Estimating interactome size: the maximum likelihood approach

When two simple random samples are drawn, with replacement, from the same population, the number of objects selected in both samples, *k*, is distributed according to the hypergeometric distribution:

$$X \sim \text{Hypergeometric}(n_1,n_2,N);\ P(X=k \mid n_1, n_2, N) = \frac{\begin{pmatrix} n_1 \\ k \end{pmatrix}\begin{pmatrix} N-n_1 \\ n_2-k \end{pmatrix}}{\begin{pmatrix} N \\ n_2 \end{pmatrix}}; \qquad (1)$$

where n_1 and n_2 are sample sizes and *N* is size of the population from which the samples were drawn. Computation on discrete distributions is often difficult; however, we are greatly aided by the observation that, for large populations, the hypergeometric distribution is well-approximated by the binomial distribution:

$$X \sim \text{Binomial}(n_1,p),\text{ where } p = \frac{n_2}{N}; \qquad (2)$$

$$P(X=k \mid n_1, p) = \begin{pmatrix} n_1 \\ k \end{pmatrix} p^k (1-p)^{n_1-k} \qquad (3)$$

From known values of *k* and n_1, we calculate the maximum likelihood estimate of *p*,

$$\hat{p} = \frac{k}{n_1} \qquad (4)$$

Substituting equation 2 and solving for *N*, we get a maximum likelihood estimate of *N*,

$$\hat{N} = \frac{n_2}{\hat{p}} = \frac{n_1 n_2}{k} \qquad (5)$$

The sampling distribution of $\hat{p}$, for large sample sizes, is approximately normal with variance given by:

$$\sigma^2 = \frac{p(1-p)}{n} \qquad (6)$$

With this, we can calculate a confidence interval around our estimate of *p*,

$$\hat{p} \pm Z_{\alpha/2}\sigma,\text{ where } \alpha = (100 - \text{CI\%});\text{ e.g., for a 95\% CI, } Z = 1.96. \qquad (7)$$

Substituting the confidence interval into equation (5) yields the corresponding confidence interval for the population estimate.

These calculations assume error-free data: the presence of assay false positives artificially inflates the sample sizes and leads to overestimation of the interactome size. In order to correct the dataset, the sample size is simply multiplied by one minus the false positive rate, yielding:

$$\hat{N} = \frac{n_1(1 - fpr_1) \times n_2(1 - fpr_2)}{k} \qquad (8)$$

The method used to determine false positive rates, as described by D'haeseleer and Church [32], is essentially the same as the maximum likelihood approach described above. The observation, shown in Figure 3a, that the ratio of regions is conserved is mathematically identical to estimating the 'population' (of true positives in the entire dataset) by maximum likelihood, although it is not explicitly described as such in [32]. Confidence intervals (95%) for false positive rate estimates are generally less than +/-5%, and for the large Gavin *et al.* [27] and Krogan *et al.* [28] sets are less than +/-2%.