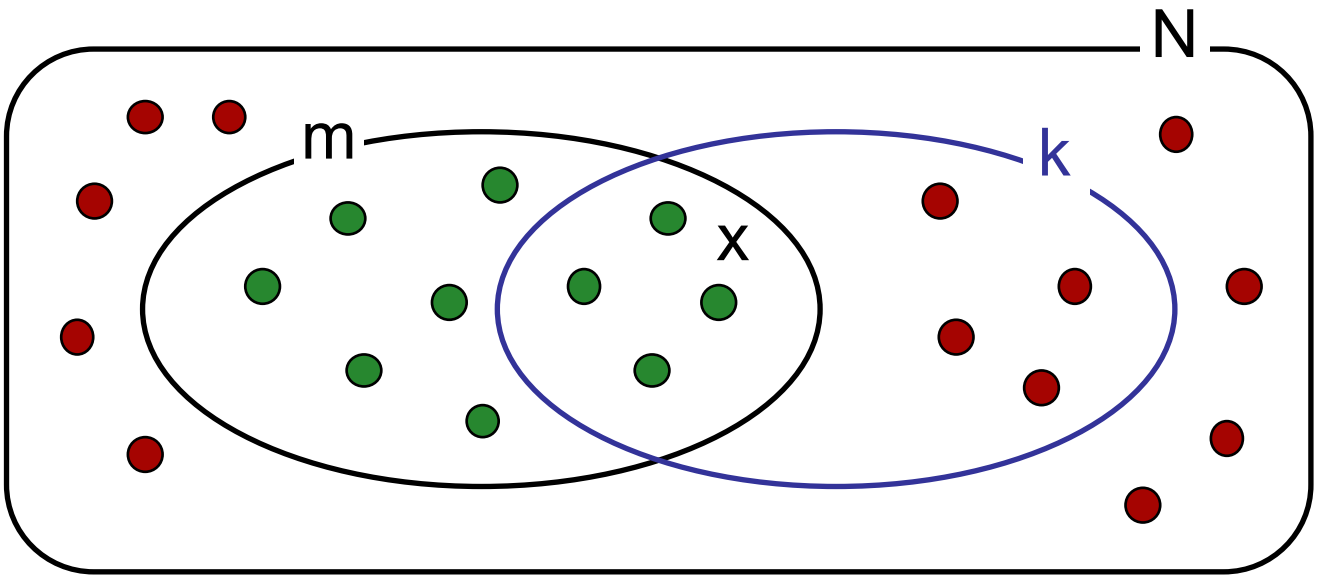
Hypergeometric distribution



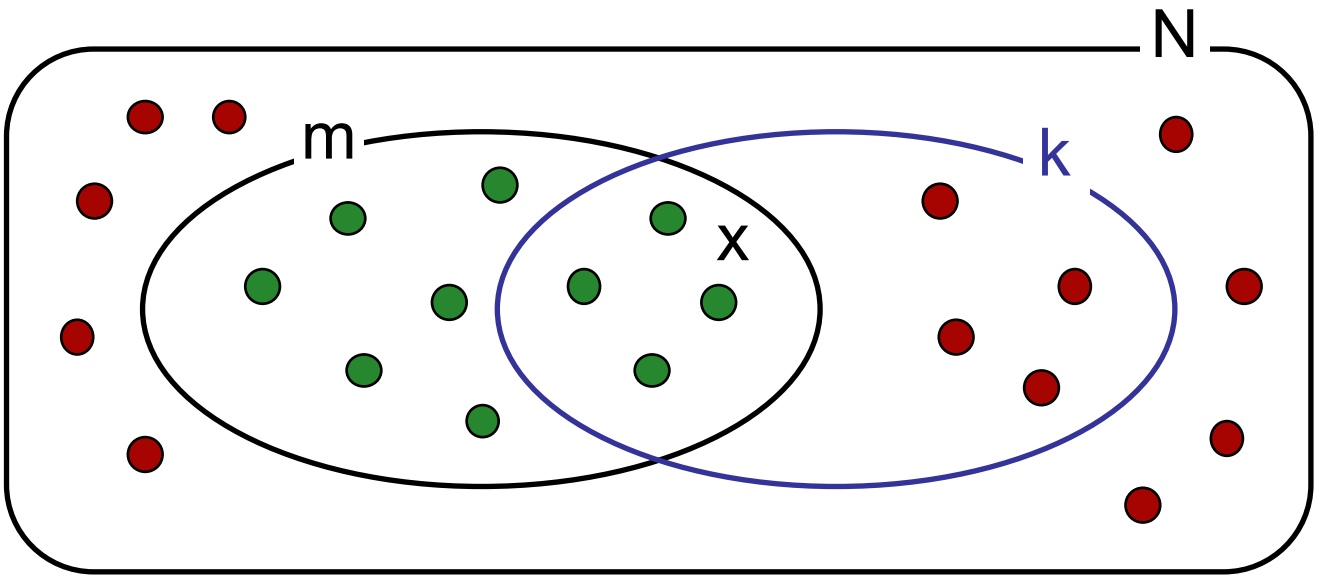
N = total number of elements

m = number of marked elements

k = number of sampled elements

x = number of marked sampled elements

Hypergeometric distribution



What is the probability to observe exactly x marked elements in the sample?

