How much variation do we expect?

Toss a coin repeatedly

\[ P(H) = \frac{1}{2} \]
\[ P(T) = \frac{1}{2} \]

consider \[ \# \text{ heads} = \frac{1}{2} \# \text{ tosses} \].

Expect this to be zero but there will be variation.

- absolute size of the variation will get bigger as \# tosses increase.

But relative variation - the size of the variation as a fraction of the \# of tosses decreases.

Consider the size of the variation with 100 tosses and the size with 10,000 tosses (100 x more). The size of the variation with 10,000 tosses will be about \( \sqrt{100} \) times larger than with 100 tosses.
# tosses increased by factor of 100
size of variation increases by factor of $\sqrt{100}$.

$$\frac{\text{Size of variation}}{\text{# tosses}} \propto \frac{\sqrt{N}}{N} \Rightarrow \frac{1}{\sqrt{N}} \text{ - gets smaller as } N \text{ increases.}$$

---

How do we study problems like this?

- make an analogy between
  the process being studied and
drawing numbers at random
  from a box.

<table>
<thead>
<tr>
<th>H</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H</th>
<th>H</th>
<th>T</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<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>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$\Rightarrow$ sum of draws $\equiv$ # sixes in $N$ rolls of a die.
Connect the variability you want to know about with the chance variability in the sum of the numbers drawn from the box.

"box model"

- which numbers go on the tickets
- how many tickets of each kind
- how many draws

Example:

Roulette: 38 numbered slots.
1-36 half coloured red
   half black.
   0
   00

Simplest Bet: $ bet Red or Black.
If bet Red, and ball lands on red get back your stake plus the same amount.
eg bet $1 and win, $2 returned.
gain on each spin is \(-1\) (lose)
\(\quad +1\) (win).

**Box model.**

- Numbers on the tickets are \(-1\) plus 1.
- How many tickets of each type?
  - 18 ways to win \(\rightarrow\) 18 tickets with +1
  - 20 ways to lose \(\rightarrow\) 20 tickets with -1

\[
\begin{array}{|c|c|}
\hline
18 \text{ tickets} & 20 \text{ tickets} \\
\hline
+1 & -1 \\
\hline
\end{array}
\]

If each of us draws (with replacement)
- Outcomes will vary
  - Vary about \underline{Expected value}.
  - Size of fluctuations \underline{Standard Error}.

\(\rightarrow\) Can calculate these quantities from the box model.
Expected value for sum of draws made at random with replacement in

\((\# \text{ draws}) \times (\text{average of box})\).

Add up numbers on all the tickets in the box and divide by the number of tickets.

Play Roulette 38 times.

\[ = 38 \text{ draws from box.} \]

Expected value

\[ 38 \times \frac{18 - 20}{38} = -2. \]

10 rolls of fair die.

\[ \frac{\text{total pips}}{\text{total possible pips}} = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = \frac{21}{6} = 3.5. \]

Expected value for total pips on 10 rolls is \(3.5\)
Quantifying the variability.

\# pips on 10 rolls = expected value + chance variability.

\[ \text{S.E. for sum of draws with replacement from box model is:} \]
\[ \sqrt{\text{# draws}} \times (\text{SD of box}) \]

Consider all the numbers on the tickets as a list of numbers, and compute the SD of that list.

Toss a coin 100 times:

- count \# Heads

<table>
<thead>
<tr>
<th>T</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ \Rightarrow \text{sum of draws} = \# H. \]

Expected \#H in 50 draws

\[ 50 \times \frac{0+1}{2} = 50 \times \frac{1}{2} = 25 \]
$$\text{S.E.} = \sqrt{50} \times (\text{SD of box})$$

Mean of box is $\frac{1}{2}$.

SD of box:
$$\sqrt{\frac{(0 - \frac{1}{2})^2 + (1 - \frac{1}{2})^2}{2}} = \frac{1}{2}.$$ 

S.E. (variance in # H in 50 tosses)
$$= \sqrt{\# \text{ draws} \times (\text{SD of box})}$$
$$= \sqrt{50} \times \frac{1}{2}.$$
$$= 7.07 \times \frac{1}{2} \approx 3.5.$$ 

In 50 tosses, we expect 25 heads $\pm 3.5$ or so.

