93.16 The coupon collector's problem with unequal probabilities

The classic coupon collector's problem, where k different items are to be collected (say by opening packets with one coupon in each) with each item having probability $\frac{1}{k}$ of occurring, is well known (see [1]). The expected number of packets which need to be opened to give a full set is

$$k \sum_{i=1}^{k} \frac{1}{i}. \tag{1}$$

The unequal probabilities case recently came up in a *Times Educational Supplement* online forum [2]. The problem was solved in 1954 by Herman Von Schelling [3]; however, the following solution uses only techniques available to A Level students and results in a different form for the formula for the expected number of packets. Comparison with the equal probability case leads to a result concerning the harmonic series.

Two kinds of coupon

Consider two types of coupon, A and B, with probabilities p and q, respectively, where $p + q = 1$.

Consider buying n packets and opening them all, one at a time. Let $P_{n,2}$ be the probability that there are at least one of each of type A and type B in n packets, that is the probability of completing a full set in n packets, or fewer.

The terms of $(A + B)^n$ give all the different combinations of A and B in the n packets and the corresponding terms of $(p + q)^n$ give the probabilities of these combinations.

$$P_{n,2} = (p + q)^n - p^n - q^n. \tag{2}$$

(This simplifies to $P_{n,2} = 1 - p^n - q^n$, which is the same as saying that getting at least one of each is the complement of them all being the same, but the unsimplified formula holds even when $p + q < 1$.)

Three kinds of coupon

Consider three types of coupon, A, B and C with probabilities p, q and r, respectively, where $p + q + r = 1$.

Consider buying n packets and opening them all, one at a time. Let $P_{n,3}$ be the probability that there are at least one of each of type in n packets, that is the probability of completing a full set in n packets, or fewer.

The terms of $(A + B + C)^n$ give all the different combinations of A, B and C in the n packets and the corresponding terms of $(p + q + r)^n$ give the probabilities of these combinations.

For at least one of each type of coupon, we need to subtract the probabilities of getting exactly two kinds and of getting all n the same. Thus

$$P_{n,3} = (p+q+r)^n - \{(p+q)^n - p^n - q^n\} - \{(q+r)^n - q^n - r^n\} - \{(r+p)^n - r^n - p^n\} - p^n - q^n - r^n$$
$$= (p+q+r)^n - \{(p+q)^n + (q+r)^n + (r+p)^n\} + \{p^n + q^n + r^n\}. \quad (3)$$

This formula still holds if $p + q + r < 1$ since it consists of those terms of $(p + q + r)^n$ that include all three of p, q, r.

Working inductively

Similarly, for four kinds of coupon with probabilities p, q, r and s

$$P_{n,4} = (p+q+r+s)^n - \Sigma_4(p+q+r)^n + \Sigma_4(p+q)^n - \Sigma_4 p^n \quad (4)$$

where Σ_4 denotes summation of all such terms which are combinations from the four probabilities p, q, r and s. This formula can be derived by writing out all the terms and collecting like terms, but the amount of writing this requires becomes increasingly daunting for larger numbers of coupons. Proceeding as shown below forms the outline of a method that can be used to prove the general case, by induction.

$$P_{n,4} = (p+q+r+s)^n - \{\Sigma_4 P_{n,3} + \Sigma_4 P_{n,2} + \Sigma_4 p^n\}$$
$$\Sigma_4 P_{n,3} = \Sigma_4(p+q+r)^n - \Sigma_4\{\Sigma_3(p+q)^n\} + \Sigma_4\{\Sigma_3 p^n\}.$$

Each $\Sigma_3(p+q)^n$ contains ${}^3C_2$ terms like $(p+q)^n$ hence $\Sigma_4\{\Sigma_3(p+q)^n\}$ contains ${}^4C_3 \times {}^3C_2$ such terms. Moreover, $\Sigma_4\{\Sigma_3(p+q)^n\}$ is a symmetric function and so it is a multiple of $\Sigma_4(p+q)^n$, which has ${}^4C_2$ terms. Therefore $\Sigma_4\{\Sigma_3(p+q)^n\} = \dfrac{{}^4C_3 \times {}^3C_2}{{}^4C_2}\Sigma_4(p+q)^n$.

