Companies often promote their products by hiding something inside and promising a reward for those who collect and send back a certain number of these items. In this paper we formulate a typical problem related to this commercial technique, and investigate the solution by means of the theory of probability.

The problem is the following. *Suppose that a manufacturer produces a certain kind of soft drink, and it marks the inside of the cap of each bottle with a coloured square. The task is to collect one of each of the six different colours for a reward. How many bottles do we expect to have to buy?* (The reader is encouraged to guess.) In the solution, the following conditions will be important: there is *exactly one* coloured square in *each cap,* the manufacturer produces a *large quantity* of the drink, and the *colours are distributed evenly.*

Throughout this article we are going to use the terminology of probability theory, and will, but only briefly, explain its basic notions in an elementary (and somewhat sloppy) way.

The first concept is that of a *random variable,* whose name refers to a function that assigns (at least for the purpose of this paper) a real number to each possible outcome of a random event. The random variable is said to be discrete if it can only assume a finite number of values or, if the set of its values is infinite but denumerable. (All random variables in this paper will be discrete.) Let $X$ denote the random variable in our discussions. The distribution of a random variable describes the probabilities that belong to the various values of the random variable, and these (necessarily non-negative) probabilities should add up to 1. Consider, for example a fair die with the number 3 written on one face, 6 on two faces and 8 on the remaining three faces. The value of the random variable $X$ is the number read on top when the die has been rolled. The table below shows the values and the distribution of $X$.

| The values of the random variable | 3 | 6 | 8 |
|---|---|---|---|
| The distribution of the variable | $\frac{1}{6}$ | $\frac{1}{3}$ | $\frac{1}{2}$ |

If an experiment is carried out a large number of times, then the arithmetic mean of the values of $X$ observed will be varying around a certain number called the *expected value,* or expectation of the random variable, and is usually denoted by $\mathrm{E}(X)$. A more precise definition of the expected value of a discrete random variable is as follows:

*If the values of a discrete random variable $X$ are $x_1, x_2, \ldots, x_n, \ldots$ and the corresponding probabilities are $p_1, p_2, \ldots, p_n, \ldots$, respectively, then the expected value of the variable is* $\mathrm{E}(X) = \sum_i p_i x_i$, *provided that the sum is finite. The expectation is said to be infinite otherwise.*

In case of the die in the above example, the expected value is

$$\mathrm{E}(X) = 3 \cdot \frac{1}{6} + 6 \cdot \frac{1}{3} + 8 \cdot \frac{1}{2} = 6.5.$$

If the trial has an infinite number of possible outcomes – for example when we are investigating the number of tosses needed for a coin to land on head – the sum in the above definition has infinitely many terms. The expected value exists if the corresponding series converges, and it is equal to the sum of the series as defined in real analysis. Convergent series with positive terms are in many ways like finite sums: their terms can be rearranged or grouped, and it is allowed to multiply each term by a number and such series can be summed up term by term. These legal transformations will be used in the paper without any further explanation.

Note that a random variable does not necessarily have a finite expectation, and even if it does, the expected value is not necessarily equal to any of the values $x_i$. Furthermore, if it happens to coincide with a certain value $x_i$, it is not necessarily the most probable value.

Now let us return to the original problem.

In **Solution 1** let the random variable $X$ denote the number of bottles we buy until we first collect all the six different colours. Since we obviously need at least six of them, the possible values of $X$ are $6, 7, 8, 9, \ldots$ . The question is the number of bottles we should expect to buy. What does that mean? We cannot expect that whenever we have bought a certain fixed number of bottles, we are going to have all the colours. The meaning of the number in question is the average of the number of purchases if we repeat the collection several times or, equivalently, if there are several people buying bottles with coloured squares in the caps. In other words, the number in question is the expected value of the random variable $X$. In order to determine this expectation, we need to know the distribution of the random variable $X$, that is the probabilities of each of the values $6, 7, 8, 9, \ldots$ $X$ may assume. In general: what is the probability that it is the $n$-th bottle we buy that makes our collection of six colours complete? The probability is found by dividing the number of favourable cases by the total number of possible outcomes. There may be six different colours found in each of the $n$ purchases. Hence the total number of cases is $6^n$. In favourable cases, only 5 colours occur during the first $(n-1)$ purchases, but not fewer.

