Random Variables and Coupon Collector

Prof. Jara Uitto *Antonio Cruciani*

1 Events and Probability

We recap basics in probability theory based on the book by Mitzenmacher and Upfal [1].

Some Definitions. A probabilistic statement is always made with respect to some probability space.

Definition 1. *A* probability space *is a triple* $(\Omega, \mathcal{F}, \Pr)$*, where:*

- Ω *is the sample space, i.e., the set of all possible outcomes of a random experiment.*
- $\mathcal{F} \subseteq 2^{\Omega}$ *is a family of events, i.e., subsets of* Ω*.*
- $\Pr : \mathcal{F} \to \mathbb{R}$ *is the probability function that assigns values to events such that:*
  - $\Pr(E) \in [0, 1]$ *for all* $E \in \mathcal{F}$*,*
  - $\Pr(\Omega) = 1$*,*
  - *For any finite or countable sequence of disjoint events* $E_1, E_2, \ldots,$ *we have* $\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr(E_i)$*.*

In this course, the sample space Ω will be mostly finite and $\mathcal{F}$ will be equal to the power set 2^{Ω}.

We use standard set notation for events:

- $E_1 \cap E_2$ means both events E_1 and E_2 occur,
- $E_1 \cup E_2$ means either or both of the events occur,
- $E_1 \setminus E_2$ means E_1 occurs, but E_2 does not,
- $\overline{E}$ means that E does not occur, which happens with probability $1 - \Pr[E]$.

Example 1. *Suppose we roll two dice, so* $\Omega = [1, 6] \times [1, 6]$*. Let* E_1 *be the event that the first die shows 1:* $E_1 = \{(1, j) : j \in [1, 6]\}$*. Let* E_2 *be the event that the second die shows 1:* $E_2 = \{(i, 1) : i \in [1, 6]\}$*.*

- $E_1 \cap E_2 = \{(1, 1)\}$ *is the event that both dice show 1.*
- $E_1 \setminus E_2$ *is the event that the first die shows 1 but the second does not.*

Lemma 2. *Let* $E_1, E_2, \ldots$ *be events.*

- $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$
- $\Pr\left[\bigcup_{i\geq 1} E_i\right] \leq \sum_{i\geq 1} \Pr(E_i)$ *this is called the* union bound.

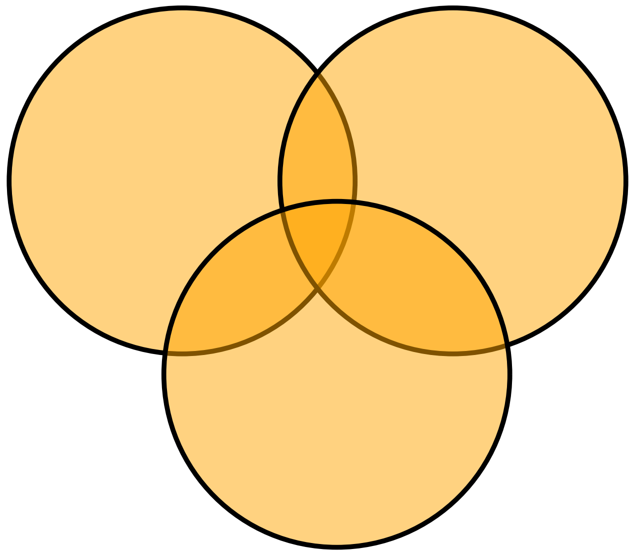

Figure 1: Area of union is bounded by sum of areas of the circles.

Surprisingly, the events are completely arbitrary, and do not need to be independent. In terms of Figure 1, the union bound just says that the area (i.e., probability mass) in the union is bounded above by the sum of the areas of the circles. The bound is tight if the events are disjoint; otherwise the right-hand side is larger, due to double-counting. (It's like inclusion-exclusion, but without any of the correction terms.). In most of the applications, the events $E_1, E_2, \ldots E_k$ are often *bad events* that we're hoping don't happen; the union bound says that as long as each event occurs with low probability and there aren't too many events, then with high probability none of them occur

Example Email Server: Assume we run a server that sends emails to 5 users: Alice, Bob, Carol, Dave and Eve.

Due to some glitch, each email independently fails to deliver with probability p (say $p = 0.01$). Now we have the following concern: What's the probability that at least one user does not receive any email?

