Random Variables
A baseball game is a random experiment

Sample statistics gathered:
  ● Number of runs, hits, and errors, per team
  ● Number of hits, walks, and outs, per player
  ● Number of strikeouts, walks, and pitches, per pitcher
  ● Etc.

These statistics are all examples of random variables.
Likewise, many of our data sets contain random variables:
  ● Iris: Petal.Length, Petal.Width, etc.
  ● Presidents: Approval Ratings
A random variable is a numerical outcome of a random experiment.

If we flip a coin 10 times, there can be multiple random variables:
- $H$ can be the random variable indicating the number of heads
- $T = 10 - H$ can be the random variable indicating the number of tails
- $X$ can represent the longest sequence of consecutive heads
- $Y$ can represent the longest sequence of consecutive tails

The random variables $H, T, X,$ and $Y$ all take on values between 0 and 10.
- $R = \{0, 1, 2, \ldots, 10\}$ is called the range of the random variable.
- The domain of the random variable is the sample space of the experiment.
- $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$
- Technically, a random variable is a function from the sample space to the reals.
Let’s look at a slightly simpler experiment, and calculate some probabilities.

We will flip a coin 2 times.

- Random variable $H$ will record the total number of heads.
- What is the range of $H$: i.e., what values can $H$ take on? $R = \{0, 1, 2\}$.

If we flip a coin twice, then the sample space is $\{HH, HT, TH, TT\}$.

- $Pr[H = 0] = |\{TT\}| / |\{HH, HT, TH, TT\}| = \frac{1}{4}$
- $Pr[H = 1] = |\{HT, TH\}| / |\{HH, HT, TH, TT\}| = \frac{1}{2}$
- $Pr[H = 2] = |\{HH\}| / |\{HH, HT, TH, TT\}| = \frac{1}{4}$

We have just found the probability distribution of the random variable $H$.

All probabilities are between 0 and 1, and together they sum to 1.
The Expected Value Operation

- Once we have specified the range of a random variable, and its probability distribution, we can calculate the expected value of that random variable.
- Assume $X$ is a random variable, $x_i$ is an element of its range, and $p_i$ is the probability of $x_i$. Then the expected value of $X$ is calculated as follows:

$$E(x) = \sum_{i=1}^{n} x_i p_i = x_1 p_1 + x_2 p_2 + x_3 p_3 + \cdots + x_n p_n$$

- $E[H] = (0)Pr[H = 0] + (1)Pr[H = 1] + (2)Pr[H = 2] = 0(\frac{1}{4}) + 1(\frac{1}{2}) + 2(\frac{1}{4}) = 1$
  - If we flip a coin two times, we should expect to see one head.
- The expected value of a random variable is the same as its mean ($\mu$).
If $X$ is a random variable representing the result of one roll of a six-sided die, then what is the expected value of $X$?

A) 3
B) 3.5
C) 4
If $X$ is a random variable for the number rolled on a six-sided die, then what is the expected value of $X$?

A) 3
B) 3.5
C) 4

$\frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) = 3.5$

This is the expected value, even though we cannot roll a 3.5 ever!
Law of Large Numbers

- You can roll a die many times.
- Each time, you will see a 1, 2, 3, 4, 5, or 6.
- If you average the value you see across trials, as you run more and more experiments, this average will approach 3.5!

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Average</td>
<td>1</td>
<td>2.5</td>
<td>8/3</td>
<td>11/4</td>
<td>16/5</td>
<td>22/6</td>
<td>24/7</td>
<td>27/8</td>
<td>31/9</td>
<td>32/10</td>
<td>38/11</td>
<td>44/12</td>
</tr>
</tbody>
</table>

- In this experiment, the sample mean is 3.667.
Law of Large Numbers

- You can toss 2 coins many times.
- Each time, you will see 0, 1, or 2 heads.
- If you average the number of heads you see across trials, as you run more and more experiments, this average will approach 1!

