Chapter 1

Random variables and their properties

1.1 What is a random variable?

Random variables are a tremendously useful concept in probability theory. In most situations of science and general life we are not really interested in the actual outcome of a random experiment. Instead, we care about particular properties of that outcome.

For starters, let us talk about the weather. Like most people living in the north of Europe, your choice of clothing probably depends strongly on the weather. Let us assume that you have a thick winter jacket that you put on when it is cold, a light soft-shell jacket that you wear when temperatures are mild and that you simply wear a T-shirt or a sweater when it is warm. Let us also assume that the forecasting site you consult every morning reports the temperatures up to one decimal. The question is: does it matter to you whether it is 5.4 or 7.3 degrees centigrade outside when you make your decision about what to wear? Most likely you are only interested in whether it is cold, mild or warm. This may obviously vary according to your own perception of warmth and cold. Here we will assume that it is cold whenever the temperature is below 10 degrees, mild between 10 and 20 degrees (inclusive) and warm whenever the temperature rises above 20 degrees. So what you are really interested in, at least as far as clothing is concerned, is whether the temperature t is $t < 10$ or $10 \leq t \leq 20$ or $t > 20$. What we are looking for is a way of transforming outcomes from the temperature scale into judgements of perceived temperature.

Another example is the lottery. Let us again consider the Dutch lottery where you have to correctly predict 6 out of 49 numbers and one out of 6 colours all which are drawn *independently* and *uniformly at random* (you should be able to understand the first term by now; after reading this chapter, you will also understand the second). You get the main prize if

you predict all these items correctly. However, pay-outs start from three correctly predicted numbers and increase for each additional number that you got right. So what do you really care about when playing the lottery? Does it matter to you whether or not you correctly predicted the number 39? Probably not. All you care about is how many of the numbers you got right.

Observe that one possible way of modelling the sample space for the lottery situation is the following:

$$(1.1) \qquad \Omega = \{(b_1, \ldots, b_7, c_1, \ldots, c_7) | b_1, \ldots, b_6 \in \{1, \ldots, 49\};$$
$$i \neq j \Rightarrow b_i \neq b_j; b_7 \in \{50, \ldots, 55\}; c_1, \ldots, c_7 \in \{0, 1\}\}$$

This rather cumbersome expression states that we draw 6 balls from an urn of 49 balls without replacement, then draw one ball from an urn of six balls (which we numbered higher than the balls from the first urn for better discrimination). We also add indicators of whether we predicted each ball drawn correctly. So again the question arises of how we can transform outcomes from this sample space into the events that we actually care about (namely how many balls we predicted correctly)?

After this little motivating section, let us now formally define what a random variable is.

Definition 1 *A (real) random variable (RV) is a function*

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}$$

We additionally require that each set $A = \{\omega | X(\omega \leq r\}$ be in $\mathcal{A}$ for all $r \in \mathbb{R}$.

Notice that the naming of the random variable is arbitrary but by convention we will often use the letters X, Y, Z. Also, as you can see, RVs operate on an entire probability space. The reason is that we want to have a probability measure associated with our sample space. In fact, this probability measure is the reason why random variables are assuming their values in a random way. Moreover, since we can only measure events from our events space, the condition on RVs ensures that we can in fact measure the probability of each real value that we may map an outcome to. In technical terms, we are thus making sure that the image of the RV is *measurable*.

Let us define a random variable for the weather situation.

$$(1.2) \qquad X(t) = \begin{cases} 1 & t < 10 \\ 2 & 10 \leq t \leq 20 \\ 3 & t > 2 \end{cases}$$

In principle we can now easily determine the probability of each of the values of this RV. However, we did not bother to define a sample space and associated probability measure, mostly because we would have to restrict the possible temperatures to those that we think are reasonable. Nevertheless, this is something that you can do yourself.

Exercise 1 *Define an upper and lower bound for temperature values (make sure you can actually get all three perceived degrees of warmth). This will give you a discrete sample space since we postulated that the website we consult only reports temperatures up to one decimal place after the comma. Then impose any probability measure on the associated event space. Finally, compute the probability of each of the values of the RV X from eq. 1.2.*

We can do the same for the lottery. Our random variable would then look as follows:

$$(1.3) \qquad X(\omega) = \sum_{i=1}^{7} c_i$$

Whenever a RV assumes a particular value, we write $X = x$. Thus, if X in eq. 1.3 assumes the value 4, i.e. we guessed for balls correctly, we would write $X = 4$. And here comes the function that will officially compute the probability that this happens for us.