Let us define the *bad* event A_i = email to user i fails, then $\Pr(A_i) = p$, and we want to know

$$\Pr\left(\bigcup_{i=1}^{5} A_i\right) = \Pr(\text{at least one email fails})$$

Using the union bound:

$$\Pr(\text{At least one failure}) \leq \sum_{i=1}^{5} \Pr(A_i) = 5 \cdot p = 5 \cdot 0.01 = 0.05$$

Independence and Conditional Probability

Definition 3. *Given some events, we say:*

- *Two events A and B are independent if and only if* $\Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$.
- *Events $A_1, A_2, \ldots, A_k$ are mutually independent if for all subsets $S \subseteq [1, k]$, it holds that*

$$\Pr\left[\bigcap_{i \in S} A_i\right] = \prod_{i \in S} \Pr(A_i).$$

Definition 4. *The conditional probability of A given B is defined as:*

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

Intuition: Look at the probability of $A \cap B$ while knowing that B happened. This corresponds to reducing the sample space to the cases in which B occurred.

Using the previous example of rolling two dice: Let B be the event that the first die shows 1. Let A be the event that both dice show 1. Then:

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} = \frac{1/36}{6/36} = \frac{1}{6}.$$

Why is that?

We are rolling two dice, thus

- The sample space $\Omega = \{1, 2, 3, 4, 5, 6\} \times \{1, 2, 3, 4, 5, 6\}$
- The total number of outcomes is 36
- Each outcome has the same probability $\Pr(\{i, j\}) = \frac{1}{6}$, $i, j \in [1, 6]$

Let us define the two events:

$$A = \text{ Both dice show 1}$$

Observe that there is only one outcome $(1, 1)$, thus $\Pr(A) = \frac{1}{36}$.

$$B = \text{ first die shows 1}$$

This includes all outcomes of the form $(1, j)$ for $j \in [1, 6]$, i.e., $B = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6)\}$, so $\Pr(B) = \frac{6}{36}$. Putting all together we have that

- $A \cap B = \{(1, 1)\}$, because both dice showing 1 implies the first die showing 1
- $\Pr(A \cap B) = \frac{1}{36}$
- $\Pr(B) = \frac{6}{36}$

If A and B are independent, then $\Pr(A \mid B) = \Pr(A)$, as in the example. The intuition is that if the events are independent, the information that B occurred does not change anything.

Law of Total Probability. The law of total probability characterizes the probability of an event via a case distinction a partition of the sample space.

Theorem 5 (Law of Total Probability)**.** *Let events $E_1, \ldots, E_n$ be mutually disjoint and satisfy $\bigcup_{i=1}^{n} E_i = \Omega$. Then for any event A,*

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \cap E_i) = \sum_{i=1}^{n} \Pr(A \mid E_i) \cdot \Pr(E_i).$$

Proof. By disjointness and the assumption $\bigcup_{i=1}^{n} E_i = \Omega$, we get

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

The second equality follows from the definition of conditional probability. □

2 Discrete Random Variables and Expectation

Random Variables Sometimes we are interested in a function of the occurring event rather than in the event itself. For example, when tossing two dice, we might be interested in the sum of the two numbers rather than in the exact outcome. Random variables model such scenarios.

Definition 6. *A* random variable $X : \Omega \to \mathbb{R}$ *is a real-valued function defined on the sample space. A* discrete random variable *takes only a finite or countably infinite number of values.*

It is common to write "$X = x$" for the event $\{\omega \in \Omega : X(\omega) = x\}$. Then:

$$\Pr(X = x) = \sum_{\omega \in \Omega : X(\omega) = x} \Pr(\omega)$$

Example: When rolling two dice, let X be the random variable representing the sum of the two dice: $X((i, j)) = i + j$ for all $(i, j) \in \Omega$. Then:

$$\Pr(X = 4) = \frac{3}{36} = \frac{1}{12},$$

since the outcomes (1,3), (2,2), and (3,1) all sum to 4.

Independence of Random Variables.

Definition 7. *Random variables X and Y are independent if and only if*

$$\Pr(X = x \ \wedge \ Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$$

for all x and y.

Definition 8. *Random variables $X_1, X_2, \ldots, X_n$ are mutually independent if for any subset $S \subseteq [1, n]$ and values x_i for $i \in S$,*

$$\Pr\left(\bigcap_{i \in S} X_i = x_i\right) = \prod_{i \in S} \Pr(X_i = x_i)$$

Expectation

The expectation of a random variable can be thought of as the "average" value it attains.

