A Brief Introduction to Probability Theory

Lecture Notes[1] for CPSC 411
Andreas Klappenecker

We collect in this chapter some basic notions and methods from probability theory. Our aim is to give a brief exposition of results that are repeatedly used in the analysis of randomized algorithms. The treatment is not exhaustive and cannot replace any standard text on probability theory. Typically, we will skip all proofs and merely remind the reader of basic definitions and useful theorems.

§1 Basic Definitions

If a coin is tossed, then there are two possible outcomes: heads and tails. Statements such as 'the chance that the outcome will be heads is 50%' are formalized in probability theory, giving a better way to reason about the odds of a certain outcome of an experiment.

The possible outcomes of an experiment are called the *sample space*. For example, the sample space of the coin tossing experiment is $\Omega = \{\text{head}, \text{tail}\}$. Certain subsets of the sample space are called *events*, and the probability of these events is determined by a *probability measure*.

For instance, if we roll a dice, then one of its six face values is the outcome of the experiment, so the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. An event is a subset of the sample space Ω. The event $\{1, 2\}$ occurs when the dice shows a face value less than three. The probability measures describes the odds that a certain event occurs, for instance $\Pr[\{1, 2\}] = 1/3$ means that the event $\{1, 2\}$ will occur with probability $1/3$.

A probability measure is not necessarily defined on all subsets of the sample space Ω, but just on all subsets of Ω that are considered events. Nevertheless, we want to have a uniform way to reason about the probability of events. This is accomplished by requiring that the collection of events form a σ-algebra. A σ-algebra $\mathcal{F}$ is a collection of subsets of the sample space Ω such that the following requirements are satisfied:

S1 The empty set is contained in $\mathcal{F}$.

S2 If a set E is contained in $\mathcal{F}$, then its complement E^c is contained in $\mathcal{F}$.

S3 The countable union of sets in $\mathcal{F}$ is contained in $\mathcal{F}$.

Henceforth we will assume that the collection of events forms a σ-algebra. This allows to talk about the complementary event, and the union of events. The empty set is often called the *impossible event*. The sample space Ω is the complement of the empty set, hence is contained in $\mathcal{F}$. The event Ω is called the *certain event*.

Let $\mathcal{F}$ be a σ-algebra over the sample space Ω. A probability measure on $\mathcal{F}$ is a function $\Pr\colon \mathcal{F} \to [0,1]$ satisfying

P1 The certain event satisfies $\Pr[\Omega] = 1$.

P2 If the events $E_1, E_2, \ldots$ in $\mathcal{F}$ are mutually disjoint, then

$$\Pr[\bigcup_{k=1}^{\infty} E_k] = \sum_{k=1}^{\infty} \Pr[E_k].$$

These axioms have a number of familiar consequences. For example, it follows that the complementary event E^c has probability $\Pr[E^c] = 1 - \Pr[E]$. In particular, the impossible event has probability zero, $\Pr[\emptyset] = 0$, as it should. Another consequence is a simple form of the inclusion-exclusion principle:

$$\Pr[E \cup F] = \Pr[E] + \Pr[F] - \Pr[E \cap F],$$

which is convenient when calculating probabilities.

Example 1 A dice has sample space $\Omega = \{1,2,3,4,5,6\}$. The four events $\textbf{impossible} = \emptyset, \textbf{head} = \{1,2\}, \textbf{tail} = \{3,4,5,6\}$, $\textbf{certain} = \{1,2,3,4,5,6\}$ form a σ-algebra $\mathcal{F}$. The event **head** occurs when the face value of the dice is less than three, and **tail** occurs otherwise. A probability measure is given by

$$\Pr[\textbf{impossible}] = 0,\ \Pr[\textbf{head}] = 1/3,\ \Pr[\textbf{tail}] = 2/3,\ \Pr[\textbf{certain}] = 1.$$

The events **head** and **tail** partition the sample space Ω. These events can be used to simulate a biased coin. Apart from the four subsets of Ω above, none of the other $2^6 - 4 = 60$ subsets of Ω is needed for this purpose.

A *probability space* is a triple $(\Omega, \mathcal{F}, \Pr)$ such that Ω is a sample space, $\mathcal{F}$ is a σ-algebra of events over Ω, and $\Pr$ is a probability measure on $\mathcal{F}$. We use this terminology to avoid lengthy descriptions.

