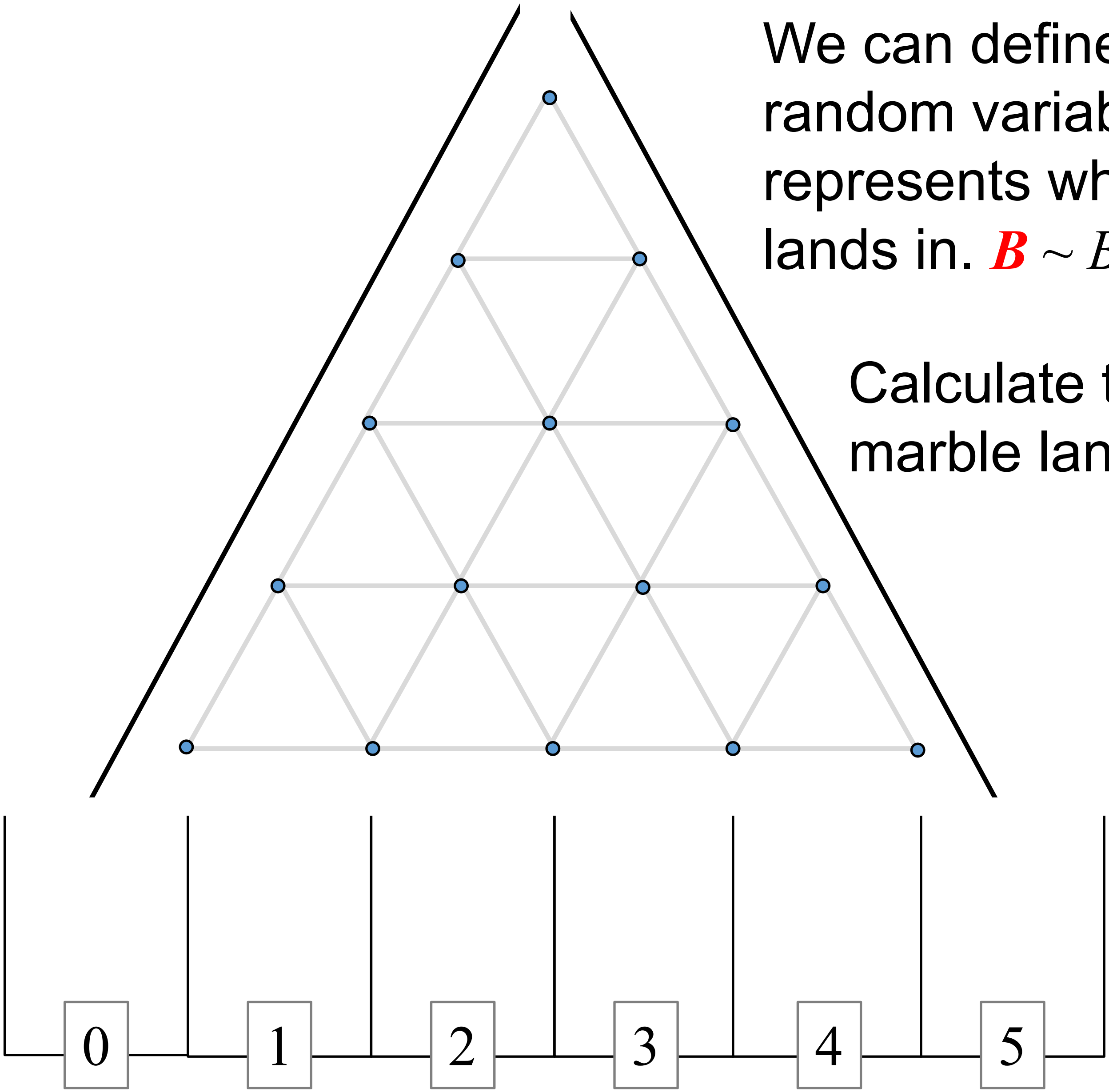
We can define an indicator random variable (***B***) which represents what bucket a marble lands in. $B \sim Bin(\text{levels}, 0.5)$

