

On Bayesian estimation of multinomial probabilities under incomplete experimental information

Pepa Ramírez^a, Brani Vidakovic^b

^a*Departamento de Estadística, Universidad Carlos III de Madrid,*

^b*Department of Biomedical Engineering, Georgia Institute of Technology*

Abstract

In this note, we discuss Bayesian estimation of multinomial probabilities associated with a finite alphabet \mathcal{A} , under incomplete experimental information. Two types of prior information are considered: (i) number of letters needed to see a particular pattern for the first time, and (ii) the fact that for two fixed words one appeared before the other.

Key words: Patterns, stopping times, incomplete experimental information.

¹ Corresponding address: Pepa Ramírez, The Wallace H.Coulter Department of Biomedical Engineering, 313 Ferst Drive, Suite 2127, Atlanta, GA 30332-0535 (jrcobo@est-econ.uc3m.es)

² Pepa Ramírez was partially supported in this research by a Fulbright grant (Ruth Lee Kennedy) to do a research stay at the Georgia Institute of Technology.

1 Introduction

In this work we take Bayesian point-of-view in the following context. We consider a discrete random variable L that, with probabilities $\mathbf{p} = (p_1, \dots, p_d)$, takes the values (l_1, \dots, l_d) , that is, $P(L = l_j) = p_j, j = 1, \dots, d$. One can think of the values l_j as letters or symbols in an alphabet \mathcal{A} . Let \mathcal{A}^n be the set of all words of length n in \mathcal{A} . To generate a particular word $\omega = \omega_1\omega_2 \dots \omega_n$, one selects letters from \mathcal{A} sequentially, starting with ω_1 , and ending with ω_n . Several problems arise from this setting. For instance, Example 7.2.(a) in Ross (1983), shows how to compute the expected value of the number of symbols N needed to be observed until some given sequence appears. In addition, the author calculates the probability that for two given patterns one occurs before the other. This is done by use of martingales and reformulating the problem in the terms of fair casino gambling. It can be proven that in this context, N represents a stopping time related to a certain martingale. Bizley (1972) also considers the problem of determining the expected number of trials (or symbols) required to obtain a specific pattern and gives expressions in terms of probability generating functions. DasGupta (1993) is also an example in which stopping times in the Bayesian context are used.

In this note, we estimate the probabilities (p_1, \dots, p_d) in Bayesian fashion, given the stopping time N_ω which is defined as the number of observed symbols until the sequence ω appears for the first time.

The organization is as follows. Section 2 provides necessary definitions and an important theorem due to Bizley (1972). In Section 3, we develop Bayesian inference for the Bernoulli model (case for $d = 2$) and generalize it to the

multinomial case, providing some examples. Section 4 considers the update of probabilities with a different limited experimental information. A brief conclusion is given in Section 5.

2 Theoretical background

We defined $\mathcal{A} = \{l_1, l_2, \dots, l_d\}$ as the alphabet consisting of d different letters and \mathcal{A}^n as the set of all words of length n in \mathcal{A} . We assumed that letter l_j is selected with probability p_j , independently of the previously selected letters. Let $\Omega = \mathcal{A}^\infty$ be the sample space, and $\mathcal{F} = \sigma(\cup_{k \in \mathbb{N}} \sigma(\mathcal{A}^k))$ a σ -algebra on Ω , where $\sigma(\mathcal{A}^k)$ is the σ -algebra generated by all subsets of \mathcal{A}^k . Then, (Ω, \mathcal{F}) is a measurable space. Let $\mathcal{F}_k = \sigma(\cup_{i \leq k} \sigma(\mathcal{A}^i))$, $k = 1, 2, \dots$ be a sequence of nested σ -algebras. In a sequential generation of word $\omega = \omega_1 \omega_2 \dots$, the statement $\rho \in \mathcal{F}_k$ means that information whether the pattern ρ appeared up to index k is available. The set of σ -algebras $\{\mathcal{F}_k\}_{k \geq 1}$ forms a *filtration* on (Ω, \mathcal{F}) , that is

$$\mathcal{F}_k \subseteq \mathcal{F} \quad \text{and} \quad k_1 \leq k_2 \Rightarrow \mathcal{F}_{k_1} \subseteq \mathcal{F}_{k_2}.$$

Definition 1 Let (Ω, \mathcal{F}) be a measurable space and $\{\mathcal{F}_k\}_{k \geq 1}$ a filtration on it. A random variable N is said to be a *stopping time* if, for every n , the event $\{N \leq n\}$ is \mathcal{F}_n -measurable.

