

In a drawing of n distinguishable objects without replacement from a set of N ($n < N$) distinguishable objects, a of which have characteristic A , ($a < N$) the probability that exactly x objects in the draw of n have the characteristic A is given by then number of different ways the x objects can be chosen from the a available times the number of different ways the $n-x$ objects in the draw which don't have A can be chosen from the $N-a$ available divided by the number of different ways n distinguishable objects can be chosen from a set of N . The resulting probability distribution for the random variable x is called the hypergeometric distribution. In symbols,

$$P(x) = \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}}.$$

The binomial coefficient $\binom{k}{j} = \frac{k!}{j!(k-j)!}$ is defined to be zero if either j or $k-j$ is negative, so that the probability of the null event of drawing more objects than those

available is zero. To prove that $\sum_{x=0}^n P(x) = \sum_{x=0}^n \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}} = 1$, consider the factorization

$(B+C)^N = (B+C)^a (B+C)^{N-a}$. From the binomial theorem,

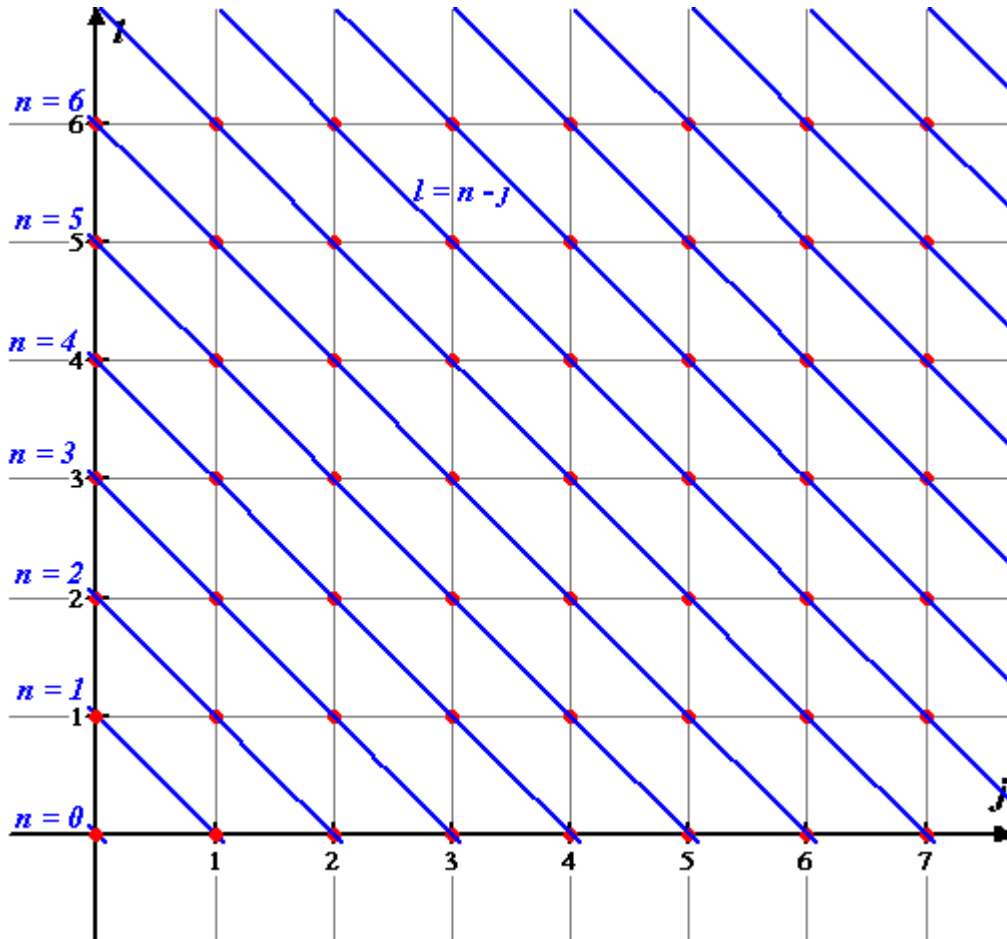
$$\begin{aligned} (B+C)^a (B+C)^{N-a} &= \sum_{j=0}^a \binom{a}{j} B^{a-j} C^j \sum_{l=0}^{N-a} \binom{N-a}{l} B^{N-a-l} C^l \\ &= \sum_{j=0}^a \sum_{l=0}^{N-a} \binom{a}{j} \binom{N-a}{l} B^{N-(l+j)} C^{l+j} \end{aligned}$$

Using the diagonal rearrangement suggested by the figure below with $l = n - j$, with the intercept n running from 0 to N and j running from 0 to a . This generates more than the $(a+1)(N-a+1)$ terms in the above sum. However, all of the new terms generated vanish since they have $l > N-a$.

$$(B+C)^a (B+C)^{N-a} = \sum_{n=0}^N \sum_{j=0}^a \binom{a}{j} \binom{N-a}{n-j} B^{N-n} C^n$$

Now, for $n > a$ extending the sum over j to n because of the $\binom{a}{j}$ factor would only add terms which are zero. Similarly, if $n < a$, the terms in the sum over j from $j = n + 1$ to $j = a$ are all zero due to the $\binom{N-a}{n-j}$ factor. Thus,

$$(B+C)^a (B+C)^{N-a} = \sum_{n=0}^N \sum_{j=0}^a \binom{a}{j} \binom{N-a}{n-j} B^{N-n} C^n = \sum_{n=0}^N \sum_{j=0}^n \binom{a}{j} \binom{N-a}{n-j} B^{N-n} C^n.$$



