EE351K—SPRING 2014

1. PROBABILITY SPACES

Definition 1.1. A *probability space* consists of three objects $(\Omega, \mathfrak{F}, \mathbb{P})$:

(1) Ω is a set called the *sample space*.

(2) $\mathfrak{F}$ is a collection of subsets of Ω (called *events*), having the properties:
  (2a) $\Omega \in \mathfrak{F}$.
  (2b) If $A \in \mathfrak{F}$, then $A^c \in \mathfrak{F}$.
  (2c) If $A_k \in \mathfrak{F}$, $1 \leq k < \infty$, is a sequence of events, then $\cup_{k=1}^{\infty} A_k \in \mathfrak{F}$.

(3) The probability $\mathbb{P}$ is a function $\mathbb{P}\colon \mathfrak{F} \to [0,1]$, satisfying
  (3a) $\mathbb{P}(\Omega) = 1$.
  (3b) If $A_j \in \mathfrak{F}$, $1 \leq j < \infty$, is a sequence of *pairwise disjoint* events, i.e., $A_i \cap A_j = \varnothing$ when $i \neq j$, then

$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)\,.$$

Remark 1. As a consequence of this definition if $\{A_k\} \subset \mathfrak{F}$ is a sequence of events, then $\cap_{k=1}^{\infty} A_k \in \mathfrak{F}$. Hence, the *countable* intersection (or union) of events is also an event. In short, we say that $\mathfrak{F}$ *is closed* under complementation and countable unions and intersections.

Definition 1.2. Any collection of subsets of Ω, which is closed under complementation, countable unions and intersections, and also contains Ω, is called a *σ-field*. The set of events $\mathfrak{F}$ is of course, also a σ-field.

If $\{A_1, \ldots, A_n\}$ is a collection of events, then we use the notation $\sigma(A_1, \ldots, A_n)$, to denote the *smallest* σ-field that contains all A_i's. We refer to $\sigma(A_1, \ldots, A_n)$ as the σ-field *generated* by $\{A_1, \ldots, A_n\}$. This of course can be formed by combining the collection $\{A_1, \ldots, A_n\}$ and their complements $\{A_1^c, \ldots, A_n^c\}$ and then generating all possible intersections and unions of these.

Exercise 1.1. Let $\{A_1, \ldots, A_n\}$ and set $\mathfrak{G} \triangleq \sigma(A_1, \ldots, A_n)$. We say that $B \in \mathfrak{G}$ is an *atom* if for any $C \in \mathfrak{G}$, either $B \cap C = B$, or $B \cap C = \varnothing$.

(i) Argue that $\mathfrak{G}$ contains at most 2^n atoms.
(ii) Show that every element of $\mathfrak{G}$ can be expressed as a union of its atoms.

Therefore $\sigma(A_1, \ldots, A_n)$ contains at most 2^{2^n} events.

Some properties of $\mathbb{P}$

(i) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$

(ii) $\mathbb{P}(\varnothing) = 0$

(iii) If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$ and $\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$.

(iv) $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$ (*subadditivity*)

(v) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

(vi) If A_n is a monotone sequence, i.e., either $A_n \uparrow A$, or $A_n \downarrow A$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)$.

Exercise 1.2. With $a \vee b$ and $a \wedge b$ denoting the maximum and the minimum of two numbers a and b, respectively, show that

$$\mathbb{P}(A \cap B \cap C) \leq \mathbb{P}(A) \wedge \mathbb{P}(B) \wedge \mathbb{P}(C)$$
$$\mathbb{P}(A \cup B \cup C) \geq \mathbb{P}(A) \vee \mathbb{P}(B) \vee \mathbb{P}(C)$$

Exercise 1.3. Show that $\mathbb{P}(A \cap B) \geq \mathbb{P}(A) + \mathbb{P}(B) - 1$.

1.1. **Definitions from Set Theory.** The set of natural numbers is denoted by $\mathbb{N}$, the set of integers by $\mathbb{Z}$ and the set of real numbers by $\mathbb{R}$. If $a \leq b$ are real numbers then the intervals $[a, b) \subset \mathbb{R}$ and $[a, b] \subset \mathbb{R}$ are defined by

$$[a, b) \triangleq \{x \in \mathbb{R} : a \leq x < b\}, \qquad [a, b] \triangleq \{x \in \mathbb{R} : a \leq x \leq b\}.$$

Analogously defined are the intervals $(a, b]$ and (a, b). Note that

$$(a, b) = \bigcup_{n=1}^{\infty} \left[a - \tfrac{1}{n}, b\right), \quad [a, b] = \bigcap_{n=1}^{\infty} \left[a, b + \tfrac{1}{n}\right), \quad \text{and} \quad (a, b] = \bigcup_{n=1}^{\infty} \bigcap_{m=1}^{\infty} \left[a - \tfrac{1}{n}, b + \tfrac{1}{m}\right).$$

Therefore, if we start with the collection of all intervals of the form $[a, b)$, we can generate the rest by countable unions and intersections.

The *difference* of two sets A and B is denoted (and defined) by $A \setminus B \triangleq A \cap B^c$, while their *symmetric difference* is the set $A \triangle B \triangleq (A \setminus B) \cup (B \setminus A)$.

We say that the sets $A_1, \ldots, A_n$ form *a partition* of A, if they are pairwise disjoint and, $A = A_1 \cup \cdots \cup A_n$. Note, for example, that $\{A \setminus B, B \setminus A, A \cap B\}$ is a partition of $A \cup B$, and so is $\{A \triangle B, A \cap B\}$. Therefore, if A and B are events we can write, for example,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B) + \mathbb{P}(A \triangle B).$$

Exercise 1.4. Show the following identities:

$$A \setminus B = A \setminus (A \cap B) = (A \cup B) \setminus B$$
$$A \cap (B \setminus C) = (A \cap B) \setminus (A \cap C)$$
$$(A \cup B) \setminus C = (A \setminus C) \cup (B \setminus C)$$
$$(A \setminus C) \cap (B \setminus C) = (A \cap B) \setminus C$$
$$(A \setminus B) \setminus C = A \setminus (B \cup C)$$
$$A \setminus (B \setminus C) = (A \setminus B) \cup (A \cap C)$$
$$(A \setminus B) \cap (C \setminus D) = (A \cap C) \setminus (B \cup D)$$

With respect to the operations $\triangle$ and $\cap$, sets form a nice algebra (think of $\triangle$ as addition and $\cap$ as multiplication)

$$\begin{gathered} A \triangle B = B \triangle A \quad \text{(symmetric rule)} \\ A \triangle (B \triangle C) = (A \triangle B) \triangle C \quad \text{(associative rule)} \\ A \cap (B \triangle C) = (A \cap B) \triangle (A \cap C) \quad \text{(distributive rule)} \\ A \triangle \varnothing = A \quad \text{(identity)}\,, \qquad A \triangle A = \varnothing \quad \text{(nilpotent)}\,. \end{gathered}$$

Exercise 1.5. Show that there is some operation $\sharp$ between two sets A and B, from which alone all operations we have defined can be derived, provided the set Ω can also be utilized in the expressions.

Exercise 1.6. Suppose $C_1, \ldots, C_n$ form a partition of Ω and $\alpha_1, \ldots, \alpha_n$ are arbitrary positive numbers. For each event $A \in \mathfrak{F}$, define

$$\tilde{\mathbb{P}}(A) \triangleq \frac{\alpha_1 \mathbb{P}(A \cap C_1) + \cdots + \alpha_n \mathbb{P}(A \cap C_n)}{\alpha_1 \mathbb{P}(C_1) + \cdots + \alpha_n \mathbb{P}(C_n)}$$

Show that $\tilde{\mathbb{P}}$ is a well defined probability on $(\Omega, \mathfrak{F})$.