The terms to be subtracted from $(p+q+r+s)^n$ are:

$$\Sigma_4 P_{n,3} = \Sigma_4(p+q+r)^n - \frac{{}^4C_3 \times {}^3C_2}{{}^4C_2}\Sigma_4(p+q)^n + \frac{{}^4C_3 \times {}^3C_1}{{}^4C_1}\Sigma_4 p^n$$
$$\Sigma_4 P_{n,2} = \Sigma_4(p+q)^n - \frac{{}^4C_2 \times {}^2C_1}{{}^4C_1}\Sigma_4 p^n$$
$$\Sigma_4 p^n = \Sigma_4 p^n$$

and the total of these terms is $\Sigma_4(p+q+r)^n - \Sigma_4(p+q)^n + \Sigma_4 p^n$, which gives formula (4).

Binomial coefficients

Expressions of the form $\dfrac{{}^aC_b \times {}^bC_d}{{}^aC_d}$ simplify.

$$\frac{{}^aC_b \times {}^bC_d}{{}^aC_d} = \frac{a!}{b!(a-b)!}\frac{b!}{d!(b-d)!}\frac{d!(a-d)!}{a!} = \frac{(a-d)!}{(a-b)!(b-d)!} = {}^{a-d}C_{a-b}.$$

Five kinds of coupon

For five kinds of coupon with probabilities p, q, r, s and t

$$P_{n,5} = (p+q+r+s+t)^n - \Sigma_5(p+q+r+s)^n + \Sigma_5(p+q+r)^n - \Sigma_5(p+q)^n + \Sigma_5 p^n. \tag{5}$$

The terms to be subtracted from $(p + q + r + s + t)^n$ are:

$$\begin{array}{lllll}
\Sigma_5 P_{n,4} = & & & & \\
\Sigma_5(p+q+r+s)^n - \dfrac{{}^5C_4 \times {}^4C_3}{{}^5C_3} & \Sigma_5(p+q+r)^n + \dfrac{{}^5C_4 \times {}^4C_2}{{}^5C_2} & \Sigma_5(p+q)^n - \dfrac{{}^5C_4 \times {}^4C_1}{{}^5C_1} & \Sigma_5 p^n \\
\Sigma_5 P_{n,3} = & \Sigma_5(p+q+r)^n - \dfrac{{}^5C_3 \times {}^3C_2}{{}^5C_2} & \Sigma_5(p+q)^n + \dfrac{{}^5C_3 \times {}^3C_1}{{}^5C_1} & \Sigma_5 p^n \\
\Sigma_5 P_{n,2} = & & \Sigma_5(p+q)^n - \dfrac{{}^5C_2 \times {}^2C_1}{{}^5C_1} & \Sigma_5 p^n \\
\Sigma_5 p^n = & & & \Sigma_5 p^n
\end{array}$$

Adding these four and simplifying binomial expressions gives

$$\Sigma_5(p+q+r+s)^n + \left(-{}^2C_1 + 1\right)\Sigma_5(p+q+r)^n + \left({}^3C_1 - {}^3C_2 + 1\right)\Sigma_5(p+q)^n + \left(-{}^4C_1 + {}^4C_2 - {}^4C_3 + 1\right)\Sigma_5 p^n.$$

This simplifies to

$$\Sigma_5(p+q+r+s)^n + \left\{(1-1)^2 - 1\right\}\Sigma_5(p+q+r)^n + \left\{(-1+1)^3 + 1\right\}\Sigma_5(p+q)^n + \left\{(1-1)^4 - 1\right\}\Sigma_5 p^n,$$

that is $\Sigma_5(p+q+r+s)^n - \Sigma_5(p+q+r)^n + \Sigma_5(p+q)^n - \Sigma_5 p^n$ and subtracting from $(p + q + r + s + t)^n$ gives (5).

*The probability of completing the set on opening the n*th *packet*

When collecting k coupons, the probability of completing the full set on opening the nth packet is

$$\begin{array}{ll} P_{n,k} - P_{n-1,k} & \text{for } n \geqslant k; \\ 0 & \text{otherwise.} \end{array} \tag{6}$$

Expected number of packets needed to complete a full set of coupons

Let X be the number of packets needed to complete a full set of coupons.

