Chapter 2

Conditional Probability and Independence

2.1 Conditional Probability

Probability assigns a likelihood to results of experiments that have not yet been conducted. Suppose that the experiment has been conducted, and we know its outcome. Then, this outcome is no longer random—we know it with certainty—whatever the probability of this outcome was, it is no longer relevant. In contrast, suppose that the experiment has been conducted, but we know nothing about the outcome. From our point of view, our predictions should be as if the experiment hasn't yet been performed. Finally, suppose that the experiment has been conducted, and we have *partial information* about the outcome. In that case, the precise outcome is still unknown, however, probabilities have to be reevaluated, based on partial knowledge. The reevaluated probability of an event is called the *conditional probability* of that event given the partial information.

(6 hrs)  (6 hrs)

Example: Two dice are tossed. What is the probability that the sum is 8? The sample space Ω comprises 36 equally-probable elements. The event "sum is 8" corresponds to the subset

$$A = \{(2,6),(3,5),(4,4),(5,3),(6,2)\}.$$

Therefore $P(A) = |A|/|\Omega| = 5/36$.

Next, suppose that someone reveals that the first die resulted in "3". How does this change our predictions? This piece of information tells us that *with certainty* the outcome of the experiment lies in set

$$F = \{(3, i) : 1 \le i \le 6\} \subset \Omega.$$

Thus, outcomes that are not in F have to be ruled out. The sample space can be restricted to F (F becomes the certain event). The event A (sum was "8") has to be restricted to its intersection with F. It seems reasonable that "the probability of A knowing that F has occurred" be defined as

$$\frac{|A \cap F|}{|F|} = \frac{|A \cap F|/|\Omega|}{|F|/|\Omega|} = \frac{P(A \cap F)}{P(F)},$$

which in the present case is $1/6$. ▲▲▲

This example motivates the following definition:

Definition 2.1 Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $F \in \mathcal{F}$ be an event for which $P(F) \neq 0$ (a non-null *event). For every $A \in \mathcal{F}$, the* conditional probability of A given that F has occurred *(*הסתברות מותנה*) (or simply, given F) is defined (and denoted) by*

$$P(A \mid F) := \frac{P(A \cap F)}{P(F)}.$$

Comment: Note that conditional probability is defined only if the conditioning event has non-zero probability.

Discussion: Like probability itself, conditional probability has different interpretations depending on whether you are a frequentist or a Bayesian. In the frequentist interpretation, we have in mind a large set of n repeated experiments. Let n_F denote the number of times event F has occurred, and let $n_{A,F}$ denote the number of times that both events A and F have occurred. In the frequentist's world,

$$P(A \mid F) = \lim_{n \to \infty} \frac{n_{A,F}}{n_F}.$$

In the Bayesian interpretation, this conditional probability is the belief that A has occurred after we learned that F has occurred.

Example: There are 10 white balls, 5 yellow balls and 10 black balls in an urn. A ball is drawn at random, what is the probability that it is yellow (answer: $5/25$)? What is the probability that it is yellow given that it is not black (answer: $5/15$)? Note how the additional information *restricts the sample space to a subset*. ▲▲▲

Example: Jeremy can't decide whether to study history or literature. If he takes literature, he will pass with probability $1/2$; if he takes history, he will pass with probability $1/3$. He made his decision based on a coin toss. What is the probability that he opted for history and passed the exam?

This is an example where the main task is to set up the probabilistic model and interpret the data. First, the sample space: we set it to be the product of the two sets

$$\{\text{history}, \text{literature}\} \times \{\text{pass}, \text{fail}\}.$$

If we define the following events:

$$A = \{\text{passed}\} = \{\text{history}, \text{literature}\} \times \{\text{pass}\}$$
$$B = \{\text{history}\} = \{\text{history}\} \times \{\text{pass}, \text{fail}\},$$

then we interpret the data as follows:

$$P(B) = P(B^c) = \frac{1}{2} \qquad P(A \mid B) = \frac{1}{3} \qquad P(A \mid B^c) = \frac{1}{2}.$$

The quantity to be calculated is $P(A \cap B)$, and this is obtained by

$$P(A \cap B) = P(A \mid B)P(B) = \frac{1}{6}.$$

▲▲▲

Example: There are 8 red balls and 4 white balls in an urn. Two balls are drawn at random. What is the probability that the second was red given that the first was red?