The number of such outcomes can be found with an application of the logical sieve:

*If each of $N$ objects may have some of the properties $a_1, a_2, \ldots, a_n$, and $N(x\ y\ z\ \ldots)$ denotes the number of objects having the properties $x, y, z, \ldots$, then the number $N^*$ of those objects having none of the listed properties is*

$$N^* = N - \big(N(a_1) + N(a_2) + \cdots + N(a_n)\big) + \big(N(a_1\ a_2) + \cdots + N(a_{n-1}\ a_n)\big)$$
$$- \big(N(a_1\ a_2\ a_3) + \cdots + N(a_{n-2}\ a_{n-1}\ a_n)\big) + \cdots + (-1)^n N(a_1\ a_2\ \ldots\ a_n).$$

How can this theorem be applied to our problem? Suppose we keep a record of our purchases by placing marks of the appropriate colour on a strip of paper. Having completed the $(n-1)$ purchases, we will have a string of $(n-1)$ coloured marks. This string of marks will play the role of the objects in the above theorem. Since each mark may have one of $(n-1)$ different colours, $5^{n-1}$ strings are possible, that is, $N = 5^{n-1}$. Let the property $a_i$ $(i = 1, 2, 3, 4, 5)$ mean that the $i$-th colour does not occur in the string. Similarly, let $a_{ij}$ denote the absence of both the $i$-th and $j$-th colours, and so on. In this notation, $N^*$ stands for the number of strings that contain all five colours. (That is what we want to calculate.)

$N(a_1) + N(a_2) + \cdots + N(a_5)$ is the number of strings missing one colour. That colour may be any one of the 5 colours, that can be selected in $\binom{5}{1} = 5$ ways. The remaining 4 colours are possible in each position, and there are $4^{n-1}$ choices for this. Thus the number of such strings is $5 \cdot 4^{n-1}$.

$N(a_1\ a_2)+\cdots+N(a_4\ a_5)$ is the number of strings missing 2 colours. Those two colours can be selected in $\binom{5}{2} = 10$ ways, with 3 choices remaining for each position. Thus there are $10 \cdot 3^{n-1}$ such strings.

Similarly, the number of possible strings missing 3 colours is $10 \cdot 2^{n-1}$, and finally there are 5 strings missing 4 colours.

Therefore $N^* = 5^{n-1} - 5 \cdot 4^{n-1} + 10 \cdot 3^{n-1} - 10 \cdot 2^{n-1} + 5$. Since the colour that does not occur in the first $(n-1)$ purchases can be any of 6 different colours, the number of favourable outcomes is $6 \cdot (5^{n-1} - 5 \cdot 4^{n-1} + 10 \cdot 3^{n-1} - 10 \cdot 2^{n-1} + 5)$.

Hence the probability $p_n$ in question is

$$p_n = \frac{6 \cdot (5^{n-1} - 5 \cdot 4^{n-1} + 10 \cdot 3^{n-1} - 10 \cdot 2^{n-1} + 5)}{6^n}$$
$$= \left(\frac{5}{6}\right)^{n-1} - 5 \cdot \left(\frac{4}{6}\right)^{n-1} + 10 \cdot \left(\frac{3}{6}\right)^{n-1} - 10 \cdot \left(\frac{2}{6}\right)^{n-1} + 5 \cdot \left(\frac{1}{6}\right)^{n-1}.$$

Now we can calculate the expected value:

$$\mathrm{E}(X) = \sum_{n=6}^{\infty} np_n = \sum_{n=6}^{\infty} n \left(\frac{5}{6}\right)^{n-1} - 5 \cdot \sum_{n=6}^{\infty} n \left(\frac{4}{6}\right)^{n-1}$$
$$+ 10 \cdot \sum_{n=6}^{\infty} n \left(\frac{3}{6}\right)^{n-1} - 10 \cdot \sum_{n=6}^{\infty} n \left(\frac{2}{6}\right)^{n-1} + 5 \cdot \sum_{n=6}^{\infty} n \left(\frac{1}{6}\right)^{n-1}.$$

