Estimating the number of categories of a multinomial distribution

Péter Kómár

October 24, 2014

Contents

1 The problem

1.1 Motivation

In a biological experiment, the complete library of a large number ($M_{\max}$) of possible different DNA sequences is known. A selection step eliminates many of the DNA sequences, and we end up with a much smaller library of the original DNA sequences. Now, we take this decimated library, and sample is N times. Each of sample contains only one type of DNA sequence, and we determine their sequence. As a result, we find $K(< N)$ different DNA sequences out of the total of N.

Considering this information, what is the probability that the decimated library has exactly M different DNA sequences?

1.2 Formalization

Let us label the DNA sequences in the complete library with positive integers:

$$\Sigma_{\max} = \{1, 2, \dots M_{\max}\}.$$

The decimated library is a subset,

$$\Sigma \subset \Sigma_{\max}.$$

Let us define the histogram of Σ as

$$S = (l_1, l_2, \dots l_{M_{\max}}), \qquad \text{where} \quad l_k \in \{0, 1\},$$

where $l_k = 1$ indicated that $k \in \Sigma$, and $l_k = 0$ that $k \notin \Sigma$.

Sampling the decimated library N times provides N outcomes,

$$\Delta = \{x_1, x_2, \dots x_N\}, \qquad \text{where} \quad x_j \in \Sigma,$$

the histogram of which can also be defined as

$$D = (n_1, n_2, \dots n_{M_{\max}}), \qquad \text{where} \quad n_k \in \{0, 1, 2 \dots\},$$

where n_k counts how many times outcome k appears in Δ. Since we have a total of N outcomes, out of which we found only K different,

$$\sum_{k=1}^{M_{\max}} n_k = N, \qquad \text{and} \qquad \sum_{k=1}^{M_{\max}} \operatorname{sgn}(n_k) = K.$$

Now, the probability in question is

$$\mathcal{P}\Big(\sum_{k=1}^{M_{\max}} l_k = M \;\Big|\; D \Big) = \sum_{S:M} \mathcal{P}\big(S \;\big|\; D\big),$$

where $S : M$ indicates that we are summing over only such S histograms which contain exactly M number of 1's.

1.3 Example

Say, the complete library has $M_{\max} = 13$ different DNA sequences:

$$\Sigma_{\max} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13\}.$$

The decimated library can be any subset of this. Some examples are

$$\begin{aligned} \Sigma_1 &= \{2, 4, 6, 8, 10, 12\}, & S &= (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0), & M &= 6 \\ \Sigma_2 &= \{1, 13\}, & S &= (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1), & M &= 2 \\ \Sigma_3 &= \{6, 7, 10\}, & S &= (0, 0, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0), & M &= 3 \end{aligned}$$

where S and M are the corresponding histogram and number of total elements.

Now, if we measure $N = 5$ samples, we can get any repetitive combination of the numbers from 1 to 13. Some examples are

$$\begin{aligned} \Delta_1 &= \{6, 6, 6, 6, 6\}, & D &= (0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, 0), & K &= 1 \\ \Delta_2 &= \{4, 6, 8, 2, 10\}, & D &= (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 0), & K &= 5 \\ \Delta_3 &= \{2, 9, 5, 9, 9\}, & D &= (0, 1, 0, 0, 1, 0, 0, 0, 3, 0, 0, 0, 0), & K &= 3 \\ \Delta_4 &= \{1, 6, 1, 4, 7\}, & D &= (2, 0, 0, 1, 0, 1, 1, 0, 0, 0, 0, 0, 0), & K &= 4 \end{aligned}$$

If we measured Δ_1, we could rule out Σ_2 since it does not have sequence 6, but on this principle we could not decide between Σ_1 and Σ_3, which correspond to different M.

2 The solution

Solving the Bayesian estimation problem requires finding the sampling distribution $\mathcal{P}(D|S)$, the prior $\mathcal{P}_0(S)$, and using Bayes theorem to compute the posterior $\mathcal{P}(S|D)$.