Let μ be a measure defined of the measurable space (Ω, \mathcal{F}) . For a word $\omega = \omega_1 \omega_2 \dots \omega_n \in \mathcal{A}^n$ assume that $P(\omega_i = l_j) = p_j$, for $i = 1, \dots, n$ with $\sum_{j=1}^d p_j = 1$. A simple measure on Ω can be defined by means of *cylindric sets*.

Definition 2 The *cylindric set* corresponding to the word $x = x_1 x_2 \dots x_n$ is

given by $\Gamma_x = \{\omega \in \Omega | \omega_1 = x_1, \dots, \omega_n = x_n\}$.

The probability of word x is in fact formally defined as the probability of $\Gamma_x \subseteq \Omega$ and then $\mu(\Gamma_x) = \prod_{i=1}^n P(x_i)$, where $P(x_i)$ is the probability of symbol x_i . The measure μ on Ω introduced as above is known as ‘Bernoulli’ measure. Various other measures are possible. The following definition can be found in Bizley (1972).

Definition 3 *The symbol x_i in the word $x = x_1x_2 \dots x_n$ is said to be critical if $x_1x_2 \dots x_i = x_{n-i+1} \dots x_n$.*

Notice that the symbol x_n is always critical as $x_1 \dots x_n = x_1 \dots x_n$. For instance, in the word 1010, second and fourth symbols are critical. In the word 1001, first, and fourth symbols are critical while in the word 100000, only the last symbol is critical. Critical symbols in the word *ABRACADABRA* are the first, fourth and eleventh.

Assume that the word ω contains m critical symbols. Denote by a_{ij} the number of letters l_i up to and including the j th critical symbol. For example, in the word 1010, $l_1 = 0, l_2 = 1, a_{11} = 1, a_{21} = 1, a_{12} = 2$, and $a_{22} = 2$.

Theorem 4 (Bizley, 1972) *The number N_ω of consecutive symbols needed to complete a given pattern ω is a stopping time with probability generating function*

$$G_\omega(s) = \left(1 + (1 - s) \sum_{j=1}^m s^{-\sum a_{ij}} \prod_{i=1}^d p_i^{-a_{ij}} \right)^{-1}, \quad (1)$$

and expectation

$$E(N_\omega) = \sum_{j=1}^m \prod_{i=1}^d p_i^{-a_{ij}}. \quad (2)$$

For example, for $\omega = 010$, $a_{11} = 1, a_{21} = 0, a_{12} = 2$, and $a_{22} = 1$. The probability generating function is

$$G_{010}(s) = \left[1 + (1-s) \left(\frac{1}{(1-p)s} + \frac{1}{(1-p)^2 ps^3} \right) \right]^{-1},$$

and

$$E(N_{010}) = \frac{1}{1-p} + \frac{1}{p(1-p)^2}. \quad (3)$$

3 Bayesian inference

Our goal is to estimate the vector of unknown probabilities $\mathbf{p} = (p_1, \dots, p_d)$, $p_j = P(\omega_i = l_j)$, in a Bayesian way. The only information we have about the experiment is the stopping time N_ω , the number of symbols observed to obtain the pattern ω for the first time. If the prior distribution for \mathbf{p} is $\pi(\mathbf{p})$, then using the Bayes formula we can compute the posterior distribution as

$$\pi(\mathbf{p}|N_\omega) = \frac{P(N_\omega|\mathbf{p})\pi(\mathbf{p})}{P(N_\omega)}. \quad (4)$$

The estimator of \mathbf{p} is the posterior mean. We will illustrate the proposed Bayesian inference on Binomial and Multinomial cases, i.e, when the alphabet contains two and more than two symbols.

