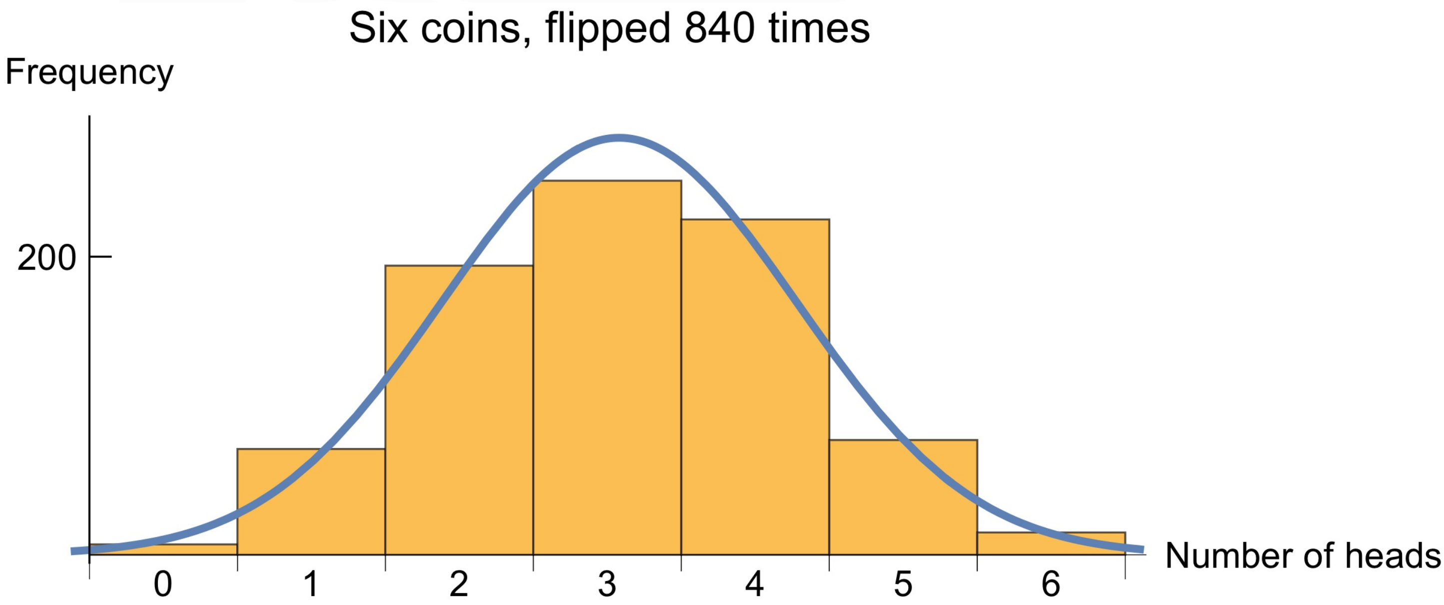
Statistics.

Consider the experiment of flipping six coins, and recording the number of heads.
We did this 840 times. Results:

