COSC 264
A Review of Discrete Probability

Andreas Willig
andreas.willig@canterbury.ac.nz

August 27, 2019

1 Introduction

The lab quizzes of the second week and the sixth week of term 3 contain some questions involving discrete probability distributions. Although you are generally expected to be familiar with this material from MATH120 or EMTH118/EMTH119, on this sheet we will review some of the relevant bits and pieces of discrete probability, just in case you feel a bit rusty in that department. Note that the comments and definitions below are not the real definitions that mathematicians would use. We have attempted to keep the technicalities to a minimum.

2 Discrete Random Variables and Expectations

We call a set Ω **discrete** if it either has:

- a finite number of elements, like for example the set $\Omega = \{1, 2, 3, 4, 5, 6\}$ of numbers on a die,
- or if it has an at most countably infinite number of elements, like for example the set $\mathbb{N}$ of (positive) integers, the set $\mathbb{N}_0$ of non-negative integers (including zero), or the set $\mathbb{Z}$ of all integers. In contrast, the set $\mathbb{R}$ of all real numbers is uncountably infinite and hence not a discrete set.

We regard a **random variable** X as a number which can take on a random value within a given set, which is called its **range**. We are particularly interested in the case of **discrete random variables**, where the range is a discrete set. A few examples:

- If we throw a coin, we can get two different outcomes, often called head and tail. In probability theory it is customary to denote the outcomes by 0 and 1, and hence the range of a "coin toss random variable" is the set $\Omega = \{0, 1\}$.

- If we throw a die, the range is the set $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- If we use a Geiger counter to count the number of radioactive decay events within one second, then theoretically every integer number (including zero) is a possible outcome, so the random variable for the number of ticks has range $\Omega = \mathbb{N}_0$.

If we are given a discrete random variable X with range Ω, then each individual outcome $\omega \in \Omega$ (e.g. the number shown by a die after actually performing a toss) occurs with a certain probability. Depending on the nature of the random variable, some outcomes might be more probable than others (e.g. when we use a biased coin). The assignment of probabilities to all the possible outcomes of a random variable is often called the **probability mass function** or the **probability distribution** of the random variable X. More specifically, we assign to each possible outcome $\omega \in \Omega$ a real number $p(\omega) := \Pr[X = \omega]$ between 0 and 1 which reflects the probability that the random variable X just takes on the value $\omega \in \Omega$. A minimum requirement for such an assignment of probabilities to outcomes minimally needs to satisfy the following criterion:

$$\sum_{\omega\in\Omega} p(\omega) = 1,$$

i.e. the sum of the probabilities over all the possible outcomes must be one.

A particularly important quantity associated to a random variable X is its **mean value**, **expected value** or simply **expectation**, often denoted by $E[X]$. The expectation is kind of the "typical value" that a random variable takes, even though the expectation value does not need to be within the range. For a discrete random variable X with range Ω the expectation is defined as

$$E[X] = \sum_{\omega\in\Omega} \omega \cdot p(\omega)$$

For any discrete random variable with finite range, the expectation always exists, i.e. is a real number. For discrete random variable with a countably infinite range it may happen that the expectation becomes ∞ or $-\infty$. However, these cases will be irrelevant to us. An important property of the expectation is that if we are given two random variables X and Y over the same range Ω, we can form a new random variable Z by adding up X and Y, and we can calculate the expectation of Z as

$$E[Z] = E[X + Y] = E[X] + E[Y],$$

i.e. by simply adding up the expectations of X and Y. Note furthermore that if we have a random variable X and we multiply it by a constant $c \in \mathbb{R}$ to give us a new random variable $Y = c \cdot X$, then its expectation can be calculated as

$$E[Y] = E[c \cdot X] = c \cdot E[X]$$

3 Conditional Probabilities and the Law of Total Probability

In the quiz for the final week of term 3 you will need the so-called law of total probability for one of the questions. To explain this, we need to introduce a few further notions. In the following, let X be a discrete random variable with range Ω.

