Course: M362K Intro to Stochastic Processes
Term: Fall 2014
Instructor: Gordan Zitkovic

Lecture 1
Probability review

RANDOM VARIABLES

A large chunk of probability is about random variables. Instead of giving a precise definition, let us just mention that a **random variable** can be thought of as an uncertain, numerical (i.e., with values in $\mathbb{R}$) quantity. While it is true that we do not know with certainty what value a random variable X will take, we usually know how to assign a number - the probability - that its value will be in some some[1] subset of $\mathbb{R}$. For example, we might be interested in $\mathbb{P}[X \geq 7]$, $\mathbb{P}[X \in [2, 3.1]]$ or $\mathbb{P}[X \in \{1, 2, 3\}]$. The collection of all such probabilities is called the **distribution** of X. One has to be very careful not to confuse the random variable (a function on a probability space) itself and its distribution (a collection of probabilities, i.e., numbers). This point is particularly important when several random variables appear at the same time[2]. When two random variables X and Y have the same distribution, i.e., when $\mathbb{P}[X \in A] = \mathbb{P}[Y \in A]$ for any set A, we say that X and Y are **equally distributed** and write $X \overset{(d)}{=} Y$.

COUNTABLE SETS

Almost all random variables in this course will take only countably many values, so it is probably a good idea to review breifly what the word *countable* means. As you might know, the countable infinity is one of many different infinities we encounter in mathematics. Simply, a set is countable if it has the same number of elements as the set $\mathbb{N} = \{1, 2, \dots\}$ of natural numbers. More precisely, we say that a set A is **countable** if there exists a function $f : \mathbb{N} \to A$ which is bijective (one-to-one and onto). You can think f as the correspondence that "proves" that there exactly as many elements of A as there are elements of $\mathbb{N}$. Alternatively, you can view f as an *ordering* of A; it arranges A into a particular order $A = \{a_1, a_2, \dots\}$, where $a_1 = f(1)$, $a_2 = f(2)$, etc. Infinities are funny, however, as the following example shows

[1] We will not worry about measurability in this class.

[2] in addition to the fact that the knowing the distributions of X and Y *does not* determine the distribution of the pair (X, Y) or, even, $X + Y$. This is not the case with (nonrandom) numbers. If I know x and y, I know $x + y$. We'll come back to this later.

Example 1.1.

1. $\mathbb{N}$ itself is countable; just use $f(n) = n$.
2. $\mathbb{N}_0 = \{0, 1, 2, 3, \dots\}$ is countable; use $f(n) = n - 1$. You can see here why I think that infinities are funny; the set $\mathbb{N}_0$ and the set $\mathbb{N}$ - which is its proper subset - have the same size[3].
3. $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, 3, \dots\}$ is countable; now the function f is a bit more complicated;

$$f(k) = \begin{cases} 2k + 1, & k \geq 0 \\ -2k, & k < 0. \end{cases}$$

   You could think that $\mathbb{Z}$ is more than "twice-as-large" as $\mathbb{N}$, but it is not. It is the same size.
4. It gets even weirder. The set $\mathbb{N} \times \mathbb{N} = \{(m, n) : m \in \mathbb{N}, n \in \mathbb{N}\}$ of all pairs of natural numbers is also countable. I leave it to you to construct the function f - it will have to meander a bit.
5. A similar argument shows that the set $\mathbb{Q}$ of all rational numbers (fractions) is also countable.
6. The set $[0, 1]$ of all real numbers between 0 and 1 is *not* countable; this fact was first proven by Georg Cantor who used a neat trick called the *diagonal argument*.

DISCRETE RANDOM VARIABLES

A random variable is said to be discrete if it takes at most countably many values. More precisely, X is said to be *discrete* if there exists a *finite* or *countable* set $S \subset \mathbb{R}$ such that $\mathbb{P}[X \in S] = 1$, i.e., if we know with certainty that the only values X can take are those in S; more precisely, what we know with certainty is that it will *not* take any values outside of S. The smallest set S with that property is called the **support** of X. If we want to stress that the support $\mathcal{S}$ corresponds to the random variable X, we write $\mathcal{S}_X$.

1. If X takes its values in the set $\mathbb{N} = \{1, 2, 3, \dots\}$, we say that X is $\mathbb{N}$**-valued**[4].
2. If we allow 0 (in addition to $\mathbb{N}$), so that $\mathbb{P}[X \in \mathbb{N}_0] = 1$, we say that X is $\mathbb{N}_0$**-valued**
3. Sometimes, it is convenient to allow discrete random variables to take the value $+\infty$. This is mostly the case when we model the waiting time until the first occurence of an event which may or may not ever happen. If it never happens, we will be waiting forever, and the waiting time will be $+\infty$. In those cases - when

[3] this points to the fact that the notion of size, while quite straighforward on finite sets, loses some of its meaning when applied to infinite sets.

[4] note that some elements of $\mathbb{N}$ may never appear as values of X. In the extreme case, the constant random variable $X = 3$ is also considered $\mathbb{N}$-valued. Its support, however, is $\mathcal{S}_X = \{3\} \subseteq \mathbb{N}$.

$S = \{1, 2, 3, \dots, +\infty\} = \mathbb{N} \cup \{+\infty\}$ - we say that the random variable is **extended** $\mathbb{N}$**-valued**. The same applies to the case of $\mathbb{N}_0$ (instead of $\mathbb{N}$), and we talk about the **extended** $\mathbb{N}_0$**-valued** random variables. Sometimes the adjective "extended" is left out, and we talk about $\mathbb{N}_0$-valued random variables, even though we allow them to take the value $+\infty$. This sounds more confusing that it actually is.

4. Occasionally, we want our random variables to take values which are not necessarily numbers (think about H and T as the possible outcomes of a coin toss, or the suit of a randomly chosen playing card). Is the collection of all possible values (like $\{H, T\}$ or $\{\heartsuit, \spadesuit, \clubsuit, \diamondsuit\}$) is countable, we still call such random variables[5] discrete. We will see more of that when we start talking about Markov chains.

