Module 7: Discrete Probability

Theme 1: Elementary Probability Theory

Probability is usually associated with an outcome of an experiment. For example, the experiment may be a throw of a coin, while the two possible outcomes are "heads" and "tails". Roughly speaking, probability will estimate our chance that the next outcome of this experiment is either a head or a tail (here we assume that tail and head are equally likely, that is, the probability of tossing a head or a tail is equal to 0.5 or 50%).

An **experiment** is a procedure that gives a set of possible outcomes. In fact, the set of **all** possible outcomes is called the **sample space** (e.g., in the experiment with a coin, the sample space$=\{\text{head},\ \text{tail}\}$). Finally, an **event** is a subset of the sample space (e.g., a *head*). When there are a finite number of *equally likely* outcomes, Laplace suggested the following definition of the probability:

> The probability of an event $E \subset S$ (which is a subset of a finite sample space S) of equally likely outcomes is
>
> $$P(E) = \frac{|E|}{|S|}$$
>
> where $|E|$ and $|S|$ are cardinalities of the sets E and S, respectively. We often call the events in E *favorable events*, while events in S *all possible events*.

Example 1: A box has 5 black balls and 7 green balls. What is the probability of selecting a green ball?

The sample space S consists of 12 balls. The event $E = \{\text{select a green ball}\}$ has seven elements. Therefore $P(E) = 7/12$.

Example 2: Let two dice be rolled (we recall that a die has six sides, and each side has one, or two, or ..., six dots). What is the probability that the sum of the numbers on the two dice is 11?

Let us first build the probability space S. It consists of pairs (i, j) where $1 \leq i, j \leq 6$, so we have $|S| = 36$ (since every die has six outcomes, so two of them must have $6 \cdot 6$ outcomes). The event $E = \{\text{sum is equal to } 11\}$ consists of

$$E = \{(5, 6), (6, 5)\},$$

therefore, $P(E) = 2/36 = 1/18$.

The counting problems encountered so far were very simple. Consider now the following problem.

Example 3: Find the probability that a hand of five cards in poker contains four cards of one kind?

We recall that there are 52 cards in a deck; there are 13 different kinds of cards, with 4 cards of each kind. These kinds are two's, three's, . . ., tens, jacks, queens, kings, and aces. There are also four suits: spades, clubs, hearts, and diamonds.

The number of ways to choose 5 cards out of 52 is $C(52, 5)$ (which is a large number). This is the cardinality of the sample space. Let us now consider the event E that a hand has four cards of one kind. By the multiplication rule, a hand of five cards with four cards of the same kind is the number of ways to pick one kind ($C(13, 1) = 13$) and the number of ways to pick the fifth card, which is $C(48, 1)$ (in words, one in every 48). Therefore, by the above definition

$$P(E) = \frac{13 \cdot 48}{C(52, 5)} \approx 0.00024$$

since there are $C(52, 5)$ possible outcomes and $C(13, 1) \cdot C(48, 1)$ "favorable" outcomes for E.

Sometimes, we know the probability of events E_1 and E_2 and need to know the probability of combinations of events such as $E_1 \cup E_2$ (i.e., at least one event occurs), $E_2 \cap E_2$ (both events must occur), or $\bar{E} = S - E$ (E does *not* occur). Let us start with the probability of the complementary event $\bar{E}$. We claim that

$$P(\bar{E}) = 1 - P(E).$$

Indeed, since $|\bar{E}| = |S| - |E|$ we obtain

$$P(\bar{E}) = \frac{|S| - |E|}{|S|} = \frac{|S|}{|S|} - \frac{|E|}{|S|} = 1 - P(E).$$

Example 4: What is the probability that among five randomly generated bits at least one is 0?

