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

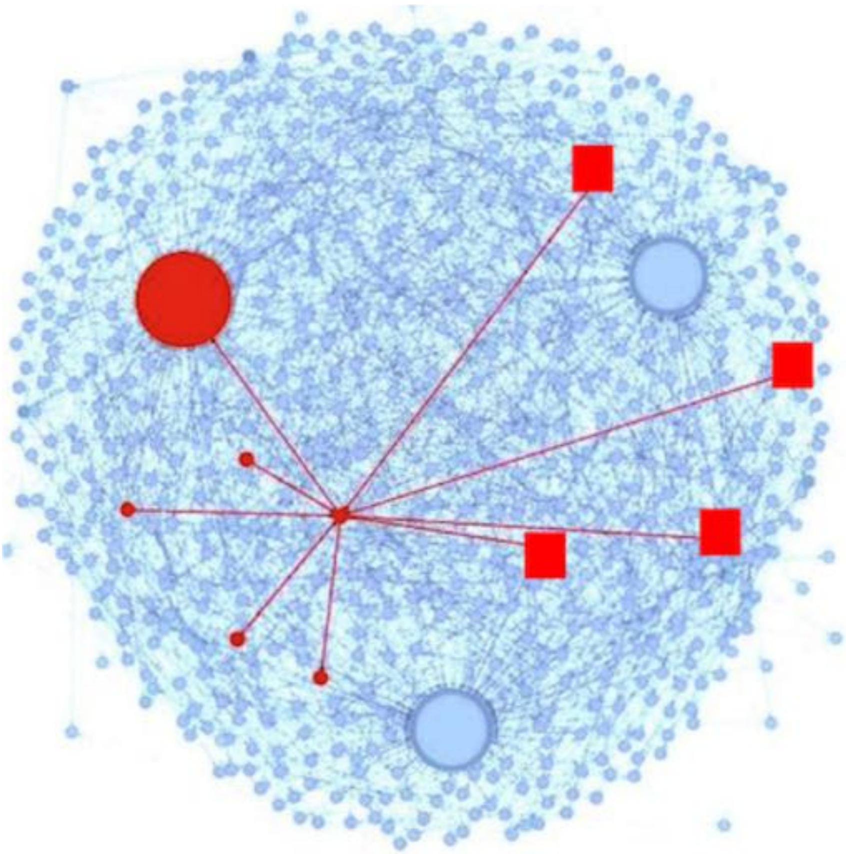
Variance

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.

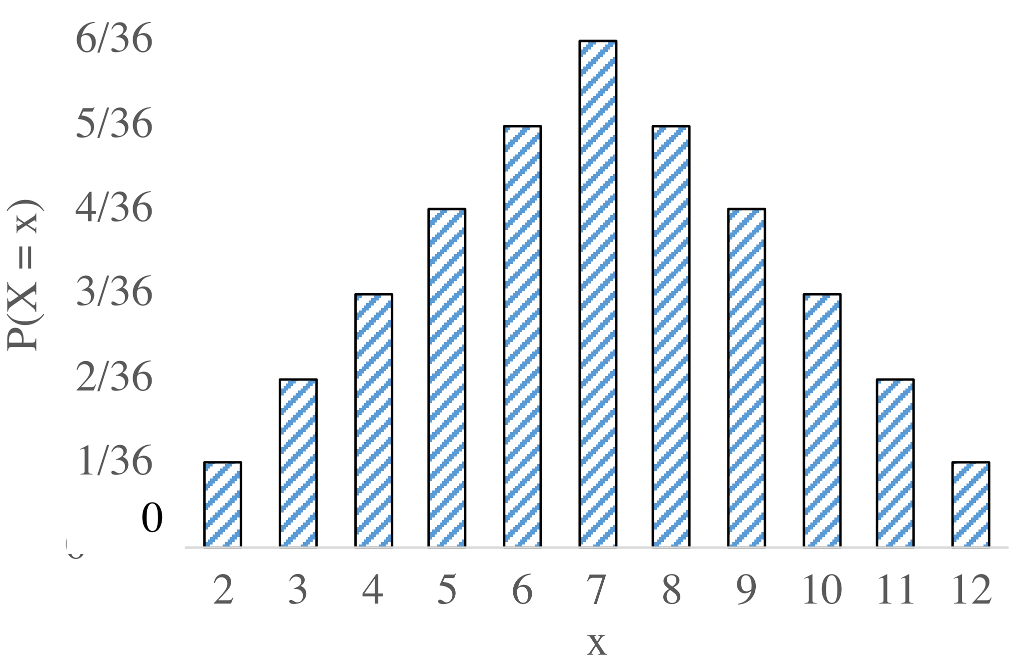
For example:
X is the sum of two dice rolled.

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

Review: Probability Mass Function

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

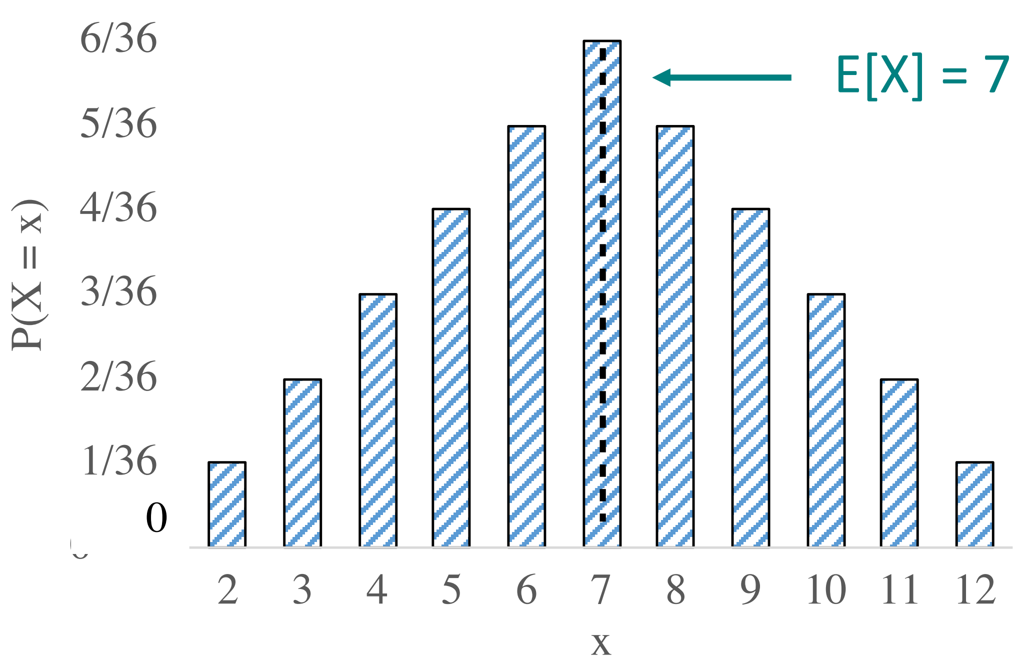
$$p_X(x) = P(\mathrm{X} = x)$$



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**:

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

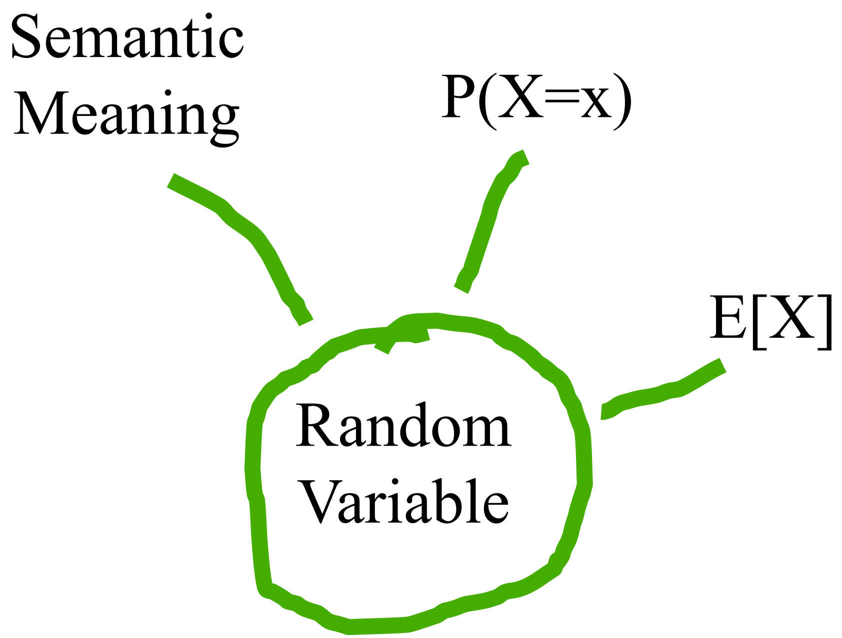
- **Expectation of a sum** is the sum of expectations

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

- **Unconscious statistician**:

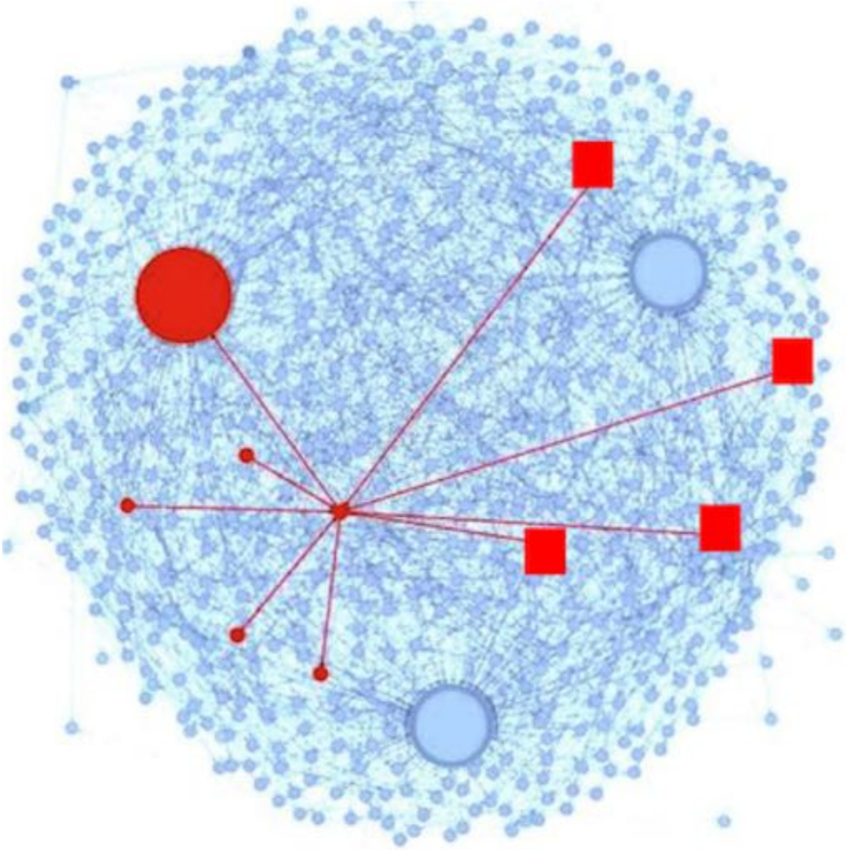
$$E[g(X)] = \sum_x g(x)p(x)$$

Fundamental Properties



Is E[X] enough?

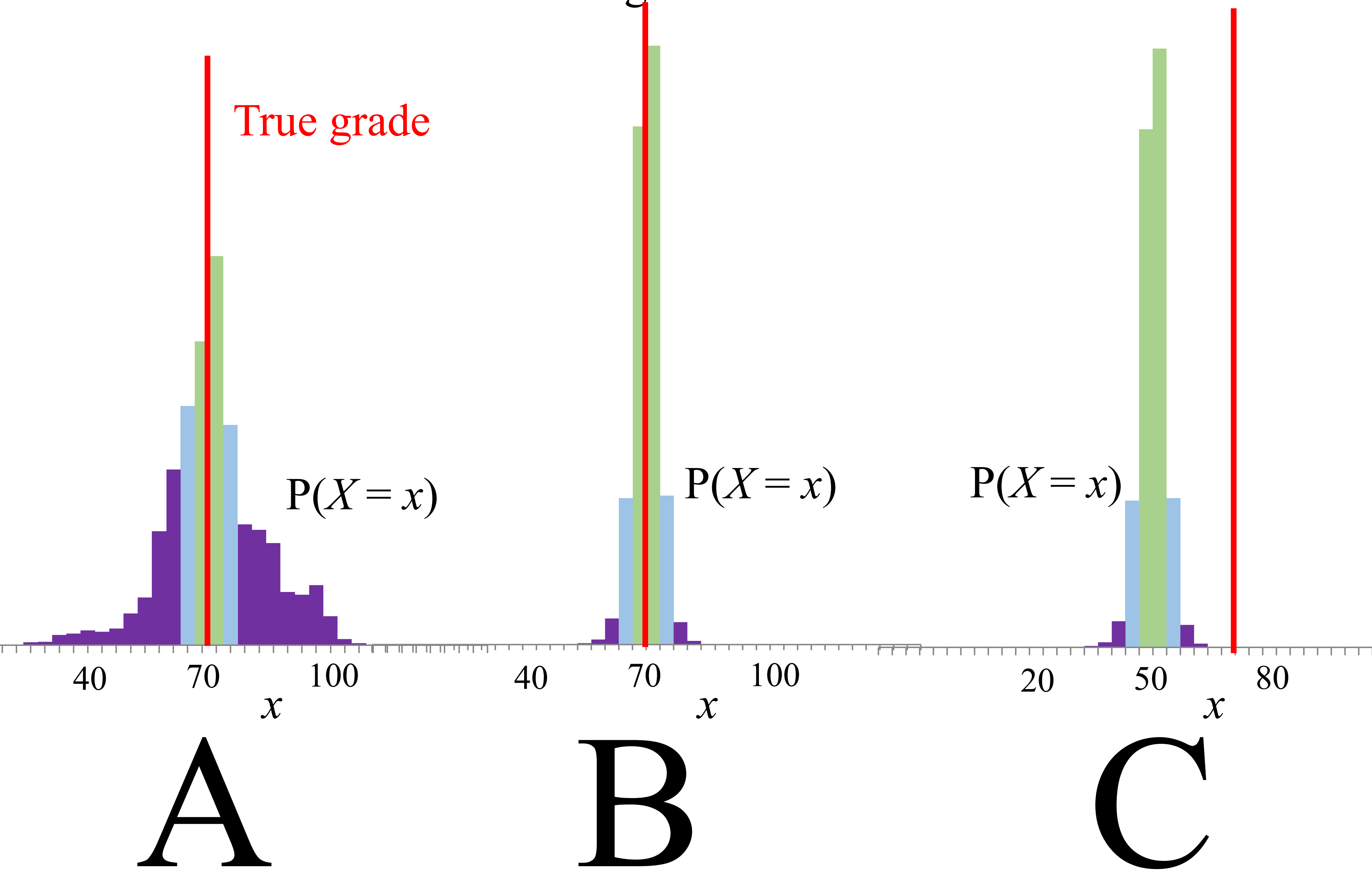
Intuition



Peer Grading on Coursera HCI.

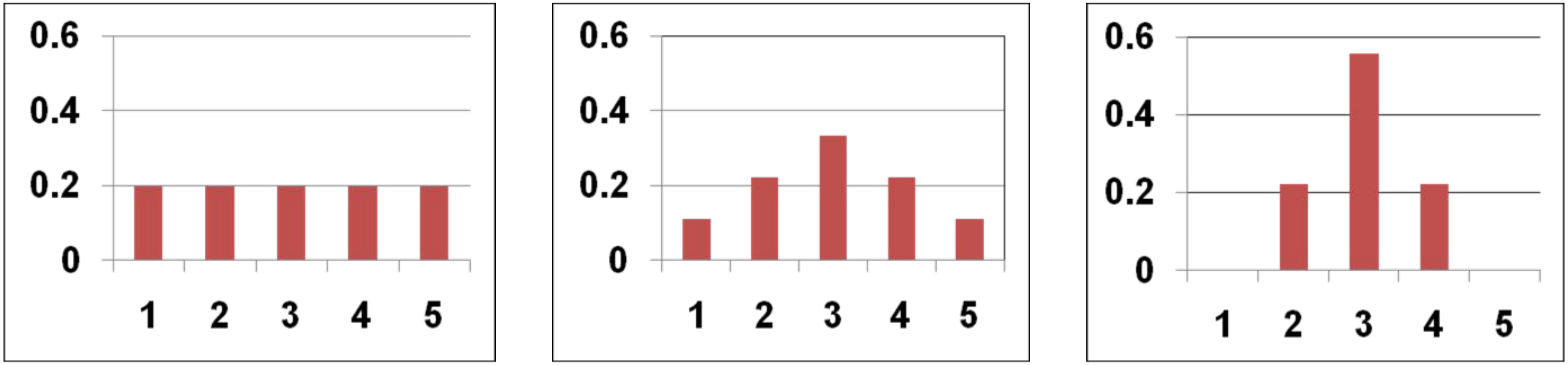
31,067 peer grades for 3,607 students.