Definition 2 *The discrete probability distribution of a RV X is the function*
$$P_X(X = x) := \mathbb{P}(\{\omega | X(\omega) = x\})$$

Now what is $P_X(X = 4)$ for the lottery example? Recall that there are $a = \binom{49}{6} \times 6$ sequences of balls that we can draw. Now in the above scenario we also have to take into account whether or not we correctly predicted each ball, leaving us with a total of 2^a outcomes. At the same time there are $b = \binom{45}{2} \times 6 + \binom{46}{3} \times 5$ outcomes in which we have made exactly 4 correct predictions. Thus we get $P_X(X = 4) = 2^b/2^a$.

We can of course also make statements like $X < x$ and $X \geq x$. This is a way of expressing that a RV assumes a value above or below a certain threshold. The way to get the probability that this would happen is through the **cumulative distribution function**.
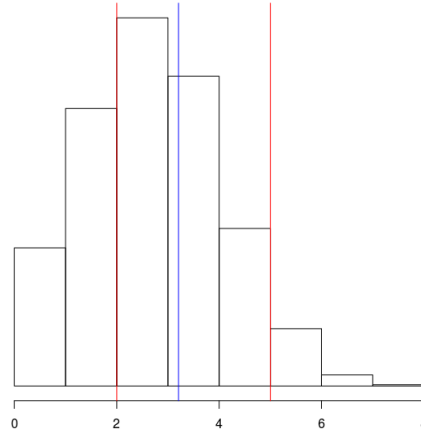
3

Figure 1.1: Probability distribution of a random variable with 8 possible values.

Definition 3 *The cumulative distribution function (cdf) of a random variable X is given as*

$$F_X(a) := P_X(X \leq a) = \sum_{x \leq a} P(X = x)$$

The cumulative distribution function is helpful in two ways. First of all, it allows us to get the probability that X takes on a value in a certain range. Second, it has applications in sampling algorithms, some of which we might see during the programming part of the course.

To better understand how we can compute the probability that $a \leq X \leq b$ for some real values a, b let us take a look at fig. 1.1. There we see a plot of the probability distribution of a random variable Z. The red lines indicate the interval that we are interested in. As a warm-up exercise, let us first compute $P_Z(Z \leq 2)$ and $P_Z(Z > 5)$. By looking at the plot we easily see which areas we need to consider, namely the bars to the left of 2 and to the right of 5. The height of the bars corresponds to their probability. The probability distribution is given as

- $P_Z(Z = 1) = 0.09$

- $P_Z(Z = 2) = 0.21$

- $P_Z(Z = 3) = 0.28$

- $P_Z(Z = 4) = 0.23$

- $P_Z(Z = 5) = 0.12$

- $P_Z(Z = 6) = 0.04$

- $P_Z(Z = 7) = 0.02$

- $P_Z(Z = 8) = 0.01$

We go on to compute the probability that Z is smaller or equal to 2 and that Z is bigger or equal to 5.

$$(1.4) \qquad P_Z(Z \leq 2) = P_Z(Z = 1) + P_Z(Z = 2) = 0.09 + 0.21 = 0.3$$
$$(1.5) \qquad P_Z(Z > 5) = P_Z(Z = 6) + P_Z(Z = 7) + P_Z(Z = 8)$$
$$= 0.04 + 0.02 + 0.01 = 0.07$$

Now it is straightforward to compute $P_Z(2 < Z \leq 5)$. If we again look at fig. 1.1 we see that this is the area delimited by the red lines. Thus, $P_Z(Z \leq 5)$ will be an overestimate of this probability. The amount by which it is an overestimate is exactly $P_Z(Z \leq 2)$. Thus we get $P_Z(2 < Z \leq 5) = P_Z(Z \leq 5) - P_Z(Z \leq 2)$.

Exercise 2 *We claim that you can compute $P_Z(2 < Z \leq 5)$ based on $P_Z(Z \leq 2)$ and $P_Z(Z > 5)$ from above. What result do you get?*

In the case of discrete random variables, i.e. the ones that we are dealing with here, there is another function that is of great interest, namely the **probability mass function**.

Definition 4 *The probability mass function (pmf) of a random variable X is given as*
$$p(x) := P_X(X = x)$$

The pmf is notationally more convenient than the probability distribution but also less general. For example, the pmf does not allow us to express that X falls in a range of values. In order to avoid confusion, we will mostly use X's probability distribution here. Be aware, however, that most papers that you are going to read in the future will use the pmf instead. Regardless of whether we are using the probability distribution or the pmf, we will call the values in $\mathbb{R}$ to which they assign positive probability, their **support**.

