Snapshots of modern mathematics
from Oberwolfach

№ 8/2021

The Enigma behind the Good–Turing formula

Fadoua Balabdaoui • Yulia Kulagina

Finding the total number of species in a population based on a finite sample is a difficult but practically important problem. In this snapshot, we will attempt to shed light on how during World War II, two cryptanalysts, Irving J. Good and Alan M. Turing, discovered one of the most widely applied formulas in statistics. The formula estimates the probability of missing some of the species in a sample drawn from a heterogeneous population. We will provide some intuition behind the formula, show its wide range of applications, and give a few technical details.

1 The species richness problem

1.1 Motivation and example

Say we have a population comprising individuals drawn from K (possibly infinite) distinct species, among which a lot of species may be rare, and a few are very common. Our goal is to either estimate the frequencies of the species in the population based on the number of their occurrences in a finite sample, or simply to estimate the number of classes K in the population from the sample.[1]

[1] The term "population frequency" commonly refers to the true (unknown) proportion. It is usually contrasted with the word "empirical" that the reader will encounter later.

We use the word "species" in a broad sense. It may refer to flora and fauna, to types of errors in a software system, to celestial bodies in the universe, to word-types in a language, to connected components in a graph, and so on. Many more applications can be found in [1].

To make the problem clearer, let us look at an example. Imagine you are traveling through a rainforest and encounter 8 armadillos, 4 pumas, 4 jaguars, 2 tapirs, 1 capybara, and 1 squirrel monkey. How can you estimate the probability distribution of all the different kinds of animals you may come across during your whole trip?

1.2 Naïve solutions and why they fail

Your intuition would suggest using what is called an "empirical frequency estimator".[2] This would assign the probabilities $\frac{2}{5}$ to armadillos, $\frac{1}{5}$ each to pumas and jaguars, $\frac{1}{10}$ to tapirs, and $\frac{1}{20}$ each to capybaras and squirrel monkeys.

But then you see a spectacled bear! Clearly, the naïve empirical frequency estimator does not give a satisfactory result. It has completely failed to account for the possibility of finding a spectacled bear in the rainforest. It is even more disappointing when we realise that in this setting, this intuitive estimator is equivalent to the "maximum likelihood estimator" (MLE). The MLE is known to have excellent statistical properties. Obviously, in this case, it needs some improvement to account for all species.

One such simple modification to the MLE suggests adding a constant to the count of each species, including the ones that have not been observed in the sample. In general, if a species has occurred r times in the sample, the modified MLE, or the "add-constant estimator", would assign to it the probability $\frac{r+c}{N+Sc+c}$. Here N is the sample size, S is the number of distinct species in this sample, and c is a constant we use for the estimation.[3]

So, for example, we can look at the *add-one* estimator, for which $c = 1$. It assigns the probabilities $\frac{8+1}{27}$ to the armadillos, $\frac{4+1}{27}$ each to pumas and jaguars, $\frac{2+1}{27}$ to tapirs, $\frac{1+1}{27}$ each to capybaras and squirrel monkeys, and $\frac{0+1}{27}$ to the unseen species.

Unfortunately, when the number of species K is large compared to the sample size, add-constant estimators perform poorly as well. A vivid example of such a failure can be found in [11], and we reproduce it here.

Say, instead of estimating the distribution of animal species in the rainforest, you are interested in evaluating the distribution of their DNA sequences. You

[2] "Empirical" simply means that the calculation is based on the observations that have been made.

[3] The denominator is chosen such that the obtained estimates are probabilities, that is, they add up to 1.

have observed the DNA sequences of a large number of animals and discovered that each of the N observed DNA sequences is unique. You would like to make an inference about the distribution of all possible DNA sequences.

Let Z denote the number of distinct species in the sample (which equals 6 in our example above). Note that in the extreme case we assume that $Z = N$, hence the add-c estimator would assign probabilities $\frac{1+c}{N+Nc+c}$ to each observed DNA sequence and $\frac{c}{N+Nc+c}$ to all unobserved sequences.