X is the score peer graders give to an assignment submission with true grade 70



Variance

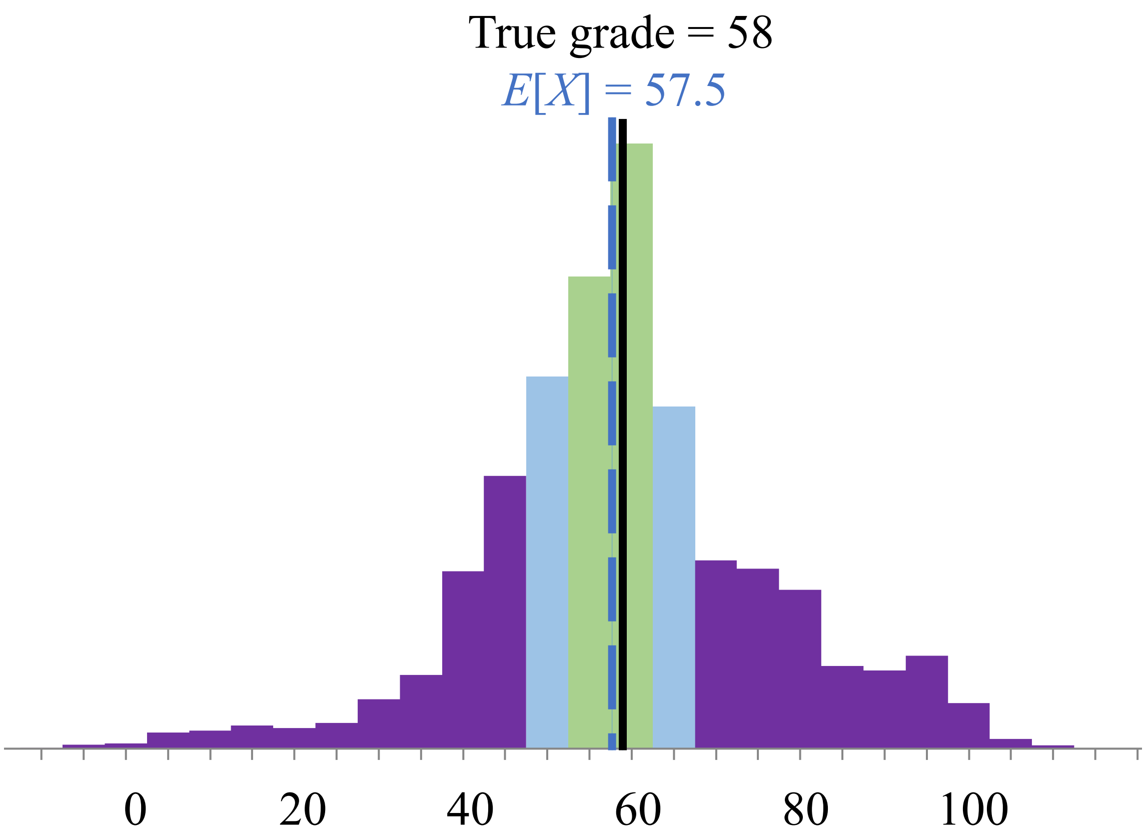
- Consider the following 3 distributions (PMFs)



- 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

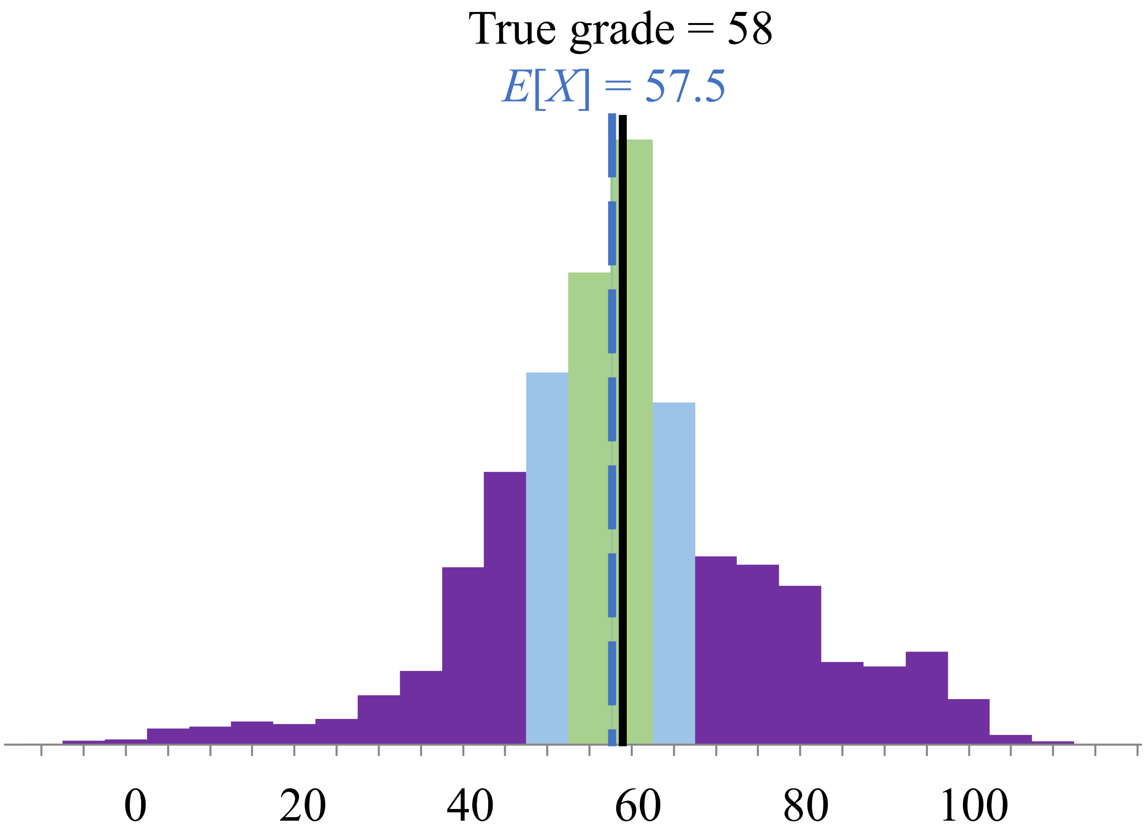
Let X be a random variable that represents a peer grade for an assignment that has a true grade of 58.



Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

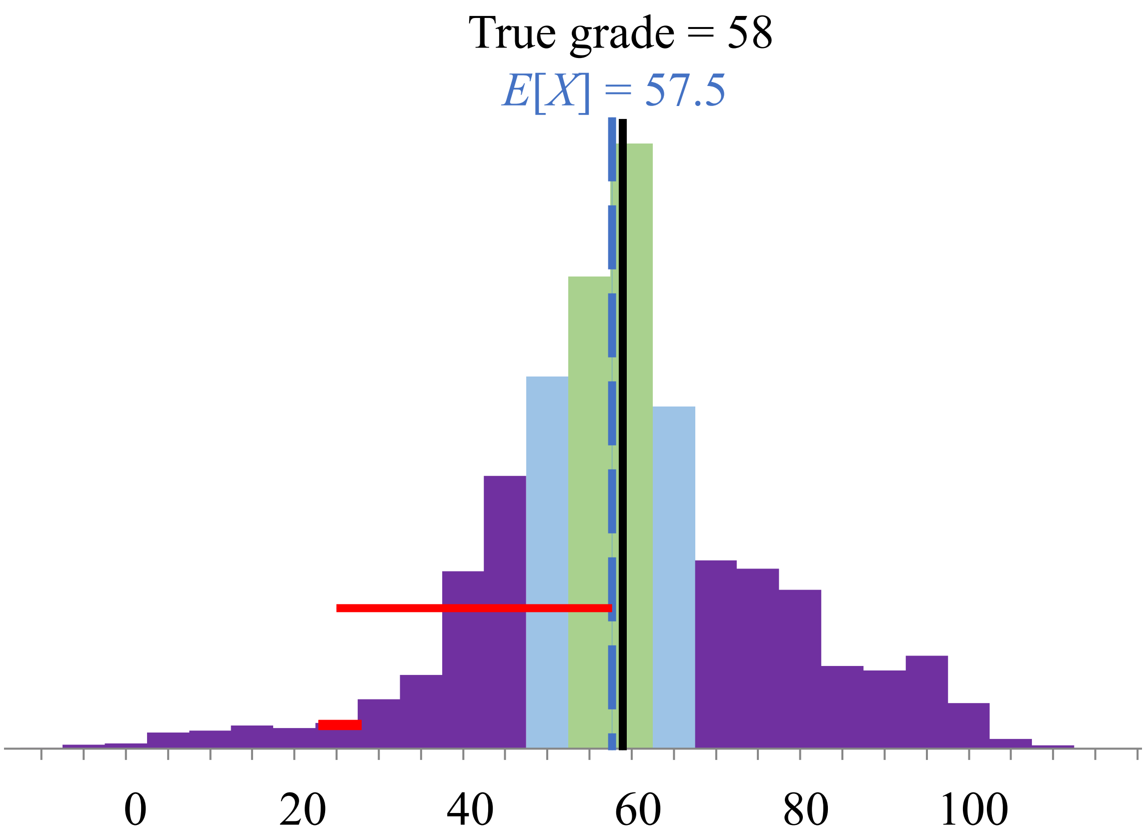
$$\mathrm{Var}(X) = E[(X - \mu)^2]$$



Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

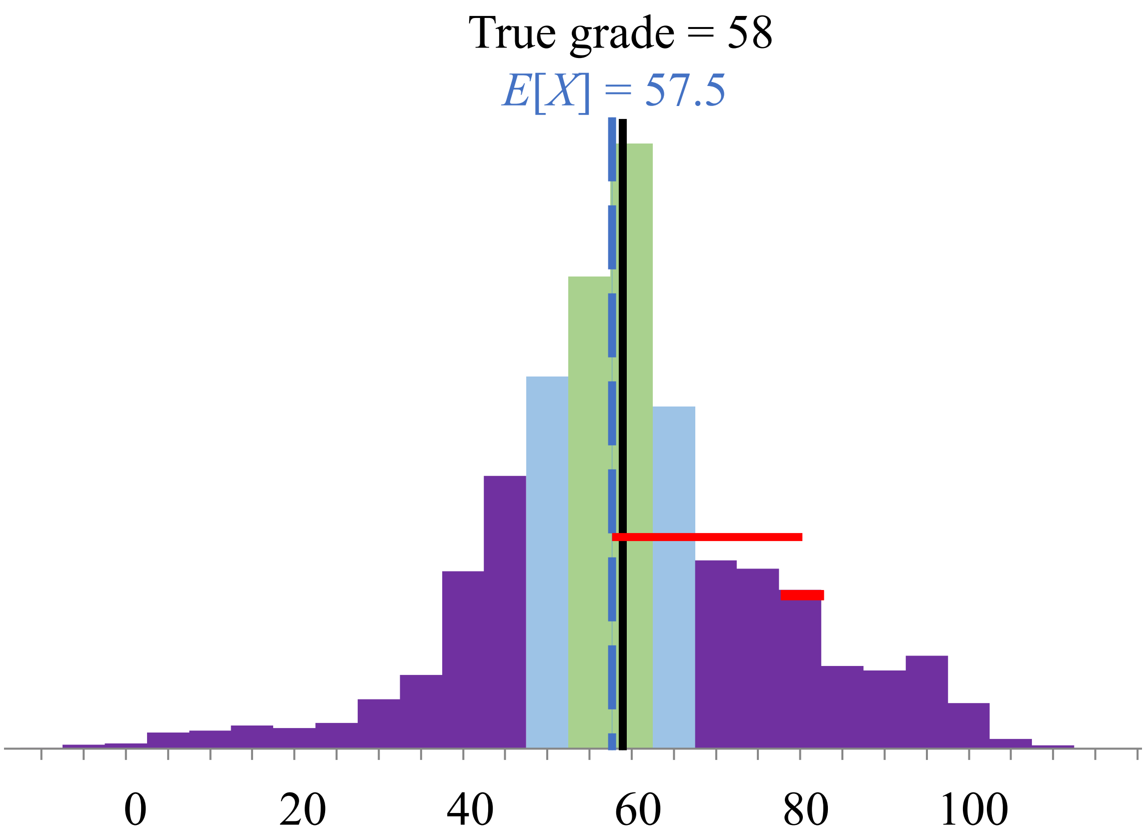


X	$(X - \mu)^2$
25 points	1056 points^2

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

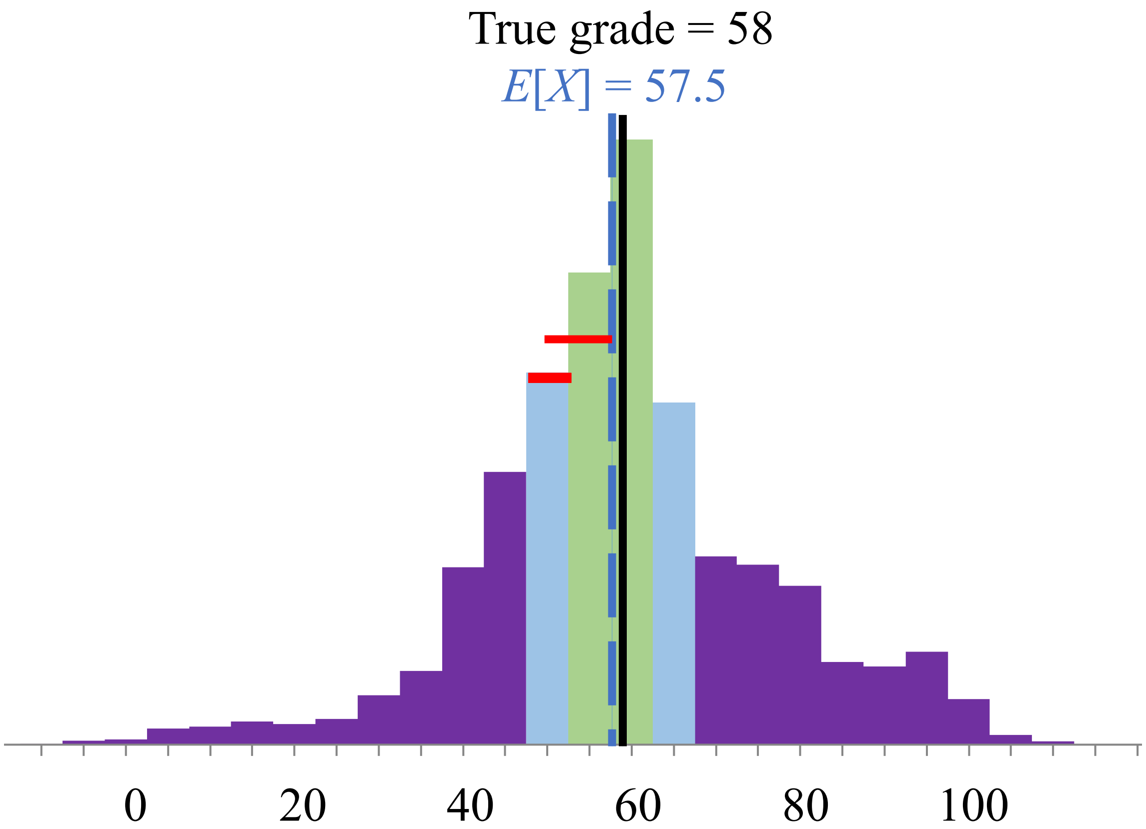


X	$(X - \mu)^2$
25 points	1056 points^2
80 points	506 points^2

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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

