1 Probability basics

Because of the reader's interest in *information theory*, it is assumed that, to some extent, he or she is relatively familiar with *probability theory*, its main concepts, theorems, and practical tools. Whether a graduate student or a confirmed professional, it is possible, however, that a good fraction, if not all of this background knowledge has been somewhat forgotten over time, or has become a bit rusty, or even worse, completely obliterated by one's academic or professional specialization!

This is why this book includes a couple of chapters on *probability basics*. Should such basics be crystal clear in the reader's mind, however, then these two chapters could be skipped at once. They can always be revisited later for backup, should some of the associated concepts and tools present any hurdles in the following chapters. This being stated, some expert readers may yet dare testing their knowledge by considering some of this chapter's (easy) problems, for starters. Finally, any parent or teacher might find the first chapter useful to introduce children and teens to probability.

I have sought to make this review of probabilities basics as simple, informal, and practical as it could be. Just like the rest of this book, it is definitely not intended to be a math course, according to the canonic theorem–proof–lemma–example suite. There exist scores of rigorous books on probability theory at all levels, as well as many Internet sites providing elementary tutorials on the subject. But one will find there either too much or too little material to approach Information Theory, leading to potential discouragement. Here, I shall be content with only those elements and tools that are needed or are used in this book. I present them in an original and straightforward way, using fun examples. I have no concern to be rigorous and complete in the academic sense, but only to remain accurate and clear in all possible simplifications. With this approach, even a reader who has had little or no exposure to probability theory should also be able to enjoy the rest of this book.

1.1 Events, event space, and probabilities

As we experience it, reality can be viewed as made of different environments or situations in time and space, where a variety of possible *events* may take place. Consider dull and boring life events. Excluding future possibilities, basic events can be anything like:

- It is raining,
- I miss the train,
- Mom calls,
- The check is in the mail,
- The flight has been delayed,
- The light bulb is burnt out,
- The client signed the contract,
- The team won the game.

Here, the events are defined in the present or past tense, meaning that they are known facts. These known facts represent something that is either true or false, experienced or not, verified or not. If I say, "Tomorrow *will* be raining," this is only an assumption concerning the future, which may or may not turn out to be true (for that matter, weather forecasts do not enjoy universal trust). Then tomorrow will tell, with rain being a more likely possibility among other ones. Thus, future events, as we may expect them to come out, are well defined facts associated with some degree of likelihood. If we are amidst the Sahara desert or in Paris on a day in November, then rain as an event is associated with a very low or a very high likelihood, respectively. Yet, that day precisely it may rain in the desert or it may shine in Paris, against all preconceived certainties. To make things even more complex (and for that matter, to make life exciting), a few other events may occur, which weren't included in any of our predictions.

Within a given environment of causes and effects, one can make a list of all possible events. The set of events is referred to as an *event space* (also called *sample space*). The event space includes anything that can possibly happen.[1] In the case of a sports match between opposing two teams, A and B, for instance, the basic event space is the four-element set:

$$S = \left\{ \begin{array}{c} \text{team A wins} \\ \text{team A loses} \\ \text{a draw} \\ \text{game canceled} \end{array} \right\}, \tag{1.1}$$

with it being implicit that if team A wins, then team B loses, and the reverse. We can then say that the events "team A wins" and "team B loses" are strictly equivalent, and need not be listed twice in the event space. People may take bets as to which team is likely to win (not without some local or affective bias). There may be a draw, or the game may be canceled because of a storm or an earthquake, in that order of likelihood. This pretty much closes the event space.

When considering a trial or an experiment, events are referred to as *outcomes*. An experiment may consist of picking up a card from a 32-card deck. One out of the 32 possible outcomes is the card being the Queen of Hearts. The event space associated

[1] In any environment, the list of possible events is generally infinite. One may then conceive of the event space as a limited set of well defined events which encompass all known possibilities at the time of the inventory. If other unknown possibilities exist, then an event category called "other" can be introduced to close the event space.

with this experiment is the list of all 32 cards. Another experiment may consist in picking up two cards successively, which defines a different event space, as illustrated in Section 1.3, which concerns *combined* and *joint* events.

The *probability* is the mathematical measure of the likelihood associated with a given *event*. This measure is called p(event). By definition, the measure ranges in a zero-to-one scale. Consistently with this definition, $p(\text{event}) = 0$ means that the event is *absolutely unlikely* or "impossible," and $p(\text{event}) = 1$ is *absolutely certain*.