Remark. The reader might wonder why not all subsets of the sample space are necessarily considered events. Indeed, if Ω is a countable set, then this is a viable option. However, if the sample space Ω is uncountable, such as

$\Omega = [0,1]$, then having all subsets of the interval as events makes it difficult to define a reasonable probability measure. The reason is that axiom **P2** imposes severe restrictions when the σ-algebra is big. Such measure-theoretic difficulties have been noticed by Borel, Ulam and others.

Exercise 1.1 *Let $\mathcal{F}$ be a σ-algebra. Show that the countable intersection of events in $\mathcal{F}$ is contained in $\mathcal{F}$.*

Exercise 1.2 *Let E and F be events contained in a σ-algebra $\mathcal{F}$. Show that the set-difference $E - F$ is contained in $\mathcal{F}$.*

Exercise 1.3 *Let E and F be events such that $E \subseteq F$. Show that*

$$\Pr[E] \leq \Pr[F].$$

Exercise 1.4 *Let $E_1, \ldots, E_n$ be events that are not necesarily disjoint. Show that*

$$\Pr[E_1 \cup \cdots \cup E_n] \leq \Pr[E_1] + \cdots + \Pr[E_n].$$

Exercise 1.5 *Let $E_1, E_2, \ldots, E_n$ be arbitrary events over a sample space Ω. Show that*

$$\Pr[\bigcup_{k=1}^{n} E_k] = \sum_{s=1}^{n} (-1)^{s+1} \sum_{k_1 < k_2 < \cdots < k_s} \Pr[\bigcap_{r=1}^{s} E_{k_r}]$$

holds. This is the so-called inclusion-exclusion principle. Hint: Consider first some small cases, such as $n = 2$ and $n = 3$, to get familiar with this formula.

Conditional Probabilities. Let E and F be events over a sample space Ω such that $\Pr[F] > 0$. The *conditional probability* $\Pr[E\,|F\,]$ of the event E given F is defined by

$$\Pr[E\,|F\,] = \frac{\Pr[E \cap F]}{\Pr[F]}.$$

The value $\Pr[E\,|F\,]$ is interpreted as the probability that the event E occurs, assuming that the event F occurs. By definition, $\Pr[E \cap F] = \Pr[E\,|F\,] \Pr[F]$, and this simple multiplication formula often turns out to be useful.

Exercise 1.6 *Suppose that $F_1, \ldots, F_n$ are events that partition the sample space Ω such that $\Pr[F_k] > 0$ holds for all k in the range $1 \leq k \leq n$. Show that*

$$\Pr[E\,] = \sum_{k=1}^{n} \Pr[E\,|F_k\,] \Pr[F_k\,]$$

holds for any event E. This fact is attributed to Reverend Thomas Bayes.

Keeping the assumptions of the previous exercise, and assuming that $\Pr[E] > 0$, we obtain as a consequence the so-called Bayes' rule

$$\Pr[F_\ell \,|E] = \frac{\Pr[F_\ell \cap E]}{\Pr[E]} = \frac{\Pr[E|F_\ell]\Pr[F_\ell]}{\sum_{k=1}^{n} \Pr[E|F_k]\Pr[F_k]},$$

which allows to compute the conditional probability $\Pr[F_\ell \,|E]$ when the probabilities $\Pr[E]$, $\Pr[F_k]$, and $\Pr[E \,|F_k]$ are known.

Example 2 We illustrate the virtue of conditional probabilities with the help of the notorious car and goats problem which got famous through the Monty Hall game show. At the end of this show, a contestant was shown three closed doors. She was told that behind one door is a new car, and behind the other two are goats. If the contestant chooses the door hiding the car, then she can keep the car, otherwise she has to marry the lucky goat.[2] Once she has made her choice, the game show host – knowing which door conceals the car – opens one of the other two doors to reveal a goat. Monty then asks her whether she would like to switch doors. Assuming that the contestant does not want to get married to a goat, the question is: Should she switch?

Without loss of generality, let us assume that the contestant has chosen door 1, Monty has opened door 2, and now the contestant has to choose between doors 1 and 3. Let C_1 denote the event that the car is behind door 1, C_3 the event that the car is behind door 3, and M_2 the event that Monty opened door 2, hence contains a goat.