Definition 1.3. The number of elements of a finite set A is called its *cardinality* and is denoted by $|A|$. The collection of all subsets (i.e., a set of sets) of a finite set A, is called the *power set* of A. It is the case that if $|A| = n$, then the cardinality of its power set is 2^n. That is the reason that the power set of A is often denoted as 2^A.

Investigating further, the number of subsets of A having cardinality k is equal to $\frac{n!}{k!(n-k)!}$, for $0 \leq k \leq n$. When $k = 0$ this refers of course to the empty set, and by convention $0! = 1$. This motivates the definition of the *binomial coefficient* with parameters n, k, as

$$\binom{n}{k} \triangleq \frac{n!}{k!(n-k)!}\,, \quad 0 \leq k \leq n\,. \tag{1.1}$$

Adding the sizes of the collections of all subsets of A of a given cardinality k, from $k = 0$ to n, we obtain

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n. \tag{1.2}$$

Exercise 1.7. Take $\Omega = \mathbb{N}$, the set of natural numbers. We want to assign a probability that agrees with the notion of frequency (arithmetical density). In other words, if D_k denotes the set of all natural numbers which are divisible by the number k, then we should have $\mathbb{P}(D_k) = \frac{1}{k}$. Note then that $\mathbb{P}(D_k \cap D_\ell) = \frac{1}{\mathrm{lcm}(k,\ell)}$, where $\mathrm{lcm}(k, \ell)$ denotes the least common multiple of k and ℓ (note that this implies that if k and ℓ are co-prime, then D_k and D_ℓ are independent). Thus

$$\begin{aligned} \mathbb{P}(D_k \cup D_\ell) &= \mathbb{P}(D_k) + \mathbb{P}(D_k) - \mathbb{P}(D_k \cap D_\ell) \\ &= \frac{1}{k} + \frac{1}{\ell} - \frac{1}{\mathrm{lcm}(k,\ell)}\,. \end{aligned}$$

A definition of $\mathbb{P}$ could perhaps use the following: if A is a subset of $\mathbb{N}$, let $N_n(A)$ denote the number of elements of A up to and including n, and considering the fraction $\frac{N_n(A)}{n}$, let $n \to \infty$. It is straightforward to show that for the sets D_k we obtain the correct answer and

$$\frac{N_n(D_k \cup D_\ell)}{n} \xrightarrow[n\to\infty]{} \frac{1}{k} + \frac{1}{\ell} - \frac{1}{\mathrm{lcm}(k,\ell)} \,.$$

But we didn't talk about $\mathfrak{F}$ yet? Can we say that sets containing a single number, i.e., $\{m\}$, or a finite set of numbers, are events? Of course, finite sets should have probability equal to 0. Investigate.

1.2. **Independence.** We start with a definition.

Definition 1.4. A finite collection of events $\{A_1, \ldots, A_n\}$ is said to be *independent* (or the events are called *mutually independent*) if for any subcollection $\{A_{k_1}, \ldots, A_{k_\ell}\}$ it holds

$$\mathbb{P}(A_{k_1} \cap \cdots \cap A_{k_\ell}) = \mathbb{P}(A_{k_1}) \times \cdots \times \mathbb{P}(A_{k_\ell}) \,.$$

An arbitrary (not necessarily finite) collection of events is said to be independent if any finite subcollection is so.

Two important properties of independent events which can be derived from this definition are the following (we have proved the first in class; try to prove the second—it is simpler):

Property 1. $A_1, \ldots, A_n$ are mutually independent if and only if $\mathbb{P}(\tilde{A}_1 \cap \cdots \cap \tilde{A}_n) = \prod_{i=1}^n \mathbb{P}(\tilde{A}_i)$ for any collection of the form $\{\tilde{A}_1, \ldots, \tilde{A}_n\}$, with $\tilde{A}_i$ taking the values A_i or A_i^c.

Property 2. If the events $A_1, \ldots, A_n$ are mutually independent and $1 \le k \le n$, then any event in $\sigma(A_1, \ldots, A_k)$ (see Definition 1.2) is independent from any event in $\sigma(A_{k+1}, \ldots, A_n)$.

We can extend the definition of independence as follows. If G' and G'' are two collections of events, we say they are independent, and denote this by $G' \perp\!\!\!\perp G''$ if every event in G' is independent from any event in G''. Then by Property 2, $\sigma(G') \perp\!\!\!\perp \sigma(G'')$.

1.3. **Borel-Cantelli.** If $\{A_n \,,\ 1 \le n < \infty\}$ is a sequence of events we define

$$\limsup_{n\to\infty} \ A_n \triangleq \bigcap_{n=1}^{\infty} \left(\bigcup_{k=n}^{\infty} A_k \right) \,, \qquad \liminf_{n\to\infty} \ A_n \triangleq \bigcup_{n=1}^{\infty} \left(\bigcap_{k=n}^{\infty} A_k \right)$$

Since $\mathfrak{F}$ is closed under countable unions and intersections, it follows that the sets $\limsup_{n\to\infty} \ A_n$ and $\liminf_{n\to\infty} \ A_n$ are also events. Also, observe that if we define

$$B_n \triangleq \bigcup_{k=n}^{\infty} A_k \,, \qquad C_n \triangleq \bigcap_{k=n}^{\infty} A_k \,,$$

then $\{B_n\}$ is an decreasing sequence of events, while $\{C_n\}$ is an increasing one.

It is easy to verify that $\liminf_{n\to\infty} \ A_n \subset \limsup_{n\to\infty} \ A_n$ and that

$$\left(\liminf_{n\to\infty} \ A_n \right)^c = \limsup_{n\to\infty} \ A_n^c \,, \quad \text{and} \quad \left(\limsup_{n\to\infty} \ A_n \right)^c = \liminf_{n\to\infty} \ A_n^c \,.$$

Note that a point $\omega \in \limsup_{n\to\infty} A_n$ if and only if ω is in *infinitely many* events A_n, and $\omega \in \liminf_{n\to\infty} A_n$ if and only if ω is in *all but finitely many* events A_n.

Therefore, we arrive at the following interpretation (viewing n as time)

$$\mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) \equiv \mathbb{P}(\text{the event } A_n \text{ occurs infinitely often})$$

$$\mathbb{P}\Big(\liminf_{n\to\infty} A_n\Big) \equiv \mathbb{P}(\text{the event } A_n \text{ always occurs, except for finitely many instances})$$

A good example to keep in mind is that if $A_n = B$ for n even, and $A_n = C$ for n odd, then $\limsup_{n\to\infty} A_n = B \cup C$ and $\liminf_{n\to\infty} A_n = B \cap C$. Try to prove that if A_n is a sequence of pairwise disjoint sets then $\limsup_{n\to\infty} A_n = \liminf_{n\to\infty} A_n = \varnothing$.

We proved the following important result in class.

Theorem (Borel-Cantelli)**.** *If $\{A_n\}$ is a sequence of events then*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \quad \Longrightarrow \quad \mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = 0\,.$$

Moreover, if the events $\{A_n\}$ ***are mutually independent*** *then*

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty \quad \Longrightarrow \quad \mathbb{P}\Big(\limsup_{n\to\infty} A_n\Big) = 1\,.$$

1.4. **Conditioning.** For A and B events, define the *conditional probability of A given B* by

$$\mathbb{P}(A \mid B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,, \quad \text{if } \mathbb{P}(B) \neq 0\,,$$

and assign an arbitrary value to it if $\mathbb{P}(B) = 0$.

