UDC 519.214

Central Limit Theorem of Turing's Formula

Molchanov S.[*†], **Zhang Zh.**[†], **Zheng L.**[‡]

[*] *Higher School of Economics, Moscow, Russia*
[†] *University of North Carolina at Charlotte, Charlotte, NC, USA*
[‡] *Tennessee Technological University, Cookeville, TN, USA*

Abstract. The paper concerns the generalized Maxwell-Boltzmann scheme: the distribution of n particles among infinitely many boxes b_k, $k \geq 1$, in a system. Let $Y_k(n)$, $k \geq 1$, be the occupation members of the boxes, $N_r = \sum_{k\geq 1} 1[Y_k = r]$, where $1[\cdot]$ is the indicator function and $r = 0, 1, \cdots, n$. The paper contains the central limit theorem for the joint distribution of N_1 and the Turing's "lost probability" statistic, $\pi_0 = \sum_{k\geq 1} p_k 1[Y_k = 0]$ under mild regularity conditions on the distribution $\{p_k; k \geq 1\}$.

Keywords: Maxwell-Boltzmann statistics, occupation numbers, missing probabilities, Turing's formula, central limit theorem.

1. Introduction

Alan Turing, along with Hödel and Cherch, is considered as one of the founders of the modern mathematical logic (Turing machine, recursive functions, etc.). But he also worked in cryptography, statistics, and information theory. The following model was proposed by Alan Turing during his famous activity in the decoding of German submarine enigmas in WWII. It is the generalization of the multivariant distribution or Maxwell–Boltzmann scheme.

2. Main Section

Let $\mathcal{X} = \{\ell_k, k \geq 1\}$ be a countable alphabet with letters $\ell_1, \ell_2, \cdots$. Consider a corresponding probability distribution $\{p_k; k \geq 1\}$ on $\mathcal{X}$, $p_k > 0$, $p_k \downarrow$, and $\sum_{k\geq 1} p_k = 1$. Let $\{Y_k; k \geq 1\}$ be the frequencies of the letters observed in an identically and independently distributed (*iid*) sample of size n under the distribution $\{p_k; k \geq 1\}$.

In the terminology of Maxwell-Boltzmann statistics, ℓ_k, $k \geq 1$, are the boxes, Y_k are the occupation numbers of the boxes in the case when n particles are allocated independently to the boxes one by one according to the distribution $\{p_k; k \geq 1\}$.

Let $N_r(n) = \sum_{k\geq 1} 1[Y_k = r]$, $r = 1, 2, \cdots, n$, *i.e.*, N_r is the number of letters repeated exactly r times in the sample. The *lost probability statistic* by Turing is given by the formula

$$\pi_0 = \sum_{k\geq 1} p_k 1[Y_k = 0],$$

the total probability associated with the unobserved letters or empty boxes.

Since Y_k, for each k, $k \geq 1$, has the binomial distribution with parameters (n, p_k), it follows that

$$\mathrm{E}(\pi_0) = \sum_{k\geq 1} p_k(1-p_k)^n$$

and

$$\mathrm{E}(N_1(n+1)) = \sum_{k\geq 1} \binom{n+1}{1} p_k(1-p_k)^n,$$

that is,

$$(n+1)\,\mathrm{E}(\pi_0(n)) = \mathrm{E}(N_1(n+1))$$

(a justification for Turing formula) and $N_1(n)$ (*i.e.*, the number of letters observed exactly once) is a good approximation for the lost probability $\pi_0(n)$ after normalization $\pi_0 \to n\pi_0$ (see additional information in the monograph [1]).

In the present talk we will discuss only random variables $\pi_0(n)$ and $N_1(n)$ as $n \to \infty$ and the corresponding central limit theorem after appropriate normalization.

Note that there are many cases when the random variable, say, $N_1(n)$ is not asymptotically normal. Consider the geometric distribution $p_k = pq^{k-1}$, $k \geq 1$, $0 < p < 1$, and $q = 1 - p$. Then

$$\begin{aligned}\mathrm{E}(N_1(n)) &= \sum_{k\geq 1} np_k(1-p_k)^{n-1}\\ &= \sum_{k\geq 1} npq^{k-1}(1-pq^{k-1})^{n-1}\\ &\sim \sum_{k\geq 1} npq^{k-1}e^{-(n-1)pq^{k-1}}.\end{aligned}$$

The asymptotics of the first moment has a form,

$$\mathrm{E}(N_1(n)) \sim \sum_{l=-\infty}^{+\infty} \frac{a(n)}{q^l} e^{-\frac{a(n)}{q^l}}(1+o(1)),$$

where

$$a(n) = p\exp\left(-\ln\frac{1}{q}\left[\frac{\ln(n-1)}{\ln(1/q)}\right]\right).$$

The important fact is not only that $N_1(n)$ is bounded in probability but also the slow oscillation of $\mathrm{E}(N_1(n))$ in the $\ln n$ scale as $n \to \infty$.

Even for slowly decreasing p_k the first two moments of $N_1(n)$ on $n\pi_0(n)$ can oscillate and be small for appropriate sub-sequences $\{n_k\}$. The proof of the central limit theorem requires some regularity assumptions.

