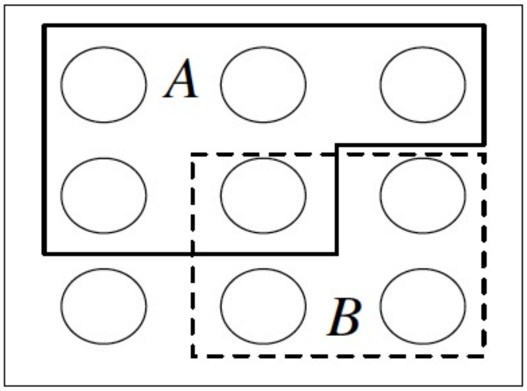


Figure 1: A sample space as Pebble World, with two events A and B spotlighted. Each pebble represents an outcome, and an event is a set of pebbles. Performing an experiment amounts to randomly selecting one pebble. A is a set of 5 pebbles, and B is a set of 4 pebbles. $A \cup B$ consists of the 8 pebbles in A or B, $A \cap B$ consists of the pebble that is in both A and B, and A^c consists of the 4 pebbles that are not in A.

1 Probability

Definition (Sample space and event)**.** The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of the sample space S, and we say that A *occurred* if the actual outcome is in A.

Definition (Naive definition of probability)**.** Let A be an event for experiment with a finite sample space S. The *naive probability* of A is

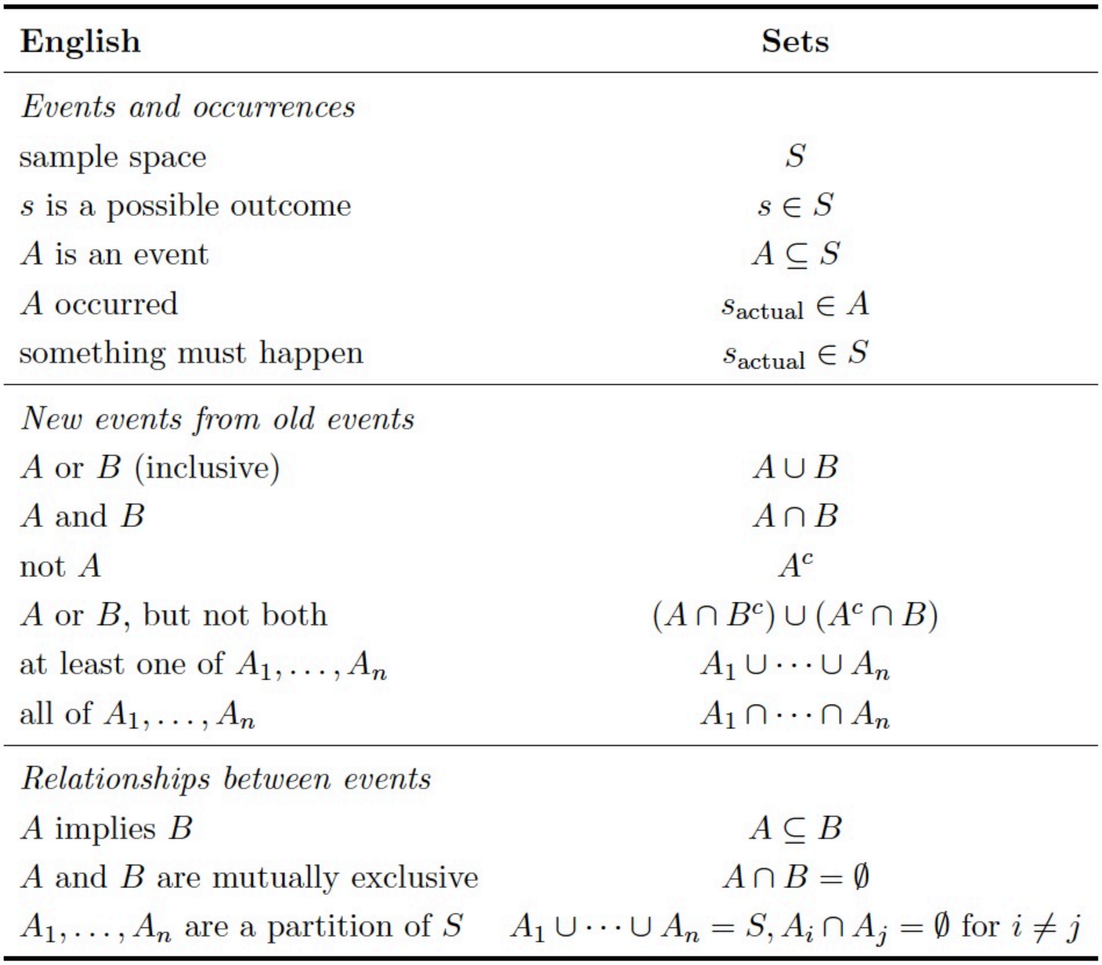
$$\mathsf{P}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to } A}{\text{total number of outcomes in } S}.$$

This definition applies only when it is reasonable to assume by design or symmetry that the outcomes are equally likely. Otherwise, the naive definition of probability does *not* apply.