Discrete random variables are very nice due to the following fact: in order to be able to compute any conceivable probability involving a discrete random variable X, it is enough to know how to compute the probabilities $\mathbb{P}[X = x]$, for all $x \in S$. Indeed, if we are interested in figuring out how much $\mathbb{P}[X \in B]$ is, for some set $B \subseteq \mathbb{R}$ (e.g., $B = [3, 6]$, or $B = [-2, \infty)$), we simply pick all $x \in S$ which are also in B and sum their probabilities. In mathematical notation, we have

$$\mathbb{P}[X \in B] = \sum_{x \in S \cap B} \mathbb{P}[X = x].$$

For this reason, the distribution of any discrete random variable X is usually described via a table

$$X \sim \begin{pmatrix} x_1 & x_2 & x_3 & \dots \\ p_1 & p_2 & p_3 & \dots \end{pmatrix},$$

where the top row lists all the elements of S (the support of X) and the bottom row lists their probabilities ($p_i = \mathbb{P}[X = x_i]$, $i \in \mathbb{N}$). When the random variable is $\mathbb{N}$-valued (or $\mathbb{N}_0$-valued), the situation is even simpler because we know what $x_1, x_2, \dots$ are and we identify the distribution of X with the sequence $p_1, p_2, \dots$ (or $p_0, p_1, p_2, \dots$ in the $\mathbb{N}_0$-valued case), which we call the **probability mass function (pmf)** of the random variable X. What about the extended $\mathbb{N}_0$-valued case? It is as simple because we can compute the probability $\mathbb{P}[X = +\infty]$, if we know all the probabilities $p_i = \mathbb{P}[X = i]$, $i \in \mathbb{N}$ ($i \in \mathbb{N}_0$). Indeed, we use the fact that

$$\mathbb{P}[X = 1] + \mathbb{P}[X = 2] + \dots + \mathbb{P}[X = \infty] = 1,$$

so that $\mathbb{P}[X = \infty] = 1 - \sum_{i=1}^{\infty} p_i$, where $p_i = \mathbb{P}[X = i]$. In other words, if you are given a probability mass function $(p_1, \dots)$ (or $(p_0, p_1, \dots)$),

[5] strictly speaking, random variables take values in $\mathbb{R}$ - anything else that is random, but whose value is not a number, should be called a **random element**. This will not be very important for us, so we will mostly disregard it in these notes.

you simply need to compute the sum $\sum_{i=1}^{\infty} p_i$. If it happens to be equal to 1, you can safely conclude that X never takes the value $+\infty$. Otherwise, the probability of $+\infty$ is positive.

The random variables X for which $\mathcal{S}_X \subseteq \{0,1\}$ are especially useful. They are called **indicators**. The name comes from the fact that you should think of such variables as signal lights; if $X = 1$ an event of interest has happened, and if $X = 0$ it has not happened. In other words, X *indicates* the occurence of an event. The notation we use is quite suggestive; for example, if Y is the outcome of a coin-toss, and we want to know whether *Heads* (H) occurred, we write

$$X = \mathbf{1}_{\{Y=H\}}.$$

Example 1.2. Suppose that two dice are thrown so that Y_1 and Y_2 are the numbers obtained (both Y_1 and Y_2 are discrete random variables with $S = \{1,2,3,4,5,6\}$). If we are interested in the probability that their sum is at least 9, we proceed as follows. We define the random variable Z - the sum of Y_1 and Y_2 - by $Z = Y_1 + Y_2$. Another random variable, let us call it X, is defined by $X = \mathbf{1}_{\{Z \geq 9\}}$, i.e.,

$$X = \begin{cases} 1, & Z \geq 9, \\ 0, & Z < 9. \end{cases}$$

With such a set-up, X signals whether the event of interest has happened, and we can state our original problem in terms of X : "Compute $\mathbb{P}[X = 1]$!". Can you compute it?

This example is, admittedly, a little contrived. The point, however, is that anything can be phrased in terms of random variables; thus, if you know how to work with random variables, i.e., know how to compute their distributions, you can solve any problem in probability that comes your way.

EXPECTATION

For a discrete random variable X with support $\mathcal{S} = \mathcal{S}_X$, we define the **expectation** $\mathbb{E}[X]$ of X by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{S}} x \mathbb{P}[X = x],$$

as long as the (possibly) infinite sum $\sum_{x \in \mathcal{S}} x \mathbb{P}[X = x]$ *absolutely converges*, i.e., as long as

$$\sum_{x \in \mathcal{S}} |x| \, \mathbb{P}[X = x] < \infty. \tag{1.1}$$

When the sum in (1.1) above diverges (i.e., takes the value $+\infty$), we say that the expectation of X **is not defined**[6].

[6] Note that, for $\mathbb{E}[X]$ to be defined, we require that $\mathbb{E}[|X|]$ be finite.

When the random variable in question is $\mathbb{N}$-valued, the expression above simplifies to

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i \times p_i,$$

where $p_i = \mathbb{P}[X = i]$, for $i \in \mathbb{N}$. Unlike in the general case, the absolute convergence of the defining series can fail in essentially one way, i.e., when

$$\lim_{n\to\infty} \sum_{i=1}^{n} i p_i = +\infty.$$

In that case, the expectation does not formally exist. We still write $\mathbb{E}[X] = +\infty$, but really mean that the defining sum diverges towards infinity.

Once we know what the expectation is, we can easily define several more common terms:

Definition 1.3. Let X be a discrete random variable.

- If the expectation $\mathbb{E}[X]$ exists, we say that X is **integrable**.
- If $\mathbb{E}[X^2] < \infty$ (i.e., if X^2 is integrable), X is called **square-integrable**.
- If $\mathbb{E}[|X|^m] < \infty$, for some $m > 0$, we say that X **has a finite** m**-th moment**.
- If X has a finite m-th moment, the expectation $\mathbb{E}[|X - \mathbb{E}[X]|^m]$ exists and we call it the m**-th central moment**.