2.1 Sampling distribution

For a given Σ (and histogram $S = (l_k)$), we can write the probability that x_j is picked as a uniform distribution over the elements of Σ,

$$\mathcal{P}(x_j|S) = \begin{cases} 1/M, & \text{if} \quad l_{x_j} = 1 \\ 0, & \text{if} \quad l_{x_j} = 0 \end{cases} \qquad \text{where} \qquad M = \sum_k l_k.$$

The probability of getting a particular collection of N outcomes, $\Delta = \{x_j \ : \ j = 1, 2, \ldots N\}$, is

$$\mathcal{P}(\Delta|S) \ = \ \prod_{j=1}^{N} \mathcal{P}(x_j|S) = \begin{cases} 1/M^N, & \text{if} \quad \forall j : \ l_{x_j} = 1 \\ 0, & \text{if} \quad \exists j : \ l_{x_j} = 0 \end{cases}$$

Since different ordering of the outcomes $\{x_j\}$ result in the same histogram $D = (n_k)$, we can write the probability of obtaining the histogram D as

$$\mathcal{P}(D|S) = \sum_{\Delta:D} \mathcal{P}(\Delta|S),$$

where $\Delta : D$ indicates summing over results Δ whose histogram is D.

Let us define the following relation between $S = (l_k)$ and $D = (n_k)$. We say that S is "consistent with" D (denoted by $S \parallel D$) if and only if,

$$S \parallel D \quad \Leftrightarrow \quad \Big[\forall k: \ (n_k > 0 \ \Rightarrow l_k = 1)\Big],$$

which means that S can explain the data D, i.e. it is not ruled out by D.

Now, if S is inconsistent with D ($S \nparallel D$), then $\mathcal{P}(D|S) = 0$, and if $S \parallel D$, then

$$\mathcal{P}(D|S) = \sum_{\Delta:D} \mathcal{P}(\Delta|S) = \sum_{\Delta:D} \frac{1}{M^N} = \frac{1}{M^N} \sum_{\Delta:D} 1,$$

where we used that M depends only on S, and N is the same for all Δ corresponding to the same D. The last term is the number of different datasets Δ that give the same D, this is

$$\sum_{\Delta:D} 1 = \binom{N}{n_1}\binom{N-n_1}{n_2}\binom{N-n_1-n_2}{n_3}\dots\binom{N-\sum_{i=1}^{k-1} n_i}{n_k}\dots\binom{N-\sum_{i=1}^{M_{\max}-1} n_i}{n_{M_{\max}}} = \frac{N!}{\prod_{k=1}^{M_{\max}} n_k!}.$$

Finally, the sampling distribution can then be written as

$$\mathcal{P}(D|S) \ = \ \begin{cases} \frac{N!}{\prod_k n_k!}\frac{1}{M^N} & , \quad \text{if} \quad S \parallel D \\ 0 & , \quad \text{if} \quad S \nparallel D \end{cases} \tag{1}$$

which is a flat ($p_k = 1/M$, $\forall k$) multinomial distribution n_k variables which can be non-zero, according to S.

2.2 Prior distribution

Before inferring from the data (D), we need to quantify our knowledge about the possible models (S) in the absence of data. This requires extra assumptions about the problem. For simplicity, let us assume that the every DNA sequence in $\Sigma_{\max}$, survives the selection process, and gets into the decimated library Σ with some unknown probability p (same for all k), i.e.

$$\mathcal{P}(l_k|p) = p^{l_k}(1-p)^{1-l_k}.$$

This assumption results in the following distributions for S and M,

$$\begin{aligned} \mathcal{P}(S|p) \ &= \ \prod_k \mathcal{P}(l_k|p) = \prod_k p^{l_k}(1-p)^{1-l_k} = p^M(1-p)^{M_{\max}-M}, \qquad \text{where} \quad M = \sum_k l_k. \\ \mathcal{P}(M|p) \ &= \ \sum_{S:M} \mathcal{P}(S|p) = \binom{M_{\max}}{M} p^M (1-p)^{M_{\max}-M}. \end{aligned}$$

Unfortunately, we don't know the value of p. (If we did, we could estimate M by $pM_{\max}$ even without the data.) To describe the vagueness of our knowledge about p, we introduce a prior for p in the form of a constant distribution over the range $[0, 1]$,

$$\mathcal{P}_0(p) = 1, \quad \text{if } 0 \le p \le 1.$$

This choice results in the following prior for S, and M

$$\mathcal{P}_0(S) \ = \ \int_0^1 dp\, \mathcal{P}(S|p)\, \mathcal{P}_0(p) = \int_0^1 dp\, p^M (1-p)^{M_{\max}-M} = \frac{M!(M_{\max}-M)!}{(M_{\max}+1)!} = \frac{1}{M_{\max}+1}\binom{M_{\max}}{M}^{-1} \tag{2}$$

$$\mathcal{P}_0(M) \ = \ \sum_{S:M} \mathcal{P}_0(S) = \binom{M_{\max}}{M} \frac{M!(M_{\max}-M)!}{(M_{\max}+1)!} = \frac{1}{M_{\max}+1}. \tag{3}$$

This is a conveniently simple result, it shows that if $p \in [0, 1]$ is uniformly distributed, then $M \in \{0, 1, 2 \dots M_{\max}\}$ is also uniformly distributed.