Example: Birthday problem

There are k people in a room. Assume each person's birthday is equally likely to be any of the 365 days of the year (we exclude February 29), and that people's birthdays are independent (we assume there are no twins in the room). What is the probability that two or more people in the group have the same birthday?

Solution: There are 365^k ways to assign birthdays to the people in the room since we can imagine the 365 days of the year being sampled k times, with replacement. By assumption, all these possibilities are equally likely, so the naive definition of probability applies. Used directly, the naive definition says we just need to count the number of ways to assign birthdays to k people such that there are two or more people who share



English	**Sets**
Events and occurrences	
sample space	S
s is a possible outcome	$s \in S$
A is an event	$A \subseteq S$
A occurred	$s_{\text{actual}} \in A$
something must happen	$s_{\text{actual}} \in S$
New events from old events	
A or B (inclusive)	$A \cup B$
A and B	$A \cap B$
not A	A^c
A or B, but not both	$(A \cap B^c) \cup (A^c \cap B)$
at least one of $A_1, \ldots, A_n$	$A_1 \cup \cdots \cup A_n$
all of $A_1, \ldots, A_n$	$A_1 \cap \cdots \cap A_n$
Relationships between events	
A implies B	$A \subseteq B$
A and B are mutually exclusive	$A \cap B = \emptyset$
$A_1, \ldots, A_n$ are a partition of S	$A_1 \cup \cdots \cup A_n = S, A_i \cap A_j = \emptyset$ for $i \neq j$

Figure 2: Mini-dictionary for converting between English and sets. Here S is the sample space and s_{actual} is the actual outcome of the experiment (e.g., the pebble that ends up getting chosen when the experiment is performed).

a birthday. But this counting problem is hard, since it could be Emma and Steve who share a birthday, or Steve and Naomi, or all three of them, or the three of them could share a birthday while two others in the group share a different birthday, or various other possibilities.

Instead, let's count the complement: the number of ways to assign birthdays to k people such that no two people share a birthday. This amounts to sampling the 365 days of the year without replacement, so the number of possibilities is $365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - k + 1)$ for $k \leq 365$. Therefore the probability of no birthday matches in a group of k people is

$$\mathsf{P}(\text{no birthday match}) = \frac{365 \cdot 364 \cdot 363 \cdot \ldots \cdot (365 - k + 1)}{365^k},$$

and the probability of at least one birthday match is

$$\mathsf{P}(\text{at least one birthday match}) = 1 - \mathsf{P}(\text{no birthday match}).$$

Definition (General definition of probability)**.** A *probability space* consists of a sample space S and a *probability function* $\mathsf{P}(\cdot)$ which takes an event $A \subseteq S$ as input and returns $\mathsf{P}(A)$, a real number between 0 and 1, as output. The probability function must satisfy the following axioms:

1. $\mathsf{P}(\emptyset) = 0$, $\mathsf{P}(S) = 1$.
2. If $A_1, A_2, \ldots$ are disjoint (or mutually exclusive) events (i.e., $A_i \cap A_j = \emptyset$ for $i \neq j$), then
$$\mathsf{P}\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} \mathsf{P}(A_j).$$

In Pebble World, the definition says that probability behaves like mass: the mass of an empty pile of pebbles is 0, the total mass of all the pebbles is 1, and if we have non-overlapping piles of pebbles, we can get their combined mass by adding the masses of the individual piles. Unlike in the naive case, we can now have pebbles of differing masses, and we can also have a countably infinite number of pebbles as long as their total mass is 1. We can even have uncountable sample spaces, such as having S be an area in the plane. In this case, instead of pebbles, we can visualize mud spread out over a region, where the total mass of the mud is 1.

Any function $\mathsf{P}(\cdot)$ (mapping events to numbers in the interval $[0, 1]$) that satisfies the two axioms is considered a valid probability function. However, the axioms don't tell us how probability should be interpreted; different schools of thought exist.

- The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment: if we say a coin has probability 1/2 of Heads, that means the coin would land Heads 50% of the time if we tossed it over and over and over.
- The *Bayesian* view of probability is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses like "candidate A will win the election" or "the defendant is guilty" even if it isn't possible to repeat the same election or the same crime over and over again.

Theorem. *Properties of probability. A probability function has the following properties, for any events A and B.*

1. $\mathsf{P}(A^c) = 1 - P(A)$.
2. *If* $A \subseteq B$, *then* $\mathsf{P}(A) \leq \mathsf{P}(B)$.
3. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$.

