- Definition: expected number of n-grams that occur r times: $E_N(N_r)$

- We have s different n-grams in corpus
  - let us call them $\alpha_1, ..., \alpha_s$
  - each occurs with probability $p_1, ..., p_s$, respectively

- Given the previous formulae, we can compute

$$\begin{aligned} E_N(N_r) &= \sum_{i=1}^{s} p(c(\alpha_i) = r) \\ &= \sum_{i=1}^{s} \binom{N}{r} p_i^r (1 - p_i)^{N-r} \end{aligned}$$

- Note again: p_i is unknown, we cannot actually compute this