$$P(x \mid N,m,k) = \frac{\binom{m}{x}\binom{N-m}{k-x}}{\binom{N}{k}}$$

where

$$\binom{a}{b} = \frac{a!}{b!(a-b)!}$$

Hypergeometric distribution

What is the probability to observe at least x marked elements in the sample?

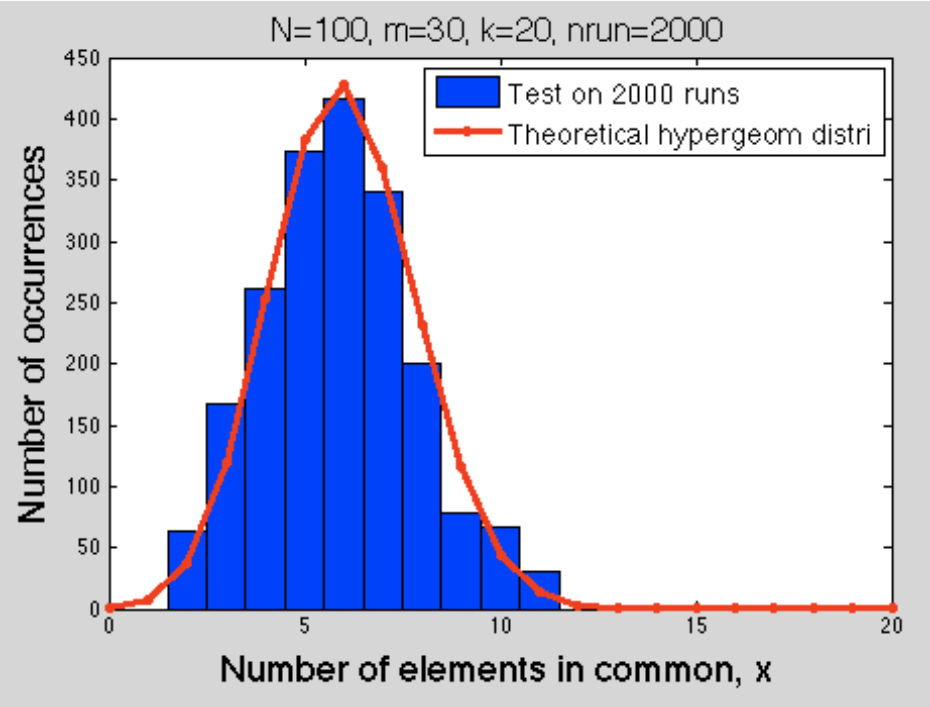
$$P(at\ least\ x \mid N,m,k) = \sum_{i=x}^{\min(k,m)} P(i \mid N,k,m) = 1 - \sum_{i=0}^{x-1} P(i \mid N,k,m)$$