It can be shown that the expectation $\mathbb{E}$ possesses the following properties, where X and Y are both assumed to be integrable:

1. $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$, for $\alpha, \beta \in \mathbb{R}$ (linearity of expectation).
2. $\mathbb{E}[X] \geq \mathbb{E}[Y]$ if $\mathbb{P}[X \geq Y] = 1$ (monotonicity of expectation).

Definition 1.4. Let X be a square-integrable random variable. We define the **variance** $\mathrm{Var}[X]$ by

$$\mathrm{Var}[X] = \mathbb{E}[(X-m)^2], \text{ where } m = \mathbb{E}[X].$$

The square-root $\sqrt{\mathrm{Var}[X]}$ is called the **standard deviation** of X.

Remark 1.5. Each square-integrable random variable is automatically integrable [7]. Also, if the m-th moment exists, then all lower moments also exist.

[7] this is due to the fact that

$$|x| \leq \tfrac{1}{2}(x^2+1), \text{ for all } x \in \mathbb{R}.$$

We still need to define what happens with random variables that take the value $+\infty$, but that is very easy. We stipulate that $\mathbb{E}[X]$ *does not exist*, (i.e., $\mathbb{E}[X] = +\infty$) as long as $\mathbb{P}[X = +\infty] > 0$. Simply put, the expectation of a random variable is infinite if there is a positive chance (no matter how small) that it will take the value $+\infty$.

EVENTS AND PROBABILITY

Probability is usually first explained in terms of the **sample space** or **probability space** (which we denote by Ω in these notes) and various *subsets* of Ω which are called events[8]. Events typically contain all **elementary events**, i.e., elements of the probability space, usually denoted by ω. For example, if we are interested in the likelihood of getting an odd number as a sum of outcomes of two dice throws, we build a probability space

$$\Omega = \{(1,1), (1,2), \dots, (6,1), (2,1), (2,2), \dots, (2,6), \dots, (6,1), (6,2), \dots, (6,6)\}$$

and define the event A which contains of all pairs $(k,l) \in \Omega$ such that $k + l$ is an odd number, i.e.,

$$A = \{(1,2), (1,4), (1,6), (2,1), (2,3), \dots, (6,1), (6,3), (6,5)\}.$$

As we already mentioned above, events can be thought of as very simple random variables. Indeed, if, for an event A, we define the random variable $\mathbf{1}_A$ by

$$\mathbf{1}_A = \begin{cases} 1, & A \text{ happened,} \\ 0, & A \text{ did not happen,} \end{cases}$$

we get the *indicator random variable* mentioned above. Conversely, for any indicator random variable X, we define the **indicated event** A as the set of all elementary events at which X takes the value 1:

$$A = \{\omega \in \Omega \,:\, X(\omega) = 1\}.$$

What does all this have to do with probability? The analogy goes one step further. If we apply the notion of expectation to the indicator random variable $X = \mathbf{1}_A$, we get the probability of A:

$$\mathbb{E}[\mathbf{1}_A] = \mathbb{P}[A].$$

Indeed, $\mathbf{1}_A$ takes the value 1 on A, and the value 0 on the complement $A^c = \Omega \setminus A$. Therefore, $\mathbb{E}[\mathbf{1}_A] = 1 \times \mathbb{P}[A] + 0 \times \mathbb{P}[A^c] = \mathbb{P}[A]$.

DEPENDENCE AND INDEPENDENCE

As we already mentioned, one of the main differences between random variables and (deterministic or non-random) quantities is that in

[8] Many times, when Ω is uncountable, not all of its subsets can be considered events, due to very strange technical reasons. We will disregard that fact for the rest of the course.

the former case the whole is more than the sum of its parts. What do I mean by that? When two random variables, say X and Y, are considered in the same setting, you must specify more than just their distributions, if you want to compute probabilities that involve both of them. Here are two examples.

1. We throw two dice, and denote the outcome on the first one by X and the second one by Y.
2. We throw two dice, and denote the outcome of the first one by X, but set $Y = 6 - X$ and forget about the second die.

In both cases, both X and Y have the same distribution

$$X, Y \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$

The pairs (X, Y) are, however, very different in the two examples. In the first one, if the value of X is revealed, it will not affect our view of the value of Y. Indeed, the dice are not "connected" in any way (they are independent in the language of probability). In the second case, the knowledge of X allows us to say what Y is without any doubt - it is $6 - X$.

This example shows that when more than one random variable is considered, one needs to obtain additional information about their relationship - not everything can be deduced only by looking at their distributions (pmfs, or ...).

One of the most common forms of relationship two random variables can have is the one of example (1) above, i.e., no relationship at all. More formally, we say that two (discrete) random variables X and Y are **independent** if

$$\mathbb{P}[X = x \text{ and } Y = y] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y],$$

for *all* x and y in the respective supports $\mathcal{S}_X$ and $\mathcal{S}_Y$ of X and Y. The same concept can be applied to events, and we say that two events A and B are independent if

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\,\mathbb{P}[B].$$

The notion of independence is central to probability theory (and this course) because it is relatively easy to spot in real life. If there is no physical mechanism that ties two events (like the two dice we throw), we are inclined to declare them independent. One of the most important tasks in probabilistic modelling is the identification of the (small number of) independent random variables which serve as building blocks for a big complex system. You will see many examples of that as we proceed through the course.

CONDITIONAL PROBABILITY

When two random variables are not independent, we still want to know how the knowledge of the exact value of one of the affects our guesses about the value of the other. That is what the conditional probability is for. We start with the definition, and we state it for events first: for two events A, B such that $\mathbb{P}[B] > 0$, the **conditional probability $\mathbb{P}[A|B]$ of A given B** is defined as:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

The conditional probability is *not defined* when $\mathbb{P}[B] = 0$ (otherwise, we would be computing $\frac{0}{0}$ - why?). Every statement in the sequel which involves conditional probability will be assumed to hold only when $\mathbb{P}[B] = 0$, without explicit mention.

The conditional probability calculations often use one of the following two formulas. Both of them use the familiar concept of partition. If you forgot what it is, here is a definition: a collection $A_1, A_2, \dots, A_n$ of events is called a **partition of Ω** if

a) $A_1 \cup A_2 \cup \dots A_n = \Omega$ and

b) $A_i \cap A_j = \emptyset$ for all pairs $i, j = 1, \dots, n$ with $i \neq j$.