1.2 Expectation and Variance

There are two major properties of probability distributions that are used to describe them. One is the center of mass, the point at which the probability mass of the distribution is split into half. If you think of the support of P_X as a plank and of the probabilities as weights, then the center of mass is the point at which you could put an infinitely thin spike into the plank from underneath such that the plank maintains perfect balance. In fig. 1.1 this so-called **expectation** is indicated by the blue line.

Definition 5 *The expectation of a random variable X with respect to the distribution P_X is defined as*

$$E_{P_X}[X] := \sum_{i=1}^{|X|} x_i P_X(x_i)$$

where $|X|$ is the size of the support of X and x_i are the members of that support.

Exercise 3 *Compute the expectation of the random variable Z from the previous section.*

It is important to point to the fact that the expectation is a theoretical quantity. This implies that it does not need to be a value in the support of X. This is for example the case in fig. 1.1 where the expectation is fractional although all support values are integers.

The expectation comes with an interesting property that is called the **linearity of expectation**. Basically whenever we multiply X with a constant a the expectation will be scaled by that constant. Moreover, when we add a constant b to X the expectation will also increase/decrease by b (depending on whether b is positive or negative). Let's proof that!

$$(1.6) \qquad E_{P_X}[aX + b] = \sum_{i=1}^{|X|} a(X = x_i + b)P_X(x_i)$$

$$(1.7) \qquad = a\sum_{i=1}^{|X|}(x_i + b)P_X(X = x_i)$$

$$(1.8) \qquad = a\sum_{i=1}^{|X|} x_i P_X(X = x_i) + bP_X(X = x_i)$$

$$(1.9) \qquad = a\sum_{i=1}^{|X|} x_i P_X(X = x_i) + \sum_{i=1}^{|X|} bP_X(X = x_i)$$

$$(1.10) \qquad = a\sum_{i=1}^{|X|} x_i P_X(X = x_i) + b$$

$$(1.11) \qquad = aE_{P_X}[X] + b$$

The equality in 1.10 follows from the fact that the sum of the probabilities of the support of a distribution is 1. The linearity of expectation is extremely useful. Basically, if you know the expectation of a RV you also know the expectation of its multiples. Furthermore, by adding a constant,

we basically shift each element in the support. However, this shifting does not do anything surprising as the centre of mass shifts by the same amount.

We can actually take the expectation of any quantity. However, this only makes sense for quantities that can be cast as functions of X. The expectation of a quantity that is constant with respect to X will just be that quantity itself. More formally:

$$(1.12) \qquad E_{P_X}[a] = \sum_{i=1}^{|X|} a P_X(X = x_i) = a$$

A class of functions of X that people are often interested in are the moments. Moments are the powers of X, so X itself is the first moment X^2 is the second moment and so on. We will not further discuss moments here but it is useful to at least know what they are in case you find the expression in a paper.

After we have discussed the centre of mass, it is also interesting to look at how far away the average outcome (or rather its image under the random variable) is from it. Assume two random variables X and Y with

- $P_X(X = -100) = 0.5$; $P_X(X = 100) = 0.5$

- $P_Y(Y = -1000) = 0.5$; $P_Y(Y = 1000) = 0.5$

The spread of Y is obviously going to be greater, although both have the same expectation. The quantity usually used to assess the spread of a RV is the **variance**.

Definition 6 *The variance of a RV X with distribution P_X is given as*

$$var(X) := E_{P_X}[(E_{P_X}[X] - X)^2]$$

The expression for $var(X)$ is rather scary-looking but let us try to make sense of it. The inner part is the difference between each value of the random variable and the expectation. Thus the out expectation is just a weighted sum of differences between values in the support and the expectation. However, since values of X can be smaller or greater than the expectation, this sum will be 0. Therefore we square it, leaving us only with positive values. In sum, the variance is the expectation of the squared differences between the expectation and the values in the support of X.

Computing the variance as an expectation can be cumbersome. We will now show how it can be done more efficiently.