The last property is called the *inclusion-exclusion* principle which generalizes to any number of events. For three events, inclusion-exclusion says

$$\begin{aligned} \mathsf{P}(A \cup B \cup C) &= \mathsf{P}(A) + \mathsf{P}(B) + \mathsf{P}(C) \\ &\quad - \mathsf{P}(A \cap B) - \mathsf{P}(A \cap C) - \mathsf{P}(B \cap C) \\ &\quad + \mathsf{P}(A \cap B \cap C). \end{aligned}$$

2 Conditional Probability

Whenever we observe new evidence (i.e., obtain *data*), we acquire information that may affect our uncertainties. A new observation that is consistent with an existing belief could make us more sure of that belief, while a surprising observation could throw that belief into question. *Conditional probability* is the concept that addresses this fundamental question: how should we update our beliefs in light of the evidence we observe? Conditional probability is essential for scientific, medical, and legal reasoning, since it shows how to incorporate evidence into our understanding of the world in a logical, coherent manner. In fact, a useful perspective is that **all probabilities are conditional**; whether or not it's written explicitly, there is always background knowledge (or assumptions) built into every probability.

Definition (Conditional probability)**.** If A and B are events with $\mathsf{P}(B) > 0$, then the conditional probability of A given B, denoted by $\mathsf{P}(A \mid B)$, is defined as:

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}.$$

Here A is the event whose uncertainty we want to update, and B is the evidence we observe. We call $\mathsf{P}(A)$ the prior probability of A (prior = before updating based on the evidence), and $\mathsf{P}(A \mid B)$ the posterior probability of A (posterior = updating based on the evidence).

Remark that, for any event A with $\mathsf{P}(A) > 0$, $\mathsf{P}(A \mid A) = \mathsf{P}(A \cap A)/\mathsf{P}(A) = 1$. After observing that A has occurred, our updated probability for A is 1. Moreover, in general, it is true that $\mathsf{P}(A \mid B) \neq \mathsf{P}(B \mid A)$.

Note (Frequentist interpretation). Recall that the frequentist interpretation of probability is based on relative frequency over a large number of repeated trials. Imagine repeating our experiment many times, generating a long list of observed outcomes. The conditional probability of A given B can then be thought of in a

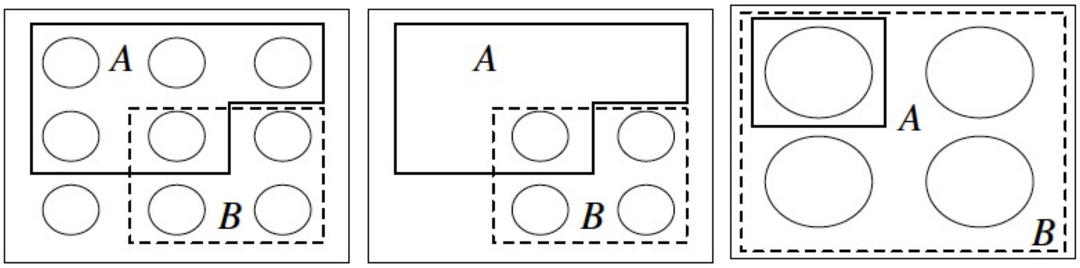


Figure 3: Pebble World intuition for $\mathsf{P}(A \mid B)$. Consider a finite sample space, with the outcomes visualized as pebbles of total mass 1. **Left panel**. Since A and B are events (subsets of the sample space), they are sets of pebbles. **Middle panel**. Once we know B has occurred, we get rid of the outcomes (pebbles) in B^c because they are incompatible with the knowledge that B has occurred. Then $\mathsf{P}(A \mid B)$ is the total mass of pebbles remaining in A. **Right panel**. In the restricted sample space, we renormalize, that is, divide all the masses by a constant so that the new total mass of the remaining pebbles is 1. This is achieved by dividing by $\mathsf{P}(B)$, the total mass of the pebbles in B. The updated mass of the outcomes corresponding to event A is the conditional probability $\mathsf{P}(A \mid B) = \mathsf{P}(A \cap B)/\mathsf{P}(B)$. In this way, our probabilities have been updated in accordance with the observed evidence. Outcomes that contradict the evidence are discarded, and their mass is redistributed among the remaining outcomes, preserving the relative masses of the remaining outcomes. For example, if pebble 2 weighs twice as much as pebble 1 initially, and both are contained in B, then after conditioning on B it is still true that pebble 2 weighs twice as much as pebble 1. But if pebble 2 is not contained in B, then after conditioning on B its mass is updated to 0.

natural way: it is the fraction of times that A occurs, restricting attention to the trials where B occurs. Let n_A, n_B, n_{AB} be the number of occurrences of $A, B, A \cap B$ respectively in a large number n of repetitions of the experiment. The frequentist interpretation is that

$$\mathsf{P}(A) \approx \frac{n_A}{n}, \mathsf{P}(B) \approx \frac{n_B}{n}, \mathsf{P}(A \cap B) \approx \frac{n_{AB}}{n}.$$

Then $\mathsf{P}(A \mid B)$ is interpreted as n_{AB}/n_B, which equals $(n_{AB}/n)/(n_B/n)$. This translates into $\mathsf{P}(A \mid B) = \mathsf{P}(A \cap B)/\mathsf{P}(B)$.