This exactly the case when it easier to compute $\bar{E}$ than E. In this case $\bar{E} = \{\text{all bits are } 1\}$. Since there are 2^5 possible binary strings of length five, only only one (i.e., $(0, 0, 0, 0, 0)$) is the "favorable" one, we find

$$P(\bar{E}) = \frac{1}{2^5}$$

since there are 2^5 binary strings of length 5 and there is only one string with all 1s. Hence

$$P(E) = 1 - P(\bar{E}) = 1 - 2^{-5} = \frac{31}{32}.$$

Let us now compute $P(E_1 \cup E_2)$. From previous modules we know that

$$|E_1 \cup E_2| = |E_1| + |E_2| - |E_1 \cap E_2|$$

therefore, by the definition of probability

$$\begin{aligned} P(E_1 \cup E_2) &= \frac{|E_1 \cup E_2|}{|S|} \\ &= \frac{|E_1| + |E_2| - |E_1 \cap E_2|}{|S|} \\ &= \frac{|E_1|}{|S|} + \frac{|E_2|}{|S|} - \frac{|E_1 \cap E_2|}{|S|} \\ &= P(E_1) + P(E_2) - P(E_1 \cap E_2). \end{aligned}$$

In summary, we prove that

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

In words, the probability of union of two events is the sum of the probability of both events minus the the probability of product of the events, When the events are disjoint (i.e., $E_1 \cap E_2 = \emptyset$), then $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

Example 5: What is the probability that a randomly selected positive integer smaller than equal to 100 is divisible either by 2 or by 5?

Let E_1 that the integer is divisible by 2, and E_2 the event that the integer is divisible by 5. Clearly $|E_1| = 50$ and $|E_2| = 20$. Observe that the event we are looking for is $E_1 \cup E_2$. In order to compute it we need $|E_1 \cap E_2| = 10$ since there are ten numbers in the range 1 to 100 that are divisible by 10. Therefore, by the definition of probability we have

$$\begin{aligned} P(E_1 \cup E_2) &= P((E_1) + P(E_2) - P(E_1 \cap E_2) \\ &= \frac{50}{100} + \frac{20}{100} - \frac{10}{100} \\ &= \frac{3}{5}. \end{aligned}$$

Exercise 7A: What is the probability of generating a binary string $(0, 0, 1, 1, 0, 1, 0)$ of length seven provided 0 and 1 are equally likely.

Theme 2: Probability Theory

In the previous section, we assumed that all outcomes of the sample space S are equally likely. This led us to the Laplace definition of probability. Here we generalize it.

Let S be a probability space. Throughout, we assume that S is finite, and often we just list all outcomes (e.g., $S = \{s_1, \ldots, s_n\}$). Any subset A of S will be called an **event**. We now define probability P as a function from the set of subsets of S into the interval $[0, 1]$. If we denote by $\Omega = \{A : A \subset S\}$, then

$$P : \Omega \to [0, 1]$$

such that the following three properties hold (below $P(A)$ denotes the probability of the event A):

1. $P(A) \geq 0$;
2. $P(S) = 1$;
3. if $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$.

The above three properties say that the probability of any event must be nonnegative, that the probability of a "sure" event (i.e., S) is equal to one, and finally that the probability of the union of disjoint events is the sum of the probabilities of the corresponding events.

Using these three assumptions one can prove many properties of probability (that we already encountered in the previous section). For example, let $\bar{A} = S - A$ be the complementary event to A (that is, $\bar{A}$ is the same as **not** A). We have $P(\bar{A}) = 1 - P(A)$. Indeed, observe that $S - A$ and A are disjoint, hence by (c) we find

$$1 = P(S) = P((S - A) \cup A) = P(S - A) + P(A) = P(\bar{A}) + P(A),$$

which proves our claim that $P(\bar{A}) = 1 - P(A)$. By the way, as a corollary we see that

$$P(\emptyset) = P(\bar{S}) = 1 - P(S) = 0.$$

Let now all outcomes in S be **equally likely**. To be more precise, let $S = \{s_1, \ldots, s_n\}$ and

$$P(s_i) = \frac{1}{n}$$

since by the second property above we have $1 = \sum_{i=1}^{n} P(s_i) = nP(s_i)$ (all events sum up to one). Let now $A = \{s_{i_1}, \ldots, s_{i_k}\}$, that is $|A| = k$. By the third property of the probability definition and the above we have

$$\begin{aligned} P(A) &= P(s_{i_1} \cup s_{i_2} \cup \ldots \cup s_{i_k}) \\ &= \sum_{j=1}^{k} P(s_{i_j}) = kP(s_1) \\ &= \frac{k}{n} = \frac{|A|}{|S|}. \end{aligned}$$