Answer: it is the probability that both were red divided by the probability that the first was red. The result is

$$\frac{\binom{8}{2}}{\binom{12}{2}} = \frac{7}{11},$$

which illuminates the fact that having drawn the first ball red, we can think of a new initiated experiment, where we draw one ball from an urn containing 7 red balls and 4 white balls. ▲▲▲

The next theorem justifies the term conditional probability:

Theorem 2.1 (Conditional probability is a probability) Let $(\Omega, \mathscr{F}, P)$ *be a probability space and let* F *be a non-null event. Define the set function* $Q(A) = P(A \mid F)$*. Then,* Q *is a probability function over* $(\Omega, \mathscr{F})$*.*

Proof: We need to show that the three axioms of a probability function are met. Clearly, Q is non-negative and

$$Q(\Omega) = \frac{P(\Omega \cap F)}{P(F)} = \frac{P(F)}{P(F)} = 1.$$

Finally, let (A_n) be a sequence of mutually disjoint events. Then the events $(A_n \cap F)$ are also mutually disjoint, and

$$\begin{aligned} Q(\bigcup_n A_n) &= \frac{P((\cup_n A_n) \cap F)}{P(F)} = \frac{P(\cup_n (A_n \cap F))}{P(F)} \\ &= \frac{1}{P(F)} \sum_n P(A_n \cap F) = \sum_n Q(A_n). \end{aligned}$$

■

In fact, the function Q is a probability function on the smaller space F, with the σ-algebra

$$\mathscr{F}|_F := \{F \cap A : \ A \in \mathscr{F}\}.$$

✎ *Exercise 2.1* Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $B, B^c \in \mathscr{F}$ have non-zero probability. Prove that if

$$P(A) < P(A \mid B)$$

then

$$P(A) > P(A \mid B^c).$$

In other words, if knowing that B has occurred increases the probability that A has occurred, then knowing that B has not occurred decreases the probability that A has occurred.

✎ *Exercise 2.2* Let A, B, C be three non-null events. We say that "A favors B" if $P(B \mid A) > P(B)$. Is it generally true that if A favors B and B favors C, then A favors C?

✎ *Exercise 2.3* Prove the *general multiplication rule*

$$P(A \cap B \cap C) = P(A)\, P(B \mid A)\, P(C \mid A \cap B),$$

with the obvious generalization for more events. Reconsider the "birthday paradox" in the light of this formula.

2.2 Bayes' rule and the law of total probability

Let $(A_i)_{i=1}^{\infty}$ be a partition of Ω. By that, we mean that the A_i are mutually disjoint and that their union equals Ω (every $\omega \in \Omega$ is in one and only one A_i). Let B be an event. Then, we can write

$$B = \dot{\bigcup_{i=1}^{\infty}} (B \cap A_i).$$

By the countable-additivity of the probability function,

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B \mid A_i) P(A_i). \tag{2.1}$$

This law is known as the law of *total probability* (חוק ההסתברות השלמה); it expresses the probability of an event as a weighted average of conditional probabilities.

The next rule is known as *Bayes' law*: let A, B be two events having positive probability. Then,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{and} \qquad P(B \mid A) = \frac{P(B \cap A)}{P(A)},$$

from which we readily deduce

$$P(A \mid B) = P(B \mid A) \frac{P(A)}{P(B)}. \tag{2.2}$$

Bayes' rule is easily generalized as follows: if $(A_i)_{i=1}^{n}$ is a partition of Ω, then

$$P(A_i \mid B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B \mid A_i) P(A_i)}{\sum_{j=1}^{n} P(B \mid A_j) P(A_j)},$$

where we have used the law of total probability. (Bayes' law is named after Rev. Thomas Bayes (1701–1761), who first provided an equation that allows new evidence to update beliefs.)

Example: Consider a lab screen for the a certain virus. A person that carries the virus is screened positive in only 95% of the cases (5% chance of *false negative*). A person who does not carry the virus is screened positive in 1% of the cases (1% chance of *false positive*). Given that 0.5% of the population carries the virus, what is the probability that a person who has been screened positive is actually a carrier?

