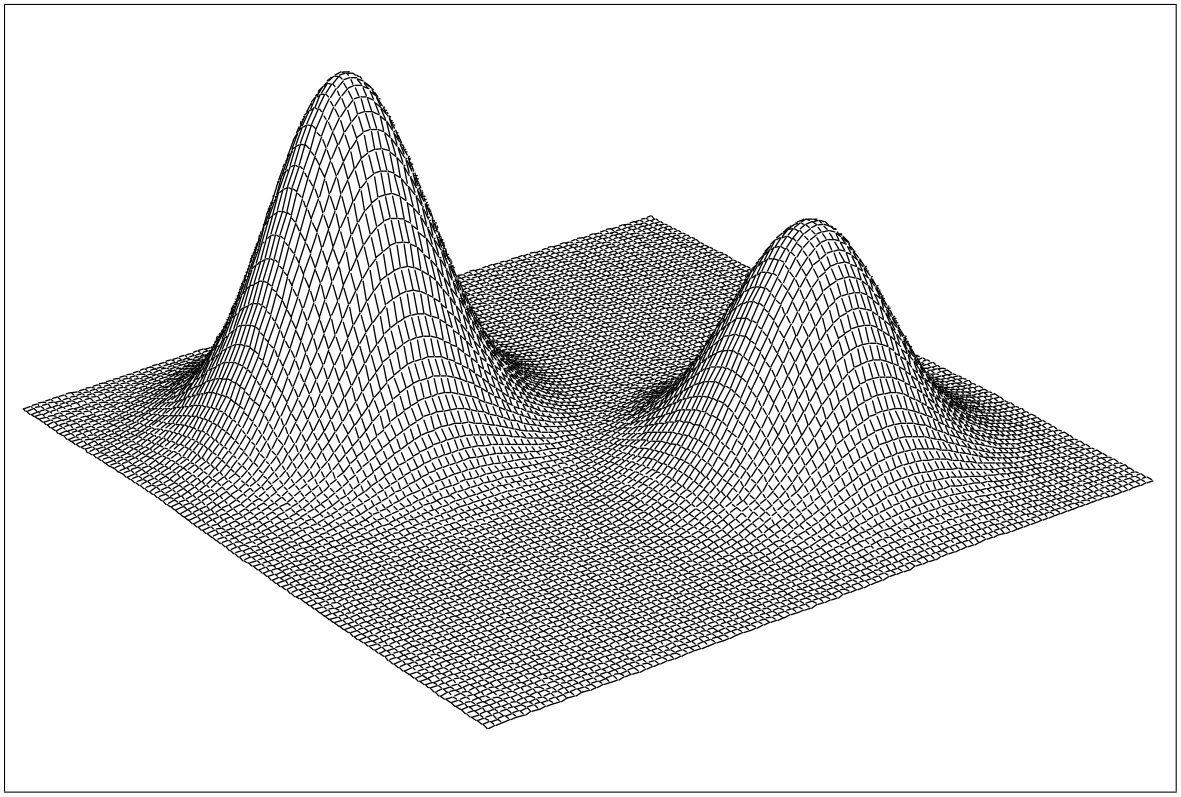
Statistics 610
Practical Theory of Probability



Daren B.H. Cline
Department of Statistics
Texas A&M University
25 August 2025

1. The Probability Measure

1.1 Randomness

The objective of probability is somehow to make sense out of the type of uncertainty known as randomness. In particular, can we use it to make some kinds of predictions?

CONCEPT 1.1 Statistical predictability means "the proportion of times an event occurs" will converge, over the long run, to a specific value representing the likelihood the event will occur at any given time.

CONCEPT 1.2 Randomness is uncertainty that is statistically predictable.

Example 1.1 We say that a fair coin is equally likely to be Heads or Tails – a prediction that favors neither possibility. We also expect that if we flip the coin many times, close to $1/2$ of the results will be Heads.

If the coin is not fair, we expect the long term proportion of Heads to be some value $p \neq 1/2$. This value would give us the odds $(p : (1-p))$ in favor of Heads.

Example 1.2 I select a cell phone battery on a production line for inspection. I may say I chose a unit "at random" from the production lot; but merely having made a choice means it may not be random. No outside observer could be certain a statistical prediction was meaningful.

Likewise, getting bumped from behind at a stop light does not occur to you "at random".

**** Uncertainty that is haphazard or whimsical is not randomness.*

Chaotic behavior, on the other hand, can mimic randomness. In fact a random number generator actually is chaotic.

Example 1.3 The "logistic" difference equation or dynamical system is

$$x_{t+1} = 4x_t(1 - x_t), \quad \text{with } x_0 \in (0, 1).$$

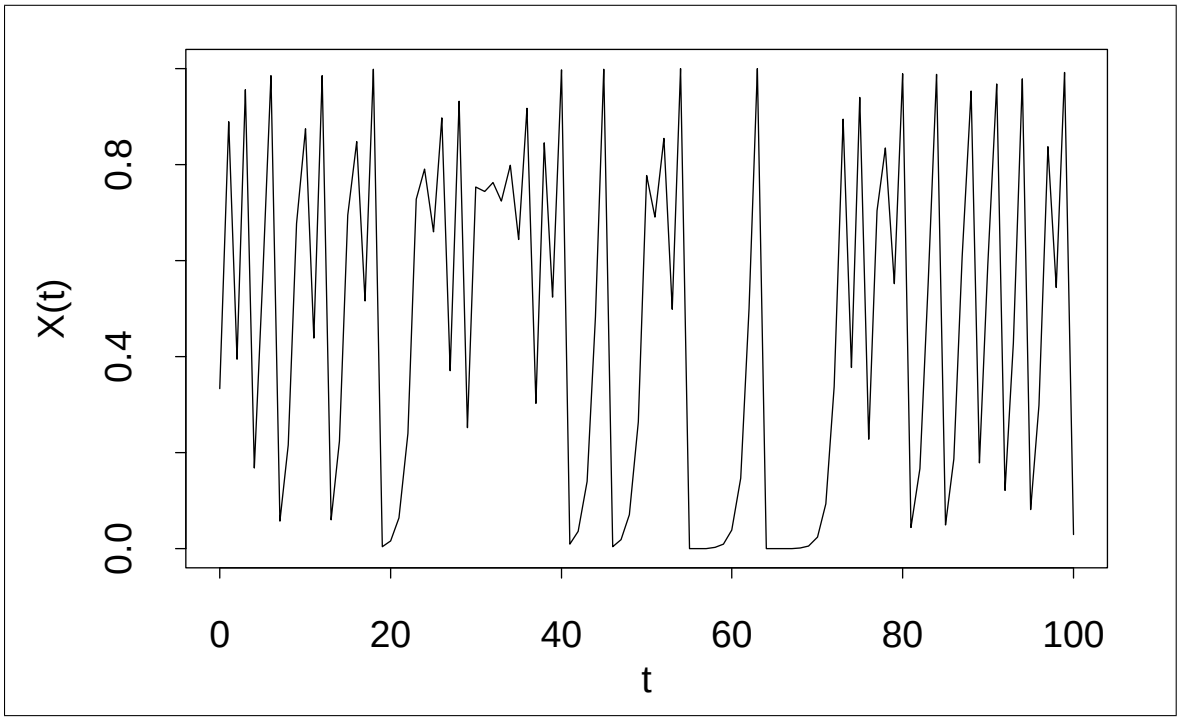


Figure 1.1 Logistic dynamical system, $t = 0, 1, \ldots, 100$, and starting at $1/3$.

