Probability

The setup

Random "experiment": A well-defined process with an uncertain outcome (for example, toss three fair coins)

Sample space, $\mathcal{S}$: The set of possible outcomes (e.g., { HHH, HHT, HTH, THH, HTT, THT, TTH, TTT })

Event: A set of outcomes (a subset of $\mathcal{S}$) (e.g., A = {exactly one head} = { HTT, THT, TTH })

An event is said to have occurred if one of the outcome it contains occurs.

Basic rules of probability

1. $1 \geq \Pr(A) \geq 0$, for any event A.
2. $\Pr(\mathcal{S}) = 1$, for the sample space $\mathcal{S}$.
3. If A and B are *mutually exclusive*, Pr(A or B) = Pr(A) + Pr(B).

More rules

Pr(not A) = 1 – Pr(A)

Pr(A or B) = Pr(A) + Pr(B) – Pr(A and B)

Conditional probability

Pr(A | B) = "probability of A given B" = Pr(A and B) / Pr(B), provided $\Pr(B) > 0$.

Independence

Events A and B are *independent* if Pr(A and B) = Pr(A) Pr(B); equivalently, Pr(A | B) = Pr(A) and Pr(B | A) = Pr(B).

Still more rules

Pr(A and B) = Pr(B) Pr(A | B) = Pr(A) Pr(B | A)

Pr(A) = Pr(A and B) + Pr(A and not B)
= Pr(B) Pr(A | B) + Pr(not B) Pr(A | not B)

Bayes rule

Pr(A | B) = Pr(A) Pr(B | A) / Pr(B)
= Pr(A) Pr(B | A) / [Pr(A) Pr(B | A) + Pr(not A) Pr(B | not A)]

Examples

1. **Mendel's peas.**

   Mendel's peas had either purple or white flowers. Flower color was due to a single gene with the purple allele (A) dominant to the white allele (a).

   The F_1 hybrid (obtained by crossing two pure-breeding lines, one with purple and one with white flowers) has purple flowers and is heterozygous, Aa.

   (a) **Self the F_1, pick an F_2 seed at random and grow it up.**

   i. **Pr(F_2 plant has white flowers)** = (1/2) × (1/2) = 1/4 = 25%.

   ii. **Pr(F_2 plant has purple flowers)** = 1 – (1/4) = 3/4 = 75%.

   iii. **Pr(F_2 plant has genotype *AA*)** = 1/4 = 25%.

   iv. **Pr(F_2 plant has genotype *AA* | it has purple flowers)**
   = Pr(F_2 has genotype *AA* and purple flowers) / Pr(F_2 has purple flowers)
   = (1/4) / (3/4) = 1/3 ≈ 33%.

   (b) **Self the F_2. Grow up ten of the F_3 seeds at random. To save writing, define the following events:**

   P_2 = { F_2 has purple flowers }
   H_o = { F_2 has genotype *AA* }
   H_e = { F_2 has genotype *Aa* }
   P_3 = { all ten F_3 plants have purple flowers }

   Note that $P_2 = H_e$ or H_o. You may want to write out the following, using words. The use of this notation makes things compact but also somewhat harder to follow.

   i. **Pr(P_2)** = 1 – (1/4) = 3/4 = 75%.

   ii. **Pr(P_3 | H_o)** = 100%.

   iii. **Pr(P_3 | H_e)** = $(3/4)^{10} \approx 5.6\%$.

   iv. **Pr(P_3 | P_2)** = Pr(P_3 and P_2) / Pr(P_2)
   = { Pr(P_3 and H_o) + Pr(P_3 and H_e) } / Pr(P_2)
   = { Pr(H_o) × Pr(P_3 | H_o) + Pr(H_e) × Pr(P_3 | H_e) } / Pr(P_2)
   = { (1/4) × 1 + (1/2) × $(3/4)^{10}$ } / (3/4) ≈ 37%

   Alternatively, Pr(P_3 | P_2) = Pr(P_3 and H_o | P_2) + Pr(P_3 and H_e | P_2)
   = Pr(H_o | P_2) × Pr(P_3 | H_o and P_2) + Pr(H_e | P_2) × Pr(P_3 | H_e and P_2)
   = (1/3) × 1 + (2/3) × $(3/4)^{10} \approx 37\%$

   v. **Pr(H_o | P_3, P_2)** = Pr(H_o and P_3 | P_2) / Pr(P_3 | P_2)
   = Pr(P_3 | H_o and P_2) × Pr(H_o | P_2) / Pr(P_3 | P_2)
   = $1 \times (1/3) / [1 \times (1/3) + (3/4)^{10} \times (2/3)]$
   ≈ 90%.

