Outline

The Sum of Draws

The Square Root Law
   Standard Error of the Sum of Draws

The Central Limit Theorem
The Sum of Draws

We can approximate the probability histogram of the sum of heads in a large number of coin tosses using the normal curve.

Q: A coin is tossed 100 times, what is the probability of getting exactly 50 heads?

A: We can look at the probability histogram for this case. We observe that the chances corresponding to 50 are equal to the area of the box that has a base from 49.5 to 50.5. The area of this box is 7.96%. 
Q: What about an approximation using the normal curve?

A: First step is to calculate the mean and standard deviation. Consider a box model where there is a zero for the tail and 1 for the head,
Average of the Box: $\frac{1}{2}$. SD of the Box: $\frac{1}{2}$

When drawing a ticket from this box 100 times with replacement, the expected value of the sum of the draws is

$$100 \times \frac{1}{2} = 50$$

**Expected Value**

In general, the expected value of the sum of the draws is given by

$$(\text{number of draws}) \times (\text{average of box})$$
The Square Root Law

The standard error of the sum of the draws is given by the square root law:

\[ \sqrt{\text{number of draws}} \times \text{(SD of box)} \]

where SD of box stands for the standard deviation of the list of numbers in the box.
Standard Error of the Sum of Draws

The standard error for the sum of the draws is given by

$$\sqrt{100} \times \frac{1}{2} = 5$$

Now we have to convert the base of the rectangle to standard units:

$$\frac{49.5 - 50}{5} = -0.1 \quad \frac{50.5 - 50}{5} = 0.1$$

So the normal approximation consists of the area under the normal curve for the interval (-0.1, 0.1). According to the table, this is equal to 7.965%.
Q: What are the approximate chances of getting between 45 and 55 heads inclusive?

A: The probability of getting between 45 and 55 heads is equal to the areas of the rectangles between 45 and 55 in the probability histogram. This is approximated by the area under the normal curve for the interval \((44.5,55.5)\). In standard units this corresponds to the interval \((-1.1,1.1)\), which has a probability of \(72.87\%\) according to the table.
Q: What are the approximate chances of getting between 45 and 55 heads exclusive?

A: This time the probability is given by the areas of the rectangles between 46 and 54, which is approximately the area under the curve corresponding to the interval \((45.5, 54.5)\), this is the interval \((-0.9, 0.9)\) in standard units, which has a probability of 63.19%.

Very often it is not specified if the end points are included or not. In that case we consider the approximation using the given interval. So, for the previous example, we would have \((45, 55)\) that is converted to \((-1, 1)\) in standard units and yields 68.27% probability.
When we can use The Normal Approximation?

Consider the box

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

then the probability histogram for the tickets in the box is:
which is far from being normal. Nevertheless, if we consider the experiment of
drawing tickets from the box and sum the results over and over again, then the
probability histogram of the sum will be approximated by the normal curve.
What if we consider the product of the tickets?

In that case the probability histogram will not be approximated by a normal curve, no matter how many draws from the box we take.
The Central Limit Theorem

In general it is true that the probability histogram of the sum of draws from a box of tickets will be approximated by the normal curve. This is a mathematical fact that can be expressed and proved as a theorem.

The Central Limit Theorem

When drawing at random with replacement from a box, the probability histogram for the sum will follow a normal curve, in the limit. This is even if the probability histogram of the contents of the box does not have a probability histogram that is approximately normal.
The reason why the CLT is used as an approximation for distributions of lists of numbers is that it often happens that the uncertainty in the data can be thought of as the sum of several sources of randomness.
Q: Four hundred draws will be made at random with replacement from the box \[1, 3, 5, 7\]. Estimate the chance that the sum of the draws will be more than 1,500.

A: The average in the box is 4 and the SD is about 2.24. The expected value for the sum is \(4 \times 400 = 1,600\) and the SE is \(\sqrt{400} \times 2.24 \approx 45\). Converting 1,500 to standard units we have

\[
\frac{1,500 - 1,600}{45} = -2.22.
\]

According to the normal curve, the chance of being above -2.22 is about 99%.
Q: Four hundred draws will be made at random with replacement from the box \([1, 3, 5, 7]\). Estimate the chance that the sum of the draws will be more than 1,500.

A: The average in the box is 4 and the SD is about 2.24. The expected value for the sum is \(4 \times 400 = 1,600\) and the SE is \(\sqrt{400} \times 2.24 \approx 45\). Converting 1,500 to standard units we have

\[
\frac{1,500 - 1,600}{45} = -2.22.
\]

According to the normal curve, the chance of being above -2.22 is about 99%.
Q: Estimate the chance that there will be fewer than 90 3's.

A: The number of 3's is like the sum of 400 draws from the box \[ \begin{array}{c} 1 \\ 3 \\ 0 \end{array} \] where the ticket marked as 1 corresponds to the 3. The average in such a box is 1/4 and the SD is about 0.43. Thus the expected number of 3's is \[ 400 \times \frac{1}{4} = 100 \] and the SE is \[ \sqrt{400 \times 0.43} = 8.66. \] Converting 90 to standard units we have

\[ \frac{90 - 100}{8.66} = -1.15. \]

According to the normal curve, the chance of being below -1.15 is about 12%.