Again, we start by setting the sample space,

$$\Omega = \{\text{carrier}, \text{not carrier}\} \times \{+, -\}.$$

Note that the sample space is not a sample of people! Define the events,

$$A = \{\text{the person is a carrier}\} \qquad B = \{\text{the person was screened positive}\}.$$

It is given that

$$P(A) = 0.005 \qquad P(B \mid A) = 0.95 \qquad P(B \mid A^c) = 0.01,$$

and we want to compute $P(A \mid B)$. Now,

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)} \\ &= \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.01 \cdot 0.995} \approx \frac{1}{3}. \end{aligned}$$

This is a nice example where fractions fool our intuition. ▲▲▲

✎ *Exercise 2.4* The following problem is known as *Pólya's urn*. At time $t = 0$ an urn contains two balls, one black and one red. At time $t = 1$ you draw one ball at random and replace it together with a new ball of the same color. You repeat this procedure at every integer time (so that at time $t = n$ there are $n + 2$ balls. Calculate

$$p_{n,r} = P(\text{there are } r \text{ red balls at time } n)$$

for $n = 1, 2, \ldots$ and $r = 1, 2, \ldots, n + 1$. What can you say about the proportion of red balls as $n \to \infty$.

Solution: In order to have r red balls at time n there must be either r or $r - 1$ red balls at time $n - 1$. By the law of total probability, we have the recursive formula

$$p_{n,r} = \left(1 - \frac{r}{n+1}\right) p_{n-1,r} + \frac{r-1}{n+1} p_{n-1,r-1},$$

with "initial conditions" $p_{0,1} = 1$. If we define $q_{n.r} = (n+1)!p_{n,r}$, then

$$q_{n,r} = (n+1-r)\, q_{n-1,r} + (r-1)\, q_{n-1,r-1}.$$

You can easily check that $q_{n,r} = n!$ so that the solution to our problem is $p_{n,r} = 1/(n+1)$. At any time all the outcomes for the number of red balls are equally likely!

2.3 Compound experiments

So far we have only "worked" with a restricted class of probability spaces—finite probability space in which all outcomes have the same probability. The concept of conditional probabilities is also a mean to define a class of probability spaces, representing compound experiments where certain parts of the experiment rely on the outcome of other parts. The simplest way to get insight into it is through examples.

Example: Consider the following statement: "the probability that a family has k children is p_k (with $\sum_k p_k = 1$). For any family size, all gender distributions have equal probabilities". What is the probability space corresponding to such a statement?

Since there is no a-priori limit on the number of children (although every family has a finite number of children), we should take our sample space to be the set of all finite sequences of type "bggbg":

$$\Omega = \{a_1 a_2 \dots a_n : \ a_j \in \{b, g\}, n \in \mathbb{N}\}.$$

This is a countable space so the σ-algebra can include all subsets. What is the probability of a point $\omega \in \Omega$? Suppose that ω is a string of length n, and let A_n be the event "the family has n children", then by the law of total probability

$$p(\omega) = \sum_{m=1}^{\infty} P(\{\omega\} \mid A_m) P(A_m) = \frac{p_n}{2^n}.$$

Having specified the probability of all singletons of a countable space, the probability space is fully specified.

We can then ask, for example, what is the probability that a family with no girls has exactly one child? If B denotes the event "no girls", then

$$P(A_1 \mid B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{p_1/2}{p_1/2 + p_2/4 + p_3/8 + \ldots}.$$

▲ ▲ ▲

Example: Consider two dice: die A has four red and two white faces and die B has two red and four white faces. One tosses a coin: if it falls Head then die A is tossed sequentially, otherwise die B is used.

What is the probability space?

$$\Omega = \{H, T\} \times \{a_1 a_2 \cdots : \ a_j \in \{R, W\}\}.$$

It is a *product* of two subspaces. What we are really given is a probability on the first space and a *conditional* probability on the second space. If A_H and A_T represent the events "head has occurred" and "tail has occurred", then we know that

$$P(A_H) = P(A_T) = \frac{1}{2},$$

and facts like

$$P(\{RRWR\} \mid A_H) = \frac{4}{6} \cdot \frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{6}$$
$$P(\{RRWR\} \mid A_T) = \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{2}{6}.$$

(No matter for the moment where these numbers come from....) ▲ ▲ ▲

The following well-known "paradox" demonstrates the confusion that can arise where the boundaries between formalism and applications are fuzzy.