Let us not discuss here what "absolutely" or "impossible" might really mean in our physical world. As we know, such extreme notions are only relative ones! Simply defined, without purchasing a ticket, it is impossible to win the lottery! And driving 50 mph above the speed limit while passing in front of a police patrol leads to absolute certainty of getting a ticket. Let's leave alone the weak possibilities of finding by chance the winning lottery ticket on the curb, or that the police officer turns out to be an old schoolmate. That's part of the event space, too, but let's not stretch reality too far. Let us then be satisfied here with the intuitive notions that impossibility and absolute certainty do actually exist.

Next, formalize what has just been described. A set of different events in a family called x may be labeled according to a series $x_1, x_2, \ldots, x_N$, where N is the number of events in the event space $S = \{x_1, x_2, \ldots, x_N\}$. The probability $p(\text{event} = x_i)$, namely, the probability that the outcome turns out to be the event x_i, will be noted $p(x_i)$ for short.

In the general case, and as we well know, events are neither "absolutely certain" nor "impossible." Therefore, their associated probabilities can be any real number between 0 and 1. Formally, for all events x_i belonging to the space $S = \{x_1, x_2, \ldots, x_N\}$, we have:

$$0 \leq p(x_i) \leq 1. \tag{1.2}$$

Probabilities are also commonly defined as percentages. The event is said to have anything between a 0% chance (impossible) and a 100% chance (absolutely certain) of happening, which means strictly the same as using a 0–1 scale. For instance, an election poll will give a 55% chance of a candidate winning. It is equivalent to saying that the odds for this candidate are 55:45, or that $p(\text{candidate wins}) = 0.55$.

As a fundamental rule, the sum of all probabilities associated with an event space S is equal to unity. Formally,

$$p(x_1) + p(x_2) + \cdots p(x_N) = \sum_{i=1}^{i=N} p(x_i) = 1. \tag{1.3}$$

In the above, the symbol Σ (in Greek, capital S or *sigma*) implies the summation of the argument $p(x_i)$ with index i being varied from $i = 1$ to $i = N$, as specified under and above the sigma sign. This concise math notation is to be well assimilated, as it will be used extensively throughout this book. We can interpret the above summation rule according to:

It is absolutely certain that one event in the space will occur.

This is another way of stating that the space includes *all* possibilities, as for the game space defined in Eq. (1.1). I will come back to this notion in Section 1.3, when considering combined probabilities.

But how are the probabilities calculated or estimated? The answer depends on whether or not the event space is well or completely defined. Assume first for simplicity the first case: we know for sure all the possible events and the space is complete. Consider then two familiar games: coin tossing and playing dice, which I am going to use as examples.

Coin tossing

The coin has two sides, heads and tails. The experiment of tossing the coin has two possible outcomes (heads or tails), if we discard any possibility that the coin rolls on the floor and stops on its edge, as a third physical outcome! To be sure, the coin's mass is also assumed to be uniformly distributed into both sides, and the coin randomly flipped, in such a way that no side is more likely to show up than the other. The two outcomes are said to be *equiprobable*. The event space is $S = \{\text{heads}, \text{tails}\}$, and, according to the previous assumptions, $p(\text{heads}) = p(\text{tails})$. Since the space includes all possibilities, we apply the rule in Eq. (1.3) to get $p(\text{heads}) = p(\text{tails}) = 1/2 = 0.5$. The odds of getting heads or tails are 50%. In contrast, a realistic coin mass distribution and coin flip may not be so perfect, so that, for instance, $p(\text{heads}) = 0.55$ and $p(\text{tails}) = 0.45$.

Rolling dice (game 1)

Play first with a single die. The die has six faces numbered one to six (after their number of spots). As for the coin, the die is supposed to land on one face, excluding the possibility (however well observed in real life!) that it may stop on one of its eight corners after stopping against an obstacle. Thus the event space is $S = \{1, 2, 3, 4, 5, 6\}$, and with the equiprobability assumption, we have $p(1) = p(2) = \cdots = p(6) = 1/6 \approx 0.166\,666\,6$.

Rolling dice (game 2)

Now play with two dice. The game consists in adding the spots showing in the faces. Taking successive turns between different players, the winner is the one who gets the highest count. The sum of points varies between $1 + 1 = 2$ to $6 + 6 = 12$, as illustrated in Fig. 1.1. The event space is thus $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, corresponding to 36 possible outcomes. Here, the key difference from the two previous examples is that the events (sum of spots) are not equiprobable. It is, indeed, seen from the figure that there exist six possibilities of obtaining the number $x = 7$, while there is only one possibility of obtaining either the number $x = 2$ or the number $x = 12$. The count of possibilities is shown in the graph in Fig. 1.2(a).

