Notes on Discrete Probability

Markus Bläser, Saarland University

1 Sample spaces, events, probability

Definition 1. *1. A finite probability space is a finite set* Ω *of* outcomes *or* elementary events *together with a function* $\Pr : \Omega \to \mathbb{R}$*, the* probability measure *or* probability distribution*, such that*

$$\Pr(\omega) \geq 0 \qquad \text{for all } \omega \in \Omega,$$
$$\sum_{\omega \in \Omega} \Pr(\omega) = 1.$$

2. An event A *is any finite subset of* Ω*. Its probability is*

$$\Pr(A) = \sum_{\omega \in A} \Pr(\omega).$$

It follows immediately from Definition 1 that

$$\Pr(\emptyset) = 0,$$
$$\Pr(\Omega) = 1,$$
$$\Pr(\omega) \leq 1 \qquad \text{for all } \omega \in \Omega.$$

Example 2. 1. If we throw a fair coin, then $\Omega = \{H, T\}$ consists of the two events heads H and tail T. Each of the two events has the probability $\Pr(H) = \Pr(T) = 1/2$. Often, we will identify H with 1 and T with 0. A coin is called p-biased, $0 \leq p \leq 1$, if $\Pr(H) = p$ and $\Pr(T) = 1 - p$ instead.

2. If we throw a fair dice, then $\Omega = \{1, 2, 3, 4, 5, 6\}$ and $\Pr(\omega) = 1/6$ for all $\omega \in \Omega$. The event of throwing an even number is $\{2, 4, 6\}$, its probability is $3 \cdots 1/6 = 1/2$.

3. If we throw ℓ fair coins independently, then $\Omega = \{0, 1\}^\ell$. We have $\Pr(x) = 2^{-\ell}$ for each $x \in \Omega$. Note that process is the same as drawing ℓ-bit strings at uniformly random. If we use a p-biased coin instead, then $\Pr(x) = p^k(1-p)^{\ell-k}$ where k is the number of 1s in x.

In all three examples above (in the first and third example only when using a fair coin), each elementary event has the same probability. Such is distribution is called *uniform.*

Proposition 3. *Let* A *and* B *be events of a finite discrete probability space. Then*

1. $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$,
2. $\Pr(A \cup B) = \Pr(A) + \Pr(B)$, *if* A *and* B *are disjoint, that is,* $A \cap B = \emptyset$,
3. $\Pr(\bar{A}) = 1 - \Pr(A)$,
4. $\Pr(A) \leq \Pr(B)$, *if* $A \subseteq B$.

Proof. The second item follows immediately from the fact that the probability of an event is the sum of the probabilities of the elementary events in it.

For the first item, note that we can write

$$\begin{aligned} A \cup B &= (A \cap B) \cup (A \setminus B) \cup (B \setminus A), \\ A &= (A \cap B) \cup (A \setminus B), \\ B &= (A \cap B) \cup (B \setminus A). \end{aligned}$$

Each of these unions is disjoint by construction. Therefore, by the second item,

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A \cap B) \cup \Pr(A \setminus B) \cup \Pr(B \setminus A), \\ \Pr(A) &= \Pr(A \cap B) \cup \Pr(A \setminus B), \\ \Pr(B) &= \Pr(A \cap B) \cup \Pr(B \setminus A). \end{aligned}$$

From this, the first item follows.

The third item follows from the second by $A \cup \bar{A} = \Omega$. The fourth item follows from

$$\Pr(B) = \Pr(\underbrace{A \cap B}_{=A}) + \Pr(B \setminus A).$$

and the fact that probabilities are nonnegative. □

The following simple looking bound is often very useful in estimating probabilities.

Corollary 4 (Union bound)**.** *For two events* A *and* B *of a finite probability space, we have*

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B).$$

Note that the union bound can be extended to arbitrary finite unions in the obvious way by an easy induction argument.

Definition 5. *Two events* A *and* B *of a finite probability space are independent if*

$$\Pr(A \cap B) = \Pr(A) \Pr(B).$$

Otherwise, A *and* B *are dependent.*

2 Conditional probabilities

Definition 6. *Let* $(\Omega, \Pr)$ *be a finite probability space. For any two events, the* conditional probability *of* A *given* B *is*

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

Example 7. Here is a popular example that intuition and conditional probabilities often do not go together. Assume a family has two kids and we know that one is a boy. What is the probability that both kids are boys. (We assume that the probability of having a boy is $1/2$ and that the sex of the first child is independent of the sex of the second child.) $\Omega = \{gg, gb, bg, bb\}$ and each elementary event has probability $1/4$. Now,

$$\Pr(\{bb\}|\{bb, bg, gb\}) = \frac{\Pr(\{bb\})}{\Pr(\{bb, gb, bg\})} = \frac{1}{3}.$$

So contrary to popular belief, the probability is $\frac{1}{3}$ and not $\frac{1}{2}$. On the other hand, the probability that both are boys given that the older kid is a boy is indeed $\frac{1}{2}$.