Example: *The sibling paradox.* Suppose that in all families with two children all the four combinations $\{bb, bg, gb, gg\}$ are equally probable. Given that a family with two children has at least one boy, what is the probability that it also has a girl? The easy answer is $2/3$.

Suppose now that one knocks on the door of a family that has two children, and a boy opens the door and says "I am the oldest child". What is the probability that he has a sister? The answer is one half. Repeat the same scenario but this time the boy says "I am the youngest child". The answer remains the same. Finally, a

boy opens the door and says nothing. What is the probability that he has a sister: a half or two thirds?

The resolution of this paradox is that the experiment is not well defined. We could think of two different scenarios: (i) all families decide that boys, when available, should open the door. In this case if a boy opens the door he just rules out the possibility of gg, and the likelihood of a girl is $2/3$. (ii) When the family hears knocks on the door, the two siblings toss a coin to decide who opens. In this case, the sample space is

$$\Omega = \{bb, bg, gb, gg\} \times \{1, 2\},$$

and all 8 outcomes are equally likely. When a boy opens, he gives us the knowledge that the outcome is in the set

$$A = \{(bb, 1), (bb, 2), (bg, 1), (gb, 2)\}.$$

If B is the event that there is a girl, then

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(\{(bg, 1), (gb, 2)\})}{P(A)} = \frac{2/8}{4/8} = \frac{1}{2}.$$

▲ ▲ ▲

✎ *Exercise 2.5* Consider the following generic compound experiment: one performs a first experiment to which corresponds a probability space $(\Omega_0, \mathscr{F}_0, P_0)$, where Ω_0 is a finite set of size n. Depending on the outcome of the first experiment, the person conducts a second experiment. If the outcome was $\omega_j \in \Omega_0$ (with $1 \le j \le n$), he conducts an experiment to which corresponds a probability space $(\Omega_j, \mathscr{F}_j, P_j)$. Construct a probability space that corresponds to the compound experiment.

2.4 Independence

Definition 2.2 Let A and B be two events in a probability space $(\Omega, \mathscr{F}, P)$. *We say that A* is independent of *B (*בלתי תלוי*) if the knowledge that B has occurred does not alter the probability that A has occurred. That is, A is independent of B if*

$$P(A \mid B) = P(A). \tag{2.3}$$

By the definition of conditional probability, this condition is equivalent to

$$P(A \cap B) = P(A)P(B), \tag{2.4}$$

which may be taken as an alternative definition of independence; the latter condition makes sense also if $P(A)$ or $P(B)$ are zero. By the symmetry of this condition we immediately conclude:

Corollary 2.1 Independence is a symmetric property: if A is independent of B then B is also independent of A

Example: A card is randomly drawn from a deck of 52 cards. Let

$$A = \{\text{the card is an Ace}\}$$
$$B = \{\text{the card is a Spade}\}.$$

Are these events independent?

We need to verify whether the definition (2.3) of independence is satisfied. The sample space is a set of cardinality 52.

$$P(A) = \frac{|A|}{|\Omega|} = \frac{4}{52} = \frac{1}{13},$$

$$P(B) = \frac{|B|}{|\Omega|} = \frac{13}{52} = \frac{1}{4},$$

and

$$P(A \cap B) = \frac{|A \cap B|}{|\Omega|} = \frac{1}{52}.$$

It follows that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{1}{13} = P(A),$$

hence A and B are independent. ▲▲▲

Example: Two dice are tossed. Let

$$A = \{\text{the first die is a } 4\}$$
$$B = \{\text{the sum is } 6\}.$$

Are these events independent (answer: no)? What if the sum was 7 (answer: yes)?
▲▲▲

Proposition 2.1 Every event is independent of Ω *and* $\varnothing$.

Proof: Immediate from the definition (note that in the latter case, we need to use the multiplicative definition (2.4) of independence). ■

Proposition 2.2 If B is independent of A then it is independent of A^c.