Galton Board



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

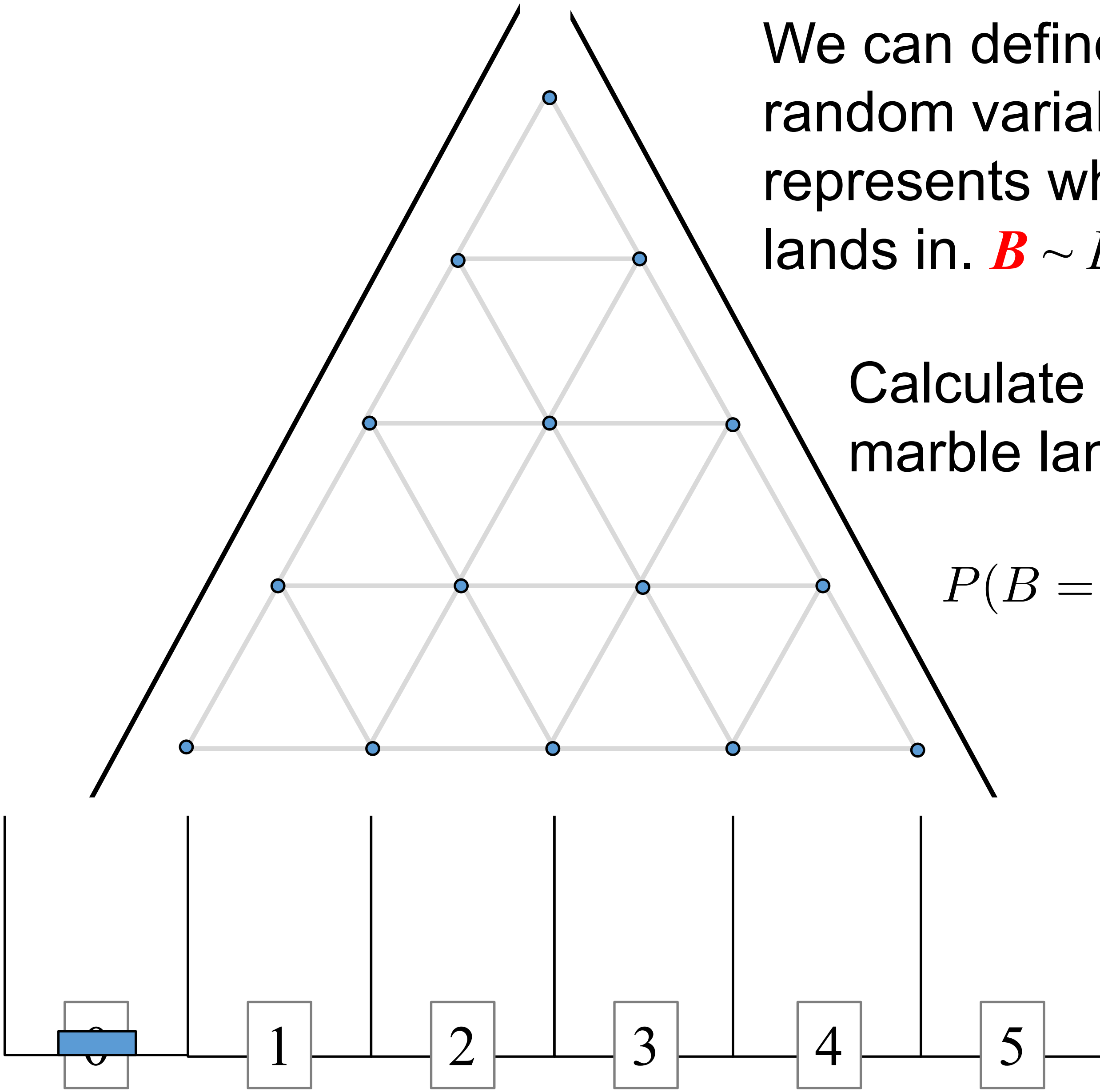
Galton Board



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

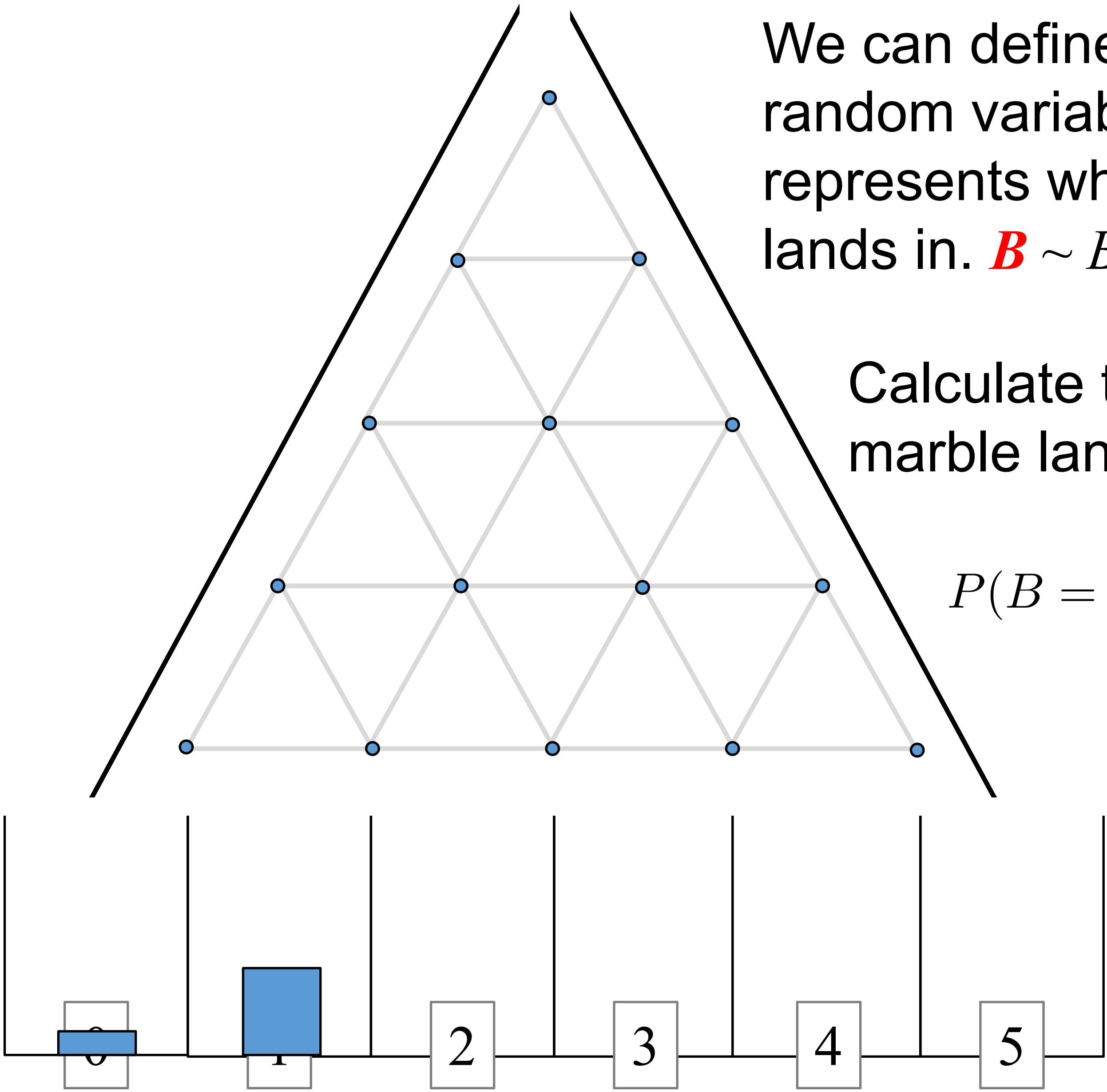


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. $B \sim Bin(5, 0.5)$