2. **Suppose a test for HIV correctly gives a positive result, if a person is infected, with probability 99.5%, and correctly gives a negative result, if a person is not infected, with probability 98%.**

   (a) **Suppose that 0.1% of a population are infected with HIV. Consider drawing a person at random and testing him or her for HIV infection.**
   Calculate Pr(infected | test is positive).

   Let I = { the person is infected } and P = { the person tests positive }.

   From the information above, on the *sensitivity* and *specificity* of the test, we have Pr(P | I) = 0.995 and Pr(not P | not I) = 0.98. From this information, Pr(not P | I) = 1 – Pr(P | I) = 0.005 and Pr(P | not I) = 1 – Pr(not P | not I) = 0.02. Note further that Pr(I) = 0.001, and so Pr(not I) = 0.999.

   We seek to calculate Pr(I | P). We use Bayes's rule. (Why use Bayes's rule here? because we want to "turn around the conditioning." We want to write Pr(I | P) in terms of things like Pr(P | I).)

   $$\Pr(I \mid P) = \Pr(I)\Pr(P \mid I) / [\Pr(I)\Pr(P \mid I) + \Pr(\text{not } I)\Pr(P \mid \text{not } I)]$$
   $$= 0.001 \times 0.995 / (0.001 \times 0.995 + 0.999 \times 0.02) \approx 4.7\%$$

   (b) **Consider a person drawn from a high-risk group, so that they have, *a priori*, probability 30% of being infected.**
   Calculate Pr(infected | test is positive).

   In this case, we change Pr(I) = 0.3 and Pr(not I) = 1 – Pr(I) = 0.7.

   Thus, $\Pr(I \mid P) = 0.3 \times 0.995 / (0.3 \times 0.995 + 0.7 \times 0.02) \approx 96\%$

3. **The "Monty Hall" problem.**

You're on a game show, and are presented with three doors. One hides a car; the other two hide goats. You're allowed to choose a door. Monty Hall then opens one of the other two doors to reveal a goat. You are now allowed to either stick with the door you chose initially, or switch to the other closed door. What should you do (assuming that you are hoping to get the car, and not a goat): stick with your original choice, or switch?

The answer depends on Monty's behavior.

(a) **If you choose the door with the car, Monty opens one of the other two doors at random. If you choose a door hiding a goat, Monty opens the other door with a goat.**

(b) **Monty Hall opens one of the other two doors at random. If he had revealed the car, you would have lost, but you happen to be in the situation where he revealed a goat.**

Let us call the doors A, B and C, and let us assume ("without loss of generality," as mathematicians like to say) that you choose door A. Moreover, let us initially *not* condition on Monty revealing the goat.

There are six possible outcomes of the exeriment, { C[G]G, CG[G], G[C]G, GC[G], G[G]C, GG[C] }. (The three letters denote the objects behind the three doors, so that CGG means the car is behind door A. The boxes indicate which door Monty opens. Since you chose door A, he won't be opening that door, but one of the other two.)

In scenario (a), Monty always opens a door with a goat. Thus, if the door you chose hid the car (that is, if the car was behind door A), he'll choose at random between the doors B and C. On the other hand, if the car is behind door B, he'll always open door C, and vice versa. Thus, under scenario A, the events { C[G]G } and { CG[G] } each have probability 1/6; events { GC[G] } and { G[G]C } each have probability 1/3, and events { G[C]G } and { GG[C] } each have probability 0.

In scenario (b), in which Monty chooses at random between the two remaining doors, the six outcomes listed above each have probability 1/6.