Discarding the coefficients, each term is an infinite series of the form $\sum_{n=6}^{\infty} nq^{n-1}$. Let $S$ denote this expression, and multiply the equality by $q$:

$$qS = 6q^6 + 7q^7 + 8q^8 + 9q^9 + 10q^{10} + \cdots + nq^n + \cdots.$$

Now subtract the two equalities:

$$(1-q)S = 6q^5 + q^6 + q^7 + q^8 + q^9 + \cdots + q^n + \cdots$$
$$= 6q^5 + q^6(1 + q + q^2 + q^3 + \cdots + q^n + \cdots).$$

It is known that for $|q| < 1$ the infinite geometric series $1 + q + q^2 + \cdots + q^n + \cdots$ is convergent and its sum is $\frac{1}{1-q}$. Hence $(1-q)S = 6q^5 + q^6 \frac{1}{1-q}$ and thus

$$S = q^5 \frac{6 - 5q}{(1-q)^2}.$$

(Here we did a little bit of cheating. Guess where, and how it can be fixed.)

Now we can plug the numbers $\frac{5}{6}, \frac{4}{6}, \frac{3}{6}, \frac{2}{6}, \frac{1}{6}$ for $q$ ($|q| < 1$ in each case) to obtain the expected value:

$$\mathrm{E}(X) = 66 \cdot \left(\frac{5}{6}\right)^5 - 120 \cdot \left(\frac{4}{6}\right)^5 + 140 \cdot \left(\frac{3}{6}\right)^5 - 97.5 \cdot \left(\frac{2}{6}\right)^5 + 37.2 \cdot \left(\frac{1}{6}\right)^5 = 14.7.$$

There is also a more elegant way to calculate the same expected value.

**Solution 2** is based on the *additive property* of the expected value. It is easy to show that if the random variable $X$ is obtained as the sum of the random variables $X_1, X_2, \ldots, X_n$, that is, $X = X_1 + X_2 + \cdots + X_n$, then $\mathrm{E}(X) = \mathrm{E}(X_1) + \mathrm{E}(X_2) + \cdots + \mathrm{E}(X_n)$.

Let the random variable $X_1$ denote the number of bottles we need to buy to get one containing the first colour in the cap, let $X_2$ denote the number of additional bottles to buy in order to have a second colour, and so on. It is clear that the sum of the random variables $X_i$ $(i = 1, 2, 3, 4, 5, 6)$ is equal to the random variable $X$. (It is also clear that the values of the variables $X_i$ are independent of each other. Independence is not a necessary condition for the additive property of the expected value, but it will play an important role later on.)

Buying the first bottle of drink we will certainly have a colour we have not had before, so $\mathrm{E}(X_1) = 1$. The probability of finding a square of a different colour in the cap of the next bottle bought is $\frac{5}{6}$. What is the expected value of the number of purchases needed to get hold of a second colour? In other words, what is $\mathrm{E}(X_2)$?

That can be determined (by using the definition of the expected value) as follows: 1 times the probability of getting it for the first time plus 2 times the probability of two purchases needed, and so on to infinity, since in principle we may keep finding the same colour in the cap forever. Therefore

$$\begin{aligned}\mathrm{E}(X_2) &= 1 \cdot \left(\frac{5}{6}\right) + 2 \cdot \left(\frac{1}{6}\right)\left(\frac{5}{6}\right) + 3 \cdot \left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) + 4 \cdot \left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right) + \cdots \\ &= \left(\frac{5}{6}\right) \cdot \left[1 + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right)^2 + 4\left(\frac{1}{6}\right)^3 + \cdots\right].\end{aligned}$$

The sum of the infinite series in the brackets is $\left(\frac{6}{5}\right)^2$, it can be evaluated in a similar way as in Solution 1. Once we have two squares of different colours, the probability of finding a third colour in the cap of another bottle is $\frac{4}{6}$, and accordingly, the expected value of the number of additional purchases to get a third colour is $\frac{6}{4}$. Likewise, we get $\frac{6}{3}$ for a fourth colour, $\frac{6}{2}$ for a fifth one, and finally $\frac{6}{1}$ for the last colour.

Hence the expected value of the total number of bottles to buy in order to get all the six different colours is

$$1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} = 14.7.$$

Solution 2 will become simpler by pointing out that the distribution of the variables $X_i$ is called a geometric distribution, whose expectation is known to be the reciprocal of the probability of „success".