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

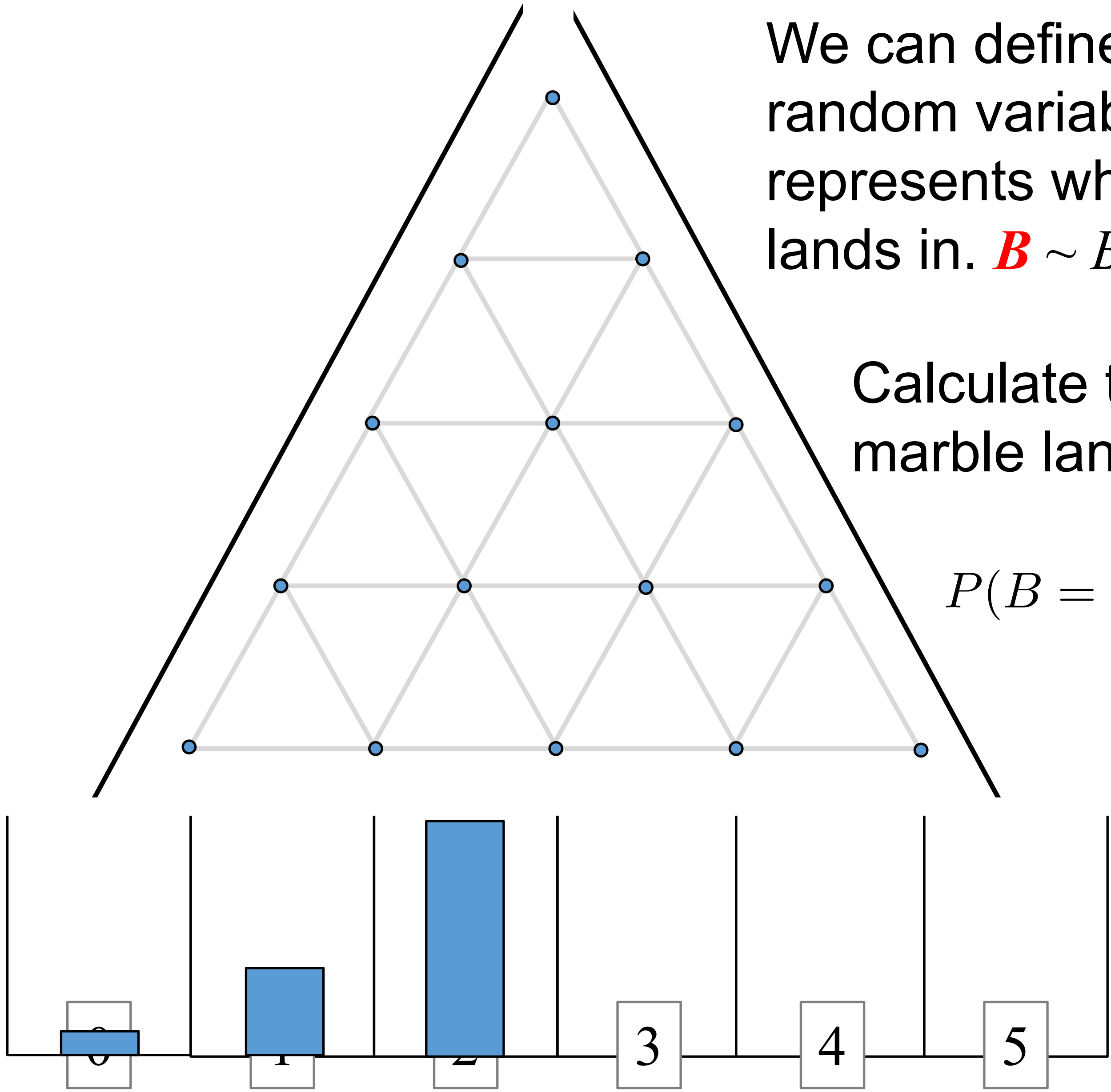


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** ~ *Bin(5, 0.5)*

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