Theorem 8 (Bayes' rule)**.** *Let* $(\Omega, \Pr)$ *be a finite probability space. For any two events* A *and* B *with* $\Pr(A), \Pr(B) > 0$*,*

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Proof. We have

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

and

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

The statement follows. □

Theorem 9 (Law of total probability)**.** *Let* $(\Omega, \Pr)$ *be a finite probability space. Let* $B_1, \ldots, B_n$ *be a partition of* Ω*, that is,* $B_1 \cup \cdots \cup B_n = \Omega$ *and* $B_i \cap B_j = \emptyset$ *for* $i \neq j$*. Assume that* $\Pr(B_i) > 0$ *for all* i*. Then for any event* A*,*

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A|B_i)\Pr(B_i).$$

Proof. We have

$$\Pr(A|B_i)\Pr(B_i) = \Pr(A \cap B_i).$$

The events $A \cap B_i$ are pairwise disjoint, therefore,

$$\sum_{i=1}^{n} \Pr(A \cap B_i) = \Pr(\bigcup_{i=1}^{n} A \cap B_i) = \Pr(A).$$

For the last equality note that each element of A appears in exactly one B_i. □

Corollary 10. *Let* $(\Omega, \Pr)$ *be a finite probability space. Let* A *and* B *events with* $0 < \Pr(B) < 1$*. Then*

$$\Pr(A) = \Pr(A|B)\Pr(B) + \Pr(A|\bar{B})\Pr(\bar{B}).$$

Together, Theorems 8 and 9 yield the following useful rule

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B|A)\Pr(A) + \Pr(B|\bar{A})\Pr(\bar{A})}.$$

Example 11. Assume that 1 out of 1000 persons has a certain disease. This disease can be tested with a certain test. If a patient is affected with the disease, the test will predict this with probability 0.99. However, it happens that with probability 0.02, the test will predict that a healthy patient is sick. You are tested and the test is positive. How scared shall you be?

Let D be the event that a patient has the disease and let P be the event that the test is positive. We have

$$\begin{aligned} \Pr(D) &= 0.001, \\ \Pr(P|D) &= 0.99, \\ \Pr(P|\bar{D}) &= 0.02. \end{aligned}$$

We want to compute $\Pr(D|P)$. We know by Bayes' rule

$$\Pr(D|P) = \frac{\Pr(P|D)\Pr(D)}{\Pr(P)}$$

and by the law of total probability

$$\Pr(P) = \Pr(P|D)\Pr(D) + \Pr(P|\bar{D})\Pr(\bar{D}) = 0.99 \cdot 0.001 + 0.02 \cdot 0.999 = 0.02097.$$

Thus

$$\Pr(D|P) = 0.99 \cdot 0.001/0.0207 = 0.0472103.$$

So there is only little reason to be scared. (Except for the fact that you might be treated by doctors that do not know probability theory.) Of course, there is an easy solution: Apply an second different (independent) test to all patients with positive first test.

Exercise 12. Do the calculations for the second test. Assume that this is successful with probability 0.98 for sick patients and errs with probability 0.03 for healthy patients.

Proposition 13. *Let* A *and* B *events of a finite discrete probability space with* $\Pr(A), \Pr(B) \neq 0$*. The following statements are equivalent:*

1. $\Pr(A \cap B) = \Pr(A)\Pr(B)$*, that is,* A *and* B *are independent.*
2. $\Pr(A|B) = \Pr(A)$*.*
3. $\Pr(B|A) = \Pr(B)$*.*

Proof. This follows immediately from

$$\Pr(A \cap B) = \Pr(A|B)\Pr(B) = \Pr(B|A)\Pr(A).$$

□

3 Random variables

Let $(\Omega, \Pr)$ be a finite probability space. A function $X : \Omega \to \mathbb{R}$ is called a *random variable*. We are usually interested in the events $X^{-1}(r)$ for some $r \in \mathbb{R}$, that is, all elementary events that are mapped to the value r. We will usually write $\Pr(X = r)$ for the probabilities of these events instead of $\Pr(X^{-1}(r))$. In the same way, we will write $\Pr(X \leq r)$ for the probability of the event $X^{-1}((-\infty, r])$.

Definition 14. *Let X be a random variable on a finite probability space.*

1. *The function $f : \mathbb{R} \to [0, 1]$ given by $f(r) = \Pr(X = r)$ is called the probability mass function of X.*
2. *The function $F : \mathbb{R} \to [0, 1]$ given by $F(r) = \Pr(X \leq r)$ is called the cumulative distribution function of X.*

Note that f only takes finitely many nonzero values. The function F is piecewise constant, monontonic increasing, and right-continuous. There are value r_0 and r_1 such that $F(r) = 0$ for all $r < r_0$ and $F(r) = 1$ for all $r > r_1$. (When the probability space is not finite or even not countably infinite, things get more complicated.)