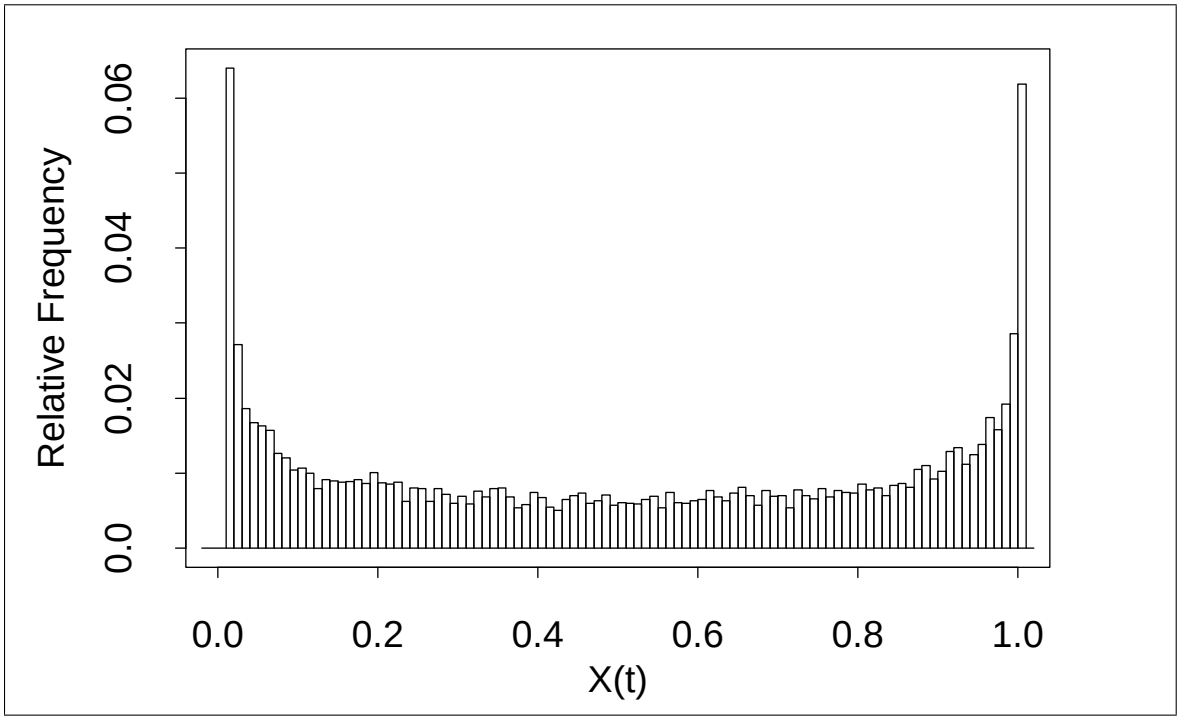


Figure 1.2 Histogram of 10,000 values of the logistic dynamical system.

This process, while completely determined by the initial condition x_0 behaves chaotically so that for large arbitrary t it is difficult to predict what x_t will be. Nevertheless, the sequence of values have a sort of statistical predictability so that if we were to assume we observe only one value from a sequence with an unknown initial condition a long time in the past, we might well think of that one value as random.

If we are collecting data for a scientific or business investigation, and our data are random, then we can use statistical predictability to our advantage.

CONCEPT 1.3 Qualified scientific conclusions require repeatable experiments, at least in concept. Statistical inference then means making conclusions from random data in such a way that one also can predict the quality (accuracy and reliability) of the conclusions.

Example 1.1 (cont.) Clearly, by flipping a coin a large number of times we can estimate the chance of Heads, and the more the flips the better the estimate. In fact, we can state clearly how likely the estimate is to be within, say, Δ of the true value.

In other words, statistical predictability gives us both the estimate and an idea of how much the estimate may be in error.

1.2 Sample Spaces and σ-algebras

We see that Concepts 1.1–1.3 implicitly require some kind of experimental context.

Definition 1.4

i. A random experiment is a well-defined, repeatable investigation in which *exactly one* of a set of possible outcomes is the experimental result, but just which outcome results is a matter of randomness.

ii. The set of possible outcomes is called the sample space and is often denoted $\mathcal{S}$.

iii. An event is any subset of $\mathcal{S}$, including $\mathcal{S}$ and $\emptyset$.

The sample space can be finite, countably infinite or uncountably infinite.

Example 1.1 (cont.)

Experiment 1: One coin flip. $\mathcal{S} = \{0, 1\}$, where 0 means Tails and 1 means Heads.

Experiment 2: n coin flips. Each outcome is a vector and

$$\mathcal{S} = \{s = (x_1, \ldots, x_n) : x_i \in \{0, 1\}\}.$$

$\mathcal{S}$ has 2^n outcomes. Possible events include

$$A = \text{“the first two flips are Heads”} = \{(x_1, \ldots, x_n) \in \mathcal{S} : x_1 = x_2 = 1\}$$

and

$$B = \text{“the total number of Heads is 12”} = \{(x_1, \ldots, x_n) \in \mathcal{S} : x_1 + \cdots + x_n = 12\}.$$

Experiment 3: Flip the coin until a Heads occurs. $\mathcal{S} = \{(1), (0, 1), (0, 0, 1), \ldots\}$ is countably infinite.

Experiment 4: An infinite sequence of flips.

$$\mathcal{S} = \{s = (x_1, x_2, \ldots) : x_i \in \{0, 1\}\}.$$

$\mathcal{S}$ is uncountably infinite. An important event is

$$\begin{aligned} C &= \text{“the limiting proportion of Heads is 1/2”} \\ &= \{(x_1, x_2, \ldots) \in \mathcal{S} : (x_1 + \cdots + x_n)/n \to 1/2\}. \end{aligned}$$

Events are subsets and adhere to standard set theory. Notation is the usual.

union: $A \cup B$, $\bigcup_{i=1}^{n} A_i$.
intersection: $A \cap B$ or AB, $\bigcap_{i=1}^{n} A_i$.
complement: A^c.

- We say "A occurs" to mean the actual experimental outcome is in A, i.e., that $s \in A$.
- "A occurs" *does not mean* every outcome in A occurs. The experiment yields exactly one outcome.
- It is important to note that events are *dynamic* because of the random outcome s. Repeating the experiment very likely will yield different outcomes.
- Indeed, it can be very helpful to express an event both verbally and in mathematical notation – as long as both are sufficiently precise.

Some useful events are

$$s \in \bigcup_n A_n \iff A_n \text{ occurs for some } n,$$

$$s \in \bigcap_n A_n \iff A_n \text{ occurs for all } n,$$

$$s \in \left(\bigcup_n A_n\right)^c \iff A_n \text{ occurs for no } n,$$

$$s \in \left(\bigcap_n A_n\right)^c \iff \text{some } A_n \text{ does not occur.}$$

Observe, moreover, (DeMorgan's rules)

$$\left(\bigcup_n A_n\right)^c = \bigcap_n A_n^c, \qquad \left(\bigcap_n A_n\right)^c = \bigcup_n A_n^c.$$

DEFINITION 1.5 A *collection* of events $\mathcal{A}$ is a σ-algebra if

i. $\mathcal{A}$ contains $\mathcal{S}$,

ii. $\mathcal{A}$ is closed under complement: $A \in \mathcal{A} \implies A^c \in \mathcal{A}$, and

iii. $\mathcal{A}$ is closed under countable unions:

$$A_1, A_2, \cdots \in \mathcal{A} \implies \bigcup_{n=1}^{\infty} A_n \in \mathcal{A}.$$

- Basically, $\mathcal{A}$ consists of all "observable" events. It will also be closed under countable intersections.
- If $\mathcal{S}$ is finite or countable then $\mathcal{A}$ could be all the events, but if $\mathcal{S}$ is uncountable then $\mathcal{A}$ must be restricted some (though not so much as to be a hindrance to us).
- $\mathcal{A}$ may also be restricted if we only observe partial information.