Now, for a fixed value of c we can see that

$$\frac{N(1+c)}{N+Nc+c} = \frac{N+Nc}{N+Nc+c}$$

approaches the value 1 as N gets bigger and bigger.

In other words, the probability this estimator assigns to all observed sequences is close to 1, whereas the probability it assigns to all unseen sequences is close to 0, which (as we know) is not at all representative of the truth.

2 The Enigma machine and cryptanalysis

2.1 Cryptanalysis during World War II

During World War II, most of the messages transmitted by the German military forces were encrypted using a device called the Enigma machine. There were several versions of the Enigma with different levels of security depending on the usage (by the Navy, the Air Force, the Secret Service and so on).

Breaking the naval Enigma code was important to the western Allies as the damage they suffered against the Axis powers at sea greatly exceeded the damage caused by the air forces and the ground troops [10, 13]. According to [10], Hitler believed that it would be the German U-boats that would win the war for him. For this reason the security requirements for the messages encrypted by the German navy were even higher than those for the army and the air force.

Meanwhile, at Bletchley Park in England, cryptanalysts were working for the British Intelligence attempting to break the naval Enigma code. In the course of this work, two of these cryptanalysts, Irving J. Good (1916-2009) and Alan M. Turing (1912-1954), faced a unique problem: estimating the distributions of bigrams and trigrams[4] used by the German navy in the encryption process. Eventually, Good and Turing came up with a non-trivial solution to the problem of decrypting previously unseen letter groupings.

Before we go on to discuss their solution, let us try to understand the procedure used to encrypt the naval messages.

[4] Bigrams and trigrams are simply sequences of two and three letters respectively.

2.2 The Enigma machine

A typical military Enigma machine was a device resembling a massive typewriter (see Figure 1).[5] Located at the front of the naval Enigma was a "plugboard", consisting of 26 sockets, one for each letter of the alphabet. The plugboard introduced an extra level of scrambling by allowing for any two letters to be swapped when connected by a cable, before entering the three rotor wheels, and once again when exiting and before reaching the lampboard.

Figure 1: Enigma M3 (isometric view)

To set up the naval Enigma, three rotor wheels had to be chosen out of a library of eight and placed into the machine in the specified order. The rings of the wheels were then adjusted to their predetermined positions, ten pairs of

[5] For a detailed description of the Enigma design an interested reader is referred to [3].

letters were connected on the plugboard, and the rotor wheels were turned to their starting position.

2.3 Settings: the daily key

The German U-boats were issued with monthly sheets that contained instructions for setting up the Enigma machine for each day of the month. The "daily key", a set of instructions with four components (wheel order, ring setting, plugboard pairs, and ground setting) would look something like this:

Date	Rotors	Ring Settings	Plugboard Settings	Ground Settings
29	I V II	B H N	GP XV CK IZ QT NO JH BW AY TR	XIO

The number of different configurations for a naval Enigma machine, in fact, exceeded 8.9×10^{20}. Without knowing the settings, it would have taken months to test for each possible combination.

2.4 Enigma's flaw and codebreaking techniques

In spite of being a powerful encrypting device, the Enigma had a flaw that provided vital clues to the codebreakers. This flaw stemmed from what is called the "reciprocal property" of the machine. What this meant was that in a given state of the machine, a letter, say, Q, encrypted as A led necessarily to A being encrypted as Q along with the fact that no letter could be enciphered as itself.