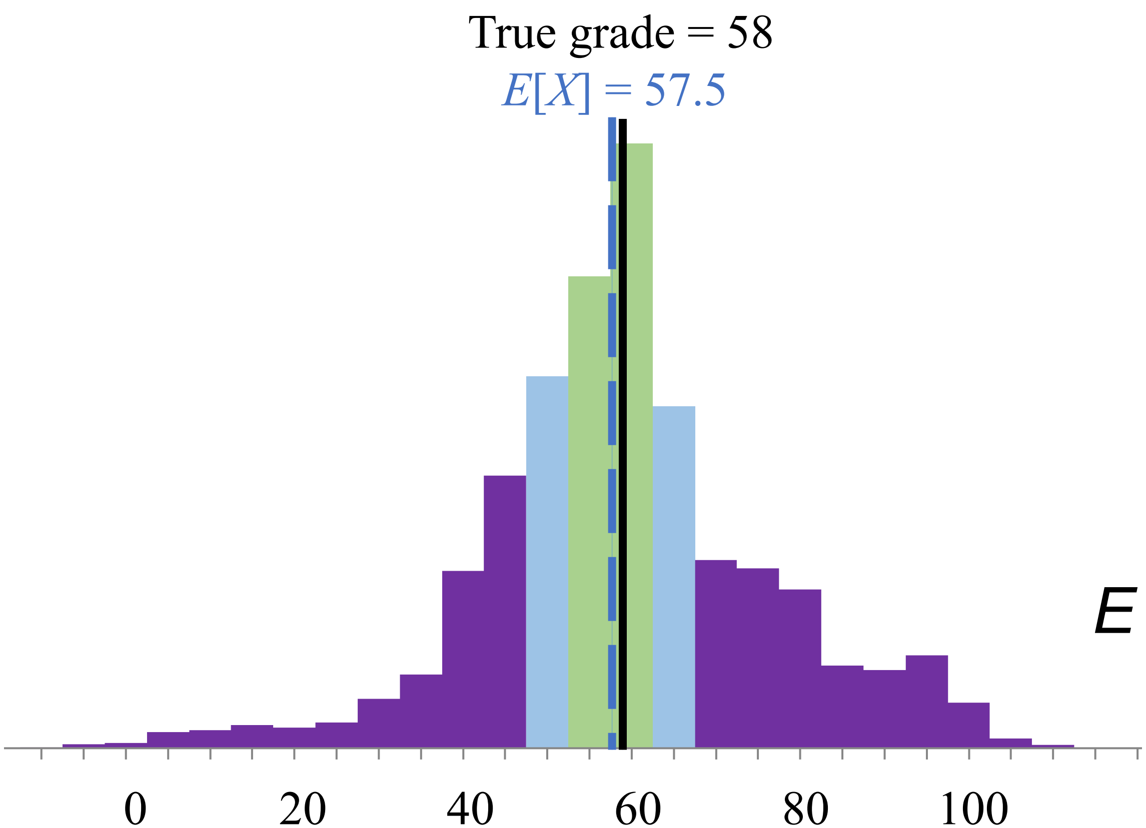


X	$(X - \mu)^2$
25 points	1056 points^2
80 points	506 points^2
50 points	56 points^2

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

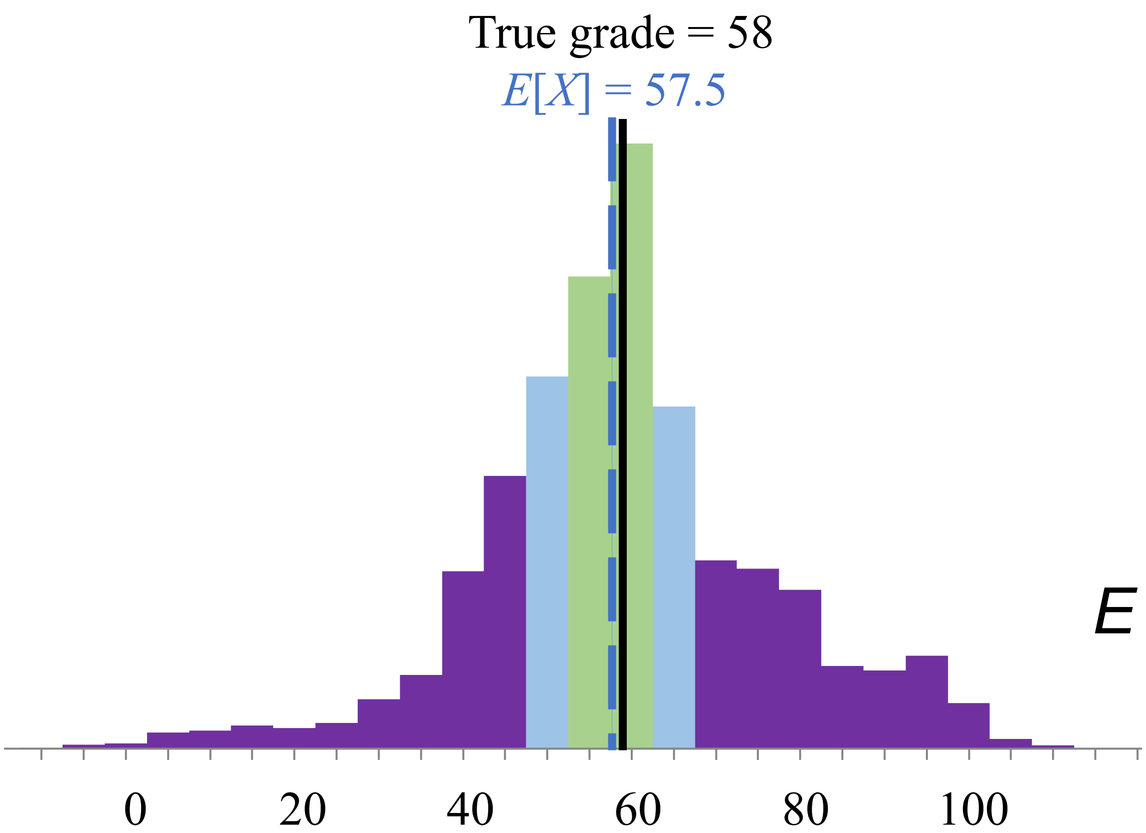
…

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

Peer Grades in Coursera HCI

Let X be a random variable that represents a peer grade

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



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

…

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

$$Std(X) = 7.2 \text{ points}$$

Variance

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

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

- Note: Var(X) ≥ 0

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

Computing Variance

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

Recall: Unconscious statistician:

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

$$\text{let } g(X) = (X-\mu)^2$$

Computing Variance

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

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]

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

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

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

Properties of Variance

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

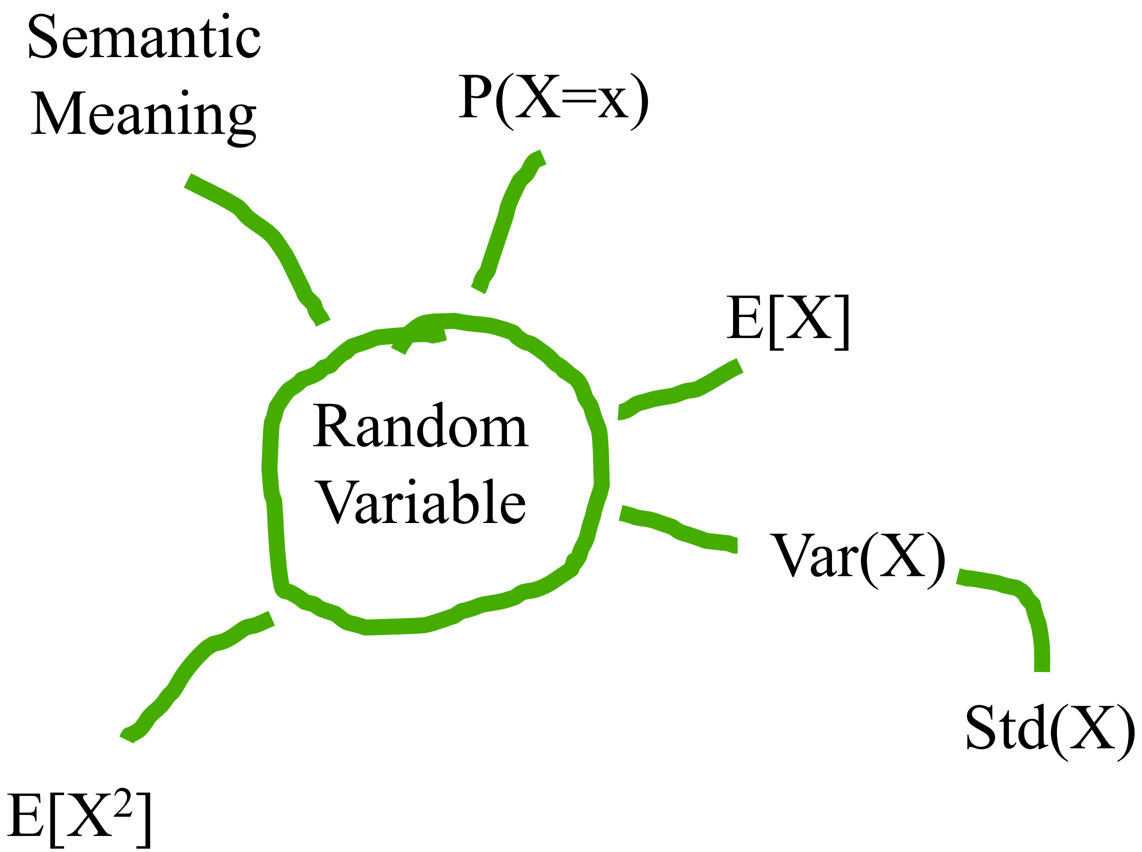
$$
\begin{aligned}
\mathrm{Var}(aX + b) &= E[(aX + b)^2] - (E[aX + b])^2 \\
&= E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\
&= a^2E[X^2] + 2abE[X] + b^2 - (a^2(E[X])^2 + 2abE[X] + b^2) \\
&= a^2E[X^2] - a^2(E[X])^2 = a^2(E[X^2] - (E[X])^2) \\
&= a^2\mathrm{Var}(X)
\end{aligned}
$$

- Standard Deviation of X, denoted SD(X), is:

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

  - 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
- Ice Cube at a renaissance fair?

Bernoulli Random Variable

- Experiment results in "Success" or "Failure"
  - X is random **indicator** variable (1 = success, 0 = failure)
  - $P(X = 1) = p$ $\qquad$ $P(X = 0) = 1 - p$
  - X is a **Bernoulli** Random Variable: $X \sim \mathrm{Ber}(p)$
  - $E[X] = p$
  - $\mathrm{Var}(X) = p(1 - p)$
- 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 probability $p = 0.1$,

works with probability $(1 - p)$

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

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

$$\boldsymbol{X} \sim \mathrm{Ber}(p = 0.1)$$

Does a User Click an Ad?

Serve an ad, clicked with probability $p = 0.01$,

ignored with prob. $(1 - p)$

C: 1 if ad is clicked

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

$$C \sim \text{Ber}(p = 0.01)$$

More!

Binomial Random Variable

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

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

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

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

Binomial distribution

From Wikipedia, the free encyclopedia

"Binomial model" redirects here. For the binomial model in options pricing, see Binomial options pricing model.

See also: Negative binomial distribution

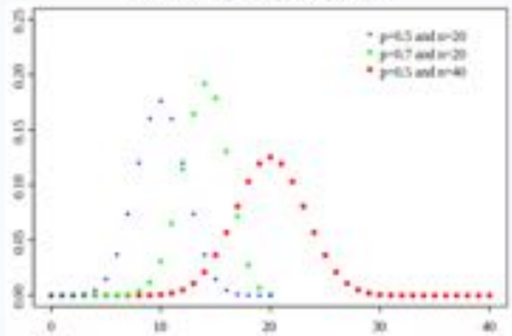
In probability theory and statistics, the **binomial distribution** with parameters *n* and *p* is the discrete probability distribu… of *n* independent experiments, each asking a yes–no question, and each with its own boolean-valued outcome: a random va… success/yes/true/one (with probability *p*) or failure/no/false/zero (with probability *q* = 1 − *p*). A single success/failure experime… Bernoulli experiment and a sequence of outcomes is called a Bernoulli process; for a single trial, i.e., *n* = 1, the binomial distr… binomial distribution is the basis for the popular binomial test of statistical significance.

The binomial distribution is frequently used to model the number of successes in a sample of size *n* drawn with replacement… is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric dis… much larger than *n*, the binomial distribution remains a good approximation, and is widely used.

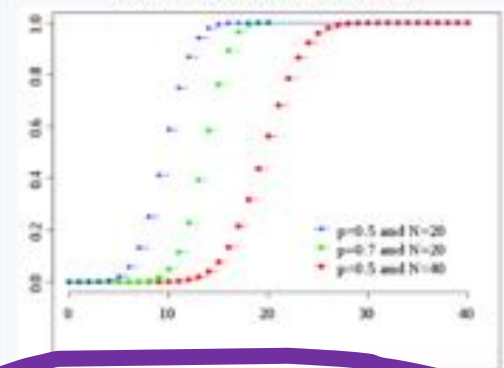
Contents [hide]

Binomial distribution

Probability mass function



Cumulative distribution function



Notation	$B(n, p)$
Parameters	$n \in \mathbf{N}_0$ — number of trials $p \in [0,1]$ — success probability in each trial
Support	$k \in \{0, \ldots, n\}$ — number of successes
pmf	$\binom{n}{k} p^k (1-p)^{n-k}$
CDF	$I_{1-p}(n-k, 1+k)$
Mean	np
Median	$\lfloor np \rfloor$ or $\lceil np \rceil$
Mode	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$
Variance	$np(1-p)$
Skewness	$\dfrac{1-2p}{\sqrt{np(1-p)}}$
Ex. kurtosis	$\dfrac{1-6p(1-p)}{np(1-p)}$
Entropy	$\frac{1}{2}\log_2(2\pi e\, np(1-p)) + O\left(\frac{1}{\ldots}\right)$

I Really Want the Proof of Var :)

$$E(X^2) = \sum_{k\ge 0}^{n} k^2 \binom{n}{k} p^k q^{n-k}$$ Definition of Binomial Distribution: $p + q = 1$

$$= \sum_{k=0}^{n} kn \binom{n-1}{k-1} p^k q^{n-k}$$ Factors of Binomial Coefficient: $k\binom{n}{k} = n\binom{n-1}{k-1}$

$$= np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)}$$ Change of limit: term is zero when $k - 1 = 0$

$$= np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j}$$ putting $j = k - 1, m = n - 1$

$$= np \left(\sum_{j=0}^{m} j \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right)$$ splitting sum up into two

$$= np \left(\sum_{j=0}^{m} m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right)$$ Factors of Binomial Coefficient: $j\binom{m}{j} = m\binom{m-1}{j-1}$

$$= np \left((n-1) p \sum_{j=1}^{m} \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right)$$ Change of limit: term is zero when $j - 1 = 0$

$$= np \left((n-1) p (p+q)^{m-1} + (p+q)^m \right)$$ Binomial Theorem

$$= np \left((n-1) p + 1 \right)$$ as $p + q = 1$

$$= n^2 p^2 + np(1-p)$$ by algebra

How Many Program Crashes?

n runs of program, each crashes with probability $p = 0.1$, works with probability $(1 - p)$.

What is the probability of exactly 2 crashes with 100 users?

H: number of crashes

$$H \sim \text{Bin}(n = 100, p = 0.1)$$

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

$$P(H = 2) = \binom{100}{2} (0.1)^2 (0.9)^{98}$$

How Many Program Crashes?

n runs of program, each crashes with probability $p = 0.1$, works with probability $(1 - p)$.

What is the probability of < 3 crashes with 100 users?