In the above we first observe that the event A is a union of the elementary events $A = s_{i_1} \cup s_{i_2} \cup \ldots \cup s_{i_k}$. All elementary events are disjoint, hence we can sum probabilities, as the second line above shows. Finally, since every event is equally likely and there are n events, hence $P(s_1) = 1/n$. We have just recovered Laplace's definition of probability for equally likely outcomes.

Example 6: Find the probability that a randomly selected k-digits *decimal* number is also a valid octal number whose digits are between 0 and 8.

First, a k digit number can be represented as $(x_1, x_2, \ldots, x_k)$ where $x_i \in \{0, 1, \ldots, 9\}$ if the number is decimal, and $x_i \in \{0, 1, \ldots, 7\}$ if the number is octal. The number of decimal numbers of length k is 10^k (just apply the multiplication rule). The number of valid octal numbers of length k is 8^k. Therefore, the probability is $\frac{8^k}{10^k} = \left(\frac{4}{5}\right)^k$.

Conditional Probability

Consider the case when you know that event B has occurred, and knowing this you want to compute the probability of event A. This is known as the **conditional probability** and denoted as $P(A|B)$.

Example 7: There are five black balls and ten green balls in a box. You select randomly a ball, and it happens to be a green ball. You do *not* return this ball to the box. What is the probability that in the second selection you pick up a green ball? If A is the event of selecting a green ball in the first pick, and B is the probability of choosing another green ball in the second pick, then the probability we are seeking is denoted as $P(B|A)$. In our case it is

$$P(B|A) = \frac{9}{14}$$

since after the first selection there are only nine green balls in the box containing 14 balls. (Here we used explicitly the fact that after picking a geen ball there are only 14 balls left with 9 green balls.) We can compute this probability in a different way. Observe that $|S| = 15$ and $|A| = 10$, hence

$$P(A) = \frac{10}{15}.$$

Let us now compute the probability of $A \cap B$. Event A can occur in 10 ways out of 15, while B can occur 9 out of 14 since one ball was already taken out from the box in the pick. Hence

$$P(A \cap B) = \frac{10}{15} \cdot \frac{9}{14},$$

and then we "define" (see below for additional explanations) the conditional probability $P(B|A)$ as

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{10}{15} \cdot \frac{9}{14} \cdot \frac{15}{10} = \frac{9}{14}.$$

Thus, we obtain the same result as the one computed directly. It suggests a definition of conditional probability that we shall discuss next.

Let us generalize this example. Consider a sample space S and two events $A, B \subset S$. Assume event B **has occurred**. Then the sample space S effectively reduces to B, therefore, we must restrict the occurrence of event A to those outcomes that fall into B. In a sense, B is the new sample space. In other words, the number of "favorable outcomes" is not $|A|$ but $|A \cap B|$. Therefore for equally likely events we compute $P(A|B)$ as follows

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

Observe, however, that

$$\begin{aligned} P(A|B) &= \frac{|A \cap B|}{|B|} \\ &= \frac{|A \cap B|}{|S|} \cdot \frac{|S|}{|B|} \\ &= \frac{\frac{|A \cap B|}{|S|}}{\frac{|B|}{|S|}} \\ &= \frac{P(A \cap B)}{P(B)}. \end{aligned}$$

In the second line above, we multiply and divide by $|S|$ and then observe in the third line that we have the probabilities $P(A \cap B)$ and $P(B)$.

Actually, the last expression is used as a definition of the conditional probability.