Example 15. We generate ℓ-bit strings at random. X assigns each string the number of 1's in it. There are $\binom{\ell}{k}$ strings with k 1's. Therefore, the mass function is

$$f(k) = \binom{\ell}{k} \cdot 2^{-\ell} \qquad k = 0, \ldots, \ell.$$

This is a so-called *binomial distribution* with parameter $1/2$. If we generate the strings by setting a bit to 1 with probability p and to 0 with probability $1 - p$, then

$$f(k) = \binom{\ell}{k} \cdot p^k (1 - p)^{\ell - k} \qquad k = 0, \ldots, \ell.$$

Example 16. Now let Y be the random variable that maps an ℓ-bit string to the position of the first 1 in it. The string 0^ℓ is mapped to $\ell + 1$. We have

$$f(k) = \begin{cases} 2^{-k} & k = 1, \ldots, \ell \\ 2^{-n} & k = \ell + 1 \end{cases}$$

This is a (truncated) *geometric distribution.*

Definition 17. *The* expected value *or* expectation *of a random variable X on a finite discrete probability space $(\Omega, \Pr)$ is*

$$\mathsf{E}[X] = \sum_{x \in X(\Omega)} x \cdot \Pr(X = x).$$

Here, $X(\Omega)$ denotes the image of X. Note that his is a finite set.

Example 18. Consider the binomial distribution with parameter p. We have

$$\mathsf{E}[X] = \sum_{k=0}^{\ell} k \cdot \binom{\ell}{k} \cdot p^k (1 - p)^{\ell - k}.$$

We want to find a closed formula for this. Recall that

$$(x+1)^\ell = \sum_{k=0}^{\ell} \binom{\ell}{k} \cdot x^k.$$

If we differentiate with respect to x on both sides, we get

$$\ell \cdot (x+1)^{\ell-1} = \sum_{k=0}^{\ell} \binom{\ell}{k} \cdot k \cdot x^{k-1}.$$

Multiplying by x, we obtain

$$\ell \cdot x(x+1)^{\ell-1} = \sum_{k=0}^{\ell} \binom{\ell}{k} \cdot k \cdot x^k.$$

Setting $x = p/(1-p)$, we get

$$\ell \cdot \frac{p}{1-p} \left(\frac{1}{1-p}\right)^{\ell-1} = \sum_{k=0}^{\ell} \binom{\ell}{k} \cdot k \cdot \left(\frac{p}{1-p}\right)^k$$

and hence $\mathsf{E}[X] = \ell p$.

Given two random variables X and Y on the same finite probability space, their product distribution is defined as

$$\Pr(X = x \wedge Y = y) = \Pr(\{\omega | X(\omega) = x \wedge Y(\omega) = y\}).$$

Given a product distribution, its marginal distributions are defined by

$$\sum_{y \in Y(\Omega)} \Pr(X = x \wedge Y = y)$$

and the corresponding expressing summing over x. Since the events $A_y = \{\omega \mid Y(\omega) = y\}$ form a partition of Ω, it follows that the marginal distribution is

$$\Pr(X = x) = \sum_{y \in Y(\Omega)} \Pr(X = x \wedge Y = y).$$

The following theorem looks innocent, but it turns out to be quite powerful.

Theorem 19. *The expected value is linear, that is, given two random variables X and Y on the same finite probability space and a scalar $\lambda \in \mathbb{R}$, we have*

$$\mathsf{E}[X+Y] = \mathsf{E}[X] + \mathsf{E}[Y],$$
$$\mathsf{E}[\lambda X] = \lambda E[X]$$

Proof. We have

$$\begin{aligned}
\mathsf{E}[X+Y] &= \sum_z z \cdot \Pr(X+Y=z) \\
&= \sum_x \sum_y (x+y) \cdot \Pr(X=x \wedge Y=y) \\
&= \sum_x \sum_y x \cdot \Pr(X=x \wedge Y=y) + \sum_y \sum_x y \cdot \Pr(X=x \wedge Y=y) \\
&= \sum_x x \sum_y \Pr(X=x \wedge Y=y) + \sum_y y \sum_x \Pr(X=x \wedge Y=y) \\
&= \sum_x x \cdot \Pr(X=x) + \sum_y y \cdot \Pr(Y=y) \\
&= \mathsf{E}[X] + \mathsf{E}[Y].
\end{aligned}$$

(Note that all sums are finite.) For the second item, we note that for $\lambda \neq 0$,

$$\begin{aligned}
\mathsf{E}[\lambda X] &= \sum_x x \cdot \Pr(\lambda X = x) \\
&= \lambda \sum_x \frac{x}{\lambda} \cdot \Pr(X = x/\lambda) \\
&= \lambda \sum_y y \cdot \Pr(X = y) \\
&= \lambda \mathsf{E}[X].
\end{aligned}$$

When $\lambda = 0$, then the statement is trivial. □

The above theorem generalizes to arbitrary finite sums in the obvious way be using induction.

Example 20. We have another look at the binomial distribution with parameter p. Let X_k be the random variable that is 1, when the kth bit is 1 and 0 otherwise. We have

$$\mathsf{E}[X_k] = 1 \cdot p + 0 \cdot (1-p) = p.$$

Let X be the random variable that assigns each string the number of 1s in it. We have $X = X_1 + \cdots + X_\ell$. By linearity of expectation,

$$\mathsf{E}[X] = \mathsf{E}[X_1] + \ldots \mathsf{E}[X_\ell] = \ell p.$$