Example 1.1 (cont.)

Experiment 1: Flip a coin once. $\mathcal{S} = \{0, 1\}$ and $\mathcal{A} = \{\emptyset, \{0\}, \{1\}, \mathcal{S}\}$, which is all possible subsets.

Experiment 2: Flip a coin n times. But suppose we only take note of the number of heads. Then $\mathcal{A}$ consists only of the empty set, of events of the form $\{x_1 + \cdots + x_n = k\}$, $k = 0, 1, \ldots, n$, and of all unions of these events.

Experiment 3: Flip the coin until a Heads occurs. Outcomes consist of $n - 1$ 0's followed by a single 1, with $n \geq 1$ arbitrary. $\mathcal{S}$ is countably infinite and $\mathcal{A}$ can be all subsets of $\mathcal{S}$.

Experiment 4: Each outcome is an infinite sequence of flips. $\mathcal{A}$ cannot be all subsets (or else it is impossible to define a coherent probability model). But we could require it to at least contain all events that depend on a finite number of flips, subject to it being a σ-algebra.

Example 1.4 Suppose we measure the lifetime X of a cell phone battery under stress. Then we could have $\mathcal{S} = [0, \infty)$. (Unbounded since perhaps we do not really know what the longest lifetime could be.)

$\mathcal{A}$ would certainly have to include events of the form $[a, b] =$ "$a \leq X \leq b$", (a, b) $=$ "$a < X < b$", etc., indicating the lifetime falls within a specified interval. The smallest σ-algebra that includes all such events is called the Borel σ-algebra and it contains everything a statistician would consider.

Events represent the ways that we can express the observable information we observe from an experiment. They are much richer and more meaningful than simply recording the precise outcome.

**** Indeed, it is the events (not the outcomes) that we define probabilities for. This is why we have defined the σ-algebra collection of events.*

1.3 Axioms and Properties

We can now actually define probability.

DEFINITION 1.6 A probability space is $(\mathcal{S}, \mathcal{A}, \mathsf{P})$ where P is a probability measure on $\mathcal{A}$ (that is, a function from subsets of $\mathcal{S}$ to the real line) satisfying Kolmogorov's axioms:

i. $\mathsf{P}(A) \geq 0$,

ii. $\mathsf{P}(\mathcal{S}) = 1$, and

iii. if $A_1, A_2, \ldots$ are disjoint ($A_m \cap A_n = \emptyset$ for $m \neq n$) then (countable additivity)

$$\mathsf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathsf{P}(A_n).$$

These are called axioms because all the other properties of probability can be derived from them.

THEOREM 1.7 P is finitely additive: if $A_1, \ldots, A_n$ are disjoint then

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

PROOF First, by considering the sequence $\mathcal{S}, \emptyset, \emptyset, \ldots$, we see that

$$1 = \mathsf{P}\Big(\mathcal{S} \cup (\bigcup_{n=2}^{\infty} \emptyset)\Big) = \mathsf{P}(\mathcal{S}) + \sum_{n=2}^{\infty} \mathsf{P}(\emptyset).$$

That is, $\mathsf{P}(\emptyset)$ must be 0.

Now let $A_{n+1} = A_{n+2} = \cdots = \emptyset$. Then

$$\mathsf{P}\Big(\bigcup_{i=1}^{n} A_i\Big) = \mathsf{P}\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mathsf{P}(A_n) = \sum_{i=1}^{n} \mathsf{P}(A_i).$$

□

Some other simple consequences are:

COROLLARY 1.8

i. $\mathsf{P}(\emptyset) = 0$,

ii. $\mathsf{P}(A^c) = 1 - \mathsf{P}(A)$,

iii. $\mathsf{P}(A) \leq 1$,

iv. $\mathsf{P}(AB) + \mathsf{P}(AB^c) = \mathsf{P}(A)$,

v. $\mathsf{P}(A) + \mathsf{P}(B) = \mathsf{P}(AB) + \mathsf{P}(A \cup B)$,

vi. $A \subset B \implies \mathsf{P}(A) \leq \mathsf{P}(B)$.

PROOF These all follow from finite additivity. For example, v. is because

$$\mathsf{P}(A) + \mathsf{P}(B) = (\mathsf{P}(AB) + \mathsf{P}(AB^c)) + (\mathsf{P}(AB) + \mathsf{P}(A^cB))$$

and

$$\mathsf{P}(AB) + \mathsf{P}(A \cup B) = \mathsf{P}(AB) + (\mathsf{P}(AB) + \mathsf{P}(AB^c) + \mathsf{P}(A^cB)).$$

□

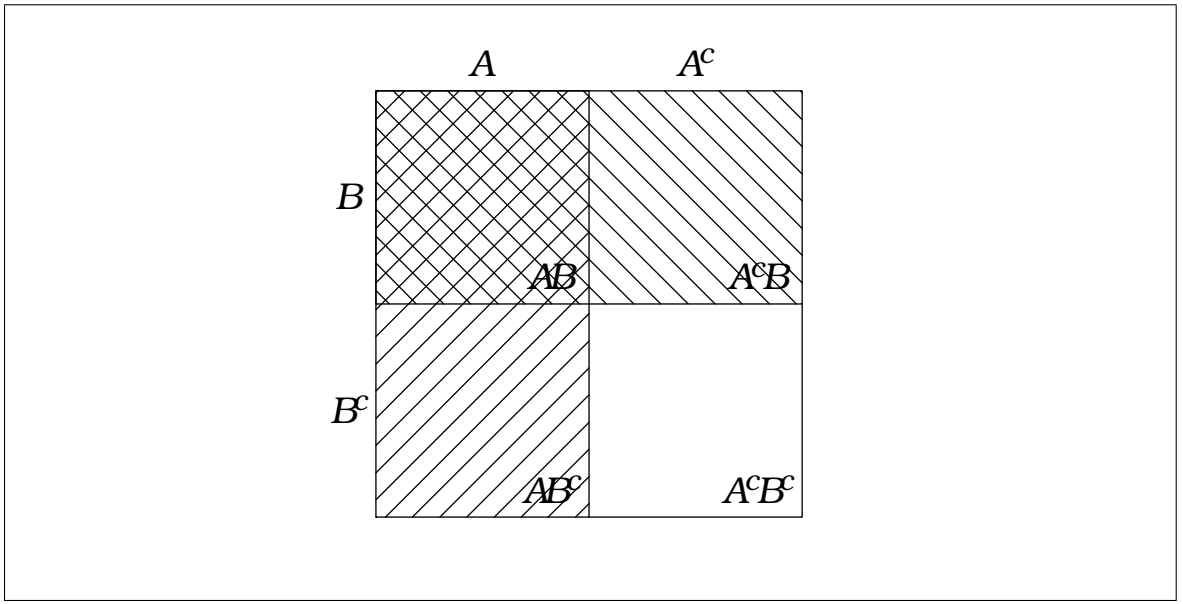


Figure 1.3 Venn diagram in tabular form.

Observe also (see the diagram above) that

$$\mathsf{P}(A \cup B) = 1 - \mathsf{P}(A^c \cap B^c),$$

which is proved formally by combining Property ii. with DeMorgan's rule.