> Let A and B be events with $P(B) > 0$. The **conditional probability** of A given B, denoted as $P(A|B)$, is defined as
>
> $$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Example 8: A box contains 5000 chips, 1000 of them made by company X, the rest by company Y. It is known that $10\% = 100$ chips made by company X are defective, while only $5\% = 200$ chips made by company Y are defective. Compute the probability that if you pick up a defective chips it comes from company X.

Let A be the event that a chip is made by company X and B that a chip is defective. We need to find $P(A|B)$, that is, the probability that provided a chip is defective it i comes from company X. For this we need $P(B)$ and $P(A \cap B)$. But

$$\begin{aligned} P(B) &= \frac{100 + 200}{5000} = 0.06, \\ P(A \cap B) &= \frac{100}{5000} = 0.02. \end{aligned}$$

Then

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.02}{0.06} = \frac{1}{3},$$

that is, one out of every three.

Independence

If $P(A|B) = P(A)$, then the knowledge of B does not change the probability of A. We say that A and B are **independent events**. Observe that the above condition is equivalent to $P(A \cap B) = P(A)P(B)$, which serves as a definition.

Two events A and B are said to be **independent** if and only if

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

Example 8: Consider a five-bit binary string. The probability of generating a zero is equal to p. Bits are generated independently. What is the probability of getting 00111?

Since we have independence we easily compute

$$P(00111) = P(0) \cdot P(0) \cdot P(1) \cdot P(1) \cdot P(1) = p^2(1-p)^3$$

since $1 - p$ is the probability of generating a one.

Exercise 7B: Show that if A and B are independent events, then $\bar{A}$ and $\bar{B}$ are also independent events.

Binomial Distribution and Bernoulli Trials

In the last example, we generated five bits and asked for the probability of getting 00111. However, if we ask for the probability of generating two 0s and three 1s, the situation is different. This time we do *not* specify where the two 0s and three 1 are located. Therefore, strings like 01011, 11001, etc. satisfy the description of the event. In fact, we have $C(5,2) = C(5,3)$ ways to select two zeros out of five. Thus this probability is equal to

$$C(5,2)p^2(1-p)3 = 10p^2(1-p)^3,$$

and this should be compared with the answer to the previous example. For instance, if $p = 0.1$, then the above becomes

$$C(5,2)0.1^2 \cdot 0.9^3 = 10 \cdot 0.01 \cdot 0.729 = 0.0729.$$

We shall generalize the last situation, and introduce the so called **Bernoulli trials** and the **binomial distribution**. Consider an experiment that has two outcomes called **successes** and **failures**. Let the probability of a success be p, while the probability of a failure $q = 1 - p$. This experiment is called the **Bernoulli trial**. Let us repeat it n times. Many problems in probability can be solved by asking what is the probability of k successes in n Bernoulli trials. The last example can be viewed as five Bernoulli trials with a success being a generation of a zero.

Let us now consider n **independent** Bernoulli trials with the probability of a success equal to p. What is the probability of obtaining k successes. Since the outcomes are independent a *particular* trial

with k successes has the probability $p^k(1-p)^{n-k}$. But we can choose on $C(n,k)$ ways k successes out of n trials, therefore, the probability of k successes in n independent Bernoulli trials is

$$C(n,k)p^k(1-p)^{n-k}. \tag{1}$$

Considered as a function of k, we call the above function the **binomial distribution** and denote it as $b(k;n,p) = C(n,k)p^k(1-p)^{n-k}$.

Observe that (1) is probability since by the definition of probability it sums up to one. More precisely, by Newton's summation formula discussed in Module 5

$$\sum_{k=0}^{n} C(n,k)p^k(1-p)^{n-k} = (p+1-p)^n = 1^n = 1$$

as needed.[1]

Example 9: A biased coin is thrown 7 times. The probability of throwing a tail is 0.4. What is the probability of throwing three tails in four trials?

Clearly, we have the Bernoulli trials with the success being a throw of a tail. Hence, the probability is equal to

$$C(7,3)(0.4)^3 \cdot 0.6 = 0.1536.$$

after substituting $p = 0.4$ in (1).

