Tail bounds for hypergeometric distribution (sampling without replacement)

October 19, 2021

The proof presented here is from [Chv79].

Suppose you have bag with $a = n(1-\epsilon)/2$ black balls and $b = n(1+\epsilon)/2$ white balls. We uniformly randomly select a subset of k balls (without replacement). We want to show that the probability that we will get at least $k/2$ black balls is quite small. The probability is given as follows

$$p = \binom{n}{k}^{-1} \sum_{j=k/2}^{k} \binom{a}{j}\binom{b}{k-j}.$$

Say, we have a real number $x \in (0,1)$. We can write

$$\begin{aligned}
p &= \binom{n}{k}^{-1} \sum_{j=k/2}^{k} \binom{a}{j}\binom{b}{k-j} \\
&\leq \binom{n}{k}^{-1} \sum_{j=k/2}^{k} \binom{a}{j}\binom{b}{k-j}(1-x)^{k/2-j} \\
&\leq \binom{n}{k}^{-1} \sum_{j=0}^{k} \binom{a}{j}\binom{b}{k-j}(1-x)^{k/2-j} \\
&= \binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{j=0}^{k} \binom{a}{j}\binom{b}{k-j}(1-x)^{k-j} \\
&= \binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{j=0}^{k} \binom{a}{j}\binom{b}{k-j}\sum_{i=0}^{k-j}\binom{k-j}{i}(-x)^{i}
\end{aligned}$$

Changing the order of the two summations.

$$\begin{aligned}
&\binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{j=0}^{k} \binom{a}{j}\binom{b}{k-j}\sum_{i=0}^{k-j}\binom{k-j}{i}(-x)^{i} \\
= &\binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^{i} \sum_{j=0}^{k-i} \binom{a}{j}\binom{b}{k-j}\binom{k-j}{i} \\
= &\binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^{i} \sum_{j=0}^{k-i} \binom{a}{j}\binom{b}{i}\binom{b-i}{k-j-i} \quad \left(\text{because } \binom{b}{k-j}\binom{k-j}{i} = \binom{b}{i}\binom{b-i}{k-j-i}\right) \\
= &\binom{n}{k}^{-1} (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^{i} \binom{b}{i}\binom{a+b-i}{k-i} \\
= &(1-x)^{-k/2} \sum_{i=0}^{k} (-x)^{i} \binom{b}{i}\binom{n-i}{k-i} \Big/ \binom{n}{k}
\end{aligned}$$

We will use $\binom{n}{k}\binom{k}{i} = \binom{n-i}{k-i}\binom{n}{i}$.

$$
\begin{aligned}
& (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^i \binom{b}{i} \binom{n-i}{k-i} / \binom{n}{k} \\
= \quad & (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^i \binom{b}{i} \binom{k}{i} / \binom{n}{i} \\
= \quad & (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^i \binom{k}{i} \frac{b(b-1)\cdots(b-i+1)}{n(n-1)\cdots(n-i+1)} \\
\leq \quad & (1-x)^{-k/2} \sum_{i=0}^{k} (-x)^i \binom{k}{i} (\frac{b}{n})^i \\
= \quad & (1-x)^{-k/2} (1-bx/n)^k
\end{aligned}
$$

To get the best bound, we will minimize $f(x) = (1-x)^{-1}(1-bx/n)^2$.

Putting $f'(x) = 0$, we get $x = 2 - n/b$. And $f(2-n/b) = 4(n-b)b/n^2 = 4ab/n^2$. The final probability bound we get is

$$p \leq (4ab/n^2)^{k/2}.$$

Put $a = n(1-\epsilon)/2$ and $b = n(1+\epsilon)/2$.

$$p \leq (1-\epsilon^2)^{k/2}.$$

Using $1 - \epsilon^2 \leq e^{-\epsilon^2}$ (standard calculus),

$$p \leq e^{-\epsilon^2 k/2}.$$

References

[Chv79] V. Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.