**** Probabilities behave just like proportions. They, in some sense, measure the size of an event.*

THEOREM 1.9

i. (Boole) $\mathsf{P}(\bigcup_{i=1}^n A_i) \leq \sum_{i=1}^n \mathsf{P}(A_i)$.

ii. (Bonferroni) If $\mathsf{P}(A_i) \geq 1 - a_i$ then $\mathsf{P}(\bigcap_{i=1}^n A_i) \geq 1 - \sum_{i=1}^n a_i$.

PROOF i. Repeatedly apply Cor. 1.8.v:

$$\begin{aligned}\mathsf{P}\left(\bigcup_{i=1}^n A_i\right) &\leq \mathsf{P}(A_1) + \mathsf{P}\left(\bigcup_{i=2}^n A_i\right) \\ &\leq \cdots \leq \sum_{i=1}^n \mathsf{P}(A_i).\end{aligned}$$

ii. Apply DeMorgan's rule, Cor. 1.8.ii and Boole's inequality:

$$\begin{aligned}\mathsf{P}\left(\bigcap_{i=1}^n A_i\right) &= 1 - \mathsf{P}\left(\bigcup_{i=1}^n A_i^c\right) \\ &\geq 1 - \sum_{i=1}^n \mathsf{P}(A_i^c) \geq 1 - \sum_{i=1}^n a_i.\end{aligned}$$

□

While not as precise, inequalities such as the above can still be very useful.

Example 1.5 Suppose three statistical methods each have 99% chance of providing an accurate result (99% confidence). Then the chance all three are accurate is at least 97%.

**** Sometimes it is more convenient to consider the probability of the complement.*

Example 1.1 (cont.) Flip a fair coin 5 times. $\mathcal{S}$ has $2^5 = 32$ equally likely outcomes. The chance of at least one Head is

$$\mathsf{P}(x_1 + \cdots + x_5 > 0) = 1 - \mathsf{P}(\{(0,0,0,0,0)\}) = 1 - \tfrac{1}{32} = \tfrac{31}{32}.$$

THEOREM 1.10 Suppose $\mathcal{S}$ is either finite or countable and $A \subset \mathcal{S}$. Then

$$\mathsf{P}(A) = \sum_{n:s_n \in A} \mathsf{P}(\{s_n\}).$$

PROOF By countable additivity, since $A = \bigcup_{n:s_n \in A} \{s_n\}$ which is a union of disjoint events. □

Example 1.6 Flip a fair coin until Heads and observe only the number of flips, denoting $\mathcal{S} = \{1, 2, \dots\}$. The chance the number of flips is even would be

$$\mathsf{P}(\text{“even number of flips”}) = \mathsf{P}(\{2\}) + \mathsf{P}(\{4\}) + \cdots = 1/3.$$

Observe the notation for singleton events. The *value* 2 is not random and has no probability. However, we can (and will) use descriptive shorthand such as writing $\mathsf{P}(s = 2)$ or $\mathsf{P}(s \geq 2)$ for $\mathsf{P}(\{2\})$ or $\mathsf{P}(\{2, 3, \dots\})$, respectively.

**** If $\mathcal{S}$ is countable we need only to know the probability of each outcome. But this is not true if $\mathcal{S}$ is uncountable.*

For uncountable $\mathcal{S}$ the probability calculus is not so intuitive. This is another reason why we deal with probabilities of events.

Example 1.4 (cont.) Measure $X =$ lifetime of a cell phone battery. We can discuss $\mathsf{P}(10 < X \leq 20)$, but we certainly *cannot* get $\mathsf{P}(10 < X \leq 20)$ by adding up probabilities for the uncountable number of possibilities between 10 and 20.

Again note the shorthand: "$10 < X \leq 20$" is *formally* the event $\{s : 10 < X(s) \leq 20\}$, where $X(s)$ is the lifetime for outcome s.

And what is $\mathsf{P}(X$ exactly equals 15$)$? Suppose, in particular, that we model $\mathsf{P}(X > t) = \mathrm{e}^{-t/10}$ for all $t > 0$. Then

$$\mathsf{P}(10 < X \leq 20) = \mathsf{P}(X > 10) - \mathsf{P}(X > 20) = \mathrm{e}^{-1} - \mathrm{e}^{-2},$$

and

$$\mathsf{P}(X = 15) = \mathsf{P}(X \geq 15) - \mathsf{P}(X > 15) = \lim_{t \uparrow 15} \mathsf{P}(X > t) - \mathsf{P}(X > 15) = 0.$$

$\mathsf{P}(X = 15) = 0$ *does not* mean 15 is not a possible outcome. Rather, it is a consequence of the uncountable number of outcomes. In practice, X could be *rounded* to 15 (e.g., $14.95 \leq X < 15.05$) which is an event with positive probability.

Another frequently used result is the following. We saw a simple version in Cor. 1.8.iv.

THEOREM 1.11 Let $B_1, B_2, \ldots$ be a partition for $\mathcal{S}$: $B_1, B_2, \ldots$ are disjoint and their union is $\mathcal{S}$. Then, for any event A, $\mathsf{P}(A) = \sum_n \mathsf{P}(A \cap B_n)$.

PROOF $A = \bigcup_n (A \cap B_n)$ and $A \cap B_1, A \cap B_2, \ldots$ are also disjoint, so the result follows by countable (or finite) additivity. □

**** We will see that partitioning one event (or value) according to another frequently reduces the work in a calculation.*

Example 1.7 Roll a fair die until a 6 appears and stop. What is the probability that a 1 does not appear?

Solution Let $A =$ "no 1" and $B_n =$ "number of rolls is n". The B_n's partition $\mathcal{S}$. Since $A \cap B_n =$ "$n-1$ rolls of 2, 3, 4 or 5 followed by one roll of 6", we get

$$\mathsf{P}(A \cap B_n) = \frac{4 \times 4 \times \cdots \times 4 \times 1}{6^n} = \frac{1}{6}\left(\frac{2}{3}\right)^{n-1},$$

using the product rule in the next section. Thus

$$\mathsf{P}(A) = \sum_{n=1}^{\infty} \mathsf{P}(A \cap B_n) = \sum_{n=1}^{\infty} \frac{1}{6}\left(\frac{2}{3}\right)^{n-1} = \frac{1}{6}\left(\frac{1}{1-2/3}\right) = \frac{1}{2}.$$

Intuition says a 1 is not more likely to appear before a 6 than a 6 is to appear before a 1, and vice versa. So the probability of A should be $1/2$.

Mere intuition is not a proof. However, in this example we can observe that all the relevant probabilities would not change if we were to relabel the sides of the die, such as $1 \leftrightarrow 6$. Thus "1 before 6" and "6 before 1" would have the same probability.

As long as such arguments are done *coherently*, they can greatly simplify computations.

**** It is also very useful to compute probabilities as limits.*

THEOREM 1.12 (Continuity)

i. Suppose $A_1 \subset A_2 \subset \cdots$ (increasing sequence of events) and $A = \bigcup_{n=1}^{\infty} A_n$. Then $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n)$.

ii. Suppose $A_1 \supset A_2 \supset \cdots$ (decreasing sequence of events) and $A = \bigcap_{n=1}^{\infty} A_n$. Then $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n)$.

This theorem applies to *monotone* sequences of events. The general concept of set limits is beyond our needs (and is not as simple as one might imagine).

PROOF

i. Let $B_1 = A_1$ and $B_n = A_n \cap \left(\bigcup_{i=1}^{n-1} A_i\right)^c$ for $n > 1$. Note that the B_n's are disjoint and $A_n = \bigcup_{i=1}^{n} B_i$. Thus

$$\lim_{n\to\infty} \mathsf{P}(A_n) = \lim_{n\to\infty} \mathsf{P}\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathsf{P}(B_i) = \sum_{i=1}^{\infty} \mathsf{P}(B_i).$$

But $s \in$ some $A_n \iff s \in$ some B_n so $A = \bigcup_{i=1}^{\infty} B_i$ and

$$\mathsf{P}(A) = \mathsf{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathsf{P}(B_i).$$

ii. (exercise).