It is apparent that $\Pr[C_1] = 1/3$ and $\Pr[C_3] = 1/3$. Assuming that Monty will choose a door at random if both doors conceal goats, we get $\Pr[M_2 \,|C_1\,] = 1/2$. We certainly have $\Pr[M_2 \,|C_3\,] = 1$, because Monty has no choice in this case. Recall that our goal is to compare the conditional probabilities $\Pr[C_1|M_2]$ and $\Pr[C_3|M_2]$. We can use Bayes' rule to determine these probabilities. Indeed,

$$\Pr[C_1|M_2] = \frac{\Pr[M_2|C_1]\Pr[C_1]}{\Pr[M_2|C_1]\Pr[C_1] + \Pr[M_2|C_3]\Pr[C_3]} = \frac{1/6}{1/6 + 1/3} = 1/3.$$

Simiarly, $\Pr[C_3|M_2] = 2/3$. In conclusion, if she sticks with her decision, then the probability to get the car is $1/3$. If she switches, then the probability is $2/3$. This means that it is advisable that she switches.

[2] Incidentally, Monty Hall was awarded the prestigious Order of Canada by the Canadian goverment for his humanitarian efforts...

Remark. There are many web sites dedicated to this problem, and one finds heated discussions about the Monty Hall problem on the internet.[3] You will notice that there exist different solutions, depending on the exact assumptions about Monty's knowledge and his strategy.

§2 Random Variables

We discuss in this section the concept of a random variable. Random variables are functions that associate a numerical value to each outcome of an experiment. For instance, if we roll a pair of dice, then the sum of the two face values is a random variable. Similarly, if we toss a coin three times, then the observed number of heads is a random variable.

Let $\mathcal{F}$ be a σ-algebra over the sample space Ω. A *random variable* X is a function $X\colon \Omega \to \mathbf{R}$ such that the set $\{z \in \Omega \mid X(z) \leq x\}$ is an event contained in $\mathcal{F}$ for all $x \in \mathbf{R}$. For brevity, we will say that X is defined on the σ-algebra $(\Omega, \mathcal{F})$. It should be clear from this definition that there is nothing *random* about a random variable, it is simply a function.

The definition ensures that a random variable can be used to specify events in a convenient way. There are a number of notational conventions which help to express events in an even more compact way. For instance, the event $\{z \in \Omega \mid X(z) \leq x\}$ is denoted shortly by $X \leq x$, an idiosyncratic but standard notation.

Example 3 If X is the random variable denoting the sum of the face values of a pair of dice, then $X \leq 3$ denotes the event $\{(1,1),(1,2),(2,1)\}$.

Example 4 If Y is the random variable counting the number of heads in three subsequent coin tosses, then $Y \leq 0$ is the event $\{$(tail,tail,tail)$\}$, and $Y \leq 1$ is the event $\{$(tail,tail,tail), (head,tail,tail), (tail,head,tail), (tail,tail,head)$\}$.

A *discrete random variable* is a random variable with countable range, which means that the set $\{X(z) \mid z \in \Omega\}$ is countable.

> We will henceforth assume that **all random variables are discrete** unless otherwise specified.

[3] The discussions are not concerned about the fact whether or not the contestant had to *marry* the goat, because this does *not* influence the probabilities. So stay focused!

The convenience of a discrete random variable X is that one can define events in terms of values of X, for instance in the form $X \in A$ which is short for $\{z \in \Omega \mid X(z) \in A\}$. If the set A is a singleton, $A = \{x\}$, then we write $X = x$. The next exercise shows that $X = x$ and $X \in A$ indeed specify events.

Exercise 1.7 *Let X be a discrete random variable defined on the σ-algebra $(\Omega, \mathcal{F})$. Show that $X = x$ is an event of $\mathcal{F}$. Hint: Prove that $y < X \leq x$ is an event for all real numbers y and x, and deduce the claim.*

Densities and Distributions. Let X be a discrete random variable defined on a σ-algebra $(\Omega, \mathcal{F})$. Let Pr be a probability measure on $\mathcal{F}$. The *density function* p_X of a discrete random variable X is defined by

$$p_X(x) = \Pr[X = x].$$

Thus, the density function describes the probabilities of the events $X = x$. Note that the density function is sometimes called *probability mass function*.

The function $F_X(x) = \Pr[X \leq x]$ is called the *distribution function* of the random variable X. Note that the distribution function can be defined in the same way for arbitrary random variables; this is not true for the density function.

