Bayes rule: updating probabilities as new information is acquired.

Abstract set-up: Partition $(B_1, B_2, \ldots)$ of "alternate possibilities". Know **prior** probabilities $P(B_i)$.
Then observe some event A happens (the "new information") for which we know $P(A|B_i)$. We want to calculate the **posterior** probabilities $P(B_i|A)$.

Bayes formula:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \ldots}.$$

Example used by psychologists in studying how people think about probability.

Two cab companies serve a city: the Green company operates 85% of the cabs and the Blue company operates 15% of the cabs. One of the cabs is involved in a hit-and-run accident at night, and a witness identifies the hit-and-run cab as a Blue cab. When the court tests the reliability of the witness under circumstances similar to those on the night of the accident, he correctly identifies the color of a cab 80% of the time and misidentifies it the other 20% of the time. What is the probability that the cab involved in the accident was Blue, as stated by the witness?

Many people answer "80%". What does Bayes formula say?

[calculation on board]

Example: Suppose a test for a disease generates the following results:
(i) If a tested patient has the disease, the test returns a positive result 99% of the time.
(ii) If a tested patient does not have the disease, the test returns a negative result 95% of the time.

Suppose also that only 0.1% of the population has that disease.

Consider a person who takes the test and gets a positive result. What is the probability the person really has the disease?

Answer: You can't say, without knowing something about why the person was taking the test. Here are two scenarios.

scenario 1. A patient with symptoms visits a doctor.
scenario 2. Mass screening of whole population.

[calculation on board]
[see http://understandinguncertainty.org/node/182 for some actual data and visualization]

Balls in boxes; conceptual model covers many different stories.

N boxes and k balls. Put each ball independently into a random box.

We'll study the event
A_k: "first k balls all in different boxes".

$$\begin{aligned} P(A_2) &= \tfrac{N-1}{N} \\ P(A_3|A_2) &= \tfrac{N-2}{N} \\ P(A_4|A_3) &= \tfrac{N-3}{N} \\ &\dots \\ P(A_k|A_{k-1}) &= \tfrac{N-(k-1)}{N} \end{aligned}$$

and so

$$P(A_3) = P(A_3|A_2) \times P(A_2) = \tfrac{N-2}{N} \times \tfrac{N-1}{N}$$

$$P(A_4) = P(A_4|A_3) \times P(A_3) = \tfrac{N-3}{N} \times \tfrac{N-2}{N} \times \tfrac{N-1}{N}$$

$$P(A_k) = P(A_k|A_{k-1}) \times P(A_{k-1}) =$$

$$= \tfrac{N-(k-1)}{N} \times \tfrac{N-(k-2)}{N} \times \ldots \times \tfrac{N-1}{N} \times \tfrac{N}{N} = [\text{ board }]$$

Birthday problem. k people in a room. What is the chance some 2 people have the same birthday?

Model: each person's birthday is equally likely to be any of the 365 days, independently.

Under this model, situation is same as in "balls in boxes" model with $N = 365$ boxes:

$$P(\text{some 2 people have the same birthday})$$

$$= 1 - P(\text{all } k \text{ people have different birthdays})$$

$$= 1 - \frac{365!}{(365-k)!365^k}.$$

A well-known **surprising fact** is that, for this chance to be $\approx 50\%$, you need only $k = 23$ people.

[Wikipedia: Birthday problem]

The birthday problem gives a nice illustration of the use of **calculus approximations**. For small x

$$e^{-x} \approx 1 - x$$

$$\log(1-x) \approx -x.$$

Looking at the "balls in boxes" formula for the event A_k: "first k balls all in different boxes":

$$\begin{aligned}
\log P(A_k) &= \sum_{i=1}^{k-1} \log(1 - \tfrac{i}{N}) \\
&\approx \sum_{i=1}^{k-1} -\tfrac{i}{N} = -\tfrac{(k-1)k}{2N}
\end{aligned}$$

and so

$$P(A_k) \approx \exp(-\tfrac{(k-1)k}{2N}).$$

Useful because we can see, for large N, how large k must be to make this chance be $1/2$ say: solve

$$\tfrac{1}{2} = \exp(-\tfrac{(k-1)k}{2N})$$

to get $k \approx 0.5 + \sqrt{2 \log 2}\ \sqrt{N} \approx 1.18\sqrt{N} + 0.5$.

Example. Throw 2 dice, add up the two numbers. The result must be between 2 and 12. We can easily calculate the probabilities $p(i)$ that the sum is exactly i.

i	2	3	4	5	6	7	8	9	10	11	12
$p(i)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

This is an example of a **probability distribution** (usually abbreviated to **distribution**). There are 4 ways we might specify a particular distribution.

(1) Via a numerical table, as above.
(2) Via a formula [board]
(3) Via a graphic – a probability histogram, in this case. Note this is similar to a data histogram.
(4) By saying the name of the distribution, if there is a standard name. [Wikipedia: List of probability distributions]