Properties of Conditional Probability

(1) If $\mathbb{P}(B) \neq 0$, and with B fixed we define $\mathbb{P}_B(A) \triangleq \mathbb{P}(A \mid B)$, for $A \in \mathfrak{F}$, then $\mathbb{P}_B$ is a well defined probability on $(\Omega, \mathfrak{F})$ (i.e., it satisfies the axioms of probability).

(2) If $\mathbb{P}(B) \neq 0$ and $A \perp\!\!\!\perp B$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.

(3) If $\{A_1, \ldots, A_r\}$ is a partition of Ω, then we have the rule of *total probability*:

$$\mathbb{P}(B) = \sum_{i=1}^{r} \mathbb{P}(B \mid A_i)\, \mathbb{P}(A_i)\,.$$

(4) Bayes' formula: If $\{A_1, \ldots, A_r\}$ is a partition of Ω, then provided $\mathbb{P}(B) \neq 0$,

$$\mathbb{P}(A_k \mid B) = \frac{\mathbb{P}(B \mid A_k)\,\mathbb{P}(A_k)}{\sum_{i=1}^{r} \mathbb{P}(B \mid A_i)\,\mathbb{P}(A_i)}\,.$$

(5) For arbitrary events $A_1, \ldots, A_n$, provided $\mathbb{P}(A_1 \cap \cdots \cap A_n) \neq 0$, we have

$$\mathbb{P}(A_1 \cap \cdots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2) \cdots \mathbb{P}(A_n \mid A_1 \cap \cdots \cap A_{n-1})\,.$$

Exercise 1.8. Find examples where $\mathbb{P}(A \cap B) < \mathbb{P}(A)\,\mathbb{P}(B)$, and $\mathbb{P}(A \cap B) > \mathbb{P}(A)\,\mathbb{P}(B)$.

Exercise 1.9. Two events A and B are *conditionally independent* given an event C if

$$\mathbb{P}(A \cap B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)\,. \tag{1.3}$$

Show that (1.4) implies

$$\mathbb{P}(A \mid C \cap B) = \mathbb{P}(A \mid C)\,, \quad \text{and} \quad \mathbb{P}(B \mid C \cap A) = \mathbb{P}(B \mid C) \tag{1.4}$$

and vice-versa.

Exercise 1.10. The intuitive "sure-thing principle" says that if

$$\mathbb{P}(A \mid C) \geq \mathbb{P}(B \mid C) \quad \text{and} \quad \mathbb{P}(A \mid C^c) \geq \mathbb{P}(B \mid C^c)$$

then $\mathbb{P}(A) \geq \mathbb{P}(B)$. Is this always true?

Exercise 1.11. We say that an event A is *favorable* to an event B, denoted by $A \parallel B$, provided $\mathbb{P}(A \cap B) \geq \mathbb{P}(A)\,\mathbb{P}(B)$. It follows from the definition that independent events are favorable to each other. Clearly the binary relation '$\parallel$' is symmetric, but not transitive (why?). Show that if $A \parallel B$, then $A^c \parallel B^c$, but $A \not\parallel B^c$ and $A^c \not\parallel B$.

An exercise in conditioning—Eddington's controversy

If A, B, C, and D each speak the truth
once in three times (independently)
and A claims that B denies that C declares that D is a liar,
what is the probability that D was speaking the truth?

Sir Arthur Stanley Eddington, who posed this problem in 1935, got the answer wrong. The correct answer is $\frac{13}{41}$, whereas Sir Eddington gave an answer to this problem as $\frac{25}{71}$. Thus this exercise has two objectives: First to derive the correct answer, and second (more difficult) to resolve the controversy, i.e., to reconstruct the argument that led Sir Eddington to his answer. (the second part might be very time consuming, so don't).

1.5. **Counting.** The starting point here is the factorial n, which equals the number of distinct permutations of n objects. Quite useful in probability is Stirling's approximation:

$$n! \approx n^n \mathrm{e}^{-n} \sqrt{2\pi n}\,,$$

which is also sometimes written in the form $\ln(n!) = n\ln(n) - n$. The latter can be easily derived by noting that

$$\ln(n!) = \sum_{k=1}^{n} \ln(k) \approx \int_1^n \ln(x)\,\mathrm{d}x = n\ln(n) - n + 1 \approx n\ln(n) - n\,.$$

It can also be derived from the integral definition of the factorial:

$$n! = \int_0^\infty \mathrm{e}^{-x} x^n\,\mathrm{d}x\,.$$

Recall the definition of the binomial coefficient in (1.1). Note then, that $\frac{n!}{(n-k)!}$ can be viewed as the number of distinct words of length k which can be formed from an alphabet of n letters. The binomial coefficient exhibits the symmetry $\binom{n}{k} = \binom{n}{n-k}$ and characterizes Pascal's triangle via its property

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}\,. \tag{1.5}$$

A useful identity is the *binomial formula*

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^n y^{n-k}, \tag{1.6}$$

holding for all real numbers x and y. Setting $x = y = 1$, (1.6) reduces to (1.2), while with $x = -y = -1$, we obtain

$$\sum_{k=0}^{n} (-1)^k \binom{n}{k} = 0 \quad \Longrightarrow \quad \sum_{k=\text{even}} \binom{n}{k} = \sum_{k=\text{odd}} \binom{n}{k} = 2^{n-1}.$$

We have obtained in class a probabilistic proof of the identity (when $k \le \min\{m,n\}$)

$$\binom{m}{0}\binom{n}{k} + \binom{m}{1}\binom{n}{k-1} + \cdots + \binom{m}{k}\binom{n}{0} = \binom{m+n}{k}\,.$$

From the Pascal triangle, it easily follows that

$$\sum_{k=0}^{n} \binom{m+k}{m} = \binom{m+n+1}{m+1}\,.$$

Try your luck with a probabilistic proof of

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}\,. \tag{1.7}$$

Apply (1.6) for an easy derivation of

$$\sum_{k=1}^{n} k\binom{n}{k} = n2^{n-1}\,, \quad \text{and} \quad \sum_{k=1}^{n}(-1)^{k+1}k\binom{n}{k} = 0\,. \tag{1.8}$$

If A is a set of cardinality n then the number of partitions $\{A_1, \ldots, A_r\}$ of A, satisfying $|A_i| = n_i$, is equal to

$$\frac{n!}{n_1!\, n_2! \cdots n_r!}\,, \qquad n_1 + n_2 + \cdots + n_r = n\,. \tag{1.9}$$

The number in (1.9) is referred to as the *multinomial coefficient* with parameters $n, n_1, \ldots, n_r$. For any real numbers $x_1, \ldots, x_r$, the *multinomial formula* applies

$$(x_1 + x_2 + \cdots + x_r)^n = \sum_{\substack{n_1,\ldots,n_r \\ n_1+\cdots+n_r=n}} \frac{n!}{n_1!\, n_2! \cdots n_r!}\, x_1^{n_1} x_2^{n_2} \cdots x_r^{n_r}\,. \tag{1.10}$$

Exercise 1.12. Use (1.10) to calculate the number of anagrams of the word "independence."