Inverse cumulative hypergeometric distribution

Hypergeometric distribution

Probability density function

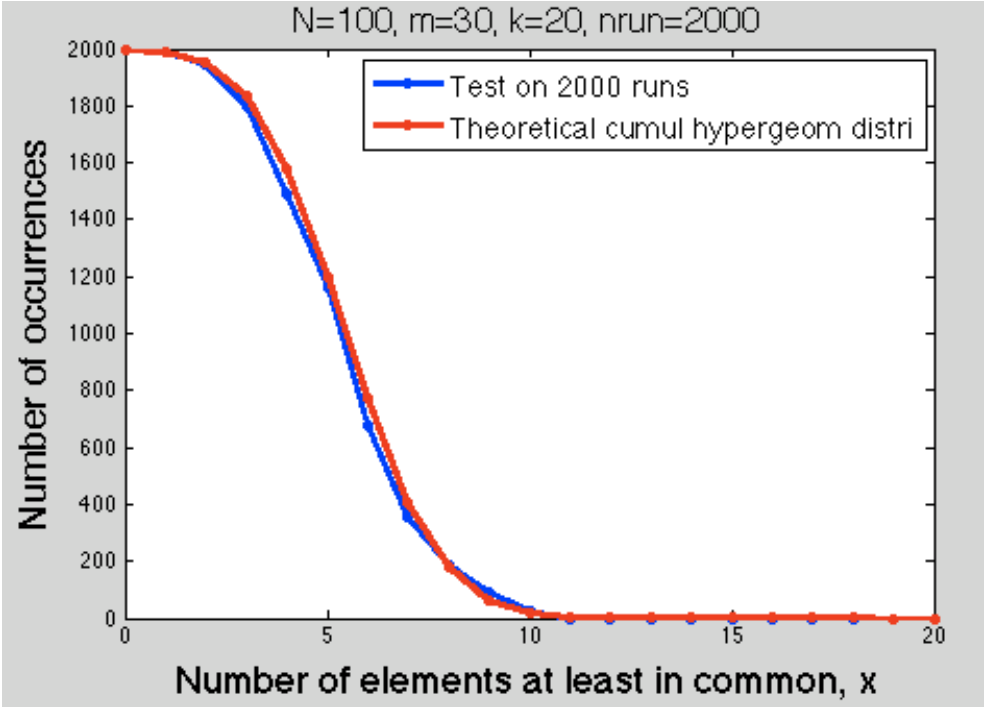
$$P(x \mid N,m,k)$$



Matlab function: hygepdf(x,N,m,k)

Inverse cumulative density function

$$P(at\ least\ x \mid N,m,k)$$



Matlab function: 1-hygecdf(x,N,m,k)

Hypergeometric distribution

Efficient computation of the hypergeometric distribution

By recursive computation, the hypergeometric formula can be computed in a much more efficient way than by computing explicitely the factorial at the numerator and denominator.

Initialization

$$P(X=0) = \frac{\binom{m}{0}\binom{n}{k-0}}{\binom{m+n}{k}}$$

$$= \frac{m!}{m!0!} \cdot \frac{n!}{k!(n-k)!} \cdot \frac{k!(m+n-k)!}{(m+n)!}$$

$$= \frac{n!}{(n-k)!} \cdot \frac{(m+n-k)!}{(m+n)!}$$

$$= \frac{n(n-1)\dots(n-k+1)}{(m+n)(m+n-1)\dots(m+n-k+1)}$$

NB: *n=N-m*

z = log(P) = log (n) + log (n-1) + ...+ log (n-k+1) - log (m+n) - ... - log (m+n-k+1)
=> P = e^z

Recursion

$$P(X=x) = P(X=x-1) \cdot \frac{(m-x+1)(k-x+1)}{x(n-k+x)}$$

The demonstration is given in the book by van Helden (see Appendix)