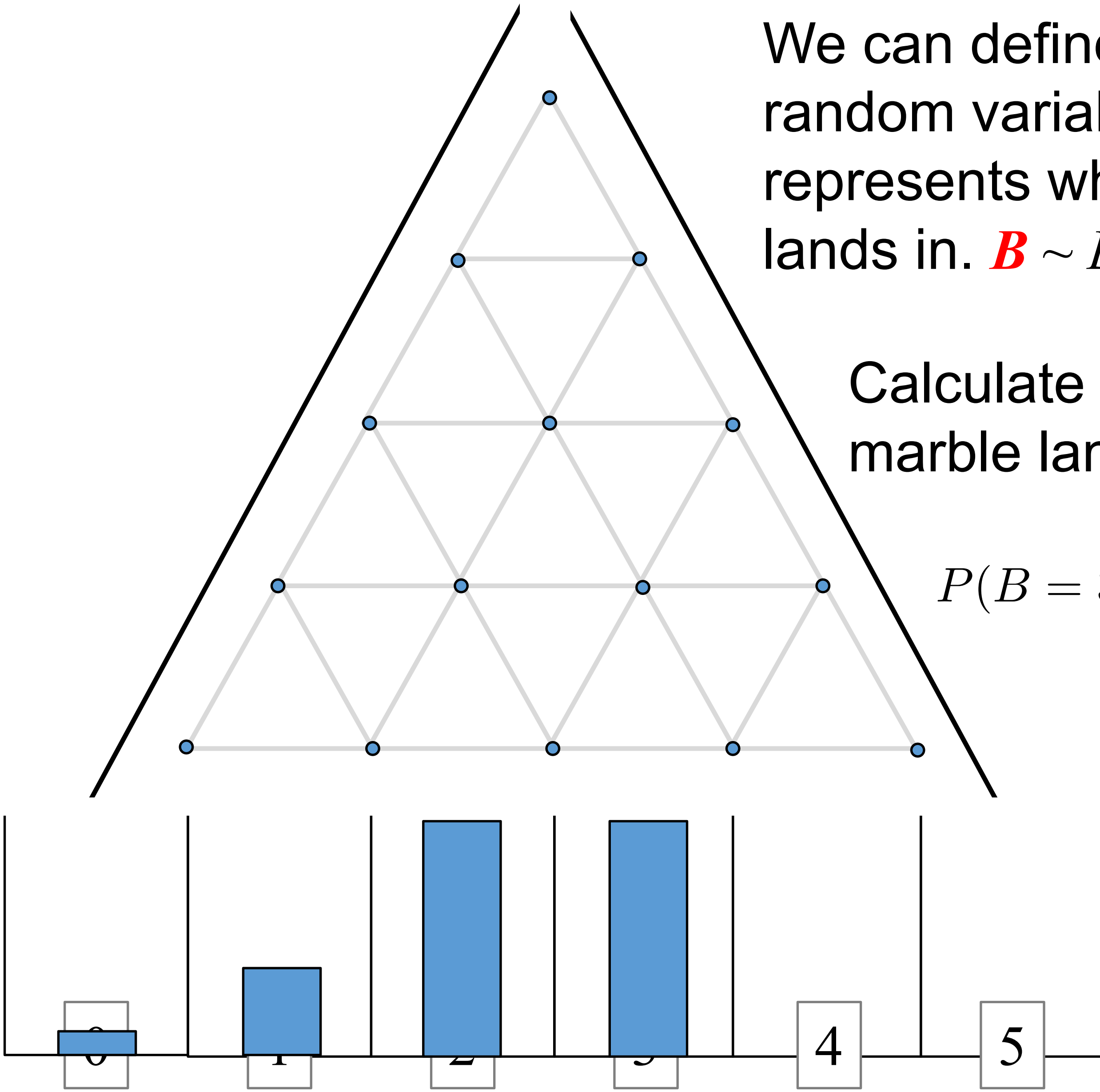


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** ~ *Bin(5, 0.5)*

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

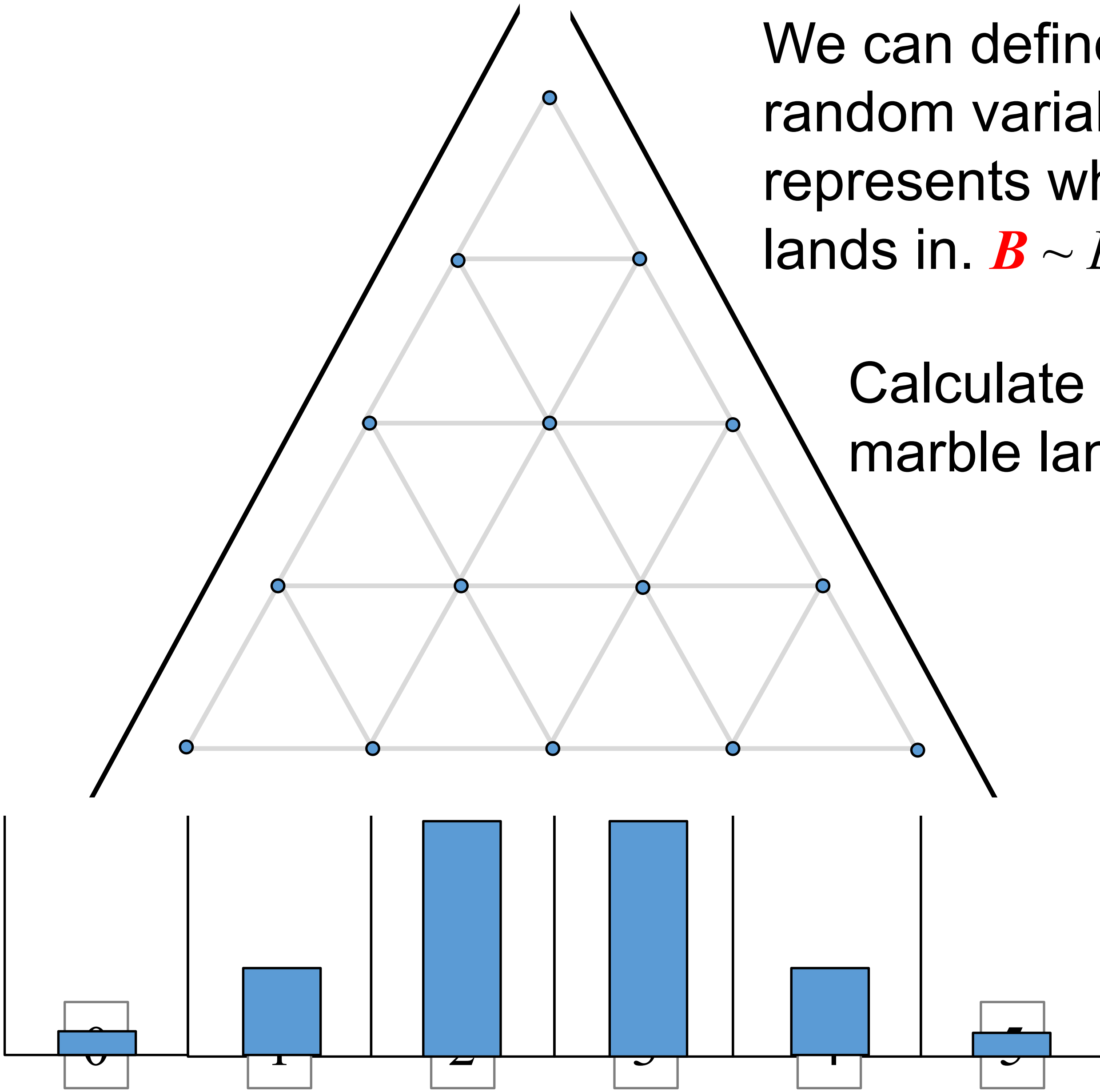


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** $\sim Bin(5, 0.5)$