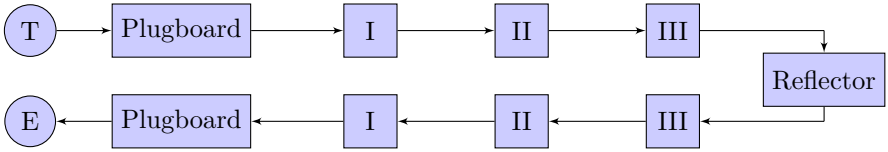
Now, say you want to decipher the message "JXATQBGGYWCRYBGD". You know that it is a part of the weather report and thus may contain the word "WETTERBERICHT"[6] with high probability. You would "slide" the word "WETTERBERICHT" along the ciphertext to find where it might "fit" by eliminating all cases with the "collisions", the instances where the letter in the code would have to be enciphered as itself:

	...	J	X	A	T	Q	B	G	G	Y	W	C	R	Y	B	G	D	...
1		W	E	T	T	E	R	B	E	R	I	C	H	T				
2			W	E	T	T	E	R	B	E	R	I	C	H	T			
3				W	E	T	T	E	R	B	E	R	I	C	H	T		
4					W	E	T	T	E	R	B	E	R	I	C	H	T	

The only possible fit occurs in the third row. This technique allowed for eliminating a lot of impossible initial settings and provided the starting point for breaking the code.

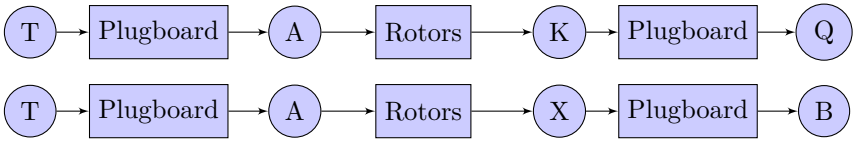
[6] *Wetterbericht* is the German word for *weather report*.

We will continue with this example to describe the idea behind working out the plugboard combinations. Observe that for T in the third row (the only row yielding a valid fit), we have the following correspondences: TE, TQ, TB, TG. The path for, say, TE, through the machine would look like this:



We can make an intial guess: T is connected with A on the plugboard. Consider the path:

From this we can deduce that P should be wired to E on the plugboard. Suppose that we continue with the other connections and get:



Hence, we can deduce the correspondences KQ and XB. At last:

It follows that T should be connected with G. But by our initial guess T is connected with A and cannot be connected with both. Thus the initial guess is incorrect and we have to make another one and repeat the process.

Technically, we would have to check 26 options.[7] If all the 26 options are incorrect, then the rotor position must be wrong. Then we choose the next rotor position and go over all possible combinations again!

Turing made a crucial observation: once a collision has been found (TA and TG in our example), all intermediate deductions (connections KQ and XB) can be rejected simultaneously and do not have to be checked again.

This principle combined with the power of electrical circuits underpinned the idea behind the "anti-Enigma" machine called the Bombe. This machine could go over all rotor positions in 18 minutes [10], eliminating all impossible positions and leaving only a few possibilities that could be checked manually.

[7] No connection on the plugboard is also an option.

2.5 Sending an encrypted naval message

Naval messages were usually transmitted in 4-letter groups. Two additional 4-letter groups, the "indicator groups", were placed at the beginning and at the end of each message.

Here is an example to make things clearer:

MMÄ 1416/27/989 38
IJTV USYX DERH RFRS OQRV DTYH QWBV HILS CXHR OPOD
GTQL DDHI KFTG EDZS WXQS EDFR HGYG EDZZ UYQV DTYY
EDGH KIRM SYBK PANX JSTP QXDT ERGP JMSX VFWI FTPZ
ADHK WDLE QPAL ALDH XNDH RYFH IJTV USYX
1231 7640

In this ciphertext, MMÄ is the identifier of the transmitting station, 1416 is the time at which the transmission began, 27 is the day of the month, 989 is the serial number of the message, and 38 is the total number of 4-letter groups.

When deciphered, the message reads: "BISMARCK MUST NOW BE ASSUMED TO HAVE SUNK. U-BOATS TO SEARCH FOR SURVIVORS IN SQUARE BE6150 AND TO NORTH WEST OF THIS POSITION."