Proof: Since $B = (B \cap A^c) \cup (B \cap A)$, it follows that

$$P(B) = P(B \cap A^c) + P(B \cap A).$$

Using the independence of A and B,

$$\begin{aligned} P(B \cap A^c) &= P(B) - P(A \cap B) \\ &= P(B) - P(A)P(B) \\ &= P(B)(1 - P(A)) \\ &= P(B)P(A^c). \end{aligned}$$

■

Thus, if B is independent of A, it is independent of the collection of sets

$$\sigma(A) = \{\Omega, A, A^c, \varnothing\},$$

which is *the* σ*-algebra generated by* A. This innocent observation will gain meaning in a moment.

(8 hrs)  (8 hrs)

Next, consider three events A, B, C: What does it mean for *three* events to be independent? For example, What does it mean for A to be independent of B and C? A first natural guess would be to say that A is independent of A and B if the knowledge that B has occurred does not affect the probability of A, and the knowledge that C has occurred does not affect the probability of A either. But does this condition imply that the probability of A is independent of *any* information regarding whether B and C occurred?

Let's examine an example:

Example: Consider the toss of two dice and define

$$A = \{\text{the sum is } 7\}$$
$$B = \{\text{the first die is a } 4\}$$
$$C = \{\text{the second die is a } 2\}.$$

A is independent of B and A is independent of C. However, if we know that both B *and* C have occurred, then A, which had originally a probability of $1/6$ becomes an impossible event! In other words, while A is independent of B and A is independent of C, A is not independent on their intersection. It would therefore be misleading to claim that A is independent of B and C. This example calls for a modified definition of independence between multiple events. ▲▲▲

Definition 2.3 The event A is said to be independent of the pair *of events B and C if it is independent of every event in the σ-algebra generated by B and C. That is, if it is independent of the collection*

$$\sigma(B,C) = \{B, C, B^c, C^c, B\cap C, B\cup C, B\cap C^c, B\cup C^c, \ldots, \Omega, \varnothing\}.$$

It seems as if there are a lot of conditions to verify in order to assert than an event is independent of a pair of events. The following proposition shows that the number of sufficient conditions is significantly smaller:

Proposition 2.3 A is independent of B and C if and only if and only if it is independent of B, C and $B\cap C$, that is, if and only if

$$P(A\cap B) = P(A)P(B) \qquad P(A\cap C) = P(A)P(C)$$
$$P(A\cap B\cap C) = P(A)P(B\cap C).$$

Proof: The "only if" part is obvious. Now to the "if" part. We need to show that A is independent of each element in the σ-algebra generated by B and C. What we already know is that A is independent of B, C, $B\cap C$, B^c, C^c, $B^c\cup C^c$, Ω, and $\varnothing$ (not too bad!). Take for example the event $B\cup C$. Using the inclusion-exclusion formula, and the given independences,

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

■

✎ *Exercise 2.6* Prove directly that if A is independent of B, C, and $B \cap C$, then it is independent of $B \setminus C$.

Corollary 2.2 The events A, B, C are mutually independent in the sense that each one is independent of the remaining pair if and only if

$$P(A \cap B) = P(A)P(B) \qquad P(A \cap C) = P(A)P(C)$$
$$P(B \cap C) = P(B)P(C) \qquad P(A \cap B \cap C) = P(A)P(B)P(C).$$

More generally,

Definition 2.4 A collection of events (A_n) is said to consist of mutually independent events *if for every subset $A_{n_1}, \ldots, A_{n_k}$,*

$$P(A_{n_1} \cap \cdots \cap A_{n_k}) = \prod_{j=1}^{k} P(A_{n_j}).$$

(9 hrs) (9 hrs)

2.5 Repeated Trials

Only now that we have defined the notion of independence we can consider the situation of an experiment being repeated again and again *under identical conditions*—a situation underlying the very notion of probability.

Consider an experiment, i.e., a probability space $(\Omega_0, \mathscr{F}_0, P_0)$. We want to use this probability space to construct a compound probability space corresponding to the idea of repeating the same experiment sequentially n times, the outcome of each trial being independent of all other trials. For simplicity (namely, to avoid measure-theoretic delicacies), we assume that the single experiment corresponds to a countable probability space.