Exercise 1.8 *Show that the distribution function $F_X(x)$ of a random variable is a non-decreasing function, i.e., that $x \leq y$ implies $F_X(x) \leq F_X(y)$.*

Example 5 Let $(\Omega, 2^\Omega, \Pr)$ be the probability space of a pair of fair dice, that is, the sample space $\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}$, and Pr is the uniform probability measure, $\Pr[A] = |A|/36$ for any subset A of Ω. Let X denote the random variable denoting the sum of the face values of the two dice. The density function and the distribution function of X are tabulated below:

x	2	3	4	5	6	7	8	9	10	11	12
$\Pr[X = x]$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
$\Pr[X \leq x]$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

Independence. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. Two events E and F are called *independent* if and only if $\Pr[E \cap F] = \Pr[E] \Pr[F]$. This means that the knowledge that F has occured does not influence the probability that E occurs, because the condition $\Pr[E \cap F] = \Pr[E] \Pr[F]$ is equivalent to

$\Pr[E\,|F\,] = \Pr[E]$. We can define the independence of random variables in a similar way.

Let $(X_1, X_2, \ldots, X_n)$ be a sequence of discrete random variables defined on the probability space $(\Omega, \mathcal{F}, \Pr)$. We say that the random variables X_k, $1 \le k \le n$, are *mutually independent* if and only if

$$\Pr[\{z \,|\, X_k(z) = x_k, 1 \le k \le n\}] = \prod_{k=1}^{n} \Pr[\{z \,|\, X_k(z) = x_k\}] \qquad (1.1)$$

holds for all $(x_1, \ldots, x_n) \in \mathbf{R}^n$. Note that condition (1.1) is usually expressed in the idiosyncratic form

$$\Pr[X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n] = \prod_{k=1}^{n} \Pr[X_k = x_k]$$

Expectation Values. Let X be a discrete random variable over the probability space $(\Omega, \mathcal{F}, \Pr)$. The *expectation value* of X is defined to be

$$\mathrm{E}[X] = \sum_{\alpha \in X(\Omega)} \alpha \Pr[X = \alpha],$$

when this sum is unconditionally convergent in $\overline{\mathbf{R}}$, the extended real numbers. The expectation value is also called the *mean* of X. If X is a random variable with nonnegative integer values, then the expectation can be calculated by

$$\mathrm{E}[X] = \sum_{x=1}^{\infty} \Pr[X \ge x],$$

which is often convenient. If X and Y are two arbitrary discrete random variables, then

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

that is, the expectation operator is linear. This is an extremely useful result.

If X and Y are independent discrete random variables, then

$$\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y].$$

Caveat: If X and Y are not independent, then this is in general false.

Example 6 Suppose that n persons give their hats to the hat check girl. She is upset because her goat has just passed away, and is handing the hats back

at random. We want to answer the following question: On average, how many persons get their own hat back?

We take $\Omega = \{1, \ldots, n\}$ as sample space, and allow all subsets of Ω as events, $\mathcal{F} = 2^\Omega$. The event $m_k = \{k\}$ means that the kth person received his own hat. Let $X_k \colon \Omega \to \mathbf{R}$ be the random variable defined by $X_k(k) = 1$ and $X_k(x) = 0$ for all $x \neq k$; hence $X_k = 1$ denotes the event m_k, and $X_k = 0$ the event $\Omega - m_k$. The probability that the kth person receives her own hat back is $1/n$, since she will receive one of n possible hats. Consequently, we define the probability measure by $\Pr[m_k] = 1/n$. Let $X = X_1 + \cdots + X_n$ denote the number of persons receiving their own hats. We have

$$\mathrm{E}[X] = \sum_{k=1}^{n} \mathrm{E}[X_k] = \sum_{k=1}^{n} 1 \cdot \Pr[X_k] = n(1/n) = 1,$$

by linearity of expectation, and by definition of the expectation value. This means that on average one person gets his own hat back. This neat illustration of the linearity of expectation is adapted from Spencer [9], but presumably this result was known since the invention of the hat.

The expectation can be used to bound probabilities, as the following simple, but fundamental, result shows:

Theorem 7 (Markov's Inequality) *If X is a random variable and t a positive real number, then*

$$\Pr[|X| \geq t] \leq \frac{\mathrm{E}[|X|]}{t}.$$

Proof. Let Y denote the random variable

$$Y(\omega) = \begin{cases} 0 & \text{if } |X(\omega)| < t, \\ 1 & \text{if } |X(\omega)| \geq t, \end{cases}$$

hence $Y = 1$ denotes the event $|X| \geq t$. The expectation value of $|X|$ satisfies

$$\mathrm{E}[|X|] \geq \mathrm{E}[tY] = t\,\mathrm{E}[Y] = t\,\Pr[|X| \geq t],$$

which proves the claim. ■

Exercise 1.9 *Let X be a discrete random variable and let $h \colon \mathbf{R} \to \mathbf{R}$ be a nonnegative function. Show that for all positive real numbers t, we have*

$$\Pr[h(X) \geq t] \leq \frac{E[h(X)]}{t}.$$

Variance. The *variance* $\mathrm{Var}[X]$ of a discrete random variable X is defined by

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2,$$

whenever this expression is well-defined. The variance measures the squared deviation from the expected value $\mathrm{E}[X]$.