□

Example 1.7 (cont.) What is the probability a 6 never appears?

Solution Let $A_n =$ "no 6 in the first n rolls". Then $A_1 \supset A_2 \supset \cdots$, $A = \bigcap_{n=1}^{\infty} A_n =$ "no 6's ever" and $\mathsf{P}(A) = \lim_{n\to\infty} \mathsf{P}(A_n) = \lim_{n\to\infty} 5^n/6^n = 0$.

But A is not the empty set. It is not even trivial. (Why not?)

1.4 Counting Rules

Classical probability and the ideas of *random sampling* depend on being able to count the number of ways a selection can be made.

THEOREM 1.13 (Counting Selections)

i. (Product Rule) If for each $i = 1, \ldots, k$, selection $\#i$ has n_i possibilities irrespective of the previous selections, then the total number of possibilities is $\prod_{i=1}^{k} n_i$.

ii. (Ordering Rule) There are $n!$ ways to order n items.

iii. (Permutations Rule) The number of ordered (order identified) selections of n items from N is $P_{N,n} = N \times (N-1) \times \cdots \times (N-n+1) = \frac{N!}{(N-n)!}$.

iv. (Combinations Rule) The number of subsets (unordered selections) consisting of n items selected from N items is $C_{N,n} = \binom{N}{n} = \frac{N!}{n!(N-n)!}$.

Note: in the product rule it is the *number* of possibilities that does not depend on previous selections. The possibilities themselves could.

Proof

i. By observation: consider the case $k = 2$ and then extend by induction.

ii. There are n possible first choices and regardless of that choice there are $n-1$ possible second choices, continuing down to the single possibility for the last choice. Now apply the product rule.

iii. Similar to ii (exercise).

iv. An ordered selection can be thought of as first making an unordered selection followed by ordering the n items. Hence $P_{N,n} = C_{N,n} \times n!$.

□

By the way, recall that $0!$ *is defined to take value 1*. While that may seem strange, it is a necessary definition.

Example 1.8 (Birthday Problem) What is the probability p_k that at least two people in a group of size k have the same birthday? (Suppose we exclude leap days and the 365 birthdays are equally likely.)

Solution The sample space has 365^k possible vectors of birthdays. But there are $P_{365,k}$ ways to select k distinct birthdays.

So the probability no two people in the group have the same birthday is $\frac{P_{365,k}}{365^k}$. Thus, the probability at least 2 have the same birthday is $p_k = 1 - \frac{P_{365,k}}{365^k}$.

k	p_k	k	p_k
10	.117	25	.569
15	.253	30	.706
20	.411	40	.891
21	.444	50	.970
22	.476	60	.994
23	.507		

[*birthday.r*]

Example 1.9 (Quality Inspection) Imagine a production lot of N cell phone batteries from which we take a random sample of $n \ll N$ to inspect. The lot has an (unknown) defective rate of $p = M/N$. If we observe X defective batteries in the sample, we can estimate p with the sample rate, $\widehat{p} = X/n$. What can we predict about the values of this estimate (or, equivalently, about X)?

Solution The sample space consists of $\binom{N}{n}$ equally likely possible samples (ignoring order). This is Sampling Without Replacement (SWOR).

As there are $\binom{M}{x}$ ways to choose x defectives from among the M defectives, and $\binom{N-M}{n-x}$ ways to select the non-defectives,

$$\mathsf{P}(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad 0 \le x \le n, n - N + M \le x \le M.$$

To determine when this probability is largest, note that

$$\begin{aligned} \frac{\mathsf{P}(X=x-1)}{\mathsf{P}(X=x)} &= \frac{x!(M-x)!(n-x)!(N-M-n+x)!}{(x-1)!(M-x+1)!(n-x+1)!(N-M-n+x-1)!} \\ &= \frac{x(N-M-n+x)}{(M-x+1)(n-x+1)} = 1 + \frac{(N+2)x-(M+1)(n+1)}{(M-x+1)(n-x+1)}. \end{aligned}$$

Thus $\mathsf{P}(X = x) > \mathsf{P}(X = x-1) \iff x < \frac{(M+1)(n+1)}{N+2}$ and the most likely value for $\widehat{p}$ is $\frac{1}{n}\lfloor\frac{(M+1)(n+1)}{N+2}\rfloor$, which is approximately $\frac{M}{N} = p$.

Example 1.10 (Independent Polling) A pollster calls n voters, selecting each at random from the entire population of N voters, irrespective of previous selections.

Suppose $p = M/N$ is the proportion in the population that would answer Yes to the pollster. Let X be the number of Yes responses in the sample.

Again, the pollster would estimate p with $\widehat{p} = X/n$. How do we predict X now?

Solution By the product rule there are N^n equally likely possible samples (ordered). This is Sampling With Replacement (SWR). The number of samples for which there are x Yes responses (and $n - x$ No's) is

$$M^x \times (N - M)^{n-x} \times \#\text{ ways to place } x \text{ Yes's among } n \text{ responses.}$$

Thus,

$$\mathsf{P}(X = x) = \frac{\binom{n}{x} M^x (N - M)^{n-x}}{N^n} = \binom{n}{x} p^x (1 - p)^{n-x}, \quad x = 0, 1, \ldots, n.$$

While the sample space takes order of selection into account, the event "$X = x$" and its probability do not require knowing the order.

Observe that any other probability that is solely about X can be computed using the expression in the previous line.

The pollster hopes to estimate p to within a small value Δ. The chance of this is

$$\mathsf{P}(n(p-\Delta) \leq X \leq n(p+\Delta)) = \sum_{n(p-\Delta)\leq x\leq n(p+\Delta)} \binom{n}{x} p^x (1-p)^{n-x}.$$

For example, if $n = 1000$, $p = .4$ and $\Delta = .03$ then

$$\mathsf{P}(370 \leq X \leq 430) = \sum_{370\leq x\leq 430} \binom{n}{x} p^x (1-p)^{n-x} = .95108.$$

(Technically, actual polling methods are much more complicated. But this example does give a relatively good idea.)

Example 1.11 (Multiple Categories) Suppose N individuals have k "traits" or are otherwise assigned to k categories. Let M_i be the number of individuals in category i, $i = 1, \ldots, k$ (so $M_1 + \cdots + M_k = N$).

Now imagine randomly sampling n from this population and observing how many are from each category. Then, once again using the product rule, we can determine that

$$\begin{aligned}
&\text{P}(\text{exactly } x_i \text{ are from category } i, \text{ for each } i) \\
&\quad = \begin{cases} \dfrac{\binom{M_1}{x_1} \times \cdots \times \binom{M_k}{x_k}}{\binom{N}{n}} & \text{sampling without replacement,} \\ \dfrac{N!}{x_1! \cdots x_k!} \left(\dfrac{M_1}{N}\right)^{x_i} \times \cdots \times \left(\dfrac{M_k}{N}\right)^{x_k} & \text{sampling with replacement,} \end{cases}
\end{aligned}$$

where $x_1 + \cdots + x_k = n$.

1.5 Conditional Probability and Bayes' Theorem

We come now to our first look at one of the most important ideas of probability.

Example 1.12 (Medical Testing) A hidden disease D inhabits 40% of the population. A medical test is available but it has a 5% false positive rate and a 20% false negative rate. Just what does this mean?