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



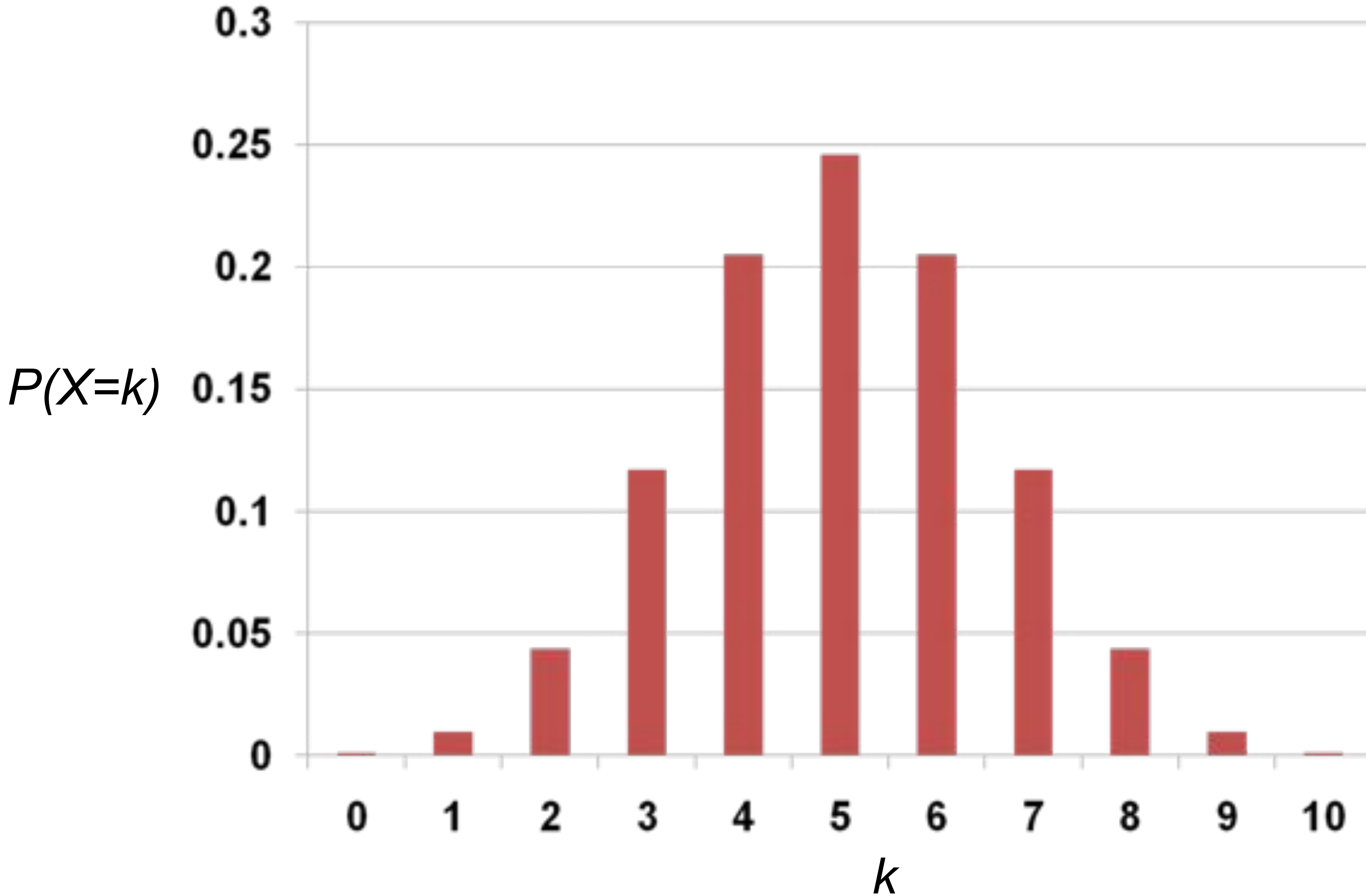
We can define an indicator random variable (B) which represents what bucket a marble lands in. $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