H: number of crashes

$$H \sim \text{Bin}(n = 100, p = 0.1)$$

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

$$P(H < 3) = \sum_{i=0}^{2} \binom{100}{i}(0.1)^i(0.9)^{100-i}$$

How Many Ads Clicked?

1000 ads served, each clicked with $p = 0.01$, otherwise ignored.
Expectation and Standard deviation of number of ads clicked?

$\boldsymbol{H}$: number of clicks

$$\boldsymbol{H} \sim \mathrm{Bin}(n = 1000, p = 0.01)$$

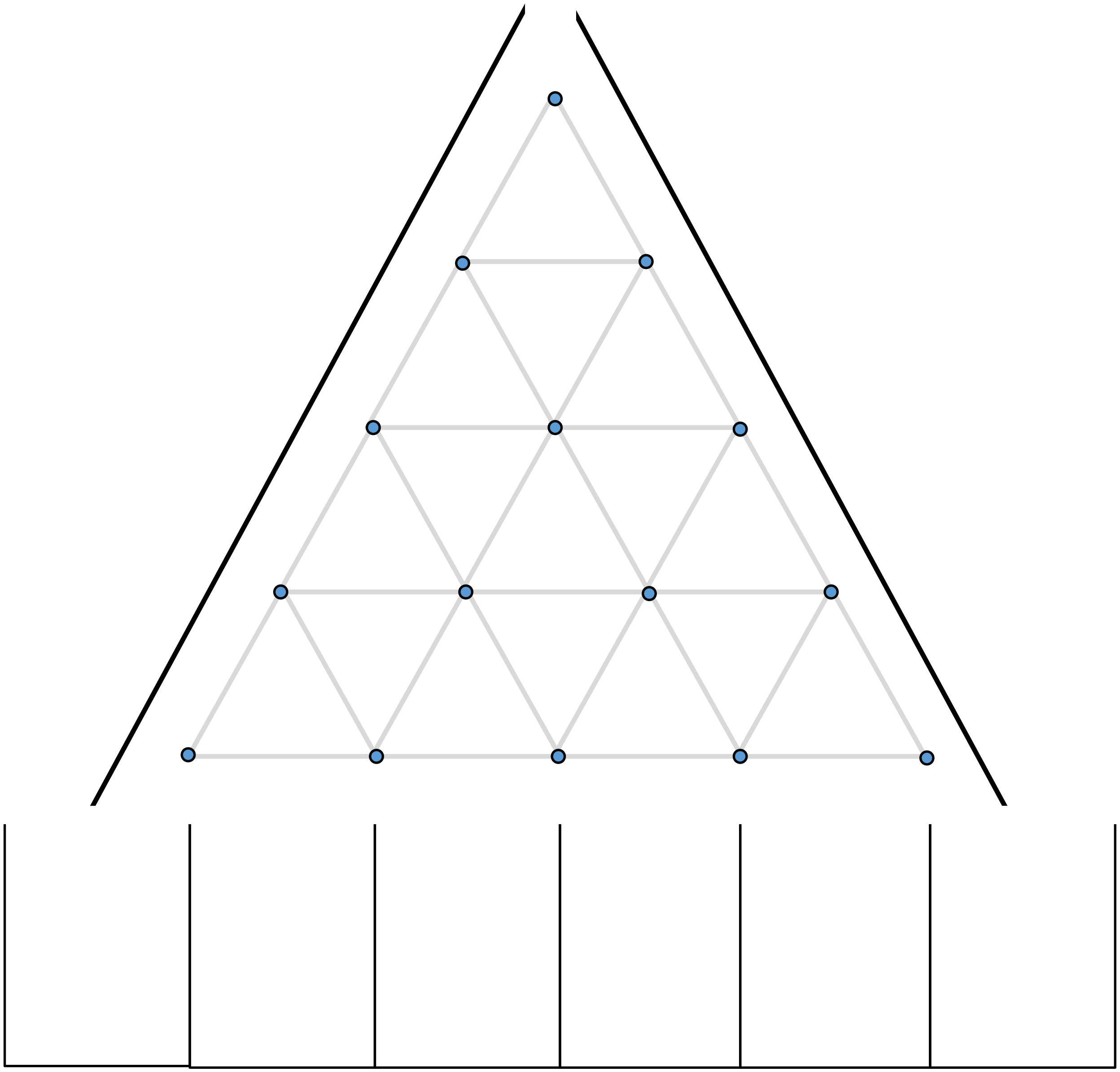
$$\mathbf{P}(\boldsymbol{H} = k) = \binom{1000}{k}(0.01)^k(0.99)^{1000-k}$$

$$\mathrm{E}(\boldsymbol{H}) = np = 10$$

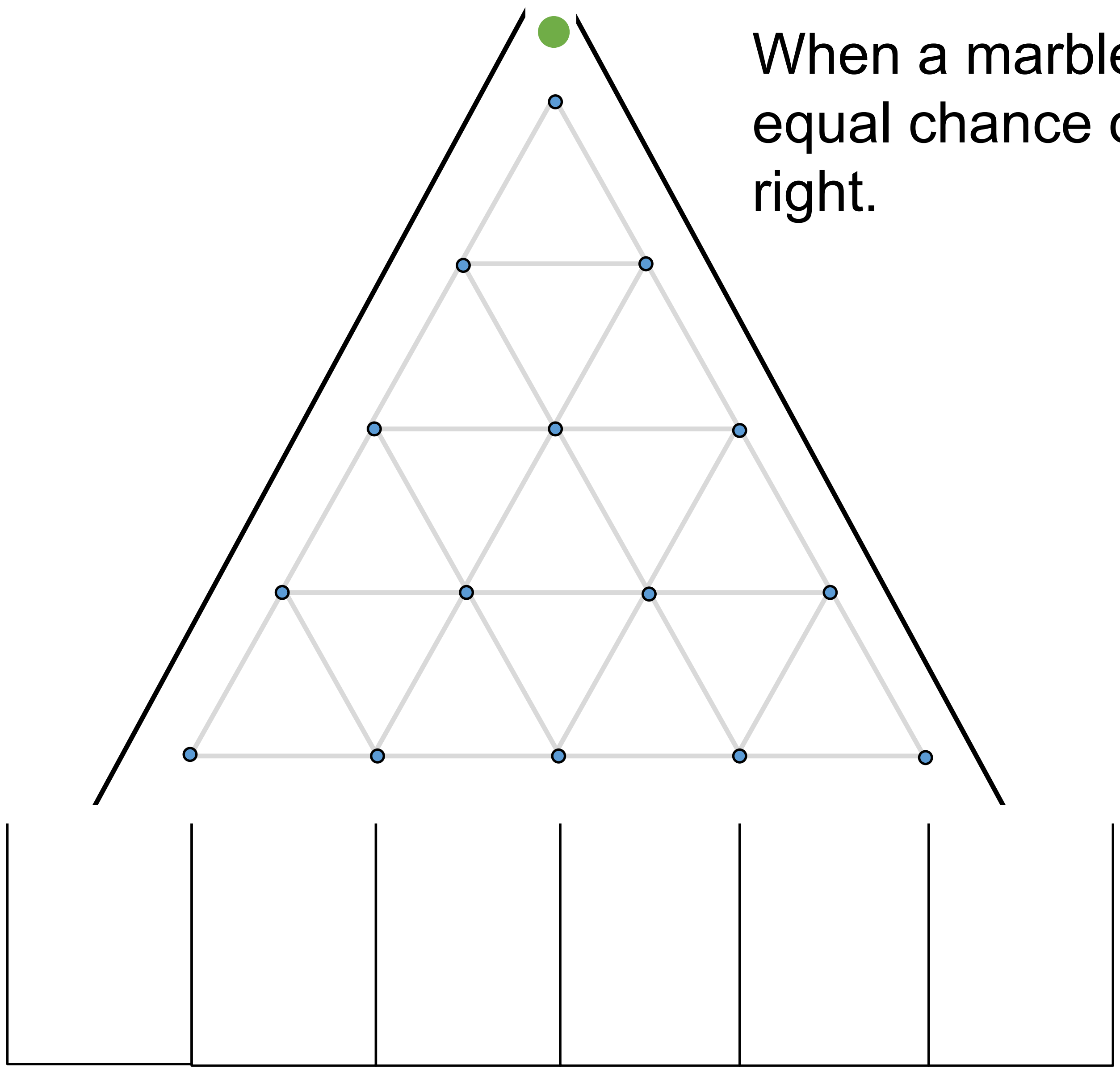
$$\mathrm{Var}(\boldsymbol{H}) = np(1\text{-}p) = 9.9$$

$$\mathrm{Std}(\boldsymbol{H}) = 3.15$$

Galton Board

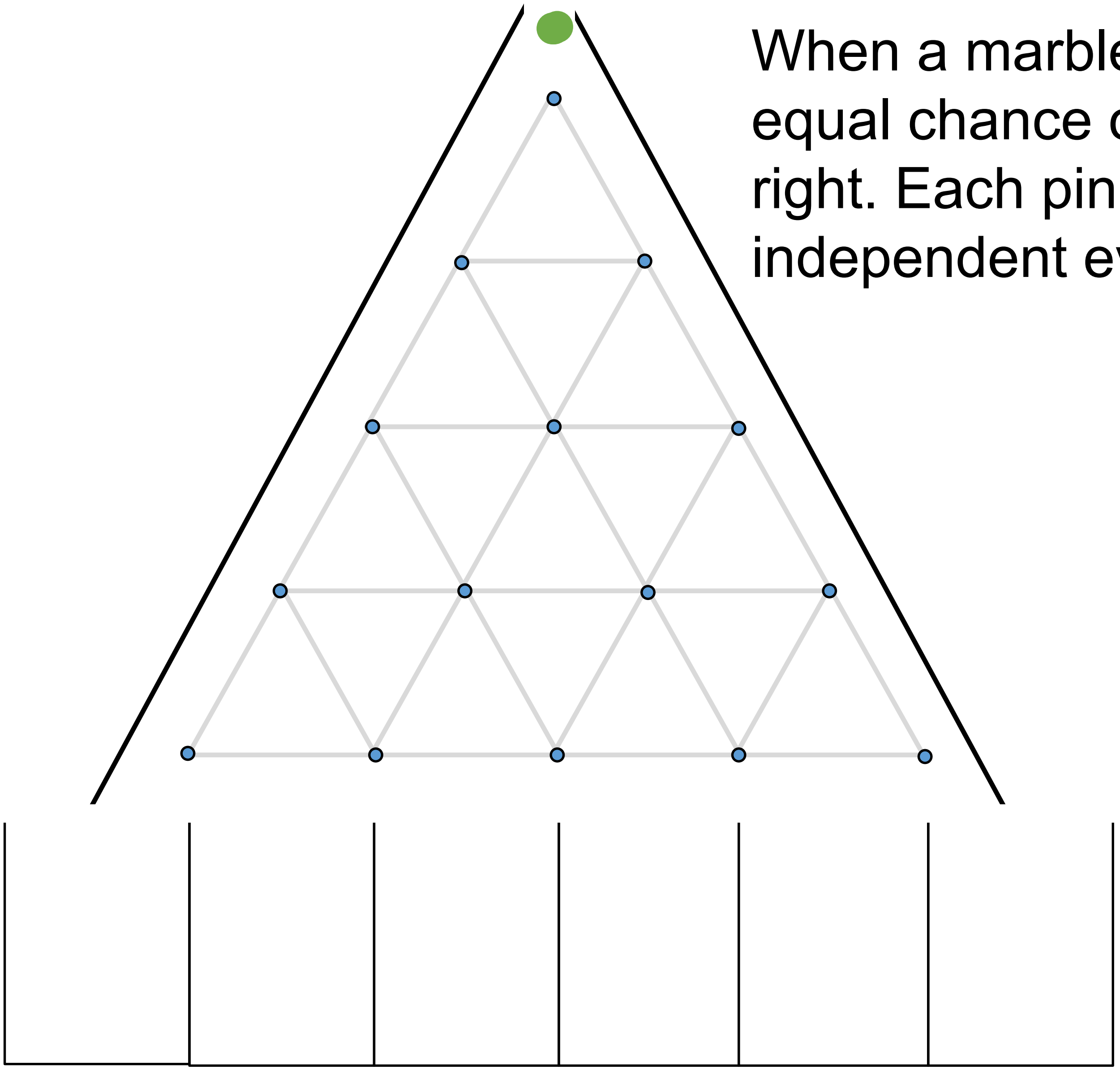


Galton Board



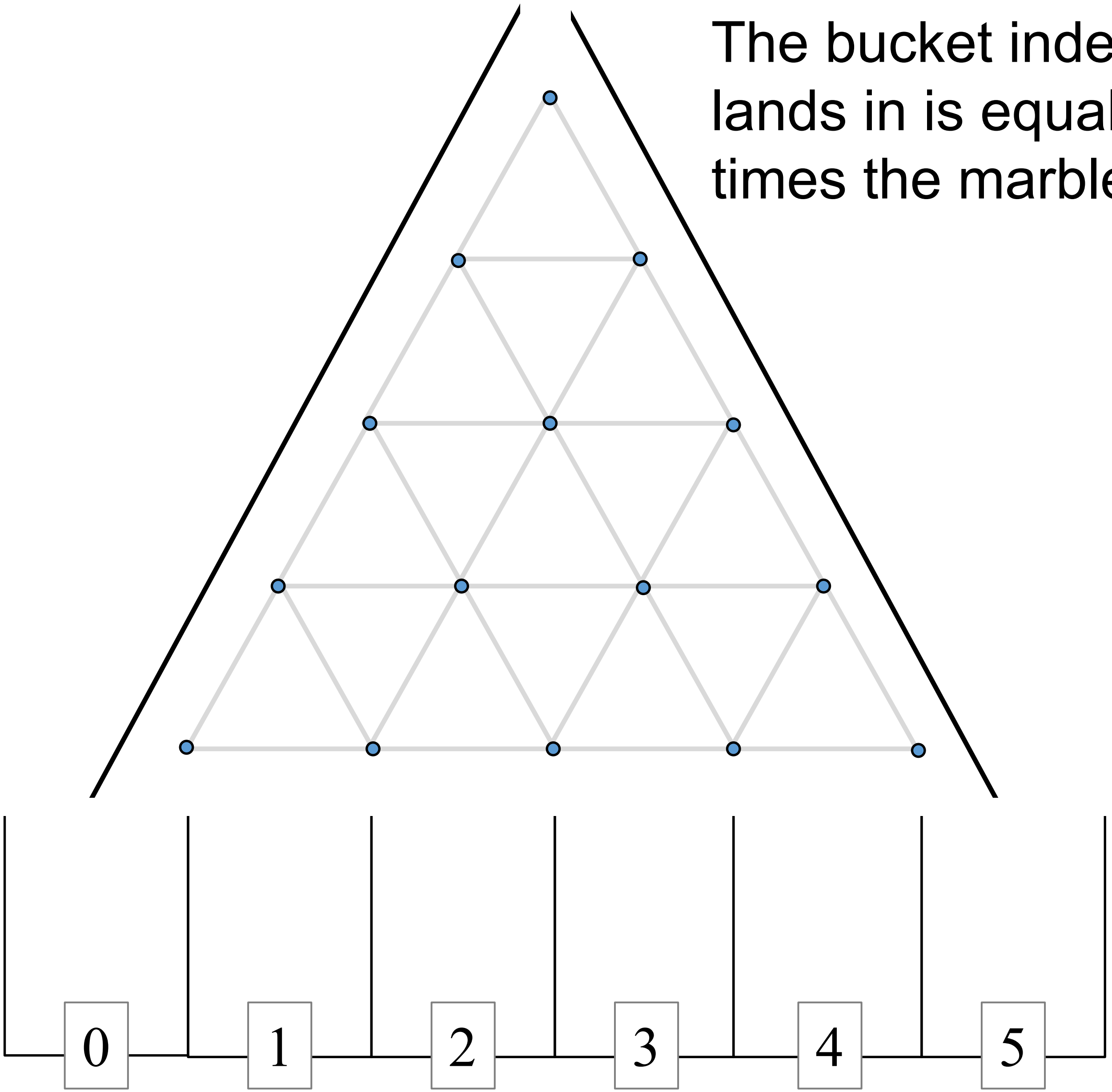
When a marble hits a pin, it has equal chance of going left or right.