But from a second use of the binomial theorem,

$$(B + C)^a (B + C)^{N-a} = \sum_{n=0}^N \sum_{j=0}^n \binom{a}{j} \binom{N-a}{n-j} B^{N-n} C^n = (B + C)^N = \sum_{n=0}^N \binom{N}{n} B^{N-n} C^n .$$

The only way the two sums can be equal for all values of B and C is for

$$\sum_{j=0}^n \binom{a}{j} \binom{N-a}{n-j} = \binom{N}{n} . \quad (1)$$

This in turn implies that the hypergeometric probabilities do indeed construct a valid

probability distribution, i.e. $\sum_{x=0}^n P(x) = \sum_{x=0}^n \frac{\binom{a}{x} \binom{N-a}{n-x}}{\binom{N}{n}} = 1.$

The mean or expected value of the hypergeometric random variable is given by

$$\mu_x = \langle x \rangle = \sum_{x=0}^n x P(x) = \binom{N}{n}^{-1} \sum_{x=0}^n x \binom{a}{x} \binom{N-a}{n-x} .$$

Now, using Equation (1),

$$\begin{aligned} \sum_{x=0}^n x \binom{a}{x} \binom{N-a}{n-x} &= \sum_{x=1}^n \frac{xa!}{x!(n-x)!} \binom{N-a}{n-x} = \sum_{x=1}^n \frac{a(a-1)!}{(x-1)![(n-1)-(x-1)]!} \binom{(N-1)-(a-1)}{(n-1)-(x-1)} \\ &= a \sum_{x=0}^{n-1} \frac{(a-1)!}{x![(n-1)-x]!} \binom{(N-1)-(a-1)}{(n-1)-x} = a \sum_{x=0}^{n-1} \binom{a-1}{x} \binom{(N-1)-(a-1)}{(n-1)-x} \\ &= a \binom{N-1}{n-1} \end{aligned}$$

This gives that $\mu_x = \langle x \rangle = \sum_{x=0}^n xP(x) = a \binom{N-1}{n-1} = a \frac{(N-1)!}{(n-1)!(N-n)!} \cdot \frac{n!(N-n)!}{N!} = \frac{na}{N}$.

Using the notation of the binomial distribution that $p = \frac{a}{N}$, we see that the expected value of x is the same for both drawing without replacement (the hypergeometric distribution) and with replacement (the binomial distribution).

$$\mu_x = \langle x \rangle = \frac{na}{N} = np \quad (2)$$

The variance of the hypergeometric distribution can be computed from the generic formula that $\sigma_x^2 = \langle [x - \langle x \rangle]^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2$. Again from Equation (1),

$$\begin{aligned} \sum_{x=0}^n x(x-1) \binom{a}{x} \binom{N-a}{n-x} &= \sum_{x=2}^n \frac{x(x-1)a!}{x!(n-x)!} \binom{N-a}{n-x} = \sum_{x=2}^n \frac{a(a-1)(a-2)!}{(x-2)![(n-2)-(x-2)]!} \binom{(N-2)-(a-2)}{(n-2)-(x-2)} \\ &= a(a-1) \sum_{x=0}^{n-2} \frac{(a-2)!}{x![(n-2)-x]!} \binom{(N-2)-(a-2)}{(n-2)-x} = a(a-1) \sum_{x=0}^{n-2} \binom{a-2}{x} \binom{(N-2)-(a-2)}{(n-2)-x} \\ &= a(a-1) \binom{N-2}{n-2} \end{aligned}$$

So,

$$\begin{aligned} \langle x(x-1) \rangle &= \binom{N}{n}^{-1} \sum_{x=0}^n x(x-1) \binom{a}{x} \binom{N-a}{n-x} = \binom{N}{n}^{-1} a(a-1) \binom{N-2}{n-2} \\ &= a(a-1) \frac{(N-2)!}{(n-2)!(N-n)!} \cdot \frac{(N-n)!n!}{N!} = \frac{a(a-1)n(n-1)}{N(N-1)} \end{aligned}$$

and

$$\langle x^2 \rangle = \langle x(x-1) \rangle + \langle x \rangle = \frac{a(a-1)n(n-1)}{N(N-1)} + \frac{an}{N} = \frac{an}{N} \left[\frac{(a-1)(n-1)}{(N-1)} + 1 \right].$$

Thus,

$$\begin{aligned}\sigma_x^2 &= \langle x^2 \rangle - \langle x \rangle^2 = \frac{an}{N} \left[\frac{(a-1)(n-1)}{N-1} + 1 - \frac{an}{N} \right] = \frac{an}{N} \left[\frac{N(a-1)(n-1)}{N(N-1)} + \frac{N(N-1)}{N(N-1)} - \frac{an(N-1)}{N(N-1)} \right] \\ &= \frac{an}{N} \left[\frac{Nan - Na - Nn + N + N^2 - N - Nan + an}{N(N-1)} \right] = \frac{an}{N} \left[\frac{N^2 - Na - Nn + an}{N(N-1)} \right] \\ &= \frac{an}{N} \left[\frac{N(N-a) - n(N-a)}{N(N-1)} \right] = \frac{an}{N} \left[\frac{(N-n)(N-a)}{N(N-1)} \right] = \frac{an}{N} \left(\frac{N-a}{N} \right) \left(\frac{N-n}{N-1} \right) \\ &= \frac{an}{N} \left(1 - \frac{a}{N} \right) \left(\frac{N-n}{N-1} \right) = np(1-p) \left(\frac{N-n}{N-1} \right)\end{aligned}$$

The last factor $\left(\frac{N-n}{N-1} \right)$ is called the “finite population correction” and is the reason that the variance of the binomial distribution $np(1-p)$ differs from the hypergeometric distribution. For N large compared to the sample size n , the two distributions are essentially identical.