An **event** is a subset of Ω. For example, when our random variable X corresponds to tossing a die, then we have $\Omega = \{1, 2, 3, 4, 5, 6\}$ and the event that our coin toss turns up an even number is given by the subset $\mathcal{E} = \{2, 4, 6\}$. The probability distribution $p(\omega) = \Pr[X = \omega]$ specifies the probabilities of individual outcomes. The probability of an event $\mathcal{E} \subset \Omega$ is given by

$$\Pr[\mathcal{E}] := \Pr[X \in \mathcal{E}] = \sum_{\omega \in \mathcal{E}} p(\omega),$$

i.e. as the sum of all the probabilities $p(\omega)$ over all ω belonging to the event $\mathcal{E}$. With this definition we can work with probabilities of events and not just individual outcomes. For these probabilities over events there are a few important rules and relationships:

- $\Pr[\emptyset] = 0$
- $\Pr[\Omega] = 1$
- If $\mathcal{E} \subset \Omega$ is an event and $\mathcal{E}^{\mathcal{C}} = \Omega - \mathcal{E}$ is its complementary set with respect to Ω, then we have $\Pr\left[\mathcal{E}^{\mathcal{C}}\right] = 1 - \Pr[\mathcal{E}]$
- If $\mathcal{A}_1, \mathcal{A}_2, \ldots$ is a finite or countably infinite sequence of mutually disjoint events, then we have
$$\Pr\left[\bigcup_i \mathcal{A}_i\right] = \sum_i \Pr[\mathcal{A}_i],$$
i.e. the probability of the union of disjoint events is just the sum of the probabilities of the individual events. Note that in expressions like these, the variable i can range over either a finite set or a countably infinite set, depending on the case at hand.

A rather important definition is the one of **independence** of two events: be X a random variable with range Ω and $\mathcal{A} \subset \Omega$ and $\mathcal{B} \subset \Omega$ be two events. These two events are called independent if

$$\Pr[\mathcal{A} \cap \mathcal{B}] = \Pr[\mathcal{A}] \cdot \Pr[\mathcal{B}],$$

i.e. when the probability of two events occuring simultaneously is just given by the product of the probabilities of the events.

For a given random variable X with range Ω we call a finite or countably infinite set of events $\mathcal{A}_1, \mathcal{A}_2, \ldots$ a **partition**, when the following three properties hold:

- All the events are mutually disjoint, i.e. $\mathcal{A}_i \cap \mathcal{A}_j = \emptyset$ for $i \neq j$.
- None of the events is the empty set.

- We have
$$\bigcup_i \mathcal{A}_i = \Omega,$$
i.e. when taken together the events $\mathcal{A}_i$ make up all of Ω.

When $\mathcal{A}$ and $\mathcal{B}$ are two events and $\Pr[\mathcal{B}] > 0$ holds, then we define the **conditional probability of event $\mathcal{A}$ given event $\mathcal{B}$** as

$$\Pr[\mathcal{A}|\,\mathcal{B}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{B}]} =: \frac{\Pr[\mathcal{A}, \mathcal{B}]}{\Pr[\mathcal{B}]}$$

You can imagine this conditional probability as the probability that event $\mathcal{A}$ occurs when we already know that event $\mathcal{B}$ has occurred. For example, if we consider the toss of a die, the probability of the event $\mathcal{A} = \{1\}$ (i.e. throwing a one) is $\frac{1}{6}$ if we know nothing else about the outcome. However, if we know up-front that the outcome is an even number (which as an event is the subset $\mathcal{B} = \{2,4,6\}$) we can say that $\Pr[\mathcal{A}|\,\mathcal{B}] = 0$ holds. There is an important relationship between the independence of two events $\mathcal{A}$ and $\mathcal{B}$ and their conditional probability $\Pr[\mathcal{A}|\,\mathcal{B}]$, assuming that $\Pr[\mathcal{B}] > 0$ holds:

$$\Pr[\mathcal{A}|\,\mathcal{B}] = \frac{\Pr[\mathcal{A} \cap \mathcal{B}]}{\Pr[\mathcal{B}]} = \frac{\Pr[\mathcal{A}] \cdot \Pr[\mathcal{B}]}{\Pr[\mathcal{B}]} = \Pr[\mathcal{A}]\,,$$

i.e. the probability that $\mathcal{A}$ occurs is not influenced by the fact that event $\mathcal{B}$ has occured.

The **law of total probability** is very useful in many circumstances, as it often helps to simplify calculating the probability of an event by using a partition and conditioning this event upon the events making up the partition. For a random variable X with range Ω the law of total probability can be formulated as follows:

Theorem 1. Be $\mathcal{A}$ an event and $\mathcal{B}_1, \mathcal{B}_2, \ldots$ a partition of Ω. Then we can express the probability of $\mathcal{A}$ as follows:

$$\Pr[\mathcal{A}] = \sum_i \Pr[\mathcal{A} \cap \mathcal{B}_i] = \sum_i \Pr[\mathcal{A}|\,\mathcal{B}_i] \cdot \Pr[\mathcal{B}_i]$$

where i ranges over all the events making up the partition of Ω.