One of the factors that made breaking the naval Enigma code so difficult was that the operator doubly enciphered a trigram of letters with which each message began and ended (marked in blue) and indicated the "message setting" (the starting position of the rotors).

Here is the procedure for using a message setting:

1. The operator chose a trigram at random, say, ARQ, from the "K-book"[8] that contained all 26^3 possible trigrams in random order (after using a trigram the operator would cross it out in their copy of the K-book to never use it again, the other operators, however, were not prevented from using that same trigram).
2. The rotors were then set to the ground setting, the three-letter group, say, JNY, fixed for the day in the daily key.
3. The operator typed in ARQ to obtain an encryption, say, LVN, which was the message setting, defining the position to which to set the wheels in order to encrypt the message itself.
4. The message setting had to be sent to the message recipient so that the latter could decrypt it using the ground setting and discover the message setting.

[8] *Kenngruppenbuch* in German.

5. Before transmitting it, the sender had to disguise ARQ by choosing another trigram from the K-book, say, YVT, and writing down the two chosen trigrams in a shifted pattern, and then filling in the blanks with 2 arbitrary letters:

$$\begin{array}{cccc} . & Y & V & T \\ A & R & Q & . \end{array} \quad \longrightarrow \quad \begin{array}{cccc} W & Y & V & T \\ A & R & Q & N \end{array}$$

6. Next, the operator consulted the day's "bigram table", an essential element for encrypting a naval message, to replace all occurrences of a given bigram, say, WA, with its equivalent, given in the table, say, IJ and vice versa.[9]
7. After having replaced all vertical pairs WA, YR, VQ, and TN by their equivalents in the table, say, IJ, TV, US, and YX, the operator placed the "indicator groups" IJTV USYX both at the beginning and at the end of the message.
8. The recipient looked up IJ, TV, US, and YX in his copy of the bigram table to obtain the initial bigrams.
9. Setting the wheels to the ground setting JNY, the trigram ARQ became the message setting LVN. This yielded the plaintext on setting the wheels accordingly and typing in the ciphertext.

The British intelligence managed to lay its hands on the K-book, but capturing the bigram tables was almost an impossible task: if a U-boat ever came under attack, the crew had strict orders to destroy the tables, which were printed in water-soluble ink [5].

3 The Enigma and the Good–Turing formula

3.1 Motivation behind the formula

Turing's method for the identification of the message settings relied upon the assumption that some trigrams were more popular with the German operators than the others. It was thus necessary to estimate the probabilities with which the operators used the trigrams.

Most likely, Turing's hypothesis was correct. In [7], Good points out that, as discovered from the captures, the trigrams printed at the top of the pages of the K-book were used more frequently than the others. A lot of letter groupings appeared only once, some not at all. The cryptanalysts wanted to learn the rare letter groupings and the groupings that had not yet appeared in

[9] There was a set of nine tables that was reissued several times during the war, and the operators had a calendar with instructions for which of the nine tables to use on a given day [3].

the collection of the intercepted German missives. Assigning a zero probability to these groupings would imply the assertion that they will not ever be used by the operators. So, Turing decided to assign those missing trigrams a small non-zero probability. By estimating the frequency of unseen species in his sample, he could then estimate the probability of the letter groupings appearing in a much larger sample of messages as well as in the very next intercepted Enigma message.

3.2 Main concepts and notation

The main goal of Good's work was to construct a good estimate of the total population frequency of the unobserved species (see [6]) using Turing's approach. Let n_r be the number of species represented by exactly r individuals in a sample of size N. Although mainly interested in finding an estimate for the probability of missing some species in the sample, Good attempted to answer a more general question: for a fixed integer $r \geq 0$, how can one estimate the true probability q_r that an arbitrary species is represented exactly r times in a random sample of size N?

The formula, suggested by Turing, and proved by Good in [6], gives the following estimator for q_r:

$$q_r \approx \frac{(r+1)}{N} \frac{n_{r+1}}{n_r}. \tag{1}$$

When one compares this estimator with $\frac{r}{N}$, one can think of $\frac{(r+1)n_{r+1}}{n_r} = r^*$ as an adjusted value of r.