Reminder: if Ω_0 is a countable set, then for every $n \in \mathbb{N}$, Ω_0^n is also countable. Moreover,

$$\bigcup_{n=0}^{\infty} \Omega_0^n$$

is countable (the space of all *finite* sequences of elements in Ω_0). On the other hand, the space $\Omega_0^{\mathbb{N}}$ of *infinite* sequences is not countable.

Consider now the compound experiment of repeating the same experiment n times. The sample space consists of n-tuples,

$$\Omega = \Omega_0^n = \{(a_1, a_2, \ldots, a_n) : \ a_j \in \Omega_0\}.$$

Since this is a countable space, the probability is fully determined by its value for all singletons. Each singleton,

$$\omega = (a_1, a_2, \ldots, a_n),$$

corresponds to an event, which is the intersection of the n events: "first outcome was a_1", "second outcome was a_2", etc. Since we assume independence between trials, its probability is the product of the individual probabilities. I.e.,

$$P(\{(a_1, a_2, \ldots, a_n)\}) = p_0(a_1)p_0(a_2)\ldots p_0(a_n), \tag{2.5}$$

where $p_0(a_j)$ is as usual shorthand for $P_0(\{a_j\})$. Note that this is *not* the only possible probability that one can define on Ω_0^n.

Proposition 2.4 Definition (2.5) *defines a probability function on* Ω_0^n.

Proof: Immediate. We only need to verify that these numbers add up to 1. ■

The following proposition shows that (2.5) does indeed correspond to a situation where different trials do not influence each other's statistics:

Proposition 2.5 Let $A_1, A_2, \ldots, A_n$ *be a sequence of events such that the* j*-th trial alone determines whether* A_j *has occurred; that is, there exists a* $B_j \subseteq \Omega_0$*, such that*

$$A_j = \Omega_0^{j-1} \times B_j \times \Omega_0^{n-j}.$$

If the probability is defined by (2.5)*, then the* A_j *are mutually independent.*

Proof: First, note that

$$A_j = \bigcup_{a_1 \in \Omega_0}^{\cdot} \cdots \bigcup_{a_j \in B_j}^{\cdot} \cdots \bigcup_{a_n \in \Omega_0}^{\cdot} \{(a_1, a_2, \ldots, a_n)\}.$$

Hence

$$P(A_j) = \sum_{a_1 \in \Omega_0} \cdots \sum_{a_j \in B_j} \ldots \cdots \sum_{a_n \in \Omega_0} p_0(a_1) p_0(a_2) \ldots p_0(a_n) = P_0(B_j).$$

Consider then a pair of such events A_j, A_k, say, $j < k$. Then,

$$A_j \cap A_k = \Omega_0^{j-1} \times B_j \times \Omega_0^{k-j-1} \times B_k \times \Omega_0^{n-k},$$

which can be written as

$$A_j \cap A_k = \bigcup_{a_1 \in \Omega_0}^{\cdot} \cdots \bigcup_{a_j \in B_j}^{\cdot} \cdots \bigcup_{a_k \in B_k}^{\cdot} \cdots \bigcup_{a_n \in \Omega_0}^{\cdot} \{(a_1, a_2, \ldots, a_n)\}.$$

Using the additivity of the probability,

$$\begin{aligned} P(A_j \cap A_k) &= \sum_{a_1 \in \Omega_0} \cdots \sum_{a_j \in B_j} \cdots \sum_{a_k \in B_k} \cdots \sum_{a_n \in \Omega_0} p_0(a_1) p_0(a_2) \ldots p_0(a_n) \\ &= P_0(B_j) P_0(B_k) = P(B_j) P(B_k). \end{aligned}$$

Thus, the pairwise independence has been proved. Similarly, we can take all of the triples, all of the quadruples, etc. ■

Example: Consider an experiment with two possible outcomes: "Success" with probability p and "Failure" with probability $q = 1 - p$ (such an experiment is called a *Bernoulli trial*). Consider now an infinite sequence of independent repetitions of this basic experiment. While we have not formally defined such a probability space (it is uncountable), we do have a precise probabilistic model for any finite subset of trials. There exists an "extension theorem" due to Kolmogorov asserting that a probability function defined on every finite n-tuple defines a unique probability function on the space of infinite sequences.

We consider now the following questions: (1) What is the probability of at least one success in the first n trials? (2) What is the probability of exactly k successes in the first n trials? (3) What is the probability of an infinite sequence of successes?

