Available online at www.sciencedirect.com

ScienceDirect

SoftwareX 5 (2016) 227–233

SoftwareX

www.elsevier.com/locate/softx

fixedTimeEvents: An R package for the distribution of distances between discrete events in fixed time

Kristian Hovde Liland[a,b,*], Lars Snipen[a]

[a] *Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Aas, Norway*
[b] *Nofima—Norwegian Institute of Food, Fisheries and Aquaculture Research, Aas, Norway*

Received 9 December 2015; received in revised form 21 September 2016; accepted 29 September 2016

Abstract

When a series of Bernoulli trials occur within a fixed time frame or limited space, it is often interesting to assess if the successful outcomes have occurred completely at random, or if they tend to group together. One example, in genetics, is detecting grouping of genes within a genome. Approximations of the distribution of successes are possible, but they become inaccurate for small sample sizes. In this article, we describe the exact distribution of time between random, non-overlapping successes in discrete time of fixed length. A complete description of the probability mass function, the cumulative distribution function, mean, variance and recurrence relation is included. We propose an associated test for the over-representation of short distances and illustrate the methodology through relevant examples. The theory is implemented in an R package including probability mass, cumulative distribution, quantile function, random number generator, simulation functions, and functions for testing.

Keywords: R Package; Distributions; Distances; Gene expression

Code metadata

Current code version	1.0
Permanent link to code/repository used of this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-15-00084
Legal Code License	GPL ≥ 2
Code versioning system used	Git
Software code languages, tools, and services used	R
Compilation requirements, operating environments & dependencies	Can run on any operating system that is supported by R.
If available Link to developer documentation/manual	http://cran.uib.no/web/packages/fixedTimeEvents/fixedTimeEvents.pdf http://cran.uib.no/web/packages/fixedTimeEvents/vignettes/fixedTimeEvents.html
Support email for questions	kristian.liland@nmbu.no

1. Introduction

In some problems, especially in the domain of bioinformatics where gene clusters and operons are often studied [1], the data can be seen as a number of Bernoulli trials (defined as R) where the number of successes (defined as $r \geq 2$) is known. The scientific questions of interest are related to the distances between the successes; are they evenly distributed in the series or do they cluster? Available statistical distributions and software today can only give approximate solutions to the problem (see the following section, Problems and Background). Therefore a formal, statistical test procedure is needed.

[*] Correspondence to: Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, P.O. Box 5003, N-1432 Aas, Norway.
E-mail address: kristian.liland@nmbu.no (K.H. Liland).

http://dx.doi.org/10.1016/j.softx.2016.09.003

In this article we propose a distribution and accompanying test to solve these problems. Since the main areas of application are likely to be in bioinformatics and computational statistics, it is natural to make the procedures available in the form of an R package. We have implemented efficient procedures with simple user interfaces that cover all facets of the theory in an R package, called *fixedTimeEvents*.

2. Problems and background

There are several distributions that describe the number of successes over some interval in time or space, for example the Poisson, exponential, negative binomial and geometric distributions [2]. The Poisson distribution describes the probability of a given number of non-overlapping successes occurring in a fixed interval. With the exponential distribution, we find the distance ('waiting time') between successful events in the Poisson distribution. In its discrete version, the negative binomial distribution is often named the Pascal distribution. This gives the number of successes in a series of Bernoulli trials before a given number of failures have occurred. Likewise, the geometric distribution gives the distance between successful Bernoulli trials when there is no upper limit to the number of trials.

With discrete, non-overlapping trials, the probability of success in each Bernoulli trial is defined to be $r/R = \#successes/\#trials$, where r is the number of successes and R is the number of trials. However, the extra limitation on the total number of trials, means the distance between two consecutive successes, X, does not have an exact geometric distribution. Here, the discrete random variable X is defined to be 1 if two consecutive successes are neighbours, 2 if there is a single failure between the successes and so on. For large R and r, the probability mass function of X can be approximated by a discretised version of the exponential distribution. However, this approximation is inaccurate for small R or r and it becomes worse as the ratio $r/R \to 1$. For these reasons, we present a distribution for X which is exact regardless of the sizes of R and r. Theorems with proofs (see the Appendix section) are included, and code samples based on the accompanying R package [3] are given in the Examples section.

Following the above definition of X as the distance between two consecutive successes in a fixed-length series of Bernoulli trials, we make some observations regarding the realisations of this distance. If one success follows the next, we observe the minimum distance: $min(X) = 1$. The maximum obtainable distance is $max(X) = R - r + 1$. This occurs if a single success is located in one end of the series of R events while all other successes are clustered at the opposite end. If $R = 5$ and $r = 3$, $R - r + 1 = 5 - 3 + 1 = 3$, leading to 3 possible realisations of X: $x \in 1, 2, 3$ (x is an observed distance in a sample). This is easily verified by looking at the results in Table 1.

We propose the following probability mass function for the distance X between consecutive successes when there are r successes in R trials:

$$P(X = x; R, r) = f(x; R, r) = \frac{\binom{R-x}{r-1}}{\binom{R}{r}} \tag{1}$$

for $x \in 1, \ldots, R - r + 1$, $r \geq 2$ and $R \geq r$, where the notation $\binom{n}{p}$ is used to signify the binomial coefficient, for example $\binom{R}{r} = R!/(r!(R - r)!)$.

This follows the ordinary argument of (#*successful outcomes*) /(#*possible outcomes*). It is easier to justify the function if one expands the fraction with $r - 1$. Starting with the denominator, this is simply the number of possible consecutive pairs in all unordered samples of r from R, i.e. $(r - 1) \cdot \binom{R}{r}$. The numerator, $(r - 1) \cdot \binom{R-x}{r-1}$, is the number of possible consecutive pairs with distance x among all possible ways of drawing r from R without replacement. In the Examples section this can be verified by counting the occurrences in a small example. We also observe that the fraction $\binom{R-x}{r-1} / \binom{R}{r}$ gives the proportion of trials where $r - 1$ is selected from the remainder of R when fixing the first success at position x.

All values of $P(X = x; R, r)$ are positive since they are defined as ratios of binomial coefficients. It can be shown that $\sum_{x=1}^{R-r+1} \binom{R-x}{r-1} = \binom{R}{r}$, such that the probability mass sums to 1 over all values of x. The most extreme values of x will have the following simple probabilities: $f(1; R, r) = r/R$ and $f(R-r+1; R, r) = 1/\binom{R}{r}$, which are the largest and smallest values of $f(x; R, r)$, respectively.

Theorem 1. *The mean value of the proposed distribution is:*

$$\mu = \frac{R+1}{r+1}. \tag{2}$$

The variance is of less interest as the distribution is highly asymmetric, but it can also be shown that a general expression is the following:

$$\sigma^2 = \frac{r(R+1)(R-r)}{(r+1)^2(r+2)}. \tag{3}$$

The most trivial example would be when the number of successful events equals the number of possible events ($R = r$) which leads to $\mu = 1$ and $\sigma^2 = 0$, which is obvious, as all distances between successes will be 1.

Theorem 2. *The recurrence relation for the proposed distribution for increasing values of x is:*

$$P(X = x + 1; R, r) = \frac{\binom{R-(x+1)}{r-1}}{\binom{R}{r}} = \left(1 - \frac{r-1}{R-x}\right) P(X = x; R, r). \tag{4}$$

Theorem 3. *The cumulative distribution function of the proposed distribution is:*