

Probability Review

CS 221

Section 3*

Olga Russakovsky

October 12, 2009

1 Random variables

Consider running a probabilistic experiment, e.g., tossing a coin. Let Ω be the set of all possible outcomes of this experiment, called the *sample space*. In this case,

$$\Omega = \{\text{Heads}, \text{Tails}\}$$

If we were to toss a coin 3 times, then

$$\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\}$$

A *random variable* is a function that map outcomes of a probabilistic experiment to a real number. More formally, a random variable X is a function $X : \Omega \rightarrow \mathbb{R}$. For example, we can define a random variable

$$X = \text{the number of heads in two tosses of a coin} \tag{1}$$

If a random variable X takes on only a finite number of values, i.e., $X \in \{0, 1, \dots, n\}$, then it is called a *discrete random variable*. Otherwise, it is called a *continuous random variable*. The example above in (1) is a discrete random variable. The simplest example of a continuous random variable is a random number generator that can return any real number between 0 and 1. In CS221 we will be mostly dealing just with discrete variables.

⁰Many thanks to Quoc Le, Arian Maleki and Tom Do. For more details, please refer to the very thorough probability review handout found at <http://cs229.stanford.edu/section/cs229-prob.pdf>

2 Probability Distributions

2.1 Discrete random variables

For a discrete random variable, a *probability distribution* is a list of probabilities associated with each of its possible values. For example, in the example above, if we have a fair coin, then

$$\begin{aligned}P(X = 0) &= P(\text{TT outcome}) = \frac{1}{4} \\P(X = 1) &= P(\text{HT or TH outcome}) = \frac{1}{2} \\P(X = 2) &= P(\text{HH outcome}) = \frac{1}{4}\end{aligned}$$

For discrete variables, this is called the *probability mass function*, or *PMF*. We denote

$$P(X = x) = P_X(x) = p(x) \quad (2)$$

The requirements imposed on the probability mass function are:

- $p(x) \geq 0$ for all $x \in \Omega$
- $\sum_{x \in \Omega} p(x) = 1$

Additionally, define an *event* A to be a subset of the sample space, i.e., $A \subseteq \Omega$. For example, A might be the event that at least one toss turns up heads, or $X \geq 1$. Then

$$p_X(A) = \sum_{x \in A} p_X(x) \quad (3)$$

So in the case where A is the event of at least one heads,

$$p_X(A) = P(X = 1) + P(X = 2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$$

Two common examples of discrete random variables are

- $X \sim \text{Bernoulli}(p)$ with $0 \leq p \leq 1$ corresponds to taking a coin which has probability p of coming up heads, and flipping it once. X is 1 if heads comes up, 0 otherwise.

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases} \quad (4)$$

- $X \sim \text{Binomial}(n, p)$ with $0 \leq p \leq 1$ corresponds to taking a coin which has probability p of coming up Heads, and flipping it n times. X is the number of heads.

$$p(x) = \binom{n}{k} p^x (1 - p)^{(n-x)} \quad (5)$$

The example discussed above is $\text{Binomial}(2, \frac{1}{2})$.

2.2 Continuous random variables

In case of a continuous random variable, the probability distribution is called the *probability density function*, or *PDF*, and is often denoted $f(x)$. Here for example, Ω can be the set of all real numbers between 0 and 1, and an event A can correspond to the set of all real numbers between $\frac{1}{5}$ and $\frac{3}{5}$. As before, the properties are

- $f(x) \geq 0$ for all $x \in \Omega$
- $\int_{x \in \Omega} f(x) = 1$, so the total probability of all outcomes is 1
- $P(A) = P(x \in A) = \int_{x \in A} f(x)$, so the probability of an event A is just computed as a part of the area under the curve $f(x)$

We won't be using continuous random variables much in CS221, but two common examples are:

- $X \sim \text{Uniform}(a, b)$ with $a < b$ corresponds a random number generator where any real number between a and b is equally likely

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

- $X \sim \text{Normal}(\mu, \sigma^2)$ is also known as a Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \quad (7)$$

2.3 Properties of probability distributions

Consider a probability distribution P and two events A and B . Then

- If $A \subseteq B$, then $P(A) \leq P(B)$
- $P(A \cap B) \leq \min(P(A), P(B))$
- *Union bound* $P(A \cup B) \leq P(A) + P(B)$
- $P(\Omega - A) = 1 - P(A)$