Definition 9. *The* expectation *(or expected value) of a discrete random variable X is:*

$$\boldsymbol{E}[X] = \sum_x x \cdot \Pr[X = x],$$

where the sum is over all values in the range of X.

Note: For countably infinite values, the expectation might be unbounded. Example: If X takes value 2^i with probability $1/2^i$ for $i \geq 1$, then:

$$\sum_{i=1}^{\infty} \frac{1}{2^i} = 1 \quad \text{(a valid probability distribution)}$$

but

$$\mathbf{E}[X] = \sum_{i=1}^{\infty} 2^i \cdot \frac{1}{2^i} = \sum_{i=1}^{\infty} 1 = \infty.$$

Dice Example: Let X be the sum of two dice. Then:

$$\mathbf{E}[X] = \frac{1}{36} \cdot 2 + \frac{2}{36} \cdot 3 + \frac{3}{36} \cdot 4 + \cdots + \frac{1}{36} \cdot 12 = 7.$$

Linearity of Expectation

Theorem 10. *Let $X_1, \ldots, X_n$ be random variables with finite expectation. Then:*

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

Proof. Here we prove the statement for two random variables X and Y, the general case follows by induction and is left as an exercise. The summations must be over the ranges of the corresponding random variables:

$$\begin{aligned}
\mathbf{E}\left[X+Y\right] &= \sum_i \sum_j (i+j) \Pr((X=i) \cap (Y=j)) \quad \text{(Definition 1)} \\
&= \sum_i \sum_j i \Pr((X=i) \cap (Y=j)) + \sum_i \sum_j j \Pr((X=i) \cap (Y=j)) \\
&= \sum_i i \sum_j \Pr((X=i) \cap (Y=j)) + \sum_j j \sum_i \Pr((X=i) \cap (Y=j)) \quad \text{(Thm. 5)} \\
&= \sum_i i \Pr(X=i) + \sum j \Pr(Y=j) = \mathbf{E}[X] + \mathbf{E}[Y]
\end{aligned}$$

□

Dice Again: Let X_i be the value shown by die i. Then $\mathbf{E}[X_i] = \frac{1}{6}\sum_{j=1}^{6} j = 21/6 = 3.5$, so $\mathbf{E}[X_1 + X_2] = 3.5 + 3.5 = 7$.

Note: This linearity holds even if the X_i are *not* independent.

Secret Santa: Imagine n friends all put their names into a hat for their Secret Santa. Everyone picks a name at random (including possibly their own). What is the expected number of people who draw their own name?

We have a lot of dependencies in this experiment: "if Alice picks Bob, then Bob cannot pick himself!! This is all dependent!"

Unfortunately, the choices are not independent, they're tied together because everyone is picking from the same hat without replacement.

So how do we compute the expected number of "self-draws" ?

This is where linearity of expectation saves the day!

Let X be the number of people who draw their own name. Define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if } i \text{ draws their own name} \\ 0 & \text{Otherwise} \end{cases}$$

Then,

$$X = \sum_{i=1}^{n} X_i \quad \text{And} \quad \mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i]$$

Even though the X_i's are dependent, linearity of expectation does not care at all. So we can just compute

$$\mathbf{E}[X_i] = \Pr(\text{person } i \text{ draws their own name}) = \frac{1}{n}$$

So

$$\mathbf{E}[X] = \sum_{i=1}^{n} \frac{1}{n} = 1$$

Even with all the dependencies, the expected number of people who draw their own name is always 1, no matter how big n is.

Magic!

A "more practical" example: Consider the following random graph $G(n, p)$, where n is the number of nodes and each edge between n nodes is added independently with probability p.

What is the expected number of isolated vertices?

$$X_i = \begin{cases} 1 & \text{if node } i \text{ is isolated} \\ 0 & \text{Otherwise} \end{cases}$$

Clearly, the total number of isolated vertices is

$$X = \sum_{i=1}^{n} X_i$$

By linearity of expectation we have

$$\mathbf{E}[X] = \sum_{i=1}^{n} \mathbf{E}[X_i]$$

Now we compute $\mathbf{E}[X_i]$. Node i is isolated if none of the $n-1$ possible edges between i and other vertices exists. Each of those $n-1$ edges is included independently with probability p, so the probability that none of them exists is

$$\Pr(\text{node } i \text{ is isolated}) = (1-p)^{n-1}$$

So

$$\mathbf{E}[X] = n \cdot (1-p)^{n-1}$$

Even though X_i's are dependent (whether one vertex is isolated affects others) the linearity of expectation lets us ignore that. We don't have to compute nasty joint probabilities.