It is remarkable that this formula estimates the probability of events by taking into account not only the number of their occurrences in the sample, but also the information about how many times other, possibly unrelated, events were seen.

In the coming sections, we will try to throw some light on this important statistical tool. We assume that the reader is familiar with the expected value of a discrete random variable, conditional probability and expectation, Bayes' theorem, and the binomial distribution.

3.3 The formula derived step by step

Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be a sample drawn from an infinitely large population of individuals. We are interested in estimating the unknown population frequency for species s, that is, $p_s = \mathbb{P}(X = s)$, $s = 1, \ldots, K$ with K denoting the unknown number of distinct species in the population. We assume that K is

finite. The sample count of the species s in the sample can be defined as

$$C(s) = \sum_{i=1}^{N} \mathbb{1}_{\{X_i = s\}}, \tag{2}$$

where $\mathbb{1}(A)$ denotes the indicator function, taking the value 1 if the event A occurs and 0 if it does not. Note that $C(s) = 0$ for any unseen species s in the random sample.

As before, we will denote by n_r the number of species which occurred exactly r times in the sample $\boldsymbol{X}$. Then,

$$n_r = n_r(\boldsymbol{X}) = \sum_{s=1}^{K} \mathbb{1}_{\{C(s)=r\}}.$$

Furthermore, we have $N = \sum_{r \geq 1} r n_r$. Let q_r denote the actual population frequency of an arbitrary species with sample count r. By definition, q_r can take its values from the finite set $\{p_s : s = 1, \ldots, K\}$.

Now, let us examine the probability distribution of q_r. For $s = 1, \ldots, K$, consider the event $\{q_r = p_s\}$. Assuming for the moment that each of the K species has a different frequency in the population[10], define

$$\begin{aligned} A_s &\equiv \{\text{The selected species is } s\} \quad \text{and} \\ B_r &\equiv \{\text{The selected species is represented} \\ &\quad \text{by exactly } r \text{ individuals in the sample}\}. \end{aligned}$$

Then, $\{q_r = p_s\} \equiv A_s | B_r$, where $|$ means that the event on the right has occurred. The expectation of q_r is given by

$$\mathbb{E}[q_r] = \sum_{s=1}^{K} p_s \mathbb{P}(q_r = p_s) = \sum_{s=1}^{K} p_s \mathbb{P}(A_s | B_r). \tag{3}$$

By the extended Bayes' theorem, we have

$$\mathbb{P}(A_s | B_r) = \mathbb{P}(A_s) \frac{\mathbb{P}(B_r | A_s)}{\sum_{k=1}^{K} \mathbb{P}(B_r | A_k) \mathbb{P}(A_k)}. \tag{4}$$

Further assuming that all species have the same probability of being selected, we get $\mathbb{P}(A_k) = 1/K$ for all $k = 1, \ldots, K$. Recalling the definition of $C(s)$ in Equation 2 and replacing s by k, it is easy to see that $B_r | A_k \equiv \{C(k) = r\}$.

[10] This assumption also appears in the original derivation by Good [6].

This alternative representation is very useful since the random variable $C(k)$ is a binomial random variable with parameters N and p_k. This implies that

$$\mathbb{P}(C(k) = r) = \binom{N}{r} p_k^r (1 - p_k)^{N-r}$$

for any $k = 1, \ldots, K$. Therefore, we can write Equation 4 as

$$\mathbb{P}(A_s | B_r) = \frac{p_s^r (1 - p_s)^{N-r}}{\sum_{k=1}^{K} p_k^r (1 - p_k)^{N-r}}.$$

The expected value of q_r in Equation 3 now has the equivalent expression

$$\mathbb{E}[q_r] = \frac{\sum_{s=1}^{K} p_s^{r+1} (1 - p_s)^{N-r}}{\sum_{s=1}^{K} p_s^r (1 - p_s)^{N-r}}, \tag{5}$$

where the summation index k in the denominator has been substituted by s.

