Statistics and Probability Letters 106 (2015) 109–112

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On distribution of the leadership time in counting votes and predicting winners

Czesław Stępniak

Department of Differential Equations and Statistics, Faculty of Mathematics and Natural Sciences, University of Rzeszów, Pigonia 1, PL-35-959 Rzeszów, Poland

ARTICLE INFO

Article history:
Received 7 March 2015
Accepted 18 June 2015
Available online 15 July 2015

MSC:
primary 60C05
secondary 62E15
05A19

Keywords:
Two-candidate election
Ballot problem
Leadership time distribution
Predicting winners

ABSTRACT

In two-candidate election the votes are counted in random order. Suppose that candidate A was the leader until the 9th vote. How may we use this information in predicting the future winner? To this aim we derive distributions of the first leadership time both for the winner and loser. Our conclusion is rather surprising.

1. Introduction

Two-candidate election such as the last round of presidential election always attracts great attention. Assume that the votes are counted in random order and we know that candidate A was the leader until the 9th vote. How may we use this information in predicting the future winner?

The classical ballot problem concerns the probability that the winner holds the leadership until counting of all votes. There is an extensive literature on this topic; see Feller (1968), Takacs (1997), Goulden and Serrano (2003) and Lengyel (2011) for the history and possible generalizations. We go off the beaten path.

Suppose that one of the candidates (future winner) will receive M votes and the other (future loser) N votes ($M > N$). Any record of the counting of votes may be represented as a path from the origin $(0, 0)$ to $(M+N, M-N)$ with steps of type $(1, -1)$ and $(1, 1)$. In particular, the first leadership time for the winner is n, where $0 < n < M+N$, if and only if, the path is touching the x-axis for $x = n+1$ and is lying above the axis for all positive integers $x \le n$. Similarly, the first leadership time for the loser is n, if the corresponding segment of the path is lying below the x-axis. It is clear that the number of paths for which the first leadership time is equal n is the same both for the winner and loser. This fact is known as reflection principle (see Feller, 1968, Goulden and Serrano, 2003, Renault, 2007 or Brémaud, 1994). In order to predict the future winner we need distributions of the first leadership time for the winner and loser.

Our problem may be classified as statistical inference in finite population (see Bolfarine and Zacks, 1992, Bolfarine and Zacks, 1991, Ghosh and Meeden, 1997, Meeden, 2000) but in difference to the commonly used approach, our population is neither associated with any auxiliary characteristic nor subject to any specific model. Classical examples of such "pure" finite

E-mail address: stepniak@umcs.lublin.pl.

http://dx.doi.org/10.1016/j.spl.2015.06.028
0167-7152/

populations are urn models. The major role in such models plays hypergeometric and negative hypergeometric distribution. Importance of such distributions was highlighted in Miller and Fridell (2007). It is worth to add that the leadership may be treated as a pattern in a sequence taken from finite population (cf. Balakrishnan and Koutras, 2001, Brémaud, 1994, Renault, 2007, Sen et al., 2006 or Stępniak, 2013).

2. Distribution of the first leadership time for winner and loser

Suppose the ballots are counted in random order that is all permutations of the ballots are equally probable.

Let X and Y be the first leadership times in the broad sense (i.e. including zero) for winner and loser, respectively. The classical ballot problem concerns only the probability that $X = M + N$. It is well known (see Brémaud, 1994, Feller, 1968, Goulden and Serrano, 2003 or Takacs, 1997) that

$$P(X = M + N) = \frac{M - N}{M + N}. \tag{1}$$

Extending this result we shall derive the distributions of the random variables X and Y. Let us introduce notations $p_n = P(X = n)$ and $q_n = P(Y = n)$.

Theorem 1. *Under the above assumption*

$$p_n = \begin{cases} \dfrac{N}{M+N}, & \text{if } n = 0, \\ \dfrac{1}{n}\dbinom{n}{\frac{n+1}{2}}\dbinom{M+N-n-1}{M-\frac{n+1}{2}}\dbinom{M+N}{M}^{-1}, & \text{if } n = 2k-1, \text{ for } k = 1, \dots, N, \\ \dfrac{M-N}{M+N}, & \text{if } n = M+N, \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

while

$$q_n = \begin{cases} \dfrac{M}{M+N}, & \text{if } n = 0, \\ \dfrac{1}{n}\dbinom{n}{\frac{n+1}{2}}\dbinom{M+N-n-1}{M-\frac{n+1}{2}}\dbinom{M+N}{M}^{-1}, & \text{if } n = 2k-1, \text{ for } k = 1, \dots, N, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

Proof. For $n = 0$ the formula (2) is evident and for $n = M + N$ it reduces to (1). For the remaining n let us introduce random variables

$$X_i = \begin{cases} 1, & \text{if the } i\text{th vote is for winner} \\ 0, & \text{if the } i\text{th vote is for loser.} \end{cases}$$

Then

$$p_n = P\left(\sum_{i=1}^{j} X_i > j - \sum_{i=1}^{j} X_i \text{ for all } j = 1, \dots, n \text{ and } X_{n+1} = 0\right)$$

and it may be presented in the form

$$P\left(\sum_{i=1}^{n} X_i = \frac{n+1}{2}, 2\sum_{i=1}^{j} X_i > j \text{ for all } j = 1, \dots, n \text{ and } X_{n+1} = 0\right).$$

By the well known formula for the hypergeometric distribution (for instance Khan, 1994),

$$P\left(\sum_{i=1}^{n} X_i = \frac{n+1}{2}\right) = \frac{\dbinom{M}{\frac{n+1}{2}}\dbinom{N}{\frac{n-1}{2}}}{\dbinom{M+N}{n}} \quad \text{for } n = 2k-1, \ k = 1, \dots, N$$