Lemma 11. *For any constant c and random variable X, we have:*

$$\boldsymbol{E}[cX] = c \cdot \boldsymbol{E}[X]$$

Conditional Expectation Analogously to the conditional probability, we can define the conditional expectation of a random variable.

Definition 12. *The conditional expectation of X given $Y = y$ is:*

$$\boldsymbol{E}[X \mid Y = y] = \sum_{x} x \cdot \Pr[X = x \mid Y = y]$$

where the sum is over all x in the range of X, i.e., all x such that $X(r) = x$ for some r.

Dice Example: Suppose we independently roll two dice. Let X_i be the number shown on the i'th die and let X be the sum of the two. Then we have

$$\mathbf{E}[X \mid X_1 = 2] = \sum_{x} x \cdot \Pr(X = x \mid X_1 = 2) = \sum_{x=3}^{8} x \cdot \frac{1}{6} = 5.5$$

We can get something analogous to the law of total probability for the expectation of a random variable X. In this case, the case distinction is over the possible values y of a second random variable Y. The *law of total expectation* states that the expectation of X can be computed as the sum over the possible outcomes y of Y of the conditional expectation of X assuming that $Y = y$ times the probability that we have $Y = y$.

Lemma 13 (Law of total expectation)**.**

$$\boldsymbol{E}[X] = \sum_y \boldsymbol{E}[X \mid Y = y] \cdot \Pr(Y = y)$$

where the sum is over all y in the range of Y, i.e., all Y such that $Y(r) = y$ for some r.

Proof.

$$\sum_y \mathbf{E}[X \mid Y = y] \cdot \Pr(Y = y) = \sum_y \Pr(Y = y) \cdot \sum_x x \cdot \Pr(X = x \mid Y = y)$$
$$= \sum_y \Pr(Y = y) \sum_x \frac{\Pr(X = x \cap Y = y)}{\Pr(Y = y)} = \sum_x \sum_y x \cdot \Pr(X = x \cap Y = y)$$
$$= \sum_x x \cdot \Pr(X = x) = \mathbf{E}[X]$$

□

Before we showed the linearity of expectation for random variables that can be written as sum of other random variables. The same holds for the conditional expectation of such random variables.

Lemma 14. *Given n random variables $X_1, \ldots, X_n$ with finite expectation, then*

$$\boldsymbol{E}\left[\sum_{i=1}^{n} X_i \mid Y = y\right] = \sum_{i=1}^{n} \boldsymbol{E}[X_i \mid Y = y]$$

We can also look at $\mathbf{E}[Y \mid Z]$ (this is different from $\mathbf{E}[Y \mid Z = z]$). We should think of this as a function from the range of Z to the reals that takes the value $\mathbf{E}[Y \mid Z = z]$ when $Z = z$. Thus, $\mathbf{E}[Y \mid Z]$ is a random variable itself.

Example (rolling two dice):

$$\mathbf{E}[X \mid X_1] = \sum_x x \cdot \Pr(X = x \mid X_1) = \sum_{x=X_1+1}^{X_1+6} x \cdot \frac{1}{6} = X_1 + 3.5$$

We can also look at the expectation of $\mathbf{E}[Y \mid Z]$. The following theorem shows that Z becomes irrelevant in this case.

Theorem 15.

$$\boldsymbol{E}[\boldsymbol{E}[Y \mid Z]] = \boldsymbol{E}[Y]$$

Proof. We have

$$\mathbf{E}[\mathbf{E}[Y \mid Z]] = \sum_z \mathbf{E}[Y \mid Z = z] \cdot \Pr(Z = z) = \mathbf{E}[Y]$$

According to the law of total expectation. □

3 Random Variables

This and the following section is based on Chapters 2.2 and 2.4 in the book by Mitzenmacher and Upfal [1]. We review three common random variables.

Bernoulli random variables. Assume to flip a coin that lands head with probability p. This can be modeled by a Bernoulli random variable:

Definition 16. *A Bernoulli random variable X with parameter p satisfies*

$$\Pr(X = i) = \begin{cases} p, & \text{if } i = 1 \\ 1 - p, & \text{Otherwise} \end{cases}$$

For the expectation of a Bernoulli random variable X, we have

$$\mathbf{E}[X] = p \cdot 1 + (1 - p) \cdot 0 = p = \Pr(X = 1)$$

A Bernoulli random variable can be used to model any scenario where we do a random experiment that succeeds with some probability p.