Theorem. *For any events A and B with positive probabilities,*

$$\mathsf{P}(A \cap B) = \mathsf{P}(B)\mathsf{P}(A \mid B) = \mathsf{P}(A)\mathsf{P}(B \mid A).$$

This theorem generalizes to an arbitrary number of events.

Theorem. *For any events $A_1, \ldots, A_n$ with positive probabilities,*

$$\mathsf{P}(A_1, A_2, \ldots, A_n) = \mathsf{P}(A_1)\mathsf{P}(A_2 \mid A_1)\mathsf{P}(A_3 \mid A_1, A_2) \ldots \mathsf{P}(A_n \mid A_1, \ldots, A_{n-1}).$$

The commas denote intersections.

Theorem (Bayes' rule)**.**

$$\mathsf{P}(A \mid B) = \frac{\mathsf{P}(B \mid A)\mathsf{P}(A)}{\mathsf{P}(B)}.$$

Definition (Odds)**.** The odds in favor of an event A are

$$\text{odds}(A) = \mathsf{P}(A)/\mathsf{P}(A^c), \mathsf{P}(A) = \text{odds}(A)/(1 + \text{odds}(A)).$$

Theorem (Odds form of Bayes' rule)**.** *For any events A and B with positive probabilities, the odds of A after conditioning on B are*

$$\frac{\mathsf{P}(A \mid B)}{\mathsf{P}(A^c \mid B)} = \frac{\mathsf{P}(B \mid A)}{\mathsf{P}(B \mid A^c)} \cdot \frac{\mathsf{P}(A)}{\mathsf{P}(A^c)}.$$

Interpretation: the posterior odds $\mathsf{P}(A \mid B)/\mathsf{P}(A^c \mid B)$ are equal to the prior odds $\mathsf{P}(A)/\mathsf{P}(A^c)$ times the factor $\mathsf{P}(B \mid A)/\mathsf{P}(B \mid A^c)$, which is known as the likelihood ratio.

Theorem (Law of total probability (LOTP))**.** *Let $A_1, \ldots, A_n$ be a partition of the sample space S with $\mathsf{P}(A_i) > 0$ for all i. Then*

$$\mathsf{P}(B) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i)\mathsf{P}(A_i).$$

Example: Testing for a rare disease

A patient named Fred is tested for a disease called conditionitis, a medical condition that affects 1% of the population. The test result is positive, i.e., the test claims that Fred has the disease. Let D be the event that Fred has the disease and T be the event that he tests positive.

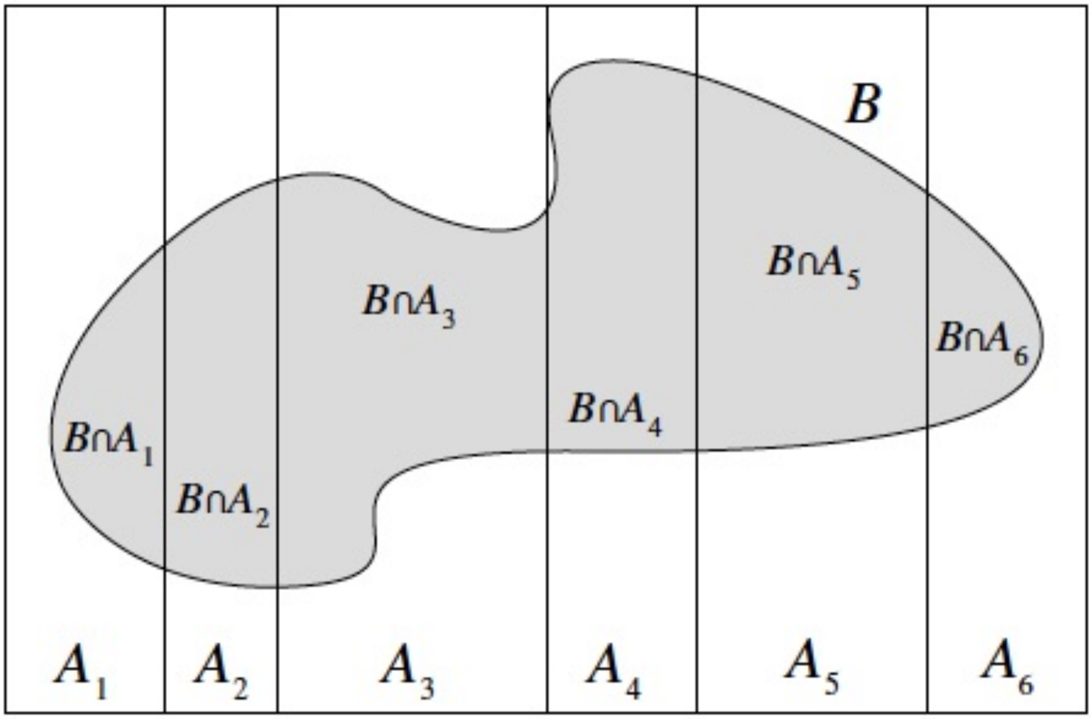


Figure 4: Graphical illustration of the law of total probability (LOTP).