3.1 Binomial case

Suppose that our alphabet consists only of two letters, say 0, 1 corresponding to *failure* and *success*. Let p be the probability of success. To be non-informative, assume that the prior distribution for p is $Beta(\alpha, \beta)$ with pa-

rameters $\alpha = \beta = 1$, which represents a uniform distribution on $(0, 1)$. Notice that a priori, $E(p) = 1/2$. Given that we know the stopping time N , we can estimate p and assess how the available information changes our prior belief. Several simple examples are provided.

Example 1. *Pattern 010 was observed for first time after $N_{010} = 10$ trials.*

Expression (3) is 10 for $p = 1/2$. The probability $P(N_{010} = 10|p)$ is computed as

$$G_{010}^{(10)}(0)/10! = (1 - p)^2 p (1 - 6p + 21p^2 - 41p^3 + 52p^4 - 39p^5 + 16p^6 - 3p^7)$$

which is plotted in Fig 1. The Bayes estimate of probability of success is equal to

$$E(p|N_{010}) = \frac{\int_0^1 p P(N_{010}|p) \pi(p) dp}{\int_0^1 P(N_{010}|p) \pi(p) dp} = 0.418838.$$

In order to check sensitivity with respect to the prior we computed the Bayes estimator for a range of values of $\alpha = \beta$. As expected, Fig 2 demonstrates that the posterior expectation for p tends to 0.5 as α increases.

Example 2. *Pattern 011 was observed first time after $N=10$ trials.* In this case, the probability generating function and expected value of N are

$$G_{011}(s) = \left(1 + (1 - s) \frac{1}{(1 - p)p^2 s^3} \right)^{-1} \quad \text{and} \quad E(N_{011}) = \frac{1}{(1 - p)p^2}.$$

$$P(N_{011} = 10|p) = p^2 - p^3 - 5p^4 + 10p^5 - 2p^6 - 9p^7 + 9p^8 - 3p^9,$$

thus, under the uniform prior, $E(p|N_{011}) = 0.573$. In this example we were also interested in comparing the Bayesian method with a classical method. The Maximum Likelihood Estimate (MLE) is easy to compute in this setting since the likelihood is a polynomial in p . However, as a polynomial, the likelihood can have multiple extrema and as a consequence, non-unique

MLE. Actually, in this example we found two local maximums at $\hat{p}_1 = .45$ and $\hat{p}_2 = .84$. So, for this problem, the MLE is problematic.

FIGURES 1 AND 2 ABOUT HERE

3.2 Multinomial case

If we consider the general case where the alphabet consists of d letters, l_1, \dots, l_d with vector of probabilities $\mathbf{p} = (p_1, \dots, p_d)$, the previous calculations can be generalized in a straightforward way. The natural prior distribution for \mathbf{p} is the *Dirichlet distribution* with a density

$$\pi(\mathbf{p}) = \pi(\mathbf{p}|\alpha_1, \dots, \alpha_d) = \frac{\Gamma(\sum_{i=1}^d \alpha_i)}{\prod_{i=1}^d \Gamma(\alpha_i)} \prod_{i=1}^d p_i^{\alpha_i-1}, \quad \alpha_i \geq 0, \sum_{i=1}^d p_i = 1.$$

Suppose that N_ω , the number of consecutive symbols to complete a given pattern ω is known to be n . Then, from (1), $P(N_\omega = n|\mathbf{p}) = G_\omega^{(n)}(0)/n!$. By Bayes formula (4), we obtain the posterior distribution:

$$\pi(\mathbf{p}|N_\omega = n) = \frac{P(N_\omega = n|\mathbf{p})\pi(\mathbf{p})}{\int P(N_\omega = n|\mathbf{p})\pi(\mathbf{p})d\mathbf{p}}. \quad (5)$$

To find the expected posterior probability for symbol l_i we compute

$$E(p_i|N_\omega = n) = \int_0^1 p_i \pi(p_i|N_\omega = n) dp_i, \quad (6)$$

where the marginal posterior $\pi(p_i|N_\omega = n)$ is obtained from (5).