(1.13)

$$var(X) = E_{P_X}[(E_{P_X}[X] - X)^2]$$

$$(1.14) \qquad = \sum_{i=1}^{|X|} P_X(X = x_i)(E_{P_X}[X] - X)^2$$

$$(1.15) \qquad = \sum_{i=1}^{|X|} P_X(X = x_i)(E_{P_X}[X])^2 - 2X E_{P_X}[X] + X^2)$$

$$(1.16) \qquad = (E_{P_X}[X])^2 - \sum_{i=1}^{|X|} P_X(X = x_i) 2X E_{P_X}[X] + \sum_{i=1}^{|X|} P_X(X = x_i) X^2$$

$$(1.17) \qquad = (E_{P_X}[X])^2 - 2(E_{P_X}[X])^2 + E_{P_X}[X^2]$$

$$(1.18) \qquad = E_{P_X}[X^2] - (E_{P_X}[X])^2$$

For most intents and purposes we will use 1.18 to compute the variance. Also, from here on, we will generally drop the subscript indicating the distribution of the the expectation in order to save ourselves some writing.

> **Exercise 4** *Compute the variance of the lottery RV from the previous section.*

Although variance is not linear, we can still try to figure out what happens when we multiply a RV by a constant or add a constant to it.

$$(1.19) \quad var(aX + b) = E[(aX + b)^2] - (E[aX + b])^2$$

$$(1.20) \qquad\qquad\quad = E[(aX + b)^2] - (aE[X] + b)^2$$

$$(1.21) \qquad\qquad\quad = E[a^2 X^2 + 2abX + b^2] - a^2 E[X]^2 - 2abE[X] - b^2$$

$$(1.22) \qquad\qquad\quad = a^2 E[X^2] + 2abE[X] + b^2 - a^2 E[X]^2 - 2abE[X] - b^2$$

$$(1.23) \qquad\qquad\quad = a^2 E[X^2] - a^2 E[X]^2 = a^2 var(X)$$

What we see here is that the variance is not affected by constant additions to the RV. This makes sense, as constant additions just shift the expectation. However, the relation of each individual value to the expectation is not affected by that. On the other hand, when we multiply each value of the RV by a constant, the variance is scaled by the square of that constant. Again, this is not surprising. Remember that we used squaring in our definition of the variance to turn all distances from the expectation into positive numbers. Now if we multiply an RV by a constant a its spread will increase in both directions by a factor of a. And since we again want only positive differences, we better square that increased spread. This intuitively justifies the above equalities.

1.3 Joint and conditional distributions

Up to now, we have only looked at cases that could be treated with one random variable. Most interesting problems involve a multitude of RVs, however. We therefore introduce the concept of **jointly distributed random variables**. What this means is that there is a distribution over tuples of values, each from a different RV.

Definition 7 *Random variables $X_1, \ldots, X_n$ (abbreviated as X_1^n) are said to be jointly distributed with probability distribution $P_{X_1^n}$ such that*

$$P_{X_1^n}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(\{\omega | X_1(\omega) = x_1, \ldots, X_n(\omega) = x_n\})$$

From the above definition we see that the random variables need to share an underlying sample space. This also gives us more insight into the real power of RVs: in applications we usually do not care at all about the underlying sample space but only about certain quantities captured by the RVs. In fact, we may not even be sure what the underlying sample space is and how we would draw samples from it. Therefore, we simply define a probability distribution over the random variables of interest. Since by definition the distribution is linked to a probability space, we can rest assured that we are working with a suitable sample space without having to bother what that space looks like.

Joint distributions are of particular interest because they make it possible to recover the probability distribution of each individual RV as well as of smaller joint distributions. The process by which this can be accomplished is known as **marginalization**. Say we are given a joint distribution P_{XY}. Now how can we determine the probability that $X = x$? Our sample space is carved up by the values that X and Y can take on. We are interested in the probability of the subset $E = \{\omega | X(\omega) = x\}$. Assuming that Y can take on n values, let us define $E_i = \{\omega | Y(\omega) = y_i\}$ for $1 \leq i \leq n$. Some (possibly all) of the E_i will overlap with E. Thus we can partition E into $E \cap E_1, \ldots, E \cap E_n$. Now by countable additivity we simply need to sum up the probabilities of these intersections. Observe that those probabilities correspond exactly to $P_{XY}(X = x, Y = y_1), \ldots, P_{XY}(X = x, Y = y_1)$. Hence, we get to the following equality:

$$(1.24) \qquad P_X(X = x) = \sum_{i=1}^{n} P_{XY}(X = x, Y = y_i)$$

Now if we set $X = (Z_1, \ldots, Z_m)$, we can recover $P_{Z_1^m}$ from $P_{YZ_1^m}$ in the same way. Likewise, if we set $Y = (Z_1, \ldots, Z_m)$ we get the generalization of 1.24. We will assume that each Z_j can assume n_j values.