Suppose that the test is "95% accurate"; there are different measures of the accuracy of a test, but in this problem it is assumed to mean that $\mathsf{P}(T \mid D) = 0.95$ and $\mathsf{P}(T^c \mid D^c) = 0.95$. The quantity $\mathsf{P}(T \mid D)$ is known as the sensitivity or true positive rate of the test, and $\mathsf{P}(T^c \mid D^c)$ is known as the specificity or true negative rate.

Find the conditional probability that Fred has conditionitis, given the evidence provided by the test result.

Solution: Applying Bayes' rule and the law of total probability, we have

$$\begin{aligned}\mathsf{P}(D \mid T) &= \frac{\mathsf{P}(T \mid D)\mathsf{P}(D)}{\mathsf{P}(T)},\\ &= \frac{\mathsf{P}(T \mid D)\mathsf{P}(D)}{\mathsf{P}(T \mid D)\mathsf{P}(D) + \mathsf{P}(T \mid D^c)\mathsf{P}(D^c)},\\ &= \frac{0.95 \cdot 0.01}{0.95 \cdot 0.01 + 0.05 \cdot 0.99},\\ &\approx 0.16.\end{aligned}$$

There is only a 16% chance that Fred has conditionitis, given that he tested positive, even though the test seems to be quite reliable.

Many people, including doctors, find it surprising that the conditional probability of having the disease given a positive test result is only 16%, even though the test is 95% accurate. The key to understanding this surprisingly high posterior probability is to realize that there are two factors at play: the evidence from the test, and our prior information about the prevalence of the disease. Although the test provides evidence in favor of disease, conditionitis is also a rare condition! The conditional probability $\mathsf{P}(T \mid D)$ reflects a balance between these two factors, appropriately weighing the rarity of the disease against the rarity of a mistaken test result. See Figure 5.

Note (Conditional probabilities are probabilities). When we condition on an event E, we update our beliefs to be consistent with this knowledge, effectively putting ourselves in a universe where we know that E occurred. Within our new universe, however, the laws of probability operate just as before. Conditional probability satisfies all the properties of probability! Therefore, any of the results we have derived about probability are still valid if we replace all unconditional probabilities with probabilities conditional on E. In particular:

- Conditional probabilities are between 0 and 1.
- $\mathsf{P}(S \mid E) = 1$, $\mathsf{P}(\emptyset \mid E) = 0$.
- If $A_1, A_2, \ldots$ are disjoint, then $\mathsf{P}(\cup_{j=1}^{\infty} A_j \mid E) = \sum_{j=1}^{\infty} \mathsf{P}(A_j \mid E)$.
- $\mathsf{P}(A^c \mid E) = 1 - \mathsf{P}(A \mid E)$.
- Inclusion-exclusion principle: $\mathsf{P}(A \cup B \mid E) = \mathsf{P}(A \mid E) + \mathsf{P}(B \mid E) - \mathsf{P}(A \cap B \mid E)$.

