Hypergeometric distribution

Josemari Sarasola

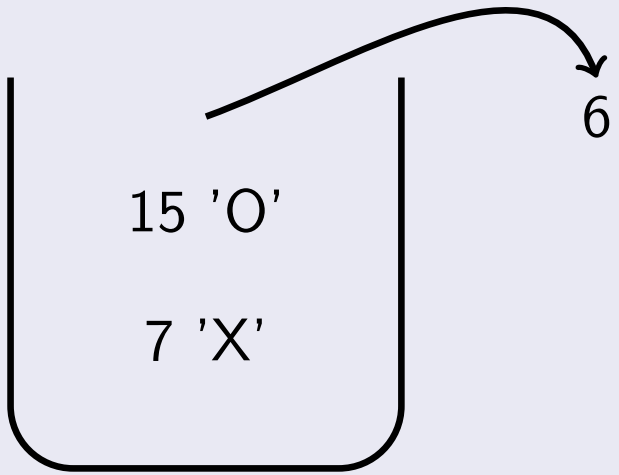
Statistics for Business

Gizapedia

Urn problems

In an urn we have 15 faultless and 7 faulty items. We take 6 items randomly. What is the probability of being among them 2 faulty items?

$$P[X = 2] = \frac{\binom{7}{2}\binom{15}{4}}{\binom{22}{6}}$$



Hypergeometric distribution

Binomial coefficients

- $n = \dfrac{n!}{k!(n-k)!}$

Some typical coefficients

- $\binom{n}{0} = 1$
- $\binom{n}{1} = n$
- $\binom{n}{n} = 1$

Trick to calculate binomial coefficients

$$\binom{8}{3} = \frac{8 \times 7 \times 6}{3!} = \frac{8 \times 7 \times 6}{6} = 8 \times 7 = 56$$

We multiply the upside number in a descending way as much times as the downside number, and in addition we divide it with the factorial of the downside number.

The urn problem again

Remember, from the [15 O, 7 X] urn we take 6 items, the probability of having 2 X:

$$P[X=2] = \frac{\binom{7}{2}\binom{15}{4}}{\binom{22}{6}} = \frac{\frac{7!}{2!5!} \times \frac{15!}{4!11!}}{\frac{22!}{6!16!}}$$

$$= \frac{\frac{7 \times 6}{2!} \times \frac{15 \times 14 \times 13 \times 11}{4!}}{\frac{22 \times 21 \times 20 \times 19 \times 18 \times 17}{6!}}$$

Other way to calculate the probability

Remember, from the [15 O, 7 X] urn we take 6 items, the probability of having 2 X:

$$P[X=2] = P[XXOOOOao]$$
$$= \frac{7}{22} \times \frac{6}{21} \times \frac{15}{20} \times \frac{14}{19} \times \frac{13}{18} \times \frac{12}{17} \times \frac{6!}{2!4!}$$

ao means 'any order' : the ways of ordering 2 X and 4 O have all the same probability, so we can take one of them (XXOOOO) and multiply with the number of permutations, that is the number of ways of ordering the elements XXOOOO. In this case, the number of the permutations for the elements XXOOOO are $\frac{6!}{2!4!}$.

Hypergeometric distribution

Calculating probabilities with R software

```
>dhyper(2,7,15,6) # from urn [15 O, 7 X], take 6,
probability for 2 X
>phyper(2,7,15,6) # from urn [15 O, 7 X], take 6,
probability for 2 X or less
>x=0:6
>dhyper(x,7,15,6) # from urn [15 O, 7 X], take 6,
probability for x X, being x=0,1,2,3,4,5,6
```

From the urn problem to the sampling without devolution

Urn problems are models for sampling in a finite population , when sampling is made without devolution.

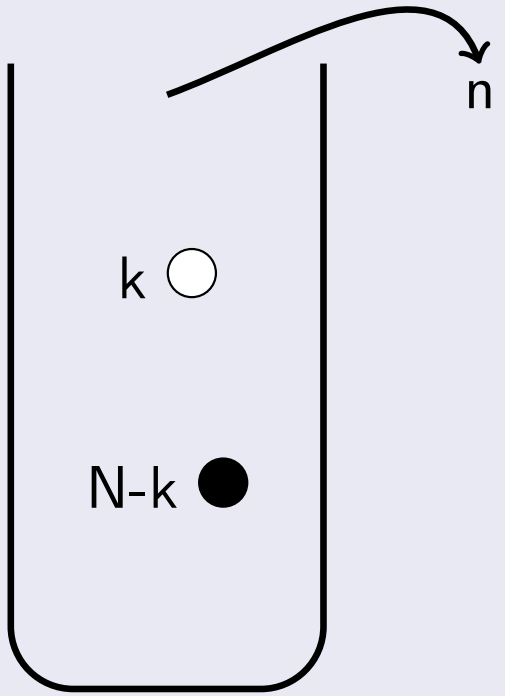
E.g., from a population fo 1000 persons with 400 women and 600 men we take a sample a 100 persons. What is the probability of having 50 men and 50 women?