It is easy to see that variance is *not* a linear operator, since

$$\mathrm{Var}[X + X] = 4\mathrm{Var}[X]$$

holds, to mention just one example. Moreover, $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$ for all $a, b \in \mathbf{R}$. If X and Y are independent random variables, then the variance satisfies

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]. \tag{1.2}$$

The random variable X will rarely deviate from the expectation value if the variance is small. This is a consequence of the Chebychev's useful inequality:

Theorem 8 (Chebychev's inequality) *If X is a random variable, then*

$$\Pr[(X - E[X])^2 \geq \beta] \leq \frac{\mathrm{Var}[X]}{\beta}. \tag{1.3}$$

Proof. We show how (1.3) can be derived to give an example of such calculations. By definition,

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \sum_{\alpha \in (X(\Omega) - \mathrm{E}[X])^2} \alpha \Pr[(X - \mathrm{E}[X])^2 = \alpha]$$

Omitting all values of α less than β from this sum, we obtain

$$\begin{aligned} \mathrm{Var}[X] &\geq \sum_{\alpha \in (X(\Omega) - \mathrm{E}[X])^2,\, \alpha \geq \beta} \alpha \Pr[(X - \mathrm{E}[X])^2 = \alpha] \\ &\geq \sum_{\alpha \in (X(\Omega) - \mathrm{E}[X])^2,\, \alpha \geq \beta} \beta \Pr[(X - \mathrm{E}[X])^2 = \alpha] \end{aligned}$$

The last sum is equal to $\beta \Pr[(X - \mathrm{E}[X])^2 \geq \alpha]$. Dividing by β on both sides yields the Chebychev's inequality. ■

The square root of the variance, $\sigma = \sqrt{\mathrm{Var}[X]}$, is called the *standard deviation* of the random variable X.

Exercise 1.10 *Show that if X is a random variable with standard deviation σ, then*

$$\Pr[\,|X - \mathrm{E}[X]| \geq c\sigma] \leq \frac{1}{c^2}$$

for any positive constant $c \in \mathbf{R}$. This formula is often also called Chebychev's inequality. Can you explain why?

Bernoulli Distribution. Tossing a biased coin can be described by a random variable X that takes the value 1 if the outcome of the experiment is **head**, and the value 0 if the outcome is **tail**. Assume that $\Pr[X = 1] = p$ and $\Pr[X = 0] = 1 - p$ for some real number $p \in (0, 1)$. The random variable X is said to have the Bernoulli distribution with parameter p. We can compute the expectation value and the variance as follows:

$$\mathrm{E}[X] = p, \qquad \mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = p - p^2 = p(1 - p).$$

Binomial Distribution. Let $X_1, \ldots, X_n$ denote independent identically distributed random variables, all having a Bernoulli distribution with parameter p. Then the random variable $X = X_1 + \cdots + X_n$ describes the number of heads in a sequence of n coin flips. The expectation of X can be immediately computed by linearity of expectation, and, since the random variables X_k are independent, we can compute the variance using (1.2):

$$\mathrm{E}[X] = np, \qquad \mathrm{Var}[X] = np(1 - p).$$

The probability of the event $X = x$, for integers in the range $0 \leq x \leq n$, is

$$\Pr[X = x] = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Indeed, choose x positions in a sequence of length n. The probability that the sequence will show heads at exactly these positions is $p^x(1 - p)^{n-x}$. The result follows, since there are $\binom{n}{x}$ ways to choose x positions in a sequence of length n.