4 Bernoulli Random Variables

The Bernoulli random variable is the most basic random variable. Conceptually, it corresponds to a single random experiment with only two outcomes, for example, a coin tossing experiment giving either head or number as outcomes. However, it is customary to take the numbers 0 (for "failure") and 1 (for "success") as the possible outcomes. The probability mass function of a random variable X having Bernoulli distribution is given by:

$$p(k) = \left\{ \begin{array}{rcl} p & : & k = 1 \\ (1-p) & : & k = 0 \end{array} \right.$$

where $p \in [0,1]$ is called *success probability* and $q := 1 - p$ is the *failure probability.*[1]

Be X a random variable having Bernoulli distribution with success parameter p. We write this as $X \sim \mathrm{B}(p)$. The expectation $E[X]$ is given by:

$$E[X] = 0 \cdot \Pr[X = 0] + 1 \cdot \Pr[X = 1] = p$$

5 Binomial Random Variables

Consider a series of n independent Bernoulli experiments with corresponding random variables $X_1, X_2, \ldots, X_n$, all having the same success probability $p \in [0,1]$. Set $q = 1 - p$. Consider a specific sequence of outcomes $X_1 = i_1, X_2 = i_2, \ldots, X_n = i_n$ with $i_\nu \in \{0,1\}$, so that there are exactly k ones among $i_1, i_2, \ldots, i_n$ and $n - k$ zeros. Due to the independence of the random variables $X_1, \ldots, X_n$ the probability that the exact sequence of outcomes $i_1, i_2, \ldots, i_n$ occurs is given by:

$$\Pr[X_1 = i_1] \cdot \Pr[X_2 = i_2] \cdot \ldots \cdot \Pr[X_n = i_n] = p^k \cdot q^{n-k}$$

There are $\binom{n}{k}$ different n-element sequences of outcomes where k successes occur in n trials and thus the probability that k successes occur in n trials no matter what their exact positions are is given by:

$$b(k; n, p) = \binom{n}{k} \cdot p^k \cdot q^{n-k} \qquad , k \in \{0, \ldots, n\} \tag{1}$$

This is the *binomial distribution.* It has the range $\{0, 1, \ldots, n\}$. The expectation of a binomially distributed random variable X with parameters n and p is given by:

$$E[X] \;=\; E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i] = np$$

As an example, the probability mass function $b(\cdot; 20, 0.2)$ is shown in Figure 1. Since $p < 0.5$ most of the probability mass concentrates in the left part (below $k = 10$), for $p > 0.5$ it would concentrate in the right part. For $p = 0.5$ the figure would look symmetric.

6 Geometric Random Variables

The geometric distribution appears in the following scenario: when doing a theoretically infinite number of identical and independent Bernoulli experiments (i.e. all having the same success parameter p), how many experiments are needed *before* you see the first success? A slightly different question is: how many experiments are needed *until* you

[1] You may find it confusing that we use the letter p in isolation as the success probability parameter of a Bernoulli random variable, and the function $p(\cdot)$ to denote the probability distribution function of a random variable. Both are somewhat customary in the literature, and you should be able to distinguish these from the context.

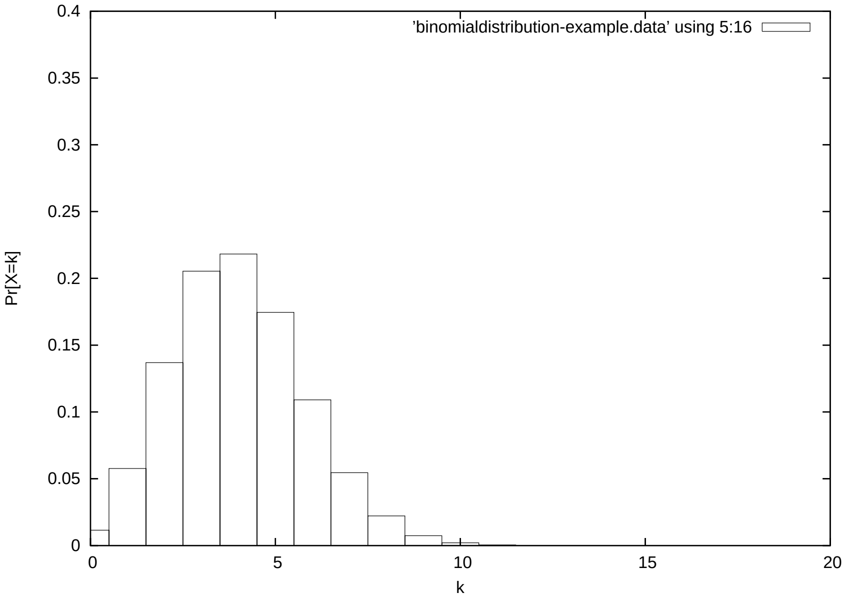

Figure 1: Probability mass function of a random variable $X \sim \text{Binomial}\,(20, 0.2)$

have seen the first success? The latter formulation includes the successful trial, the first formulation excludes it. As an example of a geometrically distributed random variable in the latter formulation is the number of throws of a die you need until you get a six for the first time.