Let A_j denote the event "the j-th trial was a success". What we know is that for all distinct natural numbers $j_1, \ldots, j_n$,

$$P(A_{j_1} \cap \cdots \cap A_{j_n}) = p^n.$$

To answer the first question, we note that the probability of having only failures in the first n trials is q^n, hence the answer is $1 - q^n$. To answer the second question, we note that exactly k successes out of n trials is a disjoint union of n-choose-k singletons, the probability of each being $p^k q^{n-k}$. Finally, to answer the third question, we use the continuity of the probability function,

$$P(\cap_{j=1}^{\infty} A_j) = P(\cap_{n=1}^{\infty} \cap_{j=1}^{n} A_j) = P(\lim_{n\to\infty} \cap_{j=1}^{n} A_j) = \lim_{n\to\infty} P(\cap_{j=1}^{n} A_j) = \lim_{n\to\infty} p^n,$$

which equals 1 if $p = 1$ and zero otherwise. Alternatively, we note that the probability of an infinite sequence of successes is less than the probability of having, for every $n \in \mathbb{N}$, n consecutive successes; since the latter equals p^n, this probability must be zero. ▲▲▲

Example: *(The gambler's ruin problem, Bernoulli 1713)* Consider the following game involving two players, which we call Player A and Player B. Player A starts the game owning i coins while Player B owns $N - i$ coins. The game is a zero-sum game, where each turn a coin is tossed. The coin has probability p to fall on Head, in which case Player B pays Player A one coin; it has probability $q = 1 - p$ to fall on Tail, in which case Player A pays Player B one coin. The game ends when one of the players is broke. What is the probability for Player A to win?

While the game may end after a finite time, the simplest sample space is that of an infinite sequence of tosses, $\Omega = \{H, T\}^{\mathbb{N}}$. The event

$$E = \{\text{“Player A wins”}\},$$

consists of all sequences in which the number of Heads exceeds the number of Tails by $N - i$ before the number of Tails has exceeded the number of Heads by i. If

$$F = \{\text{“first toss was Head”}\},$$

then by the law of total probability,

$$P(E) = P(E \mid F)P(F) + P(E \mid F^c)P(F^c) = pP(E \mid F) + qP(E \mid F^c).$$

If the first toss was a Head, then by our assumption of mutual independence, we can think of the game starting anew with Player A having $i + 1$ coins (and $i - 1$ if the first toss was Tail). Thus, if α_i denote the probability that Player A wins if he starts with i coins, then

$$\alpha_i = p\,\alpha_{i+1} + q\,\alpha_{i-1},$$

or equivalently,

$$\alpha_{i+1} - \alpha_i = \frac{q}{p}(\alpha_i - \alpha_{i-1}).$$

The "boundary conditions" are $\alpha_0 = 0$ and $\alpha_N = 1$.

This (linear!) system of equations is easily solved. We have

$$\begin{aligned}
\alpha_1 - \alpha_0 &= \alpha_1 \\
\alpha_2 - \alpha_1 &= \frac{q}{p}\alpha_1 \\
\alpha_3 - \alpha_2 &= \frac{q}{p}(\alpha_2 - \alpha_1) = \left(\frac{q}{p}\right)^2 \alpha_1 \\
&\vdots = \vdots \\
\alpha_N - \alpha_{N-1} &= \frac{q}{p}(\alpha_{N-1} - \alpha_{N-2}) = \left(\frac{q}{p}\right)^{N-1} \alpha_1.
\end{aligned}$$

Adding up,

$$1 = \left[1 + \frac{q}{p} + \cdots + \left(\frac{q}{p}\right)^{N-1}\right]\alpha_1 = \frac{(q/p)^N - 1}{q/p - 1}\,\alpha_1,$$

i.e.,

$$\alpha_1 = \begin{cases} \dfrac{q/p - 1}{(q/p)^N - 1} & p \neq q \\ \dfrac{1}{N} & p = q. \end{cases},$$

from which we get the probability that player A wins:

$$\alpha_i = \begin{cases} \dfrac{(q/p)^i - 1}{(q/p)^N - 1} & p \neq q \\ \dfrac{i}{N} & p = q. \end{cases},$$

What is the probability that Player B wins? Exchange i with $N - i$ and p with q. What is the probability that either of them wins? The answer turns out to be 1!