Random Variables

Many problems are concerned with a numerical values associated with the outcome of an experiment. For example, we can assign value 1 to the tail when throwing a coin, and value 0 when throwing a head. Such a numerical value assigned to an outcome is known as a random variable.

> A **random variable** is a function from the sample space of an experiment to the set of real numbers.

Example 10: Let us flip a coin three times. Define a random variable $X(t)$ to be the number of tails that appear when t is the outcome. We have

$$\begin{aligned} X(HHH) &= 0, \\ X(HHT) &= X(HTH) = X(THH) = 1, \\ X(TTH) &= X(THT) = X(HTT) = 2, \\ X(TTT) &= 3. \end{aligned}$$

[1] We recall that by Newton's formula

$$(a+b)^n = \sum_{k=0}^{n} C(n,k)a^k b^{n-k}.$$

Having defined a random variable, we can now introduce the **probability mass function**. Let $A_t = \{s \in S : \ X(s) = t\}$, that is, A_t is the subset of S (an event) that assigns value t of X. Then

$$P(X = t) = P(A_t) = \sum_{s \in A_t} P(s)$$

since A_t is disjoint union of elementary events s such that $X(s) = t$.

Let us now discuss an important notion of probability theory, namely, the "expected value" of an experiment. For example, one expects about 50 tails when flipping an unbiased coin 100 times. We are now in a position to definite it precisely.

> The **expected value** (also known as the **mean value**) of a random variable $X(s)$ over $s \in S$ taking values in $t \in X(s)$ is defined as
>
> $$\mathbf{E}[X] = \sum_{s \in S} P(s)X(s) = \sum_{t \in X(S)} tP(X = t).$$

The above formula extends the definition of "average value" known from high school. Indeed, let all events $X = t$ are equally likely, and assume that $t = 1, 2, \ldots, n$. WE learned in high school to compute the average (expected value) as follows

$$\mathbf{E}[X] = \frac{1 + 2 + \cdots n}{n} = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n\frac{1}{n} = \sum_{t=1}^{n} tP(X = t)$$

which coincides with the above definition.

Example 11: We shall continue Example 10 assuming that the coin is fair (i.e., probability of a head or a tail is 0.5). From the previous example we find that

$$\begin{aligned} P(X = 0) &= \frac{1}{8}, \\ P(X = 1) &= \frac{3}{8}, \\ P(X = 2) &= \frac{3}{8}, \\ P(X = 3) &= \frac{1}{8}, \end{aligned}$$

since, for example, $\{X = 1\} = \{HHT, THH, HTH\}$, thus we have three out of $2^3 = 8$ outcomes satisfying $X = 1$ (i.e., the number of tails is equal to one). Therefore,

$$\mathbf{E}[X] = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 3 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{15}{8},$$

that is, on average we have $1\frac{7}{8}$ tails per three throws.

Let us now compute the expected value of the binomial distribution defined above. We define X as the number of successes in n Bernoulli trials. Then[2]

$$\begin{aligned}
\mathbf{E}[X] &= \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} kC(n,k)p^k(1-p)^{n-k} \\
&= \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \\
&= \sum_{k=1}^{n} n\frac{(n-1)!}{(k-1)!(n-k)!}p^k(1-p)^{n-k} \\
&= np\sum_{k=1}^{n} C(n-1,k-1)p^{k-1}(1-p)^{(n-1)-(k-1)} \\
&= np\sum_{j=0}^{n-1} C(n-1,j)p^j(1-p)^{n-1-j} \\
&= np(p+1-p)^{n-1} \\
&= np.
\end{aligned}$$

The first line is just the definition of the binomial distribution and the expected value. In the third line we use the following property of the binomial coefficients (see Module 4 and 6):

$$kC(n,k) = \frac{kn!}{k!(n-k)!} = n\frac{(n-1)!}{(k-1)!(n-k)!} = nC(n-1,k-1).$$

In the fourth line above we change the index of summation from k to $j = k-1$, while in the fifth line we apply the Newton summation formula, discussed in Module 4 which we recall below

$$(a+b)^n = \sum_{k=0}^{n} C(n,k)a^k b^{n-k}.$$

(In our case, $a = p$ and $b = 1-p$.)