Galton Board



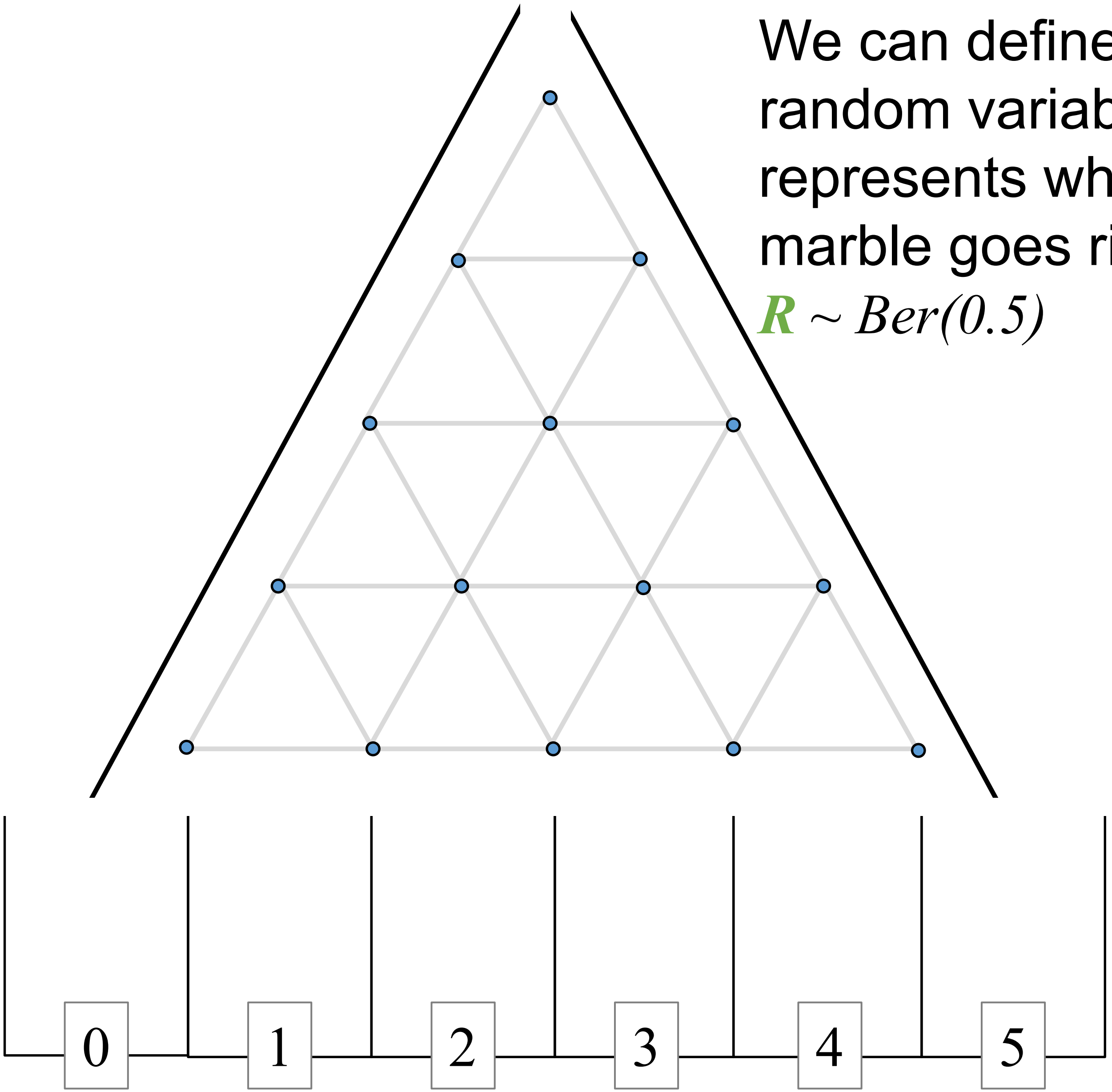
When a marble hits a pin, it has equal chance of going left or right. Each pin represents an independent event.

Galton Board



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

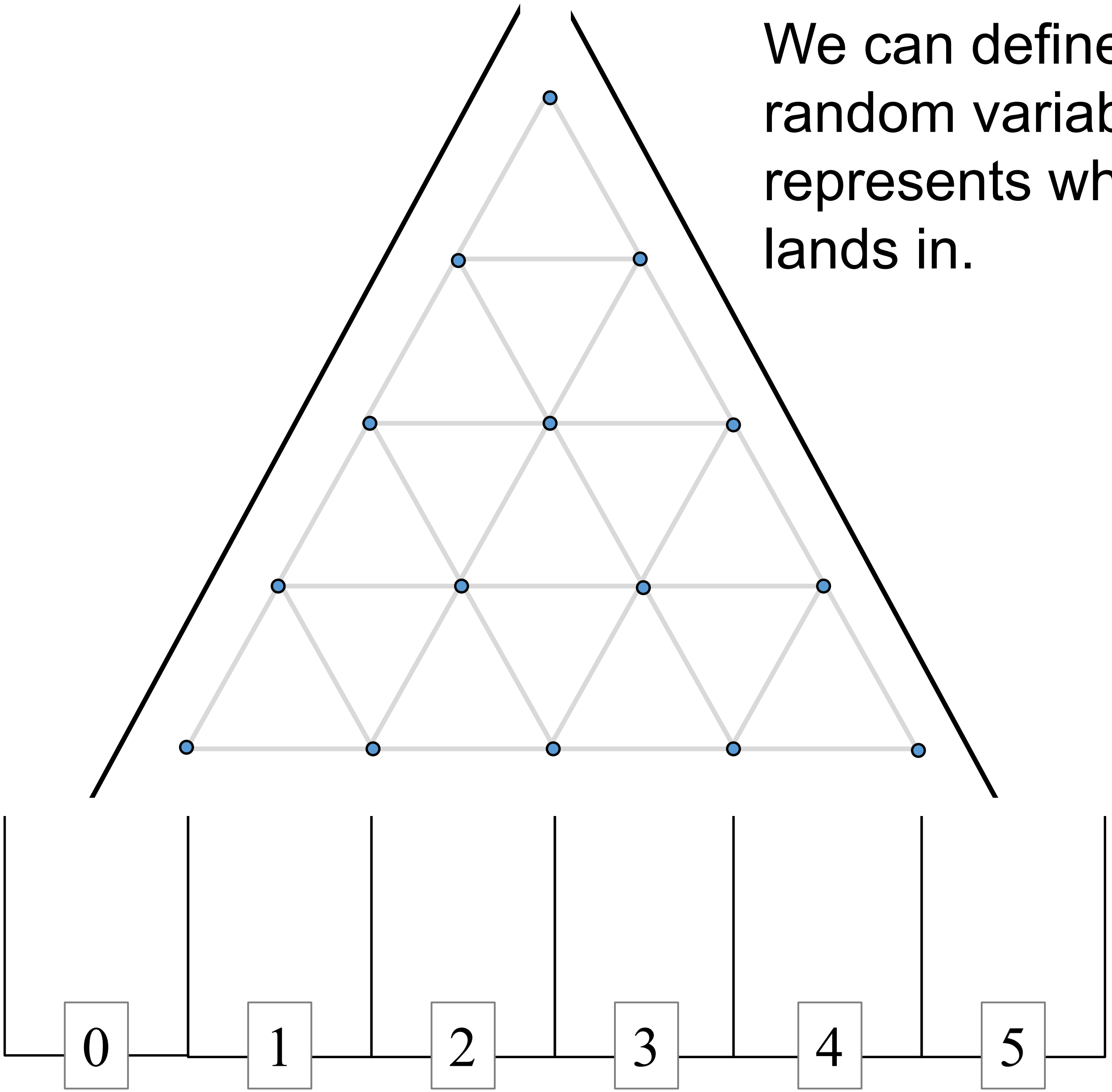
Galton Board



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

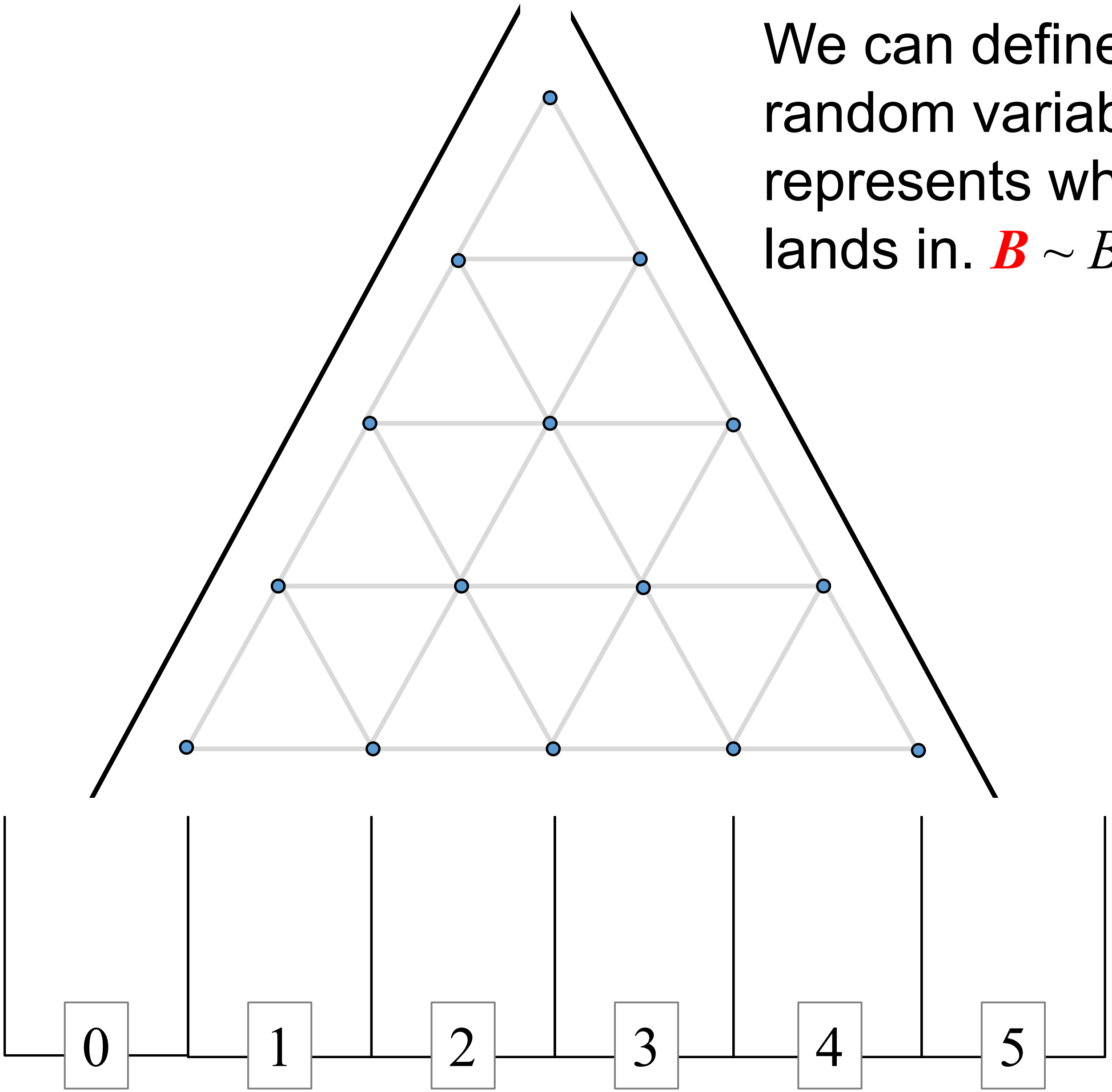
$R \sim Ber(0.5)$

Galton Board



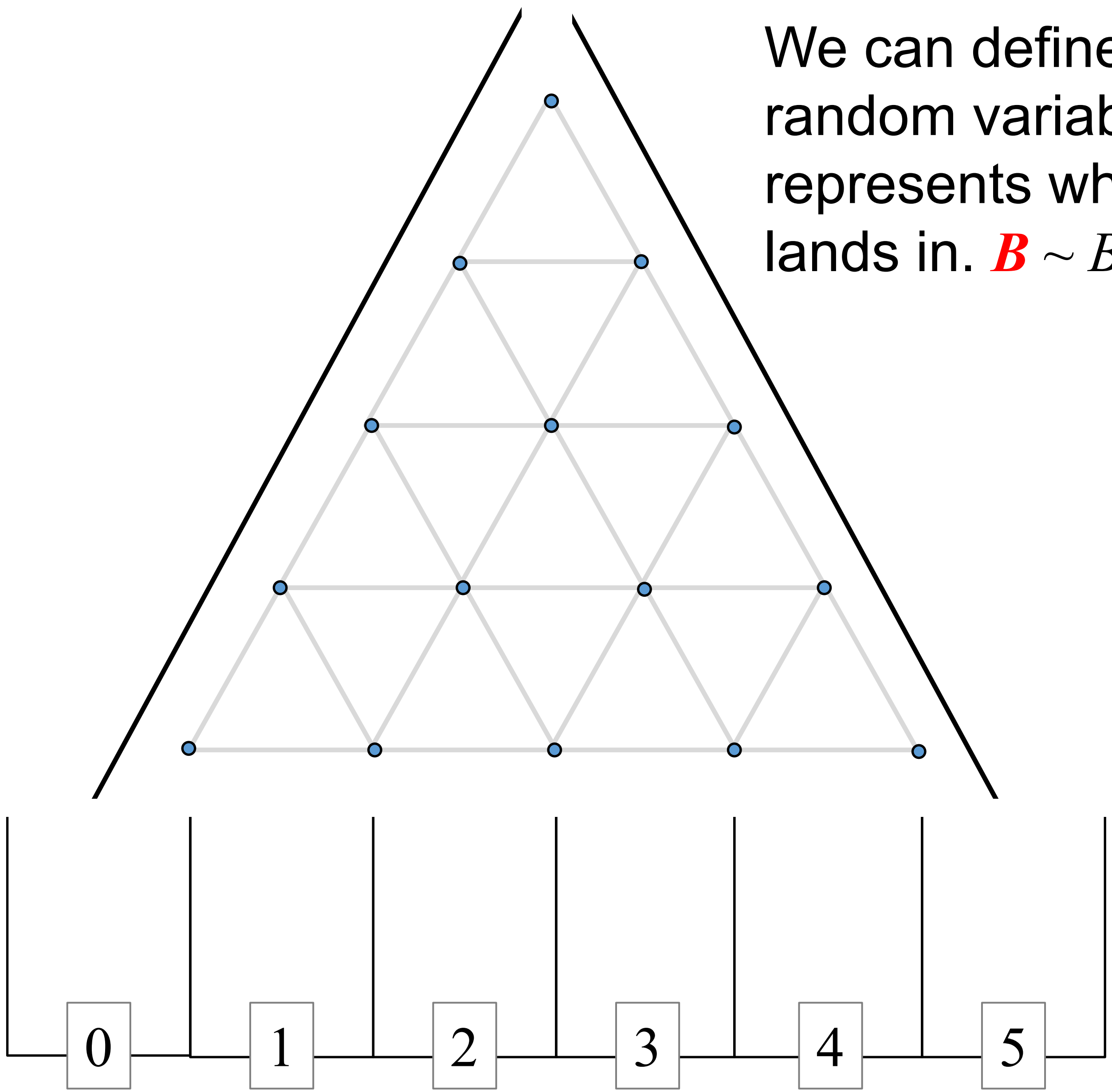
We can define an indicator random variable (***B***) which represents what bucket a marble lands in.