Now, define the events A = { the car is behind door A } = { C[G]G, CG[G] }, B = { the car is behind door B } = { G[C]G, GC[G] }, C = { the car is behind door C } = { G[G]C, GG[C] }, and D = { Monty opens a door with a goat } = { C[G]G, CG[G], GC[G], G[G]C }.

Our objective: we want to know the chance that the door we originally chose hides the car, *given that Monty has opened a door revealing a goat*. This is Pr(A | D). If this probability is $< 1/2$, we should switch doors; if it is $> 1/2$, we should stick with our first choice; if it is $= 1/2$, it doesn't matter whether we switch or stick.

The event "A and D" (the intersection between A and D) is simply event A, and Pr(A) = 1/3 in either case (a) or (b), so Pr(A and D) = 1/3.
In case (a), Pr(D) = 1, and so Pr(A | D) = 1/3. We should switch doors.
In case(b), Pr(D) = 2/3, and so Pr(A | D) = 1/2. It doesn't matter what we do.
The appropriate action to take depends on what Monty is doing!

4. **Mendel, revisited.**

   Mendel's peas had either purple or white flowers; flower color is due to a single gene, for which the purple allele (*A*) is dominant to the white allele (*a*).

   We cross two pure-breeding lines (one purple and one white) to produce the F_1 hybrid. We self the F_1 and choose an F_2 seed at random. We grow and self the F_2 and choose two F_3 seeds at random.

   Consider the following events.

$$\begin{aligned} \mathrm{P} &= \{\mathrm{F}_2 \text{ has purple flowers}\} \\ \mathrm{O} &= \{\mathrm{F}_2 \text{ is homozygous } AA\} \\ \mathrm{E} &= \{\mathrm{F}_2 \text{ is heterozygous } Aa\} \\ \mathrm{W} &= \{\mathrm{F}_2 \text{ has white flowers}\} \\ \mathrm{A}_1 &= \{\mathrm{F}_3 \text{ number 1 has purple flowers}\} \\ \mathrm{A}_2 &= \{\mathrm{F}_3 \text{ number 2 has purple flowers}\} \end{aligned}$$

   Are A_1 and A_2 independent?

   A_1 and A_2 are *not* independent! A_1 and A_2 are *conditionally independent*, given the genotype of the F_2 plant. But, as we will see, in the absence of information about the F_2 parent's genotype, the flower color of the F_3 plants are not independent.

   First, note that the events E, O and W are *mutually exclusive*, and that [Pr(E) + Pr(O) + Pr(W)] = 1.

$$\begin{aligned} \Pr(\mathrm{A}_1) = \Pr(\mathrm{A}_2) &= \Pr(\mathrm{A}_1 \mid \mathrm{E})\Pr(\mathrm{E}) + \Pr(\mathrm{A}_1 \mid \mathrm{O})\Pr(\mathrm{O}) + \Pr(\mathrm{A}_1 \mid \mathrm{W})\Pr(\mathrm{W}) \\ &= (3/4) \times (1/2) + (1) \times (1/4) + 0 \times (1/4) = 5/8 \approx 63\%. \end{aligned}$$

$$\begin{aligned} \Pr(\mathrm{A}_1 \text{ and } \mathrm{A}_2) &= \Pr(\mathrm{A}_1 \text{ and } \mathrm{A}_2 \mid \mathrm{E})\Pr(\mathrm{E}) + \Pr(\mathrm{A}_1 \text{ and } \mathrm{A}_2 \mid \mathrm{O})\Pr(\mathrm{O}) \\ &\quad + \Pr(\mathrm{A}_1 \text{ and } \mathrm{A}_2 \mid \mathrm{W})\Pr(\mathrm{W}) \\ &= (3/4)^2 \times (1/2) + 1 \times (1/4) + 0 \times (1/2) = 17/32 \approx 53\%. \end{aligned}$$

   Thus $\Pr(\mathrm{A}_2 \mid \mathrm{A}_1) = \Pr(\mathrm{A}_1 \text{ and } \mathrm{A}_2) / \Pr(\mathrm{A}_1) = (17/32) / (5/8) = 17/20 = 85\%$, which is considerably greater than $\Pr(\mathrm{A}_2)$. And so, A_2 and A_1 are not independent.