The probability mass function for the first formulation (the random variable is denoted by X_1) is given by:

$$p_1(k) = p \cdot (1-p)^k \qquad (k \in \mathbb{N}_0)$$

and hence its range is $\mathbb{N}_0$. For the second formulation (X_2) it is given by:

$$p_2(k) = p \cdot (1-p)^{k-1} \qquad (k \in \mathbb{N})$$

(with range $\mathbb{N}$). For the first formulation we get the expectation:

$$E\left[X_1\right] = \frac{q}{p}$$

which one can see after a slightly tricky calculation that is irrelevant for our purposes.[2]

[2] If you want to see it nonetheless, here goes, without many explanations:

$$\begin{aligned} E\left[X_1\right] &= \sum_{k=0}^{\infty} k p_1(k) = pq \sum_{k=1}^{\infty} k q^{k-1} = pq \sum_{k=1}^{\infty} \frac{d}{dq} q^k = pq \frac{d}{dq} \left(\sum_{k=1}^{\infty} q^k \right) \\ &= pq \frac{d}{dq} \left(\frac{1}{1-q} - 1 \right) = pq \frac{1}{(1-q)^2} = \frac{q}{1-q} = \frac{q}{p} \end{aligned}$$

This computation makes use of the geometric series and the fact that under some assumptions one can exchange the order of differentiation and summation of an infinite series.

While for the second formulation we get:

$$E[X_2] = \frac{1}{p}$$

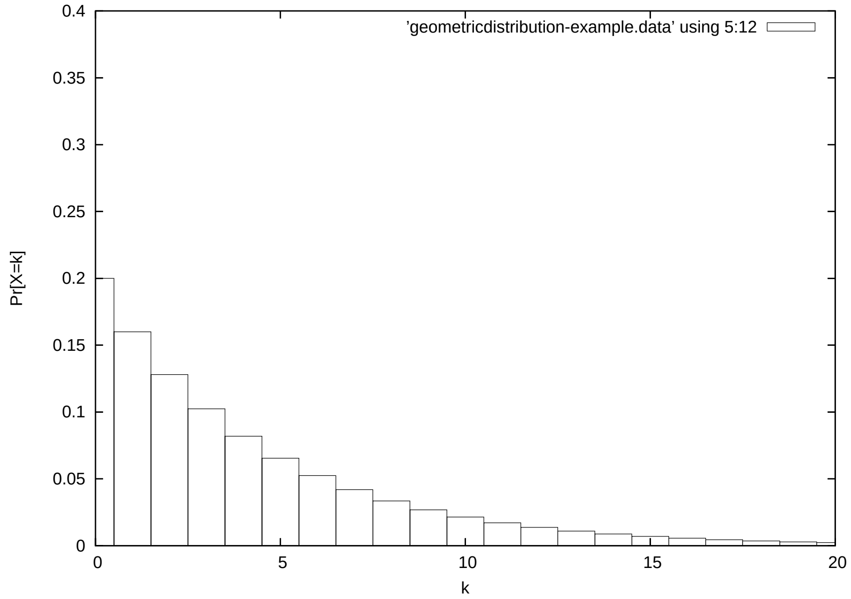

Figure 2: Probability mass function of a geometric random variable X_1 in the first formulation

As an example, the probability mass function for a random variable X_1 in the first formulationis shown in Figure 2. As is to be expected, there is a geometric decay in the probability mass function.

6.1 An Example: The Send-And-Wait Protocol

Here we discuss an example of how some of these discrete probability distributions can arise in a computer network under some idealizing assumptions. This is purely optional, somewhat dense, and perhaps best worked through after four or five weeks into the course.

A *binary symmetric channel* (BSC) is a standard error model for digital communications channels. In a nutshell, each transmitted bit is subjected to an independent Bernoulli experiment with “success” probability $p \in [0, 1]$, i.e. a bit (irrespective of its value) is received correctly with probability $1 - p$ and is received erroneously with probability p. The *bit error rate* (BER) p is assumed constant over time.