Now, using the fact that the expectation of the sum equals the sum of the expectations, we observe that

$$\mathbb{E}[n_r] = \mathbb{E}\left(\sum_{s=1}^{K} \mathbb{1}_{\{C(s)=r\}} \right) = \sum_{s=1}^{K} \mathbb{E}\left(\mathbb{1}_{\{C(s)=r\}} \right) = \binom{N}{r} \sum_{s=1}^{K} p_s^r (1 - p_s)^{N-r}.$$

Similarly,

$$\mathbb{E}(n_{r+1}) = \binom{N}{r+1} \sum_{s=1}^{K} p_s^{r+1} (1 - p_s)^{N-r-1}.$$

The trick used by Good [6] is to now imagine that the sample has been augmented by one individual. This means that N changes into $N + 1$ and the previous expectation becomes

$$\mathbb{E}[n_{r+1}] = \binom{N+1}{r+1} \sum_{s=1}^{K} p_s^{r+1} (1 - p_s)^{N-r}.$$

To specify the sample size in the calculation of this expectation, we will follow the notation of Good [6] and use $\mathbb{E}_N$ and $\mathbb{E}_{N+1}$ to indicate that the expectation is evaluated under N and $N + 1$ respectively. This gives us

$$\begin{aligned}
\mathbb{E}_N[q_r] &= \frac{\binom{N}{r}}{\binom{N+1}{r+1}} \frac{\mathbb{E}_{N+1}[n_{r+1}]}{\mathbb{E}_N[n_r]} \\
&= \frac{N!}{r!(N-r)!} \frac{(r+1)!(N-r)!}{(N+1)!} \frac{\mathbb{E}_{N+1}[n_{r+1}]}{\mathbb{E}_N[n_r]} \\
&= \frac{r+1}{N+1} \frac{\mathbb{E}_{N+1}[n_{r+1}]}{\mathbb{E}_N[n_r]}.
\end{aligned}$$

The Good–Turing formula in Equation 1 can now be obtained by replacing the expectations $\mathbb{E}_N[n_r]$ and $\mathbb{E}_{N+1}[n_{r+1}]$ by their sample-based counterparts n_r and n_{r+1}, and using the fact that for large N, $\frac{1}{N+1} \approx \frac{1}{N}$. From the formula we can easily conclude that the total probability of the occurrence of all species which are represented by exactly r individuals in the sample can be estimated by

$$\widehat{p}_r = n_r \frac{r+1}{N} \frac{n_{r+1}}{n_r} = \frac{(r+1)n_{r+1}}{N}.$$

An interesting and practically important consequence of the formula is that the total probability of missing out some species belonging to the population in the sample $\boldsymbol{X}$ can simply be estimated by

$$\widehat{p}_0 = \frac{n_1}{N}.$$

This probability, referred to as "noncoverage probability" in [14], is equal to the fraction of "singletons" in the sample, that is, the species represented by a single individual. This notion can also be understood in terms of a more concrete interpretation: $\widehat{p}_0$ gives an approximation, for large enough N, of the probability that the $(N+1)$-th species has not occurred among the first N individuals.

3.4 More applications and further research

The Good–Turing formula produces estimates for the population frequencies corresponding to the sample frequencies of the observed species as well as an estimate for the total population frequency of *all* unseen species. But it does not specify how the total probability of all unseen species is shared among them. Nor does it provide an estimate for the number of unseen species in the population. However, these quantities are often of interest in practice.

Here are some more recent examples:

1. studies of unseen genetic variations with the objective of estimating the number of unseen variants in the human genome [9], and
2. studies of password use and reuse habits with the objective of estimating how many different passwords a user types in a day and how many passwords are shared among different sites [4].