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Average</td>
<td>0</td>
<td>1</td>
<td>4/3</td>
<td>1</td>
<td>1</td>
<td>5/6</td>
<td>5/7</td>
<td>6/8</td>
<td>10/9</td>
<td>12/10</td>
<td>12/11</td>
<td>13/12</td>
</tr>
</tbody>
</table>

- In this experiment, the sample mean is 1.083.
Random Variables Vary

- What makes random variables interesting is that they vary!
- You can and will get different results running the same random experiment over and over again, as a result of the randomness.
- Expected value explains the center of a probability distribution.
- However, it does not tell us anything about the spread of the distribution.
Recall Sample Variance

Assume $x$ is a random variable, $x_i$ is an element of its range, and $\mu$ is its mean. Then the sample variance of $x$, given a sample of size $N$, is calculated as follows:

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Squared Differences</td>
<td>6.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>2.25</td>
<td>6.25</td>
<td>2.25</td>
<td>0.25</td>
<td>0.25</td>
<td>6.25</td>
<td>6.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

- In this experiment, the sample variance is 2.58, and the sample SD is 1.61.
Recall Sample Variance

Assume \( x \) is a random variable, \( x_i \) is an element of its range, and \( \mu \) is its mean. Then the sample variance of \( x \), given a sample of size \( N \), is calculated as follows:

\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Squared Differences</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

- In this experiment, the sample variance is .75, and the sample SD is .86.
Variance through the Prism of Expectation

- Actually, variance is not defined as the average squared deviation, but rather as the expected squared deviation.
- In the case of the die — actually, any time the probabilities are uniform — these are the same thing:
  - \( \left( \frac{1}{6} \right)(1 - 3.5)^2 + \left( \frac{1}{6} \right)(2 - 3.5)^2 + \left( \frac{1}{6} \right)(3 - 3.5)^2 + \left( \frac{1}{6} \right)(4 - 3.5)^2 + \left( \frac{1}{6} \right)(5 - 3.5)^2 + \left( \frac{1}{6} \right)(6 - 3.5)^2 \approx 2.9 \)
  - The sum of the squared deviations is \( (1 - 3.5)^2 + (2 - 3.5)^2 + (3 - 3.5)^2 + (4 - 3.5)^2 + (5 - 3.5)^2 + (6 - 3.5)^2 = 17.5 \). To find the average squared deviation, we divide by 17.5/6 \( \approx 2.9 \)
  - \( \left( \frac{1}{4} \right)(0 - 1)^2 + \left( \frac{1}{2} \right)(1 - 1)^2 + \left( \frac{1}{4} \right)(2 - 1)^2 = \left( \frac{1}{4} \right)(1)^2 + \left( \frac{1}{4} \right)(1)^2 = \frac{1}{2} \)
- But probabilities are not always uniform. If our random variable is the sum of two dice, 7 is a more likely outcome than 12.
- In general, to compute the variance, we need to compute an expectation.
Variance through the Prism of Expectation

- Once again, the variance is the expected value of the squared deviations between values and the mean.
  - Variance[X] = E[(X - E[X])^2]
  - We can simplify this through some algebra.
  - Variance[X] = E[X^2] - E[X]^2

- This second formula will yield the same result as the first, and it is usually easier to work with! The variance of the sum of two dice is:
  - E[X^2] = (⅙₃₆)(2)^2 + (⅓₈)(3)^2 + (⅓₁₂)(4)^2 + (⅙₉)(5)^2 + (1/₃₆)(6)^2 + (1/₉)(7)^2 + (1/₃₆)(8)^2 + (1/₉)(9)^2 + (1/₁₂)(10)^2 + (⅓₈)(11)^2 + (⅓₆)(12)^2 = 54.83
  - E[X]^2 = (7)^2 = 49
  - E[X^2] - E[X]^2 = 54.83 - 49 ~ 5.83