Such a graph is referred to as a *histogram*. If one divides the number of counts by the total number of possibilities (here 36), one obtains the corresponding probabilities. For instance, $p(x = 2) = p(x = 22) = 1/36 = 0.028$, and $p(x = 7) = 6/36 = 0.167$. The different probabilities are plotted in Fig. 1.2(b). To complete the plot, we have

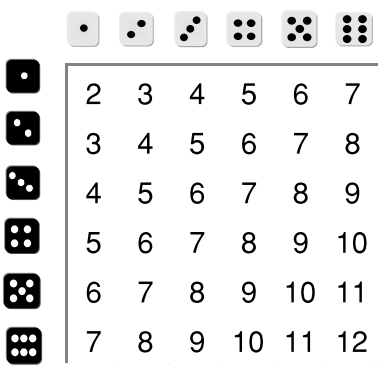

Figure 1.1 The 36 possible outcomes of counting points from casting two dice.

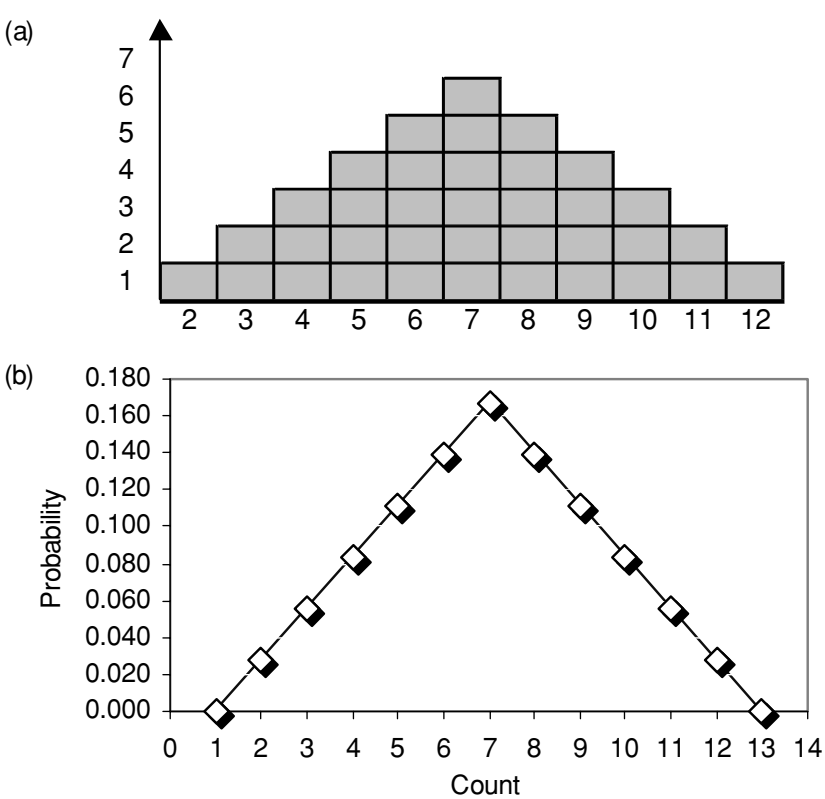

Figure 1.2 (a) Number of possibilities associated with each possible outcome of casting two dice, (b) corresponding probability distribution.

included the two count events $x = 1$ and $x = 13$, which both have zero probability. Such a plot is referred to as the *probability distribution*; it is also called the *probability distribution function* (PDF). See more in Chapter 2 on PDFs and examples. Consistently with the rule in Eq. (1.3), the sum of all probabilities is equal to unity. It is equivalent to say that the *surface* between the PDF curve linking the different points $(x, p(x))$ and

the horizontal axis is unity. Indeed, this surface is given by $s = (13-1)^{*}p(x=7)/2 = 12^{*}(6/36)/2 \equiv 1$.

The last example allows us to introduce a fundamental definition of the probability $p(x_i)$ in the general case where the events x_i in the space $S = \{x_1, x_2, \ldots, x_N\}$ do not have equal likelihood:

$$p(x_i) = \frac{\text{number of possibilities for event } i}{\text{number of possibilities for all events}}. \tag{1.4}$$

This general definition has been used in the three previous examples. The single coin tossing or single die casting are characterized by equiprobable events, in which case the PDF is said to be *uniform*. In the case of the two-dice roll, the PDF is nonuniform with a triangular shape, and peaks about the event $x = 7$, as we have just seen.