Uniform Distribution. Let X be a random variable that takes integral values in $\{1, \ldots, n\}$. Such a random variable is said to be uniformly distributed if $\Pr[X = k] = 1/n$ for all integers k in the range $1 \leq k \leq n$. The expectation value and the variance of X are respectively given by

$$\mathrm{E}[X] = \frac{n + 1}{2}, \qquad \mathrm{Var}[X] = \frac{n^2 - 1}{12}.$$

The expectation value follows from the definition. We can verify the variance by noting that

$$\mathrm{E}[X^2] = \sum_{k=1}^{n} \frac{1}{n}\, k^2 = \frac{1}{n}\frac{n(1+n)(1+2n)}{6} = \frac{(1+n)(1+2n)}{6},$$

hence $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \dfrac{(1+n)(1+2n)}{6} - \dfrac{(n+1)^2}{4} = (n^2-1)/12.$

Geometric Distribution. Suppose we keep tossing a biased coin, which has the Bernoulli distribution with parameter p, until the event **head** occurs. Let the random variable X denote the number of coin flips needed in this experiment. We say that X is geometrically distributed with parameter p. The density function of X is given by

$$p_X(x) = \Pr[X = x] = p(1-p)^{x-1}$$

for $x = 1, 2, \ldots,$ and $p_X(x) = 0$ otherwise. The expectation value and the variance of X are given by

$$\mathrm{E}[X] = \frac{1}{p}, \qquad \mathrm{Var}[X] = \frac{1-p}{p^2}.$$

It is possible to derive these facts directly from the definitions. For the expectation value this can be done without too much effort, but for the variance this is cumbersome. In the next section, we will introduce a tool that can significantly simplify such calculations.

Exercise 1.11 *Prove that a geometrically distributed random variable X with parameter p has expectation value $\mathrm{E}[X] = 1/p$ using the definition of $\mathrm{E}[X]$ and some calculus. If you are fearless, then you can also attempt to derive the variance of X.*

Negative Binomial Distribution. Let $X_1, \ldots, X_n$ be independent random variables, all having geometric distribution with parameter p. The random variable $X = X_1 + \cdots + X_n$ describes the number of coin flips that are necessary until n heads occur, when heads has probability p. The random variable X is said to have negative binomial distribution with parameters n and p. Linearity of expectation and additivity of variance for independent random variables shows that

$$\mathrm{E}[X] = \sum_{k=1}^{n} \mathrm{E}[X_k] = \frac{n}{p}, \qquad \mathrm{Var}[X] = \sum_{k=1}^{n} \mathrm{Var}[X_k] = \frac{n(1-p)}{p^2}.$$

The probability of the event $X = k$ is

$$\Pr[X = k] = \binom{k-1}{n-1} p^n (1-p)^{k-n}, \qquad k \geq n.$$

Indeed, a sequence of k coin flips that contains exactly n heads at specified positions has probability $p^n(1-p)^{k-n}$. By specification, the last position is a head, hence there are $\binom{k-1}{n-1}$ other positions that can be chosen for the heads.

Poisson Distribution. A random variable X with non-negative integer values is said to be Poisson distributed with parameter $\lambda > 0$ if

$$\Pr[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad k = 0, 1, 2, \ldots .$$

The Poisson distribution can be used to approximate the Binomial distribution if n is large and p is small. Indeed, suppose that $\lim_{n\to\infty} np_n = \lambda$, then

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = e^{-\lambda} \frac{\lambda^k}{k!}.$$

This formula is frequently used when the evaluation of the binomial distribution is not feasible.

Caveat: Note that this is an asymptotic result. Many wrong statements and conclusions can be found in the literature, which are a result of ignoring the hypothesis $\lim_{n\to\infty} np_n = \lambda$.

Exercise 1.12 *Suppose that you have a biased coin that produces heads with probability p, $0 < p < 1$, but unfortunately this probability is not known to you. Von Neumann showed that it is possible to use such a biased coin to construct a source for fair coin flips, i.e.,* $\Pr[\textbf{head}] = \Pr[\textbf{tail}] = 1/2$. *Derive a scheme such that the expected number of biased coin flips does not exceed $1/p(1-p)$. Hint: Consider consecutive pairs of biased coin flips.*

§3 Examples

We stressed in the previous section that random variables are used to specify events. Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. An event $A \in \mathcal{F}$ can be specified by its characteristic function $X_A \colon \Omega \to \mathbf{R}$,

$$X_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{otherwise.} \end{cases}$$

The random variable X_A is called the *indicator random variable* of the event A. The expectation value of X_A is given by

$$\mathrm{E}[X_A] = 1 \cdot \Pr[X = 1] + 0 \cdot \Pr[X = 0] = \Pr[A],$$

that is, the expectation value of an indicator random variable coincides with the probability of the defining event A. The variance is given by

$$\mathrm{Var}[X] = E[X_A^2] - \mathrm{E}[X_A]^2 = \mathrm{E}[X_A] - \mathrm{E}[X_A]^2 = \Pr[A](1 - \Pr[A]).$$

A common trick is to decompose a given random variable X into a sum of indicator random variables. This can simplify, for instance, the calculation of the mean. The following examples illustrate this approach.