$$P[X = 2] = \frac{\binom{400}{50}\binom{600}{50}}{\binom{1000}{100}}$$

Probability function

The hypergeometric distribution gives us,

- in a population of size N and taking from it a sample of n elements, without devolution (or all at the same time, as it's the same process)
- the probability of among the k elements we have set as success (○) having x of them.



$$P[X = x; N, k, n] = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

$$x = 0, 1, 2, \ldots, n$$

Symmetries

We have N balls in an urn, k white eta $N-k$ black, and we draw n balls:

- $P[X=x;N,k,n]=P[X=n-x;N,N-k,n]$ (e.g., we draw 10 balls, the probabilities of having 4 white balls and having 6 black balls are the same).
- $P[X=x;N,k,n]=P[X=k-x;N,k,N-n]$ (e.g., we have 30 balls in an urn - 18 black, 12 white -; after drawing 10 balls, the probability of having 4 black balls, and among the remaining 20 balls the probability of having 14 balls are the same).
- $P[X=x;N,k,n]=P[X=x;N,n,k]$ (e.g., we have 30 balls in an urn - 18 black, 12 white -; after drawing 10 balls, the probability of having 4 black balls, and among the 18 black balls the probability of having 4 drawn balls are the same).
- $P[X=x;N,k,n]=P[X=N-n-k+x;N,N-k,N-n]$, applying the second symmetry to the first one (e.g., we have 30 balls in an urn - 18 black, 12 white -; after drawing 10 balls, the probability of having 7 black balls, and among the remaining 20 balls the probability of having 9 white balls are the same).

Reference: https://statistics.stanford.edu/research/tables-hypergeometric-probability-distribution

Hypergeometric distribution

Support (possible values for the random variable)

- Generally ○ the possible values fot x go from 0 to n. But sometimes, for special valued parameters, this may be different:
- E.g., from a 3 white-4 black urn, we take 5 balls randomly. Possible values for number of drawn white balls are 1, 2 and 3 (and not 0, 1, 2, 3, ,4 and 5).
- So, more exactly the possible values are:

$$\max(0, n + k - N), \ldots, \min(k, n)$$

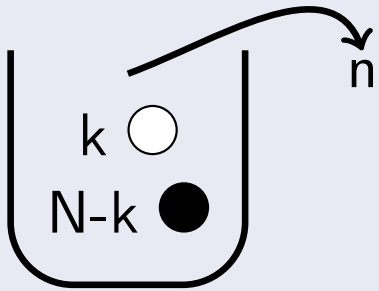
Hypergeometric distribution

Notation and parameters

$$X \sim H(k, N, n) \begin{cases} \mu = \dfrac{nk}{N} \\ \sigma^2 = \dfrac{nk}{N} \times \dfrac{N-k}{N} \times \dfrac{N-n}{N-1} \\ \text{Mo} = \left\lfloor \dfrac{(n+1)(K+1)}{N+2} \right\rfloor \end{cases}$$

Linking with the binomial distribution

- As the hypergeometric distribution, the binomial distribution gives the probability of having a concrete number of elements of a certain type in a dichotomic population, but in the hypergeometric distribution the sampling is made without devolution and in the binomial distribution is made with devolution.



- We calculate in this manner, with the binomial distribution (that is, with devolution), the probability of having x successes:

$$P[X = x] = \left(\frac{k}{N}\right)^{x} \times \left(\frac{N-k}{N}\right)^{n-x} \times \frac{n!}{x!(n-x)!}$$

Linking with the binomial distribution

In the binomial distribution the sampling units or elements are independent, that is probability for a given element is constant, but in the hypergeometric distribution sampling units are dependent, because each time we have one less element, the probability depending on the number of elements extracted so far:

$$P[X = x] = \overbrace{\frac{k}{N} \times \frac{k-1}{N-1} \times \cdots \times \frac{k-(x-1)}{N-(x-1)}}^{x \ successes} \times \underbrace{\frac{N-k}{N-x} \times \cdots \times \frac{N-(n-x)-1}{N-(n-1)}}_{n-x \ failures} \underbrace{\frac{n!}{x!(n-x)!}}_{any \ ordering}$$