$X \backslash Y$	1	2	3
1	0.15	0.08	0.07
2	0.01	0.2	0.03
3	0.28	0.17	0.09

Table 1.1: Joint probability table for X and Y.

$$(1.25) \quad P_X(X = x) = \sum_{j_1=1}^{n_1} \ldots \sum_{j_m=1}^{n_m} P_{XZ_1^m}(X = x, Z_1 = z_{j_1}, \ldots, Z_m = z_{j_m})$$

For exercise purposes, it often helps to draw a joint probability table and do marginalization from there.

Exercise 5 *Find the marginal distributions P_X and P_Y in table 1.1.*

We have already seen how the joint probability of random variables links back the the probability space that underlies them. This makes it easy for us to import certain concepts that we have gotten to know in previous chapters. In particular, we can define **conditional probability distributions**.

Definition 8 *The probability of $X = x$ conditioned on $Y = y$ is given as*
$$P_{X|Y}(X = x | Y = y) = \frac{P_{XY}(X = x, Y = y)}{P_Y(Y = y)}$$

Exercise 6 *Compute the conditional probability distributions $P_{X|Y=2}$ and $P_{Y|X=1}$ from table 1.1.*

Likewise, we can carry over the concept of independence.

Definition 9 *Two random variables X, Y are independent (write: $X \perp Y$) if $P_{XY}(X = x, Y = y) = P_X(X = x)P_Y(Y = y)$.*

As with events, independence also implies that $P_{X|Y} = P_X$. Moreover, independence makes it much easier for us to calculate the expectation and variance of jointly distributed random variables. We have to be careful, though. We cannot just compute the expectation of the joint random variables themselves as it is unclear what this would even be. What we can do is to compute the expectation of functions of them. The standard choice in

the case of expectation is to simply multiply the RVs.

$$(1.26) \qquad E[XY] = \sum_{i=1}^{n}\sum_{j=1}^{m} P_{XY}(X = x_i, Y = y_j)x_iy_j$$

$$(1.27) \qquad = \sum_{i=1}^{n}\sum_{j=1}^{m} P_X(X = x_i)P_Y(Y = y_j)x_iy_j$$

$$(1.28) \qquad = \sum_{i=1}^{n} P_X(X = x_i)x_i \sum_{j=1}^{m} P_Y(Y = y_j)y_j$$

$$(1.29) \qquad = E[X]E[Y]$$

Similarly it does not really make sense to talk about the variance of two random variables (at least not as one single value). However, we can again compute the variance of functions. For addition, independence again makes our lives much easier.

> **Exercise 7** *Show that $var(X + Y) = var(X) + var(Y)$.*

1.4 Two important distributions

We will finish this chapter by introducing two extremely useful discrete distributions. They are very often used in computer science to model the probability of bit strings and linguistics and natural language processing to model the probability of sentences.

To get us started let us consider coin tosses. When you toss a fair coin the probability that it lands on heads is 0.5 and so is the probability that it lands on tails. Let 1 stand for heads and 0 for tails in a random variable X modelling coin tosses. Then we can formalize this distribution as

$$(1.30) \qquad P_X(X = x) = 0.5^x \times 0.5^{1-x}$$

Notice though that not all coins are fair and that in general we want to allow for probabilities that are not equal. We will therefore introduce a parameter θ that regulates the probability that we would observe heads. In the above equation we set $\theta = 0.5$. What happens if we set $\theta = 0.3$?

$$(1.31) \qquad P_X(X = x) = 0.3^x \times 0.7^{1-x}$$

Where did we get the 0.7 from? Well, we know that the coin has to land on either heads or tails. Thus, if the probability to land on heads is 0.3 the probability to land on tails has to be 0.7. Moreover, we chose 1 as a stand-in for landing on heads. As we would expect, $X = 1$ will return the 0.3 probability and $X = 0$ will return the probability of 0.7.