We have calculated the expected value in two different ways. However, this is but a „theoretical" value around which the actual number of purchases will scatter. Although the expected value is 14.7, we may still have to buy more than 50 bottles before we get the sixth colour, or if we are lucky, six bottles may also be enough. But what are the odds that we need to buy so many (or so few) bottles? (Using common sense, we would say: not much.) In order to answer this we need to determine how much the actual values of the random variable scatter about the expected value. *Standard deviation* is a measure for that. Its square is called the *variance,* and is denoted by $\mathrm{D}^2(X)$. Variance expresses the expected value of the square of the deviation of $X$ from the expected value $\mathrm{E}(X)$, that is

$$\mathrm{D}^2(X) = \mathrm{E}\big(\big[X - \mathrm{E}(X)\big]^2\big).$$

It is often easier to calculate the variance by means of the equality $\mathrm{D}^2(X) = \mathrm{E}(X^2) - \big(\mathrm{E}(X)\big)^2$. We may apply this relation to determine the variance of the random variable $X$ introduced in Solution 1. Although the calculation is not difficult, it is somewhat tedious, and we leave it to the reader so that he can check his skills. It is much simpler to determine the variance of the random variable of the exemplary die. Thus we get

$$\mathrm{D}^2(X) = \left(9 \cdot \frac{1}{6} + 36 \cdot \frac{1}{3} + 64 \cdot \frac{1}{2}\right) - 6.5^2 = 3.25.$$

In Solution 2, the standard deviation is easily calculated by applying the additive property of variance to the variances of the geometrically distributed variables. (The independence of the random variables is necessary for the additivity of variances. Recall that the variables $X_i$ introduced in Solution 2 were independent.) The variance of a geometrically distributed variable in general is $\mathrm{D}^2(X) = \frac{1-p}{p^2}$, where $p$ is the probability of the event whose first occurrence is investigated in the experiment.

In our problem, the variance is 0 until the first colour is obtained (since the first colour will always occur on the

first purchase, that is $p = 1$). Then

| | | |
|---|---|---|
| the variance up to the appearance of the second colour: | 0,24, | $p = \frac{5}{6}$ |
| the variance up to the appearance of the third colour: | 0,75, | $p = \frac{4}{6}$ |
| the variance up to the appearance of the fourth colour: | 2, | $p = \frac{3}{6}$ |
| the variance up to the appearance of the fifth colour: | 6, | $p = \frac{2}{6}$ |
| the variance up to the appearance of the sixth colour: | 30, | $p = \frac{1}{6}$. |

The sum of the variances is thus

$$\mathrm{D}^2(X) = 0 + 0.24 + 0.75 + 2 + 6 + 30 = 38.99.$$

Accordingly, the standard deviation is $\mathrm{D}(X) = \sqrt{38.99} = 6.2442$.

The expected value is 14.7, and now we also know that the standard deviation is about 6.2. This result in itself still does not reveal very much, since the actual number of necessary purchases may still deviate from the expected value by much more than the standard deviation. The probability of a „large" deviation, however, is small. We can use *Chebyshev's inequality* for an estimation. It claims the following:

*If $\lambda$ is an arbitrary real number greater than 1, then the probability that a random variable will deviate from its expected value by more than $\lambda$ times its standard deviation is no greater than the reciprocal of the square of $\lambda$.*

Stated in a more formal way:

$$\mathrm{P}\big(|X - \mathrm{E}(X)| \geq \lambda \cdot \mathrm{D}(X)\big) \leq \frac{1}{\lambda^2}.$$

Hence, for example, the probability of having to buy more than 27 bottles (which differs from the expected value by about two standard deviations) is no more than 0.25. If a number $\varepsilon$ is substituted for the expression $\lambda \cdot \mathrm{D}(X)$, the inequality takes the following form:

$$\mathrm{P}\big(|X - \mathrm{E}(X)| \geq \varepsilon\big) \leq \frac{\mathrm{D}^2(X)}{\varepsilon^2}.$$

This form can be used for finding an upper bound for the probability of the number of actual purchases to fall outside a symmetrical interval centred at the expected value. In other words, we can estimate the probability that the number of purchases will be within a certain symmetrical interval. This form of the inequality is more convenient in many cases. For example, find an estimate on the probability that we have to buy at least 50 bottles. 50 differs from the expected value by $50 - 14.7 = 35.3$, and since the random variable $X$ cannot have a value less than 6, Chebyshev's inequality can be applied with $\varepsilon = 35.3$:

$$\mathrm{P}\big(|X - 14.7| \geq 35.3\big) \leq \frac{38.99}{35.3^2},$$

which is about 0.03. Hence the probability of having to buy more than 50 bottles is not large, 3 percent at most. (What is the probability that 6 bottles will be enough? It does not take Chebyshev's inequality to find that.)