Exercise 1.13. Show that

$$\binom{m}{n} = \sum_{j=n-m+k}^{k\wedge n} \binom{k}{j}\binom{m-k}{n-j}\,, \qquad \text{for } k \le n \le m\,.$$

$$\binom{m}{n} = \sum_{j=1}^{n}\binom{m-j}{n-j} = \sum_{j=1}^{n}\binom{m-j}{m-n}\,.$$

Exercise 1.14. Two players play a series of independent games, in which player A has probability p, and player B has probability $q = 1 - p$ of winning each game. Suppose that A needs m and B needs n more games to win the series. Show that the probability that A will win is given by either one of the expressions:

$$\sum_{k=m}^{m+n-1}\binom{m+n-1}{k}p^k q^{m+n-k-1}$$

$$\sum_{k=0}^{n-1}\binom{m+k-1}{k}p^m q^k$$

Exercise 1.15. On Wednesday, 29th July, 1654 Pascal wrote to Fermat a letter, containing the following passage (in translation):

"I have no time to send you the proof of a difficult point which astonished M. (de Méré) so greatly, for he has ability, but he is not a geometer (which is, as you know, a great defect) and does not even comprehend that a mathematical line is infinitely divisible and he is firmly convinced that it is composed of a finite number of points. I have never been able to get him out of it. If you could do so, he can be perfected.

He tells me then that he has found an error in the numbers for this reason: If one undertakes to throw a six with a die, there is an advantage in undertaking to do it in four rolls: 671 to 625. If one undertakes to throw double sixes with two dice, there is a disadvantage in trying to do it in 24 throws. But nonetheless, 24

is to 36 (which is the number of faces of two dice) as 4 is to 6 (which is the number of faces on one die). This is what was his great scandal which made him say haughtily that the theorems were not consistent and that arithmetic was demented. But you will easily see the reason by the principles which you have. I shall put all that I have done with this in order when I shall have finished the treatise on geometry on which I have already been working for some time."

Notwithstanding the fallacious mathematical argument, it is impressive that Chevalier de Méré could discern the very narrow odds of getting double sixes in 24 throws based only on empirical data, even with his long experience at gambling tables. Incidentally this letter together with an earlier one is regarded as founding the theory of probability.

Today you can resolve this question in a couple of minutes— but not without your calculator. But what exactly was the fallacy behind de Méré's reasoning?

The number e

We should always keep in mind the two fundamental definitions of the number e, and be ready to use them:

$$\mathrm{e}^z \triangleq 1 + \frac{z}{1!} + \frac{z^2}{2!} + \cdots + \frac{z^k}{k!} + \cdots$$

and

$$\mathrm{e}^z \triangleq \lim_{n\to\infty} \left(1 + \frac{z}{n}\right)^n .$$

Note that from the solution of the "key matching problem" we get the somewhat surprising result that even if we tried to match randomly one million keys to their hooks, the probability of not getting even a single one right is e^{-1}, and does not vanish as the number of keys $n \to \infty$. If we calculate the probability p_k that exactly k keys are matched, we obtain

$$p_k = \frac{1}{k!}\left(1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + \cdots + \frac{(-1)^{n-k}}{(n-k)!}\right) ,$$

and hence $p_k \approx \frac{\mathrm{e}^{-1}}{k!}$ for large n.

2. Discrete Random Variables

We start by describing a class of binary-valued functions, called "indicator functions" which classify $\omega \in \Omega$ by means of the dichotomy "to be or not to be a member of an event A." This notion can be extended to define "random variables." Linear combinations of indicator functions are called "simple functions" and we can adequately approximate random variables with these. Quite often if a property or theorem can be demonstrated for simple functions, it holds true for random variables in general. This is a very useful device in probability.

2.1. **Indicator functions.** We start with a given probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, and define the *indicator function of the set* $A \in \mathfrak{F}$ by

$$\mathbb{I}_A(\omega) \triangleq \begin{cases} 1 & \text{if } \omega \in A\,, \\ 0 & \text{if } \omega \notin A\,. \end{cases}$$

Recall that for two number a and b $a \vee b$ stands for their maximum and $a \wedge b$ for their minimum. It is evident from the definition that $\mathbb{I}_{A\cup B}(\omega) = \mathbb{I}_A(\omega) \vee \mathbb{I}_B(\omega)$ and since this holds for all $\omega \in \Omega$, we simply write it as

$$\mathbb{I}_{A\cup B} = \mathbb{I}_A \vee \mathbb{I}_B$$

Note also that if $A \cap B = \varnothing$, then $\mathbb{I}_{A\cup B} = \mathbb{I}_A + \mathbb{I}_B$. The following simple formulas describe the salient properties of indicator functions

$$\begin{aligned} \mathbb{I}_{A\cap B} &= \mathbb{I}_A \cdot \mathbb{I}_B = \mathbb{I}_A \wedge \mathbb{I}_B \\ \mathbb{I}_{A^c} &= 1 - \mathbb{I}_A \end{aligned}$$

Thus, we can get expressions as

$$\begin{aligned} \mathbb{I}_{A\cup B} &= 1 - \mathbb{I}_{A^c\cap B^c} \\ &= 1 - \mathbb{I}_{A^c} \cdot \mathbb{I}_{B^c} \\ &= 1 - (1 - \mathbb{I}_A)(1 - \mathbb{I}_B)\,. \end{aligned}$$

For any real numbers a and b, $a + b = (a \vee b) + (a \wedge b)$. Therefore

$$\mathbb{I}_{A\cup B} + \mathbb{I}_{A\cap B} = \mathbb{I}_A + \mathbb{I}_B\,.$$

Also using *modulo* 2 arithmetic, we obtain

$$\begin{aligned} \mathbb{I}_{A\triangle B} &= \mathbb{I}_{A\cup B} - \mathbb{I}_{A\cap B} = \mathbb{I}_A + \mathbb{I}_B - 2\,\mathbb{I}_A \cdot \mathbb{I}_B \\ &= \mathbb{I}_A + \mathbb{I}_B \mod 2 \end{aligned}$$

It is evident that two events are equal if and only if their indicator functions are equal. Therefore, indicator functions can prove very handy in operations with sets.

Example 2.1. To show $(A \triangle B) \triangle C = A \triangle (B \triangle C)$, we can use modulo 2 arithmetic to obtain

$$\begin{aligned} \mathbb{I}_{(A\triangle B)\triangle C} &= \mathbb{I}_{A\triangle B} + \mathbb{I}_C = (\mathbb{I}_A + \mathbb{I}_B) + \mathbb{I}_C \\ &= \mathbb{I}_A + (\mathbb{I}_B + \mathbb{I}_C) = \mathbb{I}_A + \mathbb{I}_{B\triangle C} = \mathbb{I}_{A\triangle(B\triangle C)} \mod 2\,. \end{aligned}$$

So algebra can sometimes be much simpler than a Ven diagram!

2.2. **Discrete random variables (definitions).** A discrete random variable is essentially a real-valued function defined on Ω, that takes values in a discrete set. A set is called *discrete* if its elements can be enumerated (i.e., is a countable set). In order to simplify the notation, we assume without loss of generality that this discrete set is the set of integers $\mathbb{Z}$. However note that the set of rational numbers $\mathbb{Q}$ is also a discrete set. A formal definition is as follows:

Definition 2.1. A discrete random variable (d.r.v.) X is a function $X\colon \Omega \to \mathbb{Z}$, such that $\{\omega \in \Omega : X(\omega) = k\} \in \mathfrak{F}$, for all $k \in \mathbb{Z}$.