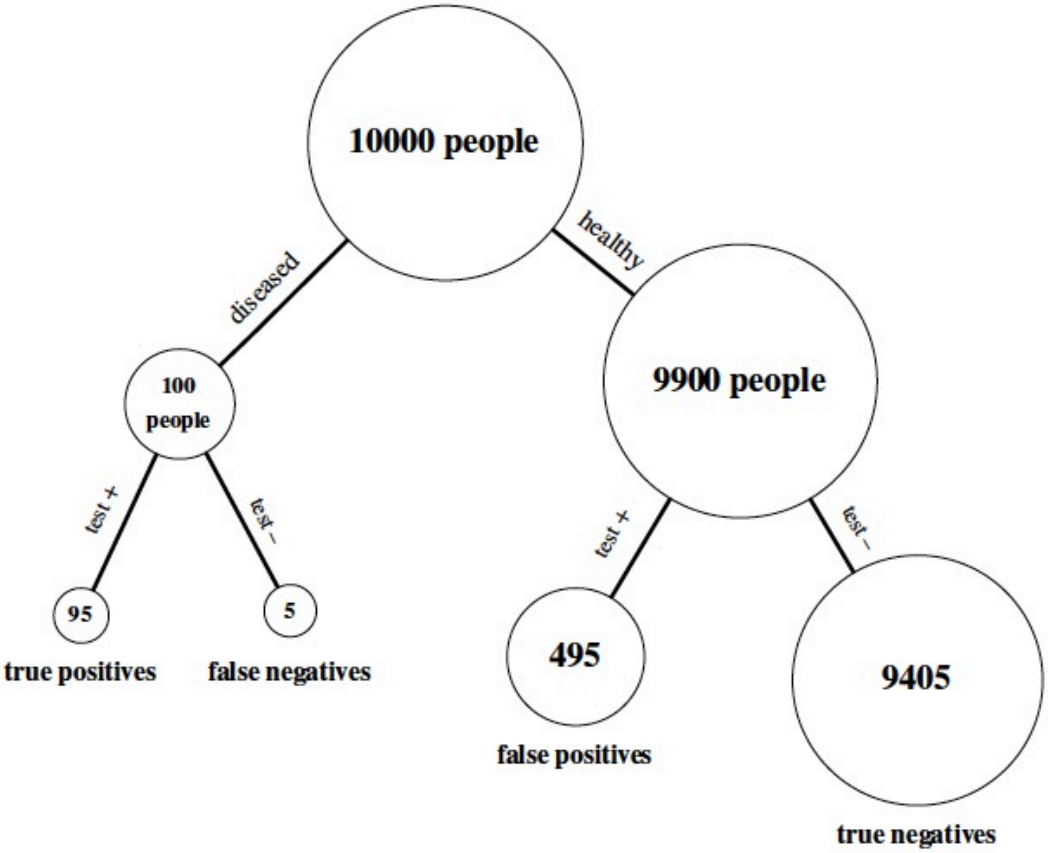


Figure 5: Testing for a rare disease in a population of 10000 people, where the prevalence of the disease is 1% and the true positive and true negative rates are both equal to 95%. For further intuition, consider a population of 10000 people as illustrated, where 100 have conditionitis and 9900 don't; this corresponds to a 1% disease rate. If we tested everybody in the population, we'd expect that out of the 100 diseased individuals, 95 would test positive and 5 would test negative. Out of the 9900 healthy individuals, we'd expect $(0.95)(9900) \approx 9405$ to test negative and 495 to test positive. The 95 true positives (i.e., the individuals who test positive and have the disease) are far outnumbered by the 495 false positives (i.e., the individuals who test positive despite not having the disease). So most people who test positive for the disease are actually disease-free!

Conversely, all probabilities can be thought of as conditional probabilities: whenever we make a probability statement, there is always some background information that we are conditioning on, even if we don't state it explicitly. Since all probabilities are conditional on background information, we can imagine that there is always a vertical conditioning bar, with background knowledge K to the right of the vertical bar. Then the unconditional probability $\mathsf{P}(A)$ is just shorthand for $\mathsf{P}(A \mid K)$; the background knowledge is absorbed into the letter P instead of being written explicitly. Remember that: *Conditional probabilities are probabilities, and all probabilities are conditional.*

Theorem (Bayes' rule with extra conditioning)**.** *Provided that* $\mathsf{P}(AcapE) > 0$ *and* $\mathsf{P}(B \cap E) > 0$*, we have*

$$\mathsf{P}(A \mid B, E) = \frac{\mathsf{P}(B \mid A, E)\mathsf{P}(A \mid E)}{\mathsf{P}(B \mid E)}.$$

Theorem (Law of total probability (LOTP) with extra conditioning)**.** *Let* $A_1, \ldots, A_n$ *be a partition of the sample space* S *with* $\mathsf{P}(A_i \cap E) > 0$ *for all* i*. Then*

$$\mathsf{P}(B \mid E) = \sum_{i=1}^{n} \mathsf{P}(B \mid A_i, E)\mathsf{P}(A_i \mid E).$$

Definition (Independence of two events)**.** Events A and B are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$. If $\mathsf{P}(A) > 0$ and $\mathsf{P}(B) > 0$, then this is equivalent with $\mathsf{P}(A \mid B) = \mathsf{P}(A)$, and also equivalent with $\mathsf{P}(B \mid A) = \mathsf{P}(B)$. Independence is a symmetric relation.