Here we are reaching a subtle point in the notion of probability, which is often mistaken or misunderstood. The known fact that, in principle, a flipped coin has equal chances to fall on heads or tails *provides no clue as to what the outcome will be*. We may just observe the coin falling on tails several times in a row, before it finally chooses to fall on heads, as the reader can easily check (try doing the experiment!). Therefore, the meaning of a probability is not the prediction of the outcome (event x being verified) but the measure of how likely such an event is. Therefore, it actually takes quite a number of trials to measure such likelihood: one trial is surely not enough, and worse, several trials could lead to the wrong measure. To sense the difference between probability and outcome better, and to get a notion of how many trials could be required to approach a good measure, let's go through a realistic coin-tossing experiment.

First, it is important to practice a little bit in order to know how to flip the coin with a good feeling of randomness (the reader will find that such a feeling is far from obvious!). The experiment may proceed as follows: flip the coin then record the result on a piece of paper (heads = H, tails = T), and make a pause once in a while to enter the data in a computer spreadsheet (it being important for concentration and expediency not to try performing the two tasks altogether). The interest of the computer spreadsheet is the possibility of seeing the statistics plotted as the experiment unfolds. This creates a real sense of fun. Actually, the computer should plot the cumulative count of heads and tails, as well as the experimental PDF calculated at each step from Eq. (1.4), which for clarity I reformulate as follows:

$$\begin{cases} p(\text{heads}) = \dfrac{\text{number of heads counts}}{\text{number of trials}} \\ p(\text{tails}) = \dfrac{\text{number of tails counts}}{\text{number of trials}}. \end{cases} \tag{1.5}$$

The plots of the author's own experiment, by means of 700 successive trials, are shown in Fig. 1.3. The first figure shows the cumulative counts for heads and tails, while the second figure shows plots of the corresponding experimental probabilities p(heads), p(tails) as the number of trials increases. As expected, the counts for heads and tails are seemingly equal, at least when considering large numbers. However, the detail shows that time and again, the counts significantly depart from each other, meaning that there are more heads than tails or the reverse. But eventually these discrepancies seem to correct