Let us consider a single transmitter and a single receiver, connected through a BSC-type channel with BER p. Data is transferred in packets. Both stations use the following procedure (known as the send-and-wait protocol) to transmit data packets: the transmitter equips each packet with a checksum (we assume it is perfect, i.e. no errors will slip through undetected) and sends it. The receiver checks the checksum. If it is wrong, the receiver discards the packet, otherwise it is accepted and an acknowledgement is sent. For simplicity, we assume that the acknowledgement is transmitted error-free. If the transmitter does not get the acknowledgement, a timer expires and the packet is retransmitted. The transmitter performs retransmissions until it receives an acknowledgement, the number of trials is not bounded. The packet under consideration has length l bits. How many trials are needed until the packet is delivered successfully? And how does this number depend on the BER p?

A packet experiences a transmission error when at least one bit is wrong. The probability that all l bits are correct and thus the packet is correct is given by:

$$Q = (1-p)^l$$

due to the independence assumption for bit errors. A trial to transmit a packet corresponds again to a Bernoulli experiment with success probability Q, and because of the BSC assumption subsequent packet trials are independent. Hence, the number X of packet trials needed until the receiver got the packet successfully is a geometric random variable of the second kind, and the expected number of trials is given by

$$E\left[X\right] = \frac{1}{Q} = (1-p)^{-l} \tag{2}$$

It is straightforward to see that for fixed packet length l we have that $E\left[X\right] \to \infty$ as $p \to 1$, and similarly, for fixed bit error probability p we have $E\left[X\right] \to \infty$ as $l \to \infty$.

Let us take this example a bit further. Packets in a communication networks cannot simply be considered as a pile of l user data bits. Instead, communication protocols have a fixed-size overhead, which is part of each packet and which reduces the amount of user data carried. The overhead takes the shape of packet headers and trailers. Let us assume that the overhead corresponds to o bits, and the user data has size s bits, thus $l = o + s$. Without considering the overhead, Equation 2 tells that we should choose l very small in order to keep the number of expected trials per packet small. However, taking the fixed overheads into account, we wish to find the optimal "goodput", i.e. we want to transmit each user data bit with as small overhead as possible.

The goodput for fixed bit error rate p and a user data size of s bits can be defined as the ratio of s to the average total amount of bits transmitted to deliver these s bits:

$$\eta(s) = \frac{s}{E\left[X\right](o+s)} = \frac{s}{(1-p)^{-(o+s)}(o+s)}$$

The optimal value s^* which maximizes $\eta(\cdot)$ can be found from standard calculus techniques (i.e. solving $\eta'(s) = 0$ for s and checking the second derivative):

$$s^* = s^*(p) = \frac{-o}{2} - \frac{1}{2\log(1-p)} \cdot \sqrt{o\log(1-p)\left(o\log(1-p)-4\right)}$$

The result is a real number. In practice, one of the neighboring integers will be taken. As a numerical example, we fix $o = 100$ bits and compute the optimal user data size s^* for varying BER p. The results are shown in the following table:

p	$s^*(p)$
0.1	8.729221
0.05	16.705116
0.01	61.57923
0.005	99.833176
0.001	270.0801
5.0E-4	399.93408
1.0E-4	951.14136
1.0E-5	3110.5203

7 A Helpful Formula

A very helpful formula which you really should know is the so-called **geometric series**: Be $x \in \mathbb{R}$ a real number with $|x| < 1$. Then we have:

$$\sum_{k=0}^{\infty} x^k = 1 + x + x^2 + x^3 + \ldots = \frac{1}{1-x}$$

Note that this formula is valid only for $|x| < 1$. To prove this formula, one can proceed as follows: we define S_n as the partial series

$$S_n = \sum_{k=0}^{n} x^k = 1 + x + x^2 + \ldots + x^n$$

Multiplying S_n by x and subtracting the result from S_n gives

$$\begin{aligned} S_n - x \cdot S_n &= (1 + x + x^2 + \ldots + x^n) - x \cdot (1 + x + x^2 + \ldots + x^n) \\ &= 1 + x + x^2 + \ldots + x^n - (x + x^2 + x^3 + \ldots + x^{n+1}) \\ &= 1 - x^{n+1} \end{aligned}$$

Now using that $S_n - x \cdot S_n = S_n(1-x)$ and dividing both sides by $1 - x$ gives

$$S_n = \frac{1 - x^{n+1}}{1-x}$$

In this formula we now let $n \to \infty$. Since $|x| < 1$, the term x^{n+1} will converge to zero and we get the formula for the geometric series.