Galton Board



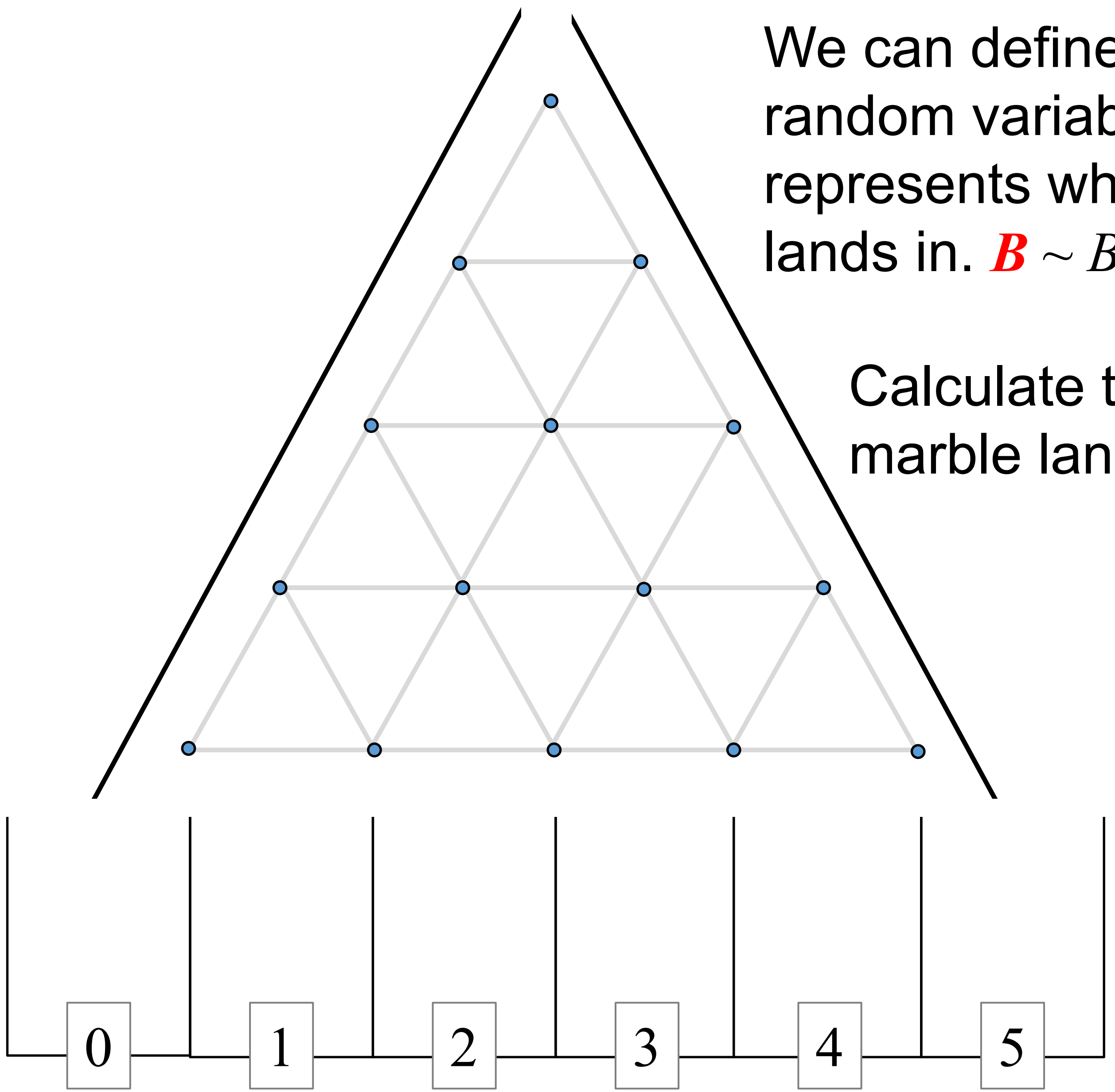
We can define an indicator random variable (***B***) which represents what bucket a marble lands in. $B \sim Bin(\text{levels}, 0.5)$

Galton Board



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

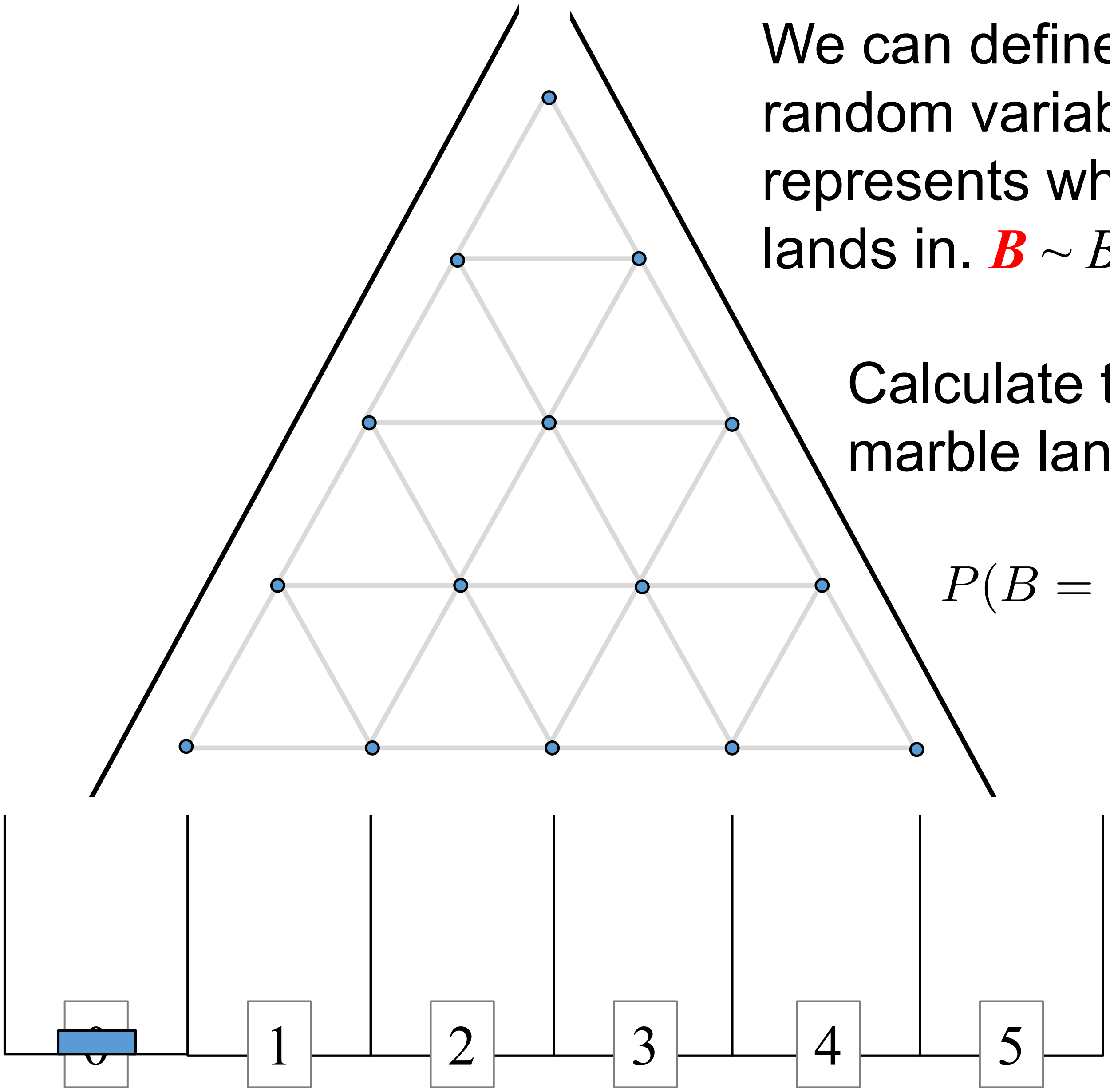
Galton Board



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.

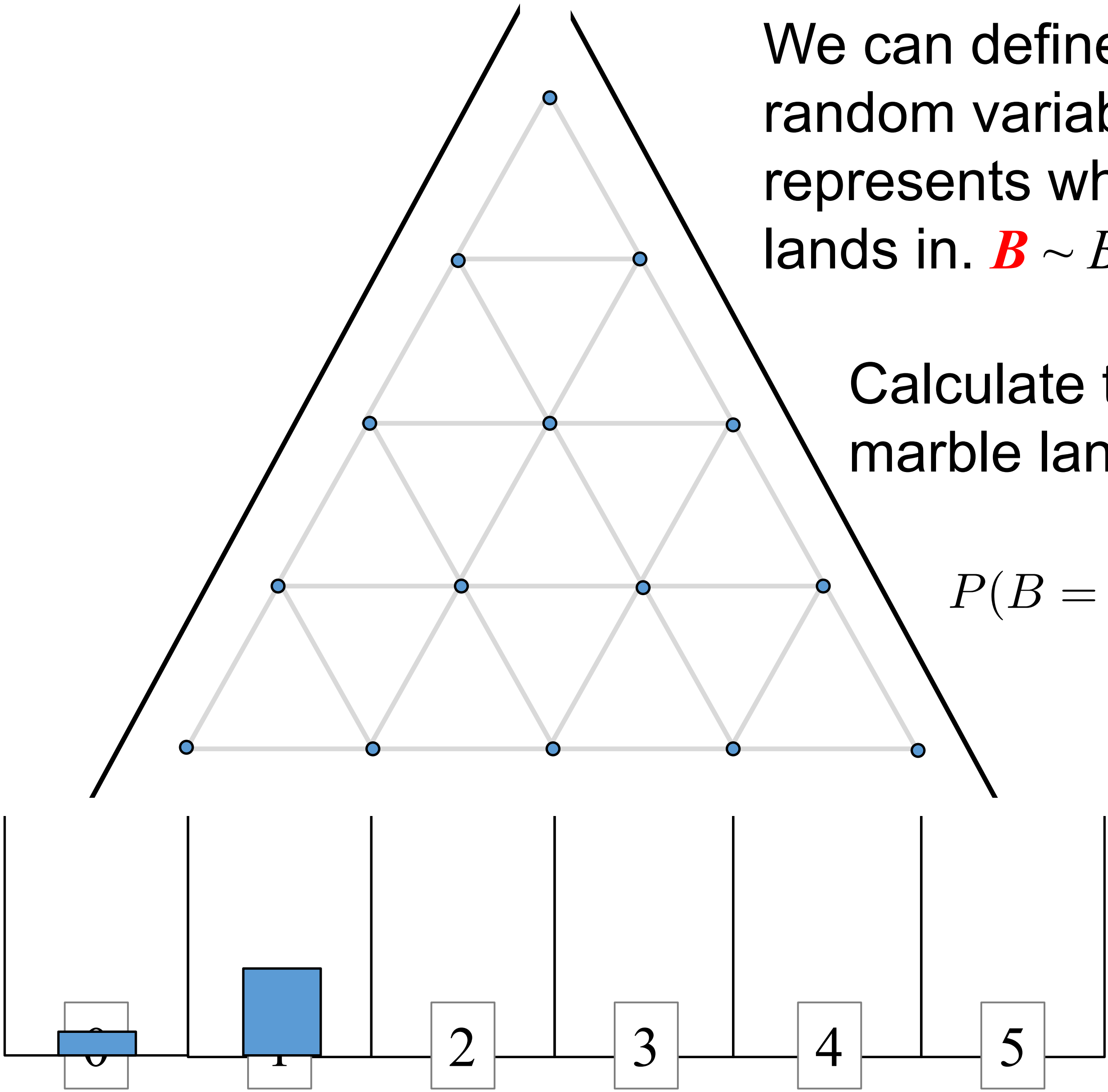
$$P(B = 0) = \binom{5}{0} \frac{1}{2}^5 \approx 0.03$$

Galton Board

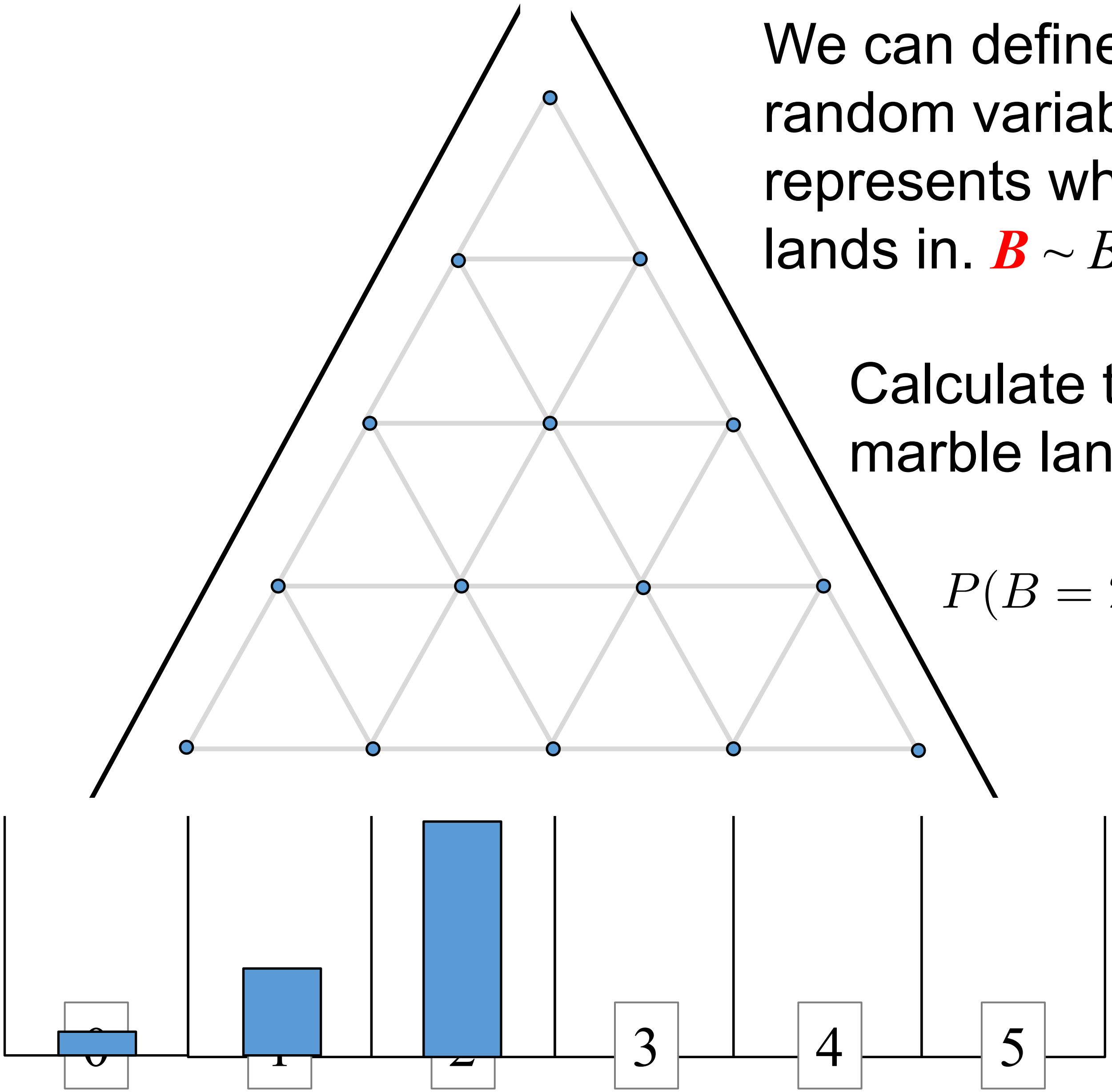
We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** $\sim Bin(5, 0.5)$