Left-to-right minima in a permutation. Let S_n denote the set of all possible permutations of the numbers $N = \{1, \ldots, n\}$. The set S_n is known as the symmetric group. An element of S_n can be written as $(a_1, a_2, \ldots, a_n)$, which can be identified with the bijective map $k \mapsto a_k$. A number a_k satisfying $a_k = \min\{a_1, \ldots, a_k\}$ is called a *left-to-right minimum*. This quantity is crucial in the analysis of minimum search algorithms [5] and some sorting algorithms [6]. Consider the random variable $X \colon S_n \to \mathbf{Z}$ defined by

$$X(a_1, \ldots, a_n) = \bigl|\bigl\{k \in \{1, \ldots, n\} \,|\, a_k = \min\{a_1, \ldots, a_k\}\bigr\}\bigr|.$$

This random variable counts the left-to-right minima. Our goal is to determine the expectation value $\mathrm{E}[X]$, assuming $\Pr[\pi] = 1/n!$ for all $\pi \in S_n$.

Let X_k, $1 \le k \le n$, denote the indicator random variable of the event $a_k = \min\{a_1, \ldots, a_k\}$. The random variable X satisfies $X = \sum_{k=1}^{n} X_k$, hence

$$\mathrm{E}[X] = \sum_{k=1}^{n} \mathrm{E}[X_k] = \sum_{k=1}^{n} \Pr[a_k = \min\{a_1, \ldots, a_k\}].$$

It is not difficult to show that $\Pr[a_k = \min\{a_1, \ldots, a_k\}] = 1/k$. Therefore, the mean of X is equal to the nth Harmonic number H_n,

$$\mathrm{E}[X] = \sum_{k=1}^{n} \frac{1}{k} = H_n.$$

Note that $\ln(n+1) \le H_n \le (\ln n) + 1$.

Exercise 1.13 *Prove that a random permutation $(a_1, \ldots, a_n)$ has a left-to-right minimum at position k with probability $1/k$, assuming that $\Pr[\pi] = 1/n!$ for all $\pi \in S_n$.*

The Coupon Collector Problem. The hat check girl is a compulsive coupon collector. Currently, she is collecting charming Harry Potter characters that are contained in overpriced serial boxes. There are n different characters, and each box contains one character. She wants to get the complete set. How many boxes does she have to buy, on average, to obtain one complete collection?

Let X denote the random variable counting the number of boxes required to collect at least one character of each type. Our goal is to determine $\mathrm{E}[X]$. Let X_k denote the random variables counting the number of boxes that the hat check girl buy to get the $(k+1)$-th character, after she has already collected k characters. The probability to draw one of the remaining characters is $p_k = (n-k)/n$. Hence X_k is a geometrically distributed random variable with parameter p_k. Consequently, $\mathrm{E}[X_k] = 1/p_k = n/(n-k)$.

The random variable X is the given by the sum $X = \sum_{k=0}^{n-1} X_k$. Linearity of expectation shows that

$$\mathrm{E}[X] = \sum_{k=0}^{n-1} \mathrm{E}[X_k] = \sum_{k=0}^{n-1} \frac{n}{n-k} = n \sum_{k=1}^{n} \frac{1}{k} = nH_n.$$

Exercise 1.14 *Calculate the variance* $\mathrm{Var}[X]$.

Let X be a random variable defined on a finite probability space $(\Omega, 2^\Omega, \Pr)$. It follows from the definition of the expectation value that there must exist elements $a, b \in \Omega$ such that $X(a) \geq \mathrm{E}[X]$ and $X(b) \leq \mathrm{E}[X]$. These simple facts are often used to show that certain combinatorial structures must exist.

Bipartite Subgraphs. A graph is *bipartite* if and only if its vertex set can be partitioned in two sets A and B such that all edges join a vertex from A with a vertex from B. Let $G = (V, E)$ be a graph with $n = |V|$ vertices and $e = |E|$ edges. We want to show that G contains a bipartite subgraph with at least $e/2$ edges.