This leads to a general formulation where we do not specify the parameter θ but leave it to be filled in. The resulting distribution, known as **Bernoulli distribution** (with parameter θ), is shown in eq. 1.32.

$$(1.32) \qquad P_X(X = x) = \theta^x \times (1 - \theta)^{1-x}$$

The Bernoulli is defined for one coin flip, or more generally for the drawing of a coloured ball from an urn containing balls of exactly two colours. Now what if we draw balls repeatedly with replacement? We will get a sequence of coloured balls. Say we draw n balls and repeat this procedure five times. Chances are that the five sequences will look rather different. Thus, we can define a probability distribution over sequences of coloured balls of length n. This can be achieved in a simple generalization of the Bernoulli. Whereas in the Bernoulli we made one draw we now make n draws.

$$(1.33) \qquad P_X(X = x) = \theta^x \times (1 - \theta)^{n-x}$$

The adjustment is slight. We basically only introduce an extra parameter, namely the number of draws n. However, there are two conditions for this distribution that we have not mentioned yet. First and foremost, the repeated Bernoulli draws are assumed to be independent. That is, for each draw in the sequence it should not matter which colours you have drawn so far and which ones you are still going to draw. More formally, if we encode each draw in the sequence as a random variable we postulate that these random variables are independent of each other.

The second point is that we cannot assign a probability to any sequence that contains more ball of one colour than there are balls in total. This is to say that we require that $0 \le x \le n$ which is a reasonable restriction.

It is also worth mentioning that in general the two values that the distributions in 1.32 and 1.33 range over are generally referred to as success and failure where the success is usually encoded as 1 and the failure as 0 (although nothing hinges on that).

Exercise 8 *Use the distribution in eq. 1.33 with $n = 10$ and $\theta = 0.8$ to determine the probability of the sequence (0,0,1,1,0,1,1,1,0,1).*

Another interesting observation is that, since the value at each position in the sequence is independent of all others, all sequences with the same number of successes and failures will have the same probability. At this point it is natural to ask for the probability to get *any* sequence with that amount of successes and failures. Say we are dealing with sequences of length n and wonder about the probability of obtaining a sequence with $0 \le x \le n$ successes. Then we simply need to count up the probabilities of all sequences that contain x successes. But how many such sequences are there? If you recall chapter 1, there we talked about how to choose a subset of k elements

out of n. This is exactly the problem we are facing here. We want to know how many ways there are to choose x out of n positions that we interpret as successes. This counting is done for us by the binomial co-efficient $\binom{n}{x}$.

We can now generalize the distribution from eq. 1.33 to a distribution that gives the probability of obtaining any sequence with x successes. This distribution is known as the **binomial distribution** (with parameters n and θ).

$$(1.34) \qquad P_X(X = x) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}$$

The binomial is of crucial importance in many fields. For example in computer science it is used to compute the probability of bit strings. Other applications include the assessment of failure rates. Say you run a company whose production relies on a machine. If the machine has a certain probability to break down while working on one item of your product, you surely are interested to know what the probability of failure is when you need to produce n items in order to meet the demand of your customers.

The second important distribution that we get to know today is the **multinomial distribution**. It is basically a generalization of the binomial. A RV that is distributed according to the multinomial distribution can take on finitely many values. Let us say that there are m such values. Furthermore let us assume that each value occurs c_i times in a sequence of length n where $1 \leq i \leq n$ and $\sum_{i=1}^{m} c_i = n$. Then the multinomial with parameters n and $\theta_1, \ldots, \theta_n$ is given as

$$(1.35) \qquad P_{X_1^m}(X_1 = c_1, \ldots, X^m = c^m) = \frac{n!}{\prod\limits_{i=1}^{m} c_i!} \prod_{i=1}^{m} \theta_i^{c_i}$$

A further condition on the multinomial is that $\sum_{i=1}^{m} \theta_i = 1$. Just as with the binomial, we can in principle infer θ_m and c_m if we know the value of the other θ_i and c_i, $1 \leq i \leq (m - 1)$. However, this would clutter notation unnecessarily and we hence do not do it here. It is noteworthy, though, that because of this fact, the multinomial has $m - 1$ and not m θ-parameters (since θ_m is determined by the others θ.

It is easy to see that the binomial is in fact just a special case of the multinomial. Because of its general importance it is usually considered a separate distribution though.

Let us finish this section by introducing some useful notation. When a RV is distributed according to some distribution, one often uses the tilde to express this fact. If, for example, we have a random variable X that is distributed according to a binomial with $n = 100$ and $\theta = 0.5$ many authors

will write

$$X \sim binom(100, 0.5)$$

Now that we know what it means for a random variable to be distributed according to some distribution, we can also clarify a question that was brought up in the beginning. What does sampling *uniformly at random* mean? It just means that we are sampling from a distribution that is the uniform distribution (all values have the same probability). The *random* comes from the fact that we are sampling the values of a random variable. Just saying that some quantity is sampled at random is not enough. You should always add which distribution underlies that randomness!