Binomial Random Variables. Assume to flip a coin n consecutive times. The overall number of heads can be modeled by a Binomial random variable.

Definition 17. *A binomial random variable X with parameters n and p, satisfies for every $0 \leq k \leq n$,*

$$\Pr(X = k) = \binom{n}{k} p^k (1 - p)^k$$

That is, the binomial random variable X is equal to k if and only if there are k heads and $n-k$ tails in the sequence of n coin flips. By using the binomial formula we can see that $\sum_{k=0}^{n} \Pr(X = k) = 1$, a necessary condition for the defined probability function to be well defined. The expectation of a binomial random variable can be written as the sum of n-Bernoulli random variables. Thus, $\mathbf{E}[X] = np$ (using the linearity of expectation).

Geometric Random Variables. Assume we flip a coin until we get our first heads. What is the distribution of the number of flips we need? How long can we expect to wait until we see heads?

Definition 18. *A* geometric random variable X *with parameter p satisfies, for every $k = 1, 2, \ldots,$*

$$\Pr(X = k) = (1 - p)^{k-1} p.$$

Using a geometric series argument (with basis $1 - p$), we get that

$$\sum_{k=1}^{\infty} \Pr(X = k) = \sum_{k=1}^{\infty} p(1 - p)^{k-1} = p \sum_{k=1}^{\infty} (1 - p)^{k-1} = p \cdot \frac{1}{1 - (1 - p)} = p \cdot \frac{1}{p} = 1.$$

Again, this is necessary for the probability function to be well defined.

Lemma 19. *A geometric random variable is memory-less, i.e., for any $k > 0$, it holds that*

$$\Pr(X = k + t \mid X > t) = \Pr(X = k).$$

Lemma 20. *For the expected value of a geometric random variable, it holds that $\mathbb{E}[X] = 1/p$.*

Proof. We have that

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k \Pr[X = k] = \sum_{k=1}^{\infty} \sum_{i=1}^{k} \Pr[X = k] = \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} \Pr[X = k] = \sum_{i=1}^{\infty} \sum_{k=i}^{\infty} (1-p)^{k-1} p.$$

$$= \sum_{i=1}^{\infty} (1-p)^{i-1} p \sum_{k=0}^{\infty} (1-p)^k = \sum_{i=1}^{\infty} (1-p)^{i-1} p \cdot \frac{1}{p} = \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

□

There is also another approach to computing this expectation using the memory-less property.

4 Coupon Collector

Consider the following scenario. We are regularly buying some item, say boxes of cereals, each of which contains one of n different coupons. Once we own all n different coupons we win a prize. Suppose that the coupons in the boxes are chosen independently and uniformly at random from all n different coupons. We are interested in the following question:

"how many times do you need to buy items until we have collected all the different coupons?"

We can model this problem by using a random variable X that denotes the number of buys we need, and we try to find $\mathbf{E}[X]$. For this purpose, we define X_i to be the number of times you buy a box while owning $i - 1$ different coupons. Then clearly

$$X = \sum_{i=1}^{n} X_i.$$

If we collected $i - 1$ different coupons so far, the probability of getting one that we do not own yet is $p_i = \frac{n-(i-1)}{n} = \frac{n-i+1}{n}$. Hence, X_i is a geometric random variable with parameter p_i, and

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}.$$

Therefore,

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \cdot \sum_{i=1}^{n} \frac{1}{i} = n \cdot H_n = n \cdot (\log n + \Theta(1)),$$

where H_n is the n-th harmonic number.

Fact 21. *The harmonic number* $H(n) := \sum_{i=1}^{n} \frac{1}{i}$ *satisfies:*

$$\log n \leq H(n) \leq \log n + 1.$$

Proof. Recall that $\log n = \int_1^n \frac{1}{x} dx$. Moreover,

$$H(n) - 1 = \sum_{i=2}^{n} \frac{1}{i} \leq \int_1^n \frac{1}{x} dx \leq \sum_{i=1}^{n} \frac{1}{i} = H(n).$$

□

In summary, the expected number of boxes we need to buy is at most $n \log n + \Theta(n)$. This is (maybe) surprisingly little, as it is only a logarithmic factor more than the obvious lower bound of n.

References

[1] Mitzenmacher, Michael and Upfal, Eli, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis, Cambridge university press, 2017*,