2.3 Posterior distribution

From $\mathcal{P}_0(S)$ and $\mathcal{P}(D|S)$, we can use Bayes' theorem to write

$$\mathcal{P}(S|D) \;\propto\; \mathcal{P}(D|S)\mathcal{P}_0(S) \quad \Rightarrow \quad \mathcal{P}(S|D) = \begin{cases} \frac{1}{A}\frac{1}{M^N}\binom{M_{\max}}{M}^{-1} & , \quad \text{if} \quad S \parallel D \\ 0 & , \quad \text{if} \quad S \nparallel D \end{cases} \tag{4}$$

The proportionality constant, A, can be determined from normalization,

$$1 = \sum_S \mathcal{P}(S|D), \qquad \text{where} \quad S \in \{0,1\}^{\times M_{\max}}.$$

To find the posterior probability that the decimated library has size M, we have to sum over all S which have $\sum_k l_k = M$.

$$\mathcal{P}(M|D) = \sum_{S:M} \mathcal{P}(S|D) = \frac{1}{A}\frac{1}{M^N}\binom{M_{\max}}{M}^{-1} \underbrace{\sum_{S:M,\parallel D} 1}_{f},$$

where $S : M, \parallel D$ indicates that the summation runs through all S histograms which have M number of 1's and are consistent with D. To find the number of such histograms, f, let us consider the following:

- If $M < K$ (where $K = \sum_k \mathrm{sgn}(n_k)$), then S has fewer non-zero elements than D and therefore $S \nparallel D$ for all S, thus $f = 0$.
- If $M = K$, then only one S is consistent with D (the one which has $l_k = \mathrm{sgn}(n_k)$, $\forall k$), thus $f = 1$.
- If $M > K$, then any compatible S should contain 1's at the K non-zero positions of D, and additional $M - K$ 1's distributed among the remaining $M_{\max} - K$ positions. The number of such consistent S histograms is $f = \binom{M_{\max}-K}{M-K}$.

The resulting distribution is

$$\mathcal{P}(M|D) = \frac{1}{A}\frac{1}{M^N}\binom{M_{\max}-K}{M-K}\binom{M_{\max}}{M}^{-1} = \frac{1}{A'}\frac{1}{M^N}\frac{M!}{(M-K)!}, \qquad \text{for} \quad M \geq K, \tag{5}$$

and $\mathcal{P}(M|D) = 0$, if $M < K$. Here A' is a different normalization constant as before, we can determine its value from the normalization condition for M,

$$1 = \sum_{M=K}^{M_{\max}} \mathcal{P}(M|D).$$

Since Eq.(5) is independent of $M_{\max}$, we can perform the limit $M_{\max} \to \infty$, and the posterior remains well defined (normalizable) as long as $N > K + 1$. This can be seen by approximating the tail of the posterior using $\log(n) \approx n \log n - n$, and $\log(n-k) \approx \log n - \frac{k}{n}$ for $n \gg k > 1$ with

$$\mathcal{P}(M|D) \sim \exp\left[-N\log M + K\log M - \frac{K^2}{M}\right] = \frac{e^{-K^2/M}}{M^{N-K}} \sim \frac{1}{M^{N-K}}, \qquad \text{for} \quad M \gg K,$$

which is normalizable over the entire domain $\{M \in \mathbb{Z} \;:\; M \geq K\}$ if and only if $N > K + 1$.