By using Chebyshev's inequality, we can establish that at least 74% of the actual numbers of purchases necessary will be between 6 and 27, and we can be at least 89% certain that 34 bottles will be enough for collecting all six colours. (It is a great advantage of Chebyshev's inequality that it provides estimates for the probability by a very simple and quick calculation. With the help of a good calculator (for example one with graphic display), we can easily get the accurate value of the probability as well. The true probability of less than 28 bottles needed to get the six colours is actually more than 0.95.) Now we are ready to answer the original question: we need to buy 15 bottles on average if we want to collect all six colours, but the number of bottles needed will only very rarely exceed 27. (The chances are less than 0.05.)

There is still one more question we have not yet studied: For what $n$ will the probability $p_n$, defined in Solution 1 be a maximum? Will that happen around the expected value 14.7, or somewhere else? Those who have ever calculated that the most probable number of rolls needed to get the first six on a fair die is 1 (however unbelievable that sounds to someone who plays *Aggravation*) will guess that $n$ cannot be very big. If we calculate the first few values of $p_n$ (starting with $n = 6$), we will find that the probability increases up to $n = 11$ and then it starts to decrease. We will show that if $n > 10$, then $p_n > p_{n+1}$ and thus the maximum of $p_n$ indeed occurs at $n = 11$.

Our statement means that for every $n$ greater than 10,

$$\left(\frac{5}{6}\right)^{n-1} - 5 \cdot \left(\frac{4}{6}\right)^{n-1} + 10 \cdot \left(\frac{3}{6}\right)^{n-1} - 10 \cdot \left(\frac{2}{6}\right)^{n-1} + 5 \cdot \left(\frac{1}{6}\right)^{n-1}$$

$$> \left(\frac{5}{6}\right)^{n} - 5 \cdot \left(\frac{4}{6}\right)^{n} + 10 \cdot \left(\frac{3}{6}\right)^{n} - 10 \cdot \left(\frac{2}{6}\right)^{n} + 5 \cdot \left(\frac{1}{6}\right)^{n}.$$

Multiply the inequality by $6^n$:

$$6 \cdot 5^{n-1} - 30 \cdot 4^{n-1} + 60 \cdot 3^{n-1} - 60 \cdot 2^{n-1} + 30 > 5^n - 5 \cdot 4^n + 10 \cdot 3^n - 10 \cdot 2^n + 5.$$

Writing $r^n = r \cdot r^{n-1}$ in each term of the right-hand side, the inequality can be rearranged as follows:

$$5^{n-1} - 10 \cdot 4^{n-1} + 30 \cdot 3^{n-1} - 40 \cdot 2^{n-1} + 25 > 0.$$

The smallest value of $n$ we consider is 11. Substituting that, we get a positive number on the left-hand side. For $n > 11$, introduce a new variable: let $n = k + 12$ where $k$ is a natural number. Thus $n - 1 = k + 11$. By substitution and applying the rules of indices, we have

$$5^{11} \cdot 5^k - 10 \cdot 4^{11} \cdot 4^k + 30 \cdot 3^{11} \cdot 3^k - 40 \cdot 2^{11} \cdot 2^k + 25 > 0.$$

Since $5^{11} > 10 \cdot 4^{11}$ and $30 \cdot 3^{11} > 40 \cdot 2^{11}$, and in each case the larger coefficient multiples a larger number, the left-hand side of the inequality is positive. Since all steps can be reversed, the original statement is also true, and the maximum of $p_n$ occurs at $n = 11$.

When contemplating our chances to win the reward, we have more than our intuitive feelings to rely on. Knowing the results, we can also guess if the behaviour of a company is fair. If experience contradicts the calculated values, the distribution of the colours may not be even. If so, it is not necessarily the manufacturer's fault. If we always go to the same shop, the delivery company may be the one to blame.