Mean and variance of Poisson-binomial distribution (PBD)

The PBD is given by:

$$P_{PBD}(X = i) = \sum_{k=1}^{\binom{N}{i}} \left(\prod_{\{m:m\in A_k\}}^{i} p_m * \prod_{\{n:n\in \overline{A_k}\}}^{N-i} (1 - p_n) \right) \quad (1)$$

Here, p_m and p_n indicate the probability of success (i.e. the co-occurrence probability of genes obtained by multiplying their occurrence probabilities in the respective lists) in the mth and nth list with m <> n. A_k denotes the k-th set of indices of the i lists where genes are co-occurring. There are $\binom{N}{i}$ possible sets and summation is carried out accordingly. $\overline{A_k}$ denotes the set of indices of N-i lists where genes are not co-occurring.

The Mean μ of PBD is calculated by multiplying the number of successes by their respective probabilities and summation over all possible successes from 1 to N where N is the number of lists. If the co-occurrence probability in the *i-th* list is designated by p_i, this calculation evaluates to:

$$\mu = \sum_{i=1}^{N} p_i \quad (2)$$

Thus, the expected number of co-occurrences is equal to the sum of co-occurrence probabilities over all lists. To prove (2), we multiply out (1) and rearrange all terms such that all product terms of the same length are shown in the same column (Fig. Proof_S1).

Fig. Proof_S1

Calculating the mean of PBD

Length of product terms →

Number of successes ↓

		1	2	3		N-1	N	
E(X)=	$1*($	$(-1)^{1+1}\binom{1}{1}\sum_{i=1}^{\binom{N}{1}} p_i$	$+(-1)^{2+1}\binom{2}{1}\sum_{i=1}^{\binom{N}{2}}\prod_{\{j:j\in A2_i\}}^{2} p_j$	$+(-1)^{3+1}\binom{3}{1}\sum_{i=1}^{\binom{N}{3}}\prod_{\{j:j\in A3_i\}}^{3} p_j$	$\ldots$	$+(-1)^{(N-1)+1}\binom{N-1}{1}\sum_{i=1}^{\binom{N}{N-1}}\prod_{\{j:j\in A(N-1)_i\}}^{N-1} p_j$	$+(-1)^{N+1}\binom{N}{1}p_{i1}p_{i2}\cdots p_{i(N-1)}p_{iN}$	$)+$
	$2*($		$+(-1)^{2+2}\binom{2}{2}\sum_{i=1}^{\binom{N}{2}}\prod_{\{j:j\in A2_i\}}^{2} p_j$	$+(-1)^{3+2}\binom{3}{2}\sum_{i=1}^{\binom{N}{3}}\prod_{\{j:j\in A3_i\}}^{3} p_j$	$\ldots$	$+(-1)^{(N-1)+2}\binom{N-1}{2}\sum_{i=1}^{\binom{N}{N-1}}\prod_{\{j:j\in A(N-1)_i\}}^{N-1} p_j$	$+(-1)^{N+2}\binom{N}{2}p_{i1}p_{i2}\cdots p_{i(N-1)}p_{iN}$	$)+$
	$3*($			$+(-1)^{3+3}\binom{3}{3}\sum_{i=1}^{\binom{N}{3}}\prod_{\{j:j\in A3_i\}}^{3} p_j$	$\ldots$	$+(-1)^{(N-1)+3}\binom{N-1}{3}\sum_{i=1}^{\binom{N}{N-1}}\prod_{\{j:j\in A(N-1)_i\}}^{N-1} p_j$	$+(-1)^{N+3}\binom{N}{3}p_{i1}p_{i2}\cdots p_{i(N-1)}p_{iN}$	$)+$
		$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$(N-1)*($				$\ldots$	$+(-1)^{(N-1)+(N-1)}\binom{N-1}{N-1}\sum_{i=1}^{\binom{N}{N-1}}\prod_{\{j:j\in A(N-1)_i\}}^{N-1} p_j$	$+(-1)^{N+(N-1)}\binom{N}{N-1}p_{i1}p_{i2}\cdots p_{i(N-1)}p_{iN}$	$)+$
	$N*($				$\ldots$		$+(-1)^{N+N}\binom{N}{N}p_{i1}p_{i2}\cdots p_{i(N-1)}p_{iN}$	$)$

Because of

$$\sum_{k=0}^{n}(-1)^{k}k\binom{n}{k}=0 \quad (3)$$

all terms in the same column cancel out except for the term in column 1 which is equal to (2). If co-occurrence probabilities are the same for all lists, (2) becomes:

$$\mu = N * p \quad (4)$$

as expected for the Binomial Distribution.

Similarly, the variance is calculated by multiplying the squared deviation of successes from the mean (#successes - μ)^^{2} by their respective probabilities and summing over all successes from 0 to N. This calculation evaluates to:

$$\sigma^2 = \sum_{i=1}^{N} p_i - \sum_{i=1}^{N} p_i^2 \quad (5)$$

The proof of (5) is performed exactly as the proof of (2). Thus, the variance is equal to the mean minus the sum of squared co-occurrence probabilities.

For all p_i being equal, we obtain:

$$\sigma^2 = N * p - N * p^2 \quad (6),$$

which is the variance of the Binomial Distribution.