• 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
  Probability that each is success

  Don’t care about the rest
  
  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
A random variable takes on values probabilistically.

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

$$P(X = 2) = \frac{1}{36}$$
The **probability mass function** (PMF) of a random variable is a function from values of the variable to probabilities.

\[ p_x(x) = P(X = x) \]
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

Random Variable

Semantics Meaning: $P(X=x)$

Expected Value: $E[X]$
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.
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”
Let $X$ be a random variable that represents a peer grade for an assignment that has a true grade of 58.
Let $X$ be a random variable that represents a peer grade

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

True grade = 58

$\mathbb{E}[X] = 57.5$
Let $X$ be a random variable that represents a peer grade.

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

True grade = 58

$E[X] = 57.5$

$X$ 

$25$ points

$(X - \mu)^2$

$1056$ points$^2$
Let $X$ be a random variable that represents a peer grade

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

True grade = 58

$E[X] = 57.5$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X - \mu)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 points</td>
<td>1056 points$^2$</td>
</tr>
<tr>
<td>80 points</td>
<td>506 points$^2$</td>
</tr>
</tbody>
</table>
Let $X$ be a random variable that represents a peer grade. The variance of $X$ is given by

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

where $\mu$ is the expected value of $X$. For this example, the true grade is 58, and the expected value $E[X]$ is 57.5. Here are some examples of grades and their squared differences from the expected value:

- 25 points: $(25 - 57.5)^2 = 1056$ points$^2$
- 80 points: $(80 - 57.5)^2 = 506$ points$^2$
- 50 points: $(50 - 57.5)^2 = 56$ points$^2$
Let $X$ be a random variable that represents a peer grade

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

True grade = 58

$E[X] = 57.5$

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X - \mu)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 points</td>
<td>1056 points$^2$</td>
</tr>
<tr>
<td>80 points</td>
<td>506 points$^2$</td>
</tr>
<tr>
<td>50 points</td>
<td>56 points$^2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

$$E[(X - \mu)^2] = 52 \text{ points}^2$$
Let $X$ be a random variable that represents a peer grade.

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

<table>
<thead>
<tr>
<th>$X$</th>
<th>$(X - \mu)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 points</td>
<td>1056 points$^2$</td>
</tr>
<tr>
<td>80 points</td>
<td>506 points$^2$</td>
</tr>
<tr>
<td>50 points</td>
<td>56 points$^2$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

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

$$\text{Std}(X) = 7.2 \text{ points}$$
• 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
Recall: Unconscious statistician:

\[ E[g(X)] = \sum_{x} g(x)p(x) \]

let \( g(X) = (X - \mu)^2 \)
Computing Variance

Var(\(X\)) = \(E[(X - \mu)^2]\)

\[
= \sum_{x} (x - \mu)^2 p(x)
\]

\[
= \sum_{x} (x^2 - 2\mu x + \mu^2) p(x)
\]

\[
= \sum_{x} x^2 p(x) - 2\mu \sum_{x} x p(x) + \mu^2 \sum_{x} p(x)
\]

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

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

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

\[
= E[X^2] - (E[X])^2
\]

Note: \(\mu = E[X]\)

Ladies and gentlemen, please welcome the 2\(^{nd}\) moment!
Variance of a 6 sided dice

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

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

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

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

- **Var(aX + b) = a^2 Var(X)**
  - **Proof:**
    \[
    \text{Var}(aX + b) = E[(aX + b)^2] - (E[aX + b])^2 \\
    = E[a^2X^2 + 2abX + b^2] - (aE[X] + b)^2 \\
    = a^2E[X^2] + 2abE[X] + b^2 - (a^2E[X]^2 + 2abE[X] + b^2) \\
    = a^2E[X^2] - a^2(E[X])^2 = a^2(E[X^2] - (E[X])^2) \\
    = a^2 \text{Var}(X)
    \]

- **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

Random Variable

Semantic Meaning

$P(X=x)$

$E[X]$  

$Var(X)$  

$E[X^2]$  

$Std(X)$
Lots of fun with Random Variables
Classics
• 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)
  - $P(X = 1) = p$    $P(X = 0) = 1 - p$
  - $X$ 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!
Run a program, crashes with probability $p = 0.1$, 
works with probability $(1 - p)$

$X$: 1 if program crashes

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

$X \sim \text{Ber}(p = 0.1)$
Serve an ad, clicked with probability $p = 0.01$, ignored with prob. $(1 - p)$

$\mathbf{C}$: 1 if ad is clicked

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

$\mathbf{C} \sim \text{Ber}(p = 0.01)$
More!
Binomial Random Variable

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

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

- 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
  - $X \sim \text{Bin}(n = 3, p = 0.5)$

\[
\begin{align*}
P(X = 0) &= \binom{3}{0} p^0 (1 - p)^3 = \frac{1}{8} \\
P(X = 1) &= \binom{3}{1} p^1 (1 - p)^2 = \frac{3}{8} \\
P(X = 2) &= \binom{3}{2} p^2 (1 - p)^1 = \frac{3}{8} \\
P(X = 3) &= \binom{3}{3} p^3 (1 - p)^0 = \frac{1}{8}
\end{align*}
\]
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** indicates here. For the binomial model in option pricing see Binomial options pricing model.