2.4 Most likely estimate

Given N and K, we can use the formula for the posterior to find the maximum posterior estimate, i.e. the mode of $\mathcal{P}(M|D)$, M^*. Using Stirling's approximation, we can write the logarithm of the posterior as

$$\log \mathcal{P}(M|D) \approx -\log A' - N \log M + M \log M - M - (M-K)\log(M-K) + (M-K).$$

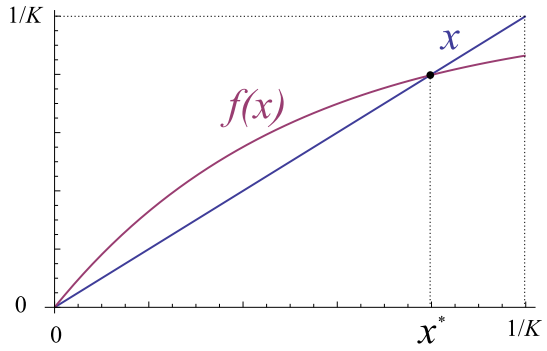
Equating the first derivative with respect to M with zero gives

$$0 = \frac{\partial}{\partial M} \log \mathcal{P}(M|D) \approx -\frac{N}{M} + \log M - \log(M-K),$$

which we can write as a transcendental equation for $x = 1/M$,

$$x = \frac{1}{K}\left(1 - e^{-Nx}\right) =: f(x),$$

which has a single solution $x^* \in (0, 1/K)$ for $N > K$, as shown on the figure below.



Since the map $x \mapsto f(x)$ has a single stable fix point, x^*, in the region $x \in (0, 1/K)$, we can obtain the solution by iterating f, $x_{n+1} = f(x_n)$,

$$x^* = \lim_{n\to\infty} x_n = \lim_{n\to\infty} f^{[n]}(x_0),$$

we can approximate the solution by choosing x_0 close enough to the solution, and iterating f only a few times (say, once). We use this approximation in two limiting cases

1. If $x^* \approx 1/K$, which happens when $N \gg K$, then we can start from $x_0 = 1/K$, and have

$$x^* \approx x_1 = f(x_0) = \frac{1}{K}\left(1 - e^{-N/K}\right) \qquad \Rightarrow \qquad M^* \approx \frac{K}{1 - e^{-N/K}} \approx K + Ke^{-N/K}. \tag{6}$$

   We can rely on this approximation as long as $M^* - K < 1$, which require $N > K\log(K+1)$.

2. If $x^* \approx 0$, which happens when $N - K \ll K$, then we can start from a small x_0 value. To find a good x_0, let's write $f(x)$ as its Taylor series around 0 up to 2nd order, and solve the quadratic equation,

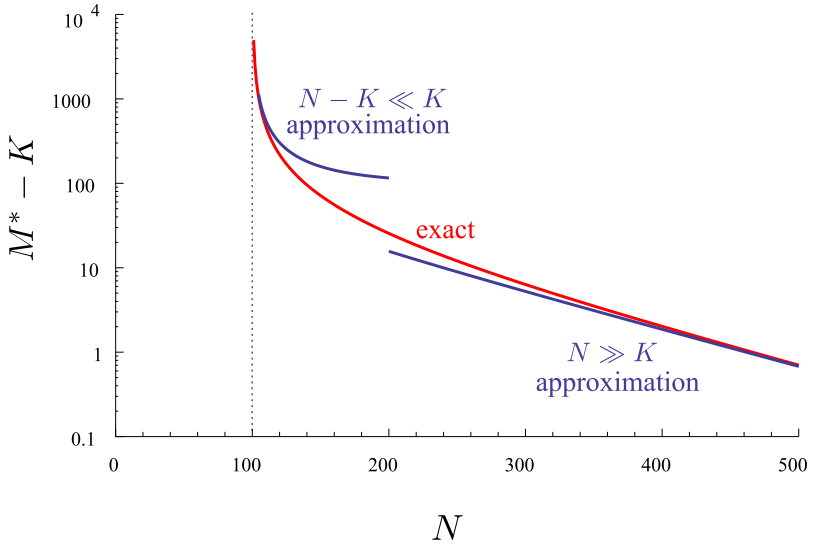
$$x = f(x) \approx \frac{N}{K}x - \frac{N^2}{2K}x^2 \qquad \to \qquad x_0 = \frac{2(N-K)}{N^2}.$$

   Then

$$x^* \approx x_1 = f(x_0) = \frac{1}{K}\left(1 - \exp\left[-\frac{2(N-K)}{N}\right]\right) \qquad \Rightarrow \qquad M^* \approx \frac{K}{1 - e^{-\frac{2(N-K)}{N}}} \approx \frac{K^2}{2(N-K)}. \tag{7}$$

   We can rely on this approximation as long as $M^* \gg K$, which is satisfied for $N - K < K/4$.