Remark 2. The second requirement in the definition is essential. Since the probability $\mathbb{P}$ is defined for events only, we wouldn't be able to even answer the question what is the probability that X takes the value k, unless the set of all $\omega \in \Omega$ such that $X(\omega) = k$ is an event.

Note that if $k_1, \ldots, k_n$ are integers, then

$$\big\{\omega \in \Omega : X(\omega) \in \{k_1, \ldots, k_n\}\big\} = \bigcup_{i=1}^{n} \{\omega \in \Omega : X(\omega) = k_i\}\,,$$

and therefore the set of all ω for which X takes a certain set of values is also an event.

It is beneficial at this point to introduce the following definition. If $C \subset \mathbb{Z}$, then

$$X^{-1}(C) \triangleq \{\omega \in \Omega : X(\omega) \in C\}\,.$$

It is evident that $X^{-1}(\mathbb{Z}) = \Omega$, and it is also straightforward to verify the following properties:

$$\begin{aligned} X^{-1}(C^c) &= \big(X^{-1}(C)\big)^c \\ X^{-1}\left(\cup_i C_i\right) &= \cup_i X^{-1}(C_i) \\ X^{-1}\left(\cap_i C_i\right) &= \cap_i X^{-1}(C_i)\,. \end{aligned}$$

This shows that the collection of all sets of the form $X^{-1}(C)$ is closed under complements and countable unions and intersections and contains Ω. Therefore is is a σ-field of events, which we denote by $\mathfrak{F}^X$ and refer to it as *the events generated by the random variable* X.

Henceforth, we simplify the notation as follows:

$$\{X = k\} \equiv \{\omega \in \Omega : X(\omega) = k\} = X^{-1}(\{k\})\,,$$

and more generally $\{X \in C\} \equiv \{\omega \in \Omega : X(\omega) \in C\} = X^{-1}(C)$.

It is clear that if $C_i \subset \mathbb{Z}$ are pairwise disjoint sets, then the events $X^{-1}(C_i)$ are disjoint. Hence for any set of distinct integers $\{k_1, k_2, \ldots\}$

$$\mathbb{P}\big(X \in \{k_1, k_2, \ldots\}\big) = \sum_i \mathbb{P}(X = k_i)\,. \tag{2.1}$$

Probability theory is not concerned with the exact function $X(\omega)$, since this is never really known. What is of interest about a d.r.v. are the values which it might take and the probabilities associated with these values, which we refer to as the *distribution* or *law* of the d.r.v. It then follows by (2.1) that all we need to know are the numbers $\mathbb{P}(X = k)$, for $k \in \mathbb{Z}$. This motivates the definition of the *probability mass function* (PMF) p_X of the d.r.v. X, as $p_X(k) \triangleq \mathbb{P}(X = k)$.

Conversely given a set of non-negative numbers π_k, $k \in \mathbb{Z}$, satisfying $\sum_i \pi_i = 1$, there exists a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ and a d.r.v. X defined on it, such that $p_X(k) = \pi_k$. [1]

We list some classes of d.r.v.s which occur frequently (in these definitions $q \equiv 1 - p$):
Bernoulli distribution (with parameter p). X is in $\{0, 1\}$, and $p_X(1) = p$, $p_X(0) = q$.

[1] This is actually very easy to prove—try it.

Binomial distribution (with parameters n and p). X takes values in $\{0, 1, \dots, n\}$, and

$$p_X(k) = \binom{n}{k} p^k q^{n-k}\,, \quad k = 0, 1, \dots, n\,.$$

Poisson distribution (with parameter $\lambda > 0$). X takes values in the non-negative integers $\mathbb{Z}_+$, and

$$p_X(k) = \frac{1}{k!}\lambda^k \mathrm{e}^{-\lambda}\,, \quad k = 0, 1, 2, \dots$$

Geometric distribution (with parameter $p > 0$). X takes values in the natural numbers $\mathbb{N}$, and

$$p_X(k) = pq^{k-1}\,, \quad k = 1, 2, 3, \dots$$

Negative binomial distribution (with parameters n and p). X takes values in $\{n, n+1, n+2, \dots\}$, and

$$p_X(k) = \binom{k-1}{n-1} p^n q^{k-n}\,, \quad k = n, n+1, n+2, \dots\,.$$

These distributions are related. If $X_1, X_2, X_3, \dots$ is a sequence of *independent* [2] Bernoulli d.r.v.s with parameter p [3], then S_n defined by $S_n(\omega) = X_1(\omega) + \cdots + X_n(\omega)$ is Binomial with parameters n and p. Also, provided $p > 0$, if we define

$$Z_n(\omega) = \min\,\{k \geq 1 : X_1(\omega) + X_2(\omega) + \cdots + X_k(\omega) = n\}\,,$$

i.e., the first time [4] that n successes occur in a sequence of independent Bernoulli trials, then Z_n has the negative binomial distribution with parameters n and p. Lastly, with fixed $\lambda > 0$, if n is very large, and $np = \lambda$, we have

$$\begin{aligned}\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} &\approx \frac{n^k}{k!}\left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &\approx \frac{1}{k!}\lambda^k \mathrm{e}^{-\lambda}\,.\end{aligned}$$

Note that

$$X(\omega) = \sum_k k\, \mathbb{I}_{\{X=k\}}(\omega)\,.$$

Functions of a d.r.v. Let $g : \mathbb{Z} \to \mathbb{Z}$. We define the d.r.v. Y, by $Y(\omega) = g(X(\omega))$, which we write in short as $Y = g(X)$. Then Y is a d.r.v. To prove that this is so, note that for any set $C \in \mathbb{Z}$, $Y^{-1}(C) = X^{-1}\big(g^{-1}(C)\big)$. Therefore, since $g^{-1}(C) \subset \mathbb{Z}$, then $X^{-1}\big(g^{-1}(C)\big) \in \mathfrak{F}$.

Also note that

$$p_Y(k) = \mathbb{P}(g(X) = k) = \mathbb{P}\big(X \in g^{-1}(\{k\})\big) = \sum_{j \in g^{-1}(\{k\})} p_X(j)\,. \tag{2.2}$$

[2] it may appear that we haven't defined independent d.r.v.s, but a general definition is that X and Y are independent if $\mathfrak{F}^X$ and $\mathfrak{F}^Y$ are so.

[3] these are often referred to as *Bernoulli trials*

[4] note that this is a random time

2.3. **Expectation.** We define the *expectation* (or *mean*, or *first moment*) of a d.r.v. X by

$$\mathbb{E}[X] \triangleq \sum_{k\in\mathbb{Z}} k\, p_X(k)\,, \tag{2.3}$$

provided $\sum_{k\in\mathbb{Z}}|k|p_X(k) < \infty$, othewise we say that X does not exist. We extend this definition to *higher moments* as follows. Provided $\sum_{k\in\mathbb{Z}}|k|^m p_X(k) < \infty$, we say that X has an m^{th} moment and we denote this

$$\mathbb{E}[X^m] = \sum_{k\in\mathbb{Z}} k^m p_X(k)\,,$$

In general if $Y = g(X)$ then we have the following important property

$$\mathbb{E}[Y] = \sum_{k\in\mathbb{Z}} g(k) p_X(k)\,, \tag{2.4}$$

provided $\sum_{k\in\mathbb{Z}}|g(k)|p_X(k) < \infty$. The proof of (2.3) is simple. Use (2.2) and the definition of expectation in (2.3) to obtain

$$\begin{aligned}
\mathbb{E}[Y] = \sum_{k\in\mathbb{Z}} k\, p_Y(k) &= \sum_{k\in\mathbb{Z}} k \sum_{j\in g^{-1}(\{k\})} p_X(j) \\
&= \sum_{k\in\mathbb{Z}} \left(\sum_{j\in g^{-1}(\{k\})} k\, p_X(j) \right) \\
&= \sum_{k\in\mathbb{Z}} \left(\sum_{j\in g^{-1}(\{k\})} g(j)\, p_X(j) \right) \\
&= \sum_{j\in \cup_k\, g^{-1}(\{k\})} g(j)\, p_X(j) = \sum_{j\in\mathbb{Z}} g(j)\, p_X(j)\,.
\end{aligned}$$

Example 2.2. A PMF which does not have a mean is $p(k) = \frac{6}{k^2\pi^2}$, for $k = 1, 2, 3, \ldots$ The reason for this is that $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$. Note that this is indeed a PMF since $\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$.