Expectation has some nice properties. For example,

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

this is, the expectation of the sum of random variables is the sum of expectations. This is very important result! Let us derive it. We have

$$\begin{aligned}
\mathbf{E}[X+Y] &= \sum_{s\in S} P(s)[X(s)+Y(s)] \\
&= \sum_{s\in S} P(s)X(s) + \sum_{s\in S} P(s)Y(s) \\
&= \mathbf{E}[X] + \mathbf{E}[Y].
\end{aligned}$$

[2]This derivation is quite long and can be omitted in the first reading. We shall re-derive the same result in Example 13 using simpler arguments.

Example 13: We just computed that $\mathbf{E}[X] = np$ for binomially distributed X. We needed a long chain of computations. But we can prove the same result using the above property in a much easier way. Observe that

$$X = X_1 + X_2 + \cdots + X_n$$

where X_i is equal to 1 when a success occurs and 0 otherwise. Such a random variable is called the Bernoulli random variable or, more precisely, Bernoulli distributed random variable. Clearly, $\mathbf{E}[X_i] = 1 \cdot p + 0 \cdot (1 - p) = p$. Since the expectation of a sum of random variables is the sum of expectations, we have

$$\mathbf{E}[X] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \cdots + \mathbf{E}[X_n] = np,$$

as before, but this time we derive it in a simple way.

However, in general $\mathbf{E}[XY]$ is not equal to $\mathbf{E}[X]\mathbf{E}[Y]$. To assure this is true one must assume X and Y are independent defined as follows:

Two random variables X and Y on the same sample space S are **independent** if

$$P(X(s) = t, Y(s) = r) = P(X(s) = t) \cdot P(Y(s) = r).$$

Example 14: Let us roll two dice. What is the probability of getting 6 on the die and 5 on the second die. Let X represent the number obtained on the first die and Y the number rolled on the second die. Since the events are independent, we have

$$P(X = 6, Y = 5) = P(X = 6) \cdot P(Y = 5) = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}.$$

We now prove the following result

Theorem 1 *Let X and Y are independent random variables. Then*

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

Proof. We have

$$\begin{aligned}
\mathbf{E}[XY] &= \sum_{t,r} t \cdot rP(X = t, Y = r) \\
&= \sum_{t,r} t \cdot rP(X = t)P(Y = r) \\
&= \sum_{t} tP(X = t) \sum_{r} rP(Y = r) \\
&= \mathbf{E}[X]\mathbf{E}[Y],
\end{aligned}$$

where in the second line we used independence, while in the third line we computed two independent sums.

Finally, we shall discuss **variance**. The expected value of a random variable tells us its average value but says nothing about variability of it. The reader should not forget that X is a random variable and it (randomly) *varies*. While we would like to find one synthetic number (e.g., the expected value) to describe this random variable, such a characterization is usually very poor. Therefore, we try to introduce some parameters that can tall us (in a simplified way) more about the random variable. The variance, roughly speaking, determines how widely a random variable is distributed around the expected value. Formally:

> Let X be a random variable defined on a sample space S. The **variance** of X, denoted as $\mathbf{Var}[X]$, is
>
> $$\mathbf{Var}[X] = \sum_{s \in S} P(s)(X(s) - \mathbf{E}[X])^2 = \mathbf{E}[(X - \mathbf{E}[X])^2].$$

That is, the variance is the *expected value* of the following random variable: $(X - \mathbf{E}[X])^2$. Since we expect that X is more likely to concentrate around $\mathbf{E}[X]$, the random variable $(X - \mathbf{E}[X])^2$ tells us about variations of X around the expected value.