So, let $A_1, \dots, A_n$ be a partition of Ω, and let B be an event.

1. **The Law of Total Probability.**

$$\mathbb{P}[B] = \sum_{i=1}^{n} \mathbb{P}[B|A_i]\mathbb{P}[A_i].$$

2. **Bayes Formula.** For $k = 1, \dots, n$, we have

$$\mathbb{P}[A_k|B] = \frac{\mathbb{P}[B|A_k]\mathbb{P}[A_k]}{\sum_{i=1}^{n} \mathbb{P}[B|A_i]\mathbb{P}[A_i]}.$$

Even though the formulas above are stated for finite partitions, they remain true when the number of the elements of the partition is countably infinite[9].

Random variables can be substituted for events in the definition of conditional probability as follows: for two random variables X and Y, the **conditional probabilty** that $X = x$, **given** $Y = y$ (with x and y in respective supports $\mathcal{S}_X$ and $\mathcal{S}_Y$) is given by

$$\mathbb{P}[X = x|Y = y] = \frac{\mathbb{P}[X = x \text{ and } Y = y]}{\mathbb{P}[Y = y]}.$$

The formula above produces a different probability distribution for each y. This is called the **conditional distribution** of X, **given** $Y = y$.

[9] naturally, the finite sums have to be replaced by infinite series.

We give a simple example to illustrate this concept. Let X be the number of *heads* obtained when two coins are thrown, and let Y be the indicator of the event that the second coin shows *heads*. The distribution of X is Binomial:

$$X \sim \begin{pmatrix} 0 & 1 & 2 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \end{pmatrix},$$

or, in the more compact notation which we use when the support is clear from the context $X \sim (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$. The random variable Y has the Bernoulli distribution $Y \sim (\frac{1}{2}, \frac{1}{2})$. What happens to the distribution of X, when we are told that $Y = 0$, i.e., that the second coin shows *tails*? In that case we have

$$\mathbb{P}[X = x | Y = 0] = \begin{cases} \frac{\mathbb{P}[X=0,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ the pattern is TT }]}{\mathbb{P}[Y=0]} = \frac{1/4}{1/2} = \boxed{\frac{1}{2}}, & x = 0 \\ \frac{\mathbb{P}[X=1,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ the pattern is HT }]}{\mathbb{P}[Y=0]} = \frac{1/4}{1/2} = \boxed{\frac{1}{2}}, & x = 1 \\ \frac{\mathbb{P}[X=2,Y=0]}{\mathbb{P}[Y=0]} = \frac{\mathbb{P}[\text{ well, there is no such pattern }]}{\mathbb{P}[Y=0]} = \frac{0}{1/2} = \boxed{0}, & x = 2 \end{cases}$$

Thus, the conditional distribution of X, given $Y = 0$, is $(\frac{1}{2}, \frac{1}{2}, 0)$. A similar calculation can be used to get the conditional distribution of X, but now given that $Y = 1$, is $(0, \frac{1}{2}, \frac{1}{2})$. The moral of the story is that the additional information contained in Y can alter our views about the unknown value of X using the concept of conditional probability.

One final remark about the relationship between independence and conditional probability: suppose that the random variables X and Y are independent. Then the knowledge of Y should not affect how we think about X; indeed, then

$$\mathbb{P}[X = x | Y = y] = \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} = \frac{\mathbb{P}[X = x]\mathbb{P}[Y = y]}{\mathbb{P}[Y = y]} = \mathbb{P}[X = x].$$

The conditional distribution does not depend on y, and coincides with the unconditional one.

The notion of independence for two random variables can easily be generalized to larger collections

Definition 1.6. Random variables $X_1, X_2, \dots, X_n$ are said to be **independent** if

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \dots X_n = x_n] = \mathbb{P}[X_1 = x_1]\, \mathbb{P}[X_2 = x_2] \dots \mathbb{P}[X_n = x_n] \text{ for } \textit{all } x_1, x_2, \dots, x_n.$$

An infinite collection of random variables is said to be **independent** if all of its finite subcollections are independent.

Independence is often used in the following way:

Proposition 1.7. *Let $X_1, \dots, X_n$ be independent random variables. Then*

1. *$g_1(X_1)$, …, $g_n(X_n)$ are also independent for (practically) all functions $g_1, \dots, g_n$,*
2. *if X_1, …, X_n are integrable then the product $X_1 \dots X_n$ is integrable and*

   $$\mathbb{E}[X_1 \dots X_n] = \mathbb{E}[X_1] \dots \mathbb{E}[X_n], \text{ and}$$

3. *if X_1, …, X_n are square-integrable, then*

   $$\text{Var}[X_1 + \dots + X_n] = \text{Var}[X_1] + \dots + \text{Var}[X_n].$$

   Equivalently

   $$\text{Cov}[X_i, X_j] = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])] = 0,$$

   for all $i \neq j \in \{1, 2, \dots, n\}$.

Remark 1.8. The last statement says that independent random variables are uncorrelated. The converse is not true. There are uncorrelated random variables which are not independent.

When several random variables $(X_1, X_2, \dots X_n)$ are considered in the same setting, we often group them together into a **random vector**. The **distribution** of the random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ is the collection of all probabilities of the form

$$\mathbb{P}[X_1 = x_1, X_2 = x_2, \dots, X_n = x_n],$$

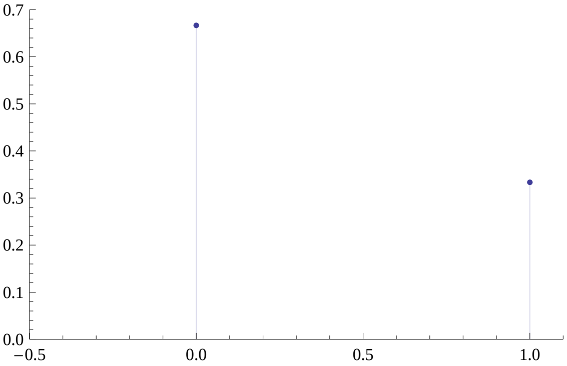
when $x_1, x_2, \dots, x_n$ range through all numbers in the appropriate supports. Unlike in the case of a single random variable, writing down the distributions of random vectors in tables is a bit more difficult. In the two-dimensional case, one would need an entire matrix, and in the higher dimensions some sort of a hologram would be the only hope.