Assume that sequence $\{p_k; k \geq 1\}$ has C^2 interpolation $p(x) \in [0, \infty)$ such that $p_k = p(k)$, $p(0) < \infty$, $p'(x) < 0$ for $x \geq 0$, *i.e.*, $\int_0^\infty p(k)dx \geq p(0) + \sum_{k \geq 1} p_k = p(0) + 1$. We consider three cases:

a) $p(x) = \frac{L(x)}{(1+x)^\alpha}$, where $\alpha > 1$ and $L(x)$ is a regularly slowly varying function on $[0, \infty)$ (it means that $L(x) \in C^1([0, \infty))$ and $L'(x) = L_1(x)/(1+x)$, where $L_1(x) = o(L(x))$.

b) $p(x) = \exp(-\ln(1+x)L(x))$, where $L(x) \to +\infty$ as $x \to \infty$ and satisfies the same condition as in a).

c) $p(x) = \exp(-x^\alpha L(x))$, where $0 < \alpha < 1$ and $L(x)$ is from the same class as in a) and b).

Theorem 1. *Consider the Turing vector $(n\pi_0(n), N_1(n))$. In all three regular situations a), b) and c) after normalization*

$$\left(\pi_0^*(n) = \frac{n\pi_0(n)}{\sqrt{\mathrm{Var}(n\pi_0(n))}}, \quad N_1^*(n) = \frac{N_1(n) - \mathrm{E}(N_1(n))}{\sqrt{\mathrm{Var}(n\pi_0(n))}}\right),$$

it follows that

$$(\pi_0^*(n), N_1^*(n)) \xrightarrow{law} N(0, B)$$

where B is a non-degenerated limiting covariance matrix (see below). The analysis of the moments of $n\pi_0(n)$ and $N_1(n)$ is a non-trivial analytic problem. Let's give first the integral representations for such moments:

$$\begin{aligned}
&\mathrm{E}(n\pi_0(n)) \sim \int_0^\infty np(x)e^{-np(x)}dx, \quad \text{as } n \to \infty \\
&\mathrm{E}(N_1(n)) = \mathrm{E}(n\pi_0(n)) + \mathcal{O}(1), \\
&\mathrm{Var}(n\pi_0(n)) \sim \int_0^\infty n^2p^2(x)\left(e^{-np(x)} - e^{-2np(x)}\right)dx, \\
&\mathrm{Var}(N_1(n)) \sim \int_0^\infty np(x)e^{-np(x)}dx - \int_0^\infty n^2p^2(x)e^{-2np(x)}dx, \\
&\mathrm{Cov}(n\pi_0(n), N_1(n)) = 0.
\end{aligned}$$

It is not difficult to understand that the main contribution to the moments is from the region where $np(x) = \mathcal{O}(1)$, *i.e.*, the occupation numbers Y_k, $k \geq 1$, have Poisson distributions. After changes of the variables and using the Laplace method, one can find the asymptotics for the moments. The formulas in the simplest cases are presented below.

I) $p(x) = \frac{L(x)}{(1+x)^\alpha}$, where $\alpha > 1$. In this case,

$$\begin{aligned}
\mathrm{E}(N_1(n+1)) = \mathrm{E}(n\pi_0(n)) &\sim C_1(\alpha) n^{1/\alpha} \tilde{L}(n), \\
\mathrm{Var}(N_1(n)) &\sim C_2(\alpha) n^{1/\alpha} \tilde{L}(n), \quad \text{and} \\
\mathrm{Var}(n\pi_0(n)) &\sim C_3(\alpha) n^{1/\alpha} \tilde{L}(n).
\end{aligned}$$

For the constants, $C_i(k)$, $i = 1, 2, 3$, and the slowly varying function $\tilde{L}(n)$ there are exact expressions.

II) $p(x) = \exp(-\ln^\beta x)$, where $\beta > 1$ (a particular case of a more general form presented above). In this case,

$$\begin{aligned}
\mathrm{E}(N_1(n)) = \mathrm{E}(n\pi_0(n)) &\sim C_1(\beta) \frac{e^{(\ln n)^{1/\beta}}}{(\ln n)^{1-1/\beta}}, \\
\mathrm{Var}(N_1(n)) &\sim C_2(\beta) \frac{e^{(\ln n)^{1/\beta}}}{(\ln n)^{1-1/\beta}}, \quad \text{and} \\
\mathrm{Var}(n\pi_0(n)) &\sim C_3(\beta) \frac{e^{(\ln n)^{1/\beta}}}{(\ln n)^{1-1/\beta}}.
\end{aligned}$$

III) $p(x) = \exp(-x^\alpha L(x))$, where $\alpha \in (0, 1)$. In this case, all first and second moments have (up to a constant factor depending only on α) the form $C_i(\alpha) \ln^{1/\alpha - 1} n \tilde{L}(n)$, where the slowly varying function $\tilde{L}$ can be expressed in terms of L and α.

3. Conclusions

The paper in progress will contain the central limit theorem for the system $\pi_0(n), N_1(n), \cdots, N_r(n)$ under the same classes of distribution $\{p_k; k \geq 1\}$ as presented above. Examples will be given to show that outside these classes there are cases for which the asymptotic normality does not hold. Statistical applications will also be considered.

References

1. *Zhang Z.* Statistical implications of Turing formula. — John Wiley & Sons, Hoboken, New Jersey, USA, 2017.