Proposition. *If A and B are independent, then A and B^c are independent, A and B^c are independent, and A^c and B^c are independent.*

Definition (Independence of three events)**.** Events A, B and C are said to be independent if all of the following relations hold:

$$\begin{aligned}
\mathsf{P}(A \cap B) &= \mathsf{P}(A)\mathsf{P}(B), \\
\mathsf{P}(A \cap C) &= \mathsf{P}(A)\mathsf{P}(C), \\
\mathsf{P}(B \cap C) &= \mathsf{P}(B)\mathsf{P}(C), \\
\mathsf{P}(A \cap B \cap C) &= \mathsf{P}(A)\mathsf{P}(B)\mathsf{P}(C).
\end{aligned}$$

If the first three conditions hold, we say that A, B, and C are pairwise independent. Pairwise independence does not imply independence: it is possible that just learning about A or just learning about B is of no use in predicting whether C occurred, but learning that both A and B occurred could still be highly relevant for C. On the other hand, $\mathsf{P}(A \cap B \cap C) = \mathsf{P}(A)\mathsf{P}(B)\mathsf{P}(C)$ does not imply pairwise independence: consider the case when $\mathsf{P}(A) = 0$, for example.

Definition (Independence of many events)**.** For n events $A_1, A_2, \ldots, A_n$ to be independent, we require any pair to satisfy $\mathsf{P}(A_i \cap A_j) = \mathsf{P}(A_i)\mathsf{P}(A_j)$ (for $i \neq j$), any triplet to satisfy $\mathsf{P}(A_i \cap A_j \cap A_k) = \mathsf{P}(A_i)\mathsf{P}(A_j)\mathsf{P}(A_k)$

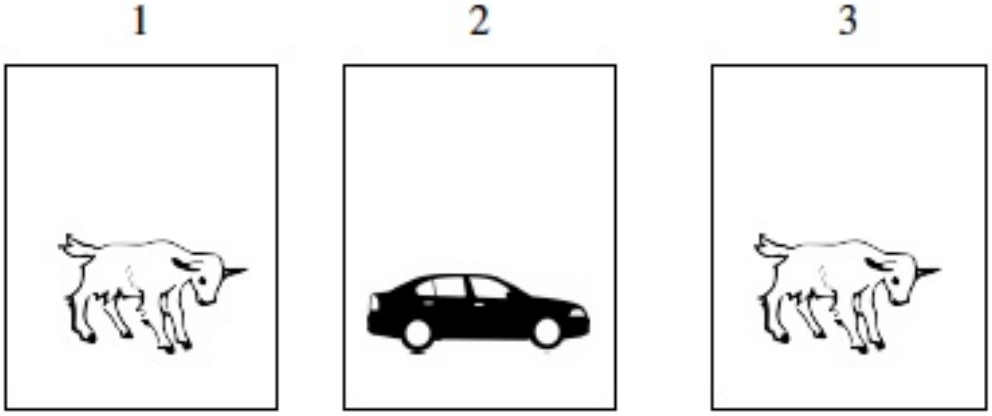


Figure 6: Month Hall game.

(for i, j, k distinct), and similarly for quadruplets, quintuplets, and so on. For infinitely many events, we say that they are independent if every finite subset of the events is independent.

Definition (Conditional independence)**.** Events A and B are said to be conditionally independent given E if $\mathsf{P}(A \cap B \mid E) = \mathsf{P}(A \mid E)\mathsf{P}(B \mid E)$.

Remember that conditional independence does not imply independence, and also that independence does not imply conditional independence.

Example: Monty Hall

On the game show Let's Make a Deal, hosted by Monty Hall, a contestant chooses one of three closed doors, two of which have a goat behind them and one of which has a car. Monty, who knows where the car is, then opens one of the two remaining doors. The door he opens always has a goat behind it (he never reveals the car!). If he has a choice, then he picks a door at random with equal probabilities. Monty then offers the contestant the option of switching to the other unopened door. If the contestant's goal is to get the car, should she switch doors?

Solution: Let's label the doors 1 through 3. Without loss of generality, we can assume the contestant picked door 1 (if she didn't pick door 1, we could simply relabel the doors, or rewrite this solution with the door numbers permuted). Monty opens a door, revealing a goat. As the contestant decides whether or not to switch to the remaining unopened door, what does she really wish she knew? Naturally, her decision would be a lot easier if she knew where the car was! This suggests that we should condition on the location of the car. Let C_i be the event that the car is behind door i, for $i = 1, 2, 3$. By the law of total probability,

$$\mathsf{P}(\text{get car}) = \mathsf{P}(\text{get car} \mid C_1) \cdot \frac{1}{3} + \mathsf{P}(\text{get car} \mid C_2) \cdot \frac{1}{3} + \mathsf{P}(\text{get car} \mid C_3) \cdot \frac{1}{3}.$$