Example 3. As an illustration, we take the case where $d = 3$, and assume that the letters are 0, 1 and 2. We are interested in Bayesian estimation of $p_1, p_2, p_3 = 1 - p_1 - p_2$, given that the word $\omega = 012$ appeared for the first

time after $N_{012} = 10$ trials. This word has only one critical symbol (the last one), so $a_{11} = a_{12} = a_{13} = 1$ and thus, the probability generating function and expected value are

$$G_{012}(s) = \frac{s^3 p_1 p_2 p_3}{1 + s^3 p_1 p_2 p_3 - s}, \quad E(N_{012}) = \frac{1}{p_1 p_2 p_3}.$$

The likelihood $P(N_{012} = 10 | p_1, p_2, p_3) = p_1 p_2 p_3 (1 - 5p_1 p_2 p_3 + 3p_1^2 p_2^2 p_3^2)$, for some values of \mathbf{p} is depicted in Fig 3.

FIGURE 3 ABOUT HERE

If prior parameters are, for instance $\alpha_1 = \alpha_3 = 3$ and $\alpha_2 = 5$, then the prior is $\pi(\mathbf{p}) \propto p_1^2 p_2^4 p_3^2$ and the expected a priori probabilities are $E(p_1) = E(p_2) = .2727$ and $E(p_3) = .4545$, that is, we favored symbol 1 against 0 and 2, before observing $N_\omega = 10$. Applying expression (6) we find that a posteriori, the expected values change to $E(p_1 | N_\omega = 10) = E(p_3 | N_\omega = 10) = .2843$ and $E(p_2 | N_\omega = 10) = .4312$. Thus, having observed the pattern 012 for the first time after 10 trials, our beliefs about p_1, p_2 and p_3 are modified: the expected values of p_1 and p_3 increase and that of p_2 decreases.

4 Updating p if the word ω appeared before the word ρ

In this Section we consider a different type of limited experimental information. The only information available is the fact that a particular word ω appeared before another fixed word ρ (we will denote this as $\omega \prec \rho$). Let N_ω and N_ρ be as before and $N_{\rho|\omega}$ be the number of additional observations necessary to obtain the word ρ when the word ω is just observed. The definition of $N_{\omega|\rho}$ is analogous. To apply Bayes formula and update p , the following result

can be utilized.

$$\textbf{Theorem 5} \quad P(\omega \prec \rho|p) = \frac{E(N_\rho) + E(N_{\omega|\rho}) - E(N_\omega)}{E(N_{\omega|\rho}) + E(N_{\rho|\omega})}. \quad (7)$$

Proof: The proof is straightforward. Denote by $a \wedge b$ the minimum of a and b . If N_ρ is represented as $N_\rho = (N_\rho \wedge N_\omega) + N_\rho - (N_\rho \wedge N_\omega)$, and $N_\rho - (N_\rho \wedge N_\omega)$ is conditioned on the event $\{\omega \prec \rho\}$, it follows

$$\begin{aligned} E(N_\rho) &= E(N_\rho \wedge N_\omega) + E(N_\rho - (N_\rho \wedge N_\omega)|\omega \prec \rho)P(\omega \prec \rho|p) \\ &= E(N_\rho \wedge N_\omega) + E(N_{\rho|\omega})P(\omega \prec \rho|p). \end{aligned} \quad (8)$$

Equivalently, $E(N_\omega) = E(N_\rho \wedge N_\omega) + E(N_{\omega|\rho})P(\rho \prec \omega|p)$.