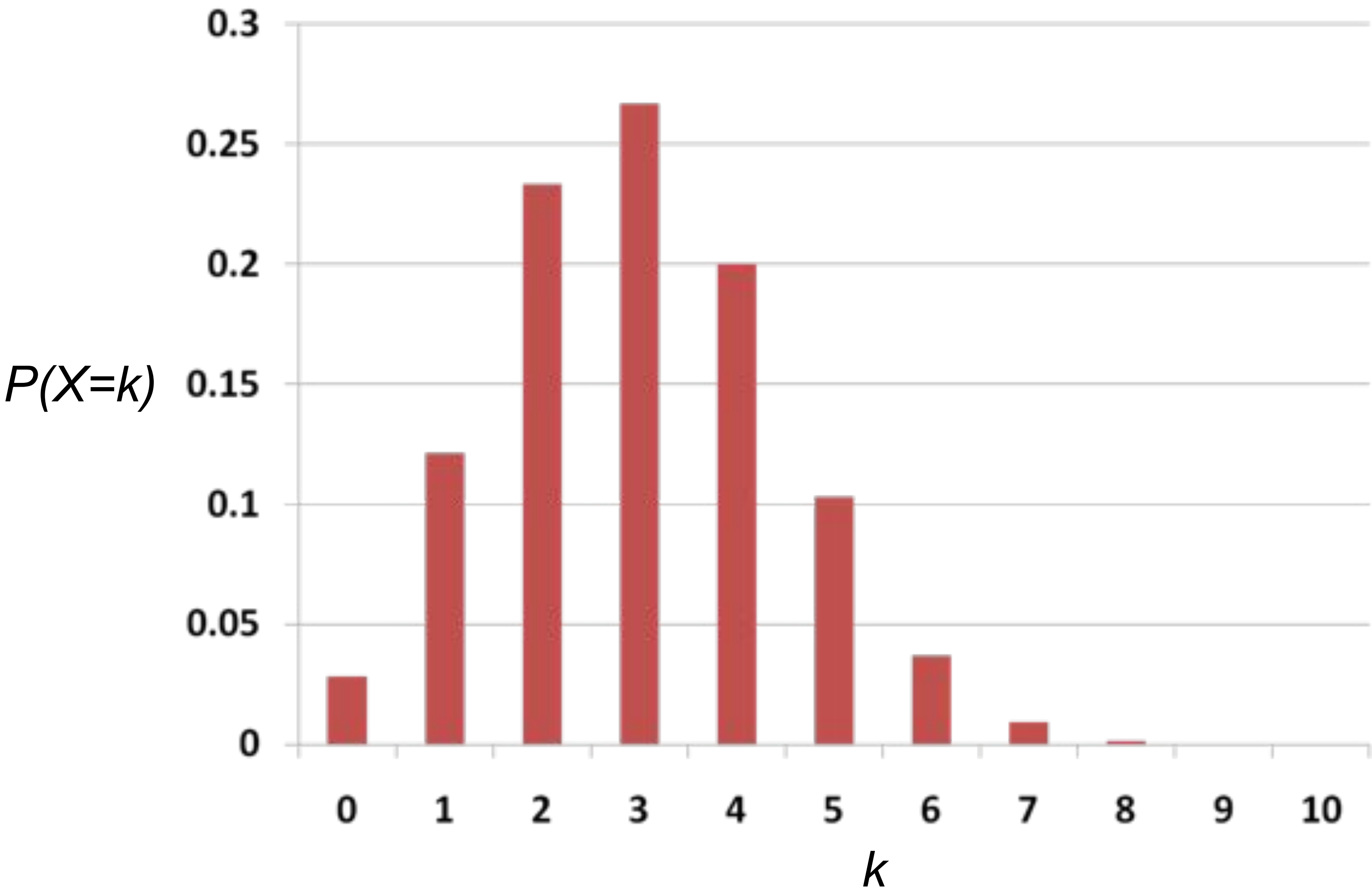
PDF

From Chaos to Order

PMF for X ~ Bin(10, 0.5)



PMF for X ~ 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 children with brown eyes)?
  - Child has blue eyes: p = (½) (½) = ¼ (2 blue genes)
  - P(child has brown eyes) = 1 – (¼) = 0.75
  - X = # of children with brown eyes. X ~ Bin(4, 0.75)

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


GOLDEN STATE
35
WARRIORS
GOLDEN STATE

Probability you win a series?

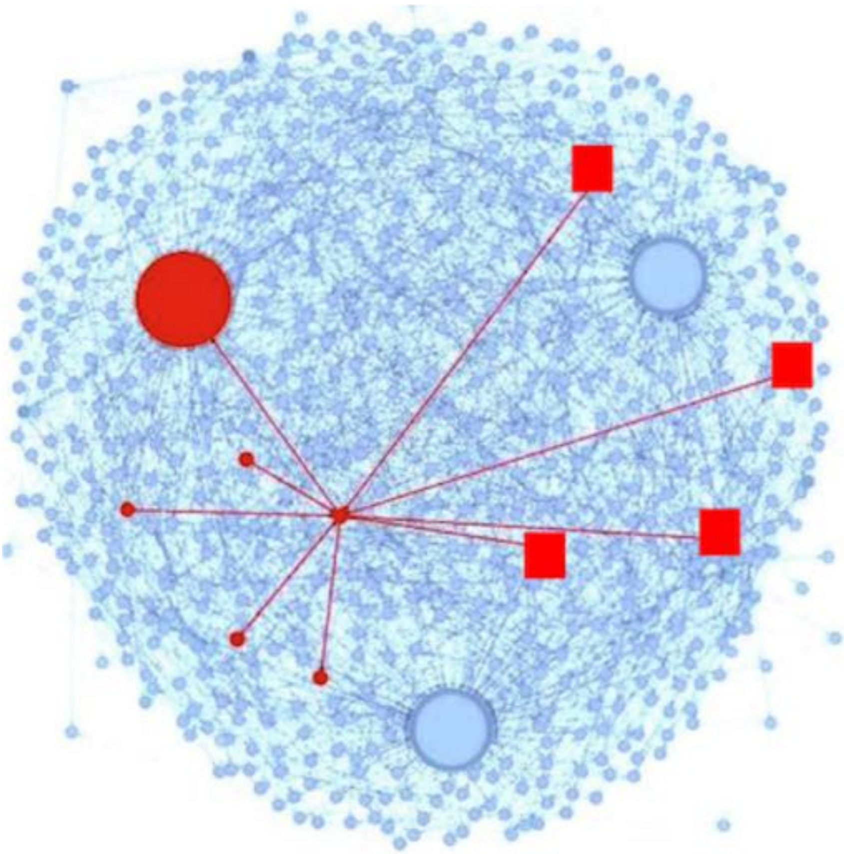
Warriors are going to play the Bucks in a best of 7 series during the 2017 NBA finals. What is the probability that the warriors win the series? Each game is **independent**. Each game, the warriors have a 0.55 probability of winning? Win series if you win at least 4 games.

Let X be the number of games won. X ~ Bin(n= 7, p = 0.55).
P(X > 3)?

$$
\begin{aligned}
P(X \geq 4) &= \sum_{i=4}^{7} P(X = i) \\
&= \sum_{i=4}^{7} \binom{7}{i} p^i (1-p)^{7-i} \\
&= \sum_{i=4}^{7} \binom{7}{i} 0.55^i (0.45)^{7-i}
\end{aligned}
$$

Is Peer Grading Accurate Enough?