We can compute the variance using the following formula

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2. \tag{2}$$

Indeed,

$$\begin{aligned} \mathbf{E}[(X - \mathbf{E}[X])^2] &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] \\ &= \mathbf{E}[X^2] - 2\mathbf{E}[X]\mathbf{E}[X] + \mathbf{E}[X]^2 \\ &= \mathbf{E}[X^2] - \mathbf{E}[X]^2 \end{aligned}$$

where above we used the fact that the expected value of a sum of random variables is the sum of the expected values and the following identity (let's call it the "square of sum identity")

$$(a + b)^2 = a^2 + 2ab + b^2$$

known from high school.

Example 15: Consider a Bernoulli random variable X taking value 1 with probability p and zero otherwise. What is the variance of X?

We observe first that $\mathbf{E}[X] = 1 \cdot p + 0 \cdot (1 - p) = p$. Then we compute

$$\mathbf{E}[X^2] = 1^2 \cdot p + 0^2 \cdot (1 - p) = p.$$

Thus, a straightforward computation gives us

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2 = p - p^2 = p(1-p) = p \cdot q.$$

Unlike the expectation the variance of a sum of two random variables is *not* the sum of variances. For this to hold, we need additional assumptions, as shown below.

Theorem 2. *Let X and Y be* **independent** *random variables. Then*

$$\mathbf{Var}[X+Y] = \mathbf{Var}[X] + \mathbf{Var}[Y].$$

In general, if X_i, $i = 1, 2, \ldots, n$ are pairwise independent random variables, then

$$\mathbf{Var}[X_1 + X_2 + \cdots + X_n] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + \cdots + \mathbf{Var}[X_n].$$

Proof. From (2) we have

$$\mathbf{Var}[X+Y] = \mathbf{E}[(X+Y)^2] - \mathbf{E}[X+Y]^2.$$

But

$$\begin{aligned}
\mathbf{E}[(X+Y)^2] &= \mathbf{E}[X^2 + 2XY + Y^2] \\
&= \mathbf{E}[X^2] + 2\mathbf{E}[XY] + \mathbf{E}[Y^2] \\
&= \mathbf{E}[X^2] + 2\mathbf{E}[X]\mathbf{E}[Y] + \mathbf{E}[Y^2]
\end{aligned}$$

where in the second line we use the identity $(a+b)^2 = a^2 + 2ab + b^2$ and in the third line we apply independence of X and Y. Summing up, we obtain

$$\begin{aligned}
\mathbf{Var}[X+Y] &= \mathbf{E}[(X+Y)^2] - \mathbf{E}[X+Y]^2 \\
&= \mathbf{E}[(X+Y)^2] - (\mathbf{E}[X] + \mathbf{E}[Y])^2 \\
&= \mathbf{E}[X^2] + 2\mathbf{E}[X]\mathbf{E}[Y] + \mathbf{E}[Y^2] - \mathbf{E}[X]^2 - 2\mathbf{E}[X]\mathbf{E}[Y] - \mathbf{E}[Y]^2 \\
&= (\mathbf{E}[X^2] - \mathbf{E}[X]^2) + (\mathbf{E}[Y^2] - \mathbf{E}[Y]^2 \\
&= \mathbf{Var}[X] + \mathbf{Var}[Y],
\end{aligned}$$

which completes the proof. In the first line we use the fact that $\mathbf{Var}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$ (derived above), then we use again the square of sum identity, then we rearrange terms of the sum, and finally obtain the desired identity.

Example 16: Let us compute the variance of the binomial distribution. We use the representation of binomial distribution from Example 13, that is,

$$X = X_1 + \cdots + X_n$$

where X_i are Bernoulli distributed with $\mathbf{Var}[X_i] = p(1-p)$ as computed in Example 15. Therefore, by the last theorem

$$\mathbf{Var}[X] = \mathbf{Var}[X_1 + \cdots + X_n] = \mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n] = np(1-p).$$

That is, the variance of the sum of Bernoulli distributed random variables is the sum of variances of individual random variables, and it is equal to $np(1-p)$.