If we impose additional restrictions on the events, we can obtain even more properties:

- If A and B are independent, then $P(A \cap B) = P(A)P(B)$
- If A and B are disjoint, so $A \cap B = \emptyset$, then $P(A \cup B) = P(A) + P(B)$
- By induction, if A_1, A_2, \dots, A_k are disjoint events, so $A_i \cap A_j = \emptyset$ whenever $i \neq j$, then

$$P(\cup_i A_i) = \sum_i P(A_i) \quad (8)$$

- *Law of total probability* If A_1, A_2, \dots, A_k are disjoint events such that $\cup_i A_i = \Omega$, then

$$\sum_i P(\cup_i A_i) = 1 \quad (9)$$

3 Joint distributions

Consider the case of two random variables X and Y . The *joint probability distribution* defines the probability of two events occurring together.

3.1 Discrete random variables

If X and Y are discrete, the *joint probability mass function* is denoted

$$p_{X,Y}(x,y) = P(X = x, Y = y) \tag{10}$$

As a specific example, let

$$\begin{aligned} X &= \text{outcome of a fair coin toss, 1 if heads, 0 if tails} \\ Y &= \text{outcome of the roll of a fair 6-sided die} \end{aligned} \tag{11}$$

Then

$$p_{X,Y}(\text{tails, roll a 3}) = P(X = \text{tails and } Y = \text{roll a 3}) = \frac{1}{12}$$

We can represent this joint distribution as a 2×6 table:

	Y = 1	Y = 2	Y = 3	Y = 4	Y = 5	Y = 6
X = 0	1/12	1/12	1/12	1/12	1/12	1/12
X = 1	1/12	1/12	1/12	1/12	1/12	1/12

The *marginal probability* is the overall probability of the event $X = x$, regardless of what happens with Y , and is obtained by

$$p(x) = \sum_{y \in \text{Domain}(Y)} p(x,y) \tag{12}$$

where $\text{Domain}(Y)$ is just the set of all possible values of Y .

This process is called *marginalization*. In the table representation, it corresponds to summing out one of the dimensions (e.g., in this case, summing over all the rows or all the columns in the table). Specifically, in example (11),

$$P(X = \text{tails}) = \sum_{i=1}^6 P(X = \text{tails}, Y = \text{roll } i) = \sum_{i=1}^6 \frac{1}{12} = \frac{1}{2}$$

The *conditional* probability mass function defines the probability of an event x occurring given that we've already observed $Y = y$ as

$$p(X = x|Y = y) = p(x|y) = \frac{p(x,y)}{p(y)} \tag{13}$$

assuming $p(y) \neq 0$. In the 2-dimensional table representation, this means we focus our attention on just a single row or column in the table and ignore the rest.

In this example, we might consider

$$P(X = \text{heads}|Y = \text{roll a 4}) = \frac{P(X = \text{heads and } Y = \text{roll a 4})}{P(Y = \text{roll a 4})} = \frac{\frac{1}{12}}{\frac{1}{6}} = \frac{1}{2}$$

3.2 Bayes' rule

In order to derive Bayes' rule, recall the definition of conditional distribution in (13) and consider two statements that follow from it:

$$p(x, y) = p(x|y)p(y) \quad (14)$$

$$p(x, y) = p(y|x)p(x) \quad (15)$$

Applying (13) first followed by (15), derive:

$$p(x|y) = \frac{p(x, y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \quad (16)$$

Another (equivalent) formulation of Bayes' rule, derived by applying (12) followed by (15) is

$$\begin{aligned} p(x|y) &= \frac{p(y|x)p(x)}{\sum_{x \in \text{Domain}(X)} p(x, y)} \\ &= \frac{p(y|x)p(x)}{\sum_{x \in \text{Domain}(X)} p(y|x)p(x)} \end{aligned} \quad (17)$$

3.2.1 Example

A typical example of an application of Bayes' rule is where

- X is a binary variable corresponding to the presence or absence of a certain disease, and
- Y is a random variable corresponding to some trait of a patient, such as blood pressure.