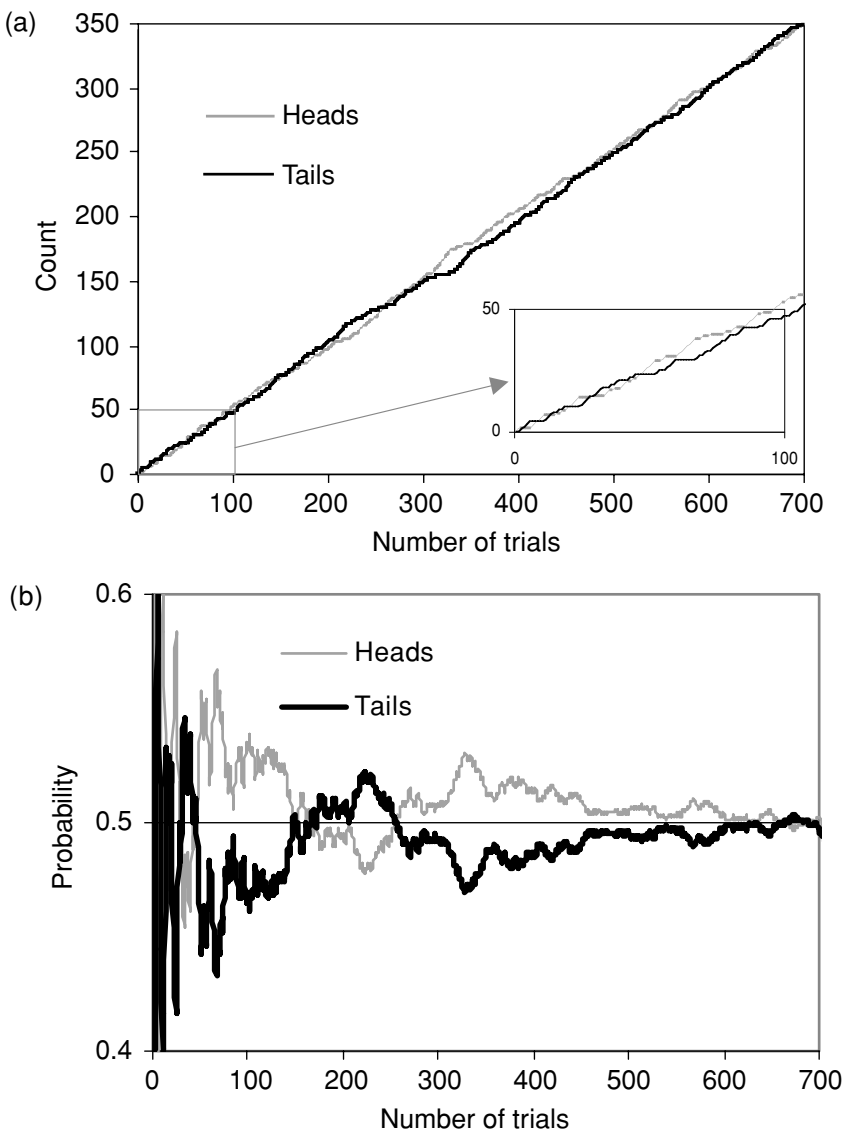

Figure 1.3 Experimental determination of the probability distribution of coin flipping, by means of 700 successive trials: (a) cumulative count of head and tails outcomes with inset showing detail for the first 100 trials, (b) corresponding probabilities.

themselves as the game progresses, as if the coin would "know" how to come back to the 50:50 odds rule. Strange isn't it? The discrepancies between counts are reflected by the wide oscillations of the PDF (Fig. 1.3(b)). But as the experiment progresses, the oscillations are damped to the point where $p(\text{heads}) \approx p(\text{tails}) \approx 0.5$, following an asymptotic behavior.[2]

[2] Performing this experiment and obtaining such results is not straightforward. Different types of coins must be tested first, some being easier to flip than others, because of size or mass. Some coins seem not to lead

The above experiment illustrates the difference between event probabilities and their actual outcomes in the physical world. The nice thing about probability theory is that the PDF gives one a sense of the unknown when it comes to a relatively large number of outcomes, as if the unknown, or "chance," were domesticated by underlying mythical principles. On the other hand, a known probability gives no clue about a single event, just a sense of what it is most likely to be. A fair way to conceive of a 10% odds is that the corresponding event "should be observed" at least once after ten outcomes, and at least ten times after 100 outcomes, and very close to $Q/10$ times after Q outcomes, the closeness being increasingly accurate as Q becomes larger. The expression "should be observed" progresses towards "should be absolutely certain." To top off this statement, we can say that *an event with a finite, nonzero probability is absolutely certain to occur at least once in the unbounded future*. Such a statement is true provided the physics governing the event (and its associated PDF) remain indefinitely the same, or is "invariant by time translation" in physicists' jargon.

1.2 Combinatorics

The examples in the previous section concerned events whose numbers of possible occurrences in a given trial are easily defined. For instance, we know for certain that a single die has exactly six faces, with their corresponding numbers of spots. But when adding spots from two dice (rolling dice, game 2), we had to go through a kind of inventory in order to figure out all the different possibilities. While this inventory was straightforward in this example, in the general case it can be much more tedious if not very complex. Studying the number of different ways of counting and arranging events is called *combinatorial analysis* or *combinatorics*. Instead of formalizing combinatorics at once, a few practical examples are going to be used to introduce its underlying concepts progressively.

Arranging books on a shelf

This is a recurrent problem when moving into a new home or office. A lazy solution is to unpack the books and arrange them on the shelf from left to right, as we randomly pick them up from the box. How many different ways can we do this?

Answer: Say the shelf can hold ten books, and number the book position from the left. Assume that the box fits in ten different (or distinguishable) books. The number of ways to pick up the first book to place in position 1 is, therefore, $Q = 10$. For the second book to put in position 2, there remain nine possibilities. So far, there have been $Q = 10 \times 9$ possibilities. Positioning next the third book, the count of possibilities becomes $Q = 10 \times 9 \times 8$. And so on, until the last book, for which there is only a single

to equiprobable outcomes, even over a number of trials as high as 1000 and after trying different tossing methods. If the coin is not 100% balanced between the two sides in terms of mass, nonuniform PDFs can be obtained.

choice left, which gives for the total number of possibilities: $Q = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1$. By definition, this product is called the *factorial* of 10 and is noted 10!