Since $P(\rho \prec \omega|p) = 1 - P(\omega \prec \rho|p)$, by solving (8) and the counterpart for $E(N_\omega)$, we get (7). \square

4.1 Calculating $N_{\rho|\omega}$

In order to calculate Expression (7) in Theorem 5 we need an efficient way of calculating $E(N_{\omega|\rho})$ and $E(N_{\rho|\omega})$. Suppose that $\omega = \omega_1\omega_2 \dots \omega_{n_1}$ is observed and assume that the expected number of trials to obtain the word $\rho = \rho_1\rho_2 \dots \rho_{n_2}$ is of interest. Let j be such that $\omega_{n_1-j+1} \dots \omega_{n_1} = \rho_1 \dots \rho_j$. We assume $j > 0$, since if $j = 0$, ρ does not overlap ω , and $E(N_{\rho|\omega}) = E(N_\rho)$. In addition, we also assume that $j < \min\{n_1, n_2\}$. If $n_2 \leq n_1$ and $j = n_2$, then we already observed ρ in ω and $N_{\rho|\omega} = 0$. If $n_1 \leq n_2$ and $j = n_1$, then $E(N_{\rho|\omega}) = E(N_{\rho_{j+1} \dots \rho_{n_2}})$.

Theorem 6

$$\begin{aligned} E(N_{\rho|\omega}) &= (n_2 - j) + (1 - P(\rho_{j+1} \dots \rho_{n_2}))(E(N_\rho) - (n_2 - j)) + \\ &\quad + \sum_{i=1}^{n_2-j} iP(\rho_{j+1}\rho_{j+2} \dots \rho_{j+i-1})Q(\rho_{j+i}). \end{aligned} \quad (9)$$

Proof: Denote by $P(\omega_i\omega_{i+1}\dots\omega_k)$ the probability of the pattern $\omega_i\omega_{i+1}\dots\omega_k$, and $Q(\cdot) = 1 - P(\cdot)$. Then,

$$\begin{aligned} E(N_{\rho|\omega}) &= Q(\rho_{j+1})(1 + E(N_\rho)) + P(\rho_{j+1})Q(\rho_{j+2})(2 + E(N_\rho)) + \dots \\ &\quad + P(\rho_{j+1}\rho_{j+2}\dots\rho_{n_2-1})Q(\rho_{n_2})(n_2 - j + E(N_\rho)) \\ &\quad + P(\rho_{j+1}\rho_{j+2}\dots\rho_{n_2-1}\rho_{n_2}). \quad \square \end{aligned} \tag{10}$$

Example 4. As an illustration, consider the case where $\omega = 1001$, and $\rho = 010$. The information we have about the experiment is that ω appeared before ρ . If the probability of the symbol ‘1’ is p , then

$$E(N_{1001}) = \frac{1}{p} + \frac{1}{p^2q^2}, \text{ and } E(N_{010}) = \frac{1}{q} + \frac{1}{pq^2}.$$

Also, $E(N_{010|1001}) = 1 + pE(N_{010})$ and $E(N_{1001|010}) = p(1 + E(N_{1001})) + q^2(2 + E(N_{1001})) + 2pq$. Thus, $P(1001 \prec 010|p) = p^2(3 - 3p + p^2)/(1 + 2p^2 - 2p^3 + p^4)$ (see Fig.4). Under the uniform prior on p , $E(p|\omega \prec \rho) = .666884$.

FIGURE 4 ABOUT HERE

5 Summary and future directions

In this work we took Bayesian point of view to infer the probabilities of letters in an alphabet \mathcal{A} . First, we considered the case where the information about the experiment is the value of a stopping time variable, representing the number of trials necessary to obtain a fixed pattern. Several examples for the Bernoulli and multinomial models are provided. Second, we update the probabilities of letters if we know that a particular pattern appeared before the other fixed pattern. Our results help in inference of binomial and multinomial probabilities when the information from the experiment is limited.

References

Bizley, M. T. L., Patterns in repeated trials., J. Inst. Actu., 1972, London, 88, 360-366.

DasGupta, A., Rinott, Y., & Vidakovic, B., Stopping times related to diagnostics and outliers. Presented at the Bayesian BOAT conference, Basel 1993.

Ross, S., Stochastic Processes, 1983, John Wiley & Sons.