Looking ahead



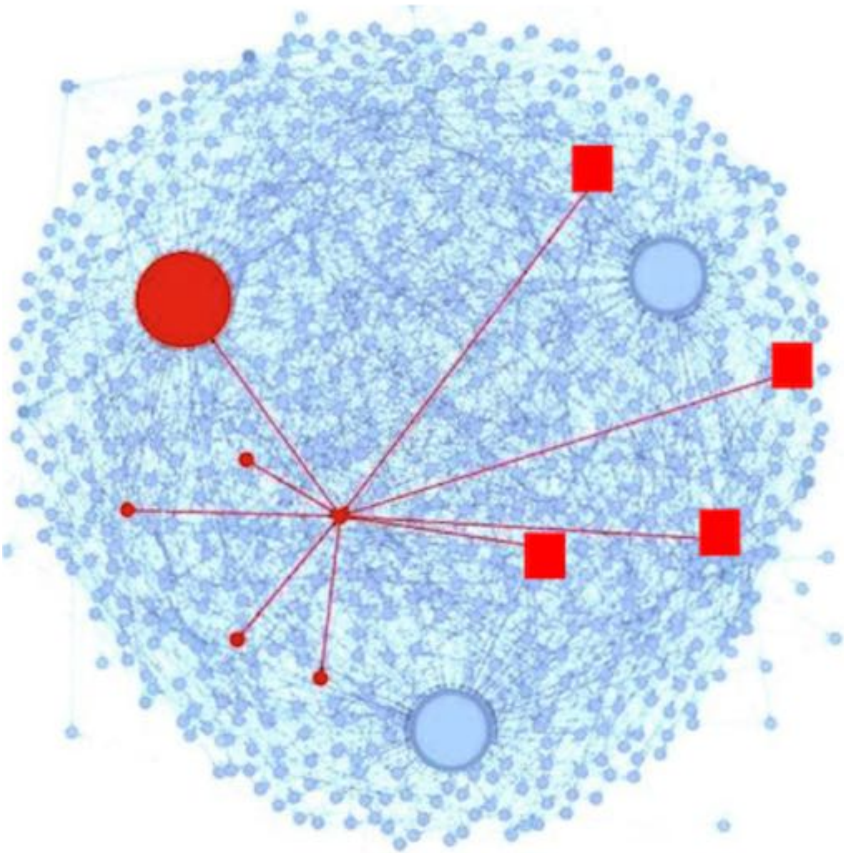
Peer Grading on Coursera HCI.

31,067 peer grades for 3,607 students.

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

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

Problem param

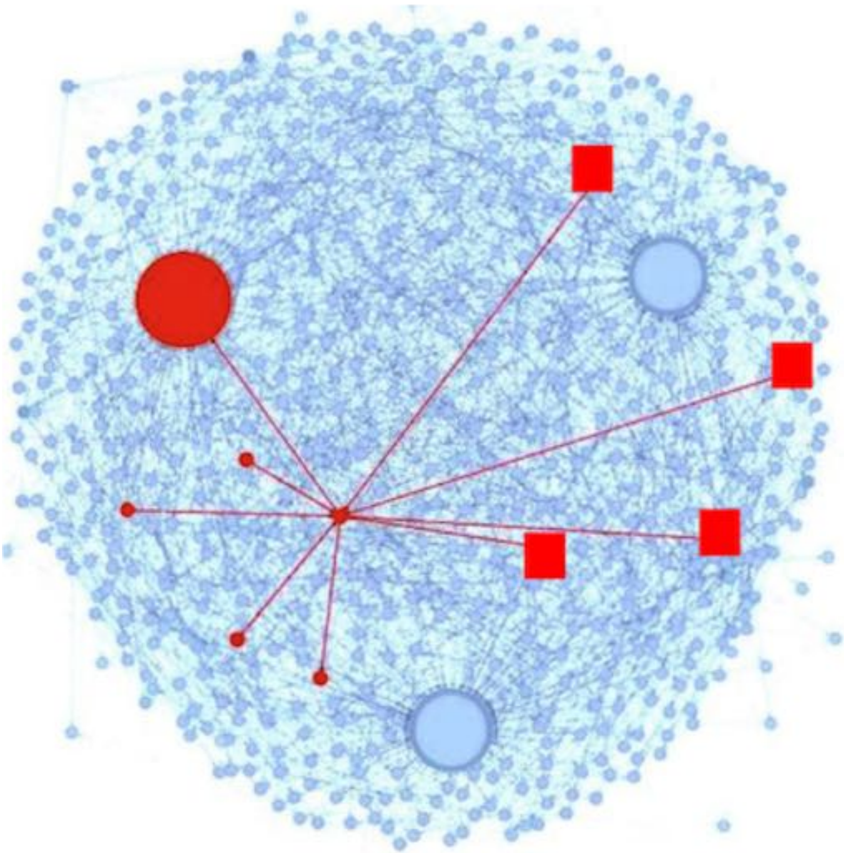
$$s_i \sim \text{Bin}(\text{points}, \theta)$$

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

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

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

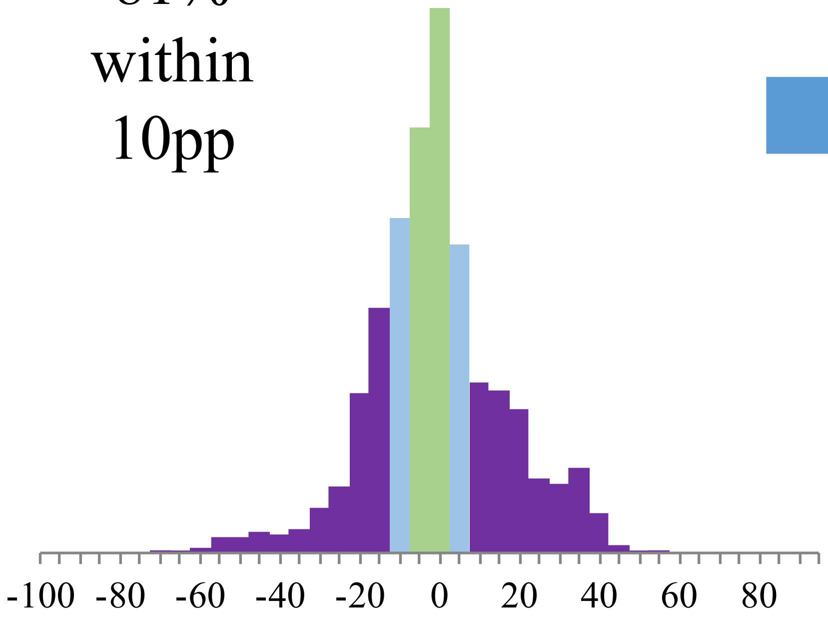
Inference or Machine Learning

Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Yes, With Probabilistic Modelling