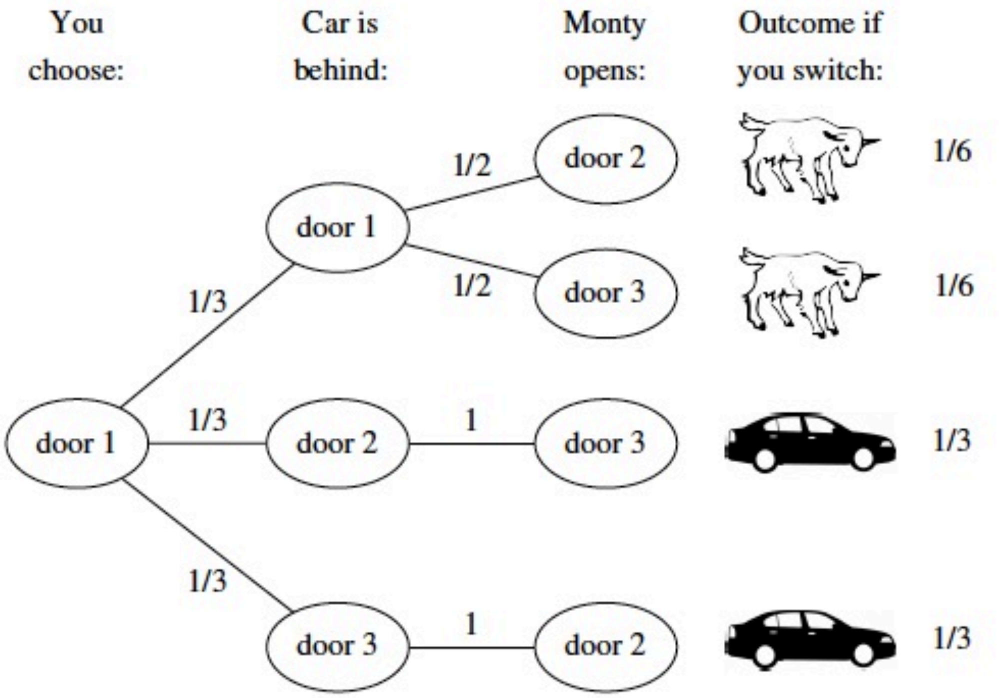


Figure 7: Tree diagram of Monty Hall problem. Switching gets the car $2/3$ of the time.

Suppose the contestant employs the switching strategy. If the car is behind door 1, then switching will fail, so $\mathsf{P}(\text{get car} \mid C_1) = 0$. If the car is behind door 2 or 3, then because Monty always reveals a goat, the remaining unopened door must contain the car, so switching will succeed. Thus,

$$\mathsf{P}(\text{get car}) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3},$$

so the switching strategy succeeds $2/3$ of the time. The contestant should switch to the other door.

Figure 7 is a tree diagram of the argument we have just outlined: using the switching strategy, the contestant will win as long as the car is behind doors 2 or 3, which has probability $2/3$. We can also give an intuitive frequentist argument in favor of switching. Imagine playing this game 1000 times. Typically, about 333 times your initial guess for the car's location will be correct, in which case switching will fail. The other 667 or so times, you will win by switching.

There's a subtlety though, which is that when the contestant chooses whether to switch, she also knows which door Monty opened. We showed that the unconditional probability of success is $2/3$ (when following the switching strategy), but let's also show that the conditional probability of success for switching, given the information that Monty provides, is also $2/3$. Let M_j be the event that Monty opens door j for $j = 2, 3$. Then

$$\mathsf{P}(\text{get car}) = \mathsf{P}(\text{get car} \mid M_2)\mathsf{P}(M_2) + \mathsf{P}(\text{get car} \mid M_3)\mathsf{P}(M_3),$$

where by symmetry $\mathsf{P}(M_2) = \mathsf{P}(M_3) = 1/2$ and $\mathsf{P}(\text{get car} \mid M_2) = \mathsf{P}(\text{get car} \mid M_3) = x$. The symmetry here is that there is nothing in the statement of the problem that distinguishes between door 2 and door 3. Then

$$\frac{2}{3} = \mathsf{P}(\text{get car}) = \frac{x}{2} + \frac{x}{2} = x.$$

Bayes' rule also works nicely for finding the conditional probability of success using the switching strategy, given the evidence. Suppose that Monty opens door 2. Using the notation and results above,

$$\mathsf{P}(C_1 \mid M_2) = \frac{\mathsf{P}(M_2 \mid C_1)\mathsf{P}(C_1)}{\mathsf{P}(M_2)} = \frac{(1/2)(1/3)}{1/2} = \frac{1}{3}.$$

So given that Monty opens door 2, there is a $1/3$ chance that the contestant's original choice of door has the car, which means that there is a $2/3$ chance that the switching strategy will succeed.

Many people, upon seeing this problem for the first time, argue that there is no advantage to switching: "There are two doors remaining, and one of them has the car, so the chances are 50-50." To build correct intuition, let's consider an extreme case. Suppose that there are a million doors, 999,999 of which contain goats and 1 of which has a car. After the contestant's initial pick, Monty opens 999,998 doors with goats behind them and offers the choice to switch. In this extreme case, it becomes clear that the probabilities are not 50-50 for the two unopened doors; very few people would stubbornly stick with their original choice. The same is true for the three-door case.