Finally, what is the probability that player A wins if player B is initially very rich (it is the casino)? Letting $N \to \infty$ we get that

$$\lim_{N\to\infty} \alpha_i = \begin{cases} 1 - (q/p)^i & p > q \\ 0 & p \leq q. \end{cases}$$

This is a very interesting result. If your chances to win in a single play is less than the casino's, then you're certain to go broke. Otherwise, you have a finite chance to win, which grows with your initial capital. ▲▲▲

(11 hrs) (11 hrs)

2.6 On Zero-One Laws

Events that have probability either zero or one are often very interesting. We will demonstrate such a situation with a funny example, which is representative of a class of problems that have been classified by Kolmogorov as 0-1 laws. The general theory is beyond the scope of this course.

Consider a monkey typing on a typing machine, each second typing a character (a letter, number, or a space). Each character is typed at random, independently of past characters. The sample space consists thus of infinite strings of typing-machine characters. The question that interests us is how many copies of the Collected Work of Shakespeare (WS) did the monkey produce. We define the following events:

$$\begin{aligned} H &= \{\text{the monkey produces infinitely many copies of WS}\} \\ H_k &= \{\text{the monkey produces at least } k \text{ copies of WS}\} \\ H_{m,k} &= \{\text{the monkey produces at least } k \text{ copies of WS by time } m\} \\ H^m &= \{\text{the monkey produces infinitely many copies of WS after time } m+1\}. \end{aligned}$$

Note the following dependencies between the various events:

(a) $H^m = H$, i.e., the event of producing infinitely many copies is not affected by any finite prefix (it is a *tail event!*).

(b) For every k, m, $H_{m,k}$ and H^m are independent, because the first m characters are independent of the characters from $m+1$ on.

(c) $H_{m,k}$ is an increasing sequence in m, and

$$\lim_{m\to\infty} H_{m,k} = \bigcup_{m=1}^{\infty} H_{m,k} = H_k.$$

Similarly,

$$\lim_{m\to\infty} (H_{m,k} \cap H) = H_k \cap H = H.$$

(d) H_k is a decreasing sequence and

$$\lim_{k\to\infty} H_k = \bigcap_{k=1}^{\infty} H_k = H.$$

Now, by the independence property, for every k, m,

$$P(H_{m,k} \cap H^m) = P(H_{m,k})P(H^m).$$

and since $H^m = H$,

$$P(H_{m,k} \cap H) = P(H_{m,k})P(H).$$

By the continuity of the probability function,

$$\lim_{m\to\infty} P(H_{m,k} \cap H) = P\left(\lim_{m\to\infty} H_{m,k} \cap H\right) = P(H),$$

and

$$\lim_{m\to\infty} P(H_{m,k}) = P\left(\lim_{m\to\infty} H_{m,k}\right) = P(H_k),$$

from which we deduce that

$$P(H) = P(H_k)P(H).$$

Finally, since

$$\lim_{k\to\infty} P(H_k) = P\left(\lim_{k\to\infty} H_k\right) = P(H),$$

we obtain

$$P(H) = P(H)P(H),$$

from which we conclude that $P(H)$ is either zero or one.

2.7 Further examples

In this section we examine more applications of conditional probabilities.

Example: The following example is actually counter-intuitive. Consider an infinite sequence of tosses of a fair coin. There are eight possible outcomes for three consecutive tosses, which are `HHH`, `HHT`, `HTH`, `HTT`, `THH`, `THT`, `TTH`, and `TTT`. It turns

out that for any of those triples, there exists another triple, which is likely to occur first with probability strictly greater than one half.

Take for example $s_1 = \texttt{HHH}$ and $s_2 = \texttt{THH}$, then

$$P(s_2 \text{ before } s_1) = 1 - P(\text{first three tosses are H}) = \frac{7}{8}.$$

Take $s_3 = \texttt{TTH}$, then

$$P(s_3 \text{ before } s_2) = P(\text{TT before } s_2) > P(\text{TT before HH}) = \frac{1}{2}.$$

where the last equality follows by symmetry. ▲▲▲

✎ *Exercise 2.7* Convince yourself that the above statement is indeed correct by examining all cases.