A random variable is called *centered*, if it has zero mean. Note that every random variable can be centered by subtracting it's mean, i.e., $X - \mathbb{E}[X]$. The *variance* of X is defined as the second moment of $X - \mathbb{E}[X]$. That is

$$\text{var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \sum_k \mathbb{E}\big[(k - \mathbb{E}[X])^2\big] p_X(k)\,.$$

By expanding the square we get

$$\text{var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2 - 2X\,\mathbb{E}[X] + (\mathbb{E}[X])^2\big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2\,.$$

Therefore, *the variance is equal to the second moment less the square of the first moment*. The square-root of the variance is called the *standard deviation* and is denoted by σ_X. Thus, if we denote the mean by μ_X then $\text{var}(X) = \sigma_X^2 = \mathbb{E}[X^2] - m_X^2$.

Exercise 2.1. Suppose $Y = aX + b$ where a and b are constants? What is the variance and mean of Y. Answer: $\sigma_Y^2 = a^2\sigma_X^2$, and $\mu_Y = a\mu_X + b$.

Exercise 2.2. What is $\min_a \ \mathbb{E}[(X-a)^2]$? Answer: $\mathrm{var}(X)$, since

$$\mathbb{E}[(X-a)^2] = \mathrm{var}(X-a) + (\mathbb{E}[X-a])^2 = \sigma_X^2 + (\mathbb{E}[X-a])^2 .$$

Mean and variance of some common distributions

Bernoulli distribution (with parameter p): $\mu_X = p$, $\sigma_X^2 = pq$

Binomial distribution (with parameters n and p): $\mu_X = np$, $\sigma_X^2 = npq$.

Poisson distribution (with parameter $\lambda > 0$): $\mu_X = \lambda$, $\sigma_X^2 = \lambda$.

Geometric distribution (with parameter $p > 0$): $\mu_X = \frac{1}{p}$, $\sigma_X^2 = \frac{q}{p^2}$.

Here we use the series $\sum_{k\geq 1} kq^{k-1} = \frac{1}{(1-q)^2}$, and $\sum_{k\geq 1} k^2 q^{k-1} = \frac{1+q}{(1-q)^3}$.

Next we state a very important theorem.

Theorem 1. *Let X be a non-negative, integer-valued random variable. Then*

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} \mathbb{P}(X \geq k) .$$

Proof.

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} k\, p_X(k) = \sum_{k=1}^{\infty}\sum_{j=1}^{k} p_X(k) = \sum_{j=1}^{\infty}\sum_{k=j}^{\infty} p_X(k) = \sum_{j=1}^{\infty} \mathbb{P}(X \geq j) . \qquad \square$$

This technique is called *Abel summation*.

Theorem (Markov Inequality)**.** *If X is a non-negative d.r.v. with finite mean, then for any $n \in \mathbb{N}$,*

$$\mathbb{P}(X \geq n) \leq \frac{\mathbb{E}[X]}{n} .$$

Proof.

$$\begin{aligned} \mathbb{E}[X] = \sum_{k=1}^{\infty} k\, p_X(k) &= \sum_{k=1}^{n-1} k\, p_X(k) + \sum_{k=n}^{\infty} k\, p_X(k) \\ &\geq \sum_{k=n}^{\infty} k\, p_X(k) \\ &\geq n \sum_{k=n}^{\infty} p_X(k) = n\, \mathbb{P}(X \geq n) . \end{aligned} \qquad \square$$

Corollary (Chebyshev's Inequality)**.** *If X has mean μ_X and finite variance σ_X^2 then*

$$\mathbb{P}(|X - \mu_X| \geq n) \leq \frac{\sigma_X^2}{n^2} .$$

A function g is called convex if $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$, for all $\lambda \in [0,1]$.

Theorem (Jensen's Inequality)**.** *If g is a non-negative convex function, then*

$$g\big(\mathbb{E}[X]\big) \leq \mathbb{E}[g(X)]\,.$$

Exercise 2.3. A fair die having two faces coloured blue, two red, and two green, is thrown repeatedly. Find the expected value of the minimum number of throws necessary for all colours to appear.

Solution: The probability that not all colours appear in the first k throws, is equal to 1 for $k = 1, 2$, and for $k \geq 3$, equals

$$3\left(\frac{2}{3}\right)^k - 3\left(\frac{1}{3}\right)^k\,.$$

Thus, if N is a random variable that takes the value k if all three colours occur in the first k throws, but only two of the colours in the first $k-1$ throws, then

$$\mathbb{E}[N] = \sum_{k=1}^{\infty} \mathbb{P}(N \geq k) = 2 + \sum_{k=3}^{\infty}\left[3\left(\frac{2}{3}\right)^k - 3\left(\frac{1}{3}\right)^k\right] = \frac{11}{2}\,.$$

Exercise 2.4. Is it ever true that $\mathbb{E}[\frac{1}{X}] = \frac{1}{\mathbb{E}[X]}$?

Solution: Yes, for example if $p_X(-1) = \frac{1}{9}$, $p_X(\frac{1}{2}) = \frac{4}{9}$, and $p_X(2) = \frac{4}{9}$.

Exercise 2.5. You roll a conventional die repeatedly. If it shows 1, you must stop, but you may choose to stop at any prior time. Your score is the outcome on the final roll. What is an optimal stopping strategy, based on the most recent roll?

Solution: Consider the strategy: stop the first time the die shows k or greater. If $S(k)$ is the expected score of this strategy, then

$$S(6) = 6 \cdot \mathbb{P}(6 \text{ appears before } 1) + 1 \cdot \mathbb{P}(1 \text{ appears before } 6) = \frac{7}{2}\,.$$

Similarly $S(5) = S(4) = 4$, $S(3) = 3.8$, $S(2) = 3.5$. So, best strategy is to stop at 4 or higher.

Exercise 2.6. If $\mathcal{G} = (V, E)$ is a finite graph (V is the set of vertices and E is the set of edges), show that there exists a subset $\tilde{V} \subset V$ such that at least half of the number edges $|E|$ connect $\tilde{V}$ to $\tilde{V}^c$.

Solution: Color the vertices red or green, independently, with equal probability, and let $\tilde{V}$ be the random set of red vertices. Let $\mathcal{I}_{\tilde{V}}(e) = 1$ if e connects $\tilde{V}$ to $\tilde{V}^c$, and 0 othewise. Set $N_{\tilde{V}} \triangleq \sum_{e\in E} \mathcal{I}_{\tilde{V}}(e)$. Then $\mathbb{E}[N_{\tilde{V}}] = \sum_{e\in E} \mathbb{E}[\mathcal{I}_{\tilde{V}}(e)] = \frac{1}{2}|E|$. Since the mean is $\frac{1}{2}|E|$, at least one realization, i.e., $N_{\tilde{V}}(\omega)$ for some ω, is not less than this value.