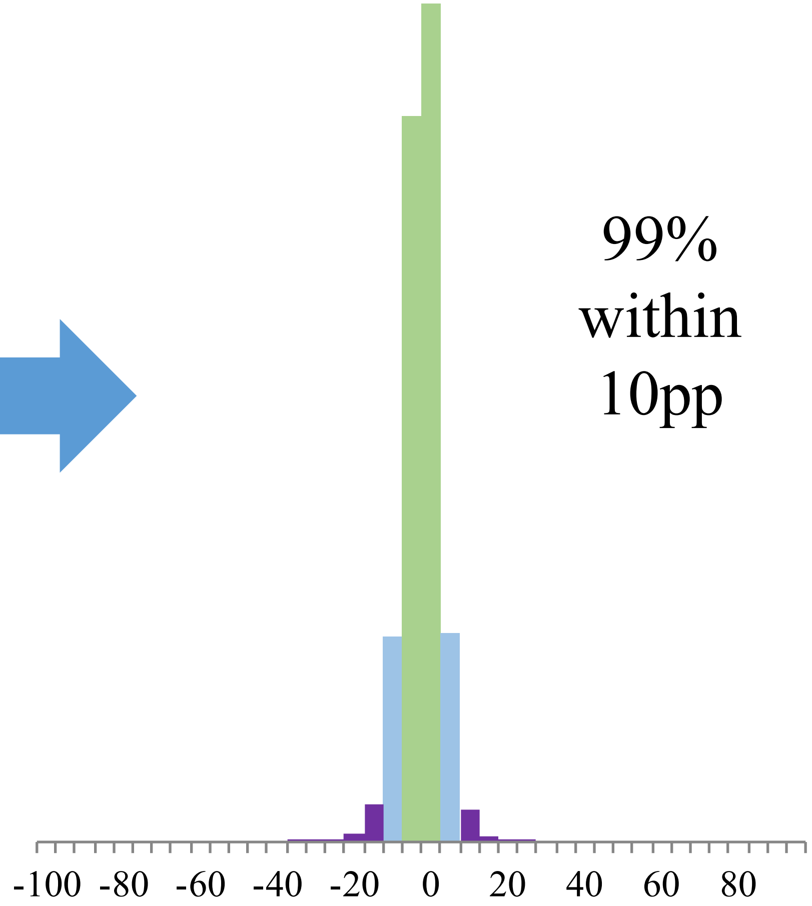
Before:

81% within 10pp



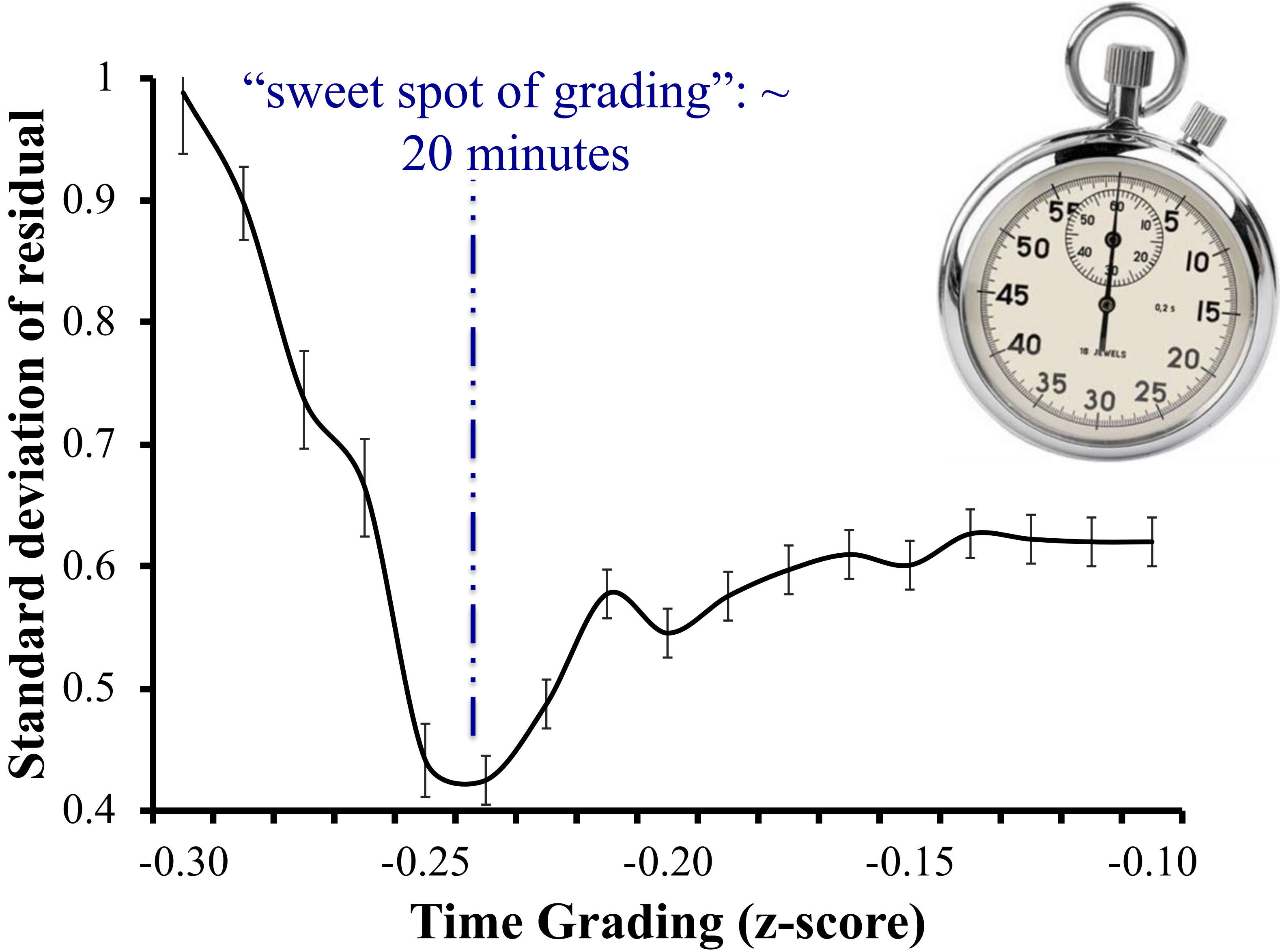
After:

99% within 10pp



Tuned Models of Peer Assessment. C Piech, J Huang, A Ng, D Koller

Grading Sweet Spot




1001

Voilà, c'est tout