**Roulette:**

<table>
<thead>
<tr>
<th>18 tickets</th>
<th>20 tickets</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 + 1 ... + 1</td>
<td>11 + 1 ... + 1</td>
</tr>
</tbody>
</table>

Mean of box:
$$\frac{18 \times 1 + 20 \times (-1)}{38} = \frac{-2}{38} = -\frac{1}{19}.$$
$\text{SD of box.}$

\[ \sqrt{\left( +1 - \left( -\frac{1}{19} \right) \right)^2 + \left( +1 - \left( -\frac{1}{19} \right) \right)^2 + \ldots + \left( +1 - \left( -\frac{1}{19} \right) \right)^2}
+ \left( -1 - \left( -\frac{1}{19} \right) \right)^2 + \left( -1 - \left( -\frac{1}{19} \right) \right)^2 + \ldots + \left( -1 - \left( -\frac{1}{19} \right) \right)^2} \]

\[ = 38. \]

\[ 18 \times \left( +1 - \left( -\frac{1}{19} \right) \right)^2 + 20 \times \left( -1 - \left( -\frac{1}{19} \right) \right)^2. \]

\[ = \frac{38}{38}. \]

10 spins. expected value is $10 \times -\frac{1}{19} = -0.526.$

$\text{SE} = \sqrt{10 \times 0.9986} = 3.16.$

In this situation the fluctuations can be very large compared to the expected value.

Observed values are rarely more than $2 - 3 \text{SE from the expected value.}$
**Dice:**

\[ \begin{array}{ccccccc} 
1 & 2 & 3 & 4 & 5 & 6 \\
\end{array} \]

expected value = 2.5

\[
\text{SD of box} = \sqrt{\frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6}}
\]

\[ \approx 1.7 \]

10 rolls, expected value = \(10 \times 3.5\) = 35. 

SE = \(\sqrt{10 \times 1.7} = \frac{5.4}{\sqrt{10}}\)

35 ± 5.4.

Expect very few values outside range.

24 → 46.
For a lot of data sets, 68% of data is within 1 standard deviation of mean.

In a lot of sampling situations, 68% of the results will be within 1 standard error of the expected value.

We can use the **normal approximation** to the values we get for the sum of the draws from a box model will (often) follow the normal curve, where the normal curve has as parameters the expected value and standard error for the particular box model being considered.
Roll a die 10 times

- What's the chance that the total # of pips on the 10 rolls < 25?

  total # pips \rightarrow \text{sum of draws from the box}

  sum of draws \sim \text{normal curve with parameters}
  \text{expected value = standard error}.

\[ \text{expected value} = 35 \]
\[ \text{SE} = 5.4. \]

Convert 25 to standard units using expected value + standard error.

\[ \frac{25 - 35}{5.4} = 1.85 \]

Area is 93.57% (from table).
On 10 rolls of a fair die, chance of getting total # pips $< 25$ is $\approx 3\%$.

In the data we collected, there were 11 values less than 225 in 331 experiments.

$$\frac{11}{331} \approx 3.3\%.$$

Here, the normal approximation has given a good result.

But can we decide in general when it is reasonable to use the normal approximation?
Computing the SE for a box that contains only 2 types of tickets:

In this situation, can compute SD of box on

\[ \text{SE}_{\text{box}} = \left( \frac{\text{to big} - \text{small}}{\text{number} - \text{number}} \right) \times \sqrt{\frac{\text{fraction of tickets with big number}}{\text{big number}} \times \frac{\text{fraction of tickets with small number}}{\text{small number}}} \]

numbers on the tickets.

\[ \text{Roulette:} \]

\[ \begin{array}{c|c|c}
18 \text{ tickets} & 20 \text{ tickets} \\
\hline
+1 & -1 \\
\end{array} \]

\[ \text{SE}_{\text{box}} = (+1 - (-1)) \sqrt{\frac{18}{38} \times \frac{20}{38}} = 0.9986 \]
How many 6's do we expect in 100 rolls of a die?

- Sum of draws represents # of sixes rolled.

\[
\begin{align*}
\text{Mean of box} &= \frac{1 \times 1 + 5 \times 0}{6} = \frac{1}{6} \\
\text{SD of box} &= \left(1 - 0\right) \sqrt{\frac{1}{6} \times \frac{5}{6}} \approx 0.3727
\end{align*}
\]

In 100 rolls, expected value is \(100 \times \frac{1}{6} \approx 17\).

\(\text{SE} = \sqrt{100 \times 0.37} \approx 3.7\).

In 100 rolls expect \(17 \pm 3.7\) sixes.