While the first problem admits the direct application of the Good–Turing frequency estimator, the second requires further refinements. A number of researchers have used the Good–Turing formula to develop effective techniques for estimating the total number of species in a population (see, for example,

[8]). Inspired by the Good–Turing formula, the very important recent works by A. Chao [2] and A. Orlitsky [12] explore the problem of estimating species richness in great depth.

Image credits

Figure 1: Figure 10.1, Chapter 10 in B. J. Copeland, J. Bowen, M. Sprevak, and R. Wilson, *The Turing guide*, Oxford University Press, 2017.

References

[1] J. Bunge and M. Fitzpatrick, *Estimating the number of species: a review*, Journal of the American Statistical Association **88** (1993), no. 421, 364–373.

[2] A. Chao, C.-H. Chiu, R. K. Colwell, L. F. S. Magnago, R. L. Chazdon, and N. J. Gotelli, *Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good–Turing theory*, Ecology **98** (2017), no. 11, 2914–2929.

[3] B. J. Copeland, J. Bowen, M. Sprevak, and R. Wilson, *The Turing guide*, Oxford University Press, 2017.

[4] D. Florencio and C. Herley, *A large-scale study of web password habits*, Proceedings of the 16th International Conference on World Wide Web, 2007, pp. 657–666.

[5] L. A. Gladwin, *Alan Turing, Enigma, and the breaking of German machine-ciphers in World War II*, Prologue: the Journal of the National Archives **29** (1997), no. 3, 202–217.

[6] I. J. Good, *The population frequencies of species and the estimation of population parameters*, Biometrika **40** (1953), no. 3-4, 237–264.

[7] I. J. Good, *Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma*, Journal of Statistical Computation and Simulation **66** (2000), no. 2, 101–111.

[8] I. J. Good and G. H. Toulmin, *The number of new species, and the increase in population coverage, when a sample is increased*, Biometrika **43** (1956), no. 1-2, 45–63.

[9] I. Ionita-Laza, C. Lange, and N. M. Laird, *Estimating the number of unseen variants in the human genome*, Proceedings of the National Academy of Sciences **106** (2009), no. 13, 5008–5013.

[10] S. B. McGrayne, *The theory that would not die: how Bayes' rule cracked the Enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*, Yale University Press, 2011.

[11] A. Orlitsky, N. P. Santhanam, and J. Zhang, *Always Good Turing: asymptotically optimal probability estimation*, Science **302** (2003), no. 5644, 427–431.

[12] A. Orlitsky, A. T. Suresh, and Y. Wu, *Optimal prediction of the number of unseen species*, Proceedings of the National Academy of Sciences **113** (2016), no. 47, 13283–13288.

[13] A. M. Turing and B. J. Copeland, *The essential Turing: seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of Enigma*, Oxford University Press (2004).

[14] Z. Zhang, *Statistical implications of Turing's formula*, Wiley Online Library, 2017.

Fadoua Balabdaoui *is a senior scientist at the Swiss Federal Institute of Technology in Zürich and an associate professor at the Université Paris-Dauphine.*

Yulia Kulagina *is a PhD student at the Swiss Federal Institute of Technology in Zürich.*

Mathematical subjects
Probability Theory and Statistics

Connections to other fields
Life Science

DOI
10.14760/SNAP-2021-008-EN

Snapshots of modern mathematics from Oberwolfach provide exciting insights into current mathematical research. They are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the interested public worldwide. All snapshots are published in cooperation with the IMAGINARY platform and can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

ISSN 2626-1995

Junior Editors
Anup Anand Singh and Sara Munday
junior-editors@mfo.de

Senior Editor
Sophia Jahns
senior-editor@mfo.de

Mathematisches Forschungsinstitut
Oberwolfach gGmbH
Schwarzwaldstr. 9 –11
77709 Oberwolfach
Germany

Director
Gerhard Huisken

Mathematisches
Forschungsinstitut
Oberwolfach

Member of Leibniz Association

IMAGINARY
open mathematics