More generally, for any integer number n we have the factorial definition:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1. \tag{1.6}$$

Thus

$$1! = 1,$$
$$2! = 2 \times 1 = 2,$$
$$3! = 3 \times 2 \times 1 = 6,$$

and so on. One can easily compute 10! (using, for instance, the factorial function of a scientific pocket calculator) to find

$$10! = 3\,628\,800 \approx 3.6 \text{ million}.$$

There is an impressive number of ways indeed to personalize one's office shelves!

When it comes to somewhat greater arguments (e.g., $n = 100$), the factorials cannot be computed by hand or through a pocket calculator, owing to their huge size. To provide an idea (as computed through a spreadsheet math function up to a maximum of 170!):

$$50! \approx 3.0 \times 10^{64},$$
$$100! \approx 9.3 \times 10^{157},$$
$$170! \approx 7.2 \times 10^{306}.$$

In this case it is possible to use *Stirling's approximation theorem*, which is written as

$$n! \approx n^n \mathrm{e}^{-n} \sqrt{2\pi n}. \tag{1.7}$$

Remarkably, the Stirling theorem is accurate within 1.0% for arguments $n \geq 10$ and within 0.20% for arguments $n \geq 40$.

To summarize, *the factorial of n is the number of ways to arrange n distinguishable elements into a given orderly fashion*. Such an arrangement is also called a *permutation*. What about the factorial of the number *zero*? By convention, mathematicians have set the odd property

$$0! = 1.$$

We shall accept here that $0! = 1$ without advanced justification. Simply put, if not a satisfactory explanation, there is only one way to arrange/permute a set containing zero element.[3]

Consider next a second example.

[3] For the math-oriented reader, it is interesting to mention that the factorial function can be generalized to *any real or even complex arguments* x, i.e., $x! = \Gamma(x+1)$, where Γ is the gamma function, which can be computed numerically.

Arranging books on a shelf, with duplicates

Assume that some of the books have one or even several duplicate copies, which cannot be distinguished from each other, as in a bookstore. For instance, the series of 10 books includes two brand-new English–Russian dictionaries. Having these two in shelf positions (a, b) or (b, a) represents the same arrangement. So as not to count this arrangement twice, we should divide the previous result (10!) by two (2!), which is the number of possible permutations for the duplicated dictionaries. If we had three identical books in the series, we should divide the result by six (3!), and so on. It is clear, then, that *the number of ways of arranging n elements containing p indistinguishable elements and $n - p$ distinguishable ones is given by the ratio:*

$$A_n^p = \frac{n!}{p!}. \tag{1.8}$$

The above theorem defines the number of possible arrangements "without repetition" (of the indistinguishable copies). Consistently, if the series contains n indistinguishable duplicates, the number of arrangements is simply $A_n^n = n!/n! = 1$, namely, leaving a unique possibility.

The next example will make us progress one step more. Assume that we must make a selection from a set of objects. The objects can be all different, all identical, or partly duplicated, which does not matter. We would only like to know the number of possibilities there are to make any random selection of these objects.

Fruit-market shopping

In a fruit market, the stall displays 100 fruits of various species and origins. We have in mind to pick at random up to five fruits, without preference. There are 100 different possibilities to pick up the first fruit, 99 possibilities to pick up the second, and so on until the last fifth, which has 96 possibilities left. The total number of possibilities to select five specific fruits out of 100 is therefore $Q = 100 \times 99 \times 98 \times 97 \times 96$. Based on the definition of factorials in Eq. (1.6), we can write this number in the form $Q = 100!/95! = 100!/(100 - 5)!$ But in each selection of five fruits we put into the bag, it does not matter in which order they have been selected. All permutations of these five specific fruits, (which are 5!), represent the same final bag selection. Therefore, the above count should be divided by 5! because of the 5! possible redundancies. The end result is, therefore, $Q = 100!/[5!(100 - 5)!]$. Most generally, *the number of ways to pick up p unordered samples out of a set of n items* is

$$C_n^p = \frac{n!}{p!(n-p)!}. \tag{1.9}$$

The number C_n^p, which is also noted $\binom{n}{p}$ or ${}_nC_p$ or $C(n, p)$ is called the *binomial coefficient*.[4]

[4] Since the factorial is expandable to a continuous function (see previous note), the binomial coefficient is most generally defined for any real/complex x, y numbers as $C_x^y = \Gamma(x+1)/\Gamma(y)[\Gamma(x-y+1)]$. Beautiful plots of C_x^y in the real x, y plane, and more on the very rich binomial