The distributions of the components $X_1, \dots, X_n$ of the random vector $\boldsymbol{X}$ are called the **marginal distributions** of the random variables $X_1, \dots, X_n$. When we want to stress the fact that the random variables $X_1, \dots, X_n$ are a part of the same random vector, we call the distribution of $\boldsymbol{X}$ the **joint distribution** of $X_1, \dots, X_n$. It is important to note that, unless random variables $X_1, \dots, X_n$ are a priori known to be independent, the joint distribution holds more information about $\boldsymbol{X}$ than all marginal distributions together.

EXAMPLES

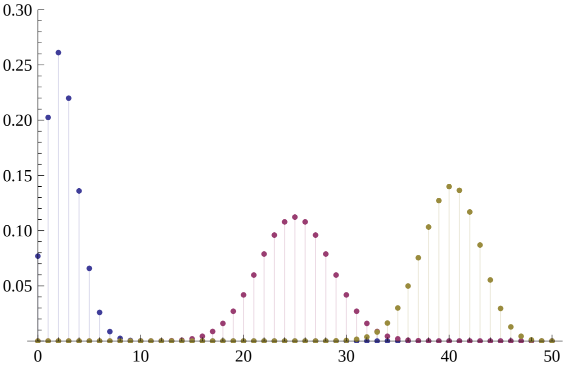
Here is a short list of some of the most important discrete random variables. You will learn about generating functions soon.

Bernoulli. *Success (1) of failure (0) with probability p*



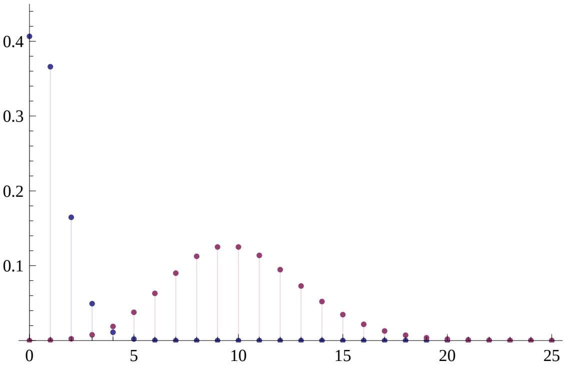
.parameters : $p \in (0,1)$ $(q = 1 - p)$
.notation : $b(p)$
.support : $\{0,1\}$
.pmf : $p_0 = q = 1 - p$ and $p_1 = p$
.generating function : $ps + q$
.mean : p
.standard deviation : $\sqrt{pq}$
.note : the variant where success is encoded by 1, failure by -1 and $p = \frac{1}{2}$ is called the **coin toss**
.figure : the mass function a Bernoulli distribution with $p = 1/3$.

Binomial. *The number of successes in n repetitions of a Bernoulli trial with success probability p.*



.parameters : $n \in \mathbb{N}$, $p \in (0,1)$ $(q = 1 - p)$
.notation : $b(n,p)$
.support : $\{0,1,\dots,n\}$
.pmf : $p_k = \binom{n}{k} p^k q^{n-k}$, $k = 0,\dots,n$
.generating function : $(ps + q)^n$
.mean : np
.standard deviation : $\sqrt{npq}$
.figure : mass functions of three binomial distributions with $n = 50$ and $p = 0.05$ (blue), $p = 0.5$ (purple) and $p = 0.8$ (yellow).

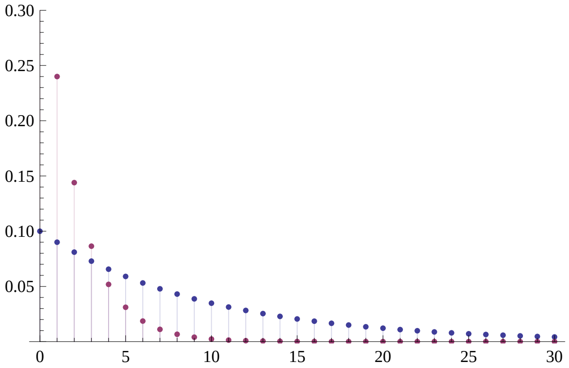
Poisson. *The number of spelling mistakes one makes while typing a single page.*



.parameters : $\lambda > 0$
.notation : $p(\lambda)$
.support : $\mathbb{N}_0$
.pmf : $p_k = e^{-\lambda} \frac{\lambda^k}{k!}$, $k \in \mathbb{N}_0$
.generating function : $e^{\lambda(s-1)}$
.mean : λ
.standard deviation : $\sqrt{\lambda}$
.figure : mass functions of two Poisson distributions with parameters $\lambda = 0.9$ (blue) and $\lambda = 10$ (purple).

Geometric. *The number of repetitions of a Bernoulli trial with parameter p until the first success.*

.parameters : $p \in (0,1)$, $q = 1 - p$
.notation : $g(p)$
.support : $\mathbb{N}_0$
.pmf : $p_k = pq^k$, $k \in \mathbb{N}_0$
.generating function : $\frac{p}{1-qs}$
.mean : $\frac{q}{p}$
.standard deviation : $\frac{\sqrt{q}}{p}$
.figure : mass functions of two Geometric distributions with parameters $p = 0.1$ (blue) and $p = 0.4$ (purple).



Note: In some textbooks, the support of the geometric random variable is taken to be $\mathbb{N}$ and, in others, it is $\mathbb{N}_0$. This depends on whether you are counting the final success as one of the trials or not. In these notes we use the $\mathbb{N}_0$-version. If confusion can arise, we write $\mathbb{N}$-geometric and $\mathbb{N}_0$-geometric for the two version.