Source: J. van Helden

Hypergeometric distribution

Link between the hypergeometric and the binomial distribution

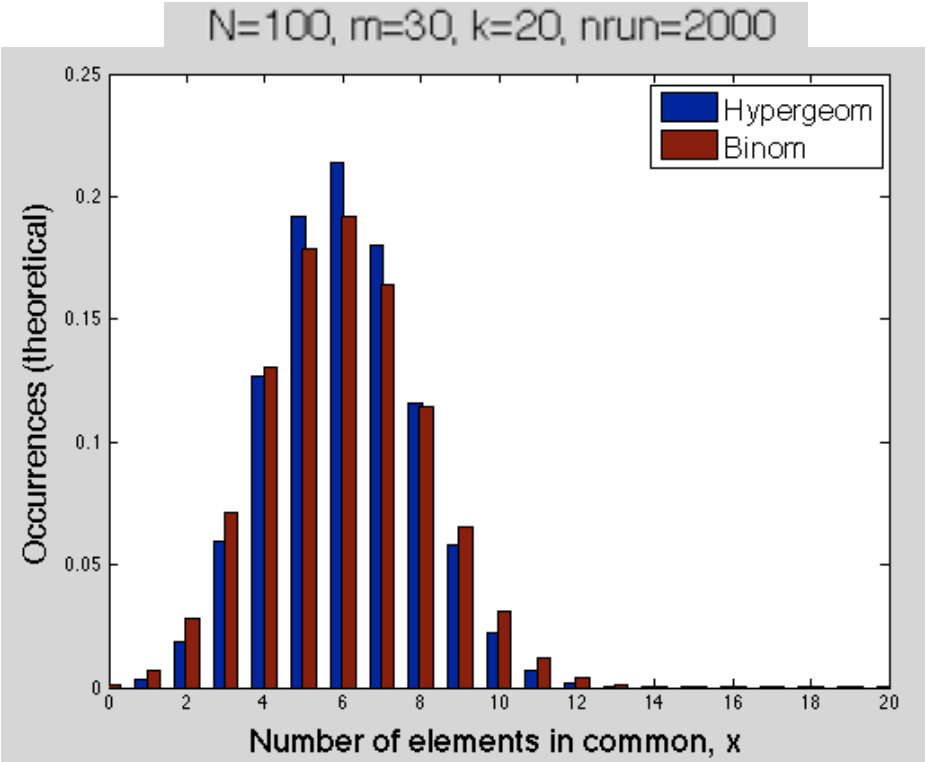
It can be demonstrated that when the size of the population ($N=m+n$) tends towards infinity whilst the proportions of labelled (m) and unlabelled (n) elements remains constant, the hypergeometric tends towards a binomial with probability $p = m/(m+n)$. This seems rather intuitive: in an infinite population, a sampling without replacement does not affect the relative proportions of remaining labelled and unlabelled elements. In practice, the binomial can be used to approximate the hypergeometric only if both m and n are sufficiently large.

$$P_h(X = x) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$$

$$\approx P_b(X = x) = \binom{k}{x}\hat{p}^x(1-\hat{p})^{k-x}$$

$$= \binom{k}{x}\left(\frac{m}{m+n}\right)^x\left(\frac{n}{m+n}\right)^{k-x}$$

The demonstration is given in the book by van Helden (see Appendix)