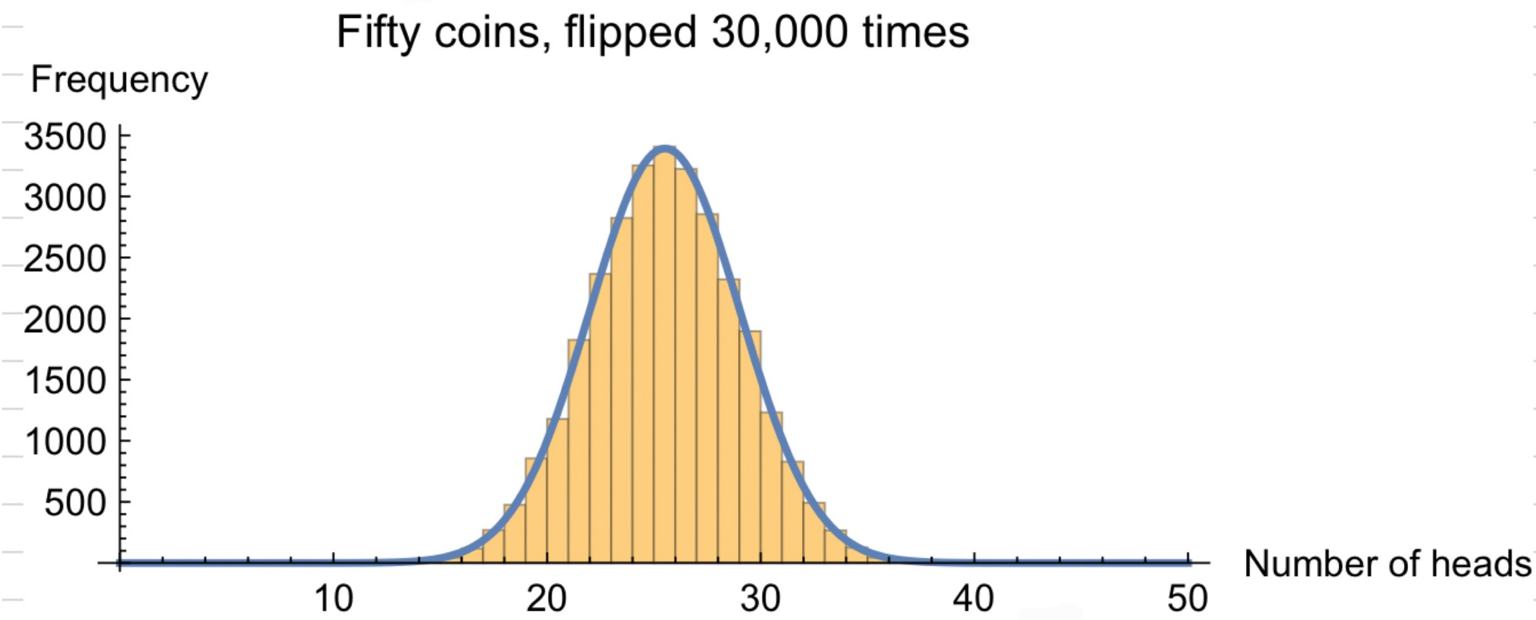
# of heads	Frequency
0	7
1	71
2	194
3	251
4	225
5	77
6	15

Here's a histogram:



Note the (rough) bell shape: a certain "normal curve" (sketched on the histogram) fits the data fairly well. But <u>which</u> normal curve (and what's a normal curve)? Answers soon.

Next, we simulated flipping 50 coins, recording the # of heads, and repeating 30,000 times. Results:



A normal curve fits the data quite closely.

This illustrates a <u>central</u> result in probability:

<u>The Central Limit Theorem (CLT).</u>

If each trial of an experiment comprises many small, independent factors, all of which behave similarly, and many trials are performed, then the outcomes of the experiment will follow a roughly <u>normal</u> distribution.

[Proof omitted.]

<u>Interlude: some formulas.</u>

Consider a data set

$$X = \{x_1, x_2, \ldots, x_n\}$$ of real numbers.

We define the <u>mean</u> $\bar{x}$ and <u>standard deviation</u> s (std dev) of the data by:

(A) General formulas.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

measures "central tendency" of the data

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

measures "spread" of the data.

(B) Formulas for grouped data.

If the data takes <u>only</u> the distinct values $y_1, y_2, \ldots, y_k$, and the value y_i happens f_i times for each i, then

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i \cdot f_i ,$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} f_i (y_i - \bar{x})^2} .$$

(These are different formulas for the same quantities as in part (A).)

Compare these formulas with formulas

$$E[X] = \sum_{\substack{\text{values} \\ x \text{ of } X}} x \cdot P(X=x);$$

$$SD[X] = \sqrt{Var[X]} = \sqrt{\sum_{\substack{\text{values} \\ x \text{ of } X}} (x-\mu)^2 \cdot P(X=x)}$$

$(\mu = E[X])$.

Example

for our "six coins, flipped 840 times" data, we have

$$\bar{x} = \frac{7 \cdot 0 + 71 \cdot 1 + \dots + 15 \cdot 6}{3814} = 3.0797,$$

$$s = \sqrt{\frac{7(0-\bar{x})^2 + \dots + 15(6-\bar{x})^2}{3813}} = 1.1978.$$

Note: we compute that

$\bar{x} - 3s = -0.5135$

$\bar{x} + 3s = 6.6731$

So: all possible data values $0, 1, 2, \dots, 6$ lie in the interval $[\bar{x} - 3s, \bar{x} + 3s]$. That is, all data is "within three standard deviations of the mean." This exemplifies the "empirical rule." More on this later.