Hypergeometric Distribution

Richard L. Smith

March 3, 2024

Consider the following problem:

There are six white balls and ten black balls in an urn. Five balls are drawn at random (without replacement). What is the probability that exactly four are white?

To fix ideas, assume the balls are numbered, so that balls 1–6 are white and the remainder 7–16 are black. We treat each possible drawing of five balls as a separate outcome, so that for instance the outcome $\{1,2,3,4,7\}$ (that these are the numbers of the five balls drawn) is a different outcome from say $\{2,3,5,6,10\}$, though both of them correspond to the outcome of four white balls and one black ball. With that understanding, the total number of possible drawings is

$$\binom{16}{5} \;=\; \frac{16!}{11!\times 5!} \;=\; \frac{16\times 15\times 14\times 13\times 12}{5\times 4\times 3\times 2\times 1} \;=\; 4368.$$

Now let's ask ourselves: out of those 4368 possible drawings, how many correspond to the one we are interested in (four white balls, one black)?

The number of ways of choosing four white balls out of 6 is $\binom{6}{4}=\frac{6\times 5}{2}=15$.

The number of ways of choosing one black ball out of 10 is $\binom{10}{1}=10$.

So there are a total of $15\times 10=150$ drawings corresponding to the outcome we are interested in.

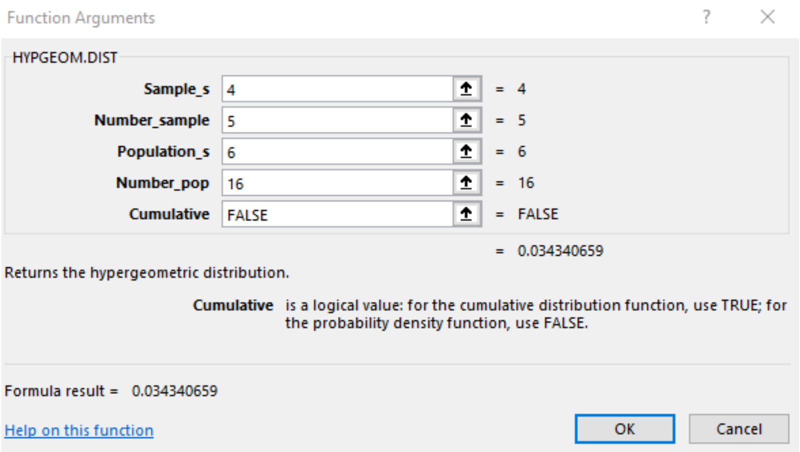
Therefore, the desired probability is

$$\frac{\binom{6}{4}\times\binom{10}{1}}{\binom{16}{5}} \;=\; \frac{15\times 10}{4368} \;=\; 0.0343 \text{ to 4dp.}$$

This is known as the *hypergeometric distribution*. The general formulation is this: suppose there are m white balls and n black balls in the urn, and we draw out k of them. The probability that exactly x balls are white (for integer x between 0 and k) is

$$\frac{\binom{m}{x}\times\binom{n}{k-x}}{\binom{m+n}{k}}.$$

Note that any time the bottom entry in the choose function is either negative or greater than the top entry, the answer is automatically 0. In particular, if $x>k$ or $k>n+x$, we have $\binom{n}{k-x}=0$ and so the probability is 0. There is nothing mysterious about this: it simply corresponds to a situation where the result of exactly x white balls, and hence $k-x$ black balls, is impossible.



HYPGEOM.DIST

Sample_s 4 = 4

Number_sample 5 = 5

Population_s 6 = 6

Number_pop 16 = 16

Cumulative FALSE = FALSE

= 0.034340659

Returns the hypergeometric distribution.

Cumulative is a logical value: for the cumulative distribution function, use TRUE; for the probability density function, use FALSE.

Formula result = 0.034340659

Figure 1: Illustration of HYPGEOM.DIST in Excel

If you know R, the relevant function is `dhyper(x,m,n,k)`. So the above example would evaluate to `dhyper(4,6,10,5)` which is 0.03434066 (which we rounded to 0.0343).

The function HYPGEOM.DIST in Excel works slightly differently: you must enter the parameters in the order x, k, m, $m+n$. Figure 1 illustrates how it works in this example. We set the value of "Cumulative" equal to FALSE, because here we want the probability of exactly 4 outcomes, not the cumulative probability of all outcomes up to 4. If you wanted the latter, you would set Cumulative to TRUE.

Exercise. Under the same values for m, n and k, find the probabilities of each of the possible x outcomes for $x = 0, 1, 2, 3, 4, 5$.

Solution. To four decimal places, the probabilities are 0.0577, 0.2885, 0.4121, 0.2060, 0.0343, 0.0014. Note that these probabilities add up to 1, which they should because one of these six outcomes has to occur (there are no other possibilities).