Suppose we have a patient with trait y , and want to compute the probability that this patient has the disease, i.e., $P(X = 1|Y = y)$. One way to do it is to collect a group of people all of which share the trait $Y = y$, and estimate $P(X = 1|Y = y)$ directly by counting the number of people in this group that have the disease. However, as you might imagine, this would be very difficult in practice, especially if Y is multi-dimensional, and, if Y is continuous (e.g, blood pressure), then even impossible.

On the other hand, using Bayes' rule, all we need to know is

- the proportion of people that have the disease, or $P(X = 1)$, which also tells us $P(X = 0)$, and
- the conditional distributions $P(Y|X = 1)$ and $P(Y|X = 0)$ corresponding to the "profiles" of sick and healthy people respectively relative to this trait

The proportion of people that have the same trait as the patient, or $P(Y = y)$ can then be computed using (13, 12)

$$P(Y = y) = P(Y = y|X = 0)P(X = 0) + P(Y = y|X = 1)P(X = 1)$$

Note that in practice finding a group of sick patients and a group of healthy patients to estimate $P(Y|X = 1)$ and $P(Y|X = 0)$, along with estimating $P(X = 1)$ for the general population, is easier than attempting to compute $P(X = 1|Y = y)$ directly.

3.3 Independence

Two variables are considered *independent* if knowing Y doesn't change your belief about X . More formally, X and Y are independent if for all values $x \in \text{Domain}(X)$ and $y \in \text{Domain}(Y)$:

$$p(X = x|Y = y) = p(X = x) \quad (18)$$

or, equivalently¹, using the definition of conditional probability,

$$p(X = x, Y = y) = p(X = x)p(Y = y) \quad (19)$$

The coin toss and the die roll random variables from the previous example are independent random variables. However, if we consider two tosses of the same fair coin and define the random variables as

$$X = \text{the number of heads in the two tosses of this coin}$$

$$Y = \begin{cases} 1 & \text{if any of the two tosses comes up heads} \\ 0 & \text{otherwise} \end{cases}$$

then the two variables are not independent, since

$$P(X = 2|Y = 0) = P(\text{HH outcome}|\text{neither toss comes up heads}) = 0$$

$$P(X = 2) = P(\text{HH outcome}) = \frac{1}{4}$$

Going back to the table representation, if X and Y are independent then we can represent their joint probability distribution by two 1-dimensional tables (one for $P(X)$ and one for $P(Y)$) instead of a full 2-dimensional table:

Y = 1	Y = 2	Y = 3	Y = 4	Y = 5	Y = 6
1/6	1/6	1/6	1/6	1/6	1/6

X = 0	X = 1
1/2	1/2

The advantage is that number of entries in this case would be

$$size(\text{textDomain}(X)) + size(\text{Domain}(Y)) = 8$$

instead of

$$size(\text{Domain}(X)) \times size(\text{Domain}(Y)) = 12$$

3.4 Conditional independence

Two variables X and Y are *conditionally independent* given a third variable Z if

$$p(X = x, Y = y|Z = z) = p(X = x|Z = z)p(Y = y|Z = z) \quad (20)$$

for all $x \in \text{Domain}(X)$, $y \in \text{Domain}(Y)$, $z \in \text{Domain}(Z)$, such that $P(Z = z) \neq 0$.

¹Formulation (19) implicitly deals with the case where $P(Y = y) = 0$. In formulation (18) to be precise we actually have to consider this as a special case, and change the definition from “for all $y \in \text{Domain}(Y)$ ” to “for all $y \in \text{Domain}(Y)$ such that $P(Y = y) \neq 0$ ”.

3.4.1 Conditional independence does not imply independence

Note that it might be the case that X and Y are not independent, but they *are* conditionally independent given Z . For example, suppose X and Y are as above, so

$$X = \text{the number of heads in the two tosses of this coin}$$

$$Y = \begin{cases} 1 & \text{if any of the two tosses comes up heads} \\ 0 & \text{otherwise} \end{cases}$$

Recall that we have already argued that these are not independent. Now we define Z to be the *same* as Y , so we obtain

$$p(X = x, Y = y | Z = z) = \begin{cases} 0 & \text{if } y \neq z \\ P(X = x | Z = z) & \text{otherwise} \end{cases}$$

This is because the variables Y and Z are identical, so it can't be the case that $y \neq z$. If $y \neq z$, since $P(Y = y | Z = z) = 0$ we have

$$p(X = x, Y = y | Z = z) = 0 = P(X = x | Z = z)P(Y = y | Z = z)$$

and conditional independence holds so far. When $y = z$, the fact that $Y = y$ gives us no more information about X than if we just knew that $Z = z$. If $y = z$, then since $P(Y = y | Z = z) = 1$,

$$P(X = x, Y = y | Z = z) = P(X = x | Z = z) = P(X = x | Z = z)P(Y = y | Z = z)$$

and we have satisfied the definition of conditional independence.