For $k = 5$ using P_r to represent $P_{r,5}$, we have $\mathrm{E}(X) = \lim_{n\to\infty} \mathrm{E}_n(X)$ where

$$\begin{aligned} E_n(X) = \quad & nP_n && -nP_{n-1} \\ & +(n-1)P_{n-1} && -(n-1)P_{n-2} \\ & +(n-2)P_{n-2} && -(n-2)P_{n-3} \\ & \vdots && \vdots \\ & +7P_7 && -7P_6 \\ & +6P_6 && -6P_5 \\ & +5P_5 \end{aligned}$$

and therefore

$$E_n(X) = nP_n - \sum_{i=5}^{n-1} P_i$$

$$= nP_n - \sum_{i=5}^{n-1} \{(p+q+r+s+t)^i - \Sigma_5(p+q+r+s)^i + \Sigma_5(p+q+r)^i - \Sigma_5(p+q)^i + \Sigma_5 p^i\}.$$

Now $\sum_{i=5}^{n-1}(p+q+r+s+t)^i = \sum_{i=5}^{n-1} 1 = n-5$ since $p+q+r+s+t = 1$

and

$$\sum_{i=5}^{n-1}(p+q+r)^i = \frac{(p+q+r)^5\{1-(p+q+r)^{n-5}\}}{\{1-(p+q+r)\}}$$

$$= \frac{(p+q+r)^5\{1-(p+q+r)^{n-5}\}}{(s+t)}.$$

Similar results hold for all the other geometric series in $E_n(X)$.

Thus $E_n(X) = nP_n - (n-5)$ + sums of geometric series which have finite sums to infinity. As $n \to \infty$, $P_n \to 1$, hence

$$E(X) = 5 + \Sigma_5 \frac{(p+q+r+s)^5}{t} - \Sigma_5 \frac{(p+q+r)^5}{(s+t)} + \Sigma_5 \frac{(p+q)^5}{(r+s+t)} - \Sigma_5 \frac{p^5}{(q+r+s+t)}. \tag{7}$$

With similar results for values of k other than 5.

The equal probability case

If $p = q = r = s = t = \frac{1}{5}$, then (7) becomes

$$E(X) = 5 + {}^5C_4\left(\frac{4}{5}\right)^5\left(\frac{5}{1}\right) - {}^5C_3\left(\frac{3}{5}\right)^5\left(\frac{5}{2}\right) + {}^5C_2\left(\frac{2}{5}\right)^5\left(\frac{5}{3}\right) - {}^5C_1\left(\frac{1}{5}\right)^5\left(\frac{5}{4}\right).$$

This has to be give the same as the answer given by (1), hence

$$5\sum_{i=1}^{5}\frac{1}{i} = 5 + {}^5C_4\left(\frac{4}{5}\right)^5\left(\frac{5}{1}\right) - {}^5C_3\left(\frac{3}{5}\right)^5\left(\frac{5}{2}\right) + {}^5C_2\left(\frac{2}{5}\right)^5\left(\frac{5}{3}\right) - {}^5C_1\left(\frac{1}{5}\right)^5\left(\frac{5}{4}\right).$$

More generally, for $n \geqslant 2$,

$$n\sum_{i=1}^{n}\frac{1}{i} = n + \sum_{i=1}^{n-1}{}^nC_i(-1)^{i+1}\left(\frac{n-i}{n}\right)^n\left(\frac{n}{i}\right).$$

References

1. F. Mosteller, *Fifty challenging problems in probability with solutions*, Dover (1965).
2. http://community.tes.co.uk/forums/t/110234.aspx?PageIndex=1
3. Herman Von Schelling, Coupon collecting for unequal probabilities, *The Amer. Math. Monthly*, Vol. 61, No. 5. (May, 1954), pp. 306-311.

STELLA DUDZIC

MEI, Monckton House, Epsom Centre, White Horse Business Park, Trowbridge BA14 0XG

93.17 Head-tail imbalance

When studying probability, one of the first situations students study is that of repeatedly throwing an unbiased coin and counting the number of heads that appear. It is well known that there are approximately the same number of heads and tails, and the result of such an experiment shows that we are unlikely to get exactly the same number of heads as tails. We are going to consider that imbalance in this article, and we shall consider the following three questions in particular.

1. For a particular number of throws, what is the distribution of the numerical difference between the numbers of heads and of tails?
2. If we throw until we have a given number of either heads or tails, how many of the other outcome will we have?
3. How long will it take for the difference in frequencies to exceed a given value?

Throughout, we shall let N stand for the total number of throws and p, q the probabilities of a head and a tail respectively. For much of the work $p = q = \frac{1}{2}$.

Question 1

As is well known, when a coin is thrown N times the number of heads X is given by the binomial distribution, so that:

$$\mathrm{P}(X = r) = {}^NC_r p^r q^{N-r}.$$