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

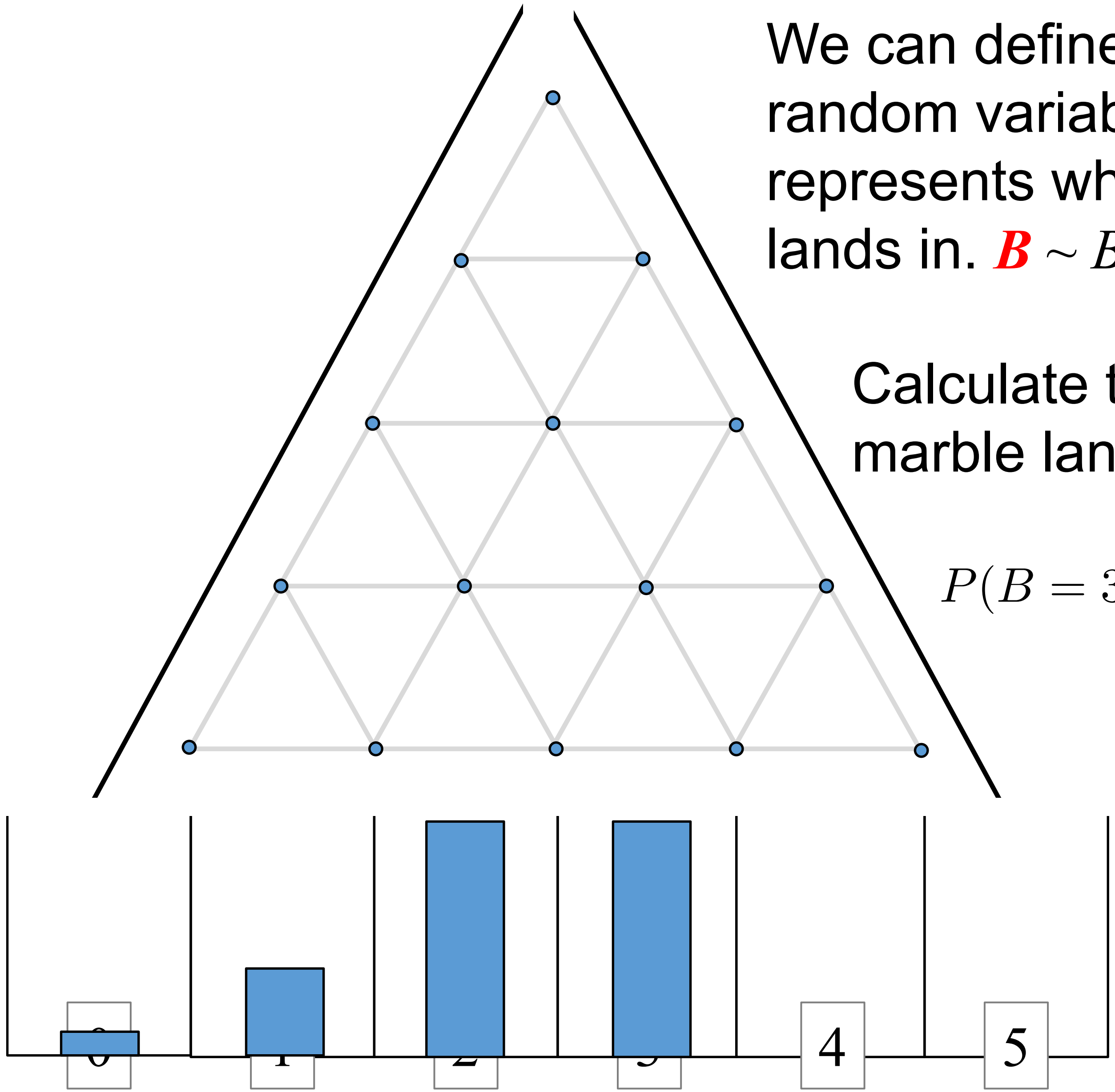


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** $\sim Bin(5, 0.5)$

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

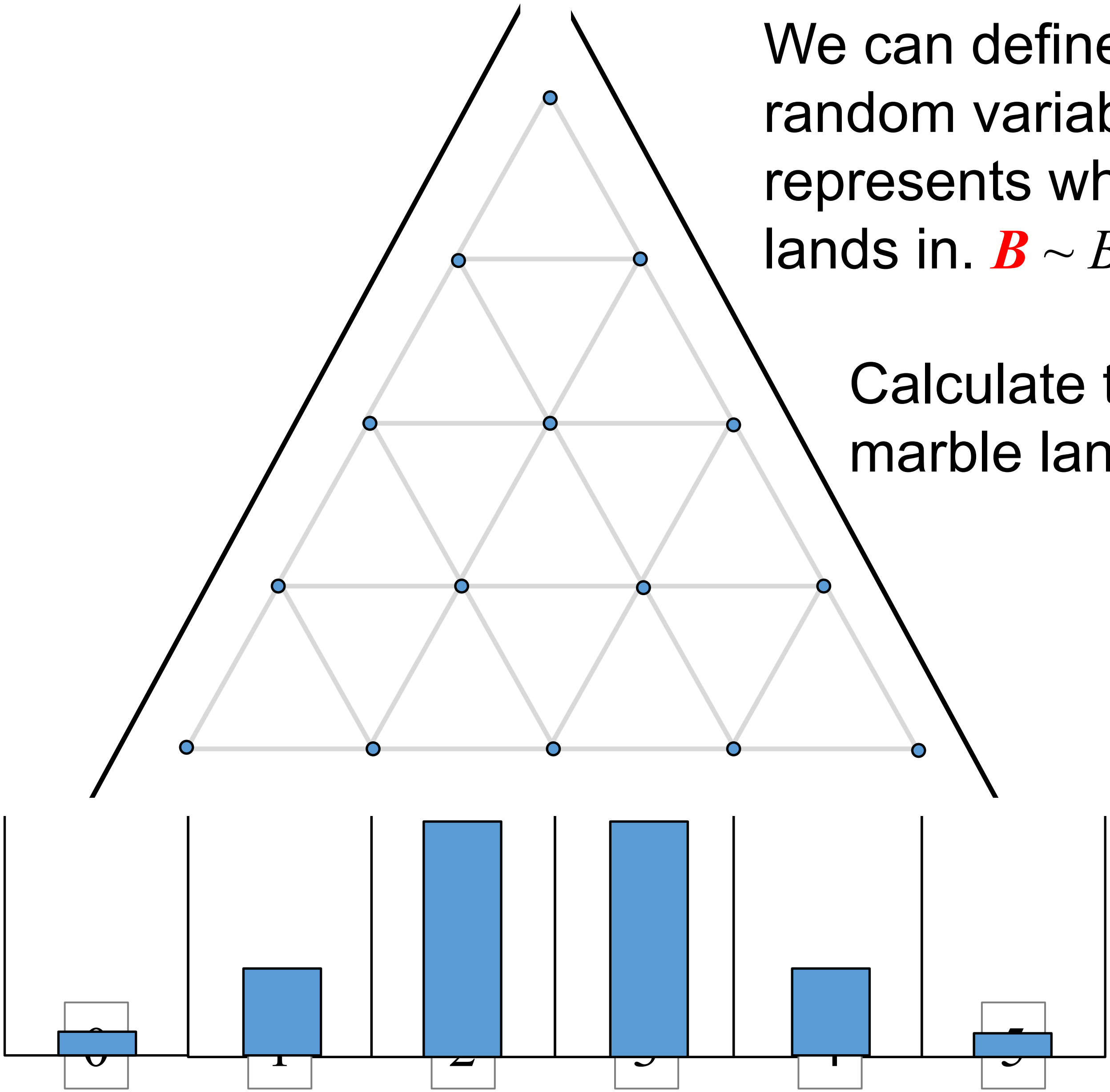


We can define an indicator random variable (***B***) which represents what bucket a marble lands in. ***B*** $\sim Bin(5,\ 0.5)$

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



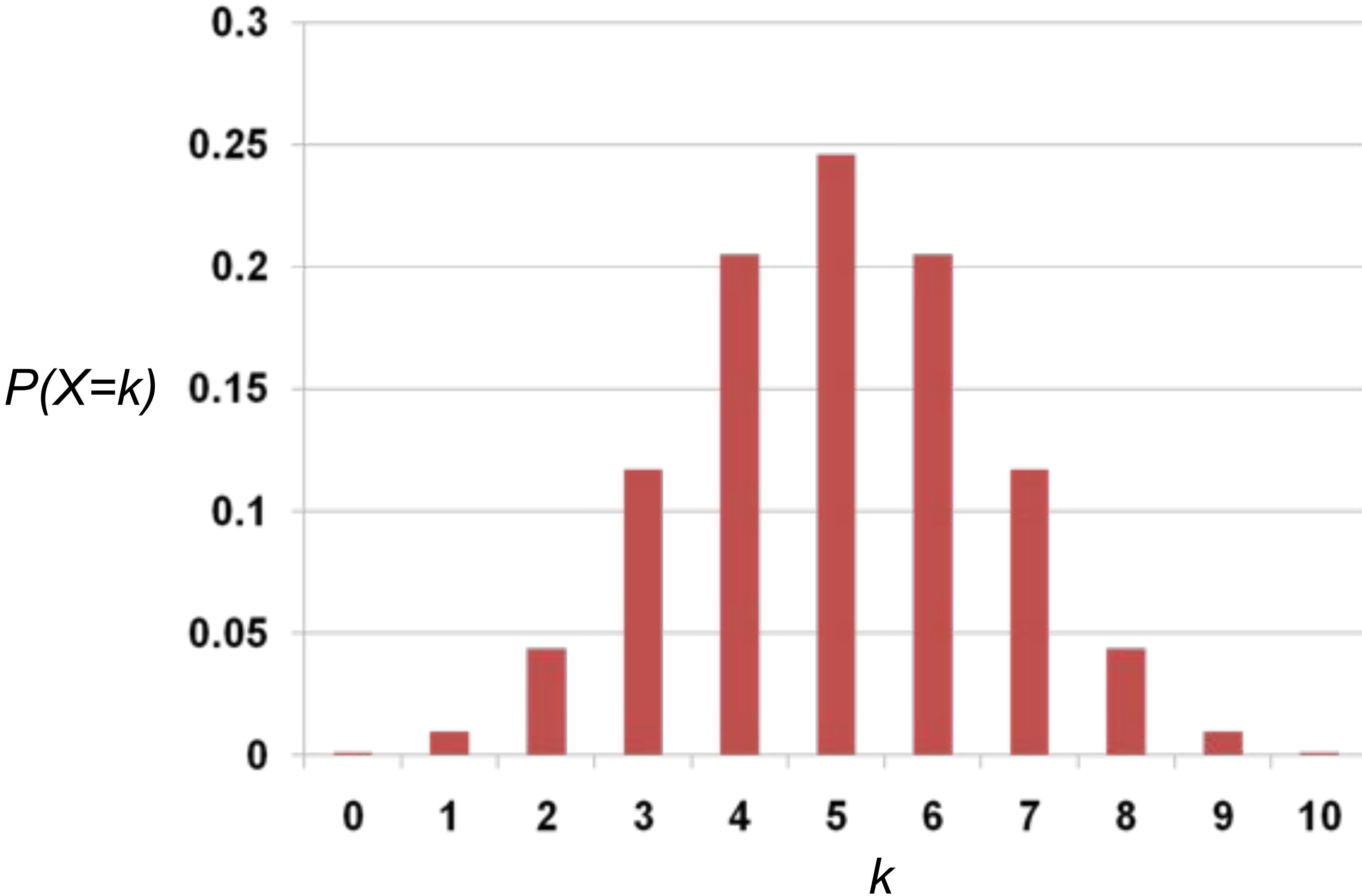
We can define an indicator random variable (***B***) which represents what bucket a marble lands in. $B \sim Bin(5, 0.5)$

Calculate the probability of a marble landing in a bucket.

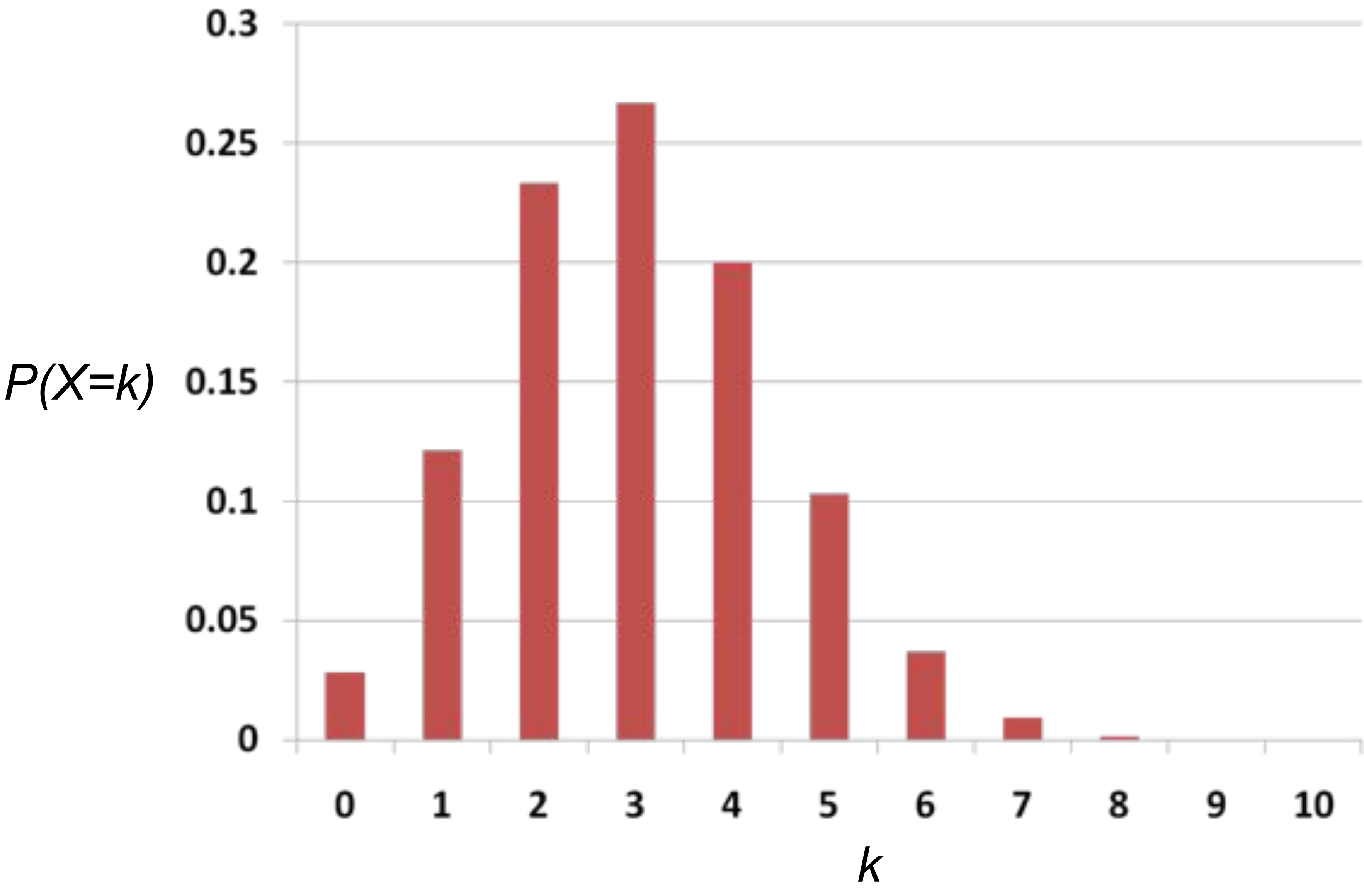
PDF

FROM CHAOS TO ORDER

PMF for $X \sim \text{Bin}(n = 10, p = 0.5)$



PMF for $X \sim \mathrm{Bin}(n = 10, p = 0.3)$



Genetic Inheritance

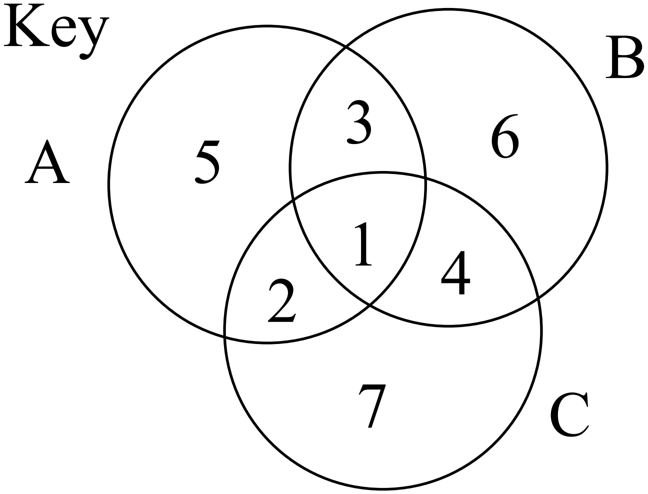
- Person has 2 genes for trait (eye color)
  - Child receives 1 gene (equally likely) from each parent
  - Child has brown eyes if either (or both) genes brown
  - Child only has blue eyes if both genes blue
  - Brown is "dominant" (d) , Blue is "recessive" (r)
  - Parents each have 1 brown and 1 blue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

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


1001

Have original 4 bit string to send over network.
Add 3 “parity” bits and send 7 bits total
Each bit independently corrupted (flipped) in transmission with probability 0.1. What is the probability of successful transmition?