This example is somewhat degenerate, but there are plenty of more complicated examples where conditional independence does not imply independence. This is an important result to keep in mind.

3.4.2 Independence does not imply conditional independence

Alternatively, note that we can also have two independent variables that become dependent when conditioned on a third one. As a simple example, consider flipping two fair coins again and let

$$\begin{aligned} X &= \begin{cases} 1 & \text{if coin1 comes up heads} \\ 0 & \text{otherwise} \end{cases} \\ Y &= \begin{cases} 1 & \text{if coin2 comes up heads} \\ 0 & \text{otherwise} \end{cases} \\ Z &= \begin{cases} 1 & \text{if both coins come up heads} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{21}$$

Clearly X and Y are independent since they are dealing with different coins. Now consider what happens when we observe that $Z = 0$ and condition on that. We know that only one of three possible outcomes have occurred: HT, TH or TT.

$$P(X = 0, Y = 0 | Z = 0) = P(\text{coin1} = T \text{ and } \text{coin2} = T | \text{HT, TH or TT}) = \frac{1}{3}$$

$$P(X = 0 | Z = 0) = P(\text{coin1} = T | \text{HT, TH or TT}) = \frac{2}{3}$$

$$P(Y = 0 | Z = 0) = P(\text{coin2} = T | \text{HT, TH or TT}) = \frac{2}{3}$$

So

$$P(X = 0, Y = 0|Z = 0) \neq P(X = 0|Z = 0)P(Y = 0|Z = 0)$$

So two variables can become correlated when conditioned on a third.

3.5 Chain rule

Using the definition of conditional distribution from (13), we can derive the *chain rule*

$$P(X = x, Y = y, Z = z) = P(X = x|Y = y, Z = z)P(Y = y|Z = z)P(Z = z) \quad (22)$$

or, more generally,

$$P(X_1, X_2, \dots, X_n) = P(X_n|X_1 \dots X_{n-1})P(X_{n-1}|X_1 \dots X_{n-2}) \dots P(X_2|X_1)P(X_1) \quad (23)$$

This holds true for all variables $X_1 \dots X_n$.

4 Statistics of random variables

4.1 Expectation

The *expectation*, or *mean* of a discrete random variable is defined as

$$\mathbb{E}[X] = \sum_{x \in \text{Domain}(X)} x P(x) \quad (24)$$

Given an arbitrary function $g : \mathbb{R} \rightarrow \mathbb{R}$, we have

$$\mathbb{E}[g(X)] = \sum_{x \in \text{Domain}(X)} g(x) P(x) \quad (25)$$

4.1.1 Expectations of standard distributions

For the Bernoulli variable defined in (4), we have

$$\mathbb{E}[X] = 1 \times p + 0 \times (1 - p) = p \quad (26)$$

For the Binomial defined in (5), the derivation is trickier, but $\mathbb{E}[X] = np$.

For the Gaussian $\mathcal{N}(\mu, \sigma^2)$ of (7), it's $\mathbb{E}[X] = \mu$.

4.1.2 Properties

Some properties of expectation are

- $\mathbb{E}[c] = c$ for any constant c
- $\mathbb{E}[cg(X)] = c\mathbb{E}[g(X)]$
- *Linearity of expectation* $\mathbb{E}[f(X) + g(X)] = \mathbb{E}[f(X)] + \mathbb{E}[g(X)]$.
This holds whether or not the variables are independent!

- For discrete random variable X , $\mathbb{E}[1\{X = k\}] = P(X = k)$

The linearity of expectation property is very important in probability theory, and will come up a lot in CS221, so let's emphasize it again:

For all random variables X and Y , whether they are independent or not,
 $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

4.1.3 Linearity of expectation examples

Since this is such an important concept, let's consider three linearity of expectation examples.