In probability theory and statistics, the **binomial distribution** with parameters \( n \) and \( p \) is the discrete probability distribution of the number of successes in \( n \) independent experiments, each asking a yes–no question, and each with its own boolean-valued outcome: a random variable \( X \) denoting the number of successes (with probability \( p \)) or failures (with probability \( q = 1 – p \)). A single success/failure experiment is also called a Bernoulli experiment and a sequence of outcomes is called a Bernoulli process; for a single trial, \( n = 1 \), the 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. If the draws are not independent and so the resulting distribution is a hypergeometric distribution much larger than \( n \), the binomial distribution remains a good approximation, and is widely used.

---

### Contents

1. Specification
   1.1 Probability mass function
   1.2 Cumulative distribution function
2. Example
   Mean
   Variance
   Mode
   Median
3. Covariance between two binomials
4. Related distributions
   8.1 Sums of binomials
   8.2 Ratio of two binomial distributions
   8.3 Conditional binomials
   8.4 Bernoulli distribution
   8.5 Poisson binomial distribution
   8.6 Normal approximation
   8.7 Poisson approximation
   8.8 Limiting distributions
   8.9 Beta distribution
5. Confidence intervals
   9.1 Wald method
   9.2 Agresti–Caffo method

---

**Notation**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>( B(n, p) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n \in \mathbb{N}_0 )</td>
<td>number of trials</td>
</tr>
<tr>
<td>( p \in [0,1] )</td>
<td>success probability in each trial</td>
</tr>
</tbody>
</table>

| Support | \( k \in \{0, \ldots, n\} \) | number of successes |
| pmf | \( \binom{n}{k} p^k (1–p)^{n–k} \) |
| CDF | \( I_{\frac{1–p}{p}}(n–k, 1+k) \) |
| Mean | \( np \) |
| Median | \( np \) or \( np \) |
| Mode | \( \frac{(n+1)p}{1+(np)} \) or \( \frac{(n+1)p}{1+(np)} - 1 \) |
| Variance | \( np(1–p) \) |
| Skewness | \( \frac{1–2p}{np(1–p)^{3/2}} \) |
| Ex. kurtosis | \( \frac{1–6p(1–p)}{np(1–p)^2} \) |
| Ex. kurtosis | \( 1 + \frac{1–6p(1–p)}{np(1–p)^2} + O(\frac{1}{n}) \) |
| Entropy | \( \frac{1}{2} \log_2 \left( \frac{2\pi e np(1–p)}{np(1–p)} \right) + O\left(\frac{1}{n}\right) \) |
\[ E(X^2) = \sum_{k \geq 0} k^2 \binom{n}{k} p^k q^{n-k} \]
\[ = \sum_{k=0}^{n} kn \binom{n-1}{k-1} p^k q^{n-k} \]
\[ = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} \]
\[ = np \sum_{j=0}^{m} (j+1) \binom{m}{j} p^j q^{m-j} \]
\[ = np \left( \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) \]
\[ = np \left( \sum_{j=0}^{m} m \binom{m-1}{j-1} p^j q^{m-j} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) \]
\[ = np \left( (n-1)p \sum_{j=1}^{m} \binom{m-1}{j-1} p^{j-1} q^{(m-1)-(j-1)} + \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} \right) \]
\[ = np \left( (n-1)p (p+q)^{m-1} + (p+q)^m \right) \]
\[ = np \left( (n-1)p + (p+q)^m \right) \]
\[ = n^2 p^2 + np (1-p) \]
$n$ 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**

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

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

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

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

\[
P(H = k) = \binom{n}{k} (p)^k (1 - p)^{n-k}
\]

\[
P(H < 3) = \sum_{i=0}^{2} \binom{100}{i} (0.1)^i (0.9)^{100-i}
\]
1000 ads served, each clicked with $p = 0.01$, otherwise ignored. Expectation and Standard deviation of number of ads clicked?

$H$: number of clicks

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

\[
\Pr(H = k) = \binom{1000}{k} (0.01)^k (0.99)^{1000-k}
\]

\[
\text{E}(H) = np = 10
\]

\[
\text{Var}(H) = np(1-p) = 9.9
\]

\[
\text{Std}(H) = 3.15
\]
When a marble hits a pin, it has equal chance of going left or right.
When a marble hits a pin, it has equal chance of going left or right. Each pin represents an independent event.
The bucket index that a marble lands in is equal to the number of times the marble went right.
We can define an indicator random variable ($R$) which represents whether a particular marble goes right as a Bernoulli distribution:

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

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

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

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

\[
P(B = 3) = \binom{5}{2} \left( \frac{1}{2} \right)^5 \approx 0.31
\]
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.
FROM CHAOS TO ORDER
PMF for $X \sim \text{Bin}(n = 10, p = 0.5)$
PMF for $X \sim \text{Bin}(n = 10, p = 0.3)$
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 = (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)

\[
P(X = 3) = \binom{4}{3} (0.75)^3 (0.25)^1 \approx 0.4219
\]
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 transmission?

Key
A
B
C

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 transmission?
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

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

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

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 \( z_{ij} \) 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

“sweet spot of grading”: ~ 20 minutes
Voilà, c'est tout