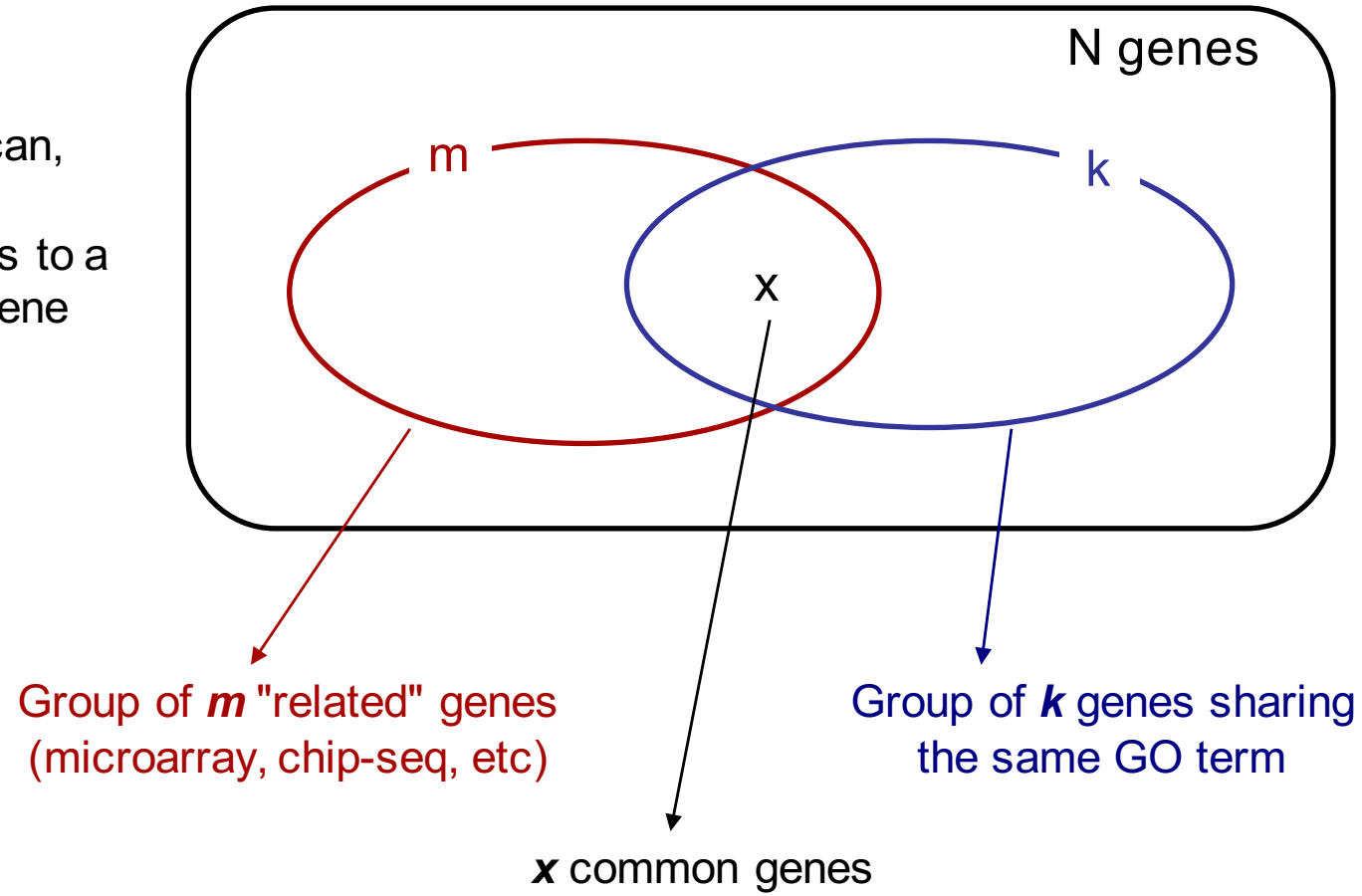


Source: J. van Helden

Hypergeometric distribution

Application

The hypergeometric test can, for example, be applied to associate a group of genes to a molecular function (e.g. Gene Ontology).



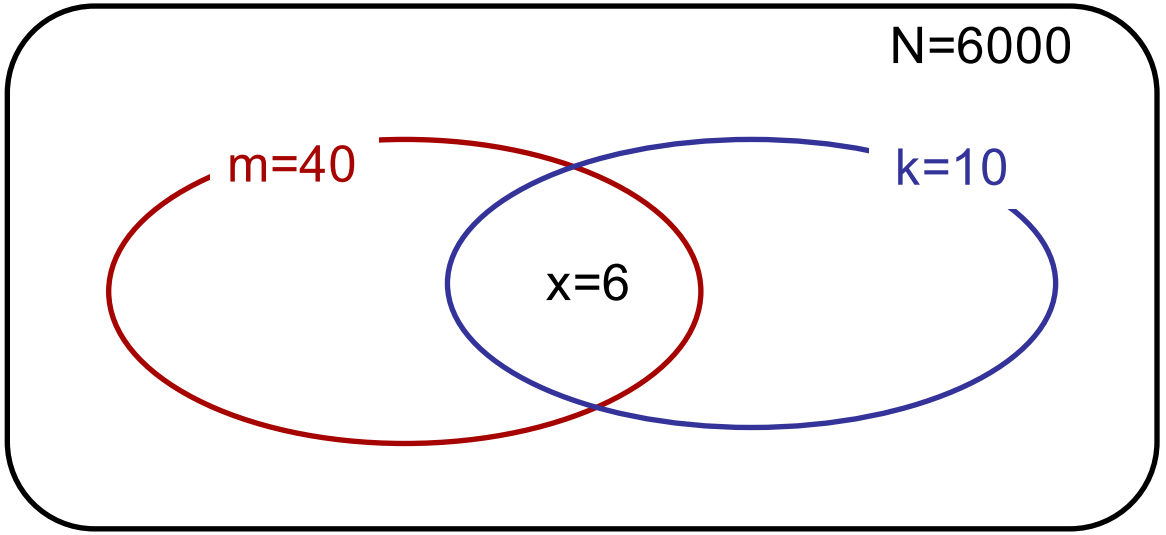
Hypergeometric distribution

Example

Let's consider the 6000 genes of yeast. Among those genes we will assume that 40 genes are known to be involved in the methionine synthesis.

Now, let's assume that an hypothetical microarray data analysis led to a cluster containing 10 genes; among which 6 belong to the methionine family.

Is such overlap significant? In other words, can we consider that the cluster can be considered as methionine-related?



The probability to have at least 6 genes in common is

$$P(at\ least\ 6 \mid 6000,40,10) = 1 - \sum_{i=0}^{5} P(i \mid 6000,40,10) =$$

This probability, called the *p-value*, is the probability to find at least 6 methionine genes in a random selection of 10 genes. Smaller the *p-value*, higher the significance of the results.