Linking with the binomial distribution

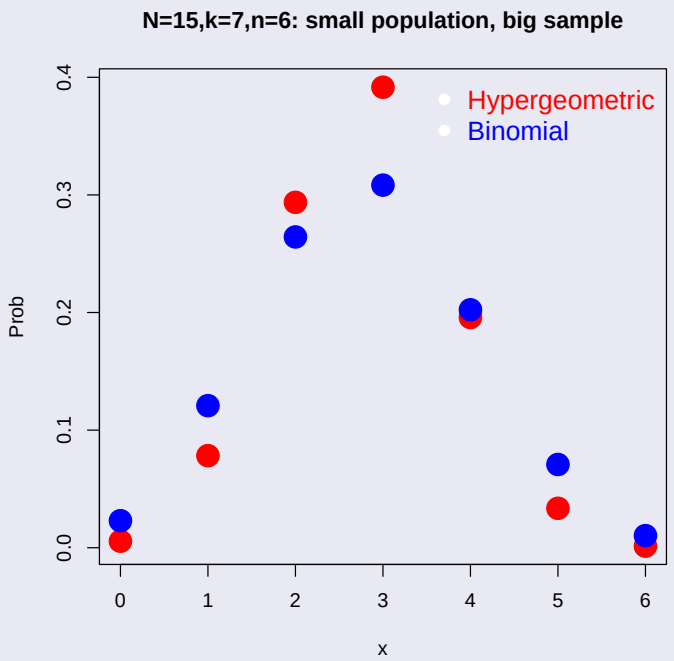
Distribution	Expected value	Variance
$B(n, p = \frac{k}{N})$	$np = \frac{nk}{N}$	$npq = \frac{nk}{N} \times \frac{N-k}{N}$
$H(k, N, n)$	$\frac{nk}{N}$	$\frac{nk}{N} \times \frac{N-k}{N} \times \frac{N-n}{N-1}$

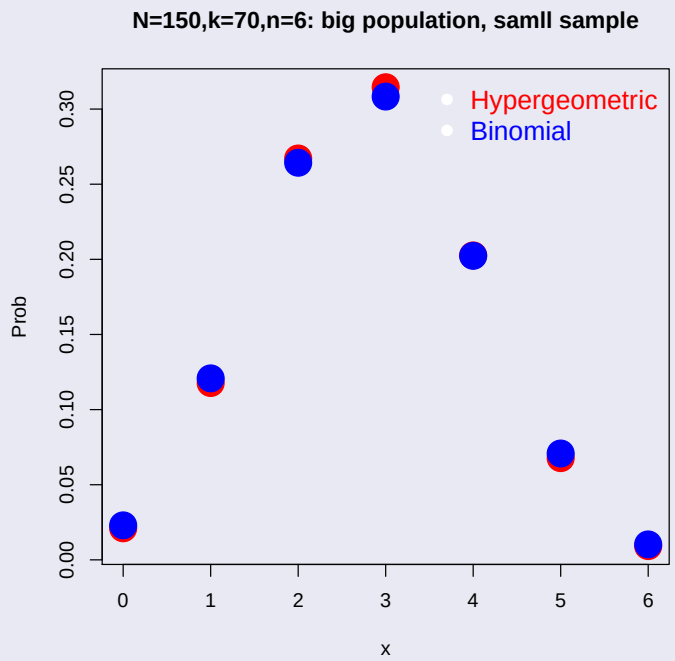
- The binomial and hypergeometric distributions have the same expected value: after extracting n elements, the mean number of successes is the same for both distributions. But the variance is smaller for the hypergeometric distribution (because we have $(N-n)/(N-1) < 1$).
- This smaller variance is quite intuitive: as we extract sampling units, population (that is, the urn) becomes smaller, and so we will have more certainty about the type of element is to be drawn.

Linking with the binomial distribution

- So, both distributions are mainly different about the $(N-n)/(N-1)$ variance factor:
  - When the population size N is big, the factor is closer to 1, so both distributions will be similar;
  - in fact, for the binomial distributions $\frac{k}{N}$ probabilities are always divided by N, but for the hypergeometric distribution are divided by N, $N-1$, ..., but for big N the difference between both types of probabilities is very small.
  - When the sampling size n is small, wtih regard to N, both distribution are similar too.
  - so finally, we may state that both distribution will be very similar when populations are big and smaples are small.

Linking with the binomial distribution





We can see that both distributions are similar when population is big and samples are small.

Linking with the binomial distribution

- So for big populations and small samples, even for sampling without devolution, is usual to take the binomial distribution and independence of probabilities as a model, in order the simplify the calculations.
- For small populations and big samples (more concretely when samples are 5-10% of the populations), sampling formulas are modified with the correction factor $(N-n)/(N-1) < 1$.