Construct a random subset A of V by tossing a fair coin for each $v \in V$, and include all vertices with outcome heads in A. Let B denote the complementary set $B = A^c$. An edge $\{x, y\}$ is said to be crossing if exactly one of x and y is contained in A. Let X be the random variable denoting the number of crossing edges. Let X_{xy} be the indicator random variable for the event that $\{x, y\}$ is a crossing edge. We have

$$X = \sum_{\{x,y\} \in E} X_{xy}.$$

Clearly, $\mathrm{E}[X_{xy}] = \Pr[$ edge $\{x, y\}$ is crossing $] = 1/2$. Therefore,

$$\mathrm{E}[X] = \sum_{\{x,y\}\in E} \mathrm{E}[X_{xy}] = \frac{e}{2}.$$

It follows that there must exists some choice of A such that $X(A) \geq e/2$, and the corresponding crossing edges form the desired bipartite subgraph.

This example illustrates the probabilistic method which proves the existence of combinatorial structures satisfying certain constraints by probabilistic arguments. This and many other examples can be found in the book by Alon and Spencer [1].

The Maximum of Geometric Random Variables. A geometric random variable describes how many times we have to flip a biased coin until we obtain heads. Suppose that we have n such coins and we toss all of them once in one round. How many rounds do we need, on average, until heads has occured at least once for all n coins?

We can model this situation as follows. Let $Y_1, \ldots, Y_n$ be independent identically distributed random variables that have a common geometric distribution with parameter q. Let X be the random variable $X = \max(Y_1, \ldots, Y_n)$. In other words, $X = k$ denotes the event that after k rounds head has occured for each coin. Our goal is to compute the expectation value $\mathrm{E}[X]$.

The random variable X has non-negative integer values, hence the mean can be computed by

$$\mathrm{E}[X] = \sum_{m=0}^{\infty} \Pr[X > m] = \sum_{m=0}^{\infty} \left(1 - \Pr[Y_1 \leq m, \ldots, Y_n \leq m]\right).$$

The last equality follows from the definition of X, expressed in terms of the complementary event. The random variables Y_k are independent and identically distributed, hence

$$\mathrm{E}[X] = \sum_{m=0}^{\infty} \left(1 - \Pr[Y_1 \leq m]^n\right).$$

For convenience, we set $p = 1-q$ to shorten some expressions. A geometrically distributed random variable satisfies $\Pr[Y_1 \leq m] = 1 - (1-q)^m = 1 - p^m$. Hence,

$$\mathrm{E}[X] = \sum_{m=0}^{\infty} \left(1 - \left(1 - (1-q)^m\right)^n\right) = \sum_{m=0}^{\infty} \left(1 - \left(1 - p^m\right)^n\right). \tag{1.4}$$

Exercise 1.15 *Derive from equation (1.4) the alternate expression*

$$\mathrm{E}[X] = \sum_{k=1}^{n} \binom{n}{k} (-1)^{k+1} \frac{1}{1-(1-q)^k}.$$

§4 Further Reading

There exist an abundance of good introductory text on probability theory. The book by Grimmett and Stirzaker [3] gives a well-rounded introduction, and contains numerous excellent exercises. The book by Ross [8] is an easily readable introduction, which contains a well-chosen number of interesting applications. The recent book by Williams [10] gives a lively discussion of many topics, and covers some computational aspects. The book by Jacod and Protter [4] gives a solid, but accessible, measure-theoretic treatment of probability theory. Probability generating functions, and much more, are discussed in Graham, Knuth, Pataschnik [2].

The book by Alon and Spencer [1] contains numerous applications illustrating the probabilistic method. The book by Motwani and Raghavan [7] gives an introduction to randomized algorithms, and is indispensable for computer scientists.

Bibliography

[1] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, New York, 2nd edition, 2000.

[2] R.L. Graham, D.E. Knuth, and O. Pataschnik. *Concrete Mathematics – A Foundation for Computer Science*. Addison-Wesley, Reading, MA, 1992. 8th corrected printing.

[3] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Oxford University Press, Oxford, 3rd edition, 2001.

[4] J. Jacod and P. Protter. *Probability Essentials*. Springer-Verlag, Berlin, 2000.

[5] D.E. Knuth. *The Art of Computer Programming – Fundamental Algorithms*, volume 1. Addison-Wesley, 3rd edition, 1997.

[6] D.E. Knuth. *The Art of Computer Programming – Sorting and Searching*, volume 3. Addison-Wesley, 2nd edition, 1998.

[7] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, Cambridge, 1995.

[8] S.R. Ross. *Introduction to Probability Models*. Academic Press, San Diego, 7th edition, 2000.

[9] J. Spencer. *The Lectures on the Probabilistic Method*. SIAM, Philadelphia, 1987.

[10] D. Williams. *Weighing the Odds – A Course in Probability and Statistics*. Cambridge University Press, Cambridge, 2001.