Let the population size be N, so $.4N$ individuals are diseased. Of those, 20% or $.08N$ will fail to show a positive response "P" to the test – they have a false negative response. Thus P("D and not P") $= \frac{.08N}{N} = .08$.

Note that this rate is relative to the entire population, while the false negative rate of 20% is relative only to the diseased subpopulation. To interpret the 20% we essentially focus only on those who are diseased and treat them as a given whole.

Note also that the calculation does not actually depend on the value of N: $.20 = \frac{.08}{.40}$.

This example leads to the following.

DEFINITION 1.14 Let A and B be events. The conditional probability of A, *given* B, is $\mathsf{P}(A|B) = \frac{\mathsf{P}(A \cap B)}{\mathsf{P}(B)}$, if $\mathsf{P}(B) > 0$.

Example 1.12 (cont.) So the false negative rate (20%) is $\mathsf{P}(P^c|D)$, the conditional probability of "not positive", given "diseased".

Likewise, the false positive rate is the conditional probability of "positive", given "not diseased":

$$.05 = \mathsf{P}(P|D^c) = \frac{\mathsf{P}(P \cap D^c)}{\mathsf{P}(D^c)}.$$

We can then calculate $\mathsf{P}(P \cap D^c) = .05(1 - .40) = .03$.

THEOREM 1.15 (Multiplication Rule)

i. $\mathsf{P}(A \cap B) = \mathsf{P}(B)\mathsf{P}(A|B)$ (with the obvious interpretation $\mathsf{P}(A \cap B) = 0$ if $\mathsf{P}(B) = 0$).

ii. $\mathsf{P}(A_1 \cap A_2 \cap \cdots \cap A_n) = \mathsf{P}(A_1)\mathsf{P}(A_2|A_1) \cdots \mathsf{P}(A_n|A_1 \cap \cdots \cap A_{n-1})$.

PROOF i. follows directly from Def. 1.14, and ii. follows by iterating i. □

Example 1.12 (cont.) Suppose the actual disease rate is unknown, but doctors are confident of the false negative and false positive rates and they know that 35% of the population tests positive. We can then "condition" on disease status:

$$.35 = \mathsf{P}(P) = \mathsf{P}(P \cap D) + \mathsf{P}(P \cap D^c) = \mathsf{P}(D)\mathsf{P}(P|D) + \mathsf{P}(D^c)\mathsf{P}(P|D^c)$$
$$= (1 - .2)\mathsf{P}(D) + .05(1 - \mathsf{P}(D)).$$

Hence $\mathsf{P}(D) = .40$.

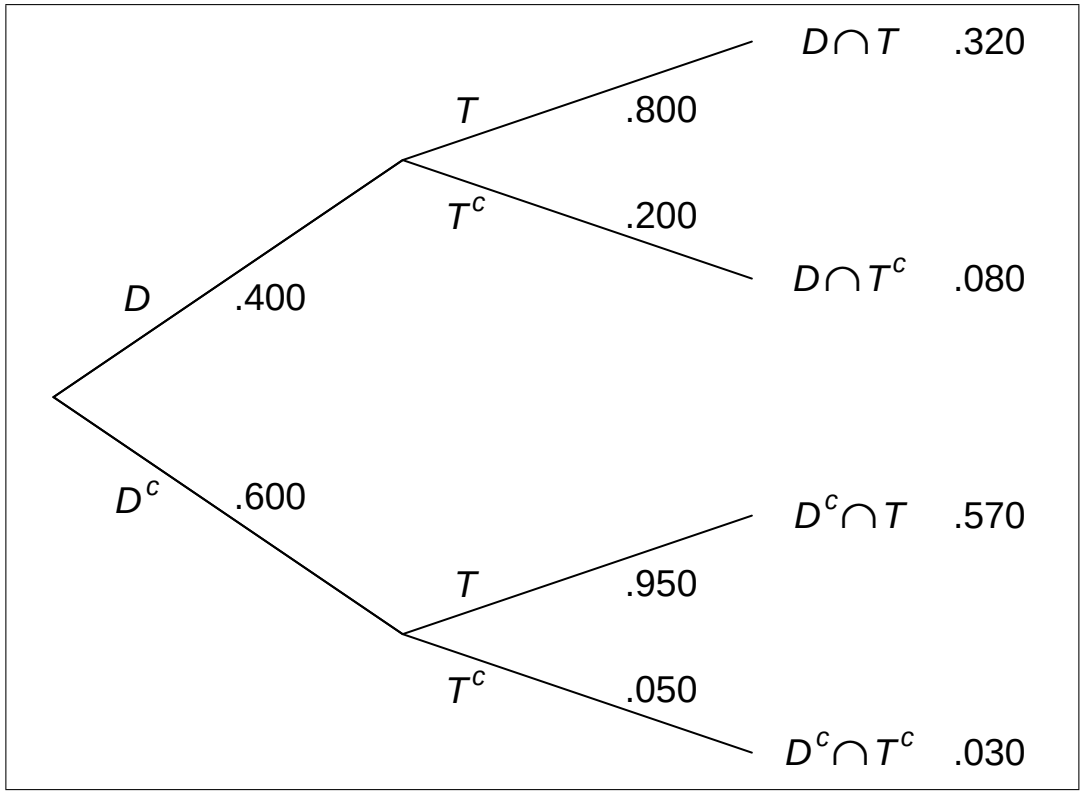


Figure 1.4 Tree diagram.

COROLLARY 1.16 Assume $\mathsf{P}(B) > 0$.

i. $\mathsf{P}(B|B) = 1$.

ii. $A \subset B \implies \mathsf{P}(A|B) = \frac{\mathsf{P}(A)}{\mathsf{P}(B)}$.

iii. $A \cap B = \emptyset \implies \mathsf{P}(A|B) = 0$.

THEOREM 1.17 Suppose $B_1, B_2, \ldots$ is a partition of $\mathcal{S}$.

i. (Total Probability) $\mathsf{P}(A) = \sum_n \mathsf{P}(A|B_n)\mathsf{P}(B_n)$.

ii. (Bayes' Rule) $\mathsf{P}(B_i|A) = \frac{\mathsf{P}(A|B_i)\mathsf{P}(B_i)}{\sum_n \mathsf{P}(A|B_n)\mathsf{P}(B_n)}$ for each i.

PROOF

i. Use the partitioning rule (Thm. 1.11) and the multiplication rule (Thm. 1.15.i).

ii. Use the law of total probability (Thm. 1.17.i) and the definition of conditional probability (Def. 1.14).

□

**** It is easier to remember the derivation of Bayes' rule, which are the computation steps anyway, than to simply memorize the formula.*

Example 1.13 (Two-State Markov Chain) Suppose the signal a radio receives has two states: 1 = "Clear", meaning a clear signal is received, and 0 = "Fade", meaning the signal is too weak. Let X_i be the value of the state at time i, where $i \geq 0$.

Now suppose α is the chance the state changes from 1 to 0, and β is the chance it changes from 0 to 1. Let $p_i = \mathsf{P}(X_i = 1)$. Then

$$\begin{aligned} p_i &= \mathsf{P}(X_i = 1|X_{i-1} = 0)\mathsf{P}(X_{i-1} = 0) + \mathsf{P}(X_i = 1|X_{i-1} = 1)\mathsf{P}(X_{i-1} = 1) \\ &= \beta(1 - p_{i-1}) + (1 - \alpha)p_{i-1}. \end{aligned}$$

This gives a recursive way to calculate the probabilities.

If the process is in equilibrium then $p_i = p_{i-1}$ and we get $p_i = \frac{\beta}{\alpha+\beta}$ for all i.