Example 1

First, consider two independent variables X and Y , defined as in (21), so corresponding to two fair coins. We compute

$$\begin{aligned}\mathbb{E}(X) + \mathbb{E}(Y) &= \frac{1}{2} + \frac{1}{2} = 1 \\ \mathbb{E}(X + Y) &= 2 \times P(\text{HH}) + 1 \times P(\text{HT or TH}) + 0 \times P(\text{TT}) \\ &= 2 \times \frac{1}{4} + 1 \times \frac{1}{2} = 1\end{aligned}$$

Example 2

Now verify that this works with dependent variables. Consider tossing just *one* fair coin, and defining

$$X = Y = \begin{cases} 1 & \text{if coin comes up heads} \\ 0 & \text{otherwise} \end{cases}$$

This time $X + Y$ can take on only two values: it's 2 if the coin comes up heads (since both X and Y will have values of 1), and 0 otherwise. Again we have

$$\begin{aligned}\mathbb{E}(X) + \mathbb{E}(Y) &= \frac{1}{2} + \frac{1}{2} = 1 \\ \mathbb{E}(X + Y) &= 2 \times P(\text{H}) + 0 \times P(\text{T}) = 2 \times \frac{1}{2} = 1\end{aligned}$$

Example 3

Finally, let's consider an even more interesting example, where we roll a single fair six-sided die and let

$$\begin{aligned}X &= \text{the roll of the die} \\ Y &= \begin{cases} 1 & \text{if die roll} \geq 3 \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

These variables consider the same die and are clearly not independent. However, we compute

$$\begin{aligned}\mathbb{E}(X) &= 1 \times P(\text{roll } 1) + 2 \times P(\text{roll } 2) + 3 \times P(\text{roll } 3) + 4 \times P(\text{roll } 4) + \\ &\quad + 5 \times P(\text{roll } 5) + 6 \times P(\text{roll } 6) \\ &= (1 + 2 + 3 + 4 + 5 + 6) \times \frac{1}{6} = 3\frac{1}{2}\end{aligned}$$

$$\mathbb{E}(Y) = 0 \times P(\text{roll } 1, 2, 3) + 1 \times P(\text{roll } 4, 5, 6) = \frac{1}{2}$$

$$\mathbb{E}(X) + \mathbb{E}(Y) = 4$$

To compute $\mathbb{E}(X + Y)$, we consider what happens on each roll of the die:

Roll	1	2	3	4	5	6
Value of $X + Y$	$0 + 1 = 1$	$0 + 2 = 2$	$0 + 3 = 3$	$1 + 4 = 5$	$1 + 5 = 6$	$1 + 6 = 7$

So

$$\begin{aligned}\mathbb{E}(X + Y) &= 7 \times P(\text{roll } 6) + 6 \times P(\text{roll } 5) + 5 \times P(\text{roll } 4) + \\ &\quad + 3 \times P(\text{roll } 3) + 2 \times P(\text{roll } 2) + 1 \times P(\text{roll } 1) \\ &= (7 + 6 + 5 + 3 + 2 + 1) \times \frac{1}{6} = 4\end{aligned}$$

Thus in this case still

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

4.2 Variance

The *variance* of a random variable is the measure of deviation from the mean, defined as

$$\text{Var}[V] = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad (27)$$

Alternatively,

$$\begin{aligned}\text{Var}[V] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - \mathbb{E}[2X\mathbb{E}[X]] + \mathbb{E}[\mathbb{E}[X]^2] \\ &= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2\end{aligned} \quad (28)$$

4.2.1 Variances of standard distributions

Again, for the Bernoulli variable defined in (4), we have

$$\begin{aligned}\mathbb{E}[X^2] &= 1^2 \times p + 0^2 \times (1 - p) = p \\ \mathbb{E}[X]^2 &= p^2 \\ \Rightarrow \text{Var}[X] &= p - p^2 = p(1 - p)\end{aligned} \quad (29)$$

For the Binomial defined in (5), $\text{Var}[X] = np(1 - p)$.

For the Gaussian $\mathcal{N}(\mu, \sigma^2)$ in (7), $\text{Var}[X] = \sigma^2$.

4.2.2 Properties

Some properties of variance are

- $\text{Var}[c] = 0$ for any constant c
- $\text{Var}[cf(X)] = c^2\text{Var}[f(X)]$