PROBLEMS

Problem 1.1. Two people are picked at random from a group of 50 and given \$10 each. After that, independently of what happened before, three people are picked from the same group - one or more people could have been picked both times - and given \$10 each. What is the probability that at least one person received \$20?

Solution: Intuitively, no person gets \$20 if there is no overlap between the group of two people who got picked first and the group of people picked after that. In other words, the three people have to end up being picked from the 48 people who *did not* get picked first. There are $\binom{50}{3}$ ways to pick 3 people from 50 and $\binom{48}{3}$ ways to pick 3 people from the 48 (that did not get picked first). Therefore, the required probability is

$$1 - \frac{\binom{48}{3}}{\binom{50}{3}} = \frac{144}{1225}$$

A more (mathematically) rigorous way of approaching the problem (and obtaining the same result) is the following. Define

$A = \{$no person picked the first time was also picked the second time$\}$,

so that the probability that at least one person received \$20 is given by $\mathbb{P}[A^c] = 1 - \mathbb{P}[A]$. In order to compute $\mathbb{P}[A]$, we note that we can write

$$A = \bigcup_{1 \le i < j \le 50} A_{ij} \cap B_{ij},$$

where

$$A_{ij} = \{\text{the first two people picked are } i \text{ and } j \text{ (not necessarily in that order)}\},$$

and

$$B_{ij} = \{i \text{ and } j \text{ are not among the next three people picked}\}.$$

The sets $A_{ij} \cap B_{ij}$ and $A_{i'j'} \cap B_{i'j'}$ are mutually exclusive whenever $i \neq i'$ or $j \neq j'$, so we have

$$\mathbb{P}[A] = \sum_{1 \leq i < j \leq 50} \mathbb{P}[A_{ij} \cap B_{ij}].$$

Furthermore, A_{ij} and B_{ij} are independent by the assumption so $\mathbb{P}[A_{ij} \cap B_{ij}] = \mathbb{P}[A_{ij}]\mathbb{P}[B_{ij}]$.

Clearly, $\mathbb{P}[A_{ij}] = \frac{1}{\binom{50}{2}}$, since there are $\binom{50}{2}$ equally likely ways to choose 2 people out of 50, and only one of these corresponds to the choice (i, j). Similarly, $\mathbb{P}[B_{ij}] = \frac{\binom{48}{3}}{\binom{50}{3}}$, because there are $\binom{50}{3}$ ways to choose 3 people out of 50, and $\binom{48}{3}$ of those do not involve i or j. Therefore,

$$\mathbb{P}[A] = \sum_{1 \leq i < j \leq 50} \frac{1}{\binom{50}{2}} \frac{\binom{48}{3}}{\binom{50}{3}}.$$

The terms inside the sum are all equal and there are $\binom{50}{2}$ of them, so

$$\mathbb{P}[A] = \binom{50}{2} \frac{1}{\binom{50}{2}} \frac{\binom{48}{3}}{\binom{50}{3}} = \frac{\binom{48}{3}}{\binom{50}{3}},$$

and the required probability is $1 - \frac{\binom{48}{3}}{\binom{50}{3}} = \frac{144}{1225}$.

Problem 1.2. The latest census has revealed the following:

- 40% of the population exercise regularly,
- 30% own a dog,
- 20% like cauliflower,
- 60% of all dog owners exercise regularly
- 10% own a dog and like cauliflower.
- 4% exercise regularly, own a dog and like cauliflower.

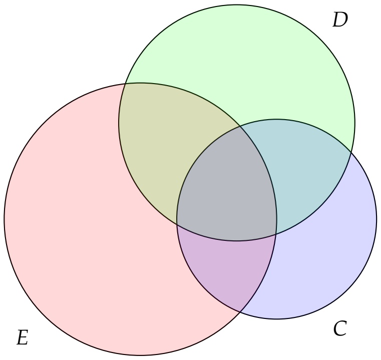
1. Draw a Venn diagram and represent all the assumptions above using probabilities and the set notation.
2. A person is selected at random. Compute the probability that he/she is a dog owner who does not exercise regularly.

3. If it is known that the randomly selected person either likes cauliflower or owns a dog (or both), what is the probability that he/she exercises regularly, owns a dog and likes cauliflower?

Solution:

1. Let E denote the event that the person chosen exercises regularly, D that he/she owns a dog and C that he/she likes cauliflower. The problem states that

$$\mathbb{P}[E|D] = 0.6, \quad \mathbb{P}[C \cap D] = 0.1, \quad \mathbb{P}[C \cap D \cap E] = 0.04,$$
$$\mathbb{P}[E] = 0.4, \quad \mathbb{P}[D] = 0.3, \quad \mathbb{P}[C] = 0.2.$$



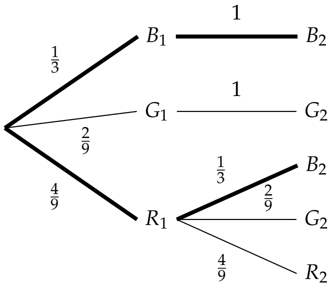
2. It follows that $\mathbb{P}[E \cap D] = \mathbb{P}[E|D] \times \mathbb{P}[D] = 0.6 \times 0.3 = 0.18$, so $\mathbb{P}[D \cap E^c] = \mathbb{P}[D] - \mathbb{P}[E \cap D] = 0.12$.

3. We are looking for $\mathbb{P}[E \cap C \cap D | C \cup D] = \mathbb{P}[E \cap C \cap D] / \mathbb{P}[C \cup D]$. We know that $\mathbb{P}[C \cup D] = \mathbb{P}[C] + \mathbb{P}[D] - \mathbb{P}[C \cap D] = 0.2 + 0.3 - 0.1 = 0.4$, so the required probability is $0.04/0.4 = 0.1$.

Problem 1.3. Two boxes are given. The first one contains 3 blue balls, 4 red balls and 2 green balls. The second one is empty. We start by picking a ball from the first box (all balls are equally likely to be picked):

- If its color is red, we return it to the first box, pick again, and place the picked ball into the second box.
- If its color is blue or green, we place it directly into the second box.