Suppose the process is "Clear" at time i. What is the chance it also was "Clear" at time $i - 1$? By Bayes' rule,

$$\mathsf{P}(X_{i-1} = 1|X_i = 1) = \frac{(1 - \alpha)p_{i-1}}{\beta(1 - p_{i-1}) + (1 - \alpha)p_{i-1}}.$$

In equilibrium, this is $1 - \alpha$.

**** Conditional probabilities do not usually announce themselves as such. This is why sometimes it seems appropriate to multiply probabilities and other times it is not.*

Example 1.14 Suppose 60% of marriages have children, 40% of marriages end in divorce and 20% of marriages with children end in divorce. (These numbers are completely made up!)

So do $24\% = .40 \times 60\%$ of marriages divorce with children?

The answer is no, because only 20% of marriages with children end in divorce. The correct value is $.20 \times 60\% = 12\%$.

On the other hand, $70\% = (.40 - .12)/(1 - .60)$ of marriages without children end in divorce.

**** Conditional probabilities are useful for updating a probability assessment as a result of partial information.*

Example 1.15 You roll two fair, but indistinguishable, dice hoping for a total of 7 (probability 1/6). After the roll one die is hidden, but you see that the other is a 4. What is the chance you have a total of 7?

Solution If you had rolled just one die and seen it was a 4, then to get a 7 you would have to roll a 3 with the second die; the chance again is 1/6. But in this case you do not know which die is hidden and which one you see.

There are 11 ways that one die could be a 4, two of which have the other die a 3. So the updated chance for "total is 7" is 2/11.

Note that this would be the new assessment for "total is 7", no matter what value you saw on the die! However, other assessments would change depending on which value was observed.

(Exercise. Find the probability the total is 5, given at least one die is 4. Compare this to the unconditional probability the total is 5 and to the probability the total is 5, given the first die is 4.)

THEOREM 1.18 Suppose $\mathsf{P}(B) > 0$ and define $\mathsf{P}^B(A) = \mathsf{P}(A|B)$. Then P^B is a valid probability measure.

PROOF P^B satisfies Kolmogorov's axioms (Def. 1.6). In particular, if $A_1, A_2, \ldots$ are disjoint then

$$\mathsf{P}^B\Big(\bigcup_n A_n\Big) = \frac{\mathsf{P}(\bigcup_n (A_n \cap B))}{\mathsf{P}(B)} = \frac{\sum_n \mathsf{P}(A_n \cap B)}{\mathsf{P}(B)} = \sum_n \mathsf{P}^B(A_n).$$

□

**** In other words, if the condition B is fixed throughout we can compute with $\mathsf{P}(\cdot|B)$ as we would with ordinary probability.*

For example,

$$\mathsf{P}(A^c|B) = 1 - \mathsf{P}(A|B),$$

$$\mathsf{P}(A \cup C|B) = \mathsf{P}(A|B) + \mathsf{P}(C|B) - \mathsf{P}(A \cap C|B),$$

etc. But, as in Ex. 1.15, the rules of probability only work right if we are always conditioning on the same event B throughout.

Example 1.7 (cont.) Roll a fair die until a 6 appears. Given that a 6 appears on the $(n+1)^{th}$ roll, what is the distribution (of probabilities) for the number of 1's that have appeared prior to the 6?

Solution On the condition a 6 first appears on roll $n+1$, we have an experiment for which there are n selections with replacement from $\{1,2,3,4,5\}$ ($N=5$, $M=1$). We can calculate as if this defines the sample space.

So by Ex. 1.10 we can state

$$\mathrm{P}(\text{“1 appears } x \text{ times before the first 6”} \,|\, \text{“6 appears first on } (n+1)^{th} \text{ roll”}) = \binom{n}{x}.2^x.8^{n-x}, \quad x = 0, 1, \ldots, n.$$

Example 1.16 Two professors have learned that if either forgets her umbrella then the chance of rain is 2/3, but if neither do then the chance of rain is only 1/3.

Suppose each forgets her umbrella half the time and they forget simultaneously one quarter of the time. What can we say about the chance of rain if they both forget? (Hint: it does not have to be 2/3.)

Solution Let $A_i =$ "i professors forget umbrella", $i = 0, 1, 2$, and let $R =$ "rain". We therefore have, by assumption, (a probability table may help here)

$$\mathsf{P}(A_0) = \mathsf{P}(A_2) = 1/4,\ \mathsf{P}(A_1) = 1/2,$$

and

$$\mathsf{P}(R|A_0) = 1/3,\ \mathsf{P}(R|A_1 \cup A_2) = 2/3.$$

Since A_1 and A_2 are disjoint, we can see that

$$\mathsf{P}(R \cap A_1) + \mathsf{P}(R \cap A_2) = \mathsf{P}(R \cap (A_1 \cup A_2)) = \mathsf{P}(A_1 \cup A_2)\mathsf{P}(R|A_1 \cup A_2) = 1/2.$$

We also have the restriction

$$0 \leq \mathsf{P}(R \cap A_2) \leq \mathsf{P}(A_2) = 1/4.$$

So $\mathsf{P}(R|A_2)$ can be anything from 0 to 1. It does not have to be 2/3.

We can also see that

$$\frac{1}{2}\mathsf{P}(R|A_1) + \frac{1}{4}\mathsf{P}(R|A_2) = 1/2,$$

so $1/2 \leq \mathsf{P}(R|A_1) \leq 1$.

The point is that, even though "both forget" implies "at least one forgets", they are not the same condition and therefore conditional probabilities need not be the same. Similarly, "exactly one forgets" also implies "at least one forgets" and, again, conditioning on them is not the same.

Exercise. What would it take for $\mathsf{P}(R|A_2) = \mathsf{P}(R|A_1 \cup A_2)$? For $\mathsf{P}(R|A_1) = \mathsf{P}(R|A_1 \cup A_2)$?

Here is another justification of the idea that conditional probability can be used to predict, based on partial information.

Example 1.12 (cont.) Suppose you will be tested for the disease. You know that if you have the disease it will cost 1 (= \$1 million), otherwise the cost will be 0, and you'd like to predict the cost, based on the test results.

For notation, let $A_1 =$ "have disease", $A_0 = A_1^c$, $B_1 =$ "test positive", $B_0 = B_1^c$. Let c_j be the predicted cost if B_j occurs, $j = 0, 1$. How should we choose c_0 and c_1?

Solution One option is to choose them to minimize the probability weighted squared prediction error:

$$\sum_{i=0}^{1}\sum_{j=0}^{1}(c_j - i)^2\mathsf{P}(A_i \cap B_j) = \sum_{j=0}^{1}(c_j^2\mathsf{P}(A_0 \cap B_j) + (c_j - 1)^2\mathsf{P}(A_1 \cap B_j)).$$

This is a quadratic in c_j. Taking the derivative (with respect to c_j) and setting it equal to 0,

$$2c_j\mathsf{P}(A_0 \cap B_j) + 2(c_j - 1)\mathsf{P}(A_1 \cap B_j) = 0.$$

This yields

$$c_j = \frac{\mathsf{P}(A_1 \cap B_j)}{\mathsf{P}(A_1 \cap B_j) + \mathsf{P}(A_0 \cap B_j)} = \mathsf{P}(A_1|B_j).$$

So the optimal prediction is just the conditional probability.

From the assumed probabilities of the example, the values are $c_0 = .08/.65 \doteq .123$ and $c_1 = .32/.35 \doteq .914$. With no test results, the prediction is $\mathsf{P}(A_1) = .400$. The predicted cost clearly depends on the test results.

1.6 Independence

If investigating the dependence between events or variables is the essence of statistics, then modeling them by starting from simple, independent terms is a major objective of probability.