Exercise 2.7. It X, Y and Z are three d.r.v.s with *distinct* values, show that $\min\big\{\mathbb{P}(X > Y), \mathbb{P}(Y > Z), \mathbb{P}(Z > X)\big\} \leq \frac{2}{3}$. Show an example where the inequality is equality. This is related to the observation that, in an election, it is possible for more than half of the voters to prefer candidate X to Y, more than half Y to Z, and more than half Z to X (voter paradox).

Solution: Work with indicator functions

$$\mathbb{P}(X > Y) + \mathbb{P}(Y > Z) + \mathbb{P}(Z > X) = \mathbb{E}[\mathbb{I}_{\{X>Y\}} + \mathbb{I}_{\{Y>Z\}} + \mathbb{I}_{\{Z>X\}}] \leq 2\,,$$

from which the result follows. Equality is attained for example, if the joint PMF is of the form $p(2,1,3) = p(3,2,1) = p(1,3,2) = \frac{1}{3}$.

Exercise 2.8. Urn R contains n red balls and urn B contains n black balls. At each step, a ball is selected at random from each urn, and they are swapped. What is the mean number of red balls in urn R after the k^{th} step? (this is a diffusion model introduced by Bernoulli)

Solution: Any given red ball is in urn R after step k, if it has been selected an even number of times. So the mean number is n times this probability, which is

$$n \sum_{m \text{ even}} \binom{k}{m} \left(\frac{1}{n}\right)^m \left(1 - \frac{1}{n}\right)^{k-m} = \frac{n}{2}\left[1 + \left(\frac{n-2}{n}\right)^k\right].$$

Exercise 2.9. If X, Y are d.r.v.s, show that $\mathbb{E}[\log(p_X(X))] \geq \mathbb{E}[\log(p_Y(X))]$. Also, show that the *mutual information*

$$\mathcal{I} \triangleq \mathbb{E}\left[\log\left(\frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right)\right]$$

satisfies $\mathcal{I} \geq 0$, with equality if and only if X and Y are independent.

Solution: We use the fact that $\log(z) \leq z - 1$ with equality if and only if $z = 1$. Thus,

$$\mathbb{E}\left[\log\left(\frac{p_Y(X)}{p_X(X)}\right)\right] \leq \mathbb{E}\left[\frac{p_Y(X)}{p_X(X)} - 1\right] = 0\,,$$

with equality only if $p_Y(X) = p_X(X)$. Second part is similarly obtained.

Exercise 2.10. Put n tokens into m boxes.

(a) What is the expected number of boxes which get k tokens?
(b) What is the expected number of tokens which do not share a box with any other tokens?

Solution: For part (a), we get

$$m\binom{n}{k}\frac{(m-1)^{n-k}}{m^n}\,, \qquad \text{and for part (b):} \quad n\left(\frac{m-1}{m}\right)^{n-1}$$

2.4. **Jointly distributed random variables.** Let X and Y be d.r.v.s defined on $(\Omega, \mathfrak{F}, \mathbb{P})$. We define their *joint PMF* by

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)\,,$$

and in general for d.r.v.s $X_1, X_2, \ldots, X_n$,

$$p_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)\,,$$

The *marginal* PMFs of X and Y can be obtained by

$$p_X(x) = \sum_y p_{X,Y}(x,y)\,, \qquad p_Y(y) = \sum_x p_{X,Y}(x,y)\,.$$

If $g\colon \mathbb{R}\times\mathbb{R}\to\mathbb{R}$ is a function, then we define $Z = g(X,Y)$, as $Z(\omega) = g(X(\omega),Y(\omega))$ and analogously for functions of more than two variables. If Z has finite mean, then

$$\mathbb{E}[Z] = \mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y)\,.$$

We view '$\mathbb{E}$' as an operator on random variables. Its most important property is that it is a *linear* operator, i.e.,

$$\mathbb{E}[aX+bY] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]\,.$$

Theorem (Cauchy-Schwartz inequality)**.** *For any d.r.v.s X and Y, with finite second moments,*

$$\big(\mathbb{E}[XY]\big)^2 \le \mathbb{E}[X^2]\;\mathbb{E}[Y^2]\,,$$

and the inequality is strict unless $Y = aX$, or $X = aY$, for some constant a.

Proof. Suppose, without loss of generality that $\mathbb{E}[X^2]\neq 0$. Since

$$\mathbb{E}[(aX-Y)^2] = a^2\,\mathbb{E}[X^2] - 2a\,\mathbb{E}[XY] + \mathbb{E}[Y^2]\,, \tag{2.5}$$

is non-negative for all a, the quadratic polynomial in a in (2.5) must have a non-positive discriminant, i.e., $\big(\mathbb{E}[XY]\big)^2 - \mathbb{E}[X^2]\;\mathbb{E}[Y^2]\le 0$. If the discriminant is zero, then for some real number a, $aX = Y$. □

Definition 2.2. Define the *covariance* $\operatorname{cov}(X,Y)$ and the *correlation coefficient* $\rho(X,Y)$ by

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]\,,\quad \text{and}\quad \rho(X,Y) \triangleq \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}\,,$$

provided $\operatorname{var}(X)\operatorname{var}(Y)\neq 0$. By the Cauchy-Schwartz inequality, $|\rho(X,Y)|\le 1$, with equality if and only if $\mathbb{P}(aX+bY=c)=1$, for some constants a, b, c (not all identically zero). When $\rho(X,Y)=0$, we say that X and Y are *uncorrelated*, and this happens Note that $\operatorname{cov}(X,Y) = \mathbb{E}[XY]-\mu_X\mu_Y$, and if X and Y are uncorrelated then $\mathbb{E}[XY]=\mu_X\mu_Y$.

Exercise 2.11. Let X, Y be d.r.v.s, and define $V \triangleq \min(X,Y)$ and $U \triangleq \max(X,Y)$. Find the PMFs of V and U.

Solution: First compute

$$\mathbb{P}(V\ge k) = \mathbb{P}(X\ge k, Y\ge k) = \sum_{x\ge k,\,y\ge k} p_{X,Y}(x,y)\,,$$

and then use $p_V(k) = \mathbb{P}(V\ge k) - \mathbb{P}(V\ge k+1)$. Similarly,

$$\mathbb{P}(U\le k) = \mathbb{P}(X\le k, Y\le k) = \sum_{x\le k,\,y\le k} p_{X,Y}(x,y)\,,$$

and then use $p_U(k) = \mathbb{P}(V\le k+1) - \mathbb{P}(U\le k)$.

2.5. **Conditioning with d.r.v.s.**

Conditioning on an event. In general we define the *conditional PMF* of X given an event A, by

$$p_{X|A}(x) \triangleq \mathbb{P}(X = x \mid A)\,,$$

provided $\mathbb{P}(A) > 0$. It is evident that $p_{X|A}$ is itself a PMF and satisfies $\sum_x p_{X|A}(x) = 1$. Also, for a partition $A_1, \ldots, A_n$, $p_X(x) = \sum_{i=1}^n \mathbb{P}(A_i) p_{X|A_i}(x)$, and more general, if B is an event,

$$p_{X|B}(x) = \sum_{i=1}^n \mathbb{P}(A_i \mid B) p_{X|A_i \cap B}(x)\,.$$

We also define *the conditional expectation of X given A* by

$$\mathbb{E}[X \mid A] \triangleq \sum_x x\ \mathbb{P}(X = x \mid A)\,.$$

For any partition $A_1, \ldots, A_n$, we have:

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X \mid A_i]\, \mathbb{P}(A_i)\,, \tag{2.6}$$

and this also holds true for non-finite partitions, i.e., a sequence $A_1, A_2, \ldots$ of pairwise disjoint events, whose union is Ω provided $\sum_i \big|\mathbb{E}[X \mid A_i]\big|\, \mathbb{P}(A_i) < \infty$.