Given that the ball in the second box is blue, what is the probability that the first ball we picked from the first box was red?



Solution: Let R_1, B_1 and G_1 denote the events where the first ball we draw is red, blue and green (respectively). Similarly, let R_2, B_2, G_2 denote the evens where the ball placed in the second box is red, blue and green (respectively).

We are looking for $\mathbb{P}[R_1|B_2]$. By the Bayes formula, we have

$$\begin{aligned}\mathbb{P}[R_1|B_2] &= \frac{\mathbb{P}[B_2|R_1] \times \mathbb{P}[R_1]}{\mathbb{P}[B_2|R_1] \times \mathbb{P}[R_1] + \mathbb{P}[B_2|B_1] \times \mathbb{P}[B_1] + \mathbb{P}[B_2|G_1] \times \mathbb{P}[G_1]} \\ &= \frac{\frac{1}{3} \times \frac{4}{9}}{\frac{1}{3} \times \frac{4}{9} + 1 \times \frac{1}{3} + 0 \times \frac{2}{9}} = \frac{4}{13}.\end{aligned}$$

Problem 1.4. Two coins are tossed and a (6-sided) die is rolled. Describe a sample space (probability space), together with the probability, on which such a situation can be modelled. Find the probability mass function of the random variable whose value is the sum of the number on the die and the total number of heads.

Solution: Each elementary event ω should track the information about three things - the outcome of the first coin toss, the outcome of the second coin toss and the number on the die. This corresponds to triplets $\omega = (c_1, c_2, d)$, where $c_1, c_2 \in \{H, T\}$ and $d \in \{1, \dots, 10\}$. Therefore, $\Omega = \{H, T\} \times \{H, T\} \times \{1, \dots, 6\}$. Since all the instruments involved are fair, the independence requirements dictate that

$$\mathbb{P}[\omega = (c_1, c_2, d)] = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{24},$$

for any $(c_1, c_2, d) \in \Omega$. In words, all elementary events are equally likely. Let C_1 be the random variable which equals to 1 is the outcome of the first coin toss if H, so that

$$C_1(\omega) = \begin{cases} 1, & c_1 = H, \\ 0, & \text{otherwise,} \end{cases} \quad \text{where } \omega = (c_1, c_2, d).$$

In other words, $C_1 = \mathbf{1}_A$ is the indicator of the event

$$A = \{\omega = (c_1, c_2, d) \in \Omega \,:\, c_1 = H\}.$$

Let C_2 and D (the number on the die) be defined analogously. Then the total number of heads M is given by $M = C_1 + C_2$. Each C_1 and C_2 are independent Bernoulli random variables with $p = \frac{1}{2}$, so M is a binomial random variable with $n = 2$ and $p = \frac{1}{2}$. Therefore, the pmf of M is

	0	1	2
p	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Let X be the random variable from the text of the problem:

$$X = D + M.$$

The values random variable X can take are $\{1, 2, \dots, 8\}$, and they correspond to the following table (the table entry is the value of X, columns go with D and rows with M):

	1	2	3	4	5	6
0	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8

A bit of accounting gives the following pmf for X:

	1	2	3	4	5	6	7	8
p	$\frac{1}{24}$	$\frac{3}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{4}{24}$	$\frac{3}{24}$	$\frac{1}{24}$

Problem 1.5. A (6-sided) die is thrown and the number shown is written down (call it X). After that, a biased coin with the probability of *heads* equal to $1/(X+1)$ is tossed until the first *heads* appears.

1. Compute the probability mass function for, as well as the expected of, the number of tosses.

2. Suppose that the number of tosses it took to get *heads* was observed, and it turned out to be equal to 5. The number on the die, on the other hand, is not known. What is the most likely number on the die? *Note:* You may want to use a calculator or a computer here.

Solution: Let Y be the number of tosses until the first head. Conditionally on X, Y is geometrically distributed with success probability $p(X) = \frac{1}{1+X}$.

1. The set of the values Y can take is $\mathbb{N}_0 = \{0, 1, \dots\}$, so the pmf is given by the sequence $p_k = \mathbb{P}[Y = k]$, $k \in \mathbb{N}_0$. By the formula of total probability

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|B_i]\mathbb{P}[B_i],$$

where $\{B_i\}_{i=1,\dots,n}$ is a partition of Ω, we have for $k \in \mathbb{N}_0$,

$$p_k = \sum_{i=1}^{6} \mathbb{P}[Y = k|X = i]\mathbb{P}[X = i] = \sum_{i=1}^{6} (1 - p(i))^k p(i) \frac{1}{6} = \frac{1}{6} \sum_{i=1}^{6} \frac{i^k}{(1+i)^{k+1}}, \; k \in \mathbb{N}_0.$$

The expectation is given by $\mathbb{E}[Y] = \sum_{k=0}^{\infty} k p_k$, so

$$\mathbb{E}[Y] = \sum_{k=0}^{\infty} k \frac{1}{6} \sum_{i=1}^{6} \frac{i^k}{(1+i)^{1+k}}.$$

This can be evaluated by using a bit of calculus, but we can compute this expectation by simple conditioning (and the corresponding law of total probability):

$$\mathbb{E}[Y] = \sum_{i=1}^{6} \mathbb{E}[Y|X = i]\mathbb{P}[X = i]].$$

Since, conditionally on X, Y is a geometric random variable with parameter $p = 1/(1+X)$, we have

$$\mathbb{E}[Y|X = i] = \tfrac{1 - 1/(1+i)}{1/(1+i)} = i,$$

and, so,

$$\mathbb{E}[Y] = \tfrac{1}{6}(1+2+3+4+5+6) = \tfrac{7}{2}.$$

2. For this part, we need to compute the conditional probabilities $\mathbb{P}[X = i|Y = 5]$, for $i = 1, \dots, 6$. We use the Bayes formula:

   Let $\{B_i\}_{i \in 1,\dots,n}$ be a partition of Ω. Then

$$\mathbb{P}[B_i|A] = \frac{\mathbb{P}[B_j \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[B_j]\mathbb{P}[A|B_j]}{\mathbb{P}[A]} = \frac{\mathbb{P}[B_j]\mathbb{P}[A|B_j]}{\sum_{k=1}^n \mathbb{P}[A|B_k]\mathbb{P}[B_k]}.$$

   In our case $B_i = \{X = i\}$ and $A = \{Y = 5\}$, so

$$r_i = \mathbb{P}[X = i|Y = 5] = \mathbb{P}[Y = 5|X = i]\frac{\mathbb{P}[X = i]}{\mathbb{P}[Y = 5]} = \frac{i^5}{(1+i)^6}\frac{1/6}{p_5}.$$

   In order to find the most likely i, we need to find the one which maximizes r_i. Since r_i and $\frac{i^5}{(1+i)^6}$ differ only by a multiplicative constant, it suffices to maximize the expression

$$f(i) = \frac{i^5}{(1+i)^6}, \text{ over } i = \{1, 2, \dots, 6\}.$$

   You can use Mathematica, for example, to get the following approximate numerical values:

i	1	2	3	4	5	6
$f(i)$	0.0156	0.0439	0.0593	0.0655	0.0670	0.0661

   so the most likely die outcome is $i = 5$.

Problem 1.6. A biased coin (probability of *heads* is 0.7) is tossed 1000 times. Write down the exact expression for the probability that more than 750 *heads* have been observed. Use the normal approximation to estimate this probability. *Note:* Before you do that, try to guess, just to test your intuition.

Solution: The random variable X which equals to the number of heads is binomial with probability $p = 0.7$ and $n = 1000$. We are interested in the probability $\mathbb{P}[X > 750]$. If we split this probability between the elementary outcomes which are > 750, we get

$$\mathbb{P}[X > 750] = \sum_{i=751}^{1000} \mathbb{P}[X = i] = \sum_{i=751}^{1000} \binom{1000}{i} (0.7)^i (0.3)^{1000-i}.$$

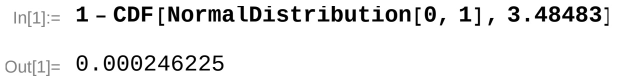
According to the Central Limit Theorem, the random variable

$$X' = \frac{X - \mathbb{E}[X]}{\sqrt{\mathrm{Var}[X]}} = \frac{X - 700}{\sqrt{1000 \cdot 0.7 \cdot 0.3}},$$

is approximately normally distributed with mean 0 and standard variation 1. Therefore (note the continuity correction),

$$\mathbb{P}[X > 750] = \mathbb{P}[X' \geq \frac{750.5 - 700}{\sqrt{210}}] \approx \mathbb{P}[Z \geq 3.48483],$$

where $Z \sim N(0,1)$ is normally distributed with mean 0 and standard variation 1. Table look-up or your computer will give you that this probability is approximately equal to 0.0002. Here is how to compute it in Mathematica:



```
In[1]:= 1 - CDF[NormalDistribution[0, 1], 3.48483]

Out[1]= 0.000246225
```

Problem 1.7. Let X be a Poisson random variable with parameter $\lambda > 0$. Compute the following:

1. $\mathbb{P}[X \geq 3]$,
2. $\mathbb{E}[X^3]$. *Note:* optional and just to challenge you a bit.

Solution:

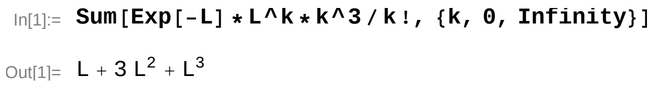
1. Using the explicit expression for the pmf of a Poisson distributions, we get

$$\mathbb{P}[X \geq 3] = 1 - \mathbb{P}[X = 0] - \mathbb{P}[X = 1] - \mathbb{P}[X = 2] = 1 - e^{-\lambda}(1 + \lambda + \tfrac{1}{2}\lambda^2).$$

2. We have

$$\mathbb{E}[X^3] = \sum_{k=0}^{\infty} k^3 \mathbb{P}[X = k] = e^{-\lambda} \sum_{k=0}^{\infty} \frac{k^3}{k!} \lambda^k.$$

We will learn how to evaluate these sums later. At this point, let me show you how to use Mathematica to do that. Simply input



```
In[1]:= Sum[Exp[-L] * L^k * k^3 / k !, {k, 0, Infinity}]

Out[1]= L + 3 L^2 + L^3
```

(don't forget to press SHIFT+ENTER), and, as you can see, the answer is $\lambda + 3\lambda^2 + \lambda^3$.

Problem 1.8. Three (independent) coins are tossed. Two of them are nickels, and the third is a quarter. Let X denote the total number of heads (H) obtained, and let Y denote the number of heads on the two nickels. (Note: For 2., 3. and 4. write down the distribution tables.)

1. (5pts) Write down the joint-distribution table of (X, Y).
2. (5pts) What is the conditional distribution of X given that $Y = 1$?

3. (5pts) What is the distribution of $Z = X - Y$?

4. (5pts) What is the distribution of $g(Z)$, where $g(x) = x^2$ and $Z = X - Y$.

Solution:

1. Here is the distribution table for (X, Y). Marginal distributions for X and Y are in the top row and the right-most column.

	1/8	3/8	3/8	1/8	
2	0	0	1/8	1/8	1/4
1	0	1/4	1/4	0	1/2
0	1/8	1/8	0	0	1/4
Y					
X	**0**	**1**	**2**	**3**	

2. (This is the second row of the joint distribution table, normalized by the marginal probability $\mathbb{P}[Y = 1]$.)

k	1	2
$\mathbb{P}[X = k \mid Y = 1]$	1/2	1/2

.

3. Z is just the outcome of the quarter, so its distribution table is given by:

]

k	0	1
$\mathbb{P}[Z = k]$	1/2	1/2

.

4. Z takes values 0 and 1, so $g(Z) = Z^2 = Z$. Therefore, its distribution table is the same as that of Z:

k	0	1
$\mathbb{P}[g(Z) = k]$	1/2	1/2

.