To begin with, we consider the special case when the conditional prediction of an event A does not depend on whether or not B occurs.

DEFINITION 1.19 A and B are independent if $\mathsf{P}(A \cap B) = \mathsf{P}(A)\mathsf{P}(B)$. Otherwise, we say A and B are dependent.

The definition for independence applies even if $\mathsf{P}(B) = 0$ or $\mathsf{P}(B) = 1$. In those cases B is independent of *every* event, including itself.

THEOREM 1.20

i. A and B are independent iff A and B^c are independent.

ii. Suppose $0 < \mathsf{P}(B) < 1$. A and B are independent iff $\mathsf{P}(A|B) = \mathsf{P}(A)$ (in which case $\mathsf{P}(A|B^c) = \mathsf{P}(A)$ also).

Proof

i. Independence of A and B implies

$$\mathsf{P}(A \cap B^c) = \mathsf{P}(A) - \mathsf{P}(A \cap B) = \mathsf{P}(A)(1 - \mathsf{P}(B)) = \mathsf{P}(A)\mathsf{P}(B^c),$$

which says A and B^c are independent. The converse follows by switching B with B^c.

ii. Independence holds if and only if

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

Applying i, we then also have that A and B are independent if and only if $\mathsf{P}(A|B^c) = \mathsf{P}(A)$.

□

Example 1.15 (cont.) Consider rolling 2 fair dice. Let $A_i =$ "1^{st} die is i" and $B_j =$ "2^{nd} die is j". The sample space consists of 36 equally likely outcomes, so

$$\mathsf{P}(A_i) = \frac{6}{36} = \frac{1}{6}, \quad \mathsf{P}(B_j) = \frac{6}{36} = \frac{1}{6}$$

and $\mathsf{P}(A_i \cap B_j) = 1/36$. Since $\mathsf{P}(A_i \cap B_j) = \mathsf{P}(A_i)\mathsf{P}(B_j)$, we see A_i and B_j are independent. Since this is true for any i and j, we can say *the two dice are independent*.

Example 1.9 (cont.) Assume M defectives in a lot of N and select two without replacement. Let $A_i =$ "i^{th} selection is defective", $i = 1, 2$. There are $N(N-1)$ possibilities (ordered) so that

$$\mathsf{P}(A_1) = \frac{M(N-1)}{N(N-1)} = \frac{M}{N}$$

and

$$\mathsf{P}(A_2) = \frac{M(M-1) + (N-M)M}{N(N-1)} = \frac{M}{N}$$

while

$$\mathsf{P}(A_1 \cap A_2) = \frac{M(M-1)}{N(N-1)} \neq \mathsf{P}(A_1)\mathsf{P}(A_2).$$

So, as expected, the first and second selections are not independent.

Example 1.10 (cont.) *(Polling)* Suppose there are N voters in the population, M of whom would respond Yes. If you sample *with* replacement then the two responses are independent. (exercise).

**** Independence is a specialized relationship and is not to be assumed lightly.*

DEFINITION 1.21 Events $A_1, A_2, \ldots, A_n$ are mutually independent if for *every subcollection* $A_{i_1}, \ldots, A_{i_k}$,

$$\mathsf{P}(A_{i_1} \cap \cdots \cap A_{i_k}) = \prod_{j=1}^{k} \mathsf{P}(A_{i_j}). \quad (2^k - k - 1 \text{ equalities})$$

The definition says this equality must hold for *any* sub-collection, not just for all n events, and not just for all pairs of events (which would be pairwise independence).

However, the definition also implies that each sub-collection is mutually independent.

Example 1.10 (cont.) Sample n voters with replacement. The n responses are independent. (exercise).

Clearly, knowing if A occurs is equivalent to knowing if A^c occurs. Thus, independence of events should encompass their complements.

THEOREM 1.22 Suppose $A_1, \ldots, A_n$ are mutually independent events. Let $A_{i_1}, \ldots, A_{i_k}$ and $A_{j_1}, \ldots, A_{j_m}$ be two subcollections such that $\{i_1, \ldots, i_k\}$ and $\{j_1, \ldots, j_m\}$ are disjoint. Then $A_{i_1}, \ldots, A_{i_k}, A_{j_1}^c, \ldots, A_{j_m}^c$ are independent and

$$\mathsf{P}(A_{i_1} \cap \cdots \cap A_{i_k} \cap A_{j_1}^c \cap \cdots \cap A_{j_m}^c) = \prod_{s=1}^{k} \mathsf{P}(A_{i_s}) \prod_{t=1}^{m} (1 - \mathsf{P}(A_{j_t})),$$

with the obvious interpretation if $k = 0$ or $m = 0$.

PROOF For independent events $B_1, \ldots, B_m$,

$$\begin{aligned}\mathsf{P}(B_1 \cap \cdots \cap B_{m-1} B_m^c) &= \mathsf{P}(B_1 \cap \cdots \cap B_{m-1}) - \mathsf{P}(B_1 \cap \cdots \cap B_{m-1} B_m) \\ &= \prod_{i=1}^{m-1} \mathsf{P}(B_i) - \prod_{i=1}^{m} \mathsf{P}(B_i) = \prod_{i=1}^{m-1} \mathsf{P}(B_i) \times (1 - \mathsf{P}(B_m)).\end{aligned}$$

$B_1, \ldots, B_m$ could be any subcollection of independent events so we can conclude, for example, that switching A_{j_1} to $A_{j_1}^c$ retains the independence. We just apply this iteratively, attaching another $A_{j_t}^c$ at each step. □

Example 1.17 Suppose $A_1, \ldots, A_n$ are *independent* events, each with the *same* probability p. Let $B_k =$ "exactly k of the A_i's occur".

Given a particular choice of which k A_i's occur (and which do not), Thm. 1.22 indicates a probability of $p^k(1-p)^{n-k}$.

Also, there are $\binom{n}{k}$ ways to select k of the n events to be the ones that occur.

Therefore,

$$\mathsf{P}(B_k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

Compare this with the probability that exactly k of n voters (sampled with replacement) will respond Yes (Ex. 1.10). Here, however, there is no sampling frame and we did not use counting to get the probability.

Example 1.18 Consider flipping an unfair coin with $p = \mathrm{P}(\text{Heads on } n^{th} \text{ flip})$, $0 < p < 1$. We cannot compute by counting now. Assume the flips are independent. How long until a Head is observed? until the k^{th} Head is observed?

Solution If the first Head is observed on flip n, there must be $n-1$ Tails beforehand. By the independence,

$$\mathrm{P}(\text{“}1^{st} \text{ Head is on } n^{th} \text{ flip”}) = p(1-p)^{n-1}, \quad n = 1, 2, \ldots$$

If the k^{th} Head is on flip n, there are $n-k$ Tails and $k-1$ other Heads placed among the previous $n-1$ flips. There are $\binom{n-1}{k-1}$ ways to do this, so

$$\mathrm{P}(\text{“}k^{th} \text{ Head is on } n^{th} \text{ flip”}) = \binom{n-1}{k-1} p^k (1-p)^{n-k}, \quad n = k, k+1, k+2, \ldots$$

Example 1.19 Imagine inspecting cell phone batteries on the production line, one after another. Supposing p is the overall defective rate, it is easy to assume that at any time i the chance of a defective is p. But are successive batteries independent? In particular, can we assume

$$\mathrm{P}(\text{“defective at time } i\text{”} \,|\, \text{“defective at time } i-1\text{”}) = p?$$

Possibly we cannot and another model, such as the two-state Markov chain of Ex. 1.13, would be more suitable.