The figure below shows the exact $M^* - K$ as a function of N for $K = 100$ alongside with the approximations for small and large N.



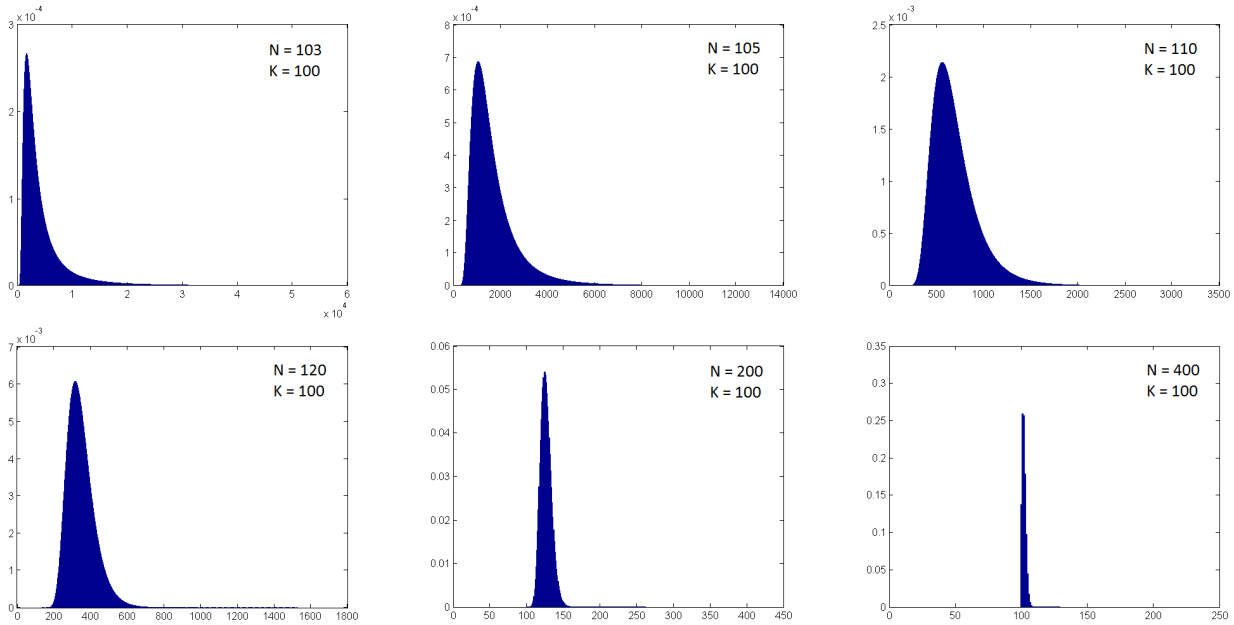
The approximate results are

$$\text{small } N: \qquad M^* \approx K\left[1 - \exp\left(-\frac{2(N-K)}{N}\right)\right]^{-1}, \qquad \text{if} \quad N - K \ll K$$

$$\text{large } N: \qquad M^* \approx K\left[1 - \exp\left(-\frac{N}{K}\right)\right]^{-1}, \qquad \text{if} \quad N \gg K$$

2.5 Examples

The figures below show the posterior distribution $\mathcal{P}(M|D)$ for $K = 100$ and $N = 103$, 105, 110, 120, 200, 400. The range of plausible M values increases dramatically as N gets close to K.



3 Numerical implementation

3.1 Motivation

Although the posterior $\mathcal{P}(M|N,K)$ is given in closed form in Eq. (5) (up to the normalization constant), using it directly has two disadvantages:

1. The straightforward method to calculate the normalization constant,

$$A' = \sum_{M=K}^{M_{\max}} \frac{1}{M^N} \frac{M!}{(M-K)!},$$

can fail for large N and K, because the evaluation of any single term is prone to numerical underflow (or overflow), due to the largeness of M^N, $M!$ and $(M-K)!$ on their own and the overall smallness of the term.

2. In the limit of $M_{\max} \to \infty$, the straightforward numerical evaluation of the expectation value and the variance,

$$\langle M \rangle = \sum_M M\, \mathcal{P}(M|N,K) \quad , \qquad \mathrm{Var}(M) = \sum_M \Big(M - \langle M \rangle\Big)^2 \mathcal{P}(M|N,K),$$

can converge slowly to the real value due to the significant weight the posterior has in the tail, especially in the case of $N \approx K$.

