How much variability do we expect?

Box model.

\[ \text{sum of draws} = \text{total in N tosses}. \]

Sum of N draws \( \equiv \) Tally in N tosses.

*Expected value* = \( N \times \text{(mean of box)} \)

*Standard error* = \( \sqrt{N} \times \text{(SD of box)} \)

68% of outcomes will be expected \( \pm 2 \text{SE} \)

95% \( EV \pm 2 \text{SE} \)

Sum of draws from a box model follows the Normal Distribution

\[ \text{mean} \equiv \text{expected value} \]

\[ \text{SD} \equiv \text{standard error} \]

When is it ok to use the Normal Approximation?
Probability Histograms.

<table>
<thead>
<tr>
<th>outcome</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2/3</td>
</tr>
<tr>
<td>3</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>1/6</td>
</tr>
</tbody>
</table>

Area of bar represents probability.

As you repeat the experiment, the empirical histogram converges to the probability histogram.
Rolling 2 dice:

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

<table>
<thead>
<tr>
<th>outcome</th>
<th>prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>$\frac{1}{36}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{2}{36}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{3}{36}$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{4}{36}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{5}{36}$</td>
</tr>
<tr>
<td>7</td>
<td>$\frac{6}{36}$</td>
</tr>
<tr>
<td>8</td>
<td>$\frac{5}{36}$</td>
</tr>
<tr>
<td>9</td>
<td>$\frac{4}{36}$</td>
</tr>
<tr>
<td>10</td>
<td>$\frac{3}{36}$</td>
</tr>
<tr>
<td>11</td>
<td>$\frac{2}{36}$</td>
</tr>
<tr>
<td>12</td>
<td>$\frac{1}{36}$</td>
</tr>
</tbody>
</table>

As you repeat the experiment (for increasing $N$), how does the experimental histogram converge to the probability histogram?
Product of #s on 2 dice.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>8</td>
<td>10</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>16</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>10</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>12</td>
<td>18</td>
<td>24</td>
<td>30</td>
<td>36</td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
13 & 14 & 15 & 16 & 17 & 18 & 19 & 20 & 21 & 22 & 23 & 24 \\
25 & 26 & 27 & 28 & 29 & 30 & 31 & 32 & 33 & 34 & 35 & 36 \\
\end{array}
\]

much less regular than sum of rolls.
What's special about the sum of draws?

\[ \text{Sum of } N \text{ draws} = \#H \text{ in } N \text{ tosses of a coin}. \]

What does the histogram of the \#H (sum of draws) converge to?

What does the probability histogram for the sum of draws look like, as \( N \) increases?

- It starts to look like the Normal curve.

Is this a general phenomenon, or does it just happen in this simple case?

\[ \begin{array}{c|c|c|c} 0 & 1 & 1 & \text{tail head} \\ \end{array} \]

\[ \begin{array}{c|c} 1 & 2 & 9 \\ \end{array} \]

prob. histogram for single draw
- Looks nothing like the Normal curve.
even though the probability histogram
for 1 draw looks nothing like the
normal curve, for 100 draws the
deviations from the normal curve have
became quite small.

This is always the case
the sum of draws always follows the
Normal curve when the number of draws
gets large.

how large depends on the contents
of the box.
Central Limit Theorem.

In general, the probability histogram of the sum of draws can be approximated by the Normal Curve.

When drawing at random with replacement from a box model, the probability histogram for the sum will follow the Normal curve in the limit (large number of draws), even if the contents of the box do not follow the Normal curve.

(Plot the histogram into standard units in "large" number of draws.)
Example

400 draws at random with replacement from:

\[
\begin{bmatrix}
1 & 3 & 5 & 7
\end{bmatrix}
\]

What's the chance that the sum $> 1500$

Mean of box $= \frac{1 + 3 + 5 + 7}{4} = \frac{16}{4} = 4$.

SD of box $= -3 -1 1 3$ deviations from mean.

\[
\begin{align*}
9 & \quad 1 & \quad 1 & \quad 9 \\
\end{align*}
\]

Squared deviations from mean $= 9 + 1 + 1 + 9 = \frac{20}{4} = 5$.

Mean squared deviation $= 5$.

SD of box $= \sqrt{\text{mean squared deviation from mean}} = \sqrt{5} \approx 2.24$.

400 draws.

Expected value of sum of draws $= 400 \times 4 = 1600$.

$SE$ $\approx 45$.
Shaded area is prob. that sum of 400 draws > 1500

Convert 1500 to standard units.

\[-Z = \frac{1500 - 1600}{45} = -\frac{100}{45} = -2.22\]

For \( Z = 2.2 \)

\( \text{shaded area} \approx 0.0275 \)

Two tails contain \( 100 - 0.9725 = 2.75\% \).

\( \Rightarrow \) 1 tail contains, 1.375\%.

\( \Rightarrow \) prob. that sum of 400 draws > 1500

\[ = 1 - 1.375 = 0.98675 \]

\( \approx 98.7\% \).