Exercise 2.12. A coin is tossed repeatedly and heads appears at each toss with probability $p \in (0, 1)$. Find the expected length of the initial run (this is the run of heads if the first toss gives heads, and of tails otherwise).

Solution: Let H be the event that the first toss gives heads, and X the length of the initial run. It is easy to see that

$$\mathbb{P}(X = k \mid H) = p^{k-1} q\,, \quad \text{and} \quad \mathbb{P}(X = k \mid H^c) = q^{k-1} p\,.$$

Thus

$$\mathbb{E}[X \mid H] = \sum_{k \geq 1} k p^{k-1} q = \frac{q}{(1-p)^2} = \frac{1}{q}\,,$$

and, similarly, $\mathbb{E}[X \mid H^c] = \frac{1}{p}$. Thus

$$\mathbb{E}[X] = \frac{p}{q} + \frac{q}{p} = \frac{1}{pq} - 2\,.$$

Definition 2.3. Define the *conditional PMF of X given $Y = y$*

$$p_{X|Y}(x \mid y) \triangleq \frac{p_{X,Y}(x, y)}{p_Y(y)} = \mathbb{P}(X = x \mid Y = y)\,.$$

The *conditional expectation of a function $g(X)$ of X given $Y = y$* is defined by

$$\mathbb{E}[g(X) \mid Y = y] \triangleq \sum_x g(x) p_{X|Y}(x \mid y)\,,$$

provided $\mathbb{E}[|g(X)|] < \infty$. We also define the *conditional expectation of X given Y* as a *random variable* which is a function of Y and is given by

$$\mathbb{E}[X \mid Y] \triangleq \psi(Y)\,, \quad \text{where} \quad \psi(y) \triangleq \mathbb{E}[X \mid Y = y]\,.$$

Theorem 2. *The conditional expectation satisfies $\mathbb{E}\big[\mathbb{E}[X \mid Y]\big] = \mathbb{E}[X]$, and more generally, for any function g such that $\mathbb{E}[Xg(Y)]$ exists,*

$$\mathbb{E}\big[\mathbb{E}[X \mid Y]g(Y)\big] = \mathbb{E}[Xg(Y)]\,.$$

We also have the property

$$\mathbb{E}[Xg(Y) \mid Y] = g(Y)\,\mathbb{E}[X \mid Y]\,.$$

Exercise 2.13. A coin shows heads with probability p. Let X_n the number of flips needed to obtain a run of n consecutive heads. Calculate $\mathbb{E}[X_n]$.

Solution: Condition on X_{n-1}, to obtain

$$\begin{aligned}\mathbb{E}[X_n] &= \mathbb{E}\big[\mathbb{E}[X_n \mid X_{n-1}]\big] \\ &= \mathbb{E}[p(X_{n-1}+1) + (1-p)(X_{n-1}+1+\tilde{X}_n)\big]\,,\end{aligned}$$

where $\tilde{X}_n$ has the same distribution as X_n. Thus,

$$\mathbb{E}[X_n] = p^{-1}\,\mathbb{E}[X_{n-1}] + p^{-1}\,,$$

and since $\mathbb{E}[X_1] = p^{-1}$, solving we have

$$\mathbb{E}[X_n] = \sum_{k=1}^{n} \frac{1}{p^k}\,.$$

Exercise 2.14. A hen lays N eggs, where N has the Poisson distribution with parameter λ, and each egg hatches with probability $p \in (0,1)$, independently of the other eggs. Let X be the number of chicks. Find $\mathbb{E}[K \mid N]$ and $\mathbb{E}[K]$.

Solution: We are given

$$p_N(n) = \frac{\lambda^n}{n!}\mathrm{e}^{-\lambda}\,, \qquad p_{X|N}(x \mid n) = \binom{n}{x} p^x (1-p)^{n-x}\,.$$

Thus,

$$\psi(n) = \mathbb{E}[X \mid N = n] = \sum_x x\, p_{X|N}(x \mid n) = pn\,,$$

and we obtain $\mathbb{E}[X \mid N] = \psi(N) = pN$. Also, $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X \mid N]\big] = \mathbb{E}[pN] = p\lambda$.

Application: Best Estimator

What follows is very useful in estimation theory. The context here is trying to estimate a random variable X from a random observation Y. In other words, the types of estimates available are the functions of Y. As a criterion of how close our estimate is we use the second moment of the error, i.e., $\mathbb{E}\big[(X-\hat{X})^2\big]$, where $\hat{X}=g(Y)$ is the estimate of X. Therefore, we would like to minimize over all candidate functions g, the expression $\mathbb{E}\left[\big(X-g(Y)\big)^2\right]$. It turns out that:

$$\min_g \ \mathbb{E}\left[\big(X-g(Y)\big)^2\right] = \mathbb{E}\left[\big(X-\mathbb{E}[X\mid Y]\big)^2\right], \tag{2.7}$$

and hence the best estimate of X is $g(Y)=\mathbb{E}[X\mid Y]$.

Proof of (2.7). To prove (2.7), we write

$$\begin{aligned}\mathbb{E}\left[\big(X-g(Y)\big)^2\right] &= \mathbb{E}\left[\big(X-\mathbb{E}[X\mid Y]+\mathbb{E}[X\mid Y]-g(Y)\big)^2\right]\\ &= \mathbb{E}\left[\big(X-\mathbb{E}[X\mid Y]\big)^2\right]+\mathbb{E}\left[\big(\mathbb{E}[X\mid Y]-g(Y)\big)^2\right]\\ &\quad +2\,\mathbb{E}\Big[\big(X-\mathbb{E}[X\mid Y]\big)\big(\mathbb{E}[X\mid Y]-g(Y)\big)\Big]. \end{aligned}\tag{2.8}$$

We claim that the third term on the right-hand side of (2.8) vanishes identically. Indeed, using the properties of conditional expectation, and setting $h(Y)=\mathbb{E}[X\mid Y]-g(Y)$, we have

$$\begin{aligned}\mathbb{E}\Big[\big(X-\mathbb{E}[X\mid Y]\big)h(Y)\Big] &= \mathbb{E}\Big[\mathbb{E}\big[\big(X-\mathbb{E}[X\mid Y]\big)h(Y)\mid Y\big]\Big]\\ &= \mathbb{E}\Big[h(Y)\,\mathbb{E}\big[X-\mathbb{E}[X\mid Y]\mid Y\big]\Big]\\ &= \mathbb{E}\Big[h(Y)\big(\mathbb{E}[X\mid Y]-\mathbb{E}[X\mid Y]\big)\Big]=0\,.\end{aligned}$$

Hence, by (2.8)

$$\mathbb{E}\left[\big(X-g(Y)\big)^2\right] = \mathbb{E}\left[\big(X-\mathbb{E}[X\mid Y]\big)^2\right]+\mathbb{E}\left[\big(\mathbb{E}[X\mid Y]-g(Y)\big)^2\right],$$

and (2.7) now easily follows. □