The aim is, therefore, to avoid underflow, and efficiently estimate the expectation value and variance contribution of the tail of the posterior.

3.2 Avoiding underflow (overflow)

Underflow can easily happen in the calculation of $\mathcal{P}(M|N,K)$, since the required normalization constant

$$A' = \sum_{M=K}^{M_{\max}} \frac{1}{M^N} \frac{M!}{(M-K)!} \approx \sum_{M=K}^{\infty} \frac{1}{M^{N-K}} \approx \int_K^{\infty} dM\, \frac{1}{M^{N-K}} = \frac{1}{(N-K)K^{N-K-1}} \sim \frac{1}{K^{N-K}},$$

which can be a small number if K and N are large. To avoid this we can recast the posterior in the following form.

$$\mathcal{P}(M|N,K) \;=\; \frac{1}{B} \frac{K^{N-K}}{M^N} \frac{M!}{(M-K)!}$$

where B is a new normalization constant, whose order of magnitude depends only weakly on N and K.

To avoid multiplying numbers with large but opposite orders of magnitude, we can evaluate the logarithms of the terms and exponentiate only at the end.

$$\mathcal{P}(M|N,K)B \;=\; \exp\Big[(N-K)\log K - N \log M + \log\Gamma(M+1) - \log\Gamma(M-K+1)\Big],$$

where $\log\Gamma(x)$ needs to be evaluated without calculating $\Gamma(x)$. This is achieved by Matlab's `gammaln` function.

3.3 Dealing with the heavy tail

Due to memory and time constraints a numerical representation of the $M_{\max} \to \infty$ case will still have a maximal M value, $\tilde{M}_{\max}$. The contribution from the region $[1, \tilde{M}_{\max}]$ to the rth moment of the posterior can be evaluated exactly, but there will always be a contribution from the tail, $[\tilde{M}_{\max} + 1, \infty)$.

$$\langle M^r \rangle = \underbrace{\sum_{M=K}^{\tilde{M}_{\max}} M^r \mathcal{P}(M|N,K)}_{b_r \text{ (exact)}} + \underbrace{\sum_{M=\tilde{M}_{\max}}^{\infty} M^r \mathcal{P}(M|N,K)}_{t_r \text{ (to be approximated)}}$$

For a large enough $\tilde{M}_{\max}$, we can approximate the posterior with $\frac{1}{B}\frac{1}{M^{N-K}}$, and therefore evaluate tail contribution of the rth moment as

$$t_r \approx \frac{K^{N-K}}{B} \sum_{M=\tilde{M}_{\max}}^{\infty} \frac{1}{M^{N-K-r}} = \frac{K^{N-K}}{B} \zeta_{N-K-r}(\tilde{M}_{\max} + 1),$$

where $\zeta_x(n) = \sum_{M=n}^{\infty} M^{-x}$. To avoid under/overflow, we write this as

$$\begin{aligned} t_r B &\approx \exp\Big[(N-K)\log K + \log \zeta_{N-K-r}(\tilde{M}_{\max}+1)\Big] \\ &\approx \exp\left[(N-K)\log K - (N-K-1-r)\log \tilde{M}_{\max} - \log(N-K-1-r) - \frac{N-K-1-r}{2\tilde{M}_{\max}}\right], \end{aligned}$$

where we approximated ζ with the first three terms of its asymptotic expansion. This expansion also shows that $\tilde{M}_{\max} \gg N - K$ is required for the approximation to work.

3.4 Matlab function

The Matlab function `EstimateFlatMultinomial` realizes the above functions. It can be found here: `https://github.com/peterkomar-hu/matlab-estimate-flat-multinomial.git`

4 Problems with the same solution

The motivation of the mathematical problem was described in terms of a biological experiment, but the same problem arises in the following problems as well:

- Taxis in a small town: After recording taxis going accross the main square of a town N times, only K different taxis were observed. What is number of active taxis in the town?
- Foreign alphabet: A small portion of a text written with an unknown alphabet contains N symbols, out of which there are K different. How big is the alphabet?
- Simple capture-recapture: While catching wild animals with a trap and releasing them, after N catches, only K different members of the population have been found. How big is the population within the range of the trap?
- Pseudo-random numbers: A pseudo-random number generating algorithm produces N numbers, out of which only K different can be found. What is the total number of possible number that the algorithm can generate?