Key



Send 1110?

Receive 1110000?

Receive 1010100?

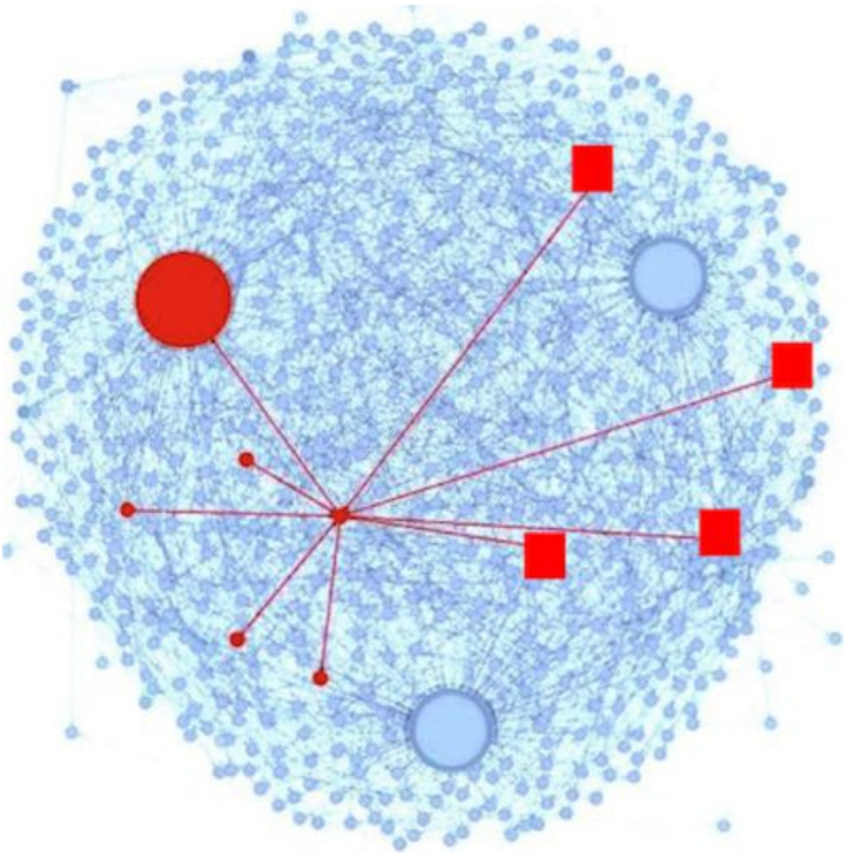
Have original 4 bit string to send over network.
Add 3 “parity” bits and send 7 bits total
Each bit independently corrupted (flipped) in transmission with probability 0.1. What is the probability of successful transmition?

Three Graders

Three peer graders (A, B, C) grade the same submission for a problem with 100 points. Each grader gives a grade which is a Binomial with $n = 100$, $p = 0.8$. What is the Expected average of their three grades?

Is Peer Grading Accurate Enough?

Looking ahead



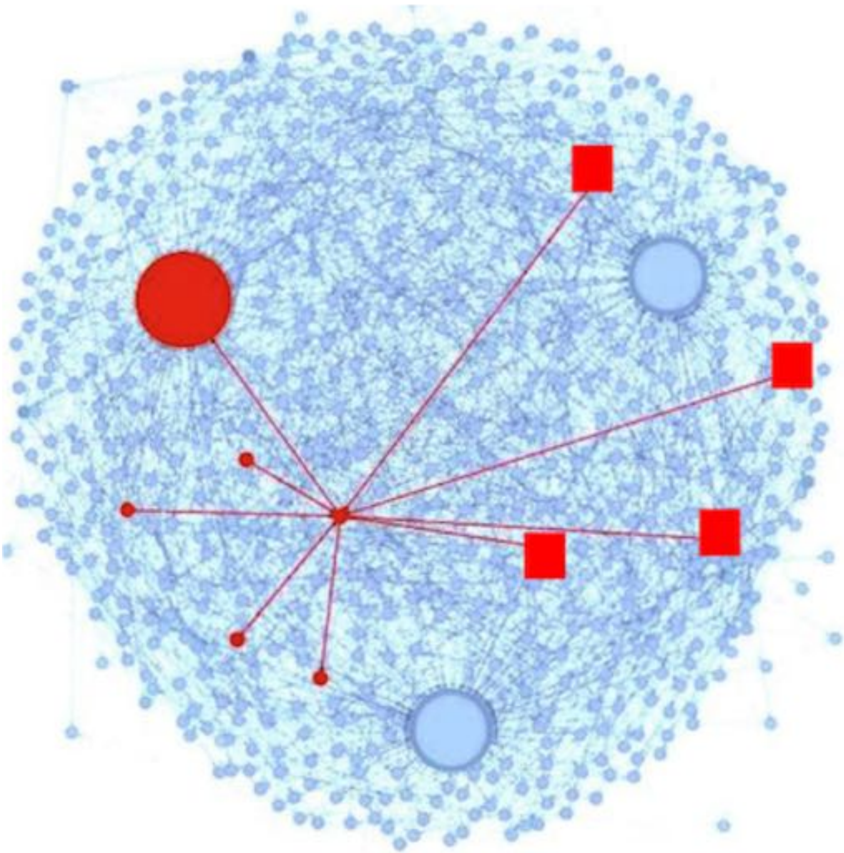
Peer Grading on Coursera HCI.

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

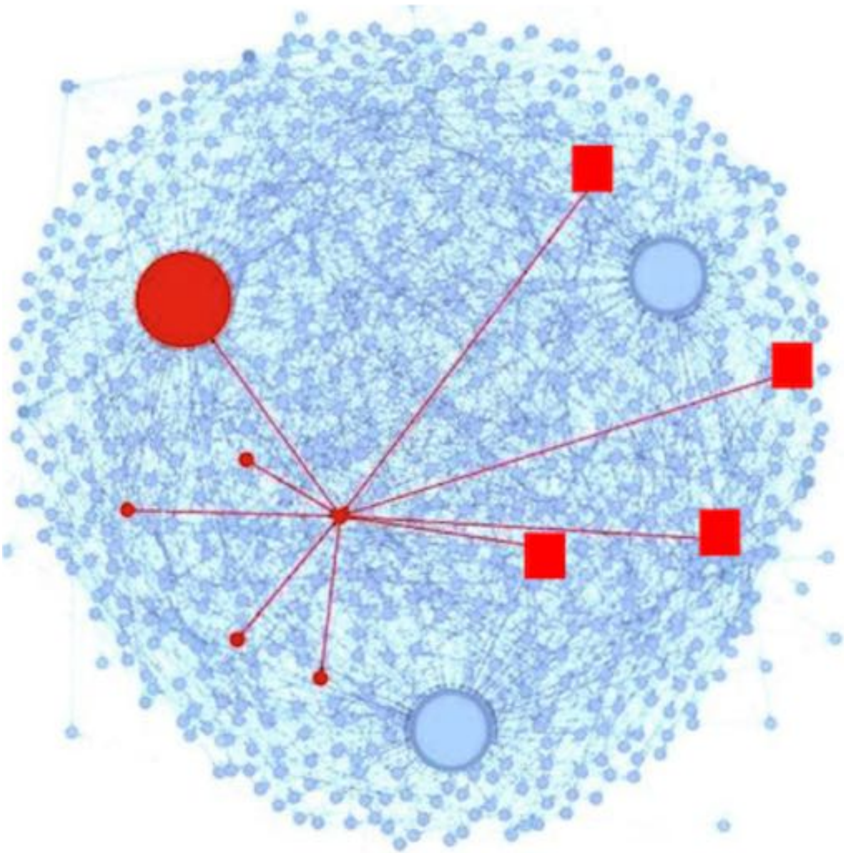
$$s_i \sim \mathrm{Bin}(\mathrm{points}, \theta)$$

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

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

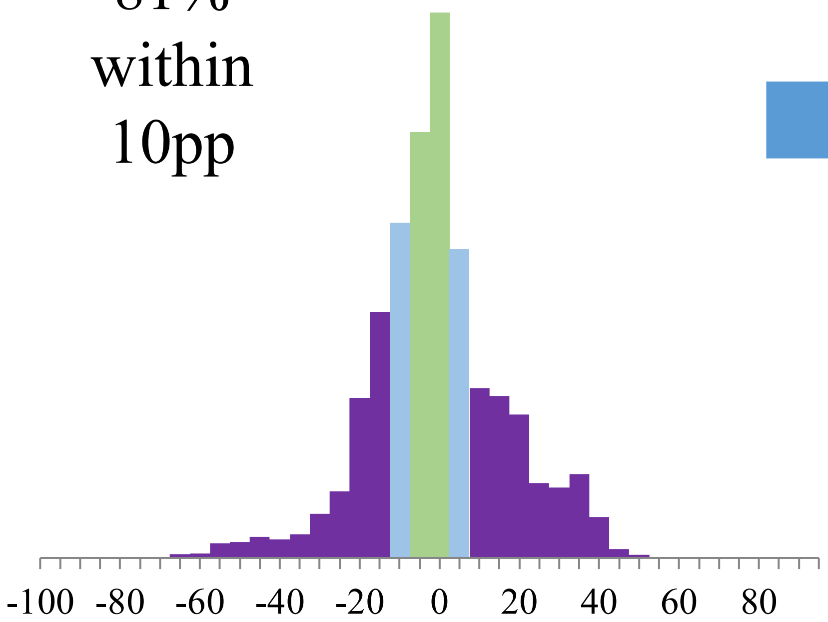
Inference or Machine Learning

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

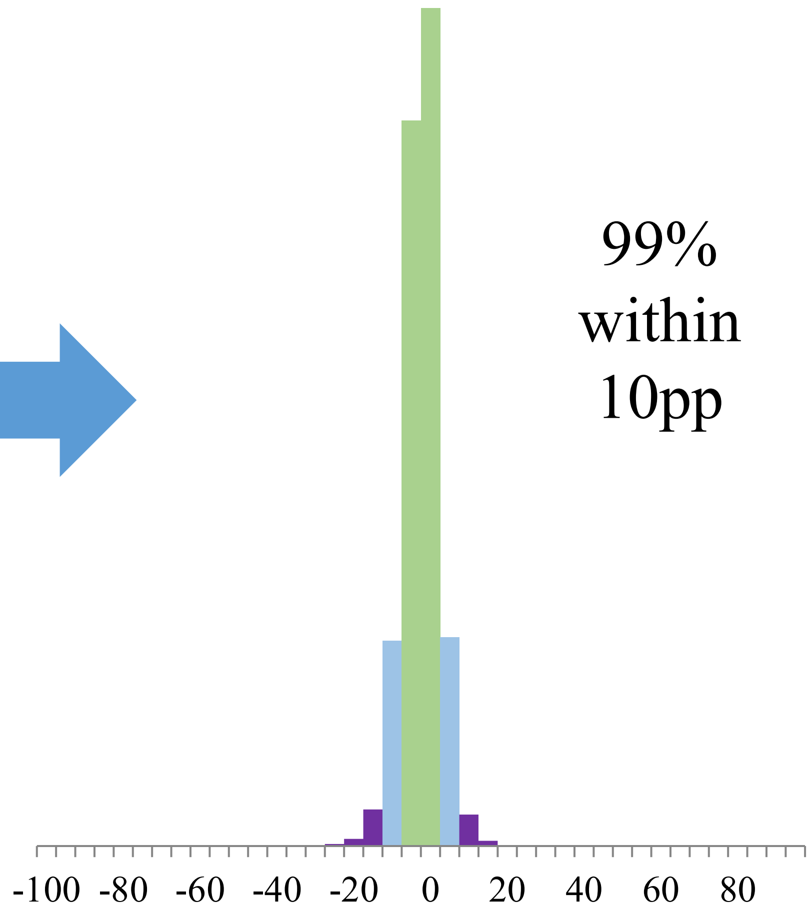
Yes, With Probabilistic Modelling

Before:

81% within 10pp



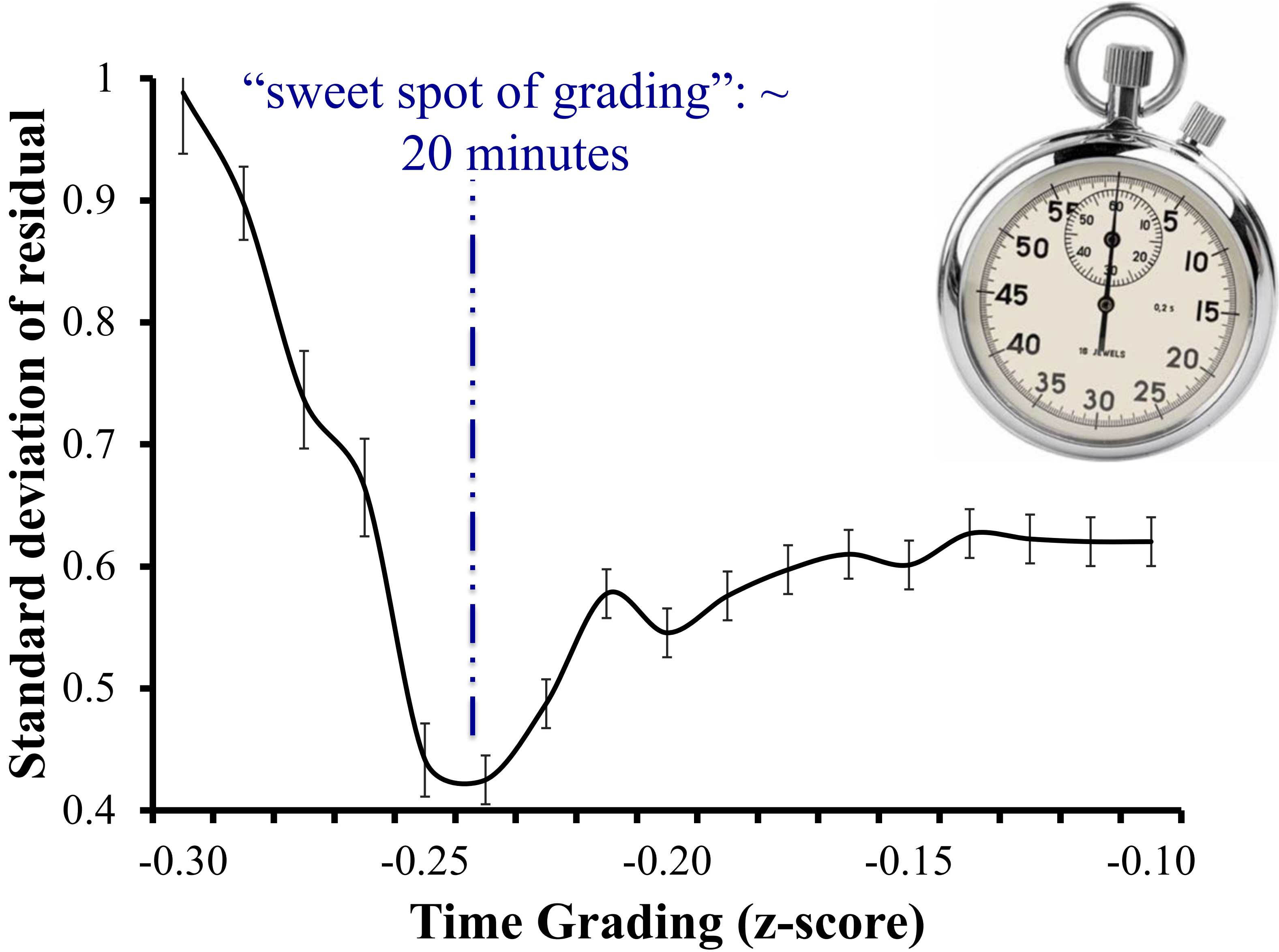
After:



99% within 10pp

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

Grading Sweet Spot



Voilà, c'est tout